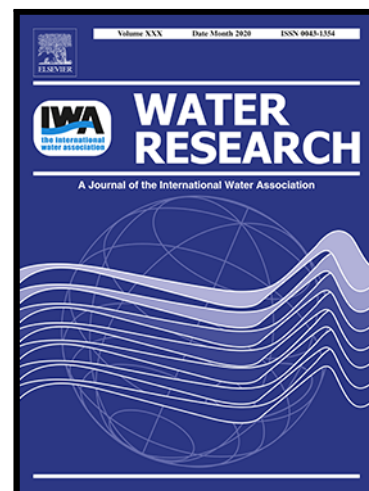


## Journal Pre-proof

Chemodiversity of riverine dissolved organic matter: Effects of local environments and watershed characteristics

Yifan Cui , Shuailong Wen , James C. Stegen , Ang Hu ,  
Jianjun Wang

PII: S0043-1354(23)01494-X  
DOI: <https://doi.org/10.1016/j.watres.2023.121054>  
Reference: WR 121054



To appear in: *Water Research*

Received date: 23 September 2023  
Revised date: 19 December 2023  
Accepted date: 21 December 2023

Please cite this article as: Yifan Cui , Shuailong Wen , James C. Stegen , Ang Hu , Jianjun Wang , Chemodiversity of riverine dissolved organic matter: Effects of local environments and watershed characteristics, *Water Research* (2023), doi: <https://doi.org/10.1016/j.watres.2023.121054>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2023 Published by Elsevier Ltd.

### Highlights

- Organic matter molecular richness decreased towards higher latitudes in sediments, but not waters.
- Precipitation and non-purgeable organic carbon strongly associated with organic matter.
- Watershed variables like land cover explained chemodiversity especially in waters.
- Relationships between watershed and organic matter molecules were dominantly positive in waters.
- Molecules positively and negatively related to watersheds had distinct stoichiometric ratios.

**Chemodiversity of riverine dissolved organic matter: Effects of local environments and watershed characteristics**

Yifan Cui<sup>a, 1</sup>, Shuailong Wen<sup>a, 1</sup>, James C. Stegen<sup>b</sup>, Ang Hu<sup>a, \*</sup>, Jianjun Wang<sup>a, \*</sup>

<sup>a</sup> State Key Laboratory of Lake Science and Environment, Nanjing Institute of Geography and Limnology, Chinese Academic of Sciences, Nanjing 210008, China

<sup>b</sup> Pacific Northwest National Laboratory, Richland, Washington 99352, United States

<sup>1</sup> Equal contribution to the manuscript

\*Corresponding author: Jianjun Wang, [jjwang@niglas.ac.cn](mailto:jjwang@niglas.ac.cn); Ang Hu, [anghu@niglas.ac.cn](mailto:anghu@niglas.ac.cn)

## Abstract

Riverine dissolved organic matter (DOM) is crucial to global carbon cycling and aquatic ecosystems. However, the geographical patterns and environmental drivers of DOM chemodiversity remain elusive especially in the waters and sediments of continental rivers. Here, we systematically analyzed DOM molecular diversity and composition in surface waters and sediments across 97 broadly distributed rivers using data from the Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS) consortium. We further examined the associations of molecular richness and composition with geographical, climatic, physicochemical variables, as well as the watershed characteristics. We found that molecular richness significantly decreased toward higher latitudes, but only in sediments ( $r = -0.24$ ,  $P < 0.001$ ). The environmental variables like precipitation and non-purgeable organic carbon showed strong associations with DOM molecular richness and composition. Interestingly, we identified that less-documented factors like watershed characteristics were also related to DOM molecular richness and composition. For instance, DOM molecular richness was positively correlated with the soil sand fraction for waters, while with the percentage of forest for sediments. Importantly, the effects of watershed characteristics on DOM molecular richness and composition were generally stronger in waters than sediments. This phenomenon was further supported by the fact that 11 out of 13 watershed characteristics (e.g., the percentages of impervious area and cropland) showed more positive than negative correlations with molecular abundance especially in waters. As the percentage of forest increased, there was a continuous accumulation of the compounds with higher molecular weight, aromaticity, and degree of unsaturation. In contrast, human activities accumulated the compounds with lower molecular weight and oxygenation, and higher bioavailability. Our findings imply that it may be possible to use a small set of broadly available data types to predict DOM molecular richness and composition across diverse river systems. Elucidation of mechanisms underlying these relationships will provide further enhancements to such

predictions, especially when extrapolating to unsampled systems.

**Keywords:** dissolved organic matter, chemodiversity, surface water, sediment, geographical pattern, molecular characteristics

Journal Pre-proof

## 1. Introduction

Global rivers carry  $0.95 \text{ Pg C yr}^{-1}$  from terrestrial inputs to coastal oceans and play a central role in biogeochemical cycles of carbon ([Battin et al., 2009](#); [Cole et al., 2007](#); [Regnier et al., 2013](#)). In rivers, dissolved organic matter is an important organic carbon component and contributes substantially to global  $\text{CO}_2$  emissions due to decomposition ([Stegen et al., 2018](#); [Ward et al., 2017](#); [Wohl et al., 2017](#)). The composition, quality and properties of riverine DOM assemblages, which determine the fate of organic carbon, are altered during transport and transformation processes along the river corridors ([Battin et al., 2008](#); [Mosher et al., 2015](#); [Riedel et al., 2016](#); [Zander et al., 2020](#)). Accordingly, the chemodiversity of DOM shows extreme heterogeneity in the vertical profiles and spatial patterns across riverine ecosystems ([He et al., 2016](#); [Li et al., 2023](#); [Wang et al., 2018](#); [Zhang et al., 2021](#)). For instance, DOM composition is significantly divergent between surface water and subsurface at a forest watershed ([Danczak et al., 2021](#)). More aromatic and unsaturated small molecules are found at higher than lower latitudes in Yenisei river ([Roth et al., 2013](#)). However, few studies report the geographical patterns and the differences of riverine DOM chemodiversity across water and sediment at large spatial scale.

The chemodiversity of DOM in aquatic ecosystems is mainly driven by geography, climate, hydrology, and physicochemical conditions. For instance, in lakes, DOM molecular composition is significantly correlated with temperature, precipitation, and water residence time ([Kellerman et al., 2014](#)). In coastal wetlands, the variation in DOM molecular composition is mainly explained by latitude, moisture, and organic matter contents ([Li et al., 2022](#)). In marine ecosystems, the diversity of DOM is found to decrease with increasing degradation time and oxygen concentration ([Chen et al., 2022](#); [Mentges et al., 2017](#)). Unlike the aforementioned habitats, rivers possess distinct features such as short water residence time and extensive connection with surrounding landscapes ([Dillon and Molot, 1997](#); [Hawkes et al., 2018](#); [Wagner et al., 2015](#)). Hydrological precipitation and snowmelt events further trigger the release of large

amounts of terrigenous DOM from basins to rivers ([Raymond et al., 2016](#)). These features potentially enhance the associations between riverine DOM chemodiversity and watershed characteristics ([He et al., 2022](#); [Sanderman et al., 2009](#); [Wagner et al., 2015](#)). Watershed characteristics such as land cover and watershed area are only recently reported to affect DOM chemodiversity. For instance, land cover plays a significant role in driving DOM diversity and composition in forested and anthropogenically impacted watershed ([Coble et al., 2019](#); [Roebuck et al., 2020](#)). The molecular diversity of DOM in a basin located in the northwestern USA increases with increasing upstream catchment area ([Danczak et al., 2023](#)). However, most of those studies are confined to water habitats at the regional scale, while left understudied is the continental-scale DOM chemodiversity in both water and sediment habitats, and specifically in context of diverse watershed characteristics.

Here, to reveal geographical patterns of riverine DOM chemodiversity and its association with environmental variables, we analyzed the DOM characteristics from 279 surface water and 272 sediment samples across 97 rivers collected by the WHONDRS consortium (Fig. 1). DOM molecules were identified using ultrahigh-resolution electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS). We examined two aspects of DOM chemodiversity: molecular richness and molecular composition. Molecular richness is one metric of molecular diversity and represents the number of molecular formulas in each sample. Molecular composition refers to the changes in abundance of each molecular formula across samples. Molecular richness could be related to microbial diversity by “diversity begets diversity” ([Hu et al., 2022](#); [Osterholz et al., 2018](#)). That is, molecular richness is relevant to how much the resources could be provided to microbes, and how much the organic matter could be produced by decomposition processes ([Judd et al., 2006](#); [Lehmann et al., 2020](#)). Importantly, molecular richness could also be related to ecosystem functions, such as the CH<sub>4</sub> emission and the decomposition rates of organic matter in lake sediments ([Tanentzap et al., 2019](#); [Wen et al., 2023](#)). We related DOM

chemodiversity to environmental variables such as geographical, climatic, physicochemical conditions, as well as watershed characteristics including watershed area, land cover, and soil texture. We addressed three main questions: (1) Are there general geographical patterns in molecular richness and composition? (2) What are the dominant environmental factors associated with molecular richness and composition of riverine DOM in waters and sediments? (3) Are there systematic differences in the chemistry of DOM molecules that are positively vs. negatively associated with watershed characteristics? Through evaluation of these questions, we could provide a more integrated understanding of DOM chemodiversity based on the largest data available so far. These findings could help predict the spatial-temporal variations of chemodiversity by a relatively small number of broadly available environmental variables in regional and global scales.

## **2. Materials and methods**

### **2.1. Sample collection and processing**

We obtained datasets from WHONDRS Summer 2019 Sampling Campaign, a part of WHONDRS consortium from Pacific Northwest National Laboratory (PNNL) (Stegen and Goldman, 2018). Details on datasets are available from the Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) (<https://data.ess-dive.lbl.gov/data>). Briefly, during 29 July and 19 September 2019, surface water and sediment samples were collected from 97 river corridor systems in eight countries: United States, Canada, Israel, Germany, Italy, Norway, United Kingdom, and South Korea (Fig. 1) (Goldman et al., 2020; Toyoda et al., 2020). The sampling sites spanned the latitude of 18.11° to 68.64° N and the elevation of -164 to 3,049 m. The mean annual temperature (MAT) at the sampling sites ranged from -9.8 to 24.5 °C, and the mean annual precipitation (MAP) ranged from 176 to 2,524 mm. The sampled rivers covered tributaries and main streams, with watershed area ranging from 0.08 to 262,525 km<sup>2</sup>. Surface water samples were collected in triplicate using 60

mL syringe and filtered through 0.22  $\mu\text{m}$  sterivex filter (EMD Millipore) into pre-acidified 40 mL amber glass vial (I-Chem amber VOA glass vials; ThermoFisher, pre-acidified with 10  $\mu\text{L}$  of 85% phosphoric acid). To cover the heterogeneity in river sediments, three replicated sediments at 1-3 cm depth were collected at three adjacent depositional zones (intervals within 10 m) within each sampling site following National Ecological Observatory Network (NEON) protocol (NEON.DOC.001193) (Jensen, 2019), and labeled as upstream, midstream, and downstream along the flow direction in rivers. A sterilized metal scoop was used to collect 125 mL saturated sediment samples from an approximately 1  $\text{m}^2$  region at each depositional zone. The samples were transported to PNNL on blue ice within 24 h of collection. Upon arrival, surface water samples were instantly frozen at  $-20\text{ }^\circ\text{C}$  until analysis, and sediment samples were individually sieved to  $< 2\text{ mm}$ , subsampled, and stored at  $-20\text{ }^\circ\text{C}$ .

## 2.2. FT-ICR MS data analysis

Prior to FT-ICR MS analysis, all samples were thawed in a dark environment at  $4\text{ }^\circ\text{C}$  for 72 h. Sediment dissolved organic matter was extracted by continuously shaking the tubes in the dark at 375 rpm and  $21\text{ }^\circ\text{C}$  for 2 h, centrifuging at 6,000 ref and  $21\text{ }^\circ\text{C}$  for 5 min and filtering supernatant through 0.22  $\mu\text{m}$  polyethersulfone membrane filter (Millipore Sterivex, USA). Non-purgeable organic carbon (NPOC) in waters and sediments was measured using a Shimadzu combustion carbon analyzer TOC-L CSH/CSN E100V with ASI-L autosampler. NPOC concentrations for all samples were normalized to  $1.5\text{ mg C L}^{-1}$  by diluting with Milli-Q deionized water. The diluted samples were further acidified to pH 2 with 85% phosphoric acid and solid-phase extracted using PPL (Priority PolLutant, Bond Elut) cartridges with methanol for final elution (Garayburu-Caruso et al., 2020). PPL was an efficient and widely used solid-phase sorbent, and the recovery rate was about 62% (Dittmar et al., 2008). It should be noted that technical limitations in approaches may affect the estimation of molecular richness due to the incomplete coverage of molecules. However, this would less likely

affect our main conclusions, largely because consistent approaches were used for all samples to allow for the comparisons across sites and regions.

To conduct highly accurate mass measurements of DOM in both waters and sediments, a 12 Tesla Bruker Solarix FT-ICR MS (Bruker, Solarix, Billerica, MA, USA) located at the Environmental Molecular Science Laboratory in Richland, WA was used with a resolution of 220 K at 481.185  $m/z$  (Garayburu-Caruso et al., 2020). The FT-ICR MS was coupled with a standard electrospray ionization source in negative ionization mode with a voltage of +4.2 kV. The instrument was externally calibrated weekly to ensure mass accuracy of less than 0.1 ppm. Data were collected with an ion accumulation of 0.05 s for waters and 0.1 or 0.2 s for sediments, and scanned 144 times over a range of 100–900  $m/z$  at 4 M (Garayburu-Caruso et al., 2020). The FT-ICR mass spectra were internally calibrated using organic matter homologous series separated by 14 Da ( $-CH_2$  groups). The mass measurement accuracy was typically within 1 ppm for singly charged ions across a broad  $m/z$  range (100–900  $m/z$ ) (Koch et al., 2007).

To convert raw spectra to a list of  $m/z$  values, Bruker Daltonics Data Analysis software (v4.2) was used. Fourier transform mass spectrometry (FT-MS) peak picker module was applied with a signal-to-noise ratio (S/N) threshold set to 7 and an absolute intensity threshold to the default value of 100. Then, chemical formulas were assigned to peaks using the software Formularity (v1.0) according to the Compound Identification Algorithm with the following criteria:  $S/N > 7$  and mass error less than 1 ppm for a given chemical formula between the measured mass and the theoretical mass (Kujawinski and Behn, 2006; Tolić et al., 2017). This algorithm considers the presence of C, H, O, N, S, and P and excludes other elements or an isotopic signature, and non-oxygen hetero-atoms are limited to  $N_{0-4}$ ,  $S_{0-2}$ , and  $P_{0-1}$  (Koch et al., 2007). All formula assignments were further screened to meet the following criteria (Hu et al., 2022; Koch et al., 2007): (1) formulae with an odd number of nitrogen atoms had an even nominal  $m/z$ , while formulae with an even number of nitrogen atoms had an odd nominal  $m/z$ ; (2) the number of hydrogen atoms was at least 1/3 of carbon but not exceed  $2C + N +$

2; (3) the number of nitrogen or oxygen atoms could not exceed the number of carbon atoms; (4) the ratio of O/C was set to  $0-1$ ,  $H/C \geq 0.3$ ,  $N/C \leq 1$ , and double bond equivalents (DBE)  $\geq 0$ .

The assigned molecules were classified into eight compound classes based on van Krevelen diagrams (Kim et al., 2003), which were lipids (O/C =  $0-0.3$ , H/C =  $1.5-2.0$ ), proteins (O/C =  $0.3-0.55$ , H/C =  $1.5-2.2$ , N/C  $\geq 0.05$ ), amino sugars (O/C =  $0.55-0.67$ , H/C =  $1.5-2.2$ , N/C  $\geq 0.05$ ), carbohydrates (Carb; O/C =  $0.67-1.1$ , H/C =  $1.5-2.0$ ), unsaturated hydrocarbons (UnsatHC; O/C =  $0-0.1$ , H/C =  $0.7-1.5$ ), lignin (O/C =  $0.1-0.67$ , H/C =  $0.7-1.5$ ), tannin (O/C =  $0.67-1.1$ , H/C =  $0.5-1.5$ ) and condensed aromatics (ConHC; O/C =  $0-0.67$ , H/C =  $0.2-0.7$ ).

### 2.3. Physiochemical, climatic and watershed variables

The physiochemical properties of each sample were examined by WHONDRS consortium. Briefly, for waters, total nitrogen (TN) and dissolved inorganic carbon (DIC) were measured as detailed in Toyoda et al. (2020). For sediments, total organic carbon content, total nitrogen content, and grain size distribution from the  $< 2$  millimeter fraction of sediment were measured as shown in Goldman et al. (2020). Moreover, water temperature at 50% depth and pH values were measured at each sampling site. More detailed information about WHONDRS and the methods used are available at <https://whondrs.pnnl.gov>. We obtained climatic variables, namely, MAT and MAP of sampling sites from the WorldClim version 2 database with  $\sim 1$  km resolution (Fick and Hijmans, 2017), which is available at <https://www.worldclim.org/>.

We collected 13 watershed characteristics that were related to watershed area, nine land cover types, and three soil textures within a catchment. Watershed boundaries and area were calculated using ArcMap v10.2. Land cover data with 30 m resolution were obtained from the Finer Resolution Observation and Monitoring of Global Land Cover (FROM-GLC) (<http://data.starcloud.pcl.ac.cn/>) (Li et al., 2017), which includes ten land cover types: cropland, forest, grassland, shrubland, wetland, water body, tundra,

impervious area, bare land, snow and ice. Note that tundra was not considered here, as less than 8% of catchments contain this land cover type. The percentage of each land cover type within a catchment was calculated in ArcMap v10.2. Soil texture, including soil sand fraction, soil clay fraction, and soil silt fraction, was extracted from the Harmonized World Soil Database v1.2 with ~1 km resolution (Fischer et al., 2008), which is available from the Food and Agriculture Organization of the United Nations (FAO) soils portal (<https://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/>).

#### 2.4. Statistical analyses

To understand the geographical patterns of DOM, we considered two aspects of chemodiversity: molecular richness and molecular composition. For molecular richness, we used linear models to explore the relationships between richness and latitude or elevation. We examined the differences in molecular richness between waters and sediments using Wilcoxon rank-sum test (Legendre and Legendre, 2012). For molecular composition, we quantified the variations along geospatial distances using a distance-based approach (Soininen et al., 2007; Tuomisto and Ruokolainen, 2006). Briefly, we examined the distance-decay relationships using least-squares linear regression models by considering Bray-Curtis similarity against the changes in geographical distance. Compositional similarity and geographical distance were log-transformed to fulfill normality criteria, and the significance of the relationships was determined using Mantel tests with Pearson correlation by 9,999 permutations (Legendre and Legendre, 2012). To visualize the molecular composition of DOM between waters and sediments, we performed non-metric multidimensional scaling (NMDS) based on Bray-Curtis distance metric (Clarke, 1993). NMDS visually represents a set of objects on a predetermined number of axes while preserving the inherent ordering relationship between them. The difference in molecular composition between waters and sediments was examined by permutational multivariate analysis of

variance (PERMANOVA) with 999 permutations ([Anderson, 2001](#)). We used the first axis of non-metric multidimensional scaling for waters and sediments to represent their molecular composition.

To identify the main drivers of DOM molecular richness and composition, we performed various statistical methods, such as Pearson correlation ([Legendre and Legendre, 2012](#)), random forest analysis ([Breiman, 2001](#)), and variation partitioning analysis (VPA) ([Borcard et al., 1992](#)). First, we calculated pairwise Pearson correlation coefficients between explanatory variables and molecular richness or composition in waters and sediments. We performed log-transformed if the data were not normally distributed. Thus, we could identify the explanatory variables significantly associated with molecular richness and composition. Second, we used random forest analysis to assess the relative importance of each explanatory variable and to identify the primary factors predicting molecular richness and composition. We examined explanatory variables including geographical, climatic, watershed, and physicochemical variables as detailed in Table S1. The relative importance of explanatory variables was determined by evaluating the increase of mean square error, which means decrease in prediction accuracy, and 2,000 trees were produced using cross-validation ([Elith et al., 2008](#)). Third, we used variation partitioning analysis to quantify the relative contributions of different categories of driving factors to DOM molecular richness and composition ([Borcard et al., 1992](#)). The explanatory variables were grouped by geographical, climatic, watershed, and physicochemical variables (Table S1). We selected explanatory variables by stepwise regression with Akaike information criterion (AIC) ([Sakamoto et al., 1986](#)).

To identify the driving factors of DOM at molecular level, we calculated the Spearman rank correlation coefficients between relative peak abundance of individual molecules and each explanatory variable ([Kellerman et al., 2014](#); [Luo et al., 2022](#)). We calculated the mean values of significant positive and negative correlation coefficients to quantify the effects of drivers for DOM assemblages. To avoid highlighting spurious

or weak correlations, we considered only the molecules that were present in at least 20% of the samples. To examine the differences in the chemistry of DOM molecules that are positively and negatively associated with watershed characteristics, we calculated 16 molecular traits. These traits included mass, the number of carbon (C) atoms, aromaticity index (AI), the modified aromaticity index ( $AI_{mod}$ ), double bond equivalents (DBE), double bond equivalents minus oxygen ( $DBE_O$ ), double bond equivalents minus aromaticity index ( $DBE_{AI}$ ), standard Gibb's Free Energy of carbon oxidation (GFE), Kendrick Defect ( $kdefect_{CH_2}$ ), nominal oxidation state of carbon (NOSC),  $Y_{met}$ : carbon use efficiency, H/C ratio, O/C ratio; N/C ratio, P/C ratio, and S/C ratio, detailed in [Hu et al. \(2022\)](#). To reveal the main differences in DOM quality in two contrasting land cover types (i.e., forest and impervious area), we performed inter sample ranking analysis as described by [Herzprung et al. \(2012\)](#) and [Herzprung et al. \(2017\)](#). The detailed descriptions can be found in Supplementary Information 1.

To explore the underlying mechanisms driving the DOM molecular richness and composition in waters and sediments, we further used structural equation model (SEM) ([Grace et al., 2012](#)). Composite variables were generated to account for the combined effects of geographical, climatic, watershed, and physicochemical factors. For geographical variables, we used principal coordinates of neighborhood matrices (PCNM) to represent original spatial distance matrices as a set of orthogonal eigenvectors ([Borcard and Legendre, 2002](#)). The formulae for calculating the composite variables obtained by multiple regression were listed in Table S2. We established an initial model that accounted for all underlying causal pathways (Fig. S14), and further performed sequential models by dropping nonsignificant paths step by step (Table S3) ([Hu et al., 2020](#)). We selected the model with the lowest AIC value as the best-fitting model ([Grace et al., 2010](#)). Thus, we could disentangle the direct and indirect effects of explanatory variables on DOM molecular richness and composition via SEM. In addition, all variables were Z-score transformed before conducting the SEM to allow for meaningful comparison among multiple predictors.

We performed the statistical analyses in R environment with the packages *vegan* V2.6-4 (Dixon, 2003), *BiodiversityR* V2.15-2 (Kindt and Coe, 2005), *rfPermute* V2.5.1 (Archer, 2016), and *lavaan* V0.6-15 (Rosseel, 2012). It should be noted that we did not include into the above statistical analyses the samples from Norway, which were considered as outliers regarding their molecular richness and composition (data not shown).

### 3. Results

#### 3.1. Geographical patterns of DOM chemodiversity

In total, we identified 8,718 and 7,204 molecular formulas in water and sediment samples, respectively (Fig. S1). This is consistent with the significantly ( $P < 0.001$ ) higher molecular richness in waters than sediments, showing an average of  $1,674 \pm 440$  and  $1,161 \pm 348$  molecular formulas, respectively (Fig. S2). We also observed more unique formulas in waters than sediments, showing 4,628 and 3,114, respectively, and 4,090 common formulas in both habitats (Fig. S3).

We further observed that the DOM composition differed in two habitats and had a significant separation as shown in the NMDS plot (PERMANOVA,  $r^2 = 0.51$ ,  $P = 0.001$ ; Fig. S4). The Bray-Curtis similarity of DOM composition was significantly ( $P < 0.001$ ) higher in waters than sediments with average values of 0.70 and 0.52, respectively (Fig. S4).

We found that there were geographical patterns regarding molecular richness and composition across the studied large spatial scales. For molecular richness, there was a significantly decreasing latitudinal pattern in sediments ( $r = -0.24$ ,  $P < 0.001$ ), but nonsignificant in waters (Fig. 2a). The observed decrease in molecular richness with increasing latitude was also true for most of compound classes in sediments (Fig. S5). The elevational patterns, however, were nonsignificant for the richness in both habitats ( $P > 0.05$ ). For molecular composition, the pairwise Bray-Curtis similarities of DOM assemblages significantly ( $P < 0.05$ ) decreased with larger geographical distances,

showing significant distance-decay relationships in both waters and sediments (Fig. 2b). Similarly, there were significant ( $P < 0.05$ ) distance-decay relationships for most of compound classes in two habitats (Fig. S6).

### 3.2. Explanatory variables for DOM chemodiversity

DOM molecular richness and composition in both habitats were associated with geographical, climatic, and physicochemical variables (Fig. 3). For molecular richness in waters, it decreased significantly ( $P < 0.05$ ) with MAP and DIC, while increased with NPOC and the carbon to nitrogen ratio (C/N) (Figs. 3a, c, S7). The environmental factors mentioned above were also confirmed as important predictors of molecular richness according to random forest analysis (Fig. 3b). In sediments, molecular richness exhibited the strongest negative correlation with NPOC ( $r = -0.62$ ,  $P < 0.001$ ; Figs. 3a, c), and increased significantly ( $P < 0.05$ ) with precipitation and temperature (Figs. 3a, c, S8). For DOM molecular composition in waters, we found that the first axis of NMDS was significantly ( $P < 0.01$ ) correlated with elevation, water temperature, NPOC, and TN (Figs. 3a, c, S9). In sediments, molecular composition (characterized by the first axis of NMDS) was strongly correlated with NPOC ( $r = 0.57$ ,  $P < 0.001$ ; Figs. 3a, c), which was identified as the most important predictor via random forest analysis (Fig. 3b). The effects of NPOC on DOM molecular composition were involved in the relative abundance of organic carbon compound classes and functional traits (i.e., stoichiometric ratios, oxygenation, unsaturation, and bioavailability) at compositional level. For instance, the relative abundance of lipids and proteins significantly ( $P < 0.05$ ) increased while lignin and tannin decreased with increasing NPOC in sediments (Fig. S11). H/C, N/C, P/C, S/C and DBE<sub>O</sub> increased while O/C and  $Y_{\text{met}}$  decreased with increasing NPOC in sediments (Fig. S12).

Interestingly, DOM molecular richness and composition were further found to be closely related to watershed characteristics, including soil texture and land cover within the catchment. Specifically, the molecular richness in waters was significantly ( $P <$

0.001) increased with soil sand fraction and decreased with soil clay fraction (Figs. 3a, c, S7). In sediments, the molecular richness was more related to land cover than soil texture. For instance, the sediment molecular richness was positively correlated with the percentage of forest within the watershed, while negatively correlated with the percentage of grassland (Figs. 3a, c, S8).

We also found that watershed characteristics were strongly associated with the DOM composition in waters and sediments, while exhibiting stronger association in the former. In waters, DOM composition represented by the first axis of NMDS was significantly ( $P < 0.001$ ) correlated with the percentage of waterbody and impervious area (Figs. 3a, c, S9). Consistently, these watershed characteristics gained the most importance in predicting DOM composition through the random forest analysis (Fig. 3b). We further found that watershed factors explained the greatest variation of DOM composition in waters by accounting for 22.2% based on the VPA (Fig. S13). In sediments, watershed factors were not the most important but non-negligible for predicting DOM composition. Specifically, DOM composition represented by the first axis of NMDS was significantly ( $P < 0.001$ ) correlated with the percentage of grassland and soil silt fraction (Figs. 3a, c, S10), and watershed factors accounted for 9.6% of the variance in molecular composition as revealed by the VPA (Fig. S13).

To synthesize the above findings, we conducted structural equation model to illustrate how watershed factors and other explanatory factors influenced the molecular richness and composition of DOM assemblages in waters and sediments (Fig. 4). For molecular richness in waters, watershed factors had both direct and indirect effects through physicochemical factors (Fig. 4a), and showed a total effect of 0.33 which was comparable to the dominant physicochemical effect of 0.35 (Table S4). By contrast, in sediments, watershed factors influenced molecular richness only through weak direct effects, while no indirect effects via physicochemical factors (Fig. 4c). For DOM composition, watershed factors were dominant in explaining the first axis of NMDS of water DOM, showing direct and total effects of 0.37 and 0.40, respectively (Fig. 4b,

Table S4). In sediments, watershed factors had a relatively weak but non-negligible direct effect of 0.23, whereas physicochemical factors were the most important with a direct effect of 0.59 (Fig. 4d). Furthermore, geographical and climatic factors generally exerted indirect or nonsignificant effects on DOM composition in both habitats (Figs. 4b, d). These results collectively show that watershed factors were important for molecular richness and composition in waters, while less important but non-negligible in sediments.

### 3.3. DOM molecular associations with watershed characteristics

We further elucidated the associations of watershed characteristics with DOM in waters and sediments from the perspective of molecular level. Specifically, we visualized the significant Spearman rank correlation coefficients between the relative abundances of individual molecules and watershed characteristics in van Krevelen diagrams, and calculated the mean values of significant positive and negative correlation coefficients, respectively (Fig. 5).

Unexpectedly, watershed characteristics exerted both positive and negative associations with DOM molecules, while the relative strength was substantially different between waters and sediments. However, we found that in waters, positive relationships between watershed characteristics and the relative abundance of DOM molecules were dominant (Fig. 5a). For instance, 11 out of 13 watershed characteristics had stronger positive than negative correlations, such as the percentages of impervious, cropland, and wetland (Fig. S15). Even if we removed the relatively high correlated variables such as the percentage of forest, soil clay and silt fraction (Fig. S16), watershed characteristics (that is, 8 out of 10) were still dominated by the positive correlations with molecular abundance in waters. These findings were further confirmed by the fact that 57–89% of watershed characteristics showed stronger positive associations for individual compound classes, except for condensed aromatics (Fig. S17). In particular, high impervious area was associated with proteins and amino

sugars, while lignin and tannin were positively correlated with the percentage of forest (Figs. 5b, c). Altogether, these results show that watershed factors generally increased the relative molecular abundances of specific molecules of DOM assemblages in waters.

Unexpectedly, we found that more than half of watershed characteristics showed stronger negative than positive associations with DOM molecules in sediments (Figs. 5d, S15). This phenomenon was consistently observed across compound classes, with 46–69% of watershed characteristics exhibiting stronger negative than positive correlations (Fig. S17). For instance, the increase in the percentages of grassland and shrubland substantially decreased the abundances of lignin, tannin, and condensed aromatics (Figs. 5e, f). Collectively, watershed showed significant but divergent associations with DOM molecules between waters and sediments, that is, dominantly positive associations in the waters.

The DOM molecules which were positively and negatively correlated with specific watershed characteristics regarding their relative abundances had distinct stoichiometric ratios. For instance, the DOM molecules positively correlated with the percentage of forest had lower H/C, N/C, and S/C, higher O/C than those with negative correlations (Figs. 5c, 6a, c). The DOM molecules positively correlated with the percentage of impervious area showed lower O/C, higher H/C, N/C, and S/C compared to those with negative correlations (Figs. 5b, 6b, d). These results were consistent with other metrics such as mass,  $AI_{mod}$ ,  $Y_{met}$ , NOSC, and DBE, which could reflect the bioavailability, oxygenation, and unsaturation of molecules. For instance, the DOM molecules positively associated with the percentage of forest showed higher mass,  $AI_{mod}$ , and DBE than the molecules with negative associations (Figs. 6, S20). In contrast, the molecules positively correlated with the percentage of impervious area had lower mass, NOSC, and  $AI_{mod}$ , and higher  $Y_{met}$  than those with negative correlations (Figs. 6, S20). Collectively, natural (e.g., forests) and human-modified (e.g., impervious area) land covers within the watershed contributed to the differences in the chemistry of DOM molecules.

## 4. Discussion

The chemodiversity of riverine DOM exhibits considerable variability at broad spatial scales and is influenced by various factors. Here, we elucidated the geographical patterns and environmental correlates of molecular richness and composition of DOM assemblages for 279 surface water and 272 sediment samples across 97 broadly distributed rivers using the ultrahigh-resolution FT-ICR MS dataset from the WHONDRS consortium ([Goldman et al., 2020](#); [Stegen and Goldman, 2018](#); [Toyoda et al., 2020](#)). We found that DOM molecular richness decreased towards higher latitudes, but only in sediments. We further illustrated the power of watershed characteristics in explaining the chemodiversity of riverine DOM, which has rarely been considered at large spatial scales in previous studies. Examining patterns in molecular abundances also revealed that the chemistry of molecules that increase with specific watershed characteristics is distinct from the chemistry of molecules that decrease with those same watershed characteristics.

### 4.1. DOM molecular richness and composition are spatially structured

DOM molecular richness significantly decreased towards higher latitudes in sediments, while showed nonsignificant latitudinal patterns in waters (Fig. 2a). To our knowledge, few studies reported the geographical patterns of riverine DOM especially for these two habitats at continental-scale. One study that has examined large-scale spatial patterns found a shift in DOM chemical properties between eastern and western parts of the United States, but formal spatial analyses were not conducted ([Garayburu-Caruso et al., 2020](#)). Our results are also partly consistent with patterns in sediment DOM from coastal wetlands in China, where molecular diversity decreases with latitude ([Li et al., 2022](#)). In marine environments, however, there is no systematic pattern of molecular richness across water depths or latitudes in the Southern or Atlantic Oceans ([Mentges et al., 2017](#)). Additional work will be needed across additional regions and ecosystem types to determine if the latitudinal patterns seen here and in [Li et al.](#)

(2022) can be generalized. If general patterns emerge, knowledge of the governing mechanisms could be useful for predicting future organic matter patterns and dynamics.

Despite divergence in the absence vs. presence of latitudinal patterns for water vs. sediment DOM molecular richness, there were distance-decay relationships for the DOM in both waters and sediments (Fig. 2b). That is, the larger geographical distances between two samples, the lower similarity in molecular composition. Thus, DOM molecular composition also follows the first law of geography (Tobler, 1970), similar to biological communities that show decreased similarity with increasing geographical distance across rivers, lakes, and soils (Nekola and White, 1999; Wang et al., 2013). The above results generally indicate that the DOM at nearby sites share more similar molecular compositions which may be shaped by similar environments (e.g., climate, land cover, physicochemical properties). Moreover, we found that the distance-decay relationship had a steeper slope in sediments than in waters, which indicates that the similarity of DOM molecular composition decreased more rapidly with larger geographical distances in sediments. This phenomenon could be explained by the greater spatial heterogeneity of sediments than riverine surface waters, as evidenced by the fact that compared to waters, sediments are further affected by geological effects like parent rock and weathering (Hu et al., 2020), in addition to watershed land covers and climate.

We would like to note that most samples were from North America, but the inclusion of the samples from other regions (i.e., Europe and Asia) did not affect our main conclusion. For instance, DOM molecular richness showed also decreasing latitudinal pattern in sediments and not in waters if only samples from North America were considered (Fig. S23). Similarly, there were significant distance-decay relationships for the molecular composition in sediments and waters without samples from other regions (Fig. S24).

#### 4.2. Watershed characteristics have significant explanatory power for DOM chemodiversity

Together, our findings indicate that the molecular diversity and composition of riverine DOM assemblages in two habitats were associated with geo-climate and physiochemical conditions, consistent with previous work ([Kellerman et al., 2014](#); [Li et al., 2022](#); [Roth et al., 2013](#)). However, our study indicates greater effects of watershed characteristics than previously recognized. Watershed characteristics had direct effects on molecular diversity and composition in both waters and sediments, especially the dominant effects on molecular composition in waters (Fig. 4). The strong linkages between watershed characteristics and DOM chemodiversity have also been reported at a recent basin-scale work, which showed water DOM chemodiversity of Yakima River increases with increasing watershed area and varies with land covers ([Danczak et al., 2023](#)). The transferability of site-level results to large spatial scales is currently unknown, but our continental-scale analysis suggests that relationships between DOM chemodiversity and watershed characteristics may be common. The understanding of the main drivers especially the watershed characteristics could help us to predict the chemodiversity of riverine DOM at broad spatial scales. This is largely because direct sampling may be challenging or impossible in remote or inaccessible regions. The predictive model has the potential to acquire the temporal variation of DOM chemodiversity in the past or future by inputting watershed variables like land use changes.

Watershed characteristics have greater explanatory power to DOM chemodiversity of waters than sediments. Watershed characteristics explained 14.7% and 22.2% of the variance in molecular richness and composition in water DOM, respectively, while they explained only 4.9% and 9.6% in sediment DOM (Fig. S13). These are supported by the correlation between watershed characteristics and DOM molecular abundance, which was generally stronger in waters than in sediments. Moreover, the average values of correlation coefficients for positive correlation were

generally higher than for negative correlation within water habitats (Figs. 5a, d). Such relationships likely reflect the fact that water DOM assemblages are the result of integrated signature across the upstream catchment. Surface runoff carries allochthonous DOM within the watershed into rivers, including DOM originating from human activities and natural production, such as the disinfection by-products, combustion-derived black carbon, and lignin-like compounds derived from plant debris (Ding et al., 2020; Riedel et al., 2016). These terrestrial DOM, especially chromophoric DOM (e.g., macromolecular aromatics), are sensitive to light and could be degraded to smaller compounds during the DOM transport along the river corridor (Hu et al., 2023). Our results also show that in waters, molecules sensitive to photo-degradation like  $C_{20}H_{16}O_{14}$ ,  $C_{18}H_{16}O_{14}$ , and  $C_{19}H_{16}O_{14}$  were significantly positively correlated with the percentage of forest (Spearman's  $\rho \geq 0.37$ ; Fig. 5c, Table S6), indicating the direct influence of land cover on riverine DOM.

In sediments, organic matter transformations are decoupled from those in waters (Stegen et al., 2022). Compared to DOM assemblages in waters, sediment DOM is more strongly influenced by nearby abiotic or biotic processes, such as mineral adsorption and microbial transformation, rather than the watershed characteristics. For instance, carboxylic-containing and polyphenolic compounds are preferentially adsorbed by minerals (e.g., iron oxides) and buried in sediments, while unsaturated aliphatic compounds prefer to remain in waters (Sowers et al., 2019). The fractionation of DOM caused by selected adsorption led to the difference in DOM molecular composition between waters and sediments. In addition, compared with waters, sediments are more heterogeneous and can provide more microhabitats for microorganisms. The high microbial diversity of sediments could enhance transformation of DOM, such as the utilization and resynthesis of protein and amino acids (Danczak et al., 2021; Stegen et al., 2018), as well as the recalcitrant compounds (e.g.,  $C_{10}H_{10}O_6$ ,  $C_{10}H_{10}O_7$ , and  $C_{11}H_{12}O_7$ ) (Hertkorn et al., 2006). These abiotic and biotic factors could collectively outweigh the influence of watershed characteristics (e.g., land cover) on sediment DOM

chemodiversity.

The quantity and quality of organic carbon changes synergistically in sediment habitats, as evidenced by the fact that NPOC was the dominant predictor of DOM molecular composition (Fig. 3). Higher organic carbon content in sediments was accompanied by higher abundance of lipids and proteins, higher bioavailability (larger H/C and lower  $Y_{met}$ ), as well as higher number of heteroatoms such as N, P, and S (Figs. S11, S12). These links contribute to the potential of carbon decomposition and greenhouse gas emissions with higher organic carbon content (Stegen et al., 2023), and may have detrimental impacts on the riverine water quality, such as eutrophication and odour problems (Paul, 2008). Given the weak influence of watershed characteristics on molecular composition of sediments, the mechanism of organic carbon turnover in sediments and riverine ecosystem functioning should be further investigated in depth.

#### **4.3. Distinct characteristics between DOM molecules positively correlated to natural and human-modified land cover**

To gain further insights into mechanisms of watershed characteristics, especially land cover, we evaluated how the DOM chemistry of individual formulas was associated with the direction of correlations between peak intensity and land cover conditions. While this analysis was done for both water and sediment DOM and across a broad range of environmental variables (Figs. 5a, d), we focus here on relationships for water DOM. This is because DOM chemodiversity was more strongly associated with watershed characteristics in waters than sediments. We also focus on the percentages of forest and impervious area as two contrasting land cover types to reflect natural and human-associated processes.

The molecules positively correlated with the percentage of forest were dominated by lignin-like (43.0%), condensed aromatics (32.2%), and tannin-like (22.5%) compounds, while the negatively correlated molecules had higher proportion of aliphatic compounds (28.5%,  $1.5 \leq H/C \leq 2.0$ ; Table S8). This is consistent with the

inter sample ranking analysis that DOM molecules within the top 50 inter sample ranks (relatively high peak intensities) were dominated by lignin-like and oxygen-rich tannin-like compounds in forest-dominated watersheds (Fig. S22). The molecules positively correlated with the percentage of forest had lower H/C ( $< 1.5$ ), higher  $AI_{mod}$ , molecular weight (most molecules  $> 400$  Da), and degree of unsaturation (higher DBE) than the molecules with negative correlations (Figs. 5c, 6a, S20), indicating that upstream catchments with more forest cover may contribute to DOM which is relatively biological inertia and has high reactivity related to photo-degradation (D'Andrilli et al., 2015; Riedel et al., 2016; Zhou et al., 2021). For instance, known photo-degraded components, such as  $C_{18}H_{16}O_{14}$  and  $C_{19}H_{16}O_{14}$  (Herzprung et al., 2023), had relatively high peak intensities (inter sample ranks in the top 80) in forest-dominated watersheds (Supplementary Information 2).

The molecules positively correlated with the percentage of impervious area displayed higher proportion of aliphatic compounds (23.5%) than the molecules with negative correlations (0.8%, Table S7). These positively correlated molecules showed lower molecular weight (most molecules  $< 400$  Da),  $AI_{mod}$ , and oxygenation (indicated by lower NOSC and O/C), higher H/C and N-content (indicated by higher N/C) (Figs. 5b, 6b, S20), and are considered to be more bioavailable and less photo-degradable (Hu et al., 2022; Li et al., 2023). In the watersheds with the percentage of impervious area greater than 25%, van Krevelen diagrams derived from inter sample rankings analysis (Fig. S21) showed similar patterns to the van Krevelen diagrams derived from Spearman rank correlations between molecular relative abundances and impervious area. Specifically, in the above watersheds, the molecules that positively correlated with impervious area were dominated by the top 100 inter sample ranks. Taken together, the above results indicate that shifts in land cover from natural (e.g., forests) to human-modified (e.g., impervious area) are associated with systematic changes in DOM quantity and quality, i.e., a shift in DOM from biologically inert and photosensitive to bioavailable and photo-insensitive.

## 5. Conclusions

Our study reveals the geographical patterns of DOM molecular diversity and composition in riverine systems at a large spatial scale, and provides new insights into the driving forces of chemodiversity in waters and sediments. Molecular richness showed decreasing latitudinal pattern only in sediments. Molecular composition showed significant distance-decay relationships in both waters and sediments. We emphasized that the connections between DOM chemodiversity and watershed characteristics were stronger in waters than in sediments. This indicates that DOM chemodiversity in waters is governed by environmental processes that are distinct from that in sediments. Furthermore, the systematic nature of patterns shown between DOM molecules and natural or human-modified landscapes emerged from data spanning broad geographical and environmental conditions. This indicates that there may be globally coherent linkages between land cover and numerous river water regarding DOM properties such as molecular stoichiometries, thermodynamic properties, and structural elements. This is remarkable given the uncontrolled nature of field sampling scheme, whereby no attempts were made to focus on land cover gradients while controlling for confounding variables that could diminish otherwise strong connections between land cover and DOM chemistry. We suggest that especially for water DOM, watershed characteristics could be used to predict highly detailed properties of DOM chemistry. If true, such predictions could be used to drive by mechanistic models aimed at predicting river corridor biogeochemical function via explicit representation of DOM chemistry. Our work provides a key step towards generating global predictions of riverine DOM chemical properties and contributes to better predictions of future patterns/dynamics of riverine carbon and nutrient cycling.

### Author contributions

A.H., and J.W. designed the research. Y.C. and S.W. analyzed the data with the contributions from A.H. and J.W. Y.C. wrote the first draft of the manuscript. Y.C., S.W.,

and J.W. finished the manuscript with the comments from J.Stegen. and A.H. All authors contributed to the intellectual development of this study. Y.C. and S.W. contributed equally to this paper.

### **Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### **Acknowledgements**

This study was supported by National Natural Science Foundation of China (42225708, 92251304, 42377122, 42077052, 42307323), and Science and Technology Planning Project of NIGLAS (NIGLAS2022GS09). JCS was supported by a United States Department of Energy (US DOE) Early Career Award (grant 74193) at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the US DOE under contract DE-AC05-76RL01830. We are grateful to Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS) consortium for field sampling and processing.

## References

- Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. *Austral. Ecol.* 26(1), 32-46.
- Archer, E. 2016. rfPermute: Estimate permutation p-values for Random Forest importance metrics. Retrieved from <https://CRAN.R-project.org/package=rfPermute>.
- Battin, T.J., Kaplan, L.A., Findlay, S., Hopkinson, C.S., Marti, E., Packman, A.I., Newbold, J.D. and Sabater, F. 2008. Biophysical controls on organic carbon fluxes in fluvial networks. *Nat. Geosci.* 1(2), 95-100.
- Battin, T.J., Luysaert, S., Kaplan, L.A., Aufdenkampe, A.K., Richter, A. and Tranvik, L.J. 2009. The boundless carbon cycle. *Nat. Geosci.* 2(9), 598-600.
- Borcard, D. and Legendre, P. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol. Model.* 153(1), 51-68.
- Borcard, D., Legendre, P. and Drapeau, P. 1992. Partialling out the spatial component of ecological variation. *Ecology* 73(3), 1045-1055.
- Breiman, L. 2001. Random forests. *Machine Learning* 45(1), 5-32.
- Chen, X., Liu, J., Chen, J., Wang, J., Xiao, X., He, C., Shi, Q., Li, G. and Jiao, N. 2022. Oxygen availability driven trends in DOM molecular composition and reactivity in a seasonally stratified fjord. *Water Res.* 220, 118690.
- Clarke, K.R. 1993. Non-parametric multivariate analyses of changes in community structure. *Aust. J. Ecol.* 18(1), 117-143.
- Coble, A.A., Koenig, L.E., Potter, J.D., Parham, L.M. and McDowell, W.H. 2019. Homogenization of dissolved organic matter within a river network occurs in the smallest headwaters. *Biogeochemistry* 143(1), 85-104.
- Cole, J.J., Prairie, Y.T., Caraco, N.F., McDowell, W.H., Tranvik, L.J., Striegl, R.G., Duarte, C.M., Kortelainen, P., Downing, J.A., Middelburg, J.J. and Melack, J. 2007. Plumbing the global carbon cycle: Integrating inland waters into the

- terrestrial carbon budget. *Ecosystems* 10(1), 172-185.
- D'Andrilli, J., Cooper, W.T., Foreman, C.M. and Marshall, A.G. 2015. An ultrahigh-resolution mass spectrometry index to estimate natural organic matter lability. *Rapid Commun. Mass Spectrom.* 29(24), 2385-2401.
- Danczak, R.E., Garayburu-Caruso, V.A., Renteria, L., McKeever, S.A., Otenburg, O.C., Grieger, S.R., Son, K., Kaufman, M.H., Fulton, S.G., Roebuck, J.A., Myers-Pigg, A.N. and Stegen, J.C. 2023. Riverine organic matter functional diversity increases with catchment size. *Front. Water.* 5, 1087108.
- Danczak, R.E., Goldman, A.E., Chu, R.K., Toyoda, J.G., Garayburu-Caruso, V.A., Tolić, N., Graham, E.B., Morad, J.W., Renteria, L., Wells, J.R., Herzog, S.P., Ward, A.S. and Stegen, J.C. 2021. Ecological theory applied to environmental metabolomes reveals compositional divergence despite conserved molecular properties. *Sci. Total Environ.* 788, 147409.
- Dillon, P.J. and Molot, L.A. 1997. Effect of landscape form on export of dissolved organic carbon, iron, and phosphorus from forested stream catchments. *Water Resour. Res.* 33(11), 2591-2600.
- Ding, Y., Shi, Z., Ye, Q., Liang, Y., Liu, M., Dang, Z., Wang, Y. and Liu, C. 2020. Chemodiversity of soil dissolved organic matter. *Environ. Sci. Technol.* 54(10), 6174-6184.
- Dittmar, T., Koch, B., Hertkorn, N. and Kattner, G. 2008. A simple and efficient method for the solid-phase extraction of dissolved organic matter (SPE-DOM) from seawater. *Limnol. Oceanogr. Methods* 6(6), 230-235.
- Dixon, P. 2003. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* 14(6), 927-930.
- Elith, J., Leathwick, J.R. and Hastie, T. 2008. A working guide to boosted regression trees. *J. Anim. Ecol.* 77(4), 802-813.
- Fick, S.E. and Hijmans, R.J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* 37(12), 4302-4315.

- Fischer, G., Nachtergaele, F., Prieler, S., Van Velthuisen, H., Verelst, L. and Wiberg, D. 2008. Global agro-ecological zones assessment for agriculture (GAEZ 2008). IIASA, Laxenburg, Austria and FAO, Rome, Italy 10.
- Garayburu-Caruso, V.A., Danczak, R.E., Stegen, J.C., Renteria, L., McCall, M., Goldman, A.E., Chu, R.K., Toyoda, J., Resch, C.T., Torgeson, J.M., Wells, J., Fansler, S., Kumar, S. and Graham, E.B. 2020. Using community science to reveal the global chemogeography of river metabolomes. *Metabolites* 10(12), 518.
- Goldman, A.E., Chu, R.K., Danczak, R.E., Daly, R.A., Fansler, S., Garayburu-Caruso, V.A., Graham, E.B., McCall, M.L., Ren, H. and Renteria, L. 2020. WHONDRS Summer 2019 Sampling Campaign: Global River Corridor Sediment FTICR-MS, NPOC, and Aerobic Respiration. Environmental System Science Data Infrastructure for a Virtual Ecosystem, Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS).
- Grace, J.B., Anderson, T.M., Olf, H. and Scheiner, S.M. 2010. On the specification of structural equation models for ecological systems. *Ecol. Monogr.* 80(1), 67-87.
- Grace, J.B., Schoolmaster Jr., D.R., Guntenspergen, G.R., Little, A.M., Mitchell, B.R., Miller, K.M. and Schweiger, E.W. 2012. Guidelines for a graph-theoretic implementation of structural equation modeling. *Ecosphere* 3(8), art73.
- Hawkes, J.A., Radoman, N., Bergquist, J., Wallin, M.B., Tranvik, L.J. and Löfgren, S. 2018. Regional diversity of complex dissolved organic matter across forested hemiboreal headwater streams. *Sci. Rep.* 8(1), 16060.
- He, D., Li, P., He, C., Wang, Y. and Shi, Q. 2022. Eutrophication and watershed characteristics shape changes in dissolved organic matter chemistry along two river-estuarine transects. *Water Res.* 214, 118196.
- He, W., Chen, M., Park, J.-E. and Hur, J. 2016. Molecular diversity of riverine alkaline-extractable sediment organic matter and its linkages with spectral indicators and molecular size distributions. *Water Res.* 100, 222-231.

- Hertkorn, N., Benner, R., Frommberger, M., Schmitt-Kopplin, P., Witt, M., Kaiser, K., Kettrup, A. and Hedges, J.I. 2006. Characterization of a major refractory component of marine dissolved organic matter. *Geochim. Cosmochim. Acta* 70(12), 2990-3010.
- Herzprung, P., Kamjunke, N., Wilske, C., Friese, K., Boehrer, B., Rinke, K., Lechtenfeld, O.J. and von Tümpling, W. 2023. Data evaluation strategy for identification of key molecular formulas in dissolved organic matter as proxies for biogeochemical reactivity based on abundance differences from ultrahigh resolution mass spectrometry. *Water Res.* 232, 119672.
- Herzprung, P., von Tümpling, W., Hertkorn, N., Harir, M., Büttner, O., Bravidor, J., Friese, K. and Schmitt-Kopplin, P. 2012. Variations of DOM quality in inflows of a drinking water reservoir: Linking of van Krevelen diagrams with EEMF spectra by rank correlation. *Environ. Sci. Technol.* 46(10), 5511-5518.
- Herzprung, P., von Tümpling, W., Wendt-Potthoff, K., Hertkorn, N., Harir, M., Schmitt-Kopplin, P. and Friese, K. 2017. High field FT-ICR mass spectrometry data sets enlighten qualitative DOM alteration in lake sediment porewater profiles. *Org. Geochem.* 108, 51-60.
- Hu, A., Choi, M., Tanentzap, A.J., Liu, J., Jang, K.-S., Lennon, J.T., Liu, Y., Soininen, J., Lu, X., Zhang, Y., Shen, J. and Wang, J. 2022. Ecological networks of dissolved organic matter and microorganisms under global change. *Nat. Commun.* 13(1), 3600.
- Hu, A., Wang, J., Sun, H., Niu, B., Si, G., Wang, J., Yeh, C.-F., Zhu, X., Lu, X., Zhou, J., Yang, Y., Ren, M., Hu, Y., Dong, H. and Zhang, G. 2020. Mountain biodiversity and ecosystem functions: interplay between geology and contemporary environments. *ISME J.* 14(4), 931-944.
- Hu, J., Kang, L., Li, Z., Feng, X., Liang, C., Wu, Z., Zhou, W., Liu, X., Yang, Y. and Chen, L. 2023. Photo-produced aromatic compounds stimulate microbial degradation of dissolved organic carbon in thermokarst lakes. *Nat. Commun.*

14(1), 3681.

- Jensen, B. 2019 AOS Protocol and Procedure: Sediment Chemistry Sampling in Wadeable Streams (NEON. DOC. 001193).
- Judd, K.E., Crump, B.C. and Kling, G.W. 2006. Variation in dissolved organic matter controls bacterial production and community composition. *Ecology* 87(8), 2068-2079.
- Kellerman, A.M., Dittmar, T., Kothawala, D.N. and Tranvik, L.J. 2014. Chemodiversity of dissolved organic matter in lakes driven by climate and hydrology. *Nat. Commun.* 5(1), 3804.
- Kim, S., Kramer, R.W. and Hatcher, P.G. 2003. Graphical method for analysis of ultrahigh-resolution broadband mass spectra of natural organic matter, the van Krevelen diagram. *Anal. Chem.* 75(20), 5336-5344.
- Kindt, R. and Coe, R. 2005. Tree diversity analysis: a manual and software for common statistical methods for ecological and biodiversity studies, World Agroforestry Centre.
- Koch, B.P., Dittmar, T., Witt, M. and Kattner, G. 2007. Fundamentals of molecular formula assignment to ultrahigh resolution mass data of natural organic matter. *Anal. Chem.* 79(4), 1758-1763.
- Kujawinski, E.B. and Behn, M.D. 2006. Automated analysis of electrospray ionization Fourier transform ion cyclotron resonance mass spectra of natural organic matter. *Anal. Chem.* 78(13), 4363-4373.
- Legendre, P. and Legendre, L.F. 2012. *Numerical Ecology*, Elsevier.
- Lehmann, J., Hansel, C.M., Kaiser, C., Kleber, M., Maher, K., Manzoni, S., Nunan, N., Reichstein, M., Schimel, J.P., Torn, M.S., Wieder, W.R. and Kögel-Knabner, I. 2020. Persistence of soil organic carbon caused by functional complexity. *Nat. Geosci.* 13(8), 529-534.
- Li, C., Gong, P., Wang, J., Zhu, Z., Biging, G.S., Yuan, C., Hu, T., Zhang, H., Wang, Q., Li, X., Liu, X., Xu, Y., Guo, J., Liu, C., Hackman, K.O., Zhang, M., Cheng, Y.,

- Yu, L., Yang, J., Huang, H. and Clinton, N. 2017. The first all-season sample set for mapping global land cover with Landsat-8 data. *Sci. Bull.* 62(7), 508-515.
- Li, J., Wang, B., Yang, M., Li, W., Liu, N., Qi, Y. and Liu, C.-Q. 2022. Geographical constraints on chemodiversity of sediment dissolved organic matter in China's coastal wetlands. *Appl. Geochem.* 147, 105506.
- Li, S., Meng, L., Zhao, C., Gu, Y., Spencer, R.G.M., Álvarez-Salgado, X.A., Kellerman, A.M., McKenna, A.M., Huang, T., Yang, H. and Huang, C. 2023. Spatiotemporal response of dissolved organic matter diversity to natural and anthropogenic forces along the whole mainstream of the Yangtze River. *Water Res.* 234, 119812.
- Luo, J., Zhou, Q., Hu, X., Zeng, H., Deng, P., He, C. and Shi, Q. 2022. Lake chemodiversity driven by natural and anthropogenic factors. *Environ. Sci. Technol.* 56(9), 5910-5919.
- Mentges, A., Feenders, C., Seibt, M., Blasius, B. and Dittmar, T. 2017. Functional molecular diversity of marine dissolved organic matter is reduced during degradation. *Front. Mar. Sci.* 4, 194.
- Mosher, J.J., Kaplan, L.A., Podgorski, D.C., McKenna, A.M. and Marshall, A.G. 2015. Longitudinal shifts in dissolved organic matter chemogeography and chemodiversity within headwater streams: a river continuum reprise. *Biogeochemistry* 124(1), 371-385.
- Nekola, J.C. and White, P.S. 1999. The distance decay of similarity in biogeography and ecology. *J. Biogeogr.* 26(4), 867-878.
- Osterholz, H., Kirchman, D.L., Niggemann, J. and Dittmar, T. 2018. Diversity of bacterial communities and dissolved organic matter in a temperate estuary. *FEMS Microbiol. Ecol.* 94(8).
- Paul, V.J. 2008. Cyanobacterial Harmful Algal Blooms: State of the Science and Research Needs. Hudnell, H.K. (ed), pp. 259-273, Springer New York, New York, NY.

- Raymond, P.A., Saiers, J.E. and Sobczak, W.V. 2016. Hydrological and biogeochemical controls on watershed dissolved organic matter transport: pulse-shunt concept. *Ecology* 97(1), 5-16.
- Regnier, P., Friedlingstein, P., Ciais, P., Mackenzie, F.T., Gruber, N., Janssens, I.A., Laruelle, G.G., Lauerwald, R., Luysaert, S., Andersson, A.J., Arndt, S., Arnosti, C., Borges, A.V., Dale, A.W., Gallego-Sala, A., Godd ris, Y., Goossens, N., Hartmann, J., Heinze, C., Ilyina, T., Joos, F., LaRowe, D.E., Leifeld, J., Meysman, F.J.R., Munhoven, G., Raymond, P.A., Spahni, R., Suntharalingam, P. and Thullner, M. 2013. Anthropogenic perturbation of the carbon fluxes from land to ocean. *Nat. Geosci.* 6(8), 597-607.
- Riedel, T., Zark, M., V h talo, A.V., Niggemann, J., Spencer, R.G.M., Hernes, P.J. and Dittmar, T. 2016. Molecular signatures of biogeochemical transformations in dissolved organic matter from ten world rivers. *Front. Earth. Sci.* 4, 85.
- Roebuck, J.A., Jr., Seidel, M., Dittmar, T. and Jaff , R. 2020. Controls of land use and the river continuum concept on dissolved organic matter composition in an anthropogenically disturbed subtropical watershed. *Environ. Sci. Technol.* 54(1), 195-206.
- Rosseel, Y. 2012. lavaan: an R package for structural equation Modeling. *J. Stat. Softw.* 48(2), 1-36.
- Roth, V.-N., Dittmar, T., Gaupp, R. and Gleixner, G. 2013. Latitude and pH driven trends in the molecular composition of DOM across a north south transect along the Yenisei River. *Geochim. Cosmochim. Acta* 123, 93-105.
- Sakamoto, Y., Ishiguro, M. and Kitagawa, G. 1986. Akaike Information Criterion Statistics, Dordrecht, The Netherlands: D. Reidel.
- Sanderman, J., Lohse, K.A., Baldock, J.A. and Amundson, R. 2009. Linking soils and streams: Sources and chemistry of dissolved organic matter in a small coastal watershed. *Water Resour. Res.* 45(3), W03418.
- Soininen, J., McDonald, R. and Hillebrand, H. 2007. The distance decay of similarity

- in ecological communities. *Ecography* 30(1), 3-12.
- Sowers, T.D., Holden, K.L., Coward, E.K. and Sparks, D.L. 2019. Dissolved organic matter sorption and molecular fractionation by naturally occurring bacteriogenic iron (oxyhydr)oxides. *Environ. Sci. Technol.* 53(8), 4295-4304.
- Stegen, J.C., Fansler, S.J., Tfaily, M.M., Garayburu-Caruso, V.A., Goldman, A.E., Danczak, R.E., Chu, R.K., Renteria, L., Tagestad, J. and Toyoda, J. 2022. Organic matter transformations are disconnected between surface water and the hyporheic zone. *Biogeosciences* 19(12), 3099-3110.
- Stegen, J.C., Garayburu-Caruso, V.A., Danczak, R.E., Goldman, A.E., Renteria, L., Torgeson, J.M. and Hager, J. 2023. Maximum respiration rates in hyporheic zone sediments are primarily constrained by organic carbon concentration and secondarily by organic matter chemistry. *Biogeosciences* 20(14), 2857-2867.
- Stegen, J.C. and Goldman, A.E. 2018. WHONDERS: a community resource for studying dynamic river corridors. *mSystems* 3(5), e00151-00118.
- Stegen, J.C., Johnson, T., Fredrickson, J.K., Wilkins, M.J., Konopka, A.E., Nelson, W.C., Arntzen, E.V., Chrisler, W.B., Chu, R.K., Fansler, S.J., Graham, E.B., Kennedy, D.W., Resch, C.T., Tfaily, M. and Zachara, J. 2018. Influences of organic carbon speciation on hyporheic corridor biogeochemistry and microbial ecology. *Nat. Commun.* 9(1), 585.
- Tanentzap, A.J., Fitch, A., Orland, C., Emilson, E.J.S., Yakimovich, K.M., Osterholz, H. and Dittmar, T. 2019. Chemical and microbial diversity covary in fresh water to influence ecosystem functioning. *Proc. Natl. Acad. Sci.* 116(49), 24689-24695.
- Tobler, W.R. 1970. A computer movie simulating urban growth in the detroit region. *Econ. Geogr.* 46, 234-240.
- Tolić, N., Liu, Y., Liyu, A., Shen, Y., Tfaily, M.M., Kujawinski, E.B., Longnecker, K., Kuo, L.-J., Robinson, E.W., Paša-Tolić, L. and Hess, N.J. 2017. Formularity: Software for automated formula assignment of natural and other organic matter

- from ultrahigh-resolution mass spectra. *Anal. Chem.* 89(23), 12659-12665.
- Toyoda, J.G., Goldman, A.E., Chu, R.K., Danczak, R.E., Daly, R.A., Garayburu-Caruso, V.A., Graham, E.B., Lin, X., Moran, J.J. and Ren, H. 2020. WHONDRS Summer 2019 Sampling Campaign: Global River Corridor Surface Water FTICR-MS and Stable Isotopes. *Environmental System Science Data Infrastructure for a Virtual Ecosystem, Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS)*.
- Tuomisto, H. and Ruokolainen, K. 2006. Analyzing or explaining beta diversity? Understanding the targets of different methods of analysis. *Ecology* 87(11), 2697-2708.
- Wagner, S., Riedel, T., Niggemann, J., Vähätalo, A.V., Dittmar, T. and Jaffé, R. 2015. Linking the molecular signature of heteroatomic dissolved organic matter to watershed characteristics in world rivers. *Environ. Sci. Technol.* 49(23), 13798-13806.
- Wang, J.-J., Lafrenière, M.J., Lamoureux, S.F., Simpson, A.J., Gélinas, Y. and Simpson, M.J. 2018. Differences in riverine and pond water dissolved organic matter composition and sources in Canadian high arctic watersheds affected by active layer detachments. *Environ. Sci. Technol.* 52(3), 1062-1071.
- Wang, J., Shen, J., Wu, Y., Tu, C., Soininen, J., Stegen, J.C., He, J., Liu, X., Zhang, L. and Zhang, E. 2013. Phylogenetic beta diversity in bacterial assemblages across ecosystems: deterministic versus stochastic processes. *ISME J.* 7(7), 1310-1321.
- Ward, N.D., Bianchi, T.S., Medeiros, P.M., Seidel, M., Richey, J.E., Keil, R.G. and Sawakuchi, H.O. 2017. Where carbon goes when water flows: Carbon cycling across the aquatic continuum. *Front. Mar. Sci.* 4, 7.
- Wen, S., Hu, A., Jiang, S., Han, L., Jang, K.-S., Tanentzap, A.J., Zhong, J. and Wang, J. 2023. Temperature sensitivity of organic carbon decomposition in lake sediments is mediated by chemodiversity. *Glob Chang Biol.*
- Wohl, E., Hall Jr, R.O., Lininger, K.B., Sutfin, N.A. and Walters, D.M. 2017. Carbon

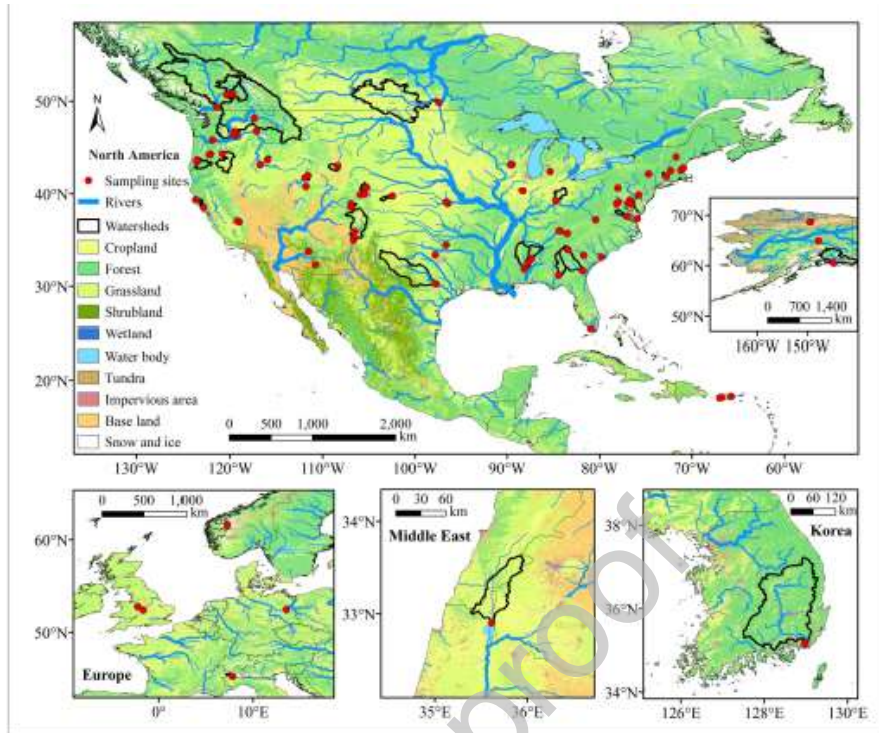
dynamics of river corridors and the effects of human alterations. *Ecol. Monogr.* 87(3), 379-409.

Zander, F., Heimovaara, T. and Gebert, J. 2020. Spatial variability of organic matter degradability in tidal Elbe sediments. *J. Soils Sediments* 20(6), 2573-2587.

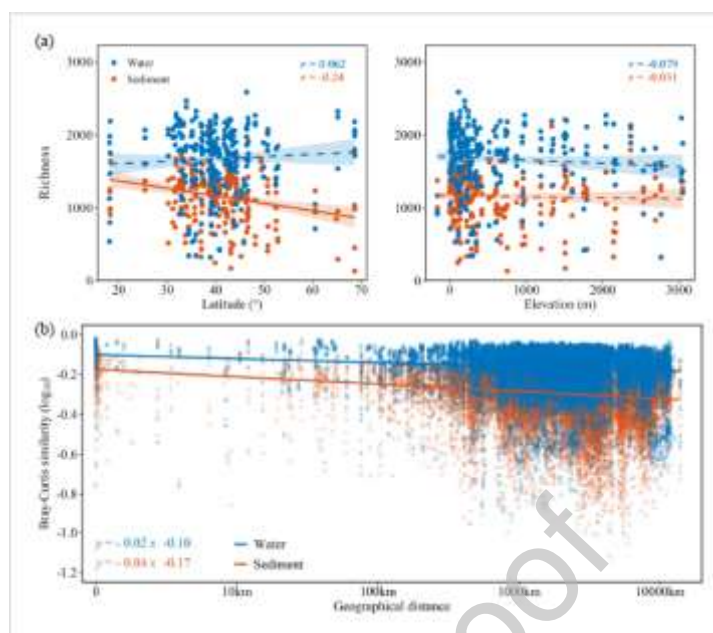
Zhang, P., Cao, C., Wang, Y.-H., Yu, K., Liu, C., He, C., Shi, Q. and Wang, J.-J. 2021. Chemodiversity of water-extractable organic matter in sediment columns of a polluted urban river in South China. *Sci. Total Environ.* 777, 146127.

Zhou, L., Zhou, Y., Tang, X., Zhang, Y., Jang, K.-S., Székely, A.J. and Jeppesen, E. 2021. Resource aromaticity affects bacterial community successions in response to different sources of dissolved organic matter. *Water Res.* 190, 116776.

## Figure legends



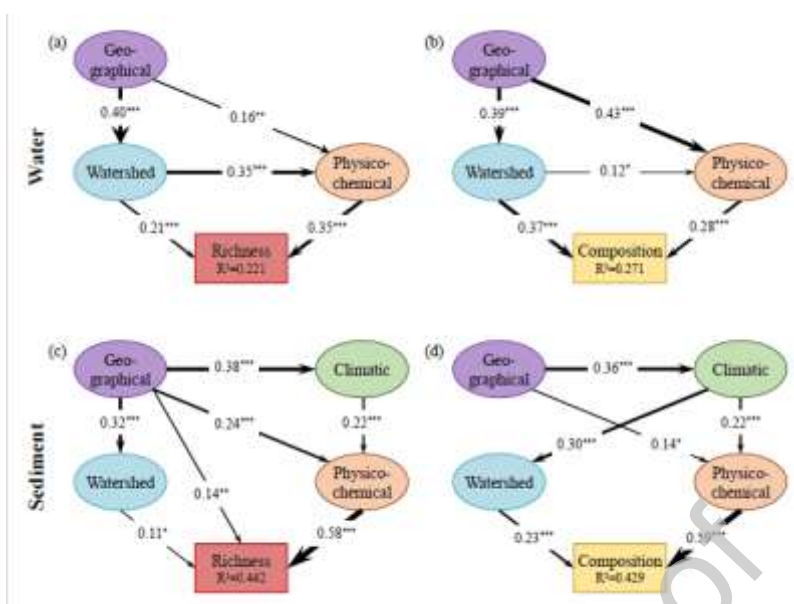
**Fig. 1.** Map of sampling sites in 97 global rivers. Sampling sites covered eight countries across North America, Europe, and Asia. Thick black lines denote the watershed boundaries of each sampling site. The categories of land cover are indicated in different colors.



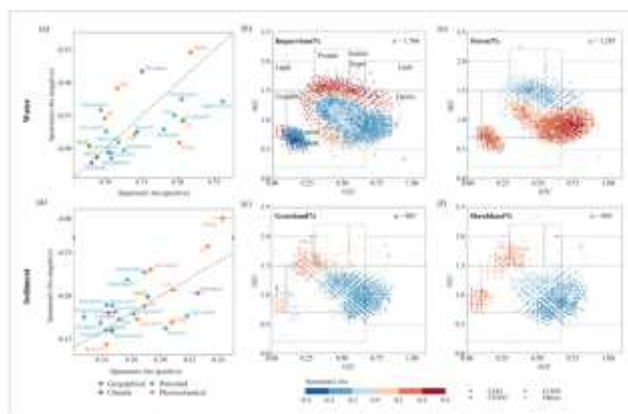
**Fig. 2.** Geographical patterns of DOM chemodiversity in waters and sediments. (a) Variations of molecular richness with latitude and elevation. Lines denote the least-squares linear regression across all water or sediment samples. Pearson correlation coefficients are listed in panels. Solid and dotted lines indicate statistically significant ( $P < 0.001$ ) and nonsignificant ( $P > 0.05$ ) relationships, respectively. The shadow area represents 95% confidence intervals. (b) Relationships between DOM molecular similarity based on Bray-Curtis metric ( $\log_{10}(X)$ ) and geographical distance ( $\log_{10}(X + 1)$ ). Solid lines indicate statistically significant ( $P < 0.05$ ) based on Mantel tests (Pearson's correlation) with 9,999 permutations.



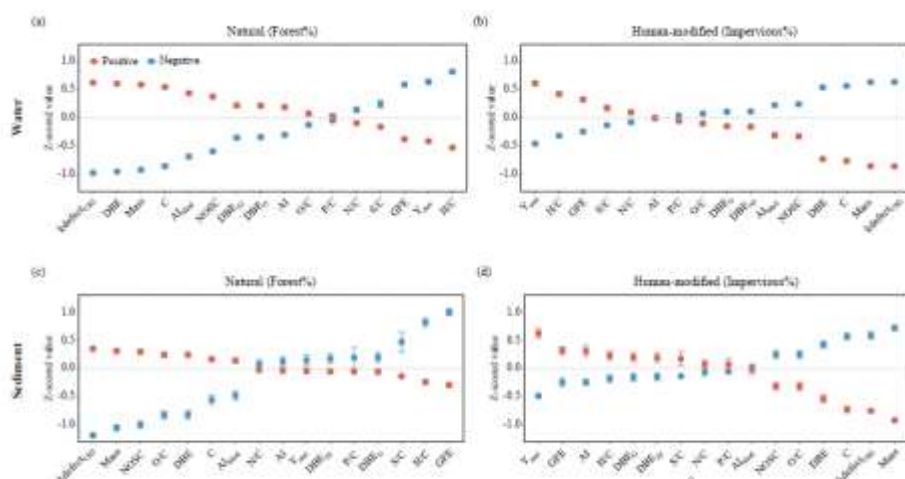
**Fig. 3.** Explanatory variables for DOM chemodiversity in waters and sediments. (a) Heatmap of Pearson correlation between DOM molecular richness or composition and explanatory variables. DOM composition is represented by the first axis of NMDS. (b) Major predictors of DOM molecular richness and composition. Increase of mean square error (MSE) of each predictive variable are acquired using random forest analysis. The predictive variables are grouped by geographical, climatic, watershed, and physicochemical variables with different colors. The asterisks denote the significance levels. \*\*\* $P < 0.001$ , \*\* $P < 0.01$ , \* $P < 0.05$ . (c) Relationships between DOM molecular richness or composition and explanatory variables. Solid and dotted lines indicate statistically significant ( $P < 0.05$ ) and nonsignificant ( $P > 0.05$ ) relationships, respectively. Pearson correlation coefficient is shown in each panel. MAT: mean annual temperature. MAP: mean annual precipitation. Area: watershed area. Temp: water temperature. NPOC: non-purgeable organic carbon. TN: total nitrogen. DIC: dissolved inorganic carbon.



**Fig. 4.** Structural equation models of DOM molecular richness and composition. Direct and indirect associations of geographical, climatic, watershed, and physicochemical factors with molecular richness and composition in waters (a, b) and sediments (c, d). Formulas to calculate composite variables are described in detail in Table S2. Model fit statistics and best-fitting model derivation process are summarized in Table S3. Black arrows indicate statistically significant ( $***P < 0.001$ ,  $**P < 0.01$ ,  $*P < 0.05$ ) relationships. Numbers near the arrows represent standardized path coefficients, and arrow width is proportional to path coefficients.  $R^2$  denotes the proportion of variance explained for endogenous variables.

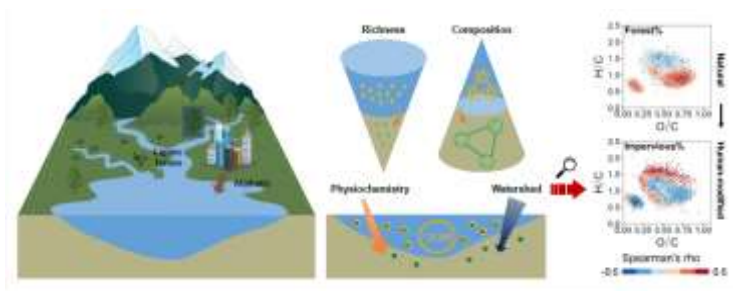


**Fig. 5.** Correlations between DOM molecules and explanatory variables. Mean significant ( $P < 0.05$ ) Spearman rank correlation coefficients of DOM molecules with each explanatory variable in waters (a) and sediments (d). The dotted lines denote the 1:1. Significant ( $P < 0.05$ ) Spearman rank correlation coefficients of individual molecules with the percentages of impervious area (b) and forest (c) in waters, and with the percentages of grassland (e) and shrubland (f) in sediments. The color indicates the direction and strength of the correlation, with positive in red and negative in blue. The shape indicates the element combination of DOM molecules: CHO (circle), CHNO (square), CHOS (triangle), and others (rhombus). The dot size is proportional to molecular mass. The dotted lines separating compound classes on van Krevelen diagrams are for visualization only. Other van Krevelen diagrams of correlations between DOM molecules and watershed characteristics could be found in Figs. S18 and S19. MAT: mean annual temperature. MAP: mean annual precipitation. Area: watershed area. Temp: water temperature. NPOC: non-purgeable organic carbon. TN: total nitrogen. DIC: dissolved inorganic carbon. Carb: carbohydrates. UnsatHC: unsaturated hydrocarbons. ConHC: condensed aromatics.



**Fig. 6.** Molecular traits of positively and negatively correlated molecular formulae with natural and human-modified land cover in waters (a, b) and sediments (c, d). The data are standardized to zero mean and unit variance. The t-test revealed that the means for all the traits within-pairs were significantly different ( $P < 0.05$ ), except for traits of P/C in panel (a), AI in panel (b), N/C,  $Y_{\text{met}}$ , AI in panel (c), P/C, N/C,  $AI_{\text{mod}}$  in panel (d).  $k\text{defect}_{\text{CH}_2}$ : Kendrick Defect. DBE: double bond equivalents. Mass: mass of DOM molecules. C: the number of carbon atoms.  $AI_{\text{mod}}$ : the modified aromaticity index. NOSC: nominal oxidation state of carbon.  $DBE_{\text{AI}}$ : double bond equivalents minus aromaticity index.  $DBE_{\text{O}}$ : double bond equivalents minus oxygen. AI: aromaticity index. O/C: O/C ratio. P/C: P/C ratio. N/C: N/C ratio. S/C: S/C ratio. GFE: standard Gibb's Free Energy of carbon oxidation.  $Y_{\text{met}}$ : carbon use efficiency. H/C: H/C ratio.

## Graphical Abstract



## Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: