

METHOD

iDOM: Statistical analysis of dissolved organic matter characterized by high-resolution mass spectrometry

Fanfan Meng^{1,2,#} , Ang Hu^{1,#} , Kyoung-Soon Jang³ , and Jianjun Wang^{1,*} **Abstract**

Dissolved organic matter (DOM) contains thousands of molecules and is key for biogeochemical cycles in aquatic and terrestrial ecosystems by interacting with microbes. Over the last decade, the study of DOM has been advanced and accelerated with the developments of instrumental and statistical approaches. However, it is still challenging in statistical analyses, data visualization, and theoretical interpretations largely due to the complexity of molecular composition and underlying ecological mechanisms. In this study, we developed an R package *iDOM* with functions for the basic and advanced statistical analyses and the visualization of DOM derived from Fourier transform ion cyclotron resonance mass spectrometer (FT-ICR MS). The package could handle various data types of DOM, including molecular compositional data, molecular traits, and uncharacterized molecules (i.e., dark matter). It could integrate explanatory data, such as environmental and microbial data, to explore the relationships between DOM and abiotic or biotic drivers. To illustrate its use, we presented case studies with an example dataset of DOM and microbial communities under experimental warming. We included case studies of basic functions for the calculation of molecular traits, the assignment of molecular classes, and the compositional analyses of chemical diversity and dissimilarity. We further showed the case studies with advanced functions to quantify DOM assembly processes, assess the effects of dark matter on molecular interactions, analyze the ecological networks between DOM and microbes, and explore their response to warming. The source code and example dataset of *iDOM* are publicly available on <https://github.com/jianjunwang/iDOM>. We expect that *iDOM* will serve as a comprehensive pipeline for DOM statistical analyses and bridge the gap between chemical characterization and ecological interpretation in a theoretical framework.

Keywords: dissolved organic matter; ecological interpretation; FT-ICR MS; R package; statistical analysis

Impact statement

Dissolved organic matter (DOM) is a key component of Earth's ecosystems, playing a vital role in nutrient cycling and interactions with microbial communities. However, studying DOM presents significant statistical challenges due to its complex molecular composition and the difficulty of connecting its composition to ecological processes. To address these challenges, we developed *iDOM*, a new R package designed to streamline the analysis, visualization, and interpretation of DOM data. By integrating DOM data with environmental and microbial datasets, *iDOM* enables researchers to better understand how DOM responds to global changes, such as climate warming. This tool bridges disciplines of chemistry and ecology, providing novel insights into the DOM's role in ecosystem functioning.

INTRODUCTION

Dissolved organic matter (DOM) is a large and complex mixture of thousands of molecules that play crucial roles in biogeochemical cycles¹. The high heterogeneity of DOM composition has previously presented a great challenge for our understanding of their reactivity, persistence, and

ecological significance^{2,3}. However, DOM studies have made substantial progress recently due to the advancements in high-resolution mass spectrometry and statistical approaches. For instance, Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) was first applied

¹State Key Laboratory of Lake and Watershed Science for Water Security, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing, China.

²University of Chinese Academy of Sciences, Beijing, China. ³Bio-Chemical Analysis Team, Korea Basic Science Institute, Cheongju, South Korea.

* **Correspondence:** Jianjun Wang, jjwang@niglas.ac.cn

[#]Fanfan Meng and Ang Hu contributed equally to this study.

Received July 11, 2024; Accepted January 25, 2025; Published online April 14, 2025

DOI: [10.1002/mlf2.70002](https://doi.org/10.1002/mlf2.70002)

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

to characterize the molecular composition of humic and fulvic acids from the Suwannee River⁴. In the last decade, FT-ICR MS has been increasingly used to examine DOM in natural environments such as lakes, rivers, oceans, and soils, and it has great potential for future studies⁵. The number of DOM studies using FT-ICR MS reached 189 in 2023, reflecting a nearly 530% increase since 2014, and the percentage of these studies among all DOM studies rose from 2.50% to 8.32% (Figure S1). Compared to conventional bulk analysis methods like absorbance and fluorescence spectroscopy, FT-ICR MS provides more detailed information on the elemental composition and molecular characteristics of DOM across global environments⁶. There were also advances in statistical approaches and relevant graphical tools, including the modified aromaticity index (AI_{mod})⁷ and the van Krevelen diagram⁸, which are widely used to assess DOM structural features and to visualize molecular composition based on H/C and O/C ratios, respectively.

New statistical and modeling approaches have been developed to substantially enhance our understanding of DOM and address key questions related to its assembly processes, environmental responses, and interactions with microbial communities. These advancements have shifted DOM research from traditional compositional analyses to a more functional perspective on DOM assemblages. One such advancement is the application of functional diversity to DOM through approaches like the community-weighted mean (CWM), which simplifies complex mass spectral data by focusing on compositional-level DOM traits^{5,8}. Furthermore, due to various traits among individual molecules, it is challenging to predict how DOM assemblages are assembled and how they respond to global change. These challenges are now being overcome by newly developed methods, such as the DOM partitioning framework (DOM_{part})⁹, the indicator of compositional-level environmental response (iCER)¹⁰, and the standardized Shannon index of DOM-microbe associations (H_2')¹¹. For instance, the DOM_{part} categorizes DOM composition by considering molecular traits such as reactivity and activity and explains how variations in DOM fractions are shaped by both deterministic and stochastic processes. By applying ecological null models, it could estimate the relative contributions of these processes to the assembly of each fraction⁹. The indicator iCER could assess the overall environmental response of DOM assemblages, such as their thermal response, by aggregating the diverse responses of molecules that are positively or negatively associated with warming^{10,12}. The H_2' index could quantify the degree of specialization between DOM and microbes by analyzing their bipartite networks, which are based on resource-consumer relationships¹¹. Notably, in previous analyses, a large proportion of DOM compounds remain uncharacterized and neglected, often referred to as chemical “dark matter”. The new indicator of dark matter effects (iDME), developed based on ecological networks, could assess how dark matter influences molecular interactions within DOM assemblages under global change¹³. So far, there are a few open-source R packages¹⁴

and pipelines¹⁵ developed for analyzing and visualizing FT-ICR MS data. However, no software is available to integrate the advanced analyses mentioned above to understand the DOM composition within the ecological context.

Here, we develop an R package *iDOM* to bridge the current gap in advanced statistical analyses of FT-ICR MS data (Figure 1). *iDOM* is a multifunctional tool that integrates both basic and advanced analytical functions for DOM research. The package's key features include basic analyses, such as the calculation of molecular traits, the assignment of molecular classes, and the calculation of chemical diversity and dissimilarity. It also offers advanced analyses to quantify the assembly processes of DOM composition⁹, the compositional-level environmental responses of DOM¹⁰, the effects of dark matter¹³, and the specialization of DOM-microbe associations¹¹. Additionally, *iDOM* provides visualization functions, such as van Krevelen diagrams to visualize FT-ICR MS data, and composition plots to show the relative abundances of different molecular classes. To illustrate its application, we applied the package to analyze an example dataset of DOM from a microcosm experiment investigating sediment DOM under experimental warming¹⁰.

RESULTS AND DISCUSSION

Calculation of molecular traits and assignment of molecular classes

The R package *iDOM* provides the *molTrait* and *molTrans* functions to calculate molecular properties and putative biochemical transformations, respectively, and the *molGroup* function to classify DOM molecules based on these traits. Specifically, the *molTrait* function calculates the chemical characteristics of molecules related to molecular weight, stoichiometry, chemical structure, and others (Tables 1 and S1). These traits include mass, the number of carbon atoms (C), Kendrick defect ($k\text{defect}_{\text{CH}_2}$), O/C ratio, H/C ratio, N/C ratio, P/C ratio, S/C ratio, the modified aromaticity index (AI_{mod}), double bond equivalent (DBE), DBE minus oxygen (DBE_{O}), DBE minus AI (DBE_{AI}), standard Gibbs Free Energy (GFE), nominal oxidation state of carbon (NOSC), and carbon use efficiency (Y_{met})^{7,16–19}. Furthermore, the *molTrans* function estimates putative biochemical transformations for each molecule by comparing the mass differences between the molecule and others with a database of known transformations²⁰.

The *molGroup* function classifies DOM composition into different groups based on various molecular traits and graphical methods, such as molecular properties, putative biochemical transformations, and van Krevelen diagrams^{8,20}. For instance, the assigned molecules can be classified into molecular formula groups based on their element composition (e.g., CHO, CHON, CHONS, CHONSP). Additionally, each molecule plotted on the van Krevelen diagram can be linked to specific natural biomolecules⁸. Molecules in different regions of the diagram are categorized into distinct

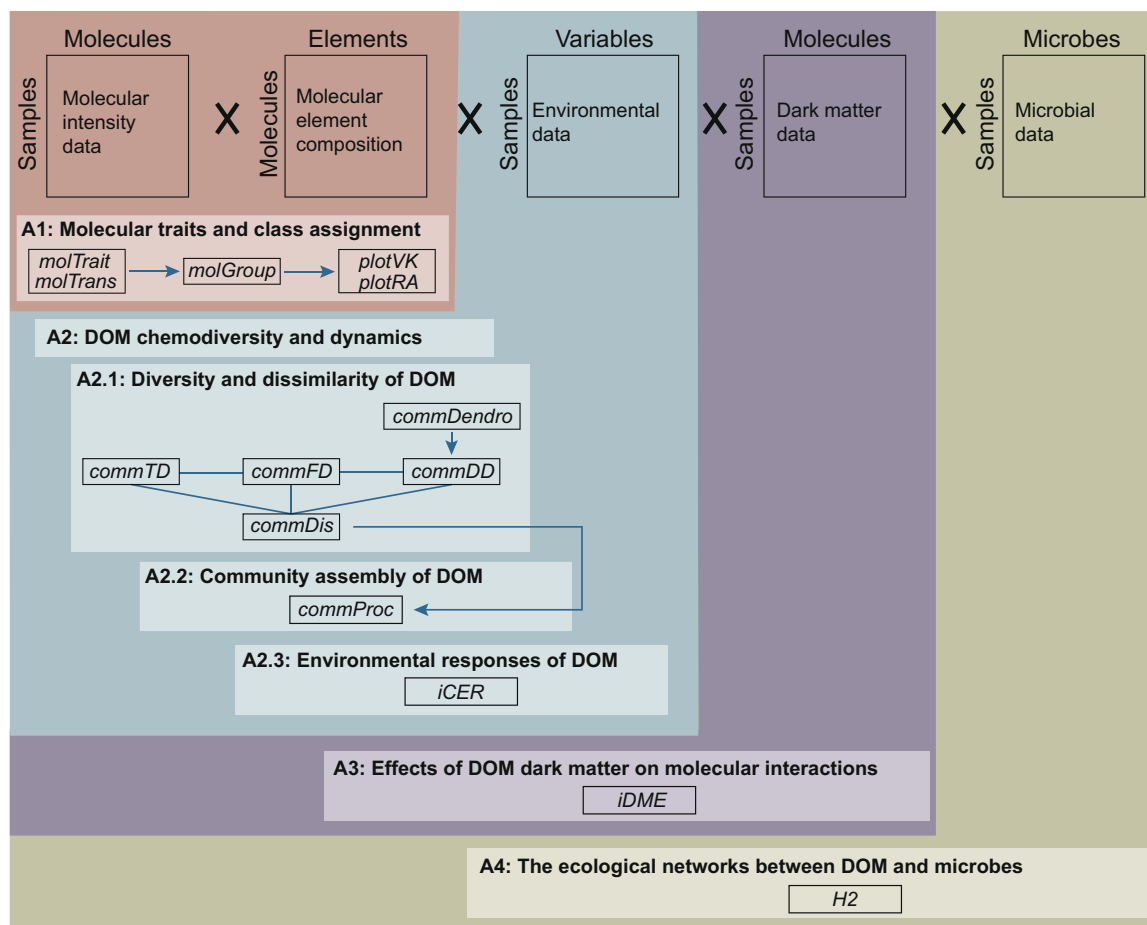


Figure 1. Conceptual scheme of R package *iDOM*. The *iDOM* package includes five example datasets: molecular intensity data, molecular element composition, environmental data, dark matter data, and microbial data. Using these datasets, the package addresses four main aims: (1) calculating molecular traits and assigning molecular classes; (2) analyzing DOM chemodiversity and dynamics, including (2.1) quantifying molecular diversity and dissimilarity, (2.2) explaining the assembly processes of DOM, and (2.3) assessing the environmental responses of DOM; (3) examining the effects of DOM dark matter on molecular interactions; and (4) quantifying DOM-microbe associations through ecological networks.

classes, such as lipids, proteins, carbohydrates, lignin, and others^{21,22}. Recently, a new method was developed to classify molecules based on molecular reactivity, determined by H/C ratios, and activity, inferred from putative biochemical transformations⁹. This classification divides molecules into four fractions: labile-active, recalcitrant-active, labile-inactive, and recalcitrant-inactive.

For the example datasets, the *molGroup* function revealed that CHO and CHON formula groups consistently exhibited higher relative abundances than other groups across the temperature gradient from 5°C to 30°C (Figure 2A). Additionally, condensed aromatics (ConHC) and lignin consistently showed dominance throughout the temperature range (Figure 2B). The *molGroup* function also classified 4150 out of 5474 molecules into four fractions based on molecular reactivity and activity: labile-active, recalcitrant-active, labile-inactive, and recalcitrant-inactive. In both the labile and recalcitrant fractions, active molecules exhibited higher relative abundances than inactive molecules (Figure S2). Compared to traditional classifications based solely on molecular reactivity, such as an H/C ratio cutoff of 1.5, molecular activity offers new

insights into the functional dynamics of molecules^{9,23}. Notably, in the van Krevelen diagram, highly active molecules are distributed across regions both above and below the H/C ratio threshold of 1.5 (Figure 2C), highlighting the limitations of such conventional criteria.

Diversity and dissimilarity of DOM

The R package *iDOM* provides the *commTD*, *commFD*, and *commDD* functions to calculate within-assembly diversity and the *commDis* function to assess between-assembly differences in DOM composition (Figure 1). To apply diversity metrics, originally designed for ecological species, to molecular data, individual molecules are treated as species, with the relative intensities of their peaks representing species abundances. The *commTD* function calculates taxonomic diversity using common alpha-diversity indices, including richness, abundance-based metrics, and evenness. It computes molecular richness as the number of molecular formulas and calculates abundance-based diversity metrics, such as Shannon

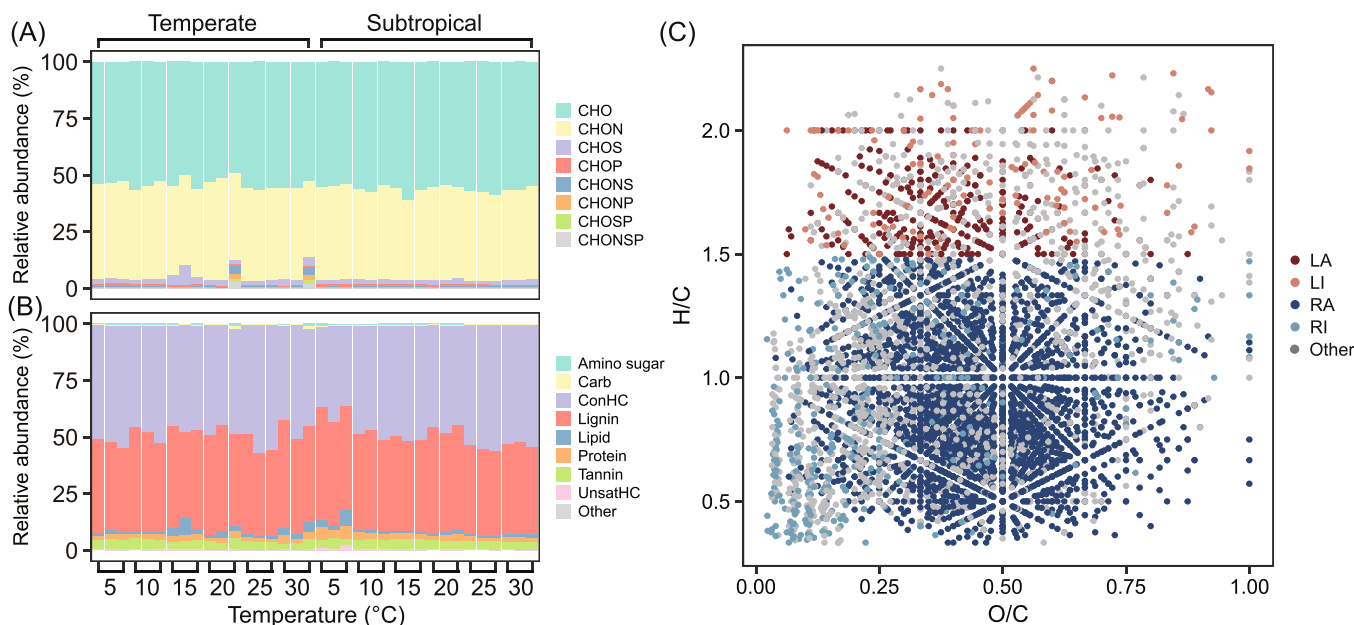


Figure 2. Molecular group composition and van Krevelen diagram. (A) Relative abundance of molecular groups based on element composition across experimental temperatures. (B) Relative abundance of molecular groups classified by H/C and O/C ratios across experimental temperatures. These groups include 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), amino sugars (O/C = 0.55–0.67, H/C = 1.5–2.2), carbohydrates (Carb; O/C = 0.67–1.2, H/C = 1.5–2), 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 = 0.67–1.2, H/C = 0.5–1.5), and condensed aromatics (ConHC; O/C = 0–0.67, H/C = 0.2–0.7)⁸. (C) Molecular distributions in van Krevelen diagram. Different colors represent molecular groups classified by molecular trait dimensions of reactivity and activity: LA (labile-active; H/C ≥ 1.5, transformations > 10), LI (labile-inactive; H/C ≥ 1.5, transformations ≤ 1), RA (recalcitrant-active; H/C < 1.5, transformations > 10), RI (recalcitrant-inactive; H/C < 1.5, transformations ≤ 1), and others⁹.

and Gini-Simpson (or Simpson's index), which integrate both molecular richness and relative intensities^{24,25}. Meanwhile, it measures evenness (e.g., Pielou's evenness), derived from molecular richness and the Shannon index, to reflect the equality of molecular intensity distributions. The *commFD* function calculates functional structure and diversity based on molecular traits, using approaches such as the CWM and Rao's quadratic entropy (RaoQ). The CWM simplifies complex mass spectral data into compositional-level DOM traits, while RaoQ measures the average abundance-weighted trait differences between pairs of molecules within an assemblage²⁶. Greater trait differences among individual molecules result in higher quadratic entropy, indicating increased functional diversity²⁷.

To further explore relationships among molecules, the *commDendro* function generates molecular dendrograms analogous to phylogenetic trees. These dendrograms are constructed based on molecular properties and putative biochemical transformations, including the molecular characteristics dendrogram (MCD), the transformation-based dendrogram (TD), and the transformation-weighted characteristics dendrogram (TWCD), to represent shared and divergent molecular traits among molecules²⁰. Using these molecular dendrograms, the *commDD* function calculates dendrogram-based diversity metrics, including dendrogram diversity (DD), mean pairwise distance (MPD), and mean nearest taxon distance (MNTD)²⁸. DD quantifies

the total branch length of a dendrogram occupied by a molecular assemblage, similar to Faith's Phylogenetic Diversity²⁹. MPD measures the average dendrogram distance between molecules, while MNTD calculates the average dendrogram distance between nearest neighbors. For instance, molecular assemblages with higher DD values reflect a broader range of molecular properties in MCD, a more extensive biochemical transformation network in TD, or a combination of both in TWCD²⁰.

As a complement to alpha diversity, the *commDis* function compares differences in DOM composition between samples by generating a dissimilarity matrix. The dissimilarity metrics include incidence-based Jaccard, abundance-based Bray-Curtis, dendrogram-based UniFrac, and others. Each dissimilarity matrix can be further analyzed using nonmetric multidimensional scaling (NMDS) or principal coordinate analysis (PCoA) to visually depict the relationships among samples.

Based on the example datasets, the *commTD*, *commFD*, and *commDD* functions revealed that DOM molecular diversity exhibited distinct relationships with experimental temperature in temperate and subtropical regions. The molecular richness decreased significantly with rising temperature in the temperate region ($p < 0.05$), while no significant change was observed in the subtropical region ($p > 0.05$) (Figure 3A). Conversely, RaoQ and MNTD decreased significantly with rising temperature in the subtropical region ($p < 0.05$), but showed no

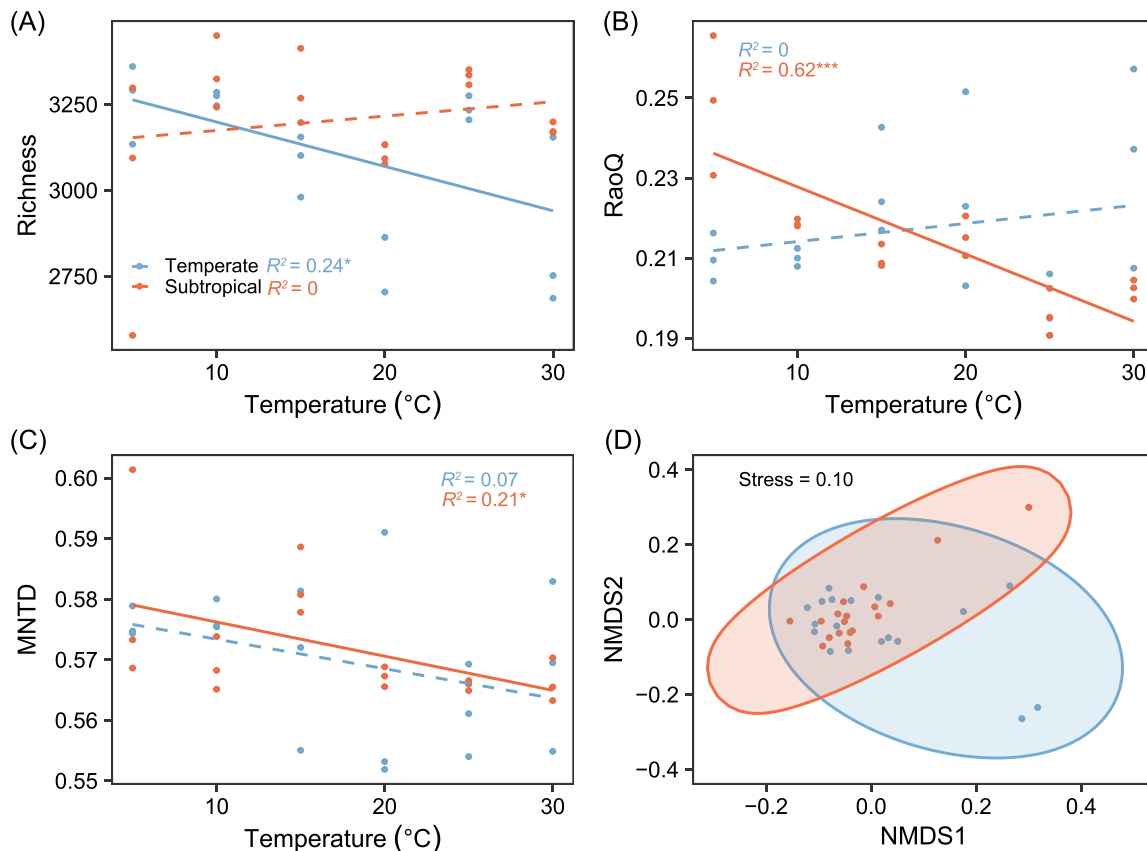


Figure 3. Relationships between DOM chemodiversity and experimental temperatures. (A–C) Molecular richness (A), Rao's quadratic entropy (RaoQ) (B) based on the modified aromaticity index (AI_{mod}), and mean nearest taxon distance (MNTD) (C) derived from the molecular characteristics dendrogram (MCD) across experimental temperatures in temperate and subtropical regions. The relationships between these indices and experimental temperatures were tested using linear models. Solid lines indicate statistically significant models ($p < 0.05$), while dashed lines represent nonsignificant models ($p > 0.05$). Significance levels are indicated as follows: *** $p < 0.001$; * $p < 0.05$. (D) Compositional dissimilarity among samples visualized using NMDS based on Bray-Curtis distance.

significant change in the temperate region ($p > 0.05$) (Figure 3B,C). Further, the *commDis* function revealed that molecular composition was similar between the subtropical and temperate regions, as indicated by Permutational Multivariate Analysis of Variance (PERMANOVA, $p > 0.05$) (Figure 3D).

Community assembly of DOM composition

Identifying how deterministic and stochastic processes influence changes in DOM assemblages is crucial for predicting their turnover and dynamics in response to global change^{20,30}. The *commProc* function quantifies the relative contributions of these processes to DOM assembly using dendrogram-based β -diversity null modeling⁹. Specifically, the function calculates the dendrogram-based β -nearest taxon index (β NTI) to measure tip-level clustering or overdispersion within a molecular dendrogram, analogous to its use in phylogenetic trees for ecological communities^{31,32}. The β NTI is calculated by comparing the observed β -mean nearest taxon distance (β MNTD_{obs}) between pairs of DOM assemblages to null expectations (β MNTD_{null}), generated by randomizing observed dendrogram associations²⁰. The formula is as follows:

$$\beta\text{NTI} = \frac{\beta\text{MNTD}_{\text{obs}} - \overline{\beta\text{MNTD}_{\text{null}}}}{\text{sd}(\beta\text{MNTD}_{\text{null}})},$$

where β MNTD_{obs} represents the observed β MNTD, $\overline{\beta\text{MNTD}_{\text{null}}}$ is the mean β MNTD for the null assemblages, and $\text{sd}(\beta\text{MNTD}_{\text{null}})$ indicates the standard deviation of null values. If the comparison between two DOM assemblages significantly deviates from the null expectations ($|\beta\text{NTI}| > 2$), deterministic processes likely drive the observed patterns. Positive β NTI values ($\beta\text{NTI} > 2$) indicate divergent molecular composition resulting from variable selection, while negative β NTI values ($\beta\text{NTI} < -2$) suggest convergent molecular composition driven by homogeneous selection⁹. Conversely, if the comparison aligns with the null expectations ($|\beta\text{NTI}| < 2$), stochastic processes are likely responsible for the observed patterns⁹.

For the example datasets, we used the *commProc* function to quantify the relative contributions of deterministic and stochastic processes to DOM assembly. The majority of β NTI values for MCD, TD, and TWCD exceeded 2, indicating that deterministic processes, particularly variable selection, predominantly governed molecular assembly (Figure 4A). To further explore the assembly mechanisms of different DOM fractions compared to whole DOM assemblages, we

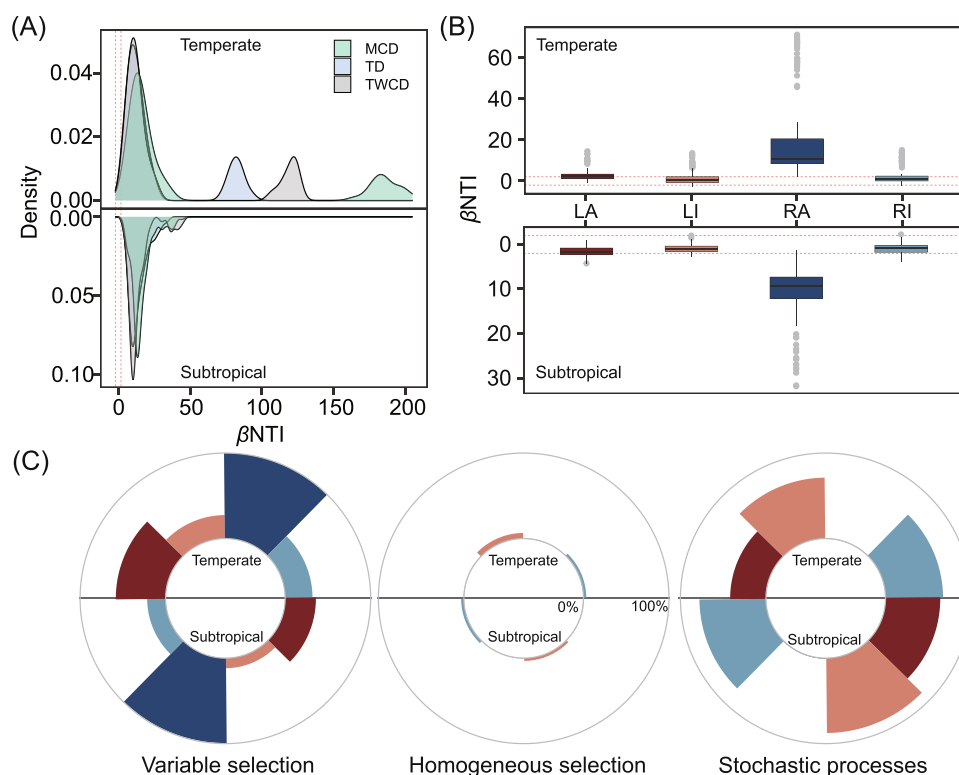


Figure 4. Community assembly of DOM composition. (A) Distribution of β -nearest taxon index (β NTI) values derived from three molecular dendrograms: MCD, transformation-based dendrogram (TD), and transformation-weighted characteristics dendrogram (TWCD). (B) Boxplot of β NTI values of different DOM fractions based on the MCD. In the example datasets, 4150 out of 5474 molecules across 36 samples were classified into four fractions based on their reactivity and activity: LA, RA, RI, and LI. (C) Relative contributions of assembly processes, including variable selection, homogeneous selection, and stochastic processes, to the four DOM fractions. For the example datasets, the assembly processes of these four fractions were quantified using incidence-based molecular data, that is, the presence or absence of molecular peaks rather than peak intensity data. This is because peak intensities could not contain the same ecological meaning as variations in microbial species abundances. The use of peak intensities from FT-ICMS is challenging for null modeling without additional information linking peak intensities to concentrations⁹.

applied the *molGroup* function to classify DOM composition into four fractions based on molecular reactivity and activity (Figure 4B). In both subtropical and temperate regions, deterministic processes driven by variable selection dominated the assembly of labile and recalcitrant molecules in the active fractions, while stochastic processes are more important for the assembly of molecules within the inactive fractions. The homogeneous selection had minimal importance across all fractions in both regions (Figure 4C).

Thermal responses of DOM composition

The diverse intrinsic traits of DOM molecules present a significant challenge in predicting their responses to climate change and, consequently, the impact on the global carbon cycle. The *iCER* function is designed to quantify the responses of DOM to environmental changes, such as warming, at both molecular and compositional levels¹⁰. The thermal responses of DOM assemblages can be quantified using a novel indicator, *iCER*, which is based on changes in molecular abundances along temperature gradients. The calculation of *iCER* involves two main steps. First, the *iCER* function determines the direction and magnitude of the molecule-specific environmental response (MER) to

temperature. MER quantifies the effect size of changes in the relative abundance of each molecule in response to temperature variations across different time, space, or treatments. Positive MER values are associated with molecules that accumulate as temperatures increase (warm-accumulating molecules), while negative MER values are linked to molecules that deplete with elevated temperatures (warm-depleting molecules). Second, the *iCER* function calculates the *iCER* indicator by taking the weighted average of the MERs within each DOM assemblage. This approach enables *iCER* to capture the overall response of a DOM assemblage to temperature changes, reflected in the balance between positive and negative MERs. A positive *iCER* value indicates that the DOM assemblage is dominated by warm-accumulating molecules, while a negative value suggests the dominance of warm-depleting molecules. An *iCER* value of zero suggests an equal representation of both groups. Since *iCER* depends on MER, it is essential to use statistically independent datasets: one for calculating MERs and the other for *iCER*s. This ensures that the relative abundances of individual molecules would not be reused, allowing *iCER* to be derived independently from MER values¹⁰.

We randomly partitioned the example datasets into two independent subsets for MER and *iCER* calculations using

the commonly applied 80:20 ratio from species distribution modeling and machine learning³³. Specifically, 80% of the samples (i.e., MER dataset) in each region were used to calculate the MERs of individual molecules, and the remaining 20% (i.e., iCER dataset) were then used to calculate the iCER for each sample based on the relative abundances of individual molecules and their corresponding MERs. In the MER dataset, we calculated MERs as Spearman's correlation coefficients (ρ) between the relative abundances of individual molecules and temperature. To avoid false correlations caused by low-occurrence molecules, we retained only those molecules detected in at least 30% of the total samples for the MER calculation. Then, in the iCER dataset, we synthesized the MERs of a suite of molecules into an iCER for each sample. The iCER was calculated as:

$$\text{iCER} = \frac{\sum(\text{MER}_i \times I_i)}{\sum(I_i)},$$

where MER_i represents the MER value of molecule i , and I_i indicates its relative abundance. This data partitioning and indicator calculations were randomly repeated 999 times, and the resulting MER and iCER values were averaged across all randomizations. For the example datasets, MER values ranged from -0.99 to 0.98 across both regions. The negative MER values had median values of -0.41 and -0.34 , while the positive MER values had medians of 0.43 and 0.42 in the subtropical and temperate regions, respectively (Figure 5A). The iCER values exhibited consistent warming responses in the temperate and subtropical regions, with positive correlations with temperature and R^2 values of 0.76

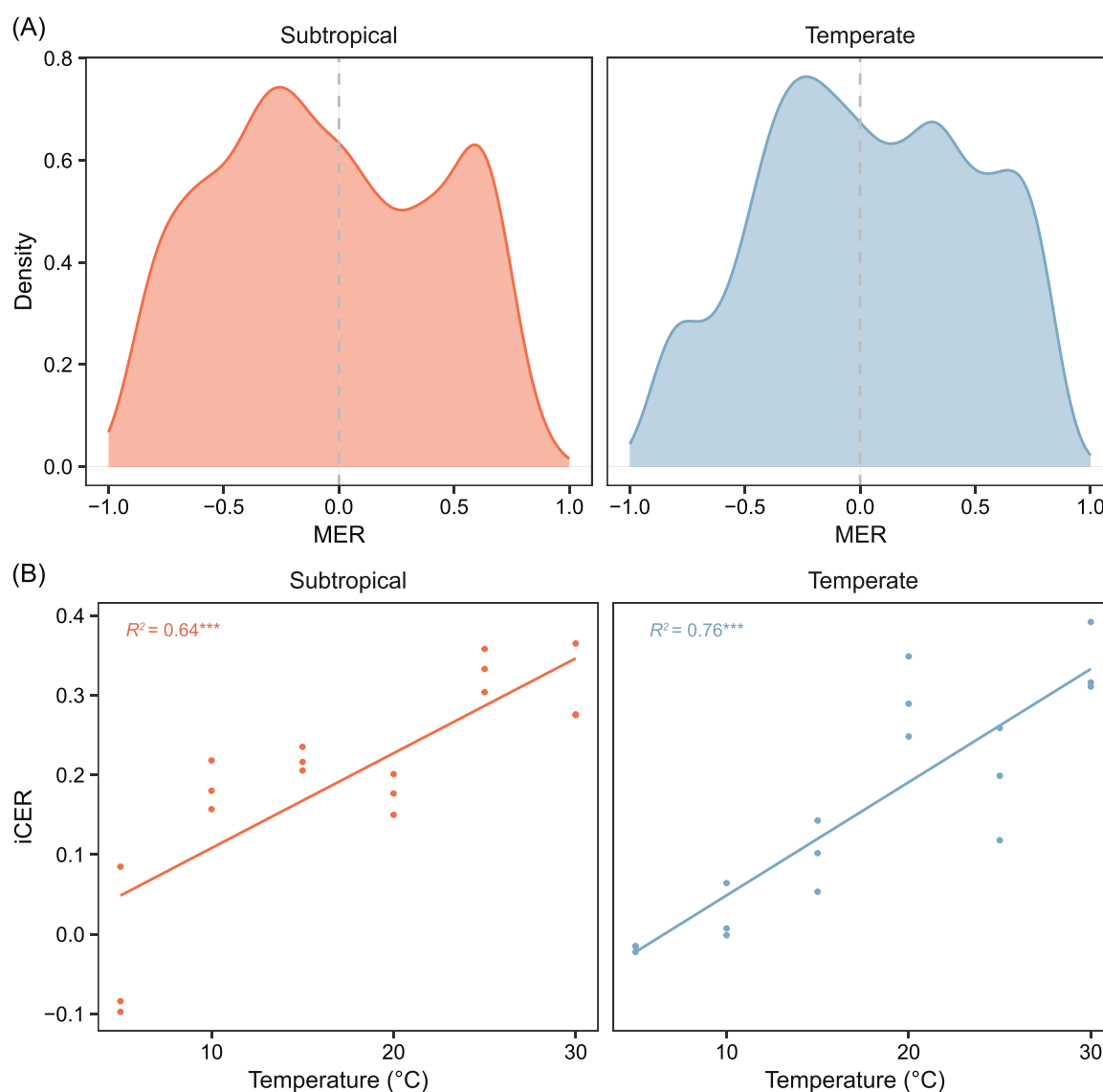


Figure 5. The molecule-specific environmental response (MER) and the indicator of compositional-level environmental response (iCER) of DOM. (A) Distribution of MERs for DOM molecules in subtropical and temperate regions. MERs were calculated as Spearman's correlation coefficients. (B) Relationships between iCERs and experimental temperatures in subtropical and temperate regions tested using linear models. iCERs were derived from molecules with statistically significant MERs ($p < 0.05$). Solid lines indicate statistically significant models ($p < 0.05$). $^{***}p < 0.001$.

and 0.64, respectively (Figure 5B). These results indicate that individual DOM molecules show varied thermal responses, while DOM assemblages consistently exhibit enhanced thermal responses under warmer conditions¹⁰.

Effects of DOM dark matter on molecular interactions

DOM molecules detected by FT-ICR MS can be assigned to identifiable molecular formulas, yet a large proportion still remains uncharacterized and is often referred to as chemical “dark matter”¹³. The interaction between chemical dark matter and assigned (i.e., known) molecules presents a significant challenge to fully understanding biogeochemical cycles. The *iDME* function calculates the *iDME* to quantify how dark matter influences molecular interactions within

DOM assemblages by constructing co-occurrence networks¹³. In each network, nodes represent individual molecules, and edges indicate the interactions between molecules. Specifically, two types of networks are constructed based on the presence or absence of dark matter: “KK” networks, which include only known molecules, and “DK” networks, which incorporate both dark matter and known molecules at a 1:1 ratio or according to the observed dark-to-known ratio in a DOM assemblage (Figure 6A). The “KK” and “DK” networks contain the same number of nodes, which are randomly subsampled from the entire DOM molecule pool and then replicated 100 times. The *iDME* function calculates the magnitude and direction of the *iDME* indicator by measuring the percentage change in the average value of a network metric M , such as degree centrality, between “KK” and “DK” networks.

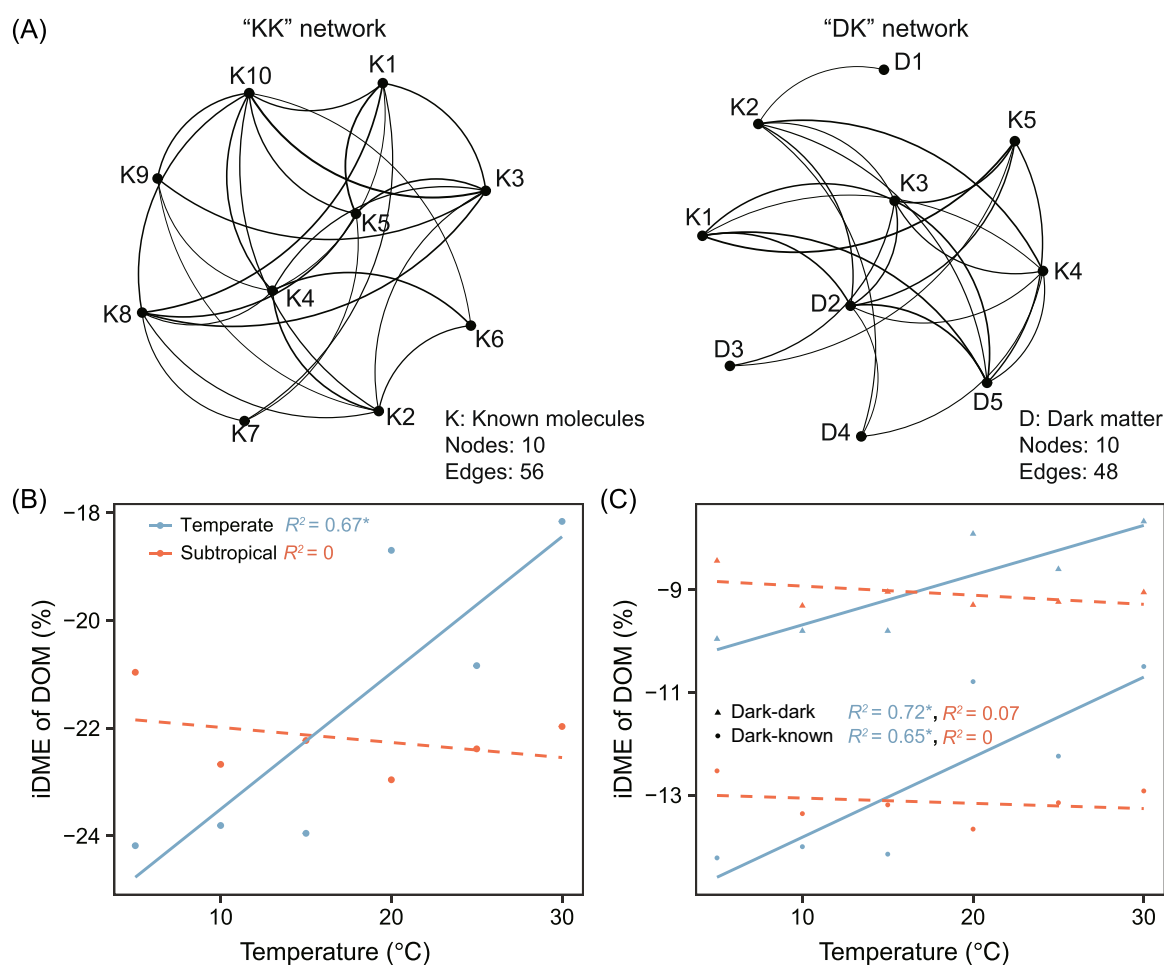


Figure 6. Molecular co-occurrence networks and the indicator of dark matter effects (*iDME*). (A) Illustration of “KK” and “DK” molecular co-occurrence networks with 10 nodes. The “KK” network includes only known molecules (“K”), while the “DK” network replaces half of the known molecules in the “KK” network with dark matter (“D”). The “KK” and “DK” networks were constructed using molecules present in more than 30% of the total samples, with interactions inferred through SparCC (Sparse Correlations for Compositional data)³⁴. SparCC correlations below the threshold of $|\rho| = 0.30$ were excluded to remove uncorrelated or weakly correlated interactions. (B) Relationships between *iDME* and experimental temperatures in subtropical and temperate regions tested using linear models. For each sample in the example datasets, the *iDME* was calculated based on the “KK” and “DK” networks, each constructed using 400 nodes randomly sampled from its DOM composition, with the sampling repeated 100 times. The *iDME* values were averaged across different temperature levels. (C) Relationships between *iDME* partitions and experimental temperatures in subtropical and temperate regions tested using linear models. Dark–dark and Dark–known links correspond to the intra- and inter-*iDME* partitions, respectively. Solid lines indicate statistically significant models ($p < 0.05$), while dashed lines represent non-significant models ($p > 0.05$). $^*p < 0.05$.

$$\text{iDME} (\%) = \left(\frac{M_{\text{DK}}}{M_{\text{KK}}} - 1 \right) \times 100.$$

Degree refers to the number of edges connecting a focal node to other nodes within a network³⁵. Molecules with a higher degree generally exhibit more interactions, indicating greater connectivity within an assemblage. Therefore, positive iDME values suggest that dark matter enhances network connectivity within a DOM assemblage, while negative values imply a reduction and an iDME of zero indicates a neutral effect. The iDME can be further divided into intra-iDME and inter-iDME to determine whether the effects of dark matter result from interactions between dark-dark nodes or dark-known nodes¹³.

For the example datasets, all iDME values calculated based on the network metric degree were negative and significantly different from zero, ranging from -24.2% to -18.2% in temperate regions and from -23.0% to -21.0% in subtropical regions. In temperate regions, iDME values increased significantly along the temperature gradient ($p < 0.05$), while subtropical regions showed nonsignificant trend ($p > 0.05$) (Figure 6B). These results suggest that DOM dark matter substantially reduced network connectivity along the temperature gradients in both regions, but its negative effects in temperate regions weakened as the temperature increased. Moreover, the partitioning of iDME showed that dark matter effects were primarily driven by changes in links between dark-known nodes, followed by changes in links between dark-dark nodes in both temperate and subtropical regions (Figure 6C).

The ecological networks between DOM and microbes

The fate of DOM is intimately linked to the metabolism of complex microbial communities, as microbes regulate the production and degradation of specific molecules^{11,36}. The associations between DOM molecules and microbial taxa can be quantified using DOM-microbe bipartite networks, which are constructed based on resource-consumer theory³⁷. In these networks, individual DOM molecules are exclusively linked to microbial taxa that utilize them, while direct interactions among molecules or taxa are not considered. Based on resource-consumer relationships, negative network interactions likely reflect the degradation of larger molecules into smaller structures, while positive interactions may indicate the production of new molecules through either degradation or biosynthetic processes¹¹.

The H_2 function calculates the H_2' index of DOM-microbe bipartite networks to quantify specialization between DOM and microbes and standardizes it using null modeling to allow direct comparisons across samples^{11,37}. Specifically, an elevated H_2' indicates a high degree of specialization between DOM and microbes, with extreme cases where a single bacterial taxon might consume or produce only one specific DOM molecule. Conversely, a lower H_2' value suggests a more generalized bipartite network, where various DOM molecules are used by a wide range of bacterial taxa¹¹.

We applied the H_2 function to the example datasets to examine how DOM-microbe associations vary under experimental temperatures. In total, the negative and positive bipartite networks contained 1108 and 1938 interactions between DOM molecules and bacteria genera, respectively (Figure 7A). The standardized H_2' values were negative and significantly lower than expected by chance ($p < 0.05$), indicating that the interactions between DOM and bacteria were nonrandom. Experimental warming had contrasting effects on the H_2' of negative and positive networks across the two regions (Figure 7B). Specifically, for the positive networks, experimental warming significantly decreased H_2' in both temperate and subtropical regions ($p < 0.05$). For the negative networks, experimental warming significantly increased H_2' for the temperate region ($p < 0.05$), while no significant correlation was observed in the subtropical region. Experimental warming may thus enhance the recalcitrance of DOM in the temperate region by increasing production (i.e., less specialized positive networks) and reducing the decomposition of molecules (i.e., more specialized negative networks)¹¹.

In summary, the package *iDOM* is a comprehensive set of functions developed to facilitate the chemical characterization and ecological interpretation of DOM based on high-resolution mass spectrometry. *iDOM* enables us to perform chemical characterization, such as molecular trait calculation, molecular class assignment, and compositional analyses of chemical diversity and dissimilarity. Further, *iDOM* integrates concepts and tools from community ecology to facilitate the theoretical interpretation of community assembly of DOM, the thermal responses of DOM assemblages, the effects of dark matter on molecular interactions, and the DOM-microbe associations. The *iDOM* is expected to promote standardized methodologies and reproducible research in DOM studies, and its extensibility makes it suitable for a wide range of applications across global environments.

MATERIALS AND METHODS

Description of functions in the R package

The *iDOM* package, developed in the R programming language, provides a variety of functions for the statistical analysis and graphical visualization of DOM, as summarized in Table 1. Key functions, such as *molGroup*, *commProc*, *iCER*, *iDME*, and H_2 , were specifically developed based on novel frameworks or indices reported in our previous literature^{9-11,13}. Other functions like *molTrait*, *molTrans*, *commTD*, *commFD*, *commDendro*, *commDD*, and *commDis* are necessary for supporting these primary functions. These additional functions rely on external packages, including *vegan*³⁸, *FD*³⁹, *picante*⁴⁰, *iCAMP*⁴¹, and *ftmsRanalysis*¹⁴, as well as R scripts from Danczak et al.²⁰.

The *iDOM* package is designed with a modular structure, with its functions focused on four main aims (Figure 1). The first aim is to use molecular intensity and element composition data to calculate molecular traits with the *molTrait* and *molTrans* functions and classify molecular groups based on these traits using the *molGroup* function. The second aim is to integrate environmental variables to (1) describe the distribution of molecular

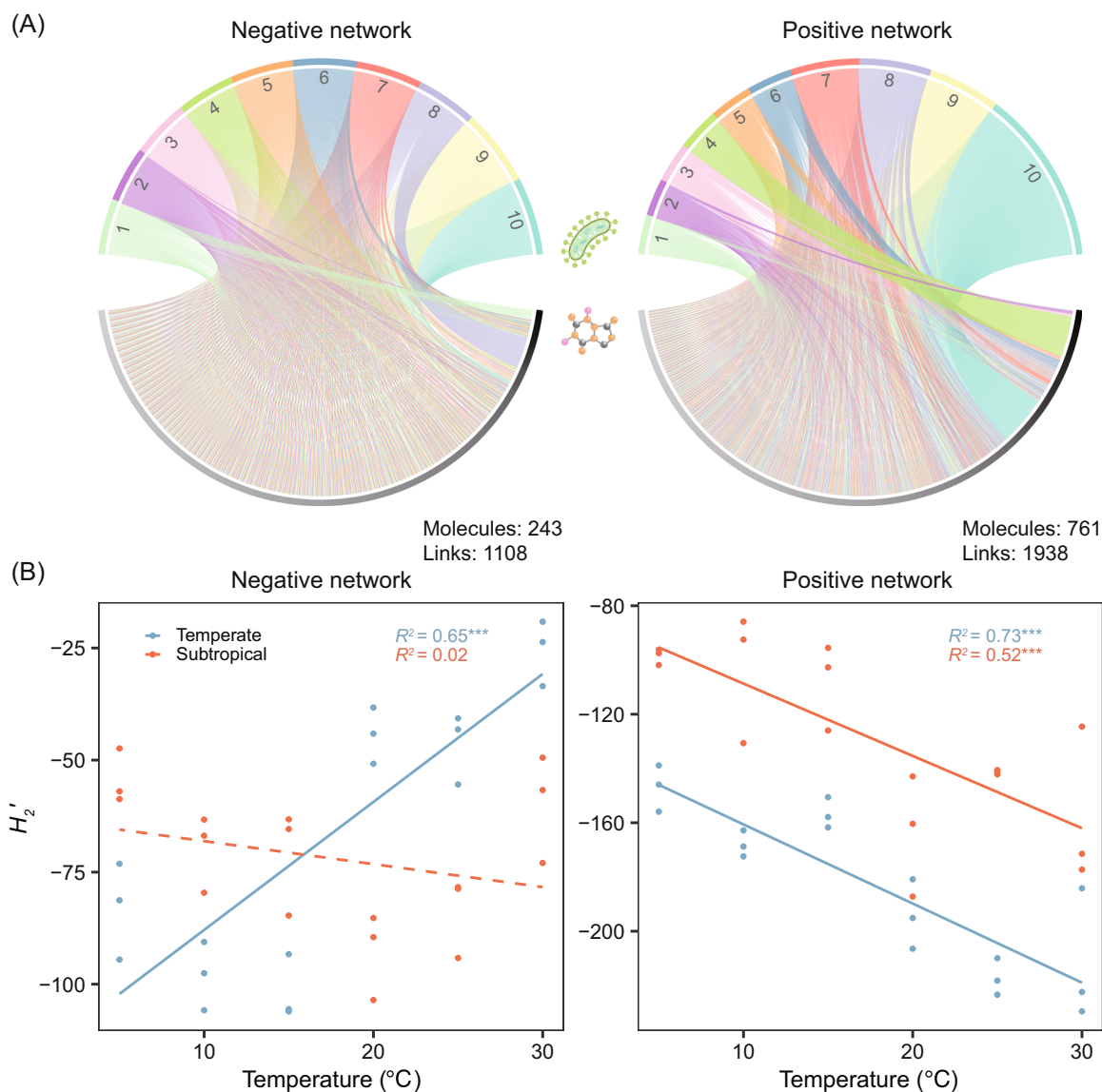


Figure 7. Specialization of DOM-microbe bipartite networks. (A) Negative and positive bipartite networks between DOM molecules and bacterial genera. Upper arcs represent the top 10 bacterial genera, which showed more associations with molecules than other bacterial genera. Each genus is assigned a unique color and number. In the negative network, the genera are: 1, *Ramlibacter*; 2, *Anaeromyxobacter*; 3, *Caulobacter*; 4, *Magnetospirillum*; 5, *GOUTA19*; 6, *Treponema*; 7, *Symbiobacterium*; 8, *Kaistobacter*; 9, *Desulfurispora*; 10, *Rhodoplanes*. In the positive network, the genera are: 1, *Opiritatus*; 2, *Sulfuricurvum*; 3, *Rhodoplanes*; 4, *Janthinobacterium*; 5, *Desulfurispora*; 6, *Azohydromonas*; 7, *Dehalobacterium*; 8, *Sphingomonas*; 9, *Paracoccus*; 10, *Kaistobacter*. Lower arcs represent DOM molecules, colored in shades from gray to black, with darker shades indicating molecules that have more associations with bacterial genera. Negative and positive networks were constructed based on DOM molecules and bacterial genera present in more than 30% of the total samples, using negative and positive SparCC correlation coefficients ($\rho < -0.50$ and $\rho > 0.50$, respectively). For each sample in the example datasets, separate negative and positive sub-networks were constructed by selecting the DOM molecules and bacterial genera in each sample based on its DOM and bacterial compositions. In each subnetwork, SparCC ρ values were multiplied by 100,000, rounded to integers, and converted to absolute values. These values were then used to calculate standardized specialization indices (H_2') through null modeling using the shuffle.web algorithm. (B) Relationships between H_2' of negative and positive networks and experimental temperatures in subtropical and temperate regions, tested using linear models. Solid lines indicate statistically significant models ($\rho < 0.05$), while dashed line represents nonsignificant models ($\rho > 0.05$). $^{***}p < 0.001$.

diversity and dissimilarity using the *commTD*, *commFD*, *commDD*, and *commDis* functions; (2) explain the assembly processes of DOM composition with the *commProc* function; and (3) analyze the environmental responses of DOM through the *iCER* function. The third aim is to incorporate dark matter data to examine its effects on molecular interactions within DOM assemblages through ecological co-occurrence networks with the

iDME function. The fourth aim is to include microbial data to quantify DOM-microbe associations based on bipartite networks using the H_2 function.

Example datasets

The *iDOM* package provides example datasets of DOM and microbial communities under experimental warming.

Table 1. The functions of R package “*iDOM*”.

Type	Function	Description
Molecular properties and class assignment	<i>molTrait</i>	To compute various molecular traits related to chemical characteristics
	<i>molTrans</i>	To estimate biochemical transformations for individual molecules
	<i>molGroup</i>	To classify DOM molecules into different groups based on various molecular traits and graphical methods
Diversity and dissimilarity of DOM	<i>commTD</i>	To calculate various taxonomic diversity and evenness metrics
	<i>commFD</i>	To calculate various functional diversity metrics
	<i>commDendro</i>	To generate relational metabolite dendrograms based on molecular traits and putative biochemical transformations
	<i>commDD</i>	To calculate various dendrogram-based diversity metrics derived from metabolite dendrograms
Community assembly of DOM	<i>commDis</i>	To calculate differences in DOM composition among samples
	<i>commProc</i>	To evaluate the influence of deterministic and stochastic processes on DOM assemblages
Thermal responses of DOM	<i>iCER</i>	To quantify the responses of DOM to environmental changes, such as warming, at both the molecular and compositional levels
Effects of DOM dark matter	<i>iDME</i>	To assess the effects of dark matter on DOM assemblages
Associations between DOM and microbes	<i>H2</i>	To calculate the network-level specialization in DOM-microbe bipartite networks
Visualization	<i>plotVK</i>	To generate van Krevelen diagrams based on molecular O/C and H/C ratios
	<i>plotRA</i>	To visualize the relative abundances of different molecular groups

These datasets were generated from a laboratory microcosm experiment using sterilized Taihu Lake sediments as the organic carbon source, inoculated with distinct microbial communities from sediments of subtropical and temperate lakes in China¹⁰. The microcosms were incubated in the dark for 1 month at six temperature levels (5°C, 10°C, 15°C, 20°C, 25°C, and 30°C), with each temperature treatment replicated three times, resulting in a total of 36 samples across two climate zones¹⁰.

Five example datasets are included: molecular intensity data, molecular element composition, environmental data, dark matter data, and microbial data (Figure 1). The molecular intensity and element composition datasets provide the intensities of 5474 molecules and their corresponding element composition assignments across 36 samples. The environmental dataset contains variables such as experimental incubation temperatures, offering additional context for analyses under varying environmental conditions. The dark matter dataset comprises the intensities of 5779 unassigned molecules, offering insights into their effects on molecular interactions within DOM assemblages. The microbial dataset includes the relative abundances of 463 bacterial genera across 36 samples and can be used to investigate the interactions between DOM and bacteria.

Data and computational requirements for the *iDOM* package

To effectively use the *iDOM* package, it is crucial to meet the minimum requirements of the mass spectral dataset, such as a sufficient sample size for robust statistical analyses. Additionally, the complete input dataset should comprise molecular intensity and element composition data, and supplementary data such as environmental variables, unassigned molecular data, and microbial data to explain the observed DOM composition. Depending on specific aims, users, however, can adjust the input data accordingly based on their dataset characteristics and

research objectives. For large datasets, users are advised to use parallel computing methods available in other R packages or software. For instance, the calculation of co-occurrence networks can be accelerated using *SpiecEasi*⁴² or *fastspar*⁴³, which can reduce the overall computation time of the *iDME* and *H2* functions.

ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (42225708, U24A20578, 42377122, 92251304, and 42077052), Basic Research Program of Jiangsu Province (BK20240111), the Research Program of Sino-Africa Joint Research Center, Chinese Academy of Sciences (151542KYSB20210007), Science and Technology Planning Project of NIGLAS (NIGLAS2022GS09), and Key Laboratory of Lake and Watershed Science for Water Security (NKL2023-QN04).

AUTHOR CONTRIBUTIONS

Fanfan Meng: Formal analysis; writing—original draft. **Ang Hu:** Data curation; methodology; writing—review and editing. **Kyoung-Soon Jang:** Data curation; methodology. **Jianjun Wang:** Data curation; methodology; project administration; supervision; writing—review and editing.

ETHICS STATEMENT

The study in this article did not involve any trials on humans or animals.

CONFLICT OF INTERESTS

The authors declare no conflict of interests.

DATA AVAILABILITY

The *iDOM* open-source software package is implemented in R and available for download via Github (<https://github.com/jianjunwang/iDOM>).


SUPPORTING INFORMATION

Additional Supporting Information for this article can be found online at <https://doi.org/10.1002/mlf2.70002>.

ORCID

Fanfan Meng  <https://orcid.org/0000-0001-5718-1621>

Ang Hu  <https://orcid.org/0000-0002-5755-2442>

Kyoung-Soon Jang  <https://orcid.org/0000-0001-5451-5788>

Jianjun Wang  <https://orcid.org/0000-0001-7039-7136>

REFERENCES

- Dittmar T, Stubbins A. Dissolved organic matter in aquatic systems. In: Holland HD, Turekian KK editors. *Treatise on geochemistry*. 2nd ed. Amsterdam: Elsevier; 2014. p. 125–56.
- Cooper WT, Chanton JC, D'Andrilli J, Hodgkins SB, Podgorski DC, Stenson AC, et al. A history of molecular level analysis of natural organic matter by FTICR mass spectrometry and the paradigm shift in organic geochemistry. *Mass Spectrom Rev*. 2022; 41:215–39.
- Ruan M, Wu F, Sun F, Song F, Li T, He C, et al. Molecular-level exploration of properties of dissolved organic matter in natural and engineered water systems: a critical review of FTICR-MS application. *Crit Rev Environ Sci Technol*. 2023;53:1534–62.
- Fievre A, Solouki T, Marshall AG, Cooper WT. High-resolution Fourier transform ion cyclotron resonance mass spectrometry of humic and fulvic acids by laser desorption/ionization and electrospray ionization. *Energy Fuels*. 1997;11:554–60.
- Hu A, Han L, Lu X, Zhang G, Wang J. Global patterns and drivers of dissolved organic matter across earth systems: insights from H/C and O/C ratios. *Fundam Res*. 2024. In press. <https://doi.org/10.1016/j.fmre.2023.11.018>
- Kellerman AM, Kothawala DN, Dittmar T, Tranvik LJ. Persistence of dissolved organic matter in lakes related to its molecular characteristics. *Nat Geosci*. 2015;8:454–7.
- Koch BP, Dittmar T. From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter. *Rapid Commun Mass Spectrom*. 2006;20:926–32.
- Kim S, Kramer RW, Hatcher PG. Graphical method for analysis of ultrahigh-resolution broadband mass spectra of natural organic matter, the van Krevelen diagram. *Anal Chem*. 2003;75:5336–44.
- Hu A, Jang KS, Meng F, Stegen J, Tanentzap AJ, Choi M, et al. Microbial and environmental processes shape the link between organic matter functional traits and composition. *Environ Sci Technol*. 2022; 56:10504–16.
- Hu A, Jang KS, Tanentzap AJ, Zhao W, Lennon JT, Liu J, et al. Thermal responses of dissolved organic matter under global change. *Nat Commun*. 2024;15:576.
- Hu A, Choi M, Tanentzap AJ, Liu J, Jang KS, Lennon JT, et al. Ecological networks of dissolved organic matter and microorganisms under global change. *Nat Commun*. 2022;13:3600.
- Hu A, Cui Y, Bercovici S, Tanentzap AJ, Lennon JT, Lin X, et al. Photochemical processes drive thermal responses of dissolved organic matter in the dark ocean. *bioRxiv*. 2024. <https://doi.org/10.1101/2024.09.06.611638>
- Hu A, Meng F, Tanentzap AJ, Jang KS, Wang J. Dark matter enhances interactions within both microbes and dissolved organic matter under global change. *Environ Sci Technol*. 2023;57:761–9.
- Bramer LM, White AM, Stratton KG, Thompson AM, Claborne D, Hofmocker K, et al. ftmsRanalysis: an R package for exploratory data analysis and interactive visualization of FT-MS data. *PLoS Comput Biol*. 2020;16:e1007654.
- Ayala-Ortiz C, Graf-Grachet N, Freire-Zapata V, Fudyma J, Hildebrand G, AminiTabrizi R, et al. MetaboDirect: an analytical pipeline for the processing of FT-ICR MS-based metabolomic data. *Microbiome*. 2023;11:28.
- Koch BP, Dittmar T. From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter. *Rapid Commun Mass Spectrom*. 2016;20:926–32.
- LaRowe DE, Van Cappellen P. Degradation of natural organic matter: a thermodynamic analysis. *Geochim Cosmochim Acta*. 2011;75:2030–42.
- Hughey CA, Hendrickson CL, Rodgers RP, Marshall AG, Qian K. Kendrick mass defect spectrum: a compact visual analysis for ultrahigh-resolution broadband mass spectra. *Anal Chem*. 2001;73:4676–81.
- Song H-S, Stegen JC, Graham EB, Lee J-Y, Garayburu-Caruso VA, Nelson WC, et al. Representing organic matter thermodynamics in biogeochemical reactions via substrate-explicit modeling. *Front Microbiol*. 2020;11:531756.
- Danczak RE, Chu RK, Fansler SJ, Goldman AE, Graham EB, Tfaily MM, et al. Using metacommunity ecology to understand environmental metabolomes. *Nat Commun*. 2020;11:6369.
- Sleighter RL, Hatcher PG. The application of electrospray ionization coupled to ultrahigh resolution mass spectrometry for the molecular characterization of natural organic matter. *J Mass Spectrom*. 2007; 42:559–74.
- Hockaday WC, Purcell JM, Marshall AG, Baldock JA, Hatcher PG. Electrospray and photoionization mass spectrometry for the characterization of organic matter in natural waters: a qualitative assessment. *Limnol Oceanogr Methods*. 2009;7:81–95.
- D'Andrilli J, Cooper WT, Foreman CM, Marshall AG. An ultrahigh-resolution mass spectrometry index to estimate natural organic matter lability. *Rapid Commun Mass Spectrom*. 2015;29:2385–401.
- Kellerman AM, Dittmar T, Kothawala DN, Tranvik LJ. Chemodiversity of dissolved organic matter in lakes driven by climate and hydrology. *Nat Commun*. 2014;5:3804.
- Li XM, Sun GX, Chen SC, Fang Z, Yuan HY, Shi Q, et al. Molecular chemodiversity of dissolved organic matter in paddy soils. *Environ Sci Technol*. 2018;52:963–71.
- Tanentzap AJ, Fitch A, Orland C, Emilson EJS, Yakimovich KM, Osterholz H, et al. Chemical and microbial diversity covary in fresh water to influence ecosystem functioning. *Proc Natl Acad Sci USA*. 2019;116:24689–95.
- Mentges A, Feenders C, Seibt M, Blasius B, Dittmar T. Functional molecular diversity of marine dissolved organic matter is reduced during degradation. *Front Mar Sci*. 2017;4:194.
- Webb CO, Ackerly DD, McPeck MA, Donoghue MJ. Phylogenies and community ecology. *Ann Rev Ecol Syst*. 2002;33:475–505.
- Faith DP. Conservation evaluation and phylogenetic diversity. *Biol Conserv*. 1992;61:1–10.
- Wu L, Yang Y, Ning D, Gao Q, Yin H, Xiao N, et al. Assessing mechanisms for microbial taxa and community dynamics using process models. *mLife*. 2023;2:239–52.
- Stegen JC, Lin X, Konopka AE, Fredrickson JK. Stochastic and deterministic assembly processes in subsurface microbial communities. *ISME J*. 2012;6:1653–64.
- Wang J, Shen J, Wu Y, Tu C, Soininen J, Stegen JC, et al. Phylogenetic beta diversity in bacterial assemblages across ecosystems: deterministic versus stochastic processes. *ISME J*. 2013;7:1310–21.
- Naimi B, Araújo MB. sdm: a reproducible and extensible R platform for species distribution modelling. *Ecography*. 2016;39:368–75.
- Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *PLoS Comput Biol*. 2012;8:e1002687.
- Proulx S, Promislow D, Phillips P. Network thinking in ecology and evolution. *Trends Ecol Evol*. 2005;20:345–53.
- Zhu Y-G, Zhu D, Rillig MC, Yang Y, Chu H, Chen Q-L, et al. Ecosystem microbiome science. *mLife*. 2023;2:2–10.
- Blüthgen N, Menzel F, Blüthgen N. Measuring specialization in species interaction networks. *BMC Ecol*. 2006;6:9.
- Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'hara R, et al. Package 'vegan'. *Community Ecology Package*. 2013;2:1–295.
- Laliberté E, Legendre P, Shipley B. Package 'FD'. Measuring functional diversity from multiple traits, and other tools for functional ecology. Version 1.0-12.3. 2014. <https://doi.org/10.32614/CRAN.package.FD>
- Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, et al. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*. 2010;26:1463–4.
- Ning D, Yuan M, Wu L, Zhang Y, Guo X, Zhou X, et al. A quantitative framework reveals ecological drivers of grassland microbial community assembly in response to warming. *Nat Commun*. 2020;11:4717.

- 42 Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. *PLOS Comput Biol*. 2015;11: e1004226.
- 43 Watts SC, Ritchie SC, Inouye M, Holt KE. FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics*. 2019;35:1064–6.

How to cite this article: Meng F, Hu A, Jang K-S, Wang J. *iDOM*: Statistical analysis of dissolved organic matter characterized by high-resolution mass spectrometry. *mLife*. 2025;4:319–331. <https://doi.org/10.1002/mlf2.70002>