Supplementary Information for

Decomposing the molecular complexity and transformation of dissolved organic matter for innovative anaerobic bioprocessing

Jun Hu[1], Chuan-Guo Liu[1], Wen-Kai Zhang[1], Xue-Wen Liu[1], Bin Dong[2], Zhan-Dong Wang[2,3], Yuan-Guo Xie[1], Zheng-Shuang Hua[1], Xian-Wei Liu[1]*

[1]Chinese Academy of Sciences Key Laboratory of Urban Pollutant Conversion, Department of Environmental Science and Engineering, University of Science and Technology of China, Hefei 230026, China.

[2]State Key Laboratory of Fire Science, University of Science and Technology of China, Hefei 230026, China.

[3]National Synchrotron Radiation Laboratory, University of Science and Technology of China, Hefei, Anhui 230029, PR China.

*Corresponding author, Email: xianweiliu@ustc.edu.cn

This PDF file includes:

Supplementary Method 1. SPE-FT-MS experiments

Solid phase extraction (SPE) was performed prior to FT-MS analysis using solid phase extraction columns (HyperSep^{TM} retain PEP 60 mg, Thermo scientific). This column provides balanced retention of polar and non-polar analytes. When preparing biological samples prior to SPE: 1) Centrifuge and filter samples to remove any solid contamination; 2) Dilute biological sample with water containing 2% formic acid to a ratio 1:1 (v/v). According to the manufacturer's guidelines, the specific procedure is slightly modified as follows: (1) the SPE columns were first conditioned using 10 mL methanol (suitable for HPLC, Sigma-Aldrich) followed by 10 mL deionized water; (2) the filtered samples (4 mL) were loaded at a rate of 2-5 mL/min; (3) the columns were washed using 4 mL 5% methanol-deionized water; (4) Elute the columns with 4 mL methanol to clean tubes. The extraction efficiencies for all samples were calculated by measuring the TOC concentration before and after extraction, resulting in an average of 49 ± 4% (n = 7). The DOM eluates were kept in acid-washed and combusted amber glass vials at -20°C until they were ready for analysis on an FT-ICR-MS (Bruker Daltonics, Solarix XR, America) spectrometer with negative mode electrospray ionization.

Supplementary Method 2. Quantitative analysis of possible chemical reactions occurring between DOM molecules.

The metabolism of carbohydrates, proteins, and lipids involves intricate biochemical reactions, which are outlined in Supplementary Fig. 19. Through hydrolysis and a series of enzymatic steps, the macromolecules are decomposed into smaller molecules, such as monosaccharides, amino acids, glycerol, and long-chain fatty acids (LCFA), primarily through dehydration (hydrolysis) reactions. The LCFA metabolism is mainly driven by β-oxidation pathway, in which, successive cycles of four enzymatic reactions occur, resulting in the cleavage of two carbon units from the acyl-CoA chain. This process is a demethylation-like reaction. The metabolism of amino acids involves dehydration, hydration, dehydrogenation, hydrogenation, deamination, and decarboxylation reactions. The metabolism of small molecules that enter TCA cycles involves dehydration, hydration, dehydrogenation, and decarboxylation reactions. All these biochemical reactions are referenced from the textbook *Biochemistry* by Zhu and Xu[1]. The quantitative analysis of chemical reactions, such as demethylation/methylation, decarboxylation/carboxylation, dehydration/hydration, dehydrogenation/hydrogenation, deamination/amination, was calculated based on the functional group differences between the formulae of PD and SD samples using the following equations of chemical reactions.

Demethylation: $C_cH_hO_oN_nP_pS_s + a[H] \rightarrow C_{c-a}H_{h-3a}O_oN_nP_pS_s + a[CH_3]$

Decarboxylation: $C_cH_hO_oN_nP_pS_s + a[H] \rightarrow C_{c-a}H_{h-a}O_{o-2a}N_nP_pS_s + a[COOH]$

Dehydration: $C_cH_hO_oN_nP_pS_s \rightarrow C_cH_{h-2a}O_{o-a}N_nP_pS_s + a[H_2O]$

Dehydrogenation: $C_cH_hO_oN_nP_pS_s \rightarrow C_cH_{h-2a}O_oN_nP_pS_s + a[H_2]$

Deamination: $C_cH_hO_oN_nP_pS_s + a[O] \rightarrow C_cH_{h-3a}O_oN_{n-a}P_pS_s + a[NH_3]$

Based on these chemical equations, we introduce demethylation index (DMI), decarboxylation index (DCI), dehydration index (DHI), dehydrogenation index (DHGI) and deamination index (DAI), which are calculated by summing the number of functional group losses in biochemical reactions of each compound class. For example,

$$Demethylation\ index\ (DMI) = \sum_{i=0}^{i=m} a_i$$

where *i* is the number of unique molecules in the DOM of PD samples and *a* is the number of functional group losses in the demethylation reaction of one single molecule. Each index

represents the number of functional group losses in reactions occurring between PD and SD samples. A positive index indicates a net loss of functional groups, while a negative index signifies a net gain of functional groups in biochemical reactions from PD to SD samples. The high index values result from several factors: (1) the broad mass range distribution of DOM molecules, (2) the large number of molecular formulas per sample contribute to a higher number of potential formula pairs and thus higher index values, and (3) the indices being calculated as the cumulative sum across seven FWTPs, each with diverse DOM composition.

The calculations were conducted using the Perl programming language, and the code is available at https://github.com/hzhengsh/DOM_biochemical_reaction. All DOM molecules were used to quantify key biochemical reactions between PD and SD at each site. The results are presented in Fig. 4 and Supplementary Tables 3-7.

It is important to note that this analysis is based on all theoretically possible reactions between molecular formula pairs detected in PD and SD, and does not imply that these reactions actually occurred. Additionally, both compound classifications and biochemical reaction predictions are based on molecular formulas, without direct structural confirmation. While this approach may introduce certain limitations due to the reliance on established biochemical knowledge associated with these formulas, it offers a novel framework for predicting reactions. Additionally, the analysis does not account for the intensity of the molecular formulas, which could lead to some discrepancies between theoretical predictions and actual reaction rates or pathways. However, these constraints highlight the potential for further refinement of our method.

Supplementary Note 1. Microbial composition and metabolism in the AD of seven FWTPs

Microbial diversity showed a significant increase from primary digestion (PD) to secondary digestion (SD) (Supplementary Fig. 4a), reflecting an expanded functional capacity for transforming diverse DOM molecules during the SD stage. This increase in diversity is often associated with greater metabolic versatility, facilitating the breakdown of complex organic matter.

Among the DOM compound classes, lipids were the only group to exhibit a significant increase from PD to SD, indicating lipid accumulation (Fig. 2b and Fig. 3c). This finding aligns with microbial functional analyses, which revealed a slight reduction in lipid metabolism during SD compared to PD (Supplementary Fig. 4b). Notably, despite the more favorable fermentation conditions in SD (pH 7.77 vs. pH 3.87 in PD), lipid metabolism remained inactive and was even significantly inhibited, contributing to the observed lipid accumulation. Lignin-like compounds, the most abundant compound class and characterized by their richness in aromatic rings, showed a significant decrease. This reduction corresponds to the enhanced microbial metabolism of aromatic compounds, including cofactors, vitamins, terpenoids, and polyketides. Proteins, the third most abundant class, showed minimal changes, likely due to ammonia-induced inhibition of microbial activity. This explanation is supported by the active metabolism of amino acids to ammonia observed during SD (Supplementary Fig. 4b). These findings provide a nuanced understanding of how microbial community dynamics and metabolic functions drive DOM transformations in AD systems.

Additionally, Mantel analysis confirmed a strong correlation between microbial composition changes and shifts in DOM compound classes such as amino sugars, carbohydrates, and lipids, as well as elemental combinations like CHO and CHONS. Molecular structure evolution, including DBE changes, was also closely linked to microbial functions (Supplementary Fig. 4c). These results demonstrate a clear alignment between microbial metabolism and DOM molecular transformations, thereby enhancing the validity and robustness of our findings.

Supplementary Note 2. Chinese food diet culture and food waste treatment

(1) The Chinese diet culture: Chinese cuisine is one of the most diverse and celebrated culinary traditions globally, renowned for its intricate flavors, an extraordinary variety of ingredients, and a broad spectrum of cooking techniques (Supplementary Fig. 3a). The culinary diversity is deeply rooted in China's vast geography, encompassing different climates, terrains, and ecosystems, which influence the availability of ingredients in various regions. Regional cuisines, such as Sichuan, Guangdong, Shandong, and Zhejiang, each have unique characteristics, from the bold spiciness of Sichuan dishes to the delicate sweetness of Guangdong dim sum. China's cultural history and the culinary practices of its 56 ethnic groups have further enriched its food culture, incorporating a wide array of spices, herbs, and flavoring techniques[2]. Chinese meals often emphasize a balance between flavors—sweet, sour, salty, bitter, and umami—while showcasing diverse textures and presentations. Common cooking techniques include stir-frying, steaming, braising, and roasting[3]. Typical Chinese diets are carbohydrate-rich, often featuring staples like rice, noodles, and steamed buns, complemented by protein sources such as pork, chicken, seafood, and tofu. Fat content is often high due to the widespread use of cooking oils in stir-fries and deep-frying methods[4]. This diversity not only reflects the cultural heritage of China but also results in food waste with a highly complex molecular composition.

(2) Food waste treatment processes: Food waste is processed in dedicated digesters through a series of intricate steps, beginning with collection from diverse urban sources and transportation to treatment facilities. At the food waste treatment plants (FWTPs), waste undergoes pretreatment, which typically includes shredding, screening, and dewatering to remove non-organic contaminants and prepare the material for further processing (Supplementary Fig. 3b). The pretreatment stage is followed by an oil recovery process, where waste oils and fats are separated for potential reuse or disposal. The remaining organic material is then allocated to anaerobic digesters for sequential PD and SD. The entire treatment cycle involves aerobic, anoxic, and anaerobic processes, each playing a critical role in degrading organic solids, transforming them into DOM, and generating biogas. The complexity of these stages, combined with the diverse origins of food waste, introduces a wide array of molecular compounds into the waste stream. As a

result, the treatment process significantly increases the molecular diversity of DOM, reflecting the broad spectrum of organic inputs and microbial transformations occurring throughout the digestion cycle. This diversity highlights the necessity of robust analytical methods to characterize and manage organic matter transformations during food waste treatment.

Supplementary Note 3. The dominance of reduced compounds in the final effluent.

The increased abundance of reduced compounds in the SD phase can be attributed to several factors related to both the nature of the substrates and the microbial processes occurring during anaerobic fermentation:

(1) Limited oxidation under anaerobic conditions: In stand-alone digesters treating food waste, the anaerobic conditions limit the extent of oxidation of organic molecules. This limitation results in the accumulation of reduced compounds, as oxidative pathways are less active. In contrast, digesters at water resource recovery facilities, such as those treating excess sludges from processes like A2O (Anaerobic-Anoxic-Oxic), handle materials that have already undergone partial oxidation during the wastewater treatment process. This difference in the treatment conditions results in a greater proportion of reduced molecules in FWTPs, as these compounds are not sufficiently oxidized during the AD process. The reduced state of the molecules in the effluent DOM in FWTPs is therefore expected.

(2) Persistence of more recalcitrant compounds: Lignin and lipids are the two most abundant components in the effluent DOM. Lignin is primarily degraded through oxidative processes, such as dioxygenase-catalyzed ring-opening, while lipids undergo degradation through β-oxidation. These degradation pathways are generally less thermodynamically favorable compared to the breakdown of carbohydrates and other easily degradable substrates. As a result, lipids and lignin under reduced states tend to persist in the organic-rich environment of the AD process.

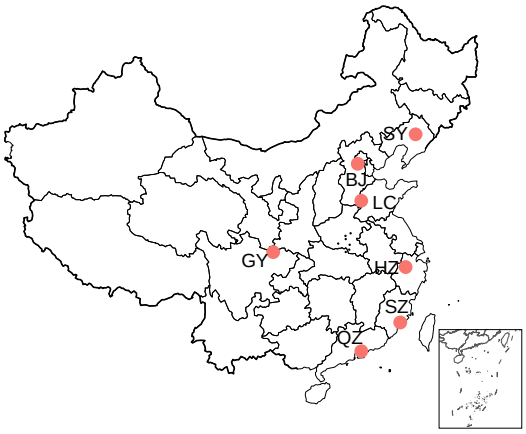
(3) Microbial biomass as a DOM source: The solubilization of microbial cell biomass is likely contributes to the effluent DOM pool, especially in the later stages of the digestion process. The shift toward more reduced DOM in secondary digestion (SD) is further influenced by the decomposition of microbial biomass, which is more resistant to degradation than more bioavailable food waste. As a result, it tends to persist as persistent DOM in the final effluent.

Supplementary Note 4. Challenges in directly determining molecular structures of DOM

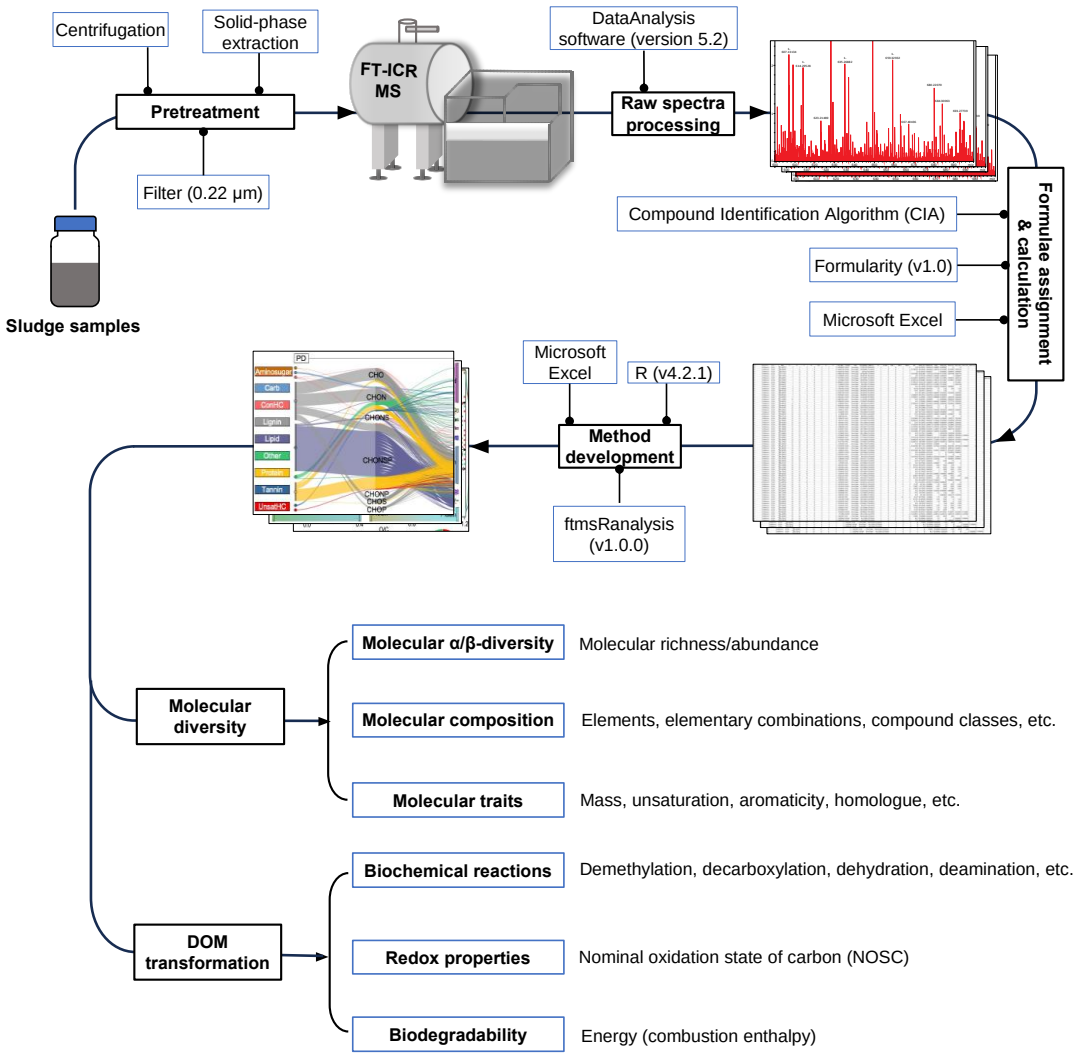
(1) Inherent complexity of DOM molecular composition. DOM from complex environmental systems comprises an enormous range of molecular sizes, functional groups, and chemical compositions. This diversity makes it difficult to characterize specific structures of individual molecules. Even with advanced techniques like MS/MS or NMR, only a fraction of such complexity can be definitively characterized. Achieving an exhaustive direct structural analysis across all compounds is therefore technically unfeasible for complex matrices like food waste digestate.

(2) Balancing broad compound detectability with in-depth structural characterization. FT-ICR MS is the only MS technique offering sub-ppm mass accuracy for various peaks in a complex mixture, making it the most advanced tool for unraveling molecular complexity with ultra-high resolution. However, FT-ICR MS alone cannot directly measure specific subsets of DOM structures without more specialized techniques (e.g., 2D-NMR), which, however, may not capture the full complexity of DOM. Therefore, it’s difficult to balance broad compound detectability with in-depth structural characterization for DOM analysis.

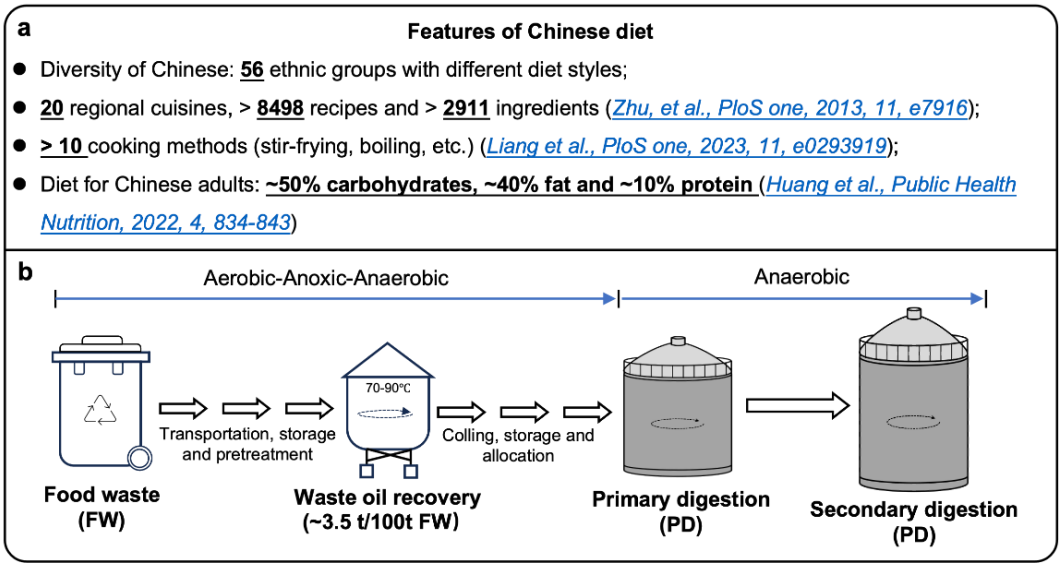
(3) Technical constraints for Integrating different MS techniques. Integrating different MS techniques for comprehensive interpretation is extremely challenging due to differences in resolution, mass accuracy, analyte detection, ionization methods, and sample preparation. Each technique provides unique insights but also generates data in different formats, often capturing distinct subsets of a sample. These inherent differences make it difficult to achieve a unified, coherent interpretation of results from multiple MS platforms, requiring specialized workflows, considerable expertise, and often significant manual effort to align the datasets.



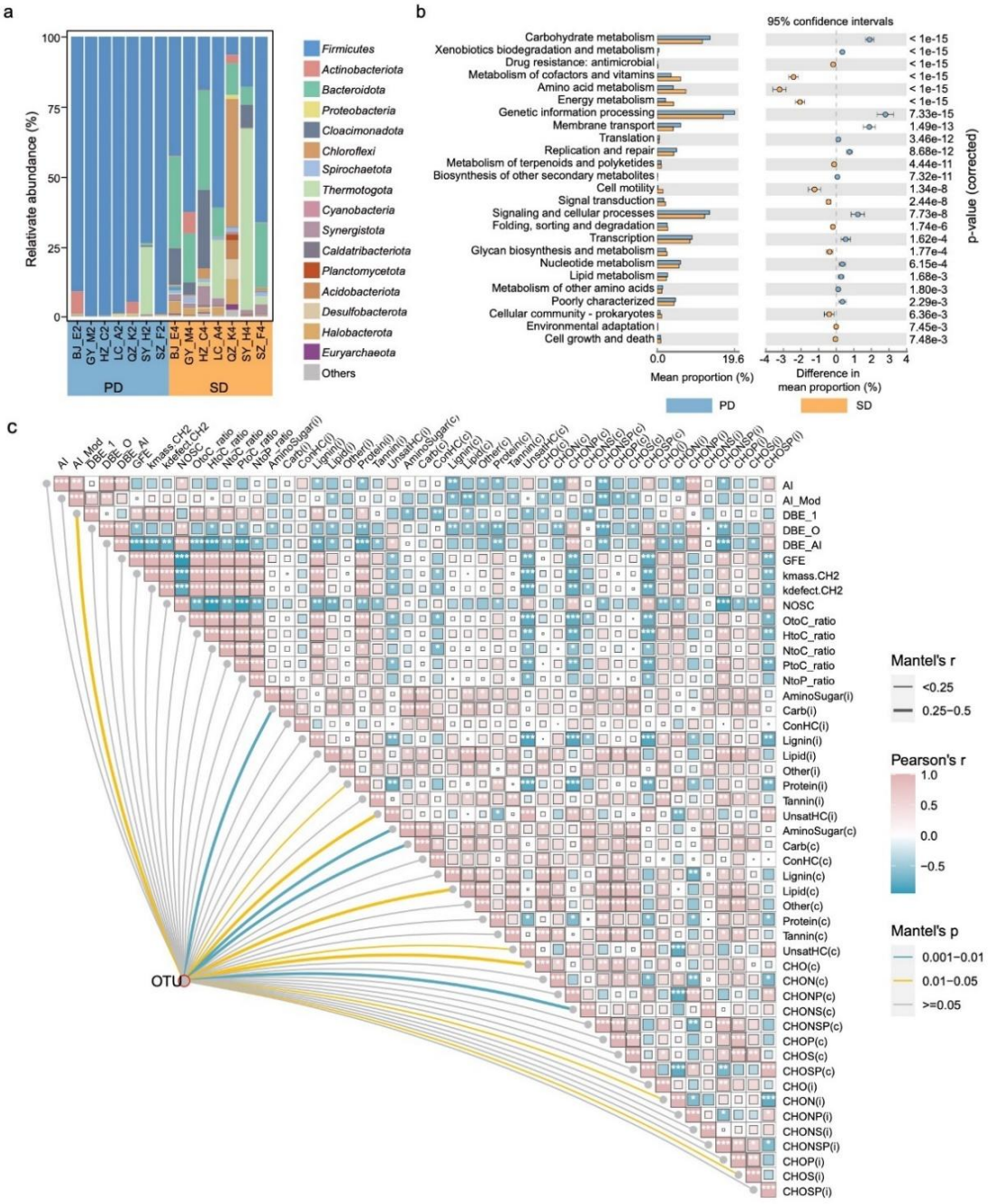
Supplementary Fig. 1 | Geographical distribution of 7 full-scale food waste treatment plants (FWTPs). Sludge samples were collected from the two-stage digesters located in Beijing (BJ), Guangyuan (GY), Hangzhou (HZ), Liaocheng (LC), Quanzhou (QZ), Shenyang (SY), Shenzhen (SZ) across China. For specific details about these regions, refer to Supplementary Table 9.



Supplementary Fig. 2 | Workflow for FT-MS experiments and data processing in this study. The pre-treatment process, involving centrifugation, filtration, and solid-phase extraction, is essential for concentrating and purifying samples to ensure precise analysis. DOM is then measured by FT-ICR MS according to the standard methods detailed in the Methods section. The resulting mass data are analyzed using both established software packages and newly developed methods, focusing on molecular diversity and DOM transformations in a complex water environment. Molecular diversity analysis includes molecular α/β diversity, composition, and structural traits, while DOM transformation encompasses biochemical reaction predictions, redox properties, and biodegradability as key components of this analysis framework.

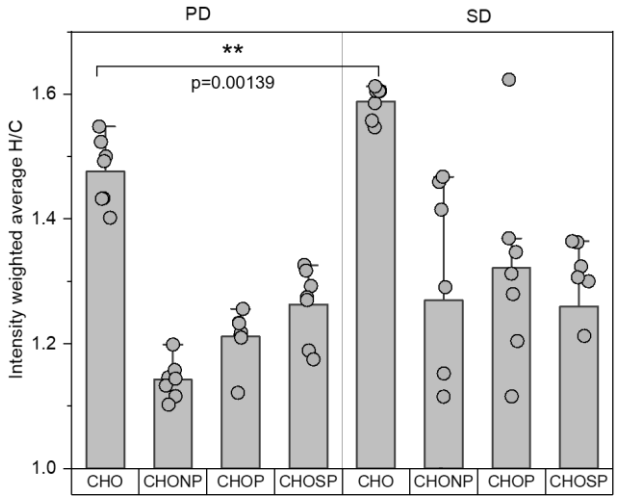


Supplementary Fig. 3 | Typical features of Chinese food and food waste treatment process. (**a**) Diversity and features of Chinese cuisines. Chinese cuisines are known for their incredible diversity, shaped by regional and cultural variations across the country. High carbohydrate and fat characteristics are evident in various regional dishes. (**b**) Overview of the food waste treatment process. Food waste is collected and transported to full-scale FWTPs. After undergoing pretreatment processes such as centrifugation, filtration, and solid-phase extraction, the waste progresses to oil recovery and subsequently to anaerobic digestion. The entire treatment cycle includes aerobic, anoxic, and anaerobic stages. The varied food diets and the complex treatment processes result in exceptionally high diversity of DOM. A more detailed discussion can be found in Supplementary Note 2.

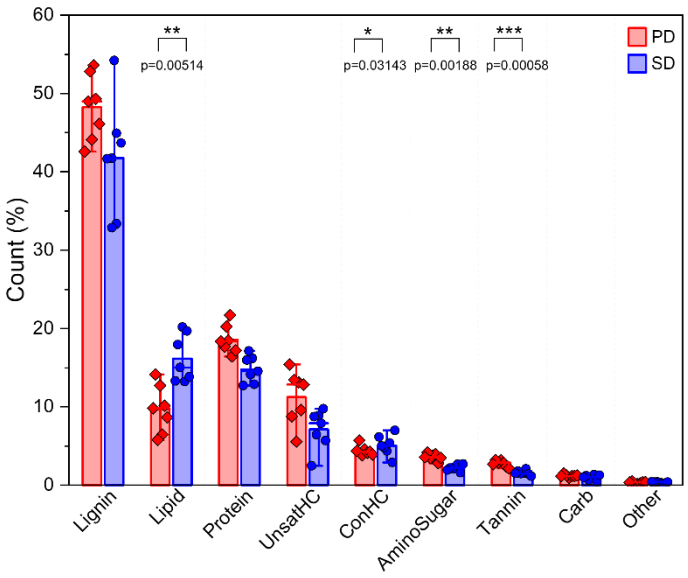


Supplementary Fig. 4 | Microbial community composition and functional analysis in the anaerobic digestion (AD) of seven food waste treatment plants (FWTPs). (**a**) Relative abundance of the major microbial groups and (**b**) functional analysis in PD-SD two-stage AD processes based on amplicon sequencing of the 16S rRNA gene; (**c**) Pairwise comparisons of DOM molecular characteristics are shown, with a color gradient denoting Spearman's correlation coefficient. Taxonomic community composition was related to each molecular parameter by Mantel tests. Edge width corresponds to the Mantel's r statistic for the corresponding distance correlations, and edge color denotes the statistical significance. The "i"

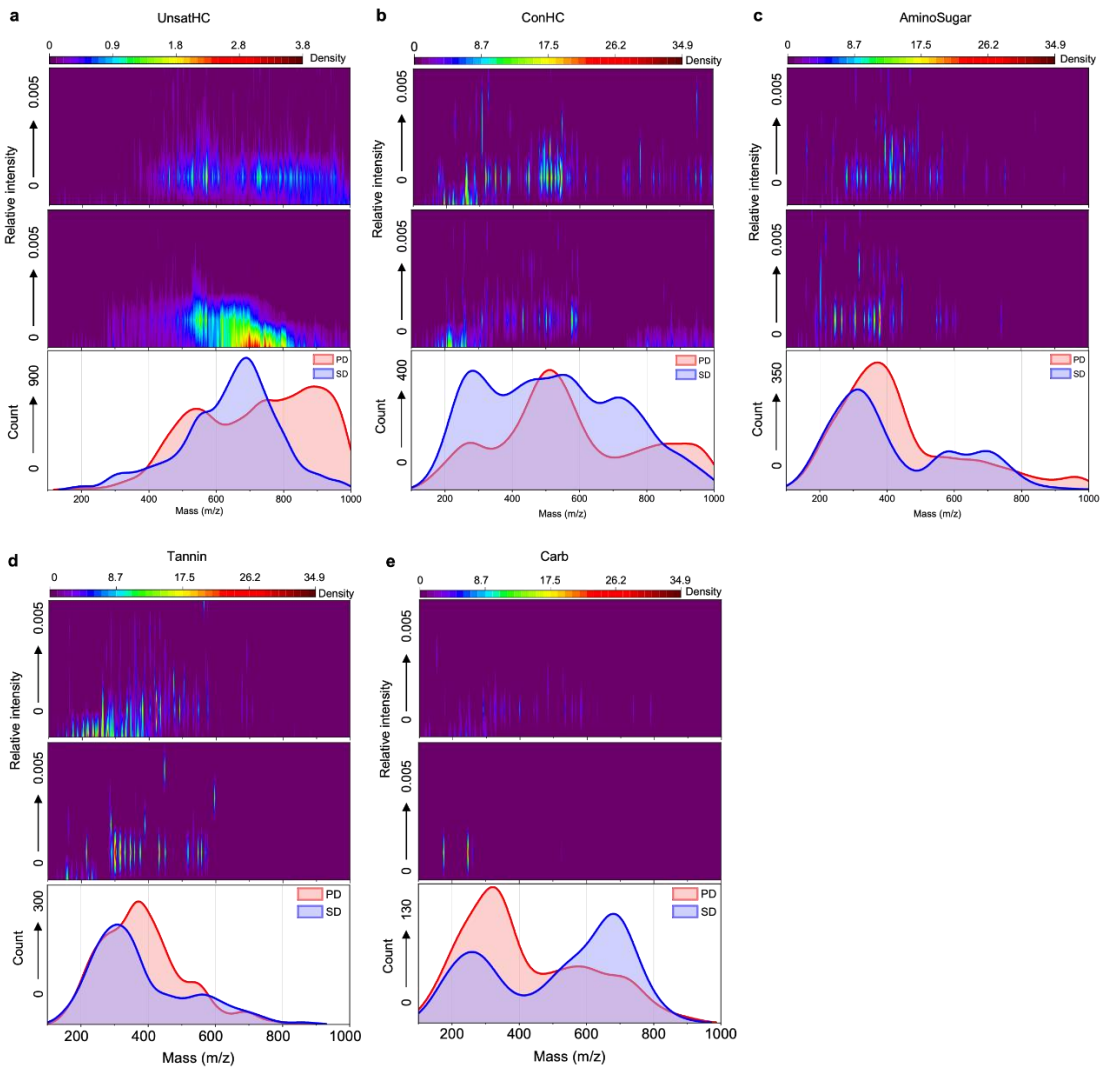
and “c” in brackets represent “molecular intensity composition” and “molecular count composition”, respectively. A more detailed discussion can be found in Supplementary Note 1.



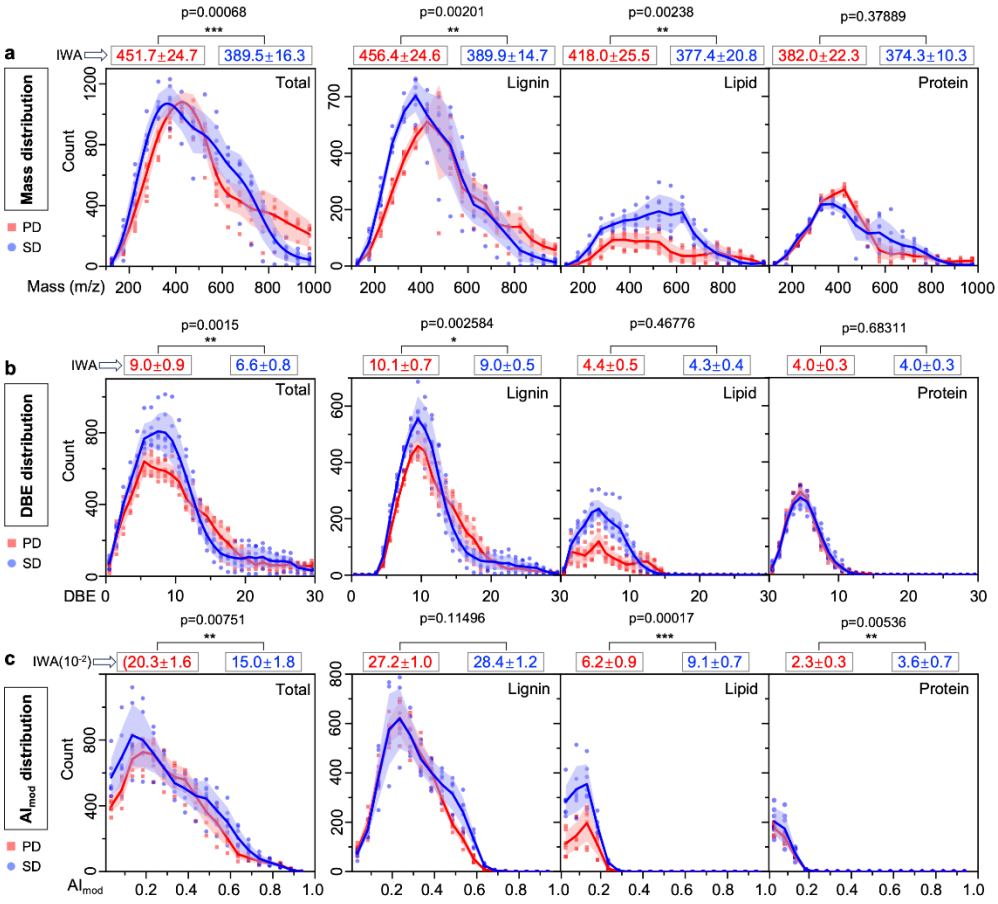
Supplementary Fig. 5 | Intensity-weighted average H/C ratio of CHO, CHONP, CHOP, and CHOSP elemental combinations in the DOM of PD and SD samples. The H/C ratio of CHO molecules is notably higher in the SD stage compared to the PD stage, and CHO molecules in both stages exhibit an obviously higher H/C ratio than those in the other elemental combinations. n = 7 biologically independent samples. Significance was evaluated by pair-sample t-test, *** represents $p<0.001$, ** represents $p<0.01$, * represents $p<0.05$, and no asterisk indicates $p > 0.05$.



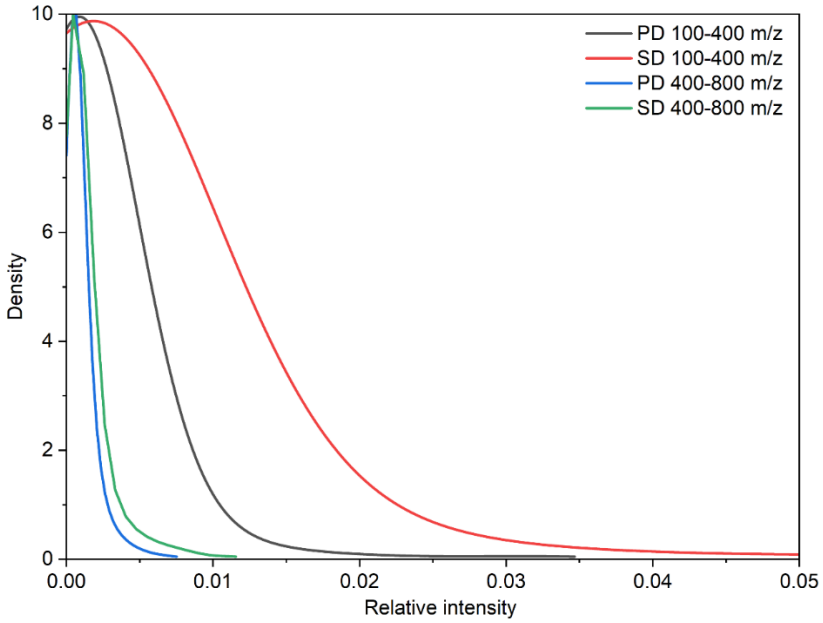
Supplementary Fig. 6 | Comparative analysis of molecular compositions between PD and SD based on molecular counts of each compound class. Lignin, lipid, and protein-like compounds are the three most abundant compound classes. Most compound classes showed a decline to different degrees from PD to SD, except for lipid- and ConHC-like compounds. n = 7 biologically independent samples. Significance was evaluated by pair-sample t-test, *** represents $p<0.001$, ** represents $p<0.01$, * represents $p<0.05$, and no asterisk indicates $p>0.05$.



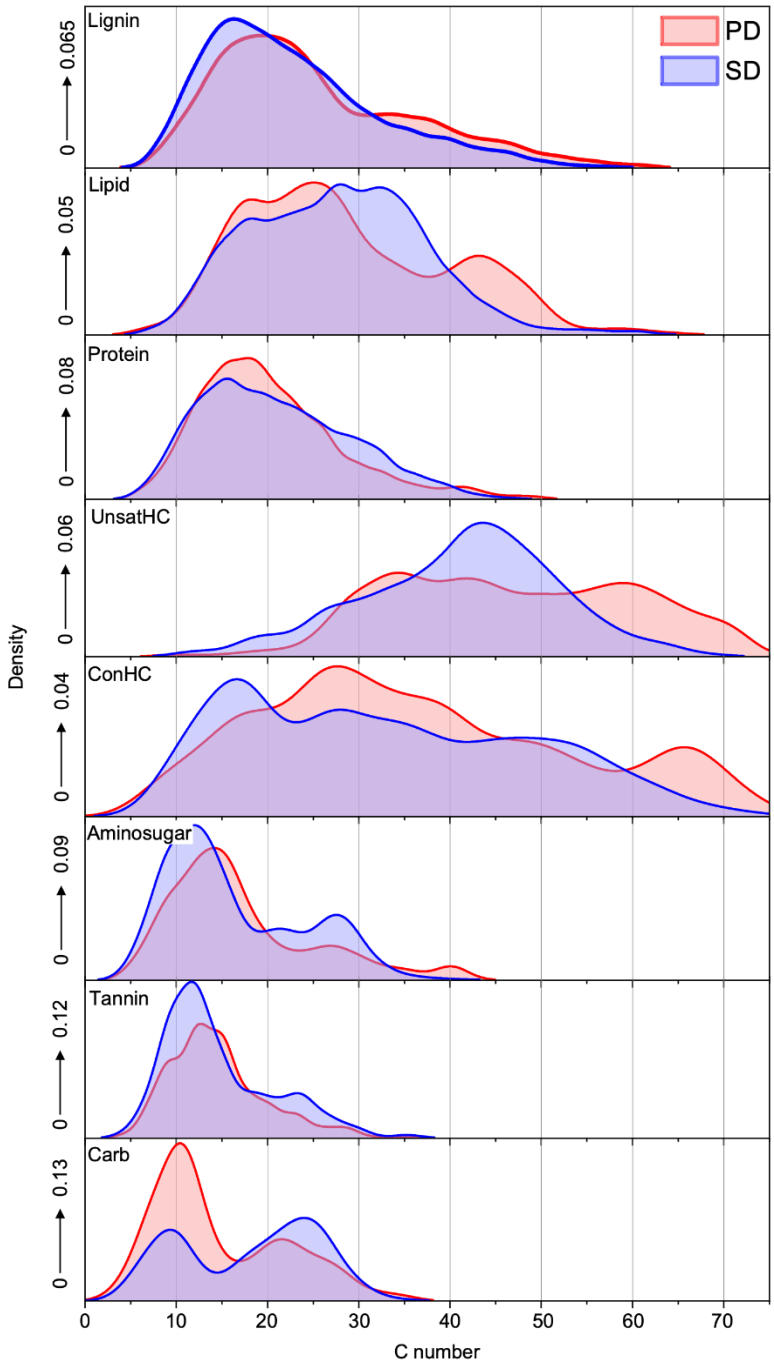
Supplementary Fig. 7 | Mass and corresponding relative intensity (RI) distribution of (a) UnsatHC, (b) ConHC, (c) AminoSugar, (d)Tannin, and (e) Carb. All DOM data from seven digesters were analyzed and compared between PD and SD. The density plots of mass-RI were generated using a 2D Kernel density plot, with the mass range from 100 to 1000 m/z and RI from 0 to 1, using a 1000 x 1000 grid.



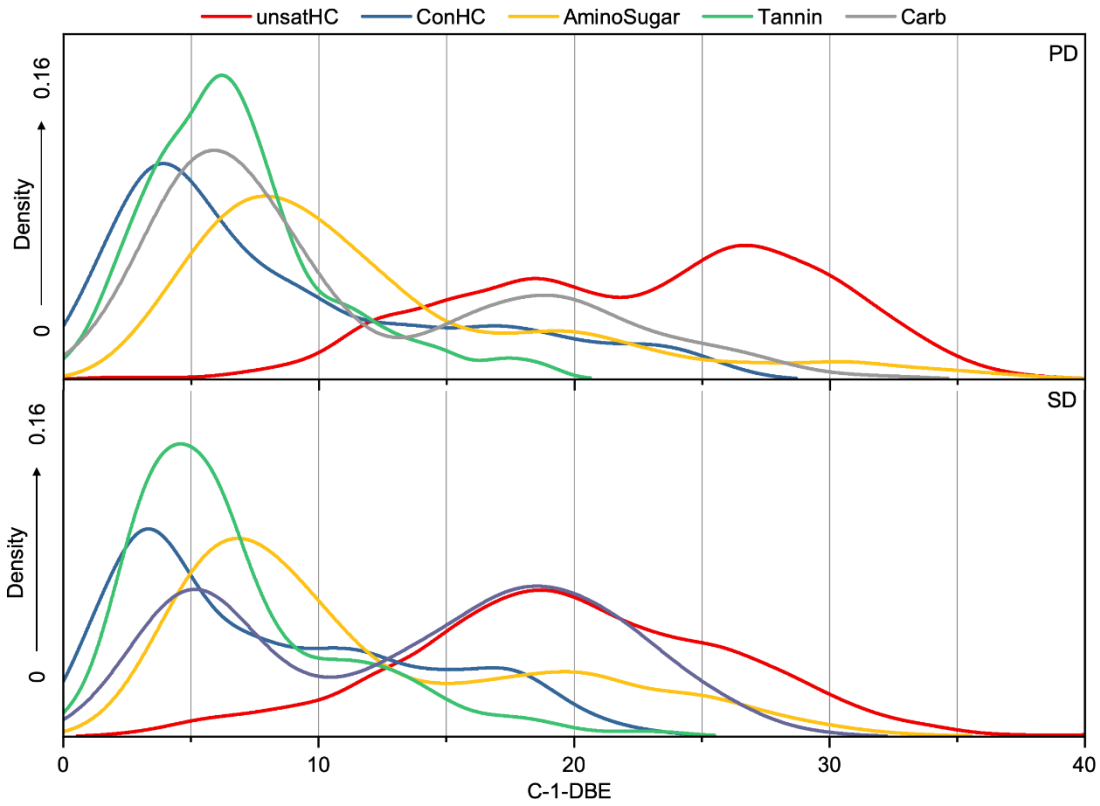
Supplementary Fig. 8 | Comparing molecular characteristics of DOM between PD and SD sludge samples. (**a**) Mass distribution, (**b**) double bound equivalent (DBE) distribution, and (**c**) modified aromaticity index (AI_{mod}) distribution for lipid, lignin, and protein compound classes. The distribution data represent average results from seven digesters and serve as supplementary comparisons to the full data set presented in **Fig. 2**. Intensity-weighted averages (IWA) of each index for lipid, lignin, and protein compound classes are calculated using n = 7 biologically independent samples. Significance was evaluated by pair-sample t-test, *** represents $p<0.001$, ** represents $p<0.01$, * represents $p<0.05$, and no asterisk indicates $p > 0.05$. Most parameters showed significant decreases from PD to SD, reflecting the effective performance of the AD process.



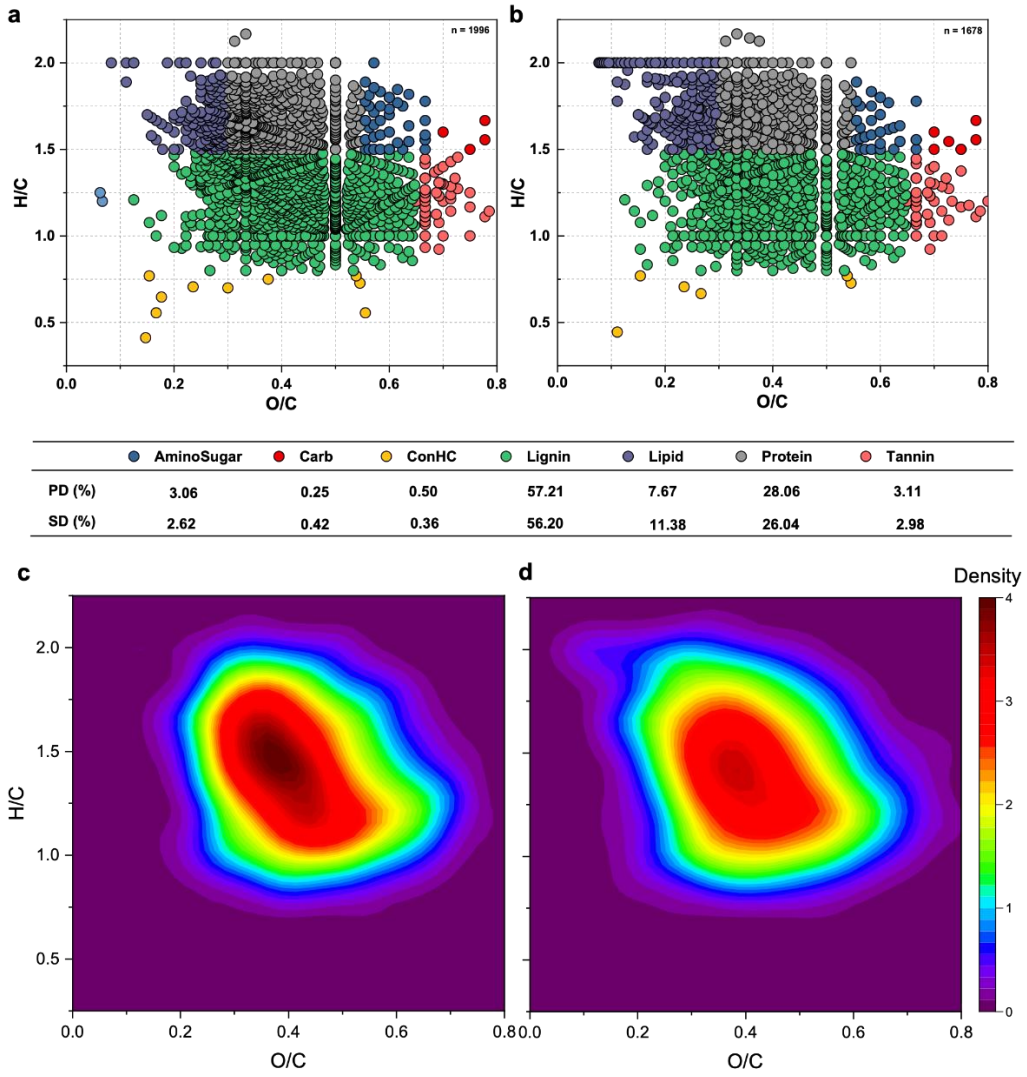
Supplementary Fig. 9 | Relative intensity distribution of high-molecular weight molecules (400-800 m/z) and smaller molecules (100-400 m/z) for lipid-like compounds in the PD and SD stages. Small lipid-like molecules display higher intensities than their high-molecular-weight counterparts. Moreover, SD samples show a significantly increased density of small molecules compared to PD samples. This pattern suggests that the SD process not only promotes molecular diversity but also leads to a higher concentration of smaller, more persistent lipid molecules in the effluent.



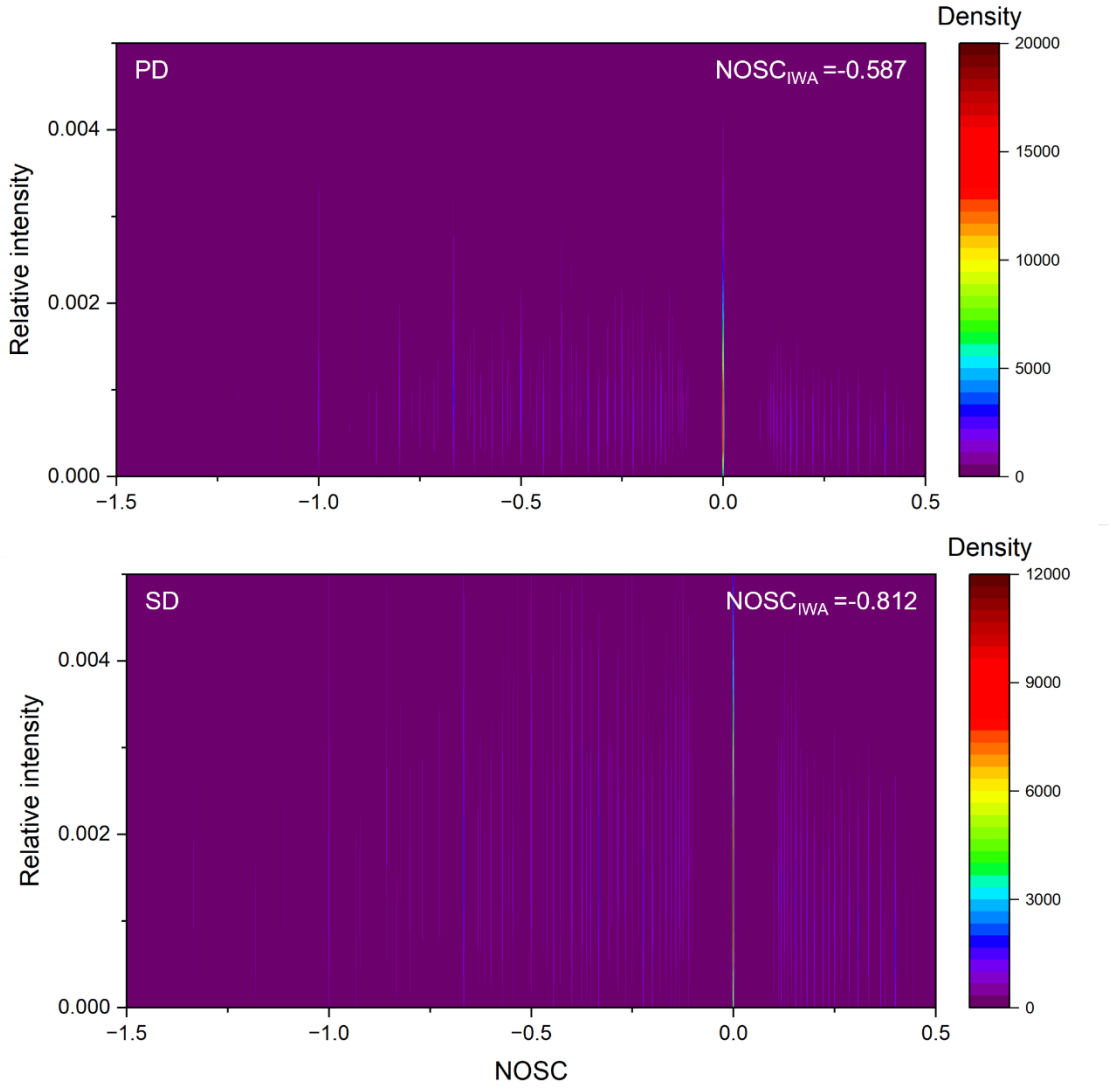
Supplementary Fig. 10 | Carbon distribution of Lignin, Lipid, Protein, UnsatHC, ConHC, AminoSugar, Tannin and Carb compounds. All compound classes showed a significant decrease in high-carbon distribution, such as lignin (C30-C60), lipids (C30-C45), and proteins (C40-50). Most compound classes, including Lignin, UnsatHC, ConHC, AminoSugar, and Tannin, exhibited an increased distribution in low-carbon molecules.



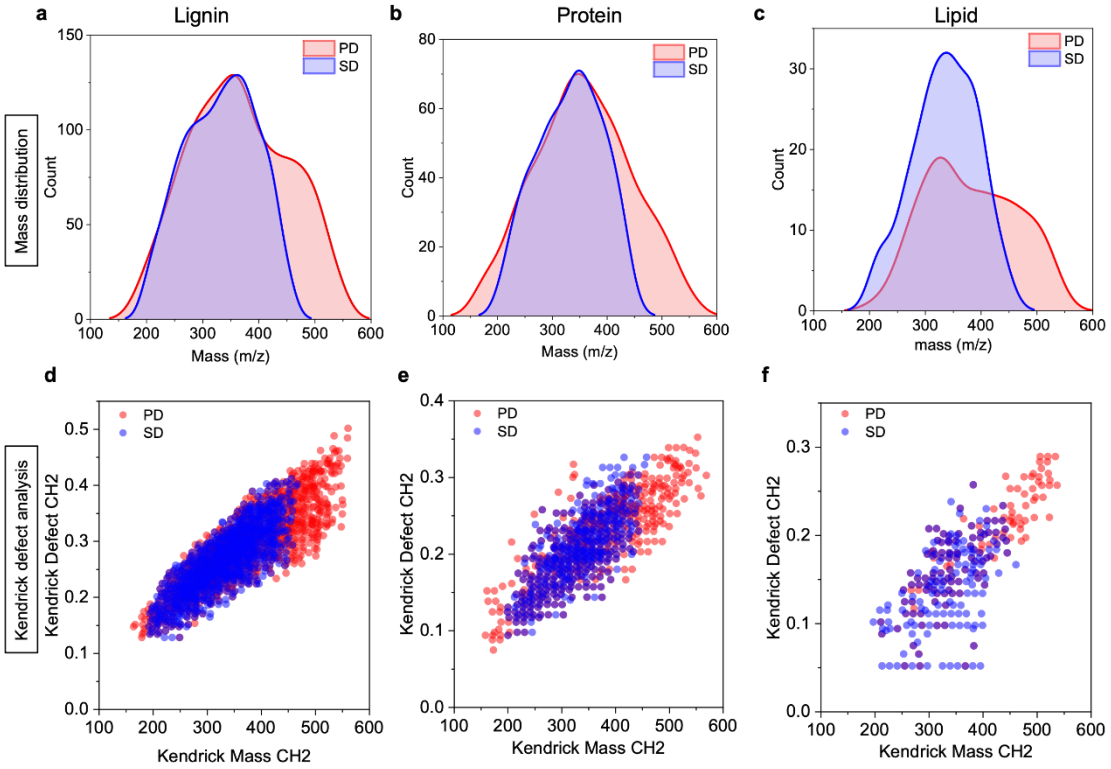
Supplementary Fig. 11 | C-C single-bond (C-1-DBE) distribution of UnsatHC, ConHC, AminoSugar, Tannin and Carb compounds in the DOM of PD and SD samples.



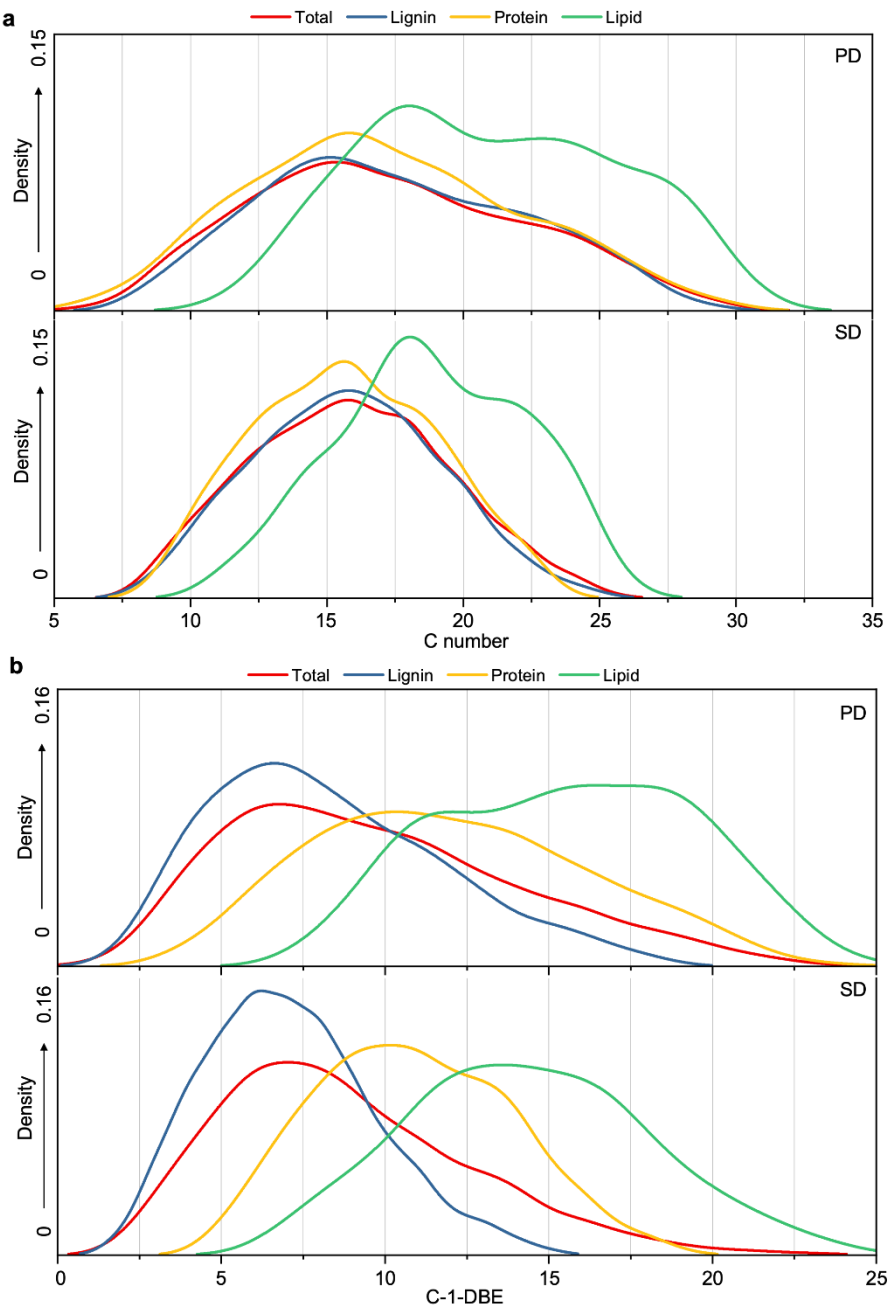
Supplementary Fig. 12 | The van Krevelen plots for core molecular communities of DOM in the (a) PD and (b) SD, with corresponding molecular distribution heatmaps for (c) PD and (d) SD. In comparison, the PD core molecular community has 1996 formulae, whereas the SD core molecular community has 1678 formulae.



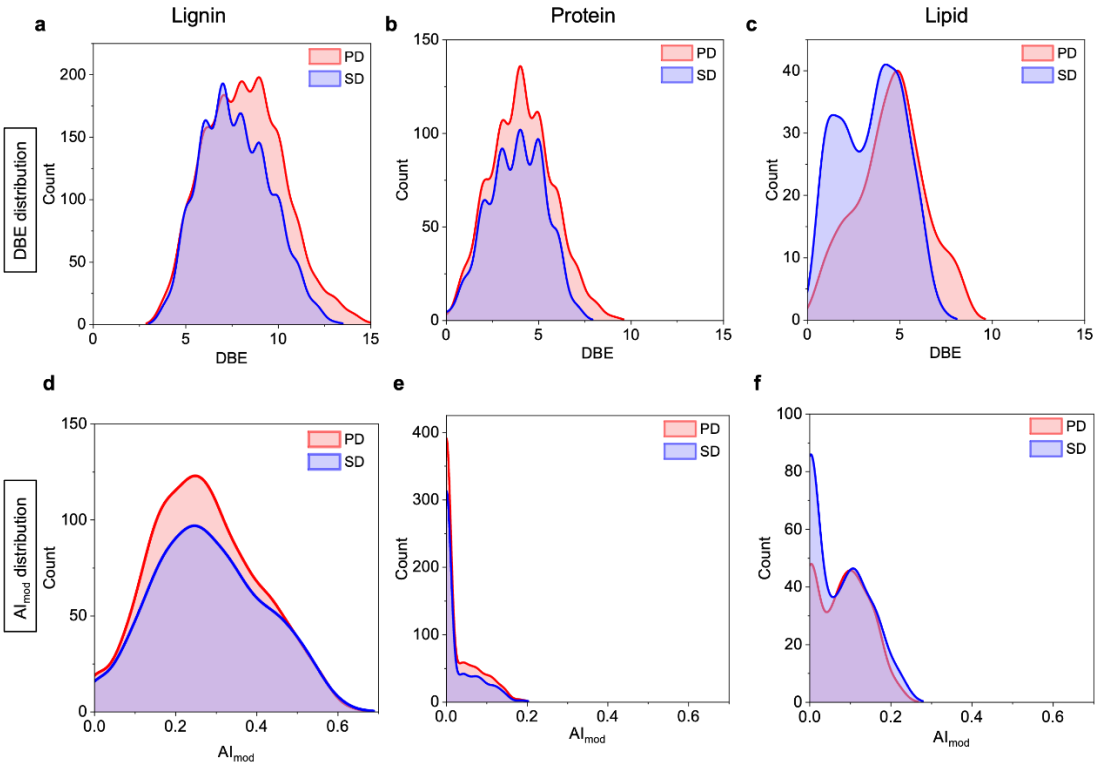
Supplementary Fig. 13 | Heatmap for NOSC-relative intensity (RI) distribution of core molecules of PD and SD samples. The density plots of NOSC-RI were generated using a 2D Kernel density plot, with the mass range from 100 to 1000 m/z and RI from 0 to 1, using a 1000 x 1000 grid. The figure shows that NOSC values in PD are generally lower than in SD, particularly at negative NOSC values, indicating a shift toward a more reduced state from PD to SD processes.



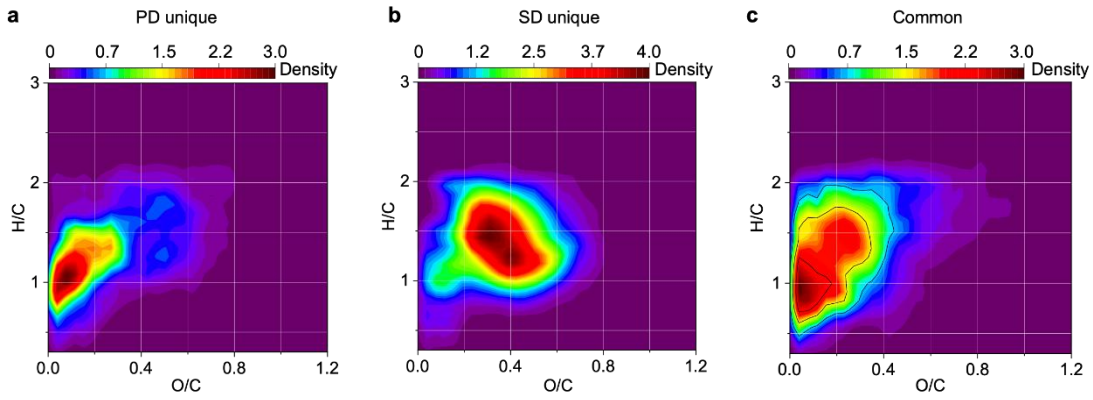
Supplementary Fig. 14 | Mass distribution and Kendrick defect analysis of key compound classes. Mass distribution of (a) lignin-, (b) protein- and (c) lipid-like molecules. Kendrick defect CH2 analysis for (d) lignin-, (e) protein- and (f) lipid-like molecules in the core molecular community of PD and SD samples.



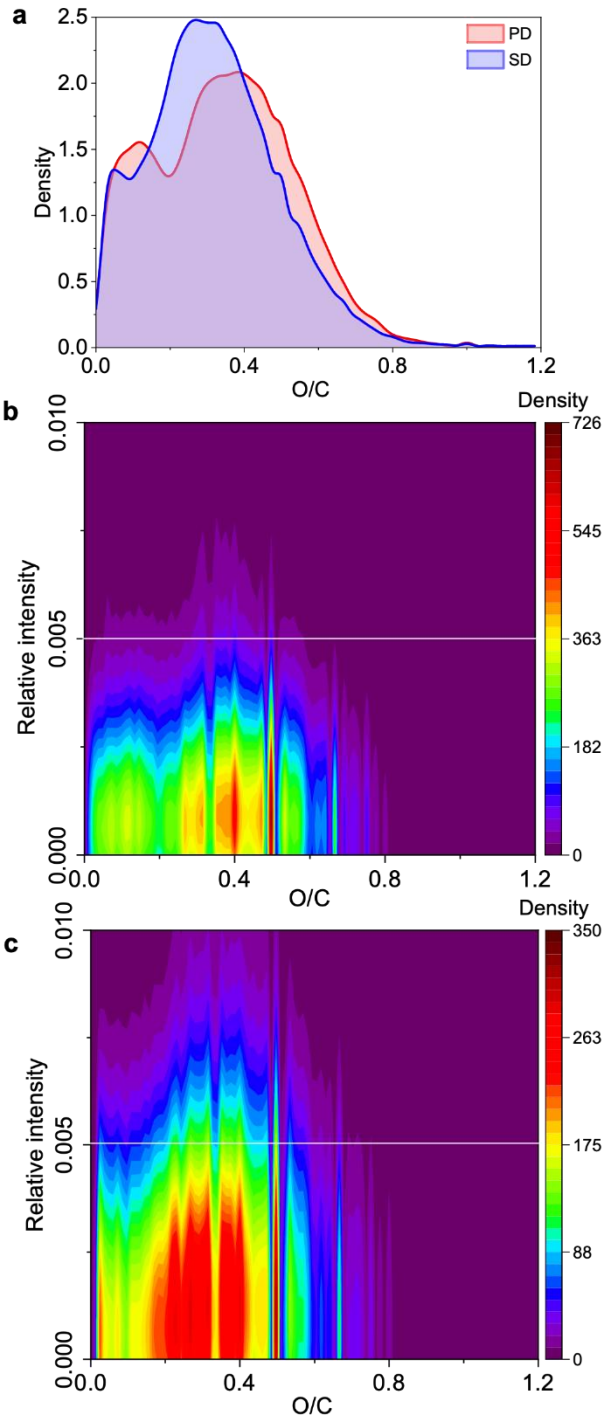
Supplementary Fig. 15 | (a) Carbon and (b) C-C single bond distribution of all molecules, lignin-, protein- and lipid-like molecules of core molecular community for PD and SD samples. The data show that carbon chain length and the proportion of C-C single bonds significantly affect the efficiency of anaerobic degradation, particularly the ratio of C-C single bonds. The peaks for C-1 DBE distribution follow the order: lipids > proteins > lignin, which corresponds to the degradation rate: lignin > proteins > lipids (Fig. 2b and Fig. 3f).



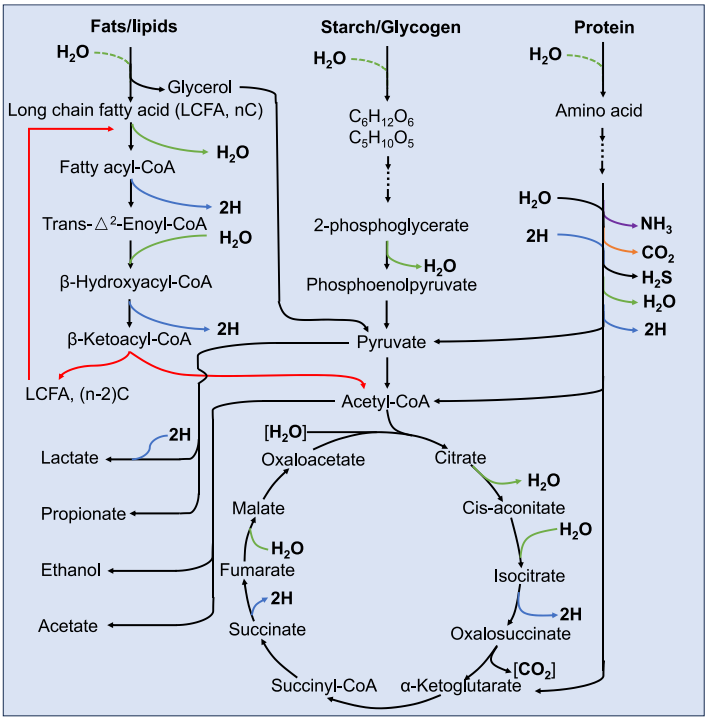
Supplementary Fig. 16 | Structural characterization of the core molecular community of PD and SD samples. DBE distribution of (**a**) lignin-, (**b**) protein- and (**c**) lipid-like molecules. AI_{mod} distribution for (**d**) lignin-, (**e**) protein- and (**f**) lipid-like molecules. The degree of unsaturation shifts toward smaller DBE distributions, while changes in aromaticity distribution are not significant for the key compound classes between PD and SD core molecular communities.



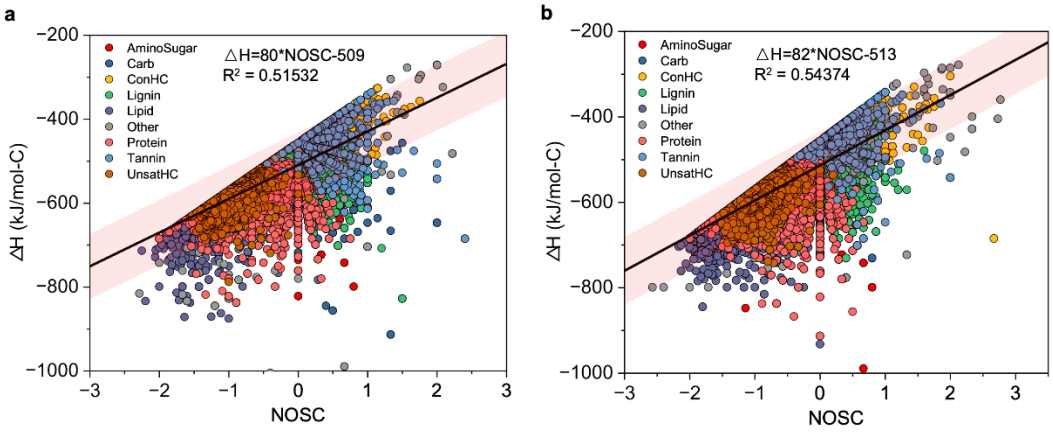
Supplementary Fig. 17 | Density plots of molecules unique to PD samples (a), unique to SD samples (b), and common to PD and SD samples (c) based on the van Krevelen plots.



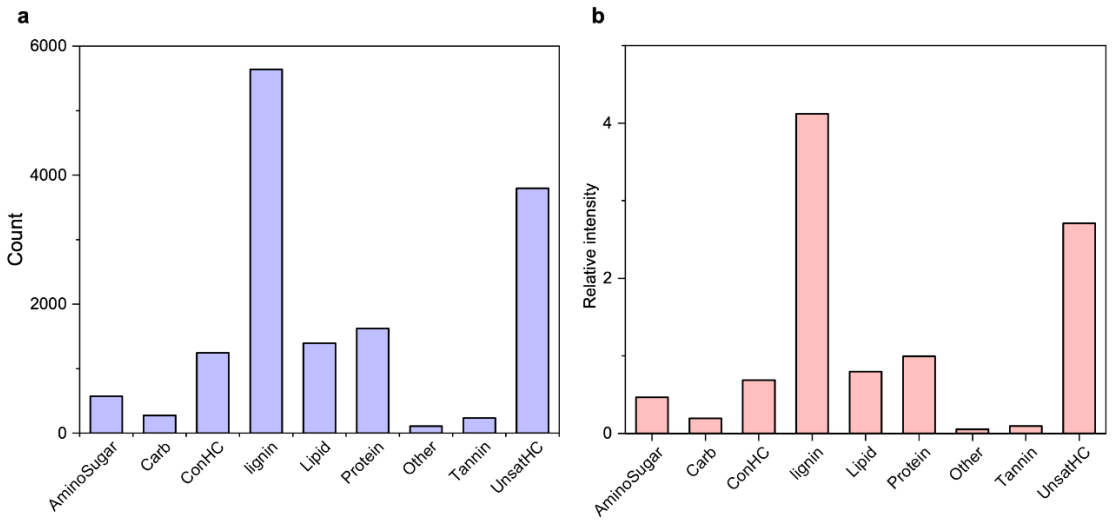
Supplementary Fig. 18 | O/C distribution of total DOM molecules in PD and SD stages. (**a**) O/C distribution based on molecular counts. Relative intensity distribution of O/C in (**b**) PD and (**c**) SD samples. Molecules with lower O/C ratios show significantly higher intensities, especially in the SD stage, suggesting that molecular composition has shifted toward lower O/C ratios.



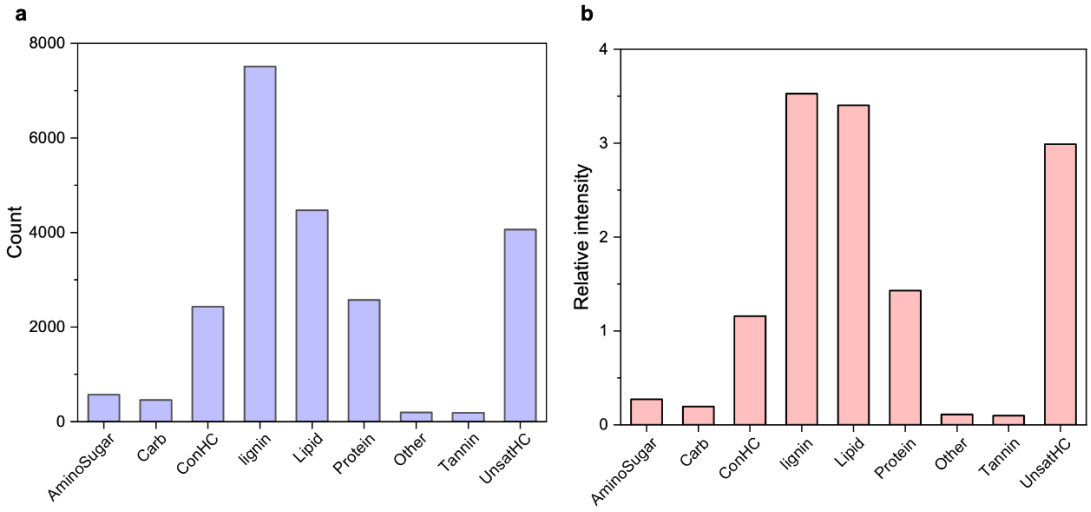
Supplementary Fig. 19 | Metabolic pathways for key organic matters that are presented in various water environments. All the metabolic pathways can be found in the textbook *Biochemistry* by Zhu and Xu[1]. Since carbohydrates, lipids, and proteins are also the primary classes in food, we believe these biochemical reactions, such as demethylation/methylation (red), decarboxylation/carboxylation (pink), dehydration/hydration (green), dehydrogenation/hydrogenation (blue), and deamination/amination (purple) reactions widely occur in the AD process.



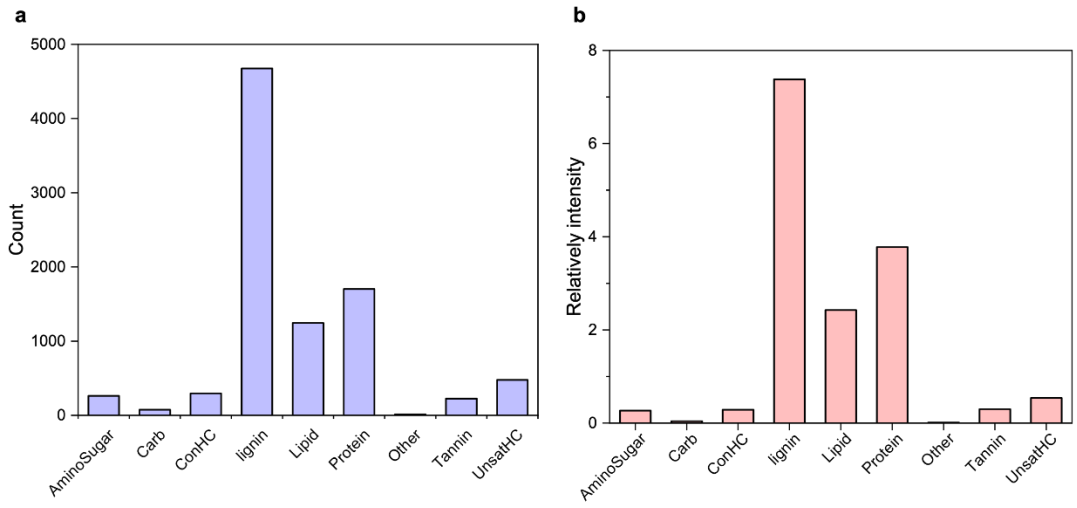
Supplementary Fig. 20 | The relationship between energy density (Δ*H*) and NOSC for whole DOM molecules in the (a) PD and (b) SD. The Δ*H* and NOSC values for all DOM molecules were calculated for both PD and SD samples and are presented in these plots. The shaded area indicates the 95% prediction interval of the fitted linear regression. Detailed calculation methods for Δ*H* and NOSC can be found in the Methods section of the main text. These relationships provide insights into the energetic properties and oxidation states of DOM molecules during the anaerobic digestion process.



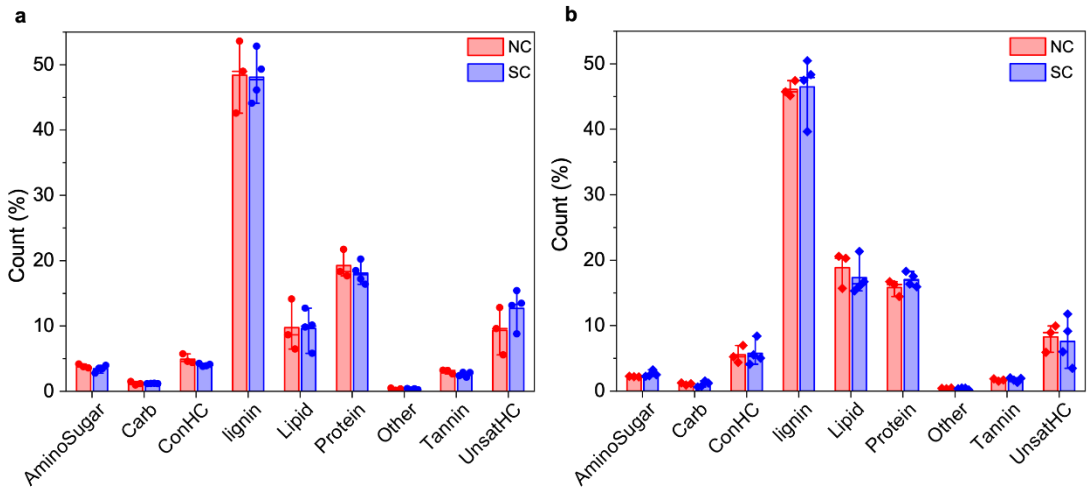
Supplementary Fig. 21 | The compound class composition of molecules unique to PD samples based on (a) molecular counts and (b) relative intensity.



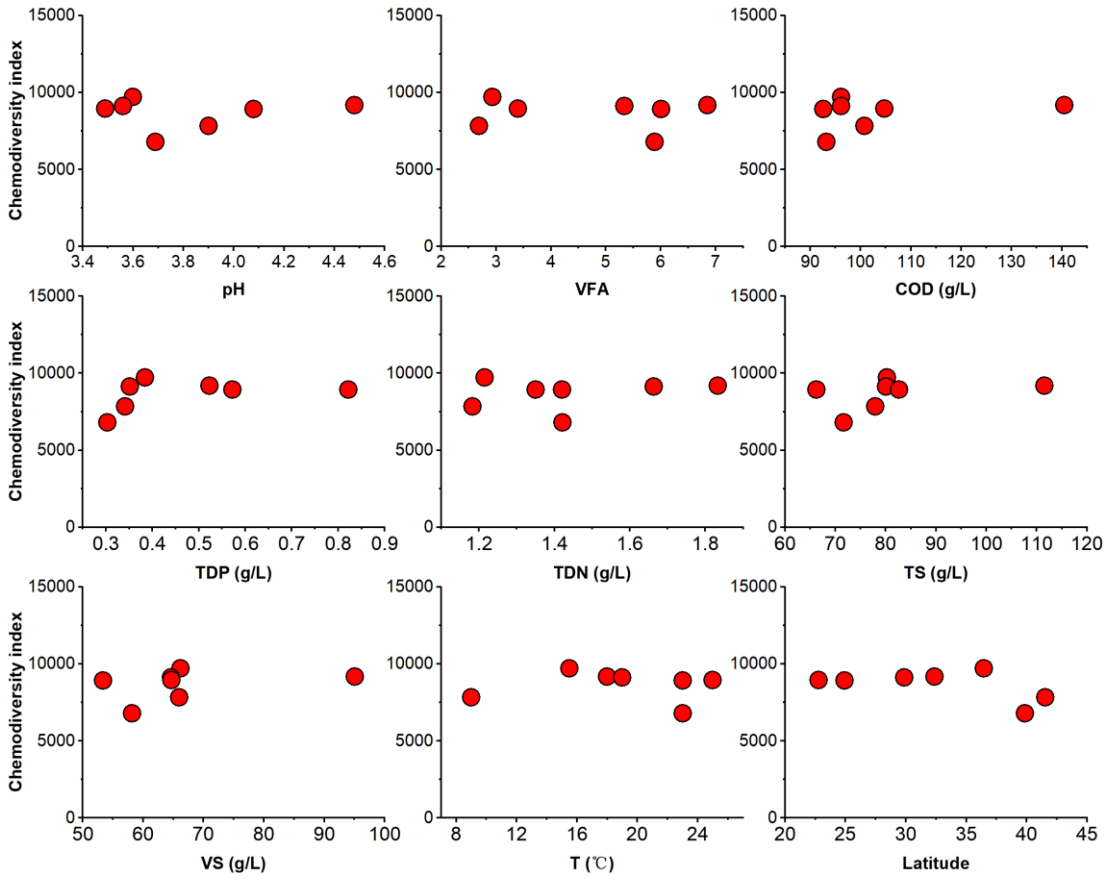
Supplementary Fig. 22 | The compound classes of molecules unique to SD samples based on (a) molecular counts and (b) relative intensity.



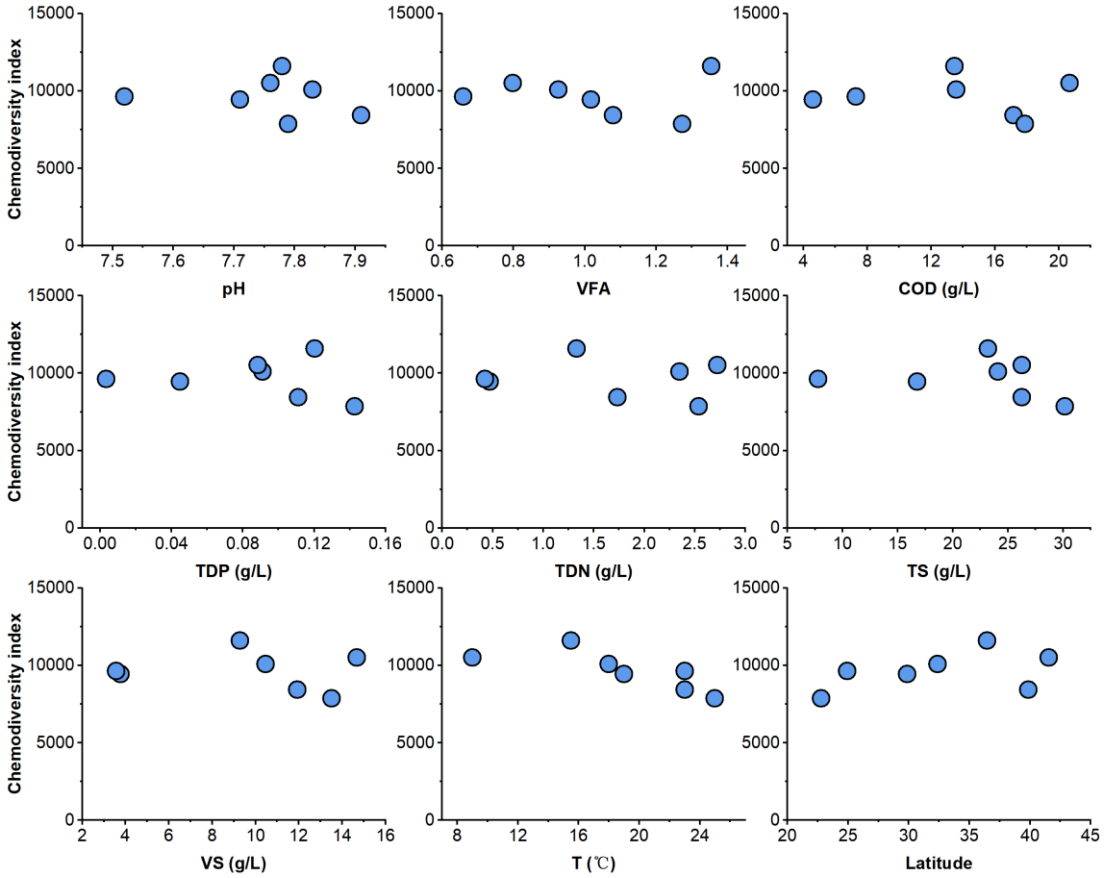
Supplementary Fig. 23 | The compound classes of molecules common to PD and SD samples based on (a) molecular counts and (b) relative intensity.



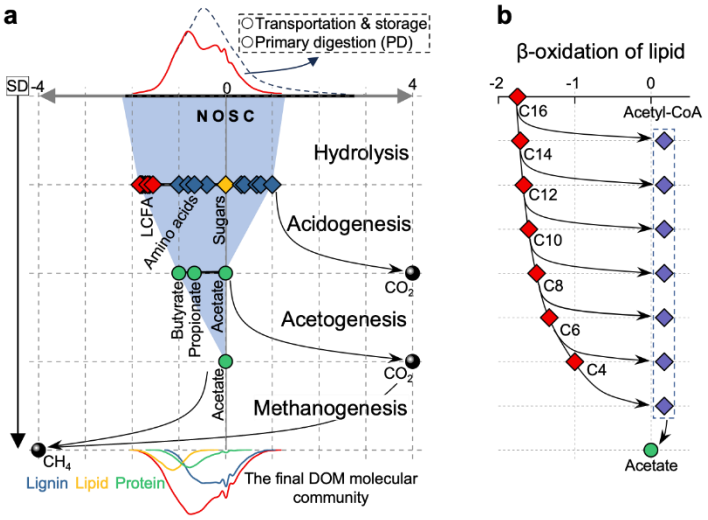
Supplementary Fig. 24 | The relations between molecular composition and regions (North China, NC; South China, SC). The bar + data overlap plot shows the mean (bar), median (center line), whiskers representing 1.5 × the interquartile range (IQR), and outliers as points beyond the whiskers. n = 3 biologically independent samples for NC and n = 4 for SC. Although the culture, habits, and customs are thought to be notably distinct between NC and SC, their molecular compositions vary slightly.



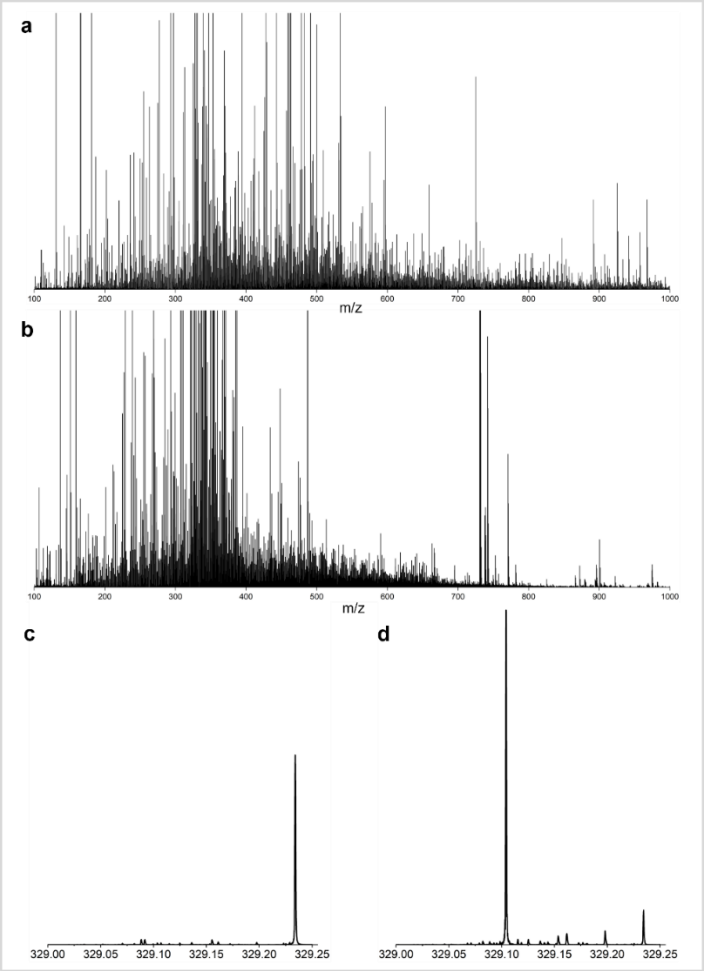
Supplementary Fig. 25 | The relations between molecular composition and physicochemical and environmental factors for PD samples. Environmental factors such as pH, volatile fatty acids (VFAs), chemical oxygen demand (COD), total dissolved phosphorus (TDP), total dissolved nitrogen (TDN), total solids (TS), volatile solids (VS), local temperature, and latitude do not have a significant impact on the DOM molecular diversity for PD samples.



Supplementary Fig. 26 | The relations between molecular composition and physicochemical and environmental factors for SD samples. Environmental factors such as pH, VFA, COD, TDP, TDN, TS, VS, local temperature, and latitude do not significantly impact the molecular diversity of DOM for SD samples.



Supplementary Fig. 27 | Redox shifts and regulation of core DOM molecules in the anaerobic digestion. NOSC evolution of (a) DOM molecules in the hydrolysis, acidogenesis, acetogenesis and methanogenesis processes and (b) the lipid β-oxidation process. The four-stage metabolism of anaerobic digestion is adapted from Zeikus et al.[5] and Batstone et al.[6]. The β-oxidation pathway is based on the Biochemistry textbook by Zhu and Xu[1].



Supplementary Fig. 28 | ESI negative FT-ICR mass spectra of DOM from (a) PD and (b) SD of Shenzhen (SZ) sludge samples. The zoomed areas of (c) PD and (d) SD at nominal mass 329 are shown to underline the differences between samples and resolution power.

Supplementary Table 1 Comparison of DOM molecular diversity from various environments

DOM sources	Number of samples	Formulae assignment	Peaks processing	Number of formulae	Composition	References
River	5	200-900 m/z	S/N[a] ≥ 7; MME[b] < 0.5 ppm	3162	---[c]	7
Sea1	6	150-650 Da; $^{12}C_{1\text{-}60}{}^{13}C_{0\text{-}1}H_{0\text{-}122}O_{0\text{-}40}N_{0\text{-}3}{}^{32}S_{0\text{-}2}{}^{34}S_{0\text{-}1}P_{0\text{-}1}$; O/C≤1, N/C≤1, H≤2C+2+N	S/N ≥ 6	3036	---	8
Sea2	60	150-850 Da; $C_{1\text{-}130}H_{1\text{-}200}O_{1\text{-}50}N_{0\text{-}4}S_{0\text{-}2}P_{0\text{-}2}$	S/N ≥ 4, MME < 0.1 ppm	4000	Highly unsaturated compounds > Highly aromatic compounds > polycyclic aromatics	9
Sea3	5	100-1000 Da; $C_{1\text{-}50}H_{1\text{-}200}O_{1\text{-}50}N_{0\text{-}4}S_{0\text{-}2}P_{0\text{-}1}$	MME < 0.1ppm	5449	Aliphatic > acetate and CRAM > carbohydrate and methoxy > olefinic > aromatic	10
Estuaries	38	$^{12}C_{1\text{-}100}{}^{13}C_{1\text{-}2}H_{1\text{-}200}O_{1\text{-}20}N_{0\text{-}4}S_{0\text{-}2}$	S/N ≥ 6, MME < 0.3 ppm	7398	Highly unsaturated > Polyphenols > unsaturated aliphatic	11
Lakes1	109	150-2000 m/z; $C_{1\text{-}100}H_{1\text{-}250}O_{1\text{-}100}N_{0\text{-}4}S_{0\text{-}1}P_{0\text{-}1}$	S/N ≥ 4	3872	---	12
Lakes2	120	150-2000 m/z; $C_{1\text{-}100}H_{1\text{-}250}O_{1\text{-}100}N_{0\text{-}4}S_{0\text{-}1}P_{0\text{-}1}$ N=4 and S=1 removed, P-/S-/N-containing (N>2) compounds not discussed.	S/N ≥ 4	4032	---	13
Lakes3	45	100-1000 m/z; $C_{1\text{-}100}H_{1\text{-}250}O_{1\text{-}100}N_{0\text{-}4}S_{0\text{-}1}P_{0\text{-}1}$ H≤2C+2+N, O/C≤1, H/C≤1, H/C>0.3; P-containing compounds were excluded	S/N ≥ 5	13952	Highly unsaturated and phenolic compounds > Aliphatic compounds > Vascular plant-derived polyphenols	14
Groundwater1	5	150–2000 Da; $C_{1\text{-}60}H_{0\text{-}\infty}O_{1\text{-}60}N_{0\text{-}4}S_{0\text{-}2}P_{0\text{-}1}$; O/C_{max} = 1, and H/C_{min} = 0.3	S/N ≥ 5, MME < 0.06 ppm	6166	Highly unsaturated>Aliphatic>Polyphenols	15
Groundwater2	35	0–1000 m/z; $C_{1\text{–}100}H_{1\text{–}100}O_{0\text{–}70}N_{0\text{–}4}S_{0\text{–}2}P_{0\text{–}1}$	NME< 0.5 ppm	12728	Lignin-like>unsaturated> unsaturated hydrocarbon-like	16
Paddy soil	88	150-2000 m/z; $C_{1\text{-}100}H_{1\text{-}150}O_{0\text{-}50}N_{0\text{-}4}P_{0\text{-}1}S_{0\text{-}1}$	S/N ≥ 5, MME < 1 ppm	12791	Highly unsaturated/aromatic hydrocarbons, phenolics, polycyclic aromatics	17
River sediment	23	$^{12}C_{1\text{-}\infty}{}^{1}H_{1\text{-}\infty}{}^{16}O_{0\text{-}\infty}{}^{14}N_{0\text{-}3}{}^{32}S_{0\text{-}3}$; DBE=1/2(2C-H+N)>0, H/C>0.3, O/C≤1, N/C≤1, S/C≤0.2	S/N ≥ 4	13931	Highly unsaturated, Aliphatics, aromatic structures	18
Deep-sea sediments	36	200-800 m/z; $^{12}C_{1\text{-}60}{}^{13}C_{0\text{-}2}H_{0\text{-}200}O_{0\text{-}40}N_{0\text{-}3}{}^{32}S_{0\text{-}2}{}^{34}S_{0\text{-}1}P_{0\text{-}1}$	S/N ≥ 5, MME < 0.2 ppm	15672	Lipids, aliphatic/proteins, lignin/CRAM	19
Lake sediments	300	150-1200 m/z; H/C≥1/3, H≤2C+N+2, 0≤O/C≤1, H/C≥0.3, N/C≤1	S/N ≥ 7; MME < 1 ppm	19538	---	20
Soil	33	150-1200 m/z; $C_{\geq 3}H_{\geq 1}O_{\geq 1}N_{0\text{–}2}S_{0\text{–}2}$; O/C≤1.2, H/C≤2.2	S/N ≥ 6, MME < 1 ppm	22746	Aliphatic compounds, proteins/aminosugars, highly unsaturated and phenolic compounds	21
AD systems	14	100-1000 m/z; $C_{1\text{–}\infty}H_{1\text{–}\infty}O_{1\text{–}\infty}N_{0\text{–}3}S_{0\text{–}2}P_{0\text{–}1}$; 0.3≤H/C≤3, O/C≤1.2, N/C≤1.3, S/C≤0.8; P/C≤0.3	S/N ≥ 7, MME < 0.5 ppm	46327	Lignin-like, lipid-like, protein-like	This study

[a] S/N represents “signal to noise ratio”. [b] MME represents “mass measurement error”. [c] --- represents “no mention in the study”

Note that comparison of numbers of formulas (“richness”) may be impeded by differences in sample number, ionization conditions, instrumental settings (including resolution power), formula assignment rules, and post-processing rules (filtering steps).

Supplementary Table 2 Concentrations of volatile free acids (VFAs) in PD-SD two-stage anaerobic digestion (g/L).

Stage	Acetic acid (C2)	Propionic acid (C3)	iso-butyric acid (C4)	n-butyric acid (C4)	iso-valeric acid (C5)	n-valeric acid (C5)	iso-caproic acid (C6)	n-caproic acid (C6)	n-heptanoic acid (C7)	n-caprylic acid (C8)
PD	3.35 ± 1.41	0.71 ± 0.24	0.05 ± 0.06	0.01 ± 0.02	0.05 ± 0.04	0.09 ± 0.06	0.02 ± 0.03	0.09 ± 0.20	0.07 ± 0.17	0.38 ± 0.07
SD	0.25 ± 0.23	0.63 ± 0.18	0	0	0	0	0	0	0	0.17 ± 0.16

Supplementary Table 3 Demethylation index (DMI) for DOM compound classes between PD and SD samples. The calculation method for DMI is detailed in Supplementary Method 2.

PD \ SD	AminoSugar	Carb	ConHC	Lignin	Lipid	Protein	Other	Tannin	UnsatHC
AminoSugar	-33	998	0	111	-46195	-42216	98	988	0
Carb	-1157	-18	0	0	-20439	-14772	109	22	0
ConHC	0	0	461	-19261	-10868	-62	0	-20	-4935
Lignin	-90	0	24134	-109680	-531136	-134759	27	21584	-2778
Lipid	20852	7058	3188	206184	-37686	132502	562	30118	825
Protein	38313	11087	0	89920	-283453	-4502	971	43381	0
Other	-148	-113	0	-55	-2160	-1678	28	-14	0
Tannin	-946	-61	11	-17245	-70153	-61990	18	-154	0
UnsatHC	0	0	11910	6258	-2316	0	0	0	569

Supplementary Table 4 Decarboxylation index (DCI) for DOM compound classes between PD and SD samples. The calculation method for DCI is detailed in Supplementary Method 2.

PD \ SD	AminoSugar	Carb	ConHC	Lignin	Lipid	Protein	Other	Tannin	UnsatHC
AminoSugar	52	-164	0	-4	1376	3156	-12	-323	0
Carb	177	-2	0	0	3	391	-6	-8	0
ConHC	0	0	59	543	0	0	0	0	22
Lignin	7	0	-488	22936	26688	7326	-40	-3139	1419
Lipid	-422	-21	0	-6076	1055	-6468	0	-1541	-2
Protein	-1613	-315	0	-3513	12886	815	-15	-3141	0
Other	13	-4	0	121	0	-3	3	0	0
Tannin	450	53	-3	3696	4056	5026	-12	72	0
UnsatHC	0	0	-98	-852	12	0	0	0	20

Supplementary Table 5 Dehydration index (DHI) for DOM compound classes between PD and SD samples. The calculation method for DHI is detailed in Supplementary Method 2.

PD \ SD	AminoSugar	Carb	ConHC	Lignin	Lipid	Protein	Other	Tannin	UnsatHC
AminoSugar	71	-212	1953	16382	9	1406	-13	0	181
Carb	218	0	805	6294	0	164	-4	93	17
ConHC	-360	-365	-78	-5737	0	-365	-8	-708	-47
Lignin	-7917	-7461	13519	19378	-1306	-20909	-442	-3979	5290
Lipid	-24	0	0	1946	2481	-5718	0	0	1113
Protein	-1209	-114	972	41933	12776	2297	-50	0	3430
Other	17	3	7	409	0	54	0	14	16
Tannin	0	-284	1743	4279	0	0	-56	49	0
UnsatHC	-454	-34	51	-3776	-1545	-5079	0	0	195

Supplementary Table 6 Dehydrogenation index (DHGI) for DOM compound classes between PD and SD samples. The calculation method for DHGI is detailed in Supplementary Method 2.

PD \ SD	AminoSugar	Carb	ConHC	Lignin	Lipid	Protein	Other	Tannin	UnsatHC
AminoSugar	-62	0	795	11578	0	0	-5	1908	0
Carb	0	82	59	0	0	0	-10	1168	0
ConHC	-669	-42	150	-6650	-6940	-5727	-30	-88	-1675
Lignin	-8974	0	14798	15935	-43107	-130014	-14	0	0
Lipid	0	0	9281	56861	823	0	0	0	2619
Protein	0	0	18065	168921	0	6574	-9	0	0
Other	17	10	15	59	4	0	-6	12	0
Tannin	-1083	-795	316	0	0	0	-21	182	0
UnsatHC	0	0	3757	0	-3268	0	0	0	1470

Supplementary Table 7 Deamination index (DAI) of DOM compound classes between PD and SD samples. The calculation method for DAI is detailed in Supplementary Method 2.

PD \ SD	AminoSugar	Carb	ConHC	Lignin	Lipid	Protein	Other	Tannin	UnsatHC
AminoSugar	-12	24	0	63	0	-220	0	303	0
Carb	-16	1	0	0	0	0	-1	14	0
ConHC	0	0	-25	-814	-9	-17	0	0	-25
Lignin	-68	0	1253	-1619	-2661	-6055	0	368	-98
Lipid	2	0	3	1852	-382	642	0	3	2
Protein	288	0	2	6982	-653	1087	0	586	0
Other	0	8	0	0	0	2	0	0	0
Tannin	-337	-13	0	-413	-3	-394	0	-12	0
UnsatHC	0	0	35	160	-64	0	0	0	-11

Supplementary Table 8 Standard molar combustion enthalpy (Δ*H*) and NOSC of typical compound classes in the food waste sludges.

Name	Formula	NOSC	Δ*H* (kJ·mol^{-1} C)	Compound class
Glucosamine	$C_6H_{13}NO_5$	0	-500.13	AminoSugar
GlcNAc	$C_8H_{15}NO_6$	0	-486.2375	AminoSugar
Galactosamine	$C_6H_{13}NO_5$	0	-500.13	AminoSugar
GalNAc	$C_8H_{15}NO_6$	0	-486.2375	AminoSugar
Muramic acid (Mur)	$C_9H_{17}NO_7$	0	-481.6066667	AminoSugar
MurNAc	$C_{11}H_{19}NO_8$	0	-474.8709091	AminoSugar
Acetate	$C_2H_4O_2$	0	-444.56	Carb
Glucose	$C_6H_{12}O_6$	0	-444.56	Carb
Fructose	$C_6H_{12}O_6$	0	-444.56	Carb
Xylose	$C_5H_{10}O_5$	0	-444.56	Carb
sucrose	$C_{12}H_{22}O_{11}$	0	-444.56	Carb
maltose	$C_{12}H_{22}O_{11}$	0	-444.56	Carb
lactose	$C_{12}H_{22}O_{11}$	0	-444.56	Carb
Cellobiose	$C_{12}H_{22}O_{11}$	0	-444.56	Carb
Methane	CH_4	-4	-889.12	CH_4
Carbon dioxide	CO_2	4	0	CO_2
Fulvic acid	$C_{14}H_{12}O_8$	0.28571429	-412.8057143	Lignin
Phenylalanine (Phe)	$C_9H_{11}NO_2$	-0.4444444	-531.0022222	Lignin
Tryptophan (Trp)	$C_{11}H_{12}N_2O_2$	-0.1818182	-525.3890909	Lignin
Tyrosine (Tyr)	$C_9H_{11}NO_3$	-0.2222222	-506.3044444	Lignin
Deoxyadenosine-MP (dAMP^{2-})	$C_{10}H_{12}N_5O_6P^{2-}$	0.8	-500.13	Lignin
Riboflavin (Vitamin B2)	$C_{17}H_{20}N_4O_6$	0.23529412	-496.8611765	Lignin
Nicotinamide (Vitamin B3)	$C_6H_6N_2O$	0.33333333	-518.6533333	Lignin
Pyridoxine (Vitamin B6)	$C_8H_{11}NO_3$	-0.25	-514.0225	Lignin
Folate (Vitamin B9)	$C_{19}H_{19}N_7O_6$	0.73684211	-485.5063158	Lignin
Indole-3-acetic acid	$C_{10}H_9NO_2$	-0.2	-500.13	Lignin
Gibberellic acid	$C_{19}H_{22}O_6$	-0.5263158	-503.0547368	Lignin
Abscisic acid	$C_{15}H_{20}O_4$	-0.8	-533.472	Lignin
Lauric acid (12:0)	$C_{12}H_{24}O_2$	-1.6666667	-629.7933333	Lipid
Myristic acid (14:0)	$C_{14}H_{28}O_2$	-1.7142857	-635.0857143	Lipid
Palmitic acid (16:0)	$C_{16}H_{32}O_2$	-1.625	-625.1625	Lipid
Stearic acid (18:0)	$C_{18}H_{36}O_2$	-1.7777778	-642.1422222	Lipid
Arachidic acid (20:0)	$C_{20}H_{40}O_2$	-1.8	-644.612	Lipid
Behenic acid (22:0)	$C_{22}H_{44}O_2$	-1.8181818	-646.6327273	Lipid
Lignoceric acid (24:0)	$C_{24}H_{48}O_2$	-1.8333333	-648.3166667	Lipid
Palmitoleic acid (16:1$\triangle^{9}$)	$C_{16}H_{30}O_2$	-1.625	-625.1625	Lipid
Oleic acid (18:1$\triangle^{9}$)	$C_{18}H_{34}O_2$	-1.6666667	-629.7933333	Lipid
Linoleic acid (18:2$\triangle^{9, 12}$)	$C_{18}H_{32}O_2$	-1.5555556	-617.4444444	Lipid
α-linolenic acid (18:3$\triangle^{9, 12, 15}$)	$C_{18}H_{30}O_2$	-1.4444444	-605.0955556	Lipid

Supplementary Table 8 (continued)

Name	Formula	NOSC	Energy (KJ·mol^{-1}C)	Compound class
Arachidonic acid (20:4△$^{5, 8, 11, 14}$)	$C_{20}H_{32}O_2$	-1.4	-600.156	Lipid
Biotin (Vitamin B7)	$C_{10}H_{16}N_2O_3S$	-0.2	-622.384	Lipid
Thioctic acid	$C_8H_{14}O_2S_2$	-0.75	-750.195	Lipid
Formate	CH_2O_2	2	-222.28	Other
Glycerol	$C_3H_8O_3$	-0.6666667	-518.6533333	Other
Histidine (His)	$C_6H_9N_3O_2$	0.66666667	-537.1766667	Protein
Isoleucine (He)	$C_6H_{13}NO_2$	-1	-611.27	Protein
Leucine (Leu)	$C_6H_{13}NO_2$	-1	-611.27	Protein
Methionine (Met)	$C_5H_{11}NO_2S$	-0.4	-733.524	protein
Proline (Pro)	$C_5H_9NO_2$	-0.4	-555.7	Protein
Valine (Val)	$C_5H_{11}NO_2$	-0.8	-600.156	Protein
Carnosine	$C_9H_{14}N_4O_3$	0.44444444	-543.3511111	Protein
Alanyl-glutamine (Ala-Glu)	$C_8H_{15}N_3O_4$	0.25	-541.8075	Protein
L-alanyl-L-alanyl-L-alanine (Ala-Ala-Ala)	$C_9H_{17}N_3O_4$	0	-555.7	Protein
L-α-aspartylglycyl-L-proline (Asp-Gly-Pro)	$C_{11}H_{17}N_3O_6$	0.36363636	-495.0781818	Protein
Humic acid	$C_9H_9NO_6$	0.66666667	-407.5133333	Tannin
Adenosine-MP (AMP^{2-})	$C_{10}H_{12}N_5O_7P^{2-}$	1	-477.902	Tannin
Guanosine-MP (GMP^{2-})	$C_{10}H_{12}N_5O_8P^{2-}$	1.2	-455.674	Tannin
Gytidine-MP (CMP^{2-})	$C_9H_{12}N_3O_8P^{2-}$	0.66666667	-456.9088889	Tannin
Uridine-MP (UMP^{2-})	$C_9H_{11}N_2O_9P^{2-}$	0.66666667	-419.8622222	Tannin
Thiamin (Vitamin B1)	$C_{12}H_{17}N_4OS^+$	-0.5833333	-629.7933333	UnsatHC
Kinetin	$C_{10}H_9N_5O$	0.8	-522.358	UnsatHC
Eicosapentaenoic acid (20:5△$^{5, 8, 11, 14,17}$)	$C_{20}H_{30}O_2$	-1.3	-589.042	UnsatHC
Docosahexaenoic acid (22:6△$^{4, 7, 11, 13,16,19}$)	$C_{22}H_{32}O_2$	-1.2727273	-586.0109091	UnsatHC
Limonene	$C_{10}H_{16}$	-1.6	-622.384	UnsatHC
β-carotene	$C_{40}H_{56}$	-1.4	-600.156	UnsatHC
Naphthalene	$C_{10}H_8$	-0.8	-533.472	ConHC
Phenanthrene	$C_{14}H_{10}$	-0.7142857	-523.9457143	ConHC
Pyrene	$C_{16}H_{10}$	-0.625	-514.0225	ConHC

Note: All molecules used in these calculations are sourced from the textbook *Food Biochemistry* by Miao Wang[22] and *Biochemistry* by Zhu and Xu[1].

Supplementary Table 9 Key parameters of anaerobic digesters across China in this study.

Sample Name	Per Capita GDP (10 thousand yuan)	Region	Longitude, Latitude	Average annual temperature (℃, 2021)	Solid loading (TS, g/L)	pH (PD-SD)
Beijing (BJ)	18.4	North China	116.048546, 39.858046	13.5	119.6022	4.12-7.91
Guangyuan (GY)	4.8	South China	105.765037, 32.379197	16.5	96.1830	3.60-7.78
Hangzhou (HZ)	11.2	South China	119.721928, 29.87747	18.5	120.255	4.48-7.83
Liaocheng (LC)	4.4	North China	115.892774, 36.474778	14.5	99.0728	3.98-7.71
Quanzhou (QZ)	7.2	South China	118.402561, 24.938461	23.0	87.7866	3.92-7.52
Shenyang (SY)	8.8	North China	123.582964, 41.563614	9.0	88.5254	4.34-7.76
Shenzhen (SZ)	15.8	South China	114.272055, 22.768711	24.5	88.1746	4.29-7.79

Supplementary Table 10 Standard methods for physicochemical analysis of anaerobic sludges in our study.

Property	Preservation	Standard Method
pH	4℃, <7 days	pH meter (Mettler Toledo)
TS (total solid)	4℃, <14 days	APHA 2054 B (Dried at 105℃)
VS (volatile solid)	4℃, <14 days	APHA2054 E (Ignited at 550℃)
COD	4℃, <14 days	HACH Method 8000
TN	4℃, <7 days	HACH Method 10208
TP	4℃, <7 days	HACH Method
VFAs (volatile fatty acids)	-20℃, <30 days	Chromatography method

Note: Detailed information on the chromatography method can be found in the Methods section of the main text.

Supplementary Table 11 List of (internal) calibrants used for FT-ICR MS

Neutral Molecular Formula	m/z (H-)	Neutral Molecular Formula	m/z (H-)
$C_2HF_3O_2$	112.985587	$C_{16}H_{24}O_8$	343.139841
$C_7H_{14}O_2$	129.092103	$C_{19}H_{22}O_6$	345.134363
$C_9H_{18}O_2$	157.123403	$C_{20}H_{34}O_3S$	353.215589
$C_{11}H_{22}O_2$	185.154704	$C_{23}H_{46}O_2$	353.342504
$C_{10}H_8O_5$	207.029898	$C_{20}H_{24}O_6$	359.150013
$C_{10}H_{14}O_3S$	213.059089	$C_{21}H_{36}O_3S$	367.231239
$C_{13}H_{26}O_2$	213.186004	$C_{24}H_{48}O_2$	367.358154
$C_{12}H_{14}O_4$	221.081933	$C_{21}H_{26}O_6$	373.165663
$C_{11}H_{16}O_3S$	227.074739	$C_{24}H_{18}O_3N_2$	381.124466
$C_{14}H_{28}O_2$	227.201654	$C_{22}H_{38}O_3S$	381.246889
$C_{10}H_{16}O_6$	231.087412	$C_{22}H_{28}O_6$	387.181313
$C_{13}H_{16}O_4$	235.097583	$C_{19}H_{24}O_9$	395.134756
$C_{12}H_{18}O_3S$	241.090389	$C_{23}H_{40}O_3S$	395.262539
$C_{15}H_{30}O_2$	241.217304	$C_{26}H_{52}O_2$	395.389454
$C_{14}H_{18}O_4$	249.113233	$C_{23}H_{30}O_6$	401.196963
$C_{13}H_{20}O_3S$	255.106039	$C_{16}H_{12}O_{13}$	411.020518
$C_{16}H_{32}O_2$	255.232954	$C_{17}H_{14}O_{13}$	425.036168
$C_{10}H_{21}O_5N_3$	262.140844	$C_{21}H_{26}O_{10}$	437.145321
$C_{15}H_{22}O_4$	265.144533	$C_{18}H_{16}O_{13}$	439.051818
$C_{14}H_{22}O_3S$	269.121689	$C_{30}H_{60}O_2$	451.452055
$C_{17}H_{34}O_2$	269.248604	$C_{19}H_{18}O_{13}$	453.067468
$C_{13}H_{20}O_6$	271.118712	$C_{25}H_{32}O_8$	459.202442
$C_{16}H_{24}O_4$	279.160183	$C_{20}H_{20}O_{13}$	467.083118
$C_{15}H_{24}O_3S$	283.137339	$C_{21}H_{22}O_{13}$	481.098768
$C_{18}H_{36}O_2$	283.264254	$C_{22}H_{24}O_{13}$	495.114418
$C_{17}H_{26}O_4$	293.175833	$C_{21}H_{24}O_{14}$	499.109329
$C_{16}H_{26}O_3S$	297.152989	$C_{23}H_{26}O_{13}$	509.130068
$C_{19}H_{38}O_2$	297.279904	$C_{24}H_{28}O_{13}$	523.145718
$C_{18}H_{28}O_4$	307.191483	$C_{25}H_{30}O_{13}$	537.161368
$C_{13}H_{10}O_9$	309.025206	$C_{27}H_{26}O_{12}$	541.135150
$C_{17}H_{28}O_3S$	311.168639	$C_{26}H_{32}O_{13}$	551.177018
$C_{20}H_{40}O_2$	311.295554	$C_{27}H_{34}O_{13}$	565.192668
$C_{19}H_{30}O_4$	321.207133	$C_{28}H_{36}O_{13}$	579.208318
$C_{18}H_{30}O_3S$	325.184289	$C_{29}H_{38}O_{13}$	593.223968
$C_{21}H_{42}O_2$	325.311204	$C_{28}H_{24}O_{15}$	599.104244
$C_{20}H_{30}O_4$	333.207133	$C_{26}H_{52}O_{15}$	603.323345
$C_{19}H_{32}O_3S$	339.199939	$C_{30}H_{40}O_{13}$	607.239618
$C_{22}H_{44}O_2$	339.326854	$C_{31}H_{42}O_{13}$	621.255268

Supplementary Table 12 Formula assignment information for all the DOM samples in our study

	Golden Rules		User Defined	BJ_PD	BJ_SD	GY_PD	GY_SD	HZ_PD	HZ_SD	LC_PD	LC_SD	QZ_PD	QZ_SD	SY_PD	SY_SD	SZ_PD	SZ_SD
Mass Range (m/z)	For mass < 1000 m/z		100-1000	117-999	114-999	101-1000	149-994	115-999	105-999	101-1000	112-1000	101-1000	115-992	111-1000	103-998	115-1000	109-999
Element Limits	C	C_{max}=78	Not restricted	4-75	4-73	3-75	4-71	4-74	4-75	3-76	3-75	3-74	4-63	3-75	4-75	4-75	6-71
	H	H_{max}=126	Not restricted	4-118	4-109	4-122	5-121	4-122	4-124	3-125	2-119	4-125	4-125	4-123	4-119	3-125	6-119
	O	O_{max}=27	O_{min}=1	1-24	1-25	4-26	1-26	1-25	1-26	1-26	1-26	1-26	1-23	1-26	1-26	1-26	1-26
	N	N_{max}=25	N_{max}=3	0-3	0-3	0-3	0-3	0-3	0-3	0-3	0-3	0-3	0-3	0-3	0-3	0-3	0-3
	S	S_{max}=14	S_{max}=2	0-2	0-2	0-2	0-2	0-2	0-2	0-2	0-2	0-2	0-2	0-2	0-2	0-2	0-2
	P	P_{max}=9	P_{max}=1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1
	H_{max} = C*2 + 2 + N + P			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Element ratios	H/C	0.2-3.1	0.3-3	0.3-2.8	0.3-2.5	0.3-2.8	0.3-2.9	0.3-2.5	0.3-2.8	0.3-2.9	0.3-2.8	0.3-2.9	0.3-2.5	0.3-2.9	0.3-2.9	0.3-2.9	0.3-2.5
	O/C	0-1.2		0.0-1.2	0.0-1.2	0.0-1.2	0.0-1.2	0.0-1.2	0.0-1.2	0.0-1.2	0.0-1.2	0.0-1.2	0.0-1.2	0.0-1.2	0.0-1.2	0.0-1.2	0.0-1.2
	N/C	0-1.3		0.0-0.4	0.0-0.5	0.0-0.8	0.0-0.5	0.0-0.4	0.0-0.6	0.0-0.8	0.0-0.7	0.0-0.8	0.0-0.4	0.0-0.8	0.0-0.6	0.0-0.5	0.0-0.4
	S/C	0-0.8		0.0-0.5	0.0-0.4	0.0-0.7	0.0-0.4	0.0-0.5	0.0-0.5	0.0-0.7	0.0-0.7	0.0-0.7	0.0-0.4	0.0-0.7	0.0-0.4	0.0-0.5	0.0-0.3
	P/C	0-0.3		0.0-0.2	0.0-0.2	0.0-0.2	0.0-0.2	0.0-0.2	0.0-0.3	0.0-0.3	0.0-0.3	0.0-0.3	0.0-0.3	0.0-0.3	0.0-0.2	0.0-0.2	0.0-0.2
DBE	integer & >=0			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
S/N			>7	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Error			0.5 ppm	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Supplementary Table 13 H/C and O/C ranges for classifying DOM molecules.

Compound classes	O:C and H:C ranges
Lipid	0 < O:C ≤ 0.3, 1.5 ≤ H:C ≤ 2.5
Unsaturated hydrocarbon (UnsatHC)	0 ≤ O:C ≤ 0.125, 0.8 ≤ H:C < 1.5
Protein	0.3 < O:C ≤ 0.55, 1.5 ≤ H:C ≤ 2.3
Amino sugar (AminoSugar)	0.55 < O:C ≤ 0.7, 1.5 ≤ H:C ≤ 2.5
Carbohydrate (Car)	0.7 < O:C ≤ 1.5, 1.5 ≤ H:C ≤ 2.5
Lignin	0.125 < O:C ≤ 0.65, 0.8 ≤ H:C ≤ 1.5
Tannin	0.65 < O:C ≤ 1.1, 0.8 ≤ H:C ≤ 1.5
Condensed hydrocarbon (ConHC)	0 ≤ O:C ≤ 0.95, 0.2 ≤ H:C < 0.8
Other	Not in the above defined ranges

Note: The compound classification based on O/C and H/C builds on the key work by Bailey et al.[23]. In this study, we employed the Formularity software[24] for formula assignment. And method details are provided in the Methods section of main text.

Supplementary References

1. Zhu and Xu, *Biochemistry*, 4th Edition, 2017, Higher Education Press, Beijing.
2. Zhu YX, Huang J, Zhang ZK, Zhang QM, Zhou T, Ahn YY. Geography and similarity of regional cuisines in China. *PloS one* **18**, e79161 (2013).
3. Liang CB, Cui B, Wang FR, Peng J, Ma JY, Xu MY, Ke J, Tian Y, Cui ZQ. Promoting healthy cooking patterns in China: Analysis of consumer clusters and the evolution of cooking pattern trends. *PloS one* **18**, e0293919 (2023).
4. Huang L, Wang L, Jiang H, Wang H, Wang Z, Zhang B, Ding G. Trends in dietary carbohydrates, protein and fat intake and diet quality among Chinese adults, 1991–2015: results from the China Health and Nutrition Survey. *Public Health Nutrition* **26**, 834-43 (2023).
5. Zeikus, J. G., R. Kerby, and Joseph Adrian Krzycki. Single-carbon chemistry of acetogenic and methanogenic bacteria. *Science* 227, 1167-1173 (1985).
6. Batstone, Damien J., et al. The IWA anaerobic digestion model no 1 (ADM1). *Water Science and technology* **45**, 65-73 (2002).
7. Danczak, R. E. *et al.* Using metacommunity ecology to understand environmental metabolomes. *Nat. Commun.* **11**, 6369 (2020).
8. Lechtenfeld, O. J., Hertkorn, N., Shen, Y., Witt, M. & Benner, R. Marine sequestration of carbon in bacterial metabolites. *Nat. Commun.* **6**, 6711 (2015).
9. Medeiros, Patricia. M. *et al.* Dissolved organic matter composition and photochemical transformations in the northern North Pacific Ocean. *Geophys. Res. Lett.* **42**, 863–870 (2015).
10. Seidel, M., Vemulapalli, S. P. B., Mathieu, D. & Dittmar, T. Marine Dissolved Organic Matter Shares Thousands of Molecular Formulae Yet Differs Structurally across Major Water Masses. *Environ. Sci. Technol.* **56**, 3758–3769 (2022).
11. Sun, X. *et al.* Linkages Between Optical and Molecular Signatures of Dissolved Organic Matter Along the Yangtze River Estuary-to-East China Sea Continuum. *Front. Mar. Sci.* **9**, 933561 (2022).

12. Kellerman, A. M., Kothawala, D. N., Dittmar, T. & Tranvik, L. J. Persistence of dissolved organic matter in lakes related to its molecular characteristics. *Nat. Geosci.* **8**, 454–457 (2015).
13. Kellerman, A. M., Dittmar, T., Kothawala, D. N. & Tranvik, L. J. Chemodiversity of dissolved organic matter in lakes driven by climate and hydrology. *Nat. Commun.* **5**, 3804 (2014).
14. Luo, J. *et al.* Lake Chemodiversity Driven by Natural and Anthropogenic Factors. *Environ. Sci. Technol.* **56**, 5910–5919 (2022).
15. Zark, M. & Dittmar, T. Universal molecular structures in natural dissolved organic matter. *Nat. Commun.* **9**, 3178 (2018).
16. Osterholz, H. *et al.* Terrigenous dissolved organic matter persists in the energy-limited deep groundwaters of the Fennoscandian Shield. *Nat. Commun.* **13**, 4837 (2022).
17. Li, H.-Y. *et al.* The chemodiversity of paddy soil dissolved organic matter correlates with microbial community at continental scales. *Microbiome* **6**, 187 (2018).
18. He, W., Chen, M., Park, J.-E. & Hur, J. Molecular diversity of riverine alkaline-extractable sediment organic matter and its linkages with spectral indicators and molecular size distributions. *Water Res.* **100**, 222–231 (2016).
19. Chen, X. *et al.* Niche differentiation of microbial community shapes vertical distribution of recalcitrant dissolved organic matter in deep-sea sediments. *Environ. Int.* **178**, 108080 (2023).
20. Hu, A. *et al.* Ecological networks of dissolved organic matter and microorganisms under global change. *Nat. Commun.* **13**, 3600 (2022).
21. Ding, Y. *et al.* Chemodiversity of soil dissolved organic matter. *Environ. Sci. Technol.* **54**, 6174–6184 (2020).
22. Miao Wang, *Food Biochemistry*, 2nd Edition, 2020, China Light Industry Press, Beijing.
23. Bailey, V. L., Smith, A. P., Tfaily, M., Fansler, S. J. & Bond-Lamberty, B. Differences in soluble organic carbon chemistry in pore waters sampled from different pore size domains. *Soil Biology Biochem.* **107**, 133–143 (2017).

24. Tolić, N. et al. Formularity: Software for Automated Formula Assignment of Natural and Other Organic Matter from Ultrahigh-Resolution Mass Spectra. *Anal Chem* 89, 12659–12665 (2017).