Aromaticity index with improved estimation of carboxyl group contribution for biogeochemical studies

Alexander Zherebker,*[a] Gleb D. Rukhovich[a], Anastasia Sarycheva[a], Oliver J. Lechtenfeld*[b], Evgeny N. Nikolaev[a]

[a] Skolkovo Institute of Science and Technology, 121205, Moscow, Russia. E-mail: a.zherebker@skoltech.ru

[b] Department of Analytical Chemistry, Research Group BioGeoOmics, Helmholtz Centre for Environmental Research - UFZ, Leipzig DE-04318, Germany. E-mail: oliver.lechtenfeld@ufz.de

Abstract

Natural organic matter (NOM) components measured with ultra-high resolution mass spectrometry are often assessed by molecular formula-based indices, particularly related to their aromaticity, which are further used as proxies to explain biogeochemical reactivity. An aromaticity index was proposed to account for hereto-atom contribution to double bonds, but relies on assumptions particularly with respect to carboxylic acids, abundant functional groups in NOM components. Here, we propose a new constrained aromaticity index (AI_{con}) based on the measured distribution of carboxylic groups among individual components of NOM which was obtained by deuteromethylation and Fourier-transform ion cyclotron resonance mass-spectrometry (FTICR MS). Labelling of carboxyl groups in NOM compounds from diverse sources (coal, marine, peat, permafrost, blackwater river, and soil) revealed that the most probable number of carboxylic groups was two, which enabled to set a reference point n=2 for carboxyl-accounted AI_{con} calculation. The proposed index was evaluated against the measured number of carboxyl groups showing that it provided the smallest errors for aromaticity calculation for all NOM samples under study as well as for individual natural compounds obtained from the Coconut

database, which were significantly less oxidized as compared to NOM. Unlike proposed AI_{con}, conventional AI and AI_{mod} resulted in significant underestimation of compound aromaticity for both NOM and individual compounds from the database. In particular, AI_{con} performed better than AI_{mod} for all compound classes in which aromatic moieties are expected: aromatics, condensed aromatics and unsaturated compounds. Therefore, AI_{con} referenced with two carboxyl groups is preferred over conventional AI and AI_{mod} for biogeochemical studies where the aromaticity of compounds is important to understand the transformations and fate of NOM compounds.

Keywords NOM, humic substances, FTICR MS, carboxylic groups, aromaticity index, deuteromethylation, isotopic labeling

Synopsis Enumeration of carboxylic groups enabled to refine aromaticity index, which is used in majority of environmental studies engaging mass spectrometry.

INTRODUCTION

Natural organic matter (NOM) and humic substances (HS) are ubiquitous in various environments, for example, freshwater and marine systems, permafrost and agricultural soils with extensive biodegradation processes, as well as in caustobiolithes. This makes them an important part of the global carbon cycle and many studies are devoted to the mechanisms of NOM formation, transformation and storage.[1,2] The rates and the extent of transformation of these forms of organic matter (OM) depend on environmental conditions (e.g. the availability of oxygen) and the source biomass composition.[3,4] However, diagenesis ultimately leads to the accumulation of aromatic and\or oxygen-rich structures, which are resistant towards further biodegradation and the extent of degradation can be evaluated by the aromaticity of NOM samples. This concept was developed from a vast number of studies utilizing the optical properties of NOM and HS (mainly absorbance and fluorescence). For example, UV/Vis absorbance is commonly used to characterize both the chemical characteristics (molecular weight or aromaticity) and the dynamics of OM.[5] Several indices were proposed to characterize aromaticity and the degree of OM transformation: ratio of absorbance values at 465 and 665 nm (E_4/E_6 ratio), long-wavelength slope of absorption spectrum, fluorescence index[6] and fluorescence spectrum asymmetry.[7] However, optical indices only reflect the mean or bulk character of OM based on empirical relationships. To reveal molecular-level transformation during biogeochemical processing and diagenesis, methods are required which directly assess the aromatic and aliphatic building blocks of NOM and HS.

In this regard, ^{1}H, ^{13}C and 2D NMR spectroscopy is highly suited to provide quantitative and qualitative information about aromatic moieties in molecules. Depending on the experiment, it is possible to determine the contribution of aromatic carbon atoms, aromatic protons and specific functional groups and their chemical environment.[8] However, in case of complex mixtures like NOM and HS this information cannot be broken down to the individual molecular level due to broad peaks

from significant overlap of chemical shifts.[9] Recently Bell with co-workers revealed possible aromatic moieties from lignin in peat humic substances via combining ^{13}C isotopic labeling and specific spectrum-acquisition conditions *(pseudo* 4D-NMR).[10]

Mass spectrometry (MS), in contrast to NMR, provides information on a single molecular composition by determination its exact molecular mass of ions and their structural fragments. For example, reduction of polar components from Suwannee River Fulvic Acid (SRFA, an OM reference standard) followed by 2D gas-chromatography (GC) coupled to time-of-flight mass spectrometry enabled to determine isomeric alicyclic hydrocarbons and terpenoids.[11] However, GC-MS analysis is mostly limited to aliphatic compound and not easily applicable for aromatic and polar species. Hence, NOM and HS studies mostly utilize direct infusion (DI) of dissolved samples using soft ionization methods like electrospray ionization (ESI) coupled to ultra-high resolution mass spectrometry UHRMS (e.g. Fourier-transform ion cyclotron resonance (FTICR) and Orbitrap mass spectrometry), which routinely resolve thousands of exact molecular compositions in a single sample.[9,12–14] Despite sufficient mass resolution provided of UHRMS, tandem mass-spectrometry analysis of NOM and HS mixtures is challenging and not routinely applied for such mixtures[15,16]. At the same time each molecular composition detected by UHRMS may correspond to a large number of structural isomers[17] and only limited formula-based structural information may be derived from these experiments.

One example is the formula-based estimation of component aromaticity, which is related to the unsaturation state of molecules. Aromaticity estimation is widely used and often connected with optical properties[18] (e.g. correlation with optical density attributed to aromatic chromophores), chemical properties[19] (e.g. photo-lability) and in biological experiments in which microorganisms transform components of NOM.[20] For highly aromatic samples (e.g. coal, petroleum, etc.), aromaticity is correlated to double bond equivalent (DBE) and the DBE/C ratio where e.g. condensed aromatic structures require DBE/C ≥ 0.7.[21] For organic aerosol the aromaticity equivalent (*Xc*) was proposed to attribute compounds to aromatic and condensed aromatic species by subtracting from

DBE the number of mathematically possible CH_2 fragments.[22] For heteroatom-rich NOM, aromaticity index (AI) and modified aromaticity index (AI_{mod}) were introduced as proxies to account for DBE and hence aromaticity attributed to carbon skeleton.[23,24] AI and AI_{mod} assume that all oxygen atoms and half of oxygen atoms form double bonds with sp^2-hybridized carbon, respectively. Sulfur and nitrogen, in contrast, are always considered to form π-bonds with carbon, both in AI and AI_{mod}, introducing additional uncertainty. The advantage of AI indices over *Xc* is the normalized range from 0 to 1. AI is considered as the most conservative aromatic system approximation for NOM/HS and despite some uncertainties of these indices, AI and AI_{mod} currently serve as major structural parameters in many geochemical studies of NOM/HS[4,18,25].

In our opinion, due to the fact that estimation of molecule aromaticity is often connected to presumed biological and chemical properties of NOM in biogeochemical studies, this approach resulted in appearance of important issues in the field. For example, AI thresholds are used as direct marker of the presence or absence of aromatic units in molecules and not as measure of "mean" aromaticity as it has been proposed in the original paper by Koch and Dittmar[23,24]. For example, phenols with saturated substituents might be excluded from the list of aromatic components assigned by AI or AI_{mod} values while from the chemical point of view they are aromatic and characterized by all chemical properties of aromatic compounds. In addition, we have shown that molecules with the same low AI_{mod} value (e.g. 0.3) may correspond to isomers with and without aromatic rings depending on the geochemical origin.[26] Hence, conclusions about aromatic character of molecules and its connection to reactivity in environment can easily be false. One of the reason for the false-assignment of structures using AI or AI_{mod} is the underestimation of phenolic and methoxy-groups, which are abundant in terrestrial NOM samples and supposedly rare in marine DOM and microbial-derived samples.[27] Another problem is that originally NMR-based distribution of carbon-oxygen bonds in marine DOM was directly used to estimate number of COOH groups for FTICR MS-based AI_{mod}.[23] However, such approximation ignores issues of using NMR versus DI ESI MS, since the latter ionizes

only a part of the sample with possible distortion of relative intensities: e.g. charge suppression during ionization[28] and selectivity of ESI[29,30] have been reported. Hence, a refinement of formula-based aromaticity index is needed.

The objective of this study was to explore the distribution of carboxyl groups among individual components of NOM and HS of different origin and to provide a better evidence-based aromaticity index calculated with the information on the COOH groups. Despite other functional groups are integrally presented in NOM[31], the detailed labeling experiments confirmed that next to carbon-backbone (C=C) carboxylic groups (C[=O]OH) contributed most to DBE. Hence, estimation of the number of carboxylic groups will enable to provide a better approximation for molecule aromaticity.

MATERIALS AND METHODS

Solvents and other reagents used in this study were commercially available. Methanol of HPLC grade (Lab-Scan) was used for elution and dissolution of sample. High-purity distilled water (18.2 MΩ) was prepared using a Millipore Simplicity 185 system. D-enrichment of deuterated methanol (CD_3OD). Bond Elut PPL (Priority PolLutant, Agilent Technologies) cartridges (50 mg, 1 mL) were used for isolation and purification of the parent and the labeled samples. PPL represents a modified styrene-divinylbenzene polymer designed for polar organic compounds extraction. Raw mass-spectrometric data for deuteromethylated samples of various origin were obtained from the previous studies.[26,31,32] Additionally, top soil pore water from a riparian zone in a headwater catchment (Bavarian Forest National Park, Germany) and North Sea water (54.132670°N, 7.891330°E) were analyzed using deuteromethylation method. The list of all samples and their description is presented in Table S1. The number of assigned CHO–only and other formulae are presented in Tables S2.

Labeling procedures

Carboxylic groups in the parent samples were selectively deuteromethylated following the previously developed regioselective method.[32] Briefly: $SOCl_2$ (60 μL) was added dropwise to a solution of 0.5

mg of OM in 1.5 mL CD_3OD under continued stirring and ice-cooling. The reaction mixture was then refluxed for 4 h (6 h for the marine sample) and dried under vacuum. Solid residue was purified using solid-phase extraction (SPE) from aqueous solution using styrene-divinylbenzene sorbens (Agilent, Bond Elut PPL) according to the procedure described for DOM samples.[33]

Determination of labeling series by FTICR mass spectrometry

Detailed information on FTICR MS measurements can be found elsewhere.[26,31,32] For COOH-groups enumeration a Python-based script has been developed which enabled fully automatic assignment. Similar to previously developed workflow[34] the algorithm includes a juxtaposition of raw FTICR mass spectra of labeled and parent samples and formulae lists of parent samples with subtraction of one proton to reproduce m/z value of negative ions. The algorithm facilitates extraction and enumeration of peak series with mass differences corresponding to the deuteromethylation (m/z difference of 17.03448) and filtration of the obtained results based in the following heuristic rules for the number of carboxylic groups (n): number of oxygen atoms (O) inmolecular formula must be $\geq 2n$; in case of a high deuteromethylation yield and the absence of the parent peak, first peak in the labeling series must correspond to $n < 3$; labeling series must be continuous without gaps; peaks corresponding to the labeling series must absent in the parent mass spectra (important for peaks with low intensities); Mass error (between peaks of a labeling series) must be below 0.0003 m/z, which was optimized based on the FTICR MS instrument performance. The applied algorithm considers stepwise small moiety addition (i.e. H vs CD_3) to the ions, enabling fast and robust detection of labeling series.

Data treatment

Visualization of data has been performed with Python library Matplotlib (https://matplotlib.org/). Statistical analysis has been performed with Python libraries numpy and pandas. Three aromaticity indices were calculated according to equations 1-3. Experimental aromaticity index - AI_{exp} was calculated according to eq. 3 with experimentally determined n.

$$AI = \frac{1+C-O-0.5H}{C-O} \quad (1)$$

$$AI_{mod} = \frac{1+C-0.5O-0.5H}{C-0.5O} \quad (2)$$

$$AI_{con} = \frac{1+C-(COOH)_n-0.5H}{C-(COOH)_n} \quad (3)$$

Natural compounds database

To assess the AI value for individual compounds, a local copy of Collection of Open Natural Products (Coconut) database[35] was established. For each compound in the database the natural product likeness (NPL) score was calculated following Ertl et al.[36] The local minimum (0.3) of the distribution of NPL vales was used to select compounds for further analysis (Fig. S1). For manipulation with structures from the database a Python library RDkit (https://www.rdkit.org/) was used. For each compound the number of COOH-groups was extracted using fully automated algorithm[37] and AI, AI_{mod} and AI_{con} were calculated according to eq. 1-3. Only CHO compounds were considered for further evaluation to exclude structures, which can not be ionized by the negative ESI.

In-silico molecular formulae dataset

To further extend the analysis to all possibilities of COOH-functionality regardless of the actual molecular existence, all possible CHO-molecular compositions were generated in-silico as it has been described elsewhere.[38] In brief: formulae were generated in the range of molecular weights from 200 to 800 Da using integer atomic weights of elements (e.g. 12, 1, 16). Subsequently, formulae were filtered according to the typical NOM and HS compositional space - $0.27 \leq H/C \leq 2.2$, $0 < O/C \leq 1$[39] and Senior's rules, which estimate the plausibility of chemical graph existence.[40] This resulted in 21,617 CHO formulae in the test dataset. Additionally, to each individual formula all possible numbers of carboxylic groups were assigned (Table S3). For example, molecule with five oxygen atoms in formula may contain zero, one and two COOH-groups in its structure and all three cases were added to the dataset. AI, AI_{mod} and AI_{con} were calculated according to eq. 1-3 for the whole dataset.

RESULTS AND DISCUSSION

Distribution of carboxylic groups in HS and NOM samples

We collected data for eleven samples from different origins (coal, marine, peat, permafrost, black-water river, and soil) cover a broad range of molecular composition and structural diversity (Table S1). The distributions of experimentally determined COOH-groups are shown in Fig. 1 for all samples. In all cases carboxylic groups were assigned to over one thousand molecular formulae based on CD_3-labeling and their distribution in the molecular H/C vs O/C space are shown in Fig. S2. For terrestrial samples isolated from peat and coal the majority of compounds possessed between zero and two carboxylic groups. A similar result was obtained for soil extracts but its fractions isolated at pH 2 and pH 3 included also higher number of formulae with three and four carboxylic groups. Surprisingly, synthetic BP-Cx sample obtained by oxidation of lignin hydrolysate was dominated by species with two to four COOH-groups, which indicates drastic oxidation of primary alcohols in lignin moieties during reaction.[41] Top soil porewater (TSDOM) contained the largest number of polycarboxylic acids while marine (NSDOM) and riverine (SRFA) DOM also contained a wide range of molecules with up to six COOH-groups. It should be noted that peak series from deuteromethylation typically consist of multiple ions with different numbers of COOH.[32] The contribution of isomers (with variable number of COOH groups) to the observed peak series cannot be elaborated without applying chromatography or other separation methods.[42–44] Therefore, the experimental setup used in this study results in an upper estimation of carboxylic group functionality for each formula.

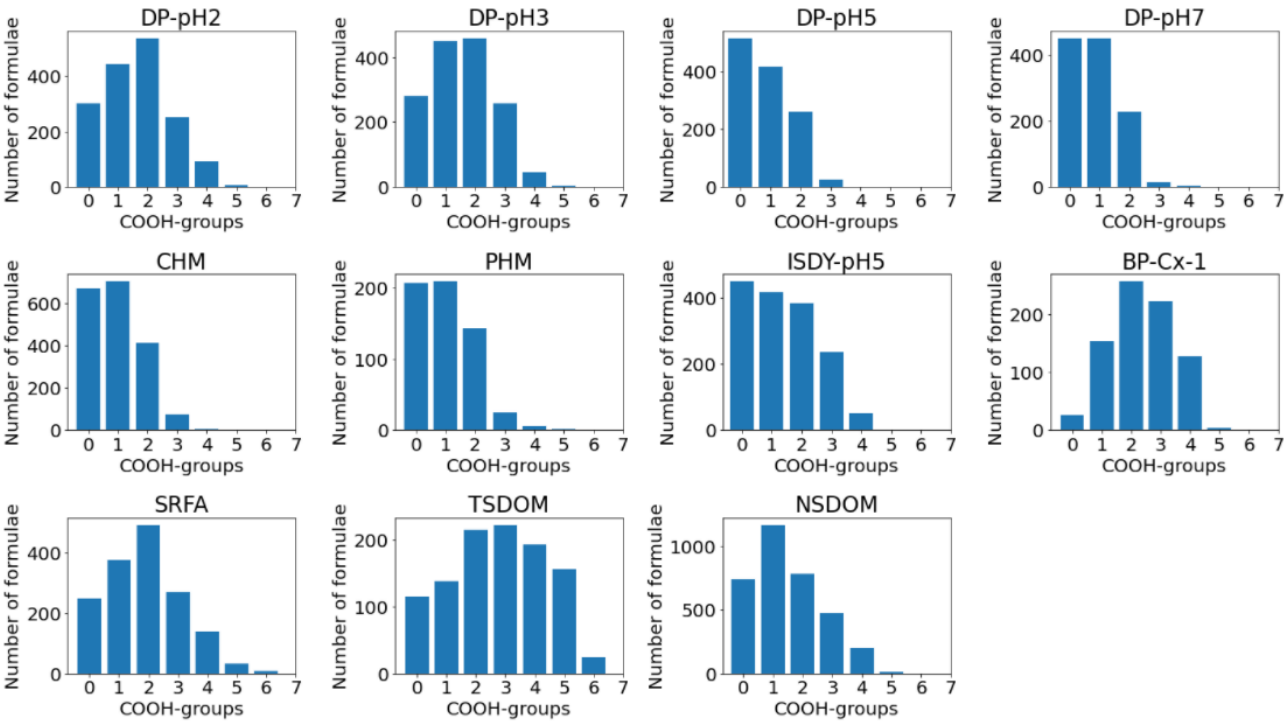

***Figure 1**. Distribution of carboxylic (COOH) groups in formulae from various NOM and HS samples as determined by deuteromethylation and FTICR MS. For a description of the samples refer to Table S1.*

Despite clear differences in carboxylic functionality between samples the maximum of COOH distribution correspond to one and two carboxylic groups per molecule for all samples, even for the most acidic, which was obtain by acid-base fractionation of soil water extract (DP-pH2; Fig 1).[26] Consequently also the overall distribution of COOH in formulae and their contribution to the total intensity across all samples revealed overall maxima at one and two COOH-groups (Fig. S3). Comparing this to the distribution of oxygen for (all and COOH-assigned) formulae in each sample (Fig. S4) and taking into account that most of the NOM components may contains no more than two carbonyl groups,[45,46] we conclude an overall high contribution of functional groups with sp^3-hybridized oxygen atoms: alcohols, phenols and ethers explaining the remainder of oxygen atoms (Fig. S5). This indicates that in most cases the actual COOH-abundance on OM is highly overestimated by AI and AI_{mod} (eq. 1-2) and the carbon skeleton aromaticity is likely significantly higher than it is deduced even from AI_{mod}.[24]

In order to deeper examine the estimation of COOH-groups used for conventional AI_{mod}, formulae distribution with different amount of assigned COOH-groups in formulae was obtained for mass ranges between 200 to 800 Da (Fig. 2). The number of COOH-groups is mostly independent on the mass and the maximum number of formulae were assigned with two COOH-groups for a broad range of masses.

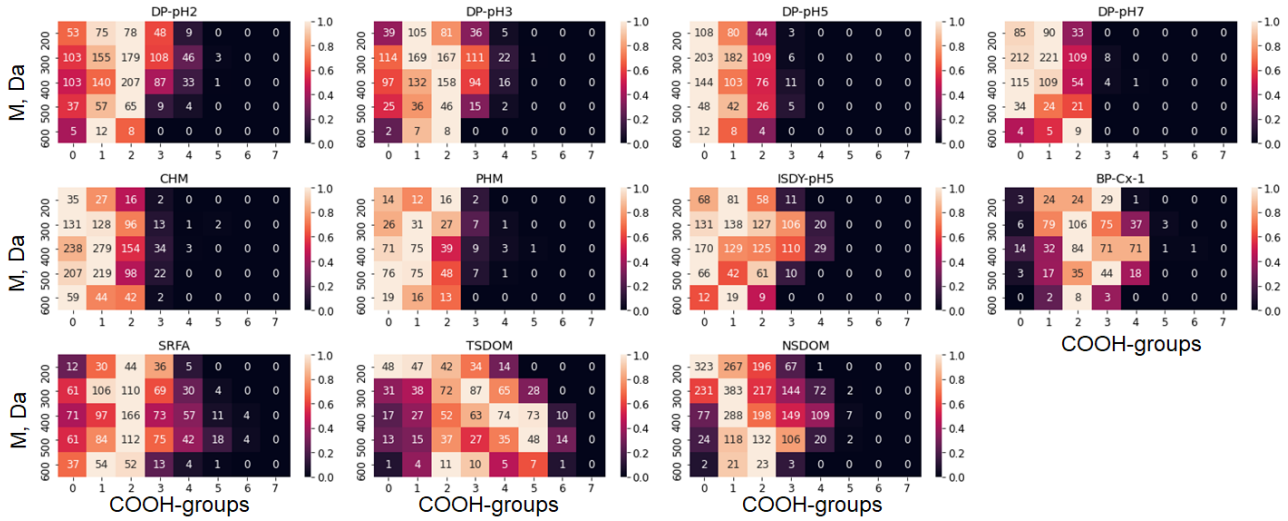

Figure 2. *Number of formulae with different number of carboxylic groups as determined by deuteromethylation and FTICR MS. Color corresponds to the proportion of the number of COOH-groups (0-7) in each mass range.*

Distribution of carboxylic groups in natural compounds

In order to further investigate the importance of COOH-functionality determination on aromaticity estimation we evaluated compounds in the largest database of individual natural compounds, the Coconut database. Only compounds with NPL-score exceeding 0.3 were considered to exclude synthetic compounds. The filtered database contained 161 316 structures (only CHO molecules) corresponding to 14 485 unique molecular formulae. In fact, most natural compounds in the mass range between 100 and 2000 Da do not contain any COOH-group and only a fraction of compounds contains one or more COOH (Fig. 3A) although these compounds are overall rich in oxygen (Fig. 3B).

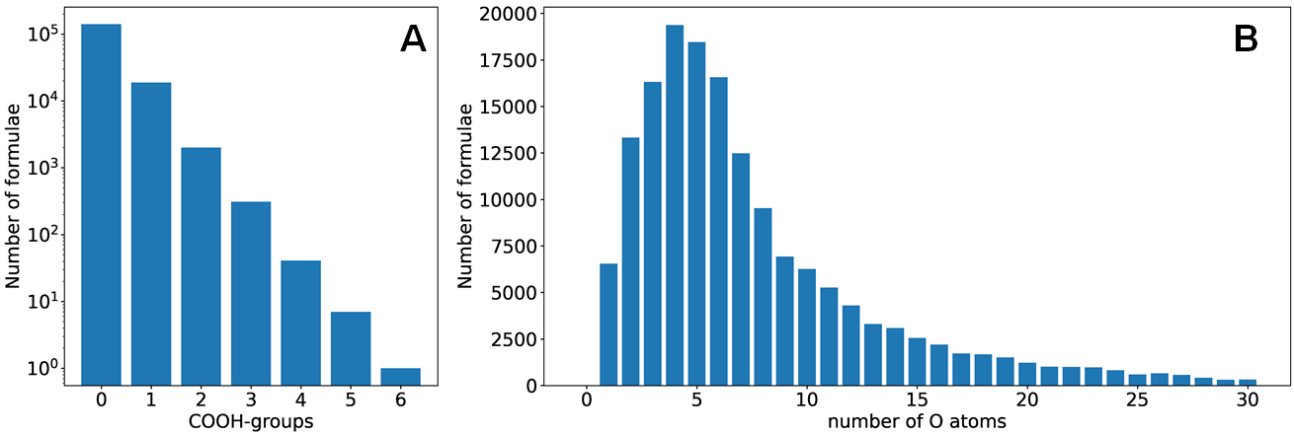

***Figure 3**. Distribution of natural compounds from Coconut database*[35] *according to the number of (A) carboxylic groups (note log scale), (B) oxygen atoms.*

Natural products in the Coconut database are by definition non-degraded compounds often containing long-aliphatic substituents adjacent to the alicyclic or aromatic cores and non-oxidized functional groups – alcohols and even aldehydes. Such saturated compounds are not resistant to biodegradation in the environment and may only be found (intact or minor transformed) in permafrost NOM, as deduced from NMR spectroscopy studies[47,48] while polyphenols without carboxylic groups present in the database are widely distributed in various eco-systems. Overall, NOM and HS were characterized by having the highest probability for *n(COOH)* = 2, while COOH-groups are mostly absent in individual structures isolated from nature. For further analysis we considered zero and two COOH groups as reference points which can be inserted into eq. 3 as number of COOH.

Carboxylic groups reference points for NOM and HS components aromaticity estimation

For the set of samples for which the number of COOH-groups were experimentally determined, an experimental aromaticity index (AI_{exp}) was derived. This index was further used as reference for statistical evaluation of different formula-based, calculated AI indices – AI, AI_{mod} and AI_{con} (eq. 1-3), the latter with two values (n = 0, 2) for COOH-groups according to the experimentally determined COOH-group distributions (Fig. S3). The calculated indices for HS and NOM components are plotted

versus experimental aromaticity index AI_{exp} (Fig. 4A-D). Expectedly, both AI and AI_{mod} mostly underestimate the compound aromaticity. Actually, original AI is calculated assuming that all oxygen atoms are bound to sp^2-hybridized carbon (i.e. C=O), which strongly reduces the available DBE to account for aromatic moieties in oxygen-rich NOM formulae. Expectedly for AI_{mod}, with the approximation of COOH groups covering all oxygen atoms (eq. 2), aromaticity can be estimated more accurately than AI, but in most cases experimentally derived aromaticity it is still significantly underestimated by AI_{mod}. Interestingly the NSDOM possessed a distinct shift toward lower number of non-carboxylic oxygen atoms when considering CHO-only formulae (Fig. S5B), corroborating NMR-based estimates of COOH functionalities for calculate AI_{mod} for marine DOM. For some samples there is indeed a connection between COOH-content and oxygen number (e.g. BP-Cx-1 and TSDOM) and the maximum of the number of formulae for each number of COOH approaches O/2 (Fig. S6), but most of the formulae are still below this upper threshold. The overall best (i.e. most accurate) approximation of AI_{exp} was obtained with the constant number of COOH groups (*n(COOH)* = 2) in eq. 3, which was chosen as reference point (Fig. 3E,F).

Calculation of AI_{con} using *n(COOH)* = 2 may result in underestimation of aromaticity, e.g. for polyphenols with a lack of carboxylic groups. Using *n(COOH)* = 0 in eq. 3 results in the transformation of AI_{con} to trivial DBE/C (in case of CHO), which ignores all oxygenated functional group with double bonds. Consequently, aromaticity of NOM and HS in that case is strongly overestimated (Fig. 3C). Despite simplification if using constant value of *n(COOH)* AI_{con} provides significantly better results as compared to AI_{mod} also for individual, biogeochemically diverse samples (Fig. S7). Only for the synthetic BP-Cx-1 sample, error distribution of AI_{mod} was comparable to the AI_{con} with *n*=2 whereas even for the marine sample, AI_{con} with *n*=2 resulted in a more accurate estimate as compared to AI_{mod}.

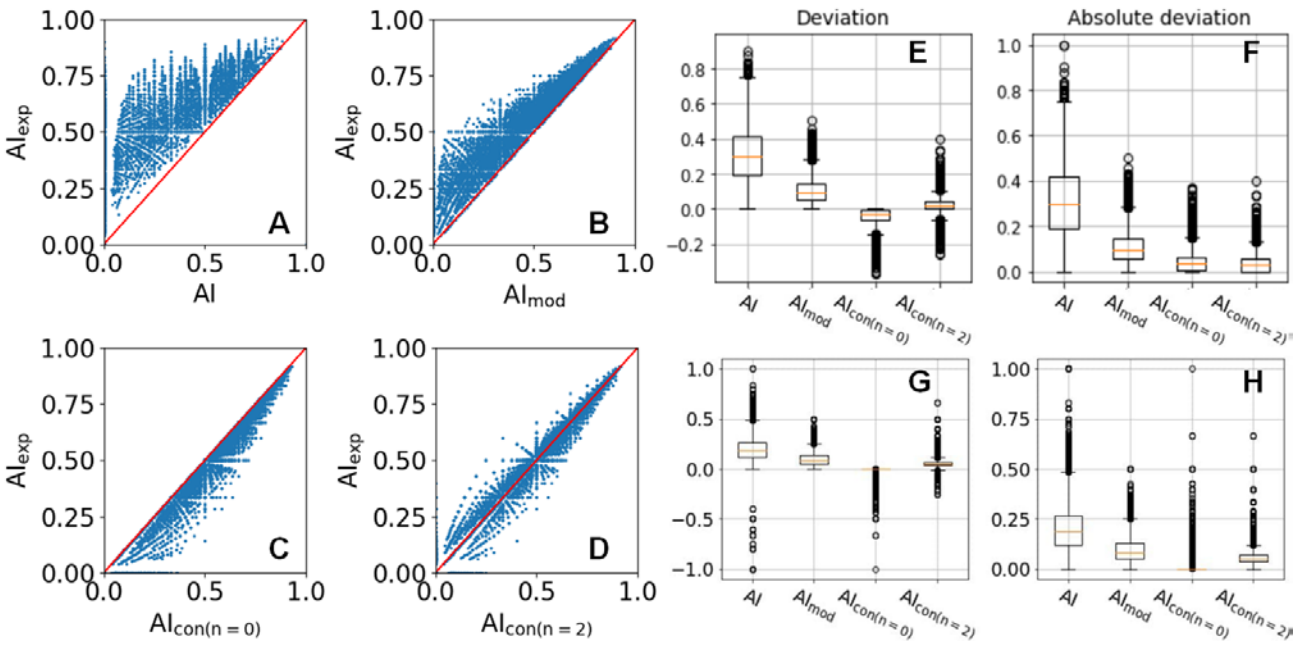

Figure 4. *A-D) Experimental AI index vs estimated aromaticity obtained by eqs. 1-3 with n(COOH) fixed to 0 and 2 for eq. 3 for all samples under study. E,F) Error distribution for calculated aromaticity indices for all NOM and HS samples combined against experimentally obtained AI. G,H) Error distribution for aromaticity indices for all Coconut CHO compounds against structure-derived AI.*

The statistical evaluation of aromaticity indices was also performed for the Coconut database, in which the number of COOH-groups was directly extracted from the structures. Exact structure-derived aromaticity index vs estimated aromaticity plots are presented in Fig. S8 and the corresponding error distribution shown in Fig. 4(G,H). Obviously, AI_{con} with *n(COOH)* = 0 yielded best results since the prevailing number of structures in Coconut are devoid of carboxylic groups. However, taking into account that most of the database compounds are less oxidized as compared to NOM and HS it was of interest to examine conventional AI and AI_{mod} for natural compounds. Clearly, approximations of AI and AI_{mod} result in significant error in aromaticity estimations. The absolute error exceed 0.3 for AI_{mod} in many cases but the third quartile of the error distribution is less for the Coconut database compounds than for the NOM components. This is explained by the maximum of the oxygen distribution (4-5) for Coconut compounds, which corresponds to a small O/2 coefficient in eq. 2 for AI_{mod}.

In order to examine the applicability of the proposed carboxyl reference point (*n(COOH)* = 2) for the entire domain of possible CHO formulae and partially account for formulae present in NOM which were absent in the current study, the in-silico formulae dataset with all possible values of COOH-groups was used. Resulting COOH-distribution and statistical assessment are presented in Fig. S9. As expected, both AI and AI_{mod} underestimated C-C accounted unsaturation of this artificial dataset, while AI_{con} with *n(COOH)* = 2 in eq. 3 resulted in adequate skewness and low median value of errors. Overall, the results for NOM samples, Coconut database and in-silico dataset for all possible variants of carboxylic functionality for CHO species strongly suggest that aromaticity can be reliably estimated for a wide range of natural species from both fresh and degraded organic matter. Overall, using AI_{con} with *n*=2 (eq. 3) results significantly smaller errors in sample comparison or even aromaticity estimate for a single molecular component than conventional AI_{mod}.

Estimation of aromaticity for different AI-based classes of NOM and HS

In biogeochemical studies of NOM and HS the suite of molecular formulae of samples is often divided into different compound classes based on atomic ratios (H/C, O/C) and aromaticity index.[49] This approach is widely applied, for example, to find correlations between molecular composition and optical properties of NOM.[19,50,51] Since these compound classes often imply a specific biogeochemical reactivity, the resulting error of aromaticity estimation of proposed AI_{con} with fixed *n(COOH)* = 2 and AI_{mod} with O/2 coefficient for different molecular classes was assessed (Fig. 5). Analysis of three most abundant compound classes, which may contain aromatic moieties ("aromatics", "condensed (aromatics)" and "unsaturated")[49] highlighted the advantages of the proposed metric (AI_{con} with *n(COOH)* = 2) over AI_{mod}. Using AI_{con} resulted in a more adequate attribution of formulae to compound classes even in case of unsaturated compounds, which are often referred as carboxyl-rich alicyclic molecules (CRAM).[52] For example, Coconut database includes a number of compounds with terpenoid scaffolds, which are unsaturated but do not contain aromatic rings. At the same time aromatic compounds with long-chain aliphatic substituents are also abundant. Without structural

elucidation it is impossible to distinguish between them, however, experimentally (by enumeration of COOH groups) or by calculation of the proposed AI_{con} with set n = 2 for COOH groups, it is possible at least to suggest aromatic moieties while AI_{mod} does not reliably indicate aromatic structures in such cases. Recently, we demonstrated that DOM from permafrost soil contain CRAM type molecules while the same molecular formulae in soil DOM from a temperate region was assigned to as aromatic compounds.[26]

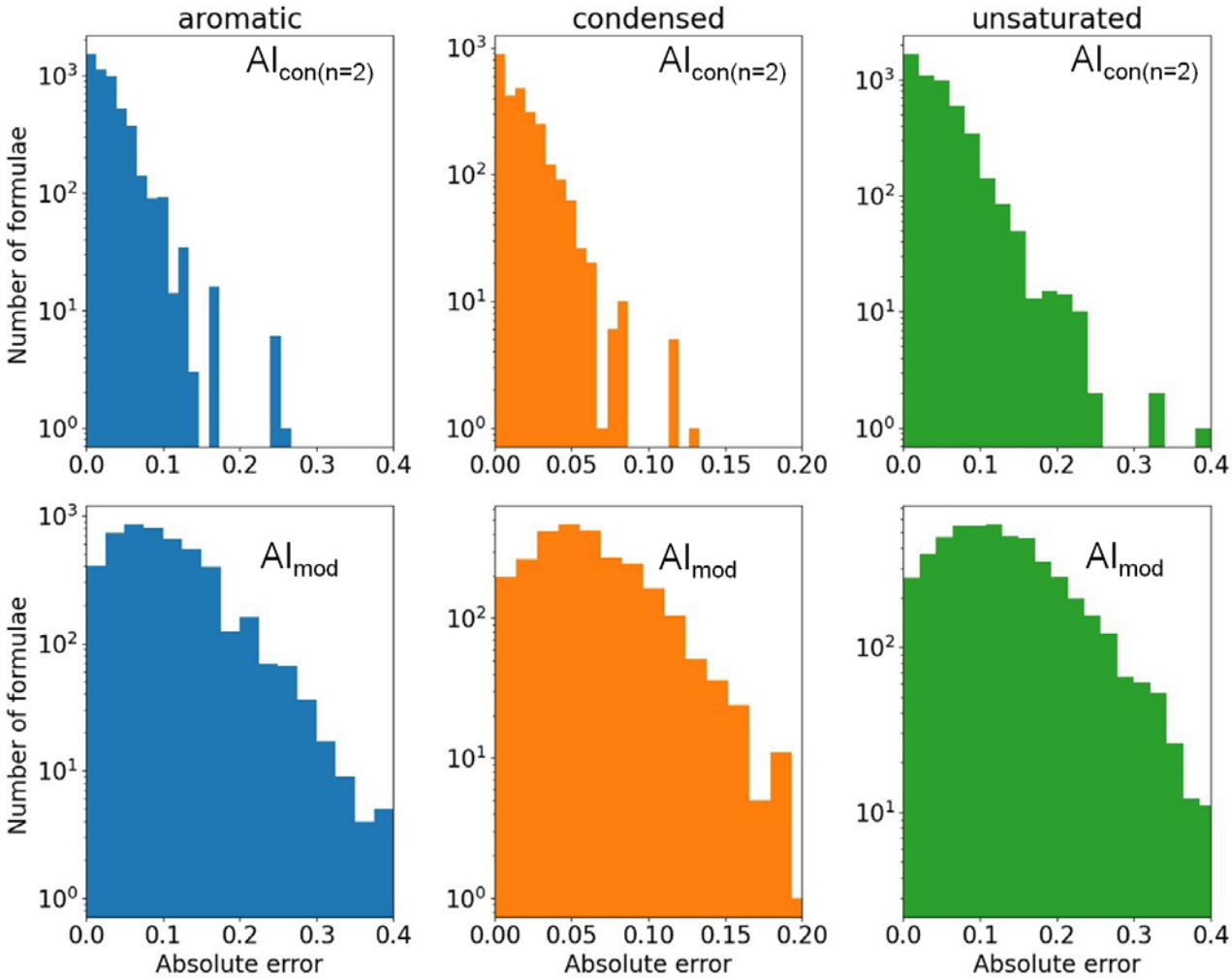

Figure 5. *Absolute error distribution of aromaticity estimation against AI_{exp} for three classes*[49] *of NOM and HS components, which may contain aromatic moieties, based on atomic ratios and (upper-row panels) AI_{con} with set COOH-group number n(COOH) = 2 or AI_{mod} (lower-row panels).*

In order to evaluate the applicability of the proposed AI_{con} with *n*=2, several points should be taken into account. Firstly, in the present work only negative ESI has been considered and applicability

of AI_{con} should be carefully used and tested in case of other ionization techniques. Secondly, setting of a fixed number for COOH-groups may still result in false conclusions when discussing the structure of compounds detected by FTICR MS. Here deuteromethylation labeling, MS/MS experiments or other techniques can be used to more precisely enumerate COOH groups. Additionally, the conservative attribution of all N,S-atoms to moieties with π-bonds remains disadvantageous and requires chemical justification. Finally, for different types of samples, the maximum of COOH-distribution can vary. Therefore, it can be expected that AI_{con} with variable *n(COOH)*-values may better describe specific types of samples. For example, considering a range of *n(COOH)*-values for CHM and SRFA samples revealed that n=1 is more suitable for the coal sample, and n=2 for the blackwater river (Fig. S10). However, in the range between 0 and 3, AI_{con} always resulted in a smaller error as compared to AI_{mod} considering experimental derived number of COOH. In conclusion, AI_{con} with *n(COOH)*=2 is a robust and precise metric for the mean aromaticity estimation, especially for CHO-only compounds. AI_{con}, which can be easily calculated from molecular formulae, can substitute conventional AI and AI_{mod} as a working metric for biogeochemical researches including NOM and HS with different degree of microbial and oxidative transformations.

ASSOCIATED CONTENT

Supporting Information

Additional experimental data, sample description, COOH and oxygen atoms distributions, including examination of aromaticity index for each sample, and Table S3 with in-silico generated CHO molecular compositions with theoretically assigned COOH-group numbers.

Acknowledgments

This work was supported by Russian Science Foundation grant 21-47-04405 and the German Research Foundation (grant 445025664). The authors are grateful to Maria P. da Silva (UFZ) for providing the soil pore-water, Boris P. Koch (AWI) for providing and Rebecca Matos (UFZ) for

labelling the marine sample, and for using the analytical facilities of the Centre for Chemical Microscopy (ProVIS) at the Helmholtz Centre for Environmental Research, Leipzig which is supported by the European Regional Development Funds (EFRE - Europe funds Saxony) and the Helmholtz Association. AZ, GDR and EN acknowledge support from the European's Horizon 2020 Research and Innovation Program under grant agreement No. 731077

Competing financial interest

The authors declare no competing financial interests.

References

(1) D'Andrilli, J.; Cooper, W. T.; Foreman, C. M.; Marshall, A. G. An Ultrahigh-Resolution Mass Spectrometry Index to Estimate Natural Organic Matter Lability. *Rapid Commun. Mass Spectrom.* **2015**, *29* (24), 2385–2401.

(2) Yuan, C.; Sleighter, R. L.; Weavers, L. K.; Hatcher, P. G.; Chin, Y. P. Fast Photomineralization of Dissolved Organic Matter in Acid Mine Drainage Impacted Waters. *Environ. Sci. Technol.* **2019**, *53* (11), 6273–6281.

(3) Waggoner, D. C.; Wozniak, A. S.; Cory, R. M.; Hatcher, P. G. The Role of Reactive Oxygen Species in the Degradation of Lignin Derived Dissolved Organic Matter. *Geochim. Cosmochim. Acta* **2017**, *208*, 171–184.

(4) Stubbins, A.; Mann, P. J.; Powers, L.; Bittar, T. B.; Dittmar, T.; McIntyre, C. P.; Eglinton, T. I.; Zimov, N.; Spencer, R. G. M. Low Photolability of Yedoma Permafrost Dissolved Organic Carbon. *J. Geophys. Res. Biogeosciences* **2017**, *122* (1), 200–211.

(5) Weishaar, J. L.; Aiken, G. R.; Bergamaschi, B. A.; Fram, M. S.; Fujii, R.; Mopper, K. Evaluation of Specific Ultraviolet Absorbance as an Indicator of the Chemical Composition and Reactivity of Dissolved Organic Carbon. *Environ. Sci. Technol.* **2003**, *37* (20), 4702–4708.

(6) Ohno, T. Fluorescence Inner-Filtering Correction for Determining the Humification Index of Dissolved Organic Matter. *Environ. Sci. Technol.* **2002**, *36* (4), 742–746.

(7) Zsolnay, A.; Baigar, E.; Jimenez, M.; Steinweg, B.; Saccomandi, F. Differentiating with Fluorescence Spectroscopy the Sources of Dissolved Organic Matter in Soils Subjected to Drying. *Chemosphere* **1999**, *38* (1), 45–50.

(8) Zherebker, A.; Shirshin, E.; Kharybin, O.; Kostyukevich, Y.; Kononikhin, A.; Konstantinov, A. I.; Volkov, D.; Roznyatovsky, V. A.; Grishin, Y. K.; Perminova, I. V; Nikolaev, E. Separation of Benzoic and Unconjugated Acidic Components of Leonardite Humic Material Using Sequential Solid-Phase Extraction at Different PH Values as Revealed by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry and Correlation Nuclear Magnetic Resonance Spectroscopy. *J. Agric. Food Chem.* **2018**, *66* (46), 12179–12187.

(9) Hertkorn, N.; Ruecker, C.; Meringer, M.; Gugisch, R.; Frommberger, M.; Perdue, E. M.; Witt, M.; Schmitt-Kopplin, P. High-Precision Frequency Measurements: Indispensable Tools at the Core of the Molecular-Level Analysis of Complex Systems. *Anal. Bioanal. Chem.* **2007**, *389* (5), 1311–1327.

(10) Bell, N. G. A.; Michalchuk, A. A. L.; Blackburn, J. W. T.; Graham, M. C.; Uhrín, D. Isotope-Filtered 4D NMR Spectroscopy for Structure Determination of Humic Substances. *Angew. Chem. Int. Ed. Engl.* **2015**, *54* (29), 8382–8385.

(11) Arakawa, N.; Aluwihare, L. Direct Identification of Diverse Alicyclic Terpenoids in Suwannee River Fulvic Acid. *Environ. Sci. Technol.* **2015**, *49* (7), 4097–4105.

(12) Spranger, T.; Pinxteren, D. Van; Reemtsma, T.; Lechtenfeld, O. J.; Herrmann, H. 2D Liquid Chromatographic Fractionation with Ultra-High Resolution MS Analysis Resolves a Vast Molecular Diversity of Tropospheric Particle Organics. *Environ. Sci. Technol.* **2019**, *53* (19), 11353–11363.

(13) Hawkes, J. A.; Dittmar, T.; Patriarca, C.; Tranvik, L.; Bergquist, J. Evaluation of the Orbitrap

Mass Spectrometer for the Molecular Fingerprinting Analysis of Natural Dissolved Organic Matter. *Anal. Chem.* **2016**, *88* (15), 7698–7704.

(14) Smith, D. F.; Podgorski, D. C.; Rodgers, R. P.; Blakney, G. T.; Hendrickson, C. L. 21 Tesla FT-ICR Mass Spectrometer for Ultrahigh-Resolution Analysis of Complex Organic Mixtures. *Anal. Chem.* **2018**, *90* (3), 2041–2047.

(15) Witt, M.; Fuchser, J.; Koch, B. P. Fragmentation Studies of Fulvic Acids Using Collision Induced Dissociation Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Anal. Chem.* **2009**, *81* (7), 2688–2694.

(16) Nebbioso, A.; Piccolo, A. Basis of a Humeomics Science: Chemical Fractionation and Molecular Characterization of Humic Biosuprastructures. *Biomacromolecules* **2011**, *12* (4), 1187–1199.

(17) Hertkorn, N.; Frommberger, M.; Witt, M.; Koch, B. P.; Schmitt-Kopplin, P.; Perdue, E. M. Natural Organic Matter and the Event Horizon of Mass Spectrometry. *Anal. Chem.* **2008**, *80* (23), 8908–8919.

(18) Gonsior, M.; Peake, B. M.; Cooper, W. T.; Podgorski, D.; D'Andrilli, J.; Cooper, W. J. Photochemically Induced Changes in Dissolved Organic Matter Identified by Ultrahigh Resolution Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Environ. Sci. Technol.* **2009**, *43* (3), 698–703.

(19) Stubbins, A.; Spencer, R. G. M.; Chen, H.; Hatcher, P. G.; Mopper, K.; Hernes, P. J.; Mwamba, V. L.; Mangangu, A. M.; Wabakanghanzi, J. N.; Six, J. Illuminated Darkness: Molecular Signatures of Congo River Dissolved Organic Matter and Its Photochemical Alteration as Revealed by Ultrahigh Precision Mass Spectrometry. *Limnol. Oceanogr.* **2010**, *55* (4), 1467–1477.

(20) Spencer, R. G. M.; Mann, P. J.; Dittmar, T.; Eglinton, T. I.; McIntyre, C.; Holmes, R. M.; Zimov, N.; Stubbins, A. Detecting the Signature of Permafrost Thaw in Arctic Rivers.

Geophys. Res. Lett. **2015**, *42* (8), 2830–2835.

(21) Hockaday, W. C.; Grannas, A. M.; Kim, S.; Hatcher, P. G. Direct Molecular Evidence for the Degradation and Mobility of Black Carbon in Soils from Ultrahigh-Resolution Mass Spectral Analysis of Dissolved Organic Matter from a Fire-Impacted Forest Soil. *Org. Geochem.* **2006**, *37* (4), 501–510.

(22) Yassine, M. M.; Harir, M.; Dabek-Zlotorzynska, E.; Schmitt-Kopplin, P. Structural Characterization of Organic Aerosol Using Fourier Transform Ion Cyclotron Resonance Mass Spectrometry: Aromaticity Equivalent Approach. *Rapid Commun. Mass Spectrom.* **2014**, *28* (22), 2445–2454.

(23) Koch, B. P.; Dittmar, T. From Mass to Structure: An Aromaticity Index for High-Resolution Mass Data of Natural Organic Matter. *Rapid Commun. Mass Spectrom.* **2006**, *20* (5), 926–932.

(24) Koch, B. P.; Dittmar, T. Erratum: From Mass to Structure: An Aromaticity Index for High-Resolution Mass Data of Natural Organic Matter (Rapid Communications in Mass Spectrometry (2006) 20 (926-932) DOI: 10.1002/Rcm.2386). *Rapid Communications in Mass Spectrometry*. 2016, p 250.

(25) Zherebker, A.; Podgorski, D. C.; Kholodov, V. A.; Orlov, A. A.; Yaroslavtseva, N. V.; Kharybin, O.; Kholodov, A.; Spector, V.; Spencer, R. G. M.; Nikolaev, E.; Perminova, I.V. The Molecular Composition of Humic Substances Isolated From Yedoma Permafrost and Alas Cores in the Eastern Siberian Arctic as Measured by Ultrahigh Resolution Mass Spectrometry. *J. Geophys. Res. Biogeosciences* **2019**, *124* (8), 2432–2445.

(26) Zherebker, A.; Shirshin, E.; Rubekina, A.; Kharybin, O.; Kononikhin, A.; Kulikova, N. A.; Zaitsev, K. V.; Roznyatovsky, V. A.; Grishin, Y. K.; Perminova, I. V.; Nikolaev, E. Optical Properties of Soil Dissolved Organic Matter Are Related to Acidic Functions of Its Components as Revealed by Fractionation, Selective Deuteromethylation, and Ultrahigh

Resolution Mass Spectrometry. *Environ. Sci. Technol.* **2020**, *54* (5), 2667–2677.

(27) Aeschbacher, M.; Graf, C.; Schwarzenbach, R. P.; Sander, M. Antioxidant Properties of Humic Substances. *Environ. Sci. Technol.* **2012**, *46* (9), 4916–4925.

(28) Rodgers, R. P.; Mapolelo, M. M.; Robbins, W. K.; Chacón-Patiño, M. L.; Putman, J. C.; Niles, S. F.; Rowland, S. M.; Marshall, A. G. Combating Selective Ionization in the High Resolution Mass Spectral Characterization of Complex Mixtures. *Faraday Discuss.* **2019**, *218*, 29–51.

(29) Hockaday, W. C.; Purcell, J. M.; Marshall, A. G.; Baldock, J. A.; Hatcher, P. G. Electrospray and Photoionization Mass Spectrometry for the Characterization of Organic Matter in Natural Waters: A Qualitative Assessment. *Limnol. Oceanogr.* **2009**, *7*, 81–95.

(30) Nebbioso, A.; Piccolo, A.; Spiteller, M. Limitations of Electrospray Ionization in the Analysis of a Heterogeneous Mixture of Naturally Occurring Hydrophilic and Hydrophobic Compounds. *Rapid Commun. Mass Spectrom.* **2010**, *24* (21), 3163–3170.

(31) Zherebker, A.; Lechtenfeld, O. J.; Sarycheva, A.; Kostyukevich, Y.; Kharybin, O.; Fedoros, E. I.; Nikolaev, E. N. Refinement of Compound Aromaticity in Complex Organic Mixtures by Stable Isotope Label Assisted Ultrahigh-Resolution Mass Spectrometry. *Anal. Chem.* **2020**, *92* (13), 9032–9038.

(32) Zherebker, A.; Kostyukevich, Y.; Kononikhin, A.; Kharybin, O.; Konstantinov, A. I.; Zaitsev, K. V.; Nikolaev, E.; Perminova, I. V. Enumeration of Carboxyl Groups Carried on Individual Components of Humic Systems Using Deuteromethylation and Fourier Transform Mass Spectrometry. *Anal. Bioanal. Chem.* **2017**, *409* (9), 2477–2488.

(33) Dittmar, T.; Koch, B.; Hertkorn, N.; Kattner, G. A Simple and Efficient Method for the Solid-Phase Extraction of Dissolved Organic Matter (SPE-DOM) from Seawater. *Limnol. Ocean. Methods* **2008**, *6*, 230–235.

(34) Kostyukevich, Y.; Kononikhin, A.; Zherebker, A.; Popov, I.; Perminova, I.; Nikolaev, E.

Enumeration of Non-Labile Oxygen Atoms in Dissolved Organic Matter by Use of ^{16}O/ ^{18}O Exchange and Fourier Transform Ion-Cyclotron Resonance Mass Spectrometry. *Anal. Bioanal. Chem.* **2014**, *406* (26), 6655–6664.

(35) Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. COCONUT Online: Collection of Open Natural Products Database. *J. Cheminform.* **2021**, *13* (1).

(36) Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural Product-Likeness Score and Its Application for Prioritization of Compound Libraries. *J. Chem. Inf. Model.* **2008**, *48* (1), 68–74.

(37) Ertl, P. An Algorithm to Identify Functional Groups in Organic Molecules. *J. Cheminform.* **2017**, *9* (1).

(38) Orlov, A. A.; Zherebker, A.; Eletskaya, A. A.; Chernikov, V. S.; Kozlovskaya, L. I.; Zhernov, Y. V.; Kostyukevich, Y.; Palyulin, V. A.; Nikolaev, E. N.; Osolodkin, D. I.; Perminova, I.V. Examination of Molecular Space and Feasible Structures of Bioactive Components of Humic Substances by FTICR MS Data Mining in ChEMBL Database. *Sci. Rep.* **2019**, *9* (1), 1–12.

(39) Sleighter, R. L.; Hatcher, P. G. *Fourier Transform Mass Spectrometry for the Molecular Level Characterization of Natural Organic Matter: Instrument Capabilities, Applications, and Limitations*; INTECH Open Access Publisher, 2011.

(40) Kind, T.; Fiehn, O. Seven Golden Rules for Heuristic Filtering of Molecular Formulas Obtained by Accurate Mass Spectrometry. *BMC Bioinformatics* **2007**, *8*.

(41) Zherebker, A.; Yakimov, B.; Rubekina, A.; Kharybin, O.; Fedoros, E. I.; Perminova, I. V.; Shirshin, E.; Nikolaev, E. N. Photoreactivity of Humic-like Polyphenol Material under Irradiation with Different Wavelengths Explored by FTICR MS and Deuteromethylation. *Eur. J. Mass Spectrom.* **2020**, *26* (4), 292–300.

(42) Hawkes, J. A.; Patriarca, C.; Sjöberg, P. J. R.; Tranvik, L. J.; Bergquist, J. Extreme Isomeric Complexity of Dissolved Organic Matter Found across Aquatic Environments. *Limnol. Oceanogr. Lett.* **2018**, *3* (2), 21–30.

(43) Leyva, D.; Tose, L. V.; Porter, J.; Wolff, J.; Jaffé, R.; Fernandez-Lima, F. Understanding the Structural Complexity of Dissolved Organic Matter: Isomeric Diversity. *Faraday Discuss.* **2019**, *218*, 431–440.

(44) Stenson, A. C.; Ruddy, B. M.; Bythell, B. J. Ion Molecule Reaction H/D Exchange as a Probe for Isomeric Fractionation in Chromatographically Separated Natural Organic Matter. *Int. J. Mass Spectrom.* **2014**, *360* (1), 45–53.

(45) Bianca, M. R.; Baluha, D. R.; Gonsior, M.; Schmitt-Kopplin, P.; Del Vecchio, R.; Blough, N. V. Contribution of Ketone/Aldehyde-Containing Compounds to the Composition and Optical Properties of Suwannee River Fulvic Acid Revealed by Ultrahigh Resolution Mass Spectrometry and Deuterium Labeling. *Anal. Bioanal. Chem.* **2020**, 1–11.

(46) Baluha, D. R.; Blough, N. V.; Del Vecchio, R. Selective Mass Labeling for Linking the Optical Properties of Chromophoric Dissolved Organic Matter to Structure and Composition via Ultrahigh Resolution Electrospray Ionization Mass Spectrometry. *Environ. Sci. Technol.* **2013**, *47* (17), 9891–9897.

(47) Pautler, B. G.; Simpson, A. J.; McNally, D. J.; Lamoureux, S. F.; Simpson, M. J. Arctic Permafrost Active Layer Detachments Stimulate Microbial Activity and Degradation of Soil Organic Matter. *Environ. Sci. Technol.* **2010**, *44* (11), 4076–4082.

(48) Perminova, I. V.; Shirshin, E. A.; Konstantinov, A. I.; Zherebker, A.; Lebedev, V. A.; Dubinenkov, I. V.; Kulikova, N. A.; Nikolaev, E. N.; Bulygina, E.; Holmes, R. M. The Structural Arrangement and Relative Abundance of Aliphatic Units May Effect Long-Wave Absorbance of Natural Organic Matter as Revealed by 1 H NMR Spectroscopy. *Environ. Sci. Technol.* **2018**, *52* (21), 12526–12537.

(49) 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.* **2014**, *5*, 3804.

(50) Mann, P. J.; Davydova, A.; Zimov, N.; Spencer, R. G. M.; Davydov, S.; Bulygina, E.; Zimov,

S.; Holmes, R. M. Controls on the Composition and Lability of Dissolved Organic Matter in Siberia's Kolyma River Basin. *J. Geophys. Res. Biogeosciences* **2012**, *117* (1).

(51) Kellerman, A. M.; Guillemette, F.; Podgorski, D. C.; Aiken, G. R.; Butler, K. D.; Spencer, R. G. M. Unifying Concepts Linking Dissolved Organic Matter Composition to Persistence in Aquatic Ecosystems. *Environ. Sci. Technol.* **2018**, *52* (5), 2538–2548.

(52) DiDonato, N.; Chen, H.; Waggoner, D.; Hatcher, P. G. Potential Origin and Formation for Molecular Components of Humic Acids in Soils. *Geochim. Cosmochim. Acta* **2016**, *178*, 210–222.