Rapid Communications in Mass Spectrometry

WILEY

Rapid Communications in Mass Spectrometry

RESEARCH ARTICLE OPEN ACCESS

From Retention Time to Functional Group Assignment: A Chemical Database-Driven Approach for High-Resolution Mass Data of Marine Dissolved Organic Matter

Fabian Moye[1,2] | Marlo Bareth[2,3] | Boris P. Koch[2,4] | Jan Tebben[2] | Tilmann Harder[1,2]

[1]Marine Chemistry, Faculty of Biology and Chemistry, University of Bremen, Bremen, Germany | [2]Department of Ecological Chemistry, Alfred-Wegener-Institut Helmholtz Zentrum für Polar- und Meeresforschung, Bremerhaven, Germany | [3]Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany | [4]University of Applied Sciences, Bremerhaven, Germany

Correspondence: Fabian Moye (fmoye@uni-bremen.de) | Tilmann Harder (t.harder@uni-bremen.de)

Received: 7 November 2024 | **Revised:** 28 March 2025 | **Accepted:** 31 March 2025

Funding: This work was supported by Universität Bremen.

Keywords: chemodiversity | DOM | LC-FT-MS | octanol–water coefficient | structure prediction

ABSTRACT

Rationale: The high chemodiversity of marine dissolved organic matter (DOM) has challenged identification of singular DOM components. To infer chemical structure features of DOM with accurately determined molecular formulas, we assessed if empirically determined elution properties of stoichiometrically identical isomers obtained from a chemical database would predict features, such as the type and number of functional groups, in structurally unknown DOM isomers.

Methods: DOM of different origin (North Sea, Southern Ocean) was analysed by LC-FT-MS using two different mass spectrometry methods. Chromatographic retention of DOM isomers was correlated with calculated retention properties of structurally known isomers in PubChem. A total of 7.7 million chemical identifiers were queried for their computed octanol–water coefficients (logP). The 50 most intense molecular formulas were queried for PubChem structure data files. The number and type of structural features was assigned to logP bins across the DOM elution window, to correlate the contribution of database-derived structure features to retention time of structurally unknown DOM isomers.

Results: The intensity-weighted average of logP for $C_xH_yO_z$ in Southern Ocean water, predicted by retention time of all molecular formulas, was in good agreement with logP values stored in PubChem. The comparatively longer retention of the same isomer in North Sea versus Southern Ocean DOM suggests a decoration with fewer alcohol groups and more ring structures and esters. Earlier eluting molecular formulas are more likely to contain rings and alcohols, while later eluting ones are more linear/branched and contain more esters.

Conclusions: We hypothesised DOM isomers belonging to comparatively older Southern Ocean water to elute earlier than young and less degraded molecules present in North Sea water. This hypothesis was verified based on three exemplarily selected molecular formulas cooccurring in all water samples. Our strategy circumvents issues of chimeric fragmentation spectra and furthermore adds retention time as a new physicochemical descriptor of DOM molecules.

1 | Introduction

Marine dissolved organic matter (DOM) is among the most complex mixtures of organic molecules on earth [1,2]. Typical mass spectra of DOM contain a wide spectrum of different molecular formulas (ca. 3000) in the mass range of *m/z* 200 to *m/z* 700, each of which has at least 30 isomers [1,2], resulting in at least 90 000 different compounds. The chemical composition and molecular diversity of DOM results from both the duration and repetition of biogeochemical transformations (e.g. microbial activity, mixing, physicochemical transformations, aggregation [3–5]) of primary produced organic matter [6,7]. Consequently, the chemical composition and diversity of DOM is an imprint of its environmental origin, history and fate. For example, the stoichiometric composition differs between marine and terrestrial DOM, with the former being more aliphatic and richer in nitrogen, and the latter mainly derived from lignin, more aromatic and poorer in nitrogen [5]. Diverse reactions lead to chemical diversification of DOM, e.g. oxidative de-aromatisation of phenolic molecules [8]. In combination with cycloadditions [8], these reactions increase the molecular weight and stoichiometric ratio of O/C and consequently alter the type and number of functional groups. During transport and aging in the open ocean, DOM is oxidised and decreases in molecular weight [9–11].

These transformations do not only result in a plethora of different molecular formulas, but importantly, the same molecular formula may also result from different sources and biogeochemical and physicochemical inputs [1,7]. In other words, the historical environmental impact on the plethora of organic molecules results in distinguishable abundances of isomers, i.e. compounds sharing the same molecular formula and mass but having different structural arrangements of atoms or groups.

DOM is typically analysed by direct injection mass spectrometry, mainly with Fourier-transform ion cyclotron resonance MS (FT-ICR-MS) [12] or high-resolution Orbitrap mass spectrometry (HR-MS) [13,14]. To avoid ionisation artefacts and ion suppression, DOM is increasingly analysed in hyphenation to liquid chromatography (LC) [15–18]. Yet, even under LC-MS conditions, the enormous structural diversity of DOM results in total ion chromatograms (TICs) with broad peak shape and without baseline resolution. This is due to the fact that individual isomers elute as a continuum of overlapping gaussian peaks along the entire polarity gradient [1]. Notably, this phenomenon is not caused by single compounds 'smearing' across the chromatogram, but rather due to the plethora of individual chromatographically resolved structural isomers that together result in one unresolved continuous mass detector signal spanning the entire chromatographic retention time window. This explanation agrees with the observation that extracted ion chromatograms of internal standards spiked into a complex DOM matrix still reveal very sharp gaussian peaks [19].

The elemental ratios of DOM constituents obtained by either one of the above analytical methods serve as valuable proxies for biochemical and physiochemical processes related to origin and diagenetic fate of DOM molecules. A number of qualifiers are calculated from the molecular formulas derived from high-resolution accurate mass spectrometry to characterise and distinguish different DOM molecules by stoichiometry, origin and transformation state, e.g. double-bond equivalents (DBEs), nominal oxidation state of carbon (NOSC) [20] and the aromaticity index (AI) [21,22]. While these qualifiers are useful to characterise and distinguish between different molecular formulas, they are mostly unable to distinguish between isomers. Moreover, the high inherent structural diversity of DOM has so far challenged the separation and structural identification of singular DOM components [23], and only very few DOM molecular structures are known (e.g. Arakawa et al. [10], Seidel et al. [24], Geuer et al. [25], Papadopoulos Lambidis et al. [26]).

In this paper, we tested if we could correlate the chromatographic retention of a priori structurally unknown DOM isomers with their decoration by functional groups that largely affect their chromatographic properties. This hypothesis was inspired by the observation that extracted ion chromatograms of internally spiked standards of known accurate mass indeed revealed sharp retention despite coelution among isomers within the DOM mixture [15]. This in turn suggested that different chromatographic behaviour (i.e. retention time) of structurally unknown molecules could be useful to predict structure motifs that govern their chromatographic behaviour, such as typical functional groups (alcohols, aldehydes, esters, carboxylic acids, ethers, alicyclic and aromatic rings). We hypothesised that the position of narrow retention time bins across the DOM chromatogram would indirectly offer chemical structure information of DOM isomers, such as the type and number of functional groups decorating these isomers.

Chromatography theory posits the three-dimensional structure and surface charge distribution on any molecule to largely govern its equilibrium between adsorption on stationary versus solubility in a mobile phase, and hence result in chromatographic separation of structurally different analytes. The octanol–water coefficient (K_{OW} or P; commonly log-transformed as logP) describes the liquid partitioning between a non-miscible apolar and polar phase and thus approximates the thermodynamic equilibrium on chromatography columns [16,27,28]. It has been successfully used to predict chromatographic retention times [16,29,30]. The coefficient logP is determined either experimentally or predicted computationally, e.g. with the XLOGP3-algorithm [31]. To test the conceptual idea outlined above, we theoretically correlated the chromatographic retention of accurate mass-derived molecular formulas of structurally unknown DOM isomers with curated structure database information of known isomers having the same molecular formula.

In this study, we analysed three DOM samples of different origin, including coastal North Sea water and shallow and deep Southern Ocean water, by LC-FT-MS. We correlated the chromatographic retention of individual DOM isomers in these samples with calculated chromatographic retention properties of structurally known isomers obtained from a compound database. Briefly, molecular formulas of DOM constituents in each sample were annotated and the assigned molecular formulas searched in PubChem (https://pubchem.ncbi.nlm.nih.gov/). More than 7.7 million chemical identifiers were filtered and queried for logP entries. The molecular formulas were

ranked by intensity, and the 50 most intense molecular formulas of each water sample were queried for PubChem structure data files resulting in a dataset of 11 237 structural isomers with known features, such as alcohols, aldehydes, esters, carboxylic acids, ethers, alicyclic and aromatic rings. The distribution of the number and type of these structural features was assigned to logP bins across the DOM sample elution window in order to correlate the contribution of database-derived structure features to the retention time of structurally unknown DOM isomers. Three of the 50 most intense molecular formulas were exemplarily chosen to test if the same isomers in comparatively young (North Sea) and old (Southern Ocean) DOM could be distinguished by chromatographic retention. Given that more recalcitrant and older DOM is generally more transformed and structurally modified by polar functional groups, such as carboxylic or hydroxy functions [10,32,33], we hypothesised the DOM isomers belonging to comparatively older Southern Ocean water to elute earlier than young and less degraded molecules present in North Sea water.

2 | Material and Methods

2.1 | Samples

Two samples of the Weddell Sea (latitude −67.5633°, longitude −55.3448°) from different depths representing marine (30 m depth) and refractory DOM (1356 m depth) were obtained as described elsewhere [34,35]. Briefly, 160 L Antarctic seawater were sampled with a rosette sampler on RV Polarstern during ANT XXII/2. The water was filtered with 0.2 μm filter cartridges, acidified to pH 2 and pumped through 60 mL solid phase extraction cartridges (PPL, 5 g). DOM was eluted with 40 mL methanol and stored at −18 °C. Fresh coastal DOM is extracted routinely by us from the Southern North Sea (latitude 54.1447°, longitude 7.8711°) and used as laboratory standard. Sea water was sampled over 0.2 μm PTFE (Whatman) filter on RV Uthörn, acidified to pH 2 and extracted with PPL cartridges. DOM was eluted with methanol and extracts were stored at −18 °C until measurement to minimise esterification in methanol [36].

2.2 | Liquid Chromatography and Mass Spectrometry

While FT-MS is well established, comparative interlaboratory experiments have shown that identified molecular formulas are not well reproduced between different instrument platforms [37]. Therefore, we chose to analyse the data with two instruments to confirm that identified analytical trends were reliable across different analytical platforms.

All samples were analysed on two instruments: (a) Fourier Transform Orbitrap mass spectrometer (FT-Orbitrap-MS; Q-Exactive Plus, Thermo Fisher Scientific, Bremen, Germany) coupled to ultra-high performance liquid chromatography (UPLC, Vanquish, Thermo Fisher Scientific, Bremen, Germany), herein referred to as LC-FT-Orbitrap-MS, and (b) 7 T scimaX MRMS system (Bruker Daltonics GmbH & Co. KG, Bremen, Germany) coupled to UPLC (Elute LC, Bruker Daltonics GmbH & Co. KG, Bremen, Germany), herein referred to as LC-FT-ICR-MS. Reversed-phase chromatography was done with a C18 column (Waters AQUITY 2 x 100 mm, 1.7 μm) at 0.3 mL min^{-1} and a linear gradient of A (ultrapure water, 4 mmol L^{-1} ammonium formate) 2 min, 99%; 11 min, 0%; 14.9 min, 99%; B (MeOH, 4 mmol L^{-1} $CHOONH_4$) 2 min, 1%; 11 min, 100%; 14.5 min, 100%; 14.9 min, 1%.

LC-FT-Orbitrap-MS was performed using heated electrospray ionisation (HESI-II, Thermo Fisher Scientific, Bremen, Germany) in negative mode with a spray voltage of 2.5 kV, capillary temperature of 256 °C, sheath gas flow of 47.5 and auxiliary gas flow rate of 11.3 (arbitrary units). Full MS mode with a resolution of 280 000 (m/z 200) and a scan range of m/z 150 to m/z 1875 was used. The ion optics and ESI source settings were tuned to maximise the intensity of the peak at m/z 400. Each scan was internally calibrated with nine lock masses (339.10854, 369.11911, 411.12967, 469.13515, 541.15628, 595.20323, 611.19814, 651.22944, 693.24001). The Xcalibur software package (Thermo Electron Corporation) was used to export the mass lists, intensity, noise level and resolution to individual peak lists.

LC-FT-ICR-MS were acquired with a capillary voltage of 4 kV. Drying gas temperature was set to 220 °C, Neb gas flow to 7 L min^{-1} and drying gas pressure to 2 bar. Measurements were performed using electrospray ionisation in negative ion mode with a mass resolution of 640 000 (m/z 200) using 4 M data points. The scan range was set to m/z 107 to m/z 2000, and eight single scans were added. The spectra were online lock mass-calibrated with the formate adduct of Hexakis(1H,1H,2H-perfluoroethoxy)phosphazene at m/z 666.019887. The method parameters were adjusted for maximum intensities at m/z 400. DataAnalysis 5.3 software package (Bruker Daltonics GmbH & Co. KG, Bremen, Germany) was used to export the mass lists, intensity and mass resolution of the added scans every 1.1 min.

2.3 | Elution Properties of Standard Compounds in a DOM Sample

Twenty-five standard substances (see Table S1) were measured under the same chromatographic conditions to determine their retention time and chromatographic peak width. Subsequently, 12 standard substances (Table S2) were selected by their ionisation efficiency in negative ESI and spiked to the North Sea water sample. Their concentrations were adjusted to achieve an intensity of ≈10^6, a common intensity range for DOM peaks in the LC-FT-Orbitrap-MS.

2.4 | Annotation of Molecular Formulas

Scans were averaged to time bins of 1 min to reduce data size and enhance S/N ratios. Each scan was externally calibrated with the Ultra-Mass-Explorer (UME, https://gitlab.awi.de/bkoch/ume; Leefmann et al. [38]) using a linear calibration and 839 known DOM molecular formulas as calibrants. Molecular formulas were annotated using UME with the following elemental composition: $^{12}C_{\leq\infty}{}^{1}H_{\leq\infty}{}^{16}O_{\leq\infty}{}^{14}N_{\leq2}{}^{32}S_{\leq1}$. The mass accuracy was 0.3 ppm for LC-FT-ICR-MS and 0.8 ppm for LC-FT-Orbitrap-MS.

The following molecular properties were calculated: DBEs, NOSC according to LaRowe et al. [20] with neutral charge and zero phosphorus and the aromaticity index [21,22]. The index of degradation (I_{DEG}, Flerus et al. [11]) was calculated if all 10 molecular formulas were present. Similarly, the terrestrial index (I_{terr}, Medeiros et al. [39]) was calculated if all 80 molecular formulas were present. Molecular formulas were filtered by a DBE-*O* (DBE minus number of oxygen atoms) to maximum of 10 [40], and molecular formulas were only considered when a ^{13}C or ^{34}S isotope mass peak was present. Formulas identified as part of a list of potential surfactants [38,41] were removed from the data set and the blank subtracted. The LC-FT-ICR-MS data did not contain any multiple assignments, while 0.12% of the LC-FT-Orbitrap-MS peaks matched either $C_xH_yO_z$ or $C_xH_yO_zN_{1-2}S_1$ (x, y and z stand for any possible number within the boundaries of annotation). Multiple assignments were filtered by $\min(|DBE - \#O|)$ [40]. The absolute intensities were converted to relative intensities by sum normalisation so that the sum of all intensities per bin was 1. Detailed equations are reported in S1.1. Ion chromatograms were extracted by converting the LC-FT-Orbitrap-MS files into .mzML with proteowizzard (version 3.0.22353) [42] and using the R API of MassQL [43].

2.5 | Calculation of Octanol–Water Coefficients

The octanol–water coefficient is defined as the partitioning ratio of a substance in octanol (c_O) and water (c_W).

$$K_{OW} = P = \frac{c_O}{c_W}$$

With the logP as:

$$\log_{10}P = \log_{10}\left(\frac{c_O}{c_W}\right) = \log_{10}c_O - \log_{10}c_W$$

LogP is not routinely measured; therefore, in databases like PubChem, logP values are predicted computationally. Calculated logP values, according to the XLOGP3-algorithm [31], are referred to as XlogP. The assigned molecular formulas were searched in the PubChem database (https://pubchem.ncbi.nlm.nih.gov/) and queried for XlogP entries. More than 7.7 million chemical identifiers (CIDs) were found and filtered to remove entries containing two molecules or peroxide structures. The 50 most intense molecular formulas detected by the mass spectrometer within all three DOM samples (LC-FT-ICR-MS) were additionally queried for their structure data files (sdf). After removing peroxides 11 580 sdf of the 50 most intense molecular formulas remained.

2.6 | Data Evaluation

Data were evaluated using R [44]. Structure-data files queried from PubChem were analysed with 'ChemmineR' [45]. Recursive feature elimination (random forest, 'ranger' package [46,47], 1000 trees, 5 times repeated cross-validation, 5 folds) was used to find relevant properties for the prediction of XlogP. Further statistics were reported according to Greenacre [48].

3 | Results

3.1 | Chromatographic Peak Width as a Measure of Chemodiversity

Under the chosen chromatographic conditions (see above), DOM eluted from 4 to 13 min. The total assigned intensity (TAI) showed a maximum at 7 min with both methods, LC-FT-ICR-MS and LC-FT-Orbitrap-MS (Figure 1a). Exemplary molecular formulas revealed the known broad DOM elution profile across several minutes, spanning the complete elution gradient (4 to 10 min, Figure 1b). However, elution maxima differed between different molecular formulas: early $C_{16}H_{18}O_9$, $C_{16}H_{20}O_9$ (RT, 4 to 6 min), mid $C_{19}H_{25}NO_9$, $C_{20}H_{26}O_{10}$ (RT, 6 to 7 min) and late $C_{20}H_{30}O_9$, $C_{21}H_{30}O_8$ (RT, 7 to 11 min) (Figure 1b). For comparison, standard substances spiked to a North Sea water DOM sample had an elution window of 0.1 to 0.4 min, typical for UHPLC applications (Figure 1c).

The chromatographic peak width of DOM molecular formulas correlated with their intensity (Figures 2, S3 and S4). The molecular formulas were separated into four quantiles (1. Q, 25%; 2. Q, 50%; 3. Q, 75%; 4. Q, 100%) based on their log-transformed intensity (base 10). Overall, $C_xH_yO_z$ molecular formulas had the highest intensities and peak widths. The peak width distribution for all quantiles spread across the complete elution gradient. The North Sea sample was chosen as an example, as the shallow and deep Southern Ocean samples behaved similarly (see Figure S5). $C_xH_yO_z$ molecular formulas (x, y and z represent any possible number of atoms within the annotation range) had the broadest elution window of up to 13 min. In the lowest intensity quartile, their elution window differed between 1 min and 13 min, while in the highest intensity quartile all $C_xH_yO_z$ molecular formulas ranged between 8 and 13 min peak width. $C_xH_yO_zN_{1-2}$ molecular formulas behaved similarly on both instruments, but only four molecular formulas occurred in the highest intensity quartile for LC-FT-ICR-MS (LC-FT-Orbitrap-MS: 255 molecular formulas). Ninety-seven $C_xH_yO_zS$ and $C_xH_yO_zN_{1-2}S$ molecular formulas were detected by LC-FT-ICR-MS. They mostly occurred in the North Sea sample (North Sea: 39, Shallow SO: 30, Deep SO: 28) with a peak width between 1 and 3 min in the first quartile and between 1 and 6 min in the fourth quartile. By LC-FT-Orbitrap-MS, we detected 814 $C_xH_yO_zS$ and $C_xH_yO_zN_{1-2}S$ molecular formulas distributed over all samples (North Sea: 294, Shallow SO: 263, Deep SO: 257). Their peak width did not steadily increase but peaked in the third quartile in the North Sea sample with an elution window between 1 and 9 min for $C_xH_yO_zS$.

3.2 | Comparison of Empirical Elution Profiles With Properties Mined From the Chemical Structure Database PubChem

The XlogP values of 27 standards measured under the same conditions were queried in PubChem (Table S8, Figure 4a). After removing $C_{26}H_{29}NO$ (tamoxiphen) and $C_{22}H_{24}N_2O_8$ (tetracycline) as outliers, there was a significant linear relationship between the XlogP value and the retention time ($XlogP = 0.58t_r - 3.72$, $F_{1,24} = 6.28$, R^2_{adj}, $p - value: < 0.001$, standard error: : (intercept) 0.92, (slope) 0.10) confirming good prediction of chromatographic behaviour.

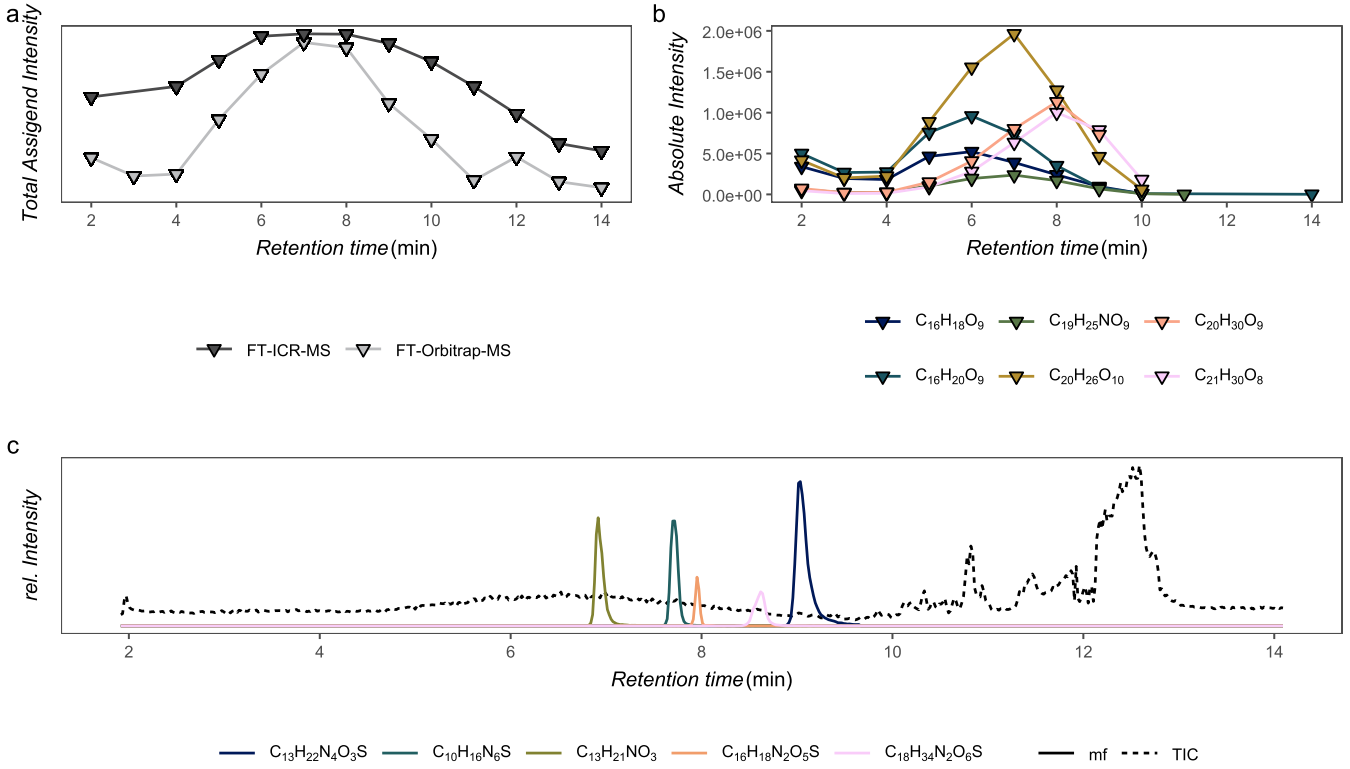

FIGURE 1 | (a) Total assigned intensity (TAI) of the North Sea water standard for LC-FT-ICR-MS and LC-FT-Orbitrap-MS. Scans were averaged into 1 min bins (triangles). (b) Extracted ion chromatograms of the LC-FT-Orbitrap-MS for typical DOM formulas had the characteristic broad elution peak (up to 6 min). Exemplary molecular formulas were chosen to represent different sections of the chromatogram (early ($C_{16}H_{18}O_9$, $C_{16}H_{20}O_9$, RT 4 to 6 min), mid ($C_{19}H_{25}NO_9$, $C_{20}H_{26}O_{10}$, RT 6 to 7 min) and late ($C_{20}H_{30}O_9$, $C_{21}H_{30}O_8$, RT 7 to 11 min) phase of the chromatogram. (c) The North Sea water standard was spiked with additional standard substances: ranitidine ($C_{13}H_{22}N_4O_3S$, 9.0 min), cimetidine ($C_{10}H_{16}N_6S$, 7.7 min), salbutamol ($C_{13}H_{21}NO_3$, 6.9 min), phenoxymethyl penicilin ($C_{16}H_{18}N_2O_5S$, 8.0 min) and lincomycin ($C_{18}H_{34}N_2O_6S$, 8.6 min). These standards were chosen by their elution maxima and their ionisation efficiency in negative ESI. Their concentration was adjusted so that each compound had an intensity maximum in a range of a typical DOM peak (absolute intensity $\approx 10^6$). The spiked standards had elutions windows of about 0.1 to 0.4 min. For a detailed list of all 25 spiked standards with the respective intensity and retention time, see Table S1. The mass tolerance of the extracted ion chromatograms was 1 ppm for the pharmaceutical standards and 0.8 ppm after external calibration for DOM molecular formulas.

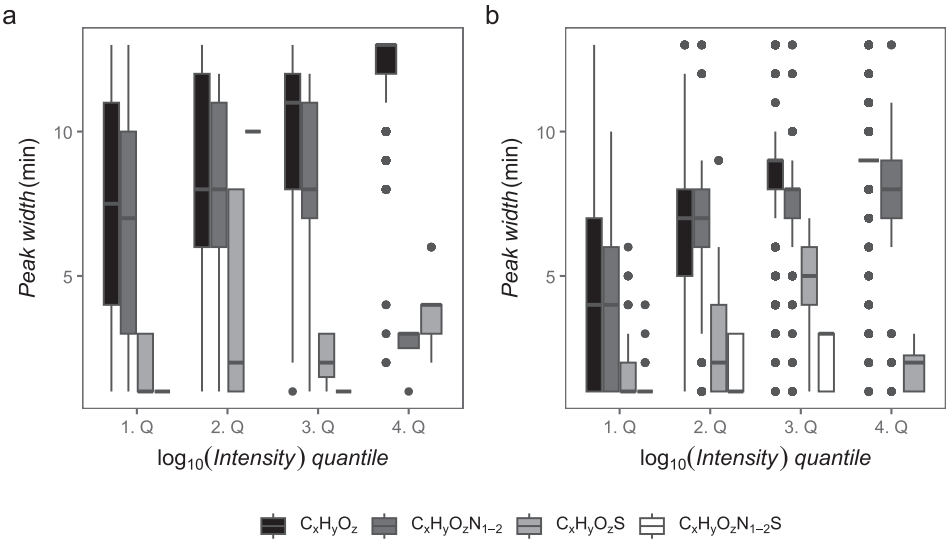

FIGURE 2 | (a) LC-FT-ICR-MS, (b) LC-FT-Orbitrap-MS. The chromatographic peak width (1 min bins) of all molecular formulas of the North Sea sample correlated with the number of heteroatoms and the signal intensity. (a) Intensity quantiles: 1. Q, $< 10^{6.46}$; 2. Q, $10^{6.46}$–$10^{6.75}$; 3. Q, $10^{6.75}$–$10^{7.13}$; 4. Q, $> 10^{7.13}$; (b) 1. Q, $< 10^{4.31}$; 2. Q, $10^{4.31}$–$10^{4.80}$; 3. Q, $10^{4.80}$–$10^{5.23}$; 4. Q, $> 10^{5.23}$. The shallow and deep Southern Ocean sample showed similar results (Figure S4).

The molecular formulas detected by LC-FT-ICR-MS were ranked by intensity, and the top 50 molecular formulas were selected (Tables S4 and S5). The same formulas were detected by LC-FT-Orbitrap-MS. Their elution maxima ranged from 2 to 12 min (median 7 min). Their mean molecular O/C ratio was 0.4799(0.0467), their mean molecular H/C ratio 1.2843(0.0925), their mean DBE was 8 (1) and their mean *m/z* 427 (33). The selected molecular formulas were queried on PubChem (https://pubchem.ncbi.nlm.nih.gov/) for structure-data file (sdf) information, which resulted in a dataset of 11 237 structural isomers after removing peroxides and adducts.

DOM molecules predicted to elute first (XlogP < −2) contained more rings (median 4 [2.5% quantile, 2; 97.5% quantile, 6; $n = 422$]), ketones (median 0 [0, 2, $n = 422$]) and alcohols (median 6 [3, 8, $n = 422$]) groups, while late eluting DOM molecules (XlogP > 3) were predicted to be more esterified (median 3 [0, 5, $n = 482$]) and aromatic (median 1 [0, 2, $n = 482$], Figure 3, Tables S6 and S7). The recursive feature elimination (random forest, 1000 trees, 5 times repeated cross validation, 5 folds) selected the following molecular properties and functional groups as most important to predict the XlogP (in this order): R-OH, aromatic structures, molecular weight, number of rings, R-O-R, RCOR (R^2 0.81, RMSE 0.63). Aldehydes and carboxylic groups contributed less to the prediction of XlogP because they did not vary along the elution gradient.

The chosen PubChem molecular formulas covered the complete elution gradient of DOM molecular formulas (Figures 3 and 4). Each of the 50 molecular formulas revealed a different PubChem-obtained isomeric assemblage behind these formulas (Figure S5). Three exemplary extracted ion chromatograms (EIC) with a good overlap of available PubChem isomers present in North Sea and Southern Ocean samples were analysed in detail ($C_{17}H_{22}O_8$, $C_{19}H_{22}O_{10}$, $C_{22}H_{28}O_{12}$) and showed shorter retention in both Southern Ocean compared to the North Sea water sample (Figure 4).

3.3 | Linear Correlation Between Empirical Retention and Theoretical logP for all Molecular Formulas

Based on the linear regression, retention times of all molecular formulas were converted to logP and average-weighted by their intensity, i.e. the chromatographic peak maximum had the highest weight. The correlation between them and the average XlogP value found in PubChem was tested for each sample and heteroatom group, with a high correlation indicating a good match to PubChem entries. The correlation was tested with a weighted least-square regression with the residuals as weights to reduce sensitivity to outliers. $C_xH_yO_z$ in the Southern Ocean samples had the highest correlation (slope: Shallow SO: 0.86, $p < 0.001$, Deep SO: 0.90, $p < 0.001$, North Sea: 0.48, $p < 0.001$). While $C_xH_yO_zN_{1-2}S$ and $C_xH_yO_zS$ molecular formulas had no significant linear relationships in SO samples, significant linear relationship was observed for $C_xH_yO_zN_{1-2}$ (slope: Shallow SO: 0. 57, $p = 0.01$, Deep SO: 0.56, $p = 0.03$, North Sea: 0.22, $p = 0.13$; Table S3). The North Sea sample had a significant correlation to all heteroatom types but with variation in strength ($C_xH_yO_z$ 0.48, $C_xH_yO_zN_{1-2}$-0.10, $C_xH_yO_zN_{1-2}S$ 0.17, $C_xH_yO_zN_{1-2}$S-0.11; all $p < 0.001$).

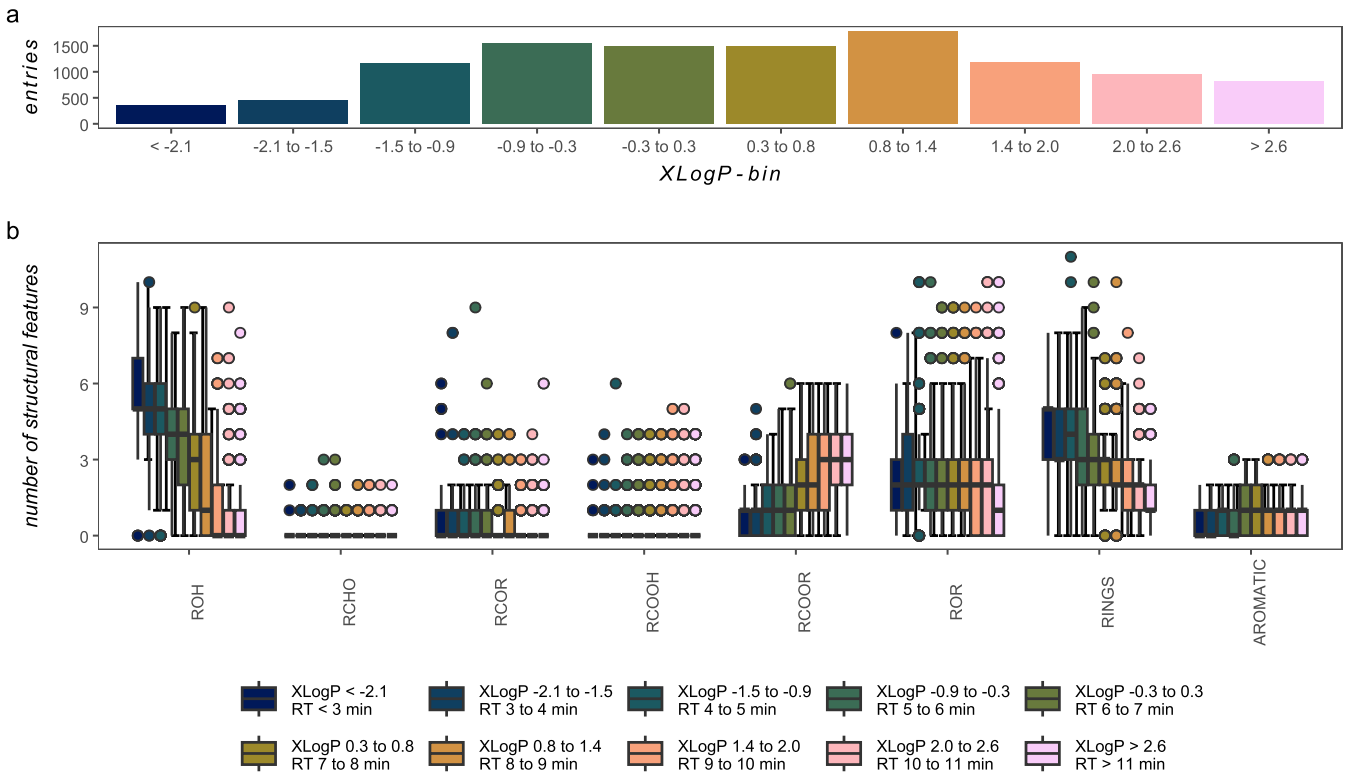

FIGURE 3 | Structural features of 50 selected molecular formulas in PubChem (11 237 structures). (a) Number of PubChem entries across the different XlogP bins. (b) Distribution of the number of structural features (points mark outlier) by XlogP bin. The legend shows the XlogP bins and the respective retention time window (for equation, see text).

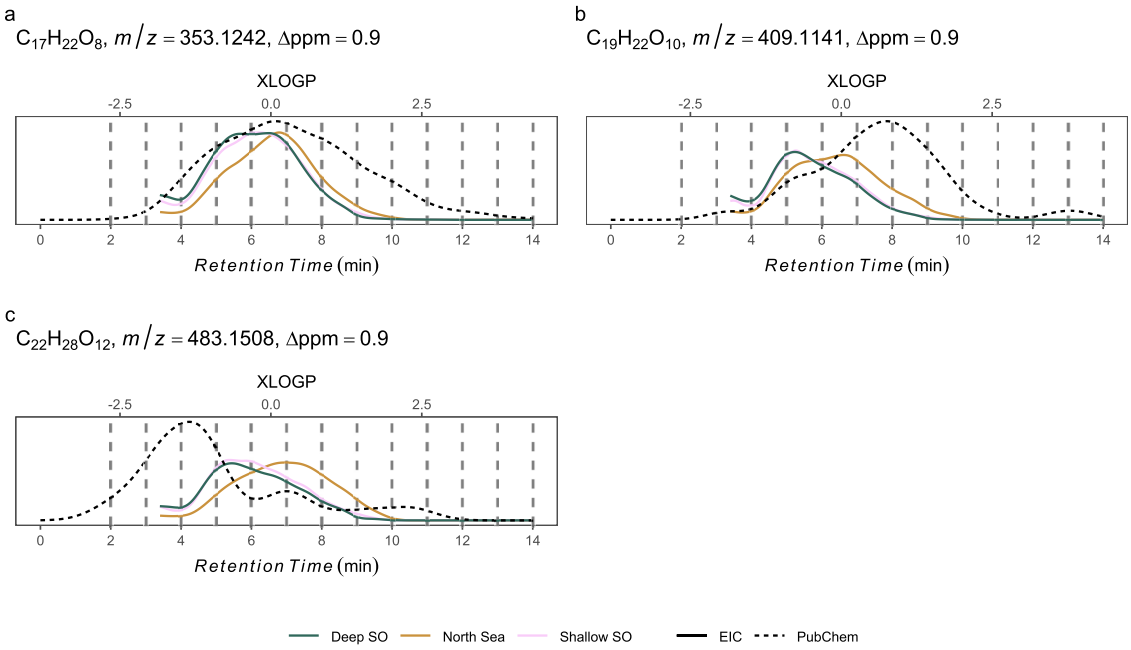

FIGURE 4 | Extracted ion chromatograms of three exemplary molecular formulas differed in retention depending on sample origin (solid lines). The isomeric assemblage behind the selected molecular formulas, as found in PubChem (dashed line), covered the DOM elution gradient of the EICs (LC-FT-Orbitrap-MS, spline smoothed, first 100 scans were cut-off). *m/z* is given for $[M-H]^-$.

4 | Discussion

To characterise DOM isomers in marine water extracts, we correlated their computed chromatographic retention with molecular structure features, such as their assignment with functional groups that largely determine the chromatographic retention. With a set of three exemplary DOM isomers, we assessed if the empirically determined elution properties of stoichiometrically identical isomers obtained from a chemical compound database would predict structural features, such as the type and number of functional groups, in structurally otherwise unknown DOM isomers.

To test this conceptual idea, we first verified if extracted ion chromatograms of internally spiked standards of known accurate mass indeed revealed sharp retention despite coelution among isomers within the DOM mixture. We screened 24 commercially available organic compounds that mimic the stoichiometry of DOM constituents in the range of $C_{16}H_{18\text{-}20}O_9$ (early elution), $C_{19\text{-}20}H_{25\text{-}26}N_{0\text{-}1}O_{9\text{–}10}$ (medium elution) 6 to 7 min) and $C_{20\text{-}21}H_{30}O_{8\text{–}9}$ (late elution). Spiking the North Sea DOM sample with 12 selected standards did indeed result in narrow peaks of each standard (Figure 1), supporting previous findings by Patriarca et al. [19]. Subsequently, the DOM molecular formulas were ranked by intensity and the first top 50 molecular formulas were selected and queried on PubChem for structure data resulting in ca. 12000 structural isomers including predicted XlogP values. The good linear relationship between XlogP values and retention time promised high predictive power. Among the identified molecular properties and functional groups, the presence of alcohols, aromatic structures, molecular weight, esters, rings and ketones correlated best with XlogP (rare feature elimination; Figure 3). Contrary, aldehydes and carboxylic groups did not support the correlation between XlogP and retention time, nor did these functional groups vary across the elution gradient. Irrespective of the analytical instrument used, the weighted average of predicted logP was a good linear predictor for the average XlogP of corresponding structures found on PubChem for $C_xH_yO_z$ molecular formulas in both Southern Ocean samples, whereas the opposite was found in the North Sea sample. Thus, the coverage of structure types in PubChem was better for the Southern Ocean than the North Sea samples.

The diagenetic imprint in any given DOM molecule co-occurring in water masses of different age or origin is caused by numerous reaction pathways that result in molecules of the same molecular formula but different decoration with functional groups [1,7]. Aged DOM molecules are more oxidised and smaller [9–11] and thus more polar, resulting in weaker chromatographic retention. The extracted ion chromatograms of three exemplary DOM isomers indeed confirmed our hypothesis that a DOM isomer belonging to comparatively older Southern Ocean water elutes earlier than its counterpart in young and less transformed North Sea water (Figure 4).

Deep Southern Ocean DOM mainly consists of aged and oxidised molecules. Due to its short retention on a hydrophobic LC column, the oxygen is likely present in oxygen-containing functional groups, such as alcohols (this study Figure 3, see also Arakawa et al. [10], Lam et al. [32]). This explanation is in analogy to prior studies demonstrating that weaker retained DOM molecules show high H_2O loss in MS^2 fragmentation experiments, indicative of high abundance of alcohol functions [1]. Conversely, the number of carboxylic groups along the DOM elution gradient, as indicated by neutral loss of CO_2 in MS^2, was shown to be rather constant and unaffected by retention time [1, 49]. These experimental observations strongly support our theoretical prediction (Figure 3). The comparatively longer retention of the same isomer in North Sea DOM thus suggests the

decoration with fewer alcohol groups and more ring structures and esters. This chemical difference between both isomers is likely explained by riverine input of terrestrial DOM into the Southern North Sea, where it mixes with freshly produced marine DOM. Terrestrial riverine DOM, with its high phenolic content, offers many precursor molecules whose dearomatisation and cycloaddition will form complex ring structures with oxygen stored in esters, anhydrides or lactones rather than alcohols [8].

The analysis of three marine DOM samples (North Sea, shallow and deep Southern Ocean) on two different LC-MS platforms revealed that chromatographic peak widths of extracted ion chromatograms correlated with the elementary composition of DOM molecular formulas, especially the number and types of heteroatoms (Figure 2). In terms of the elution window width, corresponding to the degree of chemodiversity of molecules in this time frame, the narrow and short elution of $C_xH_yO_zN_{1-2}S$ stoichiometries demonstrated this category to have the lowest chemodiversity. Since marine primary production predominately results in autochthonous organic matter with high C:N ratios [50], this is congruent with the observed high chemodiversity.

The intensity-weighted average of logP for the elementary composition $C_xH_yO_z$ in Southern Ocean samples, predicted by retention time of all molecular formulas (Figure 5), was in good agreement with XlogP values stored in PubChem (slope close to 1). For the other elementary compositions under investigation (Figure 5), the relationship was weaker, i.e. the slope was closer to zero or not significantly different from zero. This probably means that the average composition of structures stored in PubChem only partially covered the average measured structural composition of DOM molecules. This does not contradict or invalidate our conceptual approach of combining empirical retention time of DOM isomers to theoretical logP values, and in turn decoration with functional groups, because PubChem compounds cover the complete elution gradient (Figure 4). The explained variance ($R^2_{adj} = 0.23$) highlights the wide range of compounds projected on a molecular formula. We propose our approach to be advantageous over direct-injection mass-based library searches [51] as well as fragmentation-based annotation approaches [52–54], because our strategy circumvents issues of chimeric fragmentation spectra and furthermore adds retention time as a new physicochemical descriptor of DOM molecules. The combination of a constrained chemical structure feature space by retention time and fragmentation might facilitate targeted DOM fragmentation experiments, similar to Simon et al. [52].

Inter- and intra-platform comparisons of mass spectrometric data often yield variable results in the direct comparison of detected molecular formulas and their intensity [14]. Such platform-related differences are commonly assigned to different response factors, resulting from ionisation and ion transfer efficiencies prior to detection [57, 58]. However, multivariate comparisons have revealed comparable trends between instruments

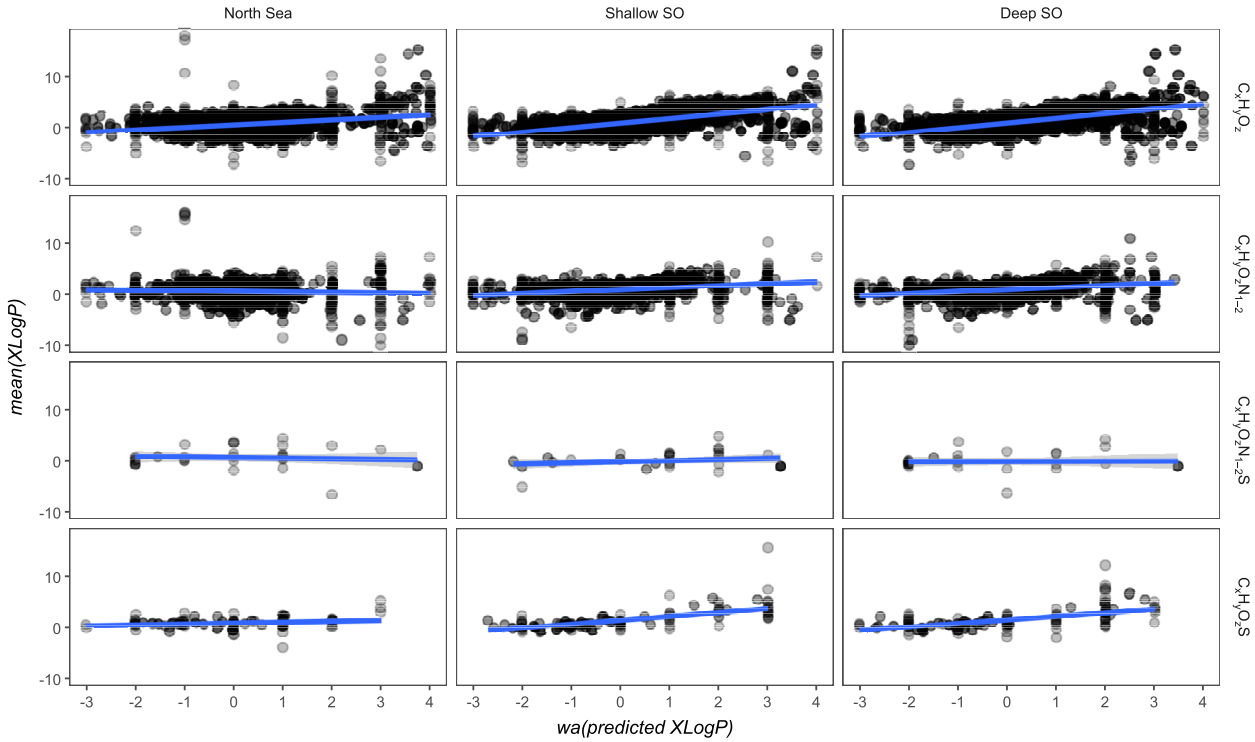

FIGURE 5 | Based on retention time, the logP value of each molecular formula was predicted and intensity-weight averaged. A linear model was constructed between these and the average XlogP value of all respective isomers available in PubChem. Depicted are all molecular formulas ($n = 5048$) detected by LC-FT-Orbitrap-MS. A perfect fit (slope = 1, intersect = 0, $R^2 \approx 1$) means that retention of the average DOM molecular formulas perfectly match the average XlogP value of PubChem structures. In this case, DOM structures would be predicted with high likelihood by structures of the same molecular formula stored in PubChem. The strength of the correlation varied between sample origin and heteroatom class. There were ca. 4 million XlogP entries for the LC-FT-ICR-MS molecular formulas and ca. 7.6 million XlogP entries for the LC-FT-Orbitrap-MS molecular formulas$\left(\text{residual weighted linear model}, R^2_{adj} = 0.23, p < 0.001, F_{23;46142} = 603.5\right)$, further statistics including the model for LC-FT-ICR-MS are summarised in Table S3.

on the multivariate level and reproduced trends (this study Figures S1 and S2) [37,55,56]. The data of this study are in good agreement with these results because assigned molecular formulas overlapped to ~31% between platforms. The percentage of the total intensity explained by the sum formulas found by both instruments was ~98% for LC-FT-ICR-MS and ~73% for LC-FT-Orbitrap-MS. Interestingly, FT-Orbitrap-MS detected more heteroatom-containing molecules (Figure 2), which may be due to platform-specific ionisation efficiency and/or a detection of more low intense signals (often heteroatom containing) due to different signal to noise cut-off.

We are aware that prediction and classification of DOM structural features by XlogP requires careful consideration, because PubChem contains known compounds with unknown similarity to DOM, and extracted PubChem data predict XlogP based on the uncharged molecule. Having measured the DOM samples in this study at pH 4–5, we can only assume that carboxylic acids were largely undissociated. Yet, as detailed structures of DOM constituents are unknown, the exact protonation and charge state cannot be estimated, thus introducing a bias of unknown magnitude since the adsorption equilibria between solid and mobile phase, especially for carboxylic rich molecules, is pH-dependent [26]. In this study, we only considered single negative charged ions. Future work will have to carefully examine the pH effect when predicting compound structure on the basis of calculated retention times.

5 | Conclusion

By correlating empirical retention time data with theoretical octanol–water partition coefficients, we were able to predict the occurrence of specific functional groups in otherwise structurally completely unknown DOM molecules. We compared the partition coefficient with structural information stored in PubChem and could thus indicate the type and abundance of distinct functional groups in DOM molecules depending on their retention time. We verified that less retained DOM isomers on a C18-LC column were likely to contain more alcohol groups and ring elements than stronger retained isomers, instead containing more esters and ethers. This approach circumvents the chimeric spectra of DOM MS fragmentation experiments. Using three exemplary DOM molecular formulas found in water samples of different origin and age, we demonstrated that the diagenetic state of these formulas aligned with changes in their retention time and thus with their decoration with different functional groups. We further demonstrated that DOM MS spectra obtained on two different LC-MS platforms, namely LC-FT-ICR-MS and LC-FT-Orbitrap-MS, were robust and suitable for this conceptual approach. We also showed that chromatographic peak width of EICs belonging to a distinct elementary composition of DOM molecules was a good measure of their underlying chemodiversity.

Author Contributions

Fabian Moye: writing – original draft, visualisation, formal analysis. **Marlo Barth:** investigation. **Boris Koch:** writing – review. **Jan Tebben, Tilmann Harder:** writing – review and editing, supervision.

Acknowledgements

This work was supported by a University of Bremen grant to TH to fund the PhD thesis of FM. MB was funded through the Helmholtz School for Marine Data Science (MarDATA), Grant No. HIDSS-0005. The authors thank Matthias Witt for the LC-FT-ICR-MS measurements. Open Access funding enabled and organized by Projekt DEAL.

Data Availability Statement

The binned peak list, and the annotated and filtered molecular formula list for both instruments have been submitted to the data publisher PANGAEA (https://pangaea.de/). As soon as curation and review are finished, the data will be made available.

Peer Review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/rcm.10043.

References

1. J. A. Hawkes, C. Patriarca, P. J. R. Sjöberg, L. J. Tranvik, and J. Bergquist, "Extreme Isomeric Complexity of Dissolved Organic Matter Found Across Aquatic Environments," *Limnology and Oceanography Letters* 3, no. 2 (2018): 21–30, https://doi.org/10.1002/lol2.10064.

2. M. Zark, J. Christoffers, and T. Dittmar, "Molecular Properties of Deep-Sea Dissolved Organic Matter Are Predictable by the Central Limit Theorem: Evidence From Tandem FT-ICR-MS," *Marine Chemistry* 191 (2017): 9–15, https://doi.org/10.1016/j.marchem.2017.02.005.

3. P. Verdugo and P. H. Santschi, "Polymer Dynamics of DOC Networks and Gel Formation in Seawater," *Deep Sea Research Part II: Topical Studies in Oceanography* 57, no. 16 (2010): 1486–1493, https://doi.org/10.1016/j.dsr2.2010.03.002.

4. B. P. Koch, G. Kattner, M. Witt, and U. Passow, "Molecular Insights Into the Microbial Formation of Marine Dissolved Organic Matter: Recalcitrant or Labile?," *Biogeosciences* 11, no. 15 (2014): 4173–4190, https://doi.org/10.5194/bg-11-4173-2014.

5. M. Seidel, P. L. Yager, N. D. Ward, et al., "Molecular-Level Changes of Dissolved Organic Matter Along the Amazon River-To-Ocean Continuum," *Marine Chemistry* 177 (2015): 218–231, https://doi.org/10.1016/j.marchem.2015.06.019.

6. N. Jiao, R. Cai, Q. Zheng, et al., "Unveiling the Enigma of Refractory Carbon in the Ocean," *National Science Review* 5, no. 4 (2018): 459–463, https://doi.org/10.1093/nsr/nwy020.

7. H. Osterholz, J. Niggemann, H. A. Giebel, M. Simon, and T. Dittmar, "Inefficient Microbial Production of Refractory Dissolved Organic Matter in the Ocean," *Nature Communications* 6, no. 1 (2015): 7422, https://doi.org/10.1038/ncomms8422.

8. S. Li, M. Harir, D. Bastviken, et al., "Dearomatization Drives Complexity Generation in Freshwater Organic Matter," *Nature* 628, no. 8009 (2024): 776–781, https://doi.org/10.1038/s41586-024-07210-9.

9. O. J. Lechtenfeld, G. Kattner, R. Flerus, S. L. McCallister, P. Schmitt-Kopplin, and B. P. Koch, "Molecular Transformation and Degradation of Refractory Dissolved Organic Matter in the Atlantic and Southern Ocean," *Geochimica et Cosmochimica Acta* 126 (2014): 321–337, https://doi.org/10.1016/j.gca.2013.11.009.

10. N. Arakawa, L. I. Aluwihare, A. J. Simpson, R. Soong, B. M. Stephens, and D. Lane-Coplen, "Carotenoids Are the Likely Precursor of a Significant Fraction of Marine Dissolved Organic Matter," *Science Advances* 3, no. 9 (2017): e1602976, https://doi.org/10.1126/sciadv.1602976.

11. R. Flerus, O. J. Lechtenfeld, B. P. Koch, et al., "A Molecular Perspective on the Ageing of Marine Dissolved Organic Matter," *Biogeosciences* 9, no. 6 (2012): 1935–1955, https://doi.org/10.5194/bg-9-1935-2012.

12. B. P. Koch, M. R. Witt, R. Engbrodt, T. Dittmar, and G. Kattner, "Molecular Formulae of Marine and Terrigenous Dissolved Organic Matter Detected by Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry," *Geochimica et Cosmochimica Acta* 69, no. 13 (2005): 3299–3308, https://doi.org/10.1016/j.gca.2005.02.027.

13. N. Cortés-Francisco and J. Caixach, "Molecular Characterization of Dissolved Organic Matter Through a Desalination Process by High Resolution Mass Spectrometry," *Environmental Science & Technology* 47, no. 17 (2013): 9619–9627, https://doi.org/10.1021/es4000388.

14. J. A. Hawkes, T. Dittmar, C. Patriarca, L. Tranvik, and J. Bergquist, "Evaluation of the Orbitrap Mass Spectrometer for the Molecular Fingerprinting Analysis of Natural Dissolved Organic Matter," *Analytical Chemistry* 88, no. 15 (2016): 7698–7704, https://doi.org/10.1021/acs.analchem.6b01624.

15. L. Han, J. Kaesler, C. Peng, T. Reemtsma, and O. J. Lechtenfeld, "Online Counter Gradient LC-FT-ICR-MS Enables Detection of Highly Polar Natural Organic Matter Fractions," *Analytical Chemistry* 93, no. 3 (2021): 1740–1748, https://doi.org/10.1021/acs.analchem.0c04426.

16. C. Patriarca, A. Balderrama, M. Može, et al., "Investigating the Ionization of Dissolved Organic Matter by Electrospray," *Analytical Chemistry* 92, no. 20 (2020): 14210–14218, https://doi.org/10.1021/acs.analchem.0c03438.

17. R. Rodrigues Matos, E. K. Jennings, J. Kaesler, T. Reemtsma, B. P. Koch, and O. J. Lechtenfeld, "Post Column Infusion of an Internal Standard Into LC-FT-ICR MS Enables Semi-Quantitative Comparison of Dissolved Organic Matter in Original Samples," *Analyst* 149 (2024): 3468–3478, https://doi.org/10.1039/D4AN00119B.

18. Y. Li, C. He, Y. Zhang, and Q. Shi, "Online LC-Orbitrap MS Method for the Rapid Molecular Characterization of Dissolved Organic Matter," *Rapid Communications in Mass Spectrometry* 38, no. 20 (2024): e9885, https://doi.org/10.1002/rcm.9885.

19. C. Patriarca, J. Bergquist, P. J. R. Sjoberg, L. Tranvik, and J. A. Hawkes, "Online HPLC-ESI-HRMS Method for the Analysis and Comparison of Different Dissolved Organic Matter Samples," *Environmental Science & Technology* 52, no. 4 (2018): 2091–2099, https://doi.org/10.1021/acs.est.7b04508.

20. D. E. LaRowe and P. Van Cappellen, "Degradation of Natural Organic Matter: A Thermodynamic Analysis," *Geochimica et Cosmochimica Acta* 75, no. 8 (2011): 2030–2042, https://doi.org/10.1016/j.gca.2011.01.020.

21. B. P. Koch and T. Dittmar, "From Mass to Structure: An Aromaticity Index for High-Resolution Mass Data of Natural Organic Matter," *Rapid Communications in Mass Spectrometry* 20, no. 5 (2006): 926–932, https://doi.org/10.1002/rcm.2386.

22. B. P. Koch and T. Dittmar, "From Mass to Structure: An Aromaticity Index for High-Resolution Mass Data of Natural Organic Matter (Vol 20, pg 926, 2006)," *Rapid Communications in Mass Spectrometry* 30, no. 1 (2016): 250–250, https://doi.org/10.1002/rcm.7433.

23. T. Dittmar and A. Stubbins, "12.6 - Dissolved Organic Matter in Aquatic Systems," in *Treatise on Geochemistry*, Second ed., eds. H. D. Holland and K. K. Turekian (Oxford: Elsevier, 2014): 125–156, https://doi.org/10.1016/B978-0-08-095975-7.01010-X.

24. M. Seidel, S. P. B. Vemulapalli, D. Mathieu, and T. Dittmar, "Marine Dissolved Organic Matter Shares Thousands of Molecular Formulae Yet Differs Structurally Across Major Water Masses," *Environmental Science & Technology* 56, no. 6 (2022): 3758–3769, https://doi.org/10.1021/acs.est.1c04566.

25. J. K. Geuer, B. Krock, T. Leefmann, and B. P. Koch, "Quantification, Extractability and Stability of Dissolved Domoic Acid Within Marine Dissolved Organic Matter," *Marine Chemistry* 215 (2019): 103669, https://doi.org/10.1016/j.marchem.2019.103669.

26. S. Papadopoulos Lambidis, T. Schramm, K. Steuer-Lodd, et al., "Two-Dimensional Liquid Chromatography Tandem Mass Spectrometry Untangles the Deep Metabolome of Marine Dissolved Organic Matter," *Environmental Science & Technology* 58 (2024): acs.est.4c07173, https://doi.org/10.1021/acs.est.4c07173.

27. K. Namjesnik-Dejanovic and S. E. Cabaniss, "Reverse-Phase HPLC Method for Measuring Polarity Distributions of Natural Organic Matter," *Environmental Science & Technology* 38, no. 4 (2004): 1108–1114, https://doi.org/10.1021/es0344157.

28. M. S. Mirrlees, S. J. Moulton, C. T. Murphy, and P. J. Taylor, "Direct Measurement of Octanol-Water Partition Coefficients by High-Pressure Liquid Chromatography," *Journal of Medicinal Chemistry* 19, no. 5 (1976): 615–619, https://doi.org/10.1021/jm00227a008.

29. P. Bonini, T. Kind, H. Tsugawa, D. K. Barupal, and O. Fiehn, "Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics," *Analytical Chemistry* 92, no. 11 (2020): 7515–7522, https://doi.org/10.1021/acs.analchem.9b05765.

30. M. Cao, K. Fraser, J. Huege, T. Featonby, S. Rasmussen, and C. Jones, "Predicting Retention Time in Hydrophilic Interaction Liquid Chromatography Mass Spectrometry and Its use for Peak Annotation in Metabolomics," *Metabolomics* 11, no. 3 (2015): 696–706, https://doi.org/10.1007/s11306-014-0727-x.

31. T. Cheng, Y. Zhao, X. Li, et al., "Computation of Octanol-Water Partition Coefficients by Guiding an Additive Model With Knowledge," *Journal of Chemical Information and Modeling* 47, no. 6 (2007): 2140–2148, https://doi.org/10.1021/ci700257y.

32. B. Lam, A. Baer, M. Alaee, et al., "Major Structural Components in Freshwater Dissolved Organic Matter," *Environmental Science & Technology* 41, no. 24 (2007): 8240–8247, https://doi.org/10.1021/es0713072.

33. N. Jiao, T. Luo, Q. Chen, et al., "The Microbial Carbon Pump and Climate Change," *Nature Reviews. Microbiology* 22, no. 7 (2024): 408–419, https://doi.org/10.1038/s41579-024-01018-0.

34. B. P. Koch, K. U. Ludwichowski, G. Kattner, T. Dittmar, and M. Witt, "Advanced Characterization of Marine Dissolved Organic Matter by Combining Reversed-Phase Liquid Chromatography and FT-ICR-MS," *Marine Chemistry* 111, no. 3–4 (2008): 233–241, https://doi.org/10.1016/j.marchem.2008.05.008.

35. S. E. D. El Naggar, G. Dieckmann, C. Haas, M. Schröder, and M. Spindler, *The Expeditions ANTARKTIS-XXII/1 and XII/2 of the Research Vessel Polarstern in 2004/2005* (Bremerhaven, Germany: Alfred-Wegener-Institut für Polar- und Meeresforschung, 2007): 1618–3193.

36. R. Flerus, B. P. Koch, P. Schmitt-Kopplin, M. Witt, and G. Kattner, "Molecular Level Investigation of Reactions Between Dissolved Organic Matter and Extraction Solvents Using FT-ICR MS," *Marine Chemistry* 124, no. 1–4 (2011): 100–107, https://doi.org/10.1016/j.marchem.2010.12.006.

37. C. Simon, V.-N. Roth, T. Dittmar, and G. Gleixner, "Molecular Signals of Heterogeneous Terrestrial Environments Identified in Dissolved Organic Matter: A Comparative Analysis of Orbitrap and Ion Cyclotron Resonance Mass Spectrometers," *Frontiers in Earth Science* 6 (2018): 138, https://doi.org/10.3389/feart.2018.00138.

38. T. Leefmann, S. Frickenhaus, and B. P. Koch, "UltraMassExplorer: A Browser-Based Application for the Evaluation of High-Resolution Mass Spectrometric Data," *Rapid Communications in Mass Spectrometry* 33, no. 2 (2019): 193–202, https://doi.org/10.1002/rcm.8315.

39. P. M. Medeiros, M. Seidel, J. Niggemann, et al., "A Novel Molecular Approach for Tracing Terrigenous Dissolved Organic Matter Into the Deep Ocean," *Global Biogeochemical Cycles* 30, no. 5 (2016): 689–699, https://doi.org/10.1002/2015gb005320.

40. P. Herzsprung, N. Hertkorn, W. von Tumpling, M. Harir, K. Friese, and P. Schmitt-Kopplin, "Understanding Molecular Formula Assignment of Fourier Transform ion Cyclotron Resonance Mass Spectrometry Data of Natural Organic Matter From a Chemical Point of View," *Analytical and Bioanalytical Chemistry* 406, no. 30 (2014): 7977–7987, https://doi.org/10.1007/s00216-014-8249-y.

41. O. J. Lechtenfeld, B. P. Koch, B. Gasparovic, S. Frka, M. Witt, and G. Kattner, "The Influence of Salinity on the Molecular and Optical Properties of Surface Microlayers in a Karstic Estuary," *Marine Chemistry* 150 (2013): 25–38, https://doi.org/10.1016/j.marchem.2013.01.006.

42. M. C. Chambers, B. Maclean, R. Burke, et al., "A Cross-Platform Toolkit for Mass Spectrometry and Proteomics," *Nature Biotechnology* 30, no. 10 (2012): 918–920, https://doi.org/10.1038/nbt.2377.

43. J. Rainer and A. Vicini, "*SpectraQL: MassQL Support for Spectra*," R package version 1.0.0 (2024), https://github.com/RforMassSpectrometry/SpectraQL.

44. R Core Team, *R: A Language and Environment for Statistical Computing [computer program]*. Version 4.2.3 (R Foundation for Statistical Computing, 2023).

45. Y. Cao, A. Charisi, L.-C. Cheng, T. Jiang, and T. Girke, "ChemmineR: A Compound Mining Framework for R," *Bioinformatics* 24, no. 15 (2008): 1733–1734, https://doi.org/10.1093/bioinformatics/btn307.

46. M. Kuhn, "Building Predictive Models in R Using the Caret Package," *Journal of Statistical Software* 28, no. 5 (2008): 1–26, https://doi.org/10.18637/jss.v028.i05.

47. M. N. Wright and A. Ziegler, "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R," *Journal of Statistical Software* 77, no. 1 (2017): 1–17, https://doi.org/10.18637/jss.v077.i01.

48. M. Greenacre, "Data Reporting and Visualization in Ecology," *Polar Biology* 39, no. 11 (2016): 2189–2205, https://doi.org/10.1007/s00300-016-2047-2.

49. M. Witt, J. Fuchser, and B. P. Koch, "Fragmentation Studies of Fulvic Acids Using Collision Induced Dissociation Fourier Transform Ion Cyclotron Resonance Mass Spectrometry," *Analytical Chemistry* 81, no. 7 (2009): 2688–2694, https://doi.org/10.1021/ac802624s.

50. C. S. Hopkinson and J. J. Vallino, "Efficient Export of Carbon to the Deep Ocean Through Dissolved Organic Matter," *Nature* 433, no. 7022 (2005): 142–145, https://doi.org/10.1038/nature03191.

51. S. A. Benk, Y. Li, V.-N. Roth, and G. Gleixner, "Lignin Dimers as Potential Markers for 14C-Young Terrestrial Dissolved Organic Matter in the Critical Zone," *Frontiers in Earth Science* 6 (2018): 168, https://doi.org/10.3389/feart.2018.00168.

52. C. Simon, K. Dührkop, D. Petras, et al., "Mass Difference Matching Unfolds Hidden Molecular Structures of Dissolved Organic Matter," *Environmental Science & Technology* 56, no. 15 (2022): 11027–11040, https://doi.org/10.1021/acs.est.2c01332.

53. D. Petras, I. Koester, R. Da Silva, et al., "High-Resolution Liquid Chromatography Tandem Mass Spectrometry Enables Large Scale Molecular Characterization of Dissolved Organic Matter," *Frontiers in Marine Science* 4 (2017): 405, https://doi.org/10.3389/fmars.2017.00405.

54. J. Patrone, M. Vila-Costa, J. Dachs, S. Papazian, P. Gago-Ferrero, and R. Gil-Solsona, "Enhancing Molecular Characterization of Dissolved Organic Matter by Integrative Direct Infusion and Liquid Chromatography Nontargeted Workflows," *Environmental Science & Technology* 58 (2024): 12454–12466, https://doi.org/10.1021/acs.est.4c00876.

55. T. N. Clark, J. Houriet, W. S. Vidar, et al., "Interlaboratory Comparison of Untargeted Mass Spectrometry Data Uncovers Underlying Causes for Variability," *Journal of Natural Products* 84 (2021): 824–835, https://doi.org/10.1021/acs.jnatprod.0c01376.

56. A. Zherebker, S. Kim, P. Schmitt-Kopplin, et al., "Interlaboratory Comparison of Humic Substances Compositional Space as Measured by Fourier Transform ion Cyclotron Resonance Mass Spectrometry (IUPAC Technical Report)," *Pure and Applied Chemistry* 92, no. 9 (2020): 1447–1467, https://doi.org/10.1515/pac-2019-0809.

57. Q. Pan, X. Zhuo, C. He, Y. Zhang, and Q. Shi, "Validation and Evaluation of High-Resolution Orbitrap Mass Spectrometry on Molecular Characterization of Dissolved Organic Matter," *ACS Omega* 5, no. 10 (2020): 5372–5379, https://doi.org/10.1021/acsomega.9b04411.

58. A. J. Craig, M. A. Ganiyu, L. W. K. Moodie, et al., "Improvement of Electrospray Ionization Response Linearity and Quantification in Dissolved Organic Matter Using Synthetic Deuterated Internal Standards," *ChemRxiv* (2025), https://doi.org/10.26434/chemrxiv-2025-6wm8s.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.