Mass difference matching unfolds hidden molecular structures of dissolved organic matter

Carsten Simon¹,†, Kai Dührkop², Daniel Petras³,⁴, Vanessa-Nina Roth¹,§, Sebastian Böcker², Pieter C. Dorrestein³, and Gerd Gleixner¹,*

¹Molecular Biogeochemistry, Department of Biogeochemical Processes, Max Planck Institute for Biogeochemistry, Hans-Knöll-Straße 10, 07745 Jena, Germany
²Chair for Bioinformatics, Friedrich-Schiller-University, Jena, Germany
³Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, San Diego, CA, USA
⁴CMFI Cluster of Excellence, Interfaculty Institute of Microbiology and Medicine, University of Tübingen, 72076, Tübingen, Germany

TOC FIGURE:

ABSTRACT: Ultrahigh-resolution Fourier transform mass spectrometry (FTMS) has revealed unprecedented detail of natural complex mixtures such as dissolved organic matter (DOM) on a molecular formula level, but we lack approaches to access the underlying structural complexity. We here explore the hypothesis that every DOM precursor is potentially linked with all emerging product ions in FTMS² experiments. The resulting mass difference (Δm) matrix is deconvoluted to isolate individual precursor Δm profiles and matched with structural information, which was derived from 42 Δm features from 14 in-house reference compounds and a global set of 11477 Δm features with assigned structure specificities, using a dataset of ~18000 unique structures. We show that Δm matching is highly sensitive in predicting potential precursor identities in terms of molecular and structural composition. Additionally, the approach identified unresolved precursors and missing elements in molecular formula annotation (P, Cl, F). Our study provides first results how Δm matching improves structural domains in Van Krevelen space, but simultaneously demonstrates the wide overlap between the structural domains. We show that this effect is likely driven by chemodiversity and offers an explanation for the observed ubiquitous presence of molecules in the center
of the Van Krevelen space. Our promising first results suggest that Δm matching can unfold the structural information encrypted in DOM and assess the quality of FTMS-derived molecular formulas of complex mixtures in general.

---

**Synopsis:** We present an approach to deconvolute and explore the structural composition of co-fragmented mixtures of organic molecules in environmental media.

**Keywords:** Natural organic matter, NOM, DI-ESI-MS/MS, FTMS, Orbitrap, tandem mass spectrometry, MS/MS, deconvolution

---

1. **INTRODUCTION**

Complex mixtures are key study objects in environmental and industrial applications, but their analysis remains challenging. One of the most complex mixtures in natural ecosystems is dissolved organic matter (DOM). DOM is a central intermediate of ecosystem metabolism and mirrors molecular imprints of interactions with its abiotic and biotic environment, which form the basis for processes such as carbon sequestration and nutrient recycling. Despite significant advances in ultrahigh-resolution mass spectrometry (FTMS) and nuclear magnetic resonance spectroscopy, scientists still struggle to decode this information on the molecular level, and novel approaches to identify distinct structures are required to translate molecular-level information into improved process understanding.

Open and living systems promote the formation of ultra-complex mixtures of thousands to millions of individual constituents that mirror large environmental gradients. As a consequence, DOM poses significant challenges to separation, isolation, and structure elucidation. Direct infusion (DI) FTMS techniques have become indispensable tools for the molecular-level analysis of DOM as they reveal unprecedented detail of molecular formulas using the exact mass (MS) data, even without prior separation. However, FTMS techniques are selective and do not resolve all structural detail observed at
the exact mass in DOM, as the presence of isobars and isomers hinders the identification of particular structures from these molecular formulas. Additionally, current structural databases cover only a small fraction of molecular formulas encountered, and typically lead to annotation rates < 5%.

One way to obtain structure information on isomers and isobars is through collision-induced dissociation (CID) in fragmentation experiments (MS², or multistage MSⁿ). The relatively wide isolation window (~1 Da) of mass filters applied for precursor selection commonly hinders the isolation and subsequent fragmentation of single exact masses, leading to mixed "chimeric" MS² spectra of co-fragmented precursors. Even though some authors have achieved isolation of single masses or improved description of chimeric tandem MS data, most studies have pointed out that fragmentation patterns were rather universal across DOM samples. Most of these studies, however, focused on the major product ion peaks (fragments), which usually make up only 60 – 70 % of the total product ion abundance, and thus disregard many low-abundance signals that may be more suitable to detect structural differences.

The major product ions encountered in tandem mass spectra of DOM relate to sequential neutral losses of common small building blocks, mainly CO₂, H₂O, or CO units. A mass difference between a precursor and a product ion in an MS² spectrum is herein called "delta mass" and referred to as Δm (plural Δm's). Many Δm's such as CO₂ or H₂O are commonly observed and are thus deemed non-indicative for the identification of structural units. In contrast, other studies found recurring low m/z product ions (e.g., at m/z 95, 97, 109, 111, 123, 125, 137, 139, 151, and 153) that were interpreted as a limited set of core structural units substituted with a set of functional groups, yet in different amounts and configurational types that would lead to highly diverse mixtures. From a stochastic standpoint, the occurrence of common neutral losses may not be surprising; for example, many structures contain hydroxyl groups that could yield H₂O losses, and CO₂ could originate from ubiquitous carboxyl groups. In contrast, the occurrence of two molecules sharing a larger substructure would be less probable, and thus less easily detected as a major peak. Signatures of DOM's structural diversity could thus prevail in the high number of low-abundance fragments usually detected below m/z 200-300, as opposed to the higher abundance of fragments connected
to losses of small substructures such as CO$_2$ or H$_2$O. Given the large number of estimated isomers and isobars underlying usual DOM data$^{18,19,31,32,39,45–48}$, we here build upon the hypothesis that every co-fragmented precursor potentially contributes to every emerging product ion signal. We interpret the resulting chimeric MS$^2$ data as a structural fingerprint that can be deconvoluted to obtain individual precursor $\Delta m$ matching profiles. The analysis of $\Delta m$'s that link precursor and product ions, in contrast to indicative product ions (fragments) alone, is independent of the masses of the unknown precursors and known reference compounds in databases of annotated $\Delta m$ features. Although this approach will sacrifice the identification of true knowns, it allows for the identification of potential structural analogs via indicative $\Delta m$’s and is suited best when annotation rates are as low as in the case of DOM, i.e., when most compounds are yet unknown.$^{18,26,27}$

Despite the unknown identity of most of the molecules present in DOM, its potential sources can be constrained reasonably well. Plants produce most of the organic matter that sustains heterotrophic food webs in natural ecosystems. Plant metabolites such as polyphenols and polyaromatic structures thus represent a major source of DOM. Therefore, an early decomposition phase likely exists when the imprint of soluble/solubilized plant metabolites is still detectable by MS$^2$ experiments using current FTMS technology. For example, lignin-related compounds show indicative methoxyl and methyl radical losses,$^{18,49,50}$ glycosides indicate the loss of a sugar unit$^{51,52}$ and hydrolyzable tannins are expected to lose galloyl units.$^{52}$ Even related compounds such as flavon-3-ols and flavan-3-ols could potentially be distinguished by their indicative retro-cyclization products.$^{51,53}$ Mass differences related to atoms such as N, S, P, Cl, Br, I and F could also help to identify unknown organic nutrient species or disinfection byproducts, thereby widening the applicability of the approach.$^{1,54}$ Lastly, indicative $\Delta m$ fingerprints could provide constraints to putative compound group annotations derived from molecular formula data alone (Van Krevelen diagrams), or allow for a more precise annotation.$^{55–57}$

We hypothesized that DOM from swamps and topsoil, in close contact to plant inputs and active microbial communities, would reflect recognizable plant-related source imprints that can be revealed by Orbitrap
tandem mass spectrometry. Specifically, we explored links between precursor Δm matching profiles and precursor characteristics such as nominal mass, mass defect, initial ion abundance, fragmentation sensitivity, oxygen-to-hydrogen ratio (O/C), heteroatom content, and structure suggestions. These properties are in part predictable from the assigned molecular formula, and thus allow for an evaluation of the approach ("proof-of-concept") while also revealing potential non-assigned molecules of special interest (e.g., P-, Cl-, Br-, I- and F-containing molecular formulas). Lastly, we hypothesized that indicative Δm features of plant phenols, e.g., lignin- and tannin-related losses, would match their yet unknown structural analogs in DOM and that these patterns would reflect commonly applied structural domain distributions.56,58

2. EXPERIMENTAL SECTION

A detailed experimental procedure is provided in the Supplemental Information of this article (Note S-1). In short, we chose a set of 14 aromatic reference compounds as representative plant metabolites in DOM (Figure S-1, Table S-1) and a forest topsoil pore water isolate59 and Suwannee River Natural Organic Matter (SRNOM)60 as exemplary DOM samples. All reference and sample solutions were directly infused into the ESI (electrospray) source of an Orbitrap Elite (Thermo Fisher Scientific, Bremen) at negative ionization mode (Table S-2) and fragmented by collision-induced dissociation (CID, MS²). We chose four nominal masses within the average mass range typically observed in terrestrial DOM samples (m/z 200 – 500) for fragmentation (m/z 241, 301, 361, and 417, herein referred to as isolated precursor ion mixtures, “IPIMs”) as a first set of data to test the approach.61 Soil DOM was analyzed at three normalized collision energy (NCE) levels (15, 20, and 25%). MS³ spectra of selected key product ions (aglycons of flavonoids and demethylated dimethoxy-methyl-benzoquinone) were acquired as well at NCE 20 or 25. After recalibration with known (Table S-3) or predicted product ions (losses of CO₂, H₂O, etc.), all major product ions were annotated with a molecular formula in reference compounds (Figure S-2, Table S-4, Table S-5) and DOM. Formula annotation was conducted with a Matlab routine recently incorporated into an open FTMS data processing pipeline.62
For MS$^2$ data analysis, we generated $\Delta m$ matrices of every pairwise combination of precursor and product ions (“$\Delta m$ fingerprints”). Every value in this matrix is referred to as a $\Delta m$ feature or simply $\Delta m$. We compared the unknown $\Delta m$ features in DOM to three lists of known $\Delta m$ features:

a) 54 $\Delta m$ features ubiquitously found in DOM (Table S-6),
b) 55 $\Delta m$ features from the set of 14 reference compounds (Table S-7), and
c) 11477 $\Delta m$ features from a negative ESI MS$^2$ library with 249916 reference spectra of 17994 unique molecular structures annotated by SIRIUS$^{63}$ (list available in the supporting datasets). Reference spectra were collected from from GNPS, MassBank, MoNA, and NIST.$^{64,65}$

The detection of a known $\Delta m$ feature in DOM is herein called “$\Delta m$ matching”, and detected $\Delta m$ features are called $\Delta m$ matches. Matching was conducted at a mass tolerance of ± 0.0002 Da (2 ppm at 200 Da). The array of $\Delta m$ matches of a single precursor is called the $\Delta m$ matching profile, and all precursor profiles of an IPIM form the subset of matched $\Delta m$’s of the $\Delta m$ matrix introduced above. The decomposition of the MS$^2$ spectrum into a $\Delta m$ matrix and therefore, individual $\Delta m$ matching profiles is what we define as the deconvolution step in this study. $\Delta m$’s of the literature- and reference-compound derived lists showed some overlap and were largely part of the SIRIUS list as well (see details in SI). The specificity of any $\Delta m$ feature in the SIRIUS list was checked by their association to compound classes as defined by ClassyFire.$^{66}$ The top 15 significantly associated classes were then obtained for each $\Delta m$ feature in list c) and included into analyses using the reference-compound derived list (list b) as well.

We assessed the probability of false positive matches and accounted for molecular formula constraints (numbers of elements in the formula), ion abundance and measures of fragmentation sensitivity to validate our approach. The matching data was combined for each NCE level and transformed into a binary format. We classified $\Delta m$ matching profiles of DOM precursors and reference compounds by two-way hierarchical clustering using Ward's method and Euclidean distance, as well as Principal Components Analysis (PCA) in PAST (v3.10).$^{67}$ We visualized numbers of individual $\Delta m$ matches and $\Delta m$ cluster matches in Van Krevelen space to analyze patterns in $\Delta m$ matching frequency (“structural domains”). We chose the structural domains reprinted in the 2014 review by Minor et al. for reference, because this represents the
general level of detail and type of classes distinguished in recent DOM studies (Figure S-3). In a separate analysis, lignin-like and N- and S-containing formulas were also classified with a more general Van Krevelen scheme besides the reference one.

Finally, we assessed the agreement between structures predicted by Δm matching and those suggested in natural product structural databases. We combined structure suggestions from different databases, including Dictionary of Natural Products, KNApSAcK, Metacyc, KEGG, and HMDB as well as their expanded in-silico annotations based on predicted enzymatic transformations in the MINEs database. Although the MINEs database covers 198 generalized chemical reaction rules it may not include all potential environmental reactions because those are not necessarily only driven by enzymes. The InChi-Key of structures was used to exclude stereoisomers and classify suggested structures into compound classes by ClassyFire.

3. RESULTS AND DISCUSSION

3.1. Tandem MS fragmentation of reference compounds and construction of Δm lists. The 14 aromatic reference compounds (Figures S-1, S-2 and S-3) yielded 42 new Δm features (i.e., not covered in the list of common Δm’s, Table S-6) but also eight that were described in DOM. These eight Δm features (namely: H$_2$O, 18.0106; CO, 27.9949; C$_2$H$_4$, 28.0313; C$_2$H$_2$O, 42.0106; CO$_2$, 43.9898; CH$_2$O$_3$, 62.0004; C$_2$O$_3$, 71.9847; and C$_3$O$_5$, 115.9746) were kept in the list to compare DOM and reference compounds (Table S-7). Besides precursor formulas #2 (Hydroxy-cinnamic acid, or p-coumaric acid; C$_9$H$_8$O$_3$, 164.0473), #3 (Gallic acid; C$_7$H$_6$O$_5$, 170.0215) and #5 (m-Guaiaicol; C$_7$H$_8$O$_2$, 124.0524), which were found among the 42 Δm’s as potential structural equivalents, five Δm’s of potential substructures likely to be found in DOM were added to the list, namely the ones of precursors #1 (Vanillic acid; C$_8$H$_6$O$_4$, 168.04226), #4 (Creosol, C$_8$H$_{10}$O$_2$, 138.0681), #8 (Ellagic acid; C$_{14}$H$_{16}$O$_6$, 302.0063) and #10 (Catechin; C$_{15}$H$_{14}$O$_6$, 290.0790), and the neutral aglycon of compounds #12 and #13 (flavonol core of Spiraeoside and Isoquercitin; C$_{15}$H$_{10}$O$_7$, 302.0427). More details on reference compound fragmentation are given in the SI (Note S-2).

3.2. Fragmentation behavior of soil DOM. DOM precursors were isolated and fragmented to obtain Δm data (Figure S-4). To find the best collision energy to fragment DOM, we analyzed soil DOM at three
NCE levels (15, 20 and 25). All IPIMs showed similar fragmentation properties (Note S-3, Table S-8). Highest numbers of product ions were found at the highest NCE (Figure S-5). Product ion spectra did not indicate abrupt structural changes upon increasing fragmentation energy, showing no separation of isomers/isobars but a continuous increase in fragmentation across all precursors. Based on the above results, NCE of 25 was chosen to fragment SRNOM as a second DOM sample for comparison.

Figure 1. Links between selected DOM precursor properties (upper panels, initial ion abundance at NCE 0; mid panels, half-life normalized collision energy (NCE) at which ion abundance has dropped by 50%; lower panels, matches of delta masses (Δm’s) of measured precursor and product ion masses (delta masses, Δm) with a list of 11477 known Δm features from SIRIUS) and each precursor’s (a, b, c) O/C ratio or (d, e, f) mass defect. O/C ratios can only be shown for precursors with an annotated molecular formula. Additional data from reference compounds (red diamonds, see also Figure S-3) and SRNOM (orange crosses) is shown in mid and lower panels, respectively. Statistical data was derived from linear fits; asterisks (*** ) denote p-value < 0.001.

Despite common differences between precursor ion abundance and O/C ratio or mass defect (Figure 1a, d), we found a significant positive link between both metrics and fragmentation sensitivity independent of nominal mass, ranging from half-life NCE (i.e., the NCE level causing 50% decrease in ion abundance) of 10 – 35 under our instrumental settings (calculated from linear fits). Remarkably, this trend was not observed in reference compounds (Figure 1b, e). Such a discrepancy has been observed also by Zark et al. for the common CO₂ loss, and was interpreted as a result of intrinsic averaging.31,45 In contrast, Dit Foque et al. described potential separation of less complex isomer mixtures by ramped fragmentation.29 Bearing the
limitation in mind that we only analyzed four IPIMs here, our results support the intrinsic averaging hypothesis and indicate that fragmentation sensitivity may be an additional property shaped by DOM complexity.\textsuperscript{18,20,45} It also supports our assumption of a high number of isomers and isobars “hidden” beneath each precursor molecular formula, which also increases the probability to detect meaningful links between precursor and product ions. A minor group of oxygen-poor formulas was non-responsive (Note S-3). Matching to the list of all SIRIUS Δm’s showed no significant relation to O/C ratio but to mass defect (Figure 1c, f). In contrast to mass defect, initial ion abundance showed no link to fragmentation sensitivity but was significantly correlated to higher numbers of Δm matches ($r = 0.41$, $R^2 = 0.17$, $n = 157$, $p < 0.001$; see also Tables S-9, S-10, S-11, S-12, and Figure S-6). DOM precursors with an average O/C ratio matched more often than low O/C, fragmentation-resistant precursors (Figure 1c; Figure S-7, Note S-3)\textsuperscript{18,19,35} or high O/C, easily fragmented precursors (Figure 1b). These observations together show that fragmentation sensitivity and Δm matching seem to be independent DOM precursor properties and that Δm matching could be driven by ion abundance. SRNOM and the soil water sample shared many molecular formulas ($n=107$; 84\% of soil DOM and 74\% of SRNOM formulas) which accounted for most of the precursor ion abundance at NCE 25 (96.5\% and 97.2\%, respectively). Despite this high similarity, SRNOM precursors showed higher numbers of Δm matches (Figure 1c, f) which could indicate that the same molecular formula is more chemodiverse, i.e. has more underlying structural formulas, in SRNOM compared to soil DOM (further discussion in section 3.5).
Figure 2. Δm matching in chemical space for soil (porewater) DOM (panels a – f) and SRNOM (panels g – l). Exemplary reference compound structures with marked indicative Δm units are shown in lower panels (m – q). Grey boxes refer to anticipated structural domains (Figure S-3). Panels a – l show precursors with an annotated molecular formula by their atomic H/C and O/C ratios (Van Krevelen plot; soil DOM, n = 127; SRNOM, n = 144); grey boxes indicate representative structural domains that are commonly used (see Figure S-3 for details). Dot size encodes the number of matches to non-indicative (a – c, g – i) vs. indicative Δm's (d – f, j – l); see legends in every plot. Colored boxes in indicative VK plots mark the expected structural region of formulas that would yield the respective Δm, and colors refer to the structural motifs marked in panels m - q. Calculations based on Δm data are presented in more detail.
in Table S-13. Highlighted red open diamonds in panels e and k indicate loss of up to three gallic acid equivalents (size not drawn to scale).

### 3.3. Evaluation of the Δm matching approach.

We used the matching data of molecular formulas in DOM for a proof-of-concept evaluation of our Δm matching approach. Specifically, we aimed to test the hypothesis that all precursors are potentially linked to all product ions in chimeric MS² spectra of ultracomplex DOM. Our analysis was congruent with previous observations, showing ubiquitous losses of non-indicative oxygen-containing functionalities (Table S-6) while also revealing more detail (Figure S-4c, Table S-7). Details are given in the Supporting Information (Note S-4); in short, we found expected trends in losses of CO₂, CO, and CH₂ in both samples (Figure 2a – c, g – i, Table S-13). The predicted heteroatom content (O, N, S) of assigned molecular formulas and a widened tolerance window were used for further analysis of the uncovered structural information. Random Δm matching would be expected if the calculated Δm values were affected by low resolution, low sensitivity, or artifacts such as reactions in the instrument (e.g., between the collision and Orbitrap cell⁷⁸). Instead, we found that 1) precursors with low ion abundance matched to less Δm features (Figure S-6), 2) non-fragmented precursors matched to less or no Δm’s (Figure S-7), and 3) identity of Δm matches agreed with molecular formula prediction (e.g., loss of S-containing Δm’s from S-containing precursors; ≤ 3 CO₂ losses from precursors containing seven O, etc.; Figures S-8 and S-9). Our evaluation also shows that Δm matching not only helps in recalibration⁷⁹ but also serves to check formula annotation routines, as it revealed unresolved precursor compositions interfering especially with CHOS precursors (related to Cl, P and F). This means 1) that these atoms should be included for better coverage of elemental composition (i.e., prioritization) in our specific sample context and that 2) higher resolution power may be required to resolve S-, Cl-, P-, or F-containing precursor compositions.¹ In summary, Δm matching revealed an inherently structured biogeochemical signal of precursors that seem to fragment individually and was highly sensitive in detecting precursor-product ion links. This suggests that chimeric DOM data can be deconvoluted to reveal differences in molecular composition not visible from MS¹ inspection.²³,⁸⁰ It should be stressed that these results will need further evaluation due to the small number of DOM precursors, m/z values and samples analyzed here (159 in soil
3.4. Structural domains emerge from clustering with reference compound Δm’s. DOM precursors from both samples were compared based on Δm matching as an indicator of structural information (Table S-7, see section 3.1). We grouped DOM precursors, reference compounds and Δm features by two-way hierarchical clustering (Table S-14). In the following, precursor clusters will be referred to by letter (A - H) and Δm clusters by number (1 – 7; Table S-15). Based on the specificity of SIRIUS Δm features (Table S-14) and clustering with 14 reference compounds we defined five of the Δm clusters found herein as being structure-specific (Table S-15, some shown in Figure 3d, e, j and k; for details, see also Table S-13).

Δm features C₄H₈O₄ (120.0423 Da, equivalent to tetrose loss) and C₆H₁₀O₅ (162.0528 Da, equivalent to hexose loss), both members of cluster 2, were found to be specific for alcohols and polyols, carbohydrates, and carbohydrate conjugates, as well as ethers (Table S-14). Reference compounds containing a polyol (quinic acid in #7) or a sugar (glucose in #12 and #13, mannose in #14) contributed Δm’s to this cluster (Table S-15). Δm equivalents of these losses matched to 18 soil DOM and 24 SRNOM precursors in the central Van Krevelen plot despite the absence of “carbohydrate-like” precursors (lilac square in Figure 2d, j and o, q). The anticipated shift towards higher O/C and H/C ratios was nonetheless apparent in both samples, especially compared to precursors associated with clusters 3, 4 and 7 (Figure 2e, f and k, l).

Δm features of clusters 3 and 4, partly specific to phenylpropanoid and benzenoid structures, were contributed by flavan-3-ols (#10, #11) and aglycones of flavon-3-ols (#12 and #13) and those containing cinnamic, coumaric or gallic acid units (#7, #9, #11). Δm features contributed by flavan-3-ols and aglycones were found in the “lignin-like” domain (orange square in Figure 2e, k; orange circles in panels o, p, q). These C- or H-rich Δm’s (e.g., C₆H₁₀O₂ or C₇H₈O₃) are likely no combinations of common O-rich losses (CO, H₂O, or CO₂) due to their low O/C and O/H ratios, but this requires further testing with model mixtures. Aliphatic chains prevail as O-poor substructures in substituted cyclic core structures and could contribute. Similar to detection of polyol-equivalent Δm matches outside the expected carbohydrate domain, gallate-equivalent losses were not matched to precursors in the
anticipated “tannic” domain (red diamonds and turquoise square in Figure 2e, k; turquoise circle in panel p).

Among the most prominent features was the methyl radical loss\textsuperscript{35,49,50} which matched to oxygen-poor DOM precursors and was one of three Δm features in cluster 7 (soil DOM: n = 18, average O/C = 0.33, SRNOM: n = 25, average O/C = 0.32, Figure 2f, l). The distribution of CH\textsubscript{3}*-yielding precursors was paralleled by CH\textsubscript{2} (soil DOM: r = 0.60, R\textsuperscript{2} = 0.35, n = 127, p < 0.001; SRNOM: r = 0.63, R\textsuperscript{2} = 0.39, n = 144, p < 0.001) and CO losses (r = 0.55, R\textsuperscript{2} = 0.30, n = 127, p < 0.001; r = 0.58, R\textsuperscript{2} = 0.34, n = 144, p < 0.001) and implied structural similarities (Figure 2f, l), e.g., condensed structures with aliphatic, lactone, or quinone moieties.\textsuperscript{14} CO and CH\textsubscript{3}* were both indicative of benzenoid structures in the SIRIUS-annotated Δm data (Table S-14). The methyl radical loss is an expected diagnostic Δm of methoxylated aromatic rings such as present in lignin (orange square in Figure 2f, l; orange circles in panels m, n; see Note S-5), but was also matched to DOM precursors not classified as “lignin-like.”\textsuperscript{18,31,35,49} The Δm features CH\textsubscript{3}* CO and C\textsubscript{2}H\textsubscript{4} were also linked to CH\textsubscript{4} vs. O series. These describe a regularity in DOM data explained by increments of 0.0364 Da, and are formally annotated as an exchange of CH\textsubscript{4} for O (Figure S-10).\textsuperscript{37,38} Concurrent losses of CO and C\textsubscript{2}H\textsubscript{4} explained the presence of CH\textsubscript{4} vs. O increments on the product ion level and were paralleled by losses of CH\textsubscript{3}*. This finding could also explain the ubiquitous presence of CH\textsubscript{4} vs O series in non-fragmented DOM; for example, concurrent β-oxidation and de-carbonylation could be enzymatic analogues of the patterns seen in MS\textsuperscript{2} data.\textsuperscript{26}

Matching to Δm features derived from a small set of reference compounds revealed emerging clusters of precursor and Δm feature families that may prove more indicative if constrained with further DOM and reference compound data.\textsuperscript{14} Anticipated structural domains were apparent but showed clear overlap, which means that the same precursor was part of more than one Δm-predicted structural domain. An extended analysis using the set of compound class-associated SIRIUS Δm features showed similar trends (Figure S-11, compare Figure 2). These findings must however be taken with caution for four reasons:
1) SIRIUS Δm features were not obtained on the same instrument and thus may include features that, although correlated with certain compound classes, may not appear in DOM under the same instrumental settings.

2) SIRIUS Δm features may be biased towards certain classes of compounds, as is our reference set of 14 aromatic compounds. Here, we only considered negative ESI mode data which is commonly employed for DOM analysis, and thus, adding positive ESI or other ionizations would extend the range of Δm features and structural classes covered and likely decrease bias.14,16,23,84 The same applies to other fragmentation techniques than CID.

3) Product ion abundance was disregarded in our analysis, but could be used to weigh probabilities of potential precursor-product ion links in future, potentially in combination with fragmentation energy gradients (fragmentation tree analysis)85, moving m/z isolation windows, or variations in ion accumulation times that influence MS^1 ion abundance.86

4) Despite a seemingly improved separation of extreme classes (high H/C ratios in fatty acids, high O/C ratios in carbohydrates, etc.), potential overlap in domain boundaries remained considerable (Figure S-11).

Categorization of precursors into multiple Δm-defined structural domains was also reflected by large differences in Δm matching between members of the same a-priori defined structural domains or classes (i.e., only based on molecular formula). Twenty-seven precursors shown in Figure 3 were classified as “lignin-like” formulas and were part of seven precursor clusters (B – H; Table S-16), thereby showing clear differences in potential structural composition. Likewise, CHNO and CHOS precursors matched with many of the S- and N-containing SIRIUS Δm features (spanning 3 – 78 S- and 4 – 251 N-containing Δm’s in soil DOM and 0 – 154/ 0 – 350 in SRNOM; Tables S-17, S-18, S-19 and S-20). These represented on average 79 ± 19% (63 ± 31% in SRNOM) of all Δm matches per CHOS precursor or 91 ± 7% (79 ± 28%) of all CHNO precursor matches (detailed analysis, see Note S-6). Many Δm features were also associated to compound classes, revealing potential structural detail (Table S-21). For example, CHNO precursor matching indicated the absence of nitrate esters, but indicated the presence of reduced forms of N partly
explained in the literature\textsuperscript{87,88}, including specific $\Delta m$'s related to aralkylamines, amino acids, carboximidamides, and dicarboximidamides/ urea-containing compounds. S-containing $\Delta m$ matches indicated the potential presence of sulfonic, thiol, thioether or aromatic S precursors.\textsuperscript{84} Taken together, our results show a wide potential diversity of N and S compounds in DOM that contradicts earlier reports of mainly aromatic N and sulfonic S.\textsuperscript{34,89,90} As most of these studies analyzed marine DOM, the detection of potentially more diverse sets of CHOS and CHNO precursors could relate to the terrestrial, less degraded DOM analyzed here.\textsuperscript{16,91–93} Further tests with N- and S-containing reference compounds and DOM samples are warranted to reveal the hidden diversity and identity of dissolved organic nitrogen and sulfur and confirm potential structures, e.g., by NMR.

All in all, our results show that it may be possible to refine Van Krevelen domains by deconvoluted MS$^2$ data, and that complementary precursor information could be used to assess false or biased $\Delta m$-based class assignments (e.g., elemental composition, DBE, ionization, fragmentation sensitivity, ion mobility, polarity index, etc.).\textsuperscript{13,56} Fluorescence or NMR spectroscopy could add valuable information if DOM would be fractionated before MS$^2$ data acquisition.\textsuperscript{21,94}

Data-dependent and data-independent acquisition (DDA, DIA) techniques could be used to cover the whole mass range of precursors in DOM mass spectra in future, and are widely employed in LC-MS of complex mixtures.\textsuperscript{16,27,95,96} For example, Ludwig et al. presented a DIA scheme (SWATH-MS) that employs one precursor scan and 32 isolation windows of 25 Da width, covering 800 Da within 3.3 seconds; similar schemes are likely transferable to acquire full mass range data of directly-injected DOM.\textsuperscript{97} Kurek et al. recently presented such data, covering product ions generated from similar isolation window ($m/z$ 392 – 408).\textsuperscript{16} Smaller isolation windows as used herein were also employed by Leyva et al. to discern fragmentation pathways and structural families in DOM (mass range $m/z$ 261 – 477)\textsuperscript{14}; this approach could be extended to include the diversity of $\Delta m$ features shown here. Together, this shows that practicable tandem MS acquisition strategies are in reach and will enable deeper analyses of $\Delta m$ features in DOM soon.

3.5. Drivers of differences in $\Delta m$ matching between soil DOM and SRNOM. Although matching among the two samples was largely consistent, slight differences were apparent from Van Krevelen
distributions (Figure 2). We therefore tested the separation of precursor clusters by ordination (Figure 3).

Precursor clusters were clearly separated on PC1 and PC2 which together held about 47% of variation. Most considered precursors were shared among samples (64%, 38 out of 59), only a small number was sample-specific (SRNOM = 14, Soil DOM = 7). Sample-specific precursors were found in clusters A (linked to carboxylic acids), B (phenols, polyols) and C (benzenoids, Table S-15), the remaining clusters D – H were dominated by the shared precursors. Out of the 38 shared precursors, 30 (79%) grouped in the same precursor cluster despite a general trend to higher numbers of matches in SRNOM, but eight grouped differently (bold precursors in Figure 3a). These differences in matching could be related to different chemistries, i.e., different isomeric/isobaric composition. For example, based on the correlation of precursor properties (Figure 3b), the cluster “switch” in C_{11}H_{14}O_{6} was largely explained by higher ion abundance and Δm matches in SRNOM, while in C_{23}H_{22}O_{4}, the effect was partly linked to higher fragmentation resistance in SRNOM. Unfortunately, we only have data on initial ion abundance and fragmentation sensitivity from the soil DOM isolate; other precursor properties, however, showed very similar trends in both samples (Figure 3b).
Figure 3. Separation of DOM precursors based on Δm matching. a) Principal Components analysis of all precursors with more than one match to indicative Δm features of the 14 reference compounds (i.e., Δm features shown in Table S-7 that are not part of Table S-6, see section 3.1). Colors of dots distinguish precursors from both samples and reference compounds (see legend). Precursors detected in both samples are connected by dotted black lines. Precursor clusters (A – H) are marked by envelops and letters (compare Tables S-14 and S-15). Eight shared precursors that switched precursor clusters are highlighted by bold molecular formula (C_{12}H_{14}O_{9}, A in soil DOM → H in SRNOM; C_{19}H_{26}O_{3}, B → C; C_{26}H_{26}O_{3} and C_{23}H_{22}O_{4}, B → D; C_{17}H_{14}O_{5}, G → E; C_{18}H_{12}O_{7} and C_{22}H_{26}O_{6}, H → E; C_{11}H_{14}O_{6}, H → G).
b) Correlations of selected precursor properties with scores of PC axes (only DOM precursors with assigned molecular formula included in the correlation). PC axes 3 and 4 are shown in addition. Correlations are indicated for all precursors (n=94) and those detected in each sample (Column “Sets”). For each combination (PC = x, property = y), Pearson’s r and significance are given (0.05 ≥ p > 0.01, “*”; 0.01 ≥ p > 0.001, “**”; p ≤ 0.001, “***”). Negative/ positive correlation is indicated also by color (blue, red); non-significant correlations are shown in lighter color or no color if no direction dominated. Matches, matches against the global list of Δm features; Structures, number of hits in natural product and in-silico databases.

Similar clustering and Δm-predicted structural classes (Figure S-11) in shared precursors could indicate a conserved structural composition. Likewise, Kurek et al. observed high similarity in APPI-ionized and IMPRD-fragmented DOM samples, but observed clear differences in CHOS fragmentation.16
similarities between DOM samples would be in line with stoichiometric principles (i.e., due to a large share in precursors between DOM samples) and could suggest that DOM processing diversifies, but also “randomizes” the molecular composition of each precursor (“universal” signal). High congruence of fragmentation patterns (and thus, Δm matching) among DOM precursors has also been interpreted as a sign of similarly substituted but slightly differing core structures. The clusters devised here were small due to the relatively small number of precursors and m/z values analyzed, and thus may not detect significant differences between samples yet. However, even with our small set of precursors, the clustering by Δm matching showed conserved differences in fragmentation between precursor clusters, and in part, even the same precursor in different samples. The fact that this could relate to differences in ion abundance (and therefore, possibly also ionization efficiency) or fragmentation sensitivity is intriguing and should be investigated across a wider range of DOM chemotypes using improved classification approaches as applied here (see also section 3.4).

3.6. Ion abundance is linked to Δm matching frequency and structural diversity. Ion abundance was the most important driver for Δm matching in both samples and highest in the structural domain usually assigned to ubiquitous lignin structures or carboxyl-rich aromatic molecules. This domain also parallels with a maximum in potential underlying chemodiversity, which could explain why these signals are ubiquitously found and especially dominant in reworked DOM. Δm matching showed potential to reveal this underlying chemodiversity effect and was therefore compared to numbers of structure suggestions and Δm-predicted compound classes per precursor (Figure 4). Numbers of Δm matches were significantly and positively related to the number of structure suggestions in absolute terms and for specific compound classes (Table S-22). The correlation between Δm-predicted and suggested compound classes was surprisingly similar in both samples and significant for almost all benzenoid-type (benzopyrans, methoxybenzenes, anisoles, phenols, etc.) and most phenylpropanoid-type structures (flavonoids, linear 1,3-diarylpropanoids). Among the organic acids, only vinylogous acids stood out (i.e., containing carboxylic acid groups with insertions of C=C bond(s)). Significant correlations were also found for pyrans, acryloyl compounds, carbohydrates, aryl ketones and alkyl aryl ethers (fatty acids and analogues only in SRNOM).
Figure 4. Agreement between chemodiversity estimates based on molecular formula (structure suggestions) and precursor-product ion links (Δm matches). Panels a, b) Correlations between numbers of SIRIUS Δm matches vs. structure suggestions (note log scale, incl. in-silico hits); a) soil DOM, b) SRNOM. Panels c, d) Number of SIRIUS Δm matches in Van Krevelen space (scales are similar but legends show different dot sizes); c) soil DOM, d) SRNOM; grey boxes refer to domains defined in Figure S-3. Panels e, f) Number of predicted classes per precursor based on SIRIUS Δm matches (color scale similar in both panels). Structural classes are associated to SIRIUS-annotated Δm features through correlation analysis of host structures and their Δm features (classification based on Classyfire); e) soil DOM, f) SRNOM.

The positive link between ion abundance and numbers of Δm matches on the one hand and predicted and suggested structures on the other indicates that ion abundance may be linked to the number of structural isomers and isobars per molecular formula in FTMS spectra of DOM and explains why Δm-defined structural domains showed strong overlap in this study. It also provides additional support to our assumption that all precursors potentially contribute to all product ions in DOM: The patterns revealed through Δm matching were largely congruent with the independent estimate of structural composition by natural product databases. The fact that only some classes of compounds (mainly benzenoids and phenylpropanoids) showed significant correlations could point to bias towards plant natural products in the databases employed here; this means that the inclusion of other structure databases and the additional assignment of Δm’s not
only to their host structures but also to host organisms (e.g., in GNPS\textsuperscript{65}) could reveal further clues about the potential sources of molecular formulas in DOM.

We propose that the number of Δm matches could be interpreted as a novel, relatively easily accessible measure to account for a precursor’s underlying potential structural diversity. Such information could help to better understand mechanisms of DOM formation and persistence in the environment. Our results encourage further studies on the Δm matching behavior of synthetic mixtures of known structures and across DOM chemotypes, and the improved bioinformatic exploitation of chimeric (LC-) FTMS\textsuperscript{n} data of complex organic mixtures.\textsuperscript{14,102–104} We acknowledge that natural product and in-silico databases are far from being complete, same as the database of annotated Δm matches we used here, despite its large coverage of ~18000 unique structures and ~11500 Δm’s. For example, precursors with low mass defects showed exceptionally few structural hits, indicating bias in natural product databases (Figure S-12).\textsuperscript{18} These structures were easily fragmented and yielded few Δm matches in our analysis; N- and S-containing precursors were double as likely to show no suggestion compared to CHO precursors. This shows that DOM contains unique molecular structures to be identified in future.

4. IMPLICATIONS

Tandem MS data of complex samples such as dissolved organic matter (DOM) is impeded by the co-fragmentation of precursors with similar nominal mass, and further complicated by the contribution of potential isomers and isobars of a precursor. We employed an approach that analyzes the pairwise mass differences between all precursor and product ions as a whole (Δm matrix). Using a very limited set of precursor features from two samples, we found potential signs of structural imprints related to benzenoids, phenylpropanoids, carbohydrates, sulfonic acids, thiols, thioethers and amino acids, amongst others. The successful matching of indicative Δm features and precursor clustering suggests a remaining – and recognizable – source imprint of primary or recycled plant remains in DOM. Tests with more DOM samples and artificial/ treated mixtures (e.g., DOM with spiked known compounds, or DOM degraded by specific enzymes) are required to test the assumptions employed here and to improve classifications of DOM precursors by Δm clusters. Our first results indicate that FTMS\textsuperscript{2} data may be useful to differentiate molecular
composition on the molecular formula level, and that ion abundance and fragmentation sensitivity are two key variables that explain differences in MS² data within and among samples. This is intriguing because a shared molecular formula could harbor a completely different set of structures but must be assessed with larger sets of DOM data which would improve detection of such differences. Generally, our findings support the view that Van Krevelen domains are associated with indicative mass losses that relate to stoichiometric differences between compound classes. The most abundant precursors however showed a mixed MS² signal that caused boundary overlap of these Δm-defined domains (Figure 4e, f). While this finding is in line with known patterns of structural diversity and partly explains the ubiquitous presence of abundant DOM signals, it introduces a new paradigm to the interpretation of DOM FTMS data by assigning unknown precursors to multiple structural categories instead of just one. Further evaluation of both natural and spiked/ treated complex mixtures, constantly growing MS databases, and comprehensive decomplexation methods (LC-MS, IMS) will together provide fundamental insights into the deconvolution of chimeric spectra from complex samples, and ultimately show the potential to unfold the hidden molecular diversity and identity of DOM.

ASSOCIATED CONTENT

Data and Software Code Accessibility
All tandem MS data can be found in on the Mass Spectrometry Interactive Virtual Environment (MassIVE) under the following links: ftp://massive.ucsd.edu/MSV000087117/ (soil DOM data), ftp://massive.ucsd.edu/MSV000088869/ (SRNOM data) ftp://massive.ucsd.edu/MSV000087133/ (reference compound data) (Data Set S-1, raw peak data, *.mzML files). All other data associated to this manuscript (extensive tables, Δm feature lists, Δm specificity, two-way clustering table, processed data used to create figures, etc.) is available as online free of charge from the PANGAEA Data Publisher under the following link: https://doi.pangaea.de/10.1594/PANGAEA.932592 (Data Set S-2, processed data as *.xlsx files).

Supporting Information
The supporting information contains 22 tables and twelve figures, six additional notes, and 69 references.
Table S-1: Information on reference compounds and solutions used in this study. Table S-2: Instrument settings for fragmentation experiments. Table S-3: Recalibration peaks used for reference compound in FTMS measurements. Table S-4: Precursor and major product ions of the 14 reference compounds. Table S-5: Results of reference
compound’s tandem MS data analysis with CSI:FingerID. Table S-6: List of reported DOM Δm features from MS1 and MS2 studies. Table S-7: List of all 50+5 Δm features extracted from the reference compound dataset. Table S-8: Properties of four isolated nominal masses (IPIMs) at different NCE levels. Table S-9: Overview of correlations between key properties of the IPIM 241. Table S-10: Overview of correlations between key properties of the IPIM 301. Table S-11: Overview of correlations between key properties of the IPIM 361. Table S-12: Overview of correlations between key properties of the IPIM 417. Table S-13: Lists of Δm values used for analysing matching patterns in Van Krevelen space. Table S-14: Matching behavior of precursor clusters against Δm features (Table S-7). Table S-15: Summary of two-way clustering of DOM precursors and reference compounds. Table S-16: Lignin-like precursor formulas and their molecular properties and clustering. Table S-17: S-containing precursor formulas in soil porewater DOM. Table S-18: N-containing precursor formulas in soil porewater DOM. Table S-19: S-containing precursor formulas in SRNOM. Table S-20: N-containing precursor formulas in SRNOM. Table S-21: Structural class-correlated Δm features matched to CHOS or CHNO precursors. Table S-22: Correlations between structure hits and specific Δm features in CHO precursors. Figure S-1: Overview of reference compounds used in the study. Figure S-2: Mass accuracy assessment based on reference compound Δm’s. Figure S-3: Distribution of exemplary known structures in chemical space. Figure S-4: Orbitrap tandem MS of soil porewater DOM. Figure S-5: Comparison of matches to the two short Δm lists in relation to m/z and NCE. Figure S-6: Δm matches in relation to precursor ion abundance in soil DOM. Figure S-7: Δm matches in relation to precursor fragmentation sensitivity in soil DOM. Figure S-8: Matching assessment with SIRIUS Δm’s (Molecular formula check). Figure S-9: Changes in Δm matching frequency upon widening of tolerance window. Figure S-10: Link between matches to CH3*, CO and C2H4 and CH4 vs. O exchange series. Figure S-11: Structural VK domains based on class-correlated SIRIUS Δm features. Figure S-12: Effect of mass defect on the number of structure suggestions. Note S-1: Supplementary experimental details. Note S-2: Detailed description of reference compound fragmentation behavior. Note S-3: Behavior of non-responsive DOM precursor ions. Note S-4: Δm matching: Proof-of-concept data and key findings. Note S-5: Potential esterification of DOM by methanol during SPE and storage. Note S-6: Structural insight into N- and S-containing DOM precursors.

The Supporting Information is available free of charge on the ACS Publications website.

**AUTHOR INFORMATION**

**Corresponding Author**

*Email: gerd.gleixner@bge-jena.mpg.de*
Present Addresses
† C.S.: Institute for Biogeochemistry and Pollutant Dynamics, ETH Zürich, Zurich, Switzerland & Swiss Federal Institute of Aquatic Science and Technology (Eawag), Department Water Resources and Drinking Water, Duebendorf, Switzerland
§ V.-N. R.: Thüringer Landesamt für Umwelt, Bergbau und Naturschutz (TLUBN), Jena, Germany

Author Contributions
CS performed the measurements. DP, VNR, PD and GG were involved in planning and supervised the work. KD and SB compiled global Δm feature data, analyzed its specificity, and performed structural classifications of Δm host structures as well as structure suggestions of DOM precursors. CS processed the experimental data, performed the downstream analyses, drafted the manuscript, and designed the figures. The manuscript was revised through the contributions of all authors. All authors have approved the final version of the manuscript.

Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENT
We thank all members of the Dorrestein lab for helpful discussions and insights. We also like to acknowledge Vivian Stefanow for initial structure database surveys. We acknowledge the International Max Planck Research School for Global Biogeochemical Cycles (IMPRS-gBGC) for sponsoring CS' research stay in the Dorrestein lab at UCSD, CA, USA. The authors acknowledge the financial support from the Max-Planck-Gesellschaft (MPG) and the German Research Foundation (DFG, Deutsche Forschungs-Gemeinschaft) as part of the CRC 1076 "Aqua Diva" and a Postdoctoral Research Fellowship PE 2600/1, and support through the CMFI Cluster of Excellence (EXC 2124) to D.P. V.-N.R. received additional funding by the Zwillenberg-Tietz Stiftung. C.S. received a Ph.D. stipend from the International Max Planck Research School for Global Biogeochemical Cycles (IMPRS-gBGC). The authors are grateful for critical comments and helpful suggestions received from the three anonymous reviewers.

REFERENCES


Engström, M. T.; Päljärvi, M.; Salminen, J. P. Rapid Fingerprint Analysis of Plant Extracts for Ellagitannins, Gallic Acid, and Quinic Acid Derivatives and Quercetin-, Kaempferol- and Myricetin-Based Flavonol Glycosides by UPLC-QqQ-MS/MS. *J. Agric. Food Chem.* 2015, 63 (16), 4068–4079. https://doi.org/10.1021/acs.jafc.5b00595.


Jeffryes, J. G.; Colastani, R. L.; Elbadawi-Sidhu, M.; Kind, T.; Niehaus, T. D.; Broadbelt,


