# Natural complex mixtures unequivocally defined in formulae difference space

Anastasia Sarycheva[1], Irina V. Perminova[2], Eugene N. Nikolaev[1], Alexander Zherebker[1*†]

[1]*Skolkovo Institute of Science and Technology, Moscow, 143026, Russia*

[2]*Department of Chemistry, Lomonosov Moscow State University, Moscow, 119991, Russia*

**Corresponding author.*

*†E-mail: a.zherebker@skoltech.ru*

**Direct comparison of high-resolution mass spectrometry (HRMS) data acquired with different instrumentation or parameters remains difficult as the derived lists of molecular species via HRMS, even for the same sample, appear distinct. This inconsistency is the result of inherent inaccuracies caused by instrumental limitations and sample conditions. We propose a method that classifies HRMS data based on the differences in the number of elements between each pair of molecular formulae within the formulae list to preserve the essence of the given sample. The novel metric, Formulae Difference Chains Expected Length (FDCEL), allowed for comparing and classifying samples measured by different instruments. FDCEL metric was implemented for both spectrum quality control and for examination of samples of various nature. We also demonstrate a web application and a prototype for a uniform database for HRMS data serving as a benchmark for future biogeochemical applications.**

## 1 Introduction

Natural organic matter (NOM) is an important part of the organic carbon pool, cycling through different environments where recalcitrant species can accumulate with time while labile organic pools are degraded[1,2]. Under specific conditions, labile NOM can be conserved in its fresh form, e.g., preserved by ice in permafrost, and released during thawing seasons or due to global warming[3,4]. The study of NOM stabilization became an intriguing topic for scientific research[5,6].

Employing powerful high-resolution analytical techniques, patterns of NOM behavior and origin are explored: from preliminary carbon sequestration to its transformations, stabilization mechanism, and mineralization[7–10]. Breakthroughs in this field were achieved by employing ultra-high resolution Fourier transform mass spectrometry (FTMS), which reaches molecular level resolution and reveals the extreme complexity of NOM: elucidating the exact elemental composition of thousands of detectable ions simultaneously[7,11–13].

HRMS application in environmental studies grows exponentially: for the last twenty years, thousands of works have been published (see Supplementary Fig. 1). The next logical step is the development of a molecular library for NOM based on this data. However, the infirmity of untargeted HRMS, like in metabolomics[14], makes this a challenging prospect. Recently, the reproducibility of NOM molecular studies was called into question: different HRMS instruments produced drastically different formulae lists for the same NOM samples[15,16]. These variations, regardless of their causative agents (e.g. instrumentation settings, different adducts formation, partial fragmentation), make the employment of formulae lists to compare/or classify samples extremely difficult, bringing into question the value of HRMS for NOM studies.

We propose a novel method of data analysis that explores robust spectral features relevant to the geochemical origin of the samples, allowing for samples to be compared based on HRMS data acquired on different instruments. Evidence of these features can be found in interlaboratory comparisons where it was determined that there is a correlation of integral number averaged values from the spectra[15,16]. NOM is the natural mixture of secondary metabolites, which underwent biochemical, photo, and chemical reactions; therefore, we hypothesize that the connection between molecular formulae can be considered as a fingerprint of such transformations and, consequently, it should be more robust to changes in instrumentation settings compared to the formulae themselves. This hypothesis is based on the intelligible principle, which has been suggested for metabolomics data treatment. Ab initio metabolite networks can be constructed directly from FTMS data[17] because all metabolites are precursors and reaction products, thus, the mass differences in the spectrum correspond to the specific organism's biochemical processes[18], characterizing the sample independent of instrumentation. Similarly, successful attempts at the construction of molecular networks for NOM have been made [19,20], but only a limited number of chemically meaningful mass differences have been considered in detail. The molecular network itself is a concatenation of complex and interwoven relationships which are impossible to directly examine. Instead, we suggest employing statistics based on all differences between formulae within the formulae list (hereinafter, formulae differences, or FDs) to reveal reliable sample features and to calculate a new FD-based measure that will enable researchers to compare samples based on their origin. This approach was applied to data published in interlaboratory studies and used to extract the necessary features for the construction of the first NOM classification database.

# 2 Results

## 2.1 Formulae difference networks

We adopt formulae as input features and not mass peaks within the raw HRMS data to allow for the inclusion of existing HRMS datasets consisting of formulae lists as well as simplifying the computations associated with the introduction of difference between features as a new space. A formulae list derived for a sample can be represented as a formulae difference network (a compositional network, or a graph): each node is a molecular formula, and edges are the numerical vectors of differences in elements count between this formula and the rest of the formulae - FDs. Similar representations have been used for various applications[20–22]. Analysis of such networks can be targeted when FDs of interest are predefined. However, this approach is not appropriate for the comparison and classification of arbitrary samples: the consideration of a limited predefined FDs would bring bias in inter-sample network analysis. We demonstrated this by comparison formulae lists of two different soil samples obtained by different instruments, considering only the FDs series of $CH_2$/$CO_2$/$H_2$/$H_2O$. Clearly, in some cases a subset of FDs can cause a false impression that the two samples are the same (see Fig. 1 and Supplementary Figs. 2-6). Therefore, to perform unbiased analysis and avoid misleading FDs targeting, we implemented the full FDs statistics considering all connectivities between all molecular species within a sample.

## 2.2 FDs statistics: distribution, series

The analysis of FDs in their entirety is not trivial. The total number of FDs within a molecular species list is a 2-combination of all formulae, i.e. $n\frac{n-1}{2}$, where $n$ is the number of formulae within a list (see Methods). The smallest number of formulae describing a sample within the available datasets was 1111 for SHA-Ctk-d with 616605 total and 54359 unique FDs, respectively. In the collected datasets, the number of unique FDs varied between 8 thousand and 2 million (see Supplementary Table 1).

We calculated occurrences of each unique FD within the formulae lists and formed vectors of unique FDs counts. Compared to the employment of formulae lists for the differentiation of samples, the counts of unique FDs allowed for a better classification of samples: the same samples analyzed at different laboratories formed more reasonable clusters based on Cosine distance measure. The advantage of FDs count is illustrated by comparison of PLFA and ESFA samples, measured by 17 different instruments. FDs counts allowed for the data to be clustered samplewise correctly, unlike the comparison based on formulae lists (see Fig. 2), despite the fact that the FDs counts analysis excludes the important information about relative intensities. This implies the robustness of FDs. It is important to note that Cosine measure analysis depends on the considered dataset: Supplementary Fig. 7 provides thresholding histogram for Cosine distance clustering.

We have also employed various dimensionality reduction techniques (Multidimensional scaling (MDS), Principal component analysis (PCA), Non-negative matrix factorization (NMF), t-distributed stochastic neighbor embedding (t-SNE), Uniform Manifold Approximation and

Projection (UMAP)) to compare formulae lists. Out of the explored samples, only formulae lists detected for PLFA formed consistent clusters in all 2D and 3D visualizations (Supplementary Fig. 8). This can be associated with the consistent relative composition throughout the formulae lists detected for this sample (Supplementary Fig. 9, 10). While compositional analysis for other samples results in a lack of reasonable formulae-based grouping (Supplementary Fig. 11). In the case of dataset A, UMAP visualization of formulae lists (Fig. 3) showed that none of the lists were clustered sample-wise. At the same time dimensionality reduction techniques, when applied to vectors of FDs counts, lead to meaningful visualizations. UMAP applied to FDs counts showed nearly perfect clusters (Fig. 3). So, FDs counts distribution analysis allowed for a better grouping of formulae lists detected for the same sample. However, FDs counts distribution is sensitive to the composition of the compounds within the formulae list and this approach fails in some cases (Supplementary Fig. 12). In case of extreme deviations in the composition of the formulae lists (Supplementary Fig. 13) both formulae lists analysis and FDs counts distribution analysis fail (Supplementary Fig. 14).

Therefore, the calculation of FDs counts has proven itself beneficial in some cases with known data labels, but apparently, its effectiveness is limited and comparable to the previously proposed number-averaged values[15,16] which are computationally less complex. However, the number-averaged approach devalues the application of sophisticated and expensive HRMS instruments for NOM analysis because the provided resolution of thousands of molecular species has been condensed to several integral values. To capture the essence of formulae lists acquired with ultra-high resolution and simultaneously overcome the fundamental issue of instrumental inconsistency, we should introduce new statistics. Such statistics have to reflect the formulae list while being robust to the changes in the detected chemical space of the sample in question, which requires the following: 1) to abandon the notion of the exact formulae values within the list, since these values are not reliable, as we demonstrated in the examples above; 2) include the intensity data to further distinguish the groups of formulae lists detected for the same sample.

To integrate the intensity data, we have to consider the formulae difference network, where FDs connect the individual formulae with corresponding relative intensities. We will not limit it to the selected FDs but consider all unique FDs as in the FDs counts distribution analysis discussed previously. To do that, we suggest the following approach, which is implemented in the developed FDS application[23] (https://nommass.com). The formulae difference network consists of subgraphs, each formed by formulae pairwise connected via a certain unique FD. Every subset of formulae with corresponding relative intensities forms an FD series. (see Methods). Series can be compared for pairs of samples containing these FDs directly using any measure (e.g. Cosine). We highlighted above that the presence of an ion in the mass spectrum is not guaranteed. Hence, the FD series should be compared in a robust way independent of exact nodes' positions within the graphs, while simultaneously accounting for the intensity of such nodes and the structure of the graph itself. The latter can be captured via statistics introduced for the connected components of the graph denoted further as chains (see Methods).

To abandon the notion of the necessity of formulae to be detected consistently throughout the measurements for the same sample, we introduce FD Chains Expected Length (FDCEL) measure (see Methods). It employs the probability measure on the set of chains to rank the chains within the series according to their significance based on the length and nodes' intensities. Then, the expected value for the length of chains within the series is calculated. In FDCEL, we use the absolute difference of expected values for the length of chains within the series for the common FDs found in the compared formulae lists. For similar series the differences is small and the expected values for the chains' lengths are close. The method is illustrated in the flowchart (Supplementary Fig. 15).

According to our hypothesis, FDs are the result of chemical transformations within a sample associated with the genesis of the latter. The number of products and their relative abundances in a mixture must depend on the source and these parameters are the integral parts of FDCEL. Respectively, we can expect that an array of similar series in terms of FDCEL measure throughout formulae lists detected for the same sample even with varying HRMS instruments can be used to represent this sample. Further, these series can be used to compare the sample with any input formulae list in terms of FDCEL to either classify it or check the quality of the input data. To demonstrate FDCEL implementation, we built the first NOM classification database[23], described below, to serve as a benchmark for future biogeochemical applications.

## 2.3 Database

To describe a sample and, store it within the database, we require a group of several formulae lists that were either detected for this sample via different HRMS setups, or when the sample was separated into fractions, and the latter were analyzed via HRMS. The series of the same FDs are compared via FDCEL within every aforementioned group. We consider only FDs which can be found throughout the group of formulae lists: this allows for the preliminary selection of the series which are characteristic to the sample regardless of the instrumentation. This reduces the time required for the calculations even though the FDs list remains excessive.

At the next step important and representative FD series are selected, which for the same sample should have consistent expected lengths of chains regardless of the composition overlap. By "consistent", we mean that the difference between expected values should be negligible rather than matching perfectly as it is shown below. Within the FDCEL measure, FD series are "weighted" for the compared lists according to the intensity of their nodes (see Methods). Note that the same FDs can be important for different samples and have different expected values. Important FDs can be additionally ranked samplewise according to their occurrence throughout the lists of selected FDs for other samples within the database.

As a result, each sample can be stored in the database as a ranked list of representative FD series with corresponding expected values for the length of chains. Any sample analyzed by HRMS can be compared to this database for two purposes exemplified below: for quality control of the given formulae list in the case a known sample is present in the database and for pairwise comparison of a new sample against the database as part of geochemical research. The latter application requires an educated guess to avoid misinterpretation. The constructed database can be expanded with new samples represented by groups of formulae lists which can be obtained as described above. The comparison algorithm is described in Fig. 4.

The assessment of the formulae list detected for an arbitrary sample in relation to the samples present within the database can provide a researcher with several insights. First, if the investigated sample is present within the database, the input formulae list should be correctly classified. If it can’t be correctly attributed, the researcher might get an idea that his particular experimental setup is way off, even according to the measure which specifically aimed at avoiding direct formulae-intensities comparison, to account for the differences in sample preparation, differences in ionization processes, and the exact HRMS instrumentation. Second, if the sample is not present within the database, the proposed method assigns several parameters to the formulae list, which can be used for further evaluation: 1) the number of FD series characterizing each database sample present within the considered formulae list; 2) the “distance” in terms of FDCEL measure calculated on each set of FD series characterizing the samples within the database. This can roughly point out the database samples related to the examined one.

Validation of the method included the comparison of the publicly available formulae lists acquired for samples present within the database (namely, SRFA) but using different ionization techniques: electrospray ionization (ESI: SRFA-r[24], SRFA-t[11], SRFA-u[25]), laser desorption/ionization (LDI: SRFA-s[26]), paper spray ionization (PSI: SRFA-v[25]), and paper spray chemical ionization (PSCI: SRFA-w[25]). In Fig. 5, the results of the comparison of formulae lists against the database are presented. Despite clear differences between the formulae lists and even a lack of formulae overlap for different ionization methods, all of the formulae lists were correctly assigned to the corresponding sample (SRFA) based on the top 9000 FD series representative of samples except for SRFA-t and SRFA-r. SRFA-t mislabeling can be associated with its sheer size as it includes 16471 molecular species while the average molecular lists throughout the samples are under 6000 formulae. The number of unique FDs within SRFA-t is more than 2 million, while the average FDs number for samples in the database was under 300 thousand. Out of 9000 representative FDs most samples from the database shared over 8000 (average is 8465, with SD of 747) with the SRFA-t. Consequently, within the FDCEL space (see raw FDCEL values in Supplementary Interactive Fig. 1) SRFA-t appears close to all of the samples within the database since there are not enough series within SRFA-t that do not match the representative sets. As to SRFA-r, its mass-spectrum was bimodal, which is not typical for SRFA and other NOM samples (Supplementary Interactive Fig. 2,3). The FDCEL measure indicated this remarkable inconsistency. This highlights the applicability of FDCEL-based comparison for quality assessment.

Another application of the database[23] is the estimation of the geochemical relevance of any HRMS analyzed samples to the samples from the database. Fig. 6 illustrates 13 formulae lists for samples of different origins compared against the database. SLNOM was isolated from the organic-rich lake (Sion Lake[27]) by solid-phase extraction using two types of resin: XAD-8 (the International Humic Substances Society (IHSS)[28] protocol) and modern Bond Elute PPL[29]. Interestingly, SLNOM-XAD was closer to the SRNOM (Suwannee River NOM), while SLNOM-PPL would be attributed to SRFA (Suwannee River Fulvic Acids (FA)). Full organic matter from the Suwannee River is richer in aromatic constituents as compared to SRFA. This is corroborative with the investigation of resin selectivity: XAD resin is slightly more selective toward aromatic species[30]. The comparison of the sample isolated from the Panikovka river on PPL resin[31] against the database gives an interesting perspective. This sample of a shallow cold-water river fed mainly by spring waters. According to FDCEL, the corresponding formulae list (PRNOM) is closest to PLFA (Pony Lake FA), which is composed exclusively of the bacterial-derived NOM[32]. This result implies that PRNOM is formed with a negligible contribution from soil drains. A clear analysis interpretation can be derived for the extract from the gray forest (GF) soil[33]. Despite the fact that other soils and even water extracts (Soil Dissolved OM: SDOM-Ctk (from the Mollisol), SDOM-PD (from the sod-podzolic soil)) are also present within the database, SDOM-GF was the closest to ESFA, which by IHSS definition was extracted from the arable gray soil[34]. Thus, despite differences in the isolation procedure, the implementation of FDCEL enabled correct geo-chemical assignment. Still, note that the extraction procedure may bias the results, especially, under harsh conditions. This is highlighted below using the examples of alkali-isolated humic substances.

The second set of samples was obtained from the thaws of yedoma ice complex deposit and from the watersheds in the Kolyma river basin as described in the previous publication[31] (AHF). According to the FDCEL measure, all samples were attributed to PLFA-like samples. This is supportive of, firstly, a high contribution of microbial-derived compounds in the permafrost thaw and, secondly, the significant presence of permafrost leachates in these Arctic rivers. The high contribution of microbial-derived compounds has been suggested according to the thorough 1H NMR study that explored long-chain aliphatic moieties. The only sample, which could be hardly unequivocally identified by the FDCEL using top-ranked 9000 FDs, was AHF-RPP-10. Surprisingly, this isolate was equally attributed to the soil FA (SFA-Ctk) which would be mistaken even taking into consideration a closeness to the Chersky city because the isolate from the same tributary to the Kolyma River extracted one year later (AHF-RPP-11) was unambiguously designated as PLFA-like by FDCEL. Yet this result is corroborative with the higher long-wave absorbance coefficient of AHF-RPP-10 compared to AHF-RPP-11[31].

The humic substances samples stand aside. Four samples were obtained using IHSS procedures – two humic acids (HA) and two fulvic acids (FA). The closest assignment of low-moor peat humic acids (PHA-TTL[35]) was to sod-podzolic HA (SHA-PD). This was relevant since, like the peat, it is acidic soil while Mollisol soil (source for SHA-Ctk) is alkaline. However, the formulae list detected for coal HA (CHA-h[36]) was not unequivocally assigned to the CHA sample from the database: based on top-ranked 9000 FDs, SHA-Ctk suggested the lowest FDCEL distance, but CHA would be a second choice.  The unconventional result was also obtained for FA samples, which were isolated from the permafrost soil[37]. Supposedly, FA1-Y-15 and FA2-Y-15 are represented by the low-transformed organic compounds well conserved under the ice shield. However, employing FDCEL, we would consider them ESFA- or SFA-Ctk-like based on top-ranking 9000 FDs. We believe that such an assignment is a result of the extraction procedure engaging strong acid and base which destroys the integrity of NOM, leading to the condition-dependent molecular ensembles rather than source-dependent. In fact, the sample integrity and secondary reactions under the IHSS extraction procedure became a topic of scientific debates[38,39].  Consequently, test permafrost FA-isolates were attributed to FA from the two temperate soils despite the drastically different nature of the parent soils.

## 3 Discussion

Our results demonstrate that the employment of FDs statistics is the key to the assessment of the NOM formulae space. The exploration of FDs space remains a crucial chemical task left for further research, however, we have already addressed some important issues, e.g. when the FD series were sorted according to their importance for the representation of samples based on the FDCEL measure. The selected FD series do contain all the information which could be extracted from the formulae lists detected for the same sample: when recreating the original formulae lists as a superposition of the representative FD series, we found that the number of formulae with the corresponding intensities which were left out is negligible (see Supplementary Fig. 16). Thus, the representation and storage of samples as a collection of FD series is valid.

The database we are showcasing here[23] includes a limited number of samples, and it exists as a proof of concept. FDCEL measure bypasses the flaws of direct comparison of formulae lists acquired via different HRMS setups. The characteristic example of FDs advantage over formulae lists is HRMS data for SRFA samples ionized by different methods (Fig. 5). The overlap of formulae detected in each experiment was negligible since LDI and ESI are drastically different ionization methods. However, the employment of the first version of the database (limited to comparison according to 9000 FD important series stored for each sample) showed that all of these ionization methods successfully managed to yield FD series, which are characteristic for SRFA. Consequently, it is possible to aggregate various molecular data regardless of the ionization method, which eases the interlaboratory studies. For the same sample, various molecular species ionized by different methods are still connected by the same important FDs which have a clear geochemical basis, a non-trivial discovery upon which the database was created. To verify this, we added three samples (ISDY, SDOM-Ctk, SDOM-Pd) to the database which were not analyzed by different HRMS instruments: for each of the samples the fractionation was performed and fractions were measured by a single instrument. The overlap between fractions of the same sample varied, and in extreme cases, there were no common molecular species at all. Still, validation showed that such entries are still eligible for the database, which highlights the importance of the FDs in describing the sample, creating an important precedent. In fact, independent of the elemental composition, aromaticity[40], or the position on the van Krevelen diagram[41], molecular species are connected by the same important FDs, which reflect biochemical and chemical processes in NOM.

Each sample in the database is built based on the superposition of 4, 7, and 17 lists depending on the availability of the data for this sample. The important feature, which is worth noting, is that records of standard samples in the database (e.g. SRFA, SRNOM) are reliable. These standards are used for FTICR instrument tuning in all laboratories working with NOM. So, the tuning may vary but, ultimately, while the sample is still recognized correctly when compared to the database via FDCEL, experiments are valid. In the example from Fig. 5, SRFA-r and SRFA-t spectra represented inconsistent instrument tuning, which requires revision.

The application of the database for explorative geochemical research is also tempting. In Fig. 6, we showed some illustrative examples. But the database should be used carefully for this kind of task. Due to computational resource limitations, the current version of the database[23] stores each sample as a vector of its 5000 most important FDs and the corresponding expected values. In most cases, the list of important FDs is much bigger. In fact, the success of sample attribution to the database depends greatly on the number of used FDs. Such a drawback is clearly illustrated in Fig. 6 on the example of FA samples from the permafrost and some riverine samples. Based on the FDCEL measure, they could be also attributed to SRNOM- and SRFA-like samples. This is the result of the heterogeneity of SRNOM and SRFA samples that happen to contain FDs which are important for FA and Arctic River samples at the top of their ranked FDs list. The current database is a prototype that requires extension with many various NOM samples while also having to limit computational time with fair sample attribution.

In this work, the FDCEL measure was calculated based on novel FDs statistics. We showed that such an approach is independent of the exact detected formulae composition, and the same FDs may be important for samples without common significant molecular formulae overlap. That means that, ultimately, we do not need molecular formulae by themselves. Clearly, the described method can and will be extended to the processed mass spectrometry peak lists instead of formulae lists. In this case, instead of a formulae difference network, a mass difference network would be constructed. Moreover, molecular compositions are never assigned to all peaks in mass spectra. Therefore, by skipping the translation of original raw m/z space into molecular space, we might be able to account for peak shifts more adequately. Employing a highly resolved m/z axis instead of a discrete molecular formulae axis could lead to important mass differences, which would unambiguously define any sample. This is beneficial because when working with formulae lists, we found that important FDs do overlap between the samples; and while they still allow for a corrected assignment based on the FDCEL measure since the expected values for the same important FD vary between the samples, the growing number of samples within the database raises a concern — the overlapping important FDs might hinder the effective FDCEL distance so we would have to increase the number of considered FDs. Hence, the next step would involve the employment of mass difference statistics instead of FDs, which allows for the ultimate HRMS data analysis independent of experimental setups.

# 4 Methods

The starting point for our approach is a list of assigned formulae for each considered sample. If the data was taken in raw MS form instead of the formulae list, it was assigned molecular species via the lab-made Transhumus software based on a total mass difference statistics algorithm[42,43].

## 4.1 Definition of FDs, Series, Chains

The formulae list can be described as $S_u = \{(f_i, I_i)\}_{i=1}^{n}$ — a set of $n$ formulae (each formulae $f_i$ is represented by $C, H, O, N, S$-elements occurrence vector, i.e. $C_{t_C}H_{t_H}O_{t_O}N_{t_N}S_{t_S}$ is $[t_C\ t_H\ t_O\ t_N\ t_S]$) and the corresponding intensities $I_i$ found in the sample $S_u$. Instead of using the formulae lists to compare samples, we calculate all the differences between formulae $\{\delta f_i\}_{i=1}^{\binom{n}{2}}$ (where $\binom{n}{2}$is 2-combinations: $\binom{n}{2} = n\frac{n-1}{2}$) for each list. These differences are counted and the resulting vectors are compared between samples (Supplementary Note 1).

We define the difference series as follows. A series is an ordered subset $S_u(\delta f\ )$ of set $S_u$, the former includes pairs $(f_i, I_i)$, for which there are neighboring formulae at $\delta f$, i.e. $S_u(\delta f\ ) = \{(f_k, I_k) \in S_u \mid \forall k \exists j : |f_k - f_j| = \delta f\ \}$. Such series can be compared directly for matching $\{\delta f\ \}_{i=1}^{p}$ throughout samples using cosine measure. The resulting distance matrices are clustered, so each distinct cluster includes a set of $\{\delta f_i\ \}_{i=1}^{q}$which either reflects similarity or difference between samples.

We consider the graph $G$ with vertex set $S_u(\delta f\ )$, where any two vertices $(f_i, I_i)$ and $(f_j, I_j)$ are connected by an edge if $|f_i - f_j| = \delta f$. The connected components of $G$ form the set of chains $\{C_k\}_{k=1}^{r}$: each chain is characterized by its length (which is introduced as the number of nodes the chain includes $l(C_k)$) and by its vertices $V_k = \{(I_j^k, f_j{}^k)\}_{j=1}^{l(C_k)}$.

### 4.2 FDs Chains Expected Length measure

The chains are assigned weights according to their length and vertex intensities as follows $\mu: \{C_k\} \rightarrow [0,1]$ by $\mu(C_k) = \frac{\sum_{j:(I_j^k, f_j^k) \in V^k}^{l(C_k)} I_j}{\sum_{i:C_i \in S_u(\delta f)} (\sum_{j:(I_j^k, f_j^k) \in V^k}^{l(C_k)} I_j)}$. This probability measure on the set of chains (this measure can be redefined, e.g., if there are a lot of short chains which are not desired for the analysis, these chains may be dropped, and $\mu(C_k)$ can be defined on the resulting set of chains) determines how significant the particular chain of the series is. It allows for the calculation of the expected value for $l(C_k)$: $E[l(C)] = \sum_{i:C_i \in S_u(\delta f)} l(C_k)\mu(C_k)$ . Thus, more significant chains within the series are defining the sample.

As mentioned above, there are many inaccuracies (associated with sample preparation, ionization process, instrumentation) that make the direct comparison of samples based on formulae lists flawed. Instead, to compare two samples ($S_u$ and $S_d$), we propose using the series which correspond to the common set $\Delta = \{(\delta f_j, w_j)\}_{j=1}^{m}$, where $\{\delta f_j \mid \delta f_j \in \{\delta f_i^{S_u}\}_i \cap \{\delta f_i^{S_d}\}_i\}$ are the formulae differences found in both samples, and $w_j \in [0,1]$is the weight of corresponding $\delta f_j$ defined as $w_j = \frac{\sum_{k:I_k \in S_u(\delta f_j)} I_k + \sum_{k:I_k \in S_d(\delta f_j)} I_k}{\sum_{k:I_k \in S_u} I_k + \sum_{k:I_k \in S_d} I_k}$ (FD series weight). $S_u$ and $S_d$ can be considered as elements of metric space which includes various formulae lists with FDCEL (FD Chains Expected Length) measure introduced as follows: $FDCEL\ \Delta(S_u, S_d) = \frac{1}{Card(\Delta)} \sum_{j:\, w_j \in \Delta} w_j |E[l(C^{u,j})] - E[l(C^{d,j})]|$, where $Card(\Delta)$ is the cardinality of set $\Delta$. The properties of the FDCEL metric are detailed in Supplementary Note 2.

The general scheme for FDCEL application is illustrated in Fig. 4.

## 4.3 Data

Both the proposed method as well as already established methods[15,16] (Supplementary Note 3) were applied to the datasets A, B, F, I, G of formulae lists (Supplementary Table 1). Dataset A includes 6 types of samples: CHA, SFA-Ctk, SFA-Pd, SHA-Ctk, SHA-Pd, SRHA; each was analyzed by 7 different instruments[16]. Dataset B consists of 4 types of samples: ESFA, PLFA, SRFA, SRNOM; each was analyzed by 17 instruments[15]. Dataset F includes 3 types of samples: SDOM-Ctk, SDOM-Pd, ISDY; each sample was separated into 4 fractions, which along with the parent sample were measured by the same FTMS instrument[44]. Dataset I includes formulae lists acquired for SRFA with various ionization techniques: ESI (SRFA-r[24], SRFA-t[11], SRFA-u[25]), LDI (SRFA-s[26]), PSI (SRFA-v[25]), PSCI (SRFA-w[25]). Dataset G for geochemical application verification includes various formulae lists: SLNOM-XAD, SLNOM-PPL[27]; PRNOM, AHF-RPP-10, AHF-RPP-11, AHF-RK5P-10, AHF-RK6P-10, AHF-AOP-10[31]; SDOM-GF[33]; PHA-TTL[35]; coal humic acid CHA-h has been described elsewhere[36] and it has been analyzed via Bruker Daltonics 12 Tesla Apex Qe FTICR-MS, housed at the College of Sciences Major Instrumentation Cluster (COSMIC) at Old Dominion University; FA1-Y-15, FA2-Y-15[37].

We selected 13 samples combining datasets A,B, and F as the database entries.

## Authors’ contributions

A.Z. conceived and designed the study. A.Z., E.N.N., and I.V.P. supervised the study. A.S. performed data analysis, developed the algorithm and web application. A.Z. and A.S. drafted the manuscript. All authors read and approved the final manuscript.

## Acknowledgements

This work was supported by Russian Science Foundation (Grant No. 21-47-04405). The authors are grateful to Dr. Natalya A. Kulikova (Lomonosov MSU) for providing soil dissolved organic matter (SDOM) samples and to Prof. Patrick Hatcher (Old Dominion University) for FTICR MS analysis of CHA sample. The authors thank Dr. Oliver Lechtenfeld (Helmholtz-Centre for Environmental Research) for his valuable comments on the manuscript.

## Competing interests

The authors declare no competing interests.

## References

1. Bianchi, T. S. The role of terrestrially derived organic carbon in the coastal ocean: A changing paradigm and the priming effect. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 19473–19481 (2011).

2. Rumpel, C. & Kögel-Knabner, I. Deep soil organic matter—a key but poorly understood component of terrestrial C cycle. *Plant Soil 2010 3381* **338**, 143–158 (2010).

3. Spencer, R. G. M. *et al.* Detecting the signature of permafrost thaw in Arctic rivers. *Geophys. Res. Lett.* **42**, 2830–2835 (2015).

4. Vonk, J. E. *et al.* Activation of old carbon by erosion of coastal and subsea permafrost in Arctic Siberia. *Nat. 2012 4897414* **489**, 137–140 (2012).

5. Lechtenfeld, O. J. *et al.* Molecular transformation and degradation of refractory dissolved organic matter in the Atlantic and Southern Ocean. *Geochim. Cosmochim. Acta* **126**, 321–337 (2014).

6. Schmidt, M. W. I. *et al.* Persistence of soil organic matter as an ecosystem property. *Nat. 2011 4787367* **478**, 49–56 (2011).

7. Hertkorn, N., Harir, M., Koch, B. P., Michalke, B. & Schmitt-Kopplin, P. High-field NMR spectroscopy and FTICR mass spectrometry: Powerful discovery tools for the molecular level characterization of marine dissolved organic matter. *Biogeosciences* **10**, 1583–1624 (2013).

8. Lechtenfeld, O. J., Hertkorn, N., Shen, Y., Witt, M. & Benner, R. Marine sequestration of carbon in bacterial metabolites. *Nat. Commun. 2015 61* **6**, 1–8 (2015).

9. Mustafa, A. *et al.* Stability of soil organic carbon under long-term fertilization: Results from 13C NMR analysis and laboratory incubation. *Environ. Res.* **205**, 112476 (2022).

10. Rakhsh, F., Golchin, A., Beheshti Al Agha, A. & Nelson, P. N. Mineralization of organic carbon and formation of microbial biomass in soil: Effects of clay content and composition and the mechanisms involved. *Soil Biol. Biochem.* **151**, 108036 (2020).

11. Hertkorn, N. *et al.* Natural organic matter and the event horizon of mass spectrometry. *Anal. Chem.* **80**, 8908–8919 (2008).

12. Moran, M. A. *et al.* Deciphering ocean carbon in a changing world. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 3143–3151 (2016).

13. Sleighter, R. L. & Hatcher, P. G. The application of electrospray ionization coupled to ultrahigh resolution mass spectrometry for the molecular characterization of natural organic matter. *J. Mass Spectrom.* **42**, 559–574 (2007).

14. Koistinen, V. M. *et al.* Interlaboratory Coverage Test on Plant Food Bioactive Compounds

and Their Metabolites by Mass Spectrometry-Based Untargeted Metabolomics. *Metab. 2018, Vol. 8, Page 46* **8**, 46 (2018).

15. Hawkes, J. A. *et al.* An international laboratory comparison of dissolved organic matter composition by high resolution mass spectrometry: Are we getting the same answer? *Limnol. Oceanogr. Methods* **18**, 235–258 (2020).

16. Zherebker, A. *et al.* Interlaboratory comparison of humic substances compositional space as measured by Fourier transform ion cyclotron resonance mass spectrometry (IUPAC Technical Report). *Pure Appl. Chem.* **92**, 1447–1467 (2020).

17. Breitling, R., Ritchie, S., Goodenowe, D., Stewart, M. L. & Barrett, M. P. Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics* **2**, 155–164 (2006).

18. Moritz, F., Kaling, M., Schnitzler, J. P. & Schmitt-Kopplin, P. Characterization of poplar metabotypes via mass difference enrichment analysis. *Plant. Cell Environ.* **40**, 1057–1073 (2017).

19. Longnecker, K. & Kujawinski, E. B. Using network analysis to discern compositional patterns in ultrahigh-resolution mass spectrometry data of dissolved organic matter. *Rapid Commun. Mass Spectrom.* **30**, 2388–2394 (2016).

20. Tziotis, D., Hertkorn, N. & Schmitt-Kopplin, P. Kendrick-Analogous Network Visualisation of Ion Cyclotron Resonance Fourier Transform Mass Spectra: Improved Options for the Assignment of Elemental Compositions and the Classification of Organic Molecular Complexity: *http://dx.doi.org/10.1255/ejms.1135* **17**, 415–421 (2011).

21. Forcisi, S. *et al.* Liquid chromatography–mass spectrometry in metabolomics research: Mass analyzers in ultra high pressure liquid chromatography coupling. *J. Chromatogr. A* **1292**, 51–65 (2013).

22. Smirnov, K. S. *et al.* Challenges of metabolomics in human gut microbiota research. *Int. J. Med. Microbiol.* **306**, 266–279 (2016).

23. Sarycheva, A. FDS application. https://nommass.com/ (2022).

24. Kew, W., Blackburn, J. W. T. & Uhrín, D. Response to Comment on 'laser Desorption/Ionization Coupled to FTICR Mass Spectrometry for Studies of Natural Organic Matter'. *Anal. Chem.* **90**, 5968–5971 (2018).

25. Kim, D., Lee, J., Kim, B. & Kim, S. Optimization and Application of Paper-Based Spray Ionization Mass Spectrometry for Analysis of Natural Organic Matter. *Anal. Chem.* **90**, 12027–12034 (2018).

26. Blackburn, J. W. T., Kew, W., Graham, M. C. & Uhrín, D. Laser Desorption/Ionization Coupled to FTICR Mass Spectrometry for Studies of Natural Organic Matter. *Anal. Chem.* **89**, 4382–4386 (2017).

27. Zherebker, A. Y. *et al.* Extraction of humic substances from fresh waters on solid-phase cartridges and their study by Fourier transform ion cyclotron resonance mass spectrometry.

*J. Anal. Chem. 2016 714* **71**, 372–378 (2016).

28. Thurman, E. M. & Malcolm, R. L. Preparative isolation of aquatic humic substances. *Environ. Sci. Technol.* **15**, 463–466 (2002).

29. Dittmar, T., Koch, B., Hertkorn, N. & Kattner, G. A simple and efficient method for the solid-phase extraction of dissolved organic matter (SPE-DOM) from seawater. *Limnol. Oceanogr. Methods* **6**, 230–235 (2008).

30. Perminova, I. V. *et al.* Molecular mapping of sorbent selectivities with respect to isolation of arctic dissolved organic matter as measured by fourier transform mass spectrometry. *Environ. Sci. Technol.* **48**, 7461–7468 (2014).

31. Perminova, I. V. *et al.* The Structural Arrangement and Relative Abundance of Aliphatic Units May Effect Long-Wave Absorbance of Natural Organic Matter as Revealed by 1 H NMR Spectroscopy. *Environ. Sci. Technol.* **52**, 12526–12537 (2018).

32. Brown, A., McKnight, D. M., Chin, Y. P., Roberts, E. C. & Uhle, M. Chemical characterization of dissolved organic material in Pony Lake, a saline coastal pond in Antarctica. *Mar. Chem.* **89**, 327–337 (2004).

33. Filippova, O. I., Kholodov, V. A., Safronova, N. A., Yudina, A. V. & Kulikova, N. A. Particle-Size, Microaggregate-Size, and Aggregate-Size Distributions in Humus Horizons of the Zonal Sequence of Soils in European Russia. *Eurasian Soil Sci. 2019 523* **52**, 300–312 (2019).

34. Source Materials for IHSS Samples. https://humic-substances.org/source-materials-for-ihss-samples/.

35. Zherebker, A. Y. Study of the structure of humic substances by isotopic exchange and mass spectrometry. (Moscow State University 'M. V. Lomonosov', 2017).

36. Zherebker, A. Y., Kostyukevich, Y. I., Kononikhin, A. S., Nikolaev, E. N. & Perminova, I. V. Molecular compositions of humic acids extracted from leonardite and lignite as determined by Fourier transform ion cyclotron resonance mass spectrometry. *Mendeleev Commun.* **26**, 446–448 (2016).

37. Zherebker, A. *et al.* The Molecular Composition of Humic Substances Isolated From Yedoma Permafrost and Alas Cores in the Eastern Siberian Arctic as Measured by Ultrahigh Resolution Mass Spectrometry. *J. Geophys. Res. Biogeosciences* **124**, 2432–2445 (2019).

38. Kleber, M. & Lehmann, J. Humic Substances Extracted by Alkali Are Invalid Proxies for the Dynamics and Functions of Organic Matter in Terrestrial and Aquatic Ecosystems. *J. Environ. Qual.* **48**, 207–216 (2019).

39. Olk, D. C. *et al.* Environmental and Agricultural Relevance of Humic Fractions Extracted by Alkali from Soils and Natural Waters. *J. Environ. Qual.* **48**, 217–232 (2019).

40. Kellerman, A. M., Dittmar, T., Kothawala, D. N. & Tranvik, L. J. Chemodiversity of dissolved organic matter in lakes driven by climate and hydrology. *Nat. Commun. 2014 51* **5**, 1–8 (2014).

41. Zherebker, A. *et al.* Separation of Benzoic and Unconjugated Acidic Components of Leonardite Humic Material Using Sequential Solid-Phase Extraction at Different pH Values as Revealed by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry and Correlation Nuclear Magnetic Resonance Spectroscopy. *J. Agric. Food Chem.* **66**, 12179–12187 (2018).

42. Kunenkov, E. V. *et al.* Total Mass Difference Statistics Algorithm: A New Approach to Identification of High-Mass Building Blocks in Electrospray Ionization Fourier Transform Ion Cyclotron Mass Spectrometry Data of Natural Organic Matter. *Anal. Chem.* **81**, 10106–10115 (2009).

43. Perminova, I. V. *et al.* Signatures of Molecular Unification and Progressive Oxidation Unfold in Dissolved Organic Matter of the Ob-Irtysh River System along Its Path to the Arctic Ocean. *Sci. Reports 2019 91* **9**, 1–16 (2019).

44. Zherebker, A. *et al.* Optical Properties of Soil Dissolved Organic Matter Are Related to Acidic Functions of Its Components as Revealed by Fractionation, Selective Deuteromethylation, and Ultrahigh Resolution Mass Spectrometry. *Environ. Sci. Technol.* **54**, 2667–2677 (2020).

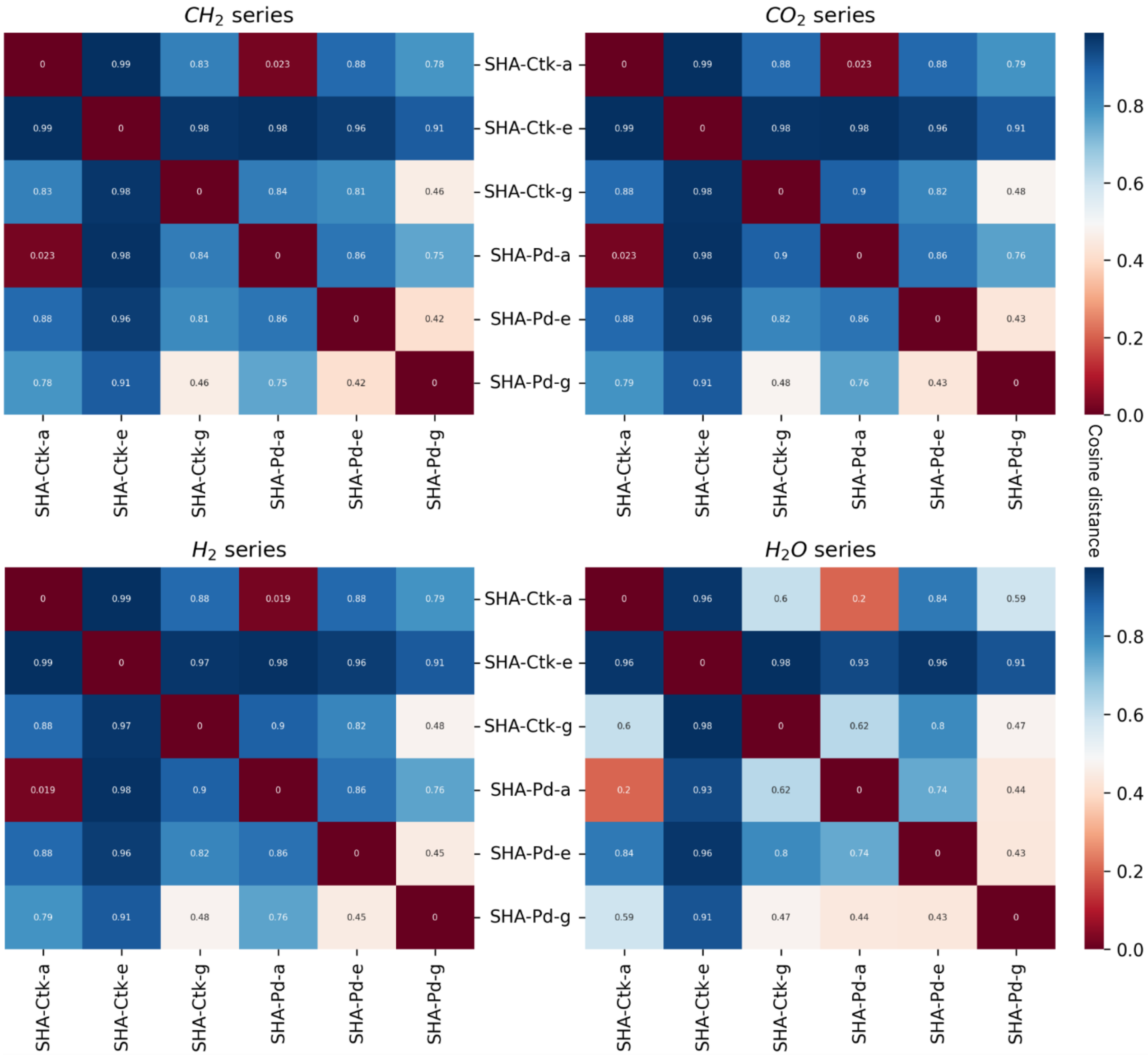

Figure 1: **Cosine distances between targeted series for the samples SHA-Ctk and SHA-Pd detected at three laboratories. These series appear relatively similar for of SHA-Pd-e and SHA-Pd-g. Misleading cases: CH2/CO2/H2/H2O series of SHA-Ctk-a are closer to corresponding series of SHA-Pd-a than to their counterpart subsets of SHA-Ctk-e.**

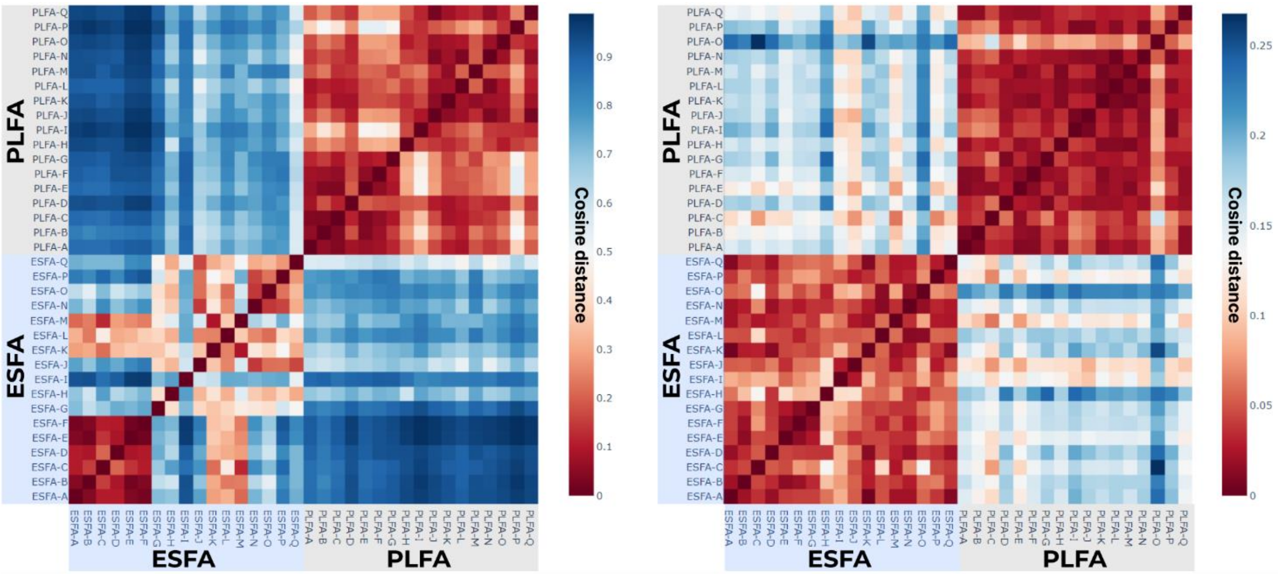

Figure 2: **Pairwise comparison in terms of Cosine distance of formulae lists detected for ESFA and PLFA datasets (left panel) and derived FDs counts (right panel). In case of formulae list comparison (left panel), ESFA-N,O,P,Q and ESFA-J,H,G are distinct from ESFA-A,B,C,D,F; ESFA-M is distinct from ESFA-N,O,P,Q, and ESFA-I is distinct from the rest formulae lists detected for ESFA sample.**

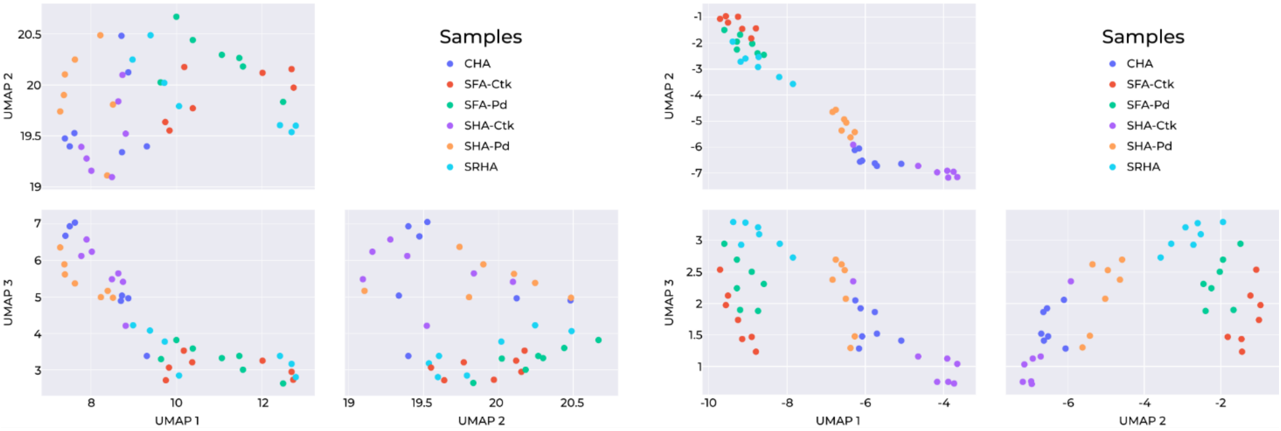

Figure 3: **UMAP applied to formulae lists (left panel) and FDs counts (right panel).**

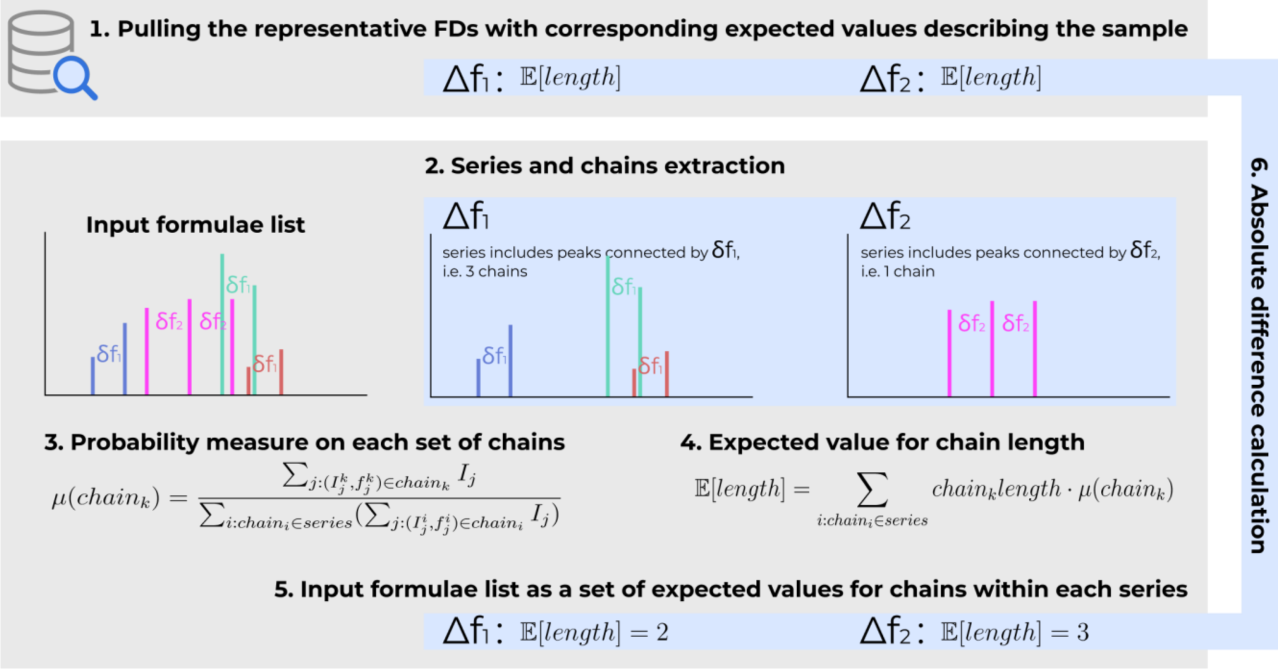
1. Pulling the representative FDs with corresponding expected values describing the sample
Δf₁: 𝔼[length]
Δf₂: 𝔼[length]
2. Series and chains extraction
Input formulae list
δf₁
δf₂
δf₂
δf₁
δf₁
Δf₁
series includes peaks connected by δf₁, i.e. 3 chains
δf₁
δf₁
δf₁
Δf₂
series includes peaks connected by δf₂, i.e. 1 chain
δf₂
δf₂
3. Probability measure on each set of chains
$\mu(chain_k) = \frac{\sum_{j:(I_j^k, f_j^k)\in chain_k} I_j}{\sum_{i:chain_i \in series}(\sum_{j:(I_j^i, f_j^i)\in chain_i} I_j)}$
4. Expected value for chain length
$\mathbb{E}[length] = \sum_{i:chain_i \in series} chain_k length \cdot \mu(chain_k)$
5. Input formulae list as a set of expected values for chains within each series
Δf₁: 𝔼[length] = 2
Δf₂: 𝔼[length] = 3
6. Absolute difference calculation

Figure 4: **FDCEL comparison (as implemented in the "compare against database" feature in FDS application[23]) of an input formulae list against a single sample within the database.**

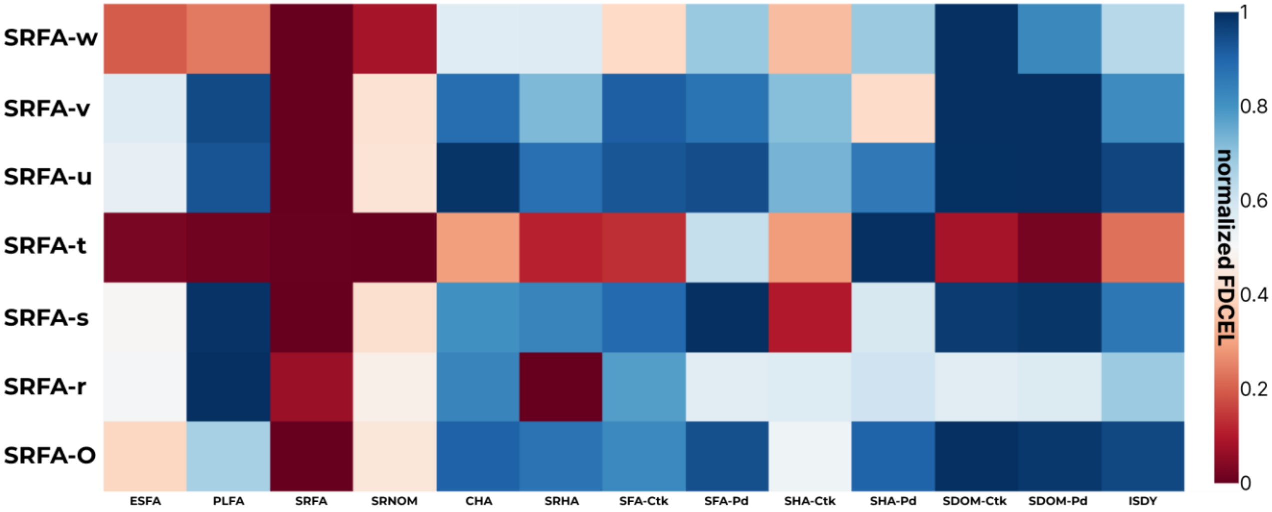

Figure 5: **The comparison of formulae lists acquired via HRMS with various mild ionization techniques as well as SRFA-O (the formulae list from the dataset employed for SRFA sample important FDs extraction) against the database (Supplementary Interactive Fig. 1).**

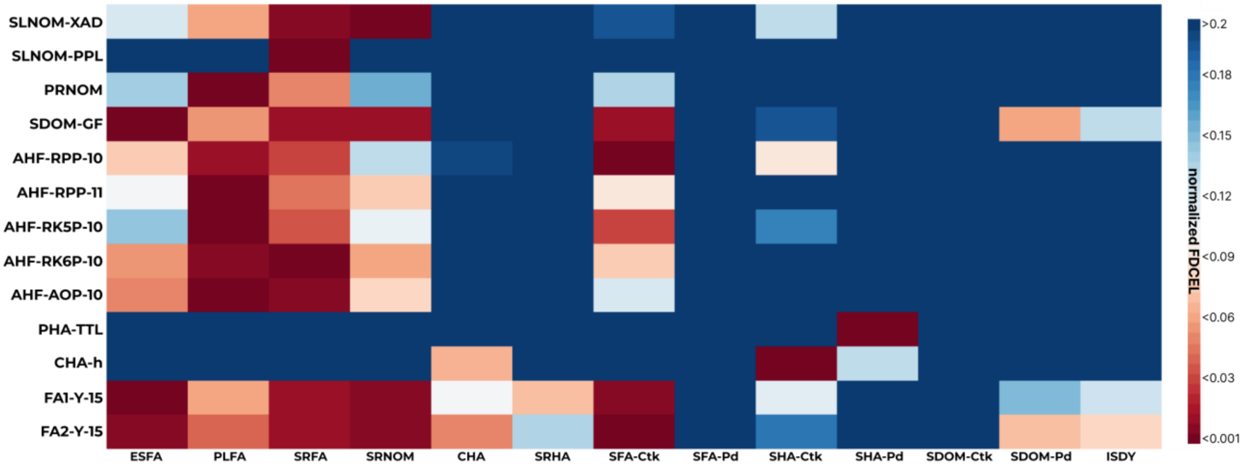

Figure 6: **The comparison of formulae lists acquired via HRMS for 13 samples (dataset G) against the database provides some geochemical insights on sample origin and source of organic compounds by pairwise similarity analysis. For details see an interactive plot (Supplementary Interactive Fig. 4).**