2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669289

Graph Theoretic Approach for the Analysis of Comprehensive Mass-Spectrometry (MS/MS) Data of Dissolved Organic Matter

Muhammad Usman Tariq*, Dennys Leyva†, Francisco Alberto Fernandez Lima‡ and *Fahad Saeed§
School of Computing and Information Sciences
Florida International University
Miami, FL 33199 USA
Email: *mtari008@fiu.edu, †dleyv008@fiu.edu, ‡fernandf@fiu.edu, §fsaeed@fiu.edu

***Abstract*—Dissolved organic matter (DOM) is a highly complex mixture of organic substances found in aquatic ecosystems. This mixture results from the degradation of primary producers within the ecosystem, groundwater, and the surrounding terrestrial sources. Understanding the chemical structure of DOM is crucial to assessing its impact on aquatic ecosystems. Although multiple studies have addressed the complexity of DOM, the molecular structure of this set of compounds remains unclear. In this work, we present a novel computational framework "Graph-DOM," to assess the comprehensive fragmentation data obtained from the analysis of DOM using the Data Independent Fragmentation strategy with ESI-FT-ICR MS/MS enabling better understanding of the structural complexity of DOM. Graph-DOM uses graph algorithms to dissect a compiled output file obtained from processing hundreds of ultra-high-resolution fragment spectra. Over half a million ordered fragmentation pathways were computed for 764 isolated precursor ions assuming up to seven vector segments categorized as neutral losses (CH2, CH3, O, CH4, H2O, CO, and CO2). Families of structurally related molecules were identified using pathway overlaps, and output files compatible with network visualization software (e.g., Cytoscape) were also generated. Graph-DOM is able to efficiently process all the pathways to discover families within only a few minutes with adjustable parameters for overlap length of fragmentation pathways as well as configuring low abundance CHOS, CHON, and CHONS compounds. Graph-DOM is available at https://github.com/Usman095/Graph-DOM.**

***Index Terms*—dissolved organic matter, dynamic programming, regularity chains, HPC.**

I. Introduction

Dissolved organic matter (DOM) is a complex mixture of organic compounds in aquatic ecosystems and mainly consists of carbon, hydrogen, and oxygen with trace amounts of other elements, e.g., Nitrogen and Sulphur. DOM is formed primarily from the degradation of primary aquatic and terrestrial producers. DOM has a significant biogeochemical impact and influence on many environmental processes, including transportation of contaminants, ecological processes, and water treatment [1]. A vast number of studies have been performed on DOM, reporting numerous new molecular formulas and novel features [2], [3], [4]. However, molecular structures of most of the molecules are still unknown due to the diversity in chemical features and complexity of the mixture [5].

Advancements in analytical techniques have helped move the field forward towards better understanding the molecular structures within the mixture [2], [6]. In particular, Fourier transform ion cyclotron Resonance-Mass Spectrometry (FT-ICR MS) and Quadrupole Time-of-Flight Mass Spectrometry (Q-TOF-MS) have aided much in the characterization of DOM due to their high-resolution capabilities and flexibility toward coupling with separation techniques. Statistical methods like linear regression have identified similarities among highly variable DOM samples collected from different sources [7]. Moreover, network analysis techniques have been used to determine the number of chemical transformations between different peaks in a sample [8]. In [9], computational methods are used to determine the 3D configuration of DOM. However, molecular structures of compounds in DOM are mostly unknown. There is a critical need to develop fast, efficient computational frameworks capable of deconvoluting complex MS2 spectra and quickly processing and extracting useful structural information.

In the present work, we design and implement graph-based algorithms for identifying fragmentation pathways of related organic molecules in DOM mass-spectrometry data and identify structural families of molecules based on partial/complete overlap of fragmentation pathways. DOM fragmentation data were collected in a Solarix X-XR 9T spectrometer by sequencing 1 Da quadrupole isolation every other m/z and using typical CID energies of 15-20 eV. 2D MS/MS data included exact masses of precursors and fragments, chemical formulae, and corresponding intensities. A customized directed acyclic-graph (DAG) based algorithm was developed to find all the fragmentation pathways and identify regular sequences of precursors with a shared core fragment. The framework provides functionalities to generate fragmentation pathways based on precursors, core fragments, and neutral losses and to identify structural commonalities in DOM-based on similarities in fragmentation patterns of precursors sharing

*Corresponding Author

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on March 25,2022 at 15:25:39 UTC from IEEE Xplore. Restrictions apply.

the same core fragment. To the best of our knowledge, this is the first time structural similarities using comprehensive fragmentation data have been assessed by developing custom graph analysis algorithms.

The rest of the paper is as follows: In section II, we present the design of the algorithm for analyzing the data. In section III, we present sample preparation, sample ionization, MS/MS analysis, and data processing. In section IV, we show results using DOM fragmentation data at 1 Da quadrupole isolation every other m/z. Section V, provides discussion of the results and section VI concludes this paper.

II. Graph-DOM

A. Constructing Pathway DAG

MS/MS is first preprocessed, and peaks are filtered to determine the intermediate fragments. All the fragmentation pathways and the associated core-fragments are identified for a given set of neutral losses by constructing a breadth-first search tree starting from the nominal mass (±1 Da) considering CH2O O, CH4, H2O, CO, and CO2 as neutral losses. The directed acyclic graph (DAG) is constructed by first selecting a precursor peak as the root node; then, seven possible children are considered at each step, each corresponding to one of the neutral losses. A child is added to the DAG if a peak is found in the spectrum with m/z value of parent minus neutral loss with an m/z tolerance of 1mDa. The process is repeated recursively till no more children can be added, at which point the child node (leaf node) is output as the core-fragment. A pathway is a path from the root node of the DAG to the leaf node with the possibility of multiple paths leading to the same leaf node. A sample spectrum for the nominal mass of 391 Da with the corresponding DAG and multiple pathways corresponding to different core fragments are shown in figure 1.

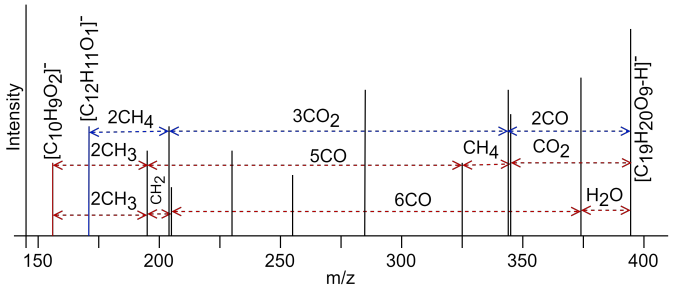

Fig. 1. Sample spectrum of nominal mass 391 Da. The spectrum is analyzed by selecting a precursor ion at the nominal mass (±1 Da) here $[C_{19}H_{20}O_9]^-$ and then constructing a directed acyclic graph (DAG) using all the neutral losses. Above, three possible pathways are shown with two (red) leading to core-fragment $[C_{10}H_9O_2]^-$ and one (blue) leading to core-fragment $[C_{12}H_{11}O]^-$. As can be seen, the paths can overlap, and multiple paths can lead to a single core-fragments generating a DAG as shown in figure 2.

The corresponding DAG for the given spectrum is shown in figure 2.

MS/MS spectra are converted into a breadth-first tree using given neutral losses and the threshold on the number of repeated neutral losses. The tree is searched using BFS to

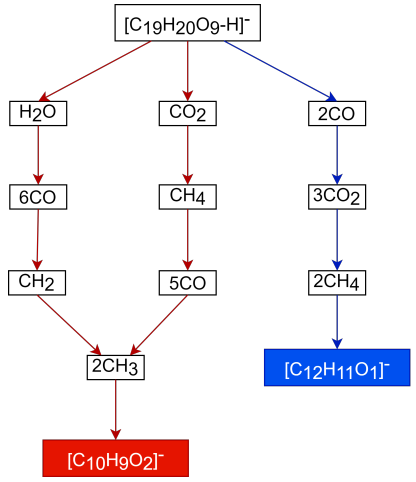

Fig. 2. DAG corresponding to the spectrum in figure 1. The branches are constructed by adding neutral losses such that a fragment of m/z of precursor minus the neutral loss is found in the spectrum. The DAG is grown recursively and splitting/joining branches wherever the paths overlap or split. As can be seen, two paths (red) start out separate but then join at the last neutral loss resulting in a common core fragment. Multiple fragmentation pathways leading to the same core fragment provide evidence for the presence of structural isomers. On the other hand (blue) might only have one branch leading to a unique core fragment.

Precursor	Intermediate Fragments
C13H15O6	C12H15O4, C10H7O4, C8H3O4
C13H15O6	C11H7O6, C9H7O4, C8H3O4
C11H9O8	C10H9O6, C9H9O4, C9H9O3, C8H5O3
C11H9O8	C10H9O6, C9H9O4, C9H9O3, C9H9O2
C11H9O8	C11H9O6, C10H9O4, C9H9O2

Fig. 3. Pathways constructed by traversing the DAG from the root (precursor) node to all the leaf (core-fragment) nodes. Note that each unique pathway is listed separately as the intermediate fragment sequence is important for identifying related families using fragmentation pathway overlap.

explore all the possible pathways up-to-the core fragments (leaves). This recursive algorithm returns all the possible pathways and core fragments interpreted from the given fragment spectra. The algorithm is generalized to run on multiple precursors and determine all the shared core fragments among precursors, shared precursors among core fragments, the similarity between the pathways of the core fragments and precursor pairs.

B. Traversing Pathway DAG

Once the DAG is constructed, it is traversed from the root (precursor) node to each leaf (core-fragment) node. Each unique pathway is listed where a pathway consists of Precursor, Core-Fragment, Neutral Losses sequence, Intermediate Fragments sequence, and the Precursor Mass. The resultant pathways are sorted according to their precursor mass in increasing order, as shown in figure 3. Multiple pathways can exist for a given precursor or even a precursor/core-fragment pair. For such cases, each pathway is listed separately with a unique identifier called "Pathway ID".

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on March 25,2022 at 15:25:39 UTC from IEEE Xplore. Restrictions apply.

C. Constructing Family DAG

Using the pathways identified in figure 3, another DAG, called Family DAG, is constructed where pathways are considered nodes. An edge from the smaller pathway P_1 (pathway with smaller precursor mass) is added to larger pathway P_2 if 1) the precursor of P_1 is same as the first intermediate fragment of P_2 and 2) intermediate fragments of P_1 and P_2 match up to a given threshold called "overlap value". One such example is shown in figure 4 where 4 pathways will get connected in the Family DAG.

Precursor	Fragmentation Pathway								
C15H15O7				C14H13O6	C13H11O6	C12H7O6	C10H5O5	C9H3O5	C8H3O3
C16H15O9			C15H15O7	C14H13O6	C13H11O6	C12H7O6	C10H5O5	C9H3O5	C8H3O3
C17H15O11		C16H15O9	C15H15O7	C14H13O6	C13H11O6	C12H7O6	C12H5O5		
C18H19O12	C17H15O11	C16H15O9	C15H15O7	C14H13O6	C13H13O4				

Fig. 4. Pathways that will be connected by edges in the Family DAG. As can be seen, for every larger precursor, the next smaller precursor is the same as its first intermediate fragment. Matching precursors and intermediate fragments are highlighted with the same color.

Note that depending on the overlap value, a pathway can be part of more than one family, as shown in figure 5. A set of root nodes is maintained separately as these nodes don't have any incoming edges, and that's where the traversal begins for identifying families.

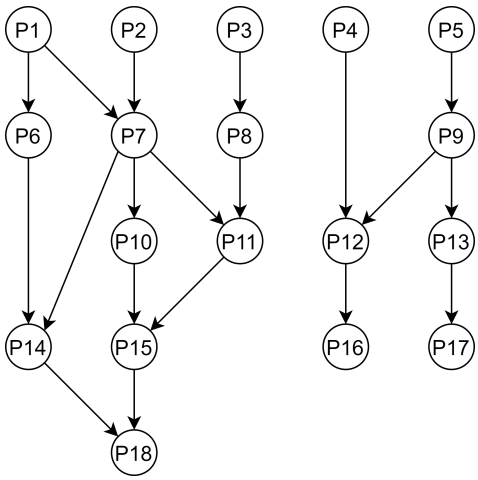

Fig. 5. Family DAG constructed from identified pathways. Each node represents a pathway while a path from the root node to a leaf node comprises a family.

The family DAG is traversed from all root nodes to find all possible paths to the leaf nodes. Each distinct path represents a family. A family of precursors can share numerous distinct sets of pathways that overlap. In this case, it is only counted as one family. The number of pathways is referred to as the confidence value as they are directly proportional to the relevance of precursors.

III. Materials and Methods

A. Sample Preparation

Surface water was collected from Pantanal (PAN) National Park – SE Brazil, one of the largest subtropical and bio-diverse freshwater wetlands in the world. Surface water was subjected to a Solid Phase Extraction (SPE) Procedure following the protocol described in [10]. The SPE-DOM methanol extract was diluted 10 times in Optima LC-MS grade denatured ethanol (Fisher Scientific, Pittsburgh, PA) prior to direct infusion in ESI-FT-ICR MS.

B. Sample Ionization

An electrospray ionization source (ESI) Apollo II ESI design (Bruker Daltonics, Inc., MA) was used in negative ion mode for all experiments. Sample solutions were introduced into the nebulizer at a rate of 360 $\mu L/h$ using a syringe pump. Typical operating conditions were 3000–3500 V capillary voltage, 4 L/min dry gas flow rate, 1.0 bar nebulizer gas pressure, and a dry gas temperature 180 °C.

C. ESI-TIMS-FT-ICR MS/MS analysis

An ESI-FT-ICR MS Solarix X-XR 9T spectrometer equipped with an infinity ICR cell was utilized. A broadband mass spectrum (115 scans) of the SPE-DOM sample was collected in the mass range 150-3000 with a resolution of 4.0E06 at m/z 400. A Continuous Accumulation of Selected Ions (CASI) was used as a Data Independent Fragmentation (DIF) method. Briefly, ions at odd nominal masses were sequentially isolated (1 Da window) in the quadrupole (m/z range 261-477), accumulated for 3 s in the collision cell, and fragmented by Collision Induced Dissociation (CID) using typical 14-15 V collision voltages. A total of 50 scans were co-added for each nominal mass MS/MS (segment), and ten single segments were stitched in each experiment using Serial LC run mode. Arginine cluster ions were utilized during instrument tuning for the external mass calibration of the broadband MS1 spectrum, and MS/MS spectra were externally calibrated using exact masses of known neutral losses in typical DOM [7].

D. Data Processing

MS and MS/MS data were processed using Data Analysis (v. 5.2, Bruker Daltonics, CA). The assignment of chemical formulae was conducted using Composer software (version 1.0.6, Sierra Analytics, CA, USA) and confirmed with Data Analysis (version 5.2, Bruker Daltonics). A formula constraint of $C_{1-100}H_{1-200}N_{0-4}O_{0-25}S_{0-2}$ was applied, and both odd and even electron configurations and mass errors below 1 ppm were allowed. An excel file containing the mass of isolated precursors and fragments, chemical formula, and abundance was created as input files for the computation of fragmentation pathways and structural families.

IV. Results

Using Graph-DOM, we calculate all possible pathways as well as core-fragments using the comprehensive fragmentation data. Similarly, Graph-DOM is able to identify all the families of related precursors along with their confidence value.

A. Experimental Setup

The experiments were run on a 24 core Xeon server with 48 GB of available memory. Jupyter Notebook was used with Python 3.7 and NumPy and Pandas data processing packages.

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on March 25,2022 at 15:25:39 UTC from IEEE Xplore. Restrictions apply.

B. Experiments

Using Graph-DOM, we identify all $0.6M$ possible pathways for comprehensive fragmentation data at a mass range of $261 - 477$ Da consisting of 764 precursors as shown in figure 6. Depending on the mass spectrometer settings, and the chemical properties of the sample, the number of identified fragmentation pathways can vary significantly and range up to more than $40k$ per sample. This also emphasizes the abundance of isomeric families in DOM.

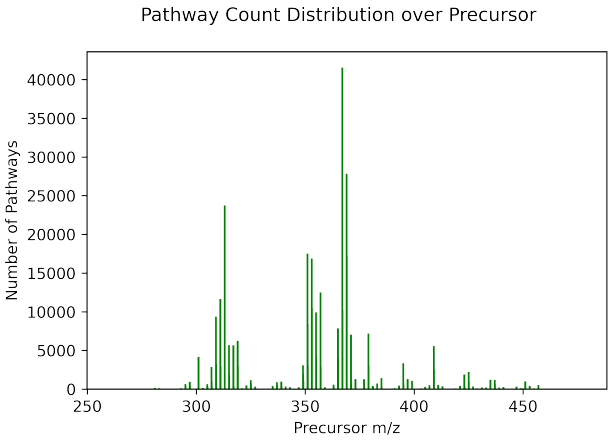

Fig. 6. Distribution of number of pathway w.r.t. to the precursor m/z. As shown in the plot, depending on the mass-spectrometer settings, and the chemical properties of the molecules, number of fragmentation pathways can vary vastly up to more than 40k per sample.

Similarly, all the core fragments (9021) for each precursor are also calculated. Figure 7 shows the distribution of the number of cores over precursor m/z. In this context, a core fragment refers to either the stable backbone structure of the molecule that does not fragment further or an intermediate fragment that could not be further fragmented due to mass-spectrometer limitations or the lower concentration of the compound within the sample.

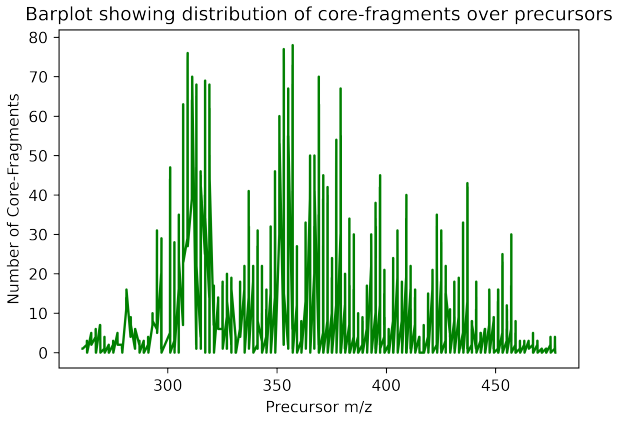

Fig. 7. Distribution of number of core-fragments w.r.t. to the precursor m/z.

By constructing the family DAG, we identify 685 unique families. Figure 8 shows the distribution of the number of families over the family size (number of precursors in a family) for varying overlap lengths.

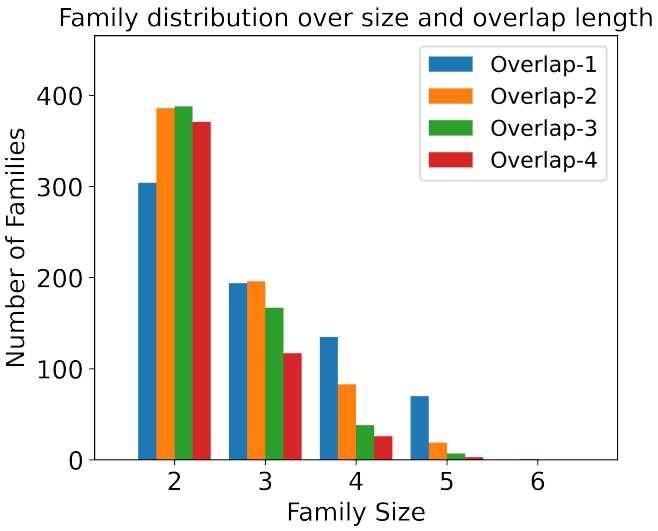

Fig. 8. Distribution of number of families w.r.t the family size and overlap length used to identify families.

In figure 9, we show the ordered distribution of family confidence.

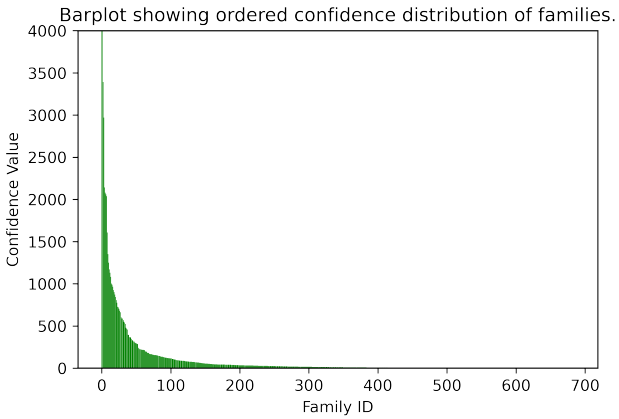

Fig. 9. Ordered distribution of confidence value for each unique family.

V. Discussion

Using Graph-DOM, we are able to process precursors for the entire range of m/z ($261 - 471$ Da) in only a few minutes, enabling swift analysis of fragmentation data and gaining structural insights for DOM. For the analysis, we generate nearly half a million distinct fragmentation pathways and based on the overlap of those pathways, precursor families of related compounds are generated. We also provide the confidence level of the generated families based on how many sets of pathways among the families overlap. The overlap length is an adjustable parameter that can be set to any positive value. For the data set used, our algorithm provides

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on March 25,2022 at 15:25:39 UTC from IEEE Xplore. Restrictions apply.

72% precursor coverage, 27% core-fragment coverage, and 42% intermediate fragment coverage, where the coverage is defined as the percentage of precursors, core-fragments, or intermediate fragments appearing in the families vs. the total number in the original data. Finally, Graph-DOM is capable of generating Cytoscape [11] input files for all the connected precursors through different families.

VI. Conclusion

In this paper, we present a graph-theoretic approach to analyzing comprehensive fragmentation data for dissolved organic matter (DOM). DOM is a complex mixture of organic compounds resulting from the decomposition of primary producers within the aquatic ecosystem or surrounding terrestrial sources. Understanding the features, properties and molecular structures of various compounds in DOM is vital for assessing its impact on the ecosystem. Although countless efforts have been made to discover novel molecules and features, the molecular structure of various compounds is still unknown. We present the first of its kind analytical tool that can assist in uncovering underlying structures of molecules by analyzing comprehensive fragmentation data for DOM. By constructing a directed acyclic graph of nearly half a million pathways, we are able to construct precursor families in which molecules differ only by a single functional group providing quantitative evidence for the presence of structural similarities within DOM. Graph-DOM is highly efficient and only takes a few minutes to process all the fragmentation pathways and generate families. Moreover, it can generate network input files for Cytoscape to enable users to visualize the resultant network of related molecules. The Jupyter Notebook for the framework is available on GitHub at https://github.com/Usman095/Graph-DOM.

VII. Acknowledgements

This research was supported by the National Science Foundations (NSF) under the Award Numbers CAREER OAC-1925960 and NIH R01GM134384. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation.

References

[1] L. Kaplan and R. Cory, "Dissolved organic matter in stream ecosystems: forms, functions, and fluxes of watershed tea," in *Stream ecosystems in a changing environment*. Elsevier, 2016, pp. 241–320.

[2] N. Hertkorn, M. Harir, B. P. Koch, B. Michalke, and P. Schmitt-Kopplin, "High-field nmr spectroscopy and fticr mass spectrometry: powerful discovery tools for the molecular level characterization of marine dissolved organic matter," *Biogeosciences*, vol. 10, no. 3, pp. 1583–1624, 2013.

[3] M. Zark, J. Christoffers, and T. Dittmar, "Molecular properties of deep-sea dissolved organic matter are predictable by the central limit theorem: Evidence from tandem ft-icr-ms," *Marine Chemistry*, vol. 191, pp. 9–15, 2017.

[4] N. Hertkorn, M. Harir, K. M. Cawley, P. Schmitt-Kopplin, and R. Jaffé, "Molecular characterization of dissolved organic matter from subtropical wetlands: a comparative study through the analysis of optical properties, nmr and fticr/ms," *Biogeosciences*, vol. 13, no. 8, pp. 2257–2277, 2016.

[5] N. W. Green, E. M. Perdue, G. R. Aiken, K. D. Butler, H. Chen, T. Dittmar, J. Niggemann, and A. Stubbins, "An intercomparison of three methods for the large-scale isolation of oceanic dissolved organic matter," *Marine Chemistry*, vol. 161, pp. 14–19, 2014.

[6] R. Jaffé, Y. Yamashita, N. Maie, W. Cooper, T. Dittmar, W. Dodds, J. Jones, T. Myoshi, J. Ortiz-Zayas, D. Podgorski *et al.*, "Dissolved organic matter in headwater streams: compositional variability across climatic regions of north america," *Geochimica et Cosmochimica Acta*, vol. 94, pp. 95–108, 2012.

[7] M. Zark and T. Dittmar, "Universal molecular structures in natural dissolved organic matter," *Nature communications*, vol. 9, no. 1, pp. 1–8, 2018.

[8] K. Longnecker and E. B. Kujawinski, "Using network analysis to discern compositional patterns in ultrahigh-resolution mass spectrometry data of dissolved organic matter," *Rapid Communications in Mass Spectrometry*, vol. 30, no. 22, pp. 2388–2394, 2016.

[9] E. A. Vialykh, G. McKay, and F. L. Rosario-Ortiz, "Computational assessment of the three-dimensional configuration of dissolved organic matter chromophores and influence on absorption spectra," *Environmental Science & Technology*, vol. 54, no. 24, pp. 15 904–15 913, 2020.

[10] T. Dittmar, B. Koch, N. Hertkorn, and G. Kattner, "A simple and efficient method for the solid-phase extraction of dissolved organic matter (spedom) from seawater," *Limnology and Oceanography: Methods*, vol. 6, no. 6, pp. 230–235, 2008.

[11] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on March 25,2022 at 15:25:39 UTC from IEEE Xplore. Restrictions apply.