UltraMassExplorer - a browser-based application for the evaluation of high-resolution mass spectrometric data

*Tim Leefmann[1], Stephan Frickenhaus[1,2], Boris P. Koch[1,2]**

[1]*Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany*

[2]*University of Applied Sciences, An der Karlstadt 8, 27568 Bremerhaven, Germany*

[3]*University of Bremen, Centre for Industrial Mathematics, Bibliothekstraße 5, 28359 Bremen, Germany*

Abstract

Rational: High-resolution mass spectrometry (HRMS) with high sample throughput has become an important analytical tool for the analysis of highly complex samples and data processing has become a major challenge for the user community. Evaluating direct-infusion HRMS data without automated tools for batch processing can be a time consuming step in the analytical pipeline. Therefore, we developed a new browser-based software tool for processing HRMS data.

Methods: The software named UltraMassExplorer (UME) was written in the R programming language using the shiny library to build the graphical user interface. The performance of the integrated formula library search algorithm was tested using HRMS data derived from analyses of up to 50 extracts of marine dissolved organic matter.

Results: The software supports the processing of lists of calibrated masses of neutral, protonated, or deprotonated molecules, respectively, with masses of up to 700 Da and a mass

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/rcm.8315

accuracy < 3 ppm. In the performance test, the number of assigned peaks per second increased with number of submitted peaks and reached a maximum rate 4,745 assigned peaks per second.

Conclusions: UME offers a complete data evaluation pipeline comprising a fast molecular formula assignment algorithm allowing for the swift reanalysis of complete datasets, advanced filter functions, and the export of data, metadata, and publication-quality graphics. Unique to UME is a fast and interactive connection between data and its visual representation. UME provides a new platform enabling an increased transparency, customization, documentation and comparability of datasets.

Introduction

A growing number of researchers in the fields of metabolomics as well as natural organic matter (NOM) and petroleum research apply high resolution Fourier Transform mass spectrometry (FT-ICR-MS[1,2]; Orbitrap[3,4]) for the chemical characterization of highly complex organic mixtures [e.g. 5-7]. Data processing is a challenging and critical step and often the bottleneck in the analytical pipeline as the time spent on data processing and evaluation can substantially exceed the time for sample preparation and spectra acquisition. The typical data evaluation comprises the molecular formula assignment process, data quality assessment, data selection, visualization, export, and documentation. Ideally, an integrated, user-friendly software environment should enable non-FT-MS specialists to perform data evaluation.

In most previous studies, individual parts of the evaluation process such as the development and improvement of molecular formula assignment [8-15] or visualization approaches [16-20] have been addressed. The existing approaches differ, however, in their scope of application, performance, and

degree of transparency and accessibility. For the efficient assignment of molecular formulas Kujawinski and Behn[12] developed the compound identification algorithm (CIA) in MatLab based on *a priori* chosen functional groups as molecular building blocks and calculation via nested loops. The algorithm of Kujawinski and Behn was recently integrated into the software Formularity[10]. Formularity is based on an extensive molecular formula database and, in addition, uses the isotope pattern algorithm (IPA) to account for halogenated compounds. An algorithm similar to CIA that circumvents an *a priori* selection of molecular building blocks, was developed by Kunenkov and colleagues[15] for their FIRAN software. Kind and Fiehn[8] published the 7GR software, an Microsoft Excel based evaluation pipeline, with a molecular formula calculator (HR) coded in C++ by Joerg Lau and automated calculation of a number of molecular parameters for further data evaluation. Another existing algorithm developed by Tziotis and colleagues[18] evaluates mass differences of ions via network analysis (netcalc) for the assignment of molecular formulas. Green and Perdue[9] generated a fast formula assignment algorithm based on low-mass moieties (CHOFIT) in Pascale that was optimized partly by replacing nested loops of the classical combinatory approaches.

The visualization of complex high-resolution mass spectrometric datasets is most commonly achieved using the van Krevelen plot.[21,22] Other approaches include the Kendrick mass defect spectrum[13,23], DBE versus O contour diagrams combined with DBE-O frequency plots[16], ratio plots to compare relative peak magnitude changes between two samples[20], mass edited H/C ratios in connection with van Krevelen diagrams[24], or the carbon vs mass (CvM) plots[17]. Most recently Kew and colleagues[19] developed an interactive version of the van Krevelen diagram (*i*-van Krevelen) for the graphical evaluation of high resolution mass spectrometric data also including scripts for formula generation and assignment. A reduced version of *i*-van Krevelen including CHO formula assignment and visualization is also available online through the GitHub repository[25]. The basic Python code of *i*-van Krevelen uses monoisotopic peak assignments, isotopologue peak assignments, and remaining

unassigned, but detected, peaks as input data. The interactive van Krevelen diagrams are visualized in a web browser. Interactive features of the diagrams include zooming, hovering, and data selection by brushing, and a linkage to the ChemSpider database for compound searches.

Summarized, almost all currently available software solutions for evaluating FT-MS data focus on specific aspects of the evaluation process. To the best of our knowledge the commercially available AutoVectis software suite, developed by Kilgour and colleagues[26,27], Composer by Sierra Analytics[28], and the PetroOrg[29] software, developed by the National High Magnetic Field Laboratory at the Florida State University and the Future Fuels Institute[30] currently represent the only approaches integrating a complete data evaluation process for high-resolution mass spectrometric data into single software packages. However, the underlying source codes of each of these software solutions are not available.

Here we aim at providing an open-access software package that is transparent as well as easily accessible and integrates the most important evaluation steps following mass spectrometric analyses and calibration. The software named UltraMassExplorer (UME)[31] is fully browser-based, i.e. the user doesn't require any programming skills. Specifically, we developed this application to provide:

- a high performance peak-based algorithm supporting data reanalysis
- an interactive and reciprocal connection between data and visual representation
- the implementation of quality assessment strategies
- a contribution to customization and transparency of the FT-MS data evaluation procedure
- an open access code in one of the most widely distributed scientific programming languages to facilitate community based improvements

With this paper, we provide the open-access link to the software and the underlying algorithms and source code, a training dataset, i.e. a list of peaks, and a short video tutorial.[31]

Methods

Application development

The code of UME was developed in R version 3.4.2. Several open-source packages (S1) were implemented for the data algorithm, the user interface, the data evaluation and the visualization. The source code licensed under GNU Affero General Public License v3.0 is available from the UltraMassExplorer website.[31]

The graphical user interface was developed using *shiny* (S1), a package that allows developing interactive web applications from R source code. The shiny application is based on two main scripts, namely the *server.R* and the *ui.R*. The *server.R* script runs on a server in the background and is handling all computational tasks of the application and contains the R code of the algorithm. The use of *data.tables* from the *data.table* package (S1) instead of the *data.frames* from the *base* package (S1) allows for high performance computation and memory efficiency, particularly for large input datasets and large molecular formula libraries. The *ui.R* contains the code for the web application. *R Shiny* uses the code from the *ui.R* script to build the graphical user interface and to generate the output directly within a web browser. UME was specifically tested with Mozilla Firefox Quantum Version 63.0b13 and Google Chrome69.0.3497.100 . For an appropriate graphical appearance, we recommend a minimum screen resolution of 1920x1080 pixels.

Molecular formula library

UME builds on static formula libraries that are used to match neutral masses to molecular formulas. The performance of UME depends on the size of the molecular formula library used. The current version of UME provides four different formula libraries, all of which include molecular formulas having $^{12}C_{>1}$, $^{13}C_{0\text{-}1}$, $^{1}H_{>1}$, ^{14}N, ^{16}O, ^{32}S, $^{34}S_{0\text{-}1}$ and ^{31}P and neutral masses up to 702 Da (Table 1). All libraries are based on exact isotopic masses compiled by the National Institute of Standards and Technology[32-34]. Each formula is assigned with a unique number ("vkey") that identifies the formula and the version of the library.

Two libraries are specifically tailored for analyzing NOM samples ("01 NOM", "02 NOM: +15N"). Libraries "04 all CHNOSP" and "05 all CHNOSP: +15N" are not restricted in the number of N, S, and P atoms (Table 1) and therefore cover almost all theoretical formulas. Libraries "02 NOM: +15N" and "05 all CHNOSP: +15N" also consider the ^{15}N isotope. The restrictions on the maximum element ratios in the libraries were set according to the limits and heuristic rules suggested by Kind and Fiehn[8]. A slight difference from the "golden rules" is that we enforced the double bond equivalent (DBE) to be an integer value, and used a modified hydrogen and halogen rule value for the H/C ratio limits (Table 2). To cover additional formulas beyond those provided with UME (e.g. Halogen formulas) we provide an R script[31] for creating new libraries in the UME offline version.

Benchmarking

The performance of the formula matching process and the calculation of the evaluation parameters was tested on a windows workstation (HP EliteDesk, Windows 10 64bit, Intel-Core i5-6500 with 3.20 Ghz and 8GB RAM, SATA 7200 rpm HDD) using the microbenchmark package (S1) for R. The lists of peaks used in benchmarking were compiled from 1, 2, 3, 4, 5, 10, 15, 20, 30, 40, or 50 FT-ICR-MS measurements of marine dissolved organic matter extracts, respectively. The benchmark was repeated ten times for each peak list using library "02 NOM: +15N".

Results

General Workflow

The minimum requirement to start an evaluation process in UME[31] is a validated (calibrated) list of peaks containing columns for (i) the mass (either protonated, deprotonated or neutral), and (ii) the peak magnitude. Also, the user must provide the predetermined mass error (± ppm) that applies to the measurement. In addition, the peak list can contain columns providing the signal/noise ratio, a unique peak identifier, and a unique sample identifier (cf. demo mode in UME and peak list example[31]). It is important to note that all files uploaded by users and corresponding evaluated data are deleted by default on logout from UME.

Filter settings can be modified using sliders, checkboxes, or selection lists located in the sidebar. Interactive data tables, reports and figures are visualized in the main panel. 25 evaluation plots are thematically arranged in the ten tabs “Reconstructed spectra”, “Quality”, “Frequency”, “Elemental frequency”, “Van Krevelen”, “Van Krevelen 3D”, “Kendrick”, “Mass”, “DBE”, and “Statistics”. In the following, we will refer to the tabs by the name of the primary and secondary tab - e.g. “Plots | Van Krevelen”. A brief introduction in the evaluation workflow is given in the Supplements (S2) and the quick start video tutorial[31].

Molecular formula algorithm and calculation of evaluation parameters

The formula assignment algorithm matches neutral mass peaks of the input dataset to a molecular formula in a library (called “library” in the following; Figure 1). For charged molecules, UME so far covers singly charged protonated or deprotonated molecules. The algorithm converts *m/z* ratios to neutral masses by addition or subtraction of the mass of one proton. The dataset is subsequently sorted by the neutral mass in ascending order. The maximum mass error (± ppm; provided by the user) defines the upper and lower limit of the mass window for each peak

assignment. During the matching procedure, the algorithm starts with first peak of the sorted dataset, screens the sorted mass column of the molecular formula library in ascending order for the first mass value larger than the lower limit of the peak mass window, and logs the corresponding row index. Subsequently, the algorithm continues screening the neutral mass column for the first mass value larger than the upper limit of the peak mass window and logs the respective row index minus one. Continuing with the next peak of the dataset, the algorithm uses the last logged row in the library as starting point for repeating the above matching process. After the algorithm has processed the last peak of the dataset, unmatched masses are removed and the numbers of isotopes (^{12}C, ^{1}H, ^{14}N, ^{16}O, ^{31}P, ^{32}S, ^{13}C, ^{15}N, ^{34}S) corresponding to the logged rows are fetched from the library and joined with the dataset. Following the initial formula assignment, valuable parameters for the evaluation of high-resolution mass spectra are automatically calculated by UME (listed with references in Table 2).

Figure 2 shows the results of the benchmark of the formula matching algorithm using library "02 NOM: +15N". The peak lists analyzed for the benchmark contained between 8,014 (=1 sample) and 413,547 (=50 samples) different peaks. The average rate of the processed peaks per second increased with increasing number of supplied peaks from 2,299 peaks s^{-1} for one sample and 4,745 peaks s^{-1} for 50 samples. Thus, a dataset of 50 samples was processed on average in 88 s.

Evaluation of isotope information (^{13}C, ^{34}S)

In a first step, all formulas containing ^{13}C or ^{34}S without an existent referring parent formula are removed. Secondly, for each parent formula, the referring isotopic formula (daughter) is verified in UME. If detected, the peak magnitude of the daughter formula will be connected with the parent formula (Tables | Filtered data: columns "Int. 13c", "Int. 34s" in the data tabs). If not detected, the

values in the “Int. 13c” and/or “Int. 34s” column(s) are set to zero. Subsequently all daughter formulas are removed from the entire dataset.

The ^{13}C stable isotope information for each formula can be further exploited by two approaches: The ^{13}C-filter (sidebar: *Analytical filter*; check box: “remove formulas not verified by ^{13}C-isotope”) can be activated to exclude all formulas that are not verified by the existence of the ^{13}C daughter. In addition, the abundance of the ^{13}C-isotope formula can be evaluated for quality control (Plots | Quality; lower left plot in main panel): The visualization compares the difference between the predicted number of C atoms based on the abundance of the ^{13}C-isotope peak in the spectrum and the number of C atoms in the assigned molecular formula.[cf cf. 14]

Filter and normalization algorithms

The underlying concept is to start with the most conservative set of data, based on all theoretical formulas fetched from the library and the application of basic chemical rules (Table 1). The ranges of 17 filters available in UME are set based on the most conservative unfiltered dataset. Most of these filters use procedures suggested in the literature (summarized in Table 2). Typical surfactants listed in the “terrabase-inc” database [see also 35] and all formulas detected in blank measurements can be excluded by the “analytical filter” in the sidebar. Another analytical filter selects those molecular formulas, for which related information is already known (sidebar filter: “Show only”). For example, the filter can be used to sub-select all formulas in the dataset, which are shared with peptides, nucleotides or formulas that were shown to have high persistence (ideg_neg)[36] or relation to terrestrial organic matter sources in the ocean (iterr_pos)[37]. By updating the underlying database of known formulas, existing knowledge can be projected on an unknown set of samples to facilitate data exploration.

Each change in the filter settings triggers an update of the data normalization and a second filter process based on optionally selected relative intensity thresholds (Figure 3). Implemented into UME are four types of normalization approaches, namely normalization by base peak, by the sum of assigned peaks, by the sum of peaks occurring in all samples of a set of samples, and by the sum of the *n* most intense peaks (Table 2).

After each filter process, UME calculates the number of occurrences of each formula in the dataset (Figure 3). The maximum number of occurrences ("n_occurrence") is therefore determined by the number of samples selected, unless peaks in the peak list fall into the same window of mass error resulting in the assignment of the same formula to different peaks in the spectrum. In addition, the number of assignments ("n_assignments") for each peak is re-determined after each filter event and displayed (Plots | Van Krevelen; right-hand side). Both, the number of assignments and occurrences, are available for the unfiltered dataset as well (Tables | Unfiltered data; columns "n_occurrence_orig", "n_assignments_orig").

Based on the filtered and normalized data set, weighted averages of isotope numbers and ratios (Table 2) are calculated for the selected samples and displayed (Data | Aggregated data).

Visualization and interactivity:

The graphical user interface of UME allows the "on-the-fly" visualization of the effect of filter settings (Figure 3). Each filter adjustment generates a formula subset (Data | Filtered data) from the unfiltered data (Data | Unfiltered data). Subsequently, the filtered dataset is the basis for all UME plots.

Data subsetting and exploration of potential structures

Seven plots in the tabs “Van Krevelen”, “Kendrick”, “DBE”, and “Mass” allow graphical subsetting and export of filtered data using the brush tool. By clicking on individual data points in the plots, respective formulas can be submitted to a PubChem compound search[38] for exploration of potential structures. As data points may overlap in the plot, a dropdown selection list of corresponding molecular formulas is displayed before submitting the formula to the PubChem compound search.

Colors, reporting and data export

For customization, most plot colors can be modified with the respective settings buttons. For those plots using color gradients, e.g. the relative intensity in the van Krevelen plot, the user can select color palettes and can switch between linear and logarithmic color scales.

After customization, plots and data tables can be selected for download in the export menu (sidebar). Publication quality plots can be exported as pdf or as png files. The three dimensional van Krevelen plots are exported as interactive html widgets that retain the entire zoom, rotate and hover functionalities of the original plot. All library and filter settings are documented in a report file, which facilitates the reproducibility and transparency of the evaluation process.

Examples: van Krevelen plots and statistics

Van Krevelen plots display a set of formulas according to their molecular H / C and O / C ratio.[21,22] For the two and three dimensional van Krevelen plots (Plots | Van Krevelen; Plots | Van Krevelen 3D), an optional data projection step is integrated into the visualization that handles the problem of formulas having identical H/C and O/C ratios. If a third dimension is represented in the plot (such as the peak magnitude) the median of the respective parameter is calculated and plotted.

The interactive three dimensional van Krevelen plot allows free rotation and zooming into the data set. Hovering over a data point displays the corresponding molecular formula at the top of the

plot. Both, in two and three dimensional representation, the color variable can be adjusted individually (settings). In addition, the z axis variable can be changed in the three dimensional representation. Three dimensional interactive van Krevelen plots can be a useful tool for data exploration, but may lose their exploratory function if they are reduced to static (printed) images. Therefore we recommend that users refrain from publishing static three dimensional van Krevelen plots, but instead add the corresponding interactive *html*-files as supplementary material to their publications.

Differences between samples can be explored interactively by statistics (Plots | Statistics). Based on the filtered dataset and normalized peak intensities, UME performs a cluster analysis and a non-metric multidimensional scaling (NMDS) to visualize sample similarities. The analyses are based on untransformed data, Bray-Curtis similarity and group-average clustering.

Discussion

Library based formula search

Similar to the formula assignment approach in the Formularity software[10], UME uses a library search approach for matching peak masses with molecular formulas. This library based approach reaches formula assignment rates (Figure 3) comparable to other high-performance formula calculator algorithms such as CHOFIT [*n-3*] *full mode*.[9]

The performance of the UME algorithm is based on the prebuilt molecular formula library. The time consuming step of calculating molecular formulas and corresponding molecular masses is implemented in the process of building the library and not the actual formula matching algorithm. Instead of spending computational power on repetitive calculations of molecular formulas, UME swiftly fetches formulas from the prebuilt library and matches them to peaks in the supplied list. As the matching process is based on screening a sorted library and comparison with a sorted list of peaks, the processing time is a function of the size of the formula library and the size of the supplied peak list. Consequently, the algorithm becomes more efficient with increasing number of supplied peaks at constant library size (Figure 2).The UME algorithm, is thus particularly suited for the fast processing of large datasets (10^5-10^6 peaks) with *a priori* fixed elemental limits. Compared to approaches based on calculating the mass for each peak via nested loops, the molecular formula library based approach of UME offers less flexibility in expanding the elemental limits or considering new elements in the formula calculation process, as it requires creating new formula libraries. New formula libraries that include additional elements or elemental limits can be created using an R script in local installations of UME. This script is available in the download section of the UME website[31]. Incorporating extreme limits for elements into a molecular formula library will slow down computation. The current standard UME library (01 NOM; Table 1) contains $3.9*10^6$ molecular

formulas and has an uncompressed size of 179 MB and can conveniently be handled by common state-of-the-art workstations. For any new element or atom included, the library size roughly increases by factor two. Furthermore, the library size increases exponentially with increasing mass range. A library e.g. containing only formulas consisting of the elements C,H,N,O,P, and S with masses up to 800 Da may already take up tens of gigabytes of memory and cannot efficiently be handled by common workstations. Such libraries may require servers with respective memory capacities or alternative approaches e.g. using a distributed memory SPARK database server. If a stronger restriction in elemental ranges or complete exclusion of certain elements from the molecular formula library can be justified, e.g. as a specific element is typically not expected in a certain type of sample, a manageable library with masses >700 Da can be created and the analysis of higher masses can become feasible on a standard workstation.

Isotope validation

In UME validation of molecular formulas by isotopologues is limited to the ^{13}C and ^{34}S due to their high relative natural abundance, and is based solely on the presence/absence of parent/daughter formulas. Isotopic pattern scoring as included in the Formularity software[10] could provide further hints for correctly assigning formulas, but must consider intensity related variability as shown in the quality control (Plots | Quality; lower left plot in main panel; cf. [14]): Low intensity as compared to high intensity peaks tend to have higher deviation between the predicted number of carbon atoms based on the abundance of the ^{13}C-isotope peak in the spectrum and the number of carbon atoms in the assigned molecular formula, and are thus more likely to mismatch with the theoretical isotope pattern. However, refraining from using an isotopic pattern scoring algorithm in UME increases the likelihood for multiple, possibly false positive formula assignments to a specific mass. This has to be considered in the data interpretation.

Compound group identification

A number of studies showed, that some biomolecular compound groups occupy specific regions in the van Krevelen plot[21]. Vice-versa this relationship has been used to assign molecular formulas from NOM to specific compound groups based on element and element ratio limits [e.g.39,40]. For NOM analysis, however, an assignment of formulas to compound groups based solely on regions in the van Krevelen plot involves the risk of over-interpretation of mass spectrometric data. Projecting commonly applied limits for the "peptide region" (N>0, O/C <= 0.9, and 1.5<=H/C<=2.0)[39] on the PubChem Compound Database[38] yielded 13.5 million hits without structural duplicates. Applying the most conservative filtering criteria for peptides, i.e. the presence of at least one R_2NH group as part of the amide bond, to the PubChem data, shows that only 63.6 % of the hits possess the necessary structural unit to be classified as peptide. While this percentage may, to a certain extent, be influenced by the PubChem users and their fields of research, it illustrates the problem with assigning regions in the van Krevelen plot to specific compound groups. Here the interactive tools of UME may help avoiding over-interpretation of data. Using the "Show only" option a data set can be filtered for molecular formulas known to occur in peptides (sidebar filter: "Show only": "Peptides"). Through the linkage of the van-Krevelen plots to the PubChem database, the filtered formulas plotting in specific compound regions can subsequently be checked for other potential compound sources.

Data projection in van Krevelen diagrams

The van Krevelen plot is probably the most common tool for the visualization of highly complex molecular formula datasets[21,22]. The plots are easy to interpret and intuitive because they project basic chemical principles such as polarity, oxidation/reduction, hydration/dehydration and multiples of CH_2. However, it must be considered that visualization in two dimensional van Krevelen plots is a projection in which several molecular formulas are displayed on the exact same position in the plot

(e.g. $C_{20}H_{20}O_{10}$ and $C_{10}H_{10}O_5$). This is particularly important when a color scale is applied as a third dimension. Without applying a grouping function (e.g. sum, mean, or median) the data representation depends on the sorting of the dataset, even more so if several samples are displayed in the same plot (Figure 4). In UME, the mean function is automatically applied for grouping in all van Krevelen plots, but can be switched off optionally in the respective plot settings menu (Plots|Van Krevelen: settings menu, tickmark “Projection”). The data reduction via a grouping function also provides an improved plotting performance.

Conclusions

UME provides a complete, powerful and fast data pipeline for high-resolution mass spectrometry data and facilitates the transparent and reproducible evaluation of complex organic matter samples. Despite the comprehensive features, we recommend considering some important aspects when using UME:

- The quality of the UME data evaluation can only be as good as the quality of the analytical data. This comprises aspects such as appropriate sample concentration, instrument settings, and appropriate mass calibration. The NOM mass list provided with UME (demo mode) might support quality control for other data sets.

- Apart from the conservative *a priori* rules implemented in the algorithms (such as the valence-based validation of formulas), the selection of the library and any *a posteriori* usage of data filters requires justification. Similar to any other molecular formula pipeline, UME might entice users into setting filter criteria in a way that the outcome matches the expectation. Statements on why filter criteria are set in the way they are is critical for a sound data interpretation. We would like to emphasize that it is not the ultimate aim of an

UME data processing to produce a dataset with solely unequivocal assignments. Differences between samples can be also explored based on a dataset that still comprises multiple formula assignments per peak.

- If UME is used for publications, we encourage users to provide the UME report file that can be exported after the final data evaluation (Data | Report). This improves transparency and reproducibility for other researchers and reviewers.

For future updates of UME, we seek to implement the consideration of other ionization adducts, the addition of a new formula library based on known compounds, additional normalization techniques and evaluation plots, and linkage to further compound databases. These updates will be provided via the online version of UME and documented on the UME website[31].

Acknowledgements

We appreciate the effort and suggestions by two anonymous reviewers and editor Prof. Volmer. TL and BPK were kindly supported by the strategy fund of the Alfred-Wegener-Institute Helmholtz Centre for Polar and Marine Research (AWI; project title: “Inorganics in organics: Chemical and biological controls on micronutrient and carbon fluxes in the polar ocean”). The authors thank Christian Schaefer-Neth, Angelo Steinbach and Jörg Matthes from AWI for setting up the Docker server for UME and invaluable inputs on developing a gateway site for UME. Ana Macario and Sylke Wohlrab helped building the website. Olav Grätz is acknowledged for Piwik integration and Jens Michael Schlüter for securing safe access to UME. Oliver Lechtenfeld, Jana Geuer and Marianna Lucio provided valuable comments on a previous version of UME, and Daniel Bollmann systematically tested options for a SPARK data warehouse.

Supplements

- S1: List of R packages used in UME
- S2: Short description of UME workflow

References

1. Comisarow MB, Marshall AG. Fourier transform ion cyclotron resonance spectroscopy. *Chem Phys Lett.* 1974;25(2):282-283.
2. Marshall AG, Hendrickson CL, Jackson GS. Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrom Rev.* 1998;17(1):1-35.
3. Makarov A. Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Anal Chem.* 2000;72(6):1156-1162.
4. Eliuk S, Makarov A. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annual Review of Analytical Chemistry.* 2015;8(1):61-80.
5. Oldenburg TBP, Brown M, Bennett B, Larter SR. The impact of thermal maturity level on the composition of crude oils, assessed using ultra-high resolution mass spectrometry. *Org Geochem.* 2014;75:151-168.
6. Kujawinski EB, Del Vecchio R, Blough NV, Klein GC, Marshall AG. Probing molecular-level transformations of dissolved organic matter: insights on photochemical degradation and protozoan modification of DOM from electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Mar Chem.* 2004;92(1):23-37.
7. Roullier-Gall C, Witting M, Gougeon RD, Schmitt-Kopplin P. High precision mass measurements for wine metabolomics. *Frontiers in chemistry.* 2014;2:102.
8. Kind T, Fiehn O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics.* 2007;8(1):1-20.
9. Green NW, Perdue EM. Fast Graphically Inspired Algorithm for Assignment of Molecular Formulae in Ultrahigh Resolution Mass Spectrometry. *Anal Chem.* 2015;87(10):5086-5094.
10. Tolic N, Liu Y, Liyu A, et al. Formularity: Software for Automated Formula Assignment of Natural and Other Organic Matter from Ultrahigh-Resolution Mass Spectra. *Anal Chem.* 2017;89(23):12659-12665.
11. Stenson AC, Marshall AG, Cooper WT. Exact Masses and Chemical Formulas of Individual Suwannee River Fulvic Acids from Ultrahigh Resolution Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectra. *Anal Chem.* 2003;75(6):1275-1284.
12. Kujawinski EB, Behn MD. Automated analysis of electrospray ionization fourier transform ion cyclotron resonance mass spectra of natural organic matter. *Anal Chem.* 2006;78(13):4363-4373.
13. Hughey CA, Hendrickson CL, Rodgers RP, Marshall AG, Qian K. Kendrick mass defect spectrum: a compact visual analysis for ultrahigh-resolution broadband mass spectra. *Anal Chem.* 2001;73(19):4676-4681.

14. Koch BP, Dittmar T, Witt M, Kattner G. Fundamentals of molecular formula assignment to ultrahigh resolution mass data of natural organic matter. *Anal Chem.* 2007;79(4):1758-1763.
15. Kunenkov EV, Kononikhin AS, Perminova IV, 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.* 2009;81(24):10106-10115.
16. Herzsprung P, Hertkorn N, von Tumpling W, Harir M, Friese K, Schmitt-Kopplin P. Understanding molecular formula assignment of Fourier transform ion cyclotron resonance mass spectrometry data of natural organic matter from a chemical point of view. *Anal Bioanal Chem.* 2014;406(30):7977-7987.
17. Reemtsma T. The carbon versus mass diagram to visualize and exploit FTICR-MS data of natural organic matter. *J Mass Spectrom.* 2010;45(4):382-390.
18. 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. *Eur J Mass Spectrom (Chichester).* 2011;17(4):415-421.
19. Kew W, Blackburn JWT, Clarke DJ, Uhrín D. Interactive van Krevelen diagrams – Advanced visualisation of mass spectrometry data of complex mixtures. *Rapid Commun Mass Spectrom.* 2017;31(7):658-662.
20. Koch BP, Kattner G, Witt M, Passow U. Molecular insights into the microbial formation of marine dissolved organic matter: recalcitrant or labile? *Biogeosciences.* 2014;11(15):4173-4190.
21. Kim S, Kramer RW, Hatcher PG. Graphical method for analysis of ultrahigh-resolution broadband mass spectra of natural organic matter, the van Krevelen diagram. *Anal Chem.* 2003;75(20):5336-5344.
22. Van Krevelen D. Graphical-statistical method for the study of structure and reaction processes of coal. *Fuel.* 1950;29:269-284.
23. Kendrick E. A Mass Scale Based on CH2 = 14.0000 for High Resolution Mass Spectrometry of Organic Compounds. *Anal Chem.* 1963;35(13):2146-2154.
24. Schmitt-Kopplin P, Gelencsér A, Dabek-Zlotorzynska E, et al. Analysis of the Unresolved Organic Fraction in Atmospheric Aerosols with Ultrahigh-Resolution Mass Spectrometry and Nuclear Magnetic Resonance Spectroscopy: Organosulfates As Photochemical Smog Constituents. *Anal Chem.* 2010;82(19):8017-8026.
25. https://github.com/wkew/FTMSVisualization. Accessed 13.10.2018.
26. Kilgour DPA, Hughes S, Kilgour SL, et al. Autopiquer - a Robust and Reliable Peak Detection Algorithm for Mass Spectrometry. *J Am Soc Mass Spectrom.* 2017;28(2):253-262.
27. Kilgour DPA, Wills R, Qi Y, O'Connor PB. Autophaser: An Algorithm for Automated Generation of Absorption Mode Spectra for FT-ICR MS. *Anal Chem.* 2013;85(8):3903-3911.
28. http://massspec.com/composer. Accessed 13.10.2018.
29. Corilo YE, Rodgers RP, Hendrickson CL, Marshall AG, Inventors. PetroOrg Software. 2013.
30. http://software.petroorg.com. Accessed 22.05.2018.
31. https://www.awi.de/en/ume. Accessed 24.05.2018.
32. Berglund M, Wieser ME. Isotopic compositions of the elements 2009 (IUPAC Technical Report). *Pure Appl Chem.* 2011;83(2):397-410.
33. Meija J, Coplen TB, Berglund M, et al. Atomic weights of the elements 2013 (IUPAC Technical Report). *Pure Appl Chem.* 2016;88(3).

34. Wang M, Audi G, Wapstra AH, et al. The Ame2012 atomic mass evaluation. *Chinese Physics C.* 2012;36(12):1603-2014.
35. Lechtenfeld OJ, Koch BP, Gašparović B, Frka S, Witt M, Kattner G. The influence of salinity on the molecular and optical properties of surface microlayers in a karstic estuary. *Mar Chem.* 2013;150:25-38.
36. Flerus R, Lechtenfeld OJ, Koch BP, et al. A molecular perspective on the ageing of marine dissolved organic matter. *Biogeosciences.* 2012;9(6):1935-1955.
37. Medeiros MP, Seidel M, Powers LC, Dittmar T, Hansell DA, Miller WL. Dissolved organic matter composition and photochemical transformations in the northern North Pacific Ocean. *Geophys Res Lett.* 2015;42(3):863-870.
38. https://pubchem.ncbi.nlm.nih.gov/. Accessed 04.12.2017, 2017.
39. Rossel PE, Bienhold C, Boetius A, Dittmar T. Dissolved organic matter in pore water of Arctic Ocean sediments: Environmental influence on molecular composition. *Org Geochem.* 2016;97:41-52.
40. Ide Ji, Ohashi M, Takahashi K, et al. Spatial variations in the molecular diversity of dissolved organic matter in water moving through a boreal forest in eastern Finland. *Scientific Reports.* 2017;7:42102.
41. Hsu CS, Qian K, Chen YC. An innovative approach to data analysis in hydrocarbon characterization by on-line liquid chromatography-mass spectrometry. *Anal Chim Acta.* 1992;264(1):79-89.
42. Koch BP, Dittmar T. From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter. *Rapid Commun Mass Spectrom.* 2006;20(5):926-932.
43. Koch BP, Dittmar T. From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter. *Rapid Commun Mass Spectrom.* 2016;30(1):250-250.
44. LaRowe DE, Van Cappellen P. Degradation of natural organic matter: A thermodynamic analysis. *Geochim Cosmochim Acta.* 2011;75(8):2030-2042.

Table 1. List of parameters and allowed ranges for calculating the molecular formula library.

Library	**01**	**02**	**04**	**05**
UME label	01 NOM	02 NOM: +15N	04 all CHNOSP	05 all CHNOSP: +15N
Neutral mass max	702	702	702	702
Heuristic filter, common range rule 4,5 (Table 2)[8]	applied	applied	applied	applied
C min	1	1	1	1
^{12}C max	not restricted**	not restricted**	not restricted**	not restricted**
^{13}C max	1	1	1	1
H min	1	1	1	1
H max	C*2+2+N+P	C*2+2+N+P	C*2+2+N+P	C*2+2+N+P
(N, O, P, S)min	0	0	0	0
(^{14}N+^{15}N)max	6	6	not restricted**	not restricted**
^{15}N max	0	1	0	1
^{23}Na max	0	0	0	0
^{16}O max	not restricted**	not restricted**	not restricted**	not restricted**
^{18}O max	0	0	0	0
^{31}P max	3	3	not restricted**	not restricted**
(^{32}S+^{34}S) max	3	3	not restricted**	not restricted**
^{34}S max	1	1	1	1
Double bond equivalents	integer & >=0	integer & >=0	integer & >=0	integer & >=0
O/(^{12}C+^{13}C) max	1.2*	1.2*	1.2*	1.2*
H/(^{12}C+^{13}C) max	not restricted**	not restricted**	not restricted**	not restricted**
(^{14}N+^{15}N)/(^{12}C+^{13}C) max	1.3*	1.3*	1.3*	1.3*
P/(^{12}C+^{13}C) max	0.3*	0.3*	0.3*	0.3*
(^{32}S+^{34}S)/(^{12}C+^{13}C) max	0.8*	0.8*	0.8*	0.8*
Total Formulas	3900896	7170597	12306406	23351423

Table 2. Selected filter parameters calculated by UME (nm = nominal mass, m = neutral mass, number of isotopes in a formula (^{12}C, ^{13}C, ^{1}H, ^{14}N, ^{15}N, ^{16}O, ^{31}P, ^{32}S, ^{34}S), $\int 12\,c$ = abundance of formula containing no ^{13}C isotope, $\int 13\,c$ = abundance of formula containing one ^{13}C isotope, I_{peak} = intensity of a peak, $I_{basepeak}$ = magnitude of base peak, I_{all} = magnitude of a peak that occurs in all samples, I_{rank} = magnitude of a peak of the *n* most intense peaks, *C, H, N, O, S* =number of atoms of the respective element, *ri* = relative intensity based on the selected normalization procedure)

Parameter	Abbreviation in UME	Calculation	Reference
Kendrick Mass Defect	kmd	$nm - \frac{14 \cdot m}{14.01565}$	13,41
z*	z	$nm\ mod\ 14\ - 14$	41
Aromaticity index	Ai	$1+^{12}C+^{13}C-^{16}O-^{32}S-^{34}S - \frac{0.5 \cdot (^{1}H+^{14}N+^{15}N+^{31}P)}{^{12}C+^{13}C-^{16}O-^{32}S-^{34}S-^{14}N-^{15}N-^{31}P}$	42,43
Average nominal oxidation state of carbon	nosc	$4 - \frac{4 \cdot (^{12}C+^{13}C)+^{1}H-3\cdot^{14}N-2\cdot^{16}O+5\cdot^{31}P-2\cdot^{32}S}{^{12}C+^{13}C}$	44
Gibbs energies for the oxidation half reactions of molecular formulas	delg0_cox	$-53.7 + 28.5 \cdot (\frac{4 \cdot (^{12}C+^{13}C)+^{1}H-3\cdot^{14}N-2\cdot^{16}O+5\cdot^{31}P-2\cdot^{32}S}{^{12}C+^{13}C})$	44
difference between the predicted number of carbon atoms based on the abundance of the ^{13}C-isotope and the number of carbon atoms in the molecular formula	dev_n_c	$\frac{107 \cdot int13c}{int12c} - 1.07\cdot^{12}C$	14
Theoretical intensity of the ^{13}C daughter	relint13c_calc	$1.07\cdot^{12}C$	-
Theoretical intensity of the ^{34}S daughter	relint32s_calc	$4.4\cdot^{32}S$	-
Relative intensity after normalization by base peak	rel_int	$\frac{I_{peak}}{I_{base\ peak}}$	-
Relative intensity after normalization by sum of all peaks	rel_sum_int	$\frac{I_{peak}}{\sum_{i=1}^{n} I_i}$	-
Relative intensity after normalization by sum of peaks occurring in all samples	rel_sum_int_opt	$\frac{I_{peak}}{\sum_{i=1}^{n} I\,all_i}$	-
Relative intensity after normalization by sum of most intense peak	rel_sum_int_rank	$\frac{I_{peak}}{\sum_{i=1}^{n} I\,rank_i}$	-
weighted average *m/z*	wa_mz	$\frac{\sum_{i=1}^{n} mz_i \cdot ri_i}{\sum_{i=1}^{n} ri_i}$	-
weighted average DBE	waDBE	$\frac{\sum_{i=1}^{n} dbe_i \cdot ri_i}{\sum_{i=1}^{n} ri_i}$	-
weighted average C	waC	$\frac{\sum_{i=1}^{n} C_i \cdot ri_i}{\sum_{i=1}^{n} ri_i}$	-
weighted average H	waH	$\frac{\sum_{i=1}^{n} H_i \cdot ri_i}{\sum_{i=1}^{n} ri_i}$	-

weighted average O	waO	$\frac{\sum_{i=1}^{n} O_i \cdot ri_i}{\sum_{i=1}^{n} ri_i}$	-
weighted average N	waN	$\frac{\sum_{i=1}^{n} N_i \cdot ri_i}{\sum_{i=1}^{n} ri_i}$	-
weighted average S	waS	$\frac{\sum_{i=1}^{n} S_i \cdot ri_i}{\sum_{i=1}^{n} ri_i}$	-
weighted average O/C	waOC	$\frac{\sum_{i=1}^{n} ri_i \cdot O_i/C_i}{\sum_{i=1}^{n} ri_i}$	-
weighted average H/C	waHC	$\frac{\sum_{i=1}^{n} ri_i \cdot H_i/C_i}{\sum_{i=1}^{n} ri_i}$	-
weighted average S/C	waSC	$\frac{\sum_{i=1}^{n} ri_i \cdot S_i/C_i}{\sum_{i=1}^{n} ri_i}$	-

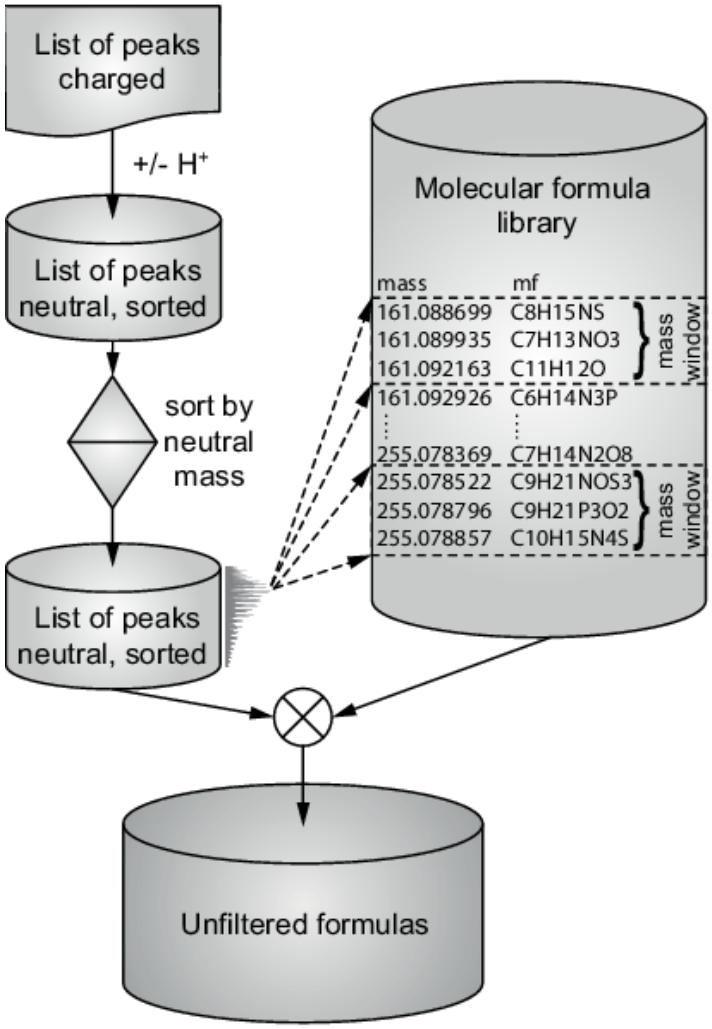

Figure 1. Processes in UME from the import of datasets to the matching of peak masses with one or more molecular formulas in the respective mass windows.

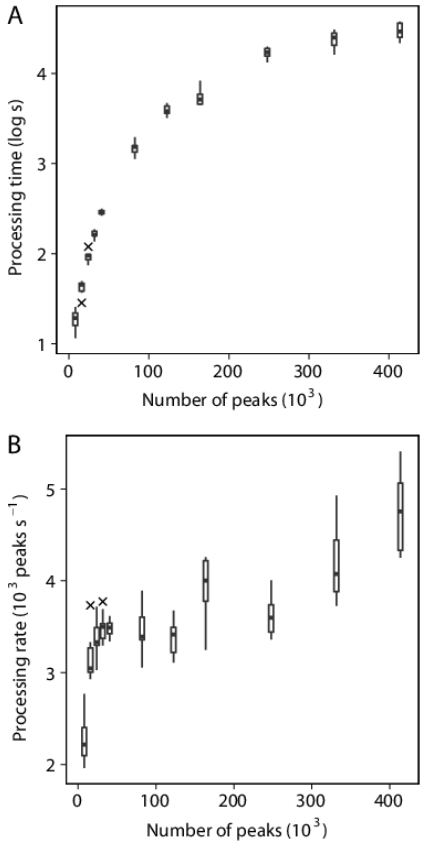

Figure 2. Formula matching algorithm benchmark. Standard boxplots (n=10) of the processing time (A) and the processing rate (B), respectively, versus the number of peak masses supplied.

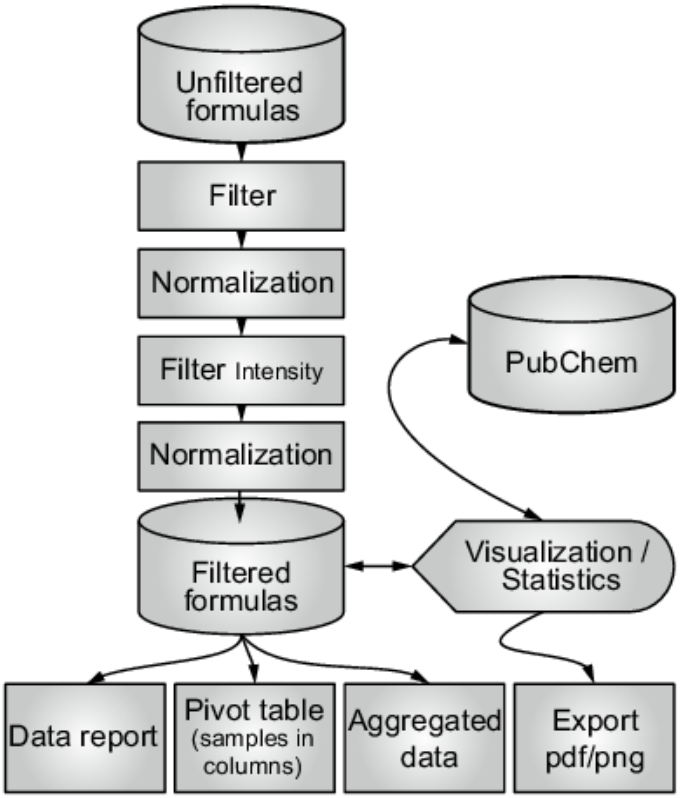

Figure 3. Processes in UME from the display of unfiltered results to the export of final results and report generation.

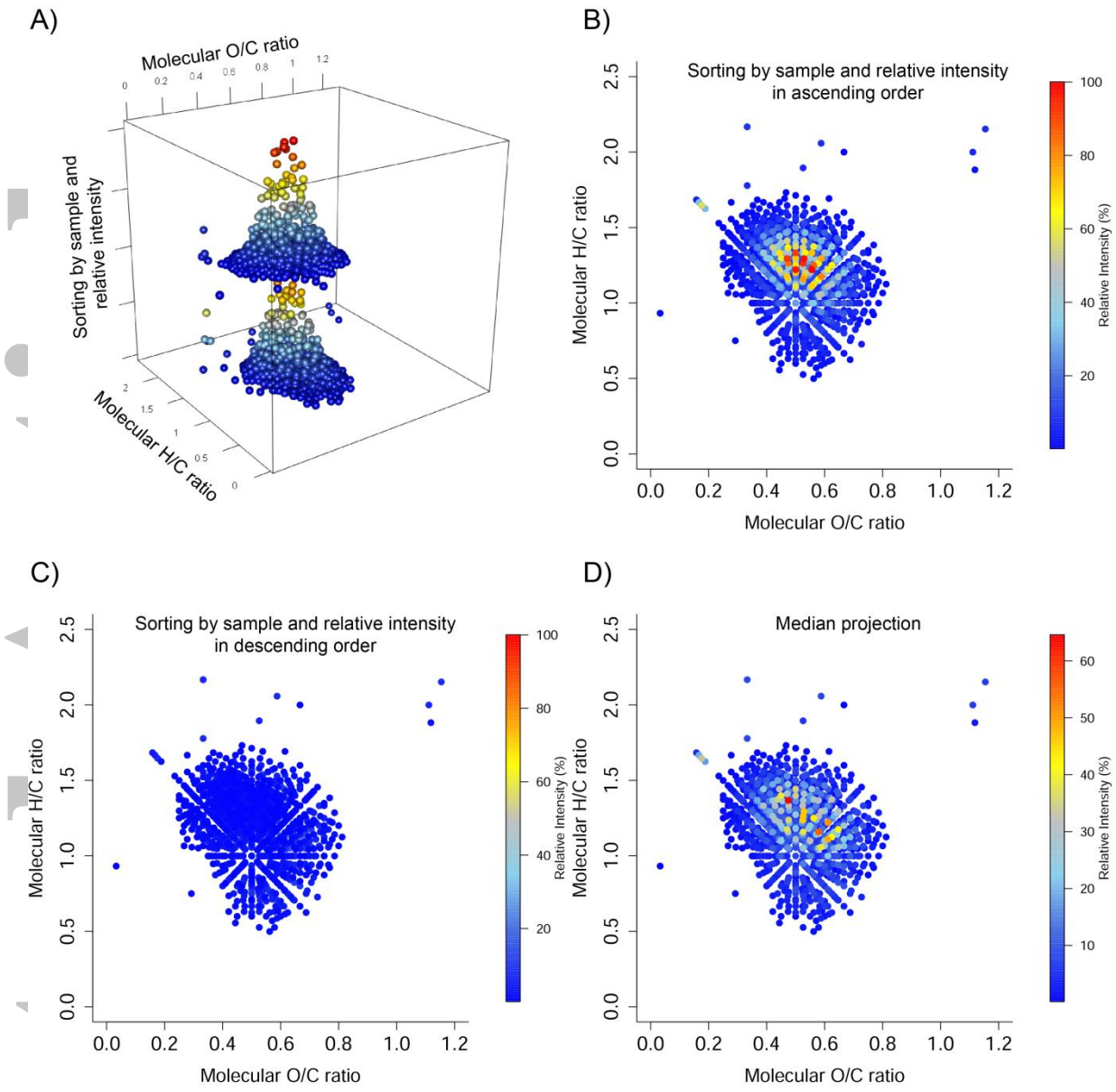

Figure 4. Visual representation of two samples in van Krevelen plots using an additional parameter for the third dimension: reducing (A) 3d representations to 2d projections must consider the dependence on data sorting: sorted by sample and relative intensity in (B) ascending and (C) descending order. The standard preset in UME (D) uses the median value for every discrete location in the plot (Plots|Van Krevelen: settings menue, tickmark “Projection”).