Received: 11 November 2016 Revised: 5 January 2017 Accepted: 5 January 2017 Published online in Wiley Online Library

Rapid Commun. Mass Spectrom. **2017**, *31*, 658–662
(wileyonlinelibrary.com) DOI: 10.1002/rcm.7823

Dear Editor,

Interactive van Krevelen diagrams – Advanced visualisation of mass spectrometry data of complex mixtures

The field of complex mixture analysis has advanced significantly in the past two decades, although its history goes much further back. When Dirk Willem van Krevelen developed his now eponymous diagram in 1950 to represent the chemical makeup of coals, he proposed that the chemical nature of samples, including the presence of structural motifs and chemical properties, could be inferred from the elemental ratios of the sample.[1] While his work, limited by the technology of the era, looked at whole samples characterised by the ratio of elements present, i.e. number of carbons-to-hydrogens within the sample, modern mass spectrometry allows us to examine in a similar manner the individual components of a complex mixture.

Since 2003, when the modern van Krevelen diagram was first used to visualise complex MS datasets,[2] every significant high-resolution mass spectrometric analysis of a complex mixture has included one.[3–5] Today's van Krevelen diagram places every assigned unique chemical formula on a 2D scatter plot of H/C ratio versus O/C ratio, although other elemental ratios can also be used. Although this represents a break from the original intentions of van Krevelen, the modified technique has become a useful tool for the interpretation and visualisation of complex data. For example, regions of the van Krevelen plot can be tentatively associated with certain compound classes,[2,6,7] such as lipids (O/C $<$ 0.2, H/C 2 – values quoted are approximate), carbohydrates (H/C 2, O/C 1), or condensed hydrocarbons (O/C $<$ 0.2, H/C $<$ 1).

In the field of complex mixture analysis, a number of methods are available to the enterprising chemist; however, Fourier transform ion cyclotron resonance mass spectrometry (FTICR MS) reigns supreme as the 'gold standard' technique.[8,9] Likewise, there exist a number of well-studied complex mixtures, including natural organic matter (NOM), i.e. dissolved organic matter,[5,10,11] soil organic matter,[12] and organic aerosols,[13–15] petroleum,[16,17] or beverages such as wine[18] or Scotch whisky.[19,20] Amongst the most complex of these, a component of NOM and the closest sample to a universal standard, is Suwannee River Fulvic Acid (SRFA) produced by the International Humic Substance Society.[21] A typical electrospray ionisation (ESI)-FTICR mass spectrum of SRFA will contain thousands of peaks across a range of masses, predominantly between *m/z* 200 and 700. Due to its ubiquity and complexity, SRFA was chosen to demonstrate the capability of the visualisation tools described herein.

With the mass accuracy of FTICR MS spectra in parts-per-billion,[22,23] routine and confident assignment of thousands of unique chemical formulae to individual peaks is now increasingly possible. The generation of this volume of data represents a significant challenge in terms of data visualisation, interrogation, and interpretation that has not been addressed so far. Here, we present a handful of tools aimed at filling this gap.

We have developed a version of the van Krevelen diagram, which introduces interactivity, and allows the analyst, or reviewer, to interrogate the data in an intuitive way. This interactive van Krevelen, or *i*-van Krevelen for short, is generated using the Bokeh Python plotting library.[24] The developed tools are fully compatible with data assigned using any software package, as the input for the *i*-van Krevelen scripts are three text files containing (1) monoisotopic peak assignments, (2) isotopologue peak assignments, (3) remaining unassigned, but detected, peaks. Example input files are included with the suite of presented tools. The Bokeh API allows for the straightforward coding, in Python, of complex JavaScript (JSON) plots as HTML5 Canvas objects. The output from this tool is a standard HTML document compatible with any modern web browser such as Google Chrome, Firefox, or Internet Explorer.

The main feature of the *i*-van Krevelen software is the generation of interactive diagrams including a centroid mass spectrum, van Krevelen, DBE vs carbon number plot and the modified Aromaticity Index vs carbon number plot.[25] The plots are linked together, such that selecting any data points in one plot highlights those same points – i.e. unique chemical formula – in the other plots. In addition, these plots are explorable, featuring zoom and pan tools, as well as a display of the key information of each point in a hover-tool. Finally, the data points can be used as hyperlinks – in our implementation, they link to a ChemSpider (The Royal Society of Chemistry, Cambridge, UK) search for their molecular formula.

The benefits of these features will be immediately obvious to any analytical chemist who has tried to make sense of complex static van Krevelen diagrams of complex mixtures.

For example, in a standard van Krevelen plot, numerous points may be superimposed if they share elemental ratios but differ in molecular formulae. As a van Krevelen plot is a specific type of scatter plot, it is susceptible to the same problems as other any other scatter plot, and can be misinterpreted when hundreds or thousands of points are plotted. Whilst the addition of colour and transparency can reduce these problems, they are not eliminated entirely.[26,27] One alternative is to plot data density, not individual data points – i.e. a histogram or kernel density plot in 1D, or a hexagonally binned data plot in 2D.[28] This allows easier visualisation of where the most (or largest, or most intense, depending on the density variable) data points are; however, this approach leads to a loss of information about specific components and their molecular formulae. With interactivity, however, a user can zoom to a region of

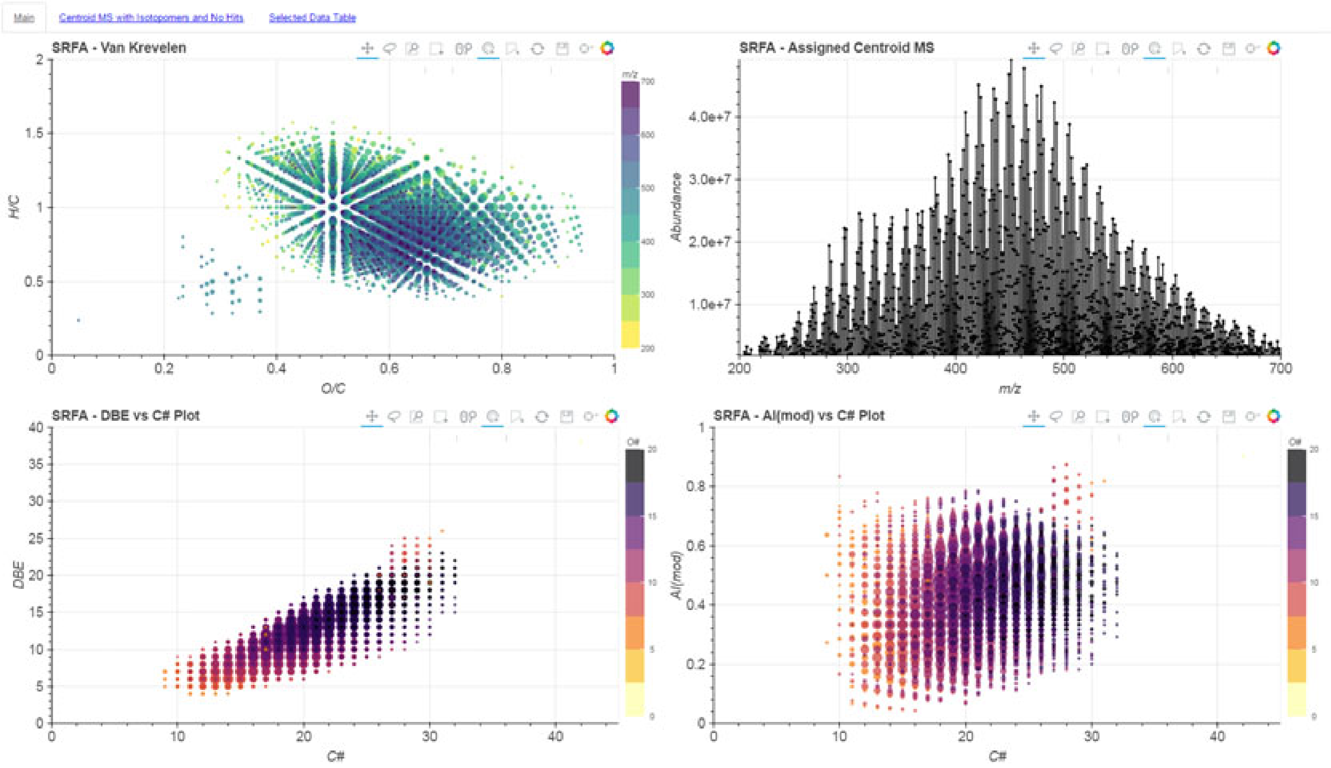

Figure 1. Overview screenshot of the *i*-van Krevelen Main Page. Four sub-plots are shown, from top left clockwise, van Krevelen, centroid mass spectrum, DBE vs C #, and AI(mod) vs C #. The scatter plots have points sized per their relative abundance, while the colour scales represent either *m/z* range (van Krevelen) or oxygen number (DBE vs C# and AI(mod) vs C#). [Colour figure can be viewed at wileyonlinelibrary.com]

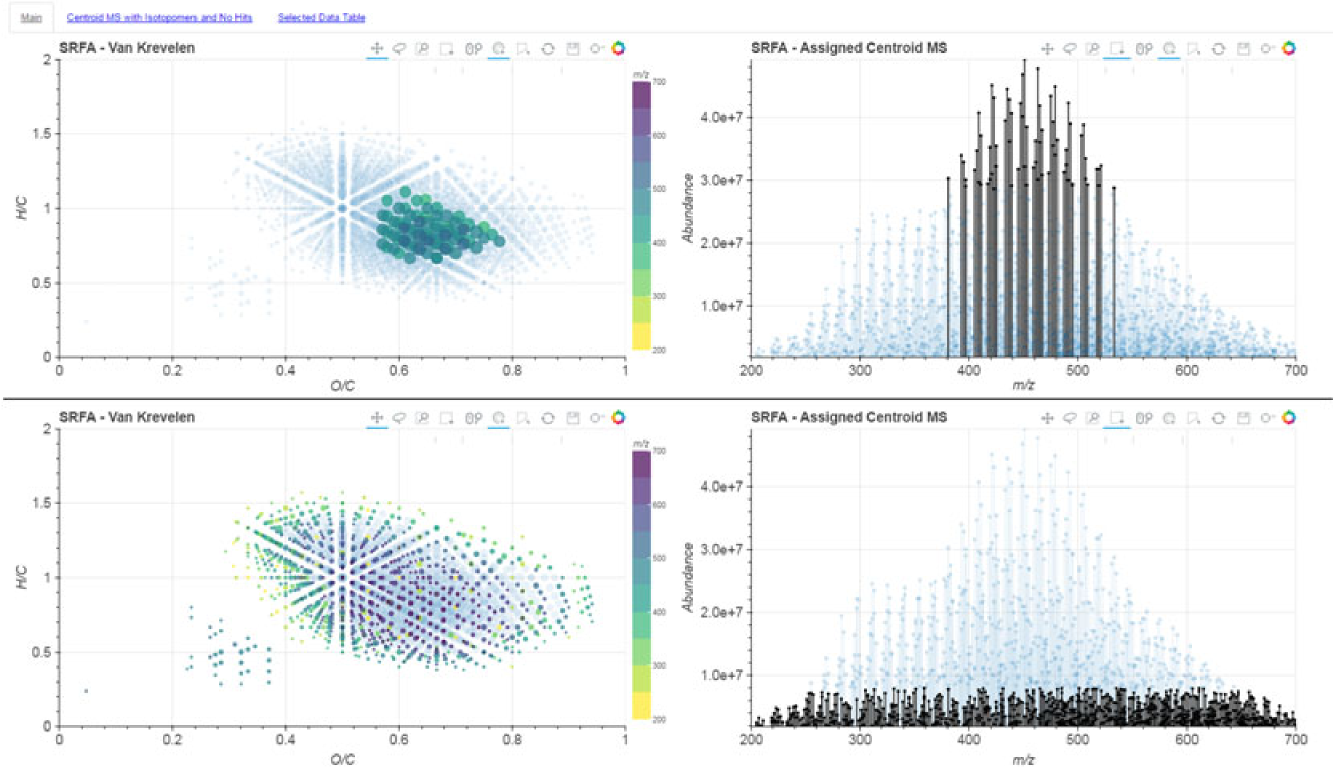
Main
Centroid MS with Isotopomers and No Hits
Selected Data Table
SRFA - Van Krevelen
H/C
O/C
m/z
SRFA - Assigned Centroid MS
Abundance
m/z
SRFA - Van Krevelen
SRFA - Assigned Centroid MS

Figure 2. Screenshots of the van Krevelen and centroid mass spectrum plots with two different selections of data points. The top frame shows that the selection of the most abundant ions only represents a small range of chemical diversity on the van Krevelen plots, whilst the bottom frame shows that the least abundant ions, “the grass”, represent the true chemical diversity of the spectrum. [Colour figure can be viewed at wileyonlinelibrary.com]

interest in the plot, and use the hover-tools to identify every component contributing to a particular point, thus removing the ambiguity caused by the overlap. Furthermore, we encode the relative abundance of a species by the size of the glyph on the plot. The colour can then be used to indicate mass, as in our van Krevelen plots, or oxygen number, as in our DBE and AI plots. This approach is illustrated in our recent paper on Scotch whisky.[20]

Reducing complex data down to a two-variable van Krevelen plot inevitably represents a loss of information. In our tool, we have therefore created several 2D plots that are linked together. An example of this layout is shown in Fig. 1. This allows for the relation of multiple variables to a single molecular formula in order to better understand the sample. For example, as shown in Fig. 2, we can select only the most intense signals in the spectrum. Here we can see that these species, whilst the dominant compounds in the mass spectrum, represent only a fraction of the diversity present in the sample as revealed by their position on the van Krevelen plot. This means that if we were to consider only the n most abundant ions – an approach utilised in some previous statistical analyses of complex spectra[19] – we would be losing the vast majority of the chemical diversity of the sample. On the contrary, by selecting only the low-abundance peaks, i.e. the "grass", we can see that these signals do describe the chemical diversity of the sample more fully. Such information, which is lost in static van Krevelen plots, will be important for comparative studies aiming to characterise multiple samples by different ionisation techniques; for example, comparing ESI with MALDI (matrix-assisted laser desorption/ionization) mass spectra, where the abundance of a species is a function of both concentration and ionisation energy. Likewise, this interactive selection of points can be used to easily link outliers on any plots to their positions on the mass

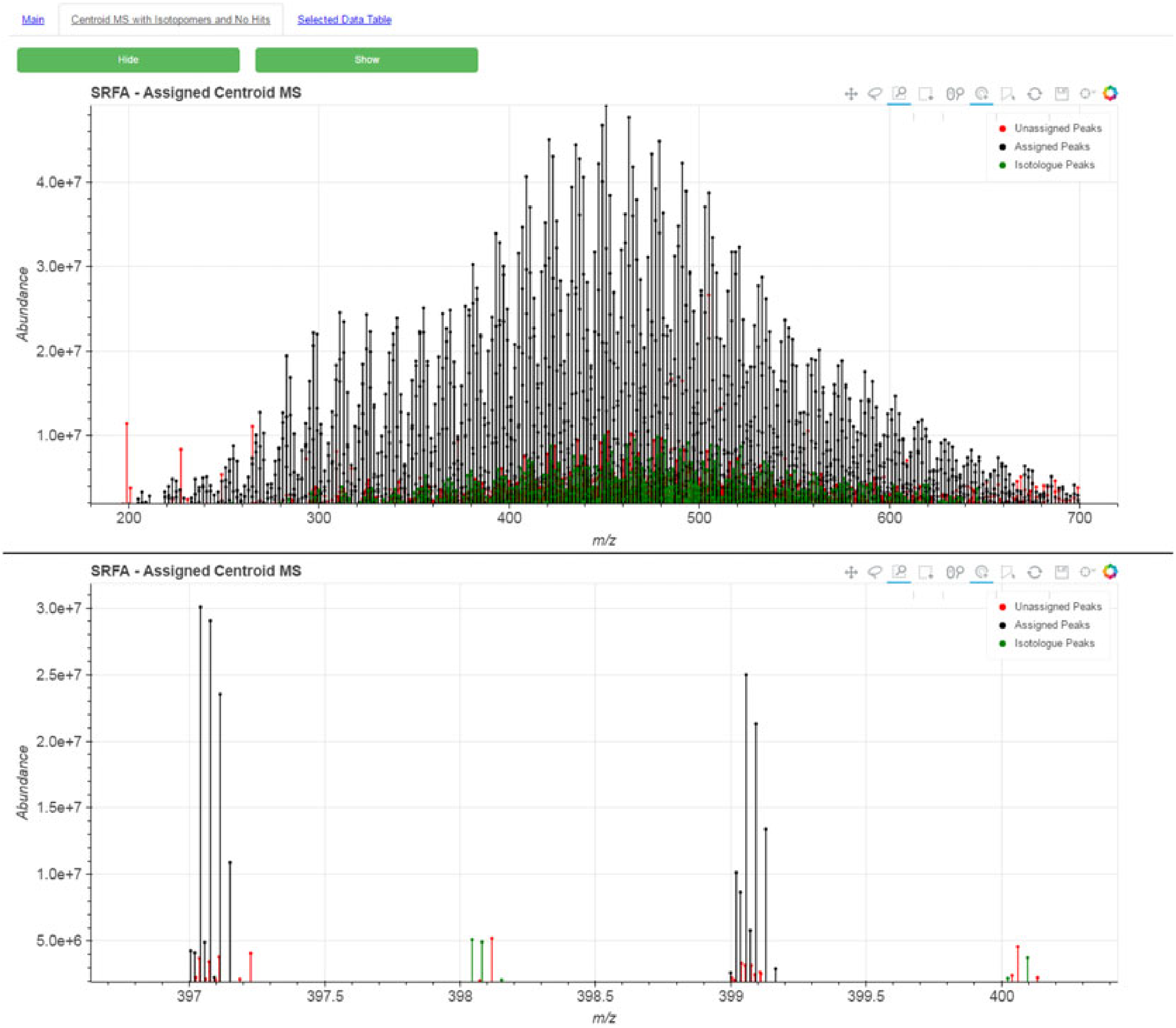

Figure 3. Screenshots of the centroid mass spectrum showing an overlay of peaks which represent isotopomers (green) and peaks which could not be assigned a molecular formula (red). The bottom frame shows a zoomed in region of m/z 397–400, clearly showing the monoisotopic peaks and their associated ^{13}C isotopomers. Those peaks not assigned may be secondary isotope peaks, e.g. with two ^{13}C atoms or a single ^{18}O atom. [Colour figure can be viewed at wileyonlinelibrary.com]

spectrum, or understand where specific regions of these plots originate from in the mass spectra.

On a second tab of the HTML page, the centroid mass spectrum is plotted with the identified isotopomers, as well as the remaining unassigned peaks. An example of this is shown in Fig. 3. This gives the analyst, and more importantly the reader or reviewer, a straightforward means to see how well the spectrum was assigned, thus validating or otherwise the assignment methodologies.

Finally, on a third tab, the data table is presented that is required to generate the plots, and it is also interactively linked to the plots, meaning that selections made on any plot are highlighted in the data table, and vice versa. This data table is downloadable as a text file.

The developed code also includes a number of related Python scripts for: (i) automated batch plotting of publication quality van Krevelen and DBE vs Carbon Number plots; (ii) heteroatomic class distribution calculation and plotting; (iii) an "all-possible-formula-generator", which calculates a list of possible, logical, chemical formulae as based on work done by Kind *et al.*;[29] (iv) a tool to batch perform automated exact mass-to-formula assignment based on Kendrick mass defect analysis and z* by looking for homologous series of compounds;[30] and (v) a tool for reformatting of PetroOrg (Florida State University, Tallahassee, FL, USA) output CSV files. Assignment files generated by the latter two tools produce, as outputs, inputs for the *i*-van Krevelen software and other included scripts. The included formula generator is especially useful for determining assignment error thresholds, for example by allowing the user to determine the minimum distance between possible compounds at a given *m/z*, and thus adding confidence to the assignment.

Overall, these interactive plots, and their combination, represent a step forward in the analysis of complex mixtures by high-resolution mass spectrometry. The tools are open-source and available freely through GitHub with a GNU General Public License v3.0, encouraging others to experiment with and build upon them. The GitHub repository[31] can be found online.[32] An online tool allowing the use of some of these tools without the need to install any specialist software has also been developed, and can be found through the GitHub repository. An example of the interactive plots enabled by this initial *i*-van Krevelen package based on the SRFA FTICR MS data can also be found online.[33]

Future work could incorporate the Datashader[34] package, which would allow the visualisation of the raw profile spectra in a web browser without the need for the end user to download large data files or install proprietary mass spectrometry software, as well as the Bokeh Server tool, allowing the user to dynamically select which variables to plot on each axis, or to choose a specific colour or size scale. Examples of code for the Datashader functionality are included as a Jupyter Notebook in the GitHub repository.

Acknowledgements

The authors gratefully acknowledge the continued support and advice of Dr Logan Mackay, and his maintenance of the FT-ICR MS instrument. This project was supported by BBSRC grant BB/L016311/1 and The Scotch Whisky Research Institute.

William Kew, John W.T. Blackburn, David J. Clarke and Dušan Uhrín**
EaStCHEM, School of Chemistry, University of Edinburgh,
Edinburgh EH9 3FJ, UK

**Correspondence to: W. Kew and D. Uhrín, EaStCHEM, School of Chemistry, University of Edinburgh, Edinburgh EH9 3FJ, UK.*
E-mail: w.kew@sms.ed.ac.uk; dusan.uhrin@ed.ac.uk

REFERENCES

[1] D. Van Krevelen. Graphical statistical method for the study of structure and reaction processes of coal. *Fuel* **1950**, *29*, 269.

[2] S. Kim, R. W. Kramer, P. G. Hatcher. Graphical method for analysis of ultrahigh-resolution broadband mass spectra of natural organic matter, the Van Krevelen diagram. *Anal. Chem.* **2003**, *75*, 5336.

[3] Z. Wu, R. P. Rodgers, A. G. Marshall. Two- and three-dimensional van Krevelen diagrams: A graphical analysis complementary to the Kendrick mass plot for sorting elemental compositions of complex organic mixtures based on ultrahigh-resolution broadband Fourier transform ion cyclotron resonance. *Anal. Chem.* **2004**, *76*, 2511.

[4] P. Herzsprung, W. von Tümpling, N. Hertkorn, M. Harir, O. Büttner, J. Bravidor, K. Friese, P. Schmitt-Kopplin. Variations of DOM quality in inflows of a drinking water reservoir: Linking of van Krevelen diagrams with EEMF spectra by rank correlation. *Environ. Sci. Technol.* **2012**, *46*, 5511.

[5] J. D'Andrilli, W. T. Cooper, C. M. Foreman, A. G. Marshall. An ultrahigh-resolution mass spectrometry index to estimate natural organic matter lability. *Rapid Commun. Mass Spectrom.* **2015**, *29*, 2385.

[6] N. Hertkorn, R. Benner, M. Frommberger, P. Schmitt-Kopplin, M. Witt, K. Kaiser, A. Kettrup, J. I. Hedges. Characterization of a major refractory component of marine dissolved organic matter. *Geochim. Cosmochim. Acta* **2006**, *70*, 2990.

[7] R. D. Gougeon, M. Lucio, M. Frommberger, D. Peyron, D. Chassagne, H. Alexandre, F. Feuillat, A. Voilley, P. Cayot, I. Gebefugi, N. Hertkorn, P. Schmitt-Kopplin. The chemodiversity of wines can reveal a metabologeography expression of cooperage oak wood. *Proc. Natl. Acad. Sci.* **2009**, *106*, 9174.

[8] L. Lim, F. Yan, S. Bach, K. Pihakari, D. Klein. Fourier transform mass spectrometry: The transformation of modern environmental analyses. *Int. J. Mol. Sci.* **2016**, *17*, 104.

[9] A. G. Marshall, C. L. Hendrickson, G. S. Jackson. Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrom. Rev.* **1998**, *17*, 1.

[10] D. Cao, H. Huang, M. Hu, L. Cui, F. Geng, Z. Rao, H. Niu, Y. Cai, Y. Kang. Comprehensive characterization of natural organic matter by MALDI- and ESI-Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chim. Acta* **2015**, *866*, 48.

[11] P. Herzsprung, N. Hertkorn, W. von Tümpling, M. Harir, K. Friese, P. Schmitt-Kopplin. Molecular formula assignment for dissolved organic matter (DOM) using high-field FT-ICR-MS: chemical perspective and validation of sulphur-rich organic components (CHOS) in pit lake samples. *Anal. Bioanal. Chem.* **2016**, *408*, 2461.

[12] B. F. Mann, H. Chen, E. M. Herndon, R. K. Chu, N. Tolic, E. F. Portier, T. Roy Chowdhury, E. W. Robinson, S. J. Callister, S. D. Wullschleger, D. E. Graham, L. Liang, B. Gu. Indexing permafrost soil organic matter degradation

using high-resolution mass spectrometry. *PLoS One* **2015**, *10*, e0130557.

[13] B. Nozière, M. Kalberer, M. Claeys, J. Allan, B. D'Anna, S. Decesari, E. Finessi, M. Glasius, I. Grgić, J. F. Hamilton, T. Hoffmann, Y. Iinuma, M. Jaoui, A. Kahnt, C. J. Kampf, I. Kourtchev, W. Maenhaut, N. Marsden, S. Saarikoski, J. Schnelle-Kreis, J. D. Surratt, S. Szidat, R. Szmigielski, A. Wisthaler. The molecular identification of organic compounds in the atmosphere: State of the art and challenges. *Chem. Rev.* **2015**, *115*, 3919.

[14] S. Tao, X. Lu, N. Levac, A. P. Bateman, T. B. Nguyen, D. L. Bones, S. A. Nizkorodov, J. Laskin, A. Laskin, X. Yang. Molecular characterization of organosulfates in organic aerosols from Shanghai and Los Angeles urban areas by nanospray-desorption electrospray ionization high-resolution mass spectrometry. *Environ. Sci. Technol.* **2014**, *48*, 10993.

[15] A. S. Wozniak, J. E. Bauer, R. L. Sleighter, R. M. Dickhut, P. G. Hatcher. Technical Note: Molecular characterization of aerosol-derived water soluble organic carbon using ultrahigh resolution electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Atmos. Chem. Phys.* **2008**, *8*, 5099.

[16] C. A. Hughey, C. L. Hendrickson, R. P. Rodgers, A. G. Marshall, K. Qian. Kendrick mass defect spectrum: A compact visual analysis for ultrahigh-resolution broadband mass spectra. *Anal. Chem.* **2001**, *73*, 4676.

[17] A. M. McKenna, J. T. Williams, J. C. Putman, C. Aeppli, C. M. Reddy, D. L. Valentine, K. L. Lemkau, M. Y. Kellermann, J. J. Savory, N. K. Kaiser, A. G. Marshall, R. P. Rodgers. Unprecedented ultrahigh resolution FT-ICR mass spectrometry and parts-per-billion mass accuracy enable direct characterization of nickel and vanadyl porphyrins in petroleum from natural seeps. *Energy Fuels* **2014**, *28*, 2454.

[18] H. J. Cooper, A. G. Marshall. Electrospray ionization Fourier transform mass spectrometric analysis of wine. *J. Agric. Food Chem.* **2001**, *49*, 5710.

[19] J. S. Garcia, B. G. Vaz, Y. E. Corilo, C. F. Ramires, S. A. S. A. Saraiva, G. B. Sanvido, E. M. Schmidt, D. R. J. J. Maia, R. G. Cosso, J. J. Zacca, M. N. Eberlin. Whisky analysis by electrospray ionization-Fourier transform mass spectrometry. *Food Res. Int.* **2013**, *51*, 98.

[20] W. Kew, I. Goodall, D. Clarke, D. Uhrín. Chemical diversity and complexity of Scotch whisky as revealed by high-resolution mass spectrometry. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 200.

[21] R. L. Malcolm, P. MacCarth. A proposal for implementing a reference collection of humic and fulvic acids, in *Trace Organic Analysis: A New Frontier in Analytical Chemistry*, (Eds: S. N. Chesier, H. S. Hertz). U.S. National Bureau of Standards, Maryland, **1979**, pp. 789–792.

[22] A. C. Stenson, A. G. Marshall, W. T. Cooper. 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*, 1275.

[23] J. B. Shaw, T.-Y. Lin, F. E. Leach, A. V. Tolmachev, N. Tolić, E. W. Robinson, D. W. Koppenaal, L. Paša-Tolić. 21 Tesla Fourier transform ion cyclotron resonance mass spectrometer greatly expands mass spectrometry toolbox. *J. Am. Soc. Mass Spectrom.* **2016**, *27*, 1929.

[24] Bokeh Development Team. Bokeh: Python library for interactive visualization. **2014**, URL http://bokeh.pydata.org (accessed 04/01/2017).

[25] B. P. Koch, T. Dittmar. From mass to structure: An aromaticity index for high-resolution mass data of natural organic matter. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 926.

[26] D. A. Keim, M. C. Hao, U. Dayal, H. Janetzko, P. Bak. Generalized scatter plots. *Inf. Vis.* **2010**, *9*, 301.

[27] A. Mayorga, M. Gleicher. Splatterplots: Overcoming overdraw in scatter plots. *IEEE Trans. Vis. Comput. Graph.* **2013**, *19*, 1526.

[28] D. B. Carr, R. J. Littlefield, W. L. Nicholson, J. S. Littlefield. Scatterplot matrix techniques for large N. *J. Am. Stat. Assoc.* **1987**, *82*, 424.

[29] T. Kind, O. Fiehn. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **2007**, *8*, 105.

[30] C. S. Hsu, K. Qian, Y. C. Chen. An innovative approach to data analysis in hydrocarbon characterization by on-line liquid chromatography-mass spectrometry. *Anal. Chim. Acta* **1992**, *264*, 79.

[31] W. Kew, J. W. T. Blackburn, D. Clarke, D. Uhrín. FTMSVisualization: verson 1 – Public. **2016**, *GitHub*, DOI: https://doi.org/10.5281/zenodo.165785.

[32] Available: https://github.com/wkew/FTMSVisualization.

[33] Available: https://wkew.github.io/FTMSViz/SRFA-plot.html.

[34] J. A. Cottam, A. Lumsdaine, P. Wang. Abstract rendering: out-of-core rendering for information visualization, in *Proceedings of SPIE – The International Society for Optical Engineering*, (Eds: P. C. Wong, D. L. Kao, M. C. Hao, C. Chen). **2013**, p. 90170K.