KairosMS: A new solution for the processing of hyphenated ultrahigh resolution mass spectrometry data

Remy Gavard,[†] Hugh E. Jones,[‡] Diana Catalina Palacio Lozano,[‡] Mary J. Thomas,[†] David Rossell,[¶,§] Simon E. F. Spencer,[¶] and Mark P. Barrow[*,‡]

MAS CDT, University of Warwick, Coventry, CV4 7AL, United Kingdom, Department of Chemistry, University of Warwick, Coventry, CV4 7AL, United Kingdom, Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom, and Department of Economics & Business, Universitat Pompeu Fabra, Barcelona, 08005, Spain

E-mail: M.P.Barrow@warwick.ac.uk

Abstract

The use of hyphenated Fourier transform mass spectrometry (FTMS) methods affords additional information about complex chemical mixtures. Co-eluted components can be resolved thanks to the ultra-high resolving power, which also allows extracted ion chromatograms (EICs) to be used for the observation of isomers. As such datasets can be large and data analyses laborious, improved tools are needed for data analyses and extraction of key information. The typical work-flow for this type of data is based upon manually dividing the total ions chromatogram (TIC) into several windows

[*]To whom correspondence should be addressed
[†]MAS CDT, University of Warwick, Coventry, CV4 7AL, United Kingdom
[‡]Department of Chemistry, University of Warwick, Coventry, CV4 7AL, United Kingdom
[¶]Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom
[§]Department of Economics & Business, Universitat Pompeu Fabra, Barcelona, 08005, Spain

of usually equal retention time, averaging the signal of each window to create a single mass spectrum, extracting a peak list, performing the compositional assignments, visualizing the results, and repeating the process for each window. By removing the need to manually divide a data set into many time windows and analyze each one, a time-consuming work-flow has been significantly simplified. An environmental sample from the oil sands region of Alberta, Canada, and dissolved organic matter samples from the Suwannee River Fulvic Acid (SRFA) and marine waters (Marine DOM) were used as a test-bed for the new method. A complete solution named KairosMS, was developed in the R language utilizing the Tidyverse packages and Shiny for the user interface. KairosMS imports raw data from common file types, processes it and exports a mass list for compositional assignments. KairosMS then incorporate those assignments for analysis and visualization. The present method increases the computational speed while reducing the manual work of the analysis when compared to other current methods. The algorithm subsequently incorporates the assignments into the processed data set, generating a series of interactive plots, EICs for individual components or entire compound classes, and can export raw data or graphics for off-line use. Using the example of petroleum related data, it is then visualized according to heteroatom class, carbon number, double bond equivalents (DBE), and retention time. The algorithm also gives the ability to screen for isomeric contributions and to follow homologous series or compound classes, instead of individual components, as a function of time.

Introduction

Complex mixtures such as petroleum, petroleum related samples and dissolved organic matter (DOM) are amongst the most complex and heterogeneous mixtures found in nature.[1,2] The study of these complex mixtures is crucial to improve refining techniques[3–5] and assess their environmental impacts.[6–8] The ultra-high resolution of Fourier transform mass spectrometry (FTMS)[9–13] has been beneficial to their study.[14–16] More recently, Orbitrap instruments were successfully used for oil-sand related samples[17,18] and DOM.[19] While Or-

bitrap instruments are more widely available and have lower costs, Fourier transform ion cyclotron mass spectrometry (FTICR MS) offers the highest performance for the study of complex mixtures.[18,19] The use of FT-based mass spectrometers have enabled researchers to observe previously unresolved molecules and gain a deep understanding of their composition.[20] Nevertheless, there still remains unexploited information to be extracted[21,22] and new techniques to observe record number of molecules are regularly developed.[23] Specifically, FTMS techniques allow researchers to resolve the mass of hundreds of thousands of molecules with a very high precision, but does not give any structural information that could be used for distinguishing molecules with the same mass but a different structure (isomers).[24] To address this issue, recent publications have used an online system of gas chromatography (GC),[24] liquid chromatography (LC)[25,26] and trapped ion mobility spectrometry (TIMS)[27] coupled to FTMS instruments.

While chromatography coupled to FTMS instruments is now well established, the data processing pipeline is struggling to keep pace with the instrumentation advances. The tools developed so far struggle to perform well with the ultra-high resolution and the complexity of the acquired spectra. Most software currently available such as OpenChrom,[28] MS-Dial,[29] XCMS,[30] MZmine,[31] MetAlign,[32] MathDAMP[33] and MS Resolver (Pattern Recognition Systems, Bergen, Norway)[34] use m/z binning in order to create a data matrix with m/z and retention time (Rt) as axes and intensity of the peak recorded by the analyzer. This strategy allows researchers to process the data rapidly, match peaks across samples and perform downstream analysis such as group comparisons, clustering, principal component analysis etcetera. However, these methods have not been designed to tackle the challenges posed by the natural variations of the m/z induced by the space-charge effects of the FTMS instruments along with the density of complex samples. Indeed, none offer the ability to recalibrate to tackle the space-charge effects. In addition, they have not been developed to work with low Signal-To-Noise (S/N) data and the need for denoising. This forces the user to raise the S/N threshold leading to potentially omitting informative peaks, losing some of

the benefits of the ultra-high resolution.

The m/z binning method used in the previous software assumes that each molecule m/z is far enough from any other so that sufficiently large bins can be used while avoiding having two different molecules in a single bin. Each bin will then contain an extracted ion chromatogram (EIC) that is comprised of peaks at a given m/z from scans spanning a retention time range. The peaks within an EIC of a molecular composition are defined by a unique m/z, intensity and retention time. Analyzing complex mixtures requires the ultra-high resolution to be able to separate molecules present within a very narrow m/z width.[4,35,36] For this reason, the use of large bins is detrimental as there is a high probability of having several EICs to appear in the same bin, losing the benefits of the ultra-high resolution. The use of small bins also poses great challenges as it increases the risk of excluding parts of the EIC, especially with FT instruments which are subject to space-charge effects resulting in shifting m/z during the experiment. The width of the bins would need to be dynamic as different sample complexities and instruments would influence the viable bin width. For example, analyzing data with an error range of 10 parts-per-million (ppm) does not pose the same challenges as techniques yielding errors of less than 1 ppm. Similarly, analyzing samples with hundreds of different molecules does not pose the same challenges as analyzing hundreds of thousands different molecules. The majority of the software cited earlier have been developed with the aim of the characterization of other sample types (e.g. biomolecules) using lower resolution instrumentation, hence they present significantly different characteristics and very different visualization tools. One issue is that they require the conversion of data to the MZxml format which multiplies the file size; an example of a hyphenated ultrahigh resolution data set in the region of 20-30 GB can become almost 100 GB, leading to increased computational overheads and making successful processing unviable. Other file export methods can sometimes place restrictions on the maximum number of peaks captured, which is not suitable for complex mixture analysis. Furthermore, the software does not allow incorporation of molecular assignments determined using external methods (e.g. in-

house algorithms or commercial software) which may be required for a researcher's workflow, especially for work in specialized fields. As a consequence, the current tools do not scale well for the particularly large and complex data sets often associated with hyphenated ultrahigh resolution experiments.

Presumably for these reasons, recent papers using hyphenated techniques with FTMS on complex mixtures have not made extensive use of the available software described previously to analyze their data.[24,25,37] MZmine, was used by Barrow et al.[24] to obtain a 3D representation of the data but not for the molecular composition analysis. Instead, we can distinguish two methods being employed to analyze hyphenated complex mixture data, and another one which has not yet been applied on complex mixtures analyzed by FTMS. The first strategy was employed by Barrow et al.[24] and Patriarca et al.[25] and relied on summing the signal for several time-frames to generate peak lists for different time ranges. Those peak lists were analyzed as individual mass spectra and molecular assignments generated for each. The information resulting from those assignments was used to create the plots for each peak list which were then used to follow the molecular evolution of the sample over time. This technique has the advantage of relying on an established workflow to analyze individual spectra but is labor-intensive and induces a loss of temporal resolution, since large time-frames are being grouped (*e.q.* 1 minute windows), meaning that variation within each averaged time frame may be lost.

The second strategy was used by Rüger et al.[37] and relies on an extensive signal processing routine coded for MATLAB which requires long computational time on a server (90-120 min with 20 to 60 GB of RAM) and a MATLAB license. The method has the advantage of not relying on other software, and performs the processing starting from raw signal. Strict filtering is applied based on the expected molecular properties[38] and makes use of a modified region of interest (ROI) algorithm to extract the EICs.[39] This ROI methods works best in the absence of noise and since it uses the recorded intensities to detect ROIs, low-intensity regions are unlikely to be detected.

The final method to isolate the EICs has been described using Kalman tracking[40] , although it hasn't been tested on hyphenated FTMS complex mixtures. This method relies on evaluating the probable position of the next data point using centroid data (discrete m/z with zero line widths) of plasma samples. While the Kalman tracking appears to perform well in the presence of hundreds of EICs, its performance hasn't yet been demonstrated with millions of data points, which are routinely obtained in centroid mode with complex mixtures.

Previous work demonstrated that peak list data and R can be used to develop new processing methods which improve the quality of the data.[41] So in order to address these issues, we have developed a method in the open source language R that can run from either a personal computer (PC) or a web application named KairosMS. The name derives from Kairos, an ancient Greek word used to describe time along with Chronos. The method uses developments in signal treatment, peak-picking and molecular assignments for complex mixtures analyzed with FTMS. It performs, when necessary, an m/z correction to compensate for the space-charge effects in complex mixtures, a quick matching of EICs together, and discarding of noise.

Once the EICs have been matched, a single mass list is generated where each EIC has been reduced to an m/z and an intensity, to facilitate assigning peaks to molecules using standard software. The short computing times allows for the trialing and optimization of different settings quickly. Peak assignments are then imported into our workflow, where we developed a suite of tools for data visualization and exploration. Standard figures such as double bond equivalents (DBE) plots[42–44] , class distributions or van Krevelen diagrams[45,46] for any specified time ranges, down to a scan-by-scan basis, can be generated within seconds. The high level of information is retained using this method, enabling new visualizations to be developed, such as the contribution of specific heteroatom classes and homologous series over each scan during the complete elution process.

Methodology

Sample Preparation

One oil sands process-affected water (OSPW) and two groundwater samples (G1 and G2) were obtained from the Athabasca region, along a groundwater flow path.[24] The samples were filtered under vacuum, acidified to pH 4.5, and extracted using Strata-X-A solid phase extraction sorbent (Phenomenex Torrance, CA, USA). The extracts were then methylated using BF3-methanol prior to analysis.

The reference material Suwannee River Fulvic Acid (SRFA) and a marine sample taken at 674 m depth from the North Pacific Ocean at the Natural Energy Laboratory of Hawaii Authority (NELHA)[47,48] used for analysis were acidified (0.01 M HCl), de-salted and concentrated by solid phase extraction.[25] The marine sample is hereafter referred as Marine DOM. The SRFA sample was diluted with ultra-pure water, enriched with 0.1% formic acid to a final concentration of 500 *ppm* in 5% methanol, 94.9% water and 0.1% formic acid. The freeze-dried SRFA powder was weighed and diluted to 500 *ppm* with 5% acetonitrile, 94.9% water and 0.1% formic acid.

A crude pyrolysis bio-oil sample with humidity less than 10% wt was produced using a mixture of softwoods as original material.[49] The samples were dissolved in acetone at a final concentration of 3 ppm and 1 mL was injected into a 30 m DB-5 column (0.25 mmID, 0.25 μ m).

Instrumentation

KairosMS capabilities and visualization tools were explored for the analysis of six hyphenated datasets acquired with different ultra-high resolution mass spectrometers. The experimental parameters and instrumentation are briefly described as follow:

GC-APCI-FTICR MS: The OSPW, G1 and G2 samples were analyzed using a 7890A GC (Agilent Technologies, Santa Clara, California, USA) connected to an atmospheric pressure

chemical ionization (APCI) source (Bruker Daltonik GmbH, Bremen, Germany) in positive mode which was itself used as ionization method and connected to a 12 T solariX FTICR mass spectrometer (Bruker Daltonik GmbH, Bremen, Germany) equipped with an Infinity Cell. The temperature was first held at 40 ℃ and increased at a rate of 20 ℃ min^{-1} until a final temperature of 280 ℃ was reached and held for 20 min. Broadband mass spectra in magnitude mode were acquired, a single zero fill and Sine-Bell apodization were applied before usage of a Fourier transform.

LC-Orbitrap: The SRFA and Marine DOM were obtained from Patriarca et al.[25] and acquired using an LTQ-Velos-Pro Orbitrap MS (Thermo Scientific, Germany) using an electrospray ionization source (ESI) in negative ion mode. The chromatography was performed using an Agilent PLRP-S poly(styrene/ divinylbenzene) column fitted with a precolumn filter (0.5 μm, Supelco Column Saver). After injection the acetonitrile percentage was increased from 5 to 20% during 2 min and maintained constant for 10 min before being increased to 40% at 13 min and held isocratic until 22 min. Finally the acetonitrile percentage was increased up to 90% and maintained for 10 min.

GC-APCI-FTICR MS 2xR: The bio-oil mass spectra were acquired using a 7890A GC (Agilent Technologies, Santa Clara, California, USA) connected to an APCI ion source in positive mode, coupled to a 7 T solariX 2xR FTICR mass spectrometer (Bruker Daltonik GmbH, Bremen, Germany) equipped with a ParaCell. It is worth noting that a 7 T FTICR equipped with a 2 ω detection has performance capabilities comparable to a 15 T instrument when operated at similar detection conditions of ω. The oven temperature was initialized at 60 ℃ and increased at a rate of 6 ℃ min^{-1} until a final temperature of 300 ℃ was reached. The oven was then maintained at 300 ℃ for 9 min. Broadband mass spectra were acquired, where a single zero fill and Sine-Bell apodization were applied before usage of a Fourier transform. In ftmsControl a processing is applied which removes 95% of the data points thanks to the removal of the electronic noise.

Statistical processing

Overview

The algorithm developed reads a mass list where each peak is defined by its m/z, intensity, and the retention time of the scan it was detected (Step 1). The mass list can be refined by cutting beginnings and/or ends of the retention time (Step 2) and/or low intensity peaks (Step 3). A method to detect and separate the EICs is applied (Step 4). After the detection of the EICs, it is possible to apply a recalibration method to compensate for any space-charge effect (Step 5). A final EIC matching is performed on the recalibrated mass list (Step 6) and a peak list is created containing only one pair of m/z and intensity for each EIC and used for molecular assignment (Step 7). The assigned peaks' information is merged with their corresponding EIC and a table containing all the EICs (assigned and unassigned) is created and used to create a large series of interactive figures (Step 8). The KairosMS workflow is shown in Figure 1 and further details of each step are described as follows. Throughout, the term "intensity" is used to refer to absolute abundances and "relative intensity" is used to refer to relative abundances.

Step 1: Extract the data

FTICR MS data were opened with Bruker DataAnalysis 4.2 (DA) software and the FTMS peak-picking method was used alongside a script to automatically perform a peak-picking for each mass spectra recorded over time. The FTMS peak picking algorithm involves the setting of a minimum S/N threshold, thus providing an initial level of noise filtering. For Orbitrap data, the ".raw" data file was converted to mzXML format and read directly into KairosMS. The information was structured into a matrix, where each row corresponds to a peak and columns to respectively retention time, m/z and intensity. The R code used for the processing made extensive use of the `Tidyverse`[50] packages and was implemented in a `Shiny`[51] interface.

The mass list is composed of peaks $i = 1, \ldots, n$, n being the number of peaks present. Each peak has an m/z, intensity and retention time respectively noted M_i, I_i and T_i, where

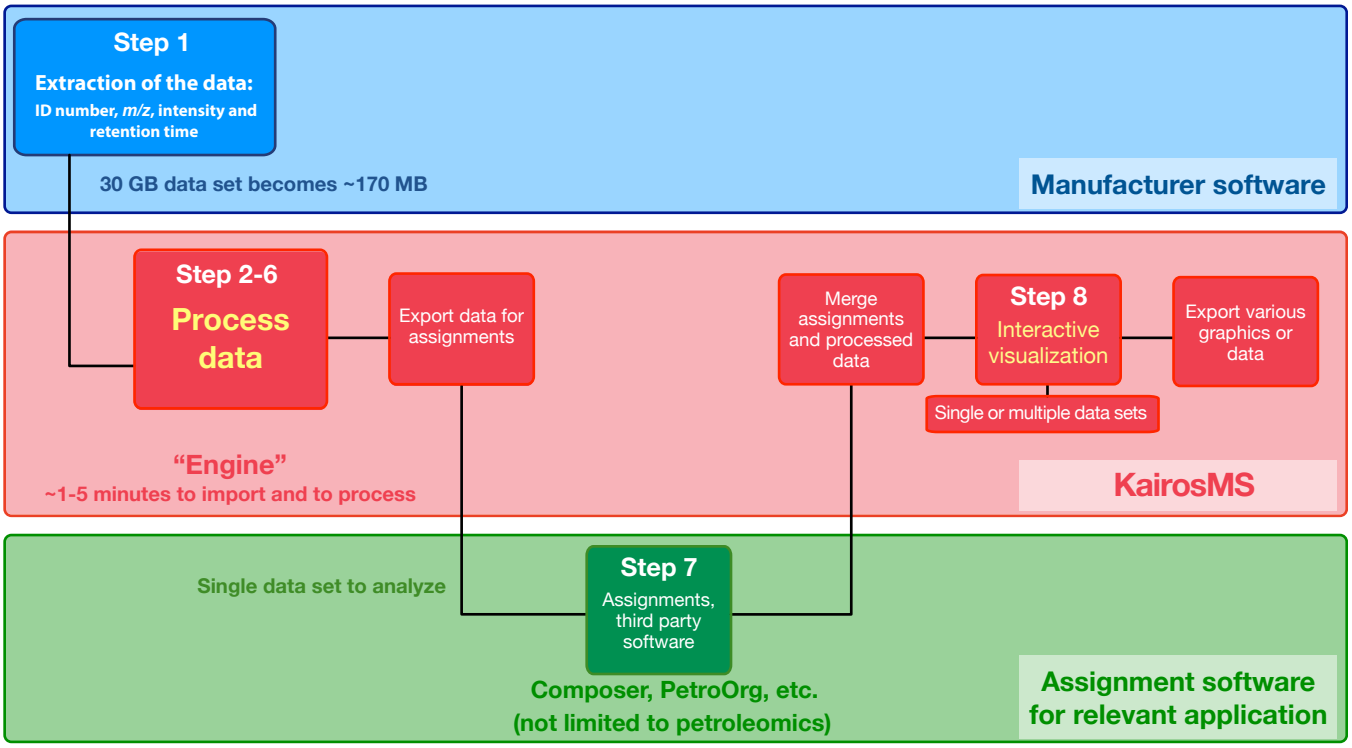

Figure 1: KairosMS workflow. In Step 1, a mass list containing an identification number, m/z, intensity, and retention time is extracted using the instrument manufacturer's software. The data processing is then performed during Steps 2-6 in KairosMS. KairosMS generates a single data set for compositional assignments (Step 7). Finally, a single data set or multiple data sets can be opened in KairosMS for interactive visualization and further data analysis (Step 8). All data and graphics are exportable.

$T_i \in t_1, t_2, t_3, \ldots, t_m$ and m is the number of scans.

Step 2: Trim the TIC

In chromatography, the beginning of acquisition often correspond to a baseline signal of noise and can optionally be removed. In consequence we offer the user the possibility to provide start and end points for the elution, and discard any peaks with T_i outside this range. This step will help to discard unnecessary information which will speed up the processing and reduce files sizes. Part of it can be retained if the user wishes to later on apply any baseline subtraction.

Step 3: Intensity Filter

Following an initial noise filtering, performed on the basis of the S/N in Step 1, a second filtering can be performed on the basis of peak intensities; Zhurov et al. [52] demonstrated that it is possible to discriminate between noise and genuine peaks using the log of intensities. A density plot of the $\log(I_i)$ was created to optionally help the user to decide on a level of intensity-filtering. Peaks with $\log(I_i)$ lower than the threshold specified by the user are discarded. Removing parts of the lowest-intensity peaks may be necessary to improve the downstream separation between noise and EICs.

Step 4: EIC matching algorithm

The Themis algorithm [41] was adapted to work through an additional dimension in order to perform the denoising and extract each EIC. As previously described, the m/z consistency was used but this time between scans to isolate the EICs. The difference was that due to intensity variations inherent to the chromatography elution, the intensity parameter had to be excluded from the equation used by Gavard et al. [41] The method performed well in those conditions but a threshold for the minimum number of consecutive peaks had to be implemented in order to reduce false positive EICs arising from the combination of too few data points. The user can adjust this parameter by considering the experiment hardware, the sample and the conditions of acquisition: in some experimental conditions combining GC and simple oil-related samples, an EIC of a low abundance specie can be as short as 3

to 5 scans. As described in Themis, a population separation threshold was automatically calculated but control was given to the user to change this value if deemed necessary.

Step 5: Recalibration

This recalibration method relies on the intra EIC variations, the first step is to perform a primary matching of the EICs as described in Step 4. The m/z was reconverted into Hz using an adaptation of equation (1)[53] taken from Barry et al..[54]

$$M_i = \frac{A}{F_i} + \frac{B}{F_i^2} \tag{1}$$

The frequency was calculated using equation (2), derived from equation (1), and using instrument-specific values A and B, provided by the user. For a Bruker FTICR MS dataset, the A and B values are respectively named ML1 and ML2 and can be found within the method file within each data directory.

$$F_i = \frac{A + \sqrt{A^2 + 4BM_i}}{2(M_i)} \tag{2}$$

Let $\underline{F}_j$ be the highest frequency in peak in the j^{th} EIC. For each peak i within EIC j we define the frequency shift $\widetilde{F}_i = F_i - \underline{F}_j$. For each time $t \in \{t_1, \ldots, t_m\}$ we compute the mean frequency shift $S(t)$

$$S(t) = \sum_{i\,:\,T_i = t} \widetilde{F}_i / \sum_{i=1}^{n} I(T_i = t) \tag{3}$$

A loess model[55] was fitted to the relation between the scan total intensity and the mean frequency shift $S(t)$. We denoted the new $S(t)$ predicted using the loess model $\widehat{S(t)}$. The modeling helps to ensure that if some scans were too shifted to be picked up, they would still get the appropriate corrections in regards to their expected shift because of the total intensity of the scan.

We search for any $\widehat{S(t)}$ which is $\widehat{S(t)} - mean(\widehat{S(t)}) > 2 \times sd(\widehat{S(t)})$. Any peak i within

the previously identified $\widehat{S(t)}$ is corrected by calculating

$$\widetilde{F_i} = F_i + \widehat{S(T_i)} - \frac{1}{m} \sum_{t \in t_1,\ldots,t_m} \widehat{S(t)} \tag{4}$$

The remaining peaks have $\widetilde{F_i} = F_i$. All the M_i are subsequently updated using (1) with the corrected frequency $\widetilde{F_i}$. Using the updated M_i the EIC matching described in Step 4 is performed again using the same parameters. We now calculate $\widetilde{S(t)}$ similarly to $S(t)$, but based on $\widetilde{F_i}$ instead of F_i. The final frequencies are obtained by calculating $F_i^* = \widetilde{F_i} + \widetilde{S(t)} - \min(\widetilde{S(t)} : t \in t_1, \ldots, t_m)$ and the corresponding m/z was calculated. The updated m/z are then used in Step 6.

If the A and B coefficient from equation (1) are not available an equivalent procedure can be applied without going into the frequency domain, by calculating the shift in *ppm* and applying the correction directly on the m/z. The described recalibration method, however, does not rely on prior knowledge of the true m/z of one or more peaks.

Step 6: Processing

If recalibration was performed in step 5, the density plot observed previously might have changed. In consequence, KairosMS offers the user the opportunity to change the settings used for the pairing (Step 4). Once the pairing described in Step 4 has been performed, the user has an overview of the number of isolated EICs. The number of EICs which had 2 or more peaks from the same retention time and went through an additional refinement is also presented, and a high value will indicate that the previous settings needs to be tightened.

Step 7: Molecular assignment

Once the EICs were isolated, a mass list was created using the sum of intensities within each EICs and the mean m/z of each EIC. This standard mass list can be read into third party molecular assignment software (*e.g.* Composer, PetroOrg, in-house scripts etc.), depending on the type of sample. The assignments for each EICs were merged with the data for the peaks within the EIC and stored as a R data table object called a tibble.[56] The

columns containing the information from the assignment remained empty for peaks within the unassigned EICs. No information is therefore removed from the original peak list and the assignments could be redone later if necessary.

Step 8: Data analysis tools

Currently, KairosMS produces a suite of visualization tools commonly used in petroleomics due to the need to visualize complex mixture data. These include displaying the DBE vs carbon number, percentage intensity contribution of the different classes, evolution of the intensity over time for each class, homologous series and molecules, van Krevelen diagrams, breakdown of the contribution of each atom present in the sample, area under the curve (quantification) for heteroatom classes to molecules, and principal component analysis. Note that in addition to using data from hyphenated mass spectrometry experiments, direct infusion data can also be analyzed, visualized, and compared. Steps 1 to 7 are performed on each sample individually, leading to the characterization of the majority of EICs. The comparison between samples is then based on the use of molecular assignments, which are determined, merged with the EICs, and compared during steps 7-8. Comparisons between several hyphenated samples that have been analyzed using KairosMS are also provided in Figures 5 and 6. KairosMS was coded in R and implemented into a Shiny interactive interface, allowing the user to see the plots as the analysis proceeds the process and adjust the parameters accordingly. KairosMS can be run either locally on a personal computer, or online through a server or a local network.

Results and discussion

Data processing

A screenshot presenting the interface of KairosMS is depicted in Figure S1. A detailed description of the processing steps in KairosMS for the OSPW sample are detailed below.

In order to process the OSPW dataset, the TIC between 0 and 9 minutes was trimmed

and the intensities below 73,130 were filtered out. This led to a reduction from 2,963,880 to 2,086,325 data points (29.61% removed). The denoising and EIC extraction method from Step 3 was applied to enable the recalibration method described in Step 4.

The optional recalibration step allowed us to correct for the space-charge effects without using any prior knowledge about the sample. The calibration was performed using matching conditions of 20 consecutive peaks. Figure 2 shows the evolution of the average frequency shift for each retention time. One can notice the similarities with the profile of the original f shift compared to the TIC in Figure 4. The recalibration performed attenuate the space-

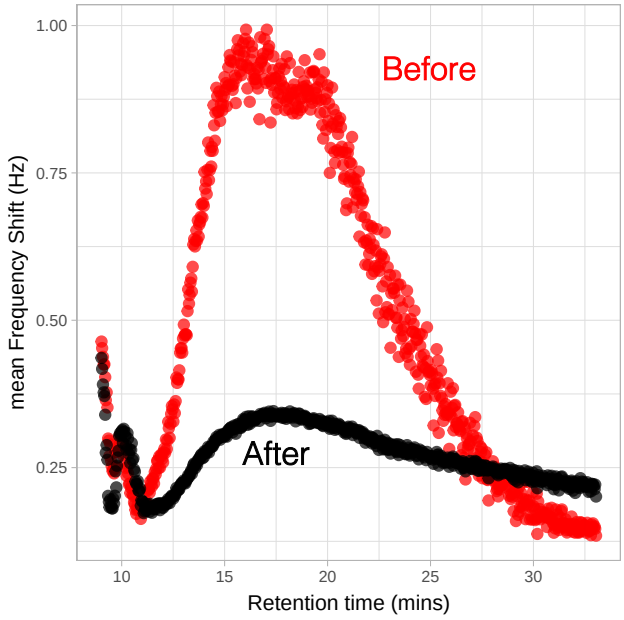

Figure 2: Average frequency shift in Hz within EICs for each retention time before recalibration (red) and after (black) for the OSPW sample.

charge effects, as the calculated frequency shifts after recalibration shown in black in Figure 2. Before recalibration, the RMS error, which was calculated using the difference between the assigned m/z and the experimental m/z of each peak of each EIC, was 2 ppm. After recalibration, the RMS error decreased to 0.4 ppm (Figure 3).

Using a minimum EIC length of 20 scans, 1,473,579 of the 2,086,325 peaks were kept (29.37% removed). The complete process typically takes tens of seconds to perform and could be improved further with the use of parallel computing. 6,540 distinct EICs were

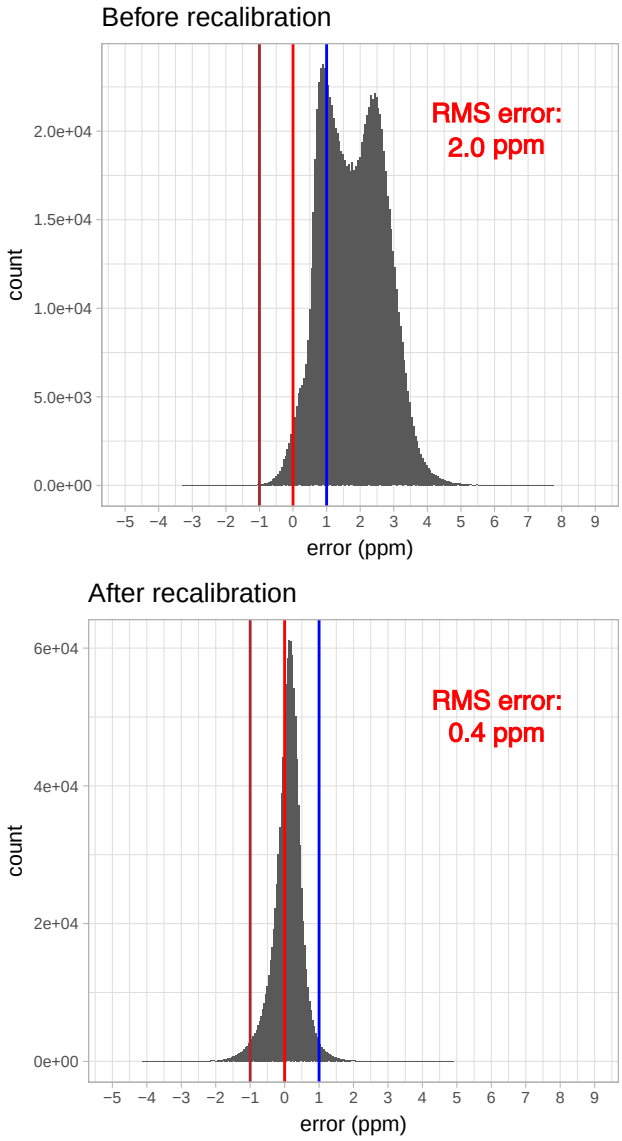

Figure 3: Histogram of the mass error in *ppm* of each peak for each EIC with the assigned m/z.

isolated with this process. We compared the TIC before and after processing to make sure that no critical features were discarded and that the shape of the TIC had been preserved. Figure 4 shows there were no noticeable differences before and after processing, ensuring that all major peaks had been preserved and matched to an EIC.

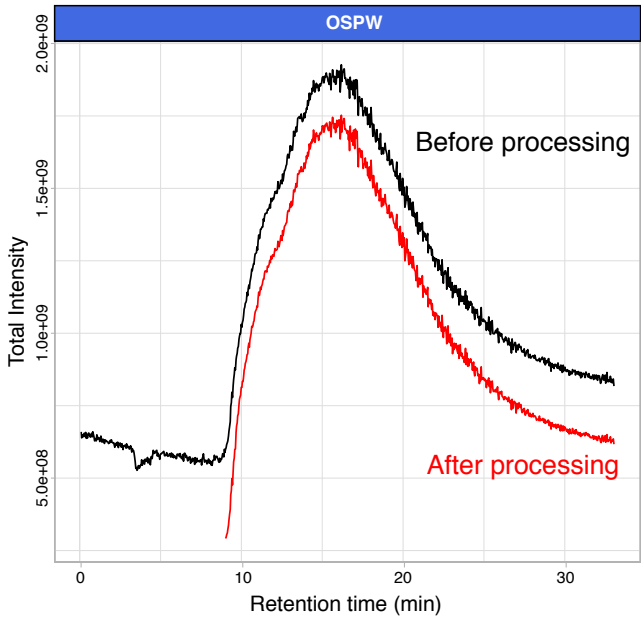

Figure 4: Comparison of TICs before (top line, black) and after (bottom line, red) processing.

Since current petroleomics software for molecular assignments (Composer, PetroOrg) were designed to assign molecular composition on a single spectrum, each EIC was summarized by a single m/z and intensity. The average m/z of each EIC was used while the intensities of the EICs were individually summed. The resulting spectrum for the OSPW sample is presented in Figure S2.

Once the assignments were performed using Composer, the molecular composition for each peak assigned was re-attributed to their respective EIC and a special data frame format (tibble) was used to store all the information. Tibbles make subsetting easier than traditional data-frames and allow the mixing of several types of data (*e.g.* characters, numeric, factor). Subsetting is crucial at a later stage as we analyze the data and explore specific classes, retention times, m/z. The unassigned EICs were preserved so that the processed data,

saved as .csv, could be reprocessed in the future.

The data processing described above for OSPW can be performed in about five minutes. The raw GC data obtained for the OSPW sample is about 24 GB in size and is reduced after steps 1-7 in KairosMS to a final data set size of 163.4 MB (0.68% of the original data set size). Similarly, the processing steps were performed to extract the EICs of the remaining data sets: G1, G2, SRFA, Marine DOM and the bio-oil. A final data set size in the range of 93-117 MB was obtained for each sample.

Data analysis tools

These data sets contain the assigned and unassigned EICs and are used for further data analysis in the final step in KairosMS. KairosMS enables the user to interpret hyphenated data and to study the molecular composition more efficiently than before, a list of some of the visualization tools available are listed below:

- TIC and mass spectra visualization at a desired retention time
- Interactive double bond equivalent (DBE) versus carbon number plots: individual or multiple classes with the possibility of the extraction of the EICs of each data point of the DBE plot. The DBE plots can be visualized in different retention time frames as desired for the user.
- Interactive class distribution: heteroatomic classes can be visualized in different retention time ranges. The heteroatomic class distribution can be fixed or automatically updated for desired retention times.
- Mass spectra, DBE and carbon number distribution: the sum intensities versus carbon number and DBE can be plotted by individual heteroatomic class.
- Interactive van Krevelen diagrams: van Krevelen diagrams can be plotted as a function of time and the heteroatomic classes can be selected for the user. The user can also

extract the EICs in each data point in the van Krevelen diagram.

- Interactive EICs visualization: a total EIC per heteroatomic class or homologous series can be extracted. The total area under the curve (AUC) per class or homologous series is calculated by KairosMS and can be exported in a .csv file. Additionally, the user can define either: an m/z, assigned molecular formula or a custom molecular formula to be visualized from the data set. The EICs can be visualized for a particular or multiple heteroatomic classes within a certain ranges of carbon number and DBEs as desired by the user.

- Filters: the data visualization includes data filters either by class, DBE, isotopic compositions, retention time, adduct type or by sample.

- Plot settings: all plots can be individually exported in .png .pdf .eps or .tiff format. The figures can be faceted by the sample-identifying name. Additionally, the data point size in DBE plots can be changed or plotted in a log10 scale. Class distribution figures can be plotted in stack bars or bar charts and the coordinates can be flipped. The data in EICs can be visualized and exported with a defined dot size and the EICs can be exported in a defined retention time domain. Alternatively, the figures can be generated in an external software by downloading the data from KairosMS in a .csv file. The figure format of the graphic can be changed by using different color schemes, different graphic resolution, figure size and data legend size.

These capabilities are shown in the Movie provided in the Supporting Information.

Applications - Data analysis visualization

GC-FTICR MS for the analysis of OSPW and groundwater

As it often becomes necessary to compare datasets[24,25] , we extended the capabilities of KairosMS so that one can compare several samples after they've been processed thanks

to the level of detail of information retained during the processing. No limits have been set to the number of datasets to compare but using more files require longer computation times and more memory. The previous OSPW sample was compared to two groundwater samples (G1 and G2). The first step was to use the classes contribution function to observe the key differences between the samples (see Figure 5A). As shown in Figure 5A, the class distribution of all samples is shifted towards higher oxygen-containing species at higher retention time. Additionally, it is noticeable that the oxygen content of the OSPW sample is comparatively different to the groundwater sample. For instance, at low retention time, the relative abundance of the oxygenated classes is higher in the OSPW in comparison to the groundwater samples and lower relative oxygen content species elute from the GC column at higher retention time in comparison with the groundwater samples.

Using the observations made in the class distribution in Figure 5A, the O_2[H] class was selected for further analysis and further broken down in order to observe independently each homologous series, where it can be seen that the predominant DBEs were 2.5, 3.5 and 4.5 (Figure 5B). An enlarged version of Figure 5B with the complete retention time is available in the SI as Figure S5. The AUC of the homologous series shows an increased contribution at higher retention times as the DBE increases. Thus, species with higher DBE have increased boiling point and therefore elute from the GC column at higher retention time.

Since we observed differences in the O_2 [H] class, we've explored it further and observed its evolution with a scan by scan resolution. In Figure S3 we notice that G1 and G2 display the exact same elution profile while the OSPW has remarkable differences. In Figure S4, the isotopes for this particular molecular composition were also observed.

KairosMS enables a rapid screening of EICs for each molecular composition identified by giving the capability to display all the EICs matching specific features such as a specific m/z or m/z range, belong to a specific heteroatom class, DBE range or range of carbon numbers. Once differences are identified for a specific heteroatom class, it becomes possible to display the EIC for individual molecular composition within that class.

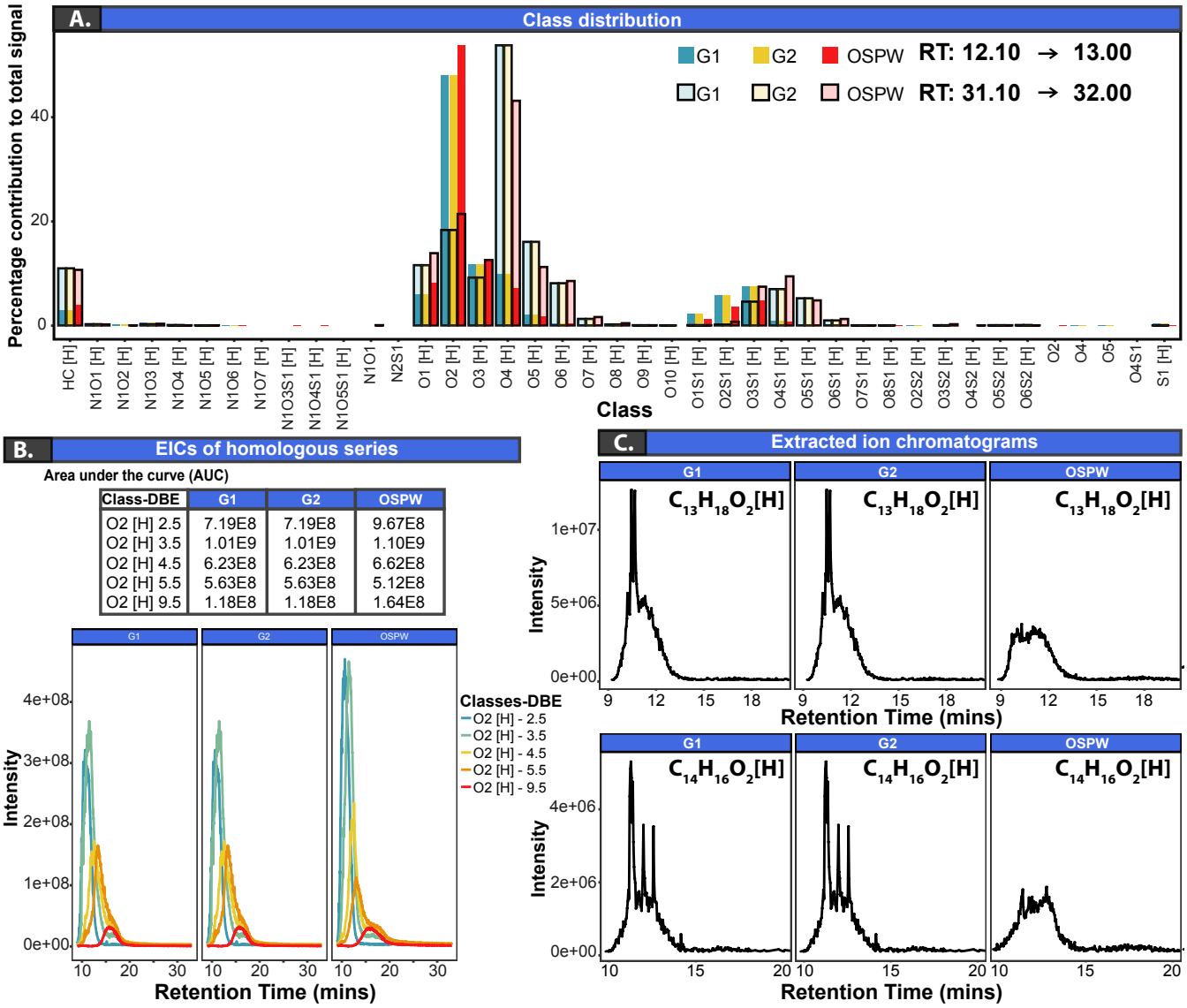

Figure 5: (**A**) Heteroatom class distribution for retention times 12.10-12 *mins* and 31.10-32 *mins* (**B**) Extracted ions chromatograms (EICs) of selected homologous series from the O_2[H] class (**C**) (EICs) of selected molecular composition for the oil sands process-affected water (OSPW) and two groundwater samples (G1 and G2).

After screening EICs corresponding to the O_2[H] class, it was found noticeable differences between the groundwater and OSPW EICs. For instance, the EICs corresponding to $C_{13}H_{18}O_2$[H] and $C_{14}H_{16}O_2$[H] in Figure 5C shown a distribution of peaks at low retention time in the groundwater samples that were not detected in the OSPW. This indicates the presence of different isomers and different ratios between isomers which can then be isolated and further investigated.

LC-Orbitrap: Dissolved Organic Matter

KairosMS capabilities to handle dissolved organic matter analyzed in an online LC - Orbitrap system were tested using a Marine DOM and SRFA samples. Each dataset was first processed using KairosMS and the results exported as .csv files. Finally, both files were loaded into KairosMS for data exploration and comparison. The TIC and the mass spectra at two different retention times of SRFA and the marine samples can be seen in Figure 6A. As shown in this figure, the compositions with higher m/z elute at higher retention time in both samples. Figure 6B and 6C shows the differences in molecular composition between the two samples for the complete run. The class distribution shown in Figure 6C can be modified within KairosMS to show any retention time range in order to highlight the difference of composition at any stage of the acquisition. It is noticeable that in contrast to GC, the components eluting from the LC column at higher retention time correspond to species with lower oxygen containing species. The major new capability enabled by this work is the ability to track specific heteroatom classes across every scan acquired. For instance, Figure 6B shows the heteroatom class O_4[H] and O_{10}[H] intensity across the complete retention time, allowing the user to immediately highlight the differences between the two samples. In similarity with the EICs by homologous series, the AUC of the classes is calculated in KairosMS.

To further explore the differences between the two samples, it is also possible to plot side by side or to overlap the van Krevelen diagrams for each sample (Figure S8). As in

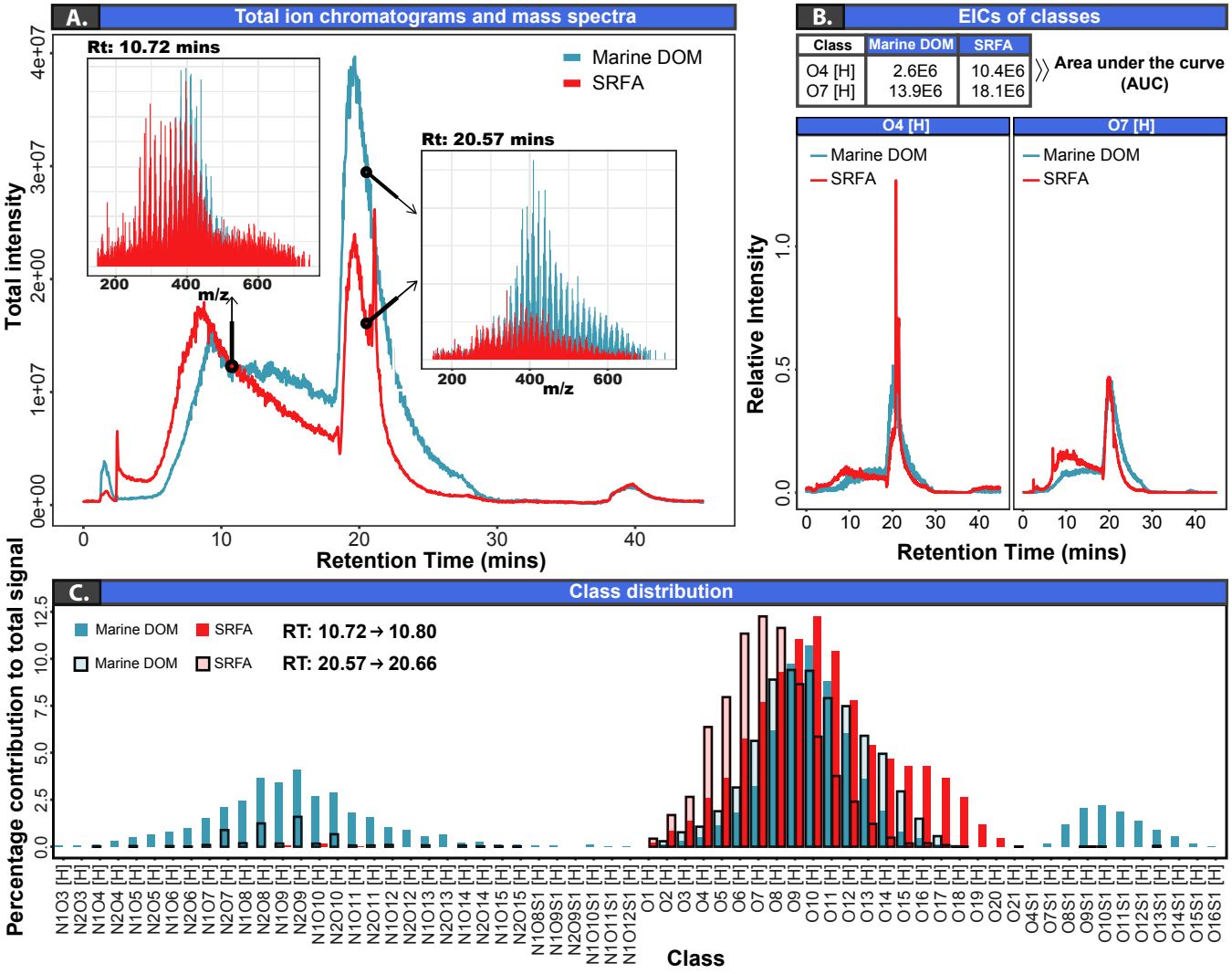

Figure 6: (**A**) Total ion chromatogram (TIC) with the mass spectra for retention times 10.72 *mins* and 20.57 *mins*. (**B**) Extracted ion chromatograms (EICs) for the heteroatom classes O_4[H] and O_7[H] with associated area under the curve (AUC). (**C**) Heteroatom class distribution for retention times 10.72-10.80 *mins* and 20.57-20.66 *mins* for the SRFA and Marine DOM samples.

Figure S6, this plot can be refined to observe any specific retention time range. The figure demonstrates significant differences between the two samples, especially in the region below H/C of 1.

GC FTICR MS 2xR

A bio-oil sample analyzed using a solariX 2xR FTICR MS. The 2ω detection from this instrument allows to either operate the instrument at twice the speed for the same resolving power or to double the resolving power if the speed is kept the same than with conventional 2ω instruments. For gas chromatography, acquiring at twice the speed is particularly useful for samples presenting a large number of isomers.

The van Krevelen diagram and the DBE plots of the classes O_2[H], O_3[H], O_4[H] and O_5[H] obtained within the total retention time in the GC column (see Figure 7A and 7B-C respectively). The EICs of the compositions in both type of plots can be visualized by using KairosMS. In contrast to DBE plots, van Krevelen data points can contain multiple EICs of compositions with the same H/C and O/C values (e.g. $C_8H_{12}O_4$[H] and $C_{12}H_{18}O_6$[H]. In comparison with the previous samples, the bio-oil has a remarkable number of potential isomers. For instance, the composition $C_8H_{12}O_4$[H] shows the presence of at least 71 potential isomers. It is important to notice that species with higher carbon number, higher DBE and higher oxygen content eluted from the column at higher retention times.

The EIC shown in Figure S9 shows the presence of at least 35 potential isomers. In order to assess KairosMS capabilities to isolate such challenging EIC, we've overlapped the EIC obtained using DA and KairosMS and it showed a complete overlap between the two with only minor differences within the noise baseline due to the necessary intensity threshold resulting from the peak picking.

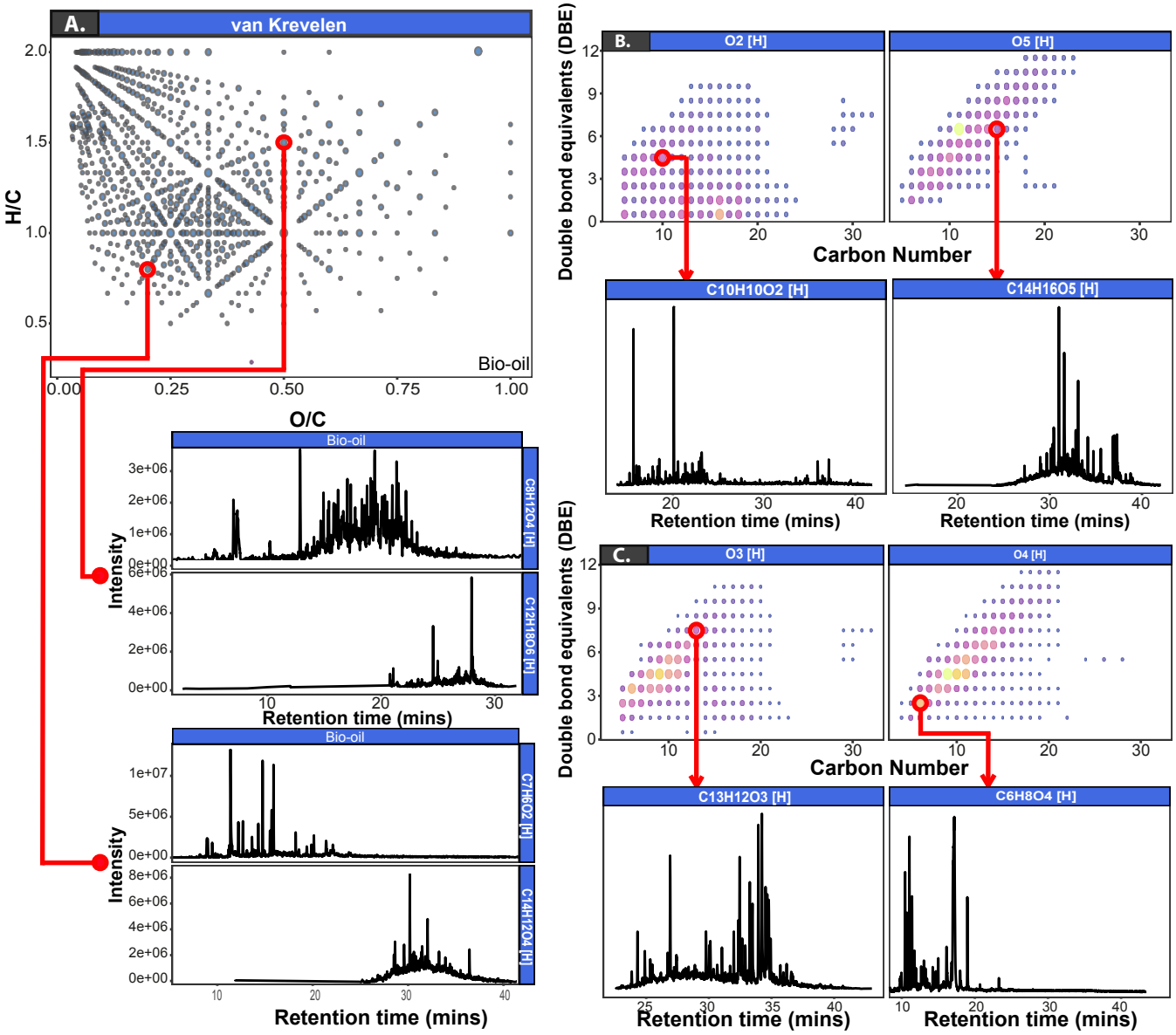

Figure 7: (**A**) van Krevelen diagram with extracted ion chromatograms (EICs) displayed for two different points on the plot. (**B**) Double bond equivalents (DBE) vs. carbon number of the O_2[H] and O_5[H] heteroatom classes with associated EICs for two data points on the plot. (**C**) DBE vs. carbon number of the O_3[H] and O_4[H] heteroatom classes with associated EICs for two data points. Note that each data point in a van Krevelen diagram represents the sum of multiple molecular formulae, whilst each data point in a DBE vs. carbon number plot is a single molecular formula.

Other applications

KairosMS also provides the ability to search for any specific EIC of an identified molecule. The user can search using m/z or molecular composition but can also display all the EICs with specific features such as heteroatom class, carbon number, DBE. This allows a researcher to quickly determine differences between elution profiles at the molecular level.

Finally, by using the intensity information of all the assigned EICs, the calculation of the elemental contribution within each samples can be swiftly obtained and the percentages for each elements calculated as depicted in Figure S7.

KairosMS was also used to process peptide digest data, analyzed by LC-FTICR MS as pictured in Figure S10. Even without providing molecular assignments, in this particular case, KairosMS was able to quickly and accurately calculate the area under the curve for all isolated EICs, providing useful quantification data.

Conclusion

Using hyphenated Fourier transform mass spectrometry, additional separation methods such as chromatography provide further insights about complex chemical mixtures, especially for the observation of isomers. The data analysis for such experiment previously relied on long and laborious manual work. The typical work-flow for this type of data is based upon manually merging mass spectra over a series of retention time ranges, extraction of each peak list, assigning compositions, visualizing the results, and repeating the process for each retention time range. KairosMS addresses those issues by removing the need to manually divide a data set into many time windows and analyze each one, while also preserving the time resolution. Data is first extracted as a mass list to reduce computing time, relying on peak picking and centroid detection existing algorithms. KairosMS then processes the data, can attenuate space-charge effects and existing peak assignments methods are reused. The recalibration is currently best suited for FTICR MS instruments analyzing complex mixtures

and needs to be optimized for datasets such as biomolecules. The method could be further improved by implementing prior knowledge about one or more peaks. KairosMS demonstrated its abilities on a wide range of samples (petroleum, environmental, dissolved organic matter results and biomolecules) from different types of analyzers and types of chromatography, turning hyphenated ultra-high resolution mass spectrometry into a regular tool for those samples. The capability to quickly visualize EICs from any class, homologous series, m/z, or molecular assignment helps the user to fully exploit the information enabled by the chromatography, especially in presence of multiple isomers, paving the way to shift from relying solely on molecular compositions to understanding the structure of the molecules in complex mixtures. Amongst the features for comparing many complex datasets, KairosMS also includes options for hierarchical clustering and principal component analysis. It should be noted KairosMS can be used for data analysis, visualization, and sample comparison for direct infusion data, in addition to hyphenated data sets. Using the same simple file format for the processed data, we were able to simultaneously browse and compare samples, saving the user from repetitive tasks. It is expected that with such information made available, additional new visualization methods can be developed to help tackle the challenges posed by the volume of data.

Acknowledgement

R.G., H.E.J. and M.J.T thanks EPSRC for a PhD studentship through the EPSRC Centre for Doctoral Training in Molecular Analytical Science, grant number EP/L015307/1. M.P.B and R.G thank Warwick Ventures for the HEIF Impact Fund. D.C.P.L thank the Newton Fund award (reference number 275910721), Research Agreement No. 5211770 UIS-ICP, and COLCIENCIAS (project No.FP44842-039-2015) for funding. M.P.B. thanks John V. Headley, Kerry M. Peru and Jason Ahad for the groundwater and OSPW samples. R.G. thanks Jeffrey Hawkes and Claudia Patriarca for the SRFA and Marine DOM data. D.R. was partially supported by Ramon y Cajal Fellowship RYC-2015-18544 from Ministerio

de Economia y Competitividad (Government of Spain), , Ayudas Investigación Cientifica Big Data (Fundación BBVA) and Programa Estatal I+D+i I+D+i PGC2018-101643-B-I00 (Government of Spain). M.P.B and R.G. also thank David Stranz (Sierra Analytics) for his valuable contributions.

Supporting Information Available

Supporting Information Available:

- Figure S1: Screenshot presenting KairosMS interface.
- Figure S2: Mass spectrum created using the EICs extracted to be used for molecular assignments.
- Figure S3: Comparison of the elution of the O_2[H] class contribution between an OSPW and two Groundwater samples using a scan by scan resolution.
- Figure S4: EICs for the monoisotopic form and isotopologues of $C_{16}H_{26}O_2$ [H] for the G1, G2 and OSPW samples (G1 and G2 perfectly overlap). The same retention for the isotopologues is further evidence for the compositional assignment.
- Figure S5: Elution of DBE series (homologeous series) comprising the O_2[H] class.
- Figure S6: Percentage of contribution to the total signal, for all the classes identified in the SRFA and Marine DOM samples.
- Figure S7: Elemental contributions for the samples SRFA and Marine DOM, based on all the assigned EICs.
- Figure S8: van Krevelen diagram of the H/C ratio vs O/C ratio for the Marine DOM and SRFA samples.
- Figure S9: Comparison of the EIC of the same molecular assignment as seen in DA and KairosMS after peak picking at S/N 1.

- Figure S10: EIC from a peptide digest of ubiquitin analyzed by LC-FTICR MS.
- Video S11: A video showing interactive data visualisation using KairosMS.

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Hertkorn, N.; Ruecker, C.; Meringer, M.; Gugisch, R.; Frommberger, M.; Perdue, E. M.; Witt, M.; Schmitt-Kopplin, P. High-precision frequency measurements: Indispensable tools at the core of the molecular-level analysis of complex systems. *Anal. Bioanal. Chem.* **2007**, *389*, 1311–1327.

(2) Barrow, M. P. Petroleomics: study of the old and the new. *Biofuels* **2010**, *1*, 651–655.

(3) Mullins, O. C., Sheu, E. Y., Hammami, A., Marshall, A. G., Eds. *Asphaltenes, Heavy Oils, and Petroleomics*; Springer New York: New York, NY, 2007.

(4) Marshall, A. G.; Rodgers, R. P. Petroleomics: Chemistry of the underworld. *Proc. Natl. Acad. Sci. U.S.A* **2008**, *105*, 18090–18095.

(5) Hsu, C. S.; Hendrickson, C. L.; Rodgers, R. P.; McKenna, A. M.; Marshall, A. G. Petroleomics: Advanced molecular probe for petroleum heavy ends. *J. Mass Spectrom.* **2011**, *46*, 337–343.

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

(7) Stubbins, A.; Dittmar, T. Illuminating the deep: Molecular signatures of photochemical alteration of dissolved organic matter from North Atlantic Deep Water. *Mar. Chem.* **2015**, *177*, 318–324.

(8) Headley, J. V.; Peru, K. M.; Barrow, M. P. Advances in mass spectrometric characterization of naphthenic acids fraction compounds in oil sands environmental samples and crude oil - A review. *Mass Spectrom. Rev.* **2016**, *35*, 311–328.

(9) Comisarow, M. B.; Marshall, A. G. Fourier transform ion cyclotron resonance spectroscopy. *Chem. Phys. Lett.* **1974**, *25*, 282–283.

(10) Comisarow, M. B.; Marshall, A. G. Selective-phase ion cyclotron resonance spectroscopy. *Can. J. Chem.* **1974**, *52*, 1997–1999.

(11) Comisarow, M. B.; Marshall, A. G. Frequency-sweep fourier transform ion cyclotron resonance spectroscopy. *Chem. Phys. Lett.* **1974**, *26*, 489–490.

(12) Amster, I. J. Fourier transform mass spectrometry. *J. Mass Spectrom.* **1996**, *31*, 1325–1337.

(13) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. Fourier Transform Ion Cyclotron Resonance Mass Spectromeyry: A Primer. *Mass Spectrom. Rev.* **1998**, *17*, 1–35.

(14) Barrow, M. P.; Burkitt, W. I.; Derrick, P. J. Principles of Fourier transform ion cyclotron resonance mass spectrometry and its application in structural biology. *The Analyst* **2005**, *130*, 18.

(15) Koch, B. P.; Ludwichowski, K. U.; Kattner, G.; Dittmar, T.; Witt, M. Advanced characterization of marine dissolved organic matter by combining reversed-phase liquid chromatography and FT-ICR-MS. *Mar. Chem.* **2008**, *111*, 233–241.

(16) Dittmar, T.; Stubbins, A. *Treatise Geochemistry Second Ed.*, 2nd ed.; Elsevier Ltd., 2013; Vol. 12; pp 125–156.

(17) Headley, J. V.; Peru, K. M.; Janfada, A.; Fahlman, B.; Gu, C.; Hassan, S. Characterization of oil sands acids in plant tissue using Orbitrap ultra-high resolution mass

spectrometry with electrospray ionization. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 459–462.

(18) Zhurov, K. O.; Kozhinov, A. N.; Tsybin, Y. O. Evaluation of high-field orbitrap fourier transform mass spectrometer for petroleomics. *Energ. Fuel.* **2013**, *27*, 2974–2983.

(19) Hawkes, J. A.; Dittmar, T.; Patriarca, C.; Tranvik, L.; Bergquist, J. Evaluation of the Orbitrap Mass Spectrometer for the Molecular Fingerprinting Analysis of Natural Dissolved Organic Matter. *Anal. Chem.* **2016**, *88*, 7698–7704.

(20) Cho, E.; Witt, M.; Hur, M.; Jung, M. J.; Kim, S. Application of FT-ICR MS Equipped with Quadrupole Detection for Analysis of Crude Oil. *Anal. Chem.* **2017**, *89*, 12101–12107.

(21) Krajewski, L. C.; Rodgers, R. P.; Marshall, A. G. 126âĂŤ264 Assigned Chemical Formulas from an Atmospheric Pressure Photoionization 9.4 T Fourier Transform Positive Ion Cyclotron Resonance Mass Spectrum. *Anal. Chem.* **2017**, acs.analchem.7b02004.

(22) Smith, D. F.; Podgorski, D. C.; Rodgers, R. P.; Blakney, G. T.; Hendrickson, C. L. 21 Tesla FT-ICR Mass Spectrometer for Ultrahigh-Resolution Analysis of Complex Organic Mixtures. *Anal. Chem.* **2018**, *90*, 2041–2047.

(23) Palacio Lozano, D. C.; Gavard, R.; Arenas-Diaz, J. P.; Thomas, M. J.; Stranz, D. D.; Mejía-Ospino, E.; Guzman, A.; Spencer, S. E. F.; Rossell, D.; Barrow, M. P. Pushing the analytical limits: new insights into complex mixtures using mass spectra segments of constant ultrahigh resolving power. *Chem. Sci.* **2019**, *10*, 6966–6978.

(24) Barrow, M. P.; Peru, K. M.; Headley, J. V. An added dimension: GC atmospheric pressure chemical ionization FTICR MS and the athabasca oil sands. *Anal. Chem.* **2014**, *86*, 8281–8288.

(25) Patriarca, C.; Bergquist, J.; SjÃűberg, P. J. R.; Tranvik, L.; Hawkes, J. A. Online HPLC-ESI-HRMS Method for the Analysis and Comparison of Different Dissolved Organic Matter Samples. *Environmental Science & Technology* **2018**, *52*, 2091–2099, PMID: 29241333.

(26) Kim, S.; Kim, D.; Kim, S.; Son, S.; Jung, M. J. Application of Online Liquid Chromatography 7 T FT-ICR Mass Spectrometer Equipped with Quadrupolar Detection for Analysis of Natural Organic Matter. *Anal. Chem.* **2019**, *91*, 7690–7697.

(27) Benigni, P.; Thompson, C. J.; Ridgeway, M. E.; Park, M. A.; Fernandez-Lima, F. Targeted high-resolution ion mobility separation coupled to ultrahigh-resolution mass spectrometry of endocrine disruptors in complex mixtures. *Anal. Chem.* **2015**, *87*, 4321–4325.

(28) Wenig Odermatt, J., P. OpenChrom: a cross-platform open source software for the mass spectrometric analysis of chomatographic data. *BMC Bioinformatics* **2010**, *11*.

(29) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature methods* **2015**, *12*, 523.

(30) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **2006**, *78*, 779–787.

(31) Katajamaa, M.; Miettinen, J.; Orešič, M. MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* **2006**, *22*, 634–636.

(32) Tolstikov, V. V.; Lommen, A.; Nakanishi, K.; Tanaka, N.; Fiehn, O. Monolithic Silica-Based Capillary Reversed-Phase Liquid Chromatography/Electrospray Mass Spectrometry for Plant Metabolomics. *Anal. Chem.* **2003**, *75*, 6737–6740.

(33) Baran, R.; Kochi, H.; Saito, N.; Suematsu, M.; Soga, T.; Nishioka, T.; Robert, M.; Tomita, M. MathDAMP: A package for differential analysis of metabolite profiles. *BMC Bioinformatics* **2006**, *7*.

(34) Idborg-Björkman, H.; Edlund, P. O.; Kvalheim, O. M.; Schuppe-Koistinen, I.; Jacobsson, S. P. Screening of biomarkers in rat urine using LC/electrospray ionization-MS and two-way data analysis. *Anal. Chem.* **2003**, *75*, 4784–4792.

(35) Marshall, A. G.; Rodgers, R. P. Petroleomics: The Next Grand Challenge for Chemical Analysis. *Acc. Chem. Res.* **2004**, *37*, 53–59.

(36) Hur, M.; Oh, H. B.; Kim, S. Optimized automatic noise level calculations for broadband FT-ICR mass spectra of petroleum give more reliable and faster peak picking results. *Bull. Korean Chem. Soc.* **2009**, *30*, 2665–2668.

(37) Rüger, C. P.; Schwemer, T.; Sklorz, M.; O'Connor, P. B.; Barrow, M. P.; Zimmermann, R. Comprehensive chemical comparison of fuel composition and aerosol particles emitted from a ship diesel engine by gas chromatography atmospheric pressure chemical ionisation ultra-high resolution mass spectrometry with improved data processing routines. *Eur. J. Mass Spectrom.* **2017**, *23*, 28–39.

(38) Schwemer, T.; Rüger, C. P.; Sklorz, M.; Zimmermann, R. Gas Chromatography Coupled to Atmospheric Pressure Chemical Ionization FT-ICR Mass Spectrometry for Improvement of Data Reliability. *Anal. Chem.* **2015**, *87*, 11957–11961.

(39) Tautenhahn, R.; Böttcher, C.; Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **2008**,

(40) Åberg, K. M.; Torgrip, R. J.; Kolmert, J.; Schuppe-Koistinen, I.; Lindberg, J. Feature detection and alignment of hyphenated chromatographic-mass spectrometric data. Extraction of pure ion chromatograms using Kalman tracking. *J. Chromatogr. A* **2008**, *1192*, 139–146.

(41) Gavard, R.; Rossell, D.; Spencer, S. E. F.; Barrow, M. P. Themis: Batch Preprocessing for Ultrahigh-Resolution Mass Spectra of Complex Mixtures. *Anal. Chem.* **2017**, *89*, 11383–11390.

(42) Hughey, C. A.; Rodgers, R. P.; Marshall, A. G.; Qian, K.; Robbins, W. K. Identification of acidic NSO compounds in crude oils of different geochemical origins by negative ion electrospray Fourier transform ion cyclotron resonance mass spectrometry. *Org. Geochem.* **2002**, *33*, 743–759.

(43) Stanford, L. A.; Kim, S.; Rodgers, R. P.; Marshall, A. G. Characterization of compositional changes in vacuum gas oil distillation cuts by electrospray ionization Fourier transform- ion cyclotron resonance (FT- ICR) mass spectrometry. *Energ. Fuel.* **2006**, *20*, 1664–1673.

(44) Barrow, M. P.; Headley, J. V.; Peru, K. M.; Derrick, P. J. Data visualization for the characterization of naphthenic acids within petroleum samples. *Energy and Fuels* **2009**, *23*, 2592–2599.

(45) Van Krevelen, D. Graphical-statistical method for the study of structure and reaction processes of coal. *Fuel* **1950**, *29*, 269–284.

(46) Kim, S.; Kramer, R. W.; Hatcher, P. G. Graphical Method for Analysis of Ultrahigh-Resolution Broadband Mass Spectra of Natural Organic Matter, the Van Krevelen Diagram. *Anal. Chem.* **2003**, *75*, 5336–5344.

(47) Green, N. W.; Perdue, E. M.; Aiken, G. R.; Butler, K. D.; Chen, H.; Dittmar, T.; Niggemann, J.; Stubbins, A. An intercomparison of three methods for the large-scale isolation of oceanic dissolved organic matter. *Mar. Chem.* **2014**, *161*, 14–19.

(48) Hawkes, J. A.; Hansen, C. T.; Goldhammer, T.; Bach, W.; Dittmar, T. Molecular alteration of marine dissolved organic matter under experimental hydrothermal conditions. *Geochim. Cosmochim. Acta* **2016**, *175*, 68–85.

(49) Palacio Lozano, D. C.; Ramírez, C. X.; Sarmiento Chaparro, J. A.; Thomas, M. J.; Gavard, R.; Jones, H. E.; Cabanzo Hernández, R.; Mejia-Ospino, E.; Barrow, M. P. Characterization of bio-crude components derived from pyrolysis of soft wood and its esterified product by ultrahigh resolution mass spectrometry and spectroscopic techniques. *Fuel* **2020**, *259*, 116085.

(50) Wickham, H. Tidyverse: Easily install and loadâĂŹtidyverseâĂŹpackages. *R package version* **2017**, *1*.

(51) Chang, W.; Cheng, J.; Allaire, J.; Xie, Y.; McPherson, J. shiny: Web application framework for R [Computer software]. *URL http://CRAN. R-project. org/package= shiny (R package version 1.0. 0)* **2017**,

(52) Zhurov, K. O.; Kozhinov, A. N.; Fornelli, L.; Tsybin, Y. O. Distinguishing analyte from noise components in mass spectra of complex samples: Where to cut the noise? *Anal. Chem.* **2014**, *86*, 3308–3316.

(53) Francl, T. J.; Sherman, M. G.; Hunter, R. L.; Locke, M. J.; Bowers, W. D.; McIver, R. T. *Experimental determination of the effects of space charge on ion cyclotron resonance frequencies*; 1983; Vol. 54; pp 189–199.

(54) Barry, J. A.; Robichaud, G.; Muddiman, D. C. Mass recalibration of FT-ICR mass spectrometry imaging data using the average frequency shift of ambient ions. *J. Am. Soc. Mass Spectrom.* **2014**, *24*, 1137–1145.

(55) Cleveland, W. S.; Grosse, E.; Shyu, W. Local regression models. *Statistical models in S, edited by John M. Chambers and Trevor J. Hastie* **1992**, 309–376.

(56) Wickham, H.; Francois, R.; Müller, K. Tibble: Simple Data Frames. 2018; `https://CRAN.R-project.org/package=tibble`, R package version 1.4.2.

Graphical TOC Entry