Addressing the challenges of petroleomics data

by

Remy Gavard

Thesis

Submitted to the University of Warwick

in partial fulfilment of the requirements

for admission to the degree of

Doctor of Philosophy

Department of Chemistry

November 2019

Contents

List of Figures

Acknowledgments

First and foremost, I would like to thank my three supervisors: Dr Mark P. Barrow, Dr Simon E. F. Spencer and Dr David Rossell, for their support and guidance throughout my PhD study. I couldn't have asked for a better team.

I would like to particularly thank Dr Barrow for being such an excellent mentor over the years, encouraging me to always improve, helping me navigate the mass spectrometry field and giving me the work environment I needed to succeed. From the beginning, Mark has given me the space and liberty for me to take ownership of my project and to manage it entirely, while making sure I never hit any dead-ends or went off track.

My thanks to Dr Spencer for being my mentor in statistics over the last five years and helping me navigate the tricky challenges of applied statistics. Simon is one of the rare people who profoundly believes in creating bridges between statistics and all other scientific fields.

To Dr Rossell I would like to express all my gratitude for taking on the role of the ”Nasty Reviewer” to push me forward and ensure a high-quality work. David's suggestions and comments have been of a tremendous help during this PhD and he continued being a critical help even after moving back to Barcelona.

I'm thankful to the Molecular Analytical Science doctoral training for taking me into the program on such short notice and for funding me. I also want to thank all the wonderful people, past and present, who organise the CDT and make it such a rich and pleasant experience over the years.

I am thankful to Dr Diana Catalina Palacio Lozano for her continuous assistance in operating the instrument and providing me with crucial data.

I would like to thank Mary J. Thomas for providing so much feedback and challenging data.

I would like to thank Hugh E. Jones for working closely with me on the software development over the last two years. I am extremely happy and proud to see you continuing the development of KairosMS.

I would like to thank the entire ICR group, especially Pete, Chris, Yuko, Meng and Cookson, for their support and valuable discussions.

I would like to thank my sisters Angeline and Amelie for their support, encouragements and last-minute proofreading.

I would like to thank my friends Rudy, Nicolas and Anne for helping me stay sane over the years.

Last but certainly not least, I would like to extend my deepest gratitude to my parents, Pierre and Odile, without whom none of this would have been possible. My Dad provided me with the aspiration and desire to pursue a PhD while also paving my way towards mass spectrometry. My Mum has been of the most tremendous support over the years, encouraging me to pursue my studies abroad, working hard to make sure money would never limit my aspirations and always putting her children before herself.

Declarations

Publications first authored part of this thesis

- Remy Gavard, David Rossell, Simon E. F. Spencer, and Mark P. Barrow. Themis: Batch Preprocessing for Ultrahigh-Resolution Mass Spectra of Complex Mixtures. *Anal. Chem.*, 89(21):11383–11390, 2017. ISSN 0003-2700. doi: 10.1021/acs.analchem.7b02345

- Remy Gavard, Diana Catalina Palacio Lozano, Alexander Guzman, David Rossell, Simon E.F. Spencer, and Mark P. Barrow. Rhapso: Automatic Stitching of Mass Segments from Fourier Transform Ion Cyclotron Resonance Mass Spectra. *Anal. Chem.*, 91(23):15130–15137, dec 2019. ISSN 15206882. doi: 10.1021/acs.analchem.9b03846. URL `https://pubs.acs.org/doi/abs/10.1021/acs.analchem.9b03846`

Publications co-authored not part of this thesis

- Diana Catalina Palacio Lozano, Remy Gavard, Juan P. Arenas-Diaz, Mary J. Thomas, David D. Stranz, Enrique Mejía-Ospino, Alexander Guzman, Simon E. F. Spencer, David Rossell, and Mark P. Barrow. Pushing the analytical limits: new insights into complex mixtures using mass spectra segments of constant ultrahigh resolving power. *Chem. Sci.*, 10(29):6966–6978, Jul 2019. ISSN 2041-6520. doi: 10.1039/C9SC02903F

- Diana Catalina Palacio Lozano, Claudia X. Ramírez, José Aristóbulo Sarmiento Chaparro, Mary J. Thomas, Remy Gavard, Hugh E. Jones, Rafael Cabanzo Hernández, Enrique Mejia-Ospino, and Mark P. Barrow. Characterization of bio-crude components derived from pyrolysis of soft wood and its esterified product by ultrahigh resolution mass spectrometry and spectroscopic techniques. *Fuel*, 259:116085, jan 2020. ISSN 00162361. doi: 10.1016/j.fuel.2019.116085

Submitted publications first authored part of this thesis

- Remy Gavard, Hugh E. Jones, Diana Catalina Palacio Lozano, Mary Joanna Thomas, David Rossell, Simon E.F. Spencer, and Mark P. Barrow. KairosMS: A new solution for the processing of hyphenated ultrahigh resolution mass spectrometry data. *Anal. Chem.*, 92(5):3775–3786, jan 2020. ISSN 0003-2700. doi: 10.1021/acs.analchem.9b05113

Publications in preparation first authored part of this thesis

- Remy Gavard, Diana Catalina Palacio Lozano, Hugh E. Jones, Mary J. Thomas, David Rossell, Simon E. F. Spencer, and Mark P. Barrow. Study of the rate of non-reproducible peaks assigned a molecular composition in petroleomics

Publications in preparation co-authored not part of this thesis

- Diana Catalina Palacio Lozano, Remy Gavard, Hugh E. Jones, Mary J. Thomas, Claudia X. Ramirez, Jos Aristbulo Sarmiento Chaparro, Matthias Witt, Enrique Mejia-Ospino, and Mark P. Barrow. Advanced Analysis of Bio-oils By Gas Chromatography Coupled To Fourier Transform Ion Cyclotron Resonance Mass Spectrometry

Oral presentations

- American Society of Mass Spectrometry (ASMS) Environmental Interest group during 65th ASMS Annual Meeting in Indianapolis, USA June 3-7 2017
- Analytical Research Forum held at the Royal Society of Chemistry, July 2017
- 38th British Mass Spectrometry Society (BMSS) Annual Meeting, Manchester, UK September 5-7 2017
- Bright Spark symposium, Bath, UK September 11 2017
- HTC-15 conference, Cardiff, UK 24-26 January 2018
- PG Symposium, University of Warwick, 30th May 2018
- useR! 2018, Brisbane, Australia, 10-13 July 2018
- International Mass Spectrometry Conference 2018, Florence, Italy, 26-31 August 2018
- 67th American Society of Mass Spectrometry (ASMS) Energy, Petroleum and Biofuels Workshop held during the Annual Meeting in Atlanta, USA June 2-6 2019
- 67th American Society of Mass Spectrometry (ASMS) Annual Meeting in Atlanta, USA June 2-6 2019

Poster presentations:

- PG Symposium, University of Warwick, 31th May 2017

- 65th American Society of Mass Spectrometry (ASMS) Annual Meeting in Indianapolis, USA June 3-7 2017
- International Conference, Exhibition and Workshops on Petroleum, Refining and Environmental Monitoring Technologies in Antwerp, Belgium 29-30 November 2017
- 39th BMSS Annual Meeting, Cambridge, UK September 11-13 2018

I hereby declare that material presented in this thesis has not been submitted in whole or in part for any other degree, diploma, or qualification in any other university.

Remy Gavard

November 2019

Abstract

Fourier transform ion cyclotron resonance mass spectrometry (FTICR MS) is currently the state-of-the-art instrument in terms of resolving power and accuracy for mass spectrometry and is able to resolve an unprecedented number of components in complex chemical mixtures, such as petroleum. The data analysis tools necessary struggle to keep pace with advancing instrument capabilities and the ever-increasing quantities of data generated. The existing workflows rely on combining different tools, not necessarily compatible between them and often generate a significant amount of manual repetitive tasks. A first issue is that the current standard practice does not utilise replicates to improve the reliability of an analysis. A second issue is that spectral stitching methods to combine data from multiple experiments performed for a single sample are not automated, and hence generate substantial manual work that precludes the routine applications of these experiments. Hyphenated ultra-high resolution, can provide structural information but the data analysis tools are lacking leading to loss retention time precision and labour-intensive workflows. A final issue explored in this thesis is that molecular assignments are performed using commercial software or in-house algorithms but currently no evaluation of the false positive assignments has been performed. During this PhD, algorithms were developed to address those needs and implemented using the R language. The tools needed to be accessible to a wide audience, not necessarily comfortable using scripted languages so interactive interfaces were created using the Shiny framework. Overall, the work presented in the thesis brings improved reliability when analysing complex mixture using Fourier transform mass spectrometry thanks to combining replicates or stitching multiple experiments, and assessing reproducibility. Further, it helps accelerate analyse hyphenated ultra-high-resolution mass spectrometry decreasing the time necessary from days to hours while bringing a deeper and more accurate insight into the data also capable to analyse and compare molecular assignments for petroleum related samples.

Sponsorships and Grants

Remy Gavard thanks the EPSRC for a PhD studentship through the EPSRC Centre for Doctoral Training in Molecular Analytical Science, grant number EP/L015307/1. Remy Gavard thanks the British Mass Spectrometry Society for the John Beynon travel grants which supported the attendance to three conferences. Remy Gavard thanks the Analytical Division of Royal Society of Chemistry for the travel grants to attend two conferences overseas. Remy Gavard thanks the Analytical Chemistry Trust Fund for the travel grant to attend a conference overseas. Remy Gavard thanks Warwick Ventures for the HEIF Impact Fund which supported the development of KairosMS.

Abbreviations

2D 2 Dimensions

3D 3 Dimensions

ACN Acetonitrile

APCI Atmospheric Pressure Chemical Ionization

APPI Atmospheric Pressure Photoionization

C Coulomb

CAD Collisionally Activated Dissociation

CI Chemical Ionization

CID Collisionally-Induced Dissociation

CLT Central Limit Theorem

Da Dalton

DA DataAnalysis (software)

DBE Double Bond Equivalents

DC Direct Current

ECD Electron Capture Dissociation

EI Electron Ionisation

ESI Electrospray Ionisation

FT-ICR MS Fourier Transform-Ion Cyclotron Mass Spectrometry

FWHM Full Width at Half Maximum

GC Gas Chromatography

GC-MS Gas Chromatography-Mass Spectrometry

Hz Hertz

ICR cell Ion Cyclotron Resonance Cell

IUPAC International Union of Pure and Applied Chemistry

KMD Kendrick Mass Defect

LC Liquid Chromatography

LC-MS Liquid Chromatography-Mass Spectrometry

MS Mass Spectrometry

m/z Mass-to-Charge Ratio

nESI Nano-Electrospray

NIST National Institute of Standards and Technology

OSPW Oil Sand Process-Affected Water

ppb Parts Per Billon

ppm Parts Per Million

RF Radio Frequency

RMS Root Mean Square

RP Resolving Power

S/N Signal-to-Noise Ratio

sd Standard Deviation

SRFA Suwannee River Fulvic Acid

T Tesla

TD Time Duration of Transient

TOF Time-of-Flight

V Volt

Symbols

π	The Sign Pi
ω	The Angular Frequency
δ_m	Full Width at Half the Maximum

Chapter 1

Introduction

1.1 Mass spectrometry

1.1.1 Theory & Terminology

Mass spectrometry (MS) is the measuring of the mass-to-charge ratio (m/z) of charged molecules. In order to measure the m/z, we need to ionise the sample. Previously charged molecules are not frequently found, hence the development of several ionisation methods that will be described in section 1.2. Following the ionisation process, the resulting ions (positively or negatively charged) are analysed and detected, and the signal's intensity for a specific m/z will vary according to a number of parameters, including their abundance within the analysed sample. Several methods to detect those charged molecules have been developed over time, each with their own specificities and will be covered in section 1.3. Isotopes can also be detected if their presence is high enough to pass the limit of detection and can be resolved by the instrument. The combination of all the m/z measured by the detector with its corresponding intensity is called a mass spectrum.

A mass spectrometer needs an ionisation source and a mass analyser, but in order to gain structural information about the molecules detected, it is possible to separate a mixture of molecules according to their physical properties before being ionised and detected by the mass spectrometer. The most common techniques used for separation coupled with MS are liquid chromatography (LC) and gas chromatography (GC) and will be described in section 1.4. MS is extensively used to analyse pharmaceuticals, biomolecules, environmental samples, crude oil and nowadays can detect analytes at very low concentration and with a mass error within the part-per-billion (ppb). However, while MS progressed and was able to chronically analyse very complex mixtures, the data analysis tools have struggled to keep pace.

1.1.2 History

The history of mass spectrometry starts in 1897 when J.J. Thomson discovered the electron and its mass-to-charge ratio (m/z). This discovery was rewarded with a Nobel prize in 1906. A first mass spectrometer was built in 1912 by J.J. Thomson. That year he succeeded in generating the first mass spectrum for O_2, N_2, CO, CO_2 and $COCl_2$ molecules [8]. In 1919, F.W. Aston built the first mass spectrometer with velocity focusing [9]. J. Beynon showed the first use of high resolution and exact mass determination in 1956 [10], and that same year, F.W. McLafferty and R.S. Gohlke presented the first mass spectrometer coupled with gas chromatography [11, 12]. A few years later, in 1967, collision-induced dissociation was introduced by F.W. McLafferty and K.R. Jennings [13, 14]. In 1974, several new advances were presented. Atmospheric-pressure chemical ionisation (APCI) was developed by E.C. Horning, D.I. Carroll, I. Dzidic, K.D. Haegele, M.D. Horning and R.N. Stillwell [15], the first high performance liquid chromatography coupled with a mass spectrometer is presented by P.J. Arpino, M.A. Baldwin and F.W. McLafferty [16] and finally, the Fourier transform ion cyclotron resonance mass spectrometer was presented by M.B. Comisarow and A.G. Marshall [17]. In 1978, the triple quadrupole mass spectrometer was developed by R.A. Yost and C.G. Enke [18]. In 1993, R.K. Julian and R.G. Cooks presented the stored-waveform inverse Fourier Transform (SWIFT) [19]. The nano electrospray ionisation source was developed in 1994 by M. Wilm and M.Mann [20]. Finally, in 1999, the high performance ion trap with electrostatic quadro-logarithmic field, commonly called Orbitrap, was invented by A.A. Makarov [21] based on the Kingdon trap [22].

1.2 Ionisation methods

Since mass spectrometers can only detect charged particles, to avoid limiting MS to molecules that are naturally charged, the first step of any analysis is the ionisation. A large variety of methods have been developed over the years to make a neutral molecule a charged one, and each method has its own specificities and affinities for different types of molecules. Pioneering methods such as electron ionisation (EI) [23] and chemical ionisation (CI) [24–26] require a volatile sample and often leads to bond breakage, particularly in presence of big molecules. Newer methods such as atmospheric pressure ionisation (API) [27–30] including electrospray ionisation (ESI) [31] are called soft ionisation as they were developed to overcome the fragmentation issue of the EI and CI.

1.2.1 Electron ionisation (EI)

Electron ionisation (EI) is one of the classical ionisation method which uses gas molecules with energetic electrons (usually 70 eV) to create ions [23]. EI was first designed by Dempster and later improved by Bleakney [32] and Nier [33]. The reaction describing the electron ionisation process is $\mathrm{M} + \mathrm{e}^{-} \longrightarrow \mathrm{M}^{+}\cdot + 2\,\mathrm{e}^{-}$. The generated beam of electrons will then expel an electron from the analyte and cause it to go from neutral to a radical cation ($\mathrm{M}^{+}\cdot$). This generates an unstable radical ion, which tends to fragment to form more stable radical elements along with some neutral species, and the fragments generated by this technique can be used to determine the structure of the molecule. Electron ionisation is a simple and stable technique but the original $\mathrm{M}^{+}\cdot$ cannot always be observed, complicating the characterisation task.

The methods can also generate a high number of fragments, leading to numerous peaks being observed and complicating the analysis.

1.2.2 Chemical ionisation (CI)

Chemical ionisation [26] was introduced by Talrose [25] and further developed by Munson and Field [24]. This ionisation method consists of the collision between analyte and a gas, such as methane or ammonia. While an electron transfer occurs in electron ionisation, this method relies on proton transfer. This method generates $[\mathrm{M} + \mathrm{H}]^+$ molecules which are more stables and result in less fragmentation than EI.

The EI reaction pathway when methane is used is $\mathrm{CH_4} + \mathrm{e^-} \longrightarrow \mathrm{CH_4}^{+\cdot} + 2\,\mathrm{e^-}$. The ion created will then fragment following two main pathways $\mathrm{CH_4}^{+\cdot} \longrightarrow \mathrm{CH_3}^+ + \mathrm{H}\cdot$ or $\mathrm{CH_4}^{+\cdot} \longrightarrow \mathrm{CH_2}^{+\cdot} + \mathrm{H_2}$ but will mostly react with other methane molecules to yield $\mathrm{CH_4}^{+\cdot} + \mathrm{CH_4} \longrightarrow \mathrm{CH_5}^+ + \mathrm{CH_3}\cdot$.

CI is a useful technique to obtain information about the molecules but it requires volatile and stable samples [24, 34] and works best with polar and semi-polar species [24]. The inconvenience is that molecules can only be single charged and that this method is not suitable for large biomolecules.

1.2.3 Electrospray ionisation (ESI)

Electrospray ionisation was developed by Fenn *et al.* in 1989 [31] and is a soft ionisation method, particularly suitable for large molecules.

ESI uses the potential difference between a capillary and a counter electrode charged between 3000 to 6000 V. The difference leads to a charge accumulation at the surface of the liquid that will be analysed at the end of the capillary. When the liquid exits the needle, it forms a Taylor cone before breaking into charged droplets creating what is called the ESI plume. The flow of the ESI plume is oriented using a gas, and a heated capillary is responsible for the final evaporation of the solvent remaining within the droplets [35]. To minimise the radial dispersion of the spray,

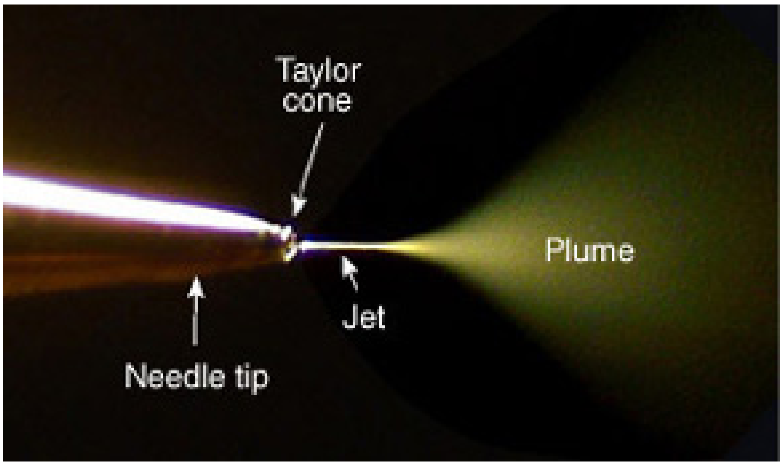


Figure 1.1: Photograph of Taylor cone and its ESI plume. Reproduced from www.newobjective.com (accessed 02/07/2019)

a so-called sheath gas can be applied coaxially. [36] ESI can operate in positive and negative mode in order to generate either positively and negatively charged ions. The ions formed are usually resulting from the addition of a hydrogen cation which are denoted $[\mathrm{M+H}]^+$ but other cations like a sodium ion can be formed ($[\mathrm{M+Na}]^+$). The removal of a proton is also possible in which case $[\mathrm{M{-}H}]^-$ ions are formed. Multiply-charged ions can be observed and will be denoted $[\mathrm{M+nH}]\mathrm{n}^+$. A mixture of solvent such as 50/50 methanol/water or toluene/methanol is often employed, with acid added such as 1% formic acid to help protonation or a base such as 0.1% ammonium hydroxide for deprotonation. Electrospray ionisation was first applied in 1968 to polymers [37] then rapidly expanded to proteins, biopolymers and complex mixtures [38–41].

1.2.4 Nano electrospray Ionisation (Nano-ESI)

Nano electrospray ionisation is a variation of the ESI described previously [42]. The objective behind nano-ESI is to use less sample while producing a sufficient signal: it uses a very thin capillary with a diameter between 10 to 100 μm and a flow rate around $\sim 1nL/min$. The fluid is not expelled thanks to the mechanical movement of a syringe like in ESI, as it uses capillary traction to move the fluid within the

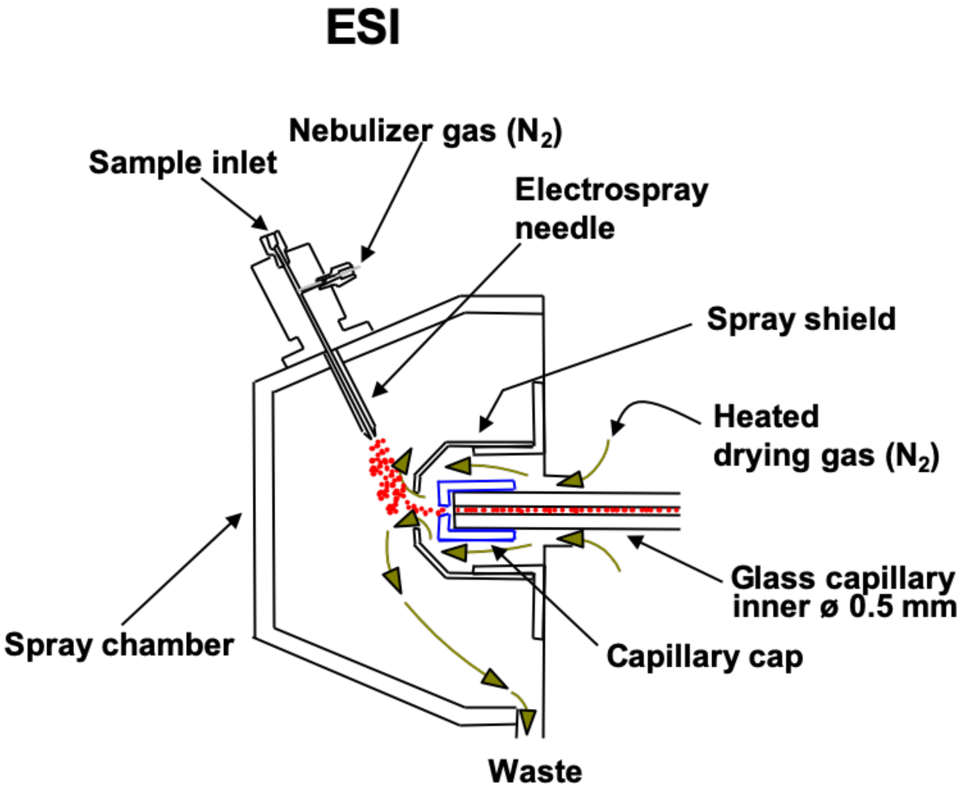


Figure 1.2: Schematic for electrospray ionisation source. Adapted from solariX training manual, Bruker Daltonik GmbH, Bremen, Germany.

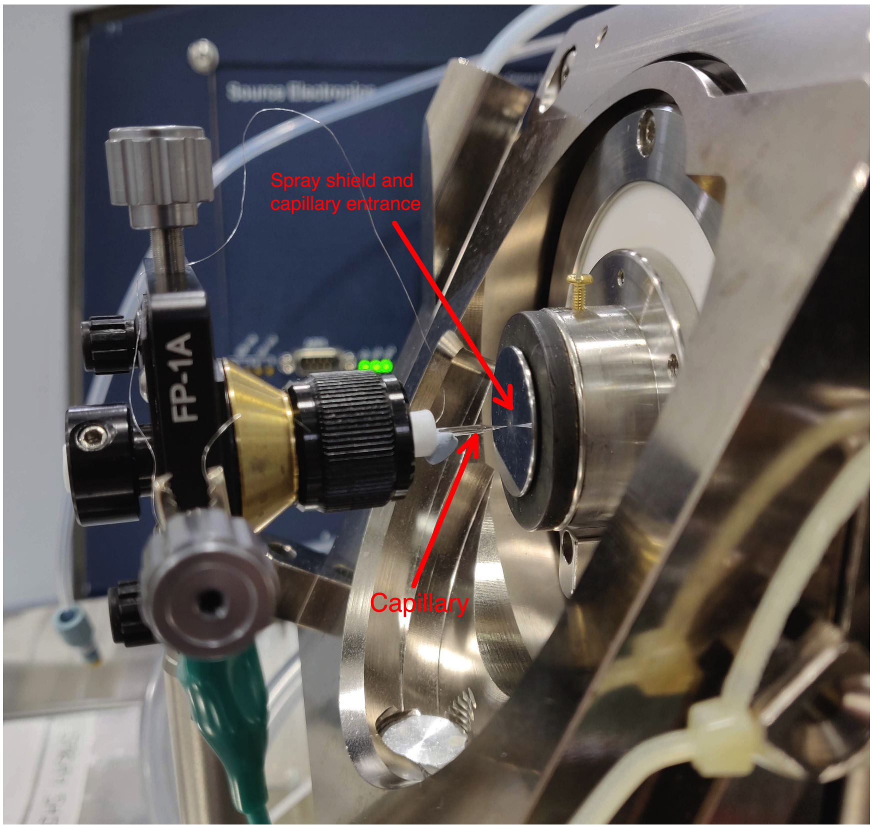


Figure 1.3: Picture of a nano electrospray ionisation source coupled to an FTICR MS.

thin capillary. As a consequence of the very thin capillary, the plume generated is invisible to the eye because of the 100 *nm* size of the droplets.

1.2.5 Atmospheric pressure chemical ionisation (APCI)

Atmospheric pressure chemical ionisation (APCI) [27, 28] shares similarities with CI, since it uses gas phase ion-molecule reaction at atmospheric pressure. This ionisation technique is particularly suited to polar and relatively non-polar molecules, usually with a molecular size up to 1500 Da, and yields singly charged ions. APCI is regularly used to analyse petroleum related samples [43] but also biological samples [44] and food samples [45, 46].

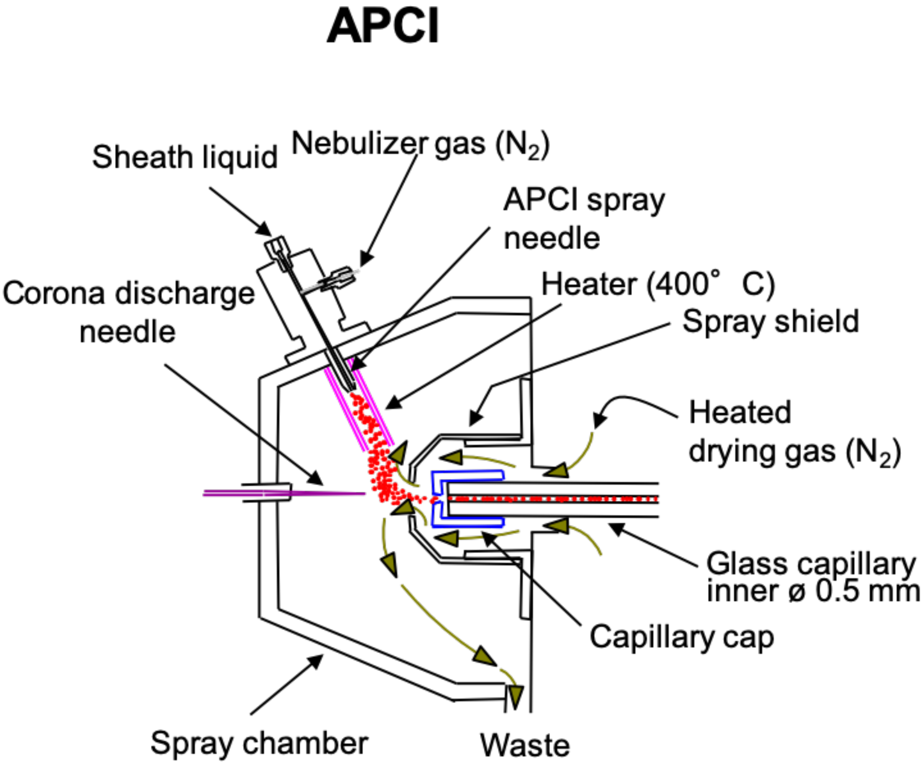


Figure 1.4: Schematic for atmospheric pressure chemical ionisation source (APCI). Adapted from solariX training manual, Bruker Daltonik GmbH, Bremen, Germany.

1.2.6 Atmospheric pressure photoionisation (APPI)

Atmospheric pressure photoionisation (APPI) [29, 30] is based on a modified APCI as it uses a discharge high-energy UV lamp (often krypton) instead of protons and electrons. The plume is exposed to the high-energy UV lamp and causes the analyte and solvent to turn into an electronically excited state so an electron transfer can occur to create ions.

The mass spectra obtained with positive mode APPI mostly contains two types of ions: the radical cation $\mathrm{M}^{+}\cdot$ and the protonated molecule $[\mathrm{M{+}H}]^{+}$. The main reaction leading to the formation of cations is $\mathrm{M} + \mathrm{hv} \longrightarrow \mathrm{M}^{+}\cdot + \mathrm{e}^{-}$ but the major presence of protonated molecules suggest the abstraction by a molecular ion of an hydrogen atom from a solvent molecule: $\mathrm{M}^{+}\cdot + \mathrm{S} \longrightarrow [\mathrm{M{+}H}]^{+} + (\mathrm{S{-}H})\cdot$. In positive ionisation mode, a dopant can be used and the first step of the ionisation process is $\mathrm{D} + \mathrm{hv} \longrightarrow \mathrm{D}^{+}\cdot + \mathrm{e}^{-}$ which creates a radical ion from the dopant. The radical cation created will then interact with the solvent $\mathrm{D}^{+}\cdot + \mathrm{S} \longrightarrow [\mathrm{S} + \mathrm{H}]^{+} + (\mathrm{D{-}H})\cdot$ which will then boost the formation of protonated molecules $\mathrm{M} + [\mathrm{S} + \mathrm{H}]^{+} \longrightarrow [\mathrm{M} + \mathrm{H}]^{+} + \mathrm{S}$. The radical cation of the dopant can also directly interact with the analyte if the ionisation energy of the analyte is lower than the one of the dopant leading to the reaction $\mathrm{D}^{+}\cdot + \mathrm{M} \longrightarrow \mathrm{M}^{+}\cdot + \mathrm{D}$.

APPI is typically used for the ionisation of non-polar compounds, but it can also work for polar species. It can be employed to ionise a wide range of compounds such as complex mixtures including petroleum [47], but since APPI allows observation of both protonated molecules and radical cation, it often generates a more complex spectrum.

1.3 Mass Analysers

After the ionisation step described earlier, the ions need to be separated according to their mass-to-charge ratio.

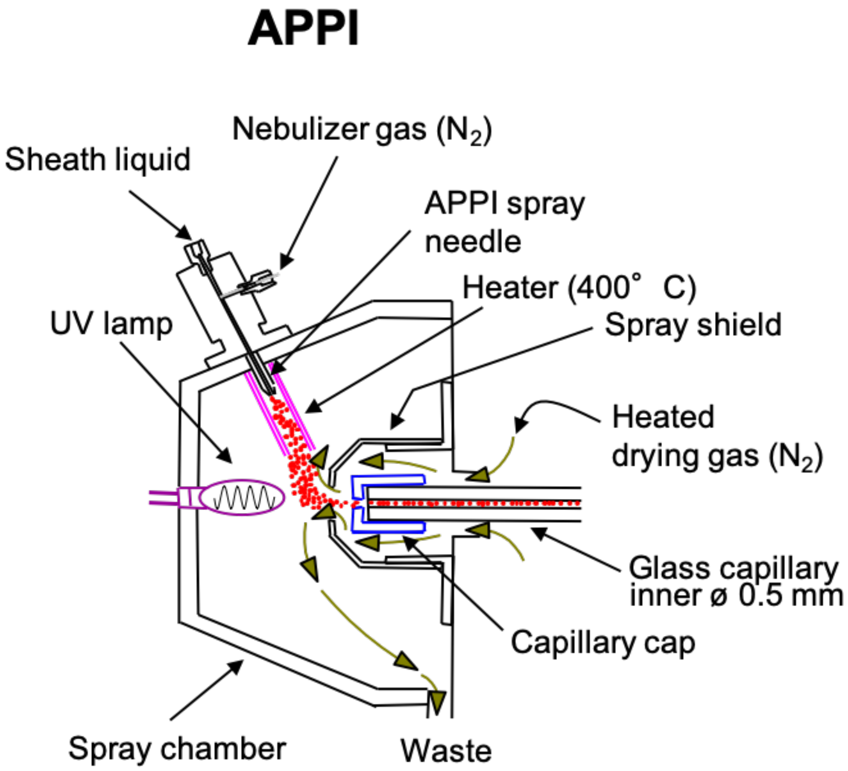


Figure 1.5: Schematic for atmospheric pressure photoionisation source (APPI). Adapted from solariX training manual, Bruker Daltonik GmbH, Bremen, Germany.

Just like the ionisation techniques, several different types of mass analysers have been developed over time, each with their own specificities. It is even possible to combine different types of mass analysers in order to achieve higher performances or improve capabilities (ex Q-TOF).

1.3.1 Time-of-flight

The time-of-flight (TOF) technique relies on the calculation of the m/z based on the flight time through a field-free flight tube [48, 49], kept under vacuum to avoid any collision with a background gas during the flight. Ions from the ion source are directed through the flight tube using electric fields. Separation will occur based on mass, as molecules of different masses will have different flight times, directly linked to their m/z values as light ions will travel faster through the flight tube than heavy ones. Equation 1.1 shows that the flight time t is obtained using the length of the flight L tube multiplied by the velocity v.

$$t = \frac{L}{v} \tag{1.1}$$

The flight time is usually under 1 ms which allows for a fast scan rate, making TOF analysers particularly suitable for LC and GC coupling. TOF analysers do not have limits to the size of the molecules they can analyse [50]. Recently, a maximum resolving power (RP) of about 90 000 was achieved using a physical tube length of 6 m and an effective length of 14 m [51] while mass accuracies below 2 ppm are being reported [52].

1.3.2 Quadrupole

The quadrupole analyser takes its origins in the 1950s [53, 54] and was later adapted to be suitable for use with ESI. It has the advantages of being robust and cheap, functioning well under high vacuums and of being fairly low cost [55]. It is also able

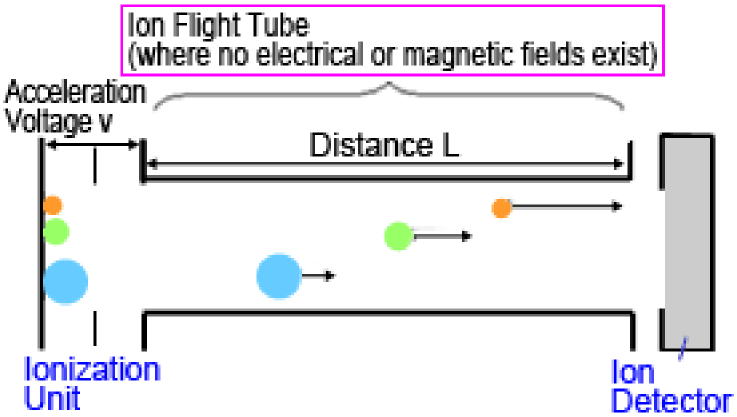


Figure 1.6: Schematic of a Time-of-flight (TOF) mass analyser. Adapted from www.shimadzu.com (02/07/2019)

to cope with a wide m/z range (4 kDa) [56–58]. They are used for ion selection as only those with a stable trajectory pass through and can later be detected, which allows analysts to select specific m/z, but the accuracy is variable.

The quadrupole consists of four charged, alternating-polarity rods and the opposite pairs are connected. The ions' trajectory is controlled using a combination of radio frequency (RF), voltage (V) and direct current (DC) applied to each pair of rods. The sign of the potential of the rods change periodically, in consequence the ions are alternatively attracted and repulsed, creating an ion oscillation. The movement of ions inside a multipole is defined by the Mathieu equation which includes a and q which are dimensionless trapping parameters. The resulting stability diagram is presented in Figure 1.7. If the ion oscillation is stable, the ion can pass through and be detected later.

Quadrupole refers to the use of four rods but more rods can be added to create hexapoles and octupoles. While quadrupoles are most often employed for transport and filtration, hexapoles, octupoles are great tools for the transport and transmission of ions but are rarely used for filtration. In FTICR MS, multipole helped to dramatically improve the transport of the ions to the ICR cell.

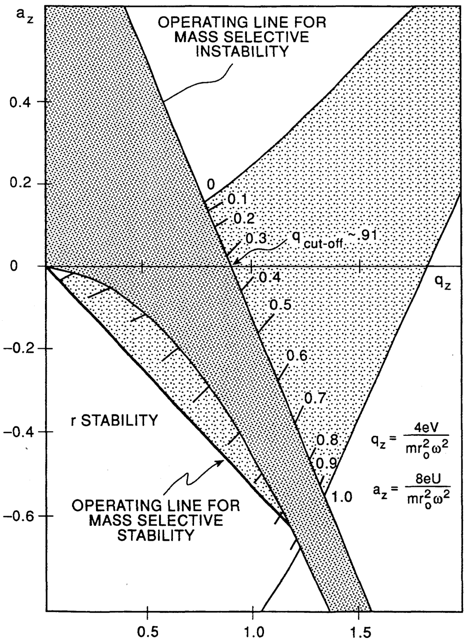


Figure 1.7: Stability diagram for a quadrupole ion trap depending of the a and q parameters which are controlled by the frequency and voltage applied to each rods. From Patent 5399857 [59].

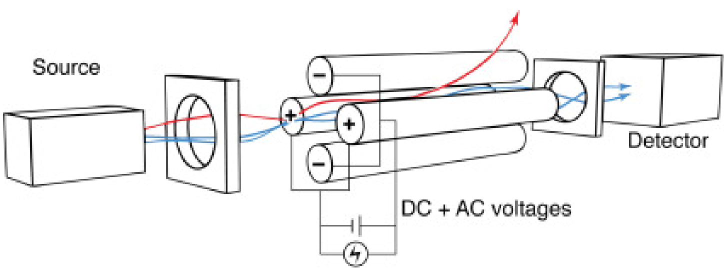


Figure 1.8: Schematic of a quadrupole. Selected ions are depicted in blue while the non-selected are in red. Reproduced from [60]

1.3.3 Ion traps

Ion traps are devices which use an oscillating electric field to store ions. Ions can be trapped in either 2 or 3 dimensions logically leading to two types of ion traps: 2D and 3D. Historically, 3D ion traps were first invented while 2D traps are more recent, we will focus on the more widespread 3D ion traps.

3D ion traps were first created by Paul and Steinwedel 1960 [61] and later improved by Stafford Jr *et al.* 1984 [62] from Finnigan Company into an exploitable mass spectrometer. Paul received the Physics Nobel Prize in 1989 for his invention. The principles behind 3D-traps are similar to the ones behind the quadrupole mass analyser but instead of simply passing through, the ions are trapped.

A 3D-trap is composed of a so-called “ring” and end-cap electrodes. The “ring” electrodes get an oscillating RF voltage while the end-cap electrodes get a static DC voltage. By changing the electric field of the end-cap electrodes, it is possible to eject the ions from the trap and to send them for detection [63].

The performances are closer to the ones of the quadrupole, with a low accuracy and resolving power but they are cheap and work well inside a vacuum. What makes the 3D traps stand out is that they can be more sensitive than the quadrupole as they can accumulate ions.

1.3.4 Orbitrap

The high performance ion trap with electrostatic quadro-logarithmic field, also called Orbitrap, uses a Fourier transform (FT) [21] and has two patents associated [64, 65]. The Orbitrap was based on the Kingdon trap [66] which was later modified [67, 68] before being adapted for mass spectrometry. The first commercial instrument utilising this new mass analysers was made available by Thermo Electron Corporation in 2005. The instrument is composed of a central electrode shaped like a spindle and surrounded by a barrel-like shaped electrode. The electrode is cut into two equal parts with a small space in between them. Ions are injected through the external

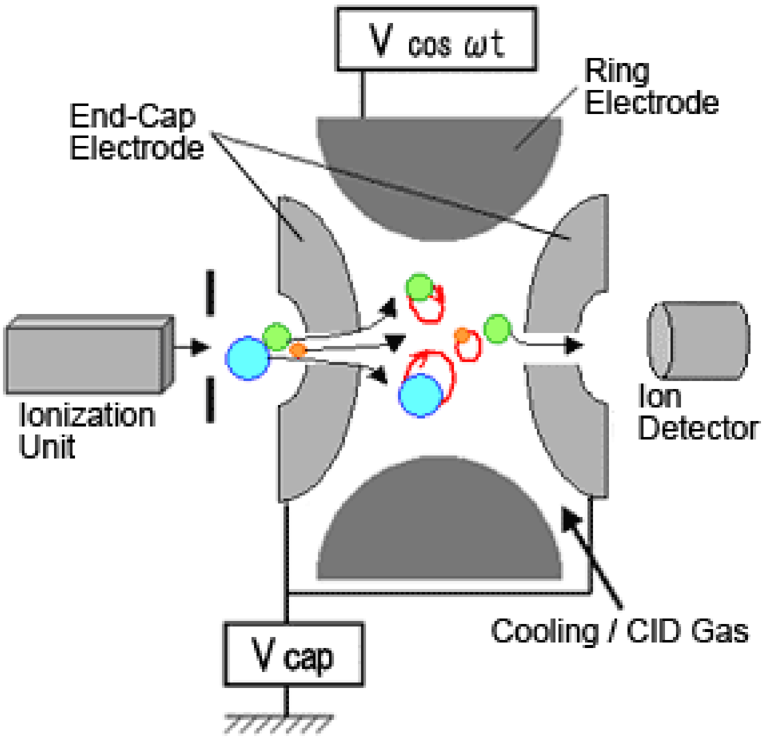


Figure 1.9: Diagram of Ion-Trap mass spectrometer. Adapted from www.shimadzu.com (02/07/2019)

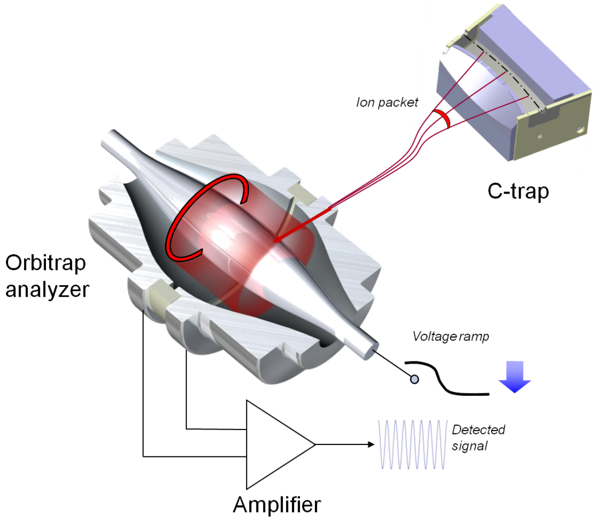


Figure 1.10: Cross-section of the C-trap and Orbitrap analyzer. Artwork courtesy of Thermo Fisher Scientific (commons.wikimedia.org/wiki/File:OrbitrapMA%26Injector.png)

electrode using a little hole with the energy of a few kilovolts. They immediately start to oscillate in the cell around the internal electrode. The oscillation of the ions is measured and transformed by Fourier transform into the frequency domain and then onto an m/z mass spectrum. The Orbitrap present the advantages of providing ultra-high resolution and of being low maintenance, as it doesn't use a superconducting magnet and so does not need liquid helium fills for example.

1.3.5 FTICR

Fourier transform ion cyclotron resonance mass spectrometry (FTICR MS) was first developed in 1974 by Comisarow and Marshall [17, 69, 70].

FTICR MS is state-of-the-art technology, with ultra-high resolving power going from 100 000 to over 20 000 000 [71]. The ions are kept in orbit using the Lorentz force, which is a centripetal force, and then the ions are separated by their frequency [72, 73]. Each ion orbits inside the magnetic field at a unique frequency inversely proportional to their m/z and the force of the magnetic field. The equation

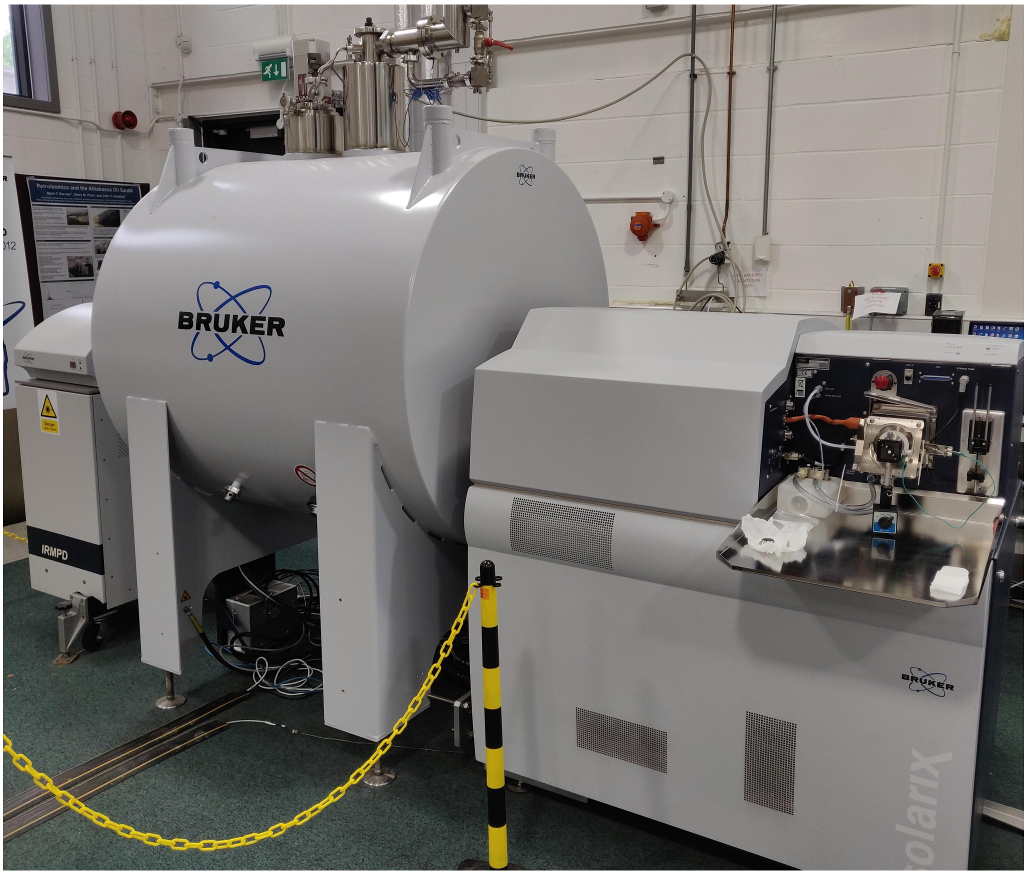


Figure 1.11: Picture of a 12 T FTICR MS.

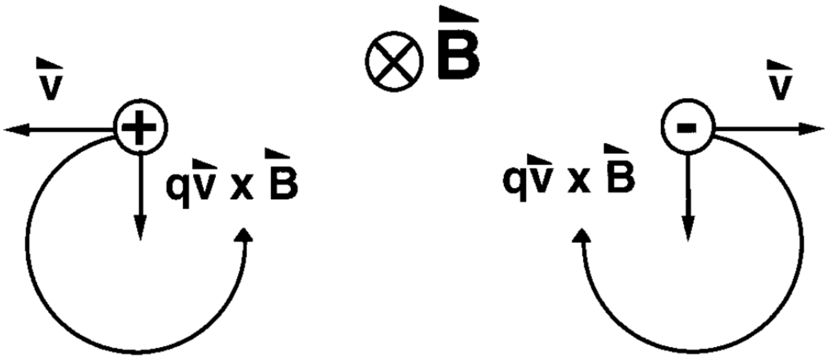


Figure 1.12: Representation of the forces applied on both positive and negatively charged ions within a magnetic field. Reproduced from Marshall *et al.* [73]

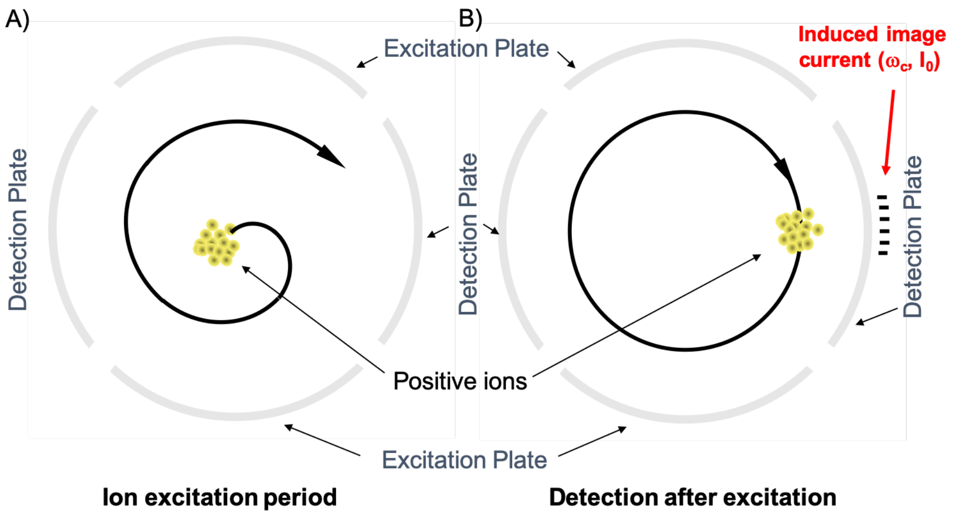


Figure 1.13: Illustration of the excitation (A) and detection (B) of a ion within an ICR cell. Courtesy of Bruker Daltonik GmbH, Bremen, Germany.

1.2 defines the cyclotron frequency f in Hz where q is expressed in coulomb (C), B is the strength of the magnetic field expressed in tesla (T) and m is the mass in kg.

$$f = \frac{qB}{2\pi m} \tag{1.2}$$

Equation 1.2 is sometimes expressed in ω_c, giving equation 1.3 where ω_c is expressed in $rad.s^{-1}$.

$$\omega_c = \frac{qB}{m} \tag{1.3}$$

In order to measure the frequency of the orbiting ions, they need to be trapped in an ICR cell. Currently the main design is composed of a total of six electrodes divided into three groups of two. There are two trapping electrodes, two detection electrodes and two excitation electrodes. Each pair of electrodes is disposed opposite to each other. Because the detected signal is very low, an amplifier is necessary to amplify the signal so it can be transmitted and analysed. Thanks to such a configuration, the ions are trapped, and the frequency can be measured.

The ions in the ICR cell are also subject to a magnetron motion. As described

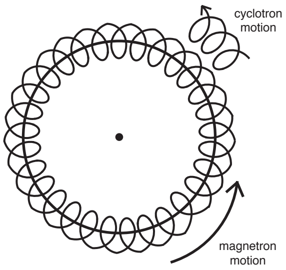


Figure 1.14: Representation of the ions moving along a cyclotron motion itself turning around a central point represented by a dot, this movement is called the magnetron motion.

in Figure 1.13, the ions move in an orbit called the cyclotron motion but this orbit itself precesses the centre in a magnetron motion as depicted in Figure 1.14. The magnetron motion is more important than the cyclotron motion and causes losses of resolution at high m/z. Thanks to the application of a quadrupolar RF field, the movement can be eliminated.

In reality, the frequency at which the ions are moving is affected by the electric field from the two trapping plates, along with the electric field of the other ions present within the cell. The ions within the cell will see their electric field interact between each other's due to the Coulombic repulsion force causing what is called a "space charge" [74] which will hamper the performance of the instrument. The measured frequency of the ions will reduce as Coulombic repulsion increases, causing the measured m/z to be higher than the true value.

Hence, the frequency measured $\omega_{measured}$ needs to be corrected using a calibration function in order to obtain an accurate m/z. The recorded time domain signal is converted using the Fourier transform into a frequency spectrum. Finally, a mass spectrum is obtained by applying one of the calibration functions. The resolution of FTICR MS can be increased with bigger magnets as the mass accuracy is proportional to the square of the magnetic field. Hence, with the availability of bigger magnets the performances of FTICR MS will increase. Other parameters such as longer acquisition time, better data processing, and segmentation can help to push the performance of existing instruments. Indeed, a 12 T FTICR MS was recently used to obtain a constant resolving power of 3 million FWHM across a broad m/z range (m/z 260-1500) and 244,779 compositional assignments without using chromatography or fragmentation [3].

1.4 Chromatography

Chromatography is defined by the IUPAC as “a physical method of separation in which the components to be separated are distributed between two phases, one of which is stationary (stationary phase) while the other (the mobile phase) moves in a definite direction”. It originated in the early 20th century with work by M. S. Tswett [75]. Chromatography and mass spectrometry have been associated since very early on and used extensively in tandem. The objective is to separate mixtures of chemical components into either pure isolated species or less complex mixtures, before being sent to the mass spectrometer for characterisation. Chromatography is based on the interactions between a mobile phase and a stationary phase as it moves through a column. Depending on the interaction between the two phases, the chemical species are separated based on their structures and chemical properties. The product eluting can be either analysed online (directly as it elutes) or off-line (fractions collected and analysed).

As researchers wanted to obtain more information, and thanks to the versatility of the FTMS instruments, the coupling with chromatography techniques enabled analysts to obtain structural insights into complex mixtures. Below, the focus will be on liquid chromatography and gas chromatography as those were the two techniques employed during this PhD, but other techniques exist and can be coupled with mass spectrometry.

1.4.1 Gas chromatography (GC)

Gas chromatography (GC) is the most straightforward technique for mass spectrometry coupling, since the ions need to be in a gas phase to be analysed in MS [76]. The columns employed are usually extremely long and the interior, coated with functionalised silica, will be used as the stationary phase. An inert gas will be used as the mobile phase and molecules elution through the column will be controlled by the column temperature. The column is placed inside an oven in order to control its temperature. A specific program to slowly elute will be set up, usually making use of one or several "heat ramps" to ensure an ideal separation. Since it is easier to control heating rather than cooling, the programs will always start at low temperatures and then increase. The temperatures employed can go up to 300 degrees Celsius for molecules with a strong affinity with the column. This technique generally provides fast elution in comparison to other techniques, which is why, when it is coupled with MS, a high scan rate is necessary to keep up with the speed of elution. GC-MS is a very widespread technique when coupled with low to average resolution MS, and has a wide range of applications. It is also possible to perform two GC back-to-back to obtain a GC/GC-MS acquisition. As the eluant is already in gas phase, it is compatible with many ionisation sources. GC-MS was successfully used with complex mixtures [77], petroleum [43, 78] environmental samples [79], pharmaceutical [80], pesticides [81], and forensic samples [82].

1.4.2 Liquid chromatography (LC)

Liquid chromatography was initiated by the invention of partition chromatography by A. J. P. Martin and R. L. M. Synge in 1941 [83] for which they obtained the Nobel Prize in 1953. Liquid chromatography (LC), unlike GC, is based on liquid rather than gas [84], following a similar principle with both a mobile phase and a stationary phase. The columns used for LC are often shorter in comparison to the GC's with a length typically between 10 - 30 *cm*, but this is not always the

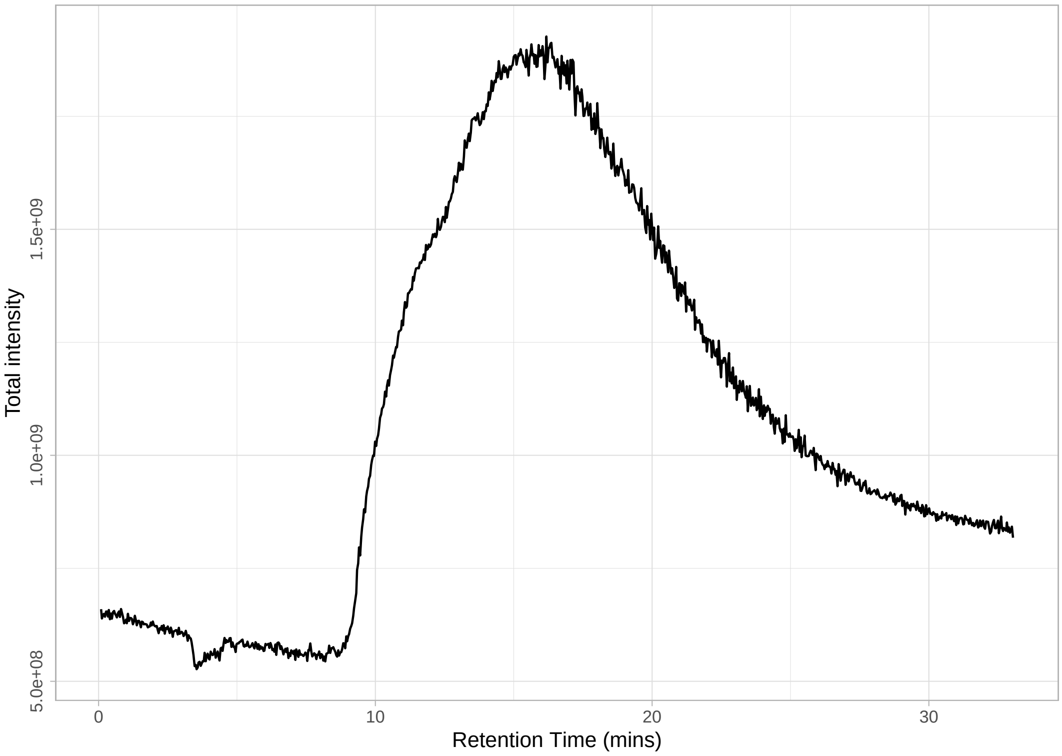


Figure 1.15: Total ion chromatogram of petroleum related sample analysed by GC-FTICR MS.

case. The elution period tends to be long, with experiments lasting between 30 to 120 minutes. Due to the advances of the field, columns packed with smaller particles able to withstand ultra high pressures have led to reductions in elution times. This time the elution gradient does not rely on temperature but on solvent mixtures which changes their ratio over time. A wide variety of stationary phases are available to the user depending on the sample being analysed. One of the most popular solvent mixtures is water with acetonitrile (ACN); this is due to the low UV absorbance of the ACN along with a lower pressure on the column. The elution usually starts with a high percentage of water and will decrease over time, until reaching a majority percentage of acetonitrile. The user will decide on an appropriate gradient to change the percentages of each solvent, with the possibility keeping the ratio stable for periods of time. The challenging part of an online LC-MS experiment is that the solvent needs to be evaporated as fast as it elutes through the column. For this reason, ionisation techniques such as ESI, APCI, APPI are best suited as they can handle the flow rate. LC-MS was successfully employed to analyse a wide range of samples such as small molecules [85], metal complexes [86], polymers [87], biomolecules [88] and dissolved organic matter [89].

1.5 Data analysis

1.5.1 Signal processing

The free induction decay (FID) is the signal obtained directly from the instrument when ions' frequencies are being measured in the ICR cell by electrodes. A Fourier transform is applied to the FID signal and an operation called zero filling is performed. Since the resolving power of the FTMS signal is proportional to the length of the time duration of transient (TD), any increase in TD will be beneficial to the quality of the signal. Unfortunately, increasing TD is not always experimentally possible. Instead, TD can be mathematically doubled by adding as many 0 as necessary at the end of the FID signal to double the size [90]. This technique is called zero-filling and allows to increase the number of data points measured without changing the peak shape and resolution but will yield better centroids for the peak picking. Finally a calibration equation is used to convert the frequency domain spectrum to the m/z domain.

1.5.2 Resolving power

The resolving power is, with mass accuracy, one of the two metrics widely used to evaluate the performance of an MS instrument. It determines the capability of an instrument to separate two peaks very close to each other on the m/z scale. This is calculated using equation 1.4 where m is the m/z and δm is the full width at half maximum (FWHM). The highest the resolving power number, the better.

$$Resolving\ power = \frac{m}{\delta m} \tag{1.4}$$

For reference, FTICR MS can routinely achieve resolving powers over 1 million [91–94] while the Orbitrap can achieve resolving powers between 100 000 and 600 000 [95]. The highest resolving power ever achieved was about 47 million [71]. For FTICR, the resolving power is inversely proportional to m/z and, for example the resolving power at m/z 400 is twice higher than at m/z 800 [73, 96]. In FTICR MS, the resolving power increases with the strength of the magnetic field. This means that a higher resolving power is achievable per second which is a useful feature for time sensitive acquisitions such as with chromatography. Another method to increase the resolving power is to have longer acquisition times.

1.5.3 Mass accuracy

The mass accuracy is defined by the deviation of a measured m/z value from the theoretically calculated m/z value and is expressed in parts-per-million (ppm) and, recently started to be expressed in parts-per-billion (ppb). Equation 1.5 defines how the mass accuracy is calculated.

$$Mass\ Accuracy\ (ppm) = \frac{(Measured\ \frac{m}{z} - Exact\ \frac{m}{z})}{Exact\ \frac{m}{z}} \times 1000000 \tag{1.5}$$

Mass spectrometry is able to provide accurate error calculation since the theoretical mass of molecules can be precisely calculated by using the exact mass of the atoms it is composed of. The highly accurate mass obtained using ultra-high resolution mass spectrometer enables us to assign formulae to unknown molecules which is a particularly desirable capability to analyse complex mixtures. The accuracy for FTICR MS instruments is proportional to the square of the magnetic field.

1.5.4 Apodisation

The apodisation is a signal processing technique which helps to improve the peaks' shape. While FTICR MS produces ultra-high resolution and narrow peaks, the base of those peaks can get quite broad. On either side of the peaks, it is common to observe smaller peaks often called "wiggles". For some high intensity peaks, the height of those "wiggles" can take over smaller genuine peaks.

In order to improve the signal, an apodisation method can be applied with the objective to iron out the "wiggles". The apodisation consists of decreasing the contribution of the extremities of the FID which are often of lower quality, and giving more weight to the centre of the FID. The apodisation can be optimised by using different functions to suit the needs of the user by giving different weights to different parts of the FID. Currently, users can choose between exponential, gaussian, sine, sine ^{2}, half-sine, shifted-sine and Kilgour methods when using an FTICR from Bruker but more methods exist such as Hanning and half-Hanning [96]. The drawback of the apodisation is that it will decrease the resolving power, but it can also dramatically improve the peaks' shapes. This leads to a cascade of improvements: as a result, the centroids will be improved, causing the peak picking algorithm to perform better, which in turns leads to improved molecular assignments. The apodisation is currently applied by default to most FTICR MS data as the benefits outweigh the costs but this remains under the user's control.

1.5.5 Absorption mode

FTICR MS data is able to produce an absorption-mode spectrum which has the advantage of an improved peak shape compared to the regular magnitude mode spectra. Due to the complexity of the phase-wrapping [97], this problem was solved recently and has since yielded major research on this topic. Phasing or phase correction is a signal-processing technique developed recently for FTICR MS data where the signal will be converted from the magnitude mode to the absorption mode [97–99]. An absorption mode spectrum presents several advantages such as a superior mass resolving power up to two fold, an increased mass accuracy and an increased sensitivity. The method was used recently in combination with other techniques to obtain the highest number of molecular assignments ever obtained [3].

1.5.6 Calibration

As mentioned earlier in the signal processing section, the instrument actually records a time domain spectrum which is then converted into a frequency spectrum which is theoretically directly converted into m/z thanks to the cyclotron frequency equation. In practise, only the reduced cyclotron frequencies are detected due to perturbations caused by electric fields and space-charge effects which affect the measurements. For this reason, it becomes necessary to perform a calibration in order to obtain an accurate mass spectrum. Many different calibration equations have been established over the years [100]. External calibration methods are not particularly suitable for FTICR MS as they do not take into account the differences in electric field, space charge and magnetron motion. It is commonly accepted that a FTICR MS spectrum

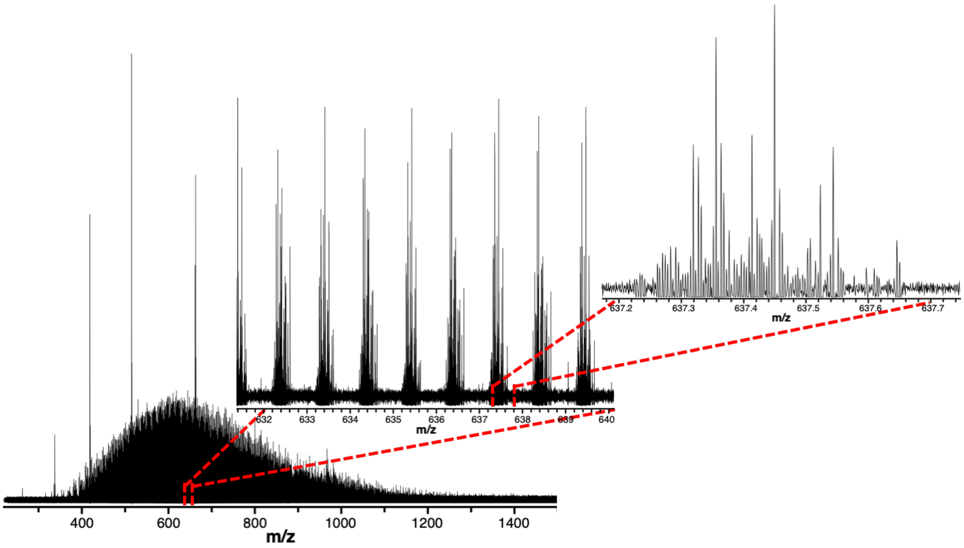


Figure 1.16: Example showing the complexity of the mass spectrum of a petroleum sample.

will be accurate within 1-2 ppm maximum while an Orbitrap spectra will be slightly higher [101]. It is sometimes necessary to perform an internal calibration using known molecular series.

1.6 Petroleomics

“Petroleomics” is the word used to designate the characterization of petroleum and its products by mass spectrometry [102, 102–113].

1.6.1 Petroleum

Petroleum is a resource created by heating and compressing over millions of years of plants and animal remains. It is an essential resource in our modern world even though efforts are made to reduce dependence upon it. Petroleum and its derivatives yield some of the most complex mass spectra observed to date, Figure 1.16 shows an example of the level of complexity and peak density that can currently be obtained.

1.6.2 Motivations

Any improvements in the characterisation methods of petroleum are needed for better understanding of petroleum composition, in order to solve the challenges posed by its production and refining [114, 115].

Today, the desirable light and sweet crude oils are becoming rarer, making the more complex and challenging varieties of crude oils more prominent as the lower quality heavy crude oils are more expensive and difficult to process. As a consequence more and more important to be able to process these crude oils as well as possible by improving the knowledge of their composition. Crude oil has a very wide range of use in everyday life. It is mostly used as a source of energy as a fuel for planes, cars, boats, *etc.* However, it is also used in solvents, plastics, dyes, waxes, lubricants, pharmaceuticals *etc.* [105]. Crude oils contain heteroatoms such as nitrogen, oxygen and sulfur but also metals that are toxic for the environment and decrease stability of the crude oil. [116]. The presence of molecules such as asphaltenes that tend to precipitate and cause blockages in the pipelines, leading to high maintenance costs [117].

Due to the complexity of the samples, petroleomics relies heavily on the ultra-high resolution mass spectrometry, particularly the high resolving power which can be obtained. In order to assign a molecular composition to the peaks detected, a low *ppm* mass error is necessary, ideally well below 1 *ppm*. The large varieties of molecules present in the petroleum samples generates the need for a large panel of ionisation methods in order to obtain a more complete picture of the composition of those samples due to the affinity of certain classes of molecules with particular ionisation methods. Finally, in order to obtain structural information from those samples, techniques such as chromatography needs to be employed.

1.6.3 Data analysis and visualisation

The petroleomics field currently employs three main criteria to classify the molecules observed in crude oils and more widely in any petroleum-related samples. Over the years, these criteria were used to create several plots and visualisation techniques to help cope with the complexity of petroleum samples.

Categorisation

Heteroatom class

The first category used to categorise the molecules of a crude oil is the heteroatom class. If a molecule only contains C and H atoms, it will be considered to belong to the CH class. But from the moment the molecule contains different additional atoms, the C and H will be ignored: for example, if the molecule only contains a N_1 in addition to the C and H, it will be classified in the N_1 class. Similarly, if the molecule contains a N_1 and an S in addition to the C and H, it will be classified in the NS class.

Double bond equivalents (DBE)

The second category is the double bond equivalents (DBE), also called hydrogen deficiency, which defines the degree of unsaturation of a molecule. This value is calculated by applying the following equation 1.6 to its molecular composition $C_cH_hN_nO_oS_s$.

$$DBE = c - \frac{h}{2} + \frac{n}{2} + 1 \tag{1.6}$$

This criterion is necessary as the long CH chains are naturally forming rings and double bonds inducing a loss of hydrogen atoms.

Carbon Number

Finally, the number of carbons atoms present within the molecule's formula assigned is used as a criterion to classify the molecules.

Visualisation

Due to the complexity of petroleum-related samples, there is a need for visualising the large amount of data generated. It quickly became necessary to find visualisation techniques in order to easily grasp the chemical composition of the samples and compare them. Several plots are extensively used in petroleomics and these popular plots have been summarised below.

DBE plot

The double bond equivalents (DBE) plot is probably the most common plot for petroleum related samples. The plot will usually focus on a specific heteroatom class and then represent the carbon number and DBE number on the axis. The dimension of the dot can be proportional to total intensity of all the peaks sharing the same DBE and carbon number. A colour scale can also be used to reflect the intensity in addition or in replacement of the dot size scale.

Class distribution

A bar plot with the relative intensity of each molecular class present in the sample is a useful representation to get a sense of the composition of a particular sample but also compare samples compositions. The relative intensity of all the samples being investigated will be displayed side by side for each class. The order of the molecular classes is also of crucial importance as they can display specific patterns as shown in figure 1.18.

Kendrick Mass Defect

The Kendrick mass defect (KMD) is a normalised mass scale unit using the CH_2 as unit instead of ^{12}C and is based on Kendrick's work [118, 119]. The Kendrick mass of each molecule is calculated using equation 1.7 while the KMD is calculated using equation 1.8.

$$Kendrick\ mass = IUPAC\ mass \times \frac{14.00000}{14.01565} \tag{1.7}$$

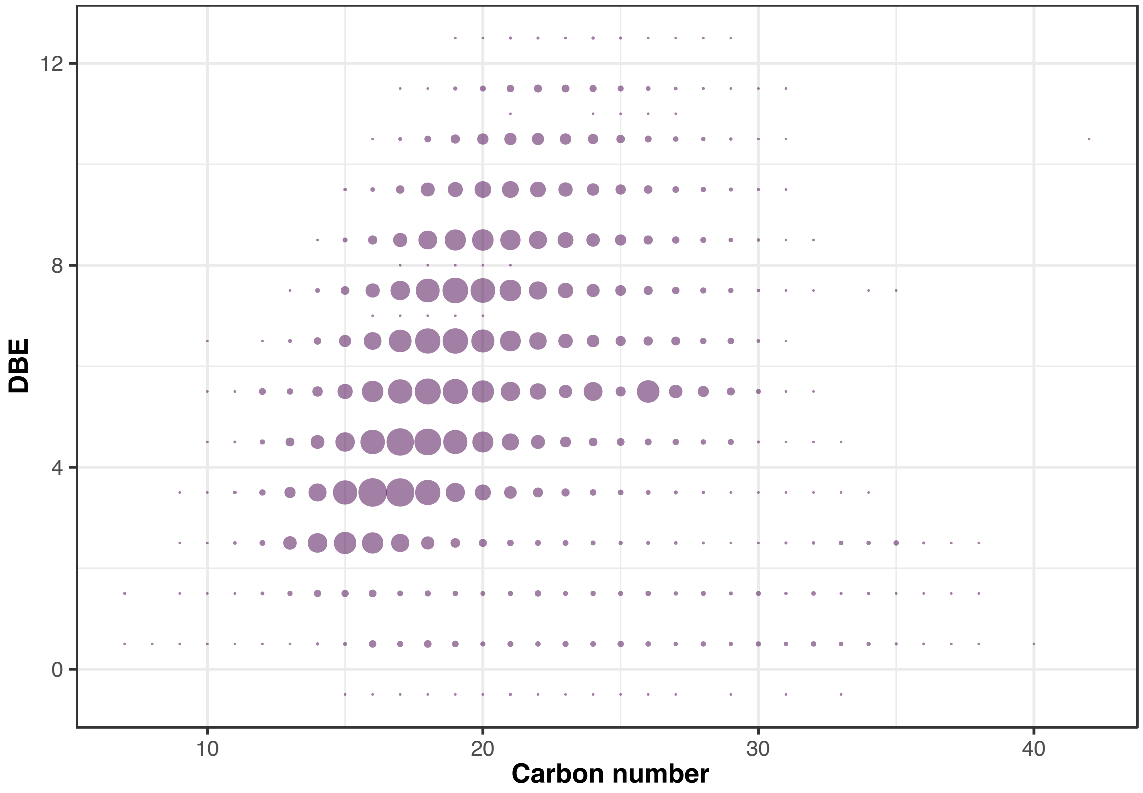


Figure 1.17: Scatter plot of the carbon number as x-axis, double bond equivalents as y-axis and the intensity as dot size for all the molecular classes of a petroleum related sample.

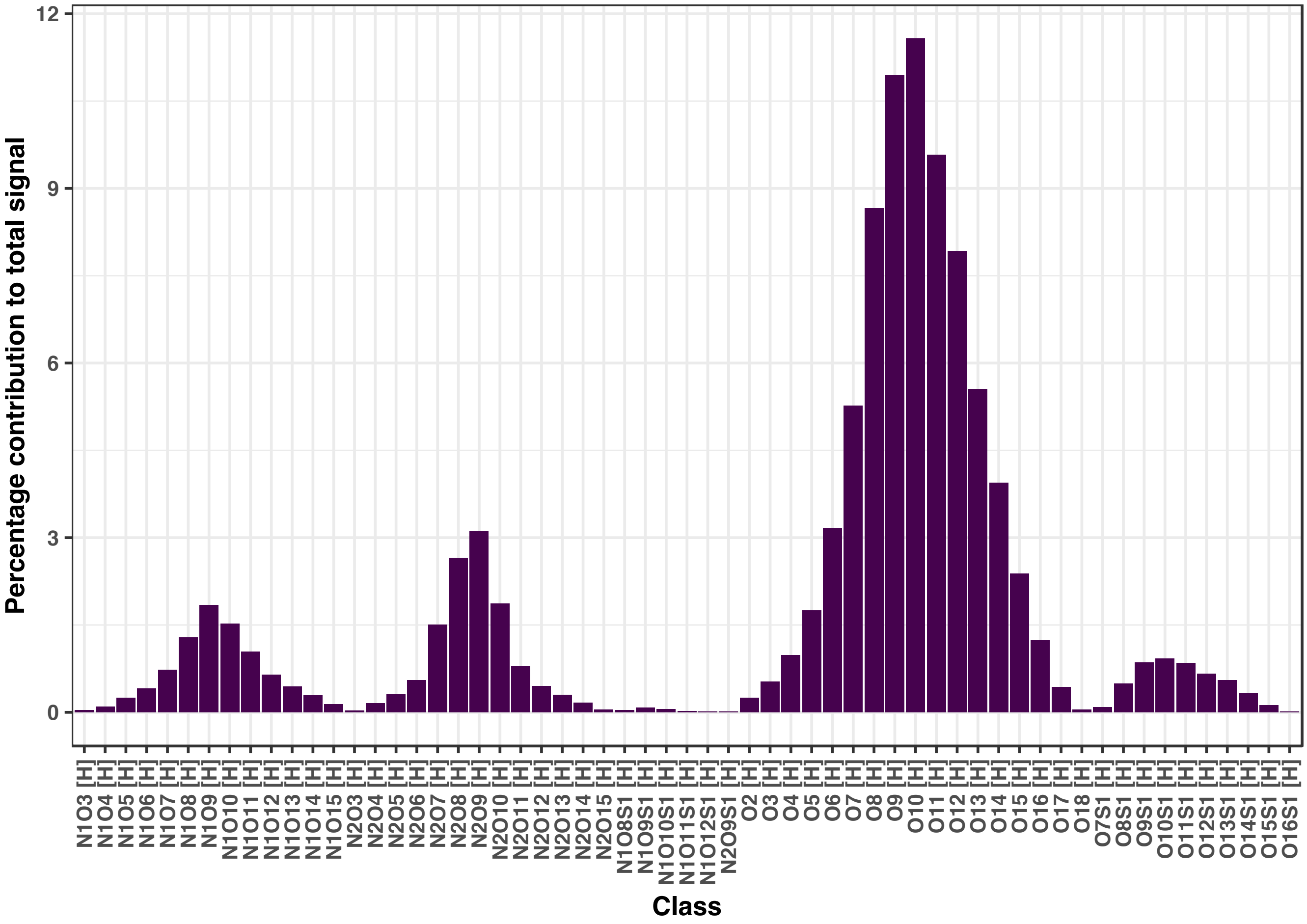


Figure 1.18: Example of barplot representing the relative contribution of each molecular classes for a disolved organic matter sample analysed by Orbitrap MS.

$$Kendrick\ mass\ defect = Nominal\ Kendrick\ mass - Exact\ Kendrick\ mass \tag{1.8}$$

The KMD has since been used to provide a new way to look into petroleomics data analysed with ultra-high resolution MS and help with molecular assignments [120]. Following the calculation of the KMD and using the nominal mass, a new type of 2D plot can be done which will allows to display a lot of information in a compact and clear figure. Thanks to this visual representation, outliers can be easily observed while it becomes possible to use the observed patterns to obtain more reliable assignments towards high masses.

van Krevelen

The van Krevelen diagram was introduced by Kim *et al.* [121] in 2003. It is a popular figure to visualise complex MS data and since it was first introduced many scientific publications analysing complex mixtures with ultra-high resolution MS have made use of it. The assigned molecules are distributed on a scatter plot with the H/C ratio vs O/C ratio as the axis. Other atoms can be used, often replacing the O/C ratio by the S/C or N/C. The dot's size is proportional to the total intensity of the molecules. Further analysis has shown that regions of the diagram can be attributed to specific compound classes.

1.6.4 Molecular assignments

Theory

In 2007, Kind and Fiehn [66] described seven golden rules which today remain a reference and are regularly used for molecular assignments in ultra-high resolution mass spectrometry and often implemented into molecular assignment software for FTICR MS [122]. The scope of each rule is listed below:

- Rule 1: Restrictions for element numbers to minimise computational time and

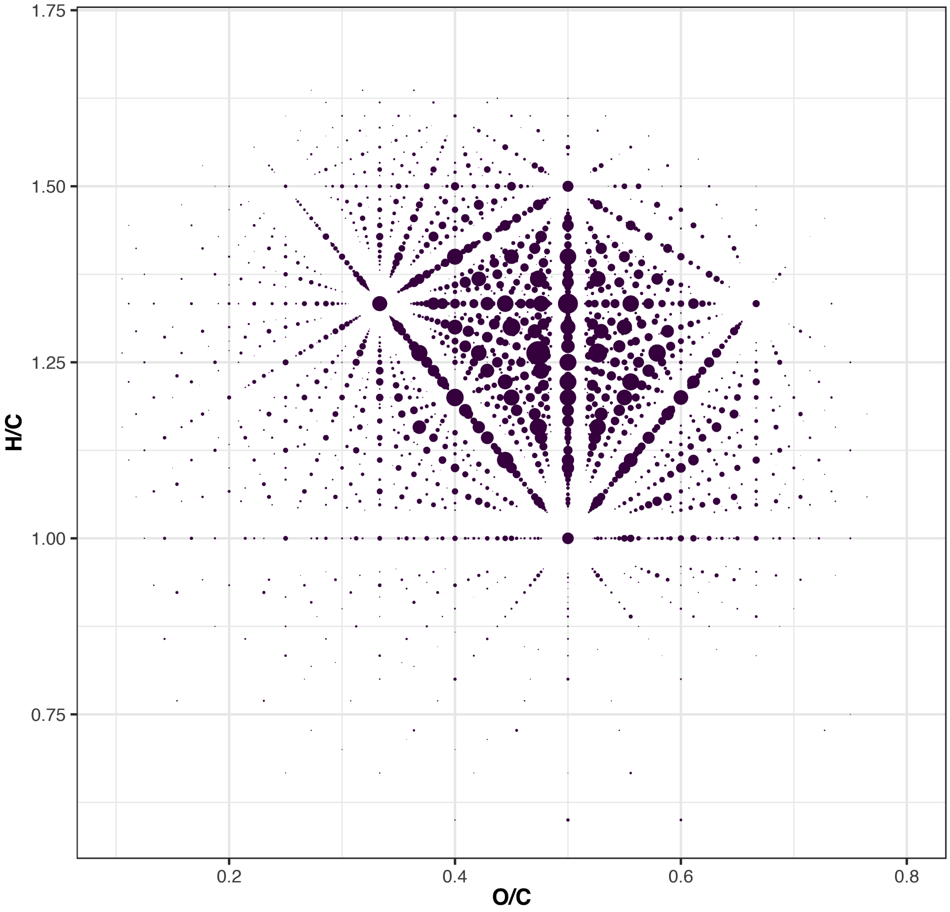


Figure 1.19: Example of a van Krevelen diagram for a dissolved organic matter sample analysed using an Orbitrap MS.

disk space

- Rule 2: LEWIS and SENIOR check

- Rule 3: Isotopic pattern filter

- Rule 4: Hydrogen/Carbon element ratio check

- Rule 5: Heteroatom ratio check

- Rule 6: Element probability check

- Rule 7: TMS (trimethylsilyl) check

Those rules are regularly used for petroleomics [105, 123–125], although not all are applicable for the molecular assignments of complex mixtures. Sometimes extra rules need to be added to take into account petroleum related samples characteristics such as the CH_2 series or, as reported by Leefmann *et al.* [126], forcing the DBE to be an integer value and specific H/C ratios.

Software

The molecular assignments of petroleum related samples can be performed using several methods listed below.

- In-house algorithms [43, 89]

- PetroOrg (Florida State University, Tallahassee, FL, U.S.A.) [127, 128]

- Composer (Sierra Analytics, Modesto, CA, U.S.A.) [77, 129]

To date there has not been a study comparing these different methods and currently only cost and personal experience drive the user to favour one or the other. Indeed, the price tag to use commercial software can be problematic for certain research groups who will then prefer to develop an in-house method.

1.6.5 Current challenges

Having appropriate, advanced and accurate tools to perform different analysis steps is important but for these to be fully practical it is also crucial that all these pieces fit well together. Issues such as data formatting, merging and transferring data can be time consuming, and can limit both the reliability and practical applicability of each existing tool. For instance, any change upstream will have to be manually repeated all along the chain. This is particularly true when analysing complex mixtures with FTICR MS, whose processing typically involved numerous steps from acquisition to the final conclusions. Four bottlenecks were identified and addressed, where new algorithms and workflows were developed as part of the research.

- In situations where a single sample is measured several replicated times, the standard was to analyse them separately and look for differences manually to ensure the conclusions were reliable and not due to acquisition variability. Most the time, only a single spectrum would be used to conclude, hence the lack of reliability.

- When acquiring multiple mass spectra of narrow ranges to improve the resolution of a single sample, the standard was to stitch segments by hand, manually trim the extremities of each mass spectrum which overlap and be forced to use a large overlap between segments to account for the "edge effect".

- When performing hyphenated ultra-high resolution of complex mixtures, the users had to manually divide a data set into as many time windows as desired and analyse each one separately resulting in a labour-intensive and error-prone task.

- After performing the molecular assignments, a number of standard figures had to be created within the molecular assignment software, the data of the figure exported, plotted again using a different software and often the figure needed to be improved with software like Inkscape or Illustrator.

1.6.6 Statistical challenges

The application of statistical methods to real world data comes with numerous challenges, even more so when working at the edge of instrumentation capabilities and ultra complex samples.

The multiple testing concept states that if we compare two samples based on many different criteria, at least one of them is bound to be different. This is particularly important to take into account as the complex mixtures studied have several thousand different molecules and a direct comparison for each of them would be the source of numerous false results. It then becomes necessary to find different ways of comparing such samples.

In mass spectrometry the intensity measured can be considered at best semi-quantitative and can vary greatly between acquisitions due to instrumental variations. As a consequence normalisation becomes necessary to adjust the range of the datasets we want to compare so that they are on a common scale.

Finally, ultra-high resolution mass spectrometry is used to analyse samples of an extreme level of complexity and a single ionisation method alone is insufficient to get the full picture of a sample. Each observer will obtain slightly different results depending of their instrumentation, parameters, sample preparation *etc.* Any progress in the field will likely result in seeing a little more information and it is not known if it will ever be possible to know the complete composition of a complex mixture sample with absolute certainty. Hence, developing data processing methods in such situations can prove to be very challenging in the absence of a gold standard to refer to and can only be as meticulous as possible to get as close as possible to the truth.

1.7 Statistical methods

Several statistical concepts were used throughout this thesis and will be briefly explained below.

1.7.1 Central Limit Theorem

Let us sample n independent observations from a population of mean μ and standard deviation σ, where μ and σ are finite. Let $\overline{X}_n = \sum_{i=1}^{n} \frac{X_i}{n}$ be a random variable representing the sample mean of the n independent observations.

The central limit theorem (CLT) says that for independent and identically distributed random variables then $\frac{\overline{X}_n - \mu}{\sigma}$ tends to $N(0, 1)$ as the number of n independent observations tends to infinity, even if the original variables are not normally distributed [130]. The CLT is important in statistics as it can be used to describe how a sample can be used to learn about the population it has been drawn from. The CLT implies that the mean of the distribution of the sample mean $\mu_{\overline{X}_n}$ will tend towards the population mean μ as n tends to infinity. Also, the standard deviation of the sampling distribution of $\overline{X}_n$, $\sigma_{\overline{X}_n}$ will tend towards satisfy $\sqrt{n}\sigma_{\overline{X}} \to \sigma$ as n tends to ∞. A common practical guideline is that $\overline{X}_n$ can often be taken to be approximately normally distributed for $n \geq 30$, although of course the actual n can change significantly from one application to another.

1.7.2 Quantile normalisation

Originally called quantile standardisation [131] and later renamed quantile normalisation [132], it is a technique extensively used to normalise the measured intensity when comparing microarray experiments. It is named “quantile normalisation” because the goal is that the measurements obtained in each sample have the same quantiles across samples, *i.e.* all samples have the same empirical distribution. Microarrays measure how active a particular gene is within a sample by measuring the intensity of different colours of light. The measurement of those intensities can be affected by experimental technical variations thus making the intensities of one microarray not comparable to another. Quantile normalisation is necessary to account for technical variabilities between experiments. This technique was adapted in this thesis to mass spectrometry in order to ensure comparable intensities across several mass spectra, *e.g.* multiple replicates obtained for a single sample.

Quantile normalisation was originally designed for microarrays, a type of data where there is an identical number of intensity measurements in each sample. The method proceeds as follows. First, the mean of the most intense value of each sample is calculated and used to replace the original intensity values within each sample. Then, the same procedure is applied for the 2nd most intense value. The process is repeated until reaching the lowest intensity value.

Quantile normalisation was applied to LC-FTICR MS data by Callister *et al.* [133] for peptide abundance measurement and compared to other normalisation techniques, namely central tendency [134], linear regression [135] and locally weighted regression [136]. The comparison relied on sufficient peak separation that allowed reliable matching of peaks between experiments and showed that while all methods reduced systematic bias there was a lack in definitive trend among the techniques.

Due to the extremely high peak density and number of intensities measured when analysing complex mixtures with ultra-high resolution mass spectrometry along with the variable number of peaks between replicates, the method had to be adapted. In order to address these issues, the intensities from all samples were combined and binned into 1000 quantiles. Then, the intensities of each sample were also divided into 1000 quantiles. A correction function was created by mapping the 1,000 quantiles from each sample to the 1,000 quantiles from the pooled data, and interpolating the resulting discrete map with the R function approxfun. Finally, individual measurements in each sample were adjusted by applying the interpolated map.

1.7.3 Clustering and mixture models

Cluster analysis or more commonly called clustering, consists of grouping a set of objects such that all the objects within the same group (cluster) are more similar to each other than the objects from other groups. In different terms, the goal is to distinguish subpopulations within a larger population. We can distinguish two main categories of clustering methods, hard clustering and soft clustering. Hard clustering methods apply when there is no overlap between the groups, they either belong to the group or they do not so that the clusters are distinct. Soft clustering methods allow for some overlap between the groups; the strength of association for each object to each cluster is calculated.

The choice of the type of distance used can have a big influence on the results, hence it is important to choose it carefully. For most methods, the default is the Euclidean distance but Manhattan distance (also called taxicab distance) is also used. Depending on the type of data, correlation-based distances may be preferable, for example, gene expression data. When using a correlation-based distance, two objects will be considered similar if their features are heavily correlated, even if when using an Euclidean distance, these two objects are very far apart. The Pearson correlation distance is the most commonly used but can give too much weight to outliers. Using the Spearman correlation distance can help mitigate the effect of outliers. Others such as Eisen cosine (a special case of the Pearson correlation) and Kendal correlation distance can also be used.

Many different algorithms exists to perform clustering. They have different specificities and will perform differently depending of the data analysed. Hence it is important to be able to pick the appropriate method. We can distinguish four main algorithms for clustering but many more exist. The K-means algorithm is a fast method which requires a user defined number of clusters to identify and can yield different results between runs as the initialisation of the clusters centres is random. The Mean-Shift clustering method is a sliding-window based algorithm which uses points with a set radius that will converge towards high density regions. Unlike the K-means algorithm there is no need to define the number of clusters, but determining the best radius can be challenging. Agglomerative Hierarchical Clustering can be either agglomerative (bottom-up) or divisive (top-down). In the first case, each data point is a cluster and the clusters that are most similar are merged successively until all have been merged into a single cluster while in the second case, all the points begin in a single cluster and the clusters are successively divided into the most different clusters until each point is a cluster. Gaussian Mixture Model clustering is the method employed in this thesis due to the advantages it offers in terms of the flexibility of the clusters covariance as well as supporting mixed membership by using probabilities to define which cluster a point belongs to. Indeed, mixture models are probabilistic models aimed at representing subpopulations within an overall population without prior information regarding the identity of any individual observation. Gaussian mixture models, also called mixture of Gaussians, are probability distributions which consists of observations drawn from a combination of Gaussians. By fitting a mixture of Gaussians to the data it becomes possible to estimate the probability of each data point of belonging to each component of the mixture of Gaussians. This allows mixture models to be used as a soft clustering method [137]. While sometimes the number of subpopulations and their characteristics (mean, variance) can be known, making the task of identifying to which population belong each objects easier, these characteristics are not necessarily required a priori when using a probabilistic model. Mixture models can be used to determine the characteristics of some subpopulations within a larger population. In this case, we do not need to know to which subpopulation a single observation belong to, the mixture model will yield a probability of each data point of belonging to a certain subpopulation. Expectation maximisation (EM) [138] is a commonly used algorithm to estimate the parameters.

The number of populations p to consider can either be set by the user or calculated. Two common strategies of choosing p are splitting the data into a training and validation set or using a criterion that balances goodness of fit against model complexity, for instance the Akaike information criterion (AIC) and the Bayes information criterion (BIC). The latter was used in this thesis. The Bayes information criterion (BIC) is also known as the Schwarz Criterion [139]. We considered a range of values for p, and compared each using the BIC. The BIC is calculated with $k \log(n) - 2\log(L(\widehat{\theta}))$ where n is the sample size/the number of observations, k is the number of parameters estimated, θ is the vector of all parameters and $L(\widehat{\theta})$ is the maximized value of the likelihood function of the model. The value of p is found by choosing the model with the smallest BIC.

1.7.4 Piecewise function

A piecewise function is defined by several sub-functions with each of them applying only to a specific part of the definition interval. They allow a function to be defined from a combination of several simple functions instead of using a single complicated function. Figure 1.20 illustrate the application of a piecewise function to isolate the central region from the edges needing intensity correction.

1.7.5 Data handling

The R programming language [140] is a free software environment first released in 1993 and originally aimed at statistical computing and graphics. A large number of packages have been developed in R to address a wide variety of needs around but not exclusively involving computation, visualisation and data manipulation. R recently benefited from a high pace of development due to the growing popularity of data science. Some of these packages have been made available under the umbrella name of tidyverse packages [141]. This is a collection of R packages with a distinctive design philosophy aimed at data science. Tidyverse packages share a similar design philosophy, grammar and data structure enabling an ultra-fast learning curve. They

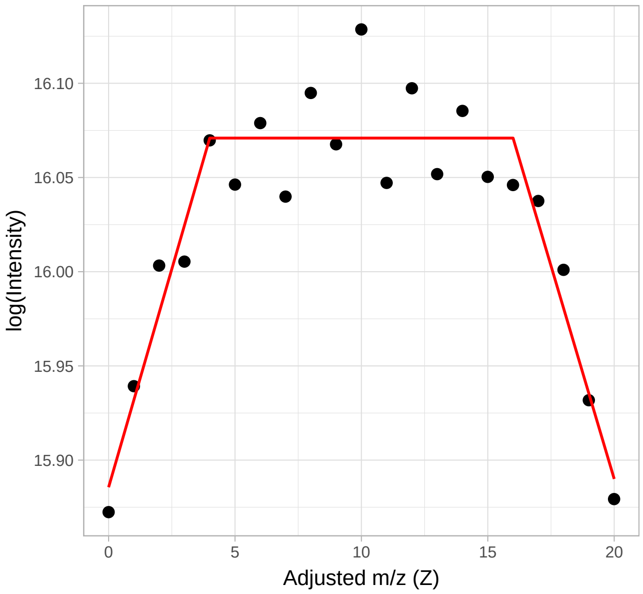


Figure 1.20: Example of the application of a piecewise function.

were crucial in giving the capability to rapidly import, clean, transform, model and visualise the vast amount of data generated during this thesis.

Finally, the Shiny package [142] made it possible to turn all the algorithms developed in this thesis into web-based interfaces, enabling users without any R knowledge to make use of the algorithms. Shiny was designed to work perfectly with the tidyverse packages.

1.8 Main software contributions

- Chapter 1: The Themis algorithm was first implemented into an R script. For ease of use, the script was combined with Rwui to provide an interface online to submit tasks by external users.

- Chapter 2: For the study, building on the previous experiences, Themis was recoded to use Tidyverse packages and optimised before being implemented into a Shiny app to provide for a higher level of interactivity.

- Chapter 3: The Rhapso algorithm was implemented into an R script before being paired with Rwui to provide an interface online to submit tasks by external users. Rhapso was later on, recoded to make use of the Tidyverse packages and optimised. Rwui was replaced by Shiny in order to provide interactivity with the user during the workflow along with visualisations of the process.

- Chapter 4: Firstly, the algorithm was developed, implemented in R and some basic visualisation capabilities were implemented. Later on, the code was implemented into a shiny app called XC-FTMS to process the raw data and obtain basic interactive visualisations. The capability to export the processed data was added and another Shiny app called CompareR was created to compare multiple datasets after processing with XC-FTMS. A derivative of compareR called PetRo-ExploreR was created to explore and compare molecular assignments from Composer without chromatography and processing with XC-FTMS. For simplicity, CompareR and PetRo-ExploreR were merged by using the implementation of Shiny modules and asynchronous computing. Finally, KairosMS was created after the merging of all the previously developed Shiny apps before being implemented on a server and used daily by Barrow's research group.

Chapter 2

Themis: Batch Preprocessing for Ultrahigh-Resolution Mass Spectra of Complex Mixtures

2.1 Context

The objective of this chapter is to better understand instrumental variability between acquisitions and to develop a method to increase the downstream reliability. It is particularly important to improve the reliability of the analysis upon which decisions can be taken. A new algorithm named Themis was developed and implemented using the R language [140]. Themis uses replicate measurement of the same sample in order to identify a consistent spectrum before molecular assignment. Themis will first identify similar peaks across the replicates prior to combining them using peak alignment and an adaptive mixture model-based strategy to separate consistent peaks from the unreliable ones. A new peak list combining all the replicates will be returned to the user which can be used for molecular assignment. The results showed that at high intensity, similar molecular assignments were obtained with and without Themis but at lower intensity, an improvement in the quality of the assignments was observed as the molecular series observed were more consistent and the RMS mass errors was smaller. Themis was demonstrated using petroleum-related samples but is expected to be applicable to a wide range of samples.

This chapter was published as an article in *Analytical Chemistry.* The research for this project was initiated during the MSc research project of the author and later improved during the PhD studies. All the code was written by the author, with advice from David Rossell, Simon E. F. Spencer and Mark P. Barrow. The NIST data was acquired by the author under Mark P. Barrow supervision while the South American crude oil data was acquired by Diana Catalina Palacio Lozano. The manuscript was written by the author.

2.2 Publication

analytical chemistry

Article

Cite This: *Anal. Chem.* 2017, 89, 11383-11390

pubs.acs.org/ac

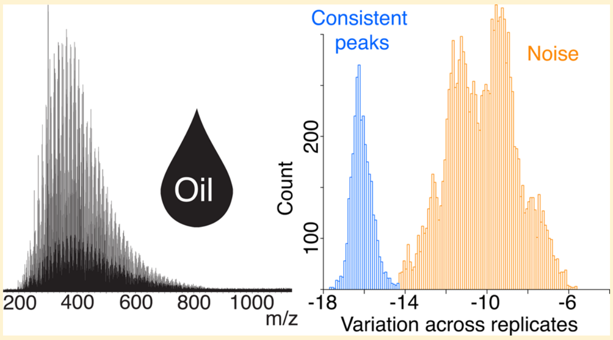
Themis: Batch Preprocessing for Ultrahigh-Resolution Mass Spectra of Complex Mixtures

Remy Gavard,[*,†] David Rossell,[‡,∥] Simon E. F. Spencer,[‡] and Mark P. Barrow[*,§]

[†]Molecular Analytical Sciences Centre for Doctoral Training, [‡]Department of Statistics, and [§]Department of Chemistry, University of Warwick, Coventry CV4 7AL, United Kingdom

[∥]Department of Economics & Business, Universitat Pompeu Fabra, Barcelona 08005, Spain

Supporting Information



ABSTRACT: Fourier transform ion cyclotron resonance mass spectrometry affords the resolving power to determine an unprecedented number of components in complex mixtures, such as petroleum. The software tools required to also analyze these data struggle to keep pace with advancing instrument capabilities and increasing quantities of data, particularly in terms of combining information efficiently across multiple replicates. Improved confidence in data and the use of replicates is particularly important where strategic decisions will be based upon the analysis. We present a new algorithm named Themis, developed using R, to jointly preprocess replicate measurements of a sample with the aim of improving consistency as a preliminary step to assigning peaks to chemical compositions. The main features of the algorithm are quality control criteria to detect failed runs, ensuring comparable magnitudes across replicates, peak alignment, and the use of an adaptive mixture model-based strategy to help distinguish true peaks from noise. The algorithm outputs a list of peaks reliably observed across replicates and facilitates data handling by preprocessing all replicates in a single step. The processed data produced by our algorithm can subsequently be analyzed by use of relevant specialized software. While Themis has been demonstrated with petroleum as an example of a complex mixture, its basic framework will be useful for complex samples arising from a variety of other applications.

Fourier transform ion cyclotron resonance mass spectrometry (FTICR MS)[1−6] represents a state-of-the-art technique for the study of complex mixtures that provides significant advantages in terms of ultrahigh resolving power and mass accuracy.[7] As a result of these performance advantages, FTICR MS affords the ability to distinguish molecules with very similar mass-to-charge ratios (m/z) on the basis of mass defects. Given the complexity of petroleum composition, these advantages are particularly relevant for the characterization of petroleum and its products by mass spectrometry,[8−14] an area of research that has become known as petroleomics. The following discussion will use application to this field as a suitable example, but it should be made clear that our methodology remains applicable to other complex samples. A variety of analytical approaches have been applied for the characterization of petroleum,[15] as well as environmental samples associated with alternative sources of oil.[16−19] Although high-field Orbitrap mass spectrometers are showing promising results for light and medium petroleum fractions, FTICR MS remains state-of-the-art for heavy fractions.[20−24] In order to address the challenges of producing and refining crude oil, one needs to develop a more detailed understanding of its composition through improvements in characterization methods.[25,26] Petroleomics is a field of growing importance because the most desirable varieties of crude oil are becoming more scarce. At the same time, the derivatives of crude oil are in everyday use and include products such as fuels, solvents, plastics, dyes, waxes, lubricants, and pharmaceuticals, among others.[27]

As the capabilities of FTICR MS have increased and produce larger and richer data sets, there has been an accompanying need for the development of more advanced software for data analysis.[28] Peak detection is a fundamental step as part of a data analysis workflow, regardless of application and instrument type. Reflecting this, a large variety of methods have been developed over time to improve peak picking.[29−37] Thus far, the development of data analysis methodologies for mass spectrometry have focused upon the characterization of biomolecules, such as peptides and proteins. In 2003, Patterson[38] argued in relation to the study of biomolecules that “data analysis is the Achilles heel of proteomics and our ability to generate data now outstrips our ability to analyze it”. Today, the ability to analyze proteomics data is considerably improved, with many software tools available. The analysis of data from complex mixtures[29−31,39−42] is different from that of proteomics, metab-

Received: June 18, 2017
Accepted: October 6, 2017
Published: October 6, 2017

DOI: 10.1021/acs.analchem.7b02345
Anal. Chem. 2017, 89, 11383−11390

olomics, or polymer data, for example, given the higher peak density (15−30 peaks in a 0.5 m/z window)[10,12,35] and different patterns within the data. While proteomics has typically involved lower resolution instrumentation and higher throughput techniques (automated systems analyzing many samples per day), of greatest need when analyzing petroleomic samples is ultrahigh resolution, making FTICR MS the tool of choice. Another difference is that software tools for biomolecule characterization are designed to match protein or peptide sequences by use of online data banks. For complex mixtures such as petroleum, the strategy is to determine series of heteroatom-containing organic components, with thousands of possible compositions ($C_cH_hN_nO_oS_s$).

One example of data analysis software is Mass-Up,[43,44] an open source mass spectrometry program that gathers functions such as normalization, peak detection and peak matching of replicated samples. It was developed specifically for proteomics matrix-assisted laser desorption/ionization (MALDI) data,[45−47] when typically a lower resolution mass analyzer was used, such as time-of-flight mass spectrometry. While a software tool designed for other varieties of mass analyzers and other sample types can be invaluable for their intended purposes, they are not appropriate for analysis of complex mixtures due to their design for use with lower resolution data and wider mass error tolerances (e.g., hundreds of parts per million, ppm). There is an emerging need for improved data analysis strategies for complex mixtures, such as for petroleomics applications, that are designed for the resolution of tens of thousands of peaks.[15]

Currently, a typical workflow for analysis using FTICR MS may consist of acquiring one spectrum per sample and processing each individual sample with specialized petroleomics software, such as Composer (Sierra Analytics, Modesto, CA)[16,20] or PetroOrg (Florida State University, Tallahassee, FL).[48] The results from individual samples can then be recalibrated with respect to m/z to compensate for electric field effects (including space-charge due to the presence of the ions) within FTICR cells.[13,49−51] As the field becomes more mature, increasing numbers of samples need to be analyzed within a practical time frame, including multiple experiments to ensure repeatability of results. A fundamental concern is to ensure that the data are reliable and false assignments are reduced by removing as much noise as possible before performing in-depth data analysis.[36]

To improve the reliability of analysis of crude oil spectra, Hur et al.[52,53] have previously highlighted the importance of the use of replicates. The need for replicates was demonstrated for FTICR MS-based metabolomics data,[54] and recently replicates were used to generate an averaged mass spectrum.[55] Our approach is based on the idea that, to fully capitalize on the advantages brought by repeat measurements, replicates should be processed together instead of separately. The first challenge is that complex mixture data sets present a high density of peaks of interest, hampering the identification of those that are consistent across replicates. A second challenge is that of the peak magnitudes: some peaks are similar in magnitude to the noise level, and it is also possible that peak magnitudes can differ significantly across replicates.

A simple strategy to avoid false positives is to use stringent parameters when making peak assignments, for example, setting a higher minimal signal-to-noise (S/N) ratio when picking peaks, or a narrower tolerance of mass error (more limited deviation on the m/z axis). There are advantages in working with such peak lists rather than full mass spectra in terms of simplicity and reduced computational cost. The problem with these strategies is that they may, at an early stage, discard low-magnitude peaks that provide valuable information and are consistently observed across replicates. That is, they may be too aggressive in reducing the number of peaks, with consequences for subsequent interpretation. In contrast, using settings that are too permissive risks including a high number of false positives. Furthermore, the fundamental issue remains that applying thresholds to individual spectra loses the opportunity to share information across samples. Ideally, one would like to preserve all potential peaks in individual samples and then use information across replicates to identify which peaks are truly reliable. Traditionally, denoising methods are based on signal magnitude, using either the shape of the peaks or their magnitudes to discriminate between noise and reliable peaks. By contrast, we propose to denoise the spectra by focusing upon the consistency on the m/z scale, with peak magnitude being used as a secondary criterion. Our algorithm ensures reproducibility of the peak list extracted from a sample and produces a single consensus list. Figure 1 provides a schematic representation. The first stage is to extract a peak list from each replicate, with a permissive S/N ratio. The second step is to detect anomalous replicates by use of quality control statistics based upon their molecular weight distributions. The third stage is the use of quantile normalization to ensure that magnitudes are comparable across replicates. Finally, the fourth step uses a statistical mixture modeling approach to distinguish reliable peaks from those due to noise.

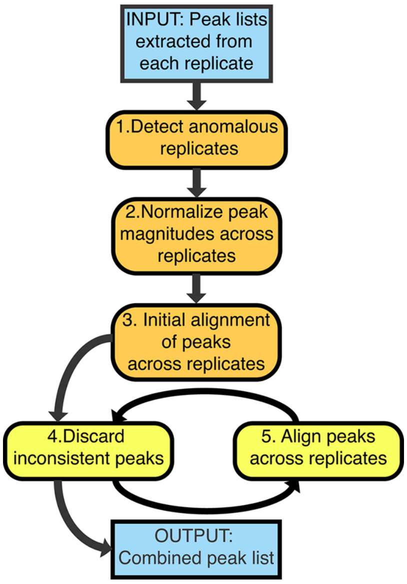

INPUT: Peak lists extracted from each replicate
1.Detect anomalous replicates
2.Normalize peak magnitudes across replicates
3. Initial alignment of peaks across replicates
4.Discard inconsistent peaks
5. Align peaks across replicates
OUTPUT: Combined peak list

Figure 1. Schematic of the Themis preprocessing algorithm.

■ METHODOLOGY

Sample Preparation. Sample A was an NIST light sour crude oil sample [National Institute of Standards and Technology, SRM 2721, crude oil (light sour)], which was dissolved at 0.1 mg/mL in an 80:20 mixture of propan-2-ol/toluene (Fisher Scientific, Loughborough, U.K.), with formic acid (Sigma−Aldrich Co. Ltd., Gillingham, U.K.) being added at 1% by volume to aid protonation. Sample B was a South American crude oil sample that was dissolved at 0.05 mg/mL in 50:50 propan-2-ol/toluene (Fisher Scientific, Loughborough, U.K.), with 0.2% formic acid (Sigma−Aldrich Co. Ltd., Gillingham, U.K.) for positive-ion mode or 0.8% ammonium hydroxide (Sigma−Aldrich Co. Ltd., Gillingham, U.K.) for negative-ion mode. Sample C was a Kodak naphthenic acid (NA) mixture (The Eastman Kodak Co., Rochester, NY) was prepared at 0.1 mg/mL in acetonitrile (VWR Chemicals, Lutterworth, U.K.) without the addition of any ammonium hydroxide.

Instrumentation. Mass spectra were acquired as 4 M data sets (i.e. approximately 4 million data points) using an Apollo II electrospray ionization (ESI) source, coupled to a 12 T solariX FTICR mass spectrometer (Bruker Daltonik GmbH, Bremen, Germany). For sample A, the instrument was operated in positive-ion mode and six repeat measurements were obtained, each of them being the result of 300 scans. Sample B was recorded in both positive- and negative-ion modes with five and six repeat measurements, respectively. The number of scans was 300 for the negative-ion mode and 210 for the positive-ion mode. Sample C was recorded in negative-ion mode with six repeat measurements and 100 scans. In all cases, replicates were obtained the same day in a single session on the instrument. Broadband mass spectra were acquired, where a single zero fill and sine-bell apodization were applied before a Fourier transform.

Statistical Processing. The spectra were exported from solariXcontrol to DataAnalysis 4.2, which was used to extract peak information by use of the following parameters: peak finder FTMS, S/N threshold 4, relative magnitude threshold (base peak) 0.01%, and absolute magnitude threshold 100%. The spectrum was not subject to any modification other than application of the default apodization before undergoing Fourier transformation.

Step 1: Detect Anomalous Replicates. The average molecular weight $\bar{W}_j$ of each replicate $j = 1, ..., r$, where r is the number of replicates, was calculated as a quality control metric to detect anomalous runs:

$$\bar{W}_j = \frac{\sum_{i=1}^{n_j} M(i, j) I(i, j)}{\sum_{i=1}^{n_j} I(i, j)} \qquad (1)$$

where $M(i,j)$ is the m/z value of peak i in sample j, $I(i,j)$ is the corresponding magnitude, and n_j is the number of peaks in sample j.

To identify what constitutes an anomalous average molecular weight, we must first characterize their reference distribution from the data. Given that the mean and the standard deviation (SD) can be heavily influenced by outliers, we used robust measures of the center and spread, namely, the median and the corrected median absolute deviation (MAD),[56−58] given by eq 2:

$$\mathrm{MAD}(x_1, ..., x_n) = b\{\mathrm{median}_i[|x_i - \mathrm{median}_j(x_j)|]\} \qquad (2)$$

with $b = 1.4826$ for Gaussian distributions. Motivated by the central limit theorem, we assume that the average molecular weights of nonanomalous samples are approximately normally distributed around a mean μ, with standard deviation σ. We wish to find an interval $(\mu - y, \mu + y)$ that, in the absence of any anomalies, should contain all n samples with probability $1 - \alpha$, where α is a user-specified error threshold (by default $\alpha = 0.05$). If it is assumed that replicates are independent, for a given μ and σ it can be seen that

$$y = \Phi^{-1}\left[\frac{(1-\alpha)^{1/r}}{2}, \mu, \sigma\right] \qquad (3)$$

where $\Phi^{-1}(x, \mu, \sigma)$ is the inverse normal cumulative distribution function.

Step 2: Normalize Peak Magnitudes across Replicates. To take into account that the dynamic range of magnitudes varies across samples, we apply quantile normalization.[59,60] This ensures that the distribution of magnitudes is identical across replicates, facilitating subsequent peak alignment.

Step 3: Initial Alignment of Peaks across Replicates. Our peak alignment strategy has two steps, a first one to initialize (step 3) and a second one used iteratively to refine the matching (step 5). For clarity, we denote any value that may change across iterations with a k superscript to indicate the value at the kth iteration. In the initialization step, $k = 0$. To initialize the peak alignment, we take the sample with the largest number of peaks as a reference and match peaks in all other replicates to the reference. Let $m^{(k)}$ denote the number of aligned peaks in iteration k and $m^{(0)}$ the number of peaks in the longest replicate at initialization. For each peak in the reference replicate, we match to the closest peak in each replicate in terms of its m/z value.

Step 4: Discarding Inconsistent Peaks. We compute the standard deviation of the m/z values matched to reference peak $i = 1, ..., m^{(k)}$, which we denote $Z_i^{(k)}$. Intuitively, peaks that are consistently observed across samples should show similar m/z values, resulting in low $Z_i^{(k)}$. That is, one typically observes a subpopulation of reliable peaks with low $Z_i^{(k)}$ and another subpopulation of less reliable peaks with high $Z_i^{(k)}$, likely due to noise. This motivated us to fit a mixture model to separate these subpopulations. Let $P_{ij}^{(k)} \in \{1, ..., n_j\}$ be the index of the peak in replicate j (for $j = 1, ..., r$) that is matched to the ith reference peak in iteration k. We define the mean m/z and magnitude for reference peak $i = 1, ..., m^{(k)}$ in eqs 4 and 5:

$$\bar{M}_i^{(k)} = \frac{1}{r}\sum_{j=1}^{r} M[P_{ij}^{(k)}, j] \qquad (4)$$

$$\bar{I}_i^{(k)} = \frac{1}{r}\sum_{j=1}^{r} I[P_{ij}^{(k)}, j] \qquad (5)$$

Equations 6 and 7 give the respective m/z and magnitude standard deviations:

$$Z_i^{(k)} = \sqrt{\frac{\sum_{j=1}^{r}\{M[P_{ij}^{(k)}, j] - \bar{M}_i^{(k)}\}^2}{r - 1}} \qquad (6)$$

$$T_i^{(k)} = \sqrt{\frac{\sum_{j=1}^{r}\{I[P_{ij}^{(k)}, j] - \bar{I}_i^{(k)}\}^2}{r - 1}} \qquad (7)$$

An important step in our algorithm is to identify the subpopulation of peaks consistently observed across replicates. To this end we fit a normal mixture model[61] to $\log\left[\frac{Z_i^{(k)}}{\bar{M}_i^{(k)}}\right]$ by use of the function `mclust`[62,63] in the R package mclust. Calculating the relative standard deviation (RSD), by dividing the standard deviation $Z_i^{(k)}$ of a peak by its $\bar{M}_i^{(k)}$, allows us to express the results in a unit equivalent to parts per million (ppm), which is a standard unit when expressing the mass error associated with the m/z of a peak. In addition, it helps to make the mixture model more reliable, as it allows to be equally stringent for high and low m/z, as the SD tends to be larger for high m/z values. We denote $G_i^{(k)} = \log\left[\frac{Z_i^{(k)}}{\bar{M}_i^{(k)}}\right]$.

In `mclust`, we set the maximum number of components to capture peak subpopulations of high and low quality, and potentially a third one of intermediate quality. We use the Bayesian information criterion (BIC) to select the final number of components in the mixture. Themis then selects the population with lowest mean $G_i^{(k)}$. When this mean is >1 ppm, a warning is given to signal that the data set may be of low quality. The first time that step 4 is performed, a conservative threshold is used: peaks are discarded if they have a probability below 0.01 of belonging to the selected subpopulation. Doing so allows the algorithm to remove the majority of the obvious noise while making sure not to discard any potentially relevant peaks. At this step, the presence of leftover noise is not problematic, as further refinement will be performed by iteratively repeating steps 4 and 5.

In each subsequent repetition of step 4 in future iterations, the 0.01 threshold is increased by 0.01, up to a maximum of 0.5. The goal is that, by the end of the iterative process, only peaks belonging to the high-quality subpopulation remain.

Step 5: Align Peaks across Replicates. After peak removal in step 4, we refine the peak matching across samples using a combined criterion that incorporates both magnitude and m/z, in contrast to step 3, where we only used m/z. Intuitively, the criterion seeks the closest peak on the basis of a score where m/z and magnitude are weighted according to their inherent variability. Given that the precision of the variance estimates in eqs 6 and 7 may suffer when the number of replicates r is low, we borrow strength across peaks by using the hierarchical empirical Bayes framework proposed by Smyth and Speed,[64] implemented in function `squeezeVar` from the Bioconductor package limma.[65] We denote $\tilde{Z}_i^{(k-1)}$ and $\tilde{T}_i^{(k-1)}$ as the refined estimates analogous to Z_i^{k-1} and T_i^{k-1}. Specifically, the score to measure the closeness of peak l in sample j to reference peak i at the kth iteration is given by eq 8:

$$S_{ijl}^{(k)} = \frac{|M_{lj} - \overline{M}_i^{(k-1)}|}{\tilde{Z}_i^{(k-1)}} + \frac{|I_{lj} - \overline{I}_i^{(k-1)}|}{\tilde{T}_i^{(k-1)}} \tag{8}$$

The highest scoring peak in each replicate replaces the one chosen in the initial matching. After this peak assignment we update $\overline{M}_i^{(k)}$, $\overline{I}_i^{(k)}$, $Z_i^{(k)}$, $T_i^{(k)}$, $\tilde{Z}_i^{(k)}$, and $\tilde{T}_i^{(k)}$. To obtain a scoring method that limits the effect of outliers and can be computed in cases where a reference peak is absent from one or a few replicates, we added the possibility to replace eqs 4, 5, 6, and 7 by trimmed means and standard deviations. That is, the replicate(s) with largest $S_{ijl}^{(k)}$ in eq 8 can be discarded.

Themis iteratively repeats steps 4 and 5 until either the BIC selects a single population or else all remaining subpopulations have a mean less than ≡1 ppm and the peak list does not change between five successive iterations.

Output Combined Peak List. The final output is a list composed of three tables containing respectively the m/z values, magnitude values, and final peak list. The m/z and magnitude tables have a $[m^K, r]$ dimension where m^K indicates a peak and r a replicate number. The final reference peak list file is an $m^{(K)} \times 4$ table, where $m^{(K)}$ is the number of reference peaks at the final iteration K. Themis stores the m/z, SD(m/z), magnitude, and SD(magnitude) of each peak as separate columns. Themis provides a function to extract columns 1 and 3 from the peak list table to a `.txt` file containing a first column with the m/z and a second with the corresponding magnitudes.

RESULTS AND DISCUSSION

The performance of the preprocessing methodology was assessed for a sample of NIST light sour crude oil, a naphthenic acid sample,[66] and a crude oil sample analyzed in both positive- and negative-ion modes. We also used a data set that was recorded by use of deliberately aberrant instrument parameters to study the ability of our framework to detect such situations. Themis is available as an online tool at http://themis.warwick.ac.uk/themis and is based on Rwui[67] to generate a web interface for the R script.

A common strategy to improve accuracy in m/z values is to apply a calibration step based upon a list of reference peaks. This step can in principle be applied to each individual peak list given as input to Themis or to the single reference peak list output by Themis. It is common that there can be minor variations in mass errors between different data sets. Calibrating each individual peak list before passing to Themis can significantly improve the quality of the processing due to improved consistency.

In order to test step 1 of the algorithm, we recorded a spectrum of NIST sample A where the ICR cell was intentionally overloaded with a high ion population and one where we deactivated ion source dissociation (ISD), which is used to minimize noncovalent aggregation. These two peak lists were extracted and included with the six others that were acquired under normal conditions. The algorithm was able to detect these two spectra as aberrant and remove them. Similarly, we then substituted the ISD off-peak list for naphthenic acid sample C for the list of replicates for sample A (NIST) used before, to verify that our method would be able to cover this potential error. Again, the algorithm successfully detected the spectrum that did not correspond to sample A (NIST) and removed it. The procedure was illustrated in Figure 2, where spectra C and E were discarded after modelization while the other were kept.

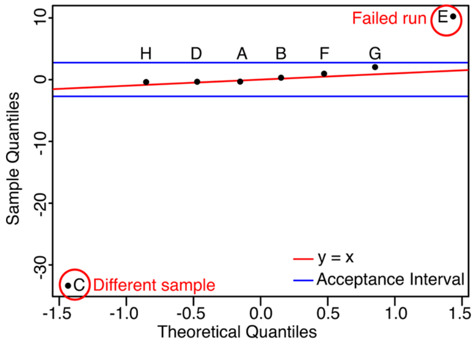


Figure 2. Automated detection of outliers and use of a series of repeat measurements to produce an averaged data set for characterization.

To assess step 2, we produced a quantile–quantile plot (q–q plot) to compare the magnitudes across samples. We observed considerable variation between replicates (Figure 3A), particularly for greater magnitudes. Low magnitudes (ranging from 0 to 0.5 × 10^8) exhibited a similar distribution across replicates. In the region from 0.5 × 10^8 to 2.0 × 10^8, we observe an inflection of the line, which demonstrates that the magnitude is different but the overall shape is similar. Also, single high-magnitude peaks such as those originating from contaminants will influence the total signal magnitude for the corresponding data set. The quantile-normalized magnitudes are shown in Figure 3B. Similar results were observed for other samples (see Figures S1–S3 in Supporting Information).

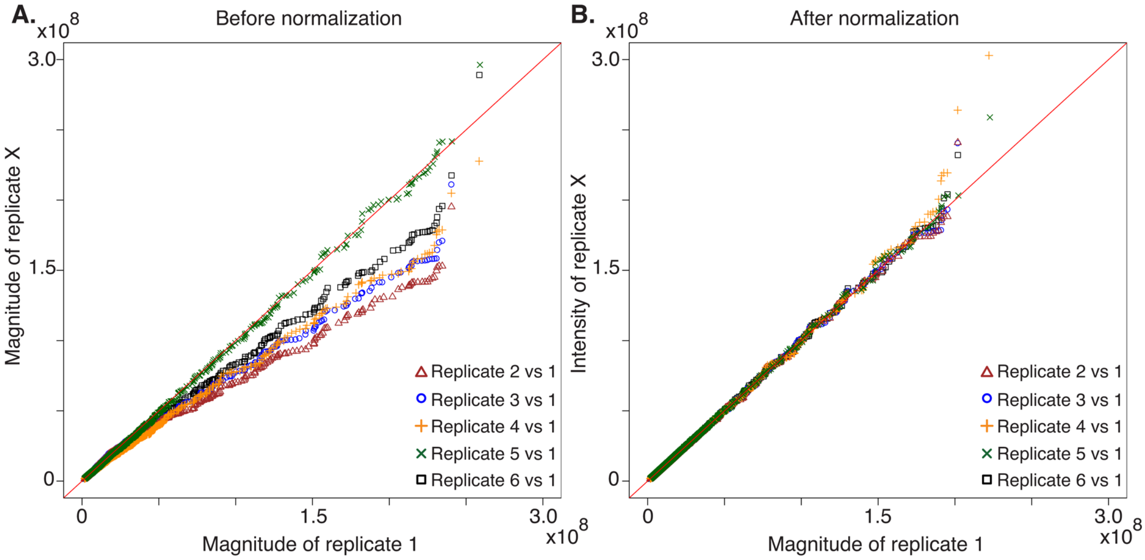


Figure 3. Quantile−quantile plots of magnitudes of six replicates, (A) before and (B) after quantile normalization, for the NIST light sour crude oil sample.

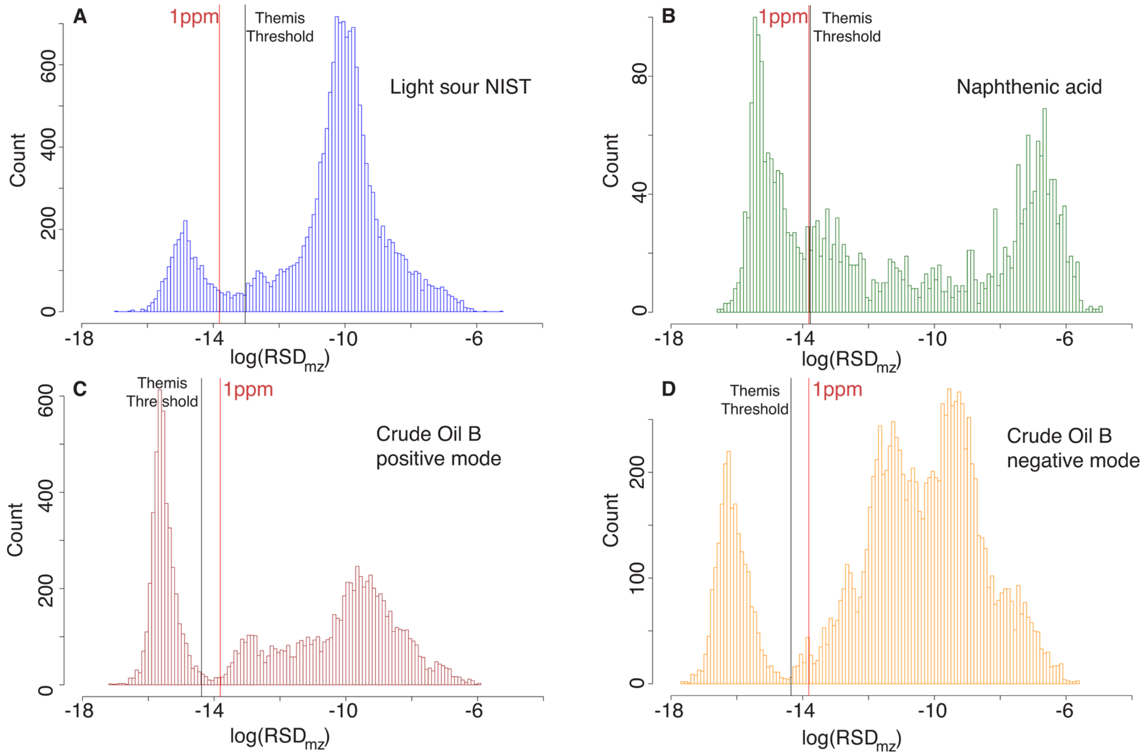


Figure 4. Histograms of log absolute relative standard deviation (RSD) for peaks matched under the initial matching of different samples. The red line, labeled 1 ppm, represents a standard deviation equivalent to 1 ppm; the black line, labeled Themis Threshold, shows the position of the threshold between noise and consistent peaks after Themis processing. (A) NIST light sour crude oil sample A, positive-ion ESI; (B) naphthenic acid sample C, negative-ion ESI; (C) South American crude oil sample B, positive-ion ESI; (D) South American crude oil sample B, negative-ion ESI.

Figure 4 was produced after the initial peak alignment in step 3. It shows a histogram of log-standard deviation within-peak m/z values for multiple data sets. It reveals the presence of a subpopulation with low log(SD/mz) corresponding to reliable peaks, that is, with similar m/z across replicates, and another subpopulation with high log(SD/mz) mostly composed of noise. Evidence of distinct subpopulations was observed in all data sets we have analyzed so far, including different samples, instruments, users, and peak list extraction methods.

Step 4 is critical because, although in all data sets there are clearly distinct subpopulations, the distributions are different. That is, the threshold used to distinguish reliable from unreliable peaks cannot be a fixed quantity but instead needs to be data-dependent. The red line, labeled 1 ppm, indicates a fixed threshold equivalent to the log of 1 ppm, a value typically used as a benchmark for accuracy of the mass measurement. For comparison, the black line, labeled Themis Threshold, indicates the final threshold identified by our mixture model framework, which is adaptive to the nature of the individual data sets. For instance, for both ionizations of the crude oil B, a more tolerant threshold was used. While for Figure 4C, the threshold immediately makes sense to the eye, Figure 4D may give the impression of selecting part of the noise population. This is because, during the refining process, the shape of the population changes due to the scoring algorithm. With the NIST data, the refinement led to the removal of peaks that ended up being present several times following the rematching performed during the iterative part of the algorithm. During this part, the peaks in between the two large populations resulting from valid peaks (on the left) and noise peaks (on the right) slowly joined these large populations. The more challenging naphthenic acid sample ended up with a threshold close to 1 ppm. This data set had considerably fewer peaks than the three other data sets, making the mixture modeling more challenging. Despite fewer peaks for the mixture modeling, the algorithm still managed to isolate a consistent population.

An example that highlights the benefits of the algorithm is given in Figure 5, with close examination of a region around the peak m/z = 248.1434, for two replicate data sets and Themis output using all replicates. It is possible for a user to manually go through every data set, adjust the parameters, and get an optimal assignment. This is a laborious task usually avoided by using default data analysis parameters across the data sets. Manual adjustments of the parameters on a case-by-case basis is the way to assign the greatest possible number of peaks but also leads to an increased risk of false assignments due to inclusion of noise peaks. In Figure 5, noise was observed between m/z 248.00 and 248.40 for the individual replicates but was not observed in the data set produced by Themis.

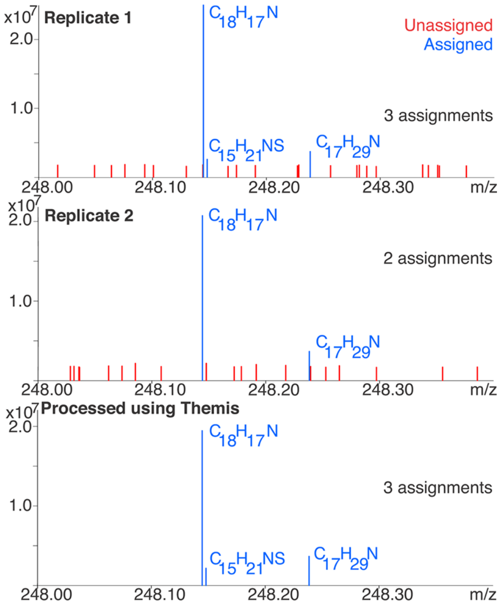


Figure 5. Peak assignment between m/z 248.00 and 248.40, showing two replicate data sets and the one produced by Themis using all replicates. In replicate 1, composition $C_{15}H_{21}NS$ is present just above the noise threshold, while in replicate 2, the peak is below the noise threshold and so not assigned.

Figure S5 shows a larger m/z region to illustrate the peak list obtained across the six replicates of the NIST sample. Our algorithm identified peaks that were consistently observed across replicates with a S/N ratio as low as 4.5 up to 15 for this section between 700 and 710 m/z. For comparison, in the region around 400 m/z the peaks are routinely observed with a S/N ratio of more than 500.

The raw peak lists for the NIST light sour crude oil sample contained an average of approximately 16 400 peaks. Out of these, Themis identified 2260 reference peaks deemed to be common among all replicates. The number of entries increased to 2523, when the peaks were allowed to be absent from one of the replicates at step 5 of our algorithm, and to 2820, when peaks could be absent from two of the replicates. Allowing peaks to be absent from one or more replicates increases the ability to detect potentially relevant peaks, at the expense of an increased risk of potentially including less reliable peaks.

We compared the chemical composition obtained from unprocessed spectra with that from the peaks list produced by our algorithm for NIST light sour crude oil. For the purposes of the comparison, the N_1 class has been used, as it is the most prevalent and the more challenging NS class because of its lower magnitudes. The data were recalibrated by use of the N_1 class and a walking algorithm.[51] The m/z match tolerance was set to 1 ppm. For the N_1 class, the results demonstrated that the reference peak list output by Themis has a similar chemical composition after processing. Plots of contributions by double-bond equivalents (DBE) and carbon number for the N_1 class are shown in Figure S6. Figure S6 demonstrates that the assignments were very similar despite the output from Themis containing a fraction of the number of peaks, indicating that information was not being lost during the processing. Themis is expected to improve picking of peaks of low S/N ratio, and therefore we next looked at the NS class, which forms a smaller contribution to the profile. Figure 6 shows the contributions of homologous series to the NS class, where the NS class included many lower-magnitude peaks, as already shown in Figure 5.

At first glance, a wider range of carbon numbers and DBE appeared to be observed when no processing was used. Closer inspection of the data, however, revealed gaps within the DBE series; this can typically be used to differentiate between likely correct and incorrect assignments within petroleum data, due to the well-known presence of homologous series. The additional assignments in the unprocessed replicates were also associated with higher mass errors, further indicating that they were of questionable validity. Furthermore, manual inspection of the data also revealed that the peaks in question were not consistently observed across the replicates. The combination of these observations provides evidence that removal of these assignments does not represent a loss of information but, in fact, a reduction in false positives. After processing with Themis, the series observed were more consistent and the associated range of mass errors was smaller. While Themis reduced the size of the peak list by differentiating noise and inconsistent peaks, information is not being lost. In fact, the processing has facilitated an improvement in data quality by reducing interference in the analysis from false positives.

Figure 7 is a histogram of the mass errors associated with assignments of the NS class for sample A (NIST) before processing (Figure 7A) and after Themis processing (Figure 7B). Typical mass errors were below 1 ppm for both data sets, with

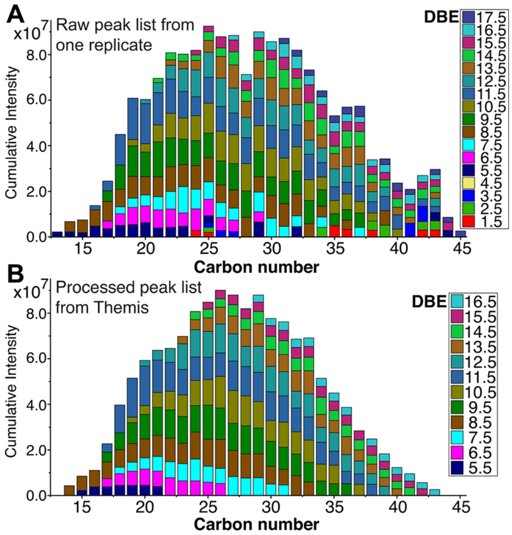


Figure 6. Stacked bar plots of the carbon number and DBE distributions for the NS class for the NIST light sour crude oil sample. The results of data analyses are shown for (A) a single replicate, where the total peak list (all classes and including noise) comprised approximately 16 400 peaks, and for (B) the output from Themis using all replicates, where the entire peak list comprised approximately 2260 peaks.

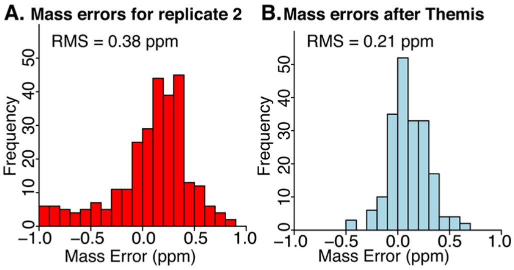


Figure 7. Histogram showing mass errors associated with assignments for the positive ESI mode NIST data for the NS class: (A) for one replicate and (B) after processing with Themis.

root-mean-square values of 0.38 ppm and 0.21 ppm, respectively. The unprocessed replicate displays larger mass errors than data resulting from processing with Themis, as also illustrated by the false positives in Figure 6A.

■ CONCLUSION

Themis capitalizes on the availability of replicated measurements to generate a single, reliable peak list, while avoiding the a priori discarding of low-magnitude peaks that typically occurs when signal-to-noise thresholds are applied. At a practical level, the user's workflow is simplified by performing downstream data analysis on a single data set produced by Themis, instead of working with replicates individually and comparing results at the end. Furthermore, the preprocessing actually led to improved assignment of low-magnitude contributions. Data set sizes and the demand for more reliable, replicated data will increase alongside technological advances in experimental methods. There is an accompanying need to simplify data sets and handle greater numbers of mass spectra. Themis currently performs its tasks within a few minutes and removes the majority of the noise, but there is scope for improvement. For instance, one could incorporate into the analysis peak shape information, such as the full width at half-maximum or some chemical prior information, to further refine the output reference peak list. In this work it has been found that it is simplistic to use a single parameter threshold, such as S/N ratio, to separate noise from valid peaks, and using *m*/*z* in combination with magnitude is a more promising approach. While the application of Themis has been demonstrated for petroleum, it is expected to also be useful for other complex samples. It is intended that Themis will be included in a workflow alongside specialized software for the analysis of different complex mixtures. The anticipated benefits include faster downstream data analysis, fewer false positives, fewer genuine peaks discarded, and hence ultimately an increased confidence in the results of the analysis, which is vital when decision-making may be based on the findings.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b02345.

> Eight figures showing quantile−quantile plots before and after quantile normalization, automated detection of outliers, enlargement of low S/N region for NIST light sour crude oil, stacked bar plots of carbon number and DBE distributions, histograms showing mass errors associated with assignments, and MDS 2D plot before and after Themis (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*(R.G.) E-mail [redacted].
*(M.P.B.) E-mail [redacted]; phone [redacted].

ORCID

Mark P. Barrow: 0000-0002-6474-5357

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

R.G. thanks EPSRC for a Ph.D. studentship through the EPSRC Centre for Doctoral Training in Molecular Analytical Science, Grant EP/L015307/1. D.R. was partially supported by Ramon y Cajal Fellowship RYC-2015-18544 from Ministerio de Economia y Competitividad (Government of Spain). R.G. also thanks David Stranz (Sierra Analytics) for his valuable contributions.

■ REFERENCES

(1) Comisarow, M. B.; Marshall, A. G. *Chem. Phys. Lett.* **1974**, *25*, 282−283.

(2) Comisarow, M. B.; Marshall, A. G. *Can. J. Chem.* **1974**, *52*, 1997−1999.

(3) Comisarow, M. B.; Marshall, A. G. *Chem. Phys. Lett.* **1974**, *26*, 489−490.

(4) Amster, I. J. *J. Mass Spectrom.* **1996**, *31*, 1325−1337.

(5) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. *Mass Spectrom. Rev.* **1998**, *17*, 1−35.

(6) Barrow, M. P.; Burkitt, W. I.; Derrick, P. J. *Analyst* **2005**, *130*, 18−28.

(7) Schaub, T. M.; Hendrickson, C. L.; Horning, S.; Quinn, J. P.; Senko, M. W.; Marshall, A. G. *Anal. Chem.* **2008**, *80*, 3985−3990.

(8) Qian, K.; Rodgers, R. P.; Hendrickson, C. L.; Emmett, M. R.; Marshall, A. G. *Energy Fuels* **2001**, *15*, 492−498.

(9) Barrow, M. P.; McDonnell, L. A.; Feng, X.; Walker, J.; Derrick, P. J. *Anal. Chem.* **2003**, *75*, 860−866.

(10) Marshall, A. G.; Rodgers, R. P. *Acc. Chem. Res.* **2004**, *37*, 53−59.

(11) Rodgers, R. P.; Schaub, T. M.; Marshall, A. G. *Anal. Chem.* **2005**, *77*, 20A−27A.

(12) Marshall, A. G.; Rodgers, R. P. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 18090−18095.

(13) Hsu, C. S.; Hendrickson, C. L.; Rodgers, R. P.; McKenna, A. M.; Marshall, A. G. *J. Mass Spectrom.* **2011**, *46*, 337−343.

(14) Griffiths, M. T.; Da Campo, R.; O'Connor, P. B.; Barrow, M. P. *Anal. Chem.* **2014**, *86*, 527−534.

(15) Rodgers, R. P.; McKenna, A. M. *Anal. Chem.* **2011**, *83*, 4665−4687.

(16) Barrow, M. P.; Witt, M.; Headley, J. V.; Peru, K. M. *Anal. Chem.* **2010**, *82*, 3727−3735.

(17) Headley, J. V.; Barrow, M. P.; Peru, K. M.; Fahlman, B.; Frank, R. A.; Bickerton, G.; McMaster, M. E.; Parrott, J.; Hewitt, L. M. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 1899−1909.

(18) Barrow, M. P.; Peru, K. M.; Headley, J. V. *Anal. Chem.* **2014**, *86*, 8281−8288.

(19) Headley, J. V.; Peru, K. M.; Barrow, M. P. *Mass Spectrom. Rev.* **2016**, *35*, 311−328.

(20) Zhurov, K. O.; Kozhinov, A. N.; Tsybin, Y. O. *Energy Fuels* **2013**, *27*, 2974−2983.

(21) Headley, J. V.; Peru, K. M.; Janfada, A.; Fahlman, B.; Gu, C.; Hassan, S. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 459−462.

(22) Marshall, A. G.; Hendrickson, C. L. *Annu. Rev. Anal. Chem.* **2008**, *1*, 579−599.

(23) Pomerantz, A. E.; Mullins, O. C.; Paul, G.; Ruzicka, J.; Sanders, M. *Energy Fuels* **2011**, *25*, 3077−3082.

(24) Smith, E. A.; Park, S.; Klein, A. T.; Lee, Y. J. *Energy Fuels* **2012**, *26*, 3796−3802.

(25) Dunning, H. N.; Moore, J. W.; Bieber, H.; Williams, R. B. *J. Chem. Eng. Data* **1960**, *5*, 546−549.

(26) Baker, E. W.; Yen, T. F.; Dickie, J. P.; Rhodes, R. E.; Clark, L. F. *J. Am. Chem. Soc.* **1967**, *89*, 3631−3639.

(27) Barrow, M. P. *Biofuels* **2010**, *1*, 651−655.

(28) Cho, Y.; Ahmed, A.; Islam, A.; Kim, S. *Mass Spectrom. Rev.* **2015**, *34*, 248−263.

(29) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 10313−10317.

(30) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 320−332.

(31) Kaur, P.; O'Connor, P. B. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 459−468.

(32) Du, P.; Kibbe, W. A.; Lin, S. M. *Bioinformatics* **2006**, *22*, 2059−2065.

(33) Mantini, D.; Petrucci, F.; Pieragostino, D.; Del Boccio, P.; Di Nicola, M.; Di Ilio, C.; Federici, G.; Sacchetta, P.; Comani, S.; Urbani, A. *BMC Bioinf.* **2007**, *8*, 101.

(34) Meuleman, W.; Engwegen, J. Y. M. N.; Gast, M.-C. W.; Wessels, L. F. a.; Reinders, M. J. T. *BMC Bioinf.* **2009**, *10* (Suppl 1), S51.

(35) Hur, M.; Oh, H.-B.; Kim, S.-H. *Bull. Korean Chem. Soc.* **2009**, *30*, 2665−2668.

(36) Zhurov, K. O.; Kozhinov, A. N.; Fornelli, L.; Tsybin, Y. O. *Anal. Chem.* **2014**, *86*, 3308−3316.

(37) Kilgour, D. P. A.; Hughes, S.; Kilgour, S. L.; Mackay, C. L.; Palmblad, M.; Tran, B. Q.; Goo, Y. A.; Ernst, R. K.; Clarke, D. J.; Goodlett, D. R. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 253−262.

(38) Patterson, S. D. *Nat. Biotechnol.* **2003**, *21*, 221−2.

(39) Chen, L.; Yap, Y. L. *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 46−54.

(40) Johnson, K. L.; Mason, C. J.; Muddiman, D. C.; Eckel, J. E. *Anal. Chem.* **2004**, *76*, 5097−5103.

(41) McIlwain, S.; Page, D.; Huttlin, E. L.; Sussman, M. R. *Bioinformatics* **2007**, *23*, i328.

(42) Park, K.; Yoon, J. Y.; Lee, S.; Paek, E.; Park, H.; Jung, H. J.; Lee, S. W. *Anal. Chem.* **2008**, *80*, 7294−7303.

(43) Lopez-Fernandez, H.; Santos, H. M.; Capelo, J. L.; Fdez-Riverola, F.; Glez-Peña, D.; Reboiro-Jato, M. *BMC Bioinf.* **2015**, *16*, 1−12.

(44) Mass-Up - mass spectrometry utility for proteomics; http://sing.ei.uvigo.es/mass-up/; Accessed April 12, 2017.

(45) Araújo, J. E.; Santos, T.; Jorge, S.; Pereira, T. M.; Reboiro-Jato, M.; Pavón, R.; Magriço, R.; Teixeira-Costa, F.; Ramos, A.; Santos, H. M. *Anal. Methods* **2015**, *7*, 7467−7473.

(46) Araújo, J. E.; Jorge, S.; Magriço, R.; Costa, T. E.; Ramos, A.; Reboiro-Jato, M.; Fdez-Riverola, F.; Lodeiro, C.; Capelo, J. L.; Santos, H. M. *Talanta* **2016**, *152*, 364−370.

(47) Santos, T.; Capelo, J. L.; Santos, H. M.; Oliveira, I.; Marinho, C.; Gonçalves, A.; Araújo, J. E.; Poeta, P.; Igrejas, G. *J. Proteomics* **2015**, *127*, 321−331.

(48) Klitzke, C. F.; Corilo, Y. E.; Siek, K.; Binkley, J.; Patrick, J.; Eberlin, M. N. *Energy Fuels* **2012**, *26*, 5787−5794.

(49) Purcell, J. M.; Merdrignac, I.; Rodgers, R. P.; Marshall, A. G.; Gauthier, T.; Guibard, I. *Energy Fuels* **2010**, *24*, 2257−2265.

(50) Xian, F.; Hendrickson, C. L.; Blakney, G. T.; Beu, S. C.; Marshall, A. G. *Anal. Chem.* **2010**, *82*, 8807−8812.

(51) Savory, J. J.; Kaiser, N. K.; McKenna, A. M.; Xian, F.; Blakney, G. T.; Rodgers, R. P.; Hendrickson, C. L.; Marshall, A. G. *Anal. Chem.* **2011**, *83*, 1732−1736.

(52) Hur, M.; Yeo, I.; Park, E.; Kim, Y. H.; Yoo, J.; Kim, E.; No, M. H.; Koh, J.; Kim, S. *Anal. Chem.* **2010**, *82*, 211−218.

(53) Hur, M.; Yeo, I.; Kim, E.; No, M.-h.; Koh, J.; Cho, Y. J.; Lee, J. W.; Kim, S. *Energy Fuels* **2010**, *24*, 5524−5532.

(54) Payne, T. G.; Southam, A. D.; Arvanitis, T. N.; Viant, M. R. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1087−1095.

(55) Pruski, P.; MacIntyre, D. A.; Lewis, H. V.; Inglese, P.; Correia, G. D. S.; Hansel, T. T.; Bennett, P. R.; Holmes, E.; Takats, Z. *Anal. Chem.* **2017**, *89*, 1540−1550.

(56) Hampel, F. R. *J. Am. Stat. Assoc.* **1974**, *69*, 383−393.

(57) Huber, P. J. *Robust statistics*; John Wiley, New York, 1981.

(58) Rousseeuw, P. J.; Croux, C. *J. Am. Stat. Assoc.* **1993**, *88*, 1273−1283.

(59) Amaratunga, D.; Cabrera, J. *J. Am. Stat. Assoc.* **2001**, *96*, 1161−1170.

(60) Bolstad, B. M.; Irizarry, R. A.; Astrand, M.; Speed, T. P. *Bioinformatics* **2003**, *19*, 185−193.

(61) Dempster, A.; Laird, N.; Rubin, D. B. *J. R. Stat. Soc., Ser. B (Methodol.)* **1977**, *39*, 1−38 (http://links.jstor.org/sici?sici=0035-9246%281977%2939%3A1%3C1%3AMLFIDV%3E2.0.CO%3B2-Z).

(62) Fraley, C.; Raftery, A. E. *J. Am. Stat. Assoc.* **2002**, *97*, 611−631.

(63) Fraley, C.; Raftery, A. E.; Murphy, T. B.; Scrucca, L. *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*; Technological Report 597, University of Washington, 2012; https://www.stat.washington.edu/research/reports/2012/tr597.pdf

(64) Smyth, G. K.; Speed, T. *Methods* **2003**, *31*, 265−273.

(65) Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K. *Nucleic Acids Res.* **2015**, *43*, e47.

(66) Da Campo, R.; Barrow, M. P.; Shepherd, A. G.; Salisbury, M.; Derrick, P. J. *Energy Fuels* **2009**, *23*, 5544−5549.

(67) Newton, R.; Wernisch, L. *R News* **2007**, *7*, 32−35 (http://sysbio.mrc-bsu.cam.ac.uk/Rwui/tutorial/Rwui_Rnews_final.pdf).

Supporting information for:

Themis: Batch Pre-processing for Ultrahigh Resolution Mass Spectra of Complex Mixtures

Remy Gavard,[*,†] David Rossell,[‡,¶] Simon E.F. Spencer,[‡] and Mark P. Barrow[*,§]

†*MAS CDT, 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*

§*Department of Chemistry, University of Warwick, Coventry, CV4 7AL, United Kingdom*

E-mail: [redacted]; [redacted]

Phone: [redacted]. Fax: [redacted]

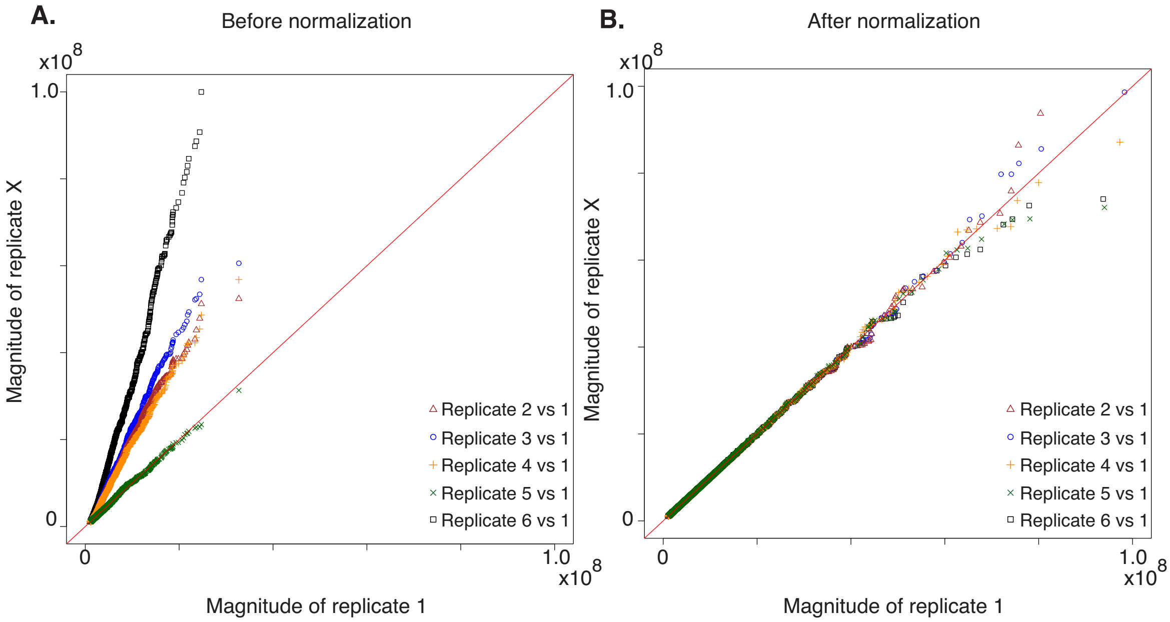


Figure S1: Quantile-quantile plots of the magnitudes of the six replicates before (**A**) and after (**B**) quantile normalization for the South American crude oil Sample B, negative-ion ESI

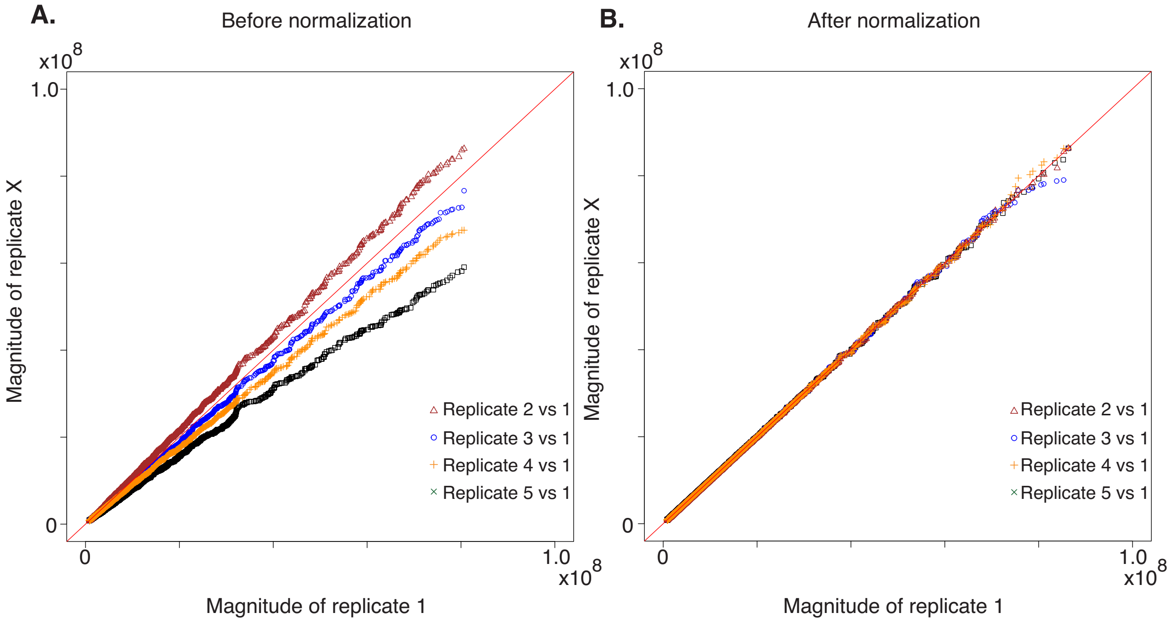


Figure S2: Quantile-quantile plots of the magnitudes of the five replicates before (**A**) and after (**B**) quantile normalization for the South American crude oil Sample B, positive-ion ESI

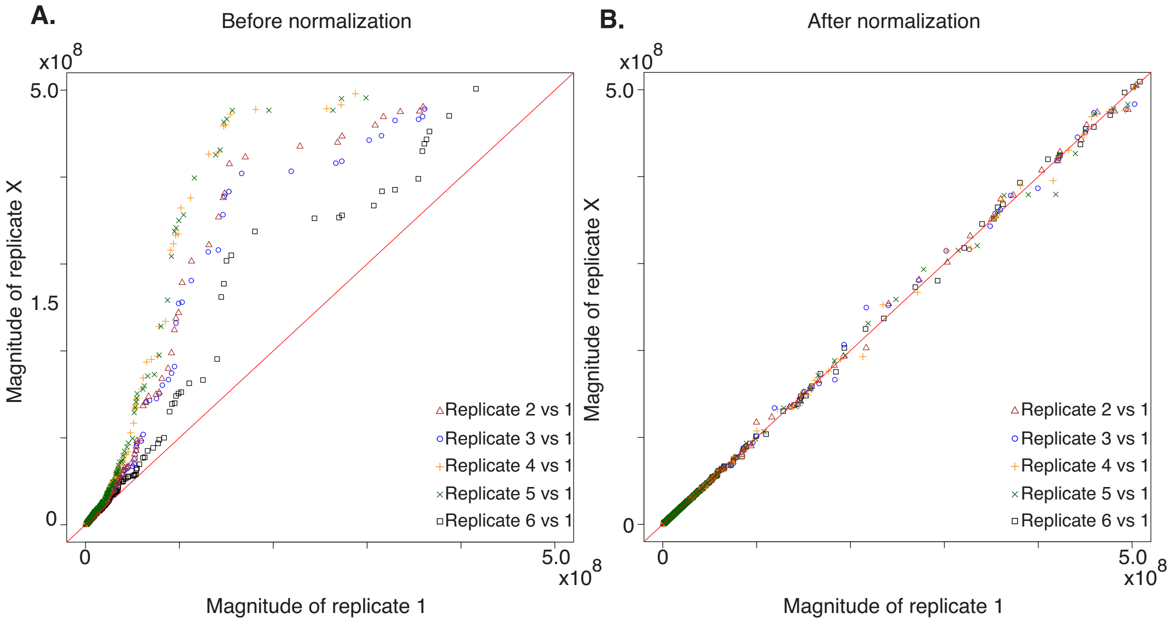


Figure S3: Quantile-quantile plots of the magnitudes of the six replicates before (**A**) and after (**B**) quantile normalization for the naphthenic acid Sample C, negative-ion ESI

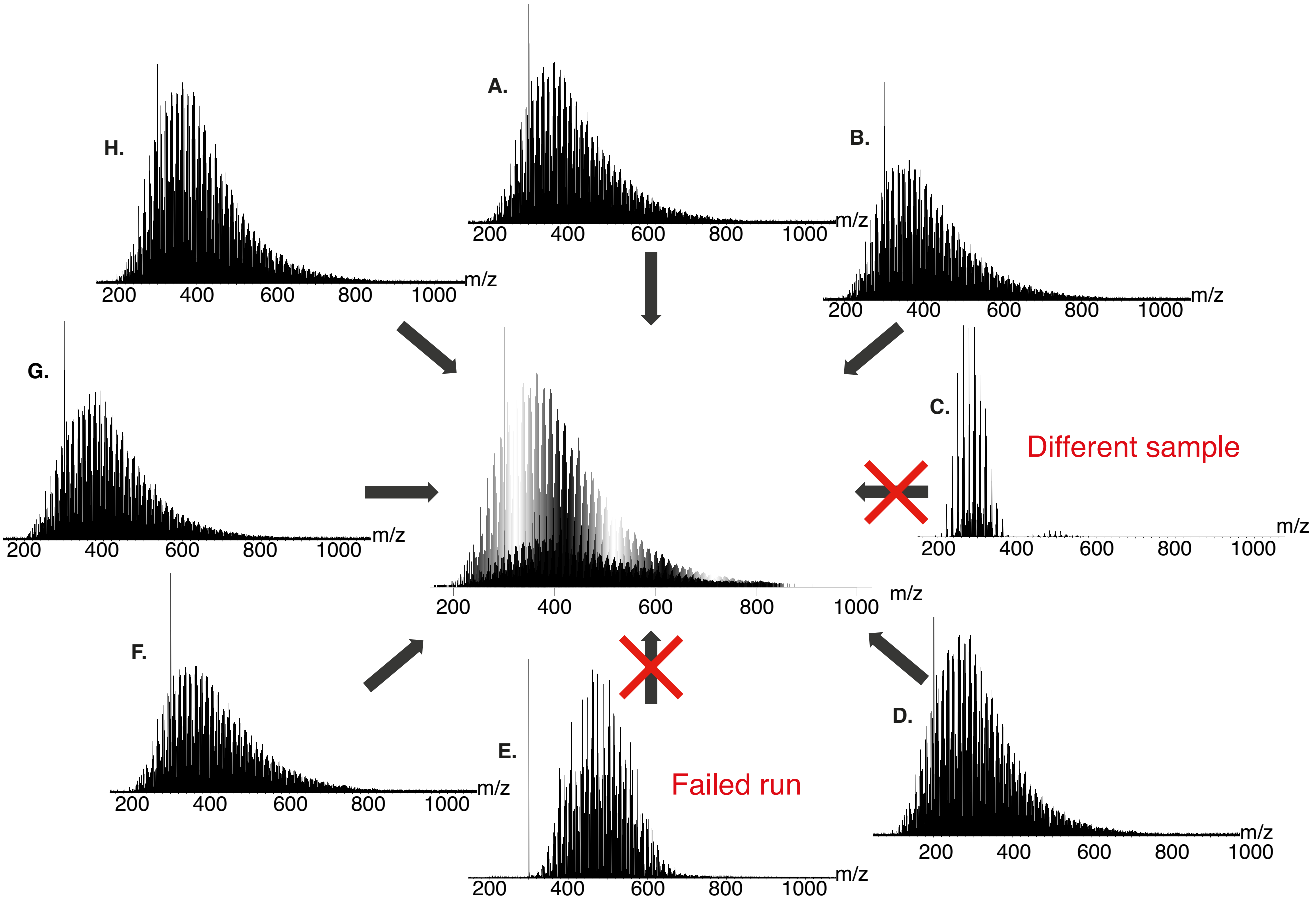


Figure S4: Automated detection of outliers and use of a series of repeat measurements to produce an averaged data set for characterization. Mass spectra A, B, D, F, G, H are replicates, mass spectrum C is for a different sample, and mass spectrum E is a failed run (using same sample as for mass spectra A, B, D, F, G, and H).

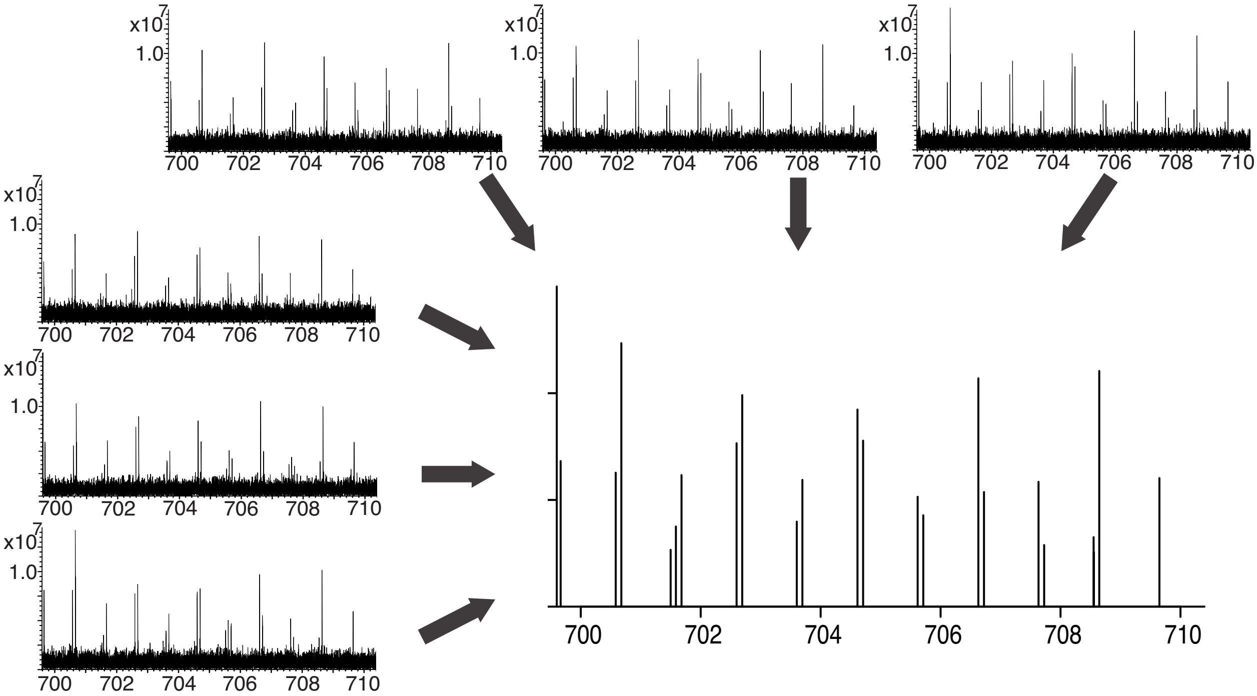


Figure S5: Enlargement of the low S/N region between m/z 700 and 710 for the NIST light sour crude oil, showing the 6 replicates and the processed dataset. Notice the presence of 3 peaks at m/z 701 and 2 peaks at m/z 703 in the processed dataset; this is due to the increase information from use of replicates, whilst it would be difficult to confidently pick all these within a single mass spectrum.

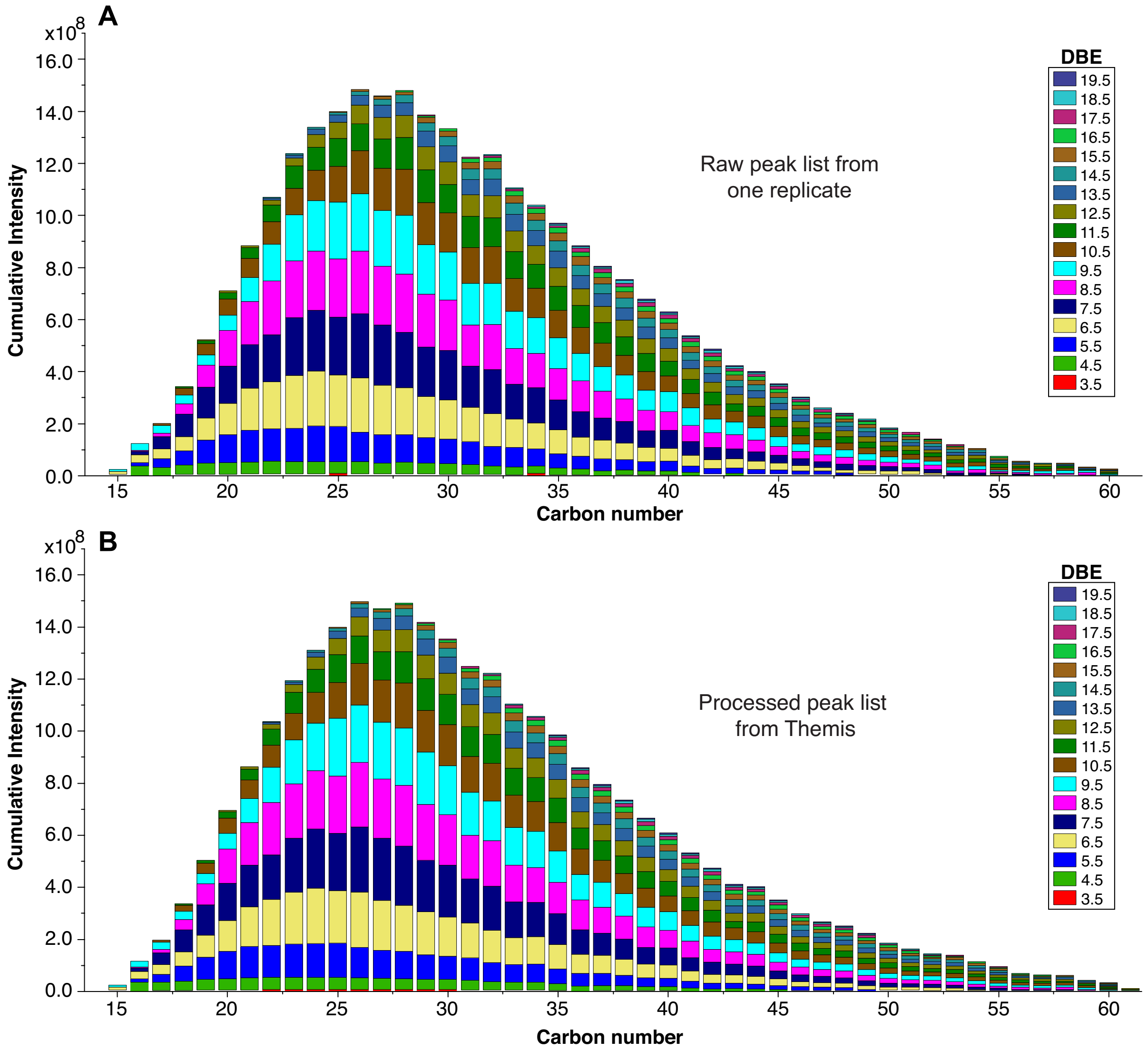


Figure S6: Stacked bar plot the carbon number and DBE distributions for the N_1 class for the NIST light sour crude oil sample. The results of the data analyses are shown for: (**A**) a single replicate, where the total peak list (all classes and including noise) comprised approximately 16,400 peaks and for (**B**) the output from Themis using all replicates, where the entire peak list comprised approximately 2,260 peaks.

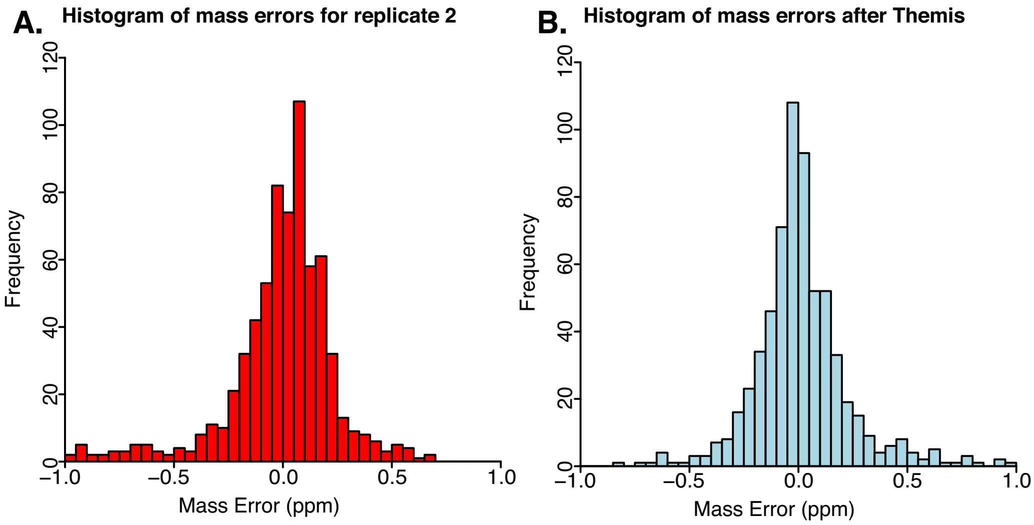


Figure S7: Histograms showing mass errors associated with assignments for the positive ESI mode NIST data for the N_1 class for one replicate (**A**) with an rms 0.24 and after processing with Themis (**B**) with an rms 0.22

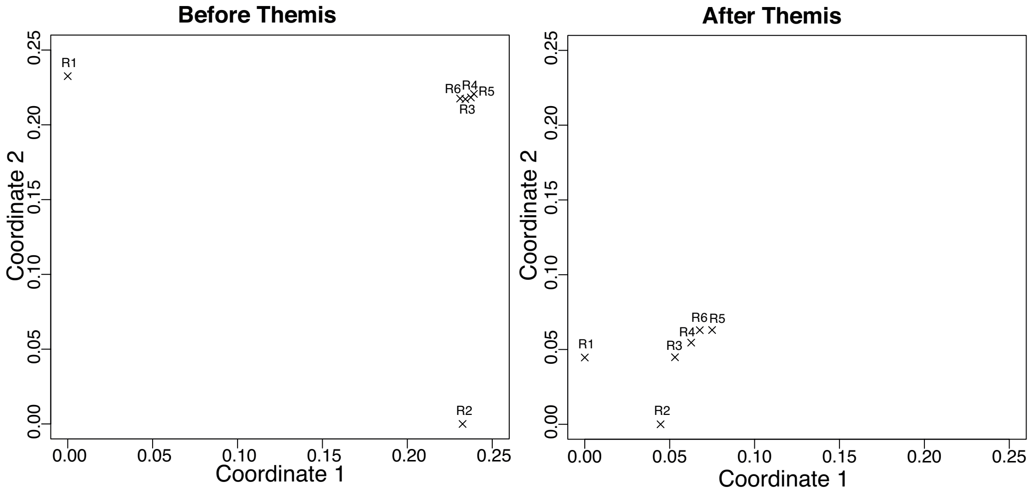


Figure S8: Multidimensional Scaling (MDS) two-dimensional plot based on the Spearman correlation between magnitudes for each pair of data sets, before and after Themis. Data points that are closer together have a stronger correlation. Prior to processing with Themis, replicates R1 and R2 appear as outliers, relative to the remaining replicates. After processing, these replicates are now much closer, indicating that Themis reduced systematic biases across samples.

Chapter 3

Repeatability, signal-to-noise ratio, mass error and molecular assignments in petroleomics

3.1 Context

In this chapter, the objective is to get a better estimate of the quantity of molecular assignments of non-repeatable peaks in petroleum-related samples. In order to get an understanding of the molecular composition of petroleum related samples, it is crucial to assign molecular formulae to the peaks detected. The methods to process the signal from Fourier transform mass spectrometry have improved over time, leading to the identification of more peaks and with greater accuracy. The tools necessary to perform the molecular assignments and reproducibility of the peaks have not been investigated. Using Themis to separate non-reproducible peaks from the reproducible ones in three types of sample, it was observed that using only reproducible peaks lead to better molecular assignments and lower root mean square mass error. The results showed that between 15 to 26% of the peaks assigned in a single replicate can be considered as non-reproducible when compared to peaks present in all replicates, depending on the S/N threshold used for the peak-picking.

This chapter will be submitted as an article to a peer-reviewed journal. The idea to study the quality of molecular assignments in complex mixtures, similar to what has been done for proteomics, came from Peter O'Connor. The data for this study was acquired by the author under the supervision of Diana Catalina Palacio Lozano. Supervisors Mark P. Barrow, Simon E. F. Spencer and David Rossell provided guidance and advice for the design of the experiment, the data analysis and the manuscript writing. The adaptation of Themis into a Shiny [142] interface was performed by the author. The manuscript was written by the author.

3.2 Publication

repeatability, signal-to-noise ratio, mass error and molecular assignments in petroleomics

Remy Gavard,[†] Diana Catalina Palacio Lozano,[‡] Hugh E. Jones,[†] 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*

[§]*Department of Economics & Business, Universitat Pompeu Fabra, Barcelona, 08005, Spain*

E-mail: [redacted]

Phone: [redacted]

Abstract

The ultra-high resolution of Fourier transform ion cyclotron resonance mass spectrometry (FTICR MS) enabled researchers to resolve the mass-to-charge ratio (m/z) of thousands of molecules within petroleum samples. Accurately assigning those m/z values to a molecular composition is crucial for the characterization of the composition of those complex samples. Over time, regular improvements in the instrumentation, the signal processing and molecular assignment software have led to better characterization but the repeatability of current methods has not been yet investigated in detail. Specifically, a peak found in a sample but not in a replicated experiment could either correspond to a false positive or be a low-intensity peak that is hard to detect. Using the Themis algorithm on peak lists which were extracted using a range of signal-to-noise (S/N) thresholds of three different samples, we were able to develop a better understanding of the assignment process and repeatability. The results showed that when analysing all peaks present in at least 2 out of 5 replicates, more compositional assignments were obtained with a significantly lower RMS mass error than when analysing individual replicates separately. It was estimated that between 15 to 26% of the compositional assignments were not fully repeatable across all replicates, depending on the S/N threshold used for the peak-picking when analysing a single replicate.

Introduction

Complex mixtures such as petroleum and dissolved organic matter (DOM) require analysis by ultra-high resolution mass spectrometry (UHRMS).[1,2] Fourier transform ion cyclotron resonance mass spectrometry (FTICR MS)[3–7] is currently the state-of-the-art technique in terms of ultrahigh resolving power and mass accuracy.[8] While light and medium petroleum fractions have been successfully analysed using high-field Orbitrap mass spectrometers, FTICR MS remains unrivalled in terms of performance.[9–13]

A deep understanding of the composition of crude oils and derived products is essential to improve the recovery of oil and its processing.[14] The term "petroleomics" has been used over the last few years to describe the characterization of petroleum and its products by mass spectrometry.[15–21] The challenges posed by the processing of crude oils continue to grow in importance as the sources of light, easy to process, crude oil is getting sparser and lead to a higher reliance on heavy crudes.[22] Characterizing the composition of crude oils also plays an important role in understanding their toxicity and their impact on the environment.[23,24]

To obtain greater insight into complex mixtures, several scientific domains have been the subject of recent advances. Increases in the magnetic field strength directly increase the resolving power (RP) achievable per second of acquisition time. A method called OCULAR was recently developed using FTICR MS to achieve a world record of unique compositional assignments within a single sample.[25] A 21 tesla FTICR MS was recently used to assign more molecules and also increase the accuracy of the data.[26] The transformation from a magnitude mode spectrum to an absorption mode spectrum has been demonstrated to help improve the resolving power by up to two fold and increase the mass accuracy.[27] Peak picking algorithms are also at the centre attention with new algorithms aiming to reduce the number of false assignments.[28] Finally, recalibration methods constitute a crucial step towards reliable assignments and several methods have been described.[29,30]

To assign a molecular composition to each measured peak from its m/z, researchers can currently rely on two commercial software offerings: Composer (Sierra Analytics, Modesto, CA, U.S.A.)[9,32] and PetroOrg (Florida State University, Tallahassee, FL, U.S.A.).[33] Some research groups are also relying on in-house algorithms.[34]

Molecules found inside petroleum and related compounds are commonly characterized by their heteroatom count (N, N_2, O, O_2 etc), their double bond equivalents (DBE), and finally the number of carbon atoms they contain. Within the same heteroatom class and DBE, homologous series of peaks will be separated by the mass of CH_2 (14.015650 Da).[22] This means that when a few peaks belong to the same heteroatom class and DBE, we can predict where the other peaks of that same family will be found.

Many data analysis workflows currently rely on peak-picking algorithms to generate a peak list. Those methods require user-defined parameters, the most prevalent being that the peak height exceeds a selected signal-to-noise ratio (S/N). Setting the S/N threshold too high leads to a conservative list of highly reliable peaks but will miss out on low intensity peaks. A lower S/N will avoid missing low intensity peaks but will include more noise and non-repeatable peaks. If the user has a particular interest in low intensity peaks, a lower S/N threshold will need to be used. In term of probabilities, the lower the S/N threshold, the more likely some of the observed peaks will be due to chance and associated to the experimental background noise.

The Themis algorithm demonstrated that by using replicates from the same sample, it was also possible to discriminate between repeatable and non-repeatable peaks.[35] Reliable peaks tended to be consistently observed across replicates, whereas spurious non-repeatable peaks were not. Thus, the use of Themis gave the ability to set a reduced S/N ratio without compromising on the reliability, keeping the root mean square mass error low and reducing the amount of non-repeatable compositional assignments. The results also showed an improvement in the profile of heteroatom classes distribution. The work also raised questions about the drawbacks of current molecular assignment methods and the rate of non-repeatable assignments associated to standard peak identification algorithms.

The goal of this paper is to address an important need in studying and correcting potential false positive assignments arising from assigning a molecular composition to peaks not consistently observed across replicates. To this end, we developed a new version of Themis which was recoded to handle larger datasets, increase speed and enable interactivity with the R package, Shiny.[36]

Three samples were used to demonstrate a range of different levels of sample complexity. For each sample, we acquired multiple replicates to enable the study of between-sample consistency. The results from extracting peak lists using different S/N was compared to the peak list obtained by processing several replicate peak lists with Themis. We then studied the differences in assignments between a regular workflow based on a single spectrum and the assignments produced with Themis designed to only keep repeatable peaks.

Methodology

Sample preparation

Three samples were selected to reflect different levels of crude oil complexity. Sample A was a crude oil from the Middle-East and was selected to act as a relatively simple sample. It was diluted to a concentration of 0.1 mg/ml with equal volumes of toluene and propanol. Sample B was a bio-oil sample and served as a medium complexity sample. It was diluted to a concentration of 0.05 mg/ml with equal volumes of toluene and methanol. Sample C was a crude oil sample from central America and was selected for its high complexity. It was diluted to a concentration of 0.05 mg/ml with 20% toluene and 80% propanol. The solvent were picked based on previous experience to achieve an optimal homogeneous dilution while the concentration was selected based on previous trials and colour of the sample. The more complex the sample, the more dilution is necessary for optimal signal during the acquisition.

Instrumentation

Mass spectra were acquired using an APPI II atmospheric pressure photoionization (APPI) source, coupled to a 12 T solariX FTICR mass spectrometer (Bruker Daltonik GmbH, Bremen, Germany). Broadband mass spectra were acquired, where a single zero fill and Sine-Bell apodization were applied before usage of a fast Fourier transform. Ten replicates of each mass spectrum were acquired with the aim to retain the five most similar to apply the Themis processing. Sample A was acquired using the accumulation of 200 scans, sample B, 100 scans and sample C, 300 scans.

Signal processing

The spectra were exported from solariXcontrol to FTMS Processing 2.1.0 software which was used offline to produce an absorption-mode mass spectra and an asymmetric apodization (`Kilgour`)[27] function was applied. DataAnalysis 4.2 was then used to extract peak information using the peak finder "FTMS" method. The peak list of each mass spectrum was exported using S/N values of 2, 3, 4, 5 and 6.

Data processing

The replicated peak list for each S/N of each sample was processed using a modified version of the Themis algorithm.[35] As a novel contribution, Themis was implemented within a Shiny interface to enable more interactivity with the users, giving the possibility, for each step, to decide whether to use the automatic settings or manual ones. It is also possible to see how many peaks are isolated in cases where certain peaks are not present in all the replicates, leaving the user the possibility to choose the desired level of replicability. Further, the processing speed was also enhanced to better cope with bigger datasets thanks to tidyverse packages and better use of functional programming. Themis takes replicates mass spectra of a same sample and combines their peak lists, only retaining the peaks present across all replicates. The peak list produced by Themis was then used to perform molecular assignments. For comparison, the peak list from each replicate of each sample was also processed for molecular assignments.

The molecular assignments were performed using Composer 1.5.6 (Sierra Analytics, Modesto, CA, U.S.A.). The same settings were used across each sample to minimize the sources of variation.

The original peak lists along with the corresponding molecular assignments were parsed using R and metrics were computed such as the number of peaks for each replicate and the S/N threshold.

Results and discussion

Throughout the study, the 5 most similar replicates, based on the average molecular weight, for each sample were retained out of the 10 acquired to ensure redundancy in case of experimental fluctuation such as signal loss. We considered that 5 was a number of replicates that could be feasibly produced by a user. The improvements in Themis now allowed the user to visualise the average molecular weight for each replicate, the confidence interval and decide whether to exclude the outliers or not. The same peak-picking settings in DataAnalysis, including the S/N threshold, were used to extract the peak list of each replicate for each sample. Using Themis, not only a peak list with only peaks present in all replicates was exported for each S/N, also with all peaks present in 4 out 5 replicates, 3 out 5 replicates and 2 out 5 replicates. The first step to explore the differences in the peak lists was to perform the molecular assignments on each replicate separately and all peak list exported from Themis. We observed differences both in the number of detected peaks and the corresponding molecular assignments. In figures 2, 3, 4, 8, those variations between replicates were highlighted using error bars for the standard deviation (SD). As the peak lists generated by Themis were based on all available replicates, there is no between-replicate variance.

Figure 1 shows the replicability of the peaks across replicates. We first notice that most peaks were found in all 5 replicates in each sample and for every S/N threshold and that the number of peaks increased when the S/N threshold decreased. The number of peaks with a lower level of repeatability increased when the S/N threshold decreased. While the number of peaks present in only present in 2,3, and 4 out of 5 replicates was similar between S/N 3 to 6, the number of peaks was significantly higher at S/N 2. Figure 1 demonstrates that the repeatability of the peaks picked with a low S/N threshold was lower. In consequences molecular assignments performed with a low S/N threshold will lead to a significant proportion of low intensity peaks not being repeatable or fully repeatable between acquisitions.

Figure 2 showed the average number of molecular formulae assigned for each replicates analysed separately (dash line) and the number of molecular formulae assigned for each peak list produced by Themis (full lines), representing different levels of repeatability. The number of molecular formulae assigned increased when including peaks present in at least 2 out of 5 replicates compared to when only including peaks present in all replicates. Comparing the values for the average number of molecular formulae assigned for each replicates analysed separately, Themis can give a higher number of assignments depending of the settings used. This suggest that when investigating peaks difficult to distinguish from the noise, reducing the presence requirement in Themis can help harvest peaks which may or may not be present when looking at a single replicate.

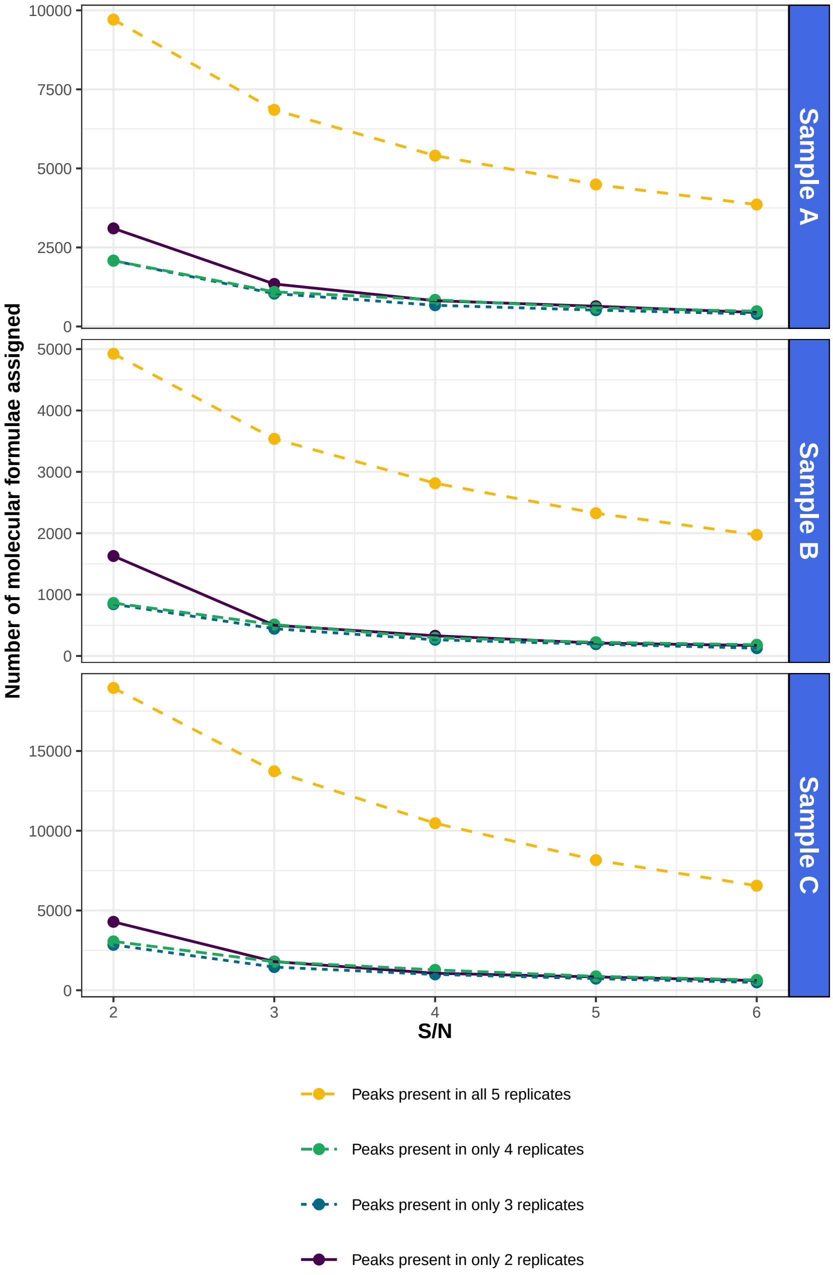


Figure 1: Decomposition of the total number of peaks found in any sample into those that were observed in 2, 3, 4 or 5 replicates, for different S/N thresholds and samples A, B and C.

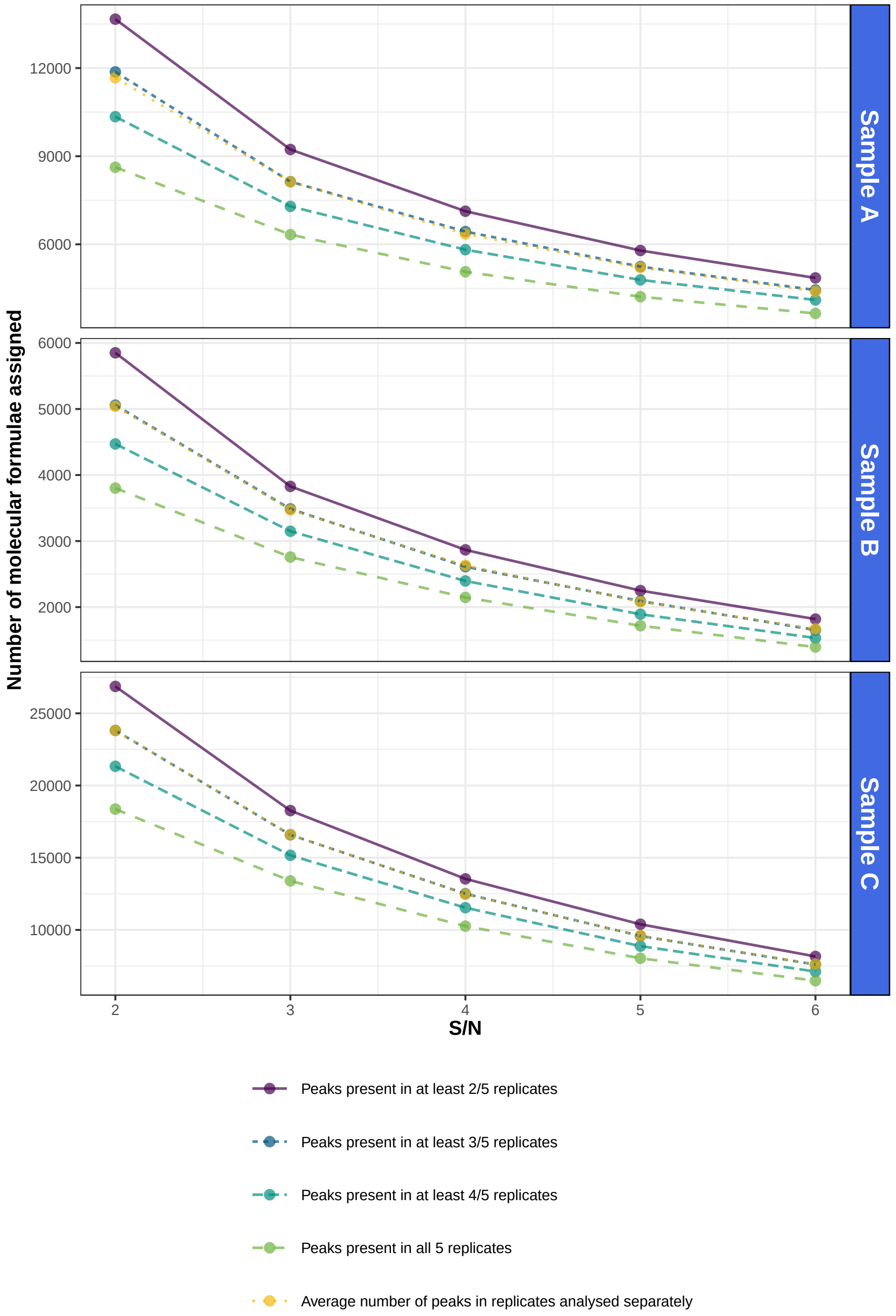


Figure 2: Number of molecular formulae assigned for peaks satisfying the criterion of being present in at least 2 out 5, 3 out 5, 4 out 5 or 5 out 5 replicates, as a function of S/N (full lines) and peaks in replicates analysed separately (dash line).

Naturally, the follow up question concerns the origin of theses changes in the number of molecular formulae assigned compared to replicates analysed separately. Gavard *et al.*[35] previously demonstrated that the decrease in the number of assigned peaks was largely due to the removal of non-repeatable information and generally important information was preserved and the RMS mass error was reduced. Figure 3 showed the root mean square (RMS) mass error of the m/z molecular assignments for all samples, using different repeatability requirements, as a function of S/N threshold.

The RMS mass error decreased when moving up from peaks present in at least 2 out of 5 replicates to only including peaks present in all replicates. There was an expected decrease in the RMS mass error of the molecular assignments as the S/N threshold was increased and replicates analysed separately appear to more affected. The RMS mass error for the assignments was consistently higher when using the average data of individual replicates compared to using Themis.

In Figure 2, we noticed that lower repeatability levels led to similar or increase in the number of molecular assignments compared to replicates analysed separately. Despite allowing peaks with a lower repeatability level, the data processed with Themis demonstrated a consistently lower RMS mass error than the replicates analysed separately.

This suggests that non-repeatable peaks with a high RMS mass error were removed but allowing lower levels of repeatability allowed to observe peaks which may have been consistent in term of m/z but sometimes above and sometimes below the intensity threshold during peak picking. The higher RMS mass error for the replicates analysed separately suggest that a number of non-repeatable peaks were assigned an incorrect molecular composition as they fell within a window of error for molecular assignment.

We then studied where the not fully repeatable peaks were located for each S/N threshold using sample C. For that we looked at the number of molecular formulae assigned per m/z width of 10 depicted in Figure 4.

The results demonstrated that at high S/N threshold the replicates have a majority of

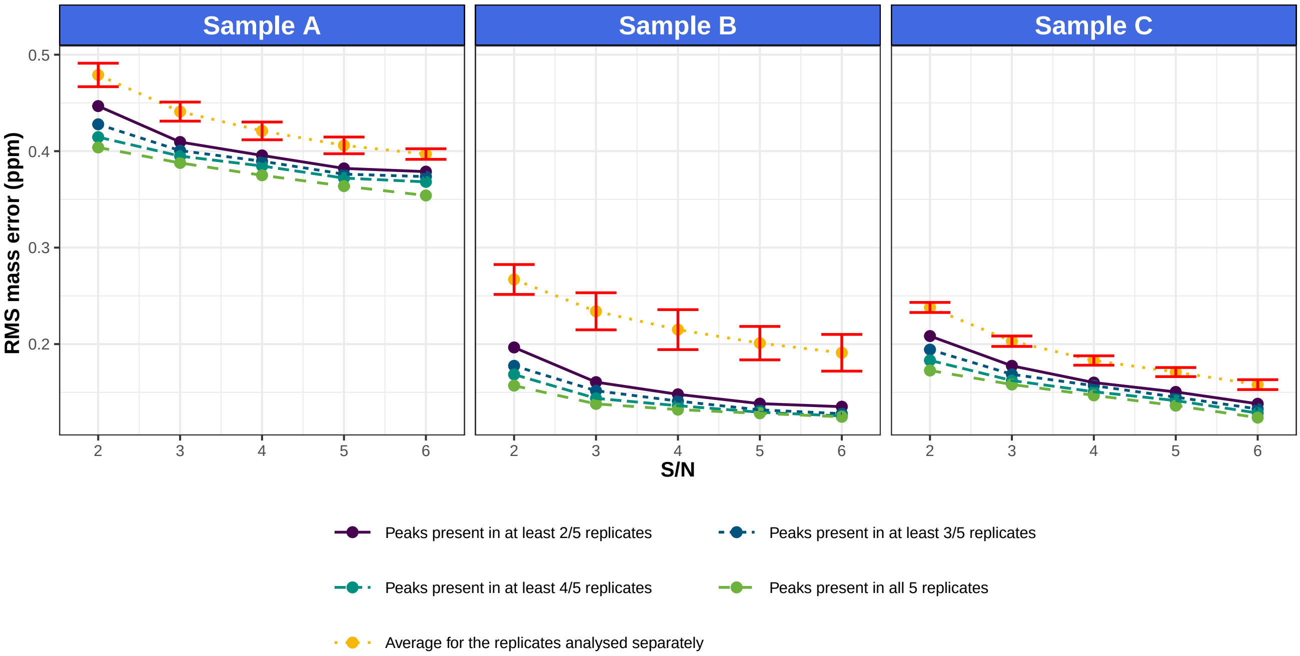


Figure 3: Root mean square (RMS) mass error for each peak list as a function of S/N. The error bars apply to the average for the replicates analysed separately.

repeatable peaks as both the error bar were small and the proximity with the number of peaks present in all 5 replicates. Those findings match what was observed earlier in Figure 1. Going towards low S/N threshold, the error bars for replicates analysed separately increased and there was a clear difference between the number of peaks present in all 5 replicates and those only present in one. The differences were more pronounced in the dense areas suggesting that the presence of non-repeatable peaks was proportional to the peak density across the mass spectrum. The non-repeatable peaks found in the most intense central region were likely to come from the noise baseline. Most non-repeatable peaks found at high m/z were probably due to the combined effect of the decrease in intensity and the decrease of resolving power which diminishes inversely proportional to the m/z. Figures S1 and S2, for respectively samples A and B, demonstrate similar trends.

In addition it was found in Figures S3, S4 and S5 that the variations in RMS mass error of the molecular assignments versus the m/z were strongly dependent of the m/z profile or spectrum so by extension the sample analysed. Sample A showed an improvement evenly

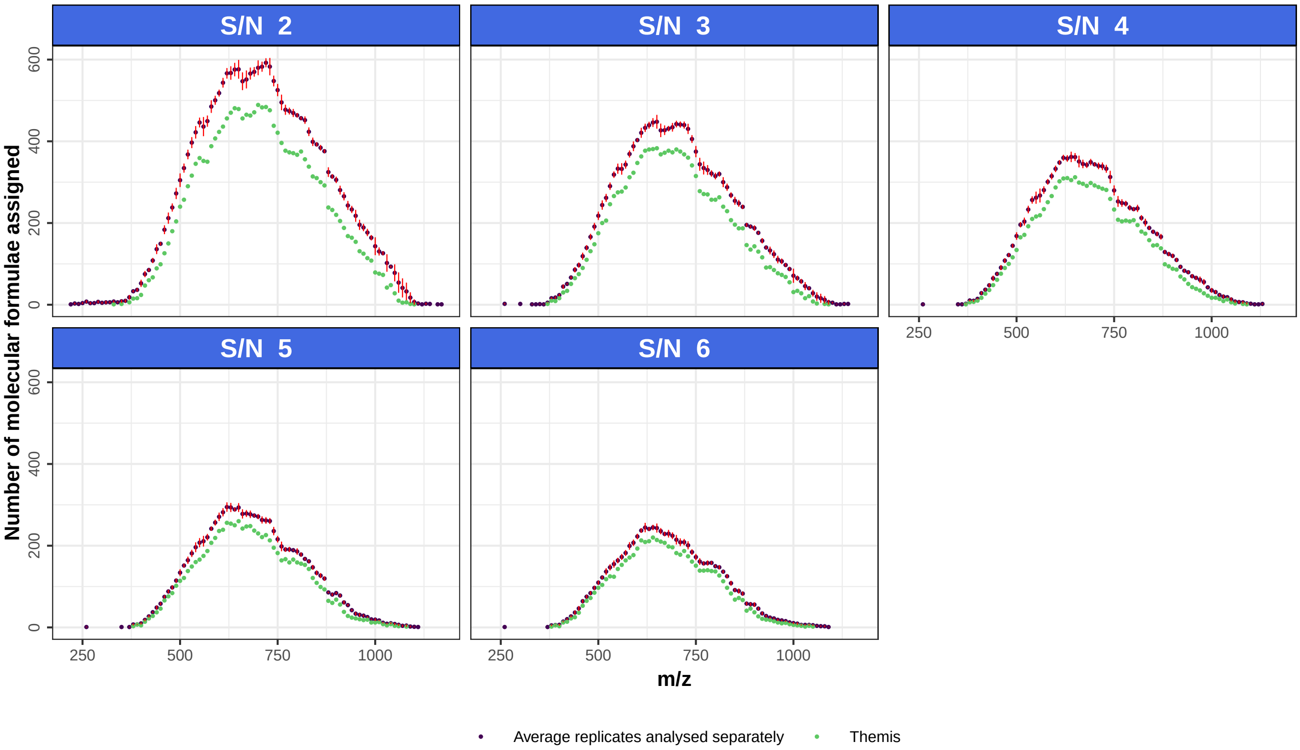


Figure 4: Number of molecular formulae assigned per m/z width of 10 for each S/N for the sample C.

distributed, sample B showed improvement in the high m/z region while for sample C, improvements were observed in both low and high m/z regions. In consequence, depending of characteristics of the mass spectra obtained for each sample, the processing with Themis lead to improvement in different regions.

Figure 5 showed the heteroatom classes distribution at S/N 2 for the peaks present in all replicates and in a single replicate. Some low contribution classes are being discarded while several high contribution classes see their percentage contribution to the total signal increase when using only peaks present in all replicates. The extremely low contributions classes displayed spurious double bond equivalents (DBE) distributions based on non-repeatable peaks which explain their removal. The heteroatom classes distribution for S/N 3, 4, 5 and 6 are depicted in Figures S6, S7, S8 and S9.

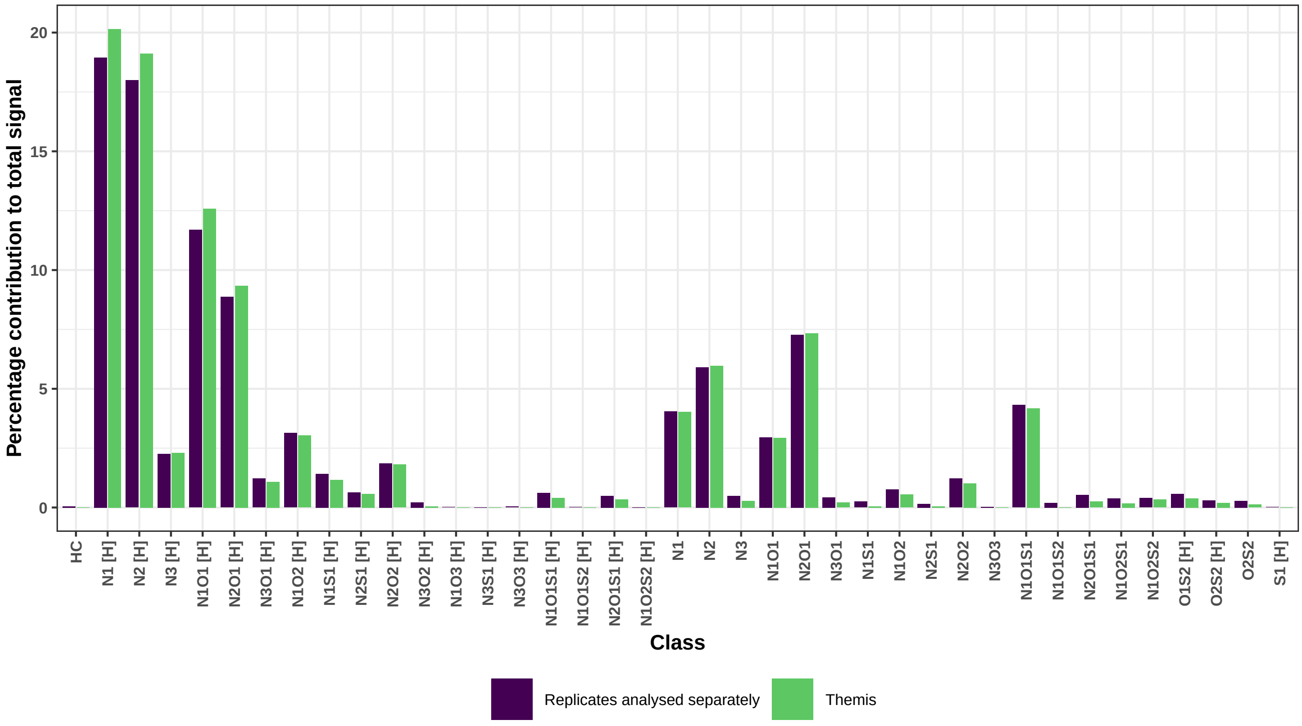


Figure 5: Heteroatom classes distribution for sample C at S/N 2. The most intense classes increased and some of the lowest were removed when only looking at peaks present in all replicates.

Figure 6 showing the double bond equivalents (DBE) vs. carbon number of the N_3[H] heteroatom class for sample C at S/N 2 for both methods showed a clear improvement in the continuity of the DBE series when only retaining the most repeatable peaks.

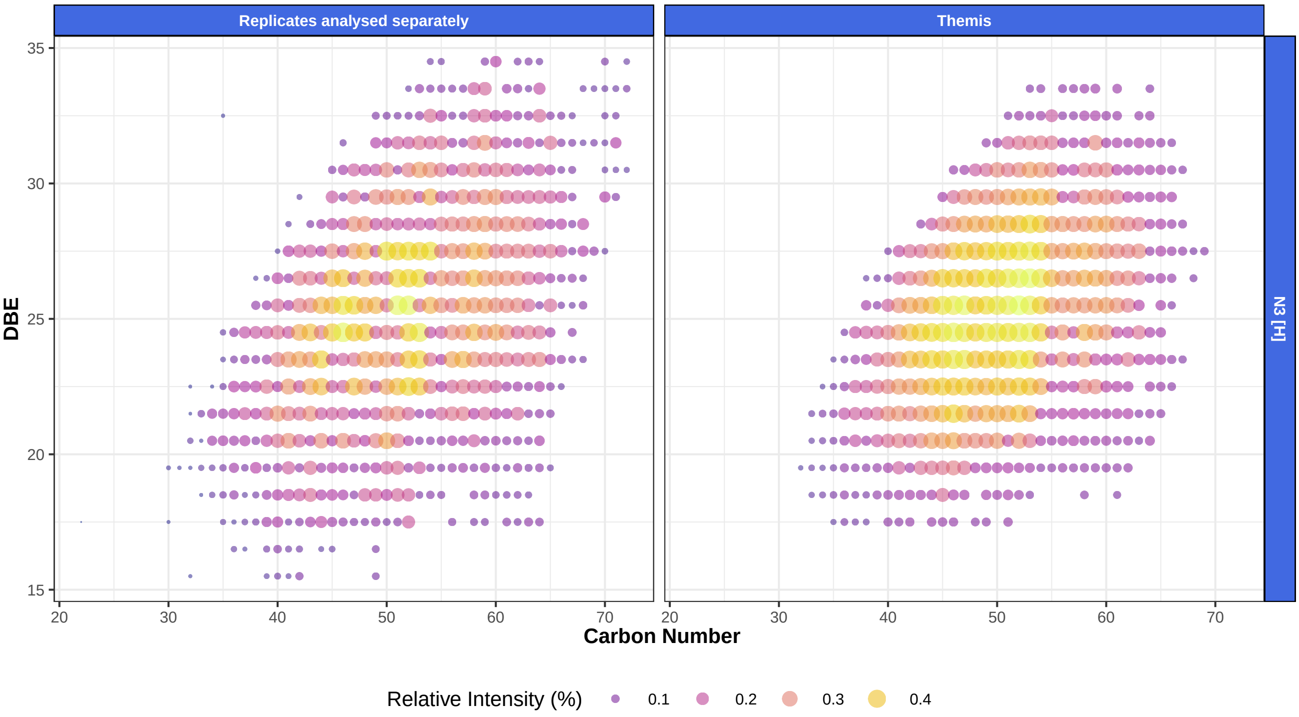


Figure 6: Double bond equivalents (DBE) vs. carbon number of the N_3 [H] heteroatom class for sample C at S/N 2.

Figure 7 showed similar improvements than in Figure 6 but this time looking at the heteroatom classes N_1 and S_3 for sample A. Indeed, for petroleum analysis, gaps within DBE series are often used to distinguish correct and incorrect assignments because of the presence of homologous series.

The previous results demonstrated the non-repeatable peaks which were been assigned a molecular composition had a mass error higher than the rest of the assigned peaks and were located within either or both low intensity regions and high m/z regions, the later being affected by a decrease in resolving power leading to reduced quality data compared to lower m/z regions. Finally, keeping only repeatable peaks led to more consistent series being observed.

Using a low S/N threshold increased the quantity of non-repeatable compositional as-

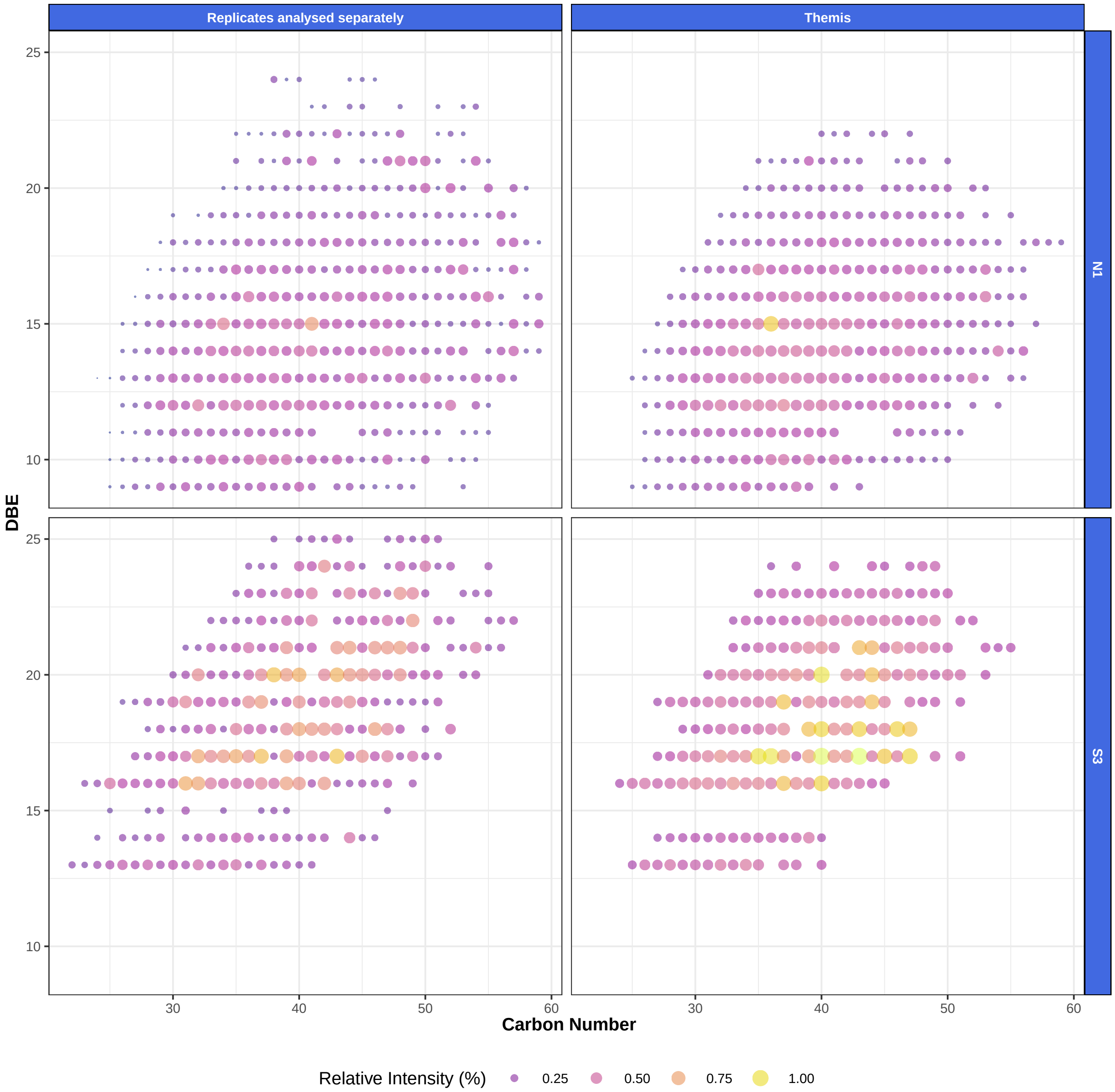


Figure 7: Double bond equivalents (DBE) vs. carbon number of the N_1 and S_3 heteroatom classes for sample A at S/N 2.

signments. We used Themis to calculate the percentage of non-repeatable compositional assignments in a single mass spectrum as a function of the S/N threshold used for each samples. Figure 8 suggests that between 15 and 17% of the compositional assignments at S/N 6 were not fully repeatable, while at S/N 2 that percentage raised to between 23 and 26%.

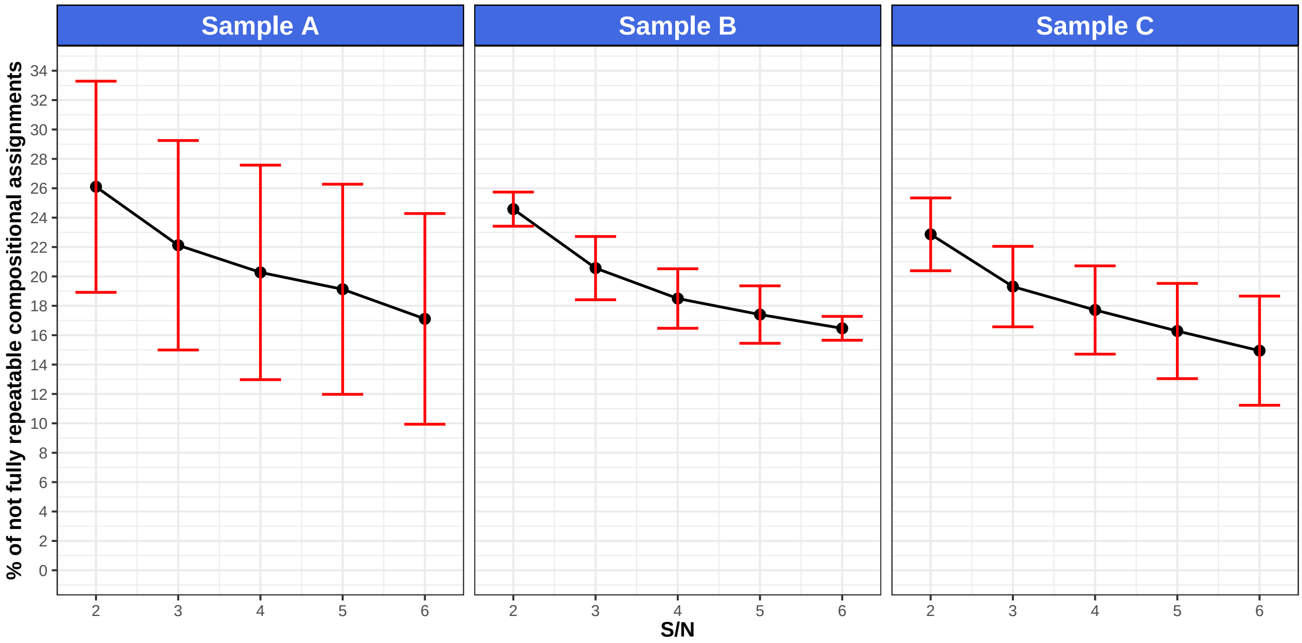


Figure 8: Percentage of compositional assignments not found across all 5 replicates as a function of S/N threshold.

Conclusion

In this study we observed that the number of non-repeatable peaks increases when reducing the S/N threshold. When analysing all peaks present in at least 2 out of 5 replicates, we were able to obtain more peaks assigned and with a significantly lower RMS mass error than when analysing replicates separately. This was due to the recovery of low intensity peaks which may have been missed in other replicates but with a consistent m/z. We demonstrated that while using peaks present in all replicates reduced the number of peaks extracted and assigned, further investigations reveal that the method provided a lower RMS mass error for the molecular assignments while preserving most of the intensity assigned. The number of not fully repeatable peaks distribute proportionally to the peak's density across the m/z range and do not appear to be focused within any specific heteroatom class. The data gathered suggests that when performing molecular assignments, no method is currently error-free or ideal. Keeping the S/N threshold high and the mass error tolerance high for assignments will cause the analyst to miss out on important information. By contrast lowering the S/N threshold and being more tolerant for the assignments will increase the number of non-repeatable compositional assignments. Even using a high S/N threshold will yield about 15 to 17% of not fully repeatable compositional assignments. The percentage rises to between 23 and 26% when reducing the S/N threshold to 2. Those numbers are very conservative as they are calculated by comparison with the peaks present in all replicates. If comparing to the peak list obtained with all peaks present in at least 3 out 5 or at least 4 out of 5 replicates, the percentage drops. Themis can be used to retain only the most repeatable peaks among replicates, leading to more confident molecular assignments or to harvest extra information by gathering peaks only present in a few replicates while still retaining benefits such as a lower RMS mass error for the molecular assignments.

Acknowledgement

R.G. and H.E.J thanks EPSRC for a PhD studentship through the EPSRC Centre for Doctoral Training in Molecular Analytical Science, grant number EP/L015307/1. 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. 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

The supporting Information is available free of charge on the ACS Publications website at:

References

(1) Barrow, M. P.; Burkitt, W. I.; Derrick, P. J. Principles of Fourier transform ion cyclotron resonance mass spectrometry and its application in structural biology. *Analyst* **2005**, *130* (1), 18–28. https://doi.org/10.1039/b403880k.

(2) Niyonsaba, E.; Manheim, J. M.; Yerabolu, R.; Kenttämaa, H. I. Recent Advances in Petroleum Analysis by Mass Spectrometry. *Anal. Chem.* **2019**, *91* (1), 156–177. https://doi.org/10.1021/acs.analchem.8b05258.

(3) Comisarow, M. B.; Marshall, A. G. Fourier transform ion cyclotron resonance spectroscopy. *Chem. Phys. Lett.* **1974**, *25* (2), 282–283. https://doi.org/10.1016/0009-2614(74)89137-2.

(4) Comisarow, M. B.; Marshall, A. G. Selective-phase ion cyclotron resonance spectroscopy. *Can. J. Chem.* **1974**, *52* (4), 1997–1999. https://doi.org/10.1139/v74-288.

(5) Comisarow, M. B.; Marshall, A. G. Frequency-sweep fourier transform ion cyclotron resonance spectroscopy. *Chem. Phys. Lett.* **1974**, *26* (4), 489–490. https://doi.org/10.1016/0009-2614(74)80397-0.

(6) Amster, I. J. Fourier transform mass spectrometry. *J. Mass Spectrom.* **1996**, *31* (12), 1325–1337. https://doi.org/10.1002/(SICI)1096-9888(199612)31:12%3C1325::AID-JMS453%3E3.0.CO;2-W.

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

https://doi.org/10.1002/(SICI)1098-2787(1998)17:1%3C1::AID-MAS1%3E3.0.CO;2-K.

(8) Schaub, T. M.; Hendrickson, C. L.; Horning, S.; Quinn, J. P.; Senko, M. W.; Marshall, A. G. High-performance mass spectrometry: Fourier transform ion cyclotron resonance at 14.5 Tesla. *Anal. Chem.* **2008**, *80* (11), 3985–3990. https://doi.org/10.1021/ac800386h.

(9) Zhurov, K. O.; Kozhinov, A. N.; Tsybin, Y. O. Evaluation of high-field orbitrap fourier transform mass spectrometer for petroleomics. *Energ. Fuel.* **2013**, *27* (6), 2974–2983. https://doi.org/10.1021/ef400203g.

(10) 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* (3), 459–462. https://doi.org/10.1002/rcm.4877.

(11) Marshall, A. G.; Hendrickson, C. L. High-Resolution Mass Spectrometers. *Annu. Rev. Anal. Chem.* **2008**, *1* (1), 579–599. https://doi.org/10.1146/annurev.anchem.1.031207.112945.

(12) Pomerantz, A. E.; Mullins, O. C.; Paul, G.; Ruzicka, J.; Sanders, M. Orbitrap mass spectrometry: A proposal for routine analysis of nonvolatile components of petroleum. *Energ. Fuel.* **2011**, *25* (7), 3077–3082. https://doi.org/10.1021/ef200359n.

(13) Smith, E. A.; Park, S.; Klein, A. T.; Lee, Y. J. Bio-oil analysis using negative electrospray ionization: Comparative study of high-resolution mass spectrometers and phenolic versus sugaric components. *Energ. Fuel.* **2012**, *26* (6), 3796–3802. https://doi.org/10.1021/ef3003558.

(14) Altgelt, K. H.; Boduszynski, M. M. *Composition and Analysis of Heavy Petroleum Fractions*, 1st ed.; CRC Press, 1993.

(15) Qian, K.; Rodgers, R. P.; Hendrickson, C. L.; Emmett, M. R.; Marshall, A. G. Reading chemical fine print: Resolution and identification of 3000 nitrogen-containing aromatic compounds from a single electrospray ionization Fourier transform ion cyclotron resonance mass spectrum of heavy petroleum crude oil. *Energ. Fuel.* **2001**, *15* (2), 492–498.

https://doi.org/10.1021/ef000255y.

(16) Barrow, M. P.; McDonnell, L. A.; Feng, X.; Walker, J.; Derrick, P. J. Determination of the nature of naphthenic acids present in crude oils using nanospray Fourier transform ion cyclotron resonance mass spectrometry: The continued battle against corrosion. *Anal. Chem.* **2003**, *75* (4), 860–866.

(17) Marshall, A. G.; Rodgers, R. P. Petroleomics: The Next Grand Challenge for Chemical Analysis. *Acc. Chem. Res.* **2004**, *37* (1), 53–59. https://doi.org/10.1021/ar020177t.

(18) Rodgers, R. P.; Schaub, T. M.; Marshall, A. G. PETROLEOMICS: MS Returns to Its Roots. *Anal. Chem.* **2005**, *77* (1), 20A–27A.

(19) Marshall, A. G.; Rodgers, R. P. Petroleomics: Chemistry of the underworld. *Proc. Natl. Acad. Sci. U.S.A* **2008**, *105* (47), 18090–18095. https://doi.org/10.1073/pnas.0805069105.

(20) 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* (4), 337–343. https://doi.org/10.1002/jms.1893.

(21) Griffiths, M. T.; Da Campo, R.; O'Connor, P. B.; Barrow, M. P. Throwing light on petroleum: Simulated exposure of crude oil to sunlight and characterization using atmospheric pressure photoionization fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **2014**, *86* (1), 527–534. https://doi.org/10.1021/ac4025335.

(22) Barrow, M. P. Petroleomics: study of the old and the new. *Biofuels* **2010**, *1* (5), 651–655. https://doi.org/10.4155/bfs.10.55.

(23) 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* (16), 8281–8288. https://doi.org/10.1021/ac501710y.

(24) Pan, Y.; Liao, Y.; Shi, Q. Variations of Acidic Compounds in Crude Oil during Simulated Aerobic Biodegradation: Monitored by Semiquantitative Negative-Ion ESI FT-ICR MS. *Energ. Fuel.* **2017**, *31* (2), 1126–1135. https://doi.org/10.1021/acs.energyfuels.

6b02167.

(25) 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* (29), 6966–6978. https://doi.org/10.1039/C9SC02903F.

(26) 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* (3), 2041–2047. https://doi.org/10.1021/acs.analchem.7b04159.

(27) Kilgour, D. P. A.; Van Orden, S. L. Absorption mode Fourier transform mass spectrometry with no baseline correction using a novel asymmetric apodization function. *Rapid Commun. Mass Spectrom.* **2015**, *29* (11), 1009–1018. https://doi.org/10.1002/rcm.7190.

(28) Kilgour, D. P. A.; Hughes, S.; Kilgour, S. L.; Mackay, C. L.; Palmblad, M.; Tran, B. Q.; Goo, Y. A.; Ernst, R. K.; Clarke, D. J.; Goodlett, D. R. Autopiquer - a Robust and Reliable Peak Detection Algorithm for Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (2), 253–262. https://doi.org/10.1007/s13361-016-1549-z.

(29) Williams, D. K.; Muddiman, D. C. Parts-per-billion mass measurement accuracy achieved through the combination of multiple linear regression and automatic gain control in a fourier transform ion cyclotron resonance mass spectrometer. *Anal. Chem.* **2007**, *79* (13), 5058–5063. https://doi.org/10.1021/ac0704210.

(30) Savory, J. J.; Kaiser, N. K.; McKenna, A. M.; Xian, F.; Blakney, G. T.; Rodgers, R. P.; Hendrickson, C. L.; Marshall, A. G. Parts-per-billion fourier transform ion cyclotron resonance mass measurement accuracy with a walking calibration equation. *Anal. Chem.* **2011**, *83* (5), 1732–1736. https://doi.org/10.1021/ac102943z.

(31) 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. **2013**, *24* (7), 1137–1145. https://doi.org/10.1007/s13361-013-0659-0.

(32) Barrow, M. P.; Witt, M.; Headley, J. V.; Peru, K. M. Athabasca oil sands process water: Characterization by atmospheric pressure photoionization and electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **2010**, *82* (9), 3727–3735. https://doi.org/10.1021/ac100103y.

(33) Klitzke, C. F.; Corilo, Y. E.; Siek, K.; Binkley, J.; Patrick, J.; Eberlin, M. N. Petroleomics by ultrahigh-resolution time-of-flight mass spectrometry. *Energ. Fuel.* **2012**, *26* (9), 5787–5794. https://doi.org/10.1021/ef300961c.

(34) 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* (15), 7698–7704. https://doi.org/10.1021/acs.analchem.6b01624.

(35) 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* (21), 11383–11390. https://doi.org/10.1021/acs.analchem.7b02345.

(36) Chang, W.; Cheng, J.; Allaire, J.; Xie, Y.; McPherson, J. *Shiny: Web Application Framework for R*; 2019.

Supporting Information: Reproducibility, signal-to-noise ratio, mass error and molecular assignments in petroleomics

Remy Gavard,[†] Diana Catalina Palacio Lozano,[‡] Hugh E. Jones,[†] 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*

[§]*Department of Economics & Business, Universitat Pompeu Fabra, Barcelona, 08005, Spain*

E-mail: [redacted]

Phone: [redacted]

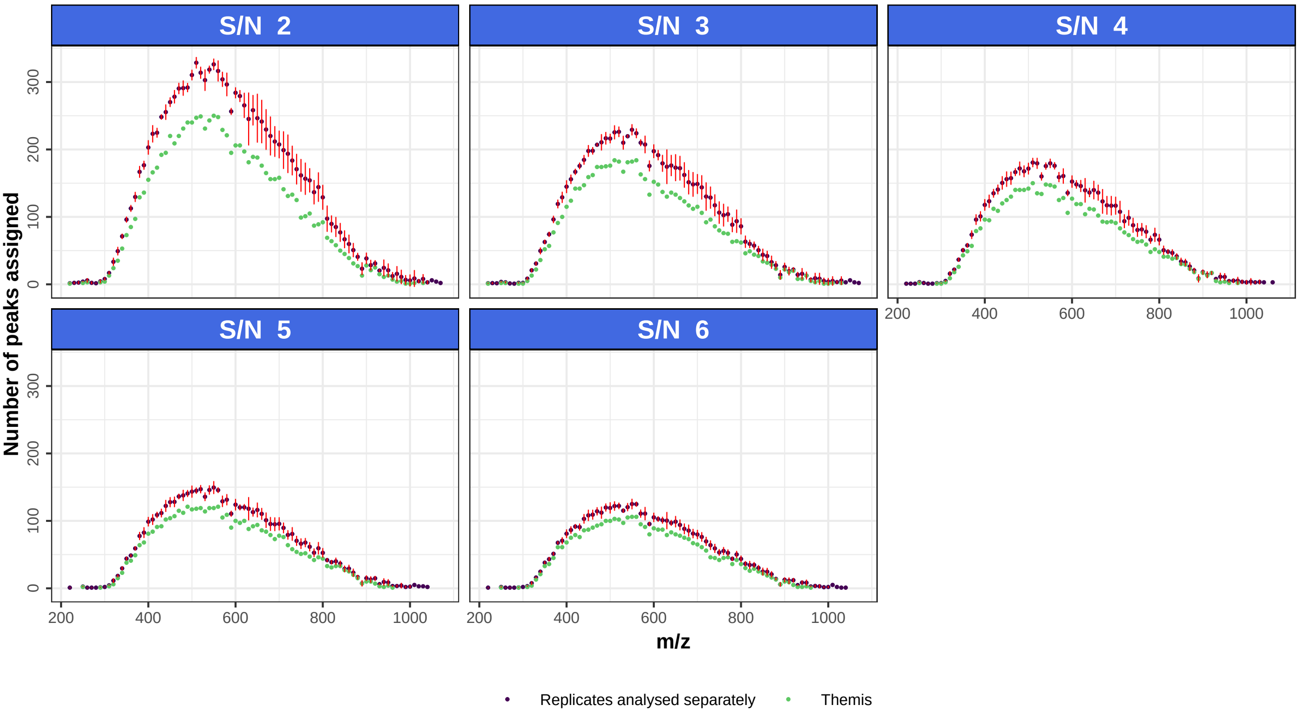


Figure S1: Number of peaks assigned per m/z width of 10 for each S/N for the sample A.

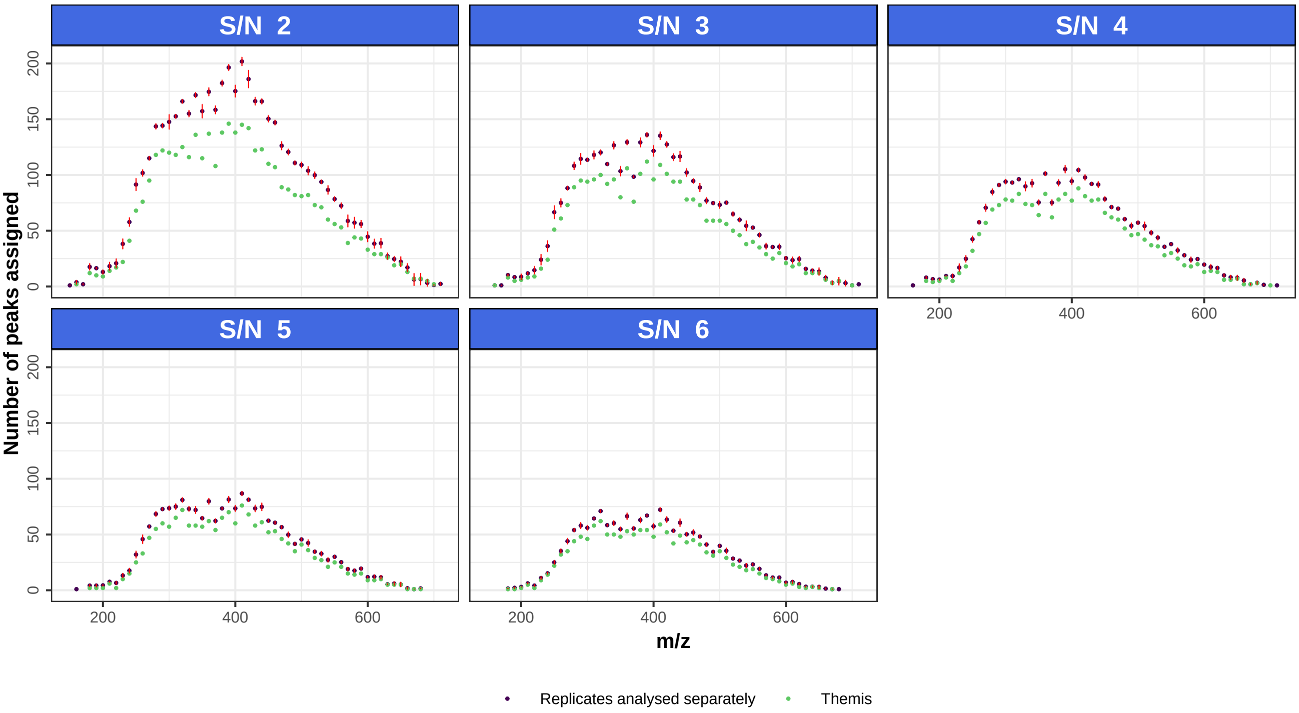


Figure S2: Number of peaks assigned per m/z width of 10 for each S/N for the sample B.

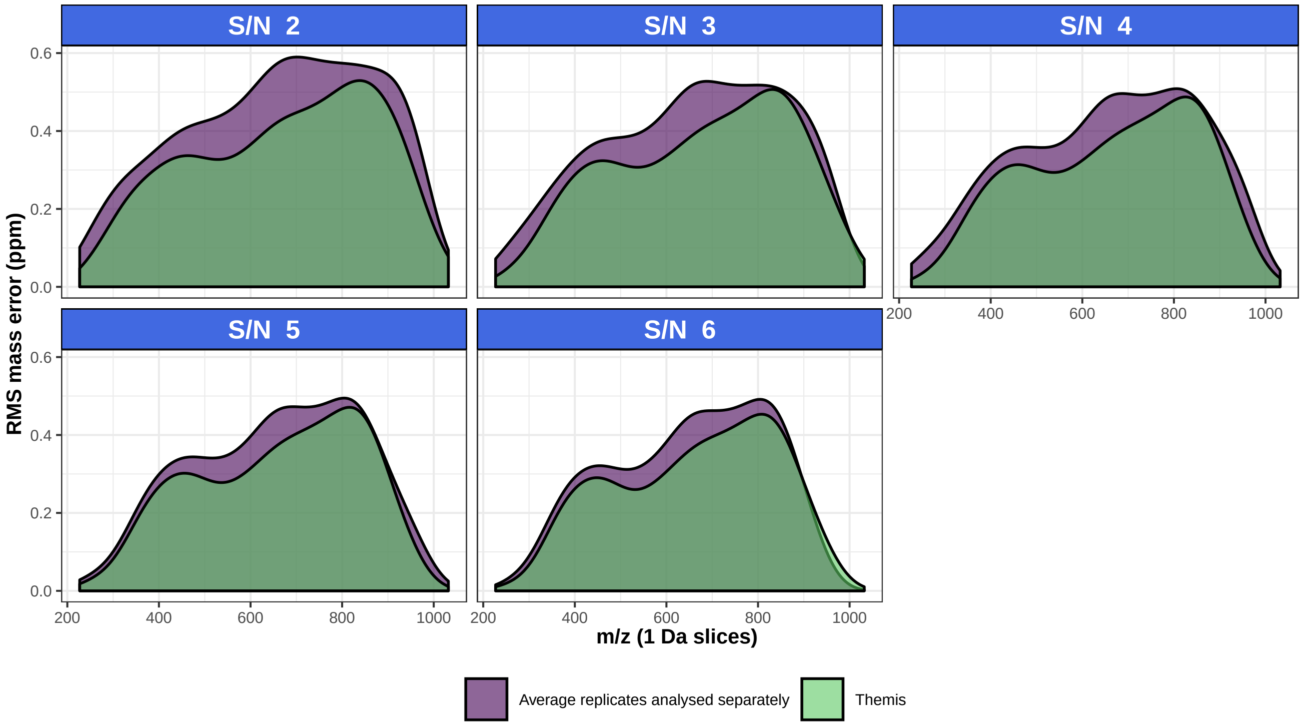


Figure S3: Evolution of the RMS mass error accross the m/z range of the Sample A

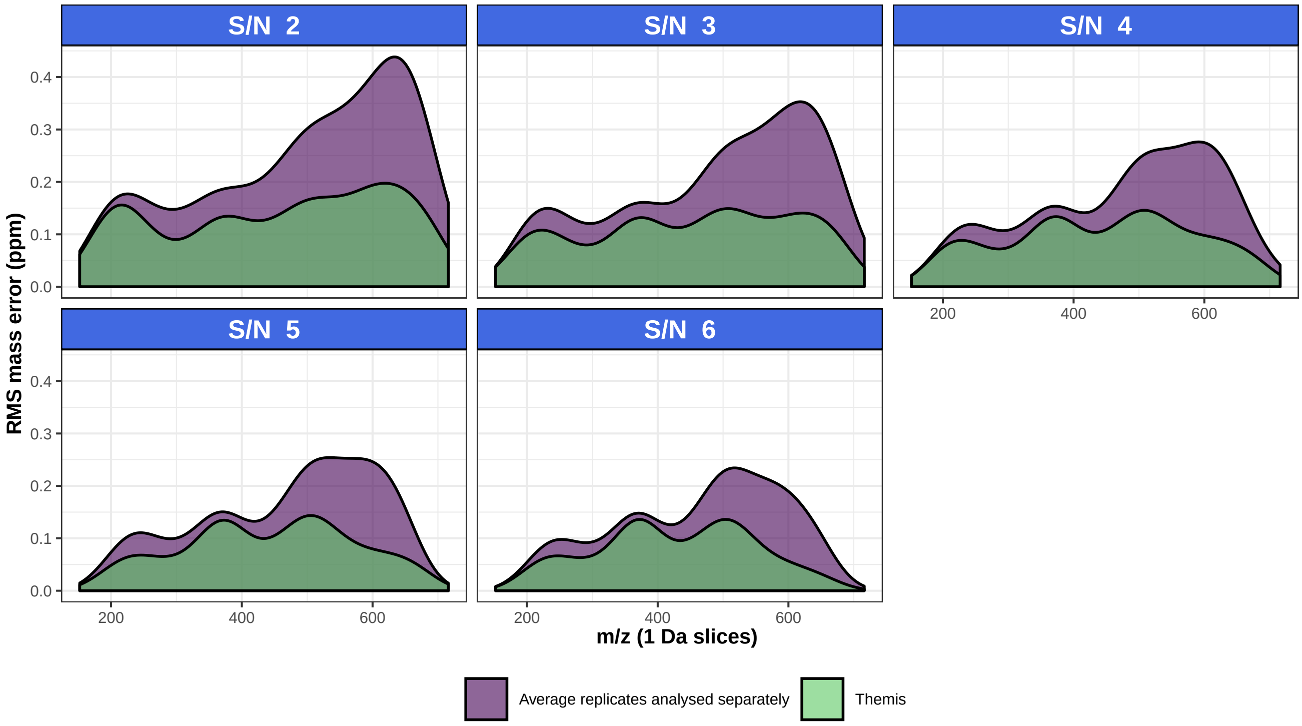


Figure S4: Evolution of the RMS mass error accross the m/z range of the Sample B

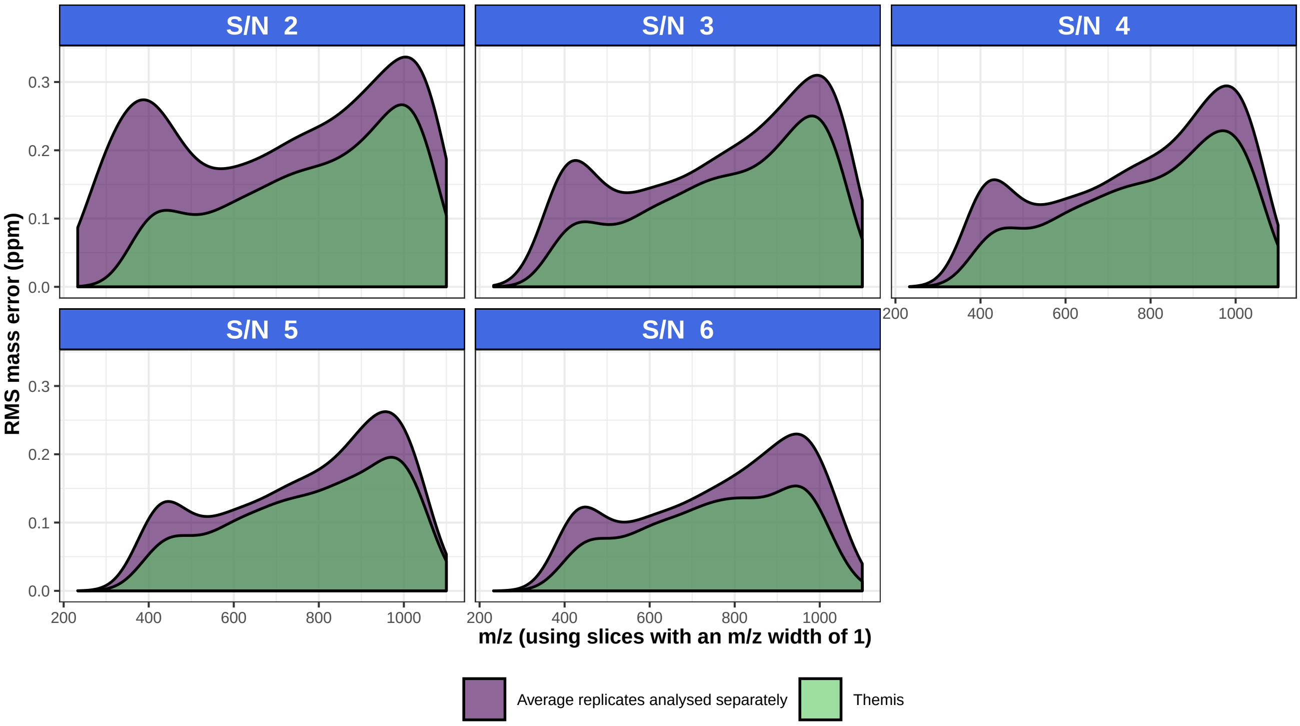


Figure S5: Evolution of the RMS mass error across the m/z range of the Sample C

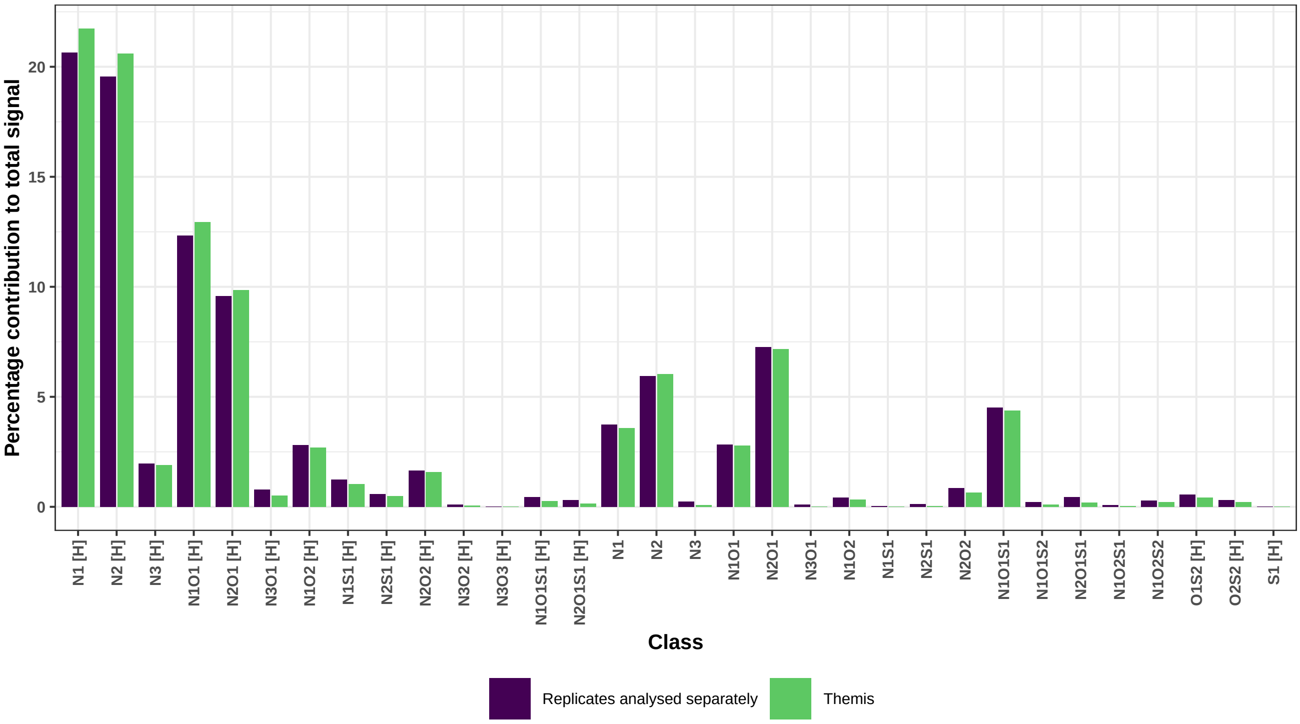


Figure S6: Heteroatom classes distribution for sample C at S/N 3.

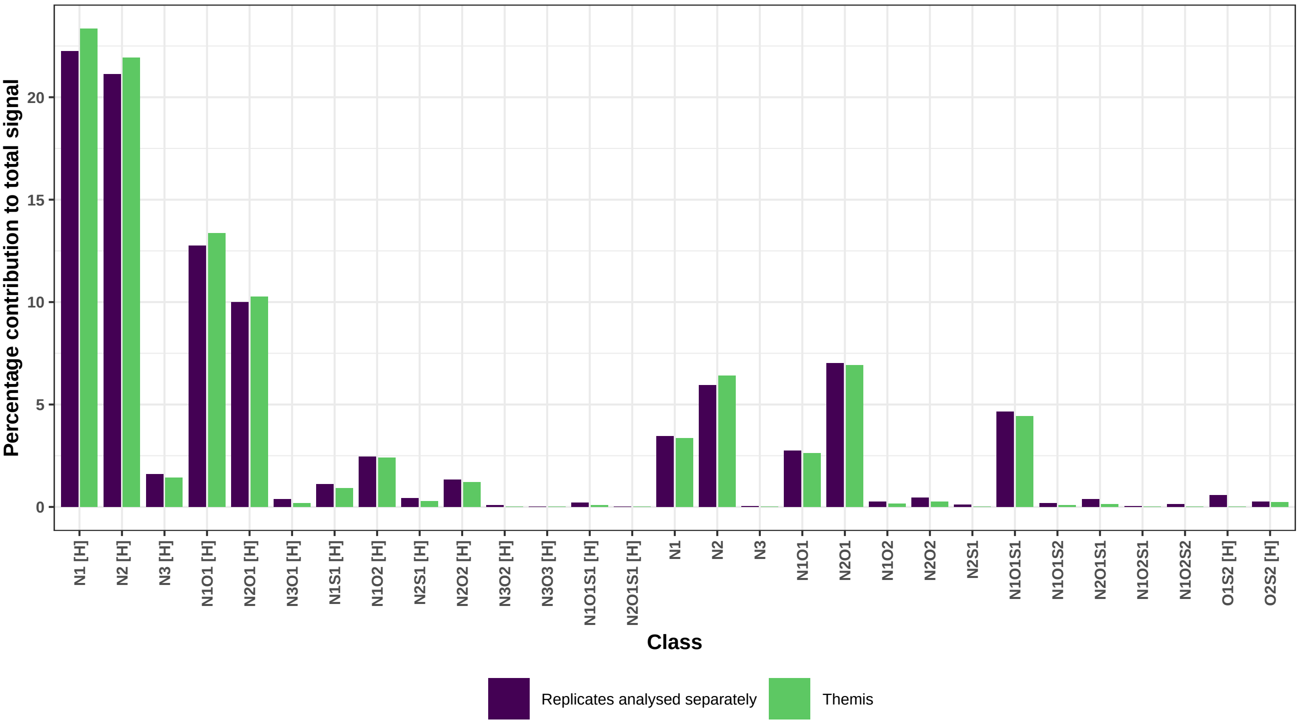


Figure S7: Heteroatom classes distribution for sample C at S/N 4.

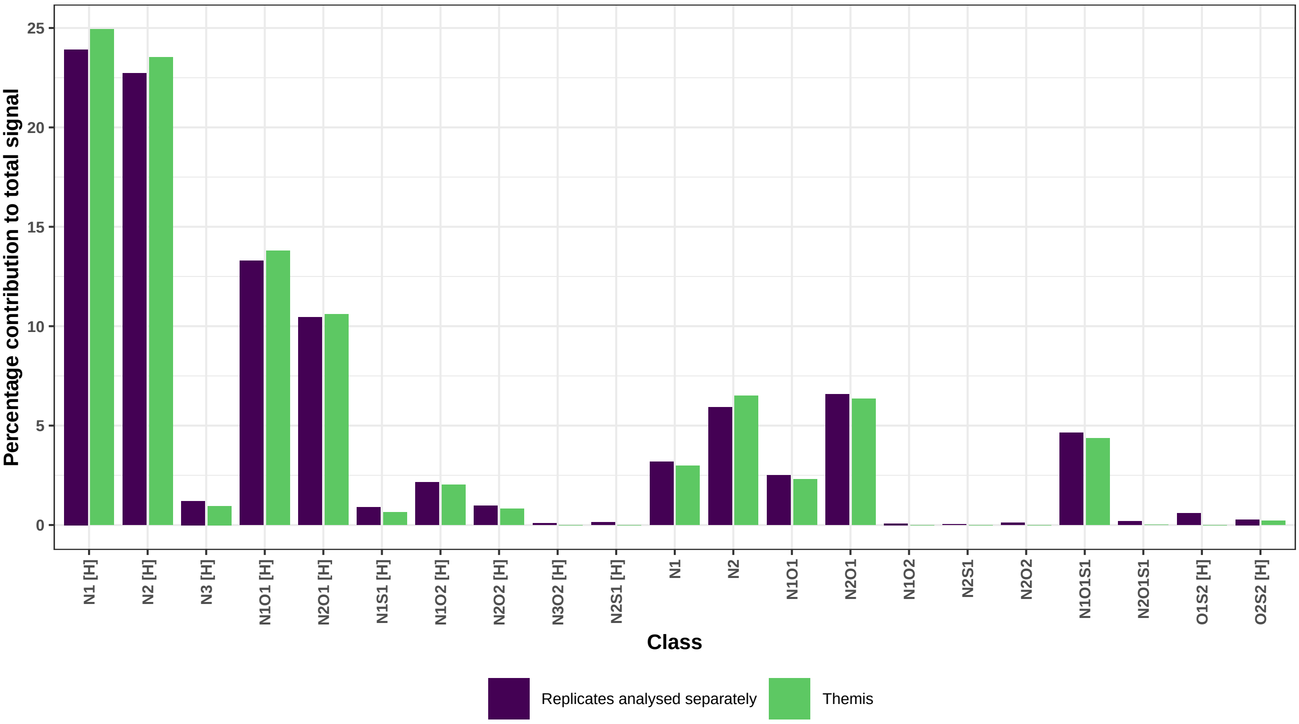


Figure S8: Heteroatom classes distribution for sample C at S/N 5.

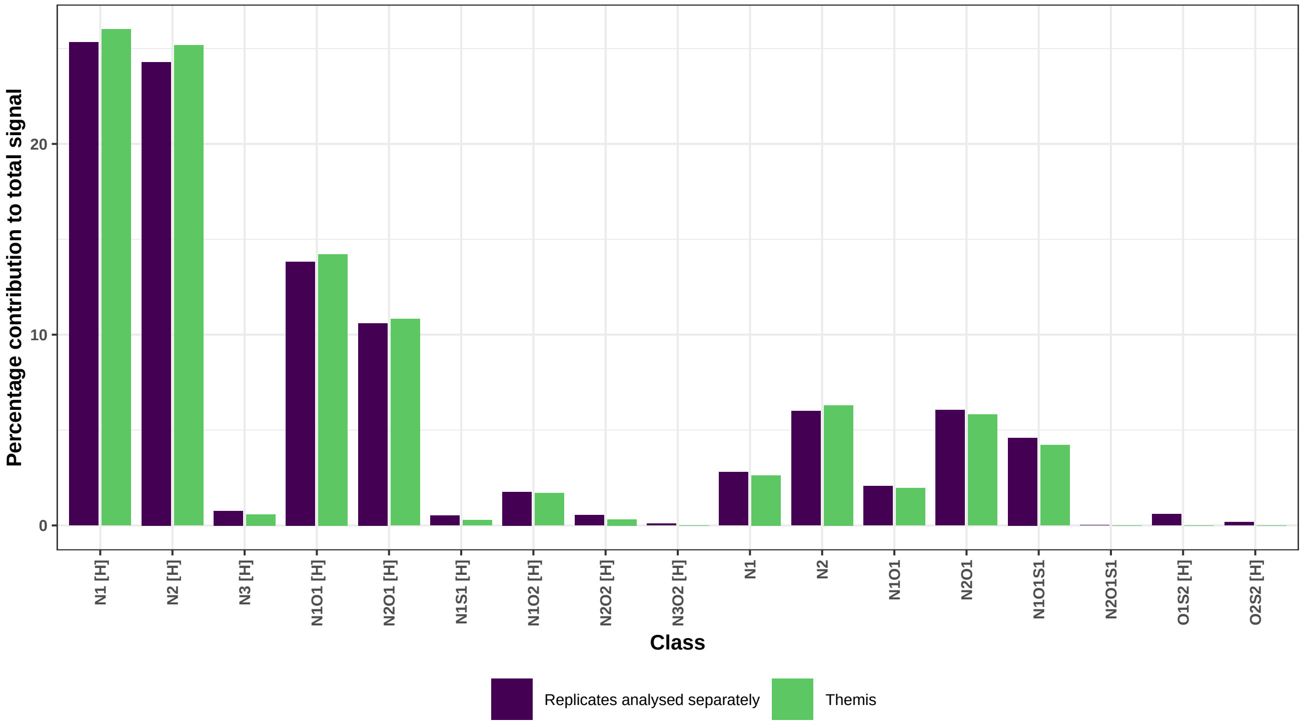


Figure S9: Heteroatom classes distribution for sample C at S/N 6.

Chapter 4

Rhapso: Automatic stitching of mass segments from Fourier transform ion cyclotron resonance mass spectra

4.1 Context

The complexity of crude oil has always pushed FTICR MS detection limits, leading to record resolution and mass accuracy. Some extremely complex samples contain so many different components that competitive effects take place within the collision cell and the ICR cell, and there is a detection limit of minimum 50 to 100 ions which prevents most ions from being detected. A new technique consisting of segmenting the acquisition into small m/z bins enabled more molecules to be observed within those samples. Unfortunately, this means that the data needs to be manually stitched together. The segments also display a decrease in intensity at the edges due to the quadrupole isolation, a phenomenon often called an “edge effect”. This segmented acquisition method offered higher accuracy and the ability to detect a record number of ions. A need emerged for a method to automatically stitch the segments together to form a spectrum but also to correct for the intensity drop at the edges. The Rhapso algorithm was developed to tackle these tasks, relieving the user from a laborious task but also reducing the need for large overlaps between segments, reducing the number of spectra needed to cover the complete m/z range of interest.

This chapter was published as an article in *Analytical Chemistry*. The Rhapso algorithm presented in this chapter was used in a publication in *Chemical Science*. Rhapso was developed and coded by the author to be part of a new method called OCULAR. The sample was provided by Alexander Guzman. The data acquisition necessary for the development of this algorithm was performed by Diana Catalina Palacio Lozano with my assistance. The manuscript for this chapter was written by the author.

4.2 Publication

analytical chemistry

Article

Cite This: Anal. Chem. 2019, 91, 15130−15137

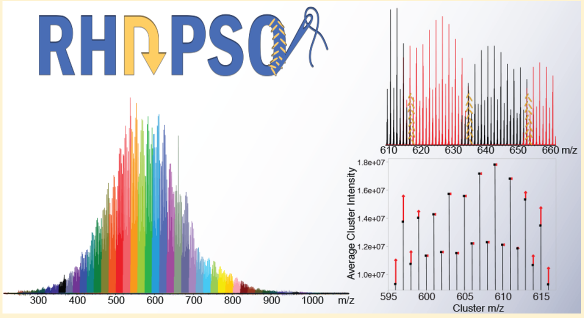
pubs.acs.org/ac

Rhapso: Automatic Stitching of Mass Segments from Fourier Transform Ion Cyclotron Resonance Mass Spectra

Remy Gavard,[†] Diana Catalina Palacio Lozano,[‡] Alexander Guzman,[¶] 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
[¶]Instituto Colombiano del Petróleo, Piedecuesta, 681011, Colombia
[§]Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom
[∥]Department of Economics & Business, Universitat Pompeu Fabra, Barcelona, 08005, Spain

Supporting Information



ABSTRACT: Fourier transform ion cyclotron resonance mass spectrometry (FTICR MS) provides the resolution and mass accuracy needed to analyze complex mixtures such as crude oil. When mixtures contain many different components, a competitive effect within the ICR cell takes place that hampers the detection of a potentially large fraction of the components. Recently, a new data collection technique, which consists of acquiring several spectra of small mass ranges and assembling a complete spectrum afterward, enabled the observation of a record number of peaks with greater accuracy compared to broadband methods. There is a need for statistical methods to combine and preprocess segmented acquisition data. A particular challenge of quadrupole isolation is that near the window edges there is a drop in intensity, hampering the stitching of consecutive windows. We developed an algorithm called Rhapso to stitch peak lists corresponding to multiple different *m*/*z* regions from crude oil samples. Rhapso corrects potential edge effects to enable the use of smaller windows and reduce the required overlap between windows, corrects mass shifts between windows, and generates a single peak list for the full spectrum. Relative to a stitching performed manually, Rhapso increased the data processing speed and avoided potential human errors, simplifying the subsequent chemical analysis of the sample. Relative to a broadband spectrum, the stitched output showed an over 2-fold increase in assigned peaks and reduced mass error by a factor of 2. Rhapso is expected to enable routine use of this spectral stitching method for ultracomplex samples, giving a more detailed characterization of existing samples and enabling the characterization of samples that were previously too complex to analyze.

Petroleum is one of the most complex mixtures found in nature and can contain hundreds of thousands of unique elemental compositions within a single sample.[1] The study of petroleum composition has become known as "petroleomics".[1−10] Developing a more detailed understanding of petroleum composition in order to address the challenges of producing and refining crude oil has become increasingly important in recent years.[11−14] Fourier transform ion cyclotron resonance mass spectrometry (FTICR MS)[15−20] is a state-of-the-art technique for petroleomics that provides a significant ultrahigh resolving power and mass accuracy to assign elemental compositions of highly complex samples.[21] FTICR cells can hold a maximum of a few million ions, and singly charged ions will be detected if their presence reaches the detectable amount of at least 50 to 100 ions.[22,23] If several thousands of molecular compositions are present, many species can fall below the detection limit.[24] To overcome this problem, we traditionally use signal averaging, summing the data over several scans (usually several hundred).[25] This method is reaching its limit for extremely complex samples, due to the space-charge effects which lowers the isolation dynamic range and mass accuracy.[26]

The space-charge effect can be addressed by segmented acquisition, a method to obtain a full-range FTICR spectrum when the instrumentation was not able to produce a broadband spectrum.[27,28] The spectral stitching method has gained interest recently as the limits of the broadband techniques start to show.[29,30] A quadrupole mass analyzer is used to select ions within a specified *m*/*z* range before being passed to the ICR cell, hence reducing the number of different molecular compositions in the cell and helping to get the molecules above the detection threshold. A complete method

Received: August 22, 2019
Accepted: October 30, 2019
Published: October 30, 2019

DOI: 10.1021/acs.analchem.9b03846

called "selected ion monitoring (SIM) windows" by its authors was established by Southam et al. for biological samples and later used to enhance relative isotopic abundance measurements by Weber et al.[32] SIM was recently improved and made widely available for mass-spectrometry-based metabolomics and lipidomics.[33] Earlier work by Rodgers et al.[34] and Zabrouskov and Senko[35] used stitched spectra after segmented acquisition; however, the data analysis methodology has not been described or made available. Every application of the SIM method reported a higher number of peaks, an increase in the number of peaks assigned, and a higher mass accuracy. Currently, a complete stitching method was only established for metabolomics and lipidomics samples, where the number of peaks observed is around a few thousand and the maximum mass width investigated is m/z 700.[31] In contrast, in complex petroleum samples, the number of peaks in broadband mass spectra can easily reach ten of thousands[30] and spana width of around 1000 m/z. This means that the methods developed for biological samples are not directly applicable to petroleum samples since the windows size must be adapted to the higher molecular density. In addition, because of the higher mass range, it is not realistic to use an overlap of width of 10 m/z between windows. The calibration tools available are different too. The strategy of Southam et al.[31] relies on having a critical number of isolation windows with an internal standard. Petroleomics researchers try to limit the use of internal calibration to avoid doping the sample, but instead use known molecular series as internal calibrants. No methods that are able to perform stitching using peak lists have been publicly described for petroleomics to date. In 2012, Gaspar and Schrader[29] used the commercial software Xcalibur (Thermo Electron, Bremen, Germany) to recreate a full spectrum by adding all the segments acquired, as well as those from the broadband. Recently Krajewski et al.[30] performed the stitching by manually trimming the best width of 20 m/z out of a width of 25 m/z acquisition and ensuring that there was no overlap with the following windows to prevent duplicating peaks.

Southam et al.[31] observed a phenomenon that they called an "edge effect" consisting of a reduction in intensity at the isolation windows' edges compared to what was expected by studying the ratio of two peaks depending of their position across the window. The strategy employed by the authors to account for this effect was to use a large overlap between windows (roughly a width of 10 m/z), so the edge of one window is covered by the central part of the subsequent window (where there is no edge effect). This strategy cannot be practically translated to petroleum samples, as these require substantially smaller windows; increasing their overlap would result in an experiment that could take days to complete.

In this paper we describe a new algorithm called Rhapso to automatically clean the peak lists, correct the edge effect via a convenient statistical model, and stitch the peak list. Rhapso was the name of a nymph in the greek mythology which derives from a greek verb meaning to stitch. Rhapso was recently successfully used to help achieve the highest resolving power and number of unique molecular assignment to date.[36]

■ METHODOLOGY

The quadrupole was used to only transmit ions within a specified m/z region for detection. This creates an isolation window of a mass spectrum; a partial mass spectrum results, which will be referred to as a "segment". The user keeps the width of the isolation window the same (e.g., spanning a width of 20 m/z) but progressively moves the center of the isolation window to higher m/z (e.g., m/z 261, 279, 297, etc.). In this way, the user acquires a large number of overlapping segments which span the entire m/z range of interest. In Rhapso each segment is trimmed at the high and low m/z ends of the observed signal to enable good subsequent stitching and prevent the inclusion of noise or low quality peaks, producing reduced-width segments. The reduced-width segments can then be combined appropriately, finding suitable regions for overlap, to produce a new mass spectrum.

Sample Preparation. A South American vacuum residue sample obtained using supercritical fluid extraction was used to illustrate our data analysis methodology. The sample has around 90% of its constituents with a boiling temperature less than 720 °C at atmospheric equivalent temperature (AET). High performance liquid chromatography (HPLC) grade toluene (Fisher Scientific, Loughborough, UK) was used to dilute the sample to a concentration of 0.05 mg/mL.

Instrumentation. Mass spectra were acquired using an Apollo II atmospheric pressure photoionization (APPI) source, coupled to a 12 T solariX FTICR mass spectrometer (Bruker Daltonik GmbH, Bremen, Germany). The injection was performed using a flow rate of 500 mL h^{-1}, vaporizer at 350 °C, drying gas at 250 °C, and capillary potential at 1200 V. Potentials of 0.4 V were applied to the front and back trap plates of the ICR cell. For the broadband mass spectrum, a data size of 4 M was used with a detection range of m/z 250–3000, and 100 time-domain transients were coadded. To produce the stitched data, the m/z range (equivalent to 1000 Da) was segmented into 41 windows, each with an m/z width of 20, with each window overlapping the adjacent windows by an m/z width of 2. A quadrupole was used to isolate these narrow m/z ranges, and 50 time-domain transients were coadded.

In order to avoid influencing the peak abundance, the mass envelope, excitation range, magnitude, and ion accumulation time were kept constant. After acquisition, a single zero fill and Sine-Bell apodization were applied before usage of a Fourier transform.

Signal Processing. The FTMS Processing 2.1.0 software was used with an asymmetric apodization ("Kilgour")[37] function for offline phasing of the segments to generate absorption-mode spectra. The spectra were then exported from solariXcontrol to DataAnalysis 4.2 and then internally recalibrated using the HC class with a mass difference of 2.01565 Da. Finally, the peak finder "FTMS" method was used to extract peak information and provide a peak list for each segment and the broadband mass spectrum.

Statistical Processing. Rhapso consists of four steps as depicted in the flowchart in Figure 1.

Step 1: *Removal of Peaks Outside of the Isolation Window.* Peak finding algorithms (e.g., the DataAnalysis 4.2 FTMS peak picking algorithm used in our examples) may identify peaks outside of the target m/z isolation window. In addition, the width of each segment can be different.

We have developed a strategy to obtain reduced-width segments that have a common width and contain the peaks of interest. One option would be to ask the user to input the theoretical m/z range targeted by each window; however, the observed m/z range can differ from the theoretical one due to the precision of isolation of the quadrupole. Instead, we developed a method to detect automatically the m/z range using as little prior information as possible. The method allows

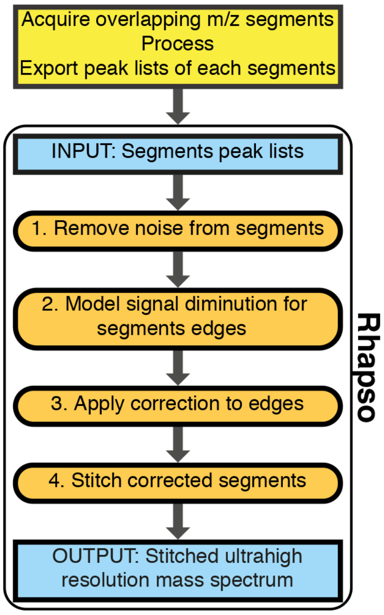

Acquire overlapping m/z segments
Process
Export peak lists of each segments
INPUT: Segments peak lists
1. Remove noise from segments
2. Model signal diminution for segments edges
3. Apply correction to edges
4. Stitch corrected segments
OUTPUT: Stitched ultrahigh resolution mass spectrum
Rhapso

Figure 1. Flowchart representing the four processing steps of Rhapso taking place after acquisition,signal processing and export to peak list for each segment.

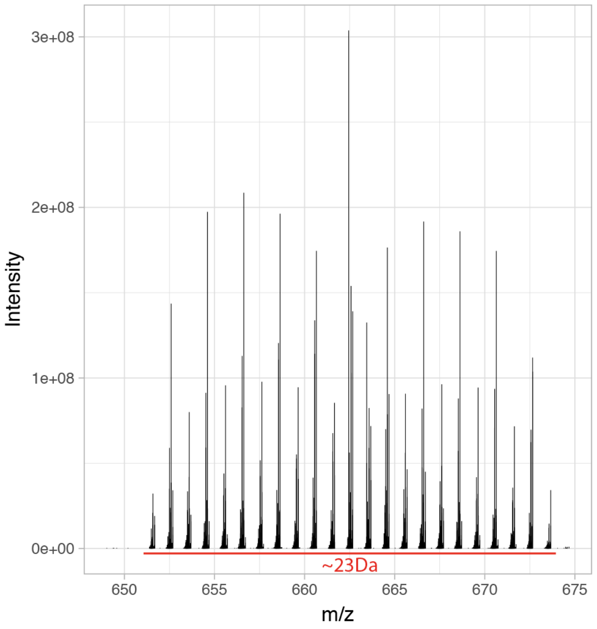


Figure 2. Mass spectrum of a segment between m/z 651 and 674 after FTMS peak picking, before processing and stitching with Rhapso.

for a maximum overlap between consecutive segments of 50% the size of the segments; e.g., for a width of 20 m/z, the maximum overlap is a width of 10 m/z.

Each segment was acquired aiming for a theoretical m/z width, which we denote by W and was kept constant for all spectra. Let r be the number of segments and n_j be the number of peaks in segment $j = 1, ..., r$. Define $M_{i,j}$ as the m/z value of peak $i = 1, ..., n_j$ in segment j and $I_{i,j}$ as the intensity value of peak $i = 1, ..., n_j$ in segment j. We define $S_j = min_i M_{i,j}$ to be the smallest m/z $M_{i,j}$ of segment j and H_j the highest m/z $M_{i,j}$ of segment j. Further, let E_j be the integer m/z at the center of the segment as specified by the user. The observed width of the signal in each segment is often larger (up to 20%) than W).

Our goal is to stitch segments using as wide a range as possible, subject to the measured intensities being high enough to ensure that measurements are accurate. Specifically, we seek the most suitable width $W + x$ to use for the stitching where x is a width adjustment for all segments to be determined as described below. Crude oil molecules are detected in clusters of peaks occurring every integer; consequently, we will investigate segments with widths $W + x$ for integer $x = 0, 1, ..., \lceil 0.2W \rceil$. The cleaning procedure described below is repeated for all the values of x. For each $(W + x)$ and within each segment j we look for the m/z value which maximizes the sum of the intensities in a window of width $(W + x)$. Hence $d_j = \arg\max_d \sum_{i \in S(d)} I_{i,j}$ where $S(d) = \{i : d \le M_{i,j} \le d + (W + x)\}$ within an m/z interval $[d, d + (W + x)]$. The range of d considered for segment j is given by the interval $d \in [E_j - (W + x) + C_j, E_j + C_j]$, with C_j being the decimal which needs to be added to the integer m/z in order to ensure that a cluster of peaks does not get split. If the peak density is too high to determine a space between them, a region with low intensity peaks will be selected. In order to calculate C_j we search for the largest 10 gaps between peaks within $[E_j - (W + x), E_j + (W + x)]$, and for those 10 gaps we calculate the decimal places of $(M_{i+1,j} + M_{i,j})/2$ to identify how far the centers of the gaps are from the integers. If the standard deviation of the decimals of those 10 values is under 0.1, then C_j is the average of those decimals; otherwise $C_j = 0$.

Step 2: Estimate the Edge Effect. To investigate and measure systematic intensity decreases at the segment edges, we combined the data from all the cleaned segments by shifting them on a common new scale. This is because individual segments display natural variability due to the chemistry of the crude oils which can be mistaken as an intensity drop if located toward the edges. In contrast, by stacking all segments one can estimate common patterns in intensity drops. The shifted m/z value for peak i in spectrum j is defined as $Z_{i,j} = M_{i,j} - S_j$. That is, the shifted m/z values range is $Z_{i,j} \in [0, W + x]$. Now we divide the $Z_{i,j}$ into k bins 1 m/z wide and let n_k denote the number of peaks for bin k and $Y_k = \log\left(\frac{1}{n_k} \sum_{i=1}^{n_k} I_{i,j}\right)$ for bin k and $X_k = min_{i,j \in k}(Z_{i,j})$ denote the floor m/z $Z_{i,j}$ of each bin. We model the log of the mean intensity of each bin using a piecewise linear model where a and b define the change points at which intensity starts to drop near the edges. That is,

$$\begin{aligned} Y_k = \beta_0 + \beta_1[(X_k - a) \times \mathbb{1}(X_k < a)] \\ + \beta_2[(X_k - b) \times \mathbb{1}(X_k > b)] + e_k \end{aligned} \qquad (1)$$

for $k = 1, ..., W + x$ where e_k is an error term and $\mathbb{1}$ the indicator function. The model is fitted by least-squares, which is finding $(a, b, \beta_0, \beta_1, \beta_2)$ that minimizes the mean square residuals (MSR). Specifically, given (a, b) the optimal $(\beta_0, \beta_1, \beta_2)$ can be found by ordinary linear regression. Hence, it suffices to consider a grid of (a, b) values for $a \in \{0, ..., [0.2(W + x)]\}$ and $b \in \{(W + x) - [0.2(W + x)], ..., (W + x)\}$, to find the MSR associated with the optimal $(\beta_0, \beta_1, \beta_2)$ and choose the (a, b) attaining the smallest MSR overall. This step was performed for each of the window sizes $W + x$ investigated. The smallest MSR obtained from all the different window sizes

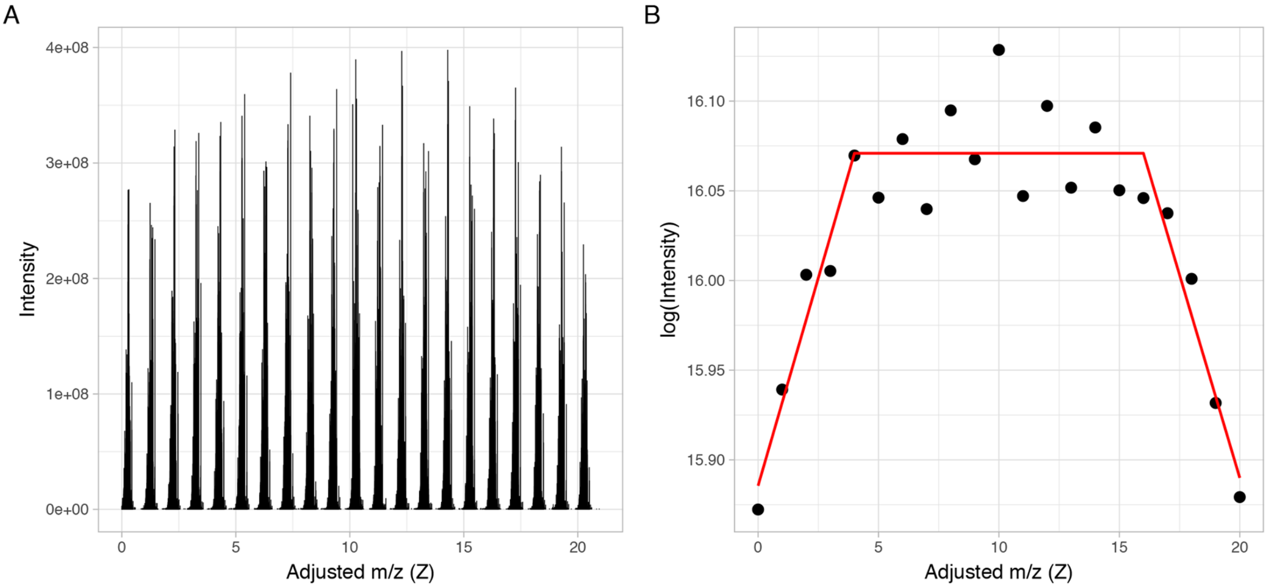


Figure 3. (A) Overlapped log-intensities across all windows in the shifted m/z scale $Z_{i,j}$ and (B) plot of the average intensity of each bin with an m/z width of 1.

is isolated, determining the value of (x, a, b) to be used in the rest of the processing.

Step 3: Correct Window Edge Effects. We used the piecewise linear model in eq 1 to correct intensity drops at the segments' edges of each individual segment. Because both the overall intensity and the magnitude of the intensity vary across segments, we estimate β_0, β_1, β_2 separately for each window j. We calculate β_1 and β_2 for each segment and use the associated segment center to fit a locally estimated scatterplot smoothing (LOESS) model[38] for each β. The models are then used to calculate a smoothed β_1 and β_2. This is done to avoid being affected by outlier peaks, especially in the presence of a low number of peaks at the extremities of the window, which can result in an inappropriate value. We denote these new estimated coefficients as $\widehat{\beta_{0j}}$, $\widehat{\beta_{1j}}$, $\widehat{\beta_{2j}}$. The corrected intensities $\widehat{I_{i,j}}$ are obtained as

$$\log(\widehat{I_{i,j}}) = \begin{cases} \log(I_{i,j}) + \max\{\beta_{1j}, 0\}((S_j + a) - M_{i,j}) & \text{if } M_{i,j} < S_j + a \\ \log(I_{i,j}) + \max\{|\beta_{2j}|, 0\}((H_j - b) - M_{i,j}) & \text{if } M_{i,j} > H_j - b \end{cases} \tag{2}$$

where S_j and H_j are the lowest and highest m/z values in segment j.

Step 4: Stitching. We have an overlap between segments j and $j + 1$ and we must choose from which window to take peaks. We wish to determine the best m/z to change from segment j to $j + 1$. We isolate the peaks from both spectra with an m/z within the interval $[S_{j+1}, H_j]$ and form a new set $u(j) = \{M_{i,j}: M_{i,j} \geq H_j\} \cup \{M_{i,j+1}: M_{i,j+1} \leq S_{j+1}\}$. Let $\widehat{M_{i,j}}$ be the elements of $u(j)$. We search for the top k largest gaps arg max$_i$ $M_{i+1,j} - M_{i,j}$ between peaks in the overlap region $u(j)$. Define P_1, ..., P_k to be the midpoints of these k gaps. In addition, we define $P_0 = (S_{j+1} + H_j)/2$ to be the center of the overlap. The merging point P^* is selected to be the midpoint closest to the center P_0 : arg min$_l$ $|P_l - P_0|$. Once all the merged peaks have been identified and the excess peaks removed from each end, the isolation windows are assembled into a unique peak list and exported for further analysis.

■ RESULTS AND DISCUSSION

The calibrated and phased mass list of each segment was exported using DataAnalysis 4.2 as text files and processed with Rhapso. Rhapso has been implemented using a Shiny web interface.

At this stage, the peak list of each isolation window can contain peaks from outside the targetted isolation due to the noise. The lower the signal-to-noise ratio (S/N) used for the peak picking, the more noise that will be included. As seen in Figure 2, different segment widths can be considered. For instance, the window illustrated in Figure 2 has an observed width of 23 m/z, larger than the theoretical width W of 20 m/z. It is in our best interest to retain as much of segments as possible to be efficient and use fewer segments. In our application, we have explored using widths of 20, 21, 22, and 23 m/z. The m/z range to include for each of the widths considered was determined as described in Step 1. In the subsequent steps we explore each of them and determine which one is the best. As explained in step 2 of the method, we have modeled and corrected a decrease in intensity at the edges of the spectra caused by the isolation by the quadrupole. This phenomenon was described by Southam et al. and called an "edge effect". Instead of using a large overlap and deleting the edges, which would require more segments and in consequence a much longer acquisition time, we decided to

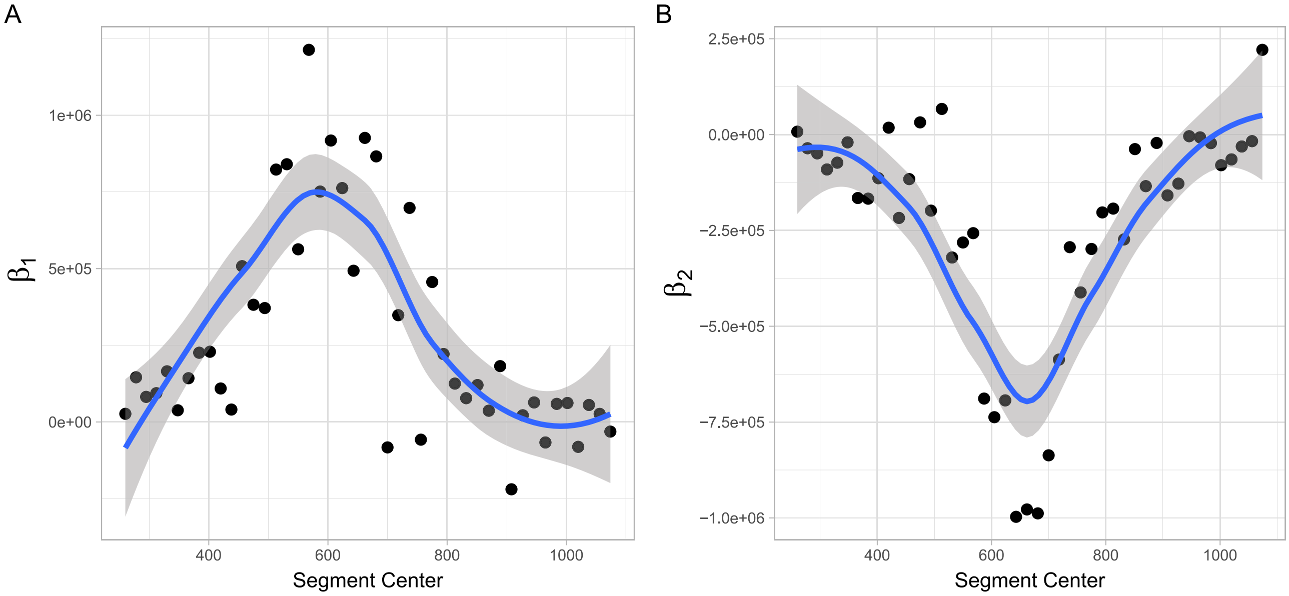


Figure 4. LOESS model estimate of (A) $\widehat{\beta}_{1j}$ and (B) $\widehat{\beta}_{2j}$ for a width $W = 21$, as a function of the segment center.

apply a correction of the intensity. In order to calculate a piecewise model of the intensity diminution at the edges, we had to find where the decrease in intensity occurs. This is specific to the type of sample analyzed, the molecular abundance, the size of the windows, and the instrument. Because of its chemical composition, crude oil displays natural undulations that need to be preserved. In Figure 3 A, we subtract the minimum m/z of each window to all the peaks present in the window and calculated the log of the mean intensities in each bin of width 1 m/z. This allows us to go beyond the natural undulations and to display a clear pattern of intensity diminutions at the edges of the windows. By looking for the minimum mean squared residuals of a piecewise model, we determine the optimum break points, where an intensity correction is needed (Figure 3 B).

Using a grid search for the best breakpoints a and b, the MSR for the model was used to determine the best values for each width considered. Finally, the width with the model yielding the minimum MSR determines the width used in the subsequent steps. After the optimal width W was determined, the β coefficients were calculated for each segment (Figure 4) and a LOESS model was fitted. Figure 4 was created without any log transformation on the intensity to highlight the similarities of the distribution of the coefficients with the intensity profile of the final stitched spectrum. It becomes clear that using the same β for all segments would not be a good fit and proves that the intensity drop at the edges is more pronounced when the intensity increases.

The correction was applied using a β coefficient calculated on the log(intensity) scale and applied to log(intensity) data before being transformed back to the original scale.

In order to assess the quality of the intensity correction performed during step 3, we looked at the mean intensity of each window of width 1 m/z. Figure 5 represents these averaged corrected and uncorrected intensities for each peak cluster. The dots represent the mean intensity without correction, and the arrows point where the mean intensity is after correction. We notice that the correction applied helps restore the natural undulation profile. In accordance to the distribution observed in Figure 4, the correction was most visible in the most intense segments while, at the edges (low and high m/z), the correction was minimal, and sometimes not necessary.

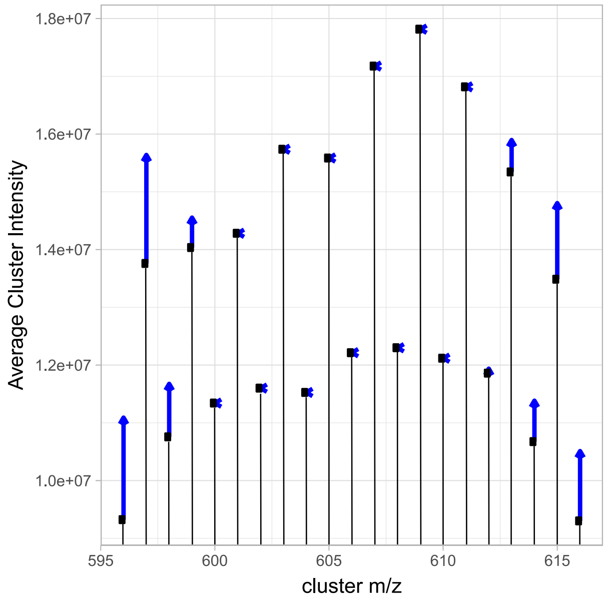


Figure 5. Mean intensity of each window with an m/z width of 1. The arrows indicate the changes to mean intensity of each peak cluster after correction.

After the intensity correction was applied, the merge between each segment was performed. As described in the methods and illustrated in Figure 6A, the overlap region was isolated (illustrated by the two vertical lines). Since we know that the quality of the peaks deteriorates at edges and we want to prevent either duplicating or losing any peaks, the best place to switch from one segment to the other will be in the center of the overlap region and between clusters.

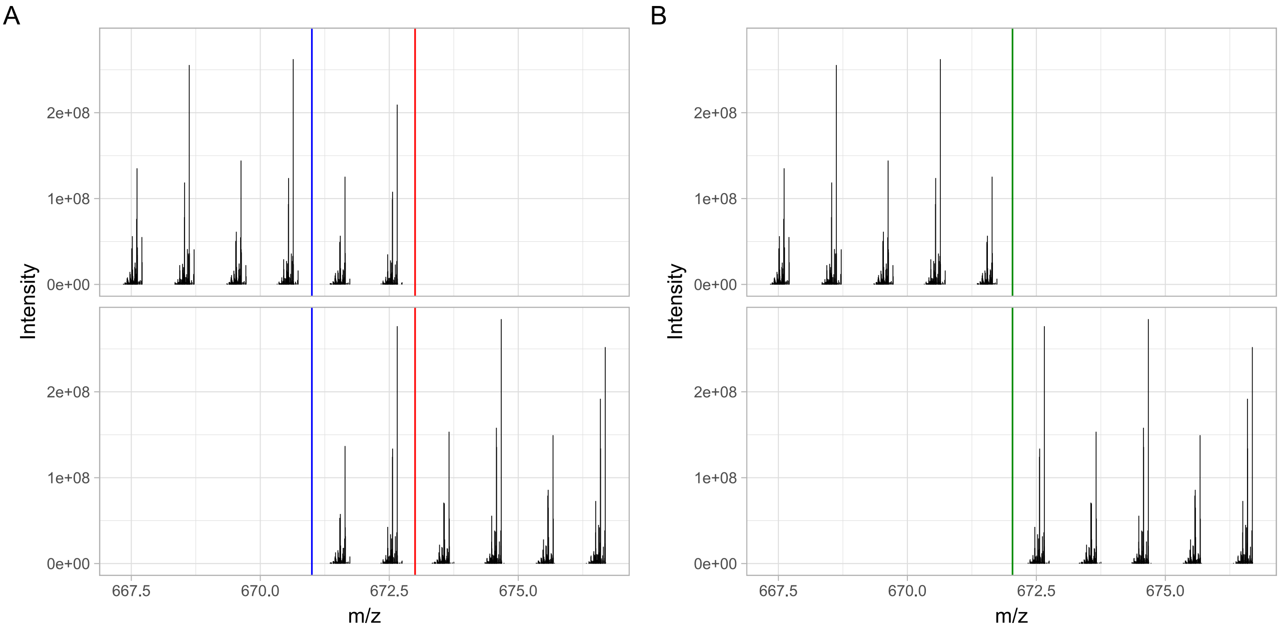


Figure 6. Illustration of the isolation of the overlap region between two segments (A) before and (B) after removal of the unnecessary peaks at the extremities, following automated calculation of the overlap position for segments.

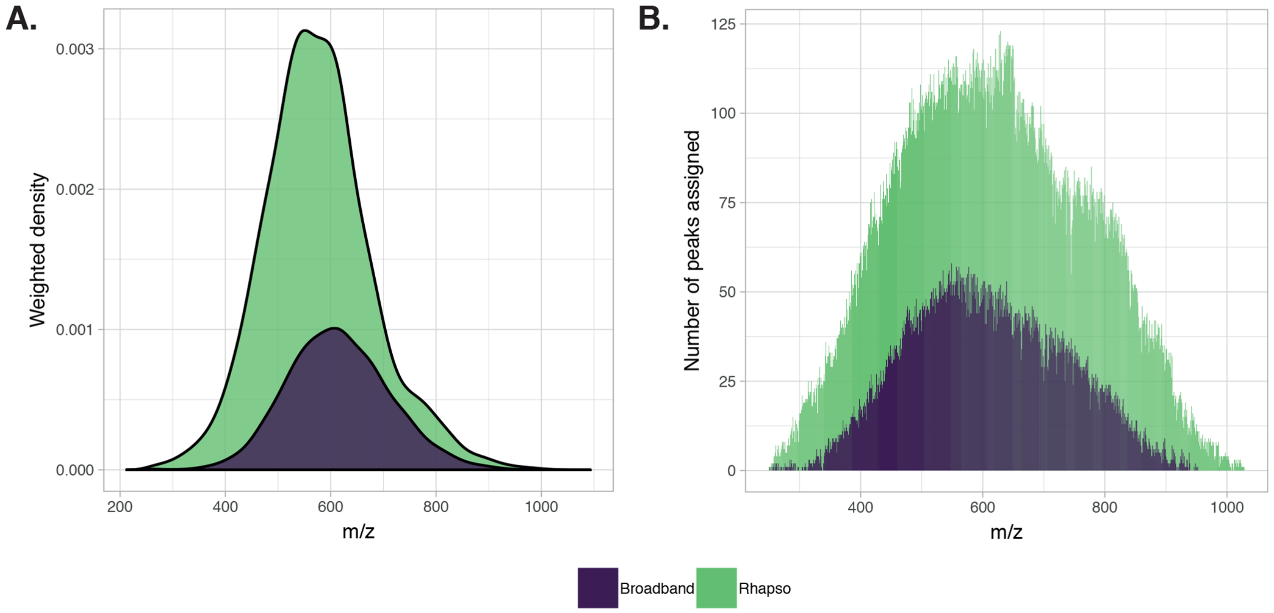


Figure 7. (A) Density plot based on all the m/z measured in both broadband and stitching mode. The plot shows the broader peak distribution in stitching mode due to the increased number of peaks in the low intensity regions. (B) Comparison of the number of assignments per m/z width of 1 using both techniques.

In Figure 6B, the green central vertical line illustrates the m/z where Rhapso found it would be the best place to perform the stitch.

With thousands of peaks, comparing two mass spectra can be challenging, so, in Figure 7A, we have represented a density plot using all the m/z measured in each data set. The density was weighted by the intensity of each m/z. The figure clearly illustrate the higher complexity of the mass spectrum, the higher intensity, and also the broader distribution of the spectrum when using a spectral stitching method.

In Figure 7B, the m/z of each assigned peak was reduced to an integer, and we counted the number of peaks of each integer. The results were plotted as a bar plot and demonstrate the increase in number of peaks assigned after using Rhapso to stitch the segments. It is worth noting the similarity of the distribution with the density plot in Figure 7.

Figure 8 shows the evolution of the main molecular classes. The bar plot clearly demonstrates an increase in the number of peaks assigned for each class. There is a 2- to 3-fold increase in

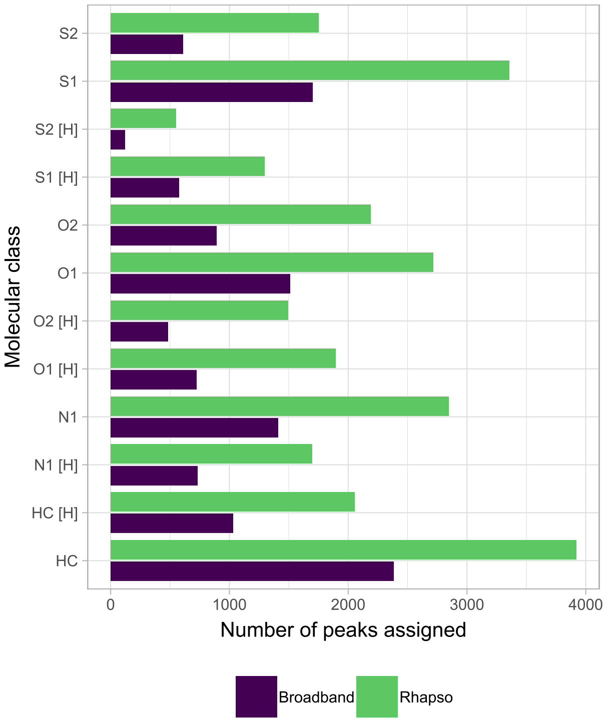


Figure 8. Number of peaks assigned for the prevalent molecular classes with the broadband mass spectrum (purple) and the mass spectrum obtained using segments stitched with Rhapso (green).

the number of peaks assigned in each class which is consistent with the higher number of peaks assigned.

While a higher number of peaks were assigned, this increase did not result in a higher mass error. Figure 9 illustrates how the mass error compares between broadband and stitching. Figure 9A shows the root-mean-square error along the m/z axis. The RMS error was calculated by pooling the mass error, in ppm, for each ppm width of 1 across the range. The data set processed with Rhapso has a systematically lower error and a wider m/z range as already illustrated before. Figure 9B illustrates the distribution of the mass error in ppm observed for all the peaks assigned. The horizontal lines respectively mark the 5%, 25%, 50%, 75%, and 95% quantiles within each violin plot. It is worth noting the much narrower distribution and median closer to 0 with the stitching method.

As presented by Palacio Lozano et al.,[36] Rhapso enabled for this vacuum residue sample a sharp increase in the number of peaks assigned (17k vs 50k) and led to a 2-fold decrease in the RMS mass error. An increase in the number of classes, highest DBE, and highest carbon number were also registered. We have also noted a sharp increase in the number of isotopic peaks assigned.

■ CONCLUSION

The acquisition process is currently very time-consuming but is expected to be automated in the future. Rhapso performs the spectral stitching of any complex spectrum acquired using selected ion monitoring windows within a few minutes. The method, implemented within a Shiny interface, allows the user to visualize each segment all along as well as check each step of the processing. Rhapso also allows analysts to speed up the acquisition process by permitting the reduction of the overlap between each window thanks to the intensity correction method. It also preserves the natural undulations of the spectra, characteristic of crude oils. The method also demonstrates that the correction allows the peaks' distribution to be closer to the ones observed in broadband mode. The spectral stitching method is crucial to increase the number of peaks observed and lower the mass error using existing instrumentation available to the users. However, any invest-

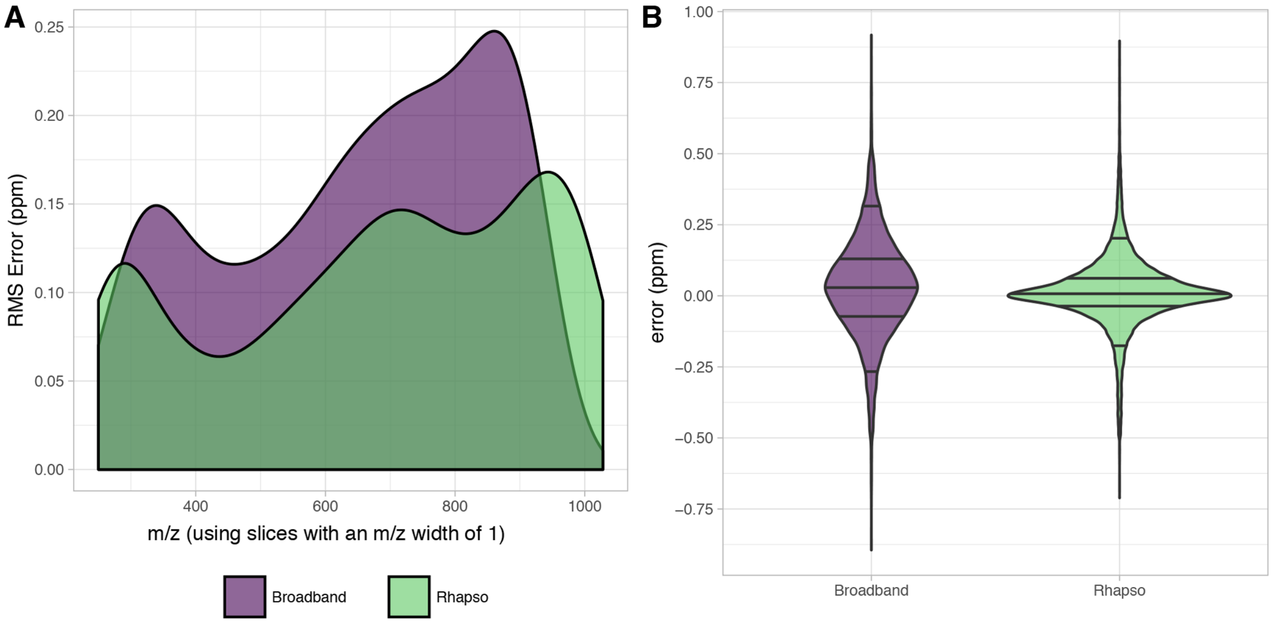


Figure 9. Root mean square error of the assigned peaks for each m/z width of 1 of the mass spectrum as (A) a density plot over the m/z range and as (B) a violin plot.

ment in more expensive equipment such as a more powerful magnet or more advanced FTICR spectrometer will also lead to further improvements. The results demonstrated not only a net increase in the number of peaks assigned but also an increase in the quality of those assignments. While the mass spectrum maintained a similar distribution, we showed there was an increase of peak density in the lower intensity regions. It is expected that this algorithm will enable a more extensive use of this technique, by relieving the user of a highly time-consuming step.

■ ASSOCIATED CONTENT

S Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.9b03846.

> Figure S1: Broadband mass spectrum of the South American vacuum residue sample, acquired using positive-ion APPI coupled to a 12 T FTICR mass spectrometer. Figure S2: Double bond equivalents (DBE) vs carbon number plots for two heteroatom classes, comparing the traditional broadband experiment and the use of Rhapso. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: [redacted].

ORCID

Remy Gavard: 0000-0001-5899-3058

Diana Catalina Palacio Lozano: 0000-0001-5315-5792

Mark P. Barrow: 0000-0002-6474-5357

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

R.G. 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. thank the Newton Fund award (Reference Number 275910721), Research Agreement No. 5211770 UIS-ICP, and COLCIENCIAS (Project No. FP44842-039-2015) for funding. 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 (Government of Spain). R.G. also thanks David Stranz (Sierra Analytics) for his valuable contributions.

■ REFERENCES

(1) Barrow, M. P. *Biofuels* **2010**, *1*, 651−655.

(2) Marshall, A. G.; Rodgers, R. P. *Acc. Chem. Res.* **2004**, *37*, 53−59.

(3) Barrow, M. P.; McDonnell, L. A.; Feng, X.; Walker, J.; Derrick, P. J. *Anal. Chem.* **2003**, *75*, 860−866.

(4) Ramírez, C. X.; Torres, J. E.; Palacio Lozano, D. C.; Arenas-Diaz, J. P.; Mejia-Ospino, E.; Kafarov, V.; Guzman, A.; Ancheyta, J. *Energy Fuels* **2017**, *31*, 13353−13363.

(5) Palacio Lozano, D. C.; Orrego-Ruiz, J. A.; Cabanzo Hernández, R.; Guerrero, J. E.; Mejía-Ospino, E. *Fuel* **2017**, *193*, 39−44.

(6) Cho, Y.; Witt, M.; Kim, Y. H.; Kim, S. *Anal. Chem.* **2012**, *84*, 8587−8594.

(7) Smith, E. A.; Lee, Y. J. *Energy Fuels* **2010**, *24*, 5190−5198.

(8) Tessarolo, N. S.; Silva, R. C.; Vanini, G.; Pinho, A.; Romão, W.; de Castro, E. V.; Azevedo, D. A. *Microchem. J.* **2014**, *117*, 68−76.

(9) Smith, D. F.; Rahimi, P.; Teclemariam, A.; Rodgers, R. P.; Marshall, A. G. *Energy Fuels* **2008**, *22*, 3118−3125.

(10) Noestheden, M. R.; Headley, J. V.; Peru, K. M.; Barrow, M. P.; Burton, L. L.; Sakuma, T.; Winkler, P.; Campbell, J. L. *Environ. Sci. Technol.* **2014**, *48*, 10264−10272.

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

(12) Marshall, A. G.; Rodgers, R. P. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 18090−18095.

(13) Hsu, C. S.; Hendrickson, C. L.; Rodgers, R. P.; McKenna, A. M.; Marshall, A. G. *J. Mass Spectrom.* **2011**, *46*, 337−343.

(14) Headley, J. V.; Peru, K. M.; Barrow, M. P. *Mass Spectrom. Rev.* **2016**, *35*, 311−328.

(15) Comisarow, M. B.; Marshall, A. G. *Chem. Phys. Lett.* **1974**, *25*, 282−283.

(16) Comisarow, M. B.; Marshall, A. G. *Can. J. Chem.* **1974**, *52*, 1997−1999.

(17) Comisarow, M. B.; Marshall, A. G. *Chem. Phys. Lett.* **1974**, *26*, 489−490.

(18) Amster, I. J. *J. Mass Spectrom.* **1996**, *31*, 1325−1337.

(19) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. *Mass Spectrom. Rev.* **1998**, *17*, 1−35.

(20) Barrow, M. P.; Burkitt, W. I.; Derrick, P. J. *Analyst* **2005**, *130*, 18.

(21) Schaub, T. M.; Hendrickson, C. L.; Horning, S.; Quinn, J. P.; Senko, M. W.; Marshall, A. G. *Anal. Chem.* **2008**, *80*, 3985−3990.

(22) Nikolaev, E. N.; Vladimirov, G.; Boldin, I. A. *Influences of non-neutral plasma effects on analytical characteristics of the top instruments in mass spectrometry for biological research*. American Institute of Physics Conference Series. 2013; pp 281−290.

(23) Nikolaev, E. N.; Kostyukevich, Y. I.; Vladimirov, G. N. *Mass Spectrom. Rev.* **2016**, *35*, 219−258.

(24) Limbach, P. A.; Grosshans, P. B.; Marshall, A. G. *Anal. Chem.* **1993**, *65*, 135−140.

(25) Purcell, J. M.; Merdrignac, I.; Rodgers, R. P.; Marshall, A. G.; Gauthier, T.; Guibard, I. *Energy Fuels* **2010**, *24*, 2257−2265.

(26) Zhang, L.-K.; Rempel, D.; Pramanik, B. N.; Gross, M. L. *Mass Spectrom. Rev.* **2005**, *24*, 286−309.

(27) Guan, S.; Marshall, A. G.; Scheppele, S. E. *Anal. Chem.* **1996**, *68*, 46−71.

(28) Senko, M. W.; Hendrickson, C. L.; Emmett, M. R.; Shi, S. D.; Marshall, A. G. *J. Am. Soc. Mass Spectrom.* **1997**, *8*, 970−976.

(29) Gaspar, A.; Schrader, W. *Rapid Commun. Mass Spectrom.* **2012**, *26*, 1047−1052.

(30) Krajewski, L. C.; Rodgers, R. P.; Marshall, A. G. *Anal. Chem.* **2017**, *89*, 11318−11324.

(31) Southam, A. D.; Payne, T. G.; Cooper, H. J.; Arvanitis, T. N.; Viant, M. R. *Anal. Chem.* **2007**, *79*, 4595−4602.

(32) Weber, R. J. M.; Southam, A. D.; Sommer, U.; Viant, M. R. *Anal. Chem.* **2011**, *83*, 3737−3743.

(33) Southam, A. D.; Weber, R. J.; Engel, J.; Jones, M. R.; Viant, M. R. *Nat. Protoc.* **2017**, *12*, 310−328.

(34) Rodgers, R. P.; Hughey, C. A.; Marshall, A. G. *Past, Present, and Future of Environmental Fourier Transform Ion Cyclotron Resonance Mass Spectrometry*. 2002.

(35) Zabrouskov, V.; Senko, M. *Direct Analysis of the Polar Fraction of Heavy Petroleum Crude Oil using a Linear Ion Trap/FTICR Hybrid Mass Spectrometer*. 2005.

(36) 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. *Chem. Sci.* **2019**, *10*, 6966−6978.

(37) Kilgour, D. P.; Van Orden, S. L. *Rapid Commun. Mass Spectrom.* **2015**, *29*, 1009−1018.

(38) Cleveland, W. S.; Grosse, E.; Shyu, W. Local regression models. *Statistical models in S*; Chambers, J. M., Hastie, T. J., Eds.; Chapman & Hall: 1992; pp 309−376.

Supporting information for: Rhapso: Automatic stitching of mass segments from Fourier transform ion cyclotron resonance mass spectra

Remy Gavard,[†] Diana Catalina Palacio Lozano,[‡] Alexander Guzman,[¶] 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, Instituto Colombiano del Petróleo, Piedecuesta, 681011, Colombia, Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom, and Department of Economics & Business, Universitat Pompeu Fabra, Barcelona, 08005, Spain

E-mail: [redacted]

[*] 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

[¶] Instituto Colombiano del Petróleo, Piedecuesta, 681011, Colombia

[§] Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom

[‖] Department of Economics & Business, Universitat Pompeu Fabra, Barcelona, 08005, Spain

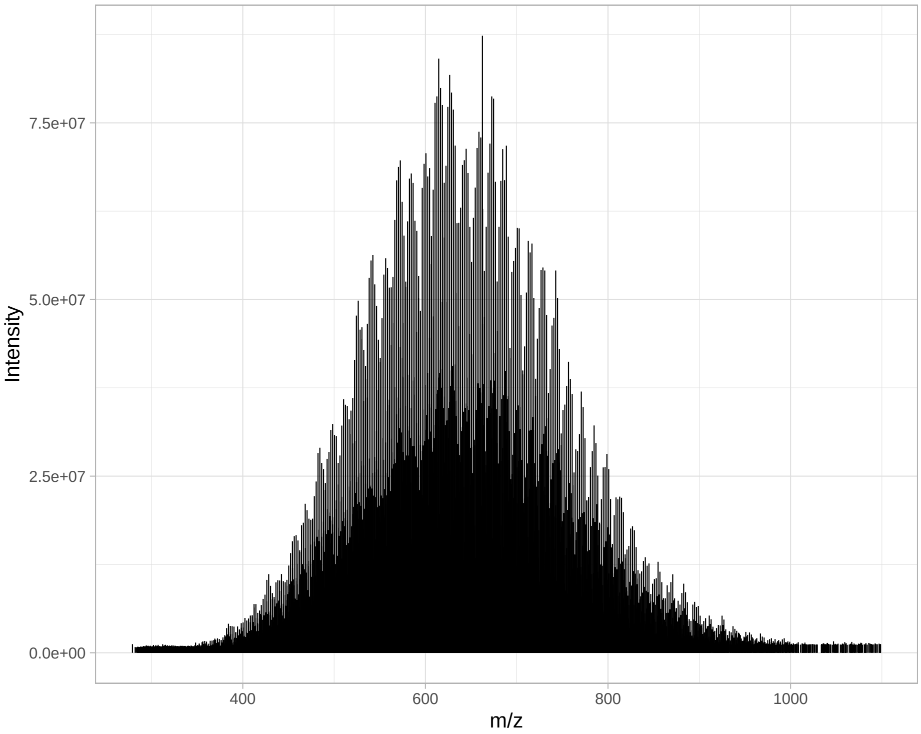


Figure S1: Broadband mass spectrum of the South American vacuum residue sample, acquired using positive-ion APPI coupled to a 12 *T* FTICR mass spectrometer.

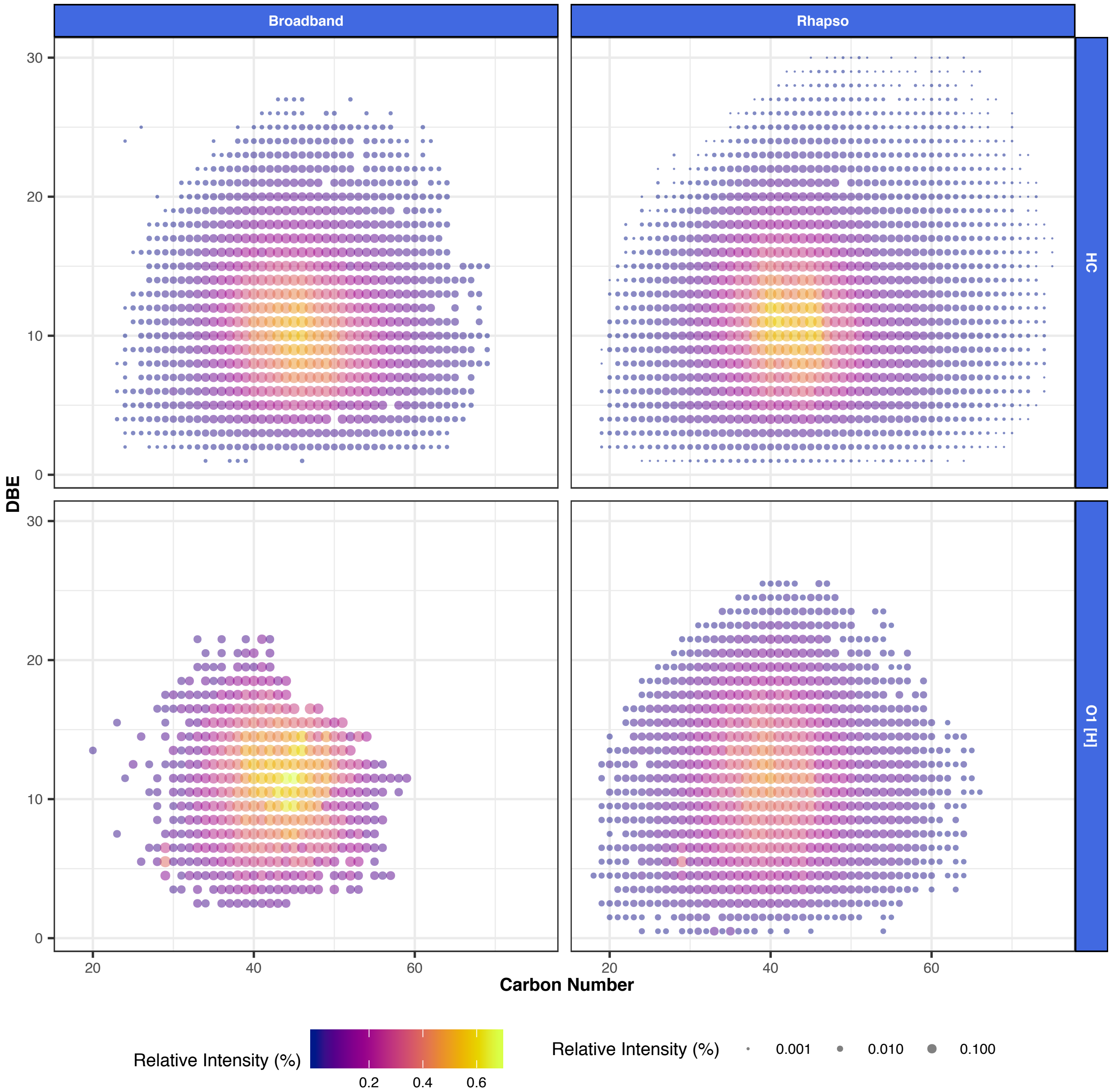


Figure S2: Double bond equivalents (DBE) vs. carbon number of the HC and O_1 heteroatom classes, produced using traditional broadband experiment and using Rhapso. The compositional space has been extended using Rhapso, as both DBE and carbon number ranges have increased.

Chapter 5

KairosMS: New solution to process complex mixture data analyzed by hyphenated - ultra-high-resolution mass spectrometry

5.1 Context

In this chapter, the challenges posed by the analysis of complex mixtures using ultra-high resolution mass spectrometry were addressed. In order to gain information regarding the structure of the molecules, it is possible to couple chromatography with ultra-high resolution mass spectrometry (UHRMS). The UHRMS is crucial to resolve all the co-eluting components and resolve each extracted ion chromatogram (EIC) to observe isomers. This method leads to large datasets and the data analysis is currently extremely laborious. The current data processing methods rely on manually segmenting the elution into as many time slices as necessary. The signal of each time slices is summed and peak list is extracted and processed individually. In order to for the method to become viable better tools are needed. KairosMS removes the need to manually divide the data and perform as many molecular assignments and plots as there are segments. KairosMS imports the data as a masslist, processes it and returns a single peak list which is used to obtain the molecular assignments. The assignments are then incorporated by KairosMS with the original data. A wide range of interactive visualisations have been implemented to assist the researcher in the exploration. It is also possible to process several samples and visualise them side-by-side within the same plots, making the comparison between samples extremely straightforward. KairosMS not only decreases the time necessary to analyse complex mixtures with hyphenated UHRMS, it also enables a more accurate analysis as no time information is lost. It also affords the ability to rapidly screen EICs and observe isomeric contributions.

This chapter was submitted as an article to *Analytical Chemistry.* The data to develop this method was sourced from various researchers across different research groups. The patterns and algorithm theory were discovered by the author. The first implementation of the algorithm was done in R by Hugh E. Jones, Masters student at the time. Initial visualisations were performed by Hugh E. Jones. Hugh E. Jones' code was then adapted and expanded by the author in a Shiny interactive interface. Extensive testing and feature requests were provided by Diana Catalina Palacio Lozano and Mary J. Thomas. The supervisors David Rossell, Simon E. F. Spencer and Mark P. Barrow provided guidance and help throughout the process. KairosMS is expected to be used in numerous upcoming publications. The manuscript was written by the author with corrections from Diana Catalina Palacio Lozano and Hugh E. Jones.

5.2 Publication

analytical chemistry

pubs.acs.org/ac Article

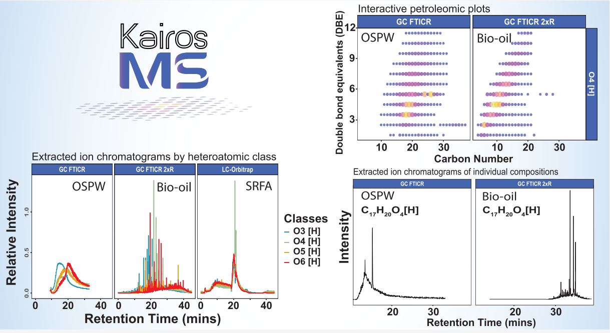
KairosMS: A New Solution for the Processing of Hyphenated Ultrahigh Resolution Mass Spectrometry Data

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

Cite This: *Anal. Chem.* 2020, 92, 3775−3786 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: The use of hyphenated Fourier transform mass spectrometry (FTMS) methods affords additional information about complex chemical mixtures. Coeluted components can be resolved thanks to the ultrahigh resolving power, which also allows extracted ion chromatograms (EICs) to be used for the observation of isomers. As such data sets can be large and data analyses laborious, improved tools are needed for data analyses and extraction of key information. The typical workflow for this type of data is based upon manually dividing the total ion chromatogram (TIC) into several windows 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. Through removal of the need to manually divide a data set into many time windows and analyze each one, a time-consuming workflow 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 incorporates 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, 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.



Complex mixtures such as petroleum, petroleum related samples, and dissolved organic matter (DOM) are among 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 ultrahigh 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 Orbitrap 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 determine the masses of thousands of ions with very high accuracy, but do 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

Received: November 10, 2019
Accepted: January 28, 2020
Published: January 28, 2020

ACS Publications © 2020 American Chemical Society

to keep pace with the instrumentation advances. The tools developed so far struggle to perform well with the ultrahigh 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 to create a data matrix with m/z and retention time (t_R) 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, etc. 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 ultrahigh resolution.

The m/z binning method is based upon the assumption 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 comprises peaks at given m/z values 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 ultrahigh resolution to be able to separate ions 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 ultrahigh 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 m/z error range of 10 parts per million (ppm) does not pose the same challenges as techniques yielding m/z 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 of different molecules. The majority of the software cited earlier was 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.g. 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 method works best in the absence of noise, and because 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 has not 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 has not 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, to address these issues, we 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 allow 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

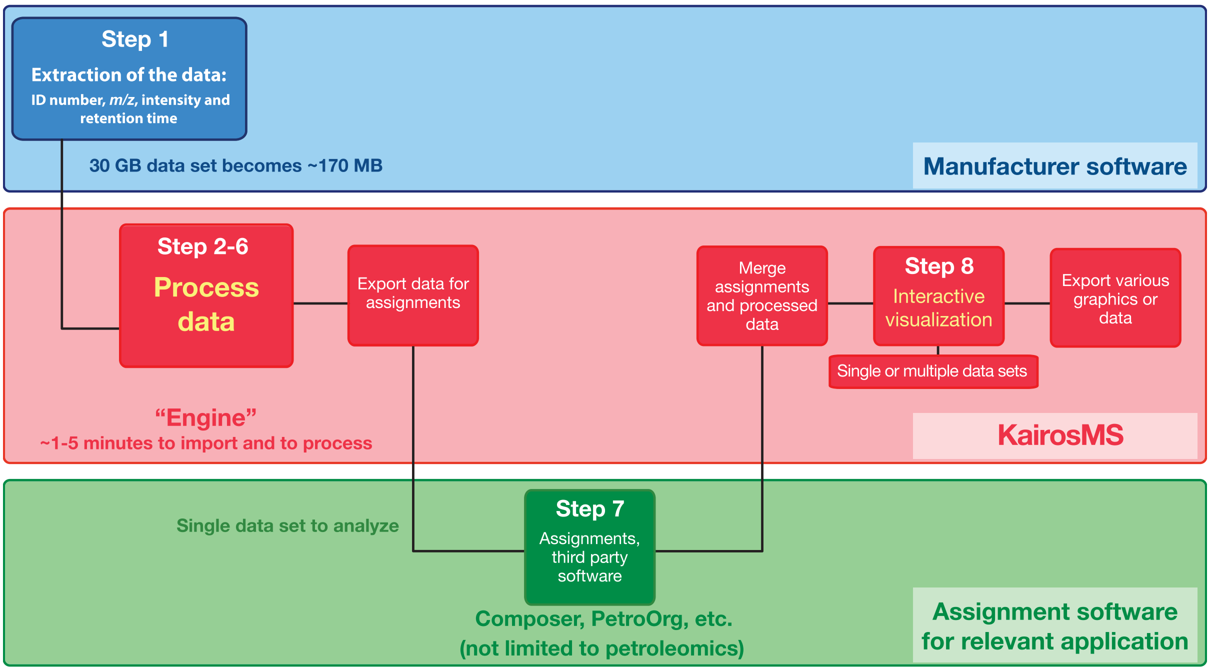

Step 1
Extraction of the data:
ID number, *m/z*, intensity and retention time
30 GB data set becomes ~170 MB
Manufacturer software
Step 2-6
Process data
Export data for assignments
Merge assignments and processed data
Step 8
Interactive visualization
Export various graphics or data
Single or multiple data sets
"Engine"
~1-5 minutes to import and to process
KairosMS
Single data set to analyze
Step 7
Assignments, third party software
Composer, PetroOrg, etc.
(not limited to petroleomics)
Assignment software for relevant application

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.

exploration. Standard figures such as double bond equivalent (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. A 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, United States). The extracts were then methylated using BF_3-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), desalted, and concentrated by solid phase extraction.[25] The marine sample is hereafter referred as Marine DOM. The SRFA sample was diluted with ultrapure water and 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 μL was injected into a 30 m DB-5 column (0.25 mmID, 0.25 μm).

Instrumentation. KairosMS capabilities and visualization tools were explored for the analysis of six hyphenated data sets acquired with different ultrahigh resolution mass spectrometers. The experimental parameters and instrumentation are briefly described as follows:

GC-APCI-FTICR MS: The OSPW, G1, and G2 samples were analyzed using a 7890A GC instrument (Agilent Technologies, Santa Clara, California, United States) connected to an atmospheric pressure chemical ionization (APCI) source (Bruker Daltonik GmbH, Bremen, Germany) in positive mode which was itself used for the 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 °C and increased at a rate of 20 °C min^{-1} until a final temperature of 280 °C was reached and held for 20 min. Broadband mass spectra in magnitude mode were acquired, and 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% over 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 Bruker 450 GC instrument (Bruker Daltonik GmbH, Bremen, Germany) 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 those of a 15 T instrument when operated at similar detection conditions of ω. The oven temperature was initialized at 60 °C and increased at a rate of 6 °C min^{-1} until a final temperature of 300 °C was reached. The oven was then maintained at 300 °C 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 was applied which removes 95% of the data points due 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 in which 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 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 of 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 spectrum 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 $i = 1, ..., n$, 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 $T_i \in t_1, t_2, t_3, ..., t_m$ and m is the number of scans.

Step 2: Trim the TIC. In chromatography, the beginning of acquisition often corresponds 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 file sizes. Part of it can be retained if the user wishes to later 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 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 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 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 species 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 eq 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 eq 2, derived from eq 1, using instrument-specific values A and B provided by the user. For a Bruker FTICR MS data set, 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 jth EIC. For each peak i within EIC j, we define the frequency shift $\tilde{F}_i = F_i - \underline{F}_j$. For each time $t \in \{t_1, ..., t_m\}$ we compute the mean frequency shift $S(t)$

$$S(t) = \left[\sum_{i:T_i=t} \tilde{F}_i\right] \Big/ \left[\sum_{i=1}^{n} I(T_i = t)\right] \tag{3}$$

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 of the M_i values 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 $S\widetilde{(t)}$ similarly to $S(t)$, but based on $\widetilde{F_i}$ instead of F_i. The final frequencies are obtained by calculating $F_i^* = \widetilde{F_i} + S\widetilde{(t)} - \min(S\widetilde{(t)}: t \in t_1, \ldots, t_m)$ and the corresponding m/z was calculated. The updated m/z are then used in Step 6.

If the A and B coefficients from eq 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 two 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 EIC 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 EIC were merged with the data for the peaks within the EIC and stored as an 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 of 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 and 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 through 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.

To process the OSPW data set, the TIC between 0 and 9 min 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 attenuates the space-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

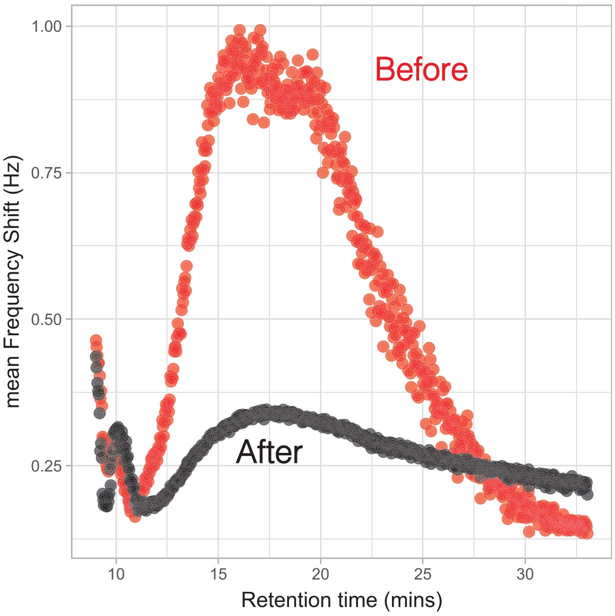


Figure 2. Average frequency shift in Hz within EICs for each retention time before recalibration (red) and after (black) for the OSPW sample.

ppm. After recalibration, the RMS error decreased to 0.4 ppm (Figure 3).

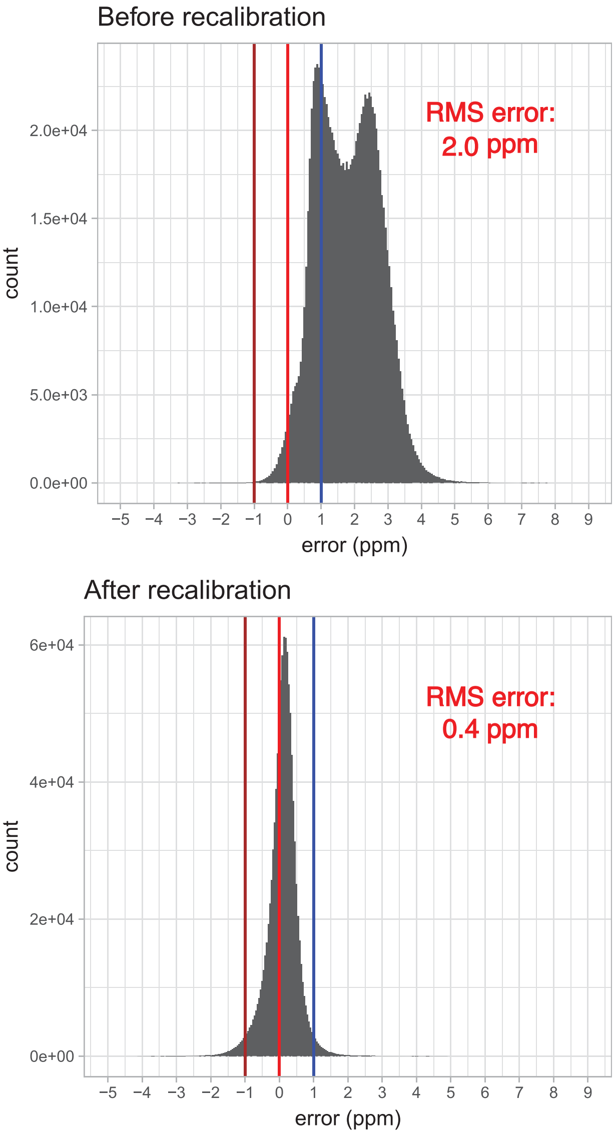


Figure 3. Histogram of the mass error in ppm of each peak for each EIC with the assigned m/z.

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. A total of 6540 distinct EICs were 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.

Because 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.

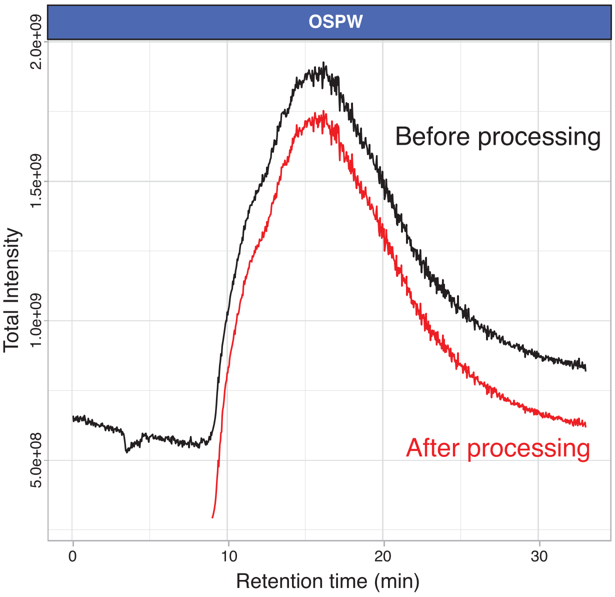


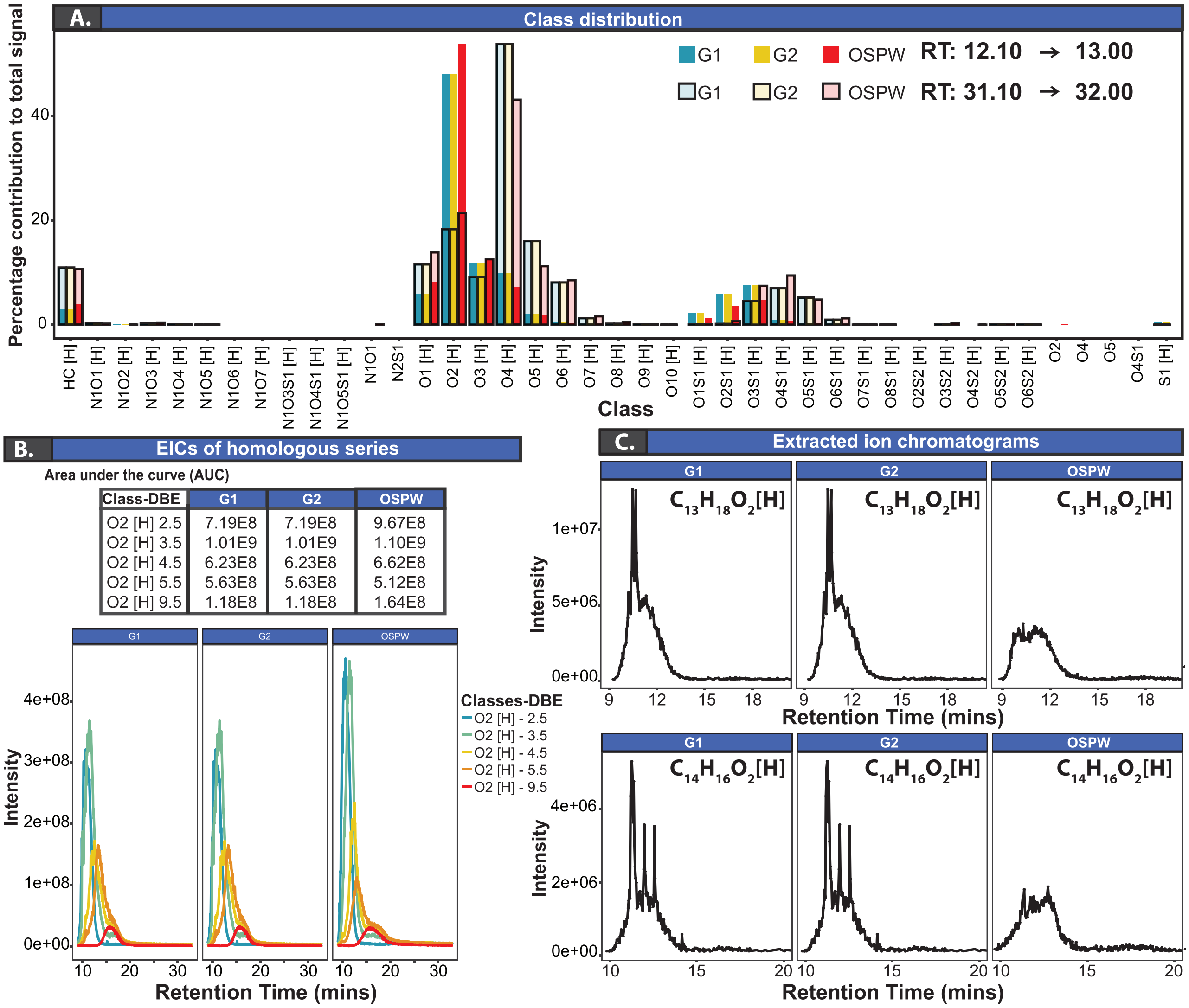
Figure 4. Comparison of TICs before (top line, black) and after (bottom line, red) processing.

Once the assignments were performed using Composer, the molecular composition for each peak assigned was reattributed to their respective EIC, and a special data frame format (tibble) was used to store all of 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 5 min. 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 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.



Class-DBE	G1	G2	OSPW
O2 [H] 2.5	7.19E8	7.19E8	9.67E8
O2 [H] 3.5	1.01E9	1.01E9	1.10E9
O2 [H] 4.5	6.23E8	6.23E8	6.62E8
O2 [H] 5.5	5.63E8	5.63E8	5.12E8
O2 [H] 9.5	1.18E8	1.18E8	1.64E8

Figure 5. (A) Heteroatom class distribution for retention times 12.10−12 min and 31.10−32 min. (B) EICs of selected homologous series from the $O_2[H]$ class. (C) EICs of selected molecular composition for the OSPW and two groundwater samples (G1 and G2).

- 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 EIC 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 data

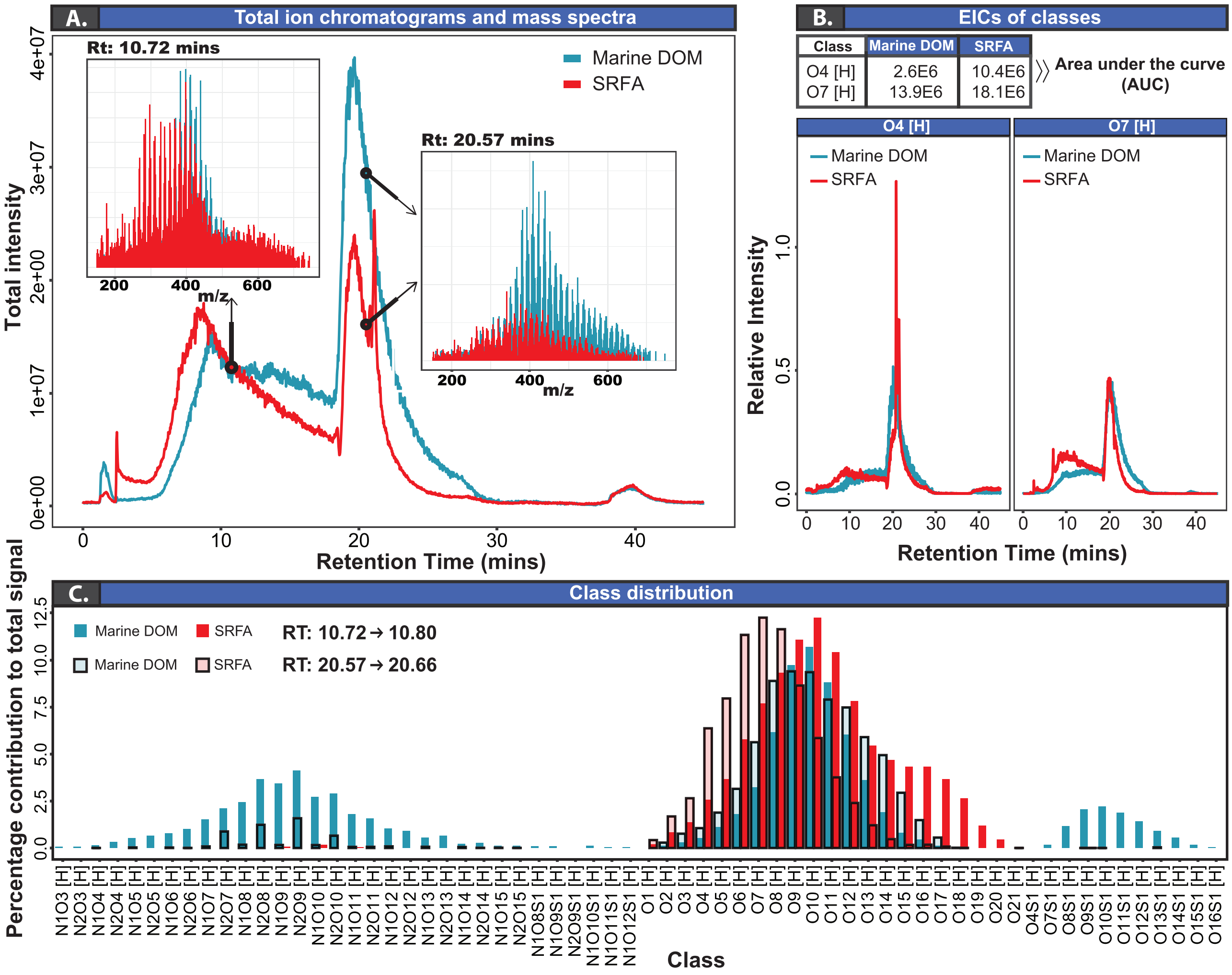


Figure 6. (A) Total ion chromatogram (TIC) with the mass spectra for retention times 10.72 min and 20.57 min. (B) EICs for the heteroatom classes $O_4[H]$ and $O_7[H]$ with associated AUC. (C) Heteroatom class distribution for retention times 10.72−10.80 min and 20.57−20.66 min for the SRFA and Marine DOM samples.

sets,[24,25] we extended the capabilities of KairosMS so that one can compare several samples after they have been processed due to the level of detail of information retained during the processing. No limits have been set to the number of data sets to compare, but using more files requires 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 class 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 toward 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 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.

Because we observed differences in the O_2 [H] class, we 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 of 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.

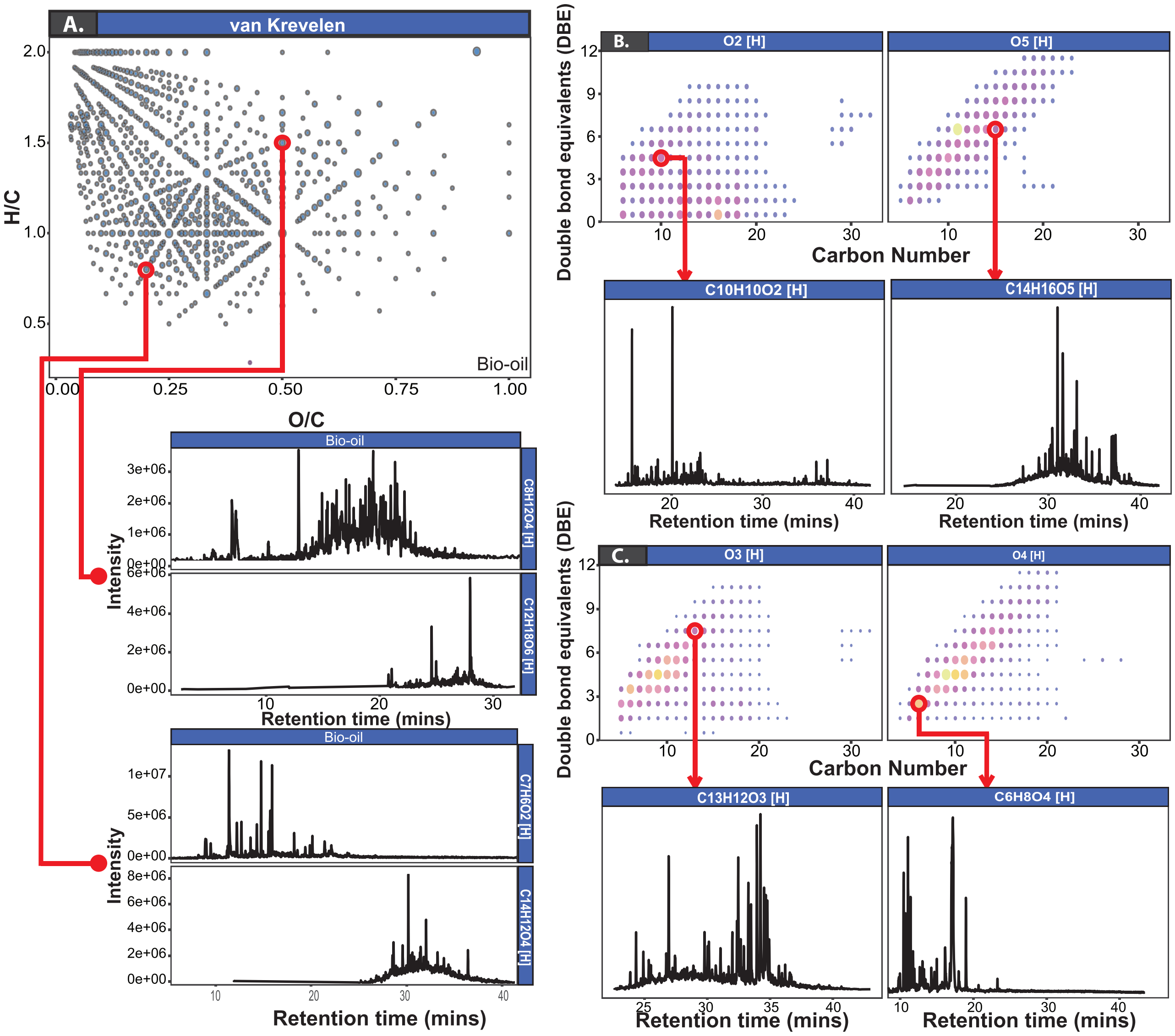


Figure 7. (A) van Krevelen diagram with EICs displayed for two different points on the plot. (B) DBE vs carbon number of the O_2[H] and O_5[H] heteroatom classes with associated EICs for two data points on the plot. (C) DBE vs carbon number of the O_3[H] and O_4[H] heteroatom classes with associated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

After screening EICs corresponding to the O_2[H] class, noticeable differences were found 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 show 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 Marine DOM and SRFA samples. Each data set 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 C 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 to highlight the difference of composition at any stage of the acquisition. In contrast to the GC experiment, the components eluting from this LC column at higher retention time correspond to species with lower oxygen containing species. A 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 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 was analyzed using a solariX 2xR FTICR MS. The 2*ω* 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 as compared to conventional 2*ω* instruments. For gas chromatography, acquisition 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, where each data point corresponds to a single composition, van Krevelen data points can correspond to multiple EICs of different 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 note 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. 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.

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 of the EICs with specific features such as heteroatom class, carbon number, and DBE. This allows a researcher to quickly determine differences between elution profiles at the molecular level.

Finally, by using the intensity information on all of the assigned EICs, the calculation of the elemental contribution within each sample can be swiftly obtained, and the percentages for each elements can be 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 experiments previously relied on long and laborious manual work. The typical workflow 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 are first extracted as a mass list to reduce computing time, relying on peak picking and centroid detection of 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 data sets such as biomolecules. The method could be further improved by implementing prior knowledge about one or more peaks. KairosMS demonstrated its abilities over a wide range of samples (petroleum, environmental, dissolved organic matter results, and biomolecules) from different types of analyzers and types of chromatography, turning hyphenated ultrahigh 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 the 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. Among the features for comparing many complex data sets, 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.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.9b05113.

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; Figure S5: Elution of DBE series (homologous series) comprising the O_2[H] class; Figure S6: Percentage of contribution to the total signal for all of the classes identified in the SRFA and Marine DOM samples; Figure S7: Elemental contributions for the samples SRFA and Marine DOM based on all of 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 (PDF)

Video showing interactive data visualization using KairosMS (MP4)

AUTHOR INFORMATION

Corresponding Author

Mark P. Barrow − *Department of Chemistry, University of Warwick, Coventry CV4 7AL, United Kingdom;* orcid.org/0000-0002-6474-5357; Email: [redacted]

Authors

Remy Gavard − *MAS CDT, University of Warwick, Coventry CV4 7AL, United Kingdom;* orcid.org/0000-0001-5899-3058

Hugh E. Jones − *Department of Chemistry, University of Warwick, Coventry CV4 7AL, United Kingdom*

Diana Catalina Palacio Lozano − *Department of Chemistry, University of Warwick, Coventry CV4 7AL, United Kingdom;* orcid.org/0000-0001-5315-5792

Mary J. Thomas − *MAS CDT, University of Warwick, Coventry CV4 7AL, United Kingdom;* orcid.org/0000-0002-6744-5413

David Rossell − *Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom; Department of Economics & Business, Universitat Pompeu Fabra, Barcelona 08005, Spain*

Simon E. F. Spencer − *Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.9b05113

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

R.G., H.E.J., and M.J.T. thank 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. is thankful for 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.

REFERENCES

(1) Hertkorn, N.; Ruecker, C.; Meringer, M.; Gugisch, R.; Frommberger, M.; Perdue, E. M.; Witt, M.; Schmitt-Kopplin, P. *Anal. Bioanal. Chem.* **2007**, *389*, 1311−1327.

(2) Barrow, M. P. *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, NY, 2007.

(4) Marshall, A. G.; Rodgers, R. P. *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. *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*. DOI: 10.1038/ncomms4804

(7) Stubbins, A.; Dittmar, T. *Mar. Chem.* **2015**, *177*, 318−324.

(8) Headley, J. V.; Peru, K. M.; Barrow, M. P. *Mass Spectrom. Rev.* **2016**, *35*, 311−328.

(9) Comisarow, M. B.; Marshall, A. G. *Chem. Phys. Lett.* **1974**, *25*, 282−283.

(10) Comisarow, M. B.; Marshall, A. G. *Can. J. Chem.* **1974**, *52*, 1997−1999.

(11) Comisarow, M. B.; Marshall, A. G. *Chem. Phys. Lett.* **1974**, *26*, 489−490.

(12) Amster, I. J. *J. Mass Spectrom.* **1996**, *31*, 1325−1337.

(13) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. *Mass Spectrom. Rev.* **1998**, *17*, 1−35.

(14) Barrow, M. P.; Burkitt, W. I.; Derrick, P. J. *Analyst* **2005**, *130*, 18.

(15) Koch, B. P.; Ludwichowski, K. U.; Kattner, G.; Dittmar, T.; Witt, M. *Mar. Chem.* **2008**, *111*, 233−241.

(16) Dittmar, T.; Stubbins, A. *Treatise Geochemistry*, 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. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 459−462.

(18) Zhurov, K. O.; Kozhinov, A. N.; Tsybin, Y. O. *Energy Fuels* **2013**, *27*, 2974−2983.

(19) Hawkes, J. A.; Dittmar, T.; Patriarca, C.; Tranvik, L.; Bergquist, J. *Anal. Chem.* **2016**, *88*, 7698−7704.

(20) Cho, E.; Witt, M.; Hur, M.; Jung, M. J.; Kim, S. *Anal. Chem.* **2017**, *89*, 12101−12107.

(21) Krajewski, L. C.; Rodgers, R. P.; Marshall, A. G. *Anal. Chem.* **2017**, *89*, 11318.

(22) Smith, D. F.; Podgorski, D. C.; Rodgers, R. P.; Blakney, G. T.; Hendrickson, C. L. *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. *Chem. Sci.* **2019**, *10*, 6966−6978.

(24) Barrow, M. P.; Peru, K. M.; Headley, J. V. *Anal. Chem.* **2014**, *86*, 8281−8288.

(25) Patriarca, C.; Bergquist, J.; Sjoberg, P. J. R.; Tranvik, L.; Hawkes, J. A. *Environ. Sci. Technol.* **2018**, *52*, 2091−2099 PMID: 29241333..

(26) Kim, S.; Kim, D.; Kim, S.; Son, S.; Jung, M. J. *Anal. Chem.* **2019**, *91*, 7690−7697.

(27) Benigni, P.; Thompson, C. J.; Ridgeway, M. E.; Park, M. A.; Fernandez-Lima, F. *Anal. Chem.* **2015**, *87*, 4321−4325.

(28) Wenig, P.; Odermatt, J. OpenChrom: a cross-platform open source software for the mass spectrometric analysis of chromatographic data. *BMC Bioinf.* **2010**, *11*. DOI: 10.1186/1471-2105-11-405

(29) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. *Nat. Methods* **2015**, *12*, 523.

(30) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.

(31) Katajamaa, M.; Miettinen, J.; Orešič, M. *Bioinformatics* **2006**, *22*, 634−636.

(32) Tolstikov, V. V.; Lommen, A.; Nakanishi, K.; Tanaka, N.; Fiehn, O. *Anal. Chem.* **2003**, *75*, 6737−6740.

(33) Baran, R.; Kochi, H.; Saito, N.; Suematsu, M.; Soga, T.; Nishioka, T.; Robert, M.; Tomita, M. *BMC Bioinf.* **2006**, *7*, 530.

(34) Idborg-Björkman, H.; Edlund, P. O.; Kvalheim, O. M.; Schuppe-Koistinen, I.; Jacobsson, S. P. *Anal. Chem.* **2003**, *75*, 4784−4792.

(35) Marshall, A. G.; Rodgers, R. P. *Acc. Chem. Res.* **2004**, *37*, 53−59.

(36) Hur, M.; Oh, H. B.; Kim, S. *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. *Eur. J. Mass Spectrom.* **2017**, *23*, 28−39.

(38) Schwemer, T.; Rüger, C. P.; Sklorz, M.; Zimmermann, R. *Anal. Chem.* **2015**, *87*, 11957−11961.

(39) Tautenhahn, R.; Böttcher, C.; Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinf.* **2008**,. DOI: 10.1186/1471-2105-9-504

(40) Åberg, K. M.; Torgrip, R. J.; Kolmert, J.; Schuppe-Koistinen, I.; Lindberg, J. *J. Chromatogr. A* **2008**, *1192*, 139−146.

(41) Gavard, R.; Rossell, D.; Spencer, S. E. F.; Barrow, M. P. *Anal. Chem.* **2017**, *89*, 11383−11390.

(42) Hughey, C. A.; Rodgers, R. P.; Marshall, A. G.; Qian, K.; Robbins, W. K. *Org. Geochem.* **2002**, *33*, 743−759.

(43) Stanford, L. A.; Kim, S.; Rodgers, R. P.; Marshall, A. G. *Energy Fuels* **2006**, *20*, 1664−1673.

(44) Barrow, M. P.; Headley, J. V.; Peru, K. M.; Derrick, P. J. *Energy Fuels* **2009**, *23*, 2592−2599.

(45) Van Krevelen, D. *Fuel* **1950**, *29*, 269−284.

(46) Kim, S.; Kramer, R. W.; Hatcher, P. G. *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. *Mar. Chem.* **2014**, *161*, 14−19.

(48) Hawkes, J. A.; Hansen, C. T.; Goldhammer, T.; Bach, W.; Dittmar, T. *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. *Fuel* **2020**, *259*, 116085.

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

(51) Chang, W.; Cheng, J.; Allaire, J.; Xie, Y.; McPherson, J. *shiny: Web application framework for R [Computer software]*, R package version 1.0.0; 2017. https://cran.r-project.org/web/packages/shiny/index.html.

(52) Zhurov, K. O.; Kozhinov, A. N.; Fornelli, L.; Tsybin, Y. O. *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. *Int. J. Mass Spectrom. Ion Processes* **1983**, *54*, 189−199.

(54) Barry, J. A.; Robichaud, G.; Muddiman, D. C. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 1137−1145.

(55) Cleveland, W. S.; Grosse, E.; Shyu, W. Local regression models. *Statistical models in S*; Chambers, J. M., Hastie, T. J.; 1992, pp 309−376.

(56) Wickham, H.; Francois, R.; Müller, K. Tibble: Simple Data Frames. 2018. https://cran.r-project.org/web/packages/tibble/index.html

Supporting information for: Supporting information for: 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: [redacted]

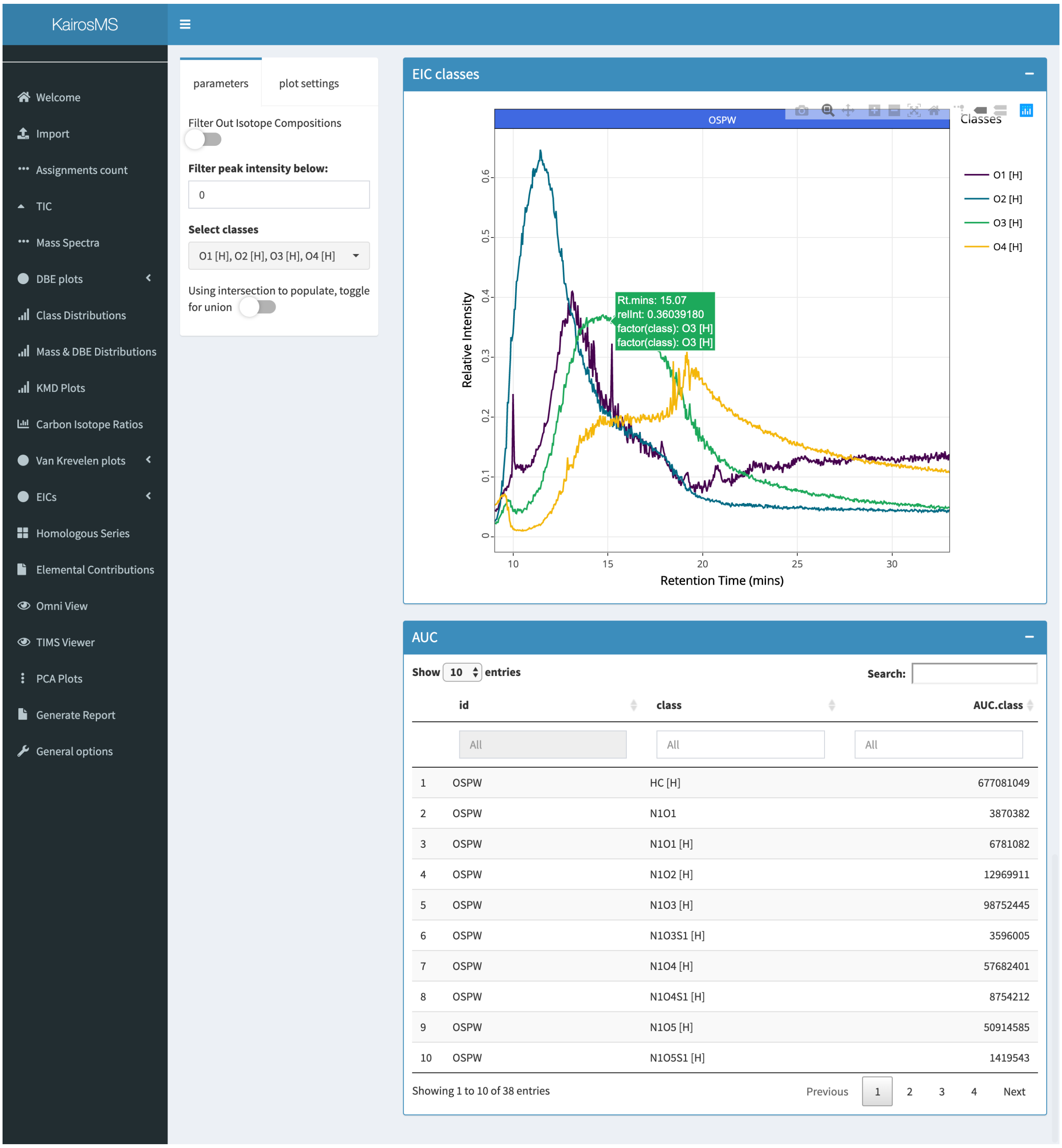
[*] To whom correspondence should be addressed

[†] 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



Filter Out Isotope Compositions

Filter peak intensity below: 0

Select classes: O1 [H], O2 [H], O3 [H], O4 [H]

Using intersection to populate, toggle for union

Show 10 **entries** **Search:**

	id	class	AUC.class
	All	All	All
1	OSPW	HC [H]	677081049
2	OSPW	N1O1	3870382
3	OSPW	N1O1 [H]	6781082
4	OSPW	N1O2 [H]	12969911
5	OSPW	N1O3 [H]	98752445
6	OSPW	N1O3S1 [H]	3596005
7	OSPW	N1O4 [H]	57682401
8	OSPW	N1O4S1 [H]	8754212
9	OSPW	N1O5 [H]	50914585
10	OSPW	N1O5S1 [H]	1419543

Showing 1 to 10 of 38 entries

Figure S1: Screenshot presenting KairosMS interface.

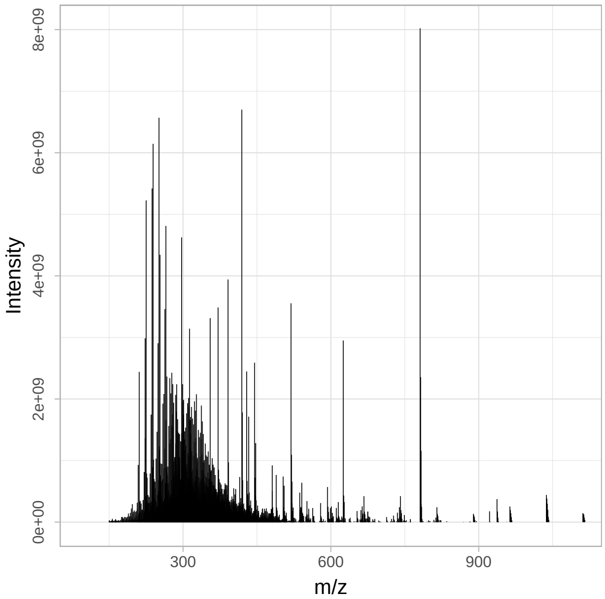


Figure S2: Mass spectrum created using the EICs extracted to be used for molecular assignments.

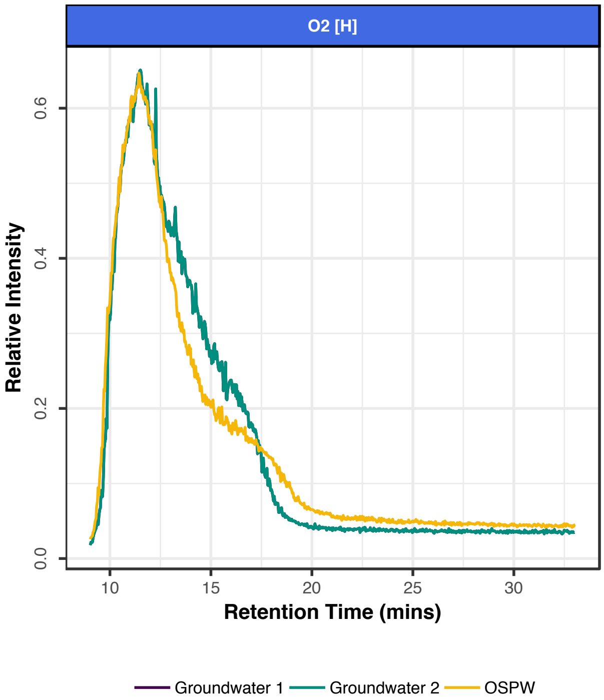


Figure S3: Comparison of the elution of the O_2[H] class contribution between an OSPW and two Groundwater samples using a scan by scan resolution.

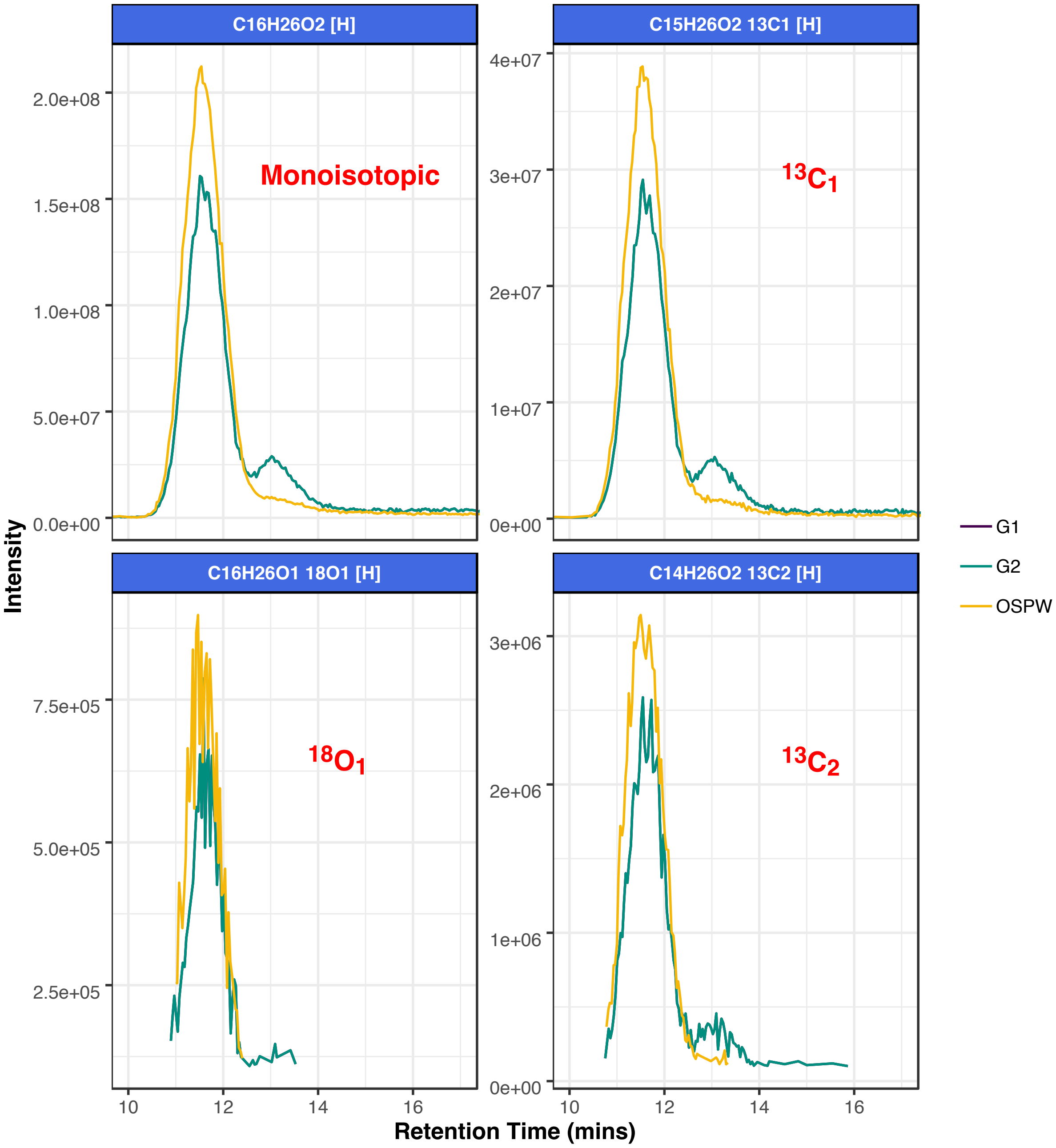


Figure S4: EICs for the monoisotopic form and isotopologues of $C_{16}H_{26}O_2$ [H] for the G1, G2 and OSPW samples (G1 and G2 perfectly overlap). The same retention for the isotopologues is further evidence for the compositional assignment.

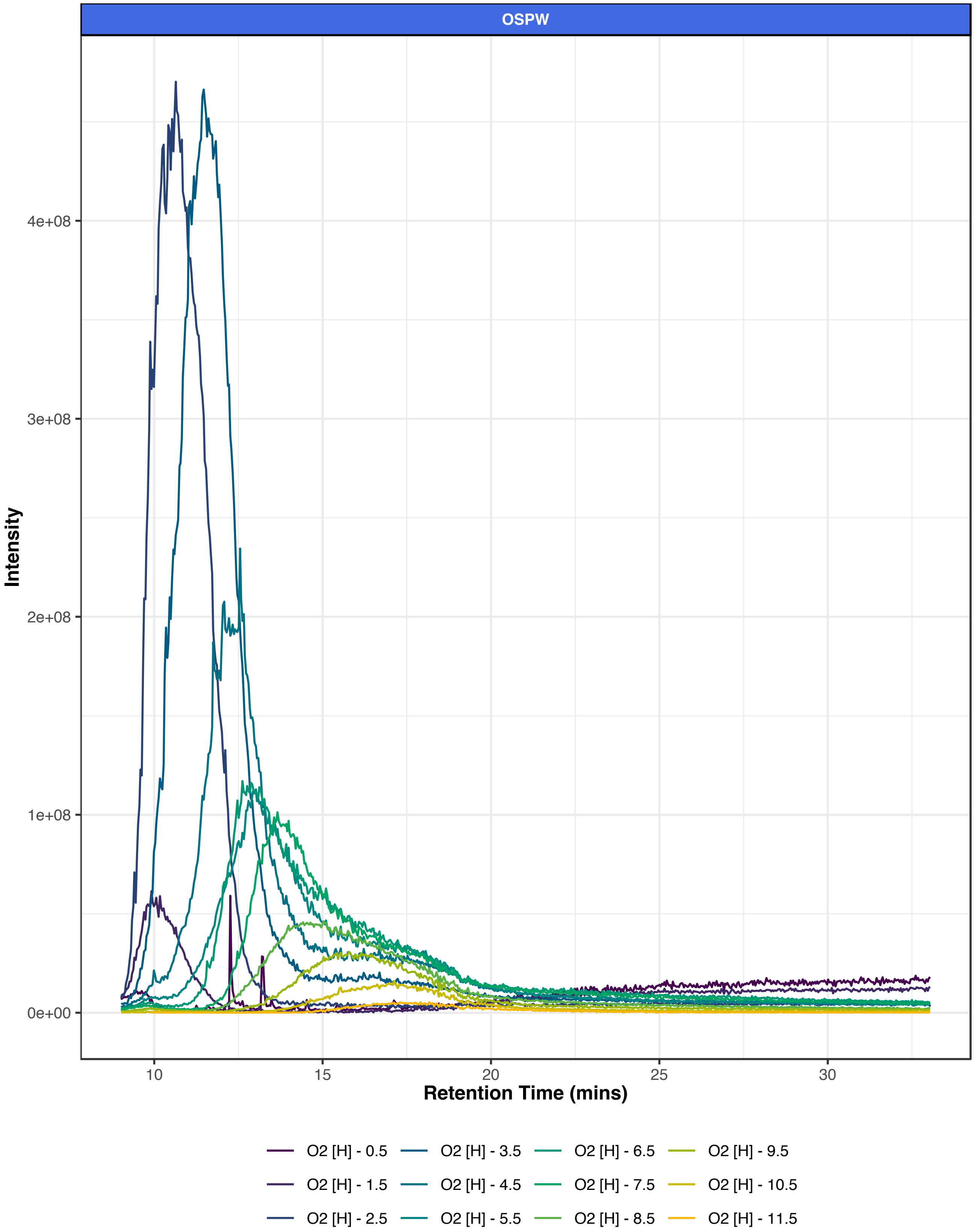


Figure S5: Elution of DBE series (homologous series) comprising the O_2[H] class.

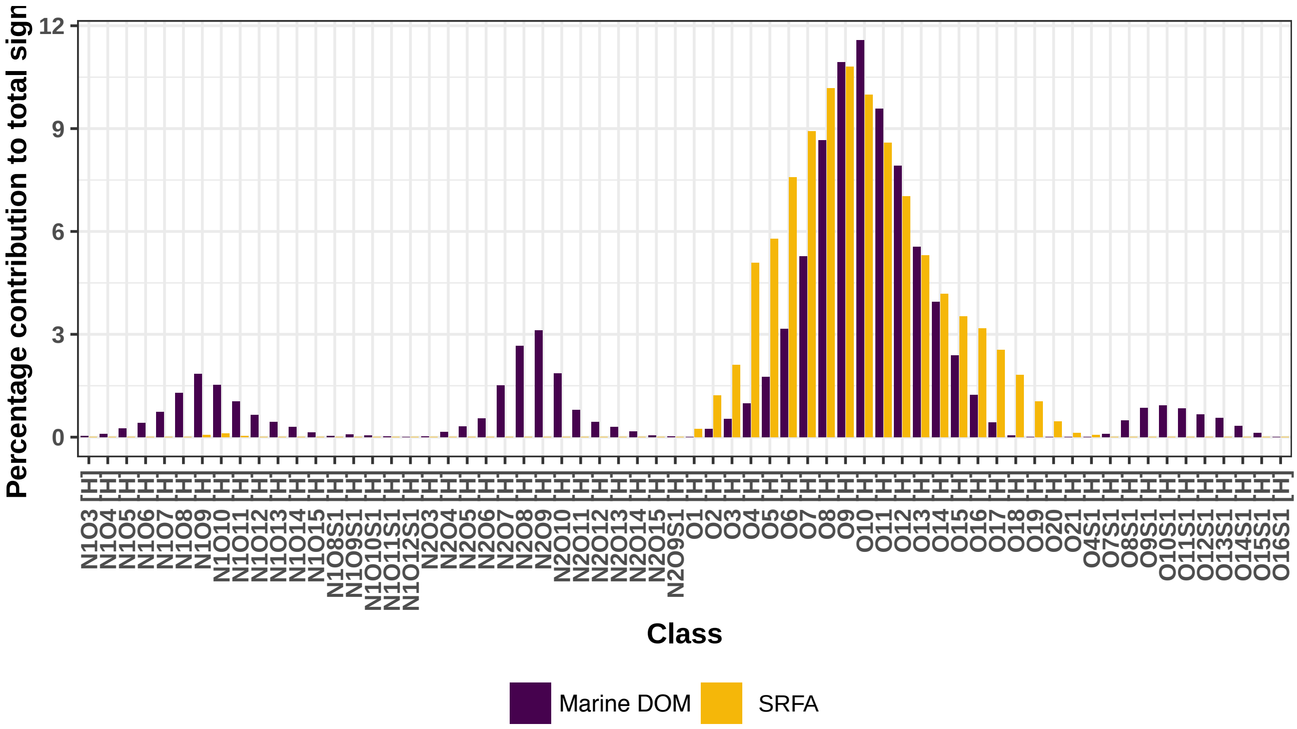


Figure S6: Percentage of contribution to the total signal, for all the classes identified in the SRFA and Marine DOM samples.

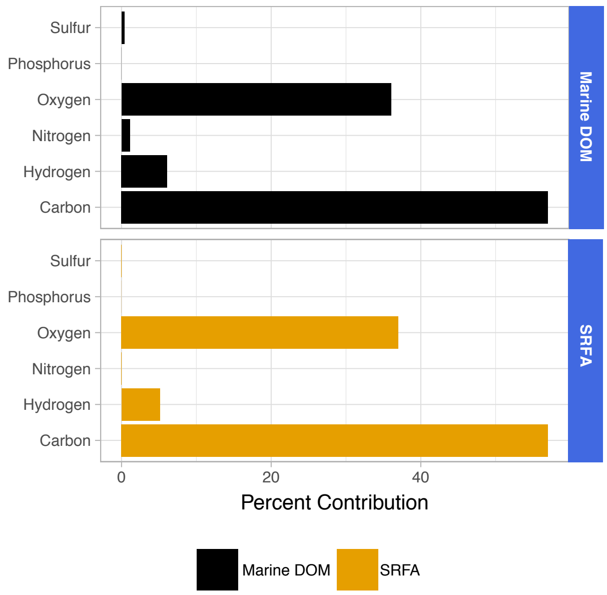


Figure S7: Elemental contributions for the samples SRFA and Marine DOM, based on all the assigned EICs.

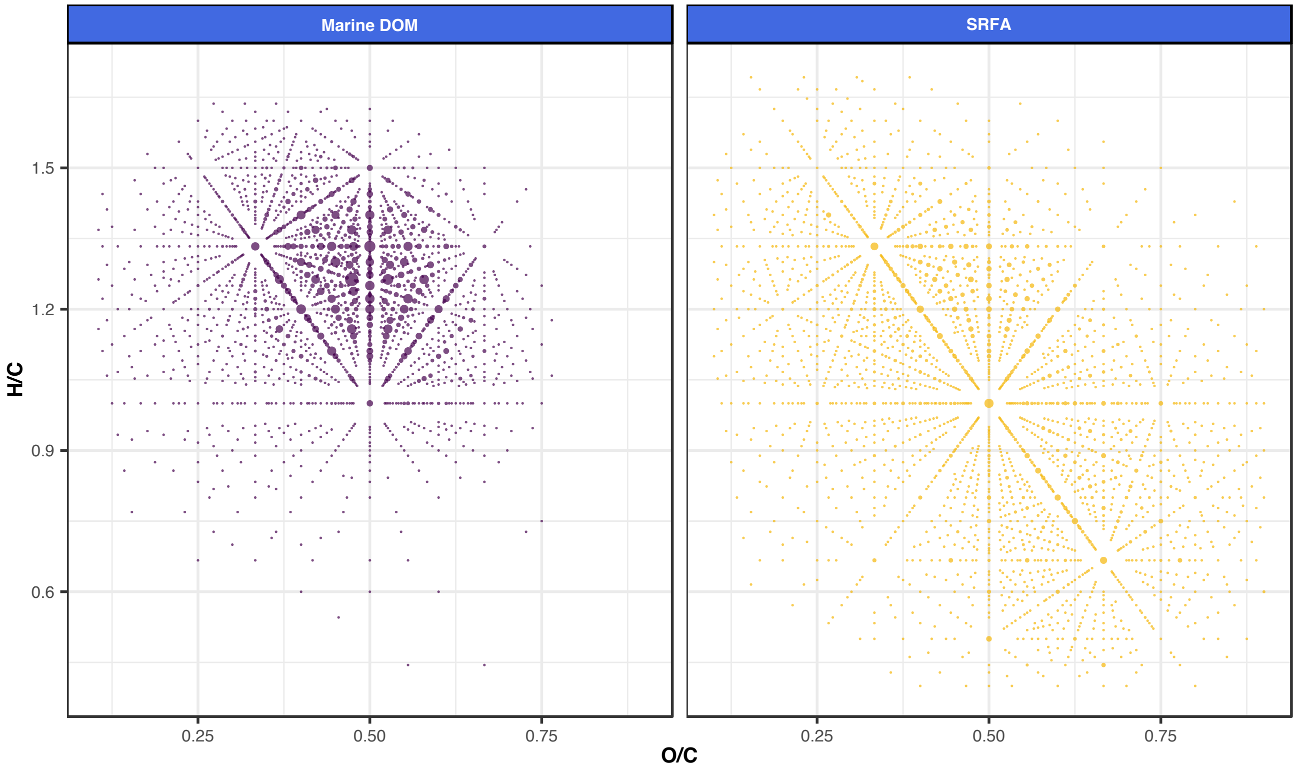


Figure S8: van Krevelen diagram of the H/C ratio vs O/C ratio for the Marine DOM and SRFA samples.

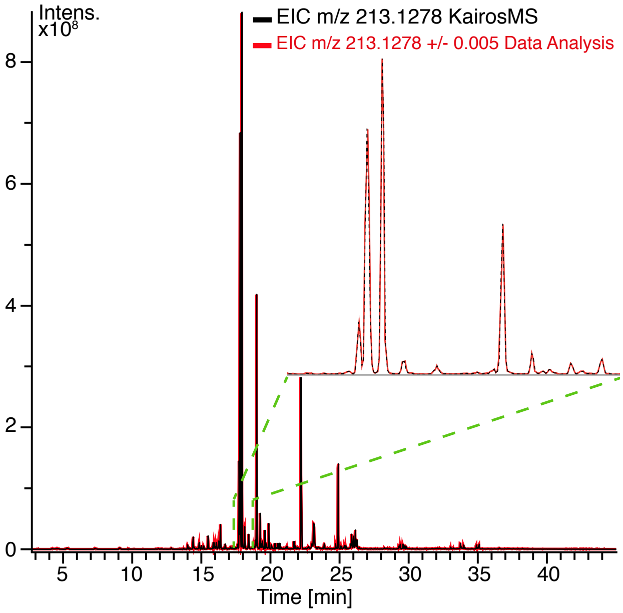


Figure S9: Comparison of the EIC of the same molecular assignment as seen in DA and KairosMS after peak picking at S/N 1.

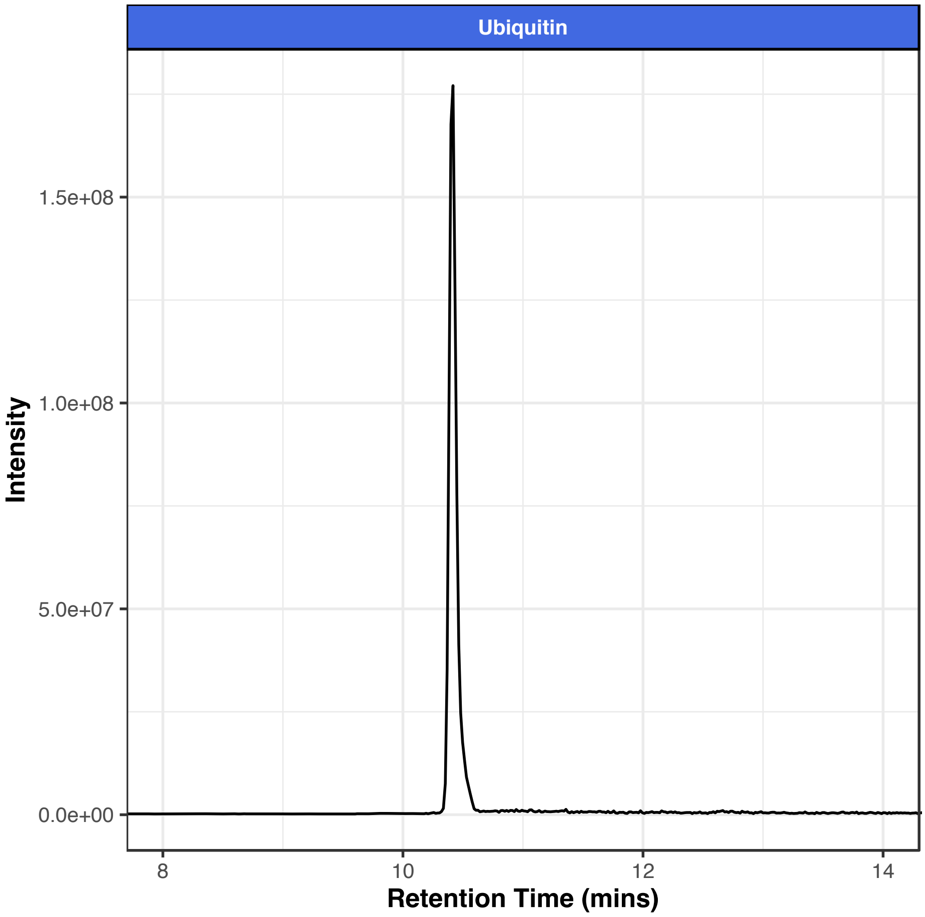


Figure S10: EIC from a peptide digest of ubiquitin analyzed by LC-FTICR MS.

Chapter 6

Summary

6.1 Chapter 2: Themis: Batch Preprocessing for Ultrahigh-Resolution Mass Spectra of Complex Mixtures

Themis uses replicate measurements of a single sample in order to create a single peak list by looking for consistent information across replicates. This method improves retention of reliable information, while avoiding the loss of low-intensity peaks which can happen when the normal signal-to-noise thresholds between 4 and 6 are applied, thanks to the retention of peaks present across multiple replicates. For the user, this translates into a more reliable peak list, without having to analyse several replicates separately or to compare the downstream results.

Themis was illustrated by analysing a light sour crude oil and a South American crude oil with an ESI 12 T FTICR MS. An intentionally low signal-to-noise threshold was used to ensure the inclusion of low intensity peaks. Themis identified 2260 peaks among the 16400 originally picked. The molecular assignments that followed were done with a dedicated software and the comparable composition was observed with and without Themis at high intensity, but only low intensity classes demonstrated improvements. The benefits of using Themis result in being able both to use a lower signal-to-noise threshold than one would when using a single mass spectrum, and reducing the chances of false positive molecular assignments. Themis provides researchers with the ability to improve their data and facilitate downstream analysis with minimal extra work.

6.2 Chapter 3: Repeatability, signal-to-noise ratio, mass error and molecular assignments in petroleomics

In order to obtain high quality molecular assignments and reduce the assignment of unreliable peaks, the important setting is the signal-to-noise threshold. Setting the S/N threshold too high leads to a conservative list of more reliable peaks but will miss out on low intensity peaks. A lower S/N will avoid missing low intensity peaks but will include more noise and unreliable peaks. The use of the Themis algorithm with replicates enables to overcome these issues by discarding non-reproducible peaks, leading to lower RMS mass error and more reliable assignments, especially towards lower intensities. These characteristics makes it a suitable tool to study the proportion of molecular assignments of not fully reproducible peaks obtained using a conventional procedure. Three different samples were studied and their peak lists exported using a range of S/N thresholds. As expected, the RMS mass error decreased and most of the original intensity was assigned after processing with Themis. The proportion of non-reproducible compositional assignments increases when reducing the S/N threshold, the highest numbers being logically for the lowest S/N threshold. These non-reproducible compositional assignments did not focus around specific molecular classes and were distributed proportionally to the peak's density across the m/z range. The distribution of the homologous series improved when analysing only peaks present in all replicates, showing an enhanced continuity, a criterion often used to distinguish correct and incorrect assignments. It has been estimated that between 15 to 26% of the compositional assignments can be considered as not fully reproducible, depending on the S/N threshold used for the peak-picking.

6.3 Chapter 4: Rhapso: Automatic stitching of mass segments from Fourier transform ion cyclotron resonance mass spectra

Some petroleum samples are so complex that even FTICR MS struggles to provide an exhaustive overview of their composition. The spectral stitching provides the capability to yield more information with a higher accuracy while using existing FTICR MS instrumentation. Rhapso was developed to allow this technique to become more common by reducing the challenge posed by data processing. The edge effect correction performed by Rhapso uses a higher proportion of the windows of acquisition while preserving the natural intensity undulations, characteristic of petroleum-related samples. Finally, Rhapso automatically performs the stitching between each spectrum by using the optimal location within the overlap of the two spectra. The Rhapso algorithm was implemented within a web-based interface capable of running on both laptops and servers and performs its task within minutes. This technique demonstrated that using spectral stitching, when compared to a broadband method, yields a net increase in the number of peaks detected but also a clearer quality gain for the molecular assignments, with a significant reduction in the RMS mass error. Rhapso was developed to be part of the OCULAR method by Palacio Lozano *et al.* which lead to the assignment of a record breaking 244,779 molecular compositions within a petroleum sample.

6.4 Chapter 5: KairosMS: New solution to process complex mixture data analyzed by hyphenated - ultra-high-resolution mass spectrometry

Chromatography coupled with ultra-high-resolution mass spectrometry has been increasingly employed over recent years as a way to gain a deeper understanding of complex mixtures. The existing data analysis methods available were not suitable for such datasets as they struggled to scale the large amount of data leading to manually laborious techniques having to be employed, often leading to loss of information. KairosMS was developed to tackle this challenge and uses existing tools while taking out most of the laborious work for the user. Existing software is used to export the large dataset to a mass list processed by KairosMS. Molecular assignments are obtained by using existent dedicated software on a peak list generated by KairosMS. After processing, the user is able to explore their data seamlessly by using the large variety of visualisations available. KairosMS was successfully used on a large variety of samples (petroleum, environmental, dissolved organic matter results and biomolecules), instruments (FTICR MS, Orbitrap) and chromatography (LC, GC). Information which was previously only attainable with a tremendous amount of work became easily accessible to users allowing them to raise more questions and find more answers, particularly in the presence of isomers. KairosMS gives one the ability to rapidly visualise EICs from any class, homologous series, m/z or molecular assignment. Thanks to a common file format for the processed data, it is also possible to simultaneously visualise multiple samples at once, saving the user from having to repeat the same tasks several times and allowing them to immediately grasp differences between samples.

Chapter 7

Conclusions and Future Work

Analysing complex mixtures with ultra-high-resolution mass spectrometry pushes not only instrumental but also data analysis techniques to their limits, highlighting the limitations of the inability of the majority of the existing tools to cope with such data. This field being at the forefront of research, the small size of the market makes it difficult for a company to develop profitable dedicated tools, and researchers had to resort to adapting multiple software and many repetitive tasks. Scientists in this area of research had to become self-sufficient in terms of data analysis tools to ensure its continued progression. This task required a highly interdisciplinary approach in order to develop new algorithms but also implement of new methods to quickly prototype and customise software that will allow rapid and efficient workflows.

The work presented in this thesis is the result of the combination of chemistry, statistics and data science in order to address the challenges of data analysis for petroleomic mass spectrometry data. Three new algorithms implemented into fully functional solutions were presented, all with friendly user interfaces allowing researchers to improve their workflow, decrease time spent on laborious tasks, improve data quality and enable access to more information. Themis, Rhapso and KairosMS are currently used regularly by Dr. Barrow's research group but their functionalities can still be expanded. In order to increase access to those tools, software engineering work has to be carried out to render these software more stable and better able to handle higher processing loads. However, this task necessitates software engineering knowledge, usually outside of the scope of a PhD research project.

In the future, further studies to better understand the optimal settings for Themis would be necessary, particularly to determine the ideal number of replicates to use according to sample complexity. Rhapso is currently still in its infancy and it is believed that more sophisticated techniques could be implemented to better handle the extremely complex situations with an abundant level of noise. Ideally, automatic sampling techniques would be implemented to allow for replicates of each window, allowing us to apply Themis prior to using Rhapso for the stitching. This would make for higher data quality but would also give Rhapso the opportunity to perform more efficiently.

KairosMS capabilities are currently being refined and extended every day as researchers use it and provide feedback. It has the potential to become the main tool for any researcher working with complex mixtures and ultra-high-resolution mass spectrometry. At the moment, KairosMS covers the needs of Dr Barrow's research group and can process data from other instruments thanks to few collaborations, but in the future there are plans to expand its capabilities beyond the lab's needs. The ability to process ion mobility data is currently being implemented, along with advanced abilities to gain more information from the isomers present in a sample. Further developments could lead to the implementation of GCxGC instrumentation along with 2D mass spectrometry.

This thesis started with the ambition of using statistics to help address the data analysis challenges faced by petroleomics. The combination of the two brought some successful developments but the data manipulation and visualisation quickly became the bottleneck. The rise of data science as a discipline helped fuel the development of specifically designed R packages to addresses these issues.

The scope of this research was refined over the years as the needs and tools were better understood, but ultimately all the different projects share the same vision of addressing difficult challenges using tools and ideas outside of the conventional chemist's toolbox. In closing, this author has gained valuable experience from this research and would like to impart some lessons learnt: spending time with statisticians from various fields, being immersed in the data science community, travelling all over the world and learning about geography, history, politics, culture, languages, sociology and photography has fuelled this multipotentialite, enabling him to look at problems in a unique way and come up with unique solutions. However, working at the intersection of many different fields comes with its own set of challenges. Being involved in multiple different disciplines means that one is constantly surrounded by specialists which can make one feel like they are never enough; there is always something that one does not know because one endeavours to become a specialist in several fields all at once while most people stick to a single field.

Finding peers can be extremely disorientating as while one theoretically belongs in multiple places, the reality is often that there is the feeling of belonging nowhere, feeling either too much or not enough. As such situations further exacerbate the well-known imposter syndrome, it is even more important to surround one's self with the right people. This author counts himself extremely lucky to have found many open-minded people who saw value in the work and who were of immense help along the way. These people were always willing to make an effort to understand the endeavour being made, and make their expertise as accessible as possible. The understanding that everyone approaches problems from different perspectives is extremely important, as is being able to translate one's ideas in a way that another can understand.

Finally, it is crucial to realise that when developing algorithms, dashboards, software or anything else aimed at users, developers need to meet the users where they are. For that, developing empathy for the users and their problems and a direct understanding of the research issues is key; walk a mile in their shoes and one will be able to develop a solution that suit them, not one's self.

The publication by Palacio Lozano *et al.* demonstrates that when chemistry, statistics and data science are combined, the analytical boundaries can be pushed further. Going forward, sciences cannot afford to ignore the benefits of interdisciplinarity and other bridges between fields that could be beneficial to everyone.

Bibliography

[1] Remy Gavard, David Rossell, Simon E. F. Spencer, and Mark P. Barrow. Themis: Batch Preprocessing for Ultrahigh-Resolution Mass Spectra of Complex Mixtures. *Anal. Chem.*, 89(21):11383–11390, 2017. ISSN 0003-2700. doi: 10.1021/acs.analchem.7b02345.

[2] Remy Gavard, Diana Catalina Palacio Lozano, Alexander Guzman, David Rossell, Simon E.F. Spencer, and Mark P. Barrow. Rhapso: Automatic Stitching of Mass Segments from Fourier Transform Ion Cyclotron Resonance Mass Spectra. *Anal. Chem.*, 91(23):15130–15137, dec 2019. ISSN 15206882. doi: 10.1021/acs.analchem.9b03846. URL `https://pubs.acs.org/doi/abs/10.1021/acs.analchem.9b03846`.

[3] Diana Catalina Palacio Lozano, Remy Gavard, Juan P. Arenas-Diaz, Mary J. Thomas, David D. Stranz, Enrique Mejía-Ospino, Alexander Guzman, Simon E. F. Spencer, David Rossell, and Mark P. Barrow. Pushing the analytical limits: new insights into complex mixtures using mass spectra segments of constant ultrahigh resolving power. *Chem. Sci.*, 10(29):6966–6978, Jul 2019. ISSN 2041-6520. doi: 10.1039/C9SC02903F.

[4] Diana Catalina Palacio Lozano, Claudia X. Ramírez, José Aristóbulo Sarmiento Chaparro, Mary J. Thomas, Remy Gavard, Hugh E. Jones, Rafael Cabanzo Hernández, Enrique Mejia-Ospino, and Mark P. Barrow. Characterization of bio-crude components derived from pyrolysis of soft wood

and its esterified product by ultrahigh resolution mass spectrometry and spectroscopic techniques. *Fuel*, 259:116085, jan 2020. ISSN 00162361. doi: 10.1016/j.fuel.2019.116085.

[5] Remy Gavard, Hugh E. Jones, Diana Catalina Palacio Lozano, Mary Joanna Thomas, David Rossell, Simon E.F. Spencer, and Mark P. Barrow. KairosMS: A new solution for the processing of hyphenated ultrahigh resolution mass spectrometry data. *Anal. Chem.*, 92(5):3775–3786, jan 2020. ISSN 0003-2700. doi: 10.1021/acs.analchem.9b05113.

[6] Remy Gavard, Diana Catalina Palacio Lozano, Hugh E. Jones, Mary J. Thomas, David Rossell, Simon E. F. Spencer, and Mark P. Barrow. Study of the rate of non-reproducible peaks assigned a molecular composition in petroleomics.

[7] Diana Catalina Palacio Lozano, Remy Gavard, Hugh E. Jones, Mary J. Thomas, Claudia X. Ramirez, Jos Aristbulo Sarmiento Chaparro, Matthias Witt, Enrique Mejia-Ospino, and Mark P. Barrow. Advanced Analysis of Bio-oils By Gas Chromatography Coupled To Fourier Transform Ion Cyclotron Resonance Mass Spectrometry.

[8] Joseph John Thomson. Bakerian lecture:rays of positive electricity. *Proc. R. Soc. Lond. A*, 89(607):1–20, 1913.

[9] AJ Dempster. A new method of positive ray analysis. *Phys. Rev.*, 11(4):316, 1918.

[10] John H Beynon. The use of the mass spectrometer for the identification of organic compounds. *Microchimica Acta*, 44(1-3):437–453, 1956.

[11] Fred W McLafferty. Mass spectrometry in chemical research and production. *Appl. Spectrosc.*, 11(4):148–156, 1957.

[12] Roland S Gohlke. Time-of-flight mass spectrometry and gas-liquid partition chromatography. *Anal. Chem.*, 31(4):535–541, 1959.

[13] F.W. McLafferty and T.A. Bryce. Metastable-ion characteristics: characterization of isomeric molecules. *Chem. Com. (London)*, (23):1215–1217, 1967.

[14] K.R. Jennings. Collision-induced decompositions of aromatic molecular ions. *Int. J. Mass Spectrom Ion Phys.*, 1(3):227–235, 1968.

[15] E.C. Horning, D.I. Carroll, I. Dzidic, K.D. Haegele, M.G. Horning, and R.N. Stillwell. Atmospheric pressure ionization mass spectrometry. solvent-mediated ionization of samples introduced in solution and in a liquid chromatograph effluent stream. *J. Chromatogr. Sci.*, 12(11):725–729, 1974.

[16] Patrick Arpino, M. A. Baldwin, and F. W. McLafferty. Liquid chromatography-mass spectrometry. ii - continuous monitoring. *Biomed. Mass Spectrom.*, 1(1): 80–82, 1974. doi: 10.1002/bms.1200010117.

[17] Melvin B. Comisarow and Alan G. Marshall. Fourier transform ion cyclotron resonance spectroscopy. *Chem. Phys. Lett.*, 25(2):282–283, Mar 1974. ISSN 00092614. doi: 10.1016/0009-2614(74)89137-2.

[18] R.A. Yost and C.G. Enke. Selected ion fragmentation with a tandem quadrupole mass spectrometer. *J. Am. Chem. Soc.*, 100(7):2274–2275, 1978.

[19] Randall K. Julian and R. Graham Cooks. Broad-band excitation in the quadrupole ion trap mass spectrometer using shaped pulses created with the inverse fourier transform. *Anal. Chem.*, 65(14):1827–1833, 1993.

[20] Matthias Mann and Matthias Wilm. Direct analysis of the polar fraction of heavy petroleum crude oil using a linear ion trap/FTICR hybrid mass spectrometer. In *Proceedings of the 42nd ASMS Conference on Mass Spectrometry and Allied Topics, Chicago,IL*, page 770, 1994.

[21] Alexander Makarov. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.*, 72(6):1156–1162, 2000.

[22] K. H. Kingdon. A method for the neutralization of electron space charge by positive ionization at very low gas pressures. *Phys. Rev.*, 21(4):408–418, Apr 1923. ISSN 0031899X. doi: 10.1103/PhysRev.21.408.

[23] G.M. Schier, B. Halpern, and G.W.A. Milne. Characterization of dipeptides by electron impact and chemical ionization mass spectrometry. *Biol. Mass Spectrom.*, 1(4):212–218, 1974.

[24] Milam S.B. Munson and F.H. Field. Chemical ionization mass spectrometry. i. general introduction. *J. Am. Chem. Soc.*, 88(12):2621–2630, 1966.

[25] Eugene Nikolaev. Victor Talrose: an appreciation. *J. Mass Spectrom.*, 33(6): 499–501, 1998.

[26] Alex G Harrison. *Chemical ionization mass spectrometry.* Routledge, 2018.

[27] D. I. Carroll, Ismet Dzidic, Richard N. Stillwell, Klaus D. Haegele, and Evan C. Horning. Atmospheric Pressure Ionization Mass Spectrometry. Corona Discharge Ion Source for use in a Liquid Chromatograph-Mass Spectrometer-Computer Analytical System. *Anal. Chem.*, 47(14):2369–2373, Dec 1975. ISSN 15206882. doi: 10.1021/ac60364a031.

[28] J.B. French and N.M. Reid. Real-time targeted compound monitoring in air using the tag a 3000 atmospheric pressure chemical ionization mass spectrometer system. *Dynamic Mass Spectrometry*, 6:220–233, 1980.

[29] Damon B. Robb, Thomas R. Covey, and Andries P. Bruins. Atmospheric pressure photoionization: An ionization method for liquid chromatography-mass spectrometry. *Anal. Chem.*, 72(15):3653–3659, 2000. ISSN 00032700. doi: 10.1021/ac0001636.

[30] Jack A. Syage and Matthew D. Evans. Photoionization mass spectrometry: A powerful new tool for drug discovery-this article describes the benefits that photoionization ms has over existing methods for performing high-speed. *Spectroscopy-Eugene*, 16(11):14–21, 2001.

[31] John B. Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craig M Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, 1989.

[32] Walker Bleakney. A new method of positive ray analysis and its application to the measurement of ionization potentials in mercury vapor. *Phys. Rev.*, 34(1): 157–160, Jul 1929. ISSN 0031899X. doi: 10.1103/PhysRev.34.157.

[33] Alfred O. Nier. A mass spectrometer for isotope and gas analysis. *Rev. Sci. Instrum.*, 18(6):398–411, Jun 1947. ISSN 00346748. doi: 10.1063/1.1740961.

[34] Sahba Ghaderi, P.S. Kulkarni, Edward B. Ledford, Charles L. Wilkins, and Michael L. Gross. Chemical ionization in fourier transform mass spectrometry. *Anal. Chem.*, 53(3):428–437, 1981.

[35] Craig M. Whitehouse, Robert N. Dreyer, Masamichi Yamashita, and John B. Fenn. Electrospray interface for liquid chromatographs and mass spectrometers. *Anal. Chem.*, 57(3):675–679, 1985.

[36] Andries P. Bruins, Thomas R. Covey, and Jack D. Henion. Ion spray interface for combined liquid chromatography/atmospheric pressure ionization mass spectrometry. *Anal. Chem.*, 59(22):2642–2646, 1987.

[37] Malcolm Dole, L. L. Mack, R. L. Hines, R. C. Mobley, L. D. Ferguson, and M. B. Alice. Molecular Beams of Macroions. *J. Chem. Phys.*, 49(5), Sep 1968. ISSN 0021-9606. doi: 10.1063/1.1670391.

[38] Swapan K. Chowdhury, Viswanatham Katta, and Brian T. Chait. An electrospray-ionization mass spectrometer with new features. *Rapid Commun. Mass Spectrom.*, 4(3):81–87, 1990.

[39] Michel W.F. Nielen and F.A. Buijtenhuijs. Polymer analysis by liquid chromatography/electrospray ionization time-of-flight mass spectrometry. *Anal. Chem.*, 71(9):1809–1814, 1999.

[40] Ray Colton and John C. Traeger. The application of electrospray mass spectrometry to ionic inorganic and organometallic systems. *Inorganica chimica acta*, 201(2):153–155, 1992.

[41] Joseph A. Loo. Electrospray ionization mass spectrometry: a technology for studying noncovalent macromolecular complexes. *Int. J. Mass Spectrom.*, 200 (1-3):175–186, 2000.

[42] Matthias Wilm and Matthias Mann. Analytical properties of the nanoelectrospray ion source. *Anal. Chem.*, 68(1):1–8, 1996.

[43] Christopher P. Rüger, Theo Schwemer, Martin Sklorz, Peter B. O'Connor, Mark P. Barrow, and Ralf Zimmermann. 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.*, 23(1):28–39, Jan 2017. ISSN 17516838. doi: 10.1177/1469066717694286.

[44] Benjamin C. Blount, K. Eric Milgram, Manori J. Silva, Nicole A. Malek, John A. Reidy, Larry L. Needham, and John W. Brock. Quantitative detection of eight phthalate metabolites in human urine using HPLC-APCI-MS/MS. *Anal. Chem.*, 72(17):4127–4134, 2000. ISSN 00032700. doi: 10.1021/ac000422r.

[45] Heng Hui Gan, Bingnan Yan, Robert S.T. Linforth, and Ian D. Fisk. Development and validation of an APCI-MS/GC-MS approach for the classification and prediction of Cheddar cheese maturity. *Food Chem.*, 190:442–447, Jan 2016. ISSN 18737072. doi: 10.1016/j.foodchem.2015.05.096.

[46] Yifei Wang, Jennifer Johnson-Cicalese, Ajay P. Singh, and Nicholi Vorsa. Characterization and quantification of flavonoids and organic acids over fruit development in American cranberry (Vaccinium macrocarpon) cultivars using HPLC and APCI-MS/MS. *Plant Sci.*, 262:91–102, Sep 2017. ISSN 18732259. doi: 10.1016/j.plantsci.2017.06.004.

[47] Mark P. Barrow, Matthias Witt, John V. Headley, and Kerry M. Peru. Athabasca oil sands process water: Characterization by atmospheric pressure photoionization and electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.*, 82(9):3727–3735, 2010. ISSN 00032700. doi: 10.1021/ac100103y.

[48] Michael Guilhaus. Special feature: Tutorial. principles and instrumentation in time-of-flight mass spectrometry. physical and instrumental concepts. *J. Mass Spectrom.*, 30(11):1519–1532, 1995.

[49] Christian Weickhardt, Friedrich Moritz, and Jürgen Grotemeyer. Time-of-flight mass spectrometry: State-of the-art in chemical analysis and molecular science. *Mass Spectrom. Rev.*, 15(3):139–162, 1996.

[50] Edmond de Hoffmann and Vincent Stroobant. *Mass Spectrometry: Principles and Applications, 3rd Edition.* Wiley-Interscience, 2007.

[51] K. G. Standing and Marvin L. Vestal. Time-of-flight mass spectrometry (TOFMS): From niche to mainstream. *Int. J. Mass Spectrom.*, 377(1):295–308, Feb 2015. ISSN 13873806. doi: 10.1016/j.ijms.2014.09.002.

[52] Mahmoud M. Yassine and Ewa Dabek-Zlotorzynska. Investigation of isomeric structures in a commercial mixture of naphthenic acids using ultrahigh pressure liquid chromatography coupled to hybrid traveling wave ion mobility-time of flight mass spectrometry. *J. Chromatogr. A*, 1572:90–99, Oct 2018. ISSN 18733778. doi: 10.1016/j.chroma.2018.08.052.

[53] R.E. Ferguson, K.E. McCulloh, and H.M. Rosenstock. Observation of the products of ionic collision processes and ion decomposition in a linear, pulsed time-of-flight mass spectrometer. *J. Chem. Phys.*, 42(1):100–106, 1965.

[54] Hermann Kienitz and Fritz Aulinger. *Massenspektrometrie*. VerlagChemie, 1968.

[55] Michael L. Gross and Richard M. Caprioli. *The encyclopedia of mass spectrometry*, volume 1. Elsevier, 2003.

[56] Klaus Biemann. The coming of age of mass spectrometry in peptide and protein chemistry. *Prot. Sci.*, 4(9):1920–1927, 1995.

[57] Ruedi Aebersold. A mass spectrometric journey into protein and proteome research. *J. Am. Soc. Mass Spectrom.*, 14(7):685–695, 2003.

[58] Ryan P. Rodgers and Amy M. McKenna. Petroleum analysis. *Anal. Chem.*, 83(12):4665–4687, 2011. ISSN 00032700. doi: 10.1021/ac201080e.

[59] Vladimir M. Doroshenko and Robert J. Cotter. Method and apparatus for trapping ions by increasing trapping voltage during ion introduction, Mar 1995.

[60] P.B. Kyle. Chapter 7 - toxicology: Gcms. In Hari Nair and William Clarke, editors, *Mass Spectrometry for the Clinical Laboratory*, pages 131 – 163. Academic Press, San Diego, 2017. ISBN 978-0-12-800871-3. doi: https://doi.org/10.1016/B978-0-12-800871-3.00007-9.

[61] W. Paul and H. Steinwedel. Apparatus for separating charged particles of different specific charges, 1960.

[62] G.C. Stafford Jr, P.E. Kelley, J.E.P. Syka, W.E. Reynolds, and J.F.J. Todd. Recent improvements in and analytical applications of advanced ion trap technology. *Int. J. Mass Spectrom. Ion Proc.*, 60(1):85–98, 1984.

[63] Raymond E March. Quadrupole ion traps. *Mass Spectrom. Rev.*, 28(6):961–989, 2009.

[64] Alexander Alekseevich Makarov. Mass spectrometer, 1999.

[65] Alexander Makarov, Mark E. Hardman, Jae C. Schwartz, and Michael W. Senko. Mass spectrometry method and apparatus, 2005.

[66] Tobias Kind and Oliver Fiehn. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, 8(1):105, Dec 2007. ISSN 14712105. doi: 10.1186/1471-2105-8-105.

[67] R. D. Knight. Storage of ions from laser-produced plasmas. *Appl. Phys. Lett.*, 38(4):221–223, Feb 1981. ISSN 00036951. doi: 10.1063/1.92315.

[68] R. Blumel. Dynamic Kingdon trap. *Phys. Rev. A*, 51(1):R30–R33, Jan 1995. ISSN 10502947. doi: 10.1103/PhysRevA.51.R30.

[69] Melvin B. Comisarow and Alan G. Marshall. Selective-phase ion cyclotron resonance spectroscopy. *Can. J. Chem.*, 52(4):1997–1999, 1974. ISSN 0008-4042. doi: 10.1139/v74-288.

[70] Melvin B. Comisarow and Alan G. Marshall. Frequency-sweep Fourier transform ion cyclotron resonance spectroscopy. *Chem. Phys. Lett.*, 26(4):489–490, 1974. ISSN 00092614. doi: 10.1016/0009-2614(74)80397-0.

[71] Eugene N. Nikolaev, Ivan A. Boldin, Roland Jertz, and Gökhan Baykut. Initial experimental characterization of a new ultra-high resolution FTICR cell with dynamic harmonization. *J. Am. Soc. Mass Spectrom.*, 22(7):1125–1133, 2011.

[72] I. Jonathan Amster. Fourier transform mass spectrometry. *J. Mass Spectrom.*, 31(12):1325–1337, Dec 1996. ISSN 10765174. doi: 10.1002/(SICI)1096-9888(199612)31:12⟨1325::AID-JMS453⟩3.0.CO;2-W.

[73] Alan G. Marshall, Christopher L. Hendrickson, and George S. Jackson. Fourier Transform Ion Cyclotron Resonance Mass Spectromeyry: A Primer. *Mass Spectrom. Rev.*, 17(1):1–35, 1998. ISSN 02777037. doi: 10.1002/(SICI)1098-2787(1998)17:1⟨1::AID-MAS1⟩3.0.CO;2-K.

[74] Robert C. Dunbar, Jyh H. Chen, and John D. Hays. Magnetron motion of ions in the cubical icr cell. *Int. J. Mass Spectrom. Ion Proc.*, 57(1):39–56, 1984.

[75] L. S. Ettre and K. I. Sakodynskii. M. S. Tswett and the discovery of chromatography I: Early work (1899-1903). *Chromatographia*, 35(3-4):223–231, feb 1993. ISSN 00095893. doi: 10.1007/BF02269707.

[76] Kateina Maštovská and Steven J. Lehotay. Practical approaches to fast gas chromatography-mass spectrometry. *J. Chromatogr. A*, 1000(1-2):153–180, Jun 2003. ISSN 00219673. doi: 10.1016/S0021-9673(03)00448-5.

[77] Mark P. Barrow, Kerry M. Peru, and John V. Headley. An added dimension: GC atmospheric pressure chemical ionization FTICR MS and the Athabasca oil sands. *Anal. Chem.*, 86(16):8281–8288, 2014. ISSN 15206882. doi: 10.1021/ac501710y.

[78] Mary J. Thomas, Emma Collinge, Matthias Witt, Diana Catalina Palacio Lozano, Christopher H. Vane, Vicky Moss-Hayes, and Mark P. Barrow. Petroleomic depth profiling of Staten Island salt marsh soil: 2ω detection

FTICR MS offers a new solution for the analysis of environmental contaminants. *Sci. Total Environ.*, 662:852–862, Apr 2019. ISSN 18791026. doi: 10.1016/j.scitotenv.2019.01.228.

[79] Patrick Louchouarn, Rainer M.W. Amon, Shuiwang Duan, Christina Pondell, Shaya M. Seward, and Noah White. Analysis of lignin-derived phenols in standard reference materials and ocean dissolved organic matter by gas chromatography/tandem mass spectrometry. *Mar. Chem.*, 118(1-2):85–97, 2010.

[80] Sílvia M. Rocha, Michael Caldeira, Joana Carrola, Magda Santos, Nádia Cruz, and Iola F. Duarte. Exploring the human urine metabolomic potentialities by comprehensive two-dimensional gas chromatography coupled to time of flight mass spectrometry. *J. Chromatogr. A*, 1252:155–163, 2012. doi: 10.1016/j.chroma.2012.06.067.

[81] Virgínia C. Fernandes, Jose L. Vera, Valentina F. Domingues, Luís M.S. Silva, Nuno Mateus, and Cristina Delerue-Matos. Mass spectrometry parameters optimization for the 46 multiclass pesticides determination in strawberries with gas chromatography ion-trap tandem mass spectrometry. *J. Am. Soc. Mass Spectrom.*, 23(12):2187–2197, 2012.

[82] Olivier L. Collin, Carolyn M. Zimmermann, and Glen P. Jackson. Fast gas chromatography negative chemical ionization tandem mass spectrometry of explosive compounds using dynamic collision-induced dissociation. *Int. J. Mass Spectrom.*, 279(2-3):93–99, 2009.

[83] A. J. P. Martin and R. L. M. Synge. A new form of chromatogram employing two liquid phases. *Biochem. J.*, 35(12):1358–1368, dec 1941. ISSN 0306-3283. doi: 10.1042/bj0351358.

[84] Emilio Gelpí. Interfaces for coupled liquid-phase separation/mass spectrometry

techniques. An update on recent developments. *J. Mass Spectrom.*, 37(3): 241–253, Mar 2002. ISSN 10765174. doi: 10.1002/jms.297.

[85] Xiao Zheng, An Kang, Chen Dai, Yan Liang, Tong Xie, Lin Xie, Yin Peng, Guangji Wang, and Haiping Hao. Quantitative analysis of neurochemical panel in rat brain and plasma by liquid chromatography–tandem mass spectrometry. *Anal. Chem.*, 84(22):10044–10051, 2012.

[86] Phanichand Kodali, Karnakar R. Chitta, Julio A. Landero Figueroa, Joseph A. Caruso, and Opeolu Adeoye. Detection of metals and metalloproteins in the plasma of stroke patients by mass spectrometry methods. *Metallomics*, 4(10): 1077–1087, 2012.

[87] Harald Pasch and Karsten Rode. Use of matrix-assisted laser desorption/ionization mass spectrometry for molar mass-sensitive detection in liquid chromatography of polymers. *J. Chromatogr. A*, 699(1-2):21–29, 1995.

[88] Jeremiah D. Tipton, John C. Tran, Adam D. Catherman, Dorothy R. Ahlf, Kenneth R. Durbin, Ji Eun Lee, John F. Kellie, Neil L. Kelleher, Christopher L. Hendrickson, and Alan G. Marshall. Nano-LC FTICR tandem mass spectrometry for top-down proteomics: routine baseline unit mass resolution of whole cell lysate proteins up to 72 kDa. *Anal. Chem.*, 84(5):2111–2117, 2012.

[89] Claudia Patriarca, Jonas Bergquist, Per J. R. Sjberg, Lars Tranvik, and Jeffrey A. Hawkes. Online hplc-esi-hrms method for the analysis and comparison of different dissolved organic matter samples. *Envir. Sci. Tech.*, 52(4):2091–2099, 2018. doi: 10.1021/acs.est.7b04508. PMID: 29241333.

[90] E. Bartholdi and R. R. Ernst. Fourier spectroscopy and the causality principle. *J. Magn. Reson.*, 11(1):9–19, Jul 1973. ISSN 00222364. doi: 10.1016/0022-2364(73)90076-0.

[91] Greg T. Blakney, Donald F. Smith, Tong Chen, Chad R. Weisbrod, Alan G. Marshall, John P. Quinn, Nathan K. Kaiser, Steven C. Beu, and Christopher L. Hendrickson. 21 Tesla Fourier Transform Ion Cyclotron Resonance Mass Spectrometer: A National Resource for Ultrahigh Resolution Mass Analysis. *J. Am. Soc. Mass Spectrom.*, 26(9):1626–1632, Sep 2015. ISSN 1044-0305. doi: 10.1007/s13361-015-1182-2.

[92] Jared B. Shaw, Tzu Yung Lin, Franklin E. Leach, Aleksey V. Tolmachev, Nikola Tolić, Errol W. Robinson, David W. Koppenaal, and Ljiljana Paša-Tolić. 21 Tesla Fourier Transform Ion Cyclotron Resonance Mass Spectrometer Greatly Expands Mass Spectrometry Toolbox. *J. Am. Soc. Mass Spectrom.*, 27(12): 1929–1936, Dec 2016. ISSN 18791123. doi: 10.1007/s13361-016-1507-9.

[93] Donald F. Smith, David C. Podgorski, Ryan P. Rodgers, Greg T. Blakney, and Christopher L. Hendrickson. 21 Tesla FT-ICR Mass Spectrometer for Ultrahigh-Resolution Analysis of Complex Organic Mixtures. *Anal. Chem.*, 90 (3):2041–2047, 2018. ISSN 15206882. doi: 10.1021/acs.analchem.7b04159.

[94] Eunji Cho, Matthias Witt, Manhoi Hur, Maeng Joon Jung, and Sunghwan Kim. Application of FT-ICR MS Equipped with Quadrupole Detection for Analysis of Crude Oil. *Anal. Chem.*, 89(22):12101–12107, 2017. ISSN 15206882. doi: 10.1021/acs.analchem.7b02644.

[95] Feng Xian, Christopher L. Hendrickson, and Alan G. Marshall. High resolution mass spectrometry. *Anal. Chem.*, 84(2):708–719, 2012.

[96] Yulin Qi and Peter B. O'Connor. Data processing in Fourier transform ion cyclotron resonance mass spectrometry. *Mass Spectrom. Rev.*, 33(5):333–352, 2014. ISSN 10982787. doi: 10.1002/mas.21414.

[97] Yulin Qi, Mark P. Barrow, Steve L. Van Orden, Christopher J. Thompson, Huilin Li, Pilar Perez-Hurtado, and Peter B. O'Connor. Variation of the

Fourier transform mass spectra phase function with experimental parameters. *Anal. Chem.*, 83(22):8477–8483, 2011. ISSN 00032700. doi: 10.1021/ac2017585.

[98] Yulin Qi, Mark P. Barrow, Huilin Li, Joseph E. Meier, Steve L. Van Orden, Christopher J. Thompson, and Peter B. O'Connor. Absorption-mode: The next generation of Fourier transform mass spectra. *Anal. Chem.*, 84(6):2923–2929, Mar 2012. ISSN 00032700. doi: 10.1021/ac3000122.

[99] David P.A. Kilgour and Steven L. Van Orden. Absorption mode Fourier transform mass spectrometry with no baseline correction using a novel asymmetric apodization function. *Rapid Commun. Mass Spectrom.*, 29(11):1009–1018, 2015. ISSN 10970231. doi: 10.1002/rcm.7190.

[100] Li-Kang Zhang, Don Rempel, Birendra N. Pramanik, and Michael L. Gross. Accurate mass measurements by Fourier transform mass spectrometry. *Mass Spectrom. Rev.*, 24(2):286–309, Mar 2005. ISSN 0277-7037. doi: 10.1002/mas.20013.

[101] Konstantin O. Zhurov, Anton N. Kozhinov, and Yury O. Tsybin. Evaluation of high-field orbitrap fourier transform mass spectrometer for petroleomics. *Energy and Fuels*, 27(6):2974–2983, jun 2013. ISSN 08870624. doi: 10.1021/ef400203g. URL http://pubs.acs.org/doi/abs/10.1021/ef400203g.

[102] Alan G. Marshall and Ryan P. Rodgers. Petroleomics: The Next Grand Challenge for Chemical Analysis. *Acc. Chem. Res.*, 37(1):53–59, 2004. ISSN 00014842. doi: 10.1021/ar020177t.

[103] Ryan P. Rodgers, Tanner M. Schaub, and Alan G. Marshall. Petroleomics: MS Returns to Its Roots. *Anal. Chem.*, 77(1):20A–27A, 2005. ISSN 0003-2700.

[104] Alan G. Marshall and Ryan P. Rodgers. Petroleomics: Chemistry of the underworld. *Proc. Natl. Acad. Sci. U.S.A*, 105(47):18090–18095, 2008. ISSN 0027-8424. doi: 10.1073/pnas.0805069105.

[105] Mark P Barrow. Petroleomics: study of the old and the new. *Biofuels*, 1(5): 651–655, 2010. ISSN 1759-7269. doi: 10.4155/bfs.10.55.

[106] Mark P. Barrow, Liam A. McDonnell, Xidong Feng, Jérémie Walker, and Peter J. Derrick. Determination of the nature of naphthenic acids present in crude oils using nanospray Fourier transform ion cyclotron resonance mass spectrometry: The continued battle against corrosion. *Anal. Chem.*, 75(4): 860–866, 2003. ISSN 00032700. doi: 10.1021/ac020388b.

[107] Claudia X. Ramírez, Juan E. Torres, Diana Catalina Palacio Lozano, Juan P. Arenas-Diaz, Enrique Mejia-Ospino, Viatcheslav Kafarov, Alexander Guzman, and Jorge Ancheyta. Molecular Representation of Petroleum Residues Using Fourier Transform Ion Cyclotron Resonance Mass Spectrometry and Conventional Analysis. *Energy and Fuels*, 31(12):13353–13363, Dec 2017. ISSN 15205029. doi: 10.1021/acs.energyfuels.7b02507.

[108] Diana Catalina Palacio Lozano, Jorge Armando Orrego-Ruiz, Rafael Cabanzo Hernández, Jáder Enrique Guerrero, and Enrique Mejía-Ospino. APPI(+)-FTICR mass spectrometry coupled to partial least squares with genetic algorithm variable selection for prediction of API gravity and CCR of crude oil and vacuum residues. *Fuel*, 193:39–44, Apr 2017. ISSN 00162361. doi: 10.1016/j.fuel.2016.12.029.

[109] Yunju Cho, Matthias Witt, Young Hwan Kim, and Sunghwan Kim. Characterization of crude oils at the molecular level by use of laser desorption ionization Fourier-transform ion cyclotron resonance mass spectrometry. *Anal. Chem.*, 84(20):8587–8594, Oct 2012. ISSN 00032700. doi: 10.1021/ac301615m.

[110] Erica A. Smith and Young Jin Lee. Petroleomic analysis of bio-oils from the fast pyrolysis of biomass: Laser desorption ionization-linear ion trap-orbitrap

mass spectrometry approach. *Energy and Fuels*, 24(9):5190–5198, Sep 2010. ISSN 08870624. doi: 10.1021/ef100629a.

[111] Nathalia S. Tessarolo, Renzo C. Silva, Gabriela Vanini, Andrea Pinho, Wanderson Romão, Eustáquio V.R. de Castro, and Débora A. Azevedo. Assessing the chemical composition of bio-oils using FT-ICR mass spectrometry and comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry. *Microchem. J.*, 117:68–76, Nov 2014. ISSN 0026265X. doi: 10.1016/j.microc.2014.06.006.

[112] Donald F. Smith, Parviz Rahimi, Alem Teclemariam, Ryan P. Rodgers, and Alan G. Marshall. Characterization of Athabasca bitumen heavy vacuum gas oil distillation cuts by negative/positive electrospray ionization and automated liquid injection field desorption ionization Fourier transform ion cyclotron resonance mass spectrometry. *Energy and Fuels*, 22(5):3118–3125, Sep 2008. ISSN 08870624. doi: 10.1021/ef8000357.

[113] Matthew R. Noestheden, John V. Headley, Kerry M. Peru, Mark P. Barrow, Lyle L. Burton, Takeo Sakuma, 19 Winkler, and J. Larry Campbell. Rapid characterization of naphthenic acids using differential mobility spectrometry and mass spectrometry. *Environ. Sci. Technol.*, 48(17):10264–10272, Sep 2014. ISSN 15205851. doi: 10.1021/es501821h.

[114] H.N. Dunning, J.W. Moore, Herman Bieber, and R.B. Williams. Porphyrin, Nickel, Vanadium, and Nitrogen in Petroleurn. *J. Chem. Eng. Data*, 497(1941): 546–549, Oct 2000. ISSN 0021-9568. doi: 10.1021/je60008a036.

[115] Earl Wayne Baker, Teh Fu Yen, John P. Dickie, Robert E. Rhodes, and Leslie F. Clark. Mass spectrometry of porphyrins. II. Characterization of petroporphyrins. *J. Am. Chem. Soc.*, 89(14):3631–3639, Jul 1967. ISSN 00027863. doi: 10.1021/ja00990a050.

[116] Yinhua Pan, Yuhong Liao, and Quan Shi. Variations of Acidic Compounds in Crude Oil during Simulated Aerobic Biodegradation: Monitored by Semi-quantitative Negative-Ion ESI FT-ICR MS. *Energy and Fuels*, 31(2):1126–1135, 2017. ISSN 15205029. doi: 10.1021/acs.energyfuels.6b02167.

[117] David Borton, David S. Pinkston, Matthew R. Hurt, Xiaoli Tan, Khalid Azyat, Alexander Scherer, Rik Tykwinski, Murray Gray, Kuangnan Qian, and Hilkka I. Kenttämaa. Molecular structures of asphaltenes based on the dissociation reactions of their ions in mass spectrometry. *Energy and Fuels*, 24 (10):5548–5559, 2010. ISSN 08870624. doi: 10.1021/ef1007819.

[118] Edward Kendrick. A mass scale based on CH_2= 14.0000 for high resolution mass spectrometry of organic compounds. *Anal. Chem.*, 35(13):2146–2154, 1963.

[119] Chang S. Hsu, Kuangnan Qian, and Yungning C. Chen. An innovative approach to data analysis in hydrocarbon characterization by on-line liquid chromatography-mass spectrometry. *Anal. Chim. Acta*, 264(1):79–89, Jul 1992. ISSN 00032670. doi: 10.1016/0003-2670(92)85299-L.

[120] J. Fernandez De La Mora. Electrospray ionization of large multiply charged species proceeds via Dole's charged residue mechanism. *Anal. Chim. Acta*, 406 (1):93–104, Feb 2000. ISSN 00032670. doi: 10.1016/S0003-2670(99)00601-7.

[121] Sunghwan Kim, Robert W. Kramer, and Patrick G. Hatcher. Graphical Method for Analysis of Ultrahigh-Resolution Broadband Mass Spectra of Natural Organic Matter, the Van Krevelen Diagram. *Anal. Chem.*, 75(20): 5336–5344, 2003. ISSN 00032700. doi: 10.1021/ac034415p.

[122] David R. Gibson and Hari Pulapaka. A fast algorithm and software for analysis of FT-ICR data. *J. Math. Chem.*, 48(2):381–394, 2010. ISSN 02599791. doi: 10.1007/s10910-010-9679-1.

[123] Christopher P. Rüger, Toni Miersch, Theo Schwemer, Martin Sklorz, and Ralf Zimmermann. Hyphenation of Thermal Analysis to Ultrahigh-Resolution Mass Spectrometry (Fourier Transform Ion Cyclotron Resonance Mass Spectrometry) Using Atmospheric Pressure Chemical Ionization for Studying Composition and Thermal Degradation of Complex Materials. *Anal. Chem.*, 87(13):6493–6499, 2015. ISSN 15206882. doi: 10.1021/acs.analchem.5b00785.

[124] Yunju Cho, Arif Ahmed, Annana Islam, and Sunghwan Kim. Developments in FT-ICR ms instrumentation, ionization techniques, and data interpretation methods for petroleomics. *Mass Spectrom. Rev.*, 34(2):248–263, Jan 2015. ISSN 10982787. doi: 10.1002/mas.21438.

[125] William Kew, John W.T. Blackburn, David J. Clarke, and Dušan Uhrín. Interactive van Krevelen diagrams - Advanced visualisation of mass spectrometry data of complex mixtures. *Rapid Commun. Mass Spectrom.*, 31(7):658–662, Apr 2017. ISSN 09514198. doi: 10.1002/rcm.7823.

[126] Tim Leefmann, Stephan Frickenhaus, and Boris P. Koch. UltraMassExplorer - a browser-based application for the evaluation of high-resolution mass spectrometric data. *Rapid Commun. Mass Spectrom.*, 2018. ISSN 09514198. doi: 10.1002/rcm.8315.

[127] Logan C. Krajewski, Ryan P. Rodgers, and Alan G. Marshall. 126264 Assigned Chemical Formulas from an Atmospheric Pressure Photoionization 9.4 T Fourier Transform Positive Ion Cyclotron Resonance Mass Spectrum. *Anal. Chem.*, page acs.analchem.7b02004, 2017. ISSN 0003-2700. doi: 10.1021/acs.analchem.7b02004.

[128] Manhoi Hur, Rebecca L. Ware, Junkoo Park, Amy M. McKenna, Ryan P. Rodgers, Basil J. Nikolau, Eve S. Wurtele, and Alan G. Marshall. Statistically Significant Differences in Composition of Petroleum Crude Oils Revealed by

Volcano Plots Generated from Ultrahigh Resolution Fourier Transform Ion Cyclotron Resonance Mass Spectra. *Energy and Fuels*, 32(2):1206–1212, 2018. ISSN 15205029. doi: 10.1021/acs.energyfuels.7b03061.

[129] Kerry M. Peru, Mary J. Thomas, Diana Catalina Palacio Lozano, Dena W. McMartin, John V. Headley, and Mark P. Barrow. Characterization of oil sands naphthenic acids by negative-ion electrospray ionization mass spectrometry: Influence of acidic versus basic transfer solvent. *Chemosphere*, 222:1017–1024, May 2019. ISSN 0045-6535. doi: 10.1016/J.CHEMOSPHERE.2019.01.162.

[130] Hans Fischer. *A history of the central limit theorem: From classical to modern probability theory*. Springer Science & Business Media, 2010.

[131] Dhammika Amaratunga and Javier Cabrera. Analysis of data from viral DNA microchips. *J. Am. Stat. Assoc.*, 96(456):1161–1170, Dec 2001. ISSN 1537274X. doi: 10.1198/016214501753381814.

[132] Ben M. Bolstad, Rafael A. Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, Jan 2003. ISSN 13674803. doi: 10.1093/bioinformatics/19.2.185.

[133] Stephen J. Callister, Richard C. Barry, Joshua N. Adkins, Ethan T. Johnson, Wei Jun Qian, Bobbie Jo M. Webb-Robertson, Richard D. Smith, and Mary S. Lipton. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.*, 5(2): 277–286, 2006. ISSN 15353893. doi: 10.1021/pr050300l.

[134] Yee Hwa Yang, Sandrine Dudoit, Percy Luu, David M. Lin, Vivian Peng, John Ngai, and Terence P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single andmultiple slide systematic variation. *Sel. Work. Terry Speed*, 30(4), 2001.

[135] Taesung Park, Sung Gon Yi, Sung Hyun Kang, Seung Yeoun Lee, Yong Sung Lee, and Richard Simon. Evaluation of normalization methods for microarray data. *BMC Bioinformatics*, 4:1–13, 2003. ISSN 14712105. doi: 10.1186/1471-2105-4-33.

[136] William S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, 74(368):829–836, 1979. ISSN 1537274X. doi: 10.1080/01621459.1979.10481038.

[137] Udi E. Makov. Mixture Models in Statistics. *Int. Encycl. Soc. Behav. Sci.*, pages 9910–9915, Jan 2004. doi: 10.1016/b0-08-043076-7/00464-2.

[138] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm . Technical Report 1, 1977.

[139] G. Kitagawa and S. Konishi. *Bayesian Information Criteria.* Springer New York, New York, NY, 2008. ISBN 978-0-387-71887-3. doi: 10.1007/978-0-387-71887-3_9.

[140] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018.

[141] Hadley Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*, 2017. R package version 1.2.1.

[142] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2019. R package version 1.3.2.