Spring 5-15-2017

Mapping Analyte-Signal Relations in LC-MS Based Untargeted Metabolomics

Nathaniel Guy Mahieu
Washington University in St. Louis

Recommended Citation

Mahieu, Nathaniel Guy, "Mapping Analyte-Signal Relations in LC-MS Based Untargeted Metabolomics" (2017). *Arts & Sciences Electronic Theses and Dissertations*. 1129.
https://openscholarship.wustl.edu/art_sci_etds/1129

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Chemistry

Dissertation Examination Committee:

Gary J. Patti, Chair

Michael L. Gross

Steven L. Johnson

Jacob Schaefer

Tim Schedl

Robert Pless

MAPPING ANALYTE-SIGNAL RELATIONS IN LC-MS BASED UNTARGETED METABOLOMICS

by

Nathaniel Guy Mahieu

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2017

St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgements

My grandparents Bill and Judy: Wilder Farm, was an idyllic setting for a childhood filled with fascination. Our exploration of the surrounding Missouri woods ground my curiosity and intuition for the workings of life. Bill, your pure hearted and rational thought is my guidepost and Judy, your loving support continues to carry me forward.

My pivotal teachers Dr. McCormick, Dr. Deakyne, and Dr. Emerich at the University of Missouri: your courses gave me the tools to make sense of the world. That exemplary teaching transformed my intuitions into the knowledge I now leverage and inspired me to discover rather than apply.

My undergraduate research advisor Dr. Gates: thank you for the freedom to fail and encouragement to continue. Rule number 26, “never pretend... say I don’t know, then go ask for help” remains key in my approach to science.

My mentors in science Dr. Pless, Dr. Johnson, and Dr. Schedl: you were exemplary scientific role models. Thank you for our insightful discussions and your personal support in my journey.

My committee members Dr. Gross and Dr. Schaefer: thank you for enduring my early talks and for your guidance. I continue to be inspired by your impact on science – the bar has been set.

My advisor Dr. Gary Patti: thank you for your unwavering support through my doubts and disagreements. You provided a space for me to develop and pursue sometimes fleeting ideas. I am further indebted to your rescue of my writing’s coherence on numerous occasions. Your companionship and guidance have allowed me to grow as an independent scientist. Thank you.

Finally, funding for this work was derived from a variety of sources. We thank the following institutions: NIH Grants R01 ES022181 and R21 CA191097 as well as grants from the Alfred P.

Sloan Foundation, the Camille & Henry Dreyfus Foundation, and the Pew Scholars Program in the Biomedical Sciences.

Nathaniel Guy Mahieu

Washington University in St. Louis

May 2017

Dedicated to my father, Martin,
whose example instilled in me the value of hard work
and whose support enabled this achievement.

ABSTRACT OF THE DISSERTATION

MAPPING ANALYTE-SIGNAL RELATIONS IN LC-MS BASED UNTARGETED METABOLOMICS

By

Nathaniel Guy Mahieu

Doctor of Philosophy in Chemistry

Washington University in St. Louis, 2017

Professor Gary J. Patti, Chair

The goal of untargeted metabolomics is to profile metabolism by measuring as many metabolites as possible. A major advantage of the untargeted approach is the detection of unexpected or unknown metabolites. These metabolites have chemical structures, metabolic pathways, or cellular functions that have not been previously described. Hence, they represent exciting opportunities to advance our understanding of biology. This beneficial approach, however, also adds considerable complexity to the analysis of metabolomics data - an individual signal cannot be readily identified as a unique metabolite. As such, a major challenge faced by the untargeted metabolomic workflow is extracting the analyte content from a dataset. Successful applications of metabolomics bypass this limitation by throwing away the 99% of the dataset that is not statistically altered between sample groups.[1] This widely accepted approach to untargeted metabolomics is functional for a very narrow set of applications, but critically, it fails to provide a comprehensive view of metabolism.

The primary thrust of this dissertation work is to overcome this fundamental barrier in metabolomic experiments and extract the unique analyte content from metabolomic datasets. To this end, three algorithms were developed.

(i) We first developed the Warpgroup algorithm to refine the features detected in replicate samples. Peak detection performed on replicate samples is highly inconsistent. Warpgroup considers all replicates in concert to determine a set of consensus signals or features – integrations that are supported by all replicates. This process improves quantitation and significantly reduces the artifact content of the dataset.[2]

(ii) Mz.unity was then developed so that one can search for any specified mass-peak relationship. Features in metabolomic data are highly degenerate and available annotation approaches have been limited to a small subset of possible degeneracies. Mz.unity addresses this deficiency. This advance enabled the systematic evaluation of complex and cross polarity adducts as well as a context-based relationship recovery approach.[3]

(iii) The credentialing approach was developed to experimentally filter non-biological features and recovers a reproducible set of biological features. While great effort had been undertaken to minimize the contribution of contaminants and informatic error to features, it was clear that many mistakes were still being made.[4]

The developed algorithms were then applied, in concert, to an untargeted analysis of *Escherichia coli*. Together, the application of these algorithms provided the first comprehensive picture of metabolomic dataset composition. Strikingly, the technologies suggest that the tens of thousands of signals detected in a typical untargeted metabolomic data set correspond to less than 1,500 analytes – a result that has large implications for the design and interpretation of untargeted metabolomic experiments.

This work constitutes a key advance in our understanding of metabolomic science, and the contributions enable more robust untargeted analyses of metabolism. Together, these concepts establish a clear course for the future development of a comprehensive metabolomic data analysis platform and bring the promise of truly untargeted metabolomics into view.

Chapter 1.

Introduction

The field of metabolomics encompasses any approach that seeks to assay many metabolite analytes in a single experiment.[5,6] The term broadly applies to both targeted and untargeted techniques. Targeted analyses seek to assay a predetermined set of analytes whose structure and characteristic signals are known. In contrast, untargeted analyses seek to assay as many analytes as possible, including unexpected or unknown species. This can encompass components represented by tens of thousands of signals.[7] The scope of metabolomics varies by practitioner; a commonly employed definition limits metabolites to any biochemically produced analyte with mass less than 1000 Da. This is a rough cutoff, though, as metabolites range in mass; for example cardiolipins reach over to 1500 Da.

Metabolomics employs a variety of instruments, each with distinct strengths.[8] Seminal metabolite profiling was performed using solution phase Nuclear Magnetic Resonance (NMR) techniques that provided concentrations of a small set (about 50) of resolvable analyte signals.[9,10] Though lacking in sensitivity, NMR is capable of elucidating molecular structure, probing pathway fluxes, and determining isotopomer patterns due to its atom-specific information. Mass spectrometry (MS) is the primary analyzer applied in metabolomics.[11] MS offers high sensitivity and can resolve many thousands of analyte species based on their mass-to-charge ratio. MS also offers some structural information with the ability to fragment analytes. This process reveals masses of the fragments that provides some insight into the structure of the analyte.

Prior to analysis by MS it is common to employ a separation technique such as liquid chromatography, gas chromatography, or capillary electrophoresis.[12–14] These techniques provide a somewhat orthogonal separation to MS and aid in the distinction of isobaric metabolite signals. Separation prior to detection also reduces ionization suppression, increasing quantitative accuracy and sensitivity.

Metabolism is the network of enzyme-catalyzed reactions that convert small molecule intermediates into the energy and building blocks that enable life.[15,16] Each reaction takes several reactant metabolites, often two to four and guide them down a reaction path to produce a distinct set of product metabolites.[17,18] A single metabolite species can participate in many of these reactions and, thus, the entire network can be represented by a large, interconnected graph. (Figure 1) The Human Metabolome Database (HMDB) currently lists 6302 metabolic reactions, a figure that underscores the scope of this network.[19]

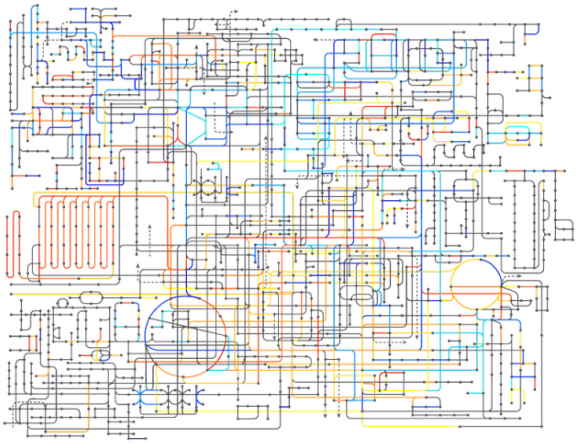


Figure 1.1. An example metabolic network. Nodes represent metabolites or multiple metabolites. Edges represent reactions, most of which are enzyme-catalyzed.*

[*] Image by J3D3 (Own work) [CC BY-SA 4.0 (http://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

Organisms cope with the demands of survival by controlling fluxes through these pathways.[20] From developing embryos to a sprinting cheetah, organisms must cope with changing energy demands, demands for biosynthetic substrate, and waste-product excretion.[21] Most of this regulation occurs via gene expression and post-translational protein modifications (on timescales ranging from seconds to days).[22–24] These and similar mechanisms act as control points in the metabolic network operating in accordance with the genetic blueprint of the organism.[25] These control points allow an organism to direct the flow of metabolic flux to satisfy the demands of growth and survival.[26,27] In this light, metabolism can be viewed as the product of many generations of natural selection that have fine-tuned the metabolic program to the organisms niche. [28–31]

In cases where the genetic program becomes damaged, a subset of cells in the organism become unable to respond appropriately to the environment – this damage results in disease such as cancer or phenylketonuria. Metabolism is the aggregate output of higher levels of regulation. As such, physical phenotypes such as disease are often accompanied by corresponding changes in the metabolic network.[32] Thus, a quantitative readout of an organism's metabolic network can be strongly predictive of disease states.[33] This has proven to be the major application of metabolomics, - clinical biomarker discovery and correlating observed phenotypes to metabolic changes.[34–36] Notably, this aggregation of the complex regulatory cascade into a metabolic state makes metabolomics an appealing experiment, but this same aggregation limits the application of metabolomics to elucidation of mechanism. In general, metabolomics provides hints as to where to investigate further, but can only provide an abstract fingerprint left by the complex process.[37] Defining mechanism requires investigation of the upstream effectors employing appropriate techniques.

1.1 Challenges to Untargeted Metabolomics

Untargeted metabolomics seeks to assess as many metabolic intermediates as possible. A major advantage of the untargeted approach is the detection of unknown metabolites. Reporting on unsuspected signals enables researchers to uncover surprising, unhypothesized metabolic interactions and previously unknown metabolic intermediates. This major benefit also adds considerable complexity to the analysis of metabolomics data.

The nature of the techniques used to analyze samples in untargeted metabolomics produces immensely complex datasets. Solvent impurities and plastic leachables appear among the metabolite signals and artifacts are introduced owing to informatic error. Chromatographic peak shapes are often non-ideal, and single analytes can appear as multiple, distinct chromatographic peaks. Degeneracy of the detected signals is a major additional source of complexity. Degeneracy refers to multiple signals arising from a single analyte. There are many causes of degeneracy including fragmentation, analyte adduction with various charge carriers (e.g., a proton, sodium, potassium, etc.), and the detection of naturally occurring isotopes (e.g., ^{13}C, ^{15}N, etc.). A final, largely under-annotated source of degeneracy is the adduction of an analyte with other species present, including other analytes and the chemical background.

A notable implication of untargeted approaches is that an individual signal cannot be readily identified. Similarly an individual signal cannot readily be discerned from the sources of complexity detailed above.[38,39] In traditional, targeted approaches the problem is sidestepped –precise masses, or fragment transitions are known prior to analysis, and that allows for rapid filtering of most or all irrelevant features. Unfortunately, application of targeted filtering removes the unhypothesized and unknown metabolites that make metabolomics such a powerful technique. As such a major challenge faced by the untargeted metabolomic workflow is extracting the analyte content from a dataset.[40,41] This limitation has significantly impeded the interpretation of metabolomic datasets and further hindered its wider adoption.

Successful applications of metabolomics begin by throwing away the 99% of the dataset that is not statistically altered between sample groups. This universally accepted approach to untargeted metabolomics is functional for a very narrow set of applications but **critically, it fails to provide a comprehensive view of metabolism**. The primary goal of this dissertation is to overcome this fundamental, and key barrier in metabolomic experiments and to extract the unique analyte content from metabolomic datasets.

Feature inflation also causes many detected signals not to be found in metabolomic databases. Investigators have interpreted the large number of unidentified signals detected in these datasets to imply that there are hundreds to thousands of unknown metabolites in these datasets.[42] This has varying implications for the experimental design of metabolomics experiments as well as biological experiments in general. A secondary goal of this dissertation is to estimate the number of analytes detected in an untargeted metabolomic experiment.

Ultimately, I posit that irrelevant signal and degeneracy in metabolomic datasets account for over 99% of the signal therein and has significantly impeded metabolomics success. To address these artifacts and degeneracies I develop three algorithms – Warpgroup, Credentialing, and mz.unity. Finally I apply the developed algorithms in concert to an analysis of *Escherichia coli* and produce the most comprehensive picture of the composition of a metabolomic dataset to date. These results demonstrate a clear path forward for the future of metabolomic analysis.

1.2 Experimental Techniques

In the last section I outlined the goals for metabolomics and the current challenges to those goals. In this section I delve into the experimental methodology we employ to perform untargeted metabolomics, and how those methods contribute to the aforementioned challenges.

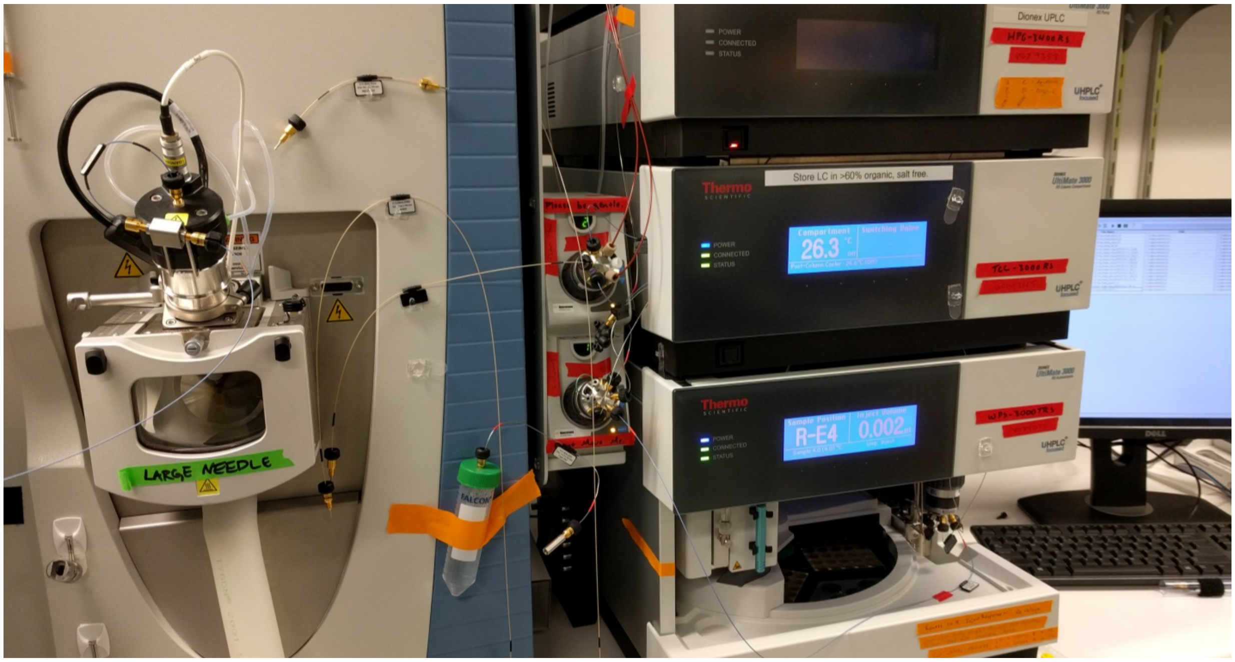


Figure 1.2. A photo of an LC-MS workstation, the components of which are discussed below. (Left) A portion of the Q-Exactive Mass Spectrometer. The prominent portion of the instrument is the electrospray source chamber. Entering the chamber from the top is the nebulizer that contains the electrospray needle. (Center) A Dionex Ultra Performance Liquid Chromatograph. Capillaries that carry solvent and analytes can be seen bridging between the LC and source. The bottom is an auto sampler that aspirates and introduces analytes into the sample flow. Above that is a temperature controlled compartment containing the column. At top is the UPLC pump that generates the high pressure gradient. (Right) A computer where the informatic processing that is a major component of the workflow takes place.

1.3 Liquid Chromatography

Very early chromatographic separations were performed by Schönbein who placed paper slips in liquid mixtures, observing the overlapping bands as they traveled at different rates.[43] This early separation of components formed the basis of modern chromatography. Paper chromatography improved throughout the 20th century eventually inspiring the use of these principles to fractionate mixtures.

Liquid chromatography enables the separation of a mixture into its constituents on the basis of their physiochemical characteristics. A mixture is introduced to the chromatographic system as a narrow band. Separation proceeds when each component moves though the system at different rates. Key to this separation are the stationary and mobile phases. The stationary phase is employed to retain or impede the progress of analytes while the mobile phase is employed to facilitate elution.[44]

Analytes possess distinct affinities for each the stationary phase and mobile phase. An analyte that has a high relative affinity for the stationary phase will spend time partitioned there and elute only after long periods of time. Conversely, an analyte with a high relative affinity for the mobile phase will partition there and move more closely to the rate of the mobile phase, eluting after a shorter time. In this way, chromatography leverages the physiochemical characteristics of analytes to separate them in time.

Owing to the wide range of polarities, not all analytes may be practical to elute using a single mobile phase.[6] Introduction of a gradient in which the mobile phase is altered throughout an experiment allows a wider range of polarities to be eluted within a reasonable time frame (see Figure 1 for a diagram of the partitioning process at two points during a gradient elution.) As the mobile phase composition changes, analytes affinities for the mobile phase will change, altering their partitioning and ultimately eluting them.

Analytical chromatography is a highly refined technique, and different approaches can separate structural isomers, regioisomers such as cis and trans double bonds, and even stereoisomers with appropriately chosen stationary and mobile phases.[45,46]

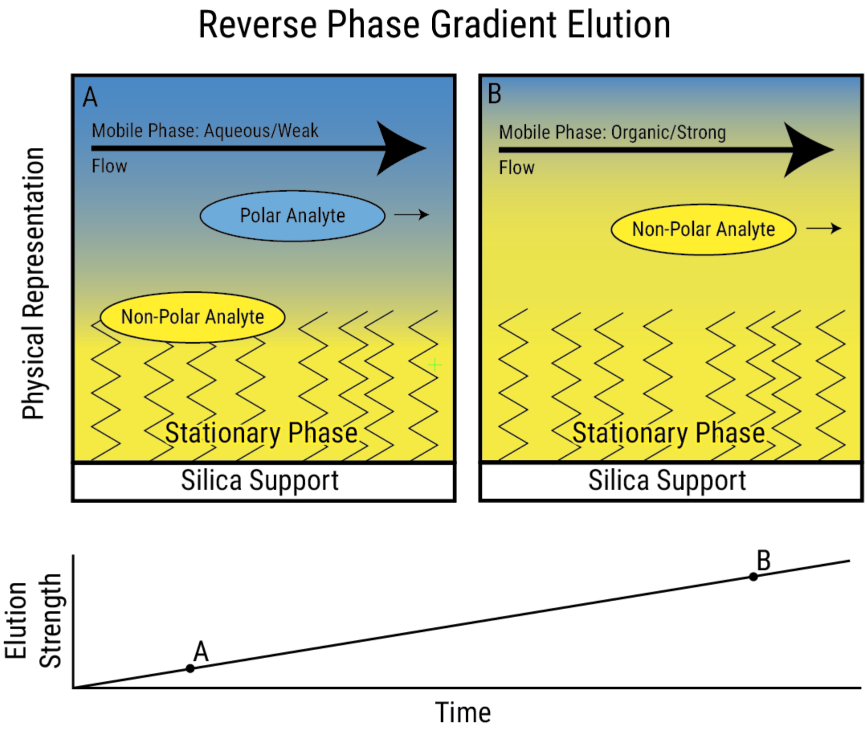


Figure 1.3. The basis for gradient chromatographic separations. (Left) Analytes are more attracted to the stationary phase than the mobile phase. At this stage the apolar analyte moves slowly through the column. (Right) At some later time the mobile phase strength has increased. The apolar analyte partitions into the mobile phase and travels more quickly through the column.

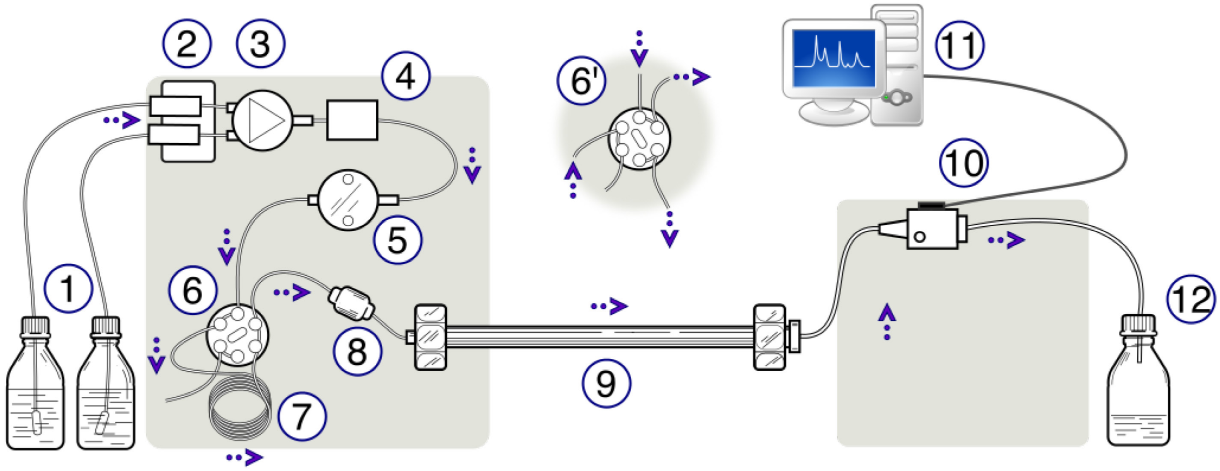


Figure 1.4. A diagram of an HPLC instrument. Mobile phase is pumped (left) through a stationary phase containing column (center) and departed analytes are detected (left).*

* By WyassineMrabetTalk. [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

The application of chromatography has become highly refined with instruments dedicated to the task of producing reproducible gradient and flow rate profiles. These reproducible techniques have enabled the use of retention time as a useful molecular descriptor. With a specific stationary phase, flow rate, and mobile phase gradient profile, multiple labs can analyze an analyte and observe the same retention time. In this way retention time is indicative of the species being observed.

Depicted is a split loop injection system in which two flow paths can be selected by a switching valve. The sample is loaded into an injection loop that the mobile phase flow is bypassing. Upon injection the valve switches, and the injection loop becomes part of the flow path, washing the mixture downstream. At this point the mobile phase and mixture reaches the column containing the stationary phase. Modern columns are manufactured with a variety of substrates, but most commonly a porous silica material is derivitized to produce a stationary phase with desired properties. Inside the column, the analytes contained in the mixture are slowed, and begin to proceed through the column at varying rates. Ultimately, the components elute from the column at varying times and are then detected. When using mass spectrometry as a detector, an ionization method is needed; both ionization and mass analysis will be discussed in the following sections.

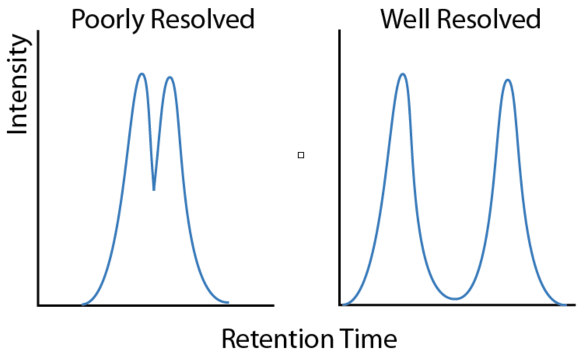


Figure 1.5. Examples of chromatographic resolution. (Left) The peaks of poorly resolved analytes overlap. (Right) The peaks of well resolved analytes do not overlap.

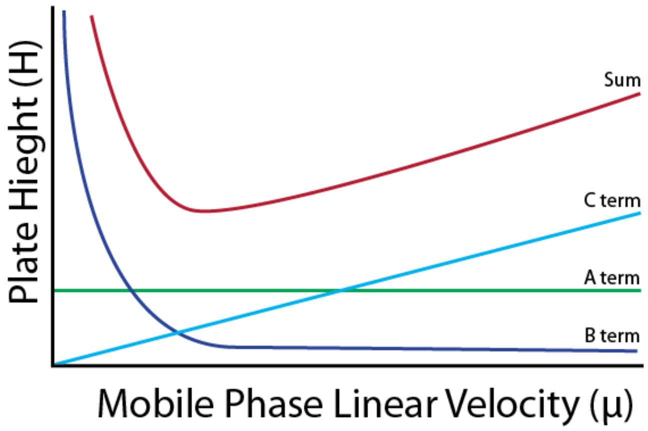


Figure 1.6. Effect of mobile phase flow rate on resolution. A Van Deemter Plot describing the efficiency of a chromatographic separation as a function of varying mobile phase flow rates. Three terms can be used to model the resolution, each due to a physical process causing peak broadening. The C term increases with flow rate, the B term decreases with flow rate, and the A term is flow rate independent.

A useful conceptualization of the physical nature of the chromatographic process is summarized with a discussion of the efficiency of a separation. In addition to selectivity (the relative partitioning of two analytes), a separation must be efficient, producing analyte bands that are narrow enough to be distinguished. Described by van Deemter and Zuiderweg[47] in 1956, the van Deemter equation outlines three terms that correspond to physical non-idealities that impact the efficiency of a separation. (Equation 1) Here, H is a measure that is inversely proportional to efficiency (peak width) and μ is the linear velocity of the mobile phase (flow rate.)[47]

$$H = A + \frac{B}{\mu} + C\mu \quad (1)$$

The A term in the van Deemter equation is the longitudinal diffusion component, that is inherent to the stationary phase. As analytes travel through the column, they will take different paths, some longer and some shorter. Higher quality packings can decrease this term but it is invariant with respect to the flow rate.

The B term represents the longitudinal diffusion of particles in the mobile phase and is dependent on the analyte, temperature, and solvent viscosity. Diffusion of the analyte bands occurs as long as the analytes are dissolved in the mobile phase – as a result, slow flow rates and longer analysis times result in more diffusion.

The C term represents an analyte's resistance to mass transfer between the mobile and stationary phases. As an analyte diffuses in and out of the pores of the mobile phase, they will spend varying amounts of time in the pore. At higher flow rates, this variance in time impacts peak widths more – this is because during the time when one particle was not moving, another has moved rapidly.

The sum of these terms gives the expected efficiency of a separation. Notably, the only penalty for flowing at higher rates is the C term. Modern columns with very small packings and core shell packings have mostly eliminated the flow rate-dependent increase of the C term. This allows for very fast flow rates with little to no efficiency penalty – a major advantage of the small particle size packings and corresponding high pressure UPLC technique.

1.4 Electrospray Ionization

After elution from a liquid chromatography experiment, analytes exist in the bulk liquid phase. All mass spectrometry experiments are performed on ions in the gas phase, as such a requisite step is the transfer of analytes from the bulk liquid to the gas phase and imparting a charge to them. The coupling of liquid chromatography and mass spectrometry was a challenging goal, the two operating under opposite extremes of pressure (liquid vs 10^{-5} torr in the gas phase.)[48] Initial work to this end was performed by Dole in 1968 using polystyrene spheres[49] – this work was eventually refined by Fenn in 1984 into the early electrospray source.[50]

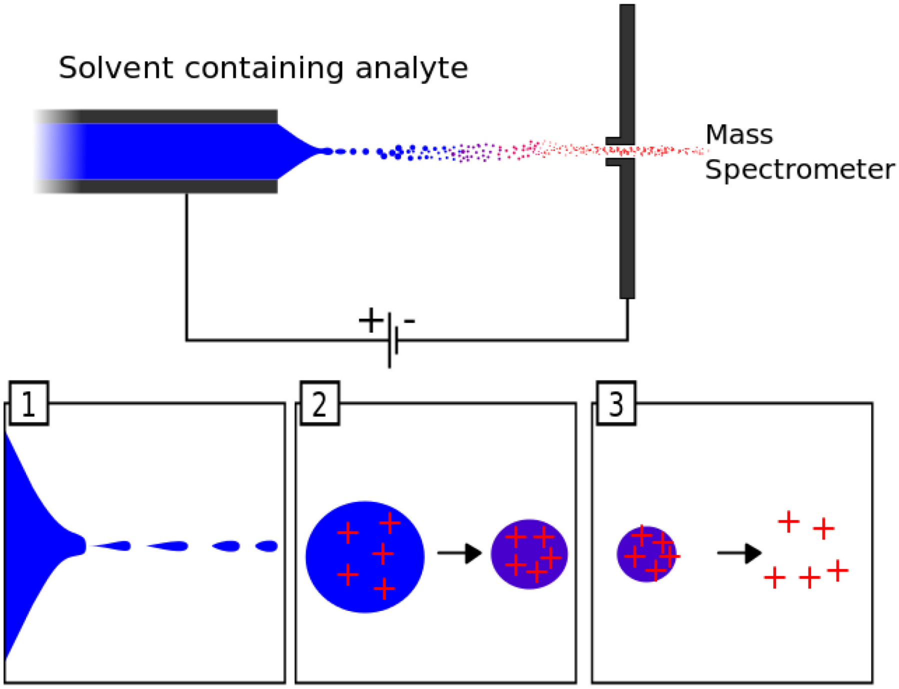


Figure 1.7. A schematic of Electrospray ionization. (Top) A potential difference of several thousand volts is applied between the capillary and mass spectrometer inlet. This high field region induces the Electrospray process. (Bottom) 1. The Taylor cone formed due to charge accumulation and surface tension. 2. Droplets shrink due to solvent evaporation and coulombic explosion. 3. Gas phase ions are produced by continued evaporation and charge expulsion from the droplets.*

The modern process of electrospray ionization proceeds by forcing the liquid through a small needle into a region with high electrostatic fields. This electrostatic region causes charges to concentrate at the surface of the liquid – the surface tension of the liquid, combined with the electrostatic repulsion of charges in solution form a cone as described by Taylor in 1964.[51] As the liquid surface grows small toward the apex of the cone droplets bud off and travel down the potential gradient.[52] (Figure 1.7) These droplets shrink, both by solvent evaporation and the expulsion of smaller charged droplets. This process continues until individual analytes are introduced to the gas phase, some with a charge.[53]

The formation of adducts is a key feature of electrospray ionization. The charge imparted to gas-phase analytes is often the result of the adduction of an analyte with a charge carrier, often a proton

[*] By Evan MAson (Own work) [CC BY-SA 4.0 (http://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

(H^+) or sodium (Na^+).[54] Adduction, though is not restricted to small charge carriers and in general any present species can adduct with the eluting analytes. This process gives rise to many degenerate ion species that are derivatives of the original analyte – a phenomena that gives rise to much of the complexity in untargeted metabolomic datasets. Addressing and overcoming this complexity is the focus of Chapter 4.

1.5 Mass Spectrometry

After ionization, analytes must be transferred from the source region at atmospheric pressure to a low pressure region of around 10^{-5} torr.[48] This is achieved by a combination of a potential difference down the ion path and pressure difference between the two regions.[55] The ion plume in the atmospheric pressure region is propelled by the electric potential difference towards the inlet of the low pressure region. As ions approach the inlet, the rapid acceleration of the bulk gas because of decreasing pressure causes a turbulent flow through the transfer capillary towards the differentially pumped regions. As the gas exits the ion transfer capillary, it rapidly expands into the low pressure region. A series of ion funnels, skimmers, and multipoles act to contain the ions as neutrals are pumped away. Throughout this process the ions are cooled, and off-axis velocity is dampened by electric fields until a beam of ions that is suitable for further analysis by a mass analyzer is produced.[56] The pressure reached is a function of the type of mass analysis performed – quadrupole-based analysis operates at pressures around 10^{-4} torr whereas analyses requiring a long mean free path (such as TOF or orbitrap experiments) operate at pressures up to 10^{-10} torr (a mean free path of several kilometers!)

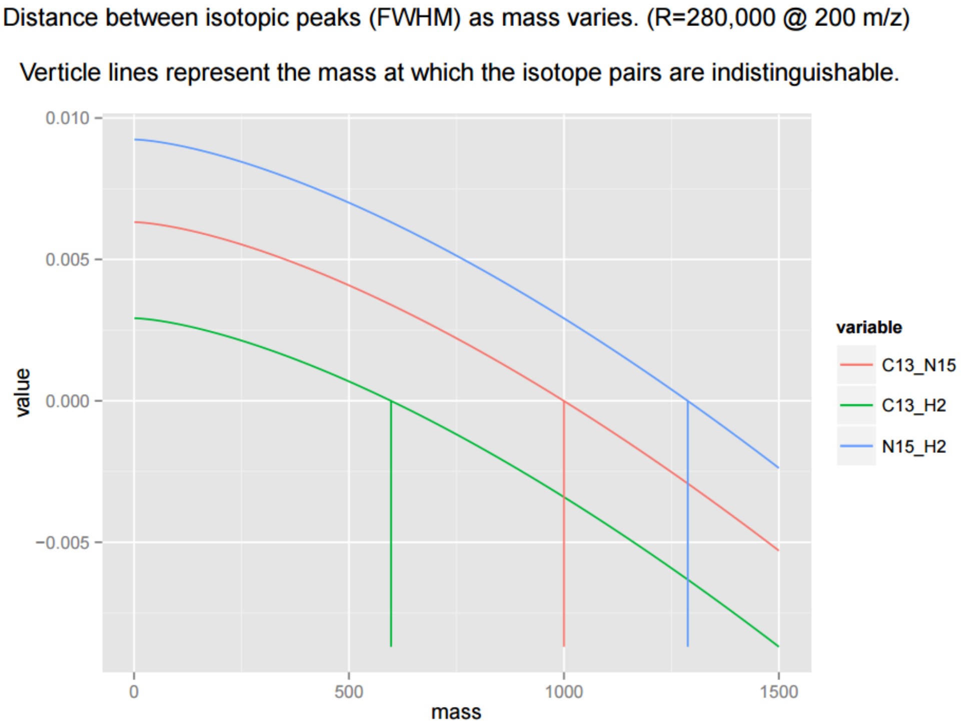


Figure 1.8. Resolution's mass dependence on the Q-Exactive illustrated with three challenging to resolve A1 species. As mass increases instrument resolution decreases. This results in distinct mass peaks becoming more poorly resolved. (Green) 13C and 2H become indistinguishable at 550 Da. (Red) 13C and 15N become indistinguishable at 1000 Da. (Blue) 15N and 2H become indistinguishable at 1300 Da. Based on experimentally observed resolutions at 280,000 resolving power.

1.5.1 The Q Exactive: Quadrupole-Orbitrap Mass Spectrometer

The Q Exactive (QE) mass spectrometer is a Fourier Transform (FT) based instrument that offers ultra-high mass resolving power, high mass accuracy and exquisite sensitivity. The mass spectrometer couples two mass analyzers, a quadrupole and an orbitrap, with a C-trap and collision cell intervening.[57] The orbitrap is an ion-trapping device, into which a narrow beam of ions is injected perpendicular too and off center from a central spindle electrode.[58] Ions are confined between the central electrode and the outer shell by electric fields, and begin to orbit perpendicular to the central electrode due to their initial velocity. The distribution of initial velocities and positions perpendicular to the central electrode cause the ion packet to spread into a ring around the central electrode. Motion of the ion rings parallel to the central electrode follows naturally owing to the off-center injection of ions into the potential well. Ion motion along this parallel axis is dependent on the potential well and critically, the mass-to-charge ratio of the ions. (Equation 2) Thus, ion motion is observed as the image current in the two outer detector plates. This time domain signal can be transformed into a frequency domain spectrum (with frequency proportional to *m/z*) by FT.[58]

$$\omega = \sqrt{\frac{k}{m/z}} \qquad (2)$$

Injection of ions into the orbitrap was a considerable design challenge – the initial distribution of position and momentum of ions parallel to the central electrode was limiting frequency determination. Focusing of ions in this dimension is achieved by the C-trap, which upon injection compresses the ion packet into a narrow ribbon for injection. As the QE is a trapping instrument, observation of the ions occurs in a pulsed manner. Depending on the desired mass resolving power, orbitrap analysis can take as long as 1000 ms (for a resolving power of 256,000.)[59] To minimize this limitation's effect on duty cycle, ions can be accumulated in the C-trap and fragmented in the collision cell with parallel acquisition of an orbitrap spectrum.

Even still, the orbitrap is inherently charge limited – as charge density in the orbitrap becomes too high dephasing of ion packets and saturation of signal amplifiers can occur. For this reason, duty cycle limitations for the QE are most often due to high ion flux rather than ion sampling limitations. This has an interesting practical result relevant to untargeted studies – the limit of detection is dependent on the ion flux during a particular scan. When a particularly abundant group of ions elute they occupy a large fraction of the charge capacity of the orbitrap. Thusly, the limit of detection for the entire scan is increased because fewer of other species will be accumulated.

The charge limitations of the orbitrap mass analyzer necessitate a rationing of space in the trap. Rationing is accomplished by quadrupole mass filtering prior to ion accumulation. In the low abundance case, when ion flux is limited, the orbitrap is an exquisitely sensitive instrument, with comparable sensitivity to triple-quadrupole type instruments. Similar to the triple quadrupole, the QE offers nearly 100% duty cycle when not charge saturated. Additionally, the QE is able to observe ions for an extended amount of time and offers high mass resolving power of possible interferences in the monitored fragments. These factors allow the QE to equal and exceed the triple quadrupole in targeted sensitivity.

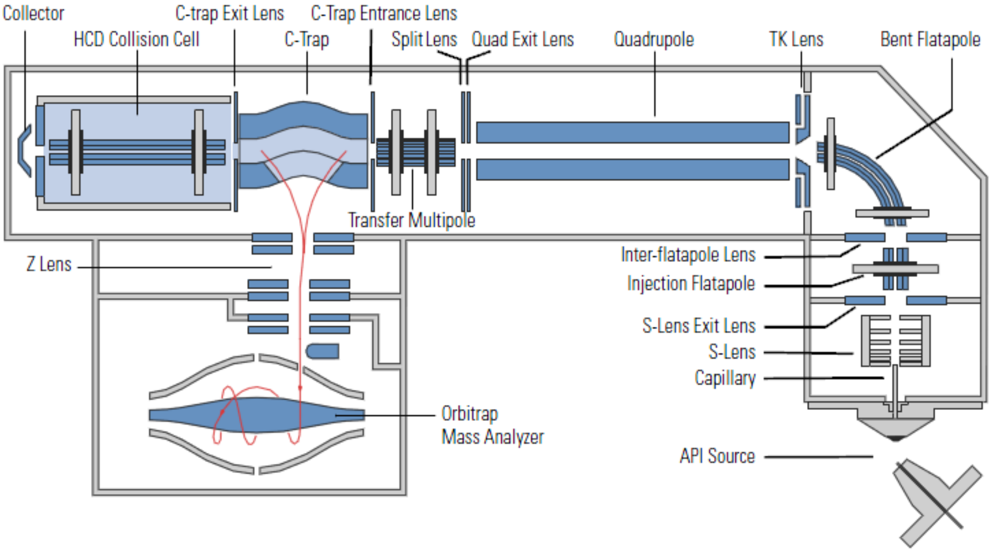


Figure 1.9. A schematic of the Q-Exactive mass spectrometer.*

1.5.2 The Quadrupole-Time-of-Flight Mass Spectrometer

The Quadrupole-Time of Flight Mass Spectrometer (QTOF) is a hybrid instrument coupling a quadrupole and collision cell to a time-of-flight mass analyzer.[60] Time-of-flight mass analysis is achieved by a high voltage pulse accelerating a packet of ions into a drift region. This pulse imparts the same amount of kinetic energy to each ion, but ions with different mass-to-charge ratios will travel at different speeds owing to the conservation of momentum. As such, ions are separated

[*] Taken from the Q-Exactive user manual, with permission.

based on the time it takes them to travel a several meter distance and detected upon impacting a detector plate. (Equation 3) Differences in the initial positions and velocities of the ions contribute to peak broadening, this is mitigated by the use of a repulsive electrostatic region that reflects the ions back in the direction they came from, focusing each mass packet.[61]

$$t = k\sqrt{m/z} \qquad (3)$$

The QTOF mass spectrometer is a pulsed instrument, requiring drifting ions to reach the detector prior to the next pulse of ions. As such, duty cycle on these instruments is around 10%. Improvements to this include using Hammond transforms and overlapped pulses, as well as gating of the ions. As opposed to the QE, the QTOF has a very large charge capacity, and is limited primarily by detector saturation. As such, other ions in the spectrum have no impact on the overall sensitivity of a scan and the QTOF is suitable for bright ion sources.

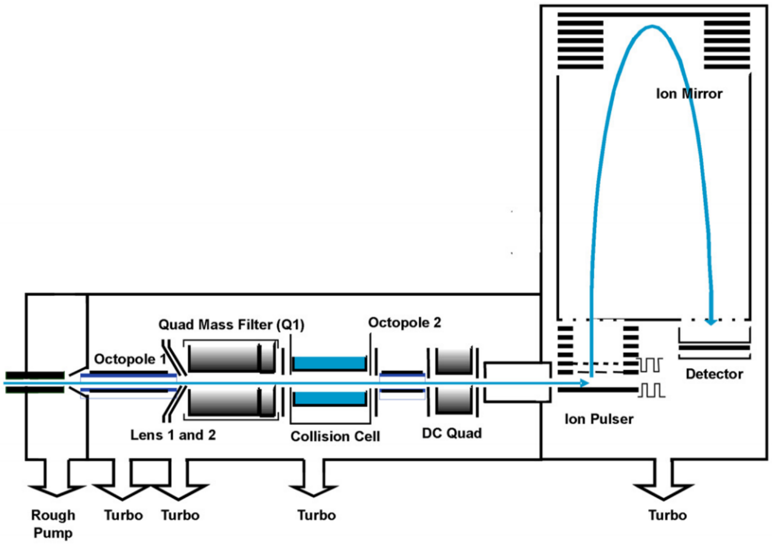


Figure 1.10. A schematic of quadrupole-time-of-flight mass spectrometer.*

* Taken from the Agilent 6500 QTOF Manual, with permission.

1.6 Informatic Techniques

The metabolomic workflow involves several processing steps. The contributions herein leverage the following two informatic techniques to improve upon this workflow.

1.6.1 Dynamic Time Warping

Dynamic Time Warping (DTW) is the name of a class of algorithms that find optimal alignment between two time series.[62] The method takes as input two time series and finds a mapping (warping) from time in series A to time in series B. These algorithms have been applied to a wide variety of alignment problems including speech recognition, genetic sequence alignment, and even data that has no true time component such as image alignment.[63,64]

Given two time series $X = (x_1 \dots x_N), Y = (y_1 \dots y_M)$ and some dissimilarity function $d(i, j) = f\left(x_i, y_j\right) \geq 0$, dynamic time warping seeks to find the warp path ϕ that minimizes the accumulated distortion $d_\phi(X, Y) = \sum_{k=1}^{T} d\left(\phi_x(k), \phi_y(k)\right) * m_\phi(k)/M_\phi$ between the two timeseries. ($m_{\phi(k)}/M_\phi$ is a normalization factor such that paths with different numbers of steps have comparible distortion values.) The result is a warp path with *k* steps $\phi(k) = (\phi_x(k), \phi_y(k))$ where $\phi_x(k) \in \{1 \dots N\}$ and $\phi_y(k) \in \{1 \dots M\}$.[62] (Figure 12)

Dynamic time warping approaches are relevant to chromatographic based techniques.[65] Between chromatographic experiments many sources of variation cause shifts in the elution time of components – sample matrix differences, temperatures, pump fluctuations, column cleanliness, and even analyte concentrations can impact retention times.[66] Prior to statistical analysis of abundances, correspondence must be determined to link the same analyte detected in each sample. DTW provides an approach to evaluate and mitigate some of these variances that benefits later correspondence determination. We apply aggressive dynamic time warping in the Warpgroup algorithm, which is developed in Chapter 3.[2]

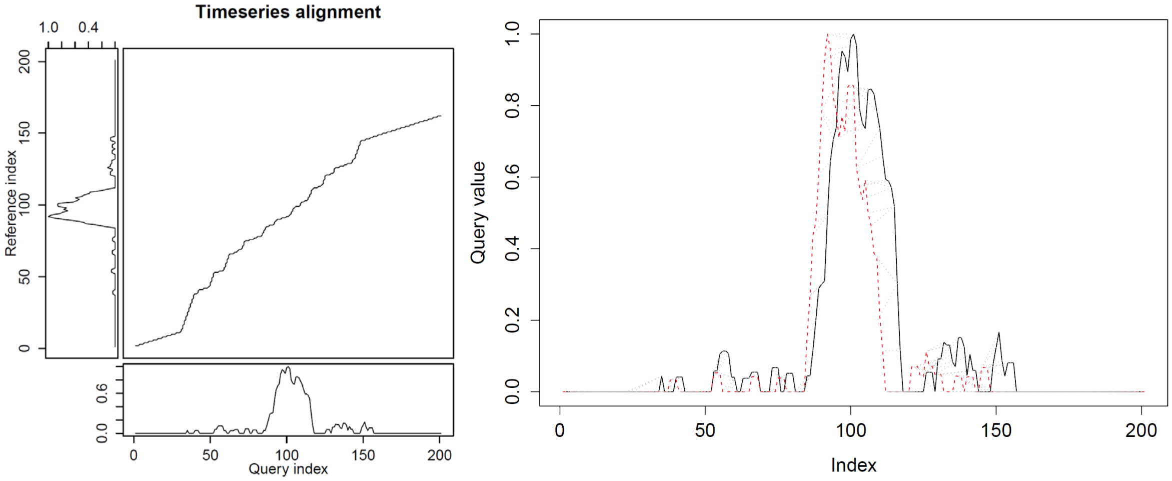


Figure 1.11. An example of the dynamic time warping alignment of two time series. (Left) Two time series are plotted on the bottom and the left. The optimal alignment which maps time in the reference to time in the query is plotted in the center. (Right) The same time series overlaid with the alignment plotted as grey connections between the points.

1.7 Graph Theory

Graph theory is the study of pairwise relationships between objects. It is a broad field that is applicable to a range of disciplines from rigorous treatment in discrete mathematics to applications in fields such as computer science and biology.[67,68]

A graph is comprised of a set of nodes and edges. Edges link exactly two vertices and can be directed or undirected.[69] Graphs can be used to represent a wide variety of problems, such as pathing through a network of roads (edges) and intersections (nodes) or finding relationships (edges) between people (nodes) in a social graph.

In Chapter 5 we utilize graphs to represent additional structure in the mass spectrometry data. Particularly interesting is the relationship graph formed by mass spectral peaks (nodes) that are transformations (edges) of a common analyte prior to detection. In general graphs can be used to represent many relational data structures. As such they are a general concept where specific details emerge in the context of specific problems.

As applied in this manuscript graphs are used in two ways. In Chapter 3, graphs are used to represent peaks which could be the same analyte across replicates. Nodes are used to represent detected peaks and edges are used to represent pairs of peaks which we posit are the same analyte. In this application, sets of well connected nodes are more likely to be the same analyte. Later, in Chapter 4, graphs are used to represent the underlying structure in mass spectra. Specifically, nodes represent detected mass-to-charge peaks and edges represent putative relationships between these peaks. (Figure 1.12)

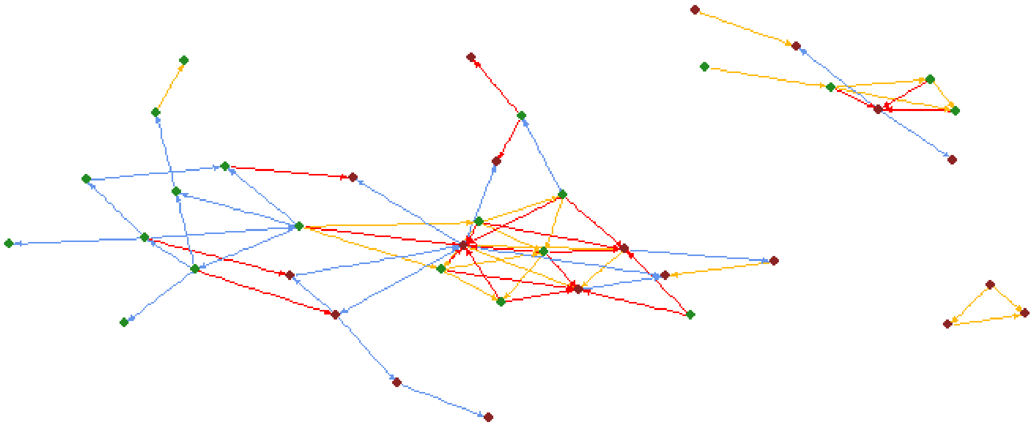


Figure 1.12. An example graph plotted from relationships detected in a mass spectrum. Nodes are the dots. Edges are the lines connecting the dots.

1.8 Raw Data

Liquid chromatography/mass spectrometry (LC/MS)-based techniques generate a sequential series of mass spectra at around 10 Hz. The result is a three dimensional dataset with axes of retention time (RT), mass-to-charge ratio, and intensity. A raw, digitized profile dataset on modern instrumentation (e.g. the Q-Exactive) contains 1x10^{7} *m/z* and intensity pairs per scan (at 140 k resolving power). This yields an impressive 1x10^{10} (ten billion) data points per 30 minute experiment (at 3 Hz). In practice, many points are zero, and peak detection on profile mass spectra is reliably performed such that only about 1x10^{6} points are used for the metabolomic analysis.

The metabolomic workflow's primary goal is the continuing reduction of raw, abstract data points into an entity representing an observed analyte.

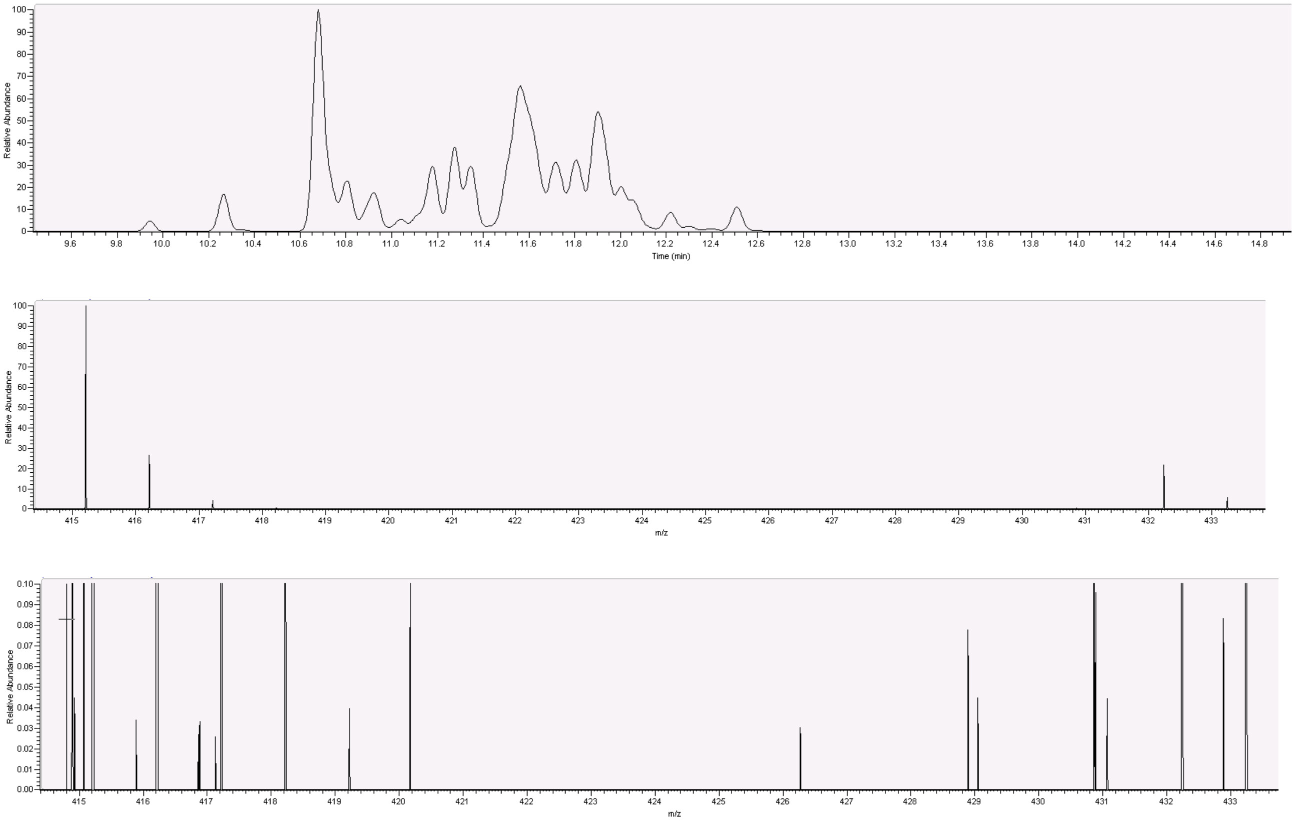


Figure 1.13. An example of raw mass spectral data. (Top) and extracted ion chromatogram of mass 415.1728-415.2599. (Middle) The mass spectrum at 10.6 minutes. (Bottom) The same mass spectrum zoomed 100x.

1.9 Feature Detection

Analytes elute across multiple scans, and mass peaks appearing in several sequential scans with a Gaussian-like profile are termed features. The first step in the processing of metabolomic data is peak detection. In general peak detection seeks to gather mass peaks produced by a single analyte, and observed in sequential scans into a single entity termed a "feature". A feature is thus a triplet (a composite of three values) consisting of the mean retention time, mean *m/z*, and integrated intensity observed.[70]

Feature detection is a common challenge and a wide variety of algorithms have been employed. One commonly employed feature detection algorithm in metabolomics research is centWave[71] - it proceeds via two steps. Initially centWave detects regions of interest, which are regions in the *m/z* and RT dimensions with a high density of mass observations – these are putative regions that may contain a chromatographic peak shape. The second phase of centWave applies a pattern wavelet based peak detection looking for a peak shape in the chromatographic domain. The ion's retention time profile is analyzed after applying a discrete wavelet transform at multiple scales. Peaks that are observed at many scales are initialized, and peak bounds are specified by descending to the nearest local minimum. The detected peaks above a specified signal to noise ratio are then retained.

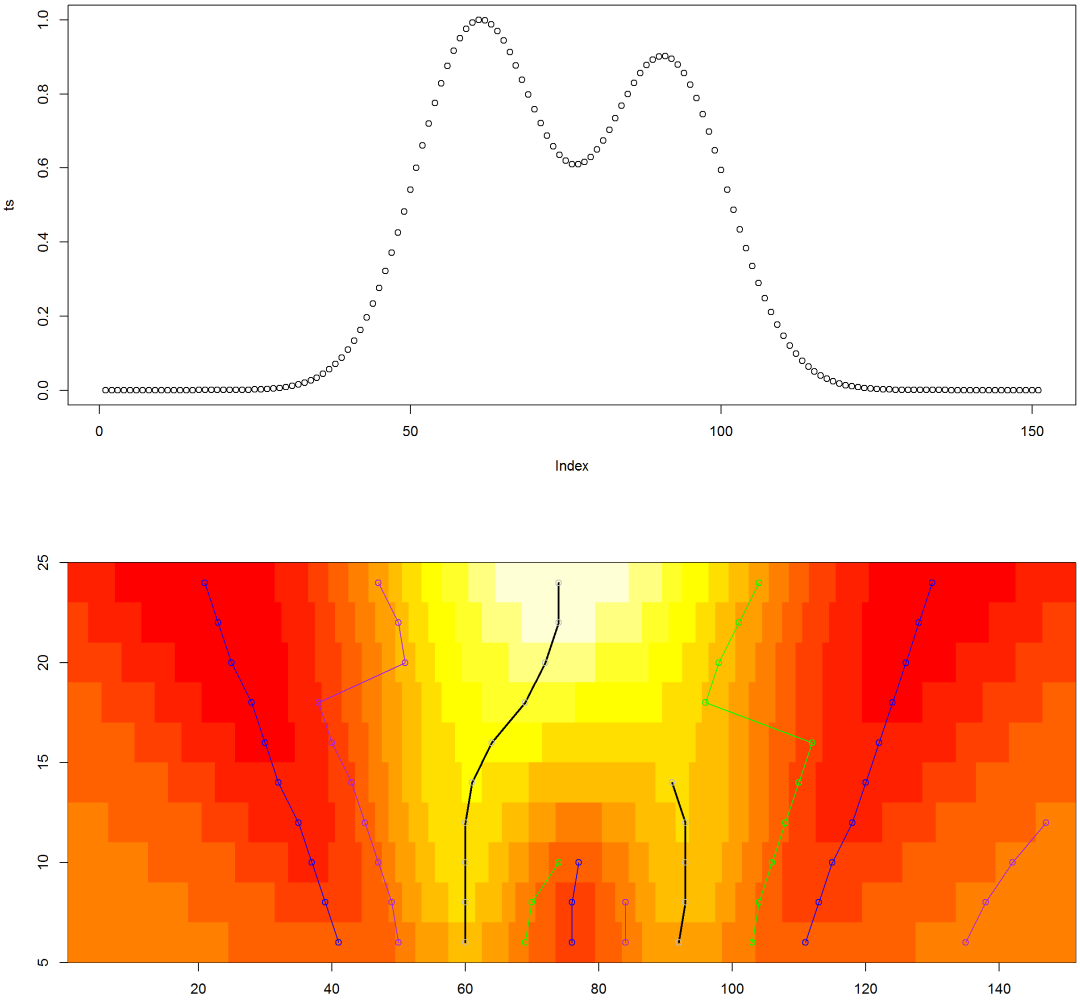


Figure 1.14. Discrete wavelet transformation of a bimodal distribution. (Top) The bimodal distribution which is to be transformed. (Bottom)The wavelet transform of this distribution. The scale of the wavelet transform is plotted on the y-axis. Note that different scales see drastically different representations.

Peak detection achieves two goals. Importantly this process removes a major fraction of the signal which does not exhibit a peak shape, and, is therefore, not relevant to the injected sample. These signals are often chemical background, or other, slowly eluting compounds. Secondly, peak detection determines the integration region of each feature, thus determining the quantitation which is later used for statistical analysis and biological inference.

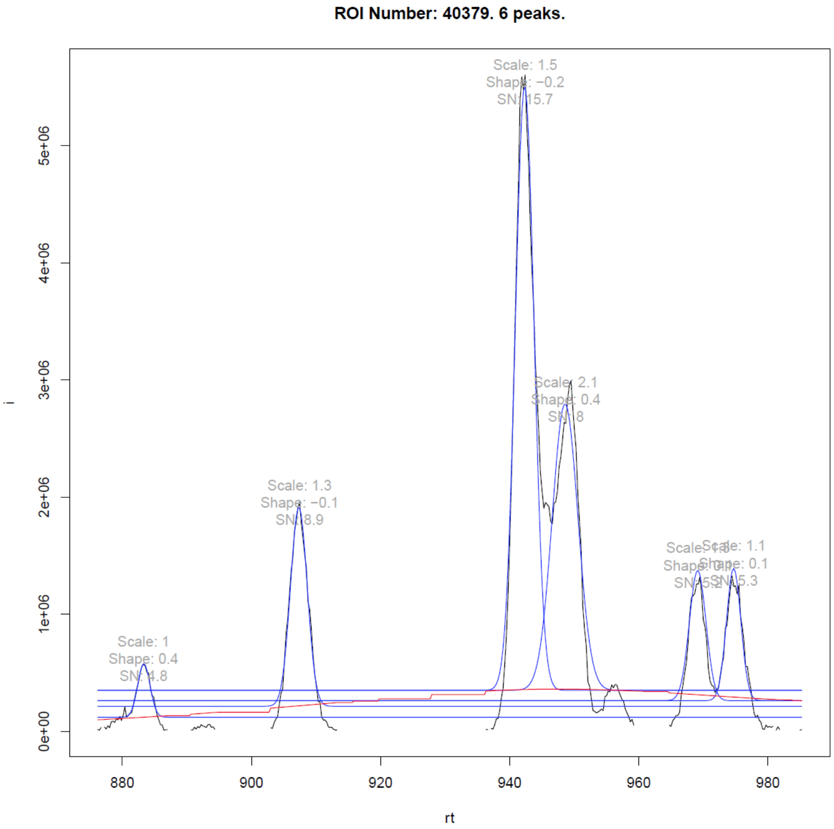


Figure 1.15. An example of detected peaks. (Black) The raw chromatographic trace is plotted for a single mass. (Red) A calculated baseline estimate. (Blue) Detected peaks.

Complicating this task are several sources of variation. From scan-to-scan a single mass-to-charge value is measured with only finite precision. As such, mass error on the order of 1-10 ppm complicates region of interest determination. The intensity of an ion signal as it is sampled also includes significant variance (consider the complex ionization process described above), thus, a chromatographic trace often fluctuates non-monotonically, complicating chromatographic feature detection. Finally, it is common in complex mixtures that eluting analytes are not fully, chromatographically resolved. Defining peak bounds for poorly resolved components is challenging even when performed manually.

The peak detection process reduces the initial 1x10^6 mass peaks to around 2x10^4 features. The output is a list of features and their corresponding *m/z*, RT, and intensity. This peak list is used for further analysis. Chapter 3 in this dissertation deals with refinements to the peak detection process while Chapter 4 details a method for further consolidating these peaks into unique analyte groups.

Chapter 2.

A Roadmap for the XCMS Family of Software Solutions in Metabolomics[*]

Global profiling of metabolites in biological samples by liquid chromatography/mass spectrometry results in datasets too large to evaluate manually. Fortunately, a variety of software programs are now available to automate the data analysis. Selection of the appropriate processing solution is dependent upon experimental design. Most metabolomic studies a decade ago had a relatively simple experimental design in which the intensities of compounds were compared between only two sample groups. More recently, however, increasingly sophisticated applications have been pursued. Examples include comparing compound intensities between multiple sample groups and unbiasedly tracking the fate of specific isotopic labels. The latter types of applications have necessitated the development of new software programs, which have introduced additional functionalities that facilitate data analysis. The objective of this review is to provide an overview of the freely available bioinformatic solutions that are either based upon or are compatible with the algorithms in XCMS, which we broadly refer to here as the "XCMS family" of software. These include CAMERA, credentialing, Warpgroup, metaXCMS, X13CMS, and XCMS Online. Together,

[*] This work is based on the following publication: "A roadmap for the XCMS family of software solutions in metabolomics". NG Mahieu, JL Genenbacher, GJ Patti, Current opinion in chemical biology, 2016. NGM provided the overview of processing algorithms, analysis of retention time drift, the credentialing algorithm and the warpgroup algorithm.

these informatic technologies can accommodate most cutting-edge metabolomic applications and offer some advantages when compared to the original XCMS program.

2.1 Introduction

In the last chapter I outlined the techniques applied in LC/MS-based metabolomics. In this chapter we provide a more in depth explanation of the informatic workflow with specific detail regarding application of the XCMS software package. XCMS is the most well-known, open-source metabolomics software and makes applying several algorithmic steps easy. The algorithms developed in chapters 3, 4, and 5 either refine or extend this functionality but are independent of XCMS. Overviews of the algorithms developed in chapters 3 and 5 are provided in sections 2.7 and 2.6, respectively. Finally, the algorithm developed in chapter 4 supersedes CAMERA, which is described in section 2.5.

Data from liquid chromatography/mass spectrometry (LC/MS)-based untargeted metabolomic experiments are highly complex. Therefore, bioinformatic software is typically required for processing of the results. At this time, there are many reliable software solutions available.[72–80] It is not the purpose of this review to comprehensively detail each, nor is it our intent to provide any type of comparative evaluation. Rather, we will exclusively focus on a selection of freely available software solutions that are interoperable with the XCMS program. Some of these software solutions bear variants of the XCMS name, while others do not. We broadly refer to the class as a whole as the “XCMS family”.

2.2 Defining the Needs: A General Bioinformatic Workflow

Historically, the bioinformatic workflow for processing untargeted metabolomic data has involved three general steps: feature detection, correspondence determination, and context-dependent analysis of the resulting measured values (Figure 1).[81,82] Each is briefly described below.

1. The first and perhaps most important step is feature detection (also known as peak detection or peak picking). The purpose of this step is to extract from the dataset signals that arise from real compounds, while attempting to exclude signals resulting from various noise sources.[83] Extracted signals with a unique mass-to-charge ratio and retention time are recorded as features. (Figure 2A)

2. The second step in the workflow is establishing correspondence between the features detected from different sample runs. Correspondence refers to establishing those features from different analytical runs that "correspond" to the same analyte. Establishing correspondence is arguably the most challenging step in the processing of untargeted metabolomic data.[82] Although the same analyte may be detected in multiple experimental runs, the measured mass-to-charge ratio and retention time of the analyte can vary in each run owing to factors such as temperature fluctuation and column degradation (Figure 3A). Importantly, many drift factors are compound specific and, therefore, global-alignment techniques cannot be used for correction (Figure 3B).[84]

In practice, the majority of investigators performing LC/MS-based metabolomics currently assert correspondence by aligning the time domains of each run with time-warping techniques. (Figure 2B) The objective is to correct for drift factors so that features can be grouped between samples by direct matching of retention time. Although the alignment approach for establishing correspondence has enabled many laboratories to analyze untargeted metabolomic data successfully, there remains a great need for robust correspondence determination algorithms, and this remains an active area of research interest.[82]

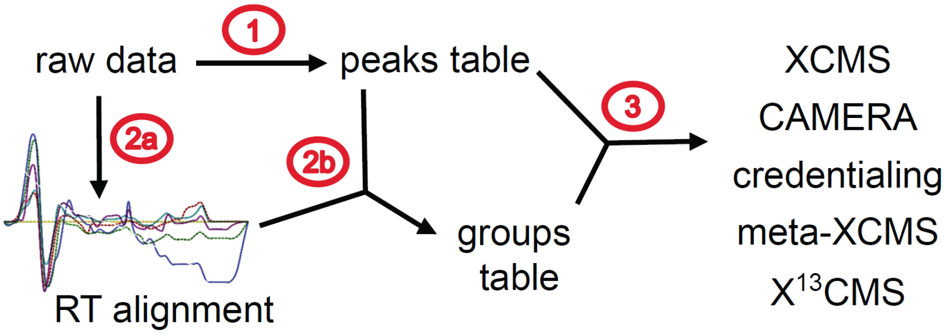


Figure 2.1. The bioinformatic workflow for processing untargeted metabolomic data with XCMS. The workflow has three general steps: 1. Feature detection, 2. Correspondence determination, and 3. Additional context-dependent analysis. These steps are numbered in red on the schematic. After acquisition of LC/MS profiling data, feature detection is performed on the raw data to generate a peaks table (step 1). Next, retention time drift is corrected (step 2a). The OBI-warp algorithm implemented within XCMS operates on the raw data to determine retention time drift. This produces a retention-time correction map that, together with the peaks table, is used to establish correspondence and generate a groups table (step 2b). The peaks table and the groups table are the input for a variety of further analyses. The third step is dependent upon experimental objectives. In the standard XCMS analysis, step 3 is statistical analysis. The other programs listed use the peaks table and groups table to achieve different aims such as adduct and artifact annotation, multiple-factor analysis, and isotopic label tacking.

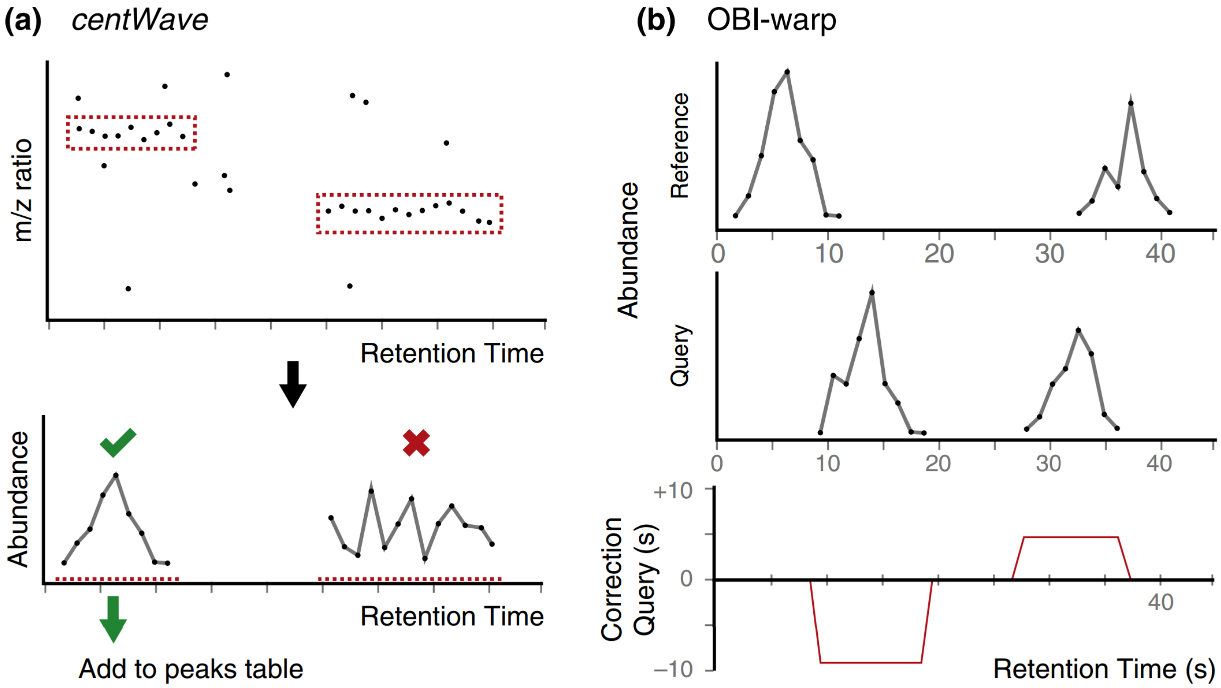


Figure 2.2. Schematic of the centWave and OBI-warp algorithms as implemented within XCMS. A. The first step in centWave is to find consecutive scans in which peaks are detected within a specific mass error (top). These are referred to as regions of interest (ROIs) Two such ROIs are displayed here and boxed in red. Second, extracted ion chromatograms are created for each ROI (bottom). Extracted ion chromatograms that display a peak shape are then added to the peaks table, as illustrated by the green checkmark and arrow. B. OBI-warp aligns a query sample to a reference sample. Here we illustrate a representative example in which two features are shifted in the query sample compared to the reference sample. Application of the correction curve to the query (bottom) brings the samples into alignment.

3. The last step of the workflow is context dependent. Analyses diverge, depending on experimental goals. In the simple cases when the objective is to compare sample classes, this step amounts to performing statistical analysis on the intensities of detected features. For more advanced objectives such as isotope tracing or tandem mass spectral analysis, additional algorithms are required.

2.3 Introducing XCMS

In 2006, the XCMS software was published as one of the first programs to provide a complete solution to the bioinformatic workflow outlined above for processing untargeted metabolomic data.[81] The "X" in the XCMS acronym is used to denote that the software can be applied to any form of chromatography. To date, however, XCMS has been predominantly used to process LC/MS-based metabolomic data. The original XCMS software used the matchedFilter algorithm to accomplish feature detection, the retcor.peakgroups algorithm to perform alignment (an application of LOESS regression to well-behaved peak groups), and the group.density algorithm to group aligned features across samples on the basis of *m/z* bins. In recent years, a new algorithm for feature detection called centWave and a new algorithm for alignment called OBI-warp have been implemented within XCMS.[65,71] It is worth noting that whereas these algorithms have led to better overall XCMS performance, there is still great opportunity for improvement. It is exciting to consider, for example, that there are hundreds of published algorithms for peak detection and correspondence determination that have not yet been implemented within XCMS for comparative evaluation.[82,85]

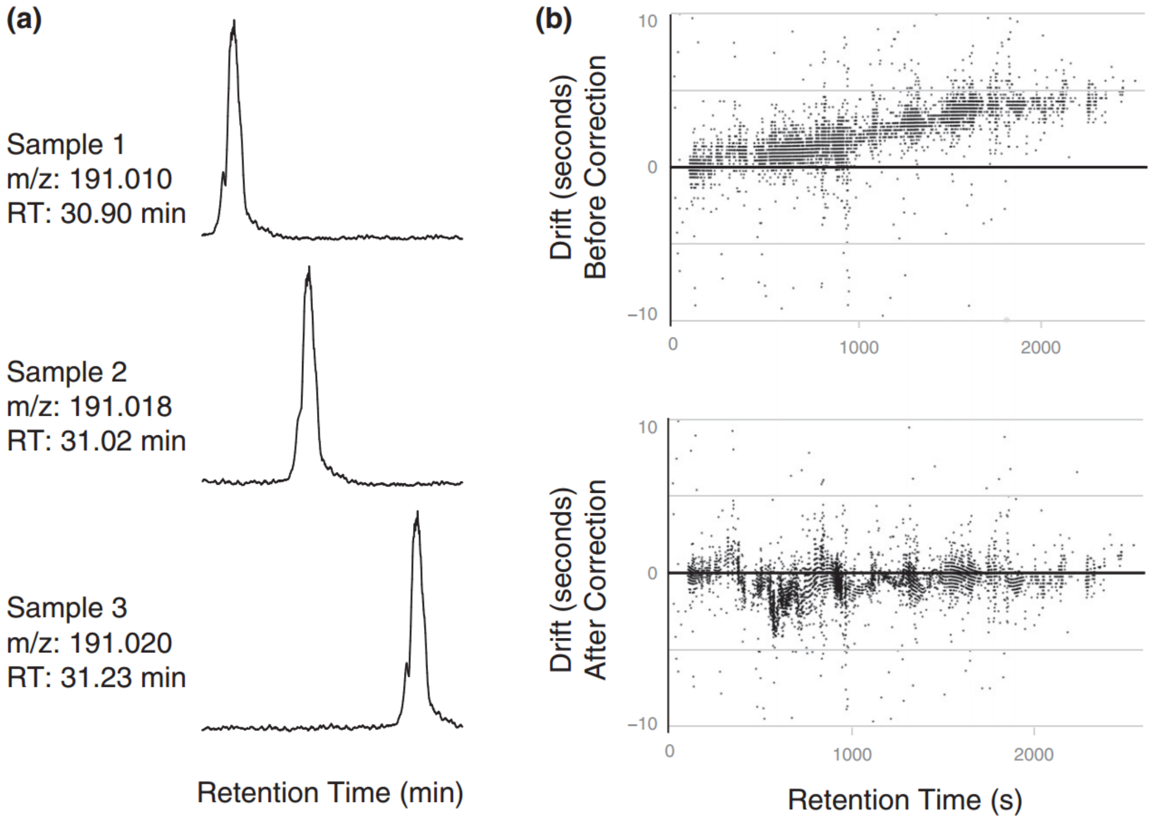


Figure 2.3. Illustrating the correspondence problem. A. Extracted ion chromatograms of citrate from three samples show that its retention time and its measured mass-to-charge values vary between three samples run back to back. B. Uncorrected retention time drift of all features detected in sample 2 as compared to sample 1 (top). Uncorrected drift remaining after OBI-warp correction. (bottom). Note that though correction reduced the overall drift, there is no global correction which will perfectly align all peaks due to multiple, compound-specific drifts occurring at a single retention time.

Applying the centWave, OBI-warp, and group.density algorithms within XCMS results in what are known as the peaks table and the groups table (Figure 1). In the standard application of XCMS, the peaks table and the groups table are then used to create a diffreport. The diffreport provides statistics on feature groups that have altered intensities between sample groups.[86] When the original XCMS software was published in 2006, generating such a diffreport in the programming language R was considered cutting edge. From the diffreport, investigators can count the number of features detected from a sample to crudely compare metabolomic workflows.[87,88] More importantly, researchers can use the determined p-values and fold changes to find features with statistically significant changes in intensity between two sample groups. However, the XCMS diffreport also has some serious limitations. It does not provide metabolite identifications, which generally require matching tandem mass spectra from the research sample to the tandem mass spectra of authentic standards.[89] Additionally, the diffreport does not provide a reliable approximation of metabolites detected as adducts, isotopes, fragments, and artifacts.[4,90] Indeed, depending on experimental conditions, more than 50% of the features on a diffreport can be fragments and artifacts.[91] As the field of metabolomics has evolved over the last decade, there has been a major push to better annotate the XCMS diffreport. Multiple bioinformatic strategies that are interoperable with the XCMS program have now emerged to enable identification of adducts, isotopes, artifacts, and in some cases even structures.[92] A selection of these resources is detailed in the sections that follow.

Also note that the XCMS diffreport was designed for evaluating features with altered intensities between only two sample classes. Yet, there are a growing number of applications with more sophisticated experimental designs involving multifactorial analysis and stable isotope labeling. These types of applications require that step 3 of the bioinformatic workflow shown in Figure 1 diverge from that of the standard XCMS program. Thus, new software solutions have been developed that operate on the peaks table and the groups table with unique algorithms (examples highlighted below.)

2.4 A Clarification on Terminology

As multiple programs have emerged with variants of the XCMS name, it may be confusing for new investigators to distinguish which software is appropriate to use for specific applications. As an example, XCMS2 was the first program to be related in name to the original XCMS software.[93] Sometimes the program's name is written as XCMS2, which may suggest that it implements a new generation of algorithms for the core functionalities of XCMS. However, XCMS2 only differs from

XCMS in its ability to process tandem mass spectral data. We will not discuss XCMS2 further in this review. Processing of tandem mass spectral data will be covered in our discussion of XCMS Online.

Below, we highlight software programs that are interoperable with XCMS and provide key solutions to some common challenges in untargeted metabolomics. Most of these programs use the XCMS peaks table and/or groups table as their inputs. Therefore, collectively, we refer to them as the XCMS family of software.

2.5 CAMERA: Annotating Isotopologues, Adducts, Clusters, and Fragments

When a metabolite is analyzed by electrospray ionization-mass spectrometry (ESI-MS), it is usually detected as more than a single ion species in the same mass spectrum owing to the presence of isotopologues, adducts, clusters, and in-source fragments.[94] Because these ion species have different mass-to-charge values, XCMS reports each as a unique feature.[5] This increases the complexity of the XCMS diffreport and complicates statistical analysis as well as compound identification.

Given that adducts, clusters, and fragments are generally formed at the source in ESI-MS, they share the same retention time as the parent compound. Similarly, isotopes usually do not influence retention.[86] Thus, a strategy widely employed to group these types of related features is evaluation of chromatographic peak shape similarity.[95] The approach has been used by several software programs, but here we describe CAMERA because it was designed for postprocessing of the XCMS output.[86] Like XCMS, CAMERA is freely available from the Bioconductor repository.

In addition to grouping related features, CAMERA also attempts to annotate ion species by applying a rule table. The rule table works for identifying isotopes, frequent adducts such as sodium and chloride, and common neutral losses or cluster-ions. Users also have the option to combine LC/MS data from positive and negative modes to improve the reliability of ion annotations.

2.6 Credentialing: Annotating Artifacts

In a conventional LC/MS-based metabolomic experiment, the XCMS diffreport includes a large number of "artifactual" features. These features significantly complicate interpretation of the data because they are not directly associated with the sample but rather arise from contaminants introduced during analysis or from chemical noise, bioinformatic noise, etc.[4] Unfortunately, information in the XCMS diffreport is insufficient to discriminate artifactual features from biological features. Artifacts are particularly problematic when attempting to interpret metabolomic data at the comprehensive level. When evaluating different analytical methods to compare metabolome coverage, for example, we demonstrated that higher feature numbers do not necessarily correlate with more detected metabolites.[4] In part, this is because artifacts are highly variable and change as a function of extraction procedure, separation technology, mobile phase, instrumentation, and mass spectrometer settings.

Currently, approaches to identify artifacts in metabolomic data rely upon stable isotopes.[4,90] Although these strategies have proven effective, we should point out that their application is limited to samples that can be cultured with labels (clinical specimens remain a challenge). One approach for removing artifacts, known as credentialing, was designed to be interoperable with the XCMS software.[4] In the credentialing scheme, artifactual features are distinguished by growing cells on heavy isotopic carbon and mixing them with natural-abundance samples at defined ratios. Notably, only features of cellular origin will have appropriate isotopic partners at the appropriate ratios. Thus, without structurally identifying every feature, artifacts can be filtered from the dataset computationally by using the credentialing software algorithms. With this platform, the number of "credentialed features" can be used (instead of total features) as a more reliable metric to benchmark analytical performance.

2.7 Warpgroup

The standard XCMS workflow employs the centWave and group.density algorithms to detect peaks in each sample independently. In this scheme, the information used to group peaks is only the average *m/z* and retention time from all samples analyzed. Further, as each sample's raw data are treated in isolation, differences in integration regions between samples contribute to increased variance in the processed dataset. We developed Warpgroup as an XCMS compatible package that addresses these limitations with consensus integration bound analysis.[2] Warpgroup applies dynamic time warping and graph analysis to improve the precision of metabolomic data processing. Warpgroup improvements include: correspondence determination that leverages the local extracted ion chromatogram topography; detection and grouping of peak subregions; selection of similar integration bounds for each group; intelligent missing value filling; and reporting of several parameters which allow the filtering of bioinformatic noise.

The benefits of Warpgroup are the retrospective combination of several independent rounds of peak detection. For an E. coli dataset, as an example, application of Warpgroup resulted in an increase in the number of unique detected analytes by 26% and halved the mean coefficient of variation of all analytes (compared to the XCMS algorithms alone).[2] Warpgroup is implemented in a general manner and is applicable to all time series data, including metabolomic data from other software packages.

2.8 metaXCMS: Finding Shared Alterations Among Multiple Sample Classes

The original XCMS algorithms were designed to compare the intensities of features from only two sample groups. The challenge of applying simple pairwise comparisons is that knocking out a single protein can lead to hundreds or thousands of changes in feature intensities because the related pathways are interconnected.[96] For instance, knocking out a protein may decrease the product of that protein. However, decreased levels of the protein's product may then itself lead to a cascade of other context-dependent metabolic alterations. Determining those metabolites that are altered directly as a result of knocking out a protein from those that are altered indirectly is challenging. Thus, it has become increasingly common in metabolomics to look for dysregulation shared among multiple sample groups as a strategy for data reduction. metaXCMS enables such multiple-factor comparisons by operating on XCMS diffreports.[97,98]

The power of assessing shared metabolic differences among multiple sample groups is perhaps best demonstrated by an example. When control *C. elegans* worms were compared to long-lived *C. elegans* worms in which the germ line had been removed by glp-1 mutation, ~44% of the total detected features (13639) were altered with a p-value <0.05 and a fold change >2.[96] From these data alone, features directly associated with increased life span could not be distinguished from those features that were altered from glp-1 mutation but that did not affect life span. Because germ-line-induced extensions in life span are dependent upon the FOXO transcription factor DAF-16, double mutant daf-16;glp-1 worms are short lived. Thus, a comparison of long-lived glp-1 worms to both wildtype worms and short-lived daf-16;glp-1 worms with metaXCMS revealed shared features that were uniquely altered in glp-1 induced longevity. By performing similar analyses of other long-lived worms with metaXCMS, the number of features directly associated with longevity was ultimately reduced to six.[96]

2.9 X^{13}CMS: Unbiased Mapping of Isotopic Fates

Although the intensities of thousands of features are measured by LC/MS-based untargeted metabolomics, these data provide only a static snapshot of cellular metabolism and do not generally capture the complex dynamics of biochemical pathways.[99] To quantitate metabolic fluxes and to determine the contribution of specific nutrients to metabolite/macromolecular synthesis, investigators typically use isotope-labeled tracers.[100] A number of robust approaches, such as metabolic flux analysis, are well established for these types of studies.[72] Most of the approaches use mass spectrometry or NMR to measure isotopic labeling in a targeted set of compounds.

In recent years, there has been a growing interest to integrate untargeted metabolomic technologies with stable isotopic tracers. One potential advantage of such an experimental design is the unbiased and comprehensive tracking of metabolite fates.[101,102] By following the metabolism of a labeled compound fed to a biological system comprehensively as a function of time by using LC/MS-based metabolomic approaches, new metabolite transformations may be discovered. Additionally, by comparing labeling patterns between different phenotypes using global metabolomic technologies, it is possible to identify relative changes in flux distributions.[103]

The XCMS software is not currently designed to support experiments involving isotopic labels. Although analysis of isotopic labels can be accomplished by using XCMS together with CAMERA, the X13CMS software was recently developed specifically to support experimental designs based on stable isotopes.[86,103] To use X13CMS, LC/MS data acquired from samples with and without isotopic label are first processed by XCMS. The XCMS results are then forwarded to X13CMS, which identifies isotopologue groups corresponding to isotopically labeled compounds. Grouping of isotopologues is performed without any a priori knowledge except input of isotopic label(s) used, instrument mass accuracy, and chromatographic drift tolerance. The labeling pattern of each compound determined to be isotopically enriched can be quantitatively compared from multiple sample groups by using the getIsoDiffReport algorithm implemented within X13CMS.

2.10 XCMS Online: Metabolomics on the Cloud

The bioinformatic resources discussed up to this point are distributed as R packages and operated through a command-line interface or customized scripts. One major advantage of this format is flexibility. Researchers can modify the XCMS algorithms to suit their own specific needs. The modular nature of the original XCMS software has made it interoperable with new generations of programs for untargeted metabolomics and enabled multiple research laboratories to improve upon the original XCMS algorithms.[65,71,104,105]

A limitation of distributing XCMS as an R package is that many users do not have the programming expertise to use a command-line interface. This can be particularly problematic for clinical and biological laboratories. In response to this issue, an intuitive graphical interface was developed to process untargeted metabolomic data; this interface implements many of the algorithms described in this review including those in XCMS, CAMERA, metaXCMS, as well as others. The platform, called XCMS Online, is cloud based.[106] Investigators upload untargeted metabolomic data by simply dragging and dropping their files into the program. Parameters are then selected and processing occurs on the cloud. Researches receive an e-mail notifying them when processing is complete. Results can then be viewed online, or downloaded for later use. An advantage unique to XCMS Online is that data are directly searched against the METLIN metabolite database.[107] When users upload both MS and MS/MS data, the matching can be performed on the basis of accurate mass and fragmentation patterns.[92] Thus, within XCMS Online, features on the diffreport can be annotated as possible isotopes, adducts, or structures.

2.11 Concluding Remarks

There are many reliable bioinformatic solutions for processing untargeted metabolomic data. The XCMS software is one platform-agnostic solution that is widely used. The success of XCMS is related to it being open source and highly modular. This has enabled multiple laboratories to contribute to its development with algorithms such as centWave and OBI-warp. There are a multitude of additional algorithms available that are relevant to the processing of untargeted metabolomic data, and it is recommended that their potential to improve XCMS performance be evaluated in the future. Given that XCMS is open source and modular, it is also interoperable with new generations of metabolomic software implemented within R and aimed at achieving advanced functionalities (e.g., better annotation of features, multifactorial analysis, unbiased tracking of isotopic labels, etc.). Consequently, the core algorithms within XCMS have become an important piece of many bioinformatic pipelines. Hopefully the roadmap for these pipelines that we have provided here will be useful in helping researchers chose a software platform most compatible with their experimental objectives.

2.12 Acknowledgements

GJP received financial support for this work from the National Institutes of Health Grants R01 ES022181 and R21 CA191097, as well as the Alfred P. Sloan Foundation, the Camille & Henry Dreyfus Foundation, and the Pew Scholars Program in the Biomedical Sciences.

Chapter 3.

Warpgroup: Increased Precision of Metabolomic Data Processing by Consensus Integration Bound Analysis[*]

Motivation: Current informatic techniques for processing raw, chromatography/mass spectrometry data break down under several common, non-ideal conditions. Importantly, hydrophilic liquid interaction chromatography (a key separation technology for metabolomics) produces data that are especially challenging to process. We identify three critical points of failure in current informatic workflows: compound specific drift, integration region variance, and naive missing value imputation. We implement the Warpgroup algorithm to address these challenges.

Results: Warpgroup adds peak subregion detection, consensus integration bound detection, and intelligent missing value imputation steps to the conventional informatic workflow. When compared to the conventional workflow, Warpgroup made major improvements to the processed data. The coefficient of variation for replicate injections of a complex Escherichia Coli extract were halved (a reduction of 19%). Integration regions across samples were much more robust. Additionally, many signals lost by the conventional workflow were "rescued" by the Warpgroup refinement, thereby resulting in greater analyte coverage in the processed data.

[*] This work is based on the following publication: "Warpgroup: increased precision of metabolomic data processing by consensus integration bound analysis". NG Mahieu, JL Spalding, GJ Patti, Bioinformatics, 2015. NGM developed and evaluated the Warpgroup algorithm, and ran all LC/MS experiments. JLS provided an initial, poorly aligned data.

Availability and Implementation: Warpgroup is an open source R package available on GitHub at github.com/nathaniel-mahieu/warpgroup. The package includes example data and XCMS compatibility wrappers for ease of use.

3.1 Introduction

In the previous chapter an overview of the informatic tasks in metabolomics was provided. In this chapter we take a deeper look at the peak detection process and develop Warpgroup, an algorithm that refines the results of peak detection in individual files by combining the results and computing consensus peak integrations – peak integrations that are supported by all replicates.

Omics-scale separation/mass spectrometry approaches (e.g., LC/MS, GC/MS, CE/MS, etc.) generate large, three-dimensional data sets consisting of elution time (rt), mass-to-charge ratio (*m*/*z*), and signal intensity information.[107] Analytes are separated by their chemical characteristics prior to being introduced into the mass spectrometer (yielding rt). The mass spectrometer acts as a second dimension of separation and a detector, providing information on the accurate mass (*m*/*z*) and amount of each analyte (signal intensity). Each sample run can generate gigabytes of data representing tens of thousands of distinct analytes.[108] The processing of raw data is a significant challenge and the conventional workflow consists of several steps. These steps include mass trace detection, chromatographic feature detection, inter-sample retention time drift correction, inter-sample grouping of common features (correspondence determination), and statistical analysis of feature groups.[5] A feature in this context refers to signal that displays a peak shape in both *m*/*z* and rt domains. The result of this data processing is quantification of all unique analytes detected across multiple sample runs.

Historically, most chromatography/mass spectrometry experiments have been performed with reversed-phase chromatography. This well-established separation technique commonly generates Gaussian peak shapes and exhibits highly reproducible retention times. A simple retention mechanism based primarily on compound polarity also minimizes compound specific drift.[109] One drawback to reversed-phase separation is a lack of retention for the highly polar compounds such as sugars and organic acids commonly of interest to metabolomic studies. As a result, many new separation chemistries have emerged under the umbrella term hydrophilic interaction liquid chromatography (HILIC), which aim to achieve separation of polar molecules.[110] Unfortunately, analytes measured by HILIC separation exhibit a wide range of non-Gaussian peak shapes as well as larger, compound-specific retention time drift.[111] Current informatic approaches were primarily developed by using reversed-phase C18 chromatography, and even today most new advances are benchmarked solely on reversed-phase datasets.[71] Thus, the performance of these algorithms degrades when applied to HILIC datasets.

Detection of features and selection of integration regions is an initial and critical step of the informatic workflow.[112] In cases where peak shapes are simple and peaks exhibit large signal-to-noise ratios, the detection and integration of peaks is reproducible. Complex metabolomic datasets, however, contain a high proportion of poorly resolved and low-abundance peaks.[113] Additionally, the non-Gaussian peak shapes exhibited by a large portion of HILIC features impede the robust selection of integration bounds. These factors complicate peak detection and result in undetected features as well as integration bounds which describe different regions of a peak in each sample. (Figure 1, A and B)

The second major informatic step is determination of correspondence. Current feature grouping techniques rely on the reproducible elution of compounds across multiple experimental runs. The elution time of *m/z*-rt pairs (i.e., features) is the key information used to associate the same compound detected in different runs.[114] In practice, elution times vary from sample to sample due to many factors. [114] This necessitates correction of retention time drift prior to grouping. Most techniques assume that drift is a function of retention time alone and thus generate a global correction curve $f(rt_A)=rt_B$.[66] This critical assumption is overly simplistic. In practice, retention time drift is compound dependent (Appendix 1.1 and 1.2).[82] Additionally, residual drift becomes greater when using more vagarious separation strategies such as HILIC, as larger groups of samples are aligned, and as research studies begin to incorporate inter-laboratory comparisons.[115]

Given the global correction assumption, most alignment techniques minimize only the average drift between samples considering all analytes equally.[65] (Appendix 1.1 is an optimistic example displaying the residual drift of technical replicates run over the course of 9 hours.) As such, the inherent compound-specific drift results in many unaligned peak remaining after correction – moreover many compounds move even further out of alignment upon global correction. (Appendix 1.2) This poor feature alignment causes major challenges for current peak grouping algorithms. The density method employed by XCMS, for example, can only group peaks if their maximum residual drift is less than the distance to the nearest group. (Figure 1C is an example of this failure.)[84] Further complexity is added by samples in which a feature is undetected, or when spurious noise is detected as a feature.

These failings of the current informatic workflow motivated our development of the Warpgroup algorithm. Warpgroup is an algorithm that utilizes dynamic time warping (DTW) and network graph decomposition. Herein we achieve five goals: (i.) accurate grouping of features between samples even in the case of deviation from the global retention time drift, (ii.) splitting of peak subregions into distinct groups, (iii.) determination of consensus integration bounds within each group such that each group represents a similar chromatographic region, (iv.) detection of the appropriate integration region in samples where no peak was detected, and (v.) reporting of several parameters that allow filtering of noise groups.

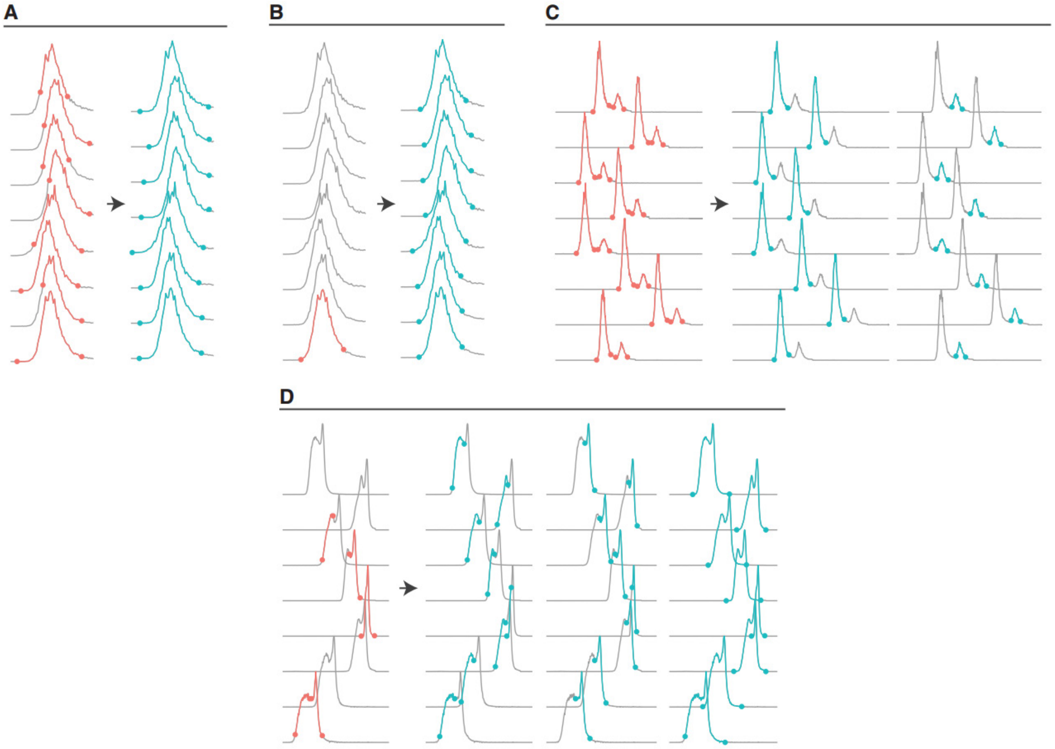


Figure 3.1. (A) Determination of integration bounds is a challenging computational problem. Independent peak detection introduces sample-to-sample variance in the integration regions (top). Peak bounds after Warpgroup (bottom). (B) Peak detection often misses peaks in some samples (top). Warpgroup detects the appropriate regions in each sample to integrate (bottom). (C) Conventional methods are unable to accurately group peaks when retention time varies more than the separation between peaks (left). Warpgroup successfully groups challenging peaks (right). (D) An extreme example in which two peaks have merged to varying degrees and peak detection identified different portions of the peak in different samples (left). Warpgroup correctly identifies the three corresponding regions in each sample (right). All examples are included in the Warpgroup R package for demonstration.

The Warpgroup algorithm establishes a correspondence between the time domains of each feature's extracted ion chromatogram (EIC) trace, utilizing dynamic time warping by default.[116,117] Based on this correspondence, Warpgroup evaluates whether all supplied peak bounds represent a similar chromatographic region using graph community detection.[69] Subsequently, it determines "consensus integration regions" for each sample and selects the appropriate integration region for samples with no detected peak. During the time warping and graph analysis, several descriptors of each group are generated and reported for use in filtering unreliable and noise-containing groups.

3.2 Methods

3.2.1 Overview of the Warpgroup Algorithm

The Warpgroup algorithm is applied after feature detection has been performed. It augments the conventional retention time correction and feature grouping steps with the addition of group splitting and consensus bound determination. The benefits of Warpgroup are derived from the combination of several peak finding rounds through the independently determined alignment between chromatograms.

The Warpgroup algorithm utilizes two pieces of information. The first is one EIC trace per sample that includes all of the masses contributing to the peak group. This trace could contain a single detected peak, or multiple peaks per sample depending on the experimental retention time drift and mass drift. These traces are used to determine the pairwise alignment between each sample's time domain for this putative group of compounds. The second piece of information is a list of peak bounds detected in the EIC traces. These must have been determined previously by a peak detection step for at least one sample. The Warpgroup algorithm will use these bounds and the aligned sample traces to split the detected peak list into groups, each of which represent a distinct chromatographic region.

The key assumption made by the Warpgroup approach is that the sample EIC traces exhibit similar topography. Though not strictly true, this is the common assumption made in current retention time alignment techniques[82] and has been shown here to be a robust basis for Warpgroup analysis. Under this assumption, we use established methods to warp (shift, expand, and contract) the time domain of the sample EIC trace such that the difference between two sample traces is minimized. In this way we establish a relationship between the two time domains, equating the scans in one sample to the scans in a second for a specific group of compounds (i.e., $f_{m,n}(scan\ in\ sample\ m) = scan\ in\ sample\ n$. This warping function is taken as the true correspondence between scans in each sample trace and is used to establish relationships between the detected peaks as well as to determine the proper integration region in samples where a peak was not detected.

To this end, the alignment between each sample scan is used to evaluate whether the supplied peak bounds delineate similar or distinct chromatographic regions of their EIC traces. Peak bounds which describe similar chromatographic regions should overlap upon transformation into a second sample's time domain. We ask, for each peak, if the transformed bounds agree. These yes/no answers are expressed as linkages between detected peaks (nodes) creating a graph structure. This graph is split using the walktrap community detection method[118] and the resulting communities are taken as peak groups (i.e., groups of peaks that describe similar chromatographic regions).

For each resulting peak group, the full set of transformed peak bounds is then filtered for outliers that do not describe a chromatographic region similar to that of the majority of detected peaks. The mean of the 75th percentile of the remaining, transformed peak bounds for each sample is taken as the "group-consensus peak bound" for each sample.

Finally, integration bounds must be determined for samples which have no detected peak remaining in the group. It is common for features to be detected in some but not all samples, especially in cases where compounds are of low abundance. Each group's consensus peak bounds are transformed into the missing sample's time domain and the median of these transformed consensus peak bounds is taken as the integration region for the missing sample.

In this way Warpgroup has assured that each peak group contains a region from every sample, each peak group describes a unique chromatographic region, and all peaks in that group describe a similar chromatographic region (Figure 1).

3.2.2 Description of the Warpgroup Algorithm

Input

The algorithm takes two pieces of information. A sample × scan sample-trace-matrix (Figure 1, traces) and a matrix of peak bounds including the peak start, and peak end, and sample index (Figure 1, dots).

Sample Trace Preprocessing

Optionally, each sample trace is smoothed, padded with 0's equal to 10% of the length of the trace, and normalized to a maximum intensity of 1.0.

Pairwise Sample Warping Matrix Generation

Each pair of sample traces is used to generate a sample × sample warping-matrix (W). Each matrix entry is a step function $W_{m,n} = f_{(m,n)}(x)$ such that $f_{m,n}(scan\ in\ sample\ m) =$

$scan\ in\ sample\ n.$ (Appendix 1.4) The notation $W_{i,j}$ represents the step function converting scans from the sample in which peak j was detected into the sample in which peak i was detected.

The warp matrices for this work are generated using dynamic time warping to determine the optimal warp path. Other techniques such as parametric time warping (PTW) have recently been applied to the correction of retention time drift[119] and in general any technique which establishes alignment between the scans in each sample can be used.

Establishing relationships between the supplied peaks

The supplied peak bounds are transformed from the originating sample's elution space into each of the other sample's elution space via the previously determined warping-matrix. Peaks which delineate the same chromatographic regions will share bounds when transformed from their time domain into the other samples time domain.

Each pair of peak bounds is compared to populate a peak × peak match-matrix (P). Pairs which differ by less than the settable cut-off sc.aligned.lim are filled as true.

$$P_{i,j} = \left|bounds_{peak\ i} - W_{i,j}\left(bounds_{peak\ j}\right)\right| < sc.alignerd.lim$$

Splitting the supplied peaks into groups which describe distinct chromatographic regions

Matrix P is represented as a graph structure where matrix indices are the nodes and matrix elements containing a true value are the edges. (Appendix 1.4) The nodes of this graph are split into communities using the walktrap community detection method.[118]

Each of the resulting communities contains one or more detected features. Within each community (i.e., group), all detected features are taken to represent the same analyte.

Determination of consensus peak bounds for each group

All detected peaks within each group contribute to the consensus peak bounds such that each peak represents the same chromatographic region. Grouped peak pairs are transformed into each sample's time domain to create a peak × peak transformed matrix (C).

$$C_{i,j} = W_{i,j}(bounds_{peak\ j})$$

The mean of the 75th percentile of each column is taken as the consensus peak bounds for the jth peak.

Determination of integration region for samples without a detected peak

For samples in which there was no detected peak remaining in a group, the consensus peak bounds are projected into that sample's time domain.

$$C_{i,j} = W_{i,j}(consensus\ bounds_{peak\ j})$$

The median of these transformed bounds are taken as the missing sample's peak bounds.

3.2.3 Output

The output of the algorithm is a list, each entry representing one peak group. A group entry is a matrix with a set of consensus peak bounds for each sample as well as descriptors of the alignment and grouping process. This output can be used for peak integration, filtering, and statistics.

3.2.4 XCMS Implementation

Warpgroup was developed as a standalone algorithm and as such it can be applied to any suitable chromatographic data. For convenience, the Warpgroup package includes integration with XCMS type objects. These functions allow application of the Warpgroup algorithm in the conventional XCMS manner by calling group.warpgroup(). The returned object is an xcmsSet object with peak bounds and groups generated by the Warpgroup algorithm. This resulting xcmsSet does not need any further fillPeaks() and is ready for statistical analysis, either manually or with XCMS's diffreport() function. Further information can be found in Appendix 1.5.

3.2.5 Datasets

Raw Data

We experimentally generated two datasets on which to benchmark Warpgrouping. To evaluate performance under a relevant set of conditions, we chose to generate one dataset with reversed-phase C18 chromatography and the second with amino propyl HILIC.[120] Each dataset contained eleven LC/MS runs of Escherichia coli (E. coli) strain K12, MG1655 metabolic extract. This design allowed us to inspect the standard error of quantitation on both ideal (C18) and non-ideal (HILIC) datasets while also observing the algorithms performance as dataset quality degrades at longer times.

Metabolic extract was generated as described previously.[4] Briefly, two cultures of E. coli were grown, one on natural-abundance glucose and a second on uniformly labeled ^{13}C-glucose as the sole carbon source. E. coli was harvested by pelleting 10 mL of culture at OD_{600} = 1.0. Pellets were extracted using 1 mL of 2:2:1 methanol:acetonitrile:water, and reconstituted in 100 µL of 1:1 acetonitrile:water.

Datasets were generated on the Thermo Q-Exactive Plus mass spectrometer interfaced with an Agilent 1260 capillary liquid chromatography system. Spectra were collected in negative ion mode with the following HESI II source settings: aux. gas, 15; sheath gas, 30; counter gas, 0; capillary temperature, 310 °C; sheath gas temperature, 200 °C; spray voltage, 3.2 kV; needle diameter, 34 ga;

s-lens, 65 V; mass range, 85-1165 Da; resolution 140,000; microscans, 1; max inj. time; 200 ms; AGC target: 3e6.

HILIC was performed as described previously[120] using the Phenomenex Luna NH_2 (1.0 mm x 150 mm x 3 um) column and a flow rate of 50 µL/minute. Solvents were: A, 95% water + 20 mM ammonium hydroxide + 20 mM ammonium acetate; B, 100% acetonitrile. An injection volume of 1 µL was used with a gradient of (minutes, %A): 0, 5; 40, 100; 50; 100; 50.5, 40; 54.5, 15; 55, 5; 65, 5.

Reversed-phase chromatography was performed as described previously [120] using the Agilent Zorbax C18 (0.5 mm x 150 mm x 3 µm) column and a flow rate of 30 µL/minute. Solvents were: A, water + 0.1% (v/v) formic acid; B, acetonitrile + 0.1% (v/v) formic acid. An injection volume of 1 µL was used with a gradient of (minutes, %A): 0, 95; 45, 0; 55; 0; 56, 95; 65, 95.

Preprocessing

The Warpgroup algorithm implements peak subregion detection, consensus/missing peak integration bound determination, and group filtering. These steps come after peak detection has been performed and putative correspondence has been determined. To generate data for comparisons, peak detection for each of the C18 and HILIC datasets was performed by the centWave algorithm as implemented in the XCMS R package.[71,84,121] Parameters were: C18, ppm=2.5, peakwidth=c(8,120), HILIC: ppm=2.5, peakwidth=c(8,120). This set of detected peaks was used as the basis for both the conventional and Warpgroup workflows as described below.

Conventional Workflow

The conventional workflow as referred to here consists of the following listed analysis steps and parameters taken from XCMS Online recommendations for the Q-Exactive Plus.[122] Global retention time correction is performed with the OBI-warp algorithm (profStep=1, center=1). Features are then grouped between samples with the density method (mzwid=0.015). Finally, missing peaks are filled by integrating the range of *m/z* and retention times in the group using fillPeaks(). The resulting filled peak groups contain at minimum one intensity value per sample, but in many instances include multiple intensity values per sample. When performing statistics, the groupval() function applies a filter to select a value which will represent each sample. By default this selects the peak which is closest to the median retention time of the group. All calculations are based on this groupval() output to make results consistent with diffreport() output as used in the conventional workflow by XCMS Online.

Warpgroup Workflow

The Warpgroup workflow consists of the following steps. Global retention time correction is performed with the OBI-warp algorithm (profStep=1, center=1). A rough grouping of features is established by grouping all features within 3 ppm and 25 scans. In our data sets this rough grouping ensured that all peaks which could possibly be the same analyte across samples remained in the same group – this also caused some groups to contain multiple peaks. Here, these rough groups were refined with the Warpgroup algorithm by a call to group.warpgroup (rt.max.drift = 20, ppm.max.drift = 3, rt.aligned.lim = 7). The resulting dataset contained one peak per sample in every group, all of which described the same region. This output xcmsSet was used for all further statistics and assessment of the Warpgroup algorithm.

Selecting Peak Groups for Comparison

The Warpgroup analysis assumes each detected peak represents a legitimate peak region. Upon Warpgroup analysis of these regions, a single group in the conventional workflow often results in multiple warp groups. (Table 2) To make a fair comparison between quantitation of the conventional and Warpgroup methods, a selection of groups had to be made. “Shared” peaks consist of any groups which contain six or more peaks in common between the two workflows. It is worth noting that the “shared” group subset masks the benefits that Warpgroup provides in low abundance peak detection and peak sub region detection.

3.2.6 Performance Evaluation

A major goal of the Warpgroup algorithm is the reduction of variance introduced by data processing. We identified several points in the conventional bioinformatic workflow for processing untargeted metabolomic data where small errors were being introduced. These small errors were compounded in each downstream step, resulting in a significant decrease in dataset quality. We evaluated the impact of Warpgrouping on the two primary errors we noticed: integration bound selection and peak grouping.

Peak Quantitation

Peak quantitation performance of each workflow was assessed by comparing the coefficient of variation (CV) across 11 replicate injections. This metric provides an assessment of the entire workflow. Warpgroup often divides conventional groups into multiple sub-groups, and thus there is not a one–to-one correspondence between warpgroups and conventional groups. To assess similar peak groups in both methods, only groups sharing more than 6 of the 11 centWave detected peaks were included in the coefficient of variation analysis. This ensures a one-to-one correspondence between groups from both methods but obscures the benefit of any additional, true groups recovered by Warpgroup.

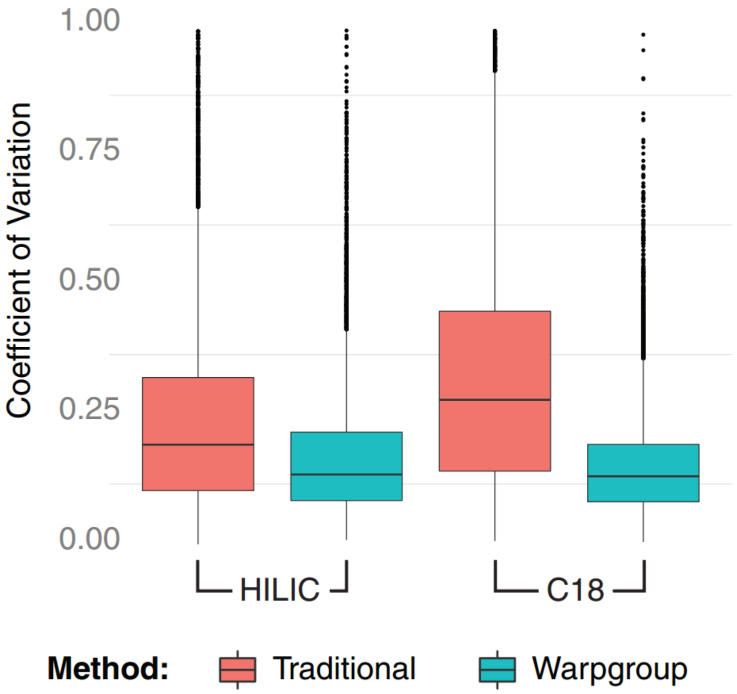


Figure 3.2. Standard error of peak quantitation comparison. The coefficient of variation for all peak groups which shared more than 6 centWave peaks from 11 replicate injections was monitored before (pink) and after (blue) warpgroup. The conventional workflow generates a large number of high variance peak groups for various reasons; upon warpgrouping these are corrected, resulting in a much lower CV for the replicates.

Grouping Quality

The quality of peak grouping was evaluated for both workflows by manually annotating the resulting groups. Automated rating of group quality is complex and remains beyond the ability of current techniques. To generate a metric for the quality of resulting groups we examined 500 groups generated by each workflow, scoring them for uniformity of included peaks. The scoring system employed was: 4, identical integration regions for every peak in the group; 3, some minor variation in the integration regions; 2, major variation in the integration regions; 1, multiple distinct peaks included in the group; 0, a noise group with no discernable correct integration. Scores were summarized for comparison of the conventional and Warpgroup workflows.

Further, the number of additional, distinct chromatographic regions the Warpgroup algorithm detected was quantified. We manually annotated each Warpgroup as redundant, noise, or unique;

Groups that shared more than 75% of their major chromatographic region, or differed in only the tails of the peak were annotated as redundant. In some cases, a peak was split into sub regions but also reported in its entirety. (Figure 1D) We considered this desired behavior and annotated all three as unique groups.

3.3 Results

3.3.1 Standard error of replicate injections

Peak picking in the conventional workflow is performed on each sample independently, causing the integration region for each peak to vary slightly from sample to sample. By considering the peak bounds detected in each sample together, we ensure that the similar integration region is selected for each peak. In addition, the Warpgroup approach reduces errors in grouping which can contribute to inaccurate quantitation and statistics. Analysis of 11 replicate injections with two chromatographies demonstrated the improvement in data processing quality using the Warpgroup method (Figure 2). The mean CV was halved (a decrease of 13% in the HILIC case and 17% in the C18 case). Pairwise comparison of each group before and after Warpgroup revealed that, in most but not all cases, the CV decreased with the application of Warpgroup. (Appendix 1.6)

Quality of resulting groups

Grouping in the conventional workflow is based solely on the assumption that common peaks will cluster in retention time. As seen in Appendix 1.1 and 1.2, this assumption is not strictly true and in many cases (such as shown in Figure 1C) groups will include two or more distinct peaks due to residual drift. We sought to evaluate the quality of peak groups returned by the Warpgroup algorithm as compared to the conventional workflow by rating groups on a scale of 0 to 4. As seen in Figure 3, the Warpgroup algorithm results in a striking increase in peak group quality.

Notably, ratings of 1 correspond to groups which contained multiple, distinct peaks. The correction of these cases represents an increase in dataset coverage, as the additional groups "rescued" by Warpgroup represent newly quantified unique signals.

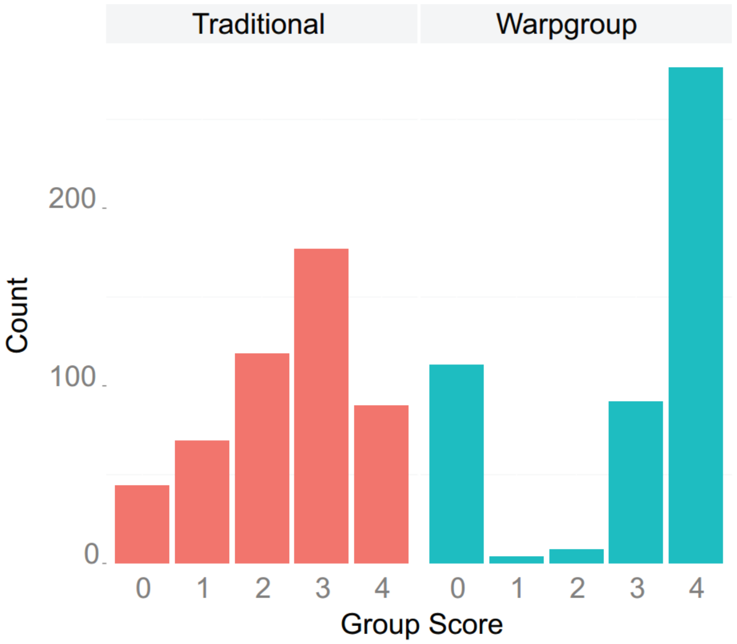


Figure 3.3. Group quality and consistency comparison. The conventional XCMS approach without Warpgroup was compared to the XCMS approach with Warpgroup. Quality of generated groups was assessed. Groups were manually inspected and rated on a scale of 0-4. Zero scores corresponded to noise groups with no discernable correct integration. The remaining scores ranged from 1 (integration regions incorporating different peaks across samples) to 4 (identical integration regions across all samples). Warpgroup (right) showed a major improvement in group quality as compared to the conventional workflow (left). Warpgroup also showed an expected increase in noise groups.

Given that Warpgroup splits peak groups into distinct regions, there is a tradeoff between the number of truly unique chromatographic regions and the number of redundant chromatographic regions that are represented. This tradeoff is controlled by the variable sc.aligned.lim, the only user-settable parameter in our algorithm. This parameter specifies the similarity two sets of peak bounds must have to be called the same peak. A smaller sc.aligned.lim results in more sensitive peak subregion detection but more orphan peaks and a larger number of redundant groups. In practice, orphaned and redundant peaks are easily filtered by removing all peak groups generated by one or a few of the originally detected peak bounds.

We evaluated the redundancy of the Warpgroups and the increase in unique signals detected by manually annotating redundant and noise-only Warpgroups in the HILIC dataset (Table 2). The conventional workflow's 18,341 peak groups resulted in 40,719 peak groups after Warpgrouping. Of these Warpgroups, we manually annotated 33% as redundant and 10% as noise. Considering these redundancies and noise, the Warpgroup approach resulted in 23,209 unique signals. The increase in peak groups by 23% represents distinct chromatographic regions that were added to the dataset and otherwise would have been lost or poorly quantified.

Table 3.1. Coefficient of variation comparisons

Method	Mean CV	90th Percentile CV	Chromatography	Shared peaks
Conventional	33%	57%	C18	15560
Warpgroup	14%	24%		
Conventional	31%	63%	HILIC	7846
Warpgroup	18%	33%		

A final, more restrictive search for rescued groups was performed. Cases such as that shown in Figure 1C were identified by searching for conventional groups in which two distinct peaks were incorrectly combined by conventional grouping due to residual drift. In these cases, peak finding detected both peaks in most samples but the conventional grouping was unable to properly separate them. This search yielded 611 peak groups in the HILIC dataset and 1,246 peak groups in the C18 dataset that were successfully rescued by Warpgroup.

Table 3.1. Group quality comparisons

Method	Groups	Percent redundant	Percent noise	Total unique signals
Conventional	18 341	0%	0%	18 341
Warpgroup	40 719	33%	10%	23 209

3.4 Discussion

The exact use cases of Warpgroup are dependent on the data and the problem at hand. We imagine four distinct goals in the next section and summarize appropriate inputs and expected outputs. Warpgroup operates optimally after inter-sample peak correspondence has been established. Though it is possible to supply ungrouped data to the Warpgroup algorithm, there are several drawbacks to this approach. First, processing time for the dynamic time warping algorithm scales with the square of the input length. Second, if a feature is present in one sample but missing from the others, this dissimilar topography can result in incorrect alignments. Finally, establishing correspondence is a complex challenge for which many more sophisticated solutions have been suggested.[82] These should be used in conjunction with the Warpgroup refinements.

While Warpgroup is not intended to determine peak correspondence, it does make a less restrictive assumption for peak alignment. Current algorithms assume that peak elution order remains monotonic across all masses. Warpgroup assumes only that peaks of indistinguishable mass retain their elution order. This more relaxed assumption allows for rudimentary correspondence to be established in more complex cases such as that shown in Figure 1C and is a major improvement to the XCMS-based workflow.[123]

3.4.1 Use Cases

There are four modes we consider for the application of Warpgroup.

Consensus bound determination.

In cases where the peak was detected in all samples, only the peak integration region with a small number of surrounding scans need be included in the EIC matrix. A large value for sc.aligned.lim can be supplied to avoid splitting of the group into sub regions. When operated in this manner, Warpgroup is relatively fast. The returned bounds are the consensus integration bounds for that group. (Figure 1A)

Peak subregion detection.

This use case is identical to case one except an appropriately small value for sc.aligned.lim is selected, allowing for subregion splitting. This use case is also relatively fast. The returned bounds will be a list of distinct chromatographic regions. (Figure 1D)

Imputation of integration bounds for samples in which no peak was detected.

In this case, both the undetected and detected peak traces must be included in the EIC matrix. Because the feature was not detected in at least one sample, the necessary scan range will be dependent on the expected range in which the undetected peak could fall (i.e., the observed drift). A large value for the argument sc.aligned.lim should be supplied to avoid splitting the detected peaks into sub groups. The peak bounds for each missing sample are then returned. (Figure 1B)

Grouping of peaks which deviate from the global retention time drift.

Warpgroup's "grouping" of peaks is a result of the subregion detection and splitting. In this mode, the EIC region supplied to Warpgroup will envelop multiple peak groups and thus take longer than the above modes. (Figure 1C top) A small value for sc.aligned.lim is supplied if peak subregion detection is desired, or a larger value if the goal is to distinguish two well separated peaks. The result will contain a group for each detected peak group.

3.4.2 Output Considerations

One major advantage of the Warpgroup workflow is the ability to detect noise groups. The warpgroup algorithm includes several peak descriptors for each group after analysis. It is important to note that Warpgroup output is dependent on the input and contains all resulting groups, including noise. Due to the splitting approach, this can result in a large increase in group number - many of which may be redundant. Further, as noise regions have an under-determined warp path, these regions are often split into distinct regions.

Two descriptors generated by the algorithm can be used to detect and filter these cases. These descriptors provide a type of quality measurement of the group. The first descriptor is "n" – the number of peaks originally detected which contribute to this group. This parameter is featured in the conventional workflow, but Warpgroup provides a more refined metric. Rather than n representing all features eluting near each other, here n represents the detected features which describe similar subregions of the chromatogram thus, the metric is much more reliable. In cases of high n, feature detection agreed upon the region of the chromatogram to call a peak. In cases of low n, the peak detection did not robustly detect the region and it is likely noise.

The second descriptor is "warp.consistency". This metric measures how much the bounds shift when projected into each time-domain and back. Chromatograms with a well-defined and conserved topography will generate highly reproducible warp paths. When bounds are projected through these warp paths, any introduced shift will be small and this metric will be low. When bounds are shifted through a poorly defined region, shifts will be greater and this metric will be high. It is recommended to monitor and filter peak groups based on these parameters prior to further analysis.

3.4.3 Challenges

A drawback of the Warpgroup approach is speed. As described in Prince et al., "warping function… [scaling]… is bounded by computational complexity (the more segmented the warp function the more computation required.)"[82] Warpgroup segments every distinct mass trace and, as such, the computational demand is high. The dynamic time warping algorithm employed scales with the length of the input as O(n^2). Thus, as correspondence confidence decreases, the length of the EIC supplied to Warpgroup increases and processing time lengthens rapidly. Conversely, in cases where correspondence confidence is high or the goal is simply consensus peak bound and subregion detection for well-grouped peaks, the algorithm remains very fast. Accordingly, the incorporation of mass and retention time drift correction as well as the establishment of correspondence prior to Warpgroup is recommended.

Prince et al. raise several limitations of current correspondence methods.[82] Though not intended as a correspondence algorithm, Warpgroup does address some of the challenges these methods face. Most importantly, Warpgroup makes more realistic assumptions about the component-specific drift expected in these datasets. Further, as a single reference sample is not used for alignment, Warpgroup remains symmetric and robust. The algorithm is easily implemented in most workflows as it relies on only one required user settable parameter.

3.4.4 Future Directions

Improving the scaling with sample number is an important goal. While the current implementation is sufficient for many published metabolomic studies, the analysis of larger datasets remains a priority. Computation can be minimized by several strategies. For many peak groups, refinement with Warpgroup will be unnecessary, making minor or no modification to the predetermined group. In these cases, Warpgroup can be omitted for all but the most complex groups. The major computational step is the establishment of a warp path between each sample. To reduce computation, the DTW algorithm can be replaced with faster warping algorithms such as PTW if the data allow. Finally, this implementation calculates the full sample x sample warping matrix. However, implementation of a sparse matrix approach could be explored.

Although Warpgroup was presented here in the context of LC/MS data, the input and output of the algorithm are of a general form (multiple time series and regions within those time-series.) As such, the method is generalizable and can find consensus regions within any time-series data. An example of Warpgrouping on echocardiogram data[124,125] can be found in Appendix 1.8.

The Warpgroup algorithm as presented addresses several major drawbacks of the current informatic workflow. Still, current processing techniques leave significant room for improvement. The development of more effective correspondence algorithms is a critical step for the advancement of the field.[82] Additionally, we see promise in leveraging the information embedded in the component-specific drift observed in these datasets. For example, the drift data may be used to cluster ions into composite spectra and to inform further identification.

3.5 Conclusion

In summary, we have found Warpgroup to be an important refinement step for current integration and correspondence methods. With Warpgroup refinement in place, data processing results remain robust across a wide range of experimental conditions. Major advantages have been noted in coverage as well as quantitation, especially in low abundance signals. Further, Warpgroup output includes additional descriptors which can be used to filter noise and unreliable groups from the final datasets. Overall we expect the addition of a Warpgrouping step to the informatic workflow to improve the quality and reliability of untargeted metabolomic analyses.

3.6 Supporting Data

The LC/MS datasets used in benchmarking of the Warpgroup algorithm can be found on our laboratory website at http://pattilab.wustl.edu/software/warpgroup/. Additional information can be found in Appendix 1.

3.7 Acknowledgements

We would like to thank Dr. Robert Pless for his valuable guidance in helping develop this work. G.J.P. received financial support for this work from the National Institutes of Health Grants R01 ES022181 and L30 AG0 038036, as well as the Alfred P. Sloan Foundation, the Camille and Henry Dreyfus Foundation, and the Pew Scholars Program in the Biomedical Sciences.

Chapter 4.

Defining and Detecting Complex Peak Relationships in Mass Spectral Data: The mz.unity Algorithm[*]

Analysis of a single analyte by mass spectrometry can result in the detection of more than one hundred degenerate peaks. These degenerate peaks complicate spectral interpretation and are challenging to annotate. In mass spectrometry-based metabolomics, this degeneracy leads to inflated false discovery rates, datasets containing an order of magnitude more features than analytes, and an inefficient use of resources during data analysis. Although software has been introduced to annotate spectral degeneracy, current approaches are unable to represent several important classes of peak relationships. These include heterodimers and higher complex adducts, distal fragments, relationships between peaks in different polarities, and complex adducts between features and background peaks. Here we outline sources of peak degeneracy in mass spectra that are not annotated by current approaches and introduce a software package called mz.unity to detect these relationships in accurate mass data. Using mz.unity, we find that datasets contain many more complex relationships than we anticipated. Examples include the adduct of glutamate and NAD, fragments of NAD detected in the same or opposite polarities, and the adduct of glutamate and a background peak. Further, the complex relationships we identify show that several assumptions

[*] This work is based on the following publication: "Defining and Detecting Complex Peak Relationships in Mass Spectral Data: The Mz. unity Algorithm". NG Mahieu, JL Spalding, SJ Gelman, GJ Patti, Analytical Chemistry, 2016. NGM developed the conceptualization of complex adduction and developed and evaluated the mz.unity algorithm, and ran all MS experiments. JLS and SJG provided additional data and insight during the writing process.

commonly made when interpreting mass spectral degeneracy do not hold in general. These contributions provide new tools and insight to aid in the annotation of complex spectral relationships and provide a foundation for improved dataset annotation. Mz.unity is an R package and is freely available at https://github.com/nathaniel-mahieu/mz.unity as well as our laboratory Web site http://pattilab.wustl.edu/software/.

4.1 Introduction

In the last chapter we refined the results of peak detection. In this chapter we investigate the nature of the detected peaks. Specifically we attempt to recover relationships between the detected peaks such that we can consolidate degeneracy and remove uninformative signals.

Adduction, fragmentation, and the natural abundance of heavy isotopes can cause a single analyte to generate more than one hundred spectral peaks in mass spectrometry-based datasets. This is referred to as peak degeneracy and it is a major source of the complexity that confounds data interpretation. Spectral peak degeneracy is challenging to annotate and its complexity can exceed the ability of manual annotation in some cases. More recently, automated solutions have been developed to aid in the annotation of mass spectral data.[86,117,126–129] However, current annotation approaches fail to account for the full gamut of possible peak relationships. Further, several common assumptions made in these annotation approaches do not hold in general. Here we present mz-sum, a complete framework for describing complex peak relationships in mass spectrometry data, and mz.unity, an R package that enables the search and exploration of these relationships.

4.1.1 Sources of Degeneracy

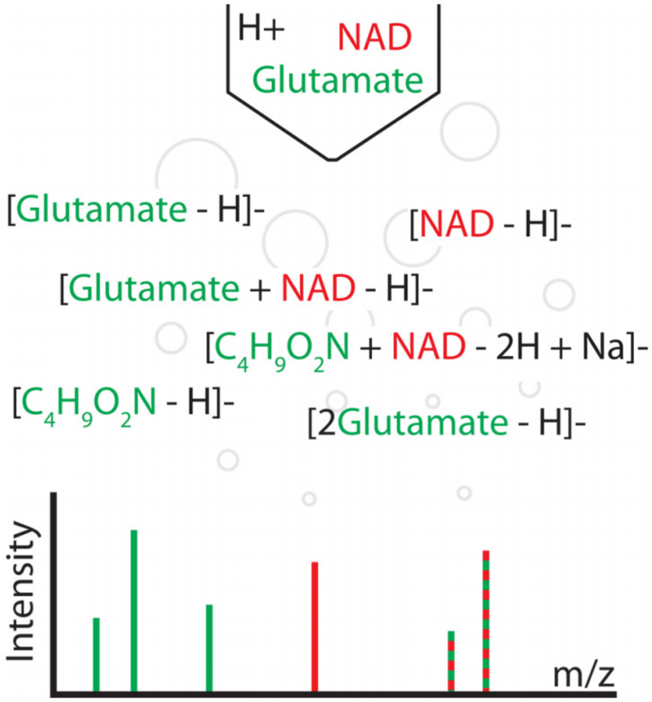


Figure 4.1. An illustration of analyte transformations resulting in degeneracy in detected, spectral peaks. Only two analytes are present, but they contribute to a total of 6 peaks.

The exact conditions under which a mass spectrum is collected have a strong influence on the peaks and types of peak relationships observed. The majority of spectral degeneracy is generated during ionization, which is the process by which analytes are converted from bulk-phase, neutral species to gas phase ions. Electrospray ionization (ESI) is one commonly employed ionization technique. Here we focus on the peak relationships associated with ESI for clarity, but these approaches can be tailored to any ionization technique.

During ESI, analytes undergo various transformations before being detected as mass spectral peaks (Figure 1). The set of possible transformations provide the scope of the peak annotation problem. ESI involves the spray of analyte solution through a charged needle generating gas phase droplets that evaporate until charged gas-phase compounds remain.[50] Two general types of analyte transformations are produced in this process, adduction and fragmentation.

Multiple chemical species that remain non-covalently bound after droplet evaporation are called an adduct. The adduct is a single gas phase ion and will give rise to a single peak, but its formula is the combination of multiple distinct species. In the simplest case, the second chemical species is a proton but others such as sodium and solvent adducts can also be formed. In general, any species present during ionization can adduct with any other species (this includes other analytes).

In contrast, fragmentation is the breakage of bonds prior to MS detection. Often only one of the portions liberated during a single bond cleavage event is detected, but in some cases both are present in the resulting mass spectrum.[130] Bond cleavages can occur at various locations in a molecule and therefore a single structure can generate many fragment species.

An important contrast between the annotation of adducts and fragments is the constraint on possible relationships. For adduction, the space of possible relationships is limited by the species present at the time of ionization. Because a mass spectrum provides an exceptional record of present species, we can reasonably limit our search to those species. In contrast, fragment relationships are limited only to subformula of the parent and are therefore more challenging to annotate.[95,131] In this work, we use mz.unity to putatively annotate two specific subsets of fragments discussed below.

Isotopes are a third source of degeneracy that are independent of the ionization process. Elements such as carbon are found in nature with varying numbers of neutrons (e.g., 12C and 13C). This natural abundance of heavy isotopes causes a single chemical formula to give rise to multiple masses, each corresponding to various numbers of heavy elements. Each of these heavy forms will be detected as a distinct mass peak.

4.1.2 Definitions

Analyte: the chemical species which is of interest in the analysis. Often a metabolite species but can include other molecules such as environmental exposures (e.g., pesticides).

Peak: a mass-to-charge ratio and intensity pair found in a mass spectrum.

Feature: a peak which has a Gaussian like shape (a signal which rises and falls smoothly around a local maximum) in the chromatographic time domain.

Background Peak: a peak which does not have a Gaussian like shape in the chromatographic time domain.

Mer: an adduct between two analytes. This includes homodimers, heterodimers, and higher n-mers.

Distal Fragment: a fragment whose corresponding neutral loss also appears as a peak in the mass spectrum.

Granular-mz: mass and charge pairs supplied by the user to the mz.unity algorithm. These represent specific analyte transformations that combine to make peak relationships.

Complex Relationships: mass spectral peak relationships between three or more detected peaks or relationships between peaks having multiple polarities (i.e., positive ions, negative ions, or neutral masses).

4.1.3 Motivation

Interpretation of a mass spectrum necessitates the annotation of degenerate peak relationships such as isotopes, adducts, and fragments. Critical to the field of metabolomics in particular is the annotation and removal of these degenerate peaks while preserving those that correspond to unique metabolites. Annotation has many benefits for metabolomics: (i) redundant features can be removed, reducing the size of the dataset by more than an order of magnitude; (ii) the concomitant reduction in statistical tests performed allows for a less stringent multiple hypothesis testing correction; (iii) confidence in the validity of a detected peak is increased when degenerate peaks are also detected; (iv) annotated relationships can inform metabolite identification steps; and (v) investigative efforts can be directed to unique analytes. Although we will focus on examples of peak degeneracy in metabolite mass spectra in this work, we point out that annotation is also important to other fields in addition to metabolomics. In proteomics, for example, annotation prior to selection of ions for MS/MS may reduce instrument cycles spent on degenerate peaks and therefore increase proteome coverage.[132–134] In trace impurity analysis, annotation can explain unknown peaks. In approaches that rely on feature counting, such as the evaluation of organic compound diversity on meteorites, annotation is critical to obtain realistic estimates of the total number of unique analytes detected.[42]

Current annotation tools utilize rule tables to describe possible peak relationships.[86,135] A rule table is a list of transformations that neutral analytes may undergo prior to detection. Rules are applied to spectral peaks and a relationship is asserted if two rule-peak pairs predict the same neutral mass. Unfortunately, this approach can only represent a subset of peak relationships. Limitations arise because many spectral peaks do not correspond to a single, underlying neutral mass. Thus, relationships between three or more peaks (as is the case for fragments and multiple-analyte-adducts) cannot be expressed or searched. Current rule tables are also not charge-aware and therefore can only annotate relationships of the same polarity. The limited scope of rule tables precludes the annotation of many putative peak relationships and, therefore, invites a more comprehensive approach to annotation.

To enable comprehensive spectral annotation, we detail two contributions here: mz-sum and mz.unity. Mz-sum is the simple concept that all peak relationships can be described as gain and/or loss of charged formulae. Mz.unity builds on this concept to enumerate all possible peak relationships in a charge-aware manner. Mz.unity is a software package implementing the peak relationship search and tools to plot and explore putative annotations. Together, mz-sum and mz.unity enable the detection of additional complex relationships (e.g., the adduct of glutamate and nicotinamide adenine dinucleotide (NAD), fragments of NAD, and peaks detected in different polarities) that are not annotated by current approaches. The purpose of mz.unity is to return all putative peak relationships within a specified mass error. While mz.unity is a functional tool for exploring spectra and programmatically evaluating relationships within them, we note that it is not an automated annotation solution and assessment of confidence in any specific peak relationship requires information beyond mass and charge. However, this contribution provides the groundwork necessary to enable automated annotation solutions to be developed in the future.

4.2 Experimental Methods

4.2.1 Notation and the Mz-sum Framework

Chemical species having mass "m" and charge "z" are denoted $[m]^{z}$. For clarity, a mass can be referred to by a chemical formula or a compound name. When names are used the neutral, monoisotopic mass is implied. Thus, the following are equivalent: $[146.0459]^{1-}$, $[C_5H_8NO_4]^{1-}$, and $[\text{Glutamate - H}]^{1-}$. Brackets used to denote chemical species can represent either detected mass spectral peaks or any additional formulae. Each set of brackets represents a distinct species.

Conversions may be noted within brackets that describe the nature of the species. In the case of $[\text{Glu-}H_2O\text{-H}]^{1-}$, we are referring to a glutamate species after water loss and deprotonation.

Annotation seeks to find relationships between detected mass and charge species. Relationships are represented by equations of brackets that balance the mass and charge on each side. These equations describe one or more $[m]^z$ peaks in terms of gained and lost mass and charge. From the gained masses and charges, specific transformations can be inferred. For example, the description of a glutamate-acetate adduct can be written as the following equation: $[C_5H_8NO_4 - H]^{1-} + [CH_4CO_2 - H]^{1-} + [H]^{1+} = [C_6H_{12}NO_6 – H]^{1-}$. Mz-sum is the basic assertion that any valid peak relationship will satisfy mass and charge balance and can be represented by such an equation. With this groundwork in place, it is now possible to define a search for all peak relationships.

4.2.2 Description of the Mz.unity Algorithm

Given a list of species with masses and charges $[m]^z$, mz.unity searches for combinations of peaks that satisfy mass and charge balance (a description of the search problem can be found in Appendix 2.1). Additional parameters specify the combinatorial depth with which to search the supplied $[m]^z$ and the acceptable mass error. As follows from the discussion of mz-sum above, this search pattern is general enough to find any type of peak relationship. Below are examples of the general relationship types detected by mz.unity. Notably, each of these lies beyond the scope of previous annotation software. Though compound names are written for clarity, the actual search is performed by using accurate mass.

Complex Adducts: $[\text{Glutamate} – \text{H}]^{1-} + [\text{NAD-H}]^{1-} + [\text{H}]^{1+} = [\text{Glutamate} + \text{NAD - H}]^{1-}$

Distal Fragments: $[\text{Fragment } 123.0453]^{1+} + [\text{Fragment } 540.0536]^{1-} - [\text{H}]^{1+} = [\text{NAD-H}]^{1-}$

Isotopes: $[\text{Glutamate - H}]^{1-} - [^{14}\text{N}]^{0} + [^{15}\text{N}]^{0} = [^{15}\text{N}_1\text{-Glutamate - H}]^{1-}$

The mz.unity search can be tailored to a specific set of relationships by supplying “granular-mz” to the search. These user supplied granular-mz represent undetected species which relate spectral peaks. In the case of adduction, many species present in solution will not be represented in the mass spectrum. This is because spectra have low and high mass cutoffs and only record ionizable species. Such granular-mz in the case of adduction would include small ions such as $[H]^{1+}$, additives such as $[Acetate]^{0}$, and solvents such as $[Acetonitrile]^{0}$. In the adduction and fragmentation examples above, $[H]^{1+}$ was a supplied granular-mz.

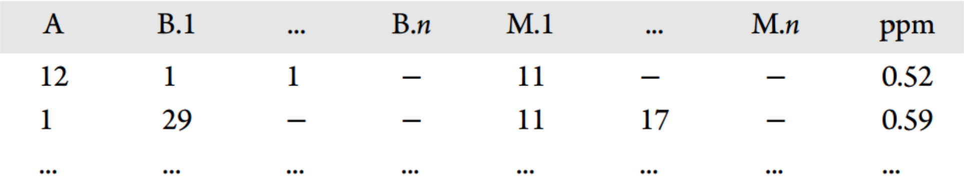
The most general relationship search would include granular-mz corresponding to the atoms C, H, N, O, P, and S, as well as an electron. This set of species would be sufficient to link every peak to every other peak but in almost all cases these relationships would be arbitrary, linking unrelated analytes. By limiting the set of granular-mz, the mz.unity search can be limited to a specific condition or relationship type. In the case of ESI spectra, we seek to relate peaks that are degenerate. This leads to the use of granular-mz that represent transformations occurring during the analysis process.

Many fragments cannot be detected by mz.unity because fragmentation is unique to each analyte and challenging to predict. There are two cases in which mz.unity can detect fragments. When a molecule has two distal charge-sites and fragmentation occurs between them, both portions of the molecule will be detected. This is especially true when spectra from both polarities are included as demonstrated below. In this case, the relationship can be detected by mz.unity, even across polarities (see the fragmentation example above). The second set of detectable fragments is those which occur often under the experimental conditions employed (i.e., common fragments). Common fragments can be supplied as granular-mz and searched like any other relationship.

4.2.3 Output of the Mz.unity Algorithm

The output of an mz.unity search is a matrix (Table 1). Cells reference the supplied [m]z pairs involved in the relationship. Each row represents a relationship. Within each row, columns prefixed with “B.” and “M.” correspond to the peaks and granular-mz that sum to the peak referenced in column “A”. The mass error associated with each relationship is also reported. A convenient visualization of this output is a graph structure (Figure 2). In this representation, nodes are peaks and edges are the detected relationships.

Table 4.1. The output of mz.unity. Row 1 contains the column headers. Cells contain references to supplied mz values. Row 2 represents a dimer relationship, this is the adduction of two glutamate monomers (1) and a proton (11) to result in the dimer (12). Row 3 represents glutamate’s (1) loss of sodium (29) and gain of a proton (11) to produce (29).



A	B.1	...	B.*n*	M.1	...	M.*n*	ppm
12	1	1	–	11	–	–	0.52
1	29	–	–	11	17	–	0.59
...	...	...	...	...	...	...	...

4.2.4 Availability and Implementation

The mz.unity project is written in R and is available at http://github.com/nathaniel-mahieu/mz.unity as well as our laboratory Web site http://pattilab.wustl.edu/software/. Installation instructions, usage examples, data, and analyses presented in this paper can be found in the repository.

4.2.5 Limitations of Mz-sum and Mz.unity

Two limitations of mass- and charge-based annotation are mass measurement error and relationships that have multiple interpretations. Overcoming these limitations requires additional information beyond mass and charge.

Imperfect Mass Information

As described above, the search for appropriately summing masses and charges is a proxy for finding sets of peaks that represent equivalent formulae. Ideally, this search would be performed by using a peak's underlying formula but in practice this is not possible. All empirical mass measurements are made with imperfect accuracy, preventing a one-to-one mapping of mass to formula.[136] Thus, a single mass can represent many possible formulae and this leads to relationships implied by formula mass that do not actually have equivalent formulae. As mass error increases, the number of false positive relationships will also increase. Similarly, the number of combinations of peaks increases rapidly as the number of peaks increases. The combinatorial explosion can quickly overwhelm the specificity offered by accurate masses. This limitation makes annotation of direct infusion data and spectra with over 5,000 peaks challenging.[137,138]

Relationship Ambiguity

Even with perfect formula information, some peak relationships have multiple interpretations that cannot be resolved without additional information. Common neutral losses such as a $[H_2O]^0$ loss could relate either a fragment analyte pair or two distinct analytes. Consider the following two interpretations of a relationship between peaks $[133.0142]^{1-}$ and $[115.0037]^{1-}$.

A. $[\text{Malate - H}]^- - [H_2O]^0 = [\text{Malate - } H_2O]^-$

B. $[\text{Malate - H}]^- - [H_2O]^0 = [\text{Fumarate - H}]^-$

In case A, the smaller peak is a fragment and the two peaks are degenerate, while in case B both peaks are distinct analytes. The two interpretations of this relationship are identical in terms of mass and charge, and additional information is required to determine which is true.

Similarly, fragments and adducts are challenging to discriminate on the basis of mass and charge alone. In both cases, two formulae sum to a third. Consider the relationship $[163.0401]^{1-} + [NH_3]^0 = [180.0666]^{1-}$. This could represent a fragment of tyrosine, in which case the $[180.0666]^{1-}$ peak would be relevant. Alternatively, this could be an ammonium adduct of coumaric acid, in which case the $[163.0401]^{1-}$ peak would be relevant. This ambiguity is true of all distal fragment and mer relationships. The two competing interpretations imply the relevance of different peaks: fragmentation events imply the heavier peak's relevance, while mer relationships imply the relevance of the two lighter peaks.

4.2.6 Dataset Generation

For evaluation of mz.unity, we experimentally generated spectra in positive and negative polarity by using the Q-Exactive Plus mass spectrometer and the HESI-II ion source coupled to an Agilent 1260 capillary flow liquid chromatography system. Spectra were collected with the following settings: aux gas, 15; sheath gas, 30; counter gas, 0; capillary temperature, 310 °C; sheath gas temperature, 200 °C; spray voltage, 3.2 kV; needle diameter, 34 ga; s-lens, 65 V; mass range, 85–1165 Da; resolution 140,000; micro scans, 1; max injection time; 200 ms; automatic gain control target: 3e6. Hydrophilic interaction liquid chromatography (HILIC) was performed as described previously with the Phenomenex Luna NH_2 (1.0mm 150 mm 3 mm) column and a flow rate of 50 μL/min.[122] Spectra were collected in negative and positive ion mode during two different injections. Solvents were: A, 95% water + 20 mM ammonium hydroxide + 20 mM ammonium acetate; B, 100% acetonitrile. An injection volume of 1 μL was used with a linear gradient of (minutes, %A): 0, 5; 40, 100; 50, 100; 50.5, 40; 54.5, 15; 55, 5; 65, 5.

Spectra were taken from a dataset of Escherichia coli (E. coli) strain K12, MG1655 metabolic extract. This design allowed us to inspect real-world data, including co-elution and background ions. Metabolic extract was generated as described previously.[88] Briefly, cultures of E. coli were harvested by pelleting 10 ml of culture at OD_{600} = 1.0. Pellets were extracted by using 1 ml of 2:2:1 methanol:acetonitrile:water, and reconstituted in 100 μL of 1:1 acetonitrile:water.

Liquid chromatography/mass spectrometry (LC/MS)-based techniques generate a series of mass spectra. Peaks that appear in several sequential scans with a Gaussian like profile are termed features (peaks whose intensity rises and falls around a regional maximum over chromatographic time). Chromatographic feature detection was performed on the dataset by using the centWave algorithm.[71] Features eluting from 21 to 22 minutes were used as a test set (FG). This included features from both positive and negative analyses. The set of background peaks (BG) was obtained by retaining all mass spectral peaks appearing in 80% of the scans within this range, regardless of peak shape. Peak lists used for annotation can be found in Appendix 2.2-6 and a spectrum can be found in Appendix 2.7.

Standards of glutamate and NAD were analyzed by direct infusion to validate the detected relationships. A solution of NAD and glutamate both at 50 μg/mL in buffer A was infused at 10 μL/min and spectra were collected at a resolving power of 280,000 in both positive and negative mode.

4.2.7 Dataset Annotation

Mass spectra from an LC/MS analysis of E. coli metabolic extract were searched for relationships by using mz.unity. Several mz.unity searches were performed, each for different relationship types. In brief, the following relationships were searched by altering the supplied granular formulae and search depth: isotopes, charge carriers, neutral gains, cross polarity, common fragments, distal fragments, and mers. Isotopes were detected and omitted from later searches. Charge states were assumed to be to 1 unless carbon isotope support for a higher charge state existed. Searches were performed with a ppm error limit of 2 ppm per observed mass. Exact parameters for each search, including supplied granular formulae and search depth, can be found in Appendix 2.7-8. Putative relationships detected by mz.unity were visualized as graphs and spectral graphs (Figure 3) by using built-in plotting functionality. The graph of relationships was parsed to reveal sets of peaks generated by a single analyte. From the relationship graph, fine isotopic patterns were extracted.

4.3 Results and Discussion

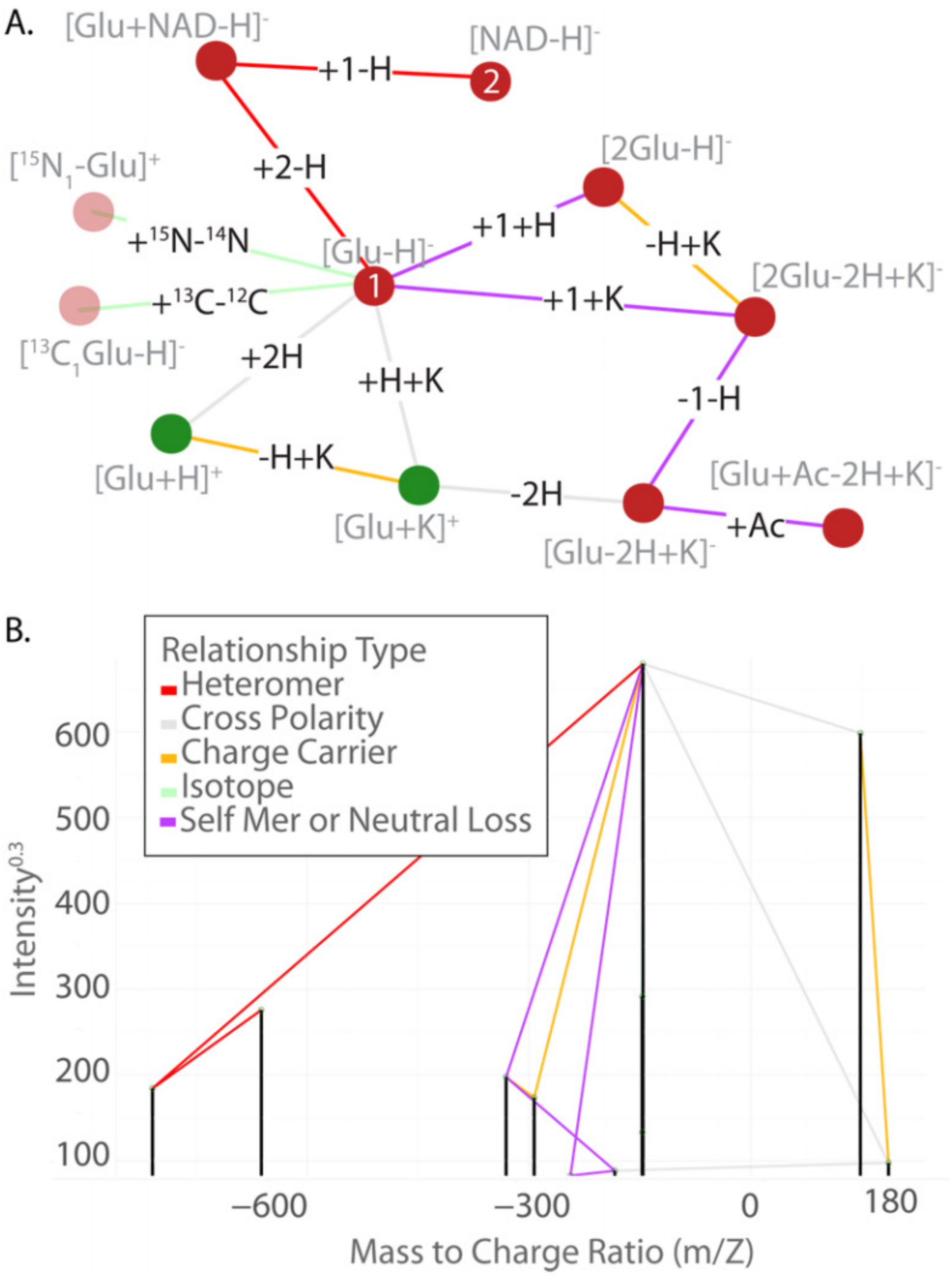


Figure 4.2. A. Output of mz.unity represented as a graph structure. Edges represent peak relationships. The modification relating the peaks is noted as text on each edge. Nodes represent detected *m/z* peaks. The identity of each is noted with grey text by each node. Nodes are colored by polarity: positive (green) and negative (red). Edges are colored by relationship type: charge carrier (yellow), cross-polarity (grey), self-mer (purple), isotopic (green), and heteromer (red). B. The graph structure in Figure 1 superimposed on the mass spectrum of the relevant peaks. Intensity in this graph is scaled as I0.3 so small peaks are visible.

4.3.1 Annotation of a Spectrum Containing Glutamate and NAD

We demonstrate mz.unity, our charge-aware framework for detecting and exploring peak relationships, with a set of peaks observed from the LC/ESI/MS analysis of an E. coli extract. The extract was a complex mixture of small molecule analytes that gave rise to approximately 46,000 total features when analyzed in both positive and negative polarities. The spectrum used to evaluate mz.unity was a composite taken from the time range 21 to 22 minutes consisting of 454 features (peaks with a Gaussian like shape in the chromatographic domain) and 2,212 background peaks. This spectrum was annotated with incremental relationship searches covering various relationship types.

Two groups of peaks were considered, features and background peaks. In LC/MS techniques, all detected analytes of interest appear as features and therefore annotation typically seeks to remove redundancy from the set of peaks that are features. Still, to fully annotate the features, background peaks must be considered as participants in adduct formation. The chromatographic domain was used only to classify mass peaks as features or background peaks, and mz.unity analysis relied only on the mass and charge of the classified peaks.

We consider three general types of relationships in this discussion of results. Simple annotations relate two detected peaks through supplied, granular-mz. Distal fragment and mer relationships relate three or more detected peaks and some number of granular-mz. Finally, background relationships are mers formed between features and the background peaks. All relationships were searched, combining both positive and negative polarities.

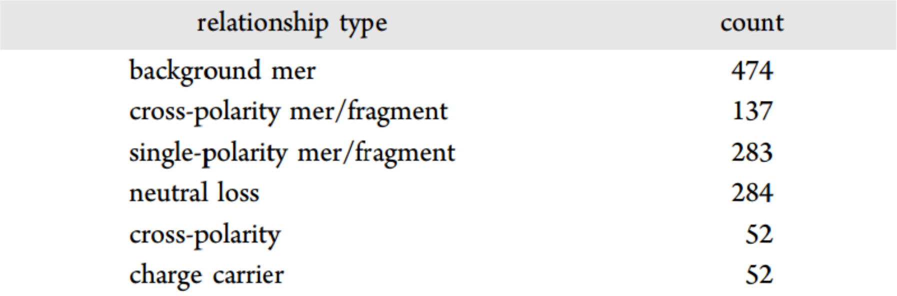
Simple Annotations

Isotope searches detected 64 monoisotopic features having isotopic support. This isotopic support consisted of 101 isotopic features identified in 141 relationships. The remaining 289 features lacked isotopes, indicating low abundance or various types of detector noise. Fine isotopic structure of analytes could be annotated below ~300 Da where resolution permitted.

Charge-aware search, as implemented in mz.unity, allowed for relationships between positive and negative mode ions to be detected simply. These included relationships like $[\text{Glu-2H+K}]^{1-} + [2\text{H}]^{2+} = [\text{Glu + K}]^{1+}$. The charge-aware search also enabled the inclusion of a neutral mass, $[\text{Glutamate}]^{0}$, in the search and easy retrieval of all transformations of this specific mass. In targeted mining approaches, the annotation search can be seeded with relevant analyte neutral masses for simple compound spectra generation. Charge carrier searches between the 64 monoisotopic features with isotopic support detected 104 relationships, 52 of which were cross-polarity relationships. (Figure 3A)

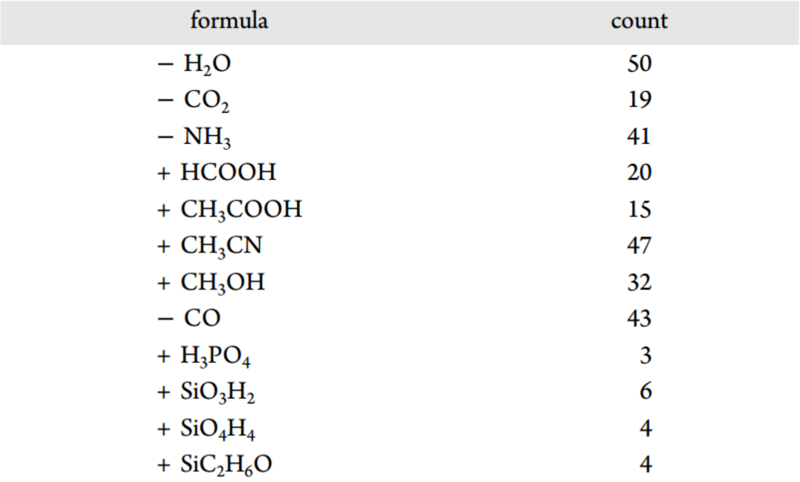
Ambiguous relationships have two interpretations that are indistinguishable by mass and charge alone. These relationships can be drawn between two distinct analytes as well as analyte-fragment or analyte-adduct pairs. We detected 91 ambiguous neutral losses corresponding to loss of $[NH_3]^0$ and $[H_2O]^0$. Manual review of these ambiguous relationships suggested that each of these were true neutral losses and not distinct analytes. Review consisted of evaluating chromatographic peak shape and the elution time of the possible derivative analytes as well as fragmentation spectra of the putative parent. An example confirmation was the relationship $[\text{Glu - H}]^{1-} - [H_2O]^0 = [128.0351]^{1-}$, which was confirmed by using the fragmentation spectrum of a glutamate standard as seen in Appendix 2.10. The automated resolution of ambiguous relationships is one of the challenges that remains to be addressed by an automated annotation solution.

Table 4.1. Breakdown of types of relationships detected



relationship type	count
background mer	474
cross-polarity mer/fragment	137
single-polarity mer/fragment	283
neutral loss	284
cross-polarity	52
charge carrier	52

Table 4.1. Breakdown of common neutral losses detected



formula	count
− H_2O	50
− CO_2	19
− NH_3	41
+ HCOOH	20
+ CH_3COOH	15
+ CH_3CN	47
+ CH_3OH	32
− CO	43
+ H_3PO_4	3
+ SiO_3H_2	6
+ SiO_4H_4	4
+ SiC_2H_6O	4

Unambiguous simple relationships included additional neutral losses and several adducts common to this chromatography such as $[CH_3CN]^0$ and $[SiO_3H_2]^0$. These relationships are unambiguous as the fragments are rare and the related formulae are unlikely to coelute. Within the 454 features, 193 additional neutral relationships were detected. A breakdown of these neutral relationships can be found in Table 2. This annotation of simple relationships reduced the 64 isotopically supported features to 34 feature groups.

Annotation thus far is similar to annotations provided by traditional rule tables. The only extension we have provided at this point is the inclusion of charge-awareness that enabled the linkage of analytes from positive and negative mode as well as neutrals. We extend annotation beyond the traditional annotation scope in the next section.

Mer and Distal Fragment Annotations

A novel set of annotated relationships included mers and distal fragments. Both of these relationship types follow the same pattern relating three or more detected features (i.e., represent complex relationships). This contrasts with approaches based on rule tables that are limited to two detected features. The distinction between mer and distal fragment is in the interpretation, distal fragments imply that the heavier feature is the original analyte while mers imply that the lighter features are the original analyte. In the absence of tools to classify relationships as mers or fragments, we have presented summaries of these searches.

Searching for analyte-analyte complex relationships asserted 420 relationships between 263 analyte peaks (analyte peaks include peaks from features and background). Examining these, examples of both distal fragmentation and analyte-analyte adduction were seen. For example, a distal fragment pair of NAD was found: $[123.0553]^{1+}$ + $[540.0536]^{1-}$ - $[H]^{1+}$ = $[NAD\text{-}H]^{1-}$ and confirmed by MS/MS. The analyte-analyte adduct $[\text{Glutamate} – H]^{1-}$ + $[NAD\text{-}H]^{1-}$ + $[H]^{1+}$ = $[\text{Glutamate} + NAD - H]^{1-}$ was also detected (Figure 3B). The reduction of complex relationships into analyte groups relies on classification of the relationship as mer or distal fragment. Accordingly, we cannot present known analyte groups.

As described above, mass measurement error contributes to false positive peak relationships. Combinatorial searching for peak relationships can rapidly exceed the specificity offered by the mass accuracy of the technique. Ultimately, a solution to probabilistically evaluate each putative relationship is needed for automated annotation. In the absence of this solution, we have manually evaluated a portion of putative relationships to control for the possibility of false positives. Known constituents of the spectrum were checked for incorrect relationships. If the search produced a significant number of false positive relationships, we expected to find these peaks implicated in incorrect relationships. The peaks corresponding to glutamate and NAD had no false positive relationships, indicating that in general these results are valid.

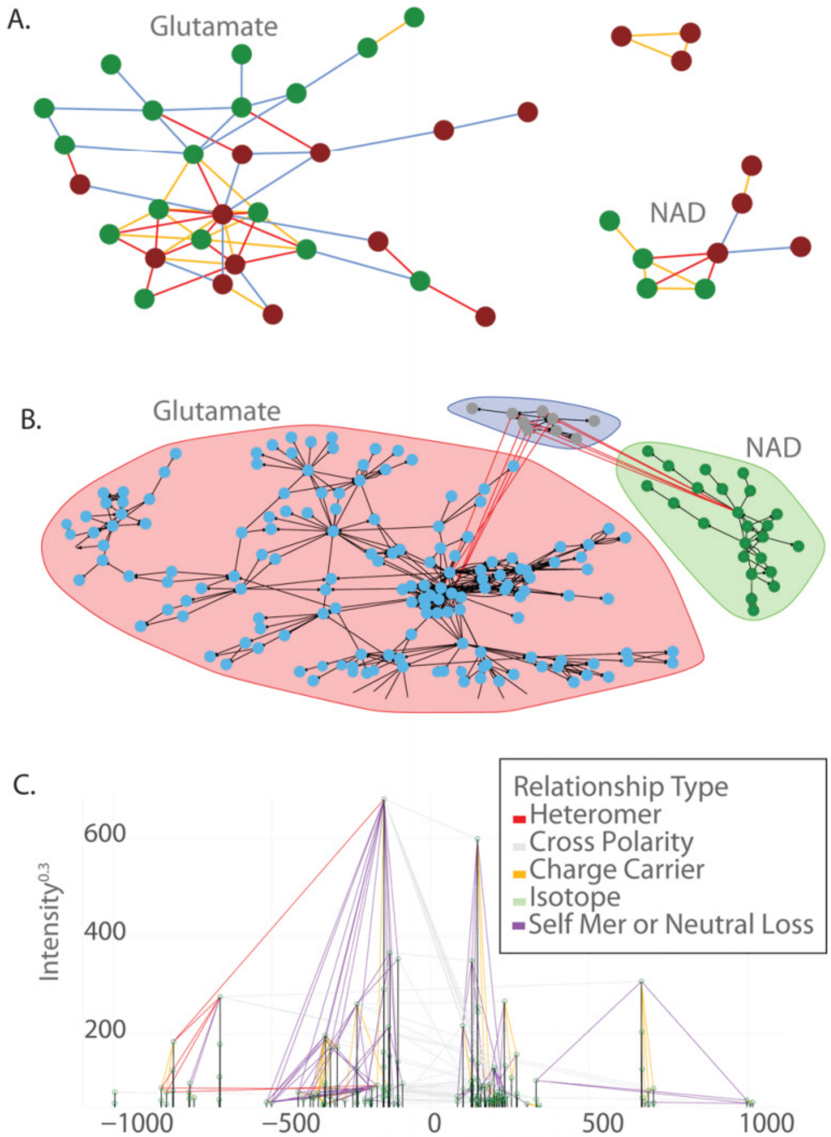


Figure 4.3. Visualizing the subset of peaks derived from analytes glutamate and NAD. A. After annotation of simple relationships. B. After annotation of complex relationships. Peaks derived from the GluNAD heteromer are shown in the blue area. Each node is an *m/z* peak and each edge is a detected relationship. This plot includes isotopes, heteromers, homomers, charge carriers, and neutral losses but omits fragments and background mers. C. The spectral graph of B.

Similarly, mers between analytes and background peaks were searched. Ideally, this search should exclude the possibility of fragment relationships because fragments would appear as features. In practice, some fragment features are detected but not recorded as features and thus enter the background pool. For this reason, we again omit the generation of analyte groups. A search of relationships with background peaks resulted in 474 relationships between 373 peaks. Of those 373 peaks, 129 were background peaks and 244 were features. We show an example of a background mer relationship later. A summary of the detected relationships is shown in Table 3.

Fragment Annotations

To examine the ability of mz.unity to detect fragments, we collected the targeted fragmentation spectrum of a neat NAD standard (Appendix 2.9). This obviates the possibility of mer formation because only the NAD precursor *m/z* was experimentally selected by the quadrupole for fragmentation. Fragment annotation is enabled by mz.unity's charge-aware complex relationship searches. Spectra from a variety of collision energies and both positive and negative polarity were de-isotoped and combined into a composite spectrum consisting of 283 peaks (two of which were the protonated and deprotonated parent peaks). Fragment relationships were detected within this composite spectrum.

Mz.unity detected 404 pairs of fragmentation relationships (Figure 4A-B). These are pairs of detected fragments that correspond to the two liberated portions of the parent ion (Figure 4A). Interestingly, mz.unity's charge-aware annotation is a major advantage for this type of search. In 250 of the detected fragment relationships, one fragment portion was detected in positive mode while the second fragment portion was detected in negative mode (Figure 4C). We also evaluated how intensity impacted the probability of finding both fragment halves. As expected, more intense fragments were more likely to result in a detected pair (Figure 4D). This implies that the number of annotated fragments will be dependent on the sensitivity of the instrument.

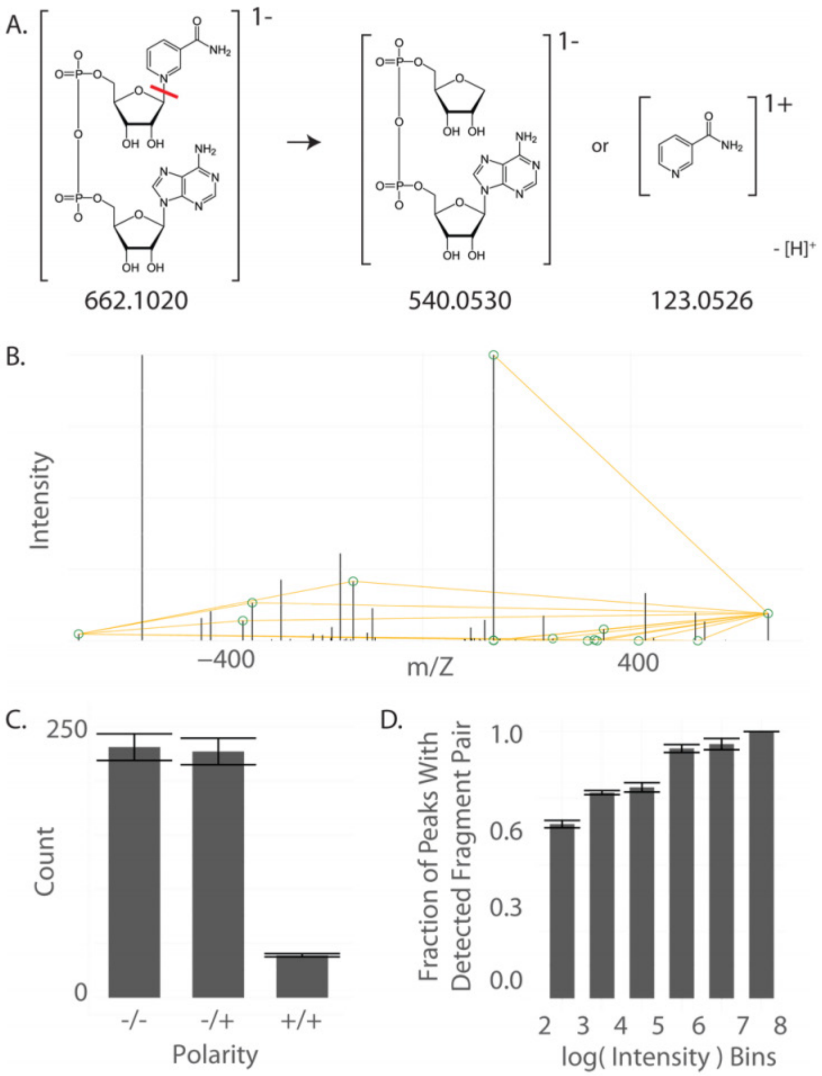


Figure 4.4. Distal fragment searches. A. Schematic of NAD fragmentation resulting in two distal fragments. B. The fragmentation spectrum of NAD and the pairs of distal fragments that sum to the positive and negative molecular ions. C. The number of fragment pairs detected in each polarity. Most fragments were detected by combining positive and negative polarities. D. The portion of peaks with detected distal fragments at varying intensity.

We supplemented the distal fragment search with several common fragments that were unable to be detected on our mass spectrometer due to their low mass. In their neutral form, these were $[H_2O]^0$, $[NH_3]^0$, $[CO_2]^0$, and $[CO]^0$. The possibility of ambiguous relationships was excluded because this was a targeted MS/MS experiment omitting other analyte species. These common neutral losses resulted in the annotation of 86 additional fragmentation relationships.

Of the original 283 peaks in the fragmentation spectrum of NAD, a combination of common neutral loss and distal fragment annotation included 171 peaks (60% of all detected fragments). The remaining fragments were both not in our list of common fragments and lacked a detectable distal second half. Annotation of this type of fragment remains an open challenge to future annotation techniques.

Annotation Summary

This work represents the most thorough annotation of a complex LC/ESI/MS spectrum to date and has important implications for the analysis of metabolomic data. We show that commonly occurring complex spectral relationships lie beyond the scope of previous annotation approaches. Consequently, the amount of spectral degeneracy in mass spectrometry-based datasets has been underestimated. The two analytes in this spectrum provide a somewhat contrasting picture of this degeneracy. Both glutamate and NAD were of relatively high abundance with intensities of $1x10^9$ and $3x10^8$, respectively. Although they were present at similar intensities, glutamate produced 98 peaks and NAD only produced 23. The results presented here underscore the need for thorough analysis of metabolomic datasets to ensure that the myriad of redundant peaks and noise sources do not obscure relevant analytes.

4.3.2 Application to 2-Hydroxyglutarate Metabolism

Mz.unity enables the most complete annotation of metabolomic features to date. Although additional work is required to implement mz.unity as an automated annotation solution on a comprehensive scale, even in its current form mz.unity provides a powerful resource for interpreting LC/MS-based untargeted metabolomic data. In this section, we provide one brief example application to highlight the utility of our mz.unity software package in processing untargeted metabolomic results.

The metabolite 2-hydroxyglutarate (2HG) is known to accumulate in several types of cancer due to gain-of-function mutations in isocitrate dehydrogenase 1 and 2.[139–141] However, the biochemical effects of 2HG accumulation are incompletely understood. We were interested in testing the hypothesis that cancer pathogenesis might be at least partially mediated by the downstream metabolism of 2HG.

We first needed to determine if 2HG is transformed into downstream products in cells. This was accomplished by comprehensively tracking the transformation of uniformly labeled 13C 2HG (U-^{13}C 2HG) into downstream metabolites.[142,143] From the thousands of features we screened by untargeted metabolomics, we found 10 features that were greater than fivefold enriched with 13C carbon compared to natural-abundance samples.

To investigate the identity of these 10 enriched features, we first analyzed the data with the rule-table based annotation package, CAMERA.[86] CAMERA indicated that 6 of the 10 features were adducts of 2HG, leaving 4 of the 10 features to represent biochemical transformations of 2HG. Importantly, this result seemed to support the metabolism of 2HG into downstream products. Therefore, we applied the conventional untargeted metabolomic workflow to identify these features as unique metabolites. When the accurate mass and MS2 data did not match those in databases, we began to explore the exciting possibility that these features might represent novel “unknown” metabolites. Fortunately, before committing to this path, we further analyzed the data with mz.unity to search for complex relationships and fragments. With mz.unity, we discovered that the remaining

4 features were indeed complex adducts and fragments of 2HG (Appendix 2.11). The mz.unity result fundamentally altered the conclusion of our experiment, showing that 2HG is not readily metabolized in the cells we tested. This brief example illustrates how the mz.unity software package can be used in untargeted metabolomic workflows to analyze and refine lists of potentially interesting features.

4.3.3 Observed Failures of Current Annotation Assumptions

In-depth analysis of the aforementioned datasets revealed several assumptions made by current annotation approaches that do not hold in practice. The application of these assumptions therefore prevents the annotation of several relationships in our datasets.

EIC Correlation

Analytes detected by LC/MS techniques elute over sequential spectra with a Gaussian like profile. A common assumption made by current annotation approaches is that related features will have similar peak shapes. This similarity is commonly measured as the Pearson product moment correlation (Pearson's r) between the extracted ion chromatograms (EICs) of the two peaks. 10 Two risks exist: high correlation and assertion of a relationship between unrelated peaks, and low correlation and segregation of related peaks. We find both of these cases to be common in our datasets. We present two cases in which related peaks exhibit low correlation.

Figure 5A shows three salt adducts of glutamate (Glu) that were annotated by mz.unity in our dataset: $[Glu\text{-}H]^{1-}$, $[Glu\text{-}2H{+}Na]^{1-}$, and $[Glu\text{-}2H{+}K]^{1-}$ corresponding to m/z 146.0455, 168.0276, and 184.0015 respectively. The EIC of the deprotonated form exhibits a smooth peak shape typical of our chromatography, but the EICs of both salt adducts exhibit a strikingly different profile. Each initially rise in tandem with the elution of the deprotonated form but quickly plateau. It is clear that each of these salt adducts is related to the [Glu-H]$^{1-}$ peak, yet their correlation is far below useful cutoffs (r of 0.59 and 0.53, respectively).[144]

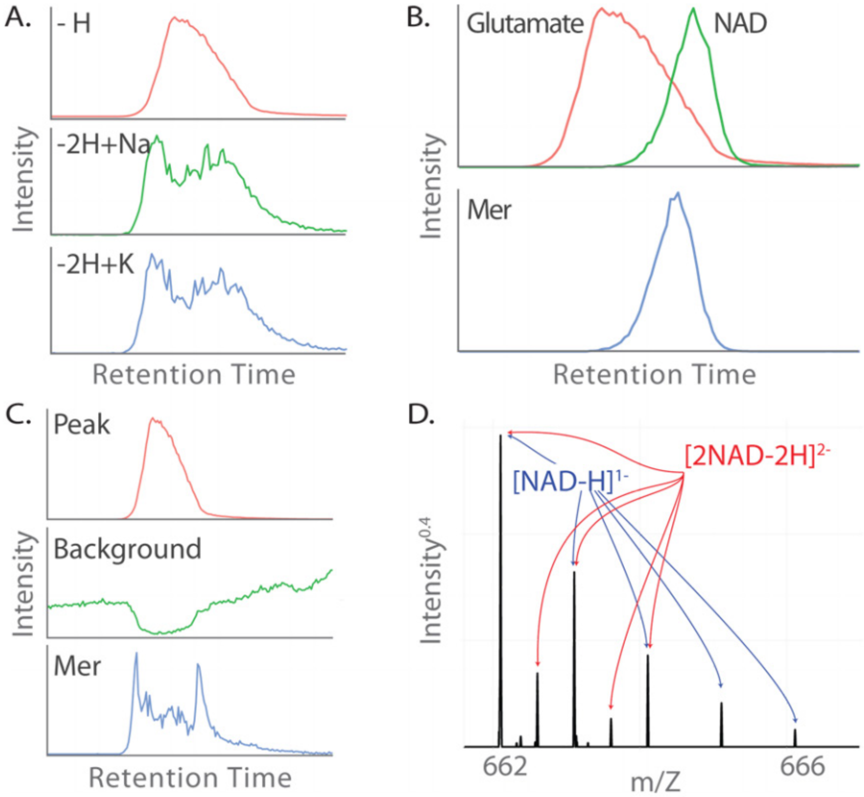


Figure 4.5. Surprising annotation examples. A. The sodiated (middle) and potassiated (bottom) forms of glutamate exhibit different peak shapes than the deprotonated form (top) (Pearson's r of 0.59 and 0.53). B. Overlapping peaks glutamate and NAD (top) adduct to form a glutamate-NAD mer (bottom). C. An artifactual peak (bottom) is formed from the adduction of glutamate (top) and a background peak that lacks a chromatographic peak shape (middle). D. A single *m/z* peak with two charge states and two formulae. The base peak at 662 is comprised of [NAD-H]- and [2NAD-2H]2- as evidenced by the annotated isotopic packet.

A second example of poor EIC correlation between related peaks occurs when two adducting species elute at different times. This is the case in the adduction of glutamate and NAD to form the GluNAD adduct. As can be seen in Figure 5B, glutamate and NAD have a very low correlation (r of 0.09) yet, these two ions are related through the glutamate-NAD mer (GluNAD). The heterodimer GluNAD also does not correlate well with either of its parent species (r of 0.34 and 0.78, respectively). Interestingly, the convolution of the glutamate and NAD EIC traces exhibits strong correlation with that of the mer (r of 0.97), suggesting a possible improvement to this test. Importantly, when EIC correlation is used to group detected features prior to relationship detection, the identification of relationships such as these is precluded.

Background Ions

Peaks lacking a chromatographic peak shape (i.e., background peaks) represent chemical species that can be involved in the ionization process. Current annotation approaches consider only ions displaying a chromatographic peak shape and in doing so they fail to annotate relationships that involve background ions. Background ions have various sources including column bleed, previously eluted compounds washing off the column, solvent impurities, and other contaminants. It is important to emphasize that background ions contribute to detected features with chromatographic peak shapes. As shown in Figure 5C, the adduction of a bonafide feature with a background ion results in a feature with a peak shape. With current annotation approaches, this background-derived artifact would be confused as an additional analyte during later processing. Annotation of this feature is only possible when background peaks are considered during the annotation process.

The adducts in figure A and the background ion in figure C demonstrate characteristics of ion suppression. This general term refers to the reduction in the intensity of a signal due to the presence of other species. It is interesting to note that reduction in the signal of the background ion is not necessarily due to the mechanisms traditionally thought to underlie ion suppression. Rather than competition for charge or alteration of droplet dynamics an additional source of "suppression" could be the scavenging of the monomer signal by other adduct signals. The result being that the same number of species are ionized and detected, but the distribution of signal among masses is altered. This is clearly visible in the background trace in which the signal of the mer necessarily takes signal from the background peak; notably this phenomena may also contribute to non-linearity as peaks reach high intensities. The complexity of this type of ion suppression is further indicated by the adducts in figure A. Adduct formation during droplet shrinkage is a dynamic chemical process involving multiple species. As concentrations change over the course of analyte elution rates and equilibria will also be altered. In the case of the salt adducts above it is possible that glutamate sequestered all available salt or alternatively dimer formation became more favorable than the monomer production. The link between adduct formation and ion suppression is worthy of further study.

Charge-States Assignment

A mass spectral peak is generally taken to represent a single species. Figure 5D demonstrates that this is not true in general but rather, it is possible to detect a single *m/z* peak which corresponds to two distinct formulae. This is common in the case of multiply charged dimers. In the spectrum of NAD found in Figure 5D, two distinct isotopic envelopes can be seen. The major pattern is the result of $[\text{NAD-H}]^{1-}$. The second pattern has spacing of (13C-12C)/2, representing a compound of charge state 2-. This pattern is produced by the ion $[2\text{NAD} - 2\text{H}]^{2-}$. The *m/z* of these two ions is identical, 662.1020, but both species have a different charge state, different formulae, and therefore different mass. The assignment of a single charge state can only explain one of the isotopic envelopes. Full annotation requires the consideration of multiple charge states.

4.3.4 Future Directions

Increases in the mass accuracy and resolving power of mass spectrometers have enabled more thorough analyses of metabolomic datasets. The tools described here, mz-sum and mz.unity, leverage these advances to provide a comprehensive list of possible spectral relationships. Still, several relationship classes require information beyond mass and charge to make definitive annotation assignments. Both ambiguous relationships and fragment/mer relationships have multiple interpretations that cannot be distinguished based on mass and charge alone.

We see four distinct challenges remaining for an automated annotation solution: (i) discrimination between distal fragments and adducts; (ii) discrimination between fragments and distinct analytes; (iii) annotation of rare, non-distal fragments; and (iv) evaluation of confidence in each asserted relationship. Metabolomic datasets offer many rich sources of information to tackle these challenges. Peak intensity, chromatographic profile, mass decomposition, isotope pattern, convolution of adduct constituent's isotopic patterns, and the web of putative relationships are all expected to offer predictive power in the context of these problems. Network based optimization problems and probabilistic assessments have addressed similar problems like fragmentation tree calculation and analyte identification with much success.[126,131,145]

A challenge distinct from annotation is the prediction of underlying neutral masses that give rise to the spectrum. The web of annotated relationships and additional information sources can be combined to assert the masses and identities of the untransformed analytes. These untransformed masses are of interest for metabolite identification and data interpretation in the context of biochemistry. Ultimately, an automated annotation solution will allow faster and more robust metabolomic data analysis while also enabling reliable analyte identification.

4.4 Conclusions

Current approaches fail to annotate a significant fraction of relationships in mass spectrometry-based datasets. We have shown that metabolites such as glutamate produce 100 or more spectral peaks, yet current approaches annotate only a fraction of these. This resulting peak degeneracy is a major challenge to the further analysis of MS data, requiring time intensive manual curation and increasing the number of false positive and misleading hits. Here we have presented mz-sum and mz.unity, which provide a novel framework for assessing these complex mass spectral relationships and enable identification of degenerate peaks that would not be found with current annotation approaches.

Referring to relationships as mz-sums accurately represents any possible analyte transformation, including complex and cross polarity relationships. Consideration of all possible analyte transformations is critical to building thorough and robust dataset annotation tools for several fields, including metabolomics.[42] Here we have expanded upon the relationship approaches based on rule tables by developing the mz.unity R package. While current annotation approaches are based on common and universal transformations, the true set of possible relationships searched for by mz.unity is much broader, encompassing both complex adducts and distal fragments. Mz.unity is both a convenient tool for manual annotation and interpretation of mass spectra as well as a step towards automated annotation of omic scale datasets.

4.5 Acknowledgements

This work was supported by funding from NIH Grants R01 ES022181 (G.J.P.) and R21 CA191097 (G.J.P.) as well as grants from the Alfred P. Sloan Foundation (G.J.P.), the Camille & Henry Dreyfus Foundation (G.J.P.), and the Pew Scholars Program in the Biomedical Sciences (G.J.P.). We thank R. Yost for insightful discussion at the preliminary stages of this project.

Chapter 5.

Credentialed Features: A Platform to Benchmark and Optimize Untargeted Metabolomic Methods*

The aim of untargeted metabolomics is to profile as many metabolites as possible, yet a major challenge is comparing experimental method performance on the basis of metabolome coverage. To date, most published approaches have compared experimental methods by counting the total number of features detected. Due to artifactual interference, however, this number is highly variable and therefore is a poor metric for comparing metabolomic methods. Here we introduce an alternative approach to benchmarking metabolome coverage which relies on mixed Escherichia coli extracts from cells cultured in regular and ^{13}C-enriched media. After mass spectrometry-based metabolomic analysis of these extracts, we "credential" features arising from E. coli metabolites on the basis of isotope spacing and intensity. This credentialing platform enables us to accurately compare the number of nonartifactual features yielded by different experimental approaches. We highlight the value of our platform by reoptimizing a published untargeted metabolomic method for XCMS data processing. Compared to the published parameters, the new XCMS parameters decrease the total number of features by 15% (a reduction in noise features) while increasing the number of true metabolites detected and grouped by 20%. Our credentialing platform relies on easily generated E. coli samples and a simple software algorithm that is freely available on our laboratory Web site

* This work is based on the following publication: "Credentialing features: a platform to benchmark and optimize untargeted metabolomic methods". NG Mahieu, X Huang, YJ Chen, GJ Patti, Analytical chemistry, 2014. NGM developed and evaluated the credentialing algorithm and ran all LC/MS experiments. XJH and YJC provided data and insight during the development process.

(http://pattilab.wustl.edu/software/credential/). We have validated the credentialing platform with reversed-phase and hydrophilic interaction liquid chromatography as well as Agilent, Thermo Scientific, AB SCIEX, and LECO mass spectrometers. Thus, the credentialing platform can readily be applied by any laboratory to optimize their untargeted metabolomic pipeline for metabolite extraction, chromatographic separation, mass spectrometric detection, and bioinformatic processing.

5.1 Introduction

In the last chapter we attempted to understand the context of detected signals in order to remove degenerate signal. This process, though, is unable to discern between analytes derived from the sample under investigation and contaminants. In this chapter we implement the credentialing algorithm, a strong experimental filter which removes contaminants and noise.

The objective of untargeted metabolite profiling is to assay as many endogenous small molecules in a biological sample as possible.[5] Mass spectrometry-based metabolomics represents an established analytical platform that has been widely applied toward this goal and has already yielded many fundamental biological insights.[146–149] Nevertheless, experimental strategies to maximize the number of metabolites profiled are still being developed.[120,150,151] A major challenge in optimizing metabolomic methodologies has been the difficulty in comparing the number of metabolites profiled in each. Given that the size and identity of the complete metabolome is unknown, it is currently not possible to assess metabolome coverage directly. Consequently, the most common metric used to compare different experimental approaches has been the number of features detected in a sample.[87,88,120,152,153]

We show here that a method detecting a maximal number of features does not necessarily provide the greatest metabolome coverage. We present a solution for the evaluation of untargeted metabolomic method performance that enables us to distinguish between two types of features: artifactual features and biologically derived features. Artifactual features are peaks in metabolomic data that arise from contaminants, chemical noise, and bioinformatic noise. In contrast, biologically derived features are peaks that arise from metabolites in the biological sample being analyzed. We refer to the process of distinguishing artifactual features from features of biological origin as "credentialing". In the credentialing workflow (Figure 1), standard samples are prepared from Escherichia coli grown in either natural-abundance media or uniformly ^{13}C (U-^{13}C) enriched media. After performing metabolomic experiments utilizing the methods to be compared, our algorithm finds and credentials features based on expected isotope-intensity ratios. This number of credentialed features represents a more reliable metric of metabolome coverage than total feature count because credentialed features are known to be of biological origin and hence are representative of true metabolites. Upon optimizing our bioinformatic workflow by counting credentialed features, we reduce noise features by 15% and increase properly detected and grouped features by 20%. Further, we select several biological features for tandem mass spectrometry (MS/MS) analysis without any prior knowledge of their identity or physiological significance. It is important to emphasize that the credentialing platform described herein is not intended to identify differences between various biological phenotypes (discovery profiling). Rather, the credentialing platform is designed only to compare the performance of different untargeted metabolomic methods. We provide a step-by-step protocol for performing credentialing with E. coli. While other cell types could potentially be used, E. coli is a simple model system whose optimized results will be applicable to the vast majority of metabolomic optimizations.

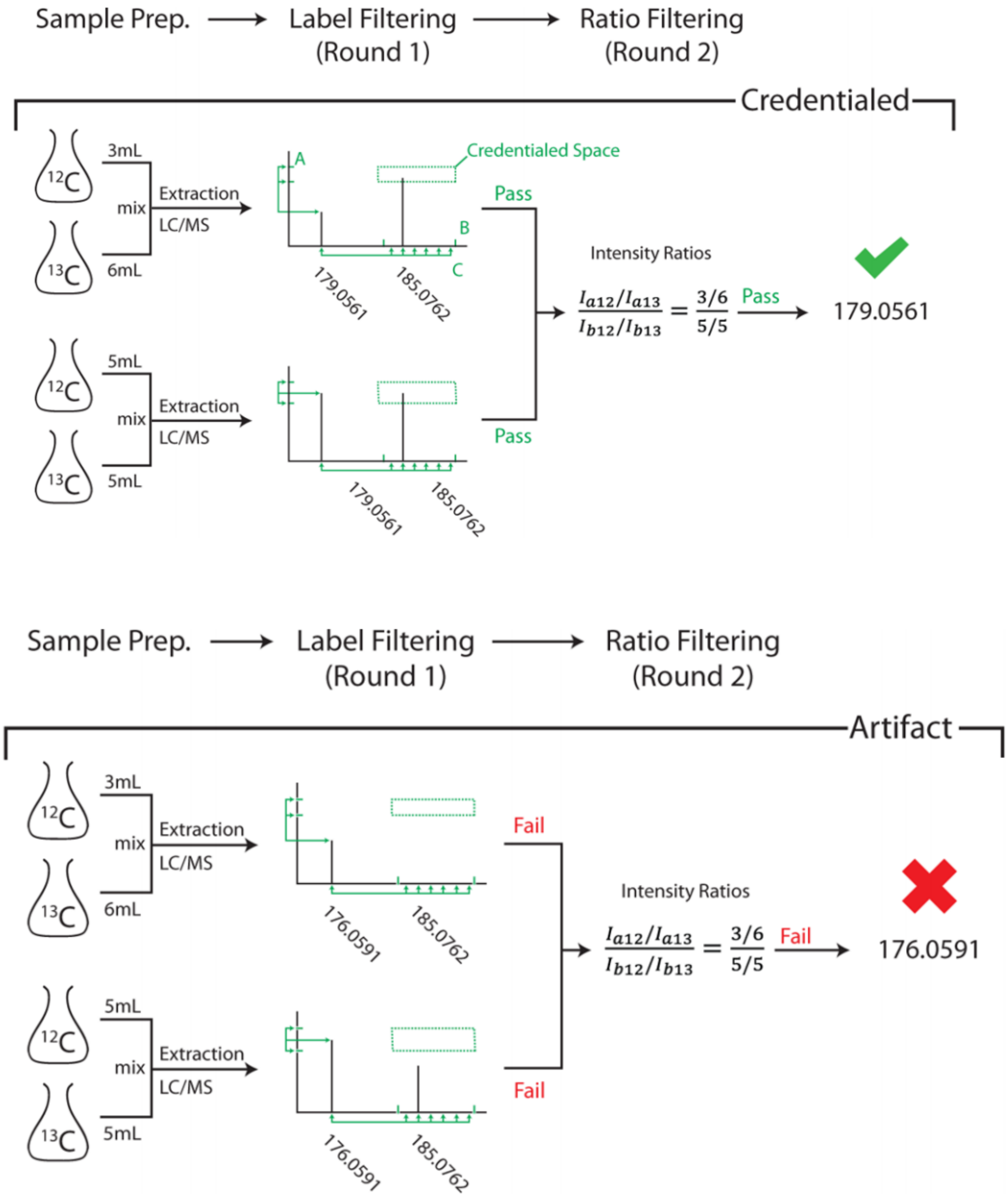


Figure 5.1. Overview of the feature credentialing process. A sample is generated from two cultures of E. coli grown in parallel, one grown on natural-abundance glucose and a second grown on 13C-glucose as the sole carbon source. These two cultures are mixed in distinct ratios prior to harvesting, here 1:1 and 1:2. Extraction and LC/MS analysis is then performed on the standard samples. The resulting data are searched for pairs of coeluting peaks which satisfy the following requirements: (i) the intensities of the peaks must reflect the mixing ratio, (ii) the U-13C peak must predict a feasible number of carbons for the mass in question, and (iii) the exact masses of the peaks must predict an integer number of carbons. These requirements define a "credentialed space" in which the apex of a second peak must be found to qualify as an acceptable isotope. These candidate peaks are then aligned and grouped between the two samples. Each peak pair is compared across samples and a second, stricter intensity check is performed. This requires that the ratios of each sample (Ia12/Ia13 and Ib12/Ib13) are proportional to the mixed ratios of each sample. Peaks that pass these filters are considered credentialed.

5.2 Background

Metabolomic studies are complex, multistep experiments with a large number of parameters to optimize. The choice of sample extraction, chromatography, and ionization method strongly influences which metabolites are detected. Establishing protocols which survey the broadest number of metabolites during untargeted profiling has received detailed attention in recent years.[88,120,154–156] Previous studies have explored a multitude of experimental variations to improve global metabolome coverage that include the addition of ammonium fluoride and ion-pairing reagents to chromatographic mobile phases, separation strategies ranging from reversed-phase to hydrophilic interaction liquid chromatography (HILIC), different mass analyzers such as time-of-flight and the Orbitrap, and various informatic software solutions for subsequent data processing.[84,88,156–158] The extensive list of mutually exclusive experimental possibilities is confounding, particularly to scientists just entering the field of untargeted metabolomics. Yet, to date, comparisons of different methods have been impractical because there is no robust metric for performance evaluation.

Most published comparisons of mass spectrometry-based, untargeted metabolomic methods are evaluated by counting the total number of features detected. A feature is defined as a peak in the metabolomic data set with a unique retention time and mass-to-charge ratio. The number of features detected depends on numerous factors including sample type, metabolite extraction protocols, analyte separation, mass analyzer, and bioinformatic processing. For liquid chromatography/mass spectrometry (LC/MS)-based metabolomics, it is common to detect thousands of features from a biological sample. Importantly, a single metabolite often leads to many features[159] due to: (i) isotopic peaks from naturally occurring 13C, (ii) adduct formation such as hydrogen, ammonium, and sodium adducts, (iii) neutral-loss fragments (loss of a hydroxyl group as water or a carboxylate as carbon dioxide), (iv) other fragmentation (breakage at labile bonds such as esters), (v) multiple-charge states, and (vi) chromatographic effects which result in a single metabolite eluting at more than one retention time.

Informatic solutions have been established to annotate isotopes, adducts, and neutral losses in untargeted metabolomic data sets[86,157,160] Although these approaches are effective, they cannot distinguish signals as endogenous or artifactual. Thus, even after data reduction, a subset of the remaining features are likely the result of contaminants introduced during sample preparation, carryover from previous experiments, chemical noise, or bioinformatic error. These highly variable artifactual signals found in untargeted metabolomic data sets make it challenging to estimate the number of true biologically derived metabolites that are assayed by a particular untargeted LC/MS-based metabolomic experiment. There is therefore a great need to develop a robust metric to the evaluate performance of untargeted metabolomic methods.

5.3 Experimental Section

Our filtering process relies on the generation of standard samples derived from a mixture of E. coli grown on 100% natural-abundance glucose and E. coli grown on 100% U-^{13}C-glucose as the sole carbon source. Two standard samples are required for the filtering process; these are generated by mixing natural-abundance E. coli cultures and U-^{13}C-glucose E. coli cultures at either 5 mL/5 mL or 3 mL/6 mL ratios, respectively. The mixed E. coli samples are then extracted, yielding a standard sample for analysis and optimization.

5.3.1 Materials

U-^{13}C-d-Glucose was purchased from Cambridge Isotope Laboratories Inc. (Andover, MA). E. coli strain K12, MG1655 was purchased from ATCC (Manassas, VA). Lennox LB broth powder, 5× M9 salts, and all LC/MS-grade solvents were purchased from Sigma-Aldrich (St. Louis, MO). Cell culture was performed with ultrapure water provided by a Milli-Q system (Millipore).

5.3.2 Growth of E. coli Standards

Cultures were grown in a rotary shaker at 37 °C and 250 rpm. A preculture of E. coli was grown in LB broth for 16 h. Prior to inoculation, 3 mL of preculture was pelleted and resuspended to OD600 = 0.6 in M9 salts. M9 salts were prepared with the following concentrations in sterile Erlenmeyer flasks: 6.8 g/L $Na_2HPO_4 \cdot 7H_2O$; 3 g/L KH_2PO_4; 1 g/L NH_4Cl; 0.5 g/L NaCl; 240 mg/L $MgSO_4$; 11 mg/L $CaCl_2$. Salts were divided into two 100 mL aliquots, and to each aliquot, 2 mL of 20% glucose was added with a fresh-filtered syringe. The filter was rinsed with 2 mL of ultrapure water to ensure complete transfer of glucose. One aliquot received U-^{13}C-glucose and the second received natural-abundance glucose. The M9 media was then inoculated with 1 mL of the resuspended preculture per 100 mL of media. Cultures were grown to OD600 = 0.6, at which point they were harvested as described below.

5.3.3 Harvesting of E. coli Standards

Upon reaching OD_{600} = 0.6, flasks were removed from the shaker and placed on ice. Appropriate volumes of the ^{12}C and ^{13}C cultures were pipetted together into 15 mL centrifuge tubes, also on ice, generating samples with ratios of 1/1 of 1/2 ^{12}C/^{13}C culture. These mixtures established two distinct ratios of ^{12}C to ^{13}C feature intensities that could then be used in our credentialing algorithm, described below. Cells were pelleted by centrifugation at 2000g for 10 min at 4 °C. The supernatant was removed via pipet, and the cell pellets were snap-frozen in liquid nitrogen. In addition to the mixed ^{12}C and ^{13}C cultures, natural-abundance (^{12}C) cultures were used as controls. We refer to the mixed samples as "labeled" and the natural-abundance extracts alone as "unlabeled."

5.3.4 Metabolite Extraction

The mixed E. coli pellets were extracted as previously described.[120] Briefly, cells were lysed by three freeze–thaw cycles in 2/2/1 methanol/acetonitrile/water along with sonication and vortexing. The soluble portion was then vacuum concentrated and reconstituted in 100 µL of 1/1 acetonitrile/water for LC/MS analysis.

5.3.5 LC/MS Analysis

The data shown herein were obtained from an Agilent 6540 UHD QTOF interfaced with an Agilent 1260 Capillary LC. The column used for separation was a Phenomenex Luna NH2 (150 mm × 1 mm, 3 µm). HILIC solvents were A, 95% water in acetonitrile with 10 mM ammonium acetate/10 mM ammonium hydroxide (pH 9.8), and B, 95% acetonitrile in water. HILIC was performed at 45 µL/min with the following linear gradient (minutes, %B): 0, 100%; 5, 100%; 45, 0%; 50, 0%; 51, 100%; 60, 100%. For all experiments, 5 µL of extract was injected. MS parameters were as follows: gas, 300 °C 9 L/min; nebulizer, 35 psi 1000 V; sheath gas, 350 °C 11 L/min; capillary, 3500 V; fragmentor, 175 V; scan rate, 1 scan/s.

To demonstrate the wide applicability of our credentialing approach to other metabolomic platforms, we also analyzed our samples and subsequently validated correct credentialing with multiple chromatographic and mass spectrometric technologies. In addition to the Agilent QTOF, we credentialed data from the Thermo QE, the AB SCIEX TripleTOF, and the LECO Pegasus GC-HRT. Chromatographic methods we credentialed include reversed-phase LC and HILIC. Effective parameters for credentialing each of these experimental platforms are listed in the Appendix 3.2.

5.3.6 Data Analysis

Analysis was performed with a custom filtering script that utilizes the XCMS[84] and CAMERA[86] R[121] packages as well as the METLIN[161] database. The script is available on our laboratory Web site at http://pattilab.wustl.edu/software/credential/. The algorithm identifies features of biological origin through two rounds of data filtering, as depicted in Figure 1. Prior to filtering, features are detected from the MS raw data with the XCMS findPeaks.centWave algorithm. In the first round of filtering, coeluting peaks within a single sample are assessed for potential isotopologue pairs differing by [(n)1.003355/z] Da in mass, where n is a whole number, z is the ion's charge, and the constant is the mass difference between ^{12}C and ^{13}C. Upper and lower bounds of n for each *m/z* in question were calculated from the distribution of mass per carbon number from the compounds in ECMDB[162] (E. coli Metabolome Database, Appendix 3.1). The ratios of the putative 12C and 13C peak intensities are then evaluated. Each measured ratio that is not within a set percentage of the mixture ratio of the ^{12}C and ^{13}C culture is disqualified. For credentialing, the default value of 400% is effective.

The two filtered samples with distinct mixture ratios of ^{12}C and ^{13}C are then taken together for a final round of filtering. Peaks from each sample are aligned and grouped. Surviving features found in both samples are evaluated such that

$$\frac{1}{e}\frac{x_1}{x_2} \leq \frac{r_1}{r_2} \leq e\frac{x_1}{x_2}$$

where x_i is the $^{12}C/^{13}C$ mixing ratio of the i^{th} sample, r_i is intensity ratio (I_{12C}/I_{13C}) of the ith sample, and e (ratio_tol) sets the acceptable tolerance for the intensity ratio relative to the mixing ratio. This two-round intensity filter allows for features with varying ^{12}C and ^{13}C intensity ratios (due to the kinetic isotope effect or carbon fixation of atmospheric CO_2) to pass the relaxed first round and stricter second round as long as their intensities vary systematically between samples. All passing features are termed credentialed. Credentialed features are output as a summary table that includes all U-^{12}C peaks determined to be of biological origin.

5.4 Results and Discussion

Each step of the untargeted metabolomic workflow can introduce artifactual signals that are not endogenous to the biological sample being analyzed. It is generally not possible to discriminate features of biological origin from artifactual features a priori, and thus, artifactual signals significantly complicate interpretation of untargeted metabolomic results. These artifactual signals can arise from sample contamination during metabolite extraction, carryover from previous experiments, background noise detected by the MS, or misannotation of data during bioinformatic processing. While efforts are made to minimize artifactual signals, it is not possible to completely eliminate them from the features list. We therefore attempted to filter out artifactual signals by using isotopic signatures of cellular metabolism that are easily identified by informatic analysis. We utilized the widely available and extensively characterized E. coli strain K12 to generate isotopically enriched biological extracts. Two cultures were prepared in parallel, one containing ^{12}C (natural-abundance) glucose and the other containing ^{13}C glucose as the sole carbon source in M9 minimal media. The cultures were mixed in defined ratios and processed through the metabolomic workflow together. By searching the resulting features list for pairs of unlabeled and fully labeled isotopologues and comparing their intensities to the values expected from the culture volume ratios, signals of biological origin can be distinguished from artifactual ones. The output of the approach is a list of credentialed features arising from the biological sample of interest. These features reflect the extent to which the methodology employed was able to capture the metabolome.

The power of stable isotope labeling in conjunction with mass spectrometry has long been leveraged to improve quantitative measurements. Mixing labeled and unlabeled samples has proven to be an effective approach to perform quantitation in proteomics,[163–165] and similar approaches have recently been extended to metabolomics.[166] Mashego et al. developed "mass isotopomer ratio analysis of U-13C labeled extracts" (MIRACLE) in which U-^{13}C labeled metabolites obtained from yeast grown in defined culture medium are mixed with unlabeled sample extracts to improve quantitation.[167] More recently, an innovative variation of ^{12}C–^{13}C metabolite mixing was developed in which cells are grown in either 5% or 95% randomly enriched ^{13}C glucose. This experimental strategy, termed isotopic ratio outlier analysis or IROA, leads to a diagnostic isotopic pattern for naturally occurring compounds that can be used for quantitation and metabolite identification during untargeted profiling.[90,168] Here, we introduce another experimental approach which involves mixing ^{12}C and ^{13}C metabolic extracts. We then use the unique isotopic signals that result from the metabolic transformation of the label as a mechanism to identify features of biological origin.

5.4.1 Contrasting the Credentialing and IROA Platforms

It is worth distinguishing IROA from our credentialing approach. Fundamental to the distinction is that mixing a natural-abundance sample with a U-13C labeled sample in a single ratio does not provide a specific enough signature to effectively discriminate features of biological origin from artifactual features. IROA introduces additional specificity to the isotopic pattern by enriching one sample with 5% 13C and a second sample with 95% 13C, instead of using natural-abundance and U-13C samples. In contrast, credentialing introduces additional specificity to the isotopic pattern by mixing different ratios of natural-abundance and U-13C samples. In credentialing, one sample is made by mixing natural-abundance and U-13C cells at a ratio of 1/1 and a second sample is made by mixing natural-abundance and U-13C cells at a ratio of 1/2. There are experimental benefits of each approach that make the platforms better suited for each of their unique experimental applications. IROA has been used to identify and quantitate differences between biological phenotypes during untargeted profiling. Given that the relative ratio of any given peak between biological phenotypes is unknown during untargeted profiling, the credentialing strategy based on defined ratios is incompatible with this type of discovery analysis. The objective of credentialing, on the other hand, is to identify features of biological origin exclusively from standard E. coli samples. While IROA could be used for this purpose in principle, the credentialing platform is not constrained by the aim of discovery analysis and therefore offers several advantages. First, media needed to produce labeled E. coli samples for credentialing is easily synthesized in any laboratory, whereas IROA media can only be obtained commercially. Second, the credentialing platform is better suited to identify low-intensity features of biological origin. In IROA, the signal intensity of any given metabolite is shifted away from the U-12C peak and the U-13C peak as a function of carbon number. For a metabolite with 10 carbons, as an example, 50% of the signal intensity is lost from the U-12C peak or the U-13C peak. This decrease in signal intensity prevents low-abundance E. coli derived metabolites that are detected in unlabeled samples from being detected with IROA. Because the credentialing platform only uses natural-abundance and U-13C samples, it is not subject to this limitation. Indeed, detection of low-abundance metabolites is of particular importance when optimizing metabolomic methods as these compounds are the most challenging to measure, but can be of great biological importance. Finally, because credentialing only uses E. coli samples, the analysis of the resulting isotopic data can exploit the known relationship between mass and carbon number derived from ECMDB (Appendix 3.1).

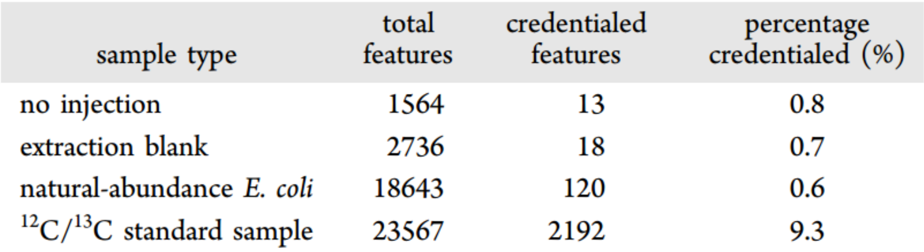
5.4.2 Parameters for Credentialing

To accomplish the filtering of artifactual signals, we created a simple R package. The core function, credential(), has several adjustable parameters allowing various chromatographic and instrumental platforms to be credentialed. These parameters include (i) iso_ppm, the ppm tolerance when searching for 13C isotopes, (ii) iso_rt, the retention-time window in which a peak and its isotope must elute, (iii) mix_tol, the tolerance for the intensity ratio of the 12C and 13C peak, (iv) ratio_tol, the tolerance for the ratio of the intensity ratios between two samples, and (v) mpc_tol, the tolerance for compounds with unusually high or low mass compared to the number of carbons they contain. (Details concerning the calculation of mass per carbon based on the ECMDB can be found in Appendix 3.1.)

We have determined effective parameters for reversed-phase and hydrophilic interaction liquid chromatography as well as for the Agilent QTOF, Thermo QE, AB SCIEX TripleTOF 5600+, and the LECO Pegasus GC-HRT. These parameters have been experimentally validated and are listed in the Appendix 3.2.

Evaluation of the filtering effectiveness was accomplished by comparing the number of credentialed features found in unlabeled and labeled extracts. In addition to the labeled extracts, natural-abundance (unlabeled) extracts were generated as controls. An unlabeled extract should have no credentialed features if it is not mixed with a labeled extract. Therefore, the number of passing features in an unlabeled extract represents the false positive rate. Initial experiments indicated that filtering based on a single mixed-extract sample was not sufficiently selective to remove the majority of artifactual peaks. We found that a two-sample, relative-intensity filter was most effective. As shown in Table 1, this filtering process is selective. The process credentialed only 0.6% of the negative-control features, whereas 9% of the 12C/13C mixture features were credentialed.

Table 5.1. A summary of the results of the credentialing process after being applied to several different data sets. The rows labeled “no injection” and “extraction blanks” represent credentialed peaks due to carryover from previous credentialing runs. Natural-abundance E. coli is a negative control that estimates the false positive rate of the credentialing process.



sample type	total features	credentialed features	percentage credentialed (%)
no injection	1564	13	0.8
extraction blank	2736	18	0.7
natural-abundance *E. coli*	18643	120	0.6
$^{12}C/^{13}C$ standard sample	23567	2192	9.3

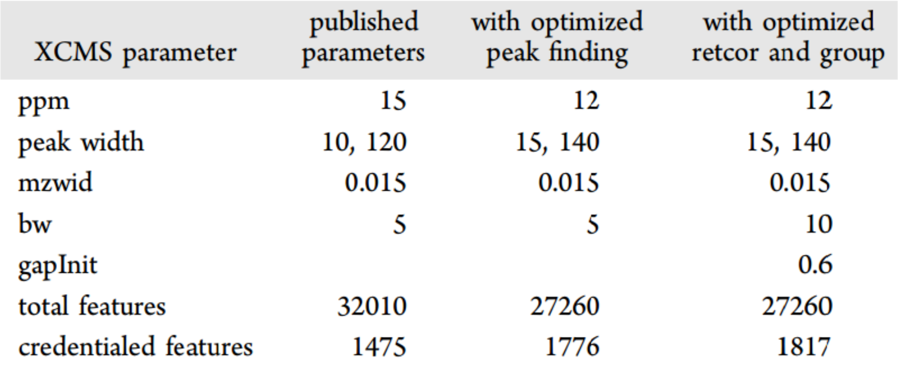
To further validate the filtering process, we examined the natural isotopic peaks that were credentialed in our 12C/13C sample. Consider that in a 12C sample many peaks will contain a natural-abundance M + 1 peak which by definition satisfies the mass requirement to be an isotope. The filtering process credentials some of these natural isotopes along with the monoisotopic peak. These are easily detected and removed by established deisotoping methods, but these peaks allowed us to assess how often an M + 1 is credentialed when the M + 0 is not. If this occurs often, it would indicate that the algorithm is inappropriately disqualifying features. We detected 385 credentialed natural isotopes in our mixture sample. Out of the 385 credentialed, natural isotopes only one did not have a corresponding U-12C in the final credentialed features list. This indicates the filtering approach is performing reliably.

5.4.3 Application: Reoptimization of a Previously Published XCMS Method

With an established method to credential features as biological in origin and exclude various noise sources, we set out to optimize our XCMS-based informatic workflow. XCMS is a widely used informatic package suited for the analysis of untargeted LC/MS data sets. The general XCMS workflow involves peak finding, peak grouping across samples, and retention-time alignment. Settings for each step in this process affect the quality of features returned and therefore the overall performance of the untargeted metabolomic workflow. For example, we found that settings for peak picking that cause the annotation of spurious noise peaks as features lower the quality of peak grouping and retention-time alignment (data not shown). Further, using poor grouping parameters can lead to XCMS splitting a single peak into multiple groups, thereby resulting in erroneous statistics.

To generate data for XCMS optimization, a previously published method was replicated.[120] The same LC/MS system, extraction method, and chromatography protocols were utilized as published and described in the Experimental Section. When processing the data, however, we varied several parameters of the XCMS functions findPeaks.centWave(), group(), and retcor(). As the filtering depends on each of these functions, the final number of credentialed features is representative of the quality of XCMS data processing. Previous approaches to optimizing untargeted metabolomic parameters such as these have relied on counting the total number of features detected. Here, we applied our filtering approach to instead count the number of credentialed features and use this as a benchmark for parameter optimization. Our results show that the published method parameters based on total number of features are suboptimal (Table 2). The published parameters do return a greater number of total features, but the number of features of biological origin accurately detected and grouped is substantially lower with these settings. These data highlight that a larger feature number does not necessarily indicate better metabolome coverage and therefore an improved untargeted metabolomic method.

Table 5.1. Parameters used and the results of each step in the optimization process are shown. Published parameters were taken from a previously published method. The column labeled "with optimized peak finding" shows results for the optimization of findPeaks.centWave().



XCMS parameter	published parameters	with optimized peak finding	with optimized retcor and group
ppm	15	12	12
peak width	10, 120	15, 140	15, 140
mzwid	0.015	0.015	0.015
bw	5	5	10
gapInit			0.6
total features	32010	27260	27260
credentialed features	1475	1776	1817

Reoptimization of XCMS parameters resulted in a substantial improvement. Our XCMS parameters led to an increase of 20% in credentialed features (an increase of 342 features), while reducing the total number of features by 15% (a decrease of 4750 features). Parameters for findPeaks.centWave() were determined to be the most critical to the analysis, while further optimization of group and retcor qualified only an additional 41 peaks. It is notable that, prior to optimizing findPeaks.centWave(), optimization of group() parameters increased the number of credentialed features, partially overcoming the negative impact of artifactual signals.

5.4.4 Characterizing Features in Untargeted Metabolomic Data Sets

To translate metabolomic data into biochemical insight, the features generated in a typical untargeted experiment must first be structurally characterized. The standard workflow for structurally characterizing features requires matching MS/MS data of the features of interest to the

MS/MS data of authentic standards. Identifying features is the most time-demanding step of the untargeted metabolomic workflow and is generally performed in a targeted manner. That is, MS/MS data are only acquired and interpreted for a handful of features determined to be interesting, usually on the basis of statistical thresholds. While this worfklow is often applied to identify tens of metabolites in a metabolomic study, attempting to identify each of the thousands of features detected in a typical sample with this approach is impractical. New technologies to reduce the time required to establish metabolite identifications are an active area of research, but high-throughput methods to structurally characterize metabolites are not widely available. Moreover, many of the MS/MS data are challenging to interpret. When the MS/MS pattern of a feature does not match any of the MS/MS patterns in metabolite databases, it is difficult to determine if the MS/MS data correspond to an unknown metabolite or merely MS/MS data from an artifactual feature.

The feature credentialing approach offers a mechanism to rapidly filter features that should not be pursued for identification, namely, those features that do not correspond to signals of biological origin. When we applied credentialing to E. coli extract, we reduced the number of features that represent candidates for MS/MS from 23 567 to 2192. The resulting subset of credentialed features can be targeted for MS/MS analysis with standard workflows. As an example, we performed targeted MS/MS on 250 compounds in a single experimental run. These data illustrate that MS/MS experiments could be performed on every feature of biological origin over a minimal and feasible number of analytical runs. Select data are presented in Figure 2A–C. The MS/MS data collected on these features were matched to the METLIN metabolite database and resulted in the identification of three metabolites: uracil, ADP (adenosine diphosphate), and UDP-GlcA (uridine diphosphate glucuronic acid). MS1 spectra and chromatograms for these compounds can be found in Appendix 3.3.

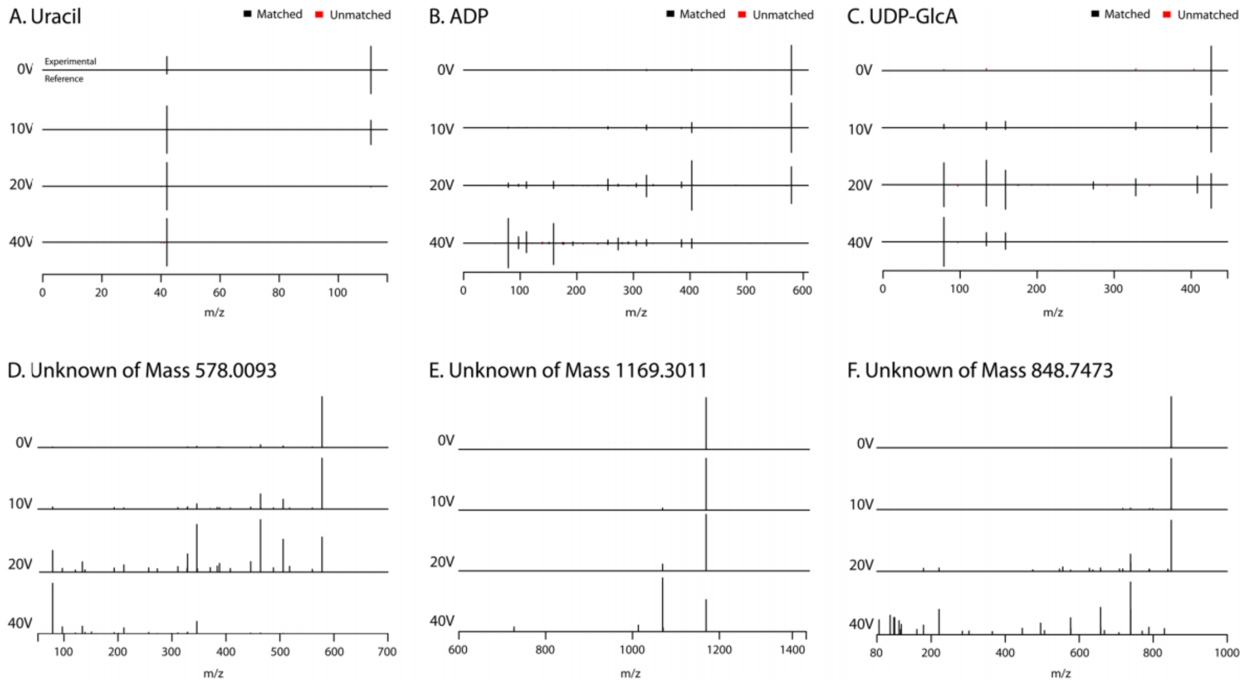


Figure 5.2. MS/MS spectra from six representative credentialed features. MS/MS spectra were collected at four collision energies (0, 10, 20, and 40 V) on six credentialed ions. Three of these ions (A) uracil, (B) ADP, and (C) UDP-GlcA were identified based on accurate mass, carbon number, and METLIN database hits. These identifications were confirmed by comparing the experimental MS/MS spectra to the METLIN MS/MS reference spectra as shown. The upper spectrum of each plot is the experimental data, and the lower spectrum is the METLIN reference data. Unmatched peaks are depicted in red. The second three ions (D) 578.0093, (E) 1169.3011, and (F) 848.7473 were classified as unknowns as they did not match any METLIN database entries as either a fragment or parent mass. The MS/MS spectrum of each ion is displayed as normalized intensity at the same four collision energies.

In addition to generating MS/MS data for metabolites included in databases, it is possible to reliably generate MS/MS data on biological peaks which currently cannot be annotated by metabolomic databases. Because credentialed features have passed our filtering rounds, we know that they are true metabolites of biological origin even if they do not return any database hits. Of the 1827 credentialed features, 392 were not found in METLIN or the METLIN fragment databases. Three such example features are seen in Figure 2D–F. Previously these features may have been discarded as artifacts, but the credentialing platform provides confidence in their authenticity such that they can be reported and referenced in future experiments.

5.5 Conclusion

The feature credentialing strategy presented here is a powerful platform to discriminate biological features from the various noise sources prevalent in untargeted metabolomic data. The process is experimentally straightforward and can be easily implemented in any metabolomic laboratory. Feature credentialing reliably removes artifactual features such as those arising from chemical and informatic noise, thereby resulting in a valuable list of features of biological origin. These credentialed features address many of the drawbacks associated with feature counting in comparing method performance on the basis of metabolome coverage. As such, counting credentialed features can be used in the development and optimization of untargeted metabolomic approaches as demonstrated by the reoptimization of XCMS parameters. Credentialing features is also an effective data reduction strategy for untargeted metabolomic results such that a smaller number of peaks can be targeted for MS/MS analysis. In summary, the feature credentialing platform introduced here represents a step toward defining optimal untargeted metabolomic platforms and provides a standard metric to facilitate collaboration between different metabolomic laboratories.

5.6 Acknowledgements

This work was supported by the National Institutes of Health Grants R01 ES022181 (G.J.P.), L30 AG0 038036 (G.J.P.), and the Alfred P. Sloan Foundation.

Chapter 6.

Contextual Annotation of Metabolomics Data Reduces 25,000 Features to Less than 1,300 Metabolites*

When using liquid chromatography/mass spectrometry (LC/MS) to perform untargeted metabolomics, it is now routine to detect tens of thousands of features from biological samples. Poor understanding of the data, however, has complicated interpretation and masked the number of unique metabolites actually being measured in an experiment. Here we place an upper bound on the number of unique metabolites detected in Escherichia coli samples analyzed with one untargeted metabolomic method. We first group multiple features arising from the same analyte, which we call "degenerate features", using a new contextual annotation approach. Surprisingly, this analysis revealed thousands of unexpected degeneracies that reduced the number of unique analytes to ~2,961. We then applied an orthogonal approach to remove non-biological features from the data by using the ^{13}C-based credentialing technology. This further reduced the number of unique analytes to < 1,000. Accurate mass, retention time, and MS/MS fragmentation data as well as annotations of credentialed features can be freely browsed and downloaded from the creDBle database (http://creDBle.wustl.edu).

* NGM ran all LC/MS experiments, extended mz.unity to full datasets, developed the creDBle database, and catologued credentialed features.

6.1 Introduction

It has become increasingly popular to perform untargeted metabolomics by using liquid chromatography/mass spectrometry (LC/MS). This is at least in part due to the large number of signals or features that are typically detected from most biological samples.[169–171] While it is often assumed that these tens of thousands of detected signals provide "global" coverage of the metabolome, the exact number of metabolites being measured in an experiment has not been rigorously assessed. The major barrier preventing this type of analysis has been the challenge of identifying metabolites[40]. To date, the overwhelming majority of the detected signals in any one untargeted metabolomics experiment have not been named. Even comprehensive efforts to identify as many metabolites as possible in a data set by using the most advanced informatic resources currently available have resulted in relatively small percentages of the total number of signals being identified.[92,172,173] Thus, the basic question of how many unique metabolites are being profiled in an untargeted metabolomics experiment has remained outstanding.

It is important to note that uncertainties related to experimental coverage have not prevented the widespread application of the untargeted metabolomics technology. Improvements in instrumentation and software have made performing untargeted metabolomics with LC/MS relatively routine.[1] Accordingly, the number of research cores offering LC/MS untargeted metabolomics services has increased dramatically over the last decade.[174] The conventional workflows used by most research facilities, however, essentially sidestep the issue of experimental coverage.[175] Their experimental output is a long list of signals or features, without thorough annotation. The data sets are either mined in a targeted fashion for specific metabolites with known retention times and fragmentation patterns, or only the small subset of signals that have a statistically significant difference between sample classes are further investigated.[98] For many of these signals altered between sample classes, further investigation does not lead to identification because their accurate mass and fragmentation patterns do not match the accurate mass and fragmentation patterns of any known reference standard in metabolomic databases.[113] Although it is common to refer to these unmatched signals as "unknowns", rarely is such a designation justified. Signals associated with contaminants, artifacts, and many adducts also do not return matches from metabolomic databases. These possibilities and others must be ruled out before gaining confidence that a signal is a bonafide, unique metabolite with an unknown structure.

The number of signals or features in an LC/MS-based metabolomics data set that result from the combination of contaminants, artifacts, and degeneracies (such as complex adduct formation) has not been comprehensively evaluated. We speculated that these may represent an underestimated portion of signals in untargeted metabolomics data. The goal of the current study was to quantitate contaminants, artifacts, and degeneracies in order to get an upper estimate of the number of unique metabolites detected in a representative LC/MS-based metabolomics experiment. For the purposes of this work, contaminant refers to a detected signal that does not originate from the biological sample being measured (e.g., solvent impurities and plastic leechables). Artifacts refer to features detected due to informatic error. As an example, artifacts can be caused by baseline fluctuations and poorly resolved components.[176,177] Finally, degeneracy refers to multiple signals arising from a single analyte. There are many causes of degeneracy including: fragmentation, analyte adduction with various charge carriers (e.g., a proton, sodium, potassium, etc.), and the detection of naturally occurring isotopes (e.g., ^{13}C, ^{15}N, etc.)[3] A final, largely under-annotated source of degeneracy is the adduction of an analyte with other species present, including other analytes or the chemical background.

Although some degenerate relationships are well known and commonly annotated, the prevalence of many degenerate relationships has not been previously estimated.[86] Here we introduce and apply an approach that recovers relationships implied by the experimental data, rather than relying on a hypothetical predetermined list. The approach allows for more comprehensive annotation, especially in the case of under-annotated adducts that may be specific to a single laboratory or experiment. To the best of our knowledge, the algorithms introduced below are the first to assess these degeneracies.

In this work, we have focused on Escherichia coli cells that were extracted and analyzed with a representative untargeted metabolomics method. In positive-ion mode, we detected 25,230 high-quality metabolomic signals or features. Strikingly, we found that more than 90% of these detected signals were due to contaminants, artifacts, and degeneracy. These results have important implications for the experimental coverage of untargeted metabolomics, which influence the design and interpretation of discovery profiling experiments. Our data indicate that caution should be employed when evaluating unidentified features from metabolomic data sets at the systems level.

6.2 Results and Discussion

6.2.1 Generating a representative untargeted metabolomic data set

In untargeted metabolomics, signals are often referred to as features, a convention we will follow here. A feature is a detected ion with a peak shape, unique m/z, and retention time. To estimate the number of unique analytes detected in a representative untargeted metabolomics data set, we set out to annotate three types of features: (i) degenerate features, (ii) contaminant features, and (iii) artifactual features. We annotated degenerate features by using mz.unity and a new contextual approach to find degeneracies implied by the data. We annotated contaminant features and artifactual features by using the credentialing approach (Mahieu et al., 2014). A requirement of the credentialing approach is uniform ^{13}C-labeling. Given that there are convenient and well-established methods to culture E. coli on a uniformly labeled carbon source, we chose to focus our work on E. coli.

Metabolites from E. coli cells were extracted and analyzed with an LC/MS-based untargeted metabolomics platform, as detailed in Methods. In brief, metabolite extraction was achieved by using a combination of methanol, acetonitrile, and water. Extracted metabolites were separated with reversed-phase chromatography prior to being analyzed in positive polarity by a Q Exactive Plus mass spectrometer. These experimental methods (or variations thereof) are commonly applied in untargeted metabolomics.[178–180,120] To process the resulting LC/MS data, we employed a custom informatic workflow (Figure 1). The workflow used an iterative, two-phase peak detection process. An in-house model-based feature detection algorithm was run on each of five individual replicates. Many of the resulting features are inconsistent between replicates due to subtle differences in the chromatograms from each file. It is common for some peaks to go undetected, or some peaks to be integrated differently between runs.[2] These errors make further analysis challenging because a one-to-one feature grouping cannot be specified between replicates, and the established groups contain artificial variation in feature areas. To refine the features detected in the five replicates, we utilized the Warpgroup algorithm.[2] Warpgroup considers all files in concert to identify "consensus features", a set of feature integrations supported by all replicates. The result is a near one-to-one matching of features between samples (Figure 2A-B) and decreased variation introduced by informatic processing (Figure 2C-D). The Warpgroup refined feature detection is highly sensitive, allowing the recovery of features that, when processed in isolation, would be challenging to detect (Figure 2E). Here, we retained only features with a signal-to-noise ratio >5 and a coefficient of variation <0.5 after Warpgrouping. This resulted in 25,230 high-quality features in our representative data set.

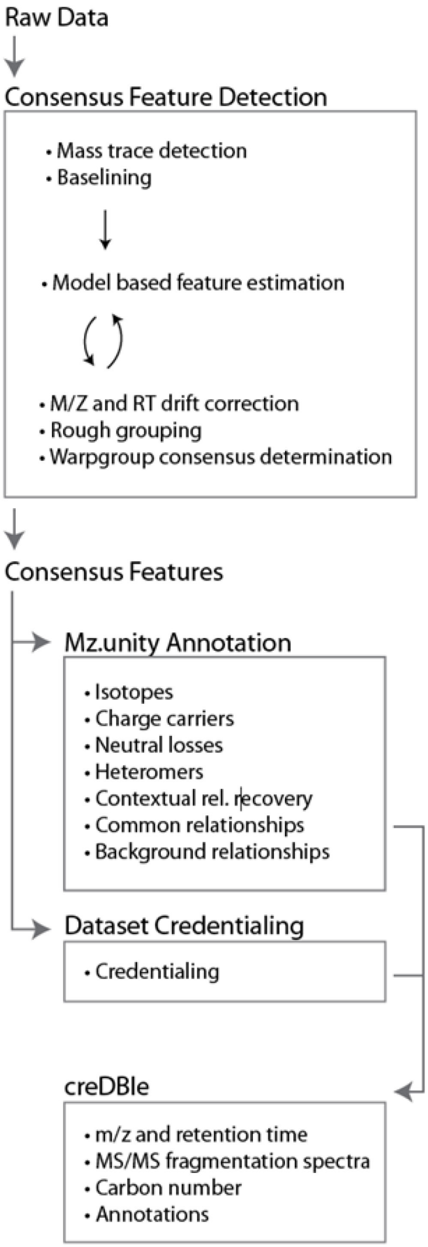


Figure 6.1. Our informatic workflow. Raw data were processed with in-house algorithms to first identify high-quality, consensus features (i.e., recurring features between replicates) and discriminate against processing artifacts. This consensus data set was further characterized by mz.unity (to estimate signal degeneracy) and credentialing (to estimate contaminants and artifacts). The resulting annotated data set was catalogued in the creDBle database.

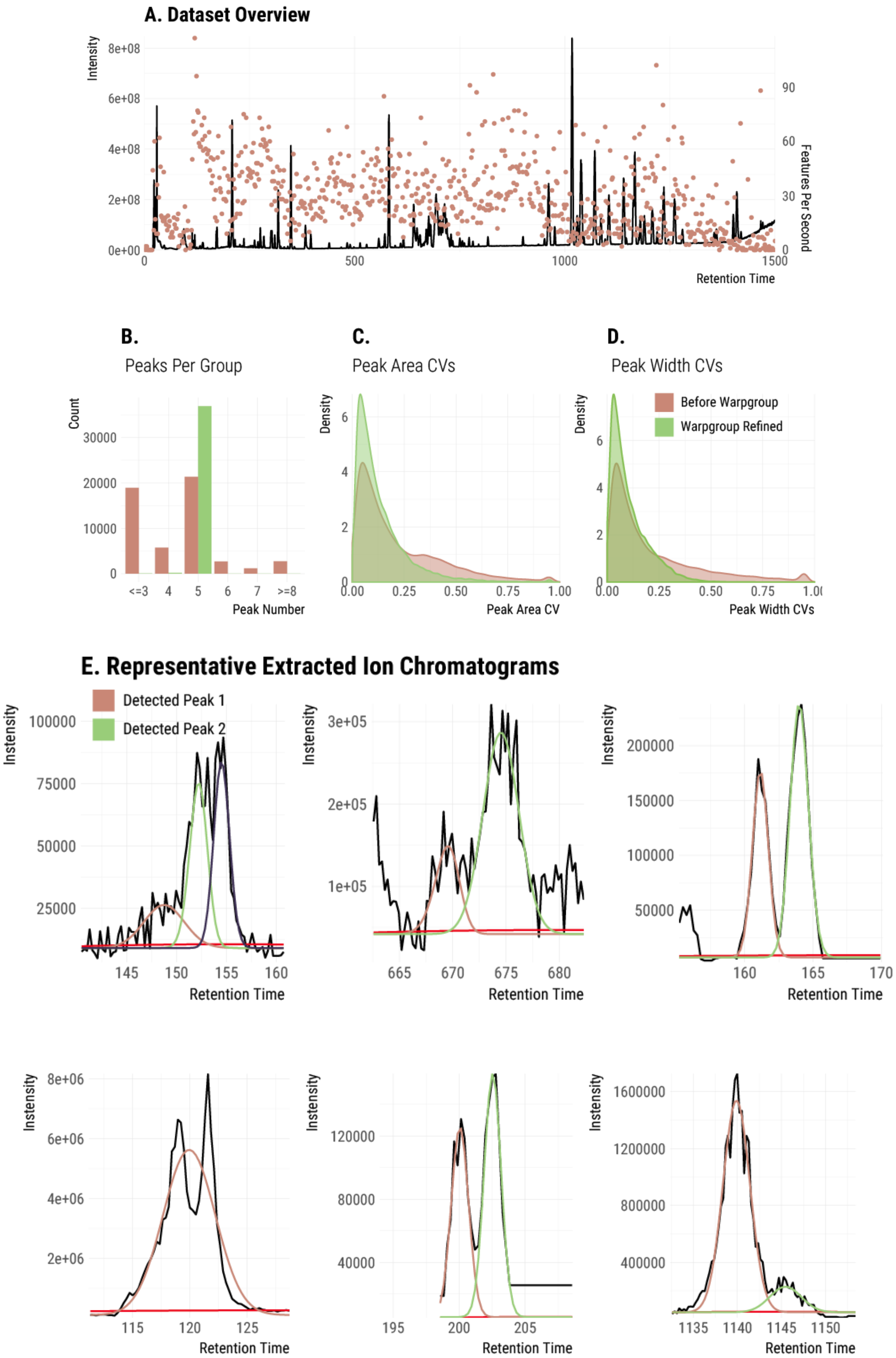


Figure 6.2. An overview of the consensus data set. (A) The base peak chromatogram of a representative run. The number of features detected during each second is overlaid. (B) The number of features detected in each group before (pink) and after (green) Warpgroup. Inconsistencies are resolved by Warpgroup. (C) The within group CVs of peak areas is decreased by Warpgroup. (D) The within group CVs of peak width are decreased by Warpgroup. (E) Several representative features detected by the informatic workflow. The estimated baseline is plotted in red.

We note that there is no universally accepted experimental platform for untargeted metabolomics at this time. The extraction techniques, chromatography, mass spectrometers, and peak detection algorithms used vary between laboratories and are often multiplexed.[181,182] However, it is routine to detect tens of thousands of signals from a biological sample in most LC/MS experiments.[77,183] Our detection of 25,230 consensus features from five replicates resulted in a data set with complexity that is typical of an untargeted metabolomics experiment.

6.2.2 Simple annotations

As a first step to place an upper bound on the number of unique metabolites detected in our experiment, we performed a background subtraction. Specifically, we filtered features that were not at least two-fold higher than the signal detected in extraction blanks. These features represent contaminants or artifacts that are introduced during the sample extraction or data-processing steps. This reduced our list of 25,230 features to 12,797 (Figure 3A).

Next, we set out to annotate degenerate features (i.e., those features arising from the same analyte). We started our analysis by identifying simple relationships that are already commonly annotated in untargeted metabolomics.[86,126,184,185] This included degeneracy due to carbon and other isotopes as well as common adducts and neutral losses. Annotations were made by using mz.unity, and degenerate features were grouped together.[3] Because features within the same group arise from the same analyte, the number of "feature groups" provides a much better estimate of the maximum number of unique analytes detected in an experiment than the number of total features (Table 1 and Figure 3 B-C). In our subsequent descriptions, we will therefore transition from counting features to counting feature groups. A feature for which no degeneracy has been identified constitutes its own feature group, which we refer to as a singlet. Figure 3B shows the progressive decrease in the number of feature groups as isotopes, common charge carriers, and common neutral losses are annotated.

Table 6.1. A breakdown of the analyte number observed after each annotation step.

	Groups with more than one feature		Singlets	
Stage	**All Features**	**Credentialed Features**	**All Features**	**Credentialed Features**
Blank Subtracted	0	0	12797	2462
Isotopes	3986	1066	5071	1326
Charge Carriers	3620	1137	4384	992
Neutral Losses	3640	1174	3678	790
Multimers	3400	1117	3381	712
Commons n>200	2809	1063	2472	495
Commons n>50	2149	864	1620	353
Background	1673	659	1288	233

When isotopes, common charge carriers, and neutral losses are annotated, the number of feature groups decreases from 12,797 to 7,318. We note that currently employed annotation approaches end here with the identification of simple relationships (see vertical line in Figure 3B). These results might suggest that there are as many as 7,318 unique analytes detected in the sample, but two observations suggested that much degeneracy still remained unannotated in our E. coli data set. First, about 50% of our feature groups still contain only a single feature (i.e., singlets with no detected relationships). Although in some cases singlets result from low-abundance analytes with no natural isotopes detected above noise level, the prevalence of singlets suggested that additional relationships remained unannotated. Second, we also know that the set of relationships annotated thus far are only a small subset of the possible degeneracies. A recent targeted study of glutamate demonstrated that many additional, complex sources of degeneracy can exist in LC/MS-based metabolomics.[3] Glutamate was found to produce over 100 spectral peaks and exhibited complex adduct formation. Our objective was to comprehensively characterize these additional sources of degeneracy within our E. coli data set.

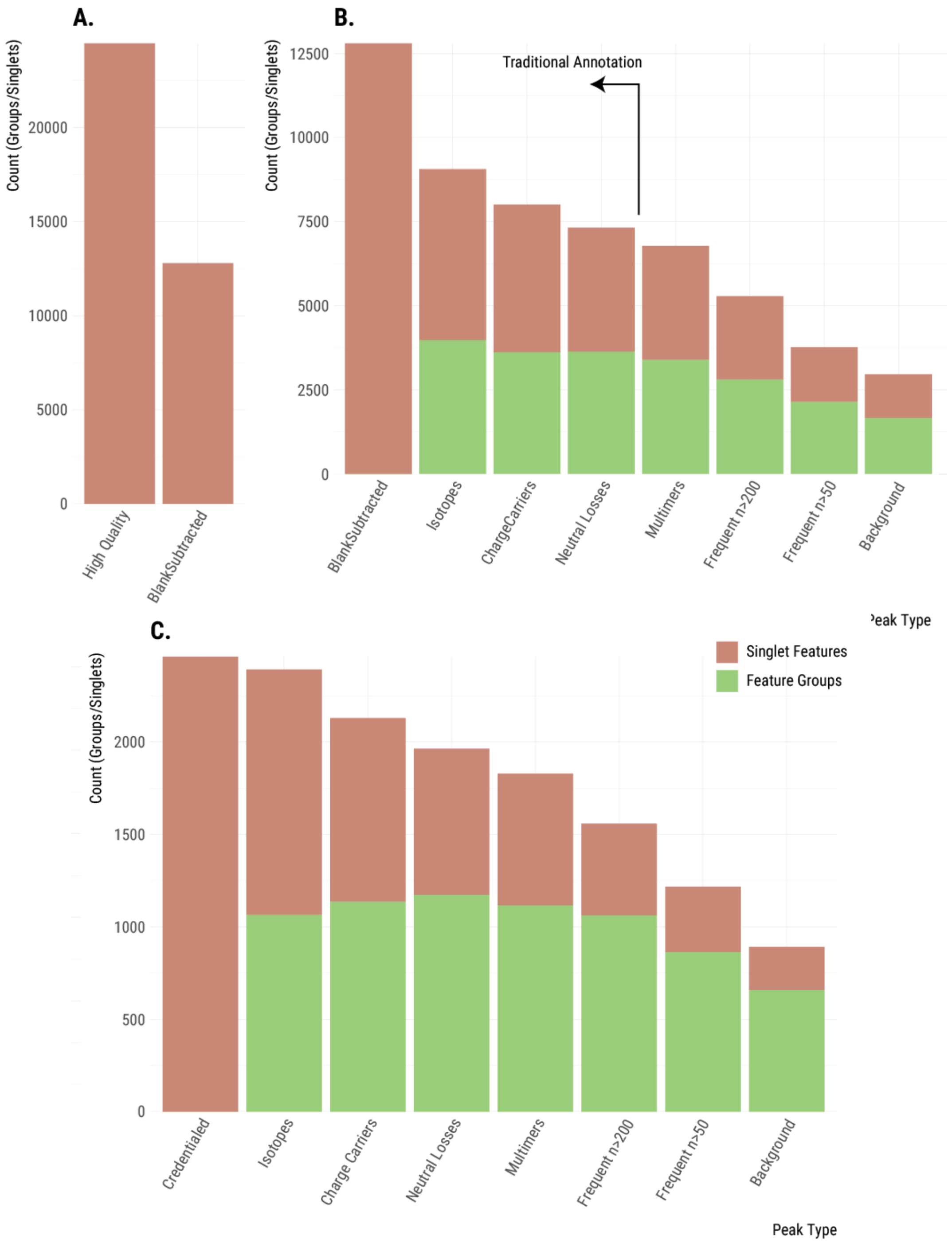


Figure 6.3. Plotting the maximum number of unique analytes detected throughout the steps of our annotation process. (A) Removal of features occurring in the blank. (B) Features are grouped as additional relationships are annotated. This reduces the maximum number of unique analytes. When a feature group contains multiple features, it is shown in green. When a feature group contains only a single feature (i.e., is a singlet), then it is shown in pink. Relationships from left to right: no relationships; isotopes; charge carriers; neutral losses; complex dimers (homo and hetero); frequent intrinsic relationships; situational adducts (background). (C) Similar annotation of features that were credentialed.

6.2.3 Homo and hetero multimers

We then expanded our search for degenerate relationships to complex adducts (i.e., two or more species non-covalently bound to one another, such as dimers, trimers, etc.). Our search included analytes adducted with themselves (homo-relationships), as well as analytes adducted with different analytes (hetero-relationships). We considered all coeluting features as potential multimer partners evaluating all [m, z] values as possible adduct formers. The charge state was specified based on observed isotopes, or assumed to be a charge state of 1. As our conditions generally form ions with a single charge, we balance the +2 charge from the observed ions with the loss of a proton [1.00783, +1] for each multimer. Thus, a complex hetero-relationship between three detected features will satisfy: $[m_1, z_1] + [m_2, z_2] – [1.00783, 1] = [m_3, z_3]$. Grouping these detected complex adducts reduced the number of feature groups in our data set to 3,400 (see “multimers” bar in Figure 3B-C).

Frequent intrinsic relationships show previously unannotated degeneracy

All current annotation approaches in untargeted metabolomics face the major challenge of determining the specific relationships to search for. While some relationships are well known and occur ubiquitously (such as the commonly annotated sodium or potassium adducts), constraining annotation to only these is significantly limiting. Other degenerate relationships are specific to experimental methodologies or the materials and reagents used during the analysis. Since there is no way to determine these relationships a priori, they have gone unannotated to date. Here we introduce an informatic approach to find data set wide, experimentally unique relationships that are implied by their context in the data. We then estimate their prevalence within our E. coli data set.

Common adducts and fragments will always coelute with the original analyte and will occur multiple times throughout the run.[3] We leverage this fact and recover “frequent intrinsic relationships” by performing a frequency analysis of mass differences between all pairs of features eluting within one second of each other. Unrelated but coeluting analytes will exhibit mass spacing that is random and, as such, will not be enriched in the frequency distribution. Thus, frequently occurring mass differences represent probable degenerate relationships. Mass differences were calculated assuming a charge state of 1, a simplification that limits the analysis to relationships that do not include a charge-state conversion. A Gaussian kernel density estimation was performed on the observed mass differences with a bandwidth of 0.00001 Da (our observed scan-to-scan mass error) (Figure 4A). The heights of the local maxima represent the frequency and mass dispersion of each mass difference. Mass differences that are frequent and similar in mass will have large density estimates. The 24 most frequently observed mass differences are listed in Table 2.

Table 6.2. Recovered frequent intrinsic relationships. Not all recovered relationships shown were used in the annotation. The local maxima of the density ordered by number of occurrences. These frequently occurring differences are good candidates for peak relationships. Several well-known relationships are present, including alternative charge carriers at the top of the list.

Δ Mass	**Δ Charge**	**Density**	**Known Species**
21.9820	0	60.4	gain:H^+ loss:Na^+
4.9554	0	55.2	gain:$NH4^+$ loss:Na^+
23.0760	0	33.6	
18.0107	0	32.5	loss:H2O
17.0266	0	30.5	loss:NH3
28.0314	0	26.7	C2H4
45.0580	0	23.4	C2H7N
14.0157	0	23.2	CH2
65.1230	0	19.6	
87.1046	0	18.2	C5H13N
42.0470	0	16.6	C3H6
44.0262	0	15.3	C2H4O
39.9926	0	13.3	C2O
7.1020	0	13.1	
15.9740	0	13.0	gain:K^+ loss:Na^+
70.0783	0	12.5	
29.0518	0	11.6	
36.0713	0	11.3	
15.9949	0	10.1	
1.9967	0	9.3	gain:k41 loss:k39
56.0627	0	9.3	
12.9952	0	8.7	
35.0373	0	8.7	
20.9292	0	8.5	gain:NH4+ loss:K+

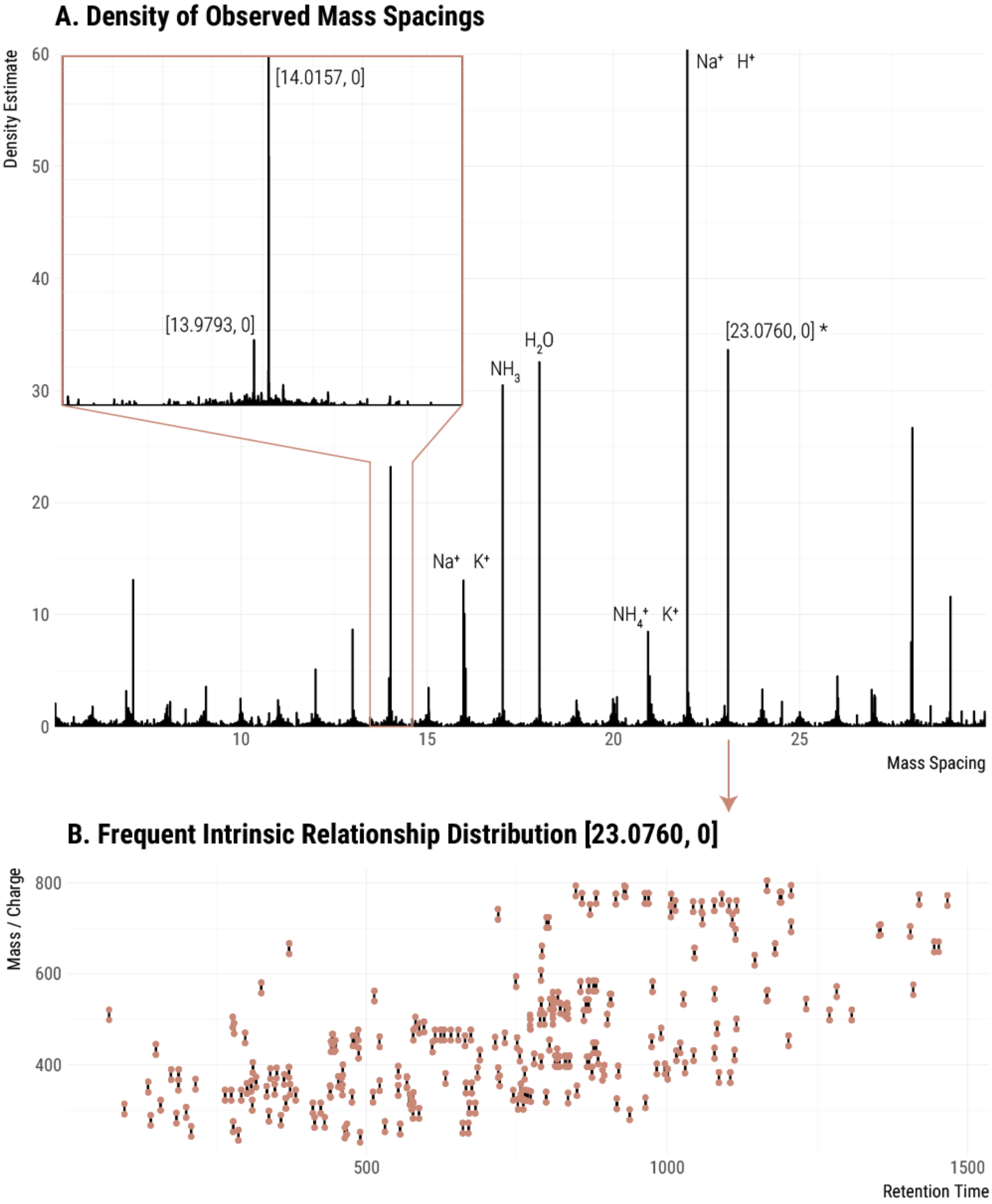


Figure 6.4. Detection of frequent intrinsic relationships. (A) The Gaussian kernel density of all pairwise peak relationships in the data set. Inset is a zoomed-in section around 14 Da. Known relationships are labeled with a formula. Unknown relationships are labeled with mass and charge transitions [m, z]. (B) Peak pairs of the recovered frequent intrinsic relationship [23.0760, 0] plotted in mass/charge and retention time (points). Line segments connect pairs with the specified spacing.

The effectiveness of the approach was confirmed by the recovery of two commonly known relationships as the most frequent relationships in our data set: the exchange of H^+ and Na^+ and the exchange of Na^+ and NH_4^+. This result indicated that the analysis of frequent intrinsic relationships offers novel insight into the nature of features detected in metabolomic data sets. Notably, the approach returned a multitude of relationships that had not been included in our prior searches. These commonly occurring relationships are likely adducts or fragments, and may be specific to our sample or experimental equipment/materials. Figure 4B shows the peak pairs observed with mass difference [23.0760, 0] throughout the data set.

We recognize that the recovery of frequent intrinsic relationships can also return relationships between commonly coeluting, non-degenerate analyte pairs. Fully saturated and partially unsaturated lipids, for example, commonly coelute and have a mass difference of [2.0156, 0] (H_2) (Han et al., 2012). We observed 176 occurrences of such a mass difference in our experiment. To minimize the risk of grouping unrelated features, we removed relationships with mass differences smaller than 15 Da and we applied two frequency cutoffs to illustrate the possible range of degeneracy. The conservative cutoff annotated and grouped frequent intrinsic relationships occurring more than 200 times (see bar labeled “commons n>200” in Figure 3B-C), while the aggressive cutoff annotated and grouped frequent intrinsic relationships occurring more than 50 times (see bar labeled “commons n>50” in Figure 3B-C). The inclusion of frequent intrinsic relationships in our data set annotation reduced the number of feature groups to 5,281 or 3,769, depending on the cutoff.

6.2.4 Situational adducts due to background ions contribute significantly to degeneracy

To further expand the scope of our annotation, we considered a source of adduct ions that are present throughout the run: the chemical background. These ions lack a chromatographic peak shape, but they are detected throughout the experiment due to the ionization of solvents, their additives, or any contaminants present. Because the background ions coelute with every feature, it is reasonable to expect that they will produce many adducts. We refer to adducts between analytes and other presently observed species (such as background ions) as “situational adducts”.

A low-mass spectrum was collected, deisotoped, and background ions appearing at intensities higher than 200,000 were used as potential participants in situational adduct formation (Figure 5). Annotation of the identified situational adducts reduced our number of feature groups to 2,961 (see bar labeled “background” in Figure 3B-C). This significant reduction in feature groups indicates that background ions are indeed a major source of feature inflation in our experiment. We also note that annotation of situational adducts reduced the number of feature groups containing only a single feature (i.e., singlets) to 1,288.

6.2.5 Background ions give rise to some frequent intrinsic relationships

Some frequent intrinsic relationships that we detected are indicative of novel adduction or fragmentation phenomena in our untargeted metabolomic data set, and we were interested in the origin of these unknown relationships. We speculated that some of the frequent intrinsic relationships that we discovered were the result of analyte adduction with the chemical background described above. In the simplest of cases, we found that some frequently occurring mass-to-charge differences between features corresponded to the mass-to-charge values of background ions. In more complex cases, however, a single analyte formed adducts with multiple background ions (Figure 5 and Figure 6) and therefore multiple situational adducts were detected for the same analyte. As the spacings between the background ions fix the spacings in the situational adduct features, we expect these repeatedly occurring spacings to be returned as frequent intrinsic relationships. Inspecting the returned frequent intrinsic relationships, we found several mass differences that also appear in the chemical background. This result is an additional confirmation of the effectiveness of frequent intrinsic relationship discovery and suggests that chemical background is a large source of feature inflation.

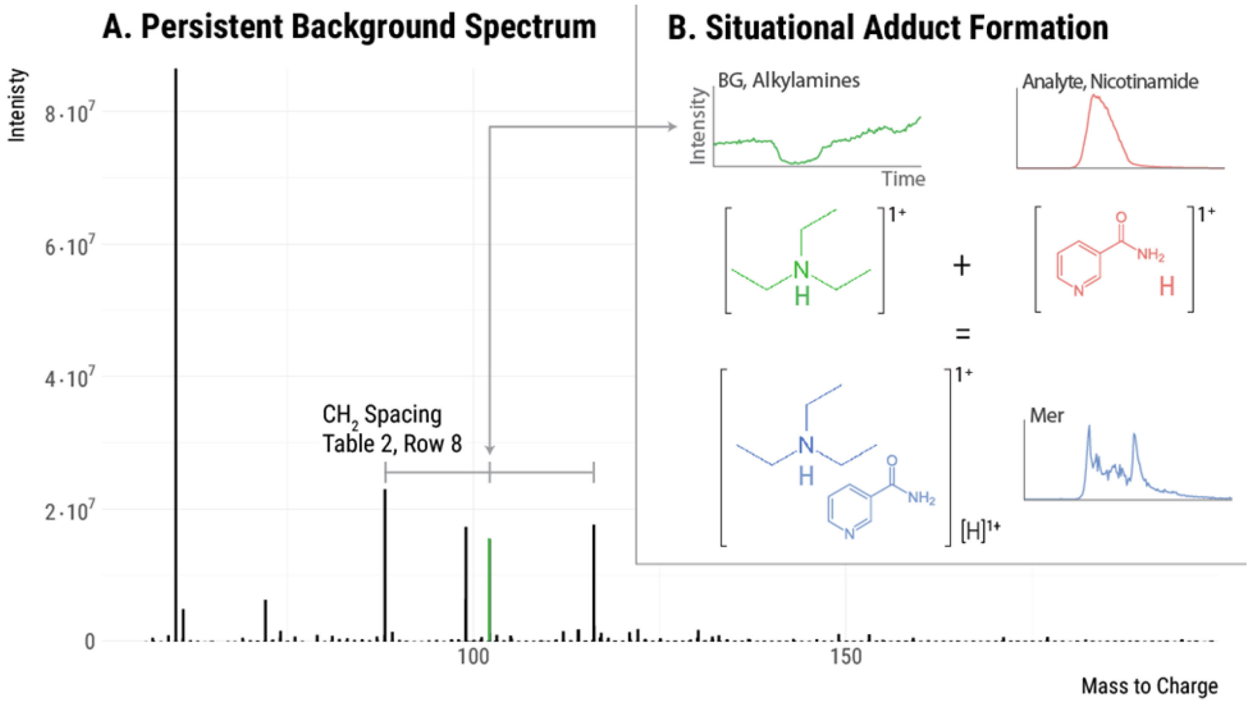


Figure 6.5. Situational adducts. (A) The persistent background spectrum observed in this experiment. The three indicated background peaks have mass spacings that correspond to a methylene group. These are likely an alkyl amine series with carbon numbers 5, 6, and 7. When these background species adduct with an analyte, situational adducts are formed. (B) An example of a situational adduct forming between background ion 102.1280 (a six carbon alkyl amine) and an eluting analyte. This process likely occurs with all three alkyl amine species throughout the run, giving rise to the frequent intrinsic relationships of mass 14.0157 (see Table 2, Row 8).

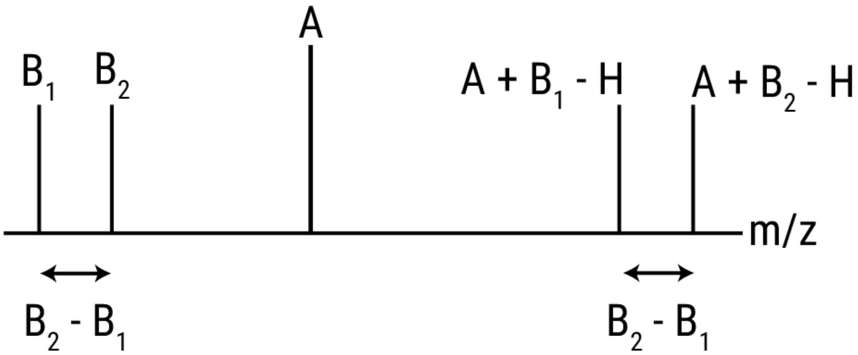


Figure 6.6. Schematic showing how background ions give rise to frequent intrinsic relationships. Analyte A is detected as an adduct of each background ion (B1 and B2). The spacing between the adducts (A+B1-H and A+B2-H) is equal to the spacing between the background ions.

We also performed formula decomposition on the frequent intrinsic relationships to further elucidate their origins. Interestingly, chemical formula CH_2, C_2H_4, and C_3H_6 were found in the frequent intrinsic relationships exhibited by the chemical background. Additional analysis of the background ions indicated that they were an alkyl amine series. These species are known to form strong adducts and are commonly found as contaminants in alcohol solvents.[94] We note that our laboratory has never performed ion-pairing experiments and the source of these reagents was solvent impurity as indicated by the series rather than sole presence of triethylamine. In developing our methods, we attempted to find solvents with the lowest possible levels of chemical background (Burdick & Jackson brand purchased from Honeywell). Unfortunately, alkyl amines seem to be ubiquitous in methanol and isopropanol LC/MS solvents.

6.2.6 Removing artifacts and contaminants by credentialing

The degenerate relationships that we annotated above led to a striking reduction in the number of feature groups, indicating that fewer than 15% of the total 25,230 features that were detected in E. coli correspond to unique analytes. Even after this extensive annotation process, however, two sources of feature inflation remained in artifacts and contaminants. We applied an alternative experimental approach called credentialing to filter these features associated with artifacts and contaminants. The credentialing process introduces an isotopic signature into biological analytes during E. coli growth.[4] Features in our data set displaying this isotopic signature are deemed "credentialed", as they are known to be of *E. coli* origin. In contrast, features that do not display this isotopic signature are annotated as artifacts or contaminants. Credentialing does not rely on any of the relationship annotation approaches that we described above, and is thus an orthogonal and highly complementary approach to data analysis.

We first filtered non-credentialed features from the raw data set on the basis of isotopic signatures. The resulting set of features is free of artifacts, noise, and contaminants. This process returned 2,462 high-quality, credentialed features. We then took these credentialed features through the same annotation process as the full data set to remove degeneracy. Annotation of degeneracy reduced the estimated number of unique E. coli analytes being measured to 832 (Figure 3C).

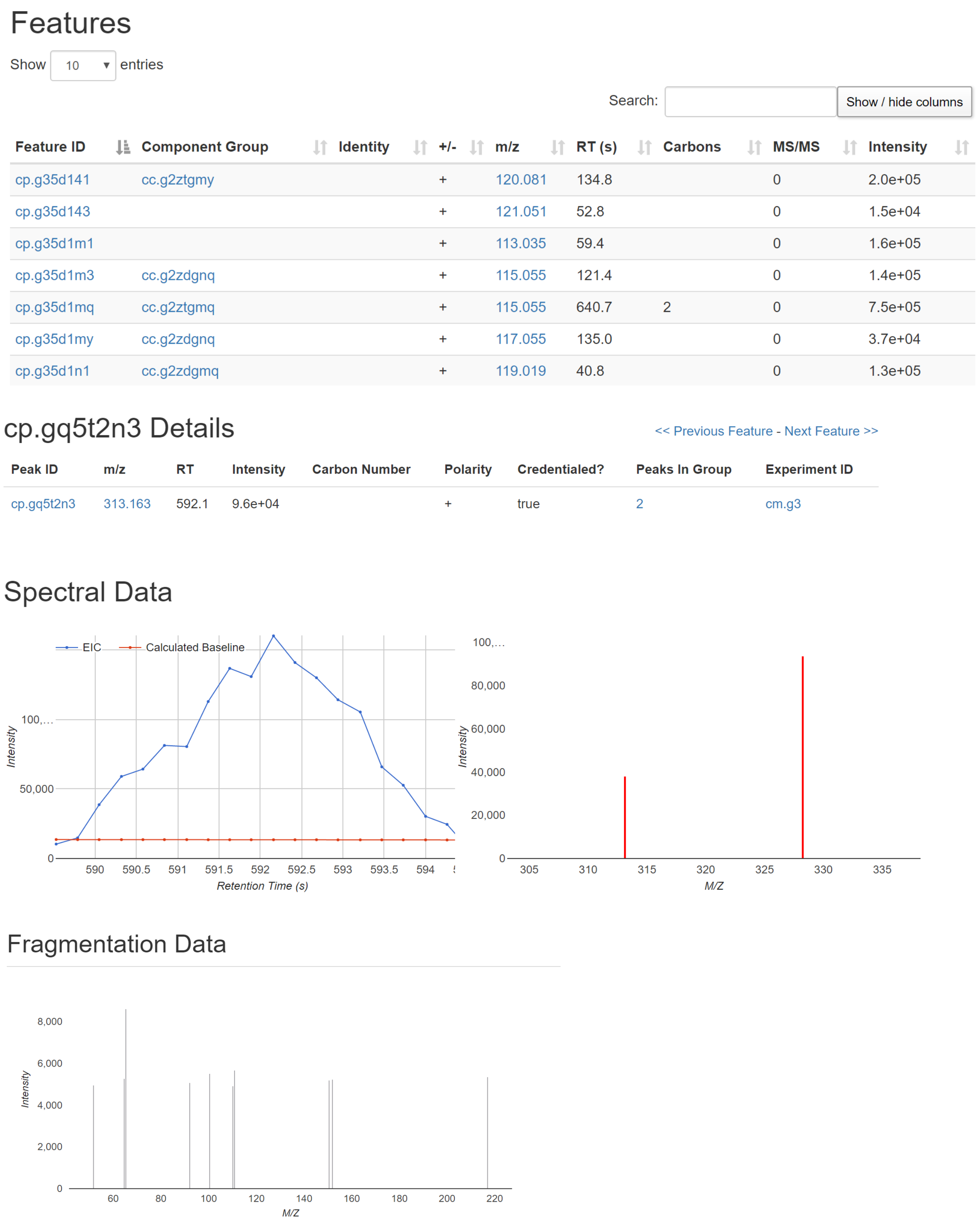
6.2.7 creDBle: a database for thoroughly annotated reference data sets

An alternative approach to each investigator having to identify the relatively small number of features corresponding to unique, bona fide metabolites from every experiment is to create thoroughly annotated reference data sets. Reference data sets have been shown to be effective in other profiling sciences, such as genomics (for example, during the EST collection era of gene identification in the 1990's).[186,187] The idea is for one laboratory to first identify all of the unique metabolites that can be detected from a given sample with a given experimental methodology. Then, other laboratories performing the same experiment benefit by having to target only these reference analytes in their subsequent experiments. Of course, the major challenge of this strategy is that there are a multitude of experimental methods currently being used in untargeted metabolomics, each of which will have to be annotated for different sample types.[181]

There may also be other benefits to having a repository of thoroughly annotated data sets. Knowing the comprehensive list of unique metabolites that can be detected with specific experimental protocols, for example, will be invaluable to designing LC/MS-based metabolomic experiments. Although the number of detected features is often used as an indicator of experimental coverage, our work suggests that this is an unreliable metric. [87,88] Instead, it would be preferred if researchers based their experimental design on the numbers of unique metabolites known to be detected. Additionally, even if the sample of interest has not been annotated, researchers might be able to use annotated data from other sample types (e.g., E. coli) as a touchstone to evaluate data from their experiments and to compare it to others.

As a first step in establishing a repository for thoroughly annotated reference data sets, we have created the creDBle database. All credentialed features for the reference E. coli data set described here have been deposited in creDBle. Degeneracy annotations as well as accurate mass, retention times, and fragmentation patterns are included. creDBle is freely available on the Web at http://creDBle.wustl.edu/ and provides a convenient companion resource for credentialed E. coli standards (Figure 7). All data within creDBle (including fragmentation patterns for identified metabolites) can be freely downloaded.

The addition of more analyses to creDBle will greatly expand its applicability. Our first goal is to repeat the annotation processes above for credentialed E. coli samples analyzed with different methods (e.g., different extraction protocols, chromatography, mass spectrometers, etc.). Notably, identification of metabolites from these annotated experiments will provide a readily available set of complex standards. As the number of credentialed E. coli experiments within creDBle increases, we hope that it will eventually provide a common reference point with enough observations in each experiment to model and normalize some of the variation that has historically prevented cross-laboratory data comparisons. This, in turn, would make data sets present in repositories, when run with a credentialed standard extract, more amenable to reprocessing and meta-analysis.



Features

Show 10 ▾ entries

Search: Show / hide columns

Feature ID	Component Group	Identity	+/-	m/z	RT (s)	Carbons	MS/MS	Intensity
cp.g35d141	cc.g2ztgmy		+	120.081	134.8		0	2.0e+05
cp.g35d143			+	121.051	52.8		0	1.5e+04
cp.g35d1m1			+	113.035	59.4		0	1.6e+05
cp.g35d1m3	cc.g2zdgnq		+	115.055	121.4		0	1.4e+05
cp.g35d1mq	cc.g2ztgmq		+	115.055	640.7	2	0	7.5e+05
cp.g35d1my	cc.g2zdgnq		+	117.055	135.0		0	3.7e+04
cp.g35d1n1	cc.g2zdgmq		+	119.019	40.8		0	1.3e+05

cp.gq5t2n3 Details

<< Previous Feature - Next Feature >>

Peak ID	m/z	RT	Intensity	Carbon Number	Polarity	Credentialed?	Peaks In Group	Experiment ID
cp.gq5t2n3	313.163	592.1	9.6e+04		+	true	2	cm.g3

Spectral Data

Fragmentation Data

Figure 6.7. Screenshots from the creDBle database. (A) The list of credentialed features showing m/z, retention time, polarity, grouping, and intensity. (B) A credentialed features page showing the extracted ion chromatogram, credentialed isotopes, and fragmentation data.

6.3 Conclusion

Detecting tens of thousands of LC/MS features from biological samples is typical in untargeted metabolomics, however, to date it has been unclear how many unique metabolites are actually being profiled. Our work here evaluated one representative untargeted metabolomics data set from E. coli to set an upper bound on the number of unique metabolites being measured. By using a new context-driven approach to identify degenerate features arising from the same metabolite, we determined that the ~25,000 features detected in our experiment corresponded to fewer than 2,961 unique analytes. An orthogonal and complimentary approach using credentialing isotope signatures to identify artifacts and contaminants similarly reduced the number of unique analytes detected. Out of the total ~25,000 features detected, only 832 passed both our degeneracy and credentialing filters. Accurate masses, retention times, fragmentation patterns, and degeneracy annotations for these 832 features have been deposited in the creDBle database.

We wish to emphasize that our work is unrelated to the size of the E. coli metabolome and should not be interpreted as an indication of the total number of intracellular metabolites present. There are certainly more than 832 E. coli metabolites.[188] The purpose of our work was only to assess how many unique metabolites are being measured in a representative untargeted metabolomics experiment. Additionally, we note that our context-driven analysis of degeneracy is not exhaustive. Relationships that are uncommon and not indicated by background ions remain unannotated and may further reduce the number of unique analytes detected. Notwithstanding, our results suggest that there are an order of magnitude more features than unique metabolites in untargeted metabolomics experiments. This has important implications for designing untargeted metabolomics experiments and influences strategies for interpreting the data produced before establishing metabolite identifications.

6.4 Methods

6.4.1 Materials

U-^{13}C-D-glucose was purchased from Cambridge Isotope Laboratories Inc. (Andover, MA). E. coli strain K12, MG1655 was purchased from ATCC (Manassas, VA). Lennox LB broth powder and 5x M9 salts were purchased from Sigma-Aldrich (St. Louis, MO). Cell culture was performed with ultrapure water provided by a Milli-Q system (Millipore). LC/MS grade, Burdick & Jackson brand water, acetonitrile, methanol, and isopropanol were purchased from Honeywell (Morris Plains, NJ). Cortecs T3 reversed phase UPLC columns and column guards were purchased from Waters Corporation (Milford, MA).

6.4.2 Generating credentialed samples

E. coli was grown in a rotary shaker at 37 ⁰C and 300 rpm as previously described (Mahieu et al., 2014). M9 minimal media was used with a glucose concentration of 2 g/L. Two cultures were grown in parallel, one using natural abundance glucose and a second using U-^{13}C-glucose as the only carbon source. Cultures were grown to OD600 = 0.7, at which point they were harvested.

For harvest, flasks were removed from the shaker and placed on ice. The contents of each flask were pipetted into 50 mL conical tubes and centrifuged at 3200 g for 10 minutes. The supernatant was decanted and remaining media was gently rinsed off the top of the pellet with 0.5 mL of water. Conical tubes were then placed in liquid nitrogen and lyophilized for 24 hours, or until dry. This powdered, credentialed E. coli standard was then extracted to generate samples for untargeted metabolomic analysis.

Several replicate extractions were performed in parallel by using a previously described method.[4] Briefly, five 2.5 mg samples of each ^{12}C and ^{13}C material were weighed out, while two empty tubes were included as extraction blanks. To these, 1,000 µL of 2:2:1 methanol:acetonitrile:water was added, followed by three freeze-thaw cycles with sonication and vortexing. After centrifugation, the supernatant was vacuum concentrated and reconstituted in 100 µL of 1:1 acetonitrile:water with internal standards. From these extracts, three samples were aliquoted for LC/MS analysis: natural abundance extract, a mix of 1:1 natural abundance extract and ^{13}C extract, and the blank extract.

6.4.3 Data set generation

Each sample was analyzed five times as analytical replicates. The untargeted LC/MS data set was generated in positive polarity on a Q Exactive Plus mass spectrometer with a HESI II source coupled to a Dionex 3000RSLC. The data set was collected with the following settings: aux gas, 5; sheath gas, 35; sweep gas, 2; capillary temperature, 300 ºC; aux gas temperature, 200 ºC; spray voltage, 3.5 kV; needle diameter, 34 ga; s-lens, 75 V; mass range, 100–1500 Da; resolution 70,000; micro scans, 1; max injection time; 100 ms; automatic gain control target: 1e6. Reversed-phase chromatography was performed with the Waters Cortecs T3 (2.1mm x 50mm, 1.6um) column at a flow rate of 300 µL/min and a column temperature of 50 ºC. Solvents were: A, water + 5mM ammonium acetate + 5uM ammonium phosphate; B, 9:1 isopropanol:methanol + 5mM ammonium acetate + 5um ammonium phosphate. An injection volume of 2 µL was used with a linear gradient of (minutes, %A): 0, 100; 28, 0; 30, 0; 30, 100; 35, 100.

Chromatographic features were detected by using a set of in-house algorithms. Mass traces were retained if they were longer than 10 scans, excluding missing peaks. Baselines for each mass trace were calculated by using the iterative restricted least squares method from the baseline R package.

Model based peak detection was performed by using the skew normal distribution as a model peak distribution. This process resulted in a set of features detected in each replicate run. Features were grouped by mass and retention time using a density based method. Retention time drift and mass drift were corrected by fitting a loess curve of degree 2 to the distance from the mean value of each group against the mean retention time of each group.

Subtle variations from run to run cause many features to be integrated differently and sometimes not integrated in each file. Further, closely eluting peaks often lead to incorrectly grouped features. To refine the individual datasets and get a set of detected peaks consistent with all replicate runs, we applied the Warpgroup algorithm.[2] Warpgroup is available at https://github.com/nathaniel-mahieu/warpgroup. Warpgroup takes as input the raw data and each file's detected features combining them to output a set of consensus features. Parameters: sc.aligned.lim, 9; pct.pad, 0.1; min.peaks, 3.

This consensus data set set is the standard output of an untargeted metabolomics experiment. As such, it was taken as a representative dataset for annotation of detected signals.

6.4.4 Mz.unity based annotation

Mz.unity was applied to the dataset to detect mass and charge ([m, z]) relationships between eluting signals derived from a single analyte.[3] We use [m, z] to denote the mass and charge of a species, where both are specified as opposed to m/z where the two are convolved. These searches find sets of features that have [m,z]s differing by a specific amount. Differences are specific to relationships, for example, loss of ^{12}C and gain of ^{13}C ([+1.003355, 0]), or loss of water ([-18.01057, 0]).

Searches were first performed for the following relationships: isotopes, common charge carriers, common neutral losses, and common adducts. We then searched for dimers between coeluting features. The dimer search posits each eluting [m, z] as a possible adduct former. The charge state was specified based on observed isotopes, or assumed to be a charge of 1. As dimers are normally formed with a charge from only one constituent, we also assumed the loss of a proton [1.00783, +1] for each pair.

Mz.unity is available at https://github.com/nathaniel-mahieu/mz.unity.

6.4.5 Frequent intrinsic relationships

Groups of features eluting within 1 second of each other were taken, and their pairwise [m, z] differences were calculated after assuming a charge state of 1. A Gaussian kernel density estimation was performed on the mass differences with a bandwidth of 0.00001 Da (our observed scan-to-scan mass error). Local maxima of the density estimate were detected along with the estimated density at those locations. The heights of the local maxima represent the frequency and mass dispersion of each mass difference. Mass differences that are more frequent and more similar in mass will have larger density estimates.

We took enriched mass differences larger than 15 Da and occurring more than 50 times throughout the dataset into the mz.unity search.

6.4.6 Situational adducts

Background ions that lack a chromatographic peak shape are an ever-present set of species that often form adducts with eluting analytes. These situational adducts are then detected as features having a chromatographic peak shape. A low mass background spectrum was collected, containing detected ions above 50 Da. This spectrum was deisotoped and background species appearing at higher than 200,000 intensity were used to seed possible adduct relationships. The [m, z]s of each background peak were included in the dimer search, as above after specifying the charge state based on observed isotopes or assuming a charge of 1.

6.4.7 Credentialing

A high-confidence set of features were recovered from the $^{12+13}$C dataset by applying version 3.0 of the credentialing algorithm, which is available at https://github.com/pattilab/credential. Credentialing searches for pairs of peaks that have precise isotopic spacing expected from U-^{12}C and U-^{13}C analytes.[4] This provides a filter against many forms of noise, contaminants, and artifact features. Credentialing was run with the parameters: ppmwid, 8; rtwid, 1.2; cd, 1.00335; mpc, c(12, 120); ratio, 1; ratio.lim, 0.1; maxnmer, 4. Credentialed features from the $^{12+13}$C data set were then matched to the ^{12}C dataset by applying retention time and mass correction as above before grouping.

6.4.8 Credentialed feature characterization

The set of credentialed features were further characterized for deposition in the creDBle database. Targeted MS/MS was performed on the credentialed features with a 0.4 Da window width and a stepped collision energy of 10, 30, and 90 V. Annotations and feature groupings of the credentialed features were taken from the previously performed mz.unity annotations.

6.4.9 The creDBle database

Characterization of all credentialed features from this data set was deposited in the creDBle database. The data are freely available at http://credble.wustl.edu/ and easily downloadable in JSON format via the REST API. This includes *m*/*z*, retention time, annotation grouping, MS/MS spectra, credentialed isotopes, and extracted ion chromatograms.

Chapter 7.

Concluding Remarks

Metabolomics remains a rapidly expanding field even 20 years after its inception. Still, the exceptional promise of untargeted analysis remains impeded by complex variance and massive dataset degeneracy.[189] Warpgroup, mz.unity, credentialing and creDBle address these critical needs and chart a course to truly systems-level metabolomics.

7.1 The Big Picture

When attempting to comprehensively understand metabolomic datasets it became apparent that preprocessing steps were critically important to downstream analysis. In particular the peak detection process preceded all feature-dependent analysis and as such any interpretation of individual features relies on accurate peak detection. Warpgroup was developed to improve the reproducibility of peak detection and decrease noise – these advances made the later steps including credentialing and mz.unity annotation tractable problems.

While forming a contextual understanding of features in metabolomic datasets it became apparent that current approaches were limited to only the simplest relationship types. Exploration of complex adducts and distal fragments required a more flexible search approach. To this end mz.unity was developed in a manner that allows one to search for any specified relationship. This advance allowed the systematic evaluation of complex and cross polarity adducts and contextual relationship recovery – two major contributions to dataset annotation.

Finally, it was clear that though great effort had been undertaken to minimize the contribution of contaminants and informatic error to features, many mistakes were still being made. As such the credentialing methodology was introduced to recover reproducible sets of biological features.

In concert these developments enabled the first unbiased catalog of analyte features from an untargeted dataset - creDBle. Additionally the combined application of these algorithms have provided insight into the analyte content and degeneracy of metabolomic datasets – a result that will guide the design of next generation metabolomic experiments.

7.2 Future Work

Though these contributions represent major conceptual advances to the metabolomic workflow, many challenges remain in the field.

Relationship annotation is the most promising approach developed herein. Current applications of mz.unity take a conservative approach in order to minimize false positive annotations. Truly comprehensive relationship annotations necessitate a statistically driven evaluation of each putative relationship. The likelihood of a relationship can be conditioned on many observations – expected mass error, prior knowledge of the likelihood of occurrence, intensities of the involved species, gradient conditions, source conditions, and other observed relationships all contribute information relevant to putative relationships. The problem of evaluating putative relationships can be stated as finding the optimum graph subsets which describe the observed signals, minimizing some measure of over-aggregation while maximizing some relationship based score. Further, evaluating self versus non-self-relationships and predicting the original analyte mass based on the observed signals are additional goals amenable to this framework. The development of mz.unity into a comprehensive, easy to use algorithm will certainly improve our ability to compute on metabolomic datasets.

Additionally, Warpgroup offers one solution to the isolated peak detection problem. It would be ideal to improve the peak detection problem in a prospective manner. To this end incorporating additional information into the peak detection step will have a major impact. The relationship search as described above offers a major unused constraint on the peak detection process. Base peaks have been used to predict and reinforce isotopic peaks during peak detection for example. This should be extended to encompass the entire relationship graph – sodium adducts, dimers, even across polarities. The detection of peaks and the annotation of relationships are interdependent, and can be co-optimized to maximize the robustness of both steps.

Current approaches treat each experiment with identical chromatographies but different polarities or ionization types as independent (ESI +/- and APCI +/- for example). This results in another large form of degeneracy that is yet to be annotated. The mz.unity approach offers the ability to search for relationships between these disparate datasets, and unifying peak detection across polarities and ionization types is only feasible with annotation driven peak detection.

Ultimately, improved computational comprehension of metabolomic datasets will enable the full power of the technology. Currently datasets are redundant and challenging to interpret. Incorporating the relationship graph and peak detection will allow an abstract representation of metabolomic datasets including the context of all detected signals. The resulting dataset will be computable, offer a strong foundation to train machine learning models for analysis, and allow for rapid extraction of biologically relevant information from these datasets. The future of metabolomics is bright.

References

(1) Mahieu, N. G.; Genenbacher, J. L.; Patti, G. J. *Curr. Opin. Chem. Biol.* **2016**, *30*, 87–93.

(2) Mahieu, N. G.; Spalding, J. L.; Patti, G. J. *Bioinformatics* **2015**, btv564.

(3) Mahieu, N. G.; Spalding, J. L.; Gelman, S. J.; Patti, G. J. *Anal. Chem.* **2016**, *88* (18), 9037–9046.

(4) Mahieu, N. G.; Huang, X.; Chen, Y.-J.; Patti, G. J. *Anal. Chem.* **2014**, *86* (19), 9583–9589.

(5) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, *13* (4), 263–269.

(6) Dettmer, K.; Aronov, P. A.; Hammock, B. D. *Mass Spectrom. Rev.* **2007**, *26* (1), 51–78.

(7) Millington, D. S.; Kodo, N.; Norwood, D. L.; Roe, C. R. *J. Inherit. Metab. Dis.* **1990**, *13* (3), 321–324.

(8) Lenz, E. M.; Wilson, I. D. *J. Proteome Res.* **2007**, *6* (2), 443–458.

(9) Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* **1999**, *29* (11), 1181–1189.

(10) Hoult, D. I.; Busby, S. J. W.; Gadian, D. G.; Radda, G. K.; Richards, R. E.; Seeley, P. J. *Nature* **1974**, *252* (5481), 285–287.

(11) Griffiths, W. J.; Wang, Y.; McCrum, E. C.; Russell, D. W.; Hamberg, M.; Alvelius, G.; Sjövall, J.; Turton, J.; Wang, Y.; Griffiths, W. J.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J. *Chem. Soc. Rev.* **2009**, *38* (7), 1882.

(12) Gates, S. C.; Sweeley, C. C. *Clin. Chem.* **1978**, *24* (10), 1663–1673.

(13) Link, H.; Fuhrer, T.; Gerosa, L.; Zamboni, N.; Sauer, U. *Nat. Methods* **2015**, *12* (11), 1091.

(14) Tomoyoshi Soga, *,†; Yoshiaki Ohashi, †; Yuki Ueno, †; Hisako Naraoka, †; Masaru Tomita, † and; Takaaki Nishioka†, ‡. **2003**.

(15) Pace, N. R. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98* (3), 805–808.

(16) Nealson, K. H.; Conrad, P. G. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **1999**, *354* (1392), 1923–1939.

(17) Alan Saghatelian; Sunia A. Trauger; Elizabeth J. Want; Edward G. Hawkins; Gary Siuzdak, and; Cravatt*, B. F. **2004**.

(18) Wimmer, M. J.; Rose, I. A. *Annu. Rev. Biochem.* **1978**, *47* (1), 1031–1078.

(19) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M.-A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.;

Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; MacInnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L. *Nucleic Acids Res.* **2007**, *35* (Database), D521–D526.

(20) Ogbaga, C. C.; Stepien, P.; Dyson, B. C.; Rattray, N. J. W.; Ellis, D. I.; Goodacre, R.; Johnson, G. N. *PLoS One* **2016**, *11* (5), e0154423.

(21) De Luca, V.; St Pierre, B. *Trends Plant Sci.* **2000**, *5* (4), 168–173.

(22) Pál, C.; Papp, B.; Lercher, M. J.; Csermely, P.; Oliver, S. G.; Hurst, L. D. *Nature* **2006**, *440* (7084), 667–670.

(23) Pilkis, S. J.; el-Maghrabi, M. R.; Claus, T. H. *Diabetes Care* **1990**, *13* (6), 582–599.

(24) Boiteux, A.; Hess, B. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **1981**, *293* (1063), 5–22.

(25) Fiehn, O. *Plant Mol. Biol.* **2002**, *48* (1/2), 155–171.

(26) Michie, K. A.; Löwe, J. *Annu. Rev. Biochem.* **2006**, *75* (1), 467–492.

(27) Hegardt, F. G. *Biochem. J.* **1999**, No. 3, 569–582.

(28) Suhre, K.; Shin, S.-Y.; Petersen, A.-K.; Mohney, R. P.; Meredith, D.; Wägele, B.; Altmaier, E.; CARDIoGRAM; Deloukas, P.; Erdmann, J.; Grundberg, E.; Hammond, C. J.; Angelis, M. H. de; Kastenmüller, G.; Köttgen, A.; Kronenberg, F.; Mangino, M.; Meisinger, C.; Meitinger, T.; Mewes, H.-W.; Milburn, M. V.; Prehn, C.; Raffler, J.; Ried, J. S.; Römisch-Margl, W.; Samani, N. J.; Small, K. S.; Wichmann, H.-E.; Zhai, G.; Illig, T.; Spector, T. D.; Adamski, J.; Soranzo, N.; Gieger, C. *Nature* **2011**, *477* (7362), 54–60.

(29) Novotny, M. V.; Soini, H. A.; Mechref, Y. *J. Chromatogr. B* **2008**, *866* (1–2), 26–47.

(30) EBENHOH, O.; Heinrich, R. *Bull. Math. Biol.* **2001**, *63* (1), 21–55.

(31) Meléndez-Hevia, E.; Waddell, T. G.; Cascante, M. *J. Mol. Evol.* **1996**, *43* (3), 293–303.

(32) Doerr, A. *Nat. Methods* **2007**, *4* (1), 8–9.

(33) Pietiläinen, K. H.; Sysi-Aho, M.; Rissanen, A.; Seppänen-Laakso, T.; Yki-Järvinen, H.; Kaprio, J.; Orešič, M. *PLoS One* **2007**, *2* (2), e218.

(34) Griffin, J. L. *Curr. Opin. Chem. Biol.* **2006**, *10* (4), 309–315.

(35) Griffin, J. L.; Shockcor, J. P. *Nat. Rev. Cancer* **2004**, *4* (7), 551–561.

(36) Holmes, E.; Wilson, I. D.; Nicholson, J. K. *Cell* **2008**, *134* (5), 714–717.

(37) Fiehn, O.; Kopka, J.; Dörmann, P.; Altmann, T.; Trethewey, R. N.; Willmitzer, L. *Nat. Biotechnol.* **2000**, *18* (11), 1157–1161.

(38) Fiehn, O.; Robertson, D.; Griffin, J.; van der Werf, M.; Nikolau, B.; Morrison, N.; Sumner, L.

W.; Goodacre, R.; Hardy, N. W.; Taylor, C.; Fostel, J.; Kristal, B.; Kaddurah-Daouk, R.; Mendes, P.; van Ommen, B.; Lindon, J. C.; Sansone, S.-A. *Metabolomics* **2007**, *3* (3), 175–178.

(39) Morris, M.; Watkins, S. M. *Curr. Opin. Chem. Biol.* **2005**, *9* (4), 407–412.

(40) Dunn, W. B.; Erban, A.; Weber, R. J. M.; Creek, D. J.; Brown, M.; Breitling, R.; Hankemeier, T.; Goodacre, R.; Neumann, S.; Kopka, J.; Viant, M. R. *Metabolomics* **2013**, *9* (S1), 44–66.

(41) Griffin, J. L.; Vidal-Puig, A. *Physiol. Genomics* **2008**, *34* (1), 1–5.

(42) Schmitt-Kopplin, P.; Gabelica, Z.; Gougeon, R. D.; Fekete, A.; Kanawati, B.; Harir, M.; Gebefuegi, I.; Eckel, G.; Hertkorn, N. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (7), 2763–2768.

(43) Schönbein, C. F. *J. für Prakt. Chemie* **1861**, *84* (1), 410–415.

(44) Still, W. C.; Kahn, M.; Mitra, A. *J. Org. Chem.* **1978**, *43* (14), 2923–2925.

(45) Eulitz, K.; Yurawecz, M. P.; Sehat, N.; Fritsche, J.; Roach, J. A. G.; Mossoba, M. M.; Kramer, J. K. G.; Adlof, R. O.; Ku, Y. *Lipids* **1999**, *34* (8), 873–877.

(46) Radhakrishna, T.; Lakshmi Narayana, C.; Sreenivas Rao, D.; Vyas, K.; Om Reddy, G. *J. Pharm. Biomed. Anal.* **2000**, *22* (4), 627–639.

(47) van Deemter, J. J.; Zuiderweg, F. J.; Klinkenberg, A. *Chem. Eng. Sci.* **1956**, *5* (6), 271–289.

(48) Pitt, J. J. *Clin. Biochem. Rev.* **2009**, *30* (1), 19–34.

(49) Dole, M.; Mack, L. L.; Hines, R. L.; Mobley, R. C.; Ferguson, L. D.; Alice, M. B. *J. Chem. Phys.* **1968**, *49* (5), 2240–2249.

(50) Fenn, J.; Mann, M.; Meng, C.; Wong, S.; Whitehouse, C. *Science (80-.).* **1989**, *246* (4926), 64–71.

(51) Taylor, G. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **1964**, *280* (1382), 383–397.

(52) Cremer, J. E.; Heath, D. F. *Biochem. J.* **1974**, *142* (3), 527–544.

(53) Kebarle, P. *J. Mass Spectrom.* **2000**, *35* (7), 804–817.

(54) Yamashita, M.; Fenn, J. B. *J. Phys. Chem.* **1984**, *88* (20), 4451–4459.

(55) Bruins, A. P. *Mass Spectrom. Rev.* **1991**, *10* (1), 53–77.

(56) Klee, S.; Derpmann, V.; Wißdorf, W.; Klopotowski, S.; Kersten, H.; Brockmann, K. J.; Benter, T.; Albrecht, S.; Bruins, A. P.; Dousty, F.; Kauppila, T. J.; Kostiainen, R.; O'Brien, R.; Robb, D. B.; Syage, J. A. *J. Am. Soc. Mass Spectrom.* **2014**, *25* (8), 1310–1321.

(57) Hu, Q.; Noll, R. J.; Li, H.; Makarov, A.; Hardman, M.; Graham Cooks, R. *J. Mass Spectrom.* **2005**, *40* (4), 430–443.

(58) Makarov*, A. **2000**.

(59) Alexander Makarov, *; Eduard Denisov; Alexander Kholomeev; Wilko Balschun; Oliver Lange; Kerstin Strupat, and; Horning, S. **2006**.

(60) Chernushevich, I. V.; Loboda, A. V.; Thomson, B. A. *J. Mass Spectrom.* **2001**, *36* (8), 849–865.

(61) Wiley, W. C.; McLaren, I. H. *Rev. Sci. Instrum.* **1955**, *26* (12), 1150–1157.

(62) Sakoe, H.; Chiba, S. *IEEE Trans. Acoust.* **1978**, *26* (1), 43–49.

(63) Vintsyuk, T. K. *Cybernetics* **1972**, *4* (1), 52–57.

(64) Aach, J.; Church, G. M. *Bioinformatics* **2001**, *17* (6), 495–508.

(65) Prince, J. T.; Marcotte, E. M. *Anal. Chem.* **2006**, *78* (17), 6140–6152.

(66) Podwojski, K.; Fritsch, A.; Chamrad, D. C.; Paul, W.; Sitek, B.; Stuhler, K.; Mutzel, P.; Stephan, C.; Meyer, H. E.; Urfer, W.; Ickstadt, K.; Rahnenfuhrer, J.; Stühler, K.; Mutzel, P.; Stephan, C.; Meyer, H. E.; Urfer, W.; Ickstadt, K.; Rahnenführer, J. *Bioinformatics* **2009**, *25* (6), 758–764.

(67) Grandjean, M. *Cogent Arts Humanit.* **2016**, *3* (1).

(68) Mashaghi, A. R.; Ramezanpour, A.; Karimipour, V. *Eur. Phys. J. B* **2004**, *41* (1), 113–121.

(69) Csardi, G.; Nepusz, T. *InterJournal* **2006**, *Complex Sy*, 1695.

(70) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinformatics* **2010**, *11*, 395.

(71) Tautenhahn, R.; Bottcher, C.; Neumann, S.; Böttcher, C.; Neumann, S.; Bottcher, C.; Neumann, S. *BMC Bioinformatics* **2008**, *9* (1), 504.

(72) Antoniewicz, M. R. *J. Ind. Microbiol. Biotechnol.* **2015**, *42* (3), 317–325.

(73) Clasquin, M. F.; Melamud, E.; Rabinowitz, J. D. In *Current Protocols in Bioinformatics*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2012; Vol. Chapter 14, p Unit14.11.

(74) Kastenmüller, G.; Römisch-Margl, W.; Wägele, B.; Altmaier, E.; Suhre, K. *J. Biomed. Biotechnol.* **2011**, *2011*, 1–7.

(75) Katajamaa, M.; Orešič, M. *BMC Bioinformatics* **2005**, *6* (1), 179.

(76) Lommen, A.; Kools, H. J. *Metabolomics* **2012**, *8* (4), 719–726.

(77) Melamud, E.; Vastag, L.; Rabinowitz, J. D. *Anal. Chem.* **2010**, *82* (23), 9818–9826.

(78) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. *Nat. Methods* **2015**, *12* (6), 523–526.

(79) Uppal, K.; Soltow, Q. A.; Strobel, F. H.; Pittard, W. S.; Gernert, K. M.; Yu, T.; Jones, D. P. *BMC Bioinformatics* **2013**, *14* (1), 15.

(80) Xia, J.; Sinelnikov, I. V.; Han, B.; Wishart, D. S. *Nucleic Acids Res.* **2015**, *43* (W1), W251–W257.

(81) Yore, M. M.; Syed, I.; Moraes-Vieira, P. M.; Zhang, T.; Herman, M. A.; Homan, E. A.; Patel, R. T.; Lee, J.; Chen, S.; Peroni, O. D.; Dhaneshwar, A. S.; Hammarstedt, A.; Smith, U.; McGraw, T. E.; Saghatelian, A.; Kahn, B. B. *Cell* **2014**, *159* (2), 318–332.

(82) Smith, R.; Ventura, D.; Prince, J. T. *Brief. Bioinform.* **2015**, *16* (1), 104–117.

(83) Rafiei, A.; Sleno, L. *Rapid Commun. Mass Spectrom.* **2015**, *29* (1), 119–127.

(84) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78* (3), 779–787.

(85) Treviño, V.; Yañez-Garza, I.-L.; Rodriguez-López, C. E.; Urrea-López, R.; Garza-Rodriguez, M.-L.; Barrera-Saldaña, H.-A.; Tamez-Peña, J. G.; Winkler, R.; Díaz de-la-Garza, R.-I. *J. Mass Spectrom.* **2015**, *50* (1), 165–174.

(86) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84* (1), 283–289.

(87) Masson, P.; Alves, A. C.; Ebbels, T. M. D.; Nicholson, J. K.; Want, E. J. *Anal. Chem.* **2010**, *82* (18), 7779–7786.

(88) Yanes, O.; Tautenhahn, R.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2011**, *83* (6), 2152–2161.

(89) Zhu, Z.-J.; Schultz, A. W.; Wang, J.; Johnson, C. H.; Yannone, S. M.; Patti, G. J.; Siuzdak, G. *Nat. Protoc.* **2013**, *8* (3), 451–460.

(90) Stupp, G. S.; Clendinen, C. S.; Ajredini, R.; Szewc, M. A.; Garrett, T.; Menger, R. F.; Yost, R. A.; Beecher, C.; Edison, A. S. *Anal. Chem.* **2013**, *85* (24), 11858–11865.

(91) Zamboni, N.; Saghatelian, A.; Patti, G. J. *Mol. Cell* **2015**, *58* (4), 699–706.

(92) Benton, H. P.; Ivanisevic, J.; Mahieu, N. G.; Kurczy, M. E.; Johnson, C. H.; Franco, L.; Rinehart, D.; Valentine, E.; Gowda, H.; Ubhi, B. K.; others; Tautenhahn, R.; Gieschen, A.; Fields, M. W.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2014**, *87* (2), 884–891.

(93) Benton, H. P.; Wong, D. M.; Trauger, S. A.; Siuzdak, G. *Anal. Chem.* **2008**, *80* (16), 6382–6389.

(94) Keller, B. O.; Sui, J.; Young, A. B.; Whittal, R. M. *Anal. Chim. Acta* **2008**, *627* (1), 71–81.

(95) Ipsen, A.; Want, E. J.; Lindon, J. C.; Ebbels, T. M. D. *Anal. Chem.* **2010**, *82* (5), 1766–1778.

(96) Patti, G. J.; Tautenhahn, R.; Johannsen, D.; Kalisiak, E.; Ravussin, E.; Brüning, J. C.; Dillin, A.; Siuzdak, G. *Metabolomics* **2014**, *10* (4), 737–743.

(97) Tautenhahn, R.; Patti, G. J.; Kalisiak, E.; Miyamoto, T.; Schmidt, M.; Lo, F. Y.; McBee, J.; Baliga, N. S.; Siuzdak, G. *Anal. Chem.* **2011**, *83* (3), 696–700.

(98) Patti, G. J.; Tautenhahn, R.; Siuzdak, G. *Nat. Protoc.* **2012**, *7* (3), 508–516.

(99) DeBerardinis, R. J.; Thompson, C. B. *Cell* **2012**, *148* (6), 1132–1144.

(100) Buescher, J. M.; Antoniewicz, M. R.; Boros, L. G.; Burgess, S. C.; Brunengraber, H.; Clish, C. B.; DeBerardinis, R. J.; Feron, O.; Frezza, C.; Ghesquiere, B.; Gottlieb, E.; Hiller, K.; Jones, R. G.; Kamphorst, J. J.; Kibbey, R. G.; Kimmelman, A. C.; Locasale, J. W.; Lunt, S. Y.; Maddocks, O. D.; Malloy, C.; Metallo, C. M.; Meuillet, E. J.; Munger, J.; Nöh, K.; Rabinowitz, J. D.; Ralser, M.; Sauer, U.; Stephanopoulos, G.; St-Pierre, J.; Tennant, D. A.; Wittmann, C.; Vander Heiden, M. G.; Vazquez, A.; Vousden, K.; Young, J. D.; Zamboni, N.; Fendt, S.-M. *Curr. Opin. Biotechnol.* **2015**, *34*, 189–201.

(101) Creek, D. J.; Chokkathukalam, A.; Jankevics, A.; Burgess, K. E. V.; Breitling, R.; Barrett, M. P. *Anal. Chem.* **2012**.

(102) Chen, Y.-J.; Huang, X.; Mahieu, N. G.; Cho, K.; Schaefer, J.; Patti, G. J. *Biochemistry* **2014**, *53* (29), 4755–4757.

(103) Huang, X.; Chen, Y.-J.; Cho, K.; Nikolskiy, I.; Crawford, P. A.; Patti, G. J. *Anal. Chem.* **2014**, *86* (3), 1632–1639.

(104) Creek, D. J.; Jankevics, A.; Breitling, R.; Watson, D. G.; Barrett, M. P.; Burgess, K. E. V. *Anal. Chem.* **2011**, *83* (22), 8703–8710.

(105) Libiseller, G.; Dvorzak, M.; Kleb, U.; Gander, E.; Eisenberg, T.; Madeo, F.; Neumann, S.; Trausinger, G.; Sinner, F.; Pieber, T.; Magnes, C. *BMC Bioinformatics* **2015**, *16* (1), 118.

(106) Patti, G. J.; Tautenhahn, R.; Rinehart, D.; Cho, K.; Shriver, L. P.; Manchester, M.; Nikolskiy, I.; Johnson, C. H.; Mahieu, N. G.; Siuzdak, G. *Anal. Chem.* **2012**, *85* (2), 798–804.

(107) Crutchfield, C. A.; Lu, W.; Melamud, E.; Rabinowitz, J. D. *Methods Enzymol.* **2010**, *470*, 393–426.

(108) Käll, L.; Vitek, O. *PLoS Comput. Biol.* **2011**, *7* (12), e1002277.

(109) Kele, M.; Guiochon, G. *J. Chromatogr. A* **2000**, *869* (1–2), 181–209.

(110) Buszewski, B.; Noga, S. *Anal. Bioanal. Chem.* **2012**, *402* (1), 231–247.

(111) Fuhrer, T.; Zamboni, N. *Curr. Opin. Biotechnol.* **2015**, *31*, 73–78.

(112) Cappadona, S.; Baker, P. R.; Cutillas, P. R.; Heck, A. J. R.; van Breukelen, B. *Amino Acids* **2012**, *43* (3), 1087–1108.

(113) Nikolskiy, I.; Mahieu, N. G.; Chen, Y.-J.; Tautenhahn, R.; Patti, G. J. *Anal. Chem.* **2013**, *85* (16), 7713–7719.

(114) Vandenbogaert, M.; Li-Thiao-Té, S.; Kaltenbach, H.-M.; Zhang, R.; Aittokallio, T.; Schwikowski, B. *Proteomics* **2008**, *8* (4), 650–672.

(115) Abate-Pella, D.; Freund, D. M.; Ma, Y.; Simón-Manso, Y.; Hollender, J.; Broeckling, C. D.; Huhman, D. V.; Krokhin, O. V.; Stoll, D. R.; Hegeman, A. D.; Kind, T.; Fiehn, O.; Schymanski, E. L.; Prenni, J. E.; Sumner, L. W.; Boswell, P. G. *J. Chromatogr. A* **2015**, *1412*, 43–51.

(116) Giorgino, T. *J. Stat. Softw.*

(117) Rabiner, L. R. *J. Acoust. Soc. Am.* **1978**, *63* (S1), S79.

(118) Pons, P.; Latapy, M. **2005**, 20.

(119) Wehrens, R.; Bloemberg, T. G.; Eilers, P. H. C. *Bioinformatics* **2015**, btv299-.

(120) Ivanisevic, J.; Zhu, Z.-J.; Plate, L.; Tautenhahn, R.; Chen, S.; O'Brien, P. J.; Johnson, C. H.; Marletta, M. A.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2013**, *85* (14), 6876–6884.

(121) R Core Team. Vienna, Austria 2014.

(122) Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. *Anal. Chem.* **2012**, *84* (11), 5035–5039.

(123) Åberg, K. M.; Alm, E.; Torgrip, R. J. O.; Aberg, K. M.; Alm, E.; Torgrip, R. J. O. *Anal. Bioanal. Chem.* **2009**, *394* (1), 151–162.

(124) Penzel, T.; Moody, G. B.; Mark, R. G.; Goldberger, A. L.; Peter, J. H. In *Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163)*; IEEE, 2000; pp 255–258.

(125) Goldberger, A. L.; Amaral, L. A. N.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; Mietus, J. E.; Moody, G. B.; Peng, C.-K.; Stanley, H. E. *Circulation* **2000**, *101* (23), e215–e220.

(126) Daly, R.; Rogers, S.; Wandy, J.; Jankevics, A.; Burgess, K. E. V.; Breitling, R. *Bioinformatics* **2014**, *30* (19), 2764–2771.

(127) Fernandez-Albert, F.; Llorach, R.; Andres-Lacueva, C.; Perera, A. *Bioinformatics* **2014**, *30* (13), 1937–1939.

(128) Zhang, W.; Chang, J.; Lei, Z.; Huhman, D.; Sumner, L. W.; Zhao, P. X. *Anal. Chem.* **2014**, *86* (13), 6245–6253.

(129) Wang, M.; Yu, G.; Mechref, Y.; Ressom, H. W. In *2013 IEEE International Conference on Bioinformatics and Biomedicine*; IEEE, 2013; pp 16–22.

(130) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. *BMC Bioinformatics* **2010**, *11* (1), 148.

(131) Böcker, S.; Dührkop, K. *J. Cheminform.* **2016**, *8* (1), 5.

(132) Zerck, A.; Nordhoff, E.; Resemann, A.; Mirgorodskaya, E.; Suckau, D.; Reinert, K.; Lehrach, H.; Gobom, J. *J. Proteome Res.* **2009**, *8* (7), 3239–3251.

(133) Hongbin Liu, †,§,∥; Rovshan G. Sadygov, †,§ and; John R. Yates, I. **2004**.

(134) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4* (3), 207–214.

(135) Draper, J.; Enot, D. P.; Parker, D.; Beckmann, M.; Snowdon, S.; Lin, W.; Zubair, H. *BMC Bioinformatics* **2009**, *10* (1), 227.

(136) Rogers, S.; Scheltema, R. A.; Girolami, M.; Breitling, R. *Bioinformatics* **2009**, *25* (4), 512–518.

(137) Junot, C.; Madalinski, G.; Tabet, J.-C. J.-C.; Ezan, E. *Analyst* **2010**, *135* (9), 2203–2219.

(138) Gross, R. W.; Han, X. **2009**, *297* (2), E297–E303.

(139) Dang, L.; White, D. W.; Gross, S.; Bennett, B. D.; Bittinger, M. A.; Driggers, E. M.; Fantin, V. R.; Jang, H. G.; Jin, S.; Keenan, M. C.; Marks, K. M.; Prins, R. M.; Ward, P. S.; Yen, K. E.; Liau, L. M.; Rabinowitz, J. D.; Cantley, L. C.; Thompson, C. B.; Vander Heiden, M. G.; Su, S. M. *Nature*.

(140) Ward, P. S.; Patel, J.; Wise, D. R.; Abdel-Wahab, O.; Bennett, B. D.; Coller, H. A.; Cross, J. R.; Fantin, V. R.; Hedvat, C. V.; Perl, A. E.; Rabinowitz, J. D.; Carroll, M.; Su, S. M.; Sharp, K. A.; Levine, R. L.; Thompson, C. B. *Cancer Cell* **2010**, *17* (3), 225–234.

(141) Xu, W.; Yang, H.; Liu, Y.; Yang, Y.; Wang, P. P.; Kim, S.-H.; Ito, S.; Yang, C.; Wang, P. P.; Xiao, M.-T.; Liu, L.; Jiang, W.; Liu, J.; Zhang, J.; Wang, B.; Frye, S.; Zhang, Y.; Xu, Y.; Lei, Q.; Guan, K.-L.; Zhao, S.; Xiong, Y. *Cancer Cell* **2011**, *19* (1), 17–30.

(142) Huang, X.; Chen, Y.-J.; Cho, K.; Nikolskiy, I.; Crawford, P. A.; Patti, G. J. **2014**.

(143) Gelman, S. J.; Mahieu, N. G.; Cho, K.; Llufrio, E. M.; Wencewicz, T. A.; Patti, G. J. *Cancer Metab.* **2015**, *3* (1), 1.

(144) Pape, J.; Vikse, K. L.; Janusson, E.; Taylor, N.; McIndoe, J. S. *Int. J. Mass Spectrom.* **2014**, *373*, 66–71.

(145) Shen, H.; Duhrkop, K.; Bocker, S.; Rousu, J.; Dührkop, K.; Böcker, S.; Rousu, J. *Bioinformatics* **2014**, *30* (12), i157-64.

(146) Patti, G. J.; Yanes, O.; Shriver, L. P.; Courade, J.-P.; Tautenhahn, R.; Manchester, M.; Siuzdak, G. *Nat. Chem. Biol.* **2012**, *8* (3), 232–234.

(147) Khan, A. P.; Rajendiran, T. M.; Ateeq, B.; Asangani, I. A.; Athanikar, J. N.; Yocum, A. K.; Mehra, R.; Siddiqui, J.; Palapattu, G.; Wei, J. T.; Michailidis, G.; Sreekumar, A.; Chinnaiyan, A. M. *Neoplasia* **2013**, *15* (5), 491–501.

(148) Tang, W. H. W.; Wang, Z.; Levison, B. S.; Koeth, R. A.; Britt, E. B.; Fu, X.; Wu, Y.; Hazen, S. L. *N. Engl. J. Med.* **2013**, *368* (17), 1575–1584.

(149) Jain, M.; Nilsson, R.; Sharma, S.; Madhusudhan, N.; Kitami, T.; Souza, A. L.; Kafri, R.; Kirschner, M. W.; Clish, C. B.; Mootha, V. K. *Science (80-.).* **2012**, *336* (6084), 1040–1044.

(150) Bajad, S. U.; Lu, W.; Kimball, E. H.; Yuan, J.; Peterson, C.; Rabinowitz, J. D. *J. Chromatogr. A* **2006**, *1125* (1), 76–88.

(151) Lu, W.; Bennett, B. D.; Rabinowitz, J. D. *J. Chromatogr. B* **2008**, *871* (2), 236–242.

(152) Zelena, E.; Dunn, W. B.; Broadhurst, D.; Francis-McIntyre, S.; Carroll, K. M.; Begley, P.; O'Hagan, S.; Knowles, J. D.; Halsall, A.; Wilson, I. D.; Kell, D. B. *Anal. Chem.* **2009**, *81* (4), 1357–1364.

(153) Geier, F. M.; Want, E. J.; Leroi, A. M.; Bundy, J. G. *Anal. Chem.* **2011**, *83* (10), 3730–3736.

(154) Nordström, A.; Want, E.; Northen, T.; Lehtiö, J.; Siuzdak, G. *Anal. Chem.* **2008**, *80* (2), 421–429.

(155) Buescher, J. M.; Moco, S.; Sauer, U.; Zamboni, N. *Anal. Chem.* **2010**, *82* (11), 4403–4412.

(156) Lu, W.; Clasquin, M. F.; Melamud, E.; Amador-Noguez, D.; Caudy, A. A.; Rabinowitz, J. D. *Anal. Chem.* **2010**, *82* (8), 3212–3221.

(157) Chokkathukalam, A.; Jankevics, A.; Creek, D. J.; Achcar, F.; Barrett, M. P.; Breitling, R. *Bioinformatics* **2013**, *29* (2), 281–283.

(158) Mishur, R. J.; Rea, S. L. *Mass Spectrom. Rev.* **2012**, *31* (1), 70–95.

(159) Brown, M.; Dunn, W. B.; Dobson, P.; Patel, Y.; Winder, C. L.; Francis-McIntyre, S.; Begley, P.; Carroll, K.; Broadhurst, D.; Tseng, A.; Swainston, N.; Spasic, I.; Goodacre, R.; Kell, D. B. **2009**, *134* (7).

(160) Alonso, A.; Julia, A.; Beltran, A.; Vinaixa, M.; Diaz, M.; Ibanez, L.; Correig, X.; Marsal, S.; Julià, A.; Beltran, A.; Vinaixa, M.; Díaz, M.; Ibañez, L.; Correig, X.; Marsal, S. *Bioinformatics* **2011**, *27* (9), 1339–1340.

(161) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther. Drug Monit.* **2005**, *27* (6), 747–751.

(162) Guo, A. C.; Jewison, T.; Wilson, M.; Liu, Y.; Knox, C.; Djoumbou, Y.; Lo, P.; Mandal, R.; Krishnamurthy, R.; Wishart, D. S. *Nucleic Acids Res.* **2013**, *41* (Database issue), D625-630.

(163) Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. *Mol. Cell. proteomics MCP* **2002**, *1* (5), 376–386.

(164) Wiese, S.; Reidegeld, K. A.; Meyer, H. E.; Warscheid, B. *Proteomics* **2007**, *7* (3), 340–350.

(165) Hebert, A. S.; Merrill, A. E.; Bailey, D. J.; Still, A. J.; Westphall, M. S.; Strieter, E. R.; Pagliarini, D. J.; Coon, J. J. **2013**, *10* (4), 332–334.

(166) Birkemeyer, C.; Luedemann, A.; Wagner, C.; Erban, A.; Kopka, J. *Trends Biotechnol.* **2005**, *23* (1), 28–33.

(167) Mashego, M. R.; Wu, L.; Van Dam, J. C.; Ras, C.; Vinke, J. L.; Van Winden, W. A.; Van Gulik, W. M.; Heijnen, J. J. **2004**, *85* (6), 620–628.

(168) de Jong, F. A.; Beecher, C. *Bioanalysis* **2012**, *4* (18), 2303–2314.

(169) Broeckling, C. D.; Afsar, F. A.; Neumann, S.; Ben-Hur, A.; Prenni, J. E. *Anal. Chem.* **2014**, *86* (14), 6812–6817.

(170) Uppal, K.; Walker, D. I.; Liu, K.; Li, S.; Go, Y.-M.; Jones, D. P. *Chem. Res. Toxicol.* **2016**, *29* (12), 1956–1975.

(171) Xia, J.; Sinelnikov, I. V.; Han, B.; Wishart, D. S. *Nucleic Acids Res.* **2015**, *43* (W1), W251–W257.

(172) Stanstrup, J.; Gerlich, M.; Dragsted, L. O.; Neumann, S. *Anal. Bioanal. Chem.* **2013**, *405* (15), 5037–5048.

(173) Tautenhahn, R.; Cho, K.; Uritboonthai, W.; Zhu, Z.; Patti, G. J.; Siuzdak, G. *Nat. Biotechnol.* **2012**, *30* (9), 826–828.

(174) Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; Sumner, S.; Subramaniam, S. *Nucleic Acids Res.* **2016**, *44* (D1), D463–D470.

(175) Cho, K.; Mahieu, N.; Ivanisevic, J.; Uritboonthai, W.; Chen, Y.-J.; Siuzdak, G.; Patti, G. J. *Anal. Chem.* **2014**, *86* (19), 9358–9361.

(176) Tong, H.; Bell, D.; Tabei, K.; Siegel, M. M. *J. Am. Soc. Mass Spectrom.* **1999**, *10* (11), 1174–1187.

(177) Zhu, J.; Cole, R. B. *J. Am. Soc. Mass Spectrom.* **2000**, *11* (11), 932–941.

(178) Cajka, T.; Fiehn, O. *Anal. Chem.* **2016**, *88* (1), 524–545.

(179) Chen, Y.-J.; Mahieu, N. G.; Huang, X.; Singh, M.; Crawford, P. A.; Johnson, S. L.; Gross, R. W.; Schaefer, J.; Patti, G. J. *Nat. Chem. Biol.* **2016**, *12* (11), 937–943.

(180) Contrepois, K.; Jiang, L.; Snyder, M. *Mol. Cell. Proteomics* **2015**, *14* (6), 1684–1695.

(181) Vinayavekhin, N.; Saghatelian, A. In *Current Protocols in Molecular Biology*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2010; Vol. Chapter 30, p Unit 30.1.1-24.

(182) Wishart, D. S. *Nat. Rev. Drug Discov.* **2016**, *15* (7), 473–484.

(183) Milne, S. B.; Mathews, T. P.; Myers, D. S.; Ivanova, P. T.; Brown, H. A. *Biochemistry* **2013**, *52* (22), 3829–3840.

(184) Kessler, N.; Walter, F.; Persicke, M.; Albaum, S. P.; Kalinowski, J.; Goesmann, A.; Niehaus, K.; Nattkemper, T. W. *PLoS One* **2014**, *9* (11), e113909.

(185) Zeng, Z.; Liu, X.; Dai, W.; Yin, P.; Zhou, L.; Huang, Q.; Lin, X.; Xu, G. *Anal. Chem.* **2014**, *86* (8), 3793–3800.

(186) Barbazuk, W. B.; Korf, I.; Kadavi, C.; Heyen, J.; Tate, S.; Wun, E.; Bedell, J. A.; McPherson, J. D.; Johnson, S. L. *Genome Res.* **2000**, *10* (9), 1351–1358.

(187) Hillier, L. D.; Lennon, G.; Becker, M.; Bonaldo, M. F.; Chiapelli, B.; Chissoe, S.; Dietrich, N.; DuBuque, T.; Favello, A.; Gish, W.; Hawkins, M.; Hultman, M.; Kucaba, T.; Lacy, M.; Le, M.; Le, N.; Mardis, E.; Moore, B.; Morris, M.; Parsons, J.; Prange, C.; Rifkin, L.; Rohlfing, T.; Schellenberg, K.; Bento Soares, M.; Tan, F.; Thierry-Meg, J.; Trevaskis, E.; Underwood, K.; Wohldman, P.; Waterston, R.; Wilson, R.; Marra, M. *Genome Res.* **1996**, *6* (9), 807–828.

(188) Sajed, T.; Marcu, A.; Ramirez, M.; Pon, A.; Guo, A. C.; Knox, C.; Wilson, M.; Grant, J. R.; Djoumbou, Y.; Wishart, D. S. *Nucleic Acids Res.* **2016**, *44* (D1), D495–D501.

(189) Orla Teahan, †,‡; Simon Gamble, †; Elaine Holmes, ‡; Jonathan Waxman, †; Jeremy K. Nicholson, ‡; Charlotte Bevan, † and; Hector C. Keun*, ‡. **2006**.

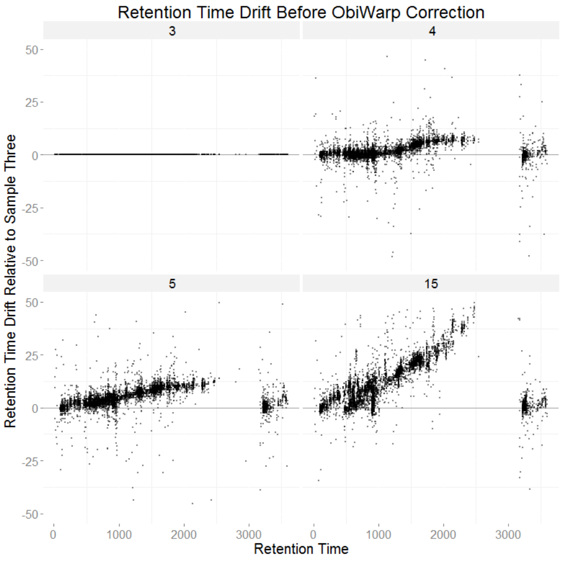
Appendix 1.

Warpgroup: increased precision of metabolomic data processing by consensus integration bound analysis

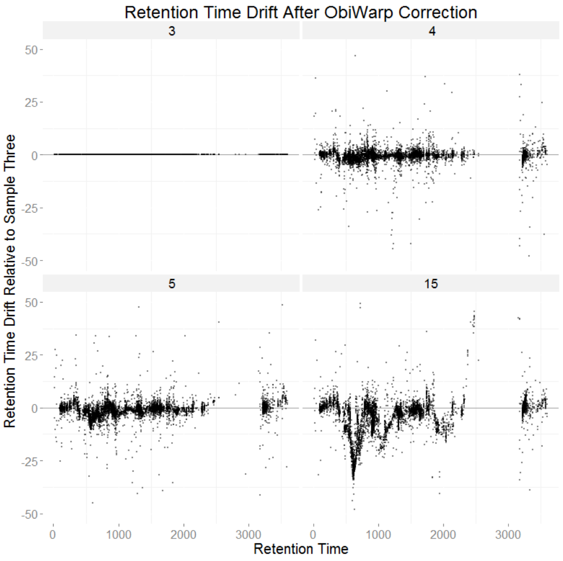
Appendix 1.1. The residual drift before and after retention time correction.

Retention time drift before (A) and after obiwarp (B) for sample numbers 3, 4, 5 and 15 from the HILIC dataset. Samples were aligned with sample 1 as the reference. It is clear that global retention time correction does shift the average drift towards zero (A and B). Importantly, even after retention time correction many peak retention times still present considerable drift. Figure C displays the change in residual drift for each peak, negative values represent a move further away from alignment while positive numbers are a shift towards alignment.

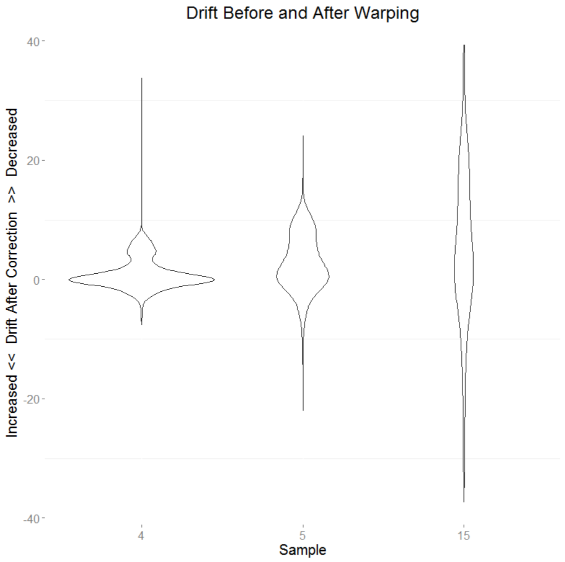
A.



B.



C.

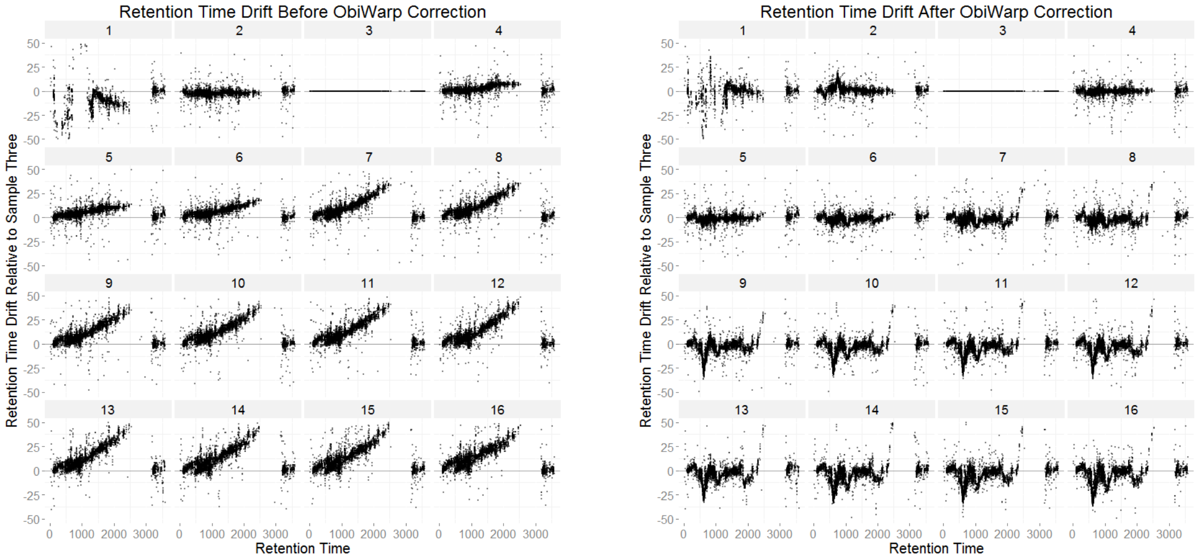


Appendix 1.2. The retention time drift of all samples in the data set.

Retention time drift of 16 samples including samples which were run to monitor equilibration of the LC system before (A) and after (B) Obiwarp alignment to sample 3.

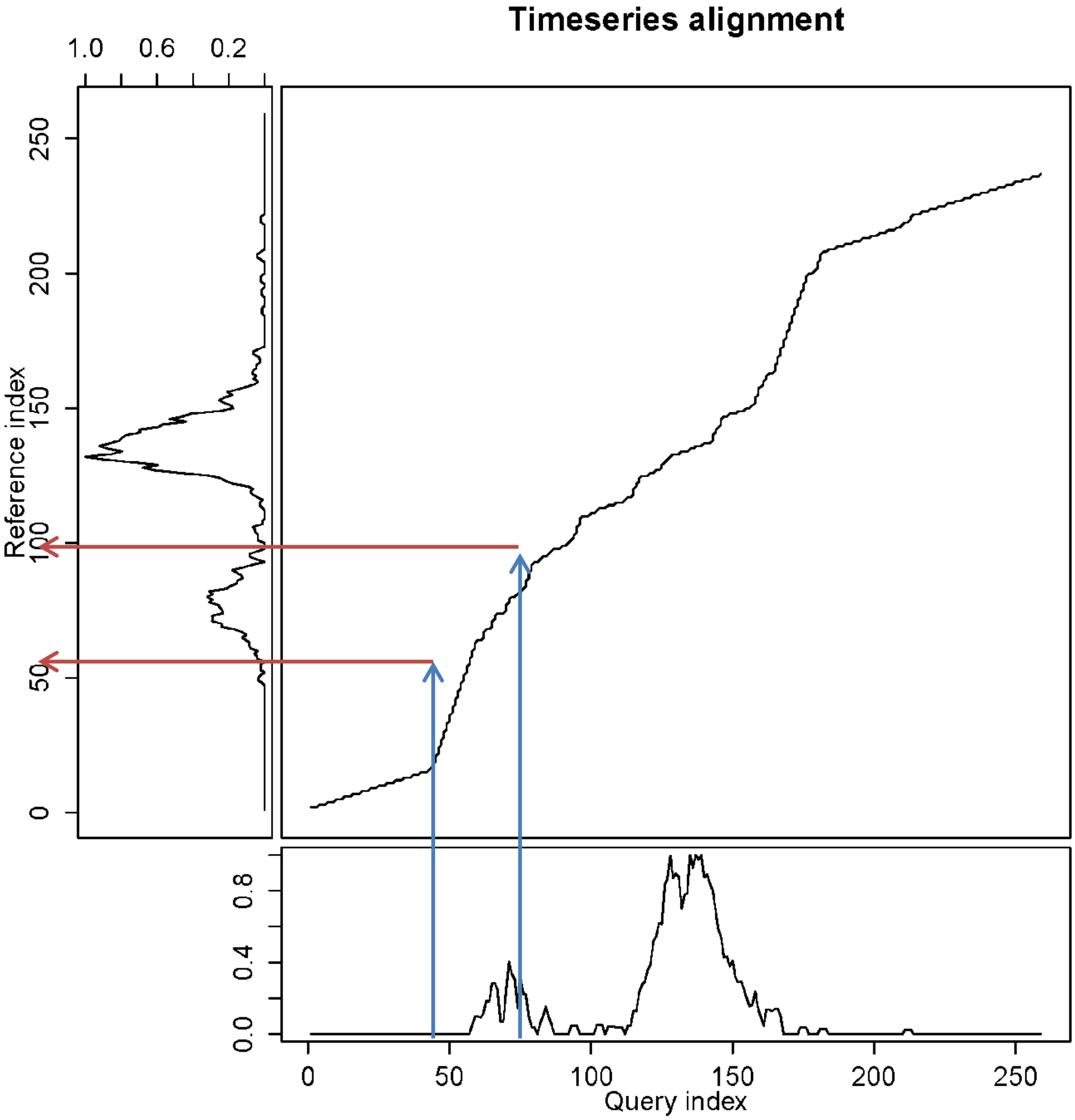
A.

B.



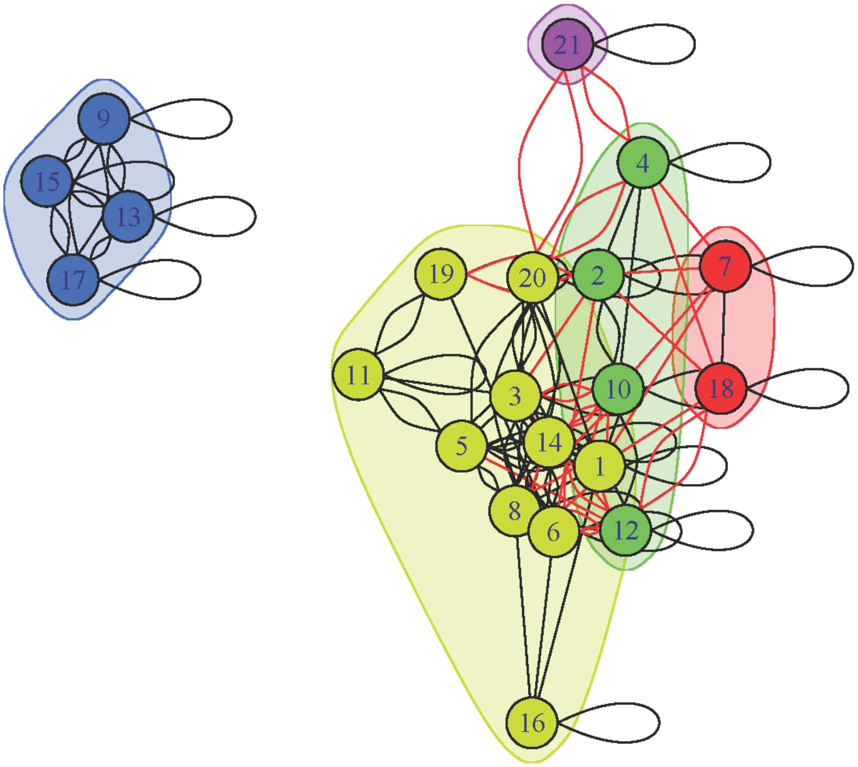
Appendix 1.3. Visualization of dynamic time warping inputs and output

A visualization of the input and output of dynamic time warping for a simple case. The two time series supplied as inputs are displayed on the X and Y axes. Dynamic time warping was performed on these, traces resulting the in "warp path" drawn as a line plot. This warp path relates the time domain of each series. Drawn arrows represent the projection of hypothetical peak bounds in the query series into the time domain of the reference series.



Appendix 1.4. Walktrap community detection on an example graph structure

In this graph structure peaks are drawn as nodes. Edges are drawn between peaks when they are determined to describe the same chromatographic region between samples. This is based on the agreement of transformed peak bounds across multiple sample pairs. Two edges are drawn per sample pair (one for A → B and a second for B → A) as DTW is not a symmetric technique. This graph structure is subjected to walktrap analysis to find communities of detected peaks which describe similar chromatographic regions.



Appendix 1.5. XCMS Integration

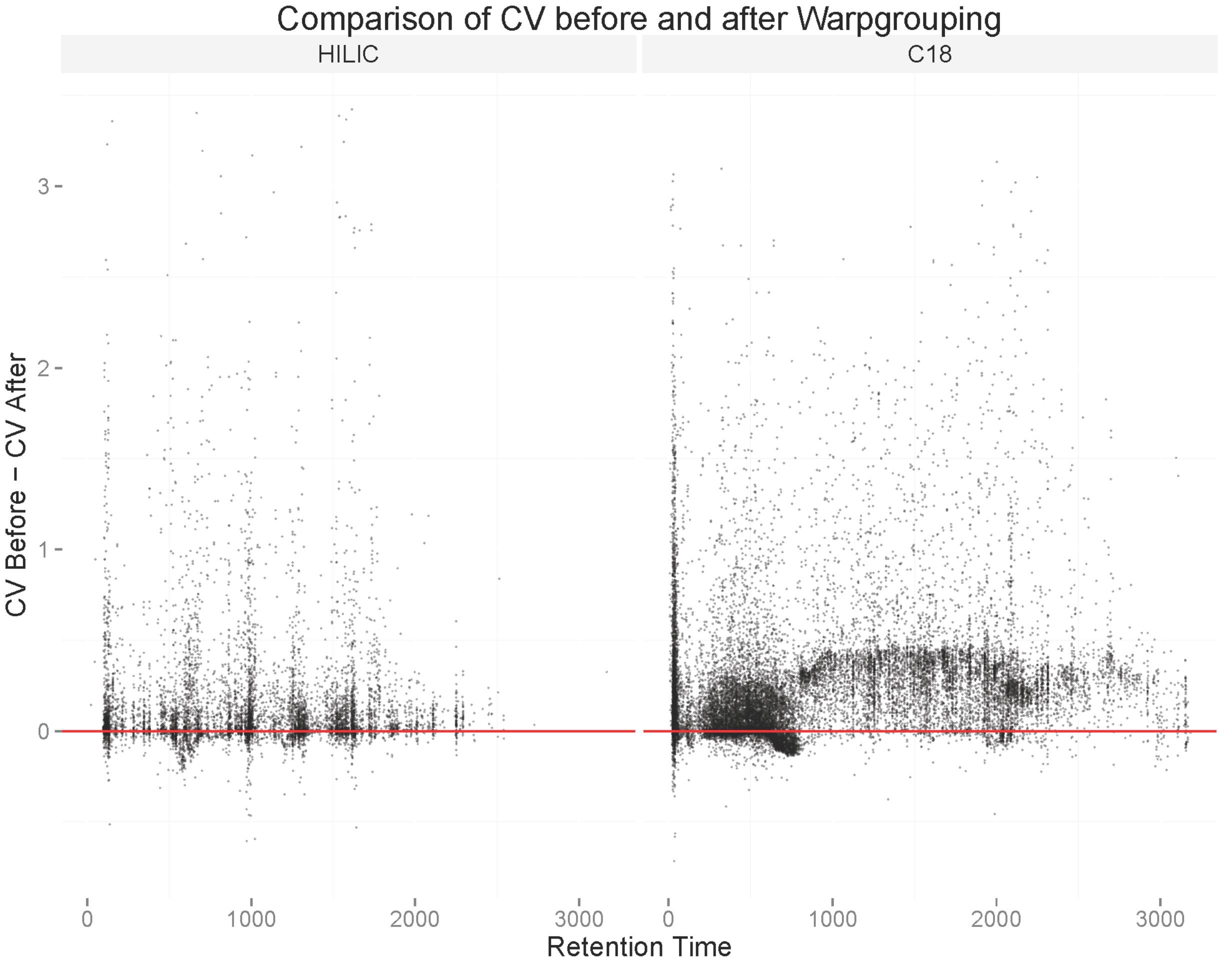
A prerequisite task to group.warpgroup() is the rough grouping of features between samples, such that all features which could possibly represent the same signal reside in a single group (as recorded in @grouped). This initial grouping should err on the side of inclusion, allowing the warpgroup algorithm to divide the rough groups into the appropriate sub-regions. This can be achieved with the default XCMS approach group.density(), but in cases of high retention time variance or small *m/z* drift a hard cutoff may be more appropriate.

The provided group.warpgroup() function iterates over each group in the xcmsSet and performs an initial setup before calling the algorithm. This setup includes the generation of an EIC trace for each sample based on the detected peak bounds and masses in that group. The EIC traces and detected peaks are then supplied to the warpgroup algorithm for processing. Returned groups are reintegrated and used to repopulate the xcmsSet.

The XMCS implementation uses the foreach package to handle parallelization. In the presence of a registered parallel backend (Eg. doRedis, doParallel) the warpgroup algorithm will be parallelized, each thread handling one warpgrouping. The generation of EIC matrices is performed in the parent thread to minimize the amount data to be transferred. When no parallel backend is registered the processing continues single threaded with a warning message.

Appendix 1.6. Pairwise comparison of CV before and after warpgroup

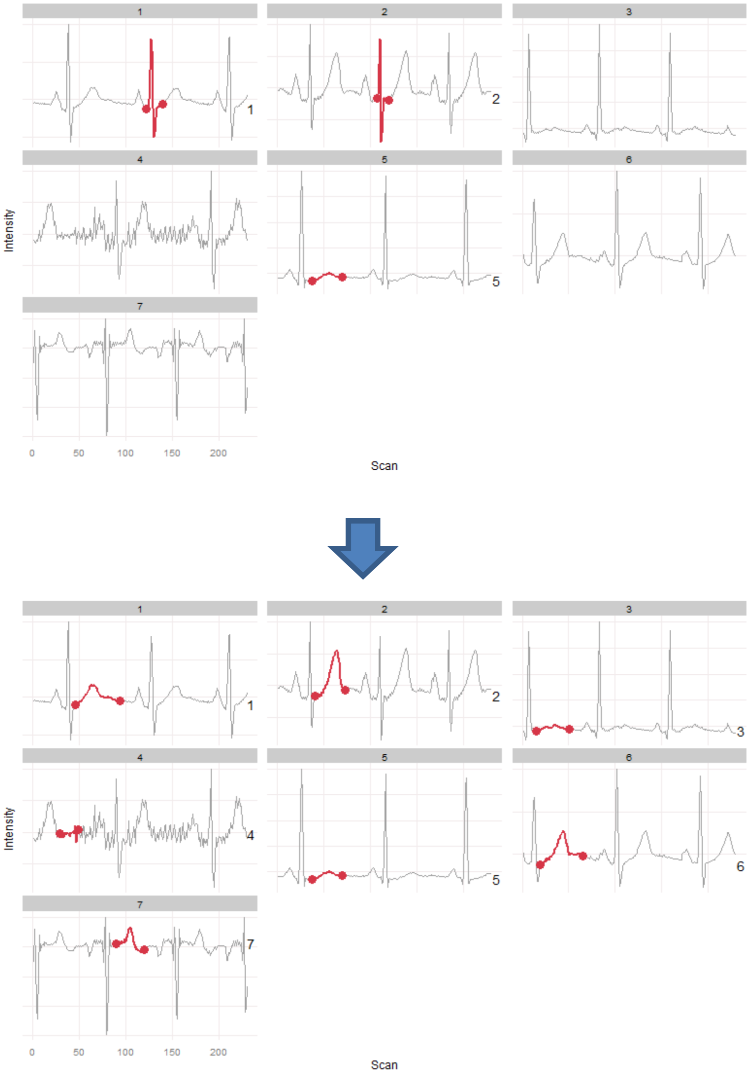
The CV of each group before and after warping was compared for each dataset by taking the difference ($CV_{before} – CV_{after}$). The red line indicates no change in CV while positive values indicate a decrease in CV. In most cases the CV was decreased by the warpgroup algorithm.

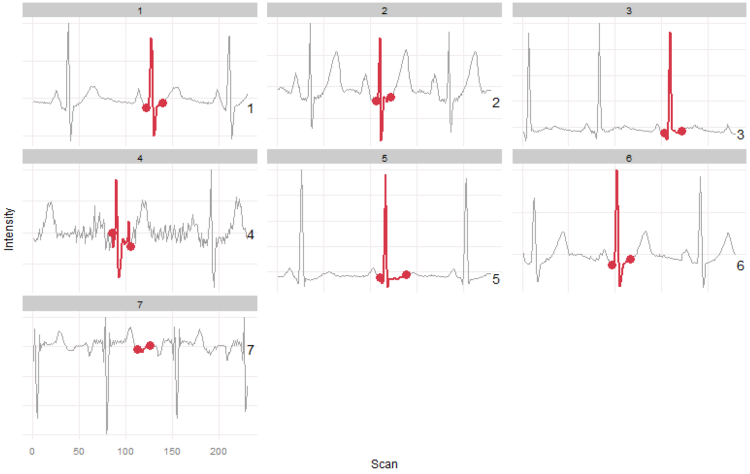


Appendix 1.7. An overview of differences between workflow outputs for the HILIC dataset with various group subsets

Subset of Peaks	All		Shared		Filtered 5 > n > 13
Workflow	**Traditional**	**Warpgroup**	**Traditional**	**Warpgroup**	**Warpgroup**
Mean CV	39%	24%	31%	18%	20%
90th Percentile CV	79%	50%	63%	33%	38%
Number of Groups	18,341	38,658	7,846	7,846	10,383

Appendix 1.8. Warpgroup of general timeseries data in the form of ecocardiograms




1
2
3
4
5
6
7
Intensity
Scan
0
50
100
150
200

Appendix 2.

Defining and Detecting Complex Peak Relationships in Mass Spectral Data: The mz.unity Algorithm

Appendix 2.1. Description of the combinatorial search problem for complex peak relationships.

Let O be the set of all observed <m,z>. Let M be a user supplied set of <m,z>. Find multisets R and S that satisfy:

$$R, S \subset O \cup M$$

$$R, S \not\subset M$$

$$R \cap S = \emptyset$$

$$\sum_{x \in R} x_m - \sum_{x \in S} x_m < \delta$$

$$\sum_{x \in R} x_z - \sum_{x \in S} x_z = 0$$

$$\delta = \frac{\varepsilon}{1E6} * \sum_{x \in R, S \cap O} \frac{x_m}{|x_z|}$$

Given a set of <m,z> pairs where m and z are positive or negative real numbers, find two sub-multisets whose summed mass is within error δ and whose summed charge is equal. Exclude pairs of multisets which share a member.

In broader terms, we have found a set of mass and charge transformations (corresponding to deprotonation, adduction, etc.) which convert a detected <m,z> value to a second <m, z> value. Thus this search ensures that the transformations of each <m,z> as described by members of each multiset results in an equal mass and charge.

Appendix 2.2. Features from negative mode included in the composite spectrum

mz	maxo	source
-146.046	2.76E+09	Psn
-128.035	3.52E+08	Psn
-102.056	3.12E+08	Psn
-147.049	1.64E+08	Psn
-662.102	1.37E+08	Psn
-231.098	1.15E+08	psn
-132.03	58812040	psn
-331.055	44957360	psn
-809.155	35751240	psn
-663.105	32867860	psn
-540.054	31911818	psn
-348.087	30048056	psn
-293.099	29273104	psn
-315.081	28147298	psn
-148.05	24049586	psn
-129.038	20168718	psn
-103.059	15625474	psn
-147.043	12138567	psn
-283.068	11558971	psn
-232.102	11044240	psn
-245.114	10539088	psn
-810.158	9959694	psn
-664.108	6937656	psn
-541.057	5288482	psn
-662.603	5039852	psn
-332.058	4757833	psn
-306.077	4592574	psn
-758.09	4585383	psn
-88.0403	4487963	psn
-168.028	3955738	psn
-357.087	3907644	psn
-129.019	3892937	psn
-333.053	3772248	psn
-294.102	3454026	psn
-320.011	3302731	psn
-847.11	3277035	psn
-184.001	3091660	psn
-349.09	3036758	psn
-316.084	2891467	psn
-133.033	2886159	psn
-148.052	2825622	psn
-811.16	2595229	psn
-358.118	2563522	psn
-228.049	2491125	psn
-244.023	2471267	psn
-146.025	2384161	psn
-416.108	2370798	psn
-993.655	2345890	psn
-831.137	2325020	psn
-760.078	2294203	psn
-353.037	2244784	psn
-130.039	2227558	psn
-874.173	2207934	psn
-146.104	2033520	psn
-151.062	1948270	psn
-378.152	1836020	psn
-994.157	1833876	psn
-317.039	1619828	psn
-146.071	1601470	psn
-146.02	1589578	psn
-129.032	1572900	psn
-244.119	1561877	psn
-104.06	1533376	psn
-103.053	1487027	psn

mz	maxo	source
-400.134	1216220	psn
-188.002	1182428	psn
-665.11	1176081	psn
-542.059	1156010	psn
-189.088	1126616	psn
-246.118	1123492	psn
-540.039	1110298	psn
-269.054	1099994	psn
-848.114	1090830	psn
-335.055	952685.8	psn
-206.012	938092.4	psn
-301.065	923233.9	psn
-516.063	921326.4	psn
-894.207	886170.4	psn
-500.09	873459.1	psn
-232.096	835217.1	psn
-345.178	834453.8	psn
-450.11	815158.2	psn
-456.246	809889.1	psn
-994.657	808162.4	psn
-881.311	808151.8	psn
-572.344	790534.1	psn
-134.034	789540.4	psn
-100.04	785416.2	psn
-187.109	781978.1	psn
-115.003	758465.8	psn
-456.166	758344.9	psn
-832.14	720119.6	psn
-333.059	683614.1	psn
-875.176	680699.4	psn
-687.107	673873.8	psn
-337.063	672261.3	psn
-598.36	661826.6	psn

-369.011	1482860	psn
-744.105	1443245	psn
-284.072	1440297	psn
-149.053	1427846	psn
-759.093	1331007	psn
-350.083	1323925	psn
-346.021	1306265	psn
-233.103	1248925	psn
-436.095	1239505	psn
-397.988	1222801	psn

Appendix 2.3. Features from positive mode included in the composite spectrum

mz	maxo	source									
148.06	1.81E+09	psp	170.0422	3883824	psp	465.3215	1884285	psp	173.0919	1058835	psp
130.0498	3.02E+08	psp	424.2588	3865671	psp	463.306	1875812	psp	425.2428	1038059	psp
233.2447	2.6E+08	psp	341.222	3865225	psp	230.1976	1848160	psp	196.1556	1034444	psp
664.1172	1.96E+08	psp	481.3163	3849028	psp	318.2982	1844858	psp	380.0776	1032436	psp
233.113	1.22E+08	psp	159.1603	3844951	psp	207.9982	1824690	psp	352.1902	1021926	psp
149.0633	91513336	psp	240.172	3813592	psp	438.2743	1823870	psp	246.2419	1021253	psp
102.0548	61472984	psp	243.2291	3693398	psp	308.1639	1808380	psp	393.2531	1016436	psp
123.0554	54190284	psp	261.276	3614619	psp	205.6433	1763577	psp	192.0241	1013983	psp
665.1202	50223776	psp	124.0587	3491213	psp	241.2133	1737384	psp	85.02834	1011348	psp
245.2447	48069948	psp	226.1565	3441997	psp	150.0665	1731420	psp	253.2133	1004417	psp
190.2025	28252376	psp	234.1538	3437938	psp	267.1369	1702268	psp	299.1636	1002699	psp
234.2479	26949514	psp	235.1617	3252422	psp	359.1025	1688196	psp	368.1851	1000992	psp
271.0688	21521252	psp	231.2293	3219105	psp	153.0405	1673078	psp	201.1903	991107.1	psp
276.2871	20748364	psp	383.2324	3199432	psp	346.1889	1665952	psp	289.2712	978187	psp
427.306	19735338	psp	702.0733	3124695	psp	483.2956	1661078	psp	350.1212	962097.6	psp
157.1447	18880884	psp	171.0762	3102233	psp	117.1263	1638529	psp	327.2175	961277.2	psp
257.2446	18655262	psp	248.1696	3073251	psp	369.2168	1625681	psp	234.1102	956079.1	psp
131.0531	17018116	psp	226.0376	2991801	psp	142.1339	1621513	psp	305.696	948127.1	psp
350.102	15073590	psp	283.132	2975036	psp	435.311	1610065	psp	260.6854	946392.5	psp
453.3216	15012519	psp	116.1184	2972269	psp	228.1819	1591366	psp	442.1668	945906.9	psp
247.1287	14912596	psp	493.3163	2965840	psp	232.1567	1547727	psp	323.215	943317.6	psp
150.0642	14645601	psp	269.2447	2952662	psp	432.1362	1543830	psp	422.2795	942465.8	psp
227.1644	13957159	psp	131.0338	2910072	psp	203.1866	1531236	psp	507.3319	936827.4	psp
191.1865	13730655	psp	103.0582	2891140	psp	273.0669	1513754	psp	225.1486	928799.9	psp
234.1163	12932350	psp	428.3099	2856149	psp	485.3112	1481398	psp	511.3267	928654.2	psp
265.0503	10756661	psp	451.306	2814969	psp	285.2398	1461184	psp	191.1024	925484.1	psp
666.1227	10674766	psp	477.3215	2653451	psp	222.1541	1426840	psp	996.1745	923613.8	psp
200.1869	10418991	psp	175.1553	2644349	psp	132.1131	1426403	psp	319.3297	923008.4	psp
295.1138	10269037	psp	686.0993	2641257	psp	667.1254	1425935	psp	412.2589	920321.3	psp
259.2604	8838102	psp	277.2905	2631271	psp	241.1616	1415203	psp	675.1086	918688	psp
304.1618	8177553	psp	450.2744	2600441	psp	440.2537	1412062	psp	289.2347	901264.7	psp
147.1604	7950535	psp	229.2137	2572000	psp	223.9721	1405962	psp	497.3113	896755.7	psp
214.2025	7811584	psp	234.2288	2559343	psp	304.2823	1398901	psp	176.0737	893954.2	psp
349.1181	7632175	psp	426.2745	2514821	psp	296.1167	1398888	psp	1014.651	888278.3	psp
314.2092	7174908	psp	379.2011	2492418	psp	271.224	1397299	psp	410.2432	887668.1	psp
434.2795	6942368	psp	99.0916	2480767	psp	235.2513	1393120	psp	427.2585	883956.6	psp
255.0949	6566525	psp	397.2115	2450742	psp	355.2012	1372936	psp	150.647	883463.1	psp
134.0447	6280206	psp	136.0618	2449351	psp	408.2639	1360118	psp	454.2693	877597.7	psp
251.0346	6058636	psp	191.2058	2432413	psp	158.1481	1357751	psp	495.2955	870804.2	psp
436.295	6045015	psp	601.3956	2407377	psp	496.3636	1350771	psp	256.1209	866752.1	psp
104.1181	5943378	psp	454.3255	2396598	psp	155.129	1348961	psp	489.3215	863220.1	psp
275.2555	5848210	psp	210.1713	2380743	psp	285.0829	1342086	psp	273.276	862540.6	psp
469.3164	5828522	psp	271.2604	2371772	psp	422.2432	1319220	psp	215.1024	861544.3	psp
261.2396	5793045	psp	495.3319	2369524	psp	353.2219	1308334	psp	392.2689	854450.4	psp
467.3007	5786659	psp	443.3009	2356544	psp	360.1324	1293417	psp	372.0838	849048	psp
234.242	5718033	psp	365.2218	2339415	psp	287.2555	1280357	psp	396.2639	846665.8	psp
410.2795	5668431	psp	323.0393	2321906	psp	335.7307	1272806	psp	267.0483	837923.8	psp
246.2481	5659415	psp	411.2635	2312104	psp	163.1552	1251915	psp	398.2431	837573.2	psp
149.0573	5610463	psp	233.1645	2304944	psp	333.0643	1236596	psp	703.077	835253.9	psp
202.2025	5554419	psp	455.3009	2298684	psp	235.1172	1228381	psp	440.29	828606.9	psp

288.2871	5536289	psp	326.1748	2271483	psp	192.19	1226547	psp	256.167	827247.1	psp
332.5624	5531230	psp	218.1975	2269794	psp	227.1979	1225753	psp	120.613	826322.6	psp
255.229	5406180	psp	258.2481	2242889	psp	129.1261	1218545	psp	151.0674	824307.5	psp
298.1796	5280998	psp	479.337	2236099	psp	310.1796	1211365	psp	342.1219	822430.4	psp
354.1695	5181576	psp	314.7111	2185420	psp	186.1713	1206633	psp	505.3164	815663.2	psp
269.1162	5132939	psp	594.4002	2179335	psp	436.2588	1201589	psp	286.1317	798008.8	psp
374.148	5068381	psp	388.1456	2168571	psp	143.1292	1200430	psp	556.3486	795780.7	psp
128.0707	5005582	psp	217.6432	2141469	psp	228.2184	1197925	psp	460.295	794064.2	psp
385.2481	4837042	psp	108.6127	2123574	psp	226.6486	1191097	psp	151.6549	793066.4	psp
384.2641	4743927	psp	272.0722	2101159	psp	131.0469	1184284	psp	239.1977	791010.1	psp
85.04767	4690118	psp	247.1618	2075201	psp	185.1761	1183860	psp	780.2363	789338.8	psp
218.6511	4430709	psp	138.6471	2071239	psp	183.1604	1172474	psp	1006.664	781422.4	psp
439.306	4337509	psp	388.0578	2069616	psp	483.332	1158942	psp	317.0835	781141.6	psp
320.1642	4316508	psp	132.0541	2067111	psp	153.0768	1157504	psp	242.1696	780253.4	psp
186.0162	4303773	psp	406.2483	2063521	psp	247.2239	1154023	psp	298.2719	780054.4	psp
322.1798	4289709	psp	468.2848	2022243	psp	995.6728	1127385	psp	148.0006	779222.2	psp
452.29	4236093	psp	351.1054	2003873	psp	285.0848	1110661	psp	406.1561	777100.1	psp
340.1903	4202912	psp	509.3111	1973029	psp	448.2951	1108775	psp	542.0681	770435.9	psp
239.1642	4089331	psp	226.2026	1969459	psp	351.2062	1105669	psp	121.6208	769528.5	psp
367.2375	4082911	psp	409.2954	1941976	psp	431.1876	1081482	psp	402.2345	767691.8	psp
198.1713	4067356	psp	100.0994	1925514	psp	395.2321	1076854	psp	418.2846	764103.7	psp
273.2396	4020422	psp	409.2477	1907389	psp	339.2062	1075096	psp	214.0916	763270.1	psp
216.2183	3967400	psp	322.207	1892691	psp	356.1375	1073961	psp	475.2136	759841.3	psp
248.2318	3908586	psp	248.1321	1892027	psp	269.1107	1062688	psp	304.0153	756906.7	psp
595.4037	753921.4	psp									
428.2903	752444.6	psp									
375.1518	751830	psp									
266.0537	749370.5	psp									
375.1989	749054.3	psp									
256.2324	740162.6	psp									
440.31	734644.9	psp									
335.2148	721214.6	psp									
352.0979	718792.9	psp									
411.2806	717596.1	psp									
687.1026	709065.9	psp									
341.1378	705273.1	psp									
152.1182	695237.9	psp									
411.2273	691263.7	psp									
148.0339	689500.3	psp									
264.1827	684786.1	psp									
203.206	678903.4	psp									
322.7088	663479.8	psp									
309.1479	649027.9	psp									
269.6907	638324.9	psp									
356.1851	623756.9	psp									
394.0932	618761.9	psp									
385.2679	615982.9	psp									
531.2981	614656.4	psp									
325.1431	606670.4	psp									
582.3642	606210.6	psp									
148.1637	593393	psp									
1015.153	591243.8	psp									
298.1513	589913.8	psp									
386.252	586079.9	psp									
314.1745	568435.3	psp									
271.1318	568325.4	psp									
481.2799	561573.8	psp									
238.1104	553831.4	psp									
404.1949	543094.6	psp									
189.1791	538244.8	psp									
235.245	531464.8	psp									
120.1133	530622.3	psp									
399.2272	530437.9	psp									
395.1958	520012.8	psp									
293.1164	515396.7	psp									
355.1735	509952.2	psp									
277.2846	505453.8	psp									
527.3216	500390	psp									
290.0722	491646	psp									

252.1722	487311.1	psp
270.1196	474292.3	psp
341.9714	469674.1	psp
212.6328	468559.9	psp

Appendix 2.4. Background peaks from postive mode

85.07597, 86.05996, 86.07934, 87.05958, 87.09155, 88.02144, 88.07563, 89.07086, 89.10725, 90.05492, 91.05417, 92.03683, 93.12071, 94.06508, 95.0603, 96.08069, 96.99513, 97.07597, 98.07123, 98.08307, 98.98, 99.05523, 99.09154, 100.07561, 100.08836, 100.09939, 100.11199, 101.05966, 101.07089, 101.079, 101.10728, 102.09128, 102.10251, 103.08651, 104.01614, 104.07051, 104.11812, 104.99225, 105.10219, 105.12099, 107.07019, 108.61268, 109.11416, 111.01075, 111.05527, 111.09161, 111.11676, 112.08948, 112.09863, 113.07085, 113.10723, 114.09157, 114.10279, 114.11087, 114.61299, 115.0212, 115.09201, 115.10702, 116.07081, 116.11841, 117.00044, 117.06609, 117.10243, 117.12156, 117.12624, 118.03207, 118.06534, 118.08644, 118.09768, 118.13401, 119.01608, 119.08167, 120.0113, 120.06568, 120.11322, 120.61293, 121.01186, 121.62073, 122.00814, 122.07141, 123.09179, 123.12619, 124.08704, 125.03632, 125.07104, 125.10743, 125.12361, 126.10266, 127.08668, 127.10709, 128.08193, 128.09011, 128.11828, 128.61031, 129.06721, 129.07548, 129.10231, 129.12098, 129.12615, 130.0863, 130.09756, 130.1339, 130.15906, 131.01594, 131.08153, 131.11858, 131.13906, 132.11315, 133.03156, 134.02723, 135.0108, 135.07242, 135.10157, 135.12607, 136.0118, 136.02156, 136.08691, 136.13386, 136.94014, 137.00786, 137.02641, 137.04574, 137.12346, 137.6393, 138.1025, 138.6471, 139.08654, 139.1229, 139.1485, 140.00183, 140.08176, 140.11816, 141.10215, 141.11338, 141.12131, 141.12599, 142.03375, 142.09748, 142.13386, 143.08149, 143.12002, 143.12912, 143.14067, 143.58723, 144.0475, 144.11311, 144.64713, 145.0315, 145.09713, 145.1178, 145.14376, 146.0268, 146.10048, 146.12877, 147.04714, 147.10153, 147.11277, 147.16038, 148.14423, 148.16268, 149.02329, 149.02641, 149.12245, 150.02785, 150.10252, 150.64706, 151.02325, 151.09657, 151.65489, 152.11815, 152.12807, 153.1134, 154.03181, 154.09741, 154.13378, 155.11813, 155.12903, 155.15419, 156.04746, 156.11312, 156.1495, 157.09714, 157.11936, 157.14472, 158.02675, 158.03935, 158.12874, 158.14175, 158.14875, 159.04713, 159.11274, 159.16035, 160.04238, 160.04885, 160.10796, 160.14435, 160.16328, 160.16815, 161.04393, 161.06277, 161.092, 161.176, 162.1236, 163.04201, 163.13277, 163.15523, 164.13632, 165.05087, 165.11229, 166.13327, 167.03699, 167.12907, 168.11308, 168.13219, 168.14949, 169.14469, 170.09636, 170.12872, 170.14839, 171.11272, 171.12396, 171.14912, 171.16031, 172.04236, 172.10797, 172.14434, 172.15241, 172.16302, 173.06283, 173.0808, 173.13967, 173.17604, 174.12368, 174.12769, 174.16006, 174.1794, 175.11884, 175.1553, 176.07364, 176.15865, 177.05769, 178.05941, 179.05017, 179.11939, 179.12908, 180.08658, 180.15934, 181.02837, 181.14467, 182.12883, 182.14002, 182.16519, 183.12404, 183.16042, 184.10809, 184.14445, 184.1572, 185.13968, 185.17607, 186.12371, 186.14282, 186.16008, 186.1713, 187.10771, 187.12666, 187.15532, 187.63268, 188.13934, 188.15886, 188.1757, 188.18692, 189.14256, 189.17097, 189.17909, 190.17403, 190.20248, 191.07326, 191.15016, 191.18648, 191.19955, 191.20588, 192.0729, 192.07663, 192.12287, 192.13817, 192.18369, 192.19, 192.63546, 193.07023, 193.07803, 193.14406, 194.08424, 194.14, 195.16039, 196.00386, 196.14419, 196.15638, 197.13006, 197.13962, 197.17604, 198.12368, 198.16005, 198.17127, 199.1553, 199.17315, 199.18045, 199.19166, 199.6326, 200.07372, 200.13934, 200.15872, 200.17577, 200.1869, 200.64022, 201.13444, 201.17093, 201.18341, 201.19028, 202.15464, 202.20253, 203.15027, 203.18662, 203.20595, 204.06863, 204.13407, 204.17066, 204.18185, 204.18999, 204.63545, 205.16596, 205.18971, 205.64328, 206.14122, 206.1975, 207.18149, 207.20089, 208.09998, 208.15572, 209.13001, 209.1396, 209.15261, 209.17605, 210.13803, 210.17129, 210.63581, 211.15529, 211.17424, 211.1916, 211.64323, 212.0948, 212.13945, 212.15023, 212.1869, 212.63283, 213.07879, 213.09428, 213.09817, 213.13451, 213.14873, 213.15972, 213.17084, 213.17876, 213.19025, 213.6406, 214.09161, 214.09928, 214.1565, 214.16874, 214.20251, 214.65799, 215.15217, 215.15837, 215.18661, 215.19963, 215.20594, 216.1427, 216.17067, 216.18184, 216.21823, 216.63547, 217.05013, 217.10706, 217.16596, 217.17918, 217.20226, 217.22087, 217.64327, 218.08436, 218.1396, 218.14574, 218.15914, 218.19751, 218.23386, 218.65107, 219.15206, 219.18127, 219.2009, 219.64064, 219.65049, 220.15656, 220.65807, 221.14953, 222.15401, 222.17124, 222.19245, 222.65578, 223.09661, 223.16716, 223.6432, 224.18692, 224.63293, 224.65101, 225.11225, 225.14906, 225.17085, 225.18178, 225.64072, 225.65032, 226.11058, 226.15659, 226.16604, 226.20263, 226.64862, 226.65824, 227.09452, 227.11292, 227.15022,

227.16439, 227.18729, 227.1979, 227.20615, 227.65944, 227.66625, 228.15422, 228.16232, 228.16944, 228.18196,
228.21841, 228.23212, 228.65962, 229.16609, 229.18622, 229.20254, 229.21365, 230.10557, 230.15002, 230.19766,
230.21678, 230.65119, 231.1491, 231.18175, 231.20077, 231.22928, 231.65925, 232.15681, 232.2133, 232.23184,
232.65825, 233.02438, 233.14719, 233.16457, 233.1987, 233.24464, 233.65661, 233.66631, 234.13359, 234.15387,
234.19217, 234.22883, 234.24205, 234.24782, 234.64771, 234.65561, 234.66968, 235.06063, 235.1516, 235.16193,
235.16631, 235.25076, 235.66326, 235.67739, 236.06246, 236.161, 236.16874, 236.17908, 236.18658, 237.02412,
237.05862, 237.1484, 237.18179, 238.11043, 238.15628, 238.20249, 238.6484, 238.65809, 239.16419, 239.18626,
239.19773, 239.6572, 239.66609, 240.08975, 240.12591, 240.15383, 240.16149, 240.17197, 240.18179, 240.21804,
240.61737, 240.65967, 240.67344, 241.16165, 241.16814, 241.17952, 241.21326, 241.22154, 241.66313, 241.67752,
242.1515, 242.16956, 242.19738, 242.21646, 242.23385, 242.66819, 243.15933, 243.18127, 243.19275, 243.20078,
243.2291, 244.12101, 244.15676, 244.17684, 244.21309, 244.23191, 245.12302, 245.16367, 245.1972, 245.20831,
245.24465, 246.07904, 246.15394, 246.1722, 246.19239, 246.22886, 246.24187, 246.24802, 246.65695, 246.67097,
247.16184, 247.17743, 247.21256, 247.2239, 247.2497, 247.26028, 247.66341, 247.67756, 248.15147, 248.1696,
248.17926, 248.208, 248.23172, 248.67112, 248.68535, 249.07634, 249.15932, 249.18051, 249.18775, 249.2395,
250.07828, 250.24198, 251.03464, 251.07009, 251.17476, 251.19762, 252.03428, 252.03791, 252.17217, 252.18178,
252.21818, 252.67116, 253.03279, 253.12126, 253.16164, 253.17849, 253.21334, 253.62562, 253.67786, 254.10541,
254.15149, 254.16949, 254.19757, 254.21688, 254.23396, 254.66789, 254.68532, 255.13691, 255.1592, 255.178,
255.18947, 255.22754, 255.66094, 255.67412, 256.12099, 256.139, 256.16712, 256.17692, 256.21303, 256.23087,
256.6827, 257.1337, 257.17442, 257.18369, 257.19763, 257.20824, 257.21646, 257.24454, 258.15903, 258.19237,
258.22875, 258.2421, 258.24801, 259.16665, 259.22411, 259.26029, 259.67765, 260.16949, 260.20795, 260.23153,
260.26289, 260.68543, 261.17771, 261.18764, 261.23959, 261.27593, 261.67401, 261.69311, 262.16712, 262.24226,
262.27937, 262.66465, 262.68286, 263.10568, 263.15769, 263.17417, 263.21891, 263.67206, 264.16438, 264.18269,
264.22663, 264.6814, 265.05028, 265.12126, 265.21334, 265.23449, 266.05218, 266.15301, 266.6855, 267.0489,
267.13697, 267.22803, 268.18746, 268.21301, 268.23164, 268.68285, 269.05054, 269.11622, 269.18161, 269.19864,
269.20832, 269.24466, 269.6907, 270.10021, 270.11834, 270.18269, 270.19228, 270.20799, 270.22869, 270.24805,
271.13185, 271.16881, 271.20036, 271.22372, 271.26031, 272.18013, 272.208, 272.22993, 272.26132, 273.19008,
273.23956, 273.27596, 274.22367, 274.24223, 274.27119, 275.10568, 275.25545, 275.27423, 275.29182, 275.69085,
276.19087, 276.25893, 276.28701, 276.68058, 277.12175, 277.17784, 277.19065, 277.19658, 277.23478, 277.2711,
277.28455, 277.29038, 277.68872, 278.04575, 278.20763, 278.29251, 279.13714, 279.15919, 279.18826, 279.2293,
279.6801, 280.13942, 280.18132, 281.13399, 281.15278, 281.20842, 282.1004, 282.1522, 282.15616, 282.22891,
283.13207, 283.14966, 283.2238, 283.26055, 284.13545, 284.20813, 284.2603, 285.1892, 285.23981, 285.27617,
286.13173, 286.19616, 286.22381, 286.24295, 286.27138, 287.21911, 287.25545, 288.19299, 288.20311, 288.23948,
288.25883, 288.28703, 289.23469, 289.27108, 289.29033, 290.07215, 290.23877, 290.26774, 291.25037, 291.27357,
292.19599, 292.2819, 292.68804, 293.1165, 293.18868, 293.28532, 293.66857, 295.13235, 295.22421, 295.26085,
296.11647, 296.1319, 296.13571, 296.16399, 297.14801, 297.23985, 297.27643, 298.15044, 298.17961, 298.22406,
298.24315, 298.27011, 299.16372, 299.18176, 299.25581, 300.17708, 300.19364, 300.25913, 300.2873, 300.6861,
301.2015, 301.27133, 301.29058, 301.69359, 301.70335, 302.18203, 302.20197, 302.26839, 302.30293, 302.67416,
303.25057, 303.30628, 304.25034, 304.2822, 304.31861, 304.68828, 305.66894, 305.69622, 306.14833, 306.19876,
306.67689, 306.68611, 307.16889, 307.20153, 307.24562, 307.28186, 307.70341, 308.09127, 308.16379, 308.18148,
308.19084, 309.14789, 309.16736, 309.1989, 309.70074, 310.17951, 310.20666, 310.70853, 311.16358, 311.18733,
311.25577, 312.15867, 312.19332, 312.28727, 313.14302, 313.20142, 313.2713, 313.69346, 313.70296, 314.17439,
314.18194, 314.2092, 314.26678, 314.30291, 314.67399, 314.70108, 314.71105, 315.19019, 315.19959, 315.21086,
315.25084, 315.68219, 315.69183, 316.18951, 316.20709, 316.25407, 316.28247, 317.19663, 317.26647, 317.31416,
318.14863, 318.20389, 318.29814, 319.20172, 319.23441, 319.28218, 319.32977, 320.05693, 320.16417, 320.2096,
320.21849, 321.14835, 321.16778, 321.19018, 321.19918, 321.25022, 321.6912, 321.70101, 322.16119, 322.17983,
322.20698, 322.70887, 322.72279, 323.16393, 323.18237, 323.18767, 323.2142, 323.22119, 323.67975, 323.71705,
324.15917, 324.19546, 324.69705, 325.14315, 325.17973, 325.1945, 325.19885, 326.17475, 326.20948, 326.30326,
327.00869, 327.17815, 327.19026, 327.21742, 327.71923, 328.19805, 328.20704, 328.28252, 328.69965, 328.72269,
329.00663, 329.18776, 329.19673, 329.21426, 330.20296, 330.29815, 330.72025, 331.00281, 331.21213, 331.24577,
331.28213, 331.32975, 332.16414, 332.19279, 332.27733, 333.08845, 333.2978, 333.309, 334.07234, 334.1798, 334.20708,
334.72284, 335.09852, 335.18775, 335.32462, 335.71669, 335.73065, 336.1591, 336.19703, 336.23243, 337.18957,
337.20292, 337.21229, 338.17472, 338.22004, 339.15872, 339.17831, 339.20624, 340.15386, 340.17151, 340.19029,
341.19309, 341.22192, 342.20607, 342.22592, 342.29807, 342.72012, 343.02814, 343.21227, 343.72801, 344.19293,
344.22943, 344.24364, 344.73619, 345.20032, 345.25045, 345.34531, 346.17987, 346.18881, 347.19201, 348.19553,
348.73844, 349.0833, 349.18939, 350.17466, 350.211, 350.32421, 351.20642, 351.22876, 351.72668, 352.19029,

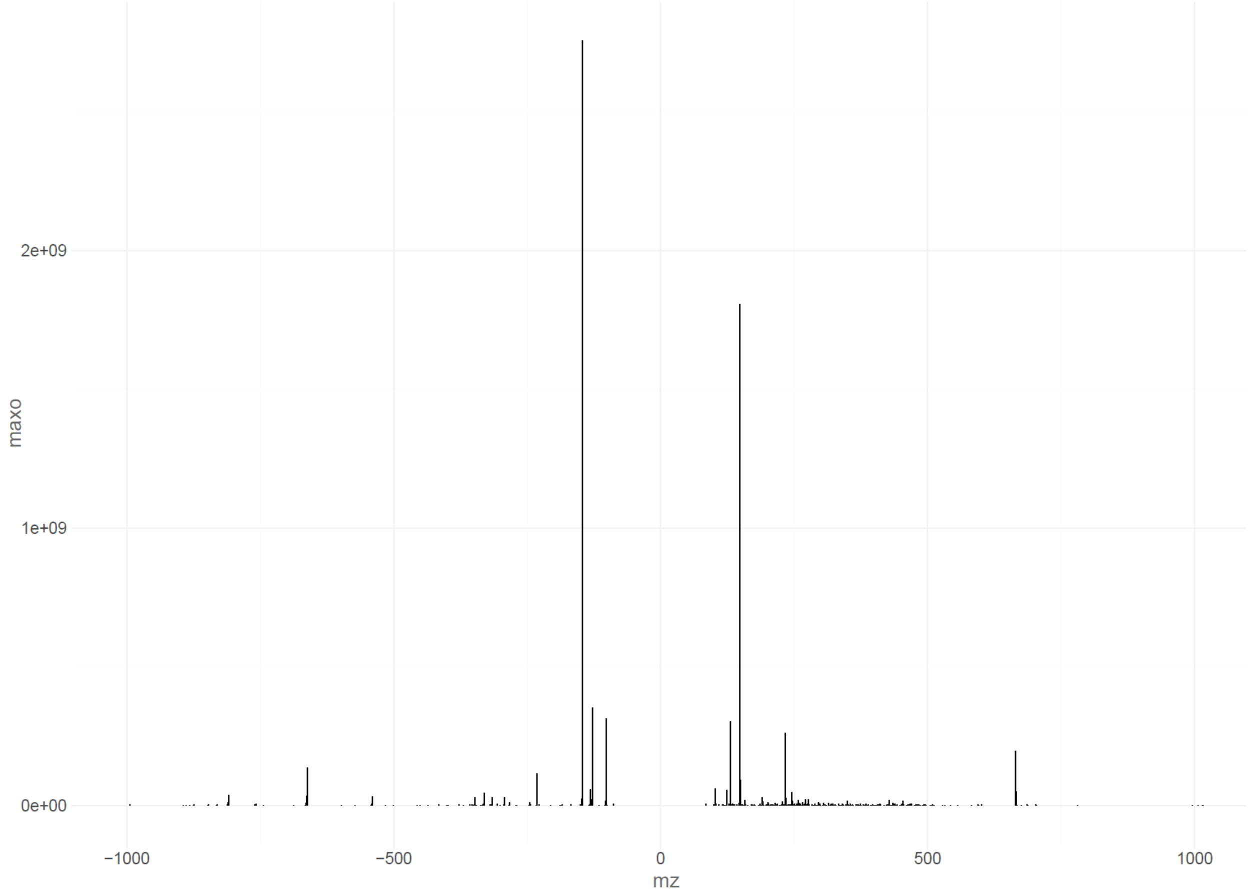
352.20907, 353.22188, 354.16954, 355.17206, 355.20119, 356.18535, 357.21686, 357.25161, 360.23312, 361.19064, 363.20625, 363.26064, 363.30825, 364.19033, 364.20989, 364.29225, 365.203, 365.22181, 366.20627, 366.23014, 366.72957, 367.12423, 367.20101, 367.23746, 368.22159, 368.24103, 369.21678, 373.24082, 373.25648, 373.74263, 375.20629, 375.23839, 375.30818, 377.2218, 379.20112, 379.2384, 379.73293, 381.21674, 381.23821, 382.2008, 382.22053, 382.24831, 383.23237, 384.23631, 384.26402, 385.24804, 385.26783, 386.18444, 386.2518, 387.18835, 387.24164, 388.1457, 388.24618, 388.74791, 389.14928, 389.22694, 389.24916, 389.72823, 390.20586, 391.20105, 391.23749, 391.28463, 392.23269, 392.24087, 392.28794, 393.21673, 393.25309, 394.24766, 395.19604, 395.23213, 396.22764, 396.24207, 396.26389, 397.21157, 397.24791, 397.75289, 398.23216, 398.24313, 399.2273, 400.25888, 401.24325, 401.25386, 401.75569, 402.23447, 402.73608, 404.23273, 405.21667, 405.31864, 406.15618, 406.24828, 407.23231, 407.25198, 407.3343, 408.23749, 408.2639, 409.2115, 409.24766, 409.26746, 409.76712, 410.23236, 410.2432, 410.25131, 410.25929, 410.27954, 411.22731, 411.2398, 411.26359, 411.28318, 411.74113, 412.25884, 412.26691, 412.27764, 413.24277, 413.26161, 418.24827, 419.2319, 419.33428, 420.22767, 420.26391, 421.21161, 421.24809, 421.2676, 422.24322, 422.27955, 423.22728, 423.24676, 423.26356, 423.27829, 424.25881, 425.24284, 425.2628, 425.29049, 426.24728, 426.27447, 427.2585, 427.27851, 427.30596, 428.29021, 428.30985, 429.22642, 429.30278, 429.31082, 430.23021, 431.18772, 432.26379, 433.24782, 433.26747, 434.243, 434.2794, 435.28324, 436.25882, 436.27694, 437.24273, 437.2793, 437.29881, 438.23799, 438.27432, 438.2918, 438.31071, 439.22204, 439.25847, 439.27804, 439.30598, 440.25366, 440.28999, 440.30989, 441.23763, 441.27397, 441.28809, 442.2692, 442.2931, 443.30094, 444.26372, 446.27942, 447.27787, 448.25868, 449.26204, 449.29015, 449.29839, 450.27437, 450.29362, 451.27811, 451.30594, 452.25366, 452.28996, 452.30965, 453.29065, 453.32153, 454.26929, 454.28784, 454.306, 454.32544, 455.24204, 455.30091, 455.31787, 455.32696, 455.33688, 456.30446, 456.32158, 457.20331, 458.2793, 459.27452, 459.29583, 460.25853, 461.29022, 461.29846, 462.14658, 462.27433, 462.29332, 462.31062, 463.26962, 463.30604, 464.25352, 464.28986, 464.30948, 465.2855, 465.32144, 466.26917, 466.28819, 466.30552, 467.30071, 468.30445, 468.3323, 469.2982, 469.31632, 470.32027, 470.34799, 471.29586, 471.31291, 471.32169, 471.3514, 472.26841, 473.2527, 473.29007, 474.2297, 474.27361, 474.31057, 475.2137, 475.2694, 475.2908, 475.30566, 475.7919, 476.27121, 476.28973, 476.309, 476.32616, 477.32141, 478.26913, 478.30556, 479.3007, 479.33703, 480.22821, 480.32115, 480.34088, 481.27995, 481.31629, 481.35256, 482.26392, 482.30048, 482.3202, 482.34784, 483.29564, 483.33191, 484.27959, 484.29799, 484.31619, 484.3358, 485.27629, 485.3112, 486.34274, 487.30603, 488.20891, 489.3214, 491.30058, 491.33696, 492.30386, 492.33185, 493.28016, 493.31629, 494.30048, 494.32014, 494.34787, 495.29589, 495.3319, 496.29817, 496.31614, 496.3359, 496.36353, 497.31127, 497.34764, 497.36724, 498.2841, 498.34273, 499.28905, 499.32682, 501.32141, 501.34254, 503.30054, 503.33694, 505.28239, 505.31629, 505.35266, 506.2437, 507.29554, 507.3318, 508.31601, 508.32754, 508.36339, 509.31109, 509.34742, 510.34264, 511.32669, 512.25818, 512.33022, 512.35825, 513.34237, 514.27939, 515.31041, 516.24001, 517.31606, 519.33175, 520.36342, 521.31104, 521.34741, 522.34262, 522.37908, 523.29034, 523.32663, 524.35818, 525.30601, 525.34231, 526.33748, 527.32162, 527.35798, 531.29806, 534.3424, 535.32666, 536.35808, 537.34231, 538.37389, 539.35803, 541.3261, 541.3372, 547.29117, 551.32165, 552.35308, 557.26712, 557.36905, 558.35315, 559.33731, 559.3563, 561.2985, 569.36916, 584.33241, 584.3689, 585.37255, 586.33019, 598.34821, 599.37975, 600.30936, 600.36385, 601.39536, 602.39869, 603.35673, 608.36893, 610.33068, 613.39597, 614.39928, 617.39086, 619.40672, 623.37891, 625.39606, 626.38008, 627.41167, 628.34136, 629.37316, 629.39054, 639.41166, 641.3909, 643.35214, 643.40674, 644.4098, 644.43824, 645.36803, 645.42242, 653.42756, 655.40663, 659.40167, 669.42247, 670.4539, 671.38357, 776.48833, 777.44678, 801.50109, 125.09311, 132.10192, 146.08126, 203.62768, 228.08982, 232.64885, 238.16619, 241.14351, 251.1615, 255.67603, 262.1905, 267.69313, 268.08485, 275.67271, 276.6987, 280.17136, 299.21931, 326.19219, 329.17434, 329.71163, 330.15849, 336.72069, 344.22006, 353.19859, 354.20606, 356.73598, 370.22028, 383.21353, 384.21719, 389.22104, 394.20079, 435.26357, 446.24197, 447.237, 459.28054, 460.29332, 468.28477, 469.28837, 470.30066, 473.27211, 480.30437, 488.26101, 488.29839, 501.30612, 506.31955, 516.31375, 539.3216, 553.33726, 194.11754, 210.13356, 285.14773, 294.14819, 307.13258, 320.71048, 321.16349, 327.25083, 437.26234, 455.27324, 602.37937, 336.22246, 449.19828, 283.1687, 350.12131,

Appendix 2.5. Background peaks from negative mode

-86.02469, -87.00343, -87.00867, -87.04507, -88.01205, -88.04032, -89.02437, -90.01876, -90.02772, -91.05041, -92.03252, -93.0345, -94.02978, -94.98056, -95.02504, -95.98015, -96.9764, -97.1843, -98.0247, -99.00797, -99.01997, -100.04037, -101.00628, -101.02429, -101.06069, -102.02006, -103.03999, -105.01917, -108.04543, -109.02929, -109.04053, -110.03587, -111.01982, -112.04024, -112.98545, -113.02427, -113.03553, -113.98874, -114.01943, -114.05585, -115.03989, -115.05112, -115.07627, -116.03591, -116.07154, -116.07965, -116.92839, -117.00118, -117.01918, -117.05557, -118.05084, -119.0168, -119.03478, -120.0179, -121.01363, -121.02934, -122.02458, -122.03272, -123.01176, -125.03549, -127.00114, -129.05563, -129.092, -130.08719, -130.99242, -131.03483, -131.07123, -132.01212, -132.99615, -133.01221, -133.05136, -134.02951, -134.0471, -134.97547, -135.02998, -136.93661, -136.99104, -137.02751, -137.03553, -138.01956, -139.04313, -140.9861, -141.01683, -141.06691, -141.09202, -142.02012, -142.05082, -142.97531, -143.02082, -143.03483, -143.04608, -143.07124, -143.10763, -144.03011, -144.11102, -144.99172, -145.05053, -145.06175, -145.08693, -145.09814, -146.02087, -146.0264, -148.9524, -149.0494, -151.04001, -152.03532, -154.06219, -154.94736, -155.00166, -155.10771, -155.94691, -156.00217, -156.94384, -156.95105, -156.99069, -156.99769, -157.00577, -157.03231, -157.06157, -157.08685, -157.1232, -157.99403, -158.04569, -158.09299, -158.1266, -158.98943, -159.07732, -159.10249, -159.99224, -160.04107, -161.02743, -161.04546, -163.06114, -165.02228, -165.04029, -165.979, -169.08703, -171.07752, -171.10274, -171.13907, -172.14253, -172.95785, -173.95753, -174.05915, -174.9543, -174.95981, -174.96711, -175.00939, -175.01731, -175.04301, -175.06107, -175.07071, -175.07691, -175.11895, -176.04273, -176.04619, -176.05593, -177.02231, -177.03986, -177.0474, -178.04317, -178.04919, -178.98346, -179.038, -179.05593, -180.0395, -180.05927, -180.9893, -180.99922, -181.07165, -183.03289, -185.05668, -185.07729, -185.09301, -185.11815, -185.1546, -187.04309, -187.0975, -188.03835, -189.02245, -189.05875, -189.08799, -190.01796, -190.05406, -192.03341, -192.06975, -193.0536, -193.0713, -194.05329, -194.05698, -194.06678, -194.08211, -194.98854, -195.03294, -195.05075, -195.05819, -196.99423, -197.01137, -197.04857, -197.99386, -199.13395, -199.1703, -200.17371, -201.02549, -202.05413, -203.05369, -203.05753, -204.03309, -204.05087, -205.05372, -205.15971, -206.05013, -206.0571, -207.03308, -207.0512, -207.06926, -208.0543, -208.06469, -208.07119, -208.93454, -209.06727, -209.07609, -210.04381, -210.06975, -210.9733, -211.06427, -211.13386, -212.06385, -212.06763, -213.06115, -213.186, -215.0328, -215.06739, -216.06963, -217.00295, -217.02977, -217.1013, -218.08584, -218.11449, -218.96332, -219.04441, -219.17532, -220.06463, -221.03029, -221.06356, -221.06777, -222.06144, -223.01999, -223.0643, -223.0823, -225.01701, -225.04366, -225.08002, -226.08247, -227.20173, -229.0304, -230.98613, -231.97848, -232.06469, -232.9247, -232.97913, -233.09839, -233.15469, -233.92427, -233.97953, -234.08034, -234.15807, -234.92133, -234.97716, -235.03943, -235.0644, -235.07781, -235.11199, -235.95363, -236.05968, -236.09602, -237.06157, -238.07533, -238.09329, -238.99375, -239.05944, -239.07732, -240.0807, -241.02088, -241.21743, -242.0519, -243.08058, -244.9855, -245.09639, -247.04109, -247.17036, -248.04282, -248.09606, -249.03817, -249.08015, -249.09137, -249.14974, -250.07565, -251.10697, -253.02069, -253.21744, -254.02174, -254.07427, -254.07842, -254.22083, -254.9612, -254.96809, -255.01766, -255.07209, -255.23308, -256.04958, -256.23648, -256.96118, -257.01576, -257.04655, -257.07002, -257.23982, -258.0717, -259.11186, -260.97409, -261.05594, -261.12776, -262.07538, -262.13042, -263.05382, -263.07809, -263.10705, -264.10945, -264.16072, -265.08649, -265.14807, -266.07057, -266.10669, -266.15146, -267.07243, -267.09078, -267.23316, -267.94553, -268.06653, -268.08679, -268.23652, -269.21248, -269.24885, -270.04732, -270.1017, -270.93524, -271.03145, -271.08563, -271.10453, -271.28679, -272.08778, -273.02044, -273.08129, -273.08973, -274.05255, -274.0602, -274.0872, -275.062, -275.10727, -276.04892, -276.05874, -276.09093, -276.10941, -277.05697, -277.12257, -277.18095, -278.15424, -279.13834, -280.08596, -280.98312, -281.06996, -281.11765, -281.24872, -282.06701, -282.25215, -283.2644, -284.08087, -284.26774, -285.047, -285.10121, -286.04808, -286.10335, -287.04319, -287.09973, -287.22015, -288.07586, -289.0781, -289.08631, -289.12237, -290.05224, -291.10201, -291.13837, -292.10163, -292.10538, -292.89169, -293.11773, -293.17765, -294.14929, -294.18017, -294.98231, -295.06749, -295.09927, -295.13312, -296.08079, -297.24355, -298.09645, -299.08043, -300.0826, -300.1386, -301.07062, -301.0961, -302.07329, -302.09839, -303.02103, -304.07074, -304.16995, -305.11795, -306.07672, -306.14923, -307.24649, -308.06295, -309.04698, -309.08325, -309.11255, -310.09946, -310.90227, -311.12828, -311.16887, -311.90183, -312.11231, -312.98764, -313.04196, -313.0963, -313.14389, -314.0987, -314.928, -315.08899, -315.1119, -315.25413, -316.10925, -317.09203, -318.14913, -319.13087, -320.09226, -321.09565, -321.11248, -321.19645, -323.24134, -325.01848, -325.05738, -325.18445, -326.09143, -327.12312, -329.23355, -330.90194, -330.95627, -330.9981, -331.08171, -331.99836, -333.2622, -334.14424, -335.12832, -335.14774, -336.15997, -337.14398, -337.19158, -339.03404, -339.07288, -339.12296, -339.15934, -340.10699, -340.16263, -341.11007, -341.13863, -342.12266, -343.08126, -344.11321, -344.97917, -345.01363, -345.06802, -345.08916, -347.21215, -347.24136, -348.15986, -349.06296, -349.10739, -349.19145, -349.25703, -350.10989, -350.28868, -351.09376, -351.12309, -352.03752, -352.08199, -352.15473, -353.08961, -353.13878, -354.12278, -354.14212, -355.15443, -356.11337, -356.18609, -358.12899, -361.01066, -361.1916,

-363.0788, -363.20726, -364.2389, -365.10937, -365.13865, -365.2226, -365.28819, -366.05303, -366.15903, -367.11801, -367.15427, -368.10189, -368.12021, -368.91205, -369.06256, -369.10447, -369.13355, -370.12876, -370.86888, -372.10812, -373.11142, -374.12378, -376.16619, -377.18648, -378.1704, -379.20204, -380.18607, -381.13375, -381.21771, -382.20174, -383.12017, -383.14937, -383.2334, -384.16984, -384.18108, -385.17951, -386.12395, -386.13115, -387.15579, -388.13152, -390.25458, -391.23858, -392.2334, -393.1523, -394.97591, -395.19685, -396.18086, -397.12857, -398.19652, -399.15547, -399.22816, -400.13929, -401.17105, -404.23363, -405.12561, -406.24931, -407.28095, -408.87925, -409.20138, -410.1967, -411.0074, -411.16263, -412.01076, -412.17595, -414.25445, -415.15063, -416.27024, -418.24949, -419.14138, -419.28114, -420.22878, -421.21262, -422.2444, -423.22844, -423.27605, -424.17609, -424.2601, -425.24413, -425.29172, -426.19177, -426.24761, -427.21183, -427.22302, -429.16902, -433.29656, -434.30039, -435.27589, -437.20749, -438.23916, -439.22315, -440.17078, -441.23883, -441.28643, -442.19777, -442.27047, -443.25448, -444.19098, -445.29664, -446.15226, -446.22906, -447.13629, -447.27601, -448.26002, -449.12792, -449.29164, -450.32321, -451.25969, -451.30724, -453.23898, -453.28656, -454.27065, -455.21828, -457.2815, -458.19278, -459.31254, -461.15203, -461.29174, -461.31234, -463.27133, -463.30744, -465.23864, -465.28629, -466.27032, -467.25432, -467.30207, -468.12112, -468.2496, -468.28597, -469.2336, -469.27009, -469.28982, -471.26045, -475.30713, -476.33871, -477.28648, -478.27046, -479.30208, -480.24969, -480.28613, -481.28133, -481.318, -482.26539, -483.29698, -484.28104, -485.23973, -485.31263, -486.31657, -487.23317, -487.30891, -488.28718, -489.29062, -491.30228, -492.28619, -493.28981, -493.31789, -494.26564, -494.30184, -495.2607, -495.3059, -496.2811, -497.27721, -497.31273, -498.2604, -498.2793, -498.31662, -499.29203, -499.32839, -501.30769, -502.31151, -506.30158, -507.29669, -508.28085, -509.31246, -510.29652, -511.32797, -512.27599, -512.33185, -513.27078, -513.30734, -513.33215, -514.30247, -516.28166, -517.30229, -520.31728, -521.3124, -522.29656, -523.29116, -523.32816, -524.27597, -524.31243, -525.30737, -525.34382, -526.29153, -526.31155, -527.32307, -528.28181, -528.32704, -529.26652, -529.30199, -529.35814, -530.29732, -533.31242, -535.32825, -537.30753, -537.34393, -538.29166, -538.32805, -538.34774, -539.28608, -539.32317, -540.31819, -540.32756, -541.30242, -541.33875, -542.29721, -543.31812, -544.34962, -549.3075, -551.32329, -553.30243, -553.33896, -554.34209, -554.37026, -555.31794, -555.37388, -556.31253, -556.33844, -565.30083, -565.33867, -567.31776, -568.34924, -569.33378, -571.34938, -580.38617, -581.3338, -581.37019, -599.38092, -601.34236, -607.38605, -611.38059, -615.37571, -617.39142, -625.39646, -627.35787, -641.39147, -642.42304, -643.40719, -659.40216, -662.37677, -671.4022, -675.39678, -685.41758, -686.42072, -687.37903, -688.39221, -701.4125, -773.4698, -775.43115, -102.04053, -201.11324, -209.98953, -226.047, -231.08067, -292.94611, -293.89124, -312.89892, -351.221, -357.09883, -499.25605, -554.29749, -252.09103, -310.95664

Appendix 2.6. Composite spectrum from 21-22 minutes of a HILIC analysis of E. coli extract



Appendix 2.7. List of granular formula used

Isotopes: M.iso

	m	z	d
C12-13	1.003355	0	1
N14-15	0.997035	0	1
O16-18	2.004245	0	1
S32-33	0.999387	0	1
S32-34	1.995796	0	1
Cl35-37	1.99705	0	1
Br79-81	1.997953	0	1
Si28-29	0.999568	0	1
Si28-30	1.996843	0	1
K41-39	1.998119	0	1

Charge Carriers: M.z

	z	m	d
H+	1	1.007825	0
Na+	1	22.98977	0
K+	1	38.96371	0
Cl-	-1	34.96885	0
Br-	-1	78.91834	0

Neutral Formula: M.n

	z	**m**	**d**
-H2O	0	-18.0106	1
-CO2	0	-43.9898	1
-NH3	0	-17.0265	1
+HCOOH	0	46.00548	1
+CH3COOH	0	60.02113	1
+CF3COOH	0	113.9929	1
+CH3CN	0	41.02655	1
+CH3OH	0	32.02622	1
-CO	0	-27.9949	1
+H3PO4	0	97.9769	1
+SiO3H2	0	77.97732	1
+SiO4H4	0	95.98789	1
+SiC2H6O	0	74.01879	1

Appendix 2.8. Mz.unity parameters used for annotation of the composite spectrum

The code used to annotate this spectrum can be found online in the repository referenced in the main text. A summary of the annotation workflow is listed here.

Find peaks with isotope support for higher charge states

All peaks in the spectrum were searched with proposal charge $z = 2 * sign(m/z)$ and mass $m = abs(m/z) * 2$. Any isotopes found support the higher charge state assignment. Search was performed on peaks of both polarities.

- M = M.iso, ppm = 1, BM.limits = cbind(M.min = c(1), M.max = c(1), B.n = c(1))

Find peaks with isotope support for charge state z = 1

All peaks in the spectrum were searched with proposal charge $z = 1 * sign(m/z)$ and mass $m = abs(m/z) * 2$. Any isotopes found support the higher charge state. Search was performed on peaks of both polarities.

- M = M.iso, ppm = 1, BM.limits = cbind(M.min = c(1), M.max = c(1), B.n = c(1))

Annotate simple relationships

Peaks with a isotope support for charge state 2 were included along with all peaks for charge state 1. Search was performed on peaks of both polarities.

- Cross Polarity: M = M.z, ppm = 10, BM.limits = cbind(M.min = c(2), M.max = c(2), B.n = c(1)
- Single Charge Carriers: M = M.z, ppm = 2, BM.limits = cbind(M.min = c(1), M.max = c(1), B.n = c(1)
- Neutral losses and adducts: M = M.n, ppm = 2, BM.limits = cbind(M.min = c(1), M.max = c(1), B.n = c(1)

Annotate analyte-analyte mers and distal fragments

Peaks with a isotope support for charge state 2 were included along with all peaks for charge state 1. Search was performed sequentially first for negative and then for positive mode.

- M = M.z (only H^+), ppm = 2, BM.limits = cbind(M.min = c(1), M.max = c(1), B.n = c(2)

Annotate analyte-analyte mers and distal fragments across polarities

Peaks with a isotope support for charge state 2 were included along with all peaks for charge state 1. Search was performed on peaks of both polarities.

- M = M.z (only H^+), ppm = 2, BM.limits = cbind(M.min = c(1), M.max = c(1), B.n = c(2)

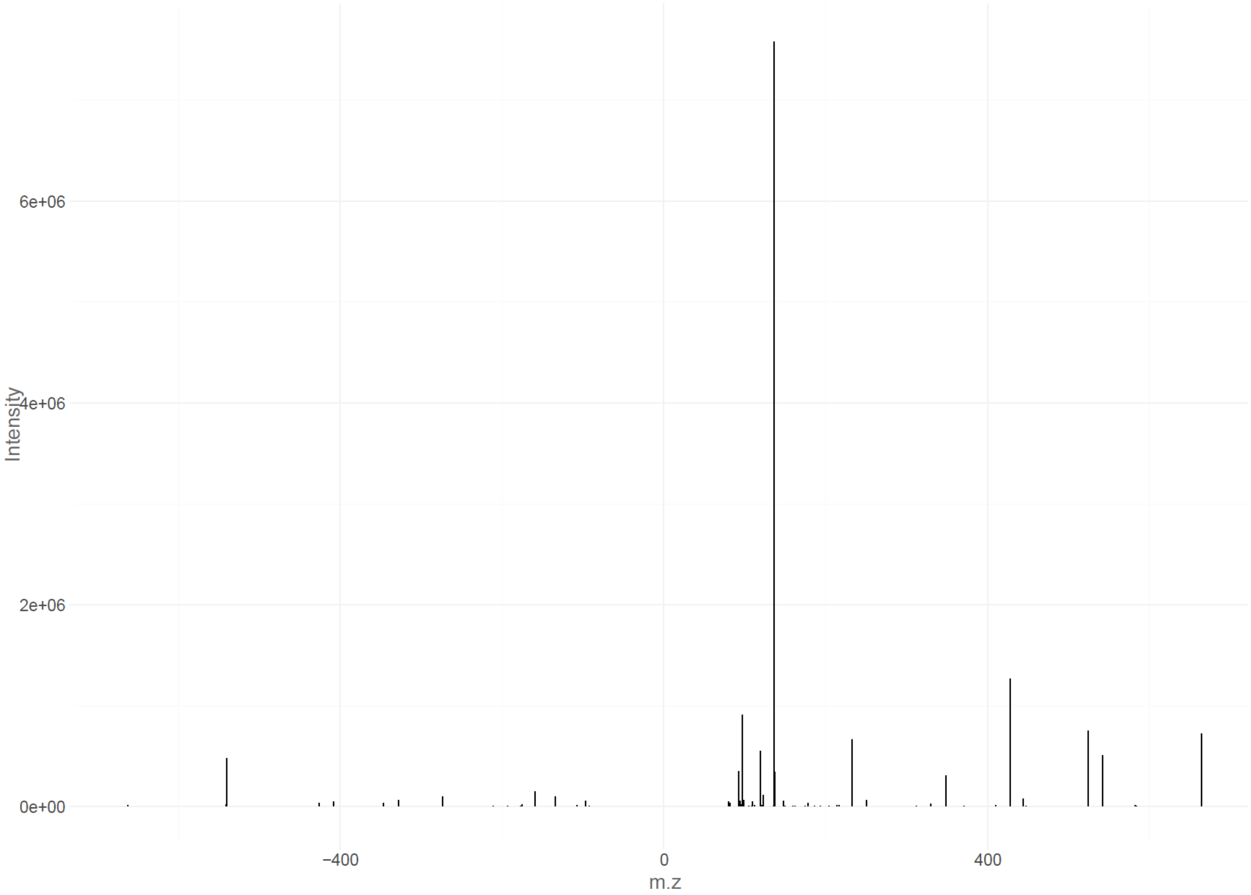
Annotate background mers

Peaks with a isotope support for charge state 2 were included along with all peaks for charge state 1. Features and mers were included in this search.

- M = M.z (only H^+), ppm = 2, BM.limits = cbind(M.min = c(1), M.max = c(1), B.n = c(2)

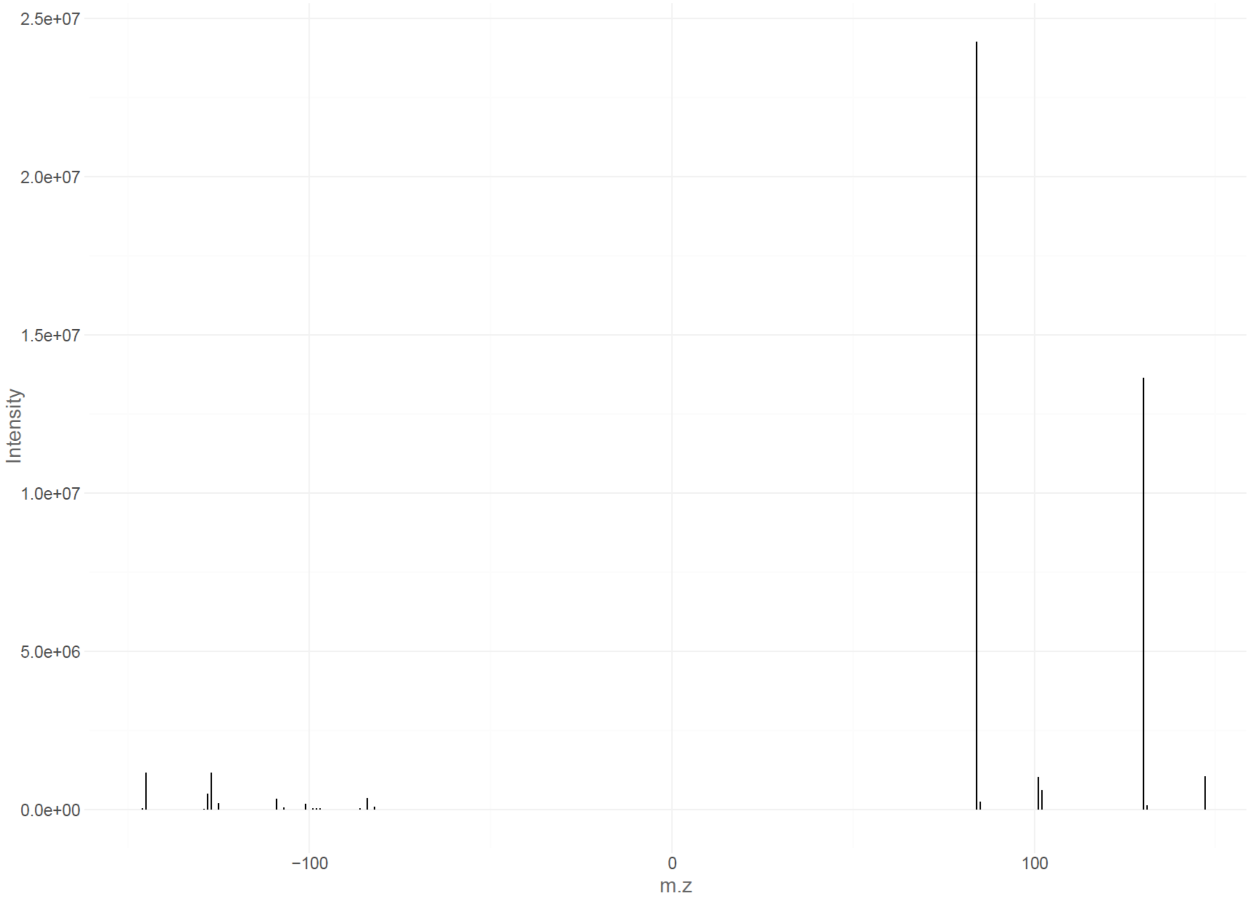
Appendix 2.9. Fragmentation spectrum of NAD

Spectra taken at collision energies 0, 10, 20, 40, 60, 90, and 120 were averaged from both positive and negative mode. The composite spectrum is shown below.



Appendix 2.10. Fragmentation spectrum of glutamate

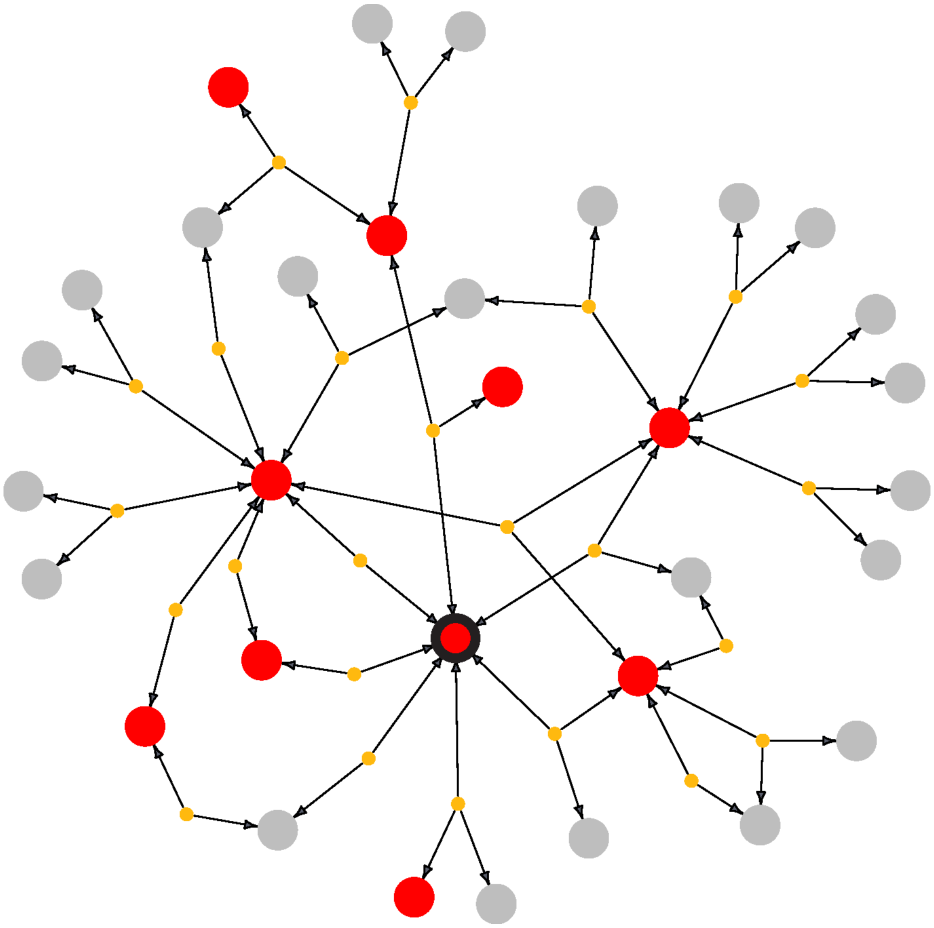
Spectra taken at collision energies 0, 10, 20, 40, 60, 90, and 120 were averaged from both positive and negative mode. The composite spectrum is shown below.



Appendix 2.11. Annotation of 2-hydroxyglutarate Metabolic Products

2-Hydroxyglutarate (2HG) corresponds to the node with a thick black border. Large nodes are features. Small nodes are relationships. Red nodes were detected as enriched by X^{13}CMS. Grey nodes were not detected as enriched by X^{13}CMS.

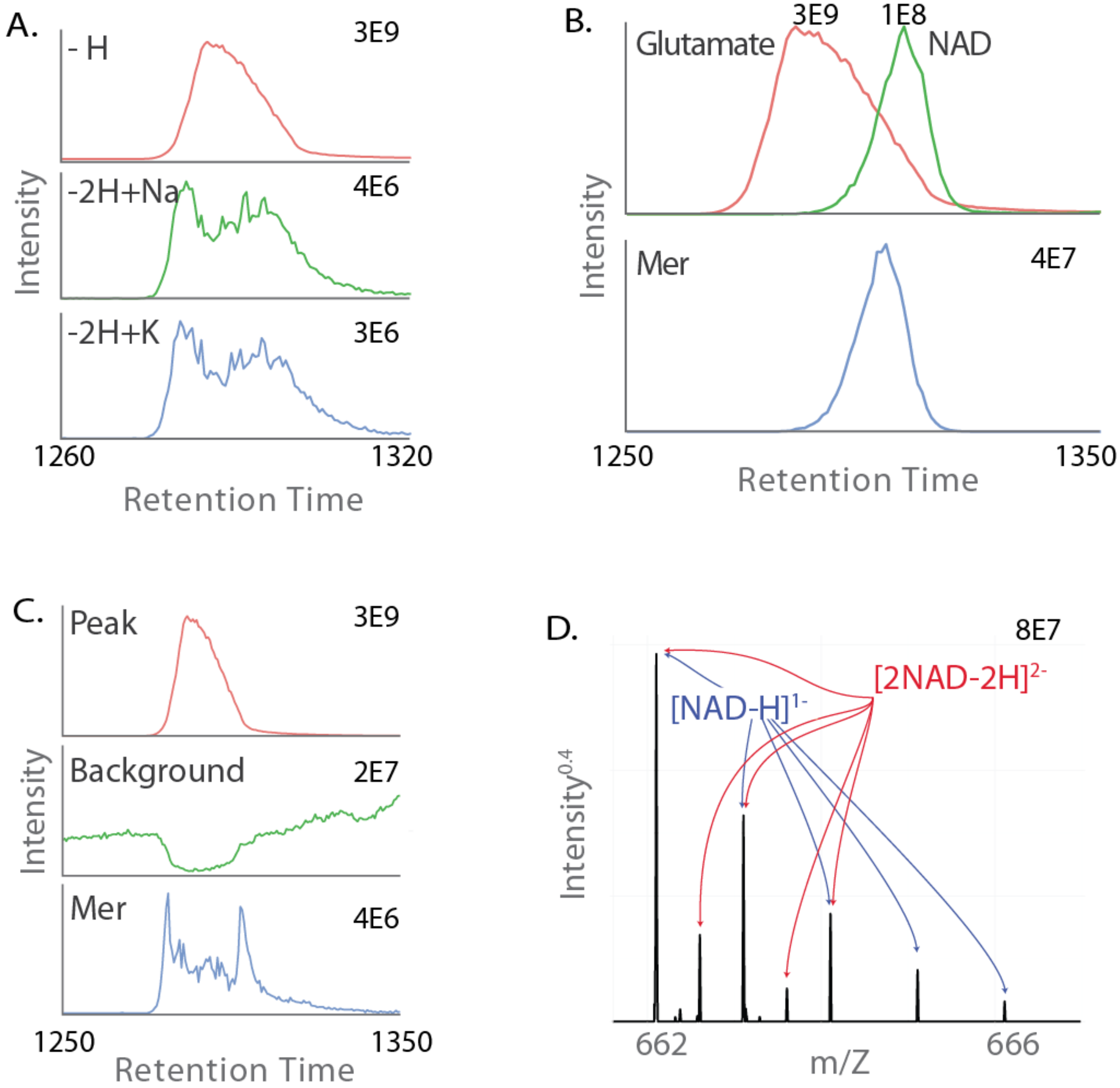
Analysis by mz.unity revealed that all enriched features were transformations of 2HG. This indicates that 2HG is not significantly metabolized in colorectal cancer cells.[14]



(14) Gelman, S. J.; Mahieu, N. G.; Cho, K.; Llufrio, E. M.; Wencewicz, T. A.; Patti, G. J. *Cancer Metab.* **2015**, *3* (1), 1.

Appendix 2.12. Intensitites, Masses, and Retention Times of Adducts

Maximum intensities from each chromatogram are inset. Retention times are in seconds. Masses from top to bottom and left to right: A, 146.0455, 168.0273, 184.0012; B, 146.0455, 662.1015, 809.1547; C, 146.0455, 98.0246, 436.0948.

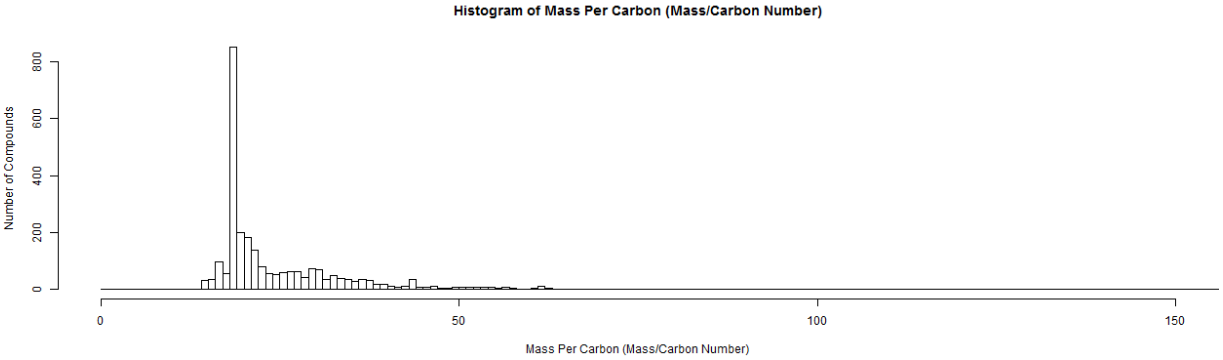


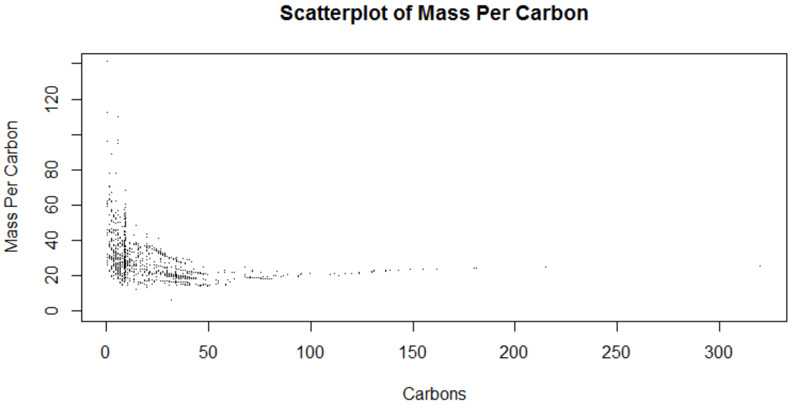
Appendix 3.

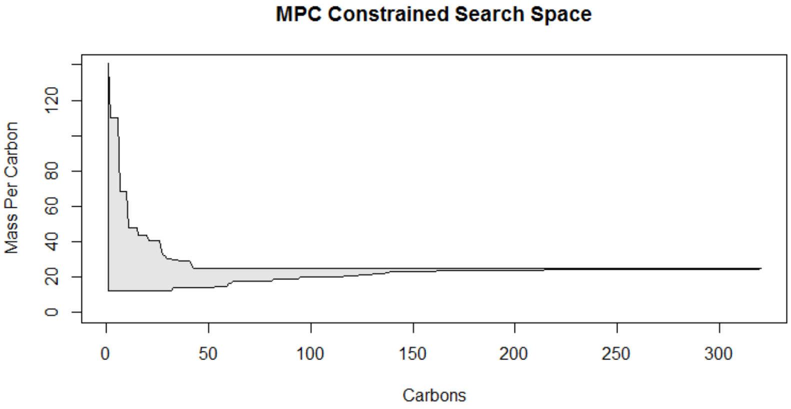
Credentialing Features: A Benchmarking Platform to Optimize Untargeted Metabolomic Methods

Appendix 3.1. Calculation of Mass Per Carbon (mpc) From ECMDB.

A histogram of mass in Daltons divided by carbon number (mpc) are shown below. The mass of a methylene (CH2, 14 Da) unit is a logical lower bound for mass per carbon. ECMDB contains four compounds which have an mpc lower than 14, all of which are more reduced (contain more rings and double bonds). An mpc of 141 is the largest in ECMDB and corresponds to carbamoyl phosphate. The most common mass per carbon is 18-19 Da/C with 850 compounds falling in this range. Based on the data, a carbon number dependent limit is placed on the mass range in which to search for isotopes. This is depicted in the lower plots.







Appendix 3.2. Suggested Parameters for Various Instrumentation Platforms.

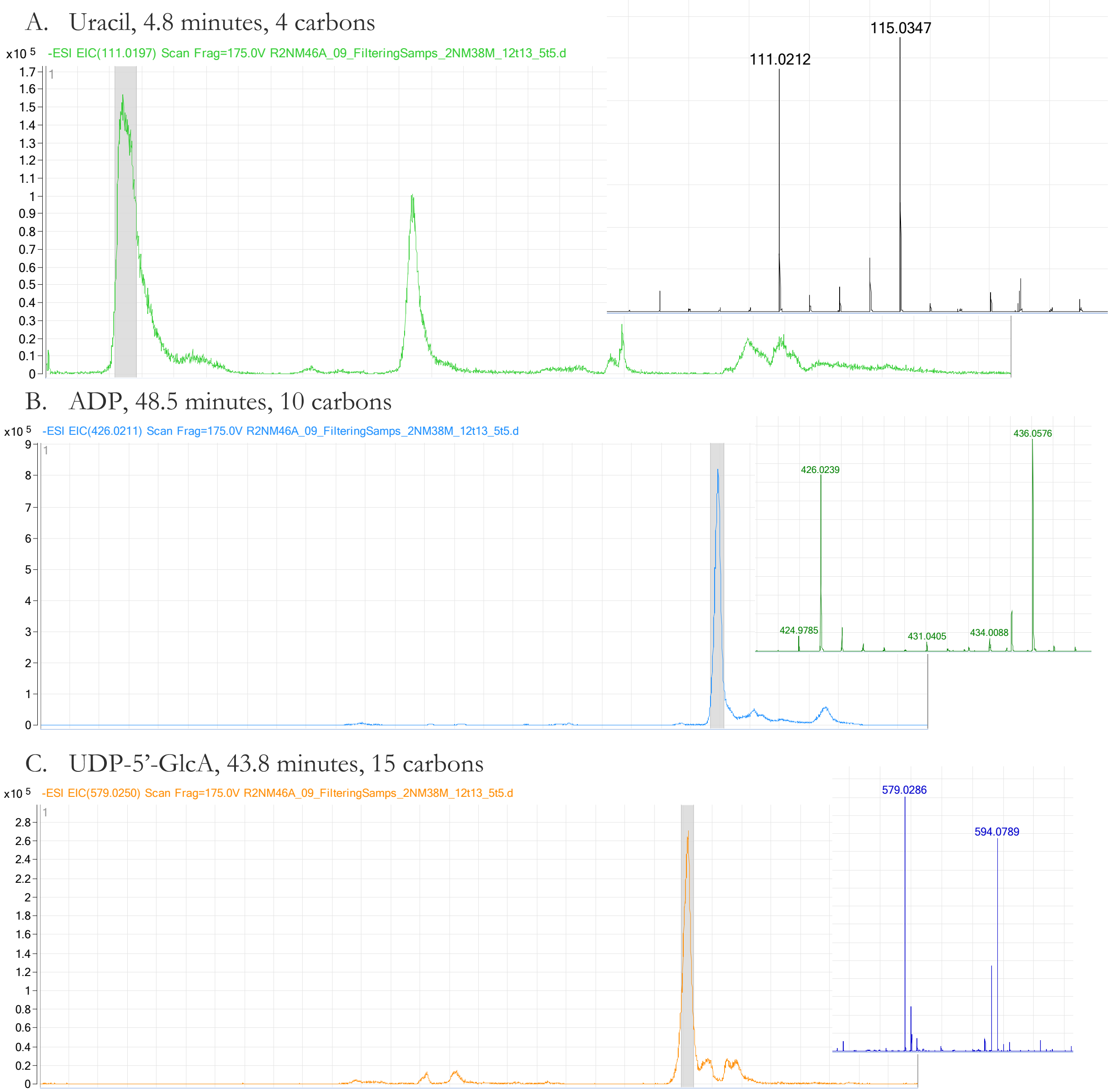
The credentialing technique is flexible and can be applied to many types of instrumentation and chromatography. Below are suggested values for different instrumentation that have been shown to be effective experimentally.

Parameter	Suggested Defaults	Explanation
iso_ppm	Time of Flight*: 4 Orbitrap**: 1 FT-ICR: 0.1	This is the mass error allowed when considering the difference between a ^{12}C and ^{13}C peak. This should be set according to the intra-scan mass error, rather than the absolute mass error of the instrument.
mix_tol	4	This is a coarse filter that ensures the ^{12}C peak and ^{13}C peak are of comparable intensity to their mixed ratios. This should allow a large error as many effects cause the $U^{12}C$ and $U^{13}C$ peaks to vary in intensity. A stricter filter is applied in the second round.
ratio_tol	1.8	This is a fine filter which ensures the intensity ratio between the two samples approaches the ratio of mixing (See Data Analysis). This is the most sensitive parameter and can be set according to the user's needs. Values approach 1 are more selective. 1.8 offers a false positive rate of approximately 0.6%
iso_rt	HILIC: 0.1 x (peak fwhm) C18: 0.05 x (peak fwhm)	This is the acceptable tolerance (in seconds) when matching a $U^{12}C$ peak to a $U^{13}C$ peak. Ideally the peaks have an identical retention times but in some cases poor peak shape causes the detected retention time to vary between isotopes. For chromatography which generates consistant peak shapes this can be lowered.
mpc_tol	1	Mass per carbon (mpc) is calculated as described above in Supplement S-2. The *mpc_tol* parameter is useful if a user is attempting to credential peaks with extremely large masses per number of carbons such as highly phosphorylated or metal containing compounds.

*Agilent QTOF, AB SCIEX TripleTOF, LECO Pegasus
**Thermo QE

Appendix 3.3. Raw Data Credentialed Features

Mass spectra and extracted ion chromatograms are shown for the three knowns targeted for MS/MS. The labeling pattern exhibited by credentialed features can be seen in the inset. (A) Uracil, (B) ADP, (C) UDP-GlcA. Inset mass spectra are averaged over the highlighted region of each chromatogram.



A. Uracil, 4.8 minutes, 4 carbons

B. ADP, 48.5 minutes, 10 carbons

C. UDP-5'-GlcA, 43.8 minutes, 15 carbons