Diss. ETH No. 23098

Mining of High-Resolution Mass Spectrometry Data to Monitor Organic Pollutant Dynamics in Aquatic Systems

A dissertation submitted to
ETH Zürich

for the degree of
Doctor of Sciences
(Dr. sc. ETH Zürich)

presented by
MARTIN JÜRGEN LOOS
MSc. Environmental Science, ETH Zürich
Dipl. Geoecology, University of Potsdam
born 25 October 1980
citizen of Germany

accepted on the recommendation of
Prof. Dr. Juliane Hollender, examiner
Heinz Singer, co-examiner
Dr. Steffen Neumann, co-examiner
Prof. Dr. Kristopher McNeill, co-examiner

2015

Contents

Summary

The ecological balance and human usage of aquatic environments are frequently affected by the widespread intentional or accidental release of anthropogenic compounds. As a particular case, the Rhine River has long been strained by a variety of emissions, whether from point sources like sewage treatment effluents and spill accidents or ubiquitous diffuse inputs from agriculture and horticulture. A long-term monitoring of the resultant pollutant levels is therefore crucial. Herein, especially the polar and therefore mobile portion of emitted organic compounds has come into focus. Termed micropollutants for their trace-level occurrence, these compounds can exert toxicological effects even at low concentrations and are composed of a complex range of pharmaceuticals, pesticides, surfactants or illicit drugs, amongst others. Many of the anthropogenic compounds are not suspected or even known to occur in the Rhine network, e.g., the sometimes hard-to-predict set of transformation products or unregistered industrial intermediates. Consequently, these compounds cannot be approached in a targeted manner. A relatively new method of choice for the chemical monitoring of such known and unknown micropollutants is thus the detection via high-performance liquid chromatography (LC) coupled to high-resolution mass spectrometry (HRMS). In this setup, electrospray ionization (ESI) offers a soft technique to transfer analytes from LC to HRMS without much fragmentation.

Among the few institutions responsible for detecting micropollutants in the named river network, the Swiss Rhine monitoring station (RÜS) in Basel has been equipped with LC-HRMS. Using a fully developed sampling strategy, the RÜS has acquired a long sequence of daily LC-HRMS measurements over the last years, comprising several hundred river samples. However, certain issues in the post-acquisition analysis have not been fully resolved yet and concern the simulation, reduction and automated trend analysis for LC-HRMS data. The aim of this thesis was to resolve these data mining issues in four steps.

A first part elaborated on the simulation of isotopic fine structures, necessary to compare the measured and theoretical mass spectra of compounds at high instrument resolution. Even for small compounds, dominating fractions of low-probable isotopologues exist, which must be efficiently pruned without

loss in computational performance. To this end, a novel transition tree approach was introduced to organize non-redundant, single-isotopic changes between isotopologues into separate tree branches. The latter can be efficiently pruned, with tree growth directed to find a relative instead of absolute pruning threshold first. The method was consequently able to outcompete existing approaches both in memory usage and computation time during an extensive performance comparison.

Benefitting from these transition trees, a second step improved the nontargeted isotopologue and adduct grouping of measured LC-HRMS signals. First, a large set of organic compounds from a public database was used to simulate isotopologue pairs and to sample defining characteristics between them, some of which are not covered by available low-resolution approaches. These characteristics were thereupon discretized in a recursive partitioning procedure and used to group coeluting measurement signals of isotopologues. A high recall and precision could therein be attested, based on a complementary evaluation with external simulation data and a targeted screening. When combined with a grouping for main ESI adducts, large fractions of measured LC-HRMS signals could thus be assorted into their chemical components.

A third step developed a first algorithm for a direct and unsupervised recognition of homologue series pattern in LC-HRMS data; that is, sets of signal series with constant shifts in mass and smooth changes in retention time. By introducing a specialized data structure, this novel approach swiftly revealed the patterns of numerous signal series even from very crowded spectra, despite the combinatorial complexity of this task. The detected series information could then be annotated to the grouped nontarget components. Further investigation also revealed multiple assignments of measured signals to different series, indicative of homologue series with different reoccurring chemical units.

Finally, the fourth step approached a primary goal of automatized trend detection to reveal riverine micropollutant spills from big data LC-HRMS sequences, incorporating the achievements in simulation and grouping. An automatized workflow for the fast extraction of chromatograms, picked signal peaks, time-intensity profiles, target and nontarget components, and trends was therefore implemented and equipped with a user interface. When tested for the routine monitoring at the RÜS, the workflow successfully prioritized numerous spill events, whereupon international alarms could be issued and the responsible emission sources of partly unknown micropollutants uncovered.

It can be concluded that environmental long-term monitoring of river systems with LC-HRMS results in data amounts which can neither be fully identified to appoint all the measured analytes nor sufficiently inspected by manual analysis alone. When fused with data mining strategies tailored to high-resolution acquisition and nontarget analysis, however, the emitted universe of polar aquatic pollutants can be routinely assessed for critical trends first and for selected chemical identification afterwards. A suite of five software tools for this former task has been made publicly available as part of this thesis (R packages *enviPat*, *enviPick*, *nontarget*, *nontargetData* and *enviMass*).

Zusammenfassung

Die umfangreiche und allgegenwärtige Emission von anthropogenen Stoffen in aquatische Systeme steht häufig im Konflikt mit der menschlichen Nutzung oder ökologischen Funktion selbiger Systeme. Ein beispielhafter Fall ist der Rhein, welcher kontinuierlich durch solche Emissionen beeinflusst wird, entweder in Form von Punktquellen (z.B. Kläranlagenausflüsse, Unfälle) oder durch diffuse Einträge aus der Landwirtschaft und dem Gartenbau. Eine langfristige Überwachung der resultierenden Schadstoffbelastungen ist daher unerlässlich, wobei der Eintrag organischer polarer Chemikalien verstärkt in den Fokus gerückt ist, zumal diese Chemikalien gut wassergängig und somit mobil sind. Letztere Substanzklasse wird aufgrund ihrer häufig niedrigen Konzentration auch als Mikroschadstoffe bezeichnet; einzelne Vertreter können zudem selbst bei starker Verdünnung noch toxische Beeinträchtigungen verursachen. Das zugrundeliegende Substanzspektrum wiederum ist vielfältig und betrifft unter anderem Stoffe wie Pharmazeutika, Pestizide, Tenside oder auch Drogen. Viele der im Rhein auftretenden Substanzen sind ausserdem noch nicht gänzlich aufgeklärt, da sie z.B. schwer hervorsagbare Transformationsprodukte oder intermediäre Industriechemikalien darstellen. Einerseits können solch unbekannten Substanzen nicht vorab eingegrenzt werden, da oft keine ausreichende Erwartungshaltung definierbar ist. Andererseits können polare Substanzen aber – wenngleich nicht identifizier – so dennoch häufig detektiert werden. In der chemischen Spurenstoffanalytik ist hierbei vor allem eine Kopplung aus (a) Hochleistungsflüssigkeitschromatographie, (b) Elektrospray-Ionisierung und (c) hochauflösender Massenspektrometrie zum Einsatz gekommen (abgekürzt LC-ESI-HRMS oder LC-HRMS).

Neben den verschiedenen mit der Detektion von Mikroschadstoffen im Rhein betrauten Einrichtungen betreibt auch die Rheinüberwachungsstation Basel (RÜS) eine solche LC-HRMS Analytik. Durch eine tägliche automatisierte Probennahmestelle konnte die RÜS hierbei über die letzten Jahre eine beachtliche LC-HRMS Messreihe aufbauen. Obgleich die analytische Seite stark ausgereift ist, ist in der Datenauswertung hinsichtlich Simulation von Messergebnissen, Datenreduktion und der automatisierten Trenddetektion jedoch Nachholbedarf ersichtlich. Die hier präsentierte Doktorarbeit behandelt diese Schwachstellen in insgesamt vier Schritten.

Ein erster Schwerpunkt zielte auf die Simulation von Isotopenfeinmustern ab, welche zum Vergleich von theoretischen und gemessenen Massenspektren benötigt werden. Da selbst kleine Moleküle einen Hauptanteil von Isotopologen mit vernachlässigbarer Auftretwahrscheinlichkeit aufweisen, sollten letztere effizient in solchen Berechnungen umgangen werden. Zu diesem Zweck wurde eine neuartige Berechnungsmethode erarbeitet, welche die Übergänge mit einzelnen Isotopen zwischen paarweisen Isotopologen in sogenannten Transitions-Bäume organisiert. Dabei wurden nicht nur redundante Übergänge vermieden, sondern auch einzelne Bereiche dieser Bäume entweder abgegrenzt oder vorranging auf wahrscheinliche Isotopologe hin überprüft. Der einhergehende Gewinn in Berechnungsgeschwindigkeit und Speicherverbrauch ist dabei besser als derjenige existierender Methoden, was durch umfangreiche Simulationsvergleiche gezeigt werden konnte.

Ein zweiter Schwerpunkt befasste sich mit einer Verbesserung der Isotopologen- und Adduktgruppierung für LC-HRMS Messdaten von unbekannten Substanzen. Unter Zuzug einer öffentlich zugänglichen Substanzdatenbank wurden hierfür zunächst umfassende Simulationen von Isotopologenpaaren berechnet. Die beobachteten Eigenschaften dieser Paare - die überdies deutlich von denjenigen aus niedrig aufgelöste Massenspektren abwichen – wurden mithilfe einer rekursiven Partitionierung dann diskretisiert und zur Sortierung von gemessenen Isotopensignalen herangezogen. Validierungsschritte mit externen simulierten Isotopologenpaaren sowie mit solchen aus gemessenen und a priori aufgeklärten Paaren wiesen dabei sowohl eine hohe Wiederfindungrate als auch eine zufriedenstellende Klassifizierungsgenauigkeit auf. In Kombination mit einer Gruppierung der wichtigsten ESI-Addukte konnten hernach grosse Signalanteile in einzelne chemische Messkomponenten zusammengeführt werden.

Ein nachfolgender dritter Teilschritt entwickelte einen ersten Algorithmus zur direkten und flexiblen Erkennung von Messmustern, welche das Vorhandensein von homologen Reihen in LC-HRMS Datensätzen nahelegen. Diese Muster charakterisierten sich einerseits durch konstante Massenabstände zwischen mehreren verketteten Messsignalen, andererseits aber auch durch gleichmässige Veränderungen in der Retentionszeit. Durch Ausarbeitung einer angepassten Datenstruktur konnten diese Muster nunmehr mit hoher Rechengeschwindigkeit detektiert werden, selbst in mit Messsignalen stark überfrachteten Messungen und trotz der Vielzahl an kombinatorischen Möglichkeiten für diese Mustererkennung.

In einem letzten übergeordneten Aufgabenbereich wurde eine automatisierte Trendüberwachung von Mikroschadstoffen durch das Filtern von LC-HRMS Datensätzen und unter Zuzug der bisher erarbeiteten Teilschritte (Simulation, Signalgruppierung und Mustererkennung von Homologen) erstellt. Der resultierende Workflow umfasste u.a. die Extraktion von Chromatogrammen, die Detektion von Signalpeaks, die Erstellung von Intensitäts-Zeitprofilen, die oben erwähnte Komponentenbildung und die Priorisierung von Intensitätsanstiegen um rasch auf plötzlich ansteigende Intensitätsveränderungen hinweisen zu können. Weitergehend wurde eine Benutzeroberfläche generiert um an der RÜS angewendet werden zu können. Dort konnten infolgedessen zahlreiche Intensitätstrends im Rhein nachgewiesen werden – und im Rahmen von internationalen Alarmfällen nicht nur etliche Verursacher, sondern auch einige bis dato unbekannte Mikroverunreinigungen zur Aufklärung gebracht werden.

Zusammengefasst lässt sich festhalten, dass eine umweltspezifische und langfristige Überwachung von Flüssen wie dem Rhein ohne LC-HRMS kaum machbar ist, jedoch die durch ausgedehnte Generierung von Messdaten anfallenden Anforderungen der Datenanalyse manuell nicht mehr hinreichend erschöpfbar sind. Neue Möglichkeiten ergeben sich daher aus der Erstellung von automatisierten Workflows. Hier können aus der Vielzahl der durch das breite Substanzspektrum hervorgerufenen Messsignale einzelne Trends herausgefiltert werden – ohne dass eine völlige Identifizierung der gemessenen Matrix nötig wäre, welche stattdessen vereinzelt nachfolgen kann. Die im Rahmen der Doktorarbeit erarbeiteten Methoden wurden in fünf Softwarepaketen implementiert und öffentlich verfügbar gemacht (R Pakete *enviPat*, *enviPick*, *nontarget*, *nontargetData* und *enviMass*).

Chapter 1 - Introduction

1.1 Micropollutants in the river Rhine

A large variety of known and unknown anthropogenic compounds from industry, agriculture, households, hospitals and transport are routinely released into riverine environments, both through point and diffuse sources.[1] Specific examples of such sources are sewage treatment plants, accidental industrial spills or diffuse inputs from plant protection applications. The diverse organic class of these compounds comprises, for example, pesticides, biocides, pharmaceuticals, plasticizers, corrosion inhibitors and surfactants, of which several are even classified as high production volume chemicals (HPVC).[2] In addition, a variety of transformation products (TPs) are formed from the emitted organic compounds through biological and abiotic processes, e.g., microbial degradation, phototransformation or hydrolysis.[3] With one parent compound often undergoing several transformations, the relative number and complexity of the formed TPs is even increased, frequently embracing yet unidentified substances. As a result, these anthropogenic chemicals occur with a larger structural diversity and in a more varied composition than natural compounds, although both the spectrum of chemical elements and their relative numbers of atom can be somewhat restricted for the former class.[4] Furthermore, toxic modes of action can be attested to many of these continuously or discontinuously released compounds. Modes range from baseline toxicity[5] to reactive toxicity[6] and can exert more complex (e.g., synergistic) effects when occurring in mixtures.[7] Termed micropollutants for their adverse effects, these contaminants are often encountered at low concentrations (ng/l - µg/l) and at masses typical for small molecules (<1000 u).

The intentional or accidental release of micropollutants often collides with the ecological function and human usage of the affected environment. One specific example is the river Rhine, a major European river system with a total length of 1233 km. With an annual discharge volume of 70 km^3 and a catchment area of more than $2x10^5$ km^2 spread over nine countries, its river bank filtrate produces drinking water for over 20 million citizens and its water body serves as an important ecological habitat. On the other hand, roughly half of the catchment is in agricultural use, 5800 sewage treatment plants (STPs) process the wastewater of 58 million inhabitants to be eventually released as effluent and intense trading frequents a fraction of 0.67 of the river length as a crucial waterway. In addition, 21 hydroelectric plants are powered by the Rhine`s discharge and various industrial factories or power

plants use its water for cooling.[8,9] Not surprisingly, chemicals such as saccharin or acesulfam (sweeteners), metformin (an antidiabetic drug) and atrazine (an abolished pesticide), to name a few, have been confirmed to occur in the Rhine.[10] Following the Schweizerhalle accident in 1986,[11] a network of seven monitoring stations was thus established along the river, chaired by the International Commission for the Protection of the Rhine (ICPR). One of these stations is the Swiss Rhine monitoring station (RÜS) downstream of the industrial hotspots of Basel. Its unique location allows the RÜS staff to monitor 68% percent of the national catchment area, which is in turn inhabited by 80% of the total population of Switzerland.

1.2 Chemical analysis of micropollutants

Given the problem of their widespread release, adverse effects and chemical diversity, a broad analytical method is required to investigate micropollutants at the mentioned monitoring station. Herein, especially the polar, ionic and therefore mobile and relevant fraction of organic pollutants emitted into the river Rhine is of interest. Being oftentimes more polar than their parent compounds, most TPs can be expected to be part of this fraction, too. As a method of choice, high-performance liquid chromatography (LC) coupled to mass spectrometry (MS) enables a rapid, reproducible and simultaneous analysis for a multitude of low-volatile compounds.[12] Within this setup, usage of soft ionization techniques such as electrospray ionization (ESI) enables the direct analysis of non-fragmented quasimolecular ions, avoiding further derivatization of the analytes. For very low concentrations and insufficient instrument sensitivity, analytes can be enriched prior to LC-MS measurements, using techniques such as solid phase extraction (SPE). This hyphenation of chromatographic and mass spectrometric separation thus characterizes analytes in the three dimensions of retention time (*RT*), mass-to-charge ratio (*m/z*) and measurement intensity, as detailed in the following.

First, the chromatographic separation from adsorption of analytes contained in a mobile solvent to a stationary phase causes different compounds to pass a chromatographic column at differing rates. Analytes are thereby characterized by their retention time to elute from this column. In reverse phase LC, for instance, a non- or semi-polar stationary phase is passed by an aqueous mobile phase and the affinity of an analyte to the stationary phase stems from hydrophobic interactions. The mobile phase can be modified through the polarity of the involved solvents: gradient elution changing from polar (e.g., water) to less polar (e.g., methanol or acetonitrile) mobile phases can be utilized to ensure elution of the more affine analyte portion after the polar (or sterically affected) one has eluted from the column at lower *RT*.

Second, MS characterizes analytes by their mass-to-charge ratios. For this purpose, and as an intermediate step between chromatographic elution and the MS inlet, analytes need to be ionized in an ion source. With ESI, the eluting solvent is dispersed from a capillary into an aerosol (spray) in a high voltage electric field. The solutes in the resulting droplets gradually evaporate and lead to a beam of charged ions which subsequently enter the mass analyzer. Besides commonly used quadropole low resolution mass spectrometers, high resolution MS is more and more used due to its higher selectivity and its ability to measure a broad mass range. Such high resolution MS instrumentation includes linear ion trap (LTQ)-Orbitrap and Time-of-Flight (TOF) analyzers. Although both instrument types are based on the differential acceleration of charged compounds, the first type derives *m/z* values from flight times in an electric field coupled to ion counting detectors, whereas the second analyzer confines charged compounds into an orbit in which their oscillation induces currents with frequencies that convert into mass spectra after Fourier transformation. As a result, instruments somewhat deviate in their defining properties, namely their mass accuracy, their scan speed, the so-called dynamic range over which ion signals remain linear with the analyte concentration, and the instrument resolution.[13] The latter in turn defines the separation of signals closely adjacent in mass and is a instrument-specific function of *m/z* itself. The separation of compounds with same nominal mass using high resolution mass spectrometry (HRMS) is nowadays available for both instrument types. Mass accuracies of up to 2 ppm can be achieved with an internal calibration and using Orbitrap instrumentation[14].

Third, and as part of the MS acquisition, analytes are characterized by their signal intensity, relatable to ion counts surpassing a limit of detection (LOD). In this aspect, MS instrumentation differs with regard to the dynamic mass range and the intensity precision. The former is usually at least a magnitude larger for Orbitrap than for TOF instrumentation under a consistent mass accuracy.[15,16] In terms of sensitivity and data acquisition rates however, TOF instruments can outperform Orbitrap setups.[17] Within these limitations, the signal intensities caused by an analyte can be traced over LC-HRMS measurements sequences corresponding to, e.g., temporal sampling campaigns. Importantly, and even if no calibration for a quantitation of concentrations exists, relative intensity variations of LC-HRMS signals over such sequences allow for inferences on the relative concentration changes of measured known or unknown analytes. Overall, LC-HRMS acquires mass spectra composed of discretized *m/z* values at certain intensities (so-called profile data), scanning the eluting sample fraction for each such a spectra with a defined frequency. Following conversion from profile to centroid data, an analyte ion is thus registered as a set of data points consecutively detected over variable stretches in retention time (Figure 1.1).

Apart from the described three dimensions of *RT*, *m/z* and measurement intensity, analytes can also be fragmented in either selective or data-

independent MS/MS experiments. The fragment mass spectra can then serve as supplementary information for identification and structure elucidation.[18]

Furthermore, different conceptual types of LC-MS analysis must be distinguished.[12] The first, targeted analysis, combines *RT* and MS/MS information determined by a reference standard with the expected mass of an analyte to unambiguously confirm the presence of the latter in a LC-HRMS measurement. The second type, suspect screening, lacks reference standards; *RT* and MS/MS information might at best be predicted to complement the mass information of the analyte. An unequivocal confirmation is therefore not possible. Third, in non-target screening neither a reference standard nor suspicion on the analyte that may constitute an observed signal are available. However, the mass and *RT* of a measured non-target signal may help to appoint candidates by means of, e.g., molecular formula-fits or database searches. Notably, the intensity variation of non-target signals can nevertheless be pursued, even without knowledge of the responsible analyte.

1.3 Challenges and research gaps

Although LC-HRMS has been established as a sensitive, selective and reproducible analytical method, certain issues with its subsequent data analysis have not been fully resolved. These issues arise especially in the context of temporal micropollutant analysis as faced by environmental monitoring stations such as the RÜS. These issues are categorized threefold in the following and are all related to the post-acquisition handling of LC-HRMS measurements.

1.3.1 Data reduction

High-throughput LC-HRMS produces a wealth of data signals, which need to be pooled at different levels in order to interpret them correctly. In fact, the total number of data points acquired over the whole mass and retention time range at high resolutions is overwhelming, as exemplified by the two small mass ranges in Figure 1.1. In addition, measurement uncertainties and potential signal interferences between different analytes cause variations in intensity, and deviations of *m/z* from the expected values of the analytes. In conjunction with the enormous diversity of molecular compositions which increases with mass, LC-HRMS measurements amount to highly heterogeneous and large data sets.

For a necessary reduction of the LC-HRMS data, a first step of automatized aggregation is the unsupervised assignment of consecutive centroid signals caused by the same (known or unknown) analyte ion, leading to extracted

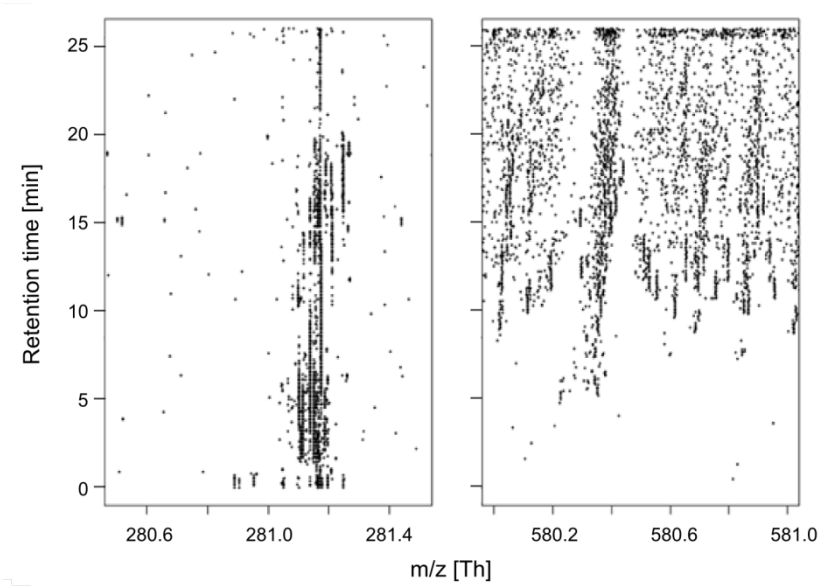


Figure 1.1 *The structure of centroided data points from the same LC-HRMS measurement depends on the m/z-region under consideration (Orbitrap XL Velos Pro, resolution 60K at m/z=400).*

ion chromatograms (EICs). Thereafter, a second step detects intensity peaks in the EICs and discards noise signals, a process commonly referred to as peak picking. The big data sets can thus be reduced by orders of magnitude to lists of picked peaks, which are characterized by a summary statistic of their *m/z*, *RT* and maximum intensity (or peak area).

Peak picking does not represent the final stage in possible reduction, considering that one analyte can form a multitude of coeluting peaks of, e.g., its different ionized species. The most common ESI species of small molecules arise from addition or subtraction of a proton during positive and negative ionization, respectively. Yet other ion species with, e.g., sodium or ammonium or additions with purposely added compounds to facilitate protonation can be formed. Multiply charged ions or dimers can also be formed, especially for larger analyte molecules. Even in targeted analysis, the relative proportions of different ion species of a compound are not always known beforehand and can be easily affected by changes in the mobile phase composition or the composition of the coeluting matrix. Furthermore, different isotopologue peaks exist for each ionized species, reflecting the differing isotopes

and atom counts of the elements an analyte is composed of. A large fraction of isotopologues per species are however of low probability and will hardly trigger measurable signals. The remaining fraction superimposes to the measurable data peaks of an analyte ion, depending on the *m/z* differences among the isotopologues and the given resolution. In some fields of research, the distribution of picked isotopologue peaks of an ion species can be approximated by average molecules or probabilistic models (a) under concise restriction of the compounds of interest (e.g., in peptidomics or lipidomics research), and (b) in low (i.e., nominal) resolution settings. Knowledge on such distributions allows for a grouping of the coeluting analyte peaks even if the responsible analyte is not a priori known or suspected to occur in a sample. Unfortunately, this grouping task has remained challenging for the intricate combination of high resolution MS and the complex universe of small molecules encountered during micropollutant monitoring. Notably, grouping neither only reduces the amount of LC-HRMS signals to monitor nor solely mitigates redundant inferences. Instead, valuable information for the identification of unknown compounds can be gained from that signal grouping, e.g., when monoisotopic masses need to be determined for molecular formula fitting or when subsequent database matches to grouped isotopologue signals are more conclusive than to solitary ones.

Yet a third level of reduction and complementary information can be reached for compounds that form homologue series (HS), foremost surfactants. The latter have been recently identified in effluents of Swiss STPs, but still await broader inspection in the receiving discharge of the Rhine.[26] HS are characterized by repeating chemical units, which lead to LC-HRMS peak patterns separated by recurring *m/z* differences and, mostly from concomitant increases in hydrophobicity, smooth shifts towards larger *RT*. Both the masses of the repeating units and the induced *RT* shifts differ widely among HS compounds and are prone to the named measurement uncertainties. In practice, these systematic patterns are impossible to discern from the multitude of picked peaks by visual inspection alone. Nonetheless, no data mining strategy has to date addressed a non-targeted recognition of these characteristic linear and nonlinear peak patterns in *m/z* and *RT*, respectively. In addition to data aggregation and to the mass information on the homologue units from detecting HS series peaks, an averaging of the systematic *m/z* differences within such series patterns can be expected to yield better *m/z* estimates and thereby assist later HS identification attempts.

1.3.2 Data simulation

To interpret mass spectrometric data, measured LC-HRMS signal peaks must be compared to the expected, i.e., theoretical mass signals for individual analytes. A reliable simulation of theoretical masses and probabilities of isotopologues from the molecular formula of an analyte as well as their su-

perposition to theoretical isotopologue peak patterns to be expected at a given resolution is hence vital. For instance, screening of target or suspect compounds would be infeasible without simulating their mass spectrometric peaks, as would be the reduction of candidate formulas during unknown identification from a comparison of their expected vs. measured isotopologue peak patterns.

The requirements for isotopologue simulation algorithms are not trivial: computational speed and memory usage should be optimized to swiftly process large batches of analyte molecular formulas, while embracing a broad range of polyisotopic elements and often vast numbers of isotopologues per formula. Again, a plethora of algorithms for low resolution simulation are available – but only few approximate the theoretical peak patterns observable at higher resolutions.[27] Among these latter strategies, different techniques for the necessary pruning of low-probable isotopologues have been employed. However, several research gaps for pruning these low-probable candidates have not been addressed, concerning (a) a dynamical adaption to the differing isotopologue distributions of different compounds at different instrument resolutions, and (b) the computational pruning performance.

1.3.3 Automated trend detection

Separated in mass and retention time, the temporal concentration changes of micropollutants in the river Rhine can be traced by intensity variations of peaks to be picked from temporal LC-HRMS measurement sequences. These measurement sequences have to date been explored in three general ways. One targeted strategy has relied on the manual or at most semi-automatized extraction of LC-HRMS signal peaks for selected target pollutants, including a quantification of their concentration changes in the environment.[19] In contrast, non-targeted strategies have intersected LC-HRMS sequences for the joint occurrence of comparable signal peaks.[20] Other non-targeted approaches use statistical data aggregation to elaborate on the principal components of the signal sequences, e.g., to point at main emission sources.[21,22] However, all three approaches have certain disadvantages when applied to riverine micropollutant monitoring. First, targeted analysis on pollutants known or suspected to occur in an aquatic system can miss a large fraction of yet unknown or emerging pollutants.[23] Extensive prior knowledge on the number and nature of non-targets to guide and extend the focus of monitoring is unfortunately rare. The complexity of chemical identification in turn tightly restricts the number of unknowns that can be converted to new targets within reasonable time frames; even a single LC-HRMS measurement contains too much data to be mined manually, not to speak of sequences. Identifying or exploring only the most intense unknown signals, on the other hand, does neither ensure focus on the pollutants of highest toxicological relevance nor - since ionization efficiencies differ among molecules - highest aquatic con-

centration. Similarly, data aggregation to principal components will not necessarily point at the intensity variations of individual signals. The non-targeted mining of temporal intensity increases has not been emphasized by any of these approaches – despite being indicative for pollutant spills of environmental relevance.

Promising alternatives have emerged in the field of metabolomics or proteomics, where LC-HRMS sequences are often mined to compare biological conditions and treatments over time.[24,25] However, these alternatives rarely account for all requirements of environmental monitoring. For example, many approaches embrace only specific compound classes (e.g., peptides or lipids) and cannot reliably deal with the full range of relevant micropollutants. Again, some approaches address solely targeted compounds. Other analysis pipelines simply lack important processing steps (e.g., background subtraction or intensity normalization), cannot integrate new modules, are restricted to low-resolution MS or do not provide a graphical interface for non-programmers. Moreover, not all extract and prioritize LC-HRMS intensity trends. Above all, none of the pipelines can quickly update processed batches of several hundred LC-HRMS measurements by a few new ones without lengthy recalculations, which is mandatory to react upon pollutant emissions within stringent timeframes.

1.4 Objectives and contents of the thesis

Given the above research gaps, the primary goal of the thesis is the implementation of a data mining strategy for an automatized spill and trend detection workflow of organic micropollutants in riverine environments, using long-term LC-HRMS measurement sequences. The strategy has to account for shortcomings in isotopologue simulation, peak grouping and HS detection. The goal is achieved in four steps:

(1) Development of an algorithm to simulate highly resolved isotopologue peak patterns.

(2) Development of an isotopologue and adduct grouping algorithm tailored to micropollutants measured at high MS resolutions.

(3) Unsupervised detection of mass and retention time peak patterns indicative of homologue series.

(4) Combination of all aspects into a user-friendly workflow for an automatized pollutant trend detection in the river Rhine.

Aspects (2) to (4) are schematically illustrated in Figure 1.2. Moreover, aspects (2) and (3) are to be thoroughly tested with STP samples that discharge into the river Rhine. While providing a more complicated and less diluted matrix for evaluation, analytes from these STP effluents likely contribute to the downstream Rhine matrix to be monitored under aspect (4).

The presented work is hence structured into four parts. Chapter 2 first details an improved method for calculating isotopologues, which is a prerequisite for developing a method to group measured LC-HRMS data in chapter 3. Chapter 4 deals with the detection of homologue series patterns. Finally, chapter 5 merges both these methods and complementary tools into one workflow and implements it at the Rhine monitoring station Basel.

Chapter 2 introduces a fast computation of the most probable isotopologues for a given molecular formula, based on so-called transition trees. In contrast to previous methods, a relative threshold is proposed to prune isotopologues of low probability. The choice of this threshold in turn influences the accuracy of a measurable isotopologue peak pattern, as quantified by simulations. Both the required number of intermediate calculations and the computational speed of the strategy are compared to an existing method.

In **Chapter 3**, the approach from Chapter 2 is used for a large-scale simulation of the isotopologue peak patterns of a large set of organic compounds at different resolutions. Pairwise characteristics among these isotopologue peaks are then described and represented by a discretized data model. The latter model can then be used to group the isotopologue peaks of other unknown compounds, validated for a test set comprising both simulated and measured compounds. In addition, the grouping of different ion species of measured analytes is discussed, as well as their full componentization in conjunction with the isotopologue peak groups.

Chapter 4 introduces a computational method to detect regular mass and retention time differences within large LC-HRMS peak sets, which can be associated with the occurrence of homologue series. Difficulties in the interpretation of the detected regular patterns are explicitly addressed.

Chapter 5 integrates the above methods as part of a workflow for the routine monitoring of micropollutant spills and trends in the river Rhine. The time profiles of detected spill cases of differing origin are presented. The peak picking methodology and some of the data processing required in chapters 3 and 4 is also used and further detailed here.

Finally, **Chapter 6** summarizes all findings, relates them and gives an outlook on open research questions.

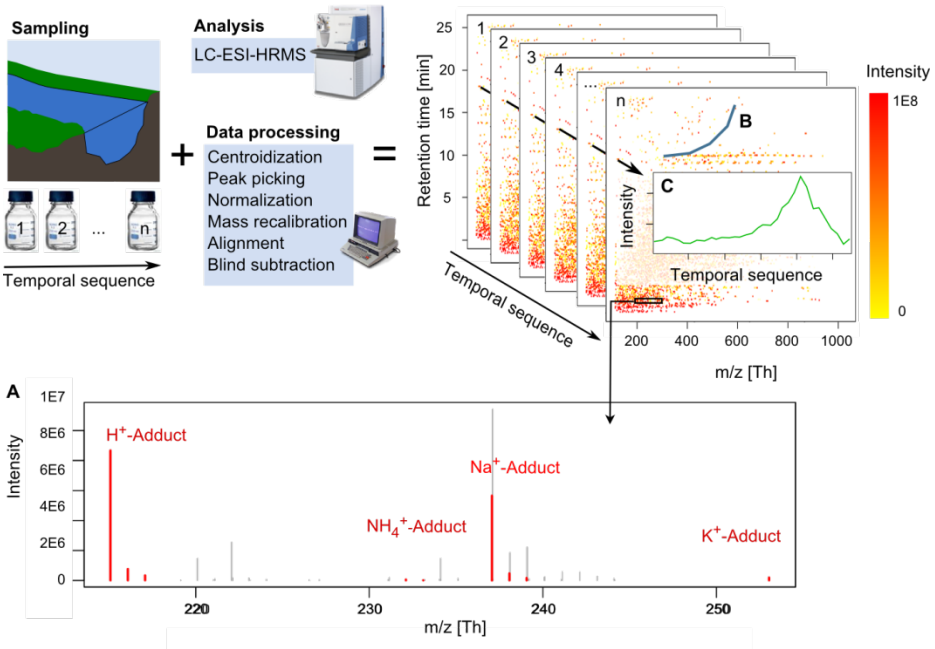


Figure 1.2 *Three main aspects of the presented research, using temporal sequences of processed LC-HRMS measurements.* ***(A)*** *Different adducts and their isotopologue peak patterns are formed by one analyte at similar retention times and need to be grouped (red bars, showing the spectrum of DMST, a transformation product of the fungicide Tolylfluanid) and thereby distinguished from the isotopologue peak patterns of other close-eluting analytes (gray bars).* ***(B)*** *Recognition of systematic shifts of mass and retention time in the measured data can aid the detection of homologue series.* ***(C)*** *Intensity increases in the time profiles of measured data help to reveal trends of concern caused by, e.g., industrial spills.*

References

(1) Luo, Y.; Guo, W.; Ngo, H. H.; Nghiem, L. D.; Hai, F. I.; Zhang, J.; Liang, S.; Wang, X. C. *Science of the Total Environment* **2014**, *473*, 619–641.

(2) UNEP. *UNEP Publications*.

(3) Fenner, K.; Canonica, S.; Wackett, L. P.; Elsner, M. *Science* **2013**, *341* (6147), 752–758.

(4) Kind, T.; Fiehn, O. *BMC bioinformatics* **2007**, *8* (1), 105.

(5) Wezel, A. P. van; Opperhuizen, A. *CRC Critical Reviews in Toxicology* **1995**, *25* (3), 255–279.

(6) Freidig, A. P.; Verhaar, H. J.; Hermens, J. L. *Environmental science & technology* **1999**, *33* (17), 3038–3043.

(7) Chèvre, N.; Maillard, E.; Loepfe, C.; Becker-van Slooten, K. *Ecotoxicology and environmental safety* **2008**, *71* (3), 740–748.

(8) Agency, E. E. Waterbase - UWWTD: Urban Waste Water Treatment Directive – reported data, 2015.

(9) Rhine (ICPR), I. C. for the Protection of the. The Rhine and its catchment: an overview, 2013.

(10) Ruff, M.; Singer, H.; Ruppe, S.; Mazacek, J.; Dolf, R.; Leu, C. *Aqua & Gas* **2013**, *93* (5), 16–25.

(11) Capel, P. D.; Giger, W.; Reichert, P.; Wanner, O. *Environmental science & technology* **1988**, *22* (9), 992–997.

(12) Krauss, M.; Singer, H.; Hollender, J. *Analytical and bioanalytical chemistry* **2010**, *397* (3), 943–951.

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

(14) Makarov, A.; Denisov, E.; Kholomeev, A.; Balschun, W.; Lange, O.; Strupat, K.; Horning, S. *Analytical chemistry* **2006**, *78* (7), 2113–2120.

(15) Blom, K. F. *Analytical chemistry* **2001**, *73* (3), 715–719.

(16) Makarov, A.; Denisov, E.; Lange, O.; Horning, S. *Journal of the American Society for Mass Spectrometry* **2006**, *17* (7), 977–982.

(17) Rousu, T.; Herttuainen, J.; Tolonen, A. *Rapid Communications in Mass Spectrometry* **2010**, *24* (7), 939–957.

(18) Kind, T.; Fiehn, O. *Bioanalytical reviews* **2010**, *2* (1-4), 23–60.

(19) Gómez, M.; Gómez-Ramos, M.; Malato, O.; Mezcua, M.; Férnandez-Alba, A. *Journal of Chromatography A* **2010**, *1217* (45), 7038–7054.

(20) Müller, A.; Schulz, W.; Ruck, W. K.; Weber, W. H. *Chemosphere* **2011**, *85* (8), 1211–1219.

(21) Karaouzas, I.; Lambropoulou, D. A.; Skoulikidis, N. T.; Albanis, T. A. *Journal of Environmental Monitoring* **2011**, *13* (11), 3064–3074.

(22) Terrado, M.; Kuster, M.; Raldúa, D.; Alda, M. L. de; Barceló, D.; Tauler, R. *Analytical and bioanalytical chemistry* **2007**, *387* (4), 1479–1488.

(23) Lahr, J.; Maas-Diepeveen, J. L.; Stuijfzand, S. C.; Leonards, P. E.; Drüke, J. M.; Lücker, S.; Espeldoorn, A.; Kerkum, L. C.; Stee, L. L. van; Hendriks, A. J. *Water Research* **2003**, *37* (8), 1691–1710.

(24) Aiche, S.; Sachsenberg, T.; Kenar, E.; Walzer, M.; Wiswedel, B.; Kristl, T.; Boyles, M.; Duschl, A.; Huber, C. G.; Berthold, M. R.; others. *Proteomics* **2015**, *15* (8), 1443–1447.

(25) Castillo, S.; Gopalacharyulu, P.; Yetukuri, L.; Orešic, M. *Chemometrics and Intelligent Laboratory Systems* **2011**, *108* (1), 23–32.

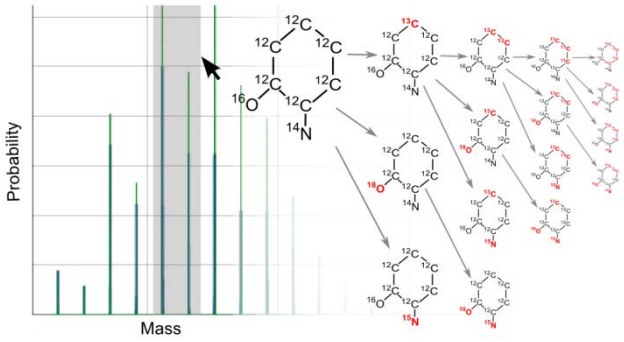
(26) Schymanski, E. L.; Singer, H. P.; Longree, P.; Loos, M.; Ruff, M.; Stravs, M. A.; Vidal, C. Ripolle s; Hollender, J. *Environmental science & technology* **2014**, *48* (3), 1811–1818.
(27) Valkenborg, D.; Mertens, I.; Lemiere, F.; Witters, E.; Burzykowski, T. *Mass spectrometry reviews* **2012**, *31* (1), 96–109.

Chapter 2

Accelerated Isotope Fine Structure Calculation Using Pruned Transition Trees

Martin Loos, Christian Gerber, Francesco Corona, Juliane Hollender, Heinz Singer

Published in Analytical Chemistry, 87(11), 5738-5744, 2015



ABSTRACT: A fast and memory-efficient calculation of theoretical isotope patterns is crucial for the routine interpretation of mass spectrometric data. For high-resolution experiments, calculations must procure the exact masses and probabilities of relevant isotopologues over a wide range of polyisotopic compounds, while pruning low-probable ones. Here, a novel albeit simple tree-like structure is introduced to swiftly derive sets of relevant sub-isotopologues for each element in a molecule, which are then combined to the isotopologues of the full molecule. In contrast to existing approaches, transitions via single replacements of the most abundant isotope per element are used in separable tree branches to derive sub-isotopologues from each other. Moreover, the underlying transition trees prevent redundant replacements and permit the detection of the most probable isotopologue in a first phase. A relative threshold can then be exploited in a second parallelized phase for a precise pre-pruning of large fractions of the remaining sub-isotopologues. The gain in performance from such early pruning and the lower variation in the distortion of simulated data when using relative rather than absolute thresholds were validated in a large-scale benchmark simulation, unprecedentedly comprising several thousand molecular formulas. Both the algorithm and a wealth of related features are freely available as R-package *enviPat* and as a user-friendly web-interface.

2.1 Introduction

An effective calculation of isotopic patterns from molecular formulas is essential to understand mass spectrometric (MS) measurements. The simulated tuples of isotopologue masses and probabilities are required for restrictions during unknown identification,[1,2] targeted screening,[3] nontargeted signal grouping[4] and the annotation of product ions in MS^n experiments,[5] among others. Herein, large batch calculations of candidate isotope patterns need often to be derived and compared, demanding computational speed and memory. On top, the measured molecules cover a wide array of polyisotopic elements and range from small molecules to polypeptides and proteins, typically with vast numbers of isotopologues per molecule even at lower masses.

Tackling the combinatorial complexity of simulating the measurable sets of such isotopologues has led to a plethora of approaches, not all of which scale well with increasing number of isotopes, elements, atoms or instrument resolution.[6] These approaches can broadly be classified into two groups. One group either aggregates the individual isotopologues to nominal and center masses or directly samples the convolution of resolution-dependent peak shapes.[7] Although results may range within measurable accuracies, the underlying isotope fine structures are not fully resolved. This structure is needed for data interpretation or when convoluting a calculated set of isotopologues for different instrument resolutions. The latter can differ in orders of magnitude at a given mass and subsequently result in widely differing measurement signals.[8] Even more concerning, computational limitations are likely to arise for methods of this first group with the advent of ever higher resolutions.[9] Given that widespread Time-of-Flight and Orbitrap MS instrumentation already achieves resolutions well above 1E5, a comparison of highly resolved measurements with rapidly simulated isotopic features is central to, e.g., complement monoisotopic mass information in molecular formula assignments. A second albeit smaller group of approaches has therefore been established, yielding isotope fine structures. These approaches are based on dynamic programming,[10] transitions in the multinomial distribution[11,12,13] or, most recently, multidimensional Fourier transforms.[9]

With isotopologue numbers easily exceeding computational resources, all these isotope fine structure approaches must prune their low-probable isotopologues. In some cases, such pruning is paid by imprecisions for the resulting isoto-pologues[10] or risks a loss of sufficiently probable ones at subsequent calculation stages.[11] But even in the precise cases, pruning has been restricted

to the late stage of combining sub-isotopologue information of the individual elements stripped from the molecule.[9,11,13] Large subsets of these sub-isotopologues have then been calculated in vain, whereas a pruning strategy at earlier stages has not been presented. The latter bears potential to significantly accelerated calculations. In addition, the named approaches base their pruning on absolute probability thresholds,[9] proportions of least probable isotopologues[10] or truncate negligible mass states based on cumulative probabilities.[11,12] A relative probability threshold to be set as a fraction of the most dominant molecular isotopologue has neither been proposed nor evaluated – simply because this most probable isotopologue is not known in the first place. One can anticipate a relative pruning strategy to scale better with the widely differing dispersion of isotopologue probabilities among the analytes of interest.

On this background, we introduce a novel approach to compute isotope fine structures, augmenting the transition methodology introduced by Yergey[11] for directly calculating isotopologue from each other by single isotope replacements. Unlike his stepwise approach or the arrangements of isotopologues into levels of increasing mass proposed by Li et al.,[12,13] our method organizes transitions between isotopologues by gradually replacing the most abundant isotope of each element with variable subsets of less abundant ones. The underlying tree-like structure takes advantage of the multinomial nature of isotopologue transitions by separating branches of decreasing probability from increasing ones. We explicitly address how this structure can be used to (a) generate all feasible isotopologues, (b) avoid redundant transitions, (c) swiftly find the most probable isotopologues and (d) enable a very efficient pruning based on relative thresholds. We also elucidate the scaling properties of our approach with an unprecedented benchmark set of several thousand molecular formulas.

2.2 Methods

2.2.1 Rationale. In general, the MS signal of a molecule can be calculated by (1) dividing the molecule into sub-molecules for each element, (2) deriving the isotopologues within these sub-molecules, (3) combining these sub-isotopologues to the exact isotopologue probabilities and masses of the full molecule and (4) convoluting to measurable spectra at a given resolution

using peak-shape functions.[9,11,13] These steps will be exemplified for the sodium adduct $C_{20}H_8O_{10}Br_4S_2Na_1^+$ of the dye bromsulphthalein.

First, the molecule is divided into sub-molecules which contain only the n_k atoms of the *k-th* of a total of k_{max} elements, i.e., C, H, O, Br, S and Na at n_1=20, n_2=8, n_3=10, n_4=4, n_5=2 and n_6=1, respectively. In turn, $n_{k,i}$ stands for the number of atoms composed of a certain isotope *i* in the *k-th* sub-molecule; $a_{k,i}$ and $m_{k,i}$ are the natural abundance and the mass of this isotope, respectively. Sorted by decreasing abundance, *i=1* denotes the monoisotopic variant and often coincides with the lightest isotope of an element in organic compounds. In the given example, *k*=3 and *i=1* refers to ^{16}O. Similarly, $n_{3,1}$=10 is the monoisotopic sub-isotopologue of oxygen.

For each sub-molecule, transitions between any two sub-isotopologues *x* and *y* that differ in only one isotope allow for a simple updating of the probability $P_{k,y}$ and mass $M_{k,y}$ of *y* from the probability $P_{k,x}$ and mass $M_{k,x}$ of *x*:[11,12]

$$P_{k,y} = P_{k,x} \frac{n_{k,i}}{n_{k,j}} \frac{a_{k,j}}{a_{k,i}} \quad (1)$$

$$M_{k,y} = M_{k,x} - m_{k,i} + m_{k,j} \quad (2)$$

Herein, *i* denotes the isotope in *x* that is replaced by isotope *j* to produce sub-isotopologue *y*, with $i \neq j$. The exact position of the replaced isotope in a sub-molecule is irrelevant; each sub-isotopologue may consist of several isotopic isomers. For example, the transition of *k*=5, *i*=1 and *j*=2 is the replacement of one ^{32}S isotope by a ^{34}S isotope in the sulfur sub-molecule. Furthermore, each set of sub-isotopologues is initialized with one entry having probability $P_{k,1} = {a_{k,1}}^{n_k}$ and mass $M_{k,1} = n_k m_{k,1}$ based on the most abundant isotope *i*=1 only. Starting from this specific sub-isotopologue, transitions are recursively applied to devise the probabilities $P_k = \{P_{k,1}, \dots, P_{k,b_k}\}$ and masses $M_k = \{M_{k,1}, \dots, M_{k,b_k}\}$ of a relevant set of *1,…,b_k* sub-isotopologues per sub-molecule *k*. Notably, different transitions can produce the same sub-isotopologue. For instance, the sub-isotopologue $^{33}S_1{}^{34}S_1$ can be formed by transition from both $^{32}S_1{}^{33}S_1$ and $^{32}S_1{}^{34}S_1$. A methodology to avoid such computational redundancies will be proposed in a section on transition trees further below.

In the third step, sub-isotopologues from the different sub-molecules are combined via the k_{max}–fold Cartesian products on their probability and mass sets:

$$P_1 \times \cdots \times P_{k_{max}} = \{\left(P_{1,j_1}, \ldots, P_{k_{max},j_{k_{max}}}\right) : P_{i,j} \in P_i\} \quad (3)$$

$$M_1 \times \cdots \times M_{k_{max}} = \{\left(M_{1,j_1}, \ldots, M_{k_{max},j_{k_{max}}}\right) : M_{i,j} \in M_i\} \quad (4)$$

The resulting k_{max}-tuples of the product set contain all ordered combinations of sub-isotopologues from the different sub-molecules. To finally derive joint probabilities of the isotopologues of the full molecule, the entries within each tuple from equation (3) are multiplied. In contrast, entries within each tuple from equation (4) are summed to molecular isotopologue masses and corrected for electron masses if charged.
Notably, each molecule has a maximum of

$$b_{tot} = \prod_{k=1}^{k_{max}} \binom{q_k + n_k - 1}{n_k} \quad (5)$$

isotopologues, where q_k is the total number of isotopes of the *k-th* element.[10] The binomial coefficient defines the number of unordered combinations in which n_k atoms can be composed of q_k isotopes. Even for small molecules, b_{tot} can become prohibitively large and often contains a dominant fraction of isotopologues with negligible probability. Indeed, the small exemplary molecule has an isotope fine structure containing $b_{tot} \approx$ 6.2E5 isotopologues. However, over 99.9% of these isotopologues have probabilities at fractions of <1E-5 of the most abundant isotopologue. Strategies to omit such low- abundant (sub-)isotopologues at the level of both transitions and Cartesian product sets are therefore subject of another section on pruning.

Finally, peak shape functions of the remaining isotopologues are superimposed in a fourth step and scaled to relate to the isotopic envelopes of measured mass spectra at a certain instrument resolution R (Figure 1). Although this theoretical envelope is continuous, it is typically sampled at regular mass

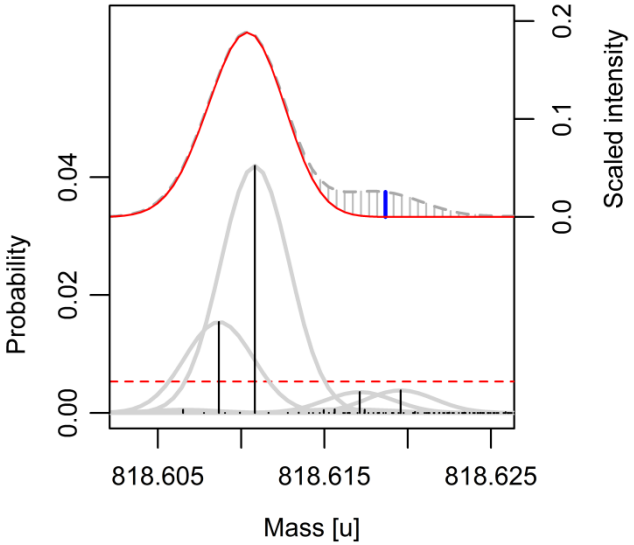


Figure 1. *The superposition of Gaussian peak shapes (grey lines) of individual isotopologues (black bars) produce an isotopic envelope (grey dashed line) to be sampled for its intensity at discrete mass intervals (grey bars), shown for the M+8 position of bromsulphthalein at a resolution of R=1.7E5. Pruning (β=0.2, red dashed line) leads to a distorted envelope (red solid line), with the maximum distortion indicated in blue. Without normalization to measured intensities, envelopes are rescaled to range within [0,1].*

intervals to match the discretized measurement spectra. Further details on deriving isotopic envelopes can be found elsewhere.[8,12]

2.2.2 Transition trees. The transitions of eq. (1) and (2) within each sub-molecule can be arranged into a tree-like structure, henceforth termed transition tree. These structures exclude redundant transitions and embrace all possible multicombinations of isotopes per element and thus the full set of sub-isotopologues. The transition trees of bromsulphthalein are shown in Figure 2. Each node in a transition tree represents a distinct sub-isotopologue. The root node at level z=1 contains the sub-isotopologue composed of only

the most abundant isotope i=1 from the ordered set of isotopes. Each edge symbolizes a transition, i.e. a replacement of one isotope i=1 with another less abundant isotope $j > i$. As a consequence, a terminal node is reached at level $z = n_k + 1$ when no further isotopes i=1 can be replaced, i.e., $n_{k,1} = 0$.

In addition, each node contains a single index I, with $1 < I \leq q_k$. This so-called transition index determines all valid transitions from a parent at any level z to its child nodes at level z+1 and must not be confused with a reference to a specific position of a node in the tree hierarchy. That is, all transitions with an isotope j under the restriction $I \leq j \leq q_k$ are feasible and must be conducted to grow the entire tree containing all possible sub-isotopologues. The transition index of a child node is in turn set to the j it was derived from and used in the very same manner for transitions to level z+2. Overall, all nodes and their transition indices are recursively developed from the root node, which is initialized at $I = 2$ if more than one isotope exists for a sub-molecule. Moreover, trees collapse to a simple sequence for elements with $q_k = 2$ (e.g., carbon) or contain none but the root node for $q_k = 1$ (e.g., sodium).

2.2.3 Pruning. The properties of transition trees allow for an efficient pruning of low-probable sub-isotopologues alias nodes, using either absolute or relative probability thresholds. For the latter case, probabilities are regarded insignificant if they range below a certain fraction $0 \leq \beta < 1$ of the highest probability P_{max} among the isotopologues of a molecule:

$$P_{max} = \prod_{k=1}^{k_{max}} P_{k,max} \tag{6}$$

where $P_{k,max}$ is the highest sub-isotopologue probability in each sub-molecule k. Consequently, the determination of all $P_{k,max}$ comes prior to any pruning, unless $\beta = 0$ is used to skip pruning. Each tree is therefore grown to reach its most probable node in a first phase. To this end, only transitions of increasing probability are made, starting from the root node. As can be seen in eq. (1), the ratio $P_{k,y}/P_{k,x}$ of a transition is driven by $n_{k,1}/n_{k,j}$, as $a_{k,j}/a_{k,1}$ is constant for any given isotope $j > 1$. Once $n_{k,1}/n_{k,j} < a_{k,1}/a_{k,j}$ is reached at a given transition, all subsequent transitions are monotonically decreasing in their probabilities for isotope j because $n_{k,1}$ can only decrease and $n_{k,j}$ can only increase with level z in a transition tree. A node-specific check of index I for reference to any transition of increasing probability with isotopes $I \leq j \leq q_k$ indicates if transitions to all child nodes must be made in this first phase.

Otherwise, all transition paths of higher level branching from this parent node are of decreasing probability; the concerned node is then made temporarily dormant during this first phase. Moreover, the ordering of isotopes by their decreasing natural abundance helps to phase out branches of solely decreasing probability at low tree levels. In small molecules, the root node is often the most abundant sub-isotopologue, when $n_{k,1}/n_{k,j} < a_{k,1}/a_{k,j}$ holds for all isotopes $j > 1$. A second phase then progresses with the dormant transitions of strictly decreasing probabilities after $P_{k,max}$ has been established for each sub-molecule (grey nodes in Figure 2). While doing so, all transitions from nodes with probability $P_{k,x}$ fulfilling

$$P_{k,x} \frac{\prod_{k=1}^{k_{max}} P_{k,max}}{P_{k,max}} < \beta P_{max} \quad (7)$$

or, after division by P_{max},

$$P_{k,x} \,/\, P_{k,max} < \beta \quad (8)$$

are halted and downstream nodes pre-pruned, even if no leaf nodes have been reached yet. The fraction in the left-hand side of equation (7) scales $P_{k,x}$ in sub-molecule k to the maximum probability it can attain in the later Cartesian product by setting probabilities in all other sub-molecules to their maximum. With no further transitions remaining, sub-isotopologue nodes that do not fulfill equation (8) are finally post-pruned in a third phase.

The procedure outlined above simplifies for absolute probability thresholds, as no $P_{k,max}$ need to be searched for in a first phase. Instead, all transitions from nodes indexing at least one transition of increasing probability have to be made. Transitions from nodes with probabilities below the absolute threshold and indexing only decreasing probabilities can be omitted. With no further transitions remaining, any developed nodes below the absolute threshold are again post-pruned. Overall, a large set of isotopologues can be pruned in most trees (Figure 2, dashed nodes), greatly reducing the computational burden.

Further pruning can be applied during a third phase of combining the sub-isotopologues from different sub-molecules via their Cartesian products. Computationally, the joint probabilities are calculated by multiplying each

sub-isotopologue probability of the first sub-molecule with each of the second sub-molecule. In turn, the resulting products are each multiplied with the sub-isotopologue probabilities of the third sub-molecule. This iteration continues until the last remaining sub-molecule has been included. The analogous summation of sub-isotopologue masses across the sub-molecules leads to the isotopologue masses of the full molecule.[9] After each of the $l=\{1,\ldots,k_{max}\text{-}1\}$ iterations, all products x with probability $P_{l,x}$ either fulfilling

$$P_{l,x} \,/\, \prod_{k=1}^{l+1} P_{k,max} < \beta \qquad (9)$$

for relative thresholds or ranging below the absolute probability threshold can be discarded, together with their related masses.

2.2.4 Parameter evaluation. The presented algorithm, henceforth called *enviPat*, was first used to elucidate differences between relative and absolute pruning thresholds on the distortion of isotopic envelopes. For this purpose, a total of 5969 unique organic molecular formulas were randomly sampled in mass bins from the PubChem database[14] and their theoretic envelopes simulated for a combination of four complementary resolution functions (Thermo Orbitrap Elite, *R=240,000@m/z400* and *60,000@m/z400* / Thermo Orbitrap Velos Pro, *R=7500@m/z400* / Waters G2 QToF *R=25,000@m/z200*) with eight different absolute and relative thresholds each. Every simulation was then compared to an approximation of the corresponding unpruned envelope, after rescaling each envelope to its most intense signal, i.e., to range in [0,1]. The distortion was then recorded as the maximum intensity difference between the two scaled envelopes over all discretization points. An illustration is given in Figure 1. This approach imitates common practice in which a pruned simulated envelope is compared to an unpruned measured envelope after the former was scaled to the latter using the most intense signal. The approximated unpruned envelope contained a minimum cumulative probability ≥0.9995 from using a very low absolute probability threshold of 1E-20. The mass discretization of all envelopes was dynamically adjusted for each formula and resolution to 1/10 the width at half maximum of a Gaussian peak shape function.

2.2.5 Performance comparison. *enviPat* was compared to two previous approaches of isotope fine structure calculation, *Isotope Calculator*[13] and *ecipex*,[9] using the above mentioned benchmark set of molecular formulas. *ecipex* has only recently been published and also yields isotopic fine struc-

tures but uses convolutions via multidimensional Fourier-transforms to generate all sub-isotopologues for each sub-molecule. Similar to *enviPat*, *Isotope Calculator* utilizes transitions between sub-isotopologues, but organizes them into mass states. Sub-isotopologues are finally combined to form the isotopologues of the full molecule in all these approaches. Memory usage was therefore quantified by the sum of these intermediate sub-isotopologues and is hence independent of the specificities of memory allocation in the various implementations. For *envi-Pat*, this amounts to the number of pruned sub-isotopologues established at the end of phase two when growing the transition trees, summed over all sub-molecules. With no pruning available at this early stage, the full count of sub-isotopologues must be reported for both *ecipex* and *Isotope Calculator*. This clearly states a lower bound of memory requirements for *ecipex*, which in fact needs $(n_k + 1)^{q_k - 1}$ intermediate terms to be summed over all *k* sub-molecules. For a runtime comparison, we only selected *ecipex* because it has already shown to outperform *Isotope Calculator* in this aspect.[9] Furthermore, both *ecipex* and *enviPat* are both implemented in the R statistical environment,[15] making a comparison more consistent.[16] For time performance, each approach was averaged over 200 repetitions for each formula, including an initial garbage collection step, a timeout of 2 s for a single isotope pattern calculation and an upper limit of 2E7 sub-isotopologues. Because *ecipex* does not allow for relative thresholds, an absolute threshold from the above parameter evaluation was selected for both approaches to calculate equal numbers of isotopologues for each formula. Calculations were run in R version 3.0.2, on a workstation with a 2.5GHz Intel Core i5-3210M Processor and 8 GB RAM under Ubuntu 14.04, 64 bit.

2.3 Results & Discussion

The evaluation of isotopic fine structure calculations has to date been based on relatively small sets of carefully selected molecular formulas, with at most ten formulas.[9] In contrast, our randomly sampled benchmark set comprises several thousand molecular formulas from 30 to 25,000 u, including large organometallic molecules with polyisotopic elements such as Zn, Ti or Gd. The size and complexity of this benchmark set enhances both the representativeness of our findings and the detection of scaling issues below.

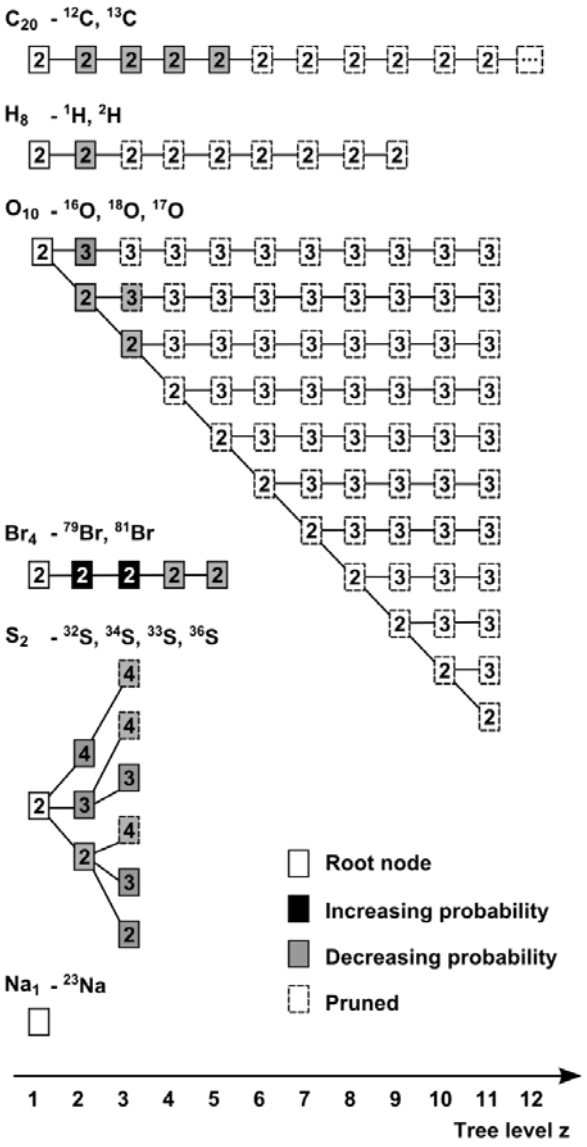

C20 - 12C, 13C
H8 - 1H, 2H
O10 - 16O, 18O, 17O
Br4 - 79Br, 81Br
S2 - 32S, 34S, 33S, 36S
Na1 - 23Na
Root node
Increasing probability
Decreasing probability
Pruned
1 2 3 4 5 6 7 8 9 10 11 12
Tree level z

Figure 2 (last page). *Transition trees for the sub-molecules of the sodium adduct of bromsulphthalein at β=1E-5. Each node represents a sub-isotopologue and lists its transition index I, which points to the ordered subset of isotopes used for its transitions to child sub-isotopologues.*

2.3.1 Parameter evaluation. Pruning of isotopologues leads to changes in the simulated isotopic envelope of a molecule and thereby to deviations from its measured and hence unpruned envelope after the simulated envelope is rescaled to match the latter by its most intense signal. Figure 3 shows this distortion for the benchmark formulas and a range of threshold values covering several orders of magnitude. Listed as the maximum difference in scaled intensity and hence constituting a fractional quantity, the absolute intensity difference between simulated and measured envelopes at a given distortion increases with increasing signal intensity. In general, an approximately log-linear relationship can be observed between the median fractional distortion and the threshold values, both for relative and absolute thresholds. The variation in distortion at a given threshold value however differs between thresholds types, pooled over the different resolutions and our extensive benchmark set of molecular formulas. Namely, the overall spread in distortion is smaller when using relative instead of commonly applied absolute thresholds. This translates to a less skewed distortion on a linear scale, i.e., a much lower tendency to generate outliers of imprecise isotopic envelopes for the single threshold choice to be made. In contrast to an absolute choice, a single relative threshold relates to widely differing cutoff probabilities βP_{max} among the different molecular formulas (data not shown). For molecules with a large spread in probability among isotopologues this cutoff is often lower and hence permits a larger number of relevant isotopologues to be included. For example, at β=1E-5, bromsulphthalein requires the calculation of 601 isotopologues, with P_{max}=0.265. On the contrary, the much larger bovine insulin, $C_{254}H_{377}N_{65}O_{75}S_6$, needs 5456 isotopologues to be calculated, using the same relative threshold for P_{max}=0.113.

2.3.2 Performance comparison. We selected an absolute probability threshold of 1E-9 for a performance comparison from the above parameter evaluation. The concomitant maximum distortion of 6.7E-5 (median: 1E-7, cp. Figure 3) lies far below the intensity uncertainties of available MS instrumen-

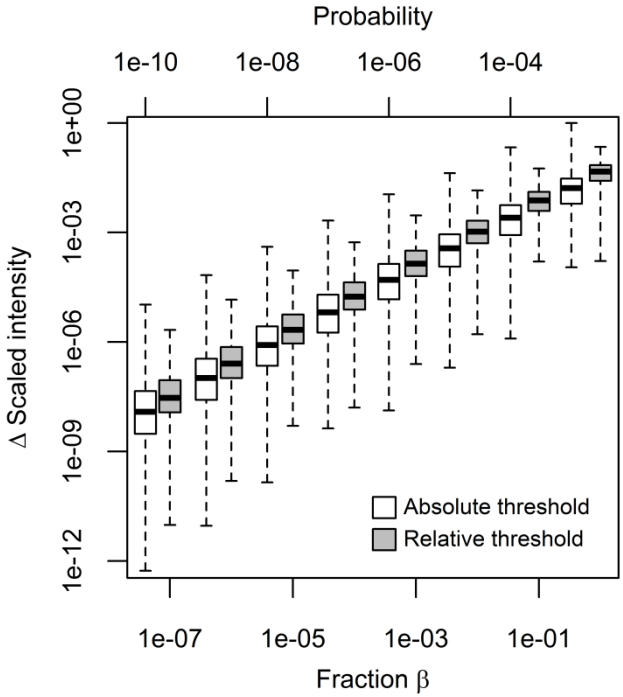


Figure 3. *Difference in scaled intensity (distortion) of theoretical envelopes for different pruning thresholds, pooled over the four instrument resolutions and the benchmark set of molecular formulas (boxplots indicate the extremes, the median and the lower and upper quartiles). Relative thresholds are given as fraction β of the most probable isotopologue, whereas absolute thresholds refer to cutoff probabilities. Note the log-scale of all axes.*

tation.[17] In any case, the general findings below were not compromised by alternative threshold choices. In combination with the chosen time and memory constraints, computation failed for only 11 and 69 of the benchmark formulas for *enviPat* and *ecipex*, respectively. The total number of molecular isotopologues given by equation (5) stretches over 14 orders of magnitude for

the remaining formulas and thus sheds light on the performance scaling with increasingly complex molecules.

First, the number of sub-isotopologues assembled in intermediate calculations increases nonlinearly with the total number of isotopologues in a molecule. In the unpruned case, the former number is approximately three times lower in orders of magnitude relative to the latter, although this ratio increases for smaller molecules (Figure 4, a). The variation in this relationship generally grows with more complex molecular formulas. In contrast, *enviPat* drastically reduces both the number of sub-isotopologues and its variation (Figure 4, b). While the mentioned ratio is still similar for small molecules, it drops stronger with rising molecular complexity than in the unpruned case. For the most complex molecules in our benchmark set, the number of sub-isotopologues can be as much as six times lower in orders of magnitude than the total number of isotopologues. Similarly, sub-isotopologue counts exceeded the numbers of molecular isotopologues above the given threshold in 37.5% of the cases (data not shown). This occurred for only 0.1% of cases for *enviPat*. Such favorable scaling is a result of the aggressive pruning which *enviPat* features. As demonstrated in Figure 2 for bromsulphthalein, branches containing large fractions of sub-isotopologues can be ignored - solely based on the probability, isotopic composition and transition index of their parent node. In this way, sub-isotopologue differences of up to two orders in magnitude can be achieved in comparison to the unpruned approaches. This does not imply that all developed sub-isotopologues range above their threshold probability. Being directed by a single transition index per node, transitions of increasing probability can be accompanied by a set of decreasing and thus irrelevant ones, which have to be post-pruned. Omission of tree branches will also be less efficient for molecules dominated by elements with several isotopes of similar natural abundance, explaining the few outliers of Figure 4 (b) containing, e.g., Pt.

The overall numbers of sub-isotopologues will seldom exceed memory resources, even for the observed outliers. Unnecessary calculations and subsequent post-pruning will rather affect computational runtimes, adding to methodological differences in deriving sub-isotopologues among the approaches. Indeed, *enviPat* outpaces *ecipex* for all benchmark molecular formulas, as shown in the bottom panels of Figure 4. Even for small molecules with similar numbers of sub-isotopologues in both approaches, *ecipex* computations take around 5 ms longer. One may argue that this constant overhead may result from implementation differences in e.g. molecular formula

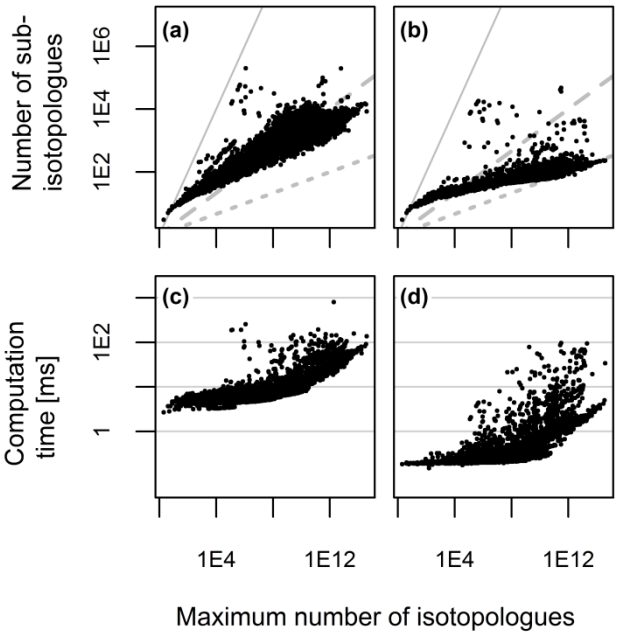


Figure 4. *Number of (a) unpruned and (b) enviPat-pruned sub-isotopologues and computation time for (c) ecipex and (d) enviPat plotted against the total isotopologue number of each molecule in the benchmark set (axes are log-scaled). Grey lines in the top plots indicate numbers of sub-isotopologues being equal (solid) or 3 (dashed) and 6 (dotted) times smaller in orders of magnitude than the total number of isotopologues.*

parsing or memory allocation instead of conceptual distinctions, giving *enviPat* an unfair advantage. However, this overhead would be small in relation to the logarithmic runtime differences between the two algorithms for more complex molecular formulas. Thus, even when subtracting the overhead from the computation time of *ecipex*, *enviPat* still consumes around one magnitude less computation time for molecules with more than 1E6 isotopologues.

2.3.3 Transition tree properties. With *enviPat*, the accuracy of an isotopologue mass and probability is only limited by the computational precision and rounding errors. The latter increase with the number of calculation steps needed to calculate sub-isotopologues. For elements with more than two isotopes, e.g., the oxygen and sulfur of Figure 2, the branching within trees results in fewer steps to reach a node by transitions from the root node than a purely sequential updating, on average. But even for elements with two isotopes, for which trees simplify to sequences, probability differences between a node at level z and its parent at level z-1 decrease monotonically with z. Any significant rounding errors will thereby accumulate in the pruned tails of these sequences long after the most probable sub-isotopologue was detected. A similar situation arises for isotopes with variable natural abundances (for instance, boron), where uncertainties in the probability of an isotopologue increase with the number of transitions. Here, finding the most dominant isotopologues within a small number of calculation steps is important to minimize uncertainties of the final envelope.

Several other properties of transition trees need to be stressed to direct future implementations. First, no information other than the state (i.e., mass, probability, isotopic composition) and transition index of a node is required to develop its transitions. The history of tree growth or the state of other nodes is irrelevant. This leaves a great potential for parallelization of the tree growth procedure ab initio (absolute threshold) or after the most abundant isotopologue has been detected (relative threshold). Similarly, parent nodes of undeveloped tree branches can be saved to hard disc for later evaluation under very stringent RAM conditions or for extremely large molecules. Second, transition trees can be built without a division into sub-molecules. Instead, transition indices can reference the full set of isotopes over all elements in a molecule - excluding those of highest abundance per element, which are used to build the root node. Tree nodes do then no longer represent sub-isotopologues of sub-molecules, but directly the isotopologues of the full molecule. Herein, a different metric for sorting the referenced set of isotopes should be considered for quickly pruning branches of decreasing probability, e.g., by the probability ratio P_y/P_x of the current transitions from node x to any of its children y. Alternatively, divisions can be made into sub-molecules containing more than one element each, too. The above aspects can also be integrated to directly superimpose the peak shapes of molecular isotopologues after they developed their transitions, without a need to store them. This would drastically shrink the memory requirements for envelope calcula-

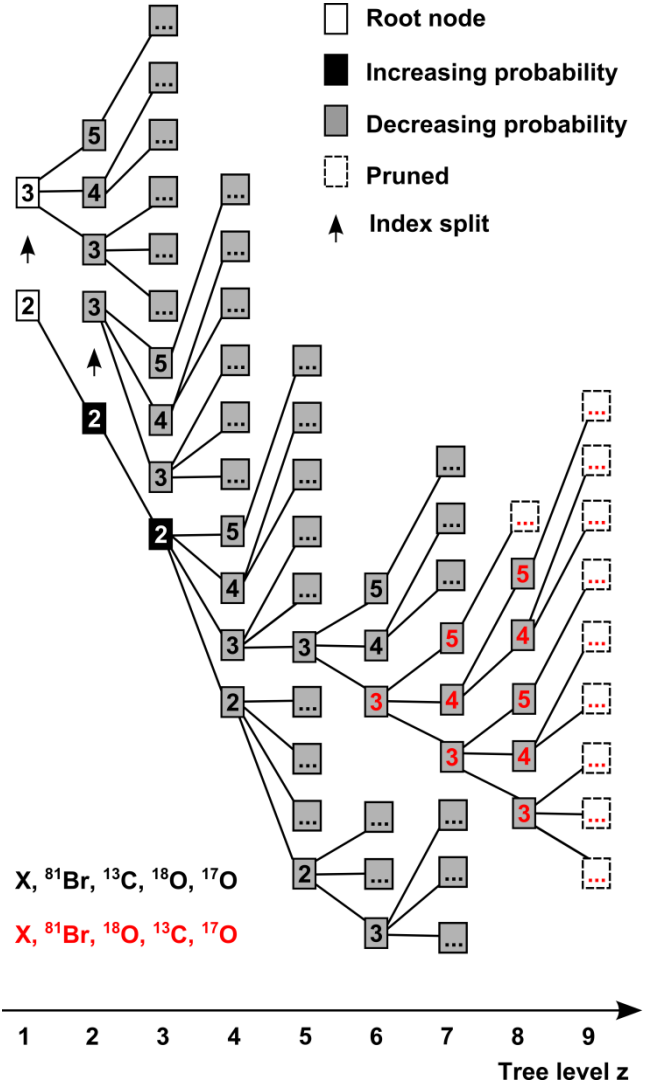

Root node
Increasing probability
Decreasing probability
Pruned
Index split
X, 81Br, 13C, 18O, 17O
X, 81Br, 18O, 13C, 17O
1
2
3
4
5
6
7
8
9
Tree level z

***Figure 5 (last page)**. Transition tree for the submolecule $C_{20}O_{10}Br_4$ of bromsulphthalein at β=1E-8. Here, transition indices of nodes refer to the joint set of all isotopes of the three elements, with X representing a single entry for the monoisotopic isotopes to be replaced. Non-monoisotopic isotopes are initially sorted by decreasing probability ratios P_y/P_x (black transition indices) and can be resorted for individual branches after several transitions (red indices). Moreover, arrows indicate a split of the transition index to separate transitions of increasing probability from decreasing ones.*

tions. Third, the order of an isotope subset referenced by transition indices does not need to be static. Rather, a set can be copied and reordered to meet the specificities of different branches, especially if transition trees for more than a single element may be built. Fourth, indices can be split. For example, given the index I=2 at q_k =5, only the transitions 2 of, e.g., increasing probability can be made, instead of the full referenced set 2,…,5. The concerned node is then set to I=3 for later evaluation of transitions with decreasing probability. Some of these properties are further illustrated in Figure 5.

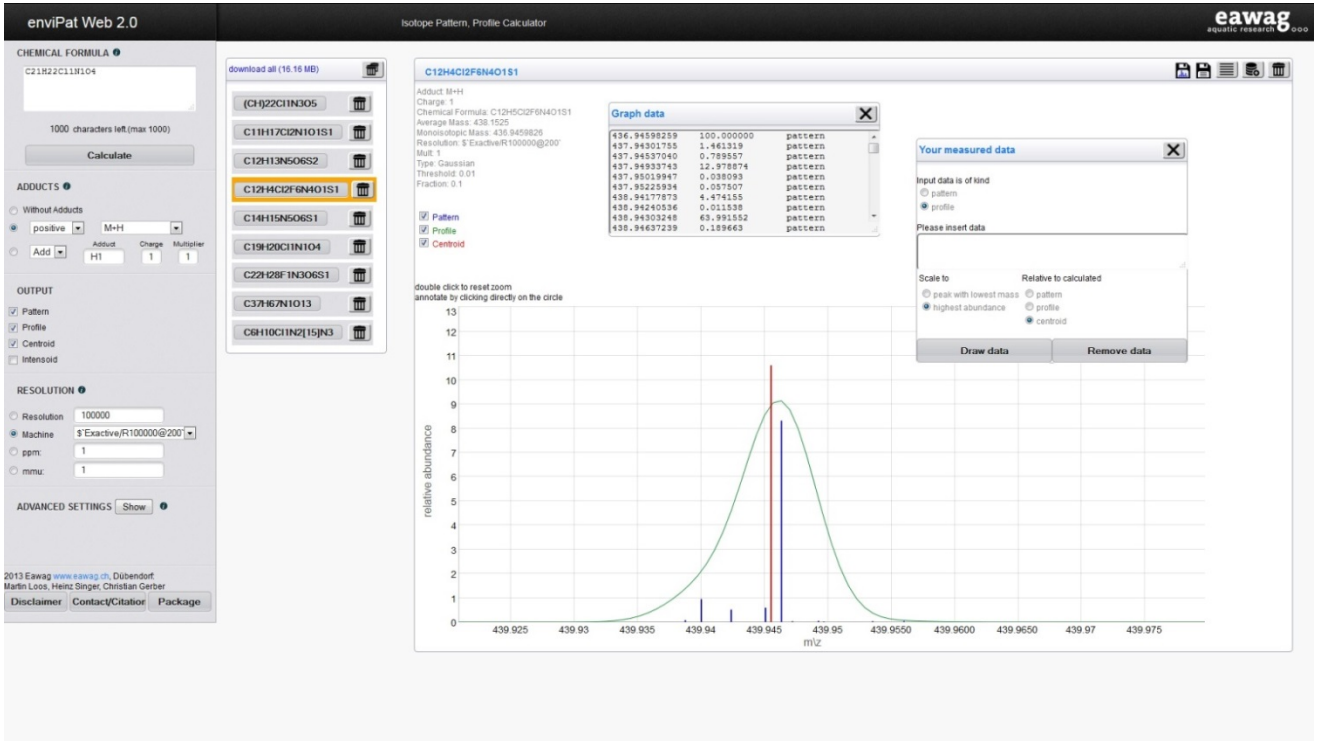
2.4 Implementation

The presented algorithm is freely available as R package enviPat. The package provides instrument-specific envelope and centroid calculation, batch processing and molecular formula parsing for a variety of adducts commonly formed during electrospray ionization (ESI). User defined inputs cover absolute and relative pruning thresholds, charge states, resolution-dependent envelope discretization and enriched isotope tables, among others. To facilitate its usage, all enviPat functionalities and simultaneous comparison with measured data can be conveniently accessed from a web-based user interface at www.envipat.eawag.ch, as shown in Figure 6.

2.5 Conclusion

Comparatively few strategies have been proposed to derive exact isotope fine structures for the simulation of mass spectrometric signals. While supporting the interpretation of highly resolved MS data, such strategies must be able to efficiently prune large fractions of low-probable isotopologues. We therefore introduced a new hierarchy to calculate child from parent isotopologues by single replacements of one most abundant isotope by a less abundant one, guided by a node-specific transition index. As opposed to previous hierarchies, our tree-like structure permits a precise pruning of large but low-probable isotopologue branches, solely based on the state of a single isotopologue at the root node of such branches. In addition, calculations can for the first time be directed to derive the most probable isotopologue in a first phase. Its probability can then be used for a relative instead of an absolute pruning threshold in a second phase of remaining calculations. On the one hand, this relative pruning adapts to the highly heterogeneous dispersion of isotopologue probabilities among the diverse and often unknown analytes. On the other hand, pruning of substantial isotopologue branches leads to a computational performance that dwarfs the runtime required to convolute the resulting isotopologues to their measurable envelopes in practical applications. Hence, and despite several yet unexploited properties of our so-called transition trees, this speed limiting step should best be addressed next.

***Figure 6 (next page)**. Web interface for using enviPat in a browser.*


enviPat Web 2.0
Isotope Pattern, Profile Calculator
eawag
aquatic research
CHEMICAL FORMULA
C21H22Cl1N1O4
1000 characters left.(max 1000)
Calculate
ADDUCTS
Without Adducts
positive
M+H
Add
H1
Charge
Multiplier
OUTPUT
Pattern
Profile
Centroid
Intensoid
RESOLUTION
Resolution
100000
Machine
$'Exactive/R100000@200'
ppm:
mmu:
ADVANCED SETTINGS
Show
2013 Eawag www.eawag.ch, Dübendorf
Martin Loos, Heinz Singer, Christian Gerber
Disclaimer
Contact/Citation
Package
download all (16.16 MB)
(CH)22Cl1N3O5
C11H17Cl2N1O1S1
C12H13N5O6S2
C12H4Cl2F6N4O1S1
C14H15N5O6S1
C19H20Cl1N1O4
C22H28F1N3O6S1
C37H67N1O13
C6H10Cl1N2[15]N3
C12H4Cl2F6N4O1S1
Adduct: M+H
Charge: 1
Chemical Formula: C12H5Cl2F6N4O1S1
Average Mass: 438.1525
Monoisotopic Mass: 436.9459826
Resolution: $'Exactive/R100000@200'
Mult: 1
Type: Gaussian
Threshold: 0.01
Fraction: 0.1
Pattern
Profile
Centroid
Graph data
436.94598259 100.000000 pattern
437.94301755 1.461319 pattern
437.94537040 0.789557 pattern
437.94933743 12.978874 pattern
437.95019947 0.038093 pattern
437.95225934 0.057507 pattern
438.94177873 4.474155 pattern
438.94240536 0.011538 pattern
438.94303248 63.991552 pattern
438.94637239 0.189663 pattern
Your measured data
Input data is of kind
pattern
profile
Please insert data
Scale to
peak with lowest mass
highest abundance
Relative to calculated
pattern
profile
centroid
Draw data
Remove data
double click to reset zoom
annotate by clicking directly on the circle
relative abundance
m/z
439.925
439.93
439.935
439.94
439.945
439.95
439.9550
439.9600
439.9650
439.97
439.975

REFERENCES

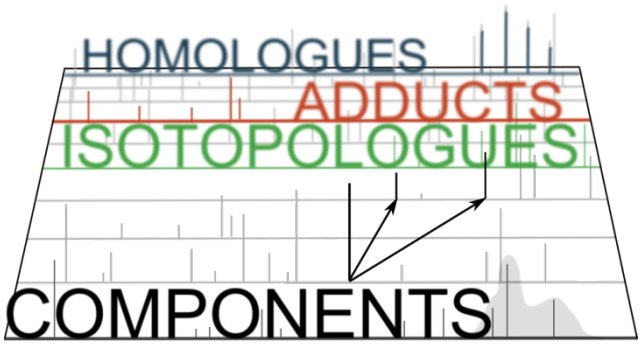
(1) Schymanski, E.L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H.P., Hollender, J. ES&T 2014, 48, 2097-2098.
(2) Roussis, S.G., Proulx, R. Anal. Chem. 2003, 75, 1470-1482.
(3) Krauss, M., Singer, H., Hollender, J. Anal. Bioanal. Chem. 2010, 397, 943-951.
(4) Kenar, E., Franken, H., Forcisi, S., Wörmann, K., Häring, H.U., Lehmann, R., Schmitt-Kopplin, P., Zell, A., Kohlbacher, O. Mol. Cell. Proteom. 2013, 13, 348-359.
(5) Rockwood, A.L., Kushnir, M.M., Gordon, J.N. J. Am. Soc. Mass Spectrom. 2003, 14, 311-322.
(6) Valkenborg, D.; Mertens, I.; Lemière, F.; Witters, E.; Burzykowski, T. Mass Spectrom. Rev. 2012, 31, 96−109.
(7) Scheubert, K.; Hufsky, F.; Böcker, S. J. Cheminf. 2013, 5:12.
(8) Werlen, R. C. Rapid Commun. Mass Spectrom. 1994, 8, 976-980.
(9) Ipsen, A. Anal. Chem. 2014, 86, 5316-5322.
(10) Snider, R.K. Am. Soc. Mass Spectrom. 2007, 18, 1511-1515.
(11) Yergey, J. A. Mass Spectrom. 1983, 52, 337-349.
(12) Li, L.; Kresh, J. A.; Karabacak, N. M.; Cobb, J. S.; Agar, J. N.; Hong, P.J. J. Am. Soc. Mass Spectrom. 2008, 19, 1867−1874.
(13) Li, L.; Karabacak, N. M.; Cobb, J. S.; Wang, Q.; Hong, P.; Agar, J. N. Rapid Commun. Mass Spectrom. 2010, 24, 2689−2696.
(14) Bolton, E.; Wang, Y.; Thiessen, P.A.; Bryant, S.H. Annu. Rep. Comput. Chem. 2008, 4, 217-241.
(15) R Core Team. R Found. for Stat. Comp. 2014.
(16) Hu, H., Dittwald, P., Zaia, J. Valkenborg, D. Anal. Chem. 2012, 84, 7052-7056.
(17) Guerrasio, R., Haberhauer-Troyer, C., Steiger, M., Sauer, M., Mattanovich, D., Koellensperger, G., Hann, S. Anal. Bioanal. Chem. 2013, 405, 5133-5146

Chapter 3

Nontargeted Peak Grouping for Chemical Component Detection in Liquid-Chromatography Mass Spectrometry Data

Martin Loos, Juliane Hollender, Francesco Corona, Heinz Singer

Paper draft



Abstract The analysis of small polar molecules in environmental samples relies heavily on mass spectrometry (MS), frequently coupled to liquid chromatography (LC) and electrospray ionization (ESI). Herein, elements such as carbon, nitrogen or chlorine combine to distinct isotopologues for a molecule, but are impossible to predict for an unknown LC-ESI-MS analyte. A post-acquisition grouping of these isotopologue LC-ESI-MS signals is highly beneficial, yet complicated by the diverse range of analytes and the superposition of each their isotopologue signals at mass-dependent resolutions. Hence, a first large-scale simulation of more than $2x10^6$ groups of centroided isotopologue signals is presented, using simplified classes of analyte ions (adducts) and three complementary Orbitrap high resolution (HR) settings. Even under relational restrictions, (a) shifts in mass differences of more than 30 ppm from theoretical isotope transitions, (b) complex mass-dependent bounds and ranges of intensity ratios and (c) pronounced specificity to each of the three resolutions were observed when linking pairs of the simulated centroids. The multidimensional characteristics of these linkages were thereupon discretized via an unsupervised partitioning algorithm into easy-to-query data structures, which can be used to group the centroids of unknown analytes. The performance of this very grouping was in turn validated with a large test set of simulated compounds, with recall values above 0.88 even under aggravated conditions; false negative linkages were mostly restricted to few adducts and low-intense centroids. Consequently, validation with identified target and spiked isotope-labelled standard compounds in samples of sewage treatment plant (STP) effluent lead to a full recall, at precisions of 0.90±0.03 and 0.88±0.03, respectively. After a second grouping for major adducts, a mean fraction of 0.61±0.02 of all centroid peaks detectable in STP samples could be assorted into nontarget components of more than one peak, equal to an average 1.8–fold data reduction. The approach is publicly available in the R package *nontarget*[1] and offers the first tool to combine isotopologue, ESI adduct and also homologue series grouping for peak-picked LC-HRMS data.

3.1 Introduction

The trace level detection, identification and quantification of small polar molecules is pivotal in research fields such as metabolomics and environmental monitoring. As a method of choice, high-resolution mass spectrometry (HRMS) has been widely applied for the joint high-throughput analysis of both known and unknown compounds.[2–4] When coupled to liquid-chromatography (LC) and soft electrospray ionization (ESI), individual analytes herein co-elute as signal groups of mass-to-charge (*m/z*) ratios of varying intensity, characteristic of their isotopologue composition and ion species (broadly termed adducts henceforth; including abstractions, dimers or doubly charged species). Unfortunately, there is no prior indication and – in nontargeted approaches – little expectation as to which of these signals should group together; several thousand analytes are often processed simultaneously and a manual grouping of the manifold of LC-HRMS signals to their chemical components is hence infeasible. Such grouping is nonetheless highly beneficial. First, the ensuing data reduction helps to cope with the large number of LC-HRMS signals. Similarly, grouping avoids redundancies in data interpretation and priorization.[5] Third, grouping facilitates blind subtraction steps. Fourth, isotopologue grouping identifies decharged monoisotopic masses, e.g., to track these masses across different samples.[6] Fifth, isotopologue groups are instrumental in restricting candidate molecular formula fits for these monoisotopic masses; as is the grouping of adducts signals to appoint the adduct ion and charge state during unknown identification.[7–10] Sixth, isotopologue groups can help decoding dissociation spectra in MS^n experiments.[11,12]

Automated isotopologue grouping has to date mostly relied upon three largely orthogonal methods, either separately or in combination: (1) similarity between chromatographic elution shapes in one sample,[13–16] (2) correlation of intensity variations of signals across different samples, conditions or replicates,[9,15–19] or (3) recognition of specific *m/z* differences and relative intensity characteristics among signals detected within similar retention time (*RT*) windows in one sample.[20–22] The first aspect must access raw LC-HRMS data and is thus incompatible with the usage of picked centroid peaks alone. The second aspect is restricted to the availability of several LC-HRMS data sets to compare and requires signals to range sufficiently often above their limit of detection (LOD). In addition, the three aspects are only partly complementary; aspect (2) is necessary, but not sufficient, for aspects (1) or (3) to hold. Aspect (3) is indeed most intricate to derive. Whereas chemical compound classes such as peptides,[23–25] saccharides,[26] lipids[27] or hydrocarbons[28] can each be approximated by averaged or simplified isotopologue models, the elemental composition of the full universe of small molecules is too diverse to be subsumed into averagines. More specifically, the differing combinations of polyisotopic elements and their varying atom counts lead to strongly varying isotopologue masses and probabilities.[29,30] To make matters

more complicated, these individual isotopologues are rarely fully resolved by LC-HRMS acquisitions, as exemplified in panel B of Figure 1. Instead, the isotopologue signals of a compound superimpose to a measurable envelope, dependent on the adduct species, its charge and on specific instrument resolutions which in turn deviate as a function of *m/z*.[31] It is noteworthy that standard instrumentation can nowadays achieve separation of such envelopes beyond mere nominal or accurate mass resolutions. Despite widespread claims to be tailored to high resolution, hitherto proposed algorithms have neither concisely described nor grouped these complex signals for small molecule applications.

To this end, an overdue strategy to restrain the feasible space of intra-group characteristics of centroided isotopologue envelopes is presented. After approximation of this space by a tailored discretization method, measured isotopologue centroids can thus be grouped by their pairwise linkages, while subject to differing measurement uncertainties and varying proportions of low-probable centroids falling below the LOD. After thorough validation, the strategy is complemented by assorting the different adducts of a compound and to finally assemble the chemical components from the vast sets of LC-HRMS centroid signals.

3.2 Methods

A grouping of centroided LC-HRMS data peaks (i.e., peak-picked lists) resulting from the different isotopologues of the same compound adduct is derived in seven stages. These are, **(1)** a large-scale simulation of isotopologues, **(2)** a convolution of the latter to centroid peaks, **(3)** a systematic extraction of pairwise centroid peak links, **(4)** a characterization of these links, **(5)** a discretization step to represent these characteristics by hyperrectangles, **(6)** organization of these rectangles into a metric data structure for efficient queries and validation with both **(7)** simulated test data and **(8)** measured and picked signal peaks. Stages (1) to (7) were completed for each of three different instrument resolution functions *R(m/z)*=70.000 (70K), 140.000 (140K) and 280.000 (280K) at *m/z*=200 applicable to Thermo Orbitrap Q-Exactive mass spectrometers, whereas stage (8) was performed for the medium resolution *R(200)*=140K function only. It must be highlighted that resolutions decreased with *m/z*, differently for each resolution function. With stages (1) to (6) outlined in the succeeding sections**,** the full componentization further embracing adduct and homologue series relations among peaks is presented thereafter, followed by details on method validation and experimental settings. All calculations were run in the R statistical environment, using specific packages as indicated.[32]

3.2.1 Centroid linkages. The first stage uses a set of $n=8.3\text{x}10^4$ unique molecular formulas of organic compounds to simulate their isotopologue fine structure. Compound molecular formulas were randomly sampled from all

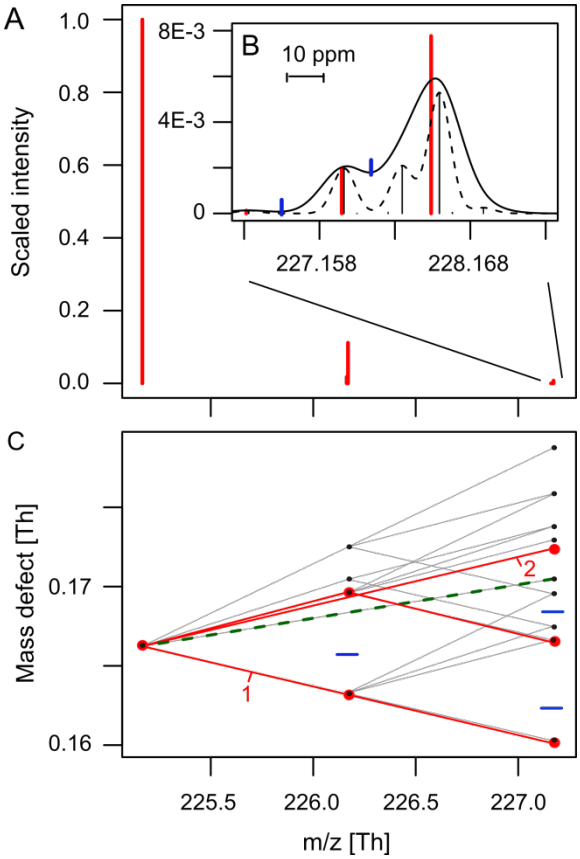


Figure 1. *Isotope pattern (black bars and dots) and resulting centroid peaks (red bars and dots) at a resolution of R(226)=65K for the positively charged pesticide Terbumeton, $C_{10}H_{19}N_5O_1^+$. While panel A depicts the full range of centroids, inset B details the M+2 position, with both the corresponding envelope (black solid profile) and another one calculated at higher R(226)=130K (black dashed profile). Panel C depicts relations: gray lines connect isotopologues which can be formed by single-isotopic transitions, whereas red lines highlight assigned linkages between centroid pairs, e.g., via a ^{14}N to ^{15}N transition (linkage 1) or a ^{16}O to ^{18}O transition (linkage 2, with the transition highlighted as green dashed line). Blue bars indicate minima (i.e., valleys) in the underlying envelope. The ordinate mass defect in the lower panel is solely used to distinguish data points of equal nominal mass.*

available instances in the PubChem database containing carbon and any of the relevant elements H, N, O, Cl, Br, K, Na, S, Si, F, P and/or I.[33] To cover both varying charges and addition of atoms during ESI adduct formation, each formula was then modified by a set of 27 adduct classes listed in table S-1. These classes were formed by a simplification of a comprehensive list of 49 adducts observable during positive and negative ESI.[34] In simplifying, elements with only one isotope (Na, F) or low natural abundance of their non-monoisotopic isotopes (H) were dropped from this list, allowing to merge several entries. Assuming that changes in resolution caused by the concomitant mass shifts as well as the influence of low-probable ^{2}H-containing isotopologues are both negligible, usage of adduct classes strongly decreased the number of required simulations as compared to calculations with the full list of adducts. Accelerated by a recently developed transition tree approach, all isotopologues above a fractional probability threshold $\beta_1 \geq 1x10^{-6}$ relative to the most probable isotopologue were then calculated for every adduct class of each molecular formula. It has been shown that distortions resulting from omission of isotopologues below this β_1 value are negligible over a wide array of compounds.[29]

In a second stage, isotopologues of every adduct class formula were combined to their so-called isotopic envelope and the latter thereupon converted to centroid peaks. The envelope results from a superposition of Gaussian peak shape functions centered around each isotopologue *i* with *m/z* value M_i and probability P_i, to be sampled at regularly spaced values m/z'. $\bar{M}$ denotes the mean *m/z* value over all M_i. For a given resolution function $R(\bar{M})$ defined as the ratio of $\bar{M}$ to the peak shape width at half maximum, this superposition is given by

$$f(m/z') = \sum_{i=1}^{k} P_i e^{\left(-\frac{(m/z'-M_i)^2 R(\bar{M})^2 \ln 256}{2M_i^2}\right)} \quad (1)$$

with k being the total number of isotopologues per formula and the terms of the Gaussian variance substituted through $R(\bar{M})$.[35] Differences $R(\bar{M}) - R(M_i)$ over the range of M_i per formula and adduct class are assumed to be negligible. For sufficient approximation, spacing of *m/z*' values was dynamically adjusted to 1/20 the half maximum width of the peak shape function at any $R(\bar{M})$. Based on the resulting envelopes, local minima in $f(m/z')$ were used to split the envelope into *1,…,j,…,n* disjoint but ordered areas and the centroid *m/z* and intensity calculated for each of them, denoted as $C_j^{m/z}$ and C_j^{int}, respectively. More precisely, $C_j^{m/z}$ is the weighted sum of *m/z'* values between two consecutive minima positions a_j and b_j of centroid j:

$$C_j^{m/z} = \frac{\sum_{i=a_j}^{b_j} m/z'_i \, f(m/z'_i)}{\sum_{i=a_j}^{b_j} f(m/z'_i)} \quad (2)$$

The concomitant C_j^{int} is an estimate of each of the disjoint profile areas, set as the average of the left and right Riemann sum of $f(m/z')$ within a_j and b_j. Akin to β_1, a second fractional threshold $\beta_2 < 1x10^{-5}$ relative to the most intense C_j^{int} was deployed to prune low-intense centroids which are unlikely measurable.

A third stage specifies and categorizes pairwise links between the centroids of each adduct class of a molecular formula. The set of such links joins all centroids of an adduct-classed formula into a tree-like structure, as exemplified by the red lines in the bottom panel of Figure 1 for the molecular formula of the pesticide Terbumeton and its first adduct class from Table S-1. Assignment of these pairwise links is run by selectively linking all centroids via their constituting isotopologues, using five steps:

(a) A listing of all single-isotopic transitions *T* between pairs of isotopologues is derived. Such a transition is defined as the replacement of one higher against a lower abundant isotope of the same element (gray lines in Figure 1, panel C). It must be noted that the higher abundant isotope may not strictly be the isotope of lowest mass per element, e.g., in a replacement of ^{33}S by ^{36}S instead of a replacement of ^{32}S.

(b) An association of each such isotopologue *i* with a centroid *j* in between envelope minima a_j and b_j is established, i.e., $m/z'_{a_j} \leq M_i \leq m/z'_{b_j}$. In the named panel, this association is valid for isotopologues (black points) and centroids (red points) jointly ranging in between the blue horizontal bars of envelope minima.

(c) Each centroid *y* is then uniquely linked through associated isotope transitions to one centroid *x* of lower *m/z*, as shown by red lines in the discussed panel C of Figure 1. Ideally, centroid *x* is the most intense among candidates. In the unlikely case of two or more centroids x of same intensity, *x* is otherwise chosen as the centroid for which the difference in *m/z* to *y* is closest to that of the underlying isotopologue transition *T*.

(d) When available, another centroid *w* with highest intensity among the remaining centroids and exceeding *x* and *y* in intensity is recorded. This so-called marker centroid indicates if the centroid pair (*x,y*) can occur without other centroids of higher intensity.

(e) The centroid link is categorized by the lower abundant isotope introduced in the associated isotope transition *T*, e.g., $T=^{18}O$ or $T=^{13}C$. If more than one transition exists to categorize the linkage *x* to *y*, again the one for which the mass difference is closest to that of the centroids is selected.

Overall, a centroid can be linked to more than one other centroid of higher *m/z*. A centroid can conversely only be linked to a single centroid of lower *m/z*, except for the monoisotopic centroid which logically lacks such a link.

The assigned linkages are each characterized in a fourth stage. Namely, for each pair *i* of centroids (*x, y*), a vector v_i with elements $v_{i,j}$ is defined as

$$v_i = (C_x^{m/z},\ \Delta m/z_{(x,y)},\ LIR_i,\ \Delta m/z_{(x,w)}) \quad (3)$$

The second vector element $v_{i,2}$ is the *m/z* difference between the two centroids, $C_y^{m/z} - C_x^{m/z}$. The third element $v_{i,3}$ is their logarithmic intensity ratio

$$LIR_i = log_{10}(C_x^{int} / C_y^{int}) \quad (4)$$

Finally, $v_{i,4}$ is the *m/z* difference between the lower mass centroid of the pair and the marker centroid *w*, i.e., $C_x^{m/z} - C_w^{m/z}$. If no centroid *w* has been recorded, this last entry in v_i is set to 0.

3.2.2 Data discretization. The four-dimensional v_i of simulated centroid pairs are converted into a discretized data model in a fifth stage. On the one hand, this model is used to approximate the feasible space of v_i values, at some level of distortion. On the other hand, and given the vast numbers of simulated centroid pairs, the model represents the data at much lower storage costs and can be efficiently used to compare measured data against. Therefore, after pooling centroid links over all molecular formulas and their adduct classes, the discretization model is built separately for disjoint subsets *v* of vectors v_i for every combination (*T,z*) of transition category *T* and charge *z*, so as to preserve information on the underlying isotope replacements and charge levels. Based on a top-down partitioning, the model represents points in *v* via their enclosing rectangles in $\mathbb{R}^4$, separating subspaces populated with simulation results from void space. The method is similar to the construction of bounding volume hierarchies, but with a different splitting heuristic to deal with large point sets.[36] Namely, the partitioning procedure searches for the largest gap in the first dimension $v_{i,1}$ and splits *v* accordingly, i.e., into two point sets of v_i with values of $C_x^{m/z}$ below and above that gap respectively. Each of the two sets are again screened for the largest gap in the next dimension and partitioned accordingly along $v_{i,2}$. This recursion cycles over the four dimensions, unless the values $v_{i,j}$ of all points in a partition are identical in a dimension *j* – the affected dimension is skipped in that case. The recursion terminates for an individual partition having reached a minimum size, i.e., the difference between the maximum and minimum values of all $v_{i,j}$ in *v* must not be larger than predefined thresholds $\tau_j \geq 0$ in each dimension

$1 \leq j \leq 4$. The point set in such a terminal partition l is then represented by its bounding rectangle $B_l^{T,z}$, with axis-parallel edges connecting the named minimum and maximum point values in each separate dimension; the original data points are discarded. Although the used split heuristic is not strictly density based, the largest gaps in a partition often fall into low-dense regions with sparse occurrence of simulated points. The method should also not be confused with binning, which uses regular intervals to discretize data, risking to miss broad gaps in the data.[37]

To account for the intermittent gaps imposed by the above partition scheme, and incomplete coverage from simulations, the size of rectangles $B_l^{T,z}$ formed in each subset v of (T,z) is increased by a distance estimate δ_j in each individual dimension j. δ_j is calculated as the 99 th percentile of the marginal distance distribution in each dimension j between a point n and its Euclidean nearest neighbor m, taken over all k points contained in all subsets v. More formally, the nearest neighbor m of a point n is defined as

$$m = argmin_i \left\{ \sqrt{\sum_{j=1}^{4} \left(\frac{\bar{v}_{i,j} - \bar{v}_{n,j}}{s_j} \right)^2} \mid i \neq n \wedge 1 \leq i \leq k \right\} \quad (5)$$

s_j is a scaling parameter, specified as the range of values for v_1 and v_3, but using $s_2 = s_4 = 1x10^{-3}\ u$ for the m/z differences v_2 and v_4. In summary, even if the centroid links of a compound molecular formula are not included in the simulation, they are assumed to rarely deviate further than δ_j from already simulated points and, as an upper bound, from their representing rectangles. Another trade-off inflicts the parameterization of τ_j to determine the maximum size of bounding rectangles. When set too small, a large fragmentation of the simulated space ensues, with aggravated query costs and increased overlap of bounding rectangles after extension by δ_j. When set too coarse, contrariwise, void space is less segregated from the underlying distribution, risking more false positive queries. Following an evaluation of this trade-off under varying parameterizations, minimum sizes of $\tau_1 = 100\ u$, $\tau_2 = \tau_4 = 2.5x10^{-3}\ u$ and $\tau_3 = 0.2$ were utilized to terminate further partitioning.

3.2.3 Data query for measured centroids. Linkages among centroid peaks of measured LC-HRMS data are compared to the above discretized data in a sixth stage. If a detected link intersects with the discretized ones within bounds of measurement and simulation uncertainties ε, the concerned centroids are annotated to stem from the same adduct of a measured but possibly unknown compound. To this end, a measured vector v'_i analogous to the simulated v_i in eq. (3) is calculated for all possible 2- and 3-combinations of measured centroids (x',y') and (x',y',w'), respectively, and under the following restrictions. First, the combined centroids deviate no more than ΔRT_{max} in

their retention times. Second, $C_{x'}^{m/z} < C_{y'}^{m/z}$. Third, and just as in the simulation routine, $v'_{i,4}=0$ in the 2-combinations lacking marker centroid w'. The points $v'_{i,j}$ of each combination are then expanded with their particular uncertainties ε to intervals, which in turn define a rectangle $M_i \in \mathbb{R}^4$

$$M_i = \left[v'_{i,1} - \varepsilon_1; v'_{i,1} + \varepsilon_2\right] \times \left[v'_{i,3} - 2\varepsilon_3; v'_{i,2} + 2\varepsilon_3\right] \times \left[v'_{i,3} - \varepsilon_4; v'_{i,3} + \varepsilon_4\right] \times \left[v'_{i,4} - 2\varepsilon_3; v'_{i,4} + 2\varepsilon_3\right] \quad (6)$$

The uncertainties ε_1 and ε_2 in the first interval account for maximum *m/z* shifts caused by the omission of single-isotopic elements during adduct class simplifications. Briefly, $\varepsilon_1=57$ *u* and $\varepsilon_2=3$ *u* for the omission of F_3 and H_3 which would at the most be added and subtracted during positively and negatively charged adduct formation, respectively. Covering the mass accuracy, ε_3 is the maximum expected deviation from the true *m/z* value of a centroid. Similarly, ε_4 quantifies the uncertainty in the LIR′ of a measured centroid pair x' and y':

$$\varepsilon_4 = log_{10}\left(\frac{1+\Delta Int}{1-\Delta Int}\right) \quad (7)$$

ΔInt specifies the fraction by which a measured centroid intensity can deviate from its true value. Notably, the uncertainty of LIR′ does not vary with absolute intensities. To swiftly test whether an observed rectangle M_i intersects with simulated subspaces, all of the *1,…,l,…n* arbitrarily overlapping rectangles $B_l^{T,z}$ were organized into multidimensional variants of ternary interval trees.[38] With a low space consumption at order *O(n)*, these trees require only *O(log n)* query time to return all $B_l^{T,z} \cap M_i \neq \emptyset$.

3.2.4 Componentization. Complementing the hitherto outlined linkage of isotopologues stemming from a specific yet unknown adduct of a compound, a search for different ESI adducts of the same analyte is conducted. For this purpose, all pairwise combinations of anticipated adducts are formed, using a predefined subset from a full list of 49 ESI adducts.[34] All possible pairs of measured centroids not deviating more than ΔRT_{max} are then checked whether their different *m/z* values can be formed from each other by any of the 2 permutations of any of these adduct pairs, within bounds of $2\varepsilon_3$.[39] If so, they are linked accordingly. Notably, more than one adduct pair can sometimes explain the same *m/z* differences, e.g., the difference between adducts $[M+H]^+$ and $[2M+NH_4]^+$ equals that of $[M+2H]^+$ and $[M+HN_4]^+$ unless information on charge state *z* is available. All measured centroids which were directly or indirectly connected via isotopologue or adduct linkages are ultimately joined into a chemical component. Herein, centroids exclusively

interrelated via isotopologue or adduct linkages are referred to as isotopologue or adduct groups, respectively. Finally, components can be tagged for their membership in a homologues series – the corresponding algorithm is described elsewhere.[32]

3.2.5 Validation. The performance of the centroid grouping approach was assessed with two complementary strategies for its recall (sensitivity) and precision.

First, an external test set of $2x10^4$ randomly selected PubChem molecular formulas was used to simulate and group their centroids for each of the three instrument resolution functions and thresholds β_1 and β_2. Formulas adhered to the same criteria as those used for above stage (1), except that they were randomly modified to represent one of the original 49 adduct species instead of simplified adduct classes and not utilized for building the discretized data model. By checking whether all centroids of each test formula could be successfully combined to a single isotopologue group, the true positive rate is defined as

$$recall = \frac{TP}{TP+FN} \tag{8}$$

where true positives (*TP*) and false negatives (*FN*) are the number of test formulas with fully and incompletely grouped centroids, respectively. Recall values must not be equated with the marginal percentiles to derive $\delta_{1\text{-}4}$. Furthermore, centroid patterns were perturbed to mimic conditions in which centroid intensities fall below the limit of detection (LOD). For this purpose, a random set of lowest-intense centroids was removed per modified formula and the recall reassessed for the $n \geq 2$ remaining centroids. Using theoretical and hence exact centroid characteristics, the extension of test points ν'_i by ε_3 and ε_4 was skipped for their query stage.

False positive rates encountered in practical applications are difficult to estimate with such test simulations. Therefore, the precision of the isotopologue grouping approach was validated with genuine LC-HRMS data, measured under positive ESI for ten STP effluent samples as detailed in the next section. Two sets of compounds were thereby investigated. On the one hand, the isotopologue grouping of a curated set of 43 ubiquitously observed target compounds was evaluated, including pharmaceuticals, biocides and transformation products. The occurrence of at least one ESI adduct of each such target in each STP sample was manually affirmed with unprocessed data via *RT* matching to reference standards, isotopologue pattern confirmation and – for a fraction 0.81 of the targets – MS/MS information, as reported elsewhere.[40] On the other hand, a spiked set of 108 isotope-labeled internal standards (IS) was investigated, some of which were labeled analogs of the

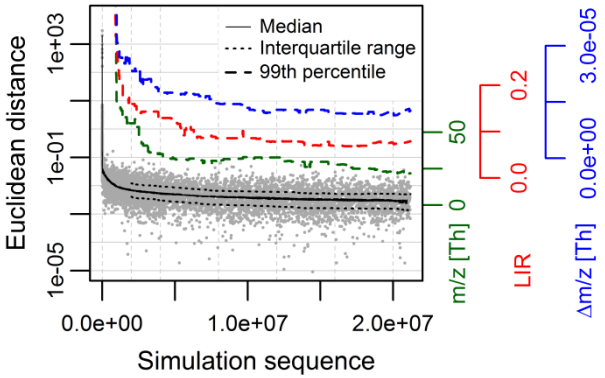


Figure 2. *Convergence of the centroid linkages distribution at R(200)=140K as a function of sequentially simulated linkage data points v_i (abscissa). Gray dots show the Euclidean nearest neighbor (NN) distance of a simulated point v_i to its previously simulated points from other molecular formulas for one particular permutation of these points (left ordinate; note the log-scale). Black dashed lines show the upper and lower quartiles thereof. Black solid lines overlay the distance medians of that and seven other random permutations of the sequence of simulated linkage points. For the NN distances, 1000 points were each sampled in the bins separated by gray dashed lines. In turn, the underlying unscaled distances δ_1 to δ_3 are drawn in color and were estimated from moving windows of ±500 binned simulation points, as were the quartiles (right ordinate).*

target compounds. In addition to a sample-wise recall, the precision for both sets was thus defined as

$$precision = \frac{TP}{TP+FP} \tag{9}$$

Akin to equation (8), *TP* refers to the number of positively screened compound adduct species per sample for which the measured centroids were correctly assorted into one isotopologue group per compound adduct species by the evaluated algorithm. For a positive screening outcome, all expected centroids of a compound adduct species above sample- and *RT*-dependent intensity thresholds (cp. Figure SI-1) had to be completely matched with measured centroids, whereas those below were treated as optional matches. The expected centroids of a compound adduct species were simulated in analogy to the above PubChem validation set, while the screening routine to

match them with measured centroids was implemented as part of the R package *enviMass* with parameters listed in Table SI-2.[41] In contrast, *FP* denotes the number of compound adduct species for which non-expected centroids were assorted into their respective isotopologue group. Again, this validation set of molecular formulas of targets and IS were a priori excluded from deriving the discretized data model. Measurement uncertainties were set to ε_3 = *2 ppm* and ε_4 alias $\Delta Int = 0.2$, as required to reproduce deviations between expected and measured centroids for the curated target set.

3.2.6 Experimental setup and data processing. Validation was conducted with the Q-Exactive measurements of flow-proportional effluent samples from ten Swiss STPs as derived and detailed by Schymanski et al.[40] In summary, sample volumes of 0.25 L were pH-adjusted, filtered and spiked with 120 ^{2}H-, ^{13}C- and/or ^{15}N-isotope labeled standards. Two enrichment steps to 1 ml ensued, one by a mixed-bed solid-phase extraction and another using a nitrogen gas stream, with intermediate steps of acidic/basic extraction and further filtering. Reconstituted aliquots of 20 μL were then analyzed with HPLC-ESI-HRMS, combining a Waters XBridge C18 column for mixed gradient reverse phase chromatography (Milford, U.S.), positive electrospray ionization (spray voltage +4; 300°C capillary temperature) and *m/z* detection with the medium 140K resolution function using a hybrid quadrupole-orbitrap Q-Exactive (Thermo Scientific, San Jose, USA) in full-scan mode. During data processing, the resultant measurement files were first converted to .mzXML open format files and centroided with ProteoWizard v3.0.7162.[42] Ion chromatogram extraction and peak picking was thereupon run with the R *enviPick* v1.2 package, with parameters specified in Tables SI-3.[43] Finally, summary function settings for the proposed isotopologue and adduct grouping incorporating some of the hitherto discussed parameters as well as the full componentization are given in Tables SI-4, as accepted by the presented algorithms.

3.3 Results & Discussion

3.3.1 Centroid linkage simulation and discretization. (Unless stated otherwise, the below outcomes are those for the medium resolution function of *R(200)*=140K, with outcomes for the lower and higher *R(200)* optionally stated in brackets). A vast total of 2.1×10^7 (1.8×10^7/3.2×10^7) simulation points were generated, with numbers increasing from *R(200)*=70K to 280K for the same adduct class formulas. The simulation progress in covering the feasible space of centroid linkages is tracked in Figure 2, expressed as the minimum Euclidean distance of each newly simulated point to its simulated predecessors. Notably, this distance decreases promptly within the initial ~ 5×10^5 simulations but on much smaller yet possibly defining scales thereafter, with new simulations located to existing ones within the percentile ranges $\delta_{1\text{-}3}$ depicted in Figure 2. At some stage, finer coverage will be paid by

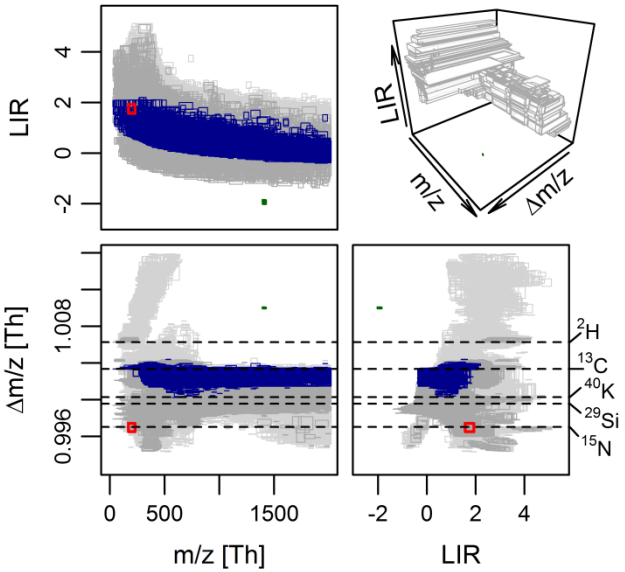


Figure 3. *Discretized distribution of centroid linkages at Δm/z ≈ 1 simulated at R(200)=140K for 83.000 compounds from the PubChem[33] database. While bounding rectangles in light gray contain all linkages, those in dark gray frame only those of centroids which differ by one isotope replacement from being monoisotopic. Blue rectangles further restrict the latter to linkages categorized by ^{13}C transitions. Dashed lines indicate the theoretical Δm/z values for relevant isotope transitions of replacing the lowest-mass isotope of an element at z=1. The small green rectangle shows the size of the nearest neighbor extension, the red the intersection query M_i for linkage 1 of Figure 1, panel C. For the 3D visualization, a much coarser discretization was utilized.*

soaring simulation costs and the latter therefore aborted. However, the until then acquired marginal percentiles give a useful estimate on how far the linkage characteristics of any uncovered formulas might range from simulated ones, assuming that simulated and uncovered formulas stem from the same population. The 99th percentiles were $\delta_1=22$ *Th*, $\delta_2=1.3x10^{-5}$ *Th*, $\delta_3=0.13$ and $\delta_4=0.33$ *Th*. Comparable distances arose for simulations with the lower and higher resolution functions. While the exact choice of this percentile for deriving δ might be regarded as arbitrary, usage of larger δ values was found to increase false positive matches, with a minor decrease in false negative cases akin to the below outlined validation.

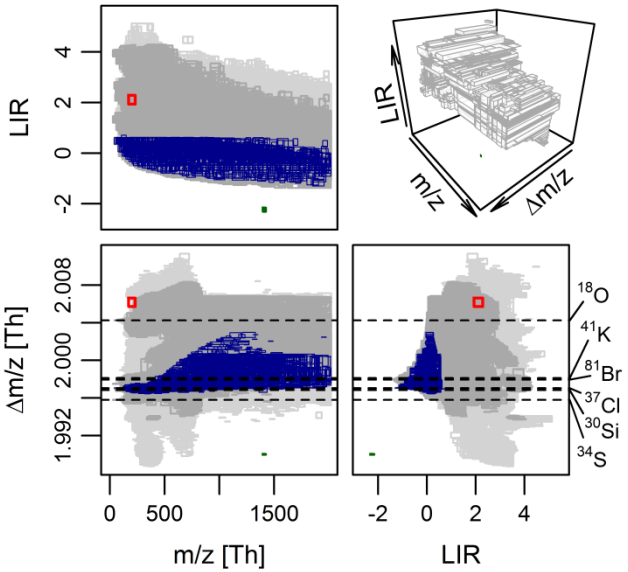


Figure 4. *Discretized distribution of simulated centroid linkages at Δm/z ≈ 2 and R(200)=140K for 83.000 compounds from the PubChem[33] database. Rectangles in light gray contain all linkages, whereas those in dark gray bound those of centroids which differ by only one isotope from being monoisotopic. Blue rectangles further reduce the latter to linkages categorized as ^{37}Cl transitions. These are not shifted towards lower Δm/z because this would premise the existence of ^{34}S transitions, in which case they are categorized differently. The red rectangle is an intersection query M_i for linkage 2 of panel C in Figure 1. All other is analogous to Figure 3.*

Having sampled the space of centroid linkages, its discretization can be approached next. As depicted in Figures 3, 4 and S-7 to S-17, a total of 4.4×10^5 (3.9×10^5/5.6×10^5) rectangles were assembled to represent all simulated centroid linkages separately for charges z and transition category T, equal to a compression ratio of 48 (46/57). Herein, adduct class simulations of positive and negative ionization were discretized jointly, so as to not further inflate the number of rectangles. In relation to the covered space, extensions by δ were small (green rectangles in the named Figures).

Based on these bounding rectangles, several findings must be highlighted. First, the overall space of centroid linkages is surprisingly complex, has dependencies between its dimensions and is in places intrinsic to individual resolution functions and the charge z (light gray rectangles in the named

Figures). For example, LIR ranges and bounds generally decrease with mass, which seems in line with (a) increasing numbers of available atoms to constitute underlying isotope transitions of rising intensity for the lower bound and (b) decreasing resolutions with mass to not resolve high LIR linkages for the higher bound. Smallest LIR ranges were in turn observed at $\Delta m/z$ locations of $z \neq 1$ in Figures S-11 and S-12. In other locations, the LIR was strictly above 0, implying that centroid x linked to y was always equal or larger in intensity (cp. Figures S-14 to S-17). On the other hand, lowest LIR values were confined to particular $\Delta m/z$ positions, e.g., around the transitions of high-abundant isotopes such as ^{37}Cl or ^{81}Br at increasing mass. These ratios firmly deviate from the simple linear or polynomial intensity relationships reported for, e.g., polypeptides.[22,44] Second, the mass difference between linked centroids frequently deviates from that of the associated isotope transitions, as caused by the superposition to envelopes from which the centroids are calculated at the mass-dependent resolution (cp. green dashed line of T=^{18}O in panel C, Figure 1). Similarly, the probabilities of the two isotopologues categorizing a linkage may not be proportional to the intensities of the paired centroids. For the example of Terbumeton in Figure 1, the transitions and centroid linkages are still of low complexity. However, this changes rapidly with slightly more complicated molecular formulas and increasing resolution, exemplified in Figures S-2 to S-4. Some strong $\Delta m/z$ shifts from the underlying isotope transitions are indeed striking. Most markedly is the one at $\Delta m/z > 1.008$ *Th* in Figure 3, which can be caused by common transitions such as T=^{13}C evidenced in Figure S-5 and which are more pronounced for higher resolution functions than low ones (cp. Figure S-7 vs. S-9; Figure S-6 gives showcases a shift for T=^{37}Cl). Thus, shifts of more than 30 ppm from the m/z of the closest theoretical isotope transition were encountered, far beyond the relative isotopic mass defects described for accurate masses at lower resolutions.[45] Limiting linkages to the $\Delta m/z$ of ^{13}C transitions[9,13,14,17] and to fixed[15,19,21,46–48] or probabilistic[20,21] error windows set around simplified m/z shifts will therefore be insufficient for a comprehensive yet discriminatory isotopologue grouping at high resolutions.

In general, the diversity of linkage characteristics reduced with decreasing resolution and was somewhat diminished for transitions from monoisotopic isotopologue compositions (dark gray rectangles in all Figures), except for higher values of z as in Figures S-10 or S-12. Finally, transition categories as those signified by blue rectangles for T=^{13}C or ^{37}Cl must not be misinterpreted. Despite usage for assigning and restricting linkages, they neither rule out the presence of other transitions involved in forming a centroid nor does their absence as category preclude their presence.

3.3.2 Validation with simulated data. To elucidate in how far these discretized models generalize to external data, their performance in correctly grouping all isotopologue centroids of an adduct was accessed with external test simulations. A recall of 0.95 (0.96/0.90) was thereby obtained, with

higher resolution functions more prone to raise *FN* cases. However, a throughout multiply charged *FN* fraction of 0.84 (0.75/0.86) was caused by only 5 of the 49 adducts (foremost $[M+H+NH_4]^+$, $[M\pm3H]^+$, $[M\pm3H]^-$, $[M+2H+Na]^+$ and $[M+H+2Na]^+$), possibly caused by omission of +H in the adduct classes. Since transitions of $T=^2H$ are of low intensity and the remaining centroids might still be correctly grouped, the practical relevance of these *FN* cases is anticipated to be small.

This first validation with simulated and hence exact data obviates any measurement uncertainties. Albeit hard to reproduce, extension of query rectangles M_i by these uncertainties can be expected though objectionable to further reduce *FN* cases. Interestingly, sufficient recall can only be achieved when a discretized model specific to the investigated resolution function is built. For instance, when validating test centroids simulated at a resolution of *R(200)*=280K with the model build at *R(200)*=70K, and vice versa, the recall deteriorated to 0.26 and 0.54, respectively. Moreover, and with regard to real-world applications, centroid sets will rarely be complete at an intensity threshold β_2, because a variable subset of them will fall below the LOD. Still, a recall of 0.95 (0.94/0.97) was achieved when removing the monoisotopic centroid and all other centroids of equal or lower intensity, i.e., for formulas for which the monoisotopic centroid was not the most intense. Furthermore, even when removing a random subset of lowest-intense peaks, the grouping recall did not decrease below 0.91 (0.91/0.88) for the at least two remaining centroids, again with high *FN* frequencies attributable to a few affected adducts.

3.3.3 Validation with measured data. Albeit extensive in size, the above recall assessment with simulated test data can neither quantify false positive findings nor mirror the nature of measured data. The latter complicates matters by inflicting measurement uncertainties, interferences between co-eluting compounds and by noise peaks, shifting detection limits, varying complexity of different matrices and automated decisions made during data processing, amongst others. Therefore, recall and precision were also evaluated with picked LC-HRMS centroid peaks of effluent samples from different STPs in Switzerland; the results are listed in the right half of Table S-5. First addressing the set of target compounds, a maximum recall was achieved for all ten samples. Put differently, the centroids of each screening match were always correctly joined, provided that at least two centroids were measured. In comparison to the recall, the average precision of 0.90±0.03 taken over all STP samples was smaller, corresponding to a mean number of 1.6±0.3 wrongly assigned peaks per *FP* case. Comparable outcomes were found for the screened set of IS compound adducts. Again, full absence of *FN* cases enabled a maximum recall, at a similar precision of 0.88±0.03 with a mean of 1.5±0.4 erroneously included peaks per *FP* case. The usage of marker peaks somewhat reduced *FP* interferences: without such a check for another centroid of higher intensity, precisions would have slightly dropped

by a coincident value of 0.02, caused by a more frequent acceptance of false centroid linkages.

Although different samples with largely varying peak numbers and matched compound intensities were used, such excellent validation outcomes might still be specific to the given instrument resolution, the validated set of compounds, the screening strategy or the quality of other data processing steps. For example, larger measurement uncertainties ε necessitate larger query rectangles M_i and will thus lead to higher *FP* rates (cp. red query rectangles in Figures 3 and 4). Similarly, the available resolution will also be decisive: the overall bounding volume for representing simulated linkages with the high resolution function *R(200)*=280K is approximately 40% larger than at the lower *R(200)*=70K, based on the scaling of equation (5). The larger this volume, the lower will be its specificity to discriminate against false linkages with, for instance, random noise peaks. Concerning their ion species, the majority of matches was attained for singly-charged adducts, for which the validation simulation from the last section proved least problematic. Finally, the quality of the peak picking will in turn not only influence the frequency of noise peaks but also the positively screened set of compounds to validate with. Yet, certain aspects of the validation are rather set to disadvantage the approach. Foremost, complex STP matrices were investigated, with an increased risk of co-eluting centroids to be falsely linked. Second, the molecular formulas of all validation compounds were a priori excluded from simulating the centroid linkage. Third, lower uncertainties ε can be attained in other experimental settings in which, for example, replicate measurements enable better mass and intensity estimates. Likewise, the *m/z* uncertainty ε_3 was estimated from matching expected and measured centroids – the uncertainties for $\Delta m/z$ differences between adjacent centroids in a mass spectrum might in fact be lower.

3.3.4 Nontargeted grouping and negative findings. On average, $(5.3\pm0.6)\text{x}10^5$ interval trees of individual rectangle sets (*T,z*) had to be queried per STP sample to decide whether each of a mean of $(2.0\pm0.5)\text{x}10^4$ candidate linkages had to be rejected or not. However, rejection of a single candidate does not imply that its two centroids are unrelated. As conditioned on the training restrictions for the discretized linkage model, the latter can nonetheless be indirectly joined via other accepted linkages into an isotopologue group. The number of true negative (*TN*) linkages must hence be estimated differently. Though crucial for the performance of a classification, *TN* rates are often ignored when evaluating isotopologue grouping tools.[49] In the first place, *TN* linkages depend on the overall candidate linkages considered. Here, only such candidate centroids pairs are queried which have a similar retention time and which fall into the separate $\Delta m/z$ ranges spanned by the discretized model - all others are outright excluded. Hence, a last analysis randomly sampled pairs of centroid peaks from any of the STP samples to record their values of *m/z*, $\Delta m/z$ and *LIR* and to attest their (in)direct linkage,

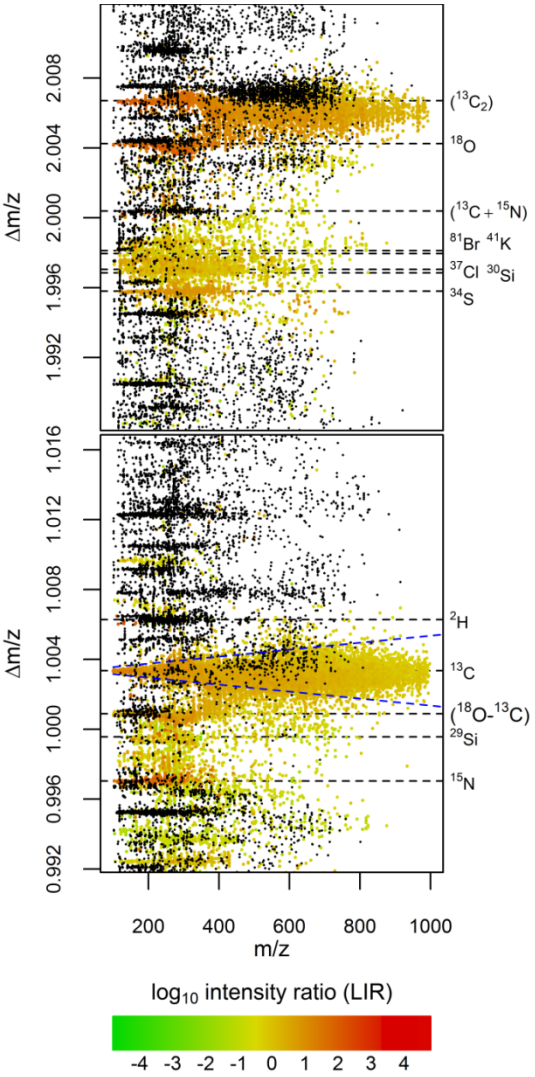


Figure 5. *Characteristics of $7x10^4$ centroids randomly paired from same STP samples and coeluting at $\Delta RT_{max} \leq 4$ s. Characteristics of pairs with a valid direct or indirect linkage are shown as points with a color coding corresponding to their LIR or, when invalid, as black dots. Dashed black lines exemplify theoretical mass distances of selected isotope transitions; blue lines signify a Δm/z window of 2 ppm located around the ^{13}C isotope transition.*

within the so far used tolerances ΔRT_{max}. With the above established recall and precision in mind, the resulting systematic distributions in Figure 5 for both the accepted (colored dots) and rejected linkages (black dots) around $\Delta m/z \approx 1$ *Th* and $\Delta m/z \approx 2$ *Th* are noteworthy.

On the one hand, while spread throughout the full feasible space, the accepted fraction of 0.77 of the linkages frequently clusters around *m/z* shifts of isotopic transitions (colored dots and dashed black lines, respectively). The most frequented is not surprisingly located around the ^{13}C isotope transition and increases in deviation with *m/z* as resolution diminishes and interferences with neighboring isotopologues could increase. On the same line of evidence, the cluster around the low-LIR ^{15}N transition fades with increasing *m/z*, most likely by getting subsumed into the envelope of the more dominant ^{13}C transition. Moreover, isotopologues formed by two ^{13}C transitions might be of comparably low intensity as those from neighboring ^{18}O transitions. Consequently, their peak shapes often integrate into envelopes with intermediate $\Delta m/z$ centroids. Yet other cluster might be explained by additional combinatorial $\Delta m/z$ distances, e.g., as a difference between ^{18}O and ^{13}C or the sum of ^{15}N and ^{13}C transitions.

On the other hand, the remaining fraction of 0.23 rejected linkages form clusters as well (black dots). Some rejections coincide in *m/z* and $\Delta m/z$ with accepted linkages and can only be rejected by their LIR and the possible occurrence of marker centroids. Situated at ^{15}N, ^{13}C and particularly ^{2}H transitions, they are likely a results of incomplete labeling of IS compounds, if not leading to false positive linkages. In general, cluster of rejected linkages were not overly sample-specific (data not shown) and can theoretically be caused by a combinatorial explosion of *m/z* differences among isotopologues of, e.g., different adducts or systematic changes in the elemental composition of co-eluting analytes. Future research should clarify their origin, so as to improve isotopologue grouping capabilities from disclosing common cluster of acceptable and partly overlapping unacceptable linkages.

3.3.5 Componentization. Having so far traversed the subtleties in isotopologue grouping, the full nontargeted combination of isotopologue and adduct linkages into their chemical components is subject of this concluding section, with outcomes for all peaks in a STP sample detailed in the left part of Table S-5.

First, although the large numbers of picked peaks somewhat varied among samples ($2.1\pm0.3 \times 10^4$), their fractions included into isotopologue groups remained markedly constant at 0.51±0.03. Put differently, 49% of the centroided peaks remained singletons. Comparably constant fractions of 0.29±0.02 peaks could also be assigned into adduct groups, using solely the most commonly reported singly charged ESI adducts $[M+H]^+$, $[M+NH_4]^+$, $[M+Na]^+$ and $[M+K]^+$.[50] Confirmingly, a fraction of 0.95±0.05 positively

screened target matches could be attributed to these adducts (column seven of the named table). For these adduct links only a minor fraction of 0.03 peaks had to be assigned with ambiguities, i.e., a centroid was only rarely related to a second centroid in one adduct relation yet with a different adduct relation to at least a third centroid peak. In contrast, ambiguities rose drastically when employing the unrestricted set of 49 adduct species. That is, in spite of an enlarge fraction of 0.79±0.02 peaks linked into adduct groups, ambiguities now inflicted fractions 0.80±0.02 of the linked peaks. These artefacts in turn complicated the full componentization. Using only the most common adducts, a fraction of 0.61±0.02 peaks could be assorted via isotopologue and adduct linkages into a mean of $1.2\pm0.1\text{x}10^4$ components – as opposed to 0.86±0.01 peak fractions assorted into only $0.40\pm0.0\text{x}10^4$ highly ambiguous components when all adducts were used. Usage of extended adduct sets may be less problematic for less complex matrices or when complemented by orthogonal information on, e.g., chromatographic peak shape similarities. Otherwise, the quantity of wrongly assigned linkages is easily misleading.

3.4 Implementation

The presented algorithms including query and discretization functionalities are made publicly available as part of the R *nontarget* package version 2.0,[1] with discretized linkage models for a particular resolution function either provided in the associated R *nontargetData* package[51] or compiled upon request. The modularized package accepts user-define parameters for all discussed variables, import of customized adducts or expected neutral losses and offers visualization of the generated groups and components, tagged by their possible membership in homologue series (cp. Figure S-18).

Avoiding the lengthy simulations to define linkage spaces and enabling the inclusion of more elements, the package also contains a previous, rule-based isotopologue grouping algorithm which has found applications elsewhere.[40,52–54] Herein, linkages between centroids are primarily defined by their underlying transitions, with rules for (a) LIR bounds based on the maximum atom counts of the underlying elements and their natural isotope abundances, (b) atom count corrections by commonly observed ratios to carbon atoms in small molecules,[55] (c) checks for predictable ^{13}C isotope peaks, (d) limits for the *m/z* range and (e) nesting of isotope patterns at different charge levels. Since fully ignoring the above outlined $\Delta m/z$ and LIR shifts induced by the superposition of isotopologues, the recall and precision of this approach is anticipated to be reduced.

3.5 Conclusion

A machine learning paradigm for delineating the joint distribution of four defining and directly interpretable LC-HRMS characteristics between pairs of isotopologue centroids is presented, in firm context of their mass spectrometric resolution. When used in the nontargeted analysis of small molecules, these distributions are highly effective in grouping the isotopologue centroids of individual analytes and isotopically labeled standards, aimed at minimizing false negative rates and adaptable to varying measurement uncertainties. Moreover, grouping via such distributions can successfully discriminate against false positive cases which in turn also appear in systematic yet largely unreported cluster. The origin of the latter may be further elucidated to improve the precision of the method. In contrast, the sole usage of mass differences frequently fails when assorting the full set of adducts potentially formed by analytes, except for the restricted subset of commonly observed single-charged adducts. In the former case, an orthogonal verification with, e.g., correlation of elution profiles seems advisable.

Future research may likely accelerate the computational performance of simulations and queries from several days and minutes to hours and seconds, respectively. The first might be achieved through a non-random sampling of molecular formulas to be simulated. The second via data-driven and hence larger bounding volumes for a comparable representation of the simulated distributions. However, given the so far achieved overall high recall and precision of the presented approach, simulation isotopologue characteristics at other instrument resolutions shall be derived, tested and discretized next for their availability in LC-HRMS investigations of small molecules.

REFERENCES

(1) Loos, M. *nontarget: Detecting Isotope, Adduct and Homologue Relations in LC-MS Data*; 2015. https://cran.r-project.org/web/packages/nontarget/index.html

(2) Dunn, W. B. *Physical biology* **2008**, *5* (1), 011001.

(3) Petrovic, M.; Farré, M.; De Alda, M. L.; Perez, S.; Postigo, C.; Köck, M.; Radjenovic, J.; Gros, M.; Barcelo, D. *Journal of Chromatography A* **2010**, *1217* (25), 4004–4017.

(4) Theodoridis, G. A.; Gika, H. G.; Want, E. J.; Wilson, I. D. *Analytica chimica acta* **2012**, *711*, 7–16.

(5) Fernandez-Albert, F.; Llorach, R.; Andres-Lacueva, C.; Perera-Lluna, A. *Analytical chemistry* **2014**, *86* (5), 2320–2325.

(6) Varghese, R. S.; Zhou, B.; Ranjbar, M. R. N.; Zhao, Y.; Ressom, H. W. *Proteome Sci* **2012**, *10* (Suppl 1), S8.

(7) Böcker, S.; Letzel, M. C.; Lipták, Z.; Pervukhin, A. *Bioinformatics* **2009**, *25* (2), 218–224.

(8) Brown, M.; Wedge, D. C.; Goodacre, R.; Kell, D. B.; Baker, P. N.; Kenny, L. C.; Mamas, M. A.; Neyses, L.; Dunn, W. B. *Bioinformatics* **2011**, *27* (8), 1108–1112.
(9) Kaever, A.; Landesfeind, M.; Possienke, M.; Feussner, K.; Feussner, I.; Meinicke, P. *BioMed Research International* **2012**, *2012*.
(10) Nakamura, Y.; Kanaya, S.; Sakurai, N.; Iijima, Y.; Aoki, K.; Okazaki, K.; Suzuki, H.; Kitayama, M.; Shibata, D. *Plant biotechnology* **2008**, *25* (4), 377–380.
(11) Pluskal, T. s; Uehara, T.; Yanagida, M. *Analytical chemistry* **2012**, *84* (10), 4396–4403.
(12) Rockwood, A. L.; Kushnir, M. M.; Nelson, G. J. *Journal of the American Society for Mass Spectrometry* **2003**, *14* (4), 311–322.
(13) Kloet, F. M. van der; Hendriks, M.; Hankemeier, T.; Reijmers, T. *Analytica chimica acta* **2013**, *801*, 34–42.
(14) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. *Analytical chemistry* **2011**, *84* (1), 283–289.
(15) Melamud, E.; Vastag, L.; Rabinowitz, J. D. *Analytical chemistry* **2010**, *82* (23), 9818–9826.
(16) Scheltema, R. A.; Decuypere, S.; Dujardin, J.-C.; Watson, D. G.; Jansen, R. C.; Breitling, R. *Bioanalysis* **2009**, *1* (9), 1551–1557.
(17) Alonso, A.; Julià, A.; Beltran, A.; Vinaixa, M.; Dıaz, M.; Ibañez, L.; Correig, X.; Marsal, S. *Bioinformatics* **2011**, *27* (9), 1339–1340.
(18) Broeckling, C. D.; Afsar, F.; Neumann, S.; Ben-Hur, A.; Prenni, J. *Analytical chemistry* **2014**, *86* (14), 6812–6817.
(19) Edmands, W. M.; Barupal, D. K.; Scalbert, A. *Bioinformatics* **2014**, btu705.
(20) Kenar, E.; Franken, H.; Forcisi, S.; Wörmann, K.; Häring, H.-U.; Lehmann, R.; Schmitt-Kopplin, P.; Zell, A.; Kohlbacher, O. *Molecular & Cellular Proteomics* **2014**, *13* (1), 348–359.
(21) McIlwain, S.; Page, D.; Huttlin, E. L.; Sussman, M. R. *Bioinformatics* **2007**, *23* (13), i328–i336.
(22) Park, K.; Yoon, J. Y.; Lee, S.; Paek, E.; Park, H.; Jung, H.-J.; Lee, S.-W. *Analytical chemistry* **2008**, *80* (19), 7294–7303.
(23) Leptos, K. C.; Sarracino, D. A.; Jaffe, J. D.; Krastins, B.; Church, G. M. *Proteomics* **2006**, *6* (6), 1770–1782.
(24) Senko, M. W.; Beu, S. C.; McLaffertycor, F. W. *Journal of the American Society for Mass Spectrometry* **1995**, *6* (4), 229–233.
(25) Sun, Y.; Zhang, J.; Braga-Neto, U.; Dougherty, E. R. *BMC bioinformatics* **2010**, *11* (1), 490.
(26) Vakhrushev, S.; Dadimov, D.; Peter-Katalinic, J. *Analytical chemistry* **2009**, *81* (9), 3252–3260.
(27) Hermansson, M.; Uphoff, A.; Käkelä, R.; Somerharju, P. *Analytical chemistry* **2005**, *77* (7), 2166–2175.
(28) Guan, S.; Marshall, A. G.; Scheppele, S. E. *Analytical chemistry* **1996**, *68* (1), 46–71.
(29) Loos, M.; Gerber, C.; Corona, F.; Hollender, J.; Singer, H. *Analytical chemistry* **2015**.
(30) Yergey, J. A. *International Journal of Mass Spectrometry and Ion Physics* **1983**, *52* (2), 337–349.
(31) Urban, J.; Afseth, N. K.; Štys, D. *TrAC Trends in Analytical Chemistry* **2014**, *53*, 126–136.
(32) Loos, M.; Singer, H. *Journal of Chemoinformatics (submitted)* **2016**.

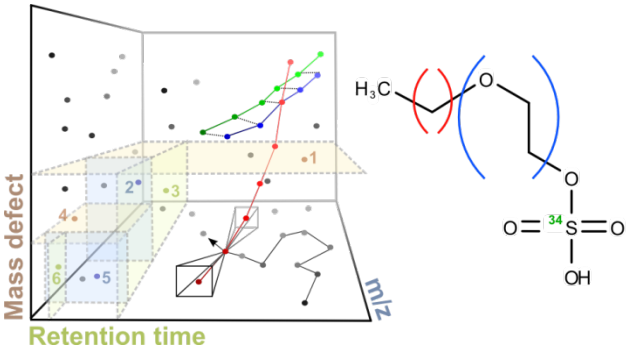
(33) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. *Annual reports in computational chemistry* **2008**, *4*, 217–241.
(34) Kind, T. Mass spectrometry adduct calculator, 2010.
(35) Li, L.; Kresh, J. A.; Karabacak, N. M.; Cobb, J. S.; Agar, J. N.; Hong, P. *Journal of the American Society for Mass Spectrometry* **2008**, *19* (12), 1867–1874.
(36) Nam, B.; Sussman, A. In *Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on*; IEEE, 2004; pp 171–180.
(37) Agrawal, R.; Gehrke, J.; Gunopulos, D.; Raghavan, P. *Data Mining and Knowledge Discovery* **2005**, *11* (1), 5–33.
(38) Samet, H. *Foundations of multidimensional and metric data structures*; Morgan Kaufmann, 2006.
(39) Bentley, J. L. *Communications of the ACM* **1975**, *18* (9), 509–517.
(40) Schymanski, E. L.; Singer, H. P.; Longree, P.; Loos, M.; Ruff, M.; Stravs, M. A.; Vidal, C. Ripolle s; Hollender, J. *Environmental science & technology* **2014**, *48* (3), 1811–1818.
(41) Loos, M. *enviMass: Utilities to Process Mass Spectrometry (LC-HRMS) Data for Environmental Trend Analysis*; 2014. https://github.com/blosloos/enviMass
(42) Holman, J. D.; Tabb, D. L.; Mallick, P. *Current Protocols in Bioinformatics* **2014**, 13–24.
(43) Loos, M. *enviPick: Peak picking for high resolution mass spectrometry data*; 2014. https://cran.r-project.org/web/packages/enviPick/index.html
(44) Valkenborg, D.; Jansen, I.; Burzykowski, T. *Journal of the American Society for Mass Spectrometry* **2008**, *19* (5), 703–712.
(45) Thurman, E. M.; Ferrer, I. *Analytical and bioanalytical chemistry* **2010**, *397* (7), 2807–2816.
(46) Aoshima, K.; Takahashi, K.; Ikawa, M.; Kimura, T.; Fukuda, M.; Tanaka, S.; Parry, H. E.; Fujita, Y.; Yoshizawa, A. C.; Utsunomiya, S.; others. *BMC bioinformatics* **2014**, *15* (1), 376.
(47) Grigsby, C. C.; Rizki, M. M.; Tamburino, L. A.; Pitsch, R. L.; Shiyanov, P. A.; Cool, D. R. *Analytical chemistry* **2010**, *82* (11), 4386–4395.
(48) Zhang, Z.; Marshall, A. G. *Journal of the American Society for Mass Spectrometry* **1998**, *9* (3), 225–233.
(49) Powers, D. M. *Journal of Machine Learning Technologies* **2011**, *2* (1), 37–63.
(50) Keller, B. O.; Sui, J.; Young, A. B.; Whittal, R. M. *Analytica chimica acta* **2008**, *627* (1), 71–81.
(51) Loos, M.; Corona, F. *nontargetData: Quantized simulation data of isotope pattern centroids*; 2014. https://cran.r-project.org/web/packages/nontargetData/index.html
(52) Hug, C.; Ulrich, N.; Schulze, T.; Brack, W.; Krauss, M. *Environmental Pollution* **2014**, *184*, 25–32.
(53) Ruff, M.; Mueller, M. S.; Loos, M.; Singer, H. P. *Water Research* **2015**, *87*, 145–154.
(54) Schymanski, E. L.; Singer, H. P.; Slobodnik, J.; Ipolyi, I. M.; Oswald, P.; Krauss, M.; Schulze, T.; Haglund, P.; Letzel, T.; Grosse, S.; others. *Analytical and bioanalytical chemistry* **2015**, 1–19.
(55) Kind, T.; Fiehn, O. *BMC bioinformatics* **2007**, *8* (1), 105.

Chapter 4

Nontargeted Homologue Series Extraction from Hyphenated High Resolution Mass Spectrometry Data

Martin Loos, Heinz Singer

Submitted to Journal of Chemoinformatics



ABSTRACT: One vast array of polar anthropogenic compounds routinely released into the environment comprises homologue series, i.e., sets of chemicals differing in a repeating chemical unit. Using analytical techniques such as liquid chromatography coupled to high-resolution mass spectrometry (LC-HRMS), these compounds are readily measurable as signal sets with characteristic shifts in mass and typically retention time. Despite such distinct characteristics, however, no computational approach for the direct untargeted detection of individual series in LC-HRMS data has to date been presented. To this end, a fast two-staged dynamic programming algorithm is introduced. In a first stage, a nearest neighbor-walk through a k-d tree representation of picked LC-HRMS peaks is used to extract feasible sub-series 3-tuples of peaks, imposing restrictions on, e.g., concomitant mass defect shifts. A second stage then gradually combines these peak tuples to larger series while ensuring smooth shifts in their retention time. This unsupervised approach was evaluated for ten effluent samples from Swiss sewage treatment plants (STPs), with both positive and negative electrospray-ionization. Substantial fractions of LC-HRMS signal peaks could subsequently be assigned to blank-subtracted series, although assignments were often not unique. The latter ambiguities were resolved using a self-organizing map technique and revealed both distinctive series meshing and rivaling combinatorial solutions in the presence of isobaric or gapped series peaks. When comparing STPs, several ubiquitous and partially low-frequent series mass shifts emerged and may prioritize future identification efforts. The presented algorithm is freely available as part of the R package *nontarget*.[1]

4.1 Introduction

Homologue compounds differing in a common chemical subunit are a widespread phenomenon. They have been focused on in fields as diverse as toxicology,[2–4] biopolymers,[5–8] food control[9,10] and oil processing[11,12]. In environmental research, natural and anthropogenic homologue sources have been addressed in various media, with Surface Active Agents (Surfactants) even classified as High Production Volume chemicals (HPVC).[13–20] Not surprisingly, the analytical detection of homologue series (HS) has therefore been of great interest. Among the methods used, liquid chromatography (LC) and high-resolution mass spectrometry (HRMS) have found abundant application to detect polar and semi-polar HS with both high sensitivity and specificity.[21] However, while most applications have targeted a priori known or suspected HS via their physicochemical characteristics, rather few nontargeted approaches have been established to point at unknown ones.[22] Concerning the latter, and as compared to non-homologous compounds, the regular patterns in LC/HRMS signals caused by the repetitive HS chemical units enable a specific fingerprinting. LC/HRMS has therefore potential to routinely single out unknown HS of, e.g., emerging contaminants, yet unidentified transformation products or differently ionized species of the same HS which otherwise evade targeted approaches. Once listed, the repetitive signals of individual or grouped HS would allow for averaged masses and additional peak relations to improve deisotoping, blank removal and finally their identification via complementary analytical methods (reference standards, MS^n).[23,24]

Using mass spectrometric information, Kendrick mass defect plots and its extensions to more than one type of chemical HS unit have been one popular method to determine the presence of unknown HS.[12,23,25] Another methodological branch has relied on extensive molecular formula fitting to detect regular patterns among measured classes of compounds, visualized by, e.g., van Krevelen diagrams or carbon versus mass plots.[26–28] Yet others have proposed a projection on regularly spaced vectors for HS pattern recognition.[29] Main drawbacks with this first mass spectrometric group arise, inter alia, from either the restriction to a fixed set of basic HS units or the requirement to derive unique molecular formulas for demanding numbers of measured masses. Any available information from the orthogonal chromatographic dimension is therein omitted – in spite of the systematic shifts in retention time (RT) among the homologues of a series.[30,31] Methods to embrace chromatographic information and to combine it with HRMS data are generally

rare in the field of nontargeted HS detection. For instance, Pietrogrande and coworkers have extensively published on autocovariance functions (ACFs) to reveal joint regularities in mass and RT differences.[32–35] Here, one major LC-related drawback is that RT shifts cannot be easily linearized to align with autocorrelated shifts in homologue masses because RT shifts are inherently nonlinear and vary significantly between different HS found in the same sample. Second, retracting and localizing single HS from an ACF is not straightforward; even the estimation of random and ordered ACF components requires assumptions on chromatographic peak properties or the maximum HS length. Third, infrequent HS may simply be masked by noise or more dominant HS. In contrast, other methods embracing both LC and HRMS information have aimed to aggregate data for comparison of samples, but will rarely elucidate individual HS.[5]

From a data mining perspective, the unsupervised extraction of regular HS patterns is indeed intricate, even from a list of picked signal peaks. As noted elsewhere,[36] an exhaustive pairwise peak comparison to find regular mass differences is a time-consuming task, not to speak of computing all possible series of such mass differences. Fortunately, changes in HS mass and RT can be restricted and their search optimized through appropriate metric data structures to obviate naïve approaches. Therefore, a fast two-staged computational strategy to extract systematically spaced peak series from electrospray-ionization (ESI) LC-HRMS measurements is presented and evaluated for ten effluent samples from sewage treatment plants (STPs). The novel approach lists all series even when (a) HS are not dominating a complex sample matrix, (b) no deisotoping or blank-subtraction was run beforehand, (c) heterogeneous measurement uncertainties need to be incorporated, (d) only limited prior HS information is available, (e) a number of different HS might occur and (f) combinatorial ambiguities can arise. The approach was successfully evaluated for several sewage treatment plant samples, revealing common patterns which have so far remained undetected.

4.2 Methods

4.2.1 Series definition. A series k of length $n \geq n_{min}$ is defined as the tuple $S_{n,k}=(p_{1,k}, \ldots, p_{n,k})$ of picked LC-HRMS peaks $p=\{m/z, RT, intensity\}$, ordered by increasing m/z values of its elements. S_n denotes the set of all series tuples with length n. Peaks being adjacent in a tuple are assumed to only differ in a repetitive but possibly unknown chemical unit or functional group, e.g., CH_2 or OH. As a result, deviations in the mass differences $\Delta m/z$ between any two adjacent series peaks $p_{j,k}$ and $p_{j+1,k}$ must remain within $[-4\varepsilon;4\varepsilon]$. For simplicity, ε here denotes the maximum ± m/z measurement error but may in principle be defined as a function of m/z or peak intensity.[37] The $\Delta m/z$ of all series in a LC-HRMS data set range within lower and upper bounds $\Delta m/z_{min}$ and $\Delta m/z_{max}$, a priori set as the considered mass range of chemical units at given charges z. Furthermore, $\Delta m/z$ restrains feasible shifts in the mass defect of adjacent series peaks. Denoted Δm, the mass defect here refers to the difference between an ion`s exact m/z value and its nearest integer.[38] For any unknown chemical unit that could constitute $\Delta m/z$, the minimum and maximum changes γ_{min} and γ_{max} in Δm from $p_{j,k}$ to $p_{j+1,k}$ in any series can be determined by the mass defects of the monoisotopic isotopes of elements introduced via this chemical unit. More precisely, let

$$\gamma_{min} = \min_{i\in\{1,\ldots,n\}} (\Delta m_i^*/m_i^*) \quad (1)$$

and

$$\gamma_{max} = \max_{i\in\{1,\ldots,n\}} (\Delta m_i^*/m_i^*) \quad (2)$$

with Δm_i^* and m_i^* denoting the mass defect and atomic mass of the monoisotopic isotopes of n elements, respectively. Then the range of permissible change of Δm with $\Delta m/z$ is set by

$$\gamma_{min} \leq \frac{\mathrm{d}\Delta m}{\mathrm{d}\Delta m/z} \leq \gamma_{max} \quad (3)$$

For example, albeit lacking knowledge of the exact composition of a chemical unit but assuming only C, H, N, O, S, Cl and Br to be present, we can expect the change in Δm from any series peak to the next to lie within [-0.0010$\Delta m/z$; 0.0078$\Delta m/z$]. The first factor γ_{min} is the Δm^* to m^* ratio of ^{79}Br, the second the γ_{max} from ^{1}H. Without such an assumption, the γ are simply calculated over all chemical elements. One must be aware, however, of the rounding involved in the calculation of mass defects: any change in Δm to values above 0.5 consequently wraps to $\Delta m-1$, values below -0.5 convert to $\Delta m+1$. Thus, ranges of Δm must be extended accordingly. A geometric representation of the discussed bounds in $\Delta m/z$ and Δm between adjacent series peaks is illustrated in Figure 1 and more formally defined in the next section. Finally, the change in retention time ΔRT between adjacent series peaks is also restricted in order to reflect reasonable chromatographic changes caused by repeated introduction of chemical units.[30,31] First, and similar to the above range of Δm, ΔRT_{min} and ΔRT_{max} define minimum and maximum bounds for ΔRT, respectively. A symmetric example of these bounds is given by the query rectangles in the top panel of Figure 1. Second, absolute changes in ΔRT from one tuple pair ($p_{j,k}$, $p_{j+1,k}$) to the next ($p_{j+1,k}$, $p_{j+2,k}$) must be smaller than a predefined value $\Delta\Delta RT$. Third, shifts in RT have to be systematic.[30,31] To this end, cubic smoothing splines are fitted to model RT as a function of m/z in each series.[39] Briefly, the model fit of each series as determined by the coefficient of determination (R^2) has to be above a certain threshold, using a preset smoothing parameter $\lambda \geq 0$.

4.2.2 Series detection. Constrained by the above definitions, series detection from a set of LC-HRMS peaks progresses in two stages. A first stage extracts the set S_3 of all possible *3*-tupels, a second one combines them to larger tuples of $n>3$ in a stepwise manner.

The first stage uses *k*-dimensional (*k*-d) trees as a metric data structure to support computationally fast peak queries.[40] Herein, each peak x is represented by a node with vector $a_x \in \mathbb{R}^4$

$$a_x=(m/z_x, \Delta m_x - \gamma_{min}\, m/z_x, \Delta m_x - \gamma_{max}\, m/z_x, RT_x) \quad (4)$$

A geometrical depiction of the elements in a_x is given by the blue lines in Figure 1. The second and third elements a_{x_2} and a_{x_3} transform the minimum and maximum change in mass defect with increasing peak mass to a scale

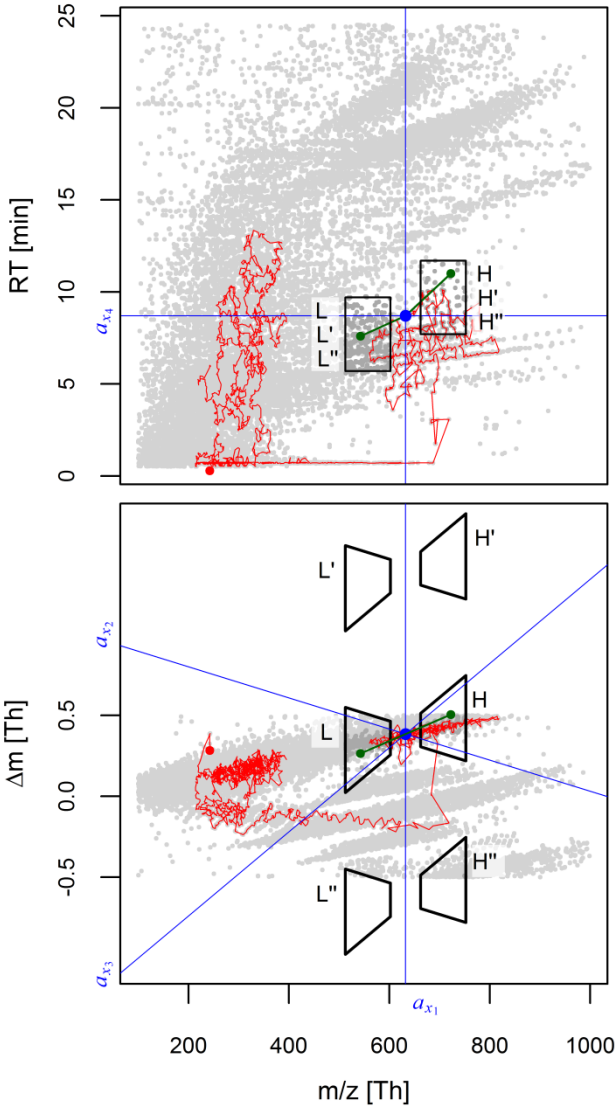

RT [min]
0
5
10
15
20
25
a_{x4}
L
L'
L"
H
H'
H"
Δm [Th]
-0.5
0.0
0.5
a_{x2}
a_{x3}
a_{x1}
200
400
600
800
1000
m/z [Th]

Figure 1 (last page). *Exemplary subspace query (black polygons) for the detection of 3-tuples, centered at the blue peak point after 2499 re-centering steps along the NN path (red line, random starting point shown as red dot) through the picked peaks (gray points). Intersection of blue lines with axes indicate the values of the four elements of a_x for the centered peak. One detected 3-tuple is exemplified in green, with the center peak as its second element. Note that the lower (L) and higher (H) query subspaces are stacked in the top panel.*

that can be represented in a *k*-d tree. In the latter, each tree node alias peak splits the space into two partitions, using the median of one of the variables in a_x. The resulting two partitions are in turn split by the next variable in a_x, each by another median peak contained in those partitions (cp. numbered splitting planes in the TOC for an arbitrary example in $\mathbb{R}^3$). Starting with the first entry of a_x for the root node and recursively cycling over entries of a_x until partitions with single peaks (i.e., terminal nodes) are reached, such a binary search tree supports fast range queries. In these queries, two subspace types *L* and *H* defined further below with lower and higher *m/z* repeatedly centered around each LC-HRMS peak are queried for all unique peak combinations that can (a) form *3*-tuples in accordance with the above series definition and (b) include the current center peak as second element of a *3*-tuple (cp. green dots in Figure 1). Aspects (a) and (b) are built on a linear search over all queried peaks in L and H, ordered by their absolute $\Delta m/z$ relative to the center peak. This computational procedure is facilitated in two ways. First, the peaks within the subspaces usually represent only a small fraction of all LC-HRMS peaks available and the check of (a) and (b) greatly improves over a check using all peaks. Second, the step size from one center peak to the next, i.e., the concomitant change of a_x, can be kept small. The amount of peaks that then leave and enter the re-centered subspaces is minimized, too. As a consequence, most queried peaks are already partially ordered, accelerating any new ordering of absolute $\Delta m/z$ relative to the next center peak.[41] Similarly, branches in the *k*-d tree which were fully enclosed in the subspace of one center peak and are fully enclosed for the next need not be traversed again. Here, the nearest neighbor (NN) algorithm with an Euclidean distance was used to construct a path for re-centering, after normalization of each dimension with its range (Figure 1, red line).[42] The above mentioned subspaces result from combining the intervals

$$I_1 = [a_{x_1} + \Delta m/z_{min};\ a_{x_1} + \Delta m/z_{max}] \quad (5)$$
$$I_2 = [a_{x_2} - 2\varepsilon; \infty] \quad (6)$$
$$I_3 = [-\infty; a_{x_3} + 2\varepsilon] \quad (7)$$
$$I_4 = \left[a_{x_4} + RT_{min}; a_{x_4} + RT_{max}\right] \quad (8)$$
$$I_5 = [a_{x_1} - \Delta m/z_{max}; a_{x_1} - \Delta m/z_{min}] \quad (9)$$
$$I_6 = [-\infty; a_{x_2} + 2\varepsilon] \quad (10)$$
$$I_7 = \left[a_{x_3} - 2\varepsilon; \infty\right] \quad (11)$$
$$I_8 = [a_{x_4} - RT_{max}; a_{x_4} - RT_{min}] \quad (12)$$

via their Cartesian products to the queried subspaces

$$H = I_1 \times I_2 \times I_3 \times I_4 \quad (13)$$
$$H' = I_1 \times (I_2 + [1;0]) \times (I_3 + [0;1]) \times I_4 \quad (14)$$
$$H'' = I_1 \times (I_2 - [1;0]) \times (I_3 - [0;1]) \times I_4 \quad (15)$$
$$L = I_5 \times I_6 \times I_7 \times I_8 \quad (16)$$
$$L' = I_5 \times (I_6 + [0;1]) \times (I_7 + [1;0]) \times I_8 \quad (17)$$
$$L'' = I_5 \times (I_6 - [0;1]) \times (I_7 - [1;0]) \times I_8 \quad (18)$$

Intervals I_1 to I_4 define bounds in each of the four dimensions of a_x for subspaces of type H; I_5 to I_8 define the subspaces of type L. The translations with $[1;0]$ and $[0;1]$ account for the mentioned rounding issue of Δm. In practice, translated subspaces need only be queried if they intersect with the feasible subspace of peak data within $-0.5 \leq \Delta m \leq 0.5$. In summary, unprimed and primed L contain potential peaks to precede the centered peak in a detected 3-tuple, H those to succeed; Figure 1 exemplifies all subspaces in $\Delta m/z$ and Δm.

The second stage successively combines all detected 3-tuples to larger ones. To this end, all pairwise combinations of tuples x and y from a set S_n which only differ in their first and last peak members, i.e.,

$$(p_{1,x}, \ldots, p_{n-1,x}) = (p_{2,y}, \ldots, p_{n,y}) \quad (19)$$

or

$$(p_{2,x}, \ldots, p_{n,x}) = (p_{1,y}, \ldots, p_{n-1,y}) \tag{20}$$

and conform to the above series definitions concerning ε, $\Delta\Delta RT$ and λ are combined to a new $(n+1)$-tuple in S_{n+1}. Having formed all combinations from S_n, the resulting tuples in S_{n+1} are in turn recombined to larger tuples in the next set S_{n+2}. This is repeated until an empty set is reached. While a n-tuple is free to combine to several different $(n+1)$-tuples, a peak cannot appear more than once in each tuple. In every recursion, n-tuples which can be combined to at least one new $(n+1)$-tuple in S_{n+1} are removed from S_n; they otherwise remain in S_n. Moreover, redundant sub-tuples generated in this procedure need to be removed at each new $S_{n\geq 5}$ formed. Namely, for any n-tuple with a given $\Delta m/z$ in $S_{n\geq 5}$ there exist another w smaller tuples with $\beta\Delta m/z$, if $\beta\Delta m/z \leq \Delta m/z_{max}$ and $n/\beta > 2$. These smaller tuples are composed by regular omission of elements in the larger n-tuple, e.g., of each second element for $\beta = 2$. Their count w for each such n-tuple is given via the modulo operation

$$w = \begin{cases} \beta & if \quad n \bmod \beta = 0 \\ n \bmod \beta & if \quad n \bmod \beta \neq 0 \end{cases} \tag{21}$$

Finally, all developed n-tuples smaller than a minimum user-defined length n_{min} are discarded.

4.2.3 Series pairing. In general, a peak may be a member of more than one series, as its containing 3-tuples might have been incorporated into several different larger tuples instead of a single one. To elucidate the underlying reasons for such ambiguities, all unique series pairs that intersect in at least one peak of a LC-HRMS sample were extracted and their properties characterized twofold.

On the one hand, the intersection angle θ was used to approximate in how far two series x and y of such a pair were superjacent in the plane of RT and m/z. θ is defined as

$$\cos\theta = \frac{u_x \cdot u_y}{\|u_x\|\|u_y\|} \tag{22}$$

In this equation, numerator and denominator state the dot product and the product of the Euclidean norm of vectors with scaled mean values

$$u_x = \left(\frac{\overline{\Delta RT_x}}{c_{\Delta RT}}, \frac{\overline{\Delta m/z_x}}{c_{\Delta m/z}}\right) \tag{23}$$

of each series, respectively. Here, $c_{\Delta RT}$ and $c_{\Delta m/z}$ are the range of $\overline{\Delta RT}$ and $\overline{\Delta m/z}$ over all series.

On the other hand, a self-organizing map (SOM) was used to visualize and cluster common patterns among the paired series.[43,44] Being an unsupervised learning strategy, SOMs allow the mapping of a large set of *m* multidimensional input vectors $v=(v_1,\ldots,v_j,\ldots,v_m)$ onto a static two-dimensional grid of SOM neurons, each having a weight vector *W* of length equal to the input vectors. The grid-like SOM can then be selectively displayed for the mapped properties. Properly trained maps preserve the topological characteristics in *v* despite a projection to a grid with numbers of neurons ≪ m. Consequently, similar input vectors are mapped to close regions in the SOM while different ones are rather separated. In the given case, each input vector

$$v_j = \left(\frac{\overline{\Delta RT_x}}{\hat{c}_{\Delta RT}}, \frac{\overline{\Delta m/z_x}}{\hat{c}_{\Delta m/z}}, \frac{\overline{\Delta RT_y}}{\hat{c}_{\Delta RT}}, \frac{\overline{\Delta m/z_y}}{\hat{c}_{\Delta m/z}}\right) \tag{24}$$

contains the mean values in a pair of series *x* and *y*, arranged by $\overline{\Delta m/z_x} \geq \overline{\Delta m/z_y}$. In contrast to the range scaling of equation (23), $\hat{c}_{\Delta RT}$ and $\hat{c}_{\Delta m/z}$ represent the expected mean measurement uncertainties in ΔRT and $\Delta m/z$, respectively. During training, vectors from *v* are sequentially used to update both their best-matching neuron *d* and other neurons in a shrinking neighborhood of *d*, with matching based on Euclidean distances. At each such sequential iteration *t* with a current v_j, the update of each node vector W_i is calculated via

$$W_i(t+1) = W_i(t) + \Phi(d,i,t)\alpha(t)[v_j - W_i(t)] \tag{25}$$

where $\alpha(t)$ denotes a learning rate which here declines linearly over the iterations. Similarly, $\Phi(...)$ defines a shrinking rectangular neighborhood around W_d; W_i outside of this neighborhood remain unchanged. Overall, the full data set v was presented multiple times to the SOM for training, which was randomly initialized from v for $t=0$ with a toroidal rectangular grid. The quality of the trained SOM was assessed by the quantization and topological errors E_q and E_t, respectively.[44,45] The first error is the mean distance of all vectors in v to their best matching node in the final map, expressed in terms of unscaled ΔRT and $\Delta m/z$ values. Thus, E_q states the accuracy with which the input vectors are represented by the SOM. Additionally, the latter metric E_t measures the continuity of the mapping by the distance between the best and second best matching unit averaged over all vectors in v and expressed in terms of grid coordinates. Finally, indication of clusters formed in the SOM was derived from the U-matrix.[46] This matrix depicts the mean Euclidean distance of the weight vector of each neuron to those of its immediate neighbors. Having the same size as the map, the U-Matrix can be superimposed on the SOM for visualization. SOM calculations were conducted with the R *kohonen* package, with parameters listed in Table S-1.[47]

4.2.4 Sampling and Analysis. Evaluation was carried out on 24 h flow-proportional samples taken from the effluent of ten Swiss sewage treatment plants in February 2010, as used and detailed in Schymanski et al.[17] In short, a sample volume of 0.25 L was each pH-adjusted, filtered, spiked with 103 isotope-labeled standards and enriched via a mixed-bed solid-phase extraction. After basic/acidic extraction, further enrichment under a nitrogen gas stream, reconstitution with HPLC water to 1 mL and a second filtering step, a final aliquot of 20 μL was analyzed with HPLC-ESI-HRMS. The chromatographic step comprised Waters XBridge C18 columns (Milford, U.S.) and a water/methanol gradient at a flow rate of 200 μL/min generated by a Rheos 2200 low pressure mixing pump (Flux instruments, Basel, Switzerland). A Q-Exactive (Thermo Fisher Scientific, San Jose, USA) was used for full-scan mass spectrometric analysis at a resolution of 140.000 at m/z = 200, following electrospray ionization in each positive and negative modes (spray voltage +4 and -4 kV, respectively; 300 °C capillary temperature). A blank measurement was run prior to each block of positive and negative sample aliquots, respectively. Further sample-specific details are provided in Table S-3.

4.2.5 Data processing. LC-HRMS full-scan data were centroided and converted to open .mzXML format files with ProteoWizard (version

3.0.7162).[48,49] All downstream analysis was then run in the R statistical environment.[50] Utilizing the R package *enviPick* (version 1.2),[51] each file of data points was partitioned, its ion chromatograms extracted and the resulting EICs then screened for signal peaks, with parameters listed in the appendix of the thesis, Table S-2. Upon peak-picking, series were detected with the above outlined algorithm, as parameterized in Table S-4. For each peak being part of a series, both a blank subtraction and a deisotoping was run with the *enviMass*[51] (version 2.0) and the *nontarget*[1] (version 1.9) packages, respectively (see Tables S-5 and S-6 for parameters). In the first case, a peak-centered *RT* and *m/z* window was checked for each sample peak to not contain raw blank data points higher than 0.1 times the maximum sample peak intensity to certify its presence in the effluent. A majority rule, i.e., a fraction of ≥ 0.5 peaks per series, was used for a final assignment of a series to be of blank origin. For deisotoping, a comparison with quantized simulation data enabled a grouping of the isotopologue peaks of an unknown compound, within given measurement uncertainties. The peaks in the individual isotopologue groups of each series peak were then ranked by increasing *m/z*. A series was assumed to be monoisotopic if the most frequent rank over all peaks in a series equaled 1.

4.3 Results & Discussion

4.3.1 Series inventory and recovery. On average (± standard deviation, SD), 21153±3052 and 10418±831 peaks were picked from the LC-HRMS measurements of the 10 STP samples in positive and negative ionization modes, respectively (Table S-3). A substantial mean fraction of 0.37±0.09 of these peaks could be assorted into series for the positive mode, whereas a smaller and less variant fraction of 0.13±0.03 was assorted in the negative mode. Overall numbers of series peaks were strongly correlated with the total number of picked peaks in a STP sample, despite dominating the measured set of picked peaks at only one location (STP ID 8, positive mode). In turn, series counts were correlated with the fraction of series peaks for both ionizations although the length of individual series varied greatly, from five and up to 30 peaks. Notably, series counts were often on the same order as the peak counts of which they were comprised, for reasons discussed in the next section. Overall, 7576±4222 and 1018±494 series were detected in positive and

negative modes, respectively. The large SD was mainly driven by one STP (ID 8, Table S-3).

To further test the presented algorithm, a set of eight HS compounds was utilized to recover their series. These compounds had each at least five of their homologues tentatively identified in a majority of the discussed STP samples in a previous study conducted by Schymanski et al.[17]. As a ground truth, the full peak series of four HS compounds were consistently recovered in all ten STP samples by the algorithm. The peak series of the remaining HS compounds were each recovered in a minimum of three samples; series in the remaining samples could not be detected because either not all peaks of a series were consistently picked at lower intensities or had erratic RT behavior. In at least six cases homologue peaks in addition to those targeted and tentatively identified in the named study were detected. In another six cases, some of the series peaks were also integrated in series other than those targeted by the named study.

Moreover, much lower series counts were observed in the two blank measurements. Only few of the STP sample series were conversely removed via majority voting during the blank subtraction step, i.e., series fractions of 0.10±0.06 (positive ionization mode) and 0.07±0.03 (negative ionization mode). Their absolute numbers correlated negatively with the total number of picked peaks in a sample, which may be explained by varying degrees of matrix suppression of blank signals in more crowded samples. Of the remaining non-blank series, fractions of 0.46±0.13 (positive) and 0.27±0.8 (negative) series contained sporadic peaks which did not pass the blank subtraction individually. On the one hand, this stresses the requirement to run the blank subtraction after peaks were assorted into series instead of before, so as to avoid sporadic series gaps which impede series detection. On the other hand, removing all sample series with sporadic blank peak assignments would overestimate counts of such sample blank series by an order of magnitude as compared to series counts found in the blank measurements. Similar uncertainties existed for the filtering of monoisotopic series, with their counts listed in column 8 of Table S-3. Fractions of 0.72 (positive) and 0.46 (negative) of monoisotopic series contained infrequent peaks with *m/z* rank ≠1, which is in line with the false positive rates of isotopologue grouping. Using ensembles of peaks in each series after the series detection step instead of an earlier deisotoping based on singular peaks can thus improve deisotoping.

4.3.2 Series computation. The various restrictions for ΔRT and $\Delta m/z$ localized at each center peak decrease the computational burden of detecting

meaningful *3*-tuples. On average, $4.3x10^5$ and $1.0x10^5$ *3*-tuples were detected, which represent massively reduced mean fractions of $2.8x10^{-7}$ and $5.0x10^{-7}$ of all possible *3*-tuple peak combinations in samples for the positive and negative ionizations, respectively. Of these triplets, fractions of only 0.13 (positive) and 0.08 (negative) passed into *4*-tuples through pairwise combinations; passed fractions then strongly increased towards higher *n*-tuple combinations. Therein, smoothing spline fits helped to exclude fractions of up to 0.14 *4*-tuple combinations with erratic $\Delta\Delta RT$, which would have been recombined and passed on otherwise. Increasingly void of such erratic combinations, the named exclusion fraction dropped mostly to zero at $n \geq 5$. Optional criteria such as the similarity of chromatographic peak shapes or the distribution of $\Delta m/z$ and intensity in a series may be approached in future versions. Overall, computation time never exceeded 4.1 minutes per sample on a standard computer, including parsing of all necessary outputs, and decreased rapidly with the number of detected triplets. For negative mode samples, computation time was hence below 0.5 minutes (Windows 7, R version 3.1.3, 2.2 GHz Intel core i7-4702 MQ processor, 32 GB RAM, 64 bit).

4.3.3 Superjacent series. The incorporation of a single peak into different series was common to all samples and ionization modes. Dominant mean fractions of 0.99±0.01 (positive) and 0.96±0.02 (negative) series thus shared peaks with other series. Often, much more than one such sharing per existed per series, leading to a multitude of series pairs with at least one peak in common (last two columns of Table S-3). A SOM was therefore trained to discern patterns in the properties of these series pairs, here exemplified for one positively ionized STP sample (ID=1, Table S-3). The resulting SOM is shown in Figure 2, with mapping errors $E_q(\overline{\Delta m/z}) = 0.02$ *Th*, $E_q(\overline{\Delta RT}) = 0.12$ *min* and $E_t = 1.6$. The top panel delineates several regions in the SOM which have similar $\overline{\Delta m/z}$ and $\overline{\Delta RT}$ properties of paired series, segregated from each other by the gray shading of U-Matrix values. Dots in the middle panel in turn size the number of series pairs *(x,y)* mapped onto individual SOM nodes, superposing the colored $\overline{\Delta m/z_x}$ distribution of these grid nodes. The lowest panel only indicates pair numbers involving monoisotopic series, superposing the $\overline{\Delta m/z_y}$ distribution. To recall, $\overline{\Delta m/z}$ is the mean $\Delta m/z$ of a single series; the concomitant SOM distributions of $\overline{\Delta RT_x}$ and $\overline{\Delta RT_y}$ can be found in Figure S-7.

Based on the SOM, several observations can be made. First, although series pairs with a wide array of different $\overline{\Delta m/z}$ and $\overline{\Delta RT}$ values exist, many pairs

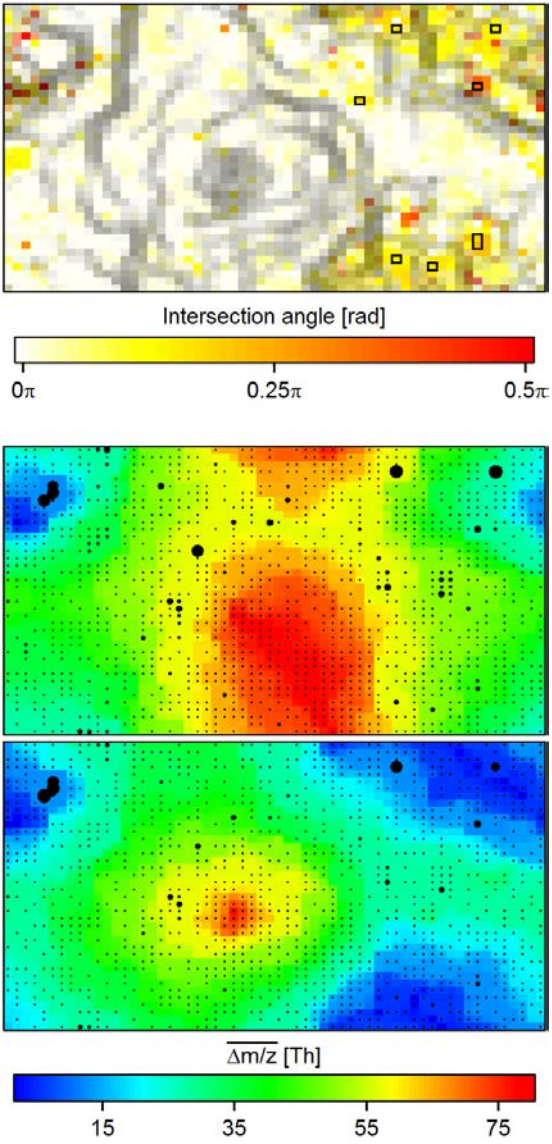

Intersection angle [rad]
0π
0.25π
0.5π
Δm/z [Th]
15
35
55
75

Figure 2 (last page). *SOM results for all series pairs from STP sample with ID=1, positive mode. The top panel shows both high U-Matrix values in gray shading and the distribution of intersection angles θ in the SOM grid in heat colors. Coloring of middle and lower panels depict $\overline{\Delta m/z_x}$ and $\overline{\Delta m/z_y}$ at these grid nodes, respectively. Dots in the middle panel indicate frequencies of series pairs clustered onto a SOM nodes, relative to the most frequent one. In contrast, lower panel dots indicates such mapping numbers after deisotoping. In turn, black squares in the top panel intersect nodes of (a) highest mapping numbers containing 50% of all monoisotopic series pairs and (b) $\theta \geq 0.08\pi$; the concerned series are plotted in Figure 3.*

nevertheless cluster at certain nodes. In fact, just 13% of the nodes are able to summarize 90% of the pairs. Second, the contribution of non-monoisotopic series herein is noteworthy, affecting 42% of the pairings. Third, a majority of series pairs map onto SOM regions with very low intersection angles and are largely superjacent, i.e., they are similarly positioned in the *RT* and *m/z* dimensions. Using a histogram-derived threshold of $\theta < 0.08\pi$, this affects a predominant 81% of series pairs in the considered STP sample (Figure S-8; cp. last column of Table S-3 for other STPs). Based on an inspection of the LC-HRMS data, it can be concluded that superjacent series frequently result from close-eluting isobaric peaks. If overlapping in the *ΔRT* window of different tuples, isobaric peaks can cause an exponential increase in the number of possible combinations for forming series from these tuples. For example, 2^n series combinations of comparable $\overline{\Delta m/z}$ arise for n pairs of isobaric peaks each located at different *m/z* values. Isobaric peaks from homologue isomers are indeed common and may require additional analytical separation to be extractable as fully non-superjacent series.[16,52] Another less frequent reason for superjacent series was the sporadic occurrence of missing peaks in otherwise continuous series, e.g., at series ends with diminishing measurement intensities. As a result, closely superjacent series with $\overline{\Delta m/z}$ being multiples of each other are detected. Because the affected series are no strict subsets of each other, they cannot be eliminated during the removal of subtuples at the end of the second stage of the algorithm. An aggravated example for illustrating superjacency caused by both the presence of isobaric peaks and series gaps is provided in Figure S-9. To clarify, $\overline{\Delta m/z}$ values being

multiples of each other can also arise for differently charged adducts of the same homologue series; these multiples are however not superjacent and can thus be distinguished. Similarly, the different series of the different isotopologues of a homologue compound are unlikely superjacent but rather parallel in orientation in the *m/z* vs. *RT* plane.

4.3.4 Meshed series. A notable 19% of series pairs were not superjacent but instructively arranged. For exemplification, a set of most strongly clustered monoisotopic series pairings with $\theta \geq 0.08\pi$ were selected from their SOM projection, as indicated by the seven black squares in the top panel of Figure 2. The chosen series are in turn plotted in Figure 3 and comprise seven distinct $\overline{\Delta m/z}$ characteristics. One first group of interrelated series embraces $\overline{\Delta m/z}$ values of 14.016, 44.026, 30.011 and 58.042 *Th*, with multiple pairings between these values. One may hypothesize that the first two values might stem from Alkyl (CH_2) and Ethoxylate ($C_2H_4O_1$) homologue units of variable length located at the same compounds. Confirmingly, chemically joint occurrence of both units has been confirmed for various surfactants in the considered STP sample via the targeted approach by Schymanski et al.[17] and has also been reported in STP effluents elsewhere.[13,16,53] With (a) both units coexisting at all their differing lengths and (b) varying *RT* increases for both the resulting chains, a mesh-like orientation of these series in the *m/z* vs. *RT* plane can be anticipated. This meshing is indeed observable in the zoom area of the upper panel of Figure 3, plotted in Figure S-10. In addition, the mutual orientation of both series types allows for further cross-meshing, formed by a difference ($C_1H_2O_1$) and a sum ($C_3H_6O_1$) of the former two homologue units. This overall hypothesis is also in agreement with observed mass defect shifts *Δm*, which are smaller for higher O to C/H ratios in these four series types (lower panel of Figure S-10). A similar meshing occurs for a second unrelated group of series, comprising $\overline{\Delta m/z}$ values of 7.008, 29.021, 51.034 and 58.042 *Th*. This second group is likely a result of adduct formation at *z=2*, considering (a) concomitant mass defect shifts, (b) the first two values being halves of the above discussed $\overline{\Delta m/z}$ values of 14.016 and 58.042 *Th* and (c) the latter two values formable by multiples and subtractions among the former two.

Several implications of the outlined meshing must further be stressed. First, series meshing does not only provide complementary information, but can also prevent false conclusions: a $\overline{\Delta m/z}$ value of 58.042 *Th* may as well suggest the occurrence of a propylene oxide unit instead of a sum of two different units.[54] As a matter of fact, other series with $\overline{\Delta m/z}$ = 58.042 *Th* not par-

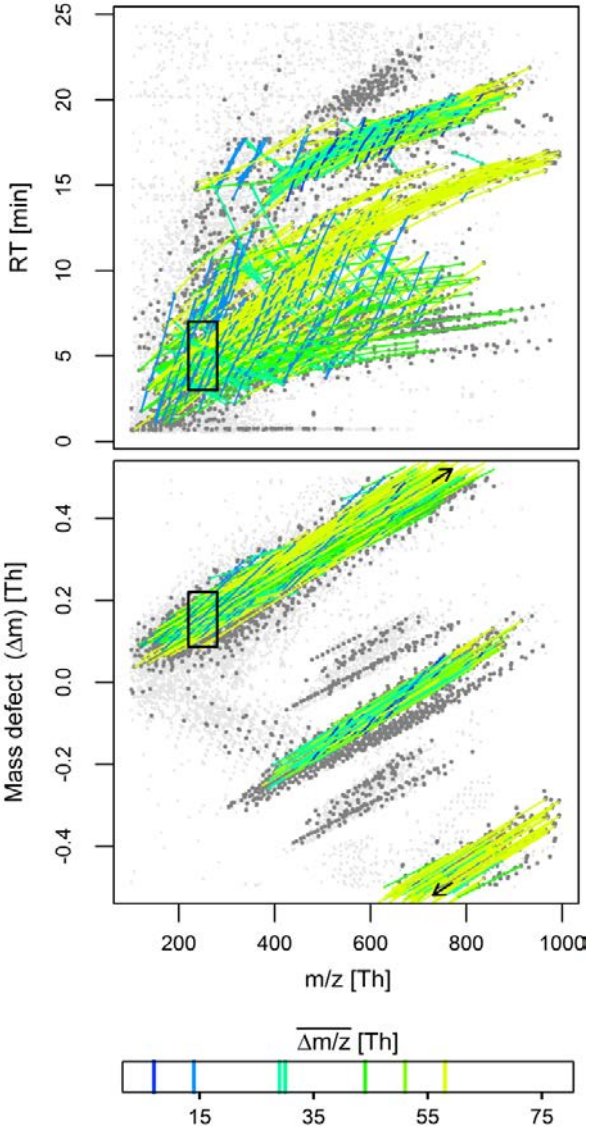

RT [min]
0
5
10
15
20
25
Mass defect (Δm) [Th]
-0.4
-0.2
0.0
0.2
0.4
200
400
600
800
1000
m/z [Th]
Δm/z [Th]
15
35
55
75

Figure 3 (last page). *Characteristics of frequently paired series selected from SOM analysis. Light gray points show all picked peaks whereas only those in dark gray were incorporated into series. Bold rectangles show the zoom areas detailed in Figure S-10.*

ticipating in any meshing occur in the very same STP sample, but have yet to be chemically identified. Second, negative *RT* changes ($\Delta RT_{min}<0$) can arise for cross-meshed series such as the above one with $\overline{\Delta m/z}$ = 30.011 *Th*, even when the *RT* is expected to increase with the length of the underlying chemical homologue chains. Third, cross-meshed series with $\overline{\Delta m/z}$ values not matching any molecular formula can arise if the atoms of the homologue units by which they differ cannot form subsets. In the above first example group, $C_2H_4O_1$ minus CH_2 equals $C_1H_2O_1$; however, a hypothetical $C_2H_4O_1$ minus CF_2 would in contrast not suggest a valid molecular formula. Fourth, meshed series may have fixed sets of $\overline{\Delta m/z}$ values but likely a more variable set of $\overline{\Delta RT}$ values. In a SOM, this latter variation is covered by several mapping nodes, which should nonetheless be close to each other in the SOM if the topological continuity holds (cp. one black square in the top panel of Figure 2 that comprises two adjacent nodes). Finally, the complexity of series meshing will rise with the number of homologous chains per compound. Even for the discussed example, further additions and subtractions from cross-meshing of $(CH_2)_2$ and $(C_2H_4O_1)_2$ units exist, but were less preponderant in the SOM and hence not selected here.

4.3.5 STP comparison. To complement the hitherto exemplification of series patterns from the single STP sample, Figure 4 ultimately stacks the $\overline{\Delta m/z}$ distributions of all blank-corrected series for every STP at both ionization modes (panels A and C, black lines) and filters for $\overline{\Delta m/z}$ values prevalent across STPs (panels B and D, gray bars). Noteworthy, a multiplicity of $\overline{\Delta m/z}$ values exist, which would not be revealed under targeted restrictions of $\Delta m/z$. Many of these values are highly conserved across the different STPs, although at different frequencies and with less diversity in the negative than in the positive ionization mode. Among the most frequent, especially in the negative mode, are the three discussed values of $\overline{\Delta m/z}$ = 14.016, 44.026 and 58.042 *Th*, partly corresponding to alkyl, ethoxylate and possibly propylene oxide units (red solid lines).[17] The larger frequency of the latter again suggests another origin than the mere addition of the former two units as pre-

sented above, both at charges *z=1* and *z=2*. Other than that, a large but still incomprehensive fraction of the remaining $\overline{\Delta m/z}$ values might be annotated via either charge- or gap-related multiples or additions/subtractions of these three units, albeit tentatively until identified as such (red dashed and gray bars). Moreover, seven of the most ubiquitous yet low-frequent $\overline{\Delta m/z}$ values among STPs in positive mode almost disappear when non-monoisotopic series are excluded from the cumulative frequency analysis (gray dashed lines, blue bars). Their values occur around major non-affected ones at mass differences equal to those between ^{12}C and ^{13}C and may involve series of different isotopologues of different carbon-rich members of homologue series. However, without further identification attempts – which can now gain from additional information on series meshing and $\overline{\Delta m/z}$ co-occurrence across STPs – such annotations remain largely speculative. Given the prevalence of some $\overline{\Delta m/z}$ values, detected series may nonetheless be engaged to cluster different STPs, to quantify the ubiquity of series across STPs or to find similarities of unpaired series arising from, e.g., transformations by a second SOM training.

4.4 Implementation

The outlined algorithm is freely available as function *homol.search()* in the R package *nontarget*.[1] Parameters $\Delta m/z_{min}$, $\Delta m/z_{max}$, ΔRT_{min}, ΔRT_{max}, $\Delta\Delta RT$, n_{min}, ε, λ, R^2 and the chemical elements to fix equation 3 can all be user-defined (cp. Table S-4). Optionally, values for $\Delta m/z$ can be specified for a more targeted series detection or to confine the numbers of computed series in samples with even higher HS contents, e.g., oil extracts. Spline smoothing can be disabled and $\Delta\Delta RT$ increased to also comprise series with erratic *RT* behavior, but will almost certainly trigger more false positive series as a trade-off. Series results can finally be tagged to adduct and isotopologue groups to derive component peak sets. Series related via a user-specified θ can be grouped; more extended clustering such as SOM is out of the package`s scope.

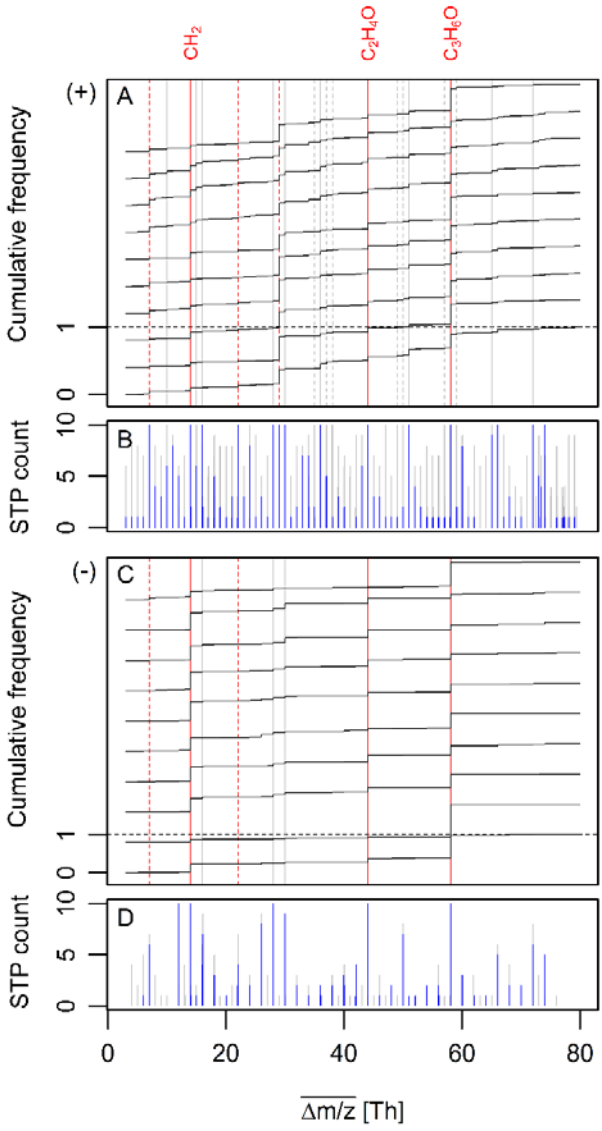

CH2
C2H4O
C3H6O
(+)
A
Cumulative frequency
1
0
B
STP count
10
5
0
(-)
C
Cumulative frequency
1
0
D
STP count
10
5
0
0
20
40
60
80
Δm/z [Th]

Figure 4 (last page). *Relative cumulative frequency of $\overline{\Delta m/z}$ values for all blank-subtracted series detected in positive (panel A) and negative (panel C) ionization modes, stacked top-down for individual STP samples in order of their IDs 1 to 10 (cp. Table S-3). Solid red lines indicate masses of three common homologue units at z=1, red dashed ones at z=2. Solid gray lines denote $\overline{\Delta m/z}$ values of possible multiples, additions or subtractions thereof. Gray dashed lines indicate isotopologue shifts of some of these masses, equal to ^{12}C vs. ^{13}C mass transitions. Moreover, gray bars in panel B and D show STP counts from a moving $\overline{\Delta m/z}$ window (±5 mu) over all stacked distributions for the positive and negative mode, respectively. In contrast, blue bars were derived after omission of non-monoisotopic series.*

4.5 Conclusion

Given the large throughput in LC-HRMS experiments, a visual detection of systematic signal patterns to pinpoint the presence of unknown homologous compounds from the accumulated data is futile. Hence, an untargeted yet efficient bottom-up computation of picked peak series with systematic changes in their mass and retention time is presented and evaluated. With just a minimum of prior information to confine this detection, the presented algorithm will reveal measurable series regardless of their specific soft-ionization species, eventual modifications during such ionization or nonlinear *RT* shifts. Without the need to bin data under variable measurement uncertainties, the algorithm enables detection of low-frequent and low-intense series even in complex matrices, provided that series peaks are properly picked and reach a minimum but adjustable series length. Furthermore, non-random inclusion of peaks into different series proved useful to discern possible ambiguities in assigning peaks to series and to identify series meshing caused by homologues with more than a single variable chemical unit. Future research may implement gap-tolerant versions of the proposed algorithm and further data mining to automatize both the digestion of the wealth of observed series interrelations and to subsequently propose unambiguous masses of the underlying chemical units for improved identification as a next step.

REFERENCES

(1) Loos, M. *nontarget: Detecting Isotope, Adduct and Homologue Relations in LC-MS Data*; 2015.

(2) DeWitt, J. C. *Toxicological Effects of Perfluoroalkyl and Polyfluoroalkyl Substances*; Springer, 2015.

(3) Martin, J. W.; Smithwick, M. M.; Braune, B. M.; Hoekstra, P. F.; Muir, D. C.; Mabury, S. A. *Environmental Science & Technology* **2004**, *38* (2), 373–380.

(4) Olsen, G. W.; Ellefson, M. E.; Mair, D. C.; Church, T. R.; Goldberg, C. L.; Herron, R. M.; Medhdizadehkashi, Z.; Nobiletti, J. B.; Rios, J. A.; Reagen, W. K.; others. *Environmental science & technology* **2011**, *45* (19), 8022–8029.

(5) Goodacre, R.; Heald, J. K.; Kell, D. B. *FEMS Microbiology Letters* **1999**, *176* (1), 17–24.

(6) Oberacher, H.; Walcher, W.; Huber, C. G. *Journal of mass spectrometry* **2003**, *38* (1), 108–116.

(7) Schneiter, R.; Brügger, B.; Sandhoff, R.; Zellnig, G.; Leber, A.; Lampl, M.; Athenstaedt, K.; Hrastnik, C.; Eder, S.; Daum, G.; others. *The Journal of cell biology* **1999**, *146* (4), 741–754.

(8) Seebach, D. In *Peptides: The Wave of the Future*; Springer, 2001; pp 569–571.

(9) Rodrigues, C. M.; Rinaldo, D.; Santos, L. C. dos; Montoro, P.; Piacente, S.; Pizza, C.; Hiruma-Lima, C. A.; Brito, A. R.; Vilegas, W. *Rapid Communications in Mass Spectrometry* **2007**, *21* (12), 1907–1914.

(10) Yassin, G. H.; Koek, J. H.; Jayaraman, S.; Kuhnert, N. *Journal of agricultural and food chemistry* **2014**, *62* (40), 9848–9859.

(11) Hughey, C. A.; Rodgers, R. P.; Marshall, A. G. *Analytical Chemistry* **2002**, *74* (16), 4145–4149.

(12) Roach, P. J.; Laskin, J.; Laskin, A. *Analytical chemistry* **2011**, *83* (12), 4924–4929.

(13) Clara, M.; Scharf, S.; Scheffknecht, C.; Gans, O. *Water Research* **2007**, *41* (19), 4339–4348.

(14) Gawlik, B.; Bidoglio, G. *European Commission, Brussels* **2006**.

(15) Lin, P.; Rincon, A. G.; Kalberer, M.; Yu, J. Z. *Environmental science & technology* **2012**, *46* (14), 7454–7462.

(16) Ruan, T.; Song, S.; Wang, T.; Liu, R.; Lin, Y.; Jiang, G. *Environmental science & technology* **2014**, *48* (8), 4289–4297.

(17) Schymanski, E. L.; Singer, H. P.; Longree, P.; Loos, M.; Ruff, M.; Stravs, M. A.; Vidal, C. Ripolle s; Hollender, J. *Environmental science & technology* **2014**, *48* (3), 1811–1818.

(18) Stenson, A. C.; Landing, W. M.; Marshall, A. G.; Cooper, W. T. *Analytical Chemistry* **2002**, *74* (17), 4397–4409.

(19) UNEP. *UNEP Publications* **2005**.

(20) Zeng, L.; Li, H.; Wang, T.; Gao, Y.; Xiao, K.; Du, Y.; Wang, Y.; Jiang, G. *Environmental science & technology* **2013**, *47* (2), 732–740.

(21) Petrovic, M.; Farré, M.; De Alda, M. L.; Perez, S.; Postigo, C.; Köck, M.; Radjenovic, J.; Gros, M.; Barcelo, D. *Journal of Chromatography A* **2010**, *1217* (25), 4004–4017.

(22) Krauss, M.; Singer, H.; Hollender, J. *Analytical and bioanalytical chemistry* **2010**, *397* (3), 943–951.
(23) Hsu, C. S.; Qian, K.; Chen, Y. C. *Analytica chimica acta* **1992**, *264* (1), 79–89.
(24) Kilgour, D. P.; Mackay, C. L.; Langridge-Smith, P. R.; O'Connor, P. B. *Analytical chemistry* **2012**, *84* (17), 7431–7435.
(25) Kendrick, E. *Analytical Chemistry* **1963**, *35* (13), 2146–2154.
(26) Kim, S.; Kramer, R. W.; Hatcher, P. G. *Analytical Chemistry* **2003**, *75* (20), 5336–5344.
(27) Reemtsma, T. *Journal of mass spectrometry* **2010**, *45* (4), 382–390.
(28) Wu, Z.; Rodgers, R. P.; Marshall, A. G. *Analytical chemistry* **2004**, *76* (9), 2511–2516.
(29) Carlson, J. E.; Gasson, J. R.; Barth, T.; Eide, I. *Chemometrics and Intelligent Laboratory Systems* **2012**, *114*, 36–43.
(30) Héberger, K. *Journal of Chromatography A* **2007**, *1158* (1), 273–305.
(31) Kaliszan, R. *Chemical Reviews* **2007**, *107* (7), 3212–3246.
(32) Marchetti, N.; Felinger, A.; Pasti, L.; Pietrogrande, M. C.; Dondi, F. *Analytical chemistry* **2004**, *76* (11), 3055–3068.
(33) Pietrogrande, M.; Perrone, M.; Sangiorgi, G.; Ferrero, L.; Bolzacchini, E. *Talanta* **2014**, *120*, 283–288.
(34) Pietrogrande, M. C.; Bacco, D.; Marchetti, N.; Mercuriali, M.; Zanghirati, G. *Talanta* **2011**, *83* (4), 1225–1232.
(35) Pietrogrande, M. C.; Zampolli, M. G.; Dondi, F. *Analytical chemistry* **2006**, *78* (8), 2579–2592.
(36) Kunenkov, E. V.; Kononikhin, A. S.; Perminova, I. V.; Hertkorn, N.; Gaspar, A.; Schmitt-Kopplin, P.; Popov, I. A.; Garmash, A. V.; Nikolaev, E. N. *Analytical chemistry* **2009**, *81* (24), 10106–10115.
(37) Brenton, A. G.; Godfrey, A. R. *Journal of the American Society for Mass Spectrometry* **2010**, *21* (11), 1821–1835.
(38) Sleno, L. *Journal of mass spectrometry* **2012**, *47* (2), 226–236.
(39) Hastie, T. J.; Tibshirani, R. J. *Generalized additive models*; CRC Press, 1990; Vol. 43.
(40) Bentley, J. L. *Communications of the ACM* **1975**, *18* (9), 509–517.
(41) Singleton, R. *Comm. ACM* **1969**, *12* (3), 185–187.
(42) Johnson, D. S.; McGeoch, L. A. *Local search in combinatorial optimization* **1997**, *1*, 215–310.
(43) Kohonen, T. *Neural Networks* **2013**, *37*, 52–65.
(44) Kohonen, T.; Schroeder, M.; Huang, T.; Maps, S.-O. *Inc., Secaucus, NJ* **2001**, *43*.
(45) Kiviluoto, K. *Topology preservation in self-organizing maps*; Helsinki University of Technology, 1995.
(46) Ultsch, A.; Siemon, H. P. *Proceedings of the International Neural Networks Conference* **1990**, 305–308.
(47) Wehrens, R.; Buydens, L. M. C. *J. Stat. Softw.* **2007**, *21* (5).
(48) Holman, J. D.; Tabb, D. L.; Mallick, P. *Current Protocols in Bioinformatics* **2014**, 13–24.

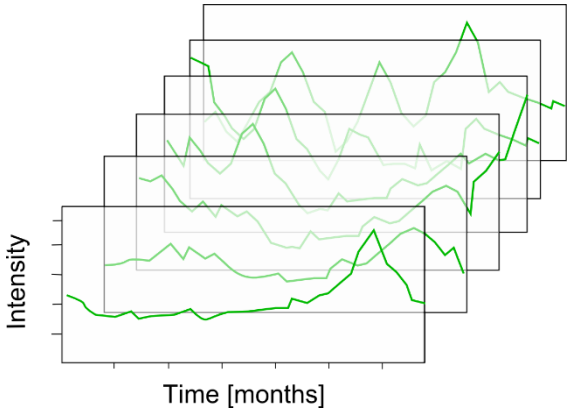
(49) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24* (21), 2534–2536.
(50) Team, R. C. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2015.
(51) Loos, M. *enviMass: Utilities to Process Mass Spectrometry (LC-HRMS) Data for Environmental Trend Analysis*; 2014.
(52) Erdem, N. S.; Alawani, N.; Wesdemiotis, C. *Analytica chimica acta* **2014**, *808*, 83–93.
(53) Corada-Fernández, C.; Lara-Martın, P. A.; Candela, L.; González-Mazo, E. *Journal of Environmental Monitoring* **2011**, *13* (7), 2010–2017.
(54) Little, J. Identification of surfactants in commercial products by mass spectrometry, 2012.

Chapter 5

enviMass 2 – a Workflow for Fast Micropollutant Spill Detection from Large LC-HRMS Measurement Sequences

Martin Loos, Matthias Ruff, Steffen Ruppe, Jan Mazacek, Heinz Singer

Paper draft



ABSTRACT: The widespread and continuing emission of a complex set of known and unknown micropollutants into riverine systems is a well-known phenomenon. With toxic effects even at low concentrations, these polar pollutants jeopardize human safety and ecological functions, and thus require a long-term monitoring. Although liquid chromatography (LC) and high-resolution mass spectrometry (HRMS) has been well-established for this task, an automatized trend and spill detection of micropollutants from the resulting measurement sequences has nevertheless remained challenging. To this end, a multi-stage data mining workflow is introduced, complemented with a versatile user interface and tailored to a sequential updating with new measurements. First stages in the workflow comprise chromatogram clustering, peak detection, comprehensive preprocessing, signal grouping, homologue series detection, and the extraction of intensity-time profiles. The latter are then mined for trends of concern and ranked. When tested with a sequence of over 700 LC-HRMS measurements at an international monitoring station for the river Rhine, a variety of alarming spill events could be detected and several non-target compounds thereby prioritized for both a subsequent identification and an investigation of the emitting sources. Chromatogram extraction and peak-picking as well as the full trend detection workflow are publicly available as R packages *enviPick* and *enviMass*, respectively.[1,2]

5.1 Introduction

Considerable quantities of anthropogenic chemicals are either continuously or sporadically emitted into our riverine environments. One ubiquitous point source of emission is the effluent of industrial, municipal and hospital wastewaters, releasing a multitude of, e.g., surfactants, personal care products, industrial agents, pharmaceuticals, illicit drugs, or perfluorinated compounds.[3,4] Another input arises by diffuse lateral transport from distant source sites, either directly through runoff and surface waters or, somewhat delayed, through groundwater percolation.[5,6] Emissions via this route include pesticides, biocides, and fuel additives, to name a few. Regarding the universe of emitted compounds, some aspects must be highlighted. First, a large, if not dominant, fraction of occurring compounds is neither suspected nor known a priori, and not even comprised in the array of emerging contaminants.[7] This, for instance, concerns unregistered industrial intermediates or transformation products (TPs). As a result, such compounds cannot be directly targeted in chemical analysis. Second, the emitted universe of man-made compounds is rather complex, while being relatively polar and hence mobile. Third, despite being mostly present at trace-level concentrations, these compounds can nonetheless exert adverse toxicological effects.[8] The levels of these so-called micropollutants in riverine systems must consequently be monitored to assess or mitigate impacts on human safety and ecosystems. For this purpose, high-performance liquid chromatography (LC) coupled to electrospray ionization (ESI) and high-resolution mass spectrometry (HRMS) has emerged as an analytical solution for high-throughput detection of the multitude of low-volatile compounds.[9]

LC-HRMS has to date been employed for routine micropollutant detection at several monitoring stations, for example in the Rhine River network.[10] There, long-term monitoring has amassed big-data LC-HRMS measurement sequences. Under focus different than for targeted compounds, non-target analysis has subsequently mined such sequences using various strategies and workflows, and with different perspectives. In the environmental context, approaches have extracted non-target signals common to several measurements or have aggregated signals to principal components to elucidate major variation among data sets.[11–14] However, neither of both strategies can reveal time-intensity trends of LC-HRMS signals that could point at the increasing presence or even spill events of individual micropollutants in the river Rhine.

On the other hand, much efforts has been invested to compile automatized methods in the research fields of metabolomics and proteomics, where LC-HRMS sequences are frequently mined to compare biological conditions and treatments over time.[15–18] Unfortunately, these alternatives rarely account for the intricacy of environmental monitoring. For example, some popular pipelines such as *MZmine*,[19] *CAMERA*,[20] *XCMS*,[21] *AStream*,[22] or *RAMClust*[23] do not explicitly include a trend analysis that could be used for spill detection.

Yet other implementations are incapable to mine larger sequences (including the predecessor to the proposed work, *enviMass* version 1.2).[24,25] Other pipelines do not reduce raw data to picked peaks, which is a requirement to process big-data LC-HRMS sequences.[26] Furthermore, another crucial step in data reduction and interpretation is the deisotoping and adduct grouping of LC-HRMS measurements. Some approaches simply lack this step.[27,28] Others use simplifications which will not embrace the patterns of chlorine, bromine or sulfur isotopes, often used as characteristic marker to single out pollutant molecules.[20,22,29,30] Other workflows provide more versatile deisotoping, but for low resolutions MS and/or only for specific compound classes, e.g., peptides or lipids.[31–34] Other approaches have only addressed subproblems, e.g., peak matching,[35] visualization,[36] or chromatogram analysis[37] and therefore remain somewhat fragmentary. Further limitations arise from the practical context of routine monitoring. Most importantly, already processed sequences of several hundred measurements must be swiftly updated on a regular basis; data processing must reveal trends of concern within short time spans to react upon them and a graphical user interface should guide the chemical analyst through decisive steps.

On this background, a first comprehensive workflow dedicated to the fast spill and trend detection of micropollutants in aquatic systems is introduced, able to process and update several hundred LC-HRMS measurements. Herein, recent advances in high-resolution signal grouping and homologues series detection are considered.[2] Coded in the R statistical environment[38] and supplemented with a versatile interface, the workflow is tested as part of the working routine at the Rhine monitoring Basel (RÜS), Switzerland.

5.2 Methods

Data processing with the proposed workflow is streamlined into five consecutive stages (A) to (E), outlined in the following sections and summarized in the scheme of Figure 1. Input to the workflow is a temporal LC-HRMS measurement sequence, which can be sequentially updated with new data. Time points in this sequence are uniquely occupied by one sample and/or one blind measurement (i.e., no replicates are involved), and processed separately for negative and positive ESI modes.

5.2.1 Stage (A) – Partitioning and peak picking. The first stage detects distinct signal peaks in centroided and baseline-corrected LC-HRMS data, for each of *i* samples and *j* blind measurements. Herein, three steps are employed. The usually vast set of recorded data points is first partitioned into smaller and unrelated subsets, so as to accelerate and potentially parallelize the subsequent steps. A clustering of these partitions to extract individual ion chromatograms (EICs) ensues in a second step, followed by peak detection in these EICs in a last step.

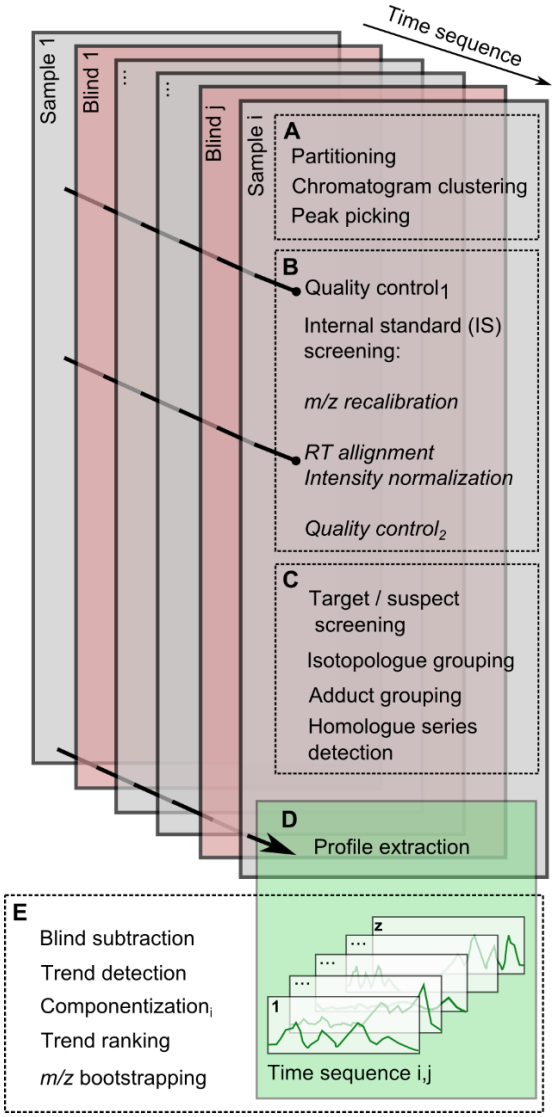


Figure 1. *The five data-processing stages embedded in the enviMass workflow.*

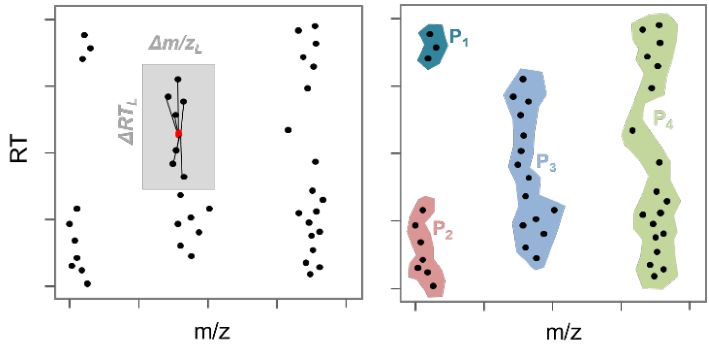


Figure 2. *Partitioning of LC-HRMS data points. First, every data point is linked to its neighbors within large tolerances of m/z and RT, here exemplified for the one red data point (left panel). Directly and indirectly data points are then grouped into partitions (P), separated by void space (right panel).*

In more detail, the set of LC-HRMS data points is first divided by linking each data point $m=\{m/z, RT, intensity\}$ to all its neighbors found within large mass and retention time (RT) tolerances $\Delta m/z_L$ and ΔRT_L, respectively (Figure 2, left panel). These tolerances must amply exceed the mass accuracy and RT gaps anticipated to occur between EICs. Next, all data points which can be directly or indirectly linked are grouped into the same partition P_x (Figure 2, right panel). With $P=\{P_{x=1}, \ldots, P_{x=n}\}$ denoting the full set of partitions, k data points $m_x=\{m_{x,1}, \ldots, m_{x,k}\}$ are placed into each partition and ranked by decreasing intensity. The number of data points contained in individual partitions is thus variable and often increases over the mass range. Although more links between data points than in, e.g., recent Delaunay triangulations[28] need to be made, the approach is faster at slightly higher memory requirements and ensures the partitions to be fully unrelated. In contrast, relations among data points within one partition are still required to be resolved, whereas data points from different partitions can be assumed to be unrelated.

Second, data points m_x in each partition are further refined for their affiliation in mass and retention time. This is achieved by first detecting distinct cluster and then merging them. The formation of cluster is initialized with an empty cluster set $C_x=\emptyset$. Any cluster y to be added to this set contains p data points $m_{x,y}=\{m_{x,y,1}, \ldots, m_{x,y,p}\}$ and is assigned with individual mass and retention time intervals $I_{m/z}(x,y)$ and $I_{RT}(x,y)$. Let ε furthermore be a ± mass accuracy, $RT_{step} << \Delta RT_L$ a small predefined retention time increment, and $m/z_{x,1}$ and $RT_{x,1}$ denote the mass-to-charge ratio and retention time of any data point $m_{x,1}$. Then, cluster formation proceeds in an intensity-descending manner:

(1) Select data point $m_{x,1}$ from the intensity-ranked data set m_x.

(2) Can $m_{x,1}$ be assigned to any cluster y of w existing cluster $C_x=\{C_{x,y=1}, \ldots, C_{x,y=w}\}$ by (a) $m/z_{x,1} \in I_{m/z}(x,y)$, (b) $RT_{x,1} \in I_{RT}(x,y)$, and (c) no other already clustered point data $m_{x,y,i\leq p}$ having an RT value identical to $RT_{x,1}$?

No:

a. Set $m_{x,w+1,1} = m_{x,1}$.
b. Add new cluster $C_{x,w+1} = \{m_{x,w+1,1}\}$ to C_x.
c. Set $I_{m/z}(x,y) = [m/z_{x,1} - 2\varepsilon;\ m/z_{x,1} + 2\varepsilon]$.
d. Set $I_{RT}(x,y) = [RT_{x,1} - RT_{step};\ RT_{x,1} + RT_{step}]$.

Yes:

e. Add $m_{x,1}$ to the cluster y with smallest difference in mean m/z, i.e., $m_{x,y,p+1} = m_{x,1}$.
f. Update the cluster mass range:
$I_{m/z}(x,y) = [max(m/z_{x,cy}-2\varepsilon);\ min(m/z_{x,cy}+2\varepsilon)]$.
g. Update the cluster RT inclusion range:
$I_{RT}(x,y) = [min(RT_{x,cy}-RT_{step});\ max(RT_{x,cy}+RT_{step})]$.

(3) Remove $m_{x,1}$ from m_x.

(4) Repeat (1) to (3) until no data points are left, i.e., $m_x=\emptyset$.

Scheme 1: Description of the intensity-descend clustering routine.

Often, high-intense data points can be associated with better mass accuracies and occur rather isolated in intensity from the bulk of medium- to low-intense data points, which in turn are more likely to contain noise peaks.[39] Thus, the former instantiate cluster first in the above intensity descend, while improving on $I_{m/z}(x,y)$ towards ε in a rather small neighborhood restricted by RT_{step}. Thus, the concept is fundamentally different from greedy uni- or bidirectional scan-to-scan assignments of data points.[28,31,40] The locally assorted data points are then used to merge some of the derived cluster, with each of the cluster constituting a full EIC after merging:

(1) List all pairs of cluster in C_x with overlapping $I_{m/z}(x,y)$ and not containing data points with same $RT_{x,1}$.

(2) Merge the cluster pair with smallest difference between their two mean *m/z* values, if available.

   a. Update $I_{m/z}(x,y)$ of the merged cluster, as in Scheme 1.

   b. Adjust the list C_x for the merged cluster pair.

(3) Repeat (1) to (2) until no more cluster can be merged. This final set of cluster represents the EICs in a partition.

Scheme 2: Description of the cluster merging routine.

Following partitioning, cluster formation and cluster merging, a final step detects distinct peaks in the resulting w EICs $C_x=\{C_{x,y=1}, \ldots, C_{x,y=w}\}$, with each EIC containing p data points. This step aims to resolve isobaric compounds, to filter out noise signals, to discriminate against analytes that form chemical baselines without sufficient *RT* separation, and to derive a final data-condensed list of picked peaks. The peak detection approach is illustrated in Figure 3 and does not presuppose a specific peak shape; a prior EIC smoothing can be run optionally. The underlying algorithm to detect up to q peaks per EIC y is hence defined as:

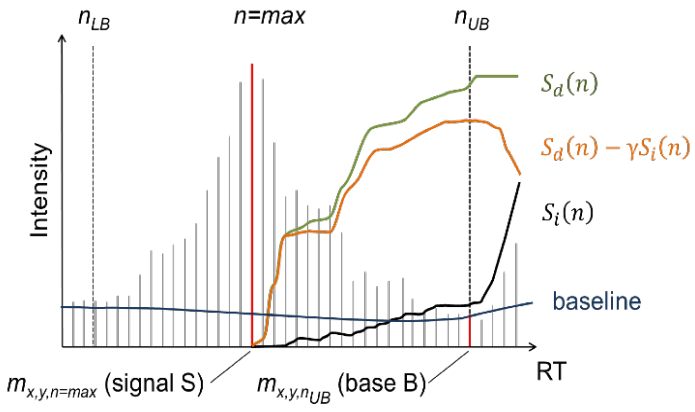


Figure 3. *Peak detection in an EIC (gray bars), initiated at the most intense data point (red bar). Notations are specified in the main text.*

(1) Order $m_{x,y}$ in RT and linearly interpolate data gaps $\leq RT_{gap}$.

(2) Select the most intense data point $m_{x,y,max}$ as candidate peak apex.

(3) Set $S_d(n)$ to be the sum of intensity decreases between $m_{x,y,max}$ and $m_{x,y,n}$.

(4) Similarly, set $S_i(n)$ as the sum of intensity increases.

(5) For EIC data points $n>max$, select an upper peak bound n_{UB} via $\text{argmax}_n\, S_d(n) - \gamma S_i(n)$.

(6) For EIC data points $n<max$, select a lower peak bound n_{LB} fulfilling $\text{argmax}_n\, S_i(n) - \gamma S_d(n)$.

(7) Retrieve the mean m/z, mean RT and maximum signal intensity S for data points $n_{LB} \leq n \leq n_{UB}$ for a preliminary peak candidate. Remove the data points from the set $m_{x,y}$ of $C_{x,y}$.

(8) Repeat (1) to (7) at most q times.

(9) Filter peak candidates by:

   a. Intensity threshold.
   b. Minimum number of data points over one predefined RT stretch.

(10) If $m_{x,y} \neq \emptyset$ and any peak candidates remain:

   a. Estimate a baseline intensity: linearly interpolate the intensity gaps caused in step (7) between the remaining points in $m_{x,y}$ and smooth intensities outside gaps with splines.
   b. Retrieve the baseline intensity B at each apex RT position and subtract it from peak candidate intensity S.
   c. Define noise N of $\text{EIC}_{x,y}$ as the median intensity deviation of $m_{x,y}$ from their baseline intensity.

(11) Filter out peak candidates ranging below threshold S/N and S/B ratios.

Scheme 3: Description of the EIC peak picking routine.

Those picked peaks fulfilling above aspects (9) and (11) are assorted into lists for each individual sample and blind measurement and passed to the downstream workflow stages.

5.2.2 Stage - (B) Data preprocessing. The second stage makes a first pass over all samples and blind peak lists for two quality-control (QC) checks and a correction of systematic deviations among the lists. The latter relies on spiked internal standard (IS) compounds to be evenly distributed over the m/z and RT ranges, in the best case isotopically labeled in order to minimize interferences with the sampled matrix. The first quality check compares the

logarithmic intensity quantile distributions of peak lists, in groups of j blinds and i samples. Single distributions with an extreme deviation from the mean of all distributions are highlighted and can hint at problems in generating the individual peak lists, e.g., in sampling, sample preparation, ionization or from parameterization issues during peak picking.

Thereafter, each peak list is screened within large tolerances $\Delta RT_{screen,large}$ and $\Delta m/z_{screen,large}$ for the centroid peaks of the spiked IS compounds and the matches filtered for their consensus with the expected isotopologue peak patterns of these IS compounds. The matches are then utilized in four ways. First, a spline fit is conducted to correct systematic offsets between the m/z values of theoretical and matched IS centroids as a function of m/z (Figure 4). Second, and taking advantage of the m/z-recalibrated data, those landmark IS peaks that can be traced over a fraction $\geq \tau$ of $i+j$ peak lists are utilized to (a) align RT shifts, and (b) normalize systematic intensity variations among the lists. Aspect (a) engages the mean RT of an IS signal matched over the peaks lists as one reference point for a spline model, in analogy to the mass recalibration step. This spline model can then be used to correct systematic RT shifts in the peak lists, as a function of elution time. In contrast, aspect (b) calculates the median M_i logarithmic intensity of a landmark signal matched over τ lists, separately for every IS compound i. With each peak list, the median log-intensity difference between the IS landmark peaks and their individual M_is is used for an intensity normalization. While such normalization retains the intensity ratios among peaks in a list, it refrains from including landmarks with too low τ coverage to avoid any systematic bias.

Finally, outliers in the difference between uncorrected minus normalized intensity versus the number of matched IS peaks per list are checked in a second quality control step. The number of IS peak matches per list can be expected to increase with the mentioned difference. Therefore, gross deviation from this pattern can point at problems in sample enrichment or simply IS compound spiking.

5.2.3 Stage (C) - Peak grouping and screening. Another stage assorts the peaks in each list into non-target isotopologue and adduct groups and detects peak series with systematic m/z and RT shifts to indicate the possible presence of homologue series (HS), as outlined in chapters 3 and 4 of this thesis, respectively.[41] In addition, peak lists are also screened for the theoretical centroids of target and suspect compounds, within a relatively small mass tolerance $\Delta m/z_{screen,small}$. In the case of targets, a tolerance $\Delta RT_{screen,small}$ around an expected RT value defined by comparison with reference standards is screened. For suspects, ΔRT_{screen} spans the full elution range as their chromatographic properties are likely unknown. Notably, the dependence on the preprocessing of stage (B) differs among the functionalities of this stage

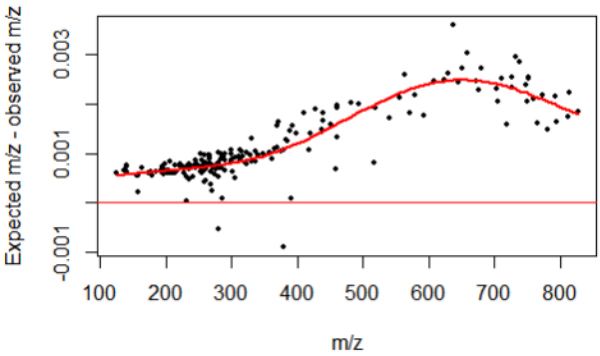


Figure 4. *Smoothing spline fit (red curve) used to interpolate and recalibrate the offset between expected and measured IS masses from a zero deviation (red straight line). In this arbitrary example, the maximum interpolated deviation amounts to ~ 3.5 ppm.*

(C). For example, whereas the target screening improves from the previous mass recalibration, the rather small *m/z* differences between non-target isotopologues to be grouped are hardly affected by this recalibration (cp. Figure 4: changes in the difference between the two red lines are relatively small for closely spaced *m/z* values).

5.2.4 Stage (D) - Profile extraction. As indicated by the black arrow in Figure 1, the fourth stage makes a final pass over all peak lists to extract peaks that are likely associated with the same analyte signal measured over consecutive LC-HRMS experiments, jointly for sample and blind peak lists. For this purpose, peaks are tagged with their membership in a list and pooled over all samples and blinds. After that, a partitioning similar to stage (A) segregates regions in this pool unrelated in *RT* and *m/z*. Thereafter, an approach analogical to the intensity-descent clustering of scheme 1 is applied to each peak region, except for two aspects. The first concerns point (2c) of scheme 1: instead of the *RT* criterion, simultaneous membership in the same list is the criterion to not add the considered peak to a cluster alias profile. Second, the update with RT_{step} in step (2g) is omitted and a fixed retention time window ΔRT_{prof} centered on the first (and hence most intense) peak to instantiate a profile is used. To summarize, stage (D) produces sets of time profiles, listing intensities over time for peaks matched in *m/z* and *RT* over the LC-HRMS sequence of blind and sample measurements.

5.2.5 Stage (E) - Trend detection and componentization. A last stage elaborates on the previously extracted profiles to detect increasing intensity trends which may be indicative of, e.g., accidental pollutant spills. The un-

derlying detection strategy is schematized in Figure 5, with the profile sample intensity I_t measured at time points t shown as green line. In the same plot, a red line quantifies the measurement intensity B_t of blind signals, which needs to be interpolated if not consecutively available for all time points t. Indicated by a gray line, a bounded mean intensity $L_{t,n}$ with lag n is then calculated for each profile time point t from

$$L_{t,n} = \begin{cases} I_t & if \ \alpha_{t,n} \geq I_t \\ \alpha_{t,n} & if \ \alpha_{t,n} < I_t \end{cases} \quad (1)$$

using an average lag intensity of

$$\alpha_{t,n} = \frac{1}{n} \sum_{i=(t-n)}^{t-1} I_t \quad (2)$$

Based on $L_{t,n}$, a candidate trend x is thereupon registered for each consecutive temporal intensity stretch having $I_t > L_{t,n}$, and the highest intensity $I_{x,max}$ in such a stretch recorded as the candidate trend intensity (cp. gray area and green dot in Figure 5). Herein, several lag values $n=\{1, ..., m\}$ are included to cover various temporal trend scales. Next, for each candidate trend intensity x, the mean μ_{-x} and the variance σ_{-x} of the trend intensities over all other candidate trends in a profile are formed; if no other candidates exist, both μ_{-x} and σ_{-x} are set to zero. Each candidate trend x is ultimately filtered to be genuine by the criteria

$$I_{x,max} > \mu_{-x} + \theta_1 \sigma_{-x} \quad (3)$$

and

$$I_{x,max} > \theta_2 B_{t,x} \quad (4)$$

where θ_1 and θ_2 specify predefined thresholds to multiply with. $B_{t,x}$ is the blind intensity registered at the time point for which an $I_{x,max}$ is found in the profile. In other words, a genuine trend must (a) exceed a lagged intensity signal, (c) be significantly higher than other candidate trends and their intensity variations found in the same profile, and (c) surpass the intensity of background contaminations present in the blind measurements.

Moreover, similarly shaped profiles can arise for the different isotopologues of an analyte and, without further data aggregation, may consequently indicate redundant trends. In addition, some trends may have shifted into the past

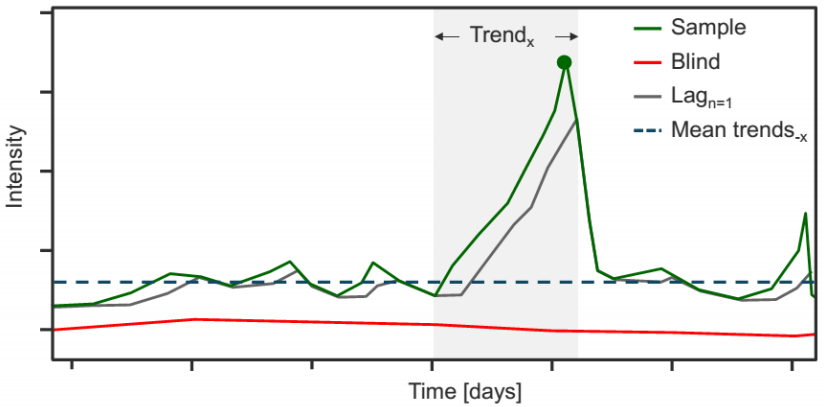


Figure 5. *Scheme of the trend detection strategy. A measured sample intensity (green line) is compared to a lagged mean intensity (gray line) and blind/blank intensities (red line) to assign a trend period (gray area) with maximum intensity $I_{x,max}$ (green dot).*

from the updating with new measurements and are not any longer of current concern. Integrating both aspects, profiles containing genuine trends are therefore combined into chemical components via the isotopologue and adduct groups generated in stage (C) from the peak list of the most recent sample *i*. Only one genuine trend of maximum intensity $I_{x,t=1}$ per component, if any, is then added to a ranked list which summarizes and prioritizes all the current trends. Moreover, if a detected trend is produced by an unknown component and its profile consists of enough peaks, (a) improved estimates of the mean *m/z* and its statistical confidence intervals, and (b) information on the monoisotopic mass can be derived from bootstrapping and the grouped adduct and isotopologue relations, respectively. Both can be instrumental for the success in identification steps outside of this workflow.

5.2.6 Sampling and analysis. A prototype of the presented workflow was integrated into the measurement routine at the RÜS Basel and there tested for Rhine River samples. A consecutive total of daily 675 sample and 75 weekly blind LC-HRMS measurements was assembled in between June 2012 and April 2014 and discussed here; further updating with new measurements has however continued since then. Daily time-proportional, cross-sectional sampling was conducted at the automated sampling station Weil am Rhein. Samples were filtered, adjusted to pH 6.7, spiked with 132 isotopically labeled IS compounds and enriched by solid phase extractions (mixed bed cartridge, ionic exchange mode) by a factor of ~ 1000.[42] Chromatographic separations

were performed on a Waters XBridge reverse phase column (Milford, U.S.), with a water to methanol gradient containing 0.1% formic acid. Full-scan mass spectrometric analysis in positive and negative ESI modes was run on a LTQ-Orbitrap at a resolution of 60.000 at m/z = 400 (Thermo Fisher Scientific, San Jose, USA). The RÜS curates a list of 288 polar organic target compounds incorporated into the workflow, including plant protection products, pharmaceuticals, industrial chemicals, corrosion inhibitors, and several of their transformation products (TPs).

5.2.7 Parametrization. Stage (A) parameterization for partitioning, clustering ($\Delta m/z_L$, ΔRT_L, ε, RT_{step}), and for peak picking (penalty γ, S/N, S/B, intensity thresholds, number of data points per RT stretch) was done both by inspection of the recovery of the below set of 132 spiked IS compounds and a visual examination of picked raw data with an interactive viewing tool introduced in section 5.3. The parameters are listed in Tables S-1. Similarly, all other parameterization (τ, ΔRT_{prof}, lags n, θ_1, θ_2, $\Delta RT_{screen,large}$, $\Delta m/z_{screen,large}$, $\Delta RT_{screen,small}$, $\Delta m/z_{screen,small}$) was achieved iteratively by examination of all outcomes and changes in profile recovery of IS compounds (Table S-2). Complementary settings for the non-target grouping and homologue series extraction are identical to the thesis chapters 3 and 4, respectively.

5.3 Implementation

The presented workflow is implemented into two software packages which can be loaded and run from the R statistical environment.[43] The first one, *enviPick*, comprises stage (A) of the workflow, i.e., partitioning, clustering, and peak picking.[2] Centroided and baseline-corrected measurements in the .mzXML open data format are required as input to the package; this format can be created from a range of vendor-specific files with the ProteoWizard *MSConvert* tool beforehand.[44] In addition, *enviPick* supports an interactive visualization of both the raw measurement data and attained results from within the R graphics device (Figure 6, A). Progressing from partitions to cluster to peaks, data points are resorted into nested subsets with an indexing for a fast data access.

The remaining stages are implemented in the R package *enviMass*, which automatically loads *enviPick* and a number of other dependencies such as the packages described in previous chapters. Given that R is based on a command-line interface, a more convenient usage of all workflow stages (e.g., management of parameters, file uploads, interactive viewing of results) is facilitated by an additional graphical user interface (GUI). The latter is run via R package *shiny* (and its dependencies) from a web browser and is thereby largely platform independent (Figure 6, B-D).[45] *enviMass* organizes measurements, parameters, stage selections, and results into projects. Each project can contain the parallel sequences of negative and positive ionization measurements simultaneously and the user can easily switch between both.

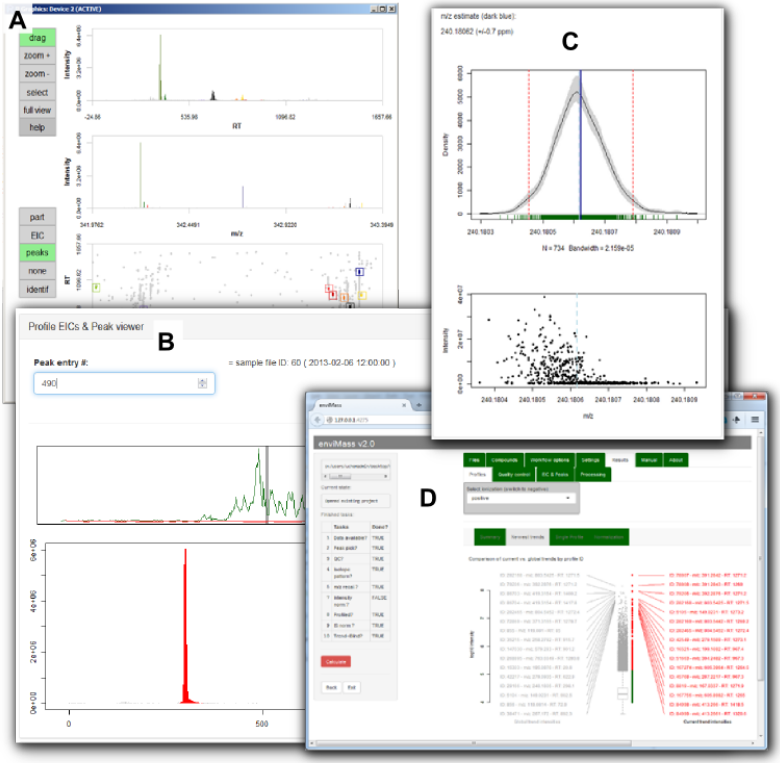


Figure 6. *Interfaces for the presented workflow. (A) enviPick enables users to access their measurement data points and picked peaks interactively. (B) The enviMass GUI lists individual profiles by their intensity ranking and lets user screen through all their peaks and EICs. (C) For a profile with sufficient peak numbers, bootstrapping can retrieve better estimates of the underlying m/z values. (D) The named GUI also lists current trend intensities and compares them against past ones.*

By keeping track of changed settings and upload of new files into these projects, the package minimizes required computation times by avoiding redundant calculations. For example, the full set of 750 measurements needs approximately 1 day for one full processing of all stages, using a computer with a minimum of ten GB RAM but rather standard hardware otherwise. In contrast, an update with one file, i.e., one file-wise calculation of stages (A) and (C) plus full recalculation of stages (B), (D), and (E) will require no more

than 15 minutes. Herein, peak picking alone consumes around three minutes for the tested Orbitrap LC-HRMS measurements. Another two minutes will approximately be spend for grouping and componentization. For IS, target and suspect screening, *enviMass* converts the molecular formulas of compounds into a range of adduct species and swiftly calculates their isotopologue pattern peaks. For this purpose, over 20 resolution functions specific to various MS instruments can be chosen from, as available in the dependency package *enviPat*.[46] Overall, a relatively simple installation of the workflow comprises (a) the R environment, (b) the *enviMass* package installed from therein (all package dependencies will be loaded automatically), and (c) a standard web browser, preinstalled on most computers. Optionally, *MSConvert* can be employed from within the interface to directly load Thermo .raw files.

5.4 Results & Discussion

The sequence of 750 positive ionization measurements resulted in a total of $1.3x10^9$ LC-HRMS data points and could be reduced to a sum of $1.0x10^7$ peaks and $3.0x10^5$ profiles with the presented *enviMass* workflow. Because only a mean of $1.3x10^4$ peaks was detected at a single time point, the comparatively high number of profiles implies that many of these were transient and short-termed. In fact, only a very small fraction of profiles was continuously present over the monitored time period, often constituted by the IS profiles spiked at constant concentrations. Moreover, the sum of profiles passing the blind subtraction as set by threshold θ_2 for at least one time point and those profiles without any blind presence at all amounted to a fraction of 0.67 profiles, which emphasizes the importance to filter against background contamination.

Furthermore, significant portions of peaks could be grouped. In the sample measurements, average fractions of 0.27±0.05 and 0.31±0.06 peaks could be assigned to isotopologue and adduct groups, respectively. Based on this grouping, a mean of 0.47±0.05 profiles could be merged to components at any current time point in the sample sequence. An example for such a component is provided in Figure 7 and even includes a low-intense isotopologue profile containing a ^{15}N isotope. Finally, another mean fraction of 0.22±0.05 peaks per sample were related via series of constant *m/z* shifts and smooth *RT* changes, which are indicators for the presence of homologue series. Notably, and in accordance to findings in chapter 4, a substantial number of such series peaks were contained in profiles that passed additional blind subtraction steps and are therefore unlikely to originate from background contaminations alone.

Backed by these processing steps, staff at the Rhine monitoring station could reveal a number of concerning trends during their daily updates with new

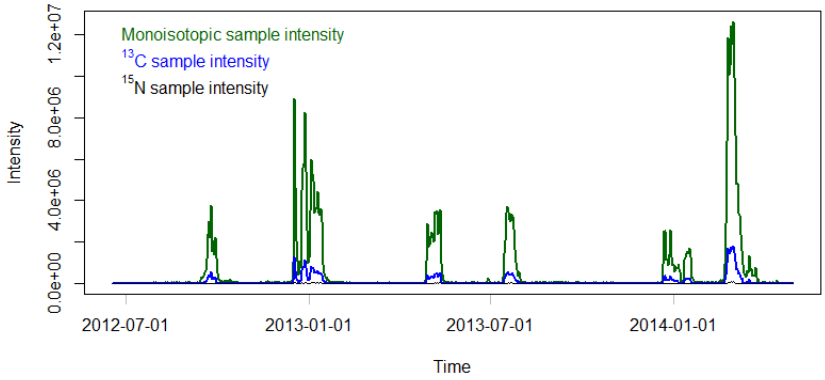


Figure 7. *Componentized enviMass intensity profiles of isotopologues of 2-(Phenyl)-2-(2-piperidin)-acetamid, causing several alarming trends, with a main spike in intensity around March 2014. The component contains the monoisotopic profile as well as two non-monoisotopic profiles of lower intensity.*

measurements and issued several alarms to track the emission sources. A few cases shall be highlighted in the following to demonstrate the capability of *enviMass*. A first non-target case is the one shown in Figure 7. The temporary intensity spikes in the time course of the profile pointed at an anthropogenic source, with a mass that could not be matched to any RÜS target compounds. After one alarming intensity trend in March 2014, molecular formula fits and MS/MS fragmentation spectra led to the candidate molecule 2-phenyl-2-(2-piperidin)-acetamid, which was finally confirmed with a reference standard and quantified. With knowledge of its identity, the pollutant could be traced back to an industrial point source. A second non-targeted case is plotted in Figure 8. Here, two striking trends initiated the identification of methadone at concerning concentrations in the river Rhine. Again, an accidental discharge from an industrial synthesis was the source to be held responsible; the total load estimated after quantification surpassed 80 kg. A third non-target case is that of the solvent tetraglyme, which appeared first in spring 2013, followed by several sporadic trends (Figure 9). While also present at low intensities in the blind measurements, a last intense spill at the end of that year necessitated an alarm; the ion species was identified, the source located and the presence of the chemical in the Rhine thereafter decreased. Another alarming trend was also recorded for the target compound isoproturon, shown in Figure 10. Being a plant protection product, its occurrence in the Rhine coincides with growing seasons and surges in profile intensity coincide with rainfall events that seemingly transfer the chemical

from agricultural sites to the riverine environment. With a concentration range of 0.002-0.18 µg/L exceeding the legal quality threshold of 0.1 µg/L (Schweizer Gewässerschutz-Verordnung), the total load of this compound was calculated to be 50 kg in 2012 alone. In contrast, other profiles hint at a rather constant consumption of, e.g., antidiabetic medication in Switzerland (Figure 11). Yet other profiles show a decline in spike intensities over time (Figure 12) or, despite their characteristic patterns, still await future identification (Figure 13).

In *enviMass*, several workflow steps are kept relatively simple and thereby result in high computational speed. Foremost, profile extraction relies on just three principles: an intensity descend, *m/z* search windows with decreasing uncertainties as peaks get added to a profile, and a fixed *RT* search window. In the vast majority of spill detection cases, this suffices. Any auxiliary information from the apparent correlation between consecutive EICs over samples adjacent in the time sequence has yet to be incorporated. In addition, and because several sections of the workflow are calculated for individual sample and blind measurements separately, updates with newly acquired measurements do not require a full workflow recalculation. Rather, results of stages (A), (C) and partly (B) remain valid for the previously processed samples and blinds. Moreover, workflow stages (B) and (C) may be skipped, except for quality control issues – but otherwise yield improved results and are needed for the later profile componentization.

The sole inclusion of IS compounds for stage (B) can be problematic if either no such compounds are available or do not adequately cover the examined *m/z* and *RT* range. Given the multitude of alignment strategies, alternatives shall find their way into future workflow versions.[16] On the other hand, mass recalibration with *enviMass* can also rely on targeted analytes, although with less confidence than for spiked compounds. Moreover, the intensity normalization of stage (B) was found to be intricate. Correction via the IS set was frequently not in line with alternative normalization methods of, e.g., using the median peak intensity distributions over all peak lists – a method which assumes that the overall relative intensity distribution in a peak list is mostly constant among the different peak lists in a measurements sequence. This observation may be explained by mean concentration changes in the overall matrix composition, for example from lower to higher dilution of the sampled compounds with increasing Rhine discharge.

Another important finding concerns the intensity correlation between profiles of the same analyte. On the one hand, and as anticipated, isotopologue profiles of the same adduct were highly correlated, both for the screened targets and, confirmingly, for the non-target isotopologue groups (data not shown). In contrast, the different adduct profiles of the same analyte were often poorly correlated, even within the individual adduct groups of the spiked IS com-

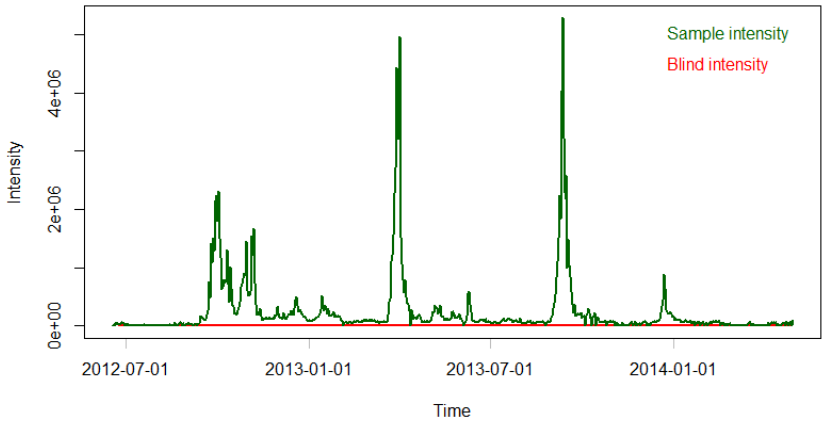


Figure 8. *Intensity time-profile of the monoisotopic mass of the hydrogen adduct of methadone (green), not relatable to any background contamination (red).*

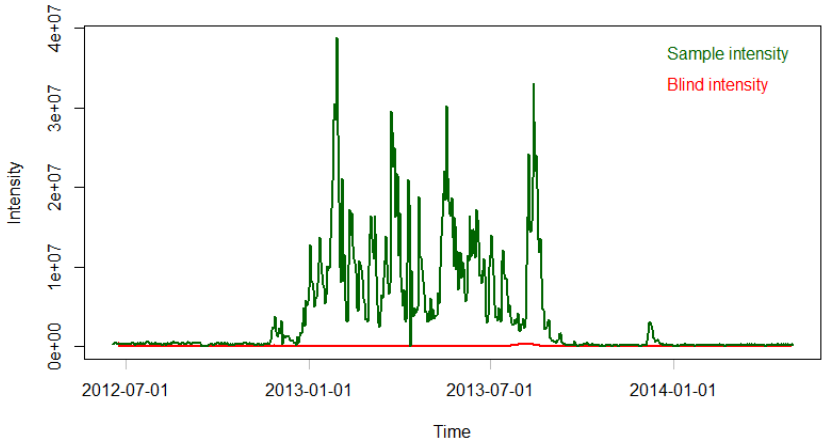


Figure 9. *Intensity time-profile of the solvent tetraglyme (green), registered with little background contamination (red).*

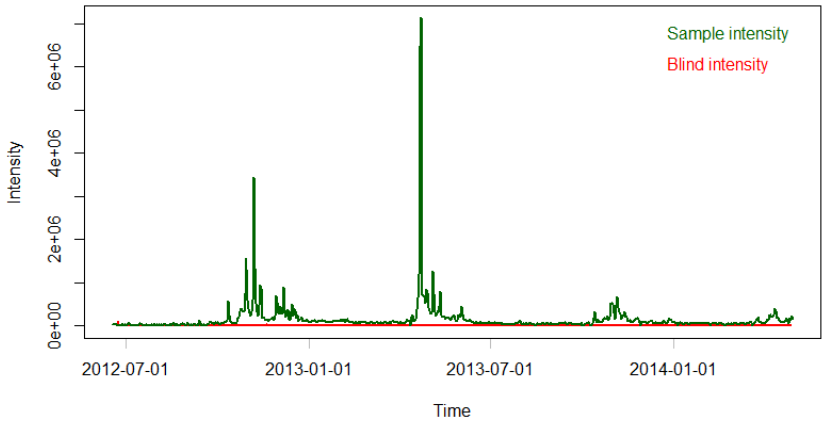


Figure 10. *Intensity time-profile of isoproturon (green), registered without background contamination (red).*

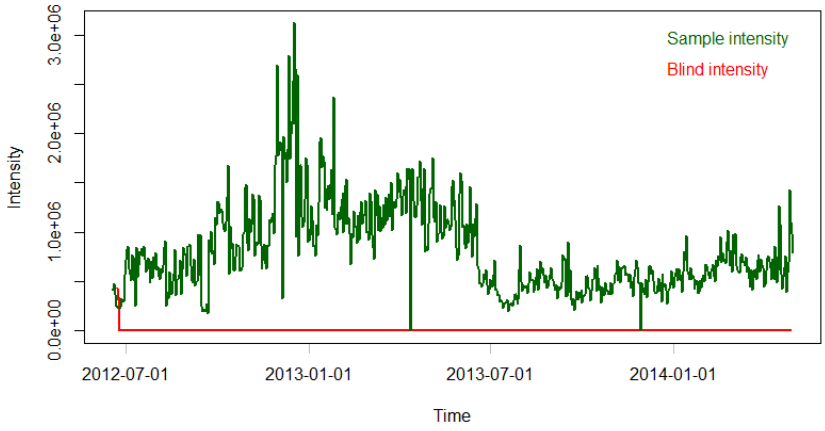


Figure 11. *Intensity time-profile of antidiabetic metformin (green), registered without background contamination (red). Covering a concentration range of 0.03-0.6 µg/L, the yearly load of 2012 in the River Rhine was 13 t.*

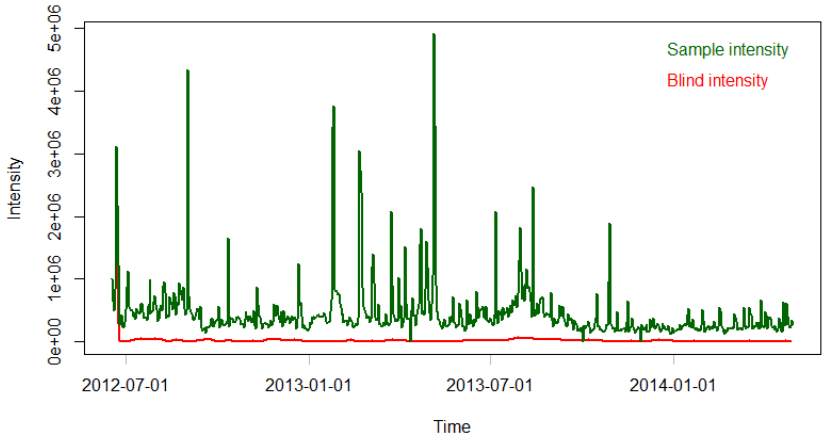


Figure 12. *Intensity time-profile for the insect repellent DEET (diethyltoluamid, green line), with little background contamination (red line).*

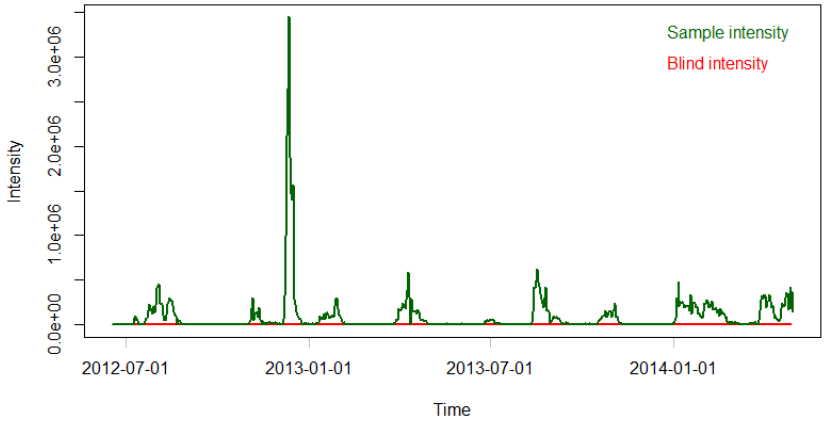


Figure 13. *Intensity time-profile of an unknown compound at m/z = 267.142 Th and RT = 8.5 min. Sample intensities are shown in green and intensities in blind measurements in red.*

pounds. Figures S-1 to S-4 in the Appendix evidence four such IS correlation pairs: peak intensity variations over measurements are either not linearly captured by the correlation of adduct pairs (Figure S-1), correlate only over certain periods (Figure S-2), or correlate only for certain adduct pairs of the same analyte and not for others (Figure S-3 vs. S-4). It may be hypothesized that these relative differences of an analyte in forming a certain adduct can be triggered by slight differences in sample preparation, contaminations or changes in the sampled matrix itself. In addition, correlations often increased with the intensity of the two involved adducts. The observed lack in correlation of adduct pairs does on the other hand not imply that concentration changes cannot be traced by the intensity variation of at least one of these adducts or for analytes with stronger intensity variations than that of the evenly spiked IS set – which is indeed a fundamental prerequisite for LC-HRMS detection of aquatic spills.

5.5 Conclusion

A versatile, fast and interactive workflow for the detection of micropollutant spills from large sets of LC-HRMS measurements is presented, streamlined for the computational speed required in routine monitoring frameworks. When tested with LC-HRMS Orbitrap data at the RÜS monitoring station Basel (CH), the workflow integrated well into the daily analysis routine and also revealed several non-targeted trends of concern which would have remained unnoticed by targeted strategies. Over ten warnings or international alarms could thus be issued in 2014 alone and several industrial point sources subsequently identified and located.

An array of complementary stages to be integrated as new features into future versions of the workflow exists, foremost the quantitation of observed intensities to their riverine concentrations. Another simple addition could then link negative and positive ionization measurements, as supported by the existing workflow architecture of *enviMass*. Besides comprising identification features such as loading of MS/MS scans, restrictions in elemental atom counts for molecular formula fits from the isotopologue grouping or links to external databases, some further analysis might also mine for decisive intensity relations among profiles. For example, temporal profile clustering may hint at pollutants stemming from similar sources, governed by similar temporal dynamics or sets of parents and their TPs. Classification of intensity patterns in profiles of target compounds might in turn help to identify profiles of unknown chemical nature by predicting their anthropogenic origin.

REFERENCES

(1) Loos, M. *enviMass: Utilities to Process Mass Spectrometry (LC-HRMS) Data for Environmental Trend Analysis*; 2015. https://github.com/blosloos/enviMass

(2) Loos, M. *enviPick: Peak picking for high resolution mass spectrometry data*; 2014. https://cran.r-project.org/web/packages/enviPick/index.html

(3) Hug, C.; Ulrich, N.; Schulze, T.; Brack, W.; Krauss, M. *Environmental Pollution* **2014**, *184*, 25–32.

(4) Petrie, B.; Barden, R.; Kasprzyk-Hordern, B. *Water research* **2015**, *72*, 3–27.

(5) Blanchard, P.; Lerch, R. *Environmental science & technology* **2000**, *34* (16), 3315–3322.

(6) Leu, C.; Singer, H.; Stamm, C.; Müller, S. R.; Schwarzenbach, R. P. *Environmental science & technology* **2004**, *38* (14), 3835–3841.

(7) Bletsou, A. A.; Jeon, J.; Hollender, J.; Archontaki, E.; Thomaidis, N. S. *TrAC Trends in Analytical Chemistry* **2015**, *66*, 32–44.

(8) Escher, B. I.; Fenner, K. *Environmental science & technology* **2011**, *45* (9), 3835–3847.

(9) Krauss, M.; Singer, H.; Hollender, J. *Analytical and bioanalytical chemistry* **2010**, *397* (3), 943–951.

(10) Ruff, M.; Singer, H.; Ruppe, S.; Mazacek, J.; Dolf, R.; Leu, C. *Aqua & Gas* **2013**, *93* (5), 16–25.

(11) Fernandez-Albert, F.; Llorach, R.; Andres-Lacueva, C.; Perera-Lluna, A. *Analytical chemistry* **2014**, *86* (5), 2320–2325.

(12) Karaouzas, I.; Lambropoulou, D. A.; Skoulikidis, N. T.; Albanis, T. A. *Journal of Environmental Monitoring* **2011**, *13* (11), 3064–3074.

(13) Müller, A.; Schulz, W.; Ruck, W. K.; Weber, W. H. *Chemosphere* **2011**, *85* (8), 1211–1219.

(14) Terrado, M.; Kuster, M.; Raldúa, D.; Alda, M. L. de; Barceló, D.; Tauler, R. *Analytical and bioanalytical chemistry* **2007**, *387* (4), 1479–1488.

(15) Aiche, S.; Sachsenberg, T.; Kenar, E.; Walzer, M.; Wiswedel, B.; Kristl, T.; Boyles, M.; Duschl, A.; Huber, C. G.; Berthold, M. R.; others. *Proteomics* **2015**, *15* (8), 1443–1447.

(16) Castillo, S.; Gopalacharyulu, P.; Yetukuri, L.; Orešic, M. *Chemometrics and Intelligent Laboratory Systems* **2011**, *108* (1), 23–32.

(17) Katajamaa, M.; Orešic, M. *Journal of chromatography A* **2007**, *1158* (1), 318–328.

(18) Listgarten, J.; Emili, A. *Molecular & Cellular Proteomics* **2005**, *4* (4), 419–434.

(19) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešic, M. *BMC bioinformatics* **2010**, *11* (1), 395.

(20) Kuhl, C.; Tautenhahn, R.; Bo ttcher Christoph; Larson, T. R.; Neumann, S. *Analytical chemistry* **2011**, *84* (1), 283–289.

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

(22) Alonso, A.; Julià, A.; Beltran, A.; Vinaixa, M.; Dıaz, M.; Ibañez, L.; Correig, X.; Marsal, S. *Bioinformatics* **2011**, *27* (9), 1339–1340.

(23) Broeckling, C. D.; Afsar, F.; Neumann, S.; Ben-Hur, A.; Prenni, J. *Analytical chemistry* **2014**, *86* (14), 6812–6817.
(24) Frenzel, T.; Miller, A.; Engel, K.-H. *European Food Research and Technology* **2003**, *216* (4), 335–342.
(25) Loos, M.; Ruff, M.; Singer, H. enviMass version 1.2. target screening software, 2013. http://www.eawag.ch/de/abteilung/uchem/software/envimass/
(26) Baran, R.; Kochi, H.; Saito, N.; Suematsu, M.; Soga, T.; Nishioka, T.; Robert, M.; Tomita, M. *BMC bioinformatics* **2006**, *7* (1), 530.
(27) Edmands, W. M.; Barupal, D. K.; Scalbert, A. *Bioinformatics* **2014**, btu705.
(28) Tengstrand, E.; Lindberg, J.; Åberg, K. M. *Analytical chemistry* **2014**, *86* (7), 3435–3442.
(29) Kaever, A.; Landesfeind, M.; Possienke, M.; Feussner, K.; Feussner, I.; Meinicke, P. *BioMed Research International* **2012**, *2012*.
(30) Kloet, F. M. van der; Hendriks, M.; Hankemeier, T.; Reijmers, T. *Analytica chimica acta* **2013**, *801*, 34–42.
(31) Kenar, E.; Franken, H.; Forcisi, S.; Wörmann, K.; Häring, H.-U.; Lehmann, R.; Schmitt-Kopplin, P.; Zell, A.; Kohlbacher, O. *Molecular & Cellular Proteomics* **2014**, *13* (1), 348–359.
(32) Monroe, M. E.; Toli, N.; Jaitly, N.; Shaw, J. L.; Adkins, J. N.; Smith, R. D. *Bioinformatics* **2007**, *23* (15), 2021–2023.
(33) Mueller, L. N.; Rinner, O.; Schmidt, A.; Letarte, S.; Bodenmiller, B.; Brusniak, M.-Y.; Vitek, O.; Aebersold, R.; Müller, M. *Proteomics* **2007**, *7* (19), 3470–3480.
(34) Strohalm, M.; Kavan, D.; Novak, P.; Volny, M.; Havlicek, V. *Analytical chemistry* **2010**, *82* (11), 4648–4651.
(35) Hansen, M. E.; Smedsgaard, J. *Journal of the American Society for Mass Spectrometry* **2004**, *15* (8), 1173–1180.
(36) Katz, J. E.; Dumlao, D. S.; Clarke, S.; Hau, J. *Journal of the American Society for Mass Spectrometry* **2004**, *15* (4), 580–584.
(37) Wenig, P.; Odermatt, J. *BMC bioinformatics* **2010**, *11* (1), 405.
(38) Team, R. C. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2015.
(39) Petyuk, V. A.; Jaitly, N.; Moore, R. J.; Ding, J.; Metz, T. O.; Tang, K.; Monroe, M. E.; Tolmachev, A. V.; Adkins, J. N.; Belov, M. E.; others. *Analytical chemistry* **2008**, *80* (3), 693–706.
(40) Tautenhahn, R.; Böttcher, C.; Neumann, S. *BMC bioinformatics* **2008**, *9* (1), 504.
(41) Loos, M. *nontarget: Detecting Isotope, Adduct and Homologue Relations in LC-MS Data*; 2015. https://cran.r-project.org/web/packages/nontarget/index.html
(42) Kern, S.; Fenner, K.; Singer, H. P.; Schwarzenbach, R. P.; Hollender, J. *Environmental science & technology* **2009**, *43* (18), 7039–7046.
(43) Team, R. C. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2015.
(44) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24* (21), 2534–2536.
(45) Chang, W.; Cheng, J.; Allaire, J.; Xie, Y.; McPherson, J. *shiny: Web Application Framework for R*; 2015.

(46) Loos, M.; Gerber, C.; Corona, F.; Hollender, J.; Singer, H. *Analytical chemistry* **2015**.

Conclusion & Outlook

The overall goal of this thesis was to devise a data mining strategy for the automatized spill and trend detection of micropollutants in riverine environments, accounting for lengthy temporal LC-HRMS measurement sequences. The strategy was to be implemented and tested at the Rhine monitoring station, Basel (CH). To achieve this primary goal, several subproblems were successfully solved. The summaries for both these subproblems and the main goal are given below, supplemented by related open research questions.

Data simulation. A first task was to improve on existing methods to rapidly calculate the isotopic fine structures of large batches of molecular formulas, as addressed in chapter 2. To this end, a novel transition tree methodology has been introduced. In such a transition tree, isotopologues are calculated from each other via unique single-isotopic modifications. The tree branches can be grown separately and hence very efficiently pruned if organized accordingly. As a result, transition trees surpass the speed, memory and precision constraints of existing methods, as tested with an extensive simulation. The method has been made available in the R package *enviPat*[1] and can be assessed through a webpage, now frequented by the mass spectrometric community.[2–6]

Open research topics

Molecular transition trees	*Apart from division into elemental sub-molecules, transition trees can also index the isotopes and build the isotopologues of a full molecule or several of its elements. Preliminary investigations have herein indicated faster performance for some larger molecules. A yet to be determined metric different from sorting transition indices by the abundances of its*

	isotopes might render molecular transitions competitive for smaller molecules, too.
Dynamic transition trees	*So far, all transition indices in parent nodes of a tree point at the same static look-up table to determine the isotopes to be replaced in transitions to daughter nodes. However, because the one initialized ranking of the indexed table might not be optimal for the transitions of all branches, different tables could be induced at different points in a transition tree. Whether computational improvements in pruning can outweigh the additional costs of dynamic table insertions might be addressed next.*
Faster envelope calculation	*Albeit crucial to derive isotopic fine structures, and to restrict and to categorize isotopologue linkages for the below grouping methods – the sampling of peak shapes at discretized mass intervals to superimpose isotopic fine structures to measurable envelopes has now become the bottleneck in terms of computational speed. A mathematical method that can analytically convert fine structures to centroids would therefore be another important step forward.*

Data grouping. Outlined in chapter 3, a second task for LC-HRMS data mining was to group (a) isotopologue centroids caused by the same adduct ion and (b) adduct ions caused by the same analyte. Taking advantage of the new transition tree methodology, a large-scale and high-resolution simulation of organic compounds from the PubChem[7] database revealed hitherto unobserved characteristics among isotopologue centroids. All characteristics were then discretized and used for a nontargeted isotopologue grouping, at high recall and precision as confirmed with both simulated data and targeted screening matches. When joint with a grouping for a set of main adducts formed in electrospray-ionization, large fractions of measured data could subsequently be reduced to chemical components, in a fully nontargeted manner. The underlying routines have all been integrated into the R package *nontarget*[8] and have contributed to recent publications.[5,6,9,10]

Open research topics

Optimized simulations	*Even when run in parallel sessions, an adequate simulation of the feasible space of linkages between isotopologue pairs can take several days or, for further increasing resolutions, presumably weeks. Herein, a large proportion of the feasible space is densely covered, whereas other subspaces receive sparse coverage. A simultaneous learning algorithm to distinguish if yet-to-be simulated molecular formulas fall either into already well-defined linkage spaces or into poorly covered regions may help decrease the simulation costs.*
Optimized data representations and queries	*In the proposed discretization method, recursive partitioning of the linkage space terminates when a maximum size is reached for a bounding box. Earlier terminations directed by statistical moments of linkage distributions rather than fixed size thresholds may however lead to lower numbers of bounding boxes and thus faster queries. Similarly, alternative splitting heuristics may be tested, e.g., density-based methods or non-perpendicular split planes.*
Peak shapes	*The exact peak shapes that make up the measured envelopes of profile mode acquisitions and the subsequent centroids differ among MS instruments.*[11] *Here, Gaussian instead of more appropriate but heavy-tailing Cauchy-Lorentzian peak shapes were used to model Orbitrap data, for the sake of performance. In how far this choice impacts the precision of the simulated linkages should be clarified next.*
True negative linkages	*When using the discretized data model to classify isotopologue linkages, systematic distributions were discovered for the rejected cases. At the validated levels of precision, these rejected cluster cannot be accounted for by false negative findings alone. To further improve on isotopologue grouping, the origin of these systematic true nega-*

tive linkages should be investigated.

Series recognition. Part of the same package is a first algorithm for the direct unsupervised detection of distinctive homologue series (HS) patterns in LC-HRMS data, i.e., sets of signal peaks with constant shifts in mass and smooth drifts in retention time. Despite the combinatorial complexity of this task, a dynamic programming approached coupled to specialized metric structures for data representation could guarantee a fast computation. When applied to STP effluents, a multitude of ubiquitous systematic series signals have thus been detected, largely surpassing the number of manually assorted series from a previous targeted analysis.[6] Together with an in-depth examination of the ambiguities for assigning peaks to different signal series, this algorithm was topic of chapter 4.

Open research topics

Series clustering	*Series have been observed which are not directly interrelated by sharing peaks but which nevertheless have identical mass shifts and similar retention time patterns. A clustering of such series and their subsequent identification shall elucidate if this observation is random or caused by, e.g., the different transformation products of the same parent compounds.*
Gap tolerance	*With the proposed recognition method, a gap-less series of homologues must be measurable to be detected. For LC-HRMS, some gapped series have been reported which otherwise fulfill the minimum series length criterion.[6] For such series, a gap-tolerant version of the presented algorithm is a prerequisite to become finally recognizable.*
Comparison of STPs	*The very comparison of cumulative frequencies between sewage treatment plants has revealed the ubiquitous occurrence of certain series mass differences in Swiss STP efflu-*

ents which have not yet been targeted. Additional congruency of the concerned series in (a) retention time shifts, (b) retention time, (c) exact homologue masses, (d) series length and (d) series interrelations should greatly refine this finding to prioritize future identification efforts on widespread nontargeted pollutants.

Usability — *The frequency at which series are detected in both effluent and Rhine samples as well as the interrelations among series can be overwhelming for users. A streamlined or simplified strategy complemented by a user interface to cope with this information may be a crucial step forward to increase the usability of the presented algorithm for the greater public.*

Spill detection. Finally, chapter 5 approaches the primary goal of riverine trend detection, aided by the above achievements. Embracing chemical componentization, HS pattern recognition, and data simulation, an analysis pipeline for swift extracting of chromatograms, peaks, time-intensity profiles and finally trends is presented. During application at the Rhine monitoring station, numerous spills and trends of concern could be detected from more than a billion data points, in a fully automatized manner and under practical time constraints. Applicable to targeted as well as nontargeted analysis, international alarms could be issued and several responsible emission sources exposed; a range of previously untargeted compounds could hereby be prioritized for identification. The functions for the novel chromatogram clustering and peak picking are embedded in the package *enviPick*.[12] The subsequent workflow functionalities, including preprocessing and a user interface, are to be publicly found in the R *enviMass* software package.[13]

Open research topics

m/z estimates	*At the final stage of the enviMass workflow, improved mass estimates can be selectively derived for individual profiles. Future implementations may consider to derive these estimates en masse for all profiles and at earlier steps, whereas the grouping of isotopologues and adducts and homologue series detection might be postponed to later stages and not run on a sample-wise basis. Taking advantage of the improved mass estimates, and given that all developed tools can indeed embrace nonconstant measurement uncertainties, the precision of the grouping and series detection might be further enhanced.*
Profile extraction	*To ensure fast processing, profiles are so far assembled within peak partitions via an intensity descend, a shrinking mass tolerance and a fixed retention time window. In contrast, future fine-tuning may incorporate, e.g., the similarity of chromatograms of measurements adjacent in a sequence or the structural information in a partition to selectively adjust retention time windows, if not compromising performance.*
Profile clustering	*enviMass has been built with focus on swift spill and rising trend detection. Based on the multitude of observed profiles, and their distinct patterns and temporal emergences, time series clustering will likely reveal interesting relationships between profile components. Anticipated relations are those between parent compounds and their transformation products or between different production intermediates of the same industrial synthesis pathway.*
Hydrology and source apportionment	*Monitoring stations such as the RÜS often intercept the discharge of different riverine tributaries. Attempts to explain variations in time-intensity profiles with differing discharge compositions and volumes may be helpful to apportion sources or to correct nontarget profiles for*

varying dilution of the underlying concentrations.

Usability — *On the one hand, the enviMass user interface is meant to summarize and prioritize findings for the chemical analyst. On the other hand, and in single cases, more information on the responsible calculations and automated decisions is needed for refinement. The integration of both is still challenging, especially if user are not able or willing to revert to command-line levels. Hence, (a) a powerful visualization engine to overlay results with selected raw data, peak partitions or supplementary data and (b) incorporation of additional analysis such as measurement replicates, MS/MS spectra, etc., is envisaged.*

Spatiotemporal analysis — *Though concentrating on temporal trends, intensity profiles can generally also be analyzed with enviMass for spatial transects or migrating waves of Rhine water. Given that several monitoring stations exist in the discussed Rhine River network, combination of LC-MS measurement data not only at one but several spatial locations over time would be a promising task.*

To conclude, high resolution LC-MS data analysis is a complex and computationally demanding task, in particular for the small molecule universe encountered in long-term monitoring programs. To this end, data reduction, data simulation and automatized trend analysis have been tailored into a software suite, dedicated to the monitoring of micropollutant dynamics in aquatic systems.

References

(1) Loos, M. *enviPat: Isotope Pattern, Profile and Centroid Calculation for Mass Spectrometry*; 2015. https://cran.r-project.org/web/packages/enviPat/index.html

(2) Chiaia-Hernandez, A. C.; Schymanski, E. L.; Kumar, P.; Singer, H. P.; Hollender, J. *Analytical and bioanalytical chemistry* **2014**, *406* (28), 7323–7335.

(3) Chu, D. B.; Troyer, C.; Mairinger, T.; Ortmayr, K.; Neubauer, S.; Koellensperger, G.; Hann, S. *Analytical and bioanalytical chemistry* **2015**, *407* (10), 2865–2875.

(4) Graf, M. M.; Sucharitakul, J.; Bren, U.; Chu, D. B.; Koellensperger, G.; Hann, S.; Furtmüller, P. G.; Obinger, C.; Peterbauer, C. K.; Oostenbrink, C.; others. *FEBS Journal* **2015**.

(5) Ruff, M.; Mueller, M. S.; Loos, M.; Singer, H. P. *Water Research* **2015**, *87*, 145–154.

(6) Schymanski, E. L.; Singer, H. P.; Longree, P.; Loos, M.; Ruff, M.; Stravs, M. A.; Vidal, C. Ripolle s; Hollender, J. *Environmental science & technology* **2014**, *48* (3), 1811–1818.

(7) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. *Annual reports in computational chemistry* **2008**, *4*, 217–241.

(8) Loos, M. *nontarget: Detecting Isotope, Adduct and Homologue Relations in LC-MS Data*; 2015. https://cran.r-project.org/web/packages/nontarget/index.html

(9) Hug, C.; Ulrich, N.; Schulze, T.; Brack, W.; Krauss, M. *Environmental Pollution* **2014**, *184*, 25–32.

(10) Schollée, J.; Schymanski, E.; Avak, S.; Loos, M.; Hollender, J. *submitted to Analytical Chemistry* **2015**.

(11) Urban, J.; Afseth, N. K.; Štys, D. *TrAC Trends in Analytical Chemistry* **2014**, *53*, 126–136.

(12) Loos, M. *R package enviPick: Peak picking for high resolution mass spectrometry data*; 2014. https://cran.r-project.org/web/packages/enviPick/index.html

(13) Loos, M. *enviMass: Utilities to Process Mass Spectrometry (LC-HRMS) Data for Environmental Trend Analysis*; 2014. https://github.com/blosloos/enviMass

Appendix

SUPPORTING INFORMATION FOR CHAPTER 3

Table S-1. *List of adduct classes used for centroid linkage simulations. Individual adduct classes are formed by multiplying the molecular formula of a compound with x, adding the formula y and division by charge z.*

ID	Formula *y*	*z*	Multipl. *x*
1	-	*1*	*1*
2	-	*2*	*1*
3	-	*3*	*1*
4	-	*1*	*2*
5	-	*1*	*3*
6	Na_2	*2*	*1*
7	Na_2	*3*	*1*
8	N_1	*1*	*1*
9	N_1	*2*	*1*
10	N_1	*1*	*2*
11	K_1	*1*	*1*
12	K_1	*2*	*1*
13	K_1	*1*	*2*
14	K_2	*1*	*1*
15	C_2O_2	*1*	*1*
16	C_2O_2	*1*	*2*
17	C_2N_1	*2*	*1*
18	C_2N_1	*1*	*1*
19	$C_2N_1Na_1$	*1*	*2*
20	C_4N_2	*2*	*1*
21	C_4N_2	*1*	*1*
22	C_6H_3	*2*	*1*
23	$C_3O_1Na_1$	*1*	*1*
24	$C_2S_1O_1$	*1*	*1*
25	O_6	*2*	*2*
26	Cl_1	*1*	*1*
27	Br_1	*1*	*1*

Table S-2. *Parameters used for function screening() from the R enviMass package.*

Parameter	Value
blanklist	*FALSE*
dmz	*2*
ppm	*TRUE*
dRT	*30 [seconds]*[1]
dRTwithin	*4 [seconds]*
dRTblank	*FALSE*
dInt	*0.2*
w1	*0.5*
w2	*0.5*
w3	*0*

[1] *Set to 120 s for one particular target with less reproducible RT behavior.*

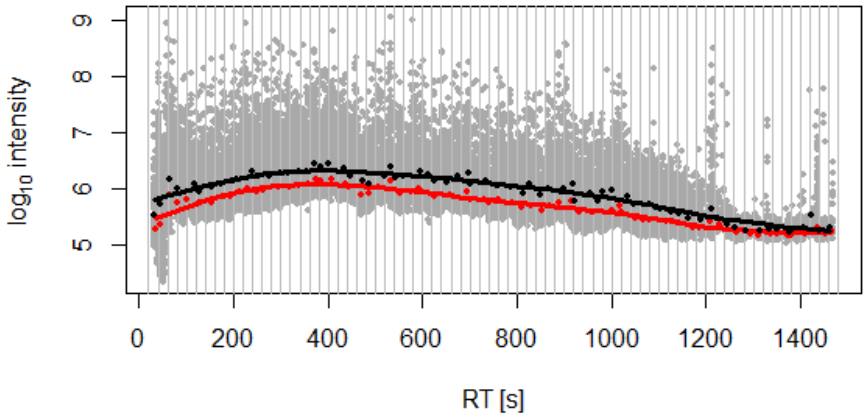


Figure S-1 (last page): *Signal peaks picked from a LC-HRMS measurement of a STP sample (gray dots), plotted over their chromatographic elution time. Gray lines separate peak bins in which the lower quartile (red dots) and the median (black dots) were extracted, to be smoothed by spline models (red and black lines, respectively); the model prediction for the lower quartile was then used as intensity threshold in the screening procedure.*

Tables S-3. *Parameters used for peak picking with the R enviPick package, functions mzagglom(), mzclust() and mzpick(), respectively. See package manual for detailed parameter descriptions.*

R function mzagglom()	
Parameter	**Value**
dmzgap	*>3.5*
ppm	*TRUE*
drtgap	*300 [seconds]*
minpeak	*4*
maxint	*1E7*

R function mzclust()	
Parameter	**Value**
dmzdens	*3.5*
ppm	*TRUE*
drtdens	*60 [seconds]*
minpeak	*4*
maxint	*1E7*

R function mzpick()	
Parameter	**Value**
minpeak	*4*
drtsmall	*20 [seconds]*
drtfill	*10 [seconds]*
drttotal	*120 [seconds]*
recurs	*2*
weight	*1*
SB	*4*
SN	*5*
minint	*1E4*
maxint	*1E7*
ended	*1*

Tables S-4. *Parameters used for isotopologue and adduct grouping and the full combination (i.e., componentization) of both groups with functions pattern.search2(), adduct.search() and combine() from the R package nontarget, respectively.*

R function pattern.search2()		
Parameter	**Article symbol**	**Value**
quantiz	*set of all* $B_l^{T,z}$	*List object from R package nontargetData*
mztol	ε_3	*2 ppm*
ppm		*TRUE*
inttol	ΔInt	*0.2*
rttol	ΔRT_{max}	*4 [seconds]*
use_isotopes		*FALSE*
use_charges		*c(1,2,3)*
use_marker		*TRUE/FALSE*
quick		*FALSE*

R function adduct.search()	
Parameter	**Value**
adducts	*default*
rttol	*4 [seconds]*
mztol	*2 ppm*
ppm	*TRUE*
use_adducts	*c("M+H","M+Na","M+NH4","M+K")*
ion_mode	*positive*

R function combine()	
Parameter	**Value**
rules	*c(FALSE,FALSE,FALSE)*
dont	*FALSE*

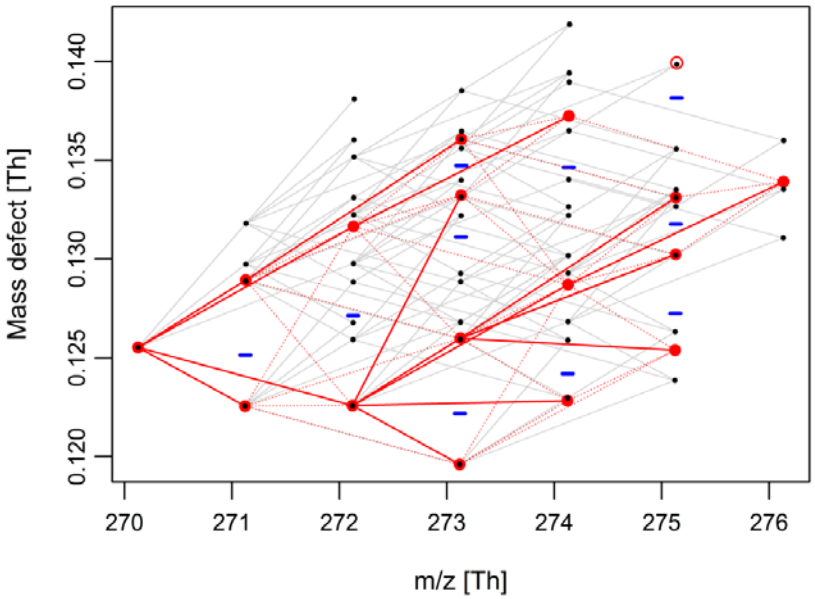


Figure S-2: *Pruned isotope pattern (black dots), resulting centroids (red solid dots) and envelope minima (blue bars) at R = 1.2x10^5 for the first adduct class of the herbicide Alachlor, $C_{14}H_{20}Cl_1N_1O_2^+$. Void red dots show pruned centroids smaller than intensity threshold β_2, calculated as a result of $\beta_2>\beta_1$. Moreover, gray lines indicate all isotopologue transitions, red dashed ones all centroid linkages and red solid lines those assigned. Note that the mass defect is solely used to distinguish data points of equal nominal mass on the ordinate.*

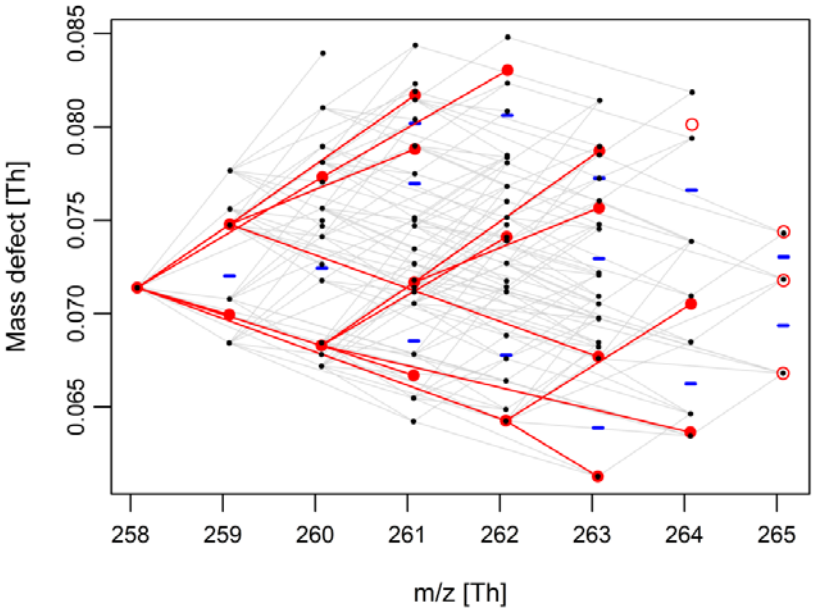


***Figure S-3**: Pruned isotope pattern (black dots), resulting centroids (red solid dots) and envelope minima (blue bars) at R = 1.2x10^5 for the first adduct class of Orbencarb, $C_{12}H_{16}Cl_1N_1O_1S_1^+$. Void red dots show pruned centroids smaller than β_2. Moreover, gray lines indicate all isotopologue transitions, red ones the assigned centroid linkages.*

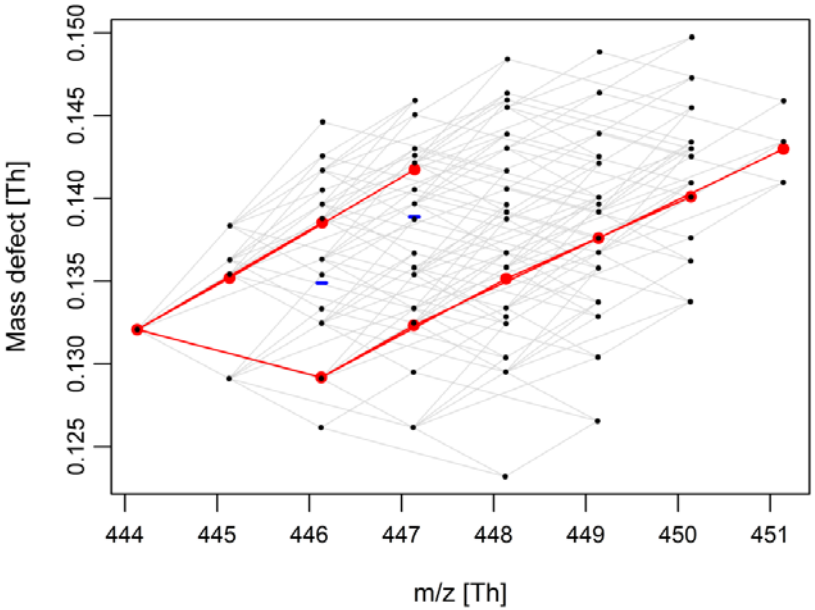


Figure S-4: *Pruned isotope pattern (black dots), resulting centroids (red solid dots) and envelope minima (blue bars) at R = 1.2x10^5 for the first adduct class of Propaquizafop, $C_{22}H_{22}Cl_1N_3O_5^+$. Moreover, gray lines indicate all isotopologue transitions, red ones the assigned centroid linkages.*

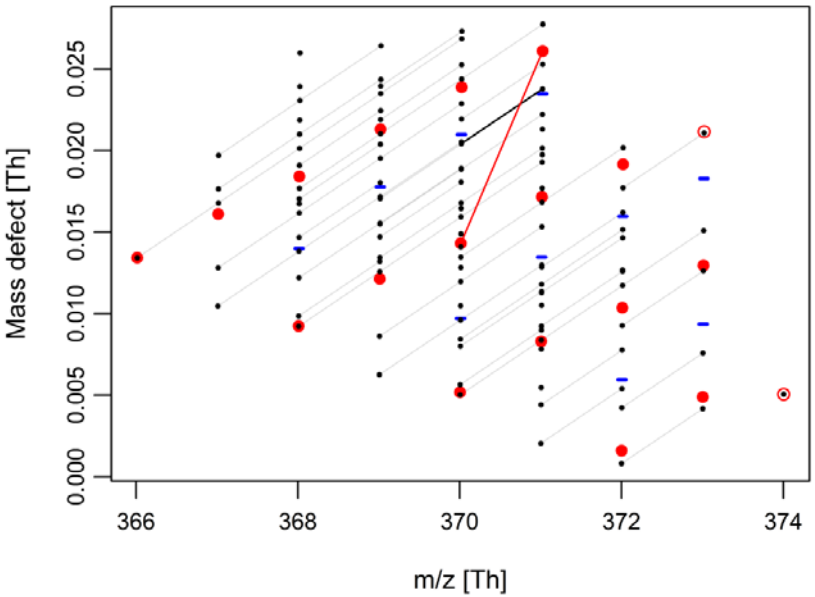


Figure S-5: *Pruned isotope pattern (black dots), resulting centroids (red solid and void dots) and envelope minima (blue bars) at R = 1.0x10^5 for the first adduct class of $C_{12}H_{16}O_6N_1S_3^+$. Gray lines exclusively indicate ^{12}C to ^{13}C isotopologue transitions, with the black one responsible for categorizing the single centroid linkage shown as red line.*

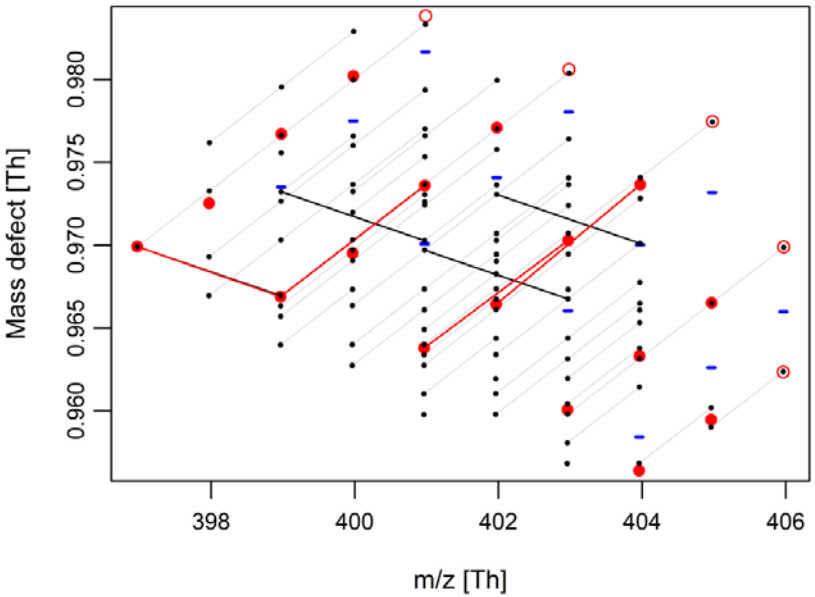


Figure S-6: *Pruned isotope pattern (black dots), resulting centroids (red solid and void dots) and envelope minima (blue bars) at R = 9.8x10^4 for the first adduct class of a fungicide, $C_{13}H_6Cl_2F_4N_4S_1^+$ (InChI-key WLONTJKIWGLQRV-UHFFFAOYSA-N). Gray lines exclusively indicate ^{35}Cl to ^{37}Cl isotopologue transitions, with the black ones responsible for categorizing the red linkages of centroids categorized by such.*

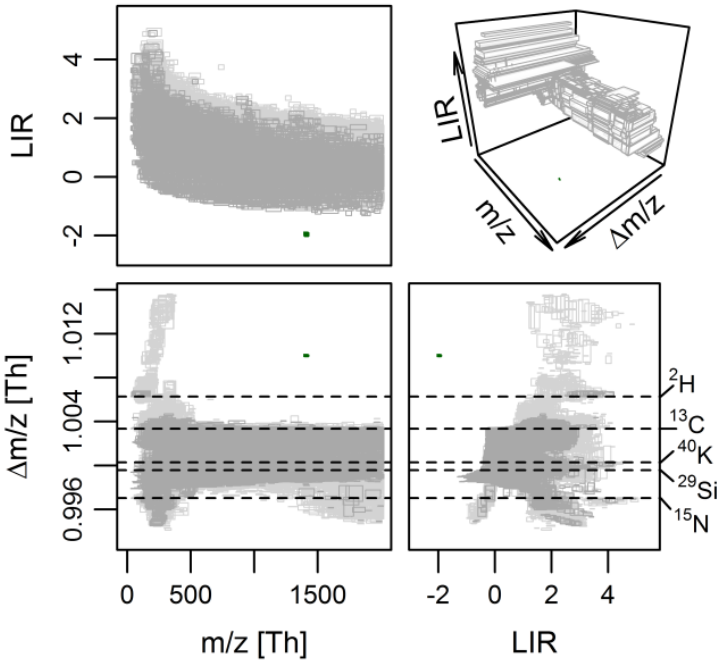


Figure S-7: *Bounding rectangles of simulated centroid linkages at Δm/z ≈ 1.0 for R=70K. While bounding rectangles in light gray contain all linkages, those in dark gray frame only those to centroids which differ by one isotope replacement from being monoisotopic. The small green rectangle shows the size of the nearest neighbor extension. For the 3D visualization, a much coarser discretization was utilized.*

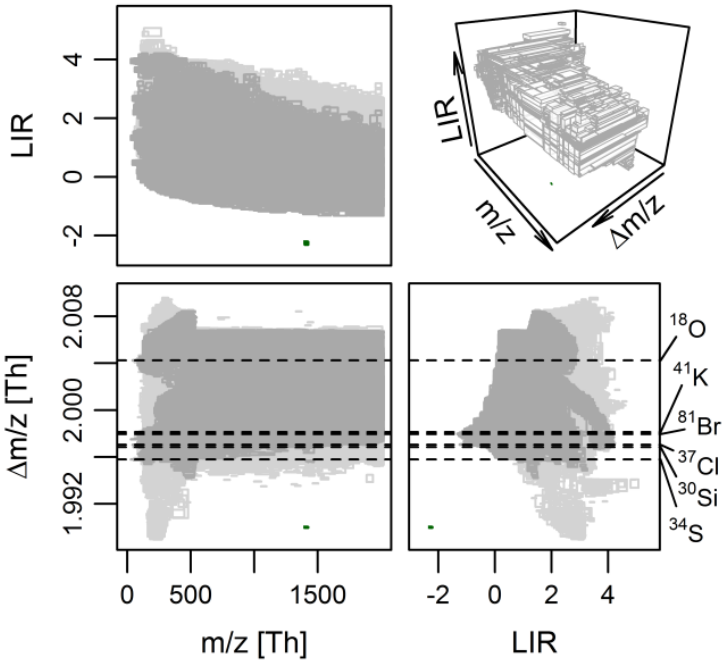


Figure S-8: *Discretized distribution of simulated centroid linkages at Δm/z ≈ 2.0 for R=70K. Check caption of Figure S-6 for further explanations.*

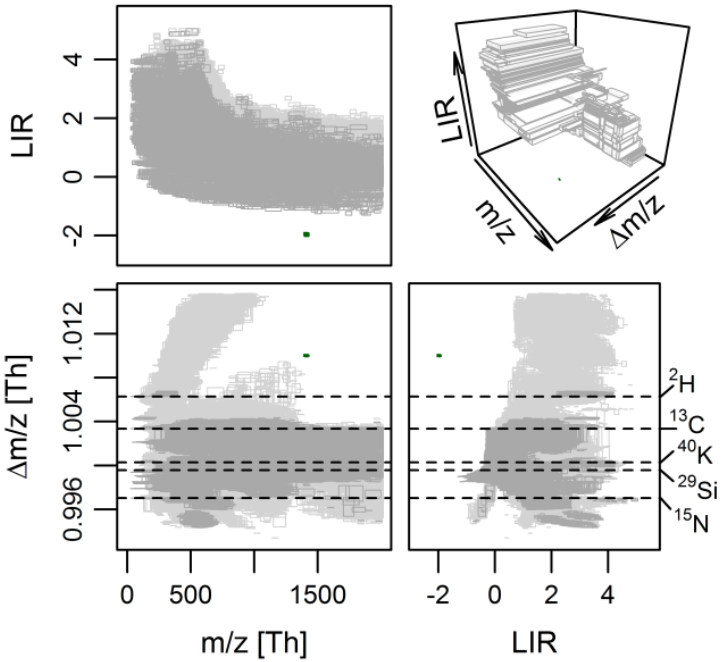


Figure S-9: *Discretized distribution of simulated centroid linkages at Δm/z ≈ 1.0 for R=280K. Check caption of Figure S-6 for further explanations.*

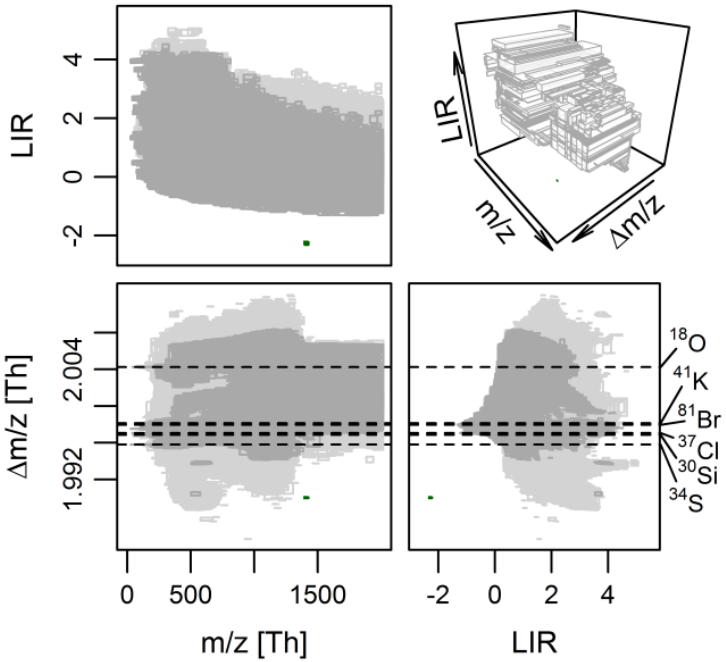


Figure S-10: *Discretized distribution of simulated centroid linkages at Δm/z ≈ 2.0 for R=280K. Check caption of Figure S-6 for further explanations.*

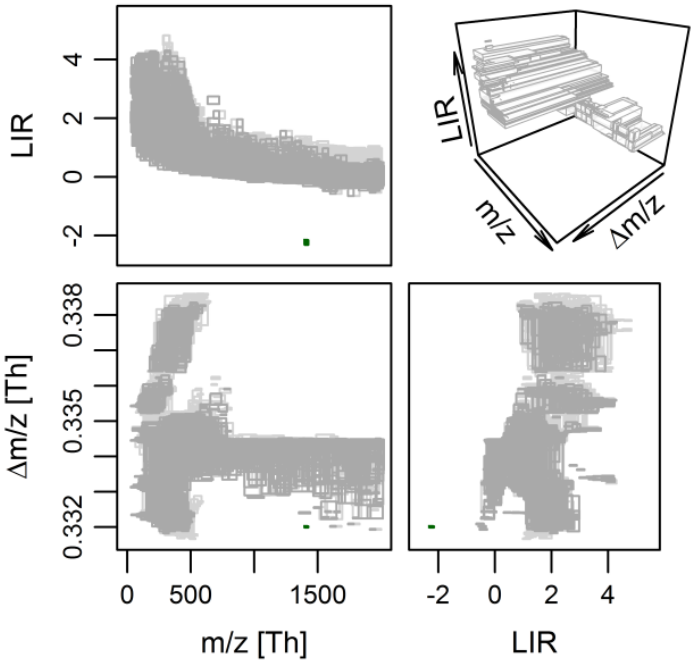


Figure S-11: *Discretized distribution of simulated centroid linkages at Δm/z ≈ 0.3 for R=140K. Check caption of Figure S-6 for further details.*

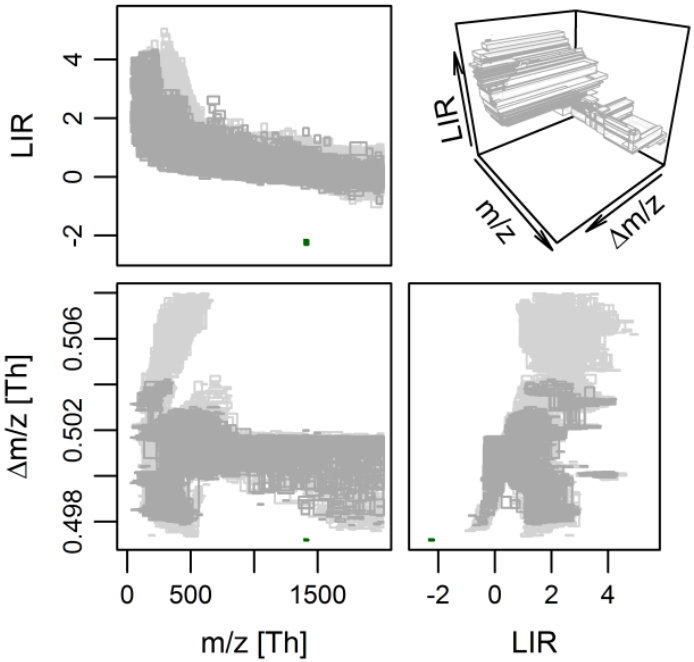


Figure S-12: *Discretized distribution of simulated centroid linkages at Δm/z ≈ 0.5 for R=140K. Check caption of Figure S-6 for further explanations.*

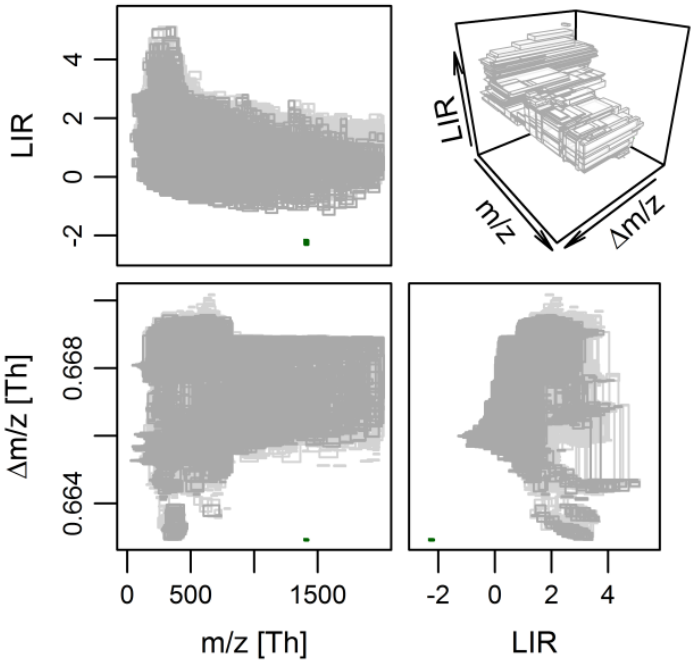


Figure S-13: *Discretized distribution of simulated centroid linkages at Δm/z ≈ 0.7 for R=140K. Check caption of Figure S-6 for further explanations.*

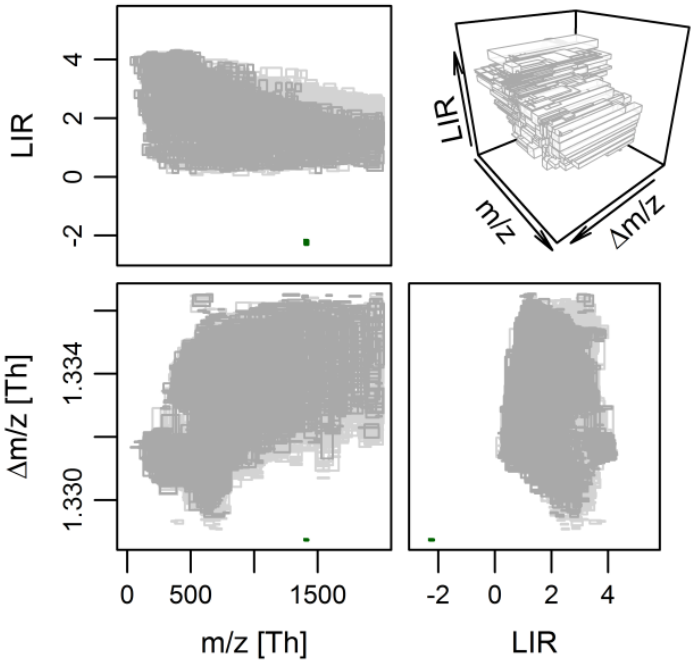


Figure S-14: *Discretized distribution of simulated centroid linkages at Δm/z ≈ 1.3 for R=140K. Check caption of Figure S-6 for further explanations.*

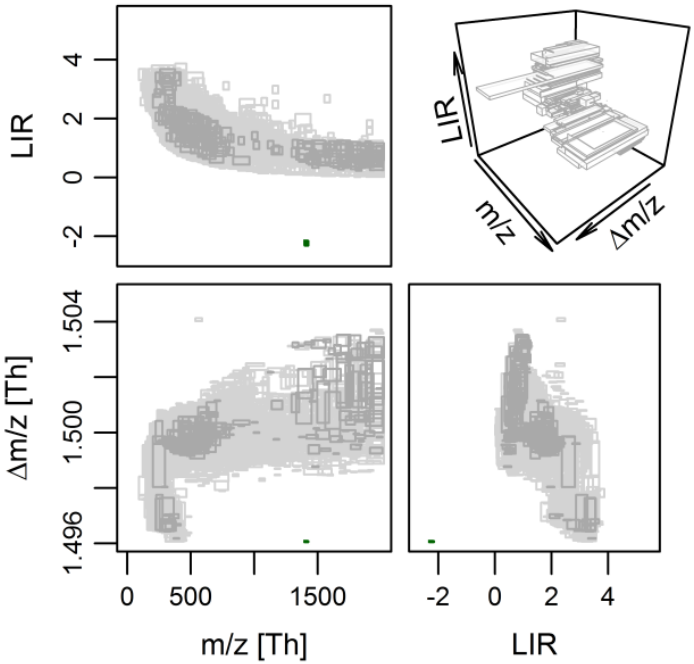


Figure S-15: *Discretized distribution of simulated centroid linkages at Δm/z ≈ 1.5 for R=140K. Compare caption of Figure S-6 for further explanations.*

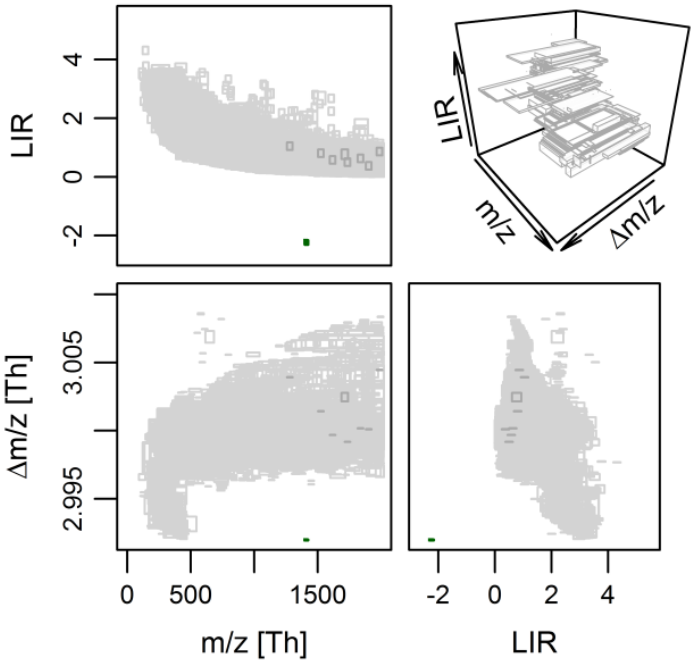


Figure S-16: *Discretized distribution of simulated centroid linkages at Δm/z ≈ 3.0 for R=140K. At this Δm/z range, transitions to form linkages mostly stem from replacing ^{33}S by ^{36}S. Check caption of Figure S-6 for further explanations.*

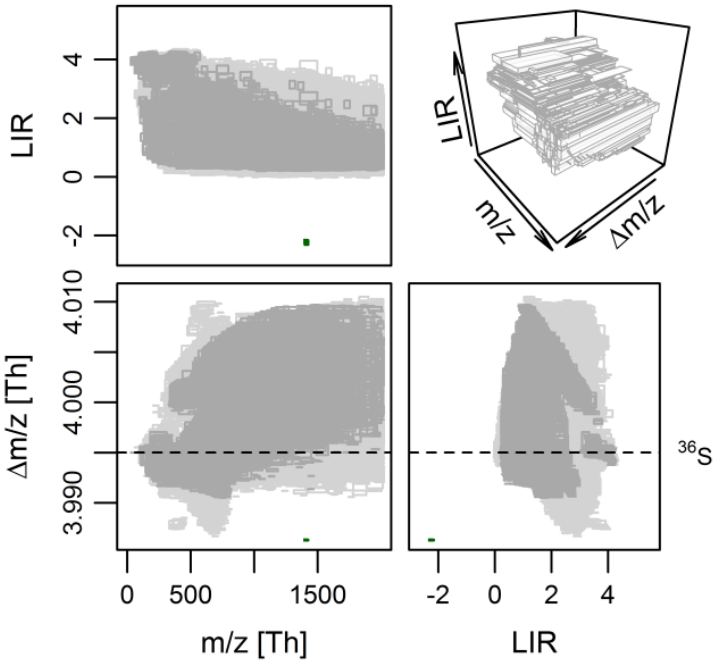


Figure S-17: *Discretized distribution of simulated centroid linkages at Δm/z ≈ 4.0 for R=140K. The dashed line indicates the theoretical Δm/z values of a isotope transitions with ^{36}S replacing the lowest-mass isotope of this element, i.e., ^{32}S. Check caption of Figure S-6 for further explanations.*

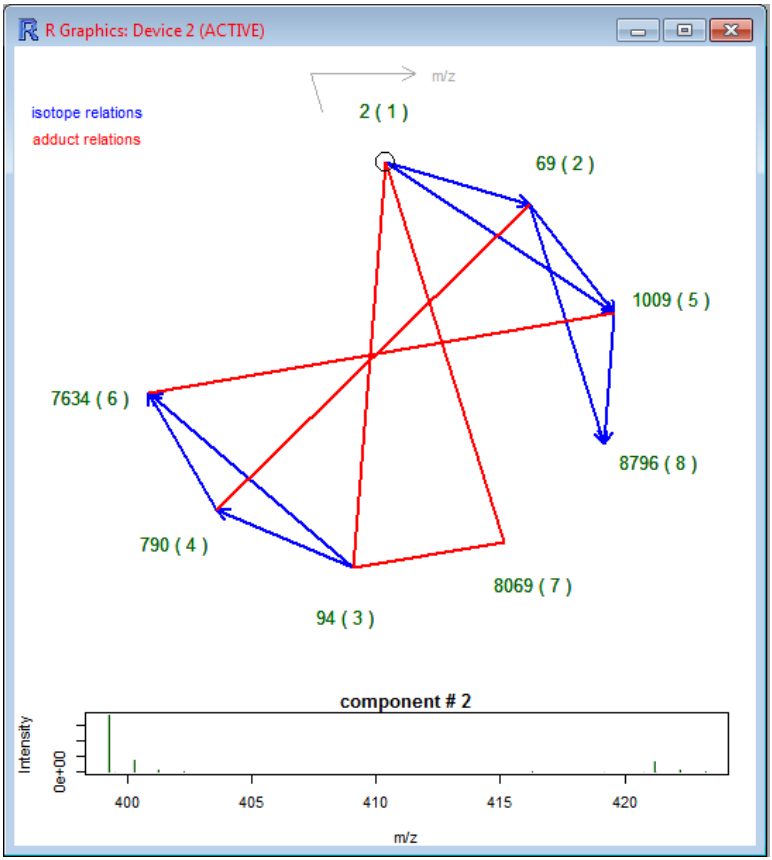


Figure S-18. *Graphical display of a component group from the R nontarget package. The top circular plot links centroids for their adduct relations (red lines) and isotopologue linkages (blue lines); numbers refer to centroid IDs, whereas numbers in brackets to the intensity rank of a centroid in that component group. Here, H- (IDs 2, 69, 1009, 8796), Na- (IDs 94, 790, 7634) and NH_4-adducts (ID 8069) are annotated to different ion species of one unknown compound. The lower inset panel displays the m/z values and intensities of the corresponding centroids.*

Table S-5. *Results for the peak componentization of ten Swiss STP effluent measurements under positive mode ESI and the validation of its isotopologue grouping.*

ID	Location	Peaks	Isotopol. groups[1]	Adduct groups[1]	Compo-nents[1]
1	Affoltern, Zwillikon	*20788*	*4025 (0.507)*	*2473 (0.288)*	*11698 (0.610)*
2	Winterthur	*18024*	*3415 (0.489)*	*2116 (0.287)*	*10409 (0.594)*
3	Werdhölzli, Zürich	*19614*	*3895 (0.509)*	*2340 (0.290)*	*11161 (0.605)*
4	Thal, Altenrhein	*21828*	*4355 (0.518)*	*2457 (0.271)*	*12353 (0.613)*
5	Uetendorf, Thun	*21105*	*4047 (0.493)*	*2471 (0.286)*	*12181 (0.593)*
6	Bioggio, Lugano	*18416*	*3607 (0.493)*	*2313 (0.306)*	*10532 (0.601)*
7	Verniere, Aïre	*23336*	*4661 (0.519)*	*2618 (0.272)*	*13207 (0.615)*
8	Bussigny-prés-Laus.	*26409*	*5731 (0.569)*	*3520 (0.331)*	*13659 (0.660)*
9	Hallau, Klettgau	*24958*	*5123 (0.536)*	*2892 (0.280)*	*13809 (0.629)*
10	Schönau, Zug	*17060*	*3272 (0.482)*	*1961 (0.277)*	*10050 (0.585)*

[1] *Values outside brackets state counts of groups/components, whereas those in brackets refer to fractions of peaks contained therein.*

Table S-5 *(continued)*

ID	Location	Validation with target compounds					
		Matches[2]	***TP***[3]	***FN***	***FP***[4]	***Recall***	***Precision***[4]
1	Affoltern, Zwillikon	*36 / 35*	*58 / 29*	*0*	*6 / 7*	*1.00*	*0.90 / .90*
2	Winterthur	*40 / 35*	*75 / 32*	*0*	*13 / 16*	*1.00*	*0.85 / 0.82*
3	Werdhölzli, Zürich	*34 / 29*	*59 / 30*	*0*	*7 / 10*	*1.00*	*0.89 / 0.86*
4	Thal, Altenrhein	*42 / 42*	*87 / 28*	*0*	*9 / 13*	*1.00*	*0.90 / 0.87*
5	Uetendorf, Thun	*43 / 42*	*85 / 33*	*0*	*8 / 9*	*1.00*	*0.91 / 0.90*
6	Bioggio, Lugano	*42 / 41*	*80 / 31*	*0*	*9 / 13*	*1.00*	*0.89 / 0.86*
7	Verniere, Aïre	*43 / 42*	*79 / 34*	*0*	*10 / 13*	*1.00*	*0.89 / 0.86*
8	Bussigny-prés-Laus.	*42 / 41*	*68 /16*	*0*	*9 / 9*	*1.00*	*0.88 /0.88*
9	Hallau, Klettgau	*35 / 32*	*57 / 27*	*0*	*5 / 6*	*1.00*	*0.92 / 0.90*
10	Schönau, Zug	*37 / 34*	*64 / 30*	*0*	*3 / 4*	*1.00*	*0.96 / 0.94*

[2] The first number states the count of compounds for which at least one full match between all expected theoretical and measured centroids of any adduct could be obtained. The second number after the backslash counts the same, but with restriction to the four main adducts [M+H⁺], [M+Na⁺], [M+NH₄⁺] and [M+K⁺].

[3] The first number states the screening count of ion species (i.e., combination of a compound and its specific adduct) for which a full match between all expected theoretical and measured centroids were obtained; this count was used for calculation of the precision. The second number after the backslash counts the same, but restricted to matches with at least two expected centroids; used for recall calculations.

[4] The first value is derived with usage of marker centroids, the second without.

Table S-5 *(continued)*

ID	Location	Validation with labeled standard compounds.					
		Matches	***TP***[3]	***FN***	***FP***[4]	***Recall***	***Precision***[4]
1	Affoltern, Zwillikon	*77*	*123 / 51*	*0*	*14 / 15*	*1.00*	*0.90 / 0.89*
2	Winterthur	*84*	*137 / 56*	*0*	*22 / 26*	*1.00*	*0.86 / 0.84*
3	Werdhölzli, Zürich	*79*	*122 / 51*	*0*	*11 / 14*	*1.00*	*0.92 / 0.990*
4	Thal, Altenrhein	*93*	*144 / 46*	*0*	*17 / 23*	*1.00*	*0.89 / 0.86*
5	Uetendorf, Thun	*86*	*138 / 46*	*0*	*17 / 22*	*1.00*	*0.89 / 0.86*
6	Bioggio, Lugano	*85*	*154 / 48*	*0*	*28 / 35*	*1.00*	*0.85 / 0.81*
7	Verniere, Aïre	*89*	*150 / 45*	*0*	*24 / 29*	*1.00*	*0.86 / 0.84*
8	Bussigny-prés-Laus.	*91*	*131 / 25*	*0*	*24 /28*	*1.00*	*0.85 / 0.82*
9	Hallau, Klettgau	*84*	*114 / 53*	*0*	*11 / 14*	*1.00*	*0.91 / 0.90*
10	Schönau, Zug	*78*	*119 / 49*	*0*	*12 / 16*	*1.00*	*0.91 / 0.88*

[3] The first number states the screening count of ion species (i.e., combination of a compound and its specific adduct) for which a full match between all expected theoretical and measured centroids were obtained; this count was used for calculation of the precision. The second number after the backslash counts the same, but restricted to matches with at least two expected centroids; used for recall calculations.

[4] The first value is derived with usage of marker centroids, the second without.

SUPPORTING INFORMATION FOR CHAPTER 4

Table S-1. *Parameters used for SOM training with the kohonen package.*

Parameter	Article symbol	Value
-	$\hat{c}_{\Delta RT}$	*0.1*
-	$\hat{c}_{\Delta m/z}$	*0.001*
xdim	-	*60*
ydim	-	*40*
rlen	-	*10.000*
alpha	α	*$5x10^{-3}$ to $5x10^{-6}$*
radius	-	*45*
init	-	*- (default)*
toroidal	-	*TRUE*
n.hood	-	*square*

Table S-2. *Parameters used for peak picking with the R enviPick package, functions mzagglom(), mzclust() and mzpick(), respectively. See package manual for detailed parameter descriptions.*

Parameter	Value
dmzgap	*>3.5*
ppm	*TRUE*
drtgap	*300 [seconds]*
minpeak	*4*
maxint	*1E7*

Parameter	Value
dmzdens	*3.5*
ppm	*TRUE*
drtdens	*60 [seconds]*
minpeak	*4*
maxint	*1E7*

Parameter	Value
minpeak	*4*
drtsmall	*20 [seconds]*
drtfill	*10 [seconds]*
drttotal	*120 [seconds]*
recurs	*2*
weight	*1*
SB	*4*
SN	*5*
minint	*1E4*
maxint	*1E7*
ended	*1*

Table S-3. *Series detection characteristics for the 10 STP effluent samples. The second-last column only counts the series which were paired with at least one other series, whereas the last column states all unique - and possibly multiple per series – pairings below the given intersection angle θ. Numbers in brackets state percentages of all observed peaks, series or series pairs.*

ID	Location	Ionization mode	Peaks	Peaks in series	Series
-	(Blind sample)	positive negative	*12843* *6768*	*1621 (12.6)* *452 (6.7)*	*573* *155*
1	Affoltern, Zwillikon	positive negative	*20788* *9533*	*7342 (35.5)* *1099 (11.5)*	*6937* *822*
2	Winterthur	positive negative	*18024* *9748*	*5173 (28.7)* *1054 (10.8)*	*4553* *1228*
3	Werdhölzli, Zürich	positive negative	*19614* *10331*	*7135 (36.4)* *1324 (12.8)*	*6692* *911*
4	Thal, Altenrhein	positive negative	*21828* *9721*	*7425 (34.0)* *1247 (12.8)*	*5641* *711*
5	Uetendorf, Thun	positive negative	*21105* *10483*	*7740 (36.7)* *1236 (11.8)*	*7806* *715*
6	Bioggio, Lugano	positive negative	*18416* *10491*	*5909 (32.1)* *1354 (12.9)*	*4871* *813*
7	Verniere, Aïre	positive negative	*23336* *11120*	*8960 (38.4)* *1419 (12.8)*	*7666* *741*
8	Bussigny-prés-Lausanne	positive negative	*26409* *12116*	*14954 (56.6)* *2401 (19.8)*	*18254* *2040*
9	Hallau, Klettgau	positive negative	*24958* *11064*	*11686 (46.8)* *2019 (18.2)*	*10090* *1706*
10	Schönau, Zug	positive negative	*17060* *9582*	*4910 (28.8)* *864 (9.0)*	*3254* *488*

Table S-3 (continued).

ID	Series, ≠ blank	Series, monoisotopic	Series, paired	Series pairs θ<0.08π
-	- -	*540 (94.2)* *143 (92.3)*	*514 (89.7)* *137 (88.4)*	*960 (54.4)* *263 (85.4)*
1	*6089 (87.8)* *749 (91.1)*	*4713 (67.9)* *714 (86.9)*	*6849 (98.7)* *792 (96.4)*	*163560 (80.6)* *23744 (88.4)*
2	*3777 (83.0)* *1129 (91.9)*	*2625 (57.7)* *1174 (95.6)*	*4481 (98.4)* *1198 (97.6)*	*224561 (93.7)* *242155 (98.7)*
3	*6263 (93.6)* *861 (94.5)*	*4404 (65.8)* *662 (72.7)*	*6609 (98.8)* *887 (97.4)*	*122861 (63.1)* *10469 (72.9)*
4	*4947 (87.7)* *641 (90.2)*	*4013 (71.1)* *619 (87.1)*	*5532 (98.1)* *675 (94.9)*	*108869 (69.3)* *3478 (46.3)*
5	*7214 (92.4)* *670 (93.7)*	*4671 (59.8)* *606 (84.8)*	*7734 (99.1)* *688 (96.2)*	*277329 (81.9)* *5844 (74.2)*
6	*4062 (83.4)* *749 (92.1)*	*3368 (69.1)* *660 (81.2)*	*4789 (98.3)* *783 (96.3)*	*57756 (57.3)* *5707 (56.4)*
7	*7023 (91.6)* *692 (93.4)*	*5259 (68.6)* *617 (83.3)*	*7563 (98.7)* *692 (93.4)*	*150723 (58.1)* *3830 (54.0)*
8	*18118 (99.3)* *2001 (98.1)*	*12211 (66.9)* *1595 (78.2)*	*18167 (99.5)* *1988 (97.5)*	*310782 (31.2)* *51026 (80.7)*
9	*9694 (96.1)* *1670 (97.9)*	*6730 (66.7)* *1331 (78.0)*	*9991 (99.0)* *1674 (98.1)*	*103496 (37.9)* *15526 (33.7)*
10	*2648 (81.4)* *440 (90.2)*	*2213 (68.0)* *373 (76.4)*	*3175 (97.6)* *460 (94.3)*	*37975 (75.7)* *5838 (92.2)*

Table S-4. *Parameters used for series extraction, R nontarget package, function homol.search(). See package manual for parameter descriptions.*

Parameter	Article symbol	Value
elements	-	*FALSE*
use_C	-	*TRUE*
minmz	$\Delta m/z_{min}$	*3*
maxmz	$\Delta m/z_{max}$	*80*
minrt	ΔRT_{min}	*-2 [minutes]*
maxrt	ΔRT_{max}	*2 [minutes]*
ppm	*related to* ε	*TRUE*
mztol	ε	*3*
rttol[1]	$\Delta\Delta RT$	*.2*
minlength	n_{min}	*5*
mzfilter	-	*FALSE*
spar	*related to* λ [2]	*0.45*
R2	R^2	*0.98*

[1] *Not to be confused with the parameter in Table S-6.*
[2] *Check R function smooth.spline() for how this parameter is defined.*

Table S-5. *Blank subtraction settings, R enviMass package, function find.raw(). See package manual for parameter descriptions.*

Parameter	Value
dmz	*5*
ppm	*TRUE*
dRT	*2 [minutes]*
int	*0.1 * peak intensity*
kdTree	*FALSE*

Table S-6. *Isotopologue grouping parameters used for the function pattern.search2(), nontarget package. See package manual for parameter descriptions.*

Parameter	Value
mztol	*3*
ppm	*TRUE*
inttol	*.3*
rttol[1]	*.1 [minutes]*
use_isotopes	*FALSE*
use_charges	*FALSE*
use_marker	*TRUE*
quick	*TRUE*

[1] *Not to be confused with the parameter in Table S-4.*

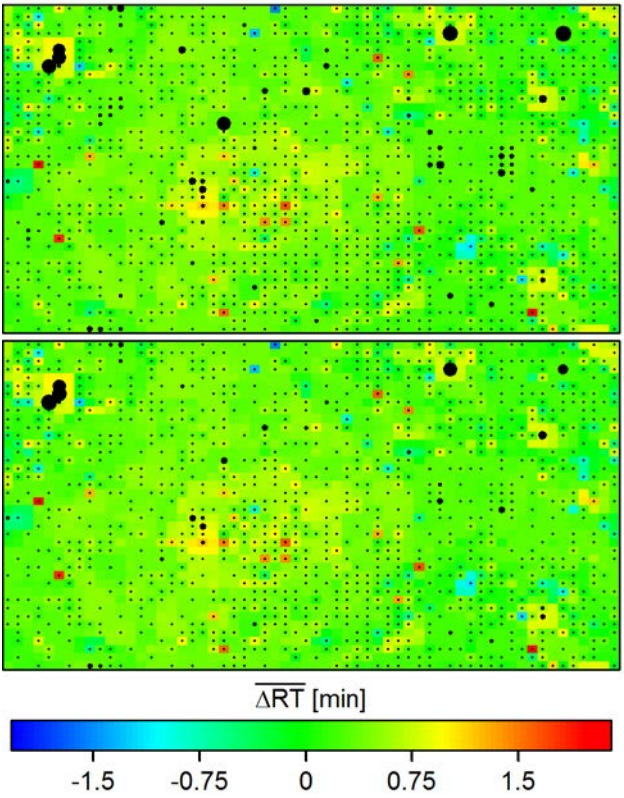


Figure S-7. *SOM results for all series pairs from STP sample with ID=1 (positive mode), complementing Figure 2 in the corresponding chapter. Coloring shows* $\overline{\Delta RT_x}$ *and* $\overline{\Delta RT_y}$ *at the SOM grid nodes, respectively. Dots in the top panel indicate frequencies of all series pairs clustered onto individual SOM nodes. In contrast, lower panel dots indicate mapping frequencies after deisotoping.*

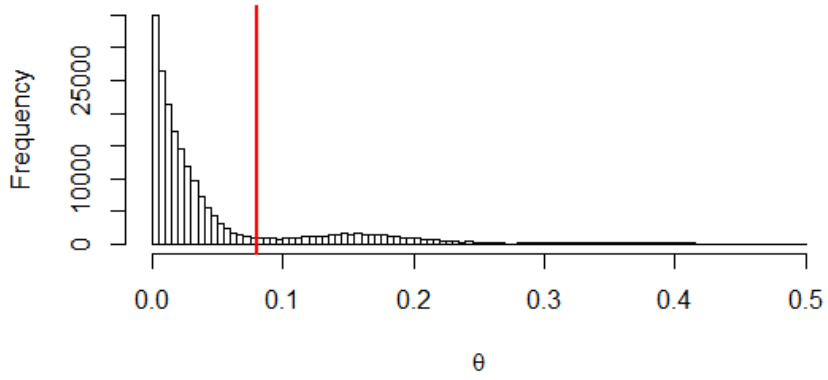


Figure S-8. *Distribution of intersection angle θ in series pairs of STP sample ID=1. The red line signifies a threshold of θ=0.08π (equivalent to 14.4°).*

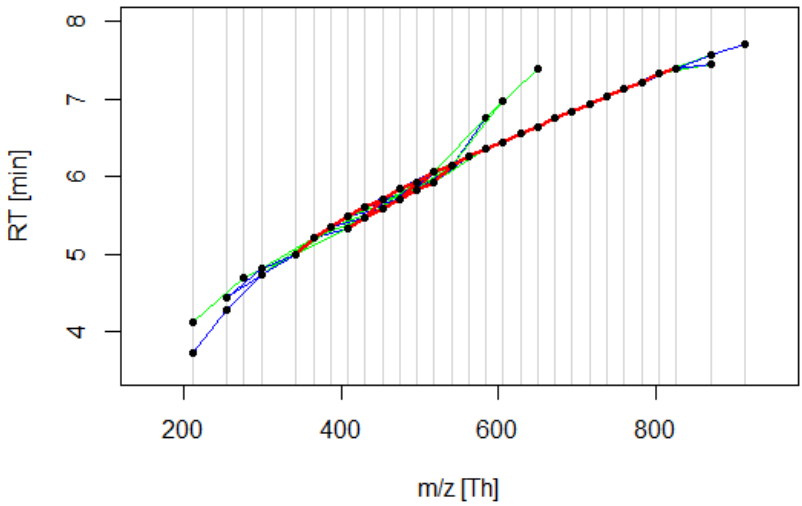


Figure S-9. *Example of 54 superjacent series for STP sample ID=1. Red, blue and green lines connect series peaks at* $\overline{\Delta m/z} \approx 22.013$, *44.026 and 66.039 [Th], respectively. Gray lines signify the spacing of series peaks in the m/z dimension, i.e., peaks on the same line are isobaric.*

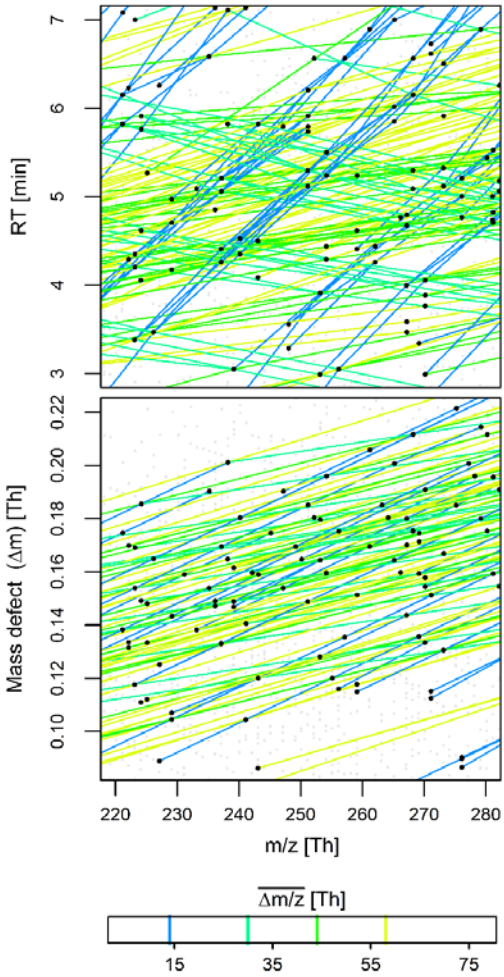


Figure S-10. *Characteristics of paired series with* $\overline{\Delta m/z}$ *= 14.016, 30.011, 44.026 and 58.042 Th, selected from the zoom window of Figure 3 in the corresponding chapter. The peaks incorporated into series are shown as black dots, others as gray ones.*

SUPPORTING INFORMATION FOR CHAPTER 5

Tables S-1. *Parameters used for peak picking with the R enviPick package, functions mzagglom(), mzclust(), and mzpick(), respectively. See enviPick package manual for detailed parameter descriptions.*

Parameter	Value
dmzgap	*>3.5*
ppm	*TRUE*
drtgap	*300 [seconds]*
minpeak	*4*
maxint	*1E7*

Parameter	Value
dmzdens	*3.5*
ppm	*TRUE*
drtdens	*60 [seconds]*
minpeak	*4*
maxint	*1E7*

Parameter	Value
minpeak	*4*
drtsmall	*20 [seconds]*
drtfill	*10 [seconds]*
drttotal	*120 [seconds]*
recurs	*3*
weight	*1*
SB	*2*
SN	*5*
minint	*1E4*
maxint	*1E6.5*
ended	*1*

Table S-2. *Parameters used in the enviMass workflow GUI.*

Parameter	Value
τ	*0.9*
ε	*3 ppm*
ΔRT_{prof}	*60 [seconds]*
lags n	*4,7,14 [days]*
θ_1	*3*
θ_2	*100*
$\Delta RT_{screen,large}$	*60 [seconds]*
$\Delta m/z_{screen,large}$	*6 ppm*
$\Delta RT_{screen,small}$	*20 [seconds]*
$\Delta m/z_{screen,small}$.	*1.5 ppm*

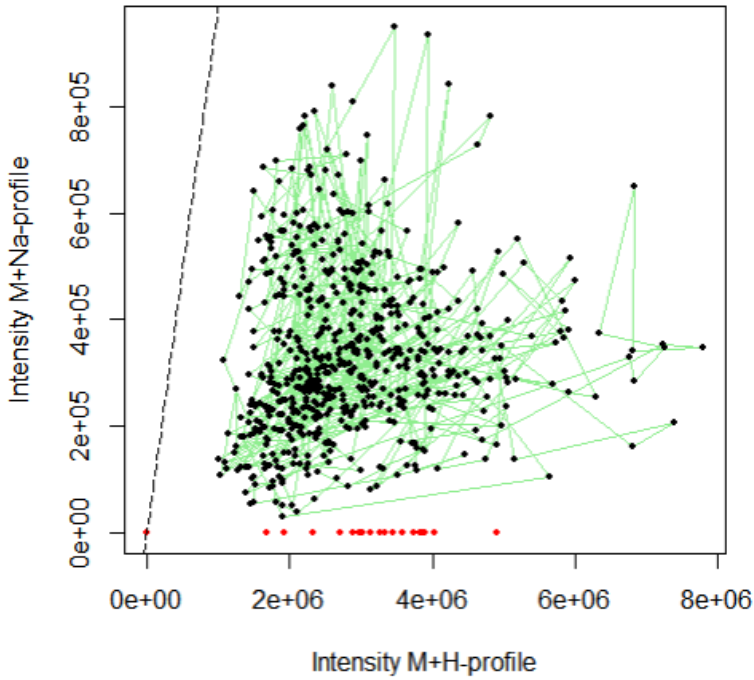


Figure S-1. *Intensity correlation between the profiles of two adducts of the same IS compound Diuron-D6. Black dots show centroid peak intensities at sequence time points for which both adducts were measurable; for red dots, only one of both profiles contained a peak. Green lines indicate a consecutive sequence time trajectory for paired profiles. The dashed line indicates a 1:1 intensity ratio.*

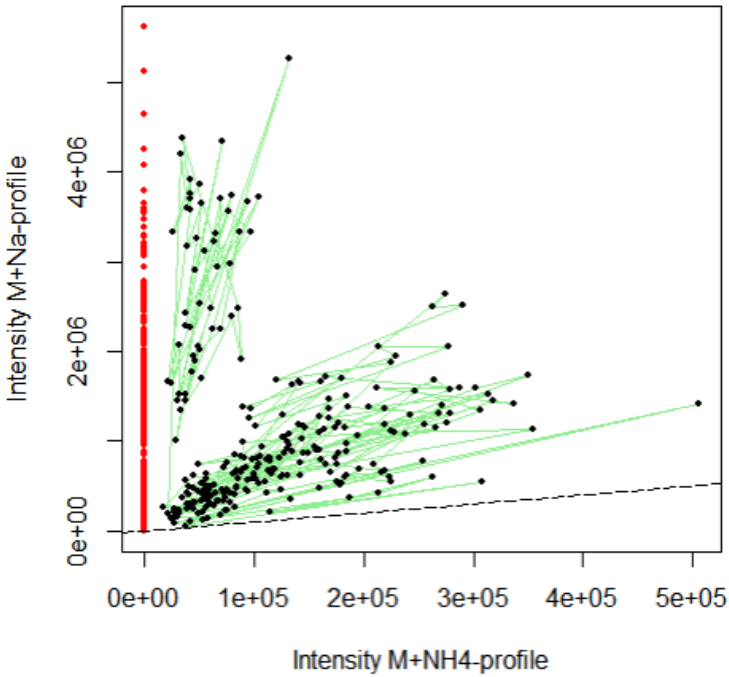


Figure S-2. *Intensity correlation between two adduct profiles of the IS compound Tipranavir–D4. Check caption of Figure S-1 for further explanation.*

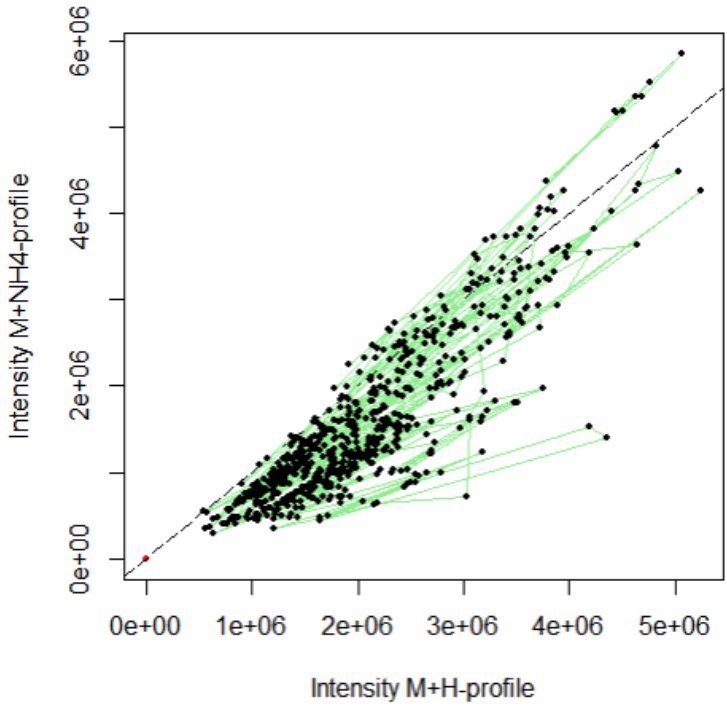


Figure S-3. *Intensity correlation between the profiles of two adducts of the same IS compound Naproxen-D3. Check caption of Figure S-1 for further explanation.*

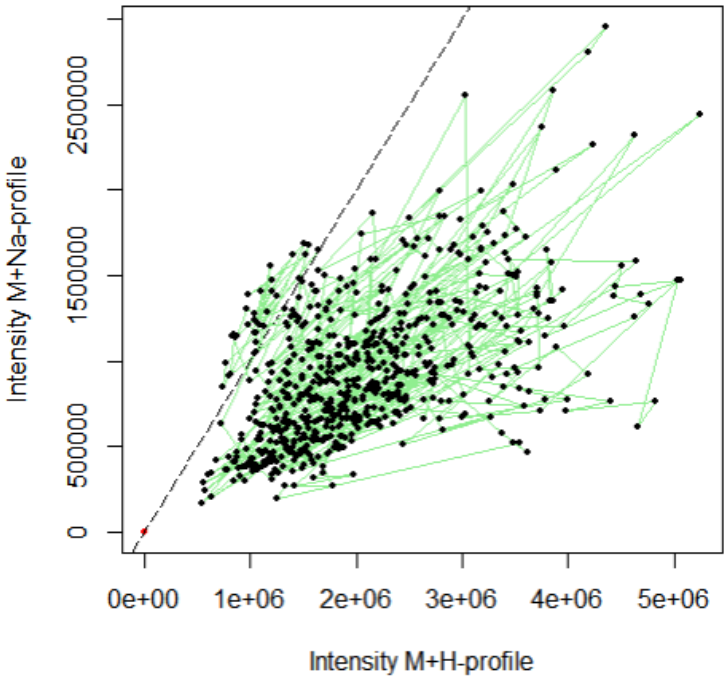


Figure S-4. *Intensity correlation between two adducts profiles of the same IS compound Naproxen-D3. Check caption of Figure S-1 for further explanation.*

Acknowledgements

This thesis would not have been possible, and sometimes bearable, without the support of many other individuals. First, I want to thank my Eawag supervisors Heinz Singer and Juliane Hollender. Heinz was patient and willing to share a lot of his expertise (and some past-7.30 pm time) on several issues, foremost analytical chemistry; and maintained a very supportive atmosphere. He was never reluctant to help, and built me up, at any time. Similarly, Juliane`s support and organizational advice was always in reach. Both made their way for several days of in-depth discussions to Helsinki, Finland. Speaking of the far north: I gratefully thank Francesco Corona for hosting me at Aalto University; his invitations; his internal revisions and helpful feedbacks and suggestions on machine learning; his positive mood and lengthy discussions on mass spectrometric topics. Thanks also to Yoan Miche, Amaury Lendasse and Alexander Grigorevskiy. In addition, thanks go to Kristopher McNeill and Steffen Neumann for participating in the defense committee; as well as Ruben Kretzschmar who volunteered to chair the defense. Valuable discussions, hints, data and exchange was also provided by Matthias Ruff, Emma Schymanski, Jennifer Schollee, Uwe Schmitt, Philip Longree, Aurea Chiaia, Rebekka Gulde and Rahel Comte. Special gratitude goes to Christian Gerber for helping at many points, without a single complaint: (re)programming *enviPat*, deriving a fabulous website or keeping the package up-to-date. Finally, I also want to thank the team at the Rhine monitoring station, Basel, i.e., Steffen Ruppe and Jan Mazacek; it was a pleasure (and will be) to program *enviMass* for them and their awesome data sets, which I guess are globally unique.

Sincere thanks go to Linda Mariko Feichtinger, who helped and advised me; and to my parents Ursula and Jürgen Loos for all they did. Miguel Loos & Co. are included there, too. Aduccia Sciacovelli supported this research with

an emergency bell, amongst other things. In a similar spirit, I want to thank Tobias Doppler (who warned me, but with little success); and of course my (former) office mates Andreas Waffel-Moser, Frederik Beuteltier-Weiss, Irene Wittmer and Devon Wemyss. I want to thank everyone at UCHEM for the great atmosphere which, surely, is above standard. I also want to thank everyone at Aalto who have not noticed the decline in computer performance at their workplace.

The presented thesis was partly financed by the Swiss Federal Office for the Environments (FOEN) and a mobility fellowship from the Swiss National Science Foundation (SNF).