A network diagram consisting of various sized light blue circles connected by thin white lines, set against a solid blue background. The circles vary in size and are distributed across the page, with some larger circles acting as hubs.

Joint Research Programme  
BTO 2022.017 | February 2022

**Improved non-target  
screening-based  
identification through  
MS online  
prioritization**

Joint Research Programme

**KWR**

Bridging Science to Practice



# Report

## Improved non-target screening-based identification through MS online prioritization

**BTO 2022.017 | February 2022**

This research is part of the Joint Research Programme of KWR, the water utilities and Vewin.

### Project number

402045/096

### Project manager

Dr. Patrick S. Bäuerlein

### Client

BTO - Thematical research - Chemical safety

### Author(s)

Nienke Meekel MSc, Dennis Vughs MSc, Frederic Béen PhD, Dr. Andrea M. Brunner

*The authors who contributed to the scientific publications are listed above the publications.*

### Quality Assurance

Dr. Thomas ter Laak

### Sent to

This report is distributed to BTO-participants.

A year after publication it is public.

### Keywords

non-target screening, chemical water quality, mass spectrometry, prioritization

[Year of publishing](#)  
2022

[More information](#)  
Nienke Meekel MSc  
T 030-6069622  
E [nienke.meekel@kwrwater.nl](mailto:nienke.meekel@kwrwater.nl)

PO Box 1072  
3430 BB Nieuwegein  
The Netherlands

T +31 (0)30 60 69 511  
F +31 (0)30 60 61 165  
E [info@kwrwater.nl](mailto:info@kwrwater.nl)  
I [www.kwrwater.nl](http://www.kwrwater.nl)

**KWR**

February 2022 ©

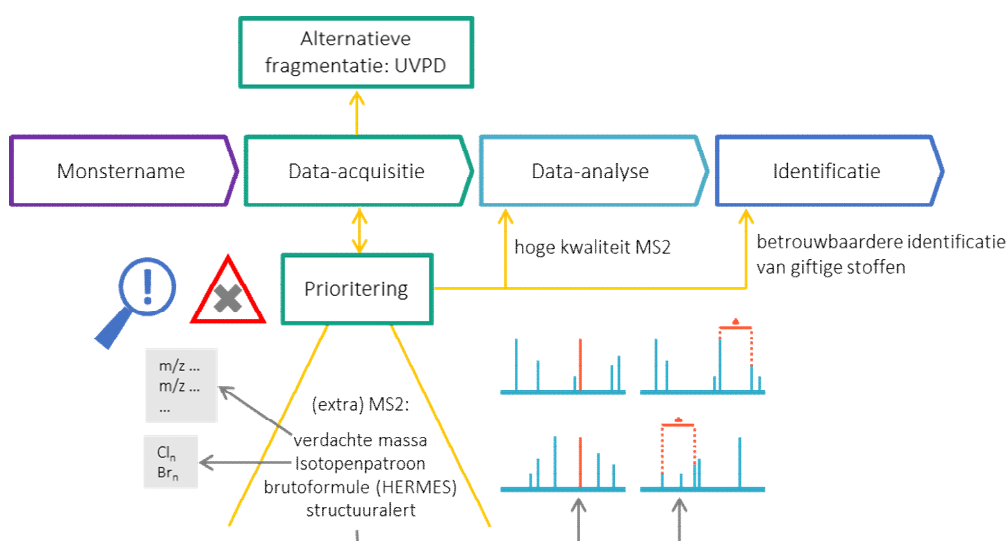
All rights reserved by KWR. No part of this publication may be reproduced, stored in an automatic database, or transmitted in any form or by any means, be it electronic, mechanical, by photocopying, recording, or otherwise, without the prior written permission of KWR.

# Managementsamenvatting

## Verbeterde prioritering en fragmentatie leiden tot betere identificatie van onbekende stoffen met non-target screening

**Auteurs** Nienke Meekel MSc, Dennis Vughs MSc, Frederic Béen PhD, Dr. Andrea M. Brunner.

Er zijn twee verschillende prioriteringsstrategieën ontwikkeld voor de identificatie van onbekende stoffen in watermonsters met non-target screening (NTS) met als doel potentieel toxische stoffen te herkennen in oppervlaktewater. De prioritering is gericht op het herkennen van substructuren van stoffen die bijdragen aan de giftigheid van een stof. Hiervoor is gebruikgemaakt van isotopenpatronen, aanwezigheid in een inclusielijst, brutoformule (met behulp van HERMES) en de aanwezigheid van een structuuralert (moleculaire substructuur die gelinkt is aan de toxiciteit van een stof). Ook is een nieuwe fragmentatietechniek op basis van fotonen (UVPD) toegepast om de identificatie van onbekende stoffen te verbeteren. Deze strategieën (HERMES, UVPD, structuuralerts en inclusie- en exclusielijsten) zijn effectief en kunnen al worden toegepast, maar dit proof of principle vraagt verdere ontwikkeling om in de toekomst de identificatie van onbekende en relevante stoffen nog verder te verbeteren en de strategieën breder toe te kunnen passen.



Workflow van de verbeterde prioriteringsstrategieën in non-target screening voor de identificatie van onbekende stoffen

### Belang: prioritering van relevante features in NTS

Non-target screening (NTS) van watermonsters resulteert vaak in een grote hoeveelheid *features*: combinaties van een retentietijd, intensiteit en accurate massa, die kunnen worden gebruikt om de identiteit van een stof te bepalen. Vanwege die grote hoeveelheid *features* in watermonsters is prioritering voor de identificatie noodzakelijk. Bij de gebruikelijke NTS data-analyse worden meestal de *features* met

de hoogste intensiteit geselecteerd. Dit zijn echter niet altijd de meest relevante *features* vanuit toxicologisch oogpunt. Bovendien is de fragmentatiescan niet altijd van voldoende kwaliteit voor identificatie. Met een betere fragmentatie én prioritering van relevante *features* zal het identificatieproces verbeteren en worden additionele analyses voor identificatie van onbekende stoffen voorkomen.

### **Aanpak: optimalisatie van prioritering en fragmentatie**

Een slimme data-acquisitiemethode is ontwikkeld om ook features met lage intensiteit te kunnen prioriteren op basis van de aanwezigheid van bepaalde eigenschappen. Tevens is AcquireX getest: een methode om niet-relevante features uit te sluiten voor fragmentatie via *exclusielijsten* waarin specifieke features die ook in de blanco voorkomen. Ook is een alternatieve fragmentatiemethode op basis van fotonen (UVPD) getest om de identificatie van (geprioriteerde) features te optimaliseren.

### **Resultaten: focus op potentieel toxische stoffen en alternatieve fragmentatiemethode**

De toegepaste strategieën bleken succesvol in de prioritering van potentieel toxische stoffen. Stoffen met specifieke substructuren (structuuralerts) konden worden herkend, waarna al tijdens de analyse een extra fragmentatiescan kon worden gestart om de stof beter te kunnen identificeren. Er is een workflow (HERMES) ontwikkeld om op basis van brutoformules *inclusielijsten* te genereren met specifieke accurate massa's die aanleiding geven tot een hogere prioritering omdat zij op de aanwezigheid van potentieel toxische stoffen duiden. Het gebruik van inclusielijsten helpt om meer relevante features te selecteren voor een fragmentatiescan. De vergelijking tussen UVPD en de reguliere (HCD, higher-energy C-trap dissociation) fragmentatie laat zien dat UVPD tot betere fragmentatie van sommige stoffen kan leiden wat identificatie mogelijk kan maken van microverontreinigingen die niet goed fragmenteren met HCD.

In dit proof of principle zijn de fundamenten gelegd voor verschillende tools om prioritering en identificatie van relevante microverontreinigingen in non-target screening te verbeteren. De volgende stap is het bepalen van parameters die de kwaliteit

van fragmentatiespectra kunnen weergeven, zodat de methode verder kan worden ontwikkeld.

### **Implementatie: (delen van) workflows projectmatig inzetten**

De prioriteringsstrategie is nog niet volledig doorontwikkeld, maar delen kunnen reeds ingezet worden in de NTS analyse. Met name exclusie- en inclusielijsten, isotopenpatronen (indicatie voor antropogene stoffen), bekende structuuralerts en de HERMES-strategie kunnen reeds toegepast worden mits de acquisitiesoftware van de massaspectrometer dit toelaat. Gegeneerde inclusielijsten kunnen worden gedeeld met andere waterbedrijven en laboratoria.

### **Rapport**

Dit onderzoek is beschreven in het rapport *Improved non-target screening based identification through MS online prioritization* (BTO-2022.017) dat bestaat uit diverse wetenschappelijke artikelen: *Ultraviolet photodissociation for non-target screening-based identification of organic micro-pollutants in water samples* van Panse et al., gepubliceerd in *Molecules* (<http://dx.doi.org/10.3390/molecules25184189>); *HERMES: a molecular formula-oriented method to target the metabolome* van Giné et al., (gepubliceerd in *Nature Methods* (<https://doi.org/10.1038/s41592-021-01307-z>)) en *Online prioritization of toxic compounds in water samples through intelligent HRMS data acquisition* van Meekel et al., gepubliceerd in *Analytical Chemistry* (<https://doi.org/10.1021/acs.analchem.0c04473?rel=cite-as&ref=PDF&jav=VoR>). Dit onderzoek is een vervolg op het werk in de BTO-rapporten *Groepsgewijze analyse en beoordeling van stoffen: implementatie van 'structural alerts' in waterkwaliteitsmonitoring* (BTO 2013.059), *Non-target screening to identify unknowns: automation and increasing confidence* (BTO 2019.032) en *Nieuwe chemische meetmethoden* (BTO 2019.029).

# Contents

<b>1</b>	<b>Nederlandse samenvatting</b>	<b>6</b>
1.1	Uitdagingen in structurele identificatie van onbekende stoffen	6
1.2	Prioritering in non-target screening door fragmentatie van relevante stoffen	6
1.3	Alternatieve fragmentatietechnieken voor verbetering van het informatiegehalte van fragmentatiespectra	7
1.4	Conclusies	7
1.5	Uitdagingen en vooruitzichten	8
<b>2</b>	<b>English summary</b>	<b>9</b>
2.1	Challenges in structural identification of unknown compounds	9
2.2	Prioritization in non-target screening through fragmentation of relevant compounds	9
2.3	Alternative fragmentation techniques to improve the information content of fragmentation spectra	10
2.4	Conclusions	10
2.5	Remaining challenges and outlook	11

# 1 Nederlandse samenvatting

Het algemene doel van dit project bestaat uit het aanpakken van de limieten van non-target screening (NTS) en het verbeteren van de identificatie van organische microverontreinigingen in water monsters door de ontwikkeling van massaspectrometrische acquisitiemethoden. De in dit project ontwikkelde methoden (1) prioriteren vóór fragmentatie stoffen die een potentieel risico vormen voor de humane gezondheid en milieu en (2) passen alternatieve fragmentatietechnieken toe om de fragmentatie te optimaliseren. Deze innovaties kunnen in de toekomst leiden tot betere prioritering van potentieel toxische features, reductie van het aantal analytische stappen dat nodig is om bepaalde ionen te fragmenteren en uitbreiding van de toepassingsmogelijkheden (scala van chemische stoffen) door verschillende fragmentatietechnieken toe te passen. De toepassing van de verschillende innovaties bij de drinkwaterlaboratoria hangt af van de (software)mogelijkheden van de gebruikte Q-TOF massaspectrometers. De opbrengsten van het project bestaan uit drie peer-reviewed publicaties en een masterscriptie, deze samenvatting beschrijft de verschillende publicaties.

## 1.1 Uitdagingen in structurele identificatie van onbekende stoffen

NTS gebaseerd op de combinatie van vloeistofchromatografie gekoppeld aan hoge resolutie massaspectrometrie (HRMS) en gegevensanalyse op maat zijn tezamen met target en suspect screening belangrijke methoden om stoffen in drinkwater en haar bronnen te identificeren en temporele en ruimtelijke analyses uit te voeren. Desalniettemin blijft de identiteit van een groot aantal stoffen onbekend. De huidige NTS aanpak bestaat uit het matchen van de accurate massa's (afkomstig uit de MS1 spectra, de volledige scan) en de fragmentatiespectra (MS2) van een onbekende feature met de spectra in één of meerdere databases. Fragmentatiespectra zijn onmisbaar voor de identificatie van stoffen, maar vaak van slechte kwaliteit of niet aanwezig omdat slechts een beperkt aantal features gefragmenteerd kan worden tijdens de analyse. In het geval van ontbrekende fragmentatiespectra of fragmentatiespectra van slechte kwaliteit, kan de stof niet worden geïdentificeerd. Het ontbreken van zo'n fragmentatiespectrum kan worden veroorzaakt door een te lage intensiteit van het MS1 spectrum omdat alleen de  $n$  meest intense ionen worden gefragmenteerd in een non-target data-afhankelijke acquisitiemethode (DDA, data-dependent analysis). Daarnaast kan de identificatie bemoeilijkt worden door de complexiteit van de MS1 spectra en de aanwezigheid van achtergrondionen. Vaak ontstaan meerdere pieken van dezelfde stof (i.e. adducten zoals  $\text{NH}_4^+$  en  $\text{Na}^+$ , in-source fragmenten, dimeren etc.), deze kunnen leiden tot overvloedige MS2-spectra van dezelfde stof. Prioritering van relevante stoffen en exclusie van achtergrondionen van fragmentatie kan deze problemen verminderen. Suboptimale fragmentatiemethoden kunnen leiden tot informatie arme fragmentatiespectra. Alternatieve fragmentatietechnieken kunnen fragmentatiespectra informatiever maken en daarmee bijdragen aan structuuropheldering. Naast een fragmentatiespectrum van voldoende kwaliteit is de structuuropheldering ook afhankelijk van een bibliotheek waarin het spectrum is opgenomen. Momenteel vergroten alternatieve fragmentatiebenaderingen de complexiteit en zijn ze beperkt beschikbaar voor alle instrumenten die door (water)laboratoria gebruikt worden. Het is echter mogelijk dat deze technieken in de toekomst beter beschikbaar worden. Het is daarom van belang hun potentiële toegevoegde waarde (opnieuw) te onderzoeken.

## 1.2 Prioritering in non-target screening door fragmentatie van relevante stoffen

Momenteel vindt prioritering na de dataverwerking plaats. Het resultaat van de prioritering is een lijst met potentieel geïdentificeerde stoffen waarvan de identiteit bevestigd moet worden met een nieuwe analyse. Dit heeft als gevolg dat er een aanzienlijke tijd kan zitten tussen de meting, (manuele) prioritering en identificatie. In data-dependent acquisitie (DDA), worden de meest aanwezige ionen in het volledige scan spectrum (MS1)

geselecteerd voor een fragmentatiestap, resulterend in een fragmentatiespectrum (MS2). Omdat deze selectie alleen op de intensiteit is gebaseerd, komt het vaak voor dat irrelevante features die in grote hoeveelheden voorkomen (bijvoorbeeld in de achtergrond) geselecteerd worden voor fragmentatie. De ionen afkomstig van potentieel toxische stoffen die in lagere concentraties voorkomen of een lagere ionisatie efficiëntie hebben worden hierdoor niet meegenomen omdat de respons lager is dan de drempelwaarde, terwijl deze wel relevant kunnen zijn. Om dit probleem in DDA aan te pakken, hebben we intelligente HRMS-data acquisitie strategieën ontwikkeld. Deze strategieën zijn gebaseerd op structuraalerts, isotopenpatronen, bruto formule en de exclusie van achtergrondionen (met behulp van de AcquireX software). Deze strategieën zijn beschreven in het artikel '[Online prioritization of toxic compounds in water samples through intelligent HRMS data acquisition](#)', de scriptie '[Improved identification of toxic compounds in drinking water sources through HRMS based intelligent data acquisition](#)', en het artikel '[HERMES: a molecular formula-oriented method to target the metabolome](#)'. Deze studies laten zien dat de strategieën succesvol zijn in het faciliteren van de prioritering van de ionen afkomstig van potentieel toxische stoffen en dat ze het percentage van de gefragmenteerde achtergrondionen reduceren. De strategieën kunnen worden geoptimaliseerd om op grotere schaal toe te kunnen passen in bijvoorbeeld monitoringsstudies, eventuele vervolgprojecten (bv. BTO Screening/NTS) zouden gericht kunnen worden op de implementatie van deze strategieën. Als de (software)mogelijkheden van de massaspectrometer het toelaten kunnen de inclusie- en exclusielijsten reeds toegepast worden.

### 1.3 Alternatieve fragmentatietechnieken voor verbetering van het informatiegehalte van fragmentatiespectra

Nadat de relevante ionen geprioriteerd zijn voor fragmentatie, is het van belang dat dit leidt tot een MS2 spectrum met voldoende informatie voor de identificatie. De optimale fragmentatiemethode en -energie resulteren in MS2 spectra met voldoende karakteristieke fragmenten en weinig ruis, en zijn afhankelijk van de structuur van de stof. In het werk gepresenteerd in '[Online prioritization of toxic compounds in water samples through intelligent HRMS data acquisition](#)' zijn verschillende acquisitie instellingen (inclusief fragmentatie energie) vergeleken voor de fragmentatie van relevante ionen. Het gebruik van verschillende instellingen verhoogt de kans op fragmentatiespectra van betere kwaliteit. Voor de stoffen die niet of nauwelijks fragmenteren met HCD (higher-energy C-trap dissociatie), werd de alternatieve fragmentatietechniek ultraviolet fotodissociatie (UVPD) toegepast om informatieve spectra te verkrijgen. UVPD is een relatief nieuwe techniek waarbij gebruik wordt gemaakt van een 213 nm UV laser. De publicatie '[Ultraviolet photodissociation for non-target screening based identification of organic micro-pollutants in water samples](#)' beschrijft de toegevoegde waarde van UVPD voor de identificatie van waterrelevante stoffen. Deze studie laat zien dat de methode in staat is om de fragmentatie van specifieke ionen te verbeteren (details zijn beschreven in de publicatie). De combinatie van de conventionele HCD fragmentatie en de UVPD techniek zorgt ervoor dat fragmentatiespectra van hoge kwaliteit verkregen kunnen worden voor een grotere set van stoffen (ook stoffen die met de klassieke methode niet of beperkt fragmenteren), en daarmee de identificatie van geprioriteerde features verbeterd kan worden. Echter is de UVPD techniek relatief nieuw en bestaan er nog geen uitgebreide databases (fragmentatiebibliotheken). Daardoor is toepassing van deze techniek in monitoringsstudies een lange-termijn traject. UVPD wordt ook in andere instrumenten, met name ion mobility massaspectrometrie, toegepast. De verwachting is dan ook dat het een algemeen gebruikte, aanvullende techniek kan worden. De volgende stap zou kunnen bestaan uit het onderzoeken van de beschikbaarheid van databanken en/of het opzetten van samenwerkingsstudies met laboratoria die UVPD beschikbaar hebben zodat gegevens gedeeld kunnen worden en de kwaliteit en deelbaarheid hiervan geëvalueerd kan worden.

### 1.4 Conclusies

De verschillende strategieën blijken veelbelovend te zijn voor wateronderzoek en -monitoring. Delen van de data acquisitiestrategie zijn reeds geïmplementeerd bij KWR. Exclusie van achtergrondionen met behulp van AcquireX is succesvol in het verminderen van het percentage gefragmenteerde achtergrondionen. Indien AcquireX niet



beschikbaar is op het gebruikte instrument (bijvoorbeeld QTOF instrumenten) dan kunnen exclusielijsten met achtergrondionen ook handmatig gegenereerd worden maar indien er vergelijkbare software beschikbaar is voor QTOF massaspectrometers, dan is de verwachting dat dit nauwkeurigere resultaten oplevert. Indien gewenst kunnen inclusielijsten eenvoudig toegepast worden met als enige vereiste dat de acquisitiesoftware van de massaspectrometer is uitgerust met deze mogelijkheid. HERMES is nog niet toegepast binnen het laboratorium van KWR maar aangezien deze software vrij toegankelijk is kan het ook door de drinkwaterbedrijven gebruikt worden. Echter is deze techniek zeer arbeidsintensief m.b.t. meettijd en dataverwerking. Dit is dus ongeschikt voor regelmatige screening, maar kan van toegevoegde waarde zijn voor een uitgebreide analyse van een monster. De alternatieve fragmentatiemethode, UVPD, kan een verbetering zijn van de identificatie van waterrelevante stoffen maar is nog niet klaar voor implementatie. Slechts een paar massaspectrometers zijn uitgerust met UVPD en er zijn nauwelijks referentiespectra beschikbaar, de verwachting is dat dit in de toekomst een meer commercieel beschikbare techniek zal worden.

## 1.5 Uitdagingen en vooruitzichten

De ontwikkelde prioriteringsstrategieën en alternatieve fragmentatiemethode dragen bij aan de reguliere NTS workflows door de identificatie van onbekende stoffen in watermonsters te verbeteren. De uitkomsten van dit onderzoek geven het belang van verder onderzoek naar en verdere ontwikkeling en implementatie van intelligente acquisitie software. De alternatieve fragmentatietechniek UVPD kan indien gewenst verder verkend worden door een samenwerking op te zetten waarbij meetresultaten uitgewisseld en vergeleken kunnen worden om zo de toegevoegde waarde voor drinkwaterlaboratoria te bepalen. Vervolgonderzoek kan zich ook richten op de implementatie van de, in dit verkennende onderzoek, ontwikkelde online prioriteringsstrategie in drinkwaterlaboratoria. De potentie van online prioriteringsstrategieën kan alleen volledig tot zijn recht komen wanneer de beoordeling van de kwaliteit van fragmentatiespectra gedurende de acquisitie kan plaatsvinden. Indien een MS2 onvoldoende spectrale informatie bevat, kan een intelligente acquisitiemethode een extra fragmentatiestap activeren met alternatieve parameters om zo de kwaliteit van het spectrum te verbeteren. Echter zijn er momenteel nog geen parameters om de kwaliteit van fragmentatiespectra te bepalen. Een spectrum is een goed spectrum wanneer het voldoende karakteristieke fragmenten bevat om een stof ondubbelzinnig te kunnen identificeren. Parameters voor spectrale kwaliteit kunnen ervoor zorgen dat de MS tijdens de acquisitie kan bepalen of een fragmentatiespectrum van voldoende kwaliteit is of dat er alternatieve fragmentatie instellingen vereist zijn. Deze parameters zouden bepaald kunnen worden met behulp van machine learning en dan worden geïmplementeerd in intelligente data acquisitie methoden. Vervolgens kan real-time herkenning van spectrale kwaliteit geïmplementeerd worden in de MS acquisitiesoftware.

## 2 English summary

The main goal of this project was to address the limitations of non-target screening (NTS) and to improve structural identification of organic micropollutants in water samples by developing mass spectrometric (MS) acquisition methods that (1) target compounds that potentially pose a risk to human health and the environment for fragmentation and (2) apply alternative fragmentation techniques to optimize fragmentation for a broader range of chemicals. These innovations can in future lead to a better prioritization of ions originating from potentially toxic compounds, shorten the number of analytical steps to fragment selected ions and broaden the window of application (the range of chemicals) by applying multiple fragmentation techniques. The extent to which the various innovations can be applied in drinking water laboratories depends on the (software) capabilities of the Q-TOF mass spectrometers used. The result of the project consists of three peer-reviewed publications and one MSc thesis, which are all included in this document.

### 2.1 Challenges in structural identification of unknown compounds

NTS based on the combination of liquid chromatography coupled to high-resolution mass spectrometry (HRMS) and tailored data analyses, together with suspect- and target screening have become important methods to identify and monitor compounds present in drinking water and its sources. However, a high number of compounds still remains unidentified with the current NTS approaches that rely on matching of the accurate mass (provided in the MS1 spectra, the full scan) and the fragmentation spectra (MS2) of a given unknown peak with those of database entries. Often fragmentation spectra, that are essential for the identification of substances, are of poor quality or are absent as only a limited number of ions can be fragmented during analysis. In the case of poor or a complete lack of MS2 fragmentation spectra, the compound can consequently not be identified. A lack of fragmentation can be due to low signal intensities of the compound, as only the  $n$  most intense peaks are fragmented in a non-target data dependent acquisition method. Moreover, the high complexity of MS1 spectra and background signals can hinder identification, e.g. due to peaks belonging to the same compound, in source fragments, adducts ( $\text{NH}_4^+$ ,  $\text{Na}^+$ ), dimers, and redundancy in MS2 spectra from the same (background) compound. Prioritizing compounds of interest for and excluding background compounds from fragmentation could alleviate this complexity. Suboptimal fragmentation methods can lead to poor fragmentation spectra. In that case, alternative fragmentation techniques can aid structural elucidation. Currently, alternative fragmentation approaches enhance the complexity and are not largely available for all instruments used by (water)laboratories. However, it is possible that these techniques become available at a larger scale in the future, therefore it is necessary to keep up with time and evaluate (again) their potential added value.

### 2.2 Prioritization in non-target screening through fragmentation of relevant compounds

To date, prioritization takes place after data processing. The result of the prioritization is often a list of potentially identified chemicals which need be confirmed by sample re-analysis. Consequently, a considerable amount of time can pass between measurement, (manual) prioritization and identification. In data-dependent acquisition (DDA), the most abundant ions detected in the full scan spectrum (MS1) are selected for a fragmentation event resulting in a fragmentation spectrum (MS2). Since this selection is based on intensity, it often occurs that highly abundant but irrelevant ions such as background ions are selected for fragmentation. Ions related to potentially toxic compounds that are present at lower concentrations and/or have a lower ionization efficiency are disregarded as their responses do not meet the threshold. To overcome this issue in DDA, we developed intelligent HRMS data acquisition strategies. These strategies are based on structural alerts, isotopic patterns, molecular formula and exclusion of background ions (the latter using the AcquireX software). The strategies are described in the paper

'[Online prioritization of toxic compounds in water samples through intelligent HRMS data acquisition](#)', the thesis '[Improved identification of toxic compounds in drinking water sources through HRMS based intelligent data acquisition](#)', and the paper '[HERMES: a molecular formula-oriented method to target the metabolome](#)'. These studies show that the strategies successfully aid prioritization of ions related to potentially toxic compounds, decrease the percentage of fragmented background ions and that these can be optimized for usage on a larger scale in for example monitoring studies, potential follow-up projects (e.g. BTO Screening/NTS) could be directed towards implementation of these strategies. It does however require a predefined list of structural alerts and their fragments and deltas, and inclusion lists that indicate that the ion is originating from a potentially toxic compound and thereby relevant to prioritize for fragmentation and further identification. If the (software) capabilities of the mass spectrometer allow, the inclusion and exclusion lists can already be applied.

### 2.3 Alternative fragmentation techniques to improve the information content of fragmentation spectra

After prioritising relevant ions for fragmentation, the triggered fragmentation event needs to result in a MS2 spectrum with sufficient information for subsequent identification. The optimal fragmentation method and fragmentation energy produce in MS2 spectra with enough characteristic fragments and little noise, and depend on the structure of the fragmented compound. In the work presented in '[Online prioritization of toxic compounds in water samples through intelligent HRMS data acquisition](#)' different acquisition settings including fragmentation energy were compared for the fragmentation of relevant ions. Using different settings increased the chance of high quality fragmentation spectra. For those compounds that fragment poorly – or not at all – using HCD (higher-energy C-trap dissociation), the alternative fragmentation technique Ultraviolet Photodissociation (UVPD) was applied to obtain informative spectra. UVPD is a relatively new fragmentation technique relying on a 213 nm UV laser. The publication '[Ultraviolet photodissociation for non-target screening based identification of organic micro-pollutants in water samples](#)' describes the added value of UVPD for the identification of water relevant compounds. This study showed that the method was able to improve fragmentation for specific ions (details are described in the publication). Thereby, the combination of the conventional HCD fragmentation and UVPD technique enables high quality fragmentation spectra for a broader set of chemicals (including chemicals which fragment hardly or poorly using classical fragmentation techniques) and thereby improving the potential for identification of prioritized features. However, UVPD is a relatively new technique and thus far no extended databases (involving fragmentation spectra) exist, yet. So usage of this technique in monitoring studies will be a long-term trajectory. UVPD is also applied in other mass spectrometry instruments, mainly in ion mobility mass spectrometry. It is therefore expected that it will become a more common technique. The next step could consist of the examination of the availability of databases and/or the set-up of collaborative studies with labs having an UVPD instrument available, so that data can be shared and the quality and shareability of these data can be examined.

### 2.4 Conclusions

The different strategies appear to be promising for water research. Parts of the data acquisition strategies are already implemented within KWR. Background exclusion using AcquireX successfully decreases the percentage of fragmented background ions. If desired, inclusion lists can be applied easily with the only requirement that the mass spectrometer acquisition software is equipped with this option. HERMES has not been applied within the KWR laboratory yet, but since this technique is open-source it can be used by the drinking water companies as well. However, this technique is very labour intensive in terms of measuring time and data processing. It is therefore unsuitable for regular screening, but might be of added value for extensive analysis of a sample. The alternative fragmentation method, UVPD, has benefits for the identification of water relevant compounds but is not yet ready for implementation. Only a few mass spectrometers are equipped with UVPD and almost no reference spectra are available, so this will become relevant on the long-term. However, it is expected that UVPD will become a more commonly and commercially available technique in the future.

## 2.5 Remaining challenges and outlook

The developed prioritization strategies and alternative fragmentation method contribute to the regular NTS workflows by increasing the probability of identification of unknown compounds in water samples. The outcomes of this study indicate the need for further research into and development and implementation of online intelligent acquisition software. The alternative fragmentation technique (UVPD) could be examined further by setting up a collaboration to share and compare measuring results in order to test the added value for drinking water laboratories. Future research could also be directed to the implementation of the, in this exploratory research, developed prioritization strategies in drinking water laboratories. The potential of online prioritization strategies can only be fully harnessed if assessment of spectral quality happens in real time during the MS acquisition. In case an MS2 contains insufficient spectral information, an intelligent acquisition method could then trigger a further mass spectrometric event with alternative parameters to increase spectral quality. However, to date there are no metrics that define spectral quality. A spectrum is a good spectrum if it contains enough characteristic fragments to unambiguously identify a compound. Spectral quality metrics could allow the mass spectrometer to decide during data acquisition whether a fragmentation spectrum is informative enough for organic micropollutant identification or whether it requires alternative fragmentation (settings). These metrics could be defined using machine learning strategies and then implemented in intelligent acquisition methods. Subsequently, real time recognition of spectral quality could be implemented into MS acquisition software to enable intelligent acquisition.

## **HERMES: a molecular formula-oriented method to target the metabolome**

Roger Giné<sup>1</sup>, Jordi Capellades<sup>1,2</sup>, Josep M. Badia<sup>1,2</sup>, Dennis Vughs<sup>3</sup>, Michaela Schwaiger-Haber<sup>4,5</sup>, Maria Vinaixa<sup>1,2</sup>, Andrea M. Brunner<sup>3</sup>, Gary J. Patti<sup>4,5</sup>, Oscar Yanes<sup>\*1,2</sup>

1. Universitat Rovira i Virgili, Department of Electronic Engineering & IISPV, Tarragona, Spain.
2. CIBER de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Instituto de Salud Carlos III, Madrid, Spain.
3. KWR Water Research Institute, Nieuwegein, The Netherlands.
4. Department of Chemistry, Washington University, St. Louis, MO, USA.
5. Department of Medicine, Washington University, St. Louis, MO, USA.

\*Corresponding author:

Oscar Yanes, PhD

Department of Electronic Engineering

Universitat Rovira i Virgili

Avinguda Països Catalans, 26, 43007 Tarragona, Spain

phone: +34 977759397

email: [oscar.yanes@urv.cat](mailto:oscar.yanes@urv.cat)

## ABSTRACT

**Comprehensive metabolome analyses are hampered by low identification rates of metabolites due to suboptimal strategies in MS and MS2 acquisition, and data analysis. Here we present a molecular formula-oriented and peak detection-free method, HERMES, that improves sensitivity and selectivity for metabolite profiling in MS and structural annotation in MS2. An analysis of environmental water, *E. coli*, and human plasma extracts by HERMES showed increased biological specificity of MS2 scans, leading to improved mass spectral similarity scoring and identification rates when compared to iterative data-dependent acquisition (DDA). HERMES is available as an R package with a user-friendly graphical interface to allow data analysis and interactive tracking of compound annotations.**

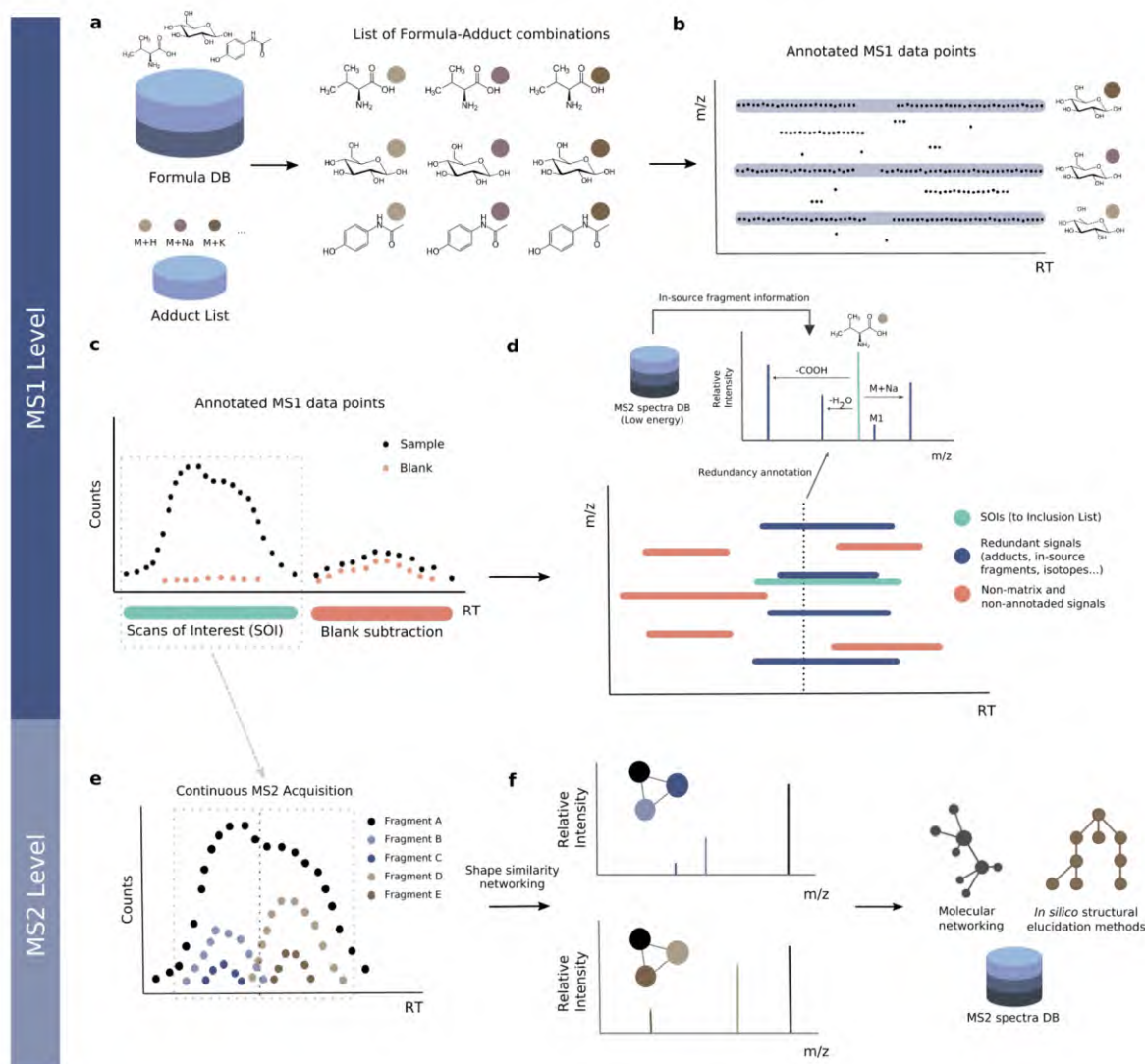
## INTRODUCTION

A single LC/MS-based metabolomic experiment generates millions of three-dimensional ( $m/z$ , retention time, intensity) data points that can be annotated and quantified into thousands of metabolite features. However, most features are either redundant ions caused by ionization-related phenomena such as cation/anion adduction, multimerization and in-source fragmentation, or unknown contaminants and artifacts<sup>1,2</sup>. Moreover, conventional untargeted metabolomic experiments lead to highly heterogeneous chromatographic peak shapes, which negatively affect the performance of peak detection<sup>3</sup> and grouping/annotation algorithms in MS1 mode<sup>4</sup>. These characteristics of MS1 data, in turn, negatively impact MS2 acquisition methods used for metabolite identification. In data-dependent acquisition (DDA) mode, MS2 spectra are automatically collected for precursor ions that exceed a predefined intensity threshold. The selection of precursor ions is a stochastic event suffering from low analytical reproducibility and favouring the selection of the most abundant, but not necessarily biologically relevant, ions. In data-independent acquisition (DIA) methods, multiple precursor ions, including redundant and biologically irrelevant ions, are simultaneously fragmented, often generating a series of complex convoluted MS2 spectra. Despite the emergence of new software to reconstruct the link between precursors and their fragments through mass spectral deconvolution<sup>5,6</sup>, MS2 spectral quality and matching scores to reference spectra are generally poorer in DIA compared to DDA<sup>7</sup>.

## RESULTS

Here we present HERMES, a novel experimental method and computational tool that improves the selectivity and sensitivity for comprehensive metabolite profiling in MS1, and identification in MS2. HERMES replaces the conventional untargeted metabolomic workflow that detects and annotates peaks<sup>8,9</sup>, for an inverse approach that directly interrogates raw LC/MS1 data points (i.e., scans) by using a comprehensive list of unique molecular formulas selected by the user. These are retrieved from large compound-centric databases (e.g., HMDB, ChEBI, NORMAN)<sup>10-12</sup>, genome-scale metabolic models, or specific metabolic pathways. Each molecular formula generates multiple 'ionic formulas' by adding or subtracting atoms from common adduct ions (Fig. 1). The resulting ionic formulas (on the order of  $10^4$ - $10^5$  from a database such as HMDB) are searched against millions of data points in an LC/MS1 experiment. HERMES calculates the theoretical isotopic pattern of each ionic formula based on a predefined experimental mass resolution value (Suppl. Fig. 1). The number of collisions between monoisotopic ionic formulas vary according to the experimental mass error (i.e., the smaller the error, the larger the percentage of non-overlapping ionic formulas; Suppl. Fig. 2). An LC/MS1 data point contains  $m/z$  and intensity information in a wide mass range (e.g.,  $m/z$  80 to 1,000) for a given instant of time. HERMES solves the limitations of peak detection by finding a series of scans, named SOI (Scans Of Interest), which are defined as clusters of data points that match an ionic formula and are concentrated within a short period of time (see Methods). SOI shapes do not necessarily fit a Gaussian-like function, as assumed in basic chromatography theory, making the process independent of the heterogeneous peak shapes commonly observed in LC/MS1 experiments from complex mixtures. SOIs are then filtered in three steps: (i) blank subtraction from the sample based on a convolutional neural network (Suppl. Fig. 3a), (ii) adduct and isotopologue grouping according to the similarity of their elution profiles (Suppl. Fig. 3b), and (iii) in-source fragment (ISF) annotation by using publicly available low-energy MS2 data (Suppl. Fig. 3c) extending on Domingo-Almenara et al.<sup>13</sup>. Finally, users can prioritize the SOIs that will constitute the inclusion list (IL) for targeted MS2 acquisition based on the following criteria: type and number of adducts, minimum intensity, isotopic fidelity, and a maximum number of overlapped precursors at any time range, which together determine the total number of MS2 runs. According to the MS2 acquisition settings, each entry in the IL may be associated with one or multiple MS2 scans: if there are more than five continuous scans, HERMES provides an optional deconvolution step (adapted from CliqueMS<sup>14</sup>) that resolves co-eluting compounds (Suppl. Fig. 4); if there are fewer scans, HERMES selects the most intense scan. The resulting curated MS2 spectra can either be identified within HERMES or exported as .mzML, .msp, or .mgf files to be used in other identification software such as NIST MS Search, SIRIUS<sup>15,16</sup>, or GNPS<sup>17</sup>.

HERMES is available as an R package (RHermes) and comes with an R Graphical User Interface (GUI) to allow data analysis, tracking of compound annotations, and visualization (Suppl. Fig. 5). RHermes accepts both CSV and XLS/XLSX files as valid molecular formula lists and can extract formulas from selected KEGG pathways for a given organism. The running time, including blank subtraction and IL generation, is <10 minutes on a six-core, 2.9 GHz CPU.

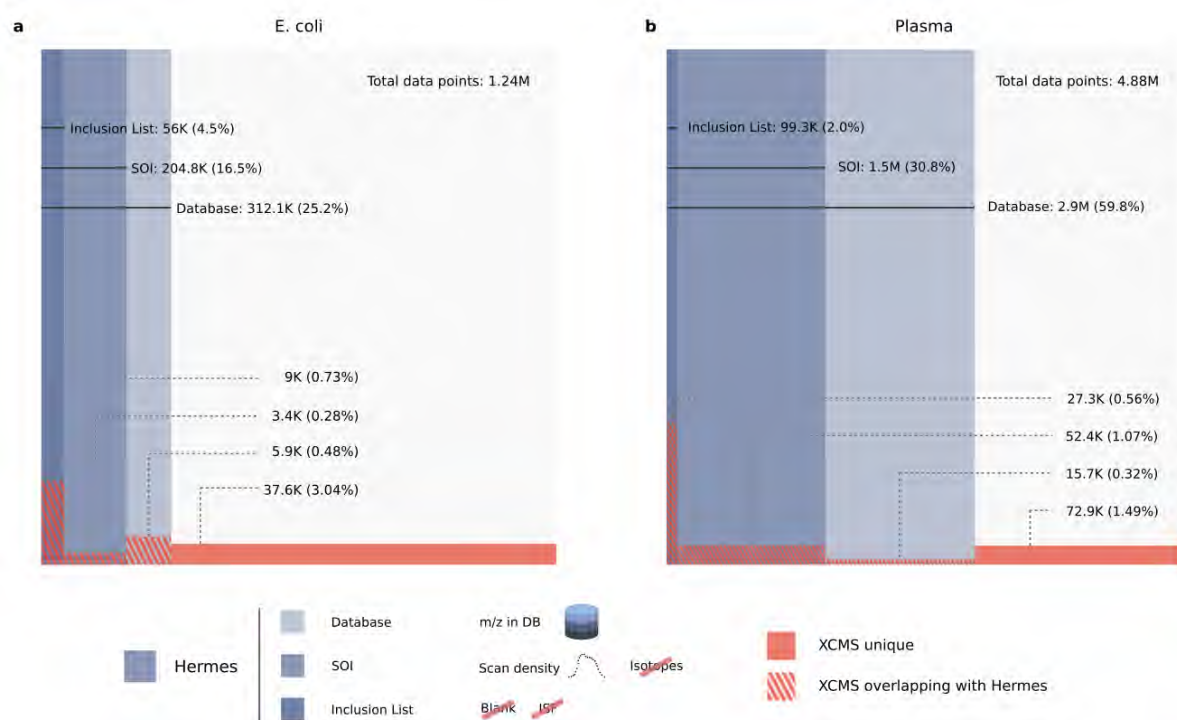


**Figure 1. The HERMES workflow.** (a) A context-specific database of molecular formulas and MS adducts generates a list of ionic formulas. (b) LC/MS1 data points are interrogated against all  $m/z$  ions corresponding to the ionic formulas and their isotopes. (c) Points with the same  $m/z$  annotation are grouped by density into retention time (RT) intervals called Scans of Interest (SOI). SOIs with similar shape and intensity in a blank sample are removed. (d) SOIs corresponding to different adducts of the same formula are grouped by their chromatographic elution profile. Similarly, in-source fragments are annotated on low intensity MS2 spectra of molecules with the same formula. The result is an inclusion list (IL) of sample-specific and non-redundant precursor ions that will be monitored in a posterior MS2 experiment. (e) The IL entries are acquired continuously along the defined RT interval and HERMES groups the resulting fragment elution profiles. (f) This results in deconvoluted spectra that can be queried against an MS2 database or exported to be used in alternative identification workflows.



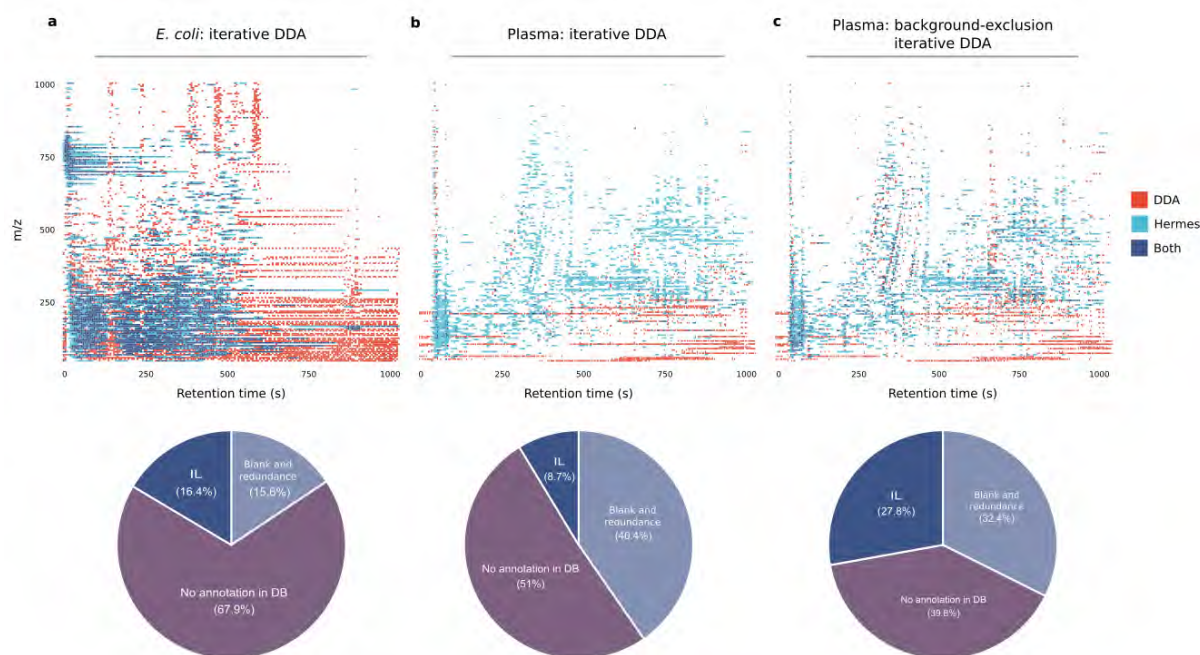
HERMES has been validated by using three (bio)chemically relevant samples of increasing complexity: (i) water collected from a canal in Nieuwegein (Netherlands), (ii) *E. coli*, and (iii) human plasma extracts. The canal water was spiked with 86 common environmental contaminants at 1 µg/L (Suppl. Table 1) and analyzed by RP/LC (C18) coupled to an Orbitrap in positive (pos) and negative (neg) ionization mode operating at 120,000 resolution. Using 118,820 (pos) and 46,809 (neg) ionic formulas calculated from 24,696 unique molecular formulas in the NORMAN database, HERMES detected and annotated all spiked compounds at the MS1 level. Certain ionic formula collisions, particularly those involving Cl, Br, S, or K, were automatically resolved by matching experimental isotopic patterns to the expected ones. This is the case, for example, of the  $[M+H]^+$  ion of chloridazon and the  $[M+K]^+$  ion of 2-amino- $\alpha$ -carboline, which overlapped at 0.27 ppm (Suppl. Fig. 6). In-source fragments that could be wrongly associated with ionic formulas were also annotated by using low-energy MS2 spectra when available. The output was a curated IL of 474 (pos) and 129 (neg) selective entries for targeted MS2 (Suppl. Fig. 7).

Next, a reference *E. coli* cell extract (Cambridge Isotope Laboratories) was analysed by HILIC coupled to an Orbitrap in positive and negative ionization mode. LC/MS1 data were analysed by HERMES by using 12,010 (pos) and 4,876 (neg) ionic formulas calculated from 2,463 unique molecular formulas obtained from the *Escherichia coli* Metabolome Database (ECMDB) and KEGG database. Interestingly, HERMES annotated ionic formulas for 25% (pos) and 22% (neg) of all data points acquired by the mass spectrometer (Fig. 2a and Suppl. Fig. 8a). In comparison with XCMS, a commonly used open-source LC/MS1 processing data tool in untargeted metabolomics<sup>9,18</sup>, 4.5% of all acquired data points were associated with an XCMS peak, 1.5% of data points in XCMS peaks matched an ionic formula from ECMDB and KEGG database, and only 0.7% of data points in XCMS peaks were represented in the final SOI list after blank subtraction, isotopic fidelity, and ISF removal. The outcome of HERMES was 2,058 (pos) and 1,081 (neg) SOIs that led to a curated IL of 1,251 and 661 entries for targeted MS2, respectively.



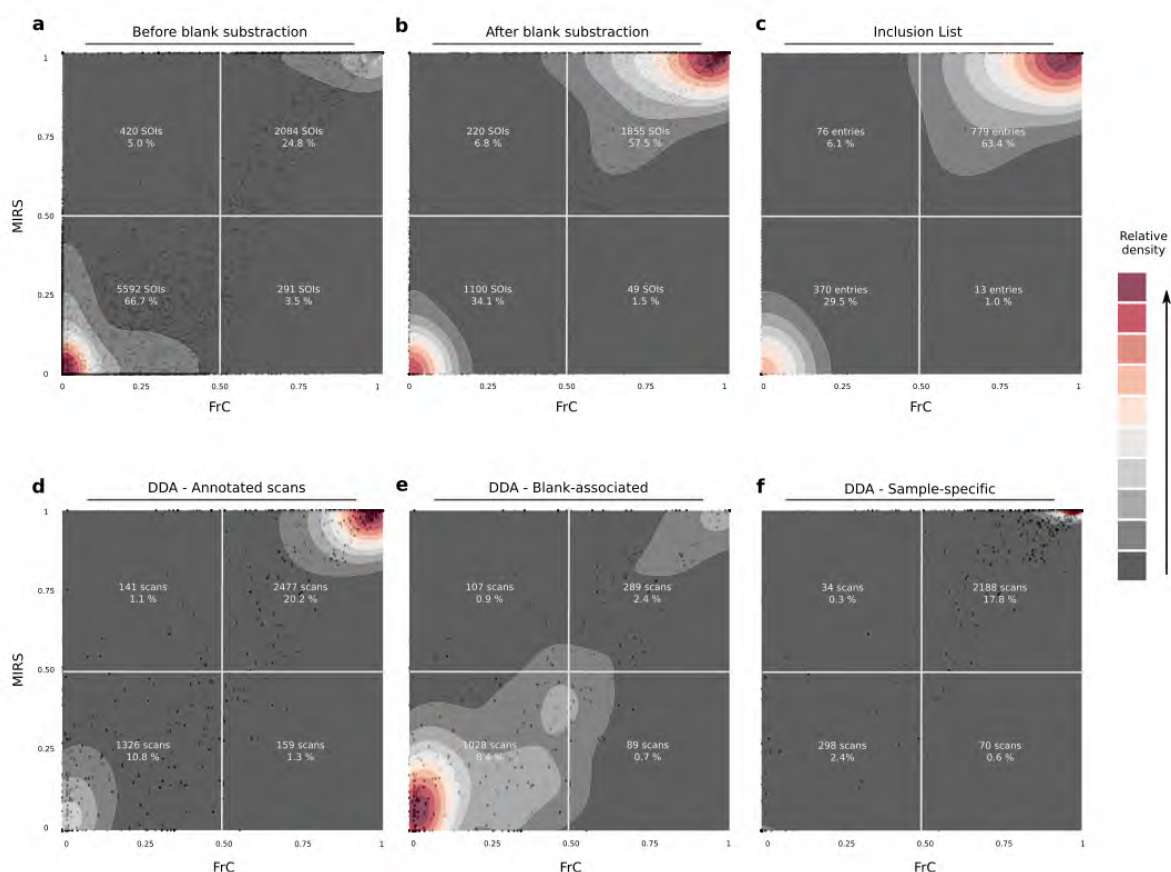
**Figure 2. Venn-like diagram of the distribution of LC/MS1 data points in different steps of the HERMES workflow and XCMS peak-associated points. a) *E. coli* extract. b) Plasma extract.** Database: Refers to all data points whose m/z matches with any m/z calculated from the ionic formula database (including isotopes). SOI: monoisotopic (M0)-annotated data points that are in Database and are also present in a SOI list that does not include blank subtraction nor any filtering. Inclusion List: data points present in Database and SOI kept through the blank subtraction, isotopic filter and ISF removal steps. Percentages refer to the total number of LC/MS1 data points. Positive ionization mode. On average, ~78% of data points in the inclusion list could not be annotated as a peak by XCMS. Conversely, ~84% scans annotated as a peak by XCMS could either not be matched to an ionic formula, were not specific of the sample or were associated with redundant signals.

The *E. coli* extract was also analysed by iterative DDA under identical analytical conditions. Remarkably, 68% of DDA scans could not be annotated as the monoisotopic signal by any ionic formula from ECMDDB and KEGG database (Fig. 3a), which indicates their exogenous or artefactual origin. After filtering out DDA precursor ions that were classified as SOIs in the blank sample, redundant adducts, and ISF by HERMES; only 16% of the DDA scans matched with any monoisotopic ionic formula in the inclusion list. In addition, HERMES included 571 inclusion list entries (46.5% of the total) that were not triggered by DDA.



**Figure 3. Distribution of MS2 scans acquired by HERMES and iterative DDA.** a) Unlabeled *E. coli* and b) human plasma samples acquired by iterative DDA. c) Human plasma sample acquired by iterative DDA with background-exclusion. The acquired scans have been binned into 5Da-5s intervals. The precursor m/z of DDA scans have been queried into the corresponding ionic formula m/z database with a 3 ppm mass error tolerance. Scans annotated in the database were further classified according to whether the m/z and retention time of the scans could be matched to the HERMES inclusion list or not. Percentages in the pie-charts refer to the total number of acquired DDA MS2 scans.

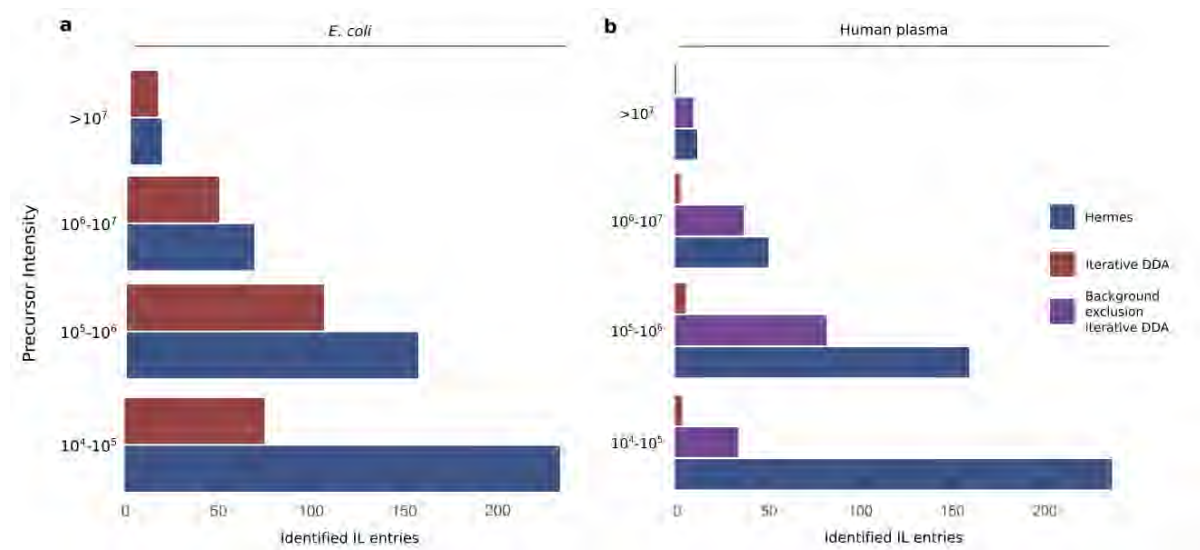
To confirm the biogenic specificity of the MS2 scans in HERMES, a reference  $^{13}\text{C}$ -labeled (at  $\geq 98\%$  from uniformly  $^{13}\text{C}$ -labeled glucose) *E. coli* credentialing extract was analysed under identical LC/MS1 conditions. For each selected precursor ion in the unlabeled *E. coli* sample, we calculated its fractional contribution (FrC)<sup>19-21</sup> and the monoisotopic ratio score (MIRS) by using the analogue  $^{13}\text{C}$ -labeled sample (see Methods). A metabolite with  $n$  carbon atoms can have zero (FrC=0) to  $n$  (FrC=1) of its carbon atoms labeled with  $^{13}\text{C}$ . In turn, similar intensity of the monoisotopic ion in the unlabeled and  $^{13}\text{C}$ -labeled *E. coli* extracts indicates no isotopic enrichment (MIRS=0), whereas loss of intensity in the  $^{13}\text{C}$ -labeled sample is associated with enrichment (MIRS=1). Around 63% of inclusion list entries in HERMES were associated with highly  $^{13}\text{C}$ -enriched metabolites (FrC and MIRS > 0.5), proving the biosynthetic origin of these ions (Fig. 4a-c). These are mainly associated with abundant ions, while unlabeled precursors relate more frequently to low-abundant ions (Suppl. Fig. 9a,b). In contrast, only 20% of all DDA scans were associated with  $^{13}\text{C}$ -labeled and annotated precursors from ECMDB and the KEGG database, pointing to ions also present in the blank sample as the main source of unlabeled precursors (Fig. 4d-f).  $^{13}\text{C}$ -labeled precursors in DDA corresponded to highly abundant ions that were also covered by IL entries in HERMES (Suppl. Fig. 9c).



**Figure 4.  $^{13}\text{C}$ -enrichment analysis in the labeled *E.coli* sample.** Each panel represents a scatterplot of two independent isotopic enrichment scores—FrC (Fractional Contribution) and MIRS (Monoisotopic Ratio Score)—and an overlaid density estimation. a) Distribution of SOIs before applying the blank subtraction filtering in HERMES. b) Same SOI list after removing most blank-related SOIs. c) SOIs in the MS2 inclusion list after removing redundant signals from b). d) Iterative DDA scans that could be matched to any  $m/z$  of the ionic formula database. e) DDA scans associated with SOIs removed during the blank subtraction step from a) to b). f) DDA scans associated with SOIs conserved during the blank subtraction. Percentages in a), b) and c) correspond to the total number of SOIs and inclusion list entries, accordingly, while percentages in d), e) and f) correspond to the total number of acquired DDA scans.

The biogenic specificity of HERMES resulted in higher similarity scores by mass spectral matching in databases (MassBankEU, MoNA, HMDB, Riken, NIST14, mzCloud)<sup>22</sup> than iterative DDA (see Methods). HERMES provided nearly double the number of confident structural metabolite annotations than iterative DDA (Fig. 5a and Suppl. Fig. 10a). The higher identification rate of HERMES was validated by using alternative spectral similarity and distance metrics (Suppl. Fig. 11). A fraction of the  $^{13}\text{C}$ -labeled compounds, however, could not be identified due to low intensity SOIs and/or the lack of reference spectra in databases. For the former, setting the maximum ion injection time at high values (1,500 ms) improved sensitivity and MS2 spectral quality in HERMES, resulting in more informative fragments and better spectral matching (Suppl. Fig. 12). Furthermore, we identified unlabeled metabolites

(FrC=0) in the  $^{13}\text{C}$ -labeled *E. coli* sample, such as choline, that we attribute to contaminants of the minimal growth medium that could not properly be removed by blank subtraction.



**Figure 5. Identified inclusion list entries according to the MS1 precursor intensity.** An inclusion list (IL) entry is considered identified if at least one MS2 scan associated with it has a compound hit in the reference MS2 database with either cosine score  $> 0.8$  (in-house database from MassBankEU, MoNA, Riken and NIST14 spectra), or Match  $> 90$  and Confidence  $> 30$  (mzCloud). Positive ionization data. a) *E. coli* extract. b) Human plasma extract.

Finally, we used a human plasma extract to compare HERMES and iterative DDA, with and without background exclusion<sup>23</sup>. Here we used 23,797 unique molecular formulas from the HMDB and Chemical Entities of Biological Interest (ChEBI) database to explore virtually all known exogenous and endogenous small molecules in this biofluid. HERMES generated 110,387 and 46,973 ionic formulas that covered 60% and 14% of all data points acquired by LC (RPC18)-Orbitrap MS in positive and negative ionization mode, respectively (Fig. 2b and Suppl. Fig 8b). Consistent with the pattern observed for *E. coli*, 3.8% of all acquired data points were associated with an XCMS peak. Only 0.5% of the data points in these peaks matched with an ionic formula from HMDB or ChEBI and were present in the inclusion list. Again, more than half of DDA precursors could not be annotated as monoisotopic ionic formulas from HMDB and ChEBI without blank subtraction (Fig. 3b). As expected, the overlap between HERMES and DDA increased upon background exclusion in iterative DDA, increasing to 29% of the number of common MS2 scans (Fig. 3c). Yet, the number of confident structural metabolite identifications with HERMES was more than three times greater than DDA because of the larger coverage of sample-specific and low abundant precursor ions (Fig. 5b and Suppl. Fig. 10b).

## DISCUSSION

Our results demonstrate that a conventional LC/MS-based untargeted metabolomic experiment can contain up to ~50 times more non-specific and redundant data points than sample-specific and selective ones, which can account for as much as 90% of the MS2 acquisition run time in an iterative DDA experiment. Current untargeted metabolomic approaches are unable to properly annotate the large number of 'junk' MS and MS2 signals, leading to false-positive identifications and an overall low number of identified metabolites. HERMES solves this problem by implementing a broad scope and molecular formula-oriented method that improves MS2 coverage by optimizing MS2 acquisition time focusing on sample-specific, MS1 pre-annotated, and biologically relevant compounds, thereby increasing the quality of MS2 spectra and the number of identified metabolites. The use of molecular formulas restricts the range of known and unknown chemical structures for *in silico* MS2 fragmentation tools, avoiding the loss of possible unknown isomeric forms in a sample, and facilitating *de novo* MS2 annotation. HERMES, in addition, provides maximum experimental flexibility by allowing users to add new molecular formulas not reported in public databases, including *in silico* secondary metabolism prediction<sup>24-26</sup> such as environmental microbial degradation, biotransformations of gut and soil/aquatic microbiota, or small peptides such as dipeptides and tripeptides. Finally, future developments should provide optimized maximum ion injection time and collision energies for each IL entry to reduce the number of MS2 scans required and improve the quality of MS2 spectra for all inclusion list entries, particularly for low intensity SOIs, as current Orbitrap mass spectrometers only allow fixed injection times.

## METHODS

**Materials.** LC/MS-grade acetonitrile, water, isopropanol, and methanol (Burdick & Jackson) were purchased from Honeywell (Muskegon, MI). LC/MS-grade ammonium acetate and ammonium hydroxide were purchased from Sigma-Aldrich (St. Louis, MO). TraceSELECT Fluka brand ammonium phosphate (monobasic) was purchased from Honeywell (Muskegon, MI). Dried down metabolic extracts of *E. coli* were purchased from Cambridge Isotope Laboratories (MSK-CRED-DD-KIT). Spike-in compounds (Suppl. Table 1) were purchased from Sigma-Aldrich (Zwijndrecht, The Netherlands), LGC Standards (Wesen, Germany) and Toronto Research Chemicals (Toronto, ON).

### Sample preparation

**Environmental water.** Surface water was obtained from the Lekkanaal at Nieuwegein (The Netherlands). The spike-in compounds were added to the surface water sample to a final concentration of 1 µg/L. Subsequently, the sample was filtered using Phenex™ reversed cellulose 15 mm Syringe Filters 0.2µ (Phenomenex, Torrance, USA) and transferred to a LC autosampler vial.

**E. coli.** Dried down *E. coli* extracts (unlabeled and uniformly <sup>13</sup>C-labeled) were reconstituted in 100 µL of acetonitrile:water (2:1), followed by 30 s vortexing, 5 min of sonication, and 30 s of vortexing.

**Human plasma.** Plasma aliquots (50 µL) were thawed at 4°C and briefly vortex-mixed. Proteins were precipitated by the addition of 200 µL cold acetonitrile/methanol/water (5:4:1, vol/vol) followed by 10 seconds vortex-mixing. Samples were subsequently maintained on ice for 30 min. After centrifugation (10 min, 15.200 rpm at 4°C), 100 µL of supernatant were transferred to a LC autosampler vial.

### LC-MS analysis

**Environmental water and human plasma.** Ultra-high performance LC (UHPLC)/MS was performed with a Thermo Scientific Vanquish UHPLC system interfaced with a Thermo Scientific Orbitrap Fusion Tribrid mass spectrometer operated in positive or negative ion mode. Reverse phase C18 liquid chromatography (RPLC) analysis was performed by using a Xbridge BEH C18 column (Waters, Etten-Leur, The Netherlands) with the following specifications: 150 mm x 2.1 mm, 2.5 µm. Mobile-phase solvents were composed of A = ultrapure water with 0.05% formic acid (v/v) and B = acetonitrile with 0.05% formic acid (v/v). The column compartment was maintained at 25 °C for all experiments. The following linear gradient was applied at a flow rate of 250 µL/min: 0-1 min: 5% B, 1-25 min: 5-100% B, 25-29 min: 100% B, 29.0-29.5 min 5% B followed by 4.5 min of re-equilibration phase. One µL of the

human plasma extract was diluted in 100  $\mu$ L of ultrapure water, and the injection volume was 100  $\mu$ L for all experiments. Data were collected with the following settings: spray voltage, 3.0 kV and -2.5 kV in positive and negative mode, respectively; sheath gas, 40; auxiliary gas, 10; sweep gas, 5; ion transfer tube temperature, 300  $^{\circ}$ C; vaporizer temperature, 300  $^{\circ}$ C; mass range, 80-1000 Da; RF lens, 50%; resolution, 120,000 (MS1), 15,000 (MS/MS); AGC target, 2e5 (MS1), 5e4 (MS2); maximum injection time, 100 ms (MS1), 50 ms (HERMES), 50 ms (DDA); isolation window, 1.6 Da. The collision energy was 35% for HCD fragmentation. With every batch run, mass calibration was performed using Pierce ESI positive and negative ion calibration solution in order to obtain a mass error of <2 ppm.

**E. coli.** LC/MS was performed with a Thermo Scientific Vanquish Horizon UHPLC system interfaced with a Thermo Scientific Orbitrap ID-X Tribrid Mass Spectrometer (Waltham, MA). Hydrophilic interaction liquid chromatography (HILIC) analysis was performed by using a SeQuant ZIC-pHILIC column (Merck Millipore, Burlington, MA) with the following specifications: 150 mm x 2.1 mm, 5  $\mu$ m. Mobile-phase solvents were composed of A = 20 mM ammonium bicarbonate, 0.1% ammonium hydroxide solution (25% ammonia in water) and 2.5  $\mu$ M medronic acid in water:acetonitrile (95:5) and B = 95% acetonitrile, 5% water, 2.5  $\mu$ M medronic acid. The column compartment was maintained at 40  $^{\circ}$ C for all experiments. The following linear gradient was applied at a flow rate of 250  $\mu$ L min<sup>-1</sup>: 0-1 min: 90% B, 1-12 min: 90-35% B, 12.5-14.5 min: 25% B, 15 min: 90% B followed by 4 min of re-equilibration phase at 400  $\mu$ L min<sup>-1</sup> and 2 min at 250  $\mu$ L min<sup>-1</sup>. The injection volume was 2  $\mu$ L for all experiments. Data were collected with the following settings: spray voltage, 3.5 kV and -2.8 kV in positive and negative mode, respectively; sheath gas, 50; auxiliary gas, 10; sweep gas, 1; ion transfer tube temperature, 300  $^{\circ}$ C; vaporizer temperature, 200  $^{\circ}$ C; mass range, 70-1000 Da; RF lens, 60%; resolution, 120,000 (MS1), 15,000 (MS/MS); AGC target, 2e5 (MS1), 5e4 (MS2); maximum injection time, 200 ms (MS1), 35 ms (HERMES, unless otherwise stated), 100 ms (iterative DDA); isolation window, 1 Da. The collision energy was 35% for HCD fragmentation.

### **Iterative DDA**

**E. coli.** After the first DDA run, the raw data file containing MS/MS spectra was converted to an .MS2 file using MS Convert<sup>27</sup>. Next, the IEomics tool<sup>28</sup> was used to generate the first exclusion list of features fragmented in the first DDA run. User inputs in the R script were RTWindow = 0.3 min, noiseCount = 25, MZWindow = 0.001. This procedure was repeated two times, which resulted in a total of three DDA data runs per polarity. The mass tolerance for exclusion lists was 5 ppm.

**Plasma.** An exclusion list of background ions was generated using the AcquireX workflow of Xcalibur data acquisition software (Thermo Fisher Scientific), by analyzing an ultrapure water sample. The exclusion list contains the exact mass, retention window and intensity (exclusion



override factor = 3) of the excluded background ions. DDA was performed for the top 6-8 most intense ions per full scan. Dynamic exclusion was used to prevent redundant acquisition of MS2 spectra for a selected precursor ion for 10 s, when two MS2 spectra were acquired within 20 s, resulting in a total of three DDA data runs per polarity. A mass tolerance of 5 ppm was used for the exclusion list and dynamic exclusion.

## HERMES algorithm

All analysis were performed using RHermes (version 0.99.0).

**MS1 data processing:** Theoretical isotopic patterns of each ionic formula were calculated by Envipat (version 2.4) and refined by RHermes, based on the predefined experimental mass resolution and mass accuracy values. Local resolution was calculated for each ionic formula as:

$$R(m) = R_{ref} \cdot \sqrt{\frac{m}{m_{ref}}}$$

Using as input a set of mzML files, SOIs were detected by RHermes using two sets of 5s bins (offset by 2.5s) and required a minimum scan density of 30% of acquired scans.

Blank subtraction was performed using an heuristic prefilter (intensity ratio sample/blank > 3) and an artificial neural network trained with >3000 manually annotated sample/blank SOI pairs. Adduct and isotopologue grouping were performed using a cosine shape similarity score and required a cosine >0.8 and >0.85, respectively.

In-source fragment (ISF) annotation was performed using an in-house MS2 database consisting of MassBankEU, MoNA, HMBD, Riken and NIST14 spectra. Low intensity spectra (<20% HCD, <20eV CID) were selected according to each SOI formula annotation. Intense fragments (>20% of maximum intensity) m/z were then queried against the SOI list. Finally, the suspected ISF SOIs elution profiles were compared to the original SOI and a cosine similarity score was calculated.

**MS2 data processing.** The program exports the IL into a csv file used to generate the MS2 acquisition method. Acquired MS2 scans were linked to each IL entry; if >5 scans were acquired, a deconvolution algorithm was applied, where fragments m/z were grouped and split with a Centwave peak picking (peakwidth = c(5,60)). A cosine shape similarity score was applied to each pair of fragment peaks to generate a similarity network. Each network was then partitioned using a greedy algorithm from *igraph* (version 1.2.4.2) and resulted in a list of deconvoluted MS2 spectra. If fewer than 5 scans were acquired, the scan with the highest TIC was selected and the fragments were filtered by intensity (> 0.5% of maximum).

## XCMS data processing

LC-MS raw data files (ESI+ and ESI- modes) were converted to open standard format mzML using Proteowizard MS-convert<sup>27</sup> and subsequently processed by HERMES and XCMS software<sup>18</sup> (version 3.8.1). XCMS settings were: `xcmsSet(method="centWave", ppm=8, peakwidth=c(1,60))`; Common data points between SOIs in HERMES and XCMS peaks were calculated by extracting the raw data points delimited by each XCMS peak ( $rt_{\min} < rt < rt_{\max}$  and  $mz_{\min} < mz < mz_{\max}$ ) and generating the set intersections using *dplyr* (version 1.0.4).

## Uniformly <sup>13</sup>C-labeled *E. coli*

Fractional contribution (FrC) was calculated using the formula:

$$FrC = \frac{\sum_{i=1}^n i \cdot M_i}{n \cdot MO_{unlabelled}}$$

Where  $M_i$  is the intensity of the <sup>13</sup>C<sub>i</sub> isotope and n is the total number of carbon atoms in the molecule.

Monoisotopic Ratio Score (MIRS) was calculated using the formula:

$$MIRS = 1 - \frac{MO_{labelled}}{MO_{unlabelled}}$$

If MIRS is smaller than zero, it is set to zero so that all points range from 0 to 1.

## Identification by MS/MS

**In-house DB.** MS/MS spectra were obtained from MassBankEU, MoNA, HMDB, Riken and NIST14 databases. All fragment m/z were discretized into 0.01Da bins. Each spectrum precursor m/z was matched against the DB spectra m/z with a 0.01Da tolerance. For the HERMES matching, the reference spectra were further filtered according to the formula database used in the MS1 analysis. A cosine similarity score was calculated between the query and reference spectra and resulting hits were filtered by requiring a score > 0.8.

**mzCloud DB.** The processed HERMES MS2 spectra were exported to the mzML file format. The DDA files were directly imported through MassFrontier version 8.0 SR1 (Thermo Scientific) and matched against the mzCloud database using three component identification types: Identity, Similarity Forward and Similarity Reverse; with the following constraints: 4.0 Tolerance Factor and Match Ion Activation Type. The resulting hits were filtered by both Match and Confidence scores (requiring a score > 90 and > 30, respectively)

Identified IL entries (Figure 5 and Supp Figure 9) were calculated as number of IL entries that resulted in a valid hit (ie. high score) against either of the two databases. For DDA, this number was calculated by matching the precursor m/z and RT of the scans to the IL and then examining if (i) any of the scans have at least one valid hit against either of the two databases and (ii) any valid hit had a molecular formula present in the HERMES formula database.

All similarity metrics were calculated using the R package *philentropy* (version 0.4.0). MS2 spectra were discretized into 0.01Da bins and their fragment intensities scaled by the sum of the intensities, so that all calculated metrics were comparable across the spectra. The query spectra (both DDA and HERMES) were matched against the previously described In-house DB. For each query, all DB hits were grouped, taking the maximum similarity (cosine and fidelity) and the lowest distance (squared chord and topsoe). Additionally, HERMES hits were restricted to compounds with formulas present in the HERMES formula database. The corresponding plots were generated using *ggplot2* (version 3.3.3).

### **Data availability**

Input mzML mass spectrometry data files and RMarkdown files are available at Zenodo with the accession number [4581662](#).

### **Code availability**

The source code of RHermes is offered to the public as a freely accessible software package under the GNU GPL, version 3 license, and is available at <https://github.com/RogerGinBer/RHermes>.

### **Acknowledgements**

We gratefully acknowledge financial support by Ministerio de Educación y Formación Profesional (Spanish Government) to R.G. (2020-COLAB-00552). O.Y. was supported by Ministerio de Economía y Competitividad (MINECO) (BFU2014-57466-P), Spanish Biomedical Research Centre in Diabetes and Associated Metabolic Disorders (CIBERDEM), an initiative of Instituto de Investigación Carlos III (ISCIII), and the European Union's Horizon 2020 program (MSCA-ITN-2015; 675610). We thank members of the Mil@b for helpful comments.

### **Author contributions**

RG and OY designed the research. RG, JC, JMB, MV and OY developed the computational method. DV and MSH performed LC-MS and MS/MS experiments. All authors applied and evaluated the method on biological samples. RG and OY wrote the manuscript, in cooperation with all authors.

### **Competing interests**

The authors declare no competing interests. A patent application for the method has been filled (P202030061).

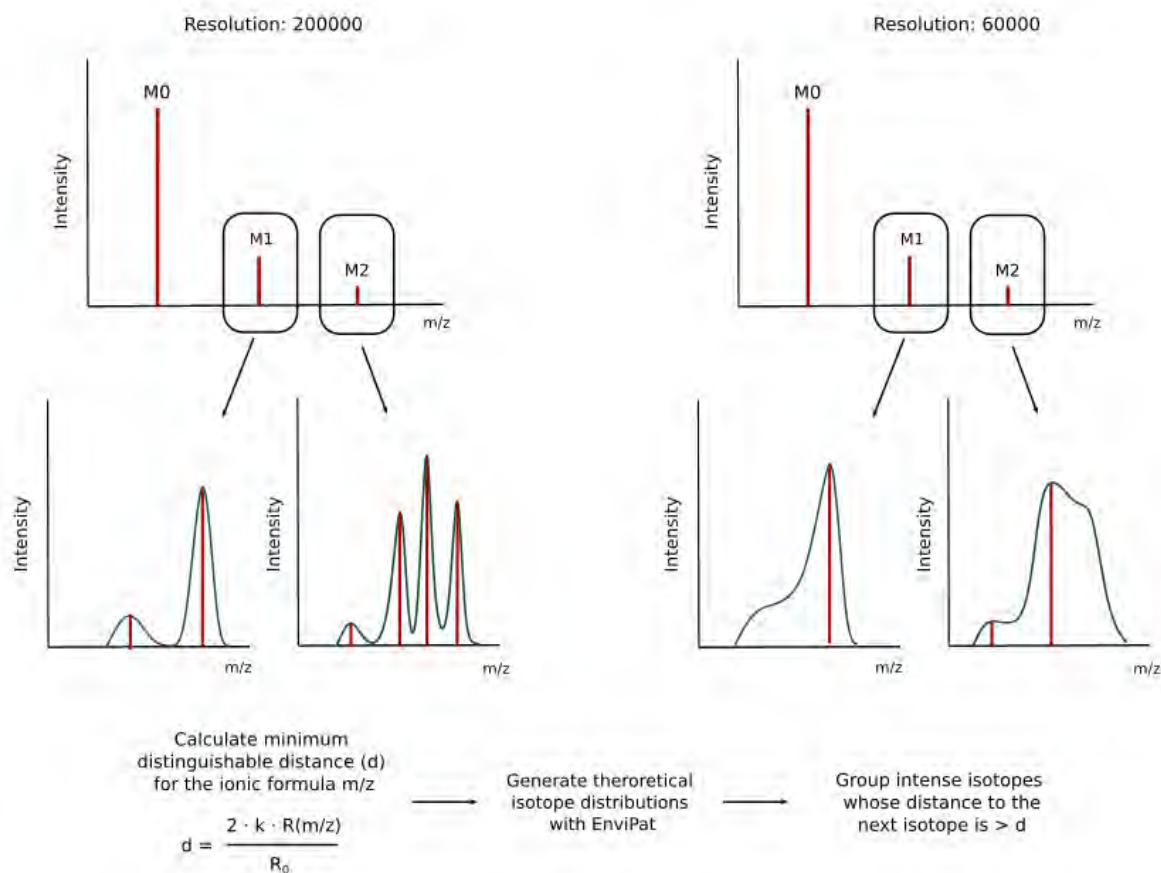
## References

1. Sindelar, M. & Patti, G. J. Chemical Discovery in the Era of Metabolomics. *J. Am. Chem. Soc.* **142**, 9097–9105 (2020).
2. Duan, L., Molnár, I., Snyder, J. H., Shen, G. & Qi, X. Discrimination and Quantification of True Biological Signals in Metabolomics Analysis Based on Liquid Chromatography-Mass Spectrometry. *Mol. Plant* **9**, 1217–1220 (2016).
3. Myers, O. D., Sumner, S. J., Li, S., Barnes, S. & Du, X. Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. *Anal. Chem.* **89**, 8689–8695 (2017).
4. Domingo-Almenara, X., Montenegro-Burke, J. R., Benton, H. P. & Siuzdak, G. Annotation: A Computational Solution for Streamlining Metabolomics Analysis. *Anal. Chem.* **90**, 480–489 (2018).
5. Tsugawa, H. *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **12**, 523–526 (2015).
6. Yin, Y., Wang, R., Cai, Y., Wang, Z. & Zhu, Z.-J. DecoMetDIA: Deconvolution of Multiplexed MS/MS Spectra for Metabolite Identification in SWATH-MS-Based Untargeted Metabolomics. *Anal. Chem.* **91**, 11897–11904 (2019).
7. Guo, J. & Huan, T. Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography–Mass Spectrometry Based Untargeted Metabolomics. *Anal. Chem.* **92**, 8072–8080 (2020).
8. Röst, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **13**, 741–748 (2016).
9. Huan, T. *et al.* Systems biology guided by XCMS Online metabolomics. *Nat. Methods* **14**, 461–462 (2017).
10. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
11. J, H. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214-9 (2015).
12. NORMAN Network *et al.* S0 | SUSDAT | Merged NORMAN Suspect List: SusDat. (2020) doi:10.5281/zenodo.4249026.

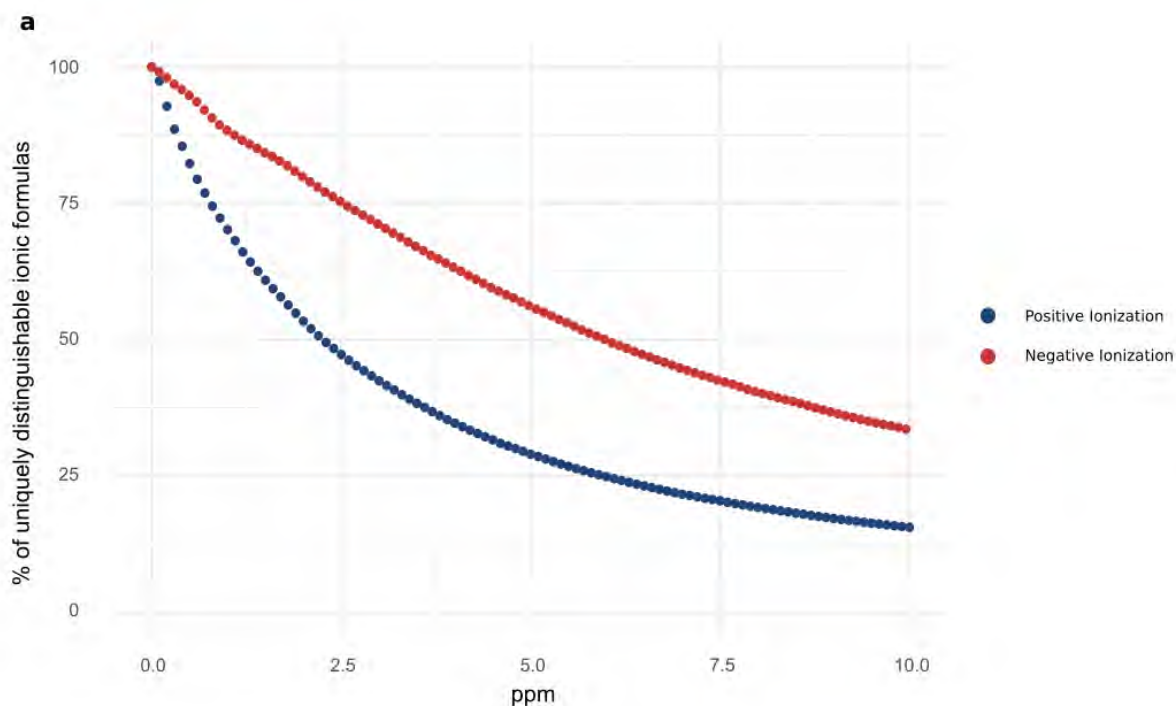
13. Domingo-Almenara, X. *et al.* Autonomous METLIN-Guided In-source Fragment Annotation for Untargeted Metabolomics. *Anal. Chem.* **91**, 3246–3253 (2019).
14. Senan, O. *et al.* CliqueMS: a computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network. *Bioinformatics* **35**, 4089–4097 (2019).
15. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
16. Dührkop, K. *et al.* Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* 1–10 (2020) doi:10.1038/s41587-020-0740-8.
17. Aron, A. T. *et al.* Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat. Protoc.* **15**, 1954–1991 (2020).
18. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **78**, 779–787 (2006).
19. Buescher, J. M. *et al.* A roadmap for interpreting <sup>13</sup>C metabolite labeling patterns from cells. *Curr. Opin. Biotechnol.* **34**, 189–201 (2015).
20. Zamboni, N., Saghatelian, A. & Patti, G. J. Defining the Metabolome: Size, Flux, and Regulation. *Mol. Cell* **58**, 699–706 (2015).
21. Jang, C., Chen, L. & Rabinowitz, J. D. Metabolomics and Isotope Tracing. *Cell* **173**, 822–837 (2018).
22. Vinaixa, M. *et al.* Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *TrAC Trends Anal. Chem.* **78**, 23–35 (2016).
23. Cho, K. *et al.* Targeting unique biological signals on the fly to improve MS/MS coverage and identification efficiency in metabolomics. *Anal. Chim. Acta* **1149**, 338210 (2021).
24. Djoumbou-Feunang, Y. *et al.* BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminformatics* **11**, 2 (2019).
25. Rutz, A. *et al.* Open Natural Products Research: Curation and Dissemination of Biological Occurrences of Chemical Structures through Wikidata. *bioRxiv* 2021.02.28.433265 (2021) doi:10.1101/2021.02.28.433265.

26. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
27. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
28. Koelmel, J. P. *et al.* Expanding Lipidome Coverage Using LC-MS/MS Data-Dependent Acquisition with Automated Exclusion List Generation. *J. Am. Soc. Mass Spectrom.* **28**, 908–917 (2017).

## Supplemental Figures

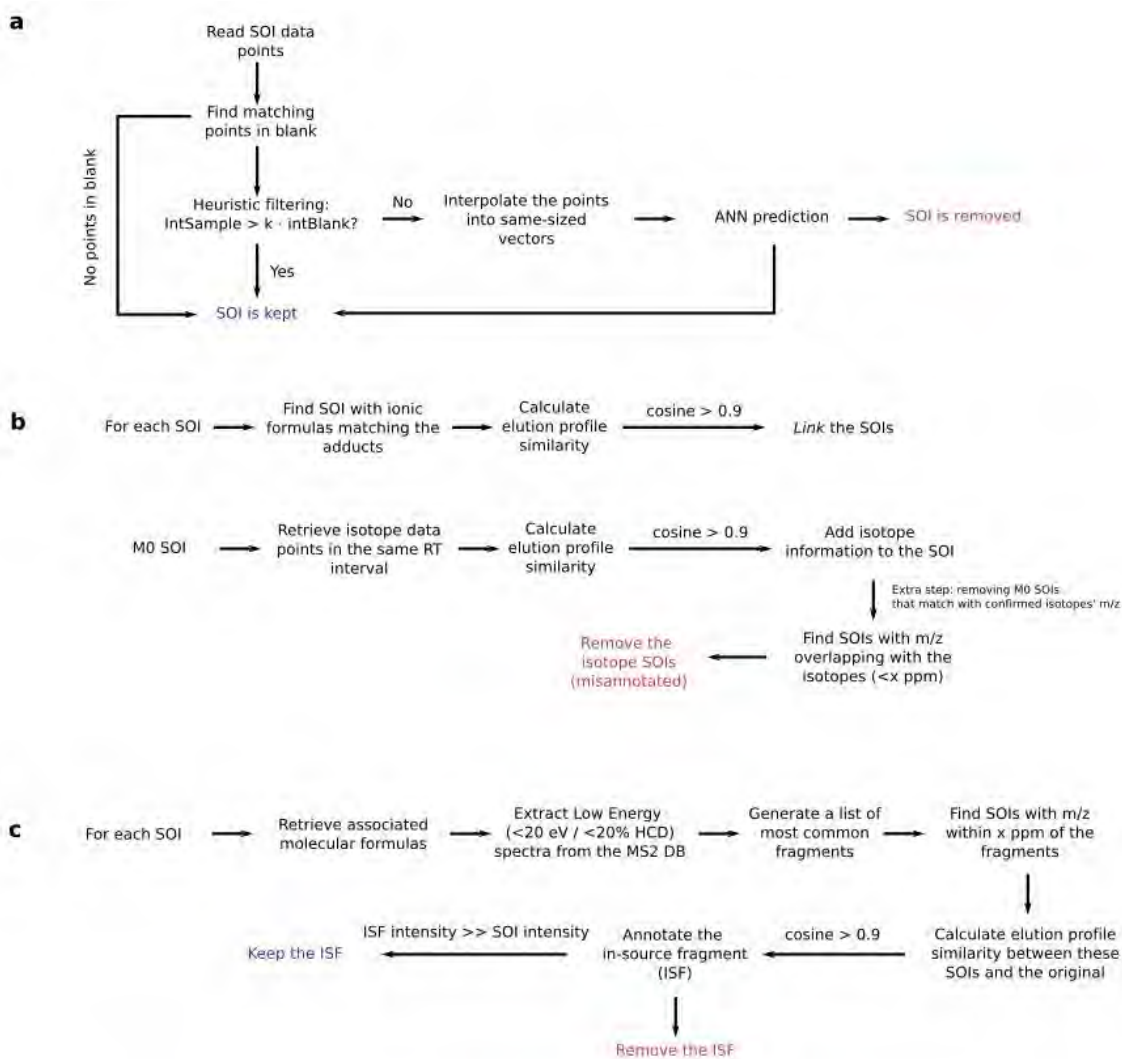


**Supplementary Figure 1. Calculation of the theoretical isotopic pattern of each ionic formula based on predefined experimental mass resolution values.** By calculating a resolution-based parameter  $d$ , it is possible to estimate which close isotopologues are likely to be distinguishable in the acquired profile MS1 data and therefore present in the centroided data.

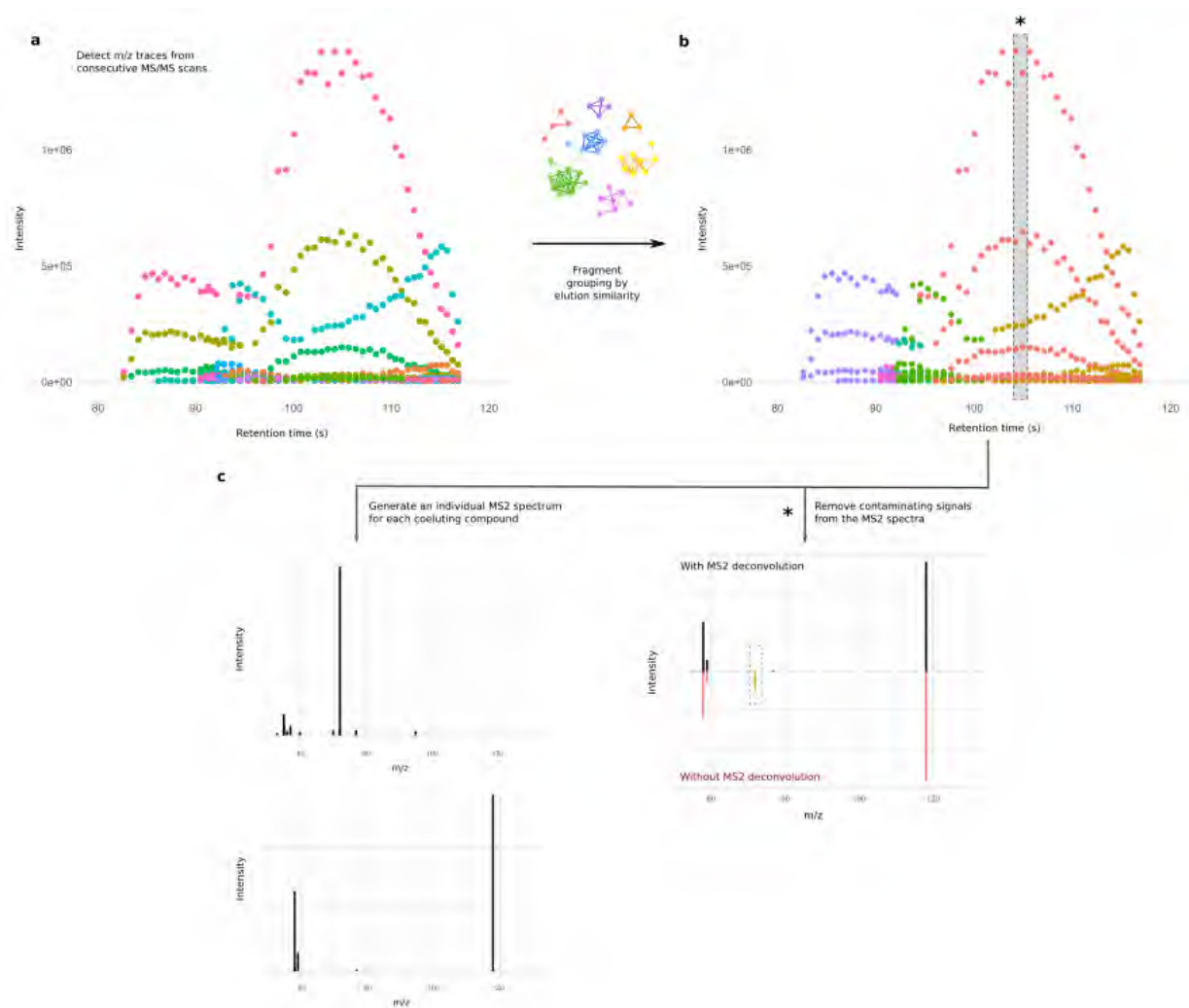


**Supplementary Figure 2:** Ionic formula collisions from the NORMAN database (24,696 unique molecular formulas). Distribution of uniquely distinguishable ionic formulas. Blue: Positive ionization taking  $[M+H]^+$ ,  $[M+Na]^+$ ,  $[M+K]^+$ ,  $[M+NH_4]^+$  and  $[M]^+$  adducts. Red: Negative ionization taking  $[M-H]^-$  and  $[M+Cl]^-$  adducts. As the ppm error of the instrument increases, the larger the percentage of overlapping ionic formulas.

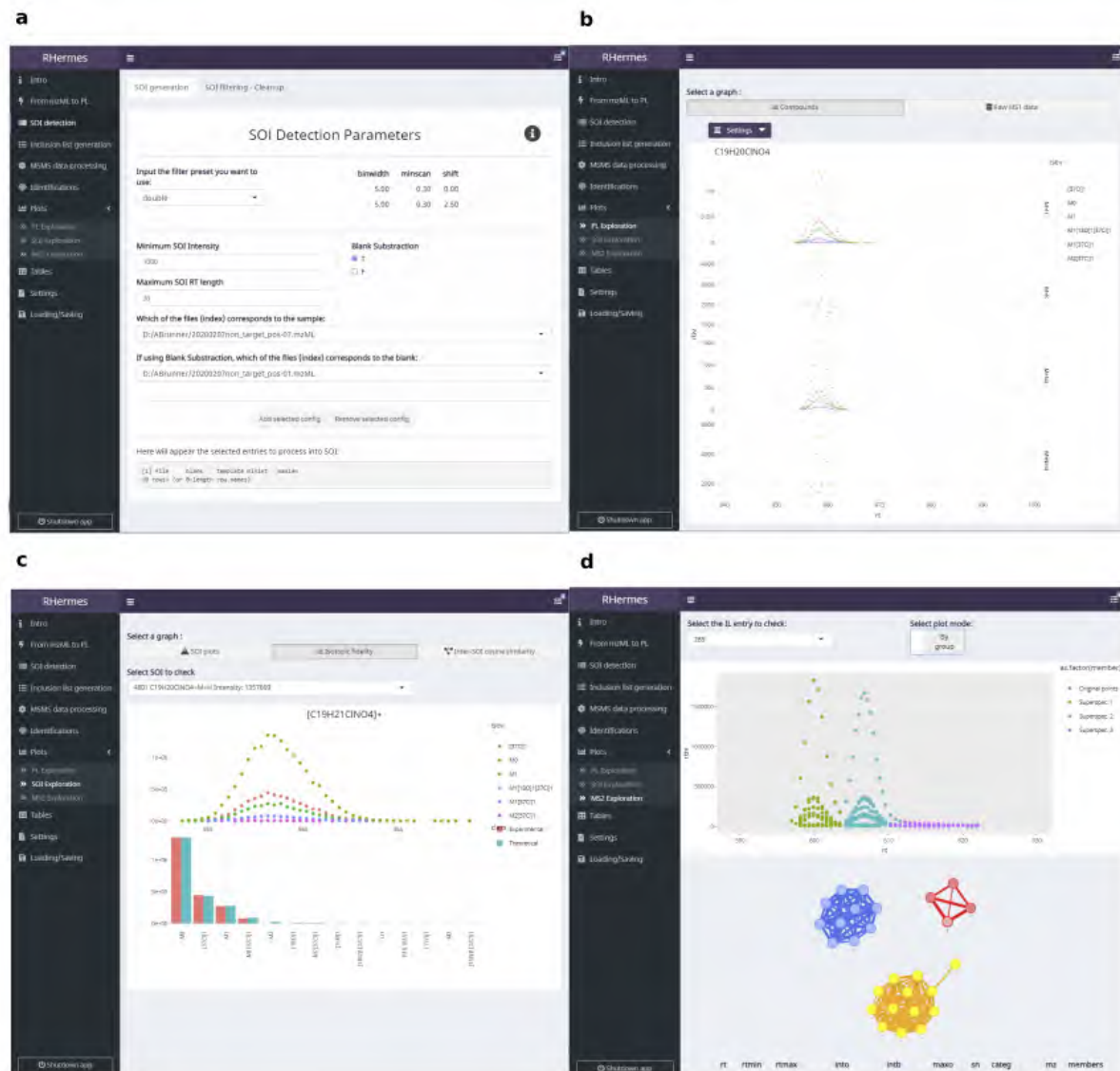




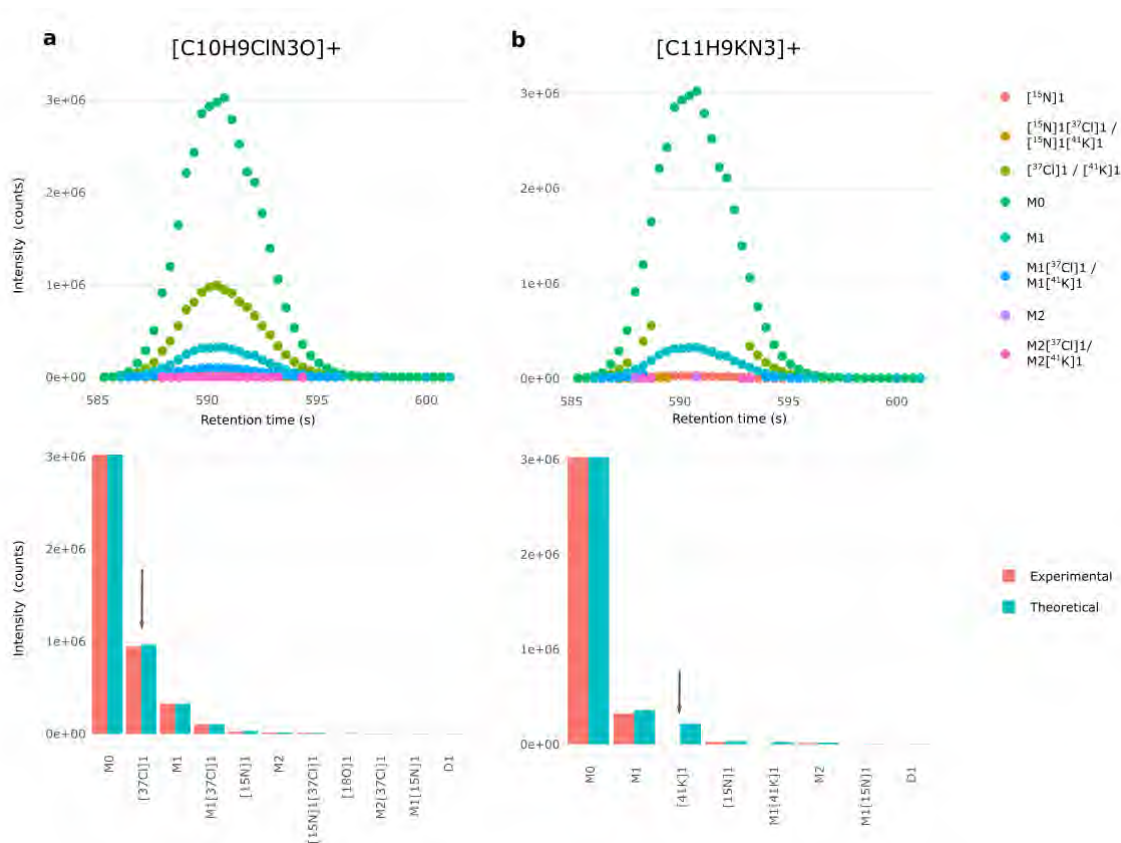
**Supplementary Figure 3. Schematic workflow of the different filtering steps in HERMES.** a) Artificial neural network (ANN) for blank subtraction. b) Adduct and isotopologue grouping according to the similarity of their elution profiles. c) In-source fragment annotation, by using publicly available low-energy MS/MS data.



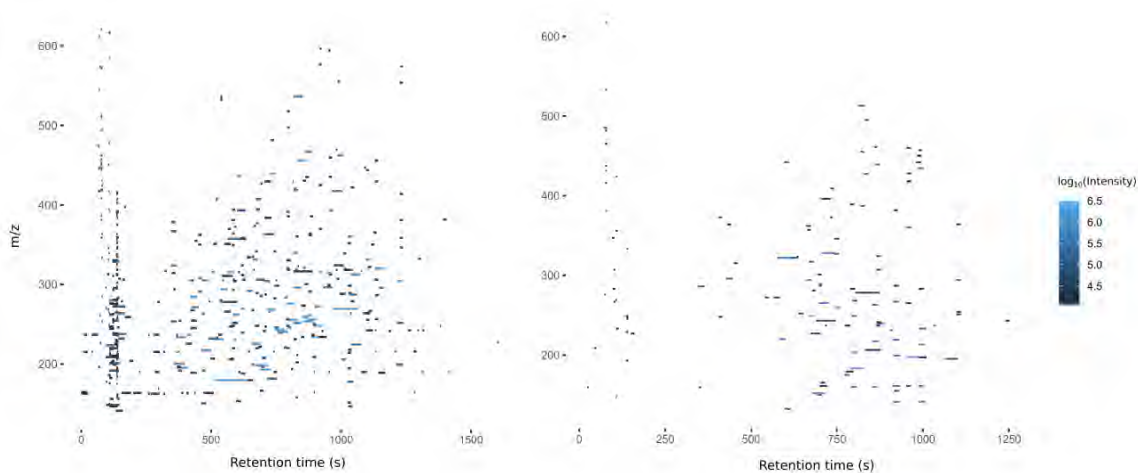
**Supplementary Figure 4. Continuous MS2 acquisition resolves co-eluting ionic species by comparing their fragment elution profile.** a) All fragment ions from continuous MS2 scans are grouped according to their  $m/z$ . b) A loose peak-picking algorithm is applied and the resulting peaks are grouped according to their elution profiles, generating a similarity network that is split by a greedy clustering algorithm. c) This grouping yields a curated MS2 spectra for each coeluting species. (\*) The shaded slice shows the impact of the algorithm on the resulting spectral quality. The delineated fragment in yellow has a different elution pattern than the rest and would contaminate the MS2 spectra if only one scan was acquired at the top of the peak. The grouping performed by HERMES confidently removes the contaminant ion and separates each group of fragments accordingly.



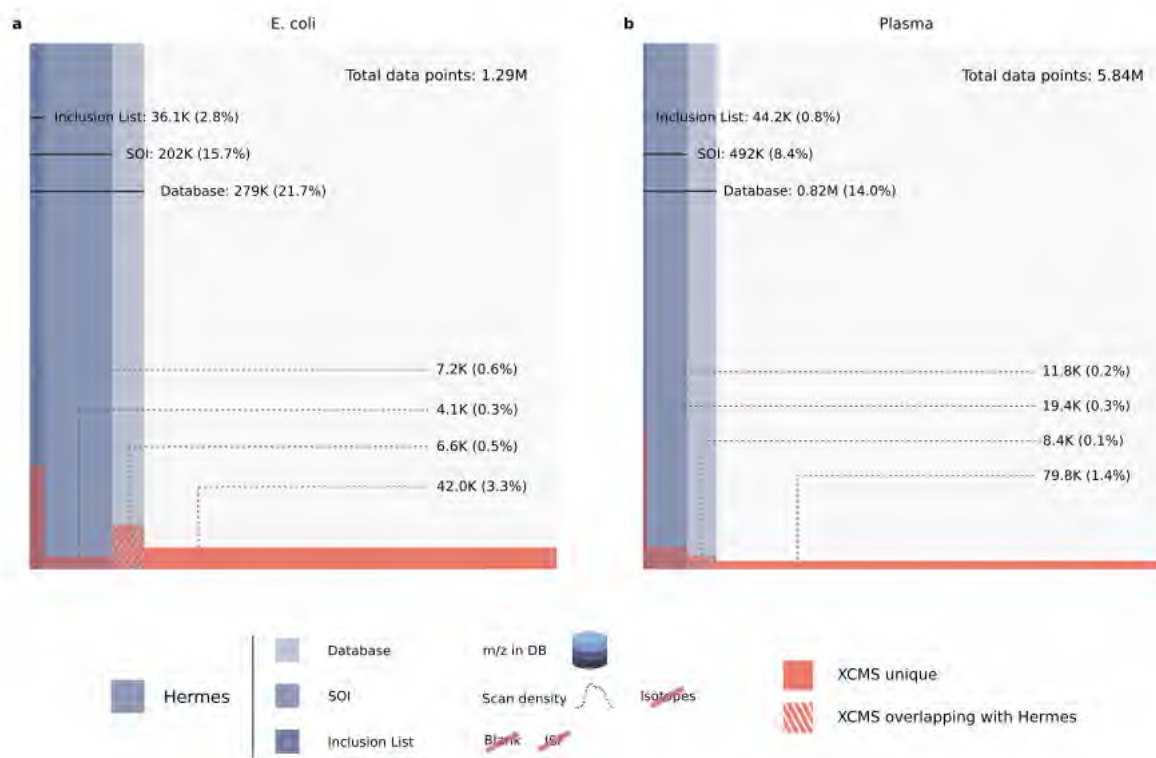
**Supplementary Figure 5. HERMES R Graphical User Interface (GUI).** a) Point-and-click selection of SOI detection parameters, with detailed explanations on their usage and optimal values. b) Visualization of isotopic profiles of different adducts of the same formula. The formula can be inputted directly or inferred from the name of a compound chosen by the user. c) Isotopic fidelity exploration of selected SOIs. d) Visualization of the continuous MS2 deconvolution step. The user can check the fragment ion elution profiles from each inclusion list entry and how they are interconnected in the corresponding profile similarity network.



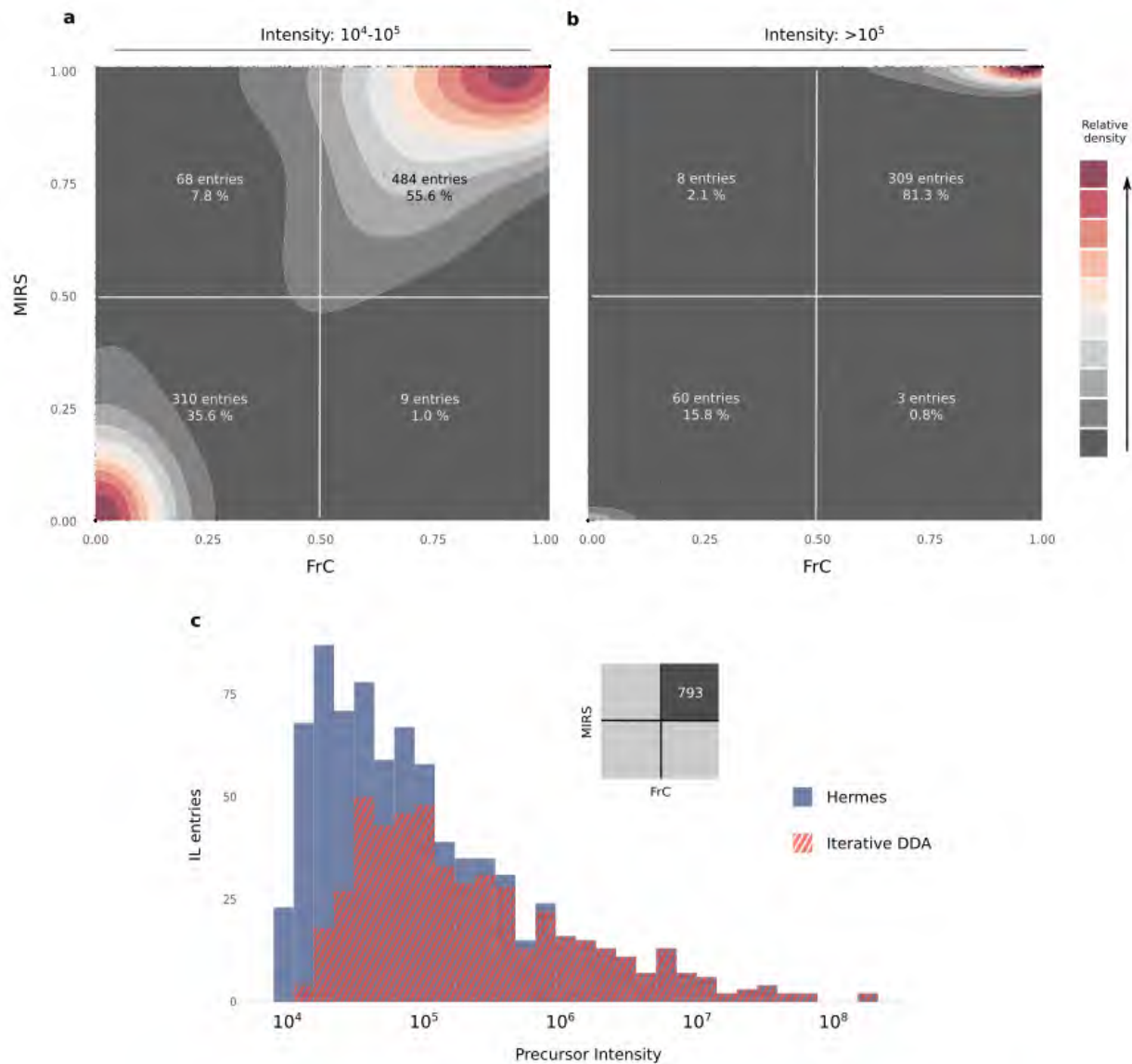
**Supplementary Figure 6. Discrimination of SOIs based on isotopic fidelity.** a)  $[M+H]^+$  ion of chloridazon and b)  $[M+K]^+$  ion of 2-Amino-alpha-carboline overlapping at 0.27 ppm. The arrows indicate the characteristic  $^{37}\text{Cl}$  isotopologue present in chloridazon and the  $[41\text{K}]$  isotopologue absent in 2-Amino-alpha-carboline. The absence of characteristic isotopologue signals (Cl, Br, K, etc.) in intense SOIs results in a low isotopic fidelity score and the removal of such SOIs.



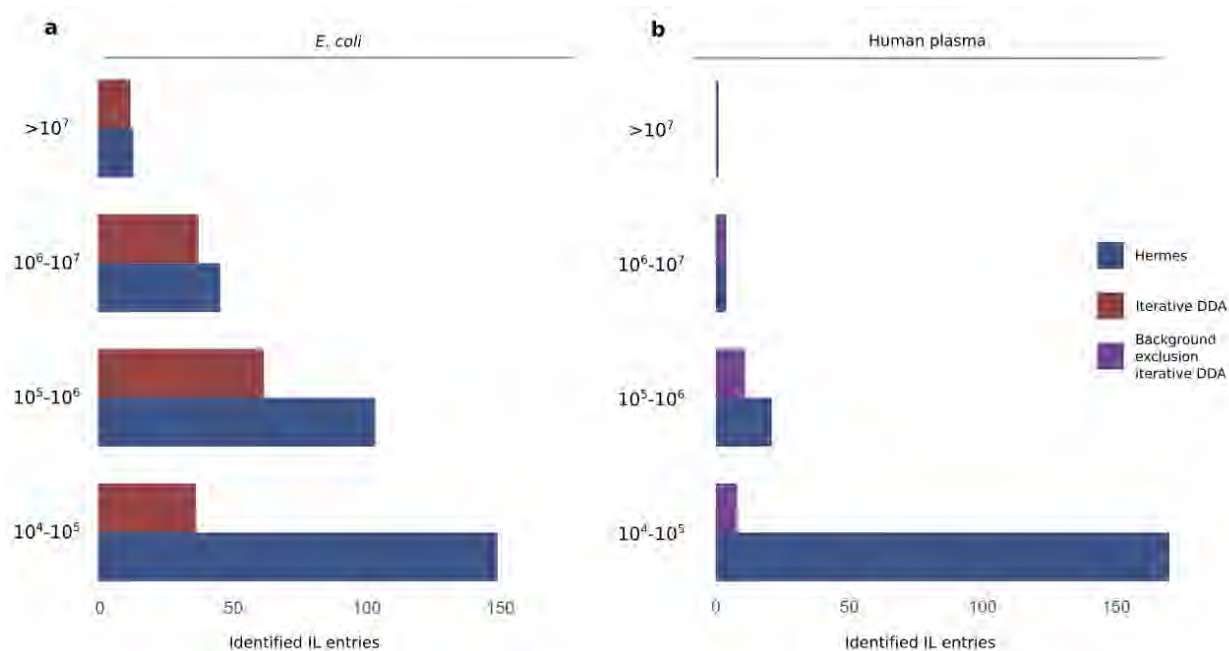
**Supplementary Figure 7.** Distribution of inclusion list entries of water in a) positive and b) negative ionization mode after blank subtraction.



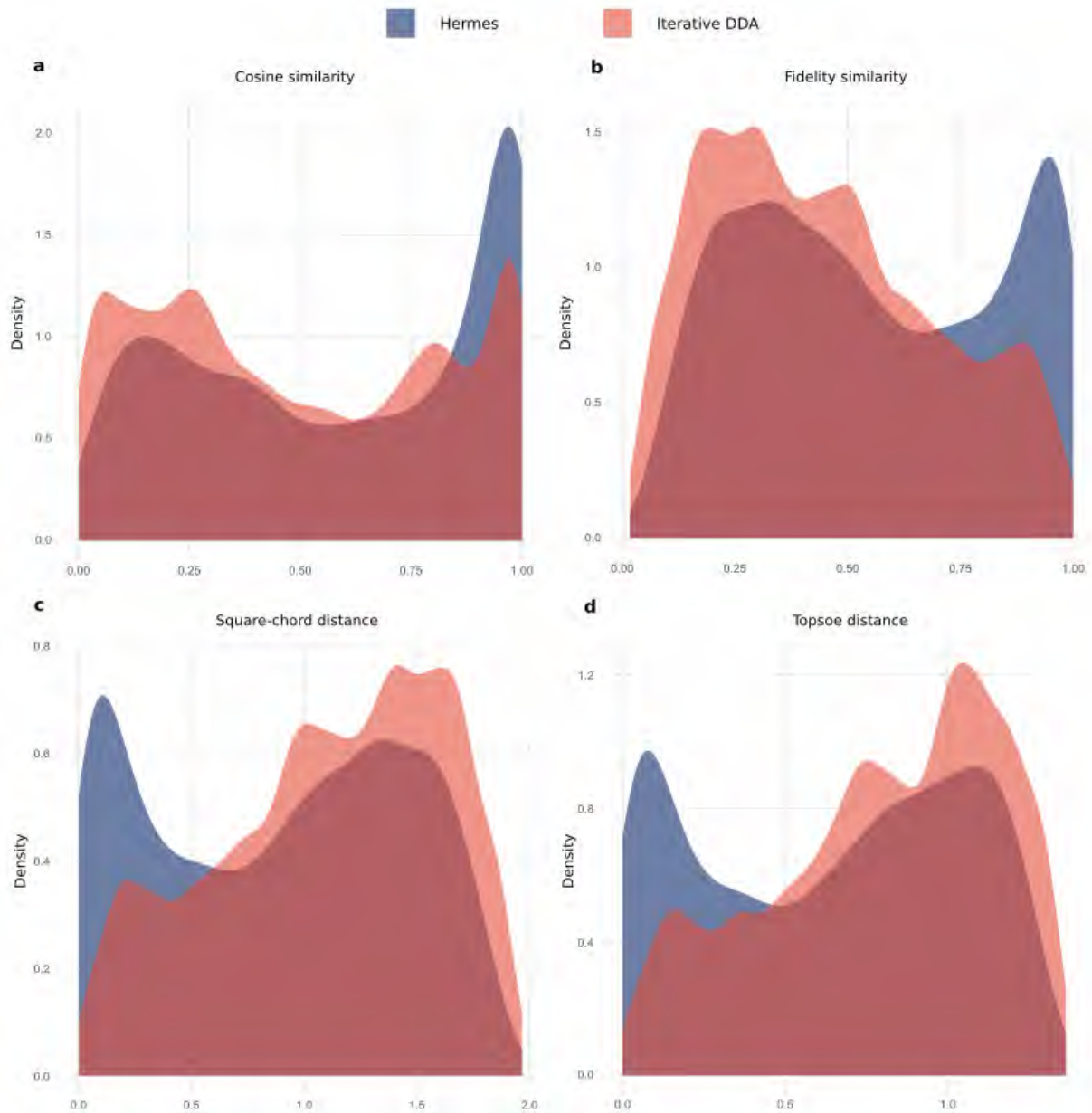
**Supplementary Figure 8. Venn-like diagram of the distribution of negative ionization LC/MS1 data points in different steps of the HERMES workflow and XCMS peak-associated points. a) *E. coli* and b) human plasma extract. Database: Refers to all data points whose m/z matches with any m/z calculated from the ionic formula database (including isotopes). SOI: monoisotopic (M0)-annotated data points that are in Database and are also present in a SOI list that does not include blank subtraction nor any filtering. Inclusion List: data points present in Database and SOI kept through the blank subtraction, isotopic filter and ISF removal steps. Percentages refer to the total number of LC/MS1 data points.**



**Supplementary Figure 9.  $^{13}\text{C}$ -enrichment distribution according to the precursor intensity.** a) and b)  $^{13}\text{C}$ -enriched metabolites (FC and MIRS  $> 0.5$ ) are mainly associated with abundant ions (intensity  $>10^5$ ), while unlabeled precursors (FC and MIRS  $< 0.5$ ) relate more frequently to low abundant ions (intensity between  $10^4$ - $10^5$ ). c)  $^{13}\text{C}$ -labeled precursors in iterative DDA corresponded to highly abundant ions that were also covered by HERMES. However, 56% of labeled low abundant ions were not covered by the iterative DDA.

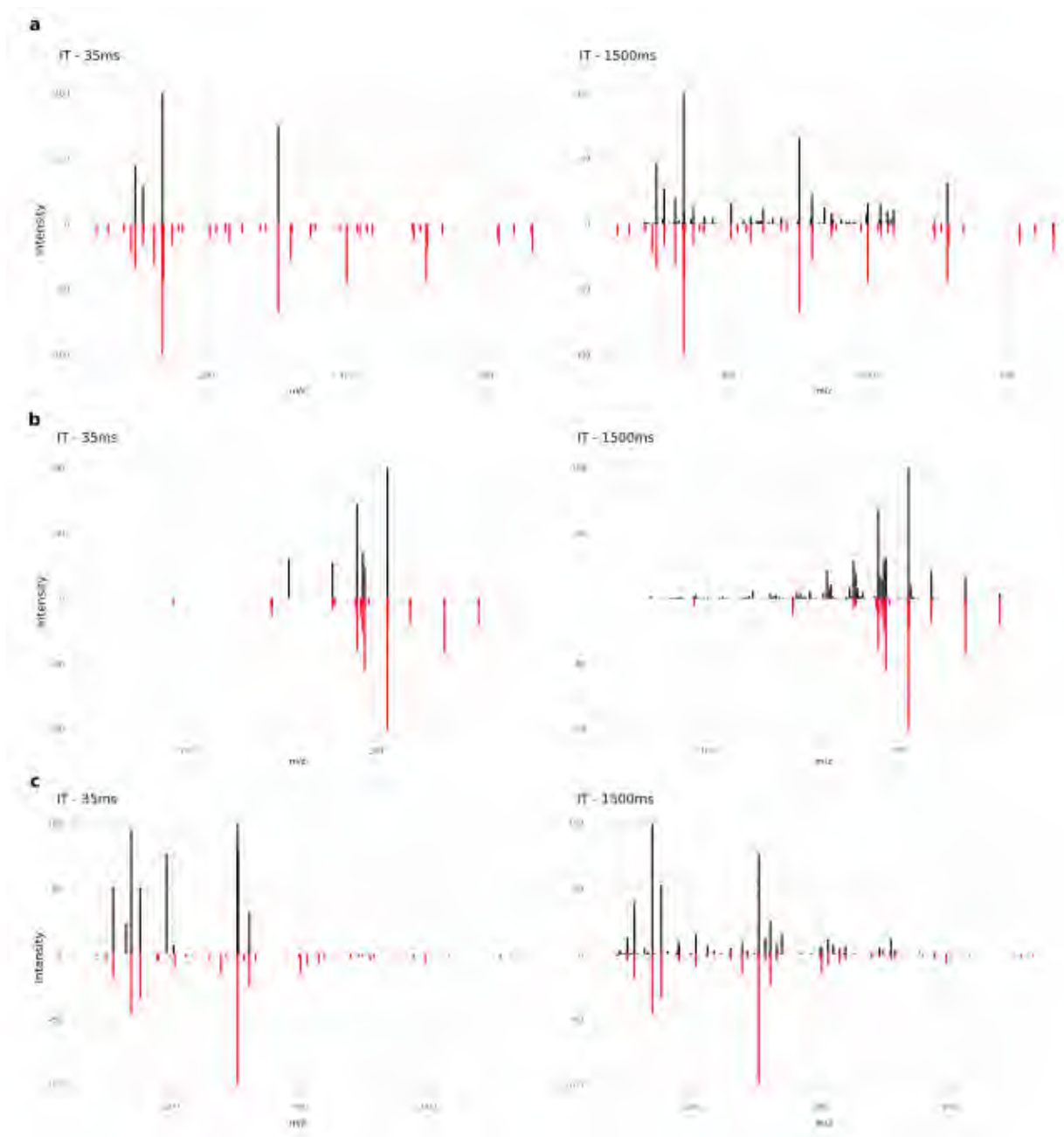


**Supplementary Figure 10. Identified IL entries according to the MS1 precursor intensity.** An inclusion list entry is considered identified if at least one MS2 scan associated with it has a compound hit in the reference MS2 database with either cosine score > 0.8 (in-house database from MassBankEU, MoNA, Riken and NIST14 spectra), or Match > 90 and Confidence > 30 (mzCloud). Negative ionization data. a) *E. coli* extract. b) Human plasma extract.



**Supplementary Figure 11. Alternative spectral similarity algorithms and spectrum-spectrum match scores.** a) Cosine similarity distribution b) Fidelity similarity distribution. c) Square-chord distance distribution. d) Topsoe distance distribution. A density estimation was calculated with ggplot2 and normalized so that the integral of the curve equals 1. HERMES spectra showed higher similarity scores (a and b) and lower spectral distances (c and d) than DDA spectra.





**Supplementary Figure 12. Injection time (IT) comparison (35 ms vs 1,500 ms).** Intensity precursor ions  $<1.0 \times 10^5$ . MS/MS (black) of a) NADH, b) Biopterin and c) NADPH against library spectra (red). A higher injection time resulted in richer spectra, with more matching fragments against the reference spectra and overall better matching scores.

**Supplementary Table 1.** List of spiked compounds

Compound name	Formula	Adduct	m/z	RT (min)
1-(3,4-dichlorophenyl)-3-methylurea	C <sub>8</sub> H <sub>8</sub> Cl <sub>2</sub> N <sub>2</sub> O	[M+H] <sup>+</sup>	219.0086	14.28
1-(3,4-Dichlorophenyl)-urea	C <sub>7</sub> H <sub>6</sub> Cl <sub>2</sub> N <sub>2</sub> O	[M+H] <sup>+</sup>	204.9930	13.29
2,4-Dichloroaniline	C <sub>6</sub> H <sub>5</sub> Cl <sub>2</sub> N	[M+H] <sup>+</sup>	161.9872	16.77
2,6-dichlorobenzamide (BAM)	C <sub>7</sub> H <sub>5</sub> Cl <sub>2</sub> NO	[M+H] <sup>+</sup>	189.9821	8.18
2-aminoacetophenone	C <sub>8</sub> H <sub>9</sub> NO	[M+H] <sup>+</sup>	136.0757	11.93
Atrazine	C <sub>8</sub> H <sub>14</sub> ClN <sub>5</sub>	[M+H] <sup>+</sup>	216.1011	14.54
Azinphos-methyl	C <sub>10</sub> H <sub>12</sub> N <sub>3</sub> O <sub>3</sub> PS <sub>2</sub>	[M+H] <sup>+</sup>	318.0131	17.17
Bezafibrate	C <sub>19</sub> H <sub>20</sub> ClNO <sub>4</sub>	[M+H] <sup>+</sup>	362.1154	15.95
Bromacil	C <sub>9</sub> H <sub>13</sub> BrN <sub>2</sub> O <sub>2</sub>	[M+H] <sup>+</sup>	261.0233	12.43
Caffeine	C <sub>8</sub> H <sub>10</sub> N <sub>4</sub> O <sub>2</sub>	[M+H] <sup>+</sup>	195.0877	6.83
Carbamazepin	C <sub>15</sub> H <sub>12</sub> N <sub>2</sub> O	[M+H] <sup>+</sup>	237.1022	13.27
Carbendazim	C <sub>9</sub> H <sub>9</sub> N <sub>3</sub> O <sub>2</sub>	[M+H] <sup>+</sup>	192.0768	6.38
Chlorpyrifos-ethyl	C <sub>9</sub> H <sub>11</sub> Cl <sub>3</sub> NO <sub>3</sub> PS	[M+H] <sup>+</sup>	349.9336	23.34
Chlortoluron	C <sub>10</sub> H <sub>13</sub> ClN <sub>2</sub> O	[M+H] <sup>+</sup>	213.0789	14.31
Chloridazon	C <sub>10</sub> H <sub>8</sub> ClN <sub>3</sub> O	[M+H] <sup>+</sup>	222.0429	9.79
Deet	C <sub>12</sub> H <sub>17</sub> NO	[M+H] <sup>+</sup>	192.1383	14.83
Desethylatrazine	C <sub>6</sub> H <sub>10</sub> ClN <sub>5</sub>	[M+H] <sup>+</sup>	188.0698	9.78
Desisopropyl Atrazine	C <sub>5</sub> H <sub>8</sub> ClN <sub>5</sub>	[M+H] <sup>+</sup>	174.0541	7.69
Diclofenac	C <sub>14</sub> H <sub>11</sub> Cl <sub>2</sub> NO <sub>2</sub>	[M+H] <sup>+</sup>	296.0240	18.37
Dimethenamid-p	C <sub>12</sub> H <sub>18</sub> ClNO <sub>2</sub> S	[M+H] <sup>+</sup>	276.0820	17.37

Dimethoate	$C_5H_{12}NO_3PS_2$	[M+H] <sup>+</sup>	230.0069	10.29
Dimethomorph (isomer 1)	$C_{21}H_{22}ClNO_4$	[M+H] <sup>+</sup>	388.1310	16.18
Dimethomorph (isomer 2)	$C_{21}H_{22}ClNO_4$	[M+H] <sup>+</sup>	388.1310	16.59
Diuron	$C_9H_{10}Cl_2N_2O$	[M+H] <sup>+</sup>	233.0243	15.07
Ethofumesate	$C_{13}H_{18}O_5S$	[M+H] <sup>+</sup>	287.0948	18.48
Phenazone	$C_{11}H_{12}N_2O$	[M+H] <sup>+</sup>	189.1022	8.66
Isoproturon	$C_{12}H_{18}N_2O$	[M+H] <sup>+</sup>	207.1492	14.93
Linuron	$C_9H_{10}Cl_2N_2O_2$	[M+H] <sup>+</sup>	249.0192	17.24
Metazachlor	$C_{14}H_{16}ClN_3O$	[M+H] <sup>+</sup>	278.1055	15.87
Metobromuron	$C_9H_{11}BrN_2O_2$	[M+H] <sup>+</sup>	259.0077	15.60
Metolachlor	$C_{15}H_{22}ClNO_2$	[M+H] <sup>+</sup>	284.1412	18.92
Metoprolol	$C_{15}H_{25}NO_3$	[M+H] <sup>+</sup>	268.1907	9.46
Metoxuron	$C_{10}H_{13}ClN_2O_2$	[M+H] <sup>+</sup>	229.0738	11.94
Metribuzin	$C_8H_{14}N_4OS$	[M+H] <sup>+</sup>	215.0961	13.19
Monuron	$C_9H_{11}ClN_2O$	[M+H] <sup>+</sup>	199.0633	12.66
Nicosulfuron	$C_{15}H_{18}N_6O_6S$	[M+H] <sup>+</sup>	411.1081	12.24
Pentoxifylline	$C_{13}H_{18}N_4O_3$	[M+H] <sup>+</sup>	279.1452	9.46
Pirimicarb	$C_{11}H_{18}N_4O_2$	[M+H] <sup>+</sup>	239.1503	9.11
Simazin	$C_7H_{12}ClN_5$	[M+H] <sup>+</sup>	202.0854	12.50
Sulfadimidine	$C_{12}H_{14}N_4O_2S$	[M+H] <sup>+</sup>	279.0910	8.38
Sulfamethoxazole	$C_{10}H_{11}N_3O_3S$	[M+H] <sup>+</sup>	254.0594	10.69

Terbutylazine	$C_9H_{16}ClN_5$	$[M+H]^+$	230.1167	16.85
Tetraglyme	$C_{10}H_{22}O_5$	$[M+H]^+$	223.1540	7.78
Triethyl Phosphate	$C_6H_{15}O_4P$	$[M+H]^+$	183.0781	10.94
Triphenylphosphine Oxide	$C_{18}H_{15}OP$	$[M+H]^+$	279.0933	15.34
Tri-n-butyl-phosphate	$C_{12}H_{27}O_4P$	$[M+H]^+$	267.1720	20.52
Tri-(2-chloroisopropyl)Phosphate	$C_9H_{18}Cl_3O_4P$	$[M+H]^+$	327.0081	17.24
Tris(2-chloroethyl)Phosphate (TCEP)	$C_6H_{12}Cl_3O_4P$	$[M+H]^+$	284.9612	14.26
2,4,6-trichlorophenol	$C_6H_3Cl_3O$	$[M+H]^-$	194.9177	17.77
2,4-dichlorophenol	$C_6H_4Cl_2O$	$[M+H]^-$	160.9566	16.52
2,4-dichlorophenoxyacetic Acid (2,4-D)	$C_8H_6Cl_2O_3$	$[M+H]^-$	218.9621	15.25
2,4-dinitrophenol	$C_6H_4N_2O_5$	$[M+H]^-$	183.0047	13.21
(4-chloro-2-methylphenoxy)Acetic Acid (MCPA)	$C_9H_9ClO_3$	$[M+H]^-$	199.0168	15.31
Bentazon	$C_{10}H_{12}N_2O_3S$	$[M+H]^-$	239.0496	14.44
Dichlorprop (2,4-DP)	$C_9H_8Cl_2O_3$	$[M+H]^-$	232.9778	16.52
Mecoprop (MCP)	$C_{10}H_{11}ClO_3$	$[M+H]^-$	213.0324	16.53
p,p-sulfonyldiphenol	$C_{12}H_{10}O_4S$	$[M+H]^-$	249.0227	11.21
N-acetyl sulfamethoxazole	$C_{12}H_{13}N_3O_4S$	$[M+H]^+$	296.0700	11.06
Metolachlor ESA	$C_{15}H_{23}NO_5S$	$[M+H]^+$	330.1370	11.24
10,11-dihydro-10,11-dihydroxy Carbamazepine	$C_{15}H_{14}N_2O_3$	$[M+H]^+$	271.1077	7.55
Gabapentin	$C_9H_{17}NO_2$	$[M+H]^+$	172.1332	6.45

Hydrochlorothiazide	$C_7H_8ClN_3O_4S_2$	[M+H] <sup>-</sup>	295.9572	7.20
Desfenylchloridazon	$C_4H_4ClN_3O$	[M+H] <sup>+</sup>	146.0116	2.25
Lamotrigine	$C_9H_7Cl_2N_5$	[M+H] <sup>+</sup>	256.0151	9.36
Metazachlor ESA	$C_{14}H_{17}N_3O_4S$	[M+H] <sup>+</sup>	324.1013	9.19
N-formyl-4-aminoantipyrine	$C_{12}H_{13}N_3O_2$	[M+H] <sup>+</sup>	232.1081	7.12
N-acetyl-4-aminoantipyrine	$C_{13}H_{15}N_3O_2$	[M+H] <sup>+</sup>	246.1237	7.08
Metazachlor OA	$C_{14}H_{15}N_3O_3$	[M+H] <sup>+</sup>	274.1186	9.32
Sitagliptin	$C_{16}H_{15}F_6N_5O$	[M+H] <sup>+</sup>	408.1254	10.27
Valsartan Acid	$C_{14}H_{10}N_4O_2$	[M+H] <sup>+</sup>	267.0877	11.78
Gabapentin-lactam	$C_9H_{15}NO$	[M+H] <sup>+</sup>	154.1226	11.22
HMMM	$C_{15}H_{30}N_6O_6$	[M+H] <sup>+</sup>	391.2300	13.24
Candesartan	$C_{24}H_{20}N_6O_3$	[M+H] <sup>+</sup>	441.1670	14.37
Irbesartan	$C_{25}H_{28}N_6O$	[M+H] <sup>+</sup>	429.2397	14.13
Valsartan	$C_{24}H_{29}N_5O_3$	[M+H] <sup>+</sup>	436.2343	16.51
Sebutylazine	$C_9H_{16}ClN_5$	[M+H] <sup>+</sup>	230.1167	16.20
Telmisartan	$C_{33}H_{30}N_4O_2$	[M+H] <sup>+</sup>	515.2442	14.07
Cetirizine	$C_{21}H_{25}ClN_2O_3$	[M+H] <sup>+</sup>	389.1627	23.33
1-H-benzotriazole	$C_6H_5N_3$	[M+H] <sup>+</sup>	120.0556	7.92
4-methyl-1H-benzotriazole	$C_7H_7N_3$	[M+H] <sup>+</sup>	134.0713	9.94
5-methyl-1H-benzotriazole	$C_7H_7N_3$	[M+H] <sup>+</sup>	134.0713	10.07
5,6-dimethyl-1H-benzotriazole	$C_8H_9N_3$	[M+H] <sup>+</sup>	148.0869	11.52

5-chloro-1H-benzotriazole	$C_6H_4ClN_3$	$[M+H]^+$	154.0167	11.30
2-aminobenzothiazole	$C_7H_6N_2S$	$[M+H]^+$	151.0324	6.43
2-hydroxybenzothiazole	$C_7H_5NOS$	$[M+H]^+$	152.0165	11.59
2-(methylthio)benzothiazole	$C_8H_7NS_2$	$[M+H]^+$	182.0093	17.38

Article

# Ultraviolet Photodissociation for Non-Target Screening-Based Identification of Organic Micro-Pollutants in Water Samples

Christian Panse <sup>1,2</sup> , Seema Sharma <sup>3</sup>, Romain Huguet <sup>3</sup>, Dennis Vughs <sup>4</sup>, Jonas Grossmann <sup>1,2</sup> and Andrea Mizzi Brunner <sup>4,\*</sup> 

<sup>1</sup> Functional Genomics Center Zurich, University of Zurich/ETH Zurich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland; cp@fgcz.ethz.ch (C.P.); jg@fgcz.ethz.ch (J.G.)

<sup>2</sup> SIB Swiss Institute of Bioinformatics, Quartier Sorge-Batiment Amphipole, 1015 Lausanne, Switzerland

<sup>3</sup> Thermo Fisher Scientific, San Jose, CA 95134, USA; seema.sharma@thermofisher.com (S.S.); romain.huguet@thermofisher.com (R.H.)

<sup>4</sup> KWR Water Research Institute, P.O. Box 1072, 3430 BB Nieuwegein, The Netherlands; dennis.vughs@kwrwater.nl

\* Correspondence: andrea.brunner@kwrwater.nl

Academic Editor: Thomas Letzel

Received: 26 August 2020; Accepted: 10 September 2020; Published: 12 September 2020



**Abstract:** Non-target screening (NTS) based on the combination of liquid chromatography coupled to high-resolution mass spectrometry has become the key method to identify organic micro-pollutants (OMPs) in water samples. However, a large number of compounds remains unidentified with current NTS approaches due to poor quality fragmentation spectra generated by suboptimal fragmentation methods. Here, the potential of the alternative fragmentation technique ultraviolet photodissociation (UVPD) to improve identification of OMPs in water samples was investigated. A diverse set of water-relevant OMPs was selected based on k-means clustering and unsupervised artificial neural networks. The selected OMPs were analyzed using an Orbitrap Fusion Lumos equipped with UVPD. Therewith, information-rich MS<sup>2</sup> fragmentation spectra of compounds that fragment poorly with higher-energy collisional dissociation (HCD) could be attained. Development of an R-based data analysis workflow and user interface facilitated the characterization and comparison of HCD and UVPD fragmentation patterns. UVPD and HCD generated both unique and common fragments, demonstrating that some fragmentation pathways are specific to the respective fragmentation method, while others seem more generic. Application of UVPD fragmentation to the analysis of surface water enabled OMP identification using existing HCD spectral libraries. However, high-throughput applications still require optimization of informatics workflows and spectral libraries tailored to UVPD.

**Keywords:** mass spectrometry; non-target screening; ultraviolet photodissociation; higher-energy collisional dissociation; organic micropollutants; water quality; small molecule fragmentation; cheminformatics; data analysis

## 1. Introduction

### 1.1. Challenges of Monitoring Drinking Water Quality

Reliable identification of organic micro-pollutants (OMPs) in drinking water and its sources is essential to risk assessment and prediction of the behavior of a substance in the environment and during water treatment. Non-target screening (NTS) based on the combination of liquid chromatography coupled to high-resolution mass spectrometry (LC-HRMS/MS) has become the key method to identify

OMPs in water samples, as it has the potential to detect all ionizable compounds that are amenable to the selected chromatographic separation, within a defined mass range [1].

The unambiguous identification of an OMP from NTS data relies on the accurate mass and isotopic pattern from the full scan MS1 spectrum to determine the elemental formula of the compound. The addition of MS2 fragmentation spectral data then allows to determine the molecular structure, given that the fragmentation event causes reproducible bond cleavages that result in diagnostic and interpretable fragment ions representative of the structure of the molecule. The MS2-based structural identification typically relies on matching of the experimental spectrum with entries in spectral libraries or *in silico* predicted fragmentation spectra. For compounds that show poor fragmentation spectra generated by higher-energy collisional dissociation (HCD) fragmentation, the fragmentation technique routinely applied in Orbitrap based NTS, confident structural elucidation often remains elusive. Alternative fragmentation techniques that cause different bond cleavages may remedy structural elucidation in these cases.

### 1.2. Interpreting Fragmentation Spectra from Ultraviolet Photodissociation

Ultraviolet photodissociation (UVPD) is a fragmentation technique achieved with a UV laser. Its main application to date is protein characterization with proteomics and intact protein MS. However, it also allows structural elucidation of small molecules that cannot be identified by HCD alone [2]. For instance, UVPD was shown to facilitate characterization of various lipid classes [3], to generate unique fragments or enhance detection of kinetically unfavorable fragments of flavonoids, phenylpropanoids and chalconoids [4,5].

In the Orbitrap Fusion Lumos mass spectrometer, a Q3 series passively Q-switched Nd:YAG laser (CryLaS GmbH) that outputs the 5th harmonic at 213 nm is interfaced to the rear of the ion trap, in the low pressure cell of which the UV photoactivation occurs [6]. The laser pulse energy is a  $1.5 \pm 0.2 \mu\text{J}/\text{pulse}$  at 2.5 kHz repetition rate. With  $450 \pm 200 \mu\text{m}$ , including the divergence at the center of the ion trap, the beam diameter is slightly larger than the simulated ion cloud diameter at normal AGC targets and no focusing optics are required. With the incorporation of this compact and robust solid state laser into a commercial instrument, UVPD could be routinely implemented in NTS workflows.

To date, however, UVPD fragmentation has not been applied to the NTS-based monitoring of small molecules. This is due to the fact that apart from the compounds mentioned above, little is known about UVPD fragmentation pathways of small molecules, the kinetics of fragment formation, and the influence of reaction time on fragmentation patterns. Moreover, as it is a novel fragmentation technique in the small molecule field, spectral library entries with UVPD spectra are still lacking, and the usability of HCD spectra for spectral matching of UVPD spectra remains to be explored. Furthermore, the aptness of *in silico* prediction algorithms for the prediction of UVPD spectra has not yet been demonstrated. Combinatorial prediction algorithms that do not apply any rules of fragmentation, but use a bond dissociation approach, could potentially be used for UVPD data.

### 1.3. UVPD Fragmentation for Water Quality Monitoring

Here, the potential of the fragmentation technique UVPD to improve the structural identification of small molecules, in particular OMPs in water samples, was evaluated using the Orbitrap Fusion Lumos Tribrid [6]. After application of a cheminformatics strategy to select water relevant OMPs that cover a wide chemical space and development of a data analysis workflow in R, HCD and UVPD fragmentation patterns of selected OMPs could be characterized and compared. The two fragmentation techniques generated both unique and common fragments, demonstrating that some fragmentation pathways are specific to the respective fragmentation method whilst others seem more generic. Application to environmental water samples showed that HCD spectral libraries can be used for UVPD based OMP identification, in particular when high collision energy (CE) spectra are



available. However, to increase successful feature annotation of NTS data with UVPD fragmentation, UVPD spectra need to be added to spectral libraries.

## 2. Results and Discussion

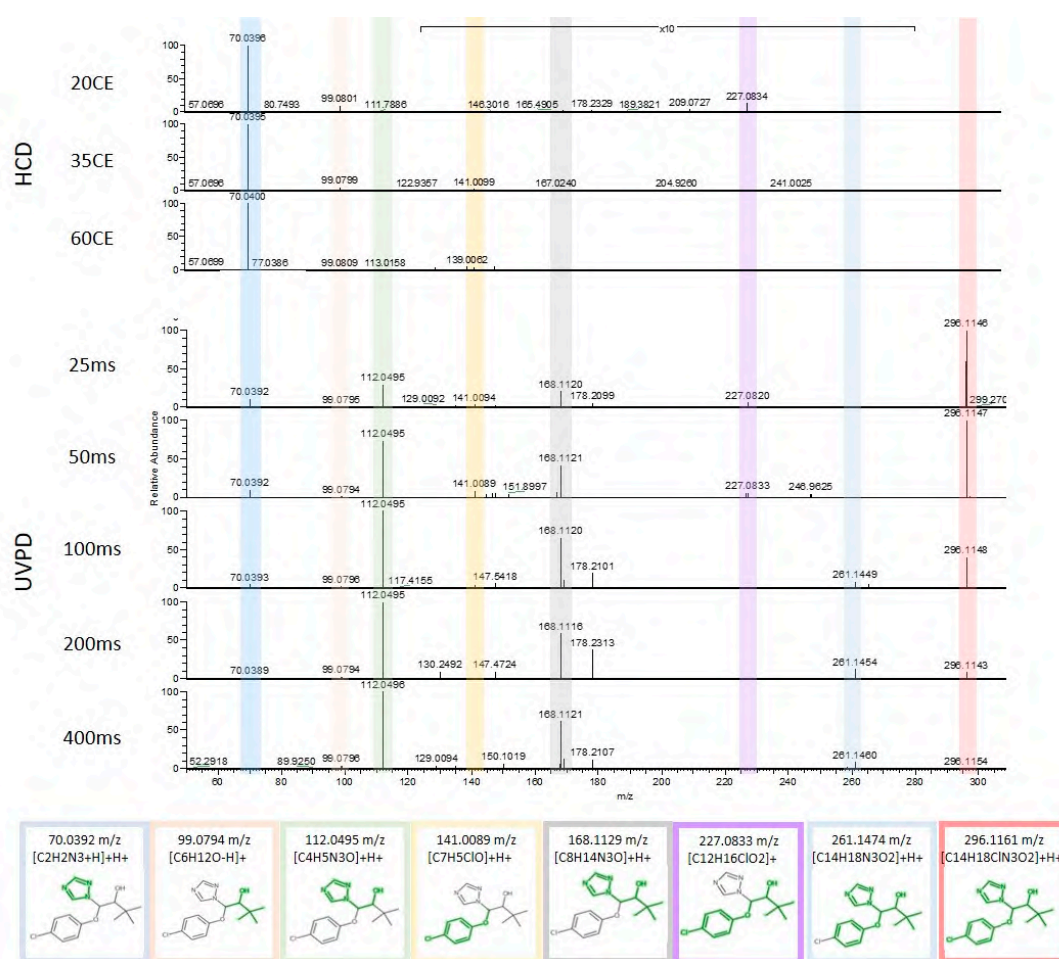
### 2.1. Proof of Principle: Manual Spectral Interpretation of Single Compounds

To investigate the potential of UVPD for OMP identification, the three compounds triadimenol, gemfibrozil and sucralose were selected as model compounds. These compounds are both relevant for the water sector and known to not fragment well using standard HCD fragmentation settings, i.e., CEs ranging from 20 to 50 (arbitrary units). As little was known about UVPD fragmentation pathways of these OMPs, we applied the combinatorial prediction algorithm of MetFrag [7], which does not rely on fragmentation rules, but uses a bond dissociation approach to predict potential fragments and matches these to the experimentally observed.

UVPD provided unique fragmentation information for structural elucidation of triadimenol, a fungicide that can be found in drinking water sources (Figure 1). HCD fragmentation at 35CE, the CE range typically used in NTS experiments, resulted in a predominant fragment at 70  $m/z$ , a minor fragment at 99  $m/z$  and a low intensity fragment (see 10 $\times$  zoom-in) at 141  $m/z$ . These fragments could be assigned to the *in silico* predicted fragments  $[C_2H_2N_3 + H] + H^+$ ,  $[C_6H_{12}O-H]^+$  and  $[C_7H_5ClO] + H^+$ . In contrast, UVPD fragmentation led to more and different fragment ions. The peaks detected with HCD were also detected in the UVPD fragmentation spectra when using shorter reaction times (25 to 100 ms), but decreased with increasing UVPD reaction times. Concurrently, peaks at 112, 168, and 261  $m/z$  increased with increasing reaction times. These could be matched to the *in silico* predicted fragments  $[C_4H_5N_3O] + H^+$ ,  $[C_8H_{14}N_3O] + H^+$ , and  $[C_{14}H_{18}N_3O_2] + H^+$ . A fragment at 227  $m/z$  was detected with UVPD at short reaction times, i.e., 25 to 50ms only and could be matched to  $[C_{12}H_{16}ClO_2]^+$ . These promising results showed that UVPD could lead to informative spectral information of an OMP that did not fragment well with HCD, and that the *in silico* prediction using MetFrag could successfully be applied for UVPD spectral annotation.

Next, UVPD fragmentation of OMPs that ionize in negative ionization mode were investigated, starting with the pharmaceutical gemfibrozil. Two peaks could be annotated in the HCD spectrum with 35CE and UVPD spectrum with 25ms reaction time, namely  $[C_8H_9O]^-$  at 121  $m/z$ , and  $[C_7H_{13}O_2-H]^-$  at 127  $m/z$  (see Figure S1a). In the UVPD spectra with longer reaction times, only the 121  $m/z$  peak could be matched. The predominant UVPD peak at 112  $m/z$  increased with reaction times. This peak was also present in the HCD spectrum. However, it could not be matched to an *in silico* predicted fragment mass, nor could any of the other UVPD peaks. The base peak in all UVPD spectra was the precursor ion, the absolute intensity of which decreased with increasing reaction times. As observed previously with UVPD of lower charged negative DNA ions, this could be due to an electron detachment-induced charge reduction [2]. Electron detachment dissociation usually involves two or more negatively charged species. For single ion negative UVPD, the mechanism for electron detachment may be more favorable than fragment generation. However, these data were based on 193 nm UVPD, and whether 213 nm UVPD would have the same effects remains unknown.

As a second OMP analyzed in negative ionization mode, the artificial sweetener sucralose was fragmented. With the standard HCD fragmentation with 35CE, six peaks could be annotated (see SI Figure 1b). Four of the annotations, i.e.,  $[C_2H_4O_2]^-$  at 59  $m/z$ ,  $[C_3H_5O_2-H]^-$  at 71  $m/z$ ,  $[C_3H_5O_2]^-$  at 73  $m/z$  and  $[C_3H_5O_3-H]^-$  at 87  $m/z$ , were low mass range fragments only detected using HCD fragmentation. The other two,  $[C_4H_7O_3-H]^-$  at 101  $m/z$  and  $[C_6H_{10}O_4-2H]^-$  at 143  $m/z$ , were present in both HCD and UVPD spectra with 50, 100, 200 ms and—in the latter—400 ms reaction time. At 25 ms, the species,  $[C_6H_{10}O_4-H]^-$  at 144  $m/z$ , was present instead. At 400 ms reaction time, the 143  $m/z$  peak was the only one that could be annotated in an overall noisy low intensity fragmentation spectrum. Corresponding to what was observed for triadimenol, long UVPD reaction times negatively affected spectral quality as well as the duty cycle in the case of sucralose.



**Figure 1.** Comparison of higher-energy collisional dissociation (HCD) (top) and ultraviolet photodissociation (UVPD) (bottom) fragmentation spectra of the fungicide triadimenol acquired with 20, 35 and 60 collision energy (CE) and 25, 50, 100, 200 and 400 ms reaction time. Annotated fragments are highlighted in color. The  $m/z$  range 130–280 is 10× enlarged.

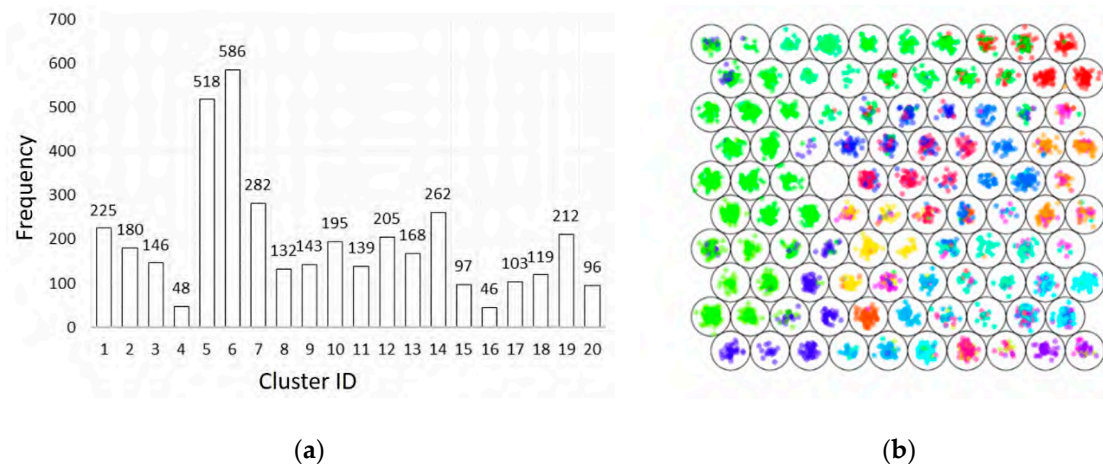
In the spectra of all shorter reaction times, i.e., 25 to 200 ms, the fragments [C<sub>6</sub>H<sub>10</sub>O<sub>5</sub>-2H]-H<sup>-</sup> at 159  $m/z$  and [C<sub>6</sub>H<sub>9</sub>Cl<sub>2</sub>O<sub>3</sub>]-H<sup>-</sup> at 197  $m/z$  could be annotated; at lower reaction times also [C<sub>6</sub>H<sub>9</sub>Cl<sub>2</sub>O<sub>3</sub>]<sup>-</sup> at 198  $m/z$  and [C<sub>6</sub>H<sub>9</sub>Cl<sub>2</sub>O<sub>3</sub>-H]-H<sup>-</sup> at 196  $m/z$ . At 50 ms, another annotated fragment, [C<sub>6</sub>H<sub>10</sub>ClO<sub>5</sub>-H]-H<sup>-</sup> at 195  $m/z$ , was present. This reaction time resulted, thus, in the most informative spectra, in particular in combination with the HCD spectrum of low mass range annotated fragments.

The UVPD fragmentation data of the three model compounds of which one ionized in positive and two in negative ionization mode suggested that UVPD could facilitate structural elucidation of some OMPs for which HCD spectra did not contain enough information. While fragmentation of the positively charged triadimenol led to more fragments and higher fragment intensities with UVPD compared to HCD, the negatively charged compound gemfibrozil fragmented poorly with both HCD and UVPD, and UVPD fragmentation of sucralose resulted in complementary fragments to HCD.

## 2.2. Selection of Reference Standards Based on Clustering

To further investigate the applicability of UVPD for OMP identification, a representative selection of compounds regarding their distribution in the chemical space and water relevance was made. First, a k-means clustering was performed using Pubchem extended fingerprints, resulting in 20 defined clusters (Figure 2a). The molecular discrimination of these clusters was confirmed using the unsupervised artificial neural network self-organizing maps (SOM, Figure 2b). In the SOM, the selected

compounds were colored according to their k-means cluster number. As compounds of the same color are in close vicinity in the SOM, this shows that the different clusters successfully separated these compounds.



**Figure 2.** Selection of organic micro-pollutants (OMPs) by two complementary approaches: (a) k-means clustering of Pubchem extended fingerprints; (b) self-organizing map. Compounds are colored according to the k-means cluster number.

Depending on in-house reference standard availability, one to four compounds were selected per cluster for fragmentation experiments, apart from cluster 19, for which no standard was available. In addition, a selection of disinfection by-products known to be relevant for drinking water treatment was added to the set of compounds, to be fragmented by HCD and UVPD.

### 2.3. An R-Based Data Analysis Workflow and Shiny Application Interface to Explore (Novel) Fragmentation Techniques

To enable high throughput data analysis of the LC-HRMS data including UVPD fragmentation spectra, an R-based workflow was developed that takes the extracted ion chromatogram (XIC) of a given compounds based on its simplified molecular-input line-entry specification (SMILES), determines the peak apex and extracts the corresponding retention time (RT). The three MS2 spectra with highest intensity neighboring the apex were used to match experimental spectra with *in silico* predicted fragmentation spectra of the given SMILES.

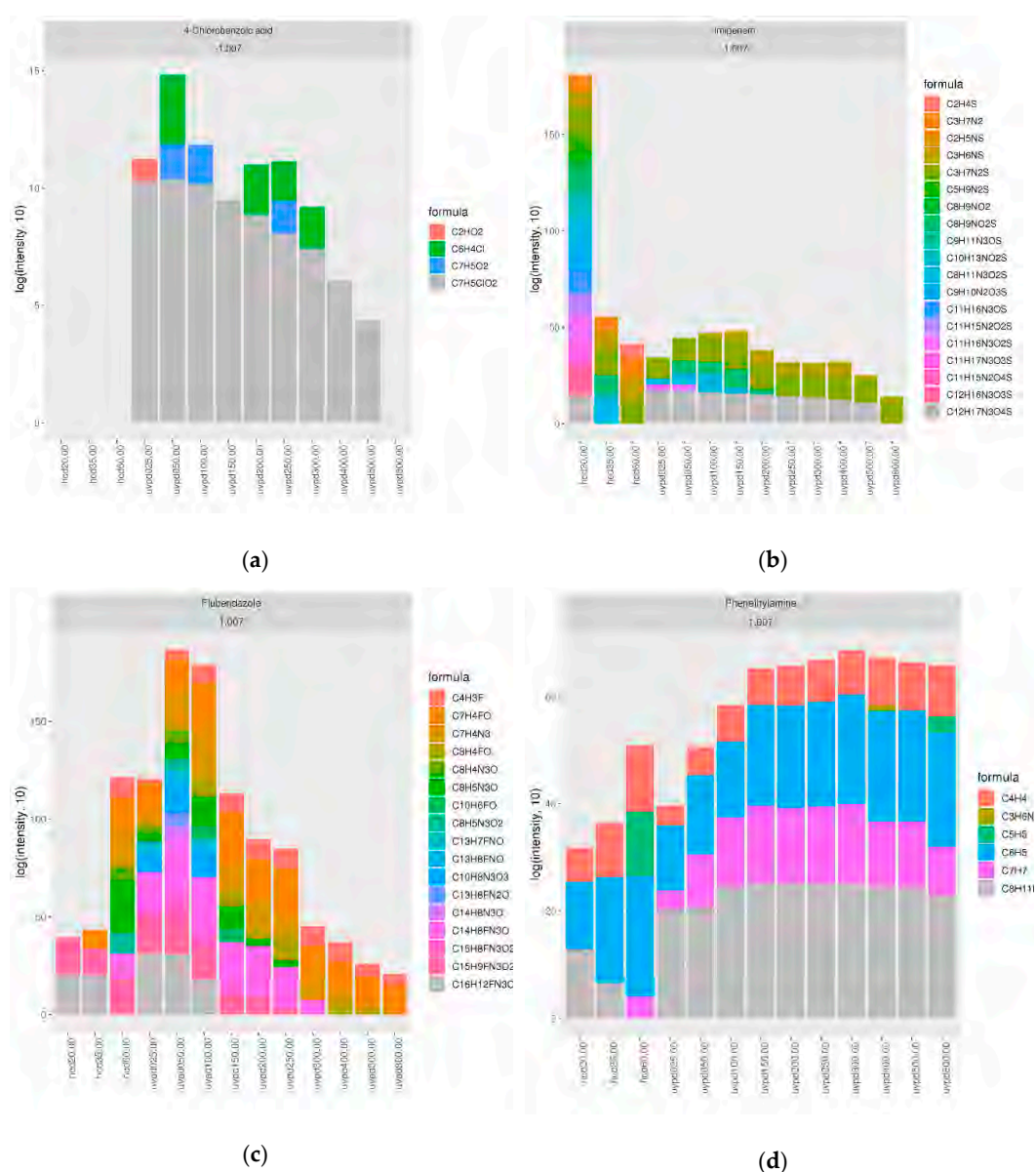
For a user friendly output and to support exploratory data analysis [8], a Shiny application based interface was created to further examine the data (<https://CRAN.R-project.org/package=shiny>). The Shiny application is provided with the 'uvpd' R package (<https://github.com/cpanse/uvpd/>) and can be accessed at <http://fgcz-ms-shiny.uzh.ch:8080/p2722-uvpd/>. Its user interface is split into an input and output part. On the left, the input panel provides a selection for the input data, compounds, ionization mode and cut-off values for the relative and absolute mass errors of the precursor mass, precursor signal removal in the MS2 spectra and cluster ID. This selection then determines the respective output in the several tab panels on the right. These tabs provide data visualization and tables of the selected compound. Table 1 describes the output tabs and whether the selected filtering is applied to a given tab.

**Table 1.** Description of Shiny application output panels and applied filtering parameters.

Tab Panel	Description	Selected Filter Option Applied to the Tab				
		Compound	Remove Precursor Items	(+/-) Ion Type	Ppm Error Cut-Off	Absolute Error Cut-Off
stacked fragments	1) Bivariate scatterplots of scores 1, 2 and 3 per fragmentation mode.					
	2) Two stacked bar charts of the logarithmically transformed fragment ion intensities of the matched fragment ions and types, respectively, per fragmentation mode.					
	3) Bivariate scatterplots of the total ion count (TIC) of the MS2 spectrum and the corresponding master intensity for the three most abundant master intensities of each raw file per fragmentation mode.	X	X	X	X	X
	4) Boxplots of the absolute error distribution (in Dalton) per fragmentation mode.					
summary	1) Statistics of the overall data and the applied filter setting.					
	2) Frequency value per fragmentation mode.					
	3) Histograms of ppm and absolute error distribution over the entire data set and selected compound, including a maximum-likelihood fitting, assuming an underlying normal distribution.	X	X	X	X	X
ms2	1) Table of detected fragment ions and ion types.	X	X	X	X	X
	2) Fragmentation spectra per fragmentation mode.					
data	All quantitative and qualitative data.	X	X	X	X	X
scores	1) Scores 1, 2 and 3.					
	2) Plots of the scores.		X	X	X	X
frequencies	Downloadable frequency table, per compound and fragmentation type		X	automatic	X	X
predicted ion	<i>In silico</i> predicted, i.e., theoretical fragment ions predicted with 'metfRag: frag.generateFragments'	X				
help	Help page					

#### 2.4. Higher-Throughput Comparison and Interpretation of UVPD and HCD Fragmentation Spectra

For a thorough comparison of UVPD and HCD fragmentation spectra, 46 selected water-relevant OMPs covering a wide chemical space were analyzed with LC-HRMS using UVPD with 25–800 ms reaction time, and HCD with 20, 35 and 60 CE. Eight of the compounds could not be detected with electrospray ionization (ESI), one eluted too early with reverse-phase (RP) LC for peak detection, one had an intensity below the cut-off threshold and two were not picked by the data analysis workflow due to a Na-adduct and Cl-salt, respectively (see Table S1). The remaining 34 compounds belonged to 11 different clusters, with one to four compounds per cluster. Their fragmentation behavior varied substantially and, based on fragmentation, four different groups of compounds could be distinguished (Table S1); such with poor fragmentation with both UVPD and HCD (Figure 3a), a preference for HCD (Figure 3b) or UVPD (Figure 3c), or good fragmentation with both UVPD and HCD (Figure 3d). These groups, however, did not seem related to the cluster number.



**Figure 3.** Fragmentation spectra annotation of the different ion types per fragmentation condition. Stacked bar plots from Shiny application output showing the summed intensities of the annotated fragments of (a) 4-chlorobenzoic acid; (b) imipenem; (c) flubendazole; (d) phenetylamine.

The first group of compounds did not generate information-rich spectra, as illustrated in Figure 3a, which shows the summed intensities of the annotated fragments from HCD and UVPD spectra of 4-chlorobenzoic acid; poor fragmentation was observed for ibuprofen, 3-nitrophenol and 4-nitrophenol with both fragmentation techniques under all conditions, for 4-chlorobenzoic acid, 4-nitrophthalic acid and 5-nitrosophthalic acid in particular with HCD. In the case of 2-methyl-4-nitrophenol and gemfibrozil, spectra were still poor, but slightly better with HCD.

A second group of compounds showed a preference for HCD compared to UVPD (Figure 3b) either with increasing CEs, for instance 3-nitroindole, with low CEs, for instance imipenem at 20CE, with a specific CE, for instance 2-methyl-4-chlorophenoxyacetic acid (MCPA) at 35CE, or with all CEs, for instance N-Desmethyl Clarithromycin. In contrast, for another group of compounds, no fragmentation was observed with HCD, but good fragmentation with a range of UVPD reaction times (Figure 3c), for instance for benzocaine and to a lesser degree 4-nitroanthranilic acid. In the case of fenofibric acid, flubendazole and triadimenol good information rich spectra were generated with a range of UVPD reaction times, but only with HCD at 60 CE.

A fourth group of compounds exhibited good fragmentation with both fragmentation techniques (Figure 3d). Some of these had an optimum at a specific fragmentation condition, for instance 2-amino-3-nitrobenzoic acid with UVPD at 50 ms, 3,5-Dinitrosalicylic acid with HCD at 35CE and UVPD at 50 ms, aflatoxin B2, dinoterb, epoxiconazole and JWH-250 with HCD at 60CE and 2-hydroxy-4-nitro-benzoic acid at 20 CE. Others fragmented well with a range of conditions, such as 2,4 Dinitrophenol, which showed more informative spectra with higher CEs, and 2-methoxy-4-nitrophenol and phenethylamine with higher HCD CEs and longer UVPD reaction times. Good fragmentation was observed for both (ranges of) UVPD and HCD conditions in the case of 2-methoxy-4,6-dinitrophenol, 4-hydroxy-3-nitrobenzenesulfonic acid, 4-nitrobenzenesulfonic acid and 5-nitrovanillin. In the case of nitrofurazone, more informative spectra were generated with UVPD.

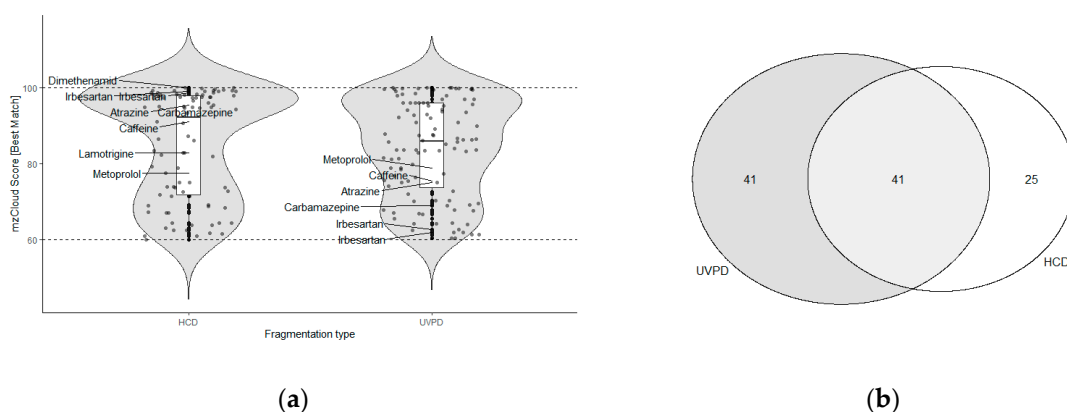
Cluster numbers could not be related to these four broad groups of fragmentation behavior. For instance, while benzocaine and 4-Nitroanthranilic acid both fragmented well with UVPD and poorly with HCD, the other two compounds from cluster 13, piperacillin and 5-Nitrosophthalic acid, showed good fragmentation for both UVPD and HCD and poor fragmentation, respectively. Regarding cluster 11, aflatoxin b2 and fenofibric acid both exhibited information rich spectra at multiple different UVPD reaction times. However, gemfibrozil, a compound of the same cluster, did not fragment well with both UVPD and HCD. This lack of similar fragmentation behavior within a cluster could indicate that fragmentation behavior depends only on a few of the descriptors used for clustering.

In particular, UV absorbing compounds such as aromatic compounds, and compounds with double bonds are expected to fragment well with UVPD. Compound class information for each of the compounds that fragmented well with UVPD, including the lipids [3], flavonoids, phenylpropanoids and chalconoids [4,5] published previously could be utilized to predict UVPD fragmentation. Furthermore, if certain compounds of classes with good UV absorbance did not fragment well, further clustering within that compound class could be utilized to improve our understanding and ultimately the prediction of UVPD fragmentation behavior.

In the UVPD spectra, fragment ion intensities decreased with increasing UVPD reaction times when normalized to precursor intensity, as illustrated in Figure S2. Overall, UVPD fragmentation was beneficial for multiple compounds, often leading to a number of annotated fragments that were unique to the fragmentation technique. This complementarity of UVPD makes it an attractive addition to HCD that can be implemented in data-dependent decision trees during NTS data acquisition. Optimal UVPD reaction times depended on the compound, analogous to HCD where the optimal CE varied amongst compounds. Interestingly, in the HCD experiments, oftentimes, CEs higher (60 CE) and lower (20 CE) than the 35 CE routinely used in NTS experiments were needed to generate informative fragmentation spectra. This should be considered in future studies to increase the confidence of OMP identification, in particular when UVPD is not available.

### 2.5. NTS of a Meuse River Surface Water Sample

To investigate the applicability of currently available spectral libraries and NTS workflows to UVPD data, a surface water sample from the river Meuse was acquired with HCD and UVPD fragmentation and analyzed using the NTS data analysis software Compound Discoverer (Thermo Fisher Scientific, San Jose, USA). This software enables suspect screening based on spectral matching with the spectral library mzCloud, which consists of collision-induced dissociation (CID) and HCD fragmentation spectra. The mzCloud score of a tentatively identified compound is a measure for the confidence of identification. It is based on the number of fragments that match the experimental and library spectra, with a score of 100 indicating a perfect match. Comparison of mzCloud scores of spectra acquired with HCD and UVPD showed that the overall score distribution was similar (no significant difference, see Appendix A), visualized in the combined box and violin plots in Figure 4a. However, individual compounds differed strongly in their scores. For instance, known water relevant OMPs on average showed a mzCloud score with UVPD fragmentation that was 15 points lower than the HCD score; atrazine scored 75.1 with UVPD versus 95.4 with HCD, caffeine 75.5 versus 91, carbamazepine 69.1 versus 96.6 and terbuthylazine 87.4 versus 98.6. UVPD spectra were matched with HCD library spectra of high CEs, i.e., 70 CE, 40 CE, 80 CE and 90 CE for the four different OMPs. In contrast, metoprolol showed a similar mzCloud score with UVPD, i.e., 78.9, when matched with an HCD 30 CE spectrum, compared to an HCD score of 77.6.



**Figure 4.** Comparison of HCD and UVPD NTS MS2 data of river Meuse water (a) mzCloud Score distribution. The scores of common water relevant compounds are labelled by compound name; (b) overlap in features annotations with an mzCloud score above 60.

Half of the compound annotations with mzCloud in the NTS data with UVPD were not assigned in the HCD data (Figure 4b). The HCD data was manually checked for features with accurate mass and retention time matching these unique compounds. Information on whether these features were detected in the HCD data and their annotation is available in Table S2. Twenty-five features were detected in the HCD data, but had no mzCloud hit, and six were annotated with a different compound based on the mzCloud matching. Ten features were not detected. This is most likely due to differences in peak picking during data analysis. Manual inspection of the UVPD assignments showed that in most cases when there was no assignment in the UVPD HCD data, the annotated UVPD spectrum consisted of only a precursor signal and—if any—low intensity ions close to the noise cut-off (see SI Table S2 Compounds detected in UVPD NTS experiments). In these cases, the match was usually against a low energy CID or HCD library spectrum (CE10 to 20) that also only contained the precursor. Consequently, these can likely be false positive assignments. The high scores for matches based on the precursor signal alone are problematic. In future studies, more appropriate scoring algorithms should be considered.

In contrast, the assignments where multiple fragments were matched, i.e., 1-methylbenzotriazole with 15, 3-hydroxyfluorene with 11, acetyl norfentanyl with three, mandipropamid with seven and metolachlor with 17 fragments, were all based on high energy HCD library spectra (60 to 130 CE)

except for 3-hydroxfluorene (20 CE). This was in correspondence with the assignments of known water relevant OMPs, and indicates that UVPD-induced fragmentation pathways in these molecules resemble those of higher energy HCD. As routinely lower HCD CEs are applied, i.e., 20–50 CE, this can explain why these assignments were only made in the case of UVPD fragmentation, emphasizing the benefit of this alternative fragmentation technique and/or higher CEs for NTS based identification of OMPs. Moreover, UVPD annotations could be used to exclude false positive annotations of sparse HCD MS2 spectra (precursor only matches) and vice versa. While HCD spectral libraries proved to be of (limited) use for UVPD spectral annotation, for the routine implementation of UVPD data in NTS workflows, spectral libraries need to be extended with UVPD spectra.

### 3. Materials and Methods

#### 3.1. Selection of Reference Standards Through Clustering

Selection criteria for OMPs to be fragmented with UVPD and HCD included their relevance for the water sector and a good coverage of the chemical space, as the compound structures were expected to affect fragmentation. To select water relevant OMPs, 4000 compounds were randomly selected from the NORMAN Substance Database, which is compiled of multiple suspect lists relevant for environmental monitoring (SusDat, <https://www.norman-network.com/nds/susdat/>). To select compounds with diverse chemical structures, these compounds were clustered [9]. To this end, the Simplified molecular-input line-entry systems (SMILES) of each compound were parsed and configured for atom typing and isotoping using the R package rcdk [10]. Next, for each compound, the extended fingerprint, a binary vector of 1024 dimensions, was extracted. A k-means clustering was conducted of the computed Tanimoto Distance matrix between all pairs of fingerprints [11]. The optimal number of clusters was determined by the elbow method [12]. To investigate molecular discrimination by the clusters, we trained a self-organizing map (SOM) [13] as a complementary approach. The SOM grid was initialized with 10 × 10 nodes. Each fingerprint selected in the training phase was colored by the corresponding k-means cluster ID for visualization. The entire *in silico* data analysis was performed using R version 3.5.1 to 4.0.1 running on Linux, Windows and MacOSX systems [14]. All code snippets are available as an R package through <https://github.com/cpanse/uvpd/>.

#### 3.2. LC-HRMS Analysis with UVPD Fragmentation

Selected reference compounds listed in SI Table S2 were prepared in ultra-pure water with a final concentration of 10 µg/L. The surface water (SW) sample was collected from the river Meuse, the Netherlands, 16.666× concentrated using Oasis-HLB SPE columns-based extraction and diluted 50× for the LC-HRMS analysis. The internal standards (IS) atrazine-d5 (CDN isotopes, Pointe-Claire, Quebec, Canada), benzotriazole-d4 and bentazon-d6 (LGC Standards, Wesen, Germany) were added to the SW sample to a final concentration of 1 µg/L. Samples were filtered using 0.2 µm Phenex™-RC 15 mm Syringe Filters (Phenomenex, Torrance, USA) prior to analysis. Blank samples were prepared correspondingly, through spike-in of IS to ultra-pure water followed by filtration. In total, 100 µL of sample were injected into the LC-HRMS.

Compounds were analyzed using reverse phase (RP) LC-HRMS/MS with a Vanquish Horizon UHPLC system (Thermo Fisher Scientific, San Jose, CA, USA) coupled to an Orbitrap Fusion Lumos equipped with ultraviolet photodissociation (UVPD) and the acquisition software AcquireX (Thermo Fisher Scientific, San Jose, CA, USA). An XBridge BEH C18 XP column (150 mm × 2.1 mm I.D., particle size 2.5 µm, Waters, Etten-Leur, The Netherlands) was used in combination with a 2.0 mm × 2.1 mm I.D. Phenomenex SecurityGuard Ultra column (Phenomenex, Torrance, CA, USA), at a temperature of 25 °C. The LC gradient started with 5% acetonitrile, 95% water and 0.05% formic acid (*v/v/v*), increased to 100% acetonitrile, 0.05% formic acid in 25 min and then remained constant for 4 min. The flow rate was 0.25 mL/min.



For the reference standards, fragmentation spectra were acquired using targeted methods with mass triggers. The fixed collision energies (CEs) 20, 35 and 60 were used for HCD fragmentation, and UVPD reaction times ranging from 25 to 800 ms for UVPD fragmentation. The full scan mass range was 100–800  $m/z$  with 120k resolution at FWHM for the MS1 scans, and 50–500  $m/z$  with 15k resolution at FWHM for the MS2 scans (due to a corrupted data file, the disinfection by-products data with 100 ms and 250 ms UVPD reaction time is lacking in the data set).

For the SW sample, NTS analyses were performed with data dependent acquisition (ddA), using the AcquireX deepscan functionality that ensures MS2 scans are acquired for most features, Top Speed and 35 CE for the HCD and 100 and 150 ms reaction time for the UVPD experiments. The full scan mass range was set at 80–1300  $m/z$  with 120k resolution, the MS2 at 50–500  $m/z$  with 15k resolution.

### 3.3. Manual Annotation of Fragmentation Spectra

Thermo Fisher Scientific raw files were viewed with Thermo Xcalibur Browser (Thermo Fisher Scientific, San Jose, CA, USA). MS2 peak lists of HCD and UVPD fragmentation spectra were exported and used for fragment annotation with the MetFrag web tool (<https://msbi.ipb-halle.de/MetFragBeta/>).

### 3.4. An R-Based LC-HRMS Data Analysis Workflow to Explore Novel Fragmentation Techniques

Fragment ions of the selected reference standard compounds were predicted with tree depth 1 and 2, using the R package MetFrag [7] in a preprocessing step. Charge configurations were derived for the predicted singly charged fragments  $[M]^+$ ,  $[M + H]^+$  and  $[M + 2H]^+$  and  $[M]^-$ ,  $[M - H]^-$  and  $[M - 2H]^-$  for the positive and negative ionization mode, respectively. The predicted fragment ions were stored and made available as a dataset in the R package UVPD.

Thermo Fisher Scientific raw files were processed with the R package rawDiag [15]. The in profile mode recorded data were centroided using the centroid method of the R package protViz (<https://CRAN.R-project.org/package=protViz>, [16]). For all compounds measured in all fragmentation modes, the retention time (RT), the area under the curve (AUC) of the APEX extracted ion chromatography (XIC) of the protonated and deprotonated precursor species, i.e.,  $[M + H]^+$  and  $[M - H]^-$  of the selected reference standard compounds, the master intensity and the total ion count (TIC) were determined (see Figure S3). The  $m/z$  of the  $[M + H]^+$  and  $[M - H]^-$  were calculated based on the compound SMILES. The top three highest intensity spectra of each reference compound per fragmentation mode were used to assess the performance of the different fragmentation modes. The peaks of the centroided fragmentation spectra were annotated with the previously predicted fragment ions if the match was within a given mass window. A default cut-off value for fragment matching was set to 1 Da, further refinement can be made in the Shiny application (1–100 ppm, 10e–4 to 0.5 Da). The default values in the Shiny application are relative and absolute cut-off of 10 ppm and 0.02 Da, respectively. These are also the tolerances used throughout the manuscript.

The quantitative, e.g., MS1 derived XIC and master intensity, and MS2 derived TIC and fragment intensities, and qualitative fragmentation data were joined by the raw file name and the scan number. To compare fragment ion annotation qualitatively and quantitatively across all compounds and fragmentation modes, we implemented three different scores:

$$\text{Score 1} = \frac{n_{\text{exp frags matched}}}{n_{\text{theor frags}}} \quad (1)$$

$$\text{Score 2} = \frac{n_{\text{exp frags matched}}}{n_{\text{exp frags}}} \quad (2)$$

$$\text{Score 3} = \frac{int_{\text{exp frags matched}}}{int_{\text{MS1 precursor}}} \quad (3)$$

where *exp frags* are the experimentally detected fragments, *exp frags* matched the experimentally detected fragment ions that could be matched to the *in silico* predicted, and *theor frags* the *in silico* predicted theoretically possible fragment ions. *n* indicates the number of fragments and *int* their intensity.

All data and results are visualized and can be interactively accessed in the R shiny application provided with the *uvpd* R package and through <http://fgcz-ms-shiny.uzh.ch:8080/p2722-uvpd/>. The entire workflow is shown in Figure S2.

### 3.5. NTS Data Analysis

NTS data were processed with Compound Discoverer 3.1 (Thermo Fisher Scientific, San Jose, CA, USA) for peak picking, componentization and suspect screening using the spectral library *mzCloud*. The output feature list, i.e., a table with accurate mass/retention time pairs (features) and their intensity, information on whether an MS2 spectrum was acquired for a given feature and the *mzCloud* spectral matching scores were imported into R Studio for further data analysis and visualization. R version 3.6.3. and R-Studio version 1.1.463 were used for the data analysis [14,17].

## 4. Conclusions

Combining the novel fragmentation technique UVPD and cheminformatics tools, we showed the potential of UVPD for structural elucidation of water-relevant OMPs in NTS data. Based on the two complementary methods *k*-means clustering and SOMs, a set of OMPs could be selected that was representative for the water cycle and a wide chemical space. An R-based LC-HRMS data analysis workflow and interactive interface for data visualization was developed to investigate UVPD fragmentation of these OMPs in a high-throughput manner.

Information-rich UVPD fragmentation spectra were achieved for 62% of the examined OMPs, in 15% of the cases also for OMPs that fragmented poorly with HCD. For 26% of the OMPs, neither fragmentation technique generated informative spectra; the remaining 12% HCD spectra were information-rich. UVPD and HCD generated both unique as well as overlapping fragments, demonstrating that some fragmentation pathways are specific to the respective fragmentation methods, while others seem to be more generic. These unique fragments provided additional information for structural identification complementary to HCD spectra. Based on these results, implementation of UVPD as a second fragmentation option in data dependent decision trees during NTS data acquisition is an attractive strategy to improve the confidence in OMP identification.

Analysis of NTS UVPD data with existing NTS software and the spectral library *mzCloud* enables annotation of features in the UVPD data using HCD library spectra of high CEs. For the routine implementation of UVPD fragmentation in NTS workflows, however, databases need to be extended with UVPD spectra, which would allow the full potential of this novel fragmentation technique to be exploited.

**Supplementary Materials:** The following are available online, Figure S1: Negative ionization mode model OMPs (a) gemfibrozil and (b) sucralose; Figure S2: Fragment intensity decreases with increasing UVPD reaction times. Total ion counts (TIC) for a given fragmentation mode versus master intensity; Figure S3: Data analysis and visualization workflow, Table S1: List of reference standards; Table S2: Compounds detected in UVPD NTS experiments.

**Author Contributions:** Conceptualization, A.M.B.; methodology, A.M.B. and C.P.; software, C.P.; validation, A.M.B. and C.P.; formal analysis, A.M.B. and C.P.; investigation, A.M.B., S.S., R.H., D.V. and C.P.; resources, A.M.B., S.S., R.H. and C.P.; data curation, C.P.; writing—original draft preparation, A.M.B. and C.P.; writing—review and editing, A.M.B., S.S., R.H., J.G., D.V. and C.P.; visualization, A.M.B. and C.P.; project administration, A.M.B.; funding acquisition, A.M.B. and C.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Joint Research Program of the Dutch and Belgian drinking water companies.

**Acknowledgments:** The authors would like to thank Robert Bijlsma, University Jaume I, for the kind contribution of reference standards, Tessa Pronk for input in the clustering analyses, Thomas ter Laak and the Dutch and Belgian drinking water companies, in particular Eelco Pieke for critical reading, and the anonymous reviewers, especially reviewer 1, for constructive comments.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

**Availability:** Used code snippets available at <https://github.com/cpanse/uvpd/>; Shiny app available at <http://fgcz-ms-shiny.uzh.ch:8080/p2722-uvpd/>; LC-HRMS raw data of set of reference standards available at <https://doi.org/10.5281/zenodo.4001653>.

## Appendix A

There was no significant difference between mzCloud scores of NTS data with HCD and UVPD fragmentation:

```
t.test(data$mzCloud.Best.Match ~ data$frag)
```

Welch Two Sample t-test

```
data: data$mzCloud.Best.Match by data$frag
```

```
t = 0.44183, df = 177.56, p-value = 0.6591
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
-2.969521 4.682821
```

```
sample estimates:
```

```
mean in group HCD mean in group UVPD
```

```
84.99091 84.13426
```

## References

1. Hollender, J.; Schymanski, E.L.; Singer, H.P.; Ferguson, P.L. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? *Environ. Sci. Technol.* **2017**, *51*, 11505–11512. [[CrossRef](#)] [[PubMed](#)]
2. Brodbelt, J.S. Photodissociation mass spectrometry: New tools for characterization of biological molecules. *Chem. Soc. Rev.* **2014**, *43*, 2757–2783. [[CrossRef](#)] [[PubMed](#)]
3. Morrison, L.J.; Parker, W.R.; Holden, D.D.; Henderson, J.C.; Boll, J.M.; Trent, M.S.; Brodbelt, J.S. UVliPiD: A UVPD-Based Hierarchical Approach for De Novo Characterization of Lipid A Structures. *Anal. Chem.* **2016**, *88*, 1812–1820. [[CrossRef](#)] [[PubMed](#)]
4. Huguet, R.; Stratton, T.; Weisbrod, C.; Berhow, M. The “ETD-like” Fragmentation of Small Molecules. In Proceedings of the ASMS, San Antonio, TX, USA, 5–9 June 2016.
5. Huguet, R.; Stratton, T.; Sharma, S.; Mullen, C.; Canterbury, J.; Zabrouskov, V. Utilizing UVPD Fragmentation for Plant Molecules: Phenylpropanoids. In Proceedings of the ASMS, Indianapolis, Indiana, 4–8 June 2017.
6. Mullen, C.; Weisbord, C.; Zhuk, E.; Huguet, R.; Schwartz, J. Implementation of 213 nm Ultra Violet Photodissociation (UVPD) on a Modified Orbitrap Fusion Lumos. In Proceedings of the ASMS, Indianapolis, Indiana, 4–8 June 2017.
7. Ruttkies, C.; Schymanski, E.L.; Wolf, S.; Hollender, J.; Neumann, S. MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *J. Cheminform.* **2016**, *8*, 3. [[CrossRef](#)] [[PubMed](#)]
8. Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley: Boston, MA, USA, 1977.
9. Karmaus, A.L.; Filer, D.L.; Martin, M.T.; Houck, K.A. Evaluation of food-relevant chemicals in the ToxCast high-throughput screening program. *Food Chem. Toxicol.* **2016**, *92*, 188–196. [[CrossRef](#)] [[PubMed](#)]
10. Guha, R. Chemical Informatics Functionality in R. *J. Stat. Softw.* **2007**, *18*, 16. [[CrossRef](#)]
11. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1979**, *28*, 100–108. [[CrossRef](#)]
12. Charrad, M.; Ghazzali, N.; Boiteau, V.; Niknafs, A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *J. Stat. Softw.* **2014**, *61*, 1–36. [[CrossRef](#)]
13. Kohonen, T. Essentials of the self-organizing map. *Neural Netw.* **2013**, *37*. [[CrossRef](#)] [[PubMed](#)]
14. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.

15. Trachsel, C.; Panse, C.; Kockmann, T.; Wolski, W.E.; Grossmann, J.; Schlapbach, R. rawDiag: An R Package Supporting Rational LC–MS Method Optimization for Bottom-up Proteomics. *J. Proteome Res.* **2018**, *17*, 2908–2914. [[CrossRef](#)] [[PubMed](#)]
16. Panse, C.; Grossmann, J. protViz: Visualizing and Analyzing Mass Spectrometry Related Data in Proteomics. R Package Version 0.6. 2020. Available online: <https://CRAN.R-project.org/package=protViz> (accessed on 12 September 2020).
17. RStudio Team. RStudio: Integrated Development for R. 2015. Available online: <https://rstudio.com/products/rstudio/>.

**Sample Availability:** Samples of the compounds are not available from the authors.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

# Online Prioritization of Toxic Compounds in Water Samples through Intelligent HRMS Data Acquisition

Nienke Meekel, Dennis Vughs, Frederic Béen, and Andrea M. Brunner\*

Cite This: *Anal. Chem.* 2021, 93, 5071–5080

Read Online

ACCESS |



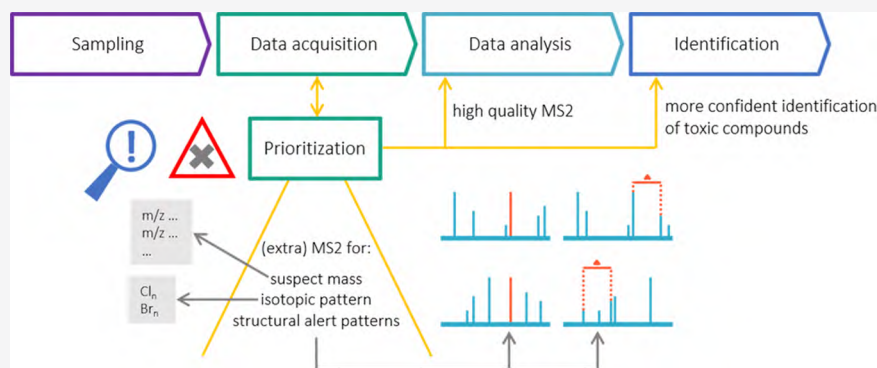
Metrics &amp; More



Article Recommendations



Supporting Information



**ABSTRACT:** LC-HRMS-based nontarget screening (NTS) has become the method of choice to monitor organic micropollutants (OMPs) in drinking water and its sources. OMPs are identified by matching experimental fragmentation (MS2) spectra with library or *in silico*-predicted spectra. This requires informative experimental spectra and prioritization to reduce feature numbers, currently performed post data acquisition. Here, we propose a different prioritization strategy to ensure high-quality MS2 spectra for OMPs that pose an environmental or human health risk. This online prioritization triggers MS2 events based on detection of suspect list entries or isotopic patterns in the full scan or an additional MS2 event based on fragment ion(s)/patterns detected in a first MS2 spectrum. Triggers were determined using cheminformatics; potentially toxic compounds were selected based on the presence of structural alerts, *in silico*-fragmented, and recurring fragments and mass shifts characteristic for a given structural alert identified. After MS acquisition parameter optimization, performance of the online prioritization was experimentally examined. Triggered methods led to increased percentages of MS2 spectra and additional MS2 spectra for compounds with a structural alert. Application to surface water samples resulted in additional MS2 spectra of potentially toxic compounds, facilitating more confident identification and emphasizing the method's potential to improve monitoring studies.

## INTRODUCTION

**Organic Micropollutants in Water.** Issues with water quality occur worldwide due to the large spread of the human population and their extensive use of chemicals, which leads to chemical pollution in a large number of water systems.<sup>1</sup> These systems cause distribution of the pollution with long-range effects, ultimately posing a threat to drinking water sources.<sup>2–4</sup> Various types of organic micropollutants (OMPs), that is, anthropogenic chemicals that are present at trace levels ( $\mu\text{g}/\text{L}$ ), have been detected in ground and surface waters used for drinking water production. These include OMPs such as pesticides, pharmaceuticals, and industrial and consumer products. Despite their low concentrations, OMPs can pose a risk to human and environmental health as they can be toxic, persistent or easily degraded into more toxic (bio)-transformation products.<sup>5</sup> Compounds that pose a potential health risk need to be monitored to be able to assess the actual risks. Monitoring is typically performed using quantitative target analyses. As target analyses are limited to a set of known

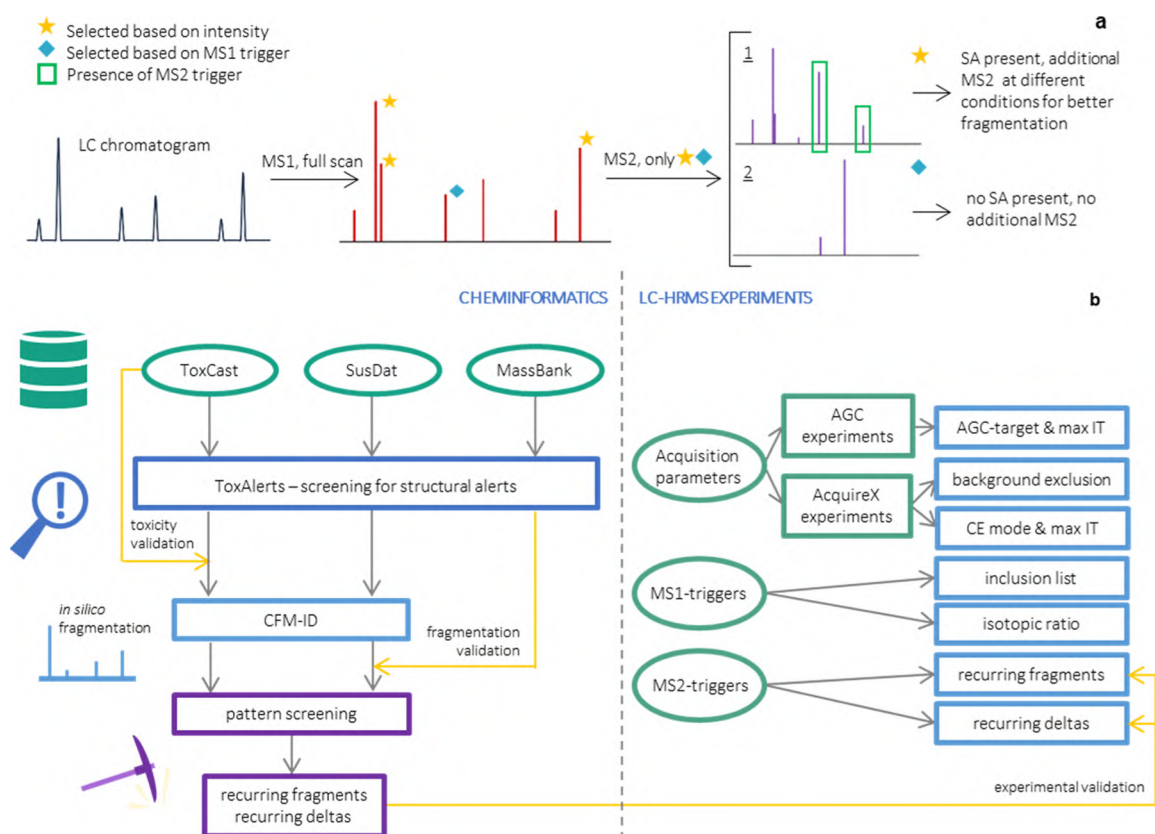
compounds, nontarget screening (NTS) based on liquid chromatography coupled with high-resolution mass spectrometry (LC-HRMS) is often applied to monitor chemical water quality more comprehensively and broaden contaminant discovery.<sup>6,7</sup>

**NTS-Based Micropollutant Identification.** The structural identification of unknown compounds from NTS data remains challenging due to the large number of signals detected per experiment—typically referred to as features (accurate mass and retention time pairs associated with a signal intensity), and the need for high quality fragmentation

Received: October 22, 2020

Accepted: February 22, 2021

Published: March 16, 2021



**Figure 1.** (A) Schematic representation of the proposed LC-HRMS/MS workflow using intelligent acquisition based on structural alerts (SAs). A full MS1 scan is taken after chromatographic separation, and the peaks are screened for their intensity and the presence of MS1-triggers (blue diamond marker). The most intense peaks (based on DDA-approach, yellow star marker) and those that contain a MS1-trigger are selected for a MS2 scan. The MS2 scans are screened for MS2-triggers, indicating the presence of a SA, resulting in two possible scenarios: 1. SA is present, so an additional MS2 scan at different conditions is taken. 2. No SA is present, so structure identification is not necessary, and no additional MS2 is taken. (B) Schematic overview of the strategy that was used to develop the intelligent acquisition method, both cheminformatics (left) and LC-HRMS experiments (right) were applied.

spectra.<sup>8,9</sup> The latter facilitates identification through spectral matching, where experimental spectra are compared to library spectra or *in silico*-predicted spectra. Software tools can connect the experimentally obtained mass spectrum with candidate structures from various sources.<sup>10–14</sup>

**Prioritization.** To limit the features that need to be identified, prioritization can be applied by selecting peaks of interest based on intensity, occurrence, persistence, or potential toxicity.<sup>9,15</sup> Prioritization is currently performed offline during data analysis (Figure S1a). This entails that the structure of prioritized features without fragmentation spectrum or with uninformative, low-quality fragmentation spectra cannot be identified in a sufficiently confident manner. Instead, the sample has to be reanalyzed to obtain high-quality fragmentation spectra requiring more measurement time and resulting in delayed identification. Here, we hypothesize that the high costs and laboriousness of NTS offline prioritization could be remedied by using online prioritization for potentially toxic compounds in the mass spectrometer during data acquisition (Figure S1b).

**Structural Alerts.** Toxic compounds often comprise one or more structural alerts, that is, molecular (sub)structures related to the toxicity of a chemical. Several databases and software programs have been developed to derive and screen molecules for the presence of a structural alert, such as ToxAlerts,<sup>16</sup> DEREK,<sup>17</sup> and MultiCASE.<sup>18</sup> Structural alerts can be specific

for a toxic end point, that is, a measured biological effect in a toxicity test.<sup>19</sup> Most are derived from the end points carcinogenicity and mutagenicity, with several lists published,<sup>20–23</sup> including a revised list by Benigni and Bossa<sup>24</sup> of 33 structural alerts included in the ToxAlerts database. Other water relevant toxic end points are examined less extensively, but some structural alerts were available in ToxAlerts for genotoxicity, endocrine disruption, and developmental toxicity.

**Intelligent Acquisition.** Structural alerts could be used for the “rough” selection of potentially toxic compounds that need to be identified in NTS methods. To this end, fragment ion masses and/or patterns indicating the presence of one or more structural alerts could be used as an MS trigger for further fragmentation events. In addition, suspect lists of toxic compounds and isotopic patterns suggesting anthropogenic origin of a compound were used to prompt a fragmentation event. This novel combination leads to an intelligent acquisition method, which would thereby prioritize (potentially) toxic compounds in contrast to the currently used data-dependent acquisition (DDA) that selects features using only the intensity in MS1 scans as selection criteria for fragmentation.

**Overview.** Here, we developed an intelligent acquisition method that utilizes online prioritization of potentially toxic compounds circumventing reanalysis of the sample due to lacking (high-quality) fragmentation spectra of features that

are prioritized post-analysis (Figure 1). Cheminformatics were applied to determine triggers for (additional) MS2 events to be used in the LC-HRMS method. MS1-triggers exploited accurate mass and isotopic ratios detected in the full scan MS1 spectrum that suggested potential toxicity. MS2-triggers were based on fragment ion masses and/or patterns detected in the MS2 spectrum and linked to the presence of a structural alert. To this end, *in silico* fragmentation predictors were used to predict fragmentation of molecules with a structural alert and screen these spectra for patterns. The derived triggers were experimentally evaluated with LC-HRMS experiments. Finally, the developed method including MS1- and MS2-triggers was compared to a regular NTS method to evaluate whether the prioritization was successful.

## METHODS AND MATERIALS

**Screening of Compounds for Structural Alerts.** The detailed workflow for the screening and fragmentation of the ToxCast<sup>13</sup> data set is given in S2.1. First, the CAS registry numbers of the 9224 compounds registered in the ToxCast data file Chemical\_Summary\_190708.csv<sup>25</sup> were converted into MS-ready SMILES using the CompTox Chemicals Dashboard (<https://comptox.epa.gov/dashboard>).<sup>12</sup> MS-ready SMILES are defined as structural representations that are observed in HRMS.<sup>26</sup> Not all CAS registry numbers could be converted, and some lead to the same MS-ready SMILES, resulting in 7571 unique MS-ready SMILES. In addition to ToxCast entries ( $n = 7571$ ), the MS-ready SMILES of the two databases NORMAN MassBank<sup>11</sup> ( $n = 2304$ ), and NORMAN SusDat<sup>14</sup> ( $n = 65,697$ ) were screened for structural alerts. NORMAN MassBank is a subset of MassBank Europe (<https://massbank.eu>) containing the majority environmental contributors. The compounds in the NORMAN MassBank are also included in NORMAN SusDat; however, MassBank contains fragmentation data and this was used for validation purposes. In the case of MassBank, only the 1903 compounds having available positive ionization HCD data were screened as this ionization mode was later used in the LC-HRMS experiments. Regarding SusDat, compounds were filtered for those with an EPISuite predicted  $\log K_{ow}$  value between  $-2.5$  and  $+3.5$  (provided in SusDat), resulting in 46,688 compounds. This filtering step was applied to eliminate compounds that are not detectable by RPLC.

Four toxic endpoints were selected for screening with ToxAlerts: “endocrine disruption” (EDC), “nongenotoxic carcinogenicity” (NGC), “genotoxic carcinogenicity, mutagenicity” (GCM), and “developmental and mitochondrial toxicity” (DMT). These end points and their corresponding 187 structural alerts were chosen based on their relevance for drinking water and potential human health risk. The endocrine disruption alerts belonged to both estrogenic and androgenic endocrine disruptors.<sup>27</sup> This selection was made based on *in vitro* and *in vivo* (mammalian) data.

The output of ToxAlerts was formatted in R<sup>28</sup> (version 3.6.1 (2019-07-05)) for fragmentation with CFM-ID. A text file was generated per structural alert containing the InChIKey and SMILES code as input for CFM-ID 2.0.

**Validation.** ToxCast assays relevant for the end points that were linked to the structural alerts were selected based on literature.<sup>9</sup> These assays are listed in Table S1. The AC<sub>50</sub> values of the ToxCast compounds with an alert were obtained from “ac50\_Matrix190708.csv” (downloaded at 04 December 2019).<sup>29</sup> In this file, inactive compounds are given an AC<sub>50</sub>

value of  $1 \times 10^6$ . Lower values indicate that the compound is active. Per toxic end point, that is, EDC, DMT, NGC, and GCM, the percentage of molecules with both a structural alert and activity in one of the specified assays was calculated. This percentage was compared to the percentage of active compounds for the total ToxCast data set, irrespective of the presence of a structural alert.

In ToxCast, MS-ready SMILES can occur multiple times but with a different DSSTox Substance identifier and in some cases, varying toxicity information. The toxicity validation was based on the DSSTox Substance ID to include all bioassay results for the same MS-ready SMILES and prevent information loss.

***In silico* Fragmentation.** Compounds with a structural alert were *in silico*-fragmented with the combinatorial fragmentation predictor CFM-ID 2.0 using single energy competitive fragmentation modeling (SE-CFM) in the command line. The main reason for using CFM-ID is that it can be accessed in batch mode. CFM-ID includes assumptions of the fragmentation process such as that the molecule needs to carry a single positive charge, removal or addition of sigma bonds during a break is not allowed, and the valence and even electron rules must be satisfied in all fragments.<sup>30</sup> Note that here, *in silico* fragmentation was not used for subsequent fragment matching but to predict spectra and screen these for patterns.

The command-line utility *cfm-predict.exe*<sup>31</sup> was used to generate fragments with CFM-ID 2.0; the standard trained CFM model and its standard configuration parameters were used (S2.1). The postprocessing option was not included, and the probability threshold was set to 0.001 (default setting). The program output consisted of three lists containing  $m/z$  values and corresponding intensities for low energy CID (10 V), medium energy CID (20 V), and high energy CID (40 V). These energies reflect the type of spectra the model is based on. CFM-ID is based on CID QTOF data, which are comparable to HCD data from an Orbitrap instrument. The output was processed in R.

**Validation.** The *in silico*-predicted fragmentation results of SusDat generated with CFM-ID were validated with experimental data obtained from NORMAN MassBank.<sup>11</sup> MassBank data was available for 2.25% of the 26,081 fragmented molecules with an alert from SusDat. The overlap in percentage of MassBank and CFM-ID fragments was calculated using eqs 1 and 2 to account for the differing total number of MassBank and CFM-ID fragments per spectrum. Since experimental data are also prone to errors, the output of these calculations must be considered as approximations.

$$P_{\text{MassBank}}^{\text{CFM-ID}} = \frac{\text{number of MassBank fragments matching with CFM-ID}}{\text{total number of MassBank fragments}} \cdot 100\% \quad (1)$$

$$P_{\text{CFM-ID}}^{\text{MassBank}} = \frac{\text{number of CFM-ID fragments matching with MassBank}}{\text{total number of CFM-ID fragments}} \cdot 100\% \quad (2)$$

**Pattern Screening.** The *in silico*-predicted fragmentation spectra of compounds with a structural alert were screened for characteristic patterns, that is, recurring fragment masses and recurring mass shifts (deltas). All structural alerts which were found in more than four molecules were included in the analysis. The CFM-ID data set was screened, with the control set being the *in silico*-predicted MS2 spectra of all molecules for each fragmentation method. To be able to compare the effect

of the three CFM-ID energy levels on the recurring fragments and deltas, an intensity threshold was set at a minimum of 5% of the maximum peak intensity (100). The energy levels had an effect on the signal intensity only and not on the type of predicted fragments. Setting this threshold led to elimination of low-intensity fragments, resulting in different fragmentation spectra for the energy levels.

The frequencies of each  $m/z$  value and delta recurring within the MS2 spectra of the molecules of one structural alert were calculated and compared to the frequencies in the total fragmented data set. An extra control step for the frequencies was performed to show the difference in frequencies between a random sample and alerts. A random set of compounds ( $n = 3953$ ) from NORMAN SusDat ( $n_{\text{total}} = 65,697$ ) that had not been screened for structural alerts was fragmented with CFM-ID. The frequencies of recurring fragment masses and recurring deltas within this random sample were then compared to the frequencies within MS2 spectra of compounds with structural alerts derived from ToxCast.

**HRMS Method Development. Sample Preparation.** The chemicals used in this study are listed in Tables S2–S8. An internal standard mixture of atrazine-d5 (CDN isotopes, Pointe-Claire, Canada) and benzotriazole-d4 (LGC Standards, Wesen, Germany) was added to each sample to a final concentration of 1  $\mu\text{g/L}$ . Surface water (SW) (Lekkanaal, the Netherlands) and wastewater treatment plant (WWTP) influent samples, with and without spike-in (see Tables S2–S8) were filtered using 0.2  $\mu\text{m}$  Phenex-RC 15 mm Syringe Filters (Phenomenex, Torrance, USA) prior to analysis. The WWTP-influent samples were 10 times diluted after spike-in and prior to filtration. The blanks used for these analyses were filtered as well. The spiking solution with water-relevant contaminants (see Table S2) was added to the samples to final concentrations of 10  $\mu\text{g/L}$ , 1  $\mu\text{g/L}$ , 100  $\text{ng/L}$ , 10  $\text{ng/L}$ , and 1  $\text{ng/L}$ .

**MS1-Trigger Experiments.** Inclusion lists for MS1-trigger experiments (SusDat,<sup>14</sup> SusDat + tR,<sup>14</sup> UBAPMT,<sup>32</sup> Sjerps,<sup>33</sup> and Spike) were retrieved from the NORMAN Suspect List Exchange (<https://www.norman-network.com/?q=suspect-list-exchange>) and an in-house database and filtered for organic compounds within the full scan mass range (80 to 1000 Da) and polarity amenable to RP-HPLC, that is,  $\log K_{\text{OW}}$  between  $-2.5$  and  $+3.5$  (see the calculation method described in S2.3).

Based on the distribution of the number of chlorine and bromine atoms in the compounds registered in the CompTox Chemicals dashboard ( $n = 869,027$ ),<sup>34</sup> the isotopic ratios covering  $\geq 99\%$  of the chlorinated compounds ( $n = 128,650$ ) and brominated compounds ( $n = 53,258$ ) were used for the MS1-triggers. The isotopic ratios of Cl up to  $\text{Cl}_6$  and Br up to  $\text{Br}_5$  were calculated with the software Xcalibur (Thermo Fisher Scientific, San Jose, USA) and are shown in Table S9. The inclusion lists and the isotopic ratio trigger were tested separately and combined. The design of the resulting acquisition decision trees is shown in Figure S2. The methods were evaluated using surface water and WWTP-influent samples spiked with water-relevant contaminants; see Table S2.

**MS2-Trigger Experiments.** The performance of four different MS2-triggers, that is, two recurring deltas and two recurring fragments, was evaluated using ultrapure water samples spiked with compounds (Tables S3–S8) predicted to exhibit these fragments or deltas in their MS2 spectra based

on the *in silico* experiments. Due to in-house availability of chemicals, only four different MS2-triggers were tested. The spike-in compounds were also added to surface water at concentrations ranging from 1  $\text{ng/L}$  to 10  $\mu\text{g/L}$  to determine sensitivity of the triggers. The MS2-trigger experiments were performed separately, together, and combined with the MS1-triggers using isotopic ratios and the Sjerps inclusion list. Detection of an MS2-trigger led to an additional MS2 event using alternative collision energies (CEs), that is, stepped CE (10, 75, 90) or assisted CE (20, 35, 50, 75, 90), or longer ITs, that is, stepped CE (20, 35, 50) with 200 ms IT instead of the regular 50 ms. These alternative fragmentation events were hypothesized to result in spectra with complementary fragments in the case of alternative energies, and higher-quality spectra in the case of longer ITs. The 11 different methods are described in Table S10 and the design of their decision trees in Figure S3. The experimental data obtained with the MS2-trigger experiments were used to validate the *in silico*-predicted fragmentation spectra and the pattern screening.

**Data Analysis.** The details of the data analysis are reported in S2.4 and S2.5. Data preprocessing and compound annotation were performed using Compound Discoverer 3.1 (Thermo Fisher Scientific, San Jose, USA). Further processing was done in R. Spectrum similarity scores were calculated using the function `SpectrumSimilarity()` from the R-package `OrgMassSpecR` (version 0.5–3).<sup>35</sup> Fragment annotation was performed with the R-package `metfRag` (version 2.4.2)<sup>36</sup> using the function `frag.generateMatchingFragments()` on the centroided MS2-spectra, using default settings. The spectrum similarity scores and number and percentage of annotated fragments and percentage of the annotated peak area were used to gain insights into the quality of the fragmentation spectra acquired with different acquisition settings.

## RESULTS AND DISCUSSION

**Screening of Compounds for Structural Alerts.** Three databases were screened with ToxAlerts for compounds with structural alerts (Figure S4). Screening of the ToxCast database revealed the presence of 139 unique structural alerts in one or more molecules (Figure S4). A total of 109 of these exceeded the pattern detection cutoff of a minimum of five molecules. Screening for structural alerts of SusDat compounds was performed accordingly, resulting in the detection of 152 unique alerts and 133 after the cutoff (Figure S4). The compounds in the NORMAN MassBank data set contained 103 unique structural alerts, of which 59 alerts were present in at least 5 compounds (Figure S4).

**Validation of Toxicity.** To validate the ToxAlerts approach for structural alert detection, we investigated whether compounds with a given structural alert were active in a bioassay linked to the toxic endpoint which was related to that alert. For all four end points, the compounds with structural alerts showed higher percentages of active chemicals in bioassays related to that alert (S3.1) than ToxCast compounds, regardless of the respective structural alert. Based on these results, structural alerts could indeed indicate toxicity, but the alerts used for screening did not cover all chemicals active in these toxic end points. Moreover, many chemicals have not been tested on all included ToxCast assays,<sup>37</sup> causing a data gap.

**In silico Fragmentation.** To be able to determine patterns in the MS2 spectra characteristic for a structural alert,



Table 1. Structural Alerts with a Recurring Fragment (Top) and Deltas (Bottom) and Their Frequencies in Each Data Set

Alert <sup>a</sup>	Name <sup>16,38</sup>	Structure	Endpoint	Recurring fragment or delta m/z or Δm/z
TA344 n <sub>TC</sub> = 23 (0.3%) n <sub>SD</sub> = 95 (0.2%)	Nitrogen and sulphur mustard (specific) X = Cl, Br, I		GCM	62.99960
TA362, TA3023, TA435 n <sub>TC</sub> = 11 (0.1%) n <sub>SD</sub> = 21 (<0.1%)	S or N mustard R = any atom/group; X = F, Cl, Br, I		GCM NGC	62.99960
TA367 n <sub>TC</sub> = 81 (1.1%) n <sub>SD</sub> = 578 (1.2%)	α, β-Unsaturated carbonyl R <sub>1</sub> and R <sub>2</sub> = any atom/group, except alkyl chains with C>5 or aromatic rings; R = any atom/group, except OH, O-		GCM	55.01784
TA401 n <sub>TC</sub> = 5 (<0.1%) n <sub>SD</sub> = 5 (<0.1%)	N-Nitroso-N-alkylureas R = aliphatic carbon or aromatic atom; R <sub>1</sub> = aliphatic carbon		GCM	62.99960 109.01632
TA414 n <sub>TC</sub> = 16 (0.2%) n <sub>SD</sub> = 67 (0.1%)	Haloethylamines R = hydrogen or carbon atom; X = F, Cl, Br, I		GCM	62.99960
TA415 n <sub>TC</sub> = 10 (0.1%) n <sub>SD</sub> = 76 (0.2%)	Haloalkylethers R = carbon atom; X = F, Cl, Br, I; only ethers containing -OCH <sub>2</sub> X (methyl) or -OCH <sub>2</sub> CH <sub>2</sub> X (ethyl) groups are included		GCM	62.99960
TA11479 n <sub>TC</sub> = 24 (0.3%) n <sub>SD</sub> = 75 (0.2%)			EDC	27.99491
TA322 n <sub>TC</sub> = 445 (5.9%) n <sub>SD</sub> = 3524 (7.5%)	Aromatic amine (general) Ar = any aromatic/heteroaromatic ring	Ar-NH <sub>2</sub>	GCM	17.02655
TA360, TA3021 n <sub>TC</sub> = 9 (0.1%) n <sub>SD</sub> = 92 (0.2%)	N-Methylol derivatives R = any atom/group		GCM NGC	30.01056
TA366, TA3027, TA332 n <sub>TC</sub> = 5 (<0.1%) n <sub>SD</sub> = 12 (<0.1%)	Alkyl nitrite R = any alkyl group	R-O-N=O	GCM NGC	47.00073
TA387 n <sub>TC</sub> = 44 (0.6%) n <sub>SD</sub> = 605 (1.3%)	Aromatic N-acyl amine Ar = any aromatic/heteroaromatic ring, R = hydrogen, methyl; chemicals with ortho-disubstitution, or with an ortho carboxylic acid substituent with respect to the N-acyl amine group are excluded; chemicals with a sulfonic acid group (-SO <sub>3</sub> H) on the same ring of the amino group are excluded.		GCM	42.01056
TA395 n <sub>TC</sub> = 52 (0.7%) n <sub>SD</sub> = 669 (1.4%)	Secondary aromatic acetamides and formamides Ar = any aromatic/heteroaromatic ring; R = H, methyl or activated methyl		GCM	42.01056
TA408 n <sub>TC</sub> = 16 (0.2%) n <sub>SD</sub> = 174 (0.4%)	Benzylic halides Ar = any aromatic/heteroaromatic ring; X = Cl, Br, I		GCM	35.97668
TA423 n <sub>TC</sub> = 20 (0.3%) n <sub>SD</sub> = 138 (0.3%)	Isocyanate R = any atom/group	R-N=C=O	GCM	15.01090 27.99491
TA424 n <sub>TC</sub> = 8 (0.1%) n <sub>SD</sub> = 68 (0.1%)	Isothiocyanate R = any atom/group	R-N=C=S	GCM	43.97207
TA433 n <sub>TC</sub> = 9 (0.1%) n <sub>SD</sub> = 92 (0.2%)	N-Methylol derivatives R = any atom/group		GCM	30.01056

<sup>a</sup>n<sub>TC</sub> and n<sub>SD</sub> represent the number of compounds in the ToxCast and SusDat data set, respectively. A description of the structural alert is given in the second column.<sup>38</sup>

fragmentation spectra were generated *in silico* using the fragmentation software CFM-ID 2.0. CFM-ID provided intensity values to filter for the most likely fragments.

*Validation with NORMAN MassBank Data.* The *in silico* fragmentation results generated by CFM-ID were validated with experimental HCD data retrieved from NORMAN

Table 2. Comparison of Percentage MS2 Scans of the Inclusion List  $m/z$  Values between Methods<sup>a</sup>

inclusion list type	sample type	method with inclusion list $\mu\%$ features with MS2	standard NTS method $\mu\%$ features with MS2	$p$ -value	test type
SusDat	WWTP-influent	95.86	91.76	0.01576	$t$ -test
	SW	97.68	97.31	0.1039	$t$ -test
UBAPMT	WWTP-influent	100.0	100.0	-	
	SW	96.97	96.97	-	
Sjerp	WWTP-influent	98.36	93.32	0.01485	$t$ -test
	SW	95.76	96.58	0.8779	$t$ -test
Spike	WWTP-influent	96.41	95.60	0.3425	$t$ -test
	SW	98.58	97.65	0.1250	Sign test
SusDat + tR	WWTP-influent	97.74	92.53	0.004934	$t$ -test
	SW	98.80	97.87	0.0005877	$t$ -test

<sup>a</sup>In one case (Spike SW), a Sign test is applied since the data was not normally distributed.

MassBank.<sup>11</sup> Positive ionization HCD data were available for 1903 compounds, 587 of which were NORMAN SusDat compounds with a structural alert. To account for the experimental error in the MassBank data, a 10 ppm mass tolerance was used to find overlapping fragments between the CFM-ID predicted and experimental MassBank fragments. Depending on the CFM-ID energy, for 144 up to 398 of the 587 compounds  $\geq 50\%$  of the CFM-ID fragments were matched with a MassBank fragment (S3.2, Table S3.2, Figure S3.2). As no CFM-ID fragmentation energy setting outperformed the others, all energies were included in the further analyses.

**Pattern Screening.** After *in silico* generation of predicted fragmentation spectra, these predicted spectra of compounds with structural alerts were screened for patterns characteristic for each structural alert for subsequent use as MS2-triggers. These patterns included recurring fragment masses and recurring mass differences between two fragments referred to as deltas. All three CFM-ID fragmentation energies were included in the pattern screening, and patterns were filtered for occurrence in the spectra of at least two fragmentation energies to remove less relevant fragments and/or deltas. To further increase specificity, only fragments and deltas with a frequency higher than 0.5 in both the ToxCast and SusDat data sets were taken into consideration. These strict requirements led to a relatively low number of alerts: 6 recurring fragments and 11 deltas exceeded this frequency cut-off (Table 1).  $m/z$  62.99960 was a recurring fragment in mustard-like structural alerts, which could correspond to  $C_2ClH^+$ , a fragment that is likely to form from these alerts. The recurring fragments  $m/z$  55.01784 and  $m/z$  109.01632 could correspond to  $C_3H_3O^+$  and  $C_2H_6ClON_2^+$ , respectively. Five structural alerts corresponded to the same recurring fragment, that is,  $m/z$  62.99960 (Table 1), and four structural alerts to two recurring deltas, that is,  $\Delta$   $m/z$  27.99491 and  $\Delta$   $m/z$  42.01056 (Table 1) due to the similarity in their structures.

For both the recurring fragments and deltas, their frequencies within an alert were significantly higher than the highest frequency observed in the three different control data sets, that is, in all fragmented molecules with an alert from ToxCast, a random sample from SusDat, regardless of the presence of an alert, and all fragmented molecules with an alert from SusDat (Tables S11–S13). This confirmed that the recurring fragments and deltas were characteristic for their structural alerts. Two deltas detected with high frequency were 2.01565 and 18.01056 Da. These were not considered as relevant deltas because there was no significant difference between their frequencies in the compounds with alerts

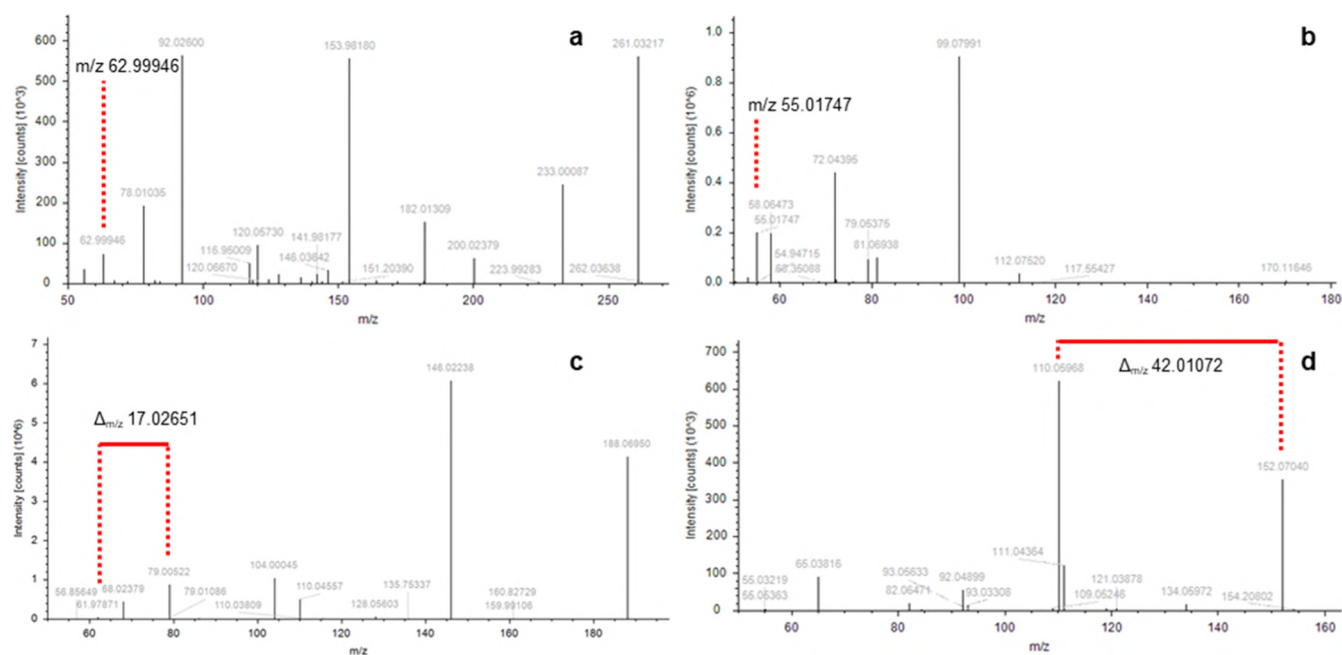
compared to the total data set. These deltas are expected to correspond to a loss of 2H and  $H_2O$ , respectively.

In order to increase the “yield” of alerts that could be used as trigger, other data mining approaches could be applied such as hierarchical clustering, random forest or multiple linear regression to find patterns characteristic for a specific structural alert. However, one has to take into account that the output of more advanced pattern recognition needs to be in a format that is suitable for implementation in acquisition software used to operate mass spectrometers. Moreover, even more reliable results could be generated when experimental fragmentation data is used instead of *in silico*-predicted fragments.

Based on in-house availability of chemicals, the recurring fragments  $m/z$  62.99960 of ToxCast alert TA344/TA362 (Table 1) and  $m/z$  55.01784 of alert TA367 and the recurring deltas  $m/z$  17.02655 of alert TA322 and  $m/z$  42.01056 of alert TA387/TA395 were selected for use in the MS2-trigger experiments.

**LC-HRMS Experiments. MS1-Trigger Experiments.** Prior to implementing MS triggers, background exclusion and selected MS acquisition parameters were optimized to maximize available cycle time for (additional) MS2 scans and MS2 spectral quality during online prioritization (S3.3 Acquisition parameter optimization). Subsequently, the potential of MS1-triggers for the prioritization of toxic compounds was assessed experimentally. The MS1-triggers consisted of five different inclusion lists and isotopic ratios for chlorinated and brominated compounds.

Based on the Cl/Br pattern, which is a parameter in Compound Discoverer stating whether a chlorine- or bromine-specific isotopic pattern is present in the MS1, there was a significant increase in the percentage of MS2 scans for the surface water ( $\mu_{NTS} = 94.2 \pm 0.4\%$ ,  $\mu_{MS1-trig} = 100 \pm 0\%$ ,  $p$ -value of 0.001292, Figure S7) but not the WWTP-influent samples ( $\mu_{NTS} = 82.7 \pm 5.2\%$ ,  $\mu_{MS1-trig} = 84.5 \pm 1.3\%$ , Figure S7). The lesser performance in the WWTP-influent samples could be due to the more complex MS1 spectra confounding isotopic ratios, in particular when low error tolerances are set. This is also supported by the pattern matches determined during the Compound Discoverer analysis. The peaks of Cl- and/or Br-containing features should contain a characteristic isotopic pattern due to the natural abundance of chlorine and bromine isotopes. For some brominated and/or chlorinated compounds, no additional MS2 was triggered because the isotopic ratio deviated more than the allowed 10% ratio tolerance. Additional experiments with increased mass tolerance (10 ppm instead of 3 ppm, which was chosen to test the extreme effect) and ratio tolerance (15% instead of



**Figure 2.** Four experimentally obtained MS2 scans with expected MS2-trigger marked in red, of which an additional MS2 was triggered. (a) MS2 spectrum of ifosfamide and its recurring fragment, (b) MS2 spectrum of diacetone acrylamide and its recurring fragment, (c) MS2 spectrum of desethylatrazine and its recurring delta, and (d) MS2 spectrum of paracetamol and its recurring delta.

10%) did not improve this. Setting priority of the decision tree to the branch with the targeted isotopic ratio node, however, led to a significant increase in percentage of Cl and/or Br containing features with an MS2 spectrum ( $p$ -value of 0.04225, one-sided  $t$ -test). Further experiments could focus on optimizing the isotopic ratio and mass tolerance of the MS1-trigger to balance a more tolerant threshold and the subsequent increase in false-positive triggers.

Based on these results, the isotopic ratio was implemented (with the narrow tolerances) in the intelligent acquisition method as MS1-trigger as it increased MS2 spectral availability for Cl-/Br- containing features which are mostly anthropogenic and often toxic, and the risk of triggering fragmentation of irrelevant features was low.

Regarding the use of inclusion lists as MS1-triggers, there was a significant increase in percentage of MS2 scans for  $m/z$  values present in the inclusion lists SusDat, SusDat + tR, and Sjerps in the WWTP-influent and SusDat + tR in the SW samples (Table 2). The lesser effect observed in SW samples can be explained by the fact that the standard NTS method without an inclusion list was able to separate and identify the features present in the SW but not WWTP influent samples. Due to the large number of compounds in SusDat (+tR), including non water-relevant ones, the Sjerps list was used for subsequent MS2-trigger experiments.

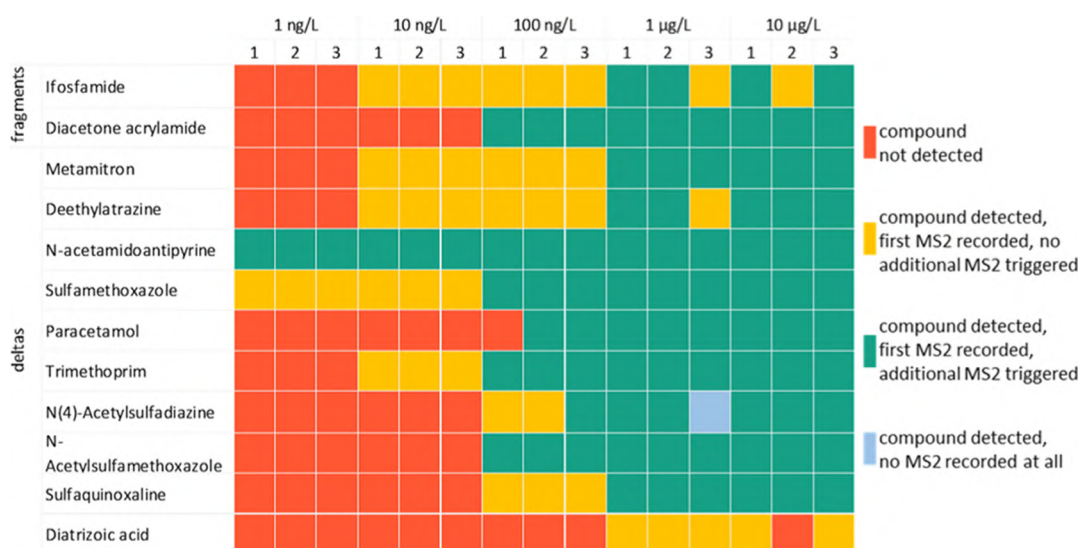
Overall, less complex matrices such as SW samples seemed to benefit more from the isotopic ratio MS1-trigger, demonstrated by the significant increase in the percentage of MS2 scans for these samples. The analysis of more complex matrices such as WWTP influent improved through the use of inclusion lists that ensured that water relevant compounds were fragmented. The inclusion list MS1-trigger showed promising results for the inclusion lists SusDat, with and without retention time estimate, and Sjerps. As the Sjerps list consisted of water-relevant compounds, this list was used in subsequent experiments in combination with the MS2-triggers.

**MS2-Trigger Experiments.** Next to MS1-triggers that trigger an MS2 scan, MS2-triggers were developed that trigger an additional MS2 scan in the presence of a structural alert, indicating a potentially toxic compound. Four specific fragment masses and deltas were used as MS2-triggers: the recurring fragments  $m/z$  62.99960 of alert TA344/TA362 and  $m/z$  55.01784 of alert TA367 and the recurring deltas  $m/z$  17.02655 of alert TA322 and  $m/z$  42.01056 of alert TA387/TA395. These alerts correspond to the toxic end points genotoxic carcinogenicity and mutagenicity. A total of 12 reference compounds were selected, which were hypothesized to contain an alert and MS2-trigger based on pattern screening (Tables S3–S7).

The recurring fragments were present in the MS2 spectra of all 12 detected compounds, thereby confirming the usefulness of the *in silico*-predicted spectra generated with CFM-ID. MS2 scans were triggered in all cases, except ifosfamide and diacetone acrylamide. For these compounds, the ppm mass error tolerance was too narrow. Increasing the tolerance to 20 ppm lead to triggering of additional MS2 scans. Therefore, a higher error tolerance or potentially a combination of a low relative tolerance and an absolute tolerance of  $m/z$  0.001 would be advantageous. Alternatively, the calibration range of the instrument could be expanded to lower  $m/z$  values.

In addition to the recurring fragments, the use of recurring deltas as MS2-triggers was investigated. The recurring delta  $m/z$  17.02650 corresponding to alert TA322 was detected in the MS2 spectra of all reference compounds that contained this alert, thereby validating the approach of using CFM-ID to *in silico* predict spectra. Additional MS2 scans were triggered for all compounds with this recurring delta.

Examples of spectra where an additional MS2 was successfully MS2-triggered are shown in Figure 2. The recurring delta  $m/z$  42.01060 corresponding to the alerts TA387 and TA395 was detected in all spectra except those of diatrizoic acid and one of the three triplicates of  $n$ -



**Figure 3.** Schematic overview of the detection of the spike-in compounds, and whether an additional MS2 was triggered or not.

acetylsulfamethoxazole. The delta  $m/z$  42.01060 triggered additional MS2 scans in all other compounds, where the recurring delta was detected. The measured MS2 spectra of diatrizoic acid did not match the *in silico*-predicted spectrum (see Figure S8), and the peaks that were expected to form the recurring delta ( $m/z$  614.7769272 and  $m/z$  572.7663625 or  $m/z$  596.7663625 and  $m/z$  554.7557979) were not present.

Next, the effect of compound concentration levels on the MS2-triggers was investigated (Figure 3). To this end, a concentration range from the 10  $\mu\text{g/L}$  used in the proof-of-principle experiments down to 1  $\text{ng/L}$  was used. At first, the precursor ion of the compound containing a structural alert has to be selected for a MS2 scan, in which the MS2-trigger can be detected. Thereafter, this trigger can prompt the consecutive MS2 scan. Generally, once a compound was detected and a MS2 scan recorded, an additional MS2 scan was triggered as well, indicating the sensitivity of the MS2-trigger. However, some exceptions were observed (marked in yellow in Figure 3). In these cases, the compound was detected, but no additional MS2 scans were triggered due to the absence of the trigger in the MS2 scan (in case of metamitron, desethylatrazine up to 100  $\text{ng/L}$ , sulfamethoxazole, trimethoprim, and sulfaquinoxaline) or the selected error tolerance (5 ppm, in case of ifosfamide and desethylatrazine in the third measurement at 1  $\mu\text{g/L}$ ). In one case, no MS2 scan was recorded. Consequently, no additional MS2 scan could be triggered. This was the case for a single measurement of *N*(4)-acetylsulfadiazine at 1  $\mu\text{g/L}$ .

MS2-triggers were applied to prompt an additional MS2 scan that would ensure more informative fragmentation spectra, that is, higher spectral quality or complementary fragments to the first MS2 scan, of features with a structural alert. Different acquisition parameters were used for this additional MS scan: stepped CE (10, 75, 90 instead of the regular 20, 35, 50), assisted CE (20, 35, 50, 75) and longer ITs (200 ms IT instead of the regular 50 ms). The effect of the acquisition parameter to increase the information content of the spectra was assessed based on the mzCloud scores assigned to the identified features because these could be easily extracted from the Compound Discoverer results. The mzCloud scores tended to increase slightly (approximately 0.1–1%) with the additional MS2 scan using assisted CE and

longer IT. As mzCloud scores are based on experimental spectra that might have not been generated with the optimal acquisition parameters, as an alternative performance evaluation MetFrag annotation was examined. This showed that generally, the additional MS2 scans using assisted CE had a higher percentage of annotated intensity (Figure S9) but no higher percentage of annotated fragments (Figure S10). However, to reach the maximum advantage of the additional MS2, higher spectral quality that facilitates identification, spectral quality metrics need to be developed and implemented online, that is, during the measurement.

**Application of Triggered Methods to SW Samples.** To compare the online prioritization methods to the standard NTS method, a SW sample spiked with water-relevant contaminants was analyzed. Three versions of the intelligent acquisition method combining the MS1- (isotopic ratio and Sjerps inclusion list) and MS2-triggers (fragment  $m/z$  62.99960, fragment  $m/z$  55.01784, delta  $m/z$  42.01060, and delta  $m/z$  17.02650) were used: with the additional MS2 with either stepped CE, ACE or longer IT. Ten of the spiked compounds contained an alert related to these MS2-triggers, and for eight of them, an additional MS2 was triggered. The spectra of 2-aminobenzothiazole and 2,4-dichloroaniline did not exhibit the expected delta  $m/z$  17.02650. Consequently, no additional MS2 was triggered. Using the regular NTS method (see Tables S14–S15), the mzCloud best match and mzVault best match scores (S1.2) ranged from 97.1 to 99.8 out of 100 and from 89.6 to 99.8, respectively. This indicates that these scores are already high. Despite these high scores, for the compounds desisopropylatrazin and desethylatrazin, the mzCloud scores increased with all three tested intelligent acquisition methods (Table S15).

## CONCLUSIONS

Overall, the intelligent acquisition method, using the Sjerps inclusion list and additional MS2's with ACE or longer IT, directed prioritization toward potentially toxic compounds. The isotopic ratio MS1-trigger significantly improved the percentage of Cl-/Br-containing compounds with a MS2 spectrum if priority was assigned in the method. The use of an inclusion list increased the percentage of MS2 spectra of

features with  $m/z$  values present in the inclusion list. The MS2-trigger method successfully triggered additional MS2 scans of molecules with a structural alert for the four alerts that were tested. Therefore, the method could prioritize these potentially toxic compounds online, and further developments will improve the added value. Once fully developed, it could be far more efficient than many current strategies involving post-acquisition processing.

Future work could expand the developed method with more structural alerts targeting different toxic endpoints, implementing the method in our laboratory, and making it available for other laboratories to use. Ultimately, application of intelligent acquisition methods in routine monitoring studies is necessary to expose the benefits in practice for safety monitoring of drinking water sources. While a clear benefit was demonstrated for MS1- and MS2-triggers, the automatic triggering of an additional MS2 scan will reach its maximum benefit once more knowledge is available on how spectral quality can be optimized in a directed manner through selection of appropriate acquisition parameters.

## ■ ASSOCIATED CONTENT

### SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.analchem.0c04473>.

Acquisition parameters, spectral libraries and chemical databases, workflow screening with ToxAlerts and fragmentation, instrument settings for LC-HRMS experiments, inclusion lists, data analysis, compound discoverer workflow parameters, toxicity validation, validation of *in silico* fragmentation, acquisition parameter optimization, NTS workflow, design of acquisition decision trees, ToxAlerts screening results, frequency distributions of recurring fragments and deltas, number of detected features per MS1-trigger method, number of detected chlorinated and brominated features per MS1-trigger method, *in silico* predicted and experimental MS2 of diatrizoic acid, comparison of annotated intensity of regular MS2 scan and triggered MS2 scan, and comparison of annotated fragments of regular MS2 scan and triggered MS2 scan (PDF)

ToxCast assays used in toxicity validation, lists of spiked compounds and sample compositions, isotopic ratios, methods of MS2-trigger experiments, frequencies of recurring fragments and deltas within compounds with an alert, frequencies of recurring fragments and deltas in control data sets, mzVault and mzCloud best match scores from the total performance analysis (XLSX)

Screening results ToxAlerts (ZIP)

## ■ AUTHOR INFORMATION

### Corresponding Author

Andrea M. Brunner – KWR Water Research Institute, 3430 BB Nieuwegein, The Netherlands; [orcid.org/0000-0002-2801-1751](https://orcid.org/0000-0002-2801-1751); Email: [andrea.brunner@kwrwater.nl](mailto:andrea.brunner@kwrwater.nl)

### Authors

Nienke Meekel – KWR Water Research Institute, 3430 BB Nieuwegein, The Netherlands

Dennis Vughs – KWR Water Research Institute, 3430 BB Nieuwegein, The Netherlands

Frederic Béen – KWR Water Research Institute, 3430 BB Nieuwegein, The Netherlands; [orcid.org/0000-0001-5910-3248](https://orcid.org/0000-0001-5910-3248)

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acs.analchem.0c04473>

## Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The authors acknowledge Astrid Reus, Tessa Pronk, and Margo van der Kooi from the KWR Water Research Institute for advice about relevant toxic end points, advice in programming in R, and preparation of the samples. Caroline Ding, Lena Becciolini, and Seema Sharma from Thermo Fisher Scientific are acknowledged for their help with the data acquisition and data processing software. Christian Panse from ETH Zürich is acknowledged for the development of the centroid function in R. Eelco Pieke from Het Waterlaboratorium and Jan van der Kooi from WLN for critical reading of the manuscript and Igor Tetko from VCCLAB for helping with ToxAlerts. This work was funded by the Joint Research Program of the Dutch and Belgian drinking water companies.

## ■ ABBREVIATIONS

AC <sub>50</sub>	concentration at 50% of maximum activity
CE	collision energy
CFM	competitive fragmentation modeling
CID	collision-induced dissociation
DDA	data-dependent acquisition
DMT	developmental and mitochondrial toxicity
EDC	endocrine disruption
GCM	genotoxic carcinogenicity, mutagenicity
HCD	higher-energy collisional dissociation
HPLC	high-performance liquid chromatography
HRMS	high-resolution mass spectrometry
InChI	International Chemical Identifier
IT	ion injection time
K <sub>OW</sub>	octanol–water partition coefficient
LC	liquid chromatography
MS/MS	tandem mass spectrometry, fragmentation spectrum
MS2	tandem mass spectrometry, fragmentation spectrum
NGC	nongenotoxic carcinogenicity
NTS	nontarget screening
OMP	organic micropollutants
QTOF	quadrupole-time-of-flight mass spectrometer
RPLC	reversed-phase liquid chromatography
tR	retention time
SA	structural alert
SMILES	simplified molecular-input line-entry specification
SusDat	NORMAN Substance Database
SW	surface water
WWTP	wastewater treatment plant

## ■ REFERENCES

- (1) Stamm, C.; Räsänen, K.; Burdon, F. J.; Altermatt, F.; Jokela, J.; Joss, A.; Ackermann, M.; Eggen, R. I. L. In *Large-Scale Ecology: Model Systems to Global Perspectives*; Dumbrell, A. J., Kordas, R. L., Woodward, G., Eds.; Academic Press, 2016, pp 183–223.
- (2) Ruff, M.; Mueller, M. S.; Loos, M.; Singer, H. P. *Water Res.* 2015, 87, 145–154.

- (3) Bernhardt, E. S.; Rosi, E. J.; Gessner, M. O. *Front. Ecol. Environ.* **2017**, *15*, 84–90.
- (4) Brack, W.; Dulio, V.; Ågerstrand, M.; Allan, I.; Altenburger, R.; Brinkmann, M.; Bunke, D.; Burgess, R. M.; Cousins, L.; Escher, B. L.; Hernández, F. J.; Hewitt, L. M.; Hilscherová, K.; Hollender, J.; Hollert, H.; Kase, R.; Klauer, B.; Lindim, C.; Herráez, D. L.; Miège, C.; et al. *Sci. Total Environ.* **2017**, *576*, 720–737.
- (5) Schwarzenbach, R. P.; Escher, B. L.; Fenner, K.; Hofstetter, T. B.; Johnson, C. A.; von Gunten, U.; Wehrli, B. *Science* **2006**, *313*, 1072.
- (6) Bletsou, A. A.; Jeon, J.; Hollender, J.; Archontaki, E.; Thomaidis, N. S. *TrAC, Trends Anal. Chem.* **2015**, *66*, 32–44.
- (7) Samanipour, S.; Martin, J. W.; Lamoree, M. H.; Reid, M. J.; Thomas, K. V. *Environ. Sci. Technol.* **2019**, *53*, 5529–5530.
- (8) Hollender, J.; Schymanski, E. L.; Singer, H. P.; Ferguson, P. L. *Environ. Sci. Technol.* **2017**, *51*, 11505–11512.
- (9) Brunner, A. M.; Dingemans, M. M. L.; Baken, K. A.; van Wezel, A. P. *J. Hazard. Mater.* **2019**, *364*, 332–338.
- (10) HighChem LLC. mzCloud Features. <https://www.mzcloud.org/Features> (accessed Nov 25, 2019).
- (11) Schymanski, E. L.; Schulze, T.; Alygizakis, N.; Meier, R. *S11 MASSBANK|NORMAN Compounds in MassBank*, version NORMAN-SLE-S1.0.1.1. Zenodo.
- (12) Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; Richard, A. M. *J. Cheminf.* **2017**, *9*, 61.
- (13) Richard, A. M.; Judson, R. S.; Houck, K. A.; Grulke, C. M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M. T.; Wambaugh, J. F.; Knudsen, T. B.; Kancherla, J.; Mansouri, K.; Patlewicz, G.; Williams, A. J.; Little, S. B.; Crofton, K. M.; Thomas, R. S. *Chem. Res. Toxicol.* **2016**, *29*, 1225–1251.
- (14) NORMAN Network. Aalizadeh, R.; Alygizakis, N.; Schymanski, E.; Slobodnik, J. *S0|SUSDAT|Merged NORMAN Suspect List: SusDat*, version NORMAN-SLE-S0.0.2.1. Zenodo.
- (15) Brunner, A. M.; Bertelkamp, C.; Dingemans, M. M. L.; Kolkman, A.; Wols, B.; Harmsen, D.; Siegers, W.; Martijn, B. J.; Oorthuizen, W. A.; Ter Laak, T. L. *Sci. Total Environ.* **2020**, *705*, 135779.
- (16) Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V. *J. Chem. Inf. Model.* **2012**, *52*, 2310–2316.
- (17) Ridings, J. E.; Barratt, M. D.; Cary, R.; Earnshaw, C. G.; Eggington, C. E.; Ellis, M. K.; Judson, P. N.; Langowski, J. J.; Marchant, C. A.; Payne, M. P.; Watson, W. P.; Yih, T. D. *Toxicology* **1996**, *106*, 267–279.
- (18) Saiakhov, R. D.; Klopman, G. *Toxicol. Mech. Methods* **2008**, *18*, 159–175.
- (19) Organisation for Economic Co-operation and Development (OECD). Appendix I - Collection of working definitions. <http://www.oecd.org/chemicalsafety/testing/49963576.pdf> (accessed Nov 18, 2019).
- (20) Ashby, J.; Tennant, R. W. *Mutat. Res.* **1988**, *204*, 17–115.
- (21) Bailey, A. B.; Chanderbhan, R.; Collazo-Braier, N.; Cheeseman, M. A.; Twaroski, M. L. *Regul. Toxicol. Pharmacol.* **2005**, *42*, 225–235.
- (22) Kazius, J.; McGuire, R.; Bursi, R. *J. Med. Chem.* **2005**, *48*, 312–320.
- (23) Kazius, J.; Nijssen, S.; Kok, J.; Bäck, T.; IJzerman, A. P. *J. Chem. Inf. Model.* **2006**, *46*, 597–605.
- (24) Benigni, R.; Bossa, C. *Mutat. Res.* **2008**, *659*, 248–261.
- (25) United States Environmental Protection Agency. Chemical \_ Summary\_190708 from invitrodb\_v3.2. <https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data> (accessed Dec 4, 2019).
- (26) McEachran, A. D.; Mansouri, K.; Grulke, C.; Schymanski, E. L.; Ruttkies, C.; Williams, A. *J. Cheminf.* **2018**, *10*, 45.
- (27) Nendza, M.; Wenzel, A.; Muller, M.; Lewin, G.; Simetska, N.; Stock, F.; Arning, J. *Environ. Sci. Eur.* **2016**, *28*, 26.
- (28) R Core Team. *R: A Language and Environment for Statistical Computing*, 3.6.1; R Foundation for Statistical Computing: Vienna, Austria, 2019.
- (29) United States Environmental Protection Agency. ac50\_Matrix\_190708 from invitrodb\_v3.2. <https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data> (accessed Dec 4, 2019).
- (30) Allen, F.; Greiner, R.; Wishart, D. *Metabolomics* **2015**, *11*, 98–110.
- (31) Allen, F. CFM-ID: Competitive Fragmentation Modeling for Metabolite Identification. <https://sourceforge.net/p/cfm-id/wiki/Home/> (accessed Dec 20, 2019).
- (32) Von der Ohe, P.; Fischer, S. *S36|UBAPMT|Potential Persistent, Mobile and Toxic (PMT) Substances*, version NORMAN-SLE-S36.0.1.0. Zenodo.
- (33) Sjerps, R. M. A. *S27|KWRSJERPS2|Extended Suspect List from Sjerps et al (KWRSJERPS)*, version NORMAN-SLE-S27.0.1.1. Zenodo.
- (34) United States Environmental Protection Agency. DSSTox MS Ready Mapping File. <https://comptox.epa.gov/dashboard/downloads> (accessed Feb 12, 2020).
- (35) Dodder, N.; Mullen, K. *Organic Mass Spectrometry*, version 0.5-3; 2017.
- (36) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. *J. Cheminf.* **2016**, *8*, 3.
- (37) Louisse, J.; Dingemans, M. M. L.; Baken, K. A.; van Wezel, A. P.; Schriks, M. *Chemosphere* **2018**, *209*, 373–380.
- (38) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Palyulin, V. A.; et al. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 533–554.

#### NOTE ADDED AFTER ISSUE PUBLICATION

This article was initially published with an incorrect copyright statement and was corrected on or around May 5, 2021.



Improved identification of toxic compounds in  
drinking water sources through HRMS based  
intelligent data acquisition

Nienke Meekel  
June 2020



UNIVERSITEIT VAN AMSTERDAM



# MSc Chemistry Analytical Sciences

Master Thesis

---

## Improved identification of toxic compounds in drinking water sources through HRMS based intelligent data acquisition

---

*by*

**Nienke Meekel**

**10719059 (UvA) / 2654274 (VU)**

*June 2020*

*48 EC*

*28 October 2019 – 26 June 2020*

*Daily supervisor*      dr. Andrea M. Brunner, KWR Water Research Institute  
*Examiner*              prof. dr. Marja H. Lamoree, VU Amsterdam  
*Second Assessor*      prof. dr. Govert W. Somsen, VU Amsterdam

Chemical Water Quality & Health

KWR Water Research Institute



## List of abbreviations

AC <sub>50</sub>	concentration at 50% of maximum activity
AGC	automatic gain control
BDE	bond dissociation energy
CALUX	chemical activated luciferase gene expression
CE	collision energy
CFM	competitive fragmentation modeling
CID	collision-induced dissociation
DDA	data-dependent acquisition
DIA	data-independent acquisition
dmt	developmental and mitochondrial toxicity
EDA	effect-directed analysis
edc	endocrine disruption
EI	electron impact ionization
ESI	electrospray ionization
gcm	genotoxic carcinogenicity, mutagenicity
HCD	higher-energy collisional dissociation
HPLC	high-performance liquid chromatography
HRMS	high-resolution mass spectrometry
HT-EDA	high-throughput effect directed analysis
InChI	International Chemical Identifier
IT	ion injection time
K <sub>ow</sub>	octanol-water partition coefficient
LC	liquid chromatography
MS/MS	tandem mass spectrometry, fragmentation spectrum
MS2	tandem mass spectrometry, fragmentation spectrum
ngc	non-genotoxic carcinogenicity
NTS	non-target screening
OMP	organic micropollutants
PFAS	poly- and perfluorinated alkylated substances
PMT	persistent, mobile and toxic
QSARs	Quantitative Structure-Activity Relationships
QTOF	quadrupole-time of flight mass spectrometer
RPLC	reversed-phase liquid chromatography
tR	retention time
SA	structural alert
SDF	structure data file
SMILES	simplified molecular-input line-entry specification
SusDat	NORMAN Substance Database
SW	surface water
TIE	toxicity identification evaluation
WWTP	wastewater treatment plant

## Abstract

The increasing occurrence of organic micropollutants (OMPs) in drinking water sources stresses the need for comprehensive chemical monitoring to ensure drinking water quality. Recently, LC-HRMS-based non-target screening (NTS) has become the method of choice for OMP monitoring. NTS enables the identification of a wide range of compounds based on their mass and isotopic pattern provided in the MS1 full scan spectrum, and the match of their MS2 fragmentation spectrum with library or *in silico* predicted spectra. However, it is challenging to acquire MS2 spectra for all compounds as the number of MS2 scans is limited by the time available between two MS1 scans. Furthermore, high quality MS2 spectra are required for confident identification based on spectral matching. To date, due to the high number the structural identification of unknown compounds requires prioritization which occurs during data analysis. Only at this point of the identification workflow, relevant compounds are selected and the availability and quality of their MS2 spectra is assessed. Lacking or low quality spectra entail re-analysis of the sample.

Here, a different prioritization strategy that ensures high quality MS2 spectra for all OMPs in water samples that potentially pose a risk for human or environmental health is proposed. The strategy is based on an innovative intelligent MS acquisition workflow and prioritizes compounds that potentially represent toxic OMPs online in the mass spectrometer. This workflow is based on MS1- and MS2-triggers; MS1-triggers prompt a MS2 event based on properties of the precursor ion in the full scan, MS2-triggers an additional MS2 event based on properties of the first MS2 spectrum, respectively.

Using a cheminformatics approach, potentially toxic compounds were selected based on the presence of structural alerts which are functional groups or substructures linked to a particular toxicological endpoint using ToxAlerts. The selected compounds were *in silico* fragmented with the fragmentation tools CFM-ID and MetFrag. Recurring masses and mass shifts (MS2-triggers) were identified in the *in silico* fragmentation spectra per structural alert using data mining strategies in R. Isotopic ratios of chlorine and bromine, and an inclusion list with m/z values of water-relevant pollutants were used as MS1-triggers.

The performance of MS1- and MS2-triggers was experimentally examined, as well as the effect of selected MS2 acquisition parameters on spectral quality and the use of a background exclusion list on spectral availability. The examined acquisition parameters included the fragmentation energy (CE) mode, the number of ions that are accumulated per scan (AGC-target) and the timespan in which they are accumulated (injection time). Higher AGC-targets resulted in an increase in the number of peaks and corresponding annotated fragments. Assisted CE increased fragment annotation compared to fixed and stepped CE. In stepped CE, an average spectrum is recorded of three different CEs whereas the optimal CE is selected by the mass spectrometer in assisted CE mode.

Regarding the MS1-triggers, the use of an inclusion list led to an increase in the percentage of compounds with an MS2, while the isotopic ratio did not have significant effects. The four tested MS2-triggers prompted additional MS2 scans for compounds that had a structural alert. Therefore, this intelligent acquisition method should be tested in high-throughput analyses and developed further to be able to apply it in routine monitoring studies for the detection of OMPs in drinking water sources.

**Keywords:** non-target screening, structural alerts, high resolution mass spectrometry, *in silico* fragmentation, prioritization, liquid chromatography, spectral quality

## Populair wetenschappelijke samenvatting

Het gebruik van door de mens gemaakte stoffen zoals medicijnen, bestrijdingsmiddelen, UV-filters in bijvoorbeeld zonnebrand en stoffen als PFAS, neemt toe. Hierdoor stijgt het voorkomen van deze chemicaliën en hun afbraakproducten in het milieu ook. De stoffen komen in de lucht, de bodem en het water terecht, vaak in lage concentraties ( $\mu\text{g/L}$ ) en worden organische microverontreinigingen genoemd. Een aantal van deze stoffen is giftig of wordt giftig zodra het afgebroken wordt in de natuur. Om te voorkomen dat deze stoffen in drinkwater terechtkomen wordt de inname van water uit de drinkwaterbronnen constant gemonitord op de aanwezigheid van giftige stoffen.

Identificatie van deze stoffen gebeurt meestal met behulp van vloeistofchromatografie gekoppeld aan massaspectrometrie. Hierbij worden de retentietijd, de massa en het fragmentatiespectrum van het molecuul gebruikt om de structuur te kunnen bepalen. Omdat dit water veel stoffen bevat die niet allemaal giftig zijn is het niet nodig om ze allemaal te identificeren. Momenteel wordt 'target-screening', 'suspect-screening' en 'non target-screening' toegepast om de focus op de giftige stoffen te leggen. Hierbij is het gangbaar dat van de stoffen die het meest intense signaal geven in het massaspectrum (MS1), een fragmentatiespectrum (MS2) wordt opgenomen. Maar ook bij deze strategieën lukt het vaak niet om van alle (giftige) stoffen een MS2 op te nemen dat goed genoeg is om de structuur van de stof op te helderen. Als gevolg moet het monster nog een keer gemeten worden om zo betere MS2's te verkrijgen. Hoe kan dit efficiënter?

Met de hulp van MS1- en MS2-triggers die samen zorgen voor de prioritering van potentieel giftige stoffen tijdens de meting. MS1-triggers activeren de opname van een MS2 scan zodra er stoffen gedetecteerd worden die een verdachte massa hebben (d.w.z. de massa staat op een lijst met giftige stoffen) en/of chloor of broom bevatten, ongeacht de intensiteit van het signaal.

MS2-triggers activeren de opname van een extra MS2 scan op basis van de aanwezigheid van 'structuur alerts'; moleculaire substructuren die gelinkt zijn aan de toxiciteit van een stof. Door tijdens de meting te bekijken of een stof een structuur alert heeft, kan op basis daarvan tijdens diezelfde meting een extra MS2 opgenomen worden in andere condities (bijv. hogere of lagere fragmentatie energie). Er wordt verwacht dat hierdoor meer stoffen geïdentificeerd kunnen worden op basis van één meting.

In dit onderzoeksproject werden stoffen uit de Amerikaanse ToxCast databank en de Europese NORMAN databank met een structuur alert virtueel (*in silico*) gefragmenteerd. De fragmentatiespectra werden gescreend op massa's en massaverschillen die karakteristiek zijn voor dat alert; de MS2-triggers. De MS1- en MS2-triggers werden toegevoegd aan de reguliere analysemethode van wateronderzoeksinstituut KWR en getest op oppervlaktewater- en waterzuiveringsinstallatie-influent monsters. Ook werd het effect getest van de verschillende instellingen voor de opname van een MS2 scan (maximale ion injectietijd, maximaal aantal ionen en verschillende fragmentatie energieën en -modi).

Uit dit onderzoek is gebleken dat een langere injectietijd en een hoger maximaal aantal ionen leiden tot meer fragmenten die geïdentificeerd kunnen worden. De MS1-triggers bleken deels effectief te zijn, maar er is meer onderzoek nodig om dit te optimaliseren. De MS2-triggers hebben extra MS2 scans geactiveerd voor de stoffen die de betreffende structuur alerts bevatten. Kortom, de ontwikkelde slimme acquisitiemethode bleek grotendeels succesvol en kan, mits verder ontwikkeld en getest, toegepast worden in de monitoring van microverontreinigingen in drinkwaterbronnen.

## Table of Contents

List of abbreviations.....	2
Abstract .....	3
Populair wetenschappelijke samenvatting .....	4
<b>1. Introduction.....</b>	<b>6</b>
<b>2. Theoretical background .....</b>	<b>7</b>
2.1 Non-target screening.....	7
2.2 High resolution mass spectrometry.....	9
2.3 Spectral libraries and chemical databases.....	12
2.4 <i>In silico</i> fragmentation.....	13
2.5 Structural alerts.....	14
2.6 Proposed method.....	15
<b>3. Materials &amp; Method.....</b>	<b>17</b>
3.1 Screening of compounds for structural alerts .....	17
3.2 <i>In silico</i> fragmentation.....	18
3.3 Pattern mining.....	18
3.4 Validation .....	19
3.5 HRMS method development.....	19
<b>4. Results .....</b>	<b>26</b>
4.1 Screening of compounds for structural alerts .....	26
4.2 <i>In silico</i> fragmentation.....	29
4.3 Pattern mining.....	30
4.4 LC-HRMS experiments.....	34
<b>5. Conclusions and future perspectives.....</b>	<b>47</b>
5.1 Conclusions .....	47
5.2 Outlook.....	47
<b>6. Acknowledgements.....</b>	<b>49</b>
<b>7. References.....</b>	<b>50</b>
Appendix A – Examples of structural alerts .....	53
Appendix B – Workflow screening with ToxAlerts and fragmentation.....	55
Appendix C – Comparison in- and output of ToxAlerts .....	56
Appendix D – Assays applied for toxicity validation.....	57
Appendix E – List of chemicals .....	59
Appendix F – Sequence lists.....	63
Appendix G – Chlorine and bromine distribution .....	71
Appendix H – Compound Discoverer workflow parameters.....	72
Appendix I – Spectral quality parameters acquisition experiments .....	76

## 1. Introduction

Issues with water quality occur worldwide due to the large spread of the human population and their extensive use of chemicals which lead to chemical pollution in a high number of water streams.<sup>1</sup> These streams cause distribution of the pollution with long-range effects, ultimately posing a threat to drinking water sources.<sup>2-4</sup> Various types of organic micropollutants (OMPs), i.e. anthropogenic chemicals that are present at trace levels ( $\mu\text{g/L}$ ) have been detected in ground and surface waters used for drinking water production. These OMPs include halogenated compounds such as poly- and perfluorinated alkylated substances (PFAS) and other types such as hormones, pharmaceuticals, UV filters and brominated flame retardants. Despite their low concentrations, OMPs can pose a risk to human and environmental health as they can be toxic, persistent or easily degraded into more toxic (bio)transformation products.<sup>5</sup>

To ensure drinking water quality, compounds that pose a potential health risk need to be monitored to be able to assess the actual human and environmental risks. Monitoring is typically performed using quantitative target analyses. As target analyses are limited to a set of known compounds, liquid chromatography coupled to high resolution mass spectrometry (LC-HRMS) based non-target screening (NTS) is often applied to more comprehensively monitor chemical water quality and broaden contaminant discovery.<sup>6-7</sup> However, the structural identification of unknown compounds from NTS data remains challenging due to the high number of features detected per experiment, and the need for high quality fragmentation spectra.<sup>8-9</sup> A feature is defined as an accurate mass combined with retention time and its intensity.

Here, these two challenges were addressed using intelligent acquisition in mass spectrometry; this strategy focuses on the features that are potentially toxic and thereby reduces the number of NTS features that need to be identified. This prioritization step occurs in the mass spectrometer during data acquisition instead of during data analysis which is the conventional offline prioritization strategy. The potentially toxic features are selected based on the presence of structural alerts. These potentially toxic features are fragmented using a number of different parameters to ensure high spectral quality. Together, this strategy can facilitate risk assessment through more efficient identification of compounds that pose a risk to human and environmental health.

The effects of selected acquisition parameters on the spectral quality and the resulting fragment annotation were tested as well. The acquisition parameters of interest were the type of collision energy (CE) (assisted or stepped), the use of a background exclusion list, the maximum ion injection time (IT) of the fragmentation scan and the automatic gain control-target (AGC-target) of this scan.

The aim of this research was to develop an online method suitable for the NTS of 'unknown unknowns' that prioritizes toxic compounds online in the mass spectrometer for fragmentation which leads to better spectra and consequently facilitates identification.

## 2. Theoretical background

### 2.1 Non-target screening

Three common approaches in the monitoring of organic micro-pollutants are target analyses, suspect screening and non-target screening. Target analyses require an in-house reference standard of the target compound, this standard must be measured under the same analytical conditions as the analyte. A calibration curve is generated on which the quantitative analysis is based on. Moreover, target methods are optimized, including buffers, LC gradient, ionization source and mass range for the selected target compounds. For both NTS and suspect screening, the whole mass range is acquired with a generic method that is possibly not optimal for every compound. All compounds that are separable with the chosen chromatography and ionizable with the selected ion source can be detected with such a method. As a result, a broad screening is achieved and a large amount of data is generated. A suspect screening can be performed on these data which is based on prior information of compounds that are expected to be present in the sample, such as exact mass, isotopic pattern and molecular structure. The NORMAN Suspect List (SusDat) is a list that contains this prior information.<sup>10</sup> When no prior information is available, NTS is performed, which is comparable to suspect screening but with a larger database (all possible molecules). Several levels of confidence for these different types of identification via HRMS data can be distinguished (see *table 1*).<sup>11-12</sup>

*Table 1 – Levels of confidence for identification of small molecules via HRMS from Schymanski et al.<sup>11</sup>*

Identification level		Minimum data requirements	
Level 1	Confirmed structure	Reference standard is available and measured under the same conditions with MS, MS/MS and retention time match.	MS, MS2, RT, reference standard
Level 2	Probable structure	a. Library: evidence via a spectrum-structure match from literature or a library spectrum.	MS, MS2, library MS2
		b. Diagnostic: no other structure fits the experimental information, but no reference standard or literature information is available.	MS, MS2, experimental data
Level 3	Tentative candidate(s)	Evidence for a few possible structure(s), but not sufficient to determine the exact structure.	MS, MS2, experimental data
Level 4	Unequivocal molecular formula	Only exact molecular formula is known.	MS isotope/adduct
Level 5	Exact mass (m/z)	Unambiguous information about structure or formula does not exist	MS

The general workflow for NTS is illustrated in *figure 1a*. It is not yet feasible to identify all hundreds to thousands, depending on type of water sample, peaks in a generated NTS spectrum of a sample, therefore prioritization strategies are needed. Prioritization is the selection of peaks of interest based on intensity, occurrence, persistence or potential toxicity. To date, prioritization strategies are applied offline during the data-analysis such as prioritization based on the presence in a suspect list or on ToxCast toxicity data.<sup>9, 13</sup> The disadvantage of offline prioritization is that this is performed during data analysis. If prioritized features lack a MS2 fragmentation spectrum or the spectrum is of insufficient quality to allow structural identification of the feature, the sample has to be re-analyzed. This research project addresses the issue of uninformative MS2 fragmentation spectra by online

prioritization of features leading to higher spectral quality, see *figure 1b*. A feature is defined as an accurate mass together with its retention time and intensity, see *figure 2*.

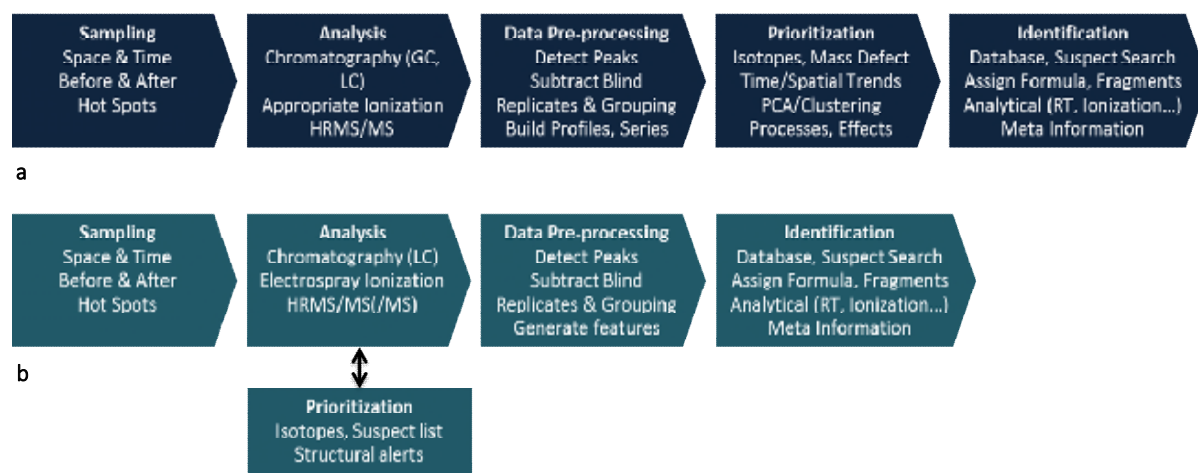


Figure 1 – a) typical workflow of NTS by Hollender et al. (2019), the prioritization step is performed offline.<sup>14</sup> b) workflow of this project, the prioritization step is performed online, during the analysis.

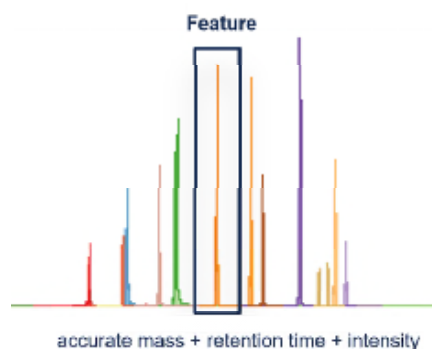


Figure 2 – Definition of a feature by Brunner (2019).<sup>15</sup>

Other approaches that are currently in use for prioritization and detection of unknown unknowns, but time- and source-consuming, are Effect Directed Analysis (EDA) and Toxicity Identification Evaluation (TIE).<sup>16</sup> The aim of TIE is to detect toxicity and ecological relevance of contaminated water and sediment samples. The aim of EDA is to detect which chemicals are responsible for the bioactivity and overall endpoint-specific activity. EDA is based on fractionation of samples to reduce the complexity of toxic mixtures and measuring the biological response in *in vitro* or *in vivo* assays. EDA is applied in water resource monitoring.<sup>16</sup> Both regular EDA and TIE are not performed online which makes them less suitable for routine monitoring. Currently, high-throughput effect-directed analysis (HT-EDA) is being developed. This technique combines NTS with EDA and is more suitable for routine monitoring as it increases the throughput and the fractionation resolution.<sup>17-18</sup>

This study tries to overcome the problems associated with the high costs and laboriousness of NTS offline prioritization by using intelligent MS acquisition methods based on computational strategies to give preference to the fragmentation of potentially toxic compounds, based on structural alerts. Structural alerts are molecular substructures that are linked to the toxicity of a molecule. A method involving structural alerts can be performed online and in routine monitoring studies. Therefore, this prioritization technique might be more cost-efficient and less time- and labor-consuming and easy to adjust when new structural alerts are discovered. Therefore, this method should make identification more efficient; potentially toxic compounds that contain structural alerts are automatically selected.

## 2.2 High resolution mass spectrometry

Mass spectrometry is a detection technique that separates ions based on their mass-to-charge ratio,  $m/z$ . A mass spectrometer is typically composed of an ion source, a separator and a detector. The ion source ionizes the analytes that enter the mass spectrometer, these ions are separated by the mass separator and detected in the detector. The output of a mass spectrometer is a mass spectrum, containing the number of ions per  $m/z$  value. High resolution mass spectrometry is a technique with high mass accuracy ( $\pm 0.001$  Da) and high mass resolution ( $m/\Delta m \geq 20\,000$ ).<sup>8</sup> This technique is very suitable for NTS of environmental contaminants as it is sensitive and thus can detect thousands of (trace) compounds in one sample within short time frames.<sup>14</sup> Moreover, the high resolution facilitates determination of the elemental formula.

### Mass analyzers

The quadrupole mass analyzer consists of four parallel hyperbolic rods to which oscillating (time-dependent) and static (time-independent) electric fields are applied.<sup>19</sup> As a result, ions are separated based on their different stabilities depending on their  $m/z$  value, only ions that have a stable trajectory towards the z-direction (which is parallel to the quadrupole rods) will reach the detector. Quadrupole mass analyzers have a limited mass range (up to  $10^3$ ), limited resolution ( $10^3$ - $10^4$ ) and accuracy (100 ppm). The quadrupole is used to get rid of noise and ions in the mass range that is out of interest.

The linear ion trap analyzer is a three dimensional version of the quadrupole where an oscillating electric field (generated by four rods) traps the ions.<sup>19</sup> The electric field is present in the axial dimension and at the two end-caps of the trap. The ions are repelled inside the trap (the closer the ion is to the rods, the more it is repelled) leading to the formation of an ion cloud in the center of the trap. The ions can be ejected selectively via axial injection (parallel to the axis of the trap) or radial ejection (perpendicular to the axis of the trap). Ion traps are sensitive, but have a low resolution and low mass accuracy (100 ppm).

The Orbitrap is a mass analyzer with a high resolution and high accuracy ( $< 5$  ppm), but the higher the accuracy, the higher the resolution has to be and the more scan-time is required, see *table 2*. The Orbitrap consists of two electrodes (see *figure 3*); an inner, spindle-shaped, electrode which generates an axial field gradient and an outer electrode that is split in two.<sup>19</sup> The ions are axially injected into the Orbitrap via the C-trap (see *figure 3*) and make circular or oval trajectories around the electrode while oscillating along the z-axis. The moving ions induce an image current that is detected by an amplifier in the split between the two halves of the outer electrode. The  $m/z$  values of the ions are determined after Fourier transformation.

*Table 2 – Trade-off between resolution and scan time from University of Washington’s Proteomics Resource (UWPR).<sup>20</sup>*

Resolution at $m/z$ 200	Transient length (ms)	Approximate scan speed [Hz]	“Free” ion time (ms)
15,000	32	NA	22
30,000	64	15	54
50,000	96	NA	86
60,000	128	7.5	118
120,000	256	4	246
240,000	512	2	502
450,000	1024	<1	1014



## Tandem mass spectrometry

Tandem mass spectrometry involves additional mass analysis stages. It is applied to obtain more information of an ion besides the  $m/z$  value obtained in a MS1 scan. In tandem mass spectrometry the precursor ion, detected in the MS1 scan, is selected and fragmented. The fragmentation pattern of a compound reflects its structure. These fragments can thus be used for structural elucidation through matching of the experimental fragmentation spectrum with spectral library entries or *in silico* predicted fragmentation spectra. The first analysis stage (in this case the quadrupole) is used to select a precursor ion, which is further fragmented in the second stage, MS2. These fragment ions can be further fragmented and analyzed in MS3 or MS<sub>n</sub> experiments.

Fragmentation reactions consist of homolytical cleavages, heterolytical cleavages and rearrangement reactions of the molecular ion, leading to different fragments. Some reactions are favored at higher energies whereas others are favored at lower fragmentation or collision energies (CEs). One way of fragmenting ions is via collision-induced dissociation (CID) where the ions are accelerated by a higher electrical potential leading to higher kinetic energies.<sup>19</sup> The accelerated ions collide with noble gas atoms such as helium or small molecules such as N<sub>2</sub> and O<sub>2</sub> and the kinetic energy is converted into internal energy. This leads to dissociation of the ion into fragments and the fragments are analyzed. Higher-energy collisional dissociation (HCD) is typically used for smaller molecules and occurs in the HCD cell (or Ion Routing Multipole IRM). The radiofrequency voltages are increased aiming to retain the maximum number of fragment ions.<sup>21</sup>

In small molecule identification, acquired fragmentation spectra are matched with *in silico* predicted- or library spectra. The more (true) fragments that can be matched, the higher the confidence of identification. Therefore, high spectral quality of the MS2 and potentially MS<sub>n</sub> spectra is required for proper structure elucidation, as it leads to a better sensitivity (more true positives) and better specificity (more true negatives). Spectral quality can be affected by e.g. signal distortion and electrical noise.<sup>19</sup> Moreover, CEs have a big influence on the number of fragments in an MS2 or MS<sub>n</sub> spectrum and thus the quality of these fragmentation spectra. At higher CEs, the ion is fragmented more heavily than at lower CEs, and different molecules have different optimal CEs. Too high CEs lead to uninformative fragments that are unspecific, whereas too low CEs lead to too few fragments that provide insufficient information for structure elucidation. CE optimization is therefore a prerequisite for high spectral quality. Also, combinations of different CEs can be used to generate more informative spectra (see Acquisition parameters section below).

Selection of peaks that have to be fragmented can be done via data-independent acquisition (DIA) or data-dependent acquisition (DDA). In data-independent acquisition all parent ions within a given  $m/z$  range are fragmented and a MS<sub>n</sub> spectrum is recorded. The spectra acquired with DIA are multiplexed due to the wide  $m/z$  range that is fragmented. As a result, data analysis and spectral matching are more complicated. Data-dependent acquisition selects the most intense peaks (parent- or precursor ions) for further fragmentation. However, compounds with low peak intensities in the mass spectrum can be highly toxic as well and thus pose a risk to human and/or environmental health. These currently risk to not be fragmented using the standard DDA-approach. Moreover, as mass spectrometry is not inherently quantitative, the signal intensity does not necessarily reflect the concentration of the compound in a sample. Actual concentrations might be higher or lower than expected from the detected signal intensity due to the compound's ionization efficiency. For instance, the poor ionization efficiency with ESI of delta-5 steroids, 5 $\alpha$ -reduced androgens and estrogens results in low signal intensities of these compounds even if present at relevant concentrations.<sup>22-23</sup>

This indicates the need for prioritization strategies that are not based on signal intensity, such as inclusion lists. Using an inclusion list for the selection of precursor ions for fragmentation is a hypothesis-driven strategy that is often applied in LC-MS/MS, for instance in directed proteomics.<sup>24</sup> It uses  $m/z$  values and retention time (optionally)<sup>24</sup> of compounds or features of interest.

### Experimental set-up used

The Orbitrap Fusion Tribrid mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) was used for NTS of organic micropollutants. This instrument was chosen for its high resolution that allows elemental formula determination, high speed and sensitivity that allow multiple scan events, and the availability of an ion trap that allows assisted CE and fragmentation trees. The instrument combines three mass analyzers (see *figure 3*): the linear quadrupole, Orbitrap and linear ion trap. It has a maximum resolution of  $m/\Delta m$  560 000. In the experiments presented here, it was operated at  $m/\Delta m$  120 000 and 240 000, the MS2 resolution was set at  $m/\Delta m$  15 000. The mass spectrometer was connected to a reversed-phase liquid chromatography (RPLC) system and equipped with a heated electrospray ionization (ESI) source, a soft ionization technique.

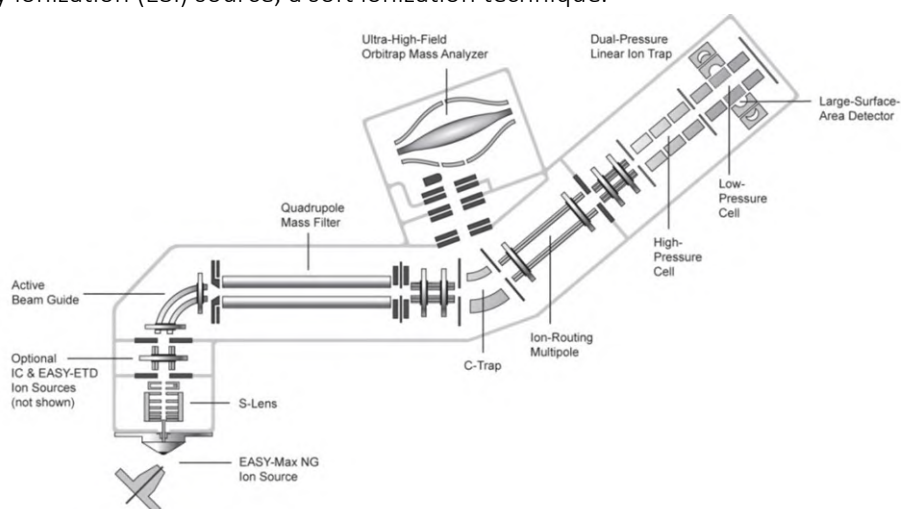


Figure 3 – Schematic representation of the Orbitrap Fusion Tribrid mass spectrometer, Thermo Fisher Scientific.<sup>25</sup>

### Acquisition parameters

There are three collision modes implemented in the Orbitrap Fusion for both CID and HCD fragmentation; fixed, stepped and assisted CE (ACE). In fixed CE mode all ions are fragmented with the same CE. In contrast, in stepped CE mode, fractions of the precursor are fragmented at a defined number of different CEs, fragments are pooled and analyzed in the Orbitrap analyzer. With ACE, the precursor is successively fragmented with multiple pre-defined HCD energies and analyzed in the ion trap. The optimal CE is determined at which the precursor ion is present at a defined intensity, for instance 10% of its original intensity, and used for a final analytical scan in the Orbitrap.<sup>26</sup>

The mass spectrometry acquisition software allows multiple acquisition parameters to adjust for a specific sample type and/or study goal. These parameters each affect the type and number of spectra that is recorded, as well as the spectral quality. The parameters automatic gain control (AGC), maximum ion injection time (IT) and CE mode are described in this section with regard to small molecule analysis in aqueous samples.

The AGC-target defines the number of ions to accumulate in the C-trap before the MS<sub>n</sub>+1 scan is performed.<sup>27</sup> The timespan in which this accumulation is allowed is defined as the maximum IT. The

AGC-target and maximum IT are related to each other; a MSn scan is triggered in case one of these values is reached. In a MS1 scan with a max IT of 200 ms and an AGC-target of  $5 \times 10^4$ , the MS2 scan is recorded after 150 ms if the AGC target is achieved at 150 ms. If the AGC-target is not reached, the MS2 scan will be recorded at 200 ms. There is a trade-off between the quality of the spectra and the number of scans available; more additional MS2 scans can be acquired if the AGC-target and/or IT are set low but also the spectral quality will be lower.<sup>27</sup> If the AGC target is set too high, space-charge effects can lead to higher mass errors.

The Orbitrap Fusion method editor also gives the possibility to in- or exclude certain m/z values (together with a retention time if desired) for/from DDA. Thermo Fisher's AcquireX<sup>28</sup> software can be used to automatically create a background exclusion list that includes all m/z values and their retention time detected in a previous blank sample. This background exclusion list is then used during the analysis to exclude background ions from fragmentation events and increase the MS2 percentage of non-background ions. Since there is a maximum number of MS2 scans that can be recorded in the duty cycle, it is disadvantageous to record MS2 scans of background ions that are not of interest.

### 2.3 Spectral libraries and chemical databases

The different levels of identification described by Schymanski et al. are shown in *table 1*.<sup>11</sup> To be able to reach confidence level 2 or 3, 'tentative candidate(s)' or 'probable structure', respectively, spectral matching is required, where experimental spectra are compared to library spectra or *in silico* predicted spectra. To this end, spectral libraries, *in silico* fragmentation predictors, and software that implements spectral matching have been developed. These tools can connect the experimentally obtained mass spectrum with candidate structures.

Spectral libraries occur in both commercial, such as mzCloud<sup>29</sup>, and open source forms, such as MassBank<sup>30</sup>. mzCloud is a spectral library that contains HRMS spectra, both MS2 and MSn, in so-called fragmentation trees.<sup>29</sup> These fragmentation trees contain several fragmentation spectra of the same precursor that are obtained at multiple different CEs. mzCloud is available online (<https://www.mzcloud.org>) and in Compound Discoverer (Thermo Fisher Scientific), both the online library as its offline version mzVault.<sup>31</sup> mzVault has the possibility to add in-house acquired spectra. A disadvantage of mzCloud is that it only contains spectra acquired with Orbitrap mass analyzers. In contrast, e.g. MassBank Europe by the NORMAN network ( $n_{\text{MassBank}} = 2304$ ) contains also data acquired with QTOF mass analyzers and is an open source spectral library that can be accessed freely and is constantly developing as well.<sup>30</sup> NORMAN MassBank is directed towards environmentally relevant contaminants<sup>32</sup> but has a lower data curation level than mzCloud.

Compounds that are not included in spectral libraries, but are present in suspect lists and chemical databases can be fragmented *in silico*. The resulting predicted fragmentation spectra can then be compared to the experimental spectra. Suspect lists and chemical databases occur in various sizes. The NORMAN network for example hosts the NORMAN Substance Database (SusDat,  $n_{\text{SusDat}} = 65697$ )<sup>33</sup> which is composed of various environmentally relevant suspect lists such as the STOFFIdent<sup>34</sup> and KWR Sjerps lists<sup>35</sup> with water relevant compounds ( $n_{\text{Sjerps}} = 5722$ ), and the UBAPMT list<sup>36</sup> with REACH substances that are (very) persistent, (very) mobile and toxic ( $n_{\text{UBAPMT}} = 240$ ). These lists provide information such as chemical names, CAS registry numbers, structure (SMILES, InChI, InChIKey), retention time index,  $[M+H]^+$  and/or  $[M-H]^-$ .

The Chemistry Dashboard<sup>37</sup> is a chemical database held by the U.S. Environmental Protection Agency and is i.a. linked to the toxicity database ToxCast.<sup>38</sup> The latter holds *in vitro* toxicity information of

many chemicals based on their response to various bioassays, this database is constantly growing. The version used in this project contained 9224 chemicals.

ChemSpider<sup>39</sup> is a chemical database that covers a larger chemical space than SusDat and currently consists of 84 million entries. ChemSpider is often used in annotation software such as MetFrag<sup>40</sup> or Compound Discoverer<sup>31</sup> to identify detected features.

## 2.4 *In silico* fragmentation

*In silico* fragmentation techniques are often applied in compound annotation; candidate structures are fragmented using computational strategies and their *in silico* fragmentation spectra are compared with the measured spectrum. *In silico* fragmentation can also be used for prediction of the fragmentation spectra of molecules with a structural alert, without comparison to experimental spectra. Rule-based fragmentation and combinatorial fragmentation are the most common types of algorithm used for *in silico* fragmentation of molecules.<sup>41</sup> The rule-based fragmentation tools predict MS2 spectra using manually created fragmentation rules. When using combinatorial fragmentation, all bonds of a molecule are broken systematically where bond energies can be taken into account. In this project, two types of fragmentation software were applied; MetFrag and CFM-ID 2.0.

### *MetFrag*

MetFrag is a combinatorial fragmentation predictor designed for the matching of experimental spectra to a candidate structure retrieved from databases such as KEGG, PubChem, ChemSpider or an uploaded structure data file (SDF).<sup>40</sup> MetFrag fragments these candidate molecules *in silico* using a bond dissociation approach. This means that each possible bond of the molecule is broken, and tree depths can be chosen. Five neutral loss rules are applied to consider rearrangement reactions. The resulting spectra are ranked based on the intensity, m/z values and bond dissociation energy of the matched peaks to find the best spectrum and compound match. In this project, no candidate matching was required. Instead, the R-package metfRag<sup>42</sup> (version 2.4.2) was used to generate fragments for molecules with a structural alert.

### *Competitive Fragmentation Modeling (CFM)*

CFM is another combinatorial fragmentation predictor but different from MetFrag. CFM consists of two methods for ESI-MS/MS CID fragmentation; single energy competitive fragmentation modeling (SE-CFM) and combined energy competitive fragmentation modeling (CE-CFM).<sup>41</sup> The SE-CFM model is a stochastic, homogeneous Markov process, where the probabilities of the fragmentation process are defined by a transition model. The model makes a few assumptions for the fragmentation process such as that the molecule needs to be singly positive charged, collision yields two fragments of which one is neutral and the other has a single positive charge, removal or addition of sigma bonds during a break is not allowed, and the valence and even electron rules must be satisfied in all fragments.<sup>41</sup> In fact, the method is similar to MetFrag apart from these assumptions. The CE-CFM model combines the information of multiple energies and is therefore more complex but does not show better results according to the designers.

Compared to MetFrag, CFM-ID 2.0 performs better at compound identification but the computation time is longer.<sup>40</sup> Ruttkies and coworkers (2016) concluded that the combination of MetFrag with CFM-ID 2.0 gave better results for compound identification.<sup>40</sup> CFM-ID 2.0 won the Critical Assessment of

Small Molecule Identification (CASMI) contest in 2014.<sup>43</sup> This contest was developed to assess the performance of annotation software for mass spectrometry data.

Recently, a new version of CFM-ID with higher performance has been developed. CFM-ID 3.0 uses a rule-based approach instead of the combinatorial approach but no Windows executables were available yet.<sup>44</sup> So CFM-ID 2.0 was applied in this project, using 'cfm-predict.exe' which predicts the spectrum for an input molecule based on the pre-trained SE-CFM model.<sup>45</sup> The program output consists of three m/z and intensity lists, for low energy CID (10 V), medium energy CID (20 V) and high energy CID (40 V). These energies reflect the type of spectra where the model is based on. CFM-ID is based on CID QTOF data which is comparable to HCD data from an Orbitrap instrument (beam-type CID).<sup>46</sup>

## 2.5 Structural alerts

Structural alerts, often referred to as toxicophores or expert rules, are molecular (sub)structures related to the toxicity of a chemical, see *figure 4*. The presence of halogens is for example often related to toxicity, such as fluor in GenX of the PFAS class.

Historically, structural alerts have been derived manually based on expert knowledge, today they can be derived computationally. To this end, several databases and software programs have been developed, such as ToxAlerts<sup>47</sup>, DEREK<sup>48</sup>, and MultiCASE<sup>49</sup>. These are often applied as a tool in the development of pharmaceuticals for prediction of potential drug toxicity using read-across. Read-across is a technique for predicting toxicity for one compound based on data of the same toxic endpoint for other compounds.<sup>50</sup>

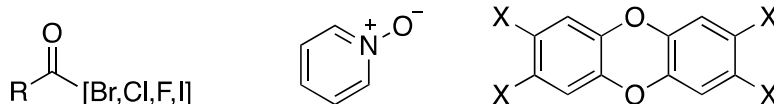


Figure 4 Examples of structural alerts; acyl halides, aromatic ring N-oxide and dioxin-like structures.<sup>51</sup>

Toxic endpoints are defined as the recorded observation or measured biological effect in a toxicity test (*in chemico*, *in vitro* or *in vivo*).<sup>52</sup> Most structural alerts are derived from the endpoints carcinogenicity and mutagenicity. Several lists of structural alerts for carcinogenicity and mutagenicity have been published, the most common are generated by Ashby and Tennant (1998)<sup>53</sup>, Baily et al. (2005)<sup>54</sup> and Kazius et al. (2005 & 2006).<sup>55-56</sup> Benigni and Bossa (2008)<sup>51</sup> compared those lists and found agreement of 65% with rodent carcinogenicity data and 75% with *Salmonella* mutagenicity data. They generated a revised list of 33 structural alerts which is given in *appendix A*. These alerts are included in the ToxAlerts database. Other toxic endpoints are examined less extensively, but they might be relevant as well. The hazardous chemicals in water are usually present at low concentrations and toxicity occurs in the long term.<sup>57</sup> Drinking water-relevant toxic endpoints are genotoxicity, carcinogenicity, mutagenicity, endocrine disruption, neurotoxicity and developmental toxicity.<sup>57</sup>

Structural alerts are useful to apply in the 'rough' selection of compounds that need to be identified in NTS methods, as compounds with a structural alert could potentially be toxic. The presence of one or more structural alerts can be used as a trigger for further fragmentation events. A disadvantage of using structural alerts in NTS for the selection of features is that initially non-toxic compounds that are bioactivated once taken up by an organism will not be selected for further identification. However, if known, these compounds could be added to the target- or suspect screening list.

Another remark on this technique is that the total molecular structure is not considered, this can result in potentially overlooking mutually interfering functional groups that affect the total potency of the compound.<sup>58</sup> Other modulating factors that influence the biological activity of the chemical are:<sup>59</sup>

- Molecular weight. The molecule is less easily absorbed with increasing molecular weight and size.
- Physical state of the compound. This determines the chances of the compound to reach its target sites.
- Solubility. Highly hydrophilic chemicals are rarely absorbed and easily excreted.
- Chemical reactivity. Reactive compounds may react with other compounds before they reach their target site.
- Geometry. This determines the fit to the target site.

Some of these limitations are less relevant, for instance solubility, as highly hydrophilic chemicals will not be detected when RPLC using a C18 column is applied because they are expected to elute at  $t_0$ .

## 2.6 Proposed method

Here, an intelligent acquisition method that online prioritizes potentially toxic features and ensures availability and good quality fragmentation spectra of the prioritized features is proposed. To this end, the concept of MS1- and MS2-triggers is introduced. MS1-triggers are specific properties of an ion detected in the full scan MS1 spectrum that suggest potential toxicity. These properties are defined as the presence of chlorine and/or bromine and masses that occur in suspect lists. If an ion is detected in the MS1 spectrum that has an isotopic pattern characteristic for chlorine and/or bromine, or when the  $m/z$  value of the ion is present in the suspect list applied, a MS2 scan is triggered. MS2-triggers are specific fragment masses or deltas in the recorded MS2 spectrum that are indications for the presence of a structural alert. If such a fragment or delta is detected in the MS2 spectrum, an additional MS2 scan is triggered.

*Figure 5* is a schematic overview of the proposed workflow of the NTS intelligent acquisition method with online prioritization. Analytes are separated by reversed-phase liquid chromatography and transferred to the mass spectrometer where a MS1 full scan is taken. Ions are selected for a MS2 scan if a MS1-trigger is present, otherwise/additionally the top  $n$  most intense peaks are selected. This MS1-trigger is defined as the presence of a chlorine or bromine isotopic pattern and/or a  $m/z$  value of the inclusion list with  $m/z$  values of suspect compounds. Once a MS2 scan is recorded, it is screened for the presence of a recurring fragment or delta corresponding to a structural alert. An additional MS2 scan is taken if such a MS2-trigger is present (*figure 5, scenario 1*). This additional MS2 scan is recorded with different settings such as a longer maximum IT, different CE or a different CE mode.

The added value of this intelligent acquisition method is that prioritization happens online, during the MS measurement and not after data analysis. Potentially toxic compounds are prioritized and have a higher probability to be fragmented (MS1-trigger) or receive an additional MS2 scan (MS2-trigger). This results in prioritized features with sufficient, good quality MS2 information which in turn facilitates their identification. Re-analysis of samples becomes obsolete and identification ultimately more efficient.

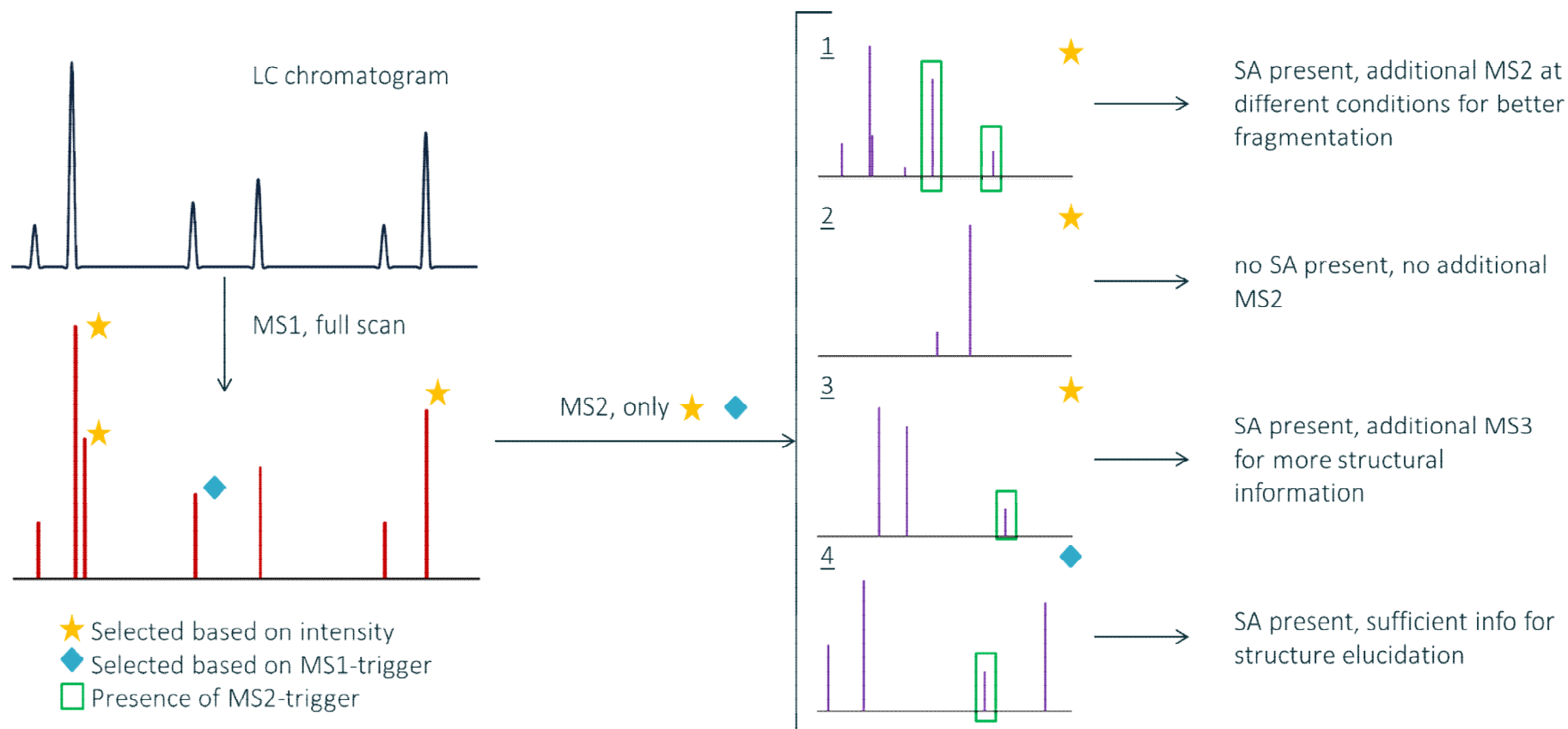


Figure 5 – Schematic representation of the proposed LC-HRMS/MS workflow using intelligent acquisition based on structural alerts. A full MS1 scan is taken after chromatographic separation and the peaks are screened for their intensity and the presence of MS1-triggers (blue diamond marker). The most intense peaks (based on DDA-approach, yellow star marker) and those that contain a MS1-trigger are selected for a MS2 scan. The MS2 scans are screened for MS2-triggers indicating the presence of a structural alert resulting in four possible scenarios:

1. SA is present, so an additional MS2 scan at different conditions is taken.
2. No SA present, structure identification is not necessary, no additional MS2 is taken.
3. SA is present, but an additional MS3 scan is necessary for complete structural elucidation (MS3 scans are outside the scope of this project).
4. SA is present, the spectral quality is sufficient for structure elucidation (online spectral quality assessment is outside the scope of this project).

### 3. Materials & Method

The complete strategy used to develop the intelligent acquisition method using both cheminformatics and LC-HRMS experiments is illustrated in *figure 6*. First, cheminformatics were applied to retrieve compounds with structural alerts, *in silico* fragment these compounds, mine for recurring fragments and deltas in the predicted spectra, and ultimately determine MS1- and MS2-triggers. LC-HRMS experiments were then performed to assess the performance of both trigger types, and in addition investigate the effect of selected acquisition parameters on MS2 spectral quality.

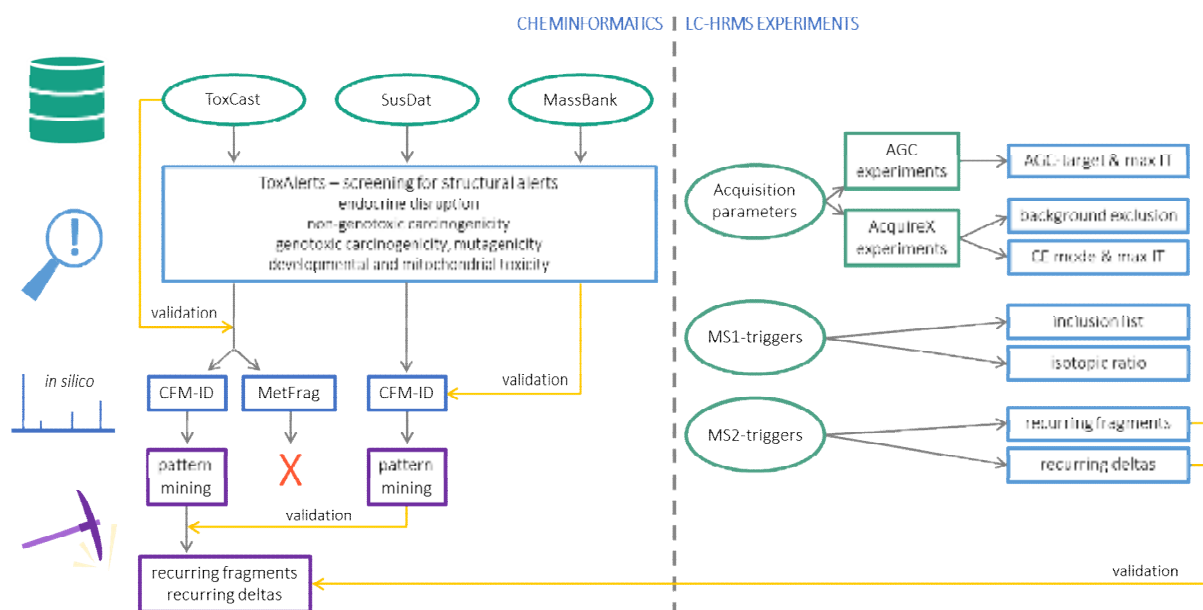


Figure 6 – Schematic overview of the strategy that was used to develop the intelligent acquisition method, both cheminformatics (left) and LC-HRMS experiments (right) were applied.

#### 3.1 Screening of compounds for structural alerts

The workflow for the screening and fragmentation of the ToxCast dataset is given in *appendix B*. First, the CAS registry numbers of the 9224 compounds registered in the ToxCast data file Chemical\_Summary\_190708.csv<sup>60</sup> were converted into 7571 unique MS-ready SMILES using the Chemistry Dashboard.<sup>37</sup>

Four toxic endpoints were selected for screening with ToxAlerts: ‘endocrine disruption’ (edc), ‘non-genotoxic carcinogenicity’ (ngc), ‘genotoxic carcinogenicity, mutagenicity’ (gcm) and ‘developmental and mitochondrial toxicity’ (dmt). These endpoints and their corresponding 187 structural alerts were chosen based on their relevance for drinking water and potential human health risk.<sup>57</sup> The endocrine disruption alerts used in this study belong to both estrogenic and androgenic endocrine disruptors.<sup>61</sup> This selection was made based on *in vitro* and *in vivo* (mammalian) data. As endocrine compounds are poorly ionizable, it is questionable whether these compounds can be detected in ESI-MS/MS.

The output of ToxAlerts was formatted in R<sup>62</sup> (version 3.6.1 (2019-07-05) -- "Action of the Toes") for fragmentation with MetFrag and CFM-ID. A .txt file was generated per structural alert containing the InChIKey and SMILES code for suitability with CFM-ID 2.0.



### 3.2 *In silico* fragmentation

The compounds with a structural alert were *in silico* fragmented with MetFrag in R using the R-package metfRag<sup>42</sup> (version 2.4.2) and CFM-ID 2.0 using the command line. The command-line utility cfm-predict.exe was used to generate fragments with CFM-ID 2.0, the standard trained CFM model and its standard configuration parameters were used, see *appendix B*. The post processing option was not included and the probability threshold was set to 0.001 (default setting). The output was processed in R. MetFrag is not designed for fragmentation of molecules only, so a separate script was written to process the generated fragments and add a proton (1.00727646677 Da) to generate the [M+H]<sup>+</sup> mass.

### 3.3 Pattern mining

The results of the *in silico* fragmentation of compounds with a structural alert were screened for characteristic patterns, that is, recurring fragments and recurring mass shifts (deltas). All structural alerts with more than 4 molecules were included in the analysis. Both the CFM-ID dataset and the MetFrag dataset were screened, with the control set being the MS2 spectra of all molecules for each method. For CFM-ID, to be able to compare the effect of the three energy levels on the recurring fragments and deltas, an intensity threshold was set at minimal 5% of the maximum peak intensity (100). The energy levels had an effect on the signal intensity only and not the m/z values of the signal. So no energy effects could be taken into account in the analysis without setting this threshold value of 5%. The spectra were screened for recurring fragments by counting the presence of all m/z values, regardless of the peak intensity, as long as it was  $\geq 5\%$  of the maximum peak intensity. A code snippet with the defined function is shown in *figure 7*.

```
# Define function to count recurring fragments
getCount <- function(x) {
  u <- unique(x)
  fragments_freq <- data.frame(
    mZ = u,
    count = pbsapply(u, function(v) {
      length(which(x == v))
    })
  )
  fragments_freq <- arrange(fragments_freq, desc(count))
}
```

Figure 7 – R-script for the function `getCount()`, where the frequency of each fragment within all MS2 spectra of the same structural alert is calculated. *x* represents a numeric vector with m/z values of all the generated fragments.

The frequency of each m/z value recurring within the MS2 spectra of the molecules of one structural alert was calculated. The same was done for the MS2 spectra of the total fragmented dataset and these frequencies were compared.

Screening for recurring deltas between fragments was performed in R as well, the frequencies of these deltas were calculated and compared with the frequencies of the deltas occurring in the total dataset. A code snippet with the defined function is shown in *figure 8*.

An extra control step for the frequencies was performed by taking a random sample ( $n = 3953$ ) from the NORMAN SusDat ( $n_{\text{total}} = 65697$ ) that was not screened for structural alerts. The sample was fragmented with MetFrag and CFM-ID as well, following the same approach. Next, the frequencies of recurring fragments and recurring deltas within this random sample were compared with the frequencies within MS2 spectra of compounds with structural alerts derived from ToxCast.

```

# Define function to calculate the delta's
getDelta <- function(y) {
  delta <- c()
  if (length(y) >= 2) {
    for (p in 1:(length(y) - 1)) {
      first_position <- p + 1
      for (q in first_position:length(y)) {
        delta_tmp <- y[q] - y[p]
        delta_tmp <- round(delta_tmp, 5)
        delta <- c(delta, delta_tmp)
      }
    }
  }
  return(unique(delta))
}

```

Figure 8 – R-script for the function *getDelta()*, where all possible deltas of each MS2 scan are calculated. *y* represents a numeric vector with *m/z* values of all the generated fragments.

### 3.4 Validation

#### *Validation of structural alerts with ToxCast toxicity data*

ToxCast assays relevant for the endpoints that were linked to the structural alerts were selected based on literature.<sup>9</sup> These assays are listed in *appendix D*. The AC<sub>50</sub> values of the ToxCast compounds with an alert were obtained from ‘ac50\_Matrix190708.csv’ (downloaded at 04 December 2019).<sup>63</sup> In this file, non-active compounds are given an AC<sub>50</sub> value of 1.000000e+06. Lower values indicate that the compound is active. Per toxic endpoint, i.e. endocrine disruption, developmental and mitochondrial toxicity, non-genotoxic carcinogenicity and genotoxic carcinogenicity, mutagenicity the percentage of molecules with both a structural alert and activity in one of the specified assays was calculated. This percentage was compared to the percentage of active compounds for the total ToxCast dataset, irrespective of the presence of a structural alert.

In ToxCast, some MS-ready SMILES codes occur multiple times (but with a different DSSTox Substance identifier) with in some cases varying toxicity information. The toxicity validation was based on these DSSTox Substance ID to include all bioassay results for the same MS-ready SMILES and prevent information loss.

#### *Validation of in silico predicted fragmentation spectra with experimental spectra*

The fragmentation results were validated with experimental data obtained from the NORMAN MassBank (MassBankEU).<sup>30</sup> As this dataset is composed of experimental data originating from various laboratories, results are not completely comparable and contain experimental errors. Therefore, it was not possible to perform the same structural alert analysis of screening for recurring fragments and deltas in the MassBank data. However, the MassBank fragments were used to validate the CFM-ID results of SusDat. MassBank data was available for 2.25% of the fragmented molecules from SusDat. Another part of the validation included the experimental data obtained during the MS2-trigger experiments. The computationally derived fragments and deltas were compared with these experimental results.

### 3.5 HRMS method development

#### *Sample preparation*

The chemicals used in this study are listed in *appendix E*. An internal standard mixture of atrazine-d5 (CDN isotopes, Pointe-Claire, Canada), benzotriazole-d4 and bentazon-d6 (LGC Standards, Wesen, Germany) at a final concentration of 1 µg/L was added to each sample. Surface water samples

(Lekkanaal, the Netherlands), wastewater treatment plant (WWTP) influent samples, spiked surface water samples and spiked WWTP-influent samples were filtered using 0.2  $\mu\text{m}$  Phenex<sup>TM</sup>-RC 15 mm Syringe Filters (Phenomenex, Torrance, USA) prior to analysis. The WWTP-influent samples were 10 times diluted. The blanks used for these analyses were filtered as well. The spiking solution 'LOA-600 + specials' was added to the samples to final concentrations of 10  $\mu\text{g/L}$ , 1  $\mu\text{g/L}$ , 100  $\text{ng/L}$ , 10  $\text{ng/L}$  and 1  $\text{ng/L}$  (see sequence list in *appendix F*).

### *Acquisition parameters*

Two sets of experiments were performed to determine the effect of a background exclusion list using the AcquireX acquisition software (Thermo Fisher Scientific Inc.), stepped and assisted CE, the AGC-target and maximum IT for the MS2 scan. The methods were edited using Thermo Xcalibur Instrument Setup (version 4.2.28.14, Thermo Fisher Scientific Inc.). All methods were based on the standard LC-HRMS KWR non-target screening reversed phase method in positive mode with the following instrument settings.

A Vanquish HPLC system (Thermo Fisher Scientific) coupled to a Tribrid Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific) was used for all experiments in this study. The LC system was composed of pumps, auto sampler (VF-A10-A) with draw- and dispense speed set to 5  $\mu\text{L/s}$ , column compartments (VH-C10-A) maintained at 25  $^{\circ}\text{C}$ . The analytical XBridge BEH C18 XP column (150mm x 2.1 mm i.d., 2.5  $\mu\text{m}$  particle size, Waters) was protected by a Phenomenex SecurityGuard Ultra column (UHPLC C18, 2.1 mm i.d.). The system was controlled with Thermo Scientific Xcalibur software (version 4.2.28.14, Thermo Fisher Scientific Inc.).

The mobile phase consisted of 0.05% formic acid (J.T. Baker, Avantor Performance Materials B.V., Deventer, the Netherlands) in ultrapure water (LiChrosolv, LC-MS grade, Merck, Darmstadt, Germany) [v/v] (mobile phase A) and 0.05% formic acid in acetonitrile (J.T. Baker, ultra-gradient HPLC grade, Avantor Performance Materials B.V., Deventer, the Netherlands) [v/v] (mobile phase B). The injection volume was set to 100  $\mu\text{L}$ . The total analysis time was set to 34 minutes: 0 to 1 min, isocratic at 5% B; 1 to 25 min linear gradient to 100% B, 25 to 29.5 min linear gradient to 5% B; 29.5 to 34 min isocratic at 5% B.

The mass spectrometer was equipped with a heated-electrospray ionization source with a spray voltage of 3000 V in positive mode. Sheath, auxiliary and sweep gas were set to 40, 10 and 5 (arbitrary units), respectively. Both the ion transfer tube and vaporizer temperature were set to 300  $^{\circ}\text{C}$ . Full scan high resolution mass spectra were recorded with the Orbitrap detector at a resolution of 120,000 FWHM from  $m/z$  80 up to  $m/z$  1000 during the first 28 minutes of the LC run. RF lens was set to 50%. For the MS1 scan, the AGC target was set at  $2.0 \times 10^5$  with a maximum IT of 100 ms. Data was acquired in profile mode.

Three different CE settings were tested in the AcquireX experiments: stepped CE 20, 35 and 50, assisted CE 20, 35 and 50, and assisted CE 20, 35, 50 and 75. These three methods were tested at a IT of 30 ms, 50 ms and 100 ms. Four different AGC-targets for the MS2 scan were tested:  $5 \times 10^4$ ,  $2 \times 10^4$ ,  $1 \times 10^4$  and  $5 \times 10^3$  and four different ITs: 30 ms, 50 ms, 100 ms and 200 ms. Leading to 16 methods which were measured in triplicate. Sequence lists are given in *appendix F*.

## MS1-trigger experiments

The effect of using an inclusion list and isotopic ratio trigger were tested in the MS1-trigger experiments. Five inclusion lists were used. These consisted of compounds within the mass range of the full scan, i.e. m/z between 80 and 1000 Da, and polarity (if available) amenable to RP-HPLC, i.e. log  $K_{OW}$  between -2.5 and +3.5. When the scan range is enlarged to > 1000 Da, the sensitivity decreases significantly. Fragmentation of molecules < 80 Da leads to fragment ions below the detection limit in MS2, ~50 Da. The mass distribution of all organic compounds present in NORMAN substance database with the selected log  $K_{OW}$  is shown in *figure 9*. It reveals that the majority of compounds is between 80 and 1000 Da, marked by the vertical red dotted lines, and is thus covered by the full scan range.

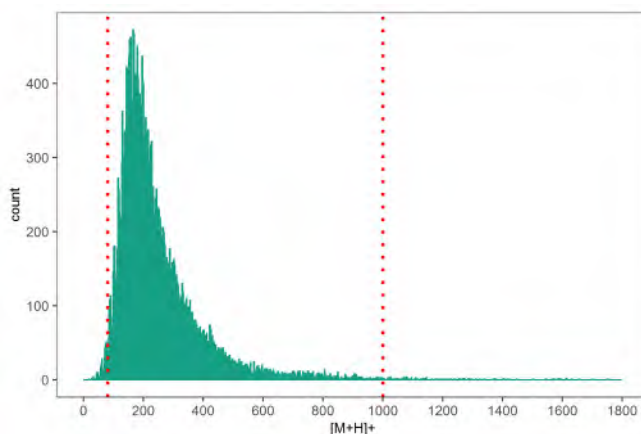


Figure 9 – Mass distribution of NORMAN Substance Database, filtered for organic compounds with predicted log  $K_{OW}$  between -2.5 and +3.5, binwidth = 2.

The five inclusion lists were:

- NORMAN Substance Database (susdat\_2020-02-03-164350.csv, downloaded at 3 February 2020).<sup>33</sup> This dataset was filtered for organic compounds with  $[M+H]^+ \geq 80$  Da and  $\leq 1000$  Da and log  $K_{OW}$  (predicted with EPISuite)  $\geq -2.5$  and  $\leq +3.5$ , resulting in 18667 compounds.
- NORMAN Substance Database with retention time prediction, the log  $K_{OW}$  values of the NORMAN Substance Database were used to predict the retention time with an experimentally derived equation, based on KWR internal data (see equation 1), this resulted in 32485 m/z values.<sup>64</sup>
  - $$t_R = \log K_{OW}/0.254 + 5.1945$$
- UBAMPT (Potential Persistent, Mobile and Toxic (PMT) substances, retrieved from NORMAN SusDat at 12 February 2020), also filtered for organic compounds with  $[M+H]^+ \geq 80$  Da and  $\leq 1000$  Da, resulting in 192 m/z values.<sup>36</sup>
- Extended KWR Sjerps list (Sjerp\_2016\_WatResManuscript\_SI-1.docx, downloaded at 19 February 2020) filtered for organic compounds with  $[M+H]^+ \geq 80$  Da and  $\leq 1000$  Da, resulting in 3399 m/z values.<sup>35</sup>
- Spiking list with all  $[M+H]^+$  of a water relevant contaminants spike, 82 m/z values. These water relevant contaminants (LOA-600 + specials) are determined in-house at KWR and listed in *appendix E*.

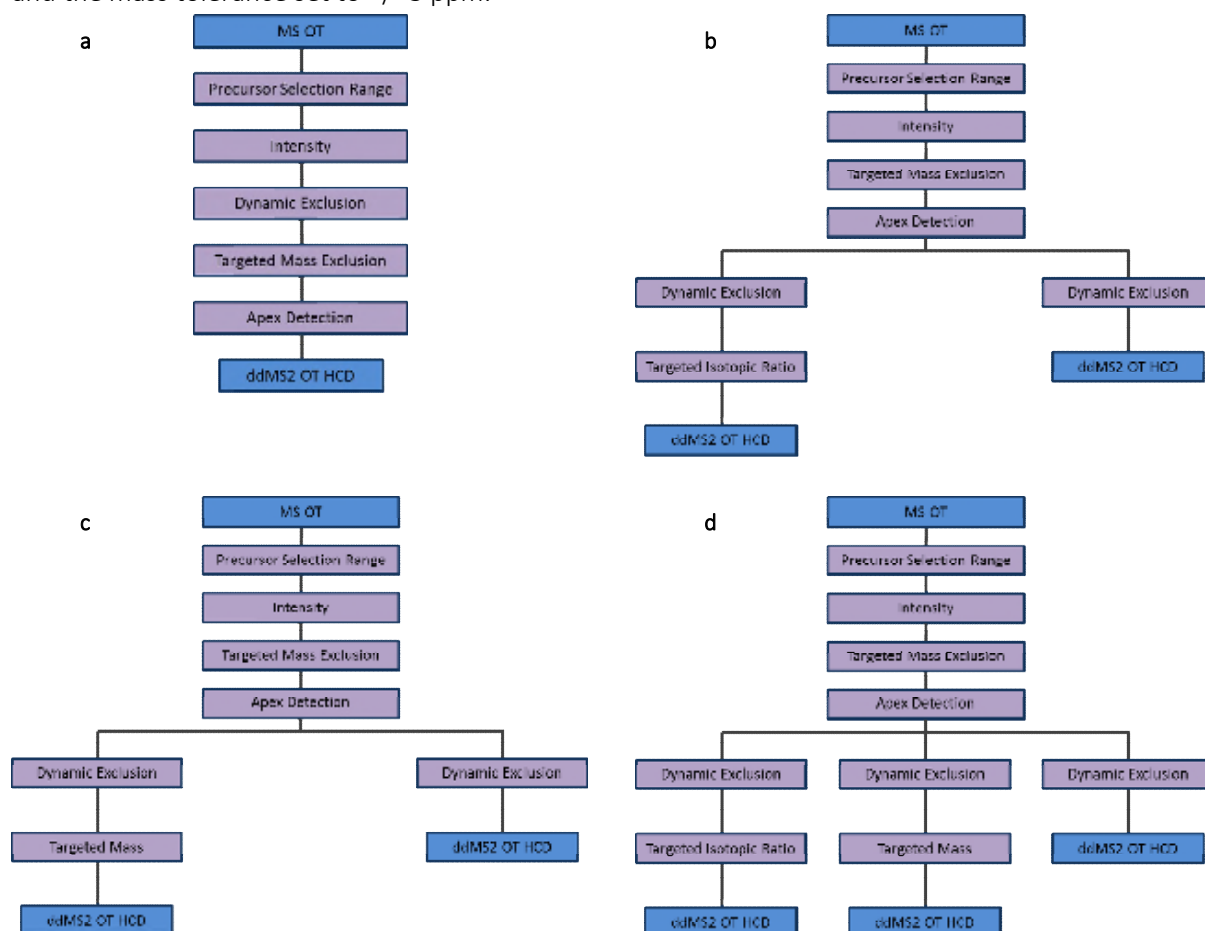
A distribution was made of the number of chlorine and bromine atoms in all 869027 compounds registered in the CompTox Chemistry dashboard, see *appendix G*.<sup>65</sup> As some outliers were present (formulas with 30 Cl or 18 Br atoms), the isotopic ratios covering  $\geq 99\%$  of the chlorinated compounds ( $n = 128650$ ) and brominated compounds ( $n = 53258$ ) were considered. The isotopic ratios of Cl up to

Cl<sub>6</sub> and Br up to Br<sub>5</sub> were calculated with Compound Discoverer software (Xcalibur) and are shown in *table 3*.

*Table 3 – Calculated ratios between the monoisotopic peak and the next peak, for Cl and Br combinations.*

delta M	expected ratio	number of Cl or Br atoms
1.99705	0.3193	Cl
1.99705	0.6394	Cl <sub>2</sub>
1.99705	0.9586	Cl <sub>3</sub>
1.99705	1.2788	Cl <sub>4</sub>
1.99705	1.5960	Cl <sub>5</sub>
1.99705	1.9206	Cl <sub>6</sub>
1.99795	0.9724	Br
1.99795	1.9455	Br <sub>2</sub>
1.99795	2.9231	Br <sub>3</sub>
1.99795	3.8939	Br <sub>4</sub>
1.99795	4.8657	Br <sub>5</sub>

The inclusion lists and the isotopic ratio trigger were tested separately and combined. The design of the resulting acquisition decision trees is shown in *figure 10*. The methods were tested on surface water and WWTP-influent samples spiked with water relevant contaminants (LOA-600 + specials, see *appendix E*). The sequence lists of the MS1-trigger experiments are given in *appendix F*. The inclusion lists were imported in the ‘Targeted Mass’ node, with the targeted mass tolerance set to +/- 5 ppm. The isotopic ratios were imported in the ‘Targeted Isotopic Ratio’ node with a ratio tolerance of 10% and the mass tolerance set to +/- 3 ppm.



*Figure 10 – Design of acquisition decision trees for MS1-trigger experiments, tree (a) is the regular KWR method, tree (b) includes the targeted isotopic ratio node, tree (c) includes the targeted mass node and tree (d) combines both.*

## MS2-trigger experiments

The effects of four different MS2-triggers, i.e. two recurring deltas and two recurring fragments, were tested as well. For this, ultrapure water samples were spiked with compounds that were predicted to exhibit these fragments and deltas in their MS2 spectra with the *in silico* experiments. The spike-in compounds were also added to surface water at different concentrations (1 ng/L up to 10 µg/L). All MS2-trigger samples are given in *appendix E*. The trigger experiments were performed separately, together and combined with the MS1-triggers using isotopic ratios and the Sjerps inclusion list. Presence of an MS2-trigger led to an additional MS2 event using alternative CEs, i.e. stepped CE (10, 75, 90) or assisted CE (20, 35, 50, 75, 90), or longer ITs, i.e. stepped CE (20, 35, 50) with 200 ms IT instead of the regular 50 ms. These alternative fragmentation events were hypothesized to result in spectra with complementary fragments in the case of alternative energies, and higher quality spectra in the case of longer ITs. The 11 different methods that were tested are described in *table 4* and the design of their decision trees in *figure 11*. The sequence lists can be found in *appendix F*.

Table 4 – Description of the methods tested in the MS2-trigger experiments.

Method	Description	Different settings additional MS2
1	regular NTS KWR method	-
2	Targeted Mass Trigger for alert TA344/TA362, m/z 62.9996	CE stepped (10, 75, 90)
3	Targeted Mass Trigger for alert TA367, m/z 55.0178	CE stepped (10, 75, 90)
4	Targeted Mass Difference for alert TA322, $\Delta M1 = 17.0265$	CE stepped (10, 75, 90)
5	Targeted Mass Difference for alert TA387/TA395, $\Delta M1 = 42.0106$	CE stepped (10, 75, 90)
6	All four alerts, m/z 62.9996; m/z 55.0178; $\Delta M1 = 17.0265$ ; $\Delta M1 = 42.0106$	CE stepped (10, 75, 90)
7	All four alerts, m/z 62.9996; m/z 55.0178; $\Delta M1 = 17.0265$ ; $\Delta M1 = 42.0106$	CE assisted (20, 35, 50, 75, 90)
8	All four alerts, m/z 62.9996; m/z 55.0178; $\Delta M1 = 17.0265$ ; $\Delta M1 = 42.0106$	max IT 200 ms
9	method 6 + MS1-triggers (isotopic ratio & Sjerps inclusion list)	CE stepped (10, 75, 90)
10	method 7 + MS1-triggers (isotopic ratio & Sjerps inclusion list)	CE assisted (20, 35, 50, 75, 90)
11	method 8 + MS1-triggers (isotopic ratio & Sjerps inclusion list)	max IT 200 ms

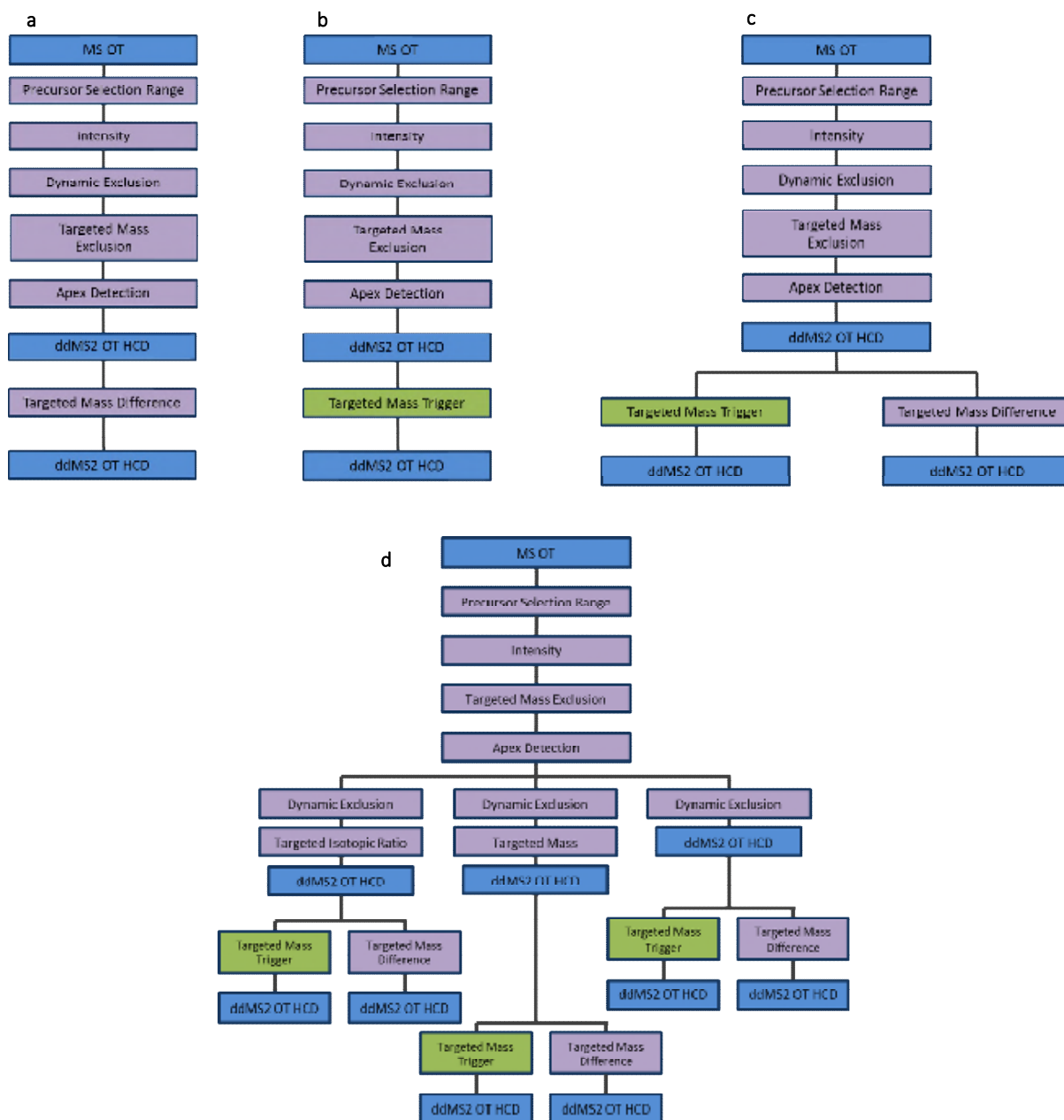


Figure 11 – Design of acquisition decision trees for MS2-trigger experiments, tree (a) corresponds to method 2 & 3, (b) corresponds to method 4 & 5, tree (c) combines both and corresponds to method 6, 7 & 8, and tree (d) combines MS1- and MS2-triggers and corresponds to method 9, 10 and 11.

## Data analysis

Compound Discoverer 3.1 (Thermo Fisher Scientific) was used to perform i.a. peak picking, retention time alignment and compound annotation. The Compound Discoverer workflow settings are given in *appendix H*. The output of the Compound Discoverer runs was exported as Compound Tables including MassList search results. For spectral quality assessment, compound annotations were removed in Compound Discoverer and the results of non-background features with  $t_R \geq 2.40$  min were exported to mzVault (Thermo Fisher Scientific). From mzVault, libraries containing feature information, CE, precursor mass and the 10 most intense peaks and their intensities were exported as .csv files. Both the mzVault libraries and the compound tables were imported in R version 3.6.1 for further processing.

The general parameters used for spectral quality assessment were obtained from the Compound Discoverer results: MassList hits, ChemSpider hits, mzCloud scores and mzLogic scores. For each MS2 scan assigned to a feature, eight more specific parameters derived by Nesvizhskii et al.<sup>66</sup> were calculated in R. These parameters consisted of the number of peaks, mean and standard deviation of the peak areas, smallest m/z ranges containing 50% and 95% of the total peak area, total ion current per m/z value, standard deviation of the sequential gaps between the peaks and the average number of neighbor peaks within a 2-Da interval around every peak.<sup>66</sup>

Spectrum similarity scores were calculated using the function `SpectrumSimilarity()` from the R-package `OrgMassSpecR`<sup>67</sup> (version 0.5-3), which is using equation (2) to calculate a similarity score between two spectra, where  $u$  and  $v$  are the aligned vectors of the two spectra. This function takes the signal intensities into account.

$$(2) \quad \cos \theta = (u \cdot v) / (\sqrt{\sum u^2} \times \sqrt{\sum v^2})$$

Fragment annotation was performed with the R-package `metfRag`<sup>42</sup> using the function `frag.generateMatchingFragments()` on the centroided MS2-spectra, using default settings. Thereby, the MS2-spectra of the four spiked compounds DEET, phenazone, primicarb and triphenylphosphine oxide could be assessed in regards to annotated peak numbers and intensities. These compounds were chosen because their MS2 scans reached the AGC-target of  $5 \times 10^4$  within the maximum IT.



## 4. Results

### 4.1 Screening of compounds for structural alerts

It appeared that the input SMILES uploaded to ToxAlerts were converted into different SMILES describing the same molecule, see *appendix C*. The use of canonical SMILES instead of MS-ready SMILES as input for ToxAlerts did not improve the overlap. Communicating the differences between in- and output to the developers led to an adjustment in the software of ToxAlerts which resulted in the possibility to connect the output SMILES to the original input MS-ready SMILES by using indexation.

In addition to ToxCast entries ( $n = 7571$ ), the MS-ready SMILES of the two databases NORMAN MassBank<sup>30</sup> ( $n = 2304$ ), and NORMAN SusDat<sup>33</sup> ( $n = 65697$ ) were screened for structural alerts. In the case of MassBank, only the 1903 compounds having available positive ionization HCD data were screened. Regarding SusDat, compounds were filtered for those with an EPISuite predicted log  $K_{ow}$  value between -3.0 and +4.5 (provided in SusDat), resulting in 46688 compounds. This filtering step was applied to eliminate compounds that are not detectable by RPLC. *Figure 12a* shows a comparison of the number of compounds before and after screening of the different datasets.

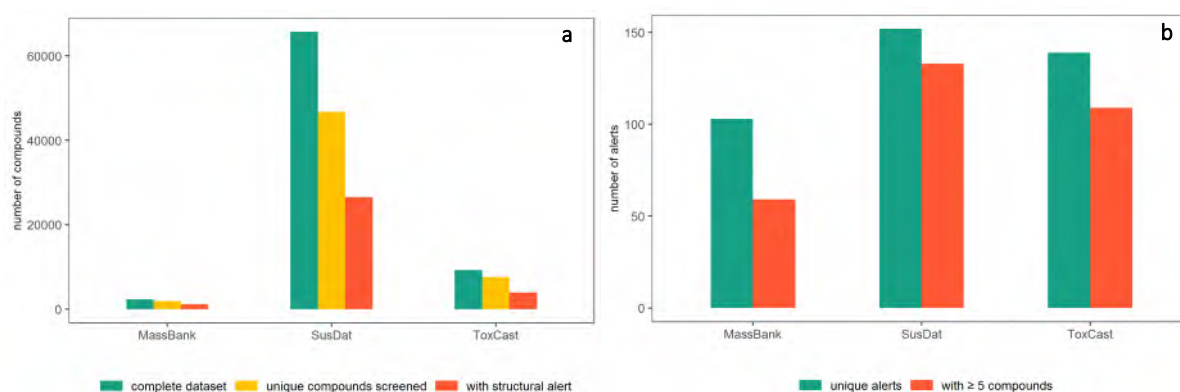


Figure 12 – a) Number of compounds per dataset, the unique compounds screened and the compounds with a structural alert. b) Schematic representation of the number of structural alerts present in the datasets.

Screening of the ToxCast database with ToxAlerts revealed the presence of 157 structural alerts, in one or more molecules. Several duplicate structural alerts were present in the ToxAlerts database. These were removed from the analysis resulting in 139 unique structural alerts. The distribution of the number of molecules per unique alert is shown in *figure 13a*. To enable pattern detection, a cut-off was set at  $n = 4$ . Consequently, only structural alerts with a minimum of 5 molecules were selected for further analysis, resulting in 109 structural alerts. Screening for structural alerts of SusDat compounds was performed accordingly, resulting in the detection of 152 unique alerts (see distribution in *figure 13b*) and 133 after the  $n = 4$  cut-off.

The compounds in the NORMAN MassBank dataset contained 103 unique structural alerts of which 59 alerts were present in at least 5 compounds. An overview of the number of structural alerts per dataset is illustrated in *figure 12b*.

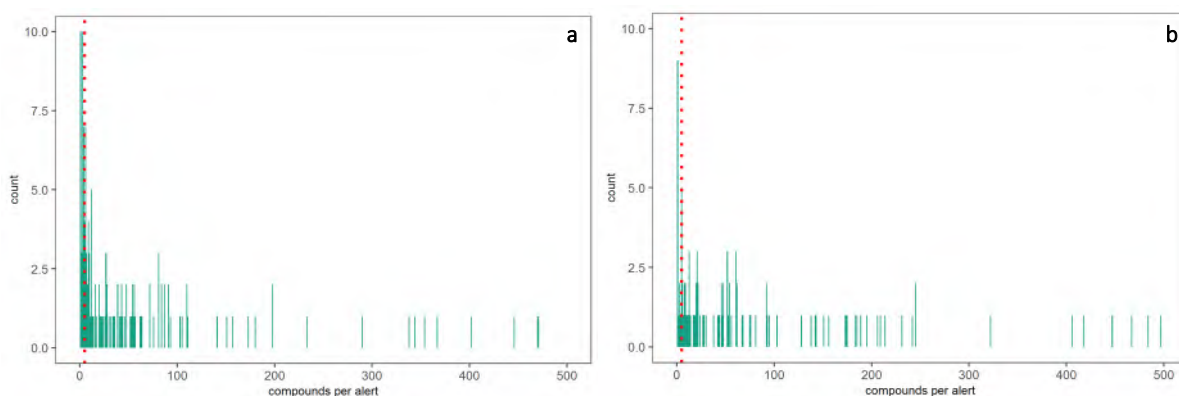


Figure 13 – Distribution of number of compounds found per alert (a) in ToxCast (TA441 with  $n = 1201$  and TA390 with  $n = 675$  were excluded from this graph) and (b) SusDat (48 alerts with  $n \geq 500$  were not included in this graph), the red dotted line is placed at  $n = 5$ . Binwidth = 1.

### Validation of toxicity

To validate the ToxAlerts approach for structural alert detection was investigated whether compounds with a given structural alert were active in a bioassay related to that alert. To this end, the percentages of compounds in the *in vitro* toxicity database ToxCast were calculated for compounds with both an alert and activity in the respective toxic endpoint(s), and for the active compounds in the total ToxCast dataset (table 5).

Overall, 55.1% of the ToxCast molecules with an alert were active in one or more bioassays corresponding to that alert, compared to 55.5% of all ToxCast molecules. These percentages are close to each other and give an indication that the alerts indeed indicate toxicity. They also indicate that the alerts used for screening were not covering all chemicals active in these toxic endpoints. Moreover, many compounds have not been tested on each bioassay. The availability of toxicity information in ToxCast has to be considered, not all chemicals are tested on all included assays.<sup>68</sup> And the U.S. EPA have not included all available assays in their program, e.g. Chemical Activated Luciferase gene eXpression (CALUX) Assays for detection of dioxins or endocrine disrupting compounds. This does not have to be a limiting factor for the derivation of structural alerts but it influences the validation of toxicity. Consequently, if a compound is not marked as active, it could be inactive or unknown.

The percentages might increase in case more tests are done. The data gap caused by missing toxicity information is shown in figure 14 per bioassay, and in figure 15 per structural alert. The data distribution for non-genotoxic carcinogenicity and genotoxic carcinogenicity and mutagenicity illustrated in figure 14 looks similar but it is not. This is due to the similarity in structural alerts linked to these toxicity endpoints and the overlap in bioassays that are linked to these endpoints. This is also represented by the percentages shown in table 5. Figure 15 suggests that there are no inactive compounds but compounds can only be marked as inactive if they are tested on all bioassays included in ToxCast and do not show activity in all of them. Since no compounds have been tested on all bioassays, none of them could be marked as inactive.

Table 5 – Toxicity of the screened molecules with a structural alert present, from screening results of the ToxCast data.

Toxic endpoint	active compounds with structural alert, % of all compounds with alert belonging to that endpoint	active compounds regardless of structural alert, % of all compounds in ToxCast (control group)
Endocrine disruption (edc)	57.2%	38.1%
Non-genotoxic carcinogenicity (ngc)	52.9%	45.8%
Genotoxic carcinogenicity, mutagenicity (gcm)	52.7%	45.8%
Developmental and mitochondrial toxicity (dmt)	11.1%	5.7%
Total toxicity	55.1%	55.5%

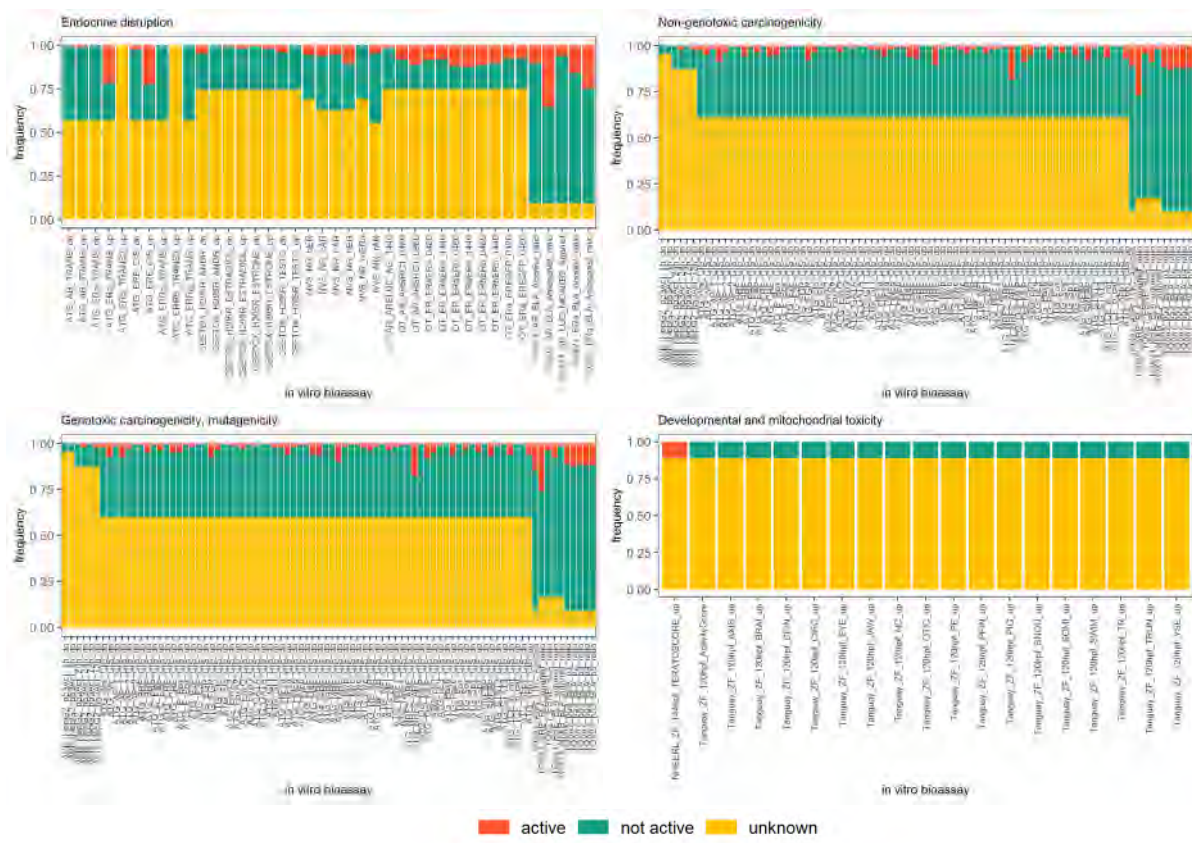


Figure 14 – Distribution of available toxicity information between in vitro bioassays per toxic endpoint.

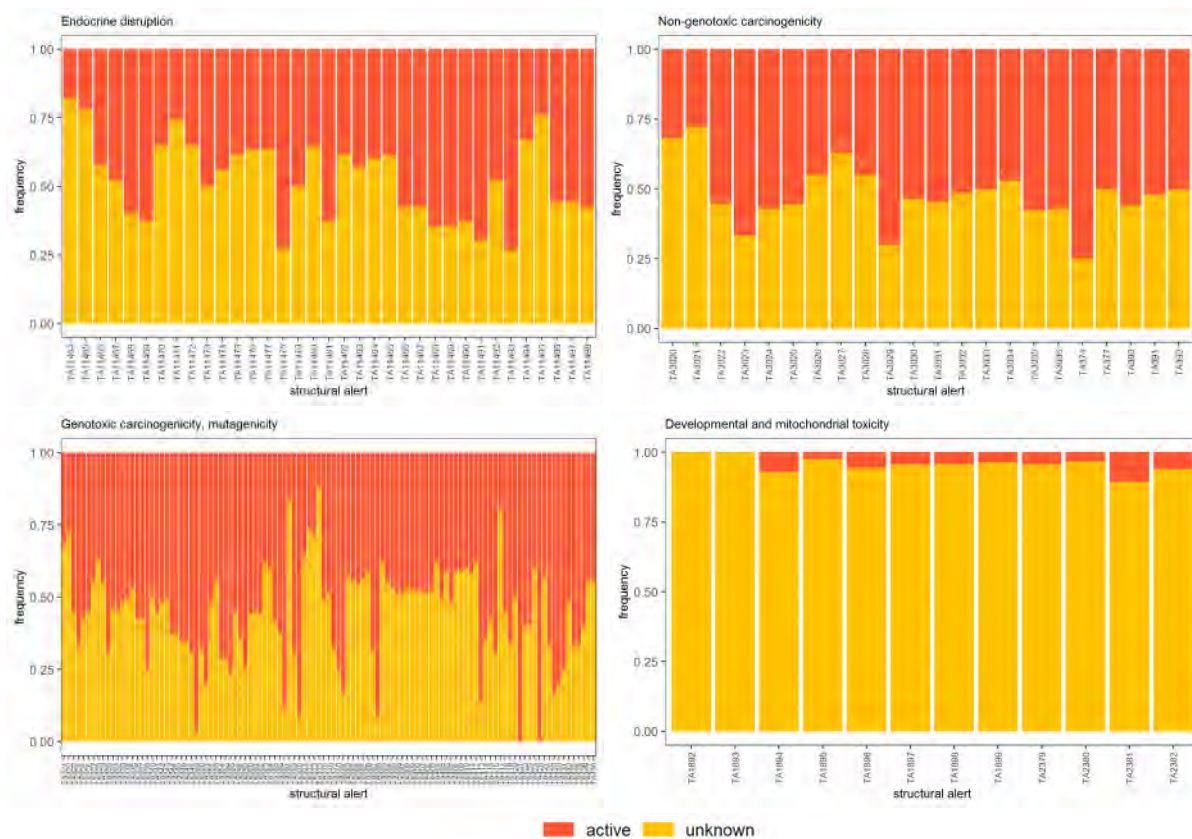


Figure 15 – Distribution of available toxicity information between in vitro bioassays per structural alert, each subplot represents a toxic endpoint.

## 4.2 *In silico* fragmentation

To be able to determine common patterns in the MS2 spectra of compounds with the same structural alert, fragmentation spectra were generated *in silico* using the fragmentation software CFM-ID 2.0 and MetFrag. As the CFM-ID software is designed to fragment molecules that are neutral or single charged (+1 or -1), multiple charged molecules resulted in an error leading to 8 molecules from the ToxCast dataset and multiple from the SusDat dataset that could not be fragmented. *In silico* fragmentation with MetFrag is based on a bond dissociation approach resulting in such a high number of predicted fragments that these could not be used for pattern mining. For example fragmentation of gonadorelin lead to 1031 MetFrag fragments and 274 CFM-ID fragments. Furthermore, CFM-ID provided intensity values to filter for the most likely fragments which was not possible for MetFrag.

### Validation with NORMAN MassBank data

The *in silico* fragmentation results generated by CFM-ID 2.0 were validated with experimental HCD data retrieved from NORMAN MassBank.<sup>30</sup> Positive ionization HCD data was available for 1903 compounds. 587 of these compounds were part of the NORMAN SusDat compounds with a structural alert. It is challenging to compare the *in silico* fragmentation results with the MassBank fragments due to experimental errors in the MassBank data. The experimental error plays a role when using the percentage of total MassBank fragments overlapping with CFM-ID results. This is due to MassBank records being composed of multiple spectra and thus similar fragments that differ by a few Dalton, these are listed as individual fragments.

To overcome this, a threshold of 10 ppm deviation was set to find overlapping fragments between the CFM-ID results and MassBank fragments. Overlap in percentage of MassBank and CFM-ID fragments was calculated using equations (3) and (4).

$$(3) \quad pct_{MassBank} = \frac{\text{number of MassBank fragments matching with CFM-ID}}{\text{total number of MassBank fragments}} \cdot 100\%$$

$$(4) \quad pct_{CFM-ID} = \frac{\text{number of CFM-ID fragments matching with MassBank}}{\text{total number of CFM-ID fragments}} \cdot 100\%$$

The results of the validation study are shown in *figure 16*. Fewer molecules have  $\geq 50\%$  of their fragments matched between the two datasets based on  $\%_{MB}$  than based on  $\%_{CFM-ID}$ . This is expected to be caused by the larger total number of fragments per molecule in MassBank than calculated by CFM-ID. This is probably due to the multiple spectra included in MassBank resulting in more different m/z values because of the experimental errors. Binning these experimental values with ranges of 10 ppm did not solve the issue because the theoretical fragment masses, where the 10 ppm deviation had to be derived from, were unknown. Therefore, all three CFM-ID fragmentation energies were included in the pattern mining and patterns were filtered for occurrence in the spectra of at least two fragmentation energies.

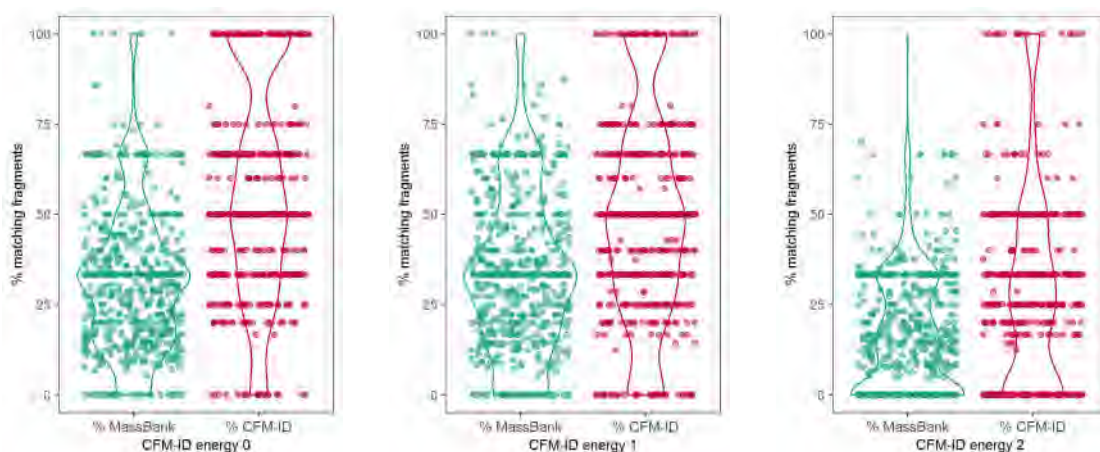


Figure 16 – Results of the validation study with MassBank and CFM-ID at the three different energy levels, shown in violin plots. Based on the 587 overlapping molecules between MassBank and CFM-ID.

### 4.3 Pattern mining

After *in silico* generation of fragmentation spectra, the predicted spectra were mined for patterns that are characteristic for each structural alert for subsequent use as MS2-triggers. These patterns included recurring fragment masses, and recurring mass differences between two fragments referred to as deltas.

Figure 17 illustrates the distributions of the recurring fragments and recurring deltas within structural alerts (blue bars) and within the total dataset (red bars) per CFM-ID fragmentation energy (different plots). The frequencies are shown on the x-axis and the number of cases where this frequency occurs, the 'count', is shown on the logarithmic y-axis. These graphs indicate that a frequency threshold of around 0.1 would be sufficient to find fragments or deltas that are specific for a certain alert, as higher frequencies hardly occur within the total dataset. However, to increase specificity only fragments and deltas with a frequency higher than 0.5 were taken into consideration. Two deltas detected with high frequency were 2.01565 Da and 18.01056 Da. These were not considered as relevant deltas because they occurred in relatively high frequencies in the total dataset. These deltas are expected to correspond to a loss of 2H's and H<sub>2</sub>O, respectively.

The mining results of recurring fragments are shown in *table 6-7* and the results of the recurring deltas are shown in *table 8-9*. *Table 6* contains the frequencies of the recurring fragments in three different control datasets; all fragmented molecules with an alert from ToxCast, a random sample from SusDat, regardless of the presence of an alert, and all fragmented molecules with an alert from SusDat. The highest frequency is 0.02588 and is thus much lower than the frequency of that fragment within an alert, shown in *table 8*. *Table 8* is similar to *table 6*, but represents frequencies of the recurring deltas, which are also lower than the frequencies reported in *table 9*. So the frequencies in *table 6* and *8* support the recurring fragments and deltas for being indicative for their structural alerts. An often recurring fragment in mustard-like structural alerts is m/z 62.99960 which could correspond to C<sub>2</sub>ClH<sup>+</sup>, a fragment that is likely to form from these alerts. The recurring fragments m/z 55.01784 and m/z 109.01632 could correspond to C<sub>3</sub>H<sub>3</sub>O<sup>+</sup> and C<sub>2</sub>H<sub>6</sub>ClON<sub>2</sub><sup>+</sup>, respectively. Some structural alerts correspond to the same recurring fragment due to the similarity in their structures which could lead to similar fragments. Based on in-house availability of chemicals was decided to test the recurring fragments m/z 62.99960 of alert TA344/TA362 and m/z 55.01784 of alert TA367, and the recurring deltas m/z 17.02655 of alert TA322 and m/z 42.01056 of alert TA387/TA395 were used as triggers in the MS2-trigger experiments.

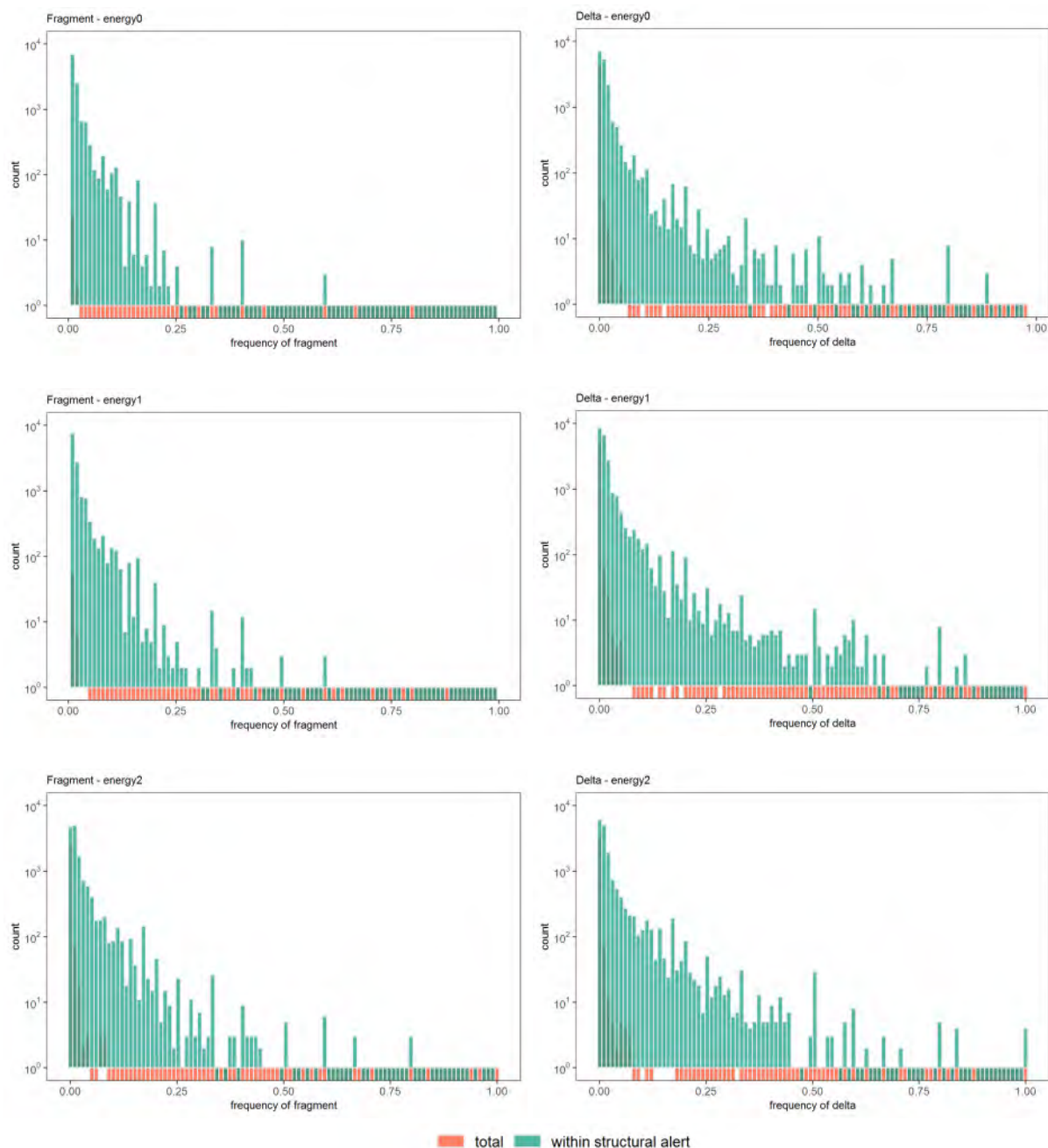


Figure 17 – Frequency distributions of recurring fragments (left) and recurring deltas (right) per CFM-ID energy. Note the logarithmic scale, the bars that look negative represent a count of 0, if no bar is visible, the count is 1.

Table 6 – All recurring fragments and their frequencies in ToxCast compounds with an alert, in a random sample of SusDat (regardless of the presence of an alert), and in SusDat compounds with an alert.

Fragment (m/z)	Frequency in fragmented part of ToxCast with alert (n = 3932)			Frequency in random sample of SusDat (n = 3953)			Frequency in fragmented part of SusDat with alert (n = 26081)		
	Energy 0	Energy 1	Energy 2	Energy 0	Energy 1	Energy 2	Energy 0	Energy 1	Energy 2
55.01784	0.012971	0.013889	0.025178	0.000253	-	-	0.013266	0.017752	0.025881
62.99960	0.004578	0.013889	0.015005	-	-	-	0.004601	0.007707	0.013803
109.01632	0.001017	0.001017	-	0.000506	0.000253	0.000253	0.000192	0.000192	-
121.02841	0.009410	0.013479	0.015514	-	-	-	0.004141	0.007247	0.009705

Table 7 – Structural alerts with a recurring fragment and its frequencies in each dataset.

Alert	Name <sup>47,69</sup>	Structure	Endpoint	Recurring fragment frequencies within compounds with the alert		
				m/z	ToxCast freq	SusDat freq
TA344 n <sub>TC</sub> = 23 (0.3%) n <sub>SD</sub> = 95 (0.2%)	Nitrogen and sulphur mustard (specific) X = Cl, Br, I		gcm	62.99960	E <sub>0</sub> : 0.174 E <sub>1</sub> : 0.783 E <sub>2</sub> : 0.957	E <sub>0</sub> : 0.189 E <sub>1</sub> : 0.705 E <sub>2</sub> : 0.853
TA362 TA3023 TA435 n <sub>TC</sub> = 11 (0.1%) n <sub>SD</sub> = 21 (<0.1%)	S or N mustard R = any atom/group; X = F, Cl, Br, I		gcm, ngc	62.99960	E <sub>0</sub> : - E <sub>1</sub> : 0.636 E <sub>2</sub> : 1.000	E <sub>0</sub> : 0.095 E <sub>1</sub> : 0.714 E <sub>2</sub> : 0.810
TA367 n <sub>TC</sub> = 81 (1.1%) n <sub>SD</sub> = 578 (1.2%)	α, β-Unsaturated carbonyl R <sub>1</sub> and R <sub>2</sub> = any atom/group, except alkyl chains with C>5 or aromatic rings; R = any atom/group, except OH, O-		gcm	55.01784	E <sub>0</sub> : 0.593 E <sub>1</sub> : 0.617 E <sub>2</sub> : 0.679	E <sub>0</sub> : 0.571 E <sub>1</sub> : 0.599 E <sub>2</sub> : 0.585
TA401 n <sub>TC</sub> = 5 (<0.1%) n <sub>SD</sub> = 5 (<0.1%)	N-Nitroso-N-alkylureas R = aliphatic carbon or aromatic atom; R <sub>1</sub> = aliphatic carbon		gcm	62.99960  109.01632	E <sub>0</sub> : 0.600 E <sub>1</sub> : 0.600 E <sub>2</sub> : 0.800  E <sub>0</sub> : 0.800 E <sub>1</sub> : 0.800 E <sub>2</sub> : -	E <sub>0</sub> : 0.800 E <sub>1</sub> : 0.400 E <sub>2</sub> : 1.000  E <sub>0</sub> : 1.000 E <sub>1</sub> : 1.000 E <sub>2</sub> : -
TA414 n <sub>TC</sub> = 16 (0.2%) n <sub>SD</sub> = 67 (0.1%)	Haloethylamines R = hydrogen or carbon atom; X = F, Cl, Br, I		gcm	62.99960	E <sub>0</sub> : - E <sub>1</sub> : 0.875 E <sub>2</sub> : 0.938	E <sub>0</sub> : 0.045 E <sub>1</sub> : 0.731 E <sub>2</sub> : 0.791
TA415 n <sub>TC</sub> = 10 (0.1%) n <sub>SD</sub> = 76 (0.2%)	Haloalkylethers R = carbon atom; X = F, Cl, Br, I; only ethers containing -OCH <sub>2</sub> X (methyl) or -OCH <sub>2</sub> CH <sub>2</sub> X (ethyl) groups are included		gcm	62.99960	E <sub>0</sub> : 0.600 E <sub>1</sub> : 0.500 E <sub>2</sub> : 0.600	E <sub>0</sub> : 0.526 E <sub>1</sub> : 0.513 E <sub>2</sub> : 0.526

Table 8 – All recurring deltas and their frequencies in ToxCast compounds with an alert, in a random sample of SusDat (regardless of the presence of an alert), and in SusDat compounds with an alert.

Delta (m/z)	Frequency in fragmented part of ToxCast with alert (n = 3932)			Frequency in random sample of SusDat (n = 3953)			Frequency in fragmented part of SusDat with alert (n = 26081)		
	Energy 0	Energy 1	Energy 2	Energy 0	Energy 1	Energy 2	Energy 0	Energy 1	Energy 2
15.01090	0.019074	0.021871	0.059257	0.025044	0.027321	0.046294	0.036118	0.034316	0.066677
17.02655	0.145473	0.156918	0.059766	0.136352	0.133569	0.047306	0.166098	0.158621	0.048579
27.99491	0.094354	0.195066	0.136826	0.103213	0.189476	0.114849	0.095740	0.187953	0.135041
30.01056	0.013225	0.048576	0.067904	0.013155	0.043005	0.056919	0.017446	0.050535	0.066945
35.97668	0.023906	0.036626	0.042981	0.017961	0.022515	0.020238	0.030827	0.039262	0.036195
42.01056	0.047050	0.047559	0.053662	0.056160	0.065520	0.055654	0.061347	0.063724	0.060044
43.97207	0.003052	0.003815	0.006104	0.003036	0.005059	0.007336	0.002876	0.005138	0.007055
47.00073	0.001018	0.001017	-	0.116367	0.024538	0.010372	0.026839	0.048963	0.024363

Table 9 - Structural alerts with a recurring delta and its frequencies in each dataset.

Alert	Name <sup>47, 69</sup>	Structure	Endpoint	Recurring delta frequencies within compounds with the alert		
				m/z	ToxCast freq	SusDat freq
TA11479 $n_{TC} = 24$ (0.3%) $n_{SD} = 75$ (0.2%)			edc	27.99491	$E_0: 0.083$ $E_1: 0.208$ $E_2: 0.875$	$E_0: 0.307$ $E_1: 0.520$ $E_2: 0.787$
TA322 $n_{TC} = 445$ (5.9%) $n_{SD} = 3524$ (7.5%)	Aromatic amine (general) Ar = any aromatic/heteroaromatic ring	Ar-NH <sub>2</sub>	gcm	17.02655	$E_0: 0.562$ $E_1: 0.524$ $E_2: 0.200$	$E_0: 0.661$ $E_1: 0.516$ $E_2: 0.131$
TA360 TA3021 $n_{TC} = 9$ (0.1%) $n_{SD} = 92$ (0.2%)	N-Methylol derivatives R = any atom/group		gcm, ngc	30.01056	$E_0: 0.889$ $E_1: 0.556$ $E_2: 0.111$	$E_0: 0.848$ $E_1: 0.620$ $E_2: 0.076$
TA366 TA3027 TA332 $n_{TC} = 5$ (<0.1%) $n_{SD} = 12$ (<0.1%)	Alkyl nitrite R = any alkyl group	R-O-N=O	gcm, ngc	47.00073	$E_0: 0.800$ $E_1: 0.800$ $E_2: -$	$E_0: 0.667$ $E_1: 0.667$ $E_2: -$
TA387 $n_{TC} = 44$ (0.6%) $n_{SD} = 605$ (1.3%)	Aromatic N-acyl amine Ar = any aromatic/heteroaromatic ring, R = hydrogen, methyl; chemicals with ortho-disubstitution, or with an ortho carboxylic acid substituent with respect to the N-acyl amine group are excluded; chemicals with a sulfonic acid group (-SO <sub>3</sub> H) on the same ring of the amino group are excluded.		gcm	42.01056	$E_0: 0.886$ $E_1: 0.864$ $E_2: 0.091$	$E_0: 0.906$ $E_1: 0.701$ $E_2: 0.064$
TA395 $n_{TC} = 52$ (0.7%) $n_{SD} = 669$ (1.4%)	Secondary aromatic acetamides and formamides Ar = any aromatic/heteroaromatic ring; R = H, methyl or activated methyl		gcm	42.01056	$E_0: 0.904$ $E_1: 0.885$ $E_2: 0.115$	$E_0: 0.916$ $E_1: 0.716$ $E_2: 0.073$
TA408 $n_{TC} = 16$ (0.2%) $n_{SD} = 174$ (0.4%)	Benzylic halides Ar = any aromatic/heteroaromatic ring; X = Cl, Br, I		gcm	35.97668	$E_0: 0.625$ $E_1: 0.625$ $E_2: 0.063$	$E_0: 0.546$ $E_1: 0.534$ $E_2: 0.155$
TA423 $n_{TC} = 20$ (0.3%) $n_{SD} = 138$ (0.3%)	Isocyanate R = any atom/group	R-N=C=O	gcm	15.01090	$E_0: 0.300$ $E_1: 0.600$ $E_2: 0.550$	$E_0: 0.239$ $E_1: 0.514$ $E_2: 0.514$
TA424 $n_{TC} = 8$ (0.1%) $n_{SD} = 68$ (0.1%)	Isothiocyanate R = any atom/group	R-N=C=S	gcm	43.97207	$E_0: 0.800$ $E_1: 0.700$ $E_2: 0.300$	$E_0: 0.754$ $E_1: 0.580$ $E_2: 0.210$
TA433 $n_{TC} = 9$ (0.1%) $n_{SD} = 92$ (0.2%)	N-Methylol derivatives R = any atom/group		gcm	30.01056	$E_0: 0.625$ $E_1: 0.625$ $E_2: 0.125$	$E_0: 0.706$ $E_1: 0.706$ $E_2: 0.074$



## 4.4 LC-HRMS experiments

### Acquisition parameters

Prior to implementing MS triggers for the online prioritization of toxic compounds, the effect of a few selected acquisition parameters on the quality of the acquired MS2 spectra was studied. The effect of using a background exclusion list was tested, the effect of the different CE modes stepped and assisted CE together with maximum IT and the effect of four different AGC-targets and maximum ITs. Once the optimal CE, AGC-target and maximum IT parameters are determined, these can be applied in the additional scans triggered by the MS2-trigger, leading to spectra of higher quality and thereby facilitating identification.

AcquireX automatically generates a background exclusion list of features that are detected in the blank sample. The AcquireX experiments showed that the use of this exclusion list lead to a decrease in the percentage of MS2 scans of features marked as background. When no exclusion list was used;  $94.7 \pm 0.9$  % of the background features were fragmented. The background exclusion list significantly reduced this number to  $21.9 \pm 1.4$  %. As a result, more time is available for fragmentation of other, more relevant, features.

Besides the use of a background exclusion list the effect of different CE modes on spectral quality was tested. A fragmentation spectrum of good quality has high signal intensity, sufficient fragments and low noise levels. However, it is difficult to determine, high-throughput, whether a spectrum is of good quality or not because it is unknown what parameters describe spectral quality for small molecules. Moreover, no spectral quality metrics are available. Currently, spectral quality has to be determined manually for each single spectrum.

To assess spectral quality, eight spectral quality parameters developed for protein analysis by Nesvizhskii et al.<sup>66</sup> were determined for the experimental data sets, and the distributions of the parameters were visualized (see *appendix I*). The spectral quality parameter plots did not reveal obvious trends; a change in the distribution-shape was visible for some parameters, but not a clear shift on the x-axis. Moreover, it remains to be shown how these spectral quality parameters correspond to spectral quality. Together, these findings stress the need for spectral quality metrics. Another strategy including mzCloud and mzLogic scores was applied to gain insight into the spectral quality of the acquired spectra per experimental condition. These two scores assess the match between the experimental spectrum and a mzCloud library spectrum, and a combination of mzCloud and structural data from ChemSpider and applied masslists, respectively. No effect of CE modes was detected in the mzLogic and mzCloud score distributions. Alternatively to the spectral quality parameters and the mzCloud based annotation scores, *in silico* predicted fragmentation spectra can be used to determine the information content of a spectrum. When the spectrum has a higher information content the more fragments could be annotated by *in silico* prediction tools. The MS2 spectra of four spike-in compounds, i.e. DEET, primicarb, phenazone and triphenylphosphine oxide were selected, annotated with MetFrag and annotation was compared for each method.

The results of the annotation experiment for the spiked compounds DEET, primicarb, phenazone and triphenylphosphine oxide are shown in *figure 18-20*. The number of annotated fragments significantly increased with longer ITs in the case of primicarb (p-value of 0.00356), phenazone (p-value of  $2.32 \times 10^{-10}$ ) and triphenylphosphine oxide (p-value of  $2.05 \times 10^{-5}$ ) (two way ANOVA, *figure 18*). The DEET-annotation dataset was not normally distributed, preventing significance testing with an ANOVA. The effect of the CE mode varied between the compounds. For triphenylphosphine oxide assisted CE 20-75 resulted in the most annotated fragments (p-value triphenylphosphine oxide  $2.21 \times 10^{-9}$ ), for

phenazone assisted CE 20-50 (p-value phenazone  $3.44 \times 10^{-7}$ ). For primicarb the best mode varied per maximum IT (p-value primicarb 0.00158).

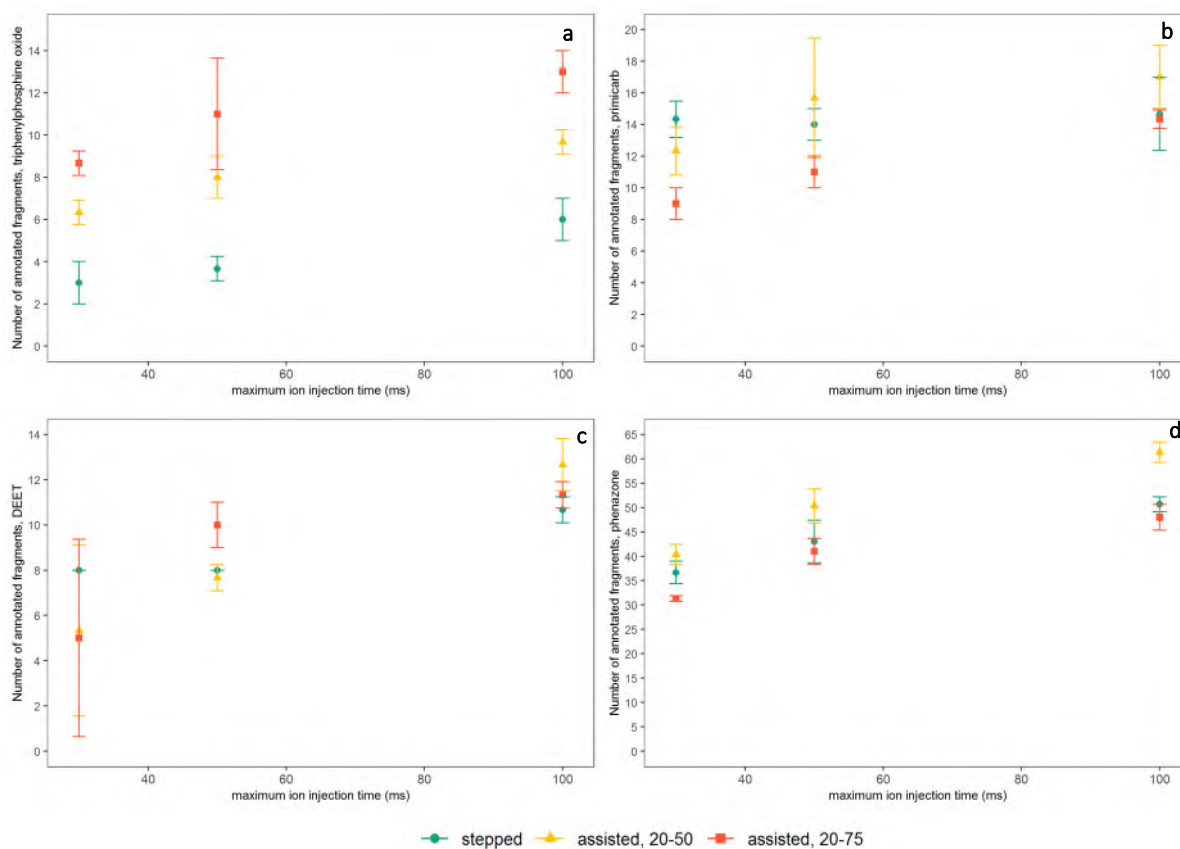


Figure 18 – Number of annotated fragments, AcquireX experiments. a) triphenylphosphine oxide, b) primicarb, c) DEET, d) phenazone.

The percentage of annotated fragments regarding the total number of peaks is shown in *figure 19*, where a significant effect of CE mode is present for primicarb (p-value of 0.00719) and triphenylphosphine oxide (p-value of  $3.14 \times 10^{-5}$ ). The effect of IT was significant for triphenylphosphine oxide (p-value of 0.0252). This suggests that both assisted CE modes led to a higher percentage of annotated fragments and higher maximum IT, at least for triphenylphosphine oxide. *Figure 20* displays the annotated percentage of the total peak area. The effects of maximum IT and CE type were significant for both triphenylphosphine oxide and phenazone. The graphs show a clear difference between both assisted CE modes and the stepped CE mode, a higher percentage of the peak area is annotated when assisted CE modes are applied.

These results suggest that more fragments can be annotated when higher maximum ITs are used, and when assisted CE is used instead of stepped CE. The advantage of assisted CE over stepped CE is that the optimal CE is determined experimentally for each precursor. The best range of assisted CE (e.g. CE 20-50 or 20-75) could not be determined from these results as this varies between the four compounds. With assisted CE a parallel scan is performed in the ion trap analyzer for each CE to determine the remaining precursor signal. Consequently, more CEs require more ion trap scans, prior to the actual MS2 acquisition scan which is performed in the Orbitrap analyzer. However, due to the speed of the ion trap scans the use of assisted CE does not have a large impact on the overall duty cycle.<sup>26</sup> This is thus not expected to cause a problem, especially in combination with the use of a background exclusion list and the developed method with MS1- and MS2-triggers, which allows more

time to focus on relevant features. Possibly, the chances of obtaining a scan on the best CE are higher if more CEs are included in assisted CE, but experiments are required to determine the best assisted CE range. Interestingly, large standard deviations were detected for some spectra from technical replicates (marked with large error bars in *figure 18-20*), possibly caused by low-intensity signals that lie around the detection limit. Other phenomena are the low percentage of annotated fragments and peak area for especially triphenylphosphine oxide and phenazone. It has to be determined whether these un-annotated peaks correspond to fragments originating from the compound or correspond to noise. In order to do so, other annotation software could be used, such as fragment annotation in mzCloud or spectra prediction using CFM-ID. These preliminary results from a small sample size indicate the need for more extensive studies on the selected acquisition parameters and their effect on spectral quality and fragment annotation.

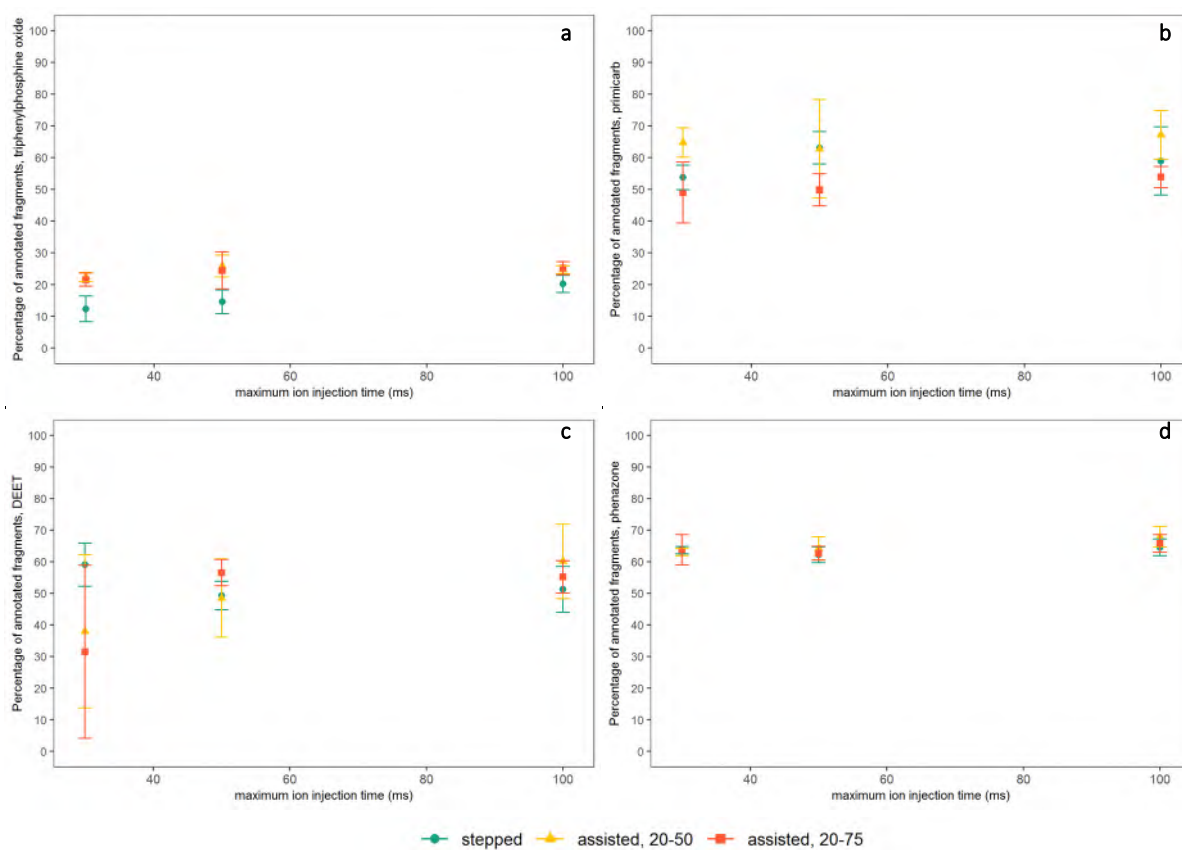


Figure 19 – Percentage of annotated peaks in the MS2 scan, AcquireX experiments. a) triphenylphosphine oxide, b) primicarb, c) DEET, d) phenazone.

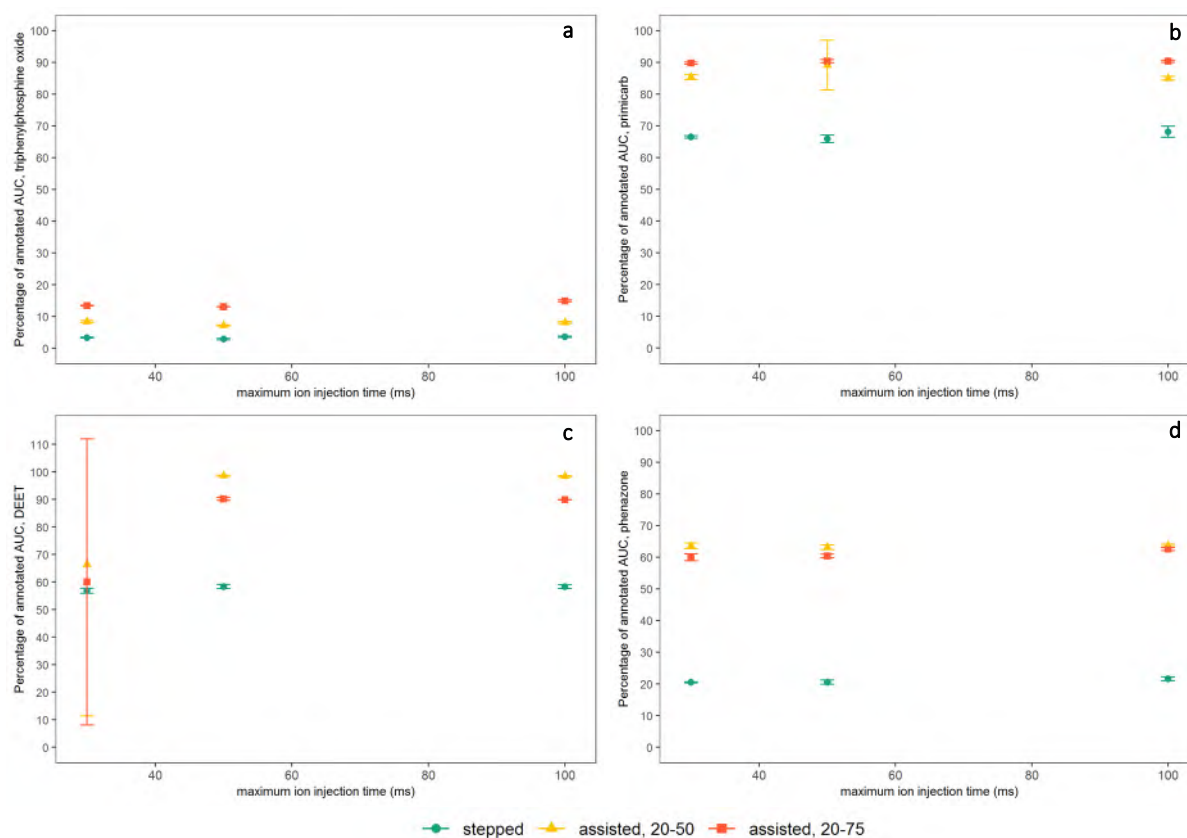


Figure 20 – Percentage of annotated peak area (AUC) in the MS2 scan, AcquireX experiments. a) triphenylphosphine oxide, b) primicarb, c) DEET, d) phenazone.

The results of the second set of AGC target experiments, in which four different AGC-targets were tested, are shown in *figure 21*. As expected, the percentage of MS2 scans reaching the AGC target before the maximum IT increases with lower AGC target and longer IT (see *figure 21c*). The decrease in number of MS1 spectra at higher ITs (see *figure 21a*) can be explained by the mass spectrometer cycle time which is fixed at 0.9 s in the applied top speed method. The mass spectrometer records the optimum number of MS2 scans which could lead to time shifts in the MS1 full scans.<sup>70</sup>

Another visible trend is the decrease in MS2 spectra with increasing maximum IT and higher AGC target, illustrated in *figure 21b*, which was expected as well. In case ions are allowed to accumulate during a larger time span, less time is available for other MS2 scans within the duty cycle. Moreover, more accumulation time is required for higher AGC targets.

The distributions of the eight spectral quality parameters for the different methods are shown in *appendix I*, where the red distributions correspond to the scans that reached the maximum AGC target before the maximum IT. Some distributions have non-normal shapes which is caused by a low sample size, especially for the scans at an AGC-target of  $5 \times 10^4$  and the lower ITs (at a maximum IT of 30 ms, none of the MS2 scans reached the AGC target of  $5 \times 10^4$ ).

Visible trends are the increasing number of peaks with increasing AGC-target and the decrease of total ion current per m/z with longer maximum IT. This could indicate that more low-intensity peaks are generated at higher AGC-targets and longer maximum ITs, which could possibly correspond to noise. The standard deviation of the consecutive m/z gaps between all peaks seems to decrease with shorter maximum ITs, suggesting that the distance between peaks becomes more regularly. It is unknown whether this corresponds to more noise or not.

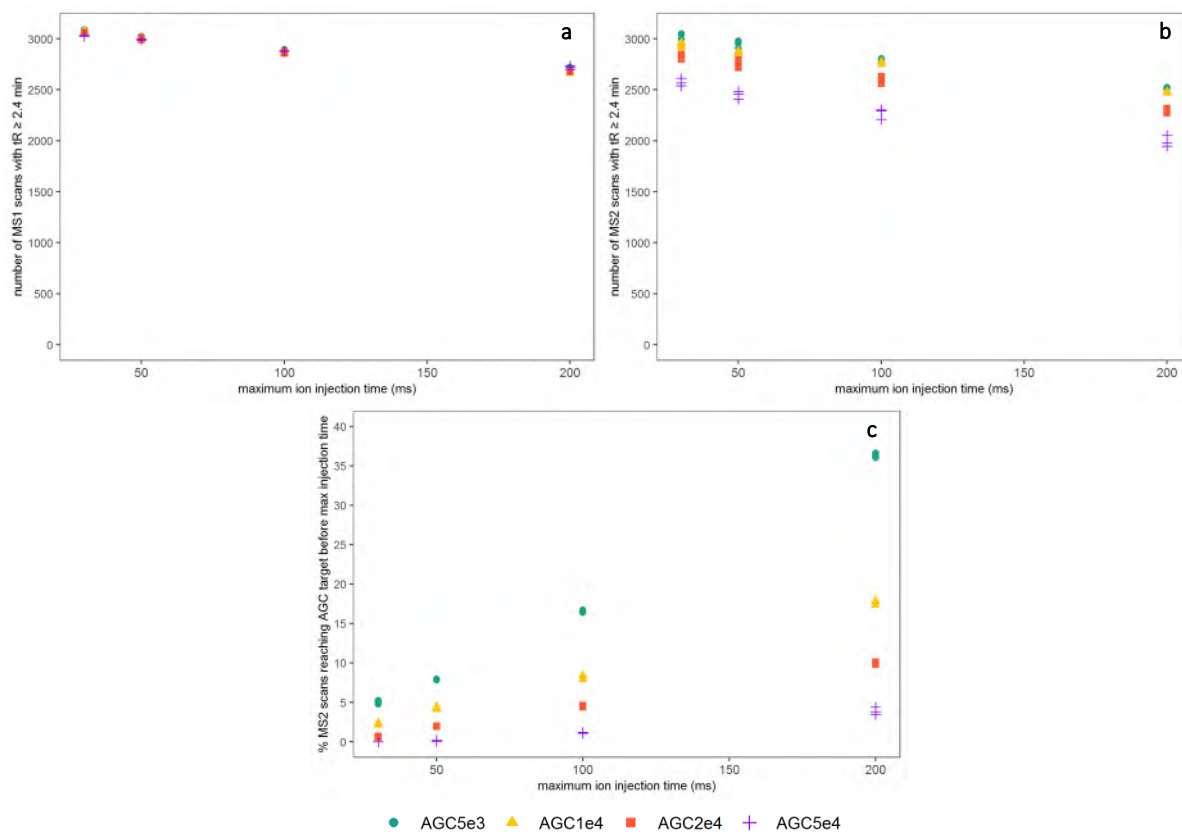


Figure 21 – a) number of MS1 scans taken per AGC target and maximum IT combination, b) number of MS2 scans taken and, c) percentage of MS2 scans reaching the AGC target before the maximum IT.

It has to be taken into account that some of these parameters are related to each other, such as the total ion current per  $m/z$  value and the  $m/z$  range of 95% and 50% of the total peak area. Currently it is unknown what values of these parameters correspond to a spectrum of high or low quality. A spectrum with noise only is expected to have many peaks of around the same intensity and thus a low standard deviation of the peak areas/intensities. But if this standard deviation is too high, it is possible that the spectrum consist of a high-intensity precursor signal and low-intensity fragment signals indicating lack of fragmentation.

To further investigate the effects of the selected acquisition parameters, the spectral similarity of four features that reached their AGC-target in the most methods was compared. These high scores (max score possible is 1) were reached because of the most intense peaks are matching, regardless of low-intensity peaks. Low intensity fragments are given less weight in the calculation of the spectral similarity score. As low intensity fragments might be the ones affected most by the acquisition parameters, another parameter was compared; fragment annotation by MetFrag.

The results of this fragment annotation test for four of the spiked compounds; DEET, primicarb, phenazone and triphenylphosphine oxide are shown in *figure 22*. A type III sum of squares two-way ANOVA was performed and showed a significant increase in the annotated fragments with increasing AGC-target for phenazone and primicarb ( $p\text{-value}_{\text{phenazone}} = 2.336\text{e-}08$ ,  $p\text{-value}_{\text{primicarb}} = 0.0001644$ ). The numbers of annotated fragments of DEET and triphenylphosphine oxide were not normally distributed and significance not tested. No other significant effects were found, but an important remark is that in the model that is generated for the ANOVA-test, 'aliased coefficients are found', suggesting singularity, which could affect the results and power of the statistical test.

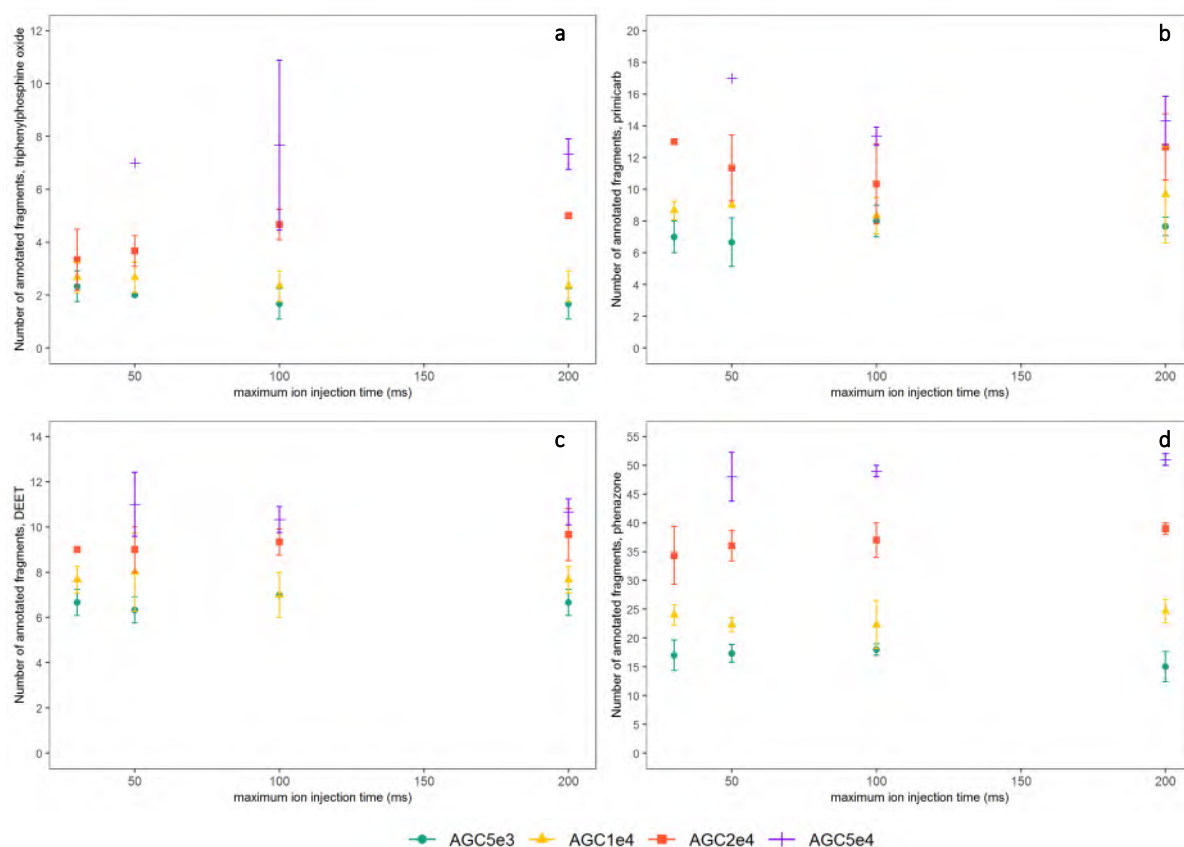


Figure 22 – Number of annotated fragments, AGC target experiments. a) triphenylphosphine oxide, b) primicarb, c) DEET, d) phenazone.

The percentage of annotated fragments regarding the total number of peaks is shown in *figure 23*. A type-III ANOVA showed no significant effects except for DEET where the maximum IT had a significant effect on the percentage of annotated fragments ( $p$ -value = 0.009109). This model encountered aliased coefficients as well. The graphs do not show a clear increase or decrease of percentage annotated peaks, suggesting that the ratio between total peaks and annotated peaks remains the same. Together with the increase in number of annotated fragments these results are suggesting that with increasing AGC-targets more peaks are generated. The percentage of annotated area under the curve (peak area) is shown in *figure 24*. No significant effects were found but the percentages of annotated peak areas differ largely between the compounds. It would be insightful to study the peaks that could not be annotated to see whether it is noise or fragments that could not be annotated by MetFrag. In order to do so, other annotation software (such as CFM-ID or Sirius<sup>71</sup>) or library matches (e.g. with mzCloud) could be applied to determine the differences.

Overall, these results suggest an increase in spectral quality with increasing AGC-targets. But further studies into spectral quality parameters and fragment annotation are required to be able to state this more confidently and to give a better indication of the change in spectral quality upon different AGC-targets and maximum IT settings. Nevertheless, in the MS2-trigger experiments an AGC-target of  $2 \times 10^4$  and maximum IT of both 50 and 200 ms were used in the alternative scan events.

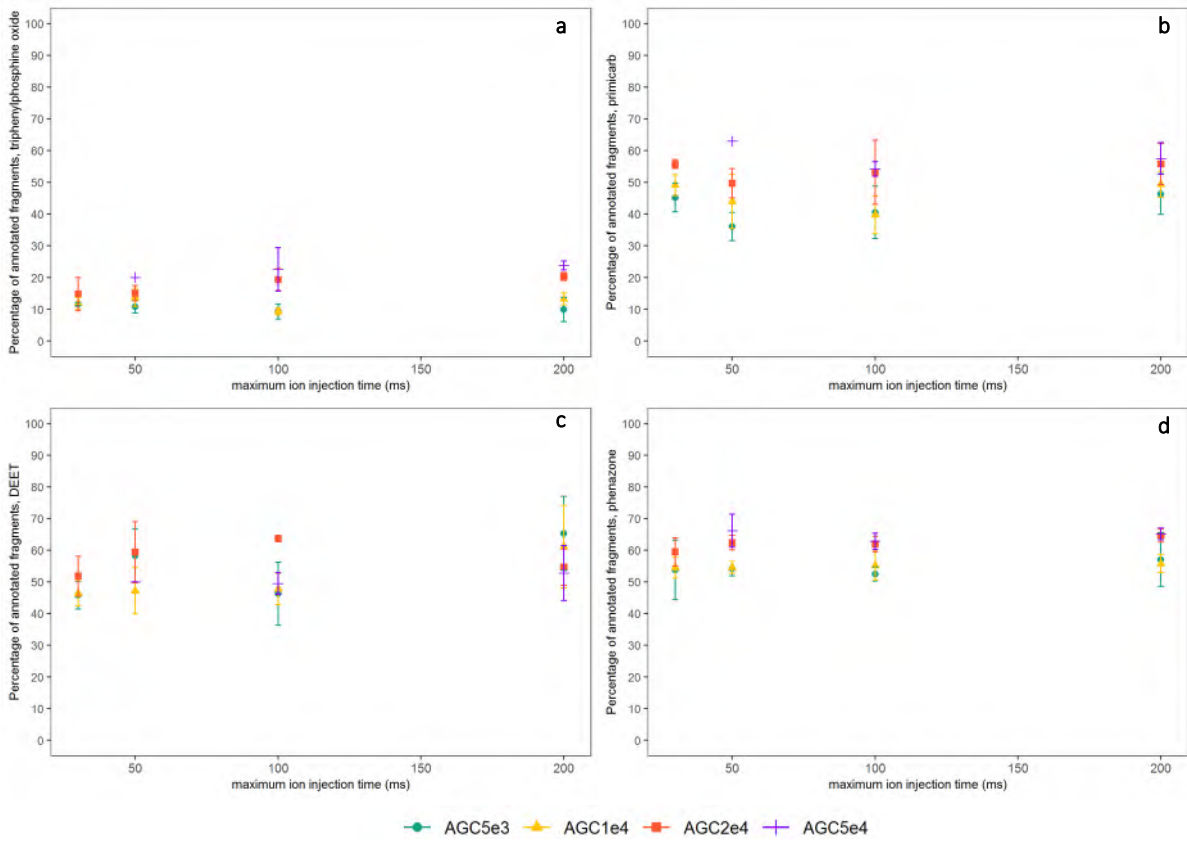


Figure 23 – Percentage of annotated peaks in the MS2 scan, AGC-target experiments. a) triphenylphosphine oxide, b) primicarb, c) DEET, d) phenazone.

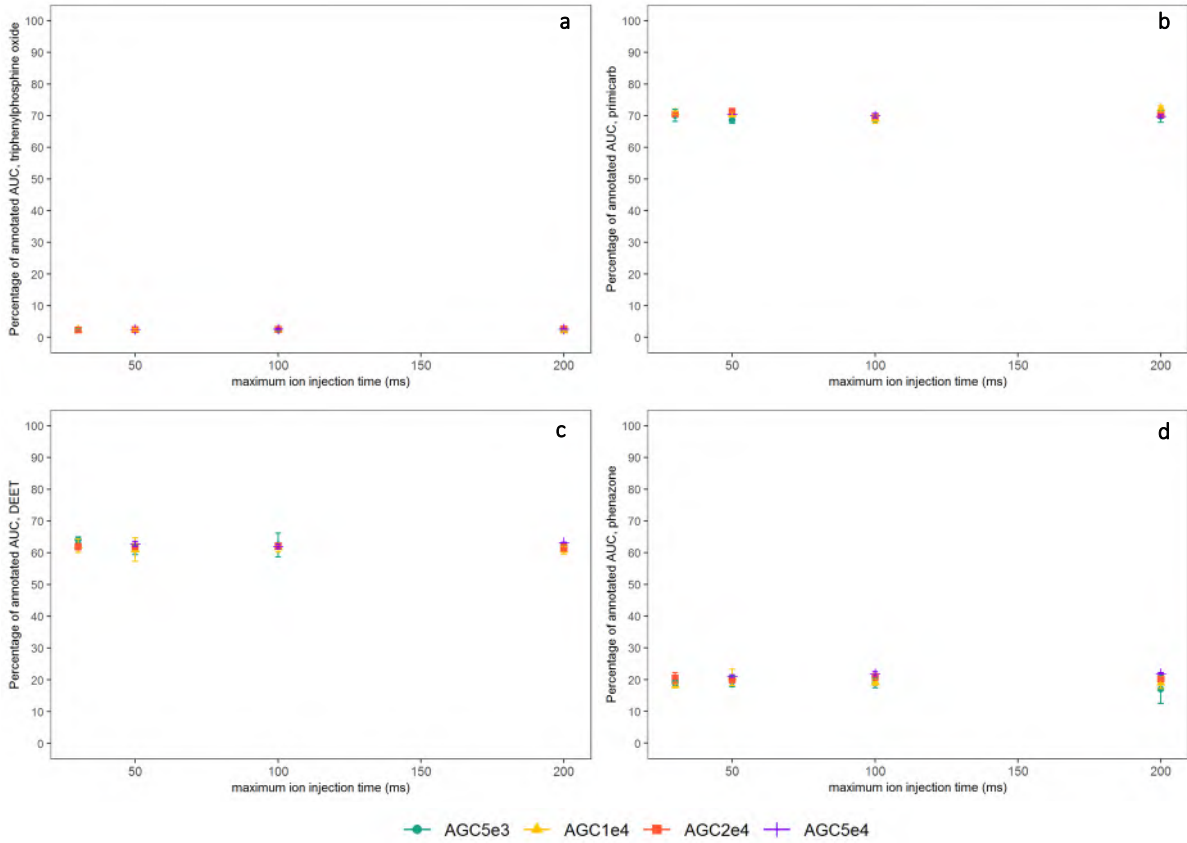


Figure 24 – Percentage of annotated peak area in the MS2 scan, AGC-target experiments. a) triphenylphosphine oxide, b) primicarb, c) DEET, d) phenazone.

## MS1-trigger experiments

Subsequently, the potential of MS1-triggers for the prioritization of toxic compounds was assessed experimentally. The MS1-triggers consisted of 5 different inclusion lists and the use of isotopic ratio triggers for chlorinated and brominated compounds. The results of both sample types, SW and WWTP-influent with a spike-in of common water relevant OMPs, are plotted in *figure 25-27*. No large differences in the number of detected features between all tested methods were visible, illustrated in *figure 25*. As expected, the WWTP-influent samples contained more features than the, much cleaner, SW samples.

The two different MS1-trigger methods with and without isotopic ratio did not result in large visible differences in percentage MS2 scans of Cl and/or Br containing features, see *figure 26*. This is mainly due to the fact that already the regular KWR method resulted in an average of 97.8% and 89.2% of the Cl and/or Br containing features with MS2 spectra in SW and in WWTP-influent, respectively.

While there was an increase in the percentage of chlorinated and/or brominated features with an MS2 spectrum visible in the SW and WWTP-influent samples for the method with isotopic ratio (see *figure 26b*), both differences were not significant ( $p\text{-value} > 0.05$ ). However, based on the Cl/Br pattern, which is a parameter in Compound Discoverer stating whether a chlorine or bromine-specific isotopic pattern is present in the MS2, there was a significant increase in the percentage of MS2 scans for the SW ( $p\text{-value}$  of 0.001292), but not the WWTP-influent samples.

A possible explanation could be that there are too many peaks in the full MS1 scan leading to difficulties in selecting proper isotopic ratios, in particular when low error tolerances are set. This is also supported by the pattern matches determined during the Compound Discoverer analysis. The peaks of Cl and/or Br containing features should contain a characteristic isotopic pattern due to the natural abundance of chlorine and bromine isotopes. Less features are selected using this approach than based on assigned formula (see *figure 27*), which could indicate that the isotopic patterns are not detectable by the software. Alternatively, it could indicate that the formulas assigned by Compound Discoverer were not correct. In Compound Discoverer, formula annotation and isotopic pattern are independent which could lead to false elemental formula annotations. A suggestion for further research would be to change these acquisition parameters (ratio- and mass tolerance) of the targeted isotopic ratio trigger in such a way that isotopic patterns are triggered at a lower threshold. However, a downside of lowering this threshold is the increase in false positive triggers.

Based on these results the isotopic ratio was implemented in the intelligent acquisition method as MS1-trigger as it focuses on Cl-/Br- containing features which are mostly anthropogenic and often toxic and the risk of triggering fragmentation of irrelevant features is low.

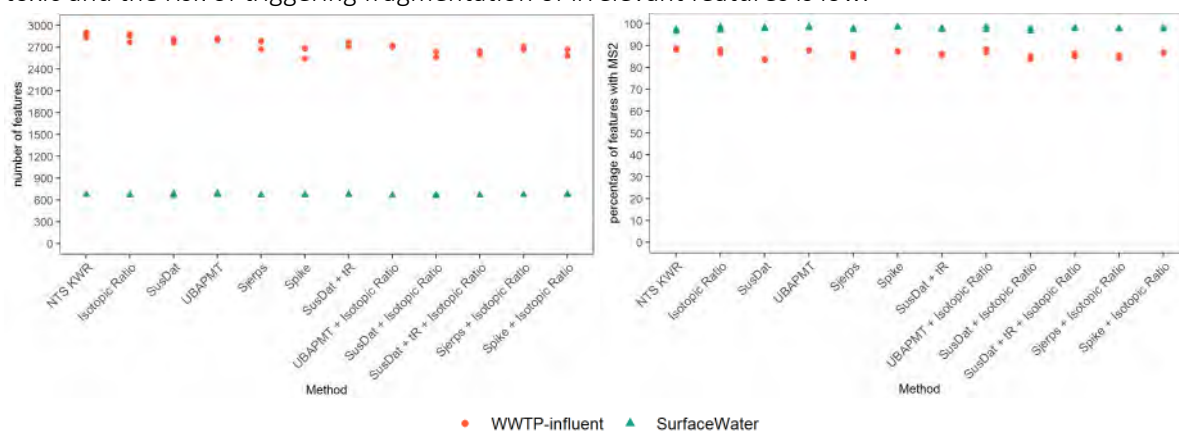


Figure 25 – Number of detected features (including background) per MS1-trigger method (left) and percentage of non-background features with MS2 (right).



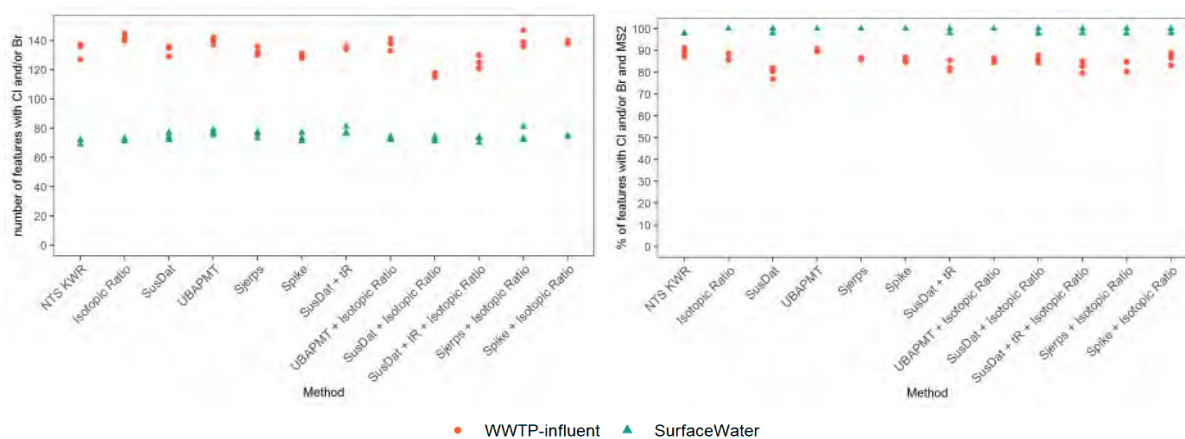


Figure 26 – Number of detected features with Cl and/or Br per MS1-trigger method (left) and percentage of non-background chlorinated and/or brominated features with MS2 (right). Presence of Cl and/or Br determined based on assigned formula.

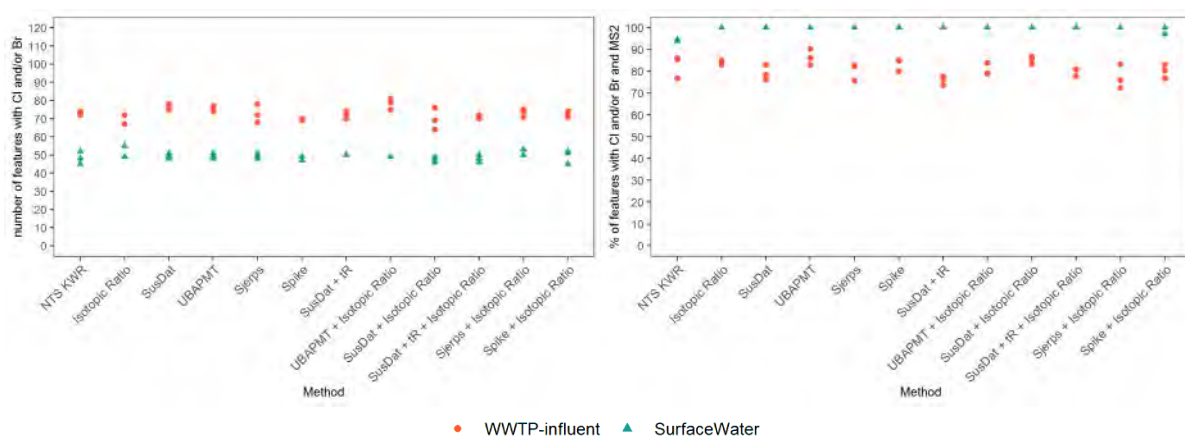


Figure 27 – Number of detected features with Cl and/or Br (including background) per MS1-trigger method (left) and percentage of non-background chlorinated and/or brominated features with MS2 (right). Presence of Cl and/or Br determined based on pattern match.

As an additional MS1-trigger, the use of inclusion lists consisting of water relevant compounds was investigated. For both SW and WWTP-influent samples, the percentage of features that matched the mass of a compound in the inclusion list (with +/- 5 ppm error tolerance) with a MS2 scan was calculated and compared with the regular KWR method. There was a significant increase ( $p$ -value = 0.0005877) in the average percentage of MS2 scans of these features for the SW samples in the method with SusDat inclusion list with retention time estimate ( $\mu$ % features with MS2 = 98.8%), compared to the KWR method ( $\mu$ % features with MS2 = 97.9%).

In the WWTP-influent samples, a significant increase in percentage of MS2 scans taken of  $m/z$  values present in the inclusion list was visible for SusDat, SusDat + tR and the Sjerps list, see *table 10*. The SusDat + tR inclusion list is very long (32485  $m/z$  values) and would also contain many compounds that are not relevant for water, which also accounts for the SusDat inclusion list without retention time estimate. Consequently, the Sjerps inclusion list was used for the MS2-trigger experiments, despite not showing a significant increase for the SW samples.

Table 10 – Comparison of percentage MS2 scans of inclusion list m/z values between methods. WWTP-influent samples.

Inclusion list type	Method with inclusion list $\mu\%$ features with MS2	KWR method $\mu\%$ features with MS2	p-value
SusDat	95.86	91.76	0.01576
UBAPMT	100.00	100.00	-
Sjerps	98.36	93.32	0.01485
Spike	96.41	95.60	0.3425
SusDat + tR	97.74	92.53	0.004934

In addition to the percentage of features with MS2 spectra, mzCloud and mzLogic scores were used to assess the performance of the different inclusion lists. mzCloud scores state how well the experimental MS2 spectrum matches with the spectrum available in the mzCloud library. mzLogic scores are similar but based on a combination of mzCloud, structural data from ChemSpider and selected masslists, resulting in scores for compounds that are not present in mzCloud. It turned out that there were no differences in the distribution of the mzCloud and mzLogic scores between the regular KWR method and the different methods with inclusion lists. The MS1-triggers were not expected to have an effect on the spectral quality as they only determine whether a MS2 spectrum is acquired or not. However, they could have an effect on the identification because higher identification and confidence levels can be reached in case a MS2 spectrum is recorded.

All in all, the isotopic ratio MS1-trigger requires further testing with larger tolerance settings. The inclusion list MS1-trigger has shown promising results for the SusDat inclusion lists, with and without retention time estimate, and Sjerps inclusion list. As the Sjerps list contains water relevant compounds was decided to use this list in combination with the MS2-triggers.

### MS2-trigger experiments

Next to MS1-triggers that trigger a MS2 scan, MS2-triggers were developed that trigger an additional MS2 scan if a structural alert is present. Four MS2-triggers were tested; the recurring fragments m/z 62.99960 of e.g. alert TA344/TA362 and m/z 55.01784 of alert TA367, and the recurring deltas m/z 17.02655 of alert TA322 and m/z 42.01056 of alert TA387/TA395. 15 spike-in compounds were measured (listed in *appendix E, table E.2-E.7*) of which acrylamide, isobornyl acrylate and 4-[2-(Acryloyloxy)ethoxy]-4-oxobutanoic acid were not detectable with the applied LC-HRMS method.

The diagnostic fragments were present in the MS2 spectra of all detected compounds, thereby confirming the *in silico* results generated with CFM-ID. However, the diagnostic fragment m/z 62.99960 did not trigger an additional MS2 scan for ifosfamide in all cases. This was due to the error tolerance settings of +/- 5 ppm which allows in this case a deviation of +/-m/z 0.00031. A typical MS2 scan of ifosfamide where no additional MS2 was triggered is represented by *figure 28*. The m/z value of 62.99913 is not covered with this small error tolerance, other MS2 scans of ifosfamide in ultrapure water contained diagnostic fragment peaks between m/z 62.99912 and m/z 62.99915. The same occurs in the MS2 spectra of diacetone acrylamide which has alert TA367 and should contain a diagnostic fragment of m/z 55.0178, with an allowed 5 ppm deviation of +/- m/z 0.00028. The MS2 spectra of diacetone acrylamide in ultrapure water contained diagnostic fragment peaks between m/z 55.01739 and m/z 55.01744, no additional MS2 scans were triggered. Therefore, an absolute error tolerance of 0.001 is suggested to use instead of a tolerance of 5 ppm.

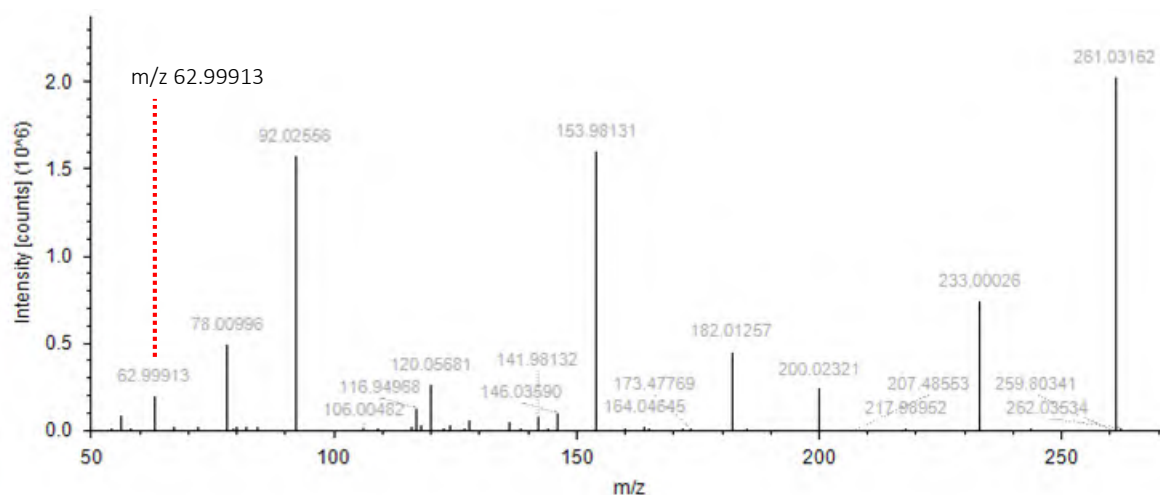


Figure 28 – MS2 spectrum of ifosfamide, the diagnostic fragment that should function as MS2-trigger is marked with red.

In addition to the diagnostic fragments, the use of diagnostic deltas as MS2-triggers was investigated. The diagnostic delta m/z 17.02650 was present in all spiked compounds in ultrapure water that contained this alert, thereby validating the *in silico* predicted spectra using CFM-ID. The diagnostic delta m/z 42.01060 was present in all spiked compounds except diatrizoic acid and one measurement of the n-acetylsulfamethoxazole spike-in. It was however present in the other two measurements of the triplicate.

The diagnostic delta m/z 17.02650, corresponding to alert TA322, did trigger additional MS2 scans for all spiked compounds in ultrapure water. An example of this trigger is visible in the MS2 spectrum of desethylatrazine, see *figure 29*. The delta m/z 42.01060 corresponding to alert TA387 and TA395 triggered additional MS2 scans in all compounds that were spiked in ultrapure water and where the diagnostic delta was detected. However, the trigger did not function for diatrizoic acid where this delta was not present in the MS2 spectra, and the single case of n-acetylsulfamethoxazole, in contrast to the *in silico* generated fragmentation spectra.

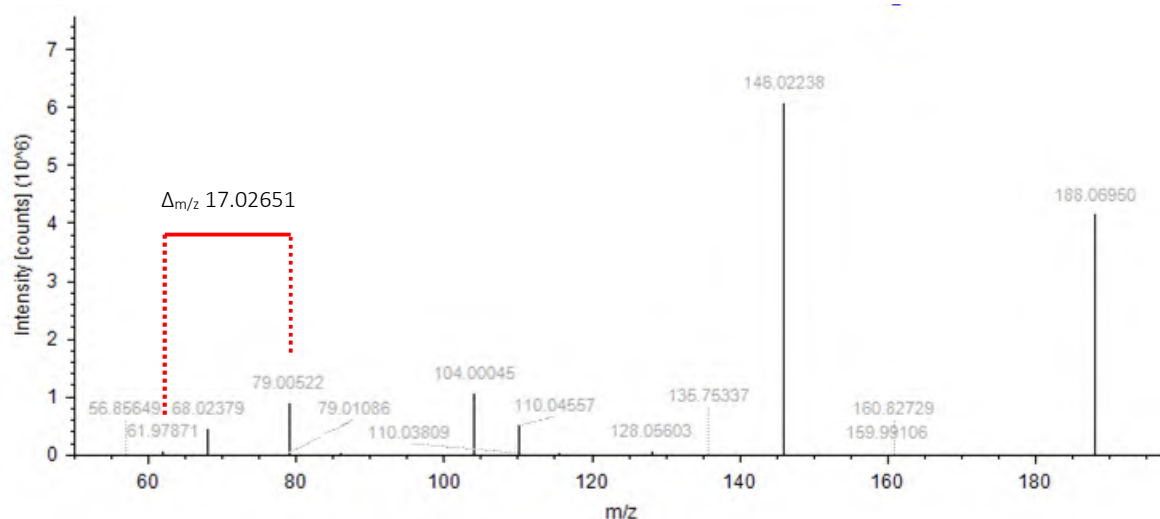


Figure 29 – MS2 spectrum of desethylatrazine, the diagnostic delta that functioned as MS2-trigger is marked with red.

The effect of concentration level of the spike-in on the MS2-triggers was tested as well, the results are illustrated in *figure 30*. At first, the precursor ion of the compound containing a structural alert has to

be selected for a MS2 scan, in which the MS2-trigger can be detected. Thereafter, this trigger can prompt the consecutive MS2 scan. Indeed, generally once a compound was detected and a MS2 scan recorded, an additional MS2 scan was triggered as well, indicating the sensitivity of the MS2-trigger. Some exceptions are marked in yellow in *figure 30*. In these measurements the compound was detected but no additional MS2 scans were triggered, probably due to the absence of the trigger in the MS2 scan or the low error tolerance (in case of ifosfamide). In one case no MS2 scan was recorded. Consequently, no additional MS2 scan could be triggered. This was the case for a single measurement of N(4)-acetylsulfadiazine at 1 µg/L.

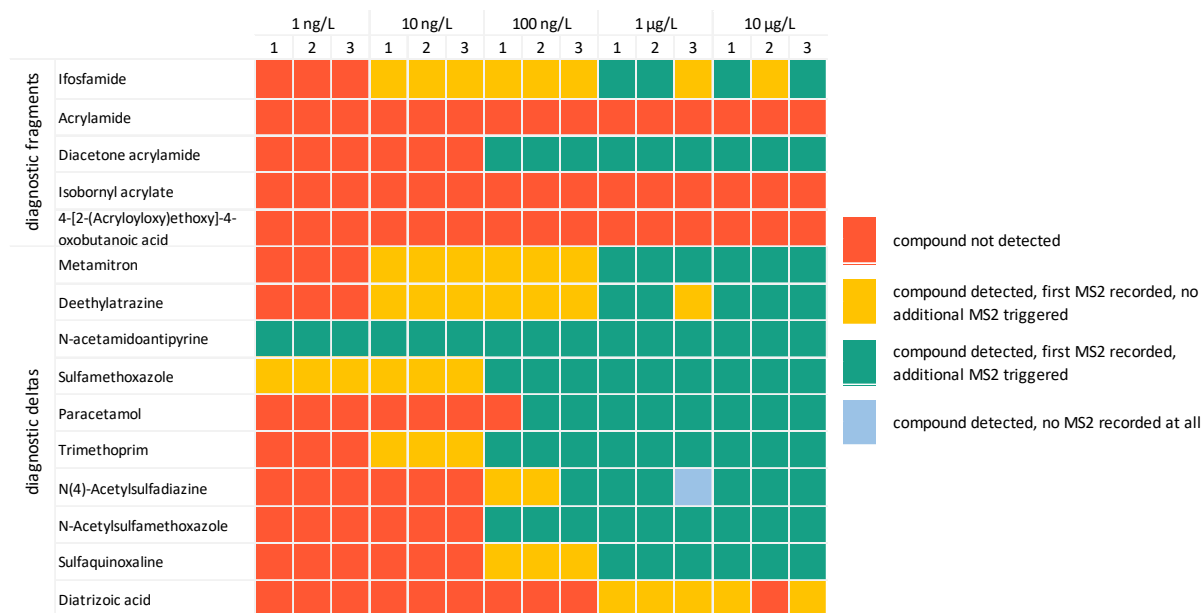


Figure 30 – Schematic overview representing the detection of the spike-in compounds, and whether an additional MS2 was triggered or not.

The acquisition parameters of the triggered MS2 scans were varied as well to determine their effects on the mzCloud scores assigned to the identified features. These settings were expected to have an effect on the spectral quality, and should thus facilitate identification. Three different settings were tested: stepped CE (10, 75, 90), assisted CE (20, 35, 50) and CE (20, 35, 50) in combination with a longer maximum IT of 200 ms instead of the regular 50 ms. Especially assisted CE and longer maximum ITs were expected to have a positive effect on the spectral quality. No clear differences were visible in the mzCloud scores, but the scores tended to increase slightly with the additional MS2 in assisted CE and the additional MS2 with longer IT. However, to state more confidently whether the additional MS2 indeed facilitates identification, spectral quality metrics or annotation studies have to be performed on these spectra. Moreover, the full potential of the MS2-triggers and the intelligent acquisition method can be reached only if spectral quality metrics are defined and the optimal acquisition parameters can be selected.

### Total performance evaluation

The addition of MS1- and MS2-triggers to the standard LC-HRMS method applied within KWR did have an effect on the data acquired during the measurements. The isotopic ratio MS1-trigger did not improve the percentage of Cl/Br containing compounds with a MS2 spectrum. The use of an inclusion list increased the percentage of MS2 spectra of features with m/z values present in the inclusion list.

The MS2-trigger method successfully triggered additional MS2 scans of molecules with a structural alert, for the four alerts that were tested. Therefore, the method could prioritize these potentially toxic compounds online. Next, the developed method needs to be compared with the regular KWR method in a NTS study to assess overall performance of the online prioritization workflow compared to offline prioritization. After identification it should be assessed which potentially toxic suspects lack a MS2 spectrum, whether this percentage is lower in the online intelligent acquisition method, and whether more potentially toxic compounds can be identified due to the additional scans prompted by the MS2-trigger. The next step would be to test the other alerts and test whether both triggers support the eventual data analysis in routine monitoring by online prioritization and yielding higher quality spectra. Despite the need for these additional tests, the new intelligent acquisition method based on structural alerts would be a promising method for the detection of organic micropollutants in drinking water sources.

## 5. Conclusions and future perspectives

### 5.1 Conclusions

This study shows that *in silico* fragmentation tools in combination with data mining in R can be used to find patterns in MS2 spectra of chemicals with the same structural alert. The derived patterns appear generally to be present in the experimental fragmentation spectra. The results of the acquisition parameter experiments suggested that more MS2 fragments could be annotated with increasing AGC-targets and higher maximum ITs and that assisted CE is preferred over stepped CE. Further research with a larger sample size is necessary to give a better indication of the change in spectral quality upon different AGC-targets and maximum IT settings.

The isotopic ratio MS1-trigger did not increase the percentage of Cl/Br containing compounds with an MS2 spectrum. More tolerant settings regarding the isotopic ratios might alleviate this. The inclusion list MS1-trigger increased the percentage of MS2 spectra of features with a mass present in the inclusion list. The MS2-triggers successfully triggered additional MS2 scans, for some compounds already at concentrations of 1 ng/L. This indicates the sensitivity of the MS2-triggers and the potential of online prioritization. Therefore, the intelligent acquisition method developed in this study is expected to be promising and should be developed further and expanded with more structural alerts to be able to apply it in routine monitoring eventually.

### 5.2 Outlook

Only a small set of structural alerts belonging to a few toxic endpoints was used in this study, as a proof of principle. The next step would be to find other structural alerts for other endpoints which could be relevant for the aquatic environment. As the *in silico* prediction of fragmentation spectra sometimes deviated from experimental results, it would also be possible to perform the pattern mining study on experimental results from a spectral library with good quality control and quality assurance.<sup>72</sup>

The pattern mining covered only recurring fragments and recurring deltas but it is also possible that a combination of both or a combination of multiple deltas and/or multiple fragments, or deltas including a specific intensity ratio is characteristic for a structural alert. These patterns could be searched using other tools such as machine learning strategies like random forest or software like the R-package mineMS2.<sup>73</sup> It also has to be researched what types are available to use as trigger in the acquisition software of the mass spectrometer.

The outcomes of this study indicate the need for further research and development of online intelligent acquisition software. This method could be more powerful if it is possible to assess the spectral quality online, so that during the acquisition can be decided whether an additional MS2 or MS3 is necessary. Interpretation of MS3 is currently also challenging to include in spectral annotation, especially for high-throughput identification. The inclusion of a MS3-trigger would be promising once this data can be handled. This trigger should activate a MS3 scan of for example the most intense ion in the MS2 scan if no sufficient peaks are present or the intensity distribution is not as good as desired (see *figure 5, scenario 3*). Furthermore, as high-throughput assessment of spectral quality is still challenging, it might be helpful to combine it with machine learning approaches to generate a model that is able to assess the quality of spectra.

The intelligent acquisition method described in this study is currently designed for an Orbitrap Fusion Tribrid mass spectrometer. Once the method is optimized and successful in prioritization of toxic

compounds in routine monitoring studies, the next step would be to develop comparable methods on other types of mass spectrometers, if possible. To cover a larger chemical space it would be fruitful to design a comparable method for GC-MS as well.

## 6. Acknowledgements

First of all, I would like to thank Andrea Brunner for supervising me and giving me the opportunity to develop myself within mass spectrometry and cheminformatics and taste a bit of what scientific research looks like. And, most of all, for being so supportive and positive (and critically) during my project, I have learned a lot and I became really motivated to continue within this field. I also thank Marja Lamoree, who was my examiner during this project and very encouraging as well. Dennis Vughs helped me a lot with the LC-HRMS experiments and all my questions regarding this, thank you very much. I also would like to thank Frederic Béen and Astrid Reus who helped me with the pattern mining and finding my way in the structural alerts, respectively. Thanks to Tessa Pronk for giving suggestions to improve my code in R and make it less time- and memory-consuming. And thanks to Margo van der Kooi for the sample preparation of the MS2-trigger experiments. Caroline Ding, Lena Becciolini and Seema Sharma helped me a lot as well, with questions regarding Compound Discoverer and the Orbitrap Fusion's method editor. Also thanks to Igor Tetko, for helping me with uploading alerts in ToxAlerts and making screening more easily by adjusting the software. Others that helped me during the cheminformatics part of this project were Christian Panse and Alexis Delabriere. And, last but not least, I would like to thank KWR Water Research Institute and in special team CWG for providing such a warm and welcoming working atmosphere and giving me the opportunity to perform my MSc thesis research within your team.



## 7. References

1. Stamm, C.; Räsänen, K.; Burdon, F. J.; Altermatt, F.; Jokela, J.; Joss, A.; Ackermann, M.; Eggen, R. I. L., Unravelling the Impacts of Micropollutants in Aquatic Ecosystems. In *Large-Scale Ecology: Model Systems to Global Perspectives*, 2016; pp 183-223.
2. Ruff, M.; Mueller, M. S.; Loos, M.; Singer, H. P., *Water Research* **2015**, *87*, 145-54.
3. Bernhardt, E. S.; Rosi, E. J.; Gessner, M. O., *Frontiers in Ecology and the Environment* **2017**, *15* (2), 84-90.
4. Brack, W.; Dulio, V.; Agerstrand, M.; Allan, I.; Altenburger, R.; Brinkmann, M.; Bunke, D.; Burgess, R. M.; Cousins, I.; Escher, B. I.; Hernandez, F. J.; Hewitt, L. M.; Hilscherova, K.; Hollender, J.; Hollert, H.; Kase, R.; Klauer, B.; Lindim, C.; Herraes, D. L.; Miede, C.; Munthe, J.; O'Toole, S.; Posthuma, L.; Rudel, H.; Schafer, R. B.; Sengl, M.; Smedes, F.; van de Meent, D.; van den Brink, P. J.; van Gils, J.; van Wezel, A. P.; Vethaak, A. D.; Vermeirssen, E.; von der Ohe, P. C.; Vrana, B., *Science of the Total Environment* **2017**, *576*, 720-737.
5. Schwarzenbach, R. P.; Escher, B. I.; Fenner, K.; Hofstetter, T. B.; Johnson, C. A.; von Gunten, U.; Wehrli, B., *Science* **2006**, *313*.
6. Bletsou, A. A.; Jeon, J.; Hollender, J.; Archontaki, E.; Thomaidis, N. S., *Trends in Analytical Chemistry* **2015**, *66*, 32-44.
7. Samanipour, S.; Martin, J. W.; Lamoree, M. H.; Reid, M. J.; Thomas, K. V., *Environmental Science and Technology* **2019**, *53* (10), 5529-5530.
8. Hollender, J.; Schymanski, E. L.; Singer, H. P.; Ferguson, P. L., *Environmental Science and Technology* **2017**, *51* (20), 11505-11512.
9. Brunner, A. M.; Dingemans, M. M. L.; Baken, K. A.; van Wezel, A. P., *Journal of Hazardous Materials* **2019**, *364*, 332-338.
10. NORMAN NORMAN Substance Database - NORMAN SusDat. <https://www.norman-network.com/nds/susdat/susdatSearchShow.php> (accessed November 12).
11. Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J., *Environmental Science and Technology* **2014**, *48* (4), 2097-8.
12. Schymanski, E. L.; Singer, H. P.; Slobodnik, J.; Ipolyi, I. M.; Oswald, P.; Krauss, M.; Schulze, T.; Haglund, P.; Letzel, T.; Grosse, S.; Thomaidis, N. S.; Bletsou, A.; Zwiener, C.; Ibanez, M.; Portoles, T.; de Boer, R.; Reid, M. J.; Onghena, M.; Kunkel, U.; Schulz, W.; Guillon, A.; Noyon, N.; Leroy, G.; Bados, P.; Bogialli, S.; Stipanicev, D.; Rostkowski, P.; Hollender, J., *Analytical and Bioanalytical Chemistry* **2015**, *407* (21), 6237-55.
13. Brunner, A. M.; Bertelkamp, C.; Dingemans, M. M. L.; Kolkman, A.; Wols, B.; Harmsen, D.; Siegers, W.; Martijn, B. J.; Oorthuizen, W. A.; Ter Laak, T. L., *Science of the Total Environment* **2020**, *705*, 135779.
14. Hollender, J.; van Bavel, B.; Dulio, V.; Farmen, E.; Furtmann, K.; Koschorreck, J.; Kunkel, U.; Krauss, M.; Munthe, J.; Schlabach, M.; Slobodnik, J.; Stroomborg, G.; Ternes, T.; Thomaidis, N. S.; Togola, A.; Tornero, V., *Environmental Sciences Europe* **2019**, *31* (1).
15. Brunner, A. M., Feature. 2019.
16. Brack, W.; Ait-Aissa, S.; Burgess, R. M.; Busch, W.; Creusot, N.; Di Paolo, C.; Escher, B. I.; Mark Hewitt, L.; Hilscherova, K.; Hollender, J.; Hollert, H.; Jonker, W.; Kool, J.; Lamoree, M.; Muschket, M.; Neumann, S.; Rostkowski, P.; Ruttkies, C.; Schollee, J.; Schymanski, E. L.; Schulze, T.; Seiler, T. B.; Tindall, A. J.; De Aragao Umbuzeiro, G.; Vrana, B.; Krauss, M., *Science of the Total Environment* **2016**, *544*, 1073-118.
17. Zwart, N.; Nio, S. L.; Houtman, C. J.; de Boer, J.; Kool, J.; Hamers, T.; Lamoree, M. H., *Environmental Science and Technology* **2018**, *52* (7), 4367-4377.
18. Zwart, N.; Jonker, W.; Broek, R. T.; de Boer, J.; Somsen, G.; Kool, J.; Hamers, T.; Houtman, C. J.; Lamoree, M. H., *Water Research* **2020**, *168*, 115204.
19. Hoffmann, E. d.; Stroobant, V., *Mass Spectrometry: Principles and Applications*. Third Edition ed.; John Wiley & Sons, Ltd: Chichester, England, 2007.
20. University of Washington's Proteomics Resource Fusion - Orbitrap Fusion Tribrid MS. <https://proteomicsresource.washington.edu/instruments/orbitrapfusion.php> (accessed 11 June 2020).
21. Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M., *Nature Methods* **2007**, *4* (9), 709-712.
22. McDonald, J. G.; Matthew, S.; Auchus, R. J., *Hormones and Cancer* **2011**, *2* (6), 324-32.
23. Sjerps, R. M. A.; Vughs, D.; van Leerdam, J. A.; Ter Laak, T. L.; van Wezel, A. P., *Water Res* **2016**, *93*, 254-264.
24. Schmidt, A.; Claassen, M.; Aebersold, R., *Current Opinion in Chemical Biology* **2009**, *13* (5-6), 510-7.
25. Thermo Fisher Scientific Schematic of the Orbitrap Fusion Tribrid MS. <https://planetorbitrap.com/orbitrap-fusion#tab:schematic> (accessed November 11, 2019).

26. Bailey, D.; McAlister, G. C.; Sharma, S.; Remes, P. M.; Tautenhahn, R.; Ntai, I., Real-time collisional energy optimization on the Orbitrap Fusion platform for confident unknown identification. *Scientific, T. F.*, Ed. 2018.
27. Kalli, A.; Smith, G. T.; Sweredoski, M. J.; Hess, S., *Journal of Proteome Research* **2013**, *12*, 3071-3086.
28. Thermo Fisher Scientific *AcquireX Data Acquisition*.
29. HighChem LLC mzCloud Features. <https://www.mzcloud.org/Features> (accessed November 25, 2019).
30. Schulze, T. European MassBank (NORMAN MassBank). [https://www.norman-network.com/sites/default/files/files/suspectListExchange/031017Update/MassBankEU\\_Cmpds\\_11042017\\_wM\\_S\\_DTXSIDs\\_03102017.xlsx](https://www.norman-network.com/sites/default/files/files/suspectListExchange/031017Update/MassBankEU_Cmpds_11042017_wM_S_DTXSIDs_03102017.xlsx) (accessed 12 December 2019).
31. Thermo Fisher Scientific *Compound Discoverer*.
32. Schulze, T.; Schymanski, E.; Stravs, M.; Neumann, S.; Krauss, M.; Singer, H.; Hug, C.; Gallampois, C.; Hollender, J.; Slobodnik, J.; Brack, W., NORMAN MassBank Towards a community-driven, open-access accurate mass spectral database for the identification of emerging pollutants. *NORMAN Bulletin* April 2012, 2012, pp 9-10.
33. NORMAN Substance Database <https://www.norman-network.com/nds/susdat/susdatSearchShow.php#> (accessed 3 February 2020).
34. Letzel, T.; Grosse, S.; Sengel, M., HSWT/LfU STOFF-IDENT Database of Water-Relevant Substances. 2017.
35. Sjerps, R. M. A.; Vughs, D.; van Leerdam, J. A.; ter Laak, T.; van Wezel, A. P. KWRSJERPS2. [https://www.norman-network.com/sites/default/files/files/suspectListExchange/190618Update/Sjerp\\_2016\\_WatResManuscript\\_Sl.docx](https://www.norman-network.com/sites/default/files/files/suspectListExchange/190618Update/Sjerp_2016_WatResManuscript_Sl.docx) (accessed 19 February 2020).
36. Von der Ohe, P.; Fischer, S. NORMAN UBAMPT. <https://www.norman-network.com/nds/SLE/> (accessed 12 February 2020).
37. National Center for Computational Toxicology Chemistry Dashboard. <https://comptox.epa.gov/dashboard/> (accessed 5 December 2019).
38. Richard, A. M.; Judson, R. S.; Houck, K. A.; Grulke, C. M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M. T.; Wambaugh, J. F.; Knudsen, T. B.; Kancherla, J.; Mansouri, K.; Patlewicz, G.; Williams, A. J.; Little, S. B.; Crofton, K. M.; Thomas, R. S., *Chemical Research in Toxicology* **2016**, *29* (8), 1225-51.
39. Royal Society of Chemistry ChemSpider. <http://www.chemspider.com>.
40. Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S., *Journal of Cheminformatics* **2016**, *8*, 3.
41. Allen, F.; Greiner, R.; Wishart, D., *Metabolomics* **2015**, *11*, 98-110.
42. Ruttkies, C.; Neumann, S.; Helmchen, A. *metfRag: Identification of metabolites using mass spectrometry data*, 2.4.2; 2017.
43. Allen, F.; Russel, G.; Wishart, D., *Current Metabolomics* **2017**, *5* (1), 35-39.
44. Djoumbou-Feunang, Y.; Pon, A.; Karu, N.; Zheng, J.; Li, C.; Arndt, D.; Gautam, M.; Allen, F.; Wishart, D. S., *Metabolites* **2019**, *9* (72).
45. Allen, F. CFM-ID: Competitive Fragmentation Modeling for Metabolite Identification. <https://sourceforge.net/p/cfm-id/wiki/Home/> (accessed 20 December 2019).
46. Brunner, A. M., QTOF and Orbitrap HCD data. Meekel, N., Ed. Nieuwegein, 2020.
47. Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V., *Journal of Chemical Information and Modeling* **2012**, *52* (8), 2310-6.
48. Ridings, J. E.; Barrat, M. D.; Cary, R.; Earnshaw, C. G.; Eggington, C. E.; Ellis, M. K.; Judson, P. N.; Langowski, J. J.; Marchant, C. A.; Payne, M. P.; Watson, W. P.; Yih, T. D., *Toxicology* **1996**, *106*, 267-279.
49. Saiakhov, R. D.; Klopman, G., *Toxicology Mechanisms and Methods* **2008**, *18* (2-3), 159-75.
50. European Chemicals Agency (ECHA) *Grouping of substances and read-across approach - an illustrative example*; Helsinki, Finland, 2013.
51. Benigni, R.; Bossa, C., *Mutat Res* **2008**, *659* (3), 248-61.
52. Organisation for Economic Co-operation and Development (OECD), Collection of working definitions. 2012.
53. Ashby, J.; Tennant, R. W., *Mutation Research* **1988**, *204*, 17-115.
54. Bailey, A. B.; Chanderbhan, R.; Collazo-Braier, N.; Cheeseman, M. A.; Twaroski, M. L., *Regul Toxicol Pharmacol* **2005**, *42* (2), 225-35.
55. Kazius, J.; McGuire, R.; Bursi, R., *Journal of Medicinal Chemistry* **2005**, *48* (1), 312-320.
56. Kazius, J.; Nijssen, S.; Kok, J.; Bäck, T.; Ilzerman, A. P., *Journal of Chemical Information and Modeling* **2006**, *46* (2), 597-605.

57. Reus, A., Structural Alerts and Toxic Endpoints. Meekel, N., Ed. Nieuwegein, 2019.
58. Alves, V.; Muratov, E.; Capuzzi, S.; Politi, R.; Low, Y.; Braga, R.; Zakharov, A. V.; Sedykh, A.; Mokshyna, E.; Farag, S.; Andrade, C.; Kuz'min, V.; Fourches, D.; Tropsha, A., *Green Chem* **2016**, *18* (16), 4348-4360.
59. Benigni, R.; Bossa, C., *Current Computer-Aided Drug Design* **2006**, *2*, 1-19.
60. United States Environmental Protection Agency Chemical\_Summary\_190708 from invitrodb\_v3.2. <https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data> (accessed 4 December 2019).
61. Nendza, M.; Wenzel, A.; Muller, M.; Lewin, G.; Simetska, N.; Stock, F.; Arning, J., *Environ Sci Eur* **2016**, *28* (1), 26.
62. R Core Team *R: A language and environment for statistical computing*, 3.6.1; R Foundation for Statistical Computing: Vienna, Austria, 2019.
63. United States Environmental Protection Agency ac50\_Matrix\_190708 from invitrodb\_v3.2. <https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data> (accessed 4 December 2019).
64. Vughs, D., logP chart. Meekel, N., Ed. Nieuwegein, 2020.
65. EPA, U. S. DSSTox MS Ready Mapping File. <https://comptox.epa.gov/dashboard/downloads> (accessed 12 February 2020).
66. Nesvizhki, A. I.; Roos, F. F.; Grossmann, J.; Vogelzang, M.; Eddes, J. S.; Gruissem, W.; Baginsky, S.; Aebersold, R., *Molecular & Cellular Proteomics* **2006**, *5* (4), 652-670.
67. Dodder, N.; Mullen, K. *OrgMassSpecR: Organic Mass Spectrometry*, 0.5-3; 2017.
68. Lousse, J.; Dingemans, M. M. L.; Baken, K. A.; van Wezel, A. P.; Schriks, M., *Chemosphere* **2018**, *209*, 373-380.
69. Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q. Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V., *J Comput Aided Mol Des* **2011**, *25* (6), 533-54.
70. Vughs, D., AGC target experimenten. Meekel, N., Ed. 2020.
71. Böcker, S.; Dührkop, K., *Journal of Cheminformatics* **2016**, *8* (5), 1-26.
72. Oberacher, H.; Sasse, M.; Antignac, J.-P.; Guitton, Y.; Debrauwer, L.; Jamin, E. L.; Schulze, T.; Krauss, M.; Covaci, A.; Caballero-Casero, N.; Rousseau, K.; Damont, A.; Fenaille, F.; Lamoree, M.; Schymanski, E. L., *Environmental Sciences Europe* **2020**, *32* (1).
73. Delabriere, A. *mineMS2*, 0.9.7; 2020.

## Appendix A – Examples of structural alerts

Table A.1 – Structural Alerts list by Benigni & Bossa (2008) with respect to mutagenicity and carcinogenicity.<sup>51</sup>

Structural Alert		Details
SA_1 acyl halides		R = any atom/group, except OH, SH
SA_2 alkyl (C<5) or benzyl ester of sulphonic or phosphonic acid		R = alkyl with C < 5 (potentially substituted by halogens), or benzyl R1 = any atom/group except OH, SH, O <sup>-</sup> , S <sup>-</sup>
SA_3 N-methylol derivatives		R = any atom/group
SA_4 monohaloalkene		R1, R2 (or R3) = H or CH3 R3 (or R2) = any atom/group except halogens
SA_5 S or N mustard		R = any atom/group
SA_6 propiolactones propiosultones		Any substance containing any of the displayed substructures
SA_7 epoxides and aziridines		R = any atom/group
SA_8 aliphatic halogens		R = any atom/group
SA_9 alkyl nitrite		R = any alkyl group
SA_10 α,β unsaturated carbonyls		R1 and R2 = any atom/group, except alkyl chains with C > 5 or aromatic rings. R = any atom/group, except OH, O <sup>-</sup>
SA_11 simple aldehyde		R = aliphatic or aromatic carbon α,β unsaturated aldehydes are excluded
SA_12 quinones		Any substance containing any of the displayed substructures
SA_13 hydrazine		R = any atom/group
SA_14 aliphatic azo and azoxy		R1 = aliphatic carbon or hydrogen R2, R3 = any atom/group R4 = aliphatic carbon
SA_15 isocyanate and isothiocyanate groups		R = any atom/group
SA_16 alkyl carbamate and thiocarbamate		R = aliphatic carbon or hydrogen R1 = aliphatic carbon
SA_17 thiocarbonyl (nongenotoxic)		R, R1, R2 = any atom/group R3 = any atom/group except OH, SH, O <sup>-</sup> , S <sup>-</sup> carbamate and thiocarbamate are excluded
SA_18 polycyclic aromatic hydrocarbons		Three or more fused rings, not heteroaromatic
SA_19 heterocyclic polycyclic aromatic hydrocarbons		Three or more fused rings, heteroaromatic
SA_20 (poly)halogenated cycloalkanes (nongenotoxic)		Any cycloalkane skeleton with three or more halogens directly bound to the same ring
SA_21 alkyl and aryl N-nitroso groups		R1 = aliphatic or aromatic carbon R2 = any atom/group

Structural Alert		Details
SA_22 azide and triazene groups		R = any atom/group
SA_23 aliphatic N-nitro group		R = aliphatic carbon or hydrogen
SA_24 α,β unsaturated aliphatic alkoxy group		R <sub>1</sub> = any aliphatic carbon R <sub>2</sub> = aliphatic or aromatic carbon
SA_25 aromatic nitroso group		Ar = any aromatic/heteroaromatic ring
SA_26 aromatic ring N-oxide		Any aromatic or heteroaromatic ring
SA_27 nitro-aromatic		Ar = any aromatic/heteroaromatic ring Aromatic nitro groups with <i>ortho</i> -disubstitution or with a carboxylic acid substituent in <i>ortho</i> position should be excluded. If a sulfonic acid group (-SO <sub>3</sub> H) is present on the ring that contains also the nitro group, the substance should be excluded.
SA_28 primary aromatic amine, hydroxyl amine and its derived esters or amine generating group		Ar = any aromatic/heteroaromatic ring R = any atom/group Aromatic amino groups with <i>ortho</i> -disubstitution or with a carboxylic acid substituent in <i>ortho</i> position should be excluded. If a sulfonic acid group (-SO <sub>3</sub> H) is present on the ring that contains also the amino group, the substance should be excluded from the alert.
SA_28 bis aromatic mono- and dialkylamine		Ar = any aromatic/heteroaromatic ring R <sub>1</sub> = hydrogen, methyl, ethyl R <sub>2</sub> = methyl, ethyl Aromatic amino groups with <i>ortho</i> -disubstitution or with a carboxylic acid substituent in <i>ortho</i> position should be excluded. If a sulfonic acid group (-SO <sub>3</sub> H) is present on the ring that contains also the amino group, the substance should be excluded from the alert
SA_28 ter aromatic N-acyl amine		Ar = any aromatic/heteroaromatic ring R = hydrogen, methyl Aromatic amino groups with <i>ortho</i> -disubstitution or with a carboxylic acid substituent in <i>ortho</i> position should be excluded. If a sulfonic acid group (-SO <sub>3</sub> H) is present on the ring that contains also the amino group, the substance should be excluded from the alert.
SA_29 aromatic diazo		Ar = any aromatic/heteroaromatic ring If a sulfonic acid group (-SO <sub>3</sub> H) is present on each of the rings that contain the diazo group, the substance should be not classified.
SA_30 coumarins and Furocoumarins		Any substance containing the displayed substructure.
SA_31a halogenated benzene (nongenotoxic)		If two halogens are present in <i>ortho</i> or <i>meta</i> positions, the substance should be not classified. If three or more hydroxyl groups are present, the substance should be not classified.
SA_31b halogenated PAH (nongenotoxic)		Ar = naphthalene, biphenyl, diphenyl
SA_31c halogenated dibenzodioxins (nongenotoxic)		X = F, Cl, Br, I Only chemicals with at least one halogen in one of the four lateral positions are included.

## Appendix B – Workflow screening with ToxAlerts and fragmentation


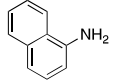
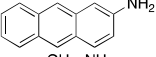
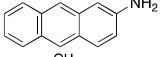










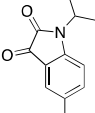
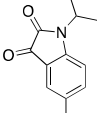
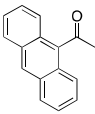
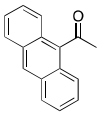
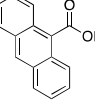
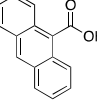
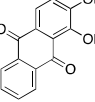
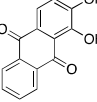
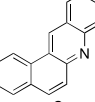
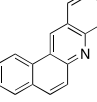


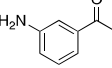

1. Subtract CAS-registry numbers (CASRN) from **Chemical\_Summary\_190708.csv** (located in INVITRODB\_V3\_2\_SUMMARY.zip, downloaded via [https://epa.figshare.com/articles/ToxCast\\_and\\_Tox21\\_Summary\\_Files/6062479](https://epa.figshare.com/articles/ToxCast_and_Tox21_Summary_Files/6062479)) and generate an Excel file. Use **Step01\_casrn\_to\_msreadysmiles.R** and follow instructions.
2. In ToxAlerts (<https://ochem.eu/alerts/home.do>):
  - a. Select “View alerts”
  - b. Select all records matching the endpoints ‘non-genotoxic carcinogenicity’, ‘genotoxic carcinogenicity, mutagenicity’ and ‘developmental and mitochondrial toxicity’ (only approved alerts).
  - c. Select “Screen compounds against alerts”
  - d. Upload the file containing MS-ready SMILES (**ToxCast\_msready\_smiles.xlsx**)
  - e. Untick all boxes in ‘Preprocessing of molecules (Chemaxon)’
  - f. Tick the box “Only 152 selected alerts” and “Only approved alerts”
  - g. Start screening.
3. After screening, export results as .csv file (Structure and Descriptors) and import file in R for further processing. Use **Step03\_import\_toxalerts.R**
4. Fragment the molecules per structural alert using CFM-ID.

```
> cfm-predict.exe ToxCast\TA322.txt 0.001 metab_se_cfm\param_output0.log
metab_se_cfm\param_config.txt 0 Output_ToxCast\TA322 0 0
```

All error messages such as “Could not ionize – already charged molecule and didn’t know what to do here” and “SMILES Parse Error: syntax error for input: XXX” were collected on screenshots of the command window.
5. Fragment the molecules per structural alert using MetFrag in R. Use **Step05\_fragmentation\_metfrag.R**

## Appendix C – Comparison in- and output of ToxAlerts

Table C.1 – Test of differences input and output ToxAlerts with compounds retrieved from MassBankEU.

Input SMILES	Structure	Output SMILES	Structure
<chem>NC1=CC=CC2=CC=CC=C12</chem>		<chem>NC1=C2C=CC=CC=C1</chem>	
<chem>NC1=CC2=CC3=CC=CC=C3C=C2C=C1</chem>		<chem>NC1=CC2=CC3=C(C=CC=C3)C=C2C=C1</chem>	
<chem>NC1=C(C=CC=C1)C(O)=O</chem>		<chem>NC1=CC=CC=C1C(O)=O</chem>	
<chem>NC1=C(C=CC=C1)C1=CC=CC=C1</chem>		<chem>NC1=C(C=CC=C1)C1=CC=CC=C1</chem>	
<chem>NC1=C(O)C=CC=C1</chem>		<chem>NC1=CC=CC=C1O</chem>	
<chem>NC1=NC=CC=C1</chem>		<chem>NC1=NC=CC=C1</chem>	
<chem>CC(=O)NC1=C(C)C=CC=C1</chem>		<chem>CC(=O)NC1=C(C)C=CC=C1</chem>	
<chem>CC(C)N1C(=O)C(=O)C2=CC(C)=CC=C12</chem>		<chem>CC(C)N1C(=O)C(=O)C2=CC(C)=CC=C12</chem>	
<chem>CC(=O)C1=C2C=CC=CC2=CC=CC=C12</chem>		<chem>CC(=O)C1=C2C=CC=CC2=CC2=C1C=CC=C2</chem>	
<chem>OC(=O)C1=C2C=CC=CC2=CC=CC=C12</chem>		<chem>OC(=O)C1=C2C=CC=CC2=CC2=CC=CC=C12</chem>	
<chem>OC1=CC=C2C(=O)C3=CC=CC=C3C(=O)C2=C1O</chem>		<chem>OC1=CC=C2C(=O)C3=C(C=CC=C3)C(=O)C2=C1O</chem>	
<chem>C1=CC=C2C(C=CC3=NC4=CC=CC=C4C=C23)=C1</chem>		<chem>C1=CC2=C(C=C1)N=C1C=CC3=CC=CC=C3C1=C2</chem>	
<chem>COC(=O)C1=CC=CC2=CC=CC=C12</chem>		<chem>COC(=O)C1=C2C=CC=CC2=CC=C1</chem>	
<chem>CC(=O)C1=CC(N)=CC=C1</chem>		<chem>CC(=O)C1=CC(N)=CC=C1</chem>	

## Appendix D – Assays applied for toxicity validation

Table D.1 – ToxCast assays used for toxicity validation listed per toxic endpoint.<sup>9</sup>

Endocrine disruption	Non-genotoxic carcinogenicity, genotoxic carcinogenicity, mutagenicity	Developmental and mitochondrial toxicity
ACEA_T47D_80hr_Positive	APR_HepG2_p53Act_1h_dn	NHEERL_ZF_144hpf_TERATOSCORE_up
ATG_ERE_CIS_up	APR_HepG2_p53Act_1h_up	Tanguay_ZF_120hpf_ActivityScore
ATG_Era_TRANS_up	APR_HepG2_p53Act_24h_dn	Tanguay_ZF_120hpf_AXIS_up
NVS_NR_bER	APR_HepG2_p53Act_24h_up	Tanguay_ZF_120hpf_BRAI_up
NVS_NR_hER	APR_HepG2_p53Act_72h_dn	Tanguay_ZF_120hpf_CFIN_up
NVS_NR_mERa	APR_HepG2_p53Act_72h_up	Tanguay_ZF_120hpf_CIRC_up
OT_ER_EraEra_0480	ATG_Ahr_CIS_up	Tanguay_ZF_120hpf_EYE_up
OT_ER_EraEra_1440	ATG_AP_1_CIS_up	Tanguay_ZF_120hpf_JAW_up
OT_ER_EraERb_0480	ATG_AP_2_CIS_up	Tanguay_ZF_120hpf_NC_up
OT_ER_EraERb_1440	ATG_BRE_CIS_up	Tanguay_ZF_120hpf_OTIC_up
OT_Era_EREGFP_0120	ATG_C_EBP_CIS_up	Tanguay_ZF_120hpf_PE_up
OT_Era_EREGFP_0480	ATG_CRE_CIS_up	Tanguay_ZF_120hpf_PFIN_up
TOX21_Era_BLA_Agonist_ratio	ATG_E_Box_CIS_up	Tanguay_ZF_120hpf_PIG_up
TOX21_Era_BLA_Antagonist_ratio	ATG_E2F_CIS_up	Tanguay_ZF_120hpf_SNOU_up
TOX21_Era_LUC_BG1_Agonist	ATG_EGR_CIS_up	Tanguay_ZF_120hpf_SOMI_up
TOX21_Era_LUC_BG1_Antagonist	ATG_Ets_CIS_up	Tanguay_ZF_120hpf_SWIM_up
OT_ER_ERbERb_0480	ATG_FoxA2_CIS_up	Tanguay_ZF_120hpf_TR_up
OT_ER_ERbERb_1440	ATG_FoxO_CIS_up	Tanguay_ZF_120hpf_TRUN_up
ATG_ERb_TRANS2_up	ATG_GATA_CIS_up	Tanguay_ZF_120hpf_YSE_up
ATG_Era_TRANS_dn	ATG_GLI_CIS_up	
ATG_ERE_CIS_dn	ATG_HIF1a_CIS_up	
ATG_ERRb_TRANS2_up	ATG_HNF6_CIS_up	
ATG_ERRa_TRANS_up	ATG_HSE_CIS_up	
ATG_ERRg_TRANS_up	ATG_ISRE_CIS_up	
CEETOX_H295R_ESTRADIOL_dn	ATG_MRE_CIS_up	
CEETOX_H295R_ESTRADIOL_up	ATG_Myb_CIS_up	
CEETOX_H295R ESTRONE_dn	ATG_Myc_CIS_up	
CEETOX_H295R ESTRONE_up	ATG_NF_kB_CIS_up	
ATG_AR_TRANS_up	ATG_NFI_CIS_up	
NVS_NR_cAR	ATG_NRF1_CIS_up	
NVS_NR_hAR	ATG_NRF2_ARE_CIS_up	
NVS_NR_rAR	ATG_Oct_MLP_CIS_up	
OT_AR_ARELUC_AG_1440	ATG_p53_CIS_up	
OT_AR_ARSRC1_0480	ATG_Pax6_CIS_up	
OT_AR_ARSRC1_0960	ATG_Sox_CIS_up	
TOX21_AR_BLA_Agonist_ratio	ATG_Sp1_CIS_up	
TOX21_AR_BLA_Antagonist_ratio	ATG_SREBP_CIS_up	
TOX21_AR_LUC_MDAKB2_Agonist	ATG_STAT3_CIS_up	
TOX21_AR_LUC_MDAKB2_Antagonist	ATG_TCF_b_cat_CIS_up	
ATG_AR_TRANS_dn	ATG_Xbp1_CIS_up	
CEETOX_H295R_ANDR_dn	TOX21_AhR_LUC_Agonist	
CEETOX_H295R_ANDR_up	TOX21_ARE_BLA_agonist_ratio	
CEETOX_H295R_TESTO_dn	TOX21_HSE_BLA_agonist_ratio	
CEETOX_H295R_TESTO_up	TOX21_p53_BLA_p1_ratio	
	TOX21_p53_BLA_p2_ratio	
	TOX21_p53_BLA_p3_ratio	
	TOX21_p53_BLA_p4_ratio	
	TOX21_p53_BLA_p5_ratio	
	TOX21_ESRE_BLA_ratio	
	TOX21_NFkB_BLA_agonist_ratio	
	ATG_Ahr_CIS_dn	
	ATG_AP_1_CIS_dn	
	ATG_AP_2_CIS_dn	
	ATG_BRE_CIS_dn	
	ATG_CRE_CIS_dn	
	ATG_C_EBP_CIS_dn	
	ATG_E2F_CIS_dn	
	ATG_EGR_CIS_dn	
	ATG_Ets_CIS_dn	
	ATG_E_Box_CIS_dn	
	ATG_FoxA2_CIS_dn	
	ATG_FoxO_CIS_dn	



	ATG_GATA_CIS_dn	
	ATG_GLI_CIS_dn	
	ATG_HIF1a_CIS_dn	
	ATG_HNF6_CIS_dn	
	ATG_HSE_CIS_dn	
	ATG_ISRE_CIS_dn	
	ATG_MRE_CIS_dn	
	ATG_Myb_CIS_dn	
	ATG_Myc_CIS_dn	
	ATG_NFI_CIS_dn	
	ATG_NF_kB_CIS_dn	
	ATG_NRF1_CIS_dn	
	ATG_NRF2_ARE_CIS_dn	
	ATG_Oct_MLP_CIS_dn	
	ATG_p53_CIS_dn	
	ATG_Pax6_CIS_dn	
	ATG_Sox_CIS_dn	
	ATG_Sp1_CIS_dn	
	ATG_SREBP_CIS_dn	
	ATG_STAT3_CIS_dn	
	ATG_TCF_b_cat_CIS_dn	
	ATG_Xbp1_CIS_dn	

## Appendix E – List of chemicals

Table E.1 – Compounds present in the LOA600 + specials spike.

Chemical name	Bruto formula	CAS Registry Number	Ionization (+/-)	Accurate mass [M+H] <sup>+</sup> or [M-H] <sup>-</sup>	RT (min)
(4-chloro-2-methylphenoxy)acetic acid (MCPA)	C9H9ClO3	94-74-6	[M+H] <sup>-</sup>	199.01675	15.31
1-(3,4-dichlorophenyl)-3-methylurea	C8H8Cl2N2O	3567-62-2	[M+H] <sup>+</sup>	219.00864	14.28
1-(3,4-Dichlorophenyl)-urea	C7H6Cl2N2O	2327-02-8	[M+H] <sup>+</sup>	204.99299	13.29
10,11-dihydro-10,11-dihydroxycarbamazepine	C15H14N2O3	35079-97-1	[M+H] <sup>+</sup>	271.10771	7.55
1-H-benzotriazol	C6H5N3	95-14-7	[M+H] <sup>+</sup>	120.05562	7.92
2-(methylthio)benzothiazool	C8H7NS2	615-22-5	[M+H] <sup>+</sup>	182.00926	17.38
2,4-Dichloroaniline	C6H5Cl2N	554-00-7	[M+H] <sup>+</sup>	161.98718	16.77
2,4-Dichlorophenoxyacetic acid (2,4-D)	C8H6Cl2O3	94-75-7	[M+H] <sup>-</sup>	218.96212	15.25
2,6-dichlorobenzamide (BAM)	C7H5Cl2NO	2008-58-4	[M+H] <sup>+</sup>	189.98210	8.18
2.4.6-trichlorophenol	C6H3Cl3O	88-06-2	[M+H] <sup>-</sup>	194.91766	17.77
2.4-dichlorophenol	C6H4Cl2O	120-83-2	[M+H] <sup>-</sup>	160.95664	16.52
2.4-dinitrophenol	C6H4N2O5	51-28-5	[M+H] <sup>-</sup>	183.00474	13.21
2-aminoacetophenone	C8H9NO	551-93-9	[M+H] <sup>+</sup>	136.07569	11.93
2-aminobenzothiazool	C7H6N2S	136-95-8	[M+H] <sup>+</sup>	151.03244	6.43
2-hydroxybenzothiazool	C7H5NOS	934-34-9	[M+H] <sup>+</sup>	152.01646	11.59
2-methyl-4.6-dinitrophenol (DNOC)	C7H6N2O5	534-52-1	[M+H] <sup>-</sup>	197.02039	19.64
4-methyl-1H-benzotriazol	C7H7N3	29878-31-7	[M+H] <sup>+</sup>	134.07127	9.94
5,6-dimethyl-1H-benzotriazol	C8H9N3	4184-79-6	[M+H] <sup>+</sup>	148.08692	11.52
5-chloor-1H-benzotriazol	C6H4ClN3	94-97-3	[M+H] <sup>+</sup>	154.01665	11.3
5-methyl-1H-benzotriazol	C7H7N3	136-85-6	[M+H] <sup>+</sup>	134.07127	10.07
atrazin	C8H14ClN5	1912-24-9	[M+H] <sup>+</sup>	216.10105	14.54
atrazin-d5	C8H9 2H5ClN5	163165-75-1	[M+H] <sup>+</sup>	221.13243	14.46
azinphos-methyl	C10H12N3O3PS2	86-50-0	[M+H] <sup>+</sup>	318.01305	17.17
bentazon	C10H12N2O3S	25057-89-0	[M+H] <sup>-</sup>	239.04958	14.44
bentazone-d6	C10H6 2H6N2O3S	n/a	[M+H] <sup>-</sup>	245.08725	14.38
Benzotriazole-d4	C6H 2H4N3	n/a	[M+H] <sup>+</sup>	124.08073	7.86
bezafibrate	C19H20ClNO4	41859-67-0	[M+H] <sup>+</sup>	362.11536	15.95
bromacil	C9H13BrN2O2	314-40-9	[M+H] <sup>+</sup>	261.02332	12.43
caffeine	C8H10N4O2	58-08-2	[M+H] <sup>+</sup>	195.08765	6.83
candesartan	C24H20N6O3	139481-59-7	[M+H] <sup>+</sup>	441.16696	14.37
carbamazepin	C15H12N2O	298-46-4	[M+H] <sup>+</sup>	237.10224	13.27
carbendazim	C9H9N3O2	10605-21-7	[M+H] <sup>+</sup>	192.07675	6.38
cetirizine	C21H25ClN2O3	83881-51-0	[M+H] <sup>+</sup>	389.16265	14
Chlooroxuron	C15H15ClN2O2	1982-47-4	[M+H] <sup>+</sup>	291.08948	17.37
chloridazon	C10H8ClN3O	1698-60-8	[M+H] <sup>+</sup>	222.04287	9.79
Chlorpyrifos-ethyl	C9H11Cl3NO3PS	2921-88-2	[M+H] <sup>+</sup>	349.93356	23.34
chlortoluron	C10H13ClN2O	15545-48-9	[M+H] <sup>+</sup>	213.07892	14.31
DEET	C12H17NO	134-62-3	[M+H] <sup>+</sup>	192.13829	14.83
desethylatrazin	C6H10ClN5	6190-65-4	[M+H] <sup>+</sup>	188.06975	9.78
Desfenylchloridazon	C4H4ClN3O	6339-19-1	[M+H] <sup>+</sup>	146.01156	2.25
desisopropylatrazin	C5H8ClN5	1007-28-9	[M+H] <sup>+</sup>	174.05410	7.69
dichlorprop (2.4-DP)	C9H8Cl2O3	120-36-5	[M+H] <sup>-</sup>	232.97777	16.52
diclofenac	C14H11Cl2NO2	15307-86-5	[M+H] <sup>+</sup>	296.02396	18.37
dimethenamid-P	C12H18ClNO2S	163515-14-8	[M+H] <sup>+</sup>	276.08195	17.37
dimethoate	C5H12NO3PS2	60-51-5	[M+H] <sup>+</sup>	230.0069	10.29
Dimethomorph (isomer 1)	C21H22ClNO4	110488-70-5	[M+H] <sup>+</sup>	388.13101	16.18
Dimethomorph (isomer 2)	C21H22ClNO4	110488-70-5	[M+H] <sup>+</sup>	388.13101	16.59
diuron	C9H10N2OCl2	330-54-1	[M+H] <sup>+</sup>	233.02429	15.07
Ethofumesate	C13H18O5S	26225-79-6	[M+H] <sup>+</sup>	287.09477	18.48
fenuron	C9H12N2O	101-42-8	[M+H] <sup>+</sup>	165.10224	9.43
Gabapentine	C9H17NO2	60142-96-3	[M+H] <sup>+</sup>	172.13320	6.45
Gabapentine-lactam	C9H15NO	64744-50-9	[M+H] <sup>+</sup>	154.12264	11.22
HMMM	C15H30N6O6	3089-11-0	[M+H] <sup>+</sup>	391.22996	13.24

Hydrochlorothiazide	C7H8ClN3O4S2	58-93-5	[M+H]-	295.95719	7.2
irbesartan	C25H28N6O	138402-11-6	[M+H]+	429.23974	14.13
isoproturon	C12H18N2O	34123-59-6	[M+H]+	207.14919	14.93
Lamotrigine	C9H7Cl2N5	84057-84-1	[M+H]+	256.01512	9.36
Linuron	C9H10Cl2N2O2	330-55-2	[M+H]+	249.01921	17.24
mecoprop (MCP)	C10H11ClO3	93-65-2	[M+H]-	213.03239	16.53
metazachlor	C14H16ClN3O	67129-08-2	[M+H]+	278.10547	15.87
Metazachlor ESA	C14H17N3O4S	172960-62-2	[M+H]+	324.10125	9.19
metazachlor OA	C14H15N3O3	1231244-60-2	[M+H]+	274.11862	9.32
Metobromuron	C9H11BrN2O2	3060-89-7	[M+H]+	259.00767	15.60
metolachlor ESA	C15H23NO5S	171118-09-5	[M+H]+	330.13697	11.24
metolachlor	C15H22ClNO2	51218-45-2	[M+H]+	284.14118	18.92
Metolachlor OA	C15H21NO4	152019-73-3	[M+H]+	280.15433	17.1
Metoprolol	C15H25NO3	37350-58-6	[M+H]+	268.19072	9.46
metoxuron	C10H13ClN2O2	19937-59-8	[M+H]+	229.07383	11.94
metribuzin	C8H14N4OS	21087-64-9	[M+H]+	215.09611	13.19
monuron	C9H11ClN2O	150-68-5	[M+H]+	199.06327	12.66
N-acetyl-4-aminoantipyrine	C13H15N3O2	83-15-8	[M+H]+	246.12370	7.08
N-acetylsulfamethoxazole	C12H13N3O4S	21312-10-7	[M+H]+	296.06995	11.06
neburon	C12H16Cl2N2O	555-37-3	[M+H]+	275.07125	19.30
N-formyl-4-aminoantipyrine	C12H13N3O2	1672-58-8	[M+H]+	232.10805	7.12
nicosulfuron	C15H18N6O6S	111991-09-4	[M+H]+	411.10813	12.24
oxypurinol	C5H4N4O2	219-570-9	[M+H]-	151.02615	2.27
p,p-sulfonyldiphenol	C12H10O4S	98388-00-2	[M+H]-	249.02270	11.21
pentoxifylline	C13H18N4O3	6493-05-6	[M+H]+	279.14517	9.46
Phenazone	C11H12N2O	60-80-0	[M+H]+	189.10224	8.66
pirimicarb	C11H18N4O2	23103-98-2	[M+H]+	239.15025	9.11
sebutylazine	C9H16ClN5	7286-69-3	[M+H]+	230.11670	16.2
simazin	C7H12ClN5	122-34-9	[M+H]+	202.0854	12.50
sitagliptine	C16H15N5OF6	486460-32-6	[M+H]+	408.12536	10.27
sulfadimidine	C12H14N4O2S	57-68-1	[M+H]+	279.09102	8.38
sulfamethoxazole	C10H11N3O3S	723-46-6	[M+H]+	254.05939	10.69
telmisartan	C33H30N4O2	144701-48-4	[M+H]+	515.24415	14.07
terbutylazin	C9H16ClN5	5915-41-3	[M+H]+	230.1167	16.85
tetraglyme	C10H22O5	143-24-8	[M+H]+	223.154	7.78
Tri-(2-chloroisopropyl)phosphate	C9H18Cl3O4P	13674-84-5	[M+H]+	327.00811	17.24
triethylphosphate	C6H15O4P	78-40-0	[M+H]+	183.07807	10.94
tri-n-butyl-phosphate	C12H27O4P	126-73-8	[M+H]+	267.17197	20.52
Triphenylphosphineoxide	C18H15OP	791-28-6	[M+H]+	279.09333	15.34
Tris(2-chloroethyl)phosphate (TCEP)	C6H12Cl3O4P	115-96-8	[M+H]+	284.96116	14.26
valsartan	C24H29N5O3	137862-53-4	[M+H]+	436.23432	16.51
valsartanzuur	C14H10N4O2	164265-78-5	[M+H]+	267.08765	11.78

Table E.2 – Chemicals and concentrations for sample TA344/TA362 (used in MS2-trigger experiment).

<b>sample name: TA344/ TA362</b>			
<b>spiked in: ultrapure water</b>			
<b>concentration: 10 µg/L</b>			
<b>filtrated: no</b>			
Chemical name	CASRN	formula	MW (g/mol)
Ifosfamide	3778-73-2	C <sub>7</sub> H <sub>15</sub> Cl <sub>2</sub> N <sub>2</sub> O <sub>2</sub> P	260.02482

Table E.3 – Chemicals and concentrations for sample TA367 (used in MS2-trigger experiment).

<b>sample name: TA367</b>			
<b>spiked in: ultrapure water</b>			
<b>concentration: 10 µg/L</b>			
<b>filtrated: no</b>			
Chemical name	CASRN	formula	MW (g/mol)
4-[2-(Acryloyloxy)ethoxy]-4-oxobutanoic acid	50940-49-3	C <sub>9</sub> H <sub>12</sub> O <sub>6</sub>	216.06339
Acrylamide	79-06-1	C <sub>3</sub> H <sub>5</sub> NO	71.03711
Diacetone acrylamide	2873-97-4	C <sub>9</sub> H <sub>15</sub> NO <sub>2</sub>	169.11028
Isobornyl acrylate	5888-33-5	C <sub>13</sub> H <sub>20</sub> O <sub>2</sub>	208.14633

Table E.4 – Chemicals and concentrations for sample TA322 (used in MS2-trigger experiment).

<b>sample name: TA322</b>			
<b>spiked in: ultrapure water</b>			
<b>concentration: 10 µg/L</b>			
<b>filtrated: no</b>			
Chemical name	CASRN	formula	MW (g/mol)
Trimethoprim	738-70-5	C <sub>14</sub> H <sub>18</sub> N <sub>4</sub> O <sub>3</sub>	290.13789
Deethylatrazine	6190-65-4	C <sub>6</sub> H <sub>10</sub> ClN <sub>5</sub>	187.06247
Metamitron	41394-05-2	C <sub>10</sub> H <sub>10</sub> N <sub>4</sub> O	202.08546
Sulfamethoxazole	723-46-6	C <sub>10</sub> H <sub>11</sub> N <sub>3</sub> O <sub>3</sub> S	253.05211
Sulfaquinoxaline	59-40-5	C <sub>14</sub> H <sub>12</sub> N <sub>4</sub> O <sub>2</sub> S	300.06810

Table E.5 – Chemicals and concentrations for sample TA387/TA395 (used in MS2-trigger experiment).

<b>sample name: TA387/TA395</b>			
<b>spiked in: ultrapure water</b>			
<b>concentration: 10 µg/L</b>			
<b>filtrated: no</b>			
Chemical name	CASRN	formula	MW (g/mol)
Acetaminophen	103-90-2	C <sub>8</sub> H <sub>9</sub> NO <sub>2</sub>	151.06333
Diatrizoic acid	117-96-4	C <sub>11</sub> H <sub>9</sub> I <sub>3</sub> N <sub>2</sub> O <sub>4</sub>	613.76965
N(4)-Acetylsulfadiazine	127-74-2	C <sub>12</sub> H <sub>12</sub> N <sub>4</sub> O <sub>3</sub> S	292.06301
N-Acetyl sulfamethoxazole	21312-10-7	C <sub>12</sub> H <sub>13</sub> N <sub>3</sub> O <sub>4</sub> S	295.06268
N-Acetylaminoantipyrine	83-15-8	C <sub>13</sub> H <sub>15</sub> N <sub>3</sub> O <sub>2</sub>	245.11643

Table E.6 – Chemicals and concentrations for sample Total MQ (used in MS2-trigger experiment).

<b>sample name: Total MQ</b>			
<b>spiked in: ultrapure water</b>			
<b>concentration: 10 µg/L</b>			
<b>filtrated: yes</b>			
Chemical name	CASRN	formula	MW (g/mol)
4-[2-(Acryloyloxy)ethoxy]-4-oxobutanoic acid	50940-49-3	C <sub>9</sub> H <sub>12</sub> O <sub>6</sub>	216.06339
Acetaminophen	103-90-2	C <sub>8</sub> H <sub>9</sub> NO <sub>2</sub>	151.06333
Acrylamide	79-06-1	C <sub>3</sub> H <sub>5</sub> NO	71.03711
Deethylatrazine	6190-65-4	C <sub>6</sub> H <sub>10</sub> ClN <sub>5</sub>	187.06247
Diacetone acrylamide	2873-97-4	C <sub>9</sub> H <sub>15</sub> NO <sub>2</sub>	169.11028
Diatrizoic acid	117-96-4	C <sub>11</sub> H <sub>9</sub> I <sub>3</sub> N <sub>2</sub> O <sub>4</sub>	613.76965
Ifofamide	3778-73-2	C <sub>7</sub> H <sub>15</sub> Cl <sub>2</sub> N <sub>2</sub> O <sub>2</sub> P	260.02482
Isobornyl acrylate	5888-33-5	C <sub>13</sub> H <sub>20</sub> O <sub>2</sub>	208.14633
Metamitron	41394-05-2	C <sub>10</sub> H <sub>10</sub> N <sub>4</sub> O	202.08546
Trimethoprim	738-70-5	C <sub>14</sub> H <sub>18</sub> N <sub>4</sub> O <sub>3</sub>	290.13789
N(4)-Acetylsulfadiazine	127-74-2	C <sub>12</sub> H <sub>12</sub> N <sub>4</sub> O <sub>3</sub> S	292.06301
N-Acetyl sulfamethoxazole	21312-10-7	C <sub>12</sub> H <sub>13</sub> N <sub>3</sub> O <sub>4</sub> S	295.06268
N-Acetylaminoantipyrine	83-15-8	C <sub>13</sub> H <sub>15</sub> N <sub>3</sub> O <sub>2</sub>	245.11643
Sulfamethoxazole	723-46-6	C <sub>10</sub> H <sub>11</sub> N <sub>3</sub> O <sub>3</sub> S	253.05211
Sulfaquinoxaline	59-40-5	C <sub>14</sub> H <sub>12</sub> N <sub>4</sub> O <sub>2</sub> S	300.06810

Table E.7 – Chemicals and concentrations for samples Total OW (used in MS2-trigger experiment).

sample names: Total OW 10 µg/L, Total OW 1 µg/L, Total OW 100 ng/L, Total OW 10 ng/L, Total OW 1 ng/L spiked in: surface water concentrations: 10 µg/L, 1 µg/L, 100 ng/L, 10 ng/L, 1 ng/L filtrated: yes			
Chemical name	CASRN	formula	MW (g/mol)
4-[2-(Acryloyloxy)ethoxy]-4-oxobutanoic acid	50940-49-3	C <sub>9</sub> H <sub>12</sub> O <sub>6</sub>	216.06339
Acetaminophen	103-90-2	C <sub>8</sub> H <sub>9</sub> NO <sub>2</sub>	151.06333
Acrylamide	79-06-1	C <sub>3</sub> H <sub>5</sub> NO	71.03711
Deethylatrazine	6190-65-4	C <sub>6</sub> H <sub>10</sub> ClN <sub>5</sub>	187.06247
Diacetone acrylamide	2873-97-4	C <sub>9</sub> H <sub>15</sub> NO <sub>2</sub>	169.11028
Diatrizoic acid	117-96-4	C <sub>11</sub> H <sub>9</sub> I <sub>3</sub> N <sub>2</sub> O <sub>4</sub>	613.76965
Ifosfamide	3778-73-2	C <sub>7</sub> H <sub>15</sub> Cl <sub>2</sub> N <sub>2</sub> O <sub>2</sub> P	260.02482
Isobornyl acrylate	5888-33-5	C <sub>13</sub> H <sub>20</sub> O <sub>2</sub>	208.14633
Metamitron	41394-05-2	C <sub>10</sub> H <sub>10</sub> N <sub>4</sub> O	202.08546
Trimethoprim	738-70-5	C <sub>14</sub> H <sub>18</sub> N <sub>4</sub> O <sub>3</sub>	290.13789
N(4)-Acetylsulfadiazine	127-74-2	C <sub>12</sub> H <sub>12</sub> N <sub>4</sub> O <sub>3</sub> S	292.06301
N-Acetyl sulfamethoxazole	21312-10-7	C <sub>12</sub> H <sub>13</sub> N <sub>3</sub> O <sub>4</sub> S	295.06268
N-Acetylaminopyrine	83-15-8	C <sub>13</sub> H <sub>15</sub> N <sub>3</sub> O <sub>2</sub>	245.11643
Sulfamethoxazole	723-46-6	C <sub>10</sub> H <sub>11</sub> N <sub>3</sub> O <sub>3</sub> S	253.05211
Sulfaquinoxaline	59-40-5	C <sub>14</sub> H <sub>12</sub> N <sub>4</sub> O <sub>2</sub> S	300.06810

## Appendix F – Sequence lists

Table F.1 – Sequence list of AcquireX experiments.

File Name	Sample Name	Instrument Method
202002017non_target_pos-01	Blank MQ filtered	NTS dda stepped 20 35 50
202002017non_target_pos-02	Blank MQ filtered	NTS dda stepped 20 35 50
202002017non_target_pos-03	Blank MQ filtered	NTS dda stepped 20 35 50
202002017non_target_pos-04	Blank MQ filtered stepped 1	NTS dda stepped 20 35 50
202002017non_target_pos-05	Blank MQ filtered stepped 2	NTS dda stepped 20 35 50
202002017non_target_pos-06	Blank MQ filtered stepped 3	NTS dda stepped 20 35 50
202002017non_target_pos-07	Blank MQ filtered ACE 1	NTS dda ACE 20 35 50
202002017non_target_pos-08	Blank MQ filtered ACE 2	NTS dda ACE 20 35 50
202002017non_target_pos-09	Blank MQ filtered ACE 3	NTS dda ACE 20 35 50
202002017non_target_pos-10	Blank OW filtered stepped 1	NTS dda stepped 20 35 50
202002017non_target_pos-11	Blank OW filtered stepped 2	NTS dda stepped 20 35 50
202002017non_target_pos-12	Blank OW filtered stepped 3	NTS dda stepped 20 35 50
202002017non_target_pos-13	OW + spike dda stepped 20 35 50 1	NTS dda stepped 20 35 50
202002017non_target_pos-14	OW + spike dda stepped 20 35 50 2	NTS dda stepped 20 35 50
202002017non_target_pos-15	OW + spike dda stepped 20 35 50 3	NTS dda stepped 20 35 50
202002017non_target_pos-16	OW + spike dda ace 20 35 50 1	NTS dda ACE 20 35 50
202002017non_target_pos-17	OW + spike dda ace 20 35 50 2	NTS dda ACE 20 35 50
202002017non_target_pos-18	OW + spike dda ace 20 35 50 3	NTS dda ACE 20 35 50
202002017non_target_pos-19	OW + spike dda ace 5 20 35 50 75 1	NTS dda ACE 5 20 35 50 75
202002017non_target_pos-20	OW + spike dda ace 5 20 35 50 75 2	NTS dda ACE 5 20 35 50 75
202002017non_target_pos-21	OW + spike dda ace 5 20 35 50 75 3	NTS dda ACE 5 20 35 50 75
202002017non_target_pos-22	Blank MQ filtered bgexcl stepped 1	NTS bgexcl stepped 20 35 50
202002017non_target_pos-23	Blank MQ filtered bgexcl stepped 2	NTS bgexcl stepped 20 35 50
202002017non_target_pos-24	Blank MQ filtered bgexcl stepped 3	NTS bgexcl stepped 20 35 50
202002017non_target_pos-25	Blank MQ filtered bgexcl ACE 1	NTS bgexcl ACE 20 35 50
202002017non_target_pos-26	Blank MQ filtered bgexcl ACE 2	NTS bgexcl ACE 20 35 50
202002017non_target_pos-27	Blank MQ filtered bgexcl ACE 3	NTS bgexcl ACE 20 35 50
202002017non_target_pos-28	OW + spike bg excl 30 ms it stepped 20 35 50 1	NTS bgexcl stepped 20 35 50 it 30 ms
202002017non_target_pos-29	OW + spike bg excl 30 ms it stepped 20 35 50 2	NTS bgexcl stepped 20 35 50 it 30 ms
202002017non_target_pos-30	OW + spike bg excl 30 ms it stepped 20 35 50 3	NTS bgexcl stepped 20 35 50 it 30 ms
202002017non_target_pos-31	OW + spike bg excl 30 ms it ace 20 35 50 1	NTS bgexcl ACE 20 35 50 it 30 ms
202002017non_target_pos-32	OW + spike bg excl 30 ms it ace 20 35 50 2	NTS bgexcl ACE 20 35 50 it 30 ms
202002017non_target_pos-33	OW + spike bg excl 30 ms it ace 20 35 50 3	NTS bgexcl ACE 20 35 50 it 30 ms
202002017non_target_pos-34	OW + spike bg excl 30 ms it ace 5 20 35 50 75 1	NTS bgexcl ACE 5 20 35 50 75 it 30 ms
202002017non_target_pos-35	OW + spike bg excl 30 ms it ace 5 20 35 50 75 2	NTS bgexcl ACE 5 20 35 50 75 it 30 ms
202002017non_target_pos-36	OW + spike bg excl 30 ms it ace 5 20 35 50 75 3	NTS bgexcl ACE 5 20 35 50 75 it 30 ms
202002017non_target_pos-37	OW + spike bg excl 50 ms it stepped 20 35 50 1	NTS bgexcl stepped 20 35 50 it 50 ms
202002017non_target_pos-38	OW + spike bg excl 50 ms it stepped 20 35 50 2	NTS bgexcl stepped 20 35 50 it 50 ms
202002017non_target_pos-39	OW + spike bg excl 50 ms it stepped 20 35 50 3	NTS bgexcl stepped 20 35 50 it 50 ms
202002017non_target_pos-40	OW + spike bg excl 50 ms it ace 20 35 50 1	NTS bgexcl ACE 20 35 50 it 50 ms
202002017non_target_pos-41	OW + spike bg excl 50 ms it ace 20 35 50 2	NTS bgexcl ACE 20 35 50 it 50 ms
202002017non_target_pos-42	OW + spike bg excl 50 ms it ace 20 35 50 3	NTS bgexcl ACE 20 35 50 it 50 ms
202002017non_target_pos-43	OW + spike bg excl 50 ms it ace 5 20 35 50 75 1	NTS bgexcl ACE 5 20 35 50 75 it 50 ms
202002017non_target_pos-44	OW + spike bg excl 50 ms it ace 5 20 35 50 75 2	NTS bgexcl ACE 5 20 35 50 75 it 50 ms
202002017non_target_pos-45	OW + spike bg excl 50 ms it ace 5 20 35 50 75 3	NTS bgexcl ACE 5 20 35 50 75 it 50 ms
202002017non_target_pos-46	OW + spike bg excl 100 ms it stepped 20 35 50 1	NTS bgexcl stepped 20 35 50 it 100 ms
202002017non_target_pos-47	OW + spike bg excl 100 ms it stepped 20 35 50 2	NTS bgexcl stepped 20 35 50 it 100 ms
202002017non_target_pos-48	OW + spike bg excl 100 ms it stepped 20 35 50 3	NTS bgexcl stepped 20 35 50 it 100 ms
202002017non_target_pos-49	OW + spike bg excl 100 ms it ace 20 35 50 1	NTS bgexcl ACE 20 35 50 it 100 ms
202002017non_target_pos-50	OW + spike bg excl 100 ms it ace 20 35 50 2	NTS bgexcl ACE 20 35 50 it 100 ms
202002017non_target_pos-51	OW + spike bg excl 100 ms it ace 20 35 50 3	NTS bgexcl ACE 20 35 50 it 100 ms
202002017non_target_pos-52	OW + spike bg excl 100 ms it ace 5 20 35 50 75 1	NTS bgexcl ACE 5 20 35 50 75 it 100 ms
202002017non_target_pos-53	OW + spike bg excl 100 ms it ace 5 20 35 50 75 2	NTS bgexcl ACE 5 20 35 50 75 it 100 ms
202002017non_target_pos-54	OW + spike bg excl 100 ms it ace 5 20 35 50 75 3	NTS bgexcl ACE 5 20 35 50 75 it 100 ms







Table F.4 – Sequence list of MS1-trigger experiments with surface water samples.

File Name	Sample Name	Instrument Method
20200414non_target_pos-01	Blank MQ	NTS_pos_method 1_KWR
20200414non_target_pos-02	Blank MQ	NTS_pos_method 1_KWR
20200414non_target_pos-03	OW + spike	NTS_pos_method 1_KWR
20200414non_target_pos-04	OW + spike	NTS_pos_method 1_KWR
20200414non_target_pos-05	Method 1 Blank MQ	NTS_pos_method 1_KWR
20200414non_target_pos-06	Method 1 OW + spike 1	NTS_pos_method 1_KWR
20200414non_target_pos-07	Method 1 OW + spike 2	NTS_pos_method 1_KWR
20200414non_target_pos-08	Method 1 OW + spike 3	NTS_pos_method 1_KWR
20200414non_target_pos-09	Method 2 OW + spike 1	NTS_pos_method 2_isotopic ratio
20200414non_target_pos-10	Method 2 OW + spike 2	NTS_pos_method 2_isotopic ratio
20200414non_target_pos-11	Method 2 OW + spike 3	NTS_pos_method 2_isotopic ratio
20200414non_target_pos-12	Method 3 OW + spike 1	NTS_pos_method 3_susdat
20200414non_target_pos-13	Method 3 OW + spike 2	NTS_pos_method 3_susdat
20200414non_target_pos-14	Method 3 OW + spike 3	NTS_pos_method 3_susdat
20200414non_target_pos-15	Method 1 Blank MQ	NTS_pos_method 1_KWR
20200414non_target_pos-16	Method 4 OW + spike 1	NTS_pos_method 4_UBAPMT
20200414non_target_pos-17	Method 4 OW + spike 2	NTS_pos_method 4_UBAPMT
20200414non_target_pos-18	Method 4 OW + spike 3	NTS_pos_method 4_UBAPMT
20200414non_target_pos-19	Method 5 OW + spike 1	NTS_pos_method 5_Sjerps
20200414non_target_pos-20	Method 5 OW + spike 2	NTS_pos_method 5_Sjerps
20200414non_target_pos-21	Method 5 OW + spike 3	NTS_pos_method 5_Sjerps
20200414non_target_pos-22	Method 6 OW + spike 1	NTS_pos_method 6_susdat_tR
20200414non_target_pos-23	Method 6 OW + spike 2	NTS_pos_method 6_susdat_tR
20200414non_target_pos-24	Method 6 OW + spike 3	NTS_pos_method 6_susdat_tR
20200414non_target_pos-25	Method 1 Blank MQ	NTS_pos_method 1_KWR
20200414non_target_pos-26	Method 7 OW + spike 1	NTS_pos_method 7_isotopic_ratio_UBAPMT
20200414non_target_pos-27	Method 7 OW + spike 2	NTS_pos_method 7_isotopic_ratio_UBAPMT
20200414non_target_pos-28	Method 7 OW + spike 3	NTS_pos_method 7_isotopic_ratio_UBAPMT
20200414non_target_pos-29	Method 8 OW + spike 1	NTS_pos_method 8_isotopic_ratio_susdat
20200414non_target_pos-30	Method 8 OW + spike 2	NTS_pos_method 8_isotopic_ratio_susdat
20200414non_target_pos-31	Method 8 OW + spike 3	NTS_pos_method 8_isotopic_ratio_susdat
20200414non_target_pos-32	Method 9 OW + spike 1	NTS_pos_method 9_isotopic_ratio_susdat_tR
20200414non_target_pos-33	Method 9 OW + spike 2	NTS_pos_method 9_isotopic_ratio_susdat_tR
20200414non_target_pos-34	Method 9 OW + spike 2	NTS_pos_method 9_isotopic_ratio_susdat_tR
20200414non_target_pos-35	Method 1 Blank MQ	NTS_pos_method 1_KWR
20200414non_target_pos-36	Method 10 OW + spike 1	NTS_pos_method 10_isotopic_ratio_Sjerps
20200414non_target_pos-37	Method 10 OW + spike 2	NTS_pos_method 10_isotopic_ratio_Sjerps
20200414non_target_pos-38	Method 10 OW + spike 3	NTS_pos_method 10_isotopic_ratio_Sjerps
20200414non_target_pos-39	Method 11 OW + spike 1	NTS_pos_method 11_Spike inclusion
20200414non_target_pos-40	Method 11 OW + spike 2	NTS_pos_method 11_Spike inclusion
20200414non_target_pos-41	Method 11 OW + spike 3	NTS_pos_method 11_Spike inclusion
20200414non_target_pos-42	Method 12 OW + spike 1	NTS_pos_method 12_isotopic_ratio_Spike inclusion
20200414non_target_pos-43	Method 12 OW + spike 2	NTS_pos_method 12_isotopic_ratio_Spike inclusion
20200414non_target_pos-44	Method 12 OW + spike 3	NTS_pos_method 12_isotopic_ratio_Spike inclusion
20200414non_target_pos-45	Method 1 Blank MQ	NTS_pos_method 1_KWR

Table F.5 – Sequence list of MS1-trigger experiments with wastewater treatment plant influent samples.

File Name	Sample Name	Instrument Method
20200501_MS1trigger_pos-01	Blank MQ	NTS_pos_method 1_KWR
20200501_MS1trigger_pos-02	Blank MQ	NTS_pos_method 1_KWR
20200501_MS1trigger_pos-03	Method 1 RWZI inf + spike 1	NTS_pos_method 1_KWR
20200501_MS1trigger_pos-04	Method 1 RWZI inf + spike 2	NTS_pos_method 1_KWR
20200501_MS1trigger_pos-05	Method 1 RWZI inf + spike 3	NTS_pos_method 1_KWR
20200501_MS1trigger_pos-06	Method 2 RWZI inf + spike 1	NTS_pos_method 2_isotopic ratio
20200501_MS1trigger_pos-07	Method 2 RWZI inf + spike 2	NTS_pos_method 2_isotopic ratio
20200501_MS1trigger_pos-08	Method 2 RWZI inf + spike 3	NTS_pos_method 2_isotopic ratio
20200501_MS1trigger_pos-09	Method 3 RWZI inf + spike 1	NTS_pos_method 3_susdat
20200501_MS1trigger_pos-10	Method 3 RWZI inf + spike 2	NTS_pos_method 3_susdat
20200501_MS1trigger_pos-11	Method 3 RWZI inf + spike 3	NTS_pos_method 3_susdat
20200501_MS1trigger_pos-12	Blank MQ	NTS_pos_method 1_KWR
20200501_MS1trigger_pos-13	Method 4 RWZI inf + spike 1	NTS_pos_method 4_UBAPMT
20200501_MS1trigger_pos-14	Method 4 RWZI inf + spike 2	NTS_pos_method 4_UBAPMT
20200501_MS1trigger_pos-15	Method 4 RWZI inf + spike 3	NTS_pos_method 4_UBAPMT
20200501_MS1trigger_pos-16	Method 5 RWZI inf + spike 1	NTS_pos_method 5_Sjerps
20200501_MS1trigger_pos-17	Method 5 RWZI inf + spike 2	NTS_pos_method 5_Sjerps
20200501_MS1trigger_pos-18	Method 5 RWZI inf + spike 3	NTS_pos_method 5_Sjerps
20200501_MS1trigger_pos-19	Method 6 RWZI inf + spike 1	NTS_pos_method 6_susdat_tR
20200501_MS1trigger_pos-20	Method 6 RWZI inf + spike 2	NTS_pos_method 6_susdat_tR
20200501_MS1trigger_pos-21	Method 6 RWZI inf + spike 3	NTS_pos_method 6_susdat_tR
20200501_MS1trigger_pos-22	Blank MQ	NTS_pos_method 1_KWR
20200501_MS1trigger_pos-23	Method 7 RWZI inf + spike 1	NTS_pos_method 7_isotopic_ratio_UBAPMT
20200501_MS1trigger_pos-24	Method 7 RWZI inf + spike 2	NTS_pos_method 7_isotopic_ratio_UBAPMT
20200501_MS1trigger_pos-25	Method 7 RWZI inf + spike 3	NTS_pos_method 7_isotopic_ratio_UBAPMT
20200501_MS1trigger_pos-26	Method 8 RWZI inf + spike 1	NTS_pos_method 8_isotopic_ratio_susdat
20200501_MS1trigger_pos-27	Method 8 RWZI inf + spike 2	NTS_pos_method 8_isotopic_ratio_susdat
20200501_MS1trigger_pos-28	Method 8 RWZI inf + spike 3	NTS_pos_method 8_isotopic_ratio_susdat
20200501_MS1trigger_pos-29	Method 9 RWZI inf + spike 1	NTS_pos_method 9_isotopic_ratio_susdat_tR
20200501_MS1trigger_pos-30	Method 9 RWZI inf + spike 2	NTS_pos_method 9_isotopic_ratio_susdat_tR
20200501_MS1trigger_pos-31	Method 9 RWZI inf + spike 3	NTS_pos_method 9_isotopic_ratio_susdat_tR
20200501_MS1trigger_pos-32	Blank MQ	NTS_pos_method 1_KWR
20200501_MS1trigger_pos-33	Method 10 RWZI inf + spike 1	NTS_pos_method 10_isotopic_ratio_Sjerps
20200501_MS1trigger_pos-34	Method 10 RWZI inf + spike 2	NTS_pos_method 10_isotopic_ratio_Sjerps
20200501_MS1trigger_pos-35	Method 10 RWZI inf + spike 3	NTS_pos_method 10_isotopic_ratio_Sjerps
20200501_MS1trigger_pos-36	Method 11 RWZI inf + spike 1	NTS_pos_method 11_Spike inclusion
20200501_MS1trigger_pos-37	Method 11 RWZI inf + spike 2	NTS_pos_method 11_Spike inclusion
20200501_MS1trigger_pos-38	Method 11 RWZI inf + spike 3	NTS_pos_method 11_Spike inclusion
20200501_MS1trigger_pos-39	Method 12 RWZI inf + spike 1	NTS_pos_method 12_isotopic_ratio_Spike inclusion
20200501_MS1trigger_pos-40	Method 12 RWZI inf + spike 2	NTS_pos_method 12_isotopic_ratio_Spike inclusion
20200501_MS1trigger_pos-41	Method 12 RWZI inf + spike 3	NTS_pos_method 12_isotopic_ratio_Spike inclusion
20200501_MS1trigger_pos-42	Method 1 Blank MQ	NTS_pos_method 1_KWR

Table F.6 – Sequence list of MS2-trigger experiments.

File Name	Sample Name	Instrument Method
20200520_MS2trigger_pos-01	Blank MQ water	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-02	Blank MQ water	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-03	Method 1 TA344/TA362	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-04	Method 1 TA344/TA362	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-05	Method 1 TA344/TA362	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-06	Method 2 TA344/TA362	NTS_pos_method 2_TA344-362
20200520_MS2trigger_pos-07	Method 2 TA344/TA362	NTS_pos_method 2_TA344-362
20200520_MS2trigger_pos-08	Method 2 TA344/TA362	NTS_pos_method 2_TA344-362
20200520_MS2trigger_pos-09	Blank MQ water	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-10	Method 1 TA367	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-11	Method 1 TA367	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-12	Method 1 TA367	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-13	Method 3 TA367	NTS_pos_method 3_TA367
20200520_MS2trigger_pos-14	Method 3 TA367	NTS_pos_method 3_TA367
20200520_MS2trigger_pos-15	Method 3 TA367	NTS_pos_method 3_TA367
20200520_MS2trigger_pos-16	Blank MQ water	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-17	Method 1 TA322	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-18	Method 1 TA322	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-19	Method 1 TA322	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-20	Method 4 TA322	NTS_pos_method 4_TA322
20200520_MS2trigger_pos-21	Method 4 TA322	NTS_pos_method 4_TA322
20200520_MS2trigger_pos-22	Method 4 TA322	NTS_pos_method 4_TA322
20200520_MS2trigger_pos-23	Blank MQ water	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-24	Method 1 TA387/TA395	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-25	Method 1 TA387/TA395	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-26	Method 1 TA387/TA395	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-27	Method 5 TA387/TA395	NTS_pos_method 5_TA387-395
20200520_MS2trigger_pos-28	Method 5 TA387/TA395	NTS_pos_method 5_TA387-395
20200520_MS2trigger_pos-29	Method 5 TA387/TA395	NTS_pos_method 5_TA387-395
20200520_MS2trigger_pos-30	Blank MQ water, filtrated	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-31	Blank MQ water, filtrated	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-32	Method 1 Total MQ, filtrated	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-33	Method 1 Total MQ, filtrated	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-34	Method 1 Total MQ, filtrated	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-35	Method 6 Total MQ, filtrated	NTS_pos_method 6_all_alerts
20200520_MS2trigger_pos-36	Method 6 Total MQ, filtrated	NTS_pos_method 6_all_alerts
20200520_MS2trigger_pos-37	Method 6 Total MQ, filtrated	NTS_pos_method 6_all_alerts
20200520_MS2trigger_pos-38	Method 7 Total MQ, filtrated	NTS_pos_method 7_all_alerts_ACE
20200520_MS2trigger_pos-39	Method 7 Total MQ, filtrated	NTS_pos_method 7_all_alerts_ACE
20200520_MS2trigger_pos-40	Method 7 Total MQ, filtrated	NTS_pos_method 7_all_alerts_ACE
20200520_MS2trigger_pos-41	Method 8 Total MQ, filtrated	NTS_pos_method 8_all_alerts_IT
20200520_MS2trigger_pos-42	Method 8 Total MQ, filtrated	NTS_pos_method 8_all_alerts_IT
20200520_MS2trigger_pos-43	Method 8 Total MQ, filtrated	NTS_pos_method 8_all_alerts_IT
20200520_MS2trigger_pos-44	Method 9 Total MQ, filtrated	NTS_pos_method 9_all_alerts_MS1trig
20200520_MS2trigger_pos-45	Method 9 Total MQ, filtrated	NTS_pos_method 9_all_alerts_MS1trig
20200520_MS2trigger_pos-46	Method 9 Total MQ, filtrated	NTS_pos_method 9_all_alerts_MS1trig
20200520_MS2trigger_pos-47	Method 10 Total MQ, filtrated	NTS_pos_method 10_all_alerts_MS1trig_ACE
20200520_MS2trigger_pos-48	Method 10 Total MQ, filtrated	NTS_pos_method 10_all_alerts_MS1trig_ACE
20200520_MS2trigger_pos-49	Method 10 Total MQ, filtrated	NTS_pos_method 10_all_alerts_MS1trig_ACE
20200520_MS2trigger_pos-50	Method 11 Total MQ, filtrated	NTS_pos_method 11_all_alerts_MS1trig_IT
20200520_MS2trigger_pos-51	Method 11 Total MQ, filtrated	NTS_pos_method 11_all_alerts_MS1trig_IT
20200520_MS2trigger_pos-52	Method 11 Total MQ, filtrated	NTS_pos_method 11_all_alerts_MS1trig_IT
20200520_MS2trigger_pos-53	Blank MQ water, filtrated	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-54	Blank MQ water, filtrated	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-55	Method 1 OW, filtrated (no spike)	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-56	Method 1 OW, filtrated (no spike)	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-57	Method 1 OW, filtrated (no spike)	NTS_pos_method 1_KWR
20200520_MS2trigger_pos-58	Method 6 OW, filtrated (no spike)	NTS_pos_method 6_all_alerts



20200520_MS2trigger_pos-119	Method 7 LOA600 OW, filtrated	NTS_pos_method 7_all_alerts_ACE
20200520_MS2trigger_pos-120	Method 7 LOA600 OW, filtrated	NTS_pos_method 7_all_alerts_ACE
20200520_MS2trigger_pos-121	Method 7 LOA600 OW, filtrated	NTS_pos_method 7_all_alerts_ACE
20200520_MS2trigger_pos-122	Method 8 LOA600 OW, filtrated	NTS_pos_method 8_all_alerts_IT
20200520_MS2trigger_pos-123	Method 8 LOA600 OW, filtrated	NTS_pos_method 8_all_alerts_IT
20200520_MS2trigger_pos-124	Method 8 LOA600 OW, filtrated	NTS_pos_method 8_all_alerts_IT
20200520_MS2trigger_pos-125	Method 9 LOA600 OW, filtrated	NTS_pos_method 9_all_alerts_MS1trig
20200520_MS2trigger_pos-126	Method 9 LOA600 OW, filtrated	NTS_pos_method 9_all_alerts_MS1trig
20200520_MS2trigger_pos-127	Method 9 LOA600 OW, filtrated	NTS_pos_method 9_all_alerts_MS1trig
20200520_MS2trigger_pos-128	Method 10 LOA600 OW, filtrated	NTS_pos_method 10_all_alerts_MS1trig_ACE
20200520_MS2trigger_pos-129	Method 10 LOA600 OW, filtrated	NTS_pos_method 10_all_alerts_MS1trig_ACE
20200520_MS2trigger_pos-130	Method 10 LOA600 OW, filtrated	NTS_pos_method 10_all_alerts_MS1trig_ACE
20200520_MS2trigger_pos-131	Method 11 LOA600 OW, filtrated	NTS_pos_method 11_all_alerts_MS1trig_IT
20200520_MS2trigger_pos-132	Method 11 LOA600 OW, filtrated	NTS_pos_method 11_all_alerts_MS1trig_IT
20200520_MS2trigger_pos-133	Method 11 LOA600 OW, filtrated	NTS_pos_method 11_all_alerts_MS1trig_IT
20200520_MS2trigger_pos-134	Blank MQ water, filtrated	NTS_pos_method 1_KWR

## Appendix G – Chlorine and bromine distribution

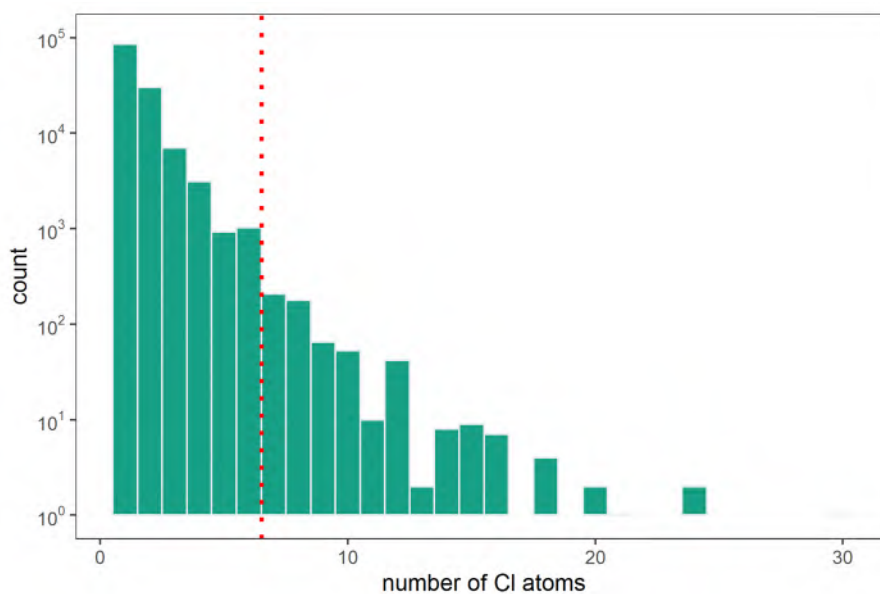


Figure G.1 – Distribution of the number of chlorine atoms in all chlorinated compounds ( $n = 1286500$ ) in the CompTox Chemistry dashboard. Note the logarithmic scale of the y-axis. The vertical red dotted line marks the 99% quantile (1 Cl atom up to 6 Br atoms).

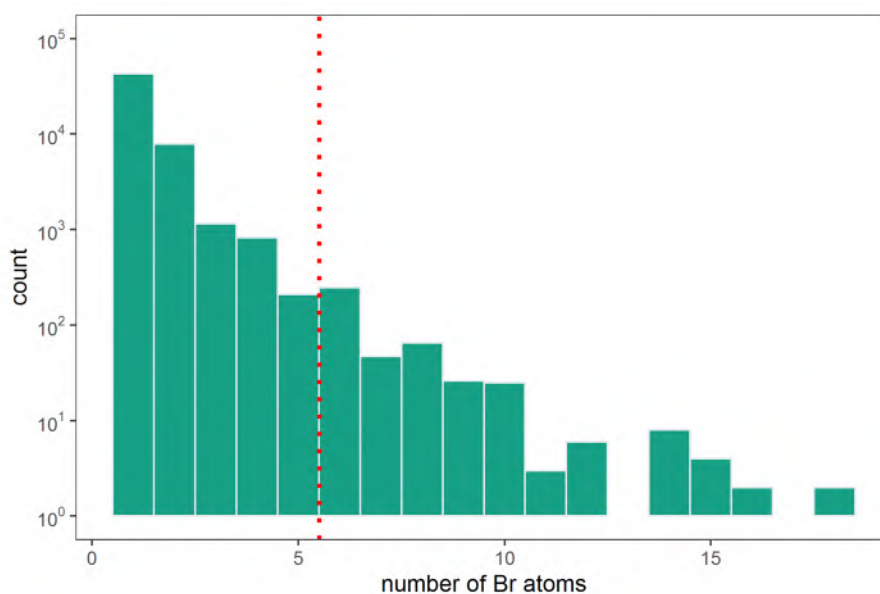


Figure G.2 – Distribution of the number of bromine atoms in all brominated compounds ( $n = 53258$ ) in the CompTox Chemistry dashboard. Note the logarithmic scale of the y-axis. The vertical red dotted line marks the 99% quantile (1 Br atom up to 5 Br atoms).

## Appendix H – Compound Discoverer workflow parameters

LOA-600 NTS workflow cd 3.1 RP Pos adjusted 20200309

### Select Spectra

- General Settings

Precursor Selection:	Use MS(n-1) Precursor
Use Isotope Pattern in Precursor Reevaluation:	True
Provide Profile Spectra:	Automatic
Store Chromatograms:	False
- Spectrum Properties Filter

Lower RT Limit:	2
Upper RT Limit:	27
First Scan:	0
Last Scan:	0
Ignore Specified Scans:	-
Lowest Charge State:	0
Highest Charge State:	0
Min. Precursor Mass:	80 Da
Max. Precursor Mass:	5000 Da
Total Intensity Threshold:	0
Minimum Peak Count:	1
- Scan Event Filters

Mass Analyzer:	(Not specified)
MS Order:	Any
Activation Type:	(Not specified)
Min. Collision Energy:	0
Max. Collision Energy:	1000
Scan Type:	Any
Polarity Mode:	(Not specified)
- Peak Filters

S/N Threshold (FT-only):	1.5
--------------------------	-----
- Replacements for Unrecognized Properties

Unrecognized Charge Replacements:	1
Unrecognized Mass Analyzer Replacements:	ITMS
Unrecognized MS Order Replacements:	MS2
Unrecognized Activation Type Replacements:	CID
Unrecognized Polarity Replacements:	+
Unrecognized MS Resolution@200 Replacements:	60000
Unrecognized MSn Resolution@200 Replacements:	30000

### Align Retention Times

- General Settings

Alignment Fallback:	Use Linear Model
Mass Tolerance:	5 ppm
Maximum Shift [min]:	1
Remove Outlier:	True
Shift Reference File:	True
Alignment Model:	Adaptive curve

### Detect Compounds

- General Settings

Ions:	2M+H], [M+2H], [M+ACN+H], [M+H], [M+H+MeOH], [M+H-H2O], [M+K], [M+Na], [M+NH4]
Base Ions:	[M+H], [M-H]
Intensity Tolerance [%]:	30
Mass Tolerance [ppm]:	3 ppm
Max. Element Counts:	C90 H190 Br3 Cl4 F6 K2 N10 Na2 O18 P3 S5
Min. Element Counts:	C H
Min. Peak Intensity	50000
S/N Threshold:	3
- Peak Detection

Filter Peaks:	True
Min. # Isotopes:	1

Min. # Scans per Peak: 5  
Max. Peak Width [min]: 0.8  
Remove Singlets: False

#### Group Compounds

1. Compound Consolidation  
Mass Tolerance: 3 ppm  
RT Tolerance [min]: 0.1
2. Fragment Data Selection  
Preferred Ions: [M+H], [M-H]

#### Merge Features

1. Peak Consolidation  
Mass Tolerance: 3 ppm  
RT Tolerance [min]: 0.1

#### Predict Compositions

1. Prediction Settings  
Mass Tolerance: 3 ppm  
Max. Element Counts: C90 H190 Br3 Cl4 F6 K2 N10 Na2 O18 P3 S5  
Max. H/C: 3.5  
Max. # Candidates: 10  
Max. # Internal Candidates: 500  
Max. RDBE: 40  
Min. Element Counts: C H  
Min. H/C: 0.1  
Min. RDBE: 0
2. Pattern Matching  
Intensity Threshold [%]: 0.1  
Intensity Tolerance [%]: 30  
Min. Pattern Cov. [%]: 80  
Min. Spectral Fit [%]: 30  
S/N Threshold: 3  
Use Dynamic Recalibration: True
3. Fragments Matching  
Mass Tolerance: 5 ppm  
S/N Threshold: 3  
Use Fragments Matching: True

#### Pattern Scoring

1. General Settings  
Intensity Tolerance [%]: 30  
Isotope Patterns: S, Cl, Br  
Mass Tolerance: 3 ppm  
Min. Spectral Fit [%]: 0  
SN Threshold: 3

#### Search Mass Lists

1. Search Settings  
Mass Lists: KWRWater\_1\_8214.massList, KWRWater\_8215\_18363.massList,  
KWRWater\_18364\_26808.massList, KWRWater\_26809\_35517.massList,  
KWRWater\_35518\_end.massList, LOA-600 suspects structures.masslist  
Mass Tolerance: 3 ppm  
RT Tolerance [min]: 0.5  
Use Retention Time: False

#### Search ChemSpider

1. Search Settings  
Result Order (for. Max # of results per compound): Order By Reference Count (DESC)  
Database(s): ACToR: Aggregated Computational Toxicology Resource, EAWAG  
Biocatalysis/Biodegradation Database; EPA DSSTox; EPA Toxcast; FDA UNII  
– NLM  
Mass Tolerance: 3 ppm



Max. # of Predicted Compositions to be searched per Compound:	3
Max. # of results per compound:	20
Search Mode:	By Formula or Mass
2. <u>Predicted Composition Annotation</u>	
Check All Predicted Compositions:	True

### Assign Compound Annotations

1. <u>General Settings</u>	
Mass Tolerance:	3 ppm
2. <u>Data Sources</u>	
Data Source #1:	mzCloud Search
Data Source #2:	mzVault Search
Data Source #3:	MassList Search
Data Source #4:	ChemSpider Search
Data Source #5:	Predicted Compositions
3. <u>Scoring Rules</u>	
SFit Range:	20
SFit Threshold:	20
Use mzLogic:	True
Use Spectral Distance:	True

### Search mzVault

1. <u>Search Settings</u>	
Apply Intensity Threshold:	True
Compound Classes:	All
FT Fragment Mass Tolerance:	10 ppm
mzVault Library:	Massbank – Fiehn HILIC.db; Massbank all.db
IT Fragment Mass Tolerance:	0.4 Da
Ion Activation Energy Tolerance:	20
Match Analyzer Type:	False
Match Ion Activation Energy:	Any
Match Ion Activation Type:	False
Match Ionization Method:	False
Match Factor Threshold:	50
Max. # Results:	10
Precursor Mass Tolerance:	10 ppm
Remove Precursor Ion:	True
RT Tolerance [min]:	2
Search Algorithm:	HighChem DP
Use Retention Time:	False

### Fill Gaps

1. <u>General Settings</u>	
Mass Tolerance:	3 ppm
S/N Threshold:	1.5
Use Real Peak Detection:	True

### Mark Background Compounds

1. <u>General Settings</u>	
Hide Background:	False
Max. Blank/Sample:	0
Max. Sample/Blank:	10

### Search mzCloud

1. <u>General Settings</u>	
Compound Classes:	All
FT Fragment Mass Tolerance:	10 ppm
IT Fragment Mass Tolerance:	0.4 Da
Library:	Autoprocessed; Reference
Max. # Results:	10
Post Processing:	Recalibrated
Precursor Mass Tolerance:	10 ppm
Annotate Matching Fragments:	False

2. DDA Search

Activation Energy Tolerance:	20
Apply Intensity Threshold:	True
Match Activation Energy:	Match with Tolerance
Match Activation Type:	True
Match Factor Threshold:	60
Identity Search:	Cosine
Similarity Search:	Confidence Forward
3. DIA Search

Activation Energy Tolerance:	100
Apply Intensity Threshold:	False
Match Activation Energy:	Any
Match Activation Type:	False
Match Factor Threshold:	20
Max. Isolation Width [Da]:	500
Use DIA Scans for Search:	False

#### Apply mzLogic

1. Search Settings

FT Fragment Mass Tolerance:	10 ppm
IT Fragment Mass Tolerance:	0.4 Da
Match Factor Threshold:	30
Max. # Compounds:	0
Max. # mzCloud Similarity Results to consider per Compound:	10

#### Apply Spectral Distance

1. Pattern Matching

Intensity Threshold [%]:	0.1
Intensity Tolerance [%]:	30
Mass Tolerance:	5 ppm
S/N Threshold:	3
Use Dynamic Recalibration:	True

## Appendix I – Spectral quality parameters acquisition experiments

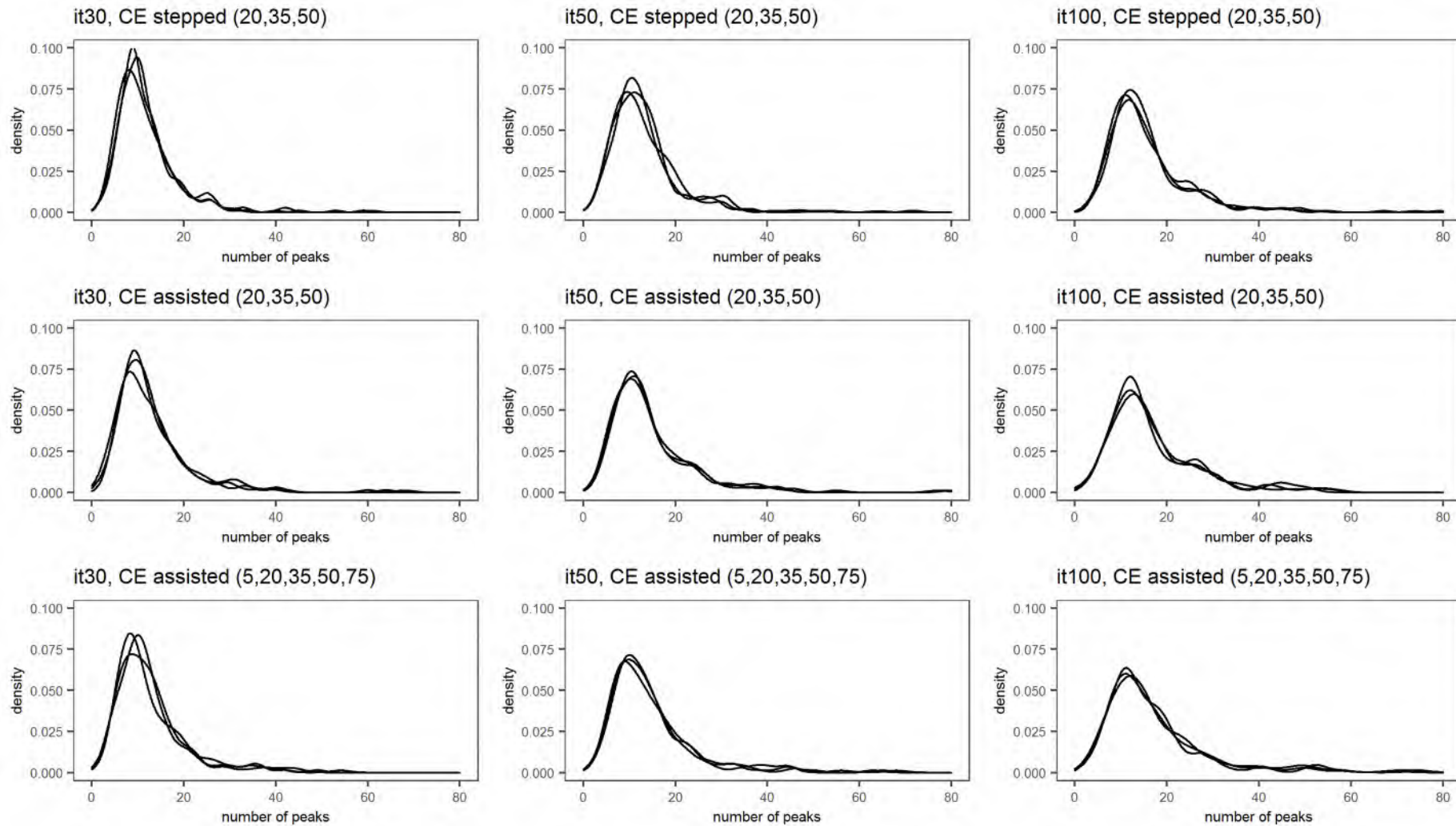


Figure I.1 – Kernel density plots of the number of peaks, square root-transformed. Data from the AcquireX experiments.

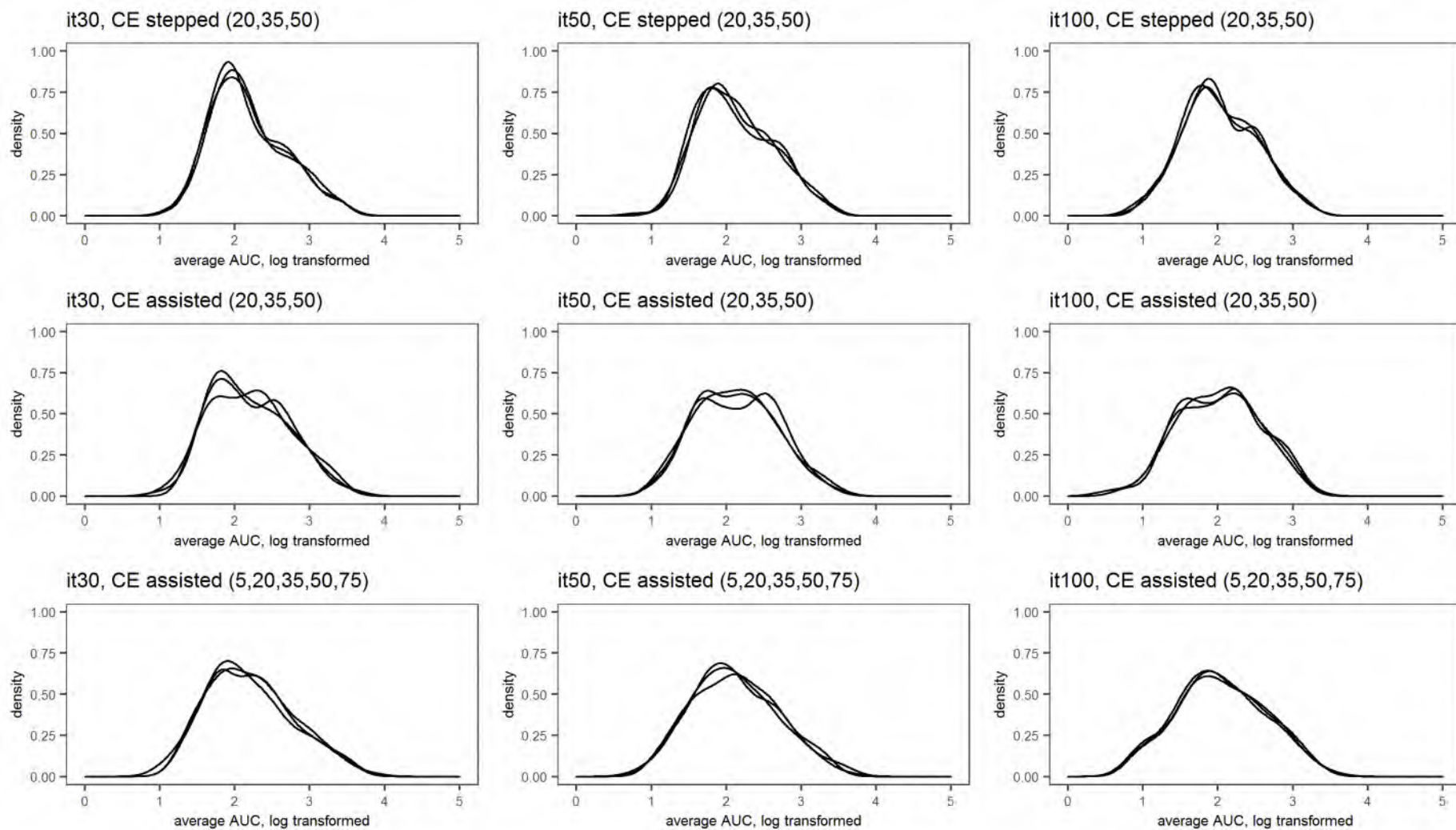


Figure 1.2 – Kernel density plots of the arithmetic mean of the peak areas, log-transformed. Data from the AcquireX experiments.

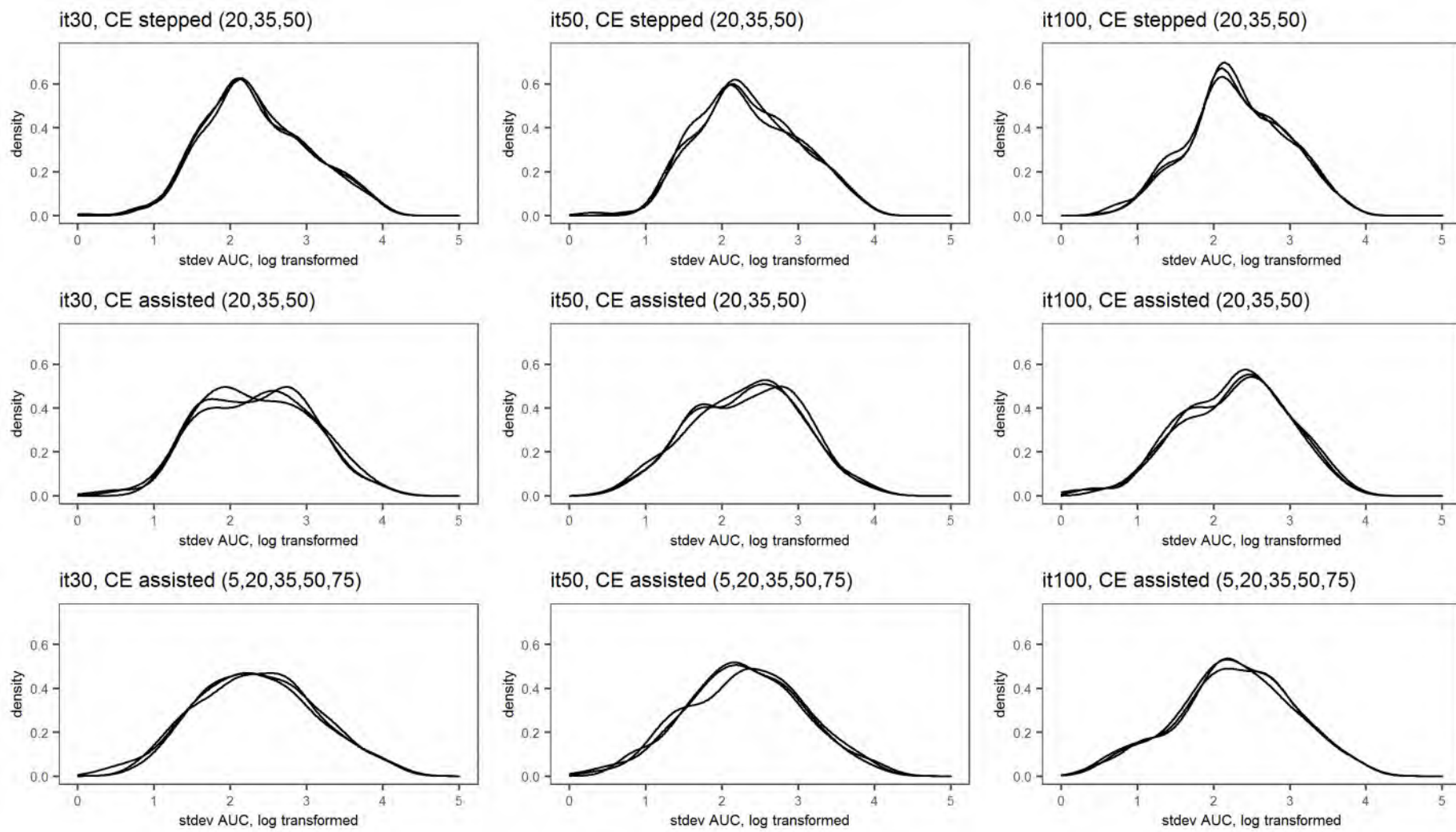


Figure I.3 – Kernel density plots of the standard deviation of the peak areas, log-transformed. Data from the AcquireX experiments.

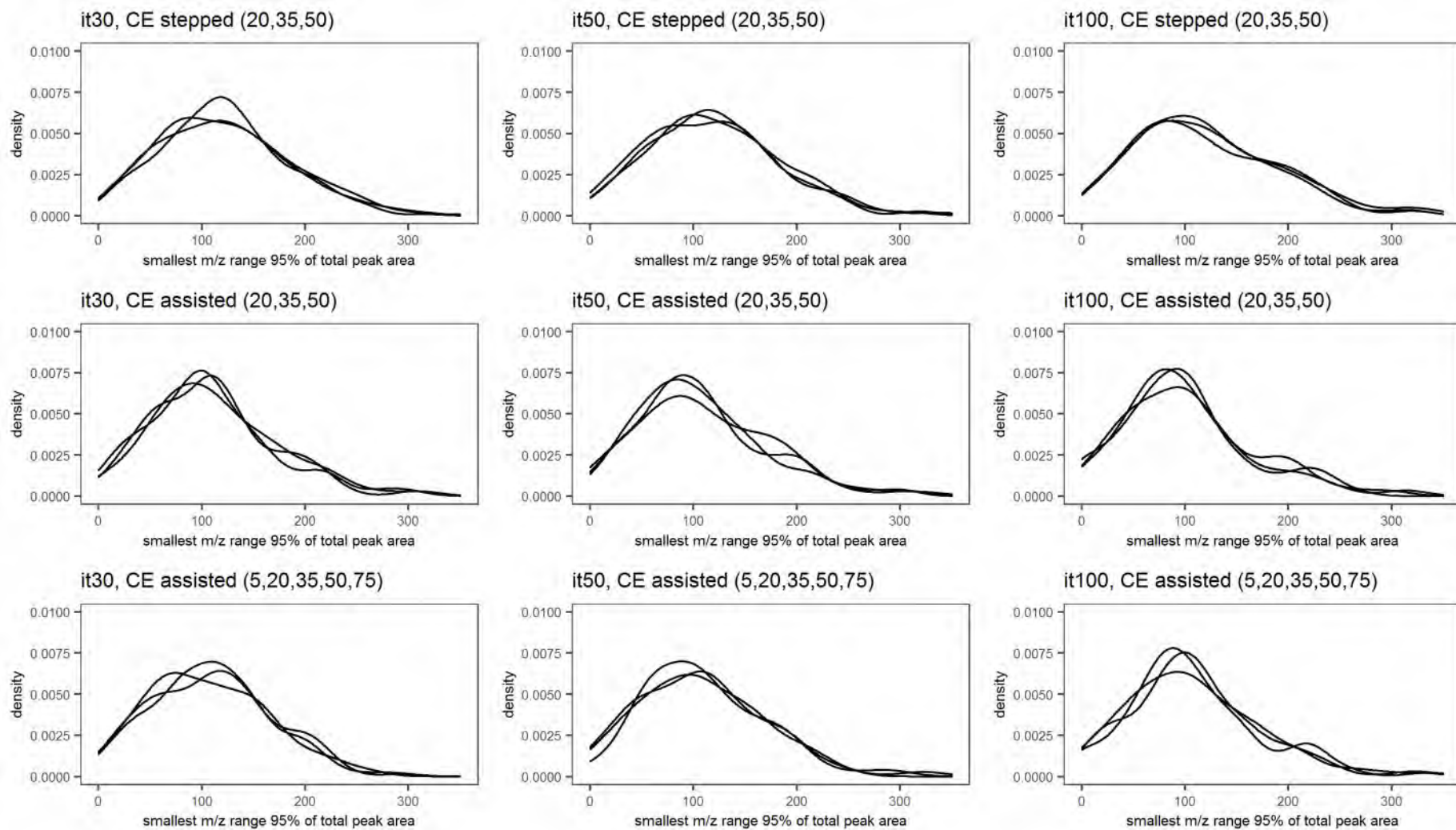


Figure 1.4 – Kernel density plots of the smallest m/z range containing 95% of the total peak area. Data from the AcquireX experiments.

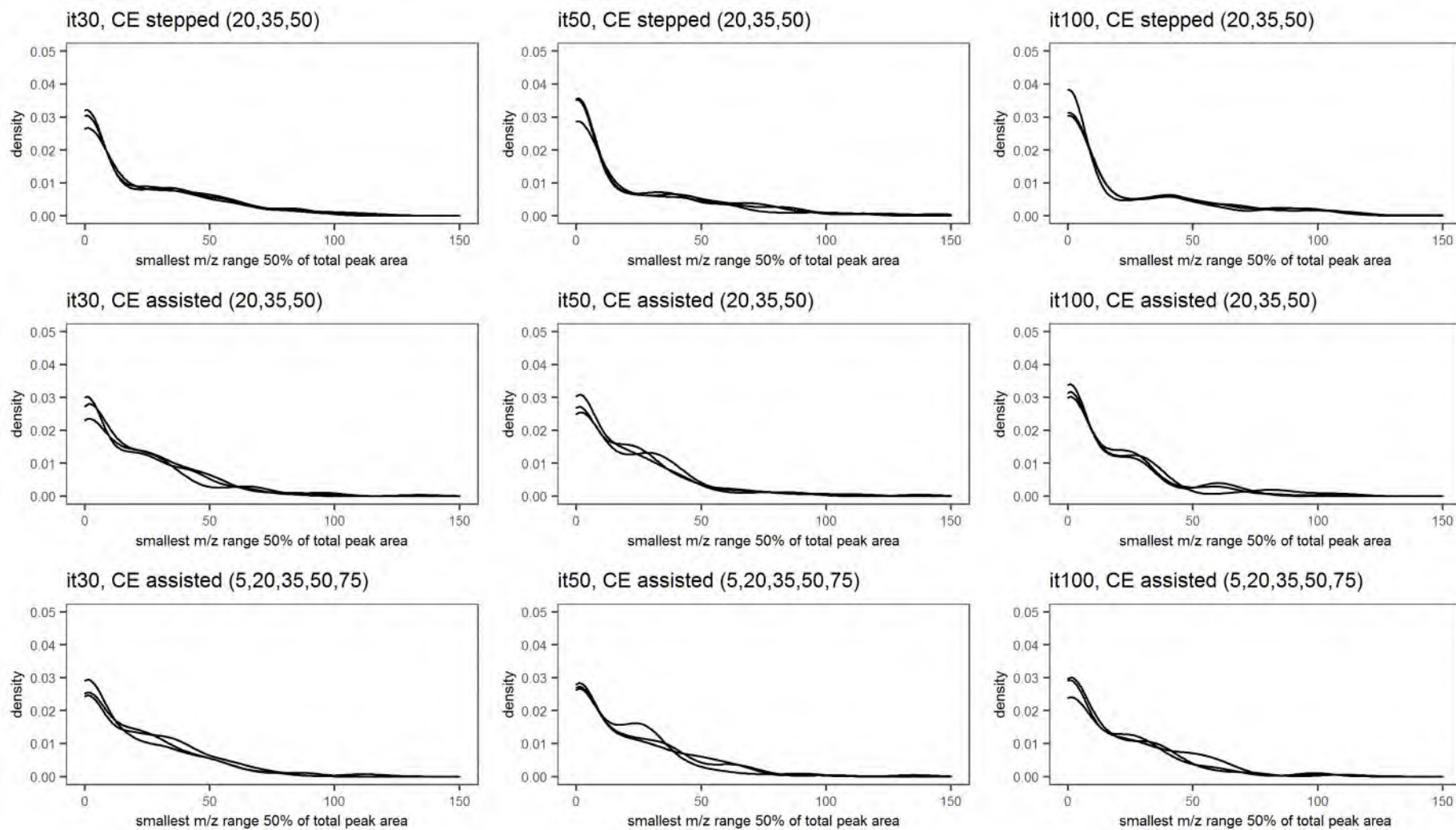


Figure 1.5 – Kernel density plots of the smallest m/z range containing 50% of the total peak area. Data from the AcquireX experiments.

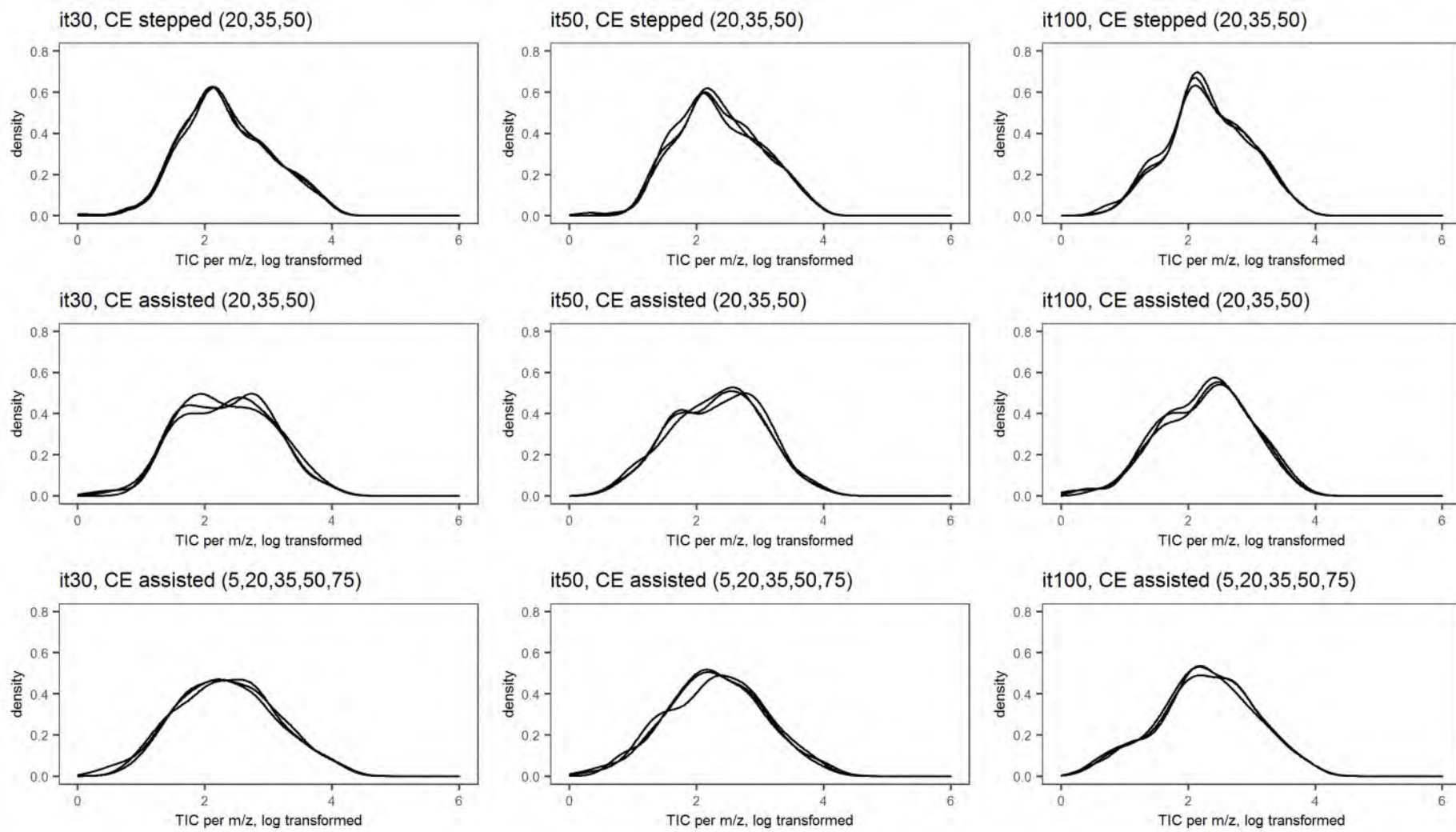


Figure 1.6 – Kernel density plots of the total ion current per m/z, log transformed. Data from the AcquireX experiments.



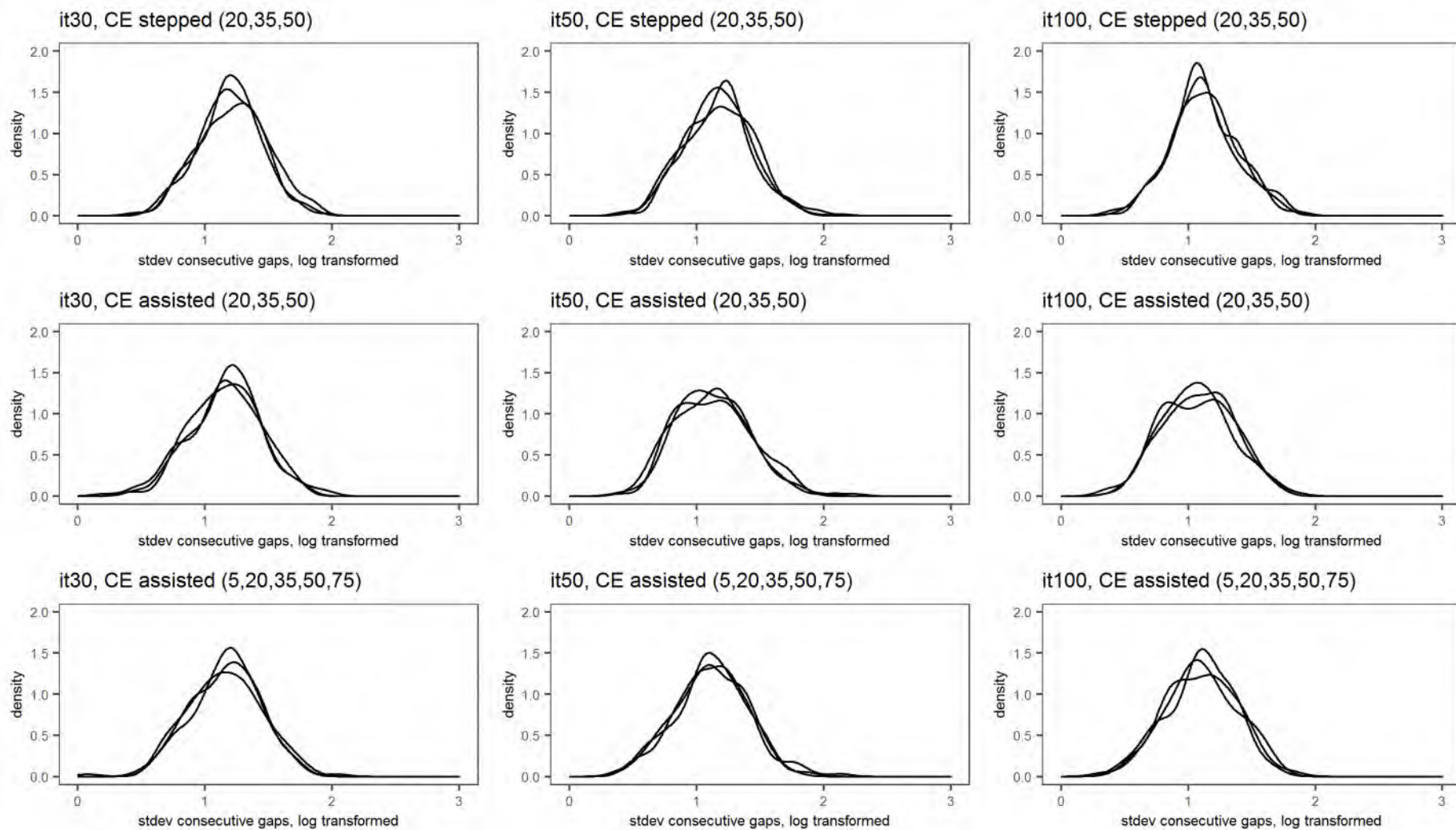


Figure I.7 – Kernel density plots of the standard deviation of the consecutive  $m/z$  gaps between all peaks, log-transformed. Data from the AcquireX experiments.

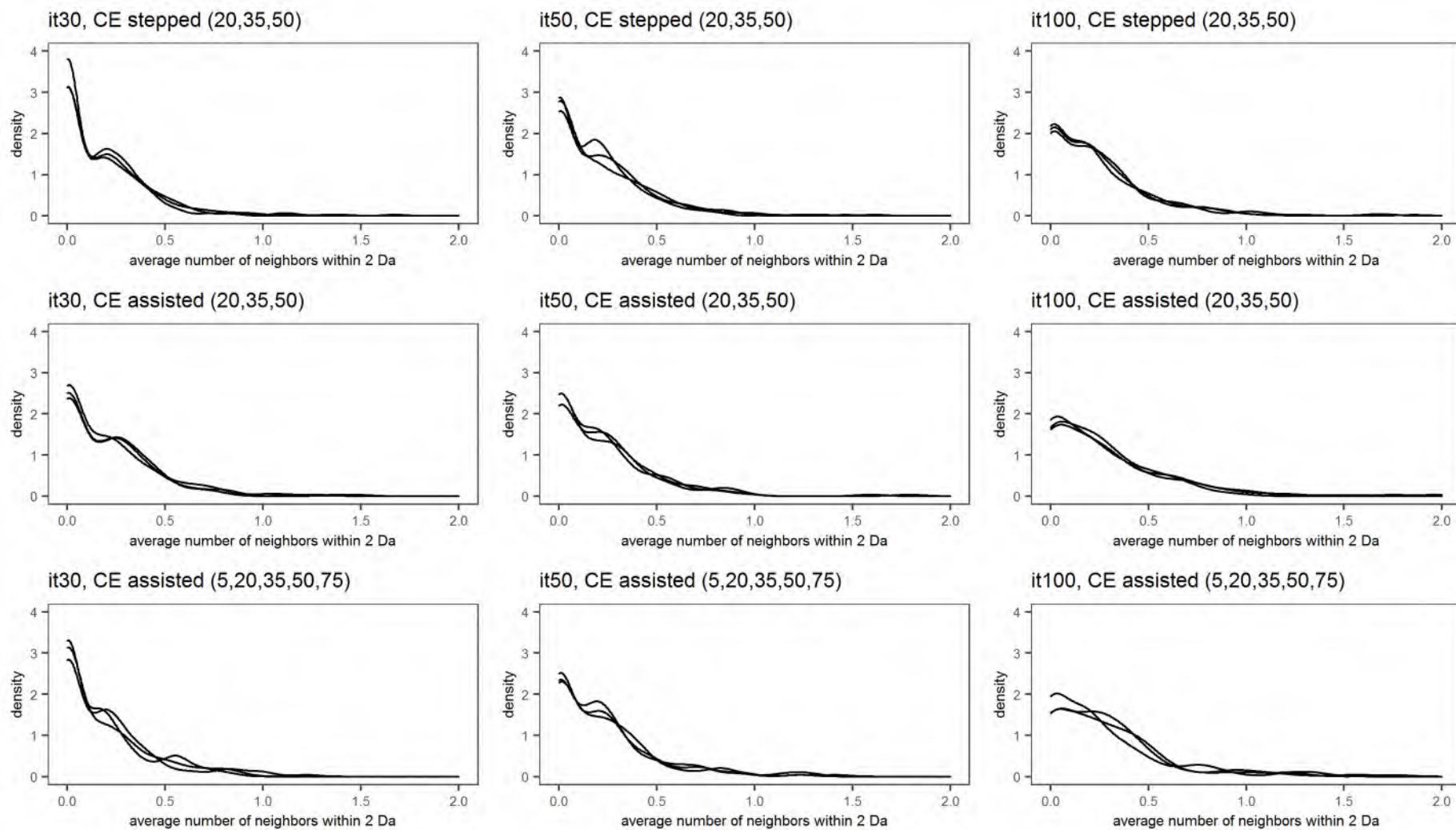


Figure 1.8 – Kernel density plots of the average number of neighbor peaks within a 2-Da interval around any peak. Data from the AcquireX experiments.

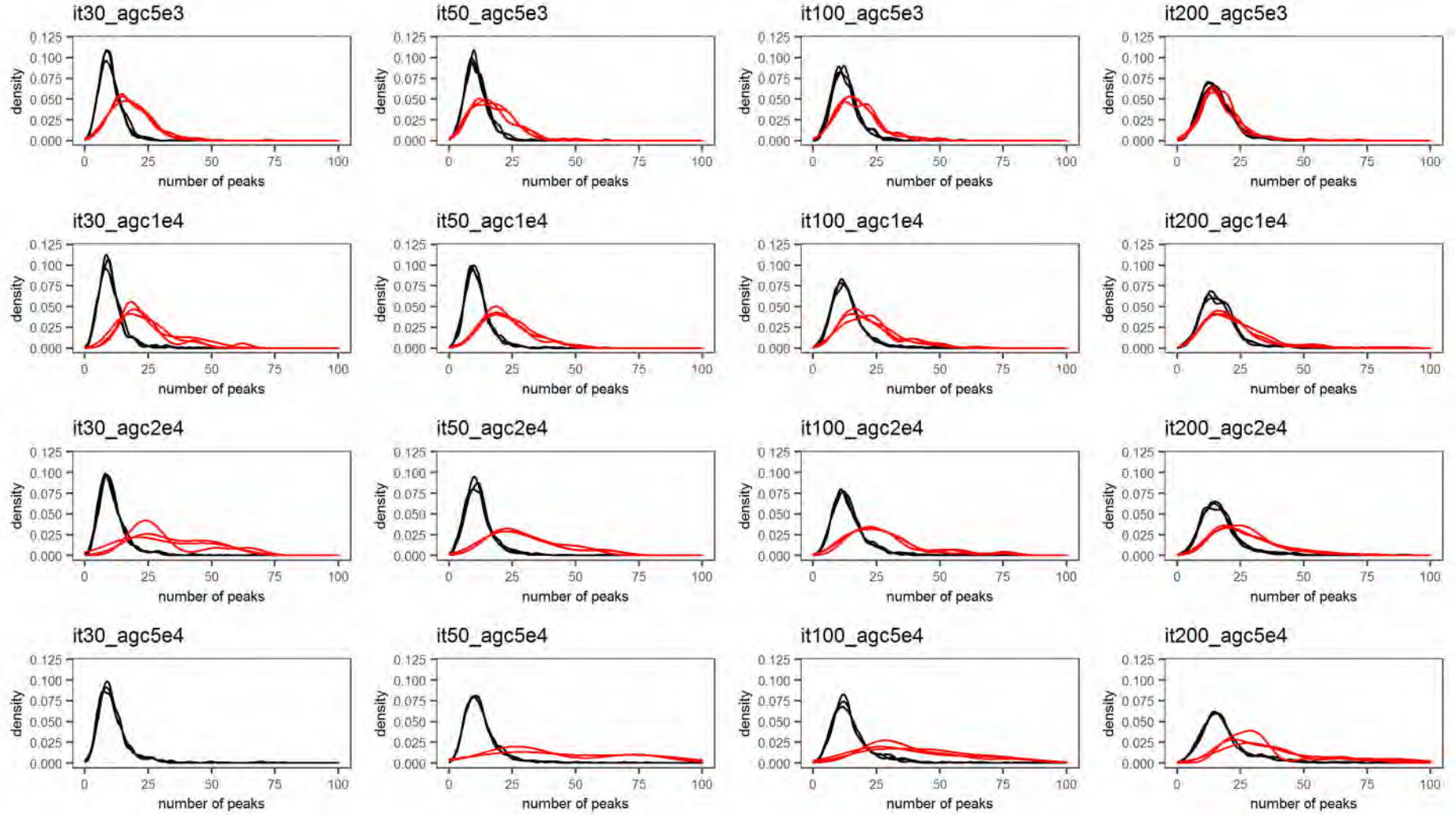


Figure I.9 – Kernel density plots of the number of peaks, square root-transformed. The red distributions correspond to the scans that reached the AGC-target before the maximum injection time. Data from the second set of AGC-target experiments.

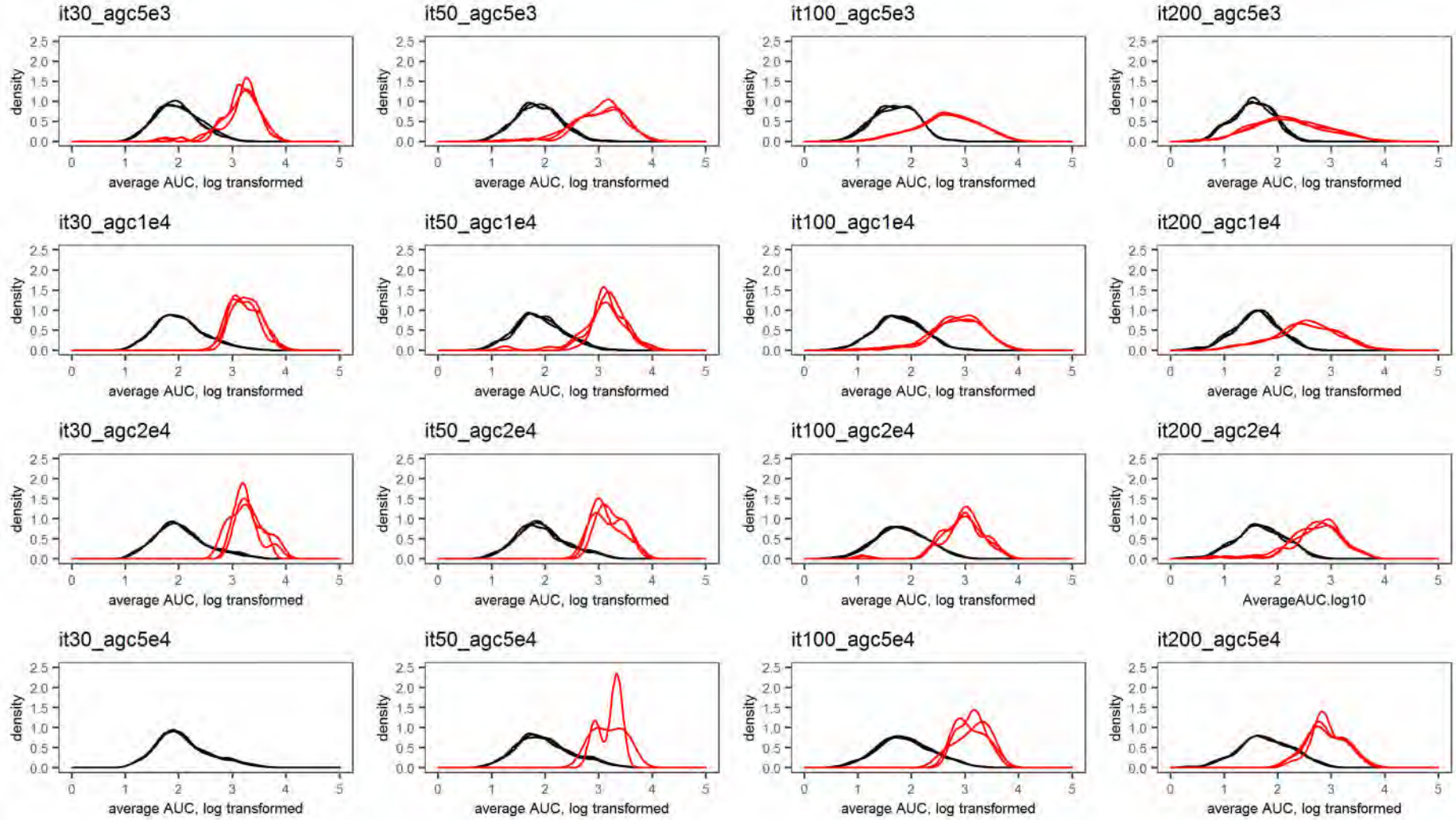


Figure I.10– Kernel density plots of the arithmetic mean of the peak areas, log-transformed. The red distributions correspond to the scans that reached the AGC-target before the maximum injection time. Data from the second set of AGC-target experiments.

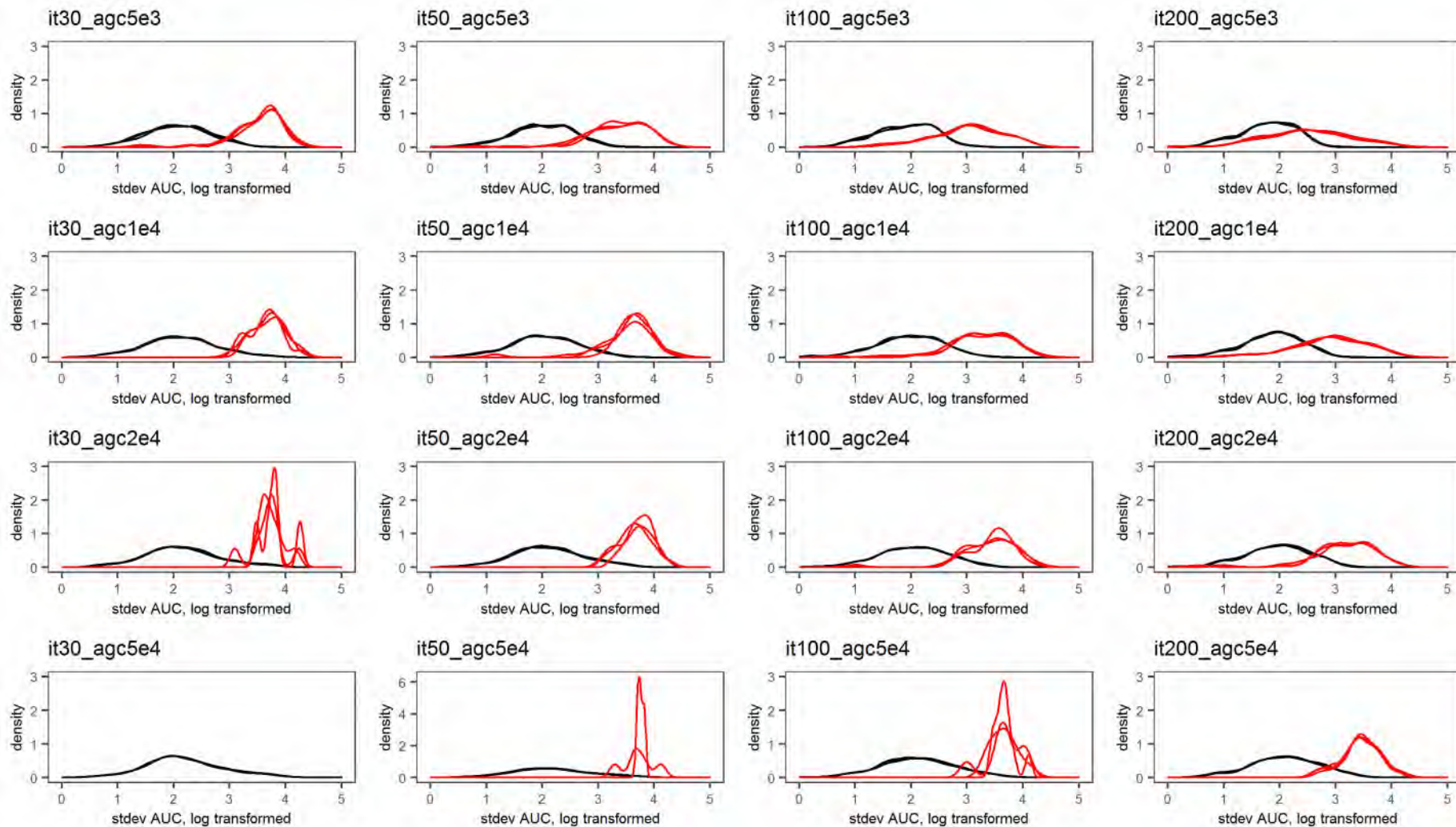


Figure I.11 – Kernel density plots of the standard deviation of the peak areas, log-transformed. Note the divergent scales of graph *it50\_agc5e4*. The red distributions correspond to the scans that reached the AGC-target before the maximum injection time. Data from the second set of AGC-target experiments.

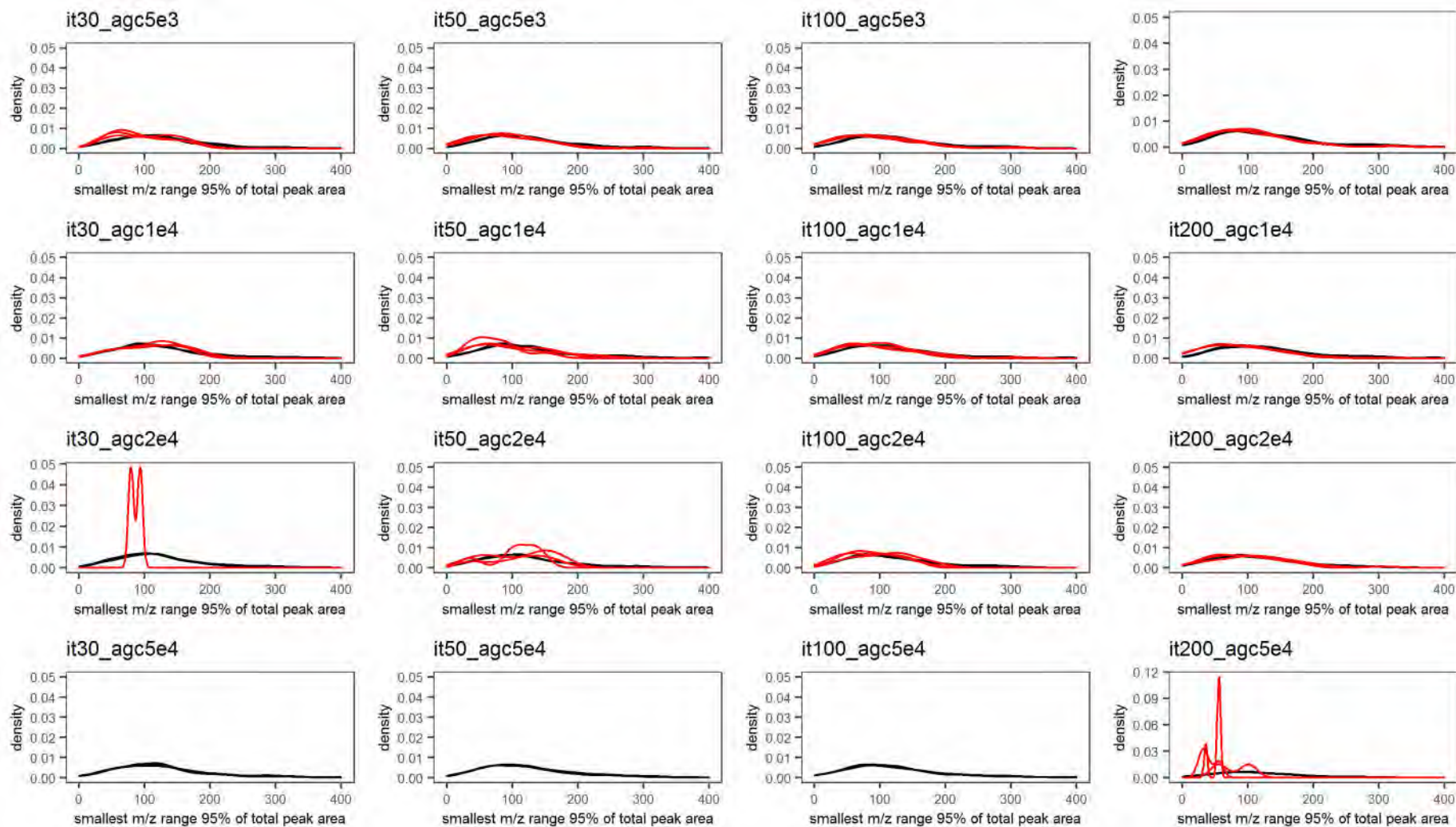


Figure I.12 – Kernel density plots of the smallest  $m/z$  range containing 95% of the total peak area. Note the divergent scales of graph *it200\_agc1e4*. The red distributions correspond to the scans that reached the AGC-target before the maximum injection time. Data from the second set of AGC-target experiments.

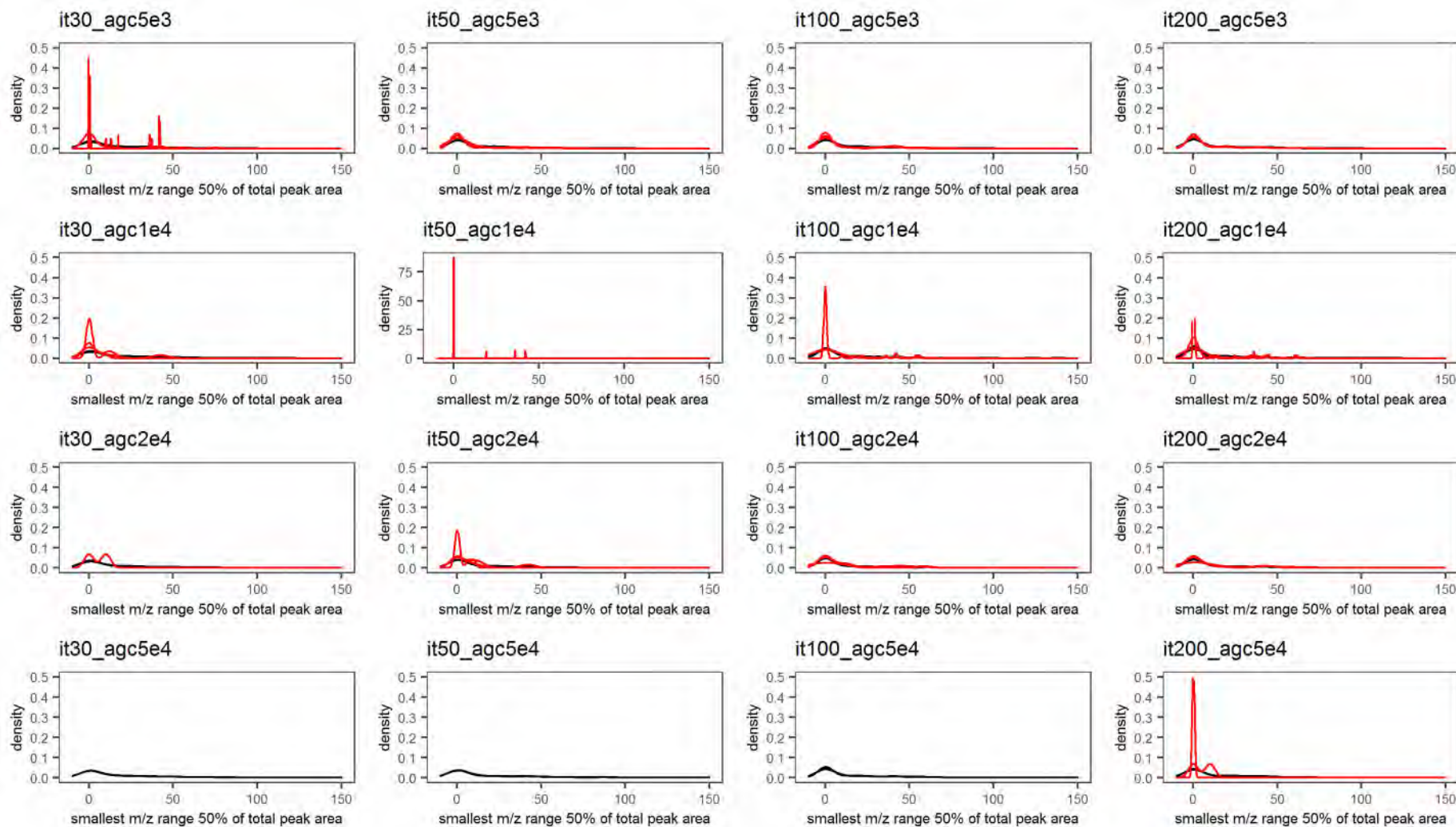


Figure I.13 - Kernel density plots of the smallest  $m/z$  range containing 50% of the total peak area. Note the divergent scales of graph *it50\_agc1e4*. The red distributions correspond to the scans that reached the AGC-target before the maximum injection time. Data from the second set of AGC-target experiments.

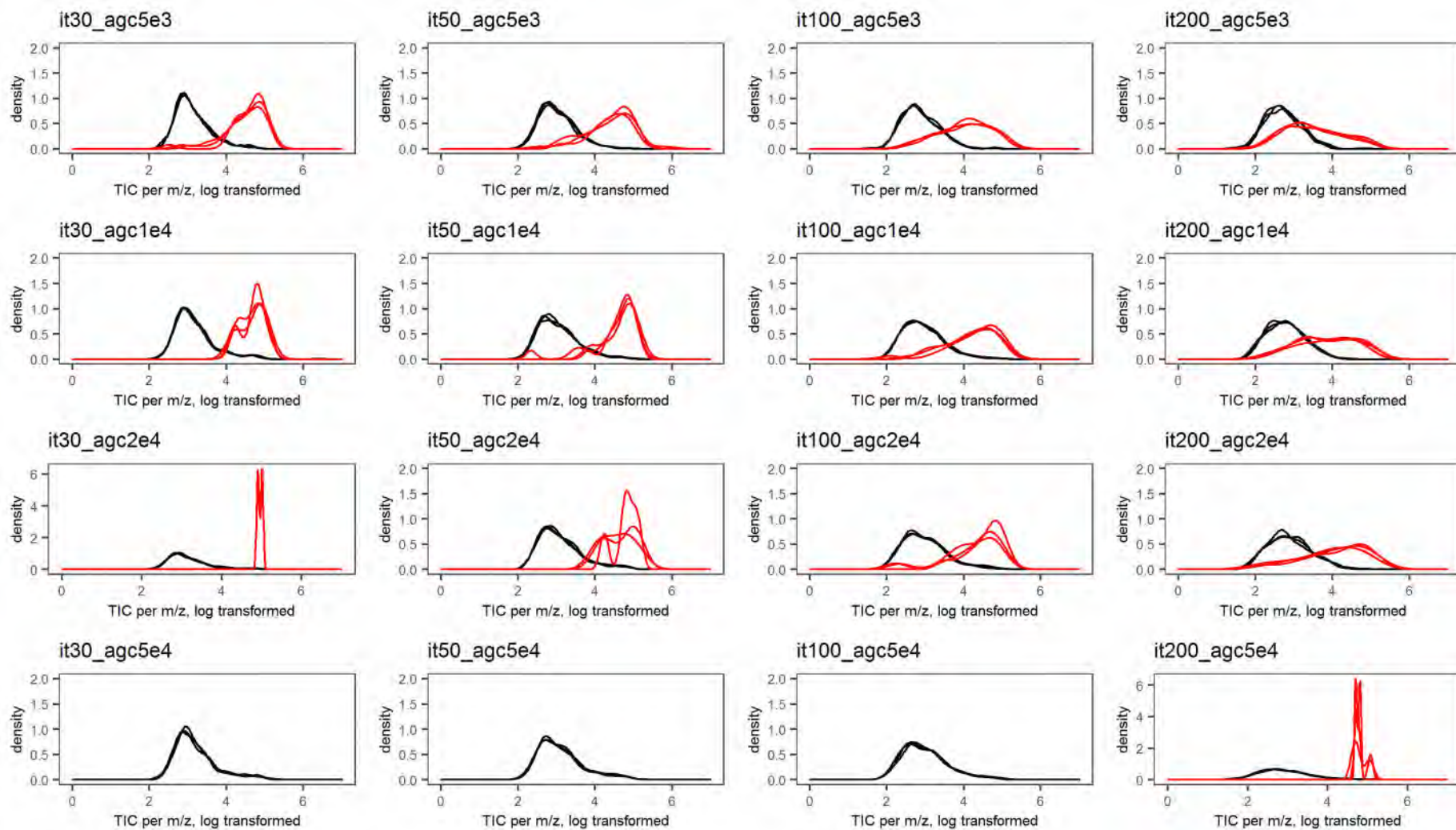


Figure I.14 – Kernel density plots of the total ion current per m/z, log transformed. The red distributions correspond to the scans that reached the AGC-target before the maximum injection time. Data from the second set of AGC-target experiments.



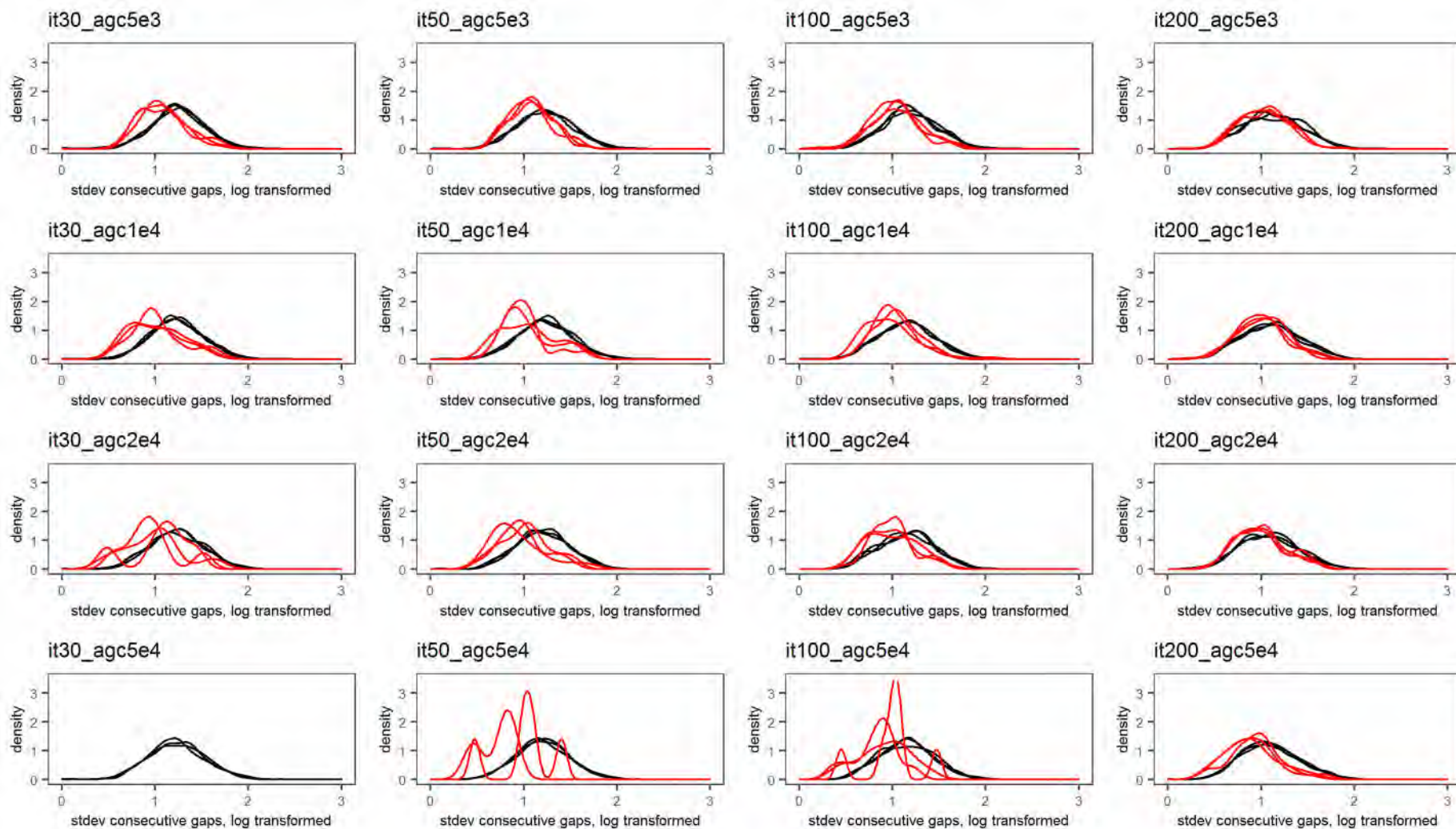


Figure I.15 – Kernel density plots of the standard deviation of the consecutive  $m/z$  gaps between all peaks, log-transformed. The red distributions correspond to the scans that reached the AGC-target before the maximum injection time. Data from the second set of AGC-target experiments.

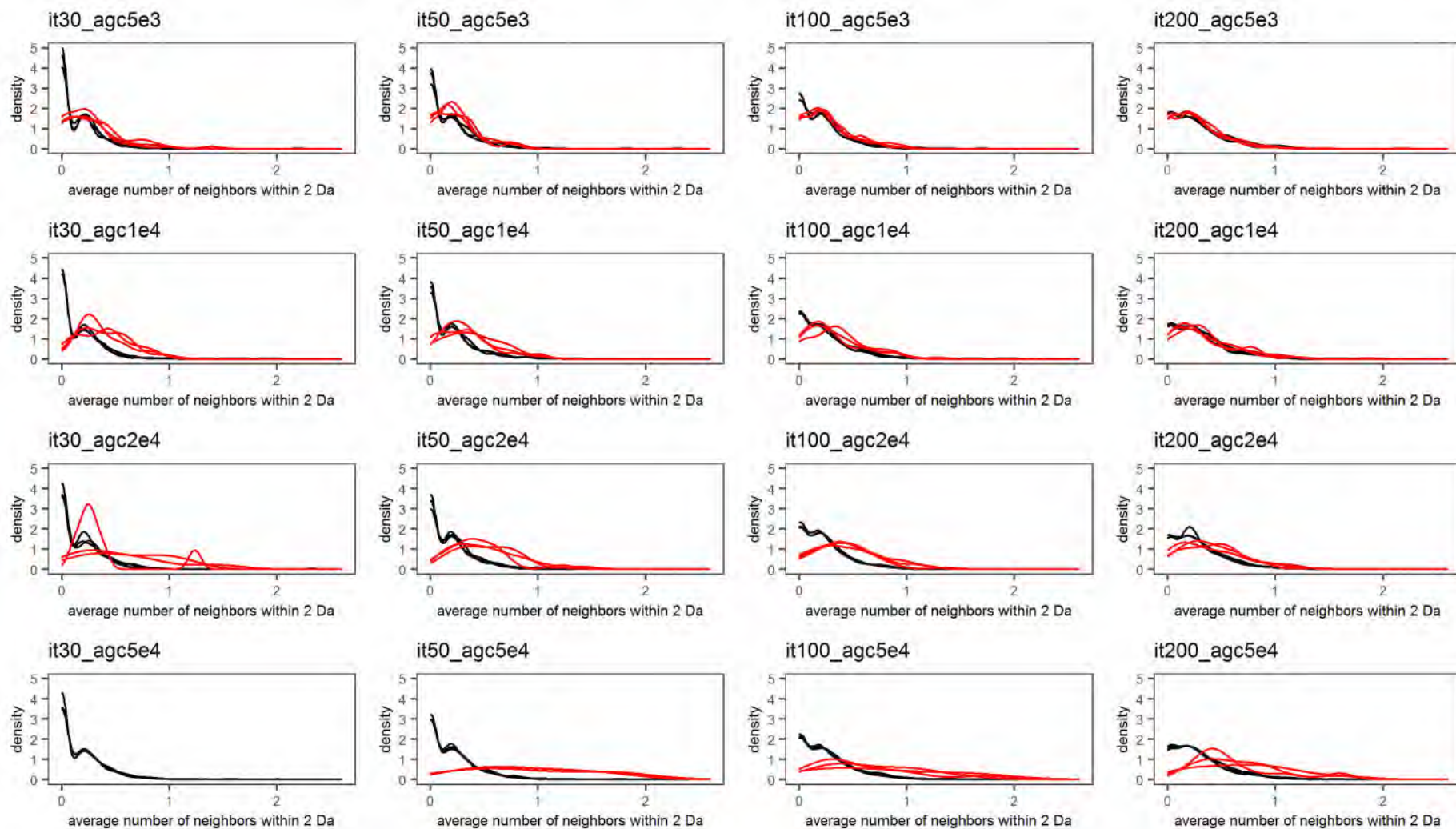


Figure I.16 – Kernel density plots of the average number of neighbor peaks within a 2-Da interval around any peak. The red distributions correspond to the scans that reached the AGC-target before the maximum injection time. Data from the second set of AGC-target experiments.