Supporting Information

Online prioritization of toxic compounds in water samples through intelligent HRMS data acquisition

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

KWR Water Research Institute, P.O. Box 1072, 3430 BB Nieuwegein, The Netherlands

Table of Contents

Table of Figures

S1.1 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.[1]

The AGC-target defines the number of ions to accumulate in the C-trap before the MSn+1 scan is performed.[2] 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; an MSn scan is triggered in case one of these values is reached. 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.[2] 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 (Thermo Fisher Scientific, San Jose, USA) 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.

S1.2 Spectral libraries and chemical databases

Spectral libraries occur in both commercial, such as mzCloud (HighChem LLC, Slovakia), and open source forms, such as MassBank[3]. mzCloud is a spectral library that contains HRMS spectra, both MS2 and MSn, in so-called fragmentation trees.[4] mzCloud is available online (https://www.mzcloud.org) and in Compound Discoverer (Thermo Fisher Scientific, San Jose, USA), both the online library as its offline version mzVault (Thermo Fisher Scientific, San Jose, USA). 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 ($n_{MassBank}$ = 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.[3] NORMAN MassBank is directed towards environmentally relevant contaminants.[5]

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_{SusDat} = 65697)[6] which is composed of various environmentally relevant suspect lists such as the STOFFIdent[7] and KWR Sjerps lists[8] with water relevant compounds (n_{Sjerps} = 5722), and the UBAPMT list[9] with REACH substances that are (very) persistent, (very) mobile and toxic (n_{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 CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard)[10] is a chemical database held by the U.S. Environmental Protection Agency and is i.e. linked to the toxicity database ToxCast.[11] 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 (www.chemspider.com) 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[12] or Compound Discoverer to identify detected features.

S2.1 Workflow screening with ToxAlerts and fragmentation

1. Subtract CAS-registry numbers (CASRN) from Chemical_Summary_190708.csv (file located in INVITRODB_V3_2_SUMMARY.zip, downloaded via https://epa.figshare.com/articles/ToxCast_and_Tox21_Summary_Files/6062479) and generate an Excel file with unique MS-ready SMILES.

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', 'developmental and mitochondrial toxicity' and 'endocrine disruption' (only approved alerts).
c. Select "Screen compounds against alerts"
d. Upload the file containing MS-ready SMILES (ToxCast_msready_smiles.xlsx)
e. Deselect all boxes in 'Preprocessing of molecules (Chemaxon)'
f. Tick the box "Only 187 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; removing duplicate alerts and alerts with less than 5 molecules and generating files suitable for fragmentation with CFM-ID 2.0.

4. Fragment the molecules per structural alert with CFM-ID in the command line using the windows executable 'cfm-predict.exe', an example is given below for structural alert TA322.

```
> 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. CFM-ID 2.0 cannot handle molecules other than singly charged.

S2.2 Instrument settings LC-HRMS experiments

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 analytical XBridge BEH C18 XP column (150mm x 2.1 mm i.d., 2.5 µm 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 µL. 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 min isocratic at 100% B, 29 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 °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 $2.0x10^5$ with a maximum IT of 100 ms. Data was acquired in profile mode.

S2.3 Inclusion lists

The five inclusion lists were:

a) NORMAN Substance Database (susdat_2020-02-03-164350.csv, downloaded at 3 February 2020).[6] This dataset was filtered for organic compounds with $[M+H]^+$ ≥ 80 Da and ≤ 1000 Da and log K_{OW} (predicted with EPISuite) ≥ -2.5 and ≤ +3.5, resulting in 18667 compounds.

b) 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.

(1) $$t_R = \log K_{OW} / 0.254 + 5.1945$$

c) UBAMPT (Potential Persistent, Mobile and Toxic (PMT) substances, retrieved from NORMAN SusDat at 12 February 2020), also filtered for organic compounds with $[M+H]^+$ ≥ 80 Da and ≤ 1000 Da, resulting in 192 m/z values.[9]

d) Extended KWR Sjerps list (Sjerp_2016_WatResManuscript_SI-1.docx, downloaded at 19 February 2020) filtered for organic compounds with $[M+H]^+$ ≥ 80 Da and ≤ 1000 Da, resulting in 3399 m/z values.[8]

e) 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 *Table S2*.

S2.4 Data analysis

Compound Discoverer 3.1 (Thermo Fisher Scientific, San Jose, USA) was used to perform i.e. peak picking, retention time alignment and compound annotation. The Compound Discoverer workflow settings are given in S2.5. 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, San Jose, USA). 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.

Spectrum similarity scores were calculated using the function SpectrumSimilarity() from the R-package OrgMassSpecR[13] (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) $$\cos\theta = (u \cdot v)/(\sqrt{\sum u^2} \times \sqrt{\sum v^2})$$

Fragment annotation was performed with the R-package metfRag[14] 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 $5x10^4$ within the maximum IT.

S2.5 Compound Discoverer workflow parameters

Select Spectra

1. General Settings

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

2. 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

3. 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)

4. Peak Filters

S/N Threshold (FT-only):	1.5

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

1. 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

1. 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

2. 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. Predicted Composition Annotation

Check All Predicted Compositions:	TRUE

Assign Compound Annotations

1. General Settings

Mass Tolerance:	3 ppm

2. Data Sources

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. Scoring Rules

SFit Range:	20
SFit Threshold:	20
Use mzLogic:	TRUE
Use Spectral Distance:	TRUE

Search mzVault

1. Search Settings

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:	HighChemDP
Use Retention Time:	FALSE

Fill Gaps

1. General Settings

Mass Tolerance:	3 ppm
S/N Threshold:	1.5
Use Real Peak Detection:	TRUE

Mark Background Compounds

1. General Settings

Hide Background:	FALSE
Max. Blank/Sample:	0
Max. Sample/Blank:	10

Search mzCloud

1. General Settings

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

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

S3.1 Toxicity validation

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 S3.1.1*). 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. This indicates that the alerts used for screening were not covering all chemicals active in these toxic endpoints. Moreover, not all chemicals are tested on all included ToxCast assays, leading to a data gap (shown in *Figure S3.1.1* per bioassay and in *Figure S3.1.2* per structural alert).[15] Consequently, if a compound is not marked as active, it could be inactive or its activity unknown. The data distribution for non-genotoxic carcinogenicity and genotoxic carcinogenicity and mutagenicity illustrated in *Figure S3.1.1* 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 S3.1.1*.
Compounds can only be marked as inactive if they are tested and do not show activity in all linked assays. This is not the case for any of them (see *Figure S3.1.5-8*), as no compound has been tested in all assays.

Table S3.1.1 – 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%

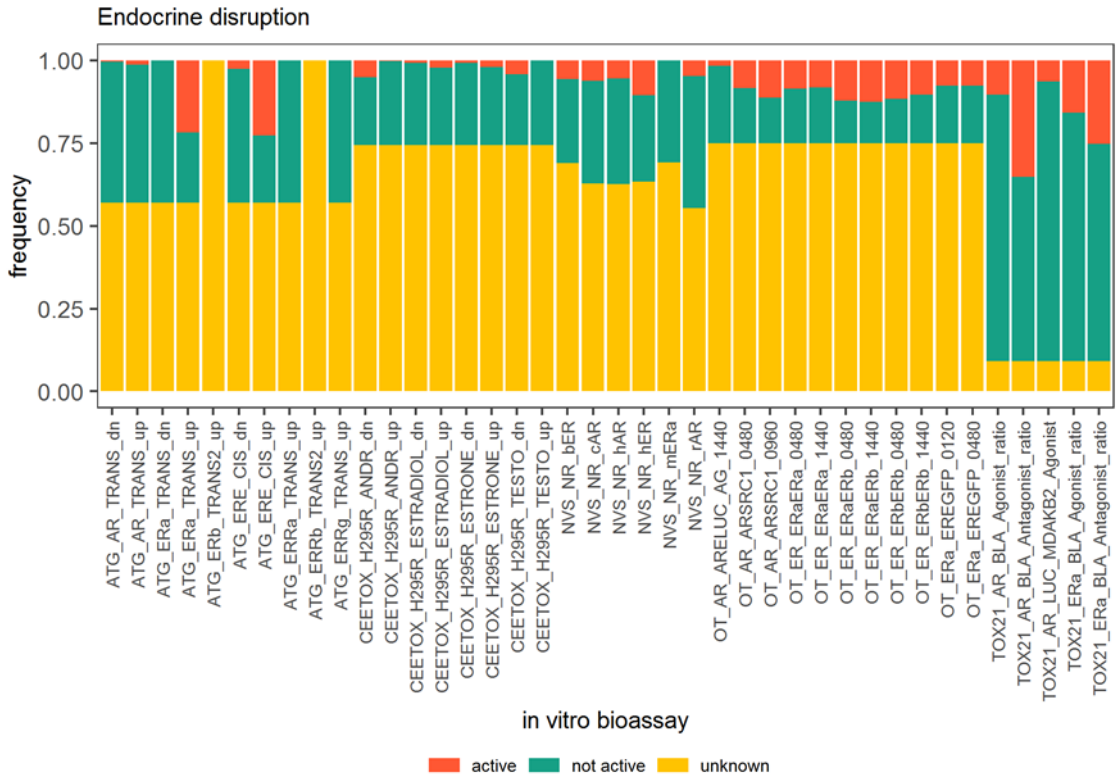

Figure S3.1.1 – Distribution of available toxicity information between in vitro bioassays for the toxic endpoint endocrine disruption.

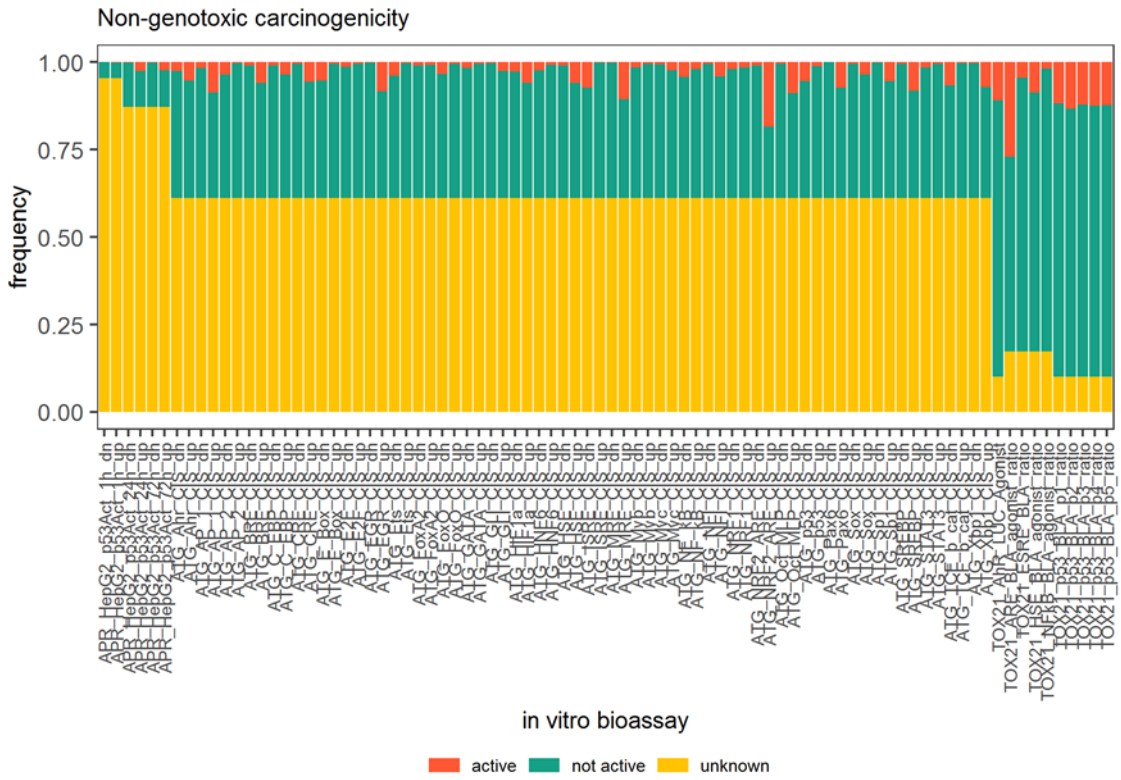

Figure S3.1.2 – Distribution of available toxicity information between in vitro bioassays for the toxic endpoint non-genotoxic carcinogenicity.

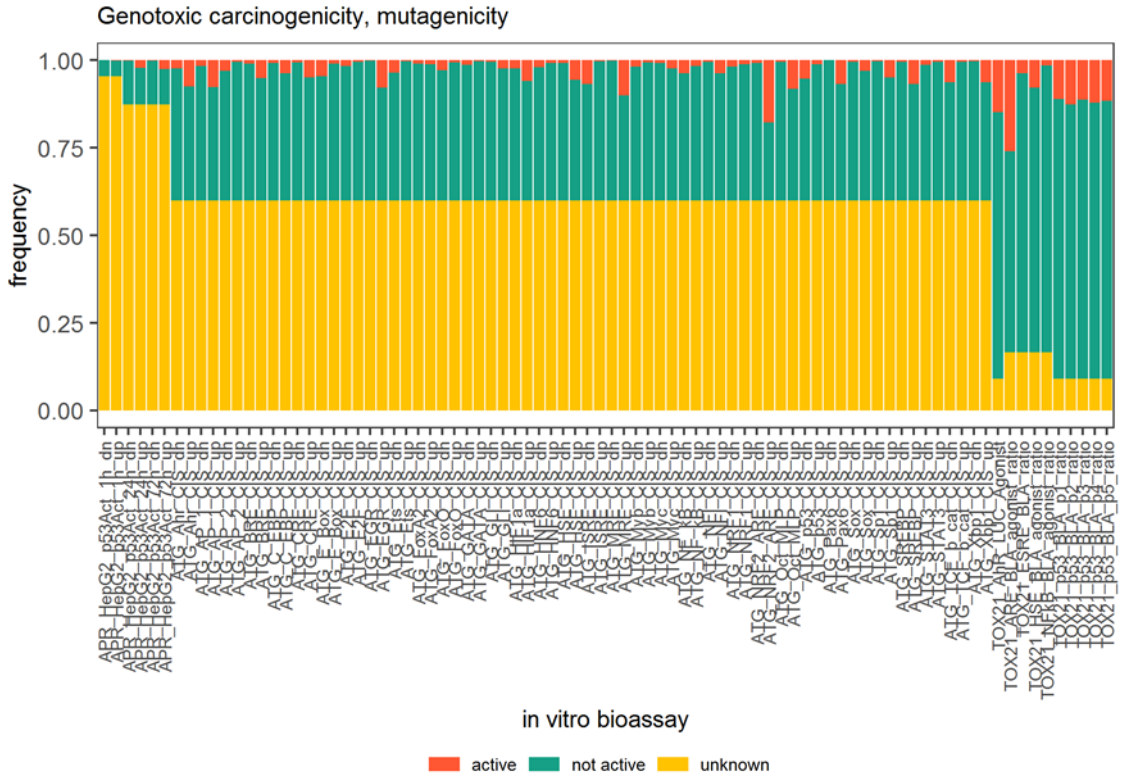

Figure S3.1.3 – Distribution of available toxicity information between in vitro bioassays for the toxic endpoint genotoxic carcinogenicity, mutagenicity.

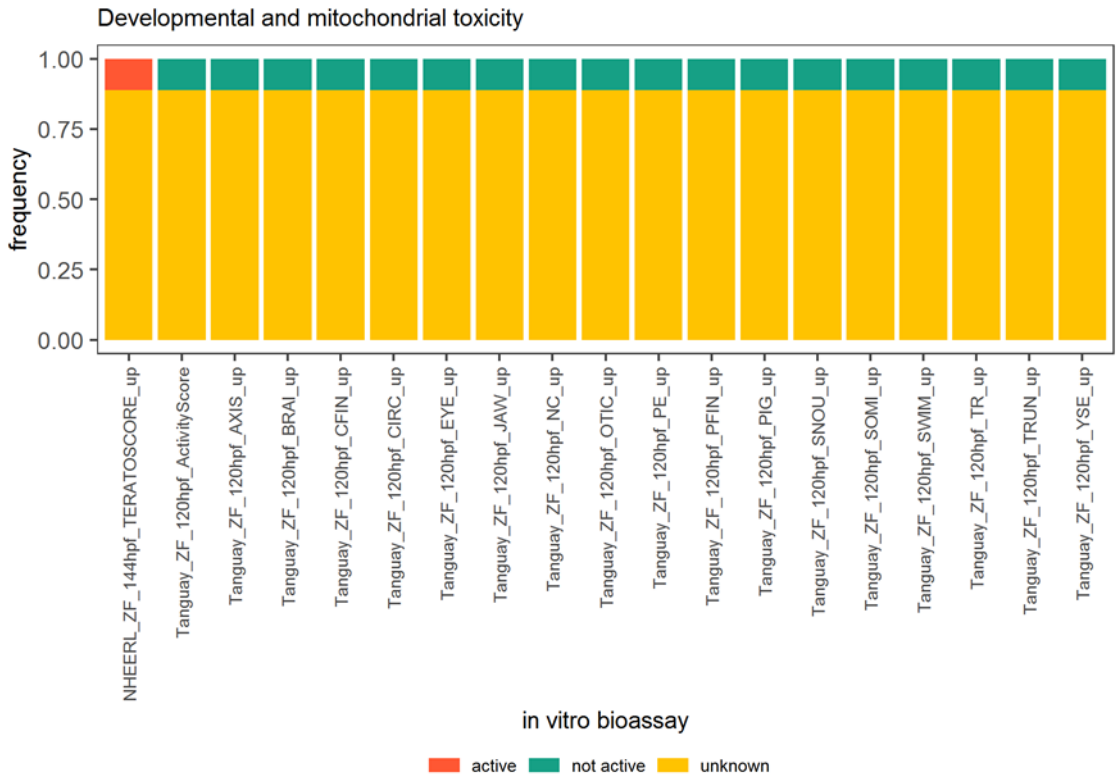

Figure S3.1.4 – Distribution of available toxicity information between in vitro bioassays for the toxic endpoint developmental and mitochondrial toxicity.

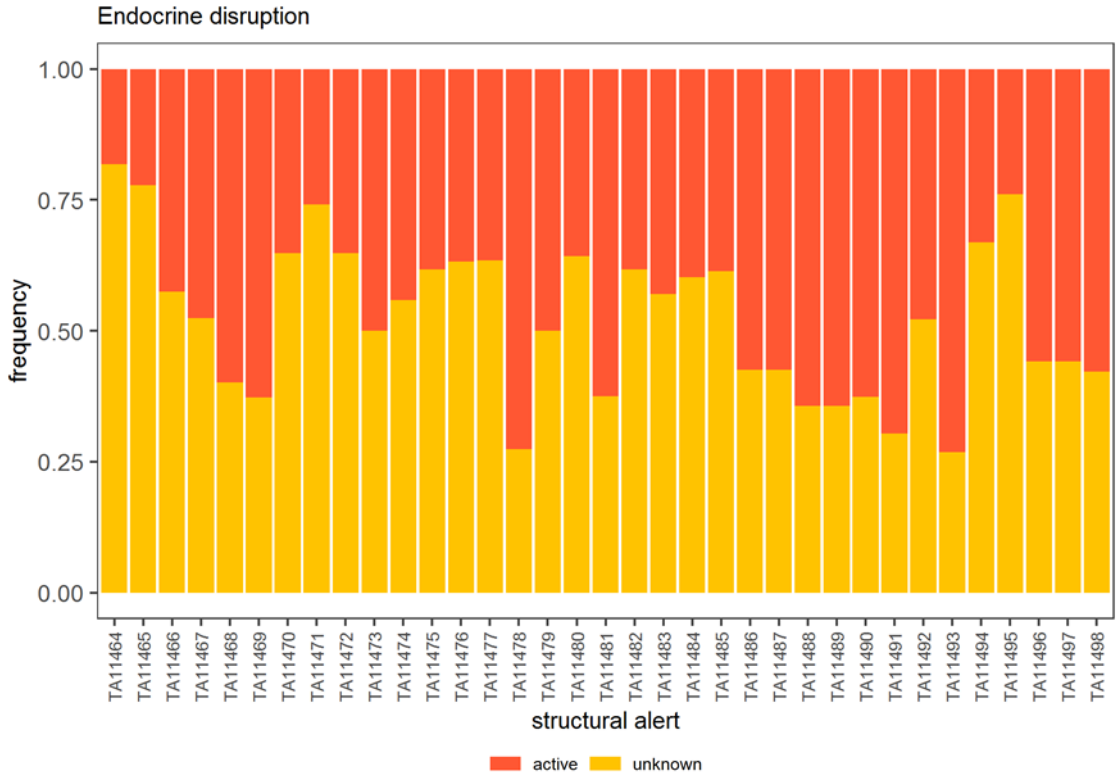

Figure S3.1.5 – Distribution of available toxicity information between in vitro bioassays per structural alert for the toxic endpoint endocrine disruption.

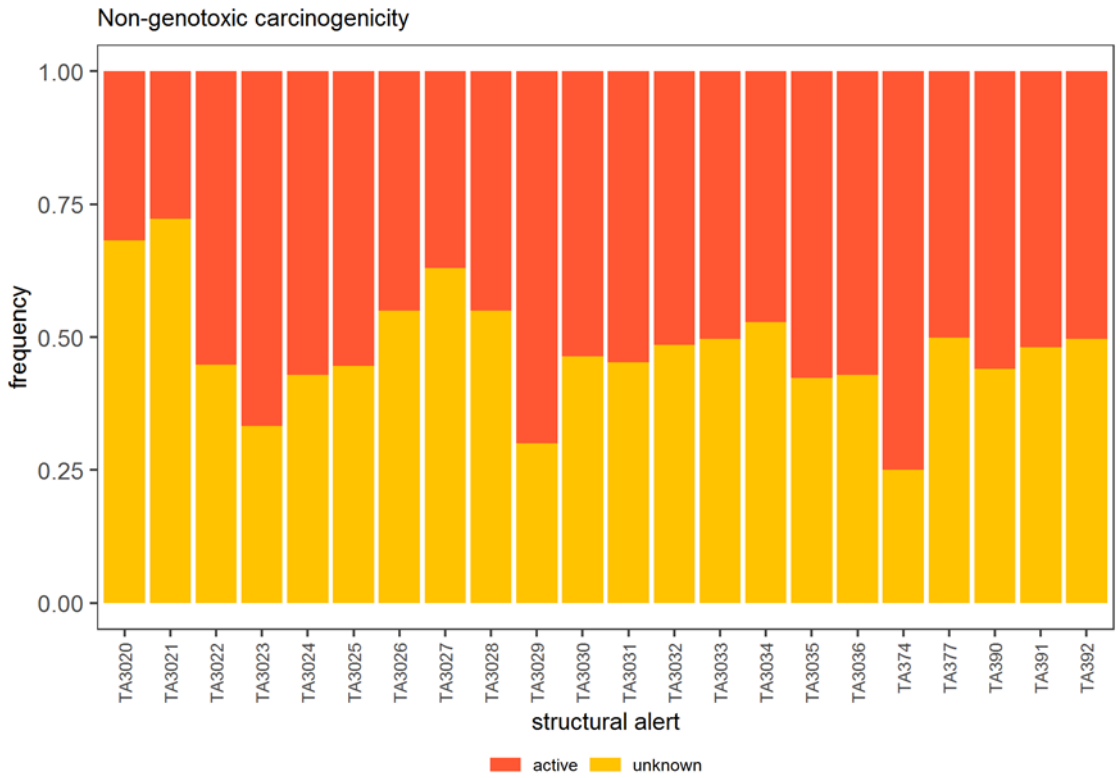

Figure S3.1.6 – Distribution of available toxicity information between in vitro bioassays per structural alert for the toxic endpoint non-genotoxic carcinogenicity.

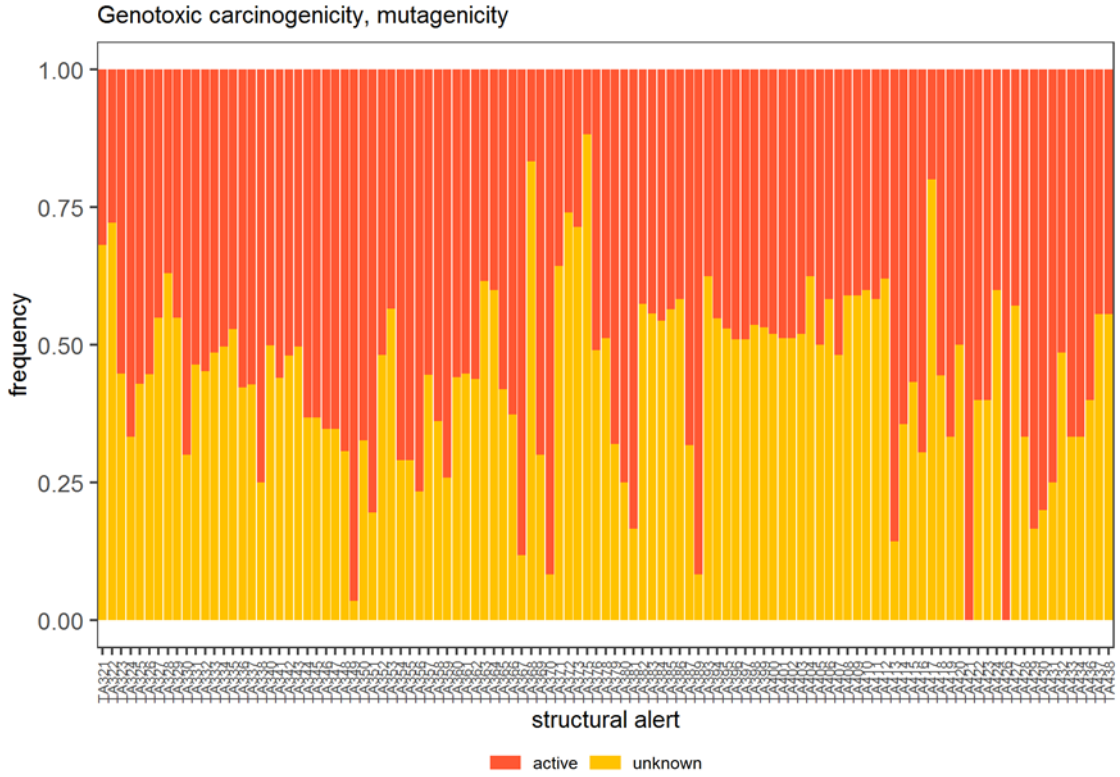

Figure S3.1.7 – Distribution of available toxicity information between in vitro bioassays per structural alert for the toxic endpoint genotoxic carcinogenicity, mutagenicity.

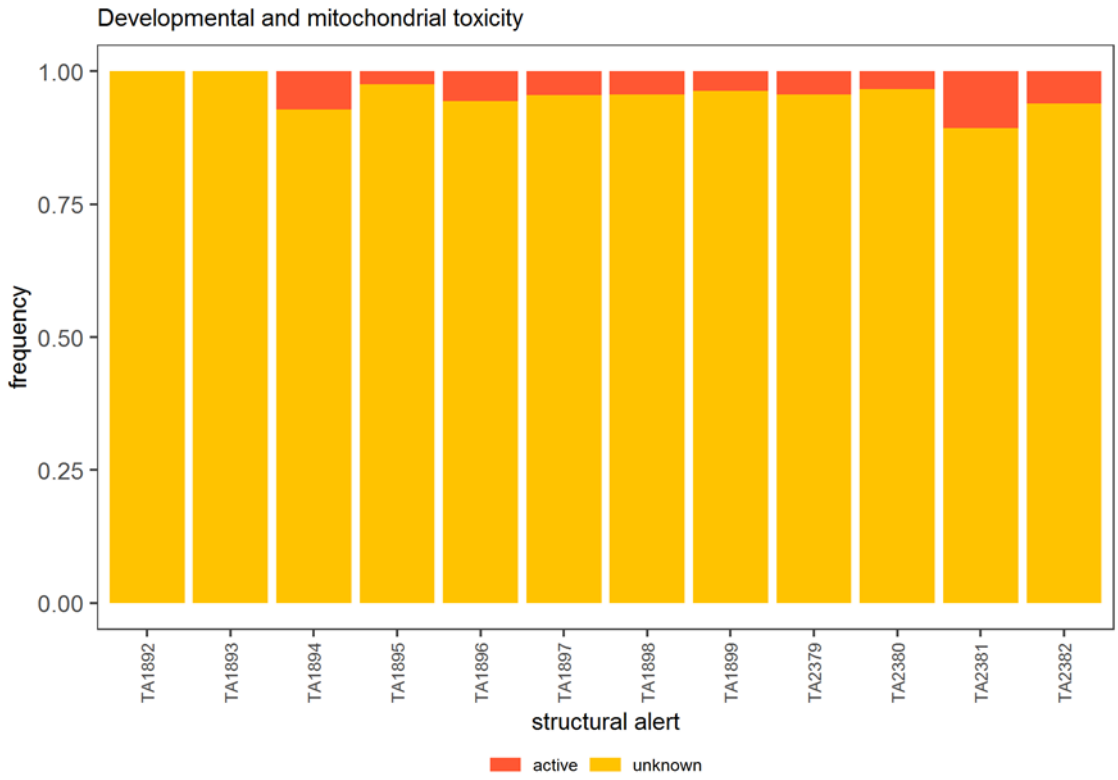

Figure S3.1.8 – Distribution of available toxicity information between in vitro bioassays per structural alert for the toxic endpoint developmental and mitochondrial toxicity.

S3.2 Validation of *in silico* fragmentation

After validation of the toxicity of the compounds with structural alerts, the *in silico* fragmentation results were validated as well using experimental fragmentation data. The results of the validation study are shown in *Table S3.2* and *Figure S3.2*. Fewer molecules have ≥50% of their fragments matched between the two datasets based on $\%_{MassBank}$ than based on $\%_{CFM\text{-}ID}$. 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. Moreover, CFM-ID is not able to predict all fragments since it is no complete prediction.

Table S3.2 – Results of the validation study with MassBank and CFM-ID at the three different energy levels. These values are based on the 587 overlapping molecules between MassBank and CFM-ID, the values represent the number of molecules.

CFM-ID energy	no CFM-ID fragments with intensity ≥ 5	0% match, based on $\%_{MB}$	≥ 50% match, based on $\%_{MB}$	0% match, based on $\%_{CFM\text{-}ID}$	≥ 50% match, based on $\%_{CFM\text{-}ID}$
energy 0	0	44	100	44	398
energy 1	0	71	140	71	325
energy 2	47	186	20	186	144

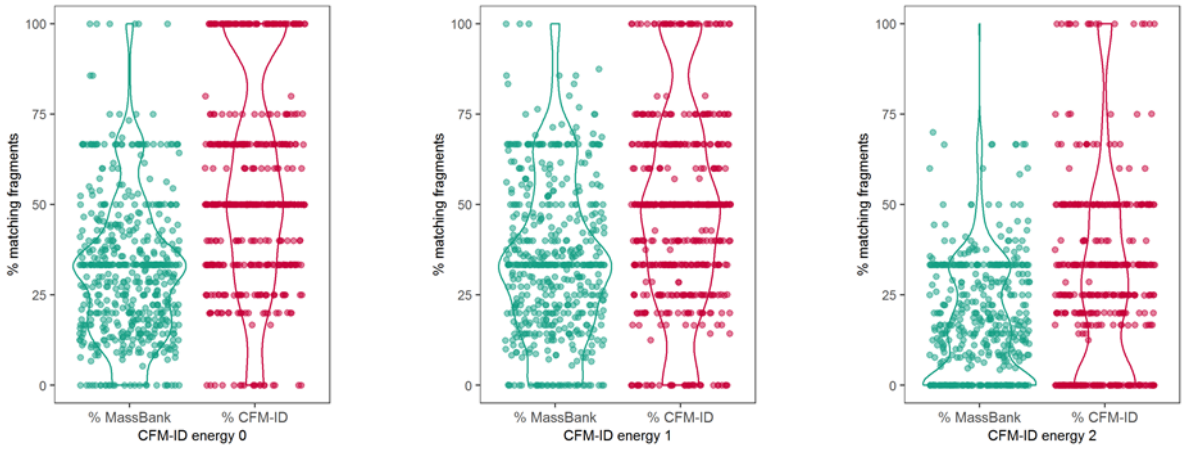

Figure S3.2 – Results of the validation study with MassBank and CFM-ID at the three different energy levels, shown in violin plots. Based on the 587 overlapping molecules between MassBank and CFM-ID.

S3.3 Acquisition parameter optimization

CE mode and maximum IT

Three different CE settings were used in the acquisition parameter experiments: stepped CE 20, 35 and 50, assisted CE 20, 35 and 50, and assisted CE 20, 35, 50 and 75, all with an maximum ion injection time (IT) of 30 ms, 50 ms and 100 ms. For the experiments with different MS2 AGC-targets, the AGC-targets $5x10^4$, $2x10^4$, $1x10^4$ and $5x10^3$ and ITs 30 ms, 50 ms, 100 ms and 200 ms were used. Detailed explanation on these acquisition parameters is given in *S1.1*. This resulted in 16 methods measured in triplicate. The novel acquisition software AcquireX (Thermo Fisher Scientific, San Jose, USA) was used to test the effect of automatic exclusion of background signals from fragmentation. The methods were edited using Thermo Xcalibur Instrument Setup (version 4.2.28.14, Thermo Fisher Scientific Inc.). All methods were based on the LC-HRMS NTS RP method in positive ionization mode as described above.

Prior to implementing MS triggers for the online prioritization of toxic compounds, it was assessed whether the use of a background exclusion list could decrease fragmentation of background signals and thereby free up extra cycle time for (additional) MS2 scans during online prioritization. To this end, the novel acquisition software AcquireX (Thermo Fisher Scientific, San Jose, USA) that automatically generates a background exclusion list based on the features that are detected in a blank sample was used. This lead to a significant decrease in the percentage of MS2 scans of features marked as background from 94.7 ± 0.9 % (n = 9) fragmented background features to 21.9 ± 1.4 % (n = 9). 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 a few selected acquisition parameters on the quality of the acquired MS2 spectra was studied, i.e. of the different CE modes stepped and assisted CE together with maximum IT, and of four different AGC-targets and maximum ITs (*S1.1*).

To gain insight into the spectral quality per experimental condition, fragment ion annotation was assessed using mzCloud and mzLogic scores as a proxy. These two scores assess the match between the experimental spectrum and an mzCloud library spectrum, and a combination of mzCloud and structural data from a selected chemical database or suspect list, respectively. As no effect of CE modes was detected in the mzLogic and mzCloud score distributions, *in silico* predicted fragmentation spectra were used to determine the information content of a spectrum. This was based on the hypothesis that a spectrum with higher information content would lead to more fragments that could be annotated. This was assessed using three different metrics: the number of annotated fragments, the percentage of annotated fragments relative to all fragments and the percentage annotated peak area relative to the total peak area. The MS2 spectra of the four spike-in compounds DEET, primicarb, phenazone and triphenylphosphine oxide (see chemical structures in *Figure S3.3.1*) were selected, annotated with MetFrag and annotation was compared for each method (*Figure S3.3.2, S3.3.3, S3.3.4*).

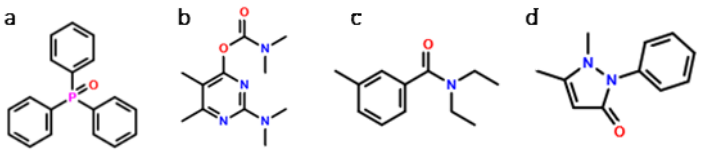

Figure S3.3.1 – Chemical structures of the four spike-in compounds triphenylphosphine oxide (a), primicarb (b), DEET (c) and phenazone (d).

The number of annotated fragments significantly increased with longer ITs for primicarb (p-value of 0.00356, two-way ANOVA), phenazone (p-value of $2.32x10^{-10}$, two-way ANOVA), triphenylphosphine oxide (p-value of $2.05x10^{-5}$, two way ANOVA) and DEET (p-value of $7.468x10^{-5}$, Kruskal-Wallis) (*Figure S3.3.2*). 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 of $2.21x10^{-9}$, two-way ANOVA), for phenazone assisted CE 20-50 (p-value of $3.44x10^{-7}$, two-way ANOVA). For primicarb the best mode varied per maximum IT but the effect was significant (p-value of 0.00158, two-way ANOVA).

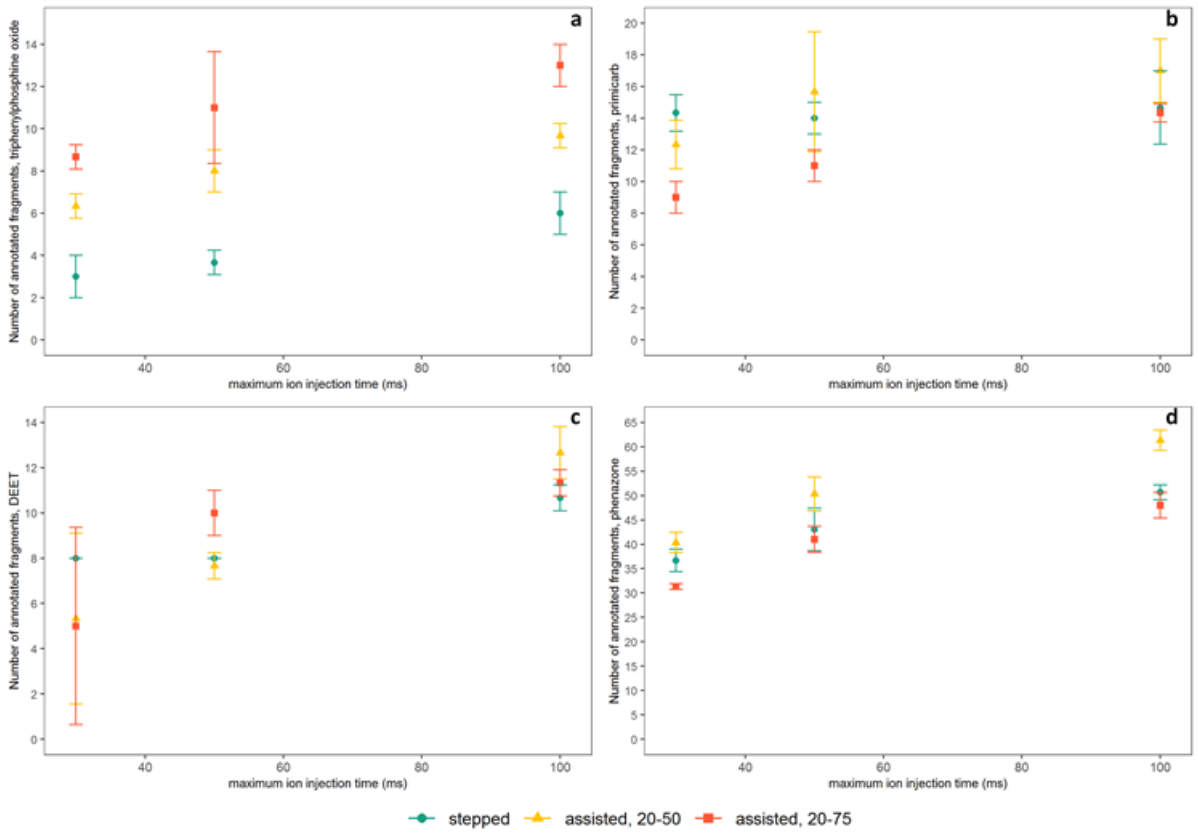

Figure S3.3.2 – Number of annotated fragments, AcquireX experiments. a) triphenylphosphine oxide, b) primicarb, c) DEET, d) phenazone.

In contrast, the percentage of annotated fragments relative to the total number of peaks significantly changes upon CE mode for primicarb (p-value of 0.00719, two-way ANOVA) and triphenylphosphine oxide (p-value of 3.14e-05, two-way ANOVA) but not for phenazone and DEET (*Figure S3.3.3).* The effect of IT was only significant in the case of triphenylphosphine oxide (p-value of 0.0252, two-way ANOVA). *Figure S3.3.4* displays the annotated percentage of the total peak area. The effects of maximum IT and CE type were significant for both triphenylphosphine oxide (p-value of 1.24×10^{-8} (IT) and $< 2 \times 10^{-16}$ (CE), two-way ANOVA) and phenazone (p-value of $< 2 \times 10^{-16}$ (IT and CE)).

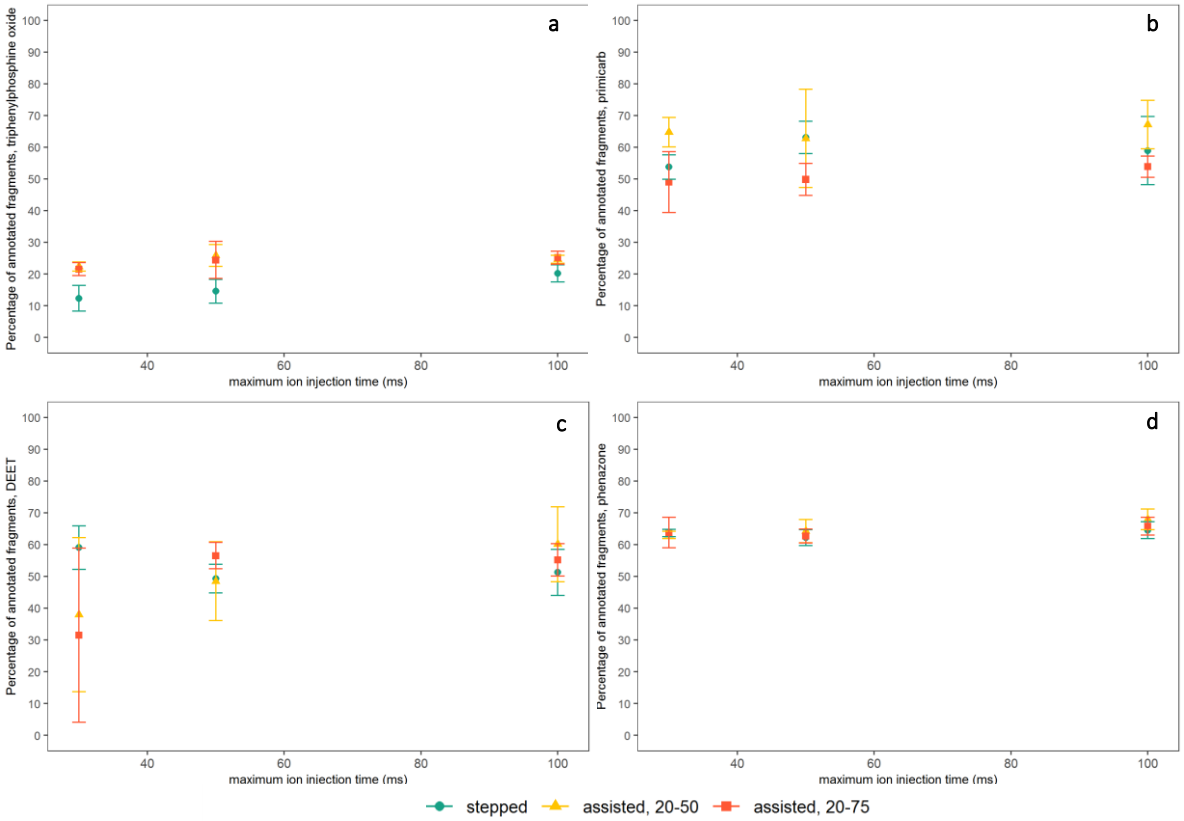

Figure S3.3.3 – Percentage of annotated peaks in the MS2 scan, AcquireX experiments. a) triphenylphosphine oxide, b) primicarb, c) DEET, d) phenazone.

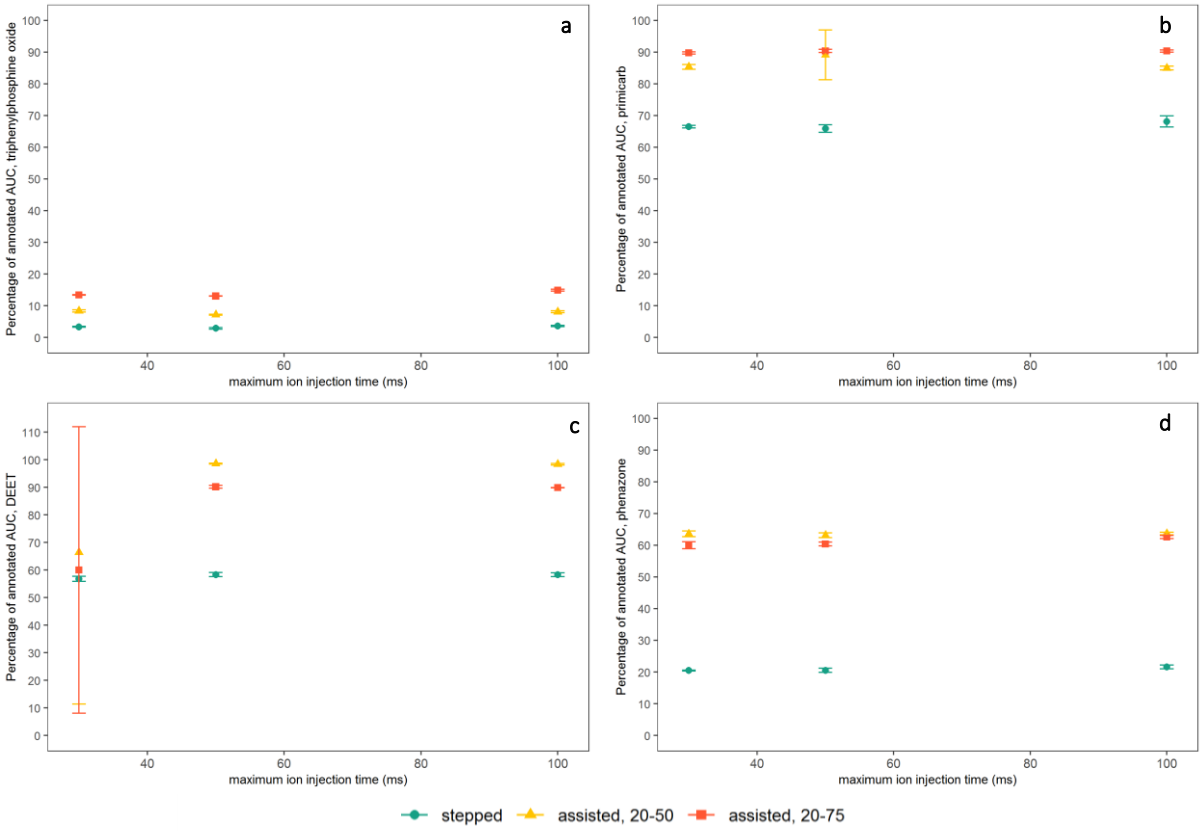

Figure S3.3.4 – Percentage of annotated peak area (area under the curve, AUC) in the MS2 scan, AcquireX experiments. a) triphenylphosphine oxide, b) primicarb, c) DEET, d) phenazone.

Overall, 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 results show that CE needs to be optimized per compound.[16] However, it was expected that fragments generated at multiple CEs would lead to more informative spectra than fragments generated at one optimal CE. The advantage of assisted CE over stepped CE is that the optimal CE is determined experimentally for each precursor (*S1.1*). This is optimization in terms of precursor intensity, not based on the information content of the spectrum. 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. 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 (close to 0 ms).[1] 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. Interestingly, large standard deviations were detected for some spectra from technical replicates (marked with large error bars), possibly caused by low-intensity signals close to 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, potential subformulas can be assigned to the fragments.[17]

AGC-target

The results of the second set of AGC target experiments showed an increase in percentage of MS2 scans reaching the AGC target before the maximum IT with lower AGC target and longer IT (see *Figure S3.3.5*). The different AGC targets do not lead to differences in the most dominant fragments in the spectra, i.e. spectral similarity scores that give more weight to the most intense signals do not indicate differences. However, when looking at annotated fragments a significant increase with increasing AGC-target was visible (see *Figure S3.3.6*). The percentages of annotated peaks and peak area (see *Figure S3.3.7* and *S3.3.8*) do not show significant differences except for DEET where the maximum IT had a significant effect on the percentage of annotated fragments (p-value = 0.009109). The peaks that could not be annotated are either noise or fragments that could not be predicted by combinatorial approaches such as used in MetFrag. All in all, these results suggest that higher AGC-targets and corresponding higher ITs lead to spectra with higher information content but there is a trade-off with the total number of MS2 scans that can be taken.

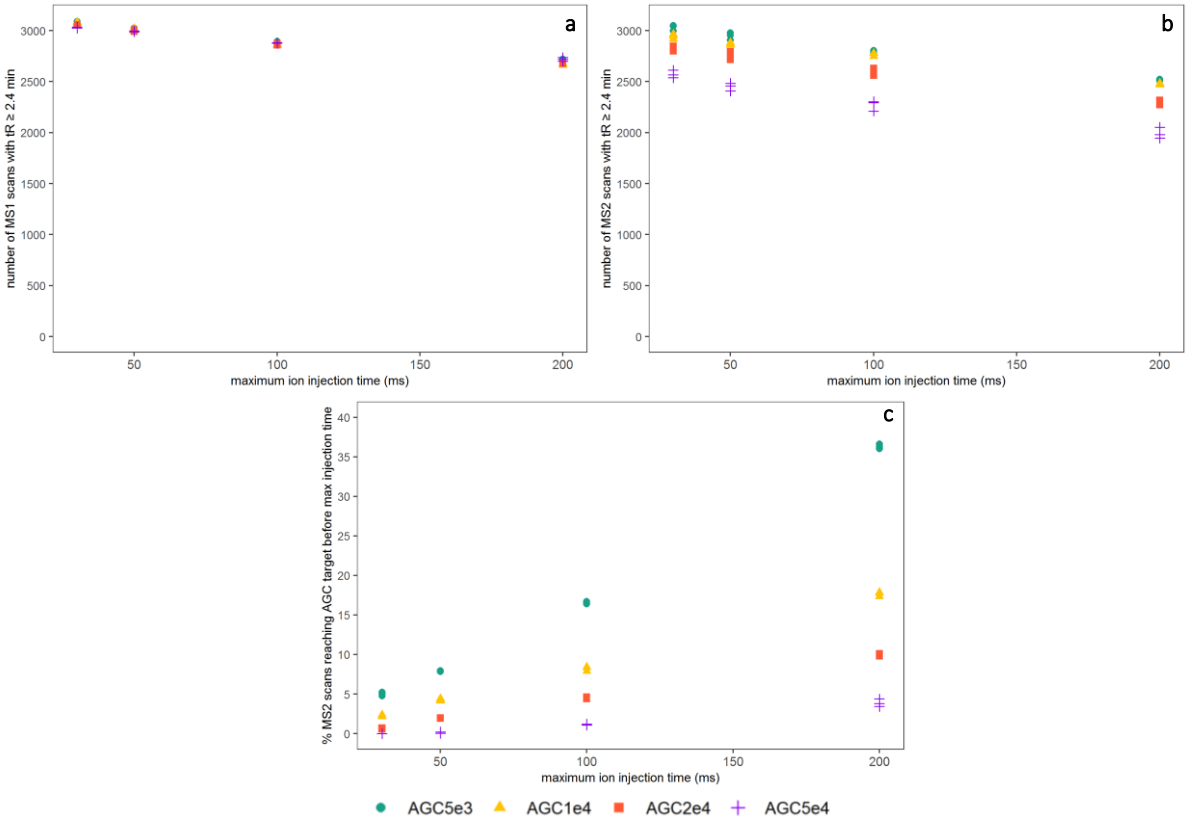

Figure S3.3.5 – a) number of MS1 scans taken per AGC target and maximum IT combination. The decrease in number of MS1 spectra at higher ITs can be explained by the mass spectrometer cycle time which is fixed at 0.9 s in the applied top speed method. The mass spectrometer records the optimum number of MS2 scans which could lead to time shifts in the MS1 full scans. b) number of MS2 scans taken. The decrease in MS2 spectra with increasing maximum IT and higher AGC target was expected. 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.
c) percentage of MS2 scans reaching the AGC target before the maximum IT.

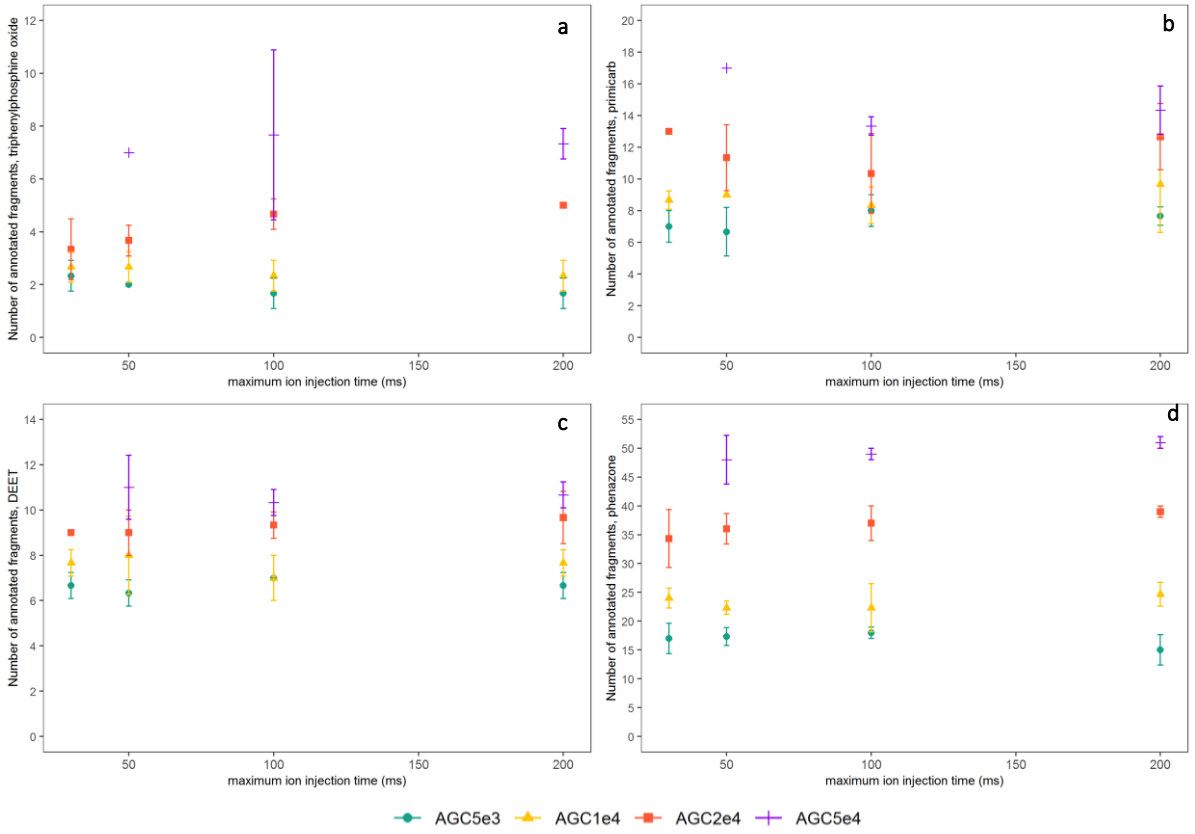

Figure S3.3.6 – Number of annotated fragments, AGC target experiments. a) triphenylphosphine oxide, b) primicarb, c) DEET, d) phenazone.

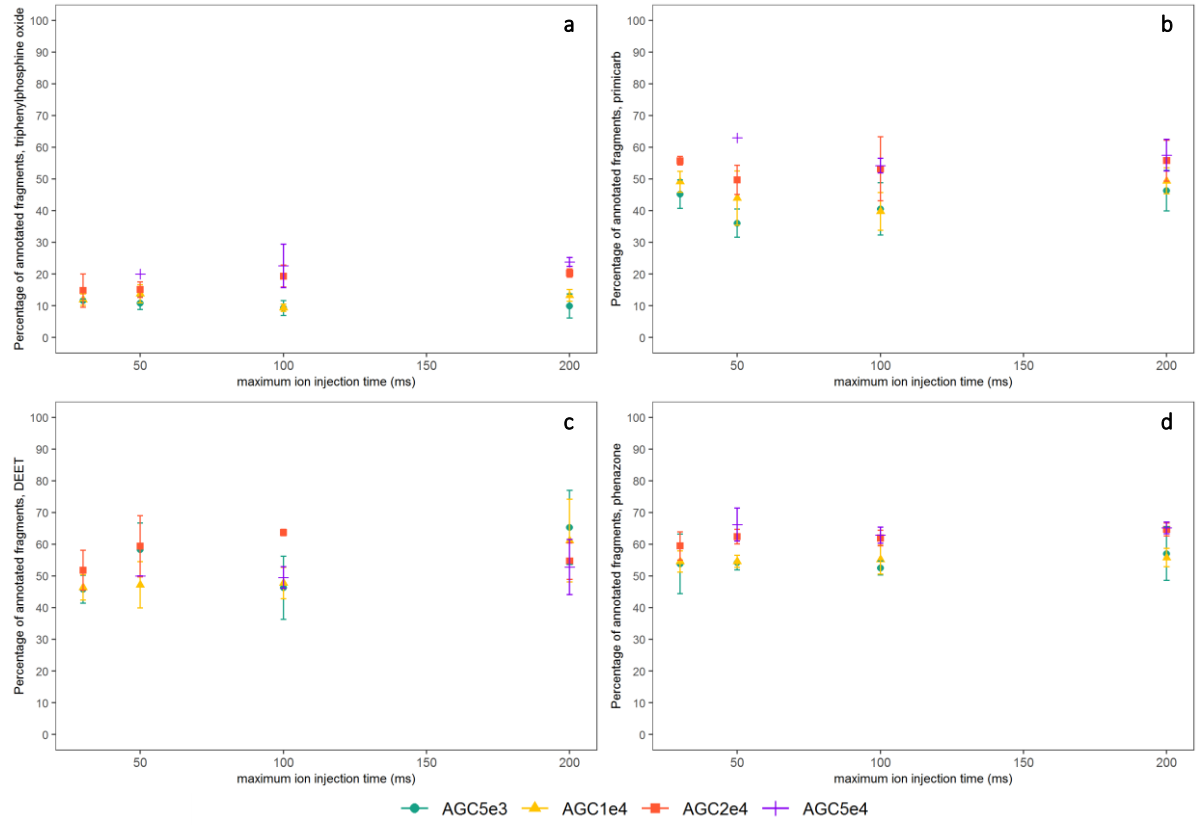

Figure S3.3.7 – Percentage of annotated peaks in the MS2 scan, AGC-target experiments. a) triphenylphosphine oxide, b) primicarb, c) DEET, d) phenazone.

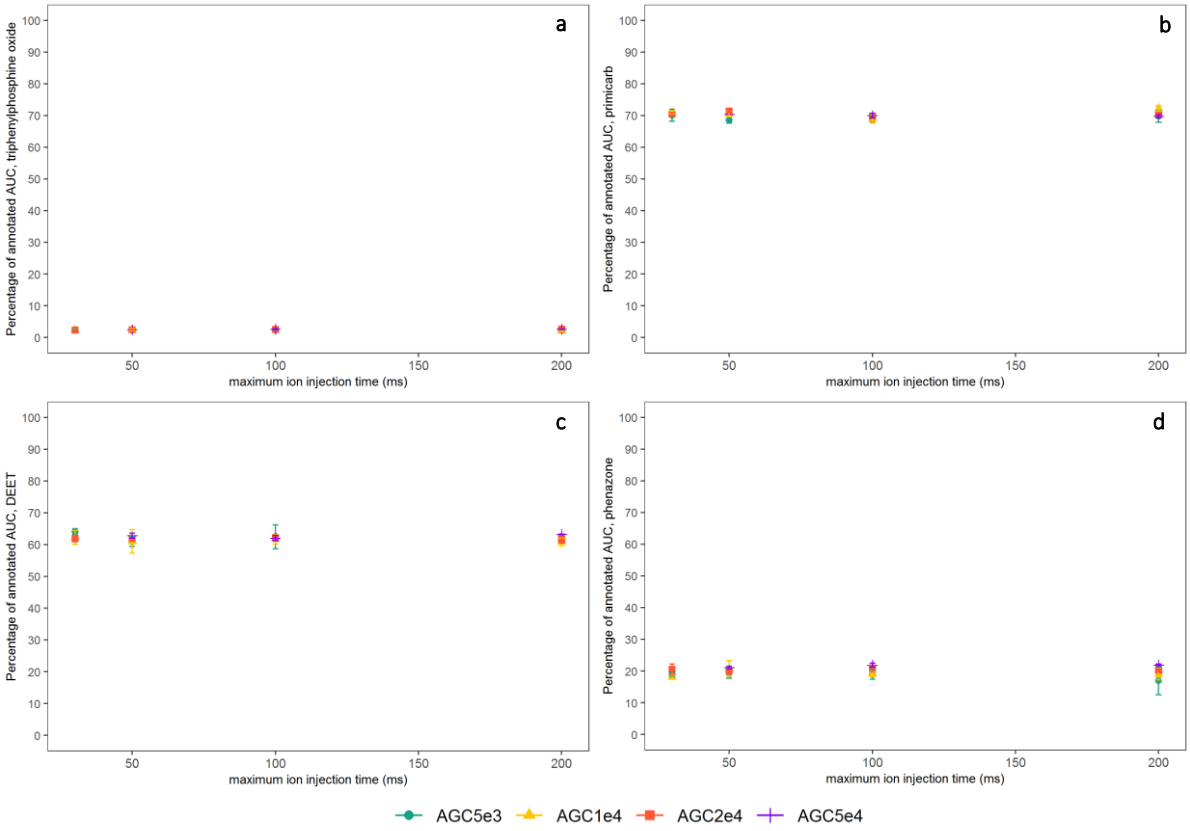

Figure S3.3.8 – Percentage of annotated peak area in the MS2 scan, AGC-target experiments. a) triphenylphosphine oxide, b) primicarb, c) DEET, d) phenazone.

S3.4 Supplementary Figures

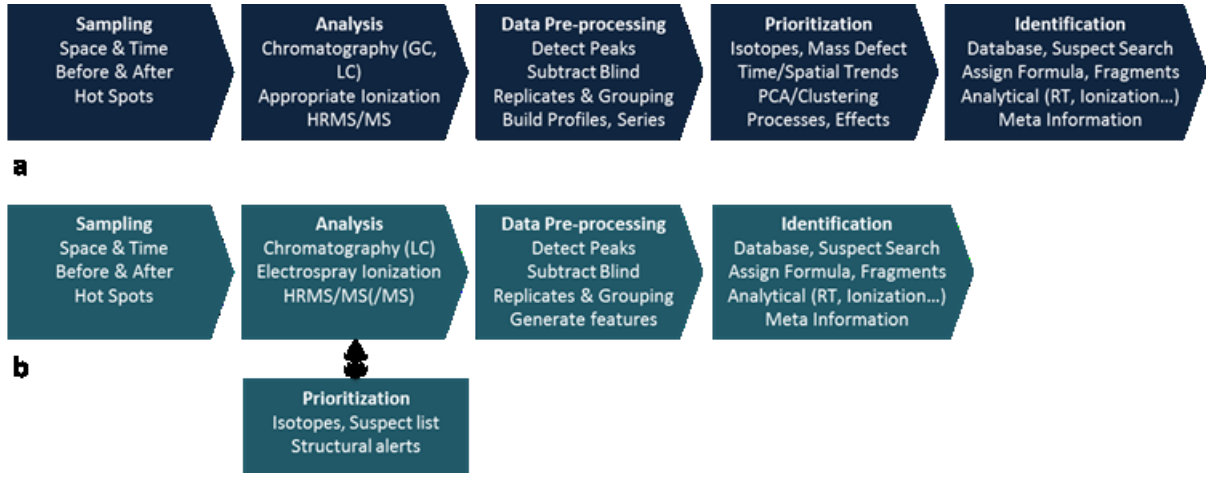

Figure S1 – a) typical NTS workflow from Hollender et al. (2019)[18], the prioritization step is performed offline. b) online prioritization that takes place during the data acquisition.

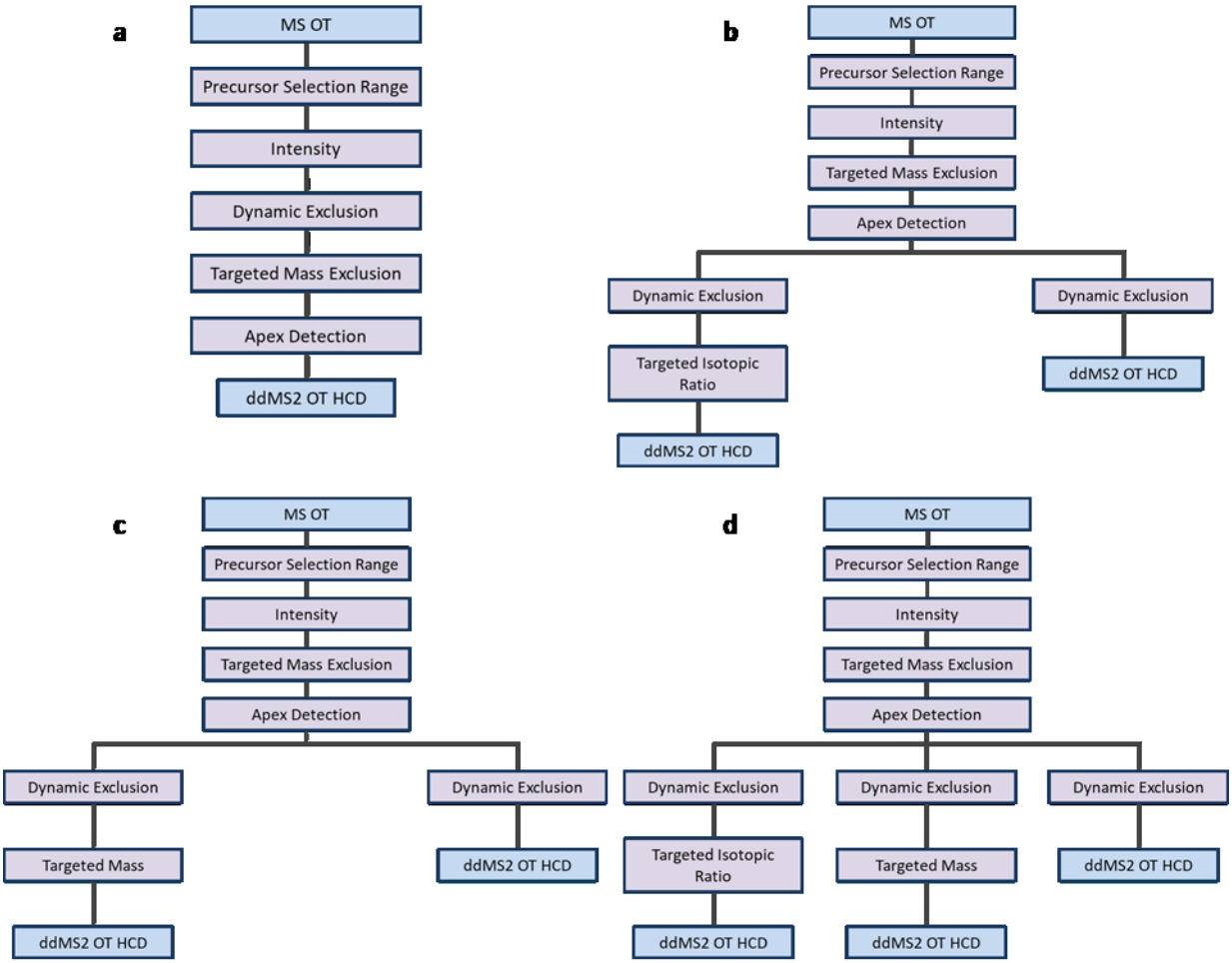

Figure S2 – Design of acquisition decision trees for MS1-trigger experiments, tree (a) is the regular KWR method, tree (b) includes the targeted isotopic ratio node, tree (c) includes the targeted mass node and tree (d) combines both. The inclusion lists were imported in the 'Targeted Mass' node, with 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.

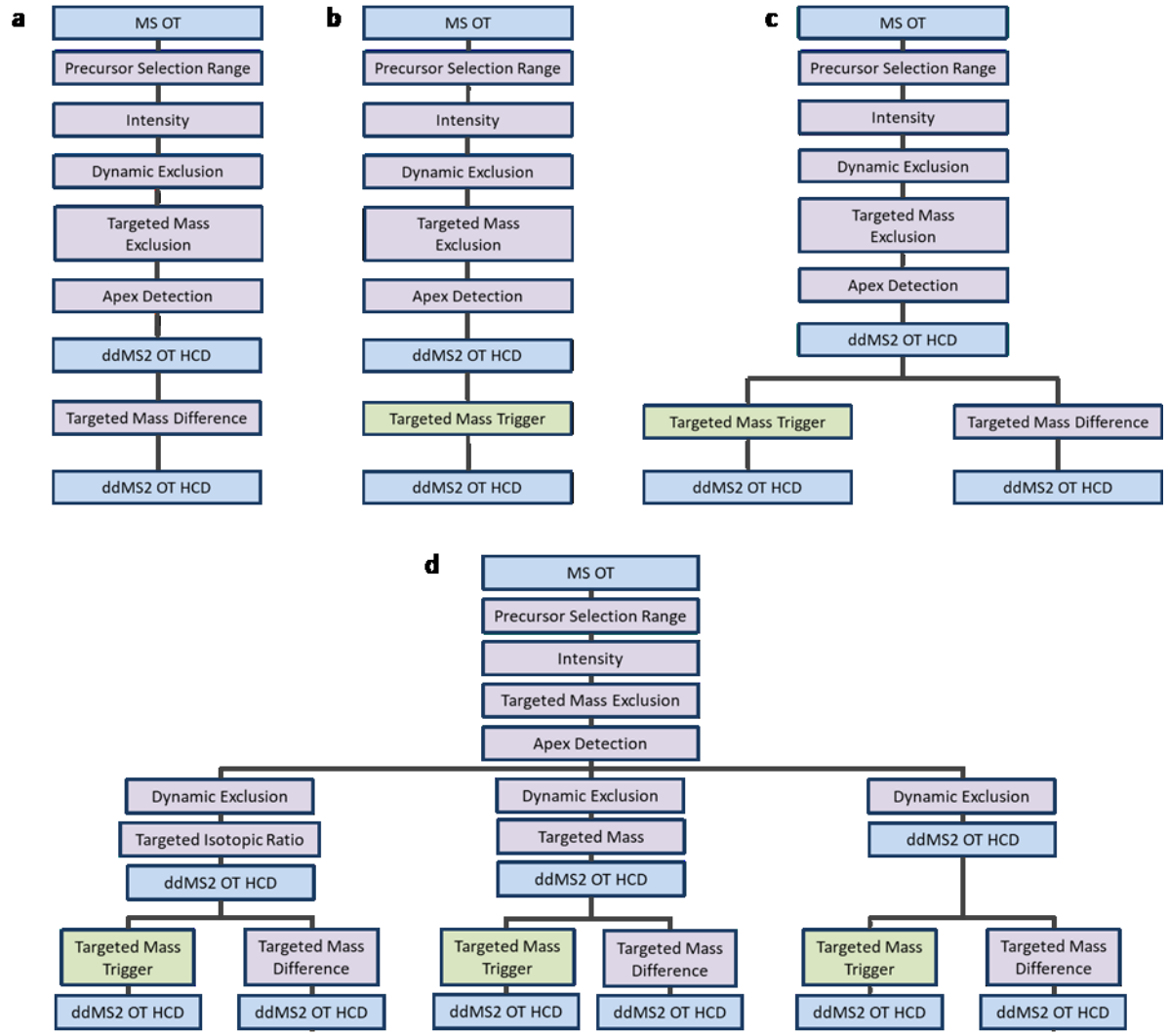

Figure S3 – Design of acquisition decision trees for MS2-trigger experiments, tree (a) corresponds to method 2 & 3, (b) corresponds to method 4 &5, tree (c) combines both and corresponds to method 6, 7 & 8, and tree (d) combines MS1- and MS2-triggers and corresponds to method 9, 10 and 11. The decision tree of the eventual intelligent acquisition method including MS1- and MS2-triggers corresponds to tree (d).

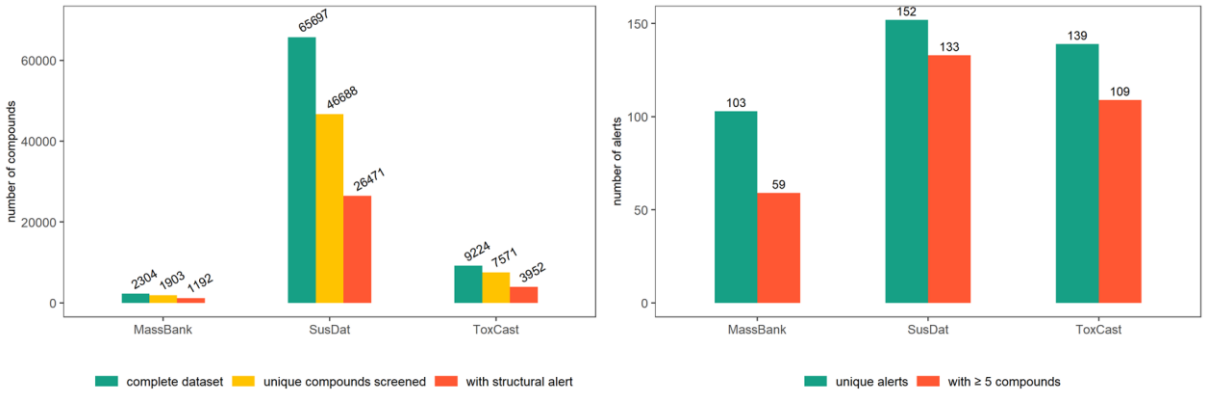

Figure S4 – Left: Number of compounds per dataset, the unique compounds screened and the compounds with a structural alert. Right: Schematic representation of the number of structural alerts present in the datasets.

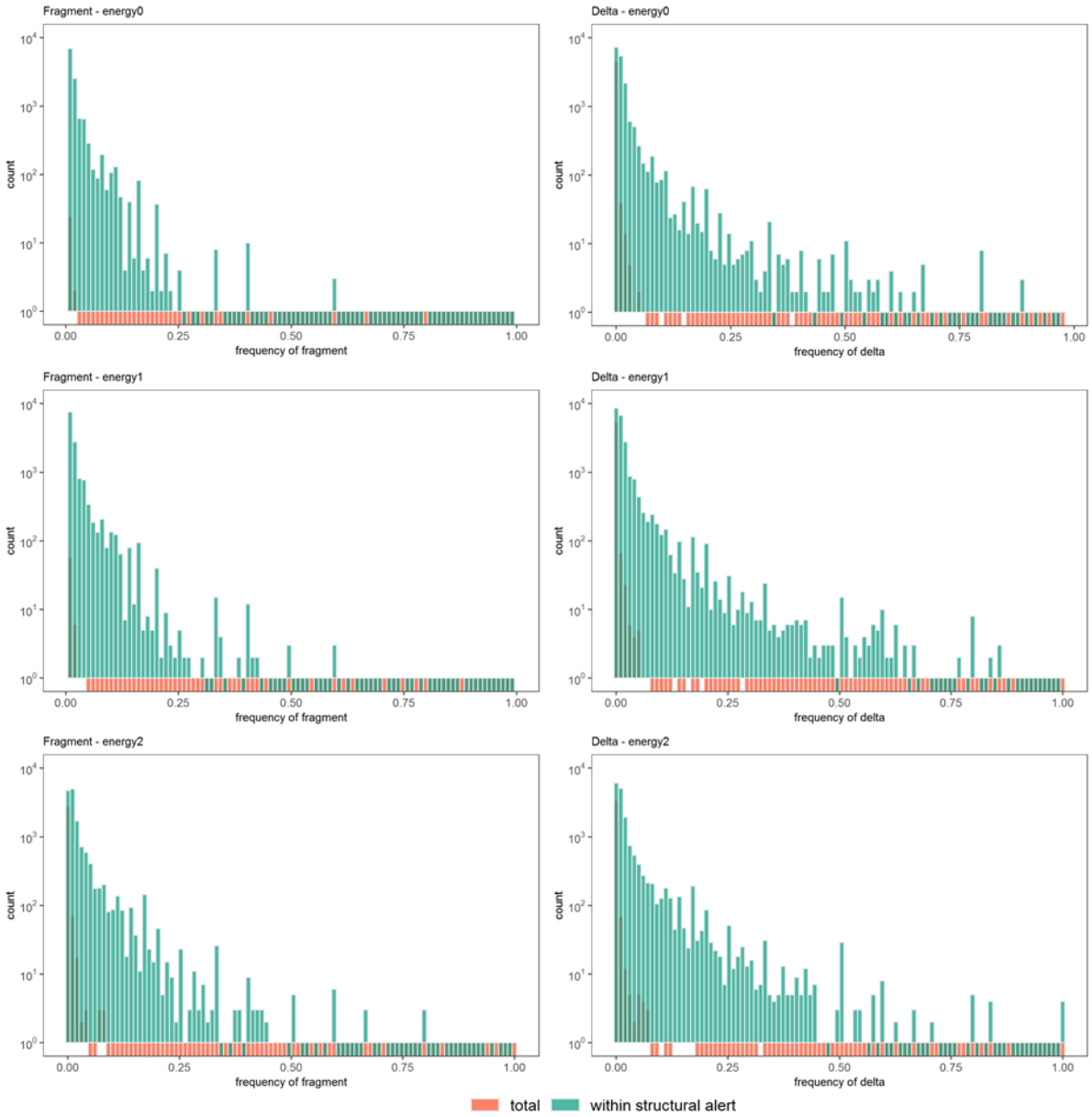

Figure S5 – Frequency distributions of recurring fragments (left) and recurring deltas (right) per CFM-ID energy. Note the logarithmic scale, the bars that look negative represent a count of 0, if no bar is visible, the count is 1.Within structural alerts (green bars) and within the total dataset (orange 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.

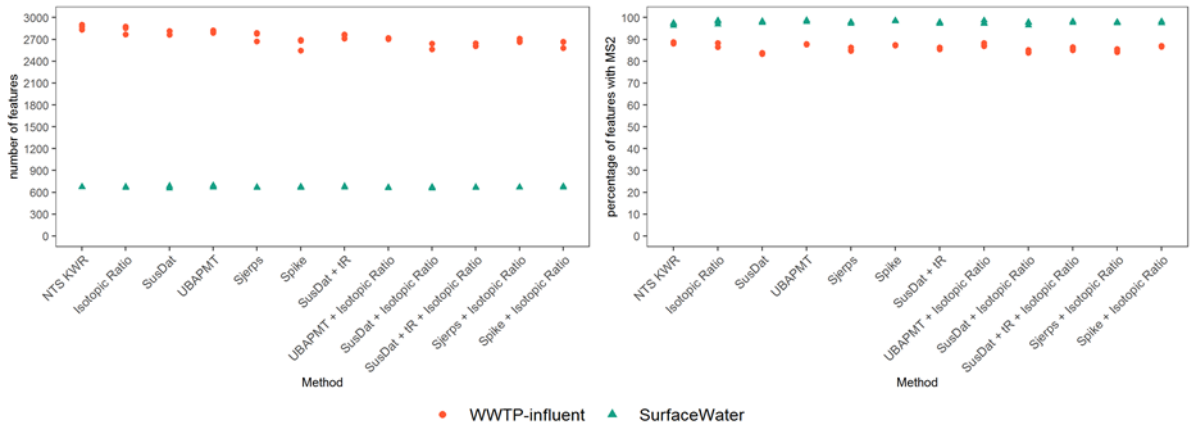

Figure S6 – Number of detected features (including background) per MS1-trigger method (left) and percentage of non-background features with MS2 (right).

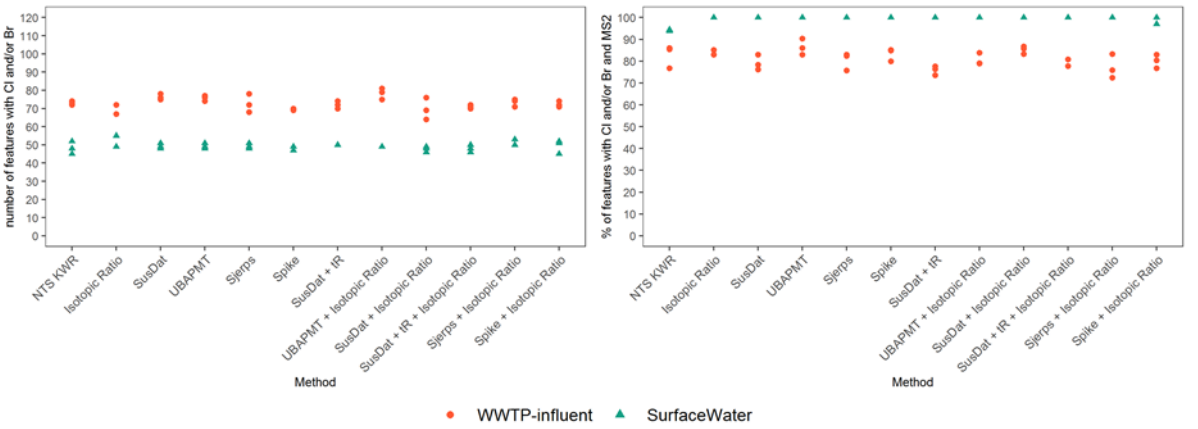

Figure S7 – Number of detected features with Cl and/or Br (including background) per MS1-trigger method (left) and percentage of non-background chlorinated and/or brominated features with MS2 (right). Presence of Cl and/or Br determined based on pattern match.

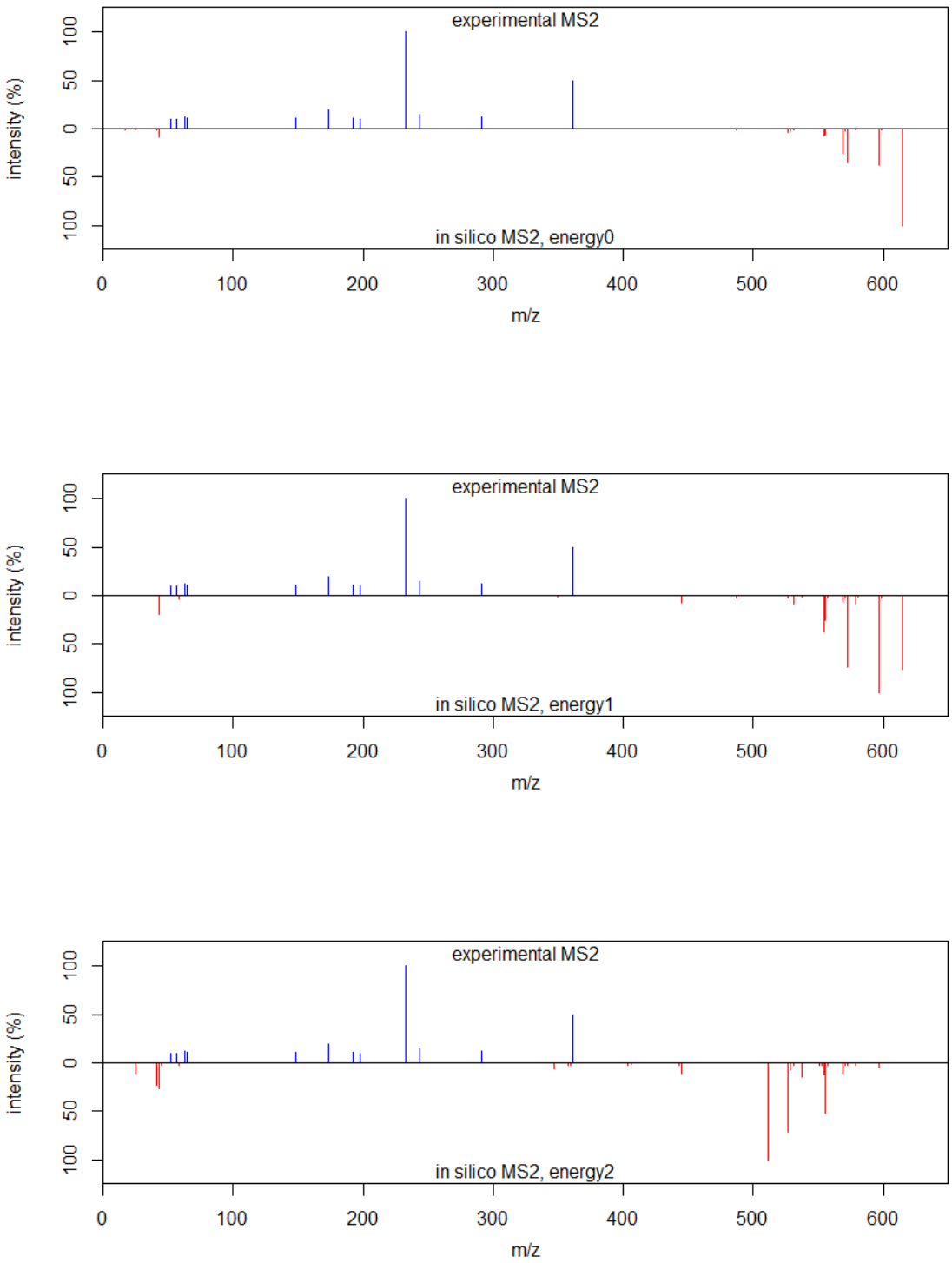

Figure S8– Experimentally obtained MS2 spectrum of diatrizoic acid and in silico *predicted MS2 spectra at three different energy levels.*[19]

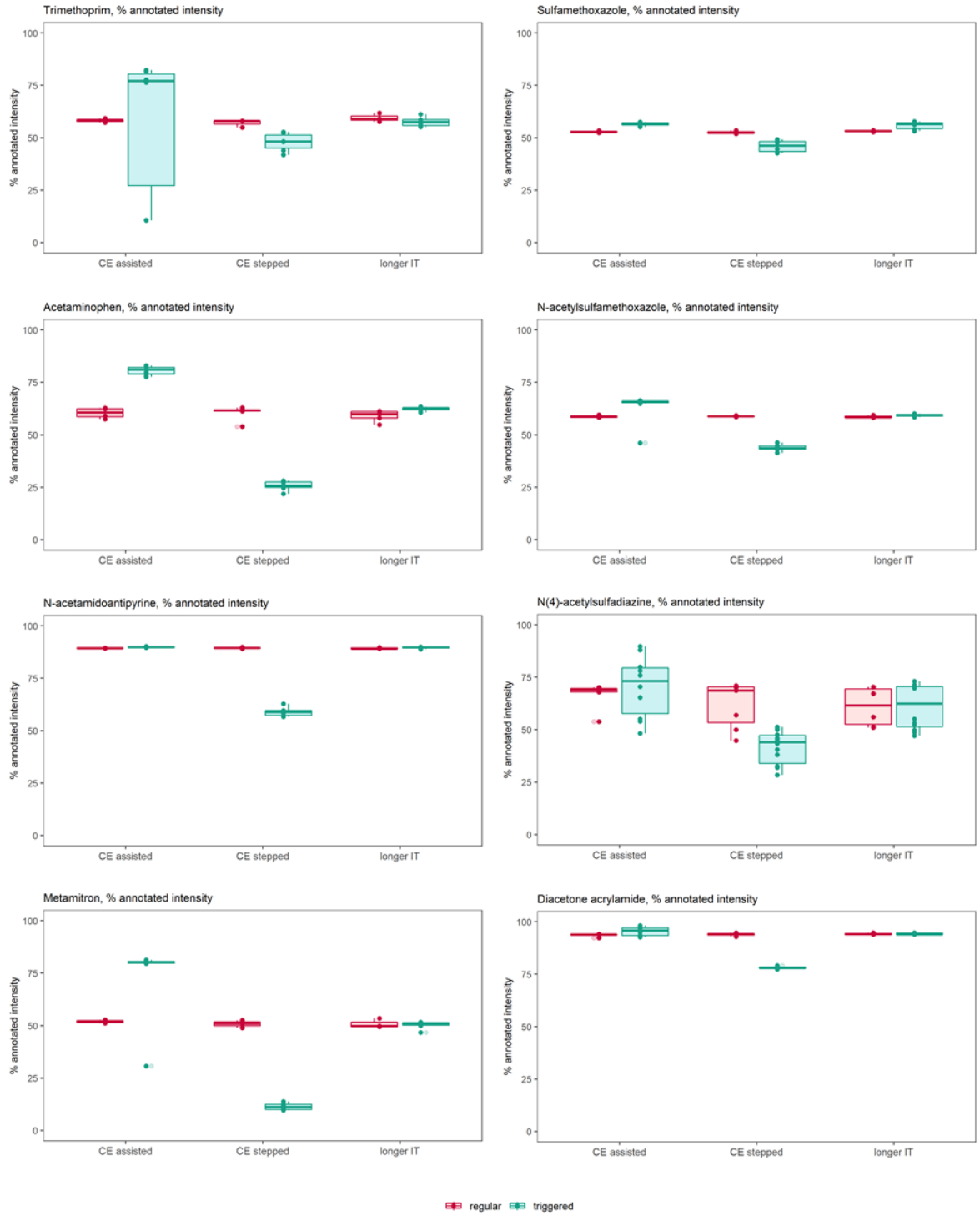

Figure S9 – Boxplots of the distribution of the percentages of annotated intensity of the peaks in the regular MS2 scan and the triggered MS2 scan of some spiked compounds with an alert.

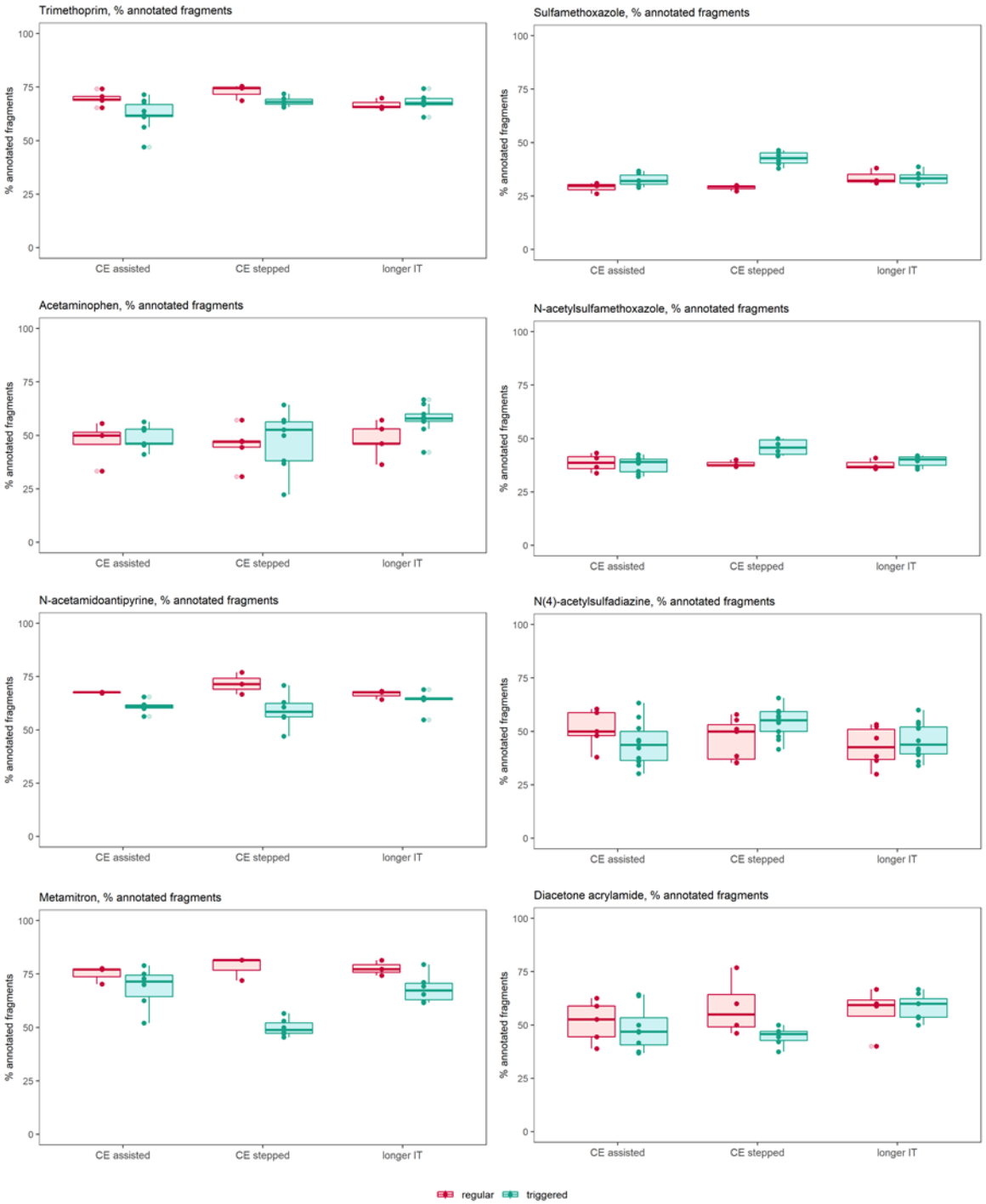

Figure S10 – Boxplots of the distribution of the percentages of annotated fragments in the regular MS2 scan and the triggered MS2 scan of some spiked compounds with an alert.

References

(1) 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, 2018. https://assets.thermofisher.com/TFS-Assets/CMD/posters/po-65258-lc-assisted-ce-asms2018-po65258-en.pdf (accessed 25 Nov 2019).
(2) Kalli, A.; Smith, G. T.; Sweredoski, M. J.; Hess, S. *J. Proteome Res.* **2013**, *12*, 3071-3086.
(3) Schymanski, E. L.; Schulze, T.; Alygizakis, N.; Meier, R., S1 | MASSBANK | NORMAN Compounds in MassBank. Version NORMAN-SLE-S1.0.1.1. Zenodo. http://doi.org/10.5281/zenodo.3247682
(4) HighChem LLC. mzCloud Features. https://www.mzcloud.org/Features (accessed Nov 25, 2019).
(5) Schulze, T.; Schymanski, E.; Stravs, M.; Neumann, S.; Krauss, M.; Singer, H.; Hug, C.; Gallampois, C.; Hollender, J.; Slobodnik, J.; Brack, W. In *NORMAN Bulletin*, 2012, pp 9-10.
(6) 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. http://doi.org/10.5281/zenodo.3548461
(7) Letzel, T.; Grosse, S.; Sengel, M., S2 | STOFFIDENT | HSWT/LfU STOFF-IDENT Database of Water-Relevant Substances. Version NORMAN-SLE-S2.0.1.2. Zenodo. http://doi.org/10.5281/zenodo.3900133
(8) Sjerps, R. M. A., S27 | KWRSJERPS2 | Extended Suspect List from Sjerps et al (KWRSJERPS). Version NORMAN-SLE-S27.0.1.1. Zenodo. http://doi.org/10.5281/zenodo.3542085
(9) Von der Ohe, P.; Fischer, S., S36 | UBAPMT | Potential Persistent, Mobile and Toxic (PMT) substances. Version NORMAN-SLE-S36.0.1.0. Zenodo. http://doi.org/10.5281/zenodo.2653213
(10) 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.
(11) 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.
(12) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. *J. Cheminf.* **2016**, *8*, 3.
(13) Dodder, N.; Mullen, K. *OrgMassSpecR: Organic Mass Spectrometry,* 0.5-3; 2017.
(14) Ruttkies, C.; Neumann, S.; Helmchen, A. *metfRag: Identification of metabolites using mass spectrometry data,* 2.4.2; 2017.
(15) Louisse, J.; Dingemans, M. M. L.; Baken, K. A.; van Wezel, A. P.; Schriks, M. *Chemosphere* **2018**, *209*, 373-380.
(16) Panse, C.; Sharma, S.; Huguet, R.; Vughs, D.; Grossmann, J.; Brunner, A. M. *Molecules* **2020**, *25*.
(17) Stravs, M. A.; Schymanski, E. L.; Singer, H. P.; Hollender, J. *J. Mass Spectrom.* **2013**, *48*, 89-99.
(18) Hollender, J.; van Bavel, B.; Dulio, V.; Farmen, E.; Furtmann, K.; Koschorreck, J.; Kunkel, U.; Krauss, M.; Munthe, J.; Schlabach, M.; Slobodnik, J.; Stroomberg, G.; Ternes, T.; Thomaidis, N. S.; Togola, A.; Tornero, V. *Environ. Sci. Eur.* **2019**, *31*.
(19) Allen F.; Greiner R.; D., W. CFM-ID 2.0 Spectrum Prediction. http://cfmid2.wishartlab.com/predict (accessed 8 Sep 2020).