BTO 2019.032 | May 2019

BTO report

Non-target screening to identify unknowns: Automation and increasing confidence

KWR Watercycle Research Institute

BTO

Non-target screening to identify unknowns: Automation and increasing confidence

BTO 2019.032 | May 2019

Project number
402045/049/001

Project manager
Stefan Kools

Client
BTO - Thematical research - New monitoring concepts

Quality Assurance
Pim de Voogt

Author(s)
Andrea Mizzi Brunner, Dennis Vughs,
Rick Helmus (UvA)

Sent to

Year of publishing
2019

More information
Dr. Andrea Mizzi Brunner
T 030 60 69 564
E andrea.brunner@kwrwater.nl

Keywords
Non-target screening, identification, chemical water quality, cheminformatics

Postbus 1072
3430 BB Nieuwegein
The Netherlands

T +31 (0)30 60 69 511
F +31 (0)30 60 61 165
E info@kwrwater.nl
I www.kwrwater.nl

Watercycle
Research
Institute

BTO 2019.032 | May 2019 © KWR

Year of publishing
2019

More information
Dr. Andrea Mizzi Brunner
T 030 60 69 564
E andrea.brunner@kwrwater.nl

Keywords
Non-target screening, identification, chemical water quality, cheminformatics

Postbus 1072
3430 BB Nieuwegein
The Netherlands

T +31 (0)30 60 69 511
F +31 (0)30 60 61 165
E info@kwrwater.nl
I www.kwrwater.nl

Watercycle
Research
Institute

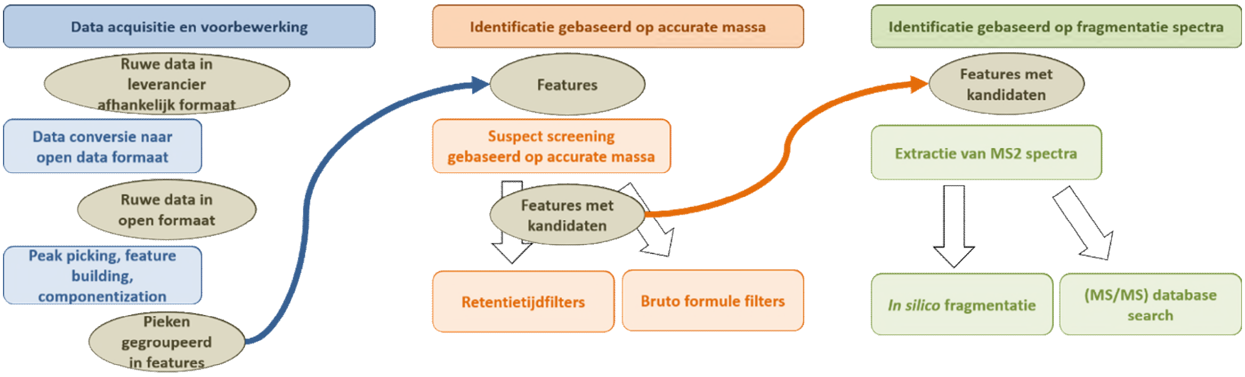
BTO 2019.032 | May 2019 © KWR

BTO *Managementsamenvatting*

Non-target screening: geautomatiseerde workflows voor identificatie van onbekende stoffen met hoge betrouwbaarheid

Auteur(s) Dr. Andrea Mizzi Brunner, Dennis Vughs MSc, Rick Helmus MSc

De betrouwbare identificatie van een onbekende microverontreiniging in water is essentieel voor de risicobeoordeling en het voorspellen van het gedrag van de verontreiniging in het milieu en in de drinkwaterzuivering. Om onbekende microverontreinigingen sneller en met een hogere betrouwbaarheid te kunnen identificeren, zijn twee geautomatiseerde workflows ontwikkeld voor non-target screening data-analyse; een is gebaseerd op het open source software package *patRoon*, de andere op de commerciële software Compound Discoverer. De twee workflows zijn vervolgens gebruikt om data van KWR en de drinkwaterbedrijven te analyseren. In een *hands-on* data-analyseworkshop bij KWR hebben medewerkers van drinkwaterlaboratoria de workflows met succes toegepast.


Data acquisitie en voorbewerking
Ruwe data in leverancier afhankelijk formaat
Data conversie naar open data formaat
Ruwe data in open formaat
Peak picking, feature building, componentization
Pieken gegroupeerd in features
Identificatie gebaseerd op accurate massa
Features
Suspect screening gebaseerd op accurate massa
Features met kandidaten
Retentietijdfilters
Bruto formule filters
Identificatie gebaseerd op fragmentatie spectra
Features met kandidaten
Extractie van MS2 spectra
In silico fragmentatie
(MS/MS) database search

Geautomatiseerde workflows om non-target screening data te analyseren combineren tools om onbekende stoffen te identificeren en de betrouwbaarheid van hun identificatie te verhogen

Belang: tijd besparen bij betrouwbare identificatie van onbekende microverontreinigingen

In 2005 vormde de introductie van hoge-resolutie massaspectrometrie een doorbraak in het onderzoek naar de aanwezigheid van organische microverontreinigingen in water. Door te techniek te combineren met vloeistofchromatografie werd het mogelijk te screenen naar een breed palet van stoffen in lage concentraties (ng/L range), een onderzoeksmethode die nu bekend staat als non-target screening. Door middel van suspect screening kan in bestaande non-target screening-data naar specifieke kandidaatstoffen worden gezocht. Uit het in 2017 uitgevoerde BTO-project *Massaspectrometrie: tools voor ID van onbekenden* werd duidelijk dat interpretatie van non-target data een tijdrovend proces is (veel handmatig werk) dat veel expertise van de onderzoeker vereist. Daarnaast is in 2016 gestart met de onderbouwing voor een wettelijke norm van brede screening. De bevindingen van beide activiteiten leidden tot de vraag om een verder gestroomlijnde interpretatie van non-target data en identificatie met hoge betrouwbaarheid.

Aanpak: tools gecombineerd in geautomatiseerde workflow voor data-analyse en identificatie

De identificatie in non-target screening gebeurt op basis van gegevens uit databanken, waaronder exacte massa, isotooppatroon, MS2-fragmentatiepatroon en metadata. Wanneer kandidaatstoffen in de beschikbare databanken

ontbreken – zoals vaak het geval bij bv. transformatieproducten - dan moeten ze handmatig worden geïdentificeerd. Afhankelijk van hoe ondubbelzinnig de identificatie is, wordt de geïdentificeerde structuur voorzien van een betrouwbaarheidsniveau. Om het identificatieproces te vereenvoudigen en te versnellen en de betrouwbaarheid van identificatie te vergroten is de eerder ontwikkelde handmatige workflow geautomatiseerd. Hiervoor werden twee softwarepakketten geëvalueerd: de commerciële software *Compound Discoverer* (alleen bruikbaar met Thermo Fisher Scientific, Orbitrap-gegevens, bètatest) en het open source softwarepakket patRoon (UvA, bruikbaar met data van alle instrumenttypen, in nauwe samenwerking met de ontwikkelaar, promovendus Rick Helmus). De bruikbaarheid van de twee geautomatiseerde identificatie-workflows werd getest door het analyseren van non-target screening data van KWR (geselecteerde voorbeelden van het DPWE-project *Robuustheid zuiveringen*) en data van twee drinkwaterlaboratoria (ringonderzoekmonsters van HWL en watermonsters voor en na behandeling van De Watergroep). Elke workflowstap van de non-target-screening data-analyse werd beoordeeld, van dataverwerking tot identificatie op basis van de accurate massa (MS1) en MS2-gebaseerde identificatie door matching met fragmentatiespectra uit bibliotheken en uit *in silico* voorspellingen.
Beide workflows zijn daarnaast met praktijkmensen beproefd tijdens een hands-on workshop bij KWR.

Resultaten: twee geautomatiseerde workflows voor non-target screening van data uit watermonsters

De workflow met de commerciële software *Compound Discoverer* (alleen bruikbaar met Thermo Fisher Scientific, Orbitrap-gegevens, bètatest) en de workflow met het open source softwarepakket patRoon (UvA, bruikbaar met data van alle instrumenttypen en vereist kennis van programmeertaal R) konden beide de suspects betrouwbaar identificeren en werden door de deelnemers aan de workshop met succes ingezet. Medewerkers van alle vier de drinkwaterlaboratoria konden tijdens het praktische gedeelte op hun eigen laptops de benchmarkresultaten reproduceren. Daartoe werden softwarepakketten geïnstalleerd en workflows uitgevoerd. Vragen konden tijdens de workshop onmiddellijk worden opgepakt.

Implementatie: beide workflows bruikbaar voor drinkwaterlaboratoria

Tijdens de workshop bleek dat beide geautomatiseerde workflows in de praktijk bruikbaar zijn voor drinkwaterlaboratoria, mits het dataformaat het toelaat (i.e. de workflow met de commerciële software *Compound Discoverer* is alleen bruikbaar met Thermo Fisher gegevens). Ook werd er getoond hoe ze kunnen worden geïmplementeerd.

Rapport

Dit onderzoek is beschreven in het rapport *Non-target screening to identify unknowns: Automation and increasing confidence* (BTO 2019.032).

More information
Dr. Andrea Mizzi Brunner
T 030 60 69 564
E andrea.brunner@kwrwater.nl

KWR
PO Box 1072
3430 BB Nieuwegein
The Netherlands

KWR Watercycle Research Institute

Contents

1 Overview of the project

1.1 What happened previously

The reliable identification of an unknown micro-pollutant in water is not only essential to good (human) risk assessment, it is also necessary to predict the behavior of a substance in the environment and in drinking water treatment. In 2005 a breakthrough took place in the investigation of the presence of organic micro-pollutants in water with the introduction of high resolution mass spectrometry (HRMS). Combined with liquid chromatography (LC), screening is thus carried out at low concentration levels (ng / L range) and for a wide range of substances, typically referred to as LC-HRMS based non-target screening (NTS).

Suspect screening can be applied to NTS data to screen for candidate substances which are suspected and/or expected to be present in a sample. This is done on the basis of data from databases, including exact mass, isotope pattern, MS2 fragmentation pattern and metadata (McEachran et al., 2017; Schymanski and Williams, 2017). If a substance is missing in available databases - as is often the case with transformation products - then manual identification is required. Depending on how certain the identification, the identified structure is provided with a defined level of confidence (Schymanski et al., 2014). In the BTO project "Mass Spectrometry: tools for unknown IDs", a workflow was developed describing these steps (BTO 2017.073). From that project it became clear that NTS data interpretation is a time- and labor- intensive process that requires a lot of expertise and manual work from the researcher (see document TG NMS 15-04-06). In addition, a trajectory towards a Dutch technical agreement to eventually substantiate a legal standard for NTS screening was started, which requires streamlined non-target data interpretation and identification with high reliability.

1.2 From software testing at KWR to implementation at the drinking water laboratories

In the present project “Non-target screening to identify unknowns: Automation and increasing confidence” we aimed at automating the manual workflow of the earlier BTO project in order to simplify and speed up the identification process, and increase the confidence of identification. For this purpose, two software packages were evaluated, namely the commercial software *Compound Discoverer* (Thermo Fisher Scientific, Orbitrap data only) as Beta tester for version 3.0, and the open source software package patRoon (UvA, data from all instrument types) in tight collaboration with its developer Rick Helmus (PhD student of Pim de Voogt and Thomas ter Laak). Each step in the workflow for the NTS data analysis was assessed, for example data treatment, identification based on the accurate mass (MS1), and MS2-based identification by searching in libraries of fragmentation spectra and '*in silico*' fragmentation databases.

The usability of the two automated identification workflows was tested by analyzing non-target screening data from KWR (selected samples from the DPWE robuustheid zuiveringen project) and non-target data from two drinking water laboratories, HWL (Round robin ground- and surface water samples) and De Watergroep (before and after treatment water samples).

As the ultimate goal of the project was the implementation of the developed workflow(s) at the drinking water laboratories, personal contact and consultancy during the whole project was key to success. KWR was visited by Nikki Janssens (De Watergroep), Mark van Huijkelom (Brabant Water), and Eelco Pieke (HWL) individually to discuss the latest strategies and progress made at KWR in the field of NTS data analysis and identification of unknowns. Milan Verwoert (WLN) spent one week at KWR and was introduced to the workflows KWR applies to this end. In turn, Andrea Brunner visited WLN and HWL to get acquainted with their instrumentation and analyses.

The evaluated software packages and the developed tailored workflows for the analysis of Orbitrap (Thermo Fisher) and QTOF (SCIEX and Bruker) data of water samples were presented during a NTS hands-on data analysis workshop at KWR on March 28 (Figure 1). The workshop was attended by members from all four BTO drinking water laboratories; Annemarie Toebak and Runa Kooper-Mookerji (AquaLab Zuid), Nikki Janssens (De Watergroep), Rob ten Broek and Eelco Pieke (HWL), and Jan van der Kooi and Milan Verwoert (WLN). The workshop started with presentations on NTS theory (Andrea Brunner), and an introduction to the two software packages *Compound Discoverer* 3.0 (Dennis Vughs) and patRoon (Rick Helmus). During the practical part of the workshop, participants were able to reproduce the benchmarking results of one data set of choice on their own laptops. To this end, software packages were installed, workflows executed and questions that arose could be tackled immediately.

FIGURE 1. HANDS-ON NTS DATA ANALYSIS WORKSHOP AT KWR.

1.3 This report

Short theoretical explanation of the limitations of the available software when the project started, as well as challenges and goals constitute Chapter 2. The commercial software *Compound Discoverer* is introduced in Chapter 3, and the open-source software patRoon in Chapter 4. These chapters both include instructions on software installation, an extended user manual and the results of the NTS data analysis of samples from KWR (Orbitrap Fusion, Thermo Fisher Scientific, selected samples from the DPWE robuustheid zuiveringen project, data analysed with *Compound Discoverer* 3.0 and patRoon) and from HWL (QTOF, Bruker, Round robin samples, data analysed with patRoon). Developments in the field of NTS-based identification of unknown substances are going fast. At the time of writing this report is almost already outdated. It was therefore crucial to keep well-informed of new developments in NTS data analysis and cheminformatics during the course of the project, including participation in mass spectrometry conferences and meetings. The abstracts and presentations of events that were (partly) funded by this project can be found in Chapter 5.

1.4 Remaining challenges and outlook

Despite the progress in NTS-based identification of unknowns, 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) and the fragmentation spectra (MS2) of a given unknown peak with those of chemical and spectral database entries such as Chemspider and mzCloud, respectively. Identification is particularly challenging for compounds with generic elemental formulas, for which many thousands of candidate substances are present, and compounds with poor fragmentation spectra or a lack thereof.

1.4.1 All ion fragmentation

A lack of fragmentation can be due to low signal intensities of the compound, as only the *n* most intense peaks are fragmented in the typical NTS method referred to as data dependent method. Data independent analysis (DIA) approaches also referred to as AIF, SWATH or MSe (depending on the instrument vendor) can circumvent this issue as every eluting peak is fragmented without discrimination or pre-selection as the instrument performs a repeating cycle of acquisitions over a set of fixed mass ranges or the full MS range (Bonner and Hopfgartner, 2018). Thereby, exact mass data for every detectable compound and its sub-structure is acquired, which can be re-interrogated at any time. However, low intensity precursors will always lead to low fragment ion intensities. Confident identification remains challenging.

1.4.2 High spectral complexity

Alternatively, a lack of fragmentation can be due to the high complexity of MS1 spectra. Multiple peaks can belong to the same compound, i.e. in source fragments, adducts, and dimers, as well as background signals. These can hinder identification when they lead to a redundancy in MS2 spectra from the same (background) compound. Prioritizing compounds of interest for and excluding background compounds from fragmentation could alleviate this. Part 2 of the BTO project “Improved non-target screening based identification through MS online prioritization” is addressing the issue of spectral complexity and redundancy. It aims at improving structural identification of organic micro-pollutants in water samples by developing intelligent MS acquisition methods that result in more informative MS spectra.

1.4.3 Novel fragmentation methods: UVPD and IR

Poor fragmentation spectra can be the result of suboptimal fragmentation methods. In that case, alternative fragmentation techniques can aid structural elucidation. Part 3 of the BTO project “Improved non-target screening based identification through MS online prioritization” is evaluating the potential of the alternative fragmentation technique ultraviolet photodissociation (UVPD) for confident identification of organic micro-pollutants. UVPD is a relatively new fragmentation technique achieved with a 213 nm UV laser, and potentially allows structural elucidation of compounds that cannot be identified by Higher collision induced dissociation (HCD) alone (Brodbelt, 2014). For instance, UVPD was shown to facilitate characterization of various lipid classes (Morrison et al., 2016), to generate unique fragments or enhance detection of kinetically unfavorable fragments of flavonoids, phenylpropanoids and chalconoids (Huguet, et al. 2016, Huguet et al. 2017).

As a second alternative fragmentation technique, we are assessing the potential of infrared ion spectroscopy (IRIS) in combination with MS for structural identification in collaboration with Jos Oomens and his group at FELIX laboratories, Radboud University (Martens et al., 2017; Martens et al., 2018). IRIS combines mass spectrometry (MS) and IR spectroscopy so that a vibrational spectrum can be recorded for an individual compound, for instance for an unknown prioritized features after fractionation of the sample with HPLC.

1.4.4 Unknown unknowns

IRIS is a particularly interesting alternative as it can reveal information on the substructure of the unknown. Thereby, it could contribute to the identification of unknown unknowns, i.e. compounds that are not listed in chemical databases, such as transformation products which to date are rarely identified. The currently running BTO project “Monitoring transformation product formation in drinking water treatment” is addressing this problem by developing a NTS-based data analysis workflow for the structural identification of transformation products. Therefore, a new strategy described by Schollee et al (Schollee et al., 2017) that is based on the structural and spectral similarity of parent compounds and their transformation products is automated and added to the workflow. Moreover, various prediction tools are evaluated and if beneficial, they will be included into the next patRoon versions. Alternatively, *Compound Discoverer* allows for similarity scoring using mzCloud and mzLogic, and the Compound Classes node.

1.4.5 NMR as a last resort

If all described approaches fail, Nuclear Magnetic Resonance (NMR) can be added to the identification process. The identification success rate mainly depends on the purity of the compound in the sample. Therefore, the sample must be concentrated in most cases, and purified to obtain the (almost) pure compound, for instant through an SPE extraction followed by preparative HPLC fractionation. However, these steps are labor intensive, and the NMR analysis must be outsourced to a specialist laboratory. Identification with NMR is therefore time-consuming and costly.

Developments in the field of NTS, in particular identification tools, software and online databases, are progressing at a high pace. It is therefore recommended to follow these developments closely, and to take an inventory at least every two years which will allow for the according adjustments of the developed NTS data analysis workflows.

2 Steps of non-target data analysis

2.1 The goal of NTS data analysis steps is confident identification

The bottleneck in LC-HRMS based non-target screening is the confident identification of the structure of an unknown feature. A feature represents a given compound and consists of a unique combination of an accurate mass and a retention time. To standardize the use of terms in identification, the so called Schymanski Levels of confidence were introduced in 2014 (Schymanski et al., 2014) that define the confidence of an identification based on available MS information (see Figure 2). In this definition, Level 5 constitutes any feature detected with HRMS and thus with an exact mass. To reach level 4 an unequivocal formula need to be attributed to the feature. This can be based on the isotopic pattern of the peak and adducts. Without MS2 fragmentation data, no level higher than level 4 can be reached. At level 3, the feature represents tentative candidate(s) that match the MS1 accurate mass and the MS2 fragmentation spectra, however, information is insufficient for one exact structure only. To increase confidence to level 2, additionally the probable structure has to be confirmed by a library spectrum match or diagnostic evidence, such as diagnostic MS/MS fragments and/or ionization behavior, parent compound information and the experimental context. Once the identity of a feature is confirmed by comparison with a reference standard, it reaches level 1. At level 1 the goal of NTS data analysis is reached, the feature is confidently identified.

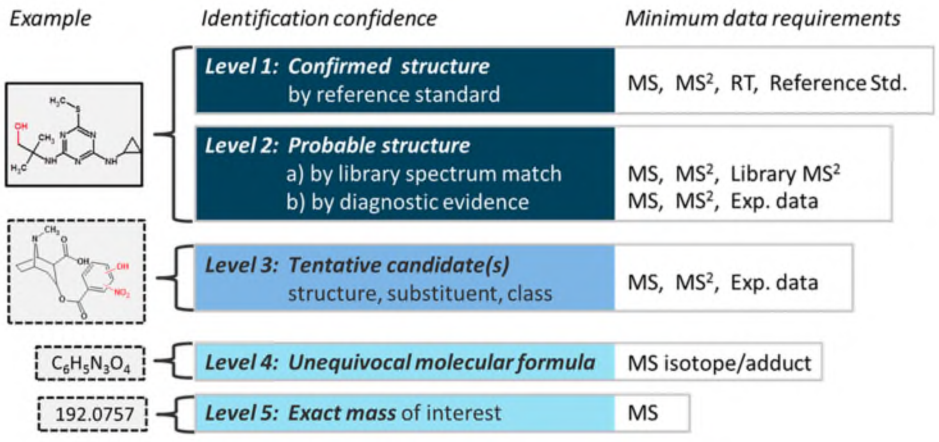

Example
Identification confidence
Minimum data requirements
Level 1: Confirmed structure
by reference standard
MS, MS^2, RT, Reference Std.
Level 2: Probable structure
a) by library spectrum match
b) by diagnostic evidence
MS, MS^2, Library MS^2
MS, MS^2, Exp. data
Level 3: Tentative candidate(s)
structure, substituent, class
MS, MS^2, Exp. data
$C_6H_5N_3O_4$
Level 4: Unequivocal molecular formula
MS isotope/adduct
192.0757
Level 5: Exact mass of interest
MS

FIGURE 2. IDENTIFICATION CONFIDENCE LEVELS PROPOSED BY SCHYMANSKI ET AL. REPRODUCED FROM (SCHYMANSKI ET AL., 2014).

2.2 Outline of the generic NTS workflow

2.2.1 Data acquisition and curation

Essentially, all NTS data analysis workflows comprise steps that lead from level 5 to higher levels of confidence, preferably level 2. In brief, the various workflow steps can be summarized as part of data acquisition and curation, MS1 based identification and MS2 based identification (see Figure 3). In the data acquisition and curation steps, LC-HRMS data is acquired in data dependent acquisition mode and if necessary converted to a data format compatible with subsequent processing steps. Then, chromatographic peaks are detected, referred to as peak picking (Alonso et al., 2011; Zhou et al., 2012). The resulting features represent a certain compound and consist of a unique combination of an accurate mass and a retention time. Some workflows comprise a componentization step that groups degenerate peaks of the same compound, i.e. isotopes, adducts and in-source fragments into one feature (Kuhl et al., 2012; Broeckling et al., 2014; Mahieu et al., 2016; Domingo-Almenara et al., 2018). Subsequently, in the feature building step the features from individual samples are grouped in order to allow comparison between samples. The end output of the data acquisition and curation step are features with a confidence level 5, i.e. exact masses of interest.

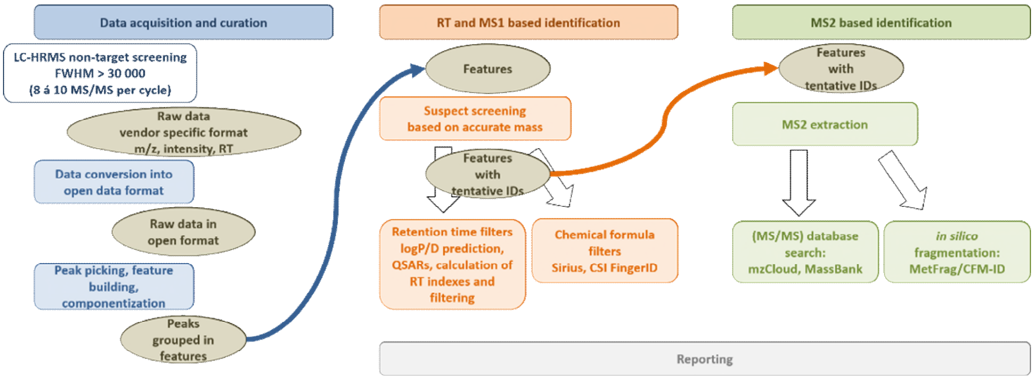

Data acquisition and curation
LC-HRMS non-target screening FWHM > 30 000 (8 á 10 MS/MS per cycle)
Raw data vendor specific format m/z, intensity, RT
Data conversion into open data format
Raw data in open format
Peak picking, feature building, componentization
Peaks grouped in features
RT and MS1 based identification
Features
Suspect screening based on accurate mass
Features with tentative IDs
Retention time filters logP/D prediction, QSARs, calculation of RT indexes and filtering
Chemical formula filters Sirius, CSI FingerID
MS2 based identification
Features with tentative IDs
MS2 extraction
(MS/MS) database search: mzCloud, MassBank
in silico fragmentation: MetFrag/CFM-ID
Reporting

FIGURE 3. SCHEMATICS OF THE STEPS IN A NTS DATA ANALYSIS WORKFLOW.

2.2.2 Suspect screening: RT and MS1 based identification

These exact masses of interest can then be subjected to a suspect screening based on accurate mass to yield tentative IDs. Large databases to specific suspect lists can be used for such suspect screenings, an overview of relevant databases and lists can be found in Table 1. Alternatively to the exact mass, also the elemental formula can be used for the suspect screening which decreases the search space and thereby improves the likelihood of the tentative candidates to be true positives. The chemical formula can be determined based on the isotopic pattern of the feature and/or its MS2 spectrum. For instance, the software packages Sirius, GenForm and CSI: fingerID determine the elemental compositions of the MS2 fragments and then rank the possible elemental formulas by probability (Pervukhin et al., 2008; Meringer et al., 2011; Dührkop et al., 2015). Attaining an elemental formula improves the level of confidence to level 4. With retention time filters derived from experimental or predicted indices (Bade et al., 2015; Aalizadeh et al., 2016) candidates that are tentatively assigned to a feature based on the suspect screening can be further refined.

TABLE 1. DATABASES AND SUSPECTS LISTS APPLIED IN SUSPECT SCREENING OF NTS DATA THAT ARE RELEVANT FOR THE WATER SECTOR

Databases	**URL**	**Number of entries**
Chemspider	http://www.chemspider.com/	67 million
Pubchem	https://pubchem.ncbi.nlm.nih.gov/	97 million
EPA CompTox Chemistry Dashboard	https://comptox.epa.gov/dashboard	875’000
Suspect lists	**URL**	**Number of entries**
NORMAN SusDat	https://www.norman-network.com/nds/susdat/	>40’000, environmentally relevant
STOFFident	https://www.lfu.bayern.de/stoffident/#!home	>10’000, water relevant

2.2.3 MS2-based identification

To increase the confidence of a feature to level 2/3 and ultimately elucidate its structure, MS2 fragmentation data is required. The MS2 spectra can be used to search against spectral libraries of experimental MS2 spectra, such as Massbank and mzCloud (see Table 2)), or against *in silico* predicted spectra generated with software tools such as MetFrag and CFM:ID (Horai et al., 2010; Kasper et al., 2012; Allen et al., 2014; Ruttkies et al., 2016; Duhrkop et al., 2019). In the case of a high scoring match, the identity can be confirmed by the analysis of a reference standard. However, if too many candidates remain, the results can be further reduced by filtering the selection of candidates for chemical / physical properties, as well as metadata (Schymanski et al., 2017).

TABLE 2. SPECTRAL LIBRARIES AND *IN SILICO* FRAGMENTATION TOOLS FOR MS2 BASED IDENTIFICATION.

Spectral libraries	**URL**	**Number of entries**
Massbank	https://massbank.eu/MassBank/	56’000
mzCloud	https://www.mzcloud.org/	17’000

***in silico* fragmentation**	**URL**
MetFrag	https://msbi.ipb-halle.de/MetFragBeta/
CFM:ID	http://cfmid.wishartlab.com
FiSH scoring	implemented in Compound Discoverer

2.3 Limitations of available software at the start of the project

Various open-source software tools are available to perform the steps described above. However, these tools typically cover only part of the workflow, and thus multiple tools often need to be combined. Moreover, there are several tools with similar functionality but differing algorithms and/or data sources. As a consequence, the analyst needs to familiarize with various software environments, the optimization of their algorithms and the tedious conversion of data formats before the different tools can be evaluated and combined.

Some software distributed by commercial parties such as *Compound Discoverer* (Thermo Fisher Scientific) is able to perform (most of) the analysis steps described in 2.2. However, these software packages are restricted to proprietary data formats. For instance, *Compound Discoverer* only allows for analysis of Orbitrap data. Moreover, these software are not open-source which can lead to difficulties in sharing data, as well as extending functionalities by the analyst her/himself.

2.4 Software of choice

To address the benefits and limitations of both the open-source and commercial packages, we chose to pursue both lines within this project. As a Beta tester for *Compound Discoverer* 3.0, we took advantage of the commercial software’s easy to use interface, and the instrument specific functionalities and optimizations. Throughout the test phase, we were contributing to the improvement of software nodes. In particular, we achieved implementation of the NORMAN SusDat suspect list of environmentally relevant chemicals including their structures into *Compound Discoverer* 3.0. An introduction to *Compound Discoverer* 3.0, a detailed tutorial on its usage and the results of its application to water samples can be found in Chapter 3.

Concerning open source software, we focused on the new R based open-source software package ‘patRoon’ (hyPhenated mAss specTROmetry nOn-target aNalysis) that aims to provide a complete NTS data analysis solution. The package is being developed by Rick Helmus, PhD student of Pim de Voogt and Thomas ter Laak at the University of Amsterdam. This allowed for tight collaboration and feedback. An introduction to patRoon, a detailed tutorial on its usage and the results of its application to water samples can be found in Chapter 4.

3 *Compound Discoverer* 3.0

3.1 About *Compound Discoverer*

Compound Discoverer is a commercial software package developed by Thermo Fisher Scientific for mass spectrometry data analysis of small molecules. It has been used for more than three years at KWR, and is currently the default software for non-target screening data processing. In this chapter the main features of *Compound Discoverer* are discussed and a step-by-step tutorial is provided. This tutorial is for novice users and was also handed out to the participants of the non-target screening workshop held at KWR. At the end of the chapter the test results of the optimized *Compound Discoverer* NTS workflow on a small dataset are presented.

3.1.1 *Compound Discoverer* main features

Compound Discoverer is a comprehensive, integrated set of libraries, databases and statistical analysis tools which can be used together in a customizable workflow. Using these automated workflows it is possible to find statistical differences (known and unknown compounds) between samples. These differences can be identified automatically using the various annotation tools within *Compound Discoverer*. The workflow itself consists of different nodes (Figure 4) which can be easily added or removed by the user. This flexibility makes the program ideal for tailoring the workflow for a specific dataset or experiment. Within each node many parameters can be adjusted or optimized.

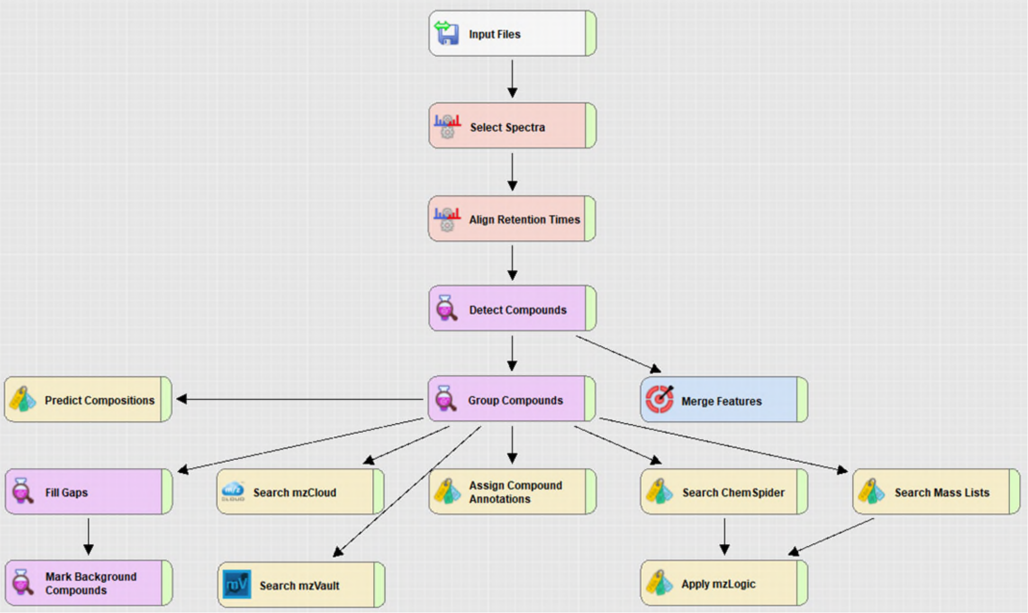


FIGURE 4. OVERVIEW OF THE NTS WORKFLOW USED AT KWR

One of the strengths of *Compound Discoverer* are the many available annotation nodes for the identification of compounds. The following annotation nodes are present in *Compound Discoverer*:

- Predict compositions node: Element formula prediction node
- Search mzCloud node: Online MS[1] and MS2 and MS^n database
- Search mzVault node: Offline MS[1] and MS2 database (including custom user libraries)
- Search Chemspider node: Online compound database node
- Search Mass Lists node: Compare detected compounds with known compounds in mass lists (with or without known RTs)
- Apply mzLogic node: *in silico* fragmentation tool that uses experimental fragmentation data in which fragments are mapped to substructures of potential structures and ranked for each unknown

All the above mentioned annotation nodes are used for the identification of compounds in the NTS workflow. The actual annotation of the compound depends on which node is set as preferred annotation node. For most data sets, the mzCloud annotation node is used as preferred annotation node. In paragraph 3.3 the confidence level of the automated identification using the mzCloud node was determined for a small dataset.

3.2 Tutorial *Compound Discoverer* 3.0

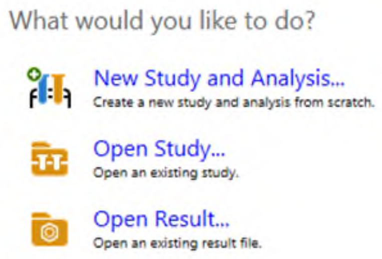
This tutorial was made for Discoverer version 3.0. It is recommended for novice users, but also contains tips and tricks for more experienced users.

3.2.1 Requirements for starting a *Compound Discoverer* experiment

In order to use the statistical capabilities of *Compound Discoverer* it is necessary to analyse samples in triplicate. This can be done by pre-treating samples in triplicate or by injecting each sample three times. Of course it is preferred to obtain samples in triplicate and then pre-treating them, because this approach provides more significant data. It is also important to include a representative blank reference sample during sample pre-treatment (in triplicate). This can be used as a reference sample for most experiments. By using a blank reference sample it is possible to determine the compounds that were unintentionally introduced during sample pre-treatment. Furthermore it is recommended to analyse a blank sample (not pre-treated) with the LC-MS analysis, which can be used during data processing for filtering the background noise/compound of the LC-MS system (singular or in triplicate).

3.2.2 Starting a new experiment

Start *Compound Discoverer* and choose New Study and Analysis (or by choosing file -> new Study and Analysis).


What would you like to do?

New Study and Analysis...
Create a new study and analysis from scratch.

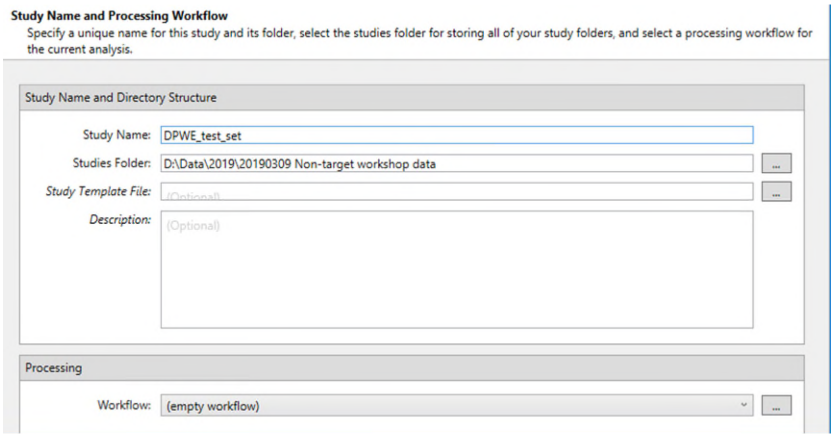
Open Study...
Open an existing study.

Open Result...
Open an existing result file.

Then a window appears with “New Study and Analysis wizard” and click next.

3.2.2.1 Step 2: of the wizard – Study Name

Pick a “Study Name” and select a folder for saving the data.
The boxes Study Template and Workflow can be skipped. Click next.



Study Name and Processing Workflow

Specify a unique name for this study and its folder, select the studies folder for storing all of your study folders, and select a processing workflow for the current analysis.

Study Name and Directory Structure

Study Name: DPWE_test_set

Studies Folder: D:\Data\2019\20190309 Non-target workshop data

Study Template File: (Optional)

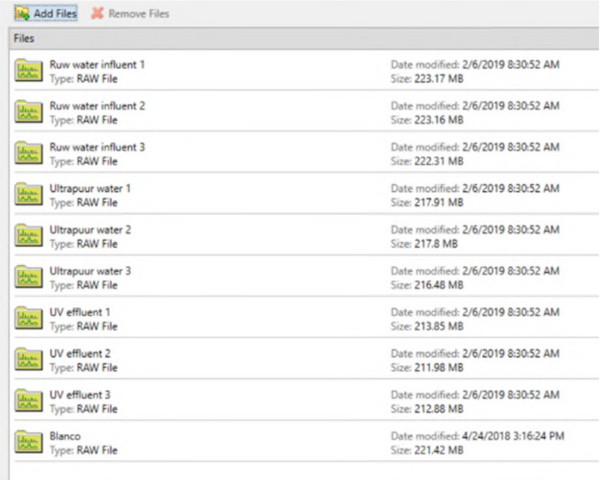
Description: (Optional)

Processing

Workflow: (empty workflow)

3.2.2.2 Step 3: of the Wizard – File Selection

Click “add files” and select the required files. Or drag the files directly using the explorer into the big white box. Click next.



Files

Ruw water influent 1
Type: RAW File
Date modified: 2/6/2019 8:30:52 AM
Size: 223.17 MB

Ruw water influent 2
Type: RAW File
Date modified: 2/6/2019 8:30:52 AM
Size: 223.16 MB

Ruw water influent 3
Type: RAW File
Date modified: 2/6/2019 8:30:52 AM
Size: 222.31 MB

Ultrapuur water 1
Type: RAW File
Date modified: 2/6/2019 8:30:52 AM
Size: 217.91 MB

Ultrapuur water 2
Type: RAW File
Date modified: 2/6/2019 8:30:52 AM
Size: 217.8 MB

Ultrapuur water 3
Type: RAW File
Date modified: 2/6/2019 8:30:52 AM
Size: 216.48 MB

UV effluent 1
Type: RAW File
Date modified: 2/6/2019 8:30:52 AM
Size: 213.85 MB

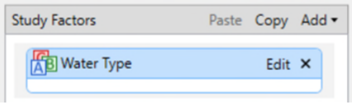
UV effluent 2
Type: RAW File
Date modified: 2/6/2019 8:30:52 AM
Size: 211.98 MB

UV effluent 3
Type: RAW File
Date modified: 2/6/2019 8:30:52 AM
Size: 212.88 MB

Blanco
Type: RAW File
Date modified: 4/24/2018 3:16:24 PM
Size: 221.42 MB

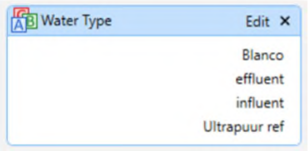
3.2.2.3 Step 4: of the Wizard – input File Characterization

Click on “add” (Study Factors) and select “Categorical Factor”. Chose a new Study Factor name such as “Water Type”



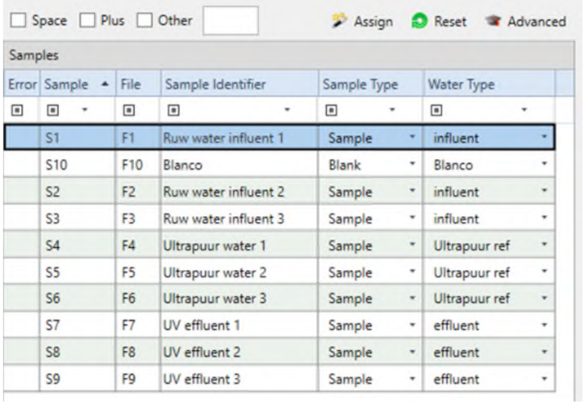
Water Type

Then add the name of the sample groups via edit (e.g. Blanco, ultrapuur ref, effluent and influent).



Blanco
effluent
influent
Ultrapuur ref

When the names of the sample groups match with a part of the file name of the .raw file, then the assign button can be used to automatically link the sample groups to the raw files. When some samples are not correctly assigned to the corresponding sample group, the pulldown menu of the corresponding sample in the "Water Type" column can be used for selecting the correct sample group.
Select in the column "Sample Type" "Blank" for the blank (blanco) sample. This sample is used by *Compound Discoverer* for the determination of the background noise.

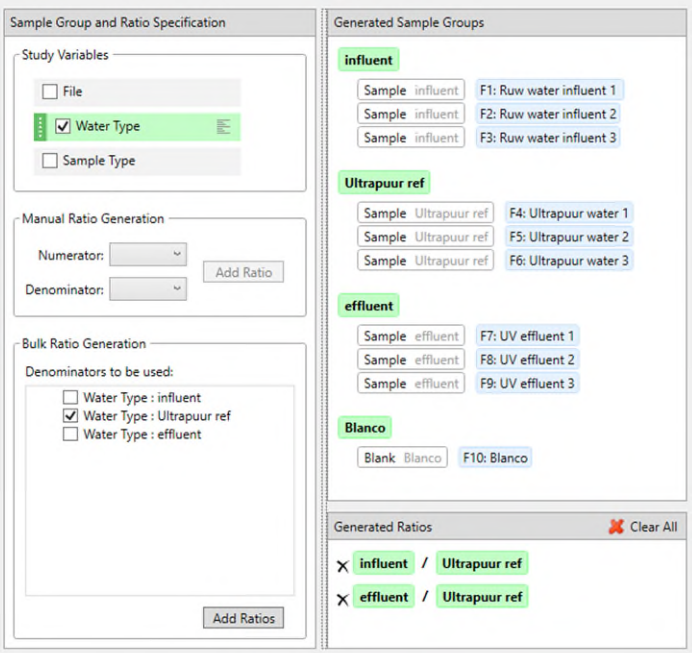


Error	Sample	File	Sample Identifier	Sample Type	Water Type
	S1	F1	Ruw water influent 1	Sample	influent
	S10	F10	Blanco	Blank	Blanco
	S2	F2	Ruw water influent 2	Sample	influent
	S3	F3	Ruw water influent 3	Sample	influent
	S4	F4	Ultrapuur water 1	Sample	Ultrapuur ref
	S5	F5	Ultrapuur water 2	Sample	Ultrapuur ref
	S6	F6	Ultrapuur water 3	Sample	Ultrapuur ref
	S7	F7	UV effluent 1	Sample	effluent
	S8	F8	UV effluent 2	Sample	effluent
	S9	F9	UV effluent 3	Sample	effluent

When all the study variables are filled in correctly, click next.

3.2.2.4 Step 5: of the Wizard – Sample Groups and Ratios

Select "Water Type" in the "Sample Group and Ratio Specification box. Select the dominator (e.g. ultrapuur ref) for the ratio calculation in the Bulk Ratio generation box and then click "add ratios". Now generated ratio appears in the Generated ratios box.



Sample Group and Ratio Specification

Study Variables

- □ File
- ☑ Water Type
- □ Sample Type

Manual Ratio Generation

Numerator:

Denominator:

Add Ratio

Bulk Ratio Generation

Denominators to be used:

- □ Water Type : influent
- ☑ Water Type : Ultrapuur ref
- □ Water Type : effluent

Add Ratios

Generated Sample Groups

influent
- Sample influent F1: Ruw water influent 1
- Sample influent F2: Ruw water influent 2
- Sample influent F3: Ruw water influent 3

Ultrapuur ref
- Sample Ultrapuur ref F4: Ultrapuur water 1
- Sample Ultrapuur ref F5: Ultrapuur water 2
- Sample Ultrapuur ref F6: Ultrapuur water 3

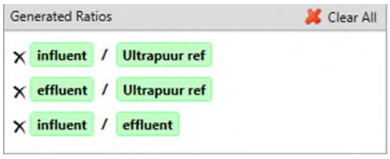
effluent
- Sample effluent F7: UV effluent 1
- Sample effluent F8: UV effluent 2
- Sample effluent F9: UV effluent 3

Blanco
- Blank Blanco F10: Blanco

Generated Ratios ✗ Clear All

- ✕ influent / Ultrapuur ref
- ✕ effluent / Ultrapuur ref

In this sample experiment it also meaningful to directly compare the influent and effluent sample which each other (e.g. for finding transformation products). Therefore also select effluent as denominator and click on “add ratios”.



Generated Ratios ✗ Clear All

- ✕ influent / Ultrapuur ref
- ✕ effluent / Ultrapuur ref
- ✕ influent / effluent

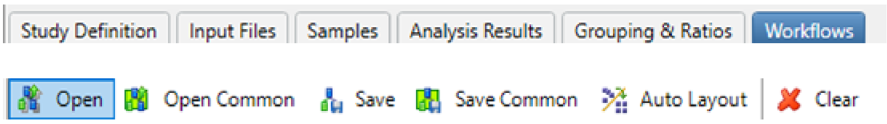
Now three ratios are calculated for data processing.

3.2.2.5 Step 6: of the Wizard – Confirm the analysis on the study page.

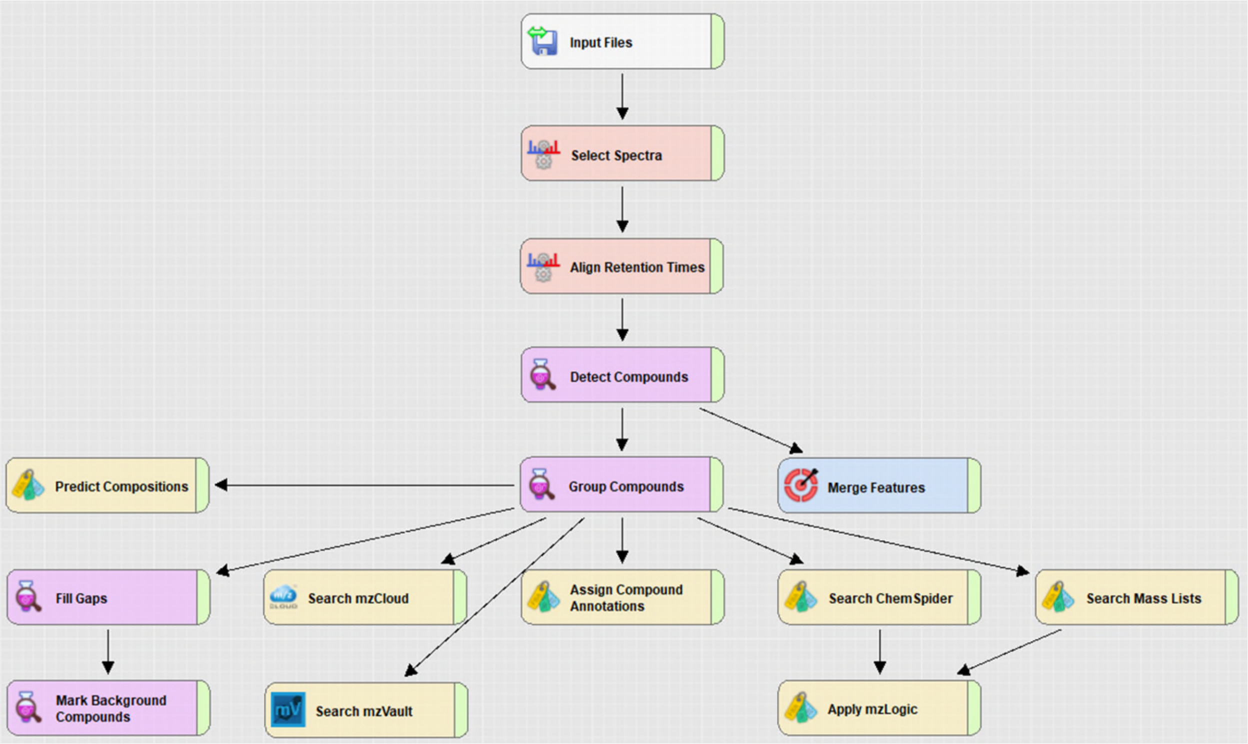
Read the text in the dialog box and click finish for finishing the study wizard.

3.2.3 Workflow and parameter selection

The last step before processing the data is the selection of a non-target screening workflow and optimisation of the workflow parameters. Click on the workflow tab and select open for opening a workflow template file.



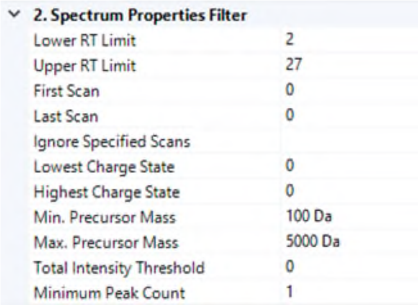
The default KWR non-target workflow is shown here.



Within this workflow it very easy to add or remove a node (by dragging a node for adding or pressing del for removing a node). For some nodes many parameters can be adjusted or optimized. In this manual only the most important parameters are discussed. Using the help function in *Compound Discoverer* extra information can be obtained about the function of a parameter.

3.2.3.1 Select spectra node

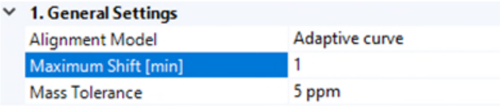
In the select spectra node, the spectra which are used for the data processing can be set here. The most important parameters that can be adjusted are in the “spectrum properties filter”. Here you can adjust the retention window which is used for data processing. It is recommended not to include the dead volume peak for data processing, because all the non-retained compounds are in here (which are a lot). It is recommended to use a Lower RT Limit of 2 min and an Upper RT Limit of 27 min. When the peak of interest is in the dead volume region, a lower RT can be used. After data processing it is still possible to adjust the RT limit using a filter and to clean up the data. But the filter can only be used within the range of the Lower and Upper RT limit. For the other parameters the default values can be used.



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	100 Da
Max. Precursor Mass	5000 Da
Total Intensity Threshold	0
Minimum Peak Count	1

3.2.3.2 Align Retention Times node

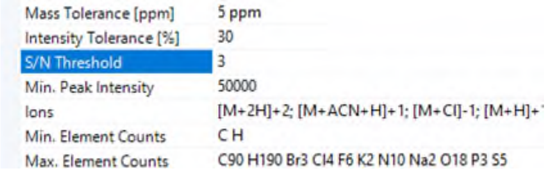
The align retention times node is useful when a retention shift has occurred during analysis. When the retention time are not shifted, it is recommended to delete the node from the workflow by pressing del. The first check for retention time shifts is to check the RT of the internal standards. When the retention times of the internal standards remain constant during analysis (< 0.05 min shift) the node can be deleted. When the retention has shifted, use a Maximum Shift of 1 min or lower.



1. General Settings	
Alignment Model	Adaptive curve
Maximum Shift [min]	1
Mass Tolerance	5 ppm

Detect Compounds node

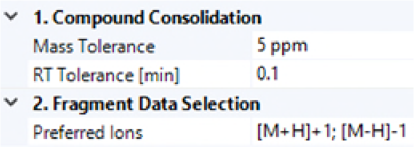
The detect compounds node is one of the most important nodes, because it is responsible for detection of unknown compounds. The number of detected compounds can be adjusted by changing the value for “Min. Peak Intensity”. It is recommend for regular water samples (i.e. not heavily contaminated) to use a “Min. Peak Intensity” of 50.000 – 100.000 counts. When the “Min. Peak Intensity” is set too low, the data processing time and amount of detected features is substantially increased (including background noise). For wastewater samples a higher “Min. Peak Intensity” should be used (100.000 – 1.000.000). Select for “Ions” the commonly observed ions for the LC-MS system. By default all are checked. Use the Max. Elements Counts as shown below. This are the elements used for the detection of the compounds including the adducts and not for the prediction of the formula (see Predict Composition node).



Mass Tolerance [ppm]	5 ppm
Intensity Tolerance [%]	30
S/N Threshold	3
Min. Peak Intensity	50000
Ions	[M+2H]+2; [M+ACN+H]+1; [M+Cl]-1; [M+H]+
Min. Element Counts	C H
Max. Element Counts	C90 H190 Br3 Cl4 F6 K2 N10 Na2 O18 P3 S5

3.2.3.3 Group Compounds node

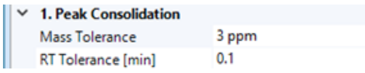
The group compounds node groups all compounds (including isotopes and adducts) by molecular weight and retention time. Use the default settings as shown below.



1. Compound Consolidation	
Mass Tolerance	5 ppm
RT Tolerance [min]	0.1
2. Fragment Data Selection	
Preferred Ions	[M+H]+1; [M-H]-1

3.2.3.4 Merge Features node

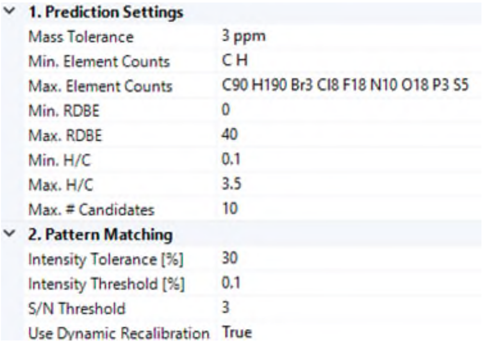
The Merge Features node merges all detected features and provides the links for the possible explanations. Use the default settings as shown below.



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

3.2.3.5 Predict compositions node

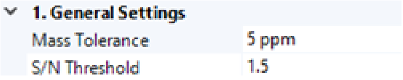
The predict compositions node predicts the elemental composition of the compounds detected. It is important to use a low Mass Tolerance for the calculation of the elemental composition. For the Orbitrap Fusion a mass tolerance of 3 ppm is recommended. A lower mass tolerance can be used (e.g. 2 ppm), but the user has to be certain that the mass error is sufficient for the whole used scan range (e.g. 80 – 1300 m/z). This can be checked by calculating the mass error of reference compounds. Use the default max. Elemental Counts for the formula prediction. Only adjust this when compounds are expected which do not fall within the Max. Element Count. Use the default settings as shown below.



1. Prediction Settings	
Mass Tolerance	3 ppm
Min. Element Counts	C H
Max. Element Counts	C90 H190 Br3 Cl8 F18 N10 O18 P3 S5
Min. RDBE	0
Max. RDBE	40
Min. H/C	0.1
Max. H/C	3.5
Max. # Candidates	10
2. Pattern Matching	
Intensity Tolerance [%]	30
Intensity Threshold [%]	0.1
S/N Threshold	3
Use Dynamic Recalibration	True

3.2.3.6 Fill Gaps Node

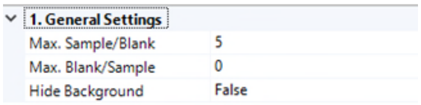
The fill gaps node fills the gaps for missing peaks. For example when a compound is only detected in one sample but is not detected in another sample because it is below the intensity threshold set in the detect compounds node, the fill gaps node will check again in that sample for the undetected compound and determine the peak area at the lowest possible intensity. Use the default settings as shown below.



1. General Settings	
Mass Tolerance	5 ppm
S/N Threshold	1.5

3.2.3.7 Mark Background compounds node

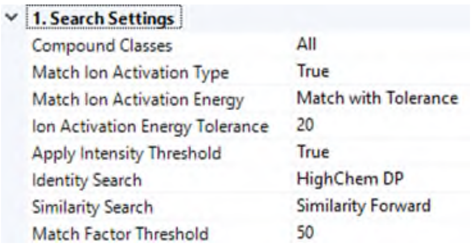
The Mark Background compounds node marks compounds as background when they are present in the blank sample and have a sample : blank ratio of < 5. These background compounds can be filtered out during data analysis.



1. General Settings	
Max. Sample/Blank	5
Max. Blank/Sample	0
Hide Background	False

3.2.3.8 Search mzCloud node

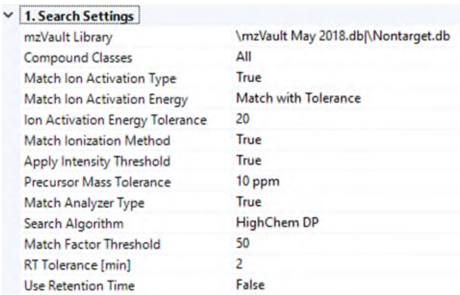
The mzCloud node identifies compound using MS2 spectra by comparing it to an online reference MS2 spectra database (mzCloud). If a MS2 spectrum has match factor > 50% the result will be stored for data analysis. Use the settings as shown below.



1. Search Settings	
Compound Classes	All
Match Ion Activation Type	True
Match Ion Activation Energy	Match with Tolerance
Ion Activation Energy Tolerance	20
Apply Intensity Threshold	True
Identity Search	HighChem DP
Similarity Search	Similarity Forward
Match Factor Threshold	50

3.2.3.9 Search mzVault node

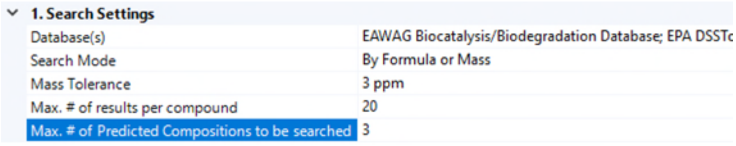
The mzVault node also identifies compound using MS2 spectra but compares it to a local database. MzVault has access to an offline version of the mzCloud database, but also can access custom user libraries (see nontarget.db below).



1. Search Settings	
mzVault Library	\mzVault May 2018.db|\Nontarget.db
Compound Classes	All
Match Ion Activation Type	True
Match Ion Activation Energy	Match with Tolerance
Ion Activation Energy Tolerance	20
Match Ionization Method	True
Apply Intensity Threshold	True
Precursor Mass Tolerance	10 ppm
Match Analyzer Type	True
Search Algorithm	HighChem DP
Match Factor Threshold	50
RT Tolerance [min]	2
Use Retention Time	False

3.2.3.10 Search ChemSpider node

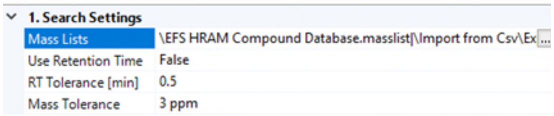
This nodes provides ChemSpider search results for the compounds detected. The unknown compounds are searched by formula or by mass when no formula is predicted. It is not recommend to use all databases within ChemSpider (slows down the search, and most libraries are not relevant). It is recommended to use the following libraries: EAWAG Biocatalysis/Biodegradation, EPA DSSTox, EPA toxcast, Drugbank, ACToR and FDA UNII – NLM. The mass tolerance should be set at a low value (< 3 ppm) for when the mass search is used (for restricting the amount of possible formulas and results).



1. Search Settings	
Database(s)	EAWAG Biocatalysis/Biodegradation Database; EPA DSSTc
Search Mode	By Formula or Mass
Mass Tolerance	3 ppm
Max. # of results per compound	20
Max. # of Predicted Compositions to be searched	3

3.2.3.11 Search Mass Lists node

With the mass list node the compounds detected are searched against a mass list with known compounds (suspect screening). Multiple mass lists can be used for this. A default mass list with relevant compounds is the EFS HRAM Compound database which contains 1634 semi-relevant compounds. Multiple mass lists from different sources are available at KWR. These include a mass list for relevant compounds in the water cycle (LOA-600 suspects) and the Norman Susdat list containing over 30.000 compounds. For regular water screening projects it is recommended to use the EFS HRAM and the LOA-600 suspect list. Use a mass tolerance of 2-3 ppm for searching the mass list. When applicable a retention time tolerance can be used for the mass lists.

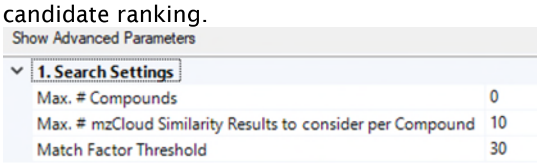


1. Search Settings

Mass Lists	\EFS HRAM Compound Database.masslist[\Import from Csv\Ex...
Use Retention Time	False
RT Tolerance [min]	0.5
Mass Tolerance	3 ppm

3.2.3.12 Apply mzLogic node

The mzLogic node compares the recorded MS2 data to the ion fragment library in mzCloud. It is an in-silico fragmentation tool that uses experimental fragmentation data in which fragments are mapped to substructures of potential structures and are ranked for each unknown. The mzLogic node uses the candidates/structures of the ChemSpider node and the mass list node (structure needs to be present in mass list in order to work) for the candidate ranking.



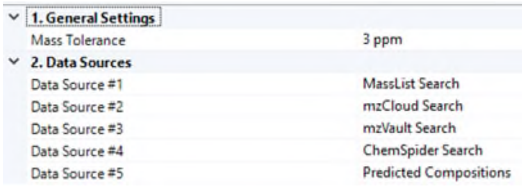
Show Advanced Parameters

1. Search Settings

Max. # Compounds	0
Max. # mzCloud Similarity Results to consider per Compound	10
Match Factor Threshold	30

3.2.3.13 Assign Compound Annotation node

This node assigns the compound annotation as it was determined or predicted by the other annotation nodes: predicted compositions, mzCloud search, mzVault search, Chemspider search and MassList search. In this node the order of the annotation which is used for the data analysis is set. Depending on the type of experiment the Mass List or the mzCLoud and mzVault search are the first data source.



1. General Settings

Mass Tolerance	3 ppm

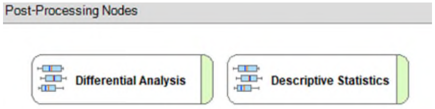
2. Data Sources

Data Source #1	MassList Search
Data Source #2	mzCloud Search
Data Source #3	mzVault Search
Data Source #4	ChemSpider Search
Data Source #5	Predicted Compositions

3.2.3.14 Post-Processing Nodes

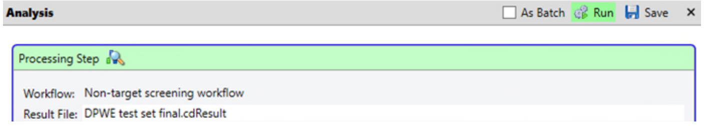
There are also two post-processing methods available for data analysis: Differential analysis and descriptive statistics. Make sure that these two nodes are present in the Post-Processing Nodes box (just below the workflow in Compound discoverer), otherwise only limited data analysis options are available.

The Descriptive Statistics node does not contain any adjustable parameters. For the Differential Analysis node only one parameter can be adjusted (Use log10 Areas, true/false)



3.2.3.15 Starting the data processing

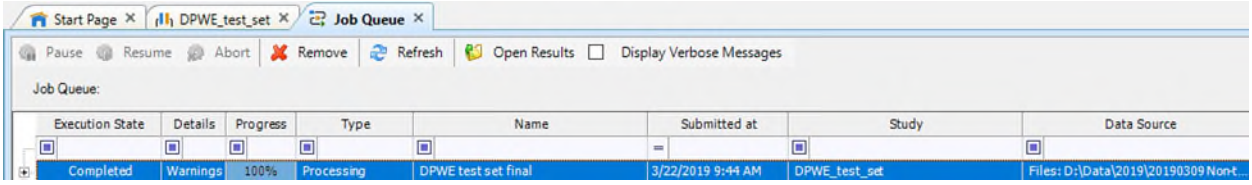
Specify a result file name and press “run” to start the data processing



Workflow: Non-target screening workflow
Result File: DPWE test set final.cdResult

3.2.4 Data Analysis

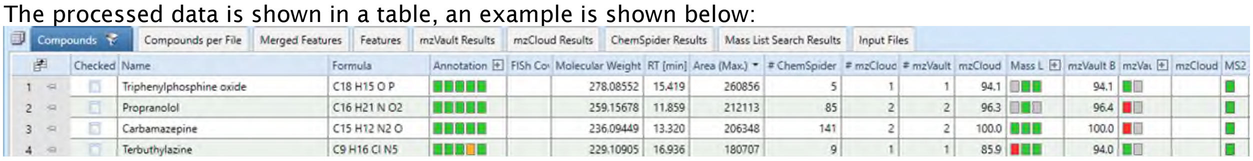
When *Compound Discoverer* is finished with the data processing, go to the job que tab and double click the completed experiment.



Job Queue:

Execution State	Details	Progress	Type	Name	Submitted at	Study	Data Source
Completed	Warnings	100%	Processing	DPWE test set final	3/22/2019 9:44 AM	DPWE_test_set	Files: D:\Data\2019\20190309 Non-t...

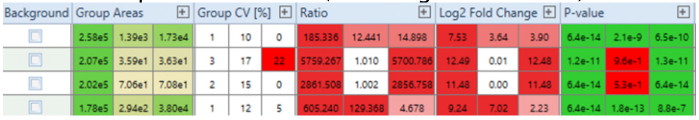
The processed data is shown in a table, an example is shown below:



	Checked	Name	Formula	Annotation	FISh Co	Molecular Weight	RT [min]	Area (Max.)	# ChemSpider	# mzCloud	# mzVault	mzCloud	Mass L	mzVault B	mzVa	mzCloud	MS2
1	☐	Triphenylphosphine oxide	C18 H15 O P	🟩🟩🟩🟩🟩		278.08552	15.419	260856	5	1	1	94.1	⬜🟩🟩	94.1	🟩⬜		🟩
2	☐	Propranolol	C16 H21 N O2	🟩🟩🟩🟩🟩		259.15678	11.859	212113	85	2	2	96.3	⬜🟩⬜	96.4	🟥⬜		🟩
3	☐	Carbamazepine	C15 H12 N2 O	🟩🟩🟩🟩🟩		236.09449	13.320	206348	141	2	2	100.0	🟩🟩🟩	100.0	🟥⬜		🟩
4	☐	Terbuthylazine	C9 H16 Cl N5	🟩🟩🟩🟧🟩		229.10905	16.936	180707	9	1	1	85.9	🟥🟩🟩	94.0	🟩⬜		🟩

The first part of the table contains the names of the compounds detected if they are identified. The presented formula is obtained from a library (i.e. mzCloud, mzVault or the mass list) or when it is not present in a library, the predicted formula is shown. The annotation source (i.e mzCloud, mzVault, predicted composition is shown in four colours: green, full match; orange, partial match; red, no match or invalid match; grey, no result). Furthermore compound information such as the molecular weight, RT and max area are shown in the table. Other important columns are the mzCloud and mzVault score, showing the library match score (%). The Mass lists results are also shown in four colours (same colour scheme as above). Another informative column is the MS2 column which shows whether or not a MS2 spectrum is present (which is needed for the annotation nodes).

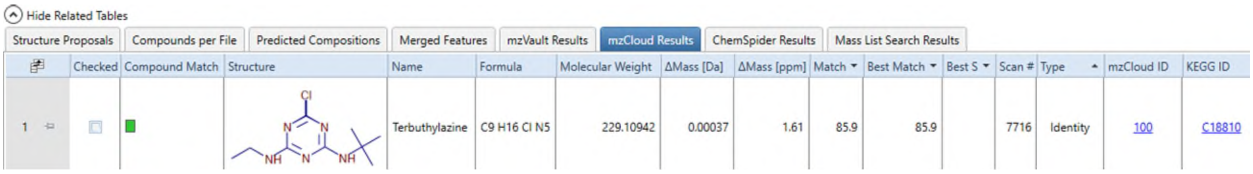
The second part of the table (to the right of the first) is shown below:



Background	Group Areas			Group CV [%]			Ratio			Log2 Fold Change			P-value		
☐	2.58e5	1.39e3	1.73e4	1	10	0	185.336	12.441	14.898	7.53	3.64	3.90	6.4e-14	2.1e-9	6.5e-10
☐	2.07e5	3.59e1	3.63e1	3	17	22	5759.267	1.010	5700.786	12.49	0.01	12.48	1.2e-11	9.6e-1	1.3e-11
☐	2.02e5	7.06e1	7.08e1	2	15	0	2861.508	1.002	2856.758	11.48	0.00	11.48	6.4e-14	5.3e-1	6.4e-14
☐	1.78e5	2.94e2	3.80e4	1	12	5	605.240	129.368	4.678	9.24	7.02	2.23	6.4e-14	1.8e-13	8.8e-7

In this part of the table a compound can be marked as background by ticking the background box. In this table the median group areas, coefficient of variation, ratio, log2 Fold Change and the p-value are shown of the compounds detected. Especially the ratios and p-values are helpful when comparing samples to each other. A low p-value (< 0.05) means that there is a significant difference between two samples. The group ratio, log2 Fold Change and p-value are parameters which are really suitable for filtering out interesting compounds from an experiment. The Group CV value is very useful for filtering background noise, because background noise (or ghost or spike peaks) have often very high CVs.

Below the compounds table, the related table is shown:



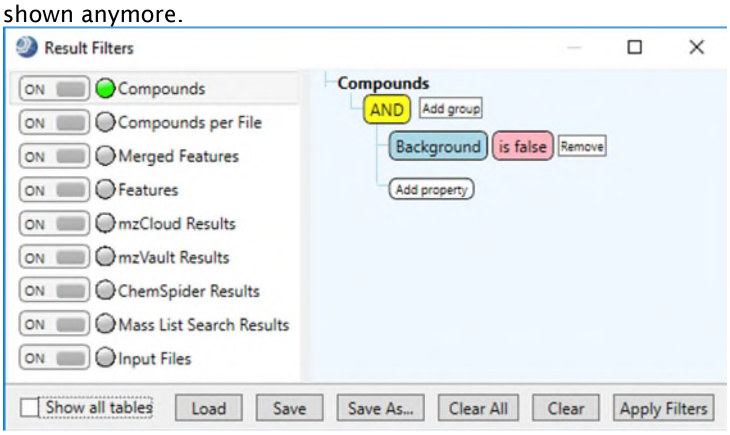
	Checked	Compound Match	Structure	Name	Formula	Molecular Weight	ΔMass [Da]	ΔMass [ppm]	Match	Best Match	Best S	Scan #	Type	mzCloud ID	KEGG ID
1	□	🟩		Terbuthylazine	C9 H16 Cl N5	229.10942	0.00037	1.61	85.9	85.9		7716	Identity	100	C18810

In this table more information is given about structural proposals, predicted compositions, merged features, mzVault results, mzCloud results (is shown above), Chemspider Results and Mass List search results.

3.2.4.1 Data analysis – Filter settings

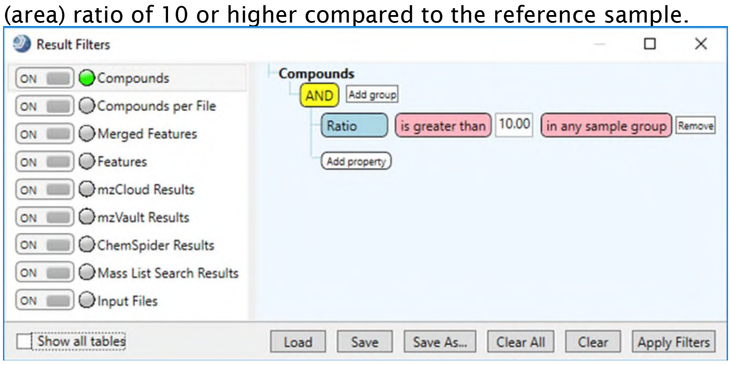
In order to find the compounds of interest or for cleaning your data, applying a filter will help reducing the amount of data. The result filter is found in the toolbar and can be accessed by pressing this icon:

One of the first filters to use is the Background filter for hiding all the background ions (detected in the LC-MS blank) in the data set. In order to do this open the result filter, click "add property" and select background from de pulldown menu and select "is false" in the red box. Then press apply filters in order to apply the filter. Now all the background ions are not shown anymore.



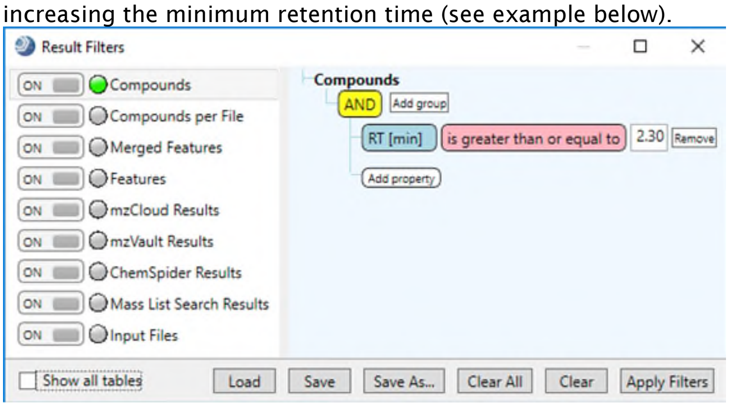
Compounds
AND
Background is false

Another very useful filter is the ratio filter, which can be used to filter out all the compounds which are also present in the reference sample. For this, a ratio > 10 in any sample group will suffice (see example below). Now only detected compounds are shown which have a (area) ratio of 10 or higher compared to the reference sample.



Compounds
AND
Ratio is greater than 10.00 in any sample group

Sometimes many compounds are present in the beginning of chromatogram, because they are not retained by the analytical column. If this is the case, a filter can be used for increasing the minimum retention time (see example below).



RT [min] is greater than or equal to 2.30

When your low intensity data is too noisy or not useful, a minimum area filter can help to filter out all the low intensity data.

If you want to compare two samples directly, you have to make sure a ratio is calculated between those two samples (see step 5: of the Wizard – Sample Groups and Ratios). Then the ratio or log2 fold change or p-value filters can be used to display the difference between the samples.

When data analysis is finished using *Compound Discoverer* the data can be exported to an .xls excel file. This can be done by right clicking on the compounds table and pressing export and selecting As excel.

3.2.4.2 Data analysis – statistic data analysis

There are also a few statistical data tools present in compound discoverer. These are: trend chart, result chart, descriptive statistics, differential analysis, PCA, PLS-DA and hierarchical cluster analysis. These can be accessed via the toolbar.

More information about these functions can be found in the *Compound Discoverer* manuals which are found via the toolbar -> help -> manuals.

3.3 Analysis of data set

In order to determine if the confidence level of the automated mzCloud identification using the optimized NTS workflow could be increased, a dataset containing “real” samples was processed using Compound Discoverer. The data set contained the following samples:

- Ultrapure water reference
- UV influent – surface water
- UV effluent

The influent and effluent samples were taken from a pilot drinking water treatment facility. The samples were originally used in a different study aimed at determining the removal of organic micro pollutants using UV treatment. Because organic micro pollutants were spiked to the UV influent, this data set is well suited for determining the confidence of the mzCloud identification for these micro pollutants.
The samples were processed with the NTS workflow. Only the compounds that were detected using the workflow and were at least 10x higher than the ultrapure water reference sample, were identified using mzCloud.

The following parameters were used for *Compound Discoverer* (most important only):

- Ionization mode: positive mode
- Detection threshold: 50.000 counts
- Mass tolerance: 5 ppm
- RT window: 2.3 – 27 min
- Annotation nodes used: Formula prediction and mzCloud search

The results of the automated identification of the spiked compounds (15) are shown in Table 3.

TABLE 3: RESULTS OF THE AUTOMATIC IDENTIFICATION OF SPIKED COMPOUNDS IN EFFLUENT AND INFLUENT WATER USING *COMPOUND DISCOVERER* AND MZCLOUD

Name	Formula	Molecular	RT	mzCloud		Confidence
		Weight (Da)	[min]	Identified	Match (%)	level
4-Methylbenzotriazole	$C_7H_7N_3$	133.0637	10.04	Yes	68.0	2
5-Methylbenzotriazole	$C_7H_7N_3$	133.0637	10.16	no	-	4
Aniline	C_6H_7N	93.05772	2.37	no	-	4
Benzotriazole	$C_6H_5N_3$	119.0480	8.00	Yes	82.5	2
Carbamazepine	$C_{15}H_{12}N_2O$	236.0945	13.32	Yes	100	2
Dimethenamid	$C_{12}H_{18}ClNO_2S$	243.0480	17.41	no	-	4
Dimethomorph	$C_{21}H_{22}ClNO_4$	387.1231	16.63	Yes	97.9	2
Gabapentin	$C_9H_{17}NO_2$	171.1255	6.39	Yes	89.5	2
Melamine	$C_3H_6N_6$	126.0651	2.15	Yes	80.0	2
Propranolol	$C_{16}H_{21}NO_2$	259.1568	11.86	Yes	96.3	2
Terbuthylazine	$C_9H_{16}ClN_5$	229.1091	16.94	Yes	85.9	2
Tetraglyme	$C_{10}H_{22}O_5$	222.1463	7.83	Yes	94.1	2
Tiamulin	$C_{28}H_{47}NO_4S$	493.3218	13.81	Yes	98.3	2
Tramadol	$C_{16}H_{25}NO_2$	263.1880	9.39	Yes	99.8	2
Triphenylphosphine oxide	$C_{18}H_{15}OP$	278.0855	15.42	Yes	94.1	2

Twelve of the fifteen spiked compounds were identified using the mzCloud database. Because the experimental spectra of these twelve compounds could be matched to the spectra in mzCloud, the confidence level could be increased from level 4 to level 2. Three compounds for which a confidence level of 4 (unequivocal formula) was obtained could not be identified. The three compounds were not identified using mzCloud for the following reasons: For 5-Methylbenzotriazole the MS2 spectrum was not present in the mzCloud database. For Aniline no MS2 spectrum was recorded, and could therefore not be identified. Dimethenamid was detected wrongly in *Compound Discoverer*. It was marked as a formic acid adduct of another detected mass (in-source fragment), and therefore no mzCloud search was performed.

These results show that the automated NTS workflow in *Compound Discoverer* works well, and that for 12 of the 15 compound the confidence level was increased to level 2. The results can be further improved by optimizing the data dependent acquisition method, in order to obtain improved MS2 coverage of the peaks detected. Current work in the BTO project “Improved non-target screening based identification through MS online prioritization” is addressing this issue. Moreover, by adding more reference compounds to custom MS2 libraries the issue can be alleviated in the future.

4 patRoon

4.1 About patRoon

'patRoon' stands for hyPhenated mAss specTROmetry nOn-target aNalysis and is R based open-source software package that aims to provide a complete NTS data analysis solution. The package is being developed by Rick Helmus, PhD student of Pim de Voogt and Thomas ter Laak. This allowed for tight collaboration and feedback. In this chapter an installation guide for R and patRoon, and a detailed tutorial on patRoon are provided. The results of its application to water samples are presented in the supplementary file "report.html".

4.2 Installation

4.2.1 R and RStudio installation

In order to use `patRoon` you need to have `R` and RStudio installed:

- Get `R` from: https://cloud.r-project.org/ (you need the base package)
- Get RStudio from: https://www.rstudio.com/products/rstudio/download/ (you need the desktop version)

NOTE: Please make sure to install `R` version 3.5.x (where x is 3 at the moment of writing).

If you already have R installed. It is highly recommended to ensure you have an up-to-date version (3.5). Furthermore, it is highly recommended to update all of your packages before installing `patRoon`, for instance by running:

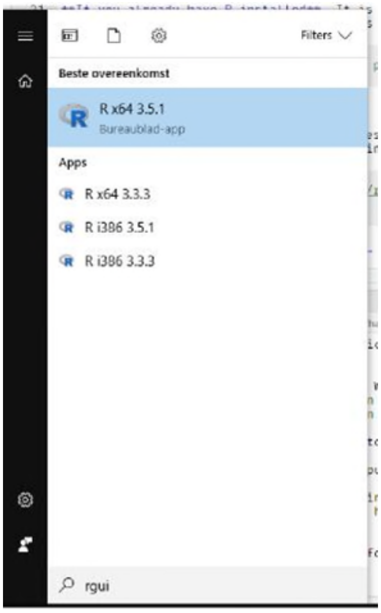
```
update.packages(ask = FALSE) # update all packages
```

4.2.2 patRoon installation

Besides `R` several other software packages and `R` packages need to be installed. The easiest option is to use the `patRoon` installation script for this. To do so, run the following commands:

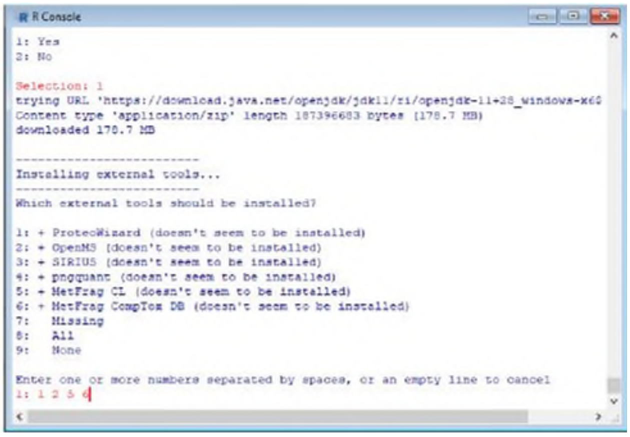
```
source("https://raw.githubusercontent.com/rickhelmus/patRoon/master/ins
tall_patRoon.R") installPatRoon()
```

NOTE it is highly recommended to *not* run the installation script in RStudio. Instead, run the commands in a 'regular' `R` console. The R console can be opened from the windows menu (it should be under the R program folder). If you cannot find it search for rgui (see below). Take care to select the x64 version!



Some hints about the installation process:

- It is recommended to install from the patRoonDeps repository (option 1 or 2). If you already have R installed it may be safer to use an isolated library (option 1).
- For the tutorial you only need the mandatory `R` packages (option 1).
- Agree if you are asked to create a personal R library.
- If the installers asks to install JDK and/or Rtools please proceed.
- *External tools*: Either simply install everything (option 8) or select the tools that are only necessary for the tutorial: ProteoWizard, OpenMS, MetFrag CL and MetFrag CompTox DB (options: 1, 2, 5, 6, see screenshot below). Please note that the ProteoWizard requires manual installation (the script will print instructions) and you need to manually go through the OpenMS installation wizard.



```
1: Yes
2: No

Selection: 1
trying URL 'https://download.java.net/openjdk/jdk11/ri/openjdk-11+28_windows-x6
Content type 'application/zip' length 187396683 bytes (178.7 MB)
downloaded 178.7 MB

--------------------------
Installing external tools...
--------------------------
Which external tools should be installed?

1: + ProteoWizard (doesn't seem to be installed)
2: + OpenMS (doesn't seem to be installed)
3: + SIRIUS (doesn't seem to be installed)
4: + pngquant (doesn't seem to be installed)
5: + MetFrag CL (doesn't seem to be installed)
6: + MetFrag CompTox DB (doesn't seem to be installed)
7:   Missing
8:   All
9:   None

Enter one or more numbers separated by spaces, or an empty line to cancel
1: 1 2 5 6
```

- Choose Yes (1) when the installer asks you if it should modify

your ~/.Rprofile file. You can re-run the installer at any time if you want to (re-)install something.

4.2.3 Verify the installation

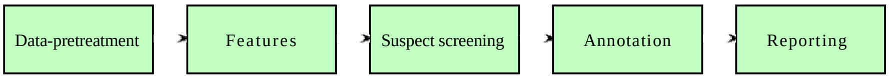
Close the `R` console that was used during the installation and open RStudio (or a restart the R console) and run the following command:

```
patRoon::verifyDependencies()
```

Take note that at least ProteoWizard, OpenMS, MetFrag CL and the MetFrag CompTox database are found.

4.3 Tutorial

This tutorial outlines how to perform suspect screening with `patRoon`. In this tutorial you will develop a screening workflow consisting of the following steps:



To summarize:

1. During data pre-treatment raw LC-MS data files (e.g. from Thermo Orbitrap or Bruker Q-TOF) are converted to an open file format needed for further processing.
2. All chromatographic peaks are automatically identified and stored as 'features'.
3. Suspect screening is performed: only features considered a suspect are retained.
4. Formulae and compound annotation is performed to ultimately verify the identity of a suspect.
5. Results are reported and interpreted.

4.3.1 R primer

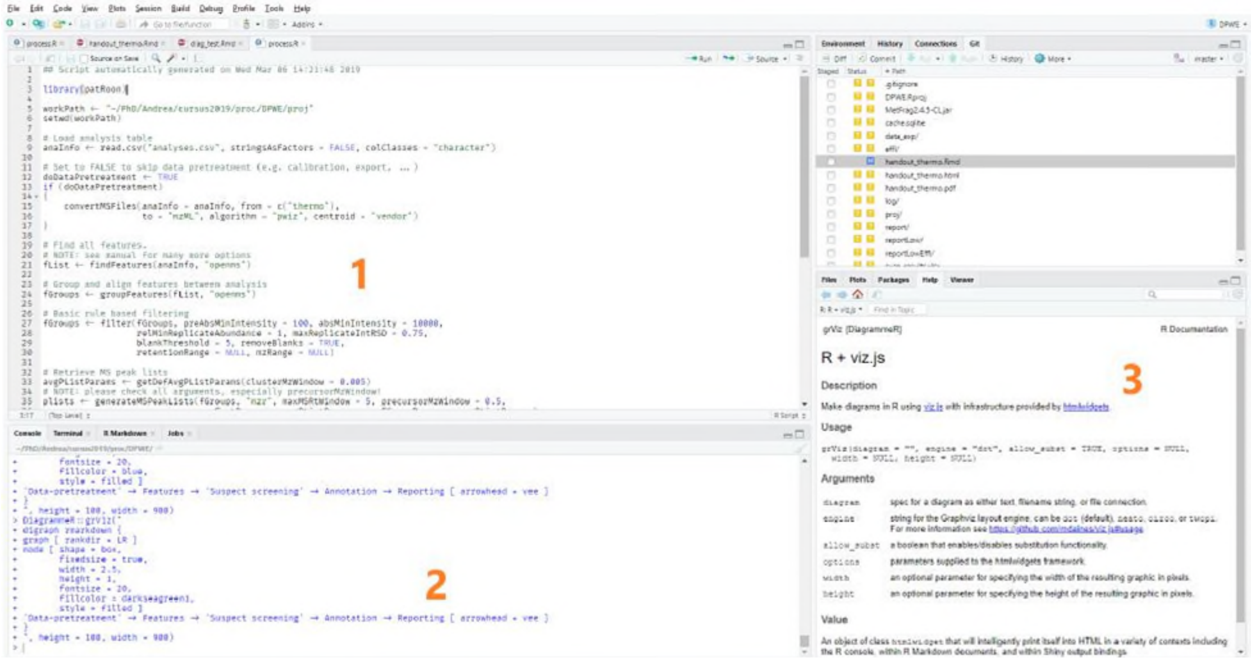
Users that are already familiar with R may want to skip this section.

The `R` programming language is nowadays commonly used to perform data science. It contains many tools to perform statistics, data transformation and tools for more specific research domains such as mass spectrometry. There are many online resources to learn more about `R`, for instance:

- Introduction to R by Monash Bioinformatics Platform
- RStudio cheat sheets such as R cheatsheet and RStudio cheatsheet
- ... and many more –> https://www.google.com

In this tutorial we will only use a small subset of R: `patRoon` will take care of most of the data processing tasks for you. Nevertheless, knowing more about `R` is a useful skill to have once you need to perform more advanced statistics, data processing or plotting of results.

For this tutorial you will be using RStudio. This software is a so called integrated development environment (IDE) for `R`. More importantly, this means that it makes working with `R` much easier and intuitive. A typical screenshot of RStudio looks like this:



```
## Script automatically generated on Wed Mar 06 14:21:48 2019

library(patRoon)

workPath ← "~/PhD/Andrea/cursus2019/proc/DPWE/proj"
setwd(workPath)

# Load analysis table
anaInfo ← read.csv("analyses.csv", stringsAsFactors = FALSE, colClasses = "character")

# Set to FALSE to skip data pretreatment (e.g. calibration, export, ...)
doDataPretreatment ← TRUE
if (doDataPretreatment)
{
    convertMSFiles(anaInfo = anaInfo, from = c("thermo"),
                   to = "mzML", algorithm = "pwiz", centroid = "vendor")
}

# Find all features.
# NOTE: see manual for many more options
fList ← findFeatures(anaInfo, "openms")

# Group and align features between analysis
fGroups ← groupFeatures(fList, "openms")

# Basic rule based filtering
fGroups ← filter(fGroups, preAbsMinIntensity = 100, absMinIntensity = 10000,
                 relMinReplicateAbundance = 1, maxReplicateIntRSD = 0.75,
                 blankThreshold = 5, removeBlanks = TRUE,
                 retentionRange = NULL, mzRange = NULL)

# Retrieve MS peak lists
avgPListParams ← getDefAvgPListParams(clusterMzWindow = 0.005)
# NOTE: please check all arguments, especially precursorMzWindow!
plists ← generateMSPeakLists(fGroups, "mzr", maxMSRtWindow = 5, precursorMzWindow = 0.5,
```

```
~/PhD/Andrea/cursus2019/proc/DPWE/
+         fontsize = 20,
+         fillcolor = blue,
+         style = filled ]
+ 'Data-pretreatment' → Features → 'Suspect screening' → Annotation → Reporting [ arrowhead = vee ]
+ }
+ ", height = 100, width = 900)
> DiagrammeR::grViz("
+ digraph rmarkdown {
+ graph [ rankdir = LR ]
+ node [ shape = box,
+         fixedsize = true,
+         width = 2.5,
+         height = 1,
+         fontsize = 20,
+         fillcolor = darkseagreen1,
+         style = filled ]
+ 'Data-pretreatment' → Features → 'Suspect screening' → Annotation → Reporting [ arrowhead = vee ]
+ }
+ ", height = 100, width = 900)
> 
```

grViz {DiagrammeR} R Documentation

R + viz.js

Description

Make diagrams in R using viz.js with infrastructure provided by htmlwidgets.

Usage

```
grViz(diagram = "", engine = "dot", allow_subst = TRUE, options = NULL,
  width = NULL, height = NULL)
```

Arguments

diagram	spec for a diagram as either text, filename string, or file connection.
engine	string for the Graphviz layout engine; can be dot (default), neato, circo, or twopi. For more information see https://github.com/mdaines/viz.js#usage
allow_subst	a boolean that enables/disables substitution functionality.
options	parameters supplied to the htmlwidgets framework.
width	an optional parameter for specifying the width of the resulting graphic in pixels.
height	an optional parameter for specifying the height of the resulting graphic in pixels.

Value

An object of class htmlwidget that will intelligently print itself into HTML in a variety of contexts including the R console, within R Markdown documents, and within Shiny output bindings.

The important areas are numbered:

1. This is the code editor. When you open an R script you will edit it here. (may not be visible if no scripts are open)
2. Here you will find the console where you can run R commands directly.
3. Here you can find help, install packages and open files present in your current project.

As mentioned above the console can be used to directly execute R commands. However, it is more common to generate an R script. These files contain a sequence of multiple R commands that together form your data processing workflow. An advantage of using a script file is that you don't forget how the data was processed (also known as 'reproducible research'). In the next sections we will automatically generate a script that will perform suspect screening on your data.

To execute code in your script from RStudio it is easiest to select the line(s) and press the Run button (shortcut: ctrl+enter). When no text is selected and the Run button is pressed the current line will be executed.

Now for some practice: launch RStudio, open a new R script file (shortcut: ctrl+shift+n) and paste and run the following code:

```
a <- 5

b <- 10 + a
print(a)
print(b)
```

```
## [1] 5

## [1] 15
```

You should see similar output in the console as is shown in the white box above.

Some more hints about `R` code:

- Text lines that start with a hash (#) are treated as text comments and are ignored by R. It is considered good practise to add comments to your code both for other readers and future you. Comments can appear on separate lines or after a line with code, for example:

```
# This is a comment line and ignored by R!
library(patRoon) # load the patRoon package
```

- Variable assignment in `R` happens with `<-` (*i.e.* not with =), for instance:

```
a <- 5.1

b <- "hello"
```

Will assign *5.1* to `a` and a text string (*"hello"*) to `b`. Note that text values always need to be quoted.

- If you want to know more about a function and its parameters you can use the help function of `R`, for instance type the following in the console to get more information on the `print` function:

```
?print
```

Note that you can also use this method to obtain more information on `patRoon` specific functions that we will use in this tutorial.

4.3.2 Suspect list

In this tutorial we will perform suspect screening. Hence, we need a database with compounds and their accurate (ionized) m/z values. To perform suspect screening with `patRoon` it is easiest to create a `.csv` (e.g. using Excel). The contents of this file should just be two columns: name and mz. The first column should contain the name of each suspect (ideally without special characters such as commas). The mz column should contain the accurate ionized (e.g. M+H or M-H) of the suspect.

The file used in this tutorial has the following format. You will need to select this file in the next section.

```
##                       name       mz
## 1   1,3-diphenylguanidine 212.1182
## 2        1H-Benzotriazole 120.0556
## 3 5-chloro-1H-benzotriazole 154.0167
## 4     Aldicarb-sulphoxide 207.0798
## 5           Amidosulfuron 370.0486
## 6             Bentazon-D6 247.1018
## 7           Bezafibrate 362.1154
## 8        Brodifancoum-A 523.0903
## 9Butocarboxim-sulphoxide 207.0798
## 10          Carbetamide 237.1234
```

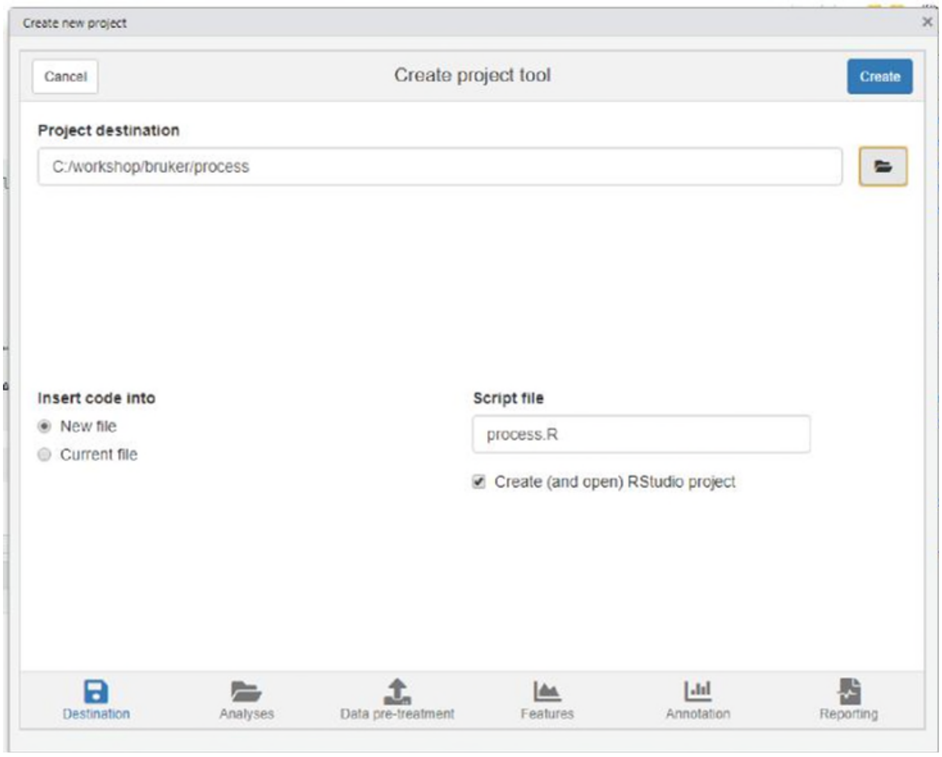
4.3.3 Create a new project

First start RStudio if you haven't already done so.

Whenever you start a new data processing project it is easiest to generate a project by running following command:

```
patRoon::newProject()
```

A tool will be launched that lets you define several settings that are used to generate a new project with a template R script. The screenshots in Figure 5 to Figure 9 summarize the settings you should use in this tutorial.



Create new project

Cancel | Create project tool | Create

Project destination

C:/workshop/bruker/process

Insert code into

● New file

○ Current file

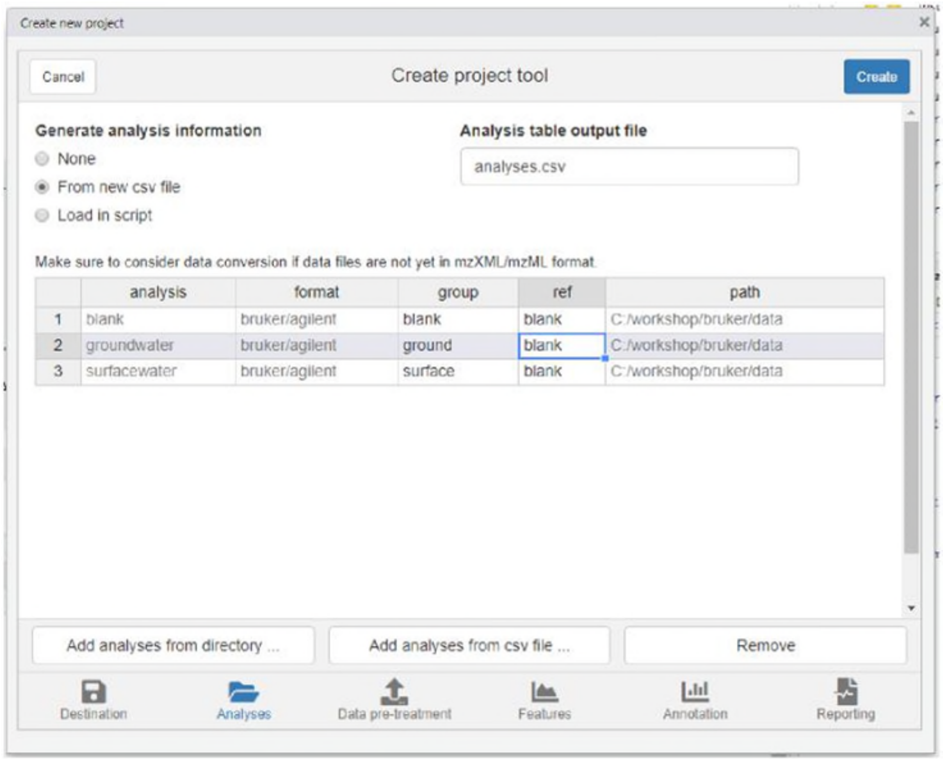
Script file

process.R

☑ Create (and open) RStudio project

FIGURE 5 DESTINATION TAB. SELECT THE PROJECT DESTINATION. THIS IS THE PATH WHERE R SCRIPTS AND WORKFLOW OUTPUT WILL BE STORED. YOU CAN

FREELY CHOOSE A LOCATION HERE. LEAVE OTHER OPTIONS IN THIS SCREEN AT THEIR DEFAULTS.



Generate analysis information

○ None
● From new csv file
○ Load in script

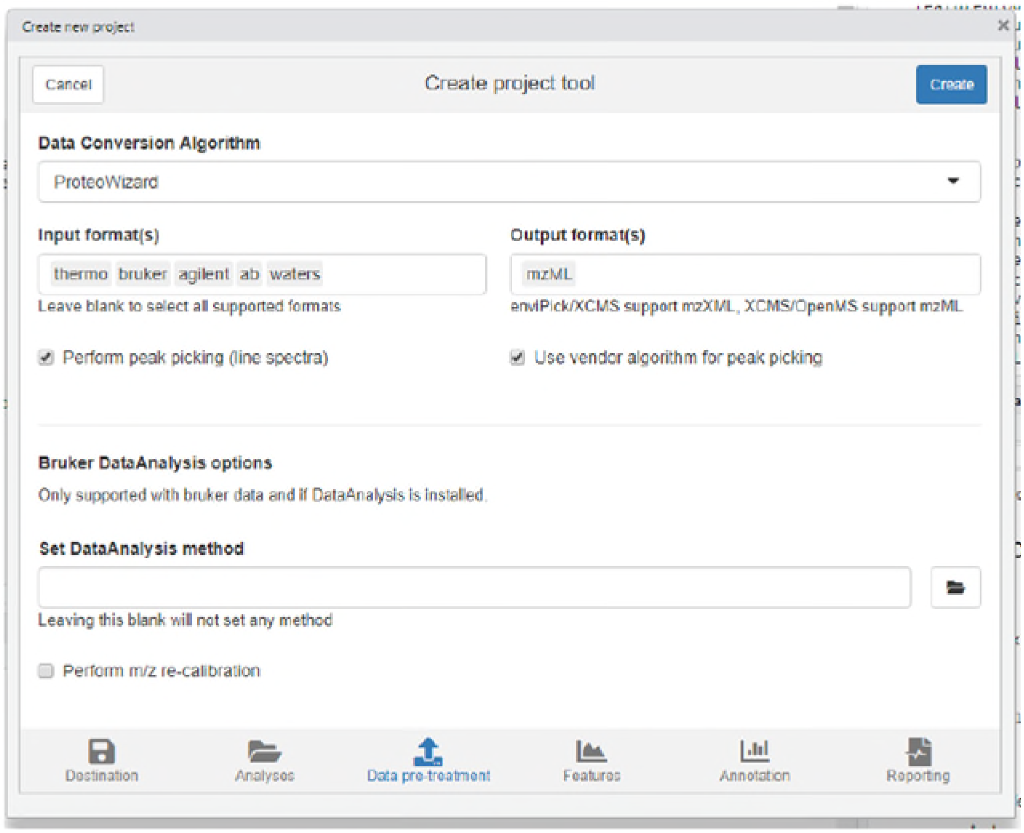
Analysis table output file: analyses.csv

Make sure to consider data conversion if data files are not yet in mzXML/mzML format.

	analysis	format	group	ref	path
1	blank	bruker/agilent	blank	blank	C:/workshop/bruker/data
2	groundwater	bruker/agilent	ground	blank	C:/workshop/bruker/data
3	surfacewater	bruker/agilent	surface	blank	C:/workshop/bruker/data

FIGURE 6. ANALYSES TAB. SELECT 'FROM NEW CSV FILE', ADD THE ANALYSES WITH THE 'ADD ANALYSES FROM DIRECTORY BUTTON' AND FILL IN THE GROUP AND REF COLUMNS AS SHOWN IN THE IMAGE. NOTE: THE GROUP COLUMN IS USED TO GROUP REPLICATE SAMPLES (NOT APPLICABLE TO BRUKER TUTORIAL),

WHEREAS THE REF COLUMN IS USED TO ASSIGN A 'REPLICATE GROUP' THAT SHOULD BE USED FOR BLANK SUBTRACTION.



Create project tool

Data Conversion Algorithm

ProteoWizard

Input format(s)

thermo bruker agilent ab waters

Leave blank to select all supported formats

Output format(s)

mzML

enviPick/XCMS support mzXML, XCMS/OpenMS support mzML

☑ Perform peak picking (line spectra)

☑ Use vendor algorithm for peak picking

Bruker DataAnalysis options

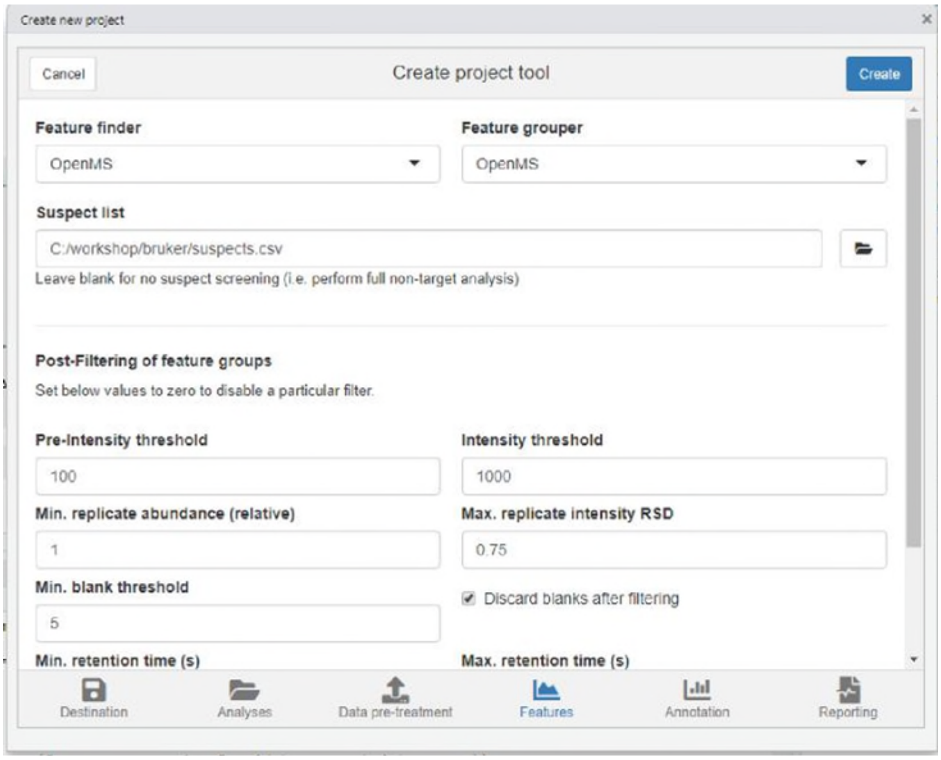
Only supported with bruker data and if DataAnalysis is installed.

Set DataAnalysis method

Leaving this blank will not set any method

□ Perform m/z re-calibration

FIGURE 7. DATA PRE-TREATMENT TAB. SELECT THE PROTEOWIZARD ALGORITHM. OTHER OPTIONS SHOULD BE LEFT AT THEIR DEFAULTS.



Feature finder: OpenMS

Feature grouper: OpenMS

Suspect list: C:/workshop/bruker/suspects.csv

Leave blank for no suspect screening (i.e. perform full non-target analysis)

Post-Filtering of feature groups

Set below values to zero to disable a particular filter.

Pre-Intensity threshold: 100

Intensity threshold: 1000

Min. replicate abundance (relative): 1

Max. replicate intensity RSD: 0.75

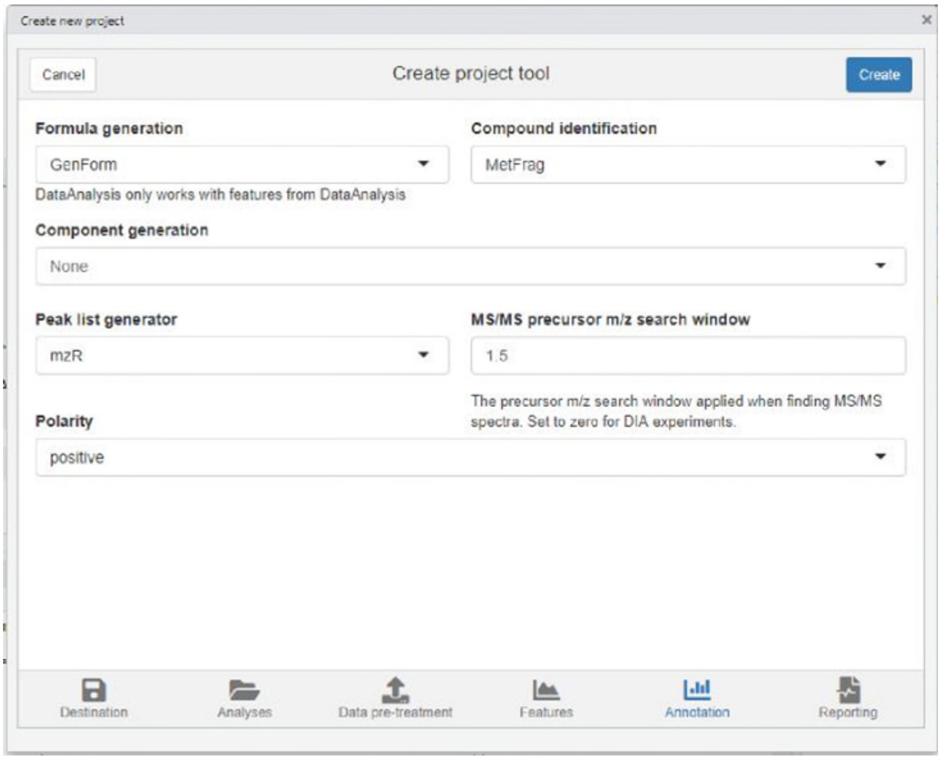
Min. blank threshold: 5

☑ Discard blanks after filtering

Min. retention time (s)

Max. retention time (s)

FIGURE 8. FEATURES TAB. OPENMS IS USED TO FIND AND GROUP FEATURES. SELECT THE SUSPECT LIST CSV FILE AND CHANGE THE INTENSITY THRESHOLD. FOR THE BRUKER DATA: SET THE MAXIMUM RETENTION TIME 975.



Formula generation: GenForm

DataAnalysis only works with features from DataAnalysis

Compound identification: MetFrag

Component generation: None

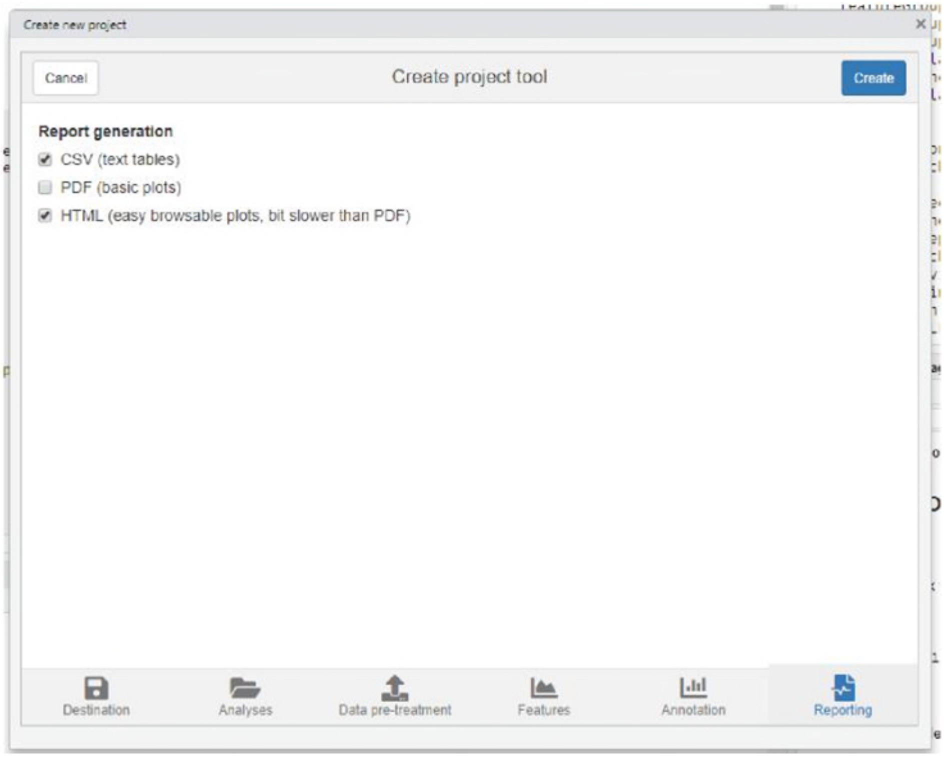
Peak list generator: mzR

MS/MS precursor m/z search window: 1.5

The precursor m/z search window applied when finding MS/MS spectra. Set to zero for DIA experiments.

Polarity: positive

FIGURE 9. ANNOTATION TAB. ENSURE THAT ALL SETTINGS ARE SET AS IS SHOWN HERE.



Report generation

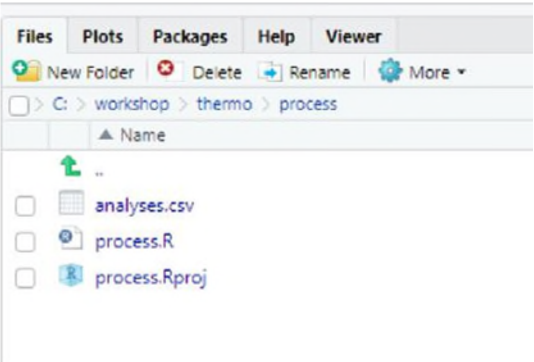
☑ CSV (text tables)

☐ PDF (basic plots)

☑ HTML (easy browsable plots, bit slower than PDF)

FIGURE 10 REPORTING TAB. YOU CAN LEAVE THE DEFAULTS HERE.

When everything is setup correctly press the “Create” button. RStudio will now automatically open the newly created project. After this is loaded you should open the generated R script (`process.R`) in your project directory.



C: > workshop > thermo > process

- ..
- analyses.csv
- process.R
- process.Rproj

FIGURE 11CLICK ON THE PROCESS.R FILE TO OPEN IT.

If everything went well the generated script (`process.R`) should look similar to this:

```
## Script automatically generated on Wed Mar 13 16:24:44 2019
library(patRoon)

 #  -----------
 # initialization

 #  -----------


 workPath <- "C:/workshop/bruker/process"
 setwd(workPath)

 # Load analysis table

 anaInfo <- read.csv("analyses.csv", stringsAsFactors = FALSE, colClasses
 = "character")


 # Set to FALSE to skip data pre-treatment

 doDataPretreatment <- TRUE

 if (doDataPretreatment)

 {

      convertMSFiles(anaInfo = anaInfo, from = c("thermo", "bruker",
                                "agilent", "ab",

     ↪ "waters"),

                    to = "mzML", algorithm = "pwiz", centroid = "vendor")

 }


 #  ------
 # features

 #  ------


 # Find all features.

 # NOTE: see manual for many more options

 fList <- findFeatures(anaInfo, "openms")

 # Group and align features between analysis
 fGroups <- groupFeatures(fList, "openms")

 # Basic rule based filtering
```

```
fGroups <- filter(fGroups, preAbsMinIntensity = 100, absMinIntensity =
1000,

                  relMinReplicateAbundance = 1, maxReplicateIntRSD =
                  0.75,

                  blankThreshold = 5, removeBlanks = TRUE,

                  retentionRange = c(0, 975), mzRange = NULL)

# Filter feature groups by suspects

suspFile <- read.csv("C:/workshop/bruker/suspects.csv", stringsAsFactors
= FALSE)

scr <- screenTargets(fGroups, suspFile, rtWindow = 12, mzWindow = 0.005)

fGroups <- groupFeaturesScreening(fGroups, scr)


#  --------
# annotation

#  --------


# Retrieve MS peak lists

avgPListParams <- getDefAvgPListParams(clusterMzWindow = 0.005)

plists <- generateMSPeakLists(fGroups, "mzr", maxMSRtWindow = 5,
precursorMzWindow = 1.5,

                             avgFeatParams = avgPListParams,
                             avgFGroupParams =

                               ,□ avgPListParams)

# uncomment and configure for extra filtering of MS peak lists

# plists <- filter(plists, absMSIntThr = NULL, absMSMSIntThr = NULL,
relMSIntThr = NULL,

#                 relMSMSIntThr = NULL, topMSPeaks = NULL, topMSMSPeaks
= NULL,

#                 deIsotopeMS = FALSE, deIsotopeMSMS = FALSE)

# Calculate formula candidates

formulas <- generateFormulas(fGroups, "genform", plists, relMzDev = 5,

                             adduct = "[M+H]+", elements = "CHNOP",
                              calculateFeatures = TRUE, featThreshold =
                              0.75)

# Find compound structure candidates

compounds <- generateCompounds(fGroups, plists, "metfrag", method = "CL",
dbRelMzDev = 5,

                                fragRelMzDev = 5, fragAbsMzDev = 0.002,

                                adduct = "[M+H]+", database = "pubchem",

                       ,□ maxCandidatesToStop = 2500)

compounds <- addFormulaScoring(compounds, formulas, TRUE)


#  -------
# reporting
    -------------
#  -------
    -------------

reportCSV(fGroups, path = "report", reportFeatures = FALSE, formulas =
formulas,

         compounds = compounds, compoundsNormalizeScores = "max",

         components = NULL)

reportMD(fGroups, path = "report", reportPlots = c("chord", "venn",
"upset", "eics",

,□ "formulas"), formula = formulas,

         compounds = compounds, compoundsNormalizeScores = "max",

         components = NULL, MSPeakLists = plists,

selfContained = FALSE, openReport = TRUE)
```

Before running this script, however, we still have to add and modify some of its code. In the next sections you will learn more about each part of the script, make the necessary changes and run its code.

4.3.4 Suspect screening workflow

4.3.4.1 Initialization

The first part of the script loads patRoon, makes sure the current working directory is set correctly and loads the analysis table generated earlier. This part in your script looks more or less like this:

```
library(patRoon)

workPath <- "C:/workshop/thermo/process"
setwd(workPath)

# Load analysis table


anaInfo <- read.csv("analyses.csv", stringsAsFactors = FALSE, colClasses
= "character")
```

Now go ahead and run this part of your script. To verify if the analysis table is correct you can inspect the contents of the `anaInfo` variable simply by running the following in the console:

```
anaInfo

##                                          path analysis group ref

## 1 C:/workshop/bruker/data blank blank blank

## 3 C:/workshop/bruker/data groundwater ground blank

## 5 C:/workshop/bruker/data surfacewater surface blank
```

The contents of `anaInfo` is a so called `data.frame`: a tabular data format, which in our case contains information about file locations and replicate and blank assignments.

After this we have to convert the analysis files from their vendor format (Thermo or Bruker) to an open data format. This is necessary because most software tools used by patRoon are only able to read open data formats. The following code in your script uses the `convertMSFiles()` function to convert the analyses to the open `mzML` format:

```
# Set to FALSE to skip data pretreatment (e.g. calibration, export, ...)

doDataPretreatment <- TRUE

if (doDataPretreatment)

{

     convertMSFiles(anaInfo = anaInfo, from = c("thermo", "bruker",
                                          "agilent", "ab",

    ,□ "waters"),

              to = "mzML", algorithm = "pwiz", centroid = "vendor")

}
```

Running this code will take some time. Afterwards, change the value of `doDataPretreatment` from `TRUE` to `FALSE` so you cannot accidentally re-convert the analyses files when (part of) the script is re-executed.

```
doDataPretreatment <- FALSE
```

4.3.4.2 Finding and grouping features

The first step of a non-target screening workflow consist of finding features. This is performed with the `findFeatures()` function. Your script should contain the following line that calls this function and stores its results in the `fList` variable (don't run this yet!):

```
fList <- findFeatures(anaInfo, "openms")
```

The `findFeatures()` function accepts many parameters that will influence its behaviour. Setting their values correctly is highly important to be able to find the many features hidden in your samples, while at the same time care has to be taken that, for instance, chromatographic noise should not be considered to assign features.

We will come back later to optimizing parameters. For now modify the code that calls `findFeatures()` and run it afterwards:

```
fList <- findFeatures(anaInfo, "openms", noiseThrInt = 500,
                      chromFWHM = 3, minFWHM = 1, maxFWHM = 30,
                      chromSNR = 3, mzPPM = 5)
```

```
## Finding features with OpenMS for 3 analyses ...

## Done!

## Feature statistics:


## blank: 481 (22.9°h)

## groundwater: 684 (32.6°h)
## surfacewater: 932 (44.4°h)
## Total: 2097
```

After the features have been found (this may take some minutes), the next step is to group features across analyses and perform basic filtering to clean your dataset. In your script these steps are performed by the the `groupFeatures()` and `filter()` functions. You don't have to change anything here. simply run it from your script.

```
fGroups <- groupFeatures(fList, "openms")

fGroups <- filter(fGroups, preAbsMinIntensity = 100,
                  absMinIntensity = 1000,
                  relMinReplicateAbundance = 1,
                  maxReplicateIntRSD = 0.75,
                  blankThreshold = 5, removeBlanks = TRUE,

                  retentionRange = c(0, 975), mzRange = NULL)
```

```
## Applying intensity filter... Done! Filtered 0 (0.00°h) groups.
Remaining: 1250.

## Applying retention filter... Done! Filtered 92 (7.36°h) groups.
Remaining: 1158.

## Applying replicate abundance filter... Done! Filtered 0 (0.00°h) groups.
Remaining: 1158.

## Applying blank filter... Done! Filtered 383 (33.07°h) groups.
Remaining: 775.

## Applying intensity filter... Done! Filtered 2 (0.26°h) groups.
Remaining: 773.

## Applying replicate abundance filter... Done! Filtered 0 (0.00°h)
groups. Remaining: 773.

## Applying replicate group filter... Done! Filtered 0 (0.00°h) groups.
Remaining: 773.
```

The filtering step consists of several steps and features will be removed if:

- Their intensity is below a defined intensity threshold (set by `absMinIntensity`).
- They are not ubiquitously present in (part of) replicate analyses. This is controlled by setting `relMinReplicateAbundance`. The value is relative, for instance, a value of `0.5` would mean that a feature needs to be present in half of the replicates. In this tutorial we use a value of `1` which means that a feature should be present in all replicate samples.
- Features that do not have a significantly higher intensity than the blank intensity are removed. This is controlled by `blankThreshold`: the given value of `5` means that the intensity of a feature needs to be at least five times higher compared to the (average) blank signal.

4.3.4.3 Suspect screening

The next step concerns suspect screening:

```
suspFile <- read.csv("C:/workshop/thermo/suspects.csv", stringsAsFactors
= FALSE)

scr <- screenTargets(fGroups, suspFile, rtWindow = 12, mzWindow = 0.005)

fGroups <- groupFeaturesScreening(fGroups, scr)
```

```
## Found 27/28 targets (96.43°h)

## Converting screening results to feature groups...
## Removing empty screening results

## Done!
```

In this code block the following three steps occur:

1. The suspect list file is loaded (from the file path you selected when the script was generated).
2. The screenTargets() function is called to find any suspects in your feature dataset. The results (an R `data.frame`) are stored in the `scr` variable.
3. The screening results are subsequently used to filter out any features in the original dataset that are not considered to be suspects.

Running this code will show you how many suspects were found. Don't worry if not all were found, this will be dealt with when optimizing the feature finding parameters in a later section.

4.3.5 Annotation

The final steps of the non-target workflow consists of annotation: here data from MS and MS/MS spectra is used to (automatically) assign possible formulae and compound structures to all features. This data is crucial to verify the chemical identity of a suspect.

4.3.5.1 MS Peak Lists

Prior to performing formulae and compound annotation we need to extract MS and MS/MS data from all features. This is performed with the `generateMSPeakLists()` function. Before running this function, please find the part in your script where this function is called and modify this part of your script so that it corresponds with the following code block:

```
avgPListParams <- getDefAvgPListParams(clusterMzWindow = 0.001)

plists <- generateMSPeakLists(fGroups, "mzr", maxMSRtWindow = 5, precursorMzWindow = 1.5,

                              avgFeatParams = avgPListParams, avgFGroupParams =
```

```
## Loading all MS peak lists for 30 feature groups in analysis
'groundwater'... ##
========================================================================
================ ## Loading all MS peak lists for 30 feature
groups in analysis 'surfacewater'... ##
========================================================================
================ ## Generating averaged peak lists for all
feature groups...

## Done!
```

During the last step the `filter()` function is called to cleanup the MS/MS spectra: all mass peaks with intensity below 2% are removed and only the ten most intense mass peaks are retained.

4.3.5.2 Formula calculation

The `generateFormulas()` function is used to automatically calculate formula candidates for each feature. Note that you need to change the `elements` parameter to this function to make sure that formulae with sulphur and chloride (S/Cl) are also accepted. Again running this code may take some time.

```
formulas <- generateFormulas(fGroups, "genform", plists, relMzDev = 5,

                             adduct = "[M+H]+", elements = "CHNOPSCl")
```

```
## Loading all MS formulas for analysis 'groundwater'...

## Loaded 2374 MS formulas for 26 features (86.67°).

## Loading all MS/MS formulas for analysis 'groundwater'...

## Loaded 1713 MS/MS formulas for 25 features (83.33°).

## Loading all MS formulas for analysis 'surfacewater'...

## Loaded 592 MS formulas for 24 features (80.00°).

## Loading all MS/MS formulas for analysis 'surfacewater'...

## Loaded 419 MS/MS formulas for 18 features (60.00°).

## Generating feature group formula consensus...

## Done!
```

4.3.5.3 Compound annotation

In order to assign structural candidates to our features we call the `generateCompounds()` function. Please modify the code in your script so it matches the following before running it:

```
compounds <- generateCompounds(fGroups, plists, "metfrag", method = "CL",

                               dbRelMzDev = 5, fragRelMzDev = 5, fragAbsMzDev = 0.002,

                               adduct = "[M+H]+", database = "comptox",
```

```
## Identifying 30 feature groups with MetFrag...

## Loaded 1131 compounds from 28 features (93.33%).

## Adding formula scoring...

## =================================================
```

In this tutorial we use the CompTox database (as set with `database = "comptox"`). While other databases such as PubChem are also possible, this database is generally more specialized towards contaminants that may be found in the environment (the database can be accessed online here: https://comptox.epa.gov/dashboard). Usage of the PubChem database is outlined at the end of this tutorial.

During the last step the `addFormulaScoring()` function is called to improve ranking of candidates by incorporating the formula calculation data from the previous step.

4.3.6 Reporting

The script ends with reporting your data:

```
reportCSV(fGroups, path = "report", reportFeatures = FALSE, formulas = formulas,

          compounds = compounds, compoundsNormalizeScores = "max",

          components = NULL)

reportMD(fGroups, path = "report", reportPlots = c("venn", "eics", "formulas"), formulas

,□ = formulas,
```

The report functions (`reportCSV` and `reportMD`) accept many parameters to influence their output. However, in this tutorial the defaults will suffice. Running this code may take a minute or two. The `reportMD()` function generates an easy to navigate report of all the data that was generated during the workflow. This file should be opened automatically when it is finished. More detailed results can be found in the CSV files that are generated with `reportCSV()`. All report files are stored in the `report` subdirectory inside your project directory.

Try to see if you can verify the identity of all suspects from the generated report that was generated by `reportMD()`.

4.3.7 Final R script

For reference: the final script should look similar as below.

```
## Script automatically generated on Wed Mar 13 16:24:44 2019

library(patRoon)
```

```
# -------------
# initialization
# -------------
```

```
workPath <- "C:/workshop/bruker/process"
setwd(workPath)

# Load analysis table

anaInfo <- read.csv("analyses.csv", stringsAsFactors = FALSE, colClasses = "character")


# Set to FALSE to skip data pre-treatment

doDataPretreatment <- FALSE

if (doDataPretreatment)

{

     convertMSFiles(anaInfo = anaInfo, from = c("thermo", "bruker", "agilent", "ab",

    ↪ "waters"),

                   to = "mzML", algorithm = "pwiz", centroid = "vendor")

}


#  ------
# features

#  ------


# Find all features.

# NOTE: see manual for many more options

fList <- findFeatures(anaInfo, "openms", noiseThrInt = 500,

                           chromFWHM = 3, minFWHM = 1, maxFWHM = 30,

                        chromSNR = 3, mzPPM = 5)

# Group and align features between analysis
fGroups <- groupFeatures(fList, "openms")

# Basic rule based filtering

fGroups <- filter(fGroups, preAbsMinIntensity = 100, absMinIntensity = 1000,

                   relMinReplicateAbundance = 1, maxReplicateIntRSD = 0.75,

                   blankThreshold = 5, removeBlanks = TRUE,
```

```
                              ↪ avgPListParams)
plists <- filter(plists, relMSMSIntThr = 0.02, topMSMSPeaks = 10)

# Calculate formula candidates
formulas <- generateFormulas(fGroups, "genform", plists, relMzDev = 5,
                             adduct = "[M+H]+", elements =
                             "CHNOPSCl", calculateFeatures
                             = TRUE, featThreshold = 0.75)
# Find compound structure candidates
compounds <- generateCompounds(fGroups, plists, "metfrag", method = "CL",
dbRelMzDev = 5,
                               fragRelMzDev = 5, fragAbsMzDev = 0.002,
                               adduct = "[M+H]+", database = "comptox",
                     □ maxCandidatesToStop = 15000)
compounds <- addFormulaScoring(compounds, formulas, TRUE)

#  -------
# reporting
#  -------

reportCSV(fGroups, path = "report", reportFeatures = FALSE, formulas =
formulas,
          compounds = compounds, compoundsNormalizeScores = "max",
          components = NULL)
reportMD(fGroups, path = "report", reportPlots = c("chord", "venn",
"upset", "eics",
□ "formulas"), formulas = formulas,
          compounds = compounds, compoundsNormalizeScores = "max",
          components = NULL, MSPeakLists = plists,
          selfContained = FALSE, openReport = TRUE)
```

4.3.8 Advanced topics

4.3.8.1 Inspecting and plotting data

The automatically generated report should provide you with a lot of information. Sometimes, however, you may have to dig a bit further in the data.

Below are some commands to inspect results for grouped features (i.e. intensities) and formula/compound candidates.

```
as.data.frame(fGroups) # convert feature data to a data.frame and show
its contents
```

```
##   group                           ret     mz groundwater surfacewater
##  1 1,3-diphenylguanidine 343.3486 212.1180       91088        45394
##  2       1H-Benzotriazole 357.0950 120.0556       42786        30835
##  3  5-chloro-1H-benzotriazole 457.4220 154.0164       13160         6206
##  4    Aldicarb-sulphoxide 275.0379 207.0795       26495        12269
##  5 Aldicarb-sulphoxide.1 283.9886 207.0796       20142        10611
##  6            Amidosulfuron 470.0035 370.0481       25080         9033
##  7              Bezafibrate 587.6653 362.1149       15428         5849
##  8          Brodifancoum-A 830.6575 523.0901       16189            0
##  9  Butocarboxim-sulphoxide 275.0379 207.0795       26495        12269
##  10 Butocarboxim-sulphoxide.1 283.9886207.0796       20142        10611

## 11              Carbetamide 278.5180 237.1229           0         3314

## 12            Carbetamide.1 198.8390 237.1229           0         5838
## 13                Carbofuran 486.2457 222.1123       67863        39890
## 14                  Di-glyme 311.3092 135.1016        7435         3507
## 15                  Etrimfos 710.6363 293.0718       69395        31207
## 16               Fenofibrate 793.5132 361.1198       66639            0
## 17                 Flonicamid 327.7561 230.0533       29074        13415
## 18            Foramsulfuron 510.4085 453.1181       15812         5978
## 19               Gabapentine 299.9987 172.1331       29333        17260
## 20               Gemfibrozil 766.9718 251.1640        8843         5462
## 21                Indoxacarb 745.4971 528.0775        4466            0
## 22                 Metformin 163.7010 130.1087           0         4365
## 23                  Methomyl 321.5968 163.0534        5969         2365
## 24               Metrafenone 737.7728 409.0644       50313         4030
## 25               Pipamperone 424.4348 376.2391       46191        18122
## 26                 Propazine 609.2565 230.1163       89302        45574
## 27            Propiconazole 713.8918 342.0768       28444         9821
## 28                Spinosyn A 727.8661 732.4668        3181            0
## 29               Tembotrione 486.2844 441.0379        1359            0
## 30  Triphenylphosphine oxide 591.8253 279.0931      134214            0
```

```
formulas[["Gabapentine"]] # formula candidates for
Gabapentine suspect
```

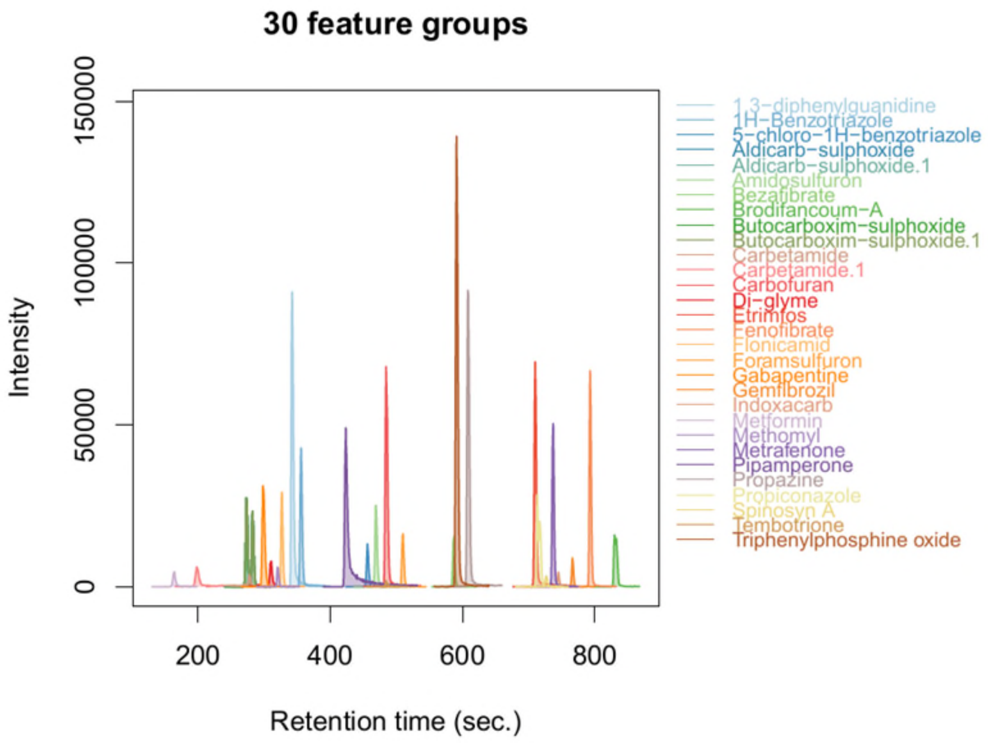
```
##   neutral_formula dbe formula_mz error isoScore byMSMS formula  frag_mz frag_error frag_
## 1:        C9H17NO2   2  172.1332   1.5  0.84377  TRUE C9H18NO2  95.08544        0.9
## 2:        C9H17NO2   2  172.1332   1.5  0.84377  TRUE C9H18NO2 137.09607        0.2
## 3:        C9H17NO2   2  172.1332   1.5  0.84377  TRUE C9H18NO2 154.12254        0.6
```

```
compounds[["Foramsulfuron"]] # compound candidates for
Foramsulfuron suspect
```

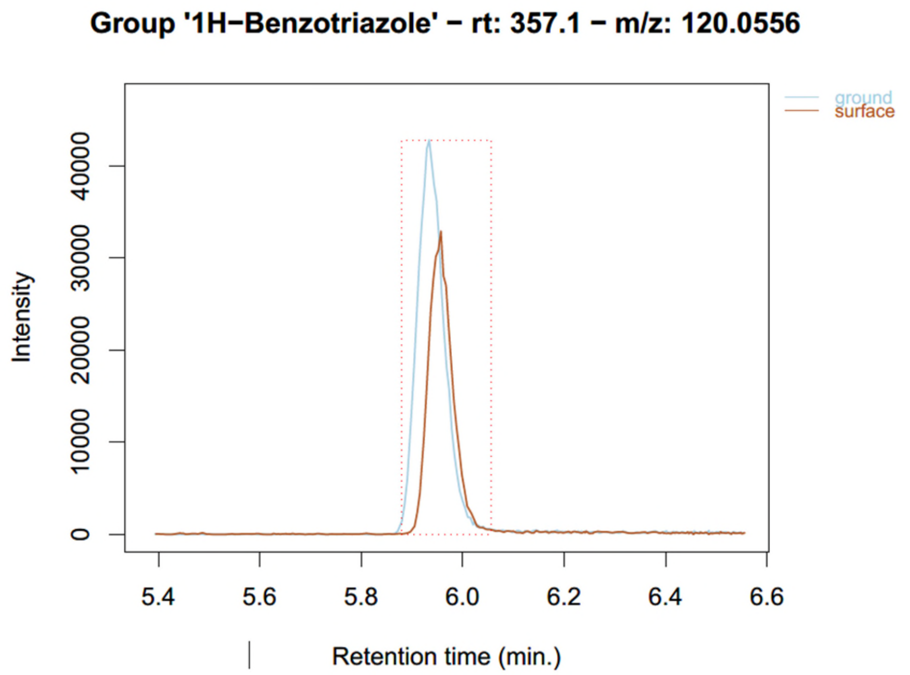
```
##     explainedPeaks      score neutralMass
## 1:               2 12.0000000   452.1115   COC1=CC(OC)=NC(NC(=O)NS(=O)(=O)C2=C(C=CC(NC=O)=C2)C(
## 2:               1  1.5779265   452.1108   OC1CC2=C(OC1C1=CC=C(O)C(O)=C1)C1=C(OC(=O)CC1C1=CC=C(O)C
## 3:               1  0.8431655  452.1092 FC1=CC=C(C=C1)C(=O)C1=CC=C(C=C1)C1=C(N=C(N1)C1=CC(Cl)=CC=C1
## 4:               0  0.6449141   452.1101          CC1=CC2=C(C=C1)C=C(CN(C1CCCCC1)C(=O)C1=CC=CC=
## 5:               1  1.0560344   452.1108   CC(=O)OC1=CC=CC=C1C(=O)OC1CC2=C(O)C=C(O)C=C2OC1C1=CC
## 6:               0  0.9080954   452.1100   ClC1C(Cl)C(=C(C2=CC=CC=C2)C2=CC=CC=C2)C1=C(C1=CC=CC=C1
## 7:               0  0.3971110   452.1130        S=C1NC2=CC=CC=C2C2=CC=CC=C2NC(=S)NC2=CC=CC=C2C
```

Below are some examples to plot data:

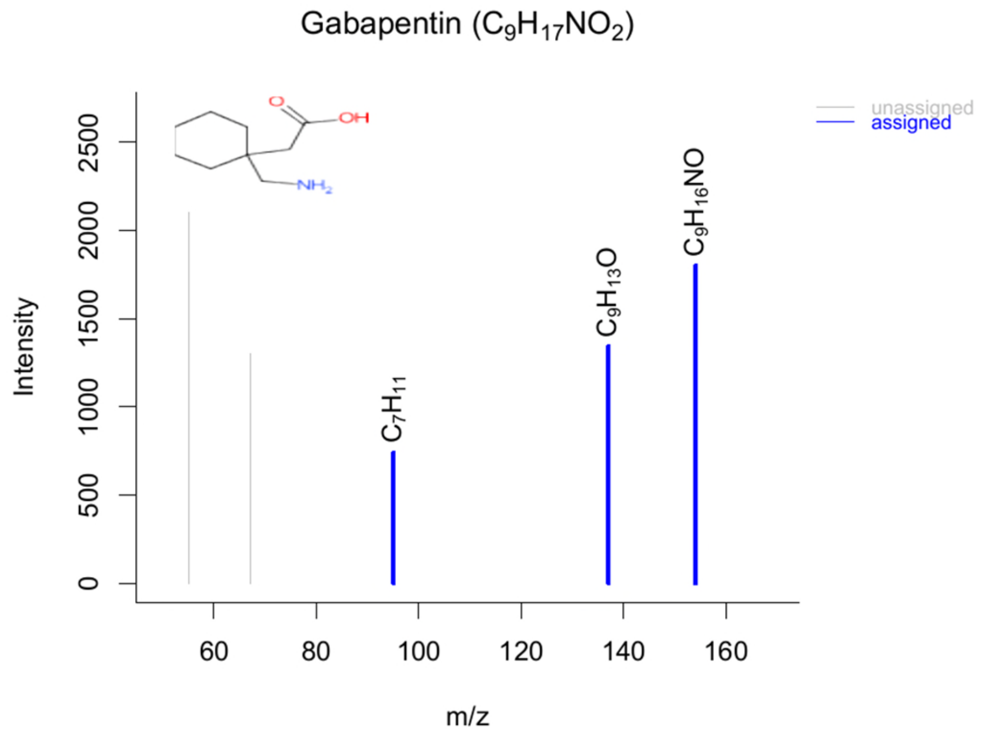
```
plotEIC(fGroups, colourBy = "fGroups", topMost = 1,
showPeakArea = TRUE, showFGroupRect = ,□ FALSE)
```



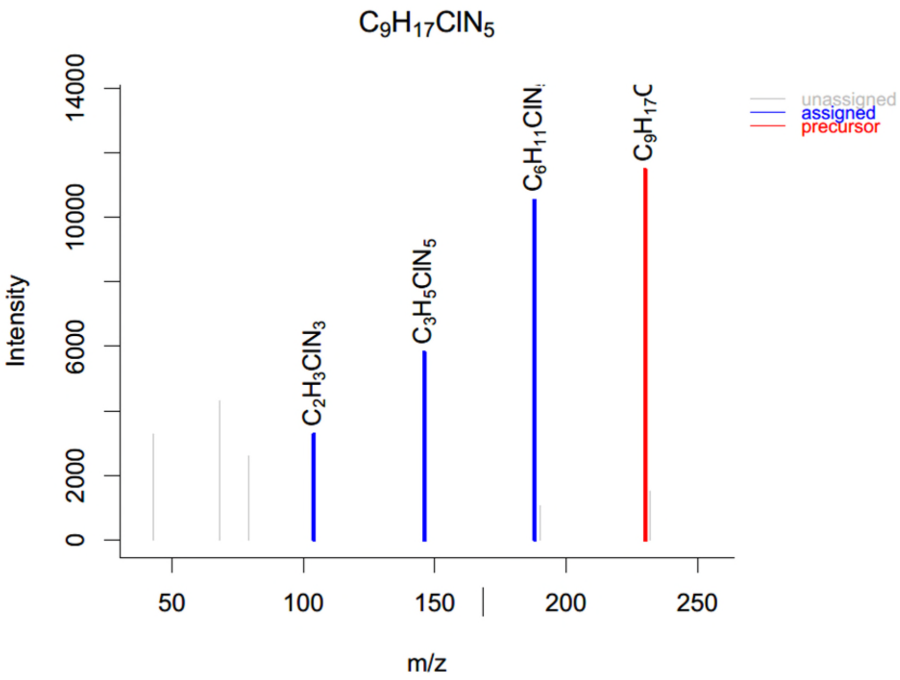
```
plotEIC(fGroups[, "1H-Benzotriazole"], colourBy = "rGroups", retMin =
TRUE)
```



```
plotSpec(compounds, 1, "Gabapentine", plists)
```



```
plotSpec(formulas, "C9H17ClN5", "Propazine", MSPeakLists = plists)
```



More documentation can be found within the R help functionality, e.g.

```
?plotEIC
?plotSpec
```

4.3.8.2 Optimizing feature finding parameters

You may have noticed that not all suspects were found (spoiler: they should all be present!). Try experimenting with different parameter settings to the `findFeatures()` function and observe the results. A summary of the settings used in this tutorial is shown below.

- `chromFWHM`: Expected full-width at height maximum (FWHM) of a chromatographic peak (in seconds)
- `minFWHM` and `maxFWHM`: Minimum and maximum FWHM values for a chromatographic peak (in seconds).

  Ensure low enough values for UHPLC!
- `noiseThrInt`: Absolute intensity cut-off. Data with intensities below this value are ingored.
- `chromSNR`: Minimum S/N ratio of a feature.

Note that each time you change a setting you have to re-run all the feature code and everything that follows (suspect screening, annotation etc). If you want to speed up the optimization process consider not updating and reporting the formula and compound annotation results. To do so, only execute everything from `findFeatures()` to `groupFeaturesScreening()` (i.e. where features are generated and converted to suspects) and run the following command afterwards to generate a simplified report which only contains feature data:

```
reportMD(fGroups)
```

4.3.8.3 Using the PubChem compound database

In this tutorial we have used to CompTox database to find candidate structures. While this database is highly applicable to environmental screening, a more thorough compound search can be performed by using the Pubchem database. This is simply done by changing the `database` parameter of the `generateCompounds()` function:

```
compounds <- generateCompounds(fGroups, plists, "metfrag", method = "CL",

                               dbRelMzDev = 5, fragRelMzDev = 5,
                               fragAbsMzDev = 0.002,

                               adduct = "[M+H]+", database = "pubchem",

                               ,□ maxCandidatesToStop = 5000)
```

Note that in the above example the `maxCandidatesToStop` parameter was lowered to *5000*: this parameter tells that a compound search should be aborted after more than a defined number of candidates (i.e. *5000*) are found. The reason for this is that, unlike the CompTox database, the compounds are searched through an online database and processing thousands of candidates takes too much time (and is not nice towards the servers running MetFrag!). An obvious drawback is that you will not get any results for features with many candidates (however, these are generally difficult to sort out anyway!).

For this same reason you may get several warnings about that results for some features could not be obtained. (See the log files in the `log` subdirectory to verify this.)

After you have generated the compounds using the PubChem database you can re-run the reporting functions to see the new results.

5 Scouting and conference report

5.1 ACS Spring meeting 2018

The American Chemical Society's spring meeting 2018 took place March 18 to 22 in New Orleans, USA and focused on chemistry-related information and research, in particular on the "Nexus of Food, Energy and Water". Of major interest to the NTS project were the sessions titled "Accurate Mass/High Resolution Mass Spectrometry for Environmental Monitoring and Remediation", and "Cheminformatics Resources & Software Tools Supporting Environmental Chemistry". The HRMS session covered the advancements in high-resolution mass spectrometry (HR/MS) instruments and software that enable identification and measurement of numerous organic chemicals in the environment. Presentations dealt with the identification of contaminants, process efficiency of treatment operations, and measurement of transformation products. In addition, toxicological and computational tools that can handle the enormous amounts of data produced by these techniques. The Cheminformatics session emphasized the enormous impact that computational chemistry has had in regards to providing environmental chemists access to data, information and software tools and algorithms. It consisted of a series of talks that provided an overview of the present state of data, tools, databases and approaches available to environmental chemists. Andrea M Brunner presented her work on "Prioritizing anthropogenic chemicals in drinking water sources through combined use of mass spectrometry based exposure data and ToxCast toxicity data" during this session (see abstract below).

ABSTRACT SYMPOSIUM NAME: Cheminformatics Resources & Software Tools Supporting Environmental Chemistry (Oral)

ABSTRACT SYMPOSIUM PROGRAM AREA NAME: CINF

CONTROL ID: 2854561

TITLE: Prioritizing anthropogenic chemicals in drinking water sources through combined use of mass spectrometry based exposure data and ToxCast toxicity data

AUTHORS (FIRST NAME, LAST NAME): Andrea M. Brunner[1], Milou M. Dingemans[1], Kirsten A. Baken[1], Annemarie P. van Wezel [1, 2]

INSTITUTIONS (ALL):

[1] Chemical Water Quality and Health, KWR Watercycle Research Institute, Nieuwegein, Netherlands, Netherlands.

[2] Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, Utrecht, Netherlands.

ABSTRACT BODY:

Abstract: Advancements in high-resolution mass spectrometry (HRMS) based screening methods have enabled a shift from target to non-target analyses to detect chemicals in water samples. The multitude of suspect chemicals resulting from such non-target screenings need to be prioritized for further identification and potential inclusion into monitoring programs. Here, we compare a strategy developed for the prioritization of chemicals in Dutch raw and drinking water samples based on semi-quantitative exposure data from HRMS (Sjerps et al. 2016) to a strategy based on high-throughput in vitro toxicity data. Between 2007-2014, 151 Dutch water samples, including waste water treatment plant effluent, surface water, ground water and drinking water, were collected. HRMS non-target screening analyses detected >7000 structures in these samples which could be linked to >1000 suspects from a curated suspect list of >5000 EU and water relevant chemicals. These suspects were subsequently prioritized based on exceedance of the Threshold of Toxicological Concern (TTC). Here, rather than using this discrete scale, we ranked suspects based on their semi-quantitative total concentration, expressed as internal standard equivalents. We then compared both the TTC prioritization and the continuous ranking of the chemicals to a prioritization based on chemical-specific in vitro toxicity data from the publicly available EPA ToxCast database. Using the 5th percentile of the AC50values from all ToxCast assays in a hypothesis-free approach, and from assays considered most water relevant in parallel, >500 suspects could be ranked based on their toxicity with respect to a total of >1000 different assay endpoints. The comparison showed that different prioritization strategies resulted in a different ranking of suspect chemicals. We therefore propose a novel prioritization scheme that combines both exposure and toxicity data and takes advantage of their complementarity to prioritize suspects in water samples.

5.2 Crash course in Cheminformatics Summer School Strasbourg

To enhance understanding and subsequently implementation of cheminformatics tools at KWR a crash course in Cheminformatics was attended at the University of Strasbourg. This course was a pre-conference session of the Cheminformatics Summer School and took place on June 25. The lectures and tutorials covered Computer representation of molecular structures such as molecular descriptors and fingerprints, chemical databases including database creation, and QSPR approaches.

5.3 IMSC Florence 2018

The 22nd International Mass Spectrometry Conference (IMSC) 2018 took place in Florence, IT , and covered all aspects of mass spectrometry, from fundamentals to instrumentation and applications. Some of the highlights of the conference were the advancements in ion spectroscopy and photodissociation that can help in identifying small molecules in cases when the more traditional non-target screening mass spectrometry approaches fail. Jos Oomen, Radboud University and his team use infrared ion spectroscopy to structurally characterize unknown molecules – a promising novel technique also for water analysis. Other advancements with high potential for the environmental sector were EIEIO presented by Gerard Hopfgartner, University Geneva, the new AquireX software to improve identification rates from Thermo Scientific and the machine learning approaches for metabolomics data analysis presented by Pierre-Marie Allard, CNRS. In addition, new techniques and methodologies from the environmental, food and petroleum sector were presented that could be applied in water research as well. During the workshop “Environmental Mass Spectrometry: from trace analysis to effect assessment” Annemieke Kolkman’s (KWR) talk on HILIC mass spectrometry for the analysis of polar micro-pollutants in drinking water and its sources complemented the presentations on effect directed analysis by Marc Suter and Marja Lamoree. Andrea M Brunner presented a poster at IMSC the abstract of which can be found below.

Title: Transformation product formation in drinking water treatment

Keywords: Non-target screening, transformation products, mass spectrometry, drinking water, water technology

Introduction: Transformation products (TPs) are formed in the water cycle through biological and technological processes. Despite their potentially altered toxicity compared to parent compounds, TPs formed by drinking water treatment are not routinely monitored and remain elusive. This is mainly due to the technical challenges in analyzing the often unknown, low concentration compounds. Their analysis requires non-target HRMS/MS methods and novel data analysis approaches.

Methods: Here, we performed lab scale experiments to monitor TP formation of the three organic micropollutants carbamazepine, clofibric acid and metolachlor during rapid sand filtration and ozonation, two readily applied biotic and abiotic drinking water treatments, respectively. TP identification from non-target data was facilitated through prediction of potential TPs based on literature and models, halogenated and/or isotopically labeled parent compounds.

Results: The experimental results showed that the degradation of parent compounds did not per se lead to mineralization of the compound, but rather to an abundance of TPs, in often low concentrations. Some of these TPs were bigger and less polar than their parent compounds, which was somewhat unexpected. The identification of peaks representing TPs was straightforward and semi-automatic with the developed workflow, based on statistical testing and peak area filters. The suspect screening based on TP suspect lists manually curated from literature mining and prediction tools was efficient for TPs in the lists. However, the majority of TPs identified did not match suspect list entries. Furthermore, the structural identification of these features, as well as of isobaric suspects remained labor and time intensive.

Conclusions: The majority of TPs remained structurally unidentified, and for the majority of identified TPs toxicological risk assessment was missing. Follow-up work should target and hopefully alleviate these issues. Finally, the developed workflow can be applied to pilot-scale experiments to allow TP monitoring in actual drinking water production.

Novel Aspect: develop and test an efficient workflow to monitor TP formation and identify drinking water treatment specific TPs on a lab-scale

6 Acknowledgments

We would like to thank Annemarie Toebak and Runa Kooper-Mookerji (AquaLab Zuid), Nikki Janssens (deWatergroep), Rob ten Broek and Eelco Pieke (HWL), and Jan van der Kooi and Milan Verwoert (WLN) for their contributions and participation at the final workshop; Emma Schymanski for advice and support; DPWE and HWL for providing NTS data for data analysis.

7 References

Aalizadeh, R., Thomaidis, N.S., Bletsou, A.A., Gago-Ferrero, P., 2016. Quantitative Structure-Retention Relationship Models To Support Nontarget High-Resolution Mass Spectrometric Screening of Emerging Contaminants in Environmental Samples. J Chem Inf Model 56, 1384-1398.

Allen, F., Pon, A., Wilson, M., Greiner, R., Wishart, D., 2014. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. Nucleic Acids Res 42, W94-99.

Alonso, A., Julia, A., Beltran, A., Vinaixa, M., Diaz, M., Ibanez, L., Correig, X., Marsal, S., 2011. AStream: an R package for annotating LC/MS metabolomic data. Bioinformatics 27, 1339-1340.

Bade, R., Bijlsma, L., Sancho, J.V., Hernández, F., 2015. Critical evaluation of a simple retention time predictor based on LogKow as a complementary tool in the identification of emerging contaminants in water. Talanta 139, 143-149.

Bonner, R., Hopfgartner, G., 2018. SWATH data independent acquisition mass spectrometry for metabolomics. TrAC Trends in Analytical Chemistry.

Brodbelt, J.S., 2014. Photodissociation mass spectrometry: new tools for characterization of biological molecules. Chem Soc Rev 43, 2757-2783.

Broeckling, C.D., Afsar, F.A., Neumann, S., Ben-Hur, A., Prenni, J.E., 2014. RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data. Analytical Chemistry 86, 6812-6817.

Domingo-Almenara, X., Montenegro-Burke, J.R., Benton, H.P., Siuzdak, G., 2018. Annotation: A Computational Solution for Streamlining Metabolomics Analysis. Analytical Chemistry 90, 480-489.

Duhrkop, K., Fleischauer, M., Ludwig, M., Aksenov, A.A., Melnik, A.V., Meusel, M., Dorrestein, P.C., Rousu, J., Bocker, S., 2019. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods 16, 299-302.

Dührkop, K., Shen, H., Meusel, M., Rousu, J., Böcker, S., 2015. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proceedings of the National Academy of Sciences 112, 12580-12585.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M.Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., Nishioka, T., 2010. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 45, 703-714.

Kasper, P.T., Rojas-Chertó, M., Mistrik, R., Reijmers, T., Hankemeier, T., Vreeken, R.J., 2012. Fragmentation trees for the structural characterisation of metabolites. Rapid Communications in Mass Spectrometry 26, 2275-2286.

Kuhl, C., Tautenhahn, R., Bottcher, C., Larson, T.R., Neumann, S., 2012. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem 84, 283-289.

Mahieu, N.G., Spalding, J.L., Gelman, S.J., Patti, G.J., 2016. Defining and Detecting Complex Peak Relationships in Mass Spectral Data: The Mz.unity Algorithm. Anal Chem 88, 9037-9046.

Martens, J., Berden, G., Bentlage, H., Coene, K.L.M., Engelke, U.F., Wishart, D., van Scherpenzeel, M., Kluijtmans, L.A.J., Wevers, R.A., Oomens, J., 2018. Unraveling the unknown areas of the human metabolome: the role of infrared ion spectroscopy. J Inherit Metab Dis 41, 367-377.

Martens, J., Berden, G., van Outersterp, R.E., Kluijtmans, L.A.J., Engelke, U.F., van Karnebeek, C.D.M., Wevers, R.A., Oomens, J., 2017. Molecular identification in metabolomics using infrared ion spectroscopy. Sci Rep 7, 3363.

McEachran, A.D., Sobus, J.R., Williams, A.J., 2017. Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. Analytical and Bioanalytical Chemistry 409, 1729-1735.

Meringer, M., Reinker, S., Zhang, J., Muller, A., 2011. MS/MS data improves automated determination of molecular formulas by mass spectrometry. MATCH Commun. Math. Comput. Chem. 65, 259-290.

Morrison, L.J., Parker, W.R., Holden, D.D., Henderson, J.C., Boll, J.M., Trent, M.S., Brodbelt, J.S., 2016. UVliPiD: A UVPD-Based Hierarchical Approach for De Novo Characterization of Lipid A Structures. Analytical Chemistry 88, 1812-1820.

Pervukhin, A., Letzel, M.C., Böcker, S., Lipták, Z., 2008. SIRIUS: decomposing isotope patterns for metabolite identification†. Bioinformatics 25, 218-224.

Ruttkies, C., Schymanski, E.L., Wolf, S., Hollender, J., Neumann, S., 2016. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminform 8, 3.

Schollee, J.E., Schymanski, E.L., Stravs, M.A., Gulde, R., Thomaidis, N.S., Hollender, J., 2017. Similarity of High-Resolution Tandem Mass Spectrometry Spectra of Structurally Related Micropollutants and Transformation Products. J Am Soc Mass Spectrom 28, 2692-2704.

Schymanski, E.L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H.P., Hollender, J., 2014. Identifying small molecules via high resolution mass spectrometry: communicating confidence. Environ Sci Technol 48, 2097-2098.

Schymanski, E.L., Ruttkies, C., Krauss, M., Brouard, C., Kind, T., Dührkop, K., Allen, F., Vaniya, A., Verdegem, D., Böcker, S., Rousu, J., Shen, H., Tsugawa, H., Sajed, T., Fiehn, O., Ghesquière, B., Neumann, S., 2017. Critical Assessment of Small Molecule Identification 2016: automated methods. Journal of Cheminformatics 9, 22.

Schymanski, E.L., Williams, A.J., 2017. Open Science for Identifying "Known Unknown" Chemicals. Environ Sci Technol 51, 5357-5359.

Zhou, B., Xiao, J.F., Tuli, L., Ressom, H.W., 2012. LC-MS-based metabolomics. Mol Biosyst 8, 470-481.