Retrospective Non-target Analysis to Support Regulatory Water Monitoring: From Masses of Interest to Recommendations via in silico workflows

Adelene Lai (adelene.lai@uni.lu)

University of Luxembourg Luxembourg Centre for Systems Biomedicine: Universite du Luxembourg Luxembourg Centre for Systems Biomedicine https://orcid.org/0000-0002-2985-6473

Randolph R. Singh

LCSB: Universite du Luxembourg Luxembourg Centre for Systems Biomedicine

Lubomira Kovalova

Amt fuer Abfall, Wasser, Energie und Luft, Baudirektion Kanton Zuerich, Switzerland

Oliver Jaeggi

Amt fuer Abfall, Wasser, Energie und Luft, Baudirektion Kanton Zuerich, Switzerland

Todor Kondic

LCSB: Universite du Luxembourg Luxembourg Centre for Systems Biomedicine

Emma L. Schymanski

LCSB: Universite du Luxembourg Luxembourg Centre for Systems Biomedicine

Research

Keywords: Non-target analysis, suspect screening, retrospective, wastewater, micropollutants, cheminformatics, identification, monitoring, regulation

Posted Date: December 31st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-136443/v1

Version of Record: A version of this preprint was published at Environmental Sciences Europe on April 4th, 2021. See the published version at https://doi.org/10.1186/s12302-021-00475-1.

Abstract

Background

Applying non-target analysis (NTA) in regulatory environmental monitoring remains challenging - instead of having exploratory questions, regulators usually already have specific questions related to environmental protection aims. Additionally, data analysis can seem overwhelming because of the large data volumes and many steps required. This work aimed to identify environmental chemical unknowns via retrospective NTA within the scope of a pre-existing Swiss environmental monitoring campaign focusing on industrial chemicals. The research question addressed immediate regulatory priorities: identify pollutants with industrial point sources occurring at the highest intensities over two time points. Samples from 22 wastewater treatment plants obtained in 2018 and measured using liquid chromatography-high resolution mass spectrometry were retrospectively analysed by i) performing peak-picking to identify masses of interest, ii) prescreening and quality-controlling spectra, and iii) tentatively identifying priority "known unknown" pollutants by leveraging environmentally-relevant chemical information provided by Swiss, Swedish, EU-wide, and American regulators. This regulator-supplied information was incorporated into MetFrag, an *in silico* identification tool replete with "post-relaunch" features used here. This study's unique regulatory context posed challenges in data quality and volume that were directly addressed with the prescreening, quality control, and identification workflow developed.

Results

One confirmed and twenty-one tentative identifications were achieved, suggesting the presence of compounds as diverse as manufacturing reagents, adhesives, pesticides, and pharmaceuticals in the samples. More importantly, an in-depth interpretation of the results in the context of environmental regulation and actionable next steps are discussed. The prescreening and quality control workflow is openly accessible within the R package Shinyscreen, and adaptable to any (retrospective) analysis requiring automated quality control of mass spectra and non-target identification, with potential applications in environmental and metabolomics analyses.

Conclusions

NTA in regulatory monitoring is critical for environmental protection, but bottlenecks in data analysis and results interpretation remain. The prescreening and quality control workflow, and interpretation work performed here are crucial steps towards scaling up NTA for environmental monitoring.

Background

Organic pollutants are well-documented in aquatic environments (Sousa et al. 2018). Traditionally, target strategies that look for chemicals known in advance have been used to identify these compounds (Krauss et al. 2010). In contrast, non-target analysis (NTA) helps discover previously undetected, unexpected and/or unknown substances. NTA has been under intense development in recent years, aided by advances in instrumentation and computational approaches (Krauss et al. 2010; Hollender et al. 2017). Considering the vast chemical space of possible environmental pollutants (Wang et al. 2020), the need for NTA is becoming more pressing in order to tackle the growing challenge of identifying chemical unknowns in samples. Yet, data analysis in NTA remains a formidable challenge. To ease the "identification burden" in NTA, simplifying approaches like Suspect Screening, where chemicals on discrete lists suspected to be present in the sample are screened, are being taken in the interim (Hollender et al. 2017).

Various successful examples of NTA (Schwarzbauer and Ricking 2010; Hug et al. 2014; Ruff et al. 2015; Carpenter et al. 2019; Albergamo et al. 2019; Sun et al. 2019; Lara-Martín et al. 2020; Beckers et al. 2020) have inevitably encouraged interest in its potential role to monitor and manage chemical pollutants in the environment (Hollender et al. 2017). As the field matures, there is some consensus that NTA is “Ready to Go”, with calls for it to be applied more widely within the regulatory frameworks of local, regional, and national authorities (Hollender et al. 2017, 2019). Data-mining routines like enviMass have contributed to such initiatives (Loos et al. 2018); enviMass facilitates NTA by peak-picking and prioritising unknown features of interest worthy of further identification efforts. It does so by connecting mass spectral features based on criteria such as having signals of sufficient intensity, grouping together isotopologues and adducts of the same component, and detecting temporal trends, ultimately giving as output a list of *m/z*-retention time pairs, plus accompanying information for further identification efforts.

However, challenges for regulators to perform NTA persist, particularly with respect to high-throughput data analysis and identification following the mass prioritization and peak-picking steps described above. For example, regulators may lack specific NTA expertise and/or resources to apply the potentially many and complicated computational workflows (Helmus et al. 2020; Ljoncheva et al. 2020) available for analysing the copious amounts of data. In addition to the time-consuming and complex nature of data interpretation, issues related to standardisation and reproducibility exist, as there is currently no ‘one size fits all’ approach to identifying compounds using NTA (Hites and Jobst 2018). As a result, NTA is currently often considered by regulators as “too much effort for too little sound evidence”.

Another more systemic obstacle to applying NTA in a regulatory context relates to the divergent interests of scientists in academia, who are (currently) responsible for driving most NTA developments, and scientists in regulatory practice, who would implement these developments towards regulatory compliance and environmental protection. While the former aim often to develop and publish novel work, the primary mandate of the latter is regulatory compliance towards environmental protection. One possible consequence of this reality is that academic research outcomes resulting from NTA may not be directly relevant or in a form that is readily usable for regulators. In other words, researchers’ questions may not be regulators’ questions - what is possibly *scientifically* interesting may not be of priority or directly useful to regulators.

Despite these aforementioned challenges, it is possible (and important) to navigate both research and regulatory needs in NTA. The present work is an example of academic research driven primarily by regulatory priorities. In this “top-down” approach, pre-existing data were used to generate results of direct environmental relevance and with immediate implications for environmental management.

Three practical challenges characteristic of applying NTA in a regulatory environmental monitoring context arose in this study: i) the study was framed by superlative questions that required a large volume of data to be analysed *i.e.*, identify unknown compounds occurring at the *highest intensities* and *highest temporal frequency* with *point sources* across all the samples of the sampling campaign; ii) there was a strict and limited timeframe allowed for the study following project management procedures of the regulatory body, and iii) the data originally collected had been repurposed for this NTA study as there was no capacity nor further resources available within the scope of the project to do additional measurements. The latter point was all the more critical as preliminary manual inspection of the available data revealed that not all measurements were fully suitable for the intended non-target identification. These challenges called for a high-throughput approach capable of processing large volumes of data of variable quality in a fast and reproducible way that would be compatible with identification approaches downstream. Additionally, unlike the seemingly increasing complexity of existing workflows (Ljoncheva et al. 2020), an uncomplicated and

'minimal, bare-bones' but fully functional approach that is transparent and easily explainable is critical given the regulatory context.

MetFrag, used in this work to support identification efforts, is an example of an open *in silico* identification approach which satisfies the aforementioned criteria. Released in 2010 (Wolf et al. 2010), it first retrieves potential candidates with matching mass from compound databases such as PubChem (Kim et al. 2019) (> 111 million compounds, December 2020), ChemSpider (Pence and Williams 2010) (> 97 million compounds, December 2020), or smaller biological databases like the Human Metabolome Database (Wishart et al. 2018) (114,214 compounds in version 4.0, July 2020). These candidates are then scored according to how well the experimental spectrum matches the *in silico* fragments generated per candidate using a bond dissociation approach (Wolf et al. 2010), and subsequently ranked according to this FragmenterScore (sometimes referred to as the Fragmentation Score or FragScore, or simply the MetFrag Score when it is the only component thereof). For the identification of environmental "known unknowns", using fragmentation information alone in this way can give mediocre results (*e.g.*, ~ 22% and 6% of 473 environmentally-relevant standards ranked first with ChemSpider and PubChem respectively (Ruttkies et al. 2016)). This outcome may have various causes: (i) the search databases used are too large and/or do not contain only environmentally relevant compounds, therefore resulting in too many candidates that are not meaningful, and/or (ii) there is simply not enough information to distinguish candidates when considering their fragmentation alone.

To address these limitations, MetFrag was 'relaunched' in 2016 to incorporate further identification strategies beyond fragmentation, such as retention time information, substructure in/exclusion, availability of literature and patent information, presence/absence in suspect lists, and user-defined scoring terms (Ruttkies et al. 2016). Over time, spectral similarity comparison with spectra from the MassBank of North America (MoNA) (Fiehn Lab) with and without a MetFusion approach (Gerlich and Neumann 2013) was also integrated into MetFrag. Since then, two further open-science/environmental chemistry developments have contributed significantly to MetFrag's extended capabilities for identifying environmental unknowns. Firstly, the release and integration of the United States Environmental Protection Agency's CompTox Chemicals Dashboard (Williams et al. 2017) (hereafter, "CompTox") into MetFrag provides a search database of > 850,000 compounds of environmental and toxicological relevance (Schymanski 2019), while allowing users to leverage the "MS-Ready" concept (McEachran et al. 2018) and various forms of chemical metadata availability in CompTox as user-defined scoring terms. Secondly, critical information from international regulatory bodies can now be exploited through MetFrag towards identifying environmental chemicals. Beyond (i) the US EPA's Chemicals and Products database (CPDat) that is already integrated via CompTox (US EPA 2016), MetFrag's user-defined scoring terms can also be configured to incorporate information on (ii) hazard and exposure from the Swedish Chemicals Agency KEMI (Fischer 2017), (iii) European chemicals registration *i.e.*, REACH (Alygizakis and Slobodnik 2018), and (iv) the NORMAN Network's merged suspect list of chemicals of emerging concern known as SusDat (NORMAN Network et al. 2020), representing knowledge gathered from > 70 regulatory and academic reference laboratories throughout Europe and North America. Used in this way, MetFrag connects disparate resources from various regulatory agencies and academic researchers towards identifying environmental unknowns, practically 'helping researchers and regulators help each other' by providing an interconnected information platform with identification functionality.

Since MetFrag's relaunch in 2016, work on the identification of environmental unknowns has used MetFrag's post-relaunch functionality to varying extents. Some research simply uses MetFrag purely for its *in silico* fragmentation capabilities, *i.e.*, not paired with any compound database (Choi et al. 2020; Purschke et al. 2020; Miaz et al. 2020). Many examples use only the FragmenterScore to rank candidates retrieved from ChemSpider alone (Li et al. 2017; Luft et al. 2017; Anliker et al. 2020), PubChem alone (Lege et al. 2019; Tian et al. 2020; Wagner et al. 2020), or a

combination of either or both with other databases (Chiaia-Hernández et al. 2017; Park et al. 2018; Oetjen et al. 2018; Köppe et al. 2020) like KEGG (Kanehisa and Goto 2000), FOR-IDENT (Letzel) and MassBank (MassBank Consortium and NORMAN Association). Several studies have begun to use one or more of MetFrag's post-relaunch capabilities such as data source, patent, and/or reference counts for the respective compound database used (Muz et al. 2017; Moschet et al. 2018; Veenaas et al. 2018; Carpenter et al. 2019; Faber et al. 2019; Beckers et al. 2020; Menger et al. 2021), spectral library similarity (Veenaas et al. 2018; Carpenter et al. 2019; Faber et al. 2019; Kandie et al. 2020; Beckers et al. 2020), and presence in suspect lists (Moschet et al. 2018; Carpenter et al. 2019; Lara-Martín et al. 2020). Albergamo and colleagues (Albergamo et al. 2019) were amongst the first to use MetFrag's post-relaunch capabilities heavily, in particular those provided via CompTox and by international regulators and scientists.

The present work aimed to exploit "post-relaunch" MetFrag and Open Science developments towards retrospectively identifying non-target environmental pollutants in a regulatory context, as summarised in Fig. 1. Here, pollutants of regulatory concern originating from industrial activities found in Swiss wastewater treatment plant (WWTP) effluents were the main subjects of this study. A prescreening and quality control workflow for high-throughput automated data processing was developed to analyse a provided list of unknown *m/z* prioritised by enviMass. The use of MetFrag in this work leverages the state-of-the-art open resources mentioned above, chief among them, regulatory information from multiple international sources, in addition to exploiting many of MetFrag's post-relaunch capabilities. The identifications provided by MetFrag were analysed with respect to the specific environmental regulatory context of this study and communicated using an established system of confidence levels, discussed in detail in the next section.

Methods

Sample Collection and Preparation

Daily water samples were collected from 25 sites based at 22 WWTPs distributed across Switzerland within sampling campaigns focusing on point sources of industrial chemicals. Of these 25 sampling sites, 22 sites are WWTP effluents (1 site per WWTP), while 3 sites are effluents of secondary clarifiers corresponding to 3 WWTPs which employ ozonation (whose effluent was also collected as above).

The samples were collected in two sampling campaigns: June and October 2018. Over seven consecutive days, 2 litres of the 24-hour flow proportional composite samples were collected daily at each sampling site. The sample was filled into two 1 litre glass bottles and kept closed at 4 °C until the last day of the respective sampling campaign. That day, all samples were transported cooled to an analytical laboratory and were filtered, flow-proportionally mixed, and sent cooled for MS-analysis. The final samples used for measurement were flow-proportional 7-day composites.

Sample Measurement

Samples were analysed without enrichment by direct injection of 100 µl into the chromatographic system. Chromatographic separation of the analytes was performed using a Waters Atlantis T3 column (150 × 3 mm, 3 µm particle size) connected to a Thermo Scientific Accela liquid chromatography system equipped with a 1250 pump, open autosampler, and Thermo Scientific Column Oven 300. The mobile phase eluent A consisted of ultrapure water (ELGA, 5 mM ammonium formate), while eluent B consisted of LC-MS grade methanol (Scharlau, 5 mM ammonium formate). The gradient programme started with 10% B, which was kept for 1 minute before a linear ramp to 95% B for 12 minutes. This condition was kept for 5 minutes before returning to starting mobile phase conditions at 18.5 min. The column was re-equilibrated for 4.5 minutes giving a total run time of 23 minutes with a flow rate of 300 µl/min.

A full scan single MS measurement was performed using a Thermo Scientific QExactive Orbitrap LC/MS system with resolving power of 70,000 (at *m/z* = 200) within 7 days of sample collection and preparation. A scan range of 100 to 1000 was used in both positive and negative electrospray ionisation modes. The samples were then stored at 4 °C.

Following the prioritisation of non-target masses (described in Part 1 of the prescreening workflow of the next section), the resulting list of non-target masses formed the inclusion list for a new round of MS/MS measurements of the same samples in February 2019. Normalised collision energy of 35 was used. The same measurement protocol as described above was applied with resolving power of 17,500 (at m/z = 200).

Computational Methods

Part 1 – enviMass prioritisation of masses of interest

enviMass was used to prioritise non-target masses of interest based on the following criteria: high intensity MS1 peaks (used as a proxy for high concentration), presumed point source (occurring at one or only a few sampling sites), multiple temporal occurrences across the sampling campaign, and existing isotopologue and adduct linkages. Initially, a list of 300 non-target masses of interest was identified and used as an inclusion list for MS/MS acquisition in the second round of measurements in February 2019 using the same samples that had been stored at 4 °C as described above. Of these 300 masses, 125 masses with associated $[M + H]^+$ and $[M-H]^-$ information from enviMass (117 and 8 respectively) were considered for further processing in the next step and constituted “List A”. A further 60 masses with associated $[M + H]^+$ and $[M-H]^-$ information (28 and 32 respectively) were also considered for the next step (“List B”), but had not been measured as part of the inclusion list. The enviMass parameters used to derive Lists A and B are detailed in the SI. These lists were the starting point for the workflows described here.

Part 2 – Prescreening and Quality Control Workflow

Data files in *.RAW* format were first converted to *.mzML* format using MSConvert from Proteowizard (Chambers et al. 2012) (v.3.0.19182-51f676fbe). The data were preliminarily inspected manually using XCalibur Qual Browser (v.4.2.28.14, Thermo Fisher Scientific, Waltham MA, USA). Then, a workflow to extract, prescreen, and quality control the spectra of the precursor masses in Lists A and B was developed and performed prior to further identification efforts.

The prescreening workflow first extracts all MS1 and MS2 ion chromatograms of each *m/z* from each *mzML* file supplied to it as input. No post-processing of mass spectral features is performed whatsoever during the extraction of spectra. Extracted MS1 precursors whose retention times are within 2 minutes of the mean retention time given by enviMass were deemed as matching the original list entries, considering possible drifts caused by wastewater matrix effects, unless specified otherwise.

A ‘case’ was defined as a measurement whose chromatograms and corresponding spectra have the same *m/z*, retention time, and file source (essentially, a single unique measurement). As part of the prescreening, each case was subject to quality control: the MS1 and MS2 ion chromatograms were checked *automatically* by an algorithm within the workflow in a stepwise fashion as per checks and thresholds 1–5 listed in Table 1. Failure to meet any of the criteria in the checks caused the case to be rejected from further identification efforts.

Table 1

Quality control checks within the prescreening workflow applied to the MS1 and MS2 spectral data for each case.

Quality Control Check	*Description*	*Positive Mode Threshold*	*Negative Mode Threshold*
1	Availability of MS1 precursor	Presence/Absence	
2	Minimum MS1 intensity	1×10^5	1×10^4
3	Maximum MS1 noise level	3x (average Baseline Intensity)	
4	Availability of MS2 corresponding to MS1 precursor	Presence/Absence	
5	MS1-MS2 alignment window	0.3 min (*i.e.*, +/- 0.15 min)	
6	Deduplication of cases	Highest MS1 intensity	
7	Minimum peak width and overall shape (manual QC)	0.1 min	

Thresholds apply to data measured using an Orbitrap instrument. Checks 1–5 are part of the automated prescreening workflow, while checks 6–7 were performed manually.

Cases that passed quality control checks 1–6 were manually inspected for peak shape and width (check 7, Table 1). Only cases that passed all quality control checks 1–7 were used as input for MetFrag identification in the next part of the workflow.

This prescreening workflow developed as part of this work has been embedded into the openly available R package Shinyscreen (Kondić et al.).

Part 3 – Identification using MetFrag

Tentative identification was performed using MetFrag (command line, version 2.4.5) (Wolf et al. 2010; Ruttkies et al. 2016). CompTox was used as the candidate database in the form of a local .csv file (Schymanski 2019). R scripts, building on the code bases of ReSOLUTION (Schymanski 2020a) and RchemMass (Schymanski 2020b), were written to accomplish the following steps.

First, the neutral monoisotopic mass corresponding to the $[M + H]^+$ or $[M-H]^-$ adducts indicated by enviMass in positive and negative mode respectively was calculated. Then, candidates of matching mass with a relative deviation of 5 ppm (selected to reflect the analytical mass error, also known as “Search ppm”) were retrieved from CompTox. Subsequently, candidates were fragmented *in silico* using the following fragmentation settings: Absolute Fragment Peak Match Deviation 0.001 Da (“Mzabs”), Relative Fragment Peak Match Deviation 5 ppm (“Mzppm”), and Maximum Tree Depth 2. Then, candidates were ranked according to the MetFrag Score, calculated as the sum of ten weighted scoring terms summarised in Table 2 and explained in detail below. These terms are either already built-in, or can easily be configured within MetFrag since its relaunch (Ruttkies et al. 2016). Candidates with identical first block InChIKeys (*i.e.*, stereoisomers, with the same structural skeleton) were grouped together.

Table 2
MetFrag scoring terms and weights used in tentative identification.

MetFrag Scoring Terms	*Weights*
Spectral Terms	
FragmenterScore	1.0
OfflineMetFusion	1.0
OfflineIndivMoNA	1.0
Total contribution to MetFrag Score:	*3.0*
Metadata Terms	
CPDAT_COUNT	1.0
DATA_SOURCES	1.0
KEMIMARKET_EXPO	1.0
KEMIMARKET_HAZ	1.0
NORMANSUSDAT	0.5*
REACH2017	0.5*
INDACT	1.0
Total contribution to MetFrag Score:	*6.0*
Maximum MetFrag Score	
Total	*9.0*
An asterisk (*) indicates these terms were given lower weights to avoid overweighting due to possible redundancy across the databases.	

Three scoring terms within the MetFrag Score reflect the contribution of the fragmentation spectra to the proposed identification: the FragmenterScore *(in silico* fragments explaining measured peaks, a function of peak count and bond dissociation energy), OfflineMetFusion (spectral similarity to entries in MassBank of North America (MoNA) using a MetFusion approach (Gerlich and Neumann 2013), and OfflineIndivMoNA (maximum spectral similarity with MoNA entries having exact InChIKey match). Four scoring terms relate to the availability of the chemical's metadata: CPDAT_COUNT (Williams et al. 2017) (number of entries within US EPA's Chemicals and Products database), DATA_SOURCES (Williams et al. 2017) (number of data sources underlying CompTox), KEMIMARKET_HAZ (Fischer 2017) (scaled and normalised hazard score calculated by the Swedish Chemicals Agency), and KEMIMARKET_EXPO (Fischer 2017) (scaled and normalised exposure score calculated by the Swedish Chemicals Agency KEMI). The remaining three terms account for the candidate's presence or absence in suspect lists, another form of metadata availability: INDACT (Industrial Activity chemicals known to be used near the sampling sites, supplied by the regulator), REACH2017 (Alygizakis and Slobodnik 2018) (chemicals registered under the European legislation framework REACH), and NORMANSUSDAT (chemicals in the merged NORMAN Suspect List Exchange (NORMAN Network) (v. 10/2019 used). All metadata scoring terms were weighted 1 except for REACH2017 and NORMANSUSDAT, which were both weighted 0.5 due to the high redundancy across the two databases.

To calculate the maximum possible MetFrag score, all the scoring terms except NORMANSUSDAT, REACH2017, INDACT, and OfflineIndivMoNA are first normalised to their respective largest values among the candidate set and scaled between 0–1. These normalised and scaled values are then summed together with the presence/absence scores of NORMANSUSDAT, REACH2017, and INDACT (0.5, 0.5, 1.0 if present, 0, 0, 0, if absent, respectively), and the similarity score from OfflineIndivMoNA (which is not scaled as it is already defined between 0 and 1).

Tentative identifications by MetFrag were communicated using an established system of levels (Schymanski et al. 2014), reiterated here with study-specific context for clarity: as MetFrag is an *in silico* method, it generally gives identifications of Level 3 confidence based on evidence for possible chemical structure using MS1, MS2 and experimental data/context. These identifications are tentative and require further validation before achieving higher confidence levels, as do Level 2a identifications of probable structure based on a library spectrum match, corresponding to a high MoNA individual similarity score (> 0.9) in the present work. Level 1 identifications require confirmation of the structure using a reference standard and includes target compounds.

Results

Prescreening and Quality Control

Preliminary manual inspection of the data using XCalibur Qual Browser (v.4.2.28.14, Thermo Fisher Scientific, Waltham MA, USA) indicated that not all measurements of each individual *m/z* were suitable for non-target identification because *e.g.*, MS1 precursors were often at low intensity, some MS2 spectra were absent, and spikes and/or noise were observed in the MS1 extracted ion chromatogram instead of actual peaks. Therefore, the prescreening workflow consisting of 7 quality control checks (Table 1) was implemented to isolate measurements that were suitable for non-target identification. Figure 2 provides examples of measurements visualised using Shinyscreen which passed all quality control checks (Panel A) and failed either one or more checks (Panels B-E) respectively. The latter were automatically eliminated from further consideration by the workflow because they were deemed unsuitable for use in non-target identification.

For identification, a total of 185 non-target *m/z* from both List A and List B were prescreened in each of the 30 *mzML* files, resulting in 5,550 cases possible for identification. For List A containing 117 *m/z* measured in positive mode, the prescreening workflow runtime was approximately 8 hours on a laptop machine with 8 GB RAM and 2 physical cores over all 30 mzML files. Runtime was estimated based on timestamps from results file generation.

Of the 5,550 cases, 899 cases satisfied checks 1–5 listed in Table 1. Duplicate cases by *m/z* (*e.g.*, if it was detected at more than one site) were eliminated by prioritising those with the highest MS1 intensity (check 6), leaving 157 cases (approximately 0.03% of total cases) to be manually inspected for peak width and shape (check 7). Of these 157 cases, only 22 passed manual inspection and qualified for further identification efforts using MetFrag (listed in full in SI Table S2). Figure 3 summarises this data reduction outcome as a result of quality control within the prescreening workflow.

Tentative Identification using MetFrag

Tentative identifications for the 22 *m/z* that passed quality control checks were obtained using MetFrag. Candidates for each *m/z* were proposed as ranked lists according to their respective MetFrag Scores comprising the ten scoring terms described in Table 2 (full MetFrag results with lists of ranked candidates available in MassIVE). Figure 4 shows the distribution of MetFrag scores classified into tertiles for the *top-ranked candidate for each of the 22 m/z*.

Interpretation of MetFrag Results

Given the background and context of this work (*i.e.*, NTA in environmental monitoring to identify high-priority unknowns), the MetFrag results described above do not represent a satisfactory end-point/end-product of this study. In other words, it does not suffice to present MetFrag's outputs (lists of ranked candidates, one list per *m/z*) alone, as these results alone do not provide sufficient direction for the next regulatory steps. Rather, it is crucial that these scientific outcomes are translated into transparent and actionable information for regulatory scientists to aid their future decision-making with respect to the following questions:

1. What does the distribution of MetFrag Scores mean and what are the implications?
2. How can this information guide evidence-based decision-making regarding further identification efforts? (*e.g.*, by adding candidates to suspect lists for future Suspect Screenings, purchasing reference standards for confirmation etc.)

The following section addresses these two questions through in-depth interpretation of MetFrag's results at two levels: at a global level across all 22 *m/z* studied, and at a candidate level per *m/z* respectively. The aim of these interpretations is to deliver information based on scientific premises that is actionable from a regulatory point of view and in doing so, present 'complex' MetFrag results in an interpretable way using Scenario Analysis.

Regarding the MetFrag Scores of the top candidates for each *m/z* (Fig. 4), this distribution arises as a result of four possible combinations of Spectral and Metadata Score components contributing toward the final MetFrag Score (Table 3).

Table 3
Four different scenarios corresponding to the four possible combinations of Spectral and Metadata Scores.

	High Metadata Score	**Low Metadata Score**
High Spectral Score	Scenario 1: High MetFrag Score (> 6)	Scenario 3: Moderate MetFrag Score (3–6)
Low Spectral Score	Scenario 2: Moderate MetFrag Score (3–6)	Scenario 4: Low MetFrag Score (< 3)

*Spectral and Metadata Scores are components of the final MetFrag Score (*Table 2*). Scores falling into the different tertiles of the MetFrag Score distribution are classified as Low, Moderate, and High respectively, as indicated in Fig. 4.*

The first combination, called Scenario 1, features both strong spectral and metadata evidence supporting a given candidate, resulting in a High MetFrag Score. Moderate MetFrag Scores result when one of these two scoring components, Spectral or Metadata, is low and the other is high, leading to Scenarios 2 and 3. Finally, Scenario 4 describes situations where both Spectral and Metadata scores are low, resulting in Low MetFrag Scores. Table 4 shows the breakdown of the MetFrag Score into its component Spectral and Metadata terms for four illustrative examples, one for each scenario. These representative examples were selected from the distribution (Fig. 4) and are the respective top-ranked candidates for 4 *m/z.*

Table 4
MetFrag Score breakdown for the top candidates of four m/z.

	MetFrag Score (Weighted)			
	7.00	**4.63**	**2.95**	**2.50**
MetFrag Score Distribution Classification	High (> 6)	Moderate (3–6)	Moderate (~ 3–6; borderline)	Low (< 3)
Scenario	Scenario 1 – *High* Spectral and Metadata Scores	2 – *Low* Spectral and *High* Metadata Scores	3 – *High* Spectral and *Low* Metadata Scores	4 - *Low* Spectral and Metadata Scores
m/z	278.1062	187.0938	152.0198	199.1050
MetFrag Score Breakdown (top candidate only)				
Spectral Terms (Raw Scores)				
FragmenterScore	95.30	7.88	217.84	19.48
OfflineMetFusion	4.64	0.88	2.06	2.81
OfflineIndivMoNA	1.00	0	0	0
Metadata Terms (Raw Scores)				
CPDAT_COUNT	0	0	0	0
DATA_SOURCES	47	42	1	1
KEMIMARKET_EXPO	16	11	0	0
KEMIMARKET_HAZ	9	2	0	0
NORMANSUSDAT	1	1	0	0
REACH2017	1	1	0	0
INDACT	0	0	0	0

Each MetFrag Score here represents one of the four scenarios in Table 3.

The implications of this distribution (Fig. 4) can guide future actions depending on whether depth or breadth of the NTA study is more important. For example, if the ultimate goal is to fully identify one or two high-priority non-target unknowns to Level 1 confidence, pursuing candidates with High MetFrag Scores (3rd tertile, dark red region in Fig. 4, Scenario 1 in Table 3) is recommended. Alternatively, if gaining a wide survey of the possibly relevant but as yet unknown environmental pollutants throughout the sampling campaign is preferred (akin to a 'first-approximation' of the situation), then even candidates with Moderate and/or Low scores can also be considered further depending on the relevance of the scoring terms to the context. Additionally, further decisions on future actions can be made based on possible limitations of the study which may be known from the outset (see Discussion).

Close inspection of the MetFrag Score, namely its component Spectral and Metadata scoring terms, enables results interpretation on the individual candidate level for each *m/z*. Irrespective of whether a breadth or depth strategy is chosen, the lists of ranked candidates should always be scrutinised for plausibility because although each identification has a top candidate ranked first by MetFrag, the top candidate may not be the only candidate worth considering (if at all) given the context of the study. Below, an in-depth analysis and results interpretation of the top 4 candidates for selected *m/z* is presented in the following tables as examples of each of the scenarios (Table 3). Distributed Structure-Searchable Toxicity Substance Identifiers from CompTox, known as DTXSIDs are given as identifiers. The choice to use DTXSID as candidate identifiers and not their compound names is addressed in the Discussion.

m/z 278.1062

Scenario 1: High Spectral and Metadata Scores (High MetFrag Score; >6)

Thirty-three compounds with matching mass were retrieved from CompTox and scored by MetFrag using the ten scoring terms (Table 2). The top-ranked candidate, DTXSID4058156, has the highest total MetFrag Score out of all the candidates proposed (Table 5). In terms of spectral information, it has the highest FragmenterScore and OfflineMetFusion score of all the candidates, as well as a MoNA library match of 0.998, while all other candidates had a MoNA library match of 0.

In terms of metadata and presence in suspect lists, DTXSID4058156 has abundant metadata, is present on many suspect lists compiled by the NORMAN Network (REACH2017, SusDat and KEMIMARKET), and has 47 underlying data sources in CompTox. Based on this aforementioned evidence, this identification has confidence level 2a.

Overall, both the spectral and metadata evidence strongly support Candidate 1 over the others, as seen in the large difference between the candidates' MetFrag Scores.

Candidate Recommendation

Candidate 1 should be strongly considered for further identification efforts.

A reference standard of DTXSID4058156 (metazachlor) provided a retention time match within 0.03 minutes, thereby confirming the identification of this unknown as metazachlor with Level 1 confidence.

Table 5
MetFrag Score breakdown by scoring term for the top 4 candidates for m/z 278.1062 (ultimately identified as metazachlor with Level 1 confidence).

MetFrag Scoring Terms	*Candidate 1* DTXSID4058156	*Candidate 2* DTXSID90916646	*Candidate 3* DTXSID40736053	*Candidate 4* DTXSID30150421
Spectral Terms (Raw Scores)				
FragmenterScore	95.30	18.00	61.52	47.52
OfflineMetFusion	4.64	3.65	3.25	2.99
OfflineIndivMoNA	1.00	0	0	0
Metadata Terms (Raw Scores)				
CPDAT_COUNT	0	0	0	0
DATA_SOURCES	47	2	1	7
KEMIMARKET_EXPO	16	0	0	0
KEMIMARKET_HAZ	9	0	0	0
NORMANSUSDAT	1	0	0	0
REACH2017	1	1	0	0
INDACT	0	0	0	0
MetFrag Score (Weighted)				
Total	7.00	*1.52*	*1.37*	*1.29*

Raw scores are given for interpretability; the maximum raw score over all candidates (used to normalize for the ranking) is indicated in bold. The final MetFrag Score is a sum of the normalised and weighted scoring terms as described in the Methods. Here, Candidate 1 has the highest overall MetFrag score, supported by both spectral and metadata scoring terms.

m/z 187.0938

Scenario 2: Low Spectral but High Metadata Scores (Moderate MetFrag Score; 3–6)

For *m/z 187.0938*, identified as a $[M + H]^+$ adduct by enviMass, the top candidate scored poorly in the Spectral terms compared to subsequent candidates. However, its strong scoring in the Metadata terms ultimately drove its high MetFrag score (Table 6).

The distribution of MetFrag Scores in Table 6 indicates that the top 3 (or even 4) candidates have relatively similar scores. Although the spectral data rather support Candidates 2 or 3 as better matching the experimental data, the high KEMIMARKET_EXPO score for Candidate 1 indicates that it may be of greater concern in a regulatory context due to the potentially large exposure volumes, and could be considered for further confirmation efforts to eliminate this from consideration in future campaigns.

Candidate Recommendation

All top four candidates should be considered for further identification efforts due to high exposure and hazard scores.

Table 6
MetFrag Score breakdown by scoring term for the top 4 candidates for m/z 187.0938.

MetFrag Scoring Terms	*Candidate 1* **DTXSID5020526**	*Candidate 2* **DTXSID70198185**	*Candidate 3* **DTXSID10185791**	*Candidate 4* **DTXSID70382365**
Spectral Terms (Raw Scores)				
FragmenterScore	7.88	65.03	50.21	40.46
OfflineMetFusion	0.88	1.04	1.01	0.86
OfflineIndivMoNA	0	0	0	0
Metadata Terms (Raw Scores)				
CPDAT_COUNT	0	0	0	0
DATA_SOURCES	42	7	5	7
KEMIMARKET_EXPO	11	2	2	6
KEMIMARKET_HAZ	2	3	3	3
NORMANSUSDAT	1	1	1	1
REACH2017	1	1	1	0
INDACT	0	0	0	0
MetFrag Score (Weighted)				
Total	*4.63*	*4.34*	*4.03*	*3.65*

Raw scores are given for interpretability; the maximum raw score over all candidates (used to normalize for the ranking) is indicated in bold. The final MetFrag Score is a sum of the normalised and weighted scoring terms as described in the Methods. Here, Candidate 1 has the highest overall MetFrag Score despite low Spectral term scores due to its high scoring Metadata. Full details on the candidates are available in MassIVE.

m/z 249.0728

Additional Example for Scenario 2: Low Spectral but High Metadata Scores (Moderate MetFrag Score; 3–6)

The information provided by high Metadata scores can serve as the discriminating factor between candidates when their Spectral scores yield little/poor information which in turn gives little indication of how to rank the candidates if only spectral evidence had been considered. In this sense, Metadata scoring terms contribute an extra layer of information beyond spectral evidence towards identifying potentially relevant unknowns.

For example, the top four candidates of *m/z 249.0728* (Table 7) have comparably poor Spectral scores meaning there is overall little spectral evidence supporting these identifications. However, Candidate 1 distinguishes itself significantly from the other candidates because of its relatively high Metadata scores, in particular its KEMIMARKET_EXPO, KEMIMARKET_HAZ, and presence in REACH2017. Therefore, it has higher environmental relevance than subsequent candidates, which explains its top ranking.

Candidate Recommendation

Candidate 1 should be considered for further identification efforts given the moderate KEMI exposure and hazard scores, indicating potential environmental relevance in Europe.

Table 7
MetFrag Score breakdown by scoring term for the top 4 candidates for m/z 249.0728.

MetFrag Scoring Terms	*Candidate 1* DTXSID50885566	*Candidate 2* DTXSID60154230	*Candidate 3* DTXSID70233803	*Candidate 4* DTXSID80278866
Spectral Terms (Raw Scores)				
FragmenterScore	0	0	0	0
OfflineMetFusion	0.67	0.64	0.63	0.70
OfflineIndivMoNA	0	0	0	0
Metadata Terms (Raw Scores)				
CPDAT_COUNT	0	0	0	0
DATA_SOURCES	6	3	3	2
KEMIMARKET_EXPO	2	0	0	0
KEMIMARKET_HAZ	3	0	0	0
NORMANSUSDAT	0	0	0	0
REACH2017	1	0	0	0
INDACT	0	0	0	0
MetFrag Score (Weighted)				
Total	*4.43*	*1.39*	*1.38*	*1.30*

Raw scores are given for interpretability; the maximum raw score over all candidates (used to normalize for the ranking) is indicated in bold. The final MetFrag Score is a sum of the normalised and weighted scoring terms as described in the Methods. Here, differences in candidates' Metadata scores allowed them to be differentiated from each other despite equally poor Spectral scores. Full details on the candidates are available in MassIVE.

m/z 142.0975

Additional Example for Scenario 2: Low Spectral but High Metadata Scores (Moderate MetFrag Score; 3–6)

Similar to the previous example, candidates for have *m/z* 142.0975 have comparable performance in the Spectral scores and would be practically indistinguishable from each other if not for the large difference in their Metadata scores (Table 8). Candidate 1 differs strongly from subsequent candidates because of its relatively high KEMIMARKET_EXPO, KEMIMARKET_HAZ and REACH2017 scores that support its top ranking.

Candidate Recommendation

Candidate 1 should be considered for further identification efforts given high Europe-relevant metadata scores.

Table 8
MetFrag Score breakdown by scoring term for the top 4 candidates for m/z 142.0975.

MetFrag Scoring Terms	*Candidate 1* DTXSID40200921	*Candidate 2* DTXSID50863460	*Candidate 3* DTXSID40233077	*Candidate 4* DTXSID90380247
Spectral Terms (Raw Scores)				
FragmenterScore	200.29	156.23	143.16	229.32
OfflineMetFusion	3.44	3.64	3.96	3.52
OfflineIndivMoNA	0	0.01	0	0
Metadata Terms (Raw Scores)				
CPDAT_COUNT	0	0	0	0
DATA_SOURCES	6	11	7	2
KEMIMARKET_EXPO	2	0	0	0
KEMIMARKET_HAZ	3	0	0	0
NORMANSUSDAT	1	1	0	0
REACH2017	1	0	0	0
INDACT	0	0	0	0
MetFrag Score (Weighted)				
Total	*5.29*	*3.11*	*2.26*	*2.07*

Raw scores are given for interpretability; the maximum raw score over all candidates (used to normalize for the ranking) is indicated in bold. The final MetFrag Score is a sum of the normalised and weighted scoring terms as described in the Methods. Here, differences in candidates' Metadata scores allowed them to be differentiated from each other despite equally good Spectral Scores. Full details on the candidates are available in MassIVE.

m/z 152.0198

Scenario 3: High Spectral Scores but low Metadata Scores (Moderate MetFrag Score; 3–6)

For the top candidates of *m/z* 152.0198, practically no metadata exists except for DATA_SOURCES - each candidate has 1, indicating that these are not particularly well-known chemicals (or, potentially newly discovered and not well documented in public databases yet). However, the FragmenterScores of the candidates differed sufficiently to discriminate between them and indicate that Candidate 1 may be the best match in this case (Table 9).

Candidate Recommendation

Candidate 1 may be considered for further identification efforts, but candidates for other masses are more promising in the regulatory context.

Table 9
MetFrag Score breakdown by scoring term for the top 4 candidates for m/z 152.0198.

MetFrag Scoring Terms	*Candidate 1* **DTXSID30534106**	*Candidate 2* **DTXSID30540904**	*Candidate 3* **DTXSID90610112**	*Candidate 4* **DTXSID40849677**
Spectral Terms (Raw Scores)				
FragmenterScore	217.84	158.82	144.54	142.75
OfflineMetFusion	2.06	2.08	2.17	2.02
OfflineIndivMoNA	0	0	0	0
Metadata Terms (Raw Scores)				
CPDAT_COUNT	0	0	0	0
DATA_SOURCES	1	1	1	1
KEMIMARKET_EXPO	0	0	0	0
KEMIMARKET_HAZ	0	0	0	0
NORMANSUSDAT	0	0	0	0
REACH2017	0	0	0	0
INDACT	0	0	0	0
MetFrag Score (Weighted)				
Total	*2.95*	*2.69*	*2.66*	*2.60*

Raw scores are given for interpretability; the maximum raw score over all candidates (used to normalize for the ranking) is indicated in bold. The final MetFrag Score is a sum of the normalised and weighted scoring terms as described in the Methods. Here, the Spectral scores provided the means for MetFrag to differentiate the candidates despite their equally poor Metadata scores. Full details on the candidates are available in MassIVE.

m/z 199.1050

Scenario 4: Low Spectral Scores, Low Metadata Scores (Low MetFrag Score; <3)

Candidates proposed for *m/z* 199.1050 had neither particularly strong spectral nor metadata information, resulting in low overall MetFrag scores. In this case, there is no strong evidence that any of the candidates available in CompTox are of particular interest in the context of the investigation.

Candidate Recommendation

Candidate 1 may be considered for further identification efforts, but candidates for other masses are more promising.

Table 10
MetFrag Score breakdown by scoring term for the top 4 candidates for m/z 199.1050.

MetFrag Scoring Terms	*Candidate 1* **DTXSID40514171**	*Candidate 2* **DTXSID00556299**	*Candidate 3* **DTXSID20776997**	*Candidate 4* **DTXSID50511555**
Spectral Terms (Raw Scores)				
FragmenterScore	19.48	2.43	8.12	6.00
OfflineMetFusion	2.808	2.809	2.800	2.810
OfflineIndivMoNA	0	0	0	0
Metadata Terms (Raw Scores)				
CPDAT_COUNT	0	0	0	0
DATA_SOURCES	1	2	1	1
KEMIMARKET_EXPO	0	0	0	0
KEMIMARKET_HAZ	0	0	0	0
NORMANSUSDAT	0	0	0	0
REACH2017	0	0	0	0
INDACT	0	0	0	0
MetFrag Score (Weighted)				
Total	*2.50*	*2.12*	*1.91*	*1.81*

Raw scores are given for interpretability; the maximum raw score over all candidates (used to normalize for the ranking) is indicated in bold. The final MetFrag Score is a sum of the normalised and weighted scoring terms as described in the Methods. Full details on the candidates are available in MassIVE.

Information for Regulatory Decision-making on Further Identification Efforts/Next Steps

Table 11
summarises the Candidate Recommendations presented above, where 7–9 candidates are recommended for further identification efforts for the 6 *m/z* presented here.

m/z	MetFrag Results Scenario	Candidates for Further Consideration	Justification for Candidate Recommendation
278.1062	Scenario 1	1	High MetFrag Score overall (high Spectral and Metadata Scores); subsequent candidates very poor in comparison.
187.0938	Scenario 2	4	Moderate MetFrag Score overall (low Spectral but high Metadata Scores); MetFrag Scores very similar across candidates, therefore all worth consideration.
249.0728	Scenario 2 (additional example)	1	Moderate MetFrag Score overall (low Spectral but high Metadata Scores); non-zero KEMIMARKET_EXPO and KEMIMARKET_HAZ, and presence in REACH2017 suspect list unlike subsequent candidates.
142.0975	Scenario 2 (additional example)	1	Moderate MetFrag Score overall (low Spectral but high Metadata Scores); non-zero KEMIMARKET_EXPO and KEMIMARKET_HAZ, and presence in REACH2017 suspect list unlike subsequent candidates.
152.0198	Scenario 3	0–1	Moderate MetFrag Score overall (high Spectral but low Metadata Scores); borderline low MetFrag Score, only worth (weakly) considering Candidate 1.
199.1050	Scenario 4	0–1	Low MetFrag Score overall (low Spectral andMetadata Scores); only worth (weakly) considering Candidate 1.

Table 11. Candidates for six m/z meriting further identification efforts based on individual evaluations.

Candidates were evaluated on an individual level for 6 m/z (selected out of 22 m/z as representative examples). Full details on further candidates are available in MassIVE.

The top four candidates for each of the remaining 16 *m/z* were analysed in the same way as discussed above, and candidates were evaluated based on the same criteria as described: prioritization according to tertile, scenario, and spectral and metadata scores, including potential exposure and hazards (SI Tables S3-S18). For these 16 *m/z*, a total of 25–49 candidates (out of possible 16 times 4 = 64) are recommended for further identification efforts (SI Table S19). Thus, for all the 22 *m/z* which underwent MetFrag identification in this study, an overall total of 32–58 candidates (out of possible 22 times 4 = 88) are recommended for further identification efforts. These candidate numbers are provided as ranges to allow for flexibility in project management and future steps, which may depend on available resources (see Discussion).

Discussion

In this study, non-target analysis was performed retrospectively on samples from Swiss WWTP effluents that had been collected as part of an existing regulatory environmental monitoring campaign. Instead of an exploratory approach that is still common amongst NTA studies, the research questions that directed this study were derived from regulatory priorities, thereby ensuring outcomes of direct and immediate relevance for environmental monitoring and protection.

Unknowns of regulatory interest were defined as those with the *highest intensities* and *highest temporal frequency* with *point sources* across all the samples of the sampling campaign. These features were prioritized in the data

using enviMass, resulting in lists of *m/z* that were the starting point of this current work. The mass spectra of these *m/z* underwent prescreening and quality control (Fig. 2) to ensure their suitability for use in non-target identification. Quality control isolated measurements worthy of further identification efforts and eliminated those of poor standard, effectively resulting in data reduction (Fig. 3). The prescreening workflow was written in R and is now openly available within the package Shinyscreen (Kondić et al.).

Then, MetFrag (Wolf et al. 2010; Ruttkies et al. 2016) was employed to provide tentative identifications for these unknowns, leveraging its extensive metadata capabilities “post-relaunch”, as well as several open resources/information sources, including chemical information from regulators around the world. MetFrag analysis was performed via the command line using scripts based on ReSOLUTION (Schymanski 2020a) and RchemMass (Schymanski 2020b).

Tentative identifications for 22 *m/z* were obtained using MetFrag (21 at Level 3, 1 at Level 2a, whose identity was eventually confirmed to Level 1). These identifications were evaluated in terms of (i) a score distribution for the top candidates (Fig. 4) and (ii) Scenario Analysis (Table 3) according to the regulatory context and research questions underlying this work. Final candidate recommendations were given based on MetFrag Score breakdowns, thereby providing in-depth and transparent analyses of the spectral and metadata evidence for proposed candidates. For the 22 *m/z* analysed, 32–58 candidates were recommended for further identification efforts.

Quality control was a critical element in the prescreening workflow, as preliminary manual inspection of the data using XCalibur revealed variable data quality. In fact, most data (> 80% cases) were not fully suitable for the intended non-target identification. R scripts (now embedded within Shinyscreen package) were written to automate most of the quality control checks (Table 1, checks 1–5). Automated quality control allowed for quick and reproducible processing of the large quantity of data needed to answer the superlative research questions guiding this work. The variable quality of the data had several likely causes: (i) List B masses were not in the inclusion list, (ii) MS2 were not measured immediately after MS1, therefore sample degradation over long storage time between MS1 and MS2 measurements could have occurred, and (iii) possibly over-restrictive enviMass prioritization criteria. Thus, the small number of cases (~ 0.03% of total) passing all quality control checks and qualifying for MetFrag identification was not unexpected.

MetFrag was configured to comprise both Spectral and Metadata scoring terms, including chemical suspect lists and scoring terms from international regulators within the latter such as KEMIMARKET_EXPO, KEMIMARKET_HAZ, REACH2017, NORMANSUSDAT, and CPDAT_COUNT. Paired with CompTox as its candidate database, MetFrag was thus specifically customised to perform non-target identification of environmental unknowns in WWTP samples within a regulatory context in this work. Beyond using fragmentation information alone, using metadata to inform MetFrag’s identifications proved to be especially important in certain situations *e.g.*, when spectral scores based on fragmentation were not informative enough to distinguish candidates from each other (Tables 7 & 8). Crucially, the information provided by metadata can serve as guidance for future regulatory actions in the context of the environmental protection aims of this study. For example, although certain candidate(s) may not be top-ranked or have strong spectral evidence (Table 6), potentially concerning hazard and exposure scores may qualify a certain candidate for serious consideration in future work in the spirit of applying the Precautionary Principle.

Regarding the components of the MetFrag Score, a total of ten scoring terms, three Spectral and seven Metadata, were used to score candidates. Compared to most previous studies which used MetFrag as mentioned in the Introduction, this number may seem large. However, adding extra scoring terms does not appear to compromise MetFrag’s identification capabilities. In fact, the additional scoring terms were beneficial because further bases for

differentiating between candidates became available. In other words, using more scoring terms can provide more granularity when distinguishing candidates, which is important for candidate evaluation and recommendation. Further scoring terms based on physical-chemical properties could be integrated in the future such as correlation of the partitioning coefficient $\log K_{ow}$ (or log P) with retention time (Ruttkies et al. 2016). Such scoring criteria would filter out any unrealistic candidates based on objective criteria like ionisability and polarity. (Insufficient information was available to perform retention time correlation via MetFrag in this study.)

With respect to the individual terms, CPDAT_COUNT, INDACT, and OfflineIndividualMoNA proved to be relatively uninformative in this particular study, evidenced by their frequent zero-value scores. As a chemical products database, CPDAT's limited applicability in wastewater studies such as the present one is unsurprising, and it instead may be more suitable for exposomics studies involving *e.g.*, household dust. INDACT, the list of industrial activity chemicals known to be used in the vicinity of the WWTPs as disclosed to the regulator, had the strongest potential to improve the identification results. However, not a single candidate across all the MetFrag results was present on this suspect list, which could suggest that the disclosures made by the industries were either incomplete or unsuitable for identification purposes (*e.g.*, parent compounds were disclosed but possibly only transformation products are present in the environment/detectable, UVCBs with unspecific chemical identities *etc.*). Lastly, while mass spectral libraries are inherently incomplete (Oberacher et al. 2020), a low OfflineIndividualMoNA score does not necessarily indicate poor spectral library matches. Rather, low OfflineIndividualMoNA scores could also signify that the candidate is not present within MoNA to begin with, or result from noisy experimental spectra even if the match would otherwise be good. Therefore, evaluating candidates on this scoring term alone must be done with these factors in mind, and improvements to its design to avoid possible faulty interpretations could constitute future work. Other future work on MetFrag itself could involve the addition of new Spectral scoring terms which do not require scaling via normalisation of the maximum value, as this maximum value is highly dependent on the candidate database chosen. For instance, a simple spectral similarity metric such as cosine similarity would evaluate how well the *in silico* and experimental fragmentation spectra align, independent of those of other candidates.

CompTox, the candidate database chosen here, remains one of the most environmentally-focused open databases of chemical compounds as it exclusively contains chemicals of environmental and toxicological relevance. Compared to other open databases like PubChem, CompTox is also smaller in size. Therefore, MetFrag paired with CompTox is likely to suggest smaller lists of candidates which are *de facto* environmentally-meaningful, making workflow runtimes shorter and candidate evaluation relatively easier. However, using CompTox has drawbacks, principally stemming from its lack of comprehensiveness when compared to PubChem, which is larger and covers a wider chemical space beyond just environmentally and toxicologically relevant chemicals. Therefore, false negatives can result should certain compounds matching the identification criteria not exist within CompTox to begin with. The forthcoming PubChemLite (Bolton et al. 2020; Schymanski et al. 2020) represents one complementary alternative to these issues, as it is by design essentially a subset of environmentally-relevant compounds based on compound classifications. Overall, the ability to subset databases based on usage and classification information of chemicals can be beneficial, as different regulatory bodies may have different mandates, and studies can be designed to align with those mandates accordingly *e.g.*, focus only on chemicals with (i) known usage in industrial manufacturing, or (ii) agricultural chemicals, or (iii) pharmaceuticals *etc*.

Using scenarios as a framework to interpret MetFrag's results was critical considering the specific regulatory aims of this work: tentatively identify pollutants of high priority (with minimum Level 3 confidence) to guide further monitoring and identification efforts.

Scenario Analysis revealed in detail whether Spectral, Metadata, or both contributed to a given MetFrag Score and in turn provided the rationale behind proposed candidates. As our evaluation has shown, multiple candidates are worth considering especially if they have very similar scores (*e.g.*, Table 6), or have more compelling evidence represented by individual scoring terms (*e.g.*, Table 13) as described above. In this way, Scenario Analysis as used here is highly suitable for transparently communicating scientific results in a regulatory context. On a larger scale, such analyses address a key weakness common to NTA studies: the current lack of ability to perform detailed data interpretation – especially in a high throughput, automatable and reproducible manner.

Furthermore, Scenario Analysis as used here can inform decision-making regarding the next steps. Besides addressing study priorities based on "depth vs. breadth" as discussed in the Results, the scenarios can be used to devise a prioritisation scheme for future work. For example, if authentic standards can only be purchased/analysed for 10 compounds due to resource limitations, those compounds should be the recommended candidates with MetFrag Scores from Scenario 1 > Scenarios 2/3 >>> Scenario 4. Alternatively, if it is known from the outset that spectral data may be poor quality, Scenario 2 candidates may take precedence over Scenario 3 candidates, as the former rely on high Metadata Scores and not high Spectral scores for their high MetFrag Scores. Additionally, applying the Precautionary Principle may motivate prioritizing identity confirmations of candidates with concerning metadata like high toxicity and/or exposure (corresponding to KEMIMARKET_HAZ and KEMIMARKET_EXPO scores), even if those candidates are not necessarily ranked highly by MetFrag.

Practically speaking, next steps in environmental monitoring based on the results here (besides identity confirmation using authentic standards) could include expanding suspect lists using the recommended candidates to improve future suspect screening activities. These new suspects could in turn be added to the inclusion lists of future measurements, thereby already gaining an analytical 'upper-hand' for future NTA studies. Expanding suspect and inclusion lists in this way, possibly in combination with using a rarity score (Krauss et al. 2019), represents an evidence-based approach towards more meaningful environmental monitoring in the long-run, as these candidate compounds were tentatively 'observed' and are therefore *site-specific*. Otherwise, suspect lists are typically expanded based on information from national or international chemical registration lists, whose applicability may be limited depending on the actual usage/exposure in the region of concern. Therefore, an additional outcome of this study is a means to *bridge Target and Non-target Analysis by supplying meaningful candidates for Suspect Screening*.

This work is one contribution to a much larger discussion surrounding (i) how NTA can support regulatory environmental monitoring and (ii) the practical feasibility of applying NTA in routine environmental monitoring. Regarding the former, this work demonstrates that NTA can be used to address the concerns of regulators by translating research questions arising from regulatory priorities into peak-picking/mass prioritisation criteria: in this case, high concentration unknown pollutants with point sources that occurred persistently were taken to be high intensity precursors found at one or few sampling sites at both sampling time points. Without the ability to perform quantification, the assumption that high ion intensity represents high concentration could be validated by using different chromatographic solvent systems as a test of ionisation efficiency in future work, or implementing ionization efficiency models (Liigand et al. 2020; Panagopoulos Abrahamsson et al. 2020).

On the feasibility of performing NTA as part of routine regulatory environmental monitoring, the overall method described here offers a highly *automated* approach via (i) feature prioritization via enviMass, (ii) prescreening and quality control (plus a manual step), and (iii) *in silico* identification, of which (ii) and (iii) were developed in this work. The results interpretation and candidate recommendation processes performed manually in this work form the basis of future efforts towards automated reporting based on Scenario Analysis, MetFrag Score distributions, and evaluation of critical parameters like thresholds for potential toxicities and exposure levels. Such automated

reporting would not only allow scalability of future regulatory NTA studies, but could also eliminate potential biases in unknown identification – analysts would not be able to 'cherry-pick' candidates based on their familiarity with certain compounds because undescriptive identifiers *e.g.*, DTXSIDs would be used up until the final results are delivered at the end of the entire method. Furthermore, while the prescreening, quality control, and identification workflow was applied retrospectively, the improvements to workflow automation detailed here could allow for quicker data analysis turnaround in the future, which would help guide future sampling and measurements planned in the short-medium term and prevent the long delays between remeasurements still commonly observed in NTA investigations – effectively, moving towards 'real-time' instead of retrospective NTA approaches. Two concrete follow-up initiatives are foreseen: (i) build an interface connecting Shinyscreen and MetFrag, including automated reporting features as previously described, and (ii) develop a set of 'default' scoring terms and settings tailored for NTA of wastewater samples. Further collaborations involving non-target wastewater studies and database hosts will help augment expert knowledge on more use cases, which would be leveraged to develop this approach further.

Conclusions

A prescreening and identification workflow for analysing non-target compounds was developed in this study to retrospectively identify unknowns detected in WWTP sites in the context of directly supporting regulatory decision-making for environmental monitoring. Using open data and open tools including the US EPA CompTox Chemicals Dashboard, NORMAN Network resources such as SusDat and the Suspect List Exchange, and MetFrag, tentative identifications for 21 unknown compounds were provided at Level 3 confidence, and 1 compound's identity was confirmed using a reference standard giving a Level 1 identification. These results were achieved despite limited data quality.

This study heavily emphasized results interpretation on two levels: on a global level across the chemical unknowns investigated, and on an individual candidate level. Through these analyses, specific candidates were recommended for further identification efforts, and transparent justifications were provided based on the MetFrag score breakdown (*i.e.*, spectral vs. metadata evidence). These recommendations, and not just MetFrag's outputs, represent the final results in the regulatory and environmental monitoring context of this study, and may serve as a template to drive future developments in NTA.

The prescreening and quality control workflow developed here is embedded in the open R package Shinyscreen (Kondić et al.), which is freely available online, as is code from ReSOLUTION (Schymanski 2020a) and RChemMass (Schymanski 2020b) used for performing command-line MetFrag identification. The CompTox database version with the metadata terms used here is likewise also publicly available (Schymanski 2019).

Abbreviations

NTA – non-target analysis; WWTP – wastewater treatment plant; US EPA – United States Environmental Protection Agency; CompTox – US EPA CompTox Chemicals Dashboard; DTXSID – DSSTox Substance Identifier (from CompTox); CPDat – Chemicals and Products Database; REACH – Registration, Evaluation, Authorisation and Restriction of Chemicals; MoNA – Mass Bank of North America; UVCB – Chemical substances of Unknown or Variable composition, Complex reaction products, and Biological materials.

Declarations

Ethics Approval and Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Availability of Data and Materials

The mass spectrometry dataset generated and analysed during the current study, including the complete MetFrag results for the 22 m/z that were tentatively identified, are available as an open MassIVE dataset (MSV000086631) via https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=14f51e6ec99a42329e7a0eeaad0e5824. (Dataset is password-protected and available to reviewers while the manuscript remains under review, username MSV000086631_reviewer. The password is provided in the Cover Letter as part of manuscript submission. Upon acceptance for publication, the dataset will be released publicly.)

All code used to run the prescreening and quality control workflow and MetFrag command-line analysis is open/publicly available via https://github.com/schymane/ReSOLUTION, https://github.com/schymane/RChemMass, and Shinyscreen (see below). All other datasets and databases used as part of MetFrag identification are open/publicly available (links in-line throughout this manuscript).

Software

Project name: Shinyscreen

Project home page: https://git-r3lab.uni.lu/eci/shinyscreen

Archived version used in this study: Shinyscreen commit version 3c08df96 (https://git-r3lab.uni.lu/eci/shinyscreen/-/commits/3c08df9631e7a81b5296b917ebd388d3667fa436)

Operating system(s): Windows, Mac OSX, Linux

Programming language: R

Other requirements: OpenJDK and other R package dependencies listed in Shinyscreen's README

License: Apache Version 2.0 (https://www.apache.org/licenses/LICENSE-2.0)

Competing Interests

The authors declare no competing interests.

Funding

ELS, AL, RRS, and TK are supported by the Luxembourg National Research Fund (FNR) for project A18/BM/12341006.

Authors' Contributions

LK conceived the study and set up the sampling campaigns; OJ measured the data; AL, ELS, RRS designed the workflow presented; AL, ELS, TK wrote the software; AL interpreted the data, AL drafted the manuscript with inputs from ELS, RRS, LK, and OJ; AL, LK, OJ, RRS, TK, and ELS revised and approved the submitted version.

Acknowledgements

The authors acknowledge Dr. Martin Loos (enviBee GmbH) for his technical support with enviMass analyses.

References

1. Albergamo V, Schollée JE, Schymanski EL, et al (2019) Nontarget Screening Reveals Time Trends of Polar Micropollutants in a Riverbank Filtration System. Environ Sci Technol. https://doi.org/10.1021/acs.est.9b01750
2. Alygizakis N, Slobodnik J (2018) S32 | REACH2017 | >68,600 REACH Chemicals. https://zenodo.org/record/3653160. Accessed 16 Aug 2020
3. Anliker S, Loos M, Comte R, et al (2020) Assessing Emissions from Pharmaceutical Manufacturing Based on Temporal High-Resolution Mass Spectrometry Data. Environ Sci Technol 54:4110–4120. https://doi.org/10.1021/acs.est.9b07085
4. Beckers L-M, Brack W, Dann JP, et al (2020) Unraveling longitudinal pollution patterns of organic micropollutants in a river by non-target screening and cluster analysis. Science of The Total Environment 727:138388. https://doi.org/10.1016/j.scitotenv.2020.138388
5. Bolton E, Schymanski E, Kondic T, et al (2020) PubChemLite for Exposomics. https://zenodo.org/record/4183801#.X89xNl4o90s. Accessed 8 Dec 2020
6. Carpenter CMG, Wong LYJ, Johnson CA, Helbling DE (2019) Fall Creek Monitoring Station: Highly Resolved Temporal Sampling to Prioritize the Identification of Nontarget Micropollutants in a Small Stream. Environ Sci Technol 53:77–87. https://doi.org/10.1021/acs.est.8b05320
7. Chambers MC, Maclean B, Burke R, et al (2012) A cross-platform toolkit for mass spectrometry and proteomics. Nature Biotechnology 30:918–920. https://doi.org/10.1038/nbt.2377
8. Chiaia-Hernández AC, Günthardt BF, Frey MP, Hollender J (2017) Unravelling Contaminants in the Anthropocene Using Statistical Analysis of Liquid Chromatography–High-Resolution Mass Spectrometry Nontarget Screening Data Recorded in Lake Sediments. Environ Sci Technol 51:12547–12556. https://doi.org/10.1021/acs.est.7b03357
9. Choi Y, Kim K, Kim D, et al (2020) Ny-Ålesund-oriented organic pollutants in sewage effluent and receiving seawater in the Arctic region of Kongsfjorden. Environmental Pollution 258:113792. https://doi.org/10.1016/j.envpol.2019.113792
10. Faber A-H, Annevelink MPJA, Schot PP, et al (2019) Chemical and bioassay assessment of waters related to hydraulic fracturing at a tight gas production site. Science of The Total Environment 690:636–646. https://doi.org/10.1016/j.scitotenv.2019.06.354
11. Fiehn Lab MassBank of North America. https://mona.fiehnlab.ucdavis.edu/. Accessed 3 Jun 2020
12. Fischer S (2017) S17 | KEMIMARKET | KEMI Market List. https://zenodo.org/record/3653175#.XrU6u26xWi4. Accessed 8 May 2020

13. Gerlich M, Neumann S (2013) MetFusion: integration of compound identification strategies. Journal of Mass Spectrometry 48:291–298. https://doi.org/10.1002/jms.3123
14. Helmus R, ter Laak TL, de Voogt P, et al (2020) Patroon: Open Source Software Platform for Environmental Mass Spectrometry Based Non-target Screening. https://doi.org/10.21203/rs.3.rs-36675/v1
15. Hites RA, Jobst KJ (2018) Is Nontargeted Screening Reproducible? Environ Sci Technol 52:11975–11976. https://doi.org/10.1021/acs.est.8b05671
16. Hollender J, Schymanski EL, Singer HP, Ferguson PL (2017) Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? Environ Sci Technol 51:11505–11512. https://doi.org/10.1021/acs.est.7b02184
17. Hollender J, van Bavel B, Dulio V, et al (2019) High resolution mass spectrometry-based non-target screening can support regulatory environmental monitoring and chemicals management. Environmental Sciences Europe 31:42. https://doi.org/10.1186/s12302-019-0225-x
18. Hug C, Ulrich N, Schulze T, et al (2014) Identification of novel micropollutants in wastewater by a combination of suspect and nontarget screening. Environmental Pollution 184:25–32. https://doi.org/10.1016/j.envpol.2013.07.048
19. Kandie FJ, Krauss M, Beckers L-M, et al (2020) Occurrence and risk assessment of organic micropollutants in freshwater systems within the Lake Victoria South Basin, Kenya. Science of The Total Environment 714:136748. https://doi.org/10.1016/j.scitotenv.2020.136748
20. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28:27–30. https://doi.org/10.1093/nar/28.1.27
21. Kim S, Chen J, Cheng T, et al (2019) PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 47:D1102–D1109. https://doi.org/10.1093/nar/gky1033
22. Kondić T, Lai A, Schymanski E, et al Environmental Cheminformatics / shinyscreen. https://git-r3lab.uni.lu/eci/shinyscreen. Accessed 16 Aug 2020
23. Köppe T, Jewell KS, Dietrich C, et al (2020) Application of a non-target workflow for the identification of specific contaminants using the example of the Nidda river basin. Water Research 178:115703. https://doi.org/10.1016/j.watres.2020.115703
24. Krauss M, Hug C, Bloch R, et al (2019) Prioritising site-specific micropollutants in surface water from LC-HRMS non-target screening data using a rarity score. Environmental Sciences Europe 31:45. https://doi.org/10.1186/s12302-019-0231-z
25. Krauss M, Singer H, Hollender J (2010) LC–high resolution MS in environmental analysis: from target screening to the identification of unknowns. Anal Bioanal Chem 397:943–951. https://doi.org/10.1007/s00216-010-3608-9
26. Lara-Martín PA, Chiaia-Hernández AC, Biel-Maeso M, et al (2020) Tracing Urban Wastewater Contaminants into the Atlantic Ocean by Nontarget Screening. Environ Sci Technol 54:3996–4005. https://doi.org/10.1021/acs.est.9b06114
27. Lege S, Eisenhofer A, Heras JEY, Zwiener C (2019) Identification of transformation products of denatonium – Occurrence in wastewater treatment plants and surface waters. Science of The Total Environment 686:140–150. https://doi.org/10.1016/j.scitotenv.2019.05.423
28. Letzel T FOR-IDENT – Fortschritte in der Identifizierung organischer Spurenstoffe: Zusammenführen der Hilfsmittel und Standardisierung der Suspected- und Non-Target Analytik. https://www.for-ident.org/. Accessed 8 Dec 2020

29. Li Z, Kaserzon SL, Plassmann MM, et al (2017) A strategic screening approach to identify transformation products of organic micropollutants formed in natural waters. Environ Sci: Processes Impacts 19:488–498. https://doi.org/10.1039/C6EM00635C
30. Liigand J, Wang T, Kellogg J, et al (2020) Quantification for non-targeted LC/MS screening without standard substances. Scientific Reports 10:5808. https://doi.org/10.1038/s41598-020-62573-z
31. Ljoncheva M, Stepišnik T, Džeroski S, Kosjek T (2020) Cheminformatics in MS-based environmental exposomics: Current achievements and future directions. Trends in Environmental Analytical Chemistry 28:e00099. https://doi.org/10.1016/j.teac.2020.e00099
32. Loos M, Schmitt U, Schollée JE (2018) blosloos/enviMass: enviMass version 3.5. https://zenodo.org/record/1213098#.X4VjSlmxWi4. Accessed 13 Oct 2020
33. Luft A, Bröder K, Kunkel U, et al (2017) Nontarget Analysis via LC-QTOF-MS to Assess the Release of Organic Substances from Polyurethane Coating. Environ Sci Technol 51:9979–9988. https://doi.org/10.1021/acs.est.7b01573
34. MassBank Consortium, NORMAN Association MassBank | MassBank Europe Mass Spectral DataBase. https://massbank.eu/MassBank/. Accessed 23 Dec 2020
35. McEachran AD, Mansouri K, Grulke C, et al (2018) “MS-Ready” structures for non-targeted high-resolution mass spectrometry screening studies. Journal of Cheminformatics 10:. https://doi.org/10.1186/s13321-018-0299-2
36. Menger F, Ahrens L, Wiberg K, Gago-Ferrero P (2021) Suspect screening based on market data of polar halogenated micropollutants in river water affected by wastewater. Journal of Hazardous Materials 401:123377. https://doi.org/10.1016/j.jhazmat.2020.123377
37. Miaz LT, Plassmann MM, Gyllenhammar I, et al (2020) Temporal trends of suspect- and target-per/polyfluoroalkyl substances (PFAS), extractable organic fluorine (EOF) and total fluorine (TF) in pooled serum from first-time mothers in Uppsala, Sweden, 1996–2017. Environ Sci: Processes Impacts 22:1071–1083. https://doi.org/10.1039/C9EM00502A
38. Moschet C, Anumol T, Lew BM, et al (2018) Household Dust as a Repository of Chemical Accumulation: New Insights from a Comprehensive High-Resolution Mass Spectrometric Study. Environ Sci Technol 52:2878–2887. https://doi.org/10.1021/acs.est.7b05767
39. Muz M, Dann JP, Jäger F, et al (2017) Identification of Mutagenic Aromatic Amines in River Samples with Industrial Wastewater Impact. Environ Sci Technol 51:4681–4688. https://doi.org/10.1021/acs.est.7b00426
40. NORMAN Network NORMAN Suspect List Exchange. https://www.norman-network.com/nds/SLE/. Accessed 24 Aug 2020
41. NORMAN Network, Aalizadeh R, Alygizakis N, et al (2020) S0 | SUSDAT | Merged NORMAN Suspect List: SusDat. https://zenodo.org/record/3695732#.XrU6K26xWi4. Accessed 8 May 2020
42. Oberacher H, Sasse M, Antignac J-P, et al (2020) A European proposal for quality control and quality assurance of tandem mass spectral libraries. Environ Sci Eur 32:43. https://doi.org/10.1186/s12302-020-00314-9
43. Oetjen K, Blotevogel J, Borch T, et al (2018) Simulation of a hydraulic fracturing wastewater surface spill on agricultural soil. Science of The Total Environment 645:229–234. https://doi.org/10.1016/j.scitotenv.2018.07.043
44. Panagopoulos Abrahamsson D, Park J-S, Singh RR, et al (2020) Applications of Machine Learning to In Silico Quantification of Chemicals without Analytical Standards. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.9b01096

45. Park N, Choi Y, Kim D, et al (2018) Prioritization of highly exposable pharmaceuticals via a suspect/non-target screening approach: A case study for Yeongsan River, Korea. Science of The Total Environment 639:570–579. https://doi.org/10.1016/j.scitotenv.2018.05.081
46. Pence HE, Williams A (2010) ChemSpider: An Online Chemical Information Resource. J Chem Educ 87:1123–1124. https://doi.org/10.1021/ed100697w
47. Purschke K, Zoell C, Leonhardt J, et al (2020) Identification of unknowns in industrial wastewater using offline 2D chromatography and non-target screening. Science of The Total Environment 706:135835. https://doi.org/10.1016/j.scitotenv.2019.135835
48. Ruff M, Mueller MS, Loos M, Singer HP (2015) Quantitative target and systematic non-target analysis of polar organic micro-pollutants along the river Rhine using high-resolution mass-spectrometry – Identification of unknown sources and compounds. Water Research 87:145–154. https://doi.org/10.1016/j.watres.2015.09.017
49. Ruttkies C, Schymanski EL, Wolf S, et al (2016) MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminform 8:3. https://doi.org/10.1186/s13321-016-0115-9
50. Schwarzbauer J, Ricking M (2010) Non-target screening analysis of river water as compound-related base for monitoring measures. Environ Sci Pollut Res 17:934–947. https://doi.org/10.1007/s11356-009-0269-3
51. Schymanski E (2019) MetFrag Local CSV: CompTox (7 March 2019 release) Wastewater MetaData File. https://zenodo.org/record/3472781#.XrU5kW6xWi5. Accessed 8 May 2020
52. Schymanski E (2020a) schymane/ReSOLUTION. https://github.com/schymane/ReSOLUTION. Accessed 16 Aug 2020
53. Schymanski E (2020b) schymane/RChemMass. https://github.com/schymane/RChemMass. Accessed 16 Aug 2020
54. Schymanski EL, Jeon J, Gulde R, et al (2014) Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. Environ Sci Technol 48:2097–2098. https://doi.org/10.1021/es5002105
55. Schymanski EL, Kondic T, Neumann S, et al (2020) Empowering Large Chemical Knowledge Bases for Exposomics: Pubchemlite Meets Metfrag | Research Square. https://doi.org/10.21203/rs.3.rs-107432/v1. Accessed 3 Dec 2020
56. Sousa JCG, Ribeiro AR, Barbosa MO, et al (2018) A review on environmental monitoring of water organic pollutants identified by EU guidelines. Journal of Hazardous Materials 344:146–162. https://doi.org/10.1016/j.jhazmat.2017.09.058
57. Sun C, Zhang Y, Alessi DS, Martin JW (2019) Nontarget profiling of organic compounds in a temporal series of hydraulic fracturing flowback and produced waters. Environment International 131:104944. https://doi.org/10.1016/j.envint.2019.104944
58. Tian Z, Peter KT, Gipe AD, et al (2020) Suspect and Nontarget Screening for Contaminants of Emerging Concern in an Urban Estuary. Environ Sci Technol 54:889–901. https://doi.org/10.1021/acs.est.9b06126
59. US EPA (2016) Chemical and Products Database (CPDat). In: US EPA. https://www.epa.gov/chemical-research/chemical-and-products-database-cpdat. Accessed 8 May 2020
60. Veenaas C, Bignert A, Liljelind P, Haglund P (2018) Nontarget Screening and Time-Trend Analysis of Sewage Sludge Contaminants via Two-Dimensional Gas Chromatography–High Resolution Mass Spectrometry. Environ Sci Technol 52:7813–7822. https://doi.org/10.1021/acs.est.8b01126
61. Wagner TV, Helmus R, Quiton Tapia S, et al (2020) Non-target screening reveals the mechanisms responsible for the antagonistic inhibiting effect of the biocides DBNPA and glutaraldehyde on benzoic acid biodegradation.

Journal of Hazardous Materials 386:121661. https://doi.org/10.1016/j.jhazmat.2019.121661

62. Wang Z, Walker GW, Muir DCG, Nagatani-Yoshida K (2020) Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. Environ Sci Technol. https://doi.org/10.1021/acs.est.9b06379
63. Williams AJ, Grulke CM, Edwards J, et al (2017) The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. Journal of Cheminformatics 9:61. https://doi.org/10.1186/s13321-017-0247-6
64. Wishart DS, Feunang YD, Marcu A, et al (2018) HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res 46:D608–D617. https://doi.org/10.1093/nar/gkx1089
65. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics 11:148. https://doi.org/10.1186/1471-2105-11-148
66. ChemSpider | Search and share chemistry. http://www.chemspider.com/. Accessed 13 Aug 2020
67. Human Metabolome Database. https://hmdb.ca/. Accessed 13 Aug 2020
68. MassBank of North America. https://mona.fiehnlab.ucdavis.edu/. Accessed 8 May 2020

Figures

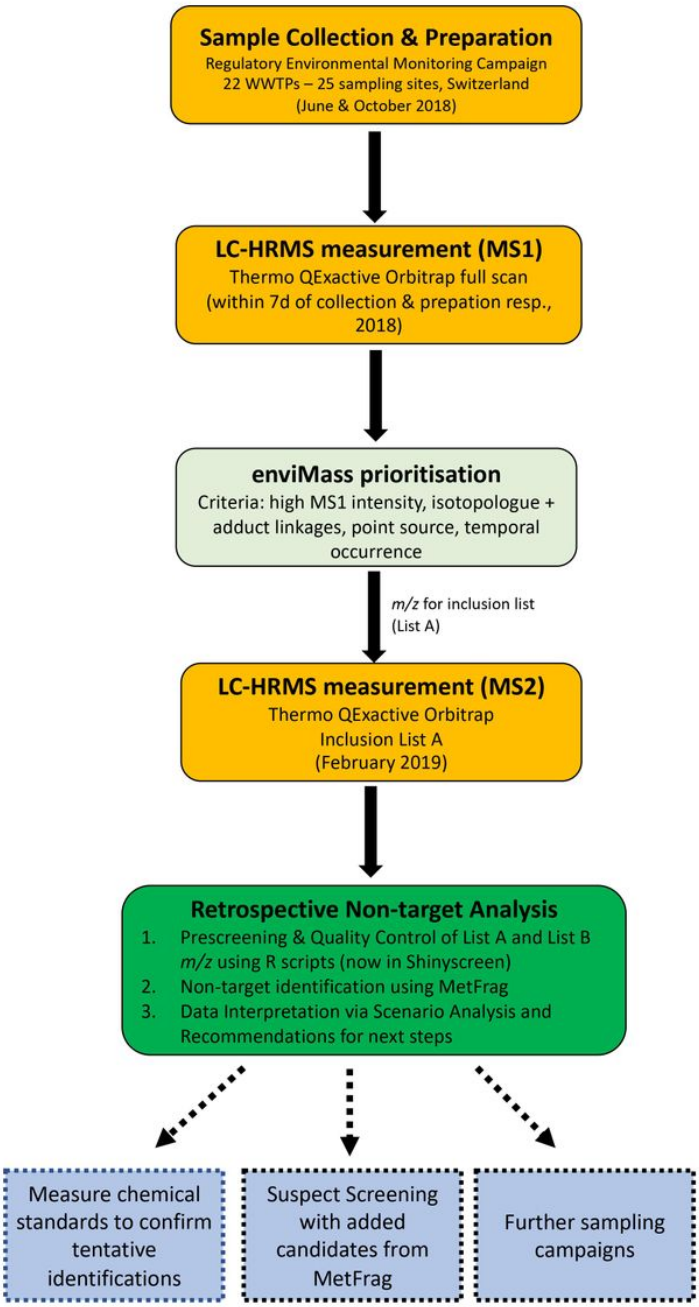

Figure 1

Visual project overview showing analytical and computational steps. Analytical “wet lab” steps are indicated in yellow, while “in silico” computational steps are indicated in green. The current study focuses on Retrospective Non-target Analysis, shown in dark green. Dotted arrows and boxes indicate possible future work based on the results of the current study, highlighted in blue to represent decisions to be made based on regulatory priorities.

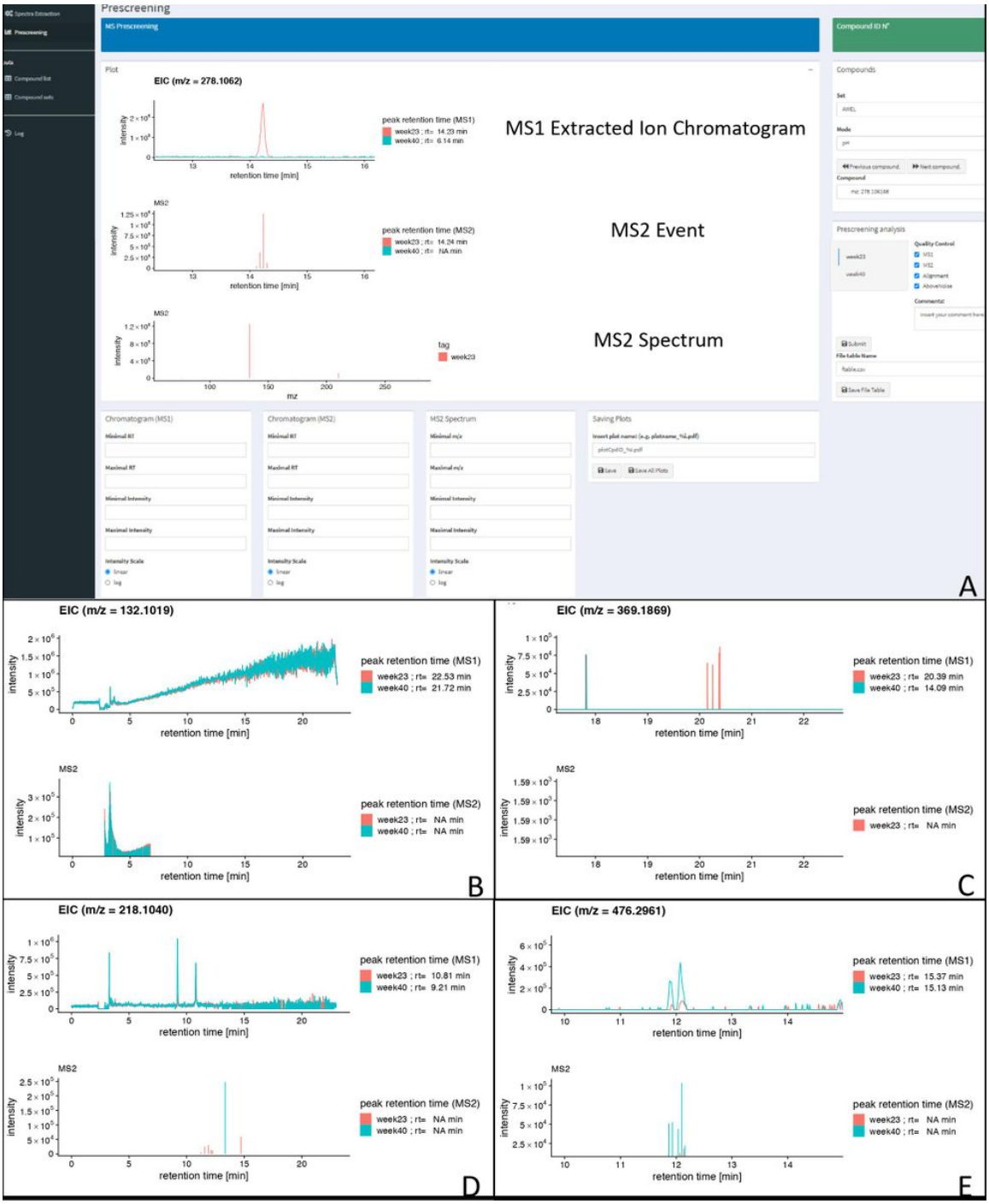

Figure 2

Examples of cases which pass and fail quality control within the prescreening workflow. Quality control helped isolate measurements which were suitable for non-target identification and discarded those which are not. Panel A shows Shinyscreen's graphical user interface and an example of a case whose MS1-MS2 measurement is suitable for non-target identification – its extracted ion chromatogram shows a MS1 peak of sufficiently high intensity, a corresponding MS2 event that is temporally well-aligned, and its MS2 spectrum. The remaining

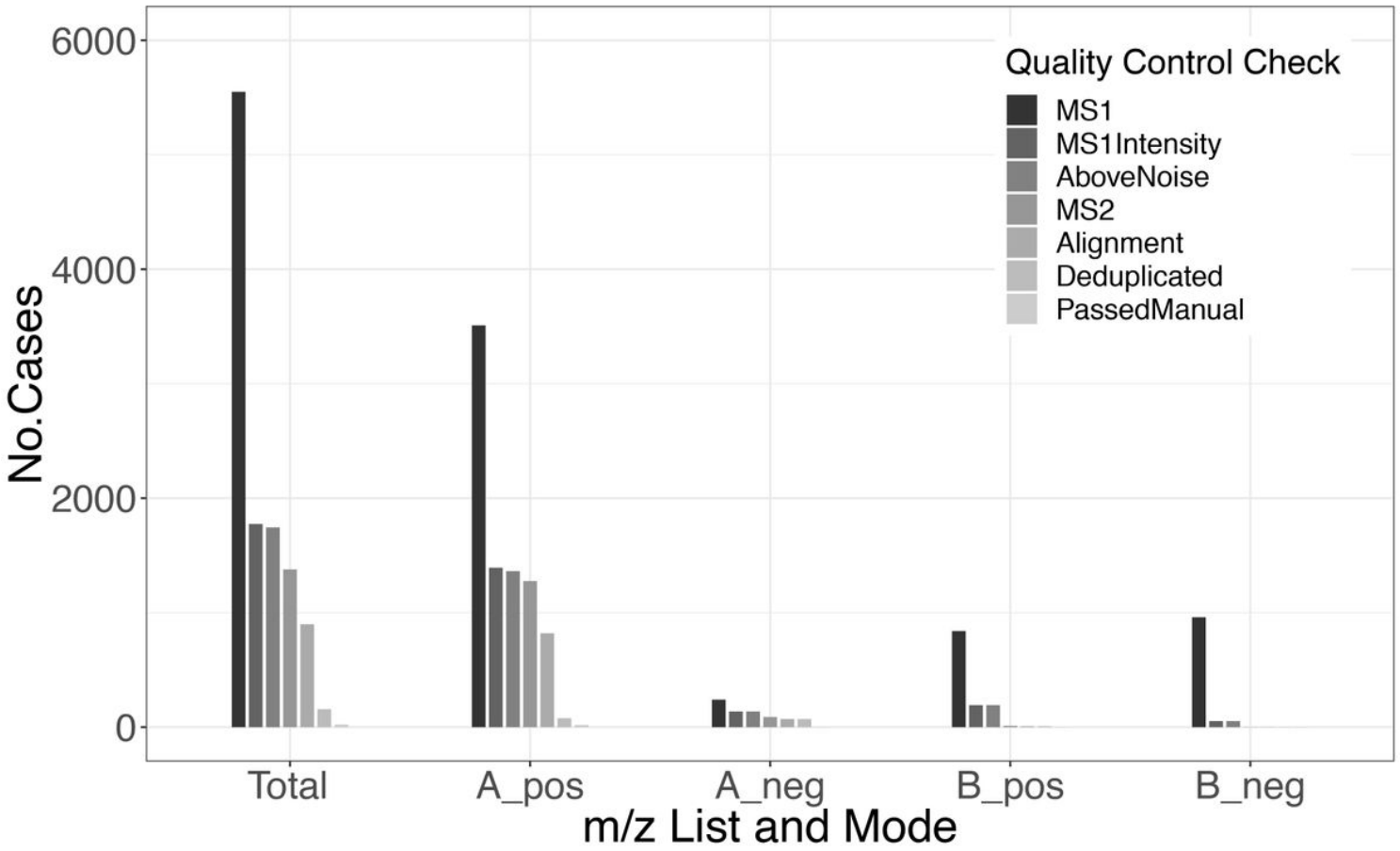

Figure 3

Quality control checks within prescreening resulted in data reduction prior to identification using MetFrag. Each check is represented by a bar whose height indicates the number of cases which passed that check. Going from left to right within each group of bars reflects the sequence of quality control checks (checks 1-7, Table 1).

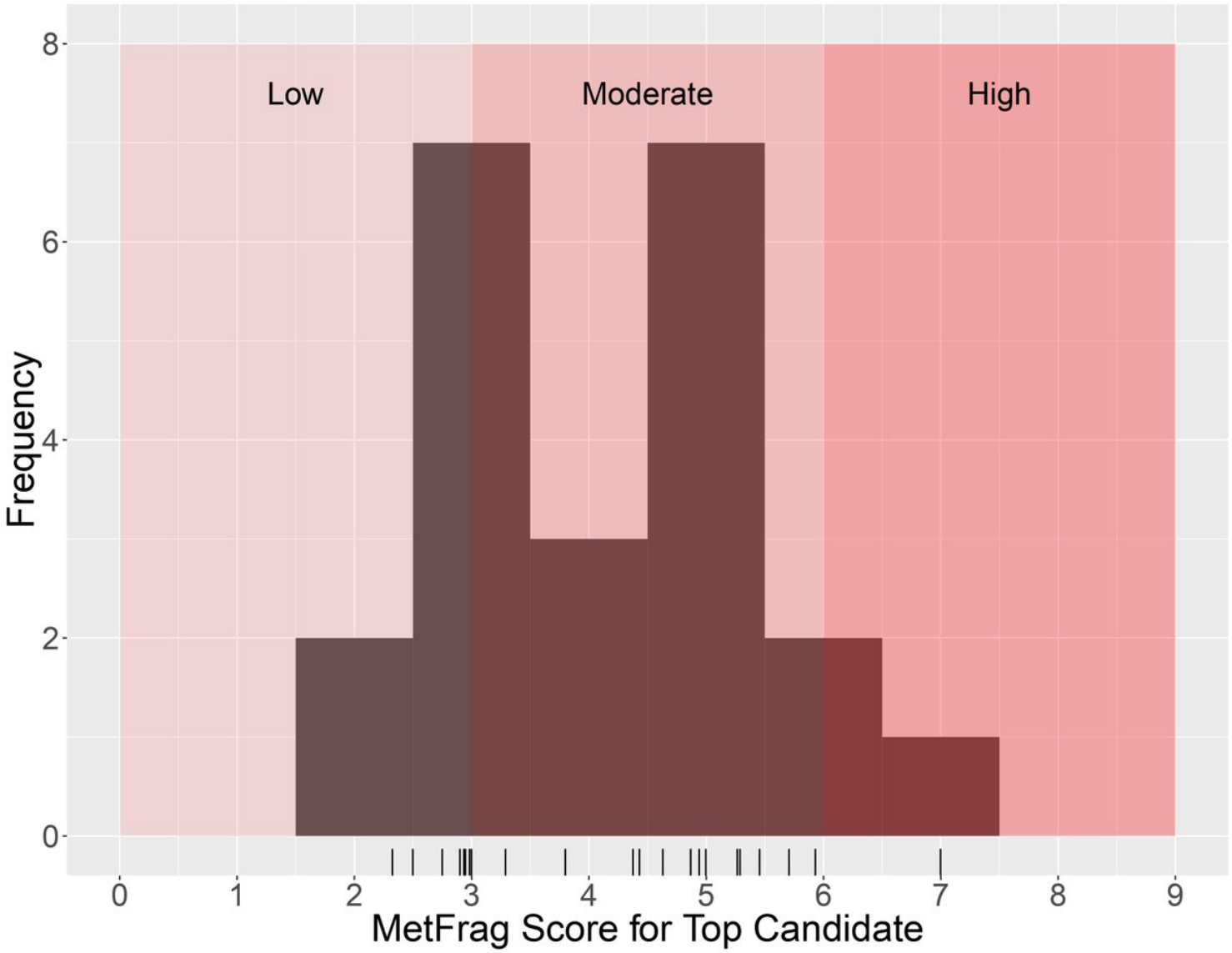

Figure 4

Distribution of MetFrag Scores for the top candidate of each m/z (n=22). Shaded regions indicate distribution tertiles corresponding to Low, Moderate, and High MetFrag Scores respectively. A rug plot is included along the x-axis to give an indication of the actual MetFrag Score values within each histogram bin.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SuppInfoLaietal2020.docx