

Discovering Pesticides and their Transformation Products in Luxembourg Waters using Open Cheminformatics Approaches

Jessy Krier

University of Luxembourg <https://orcid.org/0000-0001-6986-5545>

Randolph R. Singh

University of Luxembourg / IFREMER <https://orcid.org/0000-0003-4500-3400>

Todor Kondic

University of Luxembourg <https://orcid.org/0000-0001-6662-4375>

Adelene Lai

University of Luxembourg / FSU Jena <https://orcid.org/0000-0002-2985-6473>

Philippe Diderich

AGE <https://orcid.org/0000-0001-6969-2162>

Jian Zhang

NCBI/NLM/NIH <https://orcid.org/0000-0002-6192-4632>

Paul A. Thiessen

NCBI/NLM/NIH <https://orcid.org/0000-0002-1992-2086>

Evan E. Bolton

NCBI/NLM/NIH <https://orcid.org/0000-0002-5959-6190>

Emma L. Schymanski (✉ emma.schymanski@uni.lu)

University of Luxembourg <https://orcid.org/0000-0001-6868-8145>

Research Article

Keywords: Pesticides, Transformation Products, Suspect Screening, High Resolution Tandem Mass Spectrometry, Data mining, Non-target Screening

DOI: <https://doi.org/10.21203/rs.3.rs-478324/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

The diversity of hundreds of thousands of potential organic pollutants and the lack of (publicly available) information about many of them is a huge challenge for environmental sciences, engineering, and regulation. Suspect screening based on high-resolution liquid chromatography-mass spectrometry (LC-HRMS) has enormous potential to help characterize the presence of these chemicals in our environment, enabling the detection of known and newly emerging pollutants, as well as their potential transformation products (TPs). Here, suspect list creation (focusing on pesticides relevant for Luxembourg, incorporating data sources in 4 languages) was coupled to an automated retrieval of related TPs from PubChem based on high confidence suspect hits, to screen for pesticides and their TPs in Luxembourgish river samples. A computational workflow was established to combine LC-HRMS analysis and pre-screening of the suspects (including automated quality control steps), with spectral annotation to determine which pesticides and, in a second step, their related TPs may be present in the samples. The data analysis with Shinyscreen (<https://git-r3lab.uni.lu/eci/shinyscreen/>), an open source software developed in house, coupled with custom-made scripts, revealed the presence of 162 potential pesticide masses and 135 potential TP masses in the samples. Further identification of these mass matches was performed using the open source MetFrag (<https://msbi.ipb-halle.de/MetFrag/>). Eventual target analysis of 36 suspects resulted in 31 pesticides and TPs confirmed at Level-1 (highest confidence), and five pesticides and TPs not confirmed due to different retention times. Spatio-temporal analysis of the results showed that TPs and pesticides followed similar trends, with a maximum number of potential detections in July. The highest detections were in the rivers Alzette and Mess and the lowest in the Sûre and Eisch. This study (a) added pesticides, classification information and related TPs into the open domain, (b) developed automated open source retrieval methods - both enhancing FAIRness (Findability, Accessibility, Interoperability and Reusability) of the data and methods; and (c) will directly support “L'Administration de la Gestion de l'Eau” on further monitoring steps in Luxembourg.

1 Introduction

Human and ecosystem exposure to a broad range of substances, including a multitude of new chemicals introduced into the environment necessitates careful and increasingly high throughput characterization and examination of their effects.¹ One substance group of high relevance for human health (both via food production but also for exposure) is pesticides. Despite their usefulness, they pose potential risks to food safety, the environment, and living organisms.^{2,3,4} For this reason, there is an increasing need for approaches to detect and identify them in environmental samples. Once pesticides are released to the environment, they (parent compounds) may be degraded by biotic or abiotic processes into one or more pesticide transformation products (TPs).^{5,6} Generally these compounds are thought to have lower toxicity to biota than the parent compounds, however in some instances TPs are more persistent, more mobile, and sometimes more toxic than the parent compound itself.^{6,7} Although parent compounds are assessed in detail in many regulatory schemes, the requirements for the assessment of TPs are less well developed.⁶ While their occurrence and significance are now reasonably well-known in research circles, it

is still surprisingly difficult to access information on TPs in a central and “FAIR” (Findable, Accessible, Interoperable and Reusable^{8,9}) manner, with much valuable information documented as detailed reaction schemes (*e.g.* as images) or descriptive text in regulatory reports that are not always easily or publicly accessible. In this study, not only is the presence of pesticides in samples investigated, but also the presence of their documented TPs in openly available information sources, as well as in the samples.

Previous work by Moschet *et al.*¹⁰ and Kiefer *et al.*¹¹ both characterised the relevance of pesticide transformation products in their findings and shared their lists afterwards (SWISSPEST¹² and SWISSPEST19¹³, respectively) on the NORMAN Suspect List Exchange (NORMAN-SLE)¹⁴, thus making them more “FAIR”⁸. The SWISSPEST suspect list was a starting point for the pesticide suspect list developed in this work, with additional chemicals of local relevance added as described below (note: SWISSPEST19 was published in parallel during the early stages of this work).

For the identification of unknown contaminants in the environment, a technology that is sensitive, fast, and accurate is required, capable of confidently identifying chemical contaminants emerging at trace concentrations in complex environmental and biological matrices. High resolution mass spectrometry (HR-MS) coupled with liquid chromatography has become an established technique for the monitoring of thousands of chemicals in water (and other) samples.^{15,16} Various computational approaches can help screen non-target HR-MS measurements for large numbers of suspect chemicals using suspect lists and/or mass spectral libraries^{15,17}, or to discover and identify new, previously unknown chemicals in the environment.^{15,18} These two non-targeted analysis strategies are called suspect screening and non-targeted screening, respectively.¹⁰ Suspect screening, the strategy used in this study, uses only the information of the chemical structure and its mass (and/or spectrum) *a priori* and is, therefore, a very promising approach for the efficient tentative identification of compounds.^{10,19} Consequently, suspect screening can be used to perform extensive analytical screening for specific chemicals suspected to be in the samples without necessarily the need for reference standards in advance.¹⁰

Targeted analysis is a more classical approach for quantification providing high sensitivity and high selectivity that requires preselection of the chemicals in advance and the availability of reference standards. Nevertheless, this approach is the only way to verify and quantify the tentative candidates in the end. The increasing number of chemicals of interest in environmental and exposomics studies makes it practically impossible for target analyses dependent on individual standards to cover all potentially occurring chemicals.¹⁰ Thus, suspect screening methods are therefore developed to reveal a fuller picture of occurring chemicals and can be performed with suspect chemical lists,^{10,15,16} allowing for eventual prioritization for target analysis and confirmation efforts.¹⁰

Confidence in HR-MS-based identifications inherently varies between compounds, since it is not always possible or reasonable to synthesize each substance or confirm them via complementary methods (*e.g.* nuclear magnetic resonance) at very low environmental concentrations and in complex mixtures.²⁰ These varying levels of confidence and the need for a standardized manner to report the results were motivating

reasons for a level system that was introduced in 2014.²⁰ The system contains five identification confidence levels, which can be achieved through experimental and computational analysis of the compound(s) measured in HR-MS experiments, with the objective to achieve the highest possible identification level that is realistic with the available evidence. Suspect screening can generally be considered to start at an identification confidence of Level-3 (tentatively detected candidates following pre-screening; see below), and through data analysis compounds can obtain the confidence Level-2a, *i.e.* probable structures via a high-quality spectral library match. Should target analysis reveal a suitable match with a reference standard measured in house with the same method, this results in a Level-1 confirmed identification.

Since a suspect list is often set up based on a substance class (or classes) of interest, there is no guarantee that the suspects are present in the sample. Thus, a pre-screening step helps to determine which suspects may be present with matching MS1 and MS2 spectra of sufficient quality for further data analysis. This step was performed using Shinyscreen (<https://git-r3lab.uni.lu/eci/shinyscreen/>)²¹, a semi-automated, open-source alternative to vendor software for peak inspection, with built in quality control criteria as described recently by Lai *et al.*²² Potential suspects with MS1 and MS2 spectra passing the Shinyscreen pre-screening were promoted to further for additional identification efforts via MS2 spectra annotation using the open source *in silico* fragmentation approach MetFrag (<https://msbi.ipb-halle.de/MetFrag/>).²³ MetFrag combines compound database searching and fragmentation prediction plus other experimental and metadata terms for molecule identification using HR-MS2 fragmentation information.²³ Given a single MS2 spectrum of a suspect and the neutral mass of the parent ion, MetFrag first selects matching candidates from databases, such as PubChem (<https://pubchem.ncbi.nlm.nih.gov/>)²⁴ and CompTox (<https://comptox.epa.gov/dashboard/>)²⁵, before each of the retrieved candidates is fragmented *in silico* using a bond-disconnection method and ranked using various scoring terms (see methods for further details).²³ For this study, the US Environmental Protection Agency (US EPA) CompTox Chemicals Dashboard was used as the main compound database, consistent with Lai *et al.*²², because of its relatively small size (~ 880,000 chemicals), and the extensive environmentally-relevant metadata such as toxicity, exposure, and presence integrated in CompTox from various information sources.²⁵ The recently-released PubChemLite for Exposomics collection²⁶, which demonstrated very good performance particularly for agrochemicals (pesticides) was under development at the time that this work was performed.

The main goals for this study were (a) establishing a new high-throughput suspect screening workflow based on open resources coupled with semi-automatic screening and annotation steps (b) the discovery and FAIRification of TP information based on their parent compounds using text-mining methods and (c) application of these combined approaches on surface water samples to gain an overview of the pesticide and pesticide TP presence in Luxembourgish rivers. The resulting suspect lists, classification and permission information were uploaded to various open databases and repositories to contribute to open and “FAIR” data management for exposomics.

2 Material And Methods

The high-throughput suspect screening workflow developed here is shown in **Figure 1** and explained in the following sections.

2.1 Experimental Methods

2.1.1 Sampling and Solid Phase Extraction

Different river surface water samples, collected throughout Luxembourg, were selected by the “L'Administration de la Gestion de l'Eau” (the Luxembourgish Water Administration, hereafter AGE) for chemical monitoring; pesticides and their TPs are the specific focus of these efforts (additional activities are ongoing). Nine different locations (Fig. 2 and Supplementary Materials (SM) **Table S1**) covered the various river catchments, and the data used in this study were sampled monthly between April 2019 and October 2019 (no sampling in June 2019).

The surface water samples were filled in 1000 mL amber bottles and stored for up to one week at 5°C (\pm 3°C) in darkness until extraction. To assess possible contamination from sample handling, ultrapure water was analogously enriched and analysed as blank samples.

For the solid-phase extraction, Atlantic® HLB SPE Disks from Horizon (Salem, NH, USA) with a 47 mm diameter were used. The disks were conditioned twice for 1 minute (min) with acetonitrile, and then twice for 1 min with Milli-Q water. 1000 mL of sample was pumped through each disk at a flow rate of roughly 30 mL/min, using the SPE-DEX 47900 system from Horizon. Sample loading was followed by washing the disks twice for 1 min with Milli-Q water and drying by airflow for 15 min. The analytes were eluted for 1 min with cyclohexane, followed by an acetone elution for 1 min, then 4 times for 1 min with acetonitrile. After each elution step, the disks were air-dried for 1 min. The combined extracts were dried under nitrogen flow in a water bath heated to 40°C. The samples were resuspended in 2 mL acetonitrile/water (10/90) by sonication for 5 min and remaining particles were removed by passing the extracts through a 0.7 μ m glass-fibre filter (Sartorius, Brussels, Belgium).

2.1.2 LC-HRMS Analysis

Reversed-phase chromatography was accomplished using an Acquity Ultra Performance Liquid Chromatography (UPLC) BEH C₁₈ column (dimensions: 1.7 μ m, 2.1 x 150 mm) from “Waters”. The flow was set to 0.20 mL/min using water (0.1 % formic acid, A) and methanol (B) as the mobile phase. The mobile phase gradient started at 90 % of A and 10 % of B at 0 min and was kept for 2 min before linearly ramping to 100 % B at 15 min. This condition was kept for another 5 min before bringing back to starting mobile phase conditions after 21 min. The column was allowed to re-equilibrate for 9 min before the next injection.

The mass spectrometer Q Exactive™ HF (Thermo Scientific) was used in both positive and negative electrospray ionization. The following: full MS/data dependent (dd) MS² settings were used: resolution

(1.2×10^5 at m/z 200), automatic gain control (AGC) target (1.0×10^6), maximum injection time (IT): (70 ms), and scan range ($m/z = 60-900$). For the dd-MS2/ddSIM (data dependent selected ion monitoring) the following were used: resolution (3.0×10^4 at m/z 200), AGC target (5.0×10^5), maximum IT (70 ms), loop count (5), Top N (5), isolation window (1.0 Da), (N)CE (30). Lastly the following dd settings were used: minimum AGC target (8.0×10^3 , intensity threshold (1.1×10^5), apex trigger (4–6 s), exclude isotopes (On), and dynamic exclusion (10 s). The instrument was calibrated and optimized every time an analysis was performed using manufacturer settings to ensure consistent performance throughout the whole study.

2.2 Computational Methods

2.2.1 Pesticide Substance Selection

The plant protection product list from the Luxembourgish “Administration des Services Techniques de l'Agriculture” (ASTA)²⁷ and the SWISSPEST list of registered insecticides and fungicides in Switzerland¹⁰ were used as starting points for the suspect list. Several (multilingual) documents provided by collaborators in the Clinical & Experimental Neuroscience group at the Luxembourg Centre for Systems Biomedicine as part of previous work²⁸ were also included, as documented in the “LUXPEST” dataset available on Zenodo²⁹ and briefly below.

The final LUXPEST pesticide suspect list included 386 pesticides,²⁹ classified into different classes along with information about their use authorisation in Luxembourg.^{30,31,32} Out of the 386 pesticides, 196 are permitted to be used in Luxembourg whereas 169 are not, while for 21 pesticides, no permission information was available. The classification efforts revealed that most of them were fungicides and herbicides (96 and 93 respectively); 49 were already classified as pesticide TPs (SM **Figure S1**). As a part of “FAIRifying” this dataset, the LUXPEST list is openly available on the NORMAN-SLE¹⁴, PubChem²⁴ and CompTox^{25,33} websites, and the detailed classification information was added to the PubChem NORMAN-SLE Classification Browser (<https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101>) and into the individual records for the pesticides (see SM **Figures S2 and S3**)

2.2.2 Suspect Screening of Pesticides and Transformation Products

2.2.2.1 Pre-screening with Shinyscreen

Pre-screening was performed using Shinyscreen²¹ with the following settings for extraction and automatic quality control (explained in greater detail in Lai *et al.*²²): coarse precursor m/z error ± 0.5 Da, fine precursor m/z error ± 2.5 ppm, extracted ion chromatogram m/z error ± 0.001 Da, retention time (RT) tolerance ± 0.5 min, an MS1 intensity threshold of 1.0×10^5 and an MS2 intensity threshold relative to the MS1 peak intensity of 0.05. Features that fulfilled the following four criteria were considered as passing

the quality control: 1) MS1 peak intensity > 1×10^5 , 2) presence of MS2 spectrum, 3) alignment of MS1 and MS2 peaks within the RT tolerance, 4) signal to noise ratio > 3.

2.2.2.2 Candidate Identification with MetFrag

The features that passed the quality control were then analysed using MetFrag²³ coupled to CompTox^{25,34} to achieve tentative identifications²³, generally consistent with Lai *et al.*²² Candidates were retrieved using an (exact mass + 10 ppm) window, where the exact mass settings included the measured ion mass plus adduct species ($[M + H]^+$ for positive and $[M - H]^-$ for negative mode, automatically detected from the ShinyScreen mode output) for internal correction to neutral mass in MetFrag for candidate retrieval. The InChIKey filtering (default setting) was left on, *i.e.*, candidates that vary only in the stereochemistry are merged in the output, and the highest scoring candidate is considered. Several MetFrag scoring terms were included. The two most relevant scoring terms for this study are the MetFrag *in silico* fragmentation score (settings: $mzabs = 0.001$; $frag_ppm = 5$; adduct setting as per candidate retrieval) and the MoNA (MassBank Of North America) score.³⁵ While MetFrag compares the experimental results with *in silico* fragmentation results, it also searches the experimental data with online mass spectral records from a public spectral library, MoNA, and presents these outcomes to users via the MoNA spectral similarity scoring term (hereafter "MoNA Score"). Several additional metadata terms were used in the MetFrag calculation (generally consistent with Lai *et al.*²²) but were not considered further here, yielding in the end a maximum score of 10 where every scoring term has the same weight (10 scoring terms each with a weight of 1). However, as described below, the MoNA Score became the primary decision-making criterion in this work. The additional scoring terms were CPDAT_COUNT, PUBMED_ARTICLES, DATA_SOURCES, PUBCHEM_SOURCES, TOXCAST_PERCENT_ACTIVE_BIOASSAYS, PREDICTED_EXPOSURE, KEMIMARKET_EXPO and KEMIMARKET_HAZ.

All the chemicals that achieved a MoNA Score greater than or equal to 0.9 (scoring range between 0 and 1) were assigned as Level-2a compounds according to the scheme described by Schymanski *et al.*²⁰ and as described above. In this study, four different MoNA score scenarios were defined in the context of the results available, also in line with commonly used thresholds in the community. The four scenarios were defined as the following: 1) "very good" describes the cases with a MoNA score equal or greater to 0.9, *i.e.*, a Level-2a, 2) "good" describes the cases with a MoNA score between 0.7 and 0.9, which can be considered in some cases sufficient for Level-2a but based on experience not always sufficient; 3) "poor" describes the cases with a MoNA score between greater than 0 and smaller than 0.7 and 4) "no spectrum" describes the cases with a MoNA score equal to 0. The first scenario led to a Level-2a as described above and the three other scenarios remained at a Level-3 for further inspection.

2.2.3 Extracting Pesticide Transformation Product Information

2.2.3.1 Transformations

In a collaborative effort between PubChem and the NORMAN-SLE, several lists of chemicals including parent-TP information were mapped up into a standardized format and added into PubChem as “Transformations”, as described elsewhere²⁶ (see **Fig. 3**).

The so-called “parents” were termed “predecessor” to avoid terminology clashes (as the term “parent” has a different meaning in PubChem), and the TPs or metabolites were termed “successors” in PubChem. At the time this study was performed, the NORMAN-SLE lists included were S60 SWISSPEST19¹³ and S66 EAWAGTPS³⁶. The deposition of “Transformation” information in PubChem is automated through the NORMAN-SLE via Zenodo depositions³⁷ and mapping files in GitLab³⁸. The retrieval of this information is made possible through PubChem via a structured data query (SDQ) per PubChem Compound Identifier (CID), which can be performed *e.g.*, through the web interface via the download button (**Figure 3** top right) or via scripting queries. Custom-made R functions were designed to access this as a part of this work.³⁹

2.2.3.2 Hazardous Substance Database (HSDB) Metabolites

A further information source of TPs within PubChem is the “Metabolism and Metabolites” section which, unlike the table above, are human-readable text excerpts from several data sources, including the Hazardous Substance Database (HSDB) from the US National Library of Medicine (NLM), recently fully integrated within PubChem. As a pilot project as part of this work, a data extraction workflow was designed based on the HSDB annotation file (available in JavaScript Object Notation - JSON format). In short, text excerpts are automatically screened for recognized synonyms PubChem-side and, where detected, hyperlinked (shown as blue text in **Figure 4**, and recognizable in the annotation file by CID).

This information can be automatically retrieved from the JSON file. Additionally, the text also contains many descriptive reactions that are not suitable for automated synonym recognition, but interpretable by chemists. Thus, information was automatically extracted in a tabular form for manual curation (*e.g.*, removal of irrelevant matches, addition of new chemicals) with full provenance suitable for conversion into a “Transformations” table, coupled with an accompanying structure file to deposit new structures in PubChem. Chemical drawing and curation were performed in Chemistry Development Kit (CDK) using CDK Depict (<https://www.simolecule.com/cdkdepict/depict.html>)^{40,41}. To describe the challenges visually, the predecessor (Fig. 4, atrazine) is circled in purple and was automatically extracted, along with two TPs 2-hydroxyatrazine (red; two different synonyms mapping to the same structure) and 2-hydroxydesethylatrazine (orange, three synonyms; not each synonym was recognised fully). Text-mined entries retrieved in this manner are circled in full lines. Desethylatrazine was not automatically recognised (no blue hyperlink present) but was curated and added in manually (blue dotted lines). The synonym “hydroxy” was automatically mapped (blue hyperlink, green dashed circle) but removed in the manual curation step as an artefact of the mapping.

All HSDB TPs extracted in this manner were added to a new suspect list S68 HSDBTPS⁴² and full provenance of the curation is available on the Environmental Cheminformatics GitLab repository⁴³.

2.3 Verification and Quantification using Reference Standards

All the pesticides at a Level-2a were selected for further verification via reference standards analysed with the same chromatographic parameters and procedures as for the sample analysis. Several reference standards came from the in house available ENTACT mixtures, obtained from participation in the EPA's Non-Targeted Analysis Collaborative Trial.⁴⁴ Retention times were considered a match if the difference was less than ± 0.2 min. Additional reference standards were purchased where possible (SM **Table S2**). Where reference standards were available, the concentration of the pesticides and TPs were quantified using an external calibration curve ranging from 1ppb to 1000ppb spanning the linear dynamic range for the compounds quantified. Thermo Scientific TraceFinder™ Software (version 5.1) was used for automatic peak integration and generation of the calibration curve. Concentrations below 1ppb were reported to be below the quantifiable range.

3 Results

The numbers that will be explained in detail in the next sections are summarized in a table (SM **Table S3**), to provide an overview of the number of cases and/or compounds for each step of the workflow.

3.1 Tentatively Detected Pesticides

Shinyscreen was run with the 386 LUXPEST²⁹ suspects (SM **Table S4**) on river water samples from nine locations over six months and for two modes (positive and negative), comprising 20,844 cases for the automated quality control protocol. In total, there were 3,006 cases deemed suitable for further identification with MetFrag, corresponded to 162 unique compounds (SM **Table S5**). **Figure 5** illustrates the number of cases for each location and for each month.

For example, in April 2019 the river "Sûre" in Erpeldange revealed 44 cases that passed the quality check in Shinyscreen. These were subsequently analyzed and annotated with MetFrag to assign an identification confidence level.

3.2 Pesticide Annotation with MetFrag

The 3,006 cases were categorized into four different scenarios depending on their MoNA score, as shown in **Figure 6**.

3.3 Pesticide Transformation Products Suspect List

Out of the 386 compounds, 162 different pesticides were found (tentatively, at Level-2a or Level-3 confidence) in either one or more locations over six months. Since the manual curation of HSDB content is complex and time-consuming, only the 36 previously selected Level-2a pesticides (suspects with a MoNA score > 0.9) were selected (SM **Table S6**) for further retrieval of TP information from PubChem. Of the 36 pesticides, there were 30 that already had information in the "Transformations" section. In

addition, 22 pesticides had further information in the HSDB Metabolism and Metabolites section, while no information was available for only 3 pesticides. There were 19 pesticides that had information in both the HSDB and “Transformations” section.

In the end, a new suspect list of 181 transformation products and their parent compounds was created, including the 36 parent compounds (the Level-2a cases identified earlier) and 173 TPs related to these 36 pesticides that were added in this step. Although the parent compounds were already analysed previously, they were retained for a direct comparison between the presence of the parent compounds and their TPs (see discussion). This table is given in the SM, **Table S6**.

After manual curation, the merged data file of TPs extracted from HSDB was added to Zenodo as HSDBTPS⁴² and the newly generated information was also provided to PubChem as “Transformation” tables to update this section as well (also included in the Zenodo deposition). The HSDBTPS list is also available in CompTox.⁴⁵

3.4 Suspect Screening for the Pesticide TPs

ShinyScreen was run again for all samples with 181 pre-selected compounds (SM **Table S7**), resulting in a total of 19,548 cases. Of these, there were 1,275 cases in negative mode and 2,159 cases in positive mode that were able to pass the quality check. Since some suspects were detected in different locations in positive and negative ionization mode, these 3,434 cases corresponded to 99 transformation products (SM **Table S8**) and the 36 parent compounds (135 different compounds in total). The number of cases for each location and month is available in the SM, **Figure S4**.

The MS2 spectra of 135 tentatively identified suspects were then processed using MetFrag with the same databases and scoring terms as before and the identification confidence levels were determined based on the MoNA scores (SM **Figure S5**). Out of the 3,434 cases, there were 1,190 were able to achieve a MoNA score above 0.9 corresponding to eight unique additional TPs (SM **Table S8**).

3.5 Verification of the Tentative Candidates and Their Quantification

The 36 Level-2a pesticide identifications were selected for further confirmation efforts with reference standards (SM **Table S9**). Of these, 26 of these were verified using single standards and 10 compounds were verified with reference standards contained in the ENTACT mixtures (the work on the TPs had not yet been performed when this selection was made).

Out of the 36 parent compounds, there were 31 chemicals that achieved a Level-1, while five could not be confirmed (different retention times, see SM **Table S9**). Of the 31 Level-1 compounds, only 20 were present at quantifiable amounts (within the scope here), as presented in **Fig. 7** (see also SM **Table S10** and **Table S11**).

The classification and Luxembourgish permission information for the 20 quantified compounds are summarized in SM **Figure S6**.

3.6 Spatial and Temporal Distribution

Figure 8 visualizes: (A) the nine different river locations that were selected with the average number of detections; and (B) the number of detections over the six months. The green lines show the pesticide suspects (Level-1 through Level-3), the yellow the TP suspects (Level-1 through Level-3) and the red lines indicate the confirmed identifications (all 20 Level-1 compounds that were additionally quantified).

4 Discussion

This work aims for a more dynamic experience of suspect screening in non-target environmental HR-MS measurements, using open cheminformatics approaches and tentative detections in samples, while using Luxembourgish river samples as an example. The discussion will look into how the coupling of parent and TP information can support interpretation using the example of terbutylazine, then look at the overall implications of these results for Luxembourg, before delving into the FAIRification of TP data and the implications for further efforts.

4.1 Example of Pesticide-TP Screening: Terbutylazine

The following example of terbutylazine and three TPs visualizes how the coupling of suspect screening for pesticides and transformation products can be automated and visualized in Shinyscreen. Figure 9 shows three different plots belonging to one parent compound (terbutylazine, top, suspect list ID N° 3) with three TPs, 2-hydroxyterbutylazine (ID N° 11), desethyl-2-hydroxyterbutylazine (ID N° 4), and desethylterbutylazine (ID N° 2).

The parent compound was found in the months May, July and September at the identification Level-2a, retention time of ~ 17.41 min, with two isobars found at ~ 16.00 and 14.63 min. These isobars are speculated to be other compounds in this case; MetFrag suggested for both the compound propazine, due to highest metadata scores with the selected scoring terms in CompTox (specifically due to higher toxicity concerns and some higher reference counts); propazine was also reported as a suspect by many in the 2015 NORMAN Collaborative Trial⁴⁶, although it has not been permitted for use for many years. Interestingly, the use of PubChemLite with the optimized default scoring terms²⁶ resulted in terbutylazine appearing ahead of propazine in the metadata ranking; further addition of the "agrochemicals" category²⁶ helps up-prioritize the potentially most relevant alternative isobars for further consideration at a later stage (e.g. sebutylazine). The importance of the choice of the various CompTox metadata terms and the resulting consequences in interpretation are discussed in detail in Lai *et al.*²² and thus not discussed further here.

One of the main TPs, desethylterbutylazine (ID N° 2, 4th chromatogram in the **Figure 9**) involves the loss of the ethyl group and is detected at 15.6 min at high intensity in July and October. Since one ethyl is lost, a lower (but not dramatically lower) retention time than the parent would be expected on a reverse phase

column, thus the detection at 15.6 min is considered more plausible than other peaks reported at 9.0 min for other months. The fact that the TP peak does not occur at the same time as the parent rules out the possibility of an in-source fragmentation from the parent. After the verification with reference standards, it became clear that the retention time of desethylterbutylazine is indeed 15.67 min (the isobar, simazine, was confirmed at RT of 15.24 min, see SM **Table S9**). The third TP, desethyl-2-hydroxyterbutylazine (ID N° 4) is detected at 9 minutes in May, July, August and September (at Level-3), which coincides with parent detections plus another month where the parent was not detected. Since the chlorine has been replaced by an oxygen, combined with the ethyl group, the dramatic reduction of retention time relative to the parent is plausible, as both transformations increase the polarity and thus reduce the retention time. The last TPs of terbutylazine is terbutylazine-2-hydroxy (ID N° 11) containing an oxygen instead of a chlorine as well. This compound was found for all months and since this TP can be a degradation compound from different parent compounds (e.g. terbutylazine found at Level-3 and terbutryn found at Level-1 amongst others, it could be present due to the transformation from both, see <https://pubchem.ncbi.nlm.nih.gov/compound/135495928#section=Transformations>).

4.2 Pesticides and TPs in Luxembourgish Surface Waters

The fact that half of the detected and quantified suspects are not permitted for use in Luxembourg (see SM **Figure S6**) will be investigated further by AGE. Several reasons could contribute to this: either these pesticides were allowed in the past and their presence is due to historical use; or these pesticides are applied without permission (considered unlikely based on the results here); five of the entries were TPs that are not permitted for use. Looking at the permission information of their parent compounds revealed that for some banned TPs (e.g. 2-hydroxyatrazine) the parent compound is banned as well (atrazine), but for others (e.g. desethylterbutylazine) the parent compound is permitted (terbutylazine). As an example, the low levels of atrazine detected here (< 100 ppt) are likely to be due to historical applications still seeping into the surface waters; fresh applications would likely yield higher levels.

As shown in **Fig. 7** (all the concentrations are available in the SM, **Table S10**), the pesticide TP succinic acid was found in highest concentrations (maximum concentrations found: 773.52 parts per trillion = 0.77 ng/L) in the river samples. This high concentration is most probably due to the fact that this chemical has several "roles" in the environment and can come from both natural and anthropogenic sources. For instance, succinic acid is involved in several processes in the body (e.g., generated in mitochondria via the citric acid cycle) and is also a food additive⁴⁷; thus alternative sources are likely to be much higher contributors to the overall concentrations than this being a documented TP of the pesticides sulcotrione (present in the LUXPEST list but did not pass the pre-screening) and linuron (not present in the LUXPEST list). This shows the importance of having information about the multiple roles of chemicals available in an easily accessible and readable manner. The overall lowest concentrations were found for the compounds desethylatrazine, 2-hydroxyatrazine and simazine (minimal concentrations around 0.001 ng/L). Returning to the example from the section before (**Sect. 4.1**), desethylterbutylazine was confirmed in 8 out of 9 river samples (except for the river Alzette from Mersch-Berschbach), in all the 6 months (SM **Table S10**).

As shown in Fig. 8, the overall lowest average number of compounds were found in the rivers Eisch and Sûre, which is reassuring in the context of Luxembourg as about one-third of the drinking water originates in the river Sûre.⁴⁸

The temporal patterns (Fig. 8B) show that there is a spike in detections in late spring/beginning of the summer, with an additional smaller spike in September. The overall lowest average number of compounds was found in April, reflecting the expected seasonality of the pesticide application. All screening results presented here have been communicated with AGE for consideration in their subsequent monitoring efforts; while this article presents the results from April-October 2019, these collaborative non-target screening efforts are also still continuing.

4.3 Pre-screening and Annotation Workflow

During pre-screening, all the files were loaded into ShinyScreen, corresponding to a total of 41,688 cases and graphs (386 pesticides times two modes times six months times nine locations: $386 \times 2 \times 6 \times 9 = 41,688$) that were analysed. The manual inspection revealed that for the majority of cases, an empty graph was obtained leading to the conclusion that most suspects were not present in the samples. This demonstrates the need for such a semi-automated procedure, since it makes visualizing and checking the experimental data very efficient and easy. In the end, there were 3,006 cases retained that passed the quality checks were retained, leading to a final set of 162 different tentatively identified compounds. This means that 42 % of the compounds that were screened with ShinyScreen may be present in at least one of the samples.

Some of these 162 compounds were detected in multiple locations and the comparison between the retention times for the different locations revealed two general trends. The first trend shows a subtle difference (*e.g.*, ± 0.5 min) in the retention times, which is probably the consequence of fluctuations in the liquid chromatography. The second trend shows wide differences in retention times (several minutes) leading to the conclusion that only one of these signals could potentially belong to the suspect, whereas the other signals most likely belong to different (isobaric, *i.e.*, same mass) substances. For example, ShinyScreen suggested that the compounds 3-hydroxybenzoic acid and 4-hydroxybenzoic acid (both isobaric) are present in the samples and the automatic retrieved retention time was equal to 14.89 min (default behaviour extracts the retention time of the most intense peak). However, in the end, through the verification with reference standards, the results showed that the compound in the sample was salicylic acid since only the reference standard for this compound had a retention time of 14.9 min and the ones from 3-hydroxybenzoic acid and 4-hydroxybenzoic acid differed (12.04 min and 10.83 min respectively). ShinyScreen has subsequently been upgraded to offer more extensive isobar handling during pre-screening (release 1.0.0, 2nd April 2021); the MetFrag post-processing has also been correspondingly updated and, as discussed above, the metadata scoring terms integrated into PubChemLite have also made data interpretation of relevant isobars both easier and more powerful²⁶.

During the analysis of the MetFrag results, the months, modes and locations were considered together. At first, the MoNA score is investigated and out of the 3,006 cases: 719 cases obtained a very good, 118 a

good, and 663 a poor MoNA score. Additionally, in 1,506 cases the MoNA score was equal to 0 (no spectrum matching or available in the library). In consequence, for 719 cases an identification of Level-2a can be achieved and for the remaining 2,287 cases, a Level-3 is attained (Fig. 6). When looking at the level of unique pesticides, out of the 162 pesticides, there are 140 that remain at an identification of Level-3, while 36 obtained a Level-2a based on MoNA scores and further metadata analyses (SM **Table S5**).

For the TPs, there were 19,548 cases and graphs (181 pesticides times two modes times six months times nine locations) were analyzed. Out of these, there were 3,434 cases that passed the quality check and kept for further analysis. This leads to a final number of 99 newly identified compounds (135 compounds in total – 36 known pesticides = new compounds 99). When excluding the 36 parent compounds, this led to eight TPs with a very good, one with a good, and nine with a poor MoNA score. The remaining 81 pesticides (out of 99) had no spectrum available in MoNA, showing the importance of additional community contributions to open resources to help fill these data gaps in the future.

For the tentative identification with MetFrag, only the spectral-based scoring terms were investigated here, namely the MetFrag *in silico* fragmentation and primarily the MoNA similarity score. None of the additional metadata scores were used, as prioritization was done purely based on achieving a very good MoNA score for highest confidence. The work described here also helped contribute to the conceptual design of the PubChemLite for Exposomics collection, where the category of chemical (*e.g.* agrochemical/pesticide or pharmaceutical) can be used in interpretation and even scoring. The performance described elsewhere²⁶ demonstrated that the interpretation of results can be improved with this additional information, achieving up to 90 % annotation success for the agrochemicals (pesticides) in the benchmarking set. Efforts are underway to streamline the coupling of suspect + TP screening together with ShinyScreen, MetFrag and PubChemLite in a smooth workflow on the foundation of the work described here, including the collapsing of many "Cases" into unique compounds much earlier in the workflow.

4.4 Open Pesticide and Transformations Data

Out of the 386 selected pesticides, 196 are permitted and 169 are forbidden in Luxembourg (SM **Table S4**) and could be classified into six main categories (SM **Figure S1**). This information can be browsed in PubChem under LUXPEST at <https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101> (SM **Figure S2**) and this information is incorporated into the individual records in PubChem (Example in the SM **Figure S3**). This information flow helps create the annotation categories that form the PubChemLite for Exposomics collection (see Schymanski *et al.*²⁶ Fig. 1) and provide PubChem users with additional expert knowledge for interpretation of their results. Ensuring this continual flow of information is a major motivating factor for increasing the FAIRness of datasets and thus the upload of the datasets to different open access databases (CompTox, PubChem) and repositories (NORMAN-SLE, Zenodo), as well as the integration of the classification (SM **Figure S2**) and regulation information in Luxembourg into PubChem. Since the NORMAN-SLE compound lists are "FAIR" due to the Zenodo deposition with explicit license

declaration, they can be used by PubChem directly to create automatic workflows to build the Transformations section; other users and resources are also able (and encouraged) to re-use this data as they wish. By adding chemical identifiers to the historical information retrieved from the HSDB via text-mining methods and adding this as a new suspect list to the NORMAN-SLE, the original source (HSDB) can be credited, and the value-added data fed back into PubChem as transformations for improved automated retrieval in future screening activities, so that this information is now available in both human and machine-readable forms.

Several transformations tables have now been added to PubChem, including HSDBTPS as a part of this work. The manual curation involved with the text-mined information was the most time-consuming part of this process and was thus only performed on the 36 Level-2a pesticides that were selected from the first analysis due to their very good MoNA score. Of these, it was possible to generate transformation products for 33 compounds (no compounds were found in HSDB or the "Transformation" table in PubChem for the remaining three compounds). In the end, there were 22 entries from HSDB extracted and manually curated (files available from GitLab⁴⁹), resulting in 226 new transformation reactions with full literature provenance, and five new structural records in PubChem (CIDs 146035700, 146035701, 146035702, 146035703 and 146037633). In the end, a total of 145 transformation products were added to the 36 pesticides, which results in a suspect list of 181 compounds. Since this work was performed several other datasets have been added to the Transformations tables including MetXBioDB⁵⁰ from BioTransformer⁵¹ and it is highly likely that the numbers of pesticide TPs retrieved for screening would be higher now.

This work was only possible through the exchange of information between the NORMAN-SLE and PubChem and, at this pilot stage, willingness on both sides to develop unconventional workflows not originally foreseen for either resource. While the R scripts developed are certainly functional, several optimizations are possible. In hindsight, the created workflow with this integrated script helped the authors discover and upload relationships between pesticides and their TPs to PubChem as well as identifying areas to improve the information flow in the future. Future efforts are already underway to streamline this further based on this pilot project, to develop even more automated forms of this workflow and to ensure easy, fast and accurate suspect and TP list generation from their parent compounds. All data transfer between the NORMAN-SLE and PubChem includes full provenance to the original literature sources. Since all "Transformations" entries were based on existing suspect lists or resources, it is quite resource intensive to add existing knowledge involving only a few entries. As a result, a new list, REFTPS⁵² (currently only with very few entries) has been created to provide a pathway to add single or small numbers of transformations resulting from individual studies, such as 6PPD-quinone from Tian *et al.*⁵³ Overall, these pilot efforts have already caught the interest of several other workflows and are being integrated into the open source HR-MS workflow patRoom¹⁸, amongst others.

5 Conclusion

This study describes open cheminformatics approaches to screen for emerging contaminants (in this case pesticides) and their TPs in non-target HR-MS measurements. The coupling of major open resources such as the environmental knowledge within the NORMAN-SLE with the largest open chemical database PubChem has enabled the exchange and enhancement of information on pesticides and their TPs both in the context of Luxembourg and in the context of dynamic suspect screening (*i.e.*, the automated retrieval of TPs related to suspects detected at a Level-2a or more for subsequent screening and recognition). Through the detailed annotation content added to PubChem, it would now also be feasible to perform this in reverse, *i.e.*, form a suspect list purely on known TPs for screening proactively in samples, without the explicit presence of the parent, expanding the window beyond what was done here. The coupling of extensive suspect lists with an efficient pre-screening method such as ShinyScreen with tentative annotation approaches such as MetFrag will pave the way for higher throughput screening of exposomics samples in many contexts, as showcased here for pesticides.

In terms of local outcomes, these efforts (and parallel efforts investigating other substances classes) are continuing and the results are being exchanged with AGE to help improve monitoring efforts and thus human and environmental health in Luxembourg, above and beyond the current EU requirements.

Supplementary Material

Two supplementary data files are provided, a document containing Figures S1 through to S8, and an excel file containing Tables S1 to S11. For details about the code, software, suspect list and raw file availability, see Data Statement.

Declarations

7 CRediT Author Statement

Jessy Krier: Conceptualization, methodology, software, validation, formal analysis, investigation data curation, visualization, writing – original draft, writing - review and editing. **Randolph R. Singh:** conceptualization, methodology, validation, supervision, writing – review and editing. **Todor Kondic:** software, resources, writing – review and editing. **Adelene Lai:** software, data curation, visualization, writing – review and editing. **Philippe Diderich:** conceptualization, methodology, validation, resources, project administration, writing – review and editing. **Jian Zhang:** methodology, software, data curation, writing – review and editing. **Paul A. Thiessen:** software, validation, data curation, writing – review and editing. **Evan E. Bolton:** conceptualization, methodology, validation, resources, supervision, writing – review and editing. **Emma L. Schymanski:** conceptualization, methodology, software, validation, formal analysis, resources, data curation, supervision, project administration, funding acquisition, writing – original draft, writing - review and editing.

8 Funding

TK, AL and ELS acknowledge funding support from the Luxembourg National Research Fund (FNR) for project A18/BM/12341006. The work of EEB, PAT, and JZ was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

9 Declaration of Competing Interest

The authors declare that they have no competing interests.

10 Acknowledgements

The authors would like to acknowledge the contributions of other members of the Environmental Cheminformatics and PubChem teams to this work, as well as all contributors to the open resources used in this work. We gratefully acknowledge documents received from Dr. med. Pierre Kolber, from the Clinical and Experimental Neuroscience group at the Luxembourg Centre for Systems Biomedicine (LCSB) and the samples provided by “L'Administration de la Gestion de l'Eau” (AGE).

11 Data Statement

The suspect lists LUXPEST and HSDBTPS developed in this work are online on Zenodo (DOI: [10.5281/zenodo.3862688](https://doi.org/10.5281/zenodo.3862688) and DOI: [10.5281/zenodo.3827487](https://doi.org/10.5281/zenodo.3827487)), and CompTox (https://comptox.epa.gov/dashboard/chemical_lists/LUXPEST and https://comptox.epa.gov/dashboard/chemical_lists/HSDBTPS). Both lists are accessible in PubChem (<https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101>) and NORMAN-SLE (<https://www.norman-network.com/nds/SLE/>). The raw files are available as dataset MSV000087190 from the GNPS Massive repository (<https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp>), citable under DOI: [10.25345/C5D81C](https://doi.org/10.25345/C5D81C) and accessible via <ftp://massive.ucsd.edu/MSV000087190/> and <https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?accession=MSV000087190>. The major software used Shinyscreen (<https://git-r3lab.uni.lu/eci/shinyscreen/>) and MetFrag (<http://ipb-halle.github.io/MetFrag/>) are both open source; the code, functions and files associated with this manuscript are available from the ECI GitLab repository (<https://git-r3lab.uni.lu/eci/pubchem>). In addition, two supplementary files are provided with this manuscript, as detailed above.

References

1. Escher, B. I., Stapleton, H. M. & Schymanski, E. L. Tracking complex mixtures of chemicals in our changing environment. *Science* **367**, 388–392 (2020).
2. Calzada, J., Gisbert, M. & Moscoso, B. *The Hidden Cost of Bananas: Pesticide Effects on Newborns' Health*. <https://papers.ssrn.com/abstract=3786643> (2021) doi:10.2139/ssrn.3786643.
3. Mahmood, I., Imadi, S. R., Shazadi, K., Gul, A. & Hakeem, K. R. Effects of Pesticides on Environment. in *Plant, Soil and Microbes: Volume 1: Implications in Crop Science* (eds. Hakeem, K. R., Akhtar, M. S. & Abdullah, S. N. A.) 253–269 (Springer International Publishing, 2016). doi:10.1007/978-3-319-27455-3_13.

4. Hernández, A. F., Bennekou, S. H., Hart, A., Mohimont, L. & Wolterink, G. Mechanisms underlying disruptive effects of pesticides on the thyroid function. *Current Opinion in Toxicology* **19**, 34–41 (2020).
5. Somasundaram, L. & Coats, J. R. Pesticide Transformation Products in the Environment. In *Pesticide Transformation Products* (eds. Somasundaram, L. & Coats, J. R.) vol. 459 2–9 (American Chemical Society, 1991).
6. Sinclair, C. J. & Boxall, A. B. A. Assessing the Ecotoxicity of Pesticide Transformation Products. *Environ. Sci. Technol.* **37**, 4617–4625 (2003).
7. Olsson, O. *et al.* Fate of Pesticides and Their Transformation Products: First Flush Effects in a Semi-Arid Catchment. *Clean Soil Air Water* **41**, 134–142 (2013).
8. Sansone, S.-A. *et al.* FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology* **37**, 358–367 (2019).
9. FAIR Principles. *GO FAIR* <https://www.go-fair.org/fair-principles/>. (2021). Accessed 29/04/2021.
10. Moschet, C., Piazzoli, A., Singer, H. & Hollender, J. Alleviating the Reference Standard Dilemma Using a Systematic Exact Mass Suspect Screening Approach with Liquid Chromatography-High Resolution Mass Spectrometry. *Anal. Chem.* **85**, 10312–10320 (2013).
11. Kiefer, K., Müller, A., Singer, H. & Hollender, J. New relevant pesticide transformation products in groundwater detected using target and suspect screening for agricultural and urban micropollutants with LC-HRMS. *Water Research* **165**, 114972 (2019).
12. Moschet, C. S11 | SWISSPEST | Swiss Insecticides, Fungicides and TPs. (2017) doi:10.5281/zenodo.2623741.
13. Kiefer, K., Müller, A., Singer, H. & Hollender, J. S60 | SWISSPEST19 | Swiss Pesticides and Metabolites from Kiefer et al 2019. (2020) doi:10.5281/zenodo.3766352.
14. NORMAN Suspect List Exchange (NORMAN SLE). *NORMAN* <https://www.norman-network.com/nds/SLE/>. (2021). Accessed 29/04/2021.
15. Hollender, J., Schymanski, E. L., Singer, H. P. & Ferguson, P. L. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? *Environ. Sci. Technol.* **51**, 11505–11512 (2017).
16. Krauss, M., Singer, H. & Hollender, J. LC–high resolution MS in environmental analysis: from target screening to the identification of unknowns. *Anal Bioanal Chem* **397**, 943–951 (2010).
17. Vinaixa, M. *et al.* Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *TrAC Trends in Analytical Chemistry* **78**, 23–35 (2016).
18. Helmus, R., ter Laak, T. L., van Wezel, A. P., de Voogt, P. & Schymanski, E. L. patRoom: open source software platform for environmental mass spectrometry based non-target screening. *Journal of Cheminformatics* **13**, 1 (2021).
19. Moschet, C. *et al.* How a Complete Pesticide Screening Changes the Assessment of Surface Water Quality. *Environ. Sci. Technol.* **48**, 5423–5432 (2014).

20. Schymanski, E. L. *et al.* Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.* **48**, 2097–2098 (2014).
21. Kondic Todor *et al.* *Shinyscreen*. <https://git-r3lab.uni.lu/eci/shinyscreen/> (2020). Accessed 29/04/2021.
22. Lai, A. *et al.* Retrospective non-target analysis to support regulatory water monitoring: from masses of interest to recommendations via in silico workflows. *Environmental Sciences Europe* **33**, 43 (2021).
23. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform* **8**, 3 (2016).
24. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research* **47**, D1102–D1109 (2019).
25. Williams, A. J. *et al.* The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *Journal of Cheminformatics* **9**, 61 (2017).
26. Schymanski, E. L. *et al.* Empowering large chemical knowledge bases for exposomics: PubChemLite meets MetFrag. *Journal of Cheminformatics* **13**, 19 (2021).
27. Administration of Technical Services (ASTA) – Ministry of Agriculture, Viticulture and Rural Development // The Luxembourg Government. <https://ma.gouvernement.lu/en/administrations/asta.html>. (2021). Accessed 29/04/2021.
28. Schymanski, E. L. *et al.* Connecting environmental exposure and neurodegeneration using cheminformatics and high resolution mass spectrometry: potential and challenges. *Environ. Sci.: Processes Impacts* **21**, 1426–1445 (2019).
29. Krier, Jessy. S69 | LUXPEST | Pesticide Screening List for Luxembourg. (2020) doi:10.5281/zenodo.3862688.
30. European Commission. EU Pesticides database. <https://ec.europa.eu/food/plant/pesticides/eu-pesticides-database/public/?event=activesubstance.selection&language=EN>. (2021). Accessed 29/04/2021.
31. University of Hertfordshire. Pesticide Properties DataBase Search. <https://sitem.herts.ac.uk/aeru/ppdb/en/search.htm>. (2021). Accessed 29/04/2021.
32. University of Hertfordshire. Bio-Pesticides DataBase Search. <https://sitem.herts.ac.uk/aeru/bpdb/search.htm>. (2021). Accessed 29/04/2021.
33. Krier, Jessy. Pesticide Screening List for Luxembourg. *CompTox* https://comptox.epa.gov/dashboard/chemical_lists/LUXPEST. (2021). Accessed 29/04/2021.
34. Schymanski, E. MetFrag Local CSV: CompTox (7 March 2019 release) Wastewater MetaData File. (2019) doi:10.5281/zenodo.3472781.
35. MassBank of North America. <https://mona.fiehnlab.ucdavis.edu/>. (2021). Accessed 29/04/2021.
36. Schollee, J. & Schymanski, E. S66 | EAWAGTPS | Parent-Transformation Product Pairs from Eawag. (2020) doi:10.5281/zenodo.3754449.

37. Schymanski, E. L. & Williams, A. J. Open Science for Identifying “Known Unknown” Chemicals. *Environ. Sci. Technol.* **51**, 5357–5359 (2017).
38. Environmental Cheminformatics, GitLab Repository pubchem. <https://git-r3lab.uni.lu/eci/pubchem>. (2021). Accessed 29/04/2021.
39. Environmental Cheminformatics, GitLab Repository pubchem Annotations.R <https://git-r3lab.uni.lu/eci/pubchem/-/blob/master/annotations/tps/extractAnnotations.R>. (2021). Accessed 29/04/2021
40. Willighagen, E. L. *et al.* The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of Cheminformatics* **9**, 33 (2017).
41. Mayfield, J. The Chemistry Development Kit (CDK) Depict. <https://github.com/cdk/depict>. (2021). Accessed 29/04/2021.
42. LCSB-ECI *et al.* S68 | HSDBTPS | Transformation Products Extracted from HSDB Content in PubChem. (2020) doi:10.5281/zenodo.3830987.
43. Environmental Cheminformatics, GitLab Repository pubchem HSDB <https://git-r3lab.uni.lu/eci/pubchem/-/tree/master/annotations/tps/HSDB>. (2021) Accessed 29/04/2021.
44. Ulrich, E. M. *et al.* EPA's non-targeted analysis collaborative trial (ENTACT): genesis, design, and initial findings. *Anal Bioanal Chem* **411**, 853–866 (2019).
45. US EPA, https://comptox.epa.gov/dashboard/chemical_lists/HSDBTPS. (2021) Accessed 29/04/2021.
46. Schymanski, EL *et al.* Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis, *Anal Bioanal Chem.* **407** (21), 6237–6255(2015). DOI 10.1007/s00216-015-8681-7.
47. PubChem. Succinic acid. <https://pubchem.ncbi.nlm.nih.gov/compound/11110>. (2021). Accessed 29/04/2021.
48. Grand Duchy of Luxembourg. Woher kommt unser Trinkwasser (Where does our drinking water come from?). <http://infocrise.public.lu/de/eau-potable/informations-generales/origine-de-notre-eau-potable.html>. (2021). Accessed 29/04/2021.
49. Environmental Cheminformatics, pubchem (HSDB_done repository). https://git-r3lab.uni.lu/eci/pubchem/-/tree/master/annotations/tps/HSDB/HSDB_done. (2021). Accessed 29/04/2021.
50. Djombou-Feunang, Y., Schymanski, E., Zhang, J. & Wishart, D. S. S73 | METXBIODB | Metabolite Reaction Database from BioTransformer. (2020) doi:10.5281/zenodo.4094568.
51. Djombou-Feunang, Y *et al.* BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *Journal of Cheminformatics* **11**:2 (2019). DOI: 10.1186/s13321-018-0324-5.
52. Schymanski, E. S74 | REFTPS | Transformation Products and Reactions from Literature. (2020) doi:10.5281/zenodo.4318852.

53. Tian, Z. *et al.* A ubiquitous tire rubber-derived chemical induces acute mortality in coho salmon. *Science* **371**, 185–189 (2021).

Figures

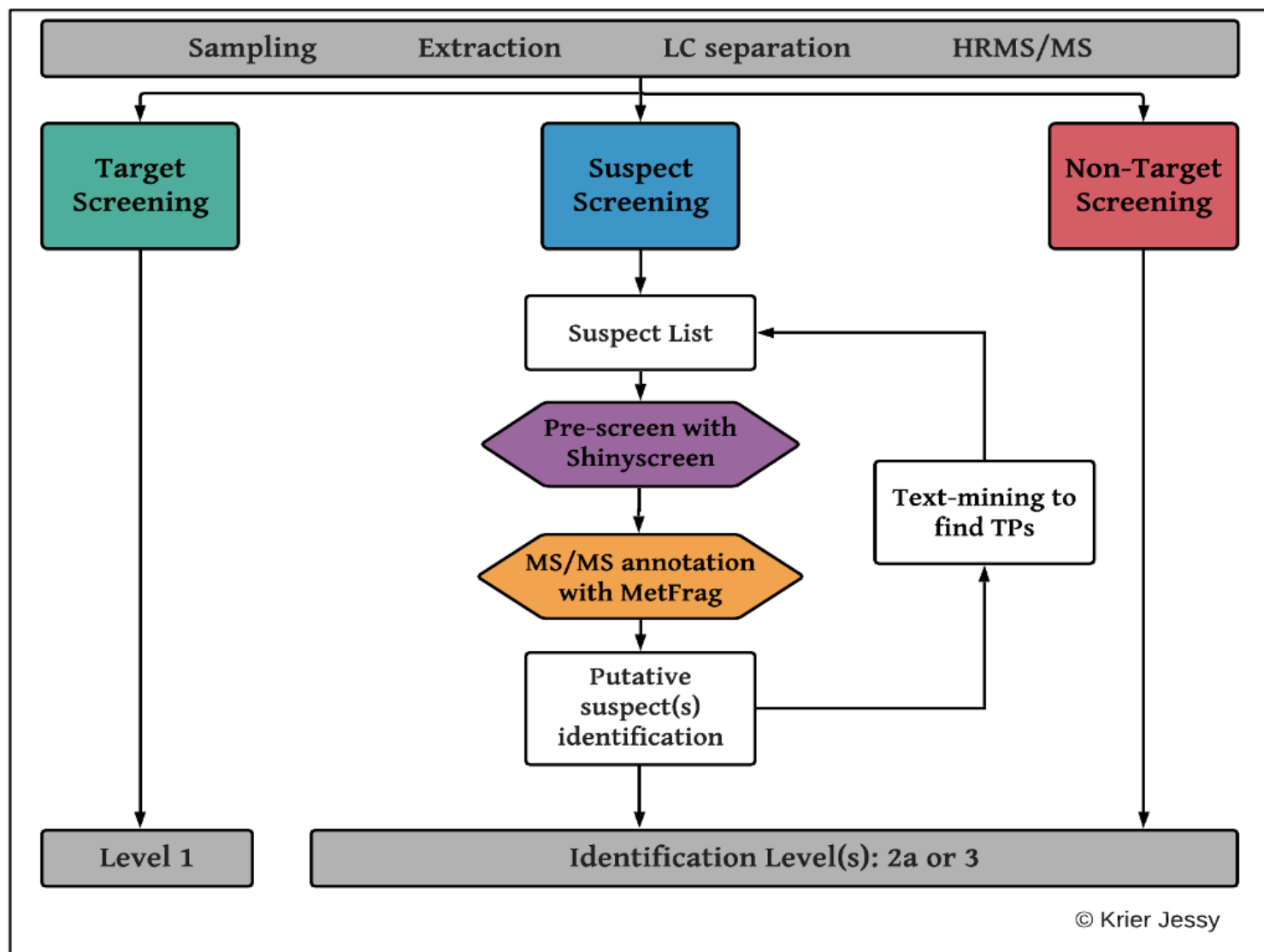


Figure 1

The newly created high-throughput suspect screening workflow, including experimental (top, grey) and computational steps. Both suspect and target screening were performed.

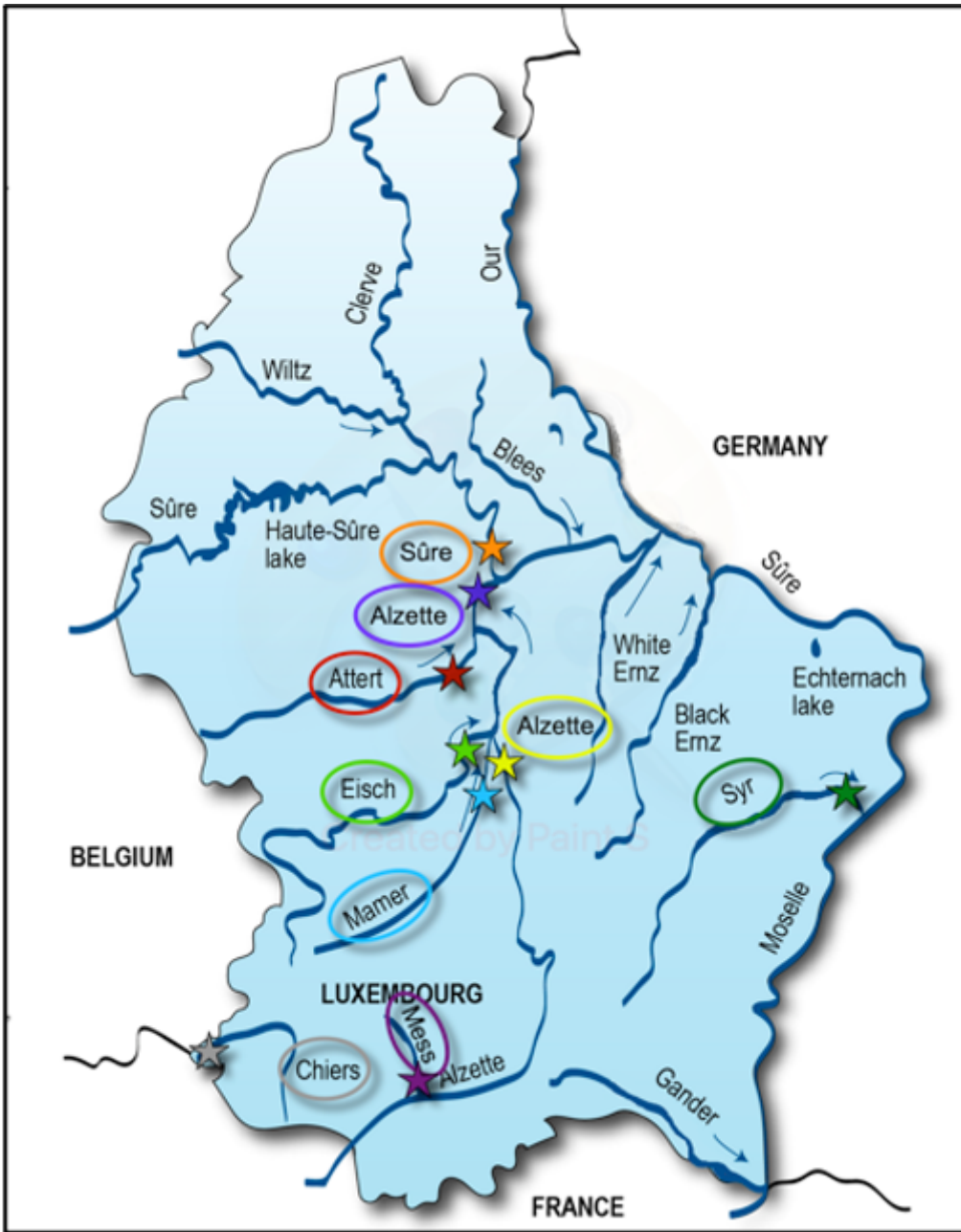


Figure 2

The Luxembourgish map with the eight selected rivers and nine sampling locations (Alzette has 2 sampling locations) marked. The sampling locations were selected by the "L'administration de la gestion de l'eau" (AGE) from Luxembourg as part of their 2019 surface water monitoring efforts. Source: LCSB. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

PubChem Terbutylazine-2-hydroxy (Compound)

8.1 Transformations

Page 2 of 9 items View More Rows & Details Download

SORT BY Please Choose One







| Predecessor Image | Predecessor Name | Transformation | Successor Image | Successor Name | Enzyme |
|---|-------------------------|----------------|---|----------------------------------|--------|
|  | Terbutryn | Oxidation |  | Terbutylazine-2-hydroxy | |
|  | Terbutylazine | Dehalogenation |  | Terbutylazine-2-hydroxy | |
|  | Terbutylazine-2-hydroxy | Deethylation |  | Terbutylazine-desethyl-2-hydroxy | |

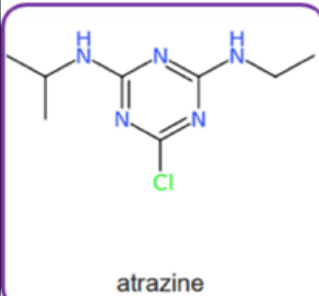
Figure 3

The “Transformations” section for Terbutylazine-2-hydroxy, CID: 135495928
<https://pubchem.ncbi.nlm.nih.gov/compound/135495928#section=Transformations>.

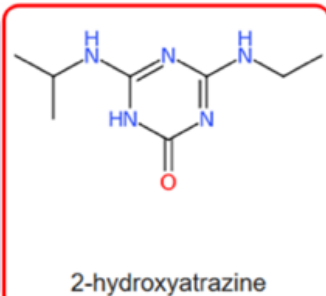
Atrazine was fed to white leghorn hens for 7 days. No residues of atrazine or its metabolites were found in eggs. In collected excreta atrazine, desethylatrazine, 2-hydroxyatrazine, and 2-hydroxy-desethylatrazine were present. Analysis of chicken tissues indicated the presence of hydroxyatrazine, desethylhydroxyatrazine & desethylatrazine. Highest levels of hydroxy- and desethyl-hydroxyatrazine were in the liver.

Menzie, C.M. Metabolism of Pesticides-Update III. Special Scientific Report- Wildlife No. 232. Washington, DC: U.S.Department of the Interior, Fish and Wildlife Service, 1980., p. 527

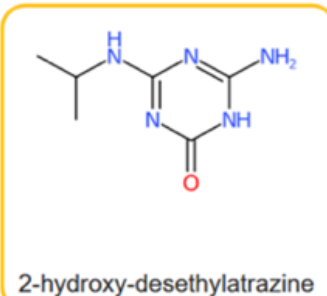
► Hazardous Substances Data Bank (HSDB)




atrazine



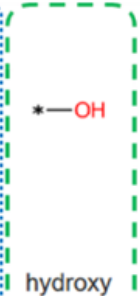
2-hydroxyatrazine



2-hydroxy-desethylatrazine



desethylatrazine



hydroxy

Figure 4

Automatic text mining (top) and manual curation (bottom) of HSDB content using one example from atrazine (<https://pubchem.ncbi.nlm.nih.gov/compound/Atrazine#section=Metabolism-Metabolites>).

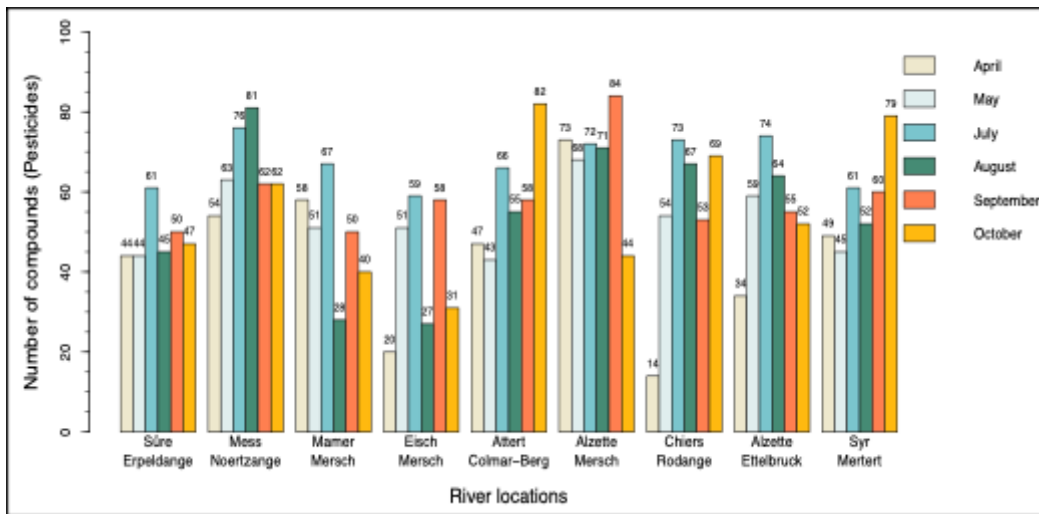


Figure 5

The results of pre-screening with ShinyScreen, showing how many pesticides passed the quality check for each sampling location and per month (positive and negative modes are visualized together).

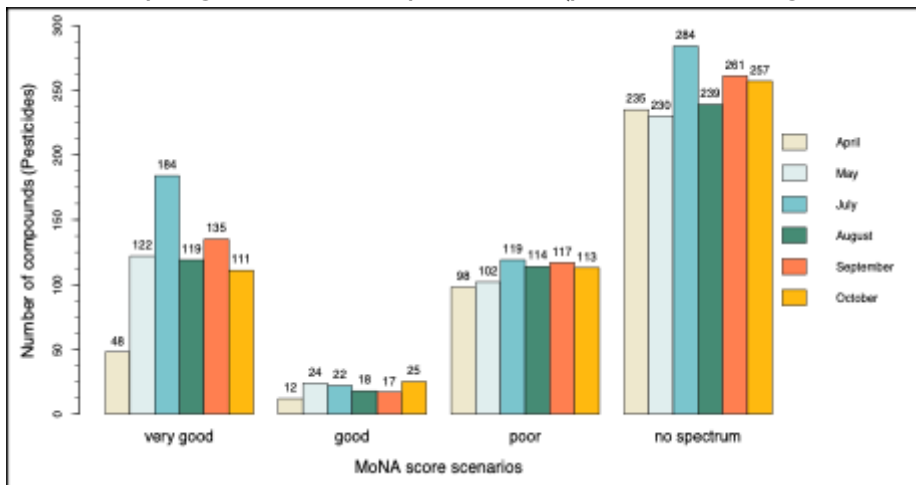


Figure 6

The results of MetFrag spectra annotation. The graph represents the 3,006 cases (162 pesticides) regrouped according to the four MoNA score scenarios for the six months (positive and negative mode together).

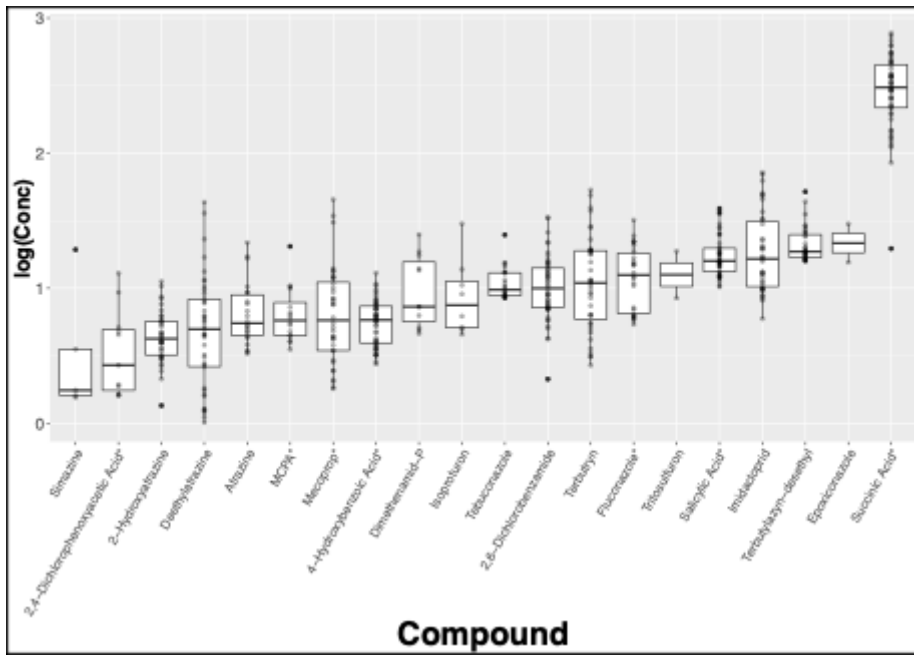


Figure 7

Boxplots showing the range of log (10) concentrations (original concentration units: parts per trillion = ppt) for the different pesticide and transformation products across all months and sampling locations in 2019. Compounds on the x-axis are sorted in ascending order of median log (10) concentration. Concentration values that were below the respective quantification range were excluded. All compounds were measured in positive mode except for those marked with an asterisk, which were measured in negative mode.

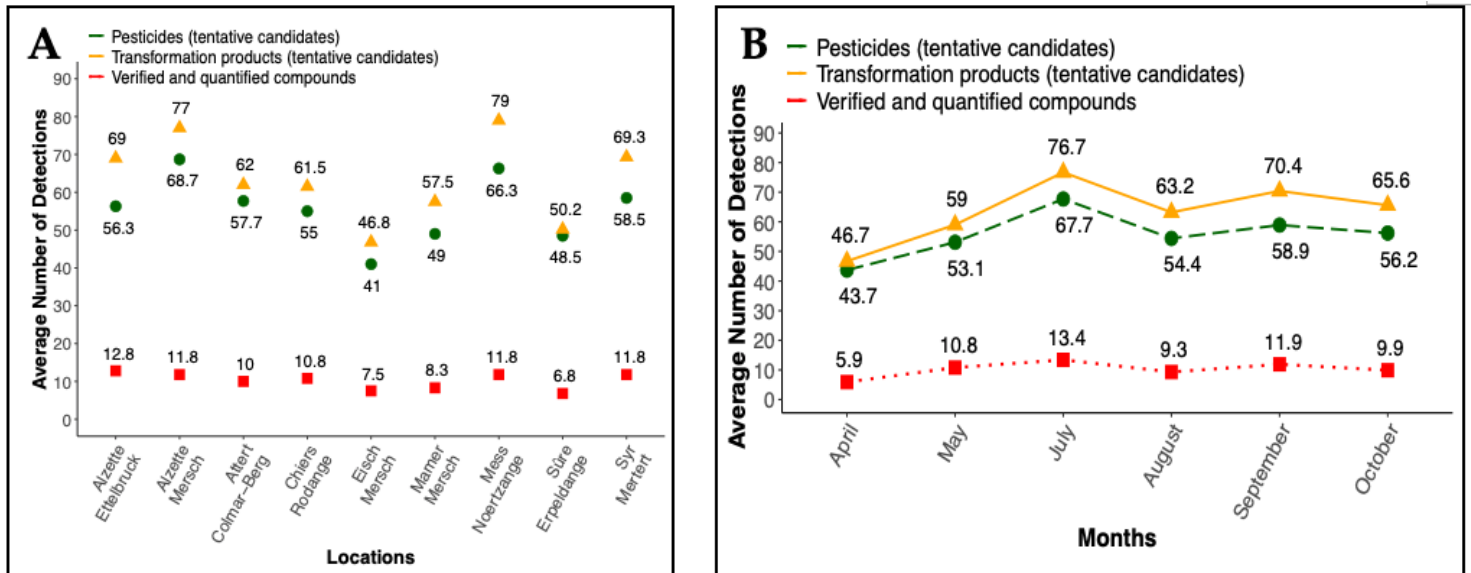


Figure 8

The spatial (A) and temporal (B) distribution of the tentatively detected pesticides and transformation products as well as for the verified and quantified compounds. No samples were available for June.

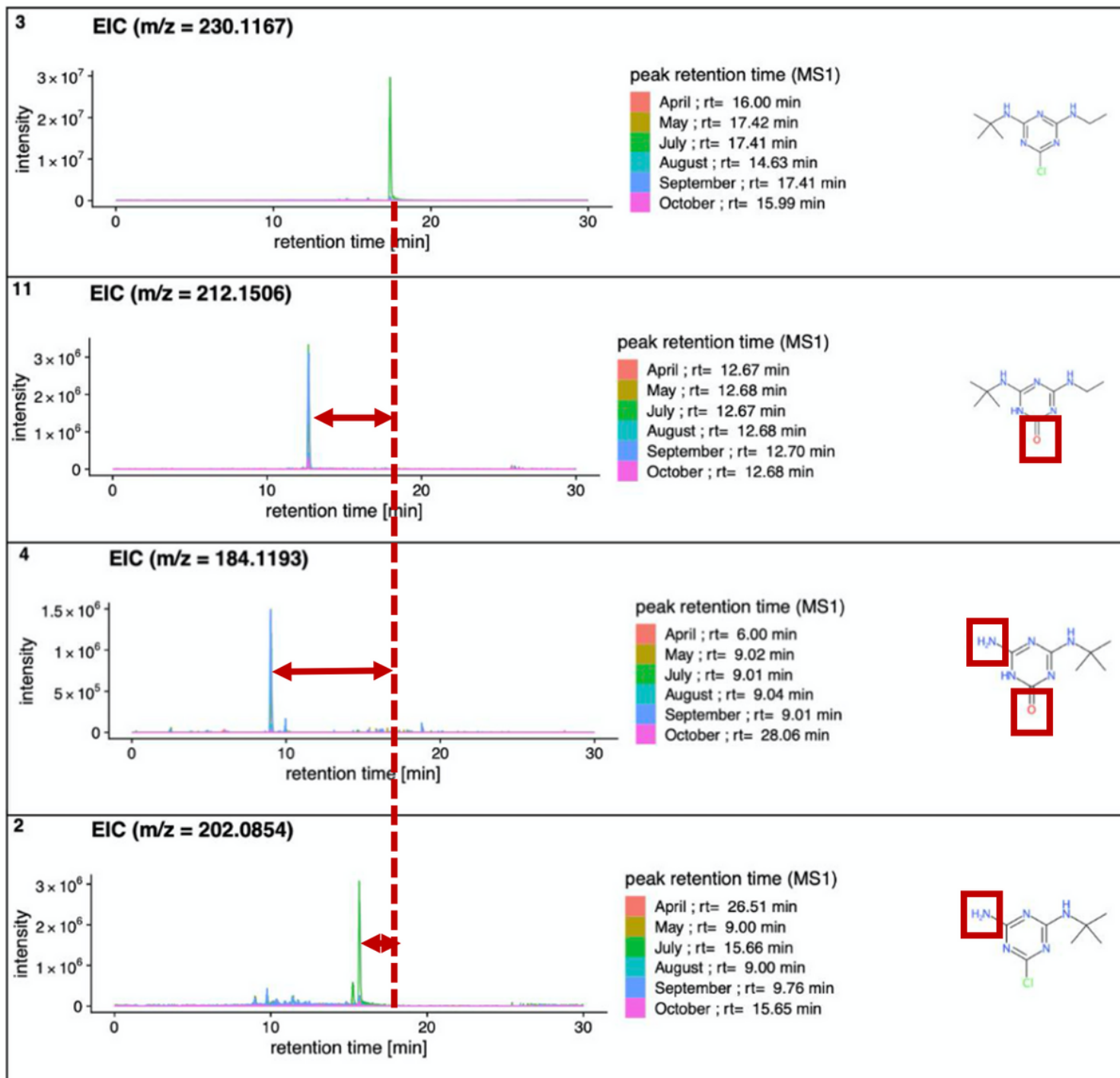


Figure 9

The spectra of the parent compound terbutylazine (top, suspect 3) with its three TPs 2-hydroxyterbutylazine (next, suspect 11), desethyl-2-hydroxyterbutylazine (next, suspect 4) and desethylterbutylazine (bottom, suspect 2) in positive mode (screenshots from ShinyScreen). The structures are shown to the right.

Supplementary Files

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

- [LUXPESTSupplementaryFigures29042021.docx](#)
- [LUXPESTSupplementaryTables05062021.xlsx](#)