Environment International 158 (2022) 106885

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Discovering pesticides and their TPs in Luxembourg waters using open cheminformatics approaches

Jessy Krier [a], Randolph R. Singh [a,1], Todor Kondić [a], Adelene Lai [a,b], Philippe Diderich [c], Jian Zhang [d], Paul A. Thiessen [d], Evan E. Bolton [d], Emma L. Schymanski [a,*]

[a] Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 Avenue du Swing, Luxembourg
[b] Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller University, Lessing Strasse 8, 07743 Jena, Germany
[c] Water Management Agency, Ministry of the Environment, Climate and Sustainable Development, 1 Avenue du Rock'n'roll, Luxembourg
[d] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

ARTICLE INFO

Handling Editor: Da Chen

Keywords:
Pesticides
Transformation products
Suspect screening
High resolution tandem mass spectrometry
Data mining
Non-target screening

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://gitlab.lcsb.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 96 potential TP masses in the samples. Further identification of these mass matches was performed using the open source approach 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. (Escher et al., 2020) 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. (Calzada et al., 2021; Mahmood et al., 2016; Hernández et al., 2020) For

* Corresponding author.
E-mail addresses: jessy.krier@uni.lu (J. Krier), randolph.singh@ifremer.fr (R.R. Singh), todor.kondic@uni.lu (T. Kondić), adelene.lai@uni.lu (A. Lai), Philippe.Diderich@eau.etat.lu (P. Diderich), jiazhang@ncbi.nlm.nih.gov (J. Zhang), thiessen@ncbi.nlm.nih.gov (P.A. Thiessen), bolton@ncbi.nlm.nih.gov (E.E. Bolton), emma.schymanski@uni.lu (E.L. Schymanski).
[1] Current affiliation: IFREMER (Institut Français de Recherche pour l'Exploitation de la Mer), Laboratoire Biogéochimie des Contaminants Organiques, Rue de l'Ile d'Yeu, BP 21105, Nantes Cedex 3, 44311, France.

https://doi.org/10.1016/j.envint.2021.106885
Received 30 April 2021; Received in revised form 30 July 2021; Accepted 15 September 2021
Available online 21 September 2021

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). (Somasundaram et al., 1991; Sinclair and Boxall, 2003) 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. (Sinclair and Boxall, 2003; Olsson et al., 2013) Although parent compounds are assessed in detail in many regulatory schemes, the requirements for the assessment of TPs are less well developed. (Sinclair and Boxall, 2003) While their occurrence and significance are now reasonably well-known in research circles, it is still difficult to access information on TPs in a central and "FAIR" (Findable, Accessible, Interoperable and Reusable) (Sansone et al., 2019) 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, the presence of both pesticides and their documented TPs (in openly available information sources) in samples is investigated.

Previous work by Moschet *et al.* (Moschet et al., 2013) and Kiefer *et al.* (Kiefer et al., 2019) both characterised the relevance of pesticide transformation products in their findings and shared their lists afterwards (SWISSPEST (Moschet, 2017) and SWISSPEST19 (Kiefer et al., 2020), respectively) on the NORMAN Suspect List Exchange (NORMAN -SLE) (NORMAN Suspect List Exchange, 2021), thus making them more "FAIR" (Sansone et al., 2019). 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. (Hollender et al., 2017; Krauss et al., 2010) 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 (Hollender et al., 2017; Vinaixa e al., 2016), or to discover and identify new, previously unknown chemicals in the environment. (Hollender et al., 2017; Helmus et al., 2021) These two non-targeted analysis strategies are called suspect screening and non-targeted screening, respectively. (Moschet et al., 2013) 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. (Moschet et al., 2013; Moschet et al., 2014) 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. (Moschet et al., 2013)

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. (Moschet et al., 2013) Thus, suspect screening methods are therefore developed to reveal a fuller picture of occurring chemicals and can be performed with suspect chemical lists, (Moschet et al., 2013; Hollender et al., 2017; Krauss et al., 2010) allowing for eventual prioritization for target analysis and confirmation efforts. (Moschet et al., 2013)

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. (Schymanski et al., 2014) 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. (Schymanski et al., 2014) 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 prescreening; 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://gitlab.lcsb.uni.lu/eci/shinyscreen/) (Kondic et al., 2020), 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.* (Lai et al., 2021) Potential suspects with MS1 and MS2 spectra passing the Shinyscreen pre-screening were considered for additional identification efforts via MS2 spectra annotation using the open source *in silico* fragmentation approach MetFrag (https://msbi.ipb-halle.de/MetFrag/). (Ruttkies et al., 2016) MetFrag combines compound database searching and fragmentation prediction plus other experimental and metadata terms for molecule identification using HR-MS2 fragmentation information. (Ruttkies et al., 2016) 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/) (Kim et al., 2019) and CompTox (https://comptox.epa.gov/dashboard/) (Williams et al., 2017), 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). (Ruttkies et al., 2016) 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.* (Lai et al., 2021), 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. (Williams et al., 2017) The recently-released PubChemLite for Exposomics collection (Schymanski et al., 2021), 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 Fig. 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 Suppl. Data Excel File **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 (±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 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 (these were all solvent options possible with this set-up, chosen to maximise the compounds eluted within the possibilities available). 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 µm 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 C_{18} column (dimensions: 1.7 µm, 2.1 × 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) (ASTA, 2021) and the SWISSPEST list of registered insecticides and fungicides in Switzerland (Moschet et al., 2013) 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 work (Schymanski

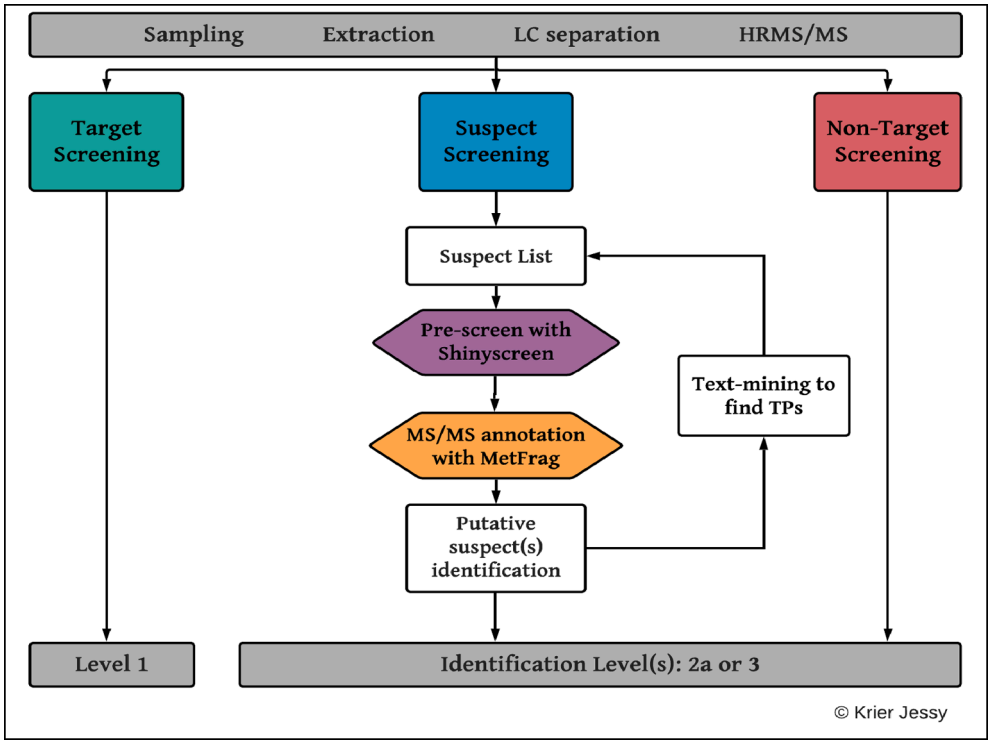

Fig. 1. The newly created high-throughput suspect screening workflow, including experimental (top, grey) and computational steps. Both suspect and target screening were performed.

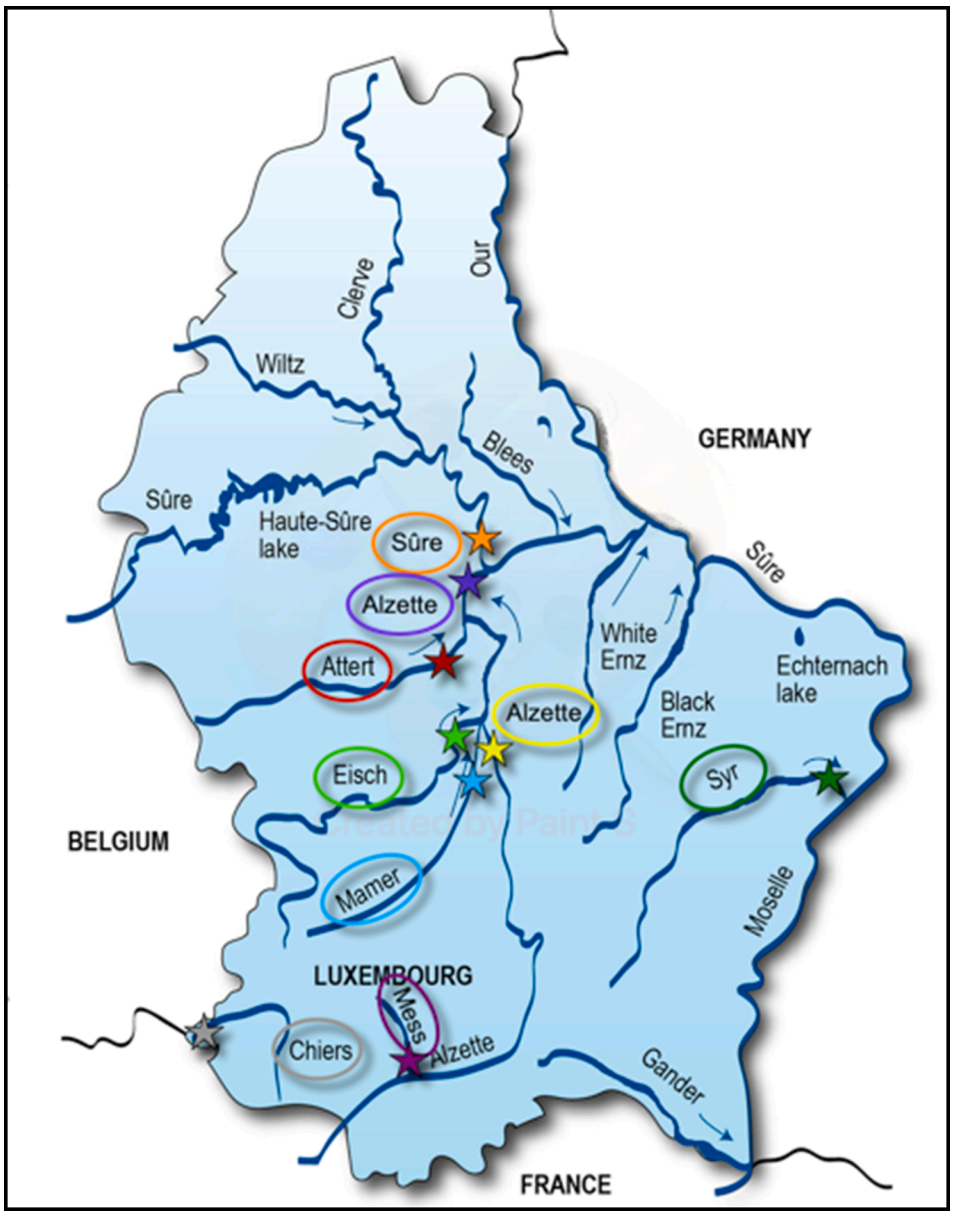

Fig. 2. The Luxembourgish map with the eight selected rivers and nine sampling locations (Alzette has 2 sampling locations) marked. The sampling locations were selected by the "L'administration de la gestion de l'eau" (AGE) from Luxembourg as part of their 2019 surface water monitoring efforts. Source: LCSB.

et al., 2019) were also included, as documented in the "LUXPEST" dataset available on Zenodo (Krier, 2020) and briefly below.

The final LUXPEST pesticide suspect list included 386 pesticides, (Krier, 2020) classified into different classes along with information about their use authorisation in Luxembourg. (European Commission, 2021; University of Hertfordshire, 2021a, 2021b) Of the 386 pesticides, 196 are permitted for use in Luxembourg, 128 are not, while for 14 pesticides and 48 TPs permission information was either not available or not applicable, respectively. The classification efforts revealed that most of them were fungicides and herbicides (96 and 93 respectively); 49 were already classified as pesticide TPs (Suppl. Data **Figure S1**). As a part of "FAIRifying" this dataset, the LUXPEST list is openly available on the NORMAN-SLE (NORMAN Suspect List Exchange, 2021), PubChem (Kim et al., 2019) and CompTox (Williams et al., 2017; Krier, 2021) 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 Suppl. Data **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 Shinyscreen (Kondic et al., 2020) with the following settings for extraction and automatic quality control based on the typical Q Exactive HF performance: 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 > $1x10^5$, 2) presence of MS2 spectrum, 3) alignment of MS1 and MS2 peaks within the RT tolerance, 4) signal to noise ratio > 3. The automatic quality control procedure is explained in greater detail in Lai *et al.* (Lai et al., 2021), including plots demonstrating "pass" and "fail" scenarios.

2.2.2.2. Candidate identification with MetFrag. The features that passed

the quality control were then analysed using MetFrag (Ruttkies et al., 2016) coupled to CompTox (Williams et al., 2017; Schymanski, 2019) to achieve tentative identifications (Ruttkies et al., 2016), generally consistent with Lai *et al.* (Lai et al., 2021) 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 was 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. (MassBank of North America, 2021) 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. The "Exact Spectral Similarity (MoNA)" term (hereafter "MoNA Score") was used, in which all MoNA spectra (if available) are retrieved using the InChIKey of the given candidate and compared with the experimental spectrum. The best spectral similarity score is reported as the result. Several additional metadata terms were used in the MetFrag calculation, 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). The additional scoring terms were TOXCAST_PERCENT_ACTIVE_BIOASSAYS, PREDICTED_EXPOSURE, PUBMED_ARTICLES, PUBCHEM_SOURCES, DATA_SOURCES, CPDAT_COUNT, KEMIMARKET_EXPO and KEMIMARKET_HAZ. For the sake of readability, further information about these scores can be found elsewhere (Lai et al., 2021), since the MoNA Score became the primary decision-making criterion in this work as described further below.

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.* (Schymanski et al., 2014) 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 elsewhere (Schymanski et al., 2021) (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

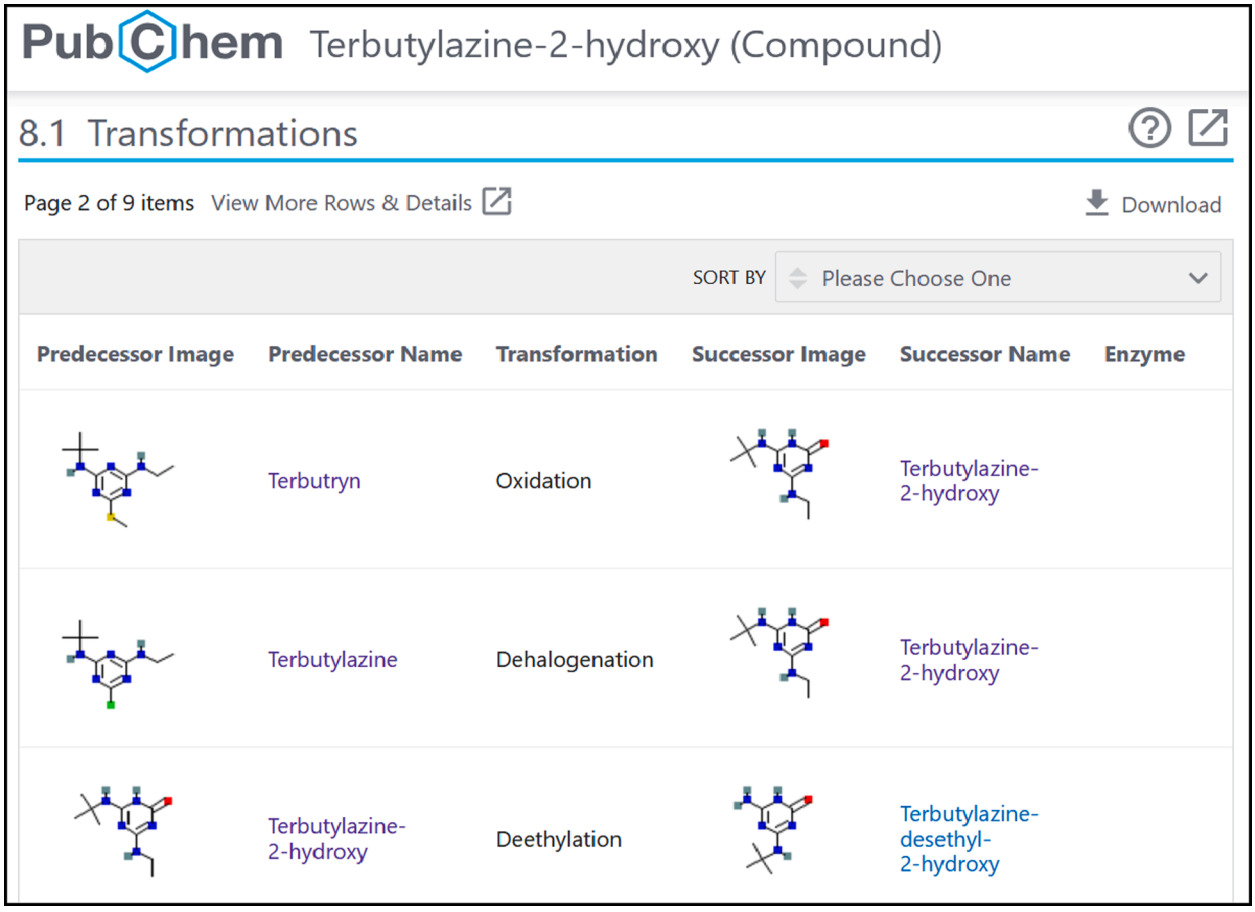

Fig. 3. The "Transformations" section for Terbutylazine-2-hydroxy, CID: 135495928. Source URL: https://pubchem.ncbi.nlm.nih.gov/compound/135495928#section=Transformations.

PubChem. At the time this study was performed, the NORMAN-SLE lists included were S60 SWISSPEST19 (Kiefer et al., 2020) and S66 EAWAGTPS (Schollee and Schymanski, 2020). The deposition of "Transformation" information in PubChem is automated through the NORMAN-SLE via Zenodo depositions (NORMAN Suspect List Exchange (NORMAN SLE) Zenodo Community, 2021) and mapping files in GitLab (Environmental Cheminformatics, 2021a). 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 (Fig. 3 top right) or via scripting queries. Custom-made R functions were designed to access this as a part of this work. (Environmental Cheminformatics, 2021b)

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 Fig. 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) (Willighagen, 2017; Mayfield, 2021). 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 HSDBTPS (LCSB-ECI et al, 2020) and full provenance of the curation is available on the Environmental Cheminformatics GitLab repository (Environmental Cheminformatics, 2021c).

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. (Ulrich et al., 2019) Retention times were considered a match if the difference was less than ± 0.2 min. Additional reference standards were purchased where possible (Suppl. Data Excel File **Table S2**). Where reference standards were available, the concentration of the pesticides and TPs were quantified using an external calibration curve ranging from 1 µg/L to 1000 µg/L 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 1 ng/L (after accounting for dilution) were reported to be below the quantifiable range. Since only an external calibration was performed, it is not possible to fully correct for several factors that may influence the concentrations such as matrix effects, and the concentrations reported here should be interpreted accordingly. Full quantification of many of these analytes is done in a routine targeted manner at AGE, and the results reported below are generally comparable. Since the extended screening performed here will also inform future targeted monitoring efforts at AGE (as described further below), extensive quantification was not the main focus of this work.

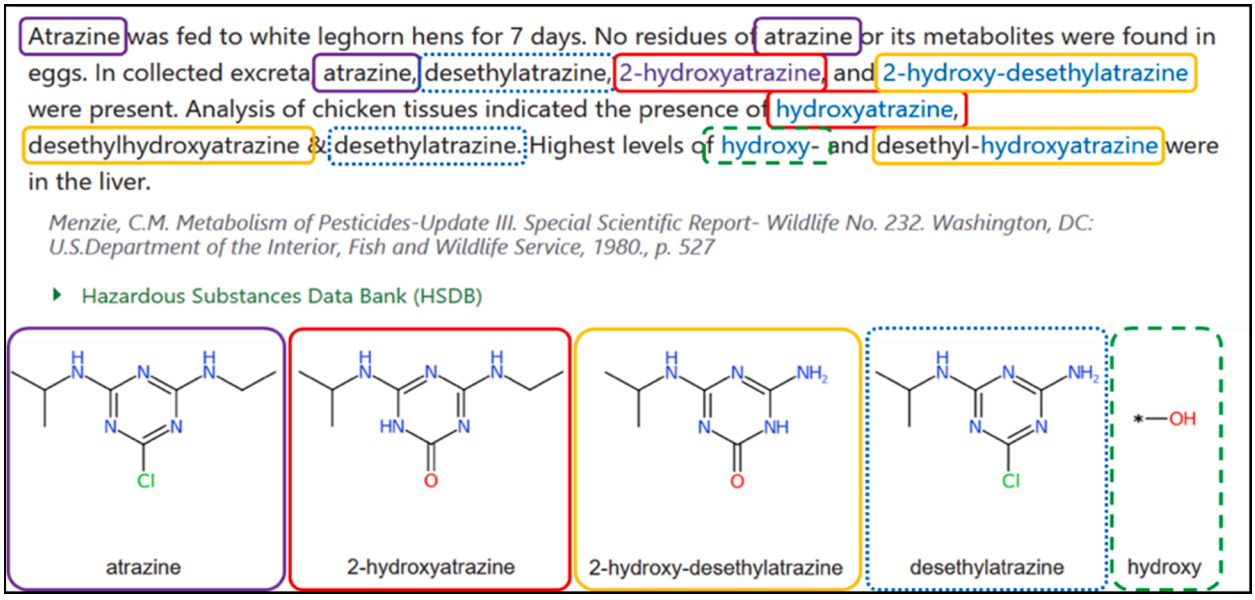

Fig. 4. Automatic text mining (top) and manual curation (bottom) of HSDB content using one example from atrazine. Source URL: https://pubchem.ncbi.nlm.nih.gov/compound/Atrazine#section=Metabolism-Metabolites.

3. Results

The numbers that will be explained in detail in the next sections are summarized in a table (Suppl. Data Excel File **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 and MetFrag annotation

Shinyscreen was run with the 386 LUXPEST (Krier, 2020) suspects (Suppl. Data Excel File **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. Each case corresponds to a unique feature per compound, location, month and mode. In total, there were 3006 cases deemed suitable for further identification with MetFrag, which corresponded to 162 unique compounds (Suppl. Data Excel File **Table S5**). Fig. 5 illustrates the number of compounds 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. All 3006 cases were subsequently analyzed and annotated with MetFrag to assign an identification confidence level. These cases were then categorized into four different scenarios depending on their MoNA score, as shown in Fig. 6. The following section describes the next steps at a "per compound" level (rather than "per case").

3.2. 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 sore $>$ 0.9) were selected (Suppl. Data Excel File **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 Suppl. Data Excel File **Table S7**.

After manual curation, the merged data file of TPs extracted from HSDB was added to Zenodo as HSDBTPS (LCSB-ECI et al, 2020) 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. (US EPA, 2021)

3.3. Suspect screening for the pesticide TPs

Shinyscreen was run again for all samples with 181 pre-selected compounds (Suppl. Data Excel File **Table S7**), resulting in a total of 19,548 cases. Of these, there were 1275 cases in negative mode and 2159 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 3434 cases corresponded to 96 transformation products (Suppl. Data Excel File **Table S8**) and the 36 parent compounds (132 different compounds in total). The number of cases for each location and month is available in the Suppl. Data **Figure S4**.

The MS2 spectra of 132 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 (Suppl. Data **Figure S5**). Out of the 3434 cases, there were 1190 that were able to achieve a MoNA score above 0.9 corresponding to eight unique additional identifications at Level-2a (Suppl. Data Excel File **Table S8**). The *m/z* 137.0244 yielded three peaks and three Level-2a candidates, salicylic acid (which can be both parent and TP in a variety of reactions, including a role as a TP of aspirin - a common pharmaceutical and not a pesticide), 3-hydroxybenzoic acid

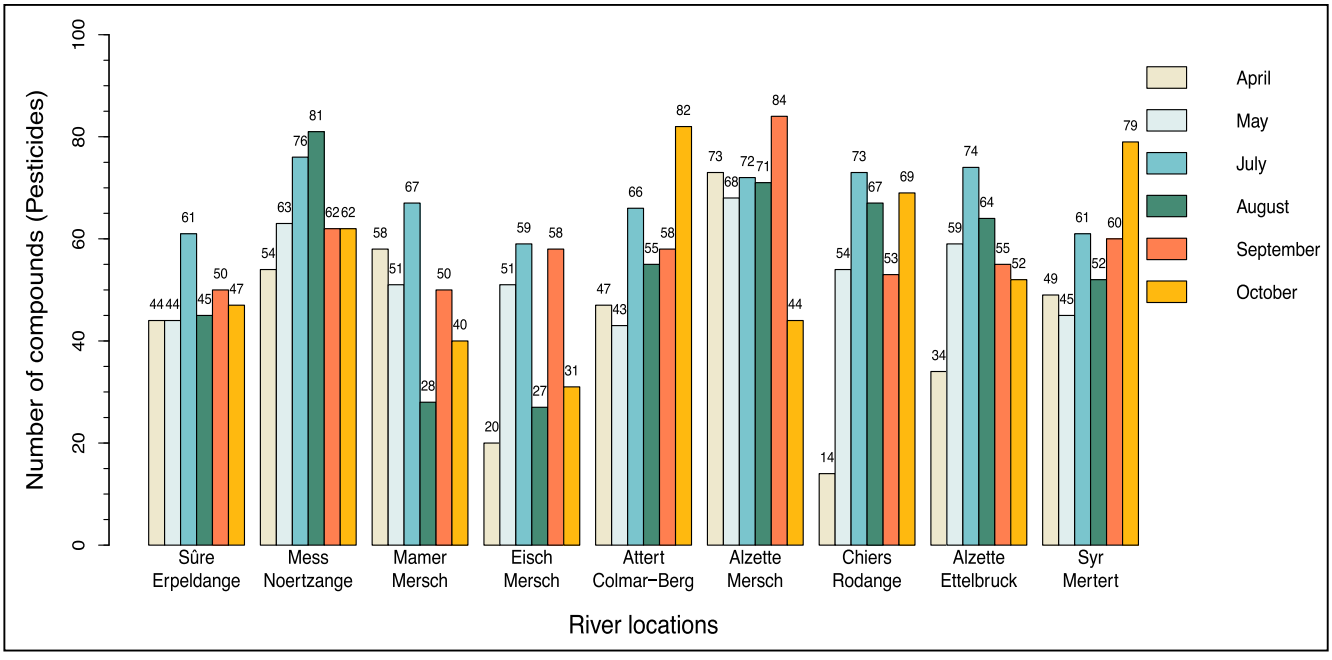

Fig. 5. The results of pre-screening with Shinyscreen, showing how many pesticides passed the quality check for each sampling location and per month (positive and negative modes are visualized together).

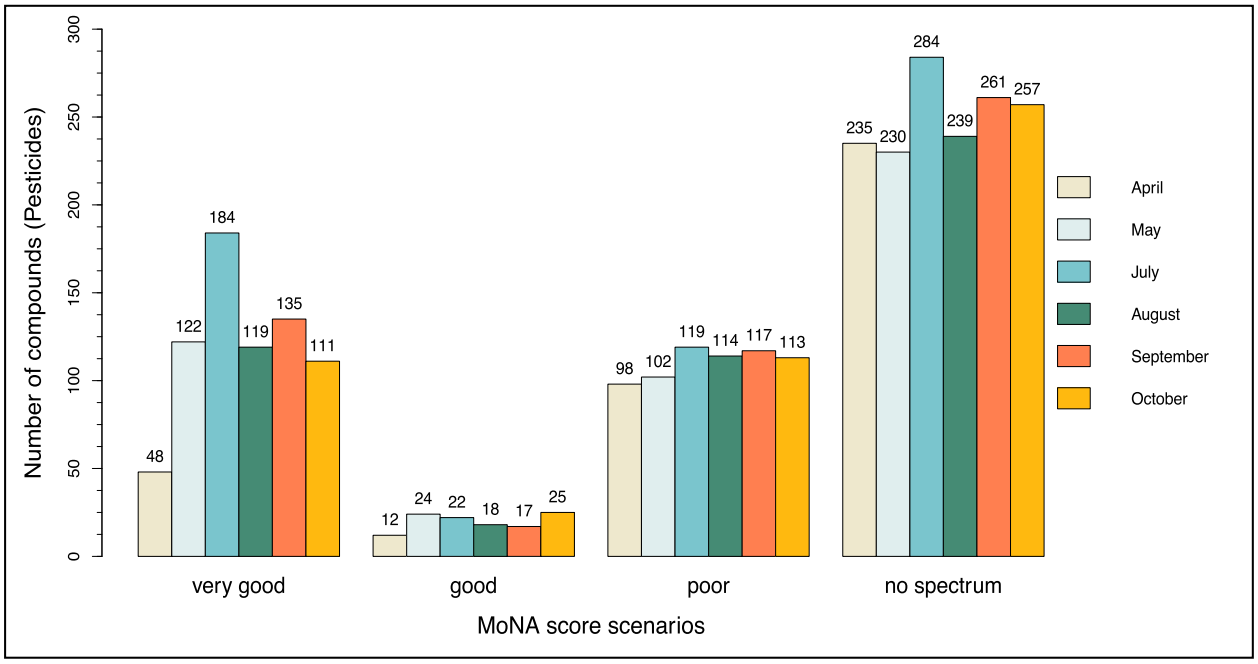

Fig. 6. The results of MetFrag spectra annotation. The graph represents the 3006 cases (162 pesticides) regrouped according to the four MoNA score scenarios for the six months (positive and negative mode together).

and 4-hydroxybenzoic acid (both TPs of benzoic acid). Further TPs included succinic acid (TP of 2,4-dichlorophenoxyacetic acid, linuron and sulcotrione), deethylterbutryne (TP of terbutryn and irgarol), terbutylazine-2-hydroxy (TP of terbuthylazine, terbutryn and terbumeton), terbutylazine-desethyl-2-hydroxy (a further TP of terbutylazine-desethyl and terbutylazine-2-hydroxy and thus also a TP related to the parents mentioned above). Finally o-phenylenediamine (TP of benomyl) had a high MoNA score for the *para*-isomer p-phenylenediamine (no TP information); neither the ortho- or meta- isomers had spectra in MoNA but could be expected to have quite similar spectra. Some of these are discussed further below.

3.4. Verification of the tentative candidates and their quantification

The 36 Level-2a pesticide identifications were selected for further confirmation efforts with reference standards (Suppl. Data Excel File **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 Suppl. Data Excel File **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 Suppl. Data Excel File **Table S10** and **Table S11**).

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

3.5. Spatial and temporal distribution

Fig. 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. Fig. 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 Trial (Schymanski et al., 2015), although it has not been permitted for use for many years. Interestingly, the use of PubChemLite with the optimized default scoring terms (Schymanski et al., 2021) resulted in terbutylazine appearing ahead of propazine in the metadata ranking; further addition of the "agrochemicals" category (Schymanski et al., 2021) 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.* (Lai et al., 2021) and thus not discussed further here.

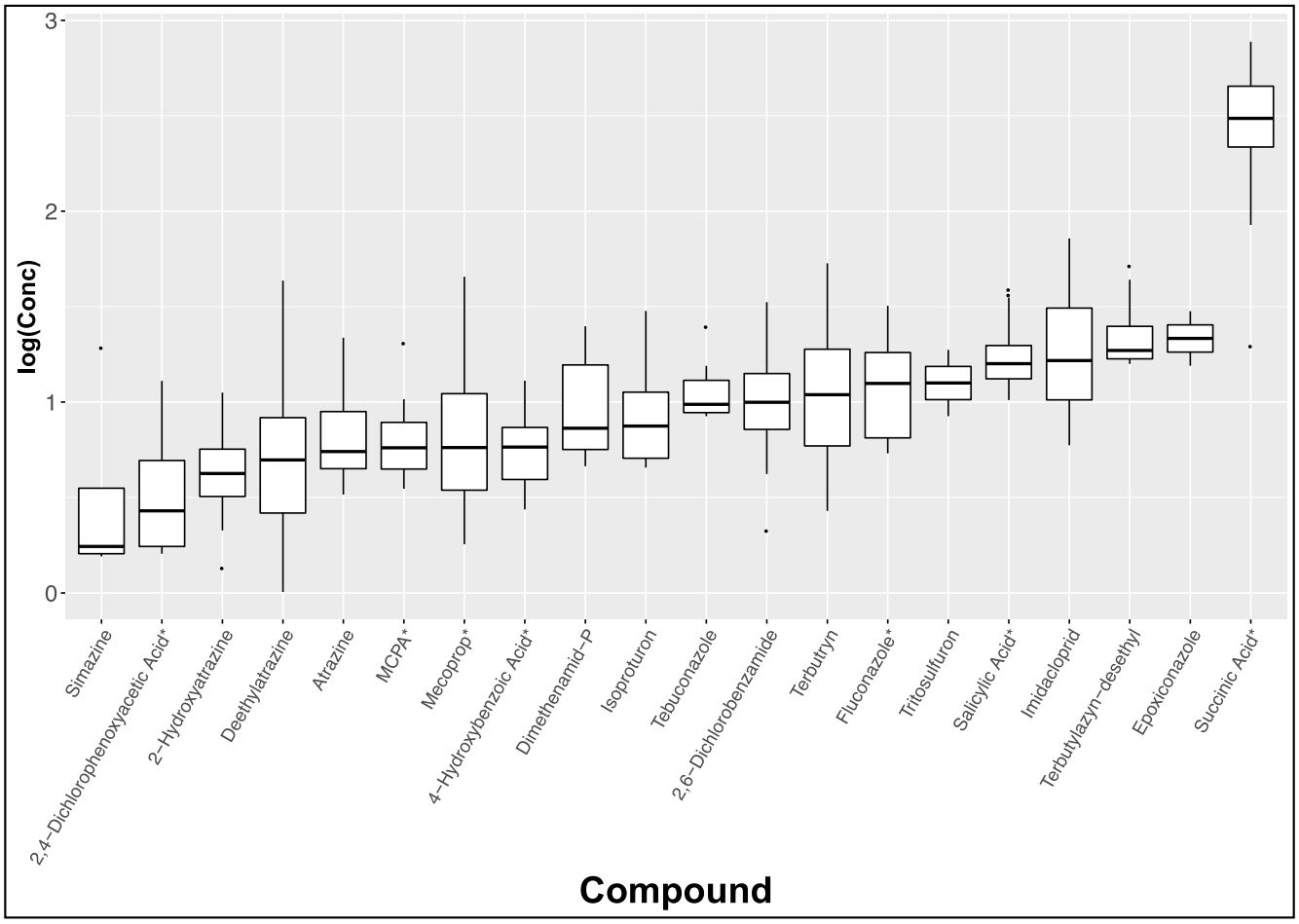

Fig. 7. Boxplots showing the range of log (10) concentrations (original concentration units: ng/L) for the different pesticide and transformation products across all months and sampling locations in 2019. Compounds on the x-axis are sorted in ascending order of median log (10) concentration. Concentration 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.

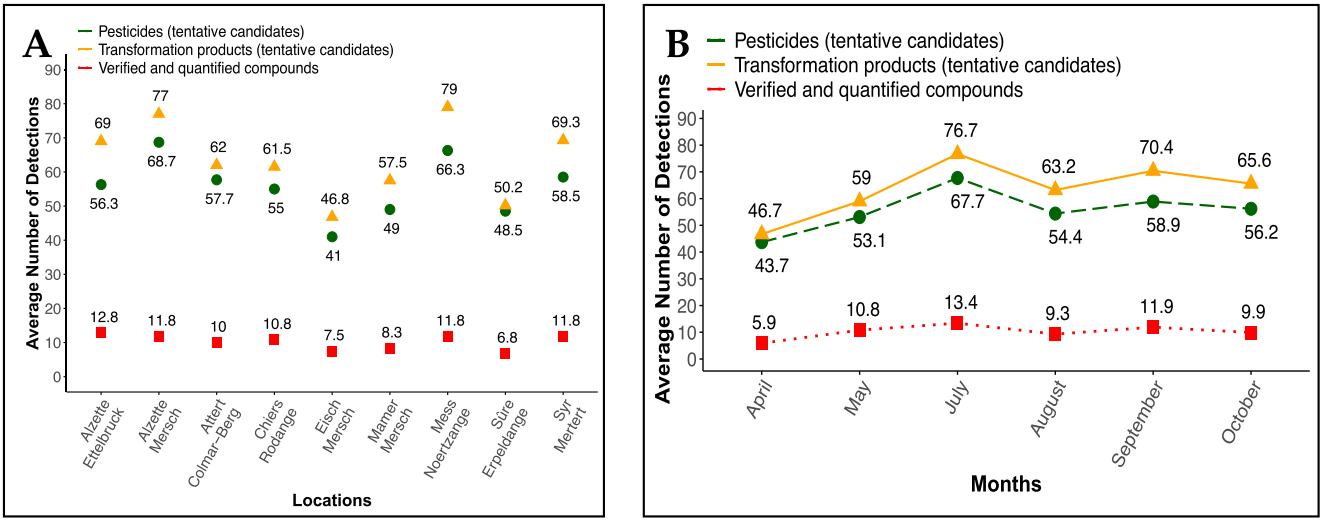

Fig. 8. The spatial (A) and temporal (B) distribution of the tentatively detected pesticides and transformation products as well as for the verified and quantified compounds. No samples were available for June.

One of the main TPs, desethylterbutylazine (ID N° 2, 4th chromatogram in the Fig. 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

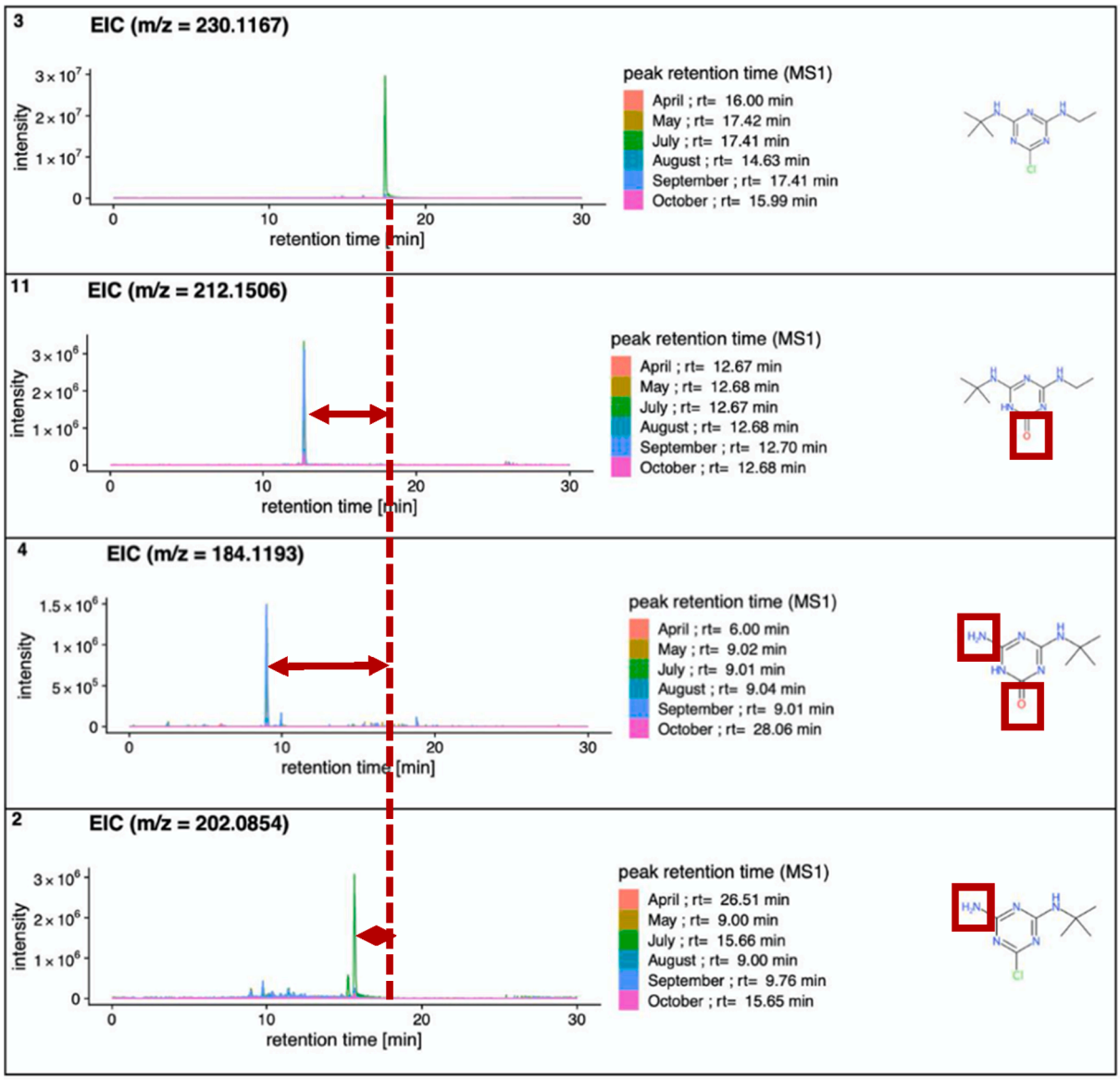

Fig. 9. The spectra of the parent compound terbutylazine (top, suspect 3) with its three TPs 2-hydroxyterbutylazine (next, suspect 11), desethyl-2-hydroxyterbutylazine (next, suspect 4) and desethylterbutylazine (bottom, suspect 2) in positive mode (screenshots from Shinyscreen). The structures are shown to the right.

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 Suppl. Data Excel File **Table S9**). The third TP, desethyl-2-hydroxyterbutylazine (ID N° 4) is detected at 9 min 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 chorine 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 occurrence of detected and quantified suspects that are not permitted for use in Luxembourg (see Suppl. Data **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 have no direct permission information. Looking at the permission information of their parent compounds revealed that for some TPs (e.g. 2-hydroxyatrazine) the parent compound is not permitted (atrazine), but for others (*e.g.* desethylterbutylazine) the parent compound is permitted (terbutylazine). As an example, the low levels of atrazine detected here (<100 ng/L) 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 Suppl.

Data Excel File **Table S10**), the pesticide TP succinic acid was found in highest concentrations (maximum concentrations found 774 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 additive (PubChem, 2021); 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 1 ng/L). Returning to the example from the section before (**Section 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 (Suppl. Data Excel File **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. (Grand Duchy of Luxembourg, 2021)

The temporal patterns (Fig. 8**B**) 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 3006 cases that passed the quality checks, leading to a final set of 162 different tentatively identified compounds. This means that 42% (162 tentatively detected compounds/386 suspects = 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 powerful (Schymanski et al., 2021).

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 3006 cases: 719 cases obtained a very good, 118 a good, and 663 a poor MoNA score. Additionally, in 1506 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 2287 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 (Suppl. Data Excel File **Table S5**).

For the TPs, 19,548 cases and graphs (181 pesticides times two modes times six months times nine locations) were analyzed. Out of these, there were 3434 cases that passed the quality check and kept for further analysis. This leads to a final number of 96 newly identified compounds (132 compounds in total − 36 known pesticides = new compounds 96). When excluding the 36 parent compounds, this led to eight TPs with a very good, two with a good, and eleven with a poor MoNA score. The remaining 75 pesticides (out of 96) 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 elsewhere (Schymanski et al., 2021) 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 128 are forbidden in Luxembourg (Suppl. Data Excel File **Table S4**) and could be classified into six main categories (Suppl. Data **Figure S1**). This information can be browsed in PubChem under LUXPEST at https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101 (Suppl. Data **Figure S2**) and this information is incorporated into the individual records in PubChem (Example in the Suppl. Data **Figure S3**). This information flow helps create the annotation categories that form the PubChemLite for Exposomics collection (see Schymanski *et al.* (Schymanski et al., 2021) 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 (Suppl. Data **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 GitLab (Environmental Cheminformatics, 2021)), 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 resulted in a suspect list of 181 compounds. Since this work was performed several other datasets have been added to the Transformations tables including MetXBioDB (Djoumbou-Feunang et al., 2020) from BioTransformer (Djoumbou-Feunang et al., 2019) and it is highly likely that the numbers of pesticide TPs retrieved for screening would be higher now.

The role of certain chemicals as a "parent" or "predecessor" versus a "TP" or "successor" is not always clear. Several entries in the original LUXPEST list are in fact themselves TPs, while several TPs can also be further transformed (for example desethylterbutylazine) such that they can become predecessors themselves. The data retrieval method used here returns any CIDs related by a Transformation to the searched CID, be they predecessor or successor, such that any "successor" in LUXPEST would result in a "predecessor" being screened in the second round. For the sake of readability of this article, this point is not belaboured in the above content. However, this information is included in Suppl. Data Excel File **Table S7** for those interested in investigating this further. Of the 36 Level-2a "parents", 22 were predecessors, 6 were successors and 8 could be both. Of the 145 retrieved "TPs", 13 were predecessors, 126 were successors and 6 could be both (according to the information sources used here).

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, REFTPS (Schymanski, 2020) (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.* (Tian et al., 2021) Overall, these pilot efforts have already caught the interest of several other workflows and are being integrated into the open source HR-MS workflow patRoon (Helmus et al., 2021), 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 prescreening 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.

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.

CRediT authorship contribution statement

Jessy Krier: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Randolph R. Singh:** Conceptualization, Methodology, Validation, Supervision, Writing – review & editing. **Todor Kondić:** Software, Resources, Writing – review & editing. **Adelene Lai:** Software, Data curation, Visualization, Writing – review & editing. **Philippe Diderich:** Conceptualization, Methodology, Validation, Resources, Project administration, Writing – review & editing. **Jian Zhang:** Methodology, Software, Data curation, Writing – review & editing. **Paul A. Thiessen:** Software, Validation, Data curation, Writing – review & editing. **Evan E. Bolton:** Conceptualization, Methodology, Validation, Resources, Supervision, Funding acquisition, Writing – review & editing. **Emma L. Schymanski:** Conceptualization, Methodology, Software, Validation, Formal analysis, Resources, Data curation, Supervision, Project administration, Funding acquisition, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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).

Data statement

The suspect lists LUXPEST and HSDBTPS developed in this work are online on Zenodo (DOI: 10.5281/zenodo.3862688 and DOI: 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 and accessible via ftp://massive.ucsd.edu/MSV000087190/ and https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?accession=MSV000087190. The major software used Shinyscreen (https://gitlab.lcsb.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://gitlab.lcsb.uni.lu/eci/pubchem/). In addition, two supplementary files are provided with this manuscript, as detailed above.

Appendix A. 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, suspect list and raw file availability, see Data Statement. Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2021.106885.

References

Administration of Technical Services (ASTA) — Ministry of Agriculture, Viticulture and Rural Development // The Luxembourg Government, 2021. https://ma.gouvernement.lu/en/administrations/asta.html (accessed 29/04/2021).

Calzada, J., Gisbert, M., Moscoso, B., 2021. The Hidden Cost of Bananas: Pesticide Effects on Newborns' Health. https://papers.ssrn.com/abstract=3786643. doi: 10.2139/ssrn.3786643.

Djoumbou-Feunang, Y., et al., 2019. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. J. Cheminf. 11, 2. https://doi.org/10.1186/s13321-018-0324-5.

Djoumbou-Feunang, Y., Schymanski, E., Zhang, J., Wishart, D., 2020. S. S73 | METXBIODB | Metabolite Reaction Database from BioTransformer. doi: 10.5281/zenodo.4094568.

Environmental Cheminformatics, 2021. pubchem (HSDB_done repository). https://gitlab.lcsb.uni.lu/eci/pubchem/-/tree/master/annotations/tps/HSDB/HSDB_done (accessed 18/09/2021).

Environmental Cheminformatics, GitLab Repository pubchem, 2021. https://gitlab.lcsb.uni.lu/eci/pubchem (accessed 29/04/2021).

Environmental Cheminformatics, GitLab Repository pubchem Annotations.R, 2021. https://gitlab.lcsb.uni.lu/eci/pubchem/-/blob/master/annotations/tps/extractAnnotations.R (accessed 29/04/2021).

Environmental Cheminformatics, GitLab Repository pubchem HSDB, 2021. https://gitlab.lcsb.uni.lu/eci/pubchem/-/tree/master/annotations/tps/HSDB (accessed 29/04/2021).

Escher, B.I., Stapleton, H.M., Schymanski, E.L., 2020. Tracking complex mixtures of chemicals in our changing environment. Science 367, 388–392.

European Commission, 2021. EU Pesticides database. https://ec.europa.eu/food/plant/pesticides/eu-pesticides-database/public/?event=activesubstance.selection&language=EN (accessed 29/04/2021).

FAIR Principles, 2021. GO FAIR https://www.go-fair.org/fair-principles/ (accessed 29/04/2021).

Grand Duchy of Luxembourg, 2021. 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 (accessed 29/04/2021).

Helmus, R., ter Laak, T.L., van Wezel, A.P., de Voogt, P., Schymanski, E.L., 2021. patRoon: open source software platform for environmental mass spectrometry based non-target screening. J. Cheminf. 13, 1.

Hernández, A.F., Bennekou, S.H., Hart, A., Mohimont, L., Wolterink, G., 2020. Mechanisms underlying disruptive effects of pesticides on the thyroid function. Current Opinion Toxicol. 19, 34–41.

Hollender, J., Schymanski, E.L., Singer, H.P., Ferguson, P.L., 2017. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? Environ. Sci. Technol. 51, 11505–11512.

Kiefer, K., Müller, A., Singer, H., Hollender, J., 2020. S60 | SWISSPEST19 | Swiss Pesticides and Metabolites from Kiefer et al 2019. doi: 10.5281/zenodo.3766352.

Kiefer, K., Müller, A., Singer, H., Hollender, J., 2019. New relevant pesticide transformation products in groundwater detected using target and suspect screening for agricultural and urban micropollutants with LC-HRMS. Water Res. 165, 114972.

Kim, S., et al., 2019. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 47, D1102–D1109.

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.

Krier, J., 2020. S69 | LUXPEST | Pesticide Screening List for Luxembourg. doi:10.5281/zenodo.3862688.

Krier, J., 2021. Pesticide Screening List for Luxembourg. CompTox https://comptox.epa.gov/dashboard/chemical_lists/LUXPEST (accessed 29/04/2021).

Lai, A., et al., 2021. Retrospective non-target analysis to support regulatory water monitoring: from masses of interest to recommendations via in silico workflows. Environ. Sci. Eur. 33, 43.

LCSB-ECI et al., 2020. S68 | HSDBTPS | Transformation Products Extracted from HSDB Content in PubChem. doi: 10.5281/zenodo.3830987.

Mahmood, I., Imadi, S.R., Shazadi, K., Gul, A., Hakeem, K.R., 2016. Effects of Pesticides on Environment. In: Hakeem, K.R., Akhtar, M.S., Abdullah, S.N.A. (Eds.), Plant, Soil and Microbes: Volume 1: Implications in Crop Science. Springer International Publishing, pp. 253–269. doi: 10.1007/978-3-319-27455-3_13.

MassBank of North America, 2021. https://mona.fiehnlab.ucdavis.edu/ (accessed 29/04/2021).

Mayfield, J., 2021. The Chemistry Development Kit (CDK) Depict. https://github.com/cdk/depict (accessed 29/04/2021).

Moschet, C., et al., 2014. How a Complete Pesticide Screening Changes the Assessment of Surface Water Quality. Environ. Sci. Technol. 48, 5423–5432.

Moschet, C., Piazzoli, A., Singer, H., Hollender, J., 2013. 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.

Moschet, C., 2017. S11 | SWISSPEST | Swiss Insecticides, Fungicides and TPs. doi: 10.5281/zenodo.2623741.

NORMAN Suspect List Exchange (NORMAN SLE), 2021. NORMAN. https://www.norman-network.com/nds/SLE/ (accessed 29/04/2021).

NORMAN Suspect List Exchange (NORMAN SLE) Zenodo Community, 2021. NORMAN Network. https://zenodo.org/communities/norman-sle. Accessed 18/09/2021.

Olsson, O., et al., 2013. Fate of Pesticides and Their Transformation Products: First Flush Effects in a Semi-Arid Catchment. Clean: Soil, Air, Water 41, 134–142.

PubChem, 2021. Succinic acid. https://pubchem.ncbi.nlm.nih.gov/compound/1110 (accessed 29/04/2021).

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

Sansone, S.-A., et al., 2019. FAIRsharing as a community approach to standards, repositories and policies. Nat. Biotechnol. 37, 358–367.

Schollee, J., Schymanski, E., 2020. S66 | EAWAGTPS | Parent-Transformation Product Pairs from Eawag. doi: 10.5281/zenodo.3754449.

Schymanski, E.L., et al., 2014. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. Environ. Sci. Technol. 48, 2097–2098.

Schymanski, E.L., et al., 2015. Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. Anal. Bioanal. Chem. 407 (21), 6237–6255. https://doi.org/10.1007/s00216-015-8681-7.

Schymanski, E., 2019. MetFrag Local CSV: CompTox (7 March 2019 release). Wastewater MetaData File. https://doi.org/10.5281/zenodo.3472781.

Schymanski, E.L., et al., 2019. Connecting environmental exposure and neurodegeneration using cheminformatics and high resolution mass spectrometry: potential and challenges. Environ. Sci.: Processes Impacts 21, 1426–1445.

Schymanski, E.L., et al., 2021. Empowering large chemical knowledge bases for exposomics: PubChemLite meets MetFrag. J. Cheminf. 13, 19.

Kondic, T., et al., 2020 Shinyscreen. https://gitlab.lcsb.uni.lu/eci/shinyscreen/ (accessed 18/09/2021).

Schymanski, E., 2020. S74 | REFTPS | Transformation Products and Reactions from Literature. doi: 10.5281/zenodo.4318852.

Sinclair, C.J., Boxall, A.B.A., 2003. Assessing the Ecotoxicity of Pesticide Transformation Products. Environ. Sci. Technol. 37, 4617–4625.

Somasundaram, L., Coats, J.R., 1991. Pesticide Transformation Products in the Environment. In: Somasundaram, L., Coats, J.R. (Eds.), Pesticide Transformation Products, vol. 459. American Chemical Society, pp. 2–9.

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

Ulrich, E.M., et al., 2019. EPA's non-targeted analysis collaborative trial (ENTACT): genesis, design, and initial findings. Anal. Bioanal. Chem. 411, 853–866.

University of Hertfordshire, 2021. Pesticide Properties DataBase Search. https://sitem.herts.ac.uk/aeru/ppdb/en/search.htm (accessed 29/04/2021).

University of Hertfordshire, 2021. Bio-Pesticides DataBase Search. https://sitem.herts.ac.uk/aeru/bpdb/search.htm (accessed 29/04/2021).

US EPA, Chemistry Dashboard | HSDBTPS Chemicals List, 2021. https://comptox.epa.gov/dashboard/chemical_lists/HSDBTPS (accessed 29/04/2021).

Vinaixa, M., et al., 2016. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. TrAC Trends Anal. Chem. 78, 23–35.

Williams, A.J., et al., 2017. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. J. Cheminf. 9, 61.

Willighagen, E.L., et al., 2017. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J. Cheminform. 9, 33.