

RESEARCH

Open Access



# Non-target screening of surface water samples to identify exposome-related pollutants: a case study from Luxembourg

Dagny Aurich<sup>1\*</sup> , Philippe Diderich<sup>2</sup> , Rick Helmus<sup>3</sup>  and Emma L. Schymanski<sup>1\*</sup> 

## Abstract

**Background** Non-target screening of surface water samples collected over an extended period can reveal interesting temporal patterns in exposome-related pollutants. Additionally, geographical data on pollution sources close to the sampling sites, chemical classification data and the consideration of flow paths can provide valuable information on the origins and potential threat of tentatively identified chemical compounds. In this study, 271 surface water samples from 20 sampling sites across Luxembourg were analysed using high-resolution mass spectrometry, complementing routine target monitoring efforts in 2019–2022. Data analysis was performed using the open source R-package *patRoan*, which offers a customizable non-target workflow. By employing open source workflows featuring scoring terms, like spectral match and applying identification levels, tentative identifications can be prioritized, e.g. based on spectral similarity. Furthermore, by utilizing supplementary database information such as *PubChemLite* annotation categories and classification software such as *classyFire*, an overall assessment of the potential threats posed by the tentatively identified chemicals was conducted, enabling the prioritization of chemicals for future confirmation through targeted approaches.

**Results** The study tentatively identified 378 compounds associated with the exposome including benzenoids, organoheterocyclic compounds, and organic phosphoric acids and derivatives (11 *classyFire* superclasses, 50 subclasses). The classification analysis not only revealed temporal variations in agrochemicals, with the majority of identifications occurring in May to July, but also highlighted the prevalence of pharmaceuticals such as venlafaxine in surface waters. Furthermore, potential sources of pollutants, like metallurgic industry or household products were explored by considering common uses and geographical information, as commercial uses of almost 100% of the identified chemicals are known. 41 chemicals were suggested for potential inclusion to governmental monitoring lists for further investigation.

**Conclusions** The findings of this study complement existing knowledge on the pollution status of surface water in Luxembourg and highlight the usefulness of non-target screening for identifying temporal and spatial trends in pollutant levels. This approach, performed in a complementary manner to routine monitoring, can help to tentatively identify chemicals of concern for potential inclusion in target monitoring methods following additional confirmation and quantification efforts.

\*Correspondence:

Dagny Aurich

dagny.aurich@uni.lu

Emma L. Schymanski

emma.schymanski@uni.lu

Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

**Keywords** Surface water, High-resolution mass spectrometry, Non-target analysis, Exposome, Cheminformatics, Luxembourg

## Background

The variety and number of chemicals of concern in the environment continue to rise. The synthesis and registration of new chemicals happens regularly [1] and their presence in the environment and potential impact are often only recognized at a late stage. Furthermore, transformation products (TPs) of well-known chemical compounds frequently go unrecognized and are often not subject to routine monitoring. While they often exhibit properties similar to their parent, they can be even more harmful or persistent. This is for example true for the pesticide dichlorodiphenyltrichloroethane (DDT) and its TPs dichlorodipenyldichloroethylene (DDE) and dichlorodipenyldichloroethane (DDD) with all three compounds having carcinogenic properties and DDE being even more potent than its parent [2, 3]. In general, the origins of pollutants vary, as they can be side products of industrial, agricultural or medical applications or chemicals resulting from households ending up in the environment. Determining the source of each pollutant is therefore a highly challenging task but can be achieved in several ways, including looking at specific samples and sources or through evaluation of long-term temporal or geographical patterns.

Environmental monitoring covers just a small fraction of concerning chemicals, such that many even well-known chemicals ('known unknowns' [4, 5]) remain undiscovered in routine analysis. Therefore, it is important to take a step back from trusting solely targeted approaches and to apply and improve existing non-target (NT) methods. State-of-the-art chemical analysis methods combined with optimized cheminformatics workflows can provide a more comprehensive picture of the chemicals contained in a variety of environmental samples. The chemical load and the potential effect on human health or wildlife are of major interest when it comes to, e.g. chemical exposure assessment and exposomics research. Non-target analysis (NTA) allows for the monitoring of chemical exposure by identification of potentially toxic and persistent chemicals, even retrospectively. Looking at the full range of chemicals in sample measurements done for several years, time trends can be discovered (e.g. seasonal changes [6, 7]) and information could be found on the source and pattern of the pollution. Suitable sample types to perform such an analysis can range from soil, water and air to biological specimens, reviewed in more detail previously [8]. Looking at the availability of samples and the sampling or analysis requirements,

monitoring the state of water bodies presents a convenient way to measure environmental pollution. Especially looking at different flow paths contributing to a spread of pollutants, the input of wastewaters and meteorological phenomena such as flooding events, water becomes a very interesting and versatile sample type.

The small country Luxembourg, with an industrial past, is covered by 102 natural (or nearly unmodified) surface water bodies, displayed in Fig. 1 with different river catchments and flow directions. Pollution sources are numerous, in addition to industry, agricultural activities, traffic and household waste contribute to the country's chemical load [9]. Chemicals such as pesticides, flame retardants like polybrominated diphenyl ethers (PBDEs) or pharmaceuticals are contaminating water bodies [10–12], resulting in poor quality evaluations [13]. In 2022, the report on water quality showed that no river in the country is in a good condition (according to European "one-out-all-out" criteria [13, 14]) with



**Fig. 1** Surface water bodies in Luxembourg with different catchments and flow directions. Modified from: [geoportail.lu](http://geoportail.lu)

concentration values exceeded for chemicals like perfluorooctanesulfonic acid (PFOS), metazachlor or anthracene [12]. Moreover, only half of the groundwater bodies were in good status in 2022 [9]. In these evaluations, the ecological condition of surface water bodies was evaluated based on biological, physical, chemical and hydromorphological parameters [9, 13]. The chemical status was based on the analysis of priority hazardous substances and substance classes of greatest (EU)-concern, specified in the Water Framework Directive Annex VIII of Directive 2000/60/EC (Annex X) [14]. The directive was transposed into Luxembourgish law by the amended Water Directive of 2008 (Directive 2008/105/EC), known as the Environmental Quality Standards or Priority Substances Directive, setting environmental quality standards for surface water pollutants [15]. In 2013 an additional European watchlist mechanism was established, setting a list of substances to be monitored by all EU members [16, 17], which is updated regularly (newest version 2022 [18]). In 2022, the EU commission adopted a proposal to revise the list of priority substances, including 25 additional substances, e.g. per- and poly-fluorinated compounds (PFAS), bisphenol A, silver and several pesticides and pharmaceuticals [16, 19]. Besides looking at priority substances, a list of catchment-specific pollutants was considered, looking at the main catchments Mosel, lower Sûre, upper Sûre, Wiltz, Our, Alzette and Chiers (see Fig. 1) [12]. Cross-border rivers, such as the Alzette (arising from France), were particularly polluted [9], already containing a high chemical load when crossing the border (see Fig. 1, blue arrows indicating the flow direction).

This analysis can be expanded looking not only at the specified compounds for routine governmental monitoring, but also performing a full NT workflow to obtain an overview of additional chemicals that may be present, but are not a part of routine monitoring efforts yet. Performing this additional screening in parallel to the routine target analysis can help to identify risks posed by new or undiscovered contaminants in Luxembourg already at an early stage. This can be then used to guide legislative decisions, e.g. to expand the list of chemicals used for target screening in routine governmental laboratories (in Luxembourg the L'Administration de la Gestion de l'Eau (AGE)) or to improve wastewater treatment filtering systems. One step further, NTA can help support determining the (geographical) origin of contamination through differences between samples (e.g. influent/effluent of wastewater treatment plants (WWTPs) [20, 21]). Even the description of processes or transformation pathways may be possible, when looking at these measurements [22].

Currently, there is not one unified workflow to perform NTA of surface water samples. Several tools are

available nowadays to screen for unknowns, but there is no one-fits-all solution. High-resolution mass spectrometry (HRMS) was the method of choice to perform the analysis of the 271 Luxembourgish surface water samples collected at 20 sampling sites in Luxembourg between April 2019 and April 2022. In addition to the routine target monitoring conducted by AGE, a NT data analysis was performed allowing for retrospective analysis, screening, for e.g. previously not discovered chemicals. In this article, the open source R-package called *patRoan* [23, 24] was used, to perform the NTA. A detailed description of the package and its functionalities can be found in Helmus et al. 2021 and 2022 [23, 24]. The package combines functionalities of many tools like XCMS [25–27] or MetFrag [28] in one 'ready to use' package to harmonize and simplify the workflow of data processing of HRMS data in environmental sciences. The utilization of this open source tool offers a potential solution for facilitating collaboration among researchers, considering the existence of various instrument types and the consequent use of different and often incompatible software. It works with the open mzML format, for which a conversion exists for each vendor.

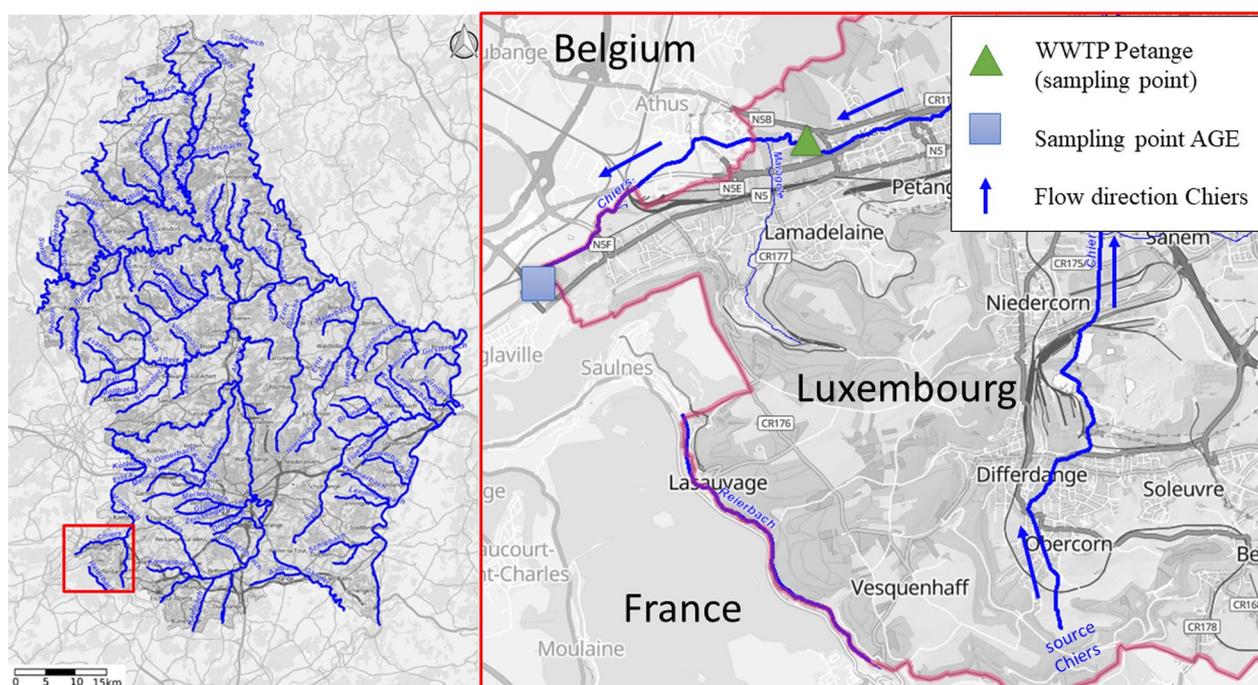
The workflow used in this study presents only one possible combination of steps (see Methods) to perform a NTA and can be expanded using, e.g. a suspect list to screen for specific substances. A novelty of *patRoan* 2.0 (compared to *patRoan* 1.0) was the possibility to perform a 'Sets' workflow processing positive and negative analyses at once, which was applied here [23]. The package workflow included peak picking, selecting relevant features, blank correction, removing the irrelevant ones and peak alignment. For the feature finding (via XCMS [25–27]) an optimization of the input parameters was performed. As shown by Libiseller et al. in 2015 [29], Albònica [30] or Tostengard and Smith [31], optimizing for the parameters *ppm* and *peakwidth* can significantly improve the results when using XCMS and the integrated *centWave* algorithm. *CentWave* is a feature detection algorithm integrated in the XCMS package applying continuous wavelet transformation (addition of Gauss-fitting is possible) to detect features even when they are partially overlapping [26]. After feature finding componentization could be performed identifying features belonging to one compound (adducts, isotopes, in-source fragments) as well as the generation of potential chemical formulas based on accurate mass and isotope patterns. For interpretation and possible identification of features, substance and spectral databases such as *PubChem* [32] and *MetFrag* [28] are required. In this study, a subset of *PubChem* called *PubChemLite* [33, 34] was used, focusing especially on exposomics-related compounds.

The resulting features were categorized by identification level [35] using the *individualMoNAScore* (spectral similarity of the candidate structure in the MassBank of North America (MoNA) [36]). Moreover, the compounds were classified using the tool *classyFire* [37] whose primary function is to classify chemical compounds based on their structural features and properties and assign them to specific chemical classes and subclasses. *ClassyFire* is widely used and can provide valuable information about the chemical composition, functional groups, and potential biological activities of a compound. In addition, *PubChem* annotation content [34] was used to estimate the environmental effect and to determine possible sources of the chemicals like agriculture, households or industry. *PubChemLite for exposomics* makes use of selected categories available in the *PubChem* Table of Contents Classification Browser [38]. These categories can help categorizing the tentative identifications. Categories used in this study included agrochemical use (*agroChemInfo*), drug and medication information (*drug-MedicInfo*), associated disorders and diseases (*disorderDisease*) and use and manufacturing (*knownUse*) [34].

A comparison to prior studies of the Environmental Cheminformatics (ECI) group at the University of Luxembourg—focussing at pesticides and pharmaceuticals in surface water [10, 11]—was conducted, looking at shared identifications to check for plausibility. These studies proved already the presence of high pesticide (even

banned compounds and transformation products) and pharmaceutical load in surface waters and complemented target monitoring efforts of AGE [10, 11] but did not look beyond these classes. For the river Chiers, located in an industrial region in the south-west of Luxembourg, an additional comparison was performed looking at compound findings from a 2022 sampling campaign, see Fig. 2. The sampling was performed from May to June 2022 at the inlet of a WWTP in Petange, located prior to the sampling point of this study. The study results were then evaluated based on the geographical information on industry, households and hospitals located in the region. Finally, the AGE target list was compared to the study findings, looking as well at catchment-specific pollutants, and discussing possible candidates to include in routine monitoring.

The primary aim of this article is to employ NTA as a complementary approach to routine target monitoring, with the objective of offering tentative insights into chemicals of concern that are currently not under surveillance to form recommendations for future target monitoring efforts. To achieve this, an open source and adaptable NT workflow is proposed as an alternative to conventional vendor software. By adopting this approach, it becomes feasible to accommodate a wide range of instruments, thereby facilitating collaboration among researchers and authorities/regulators. This applies specifically to the Luxembourgish case as AGE uses a Sciex



**Fig. 2** Two sampling sites at the river Chiers. Modified from: [geoportail.lu](http://geoportail.lu)

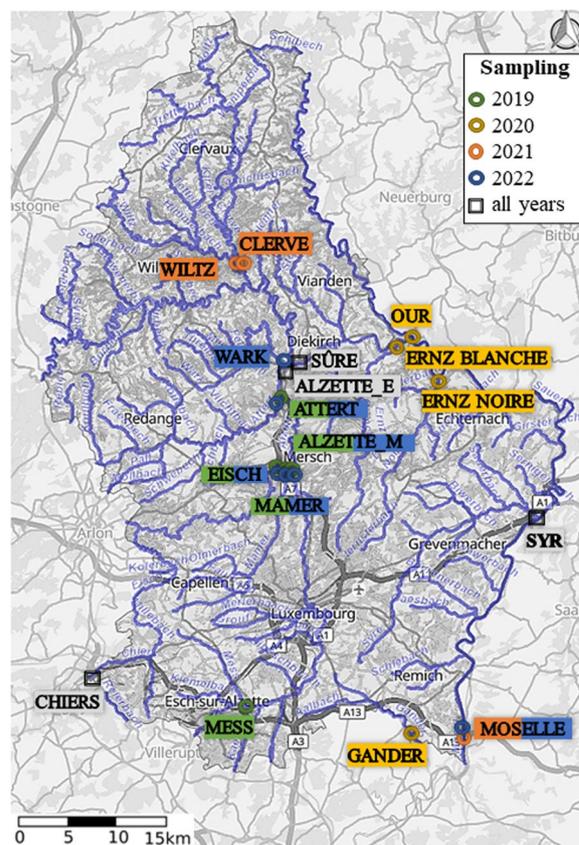
and the University of Luxembourg a Thermo Fisher Orbitrap device. The combination of various tools discussed herein represents a singular, potential method for processing HRMS data via NTA. Furthermore, this study aims to compare NT-HRMS measurements obtained from samples of Luxembourgish surface water in order to identify temporal and/or spatial patterns and to classify chemicals found using a variety of tools. Based on the results, the article aims to explore potential sources of pollutants, as well as estimate potential impacts on both the environment and human health.

## Methods

### Sample preparation and analysis

Two hundred and one water samples were collected every 4 weeks by AGE at the sampling sites indicated in Fig. 3 and in Additional file 2: Table S1. The 3-year sampling analysed in this article took place between April 2019 and April 2022 with each year having varying sampling points on a rotational basis, spread throughout Luxembourg, selected by AGE. The same four sampling points were analysed every year, the remaining river locations vary in a 3-year cycle to cover different geographic regions (or catchments) in the country (see Fig. 1). The four constant rivers—displayed as black squares in Fig. 3—were Chiers, located in the south-west (Chiers catchment), Syr in the east (Mosel catchment), and in the centre Sûre (upper Sûre catchment) plus the Alzette near Ettelbruck, hereafter ‘Alzette\_E’ (Alzette catchment). Some measurements were unavailable due to differing reasons, e.g. meteorological circumstances (excluded months: November 2019–March 2020). For March 2021 and 2022 there was an additional sampling performed at the end of the month, indicated as ‘Mar\_end’. Due to method and instrument instabilities, several months of 2021 were remeasured in 2022, resulting in increased feature and identification numbers for the remeasured analyses ‘Mar\_end-21, Apr-21, May-21’.

The 271 surface water samples were extracted (solid phase extraction, SPE) as described in Krier et al. [10] using the Atlantic<sup>®</sup> HLB SPE Disks (Horizon, Salem, NH, USA) with a 47 mm diameter and the SPE-DEX 47900 system (Horizon). The filtered extracts were spiked with a 100 ppb mix of 10 internal standards: Melamine-13C3-15N3, Carbendazim-D4, Sucralose-D6, 5-Methyl, Benzotriazol-D6, Neotame-D3, Metolachlor-D6, 5-Fluorouracil-15N2C13, Torsemide-D7, Triclosan-D3, Carbamazepine-D10 purchased from Santa Cruz Biotechnology, Heidelberg. Then they were analysed via Reversed Phase LC-HRMS using the Waters Acquity UPLC BEH C18 column and the Thermo Q Exactive HF Orbitrap Mass Spectrometer. Further analytical details



**Fig. 3** Sampling sites between 2019 and 2022 (exact location in Additional file 2: Table S1). Mixed shading corresponds to sampling done in both years. Modified from: geoportail.lu

including QA/QC procedures are given in Krier et al. 2022 [10].

### Data analysis

Several R-packages were used to perform the HRMS data processing and the following data analysis steps. The version of R and all installed dependencies of the open source package *patRoan* and other used packages are listed in the LCSB GitLab repository of the Environmental Cheminformatics Group (ECI). Moreover, the R script used for optimization and the full NT script can be found in a subfolder of the repository, while the raw data are available on GNPS (<https://doi.org/10.25345/C55X25P62>).

This study presents an NTA workflow of 271 LC-HRMS surface water measurements (in positive and negative mode), making use of a modifiable data processing workflow established by using the R package *patRoan*. The data files were converted to mzML via *ProteoWizard's MSConvert* (version 3.0.21075) [39, 40] using a peak picking (centroiding) filter. Therefore, the pre-treatment step implemented in *patRoan* was not applied here,

starting directly with the finding and grouping of features (functions `findFeatures`, `groupFeatures`) in ‘Sets’ mode. The package *XCMS* [25–27] was used to perform feature finding and later grouping. To receive best results for the feature dataset a feature optimization step was performed (as explained above) with the *patRoön*-integrated Isotopologue Parameter Optimization (IPO) [29] algorithm [function `optimizeFeatureFinding` resulting in an iterative process using Design of Experiments (DoE)]. The feature grouping using *XCMS* was followed by a basic rule-based filtering operation (filter) applying a blank, intensity and replicate group filter. To annotate those features, tables of averaged mass spectra (MS and MSMS) for each feature—so called *MS peak lists*—were created (`generateMSpeaklists`) using functionalities of the *mzR* package [39–43]. Those lists were filtered thereafter (filter) limiting the results to the top 25 MSMS peaks. Compounds were then generated using *MetFrag* [28] and the *PubChemLite for exposomics* library [33, 34] (`generateCompounds`). For each feature group possible candidate compound structures were identified and then ranked, e.g. based on the matching fragmentation (MSMS) data. Several scoring parameters can be set in this step, including the *individualMoNAScore*, which was used in the next step to determine the level of identification. For the simplicity of analysis, three levels were chosen: a good MSMS library match, i.e. level 2 scored at least 0.9, a fair match, i.e. level 3a lay between 0.7 and 0.9 and level 3b was defined to be between 0.4 and 0.7 (adapted scheme from the NTA study by Talavera Andújar et al. [44]). For each feature, the tentatively identified candidate with the highest score was selected for the final scoring and reporting. The following data analysis involved a classification of chemicals using the web interface *classyFire* [37] and four classification categories of the *PubChemLite* database (*agroChemInfo*, *drugMedicInfo*, *disorderDisease*, *knownUse*), which are available in the database file and online for each chemical record. The inter- and intra-year occurrence of compounds and compound classes was then analysed. The results were evaluated looking at the 3 years of measurements, presented in the Results, followed by a critical discussion and evaluation of the used tools, in the Discussion.

## Results

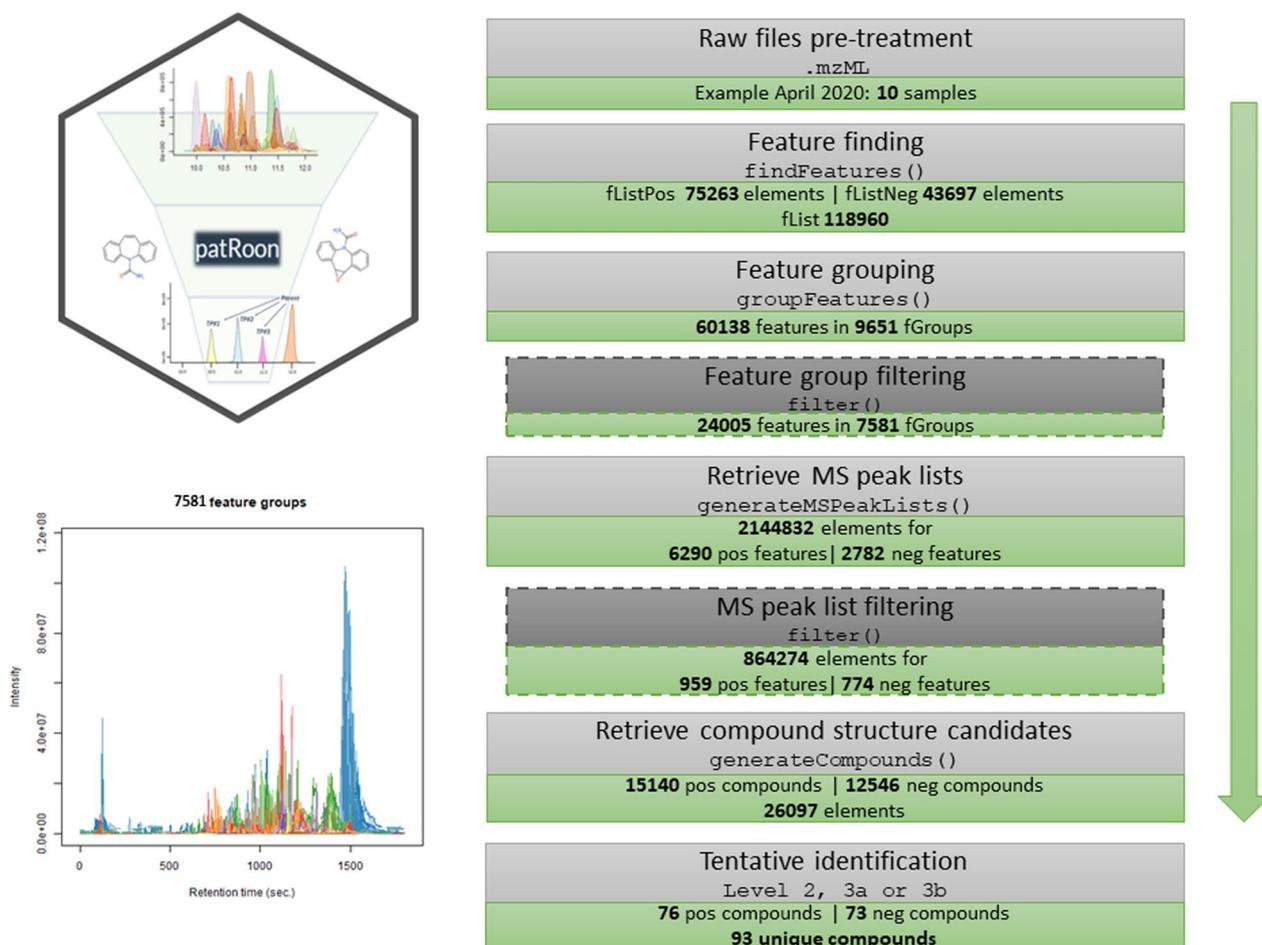
This section includes the summarized results from the NT workflow of the 271 Luxembourgish surface water samples analysed that are of solely qualitative nature. The workflow started with the optimization of the *ppm* and *peakwidth* parameters to perform feature finding, as described above. An example of different DoEs visualized by perspective plots can be found in Additional file 1: Figure S1 for the samples of April 2020 (in negative

mode). In addition, a visualization of the best parameters determined for positive and negative mode for the same month is shown in Additional file 1: Figure S2. The full list of optimized feature finding parameters for *ppm* and *peakwidth* can be found in Additional file 2: Table S2.

After optimizing the feature finding parameters, the actual NT analysis of the measured samples was performed. Figure 4 shows the applied *patRoön* workflow with data collected for the ten April 2020 samples. In total 75,263 positive and 43,697 negative features were found in the first step of the workflow, totalling to 118,960 features. After feature grouping and filtering, using the inbuilt *patRoön* functionality [23, 24] (see above), the number was reduced to 24,005 features in 7,581 feature groups. After the generation and filtering of MS peak lists, 15,140 positive compounds and 12,546 negative compounds could be assigned to the feature groups (see Fig. 4). Applying the identification scheme explained in the Data Analysis section [44], 76 positive and 73 negative compounds could be identified at levels 2, 3a and 3b, of which 93 were unique compounds and 56 were overlapping (i.e. they were tentatively identified in both positive and negative mode).

Most of the rivers are interconnected in Luxembourg and therefore the same compounds appear in several measurements. There are catchment-specific pollutants—monitored by AGE—appearing mainly in the regions indicated in Fig. 1 [12]. Figure 5A shows overlapping features (using a Venn diagram) for the four rivers monitored regularly. The most feature groups were detected for the river Syr, which overlapped most with the surface water from Chiers and Alzette\_E (975). However, all four rivers are located in different catchments with different, region-specific influences and therefore the overlap is not 100%. In Fig. 5B, a Chord plot for all feature groups in all rivers in April 2020 is presented. All rivers showed several overlapping feature groups with clear overlaps of some rivers belonging to one catchment, e.g. Gander and Mosel. However, this is not always the case, looking, e.g. at the two rivers in the Lower Sûre catchment or the large overlap between Alzette\_E and Chiers.

The analysis steps presented in Fig. 4 were accordingly performed for all 34 months and the resulting tables can be found in the GitLab repository. In Additional file 2: Tables S3 and S4 the number of 2, 3a and 3b identifications for positive and negative mode, their sum and the number of tentatively identified unique compounds per level can be found. There was a majority in level 2 identifications compared to the level 3 numbers, e.g. for the April 20 samples there were 58 level 2s, 22 3as and 17 3bs. The total number of positive, negative and unique identifications



**Fig. 4** patRoof workflow (workflow step terms described in [23, 24]) with exemplary values and feature groups plot for the April 2020 analyses, resulting in level 2, 3a or 3b identifications

(without discriminating between levels) is demonstrated in Fig. 6 based on Additional file 2: Table S4. The numbers of positive and negative (unique) identifications are presented in yellow and green and a black bar shows the total number of unique compounds. The count of identifications in positive mode is generally higher than the negative count and their overlap is shown in blue. Overall, a total of 2479 compounds were annotated. After deduplication, 378 unique chemicals remained with level 2, 3a and/or 3b. The chemicals identified per month and in total can be found in Additional file 2: Table S5 (further details are available in the ‘Output\_summary\_patRoof’ folder uploaded on GitLab), including tentative identifications of pharmaceuticals like valsartan or metformin, agrochemicals like 4,6-dinitro-o-cresol (DNOC) or their TPs like Flufenacet ESA and industrial chemicals

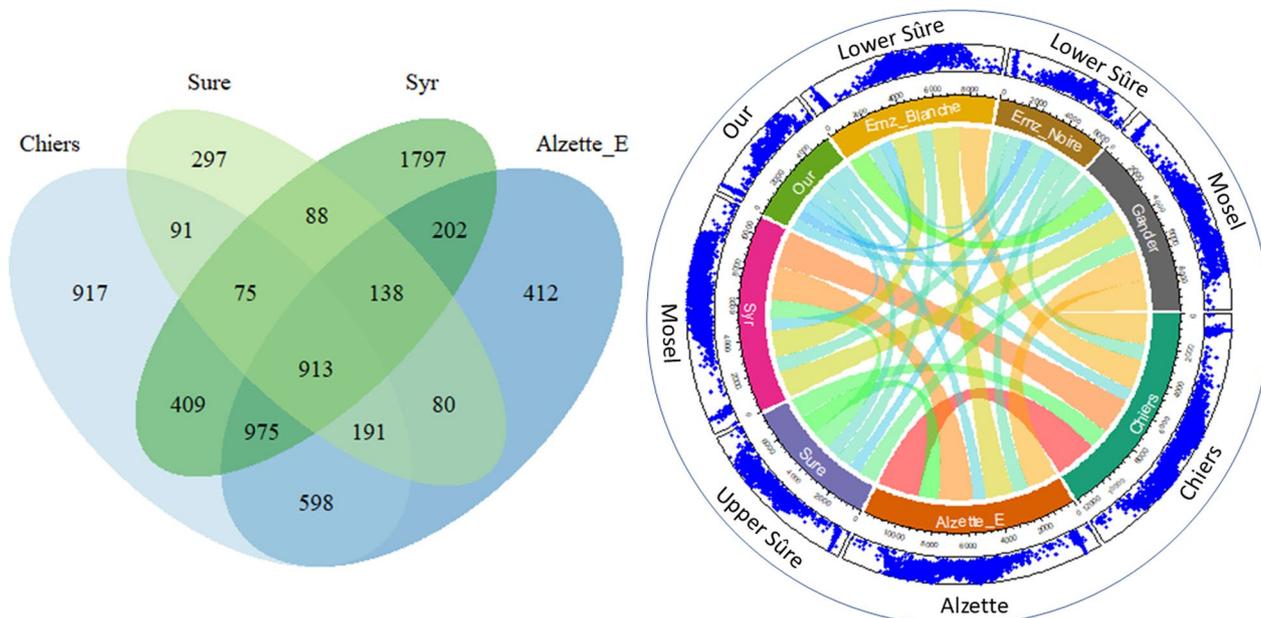
like benzotriazoles, methylbenzenesulfonamide or bisphenol S.

### Classification

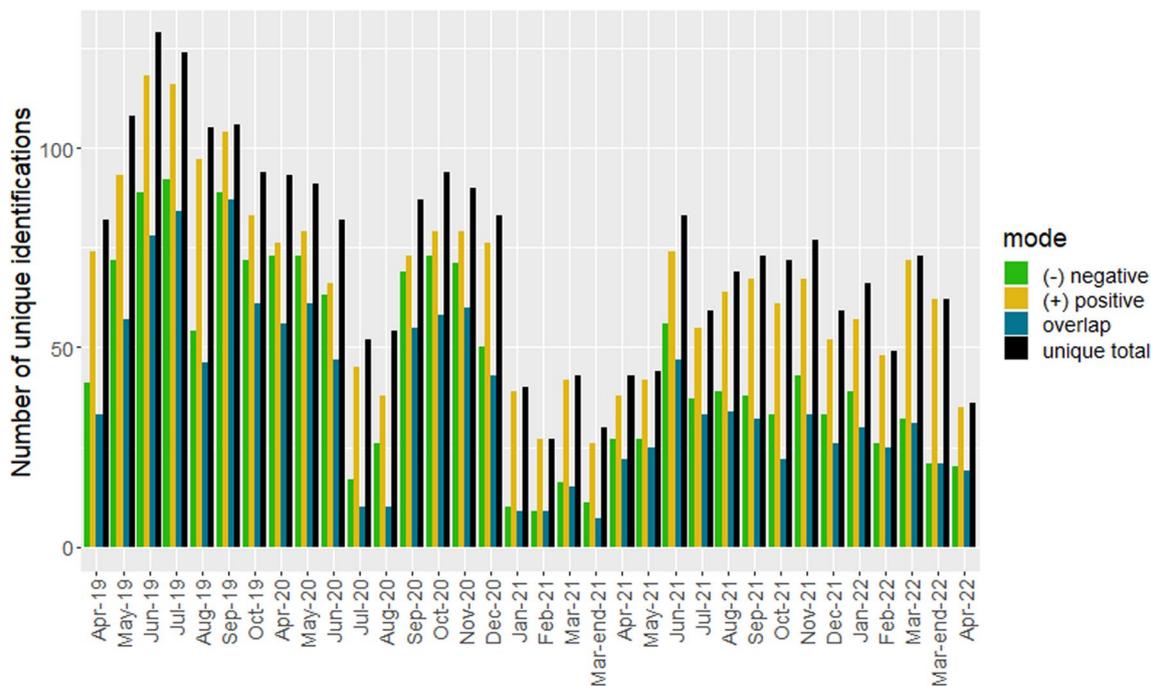
To get a better overview of and group/interpret the tentative identifications, classification steps were performed.

### ClassyFire

First, an ‘interannual’ (April results of all years) and an ‘intraannual’ (2021 results of all months) comparison was performed, looking at the number of identified compounds per *classyFire* class and parent class (superclass). The month April was the only one measured in all years and 2021 was the only year where samples were available for each month. In general, for the interannual and intraannual comparison, 11 main parent classes (superclasses) could be identified: organic oxygen compounds,



**Fig. 5** **A:** Venn plot of feature groups in April 2020 for the four rivers monitored all years; **B:** chord plot for all feature groups of all April 2020 analyses with river catchments



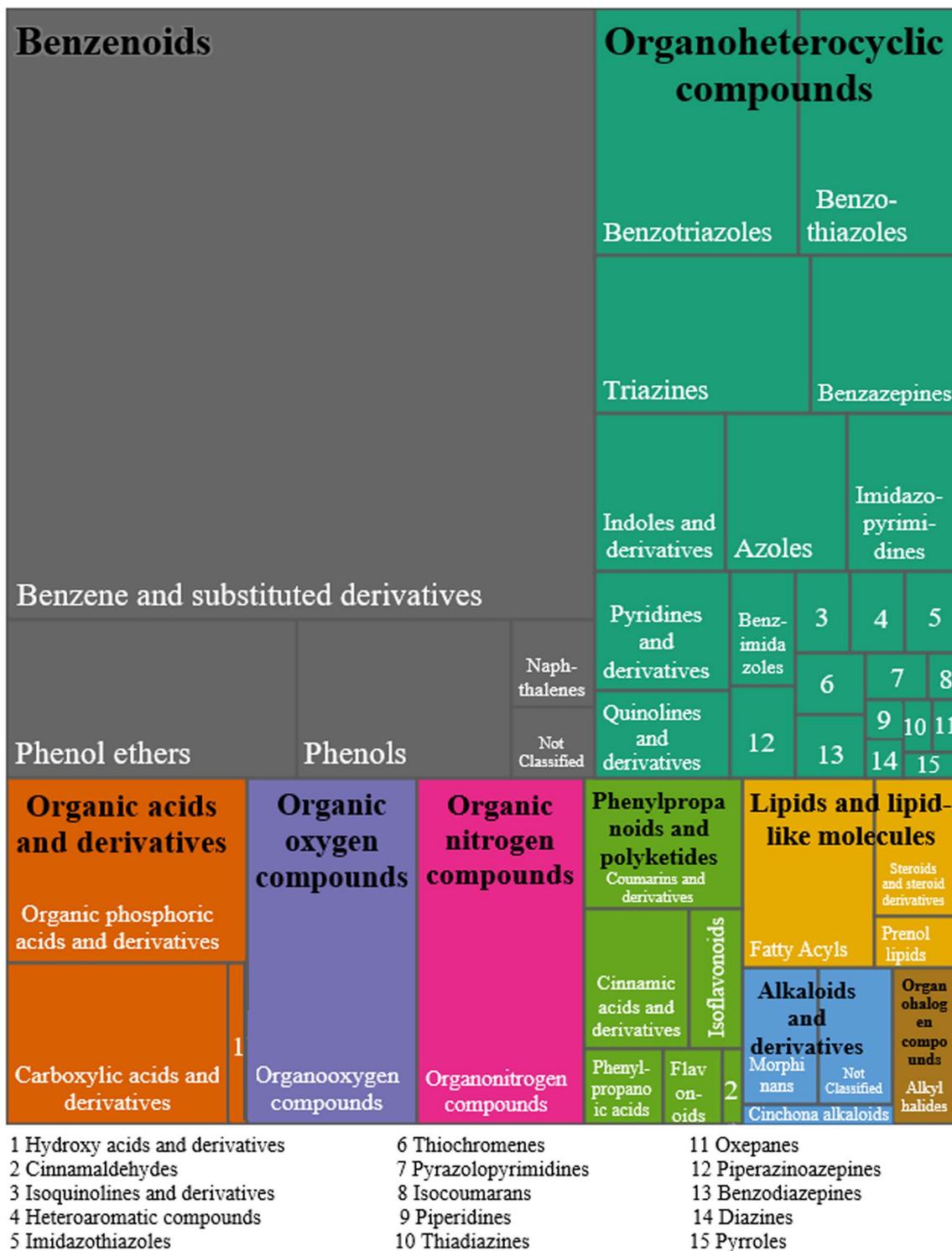
**Fig. 6** Number of unique positive, negative, overlapping and total identifications per month

organohalogen compounds, nucleosides, nucleotides, and analogues, organic nitrogen compounds, organo-sulfur compounds, lipids and lipid-like molecules, alkaloids and derivatives, benzenoids, phenylpropanoids and

polyketides, organic acids and derivatives and organo-heterocyclic compounds. 50 unique subclasses of those very general superclasses could be assigned (46 in 2021), giving a more detailed picture. The underlying data (total

numbers and percentage of compounds found per class and superclass in the inter- and intra-annual comparison) are included in Additional file 2: Table S6. An overview

of those compound classes can be seen in Fig. 7 using the summarized identification numbers of all analysed months in 2021.



**Fig. 7** Treemap of classifyFire classes and subclasses using the summarized identifications of 2021

The treemap in Fig. 7 shows that nine superclasses with several subclasses could be identified for the intraannual comparison of measurements in 2021. Most of the chemicals were categorized as benzenoids (43%) followed by organoheterocyclic compounds (26%) and organic acids and derivatives (8%). Comparing the intraannual results of 2021 with the interannual comparison of the month April between 2019 and 2022, additional chemical classes were observed. One purine nucleoside, one sulfoxide and one compound belonging to the pteridines and derivatives class were tentatively detected in 2019. Purine nucleosides are generally not considered to be harmful to the environment or human health, as they are essential components of normal cellular functioning. Some sulfoxides have been shown to have toxic effects (e.g. dimethyl sulfoxide, DMSO), particularly when they are not properly disposed of or when they enter the water supply [45]. However, looking at the measurement results of April 2019, the compound was sulforaphane (in positive mode) at the sampling points Alzette\_E, Syr, Mess, Mamer, Atert and Alzette\_M (Alzette sampling point Mersch-Berschbach), which is a naturally occurring compound that is safe for human consumption and is used in cancer treatment. The same applies for pteridines and derivatives, some chemicals of this class have been shown to have toxic effects (e.g. atrazine), but the identified compound was in this case riboflavin, also known as vitamin B2. Overall, these examples show (and it is important to remember) that the toxicity of a chemical is complex and context-dependent, and should be evaluated on a case-by-case basis. Generally, the toxicity assessment in terms of environmental and health hazards is difficult, as the toxicity of a chemical can depend on a variety of factors, including its chemical structure and specific chemical properties, concentration, mode and duration of exposure, and the susceptibility of the organism or sensitivity of the ecosystem exposed. As the concentration is not measured in this study, little can be said about the toxicity of the annotated chemicals and the chemical class as such gives only limited to no information about the environmental or health impact. Additionally, different chemical classes can have different toxicities for different organisms, and different endpoints (such as acute toxicity, chronic toxicity, carcinogenicity, mutagenicity, and reproductive toxicity) may also be relevant. It has to be considered that some compounds may have multiple classifications, and their potential impact on the environment and human health may vary depending on the specific application. The use of *classyFire* is examined further in the Discussion.

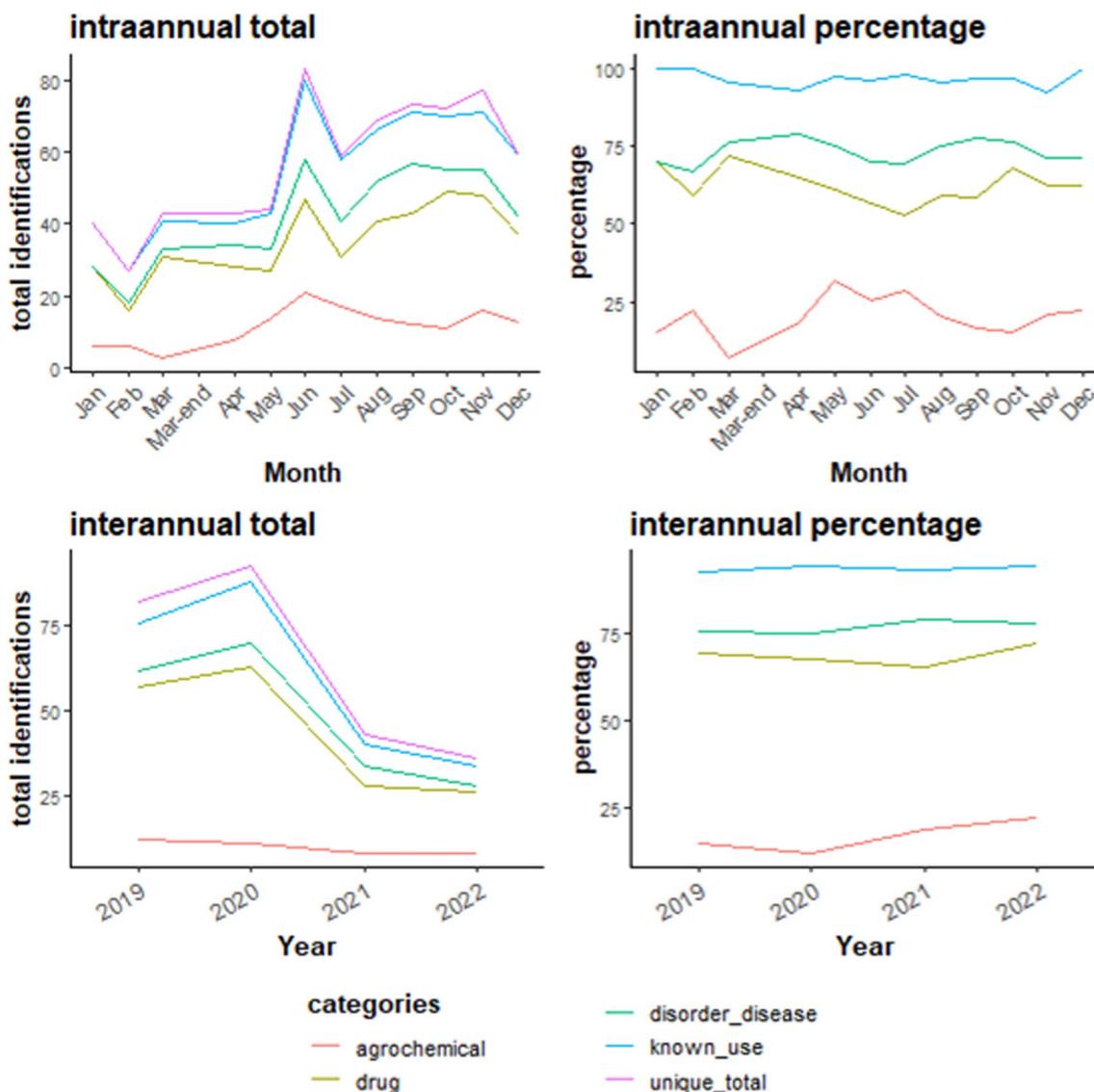
### **PubChemLite categories**

To identify possible sources and estimate the environmental impact of the exposome related chemicals, a classification of the compounds in the inter- and intra-annual comparison was performed, using the annotation content available in *PubChem* for each chemical (via the *PubChemLite* categories described in [34]). The categories *agroChemInfo* and *drugMedicInfo* were chosen to evaluate trends of agrochemical and pharmaceutical use in 1 year and over 3 years. Moreover, information about possible disorders and diseases related to a compound and known commercial uses were analysed using the *disorderDisease* and *knownUse* categories. The resulting total and percentage trends are visualized using four line charts in Fig. 8 and the raw numbers are summarized in Additional file 2: Table S7. It has to be considered that the categories identified are not exhaustive, and there may be some overlap between them (multiple uses per chemical).

Looking at the intraannual comparison of all months in 2021 an overall increase of total numbers in all categories could be monitored, but the overall percentage (relative to total numbers) stayed roughly the same. A majority of the chemicals had associated disorders and diseases content in *PubChem* (between 67 and 79%), while 53% (July) to 72% (March) of the chemicals were assigned to the class of pharmaceuticals and the percentage of agrochemicals was between 7% (March) and 32% (May). This corresponds to the usual 'spraying rhythm' of farmers who increase pesticide and herbicide spraying in May to lay a foundation for the harvest. Almost all identified chemicals (93–100%) had a documented use, with multiple matches per compound when looking at the individual case in *PubChem*. The interannual values showed a sharp decrease of total identifications in drugs, disorders and diseases and known use, either due to effects of the COVID pandemic or due to measurement variations (less likely as the agrochemical curve stayed more or less constant). The percentage values (% of total identifications) showed a constant trend between the years with nearly all annotated compounds having a known use, 75–79% being associated to disorders and diseases, ~70% being drugs and 12–22% agrochemicals according to the *PubChemLite* classification.

### **Comparison to other studies**

Besides using classification workflows, data from other water studies can be used to determine possible sources of exposome related chemicals. Former studies looking at Luxembourgish surface waters provided evidence that there are more pharmaceuticals and agrochemicals entering the environment than those included in the target monitoring by AGE that could potentially cause harm



**Fig. 8** Classification of tentatively identified compounds in the inter- and intra-annual comparison, according to PubChem

[10, 11]. Regarding agrochemical compounds in Luxembourgish rivers, a suspect and related transformation product screening study was conducted by Krier et al. [10] with the same instrumental methods and a subset of the data used here. The study identified 162 pesticides and 96 TPs in the water samples (several chemicals not allowed in Luxembourg). 31 chemicals were confirmed at level 1 [10]. Comparing these results to this study an overlap of 36 agrochemicals was seen, listed in Additional file 2: Table S9. Since that study focused exclusively on pesticides, which are often present in lower concentrations than, e.g. pharmaceuticals and industrial chemicals, it is likely that several compounds identified by Krier et al. may have other top-ranked candidates in

the current study as they have been prioritized using different scoring terms (shown to improve ranking results [34]) that were not available to Krier et al. at the time. Singh et al. [11] performed a suspect screening, identifying 94 pharmaceuticals, adding quantification steps later. The AGE monitoring however, included just five pharmaceuticals (list of AGE from 2019 and 2020 in Additional file 2: Table S8): carbamazepine, diclofenac, ibuprofen, ketoprofen, and lidocaine. All five chemicals were identified in the work of Singh et al. as well as in this study. Of the 232 pharmaceuticals tentatively identified in this study, 58 were also confirmed in the results of Singh et al. [11], including the 5 covered in the AGE monitoring. The compared lists and overlapping identifications

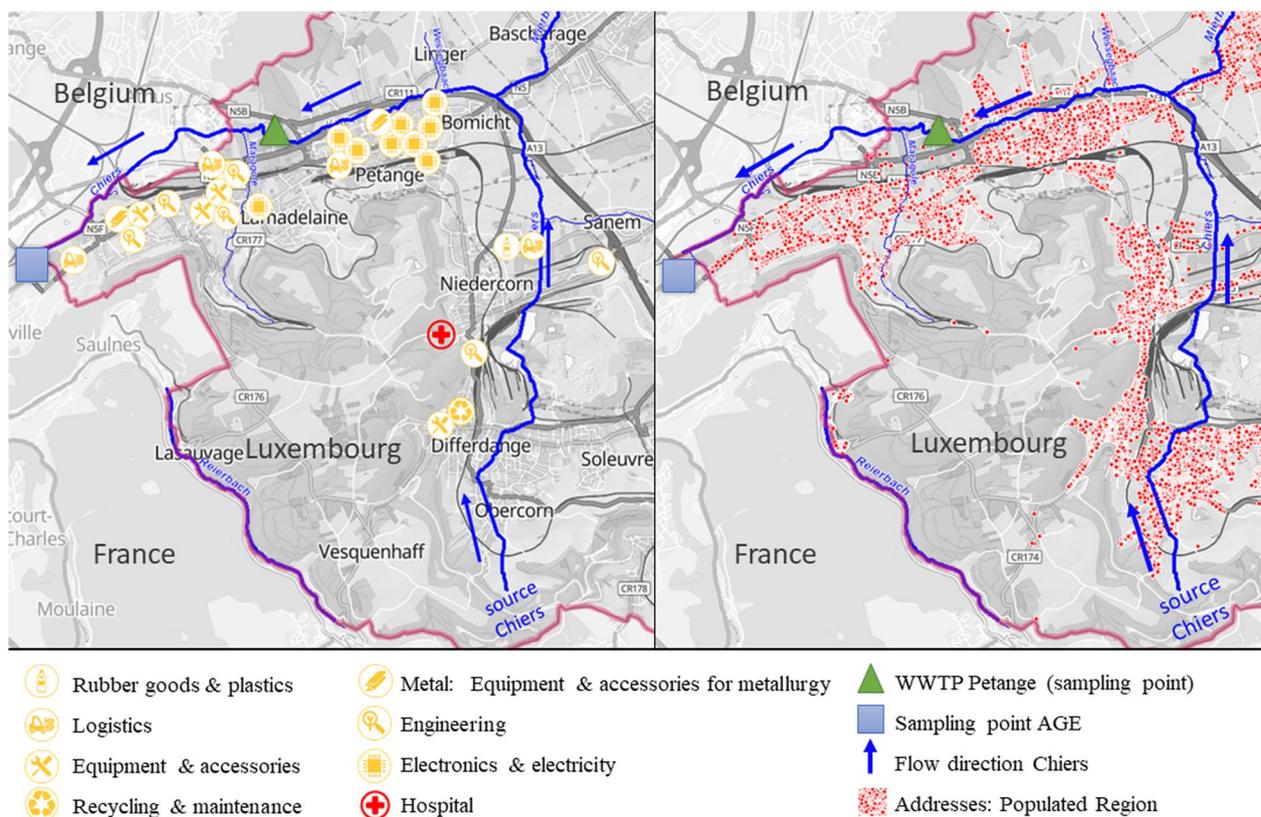
are summarized in Additional file 2: Table S10. Singh et al. also registered the trend of decreasing pharmaceutical load looking at the years 2019 and 2020, explaining it with the reduction of medical treatments due to the COVID pandemic and lower precipitation [11].

A combination of geographical information, information on flow paths and additional measurement data from the inlet of a WWTP was used to analyse possible sources of the chemicals found in the river Chiers. The river is located in the south-west area of Luxembourg, at the border to France, with exposure to a set of different sources of pollution (see Fig. 9). Its source is in Obercorn, it passes the WWTP in Petange (green) and 6 km later the sampling point of this study (blue), located at the border to France.

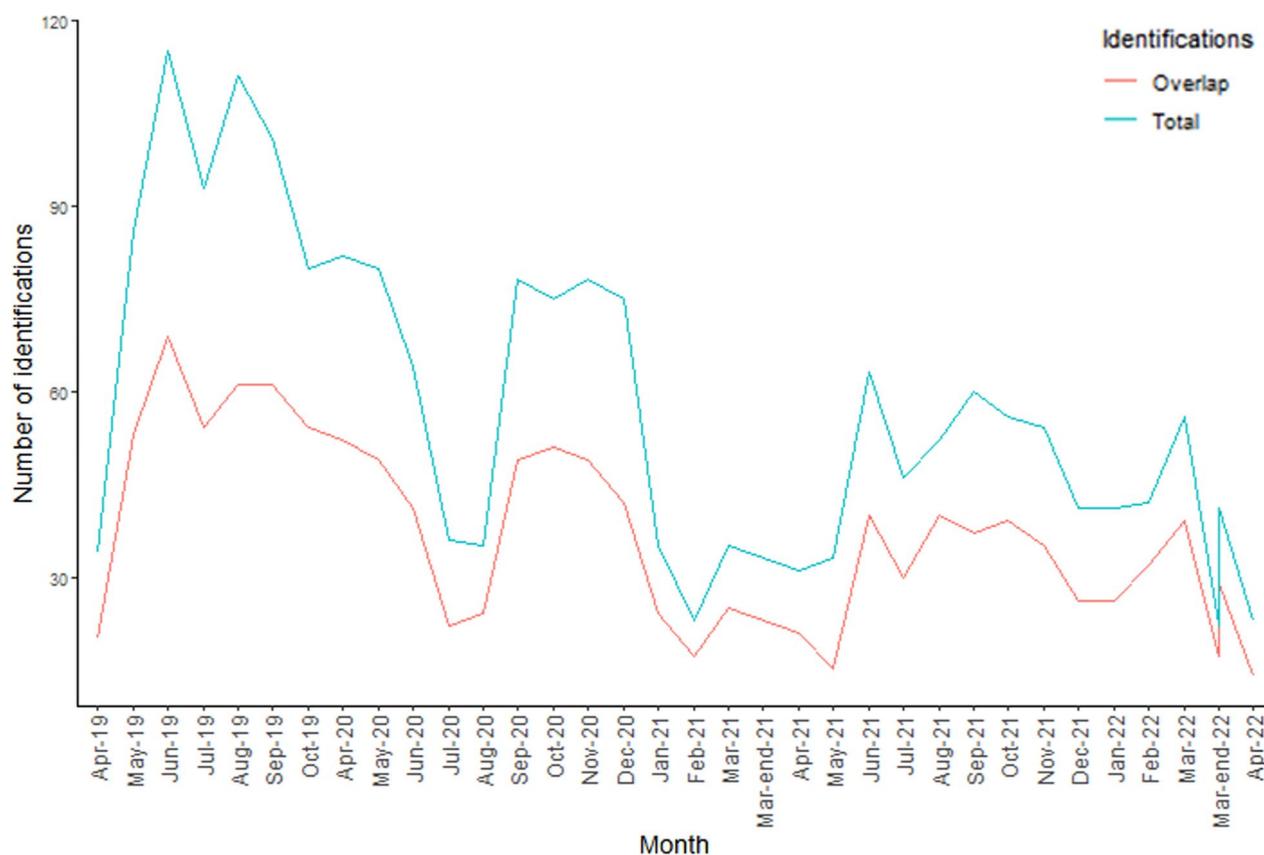
The measurements at the WWTP took place between May and June 2022, resulting in 409 tentatively identified chemicals. Comparing those findings to this study (all results from Chiers in 2019, 2020, 2021 and 2022), an overlap of 178 chemicals could be identified. Figure 10 shows the number of overlapping chemicals per month compared to the total identifications (AGE sampling point). Those chemicals were probably coming

from the WWTP with sources before this sampling point and result from incomplete filtering or there was chemical input between the WWTP and the AGE sampling point. Other chemicals were effectively filtered by the WWTP system or could not be identified at the later sampling point. For the overlapping chemicals, the four PubChemLite categories analysed above were examined as well, resulting in 36 agrochemicals, 130 pharmaceuticals, 143 compounds associated with disorders and disease and 170 known uses. Consequently, the same trend with dominating identifications of pharmaceuticals (73%) could be observed here, even after the filtering of the WWTP. Persistent synthetic chemicals, like the PFAS perfluorooctanoic acid (PFOA) or perfluorobutanesulfonic acid (PFBS) were found before the WWTP and downstream of the Chiers. However, without having quantitative data on their concentration, little can be said about their environmental effects.

Besides the overlapping chemicals, it is interesting to analyse chemicals identified only at the border to France and not at the WWTP inlet (in total 165 chemicals). Looking for example at the results from April 2022 (1 month before the Petange WWTP sampling), drugs



**Fig. 9** A: Industry and hospitals located next to the sampling points at the river Chiers; B: populated region around the river Chiers. Modified from: geoportail.lu



**Fig. 10** Number of chemicals overlapping with the Petange WWTP inlet per month compared with the total identifications at the AGE sampling point

like pregabalin (antiepileptic), tramadol (analgesic) and its TP n-desmethyltramadol (with high aquatic toxicity) were found besides other compounds like 1H-benzotriazole (anticorrosive). Again, most of the unique tentative identifications at the border to France were pharmaceuticals (56%) and related to disorders and diseases (65%). The detection of the antiepileptic pregabalin could indicate the medication being used in this area (see population distribution in Fig. 9). Industrial chemicals, like benzotriazoles, only identified after the WWTP, could result from activity in the equipment and accessories, electronics, engineering or metal industry located in the area between WWTP and border. Other PFAS identified solely at the later sampling point were perfluorononanoic acid (PFNA) and perfluoroheptanoic acid (PFHpA), known for their use as surfactants, in fire fighting foams, for the manufacturing of plastics and in the semiconductor industry. These substances are now being phased out in many applications due to their persistence in the environment and potential adverse health effects. The list of 409 chemicals compared to the Chiers results of this

study can be found in Additional file 2: Tables S11 and S12.

Lastly, a comparison to the chemicals covered by the governmental target monitoring (AGE) was performed, using the screening lists from 2019 and 2020 and the published results from the water quality report in 2022 [12] (Additional file 2: Table S8). A total of 40 identified chemicals were overlapping with the target monitoring of AGE, including eight (of 16) catchment-specific pollutants: carbamazepine, metolachlor, terbuthylazine, chlorotoluron, tebuconazole, flufenacet, metolachlor ESA and metazachlor OXA. Among the eight not detected chemicals were, e.g. metolachlor OXA and metazachlor ESA, both being TPs of metolachlor and metazachlor, just as metolachlor ESA and metazachlor OXA. 338 chemicals not covered by target monitoring remained and were ranked based on their frequency of occurrence (number of months out of 34) in the Luxembourgish rivers between 2019 and 2022. The top 54 chemicals identified, occurring in at least 13 months, were listed with their common use, the *PubChem* Chemical Identifier (CID),

**Table 1** List of chemicals with high occurrence in the 34 months analysed, not currently monitored by AGE and of interest to add to future monitoring lists

Synonym	Use	Parent name	PubChem CID	No of occurrences
Irbesartan	Pharmaceutical	–	3749	32
1H-Benzotriazole	Industry	–	7220	32
4- or 5-methyl-1H-benzotriazole (co-elution)	Industry	–	8705 or 122499	31
Amisulpride	Pharmaceutical	–	2159	31
Telmisartan	Pharmaceutical	–	65999	26
Celiprolol	Pharmaceutical	–	2663	25
Fluconazole	Pharmaceutical	–	3365	25
Trimethoprim	Pharmaceutical	–	5578	24
4-Acetamidoantipyrine	Pharmaceutical	Metamizole	65743	23
Codeine	Pharmaceutical	–	5284371	22
Desvenlafaxine	Pharmaceutical	Venlafaxine	125017	22
Tramadol	Pharmaceutical	–	33741	22
4-NP	Industrial	–	980	22
Valsartan	Pharmaceutical	–	60846	21
Fexofenadine	Pharmaceutical	–	3348	21
Salicylic acid	Industrial, pharmaceutical	–	338	20
Flecainide	Pharmaceutical	–	3356	20
Tiapride	Pharmaceutical	–	5467	20
TCEP	Flame retardant	–	8295	19
TCPP	Flame retardant	–	26176	19
Flufenamic acid	Pharmaceutical	–	3371	19
Aspirin	Pharmaceutical	–	2244	19
Sitagliptin	Pharmaceutical	–	4369359	19
Triethyl phosphate	Industrial	–	6535	18
Adipic acid	Industrial	–	196	18
Losartan	Pharmaceutical	–	3961	17
4-Formylaminoantipyrine	Pharmaceutical	Aminopyrine	72666	17
Sulfamethoxazole	Pharmaceutical	–	5329	16
Carbamazepine-10,11-epoxide	Pharmaceutical	Carbamazepine	2555	15
Bicalutamide	Pharmaceutical	–	2375	14
Dibutyl phthalate	Industrial	–	3026	14
PFOA	Industrial	–	9554	14
2-Hydroxycarbamazepine	Pharmaceutical	Carbamazepine	129274	14
Sulisobenzone	Consumer products	–	19988	13
Sulpiride	Pharmaceutical	–	5355	13
3-Hydroxypyridine	Industrial	–	7971	13
1-Methylbenzotriazole (possible co-elution)	Industrial	1H-Benzotriazole	25902	13
D617	Pharmaceutical	Verapamil	93168	13
Ensulizole	Consumer products	–	33919	13
Tributylamine	Industrial	–	7622	13
Hydrochlorothiazide	Pharmaceutical	–	3639	13

The full list of 54 compounds (including, e.g. food/natural products) is available in Additional file 2: Table S13

the number of occurrences and additional information like the CID of the parent compound (in case of TPs). Chemicals without an environmental or health effect

according to *PubChem* data were excluded, e.g. natural products, food additives or ubiquitous compounds like caffeine, reducing the list to 41 entries (Table 1).

26 of 41 chemicals tentatively identified are classified as pharmaceuticals (20) or their TPs (six), predominating the analysis results. Two TPs of carbamazepine (parent included in AGE monitoring) were tentatively identified: carbamazepine-10,11-epoxide in 15 months and 2-hydroxycarbamazepine in 14 months, whereas the parent compound was only found in 11 months of this study. Desvenlafaxine, a TP of the antidepressant venlafaxine, was detected in 22 months; the parent compound was already detected in the study by Singh et al. [11] and in 12 months of this study, having a known impact on aquatic environments even at low levels [46]. Out of the 26 pharmaceuticals, seven parent and five TP chemicals were not covered by Singh et al. [11] as well as three parents of the five TPs: metamizole, aminopyrine and verapamil. They might be measured after the study by Singh et al. was conducted or have been missed due to variations in the identification approach. The example of pharmaceuticals proves that it might be worth adding (more) TPs of monitored chemicals to routine target monitoring, as the parent is sometimes not visible and their TPs cause risks as well. Furthermore, 16 chemicals were listed covering uses in industry, consumer products and as flame retardants. The omnipresent benzotriazole class was detected in nearly all measurement months (32) with four different chemicals or TPs identified in this study. However, due to technique limitations (LC-HRMS) like insufficient separation, not all isomeric species can be correctly distinguished with the chromatographic method used, resulting in multiple possible identifications. Besides, two organophosphate flame retardants (OPFRs), namely tris(2-chloroethyl) phosphate (TCEP) and tris(1-chloro-2-propyl) phosphate (TCPP) were detected in 19 months, both known for environmental and toxic effects [47–49]. OPFRs serve as a substitution for brominated flame-retardants (BFRs) such as polybrominated diphenyl ethers (PBDEs), which have been found to cause adverse health effects in many samples recently [47, 50]. The three industrial chemicals 4-nitrophenol or 4-NP (22 months), dibutyl phthalate (DBP) and PFOA (both 14 months) were frequently detected in Luxembourgish rivers as well. 4-NP, used in many industrial applications, is known to have severe environmental and human health effects [51], just as DBP, a plasticizer with high aquatic toxicity [52] and PFOA, already listed for elimination in the Stockholm convention [53]. The chemicals in question were not previously subjected to monitoring by governmental institutions in Luxembourg. Although PFOA, other PFAS or DBP are not on the official monitoring list, these are being measured at AGE using specialized target methods and we recommend these efforts continue. 12 compounds found in this study are already included in the 2022 European watchlist [18],

e.g. fipronil, fluconazole or venlafaxine (see Additional file 2: Table S14). Further compounds, e.g. the identified PFAS are included in the 2022 proposal for a directive, amending the Water Framework Directive [19]. At this stage, the lists of tentatively annotated compounds (Table 1 and Additional file 2: Table S13) have been provided to AGE for further confirmation and quantification efforts (since targeted analysis is within their remit) to determine which compounds to include in future monitoring lists, due to their known environmental and health implications.

## Discussion

This 3-year investigation was conducted in Luxembourg to examine chemical pollutants in surface water, encompassing various sources and types of contaminants within a single medium, building on previous work. However, it is essential to also consider factors such as flow paths and meteorological phenomena, as they primarily influence the LC-HRMS peak intensities. This can lead to missing identifications when a sample is highly diluted or—during droughts—to a concentration of analyte, making it detectable at all. Other factors, such as sewer overflow due to high precipitation, lead to notably higher analyte signals. Studies show that the latter effect outweighs the dilution effects, *i.e.* rainfall could lead to increased detection of pollutants [54]. Lower precipitation was one of the reasons Singh et al. [11] cited for the decrease of pharmaceuticals in rivers, which was true for the years 2019 and 2020. This study shows the decreasing trend continuing for the following 16 months (2021 and 2022), although the average precipitation was higher in 2021 compared to prior years [55]. Consequently, it seems more likely now that the effect of the pandemic, as explained in Singh et al. [11], might be the reason for a decrease in the numbers of identified pharmaceuticals and river pollutants in general. Analytical causes were not suspected to be the cause here, as the performance of column, device and method were monitored and internal standard signals did not decrease in later analyses.

## Data analysis

The open source R package *patRoom* was used to perform the NT data analysis of the 3-year sampling of Luxembourgish surface waters with a tailor-made workflow (Fig. 4) designed for this purpose, as presented in the Data Analysis section in the Methods chapter. *PatRoom* offers the possibility to perform a componentization step, grouping related features in so called *components* based on different similarities, such as chromatographic behaviour. This step was intentionally omitted since it increased the data processing time, and several componentization algorithms lead to false associations. The current state

of componentization, particularly with the presence of numerous false positives, is acknowledged to be far from optimal. It is worth noting that the limitations of componentization predominantly stem from the algorithms available in *patRoan*, including popular ones like *CAMERA* [56], *RAMClustR* [57], and *cliqueMS* [58]. Although some tools demonstrate some improvement, most suffer from significant computational inefficiency and further developments seem necessary before these are applicable to environmental/exposomics analyses, which consider a broader range of elements than metabolomics. Moreover, *patRoan* includes the functionality to calculate chemical formulae for the feature groups, based on accurate mass and other data. Depending on the elements chosen in this step (default is C, H, N, O, P), processing time can be extremely long, and possible candidates can be excluded. As a consequence, this step was omitted since the likely presence of e.g. fluorinated compounds in the samples was already clear. Using the *GenForm* [59] algorithm can be a fast way to generate formulae. Nevertheless, when confronted with a large number of candidates, the algorithm's efficiency diminishes as the data size increases, while the use of exact mass to retrieve candidates instead of formula does not increase the number of candidates sufficiently to warrant this step. In order to address this, *patRoan* utilizes 'timeouts' to interrupt the formula generation process. Overall, the tentative identification of features and associated compounds was straightforward (as it was limited to candidates within *PubChemLite* [33, 34] and *MoNA* [36]), reducing the number of features, feature groups and MS peak lists to a minimal extent, demonstrated in the Results section. However, some features might be omitted or not identified based on experimental, algorithm or basic filtering conditions (which is always a risk). The efforts to look for (new) compounds of interest will be therefore continued, looking not only at chemicals included in the *PubChemLite for exposomics* database. However, extending the database always bears the risk of getting many candidates per mass, making data analysis even more difficult.

### Classification

In the classification steps of the tentatively identified chemicals, the widely employed tool *classyFire* was used to classify chemicals and obtain supplementary information. However, it is a purely structure-based application, giving just an overview performing a general grouping of chemicals in classes. This information is helpful in terms of establishing a structure in datasets, but does not provide much insight into chemical properties. Therefore, when evaluating *classyFire* compound classifications, it is important to acknowledge that the results may not encompass all information regarding the chemicals

under study. Certain compounds may belong to multiple classes, which could potentially result in incomplete classification. Moreover, there are multiple uses per chemical class and a distinction can only be made looking at the individual compound. A way to figure out possible sources of chemicals was the use of *PubChem* metadata for all tentatively identified chemicals and looking at temporal trends (see Fig. 8). Four categories from the *PubChemLite* database, namely *agroChemInfo*, *drug-MedicInfo*, *disorderDisease* and *knownUse* were examined to get an estimate of the count of agrochemicals, drugs, compounds associated with disorders and diseases and of those for which commercial use is known. Other categories like the *ToxicityInfo* were not considered, as they are less specific (e.g. it indicates whether information is available, but the availability of information alone does not say whether it is toxic or not). Several pharmaceuticals are associated with disorders and disease as the annotation also includes treatments. Again, the environmental effect of the tentatively detected exposome-related compounds can only be fully accessed looking at the individual case considering their concentrations. It is worrying that a majority of compounds found in the Luxembourgish environment were of pharmaceutical origin because these compounds are designed to have biological effects, which means they may have unintended effects on non-target organisms and ecosystems. Even at low concentrations, pharmaceutical compounds as venlafaxine [46] can accumulate in organisms and can have toxic effects, especially over time. They can also promote antibiotic resistance, alter gene expression, and disrupt endocrine systems (for more information in the Luxembourgish context, see Singh et al. [11]). Additionally, some pharmaceuticals are known to persist in the environment for long periods of time and can travel long distances, leading to contamination of remote areas and cross-border contamination (Germany, France, Belgium). The presence of pharmaceuticals in Luxembourgish rivers highlights the potential to enhance wastewater treatment and disposal systems, thereby safeguarding public health. Luxembourg's biological and mechanical WWTPs are not always designed to remove pharmaceuticals from wastewater, and some compounds are not effectively degraded by current treatment processes [60–62]. As a result, these compounds can end up in surface waters (as shown here), groundwater, and even drinking water sources.

### Comparison to other studies

Besides focussing on the temporal variations of chemicals and their classification, it can be highly valuable to include geographical information, especially to analyse possible origins of chemicals (and thus possible origins

of the potential threat). Therefore, the measurement results of one sampling point (Chiers) over 3 years were compared to the results of the inlet of a WWTP located ~6 km upstream. As shown in the Results section, an overlap of nearly 200 chemicals was identified and a majority of compounds identified in this study was found before in the WWTP (see Fig. 10). This is either a result of incomplete treatment in the WWTP or chemical input between the WWTP and the later sampling point. A comparison between inlet and outlet of the WWTP could clarify this. However, the identifications in both studies are of tentative nature: in this study, the aim was to use NTA to propose chemicals for further target verification at AGE (beyond the scope of this article). The verification efforts in the other study are still ongoing (likewise beyond the scope of this article, as different partners are involved).

Moreover, a comparison to former published studies on Luxembourgish surface water was performed, looking at pesticides [10] and pharmaceuticals [11]. This was done to perform a plausibility check of the results (chemical identifications and temporal trends), as both compared studies involved a target approach after screening for suspects. The list of 378 tentatively identified chemicals was then compared to the monitoring list used by the water administration in Luxembourg, showing not only the overlap of findings, but also nominating potential chemicals for future inclusion to this list (see Table 1). Several of the annotated chemicals, e.g. venlafaxine are already listed in the 2022 EU watchlist [18] or the 2022 proposal to revise the priority substances [19]. This shows that there is already work in progress to regulate these chemicals in the EU.

Monitoring and regulations are necessary steps to track contaminant distribution, identify potential sources and restrict or ban their use to stop their release to the environment. Adding NTA to track regularly potential water contaminants can help with this, as this study shows. The NTA workflow established via *patRoom* could be used and improved further by AGE, as it is also compatible with the instrument type used in their laboratories.

Future steps should involve the target analysis of the tentatively identified chemicals, with a special focus on emerging pollutants and persistent chemicals (Table 1). Giving a list of those substances to governmental institutions like AGE could improve future monitoring efforts. A quantification of these compounds, using reference standards could then help to estimate their environmental threat. Overall, the impact of these compounds on the environment and human health underscores the importance of monitoring and regulating their use and disposal to minimize their potential impact.

## Conclusions

Based on the results presented in the article, the analysis of the 3-year sampling of Luxembourgish surface waters using the *patRoom* R package tentatively identified 378 chemicals associated with the exposome. 40 chemicals were already included in routine water analysis performed by AGE. The results were plausibility checked using former studies by looking at overlapping identifications and general trends, with many of the tentative identifications here matching the confirmed identifications in the previous studies. The identified chemicals were classified to get a general overview, and the results showed that benzenoids, organoheterocyclic compounds, and organic phosphoric acids and derivatives were among the most identified classes (*classyFire*). A temporal analysis between the 3 years and in 1 year was shown, looking at the classifications obtained not only by *classyFire*, but also by the *PubChemLite* categories on agrochemicals, pharmaceuticals, drugs and diseases and known uses. Most of the chemicals detected had a known use, were classified as pharmaceuticals and are associated with disorders and diseases. A decrease in identifications was observed in 2021 and 2022. Trends in the use of agrochemicals could be seen in the monthly comparison, with May to July showing the highest number of identifications. The study examined not only temporal, but also geographical variations of chemicals to analyse their possible origins and potential threats. Chemical measurements from one sampling point were compared to those from an inlet of a WWTP located 6 km away. Overlapping and differing chemicals were identified, and the study found that pharmaceuticals dominated again the identifications. Some persistent synthetic chemicals were also found on both sampling sites. Further investigations will be done, looking at the concentrations and sources of these pollutants. A list of 41 chemicals—not yet included in the AGE monitoring—is presented with some chemicals already highlighted in former studies, e.g. venlafaxine and its TP desvenlafaxine. The study suggests adding more TP compounds to the AGE monitoring list, as many parent compounds are not detected, but rather their (active) TPs. Moreover, industrial chemicals like 4-NP and flame retardants like OPFRs were detected in a majority of the analysed months and should be considered to be added after confirmation by target efforts. The aim of including NTA lists like these to governmental monitoring is to eliminate harmful chemicals from the environment, not only by searching for their sources, but also by implementing (WWTP) treatment technologies to reduce their presence in the environment. Biological treatments and monitoring regulations are also necessary in Luxembourg to track contaminant distribution, identify potential sources, and restrict or ban their use. All in

all, this study of surface waters has shown how NTA can be used to complement routine (target) monitoring programmes and help to broaden the focus to new emerging chemicals. Future efforts will involve extending the target monitoring and working on implementing this open source workflow into AGE monitoring routines directly.

#### Abbreviations

4-NP	4-Nitrophenol
AGE	L'Administration de la gestion de l'eau
BFR	Brominated flame-retardant
CID	Chemical Identifier
DBP	Dibutyl phthalate
DDD	Dichlorodiphenyldichloroethane
DDE	Dichlorodiphenyldichloroethylene
DDT	Dichlorodiphenyltrichloroethane
DMSO	Dimethyl sulfoxide
DNOC	4,6-Dinitro-o-kresol
DoE	Design of Experiment
ECl	Environmental Cheminformatics Group
HRMS	High-resolution mass spectrometry
LC	Liquid chromatography
MoNA	MassBank of North America
NTA	Non-target analysis
NT	Non-target
OPFRs	Organophosphate flame retardants
PBDE	Polybrominated diphenyl ether
PFAS	Per- and poly-fluorinated compounds
PFBS	Perfluorobutanesulfonic acid
PFHpA	Perfluoroheptanoic acid
PFNA	Perfluorononanoic acid
PFOA	Perfluorooctanoic acid
PFOS	Perfluorooctanesulfonic acid
TCP	Tris(1-chloro-2-propyl) phosphate
TCEP	Tris(2-chloroethyl) phosphate
WWTP	Wastewater treatment plant

#### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12302-023-00805-5>.

**Additional file 1: Figure S1.** Design of Experiments (DoEs) of feature optimization workflow, with DoE3 giving best results for the April 2020 analyses. The iterative process involved the testing of parameters in three DoEs; the best slices for each DoE were visualized using perspective plots.

**Figure S2.** Results of feature optimization workflow, with the best results in positive and negative mode for the April 2020 analyses.

**Additional file 2: Table S1.** Sampling sites with weblinks. **Table S2.** Feature finding optimization results. **Table S3.** Number of 2, 3a and 3b identifications. **Table S4.** Number of positive and negative identifications. **Table S5.** Tentatively identified chemicals per month. **Table S6.** ClassyFire results. **Table S7.** PubChemLite classifications. **Table S8.** AGE monitoring comparison. **Table S9.** Krier et al. (pesticides) comparison. **Table S10.** Singh et al. (pharmaceuticals) comparison. **Table S11.** Chiers WWTP Petange comparison (SMILES). **Table S12.** Chiers results (SMILES) all months. **Table S13.** List of chemicals with high occurrence in the 34 months analysed, and not currently monitored by AGE. **Table S14.** European watchlist 2022 comparison.

#### Acknowledgements

The authors gratefully acknowledge the contribution of Parviel Chirsir for the 2022 sampling data results and discussions with others in the LuxTIME and ECl teams.

#### Author contributions

DA and ELS did the conceptualization of the study. RH contributed the changes to the software and DA developed the methodology. The formal analysis, main investigation, visualization and writing of the original draft were conducted by DA. PD provided the resources (water samples) and did parts of the investigation. All authors were responsible for review and editing. ELS was supervising and performed funding acquisition and resources.

#### Funding

DA acknowledges funding support from the University of Luxembourg through its Institute for Advanced Studies for the project LuxTIME (Audacity-2019); ELS acknowledges funding support from the Luxembourg National Research Fund (FNR) for project A18/BM/12341006.

#### Availability of data and materials

All data, code, figures and supplementary files can be found in the 'data\_luxwater\_nt\_paper\_da' repository of the ECl group in GitLab, via [https://gitlab.lcsb.unil.lu/eci/data\\_luxwater\\_nt\\_paper\\_da](https://gitlab.lcsb.unil.lu/eci/data_luxwater_nt_paper_da) under license Artistic 2.0. The required sample identifiers to link the measurement to location and date can be found in this GitLab file. The measurement files in mzML format can be found as dataset MSV000092221 from the [GNPS MassIVE repository] (<https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp>), accessible via <ftp://massive.ucsd.edu/MSV000092221/> and to be cited with <https://doi.org/10.25345/C55X25P62>. All data are also accessible via <https://doi.org/10.17881/5dk0-jv81>.

#### Declarations

##### Ethics approval and consent to participate

Not applicable.

##### Consent for publication

Not applicable.

##### Competing interests

The authors declare that they have no competing interests.

#### Author details

<sup>1</sup>Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 Avenue du Swing, 4367 Belvaux, Luxembourg. <sup>2</sup>Administration de La Gestion de L'eau (AGE), Avenue du Rock'n'Roll 1, 4361 Esch-Sur-Alzette, Luxembourg. <sup>3</sup>Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands.

Received: 3 July 2023 Accepted: 29 October 2023

Published online: 07 November 2023

#### References

1. Arp HPH, Aurich D, Schymanski EL et al (2023) Avoiding the next silent spring: our chemical past, present, and future. *Environ Sci Technol* 57:6355–6359. <https://doi.org/10.1021/acs.est.3c01735>
2. Agency for Toxic Substances and Disease Registry, U.S. Department of Health and Human Services (2022) Toxicological Profile for DDT, DDE, and DDD. In: <https://www.atsdr.cdc.gov/toxprofiles/tp35.pdf>. Accessed 23 May 2023
3. Cocco P, Kazerouni N, Zahm SH (2000) Cancer mortality and environmental exposure to DDE in the United States. *Environ Health Perspect* 108:1–4. <https://doi.org/10.1289/ehp.001081>
4. Schymanski EL, Williams AJ (2017) Open science for identifying "Known Unknown" chemicals. *Environ Sci Technol* 51:5357–5359. <https://doi.org/10.1021/acs.est.7b01908>
5. Little JL, Clevon CD, Brown SD (2011) Identification of "Known Unknowns" utilizing accurate mass data and chemical abstracts service databases. *J Am Soc Mass Spectrom* 22:348–359. <https://doi.org/10.1007/s13361-010-0034-3>
6. Wang X, Shen Z, Zeng Y et al (2018) Day-night differences, seasonal variations and source apportionment of PM10-Bound PAHs over Xi'an, Northwest China. *Atmosphere* 9:62. <https://doi.org/10.3390/atmos9020062>

7. Stamatis N, Hela D, Triantafyllidis V, Konstantinou I (2013) Spatiotemporal variation and risk assessment of pesticides in water of the lower catchment basin of Acheloos River, Western Greece. *Sci World J*. <https://doi.org/10.1155/2013/231610>
8. Aurich D, Miles O, Schymanski EL (2021) Historical exposomics and high resolution mass spectrometry. *Exposome* 1:1–15. <https://doi.org/10.1093/exposome/osab007>
9. Lickes J-P, L'Administration de la gestion de l'eau (AGE), Luxembourg (2022) Vorstellung des 3. Bewirtschaftungsplans. In: <https://gouvernement.ludam-Assetsdocumentsactualites202209-Sept.-Gest.-Dist.-Hydrogr.-3-Wasserbewirtschaftungsplan.pdf>. <https://gouvernement.lu/dam-assets/documents/actualites/2022/09-septembre/22-gestion-districts-hydrographiques/vorstellung-des-3-wasserbewirtschaftungsplan.pdf>. Accessed 26 May 2023
10. Krier J, Singh RR, Kondić T et al (2022) Discovering pesticides and their TPs in Luxembourg waters using open cheminformatics approaches. *Environ Int* 158:14. <https://doi.org/10.1016/j.envint.2021.106885>
11. Singh RR, Lai A, Krier J et al (2021) Occurrence and distribution of pharmaceuticals and their transformation products in Luxembourgish surface waters. *ACS Environ Au* 1:58–70. <https://doi.org/10.1021/acsenvironau.1c00008>
12. L'Administration de la gestion de l'eau (AGE), Luxembourg (2022) Anhang 14-Überschreitungen der UQN für prioritäre und flussgebietspezifische Stoffe. In: <https://eau.gouvernement.lu/dam-assets/administration/documents/3-cycle/elaboration-du-3e-plan-de-gestion-document-final/anhang-14-uberschreitungen-der-uqn-fur-prioritare-und-flussgebietspezifische-stoffe.pdf>. Accessed 23 May 2023
13. L'Administration de la gestion de l'eau (AGE), Luxembourg (2022) Elaboration du 3e plan de gestion. In: [http://eau.gouvernement.lu/fr/administration/directives/Directive-cadre-sur-leau/3e-cycle-\(2021-2027\)/elaboration-du-3e-plan-de-gestion-document-final.html](http://eau.gouvernement.lu/fr/administration/directives/Directive-cadre-sur-leau/3e-cycle-(2021-2027)/elaboration-du-3e-plan-de-gestion-document-final.html). Accessed 23 May 2023
14. European-Parliament, Council of the European Union (2000) Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy
15. European-Parliament, Council of the European Union (2008) Directive 2008/105/EC of the European Parliament and of the Council of 16 December 2008 on environmental quality standards in the field of water policy, amending and subsequently repealing Council Directives 82/176/EEC, 83/513/EEC, 84/156/EEC, 84/491/EEC, 86/280/EEC and amending Directive 2000/60/EC of the European Parliament and of the Council
16. Directorate-General for Environment (2023) Surface water. In: [Eur. Com. - Energy Clim. Change Environ. https://environment.ec.europa.eu/topics/water/surface-water\\_en](https://environment.ec.europa.eu/topics/water/surface-water_en). Accessed 11 Oct 2023
17. European-Parliament, Council of the European Union (2013) Directive 2013/39/EU of the European Parliament and of the Council of 12 August 2013 amending Directives 2000/60/EC and 2008/105/EC as regards priority substances in the field of water policy Text with EEA relevance
18. European-Parliament, Council of the European Union (2022) Commission Implementing Decision (EU) 2022/1307 of 22 July 2022 establishing a watch list of substances for Union-wide monitoring in the field of water policy pursuant to Directive 2008/105/EC of the European Parliament and of the Council (notified under document C(2022) 5098) (Text with EEA relevance)
19. Directorate-General for Environment (2022) Proposal amending Water Directives. In: [Eur. Com. - Energy Clim. Change Environ. https://environment.ec.europa.eu/publications/proposal-amending-water-directives\\_en](https://environment.ec.europa.eu/publications/proposal-amending-water-directives_en). Accessed 12 Oct 2023
20. Blum KM, Andersson PL, Renman G et al (2017) Non-target screening and prioritization of potentially persistent, bioaccumulating and toxic domestic wastewater contaminants and their removal in on-site and large-scale sewage treatment plants. *Sci Total Environ* 575:265–275. <https://doi.org/10.1016/j.scitotenv.2016.09.135>
21. Wang X, Yu N, Qian Y et al (2020) Non-target and suspect screening of per- and polyfluoroalkyl substances in Chinese municipal wastewater treatment plants. *Water Res* 183:12. <https://doi.org/10.1016/j.watres.2020.115989>
22. Suman T-Y, Kim S-Y, Yeom D-H, Jeon J (2022) Transformation products of emerging pollutants explored using non-target screening: perspective in the transformation pathway and toxicity mechanism—a review. *Toxics* 10:22. <https://doi.org/10.3390/toxics10020054>
23. Helmus R, van de Velde B, Brunner AM et al (2022) patRoom 2.0: improved non-target analysis workflows including automated transformation product screening. *J Open Source Softw*. <https://doi.org/10.21105/joss.04029>
24. Helmus R, ter Laak TL, van Wezel AP et al (2021) patRoom: open source software platform for environmental mass spectrometry based non-target screening. *J Cheminformatics*. <https://doi.org/10.1186/s13321-020-00477-w>
25. Smith CA, Want EJ, O'Maille G et al (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem* 78:779–787. <https://doi.org/10.1021/ac051437y>
26. Tautenhahn R, Böttcher C, Neumann S (2008) Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinform*. <https://doi.org/10.1186/1471-2105-9-504>
27. Benton HP, Want EJ, Ebbels TMD (2010) Correction of mass calibration gaps in liquid chromatography–mass spectrometry metabolomics data. *Bioinformatics* 26:2488–2489. <https://doi.org/10.1093/bioinformatics/btq441>
28. Ruttikes C, Schymanski EL, Wolf S et al (2016) MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform*. <https://doi.org/10.1186/s13321-016-0115-9>
29. Libiseller G, Dvorzak M, Kleb U et al (2015) IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics* 16:118. <https://doi.org/10.1186/s12859-015-0562-8>
30. Albóniga OE, González O, Alonso RM et al (2020) Optimization of XCMS parameters for LC–MS metabolomics: an assessment of automated versus manual tuning and its effect on the final results. *Metabolomics* 16:14. <https://doi.org/10.1007/s11306-020-1636-9>
31. Tostengard AR, Smith R (2021) A review and evaluation of techniques for improved feature detection in mass spectrometry data. *Grad Stud Theses Diss Prof Pap Univ Mont* <https://scholarworks.umt.edu/etd/11679-41>. Accessed 23 March 2023
32. Kim S, Chen J, Cheng T et al (2023) PubChem 2023 update. *Nucleic Acids Res* 51:D1373–D1380. <https://doi.org/10.1093/nar/gkac956>
33. Bolton E, Schymanski E, Kondić T, et al (2023) PubChemLite for Exposomics. <https://doi.org/10.5281/zenodo.7576412>
34. Schymanski EL, Kondić T, Neumann S et al (2021) Empowering large chemical knowledge bases for exposomics: PubChemLite meets MetFrag. *J Cheminformatics* 13:19. <https://doi.org/10.1186/s13321-021-00489-0>
35. Schymanski EL, Jeon J, Gulde R et al (2014) Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Env Sci Technol* 48:2097–2098. <https://doi.org/10.1021/es5002105>
36. Fiehnlab (2018) MassBank of North America (MoNA). In: <https://mona.fiehnlab.ucdavis.edu/>. Accessed 20 Jun 2023
37. Djoumbou Feunang Y, Eisner R, Knox C et al (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J Cheminformatics* 8:61. <https://doi.org/10.1186/s13321-016-0174-y>
38. NCBI (2023) PubChem Classification Browser TOC. In: <https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72>. Accessed 26 Jun 2023
39. Kessner D, Chambers M, Burke R et al (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 24:2534–2536. <https://doi.org/10.1093/bioinformatics/btn323>
40. Chambers MC, Maclean B, Burke R et al (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* 30:918–920. <https://doi.org/10.1038/nbt.2377>
41. Martens L, Chambers M, Sturm M et al (2011) mzML – a community standard for mass spectrometry data. *Mol Cell Proteomics* 10(R110):000133. <https://doi.org/10.1074/mcp.R110.000133>
42. Pedrioli PGA, Eng JK, Hubley R et al (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* 22:1459–1466. <https://doi.org/10.1038/nbt1031>

43. Keller A, Eng J, Zhang N et al (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* 1:8. <https://doi.org/10.1038/msb4100024>
44. Talavera Andújar B, Aurich D, Aho VTE et al (2022) Studying the Parkinson's disease metabolome and exposome in biological samples through different analytical and cheminformatics approaches: a pilot study. *Anal Bioanal Chem* 414:7399–7419. <https://doi.org/10.1007/s00216-022-04207-z>
45. Awan M, Buriak I, Fleck R et al (2020) Dimethyl sulfoxide: a central player since the dawn of cryobiology, is efficacy balanced by toxicity? *Regen Med* 15:1463–1491. <https://doi.org/10.2217/rme-2019-0145>
46. Best C, Melnyk-Lamont N, Gesto M, Vijayan MM (2014) Environmental levels of the antidepressant venlafaxine impact the metabolic capacity of rainbow trout. *Aquat Toxicol Amst Neth* 155:190–198. <https://doi.org/10.1016/j.aquatox.2014.06.014>
47. Maddela NR, Venkateswarlu K, Megharaj M (2020) Tris(2-chloroethyl) phosphate, a pervasive flame retardant: critical perspective on its emissions into the environment and human toxicity. *Environ Sci Process Impacts* 22:1809–1827. <https://doi.org/10.1039/D0EM00222D>
48. Wu H, Zhong M, Lu Z et al (2018) Biological effects of tris (1-chloro-2-propyl) phosphate (TCPP) on immunity in mussel *Mytilus galloprovincialis*. *Environ Toxicol Pharmacol* 61:102–106. <https://doi.org/10.1016/j.etap.2018.05.022>
49. Ji C, Lu Z, Xu L et al (2020) Global responses to tris(1-chloro-2-propyl) phosphate (TCPP) in rockfish *Sebastes schlegeli* using integrated proteomic and metabolomic approach. *Sci Total Environ*. <https://doi.org/10.1016/j.scitotenv.2020.138307>
50. Zhang Z-N, Yang D-L, Liu H et al (2023) Effects of TCPP and TCEP exposure on human corneal epithelial cells: oxidative damage, cell cycle arrest, and pyroptosis. *Chemosphere*. <https://doi.org/10.1016/j.chemosphere.2023.138817>
51. Naushad Mu, Ahamad T, Rizwan Khan M (2022) Remediation of wastewater containing 4-nitrophenol using ionic liquid stabilized nanoparticles: Synthesis, characterizations and applications. *Chemosphere*. <https://doi.org/10.1016/j.chemosphere.2022.135173>
52. Lin W, He Y, Li R et al (2023) Adaptive changes of swimming crab (*Portunus trituberculatus*) associated bacteria helping host against dibutyl phthalate toxification. *Environ Pollut*. <https://doi.org/10.1016/j.envpol.2023.121328>
53. Fiedler H, Kennedy T, Henry BJ (2021) A critical review of a recommended analytical and classification approach for organic fluorinated compounds with an emphasis on Per- and polyfluoroalkyl substances. *Integr Environ Assess Manag* 17:331–351. <https://doi.org/10.1002/ieam.4352>
54. Zhu L, Jiang C, Panthi S et al (2021) Impact of high precipitation and temperature events on the distribution of emerging contaminants in surface water in the Mid-Atlantic, United States. *Sci Total Environ*. <https://doi.org/10.1016/j.scitotenv.2020.142552>
55. statista (2023) Luxembourg: annual rainfall in Luxembourg-City 2021. In: <https://www.statista.com/statistics/584864/annual-rainfall-in-luxembourg-city/>. Accessed 1 Jun 2023
56. Kuhl C, Tautenhahn R, Böttcher C et al (2012) CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem* 84:283–289. <https://doi.org/10.1021/ac202450g>
57. Broeckling CD, Afsar FA, Neumann S et al (2014) RAMClust: a novel feature clustering method enables spectral-matching-based annotation for metabolomics data. *Anal Chem* 86:6812–6817. <https://doi.org/10.1021/ac501530d>
58. Senan O, Aguilar-Mogas A, Navarro M et al (2019) CliqueMS: a computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network. *Bioinformatics* 35:4089–4097. <https://doi.org/10.1093/bioinformatics/btz207>
59. Meringer M, Reinker S, Zhang J, Muller A (2011) MS/MS data improves automated determination of molecular formulas by mass spectrometry. *Match Commun Math Comput Chem* 65:259–290
60. Venditti S, Arenz-Leufen M, Köhler C, et al Treatment of pharmaceutical wastewater by O<sub>3</sub> and O<sub>3</sub>/H<sub>2</sub>O<sub>2</sub> processes: a pilot scale study in Luxembourg. <https://www.researchgate.net/publication/260697444>. Accessed 19 October 2023
61. Krein A, Pailler J-Y, Guignard C et al (2012) Determination of estrogen activity in river waters and wastewater in Luxembourg by chemical analysis and the yeast estrogen screen assay. *Environ Pollut* 1:p86. <https://doi.org/10.5539/ep.v1n2p86>
62. Pailler J-Y, Krein A, Pfister L et al (2009) Solid phase extraction coupled to liquid chromatography-tandem mass spectrometry analysis of sulfonamides, tetracyclines, analgesics and hormones in surface water and wastewater in Luxembourg. *Sci Total Environ* 407:4736–4743. <https://doi.org/10.1016/j.scitotenv.2009.04.042>

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)