UNIVERSITEIT VAN AMSTERDAM

VU VRIJE UNIVERSITEIT AMSTERDAM

MSc Chemistry

Track Analytical Sciences

Master Thesis

Semi-automated process to build a suspect list with water contaminants and their transformation products

---

by

Peli Angouraki

2773807

April 2024 60 ECs

May 2023 – March 2024

Daily Supervisor:
Nienke Meekel MSc

Examiner:
Frederic M. Béen Dr.

Second Examiner:
Marja H. Lamoree Dr.

KWR

Chemical Water Quality & Health

KWR Water Research Institute

Abstract

Various chemical compounds find their way into wastewater, such as pharmaceuticals, pesticides, and industrial chemicals. After conventional wastewater treatment, many chemical compounds do not degrade completely. On top of that, they transform to other compounds, often called as transformation products, for which little is known.

The goal of this research project was to build a workflow for an automated creation of a suspect list with chemical pollutants and their transformation products, which will support their identification in wastewater effluent samples. This workflow, utilizing a predetermined list of known chemicals, generates a suspect list with transformation products, both from the literature and prediction models. The prediction of transformation products was made with a model for environmental microbial metabolites. The compounds were prioritized based on their mobility in water and detectability with LC-ESI-MS. Starting with a list of 5,000 known chemical pollutants as an input, two suspect lists were created, each with approximately 1,400 compounds, one for positive and one for negative ionization mode in mass spectrometry. This work can set an example for the creation of other workflows, with focus on transformation products created by different transformation mechanisms and/or application of different prioritization criteria.

Layman summary

In modern society, a vast number of chemical products is used; pharmaceutical compounds in healthcare and pesticides in agriculture, given as broad examples. Municipal wastewater is collected from residential areas, and sometimes from industries and agriculture, containing a plethora of chemical compounds. The treatment of municipal wastewater is focused on the removal of organic material and microorganisms, while chemical compounds are not completely eliminated. On top of that, the initial compounds transform into other compounds (transformation products or TPs), for which very little is known.

To have a better understanding of the underlying danger of that are released in the environment, it is important to develop methods and workflows to identify them. An analytical technique, commonly used in water samples is Liquid Chromatography – ElectroSpray Ionization connected to Mass Spectrometry (LC-ESI-MS). With suspect screening, the data resulting from the analysis, is screened with a focus to identify specific compounds (suspect list), that are expected to be present in the sample. Suspect screening can be applied to identify unknown TPs, with the use of prediction models that predict their chemical structure based on possible transformation reactions.

During this internship project (May 2023 – March 2024), a (semi-)automated workflow was created, which for a given list of known chemicals as an input, returns a suspect list with the known chemicals, and their reported and predicted TPs, that are expected to be found in wastewater effluent. The compounds were prioritized based on their predicted affinity with water, and their amenability with LC-ESI-MS. The workflow was built and tested an initial dataset containing more than 5,000 chemical entities. At the end of the process, two suspect lists were created, one for each mass spectrometry ionization mode, with around ~1,400 unique structures. Extra care was put in the validation of the chemical compounds before each step of the workflow, to avoid the presence of irrelevant or erroneous chemical structures in the final suspect lists.

This work can be the base for the creation of other suspect list workflows, with access to different TPs prediction models, and using other prioritization criteria, such as biodegradability or/and toxicity.

Keywords: transformation products, wastewater, wastewater effluent, suspect screening, non-target screening, *in silico* tools, QSAR models

List of abbreviations

TP(s)	Transformation Product(s)
PMT compounds	Persistent, Mobile, and Toxic compounds
WWTP(s)	Wastewater Treatment Plant(s)
NTA	Non-Target Analysis
LC-ESI-MS	Liquid Chromatography - Electrospray Ionization - Mass Spectrometry
AOPs	Advanced Oxidation Processes
HRMS	High-Resolution Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
QSPR/QSAR	Quantitative Structure-Property Relationship/ Quantitative Structure-Activity Relationship
AD	Applicability Domain
MLR	Multiple Linear Regression
kNN	k-nearest neighbor
ML	Machine Learning
KB	Knowledge - Based
CAS	Chemical Abstract Service
SMILES	Simplified Molecular Input Line Entry System
InChI	IUPAC International Chemical Identifier
EAWAG-PPS	Swiss Federal Institute of Aquatic Science and Technology - Pathway Prediction System
PFAS	
UM-BBD	University of Minnesota Biocatalysis/ Biodegradation Database
CTS	Chemical Transformation Simulator
MCI	Molecular Connectivity Index
LogP	Logarithmic Octanol - Water Partition Coefficient
LogKoc	Logarithmic Organic Carbon - Water normalized Sorption Coefficient
pKa	Acid Dissociation Constant
CID	PubChem Compound Identification Number
logD	Logarithmic Octanol - Water Distribution Coefficient

Reading guide of the thesis

Chapter 1 explains what transformation products are and how they form with a focus on wastewater treatment. A short introduction is given on the analysis and identification of transformation products with non-target and suspect screening.

Chapter 2 provides fundamental information on chemical identifiers used to describe chemical compounds, different types of prediction models to predict chemical properties and *in silico* tools to predict transformation products.

In Chapter 3 are presented all the tools that were tested for the workflow along with the tools were included in the workflow.

In Chapter 4 are presented the workflow steps and the observed results. There are also discussed the challenges that occurred while building the workflow and how they were settled. More information about the workflow can be found in the attached Rmarkdown file.

In Chapter 5 are summarized the most important findings of the research project.

In Chapter 6 are listed the ways the workflow can be improved.

Table of Contents

1. Introduction

1.1. Micropollutants and transformation products (TPs) in wastewater

The use of existing, and the introduction of new chemical products, such as pesticides, pharmaceuticals and industrial chemicals is important to deal with modern challenges such as low productivity in agriculture [1], combat diseases [2] and help in industrial processes [3]. It is already known that chemical pollution has negative effects in the environment and on human [4–9]. However, after pollutants are released in the environment, they may transform into other chemical species (transformation products or TPs). A chemical compound might undergo abiotic transformations such as hydrolysis, photodegradation, or biotic transformations where it is being metabolized by microorganisms or larger living organisms. The kinetics of a transformation reaction might be affected by the environmental conditions and the presence of other chemical species [10,11]. Exploring and discovering all the different TPs that might be formed is a difficult task, since we do not know all the reaction mechanisms taking place or the new chemical structures thar occur [12].

The formed TPs usually are more polar, hence less bioaccumulative, with shorter lifetimes [13]. However, there are exceptions noted in the literature, where TPs have the same potency [14] or are even more toxic than the parent compound [11]. This phenomenon is especially noticed in pesticides which are purposely sold in their pro-active form, and when consumed by the undesired organism they are metabolized into their original potent form. If the active group has remained intact while the mobility of the TP is increased, thus the compound is absorbed by the sediments to a lesser extent, it might result in higher exposure to the living organisms, leading to greater risk [11]. An example is the microbial degradation of the herbicide 2,4-dichlorophenoxyacetic acid (2,4-D) to the more toxic 2,4-dichlorophenol after the degradation of the side chain [15] (Figure 1).

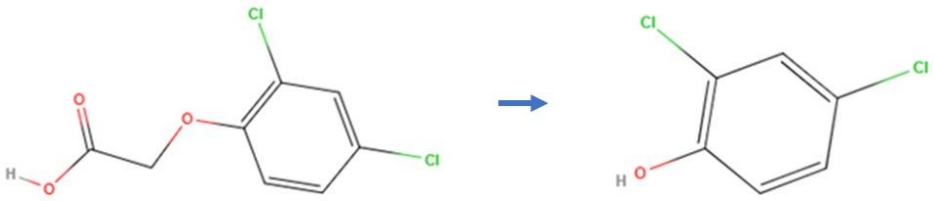


Figure 1: Microbial degradation of 2,4-D with hydrolysis, to a more potent transformation product. (structural depiction source: Molview [16])

A group of compounds defined by their high Mobility, Persistence and Toxicity (PMT) have gained interest in water quality research and European legislation [17]. This is explained by, their high mobility, -they are likely to remain in water and not be absorbed in sediment [18], and because of their persistence they tend to accumulate over time in water environments, even if their emissions are low [19]. Most of the TPs remain undetected, with their structure and properties unknown. However, a significant portion of the identified TPs are being categorized as probable PMTs [20], a fact that stresses the importance of research on TPs.

Treated wastewater is a major source of TPs pollution in the surface waters [12]. Municipal wastewater treatment plants receive influents from households, facilities such as hospitals, schools or office buildings, and in some cases, from industries. Individuals' use of pharmaceutical results in excretion of the active therapeutic substances along with their metabolism TPs [21]. In addition, various

industrial activities result in the release of various compounds in the wastewater as for example: halogenated organic compounds, such as dioxins and furans, pesticides, and pharmaceuticals [22].

The major purpose of wastewater treatment is to remove organic matter with high loads of microorganisms before discharging the effluent in the environment [23]. However, current monitoring studies in Europe show that wastewater treatment facilities do not remove chemical pollutants efficiently [24–27]. On top of that, even if a new treatment has high removal rates on known chemical pollutants (parent compounds), the removal rate of their TPs is rarely tested [28]. As a result, it is expected that traces of parents and TPs will be present in wastewater effluent during its release in freshwater [29,30]. According to the literature, parent compounds and TPs have been detected in surface water [31,32]. For example, carbamazepine-10,11-epoxide, and its parent carbamazepine have been detected in wastewater effluent [29] and in surface water where treated wastewater is being released [33].

1.2. Transportation, wastewater treatment in the Netherlands and fate of pollutants

Wastewater from households, healthcare or educational institutes and various industries is collected and transferred to wastewater treatment plants (WWTPs). The major components of wastewater are microorganisms and organic material, with metals and micropollutants such as cosmetics, product additives, medicines and pesticides having a lower share [34]. The pollutants might transform through human and microbial metabolism, or abiotic processes as for example photolysis and hydrolysis before reaching the WWTP [10].

Wastewater undergoes different treatment stages, as illustrated in Figure 2. During the primary treatment suspended solids of big size, that might cause blockage in the system, are removed [23]. After this step, the wastewater is transferred to an aerated tank with activated sludge, rich in bacteria, which break down the greatest part of organic matter [23]. Biodegradable organic material dissolved or undissolved is metabolized by the bacteria and is assimilated in the microorganisms' mass [34]. Bacteria release many different enzymes to metabolize the organic matter [35]. Parents and TPs transform to other compounds, in the presence of various enzymes responsible for hydrolysis, oxidation, conjugation reactions and more [36]. The transformation mechanisms taking place during biological treatment can be even more complex if we consider the diverse nature of wastewater, with the different pollutants and small molecules interacting with each other and with the bacterial enzymes [37,38]. Pollutants either break down in the presence of enzymes into carbon dioxide and minerals, get absorbed and immobilized into the sludge, or partially decompose into water-soluble products [34,39]. Other removal mechanisms for pollutants, but with a less important role, are volatilization from the water phase, abiotic hydrolysis and photolysis [40]. After the end of the secondary treatment, highly mobile transformation products and parents that are less likely to be absorbed by the sludge, remain in the water phase.

During the degradation of organic material, nitrogen and phosphorous are mineralized and precipitate with the sludge at the bottom of the tank. After this treatment, the wastewater has a reduced load of organics, agglomerates, and microorganisms, while some micropollutants, TPs included, are still dissolved in the liquid phase, and released in the environment [34]. It has been shown that the use of advanced oxidation processes (AOPs) such as ozonation or UV radiation decrease the concentration of known parent compounds [39]. However, there is no definite answer if the toxic effect of the effluent will decrease since the formed TPs might have greater toxic effect. These AOPs are not yet used in large scale in WWTPs in the Netherlands [41].

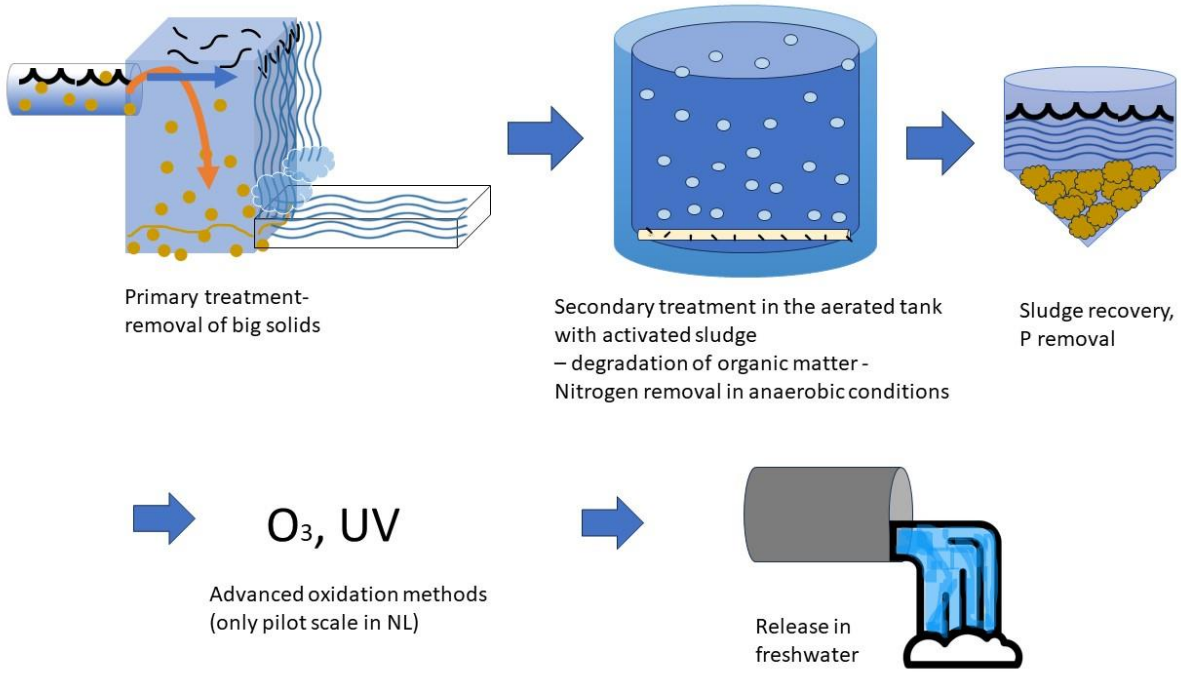


Figure 2: Schematic depiction of wastewater treatment steps.

1.3. Analysis and identification of TPs

The awareness of the dangers of TPs has resulted in a recent shift in European policy. According to the European Directive 1107/2009 [42], measuring techniques should be available to detect the levels of TPs from pesticides, when there is ecotoxicological or drinking water concern. The Urban Wastewater Treatment Directive was revised in October 2023 [43,44], according to which, special treatment should be applied to remove "substances of concern" along with their TPs. The need to adhere the current policies motivates the development of detection methods for unknown TPs.

For the analysis of TPs in water samples, a separation method (e.g., reversed phase chromatography), in combination with a mass spectrometer (LC/MS), is the most common approach [13,28,45,46]. Target analysis offers high sensitivity of detection and confident quantification for selected known target compounds with available reference materials for validation [13,47]. However, for many environmental pollutants, there are no reference standards available, or their cost is extremely high [45]. Adding to this, many of the TPs remain unknown, with no structural information.

Non-target analysis (NTA) takes advantage of high-resolution mass spectrometry (HRMS) data to identify the presence of formerly unknown chemical structures. HRMS instruments can screen a vast range of small molecular ions and make a distinction between them with precision of several decimal points [47]. While in target analysis low resolution mass spectrometry can be used, HRMS can give superior results in analysis of complex environmental samples [47].

NTA with no prior knowledge of the compounds present in the sample, is called non-target screening. In general, these steps are followed [47,48] : (1) all the data from separation and mass spectrometry that correspond to one peak (or alternatively called a feature [49]) are grouped together, (2) based on the molecular ion mass, all the possible molecular formulas are listed and the most appropriate one is selected based on heuristic rules, (3) all the possible structures for this formula are collected from

chemical databases and the appropriate structure is decided (structural annotation) based on different parameters, such as their MS/MS spectra, found in chemical databases or predicted. Additional methodologies have been applied specifically to identify unknown TPs. Schollée et al. [29] compared MS/MS spectra of wastewater samples and selected possible features of parents and TPs with mass differences corresponding to possible metabolic degradations. Brunner et al. [50] identified TPs, from simulated drinking water treatment, by comparing the features relative signal intensity before and after the treatment.

Suspect screening is part of NTA and focuses on the detection of selected compounds, expected to be found in the sample, forming a suspect list. The suspect compounds can be previously reported in the literature or predicted [13,45]. For the identification of compounds with suspect screening, similar approach is followed as with non-target screening, except that only features which their mass match at least one of the compounds in the suspect list are selected for identification. The suspect compound(s) that matches with a feature is then ranked based on credibility in the same manner as in step (4) in non-target screening mentioned above [48,51].

The use of computational prediction tools (in silico) has the advantage of predicting many different TPs structures based on known transformation reactions. In the next step, all these candidate structures can be prioritized based on relevant criteria, to decrease the number of possible candidates.

As was mentioned in chapter 1.1, PMT compounds have increased attention as water pollutants, as due to their properties, they might cause negative effects to humans and the environment. Many prioritization schemes, created in water pollution research, are related to, at least, one of the PMT criteria [52–54]. Another prioritization criterion can be the detectability with the employed analytical technique. For example, simple heuristic rules have been used to prioritize compounds detectable LC-ESI/MS [55,56]. Prediction models, based on experimental data instead of simple rules, have been developed to assess the amenability of chemical structures with various analytical techniques [57–59] with superior results [58,59].

1.4. Aim of the project

This research project aimed at building an automated workflow, which for a given list of parent compounds, can create a prioritized suspect list of the parents and their transformation products. More precisely, the workflow focuses on organic pollutants that can be analyzed with LC-ESI/MS and are relevant to wastewater treatment. Nevertheless, the basic steps of the processes used in parents and TPs curation can be applied for the creation of any suspect list with some alterations.

For the selected parent compounds, the TPs were collected from the literature and predicted *in silico*. Finally, the prioritization in the suspect list was heavily based on *in silico* models that predicted environmental fate properties (mobility in water) and detectability with LC-ESI/MS. This workflow contributes to a faster and more efficient detection of unknown TPs, which can be generalized in situations other than wastewater-related analysis.

2. Background information

2.1. Prediction models

While certain groups of compounds have been studied extensively over the years [60,61], data for many pollutants in question, such as their physical properties, toxicity, transformation products, are sparsely available in the literature [62]. To tackle this, many different computational models have been developed for the prediction of physical properties during the last decades.

2.1.1. Quantitative Structure-Property Relationship/ Quantitative Structure-Activity Relationship (QSPR/QSAR)

QSPR/QSAR models predict a compound property or (biological) activity based on the principle that the compound has similar properties with others similar in structure. More precisely, a QSPR/QSAR model describes the relationship between a group of theoretically computed variables, connected the compound structure (chemical descriptors or chemical fingerprints), and the predicted property [63,64].

A model provides accurate predictions only for similar compounds to the ones employed for the creation of the model (training dataset). This can be assessed with an Applicability Domain (AD) index. Many different approaches to define the AD of a model are found in the literature [65–68].

Different methods have been implemented for the creation of QSPR/QSAR models. Some of them are:

Multiple linear regression (MLR): The prediction model is built by linearly correlating multiple descriptors of chemicals with their known experimental property values [69]. It is one of the oldest methods to be used in QSPR/QSAR models [70].

k-nearest neighbor (kNN): The value of the compound property is based on the k number of closest neighbors (compounds) present in the training dataset [71].

Machine learning (ML) approaches: the model is built by training an algorithm to find connections between the structural characteristics of compounds and their known property values. Many different ML models are mentioned in the literature, with Neural Networks [70] and Random Forest [72] as some notable examples.

2.1.2. Transformation products prediction

Prediction tools for metabolites, and TPs in general, have been developed to make research in drug discovery, molecular synthesis, and environmental toxicology more time- and cost-effective [73].

In silico reaction models fall into these three categories [74]:

Knowledge-based (KB): Reaction rules, collected form the literature, are applied to a functional group of the parent compound and result to a product. The downside of KB models is low selectivity, with a great number of predictions (combinatorial explosion) [75] ,but only a small fraction of the TPs is likely to form [76].

ML-based: The appropriate reactions applied to the parent compound are predicted each time by a ML model. The sensitivity is related to the size of the existing data used to train the model [74,77,78].

Hybrid: The TPs are predicted based on reaction rules and reaction probabilities are calculated based on selected criteria, such as chemical classification and/or physicochemical properties [73], or reaction kinetics [74].

2.2. Chemical identifiers and structures

To study the over-increasing number of reported chemical compounds, an efficient identification system is needed, in which a particular chemical structure is correctly described without the risk of confusing it with another structure. Over the years, many different identification systems have been developed, each of them with their own advantages and weaknesses [79]. Key chemical identifiers relevant to the project, are mentioned here.

The **Chemical Abstract Service (CAS)** numbering system, is created by the American Chemical Society (ACS), in which a specific substance corresponds to a unique CAS number. CAS numbers can describe "*elements, alloys, minerals, mixtures, polymers, enzymes or polysaccharides*" [80]. It is possible for one chemical structure to have multiple CAS numbers. For example, both '12179-04-3' and '11130-12-4' describe sodium borate pentahydrate [81,82].

With **Simplified Molecular Input Line Entry System or SMILES** a chemical structure is described in one line of text. SMILES identifiers are easy to read by humans and have widespread use. On the downside, a chemical structure can be described in multiple ways with SMILES, hence it is not a unique representation system [79]. For example, both 'OCC' and 'CCO' SMILES correspond to ethanol, with the only difference being that each SMILES begins with a different atom of the molecule. Other problems exist with the lack of a consensus for stereochemistry and aromaticity representation [83,84]. Many different standardization systems have been released to solve these problems [83–85], but none of them is universally adopted.

It is possible to describe an aromatic structure by retaining the aromatic character of the bonds, by writing the aromatic carbons in lowercase "c" characters [86]. There is no widely accepted way for chemical software to recognize and describe the aromatic groups of a molecule [83]. Another way, but also not unique method, is to use kekule SMILES, where the aromatic bonds are described by a succession of double and single bonds [84].

IUPAC International Chemical Identifier or **InChI**, as the name suggest is a chemical identifier created by IUPAC. An InChI representation is organized in different layers of information. The three first layers are mandatory for an InChI and include (1) the molecular formula of the compound, (2) the connectivity between the atoms in the molecular core and (3) the connectivity between the atoms of the molecular core and the hydrogen atoms [87].

Up to three more layers can be added in a standard InChI with information on charge, stereochemistry, and isotopes (Figure 3).

The first three layers in a standardized InChI system provide a unique structure representation, in other words a chemical structure has only one InChI [79]. On the other hand, InChI standardization might change the initial input structure to obtain the desired unique representation by adding or removing proton or altering the stereochemistry [88]. InChI describes aromatic bonds in kekule form [87].

InChIKey is a digital representation of an InChI identifier, containing a standard number of 25 digits, that makes it friendly for search engines. An InChIKey is organized in three parts: The first part (InChIKey skeleton) with 14 digits describes the three mandatory layers of the InChI, the second part with 10 digits stereochemistry, isotope information etc. and the last part with one digit defines if the structure is neutral or not [87]. In Figure 3, we can see the InChI and InChIKey identifiers for the same

compound in the scenario when additional stereochemistry information is included (upper part) and with only the strictly necessary information (lower part). InChIKey is not considered completely unique representation unlike InChI. Because of this, an extremely small possibility exists of two chemical structures having the same InChIKey [79].

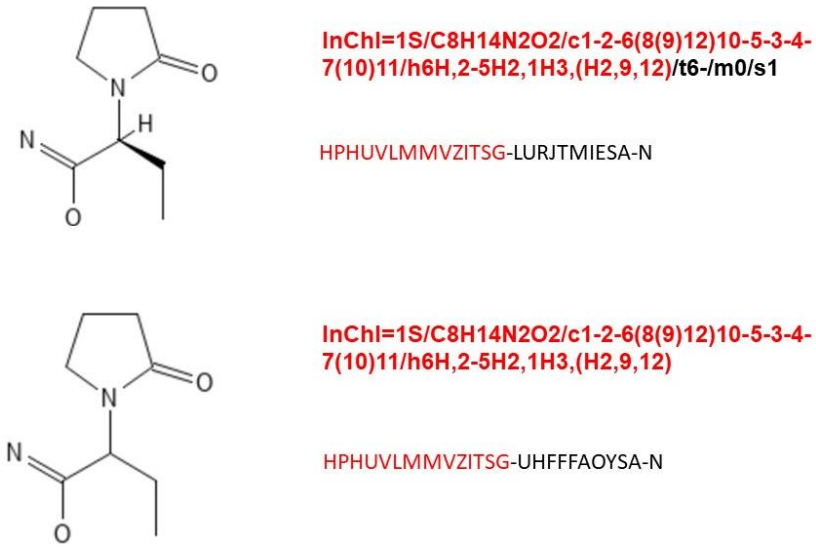


Figure 3: InChI and InChIKey identifiers for the antiepileptic drug levetiracetam. The upper representations include stereochemistry information, while in the lower part only the strictly mandatory information is retained. The three mandatory layers of the InChI's and the InChIKey skeletons are highlighted in red.

3. Methods

3.1. Workflow

The aim of the project was to build a suspect list of organic micropollutants along with their TPs in an (semi-)automated process.

First, a list of substances (with CAS numbers and trivial names) was received, in which only organic structures are selected as parents (step 1 in Figure 4). TPs are collected from the literature and from the environmental microbial degradation reaction library in BioTransformer [73,89] (step 2). The dataset is curated (step 3) by removing erroneous predicted structures with the QSAR Toolbox software [90] and compounds that are not detectable in ESI-MS based on basic heuristic rules (monoisotopic mass less than 50 Da and absence of any ionizable atoms O, N, S and P). Both parents and TPs are prioritized based on the PMT mobility criterion (Logarithmic Organic Carbon - Water normalized Sorption Coefficient or logKoc < 3) [17], their affinity to the selected separation method (Logarithmic Octanol – Water Partition Coefficient or logP, ranging between -2 to 5) and their probability to be detected from LC-ESI/MS with a ML model build from Lowe et al. [58] (step 5).

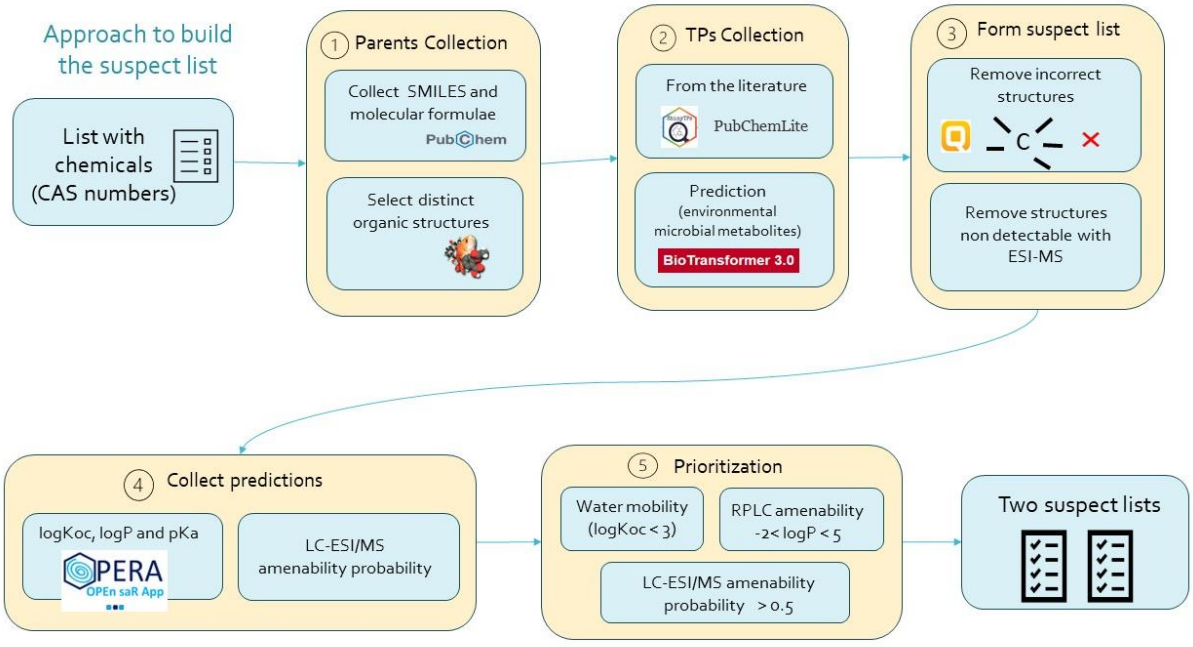


Figure 4: The final flowchart with the workflow steps.

3.2. Data and tools

The input dataset, used to test and build the workflow, was a raw list of substances, emitted by industries in various regions in the south of the Netherlands. The list was formed for the purposes of the “Van bron tot effect” (“From source to effect”) project [91]. The list contains pharmaceuticals, pesticides, and industrial chemicals.

Only freely available tools were used in this project. The process is mainly carried through a code script written in R (version 4.3.2). The software RStudio (version RStudio 2023.09.1 build 494) was used as an editor to write the script.

If there was more than one option available to complete a task, the appropriate tool was selected based on these criteria:

- Freely accessible
- User friendly
- Batch mode availability
- Time efficient

Most of the steps were carried through R function packages, however data from external software were collected manually for the validation (steps 1 and 3) and prioritization steps (step 4). When a step is performed manually out of RStudio, it is mentioned in the text. In Table 1 all the R function packages are listed, along with their use in the workflow. Processes which could not be executed in R environment, were completed with the help of additional software (Table 2).

Table 1: Function packages used in the workflow.

Package (version)	Applications	Reference (doi or source website)
webchem (1.3.0)	Collect information for compounds from PubChem	[92]
dplyr (1.1.4)	Clean the datasets	https://dplyr.tidyverse.org [93]
stringr (1.5.0)	Divide compounds into categories based on detected text	https://stringr.tidyverse.org [94]
shinyTPs (0.1.9)	Collect TPs from PubChem	[95]
patRoon (2.3.0)	• Collect TPs from PubChem, • Predict TPs with BioTransformer, • Analyze mass spectra	[96] [51]

Table 2: Software accessed out of R environment.

Software (version)	Application	References
Open Babel (3.1.1)	Analysis and editing of chemical structures	[97]
QSAR Toolbox (4.6)	Prediction and collection of experimental values for logP, logKoc and pKa. Marking of erroneous structures found in the predicted TPs.	https://qsartoolbox.org/ [90]
VEGA QSAR (1.2.3)	Prediction and collection of experimental values for logP, logKoc	https://ceur-ws.org/Vol-1107/paper8.pdf [98]
Opera (2.9)	Prediction and collection of experimental values for logP, logKoc and pKa.	[99] [100]

3.2.1. Curation of parent compounds

In this section are presented the tools tested or integrated in step 1 of the workflow.

3.2.1.1. Databases to search information for CAS numbers.

With the CAS numbers in hand, it was needed to extract the relevant chemical structures from a chemical database. The following databases were considered:

The **CAS Common Chemistry** [101], is an open-source database with 50,000 freely available CAS numbers. By adding a CAS number in the search query on the website, information for a chemical is available, such as name, other chemical identifiers, and chemical properties, for example, molecular mass and boiling point. Batch search is possible with API, after request. API stands for Application Programming Interface, and in this context sets the communication between the user computer and the server where the database is located.

The **Chemical Translation Service** [102] is a webservice which facilitates the conversion of one chemical identifier to another for more than 50 different identifiers. It is easily accessible in R with the function package webchem, through an API service[103].

CompTox Chemical Dashboard [104] is a database focused on chemical, toxicological and exposure information with over 1.2 million chemical compounds (as of March 2024). Examples of available information include chemical identifiers, physicochemical properties, toxicity concentrations, environmental fate and exposure, available literature sources, among others. The API service is available only with special permission [105].

PubChem [106] is a chemical database run by the U.S. National Library of Medicine with information for over 111 million compounds, including chemical identifiers, physicochemical properties, toxicity, applications, and transformation products. Chemical information is organized in many different domains with only two of them being relevant to this project: “Substances” are archives for chemical entities, submitted by external sources, with each submission for a chemical entity having a separate page, and “Compounds”, where all the extracted information from “Substances” on a single compound is combined to create one entry [107]. A structure found in the substance library, follows a standardization process, where the given structure is validated and in the case of a mixture, the distinct neutral compounds are derived [84,107]. With programmatic access, through the R-package webchem, it was possible to collect chemical identifiers and physicochemical properties (see Table 3).

Table 3: Properties collected from PubChem.

Property name	Description
CID	Compound identity number
Title	Compound name
InChI	Standard InChI
InChIKey	InChIKey of the given InChI
IsomericSMILES	SMILES with stereochemistry information
MolecularFormula	Molecular formula of the compound
MonoisotopicMass	Monoisotopic mass of the compound (in Da)
Charge	Charge of the molecule

3.2.1.2. Curation of parents.

The initial list contained different types of substances such as organic compounds, organic and inorganic salts, or organometallic compounds. Since the project was focused on organic compounds detectable with LC-ESI-MS, and the structures needed to follow the set criteria of the prediction software, a curation procedure for the substances had to be developed. These tools were found for this purpose:

With the use of functions from a cheminformatic toolkit such as **RDKit** [108] or **CDK** [109] it is possible to build a custom-made classification algorithm. Starting with a SMILES identifier as an input, the toolkit gives information about the elements and/or the chemical groups of the compound, and the way the elements are connected to each other. RDKit is available in python programming language. An R package [110] has been developed that transport the output of RDKit from python to R, however it is compatible only with Unix operating systems and not with Windows. CDK is available in R through the "rcdk" package [111].

Tailored made text detection algorithm: Detecting character patterns in SMILES or molecular formulas with **regular expressions (regex)** is a simple way to categorize chemical structures. The SMILES and molecular formulae of structures are recognized as text by the algorithm and are filtered based on the presence of specific elements which are described by letters of the alphabet. The recognition is performed with the package "stringr" [94] in R. This method is less complicated and less intensive for the computer processor and was used instead of the cheminformatic toolkits mentioned above.

Open Babel 3.1.1 [97] is a java-based software to convert one chemical structural identifier to another and perform simple manipulations on chemical structures. Currently, Open Babel provides conversions for over 110 different identifiers. In addition, it offers additional manipulation capabilities, as for example neutralization of charged structures with addition or removal of hydrogens, conversion between aromatic and kekule depiction or removal of stereochemistry information. The software is accessible directly through the java application, which is possible to incorporate in R or through the graphical user interface (GUI). During the project Open Babel was accessed only through the GUI.

In the final workflow the organic compounds were selected by a tailor-made text detection algorithm. The organic ions were neutralized by Open Babel.

3.2.2. Curation of predicted transformation products

In silico tools for TPs prediction might produce erroneous structures [76]. These incorrect datapoints take extra space in the dataset, cause errors in other predictive models and overall slow down the process. The following tools were considered for the validation of predicted TPs.

RDKit does not accept chemical structures with incorrect number of valence bonds as an input, thus it can be used to tag the erroneous structures. Unfortunately, RDKit is not accessible in R with a Windows operating system and for this reason it was not used.

CDK (or rcdk in R) accepts SMILES as an input and contains functions for recognizing the elements comprising a molecule and the number of their bonds. When loading the SMILES, rcdk validates the molecule, if asked, and structures with incorrect connectivity in aromatic bonds are excluded [111].

QSARS-ready [112] is a software which validates SMILES before their input in a QSAR model. The validation steps include a valence check, obtaining neutral compounds from salts, removing stereochemistry information, obtaining canonical SMILES, and removing duplicates. Originally it is part of OPERA software [99], but it is available as a standalone application.

QSAR Toolbox [90] is an application providing multiple tools (including QSAR models) to characterize molecules from a toxicological perspective. More information can be found in chapter 3.2.5. The software performs a validation process before a structure is loaded, where the duplicates and structures with incorrect number of valence bonds are marked and removed. Compared to the other tools that were readily accessible, the validation system in QSAR had the best performance and was the one eventually used.

ChemMine [113] is an online cheminformatics platform for molecule clustering and physicochemical properties prediction. Due to its user-friendly interface, it was used to visually inspect the predicted structures and check the efficiency of the validation tools that marked the erroneous structures. That was done by importing the SMILES of the structures selected as incorrect and checking if they had incorrect number of valence bonds.

In the final workflow, the incorrect structures were detected by QSAR Toolbox and then removed manually. The selected structures were visually inspected with ChemMine to understand better the nature of the structural issues.

3.2.3. Collection of TPs

In this chapter are named all the different sources found to collect TPs from the literature or predict them during the second step of the workflow. The TPs curations was heavily relied on patRoon, an R package with NTA workflows [51,96].

3.2.3.1. Search for TPs in the literature

PubChemLite [114] is a section of the PubChem compound database with compounds relevant to environmental analysis. It focuses on chemicals that are produced in large quantities, agrochemicals, and pharmaceuticals. Reported TPs from a variety of sources are included and were consequently incorporated in PubChem “Transformations” section. PubChemLite was accessed through the patRoon package. Parent and TP structures are connected based on their InChIKey skeleton. Practically this means that all the isomers of a structure are perceived as one entity when searching for TPs or parents. PubChemLite was easily accessed through patRoon, which gives the option to collect multiple generations of TPs. Generations of TPs are collected as followed: TPs resulting directly from the parents form the first generation, while TPs created form the first generation make the second generation and so on.

ShinyTPs [95,115] is an R based application which, for a given number of isomeric SMILES (collected from PubChem) of the parent compounds, returns their TPs as they are recorded in the “Transformations” section of PubChem. The application gives the option to the user to add manually TPs that are missing from the “Transformation” section but are mentioned in the “Metabolism and Metabolites” section in PubChem. This functionality was not applied in the workflow.

Both PubChemLite and ShinyTPs were used to collect TPs from the literature. With PubChemLite were collected two TPs generations and with shinyTPs only the first generation.

3.2.3.2. Transformation products prediction

EAWAG-Pathway Prediction System (EAWAG-PPS) [116,117] is a prediction system specialized in environmental microbial degradation. It is based on University of Minnesota Biocatalysis/ Biodegradation Database (UM-BBD). Nowadays, the prediction tool along with the database is hosted by EAWAG (Swiss Federal Institute of Aquatic Science and Technology) and contains more than 1500 microbial degradation reactions. The predictions are knowledge-based and as it is expected from this type of model, large number of irrelevant TPs are predicted due to combinatorial explosion [74,75]. Since its release, measures have been taken to decrease combinatorial explosion by making the reaction rules more specific and by categorizing them on how likely they occur [75]. Another limitation reported in the literature is that the reactions in the model are not representative for microbial degradation of pollutants in environmental conditions. Most of the data were collected from laboratory studies. where the bacteria were provided with large concentrations of a chemical pollutant [117], while in real conditions the concentration would be much lower, with a different reaction mechanism taking place [118]. EAWAG-PPS does not produce batch predictions and there is no automated way to export the predicted TPs from the website.

enviPath [117,119] is an API application for microbial TPs predictions based on EAWAG-BBD reaction rules but contains two additional reaction libraries: the EAWAG-SLUDGE for microbial transformations in activated sludge, and the EAWAG-SOIL for microbial degradation reactions in soil. The already existing EAWAG-BBD has been updated with new reaction rules, and for each TP a probability of prediction is included in the results [119]. An API service is available for python [120] , but it was not possible to have access it in R.

Chemical Transformation Simulator (CTS) [121,122] is a web-service by U.S. EPA, with a transformation products predictor, the Reaction Pathway Simulator (RPS), among other prediction tools. It implements knowledge-based models for abiotic hydrolysis, reduction and photolysis, mammalian metabolism, environmental microbial degradation from enviPath, and PFAS abiotic and biotic degradation. The reaction rules for the mammalian metabolism library are sourced [123] from Metabolizer, a commercial human xenobiotic phase I biotransformation library from ChemAxon [124]. Each predicted metabolite is categorized as ‘likely’, ‘probable’ or ‘unlikely’ based on the kinetics of the reaction.

Batch prediction is available for up to 10 compounds at a time from the website and from the API service [125] with no defined limit. All the reaction libraries of CTS are available in patRoon in R through API, except the enviPath environmental microbial degradation library.

BioTransformer [73,89] is a web-service and a java-based application, created by Wishart research group (University of Alberta, Canada), which predicts and identifies metabolites. The metabolite prediction tool contains human metabolism (phase I and II) libraries along with libraries for human gut microbial degradation, environmental microbial degradation (from EAWAG-BBD reaction rules) and combinational libraries (AllHuman library, predicts metabolites from all the human and human gut metabolism, while Superbio library executes 4 iterations of AllHuman library or until no new metabolites are predicted). The models for phase I and phase II metabolism are hybrid while the remaining are knowledge-based.

Only SMILES of distinct organic compounds, with molecular mass below 1500 Da, containing only the elements carbon (C), oxygen (O), nitrogen (N), sulfur (S), phosphorus (P) and halogens (X) are accepted as an input by the model. The metabolites are predicted with the use of reaction rules and machine learning algorithms (for a few of the models), which predict and select metabolites more likely to occur.

In the BioTransformer website it is possible to request for up to 10 predictions per minute. However, the java application provides predictions offline, without this limitation.

Only the environmental microbial degradation reaction library from Biotransformer was used in the workflow. The java application was easily accessible through the R package patRoon and two TPs generations were collected.

3.2.4. Properties prediction

EPI Suite™ [126,127] is a software with a collection of QSPR/QSAR models for physicochemical properties, biodegradation, and toxicity endpoints, developed by EPA. Relevant to the project, were the models for prediction of logP, KOWWIN, and the molecular connectivity index (MCI) model for logKoc, KOCWIN. The KOWWIN model was based on an atom/fragment contribution method, where partition coefficients for 150 separate elements and functional groups were calculated with a multiple regression model based on experimental data. A prediction for a compound is made based on the number of times an atom/fragment is found in the molecule, and the addition of correction factors [128]. KOCWIN originally gives predictions from two models: an estimation based on logP (predicted from KOWWIN), and a more accurate estimation based on MCI [129]. The molecular connectivity index is a molecular fingerprint which describes the area of the molecule accessible to a solvent [130].

The software does not return an AD metric for the prediction; however, the training and validation datasets can be found online [131,132]. It has been reported that EPI Suite™ predictions have decreased accuracy for compounds with molecular weight higher than 500 Da, PFAS and other highly halogenated compounds, inorganic compounds and compounds containing fragments that are not included in the model [127].

OPERA [133] is a software with prediction models for physicochemical and environmental fate properties. The predictions are based on cluster k-nearest neighbor models (k=5). The data used in the models and algorithms are freely accessible [134] and are common to a large extent with the data used by the EPI Suite™ models [99]. Two different AD metrics are provided with the predictions: a binary value (yes or no answer) for the presence of the compound inside the AD, and an AD index that is based on the similarity between the compound and the 5 nearest neighbors. Other useful information includes an AD confidence level, and all the available information for 5 nearest neighbors. In the project, OPERA predictions for logP, logKoc and pKa were implemented in the final workflow. The pKa values can be predicted only for the strongest acid and the strongest basic group of the molecule with OPERA.

VEGA QSARs [135] is a software with more than 90 QSAR models for physicochemical, environmental fate and toxicity endpoints. For the project, the three logP models (Meylan, AlogP and MlogP), one logKoc model (from Opera) and one pKa model (from OPERA) were considered for use. The Meylan logP model follows the same methodology as in EPI Suite™. AlogP is a multilinear regression model based on the hydrophobicity contribution of 120 different elements/chemical groups [136]. Similarly, MlogP is a multilinear regression model based on the contribution of 13 different molecular descriptors [137]. The AD index is given, based on two main criteria: (1) difference between experimental and predicted value for similar compounds in the training set, and (2) difference between predicted value of the input and the experimental values of similar compounds [68].

QSAR Toolbox 4.6 [90] is a software owned by the Organization for Economic co-operation and development (OECD) with workflows for risk assessment. Currently, it gives experimental values

and predictions for physicochemical properties, environmental fate, and toxicity endpoints, from OECD models, along with incorporated models from other software, such as EPI Suite™. It also contains libraries for TPs predictions. The models of interest were: for logP (from EPI Suite™), logKoc (both models from EPI Suite™), and pKa [two models from ChemAxon (commercial software) and 6 models from OECD]. There was no provided information from QSAR Toolbox about the training datasets used by the models.

CTS provides physicochemical properties predictions from the software EPI Suite™, T.E.S.T., OPERA, and ChemAxon, along with the average of these predictions as a consensus value [121]. The predictions are available through the online platform or through the API service.

ML models for LC-ESI-MS amenability prediction created by Lowe et al. [58], were applied to predict which compounds in the suspect list are likely detectable with LC-ESI-MS. They are comprised by two random forest models, one for positive and one for negative ionization mode. The training datasets comprise of compounds from the MassBank of North America database [138,139], with available data on whether they are detectable with LC-ESI-MS or not, without discriminating based on a particular separation technique. The models predict if a compound belongs in the amenable or unamenable category with the probability score (range from 0-1) that describes how likely the compound belongs to the amenable category. For probability score above 0.5 the compound is classified as amenable. The model is not uploaded yet on the internet, however the predictions for a given list with chemical structures were provided after communication with the research group.

For the prioritization of the compounds, physicochemical characteristics (logKoc, logP and pKa) were collected with OPERA. The LC-ESI/MS amenability predictions for both positive and negative ionization mode with mass spectrometry were produced by the aforementioned ML model.

4. Results and Discussion

4.1. Characterization of unique chemical structures

In the validation step for the parents, duplicates were removed based on CID and SMILES with stereochemistry markings were collected from PubChem, since this was the accepted input for shinyTPs [140]. Compounds with different CID have different isomeric SMILES, thus duplicate structures from PubChem can be removed based on isomeric SMILES as well. However, the SMILES of the predicted TPs do not follow the PubChem standardization rules [84]. Thus, the duplicates in the final suspect list, containing both parents and TPs, cannot be removed based on SMILES. Standardized InChI and InChIKey are given for all the structures, but some of the results might or might not contain additional information on stereochemistry and this can lead to confusion. Mass spectrometry is often not able to distinguish between different isomeric structures and for this reason it was decided to group all the isomers as the same compound. Indeed, tools and platforms [114,141] intended for NTA such as PubChemLite, define all stereoisomers as the same compound based on the InChI skeleton, and the same approach was applied in this project.

4.2. Validation of the compounds in the initial list

4.2.1. Collection of chemical structures from CAS numbers

The raw list of chemicals used as input contained 5,442 entries. Among the different variables for each chemical only the CAS numbers and trivial names were relevant to the research and were used. After removing the duplicates and missing entries based on the CAS numbers, 1,819 entries of unique CAS numbers remained. More information about this step can be found in Appendix I.

Only the unique CAS numbers were selected from the list and were used as an input for search in PubChem “Compound” library records.

PubChem was chosen among other available chemical databases for these reasons: (1) it is one of the greatest in size, (2) it provides unrestricted programmatic access, (3) it is accepted to search directly for chemical compounds with CAS numbers, (4) it provides information on TPs reported in the literature. However, a common problem with online databases is the existence of faulty chemical information [88,142]. Even though PubChem has adopted a standardization procedure for the “Compound” library [84], the possibility of quality issues cannot be completely excluded [143].

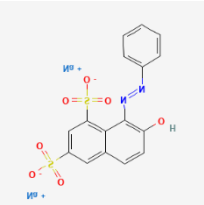
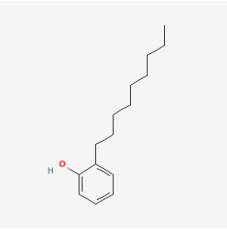
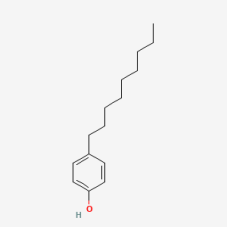
The PubChem compound record numbers (CID) which corresponded to a CAS number were collected. It was possible to obtain results for most of them, but 45% of the CAS numbers returned no result. Table 4 shows a few examples of the chemical species with no result in PubChem. Some CAS numbers described complex chemical mixtures which are hard to connect to specific compounds, while in other cases, even if a chemical compound was present in the database, the CAS number was missing from the results. In the second case, it was speculated that the CAS number was mentioned in a section of the database where the webchem package does not search into.

It was considered to search in the PubChem “Substance” database, to find mentioned compound entries. However, this information is unverified (Table 5) and might result in faulty results. No additional correction was applied in the results and all the CAS numbers with no matched CID were removed.

Table 4: A few examples of CAS numbers with no compound entry in PubChem.

CAS number	Description	Comment
91079-47-9	C9-11-phenols	Complex mixture of phenols from 9 to 10 carbon atoms.
21319-43-7	bis(2,2,6,6-tetramethyl-3,5-heptanedionato)lead(II)	There is an entry in the compound database, but no CAS number is assigned to it.
64742-53-6	"A complex collection of hydrocarbons"	
13463-39-3	Nickel tetracarbonyl	There is an entry in the compound database with the CAS number in an inaccessible section from the search function.

Table 5: Examples of irrelevant compound pages found through redirection from a substance page (figures source: PubChem).

CAS number	description	CID	description	comment
64742-65-0	Distillates (petroleum)	16015	 Orange G (colorant)	It was found in a substance entry of Orange G, that it was archived with an ID number identical to the CAS number, resulting to confusion.
84852-15-3	Phenol, 4-nonyl-, branched	67296	 2-Nonylphenol	Straight chain positioned on C2. The CAS number describes a nonylphenol with a branched chain on C4.
84852-15-3	Phenol, 4-nonyl-, branched	1752	 4-Nonylphenol	Straight chain instead of branched.

Multiple CAS numbers gave a single result from the database while there were multiple CIDs which corresponded to one CAS number. This is explained by different CAS numbers describing the same compound in different consistency or form. On the other hand, one CAS number might describe all the different isomeric forms of a compound, while in PubChem each isomeric form might have a distinct CID [84,144] and or even a structure with absence of any stereochemistry marking. For more information the reader can move to the tables 13 and 14 in Appendix I. In total 1,040 PubChem compounds were collected. All the different isomeric structures were preserved since they were used to collect TPs from PubChem in later steps. In Table 6 are listed the relationships between the CAS numbers and CIDs.

Table 6: Overview of the relationship between CAS numbers and CID's.

	CAS numbers	**CID numbers**
unique	971	993
one to multiple	32	36
missing		816

4.2.2. Automated workflow to select the parent compounds based on their structure.

It has already noticed upon receiving the raw list, that it contained a variety of different chemicals such as inorganic salts, metals, organometallic, neutral organic, and organic salts. Prediction models of TPs structures and QSAR models focus on organic compounds where all the atoms are connected with covalent bonds [73,99,121]. On top of that, analytical methods like LC-ESI/MS focus on organic compounds [145]. The appropriate selection of the collected chemical structures based on the criteria given from the *in silico* tools, prevents problem with the TPs curation and physiochemical properties predictions and can be applicable in similar workflows.

From the collected PubChem compounds, only organic structures made of one distinct entity (all atoms were connected covalent or coordination bonds), and contained only the atoms C, H, O, N, S, P, F, Cl, Br, and I were selected as parent compounds. The selection process was made by a tailor-made text recognition algorithm applied on the SMILES and molecular formulae from PubChem. More information on the selection steps can be found in Appendix II. The resulting output was visually inspected to verify the success of each filtering step.

An extra set of steps was applied in the ionic compounds datasets to obtain additional distinct organic compounds. Salts are not an acceptable input in many in silico tools such as BioTransformer, yet the organic ions are still relevant since the parents and their TPs might be detectable with LC-ESI/MS. This procedure has been described as “salt stripping” in the literature and similar automated validation processes have been developed in QSAR-ready [112] and PubChem [84] among others [142,146,147] with none of them written in R. Indeed, there is no known R package able to apply neutralization on a chemical structure, and Open Babel was used instead.

Open Babel neutralizes ions by adding or removing a proton or an electron from the structure (see examples in Figures 5 and 6). The SMILES ionic representations were loaded manually in Open Babel GUI. The process can be automated by integrating the java application in the R code, but it was not implemented due to lack of time.

A sample of 4 SMILES were depicted with Molview [16] before and after neutralization, to test that the resulting SMILES are correct. The structures that were not neutralized after this process were

removed (see examples in Figure 6), as it was observed that ionized compounds were a source of errors in BioTransformer. In addition, the compound records of the remaining compounds were searched for in PubChem and the structures without any results were excluded. This was supposed to be a validation step to make sure that the structures are correct. In total 15 distinct compounds were derived from the 229 organic salts. The steps to categorize the compounds are shown in Figure 7.

In total, 509 distinct organic compounds were extracted from the initial list and used as parents.

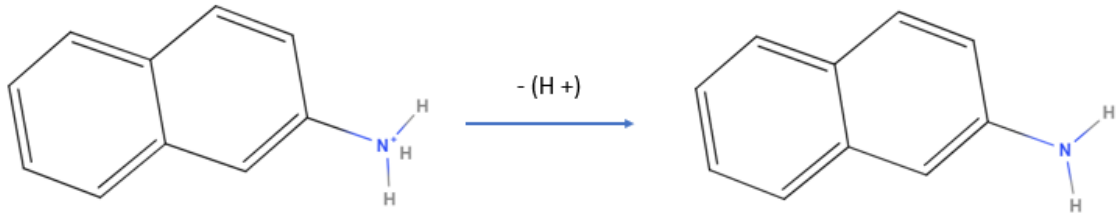


Figure 5: Neutralization by removing a proton (structural depictions source: Molview).

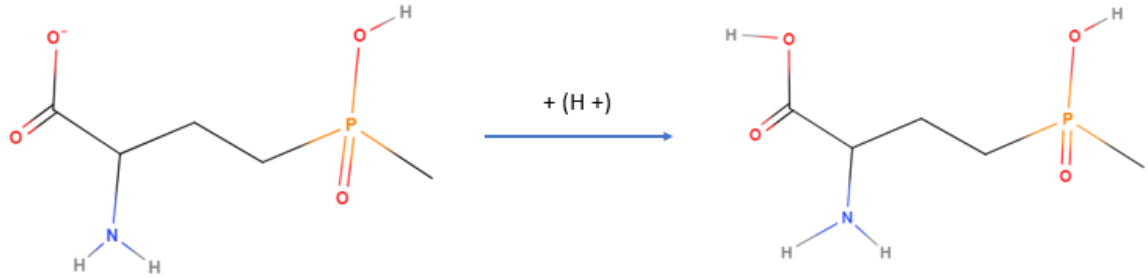


Figure 6: Neutralization by adding a proton (structural depictions source: Molview).

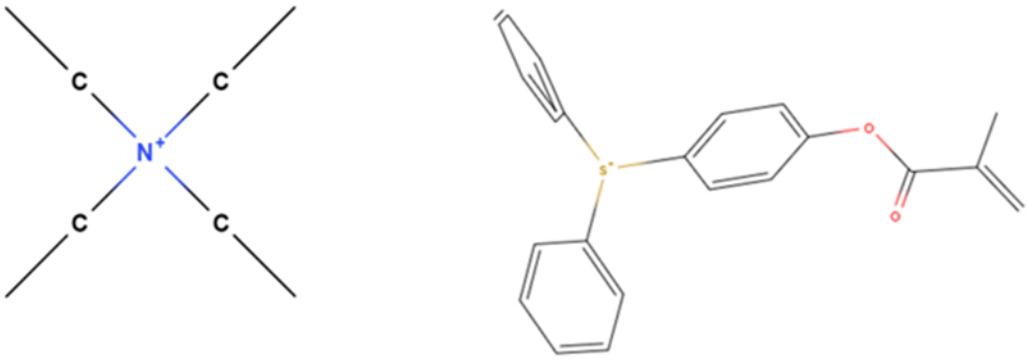


Figure 7: Examples of ions that could not be neutralized by Open Babel (structural depiction source: Molview) .

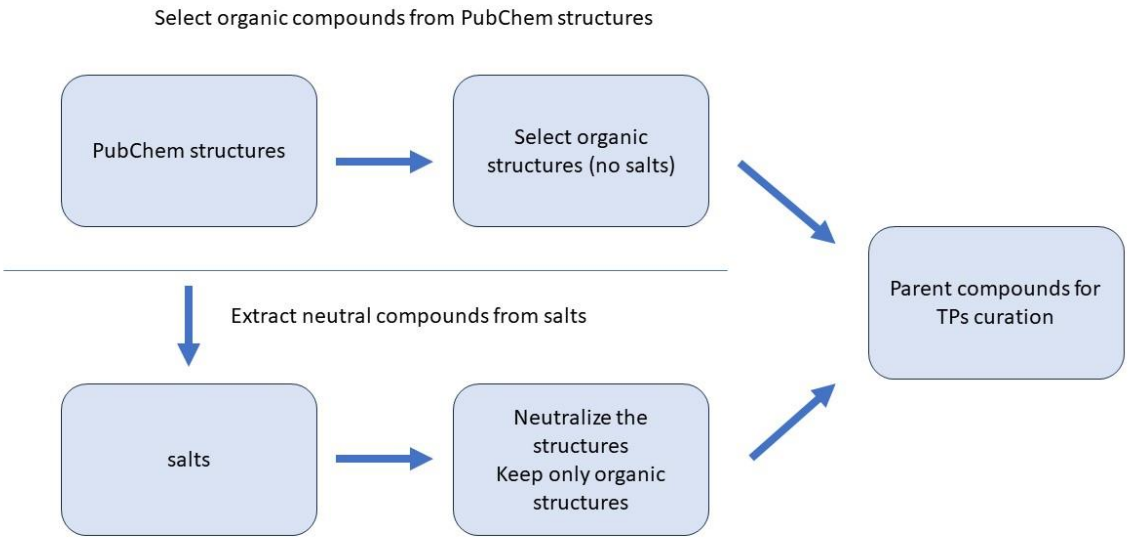


Figure 8: Steps for parent compounds curation.

4.2.3. TPs collection

4.2.3.1. Collection from the literature

Before proceeding to TPs prediction with *in silico* methods, TPs reported in the literature were collected from shinyTPs and PubChemLite. It is important to include known TPs of the parent compounds, since it is confirmed that these compounds exist. This can support the prioritization of predicted TPs, in case some of them are identical with reported TPs. The collection of references or metadata to prioritize a candidate structure is often used in NTA [148] and a similar approach might be applied to prioritize structures in a suspect list.

After searching with shinyTPs for reported TPS of the 509 collected parents, 392 results were found for 95 parents (19%). The results from shinyTPs were collected manually, as no alternative way could be found to obtain the results in an automatic manner.

The same search was repeated on PubChemLite library and returned 543 results from 102 parents (20%). It was possible to collect the results automatically with patRoon. As with shinyTPs, there are reported TPs only for a small fraction of the selected parents. This shows that not much information about TPs is available in PubChem.

PatRoon gives the option to return multiple generations of TPs, by using the search results as input to find the TPs of the next generation. The results from two generation of TPs in PubChemLite were included in the workflow (Figure 9).

PubChemLite returned a higher number of results compared to shinyTPs. Considering the unique TPs structures (defined by InChIKey skeleton), PubChemLite covers almost all of shinyTPs results while providing additional results (Figure 10). This striking difference can be explained by the way each tool collects the results. ShinyTPs search in PubChem for TPs based solely on the SMILES given as an input [140]. On the other hand, PubChemLite groups all the different isomers as the same compounds based on the InChIKey skeleton [114]. Because of this, all the recorded TPs from all the different parent isomers are given as an output. On the other hand, PubChemLite contains only a fraction of PubChem and can find hits only for parent compounds that are produced in large quantities. As shinyTPs search in the whole PubChem database, that explains the fact that there is a TP not found from PubChemLite.

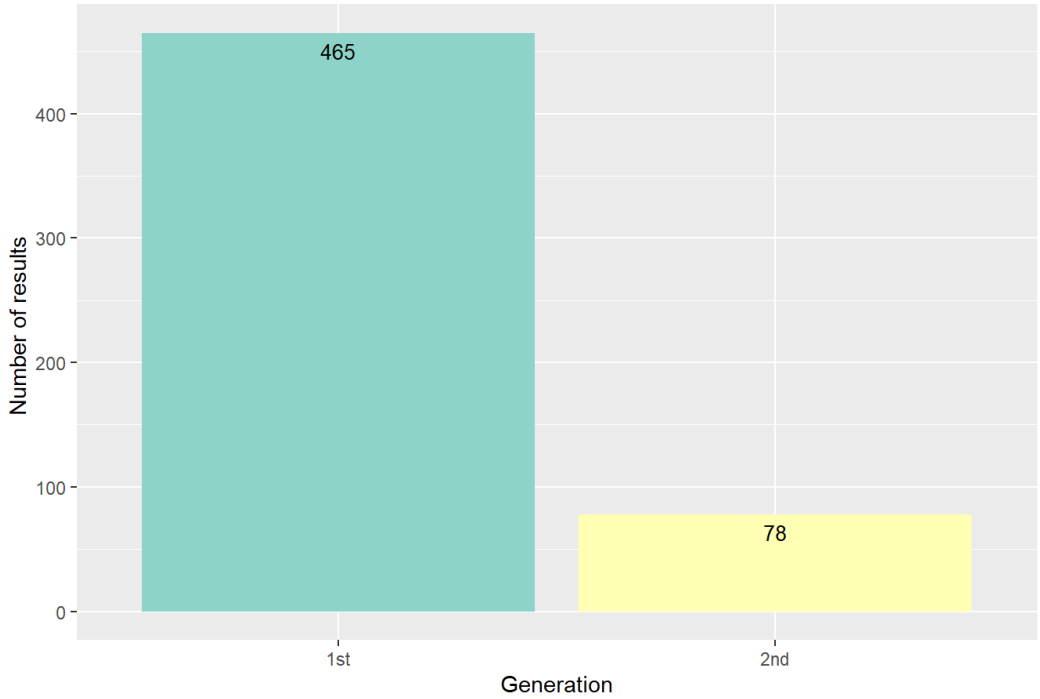


Figure 9: Number of parent-TP pairs found by 1st and 2nd iteration from PubChemLite.

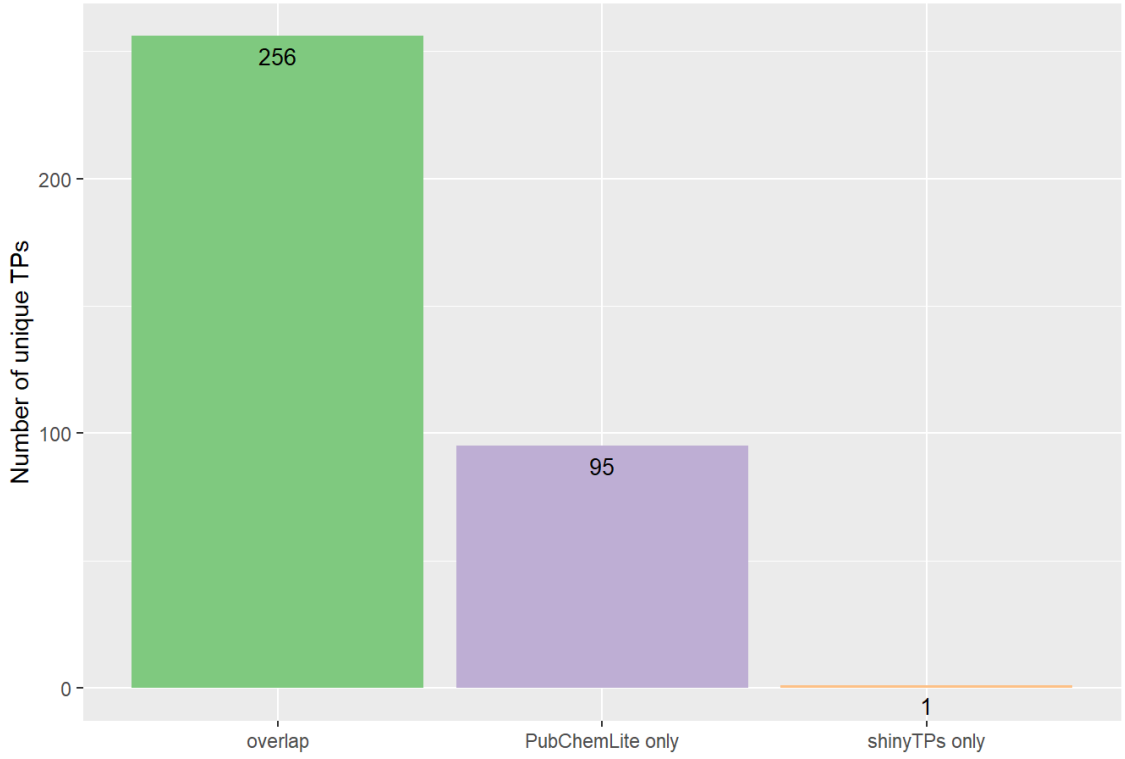


Figure 10: Overview of overlapping and exclusive unique structures from both PubChemLite and ShinyTPs. The second generation TPs from PubChemLite are included.

All the results were added together in one common file (Figure 11). The duplicate results were removed based on InChIKey skeleton, while preserving additional information about the parents where the TPs come from, the transformation mechanism, the environment where the transformation takes place and the source publications. Both shinyTPs and PubChemLite collect information from PubChem, thus their output is of similar format (Table 7).

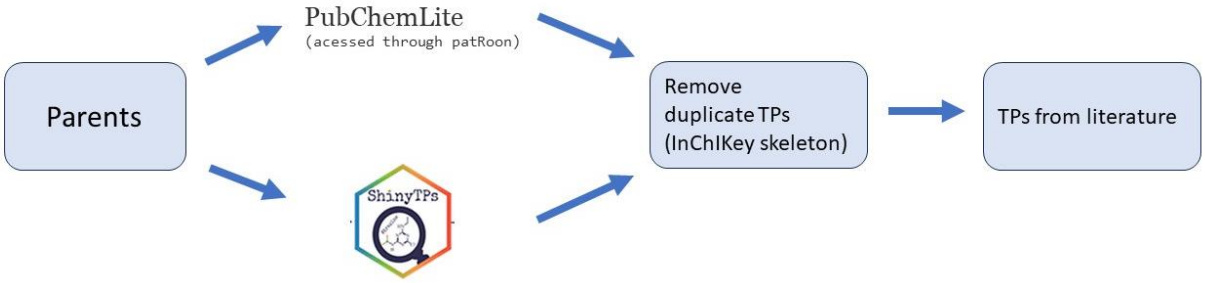


Figure 11: Steps for TPs collection from the literature.

Table 7: The common variables between shinyTPs and PubChemLite, relevant to the project.

Common variables between shinyTPs and PubChemLite	Explanation	Example
"transformation"	Transformation mechanism	"Hydroxylation of acyclic aliphatic secondary carbon / Human Phase I"
"biosystem"	Environment where the transformation takes place	"NULL/preterm neonates/monkey/Rats and mice"
"evidencedoi"	Literature sources identifiers	"10.1021/acs.est.1c00466/10.1016/j.watres.2019.114972"

The reported TPs were expected to come from a variety of sources. By collecting all the reported keywords in the "biosystem" variable for each TP, each keyword was put in one of the 6 made intuitive categories (Table 8). The categories were defined based on the wastewater treatment point of view. The "human and other eukaryotes" metabolites are relevant for the TPs that are formed before wastewater treatment, for example consumed pharmaceuticals that are being metabolized by the patient. The "bacteria" metabolites are formed by bacteria during wastewater treatment, but also in other environments, such as soil. The TPs in the "soil" category, as the name suggest are formed in soil environments and are relevant for wastewater from agricultural areas. In the "Plasma–based water treatment" category are TPs formed by an advanced oxidation method in which plasma degrades the water pollutants [149].

The TPs were assigned to one or multiple categories, considering the presence of keywords linked to a certain category. In Figure 12, is depicted a, UpSet graph presenting the size of each category, on the left down part, and the intersecting parts between them on the main part of the graph.

Table 8: Keywords to categorize each TP in a category.

Category (number of unique structures in each category)	**Keywords**
Human and other eukaryotes (211)	"Human", "Rat", "Mammal", "Plant", "Rabbit", "Fungi", "Animal", "Bird", "Rabbits", "Rats", "pig", "hamster", "dog", "pigeon", "rac cytosol", "preterm neonates", "monkey", "mice", "mouse", "Humans", "Baboons", "Liver", "baboon", "Cotton", "Cunninghamella echinulata", "soybean", "dogs", "corn", "Sugarcane", "Anodonta woodiana", "cows", "Horse", "Swine", "plants", "Mytilusedulis", "Crassostrea gigas", "Abalone", "flatfish", "Rodents", "Japanese quail", "Aspergillus", "Polyporus", "Trichoderma", "yeast", "scorpion", "peripatus", "boophilus", "eperia", "tegenaria", "mitopus", "Phalangium", "spergillus neurospora", "polystictus", "ferret", "hedgehog", "bat", "man", "monkeys", "cat", "galliformes", "anseriformes", "Mammals"
Bacteria (8)	"Bacteria", "arthrobacter sp", "flavobacterium peregrinum", "Pseudomonas"
Other (12)	"Paper", "In Vitro", "Environment"
Not assigned (176)	"NULL", "NA"
Soil (24)	"soil", "Soil microcosms"
Plasma-based water treatment (8)	"Plasma-based water treatment - gas phase", "Plasma-based water treatment"

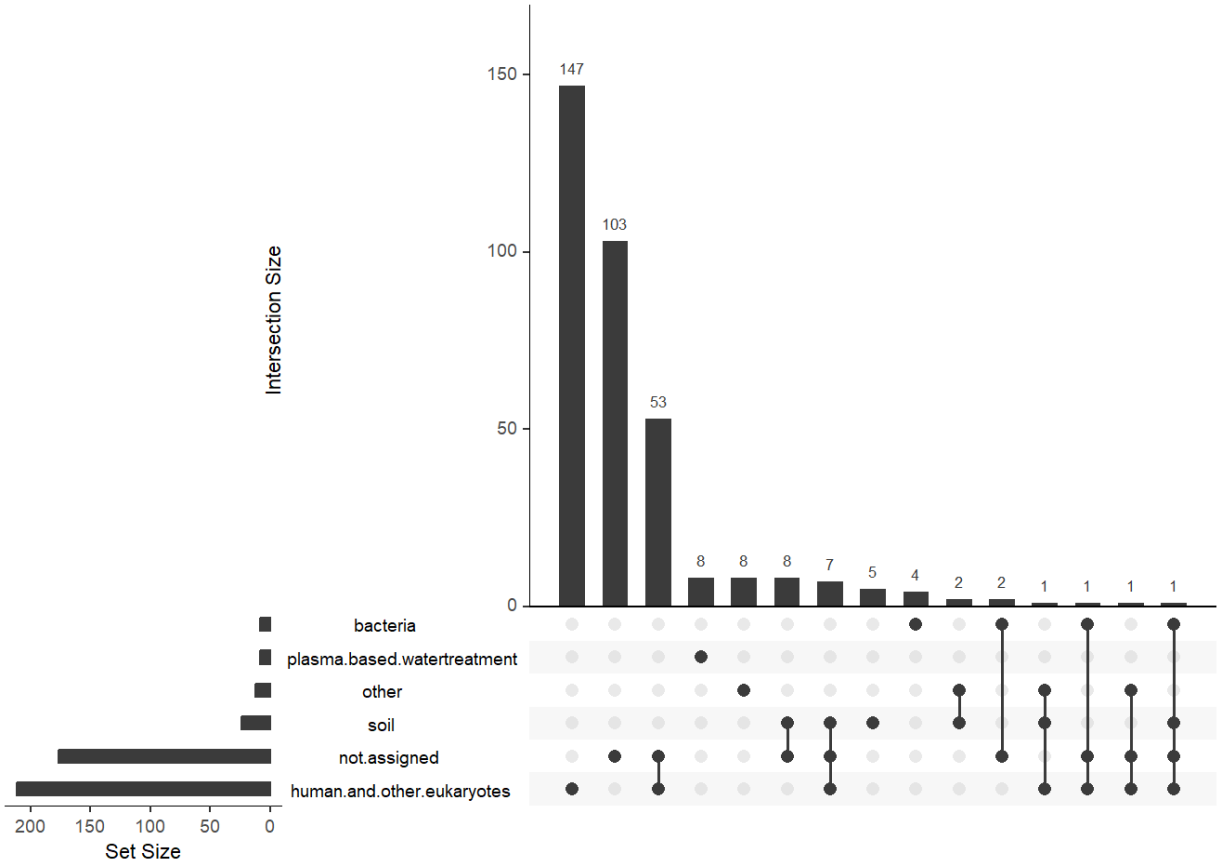


Figure 12: UpSet graph with the intersection areas for the different biosystem categories, for the sum of collected TPs from shinyTPs and patRoon. In the top right bar plot are the intersecting areas, while at the bottom left is the bar plot with the biosystem categories.

According to the graph, most TPs are metabolites produced by humans or other eukaryotic organisms. On the other hand, the second greatest category is structures with missing information related to the environment they are being detected. Another observation is that all the defined categories have overlapping areas with each other except for the “plasma-based water treatment” category. A possible explanation is that an advanced oxidation technique degrades pollutants with quite different transformation mechanisms compared to metabolic reactions. It should be noted again that the method in which the categories were defined and the way the keywords are put in each category is intuitive.

4.2.3.2. Prediction of environmental microbial TPs

The prediction of TPs was carried out with the microbial environmental degradation library (based on EAWAG-PPS) of BioTransformer. The predictions were done automatically through patRoon. This reaction library was selected since it predicts microbial degradation metabolites, which might be produced during secondary wastewater treatment. While a reaction library specialized on bacterial degradation during secondary treatment has been developed by enviPath, it was impossible in this timeframe to have access through the R environment.

The isomeric SMILES, same as the input used for shinyTPs and PubChemLite, were loaded in BioTransformer. The prediction was performed in two generations. As with the collection of TPs from PubChemLite, the TPs found from the first generation were used as an input to predict the second generation.

When setting the number of generations to a low number there is a risk of not predicting TPs that might occur in real conditions [78]. The number of generations was initially set to 4, however due to combinatorial explosion the number of predictions was extremely high (~170,000). Due to lack in computational power by dealing with such a high number of TPs, the number of generations was decreased in two. Even though some TPs were missed, most of the predicted TPs are likely irrelevant because of combinatorial explosion, where many reaction rules are activated without taking into account the kinetics of the reaction [117]. Trostel et al. [78] calculated the precision of TPs prediction for 42 pharmaceutical parents, with an activated sludge degradation experiment. It was found that the precision steadily decreases for each consequent generation.

The prediction returned 14,242 TPs for 487 parent compounds out of 509 (96%). The 22 parent compounds with no predictions were either small compounds that cannot return simpler degradation products according to the model or were polyfluorinated hydrocarbons for which no reaction rules are available from EAWAG-PPS [75].

In Figure 13, are the number of predictions from each generation of the two generations, applied in the final workflow. The number of predicted TPs in the second iteration is almost 5 times higher than the number of TPs in the first iteration, while the number of TPs in the first iteration is 4 times higher compared to the number of parents. This exponential increase of TPs in each generation is a sign of combinatorial explosion.

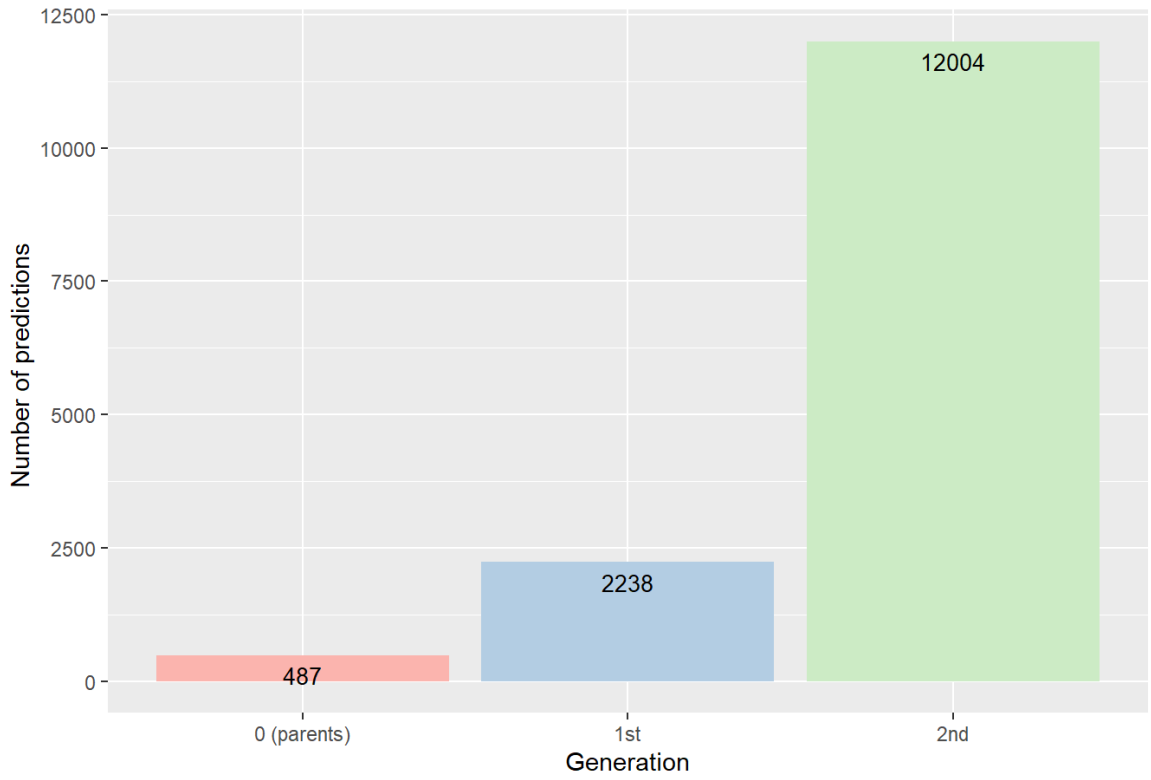


Figure 13: Number of predictions for each generation.

4.3. Removal of incorrect structures

The bacterial metabolism library from BioTransformer provides KB predictions made with empirical reaction rules found in EAWAG-BBD. This kind of models have low selectivity and can predict incorrect structures that is impossible to form. These structures made the collection of physicochemical properties in the next step of the workflow time-consuming and extremely demanding in terms of computational power. Adding to this, preserving structures in the final suspect list that are unlikely to be detected is counter-productive for the screening process.

Indeed, when importing the predicted structures in QSAR Toolbox, 13.5% of them returned an error. The problematic structures were collected and uploaded in ChemMine and selected the option to depict the structures as a validation step. The software OPERA incorporates “QSAR-ready”, a structure validation workflow which could detect some of the incorrect structures. Although “QSAR-ready” is available as an individual application it was not possible to install and use it. On the other hand, it was practically impossible to detect the faulty structures by loading the whole dataset on OPERA.

The predicted structures were loaded manually on QSAR Toolbox and the software returned the index position of the incorrect structures. This information was loaded on R and the incorrect structures were removed based on their index number. There are other workflows in the literature for chemical structures validation and standardizations [112,142] but their installation and use need advanced programming skills.

One group of erroneous structures contained at least one carbon atom with 5 valence bonds instead of 4 (Figure 14). These carbon atoms were usually part of an aromatic system. An assumption for the occurrence of these problematic predictions is limitation in the prediction model to handle aromatic groups. The imported parents with aromatic structures were written in kekule form as SMILES. It is not verified if this in silico prediction model can recognize the aromatic structures even if they are in kekule form, and if yes, what are the precise rules to define aromaticity. Aromaticity is not a well-defined concept and various cheminformatics tools might apply different rules for its definition [84].

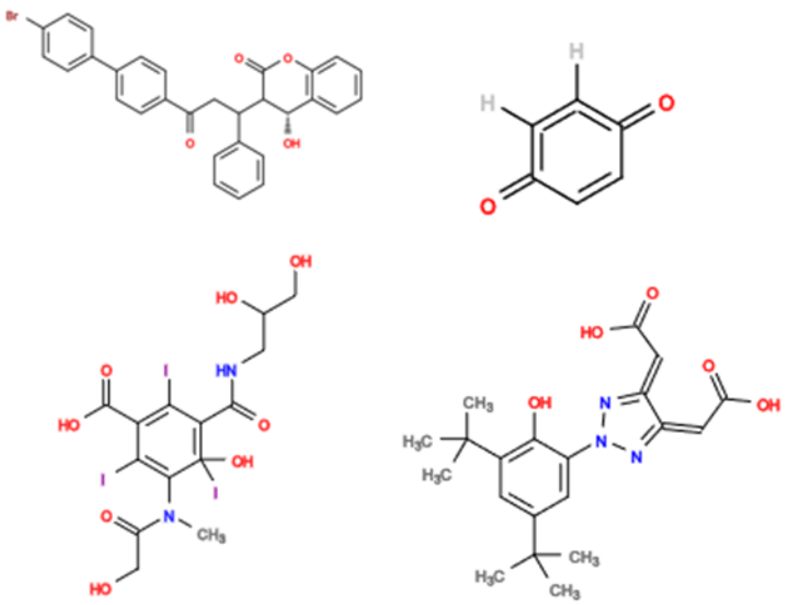


Figure14: Examples of incorrect structures with problematic carbon atoms (structural depiction source: ChemMine).

Another group of erroneous structures contained an oxygen atom with 3 valence bonds instead of 2 (figure 15).

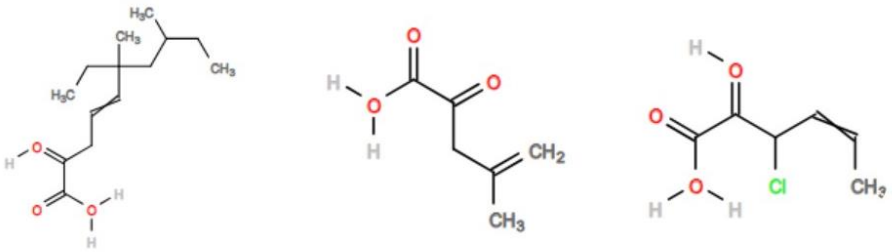
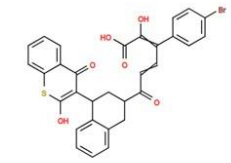
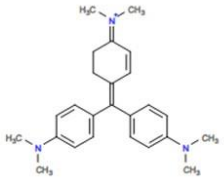


Figure15: incorrect structures with problematic oxygen atoms (structural depiction source: ChemMine).

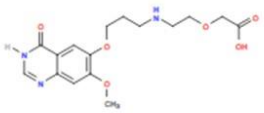
After inspecting 100 of the 773 incorrect structures with ChemMine, 3 compounds were detected (figure 16) which, although they returned an error message for incorrect valence state of a carbon atom, no issue was visible. It could not be concluded whether there is something wrong with the structures and ChemMine applies a correction to them, or that QSAR Toolbox marks them as incorrect without that being the case. They were removed with the rest of the marked structures.



C1C(Cc2ccccc2C1c1c(sc2ccccc2c1=O)O)C(=O)C=CC(=C(C(=O)O)O)c1ccc(cc1)Br



CN(C)c1ccc(cc1)C(=C1CCC(=[N+](C)C)C=C1)c1ccc(cc1)N(C)C



COc1c(cc2c(c1)nc[nH]c2=O)OCCCNCCOCC(=O)O

Figure16: Three structures marked as incorrect but with no visible incorrect valences (structural depiction source: ChemMine).

After the removal of the erroneous structures, a search was performed with the InChIKey in PubChem Compound database as an additional validation of the predictions. Matches were found for 41% of the unique predicted structures. That was a positive result, since a big portion of the predicted TPs could be successfully matched to NTA features, based on additional information found on PubChem.

4.4. Unification of the results in one list

All the parent compounds along with the TPs from the literature and BioTransformer predictions were collected in one dataset (Figure 17). The dataset included a name, chemical identifiers, if the compound is a parent, a reported or predicted TP, and if there is a PubChem compound record. Based on their InChIKey skeleton, any duplicates were removed while retaining all the relevant source information. The parents and the TPs reported in literature are considered to have a PubChem compound record by default.

In the parents and TPs datasets many small molecules and/or non-polar compounds were found which are unlikely to be detectable with LC-ESI/MS) and was deemed reasonable to remove them before obtaining the LC-ESI-MS amenability predictions.

The first step was to remove compounds with molecular mass below 50 Da, which are not detectable with spectrometry. Only 22 compounds (0.38%) were removed in this step. The next step was to remove structures without any ionizable elements, such as O, N, S and P. Subsequently, 10% of the compounds in the unified list were removed after the second step.

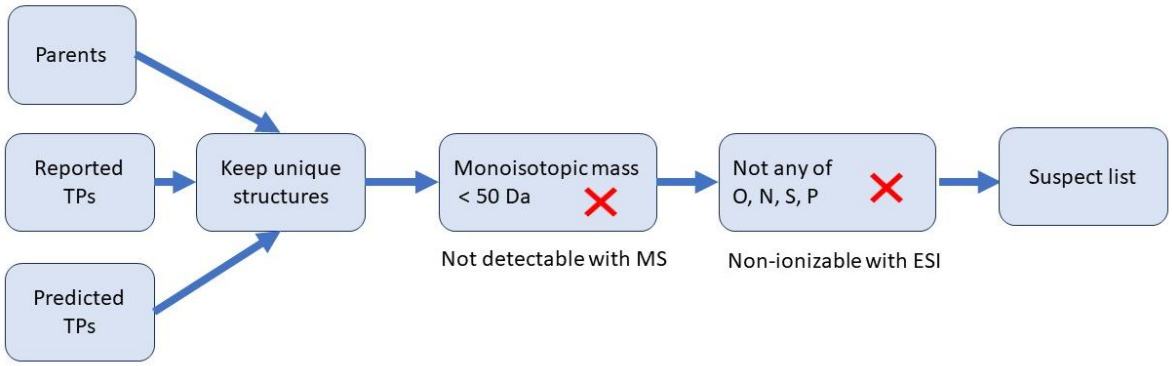


Figure 17: Steps for the formation of the suspect list with parents and TPs.

When the workflow was reviewed, it was discovered that the parent compound, 3,4-Bis(3-methylbutyl)phthalic acid, that had been collected from PubChem by CAS number search, was in its ionized form (Figure 18). Relevant data regarding this compound (SMILES, InChI, molecular formula and monoisotopic mass) were given for the ionic form, which would put an obstacle at the detection of the compound with mass spectrometry. To solve this, the values were replaced manually by the correct data of the neutral form, sourced from PubChem. The TPs in the suspect list are derived from the ionic form.

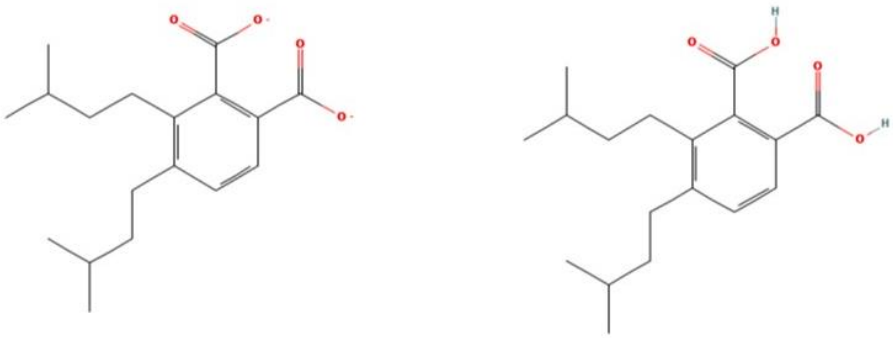


Figure 18: Ionized form (3,4-Bis(3-methylbutyl)phthalate) on the left and the neutral form (3,4-Bis(3-methylbutyl)phthalic acid) on the right (structural depiction source: PubChem).

The UpSet graph, in Figure 19, shows the different categories of compounds (parents or TPs) in the final list, along with any overlapping areas between them. It was found that 1.7% of predicted TPs were reported TPs and 1.1% were found in the parent list, summing up to 2.9%.

In total, 5004 unique structures remained in the list.

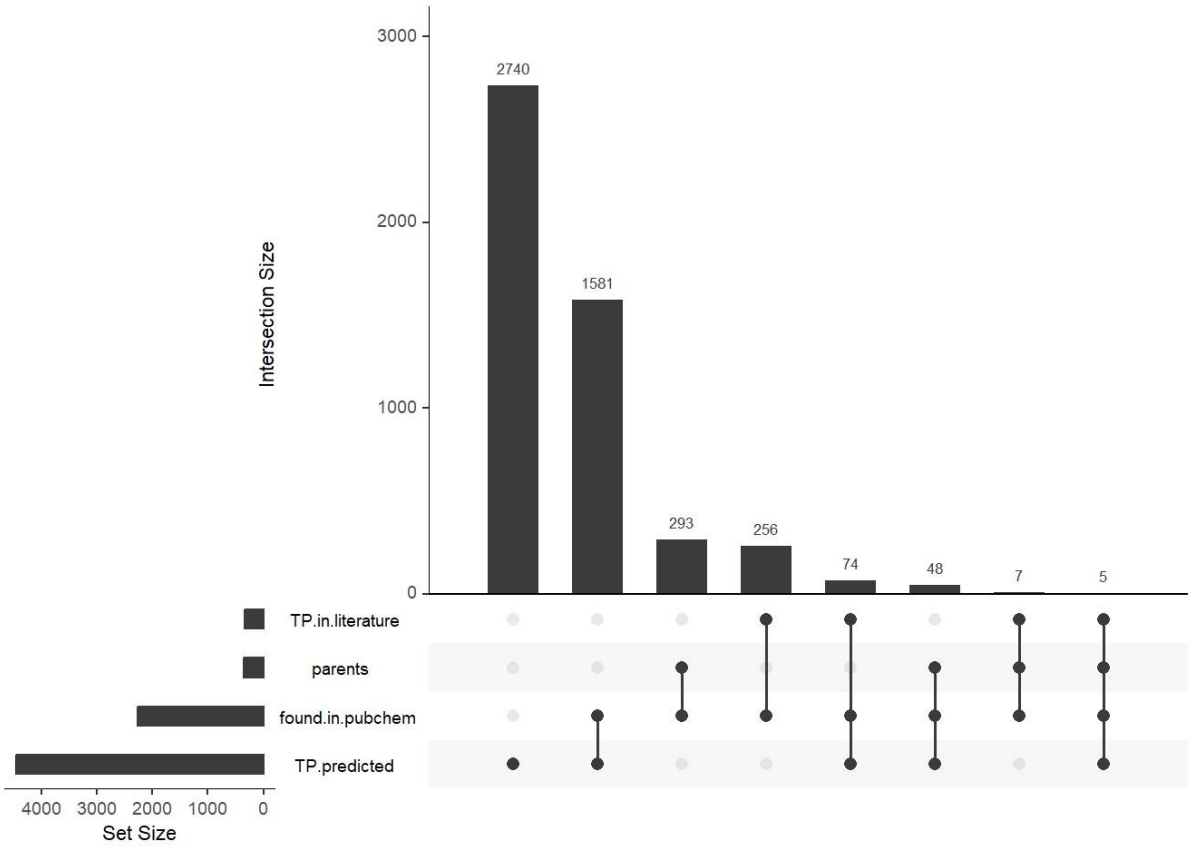


Figure 19: Overview of compounds in the unified dataset.

4.5. Physicochemical properties and LC-ESI-MS amenability predictions data

The predicted values for logarithmic octanol-water partition coefficient (logP), acid dissociation constants (pKa) and logarithmic organic carbon - water normalized sorption coefficient (logKoc) were collected from OPERA.

OPERA models were selected over other available options: models from QSAR Toolbox (Toolbox) and VEGA QSARs (VEGA) prediction software. Compared to the other mentioned software, the training and validation datasets for OPERA are open to the public. These three models from OPERA are based on the PHYSPROP database, a reportedly big database with physicochemical properties that has been open until recently. However, ambiguities and incorrect values have been pointed out in the dataset [61]. According to related publication about OPERA [112], the data used for the models have been reviewed by removing incorrect or low-quality values. Toolbox and VEGA include models from EPI Suite™ which are based on the PHYSPROP database, while for the rest of the models no information could be found. The training and validation datasets currently used in EPI Suite™ are available, but it cannot be confirmed if the latest EPI Suite™ release is identical in the models found in Toolbox and VEGA, thus it cannot be verified if the found datasets apply to them. Another advantage in OPERA is that three different metrics related to the AD of the models are provided, along with information for the five most similar compounds found in the training dataset, if needed. VEGA includes an application domain index along with the predictions, however the training datasets for the models are not available and no prediction models for pKa are provided. For all the

above reasons it was concluded that OPERA models provide the highest transparency and practicality as all the needed predictions can be collected in one step, compared to the other QSAR software.

The predictions were collected out of the R environment, through the OPERA GUI, for the 5004 compounds. A text file with the SMILES was used as an input, with enabled structure standardization, and logP, logKoc and pKa predictions were requested on the same run. The predictions were loaded in R and added to the chemicals list. It is possible to collect the predictions automatically by accessing the command line OPERA application through R, but this needs the appropriate programming knowledge.

Each prediction was tagged if it is trusted or not based on provided applicability domain metrics and guidelines in the OPERA models publication (Table 9).

Table 9: Criteria to define if the OPERA predicted values are trusted or not.

Inside the AD space (AD)	Similarity between the compound and the 5 closest neighbors (AD_index)	Is the prediction trusted? (Trusted)
No (0)	Above 0.6	Yes (TRUE)
No (0)	Below and equal to 0.6	No (FALSE)
Yes (1)	Above 0.4	Yes (TRUE)
Yes (1)	Below and equal to 0.4	No (FALSE)

In Table 10 are shown the number of missing and trusted predictions from each model:

Table 10: Number of missing and trusted predictions.

	logKoc	logP	pKa
Missing (%)	1 (0.019%)	1 (0.019%)	666[1] (13%)
Trusted (%)	3392 (68%)	4545 (91%)	1850 (43%[2])

The predicted logKoc values for all the compounds ranged from 0 to 6.5, with 2538 (50%) compounds being below 3, that classified them as mobile according to PMT criteria (Figure 20).

The logKoc values, only for the parents, had the same range as for the sum of all the compounds. However, the distribution showed a different pattern, with a lower proportion of compounds (42% of parents) having a logKoc below 3 (Figure 21). The logKoc distribution only for TPs is almost identical with the distribution for all the compounds since the suspect list contains far more TPs than parents.

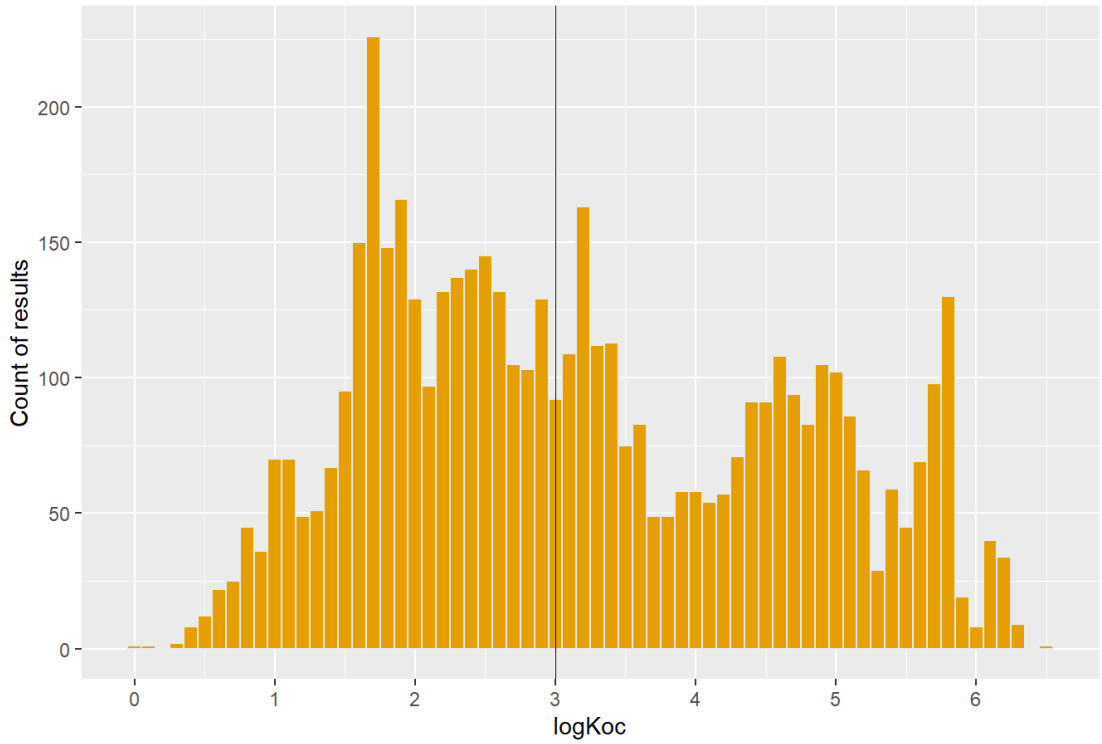


Figure 20: Distribution of predicted logKoc values. The position logKoc = 3 is marked with a line.

[1] No predictions for both higher and lower pKa values.

[2] The missing values were not taken into consideration.

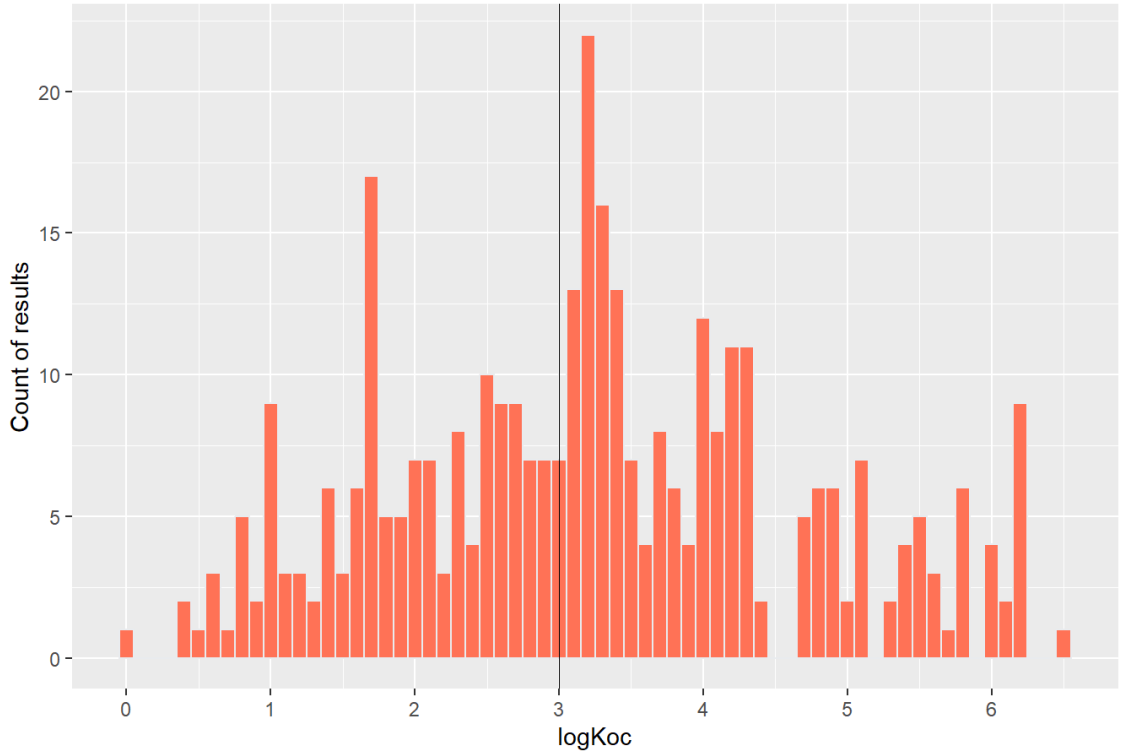


Figure 21: Distribution of predicted logKoc values only for the parents.

The predicted logP values have a range from -3.7 to 10.19 with 3804 (76%) compounds from -2 to 5 (Figure 22). The parents had a similar logP values range, with 72% of them having logP values between -2 and 5 (Figure 23).

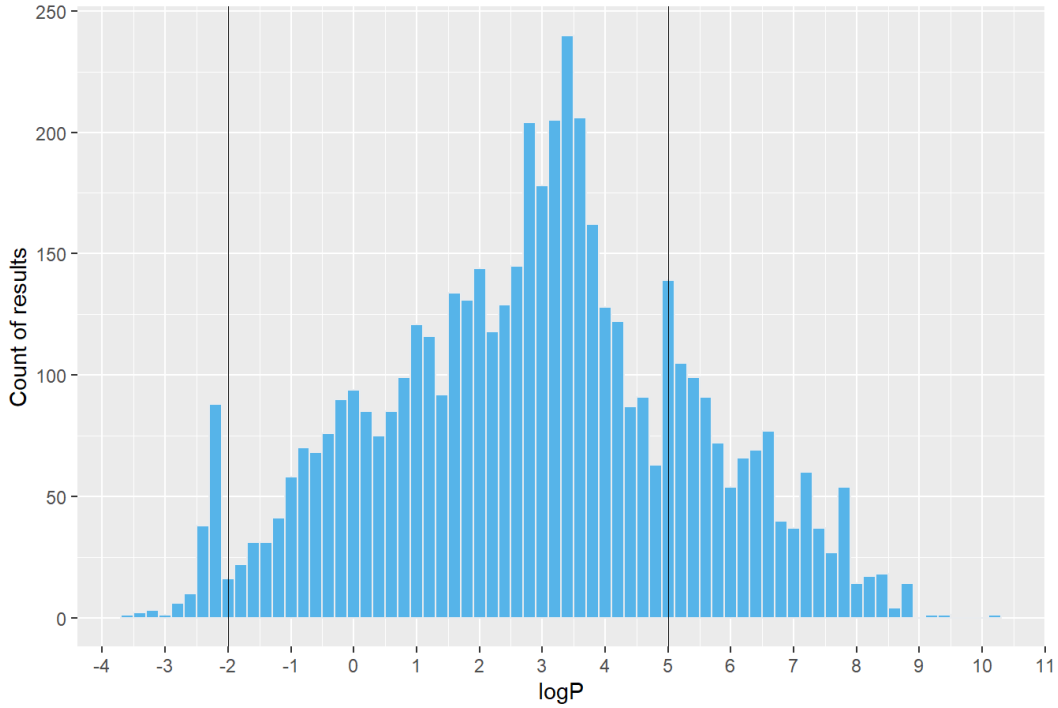


Figure 22: Distribution of predicted logP values.

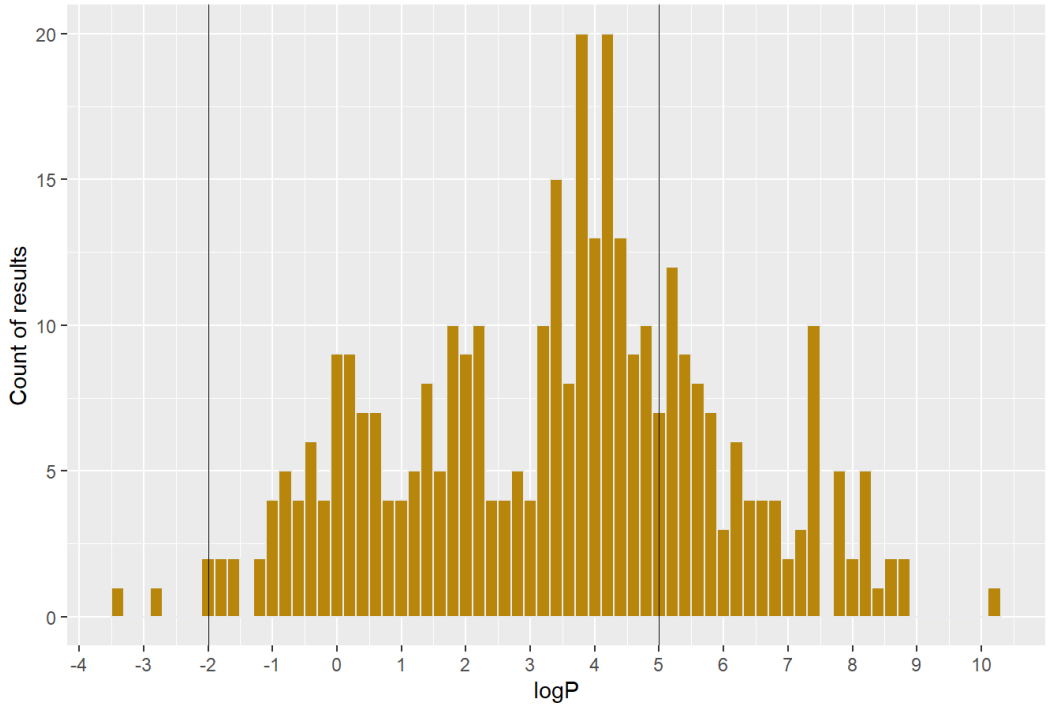


Figure 23: Distribution of predicted logP values only for the parents.

The logKoc is mainly intended describe the sorption potential of neutral lipophilic compounds in soil, sediment [150] or activated sludge [151]. Similarly, the logP is defined for neutral compounds, or ionizable compounds in pH range where they are in neutral state [152]. However, pollutants found in wastewater might have polar groups and be ionizable, thus logKoc and logP are nor appropriate to describe their environmental fate [151] and detectability [18]. For these reasons, pKa values were collected to have a better understanding of ionizable compounds in the dataset.

The dissociation constant model predicted the pKa of the most acidic and of the most basic group, if any of them are present. As mentioned in table, for 666 compounds, no pKa constants were predicted, and were categorized as non-ionizable. In table 11, are shown the number of compounds categorized based on detected acidic or basic groups:

Table 11: Overview of pKa predictions.

categories	Number of compounds (%)
No pKa defined	666 (13%)
Defined pKa of acidic group but not basic	3045 (61%)
Defined pKa of basic group but not acidic	965 (20%)
Both defined acidic and basic groups	328 (6.6%)

The octanol-water distribution coefficient (logD) is used instead of logP for ionizable compounds for the pH value of interest. The logD has also been used as an approximation of logKoc for ionizable compounds [151,153,154]. Although it has been found that logD tend to underestimate the logKoc, thus providing a conservative definition for mobility, with increased number of compounds categorized as mobile than with experimental logKoc values [155].

To approximate the mobility for the data, the normalized octanol-water distribution coefficient (logD) for pH 7 was calculated from the predicted logP and pKa values, based on the formulae in Table 12. The formulae have been derived from Ward et al [156]. In case no pKa value was predicted it was assumed that the compound is not ionizable and the logP was used instead.

Table 12: Formulae to calculate logD from logP.

No defined pKa value	$\boldsymbol{logD = logP}$
Acid (present pKa value for an acidic group but not a basic group)	$logD = logP - \log\left(1 + 10^{pH-pKa}\right)$
Base (present pKa value for a basic group but not for an acidic group)	$logD = logP - \log\left(1 + 10^{pKa-pH}\right)$
Zwitterion (Both acidic and basic groups present)	$logD = logP - \log\left(10^{pKa1-pH} + 10^{pKa2-pH}\right)$

For pH 7 it was calculated that 3125 (62%) compounds have logD below 3 (Figure 24), which is a much higher number compared to the logKoc criterion. The logD might be a more precise metric to define an ionizable compound as mobile for prioritization purposes, although the low confidence of the majority of pKa predictions should be taken into account (Table 10).

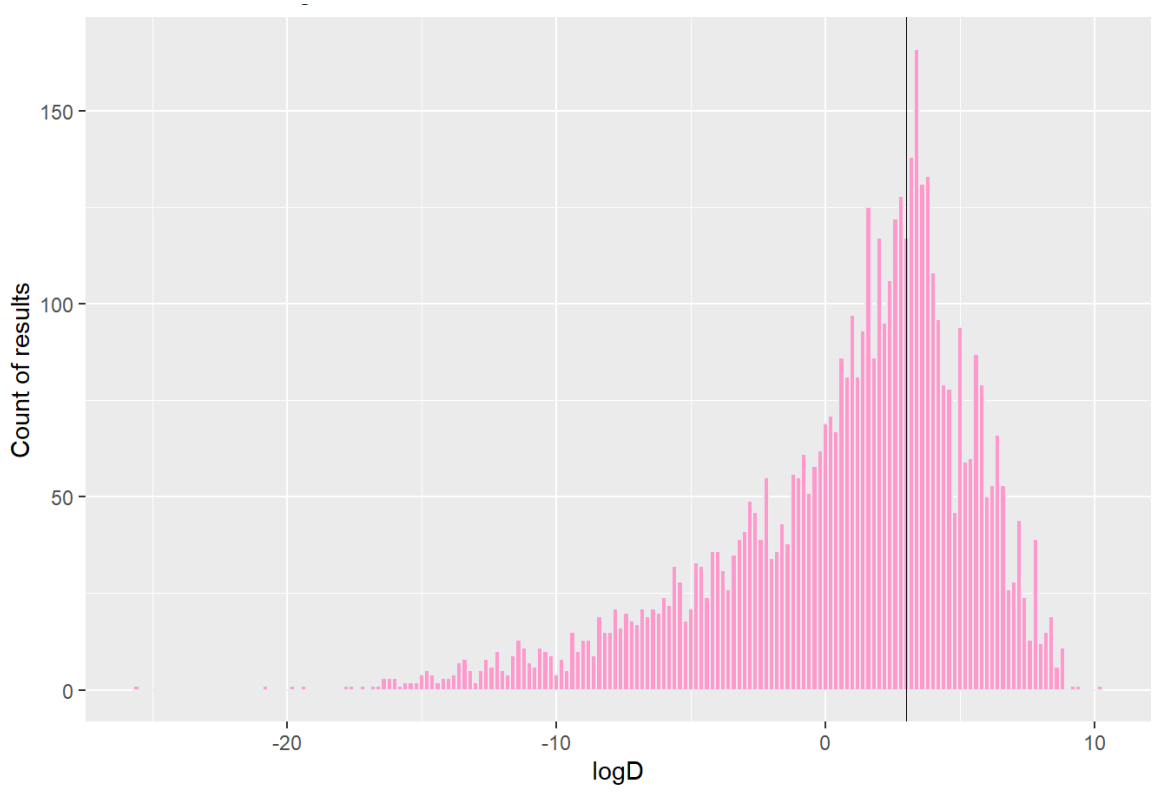


Figure 24: Distribution of logD predictions. The black vertical line intersects at the point for logD=3)

4.6. LC-ESI-MS amenability predictions

The predictions were made for an older version of the suspect list with 7235 entries. The additional compounds in the older version were duplicates, as the structures were filtered based on InChI instead of InChIKey skeleton. When it was validated that there is no variability in predictions for compounds with the same InChIKey skeleton, the predictions were added in the list.

In total, 2521 compounds were predicted as amenable in positive ionization mode and 2554 compounds as amenable in negative ionization mode. No predictions could be made for 61 unique compounds (1.2% of the suspect list), caused by problems calculating the large number of molecular fingerprints (451 in total), needed for the prediction.

Even though the ML model is highly complex in terms of the number of different implemented variables, according to its authors, the most important variable for positive ionization mode prediction was the sum of solute hydrogen bond basicity, which describes the tendency of the compound to accept hydrogens. In negative ionization mode prediction, the most important variable is related to the positions of the electronic clouds of the molecule. In both cases, the detectability prediction is related to the ionizability with ESI [58].

Figure 25 shows the density graph of the prediction probabilities of a compound being in the amenable category, in positive and negative ion mode respectively. In the density plot for positive mode, is visible a high-density area for probability score ~ 0.4, with many compounds being classified an unamenable next to the cutoff value of 0.5. This is not visible for the negative mode plot, where the predictions seem to be more evenly distributed.

The model is currently not accessible; thus, it was not possible to automate the prediction process in the workflow. An alternative to the problem would be to predict the ionization efficiency of the compounds with ESI, by calculating a small number of physicochemical properties. For example, the logP, a metric for hydrophobicity, has been chosen to describe ionization efficiency in other cases [157].

The predicted logP from the AlogP model, is one of the variables used for the amenability predictions. However, it is unknown what is the contribution of logP in the predictions since it is not mentioned in the related publication [58].

The 2D density plots between amenability probabilities and predicted logP values were made to study possible relationships for these variables.

In Figure 26, a big cluster is visible in the amenable category area for logP ~ 4 and a smaller cluster for logP -2.4. It has been mentioned that hydrophobic compounds are easily ionizable in both ionization modes while basic compounds are easily ionized in positive ionization [157,158]. Still, it is hard to find a clear pattern since multiple clusters are formed over a wide area. In Figure 27, there is less dense cluster in the amenable area for logP ~ 4, as in Figure 26. However, there is a great cluster in the unamenable category with a high logP ~ 6 for which no explanation could be found.

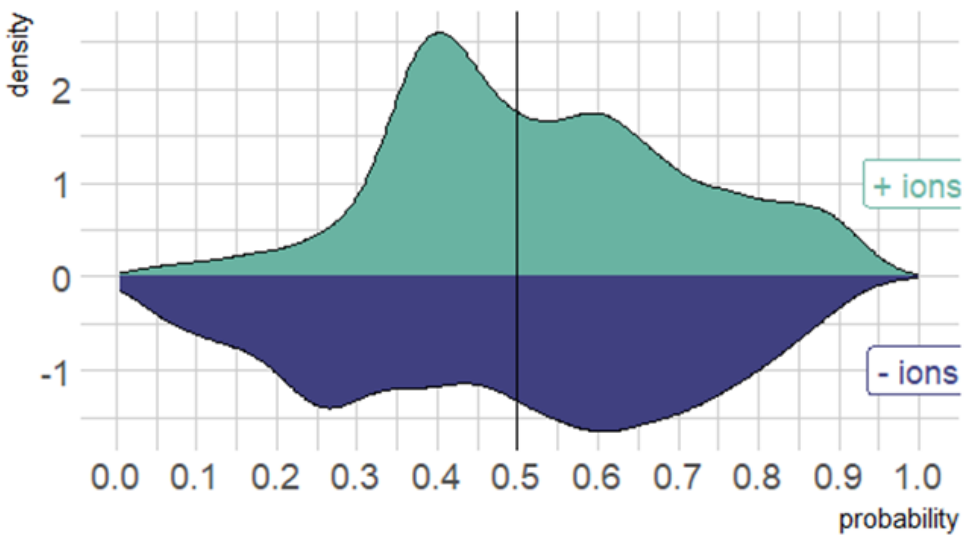


Figure 25: Density graphs for prediction probabilities of the positive ionization model (upper part), of the negative ionization mode (bottom part).

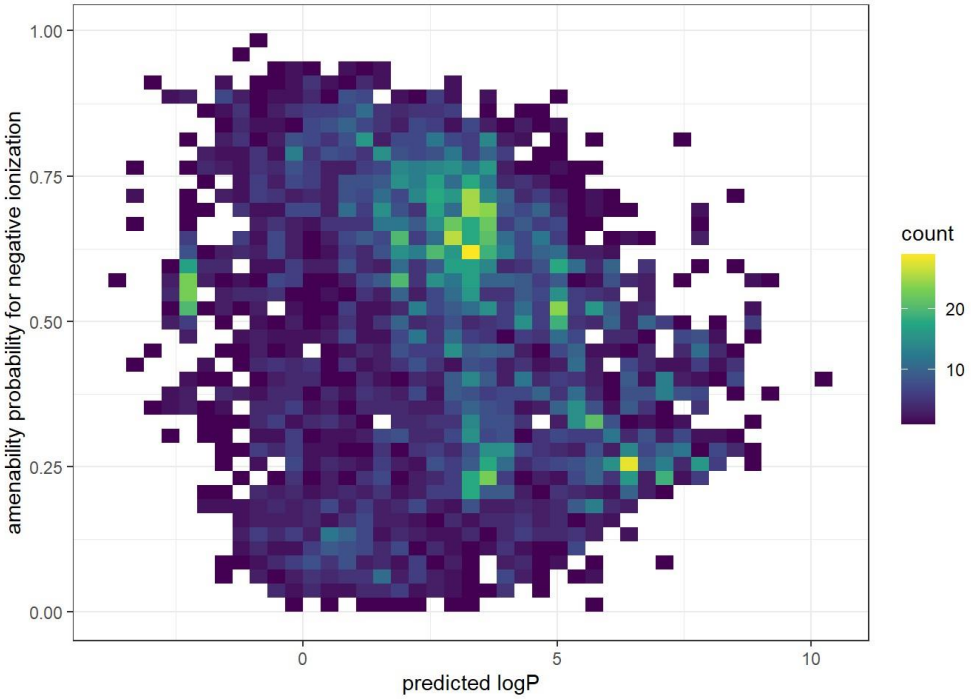


Figure 26: 2D density plot showing the relation between amenability probability and logP for negative ionization mode.

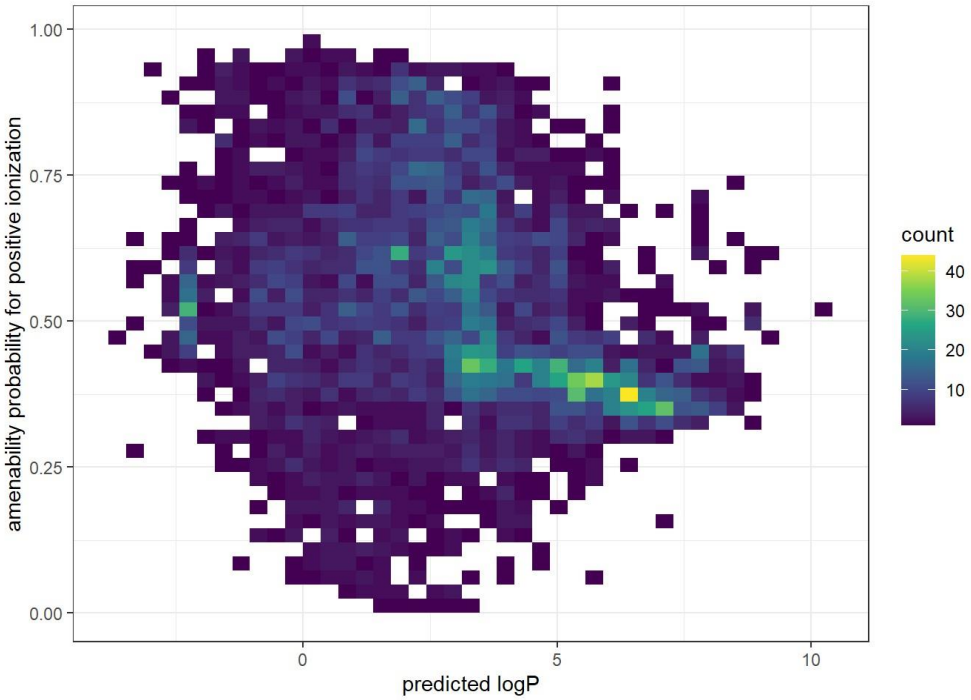


Figure 27: 2D density plot showing the relation between amenability probability and logP for negative ionization mode.

4.7. Results after prioritization

After all the relevant information were added in the suspect list for the total number of 5,004 chemical compounds, they were prioritized based on LC-ESI/MS amenability predictions, logP and logKoc. Two prioritized suspect list were created, one for each ionization mode.

After the prioritization step, the number of suspects in each list was three times lower compared to the initial suspect list.

Table 13: Number of compounds in the two final suspect lists.

Ionization mode (amenability probability > 0.5)	Number of compounds in the suspect list **(logP between -2 and 5, logKoc less than 3)**
positive	1442
negative	1313

5. Conclusions

During this research project, an R language-based workflow was developed to build a prioritized suspect list of parents and TPs.

Starting from a long and diverse list of CAS numbers, structural matches were collected from the freely accessible chemical database PubChem. Even though, PubChem has the advantage of being one of the greatest chemical databases, it could not cover the whole number of CAS numbers, either due to lack of information or difficulty of access.

A simple validation process was developed with the purpose to select only distinct organic structures. Most of the process was carried through in R environment except the salt stripping part, which was completed partly in Open Babel, - still it is possible to automate the process by integrating the Open Babel application in R. Other validation commits for chemical structures, found in the literature employ more advanced cheminformatics toolkits in other programming languages. At the end of this step, 504 parent compounds were extracted from 5,000 CAS numbers in the initial list.

TPs from the literature were collected for the selected parent compounds by using two different R-based packages: shinyTPs and PubChemLite (accessed through the R package patRoon). Both tools contain TP information that can be found in PubChem however their results were not completely identical. PubChemLite provided a greater number of results and covered almost completely the results from ShinyTPs. Another disadvantage is that there is no automated way to import the results from shinyTPs in the workflow. Each tool gave results for only 20% of the parents.

Most importantly, two generations of TPs were predicted with BioTransformer microbial environmental degradation library, based on EAWAG-PPS. As it has been noted in the literature for EAWAG-PPS, an enormous number of TPs is produced as an output, caused by the limited selectivity of the model.

The predicted compounds were manually validated with QSAR Toolbox to remove incorrect structures and consequently save computational power and decrease the number of candidates in the suspect list. It was discovered that around 10% of predicted TPs have at least one atom in the molecule with more, or less, bonds than accepted. Many of the structures, detected as incorrect, presented problems in their aromatic structure.

After all the unique remaining structures, from parents and TPs were added together, structures that are not likely detectable with ESI/MS were removed based on heuristic rules for minimum molecular mass in mass spectrometry and ionizability of the compound in ESI. In total, 10% of the structures were removed.

A few of the predicted TPs were found to be TPs from the literature or parents, but most importantly, around 40% of them have a record on PubChem, which make their identification easier.

At this point it was spontaneously found that a compound in the list was in its ionized form, instead of the expected neutral form and having incorrect molecular mass where it could not be identified in a sample. This shows that the curation steps need improvement.

Predicted values for the properties logKoc, logP and pKa were collected manually from OPERA, and the LC-ESI/MS amenability predictions were shared from the creators of the model. Based on the logKoc values, it was found that the TPs were probably categorized as mobile more often than the parent compounds. Adding to this, when the mobility was approximated with the logD in pH 7, it was noted that overall, a larger number of compounds were categorized as mobile compared to logKoc.

Two prioritized suspect lists were created based on LC-ESI/MS amenability, logKoc and logP, one for each ionization mode in mass spectrometry. After prioritization around 60% of the structures were excluded with each list containing ~1350 structures.

To conclude, the developed workflow could handle a great number of compounds and create prioritized suspect lists based on the mobility and the detectability of the compounds in LC-ESI/MS. This project gave a better understanding of the various problems of curating a suspect list with *in silico* tools and a large number of compounds, and thus the implemented solutions can be applied for the creation of different suspect lists.

6. Future Perspectives

The next step after the creation of the suspect lists is to test them in real NTA data from wastewater effluent samples. It would be also interesting to create suspect lists were the mobility criterion is based on logD and compare the number of identified compounds in samples between the two set of lists.

The workflow can be further improved by making more steps automated, for example by accessing the external software through the R language environment. By automating the structure curation steps, it will result in fewer mistakes. For example, by automating the neutralization process it is possible to apply this step to the compounds as a preventing measure and avoid the presence of structures with incorrect molecular mass in the suspect list. Another problem that was found during the project is the absence of a free and user-friendly curation algorithm to identify incorrect structures that might be produced by TPs prediction models.

References

(1) Tudi, M.; Ruan, H. D.; Wang, L.; Lyu, J.; Sadler, R.; Connell, D.; Chu, C.; Phung, D. T. Agriculture Development, Pesticide Application and Its Impact on the Environment. *Int J Environ Res Public Health* **2021**, *18* (3), 1–24. https://doi.org/10.3390/ijerph18031112.

(2) European Federation of Pharmaceutical and Associations. The Pharmaceutical Industry in Figures. *European Federation of Pharmaceutical and Associations* **2016**.

(3) UN Environment. Benefits of Chemicals Control. **2020**.

(4) European Environmental Agency. How Pesticides Impact Human Health and Ecosystems in Europe. **2023**, 1–28.

(5) Brodin, T.; Fick, J.; Jonsson, M.; Klaminder, J. Dilute Concentrations of a Psychiatric Drug Alter Behavior of Fish from Natural Populations. *Science (1979)* **2013**, *339* (6121), 814–815. https://doi.org/10.1126/science.1226850.

(6) Horký, P.; Grabic, R.; Grabicová, K.; Brooks, B. W.; Douda, K.; Slavik, O.; Hubená, P.; Santos, E. M. S.; Randák, T. Methamphetamine Pollution Elicits Addiction in Wild Fish. *Journal of Experimental Biology* **2021**, *224* (13). https://doi.org/10.1242/jeb.242145.

(7) Schulz, R.; Bub, S.; Petschick, L. L.; Stehle, S.; Wolfram, J. Applied Pesticide Toxicity Shifts toward Plants and Invertebrates, Even in GM Crops. *Science (1979)* **2021**, *372* (6537), 81–84. https://doi.org/10.1126/science.abe1148.

(8) White, S. S.; Birnbaum, L. S. An Overview of the Effects of Dioxins and Dioxin-like Compounds on Vertebrates, as Documented in Human and Ecological Epidemiology. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* **2010**, No. 919, 1–12. https://doi.org/10.1080/10590500903310047.An.

(9) Bonato, M.; Corrà, F.; Bellio, M.; Guidolin, L.; Tallandini, L.; Irato, P.; Santovito, G. Pfas Environmental Pollution and Antioxidant Responses: An Overview of the Impact on Human Field. *Int J Environ Res Public Health* **2020**, *17* (21), 1–45. https://doi.org/10.3390/ijerph17218020.

(10) Hu, D.; Henderson, K.; Coats, J. Fate of Transformation Products of Synthetic Chemicals. In *Handbook of Environmental Chemistry, Volume 2: Reactions and Processes*; 2009; Vol. 2 P, pp 103–120. https://doi.org/10.1007/698_2_018.

(11) Boxall, A. B. A.; Sinclair, C. J.; Fenner, K.; KOLPIN, D.; MAUND, S. J. When Synthetic Chemicals Degrade in the Environment. In *Environmental Science & Technology*; American Chemical Society, 2004; pp 368–375. https://doi.org/10.1201/9781482286809-37.

(12) Suman, T. Y.; Kim, S. Y.; Yeom, D. H.; Jeon, J. Transformation Products of Emerging Pollutants Explored Using Non-Target Screening: Perspective in the Transformation Pathway and Toxicity Mechanism—A Review. *Toxics*. MDPI February 1, 2022. https://doi.org/10.3390/toxics10020054.

(13) Drewes, J. E.; Letzel, T. Chemicals of Emerging Concern and Their Transformation Products in the Aqueous Environment. *ACS Symposium Series* **2016**, *1241*, 3–9. https://doi.org/10.1021/bk-2016-1241.ch001.

(14) Halling-Sørensen, B.; Sengeløv, G.; Tjørnelund, J. Toxicity of Tetracyclines and Tetracycline Degradation Products to Environmentally Relevant Bacteria, Including Selected Tetracycline-Resistant Bacteria. *Arch Environ Contam Toxicol* **2002**, *42* (3), 263–271. https://doi.org/10.1007/s00244-001-0017-2.

(15) Sinclair, C. J.; Boxall, A. B. A. Assessing the Ecotoxicity of Pesticide Transformation Products. *Environ Sci Technol* **2003**, *37* (20), 4617–4625. https://doi.org/10.1021/es030038m.

(16) *MolView v2.4*. https://molview.org/.

(17) European Parliament. *Regulation (EC) No 1272/2008*; 2023; Vol. 10, pp 1–21.

(18) Reemtsma, T.; Berger, U.; Arp, H. P. H.; Gallard, H.; Knepper, T. P.; Neumann, M.; Quintana, J. B.; Voogt, P. De. Mind the Gap: Persistent and Mobile Organic Compounds—Water Contaminants That Slip Through. *Environ Sci Technol* **2016**, *50* (19), 10308–10315. https://doi.org/10.1021/acs.est.6b03338.

(19) Arp, H. P. H.; Hale, S. E. Assessing the Persistence and Mobility of Organic Substances to Protect Freshwater Resources. *ACS Environmental Au* **2022**, *2* (6), 482–509. https://doi.org/10.1021/acsenvironau.2c00024.

(20) Hale, S. E.; Neumann, M.; Schliebner, I.; Schulze, J.; Averbeck, F. S.; Castell-Exner, C.; Collard, M.; Drmač, D.; Hartmann, J.; Hofman-Caris, R.; Hollender, J.; de Jonge, M.; Kullick, T.; Lennquist, A.; Letzel, T.; Nödler, K.; Pawlowski, S.; Reineke, N.; Rorije, E.; Scheurer, M.; Sigmund, G.; Timmer, H.; Trier, X.; Verbruggen, E.; Arp, H. P. H. Getting in Control of Persistent, Mobile and Toxic (PMT) and Very Persistent and Very Mobile (VPvM) Substances to Protect Water Resources: Strategies from Diverse Perspectives. *Environ Sci Eur* **2022**, *34* (1). https://doi.org/10.1186/s12302-022-00604-4.

(21) Boxall, A. B. A. The Environmental Side Effects of Medication: How Are Human and Veterinary Medicines in Soils and Water Bodies Affecting Human and Environmental Health? *EMBO Rep* **2004**, *5* (12), 1110–1116.

(22) Ahmed, J.; Thakur, A.; Goyal, A. Industrial Wastewater and Its Toxic Effects. *Biological Treatment of Industrial Wastewater* **2021**, No. 5, 1–14. https://doi.org/10.1039/9781839165399-00001.

(23) States, U. Wastewater Treatment Works... The Basics. **1998**, No. May.

(24) Ruff, M.; Mueller, M. S.; Loos, M.; Singer, H. P. Quantitative Target and Systematic Non-Target Analysis of Polar Organic Micro-Pollutants along the River Rhine Using High-Resolution Mass-Spectrometry - Identification of Unknown Sources and Compounds. *Water Res* **2015**, *87*, 145–154. https://doi.org/10.1016/j.watres.2015.09.017.

(25) Hug, C.; Ulrich, N.; Schulze, T.; Brack, W.; Krauss, M. Identification of Novel Micropollutants in Wastewater by a Combination of Suspect and Nontarget Screening. *Environmental Pollution* **2014**, *184*, 25–32. https://doi.org/10.1016/j.envpol.2013.07.048.

(26) Golovko, O.; Rehrl, A. L.; Köhler, S.; Ahrens, L. Organic Micropollutants in Water and Sediment from Lake Mälaren, Sweden. *Chemosphere* **2020**, *258*. https://doi.org/10.1016/j.chemosphere.2020.127293.

(27) Feo, M. L.; Bagnati, R.; Passoni, A.; Riva, F.; Salvagio Manta, D.; Sprovieri, M.; Traina, A.; Zuccato, E.; Castiglioni, S. Pharmaceuticals and Other Contaminants in Waters and Sediments from Augusta Bay (Southern Italy). *Science of the Total Environment* **2020**, *739*, 139827. https://doi.org/10.1016/j.scitotenv.2020.139827.

(28) Kern, S.; Baumgartner, R.; Helbling, D. E.; Hollender, J.; Singer, H.; Loos, M. J.; Schwarzenbach, R. P.; Fenner, K. A Tiered Procedure for Assessing the Formation of Biotransformation Products of Pharmaceuticals and Biocides during Activated Sludge Treatment. *Journal of Environmental Monitoring* **2010**, *12* (11), 2100–2111. https://doi.org/10.1039/c0em00238k.

(29) Schollée, J. E.; Schymanski, E. L.; Avak, S. E.; Loos, M.; Hollender, J. Prioritizing Unknown Transformation Products from Biologically-Treated Wastewater Using High-Resolution Mass Spectrometry, Multivariate Statistics, and Metabolic Logic. *Anal Chem* **2015**, *87* (24), 12121–12129. https://doi.org/10.1021/acs.analchem.5b02905.

(30) Gulde, R.; Rutsch, M.; Clerc, B.; Schollée, J. E.; von Gunten, U.; McArdell, C. S. Formation of Transformation Products during Ozonation of Secondary Wastewater Effluent and Their Fate in Post-Treatment: From Laboratory- to Full-Scale. *Water Res* **2021**, *200*. https://doi.org/10.1016/j.watres.2021.117200.

(31) Rodrigues, P.; Oliva-Teles, L.; Guimarães, L.; Carvalho, A. P. *Occurrence of Pharmaceutical and Pesticide Transformation Products in Freshwater: Update on Environmental Levels, Toxicological Information and Future Challenges*; Springer International Publishing, 2022; Vol. 260. https://doi.org/10.1007/s44169-022-00014-w.

(32) Eysseric, E.; Gagnon, C.; Segura, P. A. Identifying Congeners and Transformation Products of Organic Contaminants within Complex Chemical Mixtures in Impacted Surface Waters with a Top-down Non-Targeted Screening Workflow. *Science of the Total Environment* **2022**, *822*, 153540. https://doi.org/10.1016/j.scitotenv.2022.153540.

(33) Li, Z.; Sobek, A.; Radke, M. Fate of Pharmaceuticals and Their Transformation Products in Four Small European Rivers Receiving Treated Wastewater. *Environ Sci Technol* **2016**, *50* (11), 5614–5621. https://doi.org/10.1021/acs.est.5b06327.

(34) Henze, M.; van Loosdrecht, M. C. M.; Ekama, G. A.; Brdjanovic, D. *Biological Wastewater Treatment: Principles, Modelling and Design*; 2008. https://doi.org/10.2166/9781780401867.

(35) Morgenroth, E.; Kommedal, R.; Harremoës, P. Processes and Modeling of Hydrolysis of Particulate Organic Matter in Aerobic Wastewater Treatment - A Review. *Water Science and Technology* **2002**, *45* (6), 25–40. https://doi.org/10.2166/wst.2002.0091.

(36) Kennes-Veiga, D. M.; Gónzalez-Gil, L.; Carballa, M.; Lema, J. M. Enzymatic Cometabolic Biotransformation of Organic Micropollutants in Wastewater Treatment Plants: A Review. *Bioresour Technol* **2022**, *344*. https://doi.org/10.1016/j.biortech.2021.126291.

(37) Fischer, K.; Majewsky, M. Cometabolic Degradation of Organic Wastewater Micropollutants by Activated Sludge and Sludge-Inherent Microorganisms. *Appl Microbiol Biotechnol* **2014**, *98* (15), 6583–6597. https://doi.org/10.1007/s00253-014-5826-0.

(38) Su, Q.; Schittich, A. R.; Jensen, M. M.; Ng, H.; Smets, B. F. Role of Ammonia Oxidation in Organic Micropollutant Transformation during Wastewater Treatment: Insights from Molecular, Cellular, and Community Level Observations. *Environ Sci Technol* **2021**, *55* (4), 2173–2188. https://doi.org/10.1021/acs.est.0c06466.

(39) Luo, Y.; Guo, W.; Ngo, H. H.; Nghiem, L. D.; Hai, F. I.; Zhang, J.; Liang, S.; Wang, X. C. A Review on the Occurrence of Micropollutants in the Aquatic Environment and Their Fate and Removal during Wastewater Treatment. *Science of the Total Environment*. Elsevier B.V. March 1, 2014, pp 619–641. https://doi.org/10.1016/j.scitotenv.2013.12.065.

(40) Margot, J.; Rossi, L.; Barry, D. A.; Holliger, C. A Review of the Fate of Micropollutants in Wastewater Treatment Plants. *Wiley Interdisciplinary Reviews: Water* **2015**, *2* (5), 457–487. https://doi.org/10.1002/WAT2.1090.

(41) Kools, S. A. E. Large Scale Water Treatment and the Implications for the Water Cycle. **2018**, No. January, 28.

(42) Europa. Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 Concerning the Placing of Plant Protection Products on the Market and Repealing Council Directives 79/117/EEC. *Official Journal of the European Union* **2022**, *10*, 1–50.

(43) European Parliament. *Urban Wastewater Treatment Directive*; 2023; Vol. 0355. https://www.europarl.europa.eu/doceo/document/TA-9-2023-0355_EN.pdf.

(44) European Parliament. *Urban Wastewater Treatment. Recast*. https://oeil.secure.europarl.europa.eu/oeil/popups/summary.do?id=1722263&t=e&l=en.

(45) Schymanski, E. L.; Singer, H. P.; Longrée, P.; Loos, M.; Ruff, M.; Stravs, M. A.; Ripollés Vidal, C.; Hollender, J. Strategies to Characterize Polar Organic Contamination in Wastewater: Exploring the Capability of High Resolution Mass Spectrometry. *Environ Sci Technol* **2014**, *48* (3), 1811–1818. https://doi.org/10.1021/es4044374.

(46) Schulz, M.; Löffler, D.; Wagner, M.; Ternes, T. A. Transformation of the X-Ray Contrast Medium Iopromide in Soil and Biological Wastewater Treatment. *Environ Sci Technol* **2008**, *42* (19), 7207–7217. https://doi.org/10.1021/es800789r.

(47) Krauss, M.; Singer, H.; Hollender, J. LC-High Resolution MS in Environmental Analysis: From Target Screening to the Identification of Unknowns. *Anal Bioanal Chem* **2010**, *397* (3), 943–951. https://doi.org/10.1007/s00216-010-3608-9.

(48) Kiefer, K. ETH Library Polar Micropollutants and Their Transformation Products in Groundwater: Identification with LC-HRMS and Their Abatement in Water Treatment. https://doi.org/10.3929/ethz-b-000481187.

(49) Renner, G.; Reuschenbach, M. Critical Review on Data Processing Algorithms in Non-Target Screening: Challenges and Opportunities to Improve Result Comparability. *Anal Bioanal Chem* **2023**, *415* (18), 4111–4123. https://doi.org/10.1007/s00216-023-04776-7.

(50) Brunner, A. M.; Vughs, D.; Siegers, W.; Bertelkamp, C.; Hofman-Caris, R.; Kolkman, A.; ter Laak, T. Monitoring Transformation Product Formation in the Drinking Water Treatments Rapid Sand Filtration and Ozonation. *Chemosphere* **2019**, *214*, 801–811. https://doi.org/10.1016/j.chemosphere.2018.09.140.

(51) Helmus, R.; ter Laak, T. L.; van Wezel, A. P.; de Voogt, P.; Schymanski, E. L. PatRoon: Open Source Software Platform for Environmental Mass Spectrometry Based Non-Target Screening. *J Cheminform* **2021**, *13* (1). https://doi.org/10.1186/s13321-020-00477-w.

(52) Been, F.; Kruve, A.; Vughs, D.; Meekel, N.; Reus, A.; Zwartsen, A.; Wessel, A.; Fischer, A.; ter Laak, T.; Brunner, A. M. Risk-Based Prioritization of Suspects Detected in Riverine Water Using Complementary Chromatographic Techniques. *Water Res* **2021**, *204*. https://doi.org/10.1016/j.watres.2021.117612.

(53) Koronaiou, L. A.; Nannou, C.; Evgenidou, E.; Panagopoulos Abrahamsson, D.; Lambropoulou, D. A. Photo-Assisted Transformation of Furosemide: Exploring Transformation Pathways, Structure Database and Suspect and Non-Target Workflows for Comprehensive Screening of Unknown Transformation Products in Wastewaters and Landfill Leachates. *Science of the Total Environment* **2023**, *904* (August), 166599. https://doi.org/10.1016/j.scitotenv.2023.166599.

(54) Günthardt, B. F.; Schönsee, C. D.; Hollender, J.; Hungerbühler, K.; Scheringer, M.; Bucheli, T. D. “Is There Anybody Else out There?” - First Insights from a Suspect Screening for Phytotoxins in Surface Water. *Chimia (Aarau)* **2020**, *74* (3), 129–135. https://doi.org/10.2533/CHIMIA.2020.129.

(55) Getzinger, G. J.; Higgins, C. P.; Ferguson, P. L. Structure Database and In Silico Spectral Library for Comprehensive Suspect Screening of Per- and Polyfluoroalkyl Substances (PFASs) in Environmental Media by High-Resolution Mass Spectrometry. *Anal Chem* **2021**, *93* (5), 2820–2827. https://doi.org/10.1021/acs.analchem.0c04109.

(56) Sjerps, R. M. A.; Vughs, D.; van Leerdam, J. A.; ter Laak, T. L.; van Wezel, A. P. Data-Driven Prioritization of Chemicals for Various Water Types Using Suspect Screening LC-HRMS. *Water Res* **2016**, *93*, 254–264. https://doi.org/10.1016/j.watres.2016.02.034.

(57) Herwerden, D. Van; Nikolopoulos, A.; Barron, L. P.; Jake, W.; Brien, O.; Pirok, B. W. J.; Thomas, K. V; Samanipour, S. Exploring the Chemical Subspace of RPLC : A Data Driven Approach. 1–27.

(58) Lowe, C. N.; Isaacs, K. K.; McEachran, A.; Grulke, C. M.; Sobus, J. R.; Ulrich, E. M.; Richard, A.; Chao, A.; Wambaugh, J.; Williams, A. J. Predicting Compound Amenability with Liquid

Chromatography-Mass Spectrometry to Improve Non-Targeted Analysis. *Anal Bioanal Chem* **2021**, *413* (30), 7495–7508. https://doi.org/10.1007/s00216-021-03713-w.

(59) Alygizakis, N.; Konstantakos, V.; Bouziotopoulos, G.; Kormentzas, E. A Multi-Label Classifier for Predicting the Most Appropriate Instrumental Method for the Analysis of Contaminants of Emerging Concern. **2022**.

(60) Li, L.; Zhang, Z.; Men, Y.; Baskaran, S.; Sangion, A.; Wang, S.; Arnot, J. A.; Wania, F. Retrieval, Selection, and Evaluation of Chemical Property Data for Assessments of Chemical Emissions, Fate, Hazard, Exposure, and Risks. *ACS Environmental Au* **2022**, *2* (5), 376–395. https://doi.org/10.1021/acsenvironau.2c00010.

(61) Bloch, D. E. Review of PHYSPROP Database (Version 1.0). *J Chem Inf Comput Sci* **1995**, *35* (2), 328–329. https://doi.org/10.1021/ci00024a602.

(62) Dearden John; Worth Andrew. In Silico Prediction of Physicochemical Properties. *Joint European Commission Report EUR_23051_EN* **2007**, No. August.

(63) Osman, N. A. Statistical Methods for in Silico Tools Used for Risk Assessment and Toxicology. *In Silico Chemistry and Biology: Current and Future Prospects* **2022**, 157–170. https://doi.org/10.1515/9783110493955-009.

(64) Roy, K.; Kar, S.; Das, R. N. *A Primer on QSAR/QSPR Modeling: Fundamental Concepts (Springer Briefs in Molecular Sciences)*; 2015.

(65) Héberger, K. Selection of Optimal Validation Methods for Quantitative Structure–Activity Relationships and Applicability Domain. *SAR QSAR Environ Res* **2023**, *34* (5), 415–434. https://doi.org/10.1080/1062936X.2023.2214871.

(66) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17* (5), 4791–4810. https://doi.org/10.3390/molecules17054791.

(67) Aniceto, N.; Freitas, A. A.; Bender, A.; Ghafourian, T. A Novel Applicability Domain Technique for Mapping Predictive Reliability across the Chemical Space of a QSAR : Reliability - Density Neighbourhood. *J Cheminform* **2016**, 1–20. https://doi.org/10.1186/s13321-016-0182-y.

(68) Danieli, A.; Colombo, E.; Raitano, G.; Lombardo, A.; Roncaglioni, A.; Manganaro, A.; Sommovigo, A.; Carnesecchi, E.; Dorne, J. C. M.; Benfenati, E. The VEGA Tool to Check the Applicability Domain Gives Greater Confidence in the Prediction of In Silico Models. **2023**.

(69) Peter, S. C.; Dhanjal, J. K.; Malik, V.; Radhakrishnan, N.; Jayakanthan, M.; Sundar, D.; Sundar, D. *Quantitative Structure-Activity Relationship (QSAR): Modeling Approaches to Biological Applications*; Elsevier Ltd., 2018; Vol. 1–3. https://doi.org/10.1016/B978-0-12-809633-8.20197-0.

(70) Liu, P.; Long, W. Current Mathematical Methods Used in QSAR/QSPR Studies. *Int J Mol Sci* **2009**, *10* (5), 1978–1998. https://doi.org/10.3390/ijms10051978.

(71) *K-Nearest Neighbors Algorithm*. https://www.ibm.com/topics/knn#:~:text=Next steps-,K-Nearest Neighbors Algorithm,of an individual data point.

(72) Lee, K.; Lee, M.; Kim, D. Utilizing Random Forest QSAR Models with Optimized Parameters for Target Identification and Its Application to Target-Fishing Server. *BMC Bioinformatics* **2017**, *18* (Suppl 16). https://doi.org/10.1186/s12859-017-1960-x.

(73) Djoumbou-Feunang, Y.; Fiamoncini, J.; Gil-de-la-Fuente, A.; Greiner, R.; Manach, C.; Wishart, D. S. BioTransformer: A Comprehensive Computational Tool for Small Molecule Metabolism Prediction and Metabolite Identification. *J Cheminform* **2019**, *11* (1), 1–25. https://doi.org/10.1186/s13321-018-0324-5.

(74) Tam, J. Y. C.; Lorsbach, T.; Schmidt, S.; Wicker, J. S. Holistic Evaluation of Biodegradation Pathway Prediction: Assessing Multi-Step Reactions and Intermediate Products. *J Cheminform* **2021**, *13* (1). https://doi.org/10.1186/s13321-021-00543-x.

(75) Ellis, L. B. M.; Gao, J.; Fenner, K.; Wackett, L. P. The University of Minnesota Pathway Prediction System: Predicting Metabolic Logic. *Nucleic Acids Res* **2008**, *36* (Web Server issue), 427–432. https://doi.org/10.1093/nar/gkn315.

(76) Feng, F.; Lai, L.; Pei, J. Computational Chemical Synthesis Analysis and Pathway Design. *Front Chem* **2018**, *6* (JUN). https://doi.org/10.3389/fchem.2018.00199.

(77) Latino, D. A. R. S.; Wicker, J.; Gütlein, M.; Schmid, E.; Kramer, S.; Fenner, K. Eawag-Soil in EnviPath: A New Resource for Exploring Regulatory Pesticide Soil Biodegradation Pathways and Half-Life Data. *Environ Sci Process Impacts* **2017**, *19* (3), 449–464. https://doi.org/10.1039/c6em00697c.

(78) Trostel, L.; Coll, C.; Fenner, K.; Hafner, J. Combining Predictive and Analytical Methods to Elucidate Pharmaceutical Biotransformation in Activated Sludge. *Environ Sci Process Impacts* **2023**, *25* (8), 1322–1336. https://doi.org/10.1039/d3em00161j.

(79) Warr, W. A. Representation of Chemical Structures. *Wiley Interdiscip Rev Comput Mol Sci* **2011**, *1* (4), 557–579. https://doi.org/10.1002/wcms.36.

(80) American Chemical Society. Naming and Indexing of Chemical Substances for Chemical Abstracts TM 2007 Edition. *Naming and Indexing of Chemical Substances for Chemical Abstracts 2007* **2007**, No. 2006, 145–157.

(81) CAS Common Chemistry. *Sodium borate pentahydrate*. https://commonchemistry.cas.org/detail?cas_rn=11130-12-4 (accessed 2023-12-15).

(82) CAS Common Chemistry. *Sodium tetraborate pentahydrate*. https://commonchemistry.cas.org/detail?cas_rn=12179-04-3 (accessed 2023-12-15).

(83) O'Boyle, N. M. Towards a Universal SMILES Representation - A Standard Method to Generate Canonical SMILES Based on the InChI. *J Cheminform* **2012**, *4* (9), 1–14. https://doi.org/10.1186/1758-2946-4-22.

(84) Hähnke, V. D.; Kim, S.; Bolton, E. E. PubChem Chemical Structure Standardization. *J Cheminform* **2018**, *10* (1), 1–40. https://doi.org/10.1186/s13321-018-0293-8.

(85) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J Chem Inf Comput Sci* **1989**, *29* (2), 97–101. https://doi.org/10.1021/ci00062a008.

(86) *SMILES tutorial*. https://www.daylight.com/dayhtml_tutorials/languages/smiles/index.html.

(87) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. *InChI, the IUPAC International Chemical Identifier*; Journal of Cheminformatics, 2015; Vol. 7. https://doi.org/10.1186/s13321-015-0068-4.

(88) Karapetyan, K.; Batchelor, C.; Sharpe, D.; Tkachenko, V.; Williams, A. J. The Chemical Validation and Standardization Platform (CVSP): Large-Scale Automated Validation of

Chemical Structure Datasets. *J Cheminform* **2015**, *7* (1), 1–13. https://doi.org/10.1186/s13321-015-0072-8.

(89) Wishart, D. S.; Tian, S.; Allen, D.; Oler, E.; Peters, H.; Lui, V. W.; Gautam, V.; Djoumbou-Feunang, Y.; Greiner, R.; Metz, T. O. BioTransformer 3.0 - a Web Server for Accurately Predicting Metabolic Transformation Products. *Nucleic Acids Res* **2022**, *50* (W1), W115–W123. https://doi.org/10.1093/nar/gkac313.

(90) Dimitrov, S. D.; Diderich, R.; Sobanski, T.; Pavlov, T. S.; Chankov, G. V.; Chapkanov, A. S.; Karakolev, Y. H.; Temelkov, S. G.; Vasilev, R. A.; Gerova, K. D.; Kuseva, C. D.; Todorova, N. D.; Mehmed, A. M.; Rasenberg, M.; Mekenyan, O. G. QSAR Toolbox – Workflow and Major Functionalities. *SAR QSAR Environ Res* **2016**, *27* (3), 203–219. https://doi.org/10.1080/1062936X.2015.1136680.

(91) *Integrated approach to industrial problem substances in surface water*. https://www.tkiwatertechnologie.nl/projecten/van-bron-tot-effect-integrale-aanpak-van-industriele-probleemstoffen-uit-lozingen-op-het-oppervlaktewater (accessed 2024-04-03).

(92) Szöcs, E.; Stirling, T.; Scott, E. R.; Scharmüller, A.; Schäfer, R. B. Webchem: An R Package to Retrieve Chemical Information from the Web. *J Stat Softw* **2020**, *93* (13). https://doi.org/10.18637/jss.v093.i13.

(93) Wickham, H.; François, R.; Henry, L.; Müller, K.; Vaughan, D. Dplyr: A Grammar of Data Manipulation. 2023. https://dplyr.tidyverse.org.

(94) Wickham, H. Stringr: Simple, Consistent Wrappers for Common String Operations. 2023. https://stringr.tidyverse.org.

(95) Palm, E. H.; Chirsir, P.; Krier, J.; Thiessen, P. A.; Zhang, J.; Bolton, E. E.; Schymanski, E. L. ShinyTPs: Curating Transformation Products from Text Mining Results. *Environ Sci Technol Lett* **2023**, *10* (10), 865–871. https://doi.org/10.1021/acs.estlett.3c00537.

(96) Helmus, R.; van de Velde, B.; Brunner, A. M.; ter Laak, T. L.; van Wezel, A. P.; Schymanski, E. L. PatRoon 2.0: Improved Non-Target Analysis Workflowsincluding Automated Transformation Product Screening. *J Open Source Softw* **2022**, *7* (71), 4029. https://doi.org/10.21105/joss.04029.

(97) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J Cheminform* **2011**, *3* (33), 33. https://doi.org/10.1186/1758-2946-3-33.

(98) Benfenati, E.; Manganaro, A.; Gini, G. VEGA-QSAR: AI inside a Platform for Predictive Toxicology. *CEUR Workshop Proc* **2013**, *1107* (January), 21–28.

(99) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA Models for Predicting Physicochemical Properties and Environmental Fate Endpoints. *J Cheminform* **2018**, *10* (1), 1–19. https://doi.org/10.1186/s13321-018-0263-1.

(100) Mansouri, K.; Cariello, N. F.; Korotcov, A.; Tkachenko, V.; Grulke, C. M.; Sprankle, C. S.; Allen, D.; Casey, W. M.; Kleinstreuer, N. C.; Williams, A. J. Open-Source QSAR Models for PKa Prediction Using Multiple Machine Learning Approaches. *J Cheminform* **2019**, *11* (1). https://doi.org/10.1186/s13321-019-0384-1.

(101) *CAS Common Chemistry*. https://commonchemistry.cas.org/.

(102) Wohlgemuth, G.; Haldiya, P. K.; Willighagen, E.; Kind, T.; Fiehn, O. The Chemical Translation Service-a Web-Based Tool to Improve Standardization of Metabolomic Reports. *Bioinformatics* **2010**, *26* (20), 2647–2648. https://doi.org/10.1093/bioinformatics/btq476.

(103) Muench, D.; Ranke, J.; Scharmüller, A.; Scott, E. R.; Stanstrup, J.; Stirling, T. Package ‘ Webchem ’ Documentation. **2022**, No. 1.

(104) Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; Richard, A. M. The CompTox Chemistry Dashboard: A Community Data Resource for Environmental Chemistry. *J Cheminform* **2017**, *9* (1), 1–27. https://doi.org/10.1186/s13321-017-0247-6.

(105) *CompTox Dashboard API (v. 1.0.0)*. https://www.epa.gov/comptox-tools/computational-toxicology-and-exposure-data-apis.

(106) *PubChem*. https://pubchem.ncbi.nlm.nih.gov/.

(107) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res* **2016**, *44* (D1), D1202–D1213. https://doi.org/10.1093/nar/gkv951.

(108) RDKit: Open-Source Cheminformatics. https://www.rdkit.org.

(109) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J Chem Inf Comput Sci* **2003**, *43* (2), 493–500. https://doi.org/10.1021/ci025584y.

(110) RRDkit: A Pragmatic Interface to RDKit in R. https://github.com/pauca/RRDKit.

(111) Charlop-powers, Z.; Schymanski, E.; Charlop-powers, M. Z. Package ‘ Rcdk .’ **2022**.

(112) Mansouri, K.; Moreira-Filho, J. T.; Lowe, C. N.; Charest, N.; Martin, T.; Tkachenko, V.; Judson, R.; Conway, M.; Kleinstreuer, N. C.; Williams, A. J. Free and Open-Source QSAR-Ready Workflow for Automated Standardization of Chemical Structures in Support of QSAR Modeling. *J Cheminform* **2024**, *16* (1), 19. https://doi.org/10.1186/s13321-024-00814-3.

(113) Backman, T. W. H.; Cao, Y.; Girke, T. ChemMine Tools: An Online Service for Analyzing and Clustering Small Molecules. *Nucleic Acids Res* **2011**, *39* (SUPPL. 2), 486–491. https://doi.org/10.1093/nar/gkr320.

(114) Schymanski, E. L.; Kondić, T.; Neumann, S.; Thiessen, P. A.; Zhang, J.; Bolton, E. E. Empowering Large Chemical Knowledge Bases for Exposomics: PubChemLite Meets MetFrag. *J Cheminform* **2021**, *13* (1), 1–15. https://doi.org/10.1186/s13321-021-00489-0.

(115) ShinyTPs (v. 0.1.9).

(116) Gao, J.; Ellis, L. B. M.; Wackett, L. P. The University of Minnesota Biocatalysis/Biodegradation Database: Improving Public Access. *Nucleic Acids Res* **2009**, *38* (SUPPL.1). https://doi.org/10.1093/nar/gkp771.

(117) Wicker, J.; Lorsbach, T.; Gütlein, M.; Schmid, E.; Latino, D.; Kramer, S.; Fenner, K. EnviPath - The Environmental Contaminant Biotransformation Pathway Resource. *Nucleic Acids Res* **2016**, *44* (D1), D502–D508. https://doi.org/10.1093/nar/gkv1229.

(118) Tam, J. Y. C.; Lorsbach, T.; Schmidt, S.; Wicker, J. S. Holistic Evaluation of Biodegradation Pathway Prediction: Assessing Multi-Step Reactions and Intermediate Products. *J Cheminform* **2021**, *13* (1), 1–14. https://doi.org/10.1186/s13321-021-00543-x.

(119) Hafner, J.; Lorsbach, T.; Schmidt Bayer, S.; Brydon, L.; Dost, K.; Zhang, K.; Fenner, K.; Wicker, J. S.; Schmidt, S.; Wicker, J. Advancements in Biotransformation Pathway Prediction: Enhancements, Datasets, and Novel Functionalities in EnviPath. **2023**, 0–20.

(120) Github Repository: EnviPath Python (v. 0.2.2). https://github.com/enviPath/enviPath-python.

(121) United States Enviromental Protection Agency. User's Guide for the Chemical Transformation Simulator Chemical Transformation Simulator: A Cheminformatics Tool for Predicting Transformation Pathways and Physicochemical Properties. **2019**.

(122) Wolfe, K.; Pope, N.; Parmar, R.; Galvin, M.; Stevens, C.; Weber, E.; Flaishans, J.; Purucker, T. Chemical Transformation System: Cloud Based Cheminformatic Services to Support Integrated Environmental Modeling. *Environmental Modelling and Software for Supporting a Sustainable Future, Proceedings - 8th International Congress on Environmental Modelling and Software, iEMSs 2016* **2016**, *1*, 186–193.

(123) *Metabolizer Library - Scheme List*. https://qed.epa.gov/static_qed/cts_app/docs/MetabolizerHTML/MetabolizerDefaultLibrary_SchemeList.htm.

(124) *Metabolizer - ChemAxon*. https://docs.chemaxon.com/display/docs/metabolizer.md.

(125) CTS API (v. 1.0.0). https://qed.epa.gov/cts/rest/.

(126) Acrylate, I. 5 Estimating Physical / Chemical and Environmental Fate Properties with EPI Suite ™. **2012**, 1–22.

(127) Card, M. L.; Gomez-Alvarez, V.; Lee, W. H.; Lynch, D. G.; Orentas, N. S.; Lee, M. T.; Wong, E. M.; Boethling, R. S. History of EPI Suite™ and Future Perspectives on Chemical Property Estimation in US Toxic Substances Control Act New Chemical Risk Assessments. *Environ Sci Process Impacts* **2017**, *19* (3), 203–212. https://doi.org/10.1039/c7em00064b.

(128) Meylan, W. M.; Howard, P. H. Estimating Log P with Atom/Fragments and Water Solubility with Log P. *Perspectives in Drug Discovery and Design* **2000**, *19*, 67–84. https://doi.org/10.1023/A:1008715521862.

(129) William, M.; Howard, P. H.; Boethling, R. S. Molecular Topology/Fragment Contribution Method for Predicting Soil Sorption Coefficients. *Environ Sci Technol* **1992**, *26* (8), 1560–1567. https://doi.org/10.1021/es00032a011.

(130) Estrada, E. Physicochemical Interpretation of Molecular Connectivity Indices. *Journal of Physical Chemistry A* **2002**, *106* (39), 9085–9091. https://doi.org/10.1021/jp026238m.

(131) EpiSuite: KOWWIN User's Guide. https://episuite.dev/EpiWebSuite/#/help/kowwin.

(132) EpiSuite: KOCWIN User's Guide. https://episuite.dev/EpiWebSuite/#/help/kocwin.

(133) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA Models for Predicting Physicochemical Properties and Environmental Fate Endpoints. *J Cheminform* **2018**, *10* (1), 1–19. https://doi.org/10.1186/s13321-018-0263-1.

(134) OPERA (v. 2.9). https://github.com/kmansouri/OPERA.

(135) Roncaglioni, A.; Lombardo, A.; Benfenati, E. The VEGAHUB Platform: The Philosophy and the Tools. *ATLA Alternatives to Laboratory Animals* **2022**, *50* (2), 121–135. https://doi.org/10.1177/02611929221090530.

(136) Golbamaki, A.; Farmacologiche, R.; Negri, M.; Negri, V. M. ALog P Model v. 1.0.0 in VEGA v. 1.1.4. **2020**, 2–8.

(137) Golbamaki, A.; Farmacologiche, R.; Negri, M.; Negri, V. M. MLog P Model v. 1.0.0 in VEGA v. 1.1.4. **2020**, 2–8.

(138) *MoNA - MassBank of North America*. https://mona.fiehnlab.ucdavis.edu/.

(139) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. MassBank: A Public Repository for Sharing Mass Spectral Data for Life Sciences. *Journal of Mass Spectrometry* **2010**, *45* (7), 703–714. https://doi.org/10.1002/jms.1777.

(140) Palm, E. H.; Krier, J.; Schymanski, E. L. ShinyTPs : Curate Transformation Products from Text Mining Results. **2023**, 1–9.

(141) McEachran, A. D.; Mansouri, K.; Grulke, C.; Schymanski, E. L.; Ruttkies, C.; Williams, A. J. “MS-Ready” Structures for Non-Targeted High-Resolution Mass Spectrometry Screening Studies. *J Cheminform* **2018**, *10* (1). https://doi.org/10.1186/s13321-018-0299-2.

(142) Bento, A. P.; Hersey, A.; Félix, E.; Landrum, G.; Gaulton, A.; Atkinson, F.; Bellis, L. J.; De Veij, M.; Leach, A. R. An Open Source Chemical Structure Curation Pipeline Using RDKit. *J Cheminform* **2020**, *12* (1), 1–16. https://doi.org/10.1186/s13321-020-00456-1.

(143) Glüge, J.; McNeill, K.; Scheringer, M. Getting the SMILES Right: Identifying Inconsistent Chemical Identities in the ECHA Database, PubChem and the CompTox Chemicals Dashboard. *Environmental Science: Advances* **2023**, *2* (4), 612–621. https://doi.org/10.1039/d2va00225f.

(144) Grulke, C. M.; Williams, A. J.; Thillanadarajah, I.; Richard, A. M. EPA’s DSSTox Database: History of Development of a Curated Chemistry Resource Supporting Computational Toxicology Research. *Computational Toxicology* **2019**, *12* (June), 100096. https://doi.org/10.1016/j.comtox.2019.100096.

(145) McMaster, M. C. *LC/MS*; Wiley, 2005. https://doi.org/10.1002/0471736589.

(146) Gadaleta, D.; Lombardo, A.; Toma, C.; Benfenati, E. A New Semi-Automated Workflow for Chemical Data Retrieval and Quality Checking for Modeling Applications. *J Cheminform* **2018**, *10* (1). https://doi.org/10.1186/s13321-018-0315-6.

(147) Dolciami, D.; Villasclaras-Fernandez, E.; Kannas, C.; Meniconi, M.; Al-Lazikani, B.; Antolin, A. A. CanSAR Chemistry Registration and Standardization Pipeline. *J Cheminform* **2022**, *14* (1). https://doi.org/10.1186/s13321-022-00606-7.

(148) Hollender, J.; Schymanski, E. L.; Ahrens, L.; Alygizakis, N.; Béen, F.; Bijlsma, L.; Brunner, A. M.; Celma, A.; Fildier, A.; Fu, Q.; Gago-Ferrero, P.; Gil-Solsona, R.; Haglund, P.; Hansen, M.; Kaserzon, S.; Kruve, A.; Lamoree, M.; Margoum, C.; Meijer, J.; Merel, S.; Rauert, C.; Rostkowski, P.; Samanipour, S.; Schulze, B.; Schulze, T.; Singh, R. R.; Slobodnik, J.; Steininger-Mairinger, T.; Thomaidis, N. S.; Togola, A.; Vorkamp, K.; Vulliet, E.; Zhu, L.;

Krauss, M. *NORMAN Guidance on Suspect and Non-Target Screening in Environmental Monitoring*; Springer Berlin Heidelberg, 2023; Vol. 35. https://doi.org/10.1186/s12302-023-00779-4.

(149) Foster, J. E. Plasma-Based Water Purification: Challenges and Prospects for the Future. *Phys Plasmas* **2017**, *24* (5). https://doi.org/10.1063/1.4977921.

(150) Doucette, W. J. Quantitative Structure-Activity Relationships for Predicting Soil-Sediment Sorption Coefficients for Organic Chemicals. *Environ Toxicol Chem* **2003**, *22* (8), 1771–1788. https://doi.org/10.1897/01-362.

(151) Carballa, M.; Fink, G.; Omil, F.; Lema, J. M.; Ternes, T. Determination of the Solid–Water Distribution Coefficient (Kd) for Pharmaceuticals, Estrogens and Musk Fragrances in Digested Sludge. *Water Res* **2008**, *42* (1–2), 287–295. https://doi.org/10.1016/j.watres.2007.07.012.

(152) Hodges, G.; Eadsforth, C.; Bossuyt, B.; Bouvy, A.; Enrici, M. H.; Geurts, M.; Kotthoff, M.; Michie, E.; Miller, D.; Müller, J.; Oetter, G.; Roberts, J.; Schowanek, D.; Sun, P.; Venzmer, J. A Comparison of Log K Ow (n-Octanol–Water Partition Coefficient) Values for Non-Ionic, Anionic, Cationic and Amphoteric Surfactants Determined Using Predictions and Experimental Methods. *Environ Sci Eur* **2019**, *31* (1), 1–18. https://doi.org/10.1186/s12302-018-0176-7.

(153) Hyland, K. C.; Dickenson, E. R. V.; Drewes, J. E.; Higgins, C. P. Sorption of Ionized and Neutral Emerging Trace Organic Compounds onto Activated Sludge from Different Wastewater Treatment Configurations. *Water Res* **2012**, *46* (6), 1958–1968. https://doi.org/10.1016/j.watres.2012.01.012.

(154) Hartmann, J.; Rorije, E.; Wassenaar, P. N. H.; Verbruggen, E. Screening and Prioritising Persistent, Mobile and Toxic Chemicals: Development and Application of a Robust Scoring System. *Environ Sci Eur* **2023**, *35* (1). https://doi.org/10.1186/s12302-023-00749-w.

(155) Sigmund, G.; Arp, H. P. H.; Aumeier, B. M.; Bucheli, T. D.; Chefetz, B.; Chen, W.; Droge, S. T. J.; Endo, S.; Escher, B. I.; Hale, S. E.; Hofmann, T.; Pignatello, J.; Reemtsma, T.; Schmidt, T. C.; Schönsee, C. D.; Scheringer, M. Sorption and Mobility of Charged Organic Compounds: How to Confront and Overcome Limitations in Their Assessment. *Environ Sci Technol* **2022**, *56* (8), 4702–4710. https://doi.org/10.1021/acs.est.2c00570.

(156) Ward, S. E.; Davis, A. M. *The Handbook of Medicinal Chemistry: Principles and Practice, 2nd Edition Edited*; 2023. http://books.rsc.org/books/edited-volume/chapter-pdf/1494624/bk9781788018982-00001.pdf.

(157) Liigand, P.; Liigand, J.; Kaupmees, K.; Kruve, A. 30 Years of Research on ESI/MS Response: Trends, Contradictions and Applications. *Anal Chim Acta* **2021**, *1152*, 238117. https://doi.org/10.1016/j.aca.2020.11.049.

(158) Malm, L.; Palm, E.; Souihi, A.; Plassmann, M.; Liigand, J.; Kruve, A. Guide to Semi-Quantitative Non-Targeted Screening Using Lc/Esi/Hrms. *Molecules* **2021**, *26* (12), 1–21. https://doi.org/10.3390/molecules26123524.

Appendix I

The parent compounds were collected from a list of chemicals (5,442 entries) which mentioned their CAS numbers and, in most cases, a trivial name, or general description if it was a mixture. In Figure 31 it is shown that most of the entries were repeated, while some entries did not contain a CAS number at all.

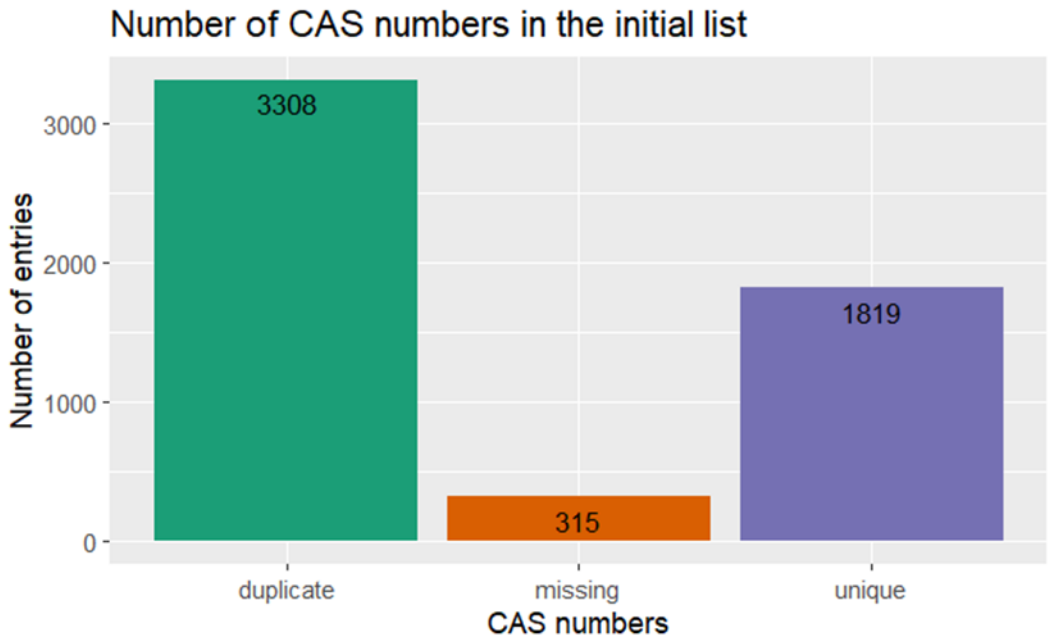


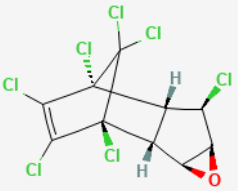
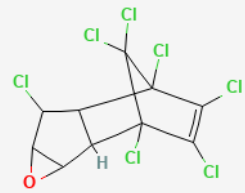
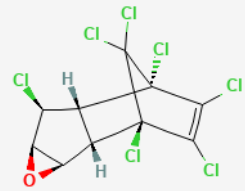
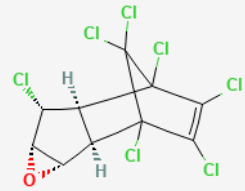
Figure 2: Number of duplicate, missing and unique values from the initial dataset.

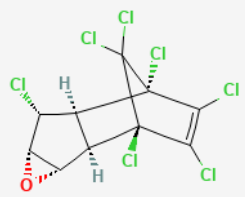
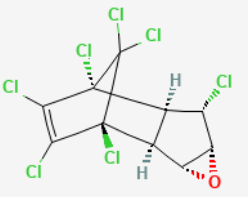
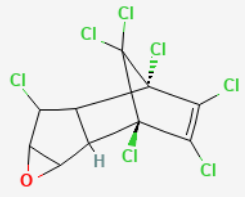
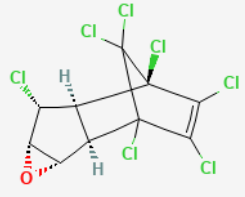
Table 14: Multiple CAS numbers for nickel sulfide (CID 28094)

CAS number	Description	Comment
1314-04-1	Millerite (NiS)	
11113-75-0	Nickel sulfide	
12137-12-1	Nickel sulfide (Ni3S4)	
16812-54-7	Nickel sulfide (NiS)	

Table 15: Multiple CIDs for the CAS number 1024-57-3 Heptachlor epoxide. (figures source: PubChem)

CID	Description	Comment

15559699	Epoxyheptachlor	
13930	Heptachlor epoxide	
12054527	cis-Heptachlor Epoxide	
16212247	Heptachlor-exo-epoxide	

15559706	Heptachloroepoxide	
15559705	Heptachlor-endo-epoxide	
122390596	(+/-)-cis-Heptachlor epoxide	
138394051	Heptachlor Epoxide B	

Appendix II

The compounds were categorized based on their molecular formula and isomeric SMILES (PubChem SMILES with stereochemistry markings), in inorganic, organometallic, organic salts, and organic compounds. The organic compounds were distinct entities (all atoms were connected with valence or coordination bonds), and contained only the atoms C, H, O, N, S, P, F, Cl, Br, and I. For detailed information on the selection criteria for each category see Table 14.

Table 16: Selection criteria in each compound category.

Category	Compound criteria
Inorganic	Compounds with less than two carbon atoms present in the structure. They can be salts or distinct structures.
Organometallic	Compounds containing additional atoms except C, H, O, N, S, P, F, Cl, Br, and I. Only distinct structures.
Organic salts	Compounds with more than two carbon atoms present. No distinct structures.
Organic	Compounds containing only C, H, O, N, S, P, F, Cl, Br, and I atoms. Only distinct structures.

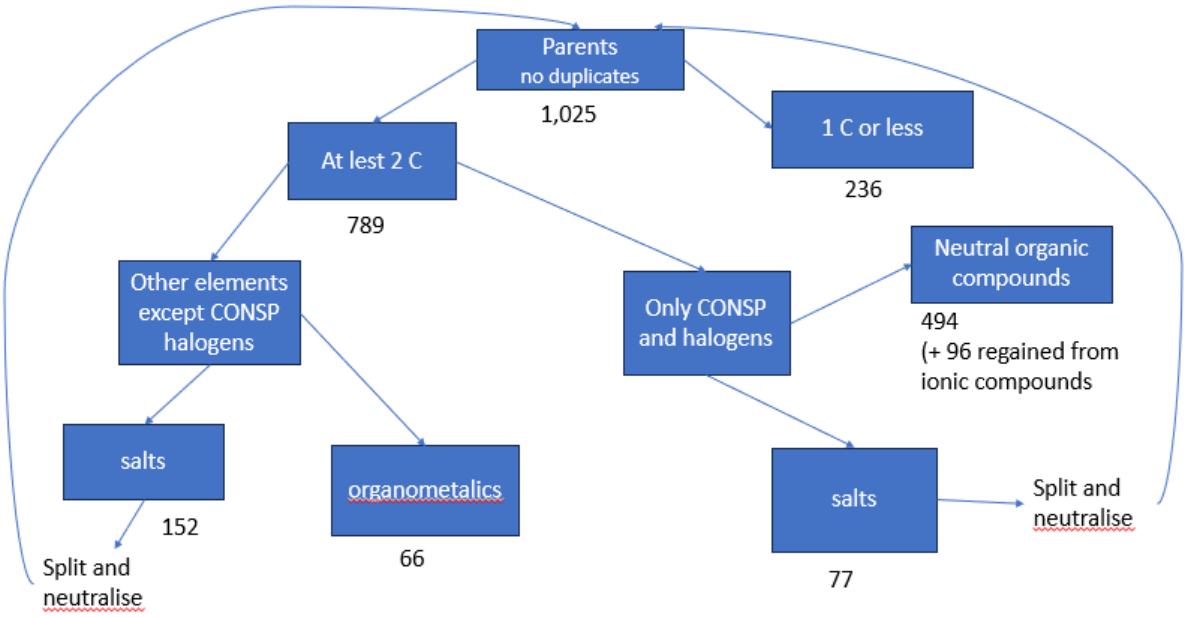


Figure 3: Sketch of the strategy for the categorization of the compounds. The number of compounds in each final category is written below.

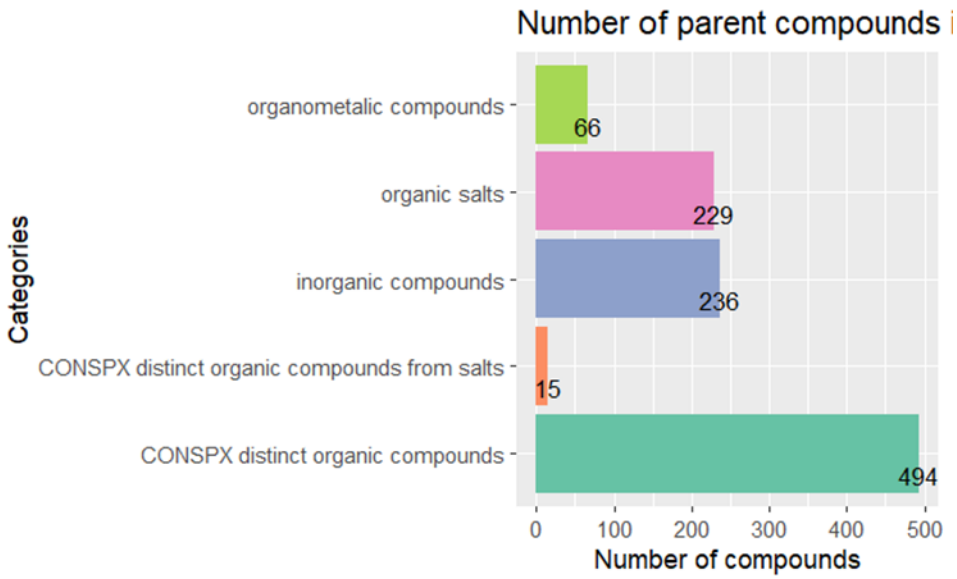


Figure 4: Bar plot with the number of compounds in each category.