Supplementary information to: Ecotoxicological HC20-values and their statistical distribution: A nonlinear weighted regression applied to thousands of chemicals

Oskar Nyberg, Reinout Heijungs, Patrik J. G. Henriksson

2023-09-19

Table of Contents

[Workflow overview 3](#_Toc146036624)

[1 Pre-processing of the chemical inventory 4](#_Toc146036625)

[1.1 The HESTIA chemical inventory 4](#_Toc146036626)

[1.2 CASRN to SMILES identification 4](#_Toc146036627)

[2 Data queries 6](#_Toc146036628)

[2.1 Physicochemical data gathering 6](#_Toc146036629)

[2.1.1 Molecular weight (MolW) 6](#_Toc146036630)

[2.1.2 Acid and base dissociation values (pKa) 6](#_Toc146036631)

[2.1.3 Octanol-water partitioning coefficient (Kow) 6](#_Toc146036632)

[2.1.4 Organic carbon-water partitioning coefficient (Koc) 7](#_Toc146036633)

[2.1.5 Vapor pressure () 7](#_Toc146036634)

[2.1.6 Water solubility (Sol25) 7](#_Toc146036635)

[2.1.7 Ultimate biodegradation models for water, soil and sediment (Kdeg) 7](#_Toc146036636)

[2.1.8 Ultimate biodegradation models for air (Kdeg) 8](#_Toc146036637)

[2.1.9 Bio accumulation factor, fish () 8](#_Toc146036638)

[2.1.10 Henry’s Law coefficient (KH25C) 8](#_Toc146036639)

[2.2 Ecotoxicological effect data 8](#_Toc146036640)

[3 Data processing 11](#_Toc146036641)

[3.1 Chemical use categorization 11](#_Toc146036642)

[3.1.1 Data gathering and processing 11](#_Toc146036643)

[3.1.2 Substance type annotations 12](#_Toc146036644)

[3.1.3 Categorizing use classification and chemical types 12](#_Toc146036645)

[3.2 Taxonomic classification 13](#_Toc146036646)

[3.3 Curation of ecotoxicological data 16](#_Toc146036647)

[3.3.1 Adjusting test concentrations reported as a range, 16](#_Toc146036648)

[3.3.2 Manual removal of data 16](#_Toc146036649)

[3.3.3 Test medium selection 18](#_Toc146036650)

[3.3.4 Selecting ecologically relevant effect criterions 18](#_Toc146036651)

[3.3.5 Selecting relevant taxonomic annotations 20](#_Toc146036652)

[3.3.6 Effect concentration and unit harmonization 22](#_Toc146036653)

[3.3.7 Test durations 23](#_Toc146036654)

[3.3.8 Assign acute or chronic test definitions. 24](#_Toc146036655)

[3.3.9 EC10eq conversions 25](#_Toc146036656)

[4 Results 27](#_Toc146036657)

[4.1 Ecotoxicological database curation summary 27](#_Toc146036658)

[File connectivity schema 29](#_Toc146036659)

[References 30](#_Toc146036660)

The following document provides detailed descriptions to the construction of the HESTIA Ecotoxicological database. Data is attached as a Microsoft Excel file (Supplementary Data) and R code is available at <https://github.com/osny1923/HESTIA_tox_platform>.

Corresponding author: [oskar.nyberg@su.se](mailto:oskar.nyberg@su.se); [nyberg.oskar@gmail.com](mailto:nyberg.oskar@gmail.com)

# Workflow overview

Below is a general schematic diagram of the database construction methodology (Figure A1). A detailed schematic for the file structure-relationships of the GitHub repository is given at the end of this document (Figure A3).

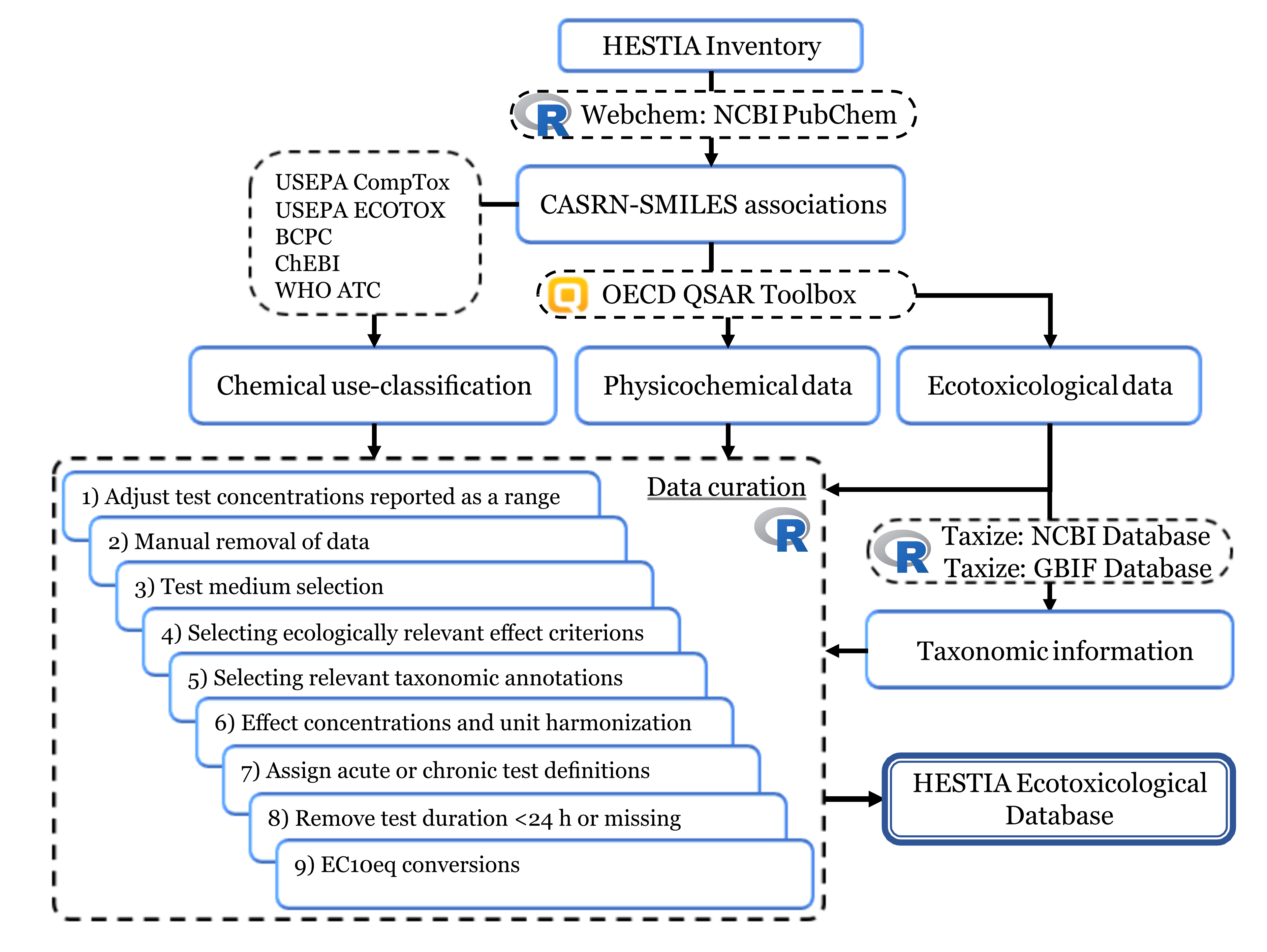


Figure A1. A schematic diagram of the database construction methodology. Dashed lines implies accessing software or other databases to process data. OECD QSAR Toolbox logotype from: <http://oasis-lmc.org/products/software/toolbox/toolbox-support.aspx>, R logotype: Hadley Wickham and others at Rstudio: <https://www.r-project.org/logo/>

# 1 Pre-processing of the chemical inventory

## 1.1 The HESTIA chemical inventory

HESTIA is a free open-access platform that provides a data repository for life cycle inventory data using a harmonized schema and glossary of terms, and calculations tools for various emissions and impact assessments. We use the full inventory of agrochemicals from the HESTIA platform as the starting point for the construction of an ecotoxicological database. This inventory is contained in the file data/excel\_references/pesticideAI.csv, with the following information for 16797 substances: term.id, term.name, term.units, term.synonyms, term.subClassOf.0.id, term.casNumber, term.pubchem, and term.chemidplus.

* “term.id” contains a string “CAS-xxxx-xx-x”, where “x” corresponds to numerials in a Chemical abstract services registry number (CASRN).
* “term.name” contains a string with a name identifier of the chemical.
* “term.units” contains the inventory unit string “kg active ingredient”.
* “term.synonyms” contains 0 to 934 other possible name identifiers for each chemical respectively.
* “term.subClassOf.0.id” contains pesticide-type specific information for 123 chemicals with 11 levels (triazine, dinitroaniline, cyclicN, coumarin, neonicotinoid, carbamate, organophosphorus, insecticideUnspecifiedAi, organochlorine, organotin, and herbicideUnspecifiedAi).
* “term.casNumber” contains the CASRN identifier.
* “term.pubchem” contains a webpage reference to the chemical-specific PubChem database page.
* “term.chemidplus” contains a webpage reference to the chemical-specific ChemIdPlus database page.

The full inventory is available in supplementary data “HESTIA\_INVENTORY”-tab.

## 1.2 CASRN to SMILES identification

The file pesticideAI.csv is read into R and because of cases where Excel tends to reformat certain CASRNs into dates, the column “term.id” is used to derive CASRN. This is performed by separating the column from “CAS-” and only keep the corresponding CASRN numerical identification, calling this column “CAS.number”. Next, we use the NCBI PubChem project to match CASRN and substance names to SMILES configuration with the intermediate step of acquiring PubChem substance IDs using the R package [Webchem](https://docs.ropensci.org/webchem/) (Szöcs et al. 2020). The script is documented in the file “*code/Translating\_CAS\_to\_SMILES\_via\_PubChem.R*”.  
First query to the PubChem (PC) project is performed to match chemical name (“term.name”-column) to PC substance IDs using the get\_cid()-function. This matched 1.2886^{4} names to PC-IDs which are used to collect information on SMILES-configuration. For the remaining 3911 chemicals, a second query is performed, here using the “CAS.number” as “from = ‘xref/RN’ inside the get\_cid()-function, making sure to enable all possible matches. We notice that chemical names are spelled different between the pesticideAI.CSV and the PC repository, i.e., spelling of”sulfate” in pesticideAI.csv and “sulphate” in the PC repository. After substituting “sulf ->”sulph” across the name-vector and subsequently name-matching between the returned query, remaining unidentified chemicals are queried using the CASRN for PC Ids using the pc\_sect-function. 1652 chemicals obtained a matching SMILES-configuration.

Lastly, 2259 chemicals remain, with multiple PC Id matches from the CASRN-query are returned from the pc\_prop()-function, and an iterative process of identifying the correct chemical per multi-match is undertaken, querying the SMILES-configuration for all possible PC Ids using pc\_sect(), then selecting correct CASRN-matches. The final result of CASRN-SMILES configuration matching is compiled to as a complete list of 16797 chemicals, with 15030 identified SMILES-configurations, available in supplementary data “CAS\_SMILES\_MATCHING”-tab. A .txt-file with 16797 rows containing only CASRN and SMILES-configuration is created to be used for importing data into the OECD QSAR Toolbox  
Lastly, the output is also exported as five .txt-files with maximum 4000 rows of CAS-SMILES matches intended for use in downstream data queries from OECD QSAR Toolbox. Due to memory issues with the VEGA QSAR-software and when compiling ecotoxicological data in OECD QSAR Toolbox, a query of maximum 4000 chemicals per turn will be used. These files are all exported as “data/excel\_references/cas\_smiles\_list[’x’k-’x’k].txt” - where x represents the row number in “data/excel\_references/CAS\_CID\_list\_final.txt”.

# 2 Data queries

The OECD QSAR Toolbox v4.5 SP1 2022 software (Dimitrov et al. 2016) is used to gather physicochemical data and ecotoxicological records for the chemicals listed in the “data/excel\_references/CAS\_CID\_list\_final.csv”-file. As input, a .txt-file with CAS-SMILES matches is read and 15030 with matching CASRN-SMILES-configurations are imported by reading the .txt-file as list object, selecting “NO” when asked to “search the database for empty SMILES with defined CAS numbers”. 9 “problematic chemicals” were identified at rows 1696, 1724, 1937, 2334, 2406, 2877, 4252, 6961, and 6502, leading to 16788 chemical structures imported to the software.

## 2.1 Physicochemical data gathering

The following 2D parameters are selected for estimations using OECD QSAR Toolbox:

Acidic pKa OASIS Consensus , Acidic pKa OASIS Regression , BAF upper trophic , Basic pKa OASIS Regression , Exp Henrys Law Constant, Exp Log P, Exp Vapor Pressure, Exp Water Solubility, Henrys Law Constant Bond Method , Koc Log Kow , Koc MCI , log Kow, Molecular Weight, OVERALL OH Half life, OVERALL OH rate constant, Selected Vapor Pressure, Ultimate biodeg Biowin 3 , Vapor Pressure Antoine method , Vapor Pressure Modified Grain Method , and Water Solubility. All of these parameters are required by USEtox v2.1 for characterizations of chemicals. The full dataset with obtained physicochemical data is available in supplementary data “PHYSCHEM”-tab. Annotations tagged with “Exp”-implies that data comes from experimental sources, others are tagged with “Est”, meaning they are estimated using *in silico* methods.

### 2.1.1 Molecular weight (MolW)

Molecular weight was calculated using OASIS software (Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria, <https://oasis-lmc.org/>). MolW is reported as Dalton (Da) but converted to g as a 1:1 ratio.

### 2.1.2 Acid and base dissociation values (pKa)

Acid and base dissociation values (pKa) are calculated using the OASIS regression model (Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria, <https://oasis-lmc.org/>), where pKa gain is defined from the “Basic pKa”, pKa loss is defined from the Acidic pKa, and the pKa chemical class parameter can be defined as “Acidic” if cells are populated for pKa loss, “Basic” if cells are populated for pKa gain, “Amphoteric” if both cells are populated and “Neutral” if both cells are blank. For multi constituent substances, the pKa chemical class was left as “undefined”.

### 2.1.3 Octanol-water partitioning coefficient (Kow)

Octanol-water partitioning coefficient KOW was estimated using the EPIsuite LOGKOW fragment constant methodology.

### 2.1.4 Organic carbon-water partitioning coefficient (Koc)

Organic carbon-water partitioning coefficient is estimated in EPIsuite using the MCI regression models, similar to in USEtox’s methodology, applying the same cut-off values for suitability. Koc is then calculated for substances classed as neutral, acids and amphoteric substances when pKa.loss ranges 0-12 and logKOW ranges -2.18 - 8.5, and bases and amphoteric substances when pKa.gain is above 2 and logKOW ranges -1.66 - 7.03.

### 2.1.5 Vapor pressure ()

Vapor pressure at 25 degrees Celcius () is estimated using available EPIsuite models and is automatically selected as by the software as the most fitting model from either the Antoine Method, the Modified Grain Method, or the Mackay Method. is reported as mm Hg, but converted into Pascal (Pa) using a multiplier of 133,322.

### 2.1.6 Water solubility (Sol25)

Where experimental water solubility (Sol25, mg L-1) data is available, these data are prioritized over solubility data estimated using EPIsuite WSKOW models which use KOW as base for regression models to generate estimates.

### 2.1.7 Ultimate biodegradation models for water, soil and sediment (Kdeg)

Ultimate biodegradation models (BIOWIN 3) are used to estimate biodegradation rates in water, soil and sediment. Estimated BIOWIN 3 output values are transformed into degradation rates (1/s) by first using the conversion specified in Table A1 available in the EPIsuite manual (P2 Framework Manual 2012 EPA-748-B12-001), transforming values into delimited time formats and thereafter converting these into biodegradation rates as 1/s according to the USEtox manual (Fantke et al. 2017).

| BIOWIN 3 Result | Time Required for Biodegradation | Assigned Half-Life (days) | Biodegradation rate (1/s) |
| --- | --- | --- | --- |
| >4.75 - 5 | Hours | 0.17 | 4.7e-05 |
| >4.25 - 4.75 | Hours - days | 1.25 | 6.4e-06 |
| >3.75 - 4.25 | Days | 2.33 | 3.4e-06 |
| >3.25 - 3.75 | Days - weeks | 8.67 | 9.3e-07 |
| >2.75 - 3.25 | Weeks | 15.00 | 5.3e-07 |
| >2.25 - 2.75 | Weeks - months | 37.50 | 2.1e-07 |
| >1.75 - 2.25 | Months | 60.00 | 1.3e-07 |
| <1.75 | Longer “recalcitrant” | 180.00 | 4.5e-08 |

Table A1. Conversions of BIOWIN 3 output into biodegradation rates according to methodology in Fantke et al. (2017)

### 2.1.8 Ultimate biodegradation models for air (Kdeg)

Degradation rates for air was calculated according to the OH rate constant method specified in the USEtox manual with the formula

Where [OH] represents the default OH concentration in air: 1.5 × 106 molecules (radicals)/cm3 per 12h of daylight (Fantke et al. 2017). The overall OH rate constant was calculated from EPIsuites AOP v1.92: HYDROXYL RADICALS (25 deg C) model.

### 2.1.9 Bio accumulation factor, fish ()

Data on bio accumulation factors in fish, Log BAF (upper trophic) values estimated using Arnot-Gobas BCF & BAF Methods in EPIsuite were used. Log-transformed log10 gives the correct unit for USEtox calculations, L .

### 2.1.10 Henry’s Law coefficient (KH25C)

Henry’s Law coefficient (KH25C) could be sourced from experimental data (LogP) or calculated by

according to Fantke et al., (2017), where Pvap25 (Pa) is the vapor pressure at 25℃, MolW is the molecular weight (g mol-1), and Sol25 is the water solubility at 25℃ (mg L-1).

## 2.2 Ecotoxicological effect data

The five files containing CASRN-SMILES matches (“data/excel\_references/cas\_smiles\_list[’x’k-’x’k].txt”) are each imported separately into OECD QSAR Toolbox. Then all available aquatic ecotoxicity data is gathered from available sources using the “data”->“gather” function, selecting “all endpoints”. The data are subsequently exported as five separate documents (to avoid system freezing due to insufficient memory).

The data output from OECD QSAR Toolbox contains 441206 records from data sources reported in Table A2.

| Database | Number of records |
| --- | --- |
| Aquatic ECETOC | 6,837 |
| Aquatic Japan MoE | 3,349 |
| Aquatic OASIS | 4,028 |
| ECOTOX | 414,724 |
| Food TOX Hazard EFSA | 2,181 |
| NA | 10,087 |
| Total number of records | 441,206 |

Table A2. Numer of records returned from OECD QSAR Toolbox. Database = NA imply that no matches were found for a chemical.

When imported into R, the script “*code/raw\_data\_read\_and\_wrangle.R*” merge all files, selects relevant columns, relevant endpoints, removes empty rows and duplicate records across all columns. The data are subsequently exported to a “.csv” file: “data/RAW\_DB/HESTIA\_HC20\_DB\_raw\_toxdata.csv”, which is used as the starting point for the ecotoxicological database. The number of records in this dataset is reported in Table A3

| Database | Endpoint | Total number of records |
| --- | --- | --- |
| Aquatic ECETOC | EC10 | 94 |
| Aquatic ECETOC | EC50 | 4,739 |
| Aquatic ECETOC | LOEC | 829 |
| Aquatic ECETOC | NOEC | 813 |
| Aquatic Japan MoE | EC50 | 1,537 |
| Aquatic Japan MoE | LC50 | 559 |
| Aquatic Japan MoE | LOEC | 33 |
| Aquatic Japan MoE | NOEC | 1,220 |
| Aquatic OASIS | EC50 | 259 |
| Aquatic OASIS | IC50 | 56 |
| Aquatic OASIS | LC50 | 2,011 |
| Aquatic OASIS | LD50 | 16 |
| ECOTOX | EC0 | 526 |
| ECOTOX | EC10 | 3,387 |
| ECOTOX | EC50 | 32,696 |
| ECOTOX | EL10 | 2 |
| ECOTOX | EL50 | 4 |
| ECOTOX | ER10 | 25 |
| ECOTOX | ER50 | 122 |
| ECOTOX | IC10 | 419 |
| ECOTOX | IC50 | 3,500 |
| ECOTOX | LC0 | 1,024 |
| ECOTOX | LC10 | 1,506 |
| ECOTOX | LC50 | 86,720 |
| ECOTOX | LD10 | 34 |
| ECOTOX | LD50 | 1,887 |
| ECOTOX | LOEC | 54,039 |
| ECOTOX | NOEC | 71,109 |
| ECOTOX | NOEL | 10,568 |
| ECOTOX | NOER | 168 |
| Food TOX Hazard EFSA | EC10 | 23 |
| Food TOX Hazard EFSA | EC50 | 930 |
| Food TOX Hazard EFSA | ER50 | 1 |
| Food TOX Hazard EFSA | IC50 | 2 |
| Food TOX Hazard EFSA | LC10 | 1 |
| Food TOX Hazard EFSA | LC50 | 394 |
| Food TOX Hazard EFSA | LD50 | 1 |
| Food TOX Hazard EFSA | NOAEC | 10 |
| Food TOX Hazard EFSA | NOEC | 791 |
| Food TOX Hazard EFSA | NOEL | 1 |
| Number of records |  | 282,056 |

Table A3. Numer of records per database and endpoint used as input to the ecotoxicological database construction after selecting relevant end points, removing chemicals without data and duplicates.

# 3 Data processing

## 3.1 Chemical use categorization

To categorize chemicals according to how they are used or applied, several processing steps are applied to obtain the best categorization per chemical. The operations utilize the R package Webchem (Szöcs et al. 2020). Starting with importing the file “data/excel\_references/CAS\_CID\_list\_final.csv” as an object called “CAS\_CID\_list” for chemical use category annotations. Code is available in “*code/Pesticide\_annotations.R*”

### 3.1.1 Data gathering and processing

Data from multiple resources are gathered using various methods:

**USEPA CompTox v2.2** (query date 2023-01-20) was queried for repositories of classified chemicals by searching “chemical lists” -> “list names” using the queries “pesticides”, “anti” (to select lists matching the terms “antibiotic”, “antifungal”, and “antiviral”), “pharmaceutical”, “herbicide”, “insecticide”, and “rodenticide”. Matches were downloaded as .csv files and imported into R and compiled into one extensive document columns defined depending on database match and contents were marked as the respective use-category:

* DRUGBANK (content marked as “Pharmaceutical”),
* United States Environmental Protection Agency: Pesticide Chemical Search Database (EPAPCS; content marked as “Pesticide”),
* Healthy Building Network (HBN; content marked as “Antibiotic”),
* NORMAN Innovative Training Network (Paulus et al. (2019); content marked as “Antibiotic”),
* Office of Pesticide Programs Information Network (OPPIN; content marked as “Pesticide”),
* Pesticide properties Database (PPDB; content marked as “Pesticide”),
* USEPA (content marked as “Antibiotic”, “Pesticide ingredient”, or “Pesticide”),
* Wikipedia (content marked as “Antibiotic”, “Antifungal”, “Antiviral”, “Insecticide”, “Herbicide”, “Rodenticide”, or “Veterinary drug”).

Each chemicals matched in either repository was categorized into the groups: Antibiotic, Antiviral, Other inorganic chemicals, Other organic chemicals, PPCP, Pesticide, Pharmaceutical, and Unknown. The resulting dataset contains information for 11093 chemicals and is exported as a .csv-file “data/excel\_references/USEPA\_CHEMLIST\_2023-01-20.csv”.

**US EPA ECOTOX DB** was accessed and downloaded on 2022-03-10 from <https://cfpub.epa.gov/ecotox/> as ASCII formatted data contained in a .zip file. The file “/validation/chemicals.txt” contains information on chemical\_name, USEPA\_ecotox\_group, dtxsid, and CAS.Number for 17477 chemicals of which 1653 are annotated by 67 different categories.

**The British Compendium of Pesticide Common Names (BCPC)** was accessed via the webchem package on 2022-10-13 using the function webchem::bcpc\_query() and querying using CASRN from the CAS\_CID\_list-object. The query returned use categorizations for 1634 chemicals and is saved to file “data/excel\_references/BCPC\_df.csv”.

**The Chemical Entities of Biological Interest (ChEBI) database** (<https://www.ebi.ac.uk/chebi/init.do>) contain chemical use classifications in the “parents” subsection following a query via webchem using the functions webchem::get\_chebiid() on the 2023-01-18, to retrieve ChEBI ids for chemicals in the CAS\_CID\_list-object, and webchem::chebi\_comp\_entity which returns a list of lists containing the full ChEBI inventory for each chemical queried. In the parents-list under each chemical is information on a use classification. Through manual validation, 1140 chemicals are matched to the use categorizations according to categories: pesticide, fungicide, fertilizer, antibacterial, antimicrobial, insecticide, agrochemical, herbicide, acaricide, antiviral, and nematocide.

**The Anatomical Therapeutic Chemical (ATC) Classification System** (<https://www.whocc.no/>) is accessible via webchem::chembl\_atc\_classes(). The full ATC class inventory was saved as an object and filtered to select the ATC classes “P” (antiparasitic products, insecticides and repellents) and “J” (antiinfectives for systemic use) and converting chemical names into PC Ids.

These data sets are joined to the CAS\_CID\_list-object and stored as object HESTIA\_Comp\_info.

### 3.1.2 Substance type annotations

Chemicals in the HESTIA\_Comp\_info-object are classified into chemical groups: Other organic chemicals, Pesticide, Unknown, Other inorganic chemicals, PPCP, Antibiotic, and Antiviral. Chemicals were defined as organic or inorganic based on presence or absence of carbon [C] in the SMILES configuration. Halogenated chemicals are defined as chemicals containing the elements [F], [Cl], [Br], [I], and heavy metals are defined as chemicals containing elements with a density 5 g per and an atomic number > 20 (Raychaudhuri et al. 2021). Multi-constituent chemicals are classified as “unknown” due to difficulties in assigning correct SMILES configurations, but if chemical names contain the string “mixt.”, chemicals are classified as “Chemical mixture”.

### 3.1.3 Categorizing use classification and chemical types

With different data sources, a chemical may have >1 use categorization and to supply the best categorization, a function is designed to count the number of various classifications for each chemical and assign the one with highest number of matches (“code/Use\_category\_function.R”). In case of a tie, “Pesticide” categorization is prioritized followed by “Antibiotic” categorization and each annotated with “tie” respectively. If no use classification is available, chemicals are categorized as either “Other organic chemicals” or “Other inorganic chemicals” whether the chemical is defined as organic or inorganic, or “Unknown” if SMILES-configurations are lacking. A summary of the number of records per chemical type and use classification is presented in Table A 4. The final data set is exported as “results/HESTIA\_chem\_prop\_list\_full.csv”. The use-classification is found in supplementary data “CHEMICAL\_PROPERTIES”-tab.

| Group | Substance\_type | Number of records |
| --- | --- | --- |
| Antibiotic | Inorganic | 5 |
| Antibiotic | Organic | 211 |
| Antibiotic | Unknown | 1 |
| Antiviral | Organic | 6 |
| Other inorganic chemicals | Chemical mixture | 170 |
| Other inorganic chemicals | Inorganic | 558 |
| Other organic chemicals | Organic | 10,122 |
| PPCP | Inorganic | 13 |
| PPCP | Organic | 891 |
| PPCP | Unknown | 1 |
| Pesticide | Chemical mixture | 27 |
| Pesticide | Inorganic | 195 |
| Pesticide | Organic | 2,838 |
| Pesticide | Unknown | 112 |
| Unknown | Chemical mixture | 353 |
| Unknown | Unknown | 1,294 |
| Total number of records |  | 16,797 |

Table A4. Numer of records categorized per chemical type and use classification.

## 3.2 Taxonomic classification

The dataset (“data/RAW\_DB/HESTIA\_HC20\_DB\_raw\_toxdata.csv”) contains *in vivo* ecotoxicological effect data based on 4206 different taxa at various taxonomic level with multiple invalid, misspelled or outdated species definitions, defined in the column “Test.organism..species.”. For treating taxonomic information and validation in the R environment, the R package Taxize is used extensively (Chamberlain and Szocs 2013). The following operations are performed to produce a coherent list of taxonomic information, scripted in the R code “*code/Taxonomy\_wrangling.R*”.

* Read the “data/RAW\_DB/HESTIA\_HC20\_DB\_raw\_toxdata.csv” and select unique records from the “Test.organism..species.”-column.
* Define binomial species definitions (“genus” + “species”), apply the tol\_resolve()-function to correct incorrect taxonomic classifications, separate correct records with taxonomic information.
* Query the NCBI database for full taxonomic information.
* Subset all species that could not be queried at NCBI into categories: “not found in the NCBI query”, “Common names”, and “non-binomial definitions”.
* Query the GBIF database for additional taxonomic information on the subset “not found in the NCBI query”.
* Query the NCBI database for common names
* Manually correct taxonomic information, e.g., remove sub-species information within the subset “non-binomial definitions”.
* Query the NCBI database for corrected taxonomic information in the last subset.
* Merge the four subsets and select relevant columns: Test.organisms..species., phylum, class, order, genus, species (updated), source, and query.
* Export results as .csv file “data/Taxonomy/Species\_taxonomy.csv” where the column “Test.organism..species.” contains the original taxonomic information, and the column “Species” contains updated taxonomic information.

The operation resolved correct taxonomic information for 3134, observe that some taxa were counted twice as both outdated and current names within the starting data. The full list of taxonomic information is supplied in supplementary data “TAXONOMIC\_INFORMATION”-tab.

| Taxonomy.Group | Phylum | Total number of records |
| --- | --- | --- |
| Algae | Bacillariophyta | 155 |
| Algae | Charophyta | 7 |
| Algae | Chlorophyta | 215 |
| Algae | Cryptophyta | 17 |
| Algae | Cyanobacteriota | 102 |
| Algae | Dinophyta | 27 |
| Algae | Haptophyta | 14 |
| Algae | Ochrophyta | 31 |
| Algae | Rhodophyta | 22 |
| Algae | Sar | 47 |
| Amphibian | Chordata | 150 |
| Annellidae | Annelida | 94 |
| Crustacean | Arthropoda | 635 |
| Fish | Chordata | 680 |
| Insect | Arthropoda | 565 |
| Mollusca | Mollusca | 408 |
| Others | Acanthocephala | 1 |
| Others | Apicomplexa | 2 |
| Others | Arthropoda | 18 |
| Others | Ascomycota | 54 |
| Others | Basidiomycota | 2 |
| Others | Bryozoa | 15 |
| Others | Chaetognatha | 2 |
| Others | Chordata | 38 |
| Others | Chytridiomycota | 2 |
| Others | Ciliophora | 57 |
| Others | Cnidaria | 47 |
| Others | Ctenophora | 2 |
| Others | Cyanobacteria | 15 |
| Others | Discosea | 5 |
| Others | Echinodermata | 39 |
| Others | Entoprocta | 1 |
| Others | Euglenozoa | 18 |
| Others | Foraminifera | 1 |
| Others | Gastrotricha | 1 |
| Others | Microsporidia | 1 |
| Others | Mucoromycota | 3 |
| Others | Nematoda | 72 |
| Others | Nematomorpha | 2 |
| Others | Oomycota | 8 |
| Others | Perkinsozoa | 2 |
| Others | Platyhelminthes | 53 |
| Others | Porifera | 3 |
| Others | Pseudomonadota | 2 |
| Others | Sarcomastigophora | 1 |
| Others | Tracheophyta | 2 |
| Others |  | 288 |
| Plant | Streptophyta | 210 |
| Rotifera | Rotifera | 70 |

Table A5. Numer of unique species identified per taxonomic group .

## 3.3 Curation of ecotoxicological data

Using the OECD QSAR Toolbox software, the list of matching CASRN to SMILES configurations (15030 chemicals with matches; see Supplementary Data, tab: “CAS\_SMILES\_MATCHING”) were queried for ecotoxicological data at the following endpoints NOEC, NOEL, LC0, EC0, NOER, NOAEC, LC50, EC50, IC50, LD50, ER50, EL50, LOEC, EC10, LC10, IC10, LD10, ER10, and EL10. A total of 282056 records were returned for 5027 chemicals. These data were subsequently curated stepwise as follows:

### 3.3.1 Adjusting test concentrations reported as a range,

Curation of values reported as ranges are done following a simplified methodology from Saouter et al. (2018). For records where a min and max range is available the following rules will be applied: In cases where a lower range value is annotated with “ca.”, “=” or no annotation, **the lowest value is selected**. But if the lower range value is annotated with “>”, the record is removed, except for records with NOEC endpoints. In cases where lower values are annotated with “<”, this value is removed (Table A6).

| Data qualifier type | Total number records reported as range | EC10 | EC50 | NOEC | Number of records removed |
| --- | --- | --- | --- | --- | --- |
| < | 759 | 340 | 16 | 403 | 759 |
| = | 540 | 3 | 2 | 535 |  |
| > | 3,369 | 303 | 2,660 | 406 | 2,963 |
| ca. | 35 | 10 | 1 | 24 |  |
| No qualifier | 69,513 | 6,062 | 56,684 | 6,767 |  |
| (Total) records where low range effect value is used | 74,216 | 6,718 | 59,363 | 8,135 |  |

Table A6. Number of records where effect concentrations are presented as ranges with or without value qualifiers.

### 3.3.2 Manual removal of data

Records with strings “Insufficient” or “Unsatisfactory” in the column Control.type are removed, as well as data containing the strings “QSAR”, “bioassay”, “quantitative” in the Title column, and the strings “QSAR” and “bioassay” in the Experimental.design column. Manual inspection of data reveals records containing inaccurate or improper data and the following records are subsequently removed (Table A7)

| Search.String | Column | Motivation |
| --- | --- | --- |
| Assessment of Aquatic Experimental Versus Predicted and Extrapolated Chronic Toxicity Data of Four Structural Analogues | Title | QSAR study |
| Effects of Chlorpyrifos, Carbendazim, and Linuron on the Ecology of a Small Indoor Aquatic Microcosm | Title | QSAR study |
| Altenburger,R., H. Walter, and M. Grote | Title | QSAR study |
| PREDICTING MODES OF TOXIC ACTION FROM CHEMICAL STRUCTURE: ACUTE TOXICITY IN THE FATHEAD MINNOW | Title | QSAR study |
| Human Cardiotoxic Drugs Delivered by Soaking and Microinjection Induce Cardiovascular Toxicity in Zebrafish | Title | Data entered incorrectly, mixing cardiovascular injections with soaking solutions, also, data reported as minimum non-lethal concentration is entered as EC10. |
| A Multi-Battery Toxicity Investigation on Fungicides | Title | Data on species-specific toxicity entered incorrectly, attributing D magna toxicity to S. capricirnutum. |
| Rainbow Trout Larvae Compared with Daphnia | Title | Bioassay test-results |
| GERISH | Title | No source available for inspection, yet substance 107-21-1 from this study cause an extreme outlier |
| Acute Effects of Binary Mixtures of Imidacloprid and Tebuconazole on 4 Freshwater Invertebrates | Title | 95% CI has one report of extrapolated negative concentrations, removing all data |
| Relative Chronic Sensitivity of Neonicotinoid Insecticides to Ceriodaphnia dubia and Daphnia magna | Title | 95% CI has one report of extrapolated negative concentrations, removing all data |
| Multi-Laboratory Hazard Assessment of Contaminated Microplastic Particles by Means of Enhanced Fish Embryo Test with the Zebrafish | Title | Presents very low effect concentrations from micrplastic study on D. rerio embryos, concentrations reported are used as an effect test result, incorrect entry into database |
| Water Toxicology and Radioecology. Acute Toxicity of Heavy Metals to Aquatic Invertebrates at Different Temperatures | Title | Study runs toxicological effect testing across a range of temperatures, leading to skewed data output. |
| Acute Toxicity of 46 Selected Dyes to the Fathead Minnow, Pimephales promelas | Title | Duplicate data exists from the same author as a conference publication |
| 'Aquatic Japan MoE' | '87-82-1' | Database | CAS.Number | The Japan MoE Ecotoxicological tests on chemical Hexabromobenzene, CASRN=87-82-1, is reported with data 6 orders of magnitude from all other data, and the specific test report is not traceable. |
| A Freshwater Mesocosm Study into the Effects of the Neonicotinoid Insecticide Thiamethoxam at Multiple Trophic Levels | Title | Species | Removal of taxonomic groups Algae and 'Others' in paper, since NOEC values are posterior attribute to non-target communities. |

Table A7. Manually removed records from the dataset

### 3.3.3 Test medium selection

The present database requires ecotoxicological effect data from freshwater organisms. But 16065 records are missing test medium annotations. The majority of records lacking test medium annotations belong to freshwater species: *Daphnia magna*, *Raphidocelis subcapitata*, *Pimephales promelas*, *Oncorhynchus mykiss* and *Oryzias latipes* account for 57.3 % of all 16065 records missing a defined test medium. By using data from identical species we can gap-fill missing test medium information, and define test media for records where no test medium is reported. The updated media type annotations are shown in Table 8 and records with test medium annotated freshwater or culture media are kept, while close to 40,000 records are removed.

| Media.type | Number of records |
| --- | --- |
| Culture | 6,422 |
| Freshwater | 202,995 |
| Nosubstrate | 14 |
| Saltwater | 39,022 |
| NA | 818 |

Table A8. Number of records per defined test medium in the input data.

### 3.3.4 Selecting ecologically relevant effect criterions

The selection of relevant effect criterions was decided on following discussions with Dr. Andreu Rico (IMDA Institute, Madrid, Spain) to maximize data inclusions, while omitting effects that display low ecological relevance (Table A9).

| Effect | Number of records | Ecological relevance |
| --- | --- | --- |
| Mortality | 114,626 | YES |
| Population | 32,205 | YES |
| Biochemistry | 21,859 | NO |
| Genetics | 17,714 | NO |
| Growth | 15,949 | YES |
| Enzyme(s) | 14,476 | NO |
| Intoxication | 11,224 | YES |
| Reproduction | 9,146 | YES |
| Physiology | 8,230 | NO |
| Development | 6,942 | YES |
| Behavior | 6,657 | YES |
| Cell(s) | 3,698 | YES |
| Morphology | 3,601 | NO |
| Hormone(s) | 2,477 | NO |
| Accumulation | 2,327 | NO |
| Histology | 1,525 | NO |
| Multiple | 1,474 | NO |
| Feeding Behavior | 1,322 | YES |
| Acute | 879 | YES |
| Growth Rate | 863 | YES |
| No Effect | 715 | NO |
| Chronic | 708 | YES |
| Area Under Growth Curve | 672 | NO |
| Injury | 486 | NO |
| Immunological | 408 | NO |
| Avoidance | 372 | NO |
| Not Reported | 308 | NO |
| Biomass | 254 | YES |
| Ecosystem Process | 241 | NO |
| Prolonged | 227 | NO |
| Other | 135 | NO |
| Immobilisation | 126 | YES |
| Frond Number | 94 | YES |
| Mobility | 71 | YES |
| Seedling Emergence | 10 | YES |
| Body Weight | 9 | YES |
| Behaviour | 7 | YES |
|  | 7 | NO |
| No Adverse Effect Observed At Single/highest Dose | 5 | NO |
| Time to Hatch | 4 | NO |
| Time to Swim Up | 2 | NO |
| Clinical Signs | 1 | NO |

Table A9. Number of records included (Ecological relevance = YES) and removed (Ecological relevance = NO) depending on test effect criterion

### 3.3.5 Selecting relevant taxonomic annotations

Binomial taxonomic descriptions are required at species level to not misrepresent the presence of an “artificial” species in the downstream calculations. The incorrect species annotations and number of records removed are shown in Table A10.

| Species | Number of records | | | Species | Number of records | | | Species | Number of records | | Species | Number of records |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Algae | | 709 | Osteichthyes | | | 18 | Goniobasis sp. | | | 5 | Megapodagrionidae | 2 |
| Invertebrates | | 399 | Turbellaria | | | 18 | Hemiptera | | | 5 | Mugilidae | 2 |
| Plantae | | 230 | Dytiscidae | | | 17 | Hydropsychidae | | | 5 | Nemertea | 2 |
| Aquatic Community | | 211 | Hydracarina | | | 17 | Neoniphargus sp. | | | 5 | Palustra laboulbeni | 2 |
| Cladocera | | 190 | Isopoda | | | 17 | Tanytarsini | | | 5 | Peritrichida | 2 |
| Chironomidae | | 165 | Chydoridae | | | 16 | Canna x generalis | | | 4 | Physidae | 2 |
| Maxillopoda | | 162 | Coleoptera | | | 15 | Cryptophyceae | | | 4 | Planariidae | 2 |
| Rotifera | | 134 | Daphniidae | | | 15 | Decapoda | | | 4 | Poaceae | 2 |
| Cyclopoida | | 109 | Bosminidae | | | 14 | Hydrachnidae | | | 4 | Trebouxiophyceae | 2 |
| Copepoda | | 106 | Clitellata | | | 14 | Limnephilidae | | | 4 | Wilhelmia equina | 2 |
| Fungi | | 82 | Macrothricidae | | | 14 | Magnoliophyta | | | 4 | Caenidae | 1 |
| Calanoida | | 81 | Bivalvia | | | 13 | Monera | | | 4 | Calanidae | 1 |
| Plankton | | 69 | Ancylidae | | | 12 | Parathelphusidae | | | 4 | Characidae | 1 |
| Animalia | | 68 | Clariidae | | | 12 | Phaeophyta | | | 4 | Chironominae | 1 |
| Chaoboridae | | 68 | Gerridae | | | 12 | Planorbidae | | | 4 | Cichlidae | 1 |
| Oligochaeta | | 65 | Orthocladiinae | | | 11 | Sagittaria sp. | | | 4 | Cnidaria | 1 |
| Ostracoda | | 63 | Phryganeidae | | | 11 | Talitridae | | | 4 | Collembola | 1 |
| Anculosa sp. | | 60 | Simuliidae | | | 11 | Acarina | | | 3 | Collothecaceae | 1 |
| Arthropoda | | 60 | Ampullariidae | | | 10 | Arachnida | | | 3 | Desmidiaceae | 1 |
| Amphipoda | | 55 | Ephemerella sp. | | | 10 | Branchiostomidae | | | 3 | Dolichorhynchus sp. | 1 |
| Tubificidae | | 47 | Nematocera | | | 10 | Cyclopidae | | | 3 | Echinosphaerium sp. | 1 |
| Insecta | | 46 | Anura | | | 9 | Diplogasteritus | | | 3 | Eichhornia sp. | 1 |
| Crustacea | | 44 | Basommatophora | | | 9 | Haitia pomilia | | | 3 | Elmidae | 1 |
| Culicidae | | 44 | Polycentropodidae | | | 9 | Harpacticoida | | | 3 | Empididae | 1 |
| Corixidae | | 38 | Tanypodinae | | | 9 | Hirudinea | | | 3 | Ephemeridae | 1 |
| Ephemeroptera | | 36 | Anisoptera | | | 8 | Texadina sp. | | | 3 | Ephydridae | 1 |
| Ceratopogonidae | | 35 | Chlorophyta | | | 8 | Veliidae | | | 3 | Erpobdellidae | 1 |
| Odonata | | 34 | Heteroptera | | | 8 | Aeshnidae | | | 2 | Gasterosteidae | 1 |
| Trichoptera | | 34 | Notonectidae | | | 8 | Anacystis alpicola | | | 2 | Hydrachnellae | 1 |
| Gastropoda | | 32 | Chlorophyceae | | | 7 | Atheriniformes | | | 2 | Hymenoptera | 1 |
| Plecoptera | | 32 | Mollusca | | | 7 | Austrophlebia sp. | | | 2 | Newnhamia sp. | 1 |
| Diptera | | 31 | Nimbocera sp. | | | 7 | Blephariceridae | | | 2 | Noteridae | 1 |
| Bacillariophyceae | | 30 | Salmonidae | | | 7 | Chloroperlidae | | | 2 | Petrothrincidae | 1 |
| Hydrophilidae | | 29 | Zygoptera | | | 7 | Chrissia halyi | | | 2 | Pisidiidae | 1 |
| Cyprinidae | | 24 | tribe Chironomini | | | 7 | Chrysophyceae | | | 2 | Pleuroceridae | 1 |
| Nuria danrica | | 24 | Amphibia | | | 6 | Ciliophora | | | 2 | Podonominae | 1 |
| Rasbora daniconius neilgeriensis | | 24 | Cambaridae | | | 6 | Conchostraca | | | 2 | Pyrrophycophyta | 1 |
| Bacillariophyta | | 22 | Characeae | | | 6 | Cryptomonadida | | | 2 | Scuticociliatida | 1 |
| Naididae | | 21 | Chrysophyta | | | 6 | Diatomaceae | | | 2 | Siluriformes | 1 |
| Baetidae | | 20 | Cobitidae | | | 6 | Dinophyceae | | | 2 | Sphaerodema sp. | 1 |
| Protozoa | | 20 | Nematoda | | | 6 | Euglenophyceae | | | 2 | Tipulidae | 1 |
| Heptageniidae | | 19 | Centrarchidae | | | 5 | Kurzia sp. | | | 2 | Umbridae | 1 |
| Coenagrionidae | | 18 | Chromista | | | 5 | Lemonniera | | | 2 |  |  |
| Hormidium sp. | | 18 | Ciliatea | | | 5 | Libellulidae | | | 2 |  |  |

Table A10. Number of records with invalid 'Species'-annotations which are removed from the dataset

### 3.3.6 Effect concentration and unit harmonization

Effect data are tested and reported with heterogenous concentrations and units. To harmonize tests respective concentration into mg/L, simple conversions are needed, where possible. Effect data comes reported with 197 different units and some cannot be converted into mg/L, i.e., “ug/cell” or “gal/acre”. Data with non-convertible units are thus removed.

| Original.Units | Operator | Conversion.factor | Final.unit |
| --- | --- | --- | --- |
| ppm; µg/cm³; mg/L; µg/mL; g/m³; mg/dm³ | × | 1 | mg/L |
| mg/200mL | × | 5 | mg/L |
| mg/100cm³ | × | 10 | mg/L |
| g/L; g/dm³; mg/mL; µg/mm³ | × | 1000 | mg/L |
| g/mL | × | 1000000 | mg/L |
| mmol/L; mmol/dm³; mol/m³; mM | × | Mol Weight | mg/L |
| mol/L; mol/dm³; M; mol | × | Mol Weight × 1000 | mg/L |
| µmol/L; µmol/dm³; uM/L; µM/L; mmol/m³; nmol/mL; µM; mmol; µm | × | Mol Weight / 1000 | mg/L |
| nmol/L; nM/L; mM; "nmol" | × | Mol Weight /1000000 | mg/L |
| pM | × | Mol Weight /1000000000 | mg/L |
| µg/3.5L | / | 0,000285714 | mg/L |
| µg/5mL | / | 0,2 | mg/L |
| µg/100mL | / | 100 | mg/L |
| µg/L; ng/mL; µg/dm³; ppb | / | 1000 | mg/L |
| µg/10L | / | 10000 | mg/L |
| ng/L;pg/mL; µg/µL | / | 1000000 | mg/L |
| pg/L | / | 1000000000 | mg/L |

Table A11. Conversions from the various units effect data comes reported in, into mg/L

### 3.3.7 Test durations

Due to the heterogeneity of the gathered data, test duration is reported as either “duration” or “exposure duration”, and in some cases both. The defined “exposure duration” is prioritized, but if missing, “duration” is used. There are 23 different units of test duration reported, yet only discrete time is used for the current study (seconds(s), hours (h), days (d), weeks (wk), months (mo), years (yr); Table A12), while all records with other units are discarded. All durations are subsequently converted into hours.

| Duration.Unit | Number of records |
| --- | --- |
| h | 174,801 |
| d | 88,756 |
| wk | 7,211 |
| min | 4,907 |
|  | 2,657 |
| mo | 1,883 |
| yr | 396 |
| Brood or litter | 250 |
| Generation | 249 |
| Stage | 214 |
| Second(s) | 190 |
| Gosner stage | 177 |
| Nieuwkoop-faber-stage | 146 |
| Instar | 86 |
| Degree days | 53 |
| Crab stage | 34 |
| Growing season | 19 |
| Zoeae - megalop | 15 |
| Fry | 6 |
| Until hatch | 3 |
| Harvest | 1 |
| Lifetime;no associated numeric value | 1 |
| Maturity | 1 |

Table A12. Number of records per reported experiment duration unit

### 3.3.8 Assign acute or chronic test definitions.

Test duration and test concentration units were harmonized into hours and respectively, and records not possible to convert were discarded. Follow the general methodology of Aurisano et al. (2019) we consider acute effects at exposure duration ≤24 h for the taxonomic groups algae (including cyanobacteria), rotifers, and microorganisms (all phyla included in the taxonomic group “Others”, except for Chordata, Arthropoda, and Cnidaria), ≤96 h for phylum crustaceans and arthropods, and ≤168 h for fish, invertebrates (noncrustaceans), vertebrates, and aquatic plants other than algae according to Aurisano et al. (2019). This implies that for the crustacean family *Daphniidae*, a data rich group of organisms throughout the data set, tests duration > 96h are defined as chronic, despite the OECD TG 211 guidelines for Daphnia magna reproduction tests are considered chronic at 21 days (504 h) (OECD 2012). Note that we chose to limit the data set to records with a test duration 24 hours, thus removing all “acute” tests for Algae and Rotifera.

| Taxonomy.Group | Phylum | Acute | Chronic | n |
| --- | --- | --- | --- | --- |
| Fish | Chordata | <= 168 | > 168 | 50,513 |
| Crustacean | Arthropoda | <= 96 | > 96 | 27,176 |
| Algae | Chlorophyta | <= 24 | > 24 | 9,278 |
| Insect | Arthropoda | <= 168 | > 168 | 8,258 |
| Amphibian | Chordata | <= 168 | > 168 | 5,504 |
| Plant | Streptophyta | <= 168 | > 168 | 4,708 |
| Mollusca | Mollusca | <= 168 | > 168 | 4,253 |
| Others | Ciliophora | <= 24 | > 24 | 1,592 |
| Algae | Cyanobacteriota | <= 24 | > 24 | 1,275 |
| Rotifera | Rotifera | <= 24 | > 24 | 1,170 |
| Algae | Bacillariophyta | <= 24 | > 24 | 1,059 |
| Annellidae | Annelida | <= 168 | > 168 | 936 |
| Others | Platyhelminthes | <= 24 | > 24 | 743 |
| Others | Nematoda | <= 24 | > 24 | 403 |
| Others | Cnidaria | <= 168 | > 168 | 238 |
| Others | Euglenozoa | <= 24 | > 24 | 163 |
| Others | Chordata | <= 168 | > 168 | 154 |
| Others | Ascomycota | <= 24 | > 24 | 131 |
| Others | Discosea | <= 24 | > 24 | 107 |
| Others | Cyanobacteria | <= 24 | > 24 | 104 |
| Others | Tracheophyta | <= 24 | > 24 | 96 |
| Algae | Ochrophyta | <= 24 | > 24 | 47 |
| Algae | Sar | <= 24 | > 24 | 47 |
| Algae | Haptophyta | <= 24 | > 24 | 32 |
| Algae | Cryptophyta | <= 24 | > 24 | 30 |
| Algae | Dinophyta | <= 24 | > 24 | 29 |
| Algae | Charophyta | <= 24 | > 24 | 17 |
| Others | Arthropoda | <= 96 | > 96 | 17 |
| Others | Chytridiomycota | <= 24 | > 24 | 16 |
| Algae | Rhodophyta | <= 24 | > 24 | 8 |
| Others | Bryozoa | <= 24 | > 24 | 7 |
| Others | Nematomorpha | <= 24 | > 24 | 5 |
| Others | Gastrotricha | <= 24 | > 24 | 4 |
| Others | Basidiomycota | <= 24 | > 24 | 3 |
| Others | Mucoromycota | <= 24 | > 24 | 3 |
| Others | Perkinsozoa | <= 24 | > 24 | 3 |
| Others | Oomycota | <= 24 | > 24 | 2 |

Table A13. Acute and chronic definitions per taxonomic group and phylum. Definitions for Acute and Chronic effect are shown as hours

### 3.3.9 EC10eq conversions

The last step of curating data from the OECD QSAR Toolbox query was to apply regression coefficients for endpoint conversion of effect data from acute EC50 and NOEC, and chronic EC50 and NOEC into chronic EC10eq following the suggested coefficients from (Aurisano et al. 2019) (Table A14).

| Assigned endpoint | Acute or Chronic | Taxonomic group | Extrapolation factor | 97.5% CI | 2.5% CI |
| --- | --- | --- | --- | --- | --- |
| EC50 | Acute | Fish | 7.44 | 18.95 | 2.92 |
| EC50 | Acute | Crustacean | 3.38 | 5.34 | 2.14 |
| EC50 | Acute | Algae | 4.00 | 6.10 | 2.60 |
| EC50 | Acute | Others | 4.00 | 6.10 | 2.60 |
| EC50 | Chronic | Fish | 1.55 | 3.66 | 0.67 |
| EC50 | Chronic | Crustacean | 1.94 | 2.41 | 1.56 |
| EC50 | Chronic | Algae | 2.24 | 2.65 | 1.90 |
| EC50 | Chronic | Others | 2.00 | 2.50 | 1.80 |
| EC10 | Acute | Fish | 1.00 | 1.00 | 1.00 |
| EC10 | Acute | Crustacean | 1.00 | 1.00 | 1.00 |
| EC10 | Acute | Algae | 1.00 | 1.00 | 1.00 |
| EC10 | Acute | Others | 1.00 | 1.00 | 1.00 |
| EC10 | Chronic | Fish | 1.00 | 1.00 | 1.00 |
| EC10 | Chronic | Crustacean | 1.00 | 1.00 | 1.00 |
| EC10 | Chronic | Algae | 1.00 | 1.00 | 1.00 |
| EC10 | Chronic | Others | 1.00 | 1.00 | 1.00 |
| NOEC | Acute | Fish | 3.97 | 17.39 | 0.90 |
| NOEC | Acute | Crustacean | 1.55 | 2.64 | 0.91 |
| NOEC | Acute | Algae | 1.80 | 2.70 | 1.00 |
| NOEC | Acute | Others | 1.80 | 2.70 | 1.00 |
| NOEC | Chronic | Fish | 0.60 | 0.70 | 0.40 |
| NOEC | Chronic | Crustacean | 0.95 | 1.16 | 0.77 |
| NOEC | Chronic | Algae | 0.44 | 0.49 | 0.39 |
| NOEC | Chronic | Others | 0.60 | 0.70 | 0.40 |

Table A14. Conversion factors applied for respective endpoints for conversion into 'EC10eq', modified from Aurisano et al., 2019

# 4 Results

## 4.1 Ecotoxicological database curation summary

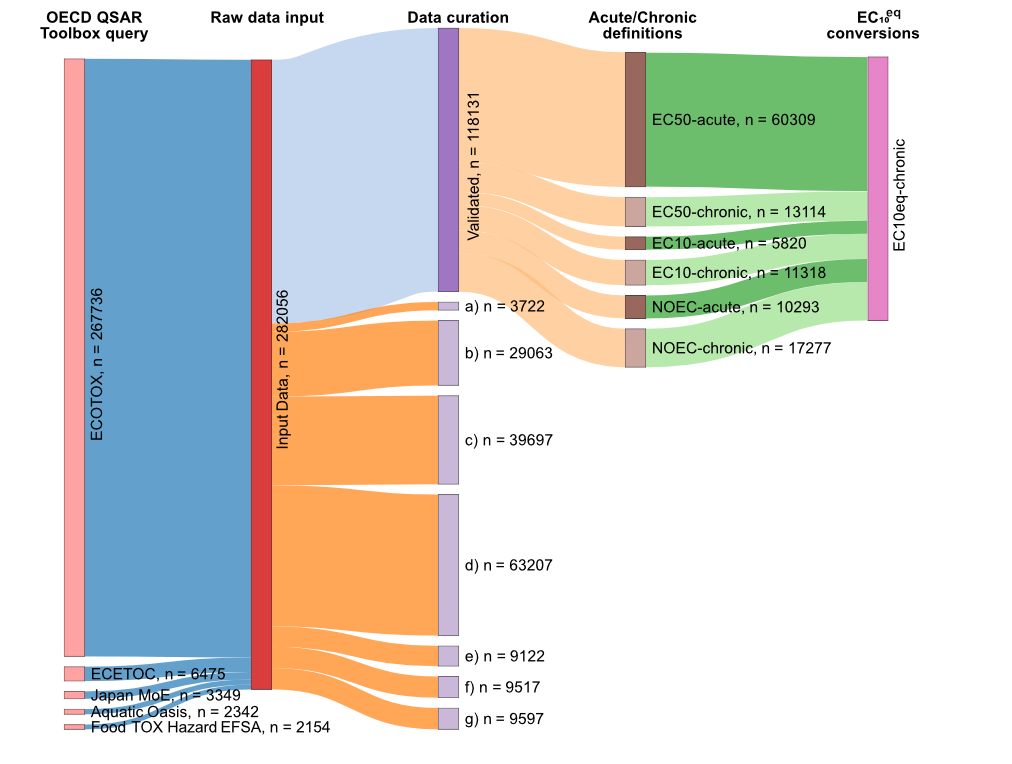


Figure A2. Overview of the data curation process from respective number of records (n) per database, stepwise removal of records, and the counts of acute or chronic records per endpoint in the ‘validated data’-category. Data removal occured in the following steps: a) Data reported as a range are conditionally excluded, b) Controls marked as insufficient or unsatisfactory, or effect data are based on QSAR estimations or bioassays, c) Non-fresh water data, d) Effect criterions irrelevant, e) Poor taxonomic descriptions, f) Effect unit or effect value missing or reported as 0, g) Test duration missing or < 24h

| Taxonomy group | EC10 acute | EC10 chronic | EC50 acute | EC50 chronic | NOEC acute | NOEC chronic | Total per taxa |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Algae | 82 | 2,278 | 454 | 6,080 | 141 | 2,787 | 11,822 |
| Amphibian | 553 | 539 | 2,183 | 272 | 631 | 1,326 | 5,504 |
| Annellidae | 76 | 39 | 676 | 32 | 33 | 80 | 936 |
| Crustacean | 915 | 2,897 | 14,466 | 2,114 | 2,269 | 4,515 | 27,176 |
| Fish | 2,936 | 3,422 | 31,598 | 1,432 | 5,903 | 5,222 | 50,513 |
| Insect | 249 | 279 | 6,316 | 347 | 358 | 709 | 8,258 |
| Mollusca | 208 | 319 | 2,596 | 185 | 295 | 650 | 4,253 |
| Others | 88 | 547 | 787 | 1,528 | 130 | 711 | 3,791 |
| Plant | 708 | 728 | 945 | 889 | 526 | 912 | 4,708 |
| Rotifera | 5 | 270 | 288 | 235 | 7 | 365 | 1,170 |
| Total | 5,820 | 11,318 | 60,309 | 13,114 | 10,293 | 17,277 | 118,131 |

Table A15. Number of records present in the curated ecotoxicological database presented by taxonomic group and toxicological endpoint.

| Chemical use category | Number of records | Number of chemicals |
| --- | --- | --- |
| Antibiotic | 943 | 48 |
| Other inorganic chemicals | 3,134 | 108 |
| Other organic chemicals | 17,288 | 1,888 |
| PPCP | 5,712 | 212 |
| Pesticide | 91,054 | 1,436 |
| Total | 118,131 | 3,692 |

Table A16. Number of toxicological records and chemicals respecively categorized into defined chemical use categories. PPCPs = Pharmaceuticals and Personal Care Products.

# File connectivity schema

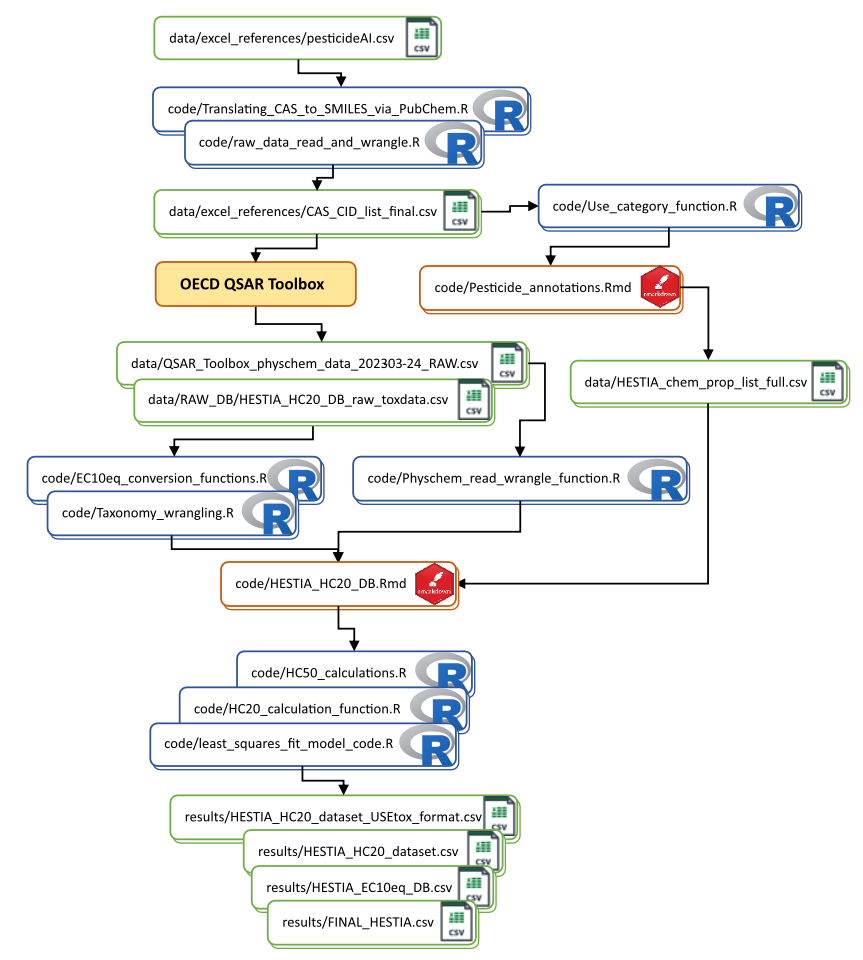


Figure A3. A schematic diagram of the database construction focusing on the connectivity of the various code-files in the GitHub repository

# References

Aurisano, Nicolò, Paola Federica Albizzati, Michael Hauschild, and Peter Fantke. 2019. “Extrapolation Factors for Characterizing Freshwater Ecotoxicity Effects.” *Environmental Toxicology and Chemistry* 38 (11): 2568–82. <https://doi.org/10.1002/etc.4564>.

Chamberlain, Scott, and Eduard Szocs. 2013. “Taxize - Taxonomic Search and Retrieval in r.” *F1000Research*. <https://f1000research.com/articles/2-191/v2>.

Dimitrov, SD, R Diderich, T Sobanski, TS Pavlov, GV Chankov, AS Chapkanov, YH Karakolev, et al. 2016. “QSAR Toolbox–Workflow and Major Functionalities.” *SAR and QSAR in Environmental Research* 27 (3): 203–19.

Fantke, P., M. Bijster, C. Guignard, M. Hauschild, M. Huijbregts, O. Jolliet, A. Kounina, et al. 2017. “USEtox 2.0, Documentation Version 1.” <https://doi.org/10.11581/DTU:00000011>.

OECD. 2012. *Test No. 211: Daphnia Magna Reproduction Test*. https://doi.org/<https://doi.org/https://doi.org/10.1787/9789264185203-en>.

Paulus, Gabriela K, Luc M Hornstra, Nikiforos Alygizakis, Jaroslav Slobodnik, Nikolaos Thomaidis, and Gertjan Medema. 2019. “The Impact of on-Site Hospital Wastewater Treatment on the Downstream Communal Wastewater System in Terms of Antibiotics and Antibiotic Resistance Genes.” *International Journal of Hygiene and Environmental Health* 222 (4): 635–44.

Raychaudhuri, Sarmistha Sen, Paulami Pramanick, Pratik Talukder, and Apaala Basak. 2021. “Polyamines, Metallothioneins, and Phytochelatins—Natural Defense of Plants to Mitigate Heavy Metals.” *Studies in Natural Products Chemistry* 69: 227–61.

Saouter, Erwan, Fabrizio Biganzoli, Lidia Ceriani, Donald Versteeg, Eleonora Crenna, Luca Zampori, Serenella Sala, and Rana Pant. 2018. “Environmental Footprint: Update of Life Cycle Impact Assessment Methods : Ecotoxicity Freshwater, Human Toxicity Cancer, and Non-Cancer.” Luxembourg: Publications Office of the European Union.

Szöcs, Eduard, Tamás Stirling, Eric R. Scott, Andreas Scharmüller, and Ralf B. Schäfer. 2020. “webchem: An R Package to Retrieve Chemical Information from the Web.” *Journal of Statistical Software* 93 (13): 1–17. <https://doi.org/10.18637/jss.v093.i13>.