Supplementary information

Oskar Nyberg

2023-07-05

Table of Contents

Oskar Nyberg

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

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

# 1. Pre-processing of the chemical inventory

## 1.1 The HESTIA chemical inventory

**The HESTIA project is cool, but needs a description here**  
… and this is where we extract the chemical inventory as the file data/excel\_references/pesticideAI.csv, containing 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.
* “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.

## 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. In finding matches, it is realized that chemical naming is different between the pesticideAI.CSV and the PC repository i.e., spelling of”sulfate” in pesticideAI.csv vs. “sulphate” in the PC repository. After substitution of “sulf ->”sulph” and 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.  
For the remaining 2259 chemicals, multiple PC ID matches from the CASRN-query is 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 the file “data/excel\_references/CAS\_CID\_list\_final.csv” as a complete list of 16797 chemicals, with 15030 identified SMILES-configurations. A .txt-file (“data/excel\_references/CAS\_CID\_list\_final.txt”) with 16797 rows containing only CASRN and SMILES-configuration is also 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 to be used in the downstream data queries from OECD QSAR Toolbox. Due to memory issues with the VEGA QSAR-software and when compiling ecotoxicological data in OECD QSAR Toolbos, 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. Querying OECD QSAR Toolbox

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, the file “data/excel\_references/CAS\_CID\_list\_final.txt” 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, and 16788 chemical structures are imported.

## 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 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 which are required by USEtox v2.1 for characerizations of chemicals. “Exp”-implying data comes from experimental sources, others are estimated using *in silico* methods, explained below.

### 2.1.1 Molecular weight (MolW)

Molecular weight are 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 S?? 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 2017).

| a | b | c | d |
| --- | --- | --- | --- |
| >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 |

### 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 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 S1

Table 1: Numer of records returned from OECD QSAR Toolbox. Database = NA imply that no matches were found for a chemical.

| Database | Total 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 |
| Number of records | 441,206 |

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 S2

Table 2: 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.

| 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 |

# 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.

## 3.2 Taxonomic classification

HESTIA\_HC20\_DB <- read.csv("data/RAW\_DB/HESTIA\_HC20\_DB\_raw\_toxdata.csv")  
  
selection\_list <- c("Test.organisms..species.", "phylum", "class", "order", "genus", "species", "source", "query")  
taxonomy <- read.csv("data/Taxonomy/Species\_taxonomy.csv")

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 (Scott Chamberlain and Eduard Szocs 2013). The following operations are performed to produce a coherent list fo 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, 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 r´nrow(taxonomy %>% distinct(Species)), observe that some taxa were counted twice as both outdated and current names within the starting data.

taxonomy %>%   
 count(Taxonomy.Group)

## Taxonomy.Group n  
## 1 Algae 548  
## 2 Amphibian 150  
## 3 Annellidae 94  
## 4 Crustacean 635  
## 5 Fish 680  
## 6 Insect 565  
## 7 Mollusca 408  
## 8 Others 561  
## 9 Plant 210  
## 10 Rotifera 70  
## 11 <NA> 285

Table of contents:

1. Pre-processing of chemical inventory  
   1.1 The HESTIA chemical inventory  
   1.2 CASRN to SMILES identification
2. Querying OECD QSAR Toolbox  
   2.1 Physicochemical data gathering  
   2.2 Ecotoxicological effect data
3. Data processing 3.1 Chemical use categorization  
   3.2 Taxonomic classification

# References

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, Bijster, P. 2017. “USEtox 2.0, Documentation Version 1.” <https://doi.org/10.11581/DTU:00000011>.

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

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>.