Toxicological database construction and uncertainty analysis of toxicological effect data used in life cycle assessment

Oskar Nyberg

2023-06-26

Oskar Nyberg, Reinout Heijungs, Patrik Henriksson

Department of Ecology, Environment and Plant Sciences, Stockholm University, Stockholm, Sweden Department of Operations Analytics, Vrije Universiteit, Amsterdam, The Netherlands  
 Institute of Environmental Sciences, Leiden University, Leiden, The Netherlands  
 Stockholm Resilience Centre, Stockholm, Sweden  
 WorldFish, Jalan Batu Maung, Penang, Malaysia  
 Beijer Institute of Ecological Economics, The Royal Swedish Academy of Science, Stockholm, Sweden

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

## Abstract

We construct a database from a large set of openly available ecotoxicological effect data for a large set of chemicals with potential negative environmental impact and calculate ecotoxicological effect values for use in life cycle impact assessment. We then model the uncertainties at the values for all chemicals that have sufficient effect data records available using a NLS model based on the cumulative normal distribution function. We finally explore the potential to complement or replace ecotoxicological effect data with estimations generated from quantitative structure-activity relationship models.

## Introduction

Increasing chemical use is a major concern for operating within the safe space of planetary boundaries, and agriculture is a major driving force behind this increase, through pesticide use, veterinarian drugs, and disinfectants (Gordon et al. 2017; Persson et al. 2022). Chemical use in agriculture is also expected to continue to increase in the coming decade, as a result of larger production volumes and more intensive production systems (Schreinemachers and Tipraqsa 2012). However, data on potentially toxic chemicals used for different farming systems, toxicity evaluations for different compounds, and variability in toxicity potentials are incomplete and fragmented, hampering accurate global comparisons of ecotoxicity impacts of different food products (Der Werf, Knudsen, and Cederberg 2020). Accurate toxicological characterization of chemicals is essential for ensuring the safety of human health and the environment, where proper toxicological characterization involves identifying and evaluating the potential adverse effects of chemicals, determining the level of exposure that is safe for humans and the environment, and assessing the risk posed by exposure to chemicals (Krewski et al. 2010).

Environmental Risk Assessment (ERA) has been the primary tool for evaluating toxicological risks at fields and farms, but it is limited to individual processes in production chains (e.g. grow-out). This can be limiting when trying to understand the overall impacts of food products, including e.g. all ingredients used to produce animal compound feeds. Subsequently, life cycle assessments (LCAs) have increasingly been used as a complementing framework to benchmark toxicological impacts of agrifood products. LCA is an ISO-standardized environmental framework that seeks to aggregate the emissions and products that are needed throughout a value chain, and characterize these towards one or more environmental impacts. The results are also scaled to a predefined unit of reference (functional unit), which allows for comparisons to be made between products (e.g. in terms of kg food), functions (e.g. in terms of kcal), or services (e.g. washing one plate). It is most commonly encountered as the methodology behind carbon footprints, but is also used to quantify freshwater consumption, land occupation, eutrophication, biodiversity loss, or toxicity impacts (Hauschild and Huijbregts 2015). Toxicity impacts, in turn, are commonly evaluated in terms of human toxicity, either including or excluding cancer cases, and ecotoxicity impacts on freshwater, marine, or terrestrial ecosystems (Hauschild and Huijbregts 2015).

LCA is cruder than ERA when it comes to evaluating toxicological effects, as it disregards temporal scales and critical concentrations, assumes inputs equal to emissions, simplifies assumptions related to release location and exposure, and disregards ‘cocktail effects’ between chemicals (Fantke et al., 2015). In the meantime, it can capture toxicological impacts throughout whole value chains, from mercury emissions from coal-fired power plants to therapeutants used in aquaculture ponds. LCAs can thereby provide useful insights into where in value chains the largest toxicity reductions can be achieved, such as that most freshwater ecotoxicological impacts related with prawn farming in Bangladesh are related to the production certain agricultural materials used for feeds, and not the prawn grow-out (Henriksson et al. 2015).

To assess ecotoxicological impacts throughout the value chain of a product, ecotoxicological characterization factors for chemicals are used to evaluate potential impact of different chemicals in the life cycle impact assessment (LCIA) phase of an LCA. Toxicological effect data for chemicals provide, together with data on fate and exposure, the basis for characterization factor calculations within LCA. These data, however, are notoriously heterogeneous since they are reported across thousands of tested species at variable concentrations, measured effects, and at various empirical or modeled endpoints, resulting in large uncertainty and variability (which we will from here on simply refer to as “uncertainty”) in characterization factor calculations. To investigate uncertainties in the ecotoxicological effect data that is used to calculate ecotoxicological characterizations we have queried the OECD QSAR Toolbox (Dimitrov et al. 2016) for ecotoxicological effect data for the 16797 chemicals included in the HESTIA inventory (<https://www.hestia.earth>) and gathered all non-proprietary records available. Additional data were sourced from EnviroTox, a curated ecotoxicological database for 4267 substances (Connors et al. 2019).

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. The ambition of HESTIA is to make environmental benchmarks of agrifood commodities more accessible and transparent, by providing a free harmonized framework. Given that the agrifood sector is a major user of potentially toxic chemicals, mainly in terms of pesticides and therapeutants, we deem it important to provide a complete and verifiable set of ecotoxicity potentials as possible.

The evaluation of toxicological impacts in LCA is limited by the number of chemical compounds characterized by impact assessment methodologies. The characterization factor is the value used to translate the amounts of chemicals used to its potential toxicity impact. There are several different impact assessment methodologies to derive these characterization factors, including USES-LCA v1&2 (Huijbregts et al. 2000; Van Zelm, Huijbregts, and De Meent 2009), IMPACT 2002 (Pennington et al. 2005), and UNEP-SETAC’s USEtox (B. Fantke P. 2017), but ultimately they all rely on fundamental data on toxicity and physicochemical properties. Among these impact assessment methodologies, the USEtox model is most widely used at present, and also the one promoted by the European Platform on Life Cycle Assessment (ILCD, 2010). The USEtox model is currently on version 2.1, and readily presents 2499 freshwater ecotoxicological characterization factors for different chemicals, derived from physicochemical properties and empirical data at the 50% (EC50) endpoint, presented as the 50th percentile response level on the species-sensitivity distribution (SSD) model curve, denominated the HC50 value (B. Fantke P. 2017). For chemicals not readily characterized, the USEtox model also allows users to derive their own characterization factors through an Excel spreadsheet, but this is, in turn, dependent on access to physicochemical and toxicological properties for each chemical.

Based on a series of recent expert workshops, updated recommendations on which toxicological input data should form the SSD slopes for characterization of substances has been published (Owsianiak et al., -Mikolaj Owsianiak et al. (2019)). The new recommendations include using effect data at concentrations within a range similar to ambient environmental concentrations to construct SSD models which translates into usage of the endpoints no observed effect concentration (NOEC), lowest observed effect concentration (LOEC), or effect concentrations at 0, 10 and 50% response level (EC0, EC10 & EC50) endpoints when constructing SSD curves. This necessitates harmonization of data given as different endpoints into one coherent effect equivalent, recommended as effect concentration at the 10% response level equivalent (EC10eq; Aurisano et al., -Aurisano et al. (2019)). Additionally, recommendations set the working point for toxicological effect factor at the 20th percentile response level using EC10eq data under the assumption that ecotoxicological data is log-normally distributed resulting in effect data being log(10)-transformed, hence this point is called . From this point value, it is easy to calculate the concentration-response slope factor at the 20% response level () which is multiplied by a severity factor to derive the effect factor (EF) which, in turn, is multiplied with environmental fate, and ecosystem exposure products to produce characterization factors for chemicals emitted to the environment through a products’ life cycle (Owsianiak et al., (2022)).

In October 2022, Sala et al. (2022) published the [EU environmental footprint (EF v3.1) database](https://eplca.jrc.ec.europa.eu/ecotox.html) database constructed from toxicological data from three repositories; the OpenFoodTox database, the Pesticide Property Database, the Registration, Evaluation and Restriction of Chemicals (REACH; European Chemicals Agency, <http://echa.europa.eu/>) database, as well as the USEtox v2.1 database for records missing in the previous repositories. The EF v3.1 database is substantially (140%) larger than the former USEtox v2.1 database of toxicological characterization factors, easily accessible online and presents substances’ physicochemical and toxicological data as well as characterization factors for 6,038 chemicals based on values (Saouter et al. 2019; Sala et al. 2022). While this is a big improvement, it is still far from the complete list of potentially toxic chemicals encountered in agrifood production. However, parts of the toxicological records used to calculate values are defined as proprietary substance registrations within REACH database, are thus protected by confidentiality agreements and have been made unavailable to the public. This is unfortunate, as one of the 19 key recommendations of the Ecotoxicity Task Force and the Pellston workshop that was held in 2018 in Valencia, Spain is to “use data that has a traceable origin” (Owsianiak et al., (2022)). Since the present HESTIA database was under construction at the release of EF v3.1, data sources overlap only to a minor extent, and original data sources are accessible within the HESTIA database, the continued assembly of the HESTIA database was motivated.

The objective of this article is to calculate values according to the most recent modeling recommendations for all openly available toxicological effect data and using this data set investigate and report statistical uncertainty for the data that form the basis for characterization factor calculations. Our approach is to apply weighted means to species-specific effect concentration averages when constructing SSD curves and to explore the uncertainty in values a using a non-linear least squares fit model to a cumulative normal distribution of data. Additionally, we estimate ecotoxicological data based on quantitative structure-activity relationships (QSAR) for evaluating the applicability of ecotoxicological effect data in covering more substances or complement toxicological data based on animal testing. Finally, data are reported as a **shiny-application at website:X** to better inform environmental impact calculations for agrifood LCAs.

## Methods

### HESTIA data collection

The starting point for constructing a database with ecotoxicological effect data to calculate characterization factors suited for the online life cycle assessment (LCA) application HESTIA ([http://HESTIA.earth](http://hestia.earth)) is the substance inventory; a list of 16797 CAS registry numbers (CASRN) and matching chemical names of potentially harmful substances *based on the United States Environmental Protection Agency’s Substance Registry Services (USEPA SRS) inventory.*

CASRN and chemical names were queried to match SMILES configurations acquired from the NCBI PubChem database using the R package Webchem (Szöcs et al. 2020). 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 and an atomic number >20 (Raychaudhuri et al. 2021).  
**multi-constituent chemicals are neglected throughout the whole project due to difficulties in collecting correct data.**  
Based on the CASRN-SMILES matches three distinct data queries were taken:

1. Information on chemical use-categories of chemicals in the HESTIA inventory were gathered from several repositories (see supporting materials):  
   *a) US EPA ECOTOX DB (downloaded on 2022-03-10) where information in “ecotox\_group” was extracted using CASRN,*  
   *b) the British Compendium of Pesticide Common Names (BCPC) accessed via R package “Webchem” (Szöcs et al. 2020) and matched to CASRN,*  
   *c) the Anatomical Therapeutic Chemical (ATC) classification system inventory was compared to chemical names in groups P (antiparasitic products, insecticides and repellents) and J (antiinfectives for systemic use) to define chemicals as “Pesticides” and “Antibiotics” respectively. The ATC classification system inventory was downloaded via R package “Webchem” function “chembl\_atc\_classes” (Szöcs et al. 2020),*  
   *d) the Chemical Entities of Biological Interest (*[*ChEBI*](https://www.ebi.ac.uk/chebi/init.do)*) was queried for the definition “has role” where use-classification was extracted and matched to CASRN using R package “Webchem” (Szöcs et al. 2020), and*   
   *e) USEPA CompTox v2.2 (query date 2023-01-20) by searching “chemical lists” -> “list names” using the queries “pesticides”, “anti” (to select lists for “antibiotic”, “antifungal”, and “antiviral”-chemicals ), “pharmaceutical”, “herbicide”, “insecticide”, and “rodenticide” where the following inventories were compiled into one data set (contents 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”), OPPIN (content marked as “Pesticide”), Pesticide properties Database (PPDB; content marked as “Pesticide”), USEPA (content marked as “Antibiotic”, “Pesticide ingredient”, or “Pesticide”), and Wikipedia (content marked as “Antibiotic”, “Antifungal”, “Antiviral”, “Insecticide”, “Herbicide”, “Rodenticide”, or “Veterinary drug”) for each substance respectively to categorize chemicals into groups: Antibiotic, Antiviral, Other inorganic chemicals, Other organic chemicals, PPCP, Pesticide, Pharmaceutical, and Unknown.*
2. Query the OECD QSAR Toolbox v4.5 (Dimitrov et al. 2016) for physicochemical properties required by USEtox v2.1: MW.g.mol, pKa.gain, pKa.loss, Kow\_L.L, Koc\_L.kg\_MCI, kH25C\_Pa.m3.mol, Vapor.Pressure\_Pa, Sol\_mg.L, KdegA, KdegW, KdegSl, KdegSd, and BAF\_L.Kg.
3. Query the OECD QSAR Toolbox v4.5 (Dimitrov et al. 2016) for available aquatic ecotoxicological records for the entire HESTIA chemical inventory from the following toxicological endpoints: EC10, EL10, IC10, LC10, LOEC (grouped as “EC10”); EC50, EL50, IC50, LC50 (grouped as “EC50”); EC0, LC0, NOAEC, NOEC, NOER, NOEL (grouped as “NOEC”), according to the methodology of (Aurisano et al. 2019). Further curation of toxicological records include the selection of relevant effect criterions (Behavior, Growth, Intoxication, Population, Reproduction, Acute, Cell(s), Growth Rate, Mortality, Feeding Behavior, Biomass, Body Weight, Chronic, Frond Number, Development, Mobility,Seedling Emergence, and Immobilization), curation of values reported as ranges following the methodology of Saouter et al., (2018), removal of records based on QSAR data and records identified as reported bioassay data, as well as inaccurate database entries. Records where test medium is marked freshwater and culture media are kept, including blank reports after confirming that at most of those records are belong to freshwater species (e.g. *Daphnia magna*, *Pseudokirchneriella subcapitata*, *Pimephales promelas* and *Oryzias latipes*).  
   Records with taxonomic information annotated as binomial species names (i.e. *Daphnia magna*) are kept, including annotations at a lower taxonomic level, such as sub-species or *varietas*. Meanwhile, any record with taxonomic annotation higher than species level, such as genus or class or higher, are discarded as to no misrepresent a presence of an “artificial” species in the downstream calculations. Complete taxonomic information was collected for all species using R package Taxize (Scott Chamberlain and Eduard Szocs 2013), and taxonomic rank at class and phylum level used to categorize organisms into “taxonomic groups” (e.g Crustacean, Fish, Algae, Others, Plant, Mollusca, Rotifera, Insect, Amphibian, and Annellidae).  
   Test duration and test concentration units were harmonized into hours and respectively, and records not possible to convert were discarded. Acute and chronic definitions of tests were defined as chronic at 168 hours for groups Fish, Plants, Insects, Mollusca, Annellidae, and Amphibians (as well as the rare chordates and cnidarians belonging to taxonomic group “others”), 96 hours for groups Crustaceans and arthropods, and 24 hours for groups Algae and Rotifera, according to the methodology of (Aurisano et al. 2019). 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.

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 1.

Table 1: Conversion factors applied for respective endpoints for conversion into 'EC10eq', modified from Aurisano et al., 2019

| 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 |

Lastly, we merge the data set with selected (non-overlapping) records from the EnviroTox toxicological database. The EnviroTox toxicological database by Connors et al., (2019) is a curated aquatic toxicological database containing a large set of toxicological data from a broad range of potentially toxic substances. Prior to merging the two databases, curation of the EnviroTox database was performed accordingly:  
Toxicological endpoints were selected and grouped as for HESTIA database (Table STX), Acute and Chronic test duration definitions were performed according to taxonomic group and duration of experiment. The EnviroTox team have thouroughly refined acute/chronic exposure selection methodology yet for the current purpose we chose to categorize acute/chronic definitions in a harmonized way according to the methodology described above. Taxonomic information from the EnviroTox database was slightly revised to match the HESTIA data set (e.g minor spelling corrections). With harmonized acute/chronic definitions, identical taxonomic descriptions and endpoint conversions, extrapolation factors for EC10eq conversions were added to effect data as for the previous data set. Lastly, records with duplicates across all rows are removed.  
Data sources overlap across the HESTIA and EnviroTox data sets, and records are selected from EnviroTox when 1) all records with CASRN unique to EnviroTox, and 2) records where species, but not CASRN, are unique to EnviroTox. For details on the processing of data, please see supporting information (code/Pesticide\_annotations.Rmd, Physchem\_read\_wrangle\_function.R, data/raw\_data\_read\_and\_wrangle.R, and HESTIA\_HC20\_DB.Rmd).

### Exploring QSAR methods

The applicability of toxicological effect data derived from quantitative structure–activity relationship (QSAR) models generated are explored using VEGA software application (Benfenati, Manganaro, and Gini 2013). Estimated effect concentrations are based on substances’ SMILES configuration and the estimated effect data is reported with the unit mg/L along with a quality evaluation of the prediction based on the similarity to compared compounds. Model estimation quality is calculated as a compounded similarity index which evaluates similarities to validated training data sets of chemicals as an “applicability domain” (AD) index and is reported as “EXPERIMENTAL”, “GOOD”, and “MODERATE”, “LOW”, or “ERROR” for each chemical respectively (Floris et al. 2014). Since training data sets only contain organic substances, multi-constituent or metal-complexes will be given “ERROR” status. Fifteen selected QSAR models were applied across the 15030 chemicals with SMILES configuration available (see Table 6). Full documentation of models, training data and AD index parameters is provided at the [VEGA QSAR online platform](https://www.vegahub.eu/portfolio-item/vega-qsar-models-qrmf/). Estimated effect concentrations were subsequently transformed into EC10eq using the same regression coefficients as for the empirical data set: 0.6 for Chronic NOEC to chronic EC10eq conversions and 2 for Acute E(L)C50 to chronic EC10eq. and values are subsequently calculated according to the methodology used for the HESTIA empirical toxicological effect data set. Here species-specific averages are applied across “algae”, “D.magna”, and “fish”, which means that the number of species data to construct SSD models is not sufficient (number of species <5), therefore no SSDs are constructed, but HC20EC10eq-values are calculated using eq.1-7. One “high-quality” data set is created from a subset of the QSAR estimations where only records with “EXPERIMENTAL”, “GOOD”, and “MODERATE” quality are selected as a complimentary analysis. Quality of all available predictions are measured as a “QSAR-Empirical match ratio” between the calculated HC20-values based on QSAR estimations () and the HC20-values based on empirical records () as .

### Exploring weighted means of averages and uncertainty through nonlinear least square fit modeling

Based on the assumption of lognormal distribution of ecotoxicological effect data, we investigate the uncertainties of the value by fitting all records per substance to a nonlinear least squares model to explore and at the 20% response level on the SSD curve. First, we define the raw data. For a specific substance , we have data for species , done in a number of experiments, labeled as . The data are thus indicated as , Throughout the analysis, we will work with the logarithm of the data, and for conciseness the data are indicated as , thus:

The per-species average over all samples are found as:

and the corresponding standard deviation as:

The average values have a standard error that is given by:

Next, we make the assumption that species sensitivity follows a lognormal distribution. This is in agreement with a mainstream practice of ecologists (ref Posthuma et al?). Because we have transformed the data with a logarithm, we now have a normal distribution for the -values. This distribution is characterized by two parameters, the mean () and the standard deviation (), which are traditionally estimated as follows:

Also, the is found as the 20-percentile value of the distribution:

where is the inverse of the standard normal distribution, which approximates -0.84.  
The uncertainty of this estimate basically depends on two elements: the uncertainty in and the uncertainty in . Both of these depend on the degree of fit with the normal distribution, and with the intra-species variation. Our approach will be to fit a normal distribution to the vector of mean values , weighted with the reciprocal of the variance, . Suppose we fit a function to the cumulative distribution of -values, each of which is associated with a standard error . We can find the optimal values of and through a least-squares minimization of the residual:

Here, denotes the order of appearance in the cumulative form. More precisely,

For F, we take the cumulative normal distribution, given by

To model estimates for and , we fit a nonlinear least-squares model (nls) to the toxicological dose-response data using R-programming language. By assuming a cumulative normal distribution for the data (Eq. 10) and defining a self-starting function for the cumulative normal distribution to start solving for the least squares residual of and , programming in R allow the use of the nls()function for this purpose. Starting points the model are defined as and respectively. In cases where only one toxicological record per species is available, is not obtainable, thus an arbitrary value of is used. If there are sufficient records for each chemical, the function models the (Eq. 5) and (Eq. 6) for each chemical, based of the and per species. We then calculate the using weighted means for each chemical (Eq. 7) as well as define the value at the 0.2 quantile of the model results using the qnorm()-function. Additionally, by extracting and from the model results, we fit each data point to a normal distribution and using Monte Carlo simulations with 100 000 iterations to populate vectors with simulated and at the 20% percentile level respectively. These samples are then used calculate a vector of 100 000 point values of the 0.2 quantile, from which we select the 2.5 and the 97.5 quantiles to reveal the 95% probability distribution within the 0.2 quantile for each chemical. Then, to enable a comparison of uncertainties across multiple chemicals, the geometric standard deviation for each probability distributions is calculated using the EnvStats::GeoSD function from the back-log-transformed vector of values in the 0.2 quantile.

The USEtox 2.1 manual requires data from three taxonomic groups to be present for ensuring some ecological relevance when calculating effect factors, while Owsianiak et al., (2022) makes a point to only include chemicals with effect data from at least five species for statistical reasons when deriving SSD’s, to ensure that at least one data point is placed below the -value. Hence we delimit our analysis to the proposed data inclusion of species from taxonomic groups.

### Data availability in the HESTIA Ecotox Explorer:

The script detailed above can generate a summary of the model results, creates a plot of the dose-response curve for each chemical, histogram over the Monte Carlo run’s data distribution, and a table containing the underlying dataset which is available for download.

### Software used

OECD QSAR Toolbox v.4.5 was used for collecting ecotoxicological data and physicochemical data, VEGA QSAR software v.1.2.0 was used to generated QSAR estimates for toxicity.R version 4.1.2 (2021-11-01) has been used for data collection, curation and visualization, text and data was published using Pandoc v.2.19.2 and Shiny R package version 1.7.4.9002 was used to construct the interactive web-based uncertainty explorer (Chang et al. 2023).

## Results

The curated data set contain toxicological data with 118770 records across 1999 species (Table 2), adapted for freshwater aquatic ecotoxicological potential (e.g. ecotoxicological effect factor) calculations at HC20EC10eq (hazard concentration where 20% effect is expected at the EC10eq concentration) for a set of 3698 chemicals. An overview of the data set curation and records kept or removed throughout the operations is given in Figure 1.

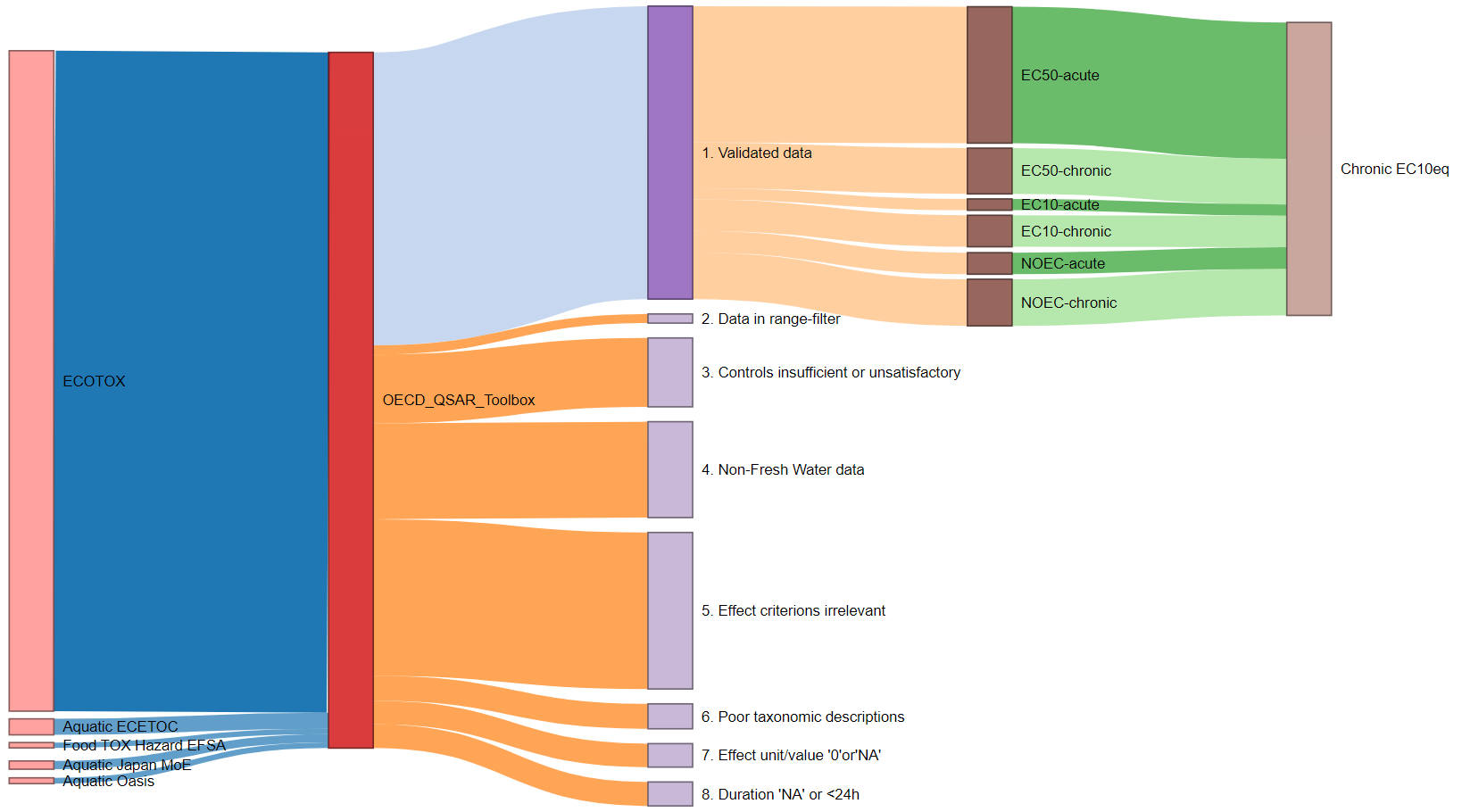


Figure 1: Overview of the data curation process from respective number of records per database, omitted records per curation step, and the counts of acute or chronic records per endpoint.

Table 2: Number of records found in the HESTIA Environmental Toxicology data set by taxonomic group and endpoint

| Taxonomy group | EC10 chronic | EC50 chronic | NOEC chronic | EC10 acute | EC50 acute | NOEC acute | Total per taxa |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Algae | 2,404 | 6,571 | 2,896 |  |  |  | 11,871 |
| Amphibian | 574 | 325 | 1,400 | 524 | 2,136 | 560 | 5,519 |
| Annellidae | 41 | 32 | 84 | 74 | 676 | 29 | 936 |
| Crustacean | 3,066 | 4,440 | 4,737 | 752 | 12,154 | 2,047 | 27,196 |
| Fish | 3,831 | 2,102 | 5,888 | 2,537 | 31,009 | 5,250 | 50,617 |
| Insect | 293 | 387 | 743 | 235 | 6,278 | 324 | 8,260 |
| Mollusca | 333 | 239 | 654 | 198 | 2,562 | 279 | 4,265 |
| Others | 598 | 2,143 | 823 | 53 | 231 | 83 | 3,931 |
| Plant | 1,324 | 1,772 | 1,300 | 212 | 233 | 147 | 4,988 |
| Rotifera | 288 | 527 | 372 |  |  |  | 1,187 |
| Total | 12,752 | 18,538 | 18,897 | 4,585 | 55,279 | 8,719 | 118,770 |

As a result of selecting records unique to the Envirotox database and curating the data as previously described, 12294 records are added to the final data set, summarized in Table 3.

Table 3: Number of records from EnviroTox data set joined to HESTIA by taxonomic group and endpoint

| Taxonomy group | EC10 chronic | EC50 acute | EC50 chronic | NOEC acute | NOEC chronic | Total per taxa |
| --- | --- | --- | --- | --- | --- | --- |
| Algae | 161 | 773 | 1 | 2 | 440 | 1,377 |
| Amphibian |  | 134 |  |  | 3 | 137 |
| Annellidae |  | 47 |  |  | 1 | 48 |
| Crustacean | 51 | 2,723 |  |  | 591 | 3,365 |
| Fish | 10 | 5,560 | 6 | 23 | 802 | 6,401 |
| Insect | 8 | 203 |  |  | 7 | 218 |
| Mollusca |  | 161 | 1 |  | 9 | 171 |
| Others | 4 | 255 |  |  | 66 | 325 |
| Rotifera |  | 208 |  |  | 44 | 252 |
| Total | 234 | 10,064 | 8 | 25 | 1,963 | 12,294 |

The HESTIA database construction was able to gather physicochemical properties required for freshwater aquatic toxicity potential characterization in USEtox for 13680 chemicals. Additionally, toxicological records for the defined substance groups Antibiotic, Antiviral, Other inorganic chemicals, Other organic chemicals, PPCP, Pesticide, Pharmaceutical, and Unknown with 1232, 1, 4269, 19897, 2414, 98914, 4101, and 236 records, respectively across 83, 1, 198, 2501, 72, 1831, 215, and 30 substances, respectively (Table 4). This data set allows for calculations of -values for 2976 chemicals along with the corresponding concentration-response slope factors according to Eq. 1-7, i.e the methodology in Owsianiak et al. (2022).

Table 4: Number of toxicological use annotations identified for chemicals

| Substance use category | Number of records | Number of chemicals |
| --- | --- | --- |
| Antibiotic | 1,232 | 83 |
| Antiviral | 1 | 1 |
| Other inorganic chemicals | 4,269 | 198 |
| Other organic chemicals | 19,897 | 2,501 |
| PPCP | 2,414 | 72 |
| Pesticide | 98,914 | 1,831 |
| Pharmaceutical | 4,101 | 215 |
| Unknown | 236 | 30 |
| Total | 131,064 | 4,931 |

For the final database, the HESTIA toxicological data set with EnviroTox unique records attached, we calculate 2976 -values out of 4931 chemicals. Moreover, we calculate HC50-values (denoted in B. Fantke P. (2017)) according to the former effect factor standard in USEtox 2.1 for 4690 chemicals.

When taking the weighted means approach and fit a nonlinear least-squares model to the database, 1132 chemicals have enough data to calculate . Data availability is a major issue here, since only 23% of all chemicals have enough data to fit the nonlinear least squares model (Table 5).

Table 5: Summary overview of the nonlinear least square model fit for the HESTIA ecotoxicological database

| Output status | Number of records | Minimum number of species per chemical | Maximum number of species per chemical | Minimum number of taxonomic groups per chemical | Maximum number of taxonomic groups per chemical |
| --- | --- | --- | --- | --- | --- |
| Convergence | 1,132 | 5 | 427 | 3 | 10 |
| not enough data | 3,763 | 1 | 17 | 1 | 4 |
| singular gradient matrix at initial parameter estimates | 15 | 5 | 15 | 3 | 4 |
| warning: step factor reduced below 'minFactor' | 21 | 5 | 12 | 3 | 5 |

The geometric standard deviation for the probability distribution at the -point are decreasing as the number of effect data increases (Figure 2). The highest geometric standard deviation belong to chemicals classified as pesticides (Figure 2*a*) with a range up to , with other groups of chemicals having far narrower probability distributions. Interestingly, at around 30-50 records per chemical, the geometric standard deviation appear to be the highest for most groups of chemicals. *The geometric standard deviation of the 40 of the top 50 chemicals with highest geometric standard deviation of the probability distribution at the -point belongs to pesticides.*

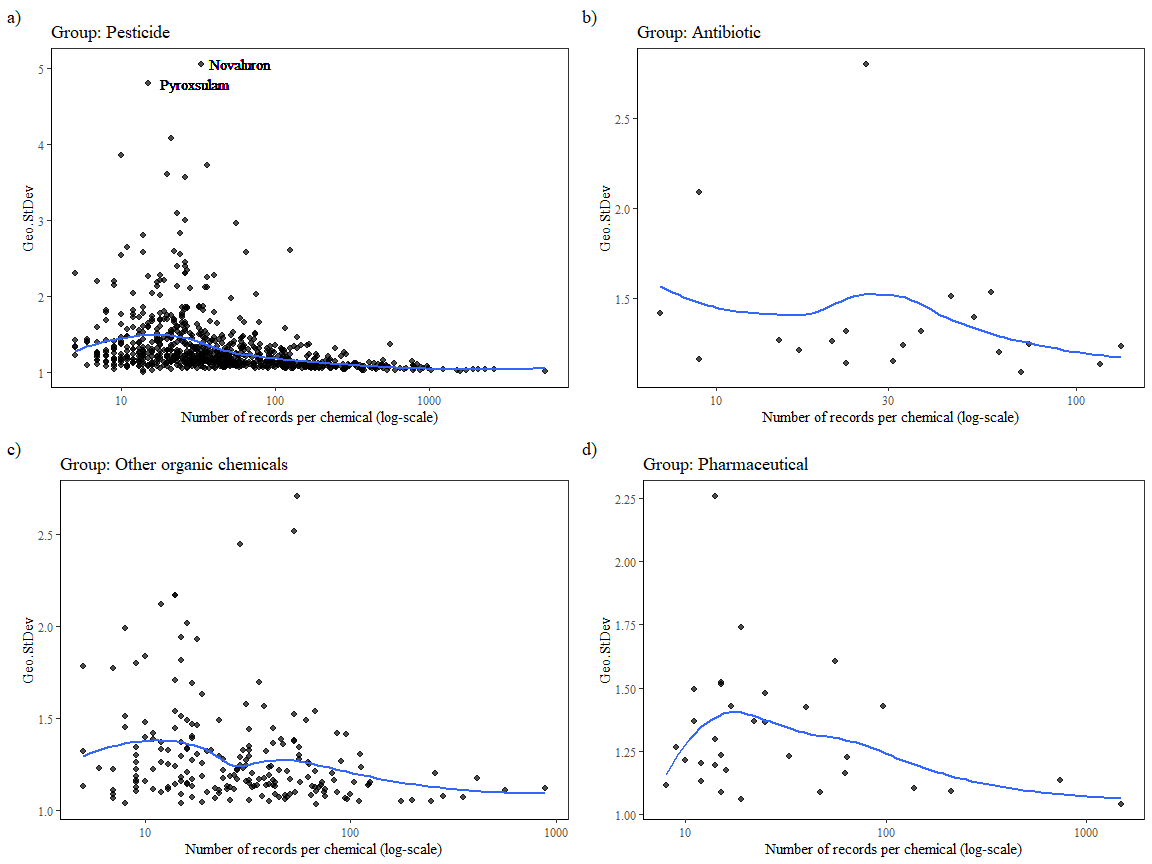


Figure 2: Geometric standard deviation of probability distribution ranges based on Monte Carlo simulations per number of effect data records per chemical. Curve fitted using loess smoothing. Geo.StDev = Geometric standard deviation.

### Comparison of ecotoxicological effects based on species sensitivity distributions (SSD) to the weighted means distribution

By fitting a linear model to the two data sets with values based on weighted means generated by the NLS model and the standard values calculated using the methodology by Owsianiak et al. (2022), we can establish a very good fit for the two datasets with close to 94 % of the variability of weighted means data is explained by the standard SSD calculations, Figure 3 **a)**. We plot the absolute differences between the two methods in a histogram to visualize the differences (**More explanation if we deem this comparison to be relevant**), referred to as Figure 3 **b)**.

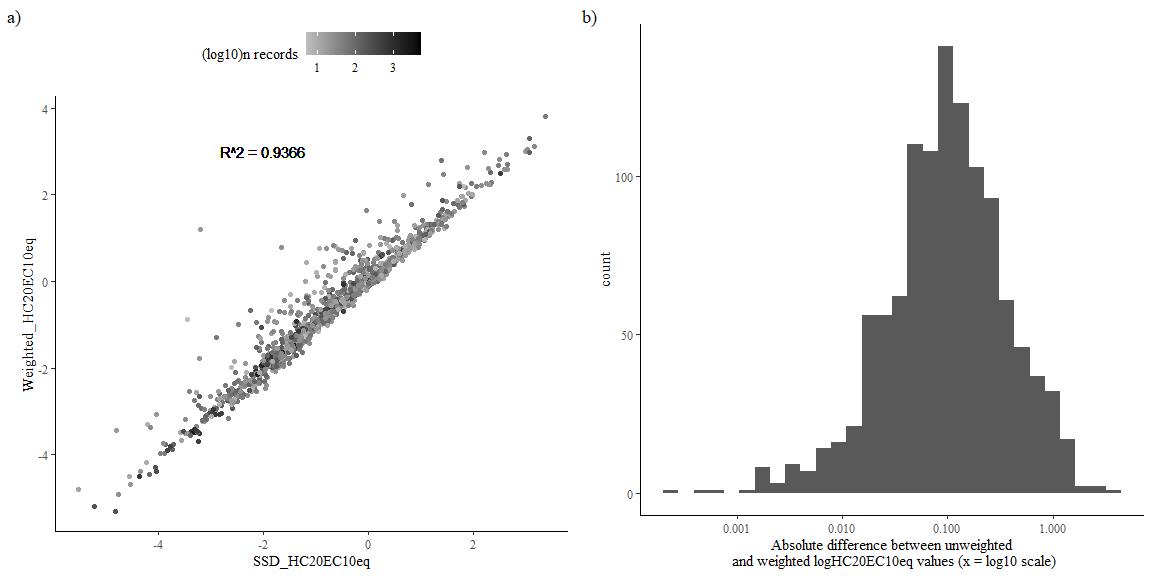


Figure 3: Test caption and explanation

### Investigating the applicability of QSAR models to estimate toxicological effect data

From the 15030 chemicals that could be matched to a SMILES annotation, which means mono-constituent chemicals, we see a generally low quality scoring for most chemicals, with KNN acute LC50 fish toxicity model producing the most good quality matches contrasted by the NIC acute LC50 fish toxicity model which produced no higher quality matches, outside of the chemicals within the model’s training set. A summarized overview of the QSAR estimations quality report in shown in Table 6.

Table 6: Summary of the QSAR models applied to the HESTIA toxicological data set, with counts of model quality per QSAR model. 'ERROR'-column implies either ERROR-reported quality score, estimates are missing, or estimate = 0 mg/L

| QSAR model | EXPERIMENTAL | GOOD | MODERATE | LOW | ERROR |
| --- | --- | --- | --- | --- | --- |
| Fish Acute KNN | 952 | 1,583 | 2,855 | 4,870 | 4,770 |
| D.magna Acute IRFMN | 398 | 1,103 | 2,979 | 7,504 | 3,046 |
| Fathead Minnow Acute EPA | 940 | 937 | 3,238 | 6,744 | 3,171 |
| D.magna Chronic IRFMN | 296 | 902 | 2,276 | 8,510 | 3,046 |
| Fish Acute IRFMN | 293 | 767 | 2,737 | 8,241 | 2,992 |
| Fathead Minnow Acute KNN.IRFMN | 652 | 652 | 2,225 | 5,093 | 6,408 |
| Algae Acute ProtoQSAR.Combase | 306 | 598 | 2,836 | 8,119 | 3,171 |
| Guppy Acute KNN.IRFMN | 229 | 530 | 385 | 949 | 12,937 |
| Algae Chronic IRFMN | 374 | 407 | 1,977 | 9,226 | 3,046 |
| Algae Acute IRFMN | 290 | 391 | 1,815 | 9,486 | 3,048 |
| D.magna Acute EPA | 442 | 328 | 2,173 | 8,916 | 3,171 |
| Fish Acute IRFMN.Combase | 150 | 58 | 1,473 | 10,357 | 2,992 |
| Fish Chronic IRFMN | 113 | 53 | 805 | 11,067 | 2,992 |
| D.magna Acute IRFMN.Combase | 170 | 3 | 976 | 10,889 | 2,992 |
| Fish Acute NIC | 943 |  |  | 10,920 | 3,167 |

Only a small subset of chemicals are eligible to include for calculations as a “high-quality” subset with 412 records where HC20-values are based on only “GOOD” and “MODERATE” quality are selected for analysis. Records marked as “EXPERIMENTAL” were removed to not compare data originating from the same source as the experimental records. When fitting a linear model to the calculated -values based on QSAR estimations () and -values based on empirical records () we see a strong relationship between the two data sets, but only 0.42 % of the variability of is explained by (Figure 4 **a)**). Worth noting is the severe under-estimation in toxicity for the substances deemed most toxic from empirical data, highlighted in the red circle in Figure 4 **a)**, where toxicity is underestimated by several orders of magnitude in the QSAR estimations. Figure 4 **b)** shows the distribution of relative difference between QSAR estimations to empirical data, presented on a log scale where the quality of all available predictions are measured as a “QSAR to empirical data ratio” () between and the as

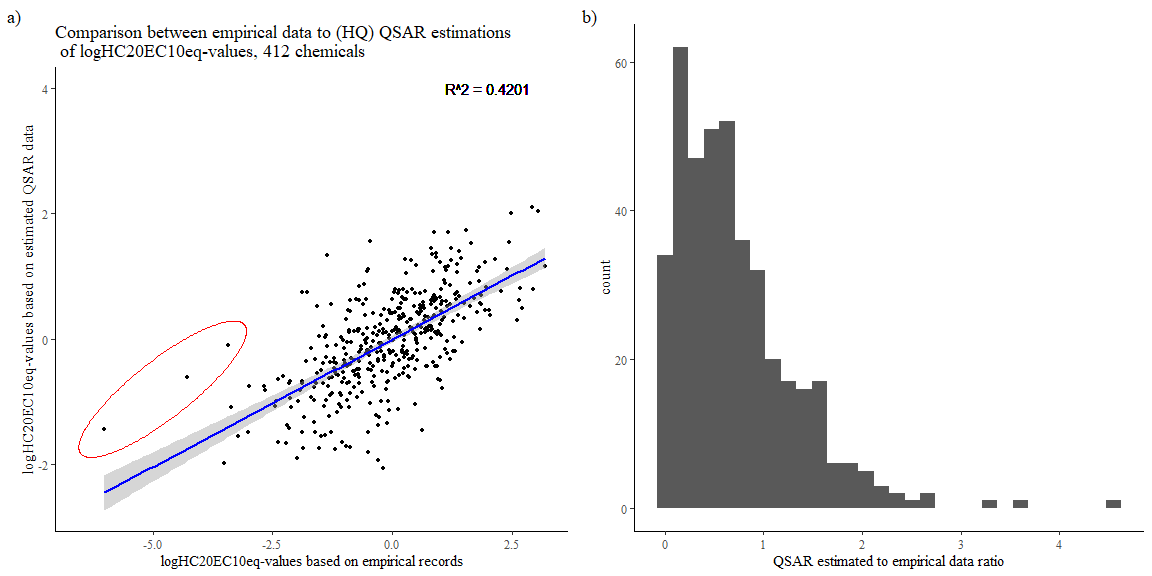


Figure 4: (TEST CAPTION)

### Discussion

*Database:*  
By gathering physicochemical and ecotoxicological effect data for a set of 16797 chemicals through querying OECD QSAR Toolbox software, we have been able to calculate concentration-response slope factors at the 20% response level () for 2976 chemicals, assess the uncertainty at the point of the SSD curve for 1132 chemicals and gather 13920 physicochemical records required for characterization of chemicals with USEtox 2.1 methodology. SSD curves with uncertainty intervals at the point can be created for 1132 of the chemicals within this database, while data for 2976 chemicals have been calculated using the methodology of Owsianiak et al. (2022). The present data set is not a fully comprehensive database for all potentially harmful substances from open sources, 16797 substances is only a small part of the full spectrum of man-made substances released into the environment and within this subset of data toxicological effect data are available for 30% of substances queried. More substances need to be tested, across a more diverse set of species than for the current characterization methodology to be able to include a larger set of potentially harmful substances.

*Uncertainties:*  
The present database contains data at both acute and chronic effect data from EC50, EC10 and NOEC endpoints. With regression coefficients available from Aurisano et al. (2019), we deemed it fruitful to include acute effect data, albeit large confidence interval ranges for some of these coefficients (regression coefficient 7.44 for Acute EC50 to Chronic EC10eq for fish, with a 95% confidence interval 2.92-18.95). The probability distributions within each data point will have marginal influence on the final -value considering the ranges of effect across all species per chemical (REF???-have not checked for this). Aurisano et al. (2023) quantify the uncertainty “POD/RDs: human impacts” for 10145 chemicals by calculating the squared geometric standard deviation of interstudy variability and intrastudy variability prior to fitting a lognormal distribution of data and bootstrapping, to generate comparable uncertainty distributions across the entire dataset.  
The present approach fits the cumulative normal distribution to the vector of , allowing to account for the variance across all species data, while the methodology recommended by Owsianiak et al., (2019) use the arithmetic mean of vectors and respectively are used to calculate a -value. The current approach is crucial for a fair estimate of probability distributions at the -value, as data availability varies greatly across species, with few species being over represented in the dataset (e.g *Daphnia magna, Pimephales promelas, and Oncorhynchus mykiss* corresponding to 28 % of the dataset).  
A problem for probabilistic uncertainty assessment of pesticides, due to the specific mode of action, are the large range of effect between target and non-target organisms (Warren-Hicks and Hart 2010). So with a distribution of effect data at two extremes, we can assume that pesticides will have large uncertainties in the calculated effect factor, which, indeed, is the case for the insecticide “Novaluron” and herbicide “Pyroxsulam”, where the geometric standard deviation is very high (5.1 and 4.8 respectively) (figure 2). This brings up the issue of how to represent pesticide ecotoxicity in LCIA modelling, should results be weighted towards the lowest quantile, or include regression coefficients for pesticide class chemicals?  
- Should i continue on this thought? *(cite P. Fantke et al. (2018): “The choice of a lower percentile than the median will also reduce the discrepancy with contemporary approaches in chemical risk assessment that ask for the use of several SSD models in the case of chemicals with a specific mode of action (most pesticides). There are numerically large differences at the level of the median value (HC50EC50), but expectedly lower numerical consequences in the tails between the nonsplit and split SSD approaches (e.g., Zajdlik et al. 2009).”)*

*Quantitative structure-activity relationships:*  
Complementing toxicological testing of animals with *in silico* simulations to derive quantitative structure-activity relationship estimates for chemical toxicity is a desirable approach for improving our understanding of the full range of impacts caused by anthropogenic activities. This as toxicological tests of chemicals are time consuming and expensive, while new chemicals are constantly introduced. Moreover, ethical aspects of exposing animals to potentially harmful chemicals also need to be considered. These QSAR models, however, are only as good as the training data and results from the current work show that when applying fifteen different QSAR models to the HESTIA inventory chemicals, many are too dissimilar to the models’ training data sets and either fail to generate estimations or are deemed as low quality estimations.

Even more worrisome is that estimations of experimental, good and moderate quality are still not providing HC20-values within a reasonable error margin. When comparing QSAR HC20-values with HC20-values based on empirical data across 412 chemicals, the estimated HC20-values deviate substantially from the corresponding HC20-values derived from empirical data with errors up to 2.02e+05 at the largest extreme, while the median difference is 0.23 . data and data for additional chemicals are available (e.g in the REACH dossier) and gathering of data for *any other* chemicals using OECD QSAR Toolbox could be a straightforward task, with the here available R-scripts adapted for reading and wrangling data structured as OECD QSAR Toolbox-output.

*Considerations of taxonomy:*  
Averaging across species can be problematic for records where taxonomic details are given at levels above species. Eq. 5 calculates the mean of all per-species EC10eq effect values () and whether using the weighted means-method or the standard non-weighted method to calculate , records defined at a higher taxonomic rank than species will influence the calculation by representing an artificial species, for example *Daphnia sp.* will count as one species alongside *D. magna*. Such above-species-level taxonomic identifiers have been removed from this data set resulting in the exclusion of 803 names (not counting common names or erroneous names). This issue seem to be overlooked, as these taxonomic ids at higher level **are present in both Connors et al. (2019) and Sala et al. (2022)**. Since data scarcity is an issue this poses an avenue for future analysis when testing how robust a methodology of averaging records across genus-level instead of species-level. Such an approach could produce more data-rich averaging calculations, but at the cost of fewer data points to construct an SSD-curve from.

**Conclusions**  
- By gathering a large set of freely available ecotoxicological records, we have been able to calculate the 95% probability distribution for the -value on SSD curves for 1132 chemicals, yet only fewer than 10% of the original data set with 16797 chemicals. - Some pesticides have notoriously high uncertainty at the -value compared to other types of substances. This can be expected, due to the nature of toxicological specificity of these chemicals.

### 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>.

Aurisano, Nicolò, Olivier Jolliet, Weihsueh A. Chiu, Richard Judson, Suji Jang, Aswani Unnikrishnan, Marissa B. Kosnik, and Peter Fantke. 2023. “Probabilistic Points of Departure and Reference Doses for Characterizing Human Noncancer and Developmental/Reproductive Effects for 10,145 Chemicals.” *Environmental Health Perspectives* 131 (3): 037016. <https://doi.org/10.1289/EHP11524>.

Benfenati, Emilio, Alberto Manganaro, and Giuseppina C. Gini. 2013. “VEGA-QSAR: AI Inside a Platform for Predictive Toxicology.” In *PAI@AI\*IA*.

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2023. *Shiny: Web Application Framework for r*. <https://shiny.rstudio.com/>.

Connors, Kristin A., Amy Beasley, Mace G. Barron, Scott E. Belanger, Mark Bonnell, Jessica L. Brill, Dick de Zwart, et al. 2019. “Creation of a Curated Aquatic Toxicology Database: EnviroTox.” *Environmental Toxicology and Chemistry* 38 (5): 1062–73. <https://doi.org/10.1002/etc.4382>.

Der Werf, Hayo MG van, Marie Trydeman Knudsen, and Christel Cederberg. 2020. “Towards Better Representation of Organic Agriculture in Life Cycle Assessment.” *Nature Sustainability* 3 (6): 419–25.

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>.

Fantke, Peter, Nicoló Aurisano, Jane Bare, Thomas Backhaus, Cécile Bulle, Peter M. Chapman, Dick De Zwart, et al. 2018. “Toward Harmonizing Ecotoxicity Characterization in Life Cycle Impact Assessment.” *Environmental Toxicology and Chemistry* 37 (12): 2955–71. <https://doi.org/10.1002/etc.4261>.

Floris, Matteo, Alberto Manganaro, Orazio Nicolotti, Ricardo Medda, Giuseppe Felice Mangiatordi, and Emilio Benfenati. 2014. “A Generalizable Definition of Chemical Similarity for Read-Across.” *Journal of Cheminformatics* 6 (1): 1–7.

Hauschild, Michael Z, and Mark AJ Huijbregts. 2015. *Introducing Life Cycle Impact Assessment*. Springer.

Henriksson, Patrik JG, Andreu Rico, Wenbo Zhang, Sk Ahmad-Al-Nahid, Richard Newton, Lam T Phan, Zongfeng Zhang, et al. 2015. “Comparison of Asian Aquaculture Products by Use of Statistically Supported Life Cycle Assessment.” *Environmental Science & Technology* 49 (24): 14176–83.

Huijbregts, Mark AJ, U Thissen, JB Guinée, T Jager, D Kalf, D Van de Meent, AMJ Ragas, A Wegener Sleeswijk, and L Reijnders. 2000. “Priority Assessment of Toxic Substances in Life Cycle Assessment. Part i: Calculation of Toxicity Potentials for 181 Substances with the Nested Multi-Media Fate, Exposure and Effects Model USES–LCA.” *Chemosphere* 41 (4): 541–73.

Krewski, Daniel, Daniel Acosta Jr, Melvin Andersen, Henry Anderson, John C Bailar III, Kim Boekelheide, Robert Brent, et al. 2010. “Toxicity Testing in the 21st Century: A Vision and a Strategy.” *Journal of Toxicology and Environmental Health, Part B* 13 (2-4): 51–138.

Owsianiak, Mikolaj, Peter Fantke, Leo Posthuma, Erwan Saouter, Martina Vijver, Thomas Backhaus, Christian Schlekat, and Michael Z. Hauschild. 2019. “Ecotoxicity.” In *Global Guidance on Environmental Life Cycle Impact Assessment*, edited by Rolf Frischknecht and Olivier Jolliet, 138–72. Paris, France: UNEP/SETAC Life Cycle Initiative.

Owsianiak, Mikołaj, Michael Z. Hauschild, Leo Posthuma, Erwan Saouter, Martina G. Vijver, Thomas Backhaus, Mélanie Douziech, Tamar Schlekat, and Peter Fantke. 2022. “Ecotoxicity Characterization of Chemicals: Global Recommendations and Implementation in USEtox.” *Chemosphere*, no. 310 (October). <https://doi.org/10.1016/j.chemosphere.2022.136807>.

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.

Pennington, David W, Manuele Margni, Christoph Ammann, and Olivier Jolliet. 2005. “Multimedia Fate and Human Intake Modeling: Spatial Versus Nonspatial Insights for Chemical Emissions in Western Europe.” *Environmental Science & Technology* 39 (4): 1119–28.

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.

Sala, Serenella, Fabrizio Biganzoli, Esther Sanye Mengual, and Erwan Saouter. 2022. “Toxicity Impacts in the Environmental Footprint Method: Calculation Principles.” *The International Journal of Life Cycle Assessment* 27 (4): 587–602. <https://doi.org/10.1007/s11367-022-02033-0>.

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.

Saouter, Erwan, Fabrizio Biganzoli, Rana Pant, Serenella Sala, and Donald Versteeg. 2019. “Using REACH for the EU Environmental Footprint: Building a Usable Ecotoxicity Database, Part I.” *Integrated Environmental Assessment and Management* 15 (5): 783–95. <https://doi.org/10.1002/ieam.4168>.

Schreinemachers, Pepijn, and Prasnee Tipraqsa. 2012. “Agricultural Pesticides and Land Use Intensification in High, Middle and Low Income Countries.” *Food Policy* 37 (6): 616–26.

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>.

Van Zelm, Rosalie, Mark AJ Huijbregts, and Dik van De Meent. 2009. “USES-LCA 2.0—a Global Nested Multi-Media Fate, Exposure, and Effects Model.” Springer.

Warren-Hicks, William J, and Andy Hart. 2010. *Application of Uncertainty Analysis to Ecological Risks of Pesticides*. CRC Press.