Semi-automated harmonization and selection of chemical data for risk and impact assessment

Aurisano, Nicolò; Fantke, Peter

Published in:
Chemosphere

Link to article, DOI:
10.1016/j.chemosphere.2022.134886

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Chemosphere 302 (2022) 134886

Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

Semi-automated harmonization and selection of chemical data for risk and impact assessment

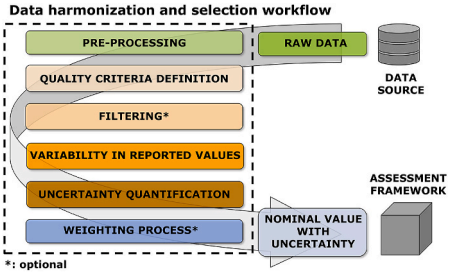
Nicolò Aurisano, Peter Fantke *

Quantitative Sustainability Assessment, Department of Environmental and Resource Engineering, Technical University of Denmark, Produktionstorvet 424, 2800, Kgs. Lyngby, Denmark

HIGHLIGHTS

- Level of required data completeness and quality varies across application domains.
- Method able to assess the quality of chemical data based on a set of criteria.
- We derive a unique nominal value with its uncertainty from a set of data points.
- Intrinsic variability and uncertainty may differ across chemicals data sources.
- Standardized methods are needed to systematically curate data across properties.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Klaus Kümmerer

Keywords:
Data quality
Uncertainty assessment
Chemical properties
REACH
Partition coefficient

ABSTRACT

Chemical data for thousands of substances are available for safety, risk, life cycle and substitution assessments, as submitted for example under the European Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) Regulation. However, to widely disseminate reported physicochemical properties as well as human and ecological exposure and toxicological data for use in various science and policy fields, systematic methods for data harmonization and selection are necessary. In response to this need, we developed a semi-automated method for deriving appropriate substance property values as input for various assessment frameworks with different requirements for resolution and data quality. Starting with data reported for a given substance and property, we propose a set of aligned data selection and harmonization criteria to obtain a representative mean value and related confidence intervals per chemical-property combination. The proposed method was tested on a set of octanol-water partition coefficients (K_{ow}) for an illustrative set of 20 substances, reported under the REACH regulation as example data source. Our method is generally applicable to any set of substances, and can assess specific distributions in quality and variability across reported data. Further research can likely extend our method for mining information from text fields and adapt it to available data reported or collected from other sources and other substance properties to improve the reliability of input data for risk and impact assessments.

1. Introduction

Over the past two decades, the need for high-quality data supporting reliable input for chemicals management and sustainability assessment frameworks of various complexity and spatiotemporal granularity has increased (Persson et al., 2022; Schenker et al., 2005). At the same time,

* Corresponding author.
E-mail address: pefan@dtu.dk (P. Fantke).

https://doi.org/10.1016/j.chemosphere.2022.134886
Received 14 October 2021; Received in revised form 3 May 2022; Accepted 5 May 2022
Available online 7 May 2022

new data sources for physicochemical substance properties, degradation, human and ecological exposure and toxicological effects (hereafter referred to as *chemical property information*) are increasingly becoming available, covering tens of thousands of chemicals (e.g., Bolton et al., 2008; Dorne et al., 2017; Pence and Williams, 2010; Sobanska and Le Goff, 2014; Williams et al., 2017). Nevertheless, for using such information in decision support frameworks, it is essential to address current concerns related to reliability and quality of the various reported experimental data and their application without meaningful prior data interpretation and curation (Fantke et al., 2018, 2021b; Igos et al., 2014; Mansouri et al., 2016; Müller et al., 2017).

For example, under the European Registration, Evaluation, Authorisation, and Restriction of Chemicals (REACH) Regulation (EC) 1907/2006, chemical property information has been collected for more than 20,000 substances in the International Uniform Chemical Information Database (IUCLID) (Fantke et al., 2020). The potential applicability of IUCLID data in different science and policy fields (e.g., chemical substitution, life cycle impact assessment, health impact assessment, chemical prioritization, high-throughput risk screening, exposure assessment) has been widely discussed (Askham, 2012; Fantke et al., 2018; Igos et al., 2014; Luechtefeld et al., 2016; Müller et al., 2017; Saouter et al., 2017a, 2017b). Findings from these discussions suggest that, in practice, it remains unclear how to select from any data source the appropriate values for a given chemical property information (Luechtefeld et al., 2016; Przybylak et al., 2012). To choose appropriate values for a given assessment or decision context, it is considered essential to identify the level of required data representativeness and quality, which both differ as function of the scope and resolution of the performed assessments (e.g., site-specific versus global supply chain assessment, or single-chemical risk assessment versus prioritization across thousands of chemicals).

More than one value can be reported for a given chemical property information in the different available chemical databases since experimental results in a database are usually directly reported as they were gathered from diverse sources. Thereby, test results are often obtained under specific test conditions (e.g., pH, temperature) and methods. In some databases, also values estimated from quantitative structure-activity relationships (QSAR) or read-across methods are reported in addition to experimental data, with new-approach methods (NAM) and machine-learning methods as emerging extrapolation approaches (e.g., Hou et al., 2020; Wambaugh et al., 2019). As a result, the available (experimental or estimated) data for a given chemical property information of a specific substance often vary in numerical values, data quality, and information completeness. Then, how do we select the appropriate result from all the available data for a given substance property and chemical? Unquestionably, standardized methods are urgently needed for systematically harmonizing and selecting the various available data to arrive at representative input values for the different application fields and models, including consideration of the uncertainty around these values for a reliable interpretation (Fantke et al., 2020; Li et al., 2003).

Yet few studies offer methods for systematically selecting or harmonizing data from publicly available databases. For example, Saouter et al. (2019a, 2019b) describe how to choose values for multiple chemical properties from data in IUCLID, the OpenFoodTox database (Barbaro et al., 2015), and the Pesticide Properties Database (PPDB) (Lewis et al., 2016). Mansouri et al. (2016) propose methods for finding and correcting errors in chemical identifier representations. Beyer et al. (2002) present an approach for selecting values of chemical properties subject to thermodynamic constraints, subsequently updated to minimize the adjustment required for thermodynamic consistency (Li et al., 2003; Schenker et al., 2005; Xiao et al., 2004). Other studies propose methods for specific aspects, such as deriving hazard properties for selected substances (e.g., Stieger et al., 2014), curating chemicals and biological data (e.g., Fourches et al., 2016), or obtaining freshwater ecotoxicological effect endpoints from measured data reported in IUCLID (Aurisano et al., 2019). However, these methods commonly use fixed filters to screen out data based on predefined criteria thresholds (e.g., for data reliability), which often do not exploit the reported data's full potential for the various assessment frameworks with their different input data requirements. Furthermore, existing methods do not consider the intrinsic variability and uncertainty across reported data, which may differ widely across chemicals and data sources, and which is important for correctly interpreting any data harmonization and selection results (Posthuma et al., 2019).

To address this gap, we propose a systematic and flexible data harmonization and selection method for deriving suitable substance property values as input for various risk and impact assessment frameworks with different resolution and data quality requirements. Our goal is to exploit reported chemical data's potential fully and to allow an appropriate interpretation of the results by characterizing related uncertainties. To achieve this goal, we focus on three specific objectives: (i) to define a set of flexible criteria for consistently evaluating reliability, quality, completeness, variability, and uncertainty of reported data for a given substance property; (ii) to develop a data harmonization and selection workflow based on the context-specific set of criteria to derive an appropriate value, along with a confidence interval, from the available data for a substance property; and (iii) to apply the data harmonization and selection workflow in a case study to derive octanol-water partition coefficient (K_{ow}) values for a set of 20 test chemicals registered under REACH and covering a wide physicochemical property space, for application in high-throughput risk screening, Life Cycle Impact Assessment (LCIA) and chemical substitution.

2. Materials and methods

2.1. Quality criteria definition

Since different data application domains require different minimum quality levels, flexible approaches are needed for the data harmonization and selection process to ensure an appropriate application of the results in the given context (Przybylak et al., 2012). For example, a strict selection of only high-quality data is required in a regulatory safety assessment context, disregarding low-quality information. Likewise, only high-quality data are considered when developing extrapolations for substances without available information, i.e., predictive approaches (Aurisano et al., 2019; Cronin and Schultz, 2003; Posthuma et al., 2019). A more inclusive approach (i.e., high data coverage but reduced average data quality) is suitable for screening level prioritization or substitution of chemicals across thousands of substances or for characterizing hundreds of chemicals associated with a given product life cycle (Aurisano et al., 2021a, 2021b, 2022; Fantke et al., 2020, 2021a; Tickner et al., 2019). To match the available data to the different application requirements, we propose to assess the quality and completeness of each reported experimental or estimated result (hereafter defined as *data point*) against quality criteria.

For quality criteria, we propose using the categorization in the original databases as a starting point, where available. For example, in REACH, quality criteria applicable to a wide range of substance properties include data reliability (Klimisch score) and Type of Information (adequacy) outlining a data point's origin (e.g., experimental study, read-across) (European Commission, 2006; Sobanska et al., 2014; Tarazona et al., 2014).

Each reported data point usually comes with information regarding different data quality criteria, $q_i, i \in \{1, \ldots, Q\}$, with for example q_1 = 'reliability', q_2 = 'type of information'. We define the different options for each quality criterion q_i as classes, $n_{i,j}$, $j \in \{1, \ldots, N\}$. For example, classes of the criterion q_1, according to the Klimisch score used in REACH, include $n_{1,1}$ = '1 (reliable without restriction)', $n_{1,2}$ = '2 (reliable with restrictions)', $n_{1,3}$ = '3 (not reliable)', $n_{1,4}$ = '4 (not assignable)', $n_{1,5}$ = 'Other', and $n_{1,6}$ = 'Missing data' (Klimisch et al., 1997).

Depending on the data quality requirements of a given assessment framework or purpose, some classes might be grouped for a single criterion, $n'_{i,j}$, $j \in \{1,\ldots,N'\}$. For example $n'_{1,1}$ = 'reliable' (i.e., $n'_{1,1} \in \{n_{1,1}, n_{1,2}\}$), and $n'_{1,2}$ = 'unreliable' (i.e., $n'_{1,2} \in \{n_{1,3}, n_{1,4}, n_{1,5}, n_{1,6}\}$). Finally, these grouped classes ($n'_{i,j}$) are combined across considered quality criteria into unique combinations of aggregate-criteria classes $c_k^{q,n'}$, $k \in \{1,\ldots,K\}$. For example $c_1^{q,n'}$ = 'reliable, experimental' represents data with a specific combination of grouped classes for each of the criteria 'reliability' (q_1) and 'type of information' (q_2). Once $c_k^{q,n'}$ are defined, each available data point for a specific substance and chemical property can be allocated to the respective $c_k^{q,n'}$, reflecting its underlying data quality and information completeness. An illustrative example of the calculation of $c_k^{q,n'}$ is presented in SI (Table S1).

2.2. Variability in reported values

Across different data sources, chemical property information can be reported either as individual numerical values or as ranges (i.e., bounded or unbounded intervals). To harmonize such information, we define for each reported test result a nominal value (x) together with a Confidence Interval (CI). If for a given data point the test result is reported as a range with two bounds, we define x as the arithmetic mean of the range and consider the two reported values as CI^l (lower CI limit) and CI^u (upper CI limit). If reported as a single value, we consider the reported value as x, and multiply and divide by a variability factor (F_V) for defining its CI, S1 summarizes the defined rules for test results reported with relational operators (qualifiers). F_V is derived from the variability of all data in the entire considered dataset to introduce additional conservatism in the assessment when only a single numerical value is reported.

After defining for each data point x and CI, we derive for each combination of aggregate-criteria classes $c_k^{q,n'}$ a x_c and its CI_c based on the pool of data points allocated to the same combination, where x_c is calculated as the arithmetic mean across reported data points(x) in the same combination $c_k^{q,n'}$ and the related CI_c is estimated as the highest CI^u and the lowest CI^l across data points allocated to that combination. S2 presents an example of deriving x_c and its CI_c based on the pool of data points allocated to the same combination for an arbitrary substance.

2.3. Uncertainty quantification

We combine two types of uncertainty, expressed as squared geometric standard deviation (GSD^2), to characterize the uncertainty around x_c. For each combination of aggregate-criteria classes $c_k^{q,n'}$, base uncertainty ($GSD^2_{\text{base},c}$) reflects reported or otherwise default variability in x_c (Frischknecht et al., 2005), and criteria uncertainty ($GSD^2_{\text{criteria},c}$) reflects the quality and completeness of the available data points (Beyer et al., 2002).

We assume approximately log-normally distributed values for strictly non-negative physicochemical properties, such as K_{ow} (MacLeod et al., 2002; Schenker et al., 2009). For some properties, other distributions might apply (e.g., Wender et al., 2018). Uncertainty can be expressed for a parameter as the 95% CI_c range with the 2.5$^{\text{th}}$ and 97.5$^{\text{th}}$ percentiles obtained from the geometric mean of x_c, generalized as $x_c^* \in P$, and the related $GSD^2 \triangleq e^{2\times\sigma}$ with $\sigma \in P,\ \sigma > 0$ as standard deviation of the natural logarithm of x_c and the probability $\{x_c^*/GSD^2 < x_c < x_c^* \times GSD^2\} \approx 0.95$ (Fantke et al., 2012; Hong et al., 2010; Slob, 1994; Stylianou et al., 2021). A $GSD^2 = 2$, for instance, denotes that the distribution of 95% of all values fall within half and twice of x_c. Whenever the exact distribution is not reported, we assume a standard log-normal distribution, where $GSD^2_{\text{base},c}$ is derived from CI_c^l and CI_c^u around x_c for each combination of aggregate-criteria classes $c_k^{q,n'}$ as (Rosenbaum et al., 2018):

$$GSD^2_{\text{base},c} = \sqrt{CI_c^u / CI_c^l} \tag{1}$$

$GSD^2_{\text{criteria},c}$ is derived from applying the Pedigree matrix approach, which was first introduced using qualitative criteria, and further refined to derive quantitative uncertainty classes as a function of categories for base and parameter uncertainty (Ciroth et al., 2016; Muller et al., 2016; Weidema and Wesnæs, 1996). In the Pedigree matrix, we assign to each grouped class of a given criterion ($n'_{i,j}$) an uncertainty factor ($GSD^2_{\text{criteria},c_n}$), reflecting quality and completeness of its elements. $GSD^2_{\text{criteria},c}$ for the combination of distinct quality and completeness across aggregate-criteria classes is then quantified by combining the different $GSD^2_{\text{criteria},c_n}$:

$$GSD^2_{\text{criteria},c} = e^{\sqrt{\sum_{c_n}\left(\ln GSD^2_{\text{criteria},c_n}\right)^2}} \tag{2}$$

The combination of the base uncertainty and the criteria uncertainty yields the overall uncertainty (GSD_c^2) around x_c for each combination of aggregate-criteria classes (Ciroth et al., 2016; Frischknecht et al., 2005; Slob, 1994):

$$GSD_c^2 = e^{\sqrt{\left(\ln GSD^2_{\text{base},c}\right)^2 + \left(\ln GSD^2_{\text{criteria},c}\right)^2}} \tag{3}$$

By estimating both x_c and GSD_c^2, we can provide a nominal value with related uncertainty per $c_k^{q,n'}$ for a specific substance property.

2.4. Weighting processes

Since, in some cases, results from more than one combination of aggregate-criteria classes are suitable for a given application context, we developed a weighting process for deriving a unique nominal value ($\overline{x}_w$) and related GSD_w^2 across available combinations of aggregate-criteria classes. For estimating $\overline{x}_w$, we first assign to each combination of aggregate-criteria classes a quality weight ($w_{Q,c}$), derived from $GSD^2_{\text{criteria},c}$ and thus directly considering its earlier defined quality and completeness as weights:

$$w_{Q,c} = \frac{1}{GSD^2_{\text{criteria},c}} \tag{4}$$

The higher the $GSD^2_{\text{criteria},c}$ of a given combination of aggregate-criteria classes, the lower its $w_{Q,c}$ and, thus, the lower its influence on the final $\overline{x}_w$. Next, we rank $w_{Q,c}$ and derive combined quality/variability weights ($w_{V,c}$), scaled based on the variability distribution P of available x_c:

$$w_{V,c} = w_{Q,c} \times \frac{x_c^P}{x_{c,\max}} \tag{5}$$

Since $w_{V,c}$ is a function of both the number of combinations of aggregate-criteria classes and their respective quality, we normalized $w_{V,c}$ to determine the overall weight for each x_c:

$$w_{V,c}^{\text{norm}} = \frac{w_{V,c}}{\sum_c w_{V,c}} \tag{6}$$

Finally, $w_{V,c}^{\text{norm}}$ are combined with x_c to yield a $\overline{x}_w$ across combinations of aggregate-criteria-classes:

$$\overline{x}_w = \sum_c \left(x_c \times w_{V,c}^{\text{norm}}\right) \tag{7}$$

For estimating related GSD_w^2 of $\overline{x}_w$, we applied the sensitivity factors (S) as described by MacLeod et al. (2002). S is calculated for each available $c_k^{q,n'}$ as $(\Delta O/O)/(\Delta I/I)$, where O is the output and I is the input

variable of interest. With that, GSD_w^2 across available $c_k^{q,n'}$ is calculated as (MacLeod et al., 2002; Slob, 1994):

$$GSD_w^2 = e^{\sqrt{\sum_c S_c^2 \left(GSD_c^2\right)^2}} \quad (8)$$

GSD_w^2 describes the spread of data around their geometric mean, and more specifically indicates that 95% of the data fall within the weighted nominal value ($\bar{x}_w$) divided by GSD_w^2 and multiplied by GSD_w^2. For example, a $GSD_w^2 = 10$ indicates that the 95% *CI* of $\bar{x}_w$ span over two orders of magnitude.

Fig. 1 graphically summarizes the presented workflow for deriving a weighted nominal value ($\bar{x}_w$) and related uncertainty (GSD_w^2) from a set of data points with different quality and completeness information for a specific substance property. An illustrative example of the calculation of $\bar{x}_w$ and GSD_w^2 is presented in SI, Table S2 for an arbitrary substance across five different $c_k^{q,n'}$.

2.5. Case study: source, chemical property, and test substances selection

As an illustrative case study to test our proposed workflow, we focus on REACH-IUCLID as data source and K_{ow} as an example chemical property as an important predictor of variables relevant for estimating environmental fate, exposure, and (eco-)toxicological effects for most organic substances. In IUCLID, the substance registration dossiers follow a standardized structure and are divided into sections (e.g., 'Physical & Chemical properties') and sub-sections (e.g., 'Partition coefficient'). Per sub-section, one or more test results for a given substance property could be gathered. Each test result presents various supporting information together with a reported value, such as test method, experimental conditions, and administrative information. We gathered data from the 'Partition coefficient' (i.e., K_{ow}) sub-section for all REACH registration dossiers and built a dataset composed of 30,312 data points covering 11,053 unique substances (see SI, Fig. S1 for more statistics on the built dataset).

To select a set of representative chemical compounds covering a wide range of physicochemical property space, we matched the CAS numbers of the list of substances registered under REACH against the scientific consensus model USEtox (Rosenbaum et al., 2008). For the matching substances, we retrieved fate and exposure information covering half-life in air ($DT50_{air}$) and soil ($DT50_{soil}$), K_{ow}, K_{aw} (air-water partition coefficient), and estimated iF (the human intake fraction) from usetox.org and annual tonnage bands (i.e., 1–10, 10–100, 100–1000, or 1000+ tonnes/year production/import/export volume in the European Economic Area) from REACH (European Commission, 2006). We binned the matching substances into nine categories based on their $DT50_{air}$ and $DT50_{soil}$. The two-dimensional binning was performed using the bivariate histogram bin counts function histcounts2 (X, Y) in MATLAB. This function uses a set of automatic binning algorithms, including Scott's normal reference rule (Scott, 2010) and the Freedman–Diaconis' rule (Freedman and Diaconis, 1981), applied depending on the structure of the data to be binned. This function returns uniform bins chosen to cover the range of values in X (i.e., $DT50_{soil}$) and Y (i.e., $DT50_{air}$) and reveals the underlying shape of the distribution. From the binning results, we selected 20 substances from the nine bins to cover the spectrum of chemicals as widely as possible, as well as all the annual tonnage bands reported under REACH. During the selection of the 20 test substances, we filtered out chemicals with the potential to ionize (dissociate), since for these substances (i.e., ionizable organic chemicals), the octanol-water distribution is a function of pH and hence the K_{ow} does not apply (IUPAC, 1997).

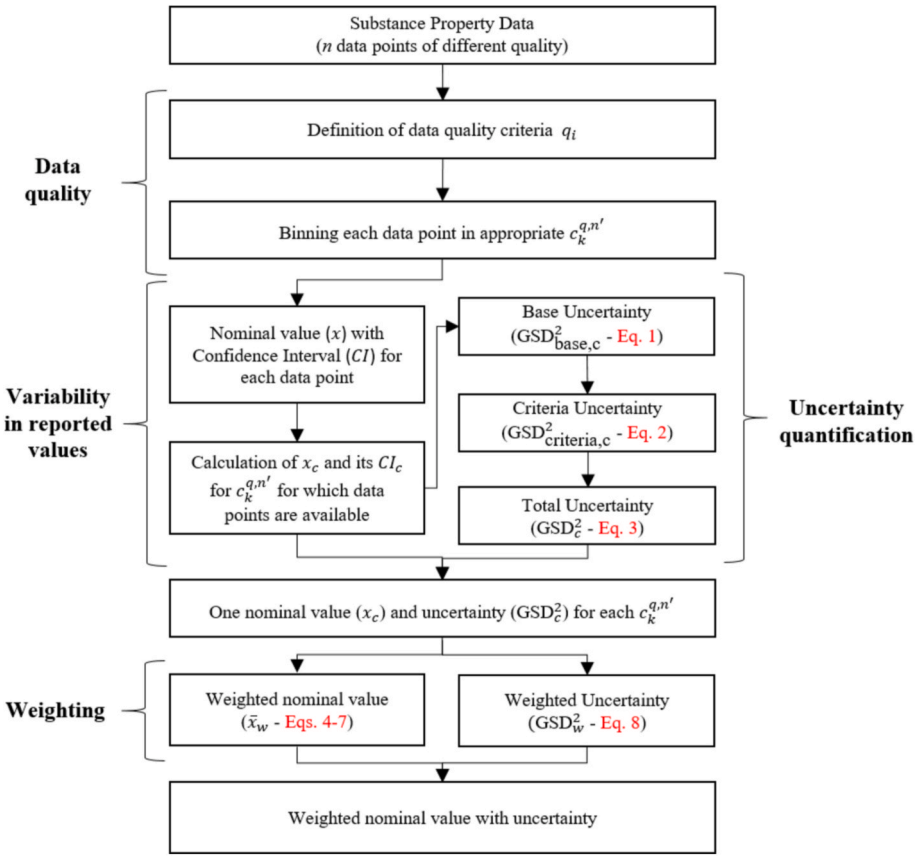


Fig. 1. Graphical overview for deriving weighted nominal values ($\bar{x}_w$) with uncertainty (GSD_w^2) from a set of data points across combinations of aggregate-criteria classes $c_k^{q,n'}$ for a specific substance property.

3. Results

3.1. Data harmonization and selection workflow

Before applying the developed workflow for deriving per chemical weighted nominal values with uncertainty for a given substance property (Fig. 1), data from specific sources might require pre-processing, including structuring, interpretation, and harmonization. In our illustrative case study, pre-processing tasks included: disregarding double entries, harmonizing units and reported values (e.g., conversion of all the reported log-scale K_{ow} values into normal scale), and matching and checking the reported CAS number with the actual tested substance (i.e., test material information). This aspect is crucial to consider, depending on the context in which REACH data are to be used, since REACH dossiers may contain data from very similar yet distinct molecules (structural analogue or surrogate substance) to, e.g., minimize laboratory experiments and support reported results. In these cases, it is essential to update the registered CAS number of the reported substance with the CAS number of the tested substance (Aurisano et al., 2019). Another pre-processing step specific to this case study on K_{ow} is to filter out chemicals with the potential to ionize, as applied for the selection of the 20 test substances. Other filtering steps might be needed for other chemical properties, depending on the specific parameter characteristics.

As first part of the proposed workflow, data quality criteria q_i and related combinations of aggregate-criteria classes $c_k^{q,n'}$ were defined. We defined three data quality criteria for our case study: Reliability, Purpose Flag, and Type of Information. Data reliability is evaluated with the Klimisch scoring system, from 1 (reliable without restrictions) to 4 (not assignable) (Klimisch et al., 1997). We note that the Klimisch scoring system was initially developed for evaluating experimental toxicological and ecotoxicology data, but may be extended to evaluate physical-chemical properties data, including K_{ow}, in certain regulatory settings (e.g., European Chemicals Agency, 2011; Ingre-Khans et al., 2019). However, when a high precision for specific decision contexts is required, we recommend to include the specific, relevant test conditions into our criteria-based approach rather than relying on more generic scoring systems. The quality criterion Purpose Flag (relevance) defines within REACH the usefulness of data for hazard/risk assessment purposes, and Type of Information (adequacy) outlines a data point's origin (European Commission, 2006; Przybylak et al., 2012; Sobanska et al., 2014). The available options when registering a substance through ECHA's IUCLID system for the three considered data quality criteria q_i were directly used to define the classes $n_{i,j}$ that are summarized in SI (Table S3). For the q_i Purpose Flag, examples of $n_{i,j}$ include Key study, Weight of Evidence (WoE), and Supporting study, among others. Table S3 furthermore presents the grouping of the different $n_{i,j}$ into grouped classes $n'_{i,j}$, while the 18 defined combinations of aggregate-criteria classes are listed in Table S4.

Not all the information reported for a given data point is necessary for our proposed method. For example, in our case study, the information summarized is CAS number, Reliability, Purpose Flag, Type of Information, Test material in raw data, Test Method, Good Laboratory Practice (GLP), Temperature, pH, physicochemical property value (or range), and qualifiers. The information considered may vary across different substance properties of interest and sources; as a basic rule, at least the information relevant for and covering the data quality criteria should be gathered.

In addition, we propose an optional filtering process. More specifically, during the filtering, only data points respecting the boundary conditions, purpose, and reliability tolerance of the application context are kept (e.g., only experimental results) by considering data points belonging only to specific $c_k^{q,n'}$. Moreover, other filters distinct from the defined $c_k^{q,n'}$ can be applied. Examples include considering only test results conducted under specific experimental conditions (e.g., certain pH, temperature), and following an appropriate, standardized testing method (e.g., 'shake-flask' method (OECD 107) for substances with log $K_{ow} < 4$ or 'slow-stirring' method (OECD 123) for substances with log $K_{ow} \geq 5$), or data from tests that follow GLP. We highlight that the filtering *per se* will not influence the actual harmonization process nor the pre-processing but is restricting (or expanding) the availability of data points based on the case-specific application requirements. All 18 $c_k^{q,n'}$ are considered in our case study not to disregard *a priori* any data point regardless of its quality (see SI, Table S4).

Once all relevant combinations of aggregate-criteria classes $c_k^{q,n'}$ are defined, the proposed workflow can be implemented. First, each available data point is allocated to the proper $c_k^{q,n'}$. In parallel, a nominal value (x) and its *CI* are derived for each data point, applying a variability factor F_V for estimating *CI* in cases where raw values are not reported already as ranges. In our case study, $F_V = 15.85$ was implemented based on the statistical analysis on the K_{ow} variability in the entire K_{ow} REACH-IUCLID dataset (~6000 data points as range and ~25,000 as single values) covering the entire chemical space of substances registered under REACH. With that, F_V represents an average variability in reported K_{ow} values across all combination of aggregate-criteria classes. Second, for each combination of aggregate-criteria classes $c_k^{q,n'}$, a x_c and its CI_c are estimated based on the data points available for the same combination. Third, for each defined $c_k^{q,n'}$, both base uncertainty ($GSD^2_{base,c}$, Eq. (1)) and criteria uncertainty ($GSD^2_{criteria,c}$, Eq. (2)) are estimated and combined to quantify the overall uncertainty (GSD^2_c, Eq. (3)) around x_c. The Pedigree matrix of criteria uncertainty factors used for quantifying $GSD^2_{criteria,c}$ in our case study is presented in Table 1. In the Pedigree matrix, we assign to each grouped class an uncertainty factor reflecting its quality and completeness. Note that using a criteria uncertainty factor of 1 in case of high-quality (e.g., Reliability: K 1, 2) is adding no criteria-related uncertainty.

Fourth, the results are summarized and reported per combination of aggregate-criteria classes. With that, the final case study results comprise the set of originally reported raw data (values and related key information), as well as x_c, its related GSD^2_c and the number of data points on which each x_c is based. By providing in the results also the set of originally reported values (i.e., raw data), we ensure traceability as well as reproducibility of the curated values and help practitioners to put them in perspective of the underlying available data.

The weighting process is implemented as a last step to deliver a unique nominal value ($\overline{x}_w$, Eq. (7)) and its related uncertainty (GSD^2_w, Eq. (8)) across considered combinations of aggregate-criteria classes. Table 2 summarizes the quality weights ($w_{Q,c}$) calculated for each

Table 1
Pedigree matrix with uncertainty factors for calculating the criteria related uncertainty ($GSD^2_{criteria,c}$) for our illustrative case study on K_{ow}.

Criterion	Grouped classes ($n'_{i,j}$)	Criteria uncertainty factors*
Reliability	K 1, 2	1
	K 3, 4 & other	10
Purpose Flag	Key study	1
	Supporting, WoE	5
	Other Study	10
Type of Information	Experimental	1
	Calculated, estimated	5
	Other	10

Uncertainty factors based on expert judgment follow the approach by Frischknecht et al. (2005). When other information on data quality becomes available, such default uncertainty factors can be refined accordingly (Slob, 1994). WoE: Weight of Evidence. Purpose flag "Other Study" and Type of Information "Other" include missing or not specified information (see Tables S3 for more examples).

Table 2
Set of quality weights ($w_{Q,c}$), calculated as inverse of $GSD^2_{criteria,c}$, and as used in our illustrative case study. $w_{Q,c}$ are assigned to each of the 18 combinations of aggregate-criteria classes for calculating a unique $\bar{x}_w$ across combinations.

Reliability	Purpose Flag	Type of Information	$GSD^2_{criteria,c}$	$w_{Q,c}$
K 1, 2	Key Study	Experimental	1.0	1.00
K 1, 2	Key Study	Calculated, estimated	5.0	0.20
K 1, 2	Key Study	Other	10.0	0.10
K 1, 2	Supporting, WoE	Experimental	5.0	0.20
K 1, 2	Supporting, WoE	Calculated, estimated	9.7	0.10
K 1, 2	Supporting, WoE	Other	16.6	0.06
K 1, 2	Other Study	Experimental	10.0	0.10
K 1, 2	Other Study	Calculated, estimated	16.6	0.06
K 1, 2	Other Study	Other	26.0	0.04
K 3, 4 & other	Key Study	Experimental	10.0	0.10
K 3, 4 & other	Key Study	Calculated, estimated	16.6	0.06
K 3, 4 & other	Key Study	Other	26.0	0.04
K 3, 4 & other	Supporting, WoE	Experimental	16.6	0.06
K 3, 4 & other	Supporting, WoE	Calculated, estimated	25.5	0.04
K 3, 4 & other	Supporting, WoE	Other	37.8	0.03
K 3, 4 & other	Other Study	Experimental	26.0	0.04
K 3, 4 & other	Other Study	Calculated, estimated	37.8	0.03
K 3, 4 & other	Other Study	Other	54.0	0.02

WoE: Weight of Evidence. Purpose flag "Other Study" and Type of Information "Other" include missing or not specified information (see Tables S3 for more examples).

combination for estimating $\bar{x}_w$ for the substances in our case study. As can be seen from Table 2, the higher the $GSD^2_{criteria,c}$, the lower its $w_{Q,c}$ and, thus, the lower its influence on the final $\bar{x}_w$.

Finally, Fig. 2 summarizes the complete harmonization and selection workflow applied in our illustrative case study, from the pre-processing of the dataset and optional filtering, the variability and uncertainty quantification and, the weighting process, to ultimately deriving for each considered substance a single K_{ow} and related uncertainty ranges as a function of the quality and completeness of the underlying, substance-specific raw data.

3.2. Case study results

3.2.1. Selected test substances

The 1125 substances registered under REACH and available in USEtox, constituting the test dataset for our illustrative case study, were binned into nine different categories based on their $DT50_{soil}$ and $DT50_{air}$. Specific ranges were assigned to DT50s to obtain a balanced distribution of substances, i.e., <0.2 days, 0.2–1 days, and >1 day for $DT50_{air}$ and <30 days, 30–75 days, and >75 days for $DT50_{soil}$. One to two substances were then selected from each bin. To simplify the selection of the test substances, we created a 3 × 3 matrix of contour plots displaying the considered physicochemical and fate/exposure information retrieved from USEtox (i.e., $DT50_{air}$, $DT50_{soil}$, K_{ow}, K_{aw}, and human intake fraction). The matrix of contour plots is presented in Fig. 3, while the list of selected case study substances, their yearly tonnage band, and the number of data points available in REACH are presented in Table 3. The selected substances range from low yearly produced/imported volume with less than 10 tonnes/year (e.g., Tetradecylamine CAS: 2016-42-4) to high-volume chemicals with more than 1000 tonnes/year (e.g., Diisononyl phthalate CAS: 28,553-12-0). Furthermore, different categories of chemicals are represented, ranging from industrial fungicides (e.g., Captan CAS: 133-06-2) to food and flavor ingredients (e.g., 2,4-Dimethylphenol CAS: 105-67-9). As a general trend, we observed in the REACH-IUCLID database that the higher the tonnage band, the higher the number of data points available.

3.2.2. Workflow results

We applied the data harmonization and selection workflow to derive octanol-water partition coefficient (K_{ow}) values for 20 test substances. For these substances, a total of 65 data points were available in REACH-IUCLID. As a first step of the pre-processing, we checked the consistency between the reported substances and the actual test material. There was a mismatch for around 35% of the gathered data points, and thus $n = 23$ data points were disregarded during the pre-processing. It is worth mentioning that this high mismatch rate is due to the fact that REACH dossiers are allowed to contain data from similar yet structurally different chemicals, which must be filtered when such data are used to provide a curated dataset that should only contain data for the specifically assessed chemicals as in our case study. No filtering was imposed in the case study; thus, all the 42 pre-processed data points were selected, potentially considering all defined combinations of aggregate-criteria classes (see SI, Table S4).

The weighted K_{ow}, i.e., the harmonization and selection process results with quantified uncertainty for the 20 test substances are summarized in Table 3. We acknowledge that in our specific example of using REACH dossiers as data source, the quantified uncertainty reflects compliance with standardized testing methods rather than actual accuracy or quality of the underlying data. Where appropriate, we suggest using other or additional quality-related criteria.

Across the weighted results, characterized GSD^2_w range from 2.82 (Diisononyl phthalate CAS: 28553-12-0) to 44.25 (2,2'-(ethylenedioxy) diethanol CAS: 112-27-6). For the former, the low GSD^2_w is driven by the high quality of the two available data points yielding low individual GSD^2_c. For the substance 4-methyl-1,3-dioxolan-2-one (CAS: 108-32-7), no high-quality data points (i.e., experimental – reliable – key study) were available; nevertheless, the high number of data points available ($n = 4$) and their low variability yielded a $GSD^2_w = 9.86$. Indeed, GSD^2_w is a function of various factors, including the number and quality of the available combinations of aggregate-criteria classes (the higher their quality, the lower the GSD^2_w), and the variability across curated K_{ow} values (the lower their variability, the lower the GSD^2_w). When for a substance, only one high-quality data point (i.e., experimental – reliable – key study) is available with K_{ow} reported as a single value, the reported $GSD^2_w = 15.85$ directly reflects F_V, since the three quality-related criteria provide no additional uncertainty on top of uncertainty related to the data variability. Thus, in such cases, the results could also be provided without the need for an actual weighting process and full uncertainty characterization. This correctly reflects lower confidence in reported point estimates, which could be both mean or outlier values, hence the high associated default variability.

The weighted K_{ow} results are presented in Fig. 4 together with the underlying raw data found in SI (Table S5), differentiating between high-quality data points (i.e., experimental – reliable - key study) and other data. Fig. 4 highlights the importance of considering a 95% *CI* around the weighted values for putting the results into perspective and how high-quality data drive the final results via the weighting process. The weighted K_{ow} from 2 data points (allocated to different aggregate-criteria classes $c_k^{q,n'}$) for Diisononyl phthalate (CAS: 28553-12-0) is a clear example of this effect. As graphically represented in Fig. 4, the estimated value is mainly driven by the available high-quality data point (black cross).

Finally, the reported data for any substance (including our 20 test substances) under REACH may be updated by registrants at any time after we retrieved the data used in the present study in July 2021 (Sobanska et al., 2014). If such additional data become available, our results can be updated following the same procedure as outlined for our case study.

4. Discussion

4.1. Applicability of the proposed method

Our workflow should be seen as a first step to create a standardized

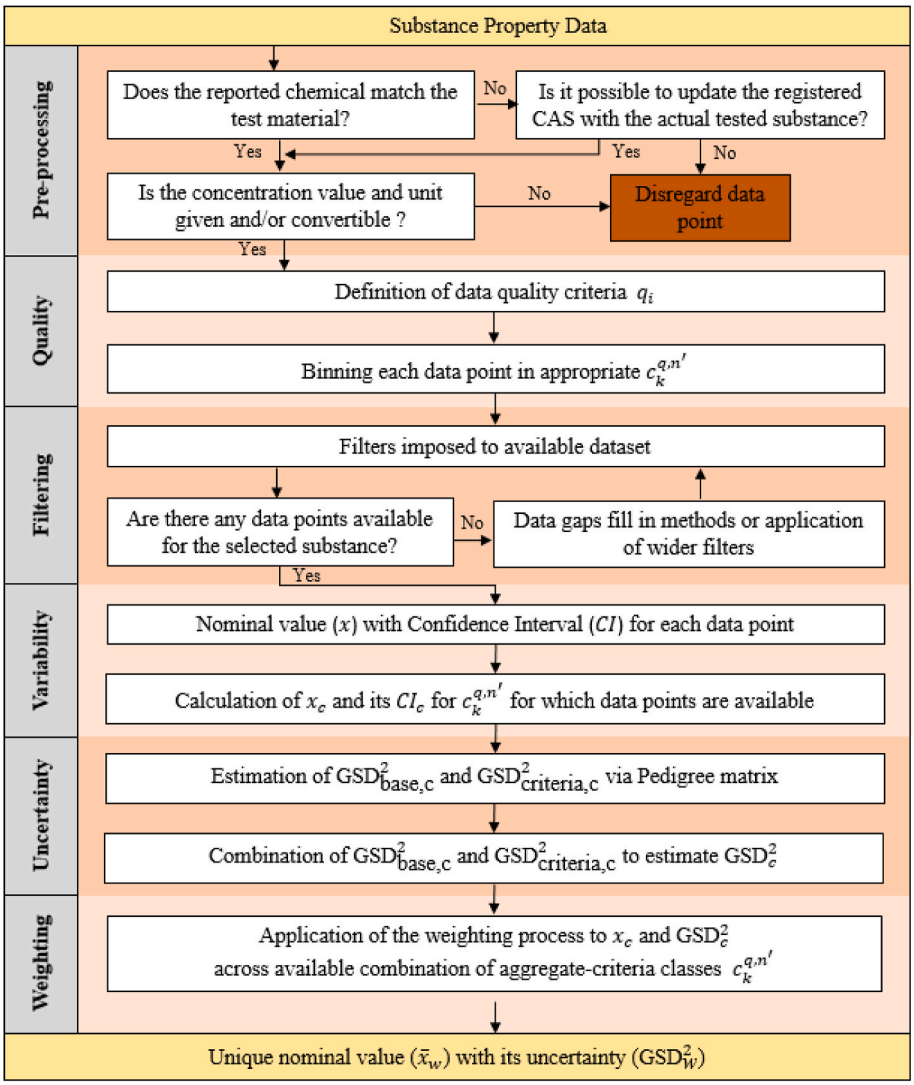


Fig. 2. Flow chart presenting the harmonization and selection workflow applied in our illustrative case study to derive a unique nominal value ($\bar{x}_w$) and its uncertainty (GSD^2_w) across combinations of aggregate-criteria classes from a set of raw data points of different quality and completeness.

method for harmonizing and selecting chemical property information across different data sources and when more than one data point is reported. The criteria and related uncertainty factors presented in this study apply to K_{ow} data available in REACH-IUCLID (since we focused our case study on this specific data source and chemical property). Even though our selected data source has a standardized structure, additional criteria and related uncertainty factors might need to be defined for applying the proposed workflow to other chemical properties within the same data source. For example, additional criteria need to be considered for oral toxicity data, such as exposure duration or species tested, which are critical for the interpretation of this property (Fantke et al., 2021a). Another example is the octanol-water distribution ratio (D_{ow}), relevant for ionizable organic chemicals, for which additional test parameters, such as pH, need to be considered (IUPAC, 1997).

Similarly, criteria and related uncertainty factors need to be defined when adapting the proposed workflow to other data sources. For example, Reliability or Purpose Flag are not provided in results reported in the US-EPA CompTox Chemistry Dashboard (Williams et al., 2017) or ChemSpider (Pence and Williams, 2010). For these databases, the source of the reported information is potentially a criterion since systematically provided. No adaptations would be needed for the filtering and the weighting processes since these processes are independent of the included data sources.

The implementation of the filtering process enables practitioners to apply our harmonization and selection workflow in different contexts. This includes, for example, Life Cycle Impact Assessment (LCIA), where numerous chemicals have to be characterized, using a wide range of underlying data of varying quality, or the development of QSAR models, where only the most reliable experimental results are usually considered for model-building (Cherkasov et al., 2014; Cronin and Schultz, 2003). The final weighting process additionally enables to account for differences across data points in terms of their data quality and completeness. However, the weighting process is only needed in cases where results from more than one combination of aggregate-criteria classes fit the relevant application context.

4.2. Limitations of our workflow

The proposed workflow comes with limitations. For example, the final results are dependent on the criteria considered for assessing the quality of each data point and on the criteria-related uncertainty factors (i.e., pedigree matrix) applied to quantify their confidence interval. More precisely, the assessed 'quality' of the data reflects compliance and consistency with reporting and standardized test guidelines; thus, it does not necessarily express the actual accuracy (even though standardized and well-conducted tests are more likely to be accurate). Vice versa, data

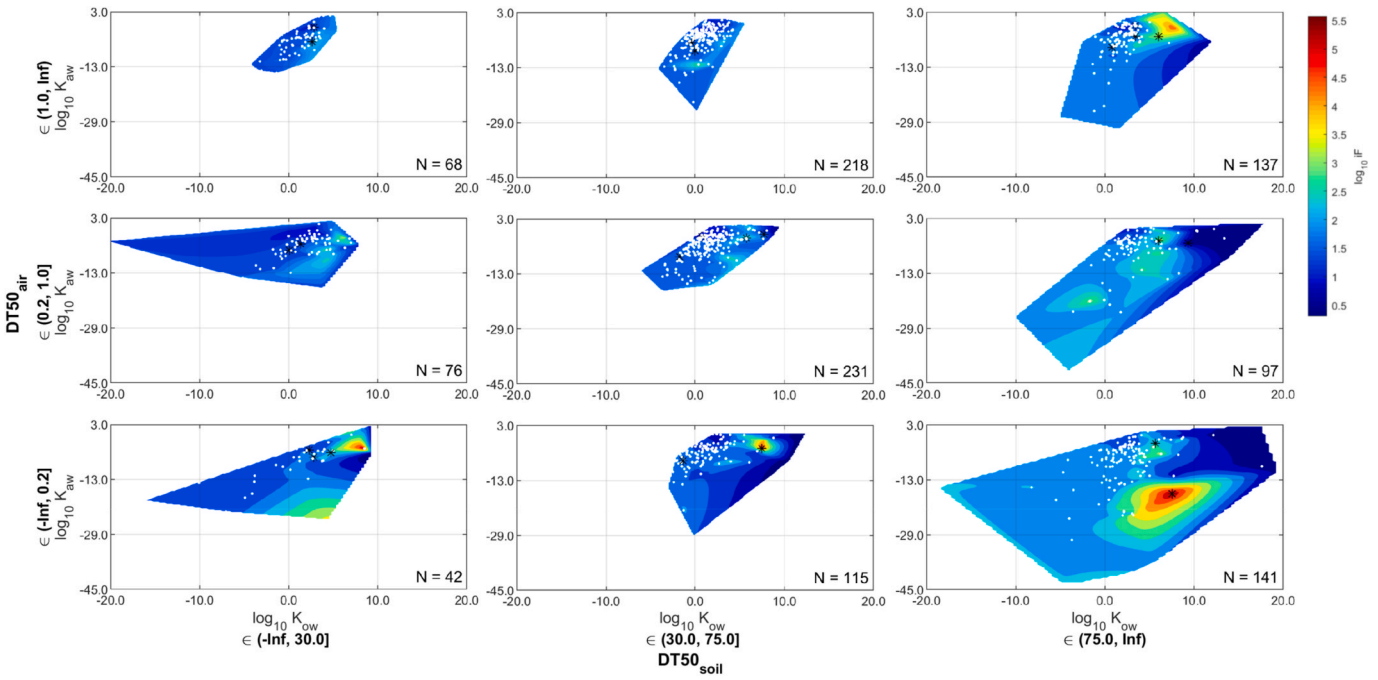


Fig. 3. Multidimensional space maps for physicochemical properties covered by the 1125 substances present in both USEtox and REACH. Black stars: Selected test substances for our illustrative case study ($n = 20$). White dots: remaining 1108 substances. $DT50_{soil}$ is increasing from left to right plots, and $DT50_{air}$ is increasing from bottom to top plots. In each subplot, $\log_{10} K_{aw}$ (y-axes) is plotted against $\log_{10} K_{ow}$ (x-axes). Intake fractions, log iF, provide an additional dimension and are represented by colors. The number of substances present per bin is shown in each subplot. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 3
Results of K_{ow} weighted based on applying the proposed harmonization and selection workflow for the 20 test substances in our illustrative case study.

CAS	Number data points[a]	Tonnage band [tonnes/y][b]	log K_{ow}	GSD_w^2
106-46-7	4	1000+	3.37	10.66
108-32-7	4	1000+	−0.45	9.86
2016-42-4	4	1–10	5.91	29.32
112-27-6	3	1000+	−1.83	44.25
94-36-0	3	1000+	3.24	9.22
7173-62-8	2	1000+	0.03	10.63
28553-12-0	2	1000+	9.25	2.82
112-92-5	2	1000+	7.38	13.17
78-84-2	2	1000+	0.77	12.94
67-68-5	2	1000+	−1.34	11.95
732-26-3	2	100–1000	7.09	15.78
133-06-2	2	100–1000	2.54	17.18
100-55-0	2	1–10	−0.86	12.64
105-67-9	2	–	2.46	33.47
5567-15-7	1	1000+	0.02	15.85
51-03-6	1	1000+	4.80	15.85
117-80-6	1	10–100	2.90	15.85
7212-44-4	1	10–100	4.50	15.85
109-21-7	1	1–10	2.16	15.85
26140-60-3	1	–	5.86	41.44

[a] After the pre-processing.
[b] For the yearly tonnage band, in case of confidential information, the cell is left empty.

points with low assessed quality, driven by, e.g., missing reporting information or tests not conducted according to GLP, do not necessarily deliver inaccurate results (Ingre-Khans et al., 2020; Przybylak et al., 2012). Consequently, the results of the quality assessment performed in our study represent a measure of confidence in the processed results rather than accuracy of the delivered values. For example, specific to the REACH-IUICLID data source, it has been recently highlighted that the reported Klimisch scores are mainly based on studies complying with GLP and test guidelines, with the risk of considering GLP and guideline studies being reliable by default and overlooking non-GLP and non-test guideline data that might be of high-quality (Ingre-Khans et al., 2019).

Currently, we included broad criteria that account for test conditions of reported data (e.g., Klimisch score), where specific conditions (e.g., certain pH range or temperature) are not considered separately. This could influence uncertainty estimates around parameter results (Beyer et al., 2002; Lei et al., 2004; OECD, 2006; Sangster, 1989). Where appropriate, such specific test conditions could be explicitly accounted for in our approach as separate criteria, by, e.g., filtering reported data points to consider only results conducted under specific experimental conditions or in compliance with specific standard testing guidelines. For example, when a chemical property information varies widely as a function of test conditions, such as pH, the large variability in test results will propagate into wider confidence intervals when a generic value is required for a given decision context. In contrast, a pH-specific value could be reported for specific contexts, where pH itself could then be used as a criterion to determine variability, and related confidence ranges across reported test data.

We have applied our proposed approach to a rather small number of test chemicals in our case study. However, our workflow can generally be applied to a large number of chemicals and data depending on the data source conditions. This is possible since the different data points are allocated to the different criteria classes and weighted results are automatically derived after quality criteria are defined manually (hence, semi-automated).

Our derived log $K_{ow} = 0.03$ for CAS: 7173-62-8 is derived from a weighted average across structurally-related compounds, with an originally reported experimental value of log $K_{ow} = 0$ specifically for our target chemical with CAS: 7173-62-8. We note that the reported experimental result for this chemical is very different from predicted values, which are in the range of log $K_{ow} = 7.08$ to 8.63 (Mansouri et al., 2018). We observe similar discrepancies also for CAS: 5567-15-7 with predicted log $K_{ow} = 4.73$ to 9.45 (Mansouri et al., 2018). In cases where

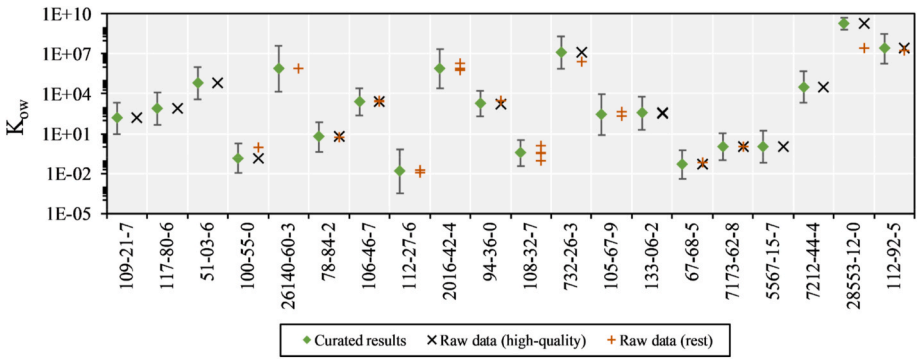


Fig. 4. Weighted K_{ow} results with estimated confidence intervals 95% *CI* (represented as error bars) and raw data used for the 20 test substances in our illustrative case study, differentiating between high-quality (i.e., experimental – reliable - key study) and other data (rest).

such discrepancies occur between measured and predicted data, we emphasize the need for further research to better understand where these differences are coming from, which includes refinement of prediction models as well as additional experimental testing, including full reporting of specific test conditions. The range in which the true value for a property will lie could be constrained by using information for all partitioning properties simultaneously rather than one at a time. This could be adapted based on available approaches (e.g., Cole and Mackay, 2000; Ma et al., 2010; Wenger et al., 2012; Xiao et al., 2004).

In addition, there are still fields in the registration dossiers where additional information can be submitted via free text. In the proposed method, we do not implement any data mining tool, disregarding all information in free text fields and thus might miss relevant additional information for some data points.

Finally, combinations of aggregate-criteria classes were also created for treating non-experimental data points (e.g., QSARs). We acknowledge that for some parameters and decision contexts, robust QSAR estimates might be more appropriate than individual experimental results reported for specific, not necessarily representative test conditions (Fantke et al., 2014). We further acknowledge that QSARs differ in terms of compliance with recommended validation approaches, applicability domain and predictive power—information that is unfortunately not always provided. Where needed, such aspects should hence be included in our approach as explicit QSAR-related quality criteria, which could also be done by considering reported uncertainty ranges for specific QSARs.

4.3. Recommendations and future research needs

From this first step toward developing overarching principles and methods for pre-processing, harmonizing, and selecting chemical property data from different data sources for various application fields, we recommend that results should always contain the set of raw data that was used to derive any processed result to ensure traceability as well as reproducibility. We note that where this is not possible when using, for example, any proprietary or otherwise protected data, this comes at the expense of data transparency. Furthermore, we recommend including uncertainty estimates around any processed data as it reflects differences in quality and completeness of the underlying raw data and ensures maximum interpretability.

Future research needs include identifying and developing sets of quality criteria to derive uncertainty factors for other substance properties and other available data sources (e.g., US-EPA CompTox Chemistry Dashboard, ChemSpider). In addition, the default uncertainty factors, currently based on expert judgment and used to quantify criteria-related uncertainty ($GSD^2_{\text{criteria,c}}$) and quality weights ($w_{Q,c}$), should be refined when more specific information on data quality becomes available, starting from available approaches to derive uncertainty from data quality and relevance such as the framework for quantitative weight-of-evidence analysis developed by Bridges et al. (2017).

Moreover, further research could extend the proposed method for mining and interpreting information from text fields and other unstructured yet relevant information, as used in some data sources, including REACH-IUCLID. However, when potentially relevant data are reported in “free text” sections, additional pre-processing, interpretation, and harmonization is needed since possible typos, ambiguity, or irrelevance in the information reported might occur.

Finally, even if we considered a small set of test substances in our illustrative case study, we have identified few discrepancies between the reported experimental values and other sources reporting predicted values. At the same time, we have also observed a lack of data completeness and transparency in the considered dossiers for some of the test substances. These inconsistencies are urging for REACH data to be subject to a significant quality review based on a clearly defined set of criteria to increase the consistency and robustness of any related data used in assessments and decision support (Fantke et al., 2020; Ingre--Khans et al., 2020).

5. Conclusions

We developed a criteria-based method to enable the use of reported chemical property information of different quality and provide quantitative uncertainty information around the resulting values. With that, our proposed workflow may serve as starting point for systematically developing data harmonization and selection tools that build on criteria and uncertainty estimates tailored toward the specific characteristics of different chemical properties and data sources. We tested our method for deriving K_{ow} values and related confidence intervals for 20 test substances from a set of underlying raw data of different quality. Our proposed workflow is suitable to assess both high- and low-information substances as input to various modeling approaches with different resolution and data quality requirements, from life cycle impact assessment to chemical substitution and high-throughput risk screening.

Credit author statement

Nicolò Aurisano: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Peter Fantke:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft; Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

Acknowledgments

We thank P. Karamertzanis and J. Provoost (European Chemicals Agency) for valuable comments and for sharing REACH data, and M. Hauschild (Technical University of Denmark) and O. Joliet (University of Michigan) for initial comments on the data curation method. This work was financially supported by the "Safe and Efficient Chemistry by Design (SafeChem)" project funded by the Swedish Foundation for Strategic Environmental Research (grant no. DIA 2018/11), and by a collaborative project between the European Chemicals Agency and the Technical University of Denmark (grant agreement no. ECHA/2017/445).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemosphere.2022.134886.

References

Askham, C., 2012. REACH and LCA-methodological approaches and challenges. Int. J. Life Cycle Assess. 17, 43–57. https://doi.org/10.1007/s11367-011-0329-z.

Aurisano, N., Albizzati, P.F., Hauschild, M., Fantke, P., 2019. Extrapolation factors for characterizing freshwater ecotoxicity effects. Environ. Toxicol. Chem. 38, 2568–2582. https://doi.org/10.1002/etc.4564.

Aurisano, N., Fantke, P., Huang, L., Jolliet, O., 2022. Estimating mouthing exposure to chemicals in children's products. J. Expo. Sci. Environ. Epidemiol. 32, 94–102. https://doi.org/10.1038/s41370-021-00354-0.

Aurisano, N., Huang, L., Milà i Canals, L., Jolliet, O., Fantke, P., 2021a. Chemicals of concern in plastic toys. Environ. Int. 146, 106194. https://doi.org/10.1016/j.envint.2020.106194.

Aurisano, N., Weber, R., Fantke, P., 2021b. Enabling a circular economy for chemicals in plastics. Curr. Opin. Green Sustain. Chem. 31, 100513. https://doi.org/10.1016/j.cogsc.2021.100513.

Barbaro, B., Baldin, R., Kovarich, S., Pavan, M., Fioravanzo, E., Bassan, A., 2015. Further development and update of EFSA's chemical hazards database. EFSA Support. Publ. EN- 823, 1–84. https://doi.org/10.2903/sp.efsa.2015.en-823.

Beyer, A., Wania, F., Gouin, T., Mackay, D., Matthies, M., 2002. Selecting internally consistent physicochemical properties of organic compounds. Environ. Toxicol. Chem. 21, 941–953. https://doi.org/10.1002/etc.5620210508.

Bolton, E.E., Wang, Y., Thiessen, P.A., Bryant, S.H., 2008. Chapter 12 PubChem: integrated platform of small molecules and biological activities. In: Annual Reports in Computational Chemistry, pp. 217–241. https://doi.org/10.1016/S1574-1400(08)00012-1.

Bridges, J., Sauer, U.G., Buesen, R., Deferme, L., Tollefsen, K.E., Tralau, T., van Ravenzwaay, B., Poole, A., Pemberton, M., 2017. Framework for the quantitative weight-of-evidence analysis of 'omics data for regulatory purposes. Regul. Toxicol. Pharmacol. 91, S46–S60. https://doi.org/10.1016/j.yrtph.2017.10.010.

Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C., Todeschini, R., Consonni, V., Kuz'Min, V.E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., Tropsha, A., 2014. QSAR modeling: where have you been? Where are you going to? J. Med. Chem. 57, 4977–5010. https://doi.org/10.1021/jm4004285.

Ciroth, A., Muller, S., Weidema, B., Lesage, P., 2016. Empirically based uncertainty factors for the pedigree matrix in ecoinvent. Int. J. Life Cycle Assess. 21, 1338–1348. https://doi.org/10.1007/s11367-013-0670-5.

Cole, J.G., Mackay, D., 2000. Correlating environmental partitioning properties of organic compounds: the three solubility approach. Environ. Toxicol. Chem. 19, 265–270. https://doi.org/10.1002/etc.5620190203.

Cronin, M.T.D., Schultz, T.W., 2003. Pitfalls in QSAR. J. Mol. Struct. THEOCHEM 622, 39–51. https://doi.org/10.1016/S0166-1280(02)00616-4.

Dorne, J.L., Richardson, J., Kass, G., Georgiadis, N., Monguidi, M., Pasinato, L., Cappe, S., Verhagen, H., Robinson, T., 2017. Editorial: OpenFoodTox: EFSA's open source toxicological database on chemical hazards in food and feed. EFSA J. 15, 15011. https://doi.org/10.2903/j.efsa.2017.e15011.

European Chemicals Agency (ECHA), 2011. Guidance on Information Requirements and Chemical Safety Assessment. Chapter R. 4: Evaluation of Available Information.

European Commission, 2006. REGULATION (EC) No 1907/2006 of the EUROPEAN PARLIAMENT and of the COUNCIL of 18 December 2006 Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), vol. 849. Off. J. Eur. Union.

Fantke, P., Aurisano, N., Bare, J., Backhaus, T., Bulle, C., Chapman, P.M., De Zwart, D., Dwyer, R., Ernstoff, A., Golsteijn, L., Holmquist, H., Jolliet, O., McKone, T.E., Owsianiak, M., Peijnenburg, W., Posthuma, L., Roos, S., Saouter, E., Schowanek, D., van Straalen, N.M., Vijver, M.G., Hauschild, M., 2018. Toward harmonizing ecotoxicity characterization in life cycle impact assessment. Environ. Toxicol. Chem. 37, 2955–2971. https://doi.org/10.1002/etc.4261.

Fantke, P., Aurisano, N., Provoost, J., Karamertzanis, P.G., Hauschild, M., 2020. Toward effective use of REACH data for science and policy. Environ. Int. 135, 105336. https://doi.org/10.1016/j.envint.2019.105336.

Fantke, P., Chiu, W.A., Aylward, L., Judson, R., Huang, L., Jang, S., Gouin, T., Rhomberg, L., Aurisano, N., McKone, T., Jolliet, O., 2021a. Exposure and toxicity characterization of chemical emissions and chemicals in products: global recommendations and implementation in USEtox. Int. J. Life Cycle Assess. 26, 899–915. https://doi.org/10.1007/s11367-021-01889-y.

Fantke, P., Cinquemani, C., Yaseneva, P., De Mello, J., Schwabe, H., Ebeling, B., Lapkin, A.A., 2021b. Transition to sustainable chemistry through digitalisation. Inside Chem. 7 https://doi.org/10.1016/j.chempr.2021.09.012.

Fantke, P., Gillespie, B.W., Juraske, R., Jolliet, O., 2014. Estimating half-lives for pesticide dissipation from plants. Environ. Sci. Technol. 48, 8588–8602. https://doi.org/10.1021/es500434p.

Fantke, P., Wieland, P., Juraske, R., Shaddick, G., Itoiz, E.S., Friedrich, R., Jolliet, O., 2012. Parameterization models for pesticide exposure via crop consumption. Environ. Sci. Technol. 46, 12864–12872. https://doi.org/10.1021/es301509u.

Fourches, D., Muratov, E., Tropsha, A., 2016. Trust, but verify II: a practical guide to chemogenomics data curation. J. Chem. Inf. Model. 56, 1243–1252. https://doi.org/10.1021/acs.jcim.6b00129.

Freedman, D., Diaconis, P., 1981. On the histogram as a density estimator:L2 theory. Z. Wahrscheinlichkeitstheor. Verwandte Geb. 57, 453–473. https://doi.org/10.1007/BF01025868.

Frischknecht, R., Jungbluth, N., Althaus, H.J., Doka, G., Dones, R., Heck, T., Hellweg, S., Hischier, R., Nemecek, T., Rebitzer, G., Spielmann, M., 2005. The ecoinvent database: overview and methodological framework. Int. J. Life Cycle Assess. 10, 3–9. https://doi.org/10.1065/lca2004.10.181.1.

Hong, J., Shaked, S., Rosenbaum, R.K., Jolliet, O., 2010. Analytical uncertainty propagation in life cycle inventory and impact assessment: application to an automobile front panel. Int. J. Life Cycle Assess. 15, 499–510. https://doi.org/10.1007/s11367-010-0175-4.

Hou, P., Jolliet, O., Zhu, J., Xu, M., 2020. Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. Environ. Int. 135, 105393. https://doi.org/10.1016/j.envint.2019.105393.

Igos, E., Moeller, R., Benetto, E., Biwer, A., Guiton, M., Dieumegard, P., 2014. Development of USEtox characterisation factors for dishwasher detergents using data made available under REACH. Chemosphere 100, 160–166. https://doi.org/10.1016/j.chemosphere.2013.11.041.

Ingre-Khans, E., Ågerstrand, M., Beronius, A., Rudén, C., 2019. Reliability and relevance evaluations of REACH data. Toxicol. Res. (Camb). 8, 46–56. https://doi.org/10.1039/c8tx00216a.

Ingre-Khans, E., Ågerstrand, M., Rudén, C., Beronius, A., 2020. Improving structure and transparency in reliability evaluations of data under REACH: suggestions for a systematic method. Hum. Ecol. Risk Assess. 26, 212–241. https://doi.org/10.1080/10807039.2018.1504275.

IUPAC, 1997. Compendium of chemical terminology. In: The "Gold Book"). Compiled by A. D. McNaught and A. Wilkinson, second ed. Blackwell Scientific Publications, Oxford, ISBN 0-9678550-9-8. https://doi.org/10.1351/goldbo. Online version (2019-) created by S. J. Chalk.

Klimisch, H.J., Andreae, M., Tillmann, U., 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul. Toxicol. Pharmacol. 25, 1–5. https://doi.org/10.1006/rtph.1996.1076.

Lei, Y.D., Wania, F., Mathers, D., Mabury, S.A., 2004. Determination of vapor pressures, Octanol−Air, and Water−Air partition coefficients for polyfluorinated sulfonamide, sulfonamidoethanols, and telomer alcohols. J. Chem. Eng. Data 49, 1013–1022. https://doi.org/10.1021/je049949h.

Lewis, K.A., Tzilivakis, J., Warner, D.J., Green, A., 2016. An international database for pesticide risk assessments and management. Hum. Ecol. Risk Assess. 22, 1050–1064. https://doi.org/10.1080/10807039.2015.1133242.

Li, N., Wania, F., Lei, Y.D., Daly, G.L., 2003. A comprehensive and critical compilation, evaluation, and selection of physical–chemical property data for selected polychlorinated biphenyls. J. Phys. Chem. Ref. Data 32, 1545–1590. https://doi.org/10.1063/1.1562632.

Luechtefeld, T., Maertens, A., Russo, D.P., Rovida, C., Zhu, H., Hartung, T., 2016. Global analysis of publicly available safety data for 9,801 substances registered under REACH from 2008-2014. ALTEX 33, 95. https://doi.org/10.14573/altex.1510052.

Ma, Y.-G., Lei, Y.D., Xiao, H., Wania, F., Wang, W.-H., 2010. Critical review and recommended values for the physical-chemical property data of 15 polycyclic aromatic hydrocarbons at 25 °C. J. Chem. Eng. Data 55, 819–825. https://doi.org/10.1021/je900477x.

MacLeod, M., Fraser, A.J., Mackay, D., 2002. Evaluating and expressing the propagation of uncertainty in chemical fate and bioaccumulation models. Environ. Toxicol. Chem. An Int. J. 21, 700–709. https://doi.org/10.1002/etc.5620210403.

Mansouri, K., Grulke, C.M., Judson, R.S., Williams, A.J., 2018. OPERA models for predicting physicochemical properties and environmental fate endpoints. J. Cheminf. 10, 10. https://doi.org/10.1186/s13321-018-0263-1.

Mansouri, K., Grulke, C.M., Richard, A.M., Judson, R.S., Williams, A.J., 2016. An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling. SAR QSAR Environ. Res. 27, 911–937. https://doi.org/10.1080/1062936X.2016.1253611.

Müller, N., de Zwart, D., Hauschild, M., Kijko, G., Fantke, P., 2017. Exploring REACH as a potential data source for characterizing ecotoxicity in life cycle assessment. Environ. Toxicol. Chem. 36, 492–500. https://doi.org/10.1002/etc.3542.

Muller, S., Lesage, P., Ciroth, A., Mutel, C., Weidema, B.P., Samson, R., 2016. The application of the pedigree approach to the distributions foreseen in ecoinvent v3.

Int. J. Life Cycle Assess. 21, 1327–1337. https://doi.org/10.1007/s11367-014-0759-5.
OECD, 2006. OECD GUIDELINES for the TESTING of CHEMICALS 123: Partition Coeficient (1-Octanol/Water): Slow-Stirring Method, OECD Guideline for the Testing of Chemicals. https://doi.org/10.1787/9789264015845-en.
Pence, H.E., Williams, A., 2010. Chemspider: an online chemical information resource. J. Chem. Educ. 87, 1123–1124. https://doi.org/10.1021/ed100697w.
Persson, L., Carney Almroth, B.M., Collins, C.D., Cornell, S., de Wit, C.A., Diamond, M.L., Fantke, P., Hassellöv, M., MacLeod, M., Ryberg, M.W., Søgaard Jørgensen, P., Villarrubia-Gómez, P., Wang, Z., Hauschild, M.Z., 2022. Outside the safe operating space of the planetary boundary for novel entities. Environ. Sci. Technol. 56, 1510–1521. https://doi.org/10.1021/acs.est.1c04158.
Posthuma, L., van Gils, J., Zijp, M.C., van de Meent, D., de Zwartd, D., 2019. Species sensitivity distributions for use in environmental protection, assessment, and management of aquatic ecosystems for 12 386 chemicals. Environ. Toxicol. Chem. 38, 905–917. https://doi.org/10.1002/etc.4373.
Przybylak, K.R., Madden, J.C., Cronin, M.T.D., Hewitt, M., 2012. Assessing toxicological data quality: basic principles, existing schemes and current limitations. SAR QSAR Environ. Res. 23, 435–459. https://doi.org/10.1080/1062936X.2012.664825.
Rosenbaum, R.K., Bachmann, T.M., Gold, L.S., Huijbregts, M.A.J., Jolliet, O., Juraske, R., Koehler, A., Larsen, H.F., MacLeod, M., Margni, M., McKone, T.E., Payet, J., Schuhmacher, M., Van De Meent, D., Hauschild, M.Z., 2008. USEtox - the UNEP-SETAC toxicity model: recommended characterisation factors for human toxicity and freshwater ecotoxicity in life cycle impact assessment. Int. J. Life Cycle Assess. 13, 532–546. https://doi.org/10.1007/s11367-008-0038-4.
Rosenbaum, R.K., Georgiadis, S., Fantke, P., 2018. Uncertainty management and sensitivity analysis. In: Hauschild, M.Z., Rosenbaum, R.K., Olsen, S.I. (Eds.), Life Cycle Assessment: Theory and Practice. Springer International Publishing, Cham, pp. 271–321. https://doi.org/10.1007/978-3-319-56475-3_11.
Sangster, J., 1989. Octanol water partition coefficients of simple organic compounds. J. Phys. Chem. Ref. Data 18, 1111–1229. https://doi.org/10.1063/1.555833.
Saouter, E., Aschberger, K., Fantke, P., Hauschild, M.Z., Bopp, S.K., Kienzler, A., Paini, A., Pant, R., Secchi, M., Sala, S., 2017a. Improving substance information in USEtox®, part 1: discussion on data and approaches for estimating freshwater ecotoxicity effect factors. Environ. Toxicol. Chem. 36, 3450–3462. https://doi.org/10.1002/etc.3889.
Saouter, E., Aschberger, K., Fantke, P., Hauschild, M.Z., Kienzler, A., Paini, A., Pant, R., Radovnikovic, A., Secchi, M., Sala, S., 2017b. Improving substance information in USEtox®, part 2: data for estimating fate and ecosystem exposure factors. Environ. Toxicol. Chem. 36, 3463–3470. https://doi.org/10.1002/etc.3903.
Saouter, E., Biganzoli, F., Pant, R., Sala, S., Versteeg, D., 2019a. Using REACH for the EU environmental footprint: building a useable ecotoxicity database, Part I. Integr. Environ. Assess. Manag. 15, 783–795. https://doi.org/10.1002/ieam.4168.
Saouter, E., Wolff, D., Biganzoli, F., Versteeg, D., 2019b. Comparing options for deriving chemical ecotoxicity hazard values for the European union environmental footprint, Part II. Integrated Environ. Assess. Manag. 15, 796–807. https://doi.org/10.1002/ieam.4169.
Schenker, U., MacLeod, M., Scheringer, M., Hungerbühler, K., 2005. Improving data quality for environmental fate models: a least-squares adjustment procedure for harmonizing physicochemical properties of organic compounds. Environ. Sci. Technol. 39, 8434–8441. https://doi.org/10.1021/es0502526.
Schenker, U., Scheringer, M., Sohn, M.D., Maddalena, R.L., McKone, T.E., Hungerbühler, K., 2009. Using information on uncertainty to improve environmental fate modeling: a case study on ddt. Environ. Sci. Technol. 43, 128–134. https://doi.org/10.1021/es801161x.
Scott, D.W., 2010. Scott's rule. Wiley Interdiscip. Rev. Comput. Stat. 2, 497–502. https://doi.org/10.1002/wics.103.
Slob, W., 1994. Uncertainty analysis in multiplicative models. Risk Anal. 14, 571–576. https://doi.org/10.1111/j.1539-6924.1994.tb00271.x.
Sobanska, M., Le Goff, F., 2014. IUCLID (international uniform chemical information database). In: Encyclopedia of Toxicology, third ed. https://doi.org/10.1016/B978-0-12-386454-3.00567-4
Sobanska, M.A., Cesnaitis, R., Sobanski, T., Versonnen, B., Bonnomet, V., Tarazona, J.V., De Coen, W., 2014. Analysis of the ecotoxicity data submitted within the framework of the REACH Regulation. Part 1. General overview and data availability for the first registration deadline. Sci. Total Environ. 470, 1225–1232. https://doi.org/10.1016/j.scitotenv.2013.10.074.
Stieger, G., Scheringer, M., Ng, C.A., Hungerbühler, K., 2014. Assessing the persistence, bioaccumulation potential and toxicity of brominated flame retardants: data availability and quality for 36 alternative brominated flame retardants. Chemosphere 116, 118–123. https://doi.org/10.1016/j.chemosphere.2014.01.083.
Stylianou, K.S., Fulgoni, V.L., Jolliet, O., 2021. Small targeted dietary changes can yield substantial gains for human and environmental health. Nat. Food 2, 616–627. https://doi.org/10.1038/s43016-021-00343-4.
Tarazona, J.V., Sobanska, M.A., Cesnaitis, R., Sobanski, T., Bonnomet, V., Versonnen, B., De Coen, W., 2014. Analysis of the ecotoxicity data submitted within the framework of the REACH Regulation. Part 2. Experimental aquatic toxicity assays. Sci. Total Environ. 472, 137–145. https://doi.org/10.1016/j.scitotenv.2013.10.073.
Tickner, J., Simon, R., Jacobs, M., Rudisill, C., Tanir, J., Heine, L., Spencer, P., Fantke, P., Malloy, T., Edwards, S., Zhou, X., 2019. Lessons from the 2018 international symposium on alternatives assessment: advances and reflections on practice and ongoing needs to build the field. Integrated Environ. Assess. Manag. 15, 909–916. https://doi.org/10.1002/ieam.4213.
Wambaugh, J.F., Bare, J.C., Carignan, C.C., Dionisio, K.L., Dodson, R.E., Jolliet, O., Liu, X., Meyer, D.E., Newton, S.R., Phillips, K.A., Price, P.S., Ring, C.L., Shin, H.M., Sobus, J.R., Tal, T., Ulrich, E.M., Vallero, D.A., Wetmore, B.A., Isaacs, K.K., 2019. New approach methodologies for exposure science. Curr. Opin. Toxicol. 15, 76–92. https://doi.org/10.1016/j.cotox.2019.07.001.
Weidema, B.P., Wesnæs, M.S., 1996. Data quality management for life cycle inventories-an example of using data quality indicators. J. Clean. Prod. 4, 167–174. https://doi.org/10.1016/S0959-6526(96)00043-1.
Wender, B.A., Prado, V., Fantke, P., Ravikumar, D., Seager, T.P., 2018. Sensitivity-based research prioritization through stochastic characterization modeling. Int. J. Life Cycle Assess. 23, 324–332. https://doi.org/10.1007/s11367-017-1322-y.
Wenger, Y., Li, D., Jolliet, O., 2012. Indoor intake fraction considering surface sorption of air organic compounds for life cycle assessment. Int. J. Life Cycle Assess. 17, 919–931. https://doi.org/10.1007/s11367-012-0420-0.
Williams, A.J., Grulke, C.M., Edwards, J., McEachran, A.D., Mansouri, K., Baker, N.C., Patlewicz, G., Shah, I., Wambaugh, J.F., Judson, R.S., Richard, A.M., 2017. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. J. Cheminf. 9, 1–27. https://doi.org/10.1186/s13321-017-0247-6.
Xiao, H., Li, N., Wania, F., 2004. Compilation, evaluation, and selection of physical-chemical property data for α-, β-, and γ-hexachlorocyclohexane. J. Chem. Eng. Data 49, 173–185. https://doi.org/10.1021/je034214i.