



LJMU Research Online

Achar, J, Firman, JW, Cronin, MTD and Öberg, G

A framework for categorizing sources of uncertainty in in silico toxicology methods: considerations for chemical toxicity predictions

<http://researchonline.ljmu.ac.uk/id/eprint/24784/>

Article

Citation (please note it is advisable to refer to the publisher's version if you intend to cite from this work)

Achar, J, Firman, JW, Cronin, MTD and Öberg, G (2024) A framework for categorizing sources of uncertainty in in silico toxicology methods: considerations for chemical toxicity predictions. Regulatory Toxicology and Pharmacology. 154. ISSN 0273-2300

LJMU has developed **LJMU Research Online** for users to access the research output of the University more effectively. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LJMU Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain.

The version presented here may differ from the published version or from the version of the record. Please see the repository URL above for details on accessing the published version and note that access may require a subscription.

For more information please contact researchonline@ljmu.ac.uk

<http://researchonline.ljmu.ac.uk/>



A framework for categorizing sources of uncertainty in *in silico* toxicology methods: Considerations for chemical toxicity predictions

Jerry Achar^{a,*}, James W. Firman^b, Mark T.D. Cronin^b, Gunilla Öberg^a

^a Institute for Resources Environment, and Sustainability, The University of British Columbia, 2202 Main Mall, BC, V6T 1Z4, Vancouver, Canada

^b School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, L3 3AF, Liverpool, UK

ARTICLE INFO

Handling Editor: Dr. Lesa Aylward

Keywords:

In silico methods
Uncertainty
Framework
QSAR
Risk assessment
QSAR assessment framework

ABSTRACT

Improving regulatory confidence and acceptance of *in silico* toxicology methods for chemical risk assessment requires assessment of associated uncertainties. Therefore, there is a need to identify and systematically categorize sources of uncertainty relevant to the methods and their predictions. In the present study, we analyzed studies that have characterized sources of uncertainty across commonly applied *in silico* toxicology methods. Our study reveals variations in the kind and number of uncertainty sources these studies cover. Additionally, the studies use different terminologies to describe similar sources of uncertainty; consequently, a majority of the sources considerably overlap. Building on an existing framework, we developed a new uncertainty categorization framework that systematically consolidates and categorizes the different uncertainty sources described in the analyzed studies. We then illustrate the importance of the developed framework through a case study involving QSAR prediction of the toxicity of five compounds, as well as compare it with the QSAR Assessment Framework (QAF). The framework can provide a structured (and potentially more transparent) understanding of where the uncertainties reside within *in silico* toxicology models and model predictions, thus promoting critical reflection on appropriate strategies to address the uncertainties.

1. Introduction

In silico toxicology methods play a central role in the risk assessment of chemicals as they are used to predict the biological activities of chemicals by drawing on the knowledge of chemical structures or physicochemical properties (Cronin and Madden, 2010). For the purpose of this paper, the term “*in silico* toxicology methods” is taken to refer to quantitative structure-activity relationship (QSAR) models, structural alerts, read-across and chemical category formation approaches that are based on any type of chemical descriptor or property. The basic tenet of *in silico* toxicology modeling is that the potential toxicity of a chemical in a biological system can be deduced from the chemical’s molecular structure/properties, where chemicals with similar structures/properties are assumed to have similar toxicological behavior (Cronin and Madden, 2010; Cronin et al., 2013; Enoch, 2010). These types of *in silico* methods are thus used to predict the properties, or activities, of data-poor chemicals by using knowledge of the biological activities induced by data-rich chemicals with similar structures/properties (Schultz et al., 2019).

Considerable research has been conducted to improve the predictive

accuracy of *in silico* toxicology methods, especially for regulatory purposes, through the characterization of the uncertainties associated with their predictions. Commonly cited sources of uncertainty include the quality of modeling data and inferences based on chemical structural similarity assumptions (Blackburn and Stuard, 2014; Parish et al., 2020; Schultz et al., 2019). Uncertainty is also inherent *in silico* models simply because they, like all models (including *in vivo* and *in vitro* tests), are surrogates of real systems. *In silico* toxicology models can consequently only approximate the potential harm posed by chemicals to a particular level of certainty. Generally, transparent analysis and communication of uncertainties of model-based quantitative assessments are considered part of good modeling practice (Benford et al., 2018). Peer-reviewed scientific publications rarely, however, include systematic and transparent accounting of associated uncertainties (Blackburn and Stuard, 2014; Cronin et al., 2019; Pham et al., 2019; Schultz et al., 2019).

Blackburn and Stuard (2014) stated that a lack of transparent communication of uncertainties hinders proper assessment of the strength and robustness of *in silico* models for toxicity predictions. It also potentially gives a false sense of confidence in the data applied, modeling process, and model prediction output. Indeed, the regulatory

* Corresponding author.

E-mail address: jerry.achar@ubc.ca (J. Achar).

<https://doi.org/10.1016/j.yrtph.2024.105737>

Received 9 April 2024; Received in revised form 26 October 2024; Accepted 11 November 2024

Available online 14 November 2024

0273-2300/© 2024 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

application of *in silico* toxicology methods would undeniably improve if uncertainties were more transparently communicated. Transparent communication of uncertainties is, however, not sufficient to gain regulatory acceptance. It is also necessary to systematically categorize the uncertainties (Achar et al., 2024d; Alexander-White et al., 2022; ECHA (European Chemical Agency), 2012). There is, therefore, a need to develop frameworks that can aid systematic categorization of uncertainties associated with *in silico* toxicology predictions, as this would provide an easier understanding of their sources within *in silico* toxicology prediction processes (Alexander-White et al., 2022).

A few studies have attempted to categorize sources of uncertainty in *in silico* toxicology methods – e.g., Benfenati et al. (2019), Blackburn and Stuard (2014), Cronin et al. (2019, 2022) and Pham et al. (2019), while others discuss uncertainties in the methods but do not explicitly categorize the uncertainty sources (e.g., Sahlin et al. (2011, 2013; 2014)). The studies that categorize sources of uncertainty, however, only focus on a limited number of sources of uncertainty within a particular method (e.g., QSAR, structural alerts/rule-based, or read-across). Consequently, none of the studies provide a general framework that covers sources of uncertainty across different methods as a means to provide a holistic picture of diverse sources of uncertainty in *in silico* toxicology methods while also facilitating communication of the uncertainty sources (ECHA, 2012; Kirchner et al., 2021). The lack of such a general framework may also result in a lack of harmonization of terminologies used to describe sources of uncertainty, thereby leading to poor communication of the sources among different stakeholders.

This investigation aimed to develop an uncertainty categorization framework that systematically categorizes general sources of uncertainty (GSU) across different *in silico* toxicology methods. This was achieved by reviewing peer-reviewed publications on *in silico* toxicology methods and verbatim recording the sources of uncertainty (VRSU) discussed in this literature. Drawing on general uncertainty concepts, we deduce GSU through iterative categorization of the VRSU (the process followed to develop the framework is shown in Fig. 1). Our assumption is that this framework can provide developers and users of *in silico* toxicology models a foundational understanding of where uncertainties reside within the broader *in silico* toxicology modeling contexts.

2. Identification and verbatim recording of sources of uncertainty (VRSU) in the literature

Peer-reviewed papers that discuss sources of uncertainty in *in silico* toxicology methods were identified through a search in the Web of Science using the following keywords and Booleans: (topic) uncertain* AND "in silico*" OR QSAR OR SAR OR read-across OR structural alerts AND chemical*. This led to the retrieval of 283 papers. These were skimmed (titles, abstracts, and, when needed, the entirety of the identified literature) to identify relevant papers based on the following criteria: (1) must be related to *in silico* toxicology predictions (and mention at least one of the methods), (2) include a discussion of uncertainty, and (3) make direct (explicit) reference to at least three sources of uncertainty. This resulted in the identification of 11 relevant publications (see Table S1). A content analysis was conducted of these publications through line-by-line reading (Sarah, 2018), identifying sources of uncertainty mentioned or discussed in these papers. We

illustrate the process used to identify sources of uncertainty by describing the analysis of Pham et al. (2019). This paper includes a section with the heading "Uncertainty in QSAR modeling", suggesting that uncertainties in QSAR modeling would be discussed in this section, which also was the case. We proceeded by verbatim recording each uncertainty source mentioned in this section, which included "choice of modeling algorithm and hyperparameter selection" and "model prediction reproducibility". The analysis of the 11 publications resulted in the identification of 87 sources of uncertainty, which were all recorded verbatim, hence the acronym VRSU ("verbatim recorded sources of uncertainty") (Table S1). In so doing, we note that other literature (e.g., Sahlin et al. (2011, 2013; 2014)) that merely discuss but do not categorize uncertainty sources were not included in Table S1.

As noted with asterisks in Table S1, all but six of the 87 VRSU were deemed irrelevant for the present paper, thus excluded. One of the excluded sources was the "acceptable level of uncertainty" mentioned by Schultz et al. (2019), which, while relevant in decision contexts, is beyond the scope of our paper. In addition, we excluded four VRSU that point primarily to variability in model systems: "error associated with biological data" (Cronin et al., 2019), "variability of biological data" (Schilter et al., 2014), "parametric variability" (e.g., the descriptors) and "observation error" (Benfenati et al., 2019), and "data variability" (Pham et al., 2019). Skinner et al. (2014a,b), ECHA (2012) and US EPA (2011) emphasize the need to treat uncertainty and variability separately. Whereas variability is stochastic in nature (thus irreducible), uncertainty is due to imperfection in knowledge about a model system, thus, may potentially be reduced by more knowledge. Similarly, potential areas of bias were not considered in this investigation. Here, bias refers to the possibility of introducing systematic errors in model predictions given the methodological criteria applied (Cronin et al., 2019). While we acknowledge that identifying areas of variability and bias in *in silico* model systems is also important, it is beyond the scope of this paper.

3. Categorizing VRSU and formulating GSU

In the present paper, we modify the framework developed by Belfield et al. (2021) for assessing the fitness-for-purpose of QSAR models to categorize the identified VRSU. The framework outlines 10 criteria (formalized as "higher-level assessment components") for evaluating QSAR models, which broadly focuses on model creation, characterization and application. There is considerable overlap between what ECHA (2012) refers to as "sources of uncertainty" and the components in this framework. As such, we conceptualized the 10 components in Belfield et al. (2021) as areas that generally characterize potential sources of uncertainty in *in silico* toxicology modeling (Table 1).

In our modification of the framework, we excluded two higher-level assessment components in Belfield et al. (2021) – Description and Usability – as they do not directly relate to areas of uncertainty in *in silico* toxicology modeling. As noted by Cronin et al. (2019), Description and Usability draw on experiences and barriers (e.g., software accessibility and intellectual property) to the practical usage of models; thus, while relevant in, for example, assessment of whether a model software is ethically developed and transparently documented, they are less relevant for characterizing uncertainties in the application of models for

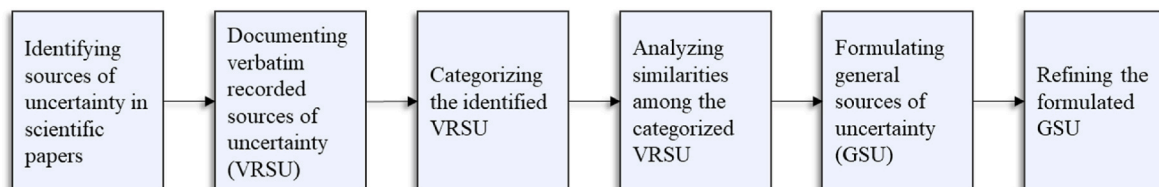


Fig. 1. A flow chart describing the steps undertaken to develop the framework that categorizes general sources of uncertainty in *in silico* models for toxicological data dap filling.

Table 1
In silico toxicology modeling phases, higher-level assessment components, and definition of the components of relevance for the present study.

Modeling phase	Higher-level assessment component	Definition of the higher-level assessment component
Model creation	Data	Quantity and quality of individual studies within the data set and the data set overall (e.g., homogeneity of the protocols) that was used for modeling
	Structure Similarity	Accuracy and/or quality of the reported chemical structures in the training (and, if applicable, test) set used for modeling Resemblance or commonality between chemical compounds, e.g., in terms of functional groups, toxicokinetic/toxicodynamic properties, and chemical structure
	Descriptors Modeling	Appropriate use and adequate definition of the descriptors used for modeling (including how and where sourced) The appropriateness and/or adequacy of the modeling approach for the endpoint with regard to complexity of the endpoint and potential use of the
Model characterization	Performance	Adequate statistical fit, predictivity and appropriate reporting
	Mechanisms	Definition and interpretation of the mechanistic significance of the model to allow for the definition of appropriate domains
	Toxicokinetics	Appropriate consideration of metabolism and toxicokinetics in the model
Model application	Applicability	The use of a model to provide data for similar prediction problem (e.g., inferring unknown values from trends in the known data)
	Relevance	Relevance of the model to its intended purpose and use

toxicity prediction. In another modification, as further discussed in Section 3.1.4, we added a new higher-level assessment component – “Similarity”. This was placed under the Model creation phase. As the framework we set out to develop includes characterizing uncertainty related to the use of *in silico* toxicology models, we also added “Applicability” as a higher-level assessment component, resulting in a total of 10 higher-level assessment components (Table 1). Cronin et al. (2019) argue that Applicability is relevant for characterizing uncertainty in the Model application phase, as it characterizes the potential use of a model to provide data for similar prediction problems. An example is the potential application of a model to predict the effect of similar target chemicals or inferring unknown values from trends in the known data (Cronin et al., 2019).

We started the categorization process by reviewing the 81 VRSU (Table S1) in light of the descriptions of the higher-level assessment components in Table 1 and the ways in which the VRSU are discussed in the analyzed texts. The VRSU were placed under the higher-level assessment component deemed most suitable (4th column in Table 2). These VRSU were then analyzed for similarities and then used to formulate the GSU shown in Table 2 (column 3). Subsections 3.1, 3.2, and 3.3 below describe the reasoning that led to the categorization of the VRSU and the formulation of the GSU.

3.1. The model creation phase

3.1.1. Data

The higher-level assessment component “Data” is in our framework described as “Quantity and quality of individual studies within the data set and the data set overall (e.g., homogeneity of the protocols) that was used for modeling” (Table 1). This description is modified from Belfield et al. (2021), who did not include quantity aspects in their description of Data. The need to consider both quality and quantity as inherent characteristics of data has been emphasized (Fu et al., 2011; Nendza et al., 2010; Stausberg et al., 2023), hence the inclusion of data quantity-related VRSU here, as further developed below.

The categorization process led us to place 37 VRSU under this component. Examples of quality-related VRSU are “quality of the data considered” (Blackburn and Stuard, 2014), “quality and robustness of the source or analogue data” (Schultz et al., 2015), and “quality of data”, “relevance of data for the endpoint of interest to its intended use”, and “completeness of data” (each from Cronin et al., 2019). Examples of quantity-related VRSU that were added include: “quantity of the data considered”, “number of analogues contributing data” and “number of the chemical analogues identified” (the full list of the VRSU categorized under Data is shown in Table 2).

The initial analysis of the 37 VRSU led us to formulate 17 GSU, which were later consolidated into six GSU: Data quantity, Data balance, Data relevance, Data reliability, Data accuracy, and Data validity. To

illustrate: the formulation of the GSU “Number of data” was based on the merging of four VRSU, each of which relates to the amount of data and contains the term “number” when mentioning the amount of data – i.e., “number of analogues contributing data” (Blackburn and Stuard, 2014; Schultz et al., 2019), “number of the chemical analogues identified” (Schilter et al., 2014), and “number of source chemicals” (Schultz et al., 2015). The formulated GSU “Number of data” was subsequently merged with the GSU “Data coverage”, which Cronin et al. (2022) define as the proportion of hits in alerts. As “Number of data” and “Coverage” both refer to the quantity of data for modeling, we formulated “Data quantity” as the common GSU that covers them (Fig. 2). The choice of the term “data quantity”, was based on its common use in describing uncertainty related to the amount of data (WHO/IPCS, 2008).

We also note that the GSU “Data quality”, as used in six papers (Table 2), refers to characteristics of data that make them fit for an intended use. These characteristics are distinctly described in the other data quality-specific GSU: Data relevance, Data suitability, Data completeness, Database deficiency, Data reliability, Data consistency, Data robustness, Data accuracy, and Data validity. As such, we excluded Data quality as a separate GSU.

Our analysis revealed an overlap in the description of three initial GSU: Data balance, Data distribution, and Data homogeneity. Pham et al. (2019), (citing He and Garcia (2009)), describe Data distribution (i.e., “distribution of the training data set”) as the characteristic of data that reflects class distribution of balanced dataset, and Data balance (i.e., “balance of the training data set”) as the distributive characteristics of data for categorical (toxic/non-toxic) endpoints. These descriptions are similar to how Cronin et al. (2019) describe Data balance (i.e., “data balance”). In another instance, Cronin et al. (2019) describe Data homogeneity as the distributive characteristics of datasets across the chemical space of the training and test sets for continuous (potency) data. In the literature, uncertainties related to data balance have been broadly described to span from potential inadequacies in partitioning data between two classes to considerations of distributive characteristics of continuous data (He and Garcia, 2009). This suggests Data distribution and Data homogeneity (as described in the analyzed studies) can be subsumed under the GSU “Data balance”.

Three data quality-related GSU – Data suitability (“suitability of analogues”; Blackburn and Stuard, 2014; Schilter et al., 2014), Data completeness (“completeness of the data set”; Cronin et al., 2019 and “completeness of the argument provided [for data quality]”; Schultz et al., 2019), and Database deficiency (“database deficiency”; Wang et al., 2012) – were subsumed under the GSU Data relevance. This is because both Data suitability and Data completeness are related to the appropriateness of chemical or biological data for predicting a toxicological endpoint (Blackburn and Stuard, 2014; Cronin et al., 2019; Schultz et al., 2019; Schilter et al., 2014) – Table 2. Notably, these descriptions align with descriptions of Data relevance by Cronin et al.

Table 2

Categories of the 81 VRSU and the formulated general sources of uncertainty (GSU, column 3). The non-bolded GSU are the tentative GSU that were subsumed under the refined GSU (bolded). Publication numbers are provided in [Table S1](#).

Modeling phase	Higher-level assessment component	General sources of uncertainty (GSU)	Verbatim recorded sources of uncertainty (VRSU)	Publication number	
Model creation	Data	Data quantity	Quantity of the data considered	1	
			Number of data	1	
		Data size	Number of analogues contributing data	3	
			Number of the chemical analogues identified	5	
		Data coverage	Number of source chemicals	8	
			Size of training data set	4	
		Data balance	Coverage [of structural alert]	7	
			Balance of the training data set	4	
		Data distribution	Data balance	6	
			Distribution of the training data set	4	
		Data homogeneity	Homogeneity of the chemical space of the training and test sets.	6	
			Data relevance	Relevance of data for the endpoint of interest	6
		Data suitability	Relevance of data	10	
			Suitability of analogues	1	
		Data completeness	Suitability of the chemical analogues identified	5	
			Completeness of the argument provided [for data quality]	3	
		Database deficiency	Completeness of the data set	6	
			Database deficiencies (e.g., lack sensitive endpoint or toxicity information)	2	
		Data reliability	Reliability of data	10	
			Data consistency	Consistency of the data set	6
		Data robustness	Consistency of data	9	
			Strength or robustness of the supporting data sets	3	
		Data accuracy	Robustness of the source or analogue data	8	
			Robustness of the supporting data sets	9	
		Data validity	Accuracy of data	10	
			[computed/not experimentally measured] Parameters used to construct the model	11	
		Data quality ⁷	Validity of data	10	
			Quality of the data	1	
		Structure	Chemical structure	Quality of the apical endpoint data	3
				Quality of data used to build model	5
		Similarity	Chemical similarity	Toxicological information found for the analogues	5
				Quality of data	6
		Descriptors	Descriptor relevance	Quality of the source or analogue data	8
				Quality of data	9
		Descriptor concordance	Descriptor concordance	Structure and its representation	8
				Structural description	7
		Modeling	Model structure	Structural similarity to target	1
				Similarity in chemistry	3
		Performance	Model performance	Toxicokinetic similarity	3
				Toxicodynamic similarity	3
		Mechanisms	Mechanistic plausibility	Similarity justification	4
				Definition and demonstration of similarity	8
		Toxicokinetics	Metabolic domain	Choice of molecular descriptors	4
				Calculated/experimentally measured properties and descriptors	6
		Activity/potency	Activity/potency	Property domain	7
				Modeling algorithm and hyperparameter	4
		Activity/potency evidence	Activity/potency evidence	Numerical errors and/or numerical approximations	11
Model bias	11				
Weight-of-Evidence	Supporting evidence	The potency of the analogues for those [toxic] effects	1		
		Nature and severity of the identified toxic effects	1		
Corroborating evidence	Corroborating evidence	Prediction of complex endpoints such as chronic toxicity	5		
		Species specificity	7		
Weight-of-evidence supporting the prediction	Weight-of-evidence supporting the prediction	Toxicity or relationship to adversity	7		
		Weight-of-Evidence	3		
Statistical performance	Statistical performance	Supporting evidence	7		
		Corroborating evidence	7		
[predictive] Performance	[predictive] Performance	Weight-of-evidence supporting the prediction	9		
		Model performance	5		
Mechanistic causality	Mechanistic plausibility	Reproducibility of model and model prediction	6		
		Adequacy of the model to make a prediction for the stated purpose	6		
Mechanistic relevance and interpretability	Mechanistic relevance	Statistical performance	6		
		Mechanistic relevance	7		
Mechanistic relevance	Mechanistic relevance	[predictive] Performance	7		
		Mechanistic plausibility	3, 8, 9		
Metabolic domain	Metabolic domain	Mechanistic causality	7		
		Mechanistic relevance and interpretability	6		
Metabolic domain	Metabolic domain	Mechanistic relevance	8		
		Metabolic domain	7		

(continued on next page)

Table 2 (continued)

Modeling phase	Higher-level assessment component	General sources of uncertainty (GSU)	Verbatim recorded sources of uncertainty (VRSU)	Publication number
		Coverage of ADME activity	Adequate coverage of Absorption, Distribution, Metabolism and Excretion effects	6
	Applicability	Applicability domain	Applicability domain	4
			Applicability domain	5
			Applicability domain	6
			Applicability domain	8
	Relevance	Extrapolation	Extrapolations (interspecies (animal-to-human))	2
			Extrapolations intraspecies (susceptible human subpopulation)	2
			Extrapolations (subchronic-to-chronic)	2
			Extrapolations (LOAEL-to-NOAEL)	2
			Extrapolation of the toxicity of the substance of interest based on data on analogues	5
		Model relevance	Relevance [of the QSAR] to the prediction or assessment goal	6
			Purpose or potential use of the structure alert	7

LOAEL = lowest observed adverse effect level; NOAEL = no observed adverse effect level; QSAR = quantitative structure-activity relationship.

^a Quality = tentatively identified GSU excluded in the final iteration of the proposed framework, as the VRSU that initially were placed under this category were subsumed under other categories, as further explained in section 3.1.

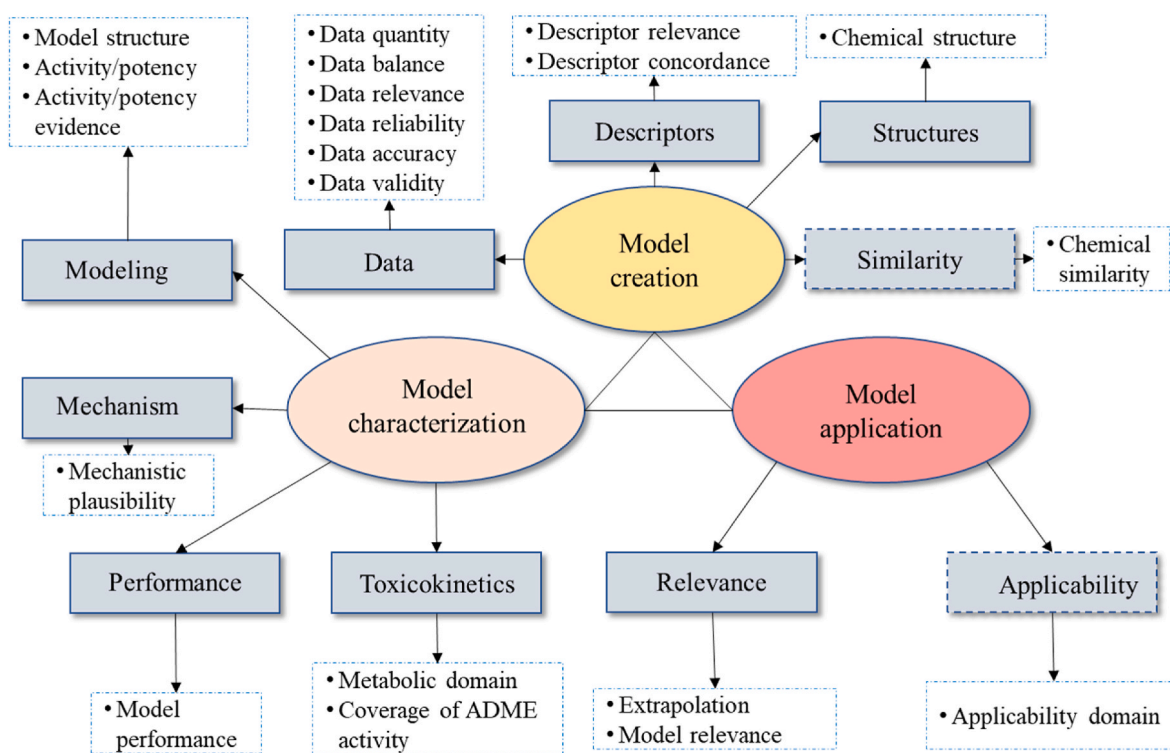


Fig. 2. The refined GSU (bulleted in the rectangles) resulting from the analysis and iterative categorization of the VRSU. The descriptions of the GSU are provided in Table S2. The grey rectangles indicate the higher-level assessment components under which the GSU are categorized, and the grey-dotted rectangles are the newly proposed higher-level assessment components. The components are, in turn, connected to one of the Modeling phases (shown in the ovals).

(2019) and Madden et al. (2020), who refer to it as the meaningfulness of data – i.e., the extent to which data are considered useful for a particular prediction context – be it endpoint, route of exposure, etc. We also note that Wang et al. (2012) use Database deficiency in reference to the incompleteness of data, which also aligns with the description of Data relevance. In other words, each of these three tentative GSU (Data suitability, Data completeness, and Database deficiency) refers to the extent to which data incorporates essential information for a given use, thus it is reasonable to subsume them under the GSU “Data relevance”.

We decided to subsume Data consistency (“consistency of the data”; Cronin et al., 2019; Pestana et al., 2021) and Data robustness (“strength or robustness of the supporting data sets”; Pestana et al., 2021; Schultz et al., 2019) under Data reliability (Madden et al., 2020), as the distinction between them is unclear and the definition of Data reliability

seems to cover the elements described in them – i.e., Data reliability refers to the comparability and reproducibility of data obtained from different laboratories under consistent test protocols or toxicity endpoints or biomarkers (Madden et al., 2020). Pestana et al. (2021) do not provide an explicit definition of Data consistency relating to read-across, but describe it as the uniformity of toxicity information in chemical datasets – one of the examples provided by the authors is “consistency in the *in vivo* effects and potency data”. Similarly, Cronin et al. (2019) describe Data consistency as the uniformity of datasets or data reproducibility between different tests. These descriptions are similar to Data robustness, which Schultz et al. (2019) describe as data consistency based on how extensive the data are measured or observed across source and target chemical categories. Given these similarities, we decided to subsume Data consistency and Data robustness under a common GSU

“Data reliability”.

The GSU Data accuracy is described by Madden et al. (2020) as the extent to which measured data deviates from its true value. The same authors relate Data validity to the acceptability of the methods used to generate modeling data relative to set guidelines or consideration of whether the methods measure what they are intended to measure. As they describe, uncertainty related to Data validity might impact data reproducibility if such guidelines are not followed. However, Data validity does not have to be always interlinked to data reproducibility, as invalid data generated using non-standardized guidelines may still be reproduced using similar non-standardized guidelines; consequently, we retained both “Data accuracy” and “Data validity” as distinct GSU (Fig. 2).

3.1.2. Structure

The higher-level assessment component “Structure” is, in this framework, described as the “Accuracy and/or quality of the reported chemical structures in the training (and, if applicable, test) set used for modeling” (Table 1). We placed two VRSU under this component (Table 2): “structure and its representation” (Schilter et al., 2014), and “structural description” (Cronin et al., 2022). Our analysis of these VRSU led us to conclude that they are similar, as they both encompass questions related to chemical structure – i.e., whether a structure (or presence of a substructure) is known and clearly described with appropriate identifiers and whether this representation is useable or suitable for a defined modeling task or for calculating, for example, descriptors. This, therefore, implies that these VRSU can be subsumed under the GSU “Chemical structure”. Taken more broadly, uncertainties within this GSU thus include the accuracy in the definition of structure, the clarity in the description of structural representation, and the measure of the fitness or relevance of the structure for a particular use.

3.1.3. Similarity

We define the higher-level assessment component “Similarity” as “Resemblance or commonality between chemical compounds, e.g., in terms of functional groups, toxicokinetic/toxicodynamic properties, and chemical structure” (Table 1). We placed six VRSU (Table 2) within: “structural similarity [of analogues] to target” (Blackburn and Stuard, 2014), “similarity in chemistry”, toxicokinetic similarity, and toxicodynamic similarity (Schultz et al., 2019), “similarity justification” (Pham et al., 2019), “definition and demonstration of similarity” (Schultz et al., 2015). Each addresses issues related to chemical similarity, which, in the context of read-across modeling, Drake et al. (2023) describe as the measure of commonality/similarity between chemical compounds in terms of their structural and physicochemical properties, toxicokinetic/toxicodynamic properties, mechanism of action, etc. The major application of this concept in *in silico* toxicology includes derivation of structure-activity relations, grouping of compounds with similar activities, and providing justification of read across (Blackburn and Stuard, 2014; Schilter et al., 2014; Schultz et al., 2015). This implies that, while the component “Similarity” depends on components such as “Structure”, it is distinct – i.e., whereas Structure relates similarity between compounds, Structure relates to chemical characteristics in the form of components like atoms and the bonds between them. As the six VRSU relate to chemical similarity, we decided to formulate “Chemical similarity” as the umbrella GSU that covers them.

3.1.4. Descriptors

As seen in Table 1, the higher-level assessment component “Descriptors” is here defined as the “Appropriate use and adequate definition of the descriptors used for modeling (including how and where sourced)”. Three VRSU were placed under this component: “choice of molecular descriptors” (Pham et al., 2019), “calculated/experimentally measured properties and descriptors” (Cronin et al., 2019), and “property domain” (Cronin et al., 2022).

In the broader *in silico* modeling literature (e.g., Chandrasekaran

et al., 2018; Cronin et al., 2013; US EPA, 2016), a descriptor is typically defined as providing a quantitative representation of the physicochemical or structural properties of a chemical, e.g., descriptors derived from molecular or atomic properties may reflect its physicochemical, topological, and surface properties. We interpret this definition as an elaboration of the description of Descriptors in Table 1, which then implies that a descriptor represents a logical transformation of chemical information encoded within its physicochemical properties. In our categorization (Table 2), we, therefore, interpreted the component Descriptors in a broad sense to not only explicitly encompass physicochemical descriptors but also include the physicochemical properties from which the descriptors are obtained. In this case, uncertainty originates from a lack of relevant physicochemical descriptors, as this could translate to an inaccurate interpretation of the properties or an inability to accurately calculate physicochemical descriptor values (Ball et al., 2016; Cronin et al., 2019).

Analysis of the three VRSU referenced above led us to formulate two GSU: “Descriptor relevance” and “Descriptor concordance”. Descriptor relevance incorporates solely the VRSU “choice of molecular descriptors” (e.g., Log P) generated from physicochemical properties (Pham et al., 2019). Pham et al. (2019) note that understanding this uncertainty source involves asking whether the physicochemical properties in question are relevant to predict the descriptors (hence “Descriptor relevance”). In this case, a lack of relevant physicochemical descriptors could translate to an inaccurate interpretation of the properties or an inability to accurately calculate physicochemical descriptors as well as inconsistent descriptor values (Ball et al., 2016; Cronin et al., 2019).

The remaining two VRSU: “calculated/experimentally measured properties and descriptors” (Cronin et al., 2019) and “property domain” (Cronin et al., 2022), were used to formulate the GSU “Descriptor concordance”. Drawing on the ways in which these are discussed in the analyzed literature (Cronin et al., 2022), it is clear that the concept of Descriptor concordance differs from Descriptor relevance. That is, Descriptor concordance provides a quantitative or qualitative description of the degree of agreement between descriptors and, for example, the toxicokinetic or toxicodynamic properties of a chemical (Cronin et al., 2019, 2022). Thus, it can be understood as a measure that demonstrates the extent of correlation between the descriptor and a variable, Y. In contrast, Descriptor relevance relates to the capacity of the descriptors to provide insight into what a model intends to predict – i.e., relevance characterizes quality dimensions like completeness and appropriateness of the descriptors.

3.2. The model characterization phase

3.2.1. Modeling

The higher-level assessment component “Modeling” is here described as the “Appropriateness and/or adequacy of the modeling approach for the endpoint with regard to the complexity of the endpoint and potential use of the model” (Table 1). We placed 11 VRSU in this category – for example, “modeling algorithm and hyperparameter” (Pham et al., 2019), “species specificity” (Cronin et al., 2022), and “prediction of complex endpoints such as chronic toxicity” (Schilter et al., 2014) (see Table 2 for the full list).

Cronin et al. (2019) elaborate on the definition of Modeling by noting that an appropriate modeling approach is one which can be gauged not only on its ability to deal with the complexity of data but also on its ability to predict activity or the toxic effects of chemicals of interest, either in simple or complex scenarios. Here, the degree of confidence in the predicted activity/potency is dependent upon the available supporting evidence (Pestana et al., 2021) or the adequacy of the modeling approach (e.g., in terms of modeling parameters and model algorithms) to predict an activity/potency (Pham et al., 2019). Taken more broadly, the component Modeling, therefore, encompasses characterizing the structure of a model (e.g., model algorithms and

parameters), prediction of activity or potency of chemicals, and consideration of the evidence that supports such predictions.

Three of the aforementioned 11 VRSU: “modeling algorithm and hyperparameter” (Pham et al. (2019), alongside “numerical errors and/or numerical approximations” and “model bias” (Benfenati et al., 2019) relate to uncertainties embedded in the model structure. Walker et al. (2003) associate uncertainty in model structure with the appropriateness or accuracy of model algorithms, mathematical formulations and parameters, etc., for particular predictions. Following this description, we decided to use “Model structure” as the umbrella GSU to group these three VRSU.

Similarity was noted among five other VRSU: “the potency of the analogues for those [toxic] effects”, “nature and severity of the identified toxic effects” (Blackburn and Stuard, 2014), “toxicity or relationship to adversity”, “species specificity” (Cronin et al., 2022), and “prediction of complex endpoints such as chronic toxicity” (Schilter et al., 2014). Uncertainties within the first four VRSU are described in the context of QSAR, read-across, and structural alerts to relate to the definition, establishing the association, or modeling the relationship between a chemical (or an alert) and particular toxicological activity or effects they elicit. Similarly, Schilter et al. (2014) relate the last VRSU to the reasonable predictions of toxicity of chemicals for complex endpoints such as chronic toxicity. Overall, our analysis led us to conclude that each of these five VRSU relates to the ability to model toxicological activities or potency of chemicals (Table 2). Considering the similarities, we thus formulated “Activity/potency”, as the umbrella GSU term for these five VRSU (Table 2).

Finally, we noted similarity among the remaining three VRSU – “corroborating evidence” (Cronin et al., 2022), “supporting evidence” (Cronin et al., 2022), and “weight-of-evidence supporting the prediction” (Pestana et al., 2021; Schultz et al., 2019). The discussion of these VRSU in the analyzed papers led us to conclude that the authors use the concept of “evidence” uniformly in reference to evidence of the toxic activity or potency of chemicals. That is, the availability of toxicological information from approaches such as *in vitro* assays, to support conclusion on activity/potency predicted by *in silico* models. Here, similar to Pestana et al. (2021), we argue that although evidence of activity/potency is closely related to the earlier formulated GSU “Activity/potency”, it is secondary to it and thus can be treated as a separate class. For example, while Activity/potency refers to the ability of a chemical to cause harm, evidence of activity/potency instead pertains to the body of information that supports whether or not toxicity is elicited and the extent of it. As seen in Table 2, therefore, we used these three VRSU to formulate the GSU “Activity/potency evidence”.

3.2.2. Performance

The higher-level assessment component “Performance” is here defined as “Adequate statistical fit, predictivity and appropriate reporting”. We placed five VRSU in this category: “model performance” (Schilter et al., 2014), “reproducibility of model and model prediction” (Cronin et al., 2019), “adequacy of the model to make a prediction for the stated purpose” (Cronin et al., 2019), “statistical performance” (Cronin et al., 2019), and “[predictive] performance” (Cronin et al., 2022). Our analysis led us to conclude that they are similar on the basis that they relate to the concept of “model performance”, which broadly refers to the measure of model predictivity of external dataset (via external validation) or of the same dataset used for model development (via internal validation), or estimation of statistical fit in the context of regression models in which a measure of overfitting in a model or statistical significance of model predictions are considered (Cronin et al., 2019; Schilter et al., 2014). Consequently, we formulated “Model performance” as the umbrella GSU for these VRSU.

3.2.3. Mechanisms

The higher-level assessment component “Mechanisms” is here defined as the “Definition and interpretation of the mechanistic

significance of the model to allow for the definition of appropriate domains”. Four VRSU mentioned from six studies were placed under this component: “mechanistic plausibility” (Pestana et al., 2021; Schultz et al., 2015, 2019), “mechanistic causality” (Cronin et al., 2022), “mechanistic relevance and interpretability” (Schultz et al., 2015) and “mechanistic relevance” (Cronin et al., 2019). Each relates to the mechanistic characterization of the effects of chemicals in biological systems (Table 2), which aligns with the description of Mechanism (Table 1).

Initial analysis led us to formulate two GSU – “Mechanistic plausibility” and “Mechanistic relevance”. The VRSU “mechanistic plausibility” by Pestana et al. (2021), Schultz et al. (2015), and Schultz et al. (2019) is based on the concept of adverse outcome pathway (AOP), where uncertainty within it is associated with the understanding of the toxic causal pathways of chemicals, involving the identification of molecular initiating events/key events causally linked to a target endpoint. Similarly, our analysis of Cronin et al. (2022), who use “mechanistic causality” in the context of structural alerts, describe this VRSU as the mechanism of action that underpins interactions of the functional group represented by the structural alert with physiological or biochemical processes in an AOP system. This led us to conclude that, as with mechanistic plausibility, uncertainty related to mechanistic causality concerns the understanding of causality as strengthened by consistency with sources or experimental data that demonstrate plausible biological or chemical reaction mechanisms. Given the similarity, we settled on using Mechanistic plausibility for the GSU, as it is used by OECD (OECD, 2019) to characterize uncertainty due to, for example, incomplete understanding of the mechanism of action or adverse outcome pathway of chemical compounds.

A second GSU, “Mechanistic relevance”, was formulated from the VRSU – “mechanistic relevance and interpretability” and “mechanistic relevance”, both of which explain the potential relevance of the causative or putative mechanism of actions of chemicals in biological systems (Table 2). However, further analysis revealed that, although described using different terminology, Mechanistic relevance also relates to Mechanistic plausibility. As seen in Table 2, uncertainty within Mechanistic relevance concerns knowledge gaps or unknowns in the understanding of causative or putative explanations of the mechanism of action of chemicals in biological systems with regard to AOPs. This suggests that both Mechanistic plausibility and Mechanistic relevance concern the understanding of causative or putative explanations of the mechanism of action of chemicals; as such, Mechanistic relevance can be subsumed under Mechanistic plausibility. Therefore, in our framework (Fig. 2), the GSU “Mechanistic plausibility” not only includes the consideration of the causative or putative mechanism of action of chemicals but also the relevance of the characterized mechanisms by drawing on the concept of AOP, including any measurable change at molecular level (molecular initiating event) or key event in biological system (see Table S2 for the description). Lastly, we note that while Mechanistic relevance, as used in the analyzed studies, warrants subsuming it under Mechanistic plausibility, this term could also be distinctly used to explain the biological relevance of a pathway to an endpoint/the test system or the relevance of the pathway to a known toxicant (Hartung et al., 2013). A detailed analysis of the difference between these two concepts was not explored in this study; thus, it remains for future studies to explore it.

3.2.4. Toxicokinetics

The higher-level assessment component “Toxicokinetics” is defined in the framework as “Appropriate consideration of metabolism and toxicokinetics in the model” (Table 1). Two of the VRSU align with this definition: “metabolic domain” (Cronin et al., 2022) and “adequate coverage of ADME effects” of metabolites (Cronin et al., 2019). Similar to Toxicokinetics, each considers the production of chemical metabolites as part of interaction with biological systems and the potential effects that result from it.

The VRSU “metabolic domain” is limited to the knowledge of whether or not the production of metabolites is part of the process of chemical interaction with biological systems or chemical reactivity (Cronin et al., 2022). Here, uncertainty could concern whether the metabolites are known or unambiguously stated – for example, in the oxidation of phenols to quinone, where uncertainty arises if quinone production is ambiguously/poorly defined or not stated at all (Cronin et al., 2022). In contrast, our analysis of the second VRSU: “adequate coverage of ADME effects” (Cronin et al., 2019), led us to conclude that this uncertainty occurs if the production of a metabolite (e.g., quinone) is known and clearly stated, but its potential activities are poorly understood, not considered, or the activities are only assumed without supporting evidence. Given the clear distinction between these two VRSU, we decided to formulate two GSU from these two VRSU: “Metabolic domain” from “metabolic domain” and “Coverage of ADME activity” from “adequate coverage of ADME effects” (Fig. 2). Here, we take Coverage of ADME activity more broadly to consider the potency, exposure, interaction with biological systems, and/or toxicity of chemical metabolites (Achar et al., 2020a, 2020b; Cronin et al., 2019).

3.3. The model application phase

3.3.1. Applicability

Drawing on Cronin et al. (2019), we here define the higher-level assessment component “Applicability” as “Use of a model to provide data for similar prediction problems (e.g., inferring unknown values from trends in the known data)”. We concluded that one of the VRSU that did not easily fit under any of the higher-level components in the original framework by Belfield et al. (2021) fit instead here: “applicability domain” mentioned in four papers (Cronin et al., 2019; Pham et al., 2019; Schilter et al., 2014; Schultz et al., 2015). These authors discuss applicability domain in reference to the adequacy of chemical space or category to predict effects of similar chemicals in a specified model prediction context (Table 2). This means that applicability domain is established prior to model application and can thus be assumed to be intrinsic to a model. Given the common term (i.e., “applicability domain”) used in these studies, we decided to formulate “Applicability domain” as the umbrella for this VRSU.

3.3.2. Relevance

“Relevance” is defined in this context as “Relevance of the model to its intended purpose and use” (Table 1). We placed seven VRSU under this component: “extrapolations (interspecies (animal-to-human), “extrapolations intraspecies (susceptible human subpopulation)”, “extrapolations (subchronic-to-chronic)”, “extrapolations (LOAEL-to-NOAEL)” mentioned by Wang et al. (2012), “extrapolation of the toxicity of the substance of interest based on data on analogues” mentioned by Schilter et al. (2014), “relevance [of QSAR] to the prediction or assessment” (Cronin et al., 2019), and “purpose [or potential use of the structure alert]” (Cronin et al., 2022), on account of our analysis suggesting that each points to the transferability of a model or model prediction towards a different context.

Our analysis revealed that the first five VRSU listed under this component relate to the concept “extrapolation” – i.e., making predictions beyond the range of the observed/known data (e.g., LOAEL data) in attempts to estimate or infer unknown properties (e.g., NOAEL) (Wang et al., 2012). Here, uncertainty arises, for example, in the use of uncertainty factors to cater for species differences in toxicological effect or where a read-across extrapolation is deemed inaccurate. As such, we used these VRSU to formulate the GSU “Extrapolation”.

The last two VRSU both describe uncertainties arising from the relevance of a model to its intended use (e.g., regulatory application). For example, while Cronin et al. (2019) discuss “relevance [of QSAR] to the prediction or assessment” as characterizing the relevance of a modeling approach for a specified endpoint or an intended use (e.g., regulatory toxicity assessment), Cronin et al. (2022) describe “purpose

[or potential use of the structure alert]” in terms of potential use (e.g., with respect to product development and regulatory applications). As such, we formulated the GSU “Model relevance” through a combination of these two VRSU. Here, Model relevance is distinct from Extrapolation – while Model relevance points to the uncertainties residing within transferability of a model or model prediction to a different prediction context (e.g., regulatory application), uncertainties within Extrapolation are closely related to making inference to data outside the range of the available data.

4. Application of the framework to prioritize areas of uncertainty

This study developed a framework (Fig. 2) to aid systematic categorization of sources of uncertainty across *in silico* toxicology methods. To evaluate its application as a tool for mapping out and prioritizing areas to consider for uncertainty during model prediction interpretation, uncertainty analysis, or data gap-filling exercises in specified prediction context(s), a case study is here used. This case study is used for illustrative purposes only and is targeted towards a simple prediction problem.

4.1. Case study

The overarching aim of the study was to evaluate the performance of Toxicity Estimation Software Tool (TEST; v5.1.2) (US EPA, 2015) in the prediction of oral rat LD₅₀ (the dose that causes death of 50% of test samples) – this kind of evaluation is useful when considering to apply a model for safety evaluation of chemicals, especially when *in vivo* data are lacking or limited (Graham et al., 2021). The performance evaluation was based on the agreement of TEST predicted LD₅₀-based Globally Harmonized System (GHS) categories with the corresponding experimental LD₅₀-based GHS categories. The GHS classification categories are presented in Table S3 (United Nations, 2021). The choice of the GHS criteria was based on its ability to provide harmonized system for classifying chemicals with respect to their degree of health concerns (expressed in LD₅₀ mg/kg bodyweight), as well as the ability to facilitate communication of chemical hazards via safety data sheets and labeling requirements (United Nations, 2021).

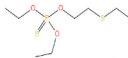
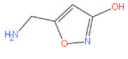
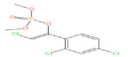
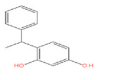
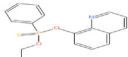
For illustrative purposes only, five organic compounds (Table 3) with oral rat experimental LD₅₀ data were used (the data is available in Firman et al. (2022)); the experimental data allows for comparison with the predicted data. These compounds were deliberately selected for this illustration as they have different experimental LD₅₀ values, which we anticipated to be useful in demonstrating model under- and over-prediction scenarios. TEST was selected for this illustration given its open-source accessibility; although it is acknowledged that similar open-access tools are available. Only predictions from TEST Consensus method (average of the Hierarchical and Nearest-neighbor model predictions) were used as they are considered more reliable than predictions from the individual methods (US EPA, 2015). The chemical CASRN identifiers were used as input and TEST prediction options were set as: endpoint – oral rat LD₅₀, method – consensus, and fragment constrain – relaxed. The predicted oral rat LD₅₀ data are shown in Table 3.

4.2. Identification of relevant GSU from the case study

We proposed a checklist to support the identification and justification of GSU deemed relevant for inclusion from the case study (see Table 4). The left-hand side column of the checklists is the GSU from the framework (Fig. 2). While not all the 20 GSU may be considered relevant in a prediction context, we recommend that all of them should be included. The right-hand side column contains spaces for justifying why a GSU is selected.

Taken together, the case study presented here, and the checklist (Table 4) indicate that the GSU within the framework can be used to

Table 3Information about the five compounds used for the illustration and their experimental and model-predicted LD₅₀ data and LD₅₀-based GHS categories.

Compound	CASRN	Structure	Experimental data		TEST Consensus data	
			Rat (mg/kg)	GHS category	Rat (mg/kg)	GHS category
Demeton-O	298-03-3		7.5	2	6.1	2
Muscimol	2763-96-4		45	2	146	2
Dimethylvinphos	2274-67-1		98	3	759.67	4
4-(1-Phenylethyl)benzene-1,3-diol	85-27-8		500	4	4324	5
Quintofos	1776-83-6		150	3	24	2

map out areas of concern for uncertainty. Notably, the checklist makes it easier to: delineate and clarify which GSU are embedded within a model or prediction (based on the areas of concern for uncertainty), and document in a structured format the rationale for including the GSU – this makes it easy to interpret the rationale and further facilitate subsequent review by others. The checklist thus offers guidance to modelers and other stakeholders seeking to make informed decisions about which area of uncertainty to consider for uncertainty analysis, incorporate in the interpretation of prediction results, and/or prioritize for data gap filling.

As shown in Table 4, not all the 20 GSU may be relevant for inclusion in a study – i.e., relevance of a GSU depends on a study context; however, whether selected or not, we argue that all the 20 GSU should remain in the checklist, as this will not only reduce the risk of modelers overlooking potentially important GSU in a study but also help during a review process the risk of not selecting specific GSU. Overall, we note that even though it may be tempting for a modeler to, for example, analyze uncertainty in a study without a systematic and an explicit indication and subsequently justification of relevant GSU (as in the checklist – Table 4), similar to Achar et al. (2024c, 2024d), Jones and Falloon (2009) and Przybylak et al. (2012), we argue that this might make it unclear whether the estimated uncertainty truly reflect relevant areas within *in silico* modeling known for potential uncertainties or whether these areas are truly justified for inclusion in the analysis.

4.3. Consideration of the framework within the OECD's QSAR Assessment Framework

The framework developed in this study addresses areas of concern for uncertainty in different contexts of *in silico* toxicology modeling. Indeed, it is anticipated that the framework will be a valuable reference tool in regulatory decision-making processes as it aligns with the principles in the OECD's proposed (QSAR Assessment Framework (QAF), as well as extends the discussions within the principles (OECD, 2023; Gissi et al., 2024)). QAF is based on the 2007 OECD principles for the validation of QSARs (OECD, 2007). The QAF provides four principles to guide the assessment of QSAR results from multiple predictions with the goal of supporting regulatory decision-making: (1) the model input(s) should be correct, (2) the substance should be within the applicability domain of the model, (3) the prediction(s) should be reliable, and (4) the outcome should be fit for the regulatory purpose (OECD, 2023). The scheme lays out two sets of assessment elements that must be considered before model predictions can be used in regulatory decisions. The first set is based on the 2007 OECD guidance principles for QSAR validation (summarized on left side of Fig. 3) (OECD, 2007), while the second set is

based on the 2023 OECD guidance on the assessment of QSAR predictions (summarized on right side of Fig. 3) (OECD, 2023).

QAF does not in itself provide a systematic categorization of diverse sources of uncertainty based on model components and modeling phases (as in our framework – Fig. 2); however, a number of issues and conditions raised in it with respect to regulatory application and acceptance of QSARs constitute the basis for the development of our framework. For example, within QAF, transparency and quality of experimental data for model building are key elements to consider in the assessment of the level of confidence in a model and predictions. As discussed in Section 1 and illustrated through the checklist (Table 4), our framework is similarly based on the understanding that *in silico* models and their predictions should be transparently reported to promote transparent evaluation of whether they are fit for defined purpose – this includes transparent accounting for uncertainty. In another example, similar to our framework, the 3rd principle in QAF recognizes that QSARs are associated with limitations with respect to, for example, physicochemical, structural and response spaces upon which they can generate reliable predictions and, at the same time, highlight issues related to model performance (OECD, 2023). In other words, these similarities suggest that the principles and our framework both recognize that the validity and uncertainty issues associated with the conceptual basis of models as well as the adequacy of model predictions are important considerations when evaluating whether models and predictions are fit-for-purpose.

Despite the similarities, our framework goes beyond the areas addressed in QAF (thus extending QAF) in different ways. For example, QAF's primary focus is the characterization of levels of uncertainty associated with the elements based on semi-qualitative uncertainty scales of “low”, “medium”, or “high”. The goal here is to determine (by balancing risk against benefits) whether the levels of uncertainty are acceptable within a given regulatory context. However, it is not always clear how much detail should be considered under the elements or what kind of uncertainty-related information is pertinent for an element to guide the characterization process. For example, despite the recommendation about the need to ensure quality of the underlying experimental data, QAF does not give a comprehensive characterization of what data quality entails (only data relevance and reliability are mentioned); rather, it open-endedly recommends that “the quality of individual data should also be assessed to the extent possible” (OECD, 2023; p14). In our framework, therefore, we not only expand consideration of data quality by proposing two additional indicators (data accuracy and validity) but also introduce two other aspects of data (i.e., data quantity and balance). In so doing, our framework aligns with the working principles of EFSA (Benford et al., 2018) and WHO/IPCS (2008), where comprehensive and explicit identification of possible

Table 4

A checklist used to highlight which GSU is relevant to consider from the case study. For each GSU selected, the corresponding justification is provided. Selected GSU is indicated by the ticked box (☑), while an empty box (☐) indicates that a GSU is not selected/considered relevant.

GSU	Justification for selecting a GSU
Data quantity	☐
Data balance	☐
Data relevance	☑ TEST is built on a large amount of data (i.e., 7413 available in ChemDplus database) collated from different studies and with multiple LD ₅₀ values for the same compound or isomers with different LD ₅₀ (US EPA, 2015), suggesting that some level of data variability is expected (Karman et al., 2022). When relying upon such heterogeneous data to predict LD ₅₀ , it thus becomes relevant to ask whether all the data are appropriate for making the predictions – e.g., do the data reflect appropriate/realistic chemical doses or exposure scenarios for rats?
Data reliability	☑ As explained under Data relevance, TEST modeling data has a high degree of variability. It should be noted that even with curation, such data may still suffer from reliability issues (e.g., in terms of data reproducibility), which ultimately impact the model prediction accuracy (Hoffmann et al., 2010; Karman et al., 2022). The fact that TEST does not report the reliability of the data suggests the need to factor in this GSU when interpreting the oral rat LD ₅₀ results.
Data accuracy	☑ Table 3 shows that the experimental LD ₅₀ -based GHS categories for the five chemicals do not match their predicted LD ₅₀ -based GHS categories (e.g., Quintiofol lies within GHS category 3 (according to its experimental LD ₅₀ value), while its predicted GHS category is 2. According to Gromek et al. (2022) and Langley (2005), such discrepancies may reflect inaccuracy in data on which a model is built. Thus, it is important to consider data accuracy as a source of uncertainty, especially when propagating uncertainty to the predicted results (Kopańska et al., 2023).
Data validity	☑ In the TEST user manual, the oral rat LD ₅₀ predictive abilities of TEST are considered as “not good” due to experimental uncertainty (US EPA (2015). Data validity is well recognized as a contributor to this type of uncertainty, attributed to the use of experimental data generated using procedures that partially or do not adhere to OECD Test guidelines or conform to good laboratory practice standards (Madden et al., 2020; Pham et al., 2019). The fact that this information is not characterized within the model presents a knowledge gap when judging the level of validity of the data.
Chemical structure	☐
Chemical similarity	☑ From the TEST prediction output, structural similarity coefficients for the five chemicals range from 0.57 to 0.81 (data not shown). Given this wide range of similarity (with as low as 0.57), it remains relevant to question, for example, whether the structurally diverse analogues may have dissimilar toxicological properties to the target compounds and how this might have influenced the accuracy of the predicted LD ₅₀ values.
Descriptor relevance	☐ TEST is developed from a pool of 797 2-dimensional descriptors, including classes such as molecular property (e.g., octanol-water partition coefficient) and molecular fragment counts (US EPA, 2015). A notable drawback in using such many descriptors is that there is no obvious way of determining whether each descriptor is relevant to the predicted LD ₅₀ , the extent to which the descriptors were considered relevant, or relative importance to the prediction output.

Table 4 (continued)

GSU	Justification for selecting a GSU
Descriptor concordance	☐
Model structure	☐
Activity/potency	☑ The predicted LD ₅₀ data assume that each chemical dose (in a statistical sense) will lead to 50% mortality in the rat population. However, according to Hoffmann et al. (2010), this assumption should be questioned, especially when asking whether the doses are realistic or representative of actual exposure scenarios involving the target animals and whether the doses will result in the recorded potency outcome.
Activity/potency evidence	☑ In decision-making contexts, where the basis on which the conclusions about the validity or reliability of the predicted rat LD ₅₀ data should be made, questions related to the weight of evidence that supports the conclusions are also pertinent to this discussion (Pestana et al., 2021; Schultz et al., 2019). For example, given the discrepancies between the <i>in vivo</i> and predicted LD ₅₀ values of the five compounds (Table 3), are there other consistent lines of evidence (e.g., similar values obtained from other methods like <i>in vitro</i> assays) to support the conclusion?
Model performance	☑ Model performance is here evaluated based on the ability of the model to accurately classify the predicted LD ₅₀ data. The overall evaluation of model performance across GHS categories indicates that the model accurately predicted GHS categories for 3/5 of the compounds (Table 3). However, the recorded under-and over-predictions (one in each case) raise questions about the level of reliability of the model for producing true positive/negative classifications (especially in the absence of <i>in vivo</i> data), or the rate at which incorrect (over- or under) predictions might occur in a large dataset.
Mechanistic plausibility	☐
Metabolic domain	☑ Consideration of chemical biotransformation is important for characterizing whether toxicity emanates from the parent compound or its metabolite (s) (Burden et al., 2016). While TEST can generate transformation products of compounds, it does not factor in any critical metabolite in the estimation of rat oral LD ₅₀ . For example, Demeton-O (Table 3) is known to produce the more toxic Demeton-S metabolite (LD ₅₀ of 1.5 mg/kg) in rats (Barnes and Denz, 1954); however, this is not accounted for in the predicted value (Table 3). During uncertainty analysis or interpretation of the predicted data, it thus remains relevant to consider the consequences of not including the influence of such toxic metabolite (Burden et al., 2016).
Coverage of ADME activity	☐
Applicability domain	☑ TEST Consensus model considers a prediction to be within its applicability domain provided the prediction is within the applicability domains of the Hierarchical clustering and Nearest neighbor models (US EPA, 2015). However, it remains unknown how representative the chemical spaces covered by the Hierarchical clustering and Nearest neighbor models are, particularly when considering the possibility that more relevant and structurally similar compounds may not have been covered by either (Zhu et al., 2009).
Extrapolation	☐
Model relevance	☑ One of the problems realized in the study above is over-and under-predictions (Table 3). This makes it important to consider uncertainty regarding the relevance of the models for hazard classification in a regulatory context. For example, where conservative predictions are desired as a health-protective strategy, under-prediction incidences (Table 3) are not

(continued on next page)

Table 4 (continued)

GSU	Justification for selecting a GSU
	desirable. On the other hand, over-prediction raises the question about whether false classification of less toxic (or safe) chemicals as more toxic (or toxic) might lead to potentially beneficial compounds being abandoned during chemical/drug development.

sources of uncertainty deemed to have the potential of altering conclusion drawn from predictions are key.

QAF is built on the premise that the usefulness of QSAR model(s) and the adequacy of their predictions are judged based on model performance in terms of the measures of goodness-of-fit and predictivity, and the consequence of the model(s) and predictions being uncertain. Our framework extends this understanding by further arguing that it might not always be beneficial to only consider these performance parameters, especially when assessing complex regulatory endpoints that are not well mechanistically understood (Cronin et al., 2019, 2022; Pestana et al., 2021; Schultz et al., 2019). Instead, the adequacy of QSARs to provide predictions with acceptable levels of confidence (which then rules out the need for any *in vivo* testing) should also be judged based on lines of evidence (presented in the framework as “Activity/potency evidence” – Fig. 2) from other decision-support methods such as *in vitro* tests. This is in line with the EFSA guidance on weight of evidence approach, which emphasizes the need to integrate and weight similar types of evidence to *in silico* predictions in order to improve confidence in the predicted outcome (EFSA et al., 2017). Such evidence can guide expert review processes aimed at determining the robustness of the models as well as the accuracy of their predictions (Pestana et al., 2021).

Within QAF, models (in the case of QSAR Model Reporting Format) and predictions (in the case of QSAR Prediction Reporting Format) are defined as separate steps (Fig. 3) when indeed, they should be defined as interconnected steps (Barber et al., 2024). This is true for reasons such as model predictions can only be trusted (and consequently accepted) if the model is suitable and robust enough to make the predictions within a defined applicability domain and if the reliability of the predictions can be ascertained (Barber et al., 2024). In attempts to incorporate this interconnection, our framework, therefore, shows connections between the proposed modeling phases (creation, characterization and application) (Fig. 2). The idea here is to promote a holistic view of the entire modeling process (which includes a model and prediction) and enable a better understanding of how different components in different phases might interact or the implications of changes in one phase on the entire modeling process.

5. Discussion and conclusion

A general uncertainty categorization framework that aids a

structured means of identifying, categorizing, and describing diverse sources of uncertainty associated with *in silico* toxicology models and their predictions could promote the alignment of terminologies for describing the sources of uncertainty and contribute to transparent communication with decision-makers about the models and their predictions (Alexander-White et al., 2022; ECHA, 2012; Kirchner et al., 2021). This study contributes to the development of such a framework.

We analyzed studies that have categorized sources of uncertainty across different *in silico* toxicology methods. Our analysis reveals that there is little overlap between the studies in terms of the kind and number of uncertainty sources they cover within as well as across the methods they describe, which, therefore, suggests the need for a general framework that covers a wide range of uncertainty sources across the methods. Additionally, as discussed in Section 3, there is little alignment in the terminologies used to describe the same sources of uncertainty. In a similar analysis of terminologies used in different uncertainty typologies in the general risk assessment literature, Skinner et al. (2014a,b) noted that such a lack of harmonization of terminologies presents a gap in the literature, as it does not only lead to confusion about the meaning of the uncertainty sources but also contributes to poor communication of the sources with stakeholders.

In an attempt to fill the highlighted gaps, we developed a framework (Fig. 2) that covers the different sources of uncertainty described in the analyzed studies and harmonizes terminologies used in describing similar uncertainty sources. While the framework is based on the framework by Belfield et al. (2021), we see our contributions in three ways. Firstly, we modified the framework by specifically tailoring it towards areas of uncertainty relevant to *in silico* toxicology modeling. This was done by introducing two new components (i.e., “Similarity” and “Applicability”), which, similar to Cronin et al. (2019), we argue to be more relevant for describing uncertainty sources in *in silico* toxicology modeling than “Description” and “Usability” proposed in the framework by Belfield et al. (2021). Secondly, we assessed, compared, and synthesized existing uncertainty sources in the analyzed studies and showed that these uncertainty sources, despite being discussed under different *in silico* toxicology methods, can be systematically categorized under the modified framework to form a more comprehensive uncertainty categorization framework. Lastly, our framework draws on diverse experiences and perspectives on sources of uncertainty in the *in silico* toxicology modeling literature as well as the recently OECD’s proposed QAF (OECD, 2023). Thus, relative to the one proposed by Belfield et al. (2021) (or the original framework by Cronin et al. (2019)), it can be argued that our framework is more representative of areas of uncertainty identified by multiple modelers. In other words, the importance of our framework is in its conceptual breadth and ability to provide a more holistic picture of the diversity of sources of uncertainty in *in silico* toxicology methods. As further illustrated in the Case study under Section 4, we have shown that the introduced general sources of uncertainty

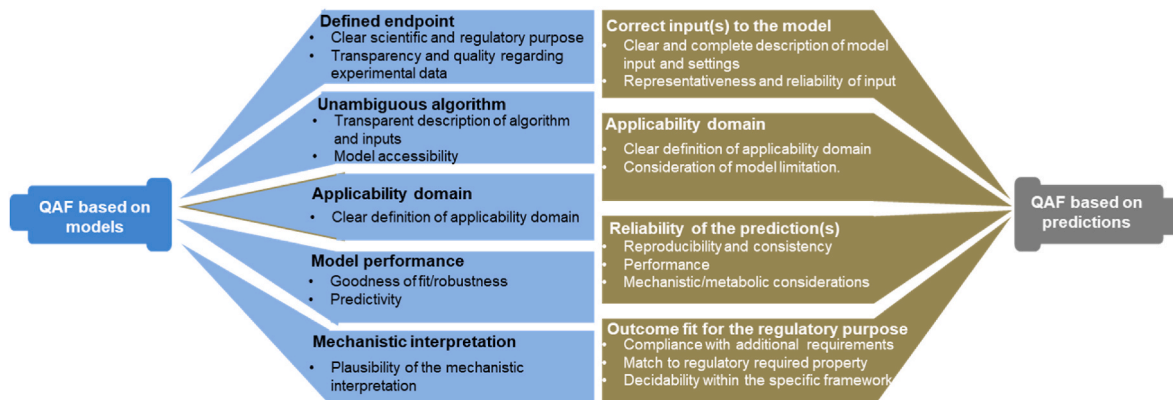


Fig. 3. A summary of the principles and elements outlined in the 2023 OECD’s proposed QAF. The elements are bulleted under the principles.

(GSU) can provide a more nuanced understanding and practical way of prioritizing specific areas within an *in silico* toxicology prediction for uncertainty analysis or for addressing uncertainty.

With the overarching aim of fostering a structured (and potentially more transparent) understanding of where uncertainties reside, our framework (Fig. 2) and the checklist (Table 4) can help modelers to reduce the risk of overlooking particular uncertainty sources during modeling, prioritize sources to dedicate efforts and resources for uncertainty analysis, and critically reflect on appropriate strategies to reduce and (where possible) eliminate uncertainties. Alternatively, the use of the framework could help increase transparency and trust in a model or modeling exercise, especially with regards to communicating uncertainties between modelers and relevant stakeholders – this is in line with the working principles of OECD (2007, 2023) and WHO/IPCS (2008), where transparency and trust are key to regulatory acceptance of models and predictions.

The proposed framework is intended to be as flexible as possible; thus, future studies may continue refining it. Moving forward, we are currently exploring other ways to test its practical application in identifying and characterizing uncertainties in the context of *in silico* predictions of a diverse and larger number of compounds. Lastly, we would like to acknowledge the difficulty of developing a framework that covers all possible sources of uncertainty; thus, while our framework covers several GSU within *in silico* toxicology modeling, we refrain from claiming to have developed a “standard” framework to this end. Additionally, we would also like acknowledge that it is possible that the literature search criteria applied in our study (under Section 2) might have led to some sources of uncertainty or the (grey) literature in *in silico* toxicology methods not being captured. Nevertheless, believe that this potential limitation did not affect the conceptual breadth of our framework.

CRedit authorship contribution statement

Jerry Achar: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **James W. Firman:** Writing – review & editing. **Mark T.D. Cronin:** Writing – review & editing. **Gunilla Öberg:** Writing – review & editing, Supervision, Funding acquisition.

Funding information

This study was supported by the Vanier Canada Graduate Scholarship (Vanier CGS), and the Social Science, and Humanities Research Council (SSHRC) grant (Grant No: 435-2019-0465).

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.yrtph.2024.105737>.

Data availability

No data was used for the research described in the article.

References

Achar, J.C., Kim, D.Y., Kwon, J.-H., Jung, J., 2020a. Toxicokinetic modeling of octylphenol bioconcentration in *Chlorella vulgaris* and its trophic transfer to *Daphnia magna*. *Ecotoxicol. Environ. Saf.* 194, 110379. <https://doi.org/10.1016/j.ecoenv.2020.110379>.

- Achar, J.C., Nam, G., Jung, J., Klammler, H., Mohamed, M.M., 2020b. Microbubble ozonation of the antioxidant butylated hydroxytoluene: degradation kinetics and toxicity reduction. *Environ. Res.* 186, 109496. <https://doi.org/10.1016/j.envres.2020.109496>.
- Achar, J., Cronin, M.T.D., Firman, J.W., Öberg, G., 2024c. A problem formulation framework for the application of *in silico* toxicology methods in chemical risk assessment. *Arch. Toxicol.* <https://doi.org/10.1007/s00204-024-03721-6>.
- Achar, J., Firman, J.W., Tran, C., Kim, D., Cronin, M.T.D., Öberg, G., 2024d. Analysis of implicit and explicit uncertainties in QSAR prediction of chemical toxicity: a case study of neurotoxicity. *Regul. Toxicol. Pharmacol.* 154, 105716. <https://doi.org/10.1016/j.yrtph.2024.105716>.
- Alexander-White, C., Bury, D., Cronin, M., Dent, M., Hack, E., Hewitt, N.J., Kenna, G., Naciff, J., Ouedraogo, G., Schepky, A., Mahony, C., Europe, C., 2022. A 10-step framework for use of read-across (RAX) in next generation risk assessment (NGRA) for cosmetics safety assessment. *Regul. Toxicol. Pharmacol.* 129, 105094. <https://doi.org/10.1016/j.yrtph.2021.105094>.
- Ball, N., Cronin, M.T.D., Shen, J., Blackburn, K., Booth, E.D., Bouhifd, M., Donley, E., Egnash, L., Hastings, C., Juberg, D.R., Kleensang, A., Kleinstreuer, N., Kroese, E.D., Lee, A.C., Luechtefeld, T., Maertens, A., Marty, S., Naciff, J.M., Palmer, J., Hartung, T., 2016. t4 report: Toward Good Read-Across Practice (GRAP) Guidance. *ALTEX* 33 (2), 149–166. <https://doi.org/10.14573/altex.1601251>.
- Barber, C., Heghes, C., Johnston, L., 2024. A framework to support the application of the OECD guidance documents on (QSAR) model validation and prediction assessment for regulatory decisions. *Computational Toxicology* 30, 100305. <https://doi.org/10.1016/j.comtox.2024.100305>.
- Barnes, J.M., Denz, F.A., 1954. The reaction of rats to diets containing octamethyl pyrophosphoramide (schradan) and 00-diethyl-S-ethylmercaptoethanol thiophosphate (“Systox”). *Br. J. Ind. Med.* 11 (1), 11–19.
- Belfield, S.J., Enoch, S.J., Firman, J.W., Madden, J.C., Schultz, T.W., Cronin, M.T.D., 2021. Determination of “fitness-for-purpose” of quantitative structure-activity relationship (QSAR) models to predict (eco)-toxicological endpoints for regulatory use. *Regul. Toxicol. Pharmacol.* 123, 104956. <https://doi.org/10.1016/j.yrtph.2021.104956>.
- Benfenati, E., Chaudhry, Q., Gini, G., Dorne, J.L., 2019. Integrating *in silico* models and read-across methods for predicting toxicity of chemicals: a step-wise strategy. *Environ. Int.* 131, 105060. <https://doi.org/10.1016/j.envint.2019.105060>.
- Blackburn, K., Stuard, S.B., 2014. A framework to facilitate consistent characterization of read across uncertainty. *Regul. Toxicol. Pharmacol.* 68 (3), 353–362. <https://doi.org/10.1016/j.yrtph.2014.01.004>.
- Burden, N., Maynard, S.K., Weltje, L., Wheeler, J.R., 2016. The utility of QSARs in predicting acute fish toxicity of pesticide metabolites: a retrospective validation approach. *Regul. Toxicol. Pharmacol.* 80, 241–246. <https://doi.org/10.1016/j.yrtph.2016.05.032>.
- Chandrasekaran, B., Abed, S.N., Al-Attraqchi, O., Kuche, K., Tekade, R.K., 2018. Chapter 21—computer-aided prediction of pharmacokinetic (ADMET) properties. In: Tekade, R.K. (Ed.), *Dosage Form Design Parameters*. Academic Press, pp. 731–755. <https://doi.org/10.1016/B978-0-12-814421-3.00021-X>.
- Cronin, M., Madden, J., 2010. In: *In Silico Toxicology: Principles and Applications*. Royal Society of Chemistry.
- Cronin, M.T.D., Madden, J., Enoch, S., Roberts, D., 2013. *Chemical Toxicity Prediction: Category Formation and Read-Across*. Royal Society of Chemistry.
- Cronin, M.T.D., Richarz, A.-N., Schultz, T.W., 2019. Identification and description of the uncertainty, variability, bias and influence in quantitative structure-activity relationships (QSARs) for toxicity prediction. *Regul. Toxicol. Pharmacol.* 106, 90–104. <https://doi.org/10.1016/j.yrtph.2019.04.007>.
- Cronin, M.T.D., Bauer, F.J., Bonnell, M., Campos, B., Ebbrell, D.J., Firman, J.W., Gutsell, S., Hodges, G., Patlewicz, G., Sapounidou, M., Spînu, N., Thomas, P.C., Worth, A.P., 2022. A scheme to evaluate structural alerts to predict toxicity – assessing confidence by characterising uncertainties. *Regul. Toxicol. Pharmacol.* 135, 105249. <https://doi.org/10.1016/j.yrtph.2022.105249>.
- Drake, C., Wehr, M.M., Zobl, W., Koschmann, J., De Lucca, D., Kühne, B.A., Hansen, T., Knebel, J., Ritter, D., Boei, J., Vrieling, H., Bitsch, A., Escher, S.E., 2023. Substantiate a read-across hypothesis by using transcriptome data—a case study on volatile diketones. *Front. Toxicol.* 5. <https://doi.org/10.3389/ftox.2023.1155645>.
- ECHA, 2012. Guidance on information requirements and chemical safety assessment. *Chapter R.19: Uncertainty analysis*. https://echa.europa.eu/documents/10162/17224/information_requirements_r19_en.pdf/d5bd6c3f-3383-49df-894e-dea410ba4335?t=1353935215756.
- EFSA, Hardy, A., Benford, D., Halldorsson, T., Jeger, M.J., Knutsen, H.K., More, S., Naegeli, H., Noteborn, H., Ockleford, C., Ricci, A., Rychen, G., Schlatter, J.R., Silano, V., Solecki, R., Turck, D., Benfenati, E., Chaudhry, Q.M., Craig, P., et al., 2017. Guidance on the use of the weight of evidence approach in scientific assessments. *EFSA J.* 15 (8), e04971. <https://doi.org/10.2903/j.efsa.2017.4971>.
- Benford, D., Halldorsson, T., Jeger, M.J., Knutsen, H.K., More, S., Naegeli, H., Noteborn, H., Ockleford, C., Ricci, A., Rychen, G., Schlatter, J.R., Silano, V., Solecki, R., Turck, D., Younes, M., Craig, P., Hart, A., Von Goetz, N., et al., EFSA, 2018. The principles and methods behind EFSA’s guidance on uncertainty analysis in scientific assessment. *EFSA J.* 16 (1), e05122. <https://doi.org/10.2903/j.efsa.2018.5122>.
- Enoch, S.J., 2010. Chemical category formation and read-across for the prediction of toxicity. In: Puzyn, T., Leszczynski, J., Cronin, M.T. (Eds.), *Recent Advances in QSAR Studies: Methods and Applications*. Springer, Netherlands, pp. 209–219. https://doi.org/10.1007/978-1-4020-9783-6_7.
- Firman, J.W., Cronin, M.T.D., Rowe, P.H., Semenova, E., Doe, J.E., 2022. The use of Bayesian methodology in the development and validation of a tiered assessment

- approach towards prediction of rat acute oral toxicity. *Arch. Toxicol.* 96 (3), 817–830. <https://doi.org/10.1007/s00204-021-03205-x>.
- Fu, X., Wojak, A., Neagu, D., Ridley, M., Travis, K., 2011. Data governance in predictive toxicology: a review. *J. Cheminf.* 3 (1), 24. <https://doi.org/10.1186/1758-2946-3-24>.
- Gissi, A., Tcheremenskaia, O., Bossa, C., Battistelli, C.L., Browne, P., 2024. The OECD (Q) SAR Assessment Framework: a tool for increasing regulatory uptake of computational approaches. *Computational Toxicology* 31, 100326. <https://doi.org/10.1016/j.comtox.2024.100326>.
- Graham, J.C., Rodas, M., Hillegass, J., Schulze, G., 2021. The performance, reliability and potential application of *in silico* models for predicting the acute oral toxicity of pharmaceutical compounds. *Regul. Toxicol. Pharmacol.* 119, 104816. <https://doi.org/10.1016/j.yrtph.2020.104816>.
- Gromek, K., Hawkins, W., Dunn, Z., Gawlik, M., Ballabio, D., 2022. Evaluation of the predictivity of Acute Oral Toxicity (AOT) structure-activity relationship models. *Regul. Toxicol. Pharmacol.* 129, 105109. <https://doi.org/10.1016/j.yrtph.2021.105109>.
- Hartung, T., Hoffmann, S., Stephens, M., 2013. Food for thought ... Mechanistic validation. *ALTEX* 30 (2), 119–130.
- He, Haibo, Garcia, E.A., 2009. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* 21 (9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>.
- Hoffmann, S., Kinsner-Ovaskainen, A., Prieto, P., Mangelsdorf, I., Bieler, C., Cole, T., 2010. Acute oral toxicity: Variability, reliability, relevance and interspecies comparison of rodent LD50 data from literature surveyed for the ACuteTox project. *Regul. Toxicol. Pharmacol.* 58 (3), 395–407. <https://doi.org/10.1016/j.yrtph.2010.08.004>.
- Jones, C., Falloon, P., 2009. Sources of uncertainty in global modelling of future soil organic carbon storage. In: Baveye, P.C., Laba, M., Mysiak, J. (Eds.), *Uncertainties in Environmental Modelling and Consequences for Policy Making*. Springer, Netherlands, pp. 283–315. https://doi.org/10.1007/978-90-481-2636-1_13.
- Karmaus, A.L., Mansouri, K., To, K.T., Blake, B., Fitzpatrick, J., Strickland, J., Patlewicz, G., Allen, D., Casey, W., Kleinstreuer, N., 2022. Evaluation of variability across rat acute oral systemic toxicity studies. *Toxicol. Sci.* 188 (1), 34–47. <https://doi.org/10.1093/toxsci/kfac042>.
- Kirchner, M., Mitter, H., Schneider, U.A., Sommer, M., Falkner, K., Schmid, E., 2021. Uncertainty concepts for integrated modeling—review and application for identifying uncertainties and uncertainty propagation pathways. *Environ. Model. Software* 135, 104905. <https://doi.org/10.1016/j.envsoft.2020.104905>.
- Kopańska, K., Rodríguez-Belenguier, P., Llopis-Lorente, J., Trenor, B., Saiz, J., Pastor, M., 2023. Uncertainty assessment of proarrhythmia predictions derived from multi-level *in silico* models. *Arch. Toxicol.* 97 (10), 2721–2740. <https://doi.org/10.1007/s00204-023-03557-6>.
- Langley, G., 2005. Acute toxicity testing without animals: more scientific and less of a gamble. Application of Alternative Methods Collection. <https://www.wellbeingintlstudiesrepository.org/appamet/1>.
- Madden, J.C., Enoch, S.J., Paini, A., Cronin, M.T.D., 2020. A review of *in silico* tools as alternatives to animal testing: Principles, resources and applications. *Alter. Lab. Animals* 48 (4), 146–172. <https://doi.org/10.1177/0261192920965977>.
- Nendza, M., Aldenberg, T., Benfenati, E., Benigni, R., Cronin, M.T.D., Escher, S., Fernandez, A., Gabbert, S., Giralt, F., Hewitt, M., Hrovat, M., Jeram, S., Kroese, D., Madden, J., Mangelsdorf, I., Rallo, R., Roncaglioni, A., Rorije, E., Segner, H., Vermeire, T., 2010. Chapter 4: Data Quality Assessment for *In Silico* Methods: A Survey of Approaches and Needs. In *Plant Dis.—PLANT DIS* 59–117.
- OECD, 2007. Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. <https://doi.org/10.1787/9789264085442-en>.
- OECD, 2023. Guiding principles and key elements for establishing a weight of evidence for chemical assessment. Organisation for Economic Co-operation and Development. https://www.oecd-ilibrary.org/environment/guiding-principles-and-key-elements-for-establishing-a-weight-of-evidence-for-chemical-assessment_f11597f6-en.
- OECD, 2023. (Q)SAR Assessment Framework: Guidance for the Regulatory Assessment of (Quantitative) Structure Activity Relationship Models and Predictions. Organisation for Economic Co-operation and Development. https://www.oecd-ilibrary.org/environment/q-sar-assessment-framework-guidance-for-the-regulatory-assessment-of-quantitative-structure-activity-relationship-models-and-predictions_d96118f6-en.
- Parish, S.T., Aschner, M., Casey, W., Corvaro, M., Embry, M.R., Fitzpatrick, S., Kidd, D., Kleinstreuer, N.C., Lima, B.S., Settivari, R.S., Wolf, D.C., Yamazaki, D., Boobis, A., 2020. An evaluation framework for new approach methodologies (NAMs) for human health safety assessment. *Regul. Toxicol. Pharmacol.* 112, 104592. <https://doi.org/10.1016/j.yrtph.2020.104592>.
- Pestana, C.B., Firman, J.W., Cronin, M.T.D., 2021. Incorporating lines of evidence from New Approach Methodologies (NAMs) to reduce uncertainties in a category based read-across: a case study for repeated dose toxicity. *Regul. Toxicol. Pharmacol.* 120, 104855. <https://doi.org/10.1016/j.yrtph.2020.104855>.
- Pham, L.L., Sheffield, T.Y., Pradeep, P., Brown, J., Haggard, D.E., Wambaugh, J., Judson, R.S., Paul Friedman, K., 2019. Estimating uncertainty in the context of new approach methodologies for potential use in chemical safety evaluation. *Curr. Opin. Toxicol.* 15, 40–47. <https://doi.org/10.1016/j.cotox.2019.04.001>.
- 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 (5–6), 435–459. <https://doi.org/10.1080/1062936X.2012.664825>.
- Sahlén, U., Filipsson, M., Öberg, T., 2011. A risk assessment perspective of current practice in characterizing uncertainties in QSAR regression predictions. *Mol. Inform.* 30 (6–7), 551–564. <https://doi.org/10.1002/minf.201000177>.
- Sahlén, U., Golsteijn, L., Iqbal, M.S., Peijnenburg, W., 2013. Arguments for considering uncertainty in QSAR predictions in hazard and risk assessments. *Altern. Lab. Anim.* 41 (1), 91–110. <https://doi.org/10.1177/026119291304100110>.
- Sahlén, U., Jeliakova, N., Öberg, T., 2014. Applicability domain dependent predictive uncertainty in QSAR regressions. *Mol. Inform.* 33 (1), 26–35. <https://doi.org/10.1002/minf.201200131>.
- Sarah, J. Tracy, 2018. A phronetic iterative approach to data analysis in qualitative research. *J. Qual. Research* 19 (2), 61–76. <https://doi.org/10.22284/QR.2018.19.2.61>.
- Schilter, B., Benigni, R., Boobis, A., Chiodini, A., Cockburn, A., Cronin, M.T.D., Lo Piparo, E., Modi, S., Thiel, A., Worth, A., 2014. Establishing the level of safety concern for chemicals in food without the need for toxicity testing. *Regul. Toxicol. Pharmacol.* 68 (2), 275–296. <https://doi.org/10.1016/j.yrtph.2013.08.018>.
- Schultz, T.W., Amcoff, P., Berggren, E., Gautier, F., Klaric, M., Knight, D.J., Mahony, C., Schwarz, M., White, A., Cronin, M.T.D., 2015. A strategy for structuring and reporting a read-across prediction of toxicity. *Regul. Toxicol. Pharmacol.* 72 (3), 586–601. <https://doi.org/10.1016/j.yrtph.2015.05.016>.
- Schultz, T.W., Richarz, A.-N., Cronin, M.T.D., 2019. Assessing uncertainty in read-across: questions to evaluate toxicity predictions based on knowledge gained from case studies. *Computational Toxicology* 9, 1–11. <https://doi.org/10.1016/j.comtox.2018.10.003>.
- Skinner, D.J.C., Rocks, S.A., Pollard, S.J.T., 2014a. A review of uncertainty in environmental risk: characterising potential natures, locations and levels. *J. Risk Res.* 17 (2), 195–219. <https://doi.org/10.1080/13669877.2013.794150>.
- Skinner, D.J.C., Rocks, S.A., Pollard, S.J.T., Drew, G.H., 2014b. Identifying uncertainty in environmental risk assessments: the development of a novel typology and its implications for risk characterization. *Hum. Ecol. Risk Assess.* 20 (3), 607–640. <https://doi.org/10.1080/10807039.2013.779899>.
- Stausberg, J., rgen, Harkener, S., 2023. Data quality and data quantity: complements or contradictions?. In: *Healthcare Transformation with Informatics and Artificial Intelligence*. IOS Press, pp. 24–27. <https://doi.org/10.3233/SHTI230414>.
- United Nations, 2021. GHS classification criteria for acute toxicity. <https://webapps.ilo.org/static/english/protection/safework/ghs/ghsfinal/ghsc05.pdf>.
- US EPA, 2011. Exposure factors handbook chapter 2—variability and uncertainty. <https://www.epa.gov/sites/default/files/2015-09/documents/efh-chapter02.pdf>.
- US EPA, O., 2015. *Toxicity estimation software tool (TEST)* [data and tools]. <https://www.epa.gov/comptox-tools/toxicity-estimation-software-tool-test>.
- US EPA, O., 2016. *(Quantitative) structure activity relationship [(Q)SAR] guidance document* [other policies and guidance]. <https://www.epa.gov/pesticide-registration/quantitative-structure-activity-relationship-qsar-guidance-document>.
- Walker, W.E., Harremoës, P., Rotmans, J., Sluijs, J.P. van der, Asselt, M.B.A. van, Janssen, P., Krauss, M.P.K. von, 2003. Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integr. Assess.* 4 (1), 5–17. <https://doi.org/10.1076/iaij.4.1.5.16466>.
- Wang, N.C.Y., Jay Zhao, Q., Wesselkamper, S.C., Lambert, J.C., Petersen, D., Hess-Wilson, J.K., 2012. Application of computational toxicological approaches in human health risk assessment. I. A tiered surrogate approach. *Regul. Toxicol. Pharmacol.* 63 (1), 10–19. <https://doi.org/10.1016/j.yrtph.2012.02.006>.
- WHO/IPCS, 2008. Part 1: Guidance Document on Characterizing and Communicating Uncertainty in Exposure Assessment, sixth ed. World Health Organization <https://www.who.int/ipcs/methods/harmonization/areas/uncertainty20.pdf>.
- Zhu, H., Martin, T.M., Ye, L., Sedykh, A., Young, D.M., Tropsha, A., 2009. QSAR modeling of rat acute toxicity by oral exposure. *Chem. Res. Toxicol.* 22 (12), 1913–1921. <https://doi.org/10.1021/tx900189p>.