

**Analyzing and Accounting for Uncertainty in Quantitative Structure-Activity Relationship  
(QSAR) Prediction of Chemical Toxicity**

by

Jerry Collince Achar

B.Sc., Kenyatta University, 2015

M.Sc., Korea University, 2020

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF  
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES  
(Resources, Environment and Sustainability)

THE UNIVERSITY OF BRITISH COLUMBIA  
(Vancouver)

January 2025

© Jerry Collince Achar, 2025

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Analyzing and Accounting for Uncertainty in Quantitative Structure-Activity Relationship (QSAR)  
Prediction of Chemical Toxicity

---

submitted by Jerry Collince Achar in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Resources, Environment and Sustainability

**Examining Committee:**

Dr. Gunilla Oberg, Professor, Institute for Resources, Environment and Sustainability, UBC  
Supervisor

Dr. Milind Kandlikar, Professor, Institute for Resources, Environment and Sustainability, UBC  
Supervisory Committee Member

Dr. Mark Cronin, Professor, School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University  
Supervisory Committee Member

Dr. Jeroen van der Sluijs, Professor, Centre for the Study of the Sciences and the Humanities, University of Bergen  
Supervisory Committee Member

Dr. Artem Cherkasov, Professor, Department of Urologic Sciences, UBC  
University Examiner

Dr. Tao Huan, Professor, Department of Chemistry, UBC  
University Examiner

Dr. Jonathan Goodman, Professor, Department of Chemistry, University of Cambridge  
External Examiner

## Abstract

Improving regulatory confidence in and acceptance of *in silico* toxicology methods and their predictions requires assessment and transparent communication of associated uncertainty to facilitate the evaluation of whether they are fit for purpose. This thesis develops frameworks and methods to facilitate systematic and transparent analysis of and accounting for uncertainty in *in silico* toxicology methods, with a central focus on QSARs. This is done through four studies. Studies 1 and 2 conduct a literature review to identify key components relevant to *in silico* toxicology problem formulation and modeling processes, which are then systematically categorized and rationalized as potential sources of uncertainty. The outcomes from the studies are four problem formulation components (as in Study 1) and 20 sources of uncertainty (as in Study 2). Study 3 focuses on analyzing implicit and explicit uncertainties expressed within QSAR studies predicting neurotoxicity of chemicals. To this end, implicit and explicit uncertainty indicators are identified, whereafter, the indicators are used to identify uncertainties. A systematic categorization of the uncertainties, according to the 20 uncertainty sources in Study 2, reveals that implicit uncertainty is expressed at a higher rate (64%) in the analyzed studies and within most uncertainty sources (65%; 13/20), indicating that uncertainty is predominantly expressed implicitly in the field. Study 4 proposes a consensus method combining TEST and CATMoS model predictions to produce conservative predictions. The level of conservativeness of the consensus predictions against the individual models is evaluated based upon the agreement of predicted LD<sub>50</sub>-based GHS categories with the corresponding experimental LD<sub>50</sub>-based GHS categories. The results show that the consensus method produces a higher over-prediction rate (39%; 2,504/6,410) than TEST (24%) or CATMoS (25%), while its under-prediction rate is lower at 8% than TEST (20%) or CATMoS (10%), which indicates that, by design, it is the most conservative. The outcomes from the four studies benefit the field of *in silico* toxicology by contributing to efforts aimed towards addressing the issue of uncertainty, promoting regulatory acceptance of models (e.g., QSARs) and their predictions, as well as reducing and (where possible) replacing use of animal in chemical toxicity assessment.

## **Lay Summary**

Improving confidence in computer-based models predicting chemical harm requires analyzing and transparently communicating uncertainties. This thesis developed two frameworks to systematically identify and categorize key components in model problem formulation and areas of uncertainty. It also analyzed how uncertainties are expressed within studies predicting chemical harm, revealing they are mainly conveyed indirectly through uncertainty indicators. Additionally, the research proposed a consensus approach that combines two model predictions and evaluated its performance in predicting acute oral toxicity for 6,410 compounds. Results showed this approach produced the highest number of over-predictions, making it the most conservative and health-protective. These frameworks and methods aim to enhance confidence in model predictions, supporting their acceptance for regulatory use.

## Preface

This thesis consists of individual research articles, with Chapters 2 through 5 structured as stand-alone manuscripts. Each manuscript was tailored for publication in different journals, which has led to some repetition in the introduction across chapters. In all the chapters, I took the lead in conducting the research with guidance from two committee members – Professor Gunilla Öberg (my supervisor) and Professor Mark Cronin – and contributions from other co-authors:

1. Chapter 2 has been published in Achar, J., Cronin, M. T. D., Firman, J. W., & Öberg, G. (2024). A problem formulation framework for the application of *in silico* toxicology methods in chemical risk assessment. *Archives of Toxicology*. <https://doi.org/10.1007/s00204-024-03721-6>.

This was a multi-author paper. I led the work and analysis in the study as recognized in the authorship contribution statement: Conceptualization, Methodology, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization. As further specified in the statement, the co-authors' contributions in this study were as follows. James W. Firman: Writing – review and editing; Mark T.D. Cronin: Writing – review and editing; Gunilla Öberg: Writing – review and editing, Supervision, Funding acquisition.

2. Chapter 3 has been published in Achar, J., Firman, J. W., Cronin, M. T. D., & 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, 105737. <https://doi.org/10.1016/j.yrtph.2024.105737>

This was a multi-author paper. I led the work and analysis in this study as recognized in the CRediT authorship contribution statement: Conceptualization, Methodology, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, and Visualization. As further specified in the statement, the co-authors' contributions in this study were as follows. James W. Firman: Writing – review and editing; Mark T.D. Cronin: Writing – review and editing; Gunilla Öberg: Writing – review and editing, Supervision, Funding acquisition.

3. Chapter 4 has been published in Achar, J., Firman, J. W., Tran, C., Kim, D., Cronin, M. T. D., & Öberg, G. (2024). Analysis of implicit and explicit uncertainties in QSAR prediction of chemical toxicity: A case study

of neurotoxicity. *Regulatory Toxicology and Pharmacology*, 154, 105716.

<https://doi.org/10.1016/j.yrtph.2024.105716>.

This was a multi-author paper. I led the work and analysis in the study as recognized in the CRediT authorship contribution statement: Conceptualization, Methodology, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization. As further specified in the statement, the co-authors' contributions in this study were as follows. James W. Firman: Writing – review and editing, Formal analysis; Chantelle Tran: Formal analysis; Daniella Kim: Formal analysis; Mark T.D. Cronin: Writing – review and editing; Gunilla Öberg: Writing – review and editing, Supervision, Funding acquisition.

4. Chapter 5 is present within Achar, J., Cronin, M. T. D., Firman, J. W. Conservative Consensus QSAR

approach for the prediction of rat acute oral toxicity. The work has been submitted and is under review.

This was a multi-author paper. I led the work and analysis in the study, through: Conceptualization, Methodology, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, and Visualization. The two co-authors (James W. Firman and Mark T.D. Cronin) each contributed to the work through review and editing.

## Table of Contents

<b>Abstract</b> .....	<b>iii</b>
<b>Lay Summary</b> .....	<b>iv</b>
<b>Preface</b> .....	<b>v</b>
<b>Table of Contents</b> .....	<b>vii</b>
<b>List of Tables</b> .....	<b>xi</b>
<b>List of Figures</b> .....	<b>xiii</b>
<b>List of Abbreviations</b> .....	<b>xv</b>
<b>Acknowledgements</b> .....	<b>xvii</b>
<b>Dedication</b> .....	<b>xviii</b>
<b>Chapter 1: Introduction</b> .....	<b>1</b>
1.1 Chemical safety and animal testing .....	1
1.1.1 Challenges in successful translation of animal testing data to humans .....	2
1.2 Non-animal methods .....	4
1.2.1 Application of in silico toxicology methods .....	6
1.2.1.1 Existing challenges in integrating in silico methods into the regulatory landscape .....	8
1.2.1.2 Considerations of uncertainty to support regulatory application of in silico methods.....	10
1.2.1.3 Addressing uncertainty in in silico predictions through conservative assumptions.....	12
1.2.1.4 Incorporating problem formulation into in silico toxicology predictions .....	13
1.3 Research objectives .....	14
<b>Chapter 2: A problem formulation framework for the application of in silico toxicology methods in chemical risk assessment</b> .....	<b>17</b>
2.1 Introduction .....	17
2.2 Materials and Methods .....	20
2.2.1 Identifying PF components in the general risk assessment literature.....	20
2.2.2 Formulating a general PF framework .....	20
2.2.3 Applying the PF framework to in silico toxicology methods.....	21
2.3 Results and Discussion.....	21
2.3.1 PFs for in silico toxicology methods.....	21
2.3.2 Applying higher-level conceptual components for in silico toxicology methods .....	23

2.3.2.1 Problem framing .....	23
2.3.2.2 Problem exploration .....	23
2.3.2.3 Conceptual model.....	25
2.3.2.4 Hypothesis formulation .....	27
2.3.3 PF components in studies of in silico toxicology methods .....	30
2.3.4 Uncertainties associated with the higher-level components of PF .....	32
2.3.4.1 Sources of uncertainty .....	33
2.3.4.1.1 Problem framing .....	33
2.3.4.1.2 Problem exploration .....	34
2.3.4.1.3 Conceptual model.....	35
2.3.4.1.4 Hypothesis formulation .....	36
2.3.4.2 Characterizing and addressing uncertainty associated with PF components .....	37
2.3.5 Further consideration of PFs.....	38
2.4 Conclusion .....	39

**Chapter 3: A framework for categorizing sources of uncertainty in in silico toxicology methods: considerations for chemical toxicity predictions.....40**

3.1 Introduction .....	40
3.2 Identification and verbatim recording of sources of uncertainty (VRSU) in the literature.....	42
3.3 Categorizing VRSU and formulating GSU .....	43
3.3.1 The Model Creation phase .....	50
3.3.1.1 Data.....	50
3.3.1.2 Structure .....	53
3.3.1.3 Similarity.....	54
3.3.1.4 Descriptors.....	55
3.3.2 The Model Characterization Phase.....	56
3.3.2.1 Modeling.....	56
3.3.2.2 Performance .....	58
3.3.2.3 Mechanisms.....	58
3.3.2.4 Toxicokinetics.....	60
3.3.3 The Model Application Phase .....	61
3.3.3.1 Applicability .....	61
3.3.3.2 Relevance.....	61
3.4 Application of the framework to prioritize areas of uncertainty.....	62

3.4.1 Case study.....	62
3.4.2 Identification of relevant GSU from the case study.....	64
3.4.3 Consideration of the framework within the OECD’s QSAR Assessment Framework.....	67
3.5 Discussion and Conclusion.....	70
<b>Chapter 4: Analysis of implicit and explicit uncertainties in QSAR prediction of chemical toxicity: a case study of neurotoxicity .....</b>	<b>74</b>
4.1 Introduction.....	74
4.2 Methodology .....	77
4.2.1 Uncertainty indicators .....	77
4.2.2 Selecting peer-reviewed papers for analysis .....	78
4.2.3 Identification of implicit and explicit uncertainty indicators .....	79
4.2.4 Identification of implicit and explicit uncertainties .....	80
4.2.5 Categorization of the identified uncertainties.....	81
4.2.5.1 The categorization process .....	83
4.3 Results.....	84
4.3.1 Intercoder agreement.....	84
4.3.2 Occurrence of uncertainty indicators .....	84
4.3.3 Variation in the occurrence of implicit and explicit uncertainty sources.....	88
4.3.4 Frequency of uncertainty sources .....	88
4.3.4.1 General distribution of uncertainty sources.....	89
4.3.4.2 Comparison of frequencies relating to implicit and explicit uncertainty .....	90
4.4 Discussion .....	91
4.4.1 Contribution of implicit versus explicit uncertainty sources to the overall uncertainty sources .....	91
4.4.2 Level of concerns raised about the uncertainty sources .....	94
4.4.3 Further consideration of the categorized uncertainty sources .....	96
4.4.4 Implication of the proposed method for uncertainty analysis in QSAR modeling .....	98
4.4.5 Potential limitations of the developed method and future work.....	99
4.5 Conclusion .....	101
<b>Chapter 5: Conservative Consensus QSAR approach for the prediction of rat acute oral toxicity.....</b>	<b>102</b>
5.1 Introduction.....	102
5.2 Methods .....	104
5.2.1 Data sourcing.....	104

5.2.2 Prediction of the oral rat acute toxicity .....	105
5.2.3 Deriving Conservative Consensus Model (CCM) predictions .....	106
5.2.4 Model predictive accuracy for hazard classification .....	107
5.3 Results.....	108
5.3.1 Applicability domain .....	108
5.3.2 Comparing model predictive accuracy for hazard classification .....	109
5.3.2.1 Agreement with experimental data.....	109
5.3.2.2 Level of conservativeness of the model predictions .....	112
5.4 Discussion and Conclusion.....	114
5.4.1 Consideration of CCM under conditions of uncertainty.....	115
5.4.2 Prioritizing chemicals based on CCM predictions.....	116
<b>Chapter 6: Conclusion .....</b>	<b>119</b>
6.1 Summary of the chapters and research contribution.....	119
6.2 Future work .....	123
6.3. Final reflection.....	106
<b>References.....</b>	<b>127</b>
<b>Appendices .....</b>	<b>149</b>
Appendix A. Supplementary materials associated with Chapter 2 .....	149
Appendix B. Supplementary materials associated with Chapter 3 .....	153
Appendix C. Supplementary materials associated with Chapter 4.....	160

## List of Tables

Table 1.1. OECD principles for (Q)SAR validation and (Q)SAR Assessment Framework.....	7
Table 1.2. Challenges in the application of <i>in silico</i> toxicology methods as reported by ICCVAM. ....	9
Table 2.1. Problem formulation components in thirteen <i>in silico</i> method-related studies that include problem formulation as part of the study. ....	31
Table 3.1. <i>In silico</i> toxicology modeling phases, higher-level assessment components, and definition of the components of relevance for the present study.....	44
Table 3.2. 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 S3.1.....	46
Table 3.3. Information about the five compounds used for the illustration and the experimental, model-predicted, and human-extrapolated LD <sub>50</sub> data. ....	63
Table 3.4. 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 ( <input checked="" type="checkbox"/> ), while an empty box ( <input type="checkbox"/> ) indicates that a GSU is not selected/considered relevant. ....	64
Table 4.1. Steps followed to identify uncertainties implicitly and explicitly expressed in the indicator-containing sentences.....	70
Table 4.2. Sources of uncertainty (arranged in alphabetical order) relatable to practices and features common to <i>in silico</i> toxicology modeling (adopted from Achar et al. (2024a)). ....	71
Table 4.3. Examples of the implicit and explicit uncertainty indicators (bolded in the sentences) identified in the 20 analyzed studies (see Table S1 for the raw data). The data is arranged in the order of the publication numbers presented in Table S4.1. ....	74
Table 5.1. GHS classification criteria for acute oral toxicity.....	92
Table 5.2. Experimental values and conservative predictions from TEST Consensus and CATMoS models and CCM of seven chemicals used for illustration .....	93

Table 5.3. Performance of TEST Consensus, CATMoS, and CCM for the classification of the predicted LD<sub>50</sub> values based on the GHS categories.....97

Table 5.4. Prioritization of chemicals based on CCM estimates. The chemicals are ranked in the order of priority according to highly toxic (LD<sub>50</sub> ≤ 50 mg/kg; orange-shaded), toxic (LD<sub>50</sub> ≤ 2000 mg/kg; blue-shaded), and non-toxic (LD<sub>50</sub> >2000 mg/kg; grey-shaded). .....107

## List of Figures

Figure 1.1. Drivers to the use of non-animal methods in chemical risk assessment (adapted from Browne (2023)). .....	5
Figure 1.2. An illustration of a hazard/risk and uncertainty assessment-driven problem formulation scheme (adapted from OECD (2018a)). .....	13
Figure 2.1. The problem formulation framework utilized in the present paper, outlining the iterative process from problem framing to hypothesis formulation (modified from Sauve-Ciencewicki et al., (2019)). .....	23
Figure 2.2. A simple conceptual model for risk assessment of a cosmetic ingredient X. The complete arrows show the assessment steps, while the dashed arrows show the possible <i>in silico</i> modelling data (in the lower box) required for each step. ....	26
Figure 2.3. Simple hypotheses formulation diagrams showing two possible pathways to skin irritation: Effect- cumulative hypothesis (upper diagram) involving absorption enhancer B and skin irritant A, and Dose-cumulative effect hypothesis (lower diagram) involving absorption enhancer B and irritant A and time and dose as the influencing factors. ....	29
Figure 2.4. A proposed step-step (shown with the complete arrows) process to characterize and address the uncertainty associated with PF components and an iteration (shown with the dashed arrow) required at one of the steps. ....	37
Figure 3.1. A flow chart describing the steps undertaken to develop the framework that categorizes general sources of uncertainty in <i>in silico</i> models for toxicological data gap filling. ....	42
Figure 3.2. The refined GSU (bulleted in the rectangles) resulting from the analysis and iterative categorization of the VRSU. The descriptions of the GSU uncertainty are provided in Table S3.2. 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). ....	53

Figure 3.3. A summary of the principles and elements outlined in the 2023 OECD’s proposed QAF. The elements are bulleted under the principles. ....68

Figure 5.1. GHS categories for the 6,410 compounds classified based upon the experimental LD50 data. ....109

Figure 5.2. Evaluation of the model predictive accuracy for GHS classification of the 6,410 compounds. (a) The match, under-and over-predictions are for each model, as well as (b) how far off each prediction were from the experimental-based GHS category. ....111

## List of Abbreviations

$\alpha_k/K$ -Alpha	Krippendorff's Alpha
3Rs	Replacement, reduction, and refinement
ACToR	Aggregated computational toxicology resource
ADME	Absorption, distribution, metabolism, and excretion
AOP	Adverse outcome pathway
BB	Blood-brain barrier
CASRN	Chemical abstracts service registry number
CATMoS	Collaborative acute toxicity modeling suite
CCM	Conservative consensus model
CI	Confidence Interval
EC	European Commission
ECHA	European Chemicals Agency
EFSA	European Food Safety Authority
EU	European Union
FDA	Food and Drug Administration
GHS	Globally harmonized system
GSU	General sources of uncertainty
ICCVAM	Interagency coordinating committee on the validation of alternative methods
LD <sub>50</sub>	Lethal dose that kills 50% of test animals
LOAEL	Lowest observed adverse effect level
NAMs	New approach methodologies
NICEATM	National Toxicity Program Interagency Center for the Evaluation of Alternative Toxicological Methods
NOAEL	No observed adverse effect level

NRC	National Research Council
OECD	Economic Cooperation and Development
PBTK	Physiologically-based toxicokinetic
PF	Problem formulation
PS	Permeability-surface
QAF	(Q)SAR Assessment Framework
QSAR	Quantitative structure-activity relationship
R <sup>2</sup>	Coefficient of determination
RAAF	Read-Across Assessment Framework
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
RISK21	Risk Assessment in the 21st Century
SAR	Structure-activity relationship
SMILE	Simplified Molecular Input Line Entry System
TEST	Toxicity Estimation Software Tool
TSCA	Toxic Substance Control Act
UN	United Nations
US EPA	United States Environmental Protection Agency
VRSU	Verbatim recorded sources of uncertainty
WHO/IPCS	World Health Organization/International Programme on Chemical Safety

## **Acknowledgements**

First and foremost, I would like to express my gratitude to the Almighty God for His strength and grace. He has enabled me to overcome challenges and achieve this milestone. I am deeply grateful for the incredible opportunity to pursue my Ph.D. at IRES and to work under the supervision of Professor Gunilla Oberg. Thank you, Gunilla, for your support and guidance throughout my Ph.D. journey. Your insightful advice has played a crucial role in shaping my approach to thinking, analyzing, and executing my research. I would also like to sincerely thank Professor Mark Cronin. You provided me with invaluable guidance, perspectives, and comments on my research, which substantially aided this accomplishment. Further, I extend my gratitude to Dr. James Firman, whose insightful feedback and contribution made it possible to achieve this. I feel incredibly fortunate to have had the mentorship of my entire committee (Professor Gunilla Oberg, Professor Mark Cronin, Professor Milind Kandlikar, and Professor Jeroen van der Sluijs) during my Ph.D. Your instructions, inspiration, and constructive advice guided me to this end. I am indebted to my wife, Cecilia, for her support. Knowing that you always had my back made me stay strong and determined even during the most challenging times. Thank you! To my mom, whose unwavering love and prayers guided me throughout this journey, I will forever be grateful. Finally, I would like to thank the Vanier Canada Graduate Scholarship (Vanier CGS) program for the financial support that made my Ph.D. research possible.

## **Dedication**

To my dear son, JJ, this Ph.D. is as much yours as it is mine. Having you during my Ph.D. has been my greatest joy; your presence has been my constant source of strength. Aheri wuoda, wuod Kamagambo!

## Chapter 1: Introduction

### 1.1 Chemical safety and animal testing

Chemicals offer numerous benefits to modern society, with about 95% of all manufactured products (e.g., pharmaceutical, cosmetic and agrochemical products and food additives) estimated to rely on chemicals (Wang et al., 2020). However, the release of chemicals during their lifecycles can lead to negative human and environmental health effects (Naidu et al., 2021; Wang et al., 2020). Indeed, chemical pollution, one of the novel entities elements within the “planetary boundaries” (i.e., the environmental limits within which humanity can operate safely), is recognized for its adverse impacts on other planetary boundaries such as biosphere integrity and climate change (Steffen et al., 2015). Understanding the potential human and environmental health effects of commercially available chemicals is important to support regulatory actions aimed at protecting public and environmental health. However, only a small fraction of these chemicals is considered well-studied and evaluated for their effects – consequently, the potential health impacts of many of them remain unknown or are insufficiently characterized (Judson et al., 2009; National Toxicology Program, 2024). This data gap is widely recognized in the literature (Egeghy et al., 2012; Johnson et al., 2020; Naidu et al., 2021; Judson et al., 2009; Wang et al., 2020; Zhang et al., 2024). For example, Egeghy et al. (2012) estimated that of the 8 million commercially available chemical substances, only about 547,088 chemicals have exposure-related information within the US EPA’s ACToR (Aggregated Computational Toxicology Resource).

Chemical risk assessments provide information about the potential harm posed by chemicals as well as help characterize the nature and scope of regulatory decisions required to address specific human and environmental health concerns. For many years, most of these assessments have relied on animal testing (i.e., the use of vertebrate and nonvertebrate animals to assess chemicals), where risk assessors are then required to extrapolate the assessment outcome to human exposure scenarios (Hackam & Redelmeier, 2006). Animal testing dates back to 1930s where experiments were driven by a desire to improve our knowledge of different diseases (Van Norman, 2019). Presently, animal testing is widely applied in, for example, biomedical and drug/chemical development

research, including chemical toxicology and safety studies, as well as compound screening processes. While these applications have undeniably provided a number of benefits to scientific and medical advancements in terms of drug development, food production, disease treatment, etc. (Akhtar, 2015; Paparella et al., 2020; van der Zalm et al., 2022; Van Norman, 2019, 2020), several concerns have been raised regarding the reliability of animals as the gold standard for informing human health, as well as regarding financial burden and ethical issues associated with animal testing. Some of the factors contributing to these concerns are discussed below in Section 1.1.1.

### **1.1.1 Challenges in successful translation of animal testing data to humans**

Akhtar (2015) described three broad areas to explain why animal testing may fail to reliably inform human health or why animals are poor predictors of human health outcomes: (1) species differences in biology and physiology, (2) discordance between animal models of diseases and human diseases, and (3) the impact of laboratory environment and procedures on study outcomes. These areas are briefly discussed below.

*Species differences in biology and physiology:* Studies have shown that interspecies differences between humans and animal models (e.g., rodents) in terms of physiology, behavior, pharmacokinetics, and genetics limit the reliability of animal models as predictors of human health. For example, Akhtar (2015) reported that rodents, such as mouse models that are widely used to assess human inflammatory diseases could be misleading, given the differences in how the animals and humans respond to inflammatory conditions. Mice are different from humans in terms of the type of genes turned on and off and in the duration and timing of gene expression, with the differences thought to be responsible, at least in part, for the drug failure rates during clinical trials. According to Akhtar (2015), about one in every 150 clinical trials on inflammatory response endpoint fails. Similar studies have shown that several drugs successfully tested on animals without any adverse health effects end up causing significant harm to humans, partly because toxicity in animals was not detected, deemed negligible/non-serious, or acceptable in light of the potential benefits (Cohen, 2017; Federer et al., 2016; Sonawane et al., 2018). In another instance, Debad et al. (2024) note that, unlike humans, rats lack gallbladders, have no emetic response, and are obligate nose-breathers, meaning that the physiological responses of rats to chemical exposure may not

accurately reflect responses in humans. Taken together, these differences lead to questions about how to discern what tests are applicable to humans and which ones are not or how to account for the differences during animal-human toxicity data extrapolation (Van Norman, 2019, 2020).

*Impact of the laboratory environment and procedures on study outcomes:* Laboratory conditions or experimental procedures can exert influence on animal behaviors and physiology that become difficult to control, thus impacting extrapolated outcomes in humans. A review of 37 chemicals by the U.S. National Toxicology Program found that non-carcinogenic toxicities were not reproducible between rats and mice, across sexes, or when compared with historical controls – for example, acute toxicity studies recorded an average positive predictive value of 55.3% for mouse-to-rat toxicity predictions, while chronic studies recorded a positive predictive value of 44.8% (Wang and Gray, 2015). Elsewhere, Hackam and Redelmeier (2006) reviewed 2000 articles (with 76 different animal studies) published between 1980 and 2000 within the field of molecular biology and found that only 37% of the studies were successfully replicated in humans and about 20% of the findings were inconsistent with human health outcomes. Perel et al. (2007) also noted that only 50% of the 221 reviewed animal studies on diseases such as acute ischaemic stroke and neonatal respiratory distress syndrome produced consistent results with human studies. Overall, it seems reasonable to conclude that these differences confound test results, interfere with extrapolation accuracy, and ultimately lower confidence in using animals as predictors of human health (Akhtar 2015; Van Norman, 2019).

*Discordance between animal models of diseases and human diseases:* Another obstacle facing the use of animal models is the lack of sufficient concordance between animal models of diseases and human diseases. For instance, as Cohen (2017) explains, most studies use healthy animals, which do not account for the comorbidities common in the human population, or where such human diseases are simulated, they are induced in animals. The problem with the latter case is that it cannot accurately replicate the complexity of human diseases (Cohen, 2017). An example in cancer research, where, while tumor induction into animals has been the standard way of assessing cancer progression and testing treatments in humans, translation of animal results to humans has not been

successful (Cohen, 2017). Van Norman (2020) explains that this failure could be due to the fact that animal models are limited in accurately reproducing the complex nature of human carcinogenesis. According to the author, this is evidenced by high rates of clinical failure of cancer drugs. In another study, Akhtar (2015) noted mouse models for amyotrophic lateral sclerosis tests produced significantly different results compared to human amyotrophic lateral sclerosis, and despite numerous clinical trials, only Riluzole drug with only marginal benefits has been approved by the FDA.

In addition to the challenges raised above, the use of animal models is further challenged by the fact that animal testing is resource-intensive because of the high costs and lengthy time required to conduct animal tests, which then limits the number of chemicals tested and animals available for a test, etc. According to Judson et al. (2009) and Zhang et al. (2024), this might explain why tens of thousands of industrial chemicals remain untested or have limited toxicity data. In attempts to overcome some of the challenges, the use of non-animal methods is increasingly being developed and favored by scientific communities as well as regulatory authorities, as further discussed below (see Section 1.2).

## **1.2 Non-animal methods**

Here, non-animal methods refer to any methodologies, technologies, or combination of approaches that generate toxicological-related information without using animals (Embry et al., 2014). The field of toxicology is continually evolving, with significant advancements in human biology and techniques for assessing the potential health effects of chemical stressors. To ensure these new initiatives are integrated into the field of toxicology, the US National Research Council (NRC) published a pivotal report over a decade ago titled "Toxicity Testing in the 21st Century" (National Research Council, 2007). This report outlined a future vision for toxicology, with the central emphasis on a long-term strategy designed to leverage new technologies that promote efficient chemical evaluation procedures that minimize the use of animals and animal suffering, as well as reduce the cost and time for chemical assessment (National Research Council, 2007). In other words, a key aspect of this strategy is reducing reliance on animal testing by shifting towards and promoting non-animal methods. These include *in vitro* tests using cultures of cells,

*in silico* methods based on, for example, modeling structure-activity relationship (SAR) and read-across, and *in chemico* approaches (Benfenati et al., 2019). This thesis focuses on *in silico* methods (see further discussion in Section 1.2.1). While methods such as *in silico* toxicology methods might not themselves be new, their integration into regulatory decision-making processes and their role in replacing animal testing represent a growing development in toxicology (van der Zalm et al., 2022).

Non-animal methods have the potential to significantly reduce and (in some cases) replace animal testing, thus contributing to efforts aimed at addressing the challenges discussed under Section 1.1.1. Browne (2023) summed up the potential benefits of using non-animal methods into five key areas (Figure 1.1). (1) Increase throughput – i.e., the methods ensure fast predictions for a large number of chemicals in a high-throughput mode. (2) Increase human relevance – this can be achieved by, for example, integrating human dosimetry modeling information or evaluating human-relevant chemical metabolites by differentiating those relevant to humans from the ones relevant to animals (van der Zalm et al., 2022). (3) Using the best science – here, data generated from non-animal methods can be evaluated against other data to allow the use of the best scientific data to support decisions. (4) Reduce decision time – a key element in the use of approaches like *in silico* methods is their ability to reduce the time required to, for example, screen chemicals for potential hazards; consequently supporting quick decisions that rely on such data. (5) Reduce animal use per study or chemical assessment – this includes complementing animal testing to prioritize chemicals and predict or guide toxicity tests, etc.

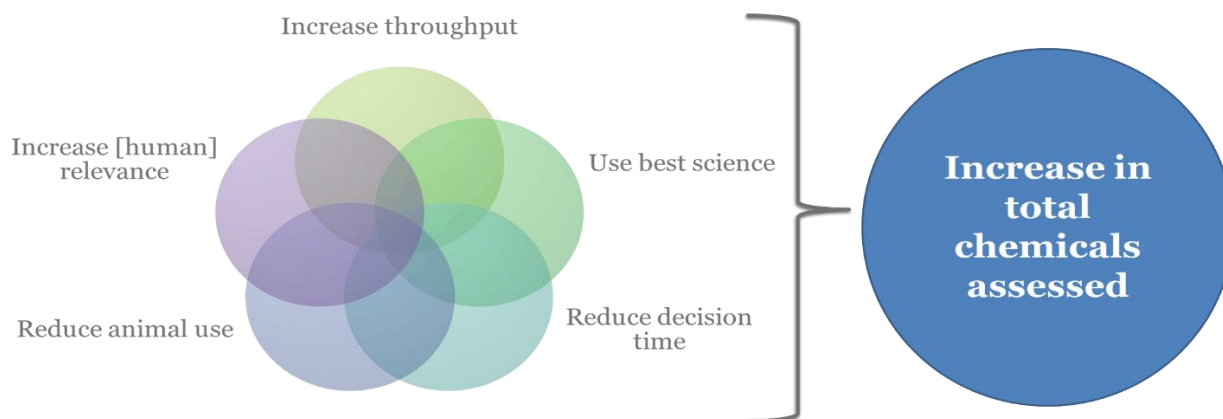


Figure 1.1. Drivers to the use of non-animal methods in chemical risk assessment (adapted from Browne (2023)).

Overall, through achieving the listed benefits, it is anticipated that non-animal methods will be able to address time constraints and ethical issues associated with animal testing, as well as potentially reduce costs involved in such testing (National Research Council, 2007).

### **1.2.1 Application of *in silico* toxicology methods**

*In silico* toxicology methods refer to computer-based techniques used to derive information on hazard or risks of compounds. By leveraging advances in computational power and the benefits required to assess a chemical, *in silico* tools offer a promising complement to *in vivo* and *in vitro* toxicity tests to potentially minimize the need for animal testing (Raies and Bajic, 2016). The European Chemical Agency (ECHA) Guidance on Information Requirements and Chemical Safety Assessment broadly categorizes *in silico* methods into: (1) (quantitative) structure-activity relationship (Q)SAR), (2) grouping approaches (e.g., read-across and category formation), and (3) expert systems (ECHA, 2008). The development as well as application of these approaches are based on the assumption that similar compounds (e.g., in terms of molecular structure) should have similar biological activities (Cronin and Madden, 2010).

QSAR is one of the most commonly applied *in silico* toxicology methods in regulatory science to predict toxicity from chemical structures (Raies and Bajic, 2016). QSARs are designed to predict activity (including toxicity) of substances based on the assumption that a chemical's biological activity is dependent upon certain structural and/or physicochemical parameters (Cronin et al., 2019; Piir et al., 2018). These models are typically developed using techniques such as statistical regression analysis, and they take the form of mathematical equations with three major elements: quantitative measure of chemical property/activity, qualitative or quantitative parameters derived from chemical structure, and an algorithm linking these elements (Nantasenamat, 2020; Piir et al., 2018; Tropsha, 2010). Another type of *in silico* method is category formation. The Organization for Economic Co-operation and Development (OECD) describes a chemical category as substances that are likely to exhibit similar

or consistent physicochemical or toxicological properties due to structural similarities (OECD, 2007). Grouping together chemicals with the same mechanism of action is one way to form a chemical category, meaning that a category can be formed based on a structural alert with a specific mechanism, with the alert expected to be present in both the target and analogue chemicals. The read-across approach is based on the hypothesis that the chemical and biological activity of structurally similar compounds will be similar; therefore, activity of a target compound can be predicted qualitatively or quantitatively by extrapolating toxicological data from analogue chemical(s) belonging to the same category as the target compound (Patlewicz et al., 2013; Schultz et al., 2015).

The central question when applying any *in silico* toxicology method for regulatory purposes is the usefulness of the method, e.g., in terms of the method's practical applicability of its predictions. OECD has developed harmonized principles to facilitate the assessment and documentation of the methods and their predictions – for example, the 2007 OECD QSAR validation principles (OECD, 2007), as well as the principles outlined in the 2023 OECD's proposed (Q)SAR Assessment Framework (QAF) (OECD, 2023) (summary of the guidelines provided in Table 1.1). Exactly how these guidance principles are applied depends on the method in question and the context in which the data generated from the method are to be used. However, none of these principles provide a practical way to identify and report areas of uncertainty associated with the methods. As such, discussion and development of schemes that support identification, assessment, and systematic reporting of uncertainty embedded within the methods would be beneficial to not only support structured regulatory decision-making but also promote transparency in the evaluation of whether the methods and their predictions are fit for purpose (Sahlin et al., 2011). This knowledge gap constitutes the basis of this thesis (see Chapters 2 and 3).

Table 1.1. OECD principles for (Q)SAR validation and (Q)SAR Assessment Framework.

<b>OECD principles for (Q)SAR validation</b>	<b>OECD (Q)SAR Assessment Framework principles</b>
1. a defined endpoint	1. the model input(s) should be correct
2. an unambiguous algorithm	2. the substance should be within the applicability domain of the model

- 
- |   |   |
|---|---|
| 3. a defined domain of applicability                                    | 3. the prediction(s) should be reliable                 |
| 4. appropriate measures of goodness-of-fit, robustness and predictivity | 4. the outcome should be fit for the regulatory purpose |
| 5. a mechanistic interpretation, if possible                            |   |
- 

As discussed in Section 1.1.1, the use of animal testing as surrogates for humans phases a number of challenges. This means that any efforts in the development and application of *in silico* methods mark a step towards the pursuit of more accurate and expedited human-relevant toxicological assessments (Worth et al., 2011a). However, the regulatory acceptance of these methods remains contingent on addressing a number of associated challenges. These challenges are discussed below (see Section 1.2.1.1).

#### **1.2.1.1 Existing challenges in integrating *in silico* methods into the regulatory landscape**

The necessity of updating chemical risk assessment practices at the regulatory level has been recognized by governments, such as Canada, the US and European Union (EU). For example, Canada has adopted laws aimed at reducing reliance on animal testing – exemplified by Bill C-47 (Budget Implementation Act, 2023) under the Food and Drugs Act, which bans cosmetic animal testing (Health Canada, 2023a). Similar laws are seen in the US – for example, Bill S.5002 of the FDA Modernization Act 2.0, which authorizes the use of non-animal testing on the safety assessment of drugs (US Congress, 2022) – and the Directive 2010/63/EU of the EU which emphasizes the need to protect animals used for scientific purposes based on the 3R (replacement, reduction, and refinement) principles (European Union, 2010). However, as regulatory authorities across these jurisdictions attempt to integrate non-animal testing approaches such as *in silico* methods into their risk assessment landscapes, a number of needs and challenges also emerge with respect to the regulatory application of the approaches.

Considering that regulatory decision-making processes ought to be anchored to the principles of protecting public health, it becomes necessary to improve scientific confidence in the *in silico* methods before they are accepted. This means that any *in silico* method must be taken through rigorous evaluation to ensure that it can reliably replicate or potentially improve the results from animal tests in order to provide a solid scientific ground for

regulatory decision-making. Through such evaluation, the commonly mentioned areas of challenges hindering acceptance of the methods must be addressed. Some of these areas of challenges (summarized in Table 1.2) were broadly described in the recent workshop report of the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) (Debad et al., 2024).

Table 1.2. Challenges in the application of *in silico* toxicology methods as reported by ICCVAM (Debad et al., 2024).

Broad areas of challenges	Description of the challenges
Validation	It is challenging to validate these approaches, especially those that cover endpoints that are typically not covered by animal testing protocols for which insufficient or no relevant data are available for use in the validation. Such a lack of reliable benchmarks makes it difficult to determine the relevance of their prediction results.
Standardization	Lack of standard protocols, frameworks, and quality-control procedures to support the assessment of the robustness and reliability of specific approaches within and across studies. Consequently, this makes it difficult to compare predicted data across regulatory submissions.
Transparency	Here, transparency entails a clear and careful documentation of the procedures, scientific rationale, uncertainties, and limitations associated with a specific approach and a demonstration of whether an approach is fit for purpose. A lack of transparency contributes to a lack of data sharing, methods, and results and makes it difficult to delineate the context of use or allow peer-review by the broader scientific community.
Biological relevance	<i>In silico</i> methods are expected to produce results that correlate with known outcomes from animal tests or <i>in vivo</i> -human data, as well as predict toxicologically relevant effects endpoints. However, challenges exist in establishing whether predicted results have

---

meaningful implications or relevance to human health and whether the predicted doses align with realistic human exposure levels.

---

Generally speaking, the areas of challenges listed in Table 1.2 revolve around what several researchers have described as sources of uncertainty or factors contributing to limited knowledge/data gaps in the development and application of *in silico* methods for chemical risk assessment (Ball et al., 2014; Belfield et al., 2021; Benford et al., 2018; Blackburn & Stuard, 2014; Cronin et al., 2019; Patlewicz et al., 2013; Pestana et al., 2021; Pham et al., 2019; Schultz et al., 2015; Worth, Fuart-Gatnik, et al., 2011). This means that any consideration to address uncertainty issues within *in silico* methods and their predictions can directly or indirectly tackle the challenges. This thesis, therefore, approaches the existing challenges from the perspective of uncertainty (further discussed in Section 1.2.1.2).

#### **1.2.1.2 Considerations of uncertainty to support regulatory application of *in silico* methods**

There seems to be a general agreement in both scientific and regulatory communities that uncertainty has a negative impact on the uptake/acceptance of *in silico* methods for regulatory applications (Ball et al., 2014; Belfield et al., 2021; Benford et al., 2018; Blackburn & Stuard, 2014; Cronin et al., 2019; Patlewicz et al., 2013; Pestana et al., 2021; Pham et al., 2019; Schultz et al., 2015; Worth, Fuart-Gatnik, et al., 2011). Here, uncertainty is defined according to EFSA (2018): “limitations in the knowledge available to assessors at the time an assessment is conducted and within the time and resources available for the assessment” – examples include uncertainties related *in silico* methodological quality, data, and relevance (Worth et al., 2011a). The US EPA (2015) distinguishes between uncertainty due to incomplete knowledge or lack of data (i.e., epistemic uncertainty) and uncertainty due to inherent randomness in data or prediction outcome (i.e., aleatoric uncertainty). The former can be reduced or eliminated with more/better data, whereas the latter cannot be reduced (US EPA, 2015). While both types of uncertainty are important in the fit-for-purpose evaluation of *in silico* models, this thesis only focuses on epistemic uncertainty, as it is regarded to be more problematic in modeling exercises (Cronin et al., 2019; Sahlin et al., 2013).

Besides the issue of uncertainty, which presents a barrier to the uptake of *in silico* methods, lack of transparency about the uncertainty makes it difficult to qualify or quantify the extent to which the uncertainty impacts conclusions drawn from a modeling exercise (Patterson and Whelan, 2017). According to EFSA (2018), addressing this issue of transparency requires a clear indication of what sources of uncertainty are present in a study, followed by characterization of its overall impact on the prediction outcome. In other words, sources of uncertainty should be reported and characterized in a clear and unambiguous manner to not only aid in the assessment of their potential impacts but also guide decisions about whether the level of uncertainty in question is acceptable in a defined decision context.

Scholars such as Alexander-White et al. (2022) and Patlewicz et al., (2015) note that, in addition to analysis and transparent identification and reporting of uncertainty, it is also necessary to develop frameworks or methods that facilitate a better understanding of where different uncertainties reside, as well as guide strategies focused on analyzing uncertainty and addressing them. Presently, however, such frameworks consider sources of uncertainty on a case-by-case basis based on, for example, the prediction purpose of a method or endpoint in question; thus, they do not provide a broader picture of sources of uncertainty across different *in silico* methods. For example, Blackburn and Stuard (2014) proposed a questionnaire that identifies sources of uncertainty in the context of read-across predictions, while the framework proposed by Cronin et al. (2019) also only targets areas of uncertainty, bias, and variability in QSARs. This isolated conceptualization of frameworks presents a research gap and an opportunity to broaden the consideration of sources of uncertainty across different *in silico* prediction contexts by integrating different perspectives around uncertainty within and across methods. Indeed, the importance of such a consideration has been underlined by different scholars as a crucial step to promoting transparent evaluation of the validity of the methods and adequacy of their predictions (Belfield et al., 2021; Cronin et al., 2019; Kirchner et al., 2021; Parish et al., 2020). Chapter 2 of this thesis attempts to address the research gap.

### 1.2.1.3 Addressing uncertainty in *in silico* predictions through conservative assumptions

It is widely acknowledged that better decisions regarding chemical safety are made when relevant uncertainties are incorporated into the decisions (EFSA et al., 2018). This is particularly important for decisions aimed at protecting human and environmental health, for which uncertainty exists, yet the available scarce information has to be used to make the decisions (WHO/IPCS, 2018). Within the context of *in silico* methods, the appropriate approach to incorporate uncertainty into decision-making involving their predictions might depend on how urgent the decision is to be made, the “severity” of the decision (guided by the “principle of proportionality”) and the possible implication of basing the decision on a “wrong” conclusion (guided by the “principle of caution/conservativeness”) (WHO/IPCS, 2018). The key question here is whether the totality of information obtained from a model prediction is sufficient to support a decision and, if not, whether the level of uncertainty in the prediction is acceptable in a given decision context.

As discussed earlier within Section 1.2.1.1, uncertainty associated with *in silico* methods contributes to the difficulty in producing reliable toxicity estimates. This difficulty has led to the emergence of the use of conservative approaches to account for uncertainty, for example, in the application of QSAR models in situations where experimental data are lacking or limited (Bishop et al., 2024; Burden et al., 2016; Graham et al., 2021). In this context of QSAR, the term “conservative” is used to refer to the acceptance of erring on the safe side based on the general sense of taking caution when there is insufficient assurance of accuracy in predictions (EFSA et al., 2018). While it is ultimately up to the end user of *in silico* predictions (e.g., regulatory authority) to define whether information obtained from a model prediction is sufficient or the level of uncertainty in the predictions is acceptable, from safety point of view, it is generally acceptable that conservative strategies can be applied when the uncertainty is considered high or when limited information is available about a compound (e.g., lack of comparative experimental data) (Cousins et al., 2016; Hansen et al., 2007; Health Canada, 2000; Tosun, 2013; Verdonck et al., 2005; WHO/IPCS, 2018). In Chapter 5 of this thesis, using QSAR as an example of *in silico* toxicology methods, I evaluate the use of conservative assumptions to account for uncertainty in the prediction of acute oral toxicity of organic chemical compounds.

#### 1.2.1.4 Incorporating problem formulation into *in silico* toxicology predictions

OECD (2018a) emphasizes that a clear formulation of the problem needs to precede efforts aimed at applying risk assessment methods (including *in silico* methods) or addressing challenges associated with their application, including the issue of uncertainty. For such efforts to be useful, factors such as the application scenario, the relevance of the proposed application, and the specific information requirements need to be contextualized. In short, chemical risk assessment and uncertainty analysis depend on the problem formulation (PF), with PF defined as a process that describes assessment questions and components within the hazard and risk assessment phases and the output, which leads to identification of uncertainty sources (see the illustration in Figure 1.2).

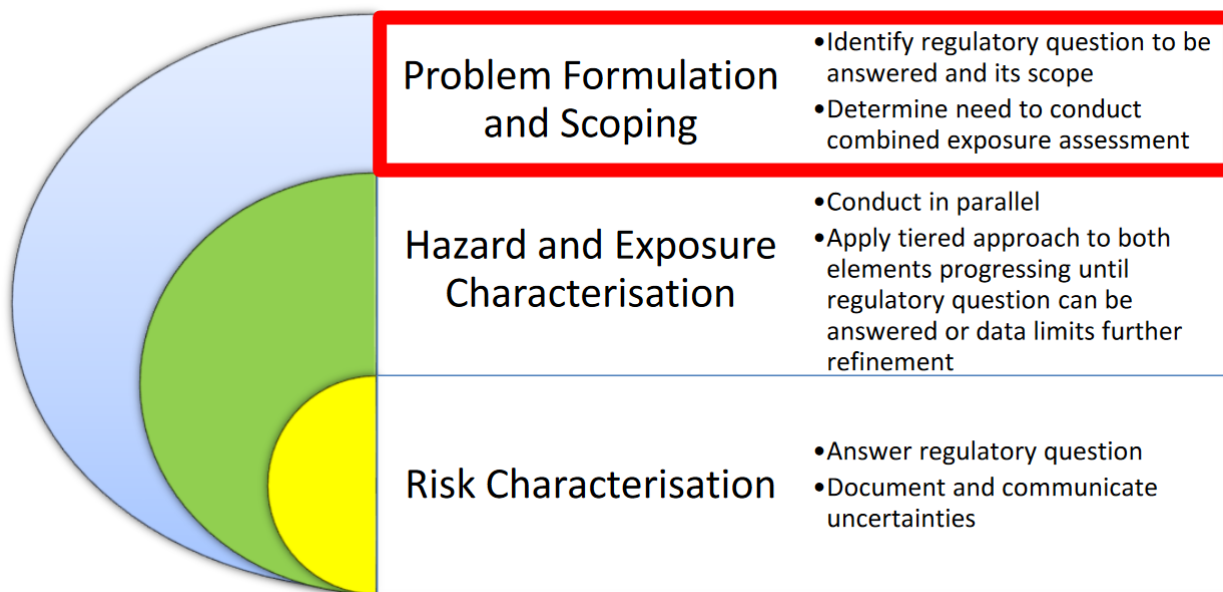


Figure 1.2. An illustration of a hazard/risk and uncertainty assessment-driven problem formulation scheme (adapted from OECD (2018a)).

Aligning with the OECD recommendation, the Risk Assessment in the 21<sup>st</sup> Century (RISK21) roadmap considers PF as a crucial step in the application of non-animal methods like *in silico* methods – it guides a systematic identification of the major factors for consideration in the application of a method and establishes a plan for

collecting information to guide a prediction process (Embry et al., 2014). Alternatively, PF can guide the demarcation of problems to justify the application of a method, such as the need to expedite screening of a large number of chemicals. In other words, the use of PF within *in silico* methods is based on the premise that each assessment step of a planned risk/hazard assessment exercise should be directed towards deriving or obtaining information that addresses a defined problem, with the ultimate aim of ensuring a method and its predictions are fit for purpose (Embry et al., 2014; WHO/IPCS, 2018).

While PF is crucial within the context of *in silico* predictions, there is presently a lack of a shared concept of PF or discussion of potential uncertainty in its conceptualization. In my view, such a lack of a harmonized concept of PF makes it difficult to, for example, determine the extent to which risk assessment and management options of particular chemicals are required. In Chapter 2 of this thesis, therefore, I synthesize the current knowledge on PF within the field of *in silico* methods for chemical risk assessment and develop a harmonized PF framework. Specifically, the discussion in the chapter sets out to answer the following fundamental questions: What do we need to know about PF in the context of *in silico* toxicology methods and what components are pertinent to it? What uncertainty issues should be considered to ensure that an *in silico* PF is fit for a particular purpose?

Taken together, Chapters 2-5 of this thesis aim to address the following overarching questions: What components and sources of uncertainty are pertinent for *in silico* problem formulation and modeling phases, and how might these sources be systematically categorized, analyzed, as well as be addressed, particularly in the context of QSAR prediction of chemical toxicity? This thesis explores these questions through the research objectives in Section 1.3.

### **1.3 Research objectives**

The field of *in silico* toxicology methods for chemical risk assessment is rapidly growing through increase in the number of models that cover different toxicological endpoints as well as improvement of the predictive ability of existing models. Evaluating this growing field requires consideration of several associated issues, including the

concept of PF, uncertainty, and frameworks to provide clear and structured means of synthesizing complex problems like uncertainty. The overarching aim of this thesis is to facilitate analysis and accounting for uncertainty associated with *in silico* toxicology methods, with a major focus on QSARs. This is achieved through four separate but interrelated studies:

1. <sup>1</sup>Chapter 2. Aim: to develop a problem formulation framework for the application of *in silico* toxicology methods in chemical risk assessment.

In this study, a systematic analysis of PF-relevant conceptual questions and components that are addressed in the general risk assessment literature is carried out. Alongside this analysis, recent peer-reviewed publications on *in silico* toxicology methods are examined to identify components integral to PF that are highlighted in these publications. A systematic combination of the concepts and components from the general risk assessment and *in silico* toxicology literature is done to develop a PF framework. In light of the developed framework, potential areas of uncertainty are highlighted by considering instances where particular components might be missing or implicitly described within the framework.

2. <sup>2</sup>Chapter 3. Aim: to develop a framework for categorizing sources of uncertainty in *in silico* toxicology methods for chemical toxicity predictions

In this study, peer-reviewed publications on *in silico* toxicology methods are reviewed to identify sources of uncertainty discussed in the publications. The scheme proposed by Belfield et al. (2021) is then drawn upon to categorize the identified sources of uncertainty. Similar sources of uncertainty are then systematically combined into what is here called “general sources of uncertainty” which are categorized through an iterative process using the scheme by Belfield et al. (2021). Lastly, the practical application of the developed framework is illustrated via

---

<sup>1</sup> Chapter 2 has been published in Achar, J., Cronin, M. T. D., Firman, J. W., & Öberg, G. (2024). A problem formulation framework for the application of *in silico* toxicology methods in chemical risk assessment. *Archives of Toxicology*. <https://doi.org/10.1007/s00204-024-03721-6>.

<sup>2</sup> Chapter 3 has been published in Achar, J., Firman, J. W., Cronin, M. T. D., & 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, 105737. <https://doi.org/10.1016/j.yrtph.2024.105737>

a case study, and the relevance of the framework is discussed within the OECD's QSAR Assessment Framework (QAF).

3. <sup>3</sup>Chapter 4. Aim: to develop a method that allows for systematic and transparent accounting for implicit and explicit uncertainties in QSAR modeling of chemical toxicity

Using neurotoxicity as an example of a toxicological endpoint, this chapter identified implicit and explicit uncertainty indicators expressed in peer-reviewed papers on the QSAR modeling of neurotoxicity endpoint. The identified indicators are then used to identify implicit and explicit uncertainties expressed within the indicator-containing sentences. To systematically categorize the identified uncertainties, the general sources of uncertainty established within the framework developed in Chapter 3 were used, allowing me to estimate the frequencies of the uncertainty sources from the analyzed studies.

4. <sup>4</sup>Chapter 5. Aim: to assess the performance of a consensus approach of QSAR models against the models individually for the prediction of a conservative oral rat acute toxicity of organic compounds.

In this assessment, two QSAR models – Toxicity Estimated Software (TEST) and Collaborative Acute Toxicity Modeling Suite (CATMoS) were applied. Additionally, oral rat LD<sub>50</sub> experimental data relating to 8,186 organic compounds were sourced from the literature, which were curated to a final list of 6,410 compounds. Using Chemical Abstract Service Registration Numbers identifiers (CASRN) as the input, the LD<sub>50</sub> of the 6,410 compounds were predicted in TEST and CATMoS. The minimum prediction concept for conservativeness was applied to select a lower (more conservative) LD<sub>50</sub> value for each chemical across TEST and CATMoS predictions as the consensus LD<sub>50</sub> value. Thereafter, the compounds were assigned into one of the Global Harmonized System (GHS) categories based upon their experimental, TEST, CATMoS and consensus LD<sub>50</sub> values. The predictive accuracy of TEST, CATMoS and the consensus approach were then estimated based on the predicted vs. experimental category comparison

---

<sup>3</sup> Chapter 4 has been published in Achar, J., Firman, J. W., Tran, C., Kim, D., Cronin, M. T. D., & Öberg, G. (2024). Analysis of implicit and explicit uncertainties in QSAR prediction of chemical toxicity: A case study of neurotoxicity. *Regulatory Toxicology and Pharmacology*, 154, 105716. <https://doi.org/10.1016/j.yrtph.2024.105716>.

<sup>4</sup> Chapter 5 is present within Achar, J., Cronin, M. T. D., Firman, J. W. Conservative Consensus QSAR approach for the prediction of rat acute oral toxicity. The work has been submitted and is under review.

## Chapter 2: A problem formulation framework for the application of *in silico* toxicology methods in chemical risk assessment

### 2.1 Introduction

*In silico* toxicology models (e.g., quantitative structure-activity relationship (QSAR) and read-across) form part of a broader collection of non-animal testing approaches that aim to reduce the reliance on animal testing and improve the prediction of potential harm caused by chemicals (Patlewicz et al., 2013; Pradeep et al., 2020; Schultz et al., 2015; Wang et al., 2012). The basic assumption is that the activity of a substance is relatable, either qualitatively or quantitatively, to its molecular structure. As such, compounds displaying similarity in terms of chemical features or properties will additionally be anticipated to exhibit likeness in toxic profiles (Cronin and Madden, 2010; Cronin et al., 2013; Enoch, 2010). *In silico* approaches are particularly important in the risk assessment of chemicals such as cosmetic ingredients within jurisdictions including the European Union (EU) (Regulation (EC) No. 1223/2009) (European Commission, 2016)), Canada (Bill S-5, Clause 16.1 (Government of Canada, 2023)), and the United States (TSCA 4(h)(2)(C) (US EPA, 2018)), where animal testing is prohibited or under consideration for prohibition for such use. However, the uncertainty associated with the model predictions is often referred to as one reason for low confidence and regulatory acceptance of *in silico* model predictions. Uncertainty in such predictions may arise from, amongst other factors, concerns regarding the quality and appropriateness of the training data, the extent of chemical applicability domain, and the interpretability of the relationship between input features and output (Blackburn and Stuard, 2014; Parish et al., 2020; Schultz et al., 2019). Accordingly, it is often advised that these methods be applied alongside other non-animal testing approaches (e.g., *in vitro* tests) to complement the weight of evidence generated by them (Gautier et al., 2020). It has been postulated that formulation of frameworks and guidelines making it possible to systematically and transparently identify the many possible sources of these uncertainties would, in turn, increase the confidence in the utility of *in silico* toxicology methods (Alexander-White et al., 2022b; Patlewicz et al., 2015). Several initiatives are thus underway to support the development of such frameworks to improve the consistency, quality, rigour,

and reliability of chemical risk assessment procedures. For example, the 2017 European Chemicals Agency's Read-Across Assessment Framework (RAAF), which aims to facilitate the development of a consistent, structured, and transparent read-across review process (European Chemicals Agency, 2017). The US Environmental Protection Agency (US EPA) also recently launched a plan to develop a scientific confidence framework to evaluate the quality, reliability, and relevance of non-animal testing approaches, including *in silico* methods for regulatory chemical risk assessment (US EPA, 2021). While considerable attention has been paid to identifying sources of uncertainty relating to the various phases of model construction and application (Ball et al., 2014; Blackburn and Stuard, 2014; Cronin et al., 2019; Escher et al., 2019; Johnson et al., 2022; Patlewicz et al., 2015; Pestana et al., 2021; Pham et al., 2019; Rathman et al., 2018; Schultz et al., 2015, 2019; Viceconti et al., 2021), comparatively little focus has been dedicated towards addressing uncertainty liable to arise during problem formulation (PF).

Ideally, the first step in the hazard or risk assessment of chemicals is to formulate the problem through a systematic and iterative process aimed at identifying and defining factors critical to the assessment (Devos et al., 2019; Embry et al., 2014). When it comes to assessing the potential for harm posed by chemicals, it is argued that the PF ought, for example, to incorporate stages covering the characterization of the scope and context of the assessment, the identification of research needs, the development of a conceptual model and the formulation of a hypothesis (Devos et al., 2019; Embry et al., 2014; Paoli et al., 2022; Raybould 2006; Sauve-Cienciewicki et al., 2019; Solomon et al., 2016; Tepfer et al., 2013; Wolt et al., 2010). The US EPA introduced the concept of PF to risk assessment in 1998, applying it within an ecotoxicological setting (US EPA, 1998). Its importance within the field is increasingly emphasized and endorsed, not only by individual scientists (Parish et al., 2020; Raybould, 2006; Sauve-Cienciewicki et al., 2019; Solomon et al., 2016; Tepfer et al., 2013; Wolt et al., 2010) but also by regulatory agencies, research organizations and international bodies. Examples of these include the US EPA and the European Food Safety Authority (EFSA) (Devos et al., 2019; US EPA, 2016), the National Research Council of the National Academy of Sciences (NRC) (Bette et al., 2013), and the Organization for Economic Cooperation and Development (OECD) (OECD, 2019). Several recent studies applying *in silico* methods or discussing the methods more generally, have also emphasized the need to include PF as the first step in the development and application of the methods

for chemical risk assessment (Alexander-White et al., 2022; Escher et al., 2019; Ouedraogo et al., 2022; Parish et al., 2020; Raybould 2006; Reynolds et al., 2021). Essentially, *in silico* toxicology methods such as QSAR and read-across differ from other non-animal testing approaches (e.g., *in vitro* tests) – for example, with regards to the complex mathematical tools and algorithms, big data, and model parameters used. This suggests the need for producers of model output (e.g., model users) to have access to a framework that allows them to define a context-specific PF that covers the complexities unique to *in silico* methods. There is, however, no general agreement on what a PF for studies applying *in silico* toxicology methods should include to strengthen the utility of such a PF and the related method itself. The lack of agreement makes it difficult to identify potential weaknesses in a PF, such as when particular components are missing. A missing central component may lead to an inadequately formulated problem, which in turn may result in an inadequate specification of risk concerns or provide insufficient clarity regarding the applicability domain of a model or scope of model predictive output. Accordingly, agreement on what a PF for an *in silico* toxicology method should include has the capacity to reduce the associated uncertainty and thus enhancing its utility.

Through providing an explicit and systematic evaluation of appropriate concepts, this study aims to contribute to the development of a PF framework relevant to the application of *in silico* methods for chemical toxicity prediction. This was performed by sourcing and examining a series of recent publications within which PF is considered in the predictive toxicology context. Components integral to the PFs in these studies – such as the endpoints addressed, the pathways of chemical exposure covered and the scope of model use intended – were analyzed in light of PF processes, as described in broader risk assessment literature [e.g., (Devos et al., 2019; Nickson, 2008; Paoli et al., 2022; Raybould 2006; Sauve-Cienciewicki et al., 2019; Solomon et al., 2016; Wolt et al., 2010; World Health Organization/International Programme on Chemical Safety (WHO/IPCS) (WHO/IPCS, 2018)]. Subsequently, these components were grouped into appropriate component categories. Once complete, I set out to answer the questions: what PF components should be considered when developing PF for *in silico* toxicology methods, and how might exclusion or implicit description of the PF components introduce uncertainty into the method's PF?

## **2.2 Materials and methods**

### **2.2.1 Identifying PF components in the general risk assessment literature**

I identified PF components described in the general risk assessment literature. Relevant documents, i.e., those describing a range of higher-level PF conceptual components potentially relevant to the application of *in silico* toxicology methods, were identified through a search in the Web of Science using two broad keywords and Boolean: (topic)"problem formulation" AND "risk assessment", identifying 221 papers. The titles and abstracts of the identified papers I skimmed to identify 12 relevant peer-reviewed papers. Three relevant grey literature sources (i.e., the OECD (OECD, 2019), the US EPA (US EPA, 2016), and the World Health Organization/International Programme on Chemical Safety (WHO/IPCS, 2018)) were also identified after skimming the reference list of the 12 papers, rendering a total of 15 papers (see Table S2.1). A content analysis (Tracy, 2018) of the 15 documents was carried out to identify higher-level PF conceptual PF components discussed in them.

### **2.2.2 Formulating a general PF framework**

One of the 15 papers (Sauve-Ciencewicki et al., 2019) explores and formalizes PF concepts. I decided to use these general concepts as representations of higher-level PF components, as I found that these concepts cover a considerable amount of the component information mentioned in the other 14 publications. For example, problem framing, as described by Sauve-Ciencewicki et al. (2019), includes defining whether an assessment is intended for hazard or risk analysis (Felter et al., 2021), what qualifies as harm (Raybould 2006; Viceconti et al., 2021), potential chemical exposure scenario (Baltazar et al., 2020; Escher et al., 2019), and scientific questions to be addressed (Paoli et al., 2022). Guided by the discussions in the other 14 publications in Table S2.1, I expanded the framework by Sauve-Ciencewicki et al. (2019) from two higher-level components – problem framing and problem exploration – to four higher-level components – problem framing, problem exploration, conceptual model, and hypothesis formulation, as each of these need to be considered as distinct phases of PF (Devos et al., 2019; OECD, 2019; Solomon et al., 2016; US EPA, 2016; Wolt et al., 2010).

### **2.2.3 Applying the PF framework to *in silico* toxicology methods**

To apply the PF framework outlined in 2.1. to *in silico* toxicology methods, I identified publications on *in silico* toxicology methods that describe PF as part of the method applications. This was done through a literature search in the Web of Science using the following broad keywords and Booleans: (topic) "*in silico*\*" OR new approach methodologies OR NAMs OR non-animal testing OR alternative to animal testing OR read-across OR QSAR OR Comput Toxicol AND (all fields) "problem formulation." Out of the 112 publications identified, 13 papers (see Table S2.2) were selected based on the following criteria: peer-reviewed, relating to *in silico* methods and describing PF associated with *in silico* toxicology methods. These papers were analyzed to identify PF components described in them, whereafter the higher-level components were identified under section 2.2.1. were discussed in light of these components. In so doing, I acknowledge that it is possible that the procedure applied here might have led to some components in the *in silico* toxicology methods literature not being captured. However, since the conceptual breadth of the framework was not based on all possible components present in the literature, I considered the components identified in this section to be sufficient for the study's discussion.

## **2.3 Results and discussion**

### **2.3.1 PFs for *in silico* toxicology methods**

In chemical risk assessment, the problem at hand is to decide whether the potential harm posed by a chemical within a given scenario is sufficient to warrant concern (Wolt et al., 2010). In order to improve clarity and reduce uncertainty as to whether or not adverse effects could realistically arise from exposure to the target substance, the broader PF literature emphasizes that three higher-level, conceptual and context-specific questions must be addressed:

1. What must happen for harm to occur?
2. What is the likelihood of harm?
3. Is there a reasonable pathway to harm?

Hypotheses drawn from answers to these questions then form a basis to identify specific PF components relevant to the assessment (Sauve-Ciencewicki et al., 2019). In other words, to define the problem, it is necessary that the

problem is first framed and explored (Sauve-Cienciewicki et al., 2019) – then that the associated research needs are identified (OECD 2019), and then subsequently that a conceptual model is developed detailing the nature of the variables (i.e., descriptors and endpoint) incorporated (Solomon et al., 2016). Subsequently, this may then guide the formulation of a specific causal pathway towards toxicity (Devos et al., 2019). When analyzing the 13 papers identified under 2.2.3. (Table S2.2), I found that these aspects are not included in these papers, even though they specifically address PF in relation to *in silico* toxicology methods. This clearly reinforces the need for a PF framework for *in silico* toxicological methods that enables users of *in silico* models to answer the three central questions outlined above and identify context-specific PF components.

Similar to Sauve-Cienciewicki et al. (2019) and others (Devos et al., 2019; Embry et al., 2014; Raybould 2006; Solomon et al., 2016; Wolt et al., 2010), I interpret PF as an iterative process that begins with problem framing and ends with hypothesis formulation (Figure 2.1). Between, it proceeds through phases including the evaluation of available data and information, the determination of a preliminary understanding of potential harm, the identification of research needs, and the development of a conceptual model. The five key stages (problem framing, problem exploration, research needs, conceptual model development, and formulation of hypothesis), alongside the connectivity present between them, are illustrated in Figure 2.1. With the exception of "research needs", aspects relevant to each of these stages are subsequently discussed below. The discussion is grounded in reference to the recognition and description of potential uncertainty liable to manifest in *in silico* toxicology methods.

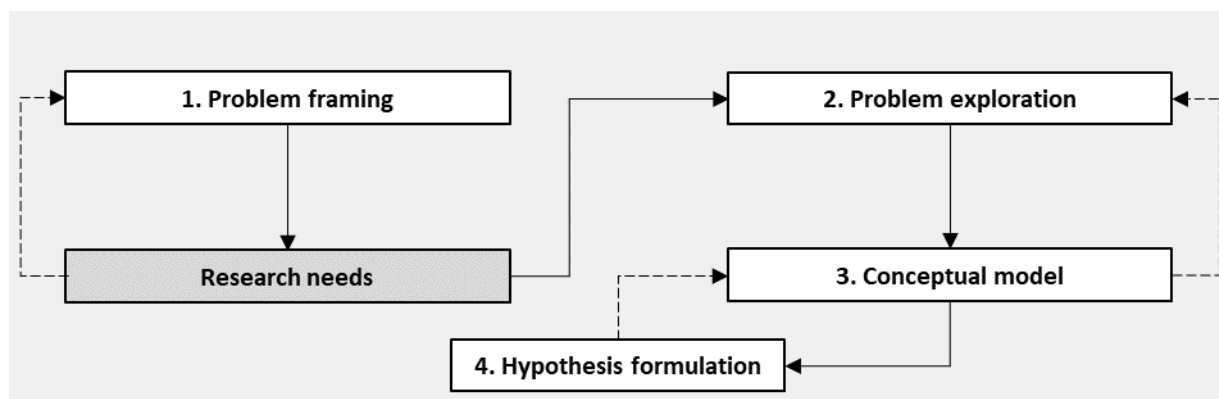


Figure 2.1. The problem formulation framework utilized in the present paper, outlining the iterative process from problem framing to hypothesis formulation (modified from Sauve-Cienciewicki et al., (2019)).

### 2.3.2 Applying higher-level conceptual components for *in silico* toxicology methods

#### 2.3.2.1 Problem framing

In the application of *in silico* toxicology methods, the framing of a problem begins when questions are raised about whether the methods are called for and, if yes, whether a selected method is suitable for a specific use case scenario, such as to predict an adverse effect that may be posed by a chemical of concern (particularly under plausible exposure scenarios) (Sauve-Cienciewicki et al., 2019).

As *in silico* methods are commonly used as data-filling techniques whose main purpose (e.g., during chemical risk assessment, classification, or prioritization) is to generate new or additional data for data-poor chemicals (Cronin and Madden, 2010; Pastoor et al., 2014; Wang et al., 2012), problem framing may initially occur at the development stage of a chemical or drug compound, where screening is done to identify and eliminate potentially toxic properties. Alternatively, new or existing substances may be screened to determine whether or not a further risk assessment is required (Wadood et al., 2013).

#### 2.3.2.2 Problem exploration

The second phase, problem exploration, includes the identification and organization of relevant knowledge and knowledge gaps, with the goal of developing a conceptual model and formulating a hypothesis. For *in silico* methods assessing the potential harm posed by chemicals such as cosmetic ingredients (e.g., coumarin (Baltazar et al., 2020)), exploring the problem (e.g., with regards to the potential of harm) could, for example, include

looking beyond the concentration of the ingredient in a product and furthermore considering use-related factors such as frequency of use, inter-individual variations in use frequency and amount encountered per use. In addition, it would extend to consideration of potentially reactive metabolites (Baltazar et al., 2020). Taking these data into account, the problem exploration phase leads to a more in-depth evaluation of whether a specified model is adequate for addressing the intended prediction problem. Take a physiologically-based toxicokinetic (PBTK) model as an example. It is necessary for a modeler to explore whether the PBTK model is suitable to predict, for example, the dose of coumarin that is causally linked to a specific toxic response in a particular organ (e.g., human lungs and heart). This exploration may include asking (1) whether the model structure reflects and can incorporate chemical-specific information (plasma protein binding, blood partition coefficients, molecular weight, solubility, hydrophobicity, etc.) and physiological information (e.g. blood flow and organ volumes) necessary for the prediction, and (2) whether the model is adaptable to predict different exposure scenarios specific to coumarin (Baltazar et al., 2020). In addition, exploration will include considering how the acceptability of the PBTK prediction results might be evaluated or how the prediction results might fit into the overall weight of evidence decisions regarding the toxic effects of coumarin.

A lack of inclusion of particular information that might give deeper insight into a problem may introduce uncertainty in understanding such a problem, especially for complex problems (e.g., reproductive and developmental toxicity) whose accurate prediction depends on the levels of details (e.g., molecular descriptors and mechanistic characteristics) included in a model (Solomon et al., 2016). A study by Low et al., (2011) explains this point, where, the poor predictive performance of a QSAR model predicting hepatotoxicity of a collection of pharmaceuticals such as acetaminophen was attributed to the model's failure to account for the influence of reactive metabolites (e.g., within acetaminophen). Accordingly, the authors suggested that such factors must be explored during the design of such QSAR model.

### 2.3.2.3 Conceptual model

Several scholars (Devos et al., 2019; Sauve-Cienciewicki et al., 2019; Solomon et al., 2016; Wolt et al., 2010) hold that a conceptual model should be iteratively developed during the framing and exploration phases (see Fig. 2.1). In theory, such a conceptual model should help in the development of testable hypotheses and operational strategies to enable data acquisition and the prioritization of information. This in turn leads to establishing the structural representation of an *in silico* model, which involves defining the model system boundary, variables, parameters and assumptions, and relationships among variables, between input and variables and between variables and output (Walker et al., 2003). Establishing the structural representation of an *in silico* model also includes clarifying the strengths and limits or suitability of the model for predicting a defined problem in a specific use case scenario (Walker et al., 2003). A conceptual model may also be held to be a useful tool when communicating the nature of a problem to stakeholders outside the PF team.

In practice, a conceptual model may take forms such as flow charts, simple statements, or diagrams (Wolt et al., 2010). For illustrative purposes only, I use a simple diagrammatic hypothetical conceptual model intended for a QSAR risk prediction of cosmetic ingredient X in humans (Fig. 2.2) to explain this. *In silico* methods like QSAR are particularly important in this illustration, as in the absence of experimental data, the methods are often applied to provide information on cosmetic ingredients through, for example, hazard identification (European Commission 2016).

In the case illustrated in Figure 2.2, the starting point is to identify the target population being investigated, especially, when setting out to estimate realistic exposures based on, for example, inter-individual variability or frequency and amount of exposure on a population scale. The population can be identified by common characteristics, such as age, gender, consumers of a given country, or a population with a unique susceptibility to the chemical (Hall et al., 2007). The next step involves outlining possible sources of the ingredient (from shampoo, facial moisturizer, body lotion, etc.) that need to be considered. Ideally, a comprehensive QSAR prediction should consider all sources of the substance X (to facilitate aggregate exposure estimation using co-use scenarios), taking into account all exposure routes and potential effects on the exposed individuals. In practice, however, it is not

uncommon for modelers to establish a boundary to limit the scope of the prediction to reduce the level of complexity of the model and its prediction, make simpler the interpretation of the model output, or take into account specific regulatory considerations (OECD, 2018a). The conceptual model is then expanded by adding all possible exposure pathway scenarios to simulate the exposure magnitude, which includes dermal, inhalation, and/or ingestion exposures. Upon exposure, toxicokinetic or toxicodynamic fates of the chemical are considered.

The next step in this conceptualization process (shown by the dashed arrows) is to identify the potential data elements (or parameter data) defining each entity being considered for the prediction of the toxicokinetic and/or toxicodynamic fates of the chemical X. The data should reflect, among others, the target endpoint, specific chemical exposure pathway(s), as relevant to each entity identified. In other words, to successfully use the QSAR model in this prediction context, a modeler needs to consider the relevance and reliability of the data, suitability of model structure with respect to the data or any specific prediction question asked (e.g. which exposure scenario and chemical mechanisms are being predicted?) and consider the sensitivity of the model to the anticipated input parameters.

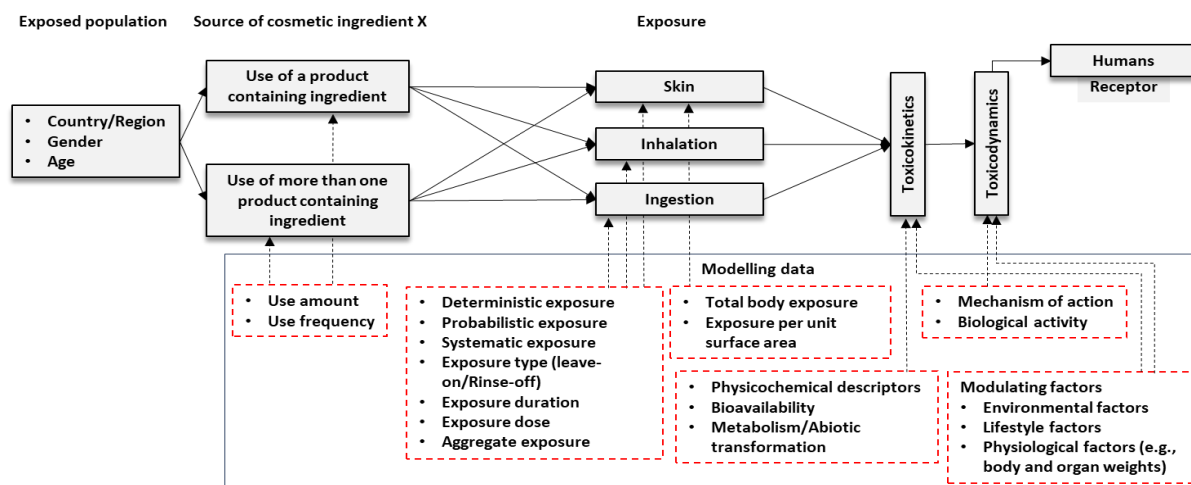


Figure 2.2. A simple conceptual model for risk assessment of a cosmetic ingredient X. The complete arrows show the assessment steps, while the dashed arrows show the possible in silico modelling data (in the lower box) required for each step.

In this context of QSAR, as explained by Cronin and Livingstone (2004), the conceptual model in Figure 2.2 is expected to describe the linkages between the independent variables (e.g., the ingredient's structure) and the dependent variables (e.g., toxic effect). Appropriate variable selection procedure should be followed to ensure the most appropriate variables that are statistically relevant in terms of the correlation between the variables are selected. It is not uncommon to use more than one variable (e.g., physiochemical descriptors for the ingredient X and, if applicable, for the ingredient's metabolites) to predict the target biological activity. Cronin and Livingstone (2004) emphasize that, in such cases, the QSAR model should specify all the variables considered and, if needed, indicate the expected order of influence on the expected biological activity. Translated to the exploration of a problem more generally in the context of *in silico* toxicology prediction, the exploration phase should lead to a conceptual model that describes plausible scenarios through which harm may arise from the chemical(s) that are being assessed – e.g., different exposure scenarios in the case illustrated in Figure 2.2.

Notably, none of the *in silico* method-related studies identified in 2.2.3 include a conceptual model. Consequently, the authors do not discuss which variables might be included in such a model. I agree with Robinson et al. (2015), who argue that no quantitative model can exist without an underlying conception of its form. According to the authors, a lack of documentation of conceptual underpinnings makes it uncertain as to how to evaluate the completeness, clarity, and consistency of the logical structure behind the predictive tool derived. For *in silico* toxicological models, the lack of an expressed concept further makes it difficult to ascertain fitness for purpose based solely on the variables included.

#### **2.3.2.4 Hypothesis formulation**

An important role of the conceptual model is to function as a basis for the creation of testable hypotheses (Sauve-Cienciewicki et al., 2019). In the context of *in silico* models for chemical risk assessment, this amounts to generating a risk hypothesis based upon credible assumptions of how exposure to a chemical might affect a biological system. Consider the example of triethanolamine (a surfactant or stabilizer used in cosmetic ingredients) – a substance associated with incidences of liver tumors in animal studies (National Toxicology Program, 2024). Consumers using triethanolamine-containing cosmetic products (e.g., moisturizers and facial cleansers) will experience systemic

exposure to the compound following its dermal absorption (National Toxicology Program, 2024). As such, it may be hypothesized that consumers regularly using the products will be at increased risk of developing cancer. As outlined in this example, the hypothesis is formulated using existing information about both exposure to and potential for triethanolamine to cause harm. Additionally, the hypothesis is based upon the National Toxicology Program (2024) classification criteria of carcinogenicity of chemicals – i.e., a chemical is reasonably anticipated to be a human carcinogen based on evidence of carcinogenicity from animal studies, which indicates the incidence of tumors at multiple tissue sites, by multiple exposure routes, etc.

In the general PF literature, it is underlined that the development of a hypothesis is an iterative process, the outcome of which has the power to increase clarity and transparency in the defining and testing of postulated harm and, thus, increase confidence in the planned model prediction (Devos et al., 2019; Raybould 2006; Solomon et al., 2016; Wolt et al., 2010). The outcome of this process might also signal the need to revisit and adjust the earlier steps carried out in the PF process, either to match the hypothesis or to develop a new model. I illustrate this by drawing on a hypothetical QSAR model predicting the skin irritation potential of a low-dose mixture of two cosmetic ingredients: skin irritant (A) and skin absorption enhancer (B). Assume the original hypothesis is *B enhances skin absorption of skin-irritant A, which induces irritation*. Here, the hypothesis is formulated to only consider an effect-cumulative model, whereby the ingredient is held to produce a distinct influence (enhancing dermal absorption or inducing irritation) – the cumulative impact of which is skin irritation (Fig. 3). However, if the hypothesis is adjusted to include dose-cumulative effects of A, as follows: *the dose and duration of exposure to skin irritant A and ingredient B, which enhances skin absorption of A, determine the level of skin irritation*, it becomes necessary to revisit and reframe the problem to include both the dose and time factors of the ingredients (Figure 2.3), as key QSAR parameter data.

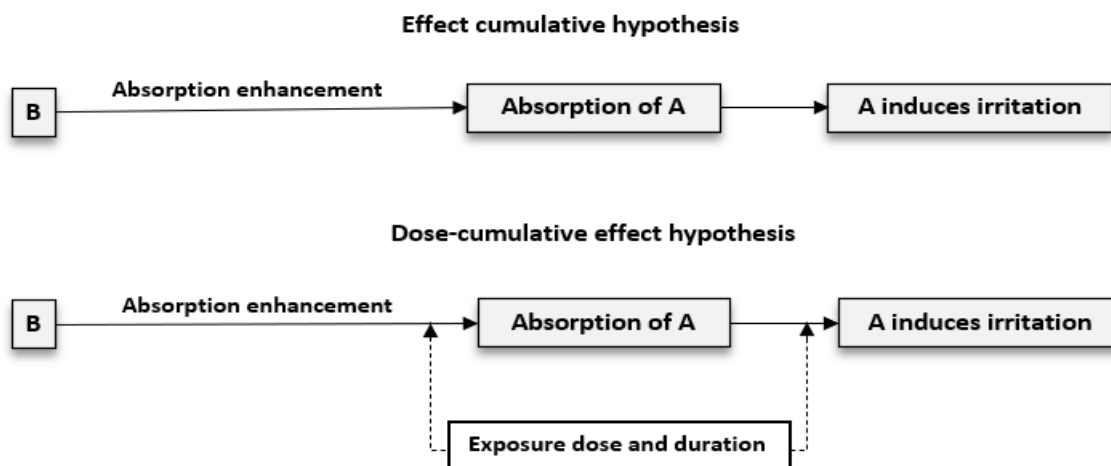


Figure 2.3. Simple hypotheses formulation diagrams showing two possible pathways to skin irritation: Effect-cumulative hypothesis (upper diagram) involving absorption enhancer B and skin irritant A, and Dose-cumulative effect hypothesis (lower diagram) involving absorption enhancer B and irritant A and time and dose as the influencing factors.

As explained in the above example (Figure 2.3), the QSAR prediction hypothesis should involve a comparison of possible hypotheses to ensure that necessary modelling data and model variables are incorporated into the model. To do this, one could first gather information about the product containing the target ingredients (in this case, A and B) to identify key properties and toxicological and toxicokinetic information about them, as these are crucial for designing the model. The idea is to formulate an initial hypothesis about the toxicological effects (and possibly underlying modes of action) and toxicokinetic fates of the ingredients, but also formulate initial thoughts on whether the model would be fit for the intended prediction. Once done, the need for one or more than one hypothesis is evaluated; whereafter, based on the available supporting evidence, one can prioritize the hypothesis to explore based on the hypothesis: supported by the evidence, tenable but not well supported by the evidence, untenable, or unable to rule out.

Overall, similar to Wolt et al. (2010), I hold that rigorous, transparent and iterative examination and consideration of various possible hypotheses, as part of the PF process (e.g., effect-cumulative hypothesis and dose-cumulative effect hypothesis in reference to Figure 2.3), is required to ensure confidence in the prediction either that harm

will result via a particular pathway, or else that this is highly unlikely (thus ruling out the need for its further analysis).

### **2.3.3 PF components in studies of *in silico* toxicology methods**

Analyzing the 13 papers in 2.2.3 led to the identification of 15 distinct components that are relatable to practices and features common to the application of predictive *in silico* toxicology (Table 2.1). In some cases, the component is explicitly mentioned, with the authors clearly describing the component in a specific assessment context, including giving detailed explanations through examples. For example, Pestana et al. (2021) explicitly mention “no observed adverse effect level (NOAEL)” for determining the risk of triazoles from oral 90-day studies in rats as the assessment endpoint. In other cases, the components are implicitly mentioned through a description in a general context without, for example, providing examples that explain it. For example, Escher et al. (2019), in a general context, mention “subchronic” as the exposure duration without specifying (e.g., by giving examples) the exact duration of the assessment.

The identified components were used to formulate broader categories that cover both specific components/with examples (i.e., component categories) based on the insights from the general PF literature. For example, the US EPA (US EPA, 2016) describes “assessment endpoint” as “an explicit expression of the environmental value that is to be protected, operationally defined by an ecological entity and its attributes”. Examples of assessment endpoints include a receptor of concern (e.g., test species) and the characteristics of the receptor to be measured (e.g., reproductive toxicity in the test species) (US EPA, 2016). From my analysis of the specific components, test species, NOAEL (Pestana et al., 2021), hepatotoxicity and reproductive toxicity (Pradeep et al., 2020), reproductive endpoints (Ouedraogo et al., 2022), and benchmark values (Escher et al., 2019) fit the description of assessment endpoints; thus I used the “assessment endpoint” as the broader component category to represent these specific components/with examples (see Table 2.1 for the other formulated component categories).

Table 2.1. Problem formulation components in thirteen in silico method-related studies that include problem formulation as part of the study.

<b>Component category</b>	<b>Specific component</b>	<b>Reference</b>
Context of use	Prioritization, hazard screening, risk assessment/classification and labelling	(Parish et al., 2020; Cronin et al., 2019)
Assessment endpoint	Test species (rat)	(Pestana et al., 2021)
	Hepatotoxicity, reproductive toxicity	(Pradeep et al., 2020)
	NOAEL	(Pestana et al., 2021)
	Reproductive endpoints	(Ouedraogo et al., 2022)
	Benchmark values	(Escher et al., 2019)
Exposure scenario	Exposure pathway (oral)	(Escher et al., 2019; Pestana et al., 2021)
	Exposure pathway (dermal)	(Ouedraogo et al., 2022; Reynolds et al., 2021; Baltazar et al., 2020)
	Exposure dose (0.1% face cream and 1% deodorant)	(Reynolds et al., 2021; Baltazar et al., 2020)
	Exposure frequency	(Sewell et al., 2017)
	Exposure duration (90-days and sub-chronic)	(Escher et al., 2019; Pestana et al., 2021)
Decision context	Acceptable level of uncertainty	(Dent et al., 2018; Belfield et al., 2021; Pallocca et al., 2022) (Schultz et al., 2019; Escher et al., 2019)
	Allowable lifetime exposures	(Escher et al., 2019)
	Acceptable safe concentrations	(Ouedraogo et al., 2022)
	Acceptable hazard	(Ouedraogo et al., 2022; Ball et al., 2022)

My analysis of the components across the 15 papers identified in 2.2.2 revealed little general consistency regarding the components addressed – with minimal overlap generally present. A majority of the components (8/15) were mentioned in only a single publication (e.g., frequency of exposure (Sewell et al., 2017)), whereas 4 were referenced in 3 or more (e.g., assessment endpoint (Enoch 2010; Escher et al., 2019; Pestana et al., 2021; Pradeep et al., 2020)), and not one appeared within more than 5 publications. Furthermore, the form in which the components manifested varied considerably, with 7/15 explicitly mentioned and the rest implicitly mentioned. An example is Pestana et al. (2021), who explicitly mention NOAEL as the assessment endpoint, while Ouedraogo et al. (2022) adopt a more general category (i.e., reproductive endpoints). Although such characterization by Ouedraogo et al. (2022) can provide a general idea of the assessment endpoint targeted by developers and users of *in silico* toxicology models, I hold that a more specific characterization could provide added value by addressing uncertainty about which specific reproductive endpoint is targeted. Notably, components describing specific model features (i.e., those relating to exposure scenarios) were addressed with greater frequency than those covering conceptual aspects (i.e., associated with the context of use).

The lack of components related to higher-level concepts (e.g. conceptual model, problem framing, and hypothesis development) in the analyzed studies presents a gap in the *in silico* toxicology PF, as the need to incorporate these concepts within such a PF has been emphasized by several authors (Callahan and Sexton 2007; Devos et al., 2019; Nickson 2008; Paoli et al., 2022; Solomon et al., 2016).

#### **2.3.4 Uncertainties associated with the higher-level components of PF**

By first analyzing and reflecting on the PF components identified in the 13 publications under section 2.2.3, I considered PF as a potential area where uncertainty can occur with respect to the higher-level components in instances where particular components are missing (thus leading to a partial inclusion of components) or only implicitly described (i.e., generality in the description of PF components such that a component does not provide any specificity in a given model prediction context). In this section, I discuss potential sources of uncertainty within PF components and propose a process that may be followed to characterize and address the uncertainty.

### 2.3.4.1 Sources of uncertainty

#### 2.3.4.1.1 Problem framing

In the problem-framing process, uncertainty generally arises, especially as there are commonly different (and not seldom conflicting) yet legitimate and plausible basis for concerns regarding a chemical. Additionally, uncertainty might arise where a problem is simplified to reduce the complexity in its interpretation. In determining whether *in silico* methods are called for, it is necessary to ascertain if there is sufficient data or scientific rationale to support the evidence about the potential harm posed (Madden et al., 2020; OECD, 2019). In cases where there is sufficient evidence or an estimate of potential harm that can be gained by other methods, *in silico* methods are generally not called for. Where *in silico* methods are called for, uncertainty still remains of whether an *in silico* model (e.g., read-across or QSAR) is robust or reliable for use (as a standalone or in an integrated system) for a particular toxicity prediction. Alternatively, uncertainty will arise from questions on whether a PF sufficiently describes the model to provide a starting point for judging the scientific validity of its predictions or the acceptability of the prediction outcome (US EPA, 2012). With respect to QSAR models, this may be identified through consideration of model form – i.e., is it quantitative (i.e., statistical regression), or qualitative (such as read-across or structural alert-based model) (US EPA, 2012)? The knowledge drawn from answers to such a question could offer meaningful insights into the strength of the models or the extent to which the models can be applied for certain applications (e.g., hazard identification, or risk assessment) (Parish et al., 2020). The knowledge could also be drawn upon to facilitate the framing of the level of importance of model features (model algorithms, descriptors, etc.) for the defined model use case scenario.

It is important to recognize that each framing will lead to the inclusion and exclusion of different aspects of the broader problem, depending on the viewpoints or assumptions considered at the PF stage. This in turn, sets the model system boundaries drawn for its assessment, such as the model structure (e.g., variables and variables relationships) (Sluijs et al., 2008). Articulation of the problem framing and related concerns of an *in silico* model is a process designed to facilitate a common understanding of the utility context of the model and its predictions. If the problem is not clearly framed, it might remain uncertain what aspects that are salient to influence the choice

of a model (e.g., its applicability domain, parameters and fitness evaluation criteria) – and what knowledge is relevant in prediction and evaluation (Sluijs et al., 2008). Ideally, the outcome of the problem-framing process is a statement that leads to describing information and research needs (Devos et al., 2019; Sauve-Cienciewicki et al., 2019; Solomon et al., 2016).

#### **2.3.4.1.2 Problem exploration**

Overall, as with problem framing, a clearly explored problem should minimize the manifestation of uncertainty by unambiguously describing the hazards or risks associated with a chemical and including the necessary details in the proposed model to help in gaining a deeper understanding of its suitability to make the prediction. This point can be illustrated by referring to the study by Moss et al. (2016) on skin sensitization of cinnamyl alcohol. The authors acknowledge the common understanding that cinnamyl alcohol is a pre-hapten whose skin sensitization can occur through conversion to protein-reactive cinnamaldehyde. However, their further exploration reveals that cinnamyl alcohol can also directly induce skin sensitization through a pathway independent of the one involving cinnamaldehyde. This conclusion was supported by observation of the formation of epoxy-alcohol and the activation of the allylic hydroxyl function. Here, uncertainty can be introduced if cinnamaldehyde data, information on possible additive/synergistic reaction of cinnamyl alcohol and cinnamaldehyde, or the influence of exposure dose and duration of each compound, etc., are not considered in a model prediction. In other words, as shown in this example and emphasized elsewhere (Sauve-Cienciewicki et al., 2019), it is possible to gain a more comprehensive understanding of a problem by organizing relevant knowledge about the chemical for *in silico* model development and prediction and ensure a well-defined applicability domain and adequacy of the model to address the prediction problem; otherwise, the model may suffer from inadequacy in terms of the input parameters and model structure used. As the central goal for the problem exploration step is to lead to the development of a conceptual model, the quality of the conceptual model will inevitably suffer, given the uncertainties in problem exploration (Devos et al., 2019; Sauve-Cienciewicki et al., 2019).

#### 2.3.4.1.3 Conceptual model

El-Ghonemy et al. (2005) as well as Zheng and Bennett (2002), emphasize the need to pay attention to possible uncertainty arising from either the under-simplification or oversimplification of a conceptualized model. Generally speaking, an oversimplified conceptual form fails to capture the crucial features necessary for the successful construction of a quantitative model, thus resulting in a model that inadequately simulates the endpoint intended. Ekins et al. (2007) illustrate this uncertainty by describing two chemical interaction systems: toxicodynamic and toxicokinetic. The former should incorporate reference to toxic responses of the biological system after chemical exposure; thus, the conceptual model should capture those elements of the biology (e.g., receptors, ion channels, nucleic acids, anabolic and catabolic enzymes) implicated in the emergence of toxicity. In the latter case, the conceptual underpinning should capture those elements of the physiological response to the xenobiotic presence (e.g., chemical-metabolizing enzymes, transporters, circulating proteins) that serve to elicit or influence either the metabolism, transportation, distribution or excretion of the chemical (Ekins et al., 2007). In these instances, oversimplification of the conceptual model might occur if, for example, the number of biological elements included is reduced to the point where connections between themselves and other variables are not captured. In contrast, under-simplification will generally introduce several variables into a model system without clearly distinguishing between them. Thus, uncertainty is introduced due to the resultant difficulties in interpreting the influence of any single feature upon the output of the model.

The above discussions highlight the importance of developing and using a conceptual model to clarify the boundaries of an *in silico* model system, as well as the variables and relationships which are conceived as relevant to it. In turn, a conceptual model will help to establish whether it is appropriate to include or exclude specific information in the quantitative model and, thus, by extension, to infer the utility of its predictions (Walker et al., 2003). This implies that it is crucial to make explicit the model boundaries, as a lack of clarity on this matter will introduce uncertainty regarding which variables are appropriate to include or exclude (Skinner et al., 2014). Furthermore, confidence in a developed quantitative approach and its predictive output will require documenting

any simplifications, assumptions, and justifications provided for the choice of information considered within the conceptual model.

#### **2.3.4.1.4 Hypothesis formulation**

A hypothesis formulation should account for possible associated uncertainties. In reference to the example in Figure 2.3, the choice of any hypothesis should clearly include the understanding of the expected level of uncertainty surrounding the exposure pathway selected and how that could translate to the level of confidence placed on the predicted skin irritation. If, for example, uncertainty is expected to be “too high”, then a decision could be iteratively made to add information (e.g., information on the exposure dose/duration) to lower the uncertainty.

Welss et al. (2004) present explanations that corroborate the above discussion on uncertainty. The authors point out that substances can operate through two distinct pathways to initiate skin irritation. In the first, damage to the barrier function of the stratum corneum can initiate irritation, with the ingredients' dose and duration of exposure determining the extent of any injury (Johnson et al., 2020a; Welss et al., 2004). The second pathway occurs when damage to the skin enhances irritants' penetration into the deeper epidermal layers, initiating irritation through interaction with living keratinocytes (Johnson et al., 2020b; Welss et al., 2004). Both routes can lead to skin irritation, whether alone or in combination. The hypotheses may be constructed based upon each of these putative pathways (pathway 1 or pathway 2) – alongside a third, combining elements of both pathway 1 and pathway 2. Uncertainty might otherwise remain in relation to factors such as the scenario anticipated in the context of the prediction, the level of confidence which should be held in the scenario chosen, the number and type of parameters which ought to be included in the model, and the robustness of the prediction output. Overall, I contend that developing and evaluating several hypotheses makes it easier to not only judge whether or not any one selected is robust but also whether it is fit for a specified hazard or risk prediction. Alternatively, the uncertainty associated with hypotheses formulation might set a precondition for rejecting *in silico* predictions for use in a regulatory setting.

### 2.3.4.2 Characterizing and addressing uncertainty associated with PF components

Following on from the discussion under section 2.3.4.1 above, I propose a process (Figure 2.4) that one can follow to characterize and address the uncertainties. Description of the PF components (and identification of areas of uncertainty) is a critical first step in this process, as it provides the context for uncertainty analysis. Three fundamental questions (shown in the light blue boxes in Figure 2.4) are then raised. The first question (“Is there uncertainty?”) seeks to determine whether uncertainty resides within any of the described components. In theory, a “no” answer can emerge, indicating that uncertainty is not a concern; thus ruling out the need for uncertainty analysis but indicating the need to directly proceed to predict chemical hazard/risk following problem formulation. If, however, the answer is “yes”, then the second question (“Is the uncertainty acceptable?”) becomes relevant to ask in a defined decision context.

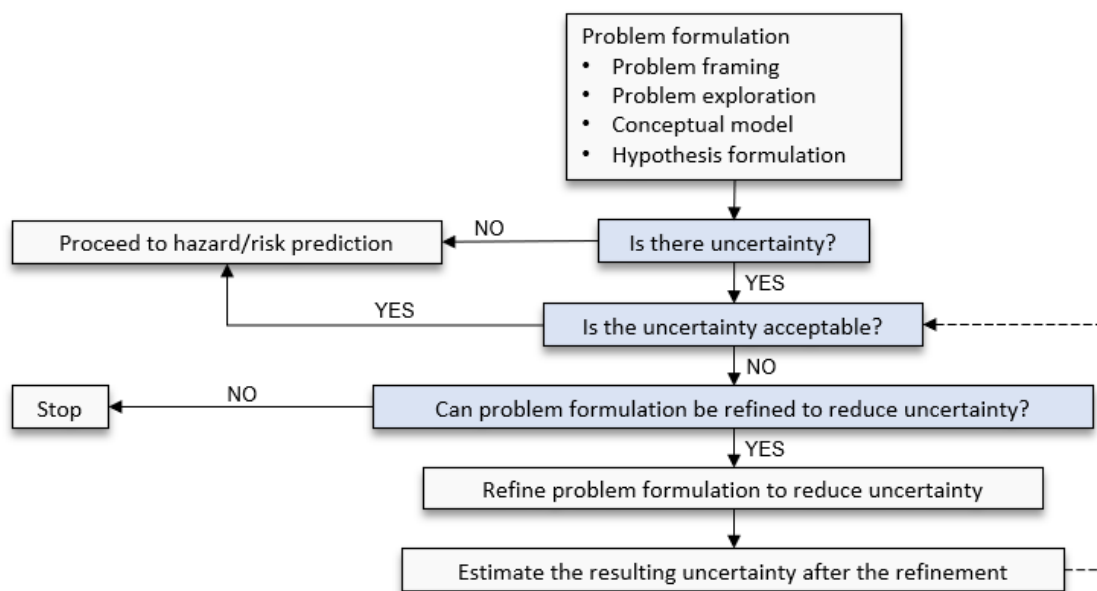


Figure 2.4. A proposed step-step (shown with the complete arrows) process to characterize and address the uncertainty associated with PF components and an iteration (shown with the dashed arrow) required at one of the steps.

To determine the acceptability of uncertainty in PF, it is necessary to first characterize (i.e., quantify/qualify) uncertainty from each component – this can be done, for example, by using a scoring scheme, such as “low” or “high” to rank uncertainty. Here, the goal is to determine how each component contributes to the overall uncertainty in PF as well as prioritize areas for uncertainty reduction (or elimination, if possible). It is necessary to

clearly spell out the context in which the level of acceptability is defined, as different uncertainty levels might be considered as being acceptable depending upon the potential consequences of inaccurate formulations. For example, a high uncertainty level can be accepted for decisions with respect to priority-setting – e.g., a machine learning based-model for screening chemical inventories to identify toxic molecules can be acceptable even with formulation that leads to relatively high false positives in the prediction, as long as the model is fit for the screening purpose (Belfield et al., 2021). On the other hand, low uncertainty in a mechanistic-based model risk prediction might be accepted when setting health-based standards (Belfield et al., 2021). Overall, if the level of uncertainty is considered acceptable, one can proceed to the hazard/risk prediction phase; otherwise, an additional question should be raised about whether PF can be refined to reduce the uncertainty to an acceptable level.

When uncertainty is considered unacceptable and irreducible, one may choose to stop the need for analysis and potentially consider not using the PF. However, if uncertainty can be reduced, one can evaluate the added value of refining the components associated with uncertainty by, for example, incorporating more information or introducing more complex models. After the refinement, the resulting uncertainty can be estimated to determine whether or not the uncertainty has been reduced to an acceptable level. The essence of this step is to allow this process to be iterative by making it possible to continuously identify, estimate, and address uncertainty within PF components for a specific prediction context.

### **2.3.5 Further consideration of PFs**

Finally, I wish to highlight an area where future research is needed. Sauve-Cienciewicki et al. (2019) emphasize that the PF process must be structured, iterative, and include *all key stakeholders*, as it is context-dependent (my emphasis). I agree that it is crucial to recognize that each problem situation is unique, and that the quality of the decision outcome depends on the perspectives and expertise of those included in the process. However, including *all key stakeholders*, including relevant experts, as proposed by the authors, is a tall task. Also, the framework proposed by the authors assumes that the deliberations will generate a consensus and that "Failure to reach consensus on the specific problem to be addressed leads to misunderstandings and an inability to create

appropriate solutions" (ibid., p. 187). I argue that this is a weakness in the PF frameworks I have reviewed, as it is well documented that in cases of societal relevance where there is large uncertainty, it is generally the case that people – including the experts – do not reach a consensus, and that more research and deliberations can lead to hardened positions (Donfrancesco et al., 2023; Funtowicz and Ravetz 1993; McIlroy-Young et al., 2021). Notably, little, if anything, is known about how different producers or users of *in silico* model predictions prioritize among different PF components or sources of uncertainty.

## 2.4 Conclusion

This study led to the discovery of a gap between the broader risk assessment literature and *in silico* toxicology method literature about how PF is conceptualized. While the general PF literature emphasizes that PF frameworks must address higher-level conceptual and context-specific questions, the studies I analyzed, which all describe PFs for *in silico* toxicology methods, do not include such components. Furthermore, there was very little consistency across the studies regarding the type of components they addressed. Drawing on a general PF framework (Sauve-Cienciewicki et al., 2019), I developed a preliminary PF framework (Figure 2.1) for *in silico* toxicology methods and described the framework in light of the components mentioned in studies that address PFs applied to *in silico* toxicology methods (Table 2.1). To my knowledge, such a framework has not been previously developed for this purpose. The framework can be used to clarify which PF components are central to a particular *in silico* toxicology method. Critical to this is to shed light on how uncertainty can manifest within the PF, if particular components are excluded or implicitly described. This study suggests that explicit-making the selection among components has the potential to clarify the perspectives used in the selection process and help avoid potential biases and blind spots in the team that developed the PF. For a growing research field such as *in silico* toxicology methods, where scholars from different disciplinary and cultural backgrounds are likely to differ in their prioritization of which components to include in a PF, chemical regulatory decisions are likely to benefit from improved transparency to that end (Devos et al., 2019).

## Chapter 3: A framework for categorizing sources of uncertainty in *in silico* toxicology methods: considerations for chemical toxicity predictions

### 3.1 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 mentioned 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 (EFSA 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 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 (Alexander-White et al., 2022; ECHA (European Chemical Agency), 2017). 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), Cronin et al. (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, I deduce GSU through iterative categorization of the VRSU (the process followed to develop the framework is shown in Figure 3.1). My 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.

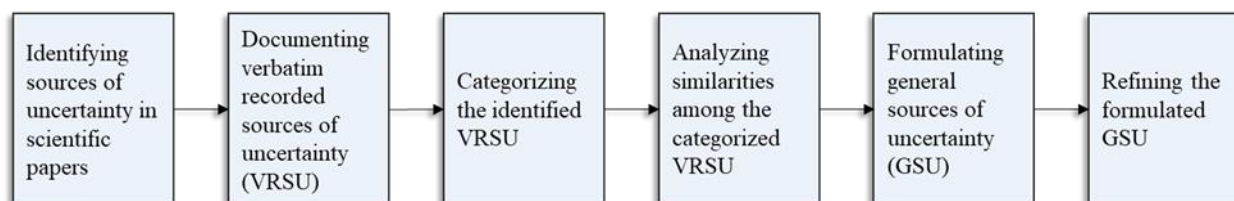


Figure 3.1. A flow chart describing the steps undertaken to develop the framework that categorizes general sources of uncertainty in *in silico* models for toxicological data gap filling.

### 3.2 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 S3.1). A content analysis was conducted of these publications through line-by-line reading (Tracy, 2018), identifying sources of uncertainty mentioned or discussed in these papers. I 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. I 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 S3.1). In so doing, I 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 S3.1.

As noted with asterisks in Table S3.1, 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 this paper. In addition, I 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. (2014), 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 I acknowledge that identifying areas of variability and bias in *in silico* model systems is also important, it is beyond the scope of this paper.

### **3.3 Categorizing VRSU and formulating GSU**

In the present paper, I 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, I conceptualized the 10 components in Belfield et al. (2021) as areas that generally characterize potential sources of uncertainty in *in silico* toxicology modeling (Table 3.1).

In my modification of the framework, I 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 toxicity prediction. In another modification, as further discussed in Section 3.3.1.4, I added a new higher-level assessment component – “Similarity”. This was placed under the Model creation phase. As the framework I set out to develop includes characterizing uncertainty related to the use of *in silico* toxicology models, I also added “Applicability” as a higher-level assessment component, resulting in a total of 10 higher-level assessment components (Table 3.1). 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).

Table 3.1. *In silico* toxicology modeling phases, higher-level assessment components, and definition of the components of relevance for the present study.

<b>Modeling phase</b>	<b>Higher-level assessment component</b>	<b>Definition of the higher-level assessment component</b>
-----------------------	--	--

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	Accuracy and/or quality of the reported chemical structures in the training (and, if applicable, test) set used for modeling
	Similarity	Resemblance or commonality between chemical compounds, e.g., in terms of functional groups, toxicokinetic/toxicodynamic properties, and chemical structure
	Descriptors	Appropriate use and adequate definition of the descriptors used for modeling (including how and where sourced)
Model characterization	Modeling	The appropriateness and/or adequacy of the modeling approach for the endpoint with regard to complexity of the endpoint and potential use of the
	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

I started the categorization process by reviewing the 81 VRSU (Table S3.1) in light of the descriptions of the higher-level assessment components in Table 3.1 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 (4<sup>th</sup> column in Table 3.2). These VRSU were then analyzed for similarities and then used to formulate the GSU shown in Table 3.2 (column 3). Subsections 3.3.1, 3.3.2, and 3.3.3 below describe the reasoning that led to the categorization of the VRSU and the formulation of the GSU.

- 1 Table 3.2. Categories of the 81 VRSU and the formulated general sources of uncertainty (GSU, column 3). The non-bolded GSU are the tentative GSU that  
 2 were subsumed under the refined GSU (bolded). Publication numbers are provided in Table S3.1.

<b>Modeling phase</b>	<b>Higher-level assessment component</b>	<b>General sources of uncertainty (GSU)</b>	<b>Verbatim recorded sources of uncertainty (VRSU)</b>	<b>Publication number</b>
<b>Model creation</b>	Data	<b>Data quantity</b>	Quantity of the data considered	1
		Number of data	Number of analogues contributing data	1
				Number of analogues contributing data
			Number of the chemical analogs identified	5
			Number of source chemicals	8
		Data size	Size of training data set data	4
		Data coverage	Coverage [of structural alert]	7
		<b>Data balance</b>	Balance of the training data set	4
			Data balance	6
		Data distribution	Distribution of the training data set	4
		Data homogeneity	Homogeneity of the chemical space of the training and test sets.	6
		<b>Data relevance</b>	Relevance of data for the endpoint of interest	6
			Relevance of data	10
		Data suitability	Suitability of analogues	1
			Suitability of the chemical analogs identified	5
		Data completeness	Completeness of the argument provided [for data quality]	3
			Completeness of the data set	6
		Database deficiency	Database deficiencies (e.g., lack sensitive endpoint or toxicity information)	2
		<b>Data reliability</b>	Reliability of data	10
		Data consistency	Consistency of the data set	6
			Consistency of data	9
		Data robustness	Strength or robustness of the supporting data sets	3
			Robustness of the source or analogue data	8
			Robustness of the supporting data sets	9
		<b>Data accuracy</b>	Accuracy of data	10

		[computed/not experimentally measured] Parameters used to construct the model	11	
	<b>Data validity</b>	Validity of data	10	
	Data quality*	Quality of the data	1	
		Quality of the apical endpoint data	3	
		Quality of data used to build model	5	
		Toxicological information found for the analogs	5	
		Quality of data	6	
		Quality of the source or analogue data	8	
		Quality of data	9	
Structure	<b>Chemical structure</b>	Structure and its representation	8	
		Structural description	7	
Similarity	Chemical similarity	Structural similarity to target	1	
		Similarity in chemistry	3	
		Toxicokinetic similarity	3	
		Toxicodynamic similarity	3	
		Similarity justification	4	
		Definition and demonstration of similarity	8	
Descriptors	Descriptor relevance	Choice of molecular descriptors	4	
	Descriptor concordance	Calculated/experimentally measured properties and descriptors	6	
		Property domain	7	
<b>Model characterization</b>	Modeling	<b>Model structure</b>	Modeling algorithm and hyperparameter	4
			Numerical errors and/or numerical approximations	11
			Model bias	11
		<b>Activity/potency</b>	The potency of the analogues for those [toxic] effects	1
			Nature and severity of the identified toxic effects	1
			Prediction of complex endpoints such as chronic toxicity	5
			Species specificity	7
			Toxicity or relationship to adversity	7
		<b>Activity/</b>	Weight-of-Evidence	3

		<b>potency evidence</b>	
		Supporting evidence	7
		Corroborating evidence	7
		Weight-of-evidence supporting the prediction	9
Performance	<b>Model performance</b>	Model performance	5
		Reproducibility of model and model prediction	6
		Adequacy of the model to make a prediction for the stated purpose	6
		Statistical performance	6
		[predictive] Performance	7
Mechanisms	<b>Mechanistic plausibility</b>	Mechanistic plausibility	3, 8, 9
		Mechanistic causality	7
	Mechanistic relevance	Mechanistic relevance and interpretability	6
		Mechanistic relevance	8
Toxicokinetics	<b>Metabolic domain</b>	Metabolic domain	7
	<b>Coverage of ADME activity</b>	Adequate coverage of Absorption, Distribution, Metabolism and Excretion effects	6
Applicability	<b>Applicability domain</b>	Applicability domain	4
		Applicability domain	5
		Applicability domain	6
		Applicability domain	8
Relevance	<b>Extrapolation</b>	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 analogs	5
	<b>Model relevance</b>	Relevance [of the QSAR] to the prediction or assessment goal	6
		Purpose or potential use of the structure alert	7

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

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

### 3.3.1 The model creation phase

#### 3.3.1.1 Data

The higher-level assessment component “Data” is in the developed 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 3.1). 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., 2015b), 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 analogs identified” (the full list of the VRSU categorized under Data is shown in Table 3.2).

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 analogs 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, I formulated “Data quantity” as the common GSU that covers them (Figure 3.2). The choice of the term

“data quantity”, was based on its common use in describing uncertainty related to the amount of data (WHO/IPCS, 2018).

I also note that the GSU “Data quality”, as used in six papers (Table 3.2), 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, I excluded Data quality as a separate GSU.

My analysis revealed an overlap in the description of three initial GSU: Data balance, Data distribution, and Data homogeneity. Pham et al. (2019), (citing Haibo 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 categoric (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 (Haibo 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 3.2. Notably, these descriptions align with descriptions of Data relevance by Cronin et al. (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. I 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”.

I 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, I 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, I retained both “Data accuracy” and “Data validity” as distinct GSU (Figure 3.2).

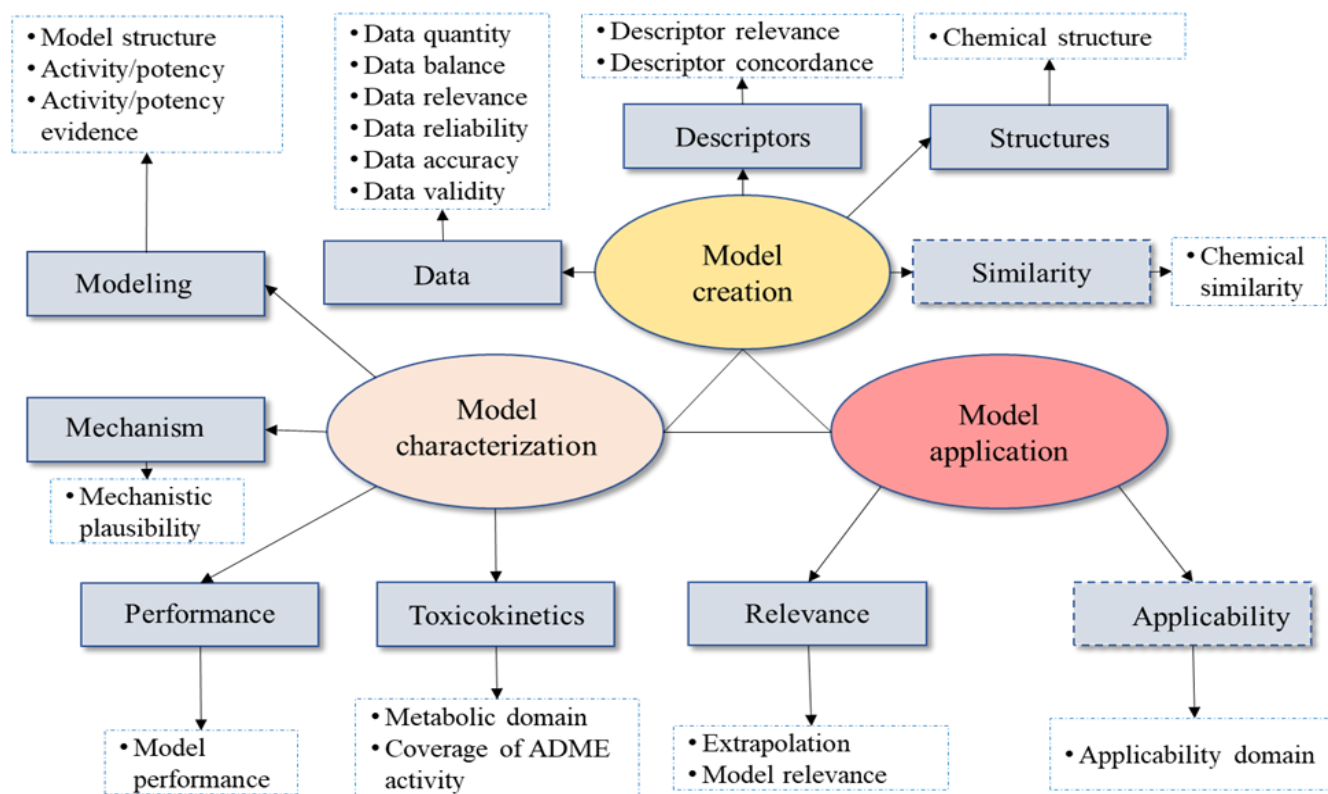


Figure 3.2. The refined GSU (bulleted in the rectangles) resulting from the analysis and iterative categorization of the VRSU. The descriptions of the GSU uncertainty are provided in Table S3.2. The grey rectangles indicate the higher-level assessment components under which the GSU are categorized, and 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).

### 3.3.1.2 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 3.1). I placed two VRSU under this component (Table 3.2): “structure and its representation” (Schilter et al., 2014),

and “structural description” (Cronin et al., 2022). My 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 usable 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.

### **3.3.1.3 Similarity**

I 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 3.1). I placed six VRSU (Table 3.2) 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 & 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 Similarity 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, I decided to formulate “Chemical similarity” as the umbrella GSU that covers them.

#### 3.3.1.4 Descriptors

As seen in Table 3.1, 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. I interpret this definition as an elaboration of the description of Descriptors in Table 3.1, which then implies that a descriptor represents a logical transformation of chemical information encoded within its physicochemical properties. In my categorization (Table 3.2), I, 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.

### **3.3.2 The model characterization phase**

#### **3.3.2.1 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 3.1). I 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 3.2 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, I 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, my analysis led us to conclude that each of these five VRSU relates to the ability to model toxicological activities or potency of chemicals (Table 3.2). Considering the similarities, I thus formulated “Activity/potency”, as the umbrella GSU term for these five VRSU (Table 3.2).

Finally, I 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), I 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 3.2, therefore, I used these three VRSU to formulate the GSU “Activity/potency evidence”.

### **3.3.2.2 Performance**

The higher-level assessment component “Performance” is here defined as “Adequate statistical fit, predictivity and appropriate reporting”. I 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). My 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, I formulated “Model performance” as the umbrella GSU for these VRSU.

### **3.3.2.3 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; Schultz et al., 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 3.2), which aligns with the description of Mechanism (Table 3.1).

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, my 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, I 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 3.2). However, further analysis revealed that, although described using different terminology, Mechanistic relevance also relates to Mechanistic plausibility. As seen in Table 3.2, 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 the developed framework (Figure 3.2), 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 S3.2 for the description). Lastly, I 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.

#### **3.3.2.4 Toxicokinetics**

The higher-level assessment component “Toxicokinetics” is defined in the framework as “Appropriate consideration of metabolism and toxicokinetics in the model” (Table 3.1). 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, my 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, I 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” (Figure 3.2). Here, I 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., 2020b; Achar et al., 2020c; Cronin et al., 2019).

### **3.3.3 The model application phase**

#### **3.3.3.1 Applicability**

Drawing on Cronin et al. (2019), I 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)”. I 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 3.2). 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, I decided to formulate “Applicability domain” as the umbrella for this VRSU.

#### **3.3.3.2 Relevance**

“Relevance” is defined in this context as “Relevance of the model to its intended purpose and use” (Table 3.1). I 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 analogs” 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 my analysis suggesting that each point to the transferability of a model or model prediction towards a different context.

My 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, I 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, I 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.

### **3.4 Application of the framework to prioritize areas of uncertainty**

This study developed a framework (Figure 3.2) 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.

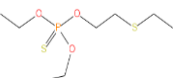
#### **3.4.1 Case study**

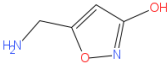
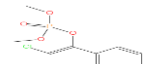
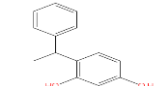
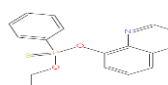
The overarching aim of the study was to evaluate the performance of Toxicity Estimation Software Tool (TEST; v5.1.2) (US EPA, 2015) in the prediction of oral rat LD<sub>50</sub> (the dose that causes death of 50% of test samples) – this

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 LD<sub>50</sub>-based Globally Harmonized System (GHS) categories with the corresponding experimental LD<sub>50</sub>-based GHS categories. The GHS classification categories are presented in Table S3.3) (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 LD<sub>50</sub> 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 3.3) with rat experimental LD<sub>50</sub> 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 LD<sub>50</sub> values, which I 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 to 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 LD<sub>50</sub>, method – consensus, and fragment constrain – relaxed. The predicted rat and human-derived LD<sub>50</sub> data are shown in Table 3.3.

Table 3.3. Information 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
Quintiofos	1776-83-6		150	3	24	2

### 3.4.2 Identification of relevant GSU from the case study

I proposed a checklist to support the identification and justification of GSU deemed relevant for inclusion from the case study (see Table 3.4). The left-hand side column of the checklists is the GSU from the framework (Figure 3.2). While not all the 20 GSU may be considered relevant in a prediction context, I recommend that all of them should be included. The right-hand side column contains spaces for justifying why a GSU is selected.

Table 3.4. 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	<input type="checkbox"/>	
Data balance	<input type="checkbox"/>	
Data relevance	<input checked="" type="checkbox"/>	TEST is built on a large amount of data (i.e., 7413 available in ChemIDplus database) collated from different studies and with multiple LD <sub>50</sub> values for the same compound or isomers with different LD <sub>50</sub> (US EPA, 2015), suggesting that some level of data variability is expected (Karmaus et al., 2022). When relying upon such heterogenous data to predict rat LD <sub>50</sub> , 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	<input checked="" type="checkbox"/>	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; Karmaus et al., 2022). The fact that TEST does not report reliability of the data suggests the need to factor in this GSU when interpreting the rat LD <sub>50</sub> results.
Data accuracy	<input checked="" type="checkbox"/>	Table 3 shows that the experimental LD <sub>50</sub> -based GHS categories for the five chemicals do not match their predicted LD <sub>50</sub> -based GHS categories (e.g., Quintiofos lies within GHS category 3

		(according to its experimental LD <sub>50</sub> 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	<input checked="" type="checkbox"/>	In the TEST user manual, the oral rat LD <sub>50</sub> 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 by the model developer presents a knowledge gap in judging level of validity of the data.
Chemical structure	<input type="checkbox"/>	
Chemical similarity	<input checked="" type="checkbox"/>	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 <sub>50</sub> values.
Descriptor relevance	<input type="checkbox"/>	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 <sub>50</sub> , the extent to which the descriptors were considered relevant or relative importance to the prediction output.
Descriptor concordance	<input type="checkbox"/>	
Model structure	<input type="checkbox"/>	
Activity/potency	<input checked="" type="checkbox"/>	The predicted LD <sub>50</sub> data assume that each chemical dose (in a statistical sense) will lead to 50% mortality in rats 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 rats and whether the doses will result in the recorded potency outcome.
Activity/potency evidence	<input checked="" type="checkbox"/>	In decision-making contexts, where the basis on which the conclusions about validity or reliability of the predicted rat LD <sub>50</sub> data should be made, questions related to the weight of evidence that support the conclusions 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 <sub>50</sub> values of the five compounds (Table 3.3), 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	<input checked="" type="checkbox"/>	Model performance is here evaluated based on the ability of the model to accurately classify the human TEST Consensus-derived LD <sub>50</sub> data under the same GHS category as the human <i>in vivo</i> -derived 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 3.3). 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	<input type="checkbox"/>	
Metabolic domain	<input checked="" type="checkbox"/>	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 <sub>50</sub> . For example, Demeton-O (Table 3.3) is known to produce the more toxic Demeton-S metabolite in rats (Barnes and Denz, 1954); however, this is not accounted for in the predicted value (Table 3.3). During uncertainty analysis or interpretation of the predicted data, it is thus remain relevant to consider consequences of not including the influence of such toxic metabolite (Burden et al., 2016).
Coverage of ADME activity	<input type="checkbox"/>	
Applicability domain	<input checked="" type="checkbox"/>	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	<input type="checkbox"/>	
Model relevance	<input checked="" type="checkbox"/>	One of the problems realized in the study above is over-and under-predictions (Table 3.3). This makes it important to consider uncertainty regarding the relevance of the models for hazard classification in regulatory context. For example, where conservative predictions are desired as a health protective strategy, under-prediction incidences (as in the case of 4-(1-Phenylethyl)benzene-1,3-diol – Table 3) are not desirable. On the other hand, the over-prediction of Quintiofos raises question about whether can also lead to false classification of less toxic (or safe) chemicals as more toxic (or toxic), which might lead to potentially beneficial compounds being abandoned during chemical/drug development.

Taken together, the case study presented here, and the checklist (Table 3.4) indicate that the GSU within the framework can be used to 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 3.4, 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, I 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, I 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 3.4), similar to Achar et al. (2024d), Jones and Falloon (2009) and Przybylak et al. (2012), I 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.

### **3.4.3 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 (Q)SAR 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 Figure 3.3) (OECD, 2007), while the second set is based on the 2023 OECD guidance on the assessment of QSAR predictions (summarized on right side of Figure 3.3) (OECD, 2023).

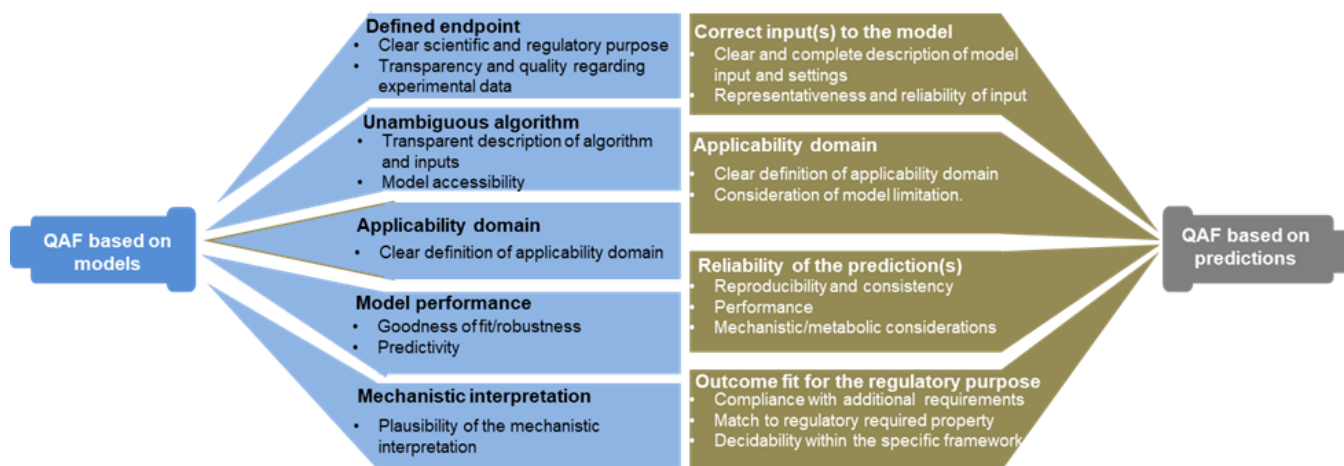


Figure 3.3. A summary of the principles and elements outlined in the 2023 OECD’s proposed QAF. The elements are bulleted under the principles.

QAF does not in itself provide a systematic categorization of diverse sources of uncertainty based on model components and modeling phases (as in the framework – Figure 3.2); 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 the 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 3.1 and illustrated through the checklist (Table 3.4), the 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 the framework, the 3<sup>rd</sup> 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 the 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, the framework developed in this thesis 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 the framework, therefore, I 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, the framework aligns with the working principles of EFSA et al. (2018) and WHO/IPCS (2018), where comprehensive and explicit identification of possible 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. The developed here 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” – Figure 3.2) 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 (Figure 3.3) 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, the developed framework, therefore, shows connections between the proposed modeling phases (creation, characterization and application) (Figure 3.2). 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.

### **3.5 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). The fact that such a framework is presently missing presents a gap within the field of *in silico* toxicology modeling.

In this study, I analyzed studies that have categorized sources of uncertainty across different *in silico* toxicology methods. My 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 3.3, 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. (2014) 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, I developed a framework (Figure 3.2) 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), I see my contributions in three ways. Firstly, I 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), I 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, I 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, the 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 the developed framework is more representative of areas of uncertainty identified by multiple modelers. In other words, the importance of the developed 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 3.4, I have shown that the introduced general sources of uncertainty (GSU) can provide a more nuanced understanding and practical way of prioritizing which areas of uncertainty to address within an *in silico* toxicology prediction context.

With the overarching aim of fostering a structured (and potentially more transparent) understanding of where uncertainties reside, the framework (Figure 3.2) and the checklist (Table 3.4) 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 (2018), 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, I explored 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 (see Chapter 4 of this thesis). Lastly, I would like to acknowledge the difficulty of developing a framework that covers all possible sources of uncertainty; thus, while the developed framework covers several GSU within *in silico* toxicology modeling, I refrain from claiming to have developed a “standard” framework to this end. Additionally, I would also

like to acknowledge that it is possible that the literature search criteria applied in the current study (under Section 3.2) might have led to some sources of uncertainty or the (grey) literature in *in silico* toxicology methods not being captured. Nevertheless, I believe that this potential limitation did not affect the conceptual breadth of the developed framework.

## Chapter 4: Analysis of implicit and explicit uncertainties in QSAR prediction of chemical toxicity: a case study of neurotoxicity

### 4.1 Introduction

Quantitative structure-activity relationships (QSARs) are an *in silico* toxicology approach that aims to establish relationships between descriptors of chemical structure and biological activities (e.g., toxicity) or properties. The implicit assumption is that structurally similar chemicals should have similar activities/properties and the trends can be identified and modeled within groups of molecules (Cronin and Madden, 2010). QSARs have the potential to support the reduction in the use of animal testing in different assessment contexts aimed toward characterizing and/or predicting chemical toxicity (Belfield et al., 2021; Cronin et al., 2019; Patlewicz et al., 2013). There are increasing calls to incorporate QSAR predictions in the assessment of chemical toxicity (e.g., within the European Food Safety Authority (EFSA), 2010), Health Canada (2023b), and the US National Research Council (2007)). However, it is recognized that it will be difficult to address complex endpoints with QSAR alone. An example is the prediction of neurotoxicity, given the unreliability of animal models in assessing this endpoint for reasons such as interspecies differences in brain morphology or differences in biological functions between humans and animals (EFSA, 2010; EFSA et al., 2021; Fritsche et al., 2018). Worth et al. (2011b) state that no single QSAR model (or in combination) seems adequate to predict the neurotoxic potentials of chemical compounds. The shortcomings of QSAR models have led to increasing demands to analyze and communicate uncertainties in QSAR models and model predictions of complex endpoints such as neurotoxicity to support efforts aimed at addressing the uncertainties (Belfield et al., 2023; Cronin et al., 2019; Piir et al., 2018; Sahlin et al., 2011; Schultz et al., 2019; Vighi et al., 2019).

Researchers express uncertainties in various ways, including the use of words that implicitly or explicitly qualify or represent the author's confidence in the content of the information communicated or alter precision implied in a measured numerical value (Flari and Wilkinson, 2011; Levin et al., 2004). As defined in Table S1, explicit uncertainties are expressed directly as gaps, unknowns, or quantitative confidence measures regarding, for

example, model predictivity (e.g., “it is uncertain” and “more data are needed”) and are consequently easily detected (Flari and Wilkinson, 2011; Sahlin et al., 2021). In contrast, implicit uncertainties are expressed indirectly in a subtle manner (or unintentionally) – e.g., through words such as “probably”, “maybe”, and “might”. While understanding the implicitly and explicitly expressed uncertainties can provide valuable nuance and guide quantitative uncertainty estimations, these uncertainties are not easily discerned and can, therefore, be easily overlooked during the analysis or interpretation of QSAR predictions (Flari and Wilkinson, 2011; Levin et al., 2004; Zerva, 2019). In other words, at present, it is difficult to elucidate whether modelers use linguistic expressions to indicate how certain they regard the current state of knowledge of QSAR models and the information regarding, for example, input data, parameters, and prediction outputs (Flari and Wilkinson, 2011). To support systematic and transparent accounting of uncertainties in QSAR predictions, I here develop a method that makes it possible to identify and, based on my previously developed framework (Achar et al., 2024a), systematically categorize implicit and explicit uncertainties expressed in texts in studies applying QSARs for chemical toxicity predictions.

Research on uncertainty indicators and how they can influence the perceived certainty of information in written statements exists across different research areas. For example, Markkanen and Schröder (1997), Varttala (2001), and Zerva (2019) explore the use and interpretation of uncertainty expressions that qualify confidence in statements in linguistic and behavioral studies, Stortenbeker et al. (2019) analyze the influence of the use of uncertainty expressions by doctors on patient anxiety during doctor-patient communication, while Ferson et al. (2015), Rubin (2007), and Zerva (2019) explore the use of machine learning to identify probability phrases and numerical uncertainty expressions. Scholars such as Flari and Wilkinson (2011) and Levin et al. (2004) highlight the utility of implicit and explicit uncertainty indicators as markers when identifying uncertain information in scientific texts or written statements. In this study, I follow what is held to be best practice when identifying uncertainty communicated within a statement, which is to first identify uncertainty indicators and then analyze the uncertainty in information in a statement based on how the indicators relate to and affect the conveyed information (Flari and Wilkinson, 2011; Hillen et al., 2017; Levin et al., 2004; Stortenbeker et al., 2019; Zerva, 2019).

Uncertainty estimation in QSAR modeling falls into two major categories: aleatoric uncertainty and epistemic uncertainty (see the definitions in Table S1). While the former cannot be eliminated through additional data, the former can (at least in theory) be eliminated through additional data or knowledge (Gajewicz et al., 2015; Scalia et al., 2020; Wang et al., 2021; Zhong et al., 2022). In this study, I focus on epistemic uncertainty, as it is regarded to be more problematic in QSAR modeling exercises – e.g., with respect to understanding whether a model is fit-for-purpose, availability of relevant and reliable data to provide more insights into chemical structure-activity relationships, or interpretation of the mechanism of action of chemical compounds (Cronin et al., 2019; Sahlin et al., 2013). Methods for analyzing epistemic uncertainties in chemical risk assessment fall into three broad tiers: qualitative, deterministic, and probabilistic methods, for which qualitative analysis of uncertainties is considered the first step in any uncertainty analysis exercise (EFSA, 2006; Sahlin et al., 2013; World Health Organization and International Programme on Chemical Safety (WHO/IPCS), 2018).

While qualitative analysis of epistemic uncertainty is an important first step in uncertainty analysis, QSAR studies predicting the toxicity of chemical compounds have hitherto focused on quantitative aspects. For example, in QSAR mutagenicity prediction, Hung and Gini (2021) applied Bayesian reasoning to quantify epistemic uncertainty in model input data – i.e., uncertainty related to the amount of data used for modeling and the availability of specific chemical information in the modeling data. Zhong et al. (2022), during QSAR development, applied the Gaussian process to quantify epistemic uncertainty with respect to the inclusion or exclusion of specific chemicals in the model training set. Similarly, Wang et al. (2021) and Zhang and Lee (2019) estimated epistemic uncertainty using Bayesian statistics – i.e., distributional uncertainty emanating from QSAR models' lack of recognition of the information in test sets, and uncertainty due to sparse or imbalanced distribution of data in model training sets. However, to my knowledge, little (if anything) has been done to qualitatively identify or categorize uncertainties that are expressed in statements in QSAR modeling studies. This is problematic as systematic and transparent accounting of uncertainties in QSAR modeling requires analyzing and addressing both quantitative and qualitative uncertainties (EFSA, 2006; Sahlin et al., 2013; WHO/IPCS, 2018).

Using the neurotoxicity endpoint as an example, the aim of this study was to develop a method that allows for systematic and transparent accounting for implicit and explicit uncertainties in QSAR modeling of chemical toxicity. The choice for neurotoxicity was based on the fact that similar to other complex toxicological endpoints, it suffers from limited experimental data, which leads to higher epistemic uncertainty (Madden et al., 2020; Worth et al., 2011a). It is also well recognized that, presently, models such as QSARs struggle to accurately predict neurotoxicity given the complex nature of the underlying biological mechanisms (Bal-Price et al., 2018; Fritsche et al., 2018; Gadeleta et al., 2022; Madden et al., 2020; Worth et al., 2011a); thus, making uncertainty analysis important for QSAR modeling of neurotoxicity. Accordingly, implicit and explicit uncertainty indicators, expressed in peer-reviewed papers on the QSAR modeling of neurotoxicity, were first identified. The indicators were then used to identify implicit and explicit uncertainties. The identified uncertainties were then systematically categorized according to the uncertainty sources proposed by Achar et al. (2024a). This allowed me to identify uncertainty sources that were most commonly highlighted by researchers. By identifying and categorizing the implicit and explicit uncertainties, I contend that this method can be used to draw attention to epistemic uncertainties in QSAR modeling of specified endpoints, here illustrated by neurotoxicity. My hope is that the information gained from this study can be used to inform decision-support initiatives by modelers and regulatory authorities when identifying research needs and the type of data required to reduce or eliminate uncertainties.

## **4.2 Methodology**

### **4.2.1 Uncertainty indicators**

Levin et al. (2004) proposed four categories of implicit uncertainty indicators in chemical risk assessment: epistemic, inferential, contentual and conditionalizing implicit uncertainty indicators (definitions of the indicators are provided in Figure S1). I considered the concepts described for the ‘implicit epistemic uncertainty indicators’ to be directly relevant to this study, as the aim is to identify epistemic uncertainties. However, the initial analysis suggested that researchers commonly do not distinguish between epistemic uncertainty (e.g., “it is presumed that ...”) and inferential uncertainty indicators (e.g., “on this basis, it is presumed that ...”). I therefore decided to also include inferential indicators as part of epistemic indicators (a detailed explanation and definition is given in Table

S4.1, Text S4.1 and Figure S4.1 of the Supplementary material). Furthermore, the concept of explicit epistemic uncertainty indicators (e.g., qualitative or quantitative statements such as “we don’t know” and it is uncertain”) described by Sahlin et al. (2021) and (Stortenbeker et al., 2019) was adopted to identify explicit uncertainty indicators (see detailed explanation in Text S1 of the Supplementary material).

Figure 4.1 outlines the process used in the present study to identify and categorize implicit and explicit uncertainties. This started with the sourcing of peer-reviewed papers (section 4.2.2), followed by the identification of epistemic explicit and implicit uncertainty indicators (section 4.2.3), whereafter, these indicators were used to highlight uncertainty in QSAR neurotoxicity papers (section 4.2.4). Finally, with the goal of supporting systematic and transparent accounting of uncertainties, I used the uncertainty sources I recently proposed (Achar et al., 2024a) (see Section 4.2.5) to categorize the identified implicit and explicit uncertainties.

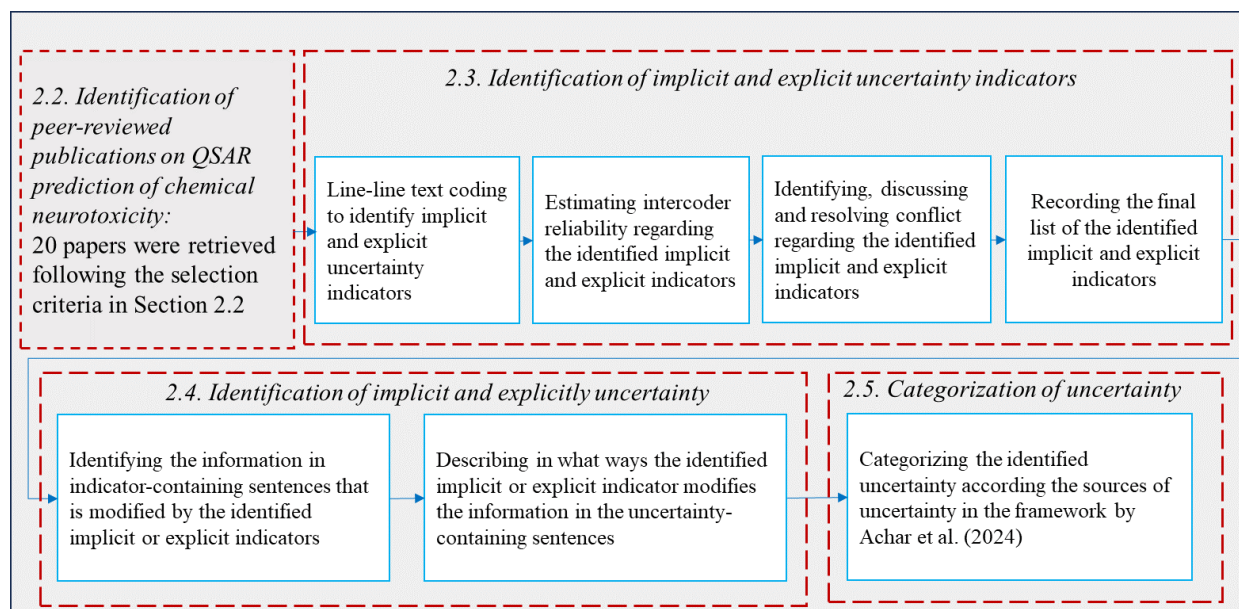


Figure 4.1. The steps undertaken to identify and categorize implicit and explicit uncertainties in peer-reviewed papers that apply QSAR to predict the neurotoxicity of chemical compounds.

#### 4.2.2 Selecting peer-reviewed papers for analysis

A search was conducted in the Web of Science Clarivate database using the following keywords and Booleans: neurotox\* (TOPIC) AND QSAR (TOPIC), while excluding review papers. This led to the identification of 75 papers.

Duplicates were excluded whereafter titles and abstracts were skimmed, and the following inclusion criteria were used: the paper had to (1) be published in a peer-reviewed journal, (2) be an original research article, (3) apply QSAR model(s), (4) address neurotoxicity assessment of a chemical (including drugs), and (5) be in the English language. After reading the papers in their entirety, 20 papers that met the selection criteria were selected for analysis (see the Supplementary material for the list of 20 papers). These articles were added to the Zotero reference management tool (version 6.0.36).

#### **4.2.3 Identification of implicit and explicit uncertainty indicators**

Implicit epistemic uncertainty indicators were identified through line-by-line text coding of the methods, results, discussion, and conclusion sections of the papers identified under section 4.2.2. These sections were selected as they are the locations in a paper where one would expect authors to describe the model(s), modeling data, parameters and variables applied in their study, report and discuss the study findings/modeling output, and draw conclusions about the study. Three coders (first, third and fourth authors) independently coded each of the 20 papers (Table S4.2), identifying, color-marking, and recording implicit and explicit indicators. Each coder reviewed each paper at least twice until no new indicators could be discovered. Thereafter, the intercoder agreement measure, calculated as the percent agreement, was estimated using Krippendorff's Alpha ( $\alpha_k$ ) (Krippendorff, 2004) on the web-based K-Alpha Calculator developed by Marzi et al. (2024). The measure relates to whether the coders agreed if the coded texts were indicators and, if yes, whether the indicators qualified to be categorized as explicit or implicit epistemic indicators. All conflicting issues were identified, discussed, and resolved before the final list of indicators was recorded (Table S4.2). I describe the uncertainty indicator identification procedure in more detail in the Supplementary materials (Text S4.2). The risk of double coding was reduced by checking whether an indicator-containing sentence was repeated in different sections of a paper (e.g., identical phrases could be used in the Results and Discussion and the Conclusion sections).

#### 4.2.4 Identification of implicit and explicit uncertainties

The indicator-containing sentences and the context in which the indicators appear were used to interpret and describe the implicit or explicit uncertainties communicated in these sentences; this was collaboratively performed by the first and second authors. Inspired by the process proposed by Zerva (2019), I started by first noting the location of uncertainty indicators in the indicator-containing sentences, followed by identifying the central piece of information communicated in the sentences. This information was identified by reading each paper in its entirety to get a general idea about the study, and then each indicator-containing sentence and, when necessary, an entire paragraph containing the indicator-containing sentence. Next, I interpreted how each indicator modified the central piece of information. The assumption is that if the central piece of information is modified by the indicator, then the information becomes uncertain (Zerva, 2019). An example of how the descriptions of the uncertainties were developed is illustrated in Table 4.1 (Step 1-3) using two of the indicator-containing sentences from two of the 20 studies (i.e., Estrada et al., 2001; Zhang et al., 2019):

Table 4.1. Steps followed to identify uncertainties implicitly and explicitly expressed in the indicator-containing sentences.

	Implicit uncertainty	Explicit uncertainty
Indicator-containing sentence	"It demonstrates that a smaller molecule is more likely to cause adverse effect to the nervous system" (Zhang et al., 2019; p 4).	"Unfortunately, there is not enough experimental data to corroborate these findings" (Estrada et al., 2001; p 457).
<b>Step 1:</b> Noting the indicator	"more likely"	"there is not enough experimental data"
<b>Step 2:</b> Identifying the information communicated	In this context, smaller molecules are more likely to cause adverse effects to the nervous systems than larger molecules	A group of chemicals were found to induce neurotoxicity in the test animal (i.e., mouse)
<b>Step 3:</b> Interpreting how the indicator modifies the information	The indicator modifies information by implicitly clarifying that it cannot be guaranteed that smaller molecules will	The indicator modifies information by explicitly clarifying that "there is not sufficient experimental data to corroborate"

---

cause more adverse effects than larger molecules. that the studied chemicals will induce neurotoxicity in mice.

---

#### 4.2.5 Categorization of the identified uncertainties

This study systematically categorized each of the uncertainties identified under Section 4.2.4 guided by the 20 uncertainty sources (Table 4.2) established within a framework I recently developed to systematically categorize general sources of uncertainty across different *in silico* toxicology methods (Achar et al., 2024a; in press). The framework conceptualizes uncertainty as a multi-source phenomenon that is associated with recognized QSAR components and modeling processes (see Figure S4.2), thus making it a valuable tool in this study to facilitate mapping out diverse sources of uncertainty in QSAR modeling exercises. In applying the framework, however, I note that Data accuracy was considered not just to entail the measure of correctness of data in relation to a “true value” (as defined in the framework – Table 4.2) but also such measure in relation to a “distribution of true values”. This adjustment was necessary to accommodate the different (implicit or explicit) descriptions of data accuracy in the 20 analyzed studies. For example, Turabekova et al. (2008), in “The bad fitting to correlation line for those compounds can be explained by possible errors [...]” (Table S4.2), implicitly mention data accuracy with respect to “distribution of true values”. In contrast, Cronin et al. (1996) in “[...] most confidence intervals are between 10% and 30% of the original value” (Table S4.2) explicitly mention it with respect to a “true value”.

Table 4.2. Sources of uncertainty (arranged in alphabetical order) relatable to practices and features common to *in silico* toxicology modeling (adapted from Achar et al. (2024a) – under review – with permission from the authors).

<b>Uncertainty sources</b>	<b>Definition of the uncertainty sources</b>
Activity/potency	Measure of elicited toxicological effect or adverse effect, degree of the effect, or ability of a chemical to exert an effect on a receptor
Activity/potency evidence	Available evidence to support the predicted activity/potency

---

---

Applicability domain	Boundaries within which a model can be applied and provide reliable and accurate predictions (e.g., adequacy of chemical structure space or category to predict effects of similar chemicals)
Chemical similarity	Resemblance or commonality between chemical compounds, e.g., in terms of functional groups and chemical structure
Chemical structure	Quality (e.g., in terms of relevance) of chemical structures or substructures with respect to a set prediction
Coverage of ADME activities	Consideration of ADME activities in biological systems, including effects of metabolites
Data accuracy	The extent to which measured data deviates from its true value
Data balance	Ratio between the number of chemicals in categories in training dataset – chemical categories with known activities (toxicants) and known non-activities (non-toxicants)
Data relevance	Data contain target information (e.g., kinetics and metabolic property) suitable for modeling or adequate for the interpretation of model predictions
Data reliability	Reproducibility of data between test approaches/sources, or reproducibility of the methodology used in generating the data
Data quantity	Amount of data - whether data is sufficiently available
Data validity	Acceptability of the method used to generate data relative to set guidelines or whether the method measured what it was intended to measure
Descriptor concordance	Degree of agreement between descriptors and other chemical features or chemical toxicokinetic or toxicodynamic properties
Descriptor relevance	Extent to which physicochemical or molecular descriptors are considered toxicologically relevant, or suitable for deriving chemical properties or for a specific prediction task
Extrapolation	Making predictions beyond the range of the observed/known data (e.g., toxicity data) in attempts to obtain new unknown data
Mechanistic plausibility	Toxic causal pathways of chemicals, involving the identification of molecular initiating events/key events linked causally to a target endpoint
Metabolic domain	Consideration of production or presence of metabolites as part of chemical interaction with biological systems
Model performance	Predictivity of a model or how well a model can predict outcomes of interest, which can be evaluated through, for example, an internal/external validation or quantitatively using the measure of statistical fit

---

Model relevance	Transferability of a model or model prediction to a different prediction context (e.g., regulatory application or prediction of new compounds)
Model structure	A model endogenous representation, such as mathematical formulations (e.g., equations or graphs), choice of algorithms, precision of numerical approximations, and relationships between variables

#### 4.2.5.1 The categorization process

I illustrate how the categorization was performed by referring to the examples in Table 4.1. For the implicit uncertainty category, the interpretation of uncertainty was not just based on the understanding of the message communicated in the indicator-containing sentence but the overall message in the paragraph containing the sentence or adjacent paragraphs as well. Accordingly, I interpreted the uncertainty (Step 3; Table 1) as: the effect of smaller molecules, relative to larger molecules, is uncertain (where molecular size is characterized using molecular weight descriptor), and (2) there is also uncertainty in knowledge about mechanistic interaction of the smaller molecule (relative to larger molecules) in the nervous system to inform judgments about their effects. In other words, as implicitly expressed by Zhang et al. (2019), uncertainty in the sentence is not only about the effect of the neurotoxicant molecules (due to exposure), but also the understanding of the mechanistic complexity of the molecules. These uncertainties fit well under the description of two of the uncertainty sources in Table 2: "Activity/potency", which is described as the "Measure of elicited toxicological effect or adverse effect, degree of the effect, or the ability of a chemical to exert an effect"; and "Mechanistic plausibility", which is described as "Toxic causal pathways of chemicals, involving the identification of molecular initiating events/key events linked causally to a target endpoint".

In the explicit uncertainty category in Table 4.1, the uncertainty noted in Step 3 relates to the evidence supporting the claim that the studied chemicals caused neurotoxic effects in mice. This fits well under the description of the uncertainty source "Activity/potency evidence", which is described as "Available evidence to support the predicted activity/potency" (Table 4.2); thus, this uncertainty was categorized as "Activity/potency evidence". The categorization process was undertaken for all uncertainties identified under section 4.2.4. The distribution and frequencies of the categorized implicit and explicit uncertainties were quantified afterward.

## 4.3 Results

### 4.3.1 Intercoder agreement

From the two papers (i.e., Amnerkar & Bhusari (2010) and Schmidt et al. (2004)) used for coding practice (see Supplementary Text S2 for details about the practice session), the three coders agreed 90% of the time on whether a statement, phrase, or word identified was an indicator of epistemic uncertainty and, if so, whether the indicator qualified to be categorized as implicit or explicit ( $\alpha_k = 0.87$ , 95% CI; N=3). The indicators identified during the practice session (Text S4.2) and the coder agreement scores are shown in Table S4.3. Similar results were obtained in the coding of the 20 papers, where the three coders, on average, agreed 87% (82 – 94%; account for each paper) of the time that the identified indicators represented implicit epistemic uncertainty ( $\alpha_k = 0.86$ , 95% CI; N=3), which, according to Krippendorff (2004), is satisfactory.

### 4.3.2 Occurrence of uncertainty indicators

Table S4.2 (second and fifth columns) shows the full list of implicit and explicit indicators (bolded in the sentences) identified from the 20 analyzed studies. A summary of their frequency of occurrence is provided in Table 4.3. A total of 406 indicators were identified: The majority (75.6%; 307) of these were implicit, and the rest (24.6%; 99) were explicit. A number of words/phrases implicitly expressing uncertainty were repeated across the 20 studies, with “suggest(s/ed/ing)”, “may”, and “may be” being the most common, while words/phrases like “implies” and “unlikely to” being among the least common (Table S4.2).

- 1 Table 4.3. Examples of the implicit and explicit uncertainty indicators (bolded in the sentences) identified in the 20 analyzed studies (see Table S2 for the  
 2 raw data). The data is arranged in the order of the publication numbers presented in Table S4.2.

Study #	Implicit uncertainty indicator (bolded in the sentence)	Page #	Frequency	Explicit uncertainty indicator (bolded in the sentence)	Page #	Frequency
1	These results <b>may indicate a certain probability</b> that compound 11 is a multitarget ligand.	1398	9	This mx-QSAR has excellent goodness-of-fit statistics [...] with <b>sensitivity (Sn), specificity (Sp), and accuracy (Ac) &gt; 80%</b> .	1394	1
2	The proposed QSAR model <b>can be a possible</b> supporting tool [...]	50	20			
3	[...] group that <b>potentially</b> explains the high number of incorrect predictions.	429	9	These results indicate that <b>it is currently difficult</b> to predict [...]	429	7
4	Surprisingly, the above equation <b>suggests</b> a lack of relationship with hydrophobicity.	105	46	[...] <b>many of these data are not available</b> at 25°C [...]	106	11
5	In the case of dioxane, that is outlier for models (2) and (3), <b>we can think</b> that neurotoxicities [...]	454	5	Unfortunately, <b>there is not enough experimental data</b> [...]	457	5
6	[...] cochlear development and <b>potentially</b> resulting in permanent auditory loss.	7	16			
7	<b>Probably</b> , the less unfavorable contacts of the ketal group inside the sub-pocket [...].	4	15	Unfortunately, <b>none of these models succeeded</b> in finding compounds more potent [...]	3	2
8	<b>To a certain extent, this indicates</b> that the structural diversity of our compounds is high [...]	167	7	Although <b>the predictive power of our model is not the best</b> [...]	168	2

9	[...] toxicokinetic properties of the chemical <b>may</b> play an important role in the neurotoxicity [...]	7	21	[...] <b>the imbalance dataset</b> was further adjusted [...]	4	12
10	[...] (62.5%) were <b>inferred to be associated with</b> Parkinson's disease [...].	3309	18	[...] <b>cannot be easily explained</b> by reduction of dopaminergic neuronal cells	3312	10
11	One compound <b>may</b> lead to 1 or more statistical cases because it <b>may</b> give different outcomes [...]	1872	7	This linear equation presented good results [...] with <b>overall Accuracy in training series above 90%</b> .	1872	1
12	The use of enzymes from different tissues and species is a <b>potential</b> limitation of the study.	232	6	However, the whole picture of influence is <b>rather complicated</b> .	236	1
13	The results obtained from the predicted model <b>could be attributed to</b> the experimental verifications.	313	12	[...] – MnAChE <b>still remain unexplored</b> .	309	2
14	<b>One tentative explanation</b> for this event <b>could be</b> related to increased hydrogen [...]	3800	11	The anticonvulsant mechanism of the semicarbazones is <b>not clearly defined</b> .	3399	2
15	The width of the REP range <b>can be roughly</b> interpreted as the lowest possible value [...]	19	10	[...] <b>experimental data are lacking adds another layer of uncertainty</b> to the NEF predictions.	14	7
16	<i>It appears that</i> no high-potency PCB congeners with EC2x values <<0.2 μM exist.	359	23	Because of the <b>poor predictivity</b> of the pEC50 QSAR, and concerns [...]	358	9
17	The bad fitting to correlation line for those compounds <b>can be explained by possible</b> errors [...]	11	23	[...] <b>no substantial features have been identified that would help</b> to distinguish [...]	5	10
18	This is <b>suggestive of the potential</b> for increased potency [...]	277	5	The source of the IC50 values [...] <b>may provide some uncertainty</b> .	230	5

19	[...] structural fragments has a <i>high possibility</i> to be neurotoxicant.	5	25	The ECFP_10 and eight molecular descriptors <b>were not able to better describe the property</b> [...]	4	7
20	The results <i>suggest</i> that nHAcc and nHDon <i>may be obviously associated</i> with drug-induced [...]	6042	17	<b>Admittedly, these methods are not perfect</b> because they [...]	6043	7
			Total =			Total = 99
			305			

3

### **4.3.3 Variation in the occurrence of implicit and explicit uncertainty sources between publications**

The use of the indicators identified under section 4.2.3, combined with the process described in section 4.2.4, allowed for the identification of implicitly and explicitly expressed uncertainties within indicator-containing statements. Each of the identified uncertainties was aligned (i.e., grouped according to) at least one of the uncertainty sources in Table 4.2 (each is henceforth identified by the uncertainty source it is categorized under) (Table S4.2, 4<sup>th</sup> and 7<sup>th</sup> columns). All but two of the uncertainty categories were not represented among the recorded uncertainty sources (Data validity and Metabolic domain). A summary of the distribution of the categorized implicit and explicit uncertainty sources across the 20 studies is provided in Table S4.4. The calculated total number of occurrences of uncertainty sources in both implicit and explicit categories was 162, of which 104 (64%) were implicit and 58 (36%) explicit. None of the uncertainty sources occurred in all the 20 analyzed studies nor was any pattern observed as related to co-occurrence. For example, although uncertainties related to Coverage of ADME effects and Extrapolation occurred in 5 of the 20 studies, they only co-occurred in two of them (Table S4.4). The most commonly occurring implicit uncertainty sources were Mechanistic plausibility and Model relevance, while Data balance and Data accuracy recorded the lowest number (each 1/104). Among the explicit uncertainty sources, Model performance was most common while seven of the sources were only mentioned once (Table S4.4).

### **4.3.4 Frequency of uncertainty sources**

EFSA et al. (2018) recommend the adoption of three tiers to describe analyzed uncertainty: (1) describe uncertainties collectively (i.e., combined uncertainty for an assessment as a whole), (2) describe uncertainties separately with regards to the main parts of the assessment, and, where possible, (3) describe uncertainties with regards to the smaller parts of the assessment. In this study, I followed these steps to describe the frequencies of the uncertainty sources.

#### 4.3.4.1 General distribution of uncertainty sources

The total number of times a specific uncertainty source (i.e. implicit and explicit) was expressed in the 20 studies is given in Figure 4.2. Mechanistic plausibility was by far the most common, with over 90 occurrences, followed by a group of four sources that were referred to in about 50 occurrences (Descriptor concordance, Model relevance, Model performance, and Activity/potency). These results suggest that a large number of uncertainties in QSAR prediction of neurotoxicity fall within these frequently mentioned sources. In contrast, Data balance, Data accuracy, and Chemical similarity were among the least frequently mentioned sources, not to mention Data validity and Metabolic domain, which were not expressed in the analyzed papers, suggesting that these uncertainty sources are not of great concern to researchers in the field.

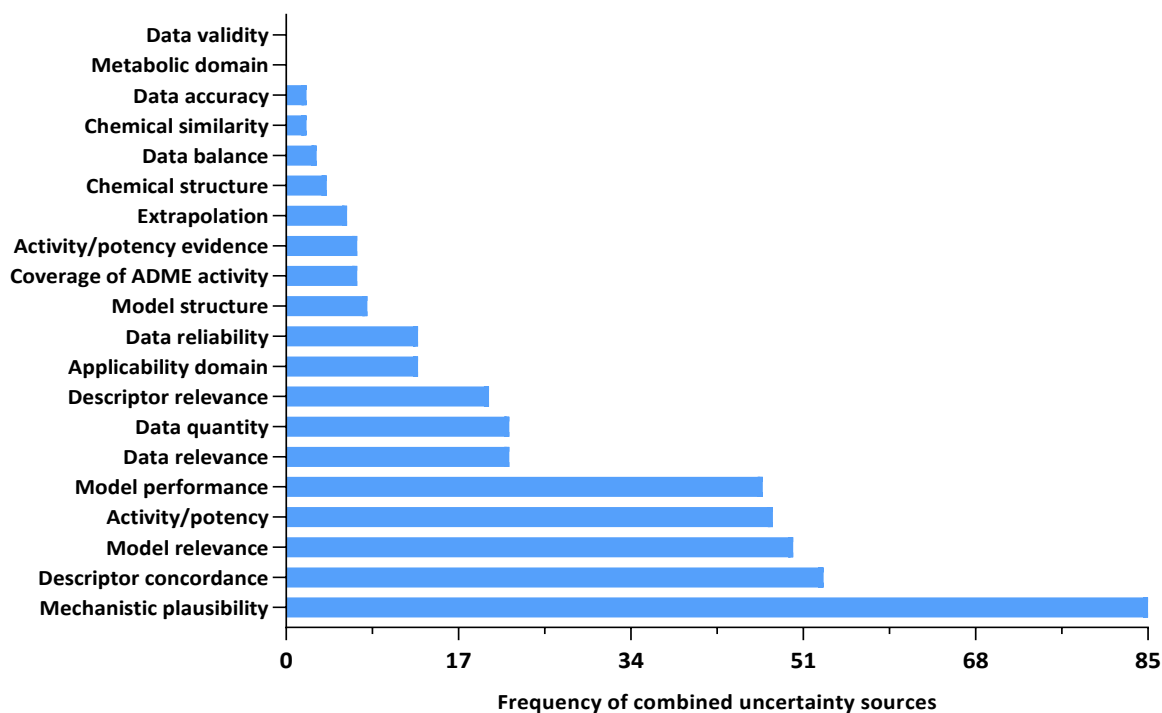


Figure 4.2. Frequencies of the combined (implicit + explicit) uncertainty sources. The data are arranged from the highest to the lowest value, according to the magnitude of the combined frequencies.

#### 4.3.4.2 Comparison of frequencies relating to implicit and explicit uncertainty

The frequencies of uncertainty sources expressed implicitly and explicitly were analyzed in order to determine the extent of any variation (Figure 3). The most frequent uncertainty source that was implicitly expressed was Mechanistic plausibility (73/310). Other high-frequency sources included Descriptor concordance (51/310), Model relevance (39/310), Activity/potency (32/310), and Model performance (27/310). As noted earlier, Data validity and Metabolic domain were not mentioned in any of the analyzed studies. Among sources that were explicitly referenced, Model performance was the most frequent one (20/102), followed by other sources such as Data quantity (18/102), Activity/potency (16/102), and Mechanistic plausibility (12/102). In contrast, Data accuracy, Chemical structure, and Model structure (each occurring only once) were among the least frequent. Similar to the implicitly expressed uncertainty sources, Data validity and Metabolic domain were not referred to at all. Implicit and explicit mentions were equally common for 2/20 of the sources (i.e., Data accuracy and Extrapolation), while explicit mentions were more common for only 4/20 sources (i.e., Data quantity, Data balance, and Activity/potency evidence) (see Figure 4.3). Taken together, the analysis shows that implicit mentions were more common for the majority of the uncertainty sources (13/20; 65%), which indicates that uncertainty is more commonly expressed implicitly in QSAR studies predicting neurotoxicity of chemicals.

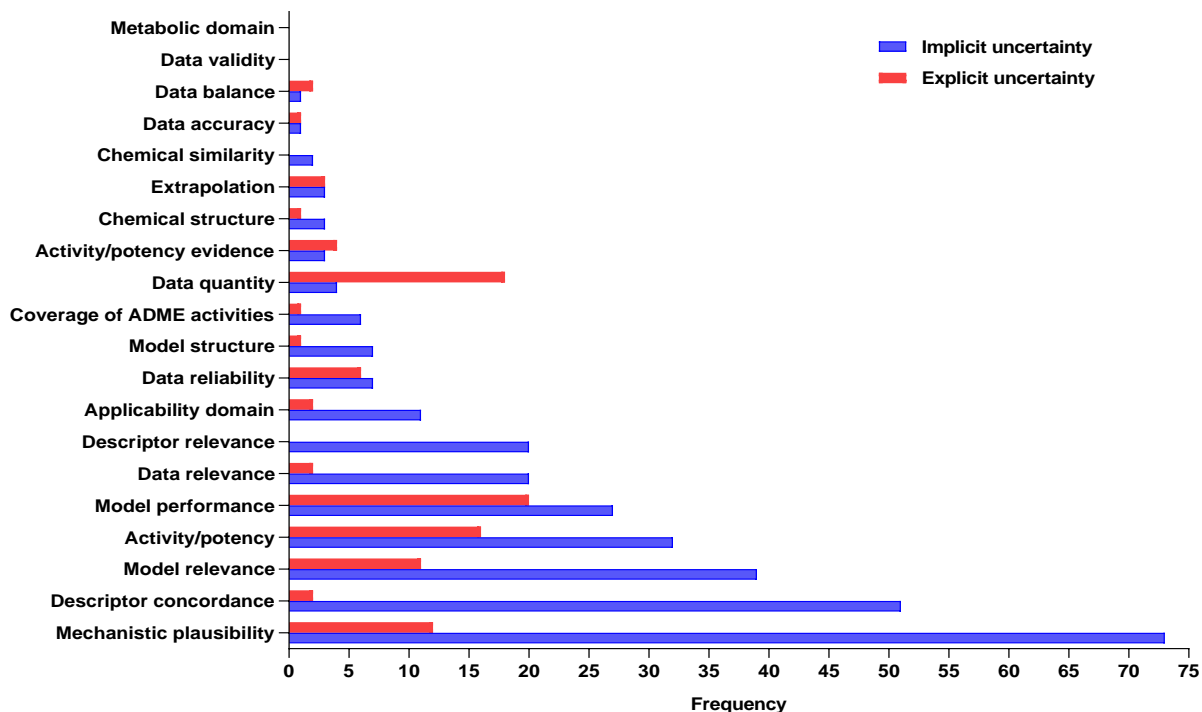


Figure 4.3. Frequency of the uncertainty sources in the implicit and explicit uncertainty categories (arranged according to the magnitude of frequency of implicit sources).

#### 4.4 Discussion

QSARs for toxicity prediction have become ubiquitous in chemical safety assessment. Understanding the uncertainties in QSAR models and their predictions is a vital and fundamental part of assessing the quality and robustness, or otherwise, of a model for its successful use, for instance, in regulatory contexts. This study evaluated 20 papers relating to QSARs for neurotoxicity prediction to assess the occurrence of implicit and explicit uncertainties expressed in them. The study results are discussed below.

##### 4.4.1 Contribution of implicit versus explicit uncertainty sources to the overall uncertainty sources

Figure 4.2 and Figure 4.3 indicate a variety of uncertainty sources and the different frequencies in which they are mentioned, ranging from relatively frequent to no mention at all. These results raise three questions: Why are the uncertainties variably expressed? What is the contribution of the implicit versus

explicit uncertainty sources to the overall uncertainty? With respect to uncertainty communication in QSAR studies predicting neurotoxicity, what is the possible explanation for why implicit uncertainty is more frequently expressed than explicit uncertainty?

According to Han et al. (2011) and Maxim (2015), the first question (i.e., Why are the uncertainties variably expressed?) can be answered by considering the tendency of studies to prioritize communicating particular uncertainties over others or the inherent difficulty of including all relevant uncertainties in a study. A similar observation has been made within modeling contexts. For example, studies have found that there is a higher tendency to communicate uncertainties related to model parameters than those related to, for example, extrapolation of laboratory experimental data to humans or field settings, suitability or robustness of models or model structure for an intended prediction, or inaccuracies in the experimental design used for data generation (Maxim, 2015; Moschandreas and Karuchit, 2002; National Research Council, 2009; Verdonck et al., 2005). According to Kirchner et al. (2021), this commonly results from modelers' assumption that the added value of including particular uncertainty sources is not worth it, as it may push models to unrealistic and extreme boundary solutions, or for reasons such as resource or computational constraints.

The answer to the second question – i.e., What is the contribution of the implicit versus explicit uncertainty sources to the overall uncertainty? – can be deduced from the distribution of the frequency datasets in Figure 4.2 and Figure 4.3, which indicates that the frequencies of the implicit uncertainty sources were generally more than explicit sources. Based on the analyzed studies and with respect to these results, it can thus be concluded that implicit uncertainties contributed more to the overall uncertainties. Indeed, these results are consistent with the frequencies of the indicators, which were more in the implicit uncertainty category than the explicit uncertainty category (Table 2). Previous research reported similar findings. For example, Flari and Wilkinson (2011), in the text analysis of uncertainties

expressed in EFSA documents on health risk assessments, found more implicit uncertainties (972/1133) than explicit uncertainties (161/1133). Stortenbeker et al. (2019) also recorded  $\approx 1.54$ -fold more implicit uncertainties than explicit uncertainties during physician explanations of medical symptoms to patients.

The third posed question is: With respect to uncertainty communication in QSAR studies predicting neurotoxicity, what is the possible explanation for why implicit uncertainty is more frequently expressed than explicit uncertainty? To answer this question, as a starting point, it might be useful to consider possible reasons why modelers would preferentially express uncertainties implicitly. Studies on uncertainty communication in scientific studies offer different explanations that may be relevant to this discussion. For example, according to Dhimi and Mandel (2022), the preference to implicitly communicate uncertainties is based on modelers' attempts to maintain the perceived credibility of their research, especially in the event that erroneous predictions are made. For instance, using the phrase "it may not occur" to express uncertainty about unexpected prediction output may lead to less credibility loss than "there is uncertainty about the predicted results". van der Bles et al. (2020) and the National Academies of Sciences, Engineering and Medicine (2017) similarly interpreted the tendency for implicit communication of uncertainty, or reluctance to communicate uncertainty, to stem from researchers' efforts to avoid signalling incompetence or inviting criticism regarding their research based upon the presence of uncertainties. In the context of this study, the reluctance to openly communicate uncertainties could also be taken more broadly to imply uncertainty communication bias among QSAR modelers of neurotoxicity (Steijaert et al., 2021); this is based on their prioritization to communicate implicit sources over explicit sources. Cronin et al. (2019) noted that such bias can negatively affect QSAR models and their use. For example, implicit recognition of uncertainties in the prediction output can give users (e.g., regulators) a false sense of accuracy in the output, or else make it difficult to identify model inputs that require additional data to reduce propagated uncertainties in the outputs.

The variation in the frequencies of implicit versus explicit uncertainties also reflects a paradox between the preference to implicitly communicate uncertainty for reasons such as the perceived negative consequences of explicit communication of uncertainty (Dhami and Mandel, 2022), and the need to explicitly communicate uncertainty to enhance transparency about risk and uncertainty (EFSA et al., 2018; Flari and Wilkinson, 2011). This study aligns with this call to explicitly communicate uncertainty. I argue that implicit communication of uncertainties associated with the QSAR prediction of neurotoxicity undermines transparent assessment of, for example, the validity of the models as well as their accuracy. An explicit expression of uncertainty is particularly important during fitness-for-purpose evaluation of QSAR models, as this can guide explicit characterization of relevant uncertainties to improve defensibility of a predicted output and provide a critical basis for informed decision-making on the need for appropriate measures to reduce potential risk of neurotoxicants and the nature of the measures (Belfield et al., 2021; Cronin et al., 2019; WHO/IPCS, 2018).

#### **4.4.2 Level of concerns raised about the uncertainty sources**

The data presented in Figure 4.2 suggest that Mechanistic plausibility constitutes the area of most concern for uncertainty for QSAR modeling of neurotoxicity. To my knowledge, no study has specifically targeted analysis of uncertainty in QSAR prediction of neurotoxicity of chemicals; nevertheless, scholars have expressed concern about uncertainties in mechanistic characterization of substances, which corroborates the findings from this study. For example, Crofton et al. (2022), in their review of the current status of the application of *in silico* approaches towards developmental neurotoxicity, suggest that the incomplete understanding of the underlying mechanisms behind the emergence of adverse outcomes seems to constitute uncertainties in understanding adverse outcomes pathways related to developmental neurotoxicity. Others have similarly suggested that the inability to identify or interpret the mechanisms of actions of neurotoxicants, to an extent, contributes to uncertainties in neurotoxicity assessment (Mundy et al., 2015; Worth et al., 2011b). This especially seems to be the case for developmental

neurotoxicity, where exposure to neurotoxicants during brain development or the developmental window further complicates the understanding of the underlying mechanisms of the adverse effects (Bal-Price et al., 2018; Fritsche et al., 2018; Mundy et al., 2015).

This study also suggests that other highly mentioned uncertainty sources (e.g., Descriptor concordance, Model relevance, and Model performance) also constitute major areas of concern for uncertainty in relation to neurotoxicity prediction. For example, a number of QSARs have been developed to support neurotoxicity prediction based on the statistical correlation between blood-brain barrier penetration of compounds, and specific neuronal bioactivities (Worth et al., 2011a). Most of these models are based on descriptors such as *in vivo* Log BB (blood-brain barrier), Log PS (permeability-surface area), unbound brain-to-plasma partitioning coefficient, as well as physicochemical descriptors (e.g., lipophilicity, hydrogen bonding, and polar surface area) (Crofton et al., 2022). However, the predictive performance (and consequently the relevance) of the individual or combined models is considered limited. Such limitations have been attributed to, for example, a lack of (relevant) data to establish robust models and inadequacy of the descriptors to support the interpretation of observed adverse effects (Crofton et al., 2022; Worth et al., 2011a). Elsewhere, Bal-Price et al. (2018) note that the fact that there are only a few QSAR studies on the effects of chemical compounds on the peripheral and central nervous systems suggests uncertainties in the characterization of hazard and risk potential neurotoxicants.

Taken collectively, given the limited research on uncertainty in QSAR prediction of neurotoxicity of chemicals, the study findings provide a tentative conclusion that the frequencies with which the four sources (Mechanistic plausibility, Descriptor concordance, Model relevance, and Model performance) are mentioned reflect the state of uncertainty of most concern in this field, and possibly a reflection of their importance in the OECD Principles for the Validation of QSARs (OECD, 2007) or their role in regulatory application of the models. It is, therefore, reasonable that when setting priorities aimed at addressing

uncertainties within the field, especially under conditions of limited resources, these sources should be the primary focus.

A number of uncertainty sources shown in Figure 4.2 had relatively low frequencies, with two of them not mentioned at all. Notably, some of these low-frequency sources are related to elements that are commonly considered in model training and test sets (i.e., data balance, accuracy, validity and chemical similarity). The importance of these uncertainty sources is well recognized, given their direct influence on the level of model predictive accuracy, reliability, and adequacy (Cronin et al., 2019; Madden et al., 2020; Pham et al., 2019; Worth et al., 2011a). An example is uncertainty related to Data validity, which emanates from the quality of the experimental studies from which QSAR modeling data are obtained (Achar et al., 2024b; Belfield et al., 2021; Cronin et al., 2019; EFSA, 2006; Karmaus et al., 2022; Madden et al., 2020; Nelms et al., 2020; Pham et al., 2019; WHO/IPCS, 2018; Worth et al., 2011b). The fact that this uncertainty source was not mentioned in the 20 studies, even though it is recognized as a high-concern uncertainty source (Worth et al., 2011b), shows that its analysis must be reviewed in light of relevant literature before drawing conclusions about how to prioritize research needs within the field of QSAR prediction of neurotoxicity; otherwise, further uncertainty might be introduced into model predictions interpretation. Furthermore, where applicable, I believe it might be most useful for modelers to refer to “ignorance”, or lack of knowledge. For example (in this case of Data validity and Metabolic domain), acknowledge that uncertainties related to these sources are not considered in a study due to modelers’ lack of knowledge around them or, if that is the case, acknowledge that these sources are not accounted for reasons such as to make model prediction less complex or easy to interpret.

#### **4.4.3 Further consideration of the categorized uncertainty sources**

In chemical risk assessment, regulatory authorities such as EFSA (2006) and World Health Organization and International Programme on Chemical Safety (2018) recommend that each uncertainty be analyzed

at one of the three tiers: qualitative, deterministic, or probabilistic. Initially, uncertainties may be analyzed qualitatively to support initial judgements about the extent of the uncertainties or support subsequent steps on quantitative estimation of the uncertainties to the extent that is scientifically feasible (EFSA, 2006; WHO/IPCS, 2018). The emphasis here is that it might not be possible to treat all uncertainties quantitatively and that qualitative consideration of uncertainties can give insights into unquantifiable uncertainties, as well as their impact on the overall confidence associated with the assessment outcomes. The critical question in handling such an impact is whether the level of uncertainty or level of confidence is acceptable. While I did not explore the level of uncertainty in the analyzed studies (based on, for example, the weight, type, and consistency of scientific evidence presented (EFSA, 2018)), for the sake of the discussion here, I assume that the magnitude of the frequencies of the uncertainty sources (Figure 2) reflects the possible level of uncertainty.

The question above cannot be answered without clearly defining the context in which a decision has to be made. The guidance provided by the OECD constitutes a conceptual basis through which one can judge the acceptability of the characterized uncertainty: the OECD principles for the validation of QSARs describe information that is considered useful for assessing the models and model predictions for regulatory purposes (OECD, 2007). The OECD's (Q)SAR Assessment Framework (QAF) also provides guidance to assess QSAR results from multiple predictions to facilitate the characterization of levels of uncertainty associated with different models and model prediction elements based on semi-quantitative uncertainty scales (i.e., "low", "medium", or "high"), as well as support the determination of whether the characterized levels of uncertainty are acceptable within a given context of regulatory decision-making (Gissi et al., 2024; OECD, 2023). However, considering that these guidance in themselves do not define criteria for characterizing uncertainty, it is challenging to define fixed acceptability criteria, as this depends on the intended regulatory use of a model or model predictions (Achar et al., 2024b; Belfield et al., 2021; Cronin et al., 2019). For example, high-level uncertainty might be tolerated in hazard screening to inform

risk assessment but not in the mechanistic characterization of a model prediction, which requires high levels of certainty, reliability and model validation (Bal-Price et al., 2018; Belfield et al., 2021). The OECD Handbook (Annex 1) provides guidance regarding areas within neurotoxicity assessment that might tolerate different levels of uncertainty levels (OECD, 2018b). The templates proposed in the literature (Belfield et al., 2021; Cronin et al., 2019; Cronin et al., 2022) also provide useful guidance to developers and users of QSARs on possible ways of judging the acceptability in regulatory decision-making contexts. Defining criteria to judge the acceptability of the characterized uncertainty sources is, however, beyond the scope of this study – it thus remains for future studies to explore this topic by defining how the different levels of uncertainty can fit into a defined decision context.

#### **4.4.4 Implication of the proposed method for uncertainty analysis in QSAR modeling**

The uncertainty identification and categorization structure proposed in this study provides a method for identifying implicit and explicit uncertainties expressed in QSAR modeling studies. The method points to one major implication for modelers and decision-makers navigating uncertainties expressed in the studies. That is, analysis of implicitly or explicitly expressed uncertainties is a three-step process. When analyzing these uncertainties, assessors should: (1) first identify uncertainty indicators and (2) then analyze the corresponding uncertainty communicated within the context of the indicators, whereafter (3) these uncertainties are categorized in a systematic manner. As discussed in this study, identifying uncertainties through this process can help identify an important blind spot in the analysis of uncertainty within QSAR modeling – i.e., lack of transparent accounting of uncertainty. Furthermore, when it comes to epistemic uncertainty, Janzwood (2023) emphasizes the need to develop methods that enable analysis of as much uncertainties as possible in order to inform decision-makers not only about the presence of the uncertainties but also regarding their sources. I believe that the three-step process described in this study is rigorous to capture a wide range of uncertainty sources expressed in QSAR modeling studies.

This study also highlights that quantitative analysis of uncertainty, which is widely considered important and recommended in the literature, can only provide a partial account of uncertainty sources within QSAR modeling. Based on the outcome of this study, it seems reasonable that QSAR modeling studies should accompany quantitative uncertainty analysis (e.g., probabilistic uncertainty analysis) with the proposed uncertainty analysis method in order to account for uncertainties that are not possible to quantify (EFSA, 2006; WHO/IPCS, 2018). To my knowledge, the present study is the first to develop such a method; the findings from this study may thus prove useful (especially in regulatory contexts) in further assessment of the level of confidence in a study outcome.

While research is important in addressing areas of uncertainties, the implicit expression of uncertainties still creates difficulties in evaluating or drawing conclusions on the level of confidence in models and their predictions (Flari and Wilkinson, 2011; Levin et al., 2004; Zerva, 2019). A possible way to improve transparency about these uncertainties is to employ systematic ways (as described in this study) of identifying and categorizing the uncertainties and, where possible, quantifying the uncertainties. Where uncertainties cannot be quantified, I recommend that, at minimum, QSAR modeler should acknowledge this uncertainty explicitly – this is in line with the working principles of EFSA, where transparency in communicating uncertainty is inextricably linked to the credibility of any reported risk assessment output (EFSA, 2006).

#### **4.4.5 Potential limitations of the developed method and future work**

While this study presents a promising and reproducible method, a few questions remain unanswered. First, I assume that implicit uncertainty is expressed through hedging words. This means I do not distinguish between expressions where the authors are genuinely uncertain about something and when it is more a matter of convention. Vold (2006) notes that hedging words and expressions in English can indicate cautiousness, tentativeness, or politeness; thus, such expressions do not necessarily signal

uncertainty; instead, they can form part of the convention of expressing politeness or a humble attitude in academic writing. Providing this distinction was out of the scope of the present study but poses a challenge for future studies intending to automate the identification of uncertainty indicators – this challenge was also highlighted by Shanahan et al. (2006). To address this weakness, like EFSA (2006), I recommend harmonizing uncertainty expressions to minimize inconsistent perceptions of uncertainty in QSAR modeling studies. This can be facilitated by including glossaries of the qualitative expressions used and defining whether each expression implies uncertainty. Additionally, it is important for each study to clearly characterize the possible impact of uncertainties attributed to the expressed uncertainties and where such impact cannot be characterized, it is necessary to report that this is the case and that the conclusions drawn from an assessment are conditional on assumptions about the expressed uncertainties.

Although this study concludes that uncertainties related to Data validity and Metabolic domain sources were not expressed in the studies (see Figure 4.2 and Figure 4.3), it is also possible that the conceptual breadth of uncertainty indicators used in this study was not adequate to capture these uncertainties. A possible way to overcome this challenge is to complement the use of the indicators with other ways of analyzing uncertainty, including a systematic analysis of the quality, type, consistency, and amount of the evidence presented by the researchers in the analyzed studies about particular claims. EFSA (2018) similarly recommended incorporating different methods for uncertainty analysis to improve the quality and robustness of the analysis.

Finally, from the definition of Model performance by Achar et al. (2024a) (see Table 4.2), I assumed that only epistemic factors influence model performance. In reality, however, it can also be the case that such an influence (in)directly arises from, for example, data variability. While it is important to make this distinction in order to accurately characterize the context in which uncertainty in model performance is considered, this was not done by Achar et al. (2024a); thus presenting a potential limitation of the use of

the study. It is also important to note that, as explained under Section 4.1, I used neurotoxicity because it is a complex endpoint that is presently difficult to predict (especially in mechanistic terms) in QSAR. This suggests that the data distributions in Figure 4.2 and Figure 4.3 might depend on the endpoint used. This potential variation was not explored in the current study but remains for future studies to investigate.

#### **4.5 Conclusion**

This study aimed to identify and categorize implicit and explicit uncertainties expressed in studies that use QSAR models to predict the neurotoxicity of chemical compounds. It was found that most of the identified indicators were implicitly expressed (310), compared to those expressed explicitly (102). This indicates that, within studies on QSAR prediction of neurotoxicity of chemicals, it is considerably more common to express uncertainties implicitly than explicitly. Four uncertainty sources were most commonly referred to: Mechanistic plausibility, Descriptor relevance, Model performance, and Model relevance. This suggests that researchers are concerned about uncertainties related to, for example, predicting the adverse health risks of compounds based on mechanistic knowledge, the applicability of models (e.g., in terms of their relevance), developing models that can adequately predict external data sets or *in vivo* data, and the descriptors used in model development and predictions. It was noted that some of the uncertainty sources that were rarely noted, or not noted at all (e.g. Data validity and Metabolic domain) are in fact flagged as a concern in the broader QSAR literature as areas that researchers commonly overlook. This implies that while analysis of expressed uncertainty can help identify areas of uncertainties, conclusions can only be drawn after the output has been analyzed in light of the broader QSAR literature. Overall, the findings from the present study cannot only be used to guide modelers during modeling processes to prioritize uncertainty sources for uncertainty analysis based on the magnitude of their frequencies but also facilitate a structured dialogue between modelers and decision-makers about the need for more research to improve existing models or develop new ones that can reduce these uncertainties.

## Chapter 5: Conservative consensus QSAR approach for the prediction of rat acute oral toxicity

### 5.1 Introduction

With the advances in toxicology towards the replacement of animal testing with suitable alternatives, quantitative structure-activity relationship (QSAR) models have been developed to support the prediction of a number of toxicity endpoints. Acute oral toxicity (measured by the lethal dose that kills 50% (LD<sub>50</sub>) of test animals) is one of such endpoints for which a number of QSARs have been developed (Bercu et al., 2021; Zhu et al., 2009; Zwickl et al., 2022). Currently, rat LD<sub>50</sub> data are commonly used as the primary benchmark to, for example, establish acceptable human exposure limits, guide the classification of chemical hazards, assess the potential risk of accidental ingestion of chemical toxicants, or set appropriate doses for repeat dose toxicity assessments (Strickland et al., 2018; Walum, 1998; Zhu et al., 2009). QSARs for rat acute toxicity are commercially and publicly available (Gonella Diaza et al., 2015; Mansouri et al., 2021). In this investigation, two models – Toxicity Estimated Software (TEST) and Collaborative Acute Toxicity Modeling Suite (CATMoS) – were considered owing to their free accessibility and their history of use in regulatory contexts (Gonella Diaza et al., 2015; Mansouri et al., 2021; US EPA, 2015). TEST (containing Hierarchical clustering, Nearest neighbor, and Consensus methods) was developed by the United States Environmental Protection Agency (US EPA) (US EPA, 2015), while CATMoS (a consensus-based tool consisting of binary, categorical, and continuous models) was developed by the US National Toxicology Program (US NTP) (Mansouri et al. (2021).

A small number of studies have assessed the reliability of the TEST and CATMoS models for the prediction of oral rat LD<sub>50</sub>. For example, Nelms et al. (2020) analyzed the predictive performance of the TEST Consensus model predictions, Firman et al. (2022) assessed the prediction performance of the TEST Hierarchical clustering model, while Bishop et al. (2024) estimated the accuracy and reliability of CATMoS

predictions. The general conclusion from these studies was that the predictive abilities of the models, in terms of accuracy and hazard classification sensitivity, were inherently hindered by associated uncertainty. Consequently, when used to assign health-protective (conservative) LD<sub>50</sub> values to compounds, it is advised that one should focus on the extent to which the models can accurately predict and differentiate between low and high hazards in chemicals. In other words, under such conditions of uncertainty, it is most useful to apply the model with low incidences of predicting chemicals to be less toxic than the corresponding experimental data imply (Graham et al., 2021; Moudgal et al., 2023; World Health Organization & International Programme on Chemical Safety (WHO/IPCS), 2018). This is particularly the case where the principle of being "conservative" is applied to account for uncertainty by giving precaution to the predicted data (EFSA et al., 2018).

In order to use TEST and CATMoS conservatively, the challenge remains to determine which model to reliably use for such a purpose. This is because each model has unique attributes which might influence the accuracy of its prediction output. Among these may be errors, limitations and biases stemming from factors including its parameters and structure, its training data or its training process (Mansouri et al., 2021; US EPA, 2015). One possible way to address this challenge is to use a consensus approach that combines individual model outputs into a single prediction (Graham et al., 2021; Moudgal et al., 2023). The underlying premise of consensus QSAR modeling is that individual models, because of their reductionist nature, only account for limited structure-activity information of chemicals (as encoded in their structures and in the molecular descriptors used); consequently, combining their predictions will potentially improve the overall reliability against the same data (Valsecchi et al., 2020). In the QSAR literature, consensus modeling has been established through combinatorial approaches that apply multiple statistical methods within a model software, or use several descriptors (Hewitt et al., 2007; Lunghini et al., 2019; Mansouri et al., 2021; Schieferdecker et al., 2024; Zhu et al., 2009). However, to my

knowledge, little (if anything) has been done to derive conservative consensus model predictions based on a simple comparison of TEST and CATMoS predictions and selection of the more conservative (more toxic) chemical-specific predictions as representative of health-protective values.

The aim of this study was, therefore, to assess the performance of a consensus approach of TEST and CATMoS against the models individually for the prediction of a conservative oral rat LD<sub>50</sub> for a large, diverse number of organic compounds. To this end, prediction accuracy for hazard classification using the consensus approach was used to evaluate performance. The application of this approach lies in its ability to establish a foundation to contextualize the use of consensus modeling for deriving health-protective oral rat LD<sub>50</sub> estimates under conditions of uncertainty, especially where experimental data are limited or lacking.

## **5.2 Methods**

### **5.2.1 Data sourcing**

Oral rat LD<sub>50</sub> data relating to 8,186 organic compounds, each with Chemical Abstract Service Registration Numbers (CASRN), were obtained from Firman et al. (2022). These data were originally collated from different sources via the efforts of the US EPA and National Toxicity Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), and consisted, in the majority of instances, of two or three distinct point estimates of experimentally derived LD<sub>50</sub> (expressed in mg/kg) (Gadaleta et al., 2019; Kleinstreuer et al., 2018; Nelms et al., 2020). The data were processed in the previous analyses, removing duplicates, compounds without defined structures and inorganic compounds, correcting transcription errors, and retrieving SMILES and CASRN from the US EPA's CompTox Chemicals Dashboard or other public resources (details about the processing steps can be found in Nelms et al., 2020). In this study, the data obtained from Firman et al. (2022) were further processed by removing organometallic

substances and entries with multiple CASRN identifiers. Additionally, only compounds that could be predicted in both TEST Consensus and CATMoS were retained. The final dataset consisted of 6,410 organic molecules (raw data available here: [Supplementary material.xlsx](#)).

Although some of these data were used to develop the TEST and CATMoS models, the empirical LD<sub>50</sub> in the models' training sets do not exactly match the experimental LD<sub>50</sub> data from Firman et al. (2022). Furthermore, as noted by Bishop et al. (2024), when making predictions for compounds already in these models, consensus models do not take the exact experimental values as the predictions. Instead, they generate predictions based upon consensus of the individual models within each. That is, the model consensus predictions are generated through multistep mathematical simulations that are not based on any specific empirical value in the model datasets, meaning that the overlapping compounds between the data from Firman et al. (2022) and the model training set do not necessarily affect the interpretation of model prediction results (Bishop et al., 2024; García-Jacas et al., 2019; Karmaus et al., 2022).

### **5.2.2 Prediction of the oral rat acute toxicity**

The oral rat LD<sub>50</sub> of the 6,410 compounds were predicted in TEST software (v5.1.2) using the CASRN identifiers as input. The compounds were first split into batches, each containing about 500 entries. Preliminary analysis indicated that, in not exceeding ~500 compounds per prediction exercise, memory issues within TEST and CATMoS were avoided. The prediction options in TEST were set as: endpoint – oral rat LD<sub>50</sub>, method – consensus, and fragment constrain – relaxed. The TEST Consensus method (average of predictions generated by Hierarchical clustering and Nearest neighbor methods) is considered the most reliable (US EPA, 2015); thus, only TEST Consensus (henceforth simply called TEST) predictions (expressed in-mg/kg) were downloaded and saved. The compound batches described above were used in CATMoS (available within the OPERA App, v2.9). Accordingly, the CASRN identifiers were first saved in text files

before importing into CATMoS. The predicted LD<sub>50</sub> (expressed in mg/kg) outputs from each batch were compiled into one Excel file.

### **5.2.3 Deriving Conservative Consensus Model (CCM) predictions**

The minimum prediction concept for conservativeness was applied to select a lower (conservative) LD<sub>50</sub> value for each compound across TEST and CATMoS predictions (Gromek et al., 2022). "Consensus", was used here to refer to the lower LD<sub>50</sub> value of a compound that was obtained by comparing a compound's predictions across and TEST and CATMoS and selecting the lower (more toxic) value of the compound (Khan et al., 2019). Accordingly, the predicted LD<sub>50</sub> values from TEST and CATMoS were compared side by side to identify the lower LD<sub>50</sub> value for every compound. The lower LD<sub>50</sub> value identified was then assigned to the compound as the "conservative consensus model (CCM) prediction". Table 5.1 shows examples of four compounds obtained from Firman et al. (2022), alongside their experimental and TEST- and CATMoS-predicted LD<sub>50</sub> values, their CCM predictions derived through a consensus of the predictions from TEST and CATMoS, and an illustration of how each model's prediction accuracy was estimated (detailed description of accuracy estimation is provided in Section 5.2.4).

Table 5.1. Examples of four compounds with their experimental, TEST and CATMoS LD50 and GHS predictions. CCM is derived by selecting the lower value of a compound across TEST and CATMoS predictions where applicable.

Compound	CASRN	Experimental		TEST		CATMoS		CCM	
		LD <sub>50</sub> (mg/kg)	GHS category	LD <sub>50</sub> (mg/kg)	GHS category	LD <sub>50</sub> (mg/kg)	GHS category	LD <sub>50</sub> (mg/kg)	GHS category
Spinosyn A	59375-67-6	3,740	5	34	2*	2,467	5†	34	2*
Enflurane	13838-16-9	5,465	NC	1,114	4*	41	2*	41	2*
Pyrimitate	5221-49-8	125	3	186	3†	114	3†	114	3†
Endothal	145-73-3	38	2	915	4**	2	1*	2	1*

NC: Not classified

†Match

\*Over-prediction

\*\*Under-prediction

#### 5.2.4 Model predictive accuracy for hazard classification

Predictive accuracy was assessed based on the extent to which TEST, CATMoS and CCM predictions agreed with corresponding experimental data (Hoffmann et al., 2010). This was performed by first assigning the compounds to one of the Globally Harmonized System (GHS) categories (see Table 5.2) (United Nations, 2021), based upon their experimental, TEST, CATMoS and CCM LD<sub>50</sub> values. Subsequently, the predicted GHS vs. experimental GHS category was compared for each compound. An accurate prediction was considered to have occurred where a model GHS classification matched the experimental classification (see the illustration in Table 5.1). An under-prediction was interpreted to occur where a compound was predicted to be less toxic (i.e., have lower hazard classification) than the corresponding experimental data indicate, while an over-prediction – synonymously called a “conservative” prediction – occurred where a

compound was predicted to be more toxic (i.e., have higher hazard classification) than the corresponding experimental data indicated (Table 5.1) (EFSA, 2018; Graham et al., 2021).

Table 5.2. GHS classification criteria and associated hazard statements for acute oral toxicity.

<b>GHS Category</b>	<b>Hazard statement</b>
1 (LD50 ≤ 5 mg/kg)	Fatal if swallowed
2 (5 < LD50 ≤ 50 mg/kg)	Fatal if swallowed
3 (50 < LD50 ≤ 300 mg/kg)	Toxic if swallowed
4 (300 < LD50 ≤ 2000 mg/kg)	Harmful if swallowed
5 (LD50 > 2000 ≤ 5000 mg/kg)	May be harmful if swallowed
NC LD50 > 5000 mg/kg	Not classified

NC: Not classified

## 5.3 Results

### 5.3.1 Applicability domain

The OECD principles for the validation of (Q)SARs require model predictions to be within a defined applicability domain in order to be considered reliable (OECD, 2007). CATMoS automatically checks the applicability domain of entries, where only compounds lying within the model's applicability domain are reported (Mansouri et al., 2021). The analysis indicated that all 6,410 compounds were within the model's applicability domain. TEST Consensus predictions are derived by averaging the outputs from Hierarchical clustering and Nearest neighbor methods, where only compounds predicted in both Hierarchical clustering and Nearest neighbor are considered as the most reliable and as being within the TEST Consensus method's applicability domain (Noga et al., 2023; US EPA, 2015). This also indicated that the full data set of 6,410 compounds were within the applicability domain of TEST Consensus. Thus, for the CCM, a compound was defined to be in its applicability domain if this was the case for both TEST Consensus and CATMoS models; thus, the complete data set was in the CCM applicability domain.

### 5.3.2 Comparing model predictive accuracy for hazard classification

#### 5.3.2.1 Agreement with experimental data

Figure 5.1 shows the distribution of the GHS categories for the 6,410 compounds classified based upon the experimental LD<sub>50</sub> data. This distribution indicates that the majority (~39%; 4506/6410) of the compounds fall within category 4, with the least number within category 1 (~3%; 220/6,410). As explained in Section 5.2.4, model predictive accuracy was evaluated based upon the agreement of predicted LD<sub>50</sub>-based GHS categories with the corresponding experimental LD<sub>50</sub>-based GHS categories depicted in Figure 5.1.

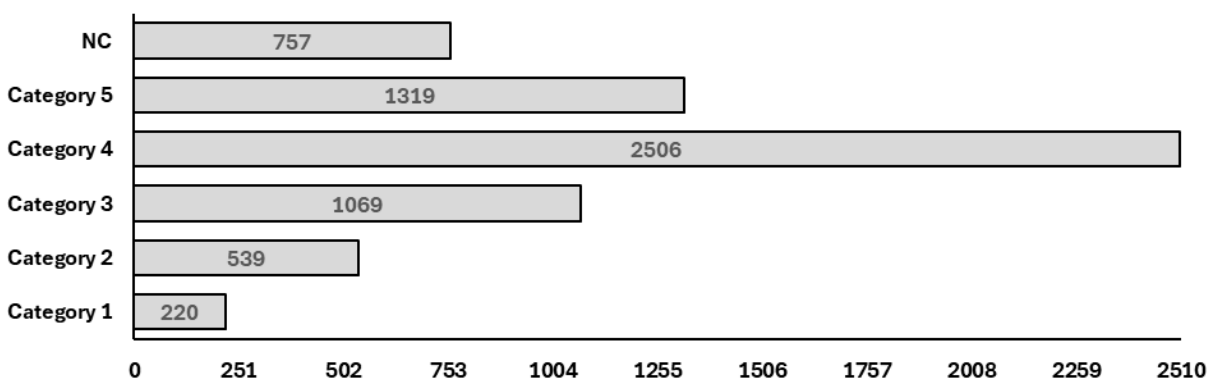
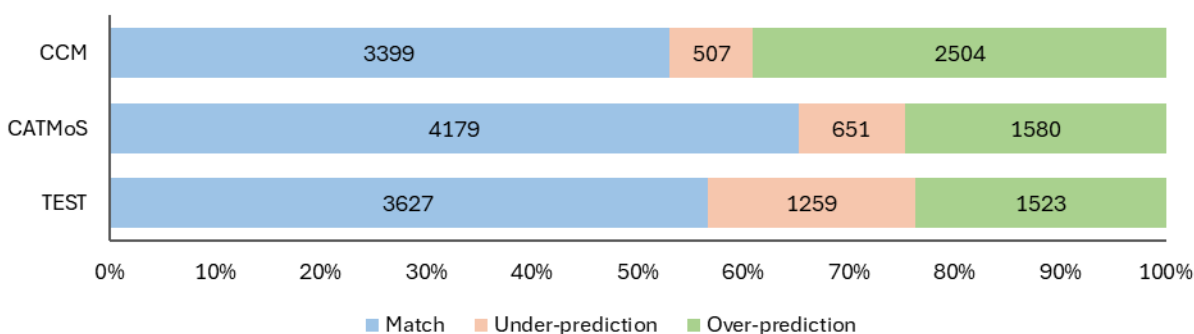


Figure 5.1. GHS categories for the 6,410 compounds classified based upon the experimental LD<sub>50</sub> data.

Three accuracy parameters were defined: match (denotes accurate prediction), under-prediction, and over-prediction. A summary of the overall prediction accuracy of the three models is shown in Figure 5.2a. About 57% of the compounds (3,627/6,410) predicted in TEST matched (i.e., were in agreement with) the experimental GSH categories, with the overall under-and over-prediction incidences in TEST being ~20% (1,259/6,410) and ~24% (1,529/6,410), respectively. As shown in Figure 5.2b, most of the matched and over-predictions were distributed within categories 3, 4, and 5 (i.e.,  $\geq 300$  LD<sub>50</sub>  $\leq 5000$  mg/kg), suggesting that TEST was mostly reliable or conservative in predicting within this range.

CATMoS exhibited about 65% (4,179/6,410) agreement with the experimental data, indicating a more accurate hazard category prediction rate than TEST (Figure 5.2a). The under- and over-prediction rates were about 10% and 25%, respectively, which indicates that CATMoS was more conservative than TEST. Similar to TEST, the majority of these accurate and over-prediction incidences occurred within GHS categories 3, 4, and 5 (i.e.,  $\geq 300 \text{ LD}_{50} \leq 5000 \text{ mg/kg}$ ) (Figure 5.2b), suggesting that this model was also mostly reliable or conservative in predicting GHS categories 3, 4, and 5.

**a**



b

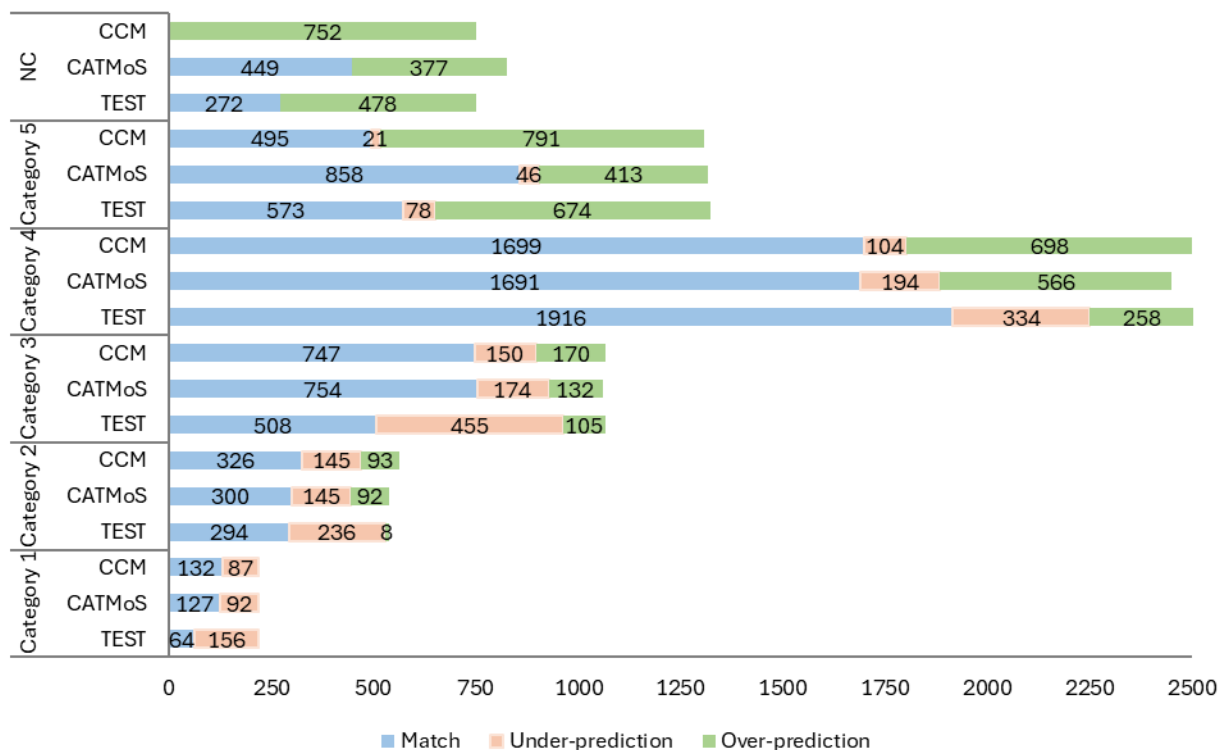


Figure 5.2. Evaluation of the model predictive accuracy for GHS classification of the 6,410 compounds. (a) The match, under-and over-predictions are for each model, as well as (b) how far off each prediction were from the experimental-based GHS category.

For CCM, the GHS classification of about 53% (3,399/6,410) of compounds were accurately predicted; this was lower than the accuracy rates recorded in TEST or CATMoS (Figure 5.2a). Graham et al. (2021) reported similar findings, where 89% (290/326) and 91% (320/353) accurate predictions in CATMoS and Leadscope, respectively, of pharmaceuticals were reduced to 77% (286/370) in a conservative consensus of the two model predictions. However, relative to TEST and CATMoS, the number of under-predictions in CCM was lowest at 8% (507/6,410), while the number of over-predictions was highest at 39% (2,504/6,410). These results indicate that, by design, CCM was the most conservative as it utilized the most conservative value from both models. As with TEST and CATMoS, most of the over-prediction incidences in CCM were within GSH categories 3, 4, and 5 (i.e.,  $\geq 300 \text{ LD}_{50} \leq 5000 \text{ mg/kg}$ ) (Figure 5.2b), which similarly suggested that CCM

was mostly conservative in predicting GHS categories 3, 4, and 5. Given this similarity, it was more informative to characterize the level of conservativeness of each model across all the GHS categories (see the discussion below in Section 5.3.2.2).

### **5.3.2.2 Level of conservativeness of the model predictions**

Maximizing the number of over-predictions and minimizing the number of under-predictions are important measures of model conservativeness (Hoffmann et al., 2010). Figure 5.2a shows that CCM resulted in about 1.6-fold more over-predictions (compared to TEST or CATMoS) and about 2.2-fold and 1.3-fold less under-predictions compared to TEST and CATMoS, respectively. To further understand the level of conservativeness of CCM predictions, the extent of coverage of its over-predictions within each GHS category was compared to those of TEST and CATMoS, as further discussed below.

Generally speaking, over-predictions in each model systematically increased from GHS category 2 to NC, while under-predictions systematically decreased from category 1 to NC (Figure 5.2b) (recognizing that Category 1 could not be overpredicted). Building upon these results, O assessed the distribution of each model's under-and over-predictions within GSH "toxic" ( $LD_{50} \leq 2000$  mg/kg; categories 1, 2, 3, and 4) and "non-toxic" ( $LD_{50} > 2000$  mg/kg; category 5 and NC) classes. This classification maybe used in regulatory contexts to communicate hazards associated with chemicals (Gonella Diaza et al., 2015). Similar to Alberga et al. (2019) and Bercu et al. (2021), the degree of conservativeness was interpreted to increase when a model's over-predictions were more frequently distributed towards the "non-toxic" class compared to the "toxic" class or when under-predictions were more frequently distributed towards "toxic" class compared to "non-toxic" class.

Table 5.3 shows the distributions of each model's over-predictions (numbers within the green-shaded cells) and under-predictions (numbers within the orange-shaded cells) – for example, 8/6410 compounds over-predicted in TEST were determined to fall under experimental category 1, whereas 108/6410 compounds

under-predicted in the same model were determined to fall under experimental category 2. The under- and over-predictions were demarcated within GHS “toxic” (outlined in the red rectangles) and “non-toxic” (outlined in the blue rectangles) classes.

Table 5.3. Confusion matrices showing the distribution of the predicted 6,410 compounds within GHS categories, organized by experimental-based GHS category.

		Experimental					
		1	2	3	4	5	NC
<b>TEST predicted</b>	<b>1</b>	64	108	28	16	4	0
	<b>2</b>	8	294	189	41	5	1
	<b>3</b>	2	103	508	426	23	6
	<b>4</b>	1	16	241	1916	309	26
	<b>5</b>	1	11	27	635	573	78
	<b>NC</b>	0	3	8	179	288	272
<b>CATMoS predicted</b>	<b>1</b>	127	74	17	1	0	0
	<b>2</b>	92	300	134	11	0	0
	<b>3</b>	2	130	754	161	12	1
	<b>4</b>	1	15	550	1691	185	9
	<b>5</b>	0	3	44	366	858	46
	<b>NC</b>	0	3	10	103	261	449
<b>CCM predicted</b>	<b>1</b>	132	77	9	1	0	0
	<b>2</b>	93	326	140	5	0	0
	<b>3</b>	3	167	747	146	4	0
	<b>4</b>	2	27	669	1699	104	0
	<b>5</b>	1	14	58	718	495	21
	<b>NC</b>	0	2	16	218	516	0

Within the “toxic” class, CCM consistently recorded the highest total number of over-predictions in each category: category 2 (TEST = 8, CATMoS = 92, CCM = 93), category 3 (TEST = 105, CATMoS = 132, CCM = 170), and category 4 (TEST = 258, CATMoS = 566, CCM = 699). A similar observation was made in the “non-toxic” categories – i.e., category 5 (TEST = 674, CATMoS = 413, CCM = 791) and NC (TEST = 478, CATMoS = 377, CCM = 752) (Table 5.3). In contrast, the number of compounds under-predicted in each model decreased with increasing GHS categories, with CCM generally recording the lowest total number of

under-predictions in each category: category 1 (TEST = 156, CATMoS = 92, CCM = 87), category 2 (TEST = 236, CATMoS = 145, CCM = 145), category 3 (TEST = 455, CATMoS = 174, CCM = 150), category 4 (TEST = 335, CATMoS = 194, CCM = 104), and category 5 (TEST = 78, CATMoS = 46, CCM = 21) (Table 5.3).

Overall, the distribution results in Table 5.3 indicate that CCM provided the highest number of over-predictions, which were more frequently distributed towards less toxic/non-toxic GSH categories, compared to TEST or CATMoS, and the fewest under-predictions, which were more frequently distributed towards more toxic/toxic GHS categories, compared to TEST or CATMoS. As such, it is reasonable to conclude that CCM had the highest level of conservativeness (Alberga et al., 2019; Bercu et al., 2021). Studies have reported that such an approach as CCM with such a relatively high capacity to classify toxic substances can be utilized to provide a more reliable means to establish safety recommendations about a chemical based on worst-case scenario considerations, prioritize compounds for further assessment, or gain assurance for non-toxic compounds before releasing them to the market (Bishop et al., 2024; EFSA, 2018; García-Jacas et al., 2019; Lunghini et al., 2019; Valsecchi et al., 2020; WHO/IPCS, 2018).

#### **5.4 Discussion and conclusion**

The aim of this study was to assess the performance of CCM against TEST and CATMoS for the prediction of conservative (health-protective) oral rat LD<sub>50</sub> by leveraging a large and diverse dataset (6,410 compounds). Existing QSAR studies on this endpoint (e.g., Firman et al. (2022) and Graham et al. (2021)) have mainly focused on model performance in general, based on statistical correlation analysis of experimental vs. predicted LD<sub>50</sub> or hazard classification sensitivity of specific models. The present study went beyond the areas addressed in these studies by primarily focusing on model conservative predictions – determining the extent to which consensus predictions of TEST and CATMoS can provide health-protective (conservative) predictions. To my knowledge, no study has applied TEST and CATMoS for this

purpose. Additionally, I used, arguably, the largest dataset (i.e., 6,410) for this exercise compared to, for example, the 371 compounds used by Graham et al. (2021) for generating consensus predictions. As further discussed below, the present study argues that the outcome of this study lays a strong foundation to contextualize the use of consensus modeling for deriving health-protective predictions under conditions of uncertainty, especially where experimental data are limited or lacking.

#### **5.4.1 Consideration of CCM under conditions of uncertainty**

Uncertainty in model predictions can be addressed through conservative approaches, e.g., by applying uncertainty factors or selecting conservative model estimates (Graham et al., 2021; EFSA, 2018; Wikoff et al., 2019). The current study proposes the need to err on the side of over-prediction that accounts for potential uncertainty arising from the individual model predictions. This proposal (established as CCM) combines the predictions from the individual models by comparing and combining the model predictions. Valsecchi et al. (2020) and Zhu et al. (2009) explain that such a method can mitigate the potential influence of outliers from individual model predictions. This mitigation potential of CCM could be particularly useful when applying TEST and CATMoS models for regulatory assessment of chemicals in the absence of experimental data, where it might become challenging to determine the degree to which any of these models can reliably predict oral rat LD<sub>50</sub> (Gonella Diaza et al., 2015; Moudgal et al., 2023; Nelms et al., 2020; Zhu et al., 2009). In other words, the proposed consensus approach has the potential to reduce or eliminate conflicting predicted LD<sub>50</sub> values between the models; consequently, addressing uncertainty due to discrepancies or inconsistency in a compound's LD<sub>50</sub> values, as earlier illustrated in Table 5.1.

As mentioned above, a key challenge in QSAR modeling is the lack of reproducibility of chemical-specific predictions between models. This is attributable to factors such as: different models using different training sets with varying degrees of experimental error, the use of different model parameters and algorithms, systematic biases within a model, or random errors embedded within a model's structure

(Graham et al., 2021; Gonella Diaza et al., 2015; Kolmar & Grulke, 2021; Mansouri et al., 2021; Moudgal et al., 2023; Nelms et al., 2020; US EPA, 2015; Zhu et al., 2009). TEST is built on the original LD<sub>50</sub> values (US EPA, 2015), whereas CATMoS is built on both original and GHS categorical LD<sub>50</sub> values (Mansouri et al., 2021). While this may not be the sole reason for the observed differences in the predictions between the two models (Figure 5.2(a and b)/Table 5.3), studies have shown that point and categorical data inputs may contribute to QSAR models yielding varying predictions (Karmaus et al., 2022; Kolmar & Grulke, 2021). This means that if conservative predictions from either of the models are used to set health-protective values, even if one model prediction is more accurate than the other, uncertainty may still exist to the extent that both models may produce conflicting predictions of the same chemical (Kirchner et al., 2021). Moreover, uncertainty might arise regarding potential disagreement about which model is more reliable to conduct a prediction (Kirchner et al., 2021). Drawing on the harmonization potential of CCM, it, therefore, becomes clear that this approach could be relied upon to reduce the effect of such uncertainties or minimize statistical type II errors (i.e., incorrectly concluding that a chemical has no health effects) in the face of the uncertainties (Chapman et al., 1998).

#### **5.4.2 Prioritizing chemicals based on CCM predictions**

While it remains for the end user of the model predictions to decide on the best practices to prioritize chemicals for further assessment or for safety-based decisions, from a health protection point of view, it is not uncommon to use GHS classifications as a factor to prioritize compounds for further assessment (Marty et al., 2022; Moudgal et al., 2023; United Nations, 2021). An example is the Canadian Identification of Risk Assessment Priorities (IRAP) scheme, which highlights the importance of using chemical hazard information and new scientific techniques (e.g., QSAR) to guide decisions regarding whether further (or no) risk assessment or more data generation is required for a chemical (Health Canada, 2017). The United States Toxic Substances Control Act (TSCA) also permits ranking substances to the extent that their health

hazards are aligned with GHS categories (US EPA, 2018). As illustrated with the examples of six compounds in Table 5.4, it is argued in this study that CCM can support these prioritization needs.

The compounds may be prioritized in the ranking order shown in Table 5.4, established based on CCM GHS categories. For example, 3-pentenenitrile might be considered the highest priority for assessment or for stringent safety measures and handling protocols due to its high potential to cause harm (the compound has the lowest LD<sub>50</sub> value of 2 (mg/kg) and falls under GHS category 1). In contrast, the potential of the non-toxic lactulose (LD<sub>50</sub> = 9511 (mg/kg); GHS category NC) to cause harm is very minimal, meaning that safety measures against it may be the most relaxed. The foundation of argument with this illustration is that CCM could have a significant value in supporting interim or internal decision-making in the initial stages of a compound development, where studies typically conduct rapid screening to detect potential toxic hazard properties/profiles that may render a compound unsuitable for development (Marty et al., 2022). In such scenarios, predicted oral rat LD<sub>50</sub> values are often considered non-conclusive; hence, they should be conservative in order to account for potential uncertainty in the prediction (Marty et al., 2022).

Table 5.4. Ranking of the six chemicals based on CCM GHS categories. The compounds are placed in an order: highly toxic (categories 1 and 2; orange-shaded), toxic (categories 3 and 4; blue-shaded), and non-toxic (category 5 and NC; grey-shaded).

Compound	CASRN	Model predictions						Ranking order
		TEST		CATMoS		CCM		
		LD <sub>50</sub> (mg/kg)	GHS category	LD <sub>50</sub> (mg/kg)	GHS category	LD <sub>50</sub> (mg/kg)	GHS category	
3-Pentenenitrile	4635-87-4	269	4	2	1	2	1	1
Spinosyn A	131929-60-7	34	2	2467	5	34	2	2

Tolclofos-methyl	57018-04-9	839	4	242	3	242	3	3
6-Methyluracil	626-48-2	908	4	1490	4	908	4	4
Heptanoic acid	111-14-8	2055	5	5827	NC	2055	5	5
Lactulose	4618-18-2	9511	NC	12095	NC	9511	NC	6

Overall, the results obtained from this study can bolster confidence that decisions based upon CCM predictions can be the most health-protective, especially in chemical screening-level assessments in the absence of experimental data. Alternatively, in demonstrating that CCM can improve the level of conservativeness in the predictions, this study argues that it can be adapted for use in regulatory contexts to contribute to a weight-of-evidence approach that justifies the need to prioritize particular chemicals for further risk assessment. Additionally, regulators can use the information from CCM to balance options or set precautionary measures aimed at reducing potential human health risks by restricting the use of particular chemicals.

## **Chapter 6: Conclusion**

This final chapter summarizes the research undertaken within Chapters 2-5 and situates the summary discussion within the context of the aim of the thesis and the broader field of *in silico* toxicology. The chapter then ends with suggestions for future work that could be built upon the research undertaken in this thesis.

### **6.1 Summary of the chapters and research contribution**

As discussed in Chapter 1, the research presented in this thesis has been conducted with an overarching aim of contributing to efforts aimed at analyzing and accounting for uncertainty in *in silico* methods for chemical toxicity predictions, particularly QSARs. As noted below, this aim has been achieved, in part at least, in Chapters 2-5 of this thesis.

Within the context of defining the scope and prediction needs in relation to relevant chemical assessment scenarios, Chapter 2 of this thesis was built on the fundamental understanding that problem formulation (PF) for *in silico* toxicology methods is the first step in determining whether a model and its predictions are scientifically valid, fit-for-purpose, and address chemical hazard or risk problems that are relevant to specific decision contexts. The developed PF framework describes an iterative process that systematically links five key stages: problem framing, problem exploration, research needs, conceptual model development, and hypothesis formulation. Through conceptualization of these stages as PF components, potential areas of uncertainty manifestation within each component are described. Specifically, uncertainty was described in relation to situations where: (1) particular components are missing, thus leading to a partial inclusion of components in a formulated problem, and/or (2) a component is included but only implicitly described in a formulated problem; thus, contributing to a general lack of specificity in its description in a given model prediction context. Following the consideration of uncertainty, this chapter ends with a proposed process that systematically outlines steps to characterize the uncertainty. Through

the discussion and characterization of uncertainty, Chapter 2 of this thesis seized the opportunity to rethink the concept of PF as a tool that not only, in a structured manner, outlines relevant components for *in silico* modeling, but also facilitates efforts aimed at accounting for potential uncertainty embedded with the components – this aligns with the overarching aim of this thesis, which, as discussed in Section 1.3, included accounting for uncertainty associated with *in silico* toxicology methods.

Chapter 3 brought to attention the framework proposed by Belfield et al. (2021), whose aim is to facilitate fit-for-purpose assessment of *in silico* models – this framework formed the foundation of the uncertainty categorization framework developed in this chapter. The framework developed in this chapter is to aid a structured means of identifying, categorizing, and describing diverse sources of uncertainty associated with *in silico* toxicology methods and their predictions. Two of the original modeling components in Belfield et al. (2021) (e.g., “Description”) were replaced with those (e.g., “Similarity”) deemed more relevant for identifying sources of uncertainty in *in silico* toxicology modeling. This modification allowed for a better description of sources of uncertainty within *in silico* toxicology modeling phases (i.e., model creation, characterization, and application). Following the synthesis of different perspectives on sources of uncertainty in the *in silico* toxicology method literature, a framework that outlines 10 modeling components and 20 general sources of uncertainty (GSU) was developed. As illustrated through the case study involving QSAR prediction of acute oral toxicity, the GSU from the developed framework proves to be valuable in guiding a practical identification of uncertainty associated with the applied QSAR and its predictions. The chapter concluded by contextualizing the relevance of the developed framework within the recent OECD’s (Q)SAR Assessment Framework (QAF) (OECD, 2023), with the underlying discussion on how the principles in QAF relate to the development of the framework, as well as how the framework extends the discussions initiated within QAF.

By completing the research in Chapters 2 and 3, this thesis advances research within *in silico* toxicology methods in different ways. First, as discussed in Chapter 1, it is argued that if *in silico* methods (e.g., QSARs) and their predictions are to gain regulatory acceptance, they need to be understandable to regulatory assessors and described in a transparent manner to form the basis for regulatory decision-making or facilitate their fit-for-purpose evaluation (Debad et al., 2024). Systematic identification and categorization of uncertainty associated with the models and their predictions can contribute to the evaluation efforts and, ultimately, improve regulatory uptake of the methods (Ball et al., 2014; Belfield et al., 2021; Benford et al., 2018; Blackburn and Stuard, 2014; Cronin et al., 2019; Patlewicz et al., 2013; Pestana et al., 2021; Pham et al., 2019; Schultz et al., 2015; Worth et al., 2011a). However, as mentioned within Chapters 2 and 3, presently, it is not clear what components should be considered in a model and modeling context and what kind of information is pertinent for uncertainty discussion under the components – i.e., in the case of problem formulation and modeling phases. Given the highlighted research gaps, I see the contribution of this thesis through the use of the framework developed in Chapter 2 as a valuable reference tool for identifying components that are pertinent for formulating *in silico* toxicology problems. Alternatively, the framework in Chapter 3 can be used by model developers and users to guide their efforts in mapping out and prioritizing areas of uncertainty, especially during uncertainty analysis, data gap-filling exercises, or when interpreting prediction results. In other words, these frameworks can foster a structured understanding of which components are central to a particular method, shed light on how uncertainty can manifest within the components, as well as help potentially increase transparency and trust in a model or modeling exercise, especially with regard to communicating uncertainties between modelers and relevant stakeholders – this is in line with the working principles of OECD (2007, 2023) and WHO/IPCS (2018), where transparency and trust are key to regulatory acceptance of models and their predictions.

The study in Chapter 4 sought to broaden the understanding of uncertainties that are important yet often implicitly expressed by modelers. As such, the focus in this chapter went beyond the general consideration of the importance of communicating uncertainty in models and predictions (as discussed in Chapters 2 and 3) to emphasize the need to broaden decision frames through consideration of implicit uncertainty. The lack of adequate considerations of implicit uncertainty has already been identified by EFSA (2006) and other scholars (Flari and Wilkinson, 2011; Levin et al., 2004; Sahlin et al., 2011) as a research gap, with these scholars calling for systematic ways of identifying and categorizing the uncertainty. Chapter 4, therefore, responded to this call and established a way to identify and categorize the uncertainty in QSAR modeling of chemical toxicity. Specifically, a method that combines the use of uncertainty indicators and the framework in Chapter 3 was developed in this chapter to support the analysis of implicit and explicit uncertainties – this directly contributes to the overarching aim of this thesis, as discussed in Section 1.3.

The analyses in Chapter 4 revealed that uncertainty was majorly expressed implicitly in the studies, with Model performance, Mechanistic plausibility, Model relevance, and Descriptor relevance being the most common sources of uncertainty. The chapter ends with a general appraisal of the implication of the proposed method for uncertainty analysis in QSAR modeling. In this appraisal, three systematic steps were emphasized, with the overarching aim of improving rigor and transparency in accounting for uncertainty in QSAR modeling processes: (1) identification of uncertainty indicators, (2) identification of uncertainty within the context of the indicators, and (3) categorization of the identified uncertainties using the framework in Chapter 3. In so doing, therefore, I anticipate that this study will inform decision-support initiatives by modelers and regulatory authorities about the need to communicate uncertainty explicitly, as well as support further research into the effects of implicit uncertainty within QSAR modeling and the development of additional methods to quantify implicit uncertainty.

Finally, Chapter 5 set out to address uncertainty in QSAR predictions through a consensus conservative approach (here referred to as “conservative consensus model” – CCM) that combines oral acute toxicity (i.e., LD<sub>50</sub>) predictions of TEST and CATMoS. The adoption of the conservative approach was based on the premise that, under conditions of uncertainty (e.g., due to conflicting chemical-specific LD<sub>50</sub> predictions from TEST and CATMoS) or in scenarios where experimental data are lacking or limited, it becomes necessary to apply the precautionary measures to account for the uncertainty to ensure that model estimates are health-protective. The study results showed that CCM considerably increased the number of over-predictions to 39%, from 24% in TEST and 25% in CATMoS and reduced under-predictions to 8%, from 20% in TEST and 10% in CATMoS. From a safety standpoint, these results bolster confidence that decisions based upon CCM predictions would be more health-protective (compared to TEST or CATMoS individually). According to EFSA, in the context of chemical hazard assessment, such conservative predictions are necessary in the general sense of dealing with uncertainty that might influence the setting of protection goals (Benford et al., 2018). Additionally, by mitigating contradictory predictions from the individual models, I believe that the approach is suited for application to large chemical inventories (e.g., within Canada’s Domestic Substance List) to improve the prediction reliability of data-poor chemicals. Furthermore, as illustrated in Table 5.4, the approach can be integrated into larger frameworks aimed at prioritizing chemicals for further assessment and ultimately contribute to reducing or eliminating animal testing for chemical hazard assessment.

## **6.2 Future work**

To advance the acceptance and use of *in silico* toxicology methods for regulatory purposes, this thesis aimed to address some of the widely recognized challenges in the field. While different strategies have been developed to enhance a transparent understanding of the methods and their predictions, there is

still work to be done. This section, informed by the findings and the highlighted limitations from the thesis, outlines some of the areas for future research focus.

*Harmonization of the frameworks.* The developed frameworks in Chapters 2 and 3 are expected to contribute to efforts that aim to standardize components for *in silico* problem formulation and modeling processes and provide a better understanding of whether biological and chemical-specific knowledge is available, or knowledge/data gaps exist in specific modeling contexts. As discussed in Chapter 1, problem formulation should be linked to hazard and risk assessment processes (expressed through the *in silico* modeling phases in Chapter 3). This suggests that, ideally, the frameworks in Chapters 2 and 3 could be linked/combined to provide a broader picture of the interlinkage between problem formulation and the modeling phases. In other words, it might be necessary to combine the frameworks in Chapters 2 and 3 into one overarching framework that makes it possible to simultaneously, in a harmonized manner, answer questions related to *in silico* problem formulation and modeling components, including potential knowledge gaps within them. Ideally, such an overarching framework would encapsulate the entirety of the central issues outlined in the two frameworks, thus ensuring a more comprehensive evaluation of the fitness-for-purpose of models or adequacy of model predictions within a defined decision context. Ultimately, such a harmonized framework may foster the production of clearly defined models that satisfy regulatory needs for case studies.

*Incorporation of information from other sources.* The field of *in silico* toxicology methods as an evolving research field with new modeling techniques being developed to improve model prediction performance. Additionally, the uncertainty characterization within the field is expected to evolve as new experiences emerge. This means that the frameworks presented in this thesis should be only considered as the initial point for discussions about sources of uncertainty as well as *in silico* modeling components. In the meantime, a variety of avenues for refining both the frameworks in Chapters 2 and 3 and the method in

Chapter 4 also exist. For example, the frameworks and the method can be refined by using additional relevant information from other methods, such as *in vitro* or *in vivo* toxicokinetic, toxicodynamic, and adverse outcome pathway information and data. Such an opportunity to use non-*in silico* information will inform the use of the weight of the evidence strategies to improve *in silico* problem formulation, chemicals hazard and risk characterization, and uncertainty assessment.

*Creation of computational technique to automate identification uncertainty indicators.* As described in Chapter 4, this thesis employed a manual approach to text analysis to identify uncertainty indicators and uncertainty expressed within the uncertainty indicator-containing sentences. In cases where a large amount of text or a relatively high number of studies are to be analyzed, this manual approach can be limiting. Computerized approaches with abilities to identify and extract specific information from texts based on natural language processing, machine learning, and text mining methods can prove to be more efficient in identifying at the heart of these new efforts to improve the identification of a large number of indicators in a short time. Such tools can help analysts quickly navigate large amounts of textual data to locate uncertain information of interest. The development of such tools might require integrating different disciplinary knowledge – for example, knowledge of key areas of uncertainty in *in silico* modeling processes, linguistic knowledge to help qualify confidence in statements, and machine learning techniques to automate the identification of indicators in texts. The method described in Chapter 4 of this thesis will provide the foundation (thus fundamental to the success) of this initiative.

*Characterization of the level of uncertainty.* Studies in the *in silico* toxicology (e.g. QSAR) literature have discussed the need to characterize (e.g., via semi-quantitative scales – “low”, “moderate”, and “high”) levels of uncertainty associated with different uncertainty sources as a means of informing judgements about whether the uncertainty is acceptable (Belfield et al., 2021; Cronin et al., 2019). Ball et al. (2014) also hold that the characterization of levels of uncertainty is important when evaluating whether

conservative models are reliable enough to produce acceptable levels of uncertainty. This thesis (e.g., in Chapter 2; Figure 2.4 and Chapter 4; Section 4.4.3) discussed the need to characterize acceptable levels of uncertainty in relation to the sources of uncertainty within the frameworks in Chapters 2 and 3; however, it did not provide a practical way to characterize of levels of potential uncertainty within the sources. As such, future studies could advance this work by, for example, characterizing whether the level of uncertainty resulting from the application of the conservative approach in Chapter 5 is acceptable in a health-protective decision-making context.

### **6.3 Final reflection**

Research undertaken in this thesis aimed to address one of the key challenges hindering the acceptance of *in silico* toxicology methods in regulatory settings – uncertainty. Throughout the chapters, four overarching questions were explored: What components constitute *in silico* modeling problem formulation and phases? Which areas within the components might uncertainty reside? How is uncertainty expressed within QSAR studies predicting chemical toxicity? To what extent can a conservative consensus approach address uncertainty in QSAR predicting acute chemical toxicity? It may be tempting to propose simple answers to these questions (e.g., using uncertainties associated with one method as representative of those in other methods); however, this thesis demonstrates that complexities associated with the questions require broader considerations. For example, addressing the issue of uncertainty requires consideration of multiple modeling components, from model input to output application, as well as unique characteristics of different *in silico* methods. Such a consideration is critical for promoting a more comprehensive and transparent communication of uncertainty as well as standardization in how different uncertainty sources are described across methods. The thesis argues that these efforts could contribute to efforts aimed at increasing acceptance of the methods and their predictions for regulatory applications, grounded within specific endpoints being evaluated and the contexts in which regulatory decisions are made.

## References

- Achar, J., Firman, J. W., Cronin, M. T. D., & Öberg, G. (2024a). A framework for categorizing sources of uncertainty in *in silico* toxicology methods: Considerations for chemical toxicity predictions. *Regulatory Toxicology and Pharmacology*, *154*, 105737. <https://doi.org/10.1016/j.yrtph.2024.105737>
- Achar, J. C., Kim, D. Y., Kwon, J.-H., & Jung, J. (2020b). Toxicokinetic modeling of octylphenol bioconcentration in *Chlorella vulgaris* and its trophic transfer to *Daphnia magna*. *Ecotoxicology and Environmental Safety*, *194*, 110379. <https://doi.org/10.1016/j.ecoenv.2020.110379>
- Achar, J. C., Nam, G., Jung, J., Klammler, H., & Mohamed, M. M. (2020c). Microbubble ozonation of the antioxidant butylated hydroxytoluene: Degradation kinetics and toxicity reduction. *Environmental Research*, *186*, 109496. <https://doi.org/10.1016/j.envres.2020.109496>
- Achar, J., Cronin, M. T. D., Firman, J. W., & Öberg, G. (2024d). A problem formulation framework for the application of *in silico* toxicology methods in chemical risk assessment. *Archives of Toxicology*. <https://doi.org/10.1007/s00204-024-03721-6>
- Akhtar, A. (2015). The flaws and human harms of animal experimentation. *Cambridge Quarterly of Healthcare Ethics*, *24*(4), 407–419. <https://doi.org/10.1017/S0963180115000079>
- 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. *Regulatory Toxicology and Pharmacology*, *129*, 105094. <https://doi.org/10.1016/j.yrtph.2021.105094>
- Ball, N., Bars, R., Botham, P. A., Cuciureanu, A., Cronin, M. T. D., Doe, J. E., Dudzina, T., Gant, T. W., Leist, M., & van Ravenzwaay, B. (2022). A framework for chemical safety assessment incorporating new approach methodologies within REACH. *Archives of Toxicology*, *96*(3), 743–766. <https://doi.org/10.1007/s00204-021-03215-9>
- Ball, N., Bartels, M., Budinsky, R., Klapacz, J., Hays, S., Kirman, C., & Patlewicz, G. (2014). The challenge of using read-across within the EU REACH regulatory framework; how much uncertainty is too much? Dipropylene glycol methyl ether acetate, an exemplary case study. *Regulatory Toxicology and Pharmacology*, *68*(2), 212–221. <https://doi.org/10.1016/j.yrtph.2013.12.007>
- Bal-Price, A., Pistollato, F., Sachana, M., Bopp, S. K., Munn, S., & Worth, A. (2018). Strategies to improve the regulatory assessment of developmental neurotoxicity (DNT) using *in vitro* methods. *Toxicology and Applied Pharmacology*, *354*, 7–18. <https://doi.org/10.1016/j.taap.2018.02.008>

- Baltazar, M. T., Cable, S., Carmichael, P. L., Cubberley, R., Cull, T., Delagrangé, M., Dent, M. P., Hatherell, S., Houghton, J., Kukic, P., Li, H., Lee, M.-Y., Malcomber, S., Middleton, A. M., Moxon, T. E., Nathanail, A. V., Nicol, B., Pendlington, R., Reynolds, G., ... Westmoreland, C. (2020). A next-generation risk assessment case study for coumarin in cosmetic products. *Toxicological Sciences*, *176*(1), 236–252. <https://doi.org/10.1093/toxsci/kfaa048>
- Barber, C., Heghes, C., & Johnston, L. (2024). A framework to support the application of the OECD guidance documents on (Q)SAR model validation and prediction assessment for regulatory decisions. *Computational Toxicology*, *30*, 100305. <https://doi.org/10.1016/j.comtox.2024.100305>
- Belfield, S. J., Cronin, M. T. D., Enoch, S. J., & Firman, J. W. (2023). Guidance for good practice in the application of machine learning in development of toxicological quantitative structure-activity relationships (QSARs). *PLOS ONE*, *18*(5), e0282924. <https://doi.org/10.1371/journal.pone.0282924>
- 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. *Regulatory Toxicology and Pharmacology*, *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. *Environment International*, *131*, 105060. <https://doi.org/10.1016/j.envint.2019.105060>
- 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., Goetz, N. V., Koutsoumanis, K., ... Hardy, A. (2018). Guidance on Uncertainty Analysis in Scientific Assessments. *EFSA Journal*, *16*(1), e05123. <https://doi.org/10.2903/j.efsa.2018.5123>
- Bishop, P. L., Mansouri, K., Eckel, W. P., Lowit, M. B., Allen, D., Blankinship, A., Lowit, A. B., Harwood, D. E., Johnson, T., & Kleinstreuer, N. C. (2024). Evaluation of *in silico* model predictions for mammalian acute oral toxicity and regulatory application in pesticide hazard and risk assessment. *Regulatory Toxicology and Pharmacology*, *149*, 105614. <https://doi.org/10.1016/j.yrtph.2024.105614>
- Blackburn, K., & Stuard, S. B. (2014). A framework to facilitate consistent characterization of read across uncertainty. *Regulatory Toxicology and Pharmacology*, *68*(3), 353–362. <https://doi.org/10.1016/j.yrtph.2014.01.004>

- Browne, P. (2023). *Introduction to QSAR Assessment Framework*.  
<https://ascct.memberclicks.net/assets/WebinarSlides/2023.10.30%20Browne.pdf>
- 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. *Regulatory Toxicology and Pharmacology: RTP*, *80*, 241–246. <https://doi.org/10.1016/j.yrtph.2016.05.032>
- Callahan, M. A., & Sexton, K. (2007). If cumulative risk assessment is the answer, what is the question? *Environmental Health Perspectives*, *115*(5), 799–806. <https://doi.org/10.1289/ehp.9330>
- Chandrasekaran, B., Abed, S. N., Al-Attraqchi, O., Kuche, K., & Tekade, R. K. (2018). Chapter 21—Computer-Aided Prediction of Pharmacokinetic (ADMET) Properties. In R. K. Tekade (Ed.), *Dosage Form Design Parameters* (pp. 731–755). Academic Press. <https://doi.org/10.1016/B978-0-12-814421-3.00021-X>
- Chapman, P. M., Fairbrother, A., & Brown, D. (1998). A critical evaluation of safety (uncertainty) factors for ecological risk assessment. *Environmental Toxicology and Chemistry*, *17*(1), 99–108. <https://doi.org/10.1002/etc.5620170112>
- Cohen, S. M. (2017). The relevance of experimental carcinogenicity studies to human safety. *Current Opinion in Toxicology*, *3*, 6–11. <https://doi.org/10.1016/j.cotox.2017.04.002>
- Cousins, I. T., Vestergren, R., Wang, Z., Scheringer, M., & McLachlan, M. S. (2016). The precautionary principle and chemicals management: The example of perfluoroalkyl acids in groundwater. *Environment International*, *94*, 331–340. <https://doi.org/10.1016/j.envint.2016.04.044>
- Crofton, K. M., Bassan, A., Behl, M., Chushak, Y. G., Fritsche, E., Gearhart, J. M., Marty, M. S., Mumtaz, M., Pavan, M., Ruiz, P., Sachana, M., Selvam, R., Shafer, T. J., Stavitskaya, L., Szabo, D. T., Szabo, S. T., Tice, R. R., Wilson, D., Woolley, D., & Myatt, G. J. (2022). Current status and future directions for a neurotoxicity hazard assessment framework that integrates in silico approaches. *Computational Toxicology (Amsterdam, Netherlands)*, *22*, 100223. <https://doi.org/10.1016/j.comtox.2022.100223>
- Cronin, M., & Madden, J. (2010). *In Silico Toxicology: Principles and Applications*. Royal Society of Chemistry. <https://books.rsc.org/books/edited-volume/1314/In-Silico-Toxicology>
- Cronin, M. T. D. (1996). Quantitative structure-Activity relationship (QSAR) analysis of the acute sublethal neurotoxicity of solvents. *Toxicology in Vitro*, *10*(2), 103–110. [https://doi.org/10.1016/0887-2333\(95\)00109-3](https://doi.org/10.1016/0887-2333(95)00109-3)
- 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. *Regulatory Toxicology and Pharmacology*, 135, 105249. <https://doi.org/10.1016/j.yrtph.2022.105249>
- Cronin, M. T. D., & Livingstone, DJ. (2004). *Predicting chemical toxicity and fate*. CRC Press. <https://doi.org/10.1201/9780203642627>
- Cronin, M. T. D., Madden, J. C., Yang, C., & Worth, A. P. (2019). Unlocking the potential of in silico chemical safety assessment – A report on a cross-sector symposium on current opportunities and future challenges. *Computational Toxicology*, 10, 38–43. <https://doi.org/10.1016/j.comtox.2018.12.006>
- 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. *Regulatory Toxicology and Pharmacology*, 106, 90–104. <https://doi.org/10.1016/j.yrtph.2019.04.007>
- Debad, S., Allen, D., Bandele, O., Bishop, C., Blaylock, M., Brown, P., Bungler, M. K., Co, J. Y., Crosby, L., Daniel, A. B., Ferguson, S. S., Ford, K., Costa, G. G. da, Gilchrist, K. H., Grogg, M. W., Gwinn, M., Hartung, T., Hogan, S. P., Jeong, Y. E., ... Fitzpatrick, S. (2024). Trust your gut: Establishing confidence in gastrointestinal models - An overview of the state of the science and contexts of use. *ALTEX - Alternatives to Animal Experimentation*. <https://doi.org/10.14573/altex.2403261>
- Dent, M., Amaral, R. T., Da Silva, P. A., Ansell, J., Boisleve, F., Hatao, M., Hirose, A., Kasai, Y., Kern, P., Kreiling, R., Milstein, S., Montemayor, B., Oliveira, J., Richarz, A., Taalman, R., Vaillancourt, E., Verma, R., Posada, N. V. O. C., Weiss, C., & Kojima, H. (2018). Principles underpinning the use of new methodologies in the risk assessment of cosmetic ingredients. *Computational Toxicology*, 7, 20–26. <https://doi.org/10.1016/j.comtox.2018.06.001>
- Devos, Y., Craig, W., Devlin, R. H., Ippolito, A., Leggatt, R. A., Romeis, J., Shaw, R., Svendsen, C., & Topping, C. J. (2019). Using problem formulation for fit-for-purpose pre-market environmental risk assessments of regulated stressors. *EFSA Journal*, 17(S1), e170708. <https://doi.org/10.2903/j.efsa.2019.e170708>
- Dhami, M. K., & Mandel, D. R. (2022). Communicating uncertainty using words and numbers. *Trends in Cognitive Sciences*, 26(6), 514–526. <https://doi.org/10.1016/j.tics.2022.03.002>
- Donfrancesco, V., Allen, B. L., Appleby, R., Behrendorff, L., Conroy, G., Crowther, M. S., Dickman, C. R., Doherty, T., Fancourt, B. A., Gordon, C. E., Jackson, S. M., Johnson, C. N., Kennedy, M. S.,

- Koungoulos, L., Letnic, M., Leung, L. K.-P., Mitchell, K. J., Nesbitt, B., Newsome, T., ... Cairns, K. M. (2023). Understanding conflict among experts working on controversial species: A case study on the Australian dingo. *Conservation Science and Practice*, 5(3), e12900. <https://doi.org/10.1111/csp2.12900>
- 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. *Frontiers in Toxicology*, 5. <https://doi.org/10.3389/ftox.2023.1155645>
- ECHA. (2008). *Guidance on Information Requirements and Chemical safety assessment*. Retrieved July 7, 2023, from <https://echa.europa.eu/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment>
- ECHA. (2012). *Guidance on information requirements and chemical safety assessment. Chapter R.19: Uncertainty analysis*. Retrieved May 6, 2022, from [https://echa.europa.eu/documents/10162/17224/information\\_requirements\\_r19\\_en.pdf/d5bd6c3f-3383-49df-894e-dea410ba4335?t=1353935215756](https://echa.europa.eu/documents/10162/17224/information_requirements_r19_en.pdf/d5bd6c3f-3383-49df-894e-dea410ba4335?t=1353935215756)
- EFSA. (2006). Opinion of the scientific committee related to uncertainties in dietary exposure assessment. *EFSA Journal, EFSA Journal*. <https://doi.org/10.2903/j.efsa.2007.438>
- EFSA. (2010). *Applicability of QSAR analysis to the evaluation of the toxicological relevance of metabolites and degradates of pesticide active substances for dietary risk assessment | EFSA*. Retrieved May 7, 2024, from <https://www.efsa.europa.eu/en/supporting/pub/en-50>
- EFSA, 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., ... Hardy, A. (2018). Guidance on Uncertainty Analysis in Scientific Assessments. *EFSA Journal*, 16(1), e05123. <https://doi.org/10.2903/j.efsa.2018.5123>
- EFSA, 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., ... Hardy, A. (2018). The principles and methods behind EFSA's Guidance on Uncertainty Analysis in Scientific Assessment. *EFSA Journal*, 16(1), e05122. <https://doi.org/10.2903/j.efsa.2018.5122>
- 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., ... Younes, M. (2017). Guidance on the use of the weight of evidence

- approach in scientific assessments. *EFSA Journal*, 15(8), e04971. <https://doi.org/10.2903/j.efsa.2017.4971>
- EFSA, Hernández-Jerez, A., Adriaanse, P., Aldrich, A., Berny, P., Coja, T., Duquesne, S., Focks, A., Marinovich, M., Millet, M., Pelkonen, O., Pieper, S., Tiktak, A., Topping, C., Widenfalk, A., Wilks, M., Wolterink, G., Crofton, K., Hougaard Bennekou, S., ... Tzoulaki, I. (2021). Development of Integrated Approaches to Testing and Assessment (IATA) case studies on developmental neurotoxicity (DNT) risk assessment. *EFSA Journal*, 19(6), e06599. <https://doi.org/10.2903/j.efsa.2021.6599>
- EFSA Scientific Committee, 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., ... Hardy, A. (2018). Guidance on Uncertainty Analysis in Scientific Assessments. *EFSA Journal*, 16(1). <https://doi.org/10.2903/j.efsa.2018.5123>
- EFSA Scientific Committee, More, S. J., Bampidis, V., Benford, D., Bennekou, S. H., Bragard, C., Halldorsson, T. I., Hernández-Jerez, A. F., Koutsoumanis, K., Naegeli, H., Schlatter, J. R., Silano, V., Nielsen, S. S., Schrenk, D., Turck, D., Younes, M., Benfenati, E., Castle, L., Cedergreen, N., ... Hogstrand, C. (2019). Guidance on harmonised methodologies for human health, animal health and ecological risk assessment of combined exposure to multiple chemicals. *EFSA Journal*, 17(3). <https://doi.org/10.2903/j.efsa.2019.5634>
- Egeghy, P. P., Judson, R., Gangwal, S., Mosher, S., Smith, D., Vail, J., & Cohen Hubal, E. A. (2012). The exposure data landscape for manufactured chemicals. *Science of The Total Environment*, 414, 159–166. <https://doi.org/10.1016/j.scitotenv.2011.10.046>
- Ekins, S., Mestres, J., & Testa, B. (2007). In silico pharmacology for drug discovery: Methods for virtual ligand screening and profiling. *British Journal of Pharmacology*, 152(1), 9–20. <https://doi.org/10.1038/sj.bjp.0707305>
- El-Ghonemy, H., Watts, L., & Fowler, L. (2005). Treatment of uncertainty and developing conceptual models for environmental risk assessments and radioactive waste disposal safety cases. *Environment International*, 31(1), 89–97. <https://doi.org/10.1016/j.envint.2004.07.002>
- Embry, M. R., Bachman, A. N., Bell, D. R., Boobis, A. R., Cohen, S. M., Dellarco, M., Dewhurst, I. C., Doerrer, N. G., Hines, R. N., Moretto, A., Pastoor, T. P., Phillips, R. D., Rowlands, J. C., Tanir, J. Y., Wolf, D. C., & Doe, J. E. (2014). Risk assessment in the 21st century: Roadmap and matrix. *Critical Reviews in Toxicology*, 44(sup3), 6–16. <https://doi.org/10.3109/10408444.2014.931924>

- Enoch, S. j. (2010). Chemical category formation and read-across for the prediction of toxicity. In T. Puzyn, J. Leszczynski, & M. T. Cronin (Eds.), *Recent Advances in QSAR Studies: Methods and Applications* (pp. 209–219). Springer Netherlands. [https://doi.org/10.1007/978-1-4020-9783-6\\_7](https://doi.org/10.1007/978-1-4020-9783-6_7)
- Escher, S. E., Kamp, H., Bennekou, S. H., Bitsch, A., Fisher, C., Graepel, R., Hengstler, J. G., Herzler, M., Knight, D., Leist, M., Norinder, U., Ouédraogo, G., Pastor, M., Stuard, S., White, A., Zdrazil, B., van de Water, B., & Kroese, D. (2019). Towards grouping concepts based on new approach methodologies in chemical hazard assessment: The read-across approach of the EU-ToxRisk project. *Archives of Toxicology*, *93*(12), 3643–3667. <https://doi.org/10.1007/s00204-019-02591-7>
- European Commission. (2016). *On the development, validation and legal acceptance of methods alternative to animal testing in the field of cosmetics (2013-2015)*. Retrieved June 4, 2021, from <http://eur-lex.europa.eu/legal-content/en/txt/pdf/?uri=celex:52016dc0599&from=en>
- European Union. (2010). *The European Union directive 2010/63/EU on the protection of animals used for scientific purposes*. Retrieved on June 25, 2022 from <https://eur-lex.europa.eu/eli/dir/2010/63/oj>
- Felter, S. P., Bhat, V. S., Botham, P. A., Bussard, D. A., Casey, W., Hayes, A. W., Hilton, G. M., Magurany, K. A., Sauer, U. G., & Ohanian, E. V. (2021). Assessing chemical carcinogenicity: Hazard identification, classification, and risk assessment. Insight from a Toxicology Forum state-of-the-science workshop. *Critical Reviews in Toxicology*, *51*(8), 653–694. <https://doi.org/10.1080/10408444.2021.2003295>
- Ferson, S., O’Rawe, J., Antonenko, A., Siegrist, J., Mickley, J., Luhmann, C. C., Sentz, K., & Finkel, A. M. (2015). Natural language of uncertainty: Numeric hedge words. *International Journal of Approximate Reasoning*, *57*, 19–39. <https://doi.org/10.1016/j.ijar.2014.11.003>
- 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. *Archives of Toxicology*, *96*(3), 817–830. <https://doi.org/10.1007/s00204-021-03205-x>
- Flari, V., & Wilkinson, D. (2011). Terminology in risk assessments used by the scientific panels and scientific committee of EFSA. *EFSA Supporting Publications*, *8*(1). <https://doi.org/10.2903/sp.efsa.2011.EN-101>
- Fritsche, E., Grandjean, P., Crofton, K. M., Aschner, M., Goldberg, A., Heinonen, T., Hessel, E. V. S., Hogberg, H. T., Bennekou, S. H., Lein, P. J., Leist, M., Mundy, W. R., Paparella, M., Piersma, A. H., Sachana, M., Schmuck, G., Solecki, R., Terron, A., Monnet-Tschudi, F., ... Bal-Price, A. (2018). Consensus

- statement on the need for innovation, transition and implementation of developmental neurotoxicity (DNT) testing for regulatory purposes. *Toxicology and Applied Pharmacology*, 354, 3–6. <https://doi.org/10.1016/j.taap.2018.02.004>
- Fu, X., Wojak, A., Neagu, D., Ridley, M., & Travis, K. (2011). Data governance in predictive toxicology: A review. *Journal of Cheminformatics*, 3(1), 24. <https://doi.org/10.1186/1758-2946-3-24>
- Funtowicz, S. O., & Ravetz, J. R. (1993). Science for the post-normal age. *Futures*, 25(7), 739–755. [https://doi.org/10.1016/0016-3287\(93\)90022-L](https://doi.org/10.1016/0016-3287(93)90022-L)
- Gajewicz, A., Schaeublin, N., Rasulev, B., Hussain, S., Leszczynska, D., Puzyn, T., & Leszczynski, J. (2015). Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies. *Nanotoxicology*, 9(3), 313–325. <https://doi.org/10.3109/17435390.2014.930195>
- García-Jacas, C. R., Marrero-Ponce, Y., Cortés-Guzmán, F., Suárez-Lezcano, J., Martínez-Rios, F. O., García-González, L. A., Pupo-Meriño, M., & Martínez-Mayorga, K. (2019). Enhancing Acute Oral Toxicity Predictions by using Consensus Modeling and Algebraic Form-Based 0D-to-2D Molecular Encodes. *Chemical Research in Toxicology*, 32(6), 1178–1192. <https://doi.org/10.1021/acs.chemrestox.9b00011>
- Gautier, F., Tourneix, F., Assaf Vandecasteele, H., van Vliet, E., Bury, D., & Alépée, N. (2020). Read-across can increase confidence in the Next Generation Risk Assessment for skin sensitisation: A case study with resorcinol. *Regulatory Toxicology and Pharmacology*, 117, 104755. <https://doi.org/10.1016/j.yrtph.2020.104755>
- 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>
- Gonella Diaza, R., Manganelli, S., Esposito, A., Roncaglioni, A., Manganaro, A., & Benfenati, E. (2015). Comparison of in silico tools for evaluating rat oral acute toxicity. *SAR and QSAR in Environmental Research*, 26(1), 1–27. <https://doi.org/10.1080/1062936X.2014.977819>
- Government of Canada. (2023). *Committee Report No. 7—ENVI (44-1)—House of Commons of Canada*. Retrieved August 1, 2024, from <https://www.ourcommons.ca/DocumentViewer/en/44-1/ENVI/report-7>
- 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.

- Regulatory Toxicology and Pharmacology*, 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. *Regulatory Toxicology and Pharmacology*, 129, 105109. <https://doi.org/10.1016/j.yrtph.2021.105109>
- Hackam, D. G., & Redelmeier, D. A. (2006). Translation of Research Evidence From Animals to Humans. *JAMA*, 296(14), 1727–1732. <https://doi.org/10.1001/jama.296.14.1731>
- Haibo He, & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- Han, P. K. J., Klein, W. M. P., & Arora, N. K. (2011). Varieties of uncertainty in health care: A conceptual taxonomy. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, 31(6), 828–838. <https://doi.org/10.1177/0272989x11393976>
- Hansen, S. F., Carlsen, L., & Tickner, J. A. (2007). Chemicals regulation and precaution: Does REACH really incorporate the precautionary principle. *Environmental Science & Policy*, 10(5), 395–404. <https://doi.org/10.1016/j.envsci.2007.01.001>
- Hartung, T., Hoffmann, S., & Stephens, M. (2013). Food for Thought ... Mechanistic Validation. *ALTEX*, 30(2), 119–130.
- Health Canada. (2000). *Health Canada Decision-Making Framework for Identifying, Assessing, and Managing Health Risks—August 1, 2000* [Transparency - other]. Retrieved August 1, 2024, from <https://www.canada.ca/en/health-canada/corporate/about-health-canada/reports-publications/health-products-food-branch/health-canada-decision-making-framework-identifying-assessing-managing-health-risks.html>
- Health Canada. (2017). *The identification of risk assessment priorities* [Education and awareness]. Retrieved June 16, 2024, from <https://www.canada.ca/en/health-canada/services/chemical-substances/fact-sheets/identification-risk-assessment-priorities.html>
- Health Canada. (2023a). *Health Canada announces the end of cosmetic animal testing in Canada* [News releases]. Retrieved July 10, 2024, from <https://www.canada.ca/en/health-canada/news/2023/06/health-canada-announces-the-end-of-cosmetic-animal-testing-in-canada.html>
- Health Canada. (2023b). *Notice of intent on the development of a strategy to guide the replacement, reduction, or refinement of vertebrate animal testing under the Canadian Environmental Protection Act, 1999 (CEPA)* [Consultations]. Retrieved August 9, 2024, from

<https://www.canada.ca/en/health-canada/programs/consultation-strategy-replace-reduce-refine-vertebrate-animal-testing/notice-intent.html>

- Hillen, M. A., Gutheil, C. M., Strout, T. D., Smets, E. M. A., & Han, P. K. J. (2017). Tolerance of uncertainty: Conceptual analysis, integrative model, and implications for healthcare. *Social Science & Medicine*, *180*, 62–75. <https://doi.org/10.1016/j.socscimed.2017.03.024>
- 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. *Regulatory Toxicology and Pharmacology*, *58*(3), 395–407. <https://doi.org/10.1016/j.yrtph.2010.08.004>
- Hung, C., & Gini, G. (2021). QSAR modeling without descriptors using graph convolutional neural networks: The case of mutagenicity prediction. *Molecular Diversity*, *25*(3), 1283–1299. <https://doi.org/10.1007/s11030-021-10250-2>
- Janzwood, S. (2023). Confidence deficits and reducibility: Toward a coherent conceptualization of uncertainty level. *Risk Analysis*, *43*(10), 2004–2016. <https://doi.org/10.1111/risa.14008>
- Johnson, A., Jin, X., Nakada, N., & Sumpter, J. P. (2020a). Learning from the past and considering the future of chemicals in the environment. *Science*, *367*(6476), 384–387. <https://doi.org/10.1126/science.aay6637>
- Johnson, C., Ahlberg, E., Anger, L. T., Beilke, L., Benigni, R., Bercu, J., Bobst, S., Bower, D., Brigo, A., Campbell, S., Cronin, M. T. D., Crooks, I., Cross, K. P., Doktorova, T., Exner, T., Faulkner, D., Fearon, I. M., Fehr, M., Gad, S. C., ... Myatt, G. J. (2020b). Skin sensitization in silico protocol. *Regulatory Toxicology and Pharmacology*, *116*, 104688. <https://doi.org/10.1016/j.yrtph.2020.104688>
- Johnson, C., Anger, L. T., Benigni, R., Bower, D., Bringezu, F., Crofton, K. M., Cronin, M. T. D., Cross, K. P., Dettwiler, M., Frericks, M., Melnikov, F., Miller, S., Roberts, D. W., Suarez-Rodriguez, D., Roncaglioni, A., Lo Piparo, E., Tice, R. R., Zwickl, C., & Myatt, G. J. (2022). Evaluating confidence in toxicity assessments based on experimental data and in silico predictions. *Computational Toxicology*, *21*, 100204. <https://doi.org/10.1016/j.comtox.2021.100204>
- Jones, C., & Falloon, P. (2009). Sources of uncertainty in global modelling of future soil organic carbon storage. In P. C. Baveye, M. Laba, & J. Mysiak (Eds.), *Uncertainties in Environmental Modelling and Consequences for Policy Making* (pp. 283–315). Springer Netherlands. [https://doi.org/10.1007/978-90-481-2636-1\\_13](https://doi.org/10.1007/978-90-481-2636-1_13)
- Judson, R., Richard, A., Dix, D. J., Houck, K., Martin, M., Kavlock, R., Dellarco, V., Henry, T., Holderman, T., Sayre, P., Tan, S., Carpenter, T., & Smith, E. (2009). The Toxicity Data Landscape for Environmental

- Chemicals. *Environmental Health Perspectives*, 117(5), 685–695.  
<https://doi.org/10.1289/ehp.0800168>
- 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. *Toxicological Sciences*, 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. *Environmental Modelling & Software*, 135, 104905. <https://doi.org/10.1016/j.envsoft.2020.104905>
- Kolmar, S. S., & Grulke, C. M. (2021). The effect of noise on the predictive limit of QSAR models. *Journal of Cheminformatics*, 13(1), 92. <https://doi.org/10.1186/s13321-021-00571-7>
- Levin, R., Hansson, S. O., & Rudén, C. (2004). Indicators of uncertainty in chemical risk assessments. *Regulatory Toxicology and Pharmacology*, 39(1), 33–43. <https://doi.org/10.1016/j.yrtph.2003.11.001>
- Lunghini, F., Marcou, G., Azam, P., Horvath, D., Patoux, R., Van Miert, E., & Varnek, A. (2019). Consensus models to predict oral rat acute toxicity and validation on a dataset coming from the industrial context. *SAR and QSAR in Environmental Research*, 30(12), 879–897. <https://doi.org/10.1080/1062936X.2019.1672089>
- 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. *Alternatives to Laboratory Animals*, 48(4), 146–172. <https://doi.org/10.1177/0261192920965977>
- Mansouri, K., Karmaus, A. L., Fitzpatrick, J., Patlewicz, G., Pradeep, P., Alberga, D., Alepee, N., Allen, T. E. H., Allen, D., Alves, V. M., Andrade, C. H., Auernhammer, T. R., Ballabio, D., Bell, S., Benfenati, E., Bhattacharya, S., Bastos, J. V., Boyd, S., Brown, J. B., ... Kleinstreuer, N. C. (2021). CATMoS: Collaborative Acute Toxicity Modeling Suite. *Environmental Health Perspectives*, 129(4), 47013. <https://doi.org/10.1289/EHP8495>
- Markkanen, R., & Schröder, H. (1997). *Hedging and Discourse: Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts*. Walter de Gruyter.
- Marzi, G., Balzano, M., & Marchiori, D. (2024). K-Alpha Calculator—Krippendorff’s Alpha Calculator: A user-friendly tool for computing Krippendorff’s Alpha inter-rater reliability coefficient. *MethodsX*, 12, 102545. <https://doi.org/10.1016/j.mex.2023.102545>

- Maxim, L. (2015). A systematic review of methods of uncertainty analysis and their applications in the assessment of chemical exposures, effects, and risks. *International Journal of Environmental Health Research*, 25(5), 522–550. <https://doi.org/10.1080/09603123.2014.980782>
- McIlroy-Young, B., Öberg, G., & Leopold, A. (2021). The manufacturing of consensus: A struggle for epistemic authority in chemical risk evaluation. *Environmental Science & Policy*, 122, 25–34. <https://doi.org/10.1016/j.envsci.2021.04.003>
- Meek ME, B., Bolger, M., Bus, J. S., Christopher, J., Conolly, R. B., Lewis, R. J., Paolini, G. M., Schoeny, R., Haber, L. T., Rosenstein, A. B., & Dourson, M. L. (2013). A framework for fit-for-purpose dose response assessment. *Regulatory Toxicology and Pharmacology*, 66(2), 234–240. <https://doi.org/10.1016/j.yrtph.2013.03.012>
- Moschandreas, D. J., & Karuchit, S. (2002). Scenario–model–parameter: A new method of cumulative risk uncertainty analysis. *Environment International*, 28(4), 247–261. [https://doi.org/10.1016/S0160-4120\(02\)00025-9](https://doi.org/10.1016/S0160-4120(02)00025-9)
- Moss, E., Debeuckelaere, C., Berl, V., Elbayed, K., Moussallieh, F.-M., Namer, I.-J., & Lepoittevin, J.-P. (2016). In Situ Metabolism of Cinnamyl Alcohol in Reconstructed Human Epidermis: New Insights into the Activation of This Fragrance Skin Sensitizer. *Chemical Research in Toxicology*, 29(7), 1172–1178. <https://doi.org/10.1021/acs.chemrestox.6b00148>
- Moudgal, C., Anger, L. T., Muster, W., Nguyen, R., Melnikov, F., Siramshetty, V. B., & Graham, J. (2023). The application of acute oral toxicity computational models in dangerous goods classification. *Toxicology and Industrial Health*, 39(12), 687–699. <https://doi.org/10.1177/07482337231209091>
- Mundy, W. R., Padilla, S., Breier, J. M., Crofton, K. M., Gilbert, M. E., Herr, D. W., Jensen, K. F., Radio, N. M., Raffaele, K. C., Schumacher, K., Shafer, T. J., & Cowden, J. (2015). Expanding the test set: Chemicals with potential to disrupt mammalian brain development. *Neurotoxicology and Teratology*, 52, 25–35. <https://doi.org/10.1016/j.ntt.2015.10.001>
- Naidu, R., Biswas, B., Willett, I. R., Cribb, J., Kumar Singh, B., Paul Nathanail, C., Coulon, F., Semple, K. T., Jones, K. C., Barclay, A., & Aitken, R. J. (2021). Chemical pollution: A growing peril and potential catastrophic risk to humanity. *Environment International*, 156, 106616. <https://doi.org/10.1016/j.envint.2021.106616>
- Nantasenamat, C. (2020). Best Practices for Constructing Reproducible QSAR Models. In K. Roy (Ed.), *Ecotoxicological QSARs* (pp. 55–75). Springer US. [https://doi.org/10.1007/978-1-0716-0150-1\\_3](https://doi.org/10.1007/978-1-0716-0150-1_3)

- National Academies of Sciences, Engineering and Medicine. (2017). *Communicating science effectively: A research agenda*. National Academies Press. <https://nap.nationalacademies.org/read/23674/chapter/1>
- National Research Council. (2007). *Read "Toxicity Testing in the 21st Century: A Vision and a Strategy" at NAP.edu*. <https://doi.org/10.17226/11970>
- National Research Council. (2009). *Science and Decisions: Advancing Risk Assessment*. National Academies Press (US). <http://www.ncbi.nlm.nih.gov/books/NBK214630/>
- National Toxicology Program. (2024). *Toxicology in the 21st Century (Tox21)*. <https://ntp.niehs.nih.gov/whatwestudy/tox21>
- Nelms, M. D., Karmaus, A. L., & Patlewicz, G. (2020). An evaluation of the performance of selected (Q)SARs/expert systems for predicting acute oral toxicity. *Computational Toxicology*, *16*, 100135. <https://doi.org/10.1016/j.comtox.2020.100135>
- 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 Disease—PLANT DIS* (pp. 59–117).
- Nickson, T. E. (2008). Planning environmental risk assessment for genetically modified crops: Problem formulation for stress-tolerant crops. *Plant Physiology*, *147*(2), 494–502. <https://doi.org/10.1104/pp.108.118422>
- Noga, M., Michalska, A., & Jurowski, K. (2023). Application of toxicology in silico methods for prediction of acute toxicity (LD50) for Novichoks. *Archives of Toxicology*, *97*(6), 1691–1700. <https://doi.org/10.1007/s00204-023-03507-2>
- OECD. (2002). *Test No. 420: Acute Oral Toxicity - Fixed Dose Procedure*. Organisation for Economic Co-operation and Development. Retrieved June 16, 2024, from [https://www.oecd-ilibrary.org/environment/test-no-420-acute-oral-toxicity-fixed-dose-procedure\\_9789264070943-en](https://www.oecd-ilibrary.org/environment/test-no-420-acute-oral-toxicity-fixed-dose-procedure_9789264070943-en)
- OECD. (2007). *Guidance document on the validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] models*. Retrieved July 22, 2024, from <https://doi.org/10.1787/9789264085442-en>
- OECD. (2018a). *Considerations for Assessing the Risks of Combined Exposure to Multiple Chemicals*. OECD. Retrieved January 11, 2024, from <https://doi.org/10.1787/ceca15a9-en>

- OECD. (2018b). *Users' Handbook supplement to the Guidance Document for developing and assessing Adverse Outcome Pathways* (OECD Series on Adverse Outcome Pathways 1; OECD Series on Adverse Outcome Pathways, Vol. 1). Retrieved October 2, 2024, from <https://doi.org/10.1787/5jlv1m9d1g32-en>
- OECD. (2019). *Guiding Principles and Key Elements for Establishing a Weight of Evidence for Chemical Assessment*. Retrieved October 2, 2024, from OECD. <https://doi.org/10.1787/f11597f6-en>
- OECD. (2023). *(Q)SAR assessment framework: Guidance for the regulatory assessment of (Quantitative) Structure – Activity Relationship models, predictions, and results based on multiple predictions*. Retrieved October 2, 2024, from [https://one.oecd.org/document/ENV/CBC/MONO\(2023\)32/en/pdf](https://one.oecd.org/document/ENV/CBC/MONO(2023)32/en/pdf)
- Ouedraogo, G., Alexander-White, C., Bury, D., Clewell, H. J., Cronin, M., Cull, T., Dent, M., Desprez, B., Detroyer, A., Ellison, C., Giammanco, S., Hack, E., Hewitt, N. J., Kenna, G., Klaric, M., Kreiling, R., Lester, C., Mahony, C., Mombelli, E., ... Cosmetics Europe. (2022). Read-across and new approach methodologies applied in a 10-step framework for cosmetics safety assessment – A case study with parabens. *Regulatory Toxicology and Pharmacology*, *132*, 105161. <https://doi.org/10.1016/j.yrtph.2022.105161>
- Pallocca, G., Moné, M. J., Kamp, H., Luijten, M., Water, B. van de, & Leist, M. (2022). Next-generation risk assessment of chemicals – Rolling out a human-centric testing strategy to drive 3R implementation: The RISK-HUNT3R project perspective. *ALTEX - Alternatives to Animal Experimentation*, *39*(3), Article 3. <https://doi.org/10.14573/altex.2204051>
- Paoli, G., Momoli, F., Tyshenko, M. G., Meek, M. E. B., & Krewski, D. (2022). Problem formulation for EFSA scientific assessments. *EFSA Supporting Publications*, *19*(7), 7349E. <https://doi.org/10.2903/sp.efsa.2022.EN-7349>
- Paparella, M., Bennekou, S. H., & Bal-Price, A. (2020). An analysis of the limitations and uncertainties of in vivo developmental neurotoxicity testing and assessment to identify the potential for alternative approaches. *Reproductive Toxicology*, *96*, 327–336. <https://doi.org/10.1016/j.reprotox.2020.08.002>
- 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. *Regulatory Toxicology and Pharmacology*, *112*, 104592. <https://doi.org/10.1016/j.yrtph.2020.104592>

- Pastoor, T. P., Bachman, A. N., Bell, D. R., Cohen, S. M., Dellarco, M., Dewhurst, I. C., Doe, J. E., Doerrer, N. G., Embry, M. R., Hines, R. N., Moretto, A., Phillips, R. D., Rowlands, J. C., Tanir, J. Y., Wolf, D. C., & Boobis, A. R. (2014). A 21st century roadmap for human health risk assessment. *Critical Reviews in Toxicology*, *44*(sup3), 1–5. <https://doi.org/10.3109/10408444.2014.931923>
- Patlewicz, G., Ball, N., Boogaard, P. J., Becker, R. A., & Hubesch, B. (2015). Building scientific confidence in the development and evaluation of read-across. *Regulatory Toxicology and Pharmacology*, *72*(1), 117–133. <https://doi.org/10.1016/j.yrtph.2015.03.015>
- Patlewicz, G., Ball, N., Booth, E. D., Hulzebos, E., Zvinavashe, E., & Hennes, C. (2013). Use of category approaches, read-across and (Q)SAR: General considerations. *Regulatory Toxicology and Pharmacology*, *67*(1), 1–12. <https://doi.org/10.1016/j.yrtph.2013.06.002>
- Perel, P., Roberts, I., Sena, E., Wheble, P., Briscoe, C., Sandercock, P., Macleod, M., Mignini, L. E., Jayaram, P., & Khan, K. S. (2007). Comparison of treatment effects between animal experiments and clinical trials: Systematic review. *BMJ*, *334*(7586), 197. <https://doi.org/10.1136/bmj.39048.407928.BE>
- 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. *Regulatory Toxicology and Pharmacology*, *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. *Current Opinion in Toxicology*, *15*, 40–47. <https://doi.org/10.1016/j.cotox.2019.04.001>
- Piir, G., Kahn, I., García-S. A. T., Sild, S., Ahte, P., & Maran, U. (2018). Best Practices for QSAR Model Reporting: Physical and Chemical Properties, Ecotoxicity, Environmental Fate, Human Health, and Toxicokinetics Endpoints. *Environmental Health Perspectives*, *126*(12), 126001. <https://doi.org/10.1289/EHP3264>
- Pradeep, P., Friedman, K. P., & Judson, R. (2020). Structure-based QSAR models to predict repeat dose toxicity points of departure. *Computational Toxicology (Amsterdam, Netherlands)*, *16*(November 2020), 10.1016/j.comtox.2020.100139. <https://doi.org/10.1016/j.comtox.2020.100139>
- 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 and QSAR in Environmental Research*, *23*(5–6), 435–459. <https://doi.org/10.1080/1062936X.2012.664825>

- Raies, A. B., & Bajic, V. B. (2016). In silico toxicology: Computational methods for the prediction of chemical toxicity. *Wiley Interdisciplinary Reviews. Computational Molecular Science*, 6(2), 147–172. <https://doi.org/10.1002/wcms.1240>
- Rathman, J. F., Yang, C., & Zhou, H. (2018). Dempster-Shafer theory for combining in silico evidence and estimating uncertainty in chemical risk assessment. *Computational Toxicology*, 6, 16–31. <https://doi.org/10.1016/j.comtox.2018.03.001>
- Raybould, A. (2006). Problem formulation and hypothesis testing for environmental risk assessments of genetically modified crops. *Environmental Biosafety Research*, 5(3), 119–125. <http://dx.doi.org/10.1051/ebr:2007004>
- Reynolds, G., Reynolds, J., Gilmour, N., Cubberley, R., Spriggs, S., Aptula, A., Przybylak, K., Windebank, S., Maxwell, G., & Baltazar, M. T. (2021). A hypothetical skin sensitisation next generation risk assessment for coumarin in cosmetic products. *Regulatory Toxicology and Pharmacology*, 127, 105075. <https://doi.org/10.1016/j.yrtph.2021.105075>
- Robinson, S., Arbez, G., Birta, L. G., Tolk, A., & Wagner, G. (2015). Conceptual modeling: Definition, purpose and benefits. *2015 Winter Simulation Conference (WSC)*, 2812–2826. <https://doi.org/10.1109/WSC.2015.7408386>
- Rubin, V. L. (2007). Stating with Certainty or Stating with Doubt: Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements. In C. Sidner, T. Schultz, M. Stone, & C. Zhai (Eds.), *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers* (pp. 141–144). Association for Computational Linguistics. <https://aclanthology.org/N07-2036>
- Sahlin, U., Filipsson, M., & Öberg, T. (2011). A Risk Assessment Perspective of Current Practice in Characterizing Uncertainties in QSAR Regression Predictions. *Molecular Informatics*, 30(6–7), 551–564. <https://doi.org/10.1002/minf.201000177>
- Sahlin, U., Golsteijn, L., Iqbal, M. S., & Peijnenburg, W. (2013). Arguments for considering Uncertainty in QSAR Predictions in Hazard and Risk Assessments. *Alternatives to Laboratory Animals*, 41(1), 91–110. <https://doi.org/10.1177/026119291304100110>
- Sahlin, U., Helle, I., & Perepolkin, D. (2021). “This Is What We Don’t Know”: Treating Epistemic Uncertainty in Bayesian Networks for Risk Assessment. *Integrated Environmental Assessment and Management*, 17(1), 221–232. <https://doi.org/10.1002/ieam.4367>
- Sahlin, U., Jeliaskova, N., & Öberg, T. (2014). Applicability Domain Dependent Predictive Uncertainty in QSAR Regressions. *Molecular Informatics*, 33(1), 26–35. <https://doi.org/10.1002/minf.201200131>

- Tracy Sarah J. (2018). A phronetic iterative approach to data analysis in qualitative research. *Journal of Qualitative Research*, 19(2), 61–76. <https://doi.org/10.22284/QR.2018.19.2.61>
- Sauve-Cienciewicki, A., Davis, K. P., McDonald, J., Ramanarayanan, T., Raybould, A., Wolf, D. C., & Valenti, T. (2019). A simple problem formulation framework to create the right solution to the right problem. *Regulatory Toxicology and Pharmacology*, 101, 187–193. <https://doi.org/10.1016/j.yrtph.2018.11.015>
- Scalia, G., Grambow, C. A., Pernici, B., Li, Y.-P., & Green, W. H. (2020). Evaluating Scalable Uncertainty Estimation Methods for Deep Learning-Based Molecular Property Prediction. *Journal of Chemical Information and Modeling*, 60(6), 2697–2717. <https://doi.org/10.1021/acs.jcim.9b00975>
- Schieferdecker, S., Rottach, F., & Vock, E. (2024). In Silico Prediction of Oral Acute Rodent Toxicity Using Consensus Machine Learning. *Journal of Chemical Information and Modeling*, 64(8), 3114–3122. <https://doi.org/10.1021/acs.jcim.4c00056>
- 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. *Regulatory Toxicology and Pharmacology*, 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. *Regulatory Toxicology and Pharmacology*, 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>
- Sewell, F., Doe, J., Gellatly, N., Ragan, I., & Burden, N. (2017). Steps towards the international regulatory acceptance of non-animal methodology in safety assessment. *Regulatory Toxicology and Pharmacology*, 89, 50–56. <https://doi.org/10.1016/j.yrtph.2017.07.001>
- Shanahan, J. G., Qu, Y., & Wiebe, J. (2006). *Computing Attitude and Affect in Text: Theory and Applications*. Springer Science & Business Media. <https://link.springer.com/book/10.1007/1-4020-4102-0>
- Skinner, D. J. C., Rocks, S. A., & Pollard, S. J. T. (2014). A review of uncertainty in environmental risk: Characterising potential natures, locations and levels. *Journal of Risk Research*, 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. (2014). Identifying Uncertainty in Environmental Risk Assessments: The Development of a Novel Typology and Its Implications for Risk Characterization. *Human and Ecological Risk Assessment: An International Journal*, 20(3), 607–640. <https://doi.org/10.1080/10807039.2013.779899>
- Sluijs, J. P. van der, Petersen, A. C., Janssen, P. H. M., Risbey, J. S., & Ravetz, J. R. (2008). Exploring the quality of evidence for complex and contested policy decisions. *Environmental Research Letters*, 3(2), 024008. <https://doi.org/10.1088/1748-9326/3/2/024008>
- Solomon, K. R., Wilks, M. F., Bachman, A., Boobis, A., Moretto, A., Pastoor, T. P., Phillips, R., & Embry, M. R. (2016). Problem formulation for risk assessment of combined exposures to chemicals and other stressors in humans. *Critical Reviews in Toxicology*, 46(10), 835–844. <https://doi.org/10.1080/10408444.2016.1211617>
- Sonawane, K. B., Cheng, N., & Hansen, R. A. (2018). Serious Adverse Drug Events Reported to the FDA: Analysis of the FDA Adverse Event Reporting System 2006-2014 Database. *Journal of Managed Care & Specialty Pharmacy*, 24(7), 682–690. <https://doi.org/10.18553/jmcp.2018.24.7.682>
- Stausberg, J., rgen, & Harkener, S. (2023). Data Quality and Data Quantity: Complements or Contradictions? In *Healthcare Transformation with Informatics and Artificial Intelligence* (pp. 24–27). IOS Press. <https://doi.org/10.3233/SHTI230414>
- Steijaert, M. J., Schaap, G., & Riet, J. V. (2021). Two-sided science: Communicating scientific uncertainty increases trust in scientists and donation intention by decreasing attribution of communicator bias. *Communications*, 46(2), 297–316. <https://doi.org/10.1515/commun-2019-0123>
- Stortenbeker, I., Houwen, J., van Dulmen, S., olde Hartman, T., & Das, E. (2019). Quantifying implicit uncertainty in primary care consultations: A systematic comparison of communication about medically explained versus unexplained symptoms. *Patient Education and Counseling*, 102(12), 2349–2352. <https://doi.org/10.1016/j.pec.2019.07.005>
- Strickland, J., Clippinger, A. J., Brown, J., Allen, D., Jacobs, A., Matheson, J., Lowit, A., Reinke, E. N., Johnson, M. S., Quinn, M. J., Mattie, D., Fitzpatrick, S. C., Ahir, S., Kleinstreuer, N., & Casey, W. (2018). Status of acute systemic toxicity testing requirements and data uses by U.S. regulatory agencies. *Regulatory Toxicology and Pharmacology*, 94, 183–196. <https://doi.org/10.1016/j.yrtph.2018.01.022>
- Tepfer, M., Racovita, M., & Craig, W. (2013). Putting problem formulation at the forefront of GMO risk analysis. *GM Crops & Food*, 4(1), 10–15. <https://doi.org/10.4161/gmcr.22906>

- Tosun, J. (2013). How the Eu Handles Uncertain Risks: Understanding the Role of the Precautionary Principle. *Journal of European Public Policy*, 20(10), 1517–1528. <https://doi.org/10.1080/13501763.2013.834549>
- Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*, 29(6–7), 476–488. <https://doi.org/10.1002/minf.201000061>
- United Nations. (2021). *GHS classification criteria for acute toxicity*. Retrieved September 21, 2024, from <https://webapps.ilo.org/static/english/protection/safework/ghs/ghsfinal/ghsc05.pdf>
- US Congress. (2022, September 29). *S.5002 - 117th Congress (2021-2022): FDA Modernization Act 2.0* (2022-09-29) [Legislation]. Retrieved January 22, 2024, from <https://www.congress.gov/bill/117th-congress/senate-bill/5002>
- US EPA. (1998). *Guidelines for ecological risk assessment*. Retrieved September 17, 2024, from [https://www.epa.gov/sites/default/files/2014-11/documents/eco\\_risk\\_assessment1998.pdf](https://www.epa.gov/sites/default/files/2014-11/documents/eco_risk_assessment1998.pdf)
- US EPA. (2012). *Quantitative structure activity relationships [(Q)SAR] guidance document*. Retrieved October 10, 2024, from <https://archive.epa.gov/pesticides/news/web/html/qsar-guidance.html>
- US EPA. (2016). *Phases of ERA - planning and problem formulation* [Collections and Lists]. Retrieved June 15, 2024, from <https://www.epa.gov/ecobox/phases-era-planning-and-problem-formulation>
- US EPA. (2018). *Strategic plan to reduce the use of vertebrate animals in chemical testing* [Other Policies and Guidance]. Retrieved July 8, 2024, from <https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/strategic-plan-reduce-use-vertebrate-animals-chemical>
- US EPA. (2021). *New Approach Methods Work Plan*. Retrieved September 11, 2024, from [www.epa.gov/research](http://www.epa.gov/research)
- US EPA, O. (2015). *Toxicity Estimation Software Tool (TEST)* [Data and Tools]. Retrieved July 1, 2024, from <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]. Retrieved October 10, 2024, from <https://www.epa.gov/pesticide-registration/quantitative-structure-activity-relationship-qsar-guidance-document>
- US EPA. (2011). *Exposure Factors Handbook Chapter 2—Variability and Uncertainty*. Retrieved July 10, 2024, from <https://www.epa.gov/sites/default/files/2015-09/documents/efh-chapter02.pdf>
- Valsecchi, C., Grisoni, F., Consonni, V., & Ballabio, D. (2020). Consensus versus Individual QSARs in Classification: Comparison on a Large-Scale Case Study. *Journal of Chemical Information and Modeling*, 60(3), 1215–1223. <https://doi.org/10.1021/acs.jcim.9b01057>

- van der Bles, A. M., van der Linden, S., Freeman, A. L. J., & Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences*, *117*(14), 7672–7683. <https://doi.org/10.1073/pnas.1913678117>
- van der Zalm, A. J., Barroso, J., Browne, P., Casey, W., Gordon, J., Henry, T. R., Kleinstreuer, N. C., Lowit, A. B., Perron, M., & Clippinger, A. J. (2022). A framework for establishing scientific confidence in new approach methodologies. *Archives of Toxicology*, *96*(11), 2865–2879. <https://doi.org/10.1007/s00204-022-03365-4>
- Van Norman, G. A. (2019). Limitations of Animal Studies for Predicting Toxicity in Clinical Trials: Is it Time to Rethink Our Current Approach? *JACC: Basic to Translational Science*, *4*(7), 845–854. <https://doi.org/10.1016/j.jacbts.2019.10.008>
- Van Norman, G. A. (2020). Limitations of Animal Studies for Predicting Toxicity in Clinical Trials: Part 2: Potential Alternatives to the Use of Animals in Preclinical Trials. *JACC: Basic to Translational Science*, *5*(4), 387–397. <https://doi.org/10.1016/j.jacbts.2020.03.010>
- Varttala, T. (2001). *Hedging in Scientifically Oriented Discourse*. <https://intapi.sciendo.com/pdf/10.1515/rjes-2019-0015#:~:text=The%20main%20functions%20of%20hedging,be%20valid%20in%20all%20circumstances.>
- Verdonck, F. a. M., Van Sprang, P. A., & Vanrolleghem, P. A. (2005). Uncertainty and precaution in European environmental risk assessment of chemicals. *Water Science and Technology: A Journal of the International Association on Water Pollution Research*, *52*(6), 227–234.
- Viceconti, M., Pappalardo, F., Rodriguez, B., Horner, M., Bischoff, J., & Musuamba Tshinanu, F. (2021). In silico trials: Verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products. *Methods*, *185*, 120–127. <https://doi.org/10.1016/j.ymeth.2020.01.011>
- Vighi, M., Barsi, A., Focks, A., & Grisoni, F. (2019). Predictive models in ecotoxicology: Bridging the gap between scientific progress and regulatory applicability—Remarks and research needs. *Integrated Environmental Assessment and Management*, *15*(3), 345–351. <https://doi.org/10.1002/ieam.4136>
- Vold, E. T. (2006). Epistemic modality markers in research articles: A cross-linguistic and cross-disciplinary study. *International Journal of Applied Linguistics*, *16*(1), 61–87. <https://doi.org/10.1111/j.1473-4192.2006.00106.x>

- Wadood, A., Ahmed, N., Shah, L., Ahmad, A., Hassan, H., & Shams, S. (2013). In-silico drug design: An approach which revolutionarised the drug discovery process. *OA Drug Design and Delivery*, 1(1). <https://doi.org/10.13172/2054-4057-1-1-1119>
- 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. *Integrated Assessment*, 4(1), 5–17. <https://doi.org/10.1076/iaij.4.1.5.16466>
- Walum, E. (1998). Acute oral toxicity. *Environmental Health Perspectives*, 106(suppl 2), 497–503. <https://doi.org/10.1289/ehp.98106497>
- Wang, B., & Gray, G. (2015). Concordance of Noncarcinogenic Endpoints in Rodent Chemical Bioassays. *Risk Analysis*, 35(6), 1154–1166. <https://doi.org/10.1111/risa.12314>
- Wang, D., Yu, J., Chen, L., Li, X., Jiang, H., Chen, K., Zheng, M., & Luo, X. (2021). A hybrid framework for improving uncertainty quantification in deep learning-based QSAR regression modeling. *Journal of Cheminformatics*, 13(1), 69. <https://doi.org/10.1186/s13321-021-00551-x>
- 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. *Regulatory Toxicology and Pharmacology*, 63(1), 10–19. <https://doi.org/10.1016/j.yrtph.2012.02.006>
- Wang, Z., Walker, G. W., Muir, D. C. G., & Nagatani-Yoshida, K. (2020). Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. *Environmental Science & Technology*, 54(5), 2575–2584. <https://doi.org/10.1021/acs.est.9b06379>
- Welss, T., Basketter, D. A., & Schröder, K. R. (2004). In vitro skin irritation: Facts and future. State of the art review of mechanisms and models. *Toxicology in Vitro*, 18(3), 231–243. <https://doi.org/10.1016/j.tiv.2003.09.009>
- WHO/IPCS. (2008). *Part 1: Guidance document on characterizing and communicating uncertainty in exposure assessment* (6th ed.). World Health Organization. <https://www.who.int/ipcs/methods/harmonization/areas/uncertainty%20.pdf>
- Wikoff, D., Haws, L., Ring, C., & Budinsky, R. (2019). Application of qualitative and quantitative uncertainty assessment tools in developing ranges of plausible toxicity values for 2,3,7,8-tetrachlorodibenzo-p-dioxin. *Journal of Applied Toxicology: JAT*, 39(9), 1293–1310. <https://doi.org/10.1002/jat.3814>

- Wolt, J. D., Keese, P., Raybould, A., Fitzpatrick, J. W., Burachik, M., Gray, A., Olin, S. S., Schiemann, J., Sears, M., & Wu, F. (2010). Problem formulation in the environmental risk assessment for genetically modified plants. *Transgenic Research*, *19*(3), 425–436. <https://doi.org/10.1007/s11248-009-9321-9>
- Worth, A., Fuart-Gatnik, M., Lapenna, S., & Serafimova, R. (2011a). Applicability of QSAR analysis in the evaluation of developmental and neurotoxicity effects for the assessment of the toxicological relevance of metabolites and degradates of pesticide active substances for dietary risk assessment. *EFSA Supporting Publications*, *8*(6). <https://doi.org/10.2903/sp.efsa.2011.EN-169>
- Worth, A., Lapenna, S., Lo Piparo, E., Mostrag-Szlichtyng, A., & Serafimova, R. (2011b). *A Framework for assessing in silico toxicity predictions: Case studies with selected pesticides*. JRC Publications Repository. <https://doi.org/10.2788/29048>
- Zerva, C. (2019). *Automatic identification of textual uncertainty*. 422. [https://research.manchester.ac.uk/files/86864517/FULL\\_TEXT.PDF](https://research.manchester.ac.uk/files/86864517/FULL_TEXT.PDF)
- Zhang, D., Liu, D., Jing, J., Jia, B., Tian, Y., Le, Y., Yu, Y., & Hu, Q.-N. (2024). Unveiling the chemical complexity of food-risk components: A comprehensive data resource guide in 2024. *Trends in Food Science & Technology*, *148*, 104513. <https://doi.org/10.1016/j.tifs.2024.104513>
- Zhang, Y., & Lee, A. A. (2019). Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chemical Science*, *10*(35), 8154–8163. <https://doi.org/10.1039/C9SC00616H>
- Zheng, C., & Bennett, G. D. (2002). *Applied contaminant transport modeling, 2nd Edition | Wiley*. Wiley.Com. <https://www.wiley.com/en-us/Applied+Contaminant+Transport+Modeling%2C+2nd+Edition-p-9780471384779>
- Zhong, S., Lambeth, D. R., Igou, T. K., & Chen, Y. (2022). Enlarging Applicability Domain of Quantitative Structure–Activity Relationship Models through Uncertainty-Based Active Learning. *ACS ES&T Engineering*, *2*(7), 1211–1220. <https://doi.org/10.1021/acsestengg.1c00434>
- Zhu, H., Martin, T. M., Ye, L., Sedykh, A., Young, D. M., & Tropsha, A. (2009). QSAR Modeling of Rat Acute Toxicity by Oral Exposure. *Chemical Research in Toxicology*, *22*(12), 1913–1921. <https://doi.org/10.1021/tx900189p>

## Appendices

### Appendix A. Supplementary materials associated with Chapter 2

Table S2.1. List of studies in the general risk assessment literature (outside of *in silico* methods or alternative to animal testing approaches in general) describing higher-level conceptual problem formulation components.

<b>Author</b>	<b>Higher-level component</b>
Meek et al. (2013)	Assessment context (prioritization and screening); Endpoint; Safe dose;
Callahan and Sexton (2007)	Conceptual model; Analysis plan
Jones and Falloon (2009)	Assessment objectives; Spatial and temporal scales; Measurement endpoints
Nickson (2008)	Assessment Endpoints; Conceptual Model; Analysis Plan
Poli et al. (2022)	Assessment question (scientific questions to be addressed); Conceptual model; Risk hypothesis; Legislative context an assessment; Research needs
Embry et al. (2014)	Scenario description (e.g., chemical use); Highlighting existing knowledge; Context description (e.g., potential exposure scenario); Defining acceptable margin of exposure; Regulatory options
Devos et al. (2019)	Conceptual model; Formulating hypothesis; Defining what qualifies as harm
Felter et al. (2021)	Problem scoping; Setting up hazard or risk analysis plan; Purpose of the assessment (hazard identification and classification or risk assessment)
OECD (2019)	Assessment scope and goals; Acceptable level of uncertainty; Defining urgency of the assessment
Pastoor et al. (2014)	Defining chemical exposure
Sauve-Ciencewicki et al. (2019)	Framing problem; Framing problem; Problem Statement; Conceptual Model

---

Solomon et al. (2016)	Planning and scoping; Identifying and characterizing stressors; Conceptual Model; Plan of Analysis
US EPA (2016)	Stressors (duration of persistence and frequency of occurrence); Sources (e.g., background levels), exposure (media and routes); Susceptibility and sensitivity of the receptor
Wolt et al. (2010)	Assessment endpoints; Risk hypotheses; Conceptual model; Exposure; Level of Uncertainty
WHO/IPCS (2018)	Acceptable levels of uncertainty and risk; Assessment endpoints; Exposure scenarios; Analysis plan and information needs; Risk management scope and assessment goals

Table S2.2: List of the thirteen *in silico* method-related published in the peer-reviewed papers that were analyzed in the present study. They were selected as they proposed or discussed the need for problem formulation in the development of *in silico* methods.

Author	Description of the study
Parish et al. (2020)	Develop a framework for fit-for-purpose evaluation of NAM's application in the regulatory context of chemical risk assessment. The proposed PF calls for the specification of the regulatory use of NAM (i.e., chemical prioritization, hazard identification, or/and risk assessment).
Pestana et al. (2021)	Appraise uncertainty in read-across using Assessment Elements (AEs) from the European Chemicals Agency's Read-Across Assessment Framework (RAAF). A 90-day oral sub-chronic toxicity of triazole in rats was used as a case study.
Escher et al. (2019)	Outline a general read-across assessment to support hazard characterization of grouped compounds by generating data on the chemicals' dynamic and kinetic properties.
Baltazar et al. (2020), Reynolds et al. (2021)	Systematic toxicity assessment of 0.1% coumarin in face cream and body lotion in an exposure-led approach and a battery of NAMs (including <i>in vitro</i> assays, physiologically based kinetic models, and Skin Allergy Risk Assessment Model).
Ouedraogo et al. (2022)	Develop an illustrative 10-step read-across framework for propylparaben cosmetic safety assessment as a proof-of-concept for the value added by NAMs in next-generation risk assessment of chemicals.
Dent et al. (2018)	Propose principles for incorporating NAMs into risk assessments of cosmetic ingredients using the guidelines of next-generation risk assessment, such as ensuring assessments are human-relevant, exposure-led, hypothesis-driven and designed to prevent harm.
Belfield et al. (2021)	Evaluate quantitative structure-activity relationships models in terms of their uncertainty, variability and potential areas of bias by mapping out the models' components onto specific regulatory uses (i.e., risk assessment, classification and labelling, and screening and prioritization).

---

Cronin et al. (2019)	Identify opportunities and challenges to implementing <i>in silico</i> methods (e.g., quantitative structure-activity relationships and read across) to assess the safety of chemicals like pharmaceuticals, personal care products, and industrial chemicals.
Sewell et al. (2017)	Review the rate of progress in regulatory acceptance of non-animal methodology into regulatory and identify ways to expedite progress
Schultz et al. (2019)	Identify major sources of uncertainty (e.g., regulatory use of the prediction and data for the apical endpoint being assessed) that has the potential to impact acceptance of read-across argument.
Ball et al. (2014)	Develop a framework that incorporates <i>in silico</i> , <i>in vitro</i> and <i>in vivo</i> methods for REACH requirements in assessing chemical hazard and exposure using a tiered approach.
Pallocca et al. (2022)	Delineate how the practical applicability of NAMs and strategies will be deployed to establish an overall next generation risk assessment framework for chemicals and other substances like drugs.

---

## Appendix B. Supplementary materials associated with Chapter 3

Table S3.1. Verbatim recorded sources of uncertainty (VRSU) in 11 peer-reviewed papers that discuss sources of uncertainty in in silico toxicology methods. The sources of uncertainty marked with asterisk (\*) were deemed irrelevant for developing the categorization framework in the present study.

Author	<i>in silico</i> method	VRSU	Publication number
Blackburn & Stuard (2014)	Read across	Quantity of the data considered	1
		Number of analogues contributing data	1
		Quality of the data	1
		Suitability of analogues	1
		Structural similarity to target	1
		Nature and severity of the identified toxic effects	1
		The potency of the analogues for those [toxic] effects	1
Wang et al. (2012)	QSAR	Database deficiencies (e.g., lack sensitive endpoint or toxicity information)	2
		Extrapolations (interspecies (animal-to-human))	2
		Extrapolations intraspecies (susceptible human subpopulation)	2
		Extrapolations (subchronic-to-chronic)	2
		Extrapolations (LOAEL-to-NOAEL)	2
Schultz et al. (2019)	Read across	Number of analogues contributing data	3
		Completeness of the argument provided [for data quality]	3
		Strength or robustness of the supporting data sets	3
		Quality of the apical endpoint data	3
		Similarity in chemistry	3
		Toxicokinetic similarity	3
		Toxicodynamic similarity	3
		Weight-of-Evidence	3
		Mechanistic plausibility	3
Acceptable level of uncertainty*	3		
Pham et al. (2019)	QSAR	Size of training data set data	4
		Balance of the training data set	4

		Distribution of the training data set	4
		Similarity justification	4
		Choice of molecular descriptors	4
		Modeling algorithm and hyperparameter	4
		Applicability domain	4
		Data variability*	4
Schilter et al. (2014)	QSAR; Grouping; Read across	Number of the chemical analogs identified	5
		Suitability of the chemical analogs identified	5
		Quality of data used to build model	5
		Toxicological information found for the analogs	5
		Structure and its representation	5
		Prediction of complex endpoints such as chronic toxicity	5
		Model performance	5
		Extrapolation of the toxicity of the substance of interest based on data on analogs	5
		Applicability domain	5
		Variability of biological data*	5
Cronin et al. (2019)	QSAR	Data balance	6
		Homogeneity of the chemical space of the training and test sets	6
		Relevance of data for the endpoint of interest	6
		Completeness of the data set	6
		Consistency of the data set	6
		Quality of data	6
		Calculated/Experimentally measured properties and descriptors	6
		Reproducibility of model and model prediction	6
		Adequacy of the model to make a prediction for the stated purpose	6
		Statistical performance	6
		Mechanistic relevance and interpretability	6
		Adequate coverage of Absorption, Distribution, Metabolism and Excretion effects	6
		Applicability domain	6
		Relevance [of the QSAR] to the prediction or assessment goal	6
		Error associated with biological data*	6
Cronin et al. (2022)		Coverage [of structural alert]	7

	Structural alerts	Structural description	7
		Property domain	7
		Species specificity	7
		Toxicity or relationship to adversity	7
		Supporting evidence	7
		Corroborating evidence	7
		Performance	7
		Mechanistic causality	7
		Metabolic domain	7
		Purpose or potential use of the structure alert	7
Schultz et al. (2015)	Read across	Number of source chemicals	8
		Robustness of the source or analogue data	8
		Quality of the source or analogue data	8
		Definition and demonstration of similarity	8
		Mechanistic plausibility	8
		Mechanistic relevance	8
		Applicability domain	8
Pestana et al. (2021)	Read across	Consistency of data	9
		Robustness of the supporting data sets	9
		Quality of data	9
		Weight-of-evidence supporting the prediction	9
		Mechanistic plausibility	9
Madden et al. (2020)	QSAR;	Relevance of data	10
	QSPR;	Reliability of data	10
	qAOP;	Accuracy of data	10
	PBPK/PBT	Validity of data	10
	K		
Benfenati et al. (2019)	Read across	[computed/not experimentally measured] Parameters used to construct the model	11
		Numerical errors and/or numerical approximations	11
		Model bias	11
		Parametric variability *	11

LOAEL = lowest observed adverse effect level; NOAEL = no-observed-adverse-effect level; QSAR = quantitative structure–activity relationship; QSPR = quantitative) structure–property relationship; qAOP = Quantitative AOP; PBPK/PBTK = Physiologically-based pharmacokinetic/physiologically-based toxicokinetic.

Table S3.2. The refined general sources of uncertainty (GSU) resulting from the analysis and iterative categorization of the verbatim recorded sources of uncertainty (VRSU). The descriptions of these GSU are deduced from the description of the corresponding VRSU in Table S3.1.

<b>Modelling Phase</b>	<b>Higher-level assessment component</b>	<b>GSU</b>	<b>Description of GSU as deduced from the description of the VRSU in the original papers</b>	
Model creation	Data	Data quantity	Amount of data - whether data is sufficiently available	
		Data balance	Ratio between the number of chemicals in categories in the training dataset – chemical categories with known activities (toxicants) and known non-activities (non-toxicants)	
		Data relevance	Data contain target information (e.g., kinetics and metabolic property) suitable for modelling or adequate for the interpretation of model predictions	
		Data reliability	Reproducibility of data between test approaches/sources or reproducibility of the methodology used in generating the data	
		Data accuracy	The extent to which measured data deviates from its true value	
		Data validity	Acceptability of the method used to generate data relative to set guidelines or whether the method measured what it was intended to measure	
	Structure	Chemical structure	Quality (e.g., in terms of relevance) of chemical structures or substructures with respect to a set prediction	
	Similarity	Chemical similarity	Resemblance or commonality between chemical compounds, e.g., in terms of functional groups and chemical structure	
	Descriptors	Descriptor relevance	Descriptor relevance	Extent to which physicochemical or molecular descriptors are considered toxicologically relevant or suitable for deriving chemical properties or for a specific prediction task
			Descriptor concordance	Degree of agreement between descriptors and other chemical features or chemical toxicokinetic or toxicodynamic properties
Model characterization	Modelling	Model structure	A model endogenous representation, such as mathematical formulations (e.g., equations or graphs), choice of algorithms, precision of numerical approximations, and relationships between variables	

		Activity/potency	Measure of elicited toxicological effect or adverse effect, degree of the effect, or ability of a chemical to exert an effect on a receptor
		Activity/potency evidence	Available evidence to support the predicted activity/potency
	Performance	Model performance	Predictivity of a model or how well a model can predict outcomes of interest, which can be evaluated through, for example, an internal/external validation or quantitatively using the measure of statistical fit
	Mechanisms	Mechanistic plausibility	Toxic causal pathways of chemicals, involving the identification of molecular initiating events/key events linked causally to a target endpoint
	Toxicokinetics	Metabolic domain	Consideration of production or presence of metabolites as part of chemical interaction with biological systems
		Coverage of ADME activity	Consideration of ADME activities in biological systems, including effects of metabolites
Model application	Applicability	Applicability domain	Boundaries within which a model can be applied and provide reliable and accurate predictions (e.g., adequacy of chemical structure space or category to predict effects of similar chemicals)
	Relevance	Extrapolation	Making predictions beyond the range of the observed/known data (e.g., toxicity data) in attempts to obtain new unknown data
		Model relevance	Transferability of a model or model prediction to a different prediction context (e.g., regulatory application or prediction of new compounds)

Table S3.3. GHS acute toxicity hazard categories (United Nations, 2021)

<b>Category</b>	<b>Hazard statement</b>	<b>Precautionary word</b>
1 (LD <sub>50</sub> ≤ 5 mg/kg)	Fatal if swallowed	Danger
2 (5 < LD <sub>50</sub> ≤ 50 mg/kg)	Fatal if swallowed	Danger
3 (50 < LD <sub>50</sub> ≤ 300 mg/kg)	Toxic if swallowed	Danger
4 (300 < LD <sub>50</sub> ≤ 2000 mg/kg)	Harmful if swallowed	Warning
5 (LD <sub>50</sub> > 2000 ≤ 5000 mg/kg)	May be harmful if swallowed	Warning
LD <sub>50</sub> > 5000 mg/kg	Not classified	No specified label

## Appendix C. Supplementary materials associated with Chapter 4

Text S4.1. Implicit and Explicit Epistemic Uncertainty Indicators used in the present study

Levin et al. (2004) define implicit epistemic uncertainty indicators as "words and phrases that indicate less than full confidence in, or commitment to, the propositional content of a statement" (e.g., "it is presumed that", "would [not] be expected to", and "it is unlikely that") and inferential uncertainty indicators as words and phrases that convey uncertainties related to inferences (e.g., "on this basis, it is presumed", "suggests", and "since ... it is unlikely that"). My preliminary analysis of peer-reviewed QSAR studies revealed that epistemic claims are often implicitly inferential. For example, "suggests" in the sentences "This observation suggests that the LD50..."(Liu et al., 2010; p 152) and "...blue regions at 4' position of 5-phenyl ring suggests that increased activity..." (Amnerkar and Bhusari, 2010; p 1912) indicates the authors' incomplete knowledge to make definitive conclusions about the specific subjects (this indicates epistemic uncertainty). From the description of inferential uncertainty indicator by Levin et al. (2004), it is the case that "suggests" in this case should also be considered to be arising because of inferences made. Given the difficulty in distinguishing whether or not epistemic claims are inferential, I decided to take epistemic uncertainty indicators more broadly to also include inferential indicators (see Figure S4.1).

In the application of Bayesian networks for risk prediction, Sahlin et al. (2021) describe explicit epistemic expressions as qualitative or quantitative statements that directly reveal uncertainties about the probability of adverse effects. They provide two examples to explain explicit epistemic uncertainty indicators: (1) "we are 90% certain that the risk of extinction is less than 5%", implying that the probability of extinction risk ( $< 5\%$ ) = 90%", and (2) "we don't know" or "can't know", which explicitly acknowledges a lack of knowledge or confidence in an estimation". Similar examples can be found in other studies – for example, "I don't know" and "it is not clear" (Stortenbeker et al., 2019; pg. 2349). The concepts of explicit uncertainty indicators described by these authors were adopted in this study.



Figure S4.1. Epistemic and inferential uncertainty indicators and their definitions, as proposed by Levin et al. (2004). Examples of these types of indicators are provided from the study by Avery et al. (2002). Inferential uncertainty indicators were subsumed within epistemic uncertainty indicator.

#### Text S4.2. Identification of implicit and explicit uncertainty indicators

The coding process followed the procedure suggested by Lichtenstein and Rucks-Ahidiana (2023). Two undergraduate students (third and fourth collaborators) with basic knowledge of uncertainty indicators received training from me about how to (1) distinguish between implicit and explicit uncertainty indicators, (2) distinguish between epistemic uncertainty indicators from the other types of uncertainty indicators, (3) conduct line-by-line text analysis in this context, (4) manually colour-code the implicit and explicit epistemic uncertainty indicators in the texts in the studies (stored as Zotero pdf documents). Colour coding was done by highlighting and marking the segments of the texts containing implicit or explicit uncertainty indicators using the yellow colour in the Zotero toolbar. One paper (i.e., Avery et al. (2002)) that did not meet the selection criteria under section 4.2.1 of Chapter 4 was picked for this training exercise. Examples of implicit and explicit uncertainty indicators identified in this paper are shown in Figure S4.1.

After the training exercise, two text coding practice sessions were conducted using two additional papers (i.e., Amnerkar and Bhusari (2010) and Schmidt et al. (2004)) that did not meet the selection criteria. Through line-by-line text analysis, the three coders (third and fourth collaborators and myself) independently identified, colour-marked and recorded implicit and explicit epistemic uncertainty indicators in these papers. Thereafter, the intercoder agreement measure, calculated as the percent agreement, was estimated using Krippendorff's Alpha ( $\alpha_k$ ) (Krippendorff, 2004) on the web-based K-Alpha Calculator (Marzi et al., 2024). The measure was on whether the coders agreed if the highlighted text was an indicator and, if yes, whether the indicator qualified to be categorized as an explicit or implicit epistemic indicator. All conflicting issues were identified, discussed, and resolved. During the practice exercise, I noted that some indicators conveyed two meanings: epistemic possibility and possibility when merely stating facts. For example, I interpreted "can" in this sentence "[...] compound 52 was statistically significant and can be used as a lead [...]" (Amnerkar and Bhusari, 2010; p. 153) to be conveying epistemic possibility of using compound 51, while "can" in the sentence "The 3D-QSAR results can be visualized using

3D plots of crucial pharmacophore regions (Fig. 3 a–e)" (Amnerkar and Bhusari, 2010; p. 159) to be conveying possibility about the fact that the results can be seen on the plotted figure. Only indicators that relate to epistemic possibility were included in this study.

The actual coding exercise was then done after the practice exercise. Accordingly, the third and fourth collaborators and myself independently coded each of the 20 papers (Table S4.1), identifying, colour-marking, and recording implicit and explicit indicators. The three of us each reviewed each paper at least twice until no new indicators could be discovered. The intercoder agreement was then estimated using Krippendorff's Alpha ( $\alpha_k$ ) (Krippendorff, 2004) on the web-based K-Alpha Calculator (Marzi et al., 2024). Thereafter, all conflicting issues were identified, discussed, and resolved before the final list of indicators was recorded (Table S4.1).

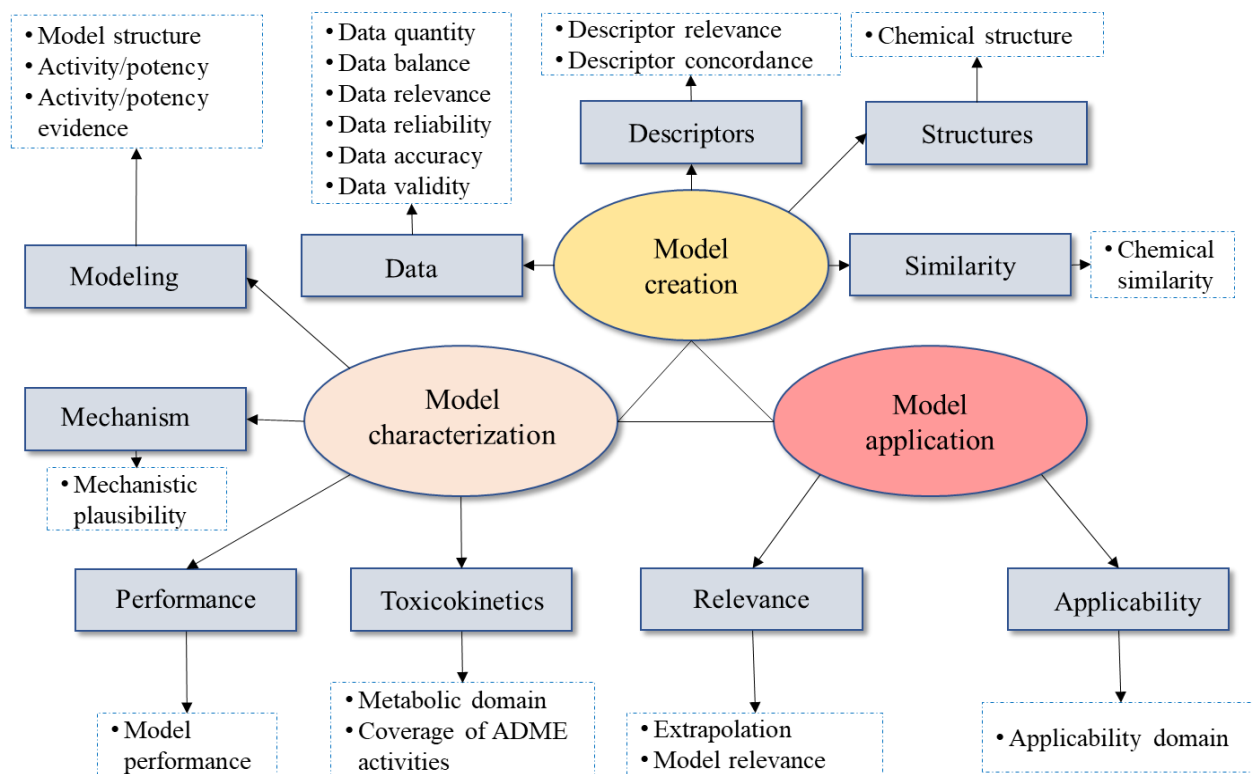


Figure S4.2. Sources of uncertainty relating to practices and features common to in silico toxicology modeling (adapted from Achar et al. (2024a) – under review – with permission from the authors).

Table S4.1. The identified implicit and explicit uncertainty indicators (bolded in the in sentences) from the 20 analyzed studies (as denoted by the publication number): Alonso et al. (2013) (1), Basant et al. (2016) (2), Chushak et al. (2018) (3), Cronin et al. (1996) (4), Estrada et al. (2001) (5), Gadaleta et al. (2022) (6), Gentile et al. (2020) (7), Jiang et al. (2020) (8), Kan et al. (2021) (9), Kan et al. (2022) (10), Luan et al. (2013) (11), Makaeva et al. (2013) (12), Marimuthu et al. (2019) (13), Mozaffari et al. (2012) (14), Pradeep et al. (2020) (15), Rayne and Forest (2010) (16), Turabekova et al. (2008) (17), Yazal et al. (2001) (18), Zhang et al. (2019) (19), Zhao et al. (2022) (20).

Study #	Implicit uncertainty indicator (bolded in the sentences)	Page #	Implicit uncertainty source	Explicit uncertainty indicator (bolded in the sentences)	Page #	Explicit uncertainty source
1	Consequently, this equation <b>may</b> become a tool to extrapolate experimental results [...]	1395	Extrapolations; Model relevance	This mx-QSAR has excellent goodness-of-fit statistics [...] with <b>sensitivity (Sn), specificity (Sp), and accuracy (Ac) &gt; 80%</b> .	1394	Model performance
	A very recent result <b>claims</b> that Fimasartan, an antihypertension drug, suppressed iNOS [...]	1398	Mechanistic plausibility			
	This <b>may be</b> in concordance with the prediction [...]	1398	Mechanistic plausibility			
	These results <b>may indicate a certain probability</b> that compound 11 is a multitarget ligand.	1398	Activity/potency			
	[...] was very recently synthesized and assayed under the hypothesis <b>implying</b> that multitarget ligands <b>may</b> provide more efficient neuroprotection [...]	1398	Activity/potency; Mechanistic plausibility			
	It is expected that network-based tools <b>may be</b> applied for the discovery of new drugs [...]	1398	Model relevance			
	New carbamate derivatives of 3-OH rasagiline <b>may be</b> prepared with interesting neuroprotective effects.	1398	Model relevance			
	[...] the model <b>may</b> become a useful tool for the identification of allonetwork drugs in the future.	1398	Model relevance			
2	reshold value of cR2 p is 0.5 and a model exceeding this value <b>might be</b> considered [...]	47	Model performance			

The results <b>suggest</b> that the constructed models are not due to chance correlation.	48	Model performance
High values for the sensitivity [...] for both the models <b>suggest</b> the robustness of the models.	48	Model performance
Fjodorova et al. (Fjodorova et al., 2010) <b>suggested</b> that classification model [...]	48	Model relevance
The values [...] were close to unity <b>suggesting</b> the adequacy of the proposed model [...]	48	Model performance
AUROC also <b>strongly supports the reliability</b> of our developed classification models.	48	Model performance
A high value of classification accuracy [...] <b>may be</b> due to the fact that the chemicals [...]	48	Data relevance; Chemical similarity
[...] Y- the original STR models are <b>unlikely to</b> arise as a result of chance correlation.	48	Model performance
[...] prediction errors <b>suggest</b> for a good-fit of the GRNN model to the data sets [...].	48	Model performance
[...] previous QSAR studies <b>may be</b> due to the linear modeling approach considered there.	49	Model structure
The anomalous behavior of these chemicals <b>may be</b> due to one of the following reasons [...]	49	Chemical structure
[...] the lipid bilayer of the cellular membrane is a <b>potential</b> site for interaction.		Mechanistic plausibility
Hence, the Log P descriptor <b>might be</b> related to the ability of solvents to penetrate through [...]	50	Descriptor concordance; Mechanistic plausibility
[...] atoms in a molecule, <b>suggesting</b> double, triple, or aromatic substituent over alkyl substituent.	50	Descriptor concordance
The proposed QAAR model <b>can be a possible</b> supporting tool [...]	50	Model relevance
[...] predicted values of the endpoint toxicity <b>suggested</b> a high correlation between them.	51	Model performance

	Visual inspection of the plot <b>suggested</b> that none of the chemical solvents considered here exhibited [...]	51	Applicability domain		
	The absence of any outlier in the test and training data <b>implies</b> a wider applicability [...]	51	Model relevance		
	[...] models here <b>may be attributed to</b> their ability to work well in noisy environments.	51	Model structure		
3	[...] structurally diverse chemicals that <b>can potentially</b> cause a range of adverse outcome effects [...]	425	Activity/potency; Mechanistic plausibility	The <b>distribution of active chemicals for these proteins was exceptionally skewed</b> , with some proteins [...]	429 Data balance
	ESR1 is <b>associated with</b> several nervous system diseases including migraine <sup>21</sup> [...]	425	Activity/potency; Mechanistic plausibility	These results indicate that <b>it is currently difficult</b> to predict [...]	429 Activity/potency
	HLA-DRA is <b>associated with</b> Parkinson's disease <sup>24</sup> and multiple sclerosis [...]	425	Activity/potency; Mechanistic plausibility	The classification <b>prediction results showed 79% accuracy while AC50 predictions demonstrated mixed accuracy</b> compared with [...]	429 Model performance
	[...] indicates the very high activity of this drug having <b>potential</b> side effects.	427	Activity/potency	[...] value <b>could not be calculated</b> from the dose–response data	429 Activity/potency
	Volinanserin was tested in clinical trials as a <b>potential</b> antipsychotic drug [...]	427	Activity/potency	[...] making it <b>difficult to identify the active/inactive boundary</b> for chemicals with low activity.	429 Activity/potency
	The primary mechanism of pyrethroid neurotoxicity is <b>proposed</b> to be via their effects [...]	427	Mechanistic plausibility	[...] <b>a different grouping schema needs to be applied</b> that will identify a different set of similar chemicals.	429 Model relevance
	Another <b>potential</b> target of pyrethroid compounds is the dopamine transporter SLC63.	428	Mechanistic plausibility	Therefore, <b>it is of interest to investigate</b> the correlation of molecular interactions [...]	429 Mechnistic plausibility
	[...] group that <b>potentially explains</b> the high number of incorrect predictions.	429	Applicability domain		

4	This is <b>thought</b> to be a measure of molecular volume	105	Descriptor concordance	Unfortunately <b>there are no truly reliable predictions available</b> for these parameters at the moment	104	Data reliability
	This <b>may</b> give a simple measure of the volatility of a compound with respect to its being able to enter the vapour phase.	105	Descriptor concordance	These are the data set [...] are <b>clearly not comparable</b> to the rest of the data set.	105	Descriptor concordance
	This <b>may</b> have obvious implication for compounds distributed in the air.	105	Descriptor relevance	<b>It is more difficult</b> to rationalize the reason [...]	106	Descriptor concordance
	An easier approach <b>seemed to be</b> to use the temperature required to achieve a vapour [...]	105	Descriptor relevance	[...] <b>many of these data are not available</b> at 25°C [...]	106	Data quantity
	Surprisingly, the above equation <b>suggests</b> a lack of relationship with hydrophobicity.	105	Descriptor concordance	[...] most <b>confidence intervals are between 10% and 30% of the original value.</b>	108	Data accuracy
	These compounds <b>may</b> be more easily trapped in lipid membranes, without a polar or polarizable blue [...]	105	Mechanistic plausibility	<b>Exact molecular mechanisms are not known</b> although [...]	108	Mechanistic plausibility
	This <b>may be</b> a factor of steric bulk with the ring molecule and <b>may at least partly explain</b> [...]	106	Descriptor concordance	QSARs such as those described are <b>intrinsically unable to model rates of biotransformation</b> [...]	108	Model relevance
	Therefore, the removal of the four outliers <b>suggest</b> that hydrophobicity is the only parameter [...]	106	Descriptor relevance	<b>Relatively little work has been performed</b> on QSAR analysis [...]	109	Activity/potency
	Further analysis of this relationship <b>suggests</b> other outliers, namely [...]	106	Descriptor relevance	<b>It is obvious that more work is required in this area</b> both [...]	109	Extrapolation
	Pyridine and dioxane <b>may be considered</b> atypical of the data set, being <b>somewhat</b> polar.	106	Descriptor relevance			
	[...] Frantik et al. (1994) do <b>suggest</b> there <b>may be</b> problems due to low aqueous solubility.	106	Descriptor relevance			
	[...] connectivity <b>suggesting</b> that the membrane permeability of large molecules <b>may be</b> reduced.	106	Descriptor concordance; Mechanistic plausibility			

---

[...] which is usually <b>associated with</b> QSARs for membrane permeability [...]	106	Descriptor concordance
Figure 2 <b>suggests it is possible</b> to fit a line to the graph that expresses the minimal toxicity [...]	106	Descriptor concordance
[...] the statistics of this equation, which <b>probably reflects</b> the large scatter of the data [...]	107	Model structure
This <b>may be due</b> to equilibrium between air and membrane <b>being more likely</b> to be achieved [...]	107	Mechanistic plausibility
[...] achieved <b>could be</b> a major source of error in these relatively short test times, and <b>may be</b> why [...]	107	Data reliability
This is an arbitrary value and the two groups <b>can be</b> labelled simply 'high neurotoxicity' [...]	107	Activity/potency
This <b>suggests</b> that all these parameters are important for neurotoxicity, despite the fact that [...]	107	Descriptor relevance
This <b>may again be related back to</b> the differences in toxicodynamics and toxicokinetics [...]	108	Activity/potency evidence
[...] a large area of physiochemical parameter space that is <b>associated with</b> a high level of neurotoxicity.	108	Descriptor concordance
[...] provides a <b>potential</b> method to separate highly neurotoxic compounds in the rat and mouse [...]	108	Model performance
This should prove useful as a <b>possible</b> qualitative prescreening method for neurotoxicity [...]	108	Model relevance
[...] error than the biological data it is <b>likely to be</b> as a result of overfitting or a spurious correlation.	108	Model performance
Also adding to the poor correlations are specific effects which <b>are likely to</b> occur <i>in vivo</i> [...]	108	Model performance
Exact molecular mechanisms are not known although <b>it is likely</b> solvents interfere with cell membranes [...]	108	Mechanistic plausibility

---

	Seeman et al. (1969) <b>suggested</b> that the depressant effects of organic solvents on the central nervous [...]	108	Mechanistic plausibility		
	Korpela and TIhti (1988) <b>proposed</b> that the anaesthetic potency of organic solvents <b>could be</b> dependent [...]	108	Mechanistic plausibility		
	Franks and Lieb (1990) <b>postulated</b> that amphiphilic pockets in the integral proteins <b>may be</b> the [...]	108	Mechanistic plausibility		
	It <b>may also be</b> true that metabolism to more toxic compounds is an important cause [...]	108	Coverage of ADME activities		
	Both of these P-450s are <b>closely associated with</b> the toxic [...] P4502E <b>appears to</b> mediate [...]	109	Mechanistic plausibility		
	Such mechanisms <b>may be relevant</b> to neurotoxic events and will not be modelled [...]	109	Mechanistic plausibility		
	Other work <b>suggests</b> that halogenated alkenes <b>may be</b> neurotoxic by /I-lyase cleavage [...]	109	Mechanistic plausibility		
	This study has demonstrated that solubility as a parameter <b>may not be</b> vitally important [...]	109	Descriptor relevance		
	Measures of volatility, such as vapour pressure <b>may, however, be</b> of greater value.	109	Descriptor relevance		
	At concentrations less than this 'minimal toxicity' there <b>may still be</b> an effect [...]		Activity/Potency		
	This is an area that <b>could be</b> linked to an integrated approach with in vitro toxicity tests	109	Model relevance		
5	This procedure <b>can be</b> useful for the identification of possible toxicophores [...]	447	Model relevance	[...] because <b>their boiling point was not available.</b>	449 Data quantity
	[...] it was observed that these two physiochemical properties <b>could be</b> of certain importance [...]	450	Descriptor relevance	[...] effects of cyclohexyl acetate <b>have not been investigated adequately.</b>	454 Activity/potency

	In the case of dioxane, that is outlier for models (2) and (3), <b>we can think</b> that neurotoxicities [...]	454	Descriptor concordance; Chemical structure	[...] some structural factor also influencing its volatility that is <b>not well accounted for</b> by the [...]	454	Chemical structure
	<b>It could be speculated</b> that the greater water solubility of dioxane compared [...]	454	Descriptor relevance	[...]some influence that it is <b>not accounted for</b> the value [...]	454	Descriptor concordance
	According to the values of fragment contributions <b>it appears that</b> alkyl halide fragments [...]	455	Activity/potency	Unfortunately, <b>there is not enough experimental data</b> [...]	457	Data quantity; Activity/potency evidence
6	[...] a protein, and <b>may</b> lead to damage of the protein and the <b>potential</b> loss of its function. This <b>may</b> affect thiol- and seleno-containing proteins, which offer antioxidant protectio.	7	Mechanistic plausibility			
	[...] (MIE D and H) <b>may</b> lead to several functional impairments, such as in learning and memory.	7	Mechanistic plausibility			
	Interference at any of these levels (MIEs G and Q-T) <b>may</b> lead to decreased thyroxine (T4) and [...]	7	Mechanistic plausibility			
	[...] cochlear development and <b>potentially</b> resulting in permanent auditory loss.	7	Mechanistic plausibility			
	[...] brain maturation is <b>likely to</b> cause such varied adverse outcomes [...]	8	Mechanistic plausibility			
	<b>One possible reason</b> for this high performance is the high structural homogeneity of the active samples.	3	Applicability domain; Data balance			
	Thyroid elements <b>seem to be</b> relevant (THR, TPO and TTR) to neurotoxicity [...]	4	Mechanistic plausibility			
	[...] AMPAR and KAR <b>seem to be</b> more linked to neurotoxicity than NMDAR.	4	Mechanistic plausibility			
	AMPA/kainite receptor-mediated neurotoxicity was reported to <b>possibly</b> play a role [...]	5	Mechanistic plausibility			
	This is <b>likely to be</b> due to the fact that RFs perform better if trained on a larger pool of variables [...]	5	Model performance			

	[...] providing insights into the <b>possible</b> mode of action of a predicted neurotoxic chemical.	5	Mechanistic plausibility			
	[...] cells are all <b>potential</b> targets that can be disrupted by neurotoxicants with different <b>possible</b> mechanisms of toxicity.	6	Activity/potency; Mechanistic plausibility			
	Predictions of MIE provided by QSARs <b>may</b> give indications of which assays to prioritise [...]	6	Model relevance			
7	[...] 0.5 Å was used as a similarity threshold, below which two conformations were <b>assumed</b> identical.		Chemical similarity	Unfortunately, <b>none of these models succeeded</b> in finding compounds more potent [...]	3	Model relevance
	[...] points (pharmacophore areas F5 and F6, Figure 1) that are <b>associated with</b> the interactions [...]	3	Descriptor concordance	[...] while being <b>aware that this limits the possibility</b> of finding structurally new molecules.	3	Model relevance
	<b>Probably</b> , the less unfavorable contacts of the ketal group inside the sub-pocket [...]	4	Mechanistic plausibility			
	[...] -8.7 kcal/mol for LC/B, <b>probably due to</b> the smaller size with respect to the other [...]		Mechanistic plausibility			
	[...] <b>highlighting the possibility</b> of developing broad-spectrum inhibitors against iatrogenic botulism.		Model relevance			
	[...] ADMET properties <b>could</b> have ruled out <b>potentially</b> interesting [...]		Data relevance			
	On the contrary, molecules 6 and 9 <b>could be</b> effluated from the central nervous system (CNS) [...]		Coverage of ADME activities			
	The values obtained <b>allow us to deduce</b> that the molecules <b>can be sufficiently</b> distributed [...]	12	Coverage of ADME activities			
	[...] compounds <b>could be</b> suitable for valid candidates as drugs and <b>could lead to</b> further [...]	12	Data reliability			
	The poses of compounds 3 and 4 <b>suggest</b> that the sub-pocket, which includes His223 and Arg363 residues, <b>may be</b> used to [...]		Mechanistic plausibility; Model relevance			

	[...] highlighting that small inhibitors (such as ZINC5729284) <b>could be</b> extremely valuable [...]		Model relevance			
8	It <b>may</b> distort the linear correlation between variables measured at the same scale [...]	166	Model structure	[...] near the diagonal, and <b>the prediction error is relatively small.</b>	168	Model performance
	<b>To a certain extent, this indicates</b> that the structural diversity of our compounds is high [...]	167	Applicability domain	Although <b>the predictive power of our model is not the best</b> [...]	170	Model performance
	The reason why we performed feature scaling was that [...], <b>it was likely</b> to dominate the objective [...]	168	Model structure			
	[...] six compounds in the training set <b>were considered</b> to be outside the applicability domain	168	Applicability domain			
	[...] 28 compounds in the training set <b>were considered</b> to have a large interference effect [...]	168	Applicability domain			
	[...] when multiple decision trees are combined, a good prediction effect <b>can be</b> achieved.	170	Model performance			
	These results <b>can be</b> used for the accurate quantitative prediction of chemical induced [...]	171	Model relevance			
9	The training dataset consisting of 681 chemicals and their <b>associated</b> neuronal cytotoxicity [...]	2	Data relevance	[...] <b>the imbalance dataset</b> was further adjusted [...]	4	Data balance
	[...] a proper adjustment of the AD of the model <b>could</b> further improve the prediction results.	5	Applicability domain	[...] reasonably good <b>sensitivity, specificity, balanced accuracy, accuracy, and AUC values of 0.577, 0.898, 0.737, 0.861 and 0.850</b> , respectively.	4	Model performance
	[...] 1% AUC value <b>were considered</b> a convergence and the selection of rules was stopped.	5	Applicability domain	[...] AUC of the <b>prediction model for the independent test dataset became 0.877 and 0.851</b> , respectively	4	Model performance

[...] 3 preservatives with high lipophilicity which <b>may</b> have high blood brain permeability [...]	5	Activity/potency	<b>Future works are desirable for characterizing their effects</b> on neural systems.	6	Activity/potency
Our model [...] <b>could be</b> useful as a first-tier screening tool for identifying chemicals [...]	5	Model relevance	There was <b>no SH-SY5Y cytotoxicity data associated with BAK, [...]</b>	6	Activity/potency
We identified 15 preservatives as <b>potential</b> neuronal cytotoxicants and were able to experimentally [...]	5	Activity/potency	<b>BAK was not identified as neuronal toxicants</b> for SH-SY5Y by our prediction model, <b>although the predicted value for BAK (0.275) was very close</b> to the decision [...]	6	Activity/potency
[...] formulation and <b>has been suggested</b> to play roles in the neurotoxicity <b>associated with</b> ketamine [...]	5	Activity/potency	<b>MI was not predicted as a neuronal cell toxicant in our model</b> (predicted value 0.17). However, MI had been shown to be cytotoxic to rat cortical neuron cells at 100 µM or higher	6	Activity/potency; Model relevance
Likewise, the chemicals <b>may</b> still exhibit toxicity in other cell types.	6	Activity/potency	[...] <b>there were still some false positives</b> in the independent [...]	7	Applicability domain
[...] preservative in eye drops <b>associated with</b> corneal neurotoxicity in animal studies.	6	Activity/potency	While <b>AUC values were only slightly improved with AD adjustment</b> , the high AUC [...]	7	Applicability domain
<b>It is reasonable that</b> the model is more suitable for predicting SH-SY5Y cytotoxicity [...]	7	Model performance	However, <b>there was no report suggesting that the chemical is neurotoxicity in vivo.</b>	7	Activity/potency
This presented model also <b>appeared to be</b> in good concordance [...] experimental evidence.	7	Model performance	<b>Future works include the evaluation for the neurotoxicity mechanisms</b> of the preservatives [...].	7	Mechanistic plausibility
Future works <b>could be</b> the curation of neuronal cytotoxicity data [...]	7	Model relevance	<b>Future studies may be needed</b> to further evaluate whether or not these preservatives [...]	7	Activity/potency



	[...] specific adverse outcome <b>could</b> enhance the specificity of the prediction.	3310	Model performance			Data quantity;
	The model <b>may be</b> further optimized when more AOPs and corresponding data [...]	3310	Data quantity; Model performance	Further experimental data [...] <b>are needed</b> to verify the model.	3312	Model performance
	This indicated that other <b>potential</b> mechanisms which <b>may</b> lead to parkinsonian motor deficits [...]	3310	Mechanistic plausibility	<b>Limited by the availability of</b> experimental data [...]	3312	Data quantity
	Each <b>may be</b> influenced by different parts of the motor control framework and cannot be [...]	3310	Mechanistic plausibility	The model may be further optimized <b>when more data measuring other KEs become available.</b>	3312	Data quantity
	[...] predicted as a parkinsonian neurotoxicant, <b>had been suggested to</b> interact with [...]	3311	Mechanistic plausibility	The other 9 neurotoxicants are <b>worthy further evaluated for PD potential.</b>	3312	Activity/potency
	The search results from ChemDIS also <b>suggested</b> that the neurotoxins not predicted to be parkinsonian neurotoxicants <b>may still be related to PD.</b>	3312	Activity/potency; Activity/potency evidence			
	The model <b>may therefore not be able</b> to predict neurotoxicants which induce parkinsonian [...]	3312	Model relevance			
	The model <b>may be</b> further optimized when more data measuring other KEs become available.	3312	Data quantity; Data relevance			
	The model <b>may also be</b> extended to combine other mechanism-based inference models [...]	3312	Model relevance			
	The integrated model presented <b>could</b> facilitate the identification of chemicals with PD potential.	3313	Model relevance			
11	One compound <b>may</b> lead to 1 or more statistical cases because it <b>may</b> give different outcomes [...]		Data reliability	This linear equation presented good results [...] with <b>overall Accuracy in training series above 90%.</b>	1872	Model performance; Model structure
	The different conditions that <b>may</b> change in the dataset are the following: organisms [...]		Data relevance			

	These models <b>are expected</b> to give different classification probabilities of the compound [...]	1872	Model relevance		
	One compound <b>may</b> lead to 1 or more statistical cases because it <b>may</b> give different outcomes for alternative biological assays carried out in diverse sets [...]	1872	Activity/potency		
	This result <b>may be</b> in coincidence with the experimental value of protection [...]	1873	Model performance		
12	The use of enzymes from different tissues and species is a <b>potential limitation</b> of the study. [...]	232	Data reliability	However, the whole picture of influence is rather <b>complicated</b> .	236 Mechanistic plausibility
	[...] liver carboxylesterase 1 (CES1), is produced in the liver, which <b>may</b> export it to plasma [...]	232	Extrapolation; Mechanistic plausibility		
	We have <b>suggested</b> that the ratios $ki(BChE)/ki(AChE)$ and $ki(CaE)/ki(AChE)$ characterize the <b>potential</b> contribution of BChE and CaE [...]	233	Descriptor concordance; Mechanistic plausibility		
	The models <b>suggest</b> that short-chain alkoxy substituents without a- and b-branching [...]	236	Mechanistic plausibility		
	[...] esterase profiles for <b>potential application</b> as selective inhibitors of CaE [...]	236	Model relevance		
13	[...] AAHHR.61 – was reliable and <b>could be</b> used further for the 3D-QSAR model development.	310	Model relevance	[...] – MnAChE <b>still remain unexplored</b> .	309 Mechanistic plausibility
	[...] R2 and Q2 is $< .15$ (.9615–.8221 = .1394), <b>suggesting</b> the close correspondence [...]	311	Model performance	Likewise, there are <b>no protein sequence and a three-dimensional structure available</b> [...]	309 Mechanistic plausibility
	Collectively, these details will provide the <b>plausible reasoning</b> for an increase or a decrease [...]	312	Mechanistic plausibility		

	This <b>suggests</b> that more hydrophobic substituents are favorable at the tail regions [...]	312	Descriptor concordance		
	[...] phosphate ion and proximal to aromatic ring <b>suggest</b> the hydrophobic substituents [...]	312	Descriptor concordance		
	This <b>suggests</b> that the presence or substitutions of polar atoms at these regions is highly favorable [...]	312	Descriptor concordance		
	[...] oxygen atoms of phosphate group <b>suggests</b> electron-withdrawal property is unfavorable [...]	312	Descriptor concordance		
	[...] nitrobenzene group of OP22 <b>suggest</b> highly favorable positive-ionic region.	313	Descriptor concordance		
	This <b>suggests</b> that the presence of positively charged ion at this position is highly favored [...]	313	Descriptor concordance		
	The results obtained from the predicted model <b>could be attributed to</b> the experimental verifications.	313	Model performance		
	Hence, <b>we propose</b> that the identified hit compounds can act as a virtual lead and <b>can be tested</b> [...]	318	Model relevance		
14	However, <b>it is proposed</b> that they act through inhibition of the GABA transaminase enzyme [...]	3799	Mechanistic plausibility	The anticonvulsant mechanism of the semicarbazones <b>is not clearly defined.</b>	3799 Mechanistic plausibility
	These results are <b>probably</b> caused by an elevation of GABA levels in the brain for –NO <sub>2</sub> –, –OH– [...]	3799	Mechanistic plausibility	High dose [...] emerged as <b>promising anticonvulsant agents</b> was further tested [...]	3800 Activity/potency
	Based on our results, <b>it can be deduced</b> that substitution of 4-Br at R position in combination [...]	3800	Mechanistic plausibility		
	One <b>tentative explanation</b> for this event <b>could be</b> related to increased hydrogen [...]	3800	Mechanistic plausibility		

	It <b>seems that</b> the lipophilicity of these molecules plays an important role [...]	3800	Descriptors concordance; Mechanistic plausibility		
	[...] 4-ethoxy and this effect <b>might</b> lead to more affinity toward hydrophobic [...].	3800	Descriptors concordance; Mechanistic plausibility		
	<b>One more possible reason</b> is the improved diffusion of 4-Br-substituted compounds [...]	3800	Descriptors concordance; Mechanistic plausibility		
	It <b>seems</b> that the presence of different substituents on the second phenyl ring [...]	3803	Mechanistic plausibility		
	[...] consequently the activity of the final compounds <b>could be</b> influenced.	3803	Descriptor concordance; Mechanistic plausibility		
	This kind of activity was <b>attributed to</b> the presence of 4-nitrophenyl and 2-hydroxyphenyl groups [...]	3803	Mechanistic plausibility		
15	<b>We considered this to be potential</b> outlier.	14	Data relevance	The experimental data from each of the sources <b>do not have high concordance</b> [...]	18 Data reliability
	[...] the NEFs derived in this study, <b>suggesting</b> that more PCBs have a higher NEF value [...]	18	Data relevance	[...] emphasizing the <b>lack of robustness in the models</b> on account of the <b>underlying experimental data quality and variability.</b>	19 Model performance; Data reliability
	The reason for this observed association <b>could be</b> because the distributions of values [...]	18	Data reliability	<b>The QSAR models developed were not robust.</b>	19 Model performance
	[...] NEF prediction can be <b>attributed to</b> two major factors [...]	18	Model performance	We believe that the <b>predictivity is limited by the quality and quantity of underlying experimental data</b> [...]	19 Data reliability; Data quantity

	The width of the REP range <b>can be roughly</b> interpreted as the lowest possible value [...]	19	Model performance	The number of <b>data points limits the ability of a machine learning algorithm</b> to learn [...]	19	Data quantity
	This <b>may</b> impact the accuracy of experimentally measured AC50 or other potency values.	19	Data accuracy	[...] issue of <b>lack of correspondence between the different assays for the same chemical.</b>	19	Data reliability
	[...] isolated PCBs <b>may</b> vary from laboratory to laboratory which <b>may</b> result in low concordance [...] The predictions for any given congener <b>appear</b> quite similar for any of the modeling approaches [...] <b>We believe</b> that the predictivity is limited by the quality and quantity [...]	19	Data reliability  Model performance  Data quantity; Data relevance	[...] assays covered <b>too few chemicals</b> for practical model building by each mode of action.	19	Data quantity
16	[...] PCB congeners outside the training set due to <b>potential</b> high unreliability of the estimates.  As a result, the full congener QSAR model <b>suggests</b> that the experimental data set [...]  <b>It appears that</b> no high-potency PCB congeners with EC2x values <<0.2 µM exist.	358	Model performance; Applicability domain	Because of the <b>poor predictivity of the pEC50 QSAR</b> , and concerns [...].	358	Model performance
	[...] Pessah et al.[42] <b>suggests</b> that, in addition to the <b>likely</b> minimum mono-ortho substitution [...]	358	Activity/potency	[...] experimental data [...] <b>only available for &lt;10 to 20% of all possible PCB congeners.</b>	359	Data quantity
	[...] values shown in Table 1 for these <b>potentially</b> inactive congeners are high (from 1.5 to 6.5 µM) [...]	359	Data reliability	[...] given the <b>poor quality optimized QSAR</b> obtained from the experimental data set. <b>If future work finds a suitable QSAR for the EC50 data set</b> , these findings can readily be integrated into our results.	359	Model relevance
				[...] antagonistic effects of PCB mixtures <b>are poorly defined</b> [...]	360	Activity/potency

[...] the resulting interpretations <b>may be</b> subject to bias where the congeners tested [...]		Extrapolation	[...] as well as the metabolites themselves, <b>are poorly defined.</b>	360	Coverage of ADMET activity
[...] multiple ortho substituents are more <b>likely to be</b> found in higher homologues.	359	Descriptors concordance	[...] leading to <b>difficulties in extrapolating the results</b> of [...]	360	Extrapolation
[...] RyR1 potency are more <b>likely to be</b> found in the high homologue groupings.	359	Descriptors concordance	Consequently, <b>in the absence of additional data, it is unclear</b> whether [...]	360	Activity/potency evidence
[...] assessment regarding the <b>potential</b> deactivating potential of a 2,4,4',6-substitution pattern.	359	Mechanistic plausibility	[...] at this point, <b>the lack of a multicongener data set</b> for establishing [...]	360	Data quantity
It is also important to stress that the <b>possible</b> synergistic and/or antagonistic effects of PCB [...]	360	Activity/potency; Data relevance			
[...] PCB mixtures are poorly defined but <b>may</b> confound efforts to develop [...]	360	Data relevance			
The <b>potentially</b> high RyR1 activity of various PCB metabolites will also complicate neurotoxicity [...]	360	Coverage of ADME activities			
[...] RyR1 enantioselectivity <b>may be</b> different among various chiral PCBs.	360	Activity/potency			
[...] (from wild type pigs),[34] <b>suggesting</b> that individuals possessing malignant hyperthermia mutations within RyR1 <b>may be</b> more susceptible [...]	360	Activity/potency; Mechanistic plausibility			
As with the <b>potential</b> enantioselective nature of chiral PCBs towards RyR1, at this point [...]	360	Activity/potency; Mechanistic plausibility			
[...] congeners in the experimental data set, <b>suggesting</b> these congeners [...]	361	Data relevance			
[...] less than 0.2 µM, <b>suggesting</b> the limited experimental data reported to date <b>appears to</b> have adequately mapped the range of <b>potential</b> PCB RyR1 activities.	361	Data relevance			
[...] number and Aroclor chlorination <b>likely</b> reflect indirect [...] substituents are more <b>likely to be</b> [...]	361	Descriptor concordance; Mechanistic plausibility			

17	The data <b>suggest</b> alkaloids inducing sodium channel activation are more toxic than those alkaloids [...]	3	Data relevance; Activity/potency; Mechanistic plausibility	[...] <b>no substantial features have been identified that would help to distinguish [...]</b>	5	Chemical structure; Mechanistic plausibility
	Any type of redundancy <b>might</b> lead to an overexploitation of a chemical property [...]	4	Descriptor concordance	[...] which however <b>can not be solely responsible</b> for opening the sodium ion channels.	5	Mechanistic plausibility
	[...] (for ground and protonated states) <b>suggests</b> that this group plays an important [...]	6	Descriptor concordance; Mechanistic plausibility			
	Additionally, it is <b>very likely</b> that benzoyl ester group defines the toxicity of the compound [...]	6	Descriptor concordance			
	One of these models <b>suggested</b> an aromatic centre, electron donor atom and hydrogen-bond [...].	7	Descriptor concordance; Mechanistic plausibility			
	The correlation observed <b>might be</b> the consequence of the fact that Authors used LUMO [...]	9	Descriptor concordance; Mechanistic plausibility			
	[...] reactivity properties is <b>associated with</b> the high antiarrhythmic activity.	9	Mechanistic plausibility			
	The correlation observed between HOMO-LUMO gap and the activity data <b>suggests</b> that formation [...]	9	Activity/potency			
	[...] antiarrhythmic activity <b>likely due to</b> formation of H-bonds between ligand and the receptor site	9	Mechanistic plausibility			
	The descriptor nHDon <b>might be considered</b> as the general case of OHt descriptor [...]	9	Descriptor concordance			
	[...] LogP <b>suggests</b> binding of alkaloids occurs on the surface of the binding site where partial desolvation <b>might occur</b> .	9	Descriptor concordance; Mechanistic plausibility			

[...] anesthetic activity <b>appeared to be considerably</b> higher (between 40.4 and 318).	9	Activity/Potency
We <b>suggested</b> curariform and antiarrhythmic alkaloids [...] are all <b>likely to be</b> classified as “non-drugs”.	10	Descriptor concordance (*2)
<b>Most likely</b> , such correlation reflects a common property of antiarrhythmic compounds [...]	10	Descriptor concordance; Mechanistic plausibility
Structurally and energetically, this <b>correlation can be explained by</b> the presence of benzoylester [...]	10	Descriptor concordance; Mechanistic plausibility
An increase in the number of aliphatic ketons <b>seems</b> to lower toxicity.	10	Descriptor concordance
[...] receptor site <b>is likely to</b> have hydrogen bond acceptor atoms according to the model 7.	10	Mechanistic plausibility
The bad fitting to correlation line for those compounds <b>can be explained by possible</b> errors [...]	11	Data Accuracy
[...] this research have confirmed the experimental findings <b>suggesting</b> neurotoxins [...]	11	Activity/potency evidence; Mechanistic plausibility
Presence of the nHDon, nOHt and nCO descriptors in the models <b>suggest</b> that [...]	11	Descriptor concordance; Mechanistic plausibility
[...] aliphatic amino groups <b>have appeared to</b> favor antiarrhythmic activity.	11	Descriptor concordance
[...]this study <b>were considered</b> as the important structure alerts (SAs) for the neurotoxicity, which <b>could be considered</b> in structural [...]	7	Model relevance
<b>It suggested</b> that the substructures produced by ECFP_10 fingerprints <b>could</b> better describe [...]	7	Descriptor relevance

	[...] which <b>suggested</b> that the naïve Bayes classifier was more suitable for the neurotoxicity prediction.	8	Model structure		
	Therefore, the established NB-03 prediction model <b>can be</b> used as a reliable [...]	10	Model relevance		
	[...] neurotoxicity were identified in this research, which <b>could</b> give an important guidance [...]	10	Model relevance		
18	[...]some higher-energy structures that <b>may be meaningful</b> for receptor binding [...]	225	Mechanistic plausibility	Database members [...] <b>not deemed</b> as hits.	229 Data relevance
	This is <b>suggestive of the potential</b> for increased potency [...]	227	Descriptor concordance;	The <b>database does not provide IC50 values</b> , however.	229 Data quantity
	These values <b>imply the potential</b> usefulness of the above pharmacophores [...]	228	Activity/potency Model relevance	<b>Limitations of the application of these findings to human toxicity need to be acknowledged.</b>	229 Extrapolation
	[...]compounds of significant biological interest and it <b>can be</b> comparatively more useful [...]	229	Data relevance	The source of the IC50 values [...] may <b>provide some uncertainty.</b> However, <b>as additional internally consistent IC50 data sets become available [...]</b>	230 Data relevance 230 Data quantity; Data reliability
19	[...] features of a molecule, which <b>could</b> capture molecular features relevant to [...]	2	Descriptor relevance	<b>The specificity (SP) was 91.5%.</b>	4 Model performance
	[...] which <b>suggested</b> the selected descriptors were <b>closely related</b> to neurotoxicity.	4	Descriptor concordance	The ECFP_10 and eight molecular descriptors <b>were not able to better describe</b> the property [...]	4 Descriptor relevance
	[...] number of C atoms have a good <b>association with</b> neurotoxicity.	4	Descriptor concordance	The NB-03 prediction model <b>generated positive predictive value (PPV), negative predictive value (NPV), balance accuracy (BA) and overall prediction accuracy (Q) for the training set</b>	6 Model performance

			<b>were 96.5%, 78.0%, 90.8% and 90.5%, respectively.</b>	
[...] smaller molecule is more <b>likely to</b> cause adverse effect to the nervous system.	4	Descriptor concordance' Activity/potency	<b>The sensitivity (SE) was 58.6%.</b>	6 Model performance
The Molecular_Weight shows a better <b>association with</b> neurotoxicity [...]	4	Descriptor concordance	<b>The specificity (SP) was 68.0%. The NB-01 model achieved 84.2% positive predictive value (PPV), 36.2% negative predictive value (NPV) and 63.3% balanced accuracy (BA).</b>	10 Model performance
<b>It suggests</b> the neurotoxicants tend to be less lipophilic than the non-neurotoxicants.	4	Descriptor concordance	[...] best prediction performance, which gave <b>90.5% overall prediction accuracy</b> for the training set and <b>82.1% concordance</b> [...].	Model performance
The neurotoxic class is distributed between [...], <b>suggesting</b> that neurotoxicants usually [...]	4	Descriptor concordance		
[...] represents there has a good <b>association</b> between number of rings and neurotoxicity.	4	Descriptor concordance		
<b>It suggested</b> the classification model based on only eight molecular descriptors [...]	4	Model relevance		
The [...]model was <b>quite reliable and could achieve satisfactory</b> capacity [...]	4	Model performance; Model relevance		
[...] structural fragments has a <b>high possibility</b> to be neurotoxicant.	5	Descriptor concordance		
[...] toxic substructures [...] <b>were considered</b> as the important structure alerts (SAs) for the neurotoxicity, which <b>could be considered</b> in structural modification [...]	5	Descriptor relevance; Model relevance		
<b>It suggested</b> that the substructures produced by ECFP_10 fingerprints <b>could better</b> describe [...]	7	Descriptor relevance		

	[...] produced by the NB-03 [...] <b>suggested</b> that the naïve Bayes classifier [...]	10	Model structure			
20	This method also <b>may</b> work on the characterization of the structural space for diverse compounds. The performance of the models on the validation set <b>could</b> provide objective evidence [...] [...] high lipid solubility are <b>more likely to</b> cross the BBB and act on the CNS, which <b>may be</b> one of [...] The results <b>suggest</b> that nHAcc and nHDon <b>may be</b> obviously <b>associated with</b> drug-induced [...] Drug-induced neurotoxicity was also significantly <b>associated with</b> nRotB (mean values of 5.33 [...] [...] chemical structures, which <b>can be considered</b> as SAs for chemical neurotoxicity. [...] atoms into a compound <b>may</b> not improve the lipophilicity of an organic compound. Toluene <b>can</b> inhibit the activity of aminopeptidase in the rat brain. Toluene <b>can</b> also adversely affect the vestibular and audiovisual motor system function In addition, toluene <b>may</b> impede the [...] from nerve endings. They <b>can</b> be served as a useful tool in the early stages of drug discovery to quickly assess[...] [...] this study, which <b>may</b> provide useful information for the understanding of the mechanism [...]	6036 6041 6041 6042 6042 6042 6042 6042 6043 6043 6043 6043 6043 6043	Applicability domain Model performance Mechanistic plausibility Descriptor concordance; Mechanistic plausibility Descriptor concordance Model relevance Chemical structure Mechanistic plausibility Mechanistic plausibility Mechanistic plausibility Model relevance Model relevance	[...] <b>it is difficult</b> to explore the mechanism [...] First, the <b>data used in the study is a little bit small.</b> [...] distinction would be <b>difficult to achieve.</b> <b>Admittedly, these methods are not perfect</b> because they [...] Third, <b>there are still several shortcomings of the methods for the SA detection.</b> [...] substructures <b>cannot be characterized,</b> [...] performance with an overall <b>prediction accuracy of 83.70% for external validation.</b>	6042 6043 6043 6043 6043 6043 6043 6043 6043 6043 6043 6043 6043 6043 6043	Mechanistic plausibility Data quantity Model relevance Model relevance Model relevance Model relevance Model relevance Model performance

---

[...] chemical neurotoxicity and <b>may</b> play an important role for the prediction of drug-induced neurotoxicity.	6043	Model relevance
The machine learning models and SAs <b>could be</b> useful tools for drug- induced neurotoxicity prediction.	6043	Model relevance
Meanwhile, the study <b>may</b> also contribute to the understanding [...]	6043	Model relevance

---

Table S4.2. The intercoder agreement scores (fourth, fifth, and sixth columns) obtained during the practice session. 1 = agreement; 2 = disagreement.

Study	Identified indicators ( <b>bolded in the sentences</b> )	Publication page #	Coder 1	Coder 2	Coder 3
Amnerkar & Bhusari (2010)	All of the compounds were active in the MES test which is an <b>indicative</b> of their ability to prevent seizure spread.	150	2	1	1
	[...] 3D-QSAR models with the aim that these models <b>could</b> provide useful pharmacophoric information for the future efforts in the development of more potent molecules in these series of chemical classes.	152	1	1	1
	The red regions in the vicinity of 30 and 40 position of 5-phenyl ring <b>suggested</b> that the substructure fragments with hydrogen-bond donor in this area <b>may</b> reduce the activity (compounds 36, 42 and 48, Fig. 3 b).	152	1	1	1
	In the negative ionic plots (Fig. 3c), blue regions at 40 position of 5-phenyl ring <b>suggests</b> that [...]	152	1	1	1
	[...] increased activity <b>may be anticipated</b> by moderate negatively charged substituents at 40 position (compounds 34 and 52) [...]	152	1	1	1
	[...] near tertiary nitrogen of pyrazoline ring defines that an increase in positive group <b>may</b> result in enhanced activity (compound 33).	153	1	1	1
	[...] 6 position of benzothiazole ring <b>signifies</b> that substitution with hydrophobic group [...]	153	1	2	1
	[...] red region in the vicinity of 40 position of 5-phenyl ring <b>suggests</b> that hydrophobic group in this area resulted in [...]	153	1	1	1
	[...] the compounds 9–32 displayed weaker anticonvulsant activity which <b>might be</b> due to the lack of positive group [...]	153	1	1	1
	<b>We believe</b> that the derived 3D-QSAR as well as clues for possible structural modifications will be of interest [...]	153	1	1	1

Schmidt et al. (2004)	[...] different types of terpenoids previously <b>suggested</b> on the basis of qualitative structure–activity [...].	4162	1	1	1
	[...] qualitative structure–activity considerations <sup>8</sup> was <b>plausible</b> .	4162	1	1	1
	The hypothetical binding site possesses two major areas where polar/electrostatic interactions <b>would</b> take place.	4162	1	1	2
	[...] large area where hydrogen bonds/salt bridges <b>can be</b> formed with the protons on OH groups [...].	4163	2	1	1
	The statistical quality is somewhat inferior to the housefly model, which <b>may, however, be explained</b> by the use of estimated DG0 values in many cases.	4164	1	1	1
	Overall, the surface of the rat binding site <b>appears to be</b> slightly more polar (see particle property distribution in Table 2).	4164	1	1	1
	[...] yet with slight differences in size or position, which <b>would be</b> in accordance with the expected similarity [...].	4164	1	1	1
	[...] differential binding affinity of the insect-selective compounds <b>apparently</b> lies in their inability to engage in interactions [...].	4165	1	2	1
	At the same time it <b>may be</b> concluded that a 7,11-d-lactone structure should not be present since the carbonyl group [...].	4166	1	1	1





