Belfield, SJ, Enoch, SJ, Firman, JW, Madden, JC, Schultz, TW and Cronin, MTD

Determination of "Fitness-for-Purpose" of Quantitative Structure-Activity Relationship (QSAR) Models to Predict (Eco-)Toxicological Endpoints for Regulatory Use

Determination of "Fitness-for-Purpose" of Quantitative Structure-Activity Relationship (QSAR) Models to Predict (Eco-)Toxicological Endpoints for Regulatory Use

Samuel J. Belfield[1], Steven J. Enoch[1], James W. Firman[1], Judith C. Madden[1], Terry W. Schultz[2], Mark T.D. Cronin[1]*

[1]School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK

[2]University of Tennessee, College of Veterinary Medicine, Knoxville, TN 37996-4500, USA

*Author for Correspondence:

Mark Cronin

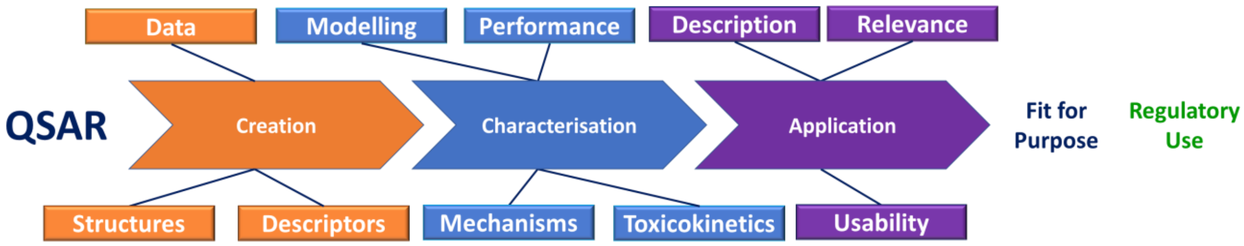
Email: m.t.cronin@ljmu.ac.uk

Abstract

In silico models are used to predict toxicity and molecular properties in chemical safety assessment, gaining widespread regulatory use under a number of legislations globally. This study has rationalised previously published criteria to evaluate quantitative structure-activity relationships (QSARs) in terms of their uncertainty, variability and potential areas of bias, into ten assessment components, or higher level groupings. The components have been mapped onto specific regulatory uses (i.e. data gap filling for risk assessment, classification and labelling, and screening and prioritisation) identifying different levels of uncertainty that may be acceptable for each. Twelve published QSARs were evaluated using the components, such that their potential use could be identified. High uncertainty was commonly observed with the presentation of data, mechanistic interpretability, incorporation of toxicokinetics and the relevance of the data for regulatory purposes. The assessment components help to guide strategies that can be implemented to improve acceptability of QSARs through the reduction of uncertainties. It is anticipated that model developers could apply the assessment components from the model design phase (e.g. through problem formulation) through to their documentation and use. The application of the components provides the possibility to assess QSARs in a meaningful manner and demonstrate their fitness-for-purpose against pre-defined criteria.

Keywords: *In silico* models; QSAR; Toxicity prediction; Uncertainty; Regulatory use

Graphical Abstract



The three phases of QSAR development with related components associated with uncertainty, variability and bias.

Highlights

- Ten components, or groups of assessment criteria, of QSARs are defined
- The components were mapped onto three phases of QSAR development and use
- QSARs assessed using the components with strategies to reduce uncertainty proposed
- Different uses of QSARs require different types of models
- The assessment components demonstrate fitness-for-purpose of QSARs

Abbreviations: log P, logarithm of the octanol-water partition coefficient; MLR, Multiple linear regression; N/A, not applicable; QMRF, QSAR Model Reporting Format; QPRF, QSAR Prediction Reporting Format; QSARs, quantitative structure-activity relationships; QSPR, quantitative structure-property relationship; RBFNN, Radial Basis Function Neural Networks.

Introduction

Computational approaches are at the heart of 21st century toxicology and, with the increase in data availability, they are becoming easier to create and utilise. They also offer the possibility of linking new "big" data resources to chemic

al safety assessment and new methods of modelling, e.g. machine learning technologies (Worth, 2020). Modelling data serves many purposes, and in chemical safety assessment much of the focus has been to predict hazard and exposure, with particular applications in product development and regulatory assessment. Other purposes include the interrogation of, and learning from, data, as well as evaluation of (structure-activity) hypotheses. For specific purposes, notably regulatory applications, there are varied uses such as data gap filling, classification and labelling, screening and prioritisation, amongst others. Whilst the number, type and application of models has steadily grown in the past few years, means of their evaluation has not developed at the same pace. At the current time models for chemical safety assessment are evaluated using the same criteria, such as the OECD Principles for the Validation of QSARs (2007), regardless of purpose. However, there is an opportunity to update our way of thinking by considering the purpose of a model, use of new approaches to understand what type of model is appropriate for a particular application and how best to assess model fitness-for-purpose (Patterson and Whelan, 2017; Patterson et al., 2021).

This article focusses on understanding the purpose of and evaluating quantitative structure-activity relationships (QSARs) that can be used to predict toxicity. Broadly speaking, QSAR models define the relationship between factors relating to chemical structure and/or molecular descriptors of a series of chemicals to their properties e.g. activity or toxicity. As such, they offer the possibility of making predictions of toxicity directly from chemical structure or using knowledge derived from similar chemical(s). Many such computational models have been developed; for ecotoxicological endpoints QSARs may be based upon well-established mechanisms of action (Cronin 2006; 2017; Cronin et al., 2002) whilst for human health effects, mechanistically-interpretable models may be less feasible due to the complexity of the endpoints (Madden et al., 2020). It is also noted that the approaches described in this paper could additionally be applied to quantitative structure-property relationships (QSPRs), although this was not the focus of this study.

There are many potential roles for QSARs in toxicology. For the purposes of this investigation the applications are considered to be broadly related to “industrial” or “regulatory” use. Other uses of QSARs include data investigation such as in-house model development (e.g. for preliminary screening of inventories) and education, however, these do not require such rigorous model evaluation. Table 1 summarises some of the main use case scenarios for *in silico* models to predict toxicity, focusing on industrial and regulatory use but also data investigation, knowledge creation and for education. It is acknowledged that this is not a comprehensive list of uses but is illustrative of the range of uses in *in silico* toxicology. In this context, industrial uses may be the development of new substances, as well as the evaluation of existing ones for potential use as ingredients. Regulatory uses of QSARs are in response to legislation and may be undertaken by the registrant, i.e. the manufacturer, as part of a dossier presented to a regulatory agency, or they may be utilised by the governmental (regulatory) agency itself for a variety of purposes. Whilst a complete description of all potential uses of QSARs is beyond the scope of this paper, it is true to say that in some cases broadly applicable models will suffice, whereas for others more localised or bespoke models for a given purpose are required. These differing requirements and applications contrast with the historical culture of a “one size fits all” for QSAR development, with the expectation that one model can serve multiple purposes. This contradiction has been exacerbated by the lack of clarity concerning the requirements to establish the validity of *in silico* model for specific purposes.

Table 1. Potential use case scenarios and characteristics of *in silico* models to predict toxicity

Use	Brief Description	Desirable characteristics of the model	Proposed level of uncertainty in a model and / or prediction considered acceptable
Data Investigation			
Investigation of “small”, or “local” data sets	E.g. analysis of congeneric series to determine mechanisms	Transparent, with a small number of mechanistically relevant descriptors	High
Investigation of “big data” sets	Investigation of chemical space, global QSAR models	Rapid and suitable for machine learning approaches	High
Knowledge and hypothesis generation and testing	Ability to use existing data resources to gain new insight from data e.g. mechanistic understanding	Any model is appropriate up to the investigation of big data using Artificial Intelligence approaches	High
Education, training and capacity building	Any type of modelling for educational and other purposes	Any model is appropriate	High
Development of new approaches	Investigation of data sets, in a comparative manner to illustrate the	Wide range of models applicable	High

	performance of a new modelling approach, descriptors etc.		
Industrial Use			
Screening of lead compounds	Identification of potential toxicity in candidate compounds through the screening of very large inventories	Rapid / automated application. Broad coverage	High
Evaluation or optimisation of a lead compound or ingredient	Assessment of the safety of an individual Ingredient or development of a new compound with improved safety profile	Specific mechanistically based and justified models	Low
Safety/ hazard assessment of a compound in a product	Assessment of the safety of an established or new compound in a product or formulation	Specific mechanistically based and justified models	Low
Regulatory Use			
Prioritisation	Prioritisation of compounds for testing according to legislative needs, e.g. Canadian Domestic Substance List	Rapid / automated application. Broad coverage	High

Classification and Labelling	Identification of hazard to allow for classification, e.g. EU Classification, Labelling and Packaging (CLP) Regulation	Broadly applicable. Capable of rapid hazard characterisation	Moderate
Hazard identification (e.g for risk assessment)	Risk assessment of the safety of a substance, e.g. EU REACH	Specific mechanistically based and justified models. Transparent and well documented	Low

In order to have confidence in the use of a QSAR model, its fitness for the purpose intended must be established. This is especially true where QSAR predictions are used to inform regulatory decisions. Generally speaking, there are three key regulatory uses for QSAR predictions: hazard identification informing risk assessment; classification and labelling; and prioritisation and screening (Cronin et al., 2003). The exact definition and implication of each of these depends on the legislation under which they are implemented. In terms of assessing whether a model is "fit for purpose", there is no method of assessment that is globally applicable, especially in terms of differentiating between the requirements of the different use cases. The most commonly applied approach to determine whether a QSAR can be used for regulatory applications, is to understand whether a model (and hence its predictions) can be considered valid. The OECD Principles for the Validation of (Q)SARs were established as a means to evaluate (Q)SARs (OECD 2007). These have been utilised for almost 15 years and, on the whole, have served the scientific community very well. They have provided a framework by which to evaluate QSAR models for toxicity according to their characterisation through documentation, performance, applicability domain and mechanistic interpretation. They have also formed the basis by which to record requisite information for QSAR models and predictions, such as the QSAR Model Reporting Format (QMRF) and QSAR Prediction Reporting Format (QPRF) respectively, which may be used for regulatory submissions (Worth, 2010).

Whilst the OECD Principles for the Validation of QSARs have been applied widely, various shortcomings have become apparent. The principles were not developed with new statistical methods, such as machine learning, in mind. They are often used to evaluate a QSAR for a specific purpose, rather than assisting in the assessment of the strengths and weaknesses of the model in a particular context. In addition, since their conception, the sciences of toxicology and risk assessment have developed greater appreciation of how uncertainties influence decision making (Thomas et al., 2019). Specifically, the Principles do not assign a particular level of confidence, neither do they address the relevance for a particular purpose, such that may be required for a regulatory application, to demonstrate whether it is fit for a regulatory use. Patlewicz (2020) has raised this as a challenge, relating in part to how informatics will be applied to larger datasets; embracing this challenge we have considered a more holistic approach to evaluating the whole life of a QSAR from its conception to implementation.

In addition, whilst useful, the implementation of the OECD QSAR Principles only provides a binary classification of whether they are met or not for a particular model, the judgement of which, in itself, can be subjective. As such, they are not entirely appropriate for consideration of whether a model is fit for a purpose or, indeed, relevant for a specific application. The situation is made more complex as there is no formal definition of fitness-for-purpose for an *in silico* model. However, a fit-for-purpose model can be taken as one that has been appropriately developed and is transparent, suitably documented and, as required, compliant with the OECD Principles (Cronin et al., 2019). Supplementing this there are proposals for Good Computer Modelling Practice (Judson et al., 2015), proposals for the use of Artificial Intelligence to assist in chemical risk assessment (Wittwehr et al., 2020), as well as protocols for the development of *in silico* models being developed for various toxicological endpoints (Myatt et al., 2018; Hasselgren et al., 2019; Johnson et al., 2020). As well as no formal definition, currently the concept of an *in silico* model being fit-for-purpose is poorly developed. However, it is acknowledged, if seldom explicitly stated, that different levels of confidence are required for different regulatory uses (Dent et al., 2018; Kulkarni et al., 2016; Taylor and Rego Alvarez, 2020). This is easier to consider in terms of the uncertainty associated with a model, for instance, risk assessment where a prediction may provide information to assist in the replacement of an *in vivo* animal test requires low uncertainty, whereas classification may accommodate moderate uncertainty; for screening and prioritisation higher levels of uncertainty may be tolerated. Thus, when considered in terms of relative uncertainty, a model and its predictions may be fit-for-purpose for one application (e.g. prioritisation), but not necessarily for another (e.g. risk assessment).

With the need to better evaluate QSARs for potential regulatory, and other, uses, Cronin et al. (2019) developed a scheme to evaluate the uncertainty, variability and areas of bias of a QSAR model. The purpose of this scheme was not to provide a definitive conclusion as to whether the model was validated or not validated, rather it was to identify areas of uncertainty in a QSAR. Identifying areas of uncertainty enables them to be addressed, either by seeking additional information to reduce the uncertainty, hence increasing confidence (and regulatory applicability) of the model, or ensuring that any residual uncertainty is clearly communicated and use of the QSAR for a given purpose is appropriate. The scheme centred around 49 aspects of a model, broadly focusing on its creation, characterisation and application. The development of criteria for the evaluation of QSARs was informed by recent progress and guidance from IPCS (2014), EFSA (2018) and elsewhere (Sahlin 2013, Pestana et al., 2021). Whilst two exemplar QSAR studies were evaluated using the scheme (Cronin et al., 2019), its full applicability has not yet been demonstrated and this will be required if such an approach could have broad regulatory application. In addition, it may be considered that assessing 49 criteria is both unwieldy and unlikely to provide a succinct evaluation of the key areas of uncertainty in a QSAR. These disadvantages mean that, in the format proposed by Cronin et al. (2019), the scheme is unlikely to provide insight into the characteristics of a QSAR that are required or desirable for a particular purpose.

The aim of this study was, therefore, to demonstrate how the scheme previously reported by Cronin et al. (2019) could be utilised to assess an *in silico* model, such as a QSAR, to determine whether it is fit for a specific purpose. To achieve this the 49 criteria were rationalised into higher level "assessment components" which were subsequently linked to one of the three phases of QSAR development. The assessment components were then mapped onto three potential regulatory uses to determine a) the levels of uncertainty that may be acceptable and b) the possible characteristics of a model for a particular purpose. Finally, 12 QSARs for (eco-)toxicological endpoints, recently published in the open scientific literature, were evaluated according to the assessment criteria to demonstrate the uncertainties within such models and provide strategies so that, in accordance with the assessment components, they could be improved and potential regulatory uses (if required) could be identified.

2. Methods

2.1 Evaluation of the previously published scheme for its potential to assess the fitness-for-purpose of in silico models for regulatory use

The 13 main areas of concern, made up of the 49 criteria in the scheme for the evaluation of QSARs proposed by Cronin et al. (2019), were consolidated into ten distinct assessment components that characterise *in silico* models. Each assessment component (referred to hereon as "components") was aligned to one of the three phases in the development of a QSAR.

2.2 Mapping of the QSAR components onto potential regulatory use

The QSAR components were considered in terms of the acceptable levels of uncertainty, variability or bias that would be appropriate for different regulatory uses. This enabled the QSARs selected to be considered in terms of their potential regulatory applicability, both before and after application of strategies to reduce uncertainty, variability and bias (Sections 2.3 and 2.4). As part of this process, the needs of regulatory uses were considered in the context of what may make the QSARs fit for this purpose.

2.3 Selection and initial assessment of QSAR models to be analysed using the QSAR components

From the outset, it should be appreciated that the purpose of the assessment of published QSARs was not to be critical or attempt to validate a particular model. All models had been published in the scientific literature, will have undergone peer review and it is, therefore, implicit that the models are sufficiently robust. The current investigation was undertaken in order to identify any areas associated with greater uncertainty, variability or potential bias and to propose strategies to reduce these, where appropriate, to ameliorate these issues, such that the models' fitness-for-purpose for regulatory applications could be enhanced. QSAR models were selected for analysis based on the following criteria:

- Available in a peer-reviewed publication published in 2018 or 2019
- Relating to (eco-)toxicity
- Representing a variety of approaches

To identify suitable QSARs, publications were searched for in Web of Science using two keywords "QSAR" and "toxic*" as part of the "topic". The publications for analysis were selected manually. In order to assist in the selection of QSARs, models were pre-screened initially to characterise them in terms of:

- Species
- Protocol (e.g., duration of study, endpoint, etc.)
- Number and type of chemicals (multi-constituent substances were omitted)
- Descriptors included in the QSAR
- Statistical method applied in the QSAR
- Potential mechanistic basis

Twelve publications were chosen to represent QSARs for (eco-)toxicological endpoints with a variety of modelling approaches, chemicals, data set sizes, descriptors and mechanisms of action.

The criteria to evaluate QSARs, as defined by the scheme for the evaluation of uncertainty, variability and areas of bias (Cronin et al., 2019) and summarised in Supplementary Information Table S1, were applied to the QSAR models identified. This was performed by expert analysis of the information provided in the publications associated with the QSARs, as well as other relevant information, e.g. retrieval of source information. Expert analysis was undertaken by a lead researcher, with subsequent verification by another researcher. At the time of undertaking the analysis the developers of the QSARs were not contacted for further information or clarification; if this process is to be more widely applicable it is essential that analysis can be carried out without recourse to model developers

The questions set out within the scheme defined within Cronin et al. (2019) were used to assess each of the QSARs. Responses were reported using a semi-quantitative scale of 1, 2 or 3, (representing low, moderate and high uncertainty respectively) or not applicable (N/A). All scores and associated comments were reported using the templates provided in Cronin et al. (2019).

2.4 Recommendations for strategies to reduce uncertainty, variability and areas of bias of the selected QSARs and identification of possible regulatory use

Potential strategies to reduce areas of significant uncertainty, variability and potential areas of bias of the selected QSARs were proposed. The purpose of the strategies was to provide a structured means to reduce the uncertainty associated with a QSAR.. In certain circumstances, the toxicological data used in the QSARs were re-evaluated from a mechanistic perspective to reduce uncertainty in this component e.g. the inclusion of mechanistically based descriptors, such as the logarithm of the octanol-water partition coefficient (log P) for acute ecotoxicological effects (Könemann, 1981). The levels of uncertainty associated with the components, as well as the characteristics, of the QSARs were compared against those proposed for regulatory purposes in an attempt to identify any regulatory use.

3. Results

3.1 Scheme for "Components of QSARs" on the basis of criteria for reducing uncertainty, variability and bias.

Evaluation of the scheme for assessing *in silico* models published by Cronin et al. (2019) allowed for the establishment of an overview of the types of uncertainty, variability and bias (summarised as "variability" herein) observed across QSAR models; the uncertainty criteria were grouped into components as shown in Figure 1. In this way the components summarise the original assessment criteria into logical groupings that can be used to identify the main characteristics of a QSAR. The ten components represent the main areas required for consideration of fitness-for-purpose of an *in silico* model for toxicity prediction. Each component is associated with one of the three phases of QSAR development - creation, characterisation and application. The components are described in Table 2, with details of the individual uncertainty criteria, represented within each component, being denoted in Supplementary Information Table S1. As well as being functional to evaluate QSARs, they can also be applied to help assess the qualities of a model that may be required for a particular purpose. The components cover all aspects of the creation, characterisation and application of QSAR models, they are designed to be flexible and updateable as required. Certain criteria (Table S1) within the components may not be required for a particular model, depending on the purpose of the model/ endpoint under consideration.

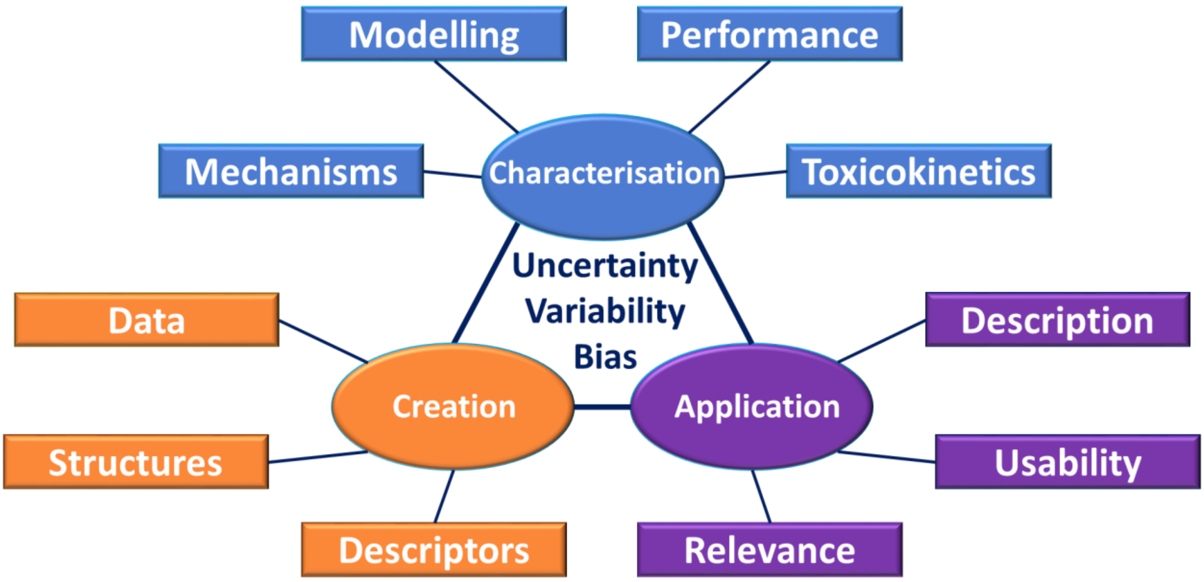


Figure 1. Scheme summarising the ten "components" of QSAR models required to be considered for toxicity prediction purposes. The components, denoted in the rectangular boxes, are linked to the phases, denoted in the oval shapes and defined for each of the three broad areas of QSAR uncertainty, variability and bias.

Table 2. Key features of the proposed ten components for QSARs.

Component	**Key Features Used to Assess the Components**
Model Creation	
1. Data	Quality of individual studies within the data set and the data set overall (e.g. homogeneity of the protocols) that was used for modelling
2. Structures	Accuracy and/ or quality of the reported chemical structures in the training (and, if applicable, test) set used for modelling
3. Descriptors	Appropriate use and adequate definition of the descriptors used for modelling (including how and where sourced)
Model Characterisation	
4. Modelling	The appropriateness and / or adequacy of the modelling approach for the endpoint with regard to complexity of the endpoint and potential use of the model
5. Performance	Adequate statistical fit, predictivity and appropriate reporting
6. Mechanisms	Definition and interpretation of the mechanistic significance of the model to allow for the definition of appropriate domains
7. Toxicokinetics	Appropriate consideration of metabolism and toxicokinetics in the model
Model Application	

8. Description	Appropriate documentation, reporting including applicability domain and transparency of the model and predictions
9. Usability	Implementation of the model; accessibility of required software (e.g. commercial, freely available, sustainable sources)
10. Relevance	Relevance of the model to its intended purpose and use

3.2 Mapping components of QSARs to define fitness-for-purpose for specific regulatory uses

In silico models for toxicity prediction have a number of potential industrial and regulatory uses. Whilst it is acknowledged that certain types of *in silico* model are more suited for some purposes than others, it has not yet been established how the suitability can be qualified in terms of the acceptable level of uncertainty. Using the components of QSARs as an investigative tool provides an opportunity to identify areas of uncertainty, variability or bias that, if reduced, would lead to greater acceptability of the models for a given regulatory purpose.

It is also important to consider which components of an *in silico* model may be associated with higher or differing levels of uncertainty depending on the purpose of the model. In terms of regulatory use, an attempt can be made to identify the different levels of uncertainty in the different components that may be associated with models for different uses. Figure 2 summarises the possible levels of uncertainty that may be associated with different regulatory uses of QSARs to predict toxicity – acceptable levels of uncertainty require discussion and debate before being implemented. Whatever the exact levels of uncertainty required, the lowest would be expected for hazard identification informing risk assessment, with all components expected to show low uncertainty. This would inevitably restrict the use of many types of QSARs for risk assessment and favour those local models based on a clear mechanistic basis with transparency a key factor in the model. As other regulatory uses are considered, going from classification and labelling to screening and prioritisation, greater uncertainty maybe acceptable in terms of being able to develop models that are usable for the purpose intended, i.e. models that can be rapidly applied to large numbers of molecules. In particular, models are likely to be automated for rapid use and have broad chemical coverage across various chemical and mechanistic domains i.e. they are global in nature. As such, it would be unrealistic to expect that the characteristics of these models would all have low uncertainty, e.g. to have a full mechanistic basis due to their inherent difficulty in definition, although mechanisms of action underpinning the model could be proposed. Likewise, less appreciation of toxicokinetics would be expected and greater flexibility in the modelling approach acceptable. It would be expected, however, that the performance of the model would be reported and that it is appropriate for the quality of the data set, regardless of the approach taken for modelling. With regard to the components associated with the application of the model, certain aspects such as description of the model, may be associated with moderate uncertainty for screening and prioritisation i.e. the full definition of a model based on machine learning may not be possible.

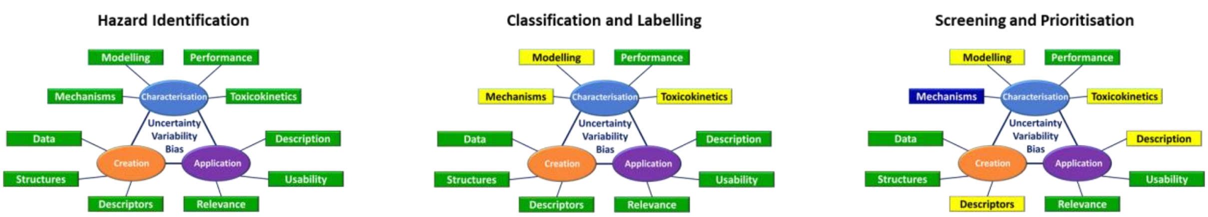


Figure 2. Levels of uncertainty of models and predictions considered acceptable for QSAR components associated with different regulatory uses; green indicates low uncertainty; yellow indicates moderate uncertainty and blue indicates high uncertainty.

3.3 Application of the components and criteria for assessment of published QSARs to assess their fitness-for-purpose

The literature search identified a large number of papers in Web of Science published in 2018-2019 that contained the words “QSAR” and “toxic*” as part of the topic. This represents the full diversity of papers now published in this area, emphasising the importance for proper evaluation. The scope of the papers included a wide spectrum of environmental and human health endpoints as well as methodological papers and opinions. The papers were screened manually using expert judgement to identify twelve publications for analysis in this study. The data sets and modelling techniques from the twelve selected recent publications are summarised in Table 3. They were chosen on the basis of representing a range of both environmental and human-health endpoints. In addition, they were chosen to include representative dataset sizes and methodological variety of QSARs. No inference, positive or negative should be implied by the inclusion or exclusion of QSAR studies in this investigation. Several of the studies implied they were compliant with the OECD QSAR Principles, but no studies stated which specific regulatory, or other, uses they could address. The datasets represent the results of toxicity tests to a variety of aquatic species including an alga, an invertebrate, an amphibian, fish and endpoints relevant to human health. Two publications (#3. de Morais e Silva et al., (2018) and #4. Toropova and Toropov (2018)) analysed the same data set, or parts of it, using different approaches and methods. The data sets generally contained fewer than 100 compounds and were made up of small molecules representative of industrial chemicals, however, some larger datasets were available for human health endpoints comprising drug-like molecules; one dataset was for nanoparticles. Descriptors utilised were mainly calculated directly from molecular structure by the authors of the publications predominantly representing hydrophobicity and electronic properties, as well as topological and steric parameters to a lesser extent. The statistical analyses published ranged from multiple linear regression to partial least squares and neural networks.

Table 3. Summary of QSAR data sets assessed in this study.

Study	Endpoint	Species	Number and type of chemicals	Descriptors included in the QSAR	Statistical method applied in the QSAR	Reference
1	40 hour inhibition of growth	Ciliated protozoan (*Tetrahymena pyriformis*)	160 substituted aromatic compounds	Various calculated properties, e.g. log P and molecular descriptors	Multiple linear regressions (MLR) in comparison to Radial Basis Function Neural Networks (RBFNN)	Luan et al., 2018
2	96 hour LC_{50}	Fathead minnow (*Pimephales promelas*)	15 substituted benzenes	Log P and electrophilicity index and squared terms	Linear regression	Pal et al., 2018
3	Acute aquatic toxicity	Fish (species not defined)	61 compounds associated with non-polar narcosis	Theoretical Volsurf molecular descriptors	Partial Least Squares	de Morais e Silva et al., 2018

4	Acute aquatic toxicity	Fish (species not defined)	111 compounds	CORAL descriptors	Monte Carlo optimisation of target functions	Toropova and Toropov, 2018
5	Inhibition of growth	Tadpoles (*Rana temporaria*)	110 “small” organic molecules	Theoretical molecular descriptors	Multiple linear regression, partial least squares, support vector regression	Wang et al., 2019
6	96-h 20% and 50% inhibitory concentrations, Lowest and No Observed Effect Concentration (LOEC and NOEC)	Alga (*Chlorella vulgaris*)	67 substituted phenols and anilines	Theoretical / molecular orbital descriptors	Multiple linear regression	Yan et al., 2019

7	Hepatotoxicity	Not stated	1,254 “unique” compounds	Topological geometry and physicochemical descriptors	Naïve Bayes, k-nearest neighbour, Kstar, AdaBoostM1, Bagging, decision tree, random forest, and Deeplearning4j	He et al., 2019
8	Reproductive toxicity	Not stated	1,823 organic compounds	Molecular fingerprints	Artificial neural network, C4.5 decision tree, k-nearest neighbour, naïve Bayes, support vector machine, and random forest	Jiang et al., 2018
9	Activity, activity score, potency, and efficacy	Androgen receptor	10,273 drug molecules	Various properties calculated with PaDEL	Random forest, decision tree, neural network, and linear model	Gupta and Rana, 2019

10	50% inhibitory concentration	Oestrogen receptor	55 persistent organic compounds	2D topological based descriptors	Genetic function algorithm	Ibrahim et al., 2019
11	Mutagenic potency logTA100	*Salmonella typhimurium* TA100 strain	48 nitroaromatic compounds	Theoretical and molecular orbital descriptors	Genetic algorithm and multiple linear regression	Hao et al., 2019
12	Cytotoxicity, cell viability (%)	Human breast cancer cell line MCF-7, human fibrosarcoma cell line HT-1080, human liver carcinoma cell line HepG2, human colon carcinoma	8 metal oxide nanoparticles	CORAL descriptors	Monte Carlo optimisation of target functions	Ahmadi, 2020

		cells HT-29, and rat adrenal pheochromocytoma cell line PC-12				

3.4 Strategies to reduce uncertainty, variability and areas of bias of the selected QSARs and identification of possible regulatory use

The evaluation of each model, by application of the assessment criteria, highlights which of the components are associated with higher uncertainty and therefore reduce the suitability of the model for regulatory purposes associated with the most stringent criteria. The results of this analysis are summarised in Figure 3 and described in detail in Supplementary Information Table S2. The overall levels of uncertainty for the 12 QSAR studies provided in Figure 3 are intended to be illustrative, rather than definitive and, as such, they highlight key areas of uncertainty for the different models. Clear areas of high uncertainty can be established across all QSARs, regardless of the endpoint and type of model. For instance, Figure 3 shows that aspects of the biological data, or their description, are associated with high uncertainty. This is a useful finding as it would suggest that no model with high uncertainty for these characteristics would be suitable for any regulatory use (as defined in Figure 2). Further areas routinely associated with high uncertainty are the mechanistic interpretation of the models, incorporation or appreciation of the toxicokinetic properties required to correctly predict toxicity and their relevance for regulatory endpoints. Other criteria associated with higher uncertainty included the unambiguous identification of chemical structures in the model, the overall description of the model such that it could be repeated and its potential usability. Areas where models showed low uncertainty typically were with regard to the description and/ or the availability of descriptors in the model and the stated performance of the model.

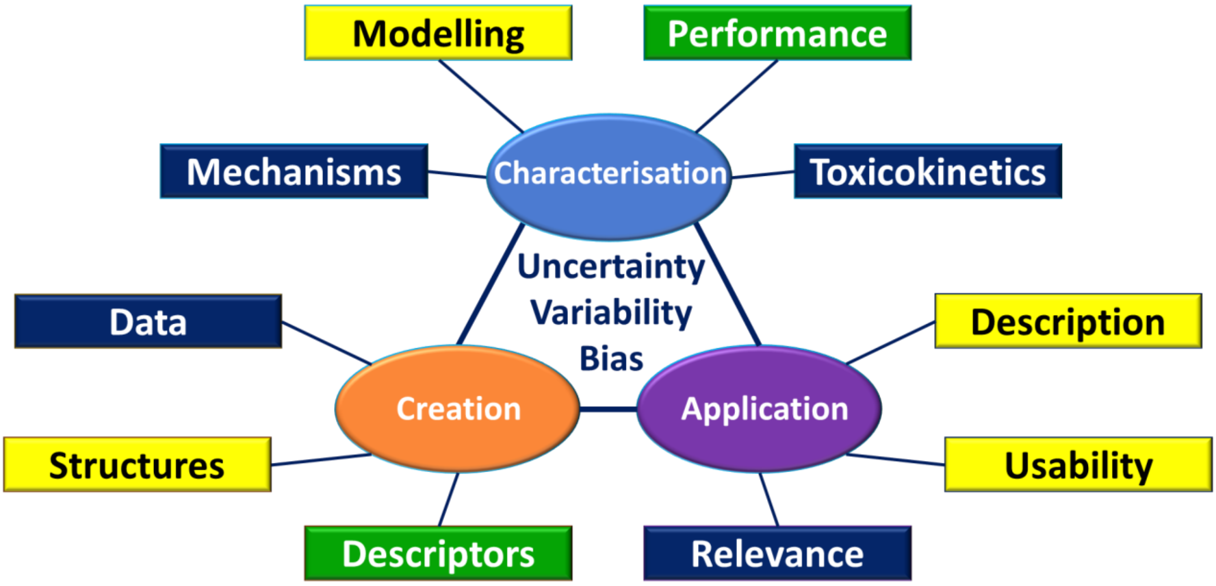


Figure 3. A summary of the levels of uncertainty associated with QSAR components for the 12 QSAR studies evaluated; green indicates low uncertainty for that component, yellow moderate uncertainty and blue high uncertainty. This figure is for illustration only and indicates the median level of uncertainty for these 12 QSAR studies. A full breakdown on the uncertainty associated with each component is provided in Supplementary Information Table S3.

As previously noted, the purpose of the evaluation of uncertainties is not to suggest that a specific model could not be used, but to understand its potential limitations allowing the developer and/ or user to reduce uncertainties. For instance, the uncertainty of many of the areas of QSARs identified as high by the assessment components could be rapidly reduced through the provision of extra information. A summary of the possibilities to enhance the suitability of the models is given in Table 4. Thus, where the description of the biological data was a significant uncertainty, this could be addressed by better reporting in the methods, etc. Likewise, for the incorporation of mechanistic and toxicokinetic information, uncertainty could often be reduced by appropriate discussion and evaluation of the model. In addition, areas of good practice within model development can be highlighted through components with low uncertainty.

Table 4 also describes the potential regulatory use for the QSAR once the uncertainties have been reduced. In order to illustrate this concept, QSAR Study #2 was assessed here as having higher uncertainties in relation to chemical structures description of the data and mechanistic interpretability and usability (component analysis summarised in Table 4). The uncertainty in the published model makes it unsuitable for regulatory use in its current form. However, regulatory suitability could be enhanced by reducing the uncertainty associated with these aspects as described in Supplementary Information Table S4. In terms of the biological data, these are from a well-established data resource, i.e. for the fathead minnow (Russom et al., 2007). The chemical structures can be defined definitively and a full mechanistic interpretation can be applied, i.e. the role of non-polar narcosis. Thus, one possibility is to provide a mechanistic interpretation of the QSAR in terms of how the descriptors relate to the underlying molecular initiating event and, for a well-studied mechanism such as non-polar narcosis, place this model in the context of existing knowledge, e.g. the role of hydrophobicity (Könemann, 1981).

Table 4. The potential suitability for regulatory use before and after implementation of strategies to reduce uncertainties as identified by the components for the 12 QSARs evaluated in this study.

Study	Scope of Model: Local vs Global	Potential Mechanistic Interpretability	Summary of Key Uncertainties in Publication	Key elements of strategy to reduce uncertainty to enhance acceptability	Potential regulatory use of QSAR following enhancements
1	Global	Low	Biological data not described / evaluated. Descriptors not provided. Complex models. Lack of mechanistic interpretation.	Provide details on biological data and descriptor set. Apply mechanistic interpretation (if possible).	Screening

2	Local	High	Biological data not described / evaluated. Descriptors not provided. Complex models. Lack of mechanistic interpretation.	Provide details on biological data. Ensure mechanistic interpretation and context of model reported.	Hazard identification
3	Local	High	Biological data not described / evaluated. Descriptors not provided. Replicate values present in both training and test sets.	Provide details on biological data and descriptor set. Remove duplicates from the training and test sets.	Classification and Labelling
4	Global	Low	Biological data not described / evaluated. Descriptors not provided. Replicate values present in both training and	Provide details on biological data and descriptor set. Remove duplicates from the training and test sets. Apply	Screening

			test sets. Lack of mechanistic interpretation.	mechanistic interpretation (if possible).	
5	Global	Low	Chemical structures not defined. Biological data not described / evaluated. Descriptors not provided. Lack of mechanistic interpretation.	Supplementation of unambiguous chemical structures. Provide details on biological data and descriptor set. Apply mechanistic interpretation.	Screening
6	Local	High	Chemical structures not defined. Biological data not described / evaluated. Lack of mechanistic interpretation.	Supplementation of unambiguous chemical structures. Provide details on biological data. Apply mechanistic interpretation.	Hazard Assessment
7	Global	Low	Biological data not described / evaluated. Descriptors not provided. Models are not	Provide details on biological data and descriptor set. Inclusion of each	Screening

			transparent. Lack of mechanistic interpretation.	models' algorithms. Apply mechanistic interpretation.	
8	Global	Low	Biological data not described / evaluated. Calculated parameters not completely described. Models are not transparent. Lack of mechanistic interpretation.	Provide details on biological data and calculated parameters. Inclusion of each models' algorithms. Apply mechanistic interpretation.	Classification and Labelling
9	Global	High	Chemical structures not defined. Biological data not described / evaluated. Physicochemical properties not provided. Highly	Supplementation of unambiguous chemical structures. Provide details on biological data and physicochemical properties. Balance actives vs inactives in data set. Apply mechanistic interpretation.	Classification and Labelling

			imbalanced data set. Lack of mechanistic interpretation.		
10	Global	High	Biological data not described / evaluated. Descriptors not provided. Descriptor calculation methodology not complete. Lack of mechanistic interpretation.	Provide details on biological data and descriptor set. Fully describe all process employed throughout development. Apply mechanistic interpretation.	Classification and Labelling
11	Local	High	Biological data not described / evaluated. Descriptors not provided. Lack of pharmacokinetic interpretation.	Provide details on biological data and descriptor set. Apply pharmacokinetic interpretation.	Hazard identification and possible support of risk assessment

12	Local	Low	Chemical structures not defined. Biological data not described / evaluated. Descriptors not provided. Lack of mechanistic interpretation.	Describe nanoparticles following ECHA guidance (ECHA 2017). Assess usage of various cell lines for single model. Provide details on biological data and descriptor set. Apply mechanistic interpretation.	Possible Classification and Labelling

4. Discussion

As computational modelling becomes commonplace in toxicology, there is a strong and increasing need to demonstrate the quality, usefulness and fitness for particular purpose of any model. This is amplified by the breadth of models in terms of complexity, endpoints, numbers of compounds and modelling technique. The aim of this study was to gain a greater understanding of fitness-for-purpose of *in silico* models for regulatory adoption, and how this could be assessed. The scheme, described herein, was evaluated for its applicability to models for ecotoxicity and human health effects – although it is noted from the outset that these models did not claim any specific regulatory use. The analysis showed that the scheme was widely applicable, flexible and could be applied to different types of models, species, endpoints and chemical space coverage. Using the criteria noted above, it was possible to determine which aspects of the models were associated with the greatest uncertainties, variability and potential for bias and how all of these could be reduced. This does not constitute a formal validation process, but does provide information on how to assess the applicability, utility and potential for constructive modification of a particular model.

4.1 "Components" of QSARs as the means to assess and reduce uncertainty, variability and bias.

Analysis of the criteria in the scheme for the evaluation of QSARs proposed by Cronin et al. (2019) allowed for the identification of ten components as summarised in Figure 1 and summarised in Table 2. The components have rationalised the 49 original criteria into fundamental properties of an *in silico* model that will allow (semi-)quantification of uncertainty. The components are designed to be flexible and, as such, applicable to any type of model from a simple QSAR with a small number of components up to machine learning approaches based on large datasets. The components address all aspects of the three phases - creation, characterisation and application of an *in silico* model and allowed for uncertainty to be assigned to them.

The consolidation of the original 49 criteria described by Cronin et al. (2019) into the general ten assessment components provides a much clearer and comprehensible overview of the uncertainty in an individual QSAR (as shown in Figure 1). It is anticipated that this type of analysis will have at least two clear uses, as described below: a better understanding of the characteristics of a model for a particular purpose (here illustrated with reference to regulatory application); and for the assessment of an individual model from the problem formulation statement through to its application.

4.2 Understanding fitness-for-purpose of QSARs for specific regulatory uses with the components

The rationale behind of the creation of the components was to enable identification of areas of uncertainty such that uncertainty could be reduced to a level that would allow a model to be considered “fit-for-purpose”. One of the most demanding and pressing uses of a model is for regulatory application, thus fitness-for-purpose was evaluated for different regulatory uses. Figure 2 gives an indication of the levels of uncertainty that may be associated with a particular regulatory use. In addition to these, unspecified applications could also be assessed in the same manner through considered adjustment of the uncertainty requirements in particular areas. For instance, using a QSAR to investigate a data set to generate hypothesis or gain mechanistic insight may allow for higher uncertainty in many areas e.g. performance may indeed not require any consideration of the Application-characteristics of the QSAR, as it would not be used for a particular predictive or regulatory purpose.

Analysis of Figure 2 demonstrates the levels of uncertainty, variability and bias that may be acceptable for a particular regulatory purpose. From the trichrome components of screening and prioritisation through the dichrome components of classification and labelling to the monochrome components of risk assessment, several aspects become apparent. Firstly, both the Creation and Application phases allow no high uncertainty, whilst only moderate uncertainty is permitted with regard to the descriptors used, documentation, transparency etc. of the model. To accomplish this, there should be a defined data set of high quality in terms of the description of chemical structures, biological data and descriptors, all of which must be unambiguous in any model, even if not completely transparent, regardless of the purpose (Young et al., 2008; Piir et al., 2018). Often, the uncertainty associated with these two components can be reduced with additional clarification although the relevance of the endpoint to the stated purpose is definitive. Secondly, the greatest acceptability of variability and bias is associated with the Characterisation phase of a QSAR. Flexibility, and an increase in uncertainty, is likely in the characterisation stage of modelling, most notably mechanistic interpretation which relates to all types of *in silico* models. While the performance component requires low uncertainty regardless of the purpose, the acceptable uncertainty of the other three Characteristics-related components are fit-for-purpose dependent. In the case of Mechanisms, Modelling and/or Toxicokinetics it is typically not possible to move to a more demanding fit-for-purpose application, i.e. reduce the uncertainty, without reverting to the Creation phase – essentially starting the development of a model again.

Fundamentally, uses for *in silico* toxicology range from the need for the rapid screening of large inventories of chemical structures to detailed hazard identification of a single substance. Screening may require assessing structurally diverse inventories in the 10-100,000s or millions of compounds; in contrast, a detailed analysis of a single compound may only require assessing 10 or fewer highly similar substances. It is intuitive that the needs for the different types of applications will be different and thus, should be considered. When screening a large chemical inventory, a rapid automated approach is ideal and approaches using machine learning, with automated data entry, prediction and analyses being required. More detailed risk assessment of a single substance will require a detailed and mechanistically derived model, such as a local, transparent QSAR based on a small number of mechanistically interpretable descriptors. The use of highly localised models also explains the high level of use for read-across for risk assessment (ECHA, 2020), whereas it finds little application for screening and prioritisation.

In terms of acceptable uncertainties, it can be proposed that there are different levels of uncertainties that might be considered as being acceptable, dependent on the potential consequence of an inaccurate prediction. For instance, it could be possible that a model based around a machine learning method, optimised to identify toxic molecules, could be acceptable with a relatively high false positive rate if it were to be used in the screening of chemical inventories for lead identification. Such a scenario may allow for relatively high uncertainty to be associated with a model, on the proviso that it is fit for its stated purpose. At the other end of the regulatory use spectrum, risk assessment requires demonstrably low uncertainty in the *in silico* approach, which is likely to be characterised only by mechanistic models based on limited chemical domains, e.g. a defined chemical class or mechanism of action, and is thus associated with the relatively high uptake and success of using read-across for toxicity prediction (ECHA, 2020).

Figure 4 demonstrates how a data resource could be utilised according to the needs of regulatory use. Taking as an example a relatively large data source, such as may be extracted from a regulatory inventory or the ChEMBL database (https://www.ebi.ac.uk/chembl/), it is assumed that there would be a process of data curation to ensure the quality of chemical structures and biological data is high, i.e. low uncertainty. Following this, it is probable that initial analyses would be rapid and use machine learning approaches, possibly with many descriptors. The machine learning approaches should provide an indication of the feasibility of modelling the data and any inconsistencies in the data matrix, if they have not already been identified through the data curation. It is likely that there will be high uncertainties at this stage, especially in aspects such as mechanistic understanding and interpretation. Such models would be global in nature and thus, suited only to screening and prioritisation.

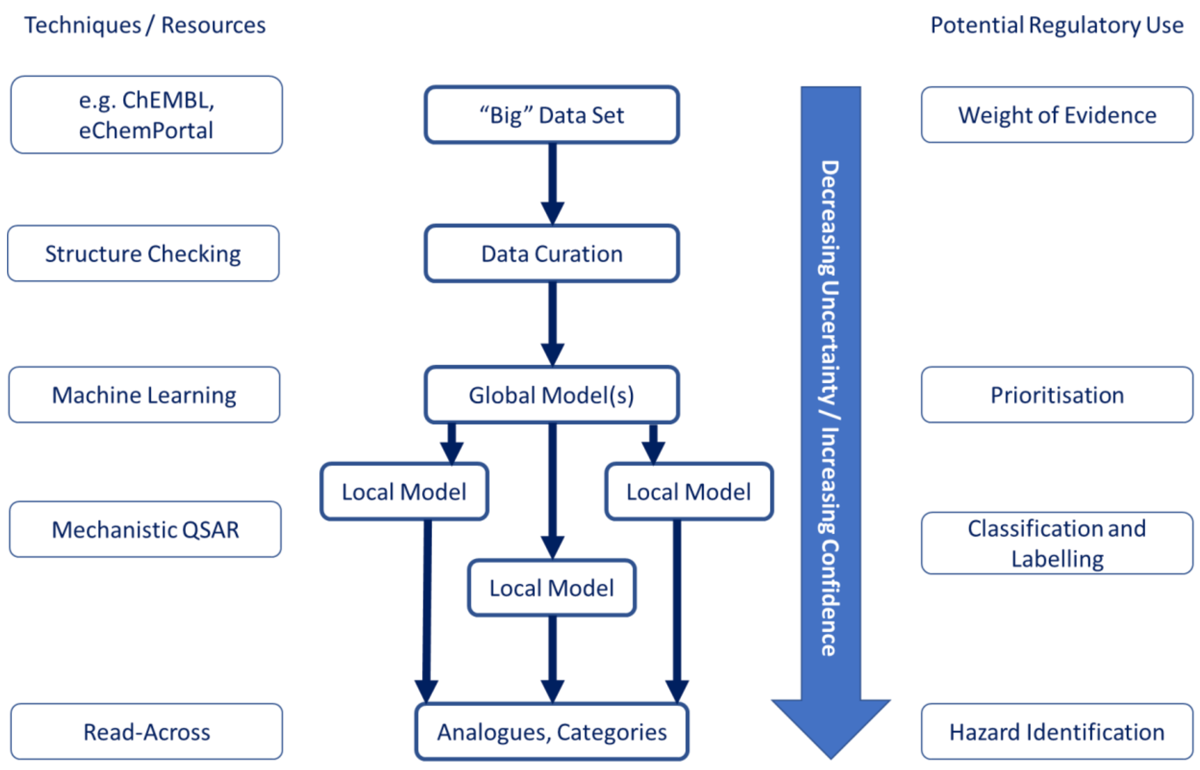


Figure 4. Potential regulatory use of different types of QSARs and *in silico* models that could be derived from a "big" data set. Models range from global machine learning to read-across from close analogues.

Subsequent analysis of the complete data set would allow for consideration of chemical space and identification of structurally-limited areas, or chemical classes, that are well populated. Therefore enabling the construction of models with reduced uncertainty in the components of Descriptors, Mechanisms and Description (see Figure 2) that are suitable for the purpose of classification and labelling. Continuous development may also lead to models deemed sufficient for hazard assessment, potentially informing risk assessment. Even within these class- or mechanism-based QSARs further refinement could be achieved to identify one, or a small number, of analogues that may be suitable for read-across or trend analysis (Date et al., 2020). Such high quality, mechanistically derived analogues can be considered to be of low uncertainty and thus useful for risk assessment.

4.3 Application of the components and criteria for assessment of published QSARs to assess their fitness-for-purpose

The assessment of the 12 QSARs selected using the components demonstrated that the criteria can be applied to a wide variety of models. The full analysis of individual QSARs (Table S2) is overwhelming, such that the use of the components to gain an overview is valuable. Also illustrative is the summary of the uncertainties across all the QSARs analysed (Figure 3). This shows consistently high levels of uncertainty associated with four of the components, namely Data, Mechanisms, Toxicokinetics and Relevance. Whilst it is recognised that the QSARs assessed may not have been developed for purpose of regulatory use, it is informative to consider them in more detail to investigate to which purpose they could be applied (Table 4) and what measures may be required to achieve this (Section 4.3 and Table S4). Comparison of the summary of results in Table 3 with the suggested levels of acceptable uncertainty for different purposes clearly shows that none would be acceptable for these purposes as they are currently presented.

As noted above, full data curation is likely to be a pre-requisite for any regulatory use of a model. Without knowledge of the data, transparency of the model cannot be demonstrated and, more importantly, the domain of a model cannot be defined. More difficult to define is the mechanistic basis. There is a long-appreciated spectrum of models from purely mechanistic to statistical based, i.e. localised QSARs to machine learning (Enoch et al., 2008). As models become global in their applicability, this will require larger datasets with more and varied compounds. Accompanying this complexity in chemistry is the increased likelihood of multiplicity of probable and plausible mechanisms of action. The types of approaches capable of modelling such datasets often use many descriptors, typically without direct mechanistic interpretation. The compromise between the need for mechanistic interpretability and practical tools for largescale screening of compounds means that higher uncertainty, in terms of defining mechanisms, will need to be acceptable. There will also be greater uncertainty associated with assignment of mechanisms of action to chemicals, and this will need to be accepted. Taking acute environmental toxicity as an example, in reality it is very difficult to associate a mechanism of action definitively with a chemical. Historical attempts were made for a relatively small number of chemicals (approximately 40) using Fish Acute Toxicity Syndromes (McKim et al., 1987). These learnings have been extrapolated up to the full spectrum of industrial chemicals and, along with a variety of other evidence, are routinely used to categorise chemicals, for instance for the application of QSARs (Cronin, 2017). Until omics responses to support grouping are robust and understood, there is likely to be on-going uncertainty in the assignment of mechanisms of action for environmental effects. Mechanisms relating to human health effects also vary widely in their level of fundamental understanding, assignment to specific chemicals and relationship to chemistry. Whilst it is a gross oversimplification, it is true to say that regulatory endpoints such as skin sensitisation have a higher degree of mechanistic understanding than, for instance, chronic toxicity. Thus, with regard to modelling and QSARs in particular, we are better able to assign a compound to a mechanistic domain associated with skin sensitisation than we are able to define many mechanisms of organ level toxicity associated with chronic toxicity. Again, until we have a better grasp of using omics data and applying knowledge from Adverse Outcome Pathways, this uncertainty at the mechanistic level is likely to remain (Brockmeier et al., 2017; Cronin et al., 2017).

Toxicokinetics, in other words the appreciation of ADME properties affecting bioavailablity, is also very difficult to address in *in silic*o modelling of toxicity. The toxicokinetics are normally part of the experimental data and would be provided as such, for instance whether there is significant metabolism of a compound, if this is consistent across the training set and if it is defined e.g. such that it can be assumed in an untested molecule for which a prediction is made. Toxicokinetics have also been shown to be an area of uncertainty in read-across (Schultz and Cronin, 2017). There is no easy solution to this issue, other than to acknowledge it as a significant area of uncertainty.

Relevance of an endpoint, and hence prediction, although often overlooked by modellers, is vital for regulatory application. In order for a prediction from a model to be relevant it must address the endpoint of interest. From the outset it would be good practice for the modeller to identify the purpose of the model and undergo a suitable process of the problem formulation. As part of the problem formulation, an objective assessment of the level of acceptable uncertainty should be set out. For instance, if the purpose of the model was to provide predictions for a particular legislation, then the model should be capable of predicting a relevant endpoint. It should be noted that most relevant endpoints for regulatory use, with the exception of creating a Weight of Evidence, are OECD Test Guideline studies. Thus, a model would be fully relevant (and have low certainty) if it made a direct prediction of the relevant OECD Test Guideline Study. In terms of the QSARs investigated in this study, QSAR #7 (hepatotoxicity) may provide support to an overall decision on chronic toxicity, but is not a direct prediction of that endpoint and further information would be required e.g. for other organ level effects; QSAR #8 (reproductive toxicity) would not be sufficient to fill a data gap as it is not defined sufficiently; QSARs #9 and #10 (androgen and oestrogen receptor binding respectively) may support a decision on reproductive toxicity and / or endocrine disruption etc., but they do not replace the need for further information on this endpoint. QSAR #11 is for a regulatory endpoint (*Salmonella typhimurium* TA100), however as only a single strain it would not meet the requirements for *in vitro* mutagenicity which require, usually, five strains to be considered.

4.4 Reducing uncertainty of QSARs using the assessment components

Assessment of QSAR models in the described manner above provides an interesting insight into areas where model developers may wish to concentrate their efforts. For all of the QSARs considered, uncertainty could be reduced by easy to implement strategies (Table S4). For instance, there were a number of issues with the provenance of biological data utilised in the QSARs including: 1) a lack of clarity over the exact description of the data (i.e. protocols) that were utilised, 2) selection of small data sets from larger data compilations without full explanation, 3) a lack of assessment of the quality of the toxicity data utilised, 4) not assessing the relevance of data for regulatory purpose, as well as other related issues. All of these issues can be addressed easily in the QSARs assessed to an appropriate level to improve possible acceptance of the models.

The scheme also highlighted issues relating to the component “Mechanisms”. While the correct identification of mechanism of action of a chemical and its associated applicability domain is the aim of this component, the reality is QSARs often deal with, at best, probable or plausible toxic mechanistic information. The level of mechanistic understanding needed to attain low uncertainty is often endpoint-specific and may vary with the experience, and even opinion, of the model developer. As noted above, there is also the current lack of knowledge of many mechanisms of toxic action – across species and effects – so pragmatism in model development and evaluation may be required in order to reduce the uncertainty associated with this component.

It proves more difficult to reduce uncertainty relating to the toxicokinetics component. However, strategies could be put in place to determine whether metabolism is relevant – a good example, for instance, being with the metabolic component of the Ames Test model (QSAR #11). Relevance to regulatory endpoints is intrinsic to the endpoint and, obviously, cannot be changed. The analysis also highlighted the complexity of some models in comparison to the data being modelled, e.g. the use of highly multivariate statistical analysis to model relatively simple mechanisms of action. Thus models could, in theory at least, be simplified to reduce this uncertainty (as demonstrated in Table S4).

Many issues with uncertainty will be overcome through adequate problem formulation in the development of a QSAR. The statement of problem formulation could be based around defined uncertainty criteria for the QSAR components, such that good modelling can be achieved from the outset. This will allow models to be designed, through the proper problem formulation, to be fit-for-purpose even before they are created. For instance, a modeller can apply the QSAR components to understand the characteristics of the model to be built e.g. the relevance and quality of the data, mechanistic understanding, coverage of descriptors etc. This should not be an onerous process, however, it is one that can be completed before model creation. In this regard, the QSAR developer could incorporate this information easily into the documentation associated with the model. In this way, the model will be assured of appropriate levels of uncertainty relating to purpose for these components. For existing QSARs, models would need to be assessed against the criteria, whether by the developer or user to demonstrate fitness-for-purpose. Overall, the opportunity is for the modeller and user to investigate and hence define the relevance of a particular model for regulatory use as part of the development process.

4.5 Using the components to improve acceptability of QSARs

A fundamental aim of a QSAR is to provide a meaningful, relevant and robust *in silico* model that is fit-for-purpose. Table 1 indicates some of the uses of models, ranging from data investigation and knowledge generation, demonstration of new techniques or descriptors to specific use in industry or regulation. The use of a model could be considered against the requirements of a model to meet a particular purpose. As the spectrum of models increases, from the analogue approach to high level, multidimensional representations of big data, it is important to appreciate that few models are suitable for more than one purpose. Thus, there is a place for all types of models and a means is required to determine whether it is suitable for the purpose proposed (Richarz, 2020).

If the purpose is for regulatory use, the QSAR must provide predictions that are acceptable according to predefined (often legislative rather than scientific) criteria. With regard to data gap filling, the most stringent criteria for the acceptable replacement of an animal test are likely to be required (shown as Risk Assessment in Figure 2). Due to the many uncertainties that may be present in a QSAR – as demonstrated in the analyses in this study – it has been increasingly difficult to gain acceptance of QSAR predictions and more fundamental and justifiable approaches, such as read-across, have been applied more commonly (ECHA, 2020).

The application of the component scheme described in the study allowed for a better understanding of the requirements for different types of regulatory use of QSAR, demonstrated a realistic assessment of QSAR models, provided strategies for their improvement, and is a means of providing evidence to the user of good model development. Future use of such components is foreseen from the very first stages of model design and data harvesting, through to the documentation of the final model.

It is foreseen that the application of such criteria will not replace the use of OECD Principles, but will supplement the information and should be used hand-in-hand with reporting formats such as the QMRF and QPRF.

5. Conclusions

Ten assessment components have been described in this study which are designed to assess uncertainties, but also variabilities and areas of bias of QSARs. These components rationalise and organise the larger number of criteria on which they are based. The ten components summarise the three key phases of *in silico* modelling – creation, characterisation and application. These components have been used to demonstrate and, to a certain extent, semi-quantify the key characteristics of uncertainty that are required for different regulatory purposes, and that different types of models should be applied for different purposes.

As a proof of concept, the components were applied to twelve recently published QSAR studies for various (eco-)toxicological endpoints. The purpose was to identify areas of potential uncertainty, variability or bias that may reduce a QSAR's applicability in a regulatory context. For the QSARs considered, most uncertainties centred around four factors: 1) the quality and / or reproducibility of the toxicity data modelled, 2) transparency of the descriptors and the model, 3) the consideration of mechanisms of action and toxicokinetics and 4) relevance for regulatory use. The analysis of the 12 QSARs demonstrated that they provide a means to assess uncertainty, identifying areas where strategies can be implemented to reduce uncertainty to an acceptable level. It is anticipated that this form of assessment could be initiated at the problem formulation stage of QSAR development to ensure the model is fit-for-purpose. In this way, the scheme provided a usable, practical and flexible means of evaluating a QSAR that extends the OECD Principles. .

Acknowledgement

This work benefitted greatly from the comments of Dr Andrea-Nicole Richarz (formerly of the European Commission Joint Research Centre (JRC), Ispra, Italy and now at the European Chemicals Agency, Helsinki, Finland) who, whilst at the JRC, was a co-author on the original uncertainty papers. SB is grateful to the Faculty of Science, Liverpool John Moores University, for a Scholarship to fund his doctoral studies.

Declaration of Interest

The authors declare no conflicts of interest.

References

Ahmadi, S., 2020. Mathematical modeling of cytotoxicity of metal oxide nanoparticles using the index of ideality correlation criteria. Chemosphere, 242, e125192.

Brockmeier, E.K., Hodges, G., Hutchinson, T.H., Butler, E., Hecker, M., Tollefsen, K.E., Garcia-Reyero, N., Kille, P., Becker, D., Chipman, K., Colbourne, J., Collette, T.W., Cossins, A., Cronin, M., Graystock, P., Gutsell, S., Knapen, D., Katsiadaki, I., Lange, A., Marshall, S., Owen, S.F., Perkins, E.J., Plaistow, S., Schroeder, A., Taylor, D., Viant, M., Ankley, G., Falciani F., 2017. The role of omics in the application of Adverse Outcome Pathways for chemical risk assessment. Toxicol. Sci. 158, 252–262.

Cronin, M.T.D., 2006. The role of hydrophobicity in toxicity prediction. Curr. Comput-Aided Drug Des. 2, 405-413.

Cronin, M.T.D., 2017. (Q)SARs to predict environmental toxicities: current status and future needs. Environ. Sci.-Proc. Imp. 19, 213-220.

Cronin, M.T.D., Dearden, J.C., Duffy, J.C., Edwards, R., Manga, N., Worth, A.P., Worgan, A.D.P., 2002. The importance of hydrophobicity and electrophilicity descriptors in mechanistically-based QSARs for toxicological endpoints. SAR QSAR Environ. Res. 13, 167-176.

Cronin, M.T.D., Enoch, S.J., Mellor, C.L., Przybylak, K.R., Richarz, A.-N., Madden, J.C., 2017. *In silico* prediction of organ level toxicity: Linking chemistry to adverse effects. Toxicol. Res. 33, 173-182.

Cronin, M.T.D., Jaworska, J.S., Walker, J.D., Comber, M.H.I., Watts, C.D., A.P., 2003. Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. Environ. Health. Perspect. 111, 1391-1401.

Cronin, M.T.D., Richarz, A.-N., Schultz, T.W., 2019. Identification and description of the uncertainty, variability, bias and influence in quantitative structure-activity relationships (QSARs) for toxicity prediction. Regul. Toxicol. Pharmacol. 106, 90-104.

Date, M.S., O'Brien, D., Botelho, D.J., Schultz, T.W., Liebler, D.C., Penning, T.M. and Salvito, D.T., 2020. Clustering a chemical inventory for safety assessment of fragrance ingredients: Identifying read-across analogs to address data gaps. Chem. Res. Toxicol. 33, 1709-1718.

de Morais e Silva, L., Feitosa Alves, M., Scotti, M., Silva Lopes, W., Tullius Scotti, M., 2018. Predictive ecotoxicity of MoA 1 of organic chemicals using *in silico* approaches. Ecotoxicol. Environ. Saf. 153, 151-159.

Dent, M., Teixeira Amaral, R., Amores Da Silva, P., 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., Vieira O'Reilly Cabral Posada N., Weiss, C., Kojima, H., 2018. Principles underpinning the use of new methodologies in the risk assessment of cosmetic ingredients. Comput. Toxicol. 7, 20-26,

ECHA. 2017. Appendix R7-1 for nanomaterials applicable to Chapter R7a (Endpoint specific guidance). Available at: https://echa.europa.eu/documents/10162/13632/appendix_r7a_nanomaterials_en.pdf

ECHA. 2020. The use of alternatives to testing on animals for the REACH Regulation fourth report (2020) under Article 117(3) of the REACH Regulation, ECHA-20-R-08-EN Cat. Number: ED-03-20-352-EN-N, ISBN: 978-92-9481-594-1.

EFSA (European Food Safety Authority) Scientific Committee, Benford, D., et al., 2018. Guidance on uncertainty analysis in scientific assessments. EFSA J. 16, 5123, pp. 39 https://doi.org/10.2903/j.efsa.2018.5123

Enoch, S.J., Cronin, M.T.D., Schultz, T.W., Madden, J.C., 2008. An evaluation of global QSAR models for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*. Chemosphere 71, 1225-1232.

Gupta, V.K., Rana, P.S., 2019. Toxicity prediction of small drug molecules of androgen receptor using multilevel ensemble model. J. Bioinf. Comput. Biol. 17, 1950033.

Hao, Y., Sun, G., Fan, T., Sun, X., Liu, Y., Zhang, N., Zhao, L., Zhong, R., Peng, Y., 2019. Prediction on the mutagenicity of nitroaromatic compounds using quantum chemistry descriptors based QSAR and machine learning derived classification methods. Ecotoxicol. Environ. Saf. 186, 109822.

Hasselgren, C., Ahlberg, E., Akahori, Y., Amberg, A., Anger, L.T., Atienzar, F., Auerbach, S., Beilke, L., Bellion,P., Benigni, R., Bercu, J., Booth, E.D., Bower, D., Brigo, A., Cammerer, Z., Cronin, M.T.D, Crooks, I., Cross, K.P., Custer, l., Dobo, K., Doktorova, T., Faulkner, D., Ford, K.A., Fortin, M.C., Frericks, M., Gad-McDonald, S.E., Gellatly, N., Gerets, H., Gervais, V., Glowienke, S., Van Gompel, J., Harvey, J.S., Hillegass, J., Honma, M., Hsieh, J.-H., Hsu, C.-W., Barton-Maclaren, T.S., Johnson, C., Jolly, R., Jones, D., Kemper, R., Kenyon, M.O., Kruhlak, N.L., Kulkarni, S.A., Kümmerer, K., Leavitt, P., Masten, S., Miller, S., Moudgal, C., Muster, W., Paulino, A., Lo Piparo, E., Powley, M., Quigley, D.P., Reddy, M.V., Richarz, A.-N., Schilter, B., Snyder, R.D., Stavitskaya, L., Stidl, R., Szabo, D.T., Teasdale, A., Tice, R.R., Trejo-Martin, A., Vuorinen, A., Wall, B.A., Watts, P., White, A.T., Wichard, J., Witt, K.L., Woolley, A., Woolley, D., Zwickl, C., Myatt, G.J., 2019. Genetic toxicology *in silico* protocol. Regul. Toxicol. Pharmacol. 107, e104403.

He, S., Ye, T., Wang, R., Zhang, C., Zhang, X., Sun, G., Sun, X., 2019. An *in silico* model for predicting drug-induced hepatotoxicity. Int. J. Mol. Sci. 20, e1897.

Ibrahim, I.T., Uzairu, A., Sagagi, B. 2019., QSAR, molecular docking approach on the estrogenic activites of persistent organic pollutants using quantum chemical disruptors. SN Appl. Sci. 1, e1599.

IPCS, International Programme on Chemical Safety (2014) Guidance document on evaluating and expressing uncertainty in hazard characterization. Geneva: World Health Organization, International Programme on Chemical Safety (Harmonization Project Document No. 11) Available from: http://www.inchem.org/documents/harmproj/harmproj/harmproj11.pdf

Jiang, C., Yang, H., Di, P., Li, W., Tang, Y., Liu, G., 2019. *In silico* prediction of chemical reproductive toxicity using machine learning. J. Appl. Toxicol. 39, 844-854.

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, Fearon, I.M., Fehr, M., Gad, S.C., Gervais, V., Giddings, A., Glowienke, S., Hardy, B., Hasselgren, C., Hillegass, J., Jolly, R., Krupp, E., Lomnitski, L., Magby, J., Mestres, J., Milchak, L., Miller, S., Muster, W., Neilson, L., Parakhia, R., Parenty, A., Parris, P., Paulino, A., Paulino, A.T., Roberts, D.W., Schlecker, H., Stidl, R., Suarez-Rodrigez, D., Szabo, D.T., Tice, R.R., Urbisch, D., Vuorinen, A., Wall, B., Weiler, T., White, A.T., Whritenour, J., Wichard, J., Woolley, D., Zwickl, C., Myatt, G.J., 2020. Skin sensitization *in silico* protocol. Regul. Toxicol. Pharmacol. 2020, 116, e104688.

Judson, P.N., Barber, C., Canipa, S.J., Poignant, G., Williams, R., 2015. Establishing Good Computer Modelling Practice (GCMP) in the prediction of chemical toxicity. Mol. Inform. 34, 276-283.

Könemann, H., 1981. Quantitative Structure-Activity Relationships in fish toxicity studies. 1. Relationship for 50 Industrial pollutants. Toxicology 19, 209-221.

Kulkarni, S A., Benfenati, E., Barton-Maclaren, T.S. 2016. Improving confidence in (Q)SAR predictions under Canada's Chemicals Management Plan – a chemical space approach. SAR QSAR Environ. Res., 27, 851-863.

Luan, F., Wang, T., Tang, L., Zhang, S., Dias Soeiro Cordeiro, N.M., 2018. Estimation of the toxicity of different substituted aromatic compounds to the aquatic ciliate *Tetrahymena pyriformis* by QSAR approach. Molecules 23, e1002.

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. ATLA. 48, 146-172.

McKim, J.M., Bradbury, S.P., Niemi, G.J., 1987. Fish acute toxicity syndromes and their use in the QSAR approach to hazard assessment. Environ. Health Perspect. 71, 171-186.

Myatt, G.J., Ahlberg, E., Akahori, Y., Allen, D., Amberg, A., Anger, L.T., Aptula, A., Auerbach, S., Beilke, L., Bellion, P., Benigni, R., Bercu, J., Booth, E.D., Bower, D., Brigo, A., Burden, N., Cammerer, Z., Cronin, M.T.D., Cross, K.P., Custer, L., Dettwiler, M., Dobo, K., Ford, K.A., Fortin, M.C., Gad-McDonald, S.E., Gellatly, N., Gervais, V., Glover, K.P., Glowienke, S., Van Gompel, J., Gutsell, S., Hardy, B., Harvey, J.S., Hillegass, J., Honma, M., Hsieh, J.-H., Hsu, C.-W., Hughes, K., Johnson, C., Jolly, R., Jones, D., Kemper, R., Kenyon, M.O., Kim, M.T., Kruhlak, N.L., Kulkarni, S.A., Kümmerer, K., Leavitt, P., Majer, B., Masten, S., Miller, S., Moser, J., Mumtaz, M., Muster, W., Neilson, L., Oprea, T.I., Patlewicz, G., Paulino, A., Lo Piparo, E., Powley, M., Quigley, D.P., Reddy, M.V., Richarz, A.-N., Ruiz, P., Schilter, B., Serafimova, R., Simpson, W., Stavitskaya, L., Stidl, R., Suarez-Rodriguez, D., Szabo, D.T., Teasdale, A., Trejo-Martin, A., Valentin, J.-P., Vuorinen, A., Wall, B.A., Watts, P., White, A.T., Wichard, J., Witt, K.L., Woolley, A., Woolley, D., Zwickl, C., Hasselgren, C., 2018. *In silico* toxicology protocols. Regul. Toxicol. Pharmacol. 96, 1-17

OECD (Organisation for Economic Cooperation and Development), 2007. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships. ENV/JM/MONO(2007)2. OECD, Paris, pp. 154.

Pal, R., Jana, G., Sural, S., Chattaraj, P.K., 2018. Hydrophobicity versus electrophilicity: A new protocol toward quantitative structure–toxicity relationship. Chem. Biol. Drug Des. 93: 1083– 1095.

Patlewicz, G., 2020. Navigating the minefield of computational toxicology and informatics: Looking back and charting a new horizon. Front. Toxicol. 2, 2. DOI=10.3389/ftox.2020.00002

Patterson, E.A., Whelan, M.P., 2017. A framework to establish credibility of computational models in biology. Prog. Biophys. Mol. Biol. 129, 13-19.

Patterson, E.A., Whelan, M.P., Worth, A.P. 2021. The role of validation in establishing the scientific credibility of predictive toxicology approaches intended for regulatory application. Comp. Toxicol. 17, e100144.

Pestana, C.B, Firman, J.W., Cronin, M.T.D. 2021. Incorporating lines of evidence from New Approach Methodologies (NAMs) to reduce uncertainties in a category based read-across: A case study for repeated dose toxicity. Regul. Toxicol. Pharmacol. *accepted*

Piir, G., Kahn, I., García-Sosa, 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. Environ. Health Persp. 126, e126001.

Richarz, A.-N., 2020. Big data in predictive toxicology: Challenges, opportunities and perspectives. In Neagu, D., Richarz, A.-N. (Eds.). Big Data in Predictive Toxicology. Royal Society of Chemistry, Cambridge, UK, pp. 1-37.

Russom, C.L., Bradbury, S.P., Broderius, S.J., Hammermeister, D.E., Drummond, R.A., 1997. Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (*Pimephales promelas*). Environ. Toxicol. Chem. 16, 948-967.

Sahlin, U., 2013. Uncertainty in QSAR predictions. ATLA 41, 111-125.

Schultz, T.W., Cronin, M.T.D., 2017. Lessons learned from read-across case studies for repeated-dose toxicity. Regul. Toxicol. Pharmacol. 88, 185-191.

Taylor, K., Rego Alvarez, L., 2020. Regulatory drivers in the last 20 years towards the use of in silico techniques as replacements to animal testing for cosmetic-related substances. Comput. Toxicol. 13, e100112.

Thomas, R.S., Bahadori, T., Buckley, T.J., Cowden, J., Deisenroth, C., Dionisio, K.L., Frithsen, J.B., Grulke, C.M., Gwinn, M.R., Harrill, J.A., Higuchi, M., Houck, K.A., Hughes, M.F., Hunter, E.S., III, Isaacs, K.K., Judson, R.S., Knudsen, T.B., Lambert, J.C., Linnenbrink, M., Martin, T.M., Newton, S.R., Padilla, S., Patlewicz, G., Paul-Friedman, K., Phillips, K.A., Richard, A.M., Sams, R., Shafer, T.J., Setzer, R.W., Shah, I., Simmons, J.E., Simmons, S.O., Singh, A., Sobus, J.R., Strynar, M., Swank, A., Tornero-Valez, R., Ulrich, E.M., Villeneuve, D.L., Wambaugh, J.F., Wetmore, B.A., Williams, A.J., 2019. The next generation blueprint of computational toxicology at the U.S. Environmental Protection Agency. Toxicol. Sci. 169, 317–332.

Toropova, A.P., Toropov, A.A., 2018. Use of the index of ideality of correlation to improve models of eco-toxicity. Environ. Sci. Poll. Res. 25, 31771–31775.

Wang, L., Xing, P., Wang, C., Zhou, X., Dai, Z., Bai, L., 2019. Maximal Information Coefficient and Support Vector Regression based nonlinear feature selection and QSAR modeling on toxicity of alcohol compounds to tadpoles of *Rana temporaria*. J. Braz. Chem. Soc. 30, 279-285.

Wittwehr, C., Blomstedt, P., Gosling, J.P., Peltola, T., Raffael, B., Richarz, A.-N., Sienkiewicz, M., Whaley, P., Worth, A., Whelan, M., 2020, Artificial Intelligence for chemical risk assessment. Comput. Toxicol. 13, e100114,

Worth, A.P., 2020. Computational modelling for the sustainable management of chemicals. Comput. Toxicol. 14, e100122.

Worth, A.P., 2010. The role of QSAR methodology in the regulatory assessment of chemicals. In Puzyn, T., Lesczynski, J., Cronin, M.T.D. (Eds.). Recent Advances in QSAR Studies: Methods and Applications. Springer, Dordrecht, The Netherlands, pp. 367-382.

Yan, F., Liu, T., Jia, Q., Wang, Q., 2019. Multiple toxicity endpoint–structure relationships for substituted phenols and anilines. Sci. Tot. Environ. 663: 560–567.

Young, D., Martin, T., Venkatapathy, R., Harten, P., 2008. Are the chemical structures in your QSAR correct? QSAR Comb. Sci. 27, 1337-1345.