Increasing the Confidence of *In Silico* Modelling in Toxicology

Samuel John Belfield

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Doctor of Philosophy.

December 2023

Acknowledgements

First and foremost, I would like to thank my director of studies, Prof Mark Cronin, for all the support and guidance you have provided me over the years. I can truly say that Mark has gone above and beyond in his mentorship, and it has been a pleasure to work with you. I am extremely grateful for your patience, time, and effort that has gone into this thesis. Similarly, I would like to extend my gratitude to both my co-supervisors, Dr Judith Madden and Dr Steve Enoch. Thank you for sharing your expertise and always encouraging me throughout my studies.

Special thanks to Dr James Firman, you have been an invaluable support throughout my PhD, and have been akin to an additional supervisor. Thank you for all the effort and advice you have given me. I am also extremely grateful to Adam Waqar Tahir, thank you for guiding me through my programming and machine learning journey – your support has been incredible.

To my family, thank you for motivating me over all these years. I may not express it enough, but I am truly lucky to have your support. To my partner, Eliza, you have been my rock throughout this, and I cannot express how much your belief in me has meant. Thank you for always being there for me, I sincerely look forward to our future together.

To my good friend Thomas Graver, studying alongside you during our undergraduate was a great pleasure, and without you I would never have discovered this wonderful field. Likewise, I would also like to extend my gratitude to Dr Mark Hewitt, whom I originally began this path with, thank you for all the support and motivation you haven given to me.

Lastly, to the Chemoinformatics Research Group both past and present: Dr David Ebbrell, Dr Julia Pletz, Dr Maria Sapounidou, Dr Nicoleta Spînu, Dr William Masinja, Dr Michelle Assante, Dr Courtney Thompson, and Zuzana Hasarova thank you for being so friendly. I have enjoyed all our conversations and cherish our time working together. It has been a pleasure to get to know each and every one of you.

Abstract

Consideration of all chemicals that we are exposed to on a daily basis is a daunting task, which has been traditionally assessed through animal testing procedures. However, the ethical and financial considerations associated with such testing has long been a topic of concern, with the desire to pursue alternative methods evident. Towards this, the vision of 21st century toxicology actively promoted the use of new approach methodologies (NAMs) that avoid the usage of animal testing, as well as fostering a more efficient means for toxicological assessment. Captured within these NAMs are *in silico* methods which include a range of *in silico* (or computational) approaches, one of the most popular being Quantitative Structure-Activity Relationships (QSARs). Although it is acknowledged that the majority of these *in silico* methods are by no means novel, it is the consideration of such within regulatory decision-making frameworks that is. Whilst these methods are being promoted for usage within regulatory settings, fundamental issues regarding assessment of confidence as well as knowledge sharing need to be addressed to further promote acceptance.

Therefore, the aim of this thesis was to provide detailed analysis of methods for *in silico* model validation, and knowledge-sharing efforts that incorporate the state-of-the-art practices, which could potentially bolster their acceptance within regulatory settings. Recently developed uncertainty assessment criteria for the evaluation of QSARs were analysed with a particular focus on how they can be employed to demonstrate fitness-for-purpose. These uncertainty assessment criteria were subsequently developed further, with considerations of challenges in QSAR, such as mixture assessment and machine learning (ML) approaches. To facilitate this, a review was conducted of the key characteristics of QSAR methods applied to mixtures, using the knowledge gathered to identify areas for additional consideration within the criteria. ML approaches were studied, with six models developed to address ML-specific considerations within the criteria. The concept of model sharing has been promoted through the application of the FAIR (Findable, Accessible, Interoperable, Reusable) principles to *in silico* methods. Outcomes from each chapter and the overall thesis promote the advancement of regulatory acceptance of QSAR models and predictions, through development of improved reporting strategies and sharing methodologies. The thesis additionally benefits the field

through thorough considerations of the most challenging aspects of QSARs, and how these subfields, such as mixture assessment and ML approaches, can gain credibility.

List of Abbreviations:

ADME	Absorption, Distribution, Metabolism and Excretion
AI	Artificial Intelligence
AOP	Adverse Outcome Pathway
API	Application Programming Interface
CA	Concentration Addition
CLP	Classification, Labelling and Packaging
CV	Cross-Validated
DHFR	Dihydrofolate reductase
DL	Deep Learning
DNN	Deep Neural Network
DOI	Digital Object Identifier
ECHA	European Chemicals Agency
EDCs	Endocrine Disrupting Chemicals
E_{LUMO}	Energy of the lowest unoccupied molecular orbital
$E_{LUMO} + 1$	Energy of the second lowest unoccupied molecular orbital
EPAA	European Partnership for Alternative Approaches to Animal Testing
eTOX	European Union IMI Project: "Integrating bioinformatics and chemoinformatics approaches for the development of expert systems allowing the *in silico* prediction of toxicities"
eTRANSAFE	European Union IMI Project: "Enhancing TRANslational SAFEty Assessment through Integrative Knowledge Management"
EU	European Union
EU-ToxRisk	An Integrated European 'Flagship' Programme Driving Mechanism-based Toxicity Testing and Risk Assessment for the 21st Century
EURL ECVAM	EU Reference Laboratory for Alternatives to Animal Testing
FAIR	Findable, Accessible, Interoperable, Reusable
GNN	Graph Neural Network
HCA	Hierarchical Cluster Analysis
IA	Independent Action
IATA	Integrated Approaches to Testing and Assessment
INFCIM	INtegrated Concentration Addition-Independent action Model

K	Number of Folds
KNN	K-Nearest Neighbours
LIME	Local Interpretable Model-agnostic Explanations
log P	Logarithm of the octanol-water partition coefficient
ML	Machine Learning
MLR	Multiple Linear Regression
MSE	Mean Squared Error
N/A	Not applicable
NAMs	New Approach Methodologies
NGRA	Next Generation Risk Assessment
NICEATM	National Toxicology Program Interagency Centre for the Evaluation of Alternative Toxicological Methods
NN	Neural Network
NOEL	No observed effect level
OECD	Organization for Economic Cooperation and Development
PAHs	Polycyclic aromatic hydrocarbons
PBK	Physiologically-based kinetic
PBPK	Physiologically-based pharmacokinetic
PCA	Principal Component Analysis
PFAS	Per- and polyfluoroalkyl substances
PLS	Partial Least Squares
q^2	Leave-one-out coefficient of determination
QMRF	QSAR Model Reporting Format
QPRF	QSAR Prediction Reporting Format
QSAR	Quantitative Structure-Activity Relationship
QSI	Quorum sensing inhibitor
QSPR	Quantitative Structure-Property Relationship
R^2	Coefficient of determination
RBFNN	Radial Basis Function Neural Networks
RDMkit	Research Data Management toolkit for Life Sciences
REACH	Registration, Evaluation, Authorisation (restriction) of Chemicals
ReLU	Rectified Linear Units
RF	Random Forest

SEURAT-1	Safety Evaluation Ultimately Replacing Animal Testing - Phase 1
SHAP	SHapley Additive exPlanations
SiRMS	Simplex Representation of Molecular Structure
SMILES	Simplified Molecular Input Line Entry System
SVM	Support Vector Machine
TMP	Trimethoprim
Tox21	Toxicology in the 21st Century
TSCA	Toxic Substances Control Act
US	United States
US EPA	United States Environmental Protection Agency
UVCB	Unknown or variable composition, complex reaction products or biological materials
XGBoost	Extreme Gradient Boosting

Table of Contents

Chapter 1. Introduction

1.1. Background

Throughout our daily lives, we are continually exposed to a multitude of chemicals, the potential effects of most of these are not yet fully understood. As such, schemes for addressing the dangers chemicals present to both individuals and the environment have been in place for almost a century (Hartung, 2009). During this time, it has been estimated that between 10-20,000 substances, such as pharmaceuticals, pesticides and many other products, have been tested. However, only a small proportion of the total number of substances can be considered to be well-studied and thoroughly assessed, with many of these receiving such focus due to possible health concerns (Krewski et al., 2019). Historically, chemicals that have been subjected to thorough testing are those that have been identified to be of significant health concern, for example carcinogens and, more recently, endocrine disrupting chemicals (EDCs) (Hartung, 2009). EDCs are an example of one of the significant issues faced by chemical safety assessment. Initial research into these substances was conducted in the mid-twentieth century following a study where researchers linked prenatal exposure, to a later defined EDC, with cancer of the cervix (Herbst et al., 1971). Similarly, a more recent issue that has become a focus of chemical risk assessment is that of per- and polyfluoroalkyl substances (PFAS). PFAS, otherwise referred to as forever chemicals, due to their lengthy persistence within the environment, have come under scrutiny due to their vast prevalence causing global health effects worldwide (Fenton et al., 2021). Both EDCs and PFAS represent a small handful of chemicals that can cause a breadth of effects to both humans and the environment that have, and will continue to be, a focus of research for the foreseeable future. However, there still exists a great need for information for millions of other chemicals that we are exposed to with unknown effects (Krewski et al., 2019).

1.2. Chemical legislation and animal testing

Determining the potential risk of exposure to chemicals not only for humans, but also the environment has been undertaken through various regulatory bodies governed by legislations; with the earliest systems for determining such hazards being introduced as far back as almost a century ago (Hartung, 2009). Since conception, there now exists over 40 pieces of key chemical legislation globally such as the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) within the EU and the Toxic Substances Control Act (TSCA) in the US. Specifically in the EU, legislation is fronted by REACH and Classification, Labelling and Packaging (CLP), which is supported by individual policies for specific groups of chemicals such as biocides, pesticides, pharmaceuticals, and cosmetics (Mahony et al., 2020). Designed as the major policy to protect both human health and the environment, REACH places strict requirements for hazard assessment upon chemicals that are produced or imported into the EU in quantities exceeding one tonne, which understandably applies pressure on the usage of animal testing. Throughout traditional risk assessment the underlying assumptions have been that whole animal testing is a sufficient predictor of adverse effects towards human/environment (Knight et al., 2021). Nevertheless, animal studies alone are unlikely to fully capture the scope of adverse effects caused, with the relevance of such results towards humans also being arguable. Thus, the reliance upon animal testing alone is ultimately outdated, with such methods simply not able to test all existing chemicals, whilst also being costly, time-consuming, and highly ethically debatable. Therefore, REACH, along with other EU legislations, actively promote the usage of alternative approaches, with there being an evident desire for new, non-animal approaches within safety assessment.

1.3. 21st Century Toxicology and NAMs

The field of toxicology is a continually progressing and developing practice, with advancements in human biology and tools for determining adverse effects of chemicals, and other stressors, growing exponentially. Ensuring that such new technologies are actively being employed within the field motivated the US National Research Council (NRC) to publish a landmark report over a decade ago labelled Toxicity Testing in the 21st Century (US NRC, 2007; Krewski et al., 2020). This report provided a vision for the future of toxicology forming a long-term strategy designed specifically to take advantage of newly introduced technologies; thus, increasing the efficiency of testing procedures enabling acceleration of the rate at which chemicals could be considered (US NRC, 2007). Reducing the reliance upon animal testing understandably plays a pivotal role within the strategy, shifting dependence towards animal alternatives and promoting the usage of new approach methodologies

(NAMs). NAMs represent any technology, methodology, approach, or combination of approaches that produces information avoiding the usage of animals. Captured within the NAMs terminology include approaches such as *in vitro* systems as well as *in silico* methods (the latter being the focus of the current thesis). These may not be novel themselves, however, their application within regulatory decision-making processes and replacement of traditional testing is a newer development (van der Zalm et al., 2022).

This need for updating chemical safety assessment practices at a regulatory level has long been understood, with the EU also adopting policies that reduce dependencies upon animal testing with the adoption of the Directive 2010/63/EU on the protection of animals used for scientific purposes being firmly rooted in the 3Rs principles (replacement, reduction, and refinement) (European Parliament and Council, 2010). Reflecting upon the EUs priorities to reduce animal testing has additionally been demonstrated by the substantial funding contributed towards various research programmes, such as SEURAT-1, EU-ToxRisk, eTOX, and eTRANSAFE (http://www.seurat-1.eu/; http://www.eu-toxrisk.eu/; http://www.etoxproject.eu/; https://etransafe.eu/). Similar efforts are also being undertaken outside of Europe, with the toxicology in the 21st century (Tox21) consortium being formed through a collaborative effort between US regulatory agencies (https://tox21.gov/). Support from such research initiatives has undoubtedly accelerated the growth of animal alternative approaches, i.e., NAMs, in both *in silico* and *in vitro* disciplines (EPA, 2018).

Replacing animal studies in safety assessment on a case-by-case basis would present an ideal scenario, but with the understanding of NAMs being somewhat in their infancy this approach is currently not feasible. Thus, combining data from a variety of approaches in a weight-of-evidence manner presents a logical and robust strategy for utilisation of NAMs (Mahony et al., 2020; Laroche et al., 2019). Though such work can provide short term gains towards the further implementation of NAMs, broader considerations must be undertaken to update current regulatory practices (Knight et al., 2021). Shifting from supporting information towards a more direct replacement indeed requires greater acceptance in risk assessment, which may only be facilitated through the revision of relevant legislation. Amending legislation is a cumbersome and slow process, and is additionally hampered where approaches lack scientific validity (Eskes and Whelan, 2016). Evidently, for NAMs, greater effort needs to be placed into further maturing the approaches. Recent workshops held by

The European Partnership for Alternative Approaches to Animal Testing (EPAA) identified key challenges, as shown in Table 1.1, that need to be addressed in order to promote acceptance (Westmoreland et al., 2022; Mahony et al., 2020). Therefore, to capitalise upon the vast research effort and investments that have progressed the development of NAMs, it is essential that such acceptance issues are addressed; thus, encouraging a re-assessment of the current safety paradigm (Cronin et al., 2021).

Table 1.1. Overview of the key challenges hindering acceptance of NAMs identified during EPAA collaborative workshops (adapted from (Westmoreland et al., 2022; Mahony et al., 2020).

Area of challenge toward NAMs	Description
Legislation	There is a clear lack of experience validating NAMs, with no one agreed upon approach being employed in regulatory science. The potential utility of NAMs in a safety assessment framework are not fully realised.
Data Sharing	Adherence to a unified approach for data sharing and management needs to be upheld. Acknowledgement that the Findable, Accessible, Interoperable, and Reusable (FAIR) principles should be applied.
Computational Approaches	There is apprehension in shifting from traditional modelling paradigms towards state-of-the-art technologies. The uptake of ML and AI methods should be promoted.
Decision Making Frameworks	Confidence towards a prediction or model are at best unclear hindering the acceptance. Addressing the acceptable level of uncertainty and scenarios that the expectations may be lowered and deemed acceptable for purpose should be defined.
Acceptance	The understandings of NAMs as a whole are lacking, severely impeding acceptance of their models and respective data in regulatory decisions. Further confusion arises from the issues of demonstrating fitness-for-purpose, the lack of consistent performance standards, and how NAMs can be utilised.

1.4. Quantitative Structure-Activity Relationships

One of the most fundamental NAMs that may be employed within 21st Century Toxicology, and previously alluded to within Section 1.3, are *in silico* methods. *In silico* methods refer to experimentation through the usage of computational means, with the procedures additionally being referred to as computational methods within literature (Ekins et al., 2007). Included within this field are a plethora of methods, such as read-across, physiologically based pharmacokinetic (PBPK) models, and quantitative adverse outcome pathways (qAOPs) to name a few. However, the focus of research throughout this thesis is related to Quantitative Structure-Activity Relationships (QSARs) (Ram et al., 2022). QSARs are a well-established *in silico* modelling technique that was originally popularised by the seminal work published by Hansch et al. (1962). Since conception, the value of QSARs as a predictive technique has been well proven, especially within scenarios dealing with toxicity predictions and data gap filling (Cronin and Yoon, 2019). By definition, a QSAR model is able to make predictions in the absence of data through defining the relationship between chemical descriptors (such as molecular structure and physicochemical properties) and the toxicological endpoint (Cronin et al., 2019). In general, the workflow of QSAR modelling can be outlined following its three fundamental requirements: data, descriptors, and statistical technique (Madden et al., 2020). Curation of a dataset with a "defined endpoint" for a series of related chemicals of good quality is crucial, where model validity may become flawed by erroneous data (De et al., 2022). Predefined endpoints can be categorised as: physico-chemical properties, such as the octanol-water partition coefficient (log P), environmental fate parameters, such as bioaccumulation, ecotoxic effects, such as acute toxicity, and human health effects, such as skin sensitisation (ECHA, 2008).

Molecular descriptors are the second core requirement for QSAR modelling, with a vast number of different types being available that provide detailed information concerning chemical structure and properties. In essence, molecular descriptors enable molecules' properties to be expressed as a mathematical representation, with these numerical values being employed to quantitatively describe both physical and chemical information (Chandrasekaran et al., 2018). Acknowledged as one of the most crucial aspects of QSAR modelling, information that is captured by descriptors is largely dependent upon either the molecular representation or algorithm used for calculation. A broad classification of the different types of descriptors can be seen in Table 1.2. Within these different classifications exists a vast quantity of descriptors, which require careful pruning, during the modelling process, to ensure the removal of redundant, noisy, and irrelevant information that may affect model performance (Xia et al., 2019). Selection of the descriptors to be used is largely dependent upon the intended use case of the model. In general, easily interpretable descriptors are preferred for risk assessment, whereas descriptors solely based upon statistical correlation are traditionally utilised in screening procedures (Madden et al., 2020).

Table 1.2. Description of the different categories of molecular descriptors (adapted from Danishuddin and Khan, 2016).

Descriptor classification	Overview
Physicochemical	Physical and chemical information from a molecule that can be determined through examination of its 2D structure.
Constitutional	Simplistic representations of molecular composition without the use of topological information.
Topological	2D descriptors utilising molecular graphs that capture compounds' internal atomic arrangement.
Geometrical	Determined from a given molecule's three-dimensional coordinates based upon all atoms.
Thermodynamic	Relationship between the chemical structure and chemical behaviour observed.
Electronic	Description of electronic properties of either the full molecule, atomic bonds, or molecular fragments.

The statistical technique that is employed to express the relationship between the selected molecular descriptors and endpoint of interest is the last requirement of QSAR modelling. As the field of QSAR has matured over the years, owing to progress in computational power, data availability and chemoinformatics, so too has the complexity of statistical techniques (such as machine learning methods) used (Cherkasov et al., 2014). This is discussed further in Chapter 4. Such statistical methods employed in modelling can be separated depending upon the expected output variable. Regression-based models are used in the prediction of continuous values (quantitative), whereas classification-based methods categorise data into different groups, such as active and inactive (qualitative) (Roy et al., 2015). Some of the most commonly used statistical methods for regression include multiple linear regression (MLR) and partial least squares (PLS). Classification modelling is traditionally performed using principal component analysis (PCA) and hierarchical cluster analysis (HCA) (Pirhadi et al., 2015). However, many machine learning algorithms, such as random forest, support vector machines, and neural networks, are unspecific and so may be used for the prediction of either output (Roy et al., 2015). Irrespective of the model developed, it is crucial that the performance of the selected technique is sufficiently evaluated. To this end, it is essential to define the difference between model fit and predictive performance, otherwise referred to as internal and external validation, respectively. Firstly, model fit reports the ability of the model to mathematically reproduce the output of the training set, which due to the model being developed using known data can be arbitrarily manipulated, with enough free parameters, to provide seemingly perfect predictive scorings (Eriksson et al., 2003). Unlike goodness of fit, predictive ability enables a measurement of how well data not previously seen by the algorithm can be estimated. This measurement is typically reported as a goodness of prediction parameter (such as q^2) and may be evaluated from a variety of proposed validation methodologies, although ultimately will provide an evaluation of the same outcome – external validation (Chirico and Gramatica, 2011). Parameters utilised within these validation strategies differ depending upon the type of QSAR developed. Regression-based QSARs are assessed based upon considerations such as standard error of estimate, determination coefficient, and explained variance (Roy and Kar, 2015). Whereas, classification-based can be evaluated depending on the sensitivity and specificity, which express the model's ability to predict a true positive or a true negative, respectively (Walker et al., 2003).

1.5. Acceptance of QSAR models and predictions

For a chemical to achieve regulatory acceptance it is imperative that the underlying risks associated with it are fully understood. To assist with this, a multitude of regulatory programmes have been conceived that enable the assessment and management of chemicals based upon a vast amount of chemical information, such as physiochemical, environmental fate, as well as adverse effects on human and environmental species (Worth, 2010). In particular, the information required for chemicals has been detailed in various legislation such as REACH and TSCA in the EU and US, respectively. Although the types and quantity of information required for these chemical safety assessment programmes vary, satisfying all requirements with available data from traditional approaches is highly unlikely. Therefore, the use of alternative approaches such as QSAR modelling offer a potential replacement for traditional testing, this may be as a support to priority setting procedures, supplementation to experimental data in weight-of-evidence approaches, or as a stand-alone replacement to experimental data (Worth, 2010).

Whilst using information obtained from QSARs in a supporting manner may only impact the outcome of an assessment indirectly, resulting in greater flexibility in the confidence of the models required, fully substituting experimental data undoubtedly requires greater certainty. Yet, with REACH actively advocating the use of QSARs, a framework to enable the acceptance of data from such methods has been devised. In essence, this framework can be summarised as: the model being proven to be scientifically valid, the model demonstrating applicability to the chemical(s) of interest, the prediction being shown to be adequate for the purpose, and lastly the method and result are suitably documented. Satisfying all such requirements will inherently provide confidence in the use QSARs as direct replacements of experimental data, while at the same time flexibility is possible, at the discretion of relevant judgement, in scenarios whereby the data are instead used in supporting roles.

Fulfilling these considerations to ensure the quality of a QSAR requires appreciation of the prerequisite information, statistical procedures, and mechanistic basis used to develop the model (Cronin et al., 2019). This awareness resulted in the definition of an initial six principles for the validation of QSARs, that were later condensed to five once adopted by the OECD Principles for the Validation of QSARs for Regulatory Use (OECD, 2007). These principles aim to facilitate the use of a QSAR model for regulatory applications, with these requirements being defined as:

1. Associated with a defined endpoint.
2. Developed using an unambiguous algorithm.
3. Boundaries of limitation outlined using a defined domain of applicability.
4. Performance of the model determined using appropriate statistical measures such as goodness-of-fit, robustness and predictivity.
5. Mechanistic interpretation to be provided (where possible) between the descriptors employed and the endpoint modelled.

Employing these principles as a framework, particularly in context when applied using the QSAR Modelling Reporting Format (QMRF), can enable conclusions of validation in terms of regulatory acceptance to be drawn. The usage of such reporting procedures has served the wider QSAR community well. However, the field of QSAR has developed exponentially since these initial frameworks were developed, with significantly more complex models now being produced. Fully evaluating such models using traditional frameworks may give an indication of validity but can no longer be assumed to be sufficient. Additionally, within toxicology a shift towards the use of weight-of-evidence based approaches, coinciding with an emphasis on defining uncertainty has occurred in recent years. Presenting data with defined levels of uncertainty can be highly beneficial due to their intrinsic diagnostic nature, enabling information deficits of the model to be addressed (Patterson and Whelan, 2017).

Acknowledging this need to update QSAR evaluation approaches, a recent study by Cronin et al. (2019) developed a framework enabling the uncertainties associated with QSAR models and predictions to be fully characterised. Within this framework a list of 49 assessment criteria were defined accounting for uncertainties arising throughout the entire development of a QSAR – including uncertainties in data, modelling approach, description and application. Organising information in this manner not only enabled adequacy towards the intended purpose to be defined, based upon semi-quantification of uncertainty, but additionally provided an opportunity for developers to identify issues that could be rectified using mitigation strategies. The layout of the framework undoubtedly provided a route towards the assessment of more complex QSAR approaches. As such, an opportunity exists to demonstrate such applicability following targeted case studies that capture these current challenges.

1.6. Research aims and contributions to knowledge

The field of QSAR is continually growing in interest, bringing rapid expansion within the approaches utilised, as well as the predictive problems faced. Evaluating this expanding field using traditional assessment procedures is limited, requiring further considerations to be addressed. The recently developed QSAR uncertainty framework by Cronin et al. (2019) provides a flexible foundation that can sustain the active growth. As such, this thesis aimed to expand upon the current framework, as well as further develop it in regard to issues such as chemical mixtures and ML. The objectives to achieve this aim were addressed in the following chapters:

Chapter 2: Determine fitness-for-purpose of QSARs through the usage of the uncertainty assessment criteria.

- This involved the definition of ten components, through grouping of the original 49 assessment criteria. Components were then related to the phases of QSAR development used to assess QSARs' fitness-for-purpose with the proposal of mitigation strategies.

Chapter 3: Review current practices in developing QSARs for mixtures, mapping key characteristics and challenges within the approaches onto the uncertainty assessment criteria.

- This involved performing a review of studies related to QSARs and mixtures, curating a list of relevant literature. Characteristics of each QSAR studied were identified and discussed, which were later mapped onto the original uncertainty assessment criteria improving mixture-specific considerations.

Chapter 4: Investigate common ML methods within QSAR and determine how these can be addressed using the uncertainty assessment criteria.

- This involved the development of six models using the most frequently employed QSAR ML algorithms. Assessment of the models with respect to the uncertainty criteria was then conducted, bolstering the criteria with ML-specific considerations.

Chapter 5: Apply the FAIR (Findable, Accessible, Interoperable, Reusable) principles, to *in silico* predictive models.

- This involved the identification of FAIR principles that enable the FAIRification of *in silico* methods. The principles were later applied to the previously developed ML models from Chapter 4.

Through the completion of research outlined above, the thesis aims to advance current uncertainty analysis schemes for the assessment of QSARs. Such contributions are observed through the initial utilisation of the uncertainty assessment criteria as a tool to enable the definition of fitness-for-purpose following the grouping of components. Subsequent studies further expand the knowledge within this field, through the identification of model- and approach-dependent considerations. Lastly, the thesis provides direction to the improvement of model sharing through the definition and interpretation of FAIR principles. As such, utilisation of the information gained throughout the thesis will enable for improved assessments of QSARs, irrespective of complexity, promoting the usage of such models within their respective applications.

Chapter 2. Determination of “fitness-for-purpose” of quantitative structure-activity relationship (QSAR) models to predict (eco-)toxicological endpoints for regulatory use

Preface:

This work has been published in: Belfield SJ et al., (2021). Determination of “fitness-for-purpose” of quantitative structure-activity relationship (QSAR) models to predict (eco-)toxicological endpoints for regulatory use. Regul. Toxicol. Pharmacol. 123: 104956. doi: 10.1016/j.yrtph.2021.104956

This was a multi-author paper. Belfield led the work and analysis in this study as recognised in the CRediT authorship contribution statement: Conceptualization, Methodology, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization.

2.1. 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 chemical 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 Chapter 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 2.1 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 chapter, 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* models for specific purposes.

Table 2.1. 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
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 performance of a new modelling approach, descriptors etc.	Wide range of models applicable	High
		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	Assessment of the safety of an	Specific mechanistically based and justified models	Low

compound in a product	established or new compound in a product or formulation		
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
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 over 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 requires consideration of 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 – creation, characterisation, and application. 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.

2.2. Methods

2.2.1. 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 herein as "components") was aligned to one of the three phases in the development of a QSAR.

2.2.2. 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 2.2.3 and 2.2.4). As part of this process, the needs of regulatory users were considered in the context of what may make the QSARs fit for this purpose.

2.2.3. 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, or 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 *Appendix I*, 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).

2.2.4. 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, 1981a). 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.

2.3. Results

2.3.1. 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 2.1. 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 2.2, with details of the individual uncertainty criteria, represented within each component, being denoted in

Appendix I 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 (*Appendix I* Table S1) within the components may not be required for a particular model, depending on the purpose of the model/endpoint under consideration.

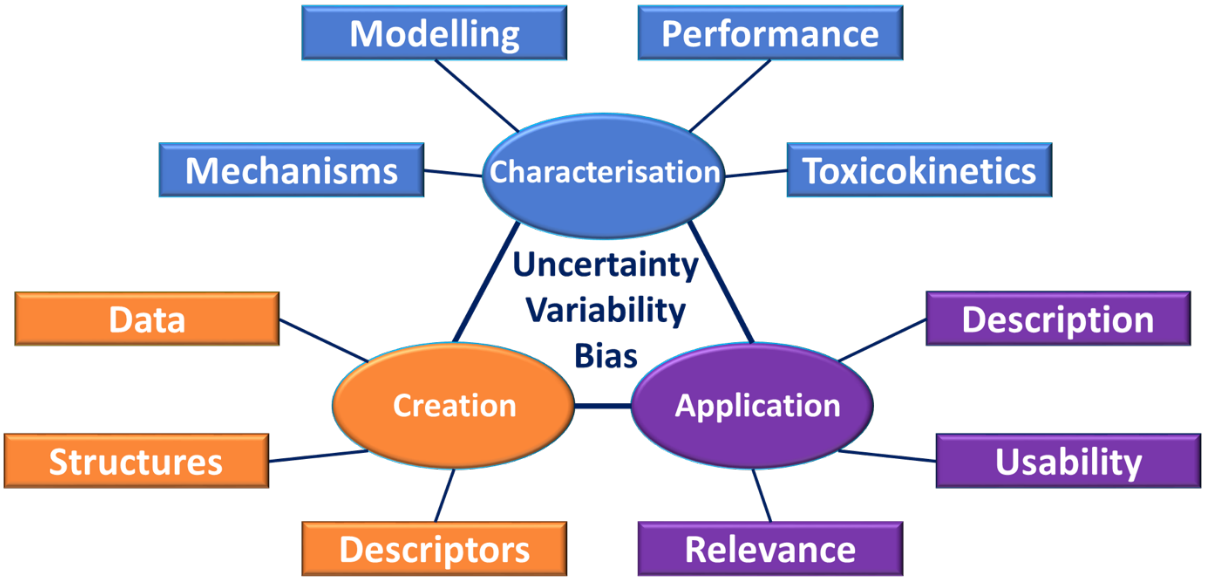


Figure 2.1. Scheme summarising the ten "components" of QSAR models required to be considered for toxicity prediction purposes. The components, denoted in the rectangular boxes, are linked to the phases, denoted in the oval shapes and defined for each of the three broad areas of QSAR uncertainty, variability and bias.

Table 2.2. Key features of the proposed ten components for QSARs.

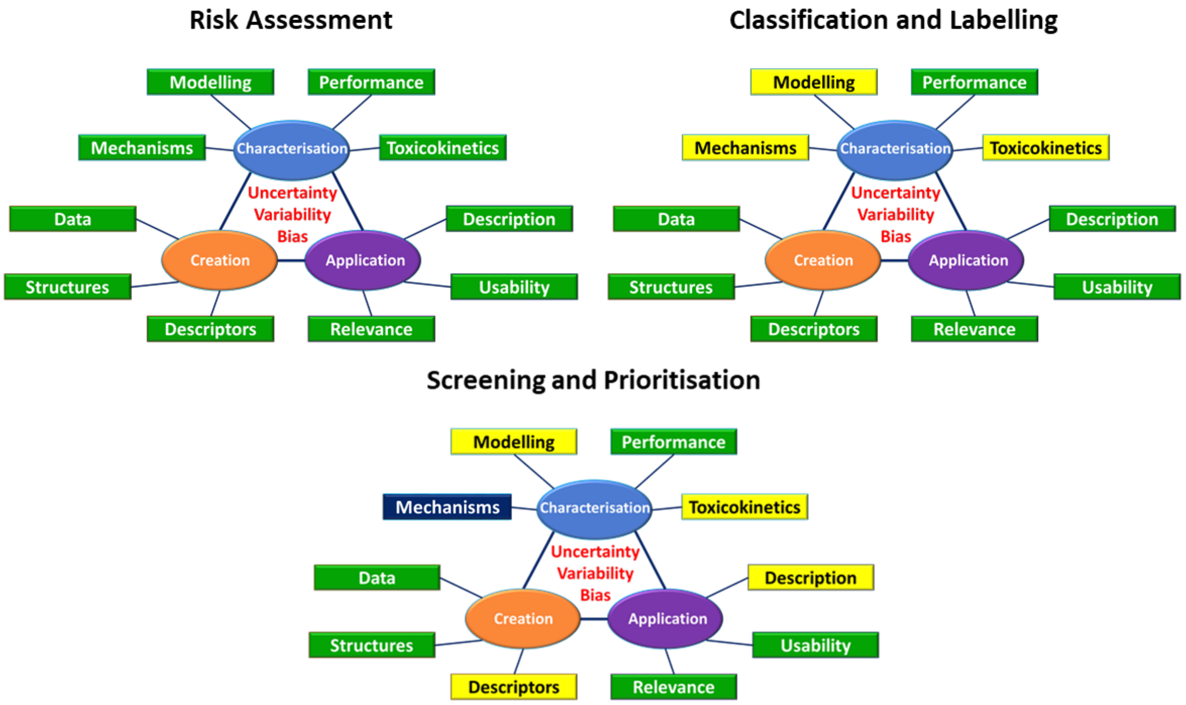
Component	Key Features Used to Assess the Components
Model Creation	
1. Data	Quality of individual studies within the data set and the data set overall that was used for modelling
2. Structures	Accuracy of the reported chemical structures in the training (and, if applicable, test) set used for modelling
3. Descriptors	Appropriate use and adequate definition of the descriptors used for modelling (including how and where sourced)
Model Characterisation	

4. Modelling	The appropriateness 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 and transparency of the model
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

2.3.2. 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 2.2 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.


Risk Assessment

Classification and Labelling

Screening and Prioritisation

Figure 2.2. Levels of uncertainty considered acceptable for QSAR components associated with different regulatory uses; green indicates low uncertainty; yellow indicates moderate uncertainty and blue indicates high uncertainty.

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

The literature search identified 150 papers in Web of Science published 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 2.3. 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 2.3. 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., 2019a
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 neighbor, 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 cells HT-29, and rat adrenal pheochromocytoma cell line PC-12	8 metal oxide nanoparticles	CORAL descriptors	Monte Carlo optimisation of target functions	Ahmadi, 2020

2.3.4. 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 2.3 and described in detail in *Appendix I* Table S2. The overall levels of uncertainty for the 12 QSAR studies provided in Figure 2.3 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 2.3 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 2.2). 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.

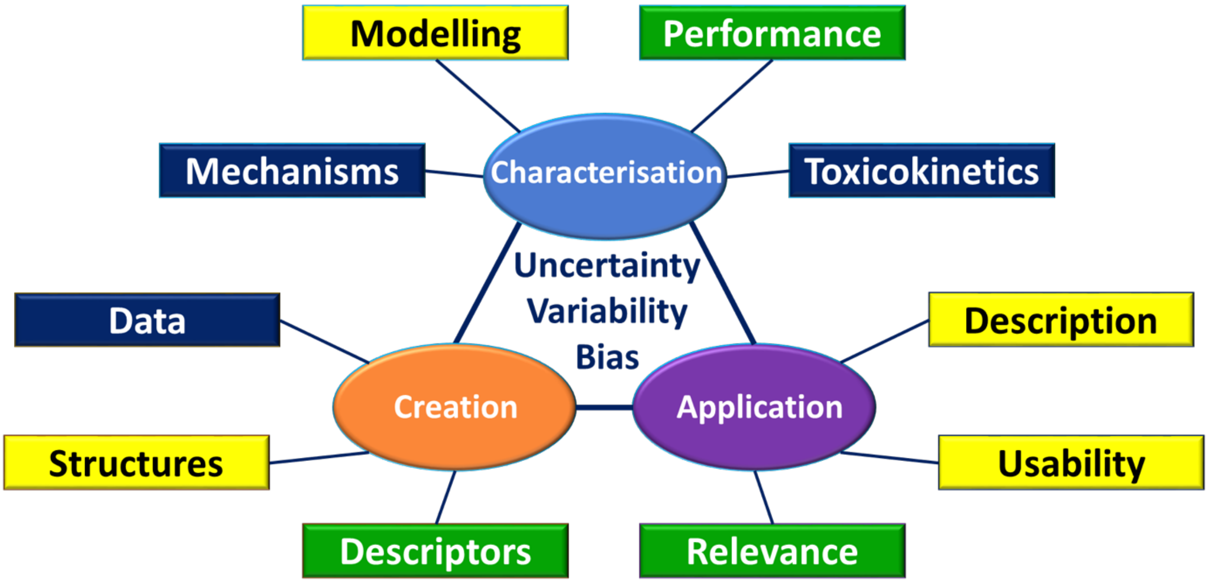


Figure 2.3. A summary of the levels of uncertainty associated with QSAR components for the 12 QSAR studies evaluated; green indicates low uncertainty for component, yellow moderate uncertainty and blue high uncertainty. A full breakdown on the uncertainty associated with each component is provided in *Appendix I* 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 2.4. 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 2.4 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 2.4). 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 *Appendix I* 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, 1981a).

Table 2.4. 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 assessment
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 test sets. Lack of mechanistic interpretation.	Provide details on biological data and descriptor set. Remove duplicates from the training and test sets. Apply mechanistic interpretation (if possible).	Screening

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 transparent. Lack of mechanistic interpretation.	Provide details on biological data and descriptor set. Inclusion of each models' algorithms. Apply mechanistic interpretation.	Screening
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 imbalanced data set. Lack	Supplementation of unambiguous chemical structures. Provide details on biological data and physicochemical properties. Balance actives vs inactives in	Classification and Labelling

			of mechanistic interpretation.	data set. Apply 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 and 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, 2017a). Assess usage of various cell lines for single model. Provide details on biological data and descriptor set. Apply mechanistic interpretation.	Possible Classification and Labelling

2.4. 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 now available 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.

2.4.1. "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 2.1 and summarised in Table 2.2. 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 2.1). 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.

2.4.2. 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 2.2 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 a 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 2.2 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 areas of 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 are 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 2.4 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.

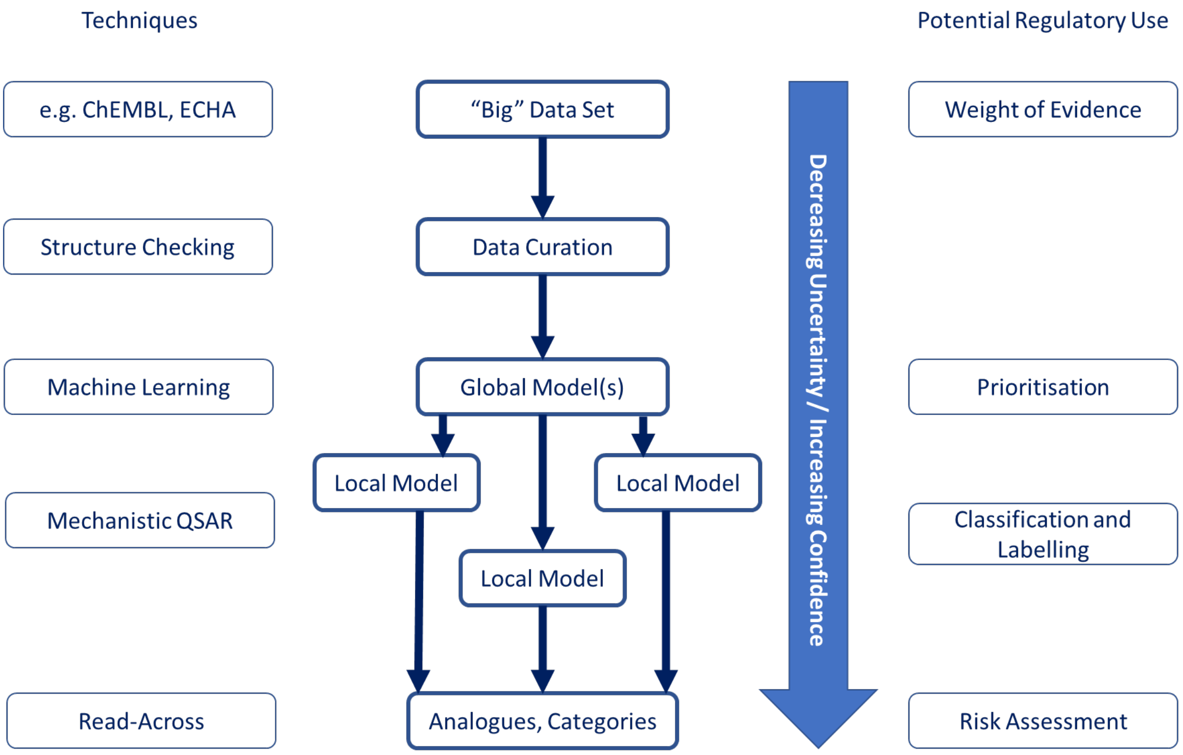


Figure 2.4. Potential regulatory use of different types of QSARs and *in silico* models that could be derived from a "big" data set. Models range from global machine learning to read-across from close analogues.

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 2.2) 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.

2.4.3. 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 (*Appendix I* Table S2) would be overwhelming, so the use of a reduced number of components to gain an overview, is valuable. Also illustrative is the summary of the uncertainties across all the QSARs analysed (Figure 2.3). Assignment of these uncertainties for each component have been based upon expert judgement, thus the occurrence of human bias throughout the procedure should be taken into account. Whilst this may be unnecessary to be considered for the purpose of this study, mitigation of this factor could be achieved through the use of external reviewers. 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 the purpose of regulatory use, it is informative to consider them in more detail to investigate to which purpose they could be applied (Table 2.4) and what measures may be required to achieve this (*Appendix I* Table S4). Comparison of the summary of results in Table 2.3 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 (time-dependent) ADME properties affecting bioavailability, 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, 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 (such as TA1535, TA1537, TA97a or TA97, TA98) to be considered.

2.4.4. Reducing uncertainty of QSARs using the assessment components

Assessment of QSAR models in the manner described 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 (*Appendix I* 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 *Appendix I* 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.

2.4.5. 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 2.1 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 2.2). 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, for regulatory purposes, 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.

2.5. Conclusions

Ten assessment components have been described in this study which are designed to assess not only uncertainties, but also variabilities and areas of bias of QSAR models. These components rationalise and organise the original 49 criteria from Cronin et al. 2019 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 need to be considered, when applying QSARs for regulatory purposes, and demonstrate 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 model's applicability in a regulatory context. For the QSAR models 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 QSAR models 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.

As exemplified through this study, the uncertainty criteria serve as an extremely valuable tool that can not only improve models through the identification of shortcomings, but additionally provide supporting evidence that a model is fit for purpose. Whilst the current study demonstrated the criteria were successful at determining the uncertainties associated with traditional modelling problems, the field of QSAR is ever-expanding with current state-of-the-art approaches utilising AI techniques, as well as the utilisation of models to predict the adverse effects of complex mixtures. Such problems require careful consideration before acceptance can be achieved, which using traditional practices may be unfeasible; thus, the importance of the uncertainty criteria to provide supporting evidence for a constantly evolving field is deemed essential. As stated above, development and evaluation of QSAR models for mixtures, is associated with additional complexity. In the next chapter the state-of-the-art of QSAR models as applied to assessment of toxicity for mixtures is investigated.

Chapter 3. A review of quantitative structure-activity relationship modelling approaches to predict the toxicity of mixtures

Preface:

This work has been published in: Belfield SJ et al., (2022). A review of quantitative structure-activity relationship modelling approaches to predict the toxicity of mixtures. Comput. Toxicol. 25: 100251. doi: 10.1016/j.comtox.2022.100251

This was a multi-author paper. Belfield led the work and analysis in this study as recognised in the CRediT statement: Conceptualization, Methodology, Investigation, Data Curation, Writing – Original Draft, Visualization.

3.1. Introduction

A significant proportion of toxicological and physicochemical analysis is performed upon single compounds, yet the scenario of one being exposed to a single chemical in isolation is unrealistic (Yang et al., 1998). In reality, both humans and environmental species face various, ever-changing mixtures of chemicals throughout daily life (European Commission, 2012a). Most, if not all, chemicals are encountered as mixtures, for instance specifically marketed formulated mixtures such as pesticides, food and feed additives and cosmetics (typically referred to as intentional mixtures). In addition, exposure to mixtures of chemicals that may interact is not limited to manufactured products. For example, co-administration of drugs may lead to drug-drug interactions and environmental pollutants may also present themselves unintentionally as mixtures from different sources (Kienzler et al., 2016; Palleria et al., 2013). The prevalence of exposure to mixtures, occurring either intentionally or unintentionally, is evidently large, although only partial regulation of intentional mixture is currently provided (Hassold et al., 2021).

Chemical mixtures can be defined as combinations of two or more chemicals that retain their individual, unaltered chemical identities (European Commission, 2012a). In certain circumstances, mixtures may be more problematic when compared to single compounds; a significant concern arises where the individual components are present in mixtures at concentrations where no effect would be anticipated e.g., lower than the no-observed-effect level (NOEL), yet in combination may have the potential to exert unexpected toxicological effects (European Commission, 2012b; Conley et al., 2021). In addition, one of the key actions of the European Union's (EU's) recent "Chemicals Strategy for Sustainability Towards a Toxic-Free Environment" is to take account of the effects of chemical mixtures (European Commission, 2020). However, as the ability to assess the vast number of potential combinations of substances using traditional experimental toxicity testing is unfeasible (European Commission, 2012a), the value that predictive approaches can provide to mixture toxicity is anticipated to play an increasingly important role in toxicity assessment. Traditional approaches for hazard assessment of chemical mixtures may either consider the mixture as a whole (top-down), or contributions from the individual components (bottom-up). In general, assessments are typically driven by bottom-up frameworks, where the individual toxicities of all components are known and then modelled mathematically to predict the combined effect of a mixture (Hernández et al., 2017). In such bottom-up or component-based approaches, it is essential to consider the influence of interactions which may arise between individual components. Where it is presumed that each constituent compound does not impact upon the biological activity of the other, the combined toxicity of a mixture is estimated according to the principle of additivity (European Commission, 2012a; WHO 2017). Should components be understood to operate through similar modes of action, this is typically framed through application of concentration addition (CA) (Loewe and Muischnek, 1926). Alternatively, those with dissimilar modes may be modelled with the assumption of independent action (IA) (Bliss, 1939). These have since been termed "first generation" techniques (Kim et al., 2013a). Whilst the decision on which procedure to adopt is dependent upon the nature of the mixture under examination, the enhanced conservatism inherent within CA has led to its emergence as the generic methodology particularly favoured by risk assessors (Belden et al., 2007, European Commission, 2009a; Kim and Kim, 2015). "Second generation" models, further accounting for variation in mode of action and in turn combining elements of both approaches (integrated addition) later emerged – with uptake generally restricted on account of the greater quantities of empirical data required in their training (Kim et al., 2013a).

Deviations from the ideal of additivity may be noted in instances whereby inter-component interactions do occur. The prevalence of such non-additive effects must not be understated, with a recent literature review by Martin et al. (2021) observing such behaviours within almost half the experimental mixture studies they reviewed (n=1220). The term “synergy” describes the phenomenon through which mixture activity is observed as greater than that predicted by simple additivity, and “antagonism” the inverse in which it is less than that predicted (Ashford, 1981; Bopp et al., 2015; Hernàndez et al., 2017; Rodea-Palomares et al., 2015). Neither CA nor IA is equipped to handle such eventualities, and as such the potential occurrence of either serves to contribute greatly towards uncertainty surrounding estimation of overall mixture toxicity – notably at very low exposure levels (Cedergreen, 2014; Hernàndez et al., 2017). Whilst the concept of the “funnel hypothesis” has been forwarded as a means of rationalising the observation that deviation from additivity is less common amongst multi-component (greater-than-binary) mixtures (Warne 1995), the occurrence of both synergy and antagonism remains challenging to forecast.

In order to assess the toxicity of a greater number and form of mixtures, both additive and non-additive, there is scope for the application of further modelling approaches. One such class of models are quantitative structure-activity relationships (QSARs). QSARs have been used widely in various industrial sectors to predict a range of toxicity endpoints, as well as enabling data gap filling (Madden et al., 2020). Predictions are formulated through identifying the correlation between quantifiable properties of the chemical, and the endpoint of concern – thus a model may allow for estimation of missing data by making use of structural information (Cronin et al., 2019). One of the earliest applications of QSARs towards mixtures was reported by Könemann (1981b), where it was recognised that the additive toxicity of mixtures could be predicted without use of empirical mechanism of toxic action data. Following this, much effort has been put into further development of related methods – since labelled “third generation”. Significant scope exists for utilisation of such approaches, on account both of their practicality and potential predictive power. Ready generation of input parameters through employment of computational techniques may allow for data generation and broadening of applicability domain.

With regard to safety assessment, there is an ever-growing need for the harmonisation of approaches that address the effects of mixtures on human health and the environment. The role of *in silico* methods within the determination of mixture toxicity is deemed essential yet requires careful consideration of the array of challenges and gaps that currently exist (Chatterjee and Roy, 2022). For example, deficiencies in appreciation of realistic co-exposure scenarios, component interactions, mechanistic knowledge and grouping criteria may each impede progress (Bopp et al., 2018). Ensuring resolution of these issues will undoubtedly require "extensive strategic transdisciplinary initiatives", and as such it is inevitable that *in silico* approaches will be of immense value within mixture safety assessment (Drakvik et al., 2020). However, it is acknowledged that available QSAR workflows for the analysis of mixtures are insufficient (Muratov et al., 2012). To enable a better understanding of the state-of-the-art, this study presents a narrative review of the different QSAR approaches to predict mixture effects within chemical safety assessment (i.e., toxicological studies). Knowledge identified from the review can be utilised to supplement current QSAR uncertainty assessment schemes.

3.2. Materials and methods

3.2.1. Collection of literature

Literature relating to the use of QSAR for the assessment of mixture toxicity was identified using the Web of Science database. To ensure that all relevant work was captured, a broad search was conducted for studies from 1970 onwards. Keywords selected within the initial search (performed 25/10/2020) included "QSAR" and "mixture" – this returning 434 publications. The search criteria used resulted in many articles not relevant to this specific topic being identified. These were removed following manual screening of abstracts. Only articles focusing on QSAR development for mixtures were retained, so reducing the list to 134 taken forward for full text review (for graphical overview of workflow, please refer to Figure 3.1).

3.2.2. Compilation of information

A detailed analysis of the publications identified was undertaken, resulting in a further reduction of the number of articles for reasons including: unavailability of key information, models developed for single chemicals, non-toxicological endpoints, studies on essential oils/nanoparticles, and mixtures predicted solely through either concentration addition or independent action. Although CA and IA are both currently accepted methods used within regulatory approaches (European Commission, 2012a), the focus of the present study is upon QSAR protocols, and as such the decision was made to remove them. The final list comprised 40 studies, with these being additionally characterised with regards to: mixture composition (number of components, e.g., binary), chemical classification, taxa or testing system, endpoint examined, descriptors adopted (both class of, and conceptual approach applied in generation of mixture descriptors), and finally modelling or statistical technique applied. Table 3.1 contains an overview of the standardised terminology adopted relating to this characterisation.

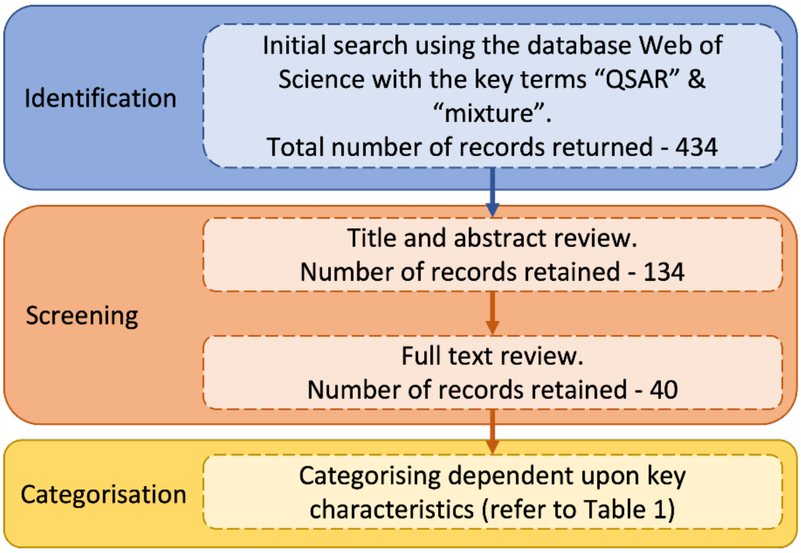

Identification
Initial search using the database Web of Science with the key terms "QSAR" & "mixture".
Total number of records returned - 434
Screening
Title and abstract review.
Number of records retained - 134
Full text review.
Number of records retained - 40
Categorisation
Categorising dependent upon key characteristics (refer to Table 1)

Figure 3.1. Overview of workflow adopted in the recovery and screening of literature for inclusion within this study.

Table 3.1. Summary of defined QSAR characteristics and the categories within

QSAR Characteristics	Categories
Chemical classification	Biocides, industrial, pharmaceuticals, priority pollutants
Mixture composition	Binary, ternary, quaternary, quinary, supra-quinary[1]
Taxa or testing system	Algae, amphibian, bacteria, cell line, embryos, insect
Endpoint	Acute, chronic, developmental, drug efficacy, growth inhibition, inhibition of reproduction
Descriptor formulation (approach)	Distribution coefficient, fragment non-additive, integral additive, integral non-additive, single variable component, structural similarity
Descriptor formulation (class)	Molecular docking, molecular fragment, molecular structure, physicochemical, quantum chemical
Modelling or statistical technique	CA and IA, CORAL, machine learning, partial order ranking, regression analysis, regression analysis (assumed)

[1] Mixtures containing greater than five components.

3.3. Results and Discussion

Evaluation of the literature resulted in identification of 40 relevant publications. As summarised in Table 3.2 and Figure 3.2, the majority of studies could be classified into groupings dependent upon methodology, endpoint, etc. The number of characteristics assigned for each grouping was not limited, with multiple classifications given where applicable. Further investigation of these characteristics has enabled the focus of current approaches to be outlined.

3.3.1. Chemical classification

Classes of chemicals considered in these articles could be classified broadly as belonging to one of four families: industrial chemicals (reported in 22 articles), pharmaceuticals (n=9), biocides (n=6) and priority pollutants (n=5). In general, the majority of articles related to environmental studies, including those for pharmaceuticals, with only a limited number of investigations considering human health effects. Future work into mixture assessments, therefore, should focus upon extending studies of the lesser examined groups, with a particular focus given to human health effects. Cell lines could provide a route towards realising this.

3.3.2. Mixture composition

Different varieties of mixtures were investigated, ranging from binary to complex. Binary mixtures made up the majority (n=38) of studies recovered, with comparatively few utilising multi-component combinations, i.e., ternary (n=10), quaternary (n=7), quinary (n=4) and the more realistic supra-quinary (n=3) – the latter term referring to those containing greater than five constituents. In addition to the number of components within the mixture, it is also important to consider the relative proportions of each, i.e., their ratios. Excluding supra-quinary, there are ten articles that investigated multi-component mixtures. Most of these were of fixed ratio design with some exceptions allowing varied ratios (Kar et al., 2018; Qin et al., 2018; Wang et al., 2018b; Kim et al., 2014; Lu et al., 2009; Duchowicz et al., 2008; Wei et al., 2004; Huang et al., 2003). Fixed ratio designs have been demonstrated as favourable within mixture studies, allowing for the distribution of the effect concentration range to be maximised, whilst additionally reducing number of experiments required (Kim et al., 2013b). Equitoxic ratios were most commonly used - this referring to mixtures where each component exists at the concentration that would result in identical effect if examined separately (Fulladosa et al., 2005). The likelihood of a mixture occurring naturally as equitoxic is very small, hence non-equitoxic ratios provide a more realistic representation (Warne, 2003). Additionally, it has been demonstrated, dependent upon the ratios of chemicals within a mixture, that the type of joint action observed can vary (Warne, 2003; Jin et al., 2014). As a result, studies involving the investigation into non-equitoxic mixtures can ensure that changes in joint action are captured.

Table 3.2. Summary and main characteristics of QSARs used in the mixture toxicity studies identified.

Chemical classification	Mixture composition	Taxa or test system	Endpoint	Molecular descriptor formulation		Modelling or statistical technique	Reference
				Conceptual approach	Descriptor class		
Biocides	Binary	Insect	Acute	Fragment non-additive	Molecular fragment	CORAL	Carnesecchi et al., 2020
Priority pollutants	Binary	Cell line	Acute	Integral additive	Molecular structure	Regression analysis	Hoover et al., 2019
Industrial	Binary	Bacteria	Acute	Integral additive	Molecular structure	Regression analysis	Chen et al., 2019
Industrial	Binary	Bacteria	Acute	Single variable component	Molecular structure	Regression analysis	Zhang et al., 2019
Biocides	Binary	Bacteria	Acute	Integral additive	Molecular structure	Regression analysis and machine learning	Wang et al., 2018a
Priority pollutants	Binary and ternary	Embryos	Developmental	Integral additive	Molecular structure	Regression analysis	Kar et al., 2018
Pharmaceuticals and biocides	Binary, ternary and quaternary	Bacteria	Acute	Integral additive	Molecular structure	Regression analysis	Qin et al., 2018
Pharmaceuticals	Binary and ternary	Bacteria	Acute	Integral additive	Molecular docking	Regression analysis	Wang et al., 2018b
Pharmaceuticals	Binary	Bacteria	Acute	Integral additive	Molecular docking	Regression analysis	Wang et al., 2018c
Pharmaceuticals	Binary	Bacteria	Acute and chronic	Integral additive	Molecular docking	Regression analysis	Wang et al., 2017
Pharmaceuticals	Binary	Bacteria	Acute	Integral additive	Molecular docking	Regression analysis	Long et al., 2016

Pharmaceuticals	Binary	Bacteria	Chronic	Integral additive	Molecular docking and physicochemical	Regression analysis	Fang et al., 2016
Priority pollutants	Binary	Cell line	Acute	Integral additive	Molecular structure and physicochemical	Regression analysis	Gaskill and Bruce, 2016
Industrial	Binary	Bacteria and algae	Acute	Integral additive	Quantum chemical	Regression analysis	Chang et *al.*, 2016
Industrial	Binary and ternary	Cell line	Organ-level effects	Unclear	Physicochemical	Regression analysis	Kim et al., 2014
Industrial	Binary	Bacteria	Acute	Single variable component	Quantum chemical	Regression analysis	Jin et al., 2014
Biocides	Supra-quinary	Bacteria	Acute	Structural similarity	Molecular structure	Machine learning and CA and IA	Kim et al., 2013b
Pharmaceuticals	Binary	Virus	Drug efficacy	Fragment non-additive	Molecular fragment	Machine learning	Muratov et al., 2013
Pharmaceuticals	Binary	Bacteria	Chronic	Integral additive	Molecular docking and physicochemical	Machine learning	Zou et al., 2013
Industrial	Binary	Not Stated	Chronic	Integral additive	Molecular structure and quantum chemical	Regression analysis and machine learning	Luan et al., 2013
Industrial	Binary	Bacteria	Acute	Single variable component	Physicochemical and quantum chemical	Regression analysis	Su et al., 2012
Industrial	Binary	Bacteria	Acute	Fragment non-additive	Molecular fragment	CORAL	Toropova et al., 2012
Priority pollutants and industrial	Binary	Not Stated	Acute	Integral additive	Molecular docking and physicochemical	Assumed regression	Wang et al., 2012

Pharmaceuticals	Binary	Bacteria	Acute and chronic	Integral additive	Molecular docking and quantum chemical	Assumed regression	Zou et al., 2012
Priority pollutants	Binary	Bacteria	Acute	Integral additive and distribution coefficient	Physicochemical	Assumed regression	Wang et al., 2011a
Biocides	Binary, ternary, quaternary and quinary	Embryos	Developmental	Unclear	Physicochemical	Regression analysis	Wang et al., 2011b
Industrial	Binary	Bacteria	Acute	Single variable component	Physicochemical and quantum chemical	Regression analysis	Su et al., 2010
Industrial	Binary, ternary and quaternary	Bacteria	Acute	Integral additive	Physicochemical and quantum chemical	Regression analysis	Lu et al., 2009
Industrial	Binary	Algae	Growth inhibition	Distribution coefficient	Physicochemical	Regression analysis	Zeng et al., 2008
Industrial	Binary, ternary, quaternary and quinary	Bacteria	Acute	Distribution coefficient	Physicochemical	Partial order ranking	Duchowicz et al., 2008
Industrial	Binary	Algae	Growth inhibition	Integral additive	Physicochemical and quantum chemical	Regression analysis	Wang et al., 2008
Industrial	Binary	Bacteria	Acute	Integral non-additive	Quantum chemical	Regression analysis	Zhang et al., 2007
Industrial	Binary, ternary, quaternary, quinary and supra-quinary	Bacteria	Acute	Integral additive	Physicochemical	Regression analysis	Wang et al., 2006
Biocides	Supra-quinary	Algae	Inhibition of reproduction	Structural similarity	Molecular structure	CA and IA	Mwense et al., 2006

Industrial	Binary, ternary, quaternary and quinary	Bacteria	Acute	Distribution coefficient	Physicochemical	Assumed regression	Wei et al., 2004
Industrial	Binary, ternary and quaternary	Amphibian	Acute	Integral additive	Physicochemical	Regression analysis	Huang et al., 2003
Industrial	Binary	Bacteria	Acute	Distribution coefficient	Physicochemical	Regression analysis	Lin et al., 2003
Industrial	Binary	Bacteria	Acute	Distribution coefficient	Physicochemical	Assumed regression	Lin et al., 2002
Industrial	Binary	Bacteria	Acute	Single variable component	Quantum chemical	Regression analysis	Yuan et al., 2002
Industrial	Binary	Bacteria	Acute	Distribution coefficient	Physicochemical	Regression analysis	Yu et al., 2001

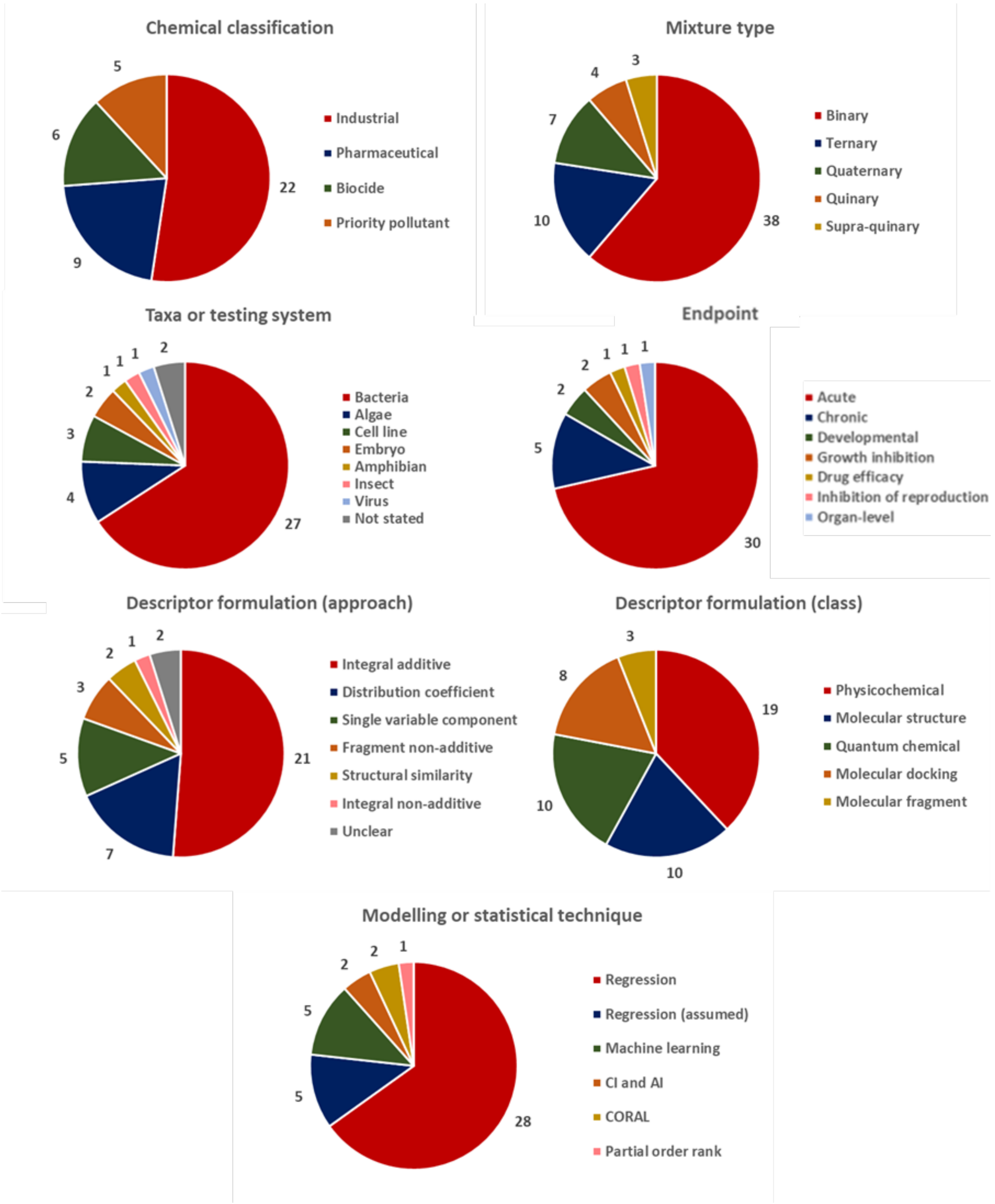


Figure 3.2. Quantification of features present amongst those parameters defining key QSAR characteristics.

Binary mixtures studies are limited to predictions of only binary combinations, unless validated otherwise. It is acknowledged that they may serve as an imperfect representation of real-world exposure scenarios (Kim and Kim, 2015). As such, the importance of developing models that can predict the effects of not only binary, but more importantly multi-component mixtures, is crucial. Nevertheless, assessments of binary mixtures can provide invaluable insights into methodology for modelling, as well as being utilised to gain information on mode of action (Hodges et al., 2006).

3.3.3. Taxa or testing system

A variety of species were used in the toxicological studies; however, the majority investigated bacterial-based bioassays (n=27). Within this group, use of bioluminescent bacterium *Aliivibrio fischeri* (formerly *Photobacterium phosphoreum*) predominated. Such tests are relatively inexpensive and enable large quantities of consistent data to be generated rapidly. Accordingly, they have been routinely employed as a first screening method within test batteries (Qu et al., 2013; Girotti et al., 2008). However, for these tests to effectively monitor an ecosystem, they must be used in combination with other biotests as well as chemical analysis (Girotti et al., 2008).

Various species other than bacteria have nevertheless been subject to investigation. Data from algae, cell lines (mammalian and amphibian), embryos, insects, amphibians, and viruses have all been used to develop mixture QSARs. Algal bioassays make up the second most common grouping (n=4), with testing upon algae providing an important insight into the balance of aquatic ecosystems as a result of them being primary food producers (Luan et al., 2020). Cell lines have been used in only a small number of studies, with such examinations potentially providing insight into specific simple mechanisms of interest. Cell line studies are an important testing procedure enabling the key processes towards a desired endpoint to be captured (Pistollato et al., 2020), however, the extrapolation of such information to entire organisms may prove difficult (Zucco et al., 1998). In general, QSAR models developed to investigate the toxicological effects of mixtures have focused upon environmentally-relevant species, with fewer considering human health.

3.3.4. Endpoint

The majority of toxicological endpoints for which mixture QSARs were developed related to acute effects. In total, 30 studies have investigated acute toxicity, in comparison to only a few chronic. Examination into the acute effects of chemicals can provide useful and fundamental information, with testing being comparatively simple, interpretable and high throughput. Moreover, such tests can enable underlying mechanisms of toxic action to be defined (Erhirhie et al., 2018). However, the use of acute toxicity data for QSAR modelling is not without its limitations. Adverse effects can result from an array of physiological, biokinetic, cellular and molecular events that span different levels of biological organisation. Measuring such complex systems in isolation will inevitably result in a loss of information (Lapenna et al., 2010). In comparison, toxicity following chronic exposure can better provide a realistic contribution to risk assessment of chemicals, particularly within environmental settings where organisms are exposed to the long-term effects of pollutants (Wang et al., 2017). However, knowledge of the chronic effects towards organisms of mixture exposure is sparse due to the intricacies of processes required for their determination – compounded by their duration and the costs of analyses (Zou et al., 2013). Accordingly, within the scope of this review, few studies utilised QSARs to predict chronic toxicity. However, a small number of successful applications have demonstrated that molecular docking based QSAR models may prove a valuable tool for predicting such endpoints (Zou et al., 2013; Fang et al., 2016; Wang et al., 2017). The current literature available for QSAR models for chronic mixture toxicity provides a solid foundation to be developed upon, with further research being required in areas of multi-component mixtures, as well as in higher-order species.

3.3.5. Mixture descriptor formulation

3.3.5.1. Conceptual approach

A fundamental distinction between the handling of single compounds and chemical mixtures when constructing a QSAR model lies in the nature of the descriptors which must be employed for each purpose. Whilst generation of molecular descriptors relating to discrete organic substances is generally a trivial process, provision of equivalents suitable for characterising mixtures is an issue of greater complexity. A variety of approaches are attested to within literature, based upon differing assumptions regarding the nature and relevance of interactions between member substances (Muratov et al., 2012).

3.3.5.1.1. Integral additive

The single most popular approach amongst those studies recovered (present within 21 of 40) – formation of integral additive descriptors, rests upon the intuitive premise that the properties of a mixture may be determined simply through summing those of its individual components – accounting for their relative prevalence and assuming occurrence of no meaningful interaction between each.

$$d_{mix} = \sum x_i d_i \qquad \textit{Equation 3.1.}$$

Where d_{mix} is a mixture descriptor, d_i the descriptor relating to chemical i, and x_i the fraction of the mixture composed by chemical i.

Application of the methodology in its simplest form is exemplified in the work of Huang et al. (2003), whereby toxicity of substituted phenol combinations is inferred solely through reference to a mixture octanol/water partition coefficient log_{kowmix} calculated via fractional addition of the log_{kow} belonging to each component. Versatility of the approach is such that there exist few limitations with respect to the nature of descriptors which may be used alongside it (refer to *Section 3.3.5.2* and Table 3.3 for examples). Accordingly, its adoption is noted in investigations employing molecular docking and quantum chemical techniques.

Despite widespread utilisation, shortcomings of this framework remain apparent. Disregarding of the potential impact of inter-component interactions (toxicodynamic, toxicokinetic or physicochemical) when inferring mixture adverse effects is most noteworthy amongst these. Such a limitation almost certainly renders it inapplicable for instances in which non-additivity is present – whilst in principle (despite favourable results) harming its capacity to model even general additive effects.

3.3.5.1.2. Integral non-additive

By contrast to the above, non-additive approaches envisage the mixture not merely as an agglomeration of mutually-inert components. Instead, they seek to integrate consideration of interactions existing between the molecules within – essentially modelling the mixture as a unit with bulk properties distinct to it (representing a more appropriate approximation of reality). Although appealing as a route towards addressing the issues inherent within additive methodologies, adoption has been limited.

A single study (Zhang et al., 2007) employing an integral, non-additive approach was retrieved. Within, toxicity of a series of binary 1:1 combinations consisting of simple substituted benzenes was modelled through use of quantum chemical descriptors. Properties of a mixture were derived through direct calculation of parameters of the appropriate pooled molecular pair – thus allowing for influence of electronic interactions between members to be accounted for. The rationale behind the lack of widespread uptake of this technique, despite conceptual promise, may lie in the restrictions placed upon its practical application: not only is scope of eligible mixtures constrained to those exhibiting 1:1 component ratio, but requirement to initiate unique calculations relating to each potential combination of substituents is potentially unwieldy.

3.3.5.1.3. Fragment non-additive

The non-additive principle is extended for application within fragment-based approaches to characterising activity of binary mixtures – forming the basis of three toxicologically-relevant studies. Whilst a thorough overview of core techniques is presented within *Section 3.3.5.2.4*, it is sufficient when considering generation of mixture descriptors to recognise the parallels which are present between this and "integral non-additive" methodology. In much the same manner, the molecular pair is treated as a unit. Individual fragments may incorporate atoms from either one or both components, and as such may provide descriptors relating both to individual compounds and to the unitary mixture.

3.3.5.1.4. Distribution coefficient-based

This approach remains suitable for instances in which activity of a mixture is modelled as a function of its partitioning between lipophilic and aqueous phases. Verhaar et al. (1995) reported derivation of a formula through which the distribution coefficient representing a mixture may be determined from those of its constituent chemicals.

$$K_{mix} = \frac{W}{V} \times \frac{\sum_{i=1}^{n} \frac{Q_{water,i}^{0}}{1 + (\frac{W}{VK_{SDi}})}}{\sum_{i=1}^{n} Q_{water,i}^{0} - \sum_{i=1}^{n} \frac{Q_{water,i}^{0}}{1 + (\frac{W}{VK_{SDi}})}} \qquad \textit{Equation 3.2}$$

Where K_{mix} is the lipoid/water partition coefficient of the mixture (substances such as n-octanol, chloroform and C18-Empore discs having been employed for this function), W the volume of the aqueous phase, V the volume of lipoid, $Q_{water,i}^{0}$ the initial amount of chemical i in water, K_{SDi} the partition coefficient of chemical i, and n the total number of chemical components in the mixture. Seven relevant studies adopting this approach were retrieved, with modifications to the methodology offered on occasion (please refer also to *Section 3.3.5.2.1*).

3.3.5.1.5. Single variable component

Each of the aforementioned techniques seeks to characterise toxicity of mixtures through consideration of the contributions of all substances within. However, there exist several studies (five retrieved from literature) in which activity is instead inferred through reference to properties of only a single constituent. In all instances, sequences of binary combinations were examined, whereby one component was held in common and the other was varied. Typical is the examination by Su et al. (2010), within which electronic and physicochemical parameters of a selection of substituted phenols were alone employed in order to model the toxicity of its mixtures alongside elemental lead. Whilst the majority of investigations have focused upon metallic-organic combinations, it should be noted that an early study by Yuan et al. (2002) featured solely organic components.

3.3.5.1.6. Similarity

A minority of studies adopt QSAR models not as a means of directly inferring the toxic potential of a mixture from the properties of its components, but instead as a means of assessing the similarity of screened compounds against those for which experimental data are present. Both Mwense et al. (2006) and Kim et al. (2013b) have put forward variations on this theme. Such similarity-based approaches enabled the mixture's components to be separated into clusters, which could then be subjected to CA and IA calculations (see *Section 3.3.6* for further information).

3.3.5.2. Descriptor class

Many different varieties of molecular descriptors exist, indicating the differing complexity levels of chemical structural representation (Cherkasov et al., 2014). In principle, any intrinsic molecular property appropriate for adoption as a descriptor within standard, single-component QSAR is further amenable to application within the domain of the mixture. As such, the range of properties referenced explicitly across the following subsections (on account of appearance within the existing literature) should not be taken as exhaustive.

3.3.5.2.1. Physicochemical

Considering the modelling of mixture toxicity, physicochemical descriptors have been employed from the very earliest studies. Of particular prominence are those based upon quantitative expression of the distribution of a substance between aqueous and representative lipophilic phases – this in short owing to their applicability in modelling compounds which exhibit a narcotic mode of action. Exemplified by logarithm of the octanol-water partition coefficient, these are acknowledged as being amongst the most effective general parameters to predict toxicity; having seen widespread use in many models for both single chemicals and mixtures (Kim and Kim 2015; Lin et al., 2002). It should be noted, however, that utility in handling toxicity mediated through means of chemical reactivity or receptor interaction may be diminished.

Application to mixtures is typically facilitated through adoption of one of two techniques introduced within *Section 3.3.5.1*: the dedicated method of Verhaar et al. (1995), or the more general integral additive approach. Employing the former, models were successfully developed to predict mixture toxicity of non-polar narcotic (Yu et al., 2001; Lin et al., 2002) and polar narcotic (Lin et al., 2003) chemicals. Following on, Wei et al. (2004) reported formulation of a simplified model demonstrating strong predictive power for both polar and non-polar mixtures. The aforementioned approaches have been limited to bacterial toxicity with regression-based models. However, additional studies have validated the methodology within algae studies, as well as with Partial Order Ranking methodology (Zeng et al., 2008; Duchowicz et al., 2008).

Considered by Roberts (1991) and by Altenburger et al. (2003), the employment of the integral additive approach towards formulation of mixture partition coefficients has since been demonstrated in various environmental studies (Huang et al., 2003; Wang et al., 2008; Lu et al., 2009; Wang et al., 2006). One of the few studies to compare both Verhaar and integral additive methodologies directly was completed by Wang et al. (2011a), in which the mixture toxicity of perfluorinated carboxylic acid was assessed. Results demonstrated that the equivalent Verhaar-adapted approach provided, in this instance, the better results for describing the hydrophobicity of mixtures.

3.3.5.2.2. Molecular docking

Information gathered from molecular docking of chemicals into receptors has been used routinely, particularly as a drug discovery tool enabling the early identification of potentially active candidate molecules. These techniques facilitate the development of mechanism-based models, with interactions between chemicals and receptors being simulated. Specifically, such studies could relate to receptor-mediated molecular initiating events (Cronin and Richarz, 2017). These simulations enable the interaction energy required for a chemical to bind to its target protein ($E_{binding}$) to be determined (Rabinowitz et al., 2008). In each of the examples subsequently presented, $E_{binding}$ relating to individual components are summed to form mixture descriptors through adoption of the integral additive approach.

Wang et al. (2012) were amongst the first to propose the use of binding energy descriptors in modelling mixture toxicity – examining the feasibility of substituting $log\, K_{owmix}$ with the molecular docking descriptor $E_{binding}$, owing to the linear trend observed between the two. Zou et al. (2012) investigated both the acute and chronic toxicities of antibiotics from the sulfonamide family, alongside the sulfonamide potentiator trimethoprim. The study initially identified the receptors responsible for both their acute and chronic effects towards *Aliivibrio fischeri*; determining them to be luciferase, dihydropteroate synthase and dihydrofolate reductase. Models using the binding energies towards each protein, supplemented by pKa, were shown to successfully predict the toxicities of mixtures for both exposures. Further to this study, Zou et al. (2013), employed docking in order to curate a library of simulated antibiotic-receptor interactions, spanning several prominent mechanisms of action. Through this, the ready construction of mechanistically-grounded QSAR models relevant to a wide range of potential antibiotic combinations was facilitated.

More recently, Wang et al. (2017) also investigated chronic effects of antibiotics. A mechanism-based QSAR model was developed whereby the chronic toxicity of sulfonamides, sulfonamide potentiators, and tetracyclines could be extrapolated from acute toxicity. Unlike previous extrapolation models, understanding of the differing toxic mechanisms between acute and chronic exposures was considered. In a variation from Zou et al. (2012), in which DHFR (Dihydrofolate reductase) served as the sole mediator of TMP (Trimethoprim) toxicity, the targets for the antibiotics reported in this study were represented by surrogate luciferase proteins. Due to a specific target not being considered and instead characterised by surrogates, the model demonstrated promise in predicting the toxicity of chemicals for which mechanisms are unknown.

Molecular docking studies have introduced new concepts to the field of QSAR mixture toxicity. Fang et al. (2016), Long et al. (2016) and Wang et al. (2018b) developed mechanistic models derived from binding energies of antibiotics towards target proteins from which they were able to theoretically identify the effective concentration of the mixtures. Wang et al. (2018b) also proposed equivalent findings but included ternary mixtures. Each study incorporated terms describing the extent to which each specific component contributed towards protein binding, i.e., the effective concentration. Wang et al. (2018b) further commented upon this, stating that such terms could be interpreted as representing the processes of a component passing through the cell membrane and reaching its target protein. Thus, the component which had a higher probability of interacting with its target protein could be identified depending upon the value of the coefficient attached to the term. The authors utilised this knowledge to enable calculation of the actual toxicity ratio – a value which was subsequently used to aid in determining which component had the greater contribution to toxicity.

Wang et al. (2018c) further employed docking techniques in the investigation of mixture effects of the recently popularised antibiotic alternative - quorum sensing inhibitors (QSIs). However, current research remains largely focused upon simple binary mixtures of antibiotics – with only Wang et al. (2018b) extending examination into multi-component mixtures. It is further noted that existing studies have yet to integrate consideration of mixture toxicokinetics in a manner which would allow conclusions to be drawn regarding likely absolute exposure of targets to components.

3.3.5.2.3. Molecular structure

Structure-based descriptors (otherwise known as 2D or topological), provide simplistic, interpretable information about molecular structure, as well as being easy and quick to generate (Cherkasov et al., 2014). A variety of software is available to calculate these parameters, with DRAGON (previously available at: https://chm.kode-solutions.net/pf/dragon-7-0/) used in several reported mixture studies. DRAGON software calculated over 5,000 molecular descriptors, with these being organised into logical blocks. A range of different blocks exists, with these including, but not limited to: constitutional, ring descriptors, topological indices, walk and path counts, and connectivity indices. The DRAGON software was used to obtain descriptors in six studies identified in this analysis, as summarised in Table 3.3. Due to the range of chemical mixtures and species examined within, it is inevitable that a variety of descriptors, with varying degrees of mechanistical interpretability, were used. For example, Chen et al. (2019) and Zhang et al. (2019) both utilised edge adjacency indices derived from H-depleted molecular graphs. Both studies utilised toxicity data for bioluminescent bacteria, with Chen et al. (2019) investigating aromatic halogenated chemicals and Zhang et al. (2019) nitro-substituted benzenes and zinc. These parameters were successful in both instances, additionally proving their worth within mixtures of different mixing ratios (Zhang et al., 2019). Gaskill and Bruce (2016) further found that information indices were able to predict mixture toxicity. The authors developed various models to predict impact of polycyclic aromatic hydrocarbon mixtures towards liver cells, with additional topological descriptors being utilised. These topological descriptors, particularly with respect to planar PAHs (Polycyclic aromatic hydrocarbons), proved to be significant in predicting effects, highlighting the role planar characteristics and bond orientation play in causing toxicity.

Table 3.3. Descriptors calculated using DRAGON software, displayed within their respective blocks.

Descriptor	Title	Block	Publication
piPC06	Molecular multiple path count of order 6	Walk and path counts	Hoover et al., 2019
Mor12m	Signal 12 / weighted by mass	3D-MoRSE descriptors	Chen et al., 2019
Mor13s	Signal 13 / weighted by I-state	3D-MoRSE descriptors	
L/Bw	Length-to-breadth ratio by WHIM	Geometrical descriptors	
Eig08_EA(ed)	Eigenvalue n. 8 from edge adjacency mat. weighted by edge degree	Edge adjacency indices	
Eig09_EA(ed)	Eigenvalue n. 9 from edge adjacency mat. weighted by edge degree	Edge adjacency indices	
Eig09_AEA(dm)	Eigenvalue n. 9 from augmented edge adjacency mat. weighted by dipole moment	Edge adjacency indices	
RDF045s	Radial Distribution Function – 045 / weighted by I-state	RDF descriptors	
J_RG	Balaban-like index from reciprocal squared geometrical matrix	3D matrix-based descriptors	
VE2_B(p)	Average coefficient of the last eigenvector from Burden matrix weighted by polarisability	2D matrix-based descriptors	Zhang et al., 2019
TIC3	Total Information Content index (neighbourhood symmetry of 3-order)	Information indices	
Eig06_AEA(dm)	Eigenvalue n. 6 from augmented edge adjacency mat. weighted by dipole moment	Edge adjacency indices	
PJI2	2D Petitjean shape index	Topological indices	Kar et al., 2018
$^{2}\chi^{v}$	Valence connectivity index of order 2	Connectivity indices	
$^{0}\chi^{v}$	Valence connectivity index of order 0	Connectivity indices	
RDF035m	Radial Distribution Function - 035 / weighted by mass	RDF descriptors	Qin et al., 2018
HATSs	Leverage-weighted total index / weighted by I-state	GETAWAY descriptors	
H-047	H attached to $C^1(sp3)/C^0(sp2)$	Atom-centred fragments	
	Independent components[a]	N/A	Mwense et al., 2006

3.3.5.2.4. Molecular fragments

Fragment-based descriptors have been described as a promising method for the QSAR modelling of mixtures (Cherkasov et al., 2014). However, there are relatively few examples of their use in practice. Muratov et al. (2013), predicted combination effects of antivirals against poliovirus-1 through use of Simplex Representation of Molecular Structure (SiRMS) - a framework which enables molecular structures to be represented as a system of simplexes (tetratomic fragments), capable of capturing features at the topological level. Modifications to the approach were undertaken to enable extension for analysis of binary systems, generating descriptors applicable either to single components (bounded simplex), or else drawing elements from across both (unbounded simplex). The latter can be considered as structural descriptors of the mixture as a unit and as such "non-additive". Whilst this approach is highly desirable, that no other recent toxicological report has utilised this methodology suggests that it may only be applicable within certain cases.

Other fragment-based descriptors were utilised by Toropova et al. (2012), who demonstrated the ability of the CORAL software (http://www.insilico.eu/coral) to again predict toxicity of binary mixtures. Molecular structures of components were represented by SMILES, using a disconnected approach with a marker (i.e., ".") separating each string. Recently, Carnesecchi et al. (2020) further extended this approach, making use of expanded "quasi-SMILES". In this case, the toxic units of each chemical in the binary mixture are incorporated. A classification model predicting potential for non-additivity (either synergism or non-synergism) was simultaneously reported. Results obtained indicated that consideration of toxic units not only enabled greater interpretability of the models, but also improved the statistical performance. In general, models developed by the CORAL software enable frequently occurring molecular features that cause binary mixture toxicity to be identified. However, studies thus far using these procedures (and SiRMS) have only been limited only to binary mixtures.

3.3.5.2.5. Quantum chemical descriptors

Quantum chemical descriptors are able to describe the electronic and geometric properties, and interactions, of molecules. Although potentially intensive as regards demands upon computational power and running time, they offer greater detail with respect to electronic effects than do traditional empirical methods (Karelson et al., 1996; Schüürmann, 2004). The most commonly applied quantum chemical descriptors utilised for modelling mixture toxicity were the molecular orbital energies, with energy of the lowest unoccupied molecular orbital (E_{LUMO}), or slight adaptions, being routinely used. This metric accounts for the electrophilicity of a molecule (Schüürmann, 2004), correlated as it is to its electron affinity. Studies extended this parameter to multi-component mixtures (Lu et al., 2009), and the variation $E_{LUMO} + 1$ (energy of the second lowest unoccupied molecular orbital), in combination with total charge weighted partial positive surface area (PPSA), have proven superior to previous hydrophobicity-dependent QSARs for non-polar narcotics (Luan et al., 2013). Additionally, the difference between the lowest and highest frontier molecular orbitals, i.e., $E_{LUMO} - E_{HOMO}$, or *vice versa*, have been proven effective in mixture calculations. Wang et al. (2008) first used this parameter, which is able to determine the stability of the molecule, collectively within a traditional hydrophobicity-based model to enable better predictions of the joint toxicity of polar narcotics.

In each of the aforementioned instances, orbital mixture descriptors were generated through integral additive means. Quantum chemical descriptors have, however, additionally found employment in a distinct collection of studies introduced within *Section 3.3.5.1.5*, under the heading "Single variable component". A typical example is provided through Jin et al. (2014), whereby models are created considering the energy difference between molecular orbitals- a parameter termed the relative hardness index ($E_{HOMO} - E_{LUMO}$). Multi-pointwise toxicological models (i.e., approaches for mixtures predicting varying effect concentrations) are an under-researched area, although interestingly an additional report studying them, that of Su et al. (2012), did employ quantum chemical descriptors. Within the joint toxicity of nitroaromatics with copper at low, medium, and high concentrations was modelled. The results were similar to those of Jin et al. (2014), in that varying the concentrations of the components played a pivotal role on the joint effects within the mixture.

Currently, the majority of literature describing use of quantum descriptors is focused exclusively on single mixture ratios - typically equitoxic. Realistically-encountered combinations of molecules are expected to deviate from this ideal, thus suggesting that a range of compositions would provide for stronger predictions. These studies, furthermore, concentrate almost exclusively upon industrial compounds – thus serving only a restricted area of chemical space.

3.3.6. Methods for model development

A variety of statistical approaches were reported across the reviewed literature with regression analysis dominant. Comparatively simple to establish and interpret, regression has been the classical approach in QSAR modelling since its inception. It is, however, not without limitations, with consideration of parameter collinearity required in order to ensure that robust models are developed (Lo et al., 2018). As an alternative, machine learning approaches permit nonlinear relationships to be better modelled, which is attractive in mixture toxicity due to the varying nature of underlying combination effects. Two studies developed models using both regression and machine learning, enabling direct comparisons between the performance of both. Results suggested that machine learning approaches, specifically radial basis function neural networks, enable improvements in statistical fit (Luan et al., 2013; Wang et al., 2018a). Although, machine learning is a current trend in the area of *in silico* prediction, it is not without its limitations: ensuring that models are well established typically requires a high volume of data. Potential for overfitting must be taken into account, and difficulties in interpretation, owing to their black box nature, typically hinder derivation of mechanistic knowledge (Lo et al., 2018).

Whilst studies incorporating exclusively either CA or IA (first generation) are considered beyond the scope of this review, a small quantity of second-generation models are eligible for inclusion on account of their integration of QSAR methodology. Each of the following techniques may be distinguished by the conditional adoption of CA or IA in modelling of inter-component interactions, dependent upon the extent of similarity either in molecular structure or mode/mechanism of action between substances. As such, the combined toxicity of like compounds is determined through the principle of CA, and dissimilar through IA – with ultimate mixture effect being derived from the contributions of both. Mwense et al. (2004) introduced an approach termed INtegrated Concentration Addition-Independent action Model (INFCIM), whereby this similarity was determined using computed molecular descriptors. The following equation was employed to calculate overall toxicity:

$$EC_{x,mix} = \omega_A \cdot (CA) + \omega_B \cdot (IA) \qquad \textit{Equation 3.3.}$$

where coefficients ω_A and ω_B are the weightings for the contributions of CA and IA.

Although this initial model had no theoretical capabilities to provide predictions that would exceed concentration addition, the model was later revised in order to address these limitations (Mwense et al., 2006). Analogously, Kim et al. (2013b) developed an approach which incorporated both CA and IA known as a two-stage prediction model. Unlike previous two-stage prediction models which relied on knowledge of modes of toxic action for all components, the authors utilised machine learning clustering techniques to group the constituents – employing CA within-group (stage 1) and IA between-group (stage 2) in determination of absolute mixture effect. Excellent performance against realistic environmental mixtures was reported, highlighting the possibility of success even in the absence of mechanistic information. Such models, however, remain at present limited to non-interacting mixtures.

3.3.7. Uncertainty criteria and assessment for mixture studies

The assessment of chemical mixtures by means of QSAR methodologies is continually generating greater interest. In ensuring that such work is up taken in regulatory settings, it is essential that potential uncertainty associated with models are defined. Cronin et al., (2019) recently developed a set of criteria that enabled the full assessment of QSAR models from conception to application, facilitating all aspects of uncertainty to be defined and scored. This was further expanded upon by Belfield et al. (2021), where it was demonstrated that the criteria could also be employed to determine fitness-for-purpose. Although these criteria have been developed in order to account for all potential usages of QSAR, completion of the present literature review has elucidated further areas of consideration specifically relevant to construction of QSAR models for prediction of mixture effects. As such, areas have been identified that can be bolstered with lessons learnt to improve the assessment of QSARs for mixtures. Specifically, it can be defined that these additional considerations relate to chemical description, descriptor calculation, and statistical performance. These are discussed below and reported in Table 3.4 – with accessory detail provided in *Appendix II*.

Firstly, worthy of note is that within the current structure of the QSAR uncertainty criteria, the consideration of chemical mixtures is approached (as clearly defined under criterion 1.1b – "Assessment of significant impurities or mixtures"). However, unambiguous guidance ought to be provided for the assistance of users unfamiliar with mixture handling. To ensure that scorings are assigned correctly, further information on what is to be expected is suggested within the comment section, as seen in Table 3.4. Not only is it vital that all components within mixtures are fully identifiable, but additionally that the proportion represented by each must be reported. Clearly, measured endpoints will be dependent upon the ratio at which mixtures are investigated, but such information is additionally required to enable accurate calculation of mixture descriptors. Omission of mixture ratios will therefore restrict external reproducibility. Further to this, and in a similar vein (although not discussed further in the present review), guidance to correct reporting techniques for substances of Unknown or Variable composition, Complex reaction products or Biological materials (UVCB) as detailed by the European Chemicals Agency (ECHA) are provided (ECHA, 2017b).

Arguably, the most important aspect that changes from modelling single chemicals to mixtures is the handling of descriptors. An entire section of the original uncertainty criteria has been devoted to the consideration of the varieties of descriptors a user may employ (this being 1.3 – "Measurement and/or Estimation of Physico-Chemical Properties and Structural Descriptors"), yet methodologies to convert such features into mixture descriptors are needed. As reviewed in *Section 3.3.5* many approaches are used to define mixture descriptors. Selection of the correct method in characterising these is not only dependent upon the type of descriptors chosen (such as fragment-based, compared to physicochemical), but additionally by the interaction effects within the mixture. Capturing such complex processes and concerns by updating comment guidance to existing criteria would clearly be insufficient; thus, an additional topic must be supplied to fulfil the need. The current structure of the criterion 1.3 enables all plausible descriptors to be considered, relying upon user discretion to evaluate only relevant features that have been employed. As such, supplementing a new point into this section will not alter the validation process, but instead extend applicability of models that may be evaluated. A further criterion 1.3d ("Calculation of mixture descriptors, if utilised") is proposed that will enable the uncertainty level of mixture descriptors to be defined. The main aspect needed to satisfy this recommended criterion is that the selected approach has been derived through thorough consideration of potential interaction effects. Calculating these effects is a topic well studied, with a variety of methods alluded to in the comments for user guidance.

The final section that would benefit from further guidance relates to external validation. Within QSAR modelling, exhaustive validation is required to ensure that predictive performance is correctly evaluated. However, compared to that of traditional QSAR procedures, validation methods for mixtures require further deliberation. Mixtures present further challenges whereby the same components may exist inside different mixtures. Splitting the dataset without consideration of this fact will undoubtedly result in datapoints from the same mixture appearing within both training and testing sets, thus resulting in over-optimistic estimations (Muratov et al., 2012). To combat such occurrences, various strategies have been developed, namely: “points out”, “mixtures out”, “compounds out”, and “everything out” (for detailed discussion of these, please refer to Oprisiu et al., 2012 and Muratov et al., 2014). Validating mixture models without consideration of these facts will certainly affect the legitimacy of predictions, as well as the associated uncertainty. As selection of appropriate validation methods is already well defined within criterion 2.2a (“Statement of statistical fit, performance and predictivity”), providing further guidance under the “comment or other information” heading will ensure that mixture strategies can be fully considered.

Table 3.4. Specific assessment points from the uncertainty criteria (previously discussed in Section 3.3.7 and originally presented in Cronin et al., 2019) that require further guidance for the assessment of mixture-based studies and their proposed updated guidance. Updates to text under heading “comment or other information” are displayed in italics. Please refer to *Appendix V* for presentation in context of unabridged scheme.

ID	Assessment criteria	Comment or other information
1.1b	Assessment of significant impurities or mixtures	*If mixtures are being modelled, each component needs to be fully identified and defined with respect to concentration present. For substances of Unknown or Variable composition, Complex reaction products or Biological materials (UVCBs) see European Chemicals Agency (2017b).*
1.3d	*Calculation of mixture descriptors, if utilised*	*Interaction effects can be identified through various methods (TU etc.) with this aiding in developing appropriate mixture descriptors for the model*
2.2a	Statement of statistical fit, performance and predictivity	The use of appropriate validation methods and/or external test sets should be demonstrated, different metrics may be required for different models. *In regard to the assessment of mixtures, external validation must*

		consider more rigorous strategies such as: "points out", "mixtures out", or "compounds out" (Muratov et al.*, 2012)*

3.4. Key Findings

The purpose of the current review was not only to identify current trends in QSAR mixture modelling, but also to determine whether existing modelling practices are sufficient to accurately address issues that mixtures present. Regardless of the source of the model or modelling approach, a number of commonalities can be identified. These form a general appraisal, or overview, of the state-of-the-art of QSAR mixture modelling:

3.4.1. Need for QSAR models

- Modelling is a vital approach to assess the toxicity of mixtures. It is inconceivable that all possible combinations of chemicals (and at varied ratios) can be experimentally measured. Therefore, there needs to be a much greater emphasis on modelling approaches for mixture toxicity. A particular direction of interest for the modelling of mixtures would be through the employment of graph neural networks (GNNs). Such methods have gained recent popularity due to their ability to learn molecular representations in the form of graphs bypassing the need to manually generate descriptors (Wang et al., 2023). Utilising GNNs to model mixtures therefore would enable the opportunity to incorporate molecular interactions within the model architecture itself, avoiding the need to generate mixture descriptors (Qin et al., 2023).

3.4.2. Need for proper problem formulation

- Much of the current modelling of mixture toxicity has been performed on an *ad hoc* basis. There needs to be greater organisation of these modelling studies to make them realistic of real-life exposures and able to address the problems associated with ensuring environmental and human safety. Utilising the uncertainty criteria proposed by Cronin et al. (2019), with guidance previously suggested, would provide a rational foundation for addressing such issues.

3.4.3. Availability of data for modelling

- This review has demonstrated the paucity of data available for mixtures. Repositories such as PubChem (https://pubchem.ncbi.nlm.nih.gov/), ChEMBL (https://www.ebi.ac.uk/chembl/), DrugBank (https://go.drugbank.com/), IPCheM (https://ipchem.jrc.ec.europa.eu/) and ChemTHEATRE (https://chem-theatre.com/) have been postulated to resolve this issue, yet collating a reliable dataset from such sources is currently unfeasible (Muratov et al., 2012). As such, gathering a larger dataset would likely be reliant upon literature, with the current review highlighting a breadth of publications containing compatible information. It is evident that not only is more data required, but that a more systematic means of storing, distributing and retrieving these data is also essential.

3.4.4. Understanding data relevance and quality

- There must be greater appreciation of what types of study are useful to assist in environmental risk assessment and will assist in the characterisation of real-life exposure scenarios. Linked to this is the lack of assessment of data quality, with few of the studies being performed to OECD Guidelines or Good Laboratory Practice. If future testing materialises, then there should be a greater emphasis on determining the relevance of experimental studies and ensuring that their quality is suitable for all purposes, including regulatory adoption.

3.4.5. Identification and incorporation of interaction effects into QSAR models

- As yet, there is no consensus on how to approach the inclusion of interaction effects, where they exist, into QSAR models. A better and more complete understanding is required of whether we need to go beyond the typical additive approach. One place where such knowledge could be identified and compiled is via a more extensive review and compilation of drug interaction effects. In addition, there could be a greater understanding and application of our knowledge of mechanisms of toxic action, particularly for acute environmental toxicities. Linked to this, there are obvious opportunities to incorporate knowledge and understanding from Adverse Outcome Pathways (AOPs) into our schemes (Cronin and Richarz, 2017). In particular, the insight gained from the structure of AOPs can supplement the understanding of how mixture components interact. This knowledge can then be used to identify specific key events of interest, or alternatively, provide better informed mixture approaches that are to be employed (Lambert, 2023; Nelms et al., 2018).

3.4.6. Modelling approach (descriptors and statistical methods)

- Models identified within this review used the full range of QSAR descriptors from physicochemical properties to 2D and quantum chemical calculations. There is no ideal descriptor for use in a mixture QSAR study, but those chosen should be pragmatic and give credibility to the model, notably by allowing full mechanistic interpretation. Ideally such descriptors should be simple, unambiguous and easy to calculate. Likewise, there is no consensus on how descriptors can be formalised to account for the mixture contributions and constitution.
- Statistical approaches applied in development of models for mixture toxicity range from simple regression analyses to machine learning. No ideal technique can be recommended at this time. It is appreciated that as the mixtures become more complex, there is likely to be a greater need to adopt machine learning approaches. Whilst rapid and potentially accurate, these typically lack transparency and interpretability, in turn hindering uptake and acceptance.
- A possibility that has yet to be explored fully in terms of mixture toxicity modelling is use of read-across such that effects and even potency may be established from similar or analogous mixtures. Such approaches have seen great acceptance for single chemicals and are increasingly being considered for botanical substances, natural products and UVCBs.

3.4.7. Towards a unified approach to model meaningful effects for realistic environmental and other mixtures

- Many currently available mixture toxicity QSAR models have limited practical application towards realistic exposure scenarios. Despite this, they have provided a wealth of knowledge on which we can build new frameworks and approaches to model such endpoints. Given the possibilities and the appreciated challenges associated with modelling toxicity, there is a great need to develop a unified approach to understanding its application towards mixtures, alongside practical means to developing, evaluating and applying such models to realistic environmental exposures of relevant chemical combinations.

3.5. Conclusion

The present review has provided a detailed analysis of the differing approaches that have been used throughout QSAR model development to predict the effects of mixtures. In general, reoccurring trends presented themselves throughout toxicological-based publications, in which binary mixtures at a single concentration ratio have been examined in an additive manner. In addition, molecular descriptors have commonly been employed to describe the mixtures using molar weightings, and resulting QSAR models are traditionally developed using regression analysis. The overwhelming majority of research on mixtures has been conducted for environmental effects, while other fields, for instance human health, have been understudied. It is expected that to increase the uptake of QSAR predictions, greater respect for potential interaction effects should be considered, although firstly, it is imperative that current modelling practices are to be extended enabling the assessment of realistic mixture scenarios. In general, research up to the current time has provided an excellent foundation, where future work that addresses current limitations may not only improve uptake of predictions, but additionally increase our knowledge in the field of mixture studies. Expanding upon this, the viability of the uncertainty criteria to evaluate QSAR models for the prediction of mixtures has been shown. Inclusion of supplementary considerations have demonstrated the ease and flexibility of the criteria to be able to capture the additional areas that a mixture study presents enabling a thorough assessment. Whilst the current study has proven the ability of the uncertainty criteria (introduced in Chapter 2) to be extended to various types of data, it has also highlighted that there is a need for more complex modelling procedures, such as AI; this is the subject of the next chapter.

Chapter 4. Good practice for machine learning methods in predictive toxicology

Preface:

This work has been published in Belfield SJ et al., (2023). Guidance for good practice in the application of machine learning in development of toxicological quantitative structure-activity relationships (QSARs). PLoS ONE. 18: e0282924. https://doi.org/10.1371/journal.

pone.0282924

This was a multi-author paper. Belfield led the work and analysis in this study as recognised in the CRediT authorship contribution statement: Conceptualisation, Data Curation, Investigation, Methodology, Visualisation, Writing – Original Draft.

4.1. Introduction

The use of computational approaches to predict adverse effects in toxicology, to support chemical safety assessment, has become standard practice. Quantitative structure-activity relationships (QSARs) are one of the most well-established methods within the field of *in silico* toxicology and, as such, have been used extensively to identify hazard and predict potency (Cherkasov et al., 2014). QSAR models attempt to formalise the relationship between descriptors calculated from chemical structures or physico-chemical properties and the desired endpoint (Madden et al., 2020). Traditional QSAR modelling was predominantly based around regression analysis, however as far back as the 1980s a variety of other multivariate statistical approaches were being applied (Wold and Dunn, 1982), with the uptake of neural networks in the early 1990s (Rose et al., 1991). The past decade has seen a much greater shift towards machine learning (ML) strategies to develop models in predictive toxicology. There is no one reason for the increased use of ML, but increased availability of data, more easily accessible informatics and statistics tools, as well as greater computational power have all contributed.

ML methods originated in the early to mid-20th century from mathematical considerations of data matrices. More recently, ML approaches are considered to be a subset of artificial intelligence (AI), which broadly refers to computational systems that are able to mimic human intelligence (Robinson and Akins, 2021). Since their conception, ML techniques have been developed within the field of computer science and have been identified as one of the most vital and rapidly evolving areas in chemoinformatics (Varnek and Baskin, 2012). Emerging from pattern recognition studies and the concept of computational learning, ML algorithms can learn and adapt without being explicitly programmed to do so, thus, in turn, improving the accuracy of generated predictions (Barros et al., 2020). ML methods can be broadly separated into two classes, either supervised or unsupervised. In this regard the majority of QSAR applications apply supervised learning approaches, where the data are labelled such that both the chemical information and investigated property are known, in contrast to unsupervised techniques in which patterns are identified from unlabelled data (Gini and Zanoli, 2020; Lo et al., 2018). Many ML approaches have been applied, with the main strategies in QSAR reviewed by Lo et al. (2018), ML methods that have been employed in QSAR are summarised in Figure 4.1, and described in more detail in Section 4.2. Of these approaches, it is deep learning (DL) that has captured the imagination and has been identified as one of the most exciting ML strategies of the past few years, with these utilising multiple layers of interconnected neural networks to self-train (Robinson and Akins, 2021; Muratov et al., 2020). DL has widespread applications in many areas of research, such as computer vision, speech recognition among many others (Hochreiter et al., 2018; Mater and Coote, 2019). Although the concepts of DL have been around for many years, applications to QSAR mainly began after the approaches were employed to win the Merck Molecular Activity Challenge in 2012 (Merck, 2012). As a result of the team's usage of DL to outperform other methods in the QSAR challenge, a renewed interest in the approaches was observed (Muratov et al., 2020; Dahl et al., 2014).

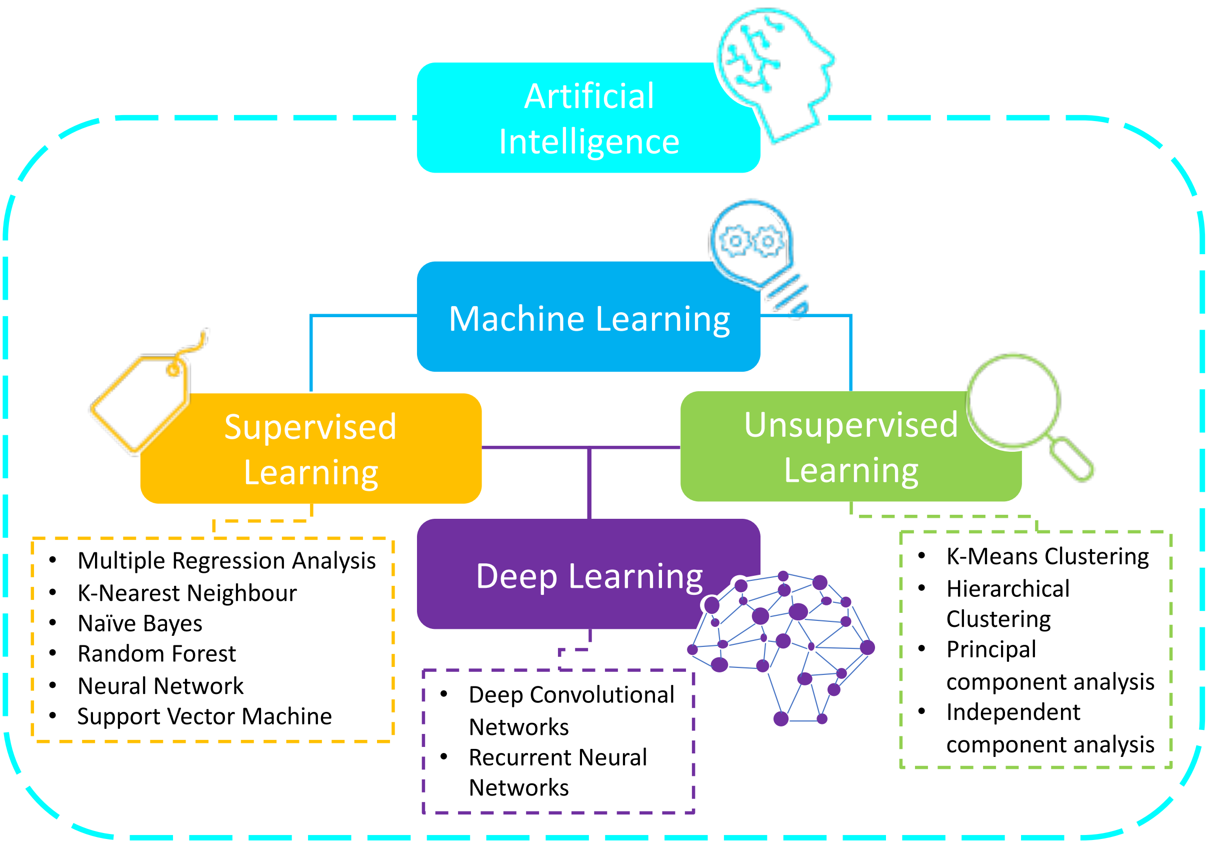


Figure 4.1. General approaches encompassed within the umbrella of AI and ML that are relevant to predictive toxicology.

There are many potential uses for *in silico* approaches to predict toxicity. These range from the rapid screening of large chemical libraries and inventories to the identification of potential hazards contributing to risk assessment of individual compounds by either providing a replacement for a test or contributing to a weight of evidence. A key aspect of the use of QSAR models to predict toxicity is the acceptance of the predictions for a particular purpose, with different characteristics of QSAR models being associated with different uses (Belfield et al., 2021). Regarding the legal interpretation of legislation such as EU Regulation on Registration, Evaluation, Authorisation (restriction) of Chemicals (REACH), there is a strict requirement that the prediction should provide the same information as the test it is replacing (the so-called process of adaptation of a testing requirement). To achieve this, amongst other criteria, the model must be shown to be “scientifically valid”. This is currently achieved using approaches to evaluate QSARs, such as the OECD Principles for the Validation of QSARs (OECD, 2007). However, ML models of toxicity can be difficult to evaluate with these principles as they are perceived to lack: 1) a defined and transparent algorithm as compared to regression analysis (OECD Principle 2), 2) mechanistic interpretability (OECD Principle 5) and 3) conclusive documentation. Specific issues regarding the application of ML to predict toxicity for regulatory use also includes overfitting (Ying, 2019). These may have a significant impact on the acceptance of models and their predictions.

To further support the growth of ML in the field of QSAR, considerations of the challenges faced need to be addressed. Recent work on uncertainty assessment of QSARs, which is based around the OECD QSAR Principles, could provide a different insight into ML models for toxicity prediction (Cronin et al., 2019). The use of uncertainty was intended to be applied to provide assessment schemes to enable authors/users to understand strengths and limitations of predictive toxicology models. Although the current scheme provides applicability for a vast range of QSAR modelling practices, additional supplemental guidance for the specific consideration of ML methods will undoubtedly provide greater confidence in such inherently "difficult to interpret" models.

The aim of this investigation was to identify good practice in ML methods for predictive toxicology with a view to improving their acceptance. To achieve this, two toxicity datasets with potency data of varying complexity and quality were modelled. Modelling was undertaken using differing ML algorithms that had been produced with state-of-the-art optimisation and interpretability techniques. Good practice for ML modelling in predictive toxicology was identified following evaluation of the models, supplemented through consideration of the key uncertainties which were characterised according to, and thus in turn extending, the scheme published by Cronin et al. (2019).

4.2. Methods

4.2.1. Data curation

Two data sets were assessed in this analysis. With regard to QSAR modelling, the datasets represent relatively large compilations with one for a cytotoxicity endpoint in an aquatic ciliated protozoan and the second acute rodent toxicity.

4.2.1.1. Inhibition of growth to Tetrahymena pyriformis dataset

The *Tetrahymena pyriformis* dataset was harvested from Ruusmann and Maran (2013). This publication collated and curated data relating specifically to the acute toxicity of compounds towards the aquatic ciliated protozoan *Tetrahymena pyriformis* as performed and reported in a plethora of publications by Prof Terry Schultz, University of Tennessee, Knoxville TN, USA (a general description of the method is provided by Schultz (1997)). Explicitly, the toxicity endpoint used was *Tetrahymena pyriformis* population growth inhibition, expressed as the inverse logarithm on the millimolar concentration that caused 50% growth inhibition after 40 hours (log 1/IGC50). In total, data for 2,072 substances were retrieved from Ruusmann and Maran (2013). These data were reduced to 1,995 substances following the removal of duplicates. Lastly, SMILES for the compounds were obtained and canonicalised using the OpenBabel software (v. 2.4.0; O'Boyle et al., 2011; http://openbabel.org), with salts and secondary fragments excluded.

4.2.1.2. Rat oral acute toxicity dataset

8,448 substances with 50% acute (24 hour) oral lethality data (LD50) (expressed in mmol/kg_{bw}), originally sourced from the NTP Interagency Centre for the Evaluation of Alternative Toxicological Methods (NICEATM) and United States Environmental Protection Agency (US EPA), and presented in Gadaleta et al. (2019) were utilised. This number of substances was then reduced to 8,186 substances following removal of duplicates, mixtures, polymers, inorganics and organometallics. SMILES were obtained and canonicalised through OpenBabel, with salts and secondary fragments being excluded. For the purpose of modelling, LD50 values were logarithmically transformed.

4.2.2. Molecular descriptors

4.2.2.1. Calculation of molecular descriptors

Physico-chemical and structural descriptors for the chemicals in both datasets were acquired using the PaDEL software (v. 2.21; http://www.yapcwsoft.com/dd/padeldescriptor/; Yap, 2011). In total, 1,441 descriptors were calculated that represented 1D and 2D structure. Redundant descriptors were removed that were uninformative. Descriptors that contained missing outputs were removed firstly, followed by features of low-variance (<0.01) using

VarianceThreshold from the feature selection function in the Python *sci-kit learn* library. Subsets of the original dataset were then curated through the exclusion of collinear descriptors, with descriptors that surpassed a specific pairwise correlation coefficient being removed. Pairwise correlation coefficient values used to limit collinearity and create the subsets were: 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, and 0.3. The descriptor of the pair that reported the weakest correlation to the target was omitted. Lastly, when modelling using non-decision tree algorithms, feature values were standardised. This was achieved using the StandardScaler from the preprocessing function in the Python *sci-kit learn* library, by removing the mean and scaling to unit variance.

4.2.3. Modelling algorithms

This analysis allowed for the comparison of a variety of well-used ML QSAR modelling techniques ranging from decision tree-based algorithms to neural networks. Regression models (mathematical methods for the prediction of a continuous outcome) for both datasets were built using six ML algorithms in Python (v. 3.7.6; https://www.python.org/). Random Forest, Support Vector Machine, and K-Nearest Neighbours were developed using the *sci-kit learn* library (v. 0.22.1; Pedregosa et al., 2011), Extreme Gradient Boosting with the package *xgboost* (v. 1.2.1; Chen and Guestrin, 2016), and Neural Networks and Deep Neural Networks by the open-source libraries *keras* (v. 2.4.0; Chollet, 2015) and *tensorflow* (v. 2.3.1; Abadi et al., 2015). The optimiser Adam (Kingma and Ba, 2014) and activation function Rectified Linear Units (ReLU) (Agarap, 2018) were employed within Neural Networks. Each individual method is introduced briefly below.

4.2.3.1. Random Forest

A Random Forest (RF) is an ensemble learning model that is based upon decision trees (Breiman, 2001). Decision trees work by allocation of data into nodes through conditional rules. Beginning at the root node, data are then partitioned into internal nodes that are continually split (until variance is sufficiently reduced) concluding in leaf nodes where the outcome is determined. Within RF, each decision tree is constructed independently from a random subset of available features. Once all trees are trained, predictions are achieved through the aggregation of the outputs of each individual decision tree, with these being grown through bootstrap sampling. Dissimilar and uncorrelated decision trees are produced through the random nature of the algorithm achieving superior robustness, comparatively, to single decision trees (Polishchuk et al., 2009).

4.2.3.2. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a progression of gradient boosting techniques. Unlike RF, gradient boosting combines a series of shallow trees sequentially that are tasked with correcting the errors produced by their preceding trees (Sheridan et al., 2016). XGBoost improves upon the standard gradient boosting framework through innovations in regularisation, parallel processing, and tree-pruning techniques. Such developments enable loss functions to be reduced and model complexity to be penalised – achieved in particular through the incorporation of L1 and L2 regularisation that punishes large coefficients (Chen and Guestrin, 2016).

4.2.3.3. Support Vector Machine

A Support Vector Machine (SVM) is a technique that fits a hyperplane that best separates data points from two different classes, with the hyperplane being positioned at the point that maximises the margin, which refers to the distance between the nearest data points from each class and the hyperplane itself (Cortes and Vapnik, 1995). To enable nonlinear data to be dealt with, SVMs utilise what is known as a kernel function. Such kernel functions (e.g., linear, polynomial, and radial basis function) allow for the linear separation of nonlinear data through the mapping of input data into higher-dimensional spaces (Ivancius, 2007). Whilst this method was initially developed for classification problems, the same concepts can be applied in regression tasks. In such scenarios, the objective instead is to identify a function that best fits the data, whilst reducing the error within a specified margin.

4.2.3.4. K-Nearest Neighbours

K-Nearest Neighbours (KNN) is a simple distance learning approach where the activity value of a target object is classified dependent upon its nearest neighbours in the training set. The space between neighbours is measured by an appropriate distance metric which calculates the similarity. The target is then classified to the group that the majority of neighbours belong to (Zheng and Tropsha, 2000; Gunturi et al., 2008).

4.2.3.4. Neural Network

Neural Networks (NNs) are a machine learning method that were inspired by the function and structure of the human brain. NNs are comprised of a collection of interconnected nodes, otherwise referred as neurons or units, consisting of three essential components: node character, network topology, and learning rules (Darnag et al., 2017). Specifically, node character defines how data are processed by the node, with this including information regarding the quantity of input and outputs and their respective weights associated with the node, as well as the activation function utilised. Next, the network topology refers to how nodes are organised, with these typically being structured into layers including an input layer, hidden layer(s), and an output layer. Lastly, learning rules are utilised to train the network itself, with these processes being responsible for how the weights are initialised and subsequently adjusted throughout training (Zou et al., 2008). Error correction methods are among the most commonly employed learning approaches, with these typically consisting of a back-propagation algorithm that aims to minimise the loss function through the adjustment of weights and biases (Freeman and Skapura, 1991).

4.2.3.5. Deep Neural Network

Deep Neural Networks (DNNs) are conceptually similar to NNs, although contain multiple hidden layers between the input and output. The resultant architecture enables the raw inputs, that can be thought of as level representations, to be transformed into higher-level concepts. For example, in image classification lower layers may identify the edges from a pixel array, whilst higher layers could combine such information into familiar objects, such as facial features. Therefore, aspects of the input that are important are amplified within the higher levels, and so maximising the accuracy (LeCun et al., 2015; Mansouri et al., 2019).

4.2.4. Model optimisation

Hyperparameters (parameter values that are set prior to training that govern the learning process) of all six ML algorithms were optimised for the reduced descriptor subset of the *T. pyriformis* dataset, TH_90 (See *Section 4.3.1* for subset selection rationale and Table 4.1 for further information). Definitions of all the hyperparameters used, as seen in Table 4.1, can be found in the official documentation for each algorithm (https://scikitL-learn.org/stable/supervised_learning.html;

https://xgboost.readthedocs.io/en/stable/parameter.html). Hyperparameters for each algorithm were initially optimised manually, followed by the randomised search algorithm from the model selection function in *sci-kit learn*, and finally by the Bayesian optimisation software *optuna* (v. 2.2.0; Akiba et al., 2019). Implementation of all strategies and resulting optimum hyperparameters was evaluated using cross-validation with metrics including the mean squared error (MSE) and coefficient of determination (R^2) additionally sourced from *sci-kit learn*. The range of hyperparameter values used within each algorithm are provided in Table 4.1. A manual search was conducted first, where each hyperparameter was evaluated in a stepwise manner. Hyperparameter spaces that resulted in significant performance drop-offs were used to update the ranges inputted into the autonomous approaches. A total of 50 trials was conducted during both automated strategies, with Bayesian optimisation being evaluated through error minimisation. Graphical plots of hyperparameter ranges evaluated by cross-validation measures for each strategy were produced and visually inspected to combat overfitting.

Table 4.1. Information relating to hyperparameters applicable in each algorithm. Title of parameter is listed, alongside the default quantities present within the adopted training software. Value ranges examined during processes of manual and automated optimisation (where appropriate) are listed – as are their preferred quantities, as identified through each tuning approach.

Modelling approach	**Hyperparameter**	**Default quantities**	**Quantity ranges examined**		**Optimised quantities**		
			In manual optimisation	**In automated optimisation**	**Manual**	**Automated**	
						Random search	**Optuna**
RF	max_depth	Automatic[b]	1 - 50	10 - 30	15	30	27
	n_estimators	100	50 - 500	100 - 500	490	490	499
	min_samples_split[a]	2	2 - 20	-	3	-	-
	min_samples_leaf[a]	1	1 - 100	-	1	-	-
	max_leaf_nodes[a]	Automatic[b]	2 - 202	-	Automatic	-	-
	max_samples[a]	Automatic[b]	0.1 - 0.99	-	0.99	-	-
SVM	gamma	scale[d]	0.0001 - 0.01	0.0012 - 0.003	0.00168	0.0012	0.00121
	C[c]	1	0.5 - 50	1 - 10	5	8.58	9.39
	Epsilon	0.1	0.001 - 1	0.001 - 0.02	0.418	0.018	0.00852
k-NN	n_neighbors	5	1 - 20	1 - 15	6	3	3
	p	2	1 - 5	1 - 3	1	1	1
XGB	eta	0.3	0.005 - 0.5	0.1 - 0.15	0.107	0.1	0.103
	min_child_weight	1	1 - 20	1 - 10	7	4	2
	max_depth	6	1 - 50	2 - 8	4	4	5
	gamma	0	0 - 3	0 - 0.3	0.103	0.1	0.00145
	n_estimators	100	50 - 500	100 - 250	250	250	205
	subsample	1	0.1 - 1	0.8 - 1	1	0.8	0.816
	colsample_bytree	1	0.1 - 1	0.5 - 1	0.6	0.9	0.962
	max_delta_step[a]	0	0 - 10	-	0	-	-
	lambda[a]	1	0 - 1	-	0.778	-	-

	alpha[a]	0	0 - 10	-	3	-	-
NN[e]	neurons	512	50 - 1000	50 - 1000	400	550	601
	dropout_rate	0	0 - 0.5	0 - 0.5	0.1	0.2	0.444
	epochs	100	50 - 500	50 - 500	100	250	236
	batch_size[f]	128	32 - 512	32 - 512	64	64	197
	learn_rate	0.001	0.0001 - 0.003	0.0001 - 0.001	0.001	0.0003	0.000376
DNN[g]	neurons (hidden layer 1)	512	50 - 1000	50 - 1000	750	650	944
	neurons (hidden layer 2)[h]	512	50 - 1000	50 - 1000	750	50	784
	dropout_rate (hidden layer 1)	0	0 - 0.5	0 - 0.5	0.2	0.3	0.161
	dropout_rate (hidden layer 2)[h]	0	0 - 0.5	0 - 0.5	0.2	0.4	0.494
	epochs	100	50 - 500	50 - 500	100	500	498
	batch_size[f]	128	32 - 512	32 - 512	64	32	75
	learn_rate	0.001	0.0001 - 0.003	0.0001 - 0.001	0.001	0.0003	0.000321

a. Parameters not subject to automated optimisation.

b. Value of parameter defined by algorithm should the term “None” be entered (please refer to official scikit-learn documentation, linked within Section 2.5).

c. Within automated procedure, range 1 - 10 applicable to randomised search only (1 - 20 instead examined in Optuna).

d. Value of parameter defined automatically by algorithm (please refer to official scikit-learn documentation, linked within Section 2.5).

e. Incorporates single hidden layer.

f. Within automated procedure, range 32 - 512 applicable to randomised search only (10 - 500 instead examined in Optuna).

g. Incorporates two hidden layers.

h. For each iteration of manual optimisation (only), parameter value adopted at layer 2 is identical to that corresponding in layer 1 (within automated protocols, the two are each fully independent).

4.2.5. Statistical performance and model validation

Model performance was evaluated using the metrics R^2, MSE, RMSE and MAE, which are defined in Table 4.2. These metrics were sourced from the model selection function in *sci-kit learn*. Cross-validation of results and processes was employed to limit overfitting. A number of folds (K) 2 to 25 were individually assessed for each ML algorithm (prior to optimisation, i.e., using default hyperparameter values) on the TH_90 subset. Cross-validated scorings with increasing number of folds were then visually inspected for each algorithm to identify the K value that best balanced the bias-variance trade-off. This optimal K value was then used in all future modelling procedures.

Table 4.2. Definition of error metrics used to evaluate models, where $\hat{y}$ is the predicted value of y, and $\bar{y}$ is the mean value of y.

Evaluation metric	Abbreviation	Equation
Coefficient of determination	R^2	$1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$
Mean Squared Error	MSE	$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$
Root Mean Squared Error	RMSE	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$
Mean Absolute Error	MAE	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}\rvert$

4.2.6. Model interpretation

Interpretations of how each ML algorithm related the descriptors to the modelled endpoint was determined through feature importance methods, identifying the contributions of descriptors to the outcome. Descriptors that had the strongest impact on model performance were then used to infer mechanistic rationales; thus, providing an insight into how each model arrived at their respective outcomes. These inspections were carried out by the methods of permutation feature importance from the permutation importance function in *sci-kit learn*, and the SHAP (SHapley Additive exPlanations) method implemented using the *shap* Python package (v. 0.39.0; Lundberg and Lee, 2017). With regards to SHAP, both decision tree-models were determined using the model-dependent Tree SHAP algorithm, whilst the model-independent approach Kernel SHAP was employed for the remaining ML models.

4.2.6.1. Permutation feature importance

Permutation feature importance is a model inspection technique that is model agnostic, thus enabling the calculation of descriptor importance for all ML algorithms. In this method, a single feature is randomly shuffled with the decrease in model performance observed being defined as the permutation feature importance (Breiman, 2001). In other words, the relationship between the feature and outcome are separated, therefore the reduction in performance can be indicative of the feature dependence upon the model.

4.2.6.2. Shapley additive explanations

SHapley Additive exPlanations (SHAP) is a recently developed method at the cutting-edge of model interpretability originating from the Shapley values of cooperative game theory (Lundberg and Lee, 2017). Shapley values provide a unique method to attribute a model's outputs towards feature contribution, and therefore guarantees the satisfaction of the three important properties: local accuracy, missingness, and consistency (Rodríguez-Pérez and Bajorath, 2020). SHAP values are assigned for each feature for individual predictions, with these representing their respective influence. These values can be calculated by removing a particular feature and comparing the performance difference of the model to when it was present (Wojtuch et al., 2021). A positive SHAP value indicates that the specific feature increases the model's output, with the opposite being true for a negative value. As such, the greater the absolute SHAP value the more impactful that feature is upon the model prediction (Ding et al., 2021).

Kernel SHAP is an extension of Local Interpretable Model-agnostic Explanations (LIME), with this approach aiming to train local surrogate models to explain individual predictions (Ribeiro et al., 2016). Specifically, feature contributions are approximated as Shapley values, whilst the locality of an instance to be explained is defined by LIME. A weighted linear regression model can then be trained as an explanation model, where the coefficients are the SHAP values determining feature importance (Rodríguez-Pérez and Bajorath, 2020). Comparatively, Tree SHAP is a variant of SHAP explicitly for decision tree-based models which boast a significantly reduce computation time, additionally employing a polynomial time algorithm that enables exact Shapley values to be calculated (Molnar, 2019; Lundberg et al., 2020).

4.2.7. Evaluation of uncertainty scheme towards ML methods

The scheme for the evaluation of QSARs models developed by Cronin et al. (2019), comprising of 49 uncertainty assessment criteria, was applied to the models of each ML method. Furthermore, criteria within the scheme that related specifically to ML development and understanding were identified. The ability of such criteria to effectively evaluate all aspects of uncertainty of the developed ML models was then addressed. Criteria were grouped into three categories specifically important for ML assessment – reproducibility, interpretability, and generalisation. Scorings for all criteria within each category were determined for ML models as a whole. Following the evaluation, supplementary guidance for each criterion was then proposed. It must be stressed that such suggestions do not aim to discredit the ability of the scheme for the evaluation of QSAR models in its current state to evaluate ML models, instead they provide recommendations for further analysis that ensures all aspects of model uncertainty are understood by both the developer and user.

4.3. Results and Discussion

This investigation has developed a series of ML models for two toxicity datasets, with a view to evaluating the models in terms of statistical performance and interpretability. The models have also been evaluated in terms of their associated uncertainties. From the evaluation of the models, criteria for good practice of ML modelling in predictive toxicology are reported in Section 4.4.

4.3.1. Analysis of datasets

The ML models were developed initially on two datasets. The datasets differ in terms of their size, coverage, consistency and probable quality.

The *Tetrahymena* dataset was curated by Ruusmann and Maran (2013) following a rigorous workflow ensuring correctness of data and so minimising errors present. Furthermore, the quality of the original data themselves has been reported to be highly reliable, utilising a single cell assay, following standardised procedures, and performed solely in one laboratory with experimental variability considered to be between 0.2 - 0.5 log units (Hewitt et al., 2011). The dataset has been developed on a mechanistic basis with a strong emphasis on the narcosis and reactive modes of action. It contains few, or no, specifically acting compounds such as pesticides and pharmaceuticals.

In comparison, the LD50 dataset has been compiled from a wide array of data sources, where in part due to the scale of the dataset, irrelevant, noisy, and redundant data are still present, these issues are discussed in detail for rat acute oral toxicity data by Karmaus et al. (2022). Further, Karmaus quantified a margin of uncertainty of ±0.24 log units (mg/kg) for discrete *in vivo* rat acute oral LD50. The LD50 dataset covers a broad range of chemical classes, including specifically acting substances, such as pesticides, although there is limited knowledge on the modes of action within the data set.

This inherent difference in quality of data modelled (i.e., the associated error for each datapoint) is strongly correlated to the performance of the ML algorithms as reported below, confirming the essential requirement of data cleaning and preparation before modelling (Cocu et al., 2008).

4.3.2. Descriptor selection

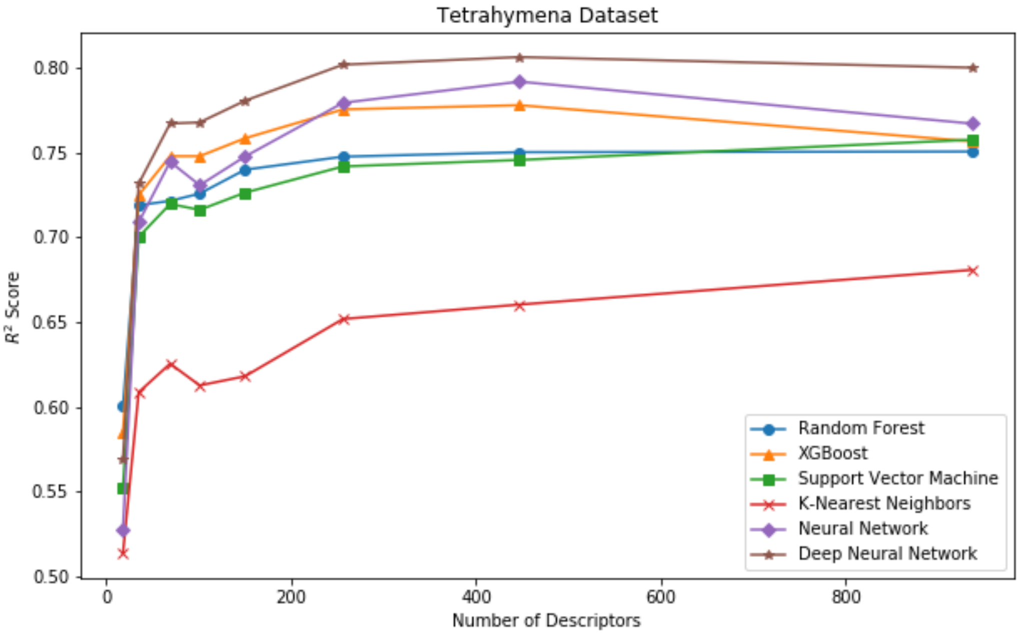
Over 1,000 descriptors were calculated very rapidly in this study, as is common with most ML models for toxicity prediction. It is known that descriptor selection can have an effect on model performance, with too many descriptors potentially introducing noise into a dataset and/or masking the influence of important descriptors (Ghafourian and Cronin, 2005). The descriptor selection process initially eliminated descriptors that contained no or little information or were otherwise redundant. This identified 936 significant descriptors for the *Tetrahymena* dataset, and 1,087 descriptors for the LD50 dataset, the datasets are available on GitHub (https://github.com/LJMU-Chemoinformatics/Best-Practice-Supplementary). The data were further reduced into seven individual subsets developed following stricter exclusion of descriptors based upon collinearity. Identifiers for each subset were labelled with either TH or LD, referring to either the *Tetrahymena* or rat acute datasets respectively, followed by the suffix of either ‘Full’ or a numerical value. ‘Full’ indicates that the subset has undergone no collinear descriptor removal, whilst a numerical value references the percentage threshold at which collinear descriptors were removed for that particular set. Table 4.3 provides the number of descriptors contained within each subset.

Table 4.3. Algorithm performance presented by cross-validated R^2 Test (k = 10) using default ML algorithm hyperparameters for each reduced descriptor subset for both the *Tetrahymena* and LD50 datasets.

Subset	**Number of Descriptors**	**R^2 Test**					
		RF	**SVM**	**KNN**	**XGBoost**	**NN**	**DNN**
Tetrahymena							
TH_Full	936	0.751	0.758	0.681	0.757	0.767	0.800
TH_90	447	0.750	0.746	0.660	0.778	0.792	0.806
TH_80	256	0.748	0.742	0.652	0.776	0.779	0.802
TH_70	150	0.740	0.726	0.618	0.758	0.748	0.781
TH_60	101	0.726	0.716	0.613	0.748	0.731	0.768
TH_50	69	0.722	0.720	0.625	0.748	0.745	0.767
TH_40	35	0.719	0.700	0.609	0.725	0.709	0.732
TH_30	18	0.600	0.552	0.513	0.585	0.528	0.569
LD50							
LD_Full	1087	0.567	0.559	0.511	0.549	0.517	0.583
LD_90	546	0.563	0.562	0.508	0.546	0.507	0.583
LD_80	353	0.567	0.565	0.517	0.538	0.502	0.578
LD_70	231	0.563	0.556	0.508	0.537	0.492	0.577
LD_60	141	0.564	0.547	0.506	0.530	0.475	0.565
LD_50	98	0.542	0.519	0.482	0.520	0.444	0.544
LD_40	59	0.504	0.450	0.435	0.467	0.370	0.470
LD_30	30	0.381	0.316	0.304	0.349	0.281	0.290

The performance of each modelling algorithm trained on the full datasets, as well as the seven subsets, for both *Tetrahymena* and LD50 are also reported in Table 4.3. The general results from the analysis demonstrate that in all cases models for the *Tetrahymena* dataset outperformed those for the LD50 dataset. Furthermore, specifically focusing upon the individual algorithms, it was found that the highest performing models for both datasets were produced using DNN, while the poorest performing with KNN. As demonstrated in Figure 4.2 for both datasets, the performance of all models increases dependent upon the number of descriptors available to be modelled. However, this growth plateaus once the number of descriptors passes 200, and there is no gain in including further descriptors. Although the trends observed between the two datasets remain the same, prominent differences in performance separate the results. Such notable differences can be accredited to the contrast in quality of both datasets from their respective sources.

In the case of the *Tetrahymena* dataset, optimal performance of RF, SVM, and KNN was reported at TH_Full, while XGBoost, NN, and DNN peaked at TH_90 (collinearity threshold=>0.9). Similar results were achieved for the LD50 dataset where optimal performance for all algorithms can be observed where more descriptors are used, while the plateauing of model performance is also seen.



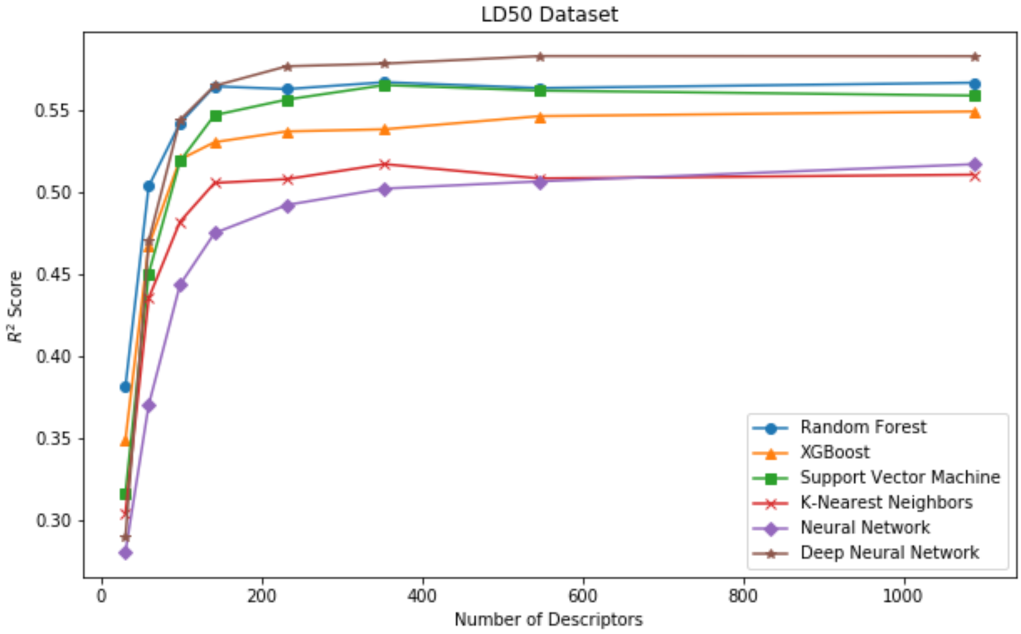


Figure 4.2. Performance of the ML methods for the *Tetrahymena* and LD50 datasets of each reduced descriptor subset.

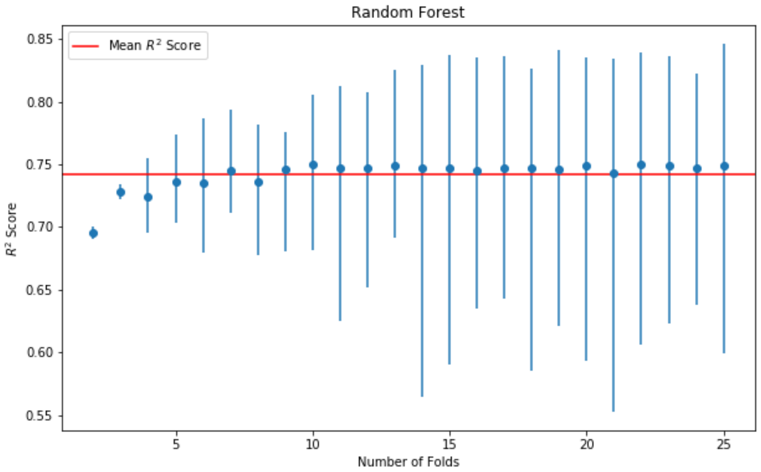
With regard to the selection of descriptors, calculation of collinearity between sets of descriptors with the pairwise correlation coefficient is a standard approach, yet the decision of which descriptor to remove from the pair may cause difficulty. Removal of the descriptors with least correlation towards the output is the most logical approach. However, the potential to remove descriptors that individually are not as statistically relevant to the outcome, but have a greater impact when modelling utilising the entire dataset, may still occur (Dormann et al., 2013).

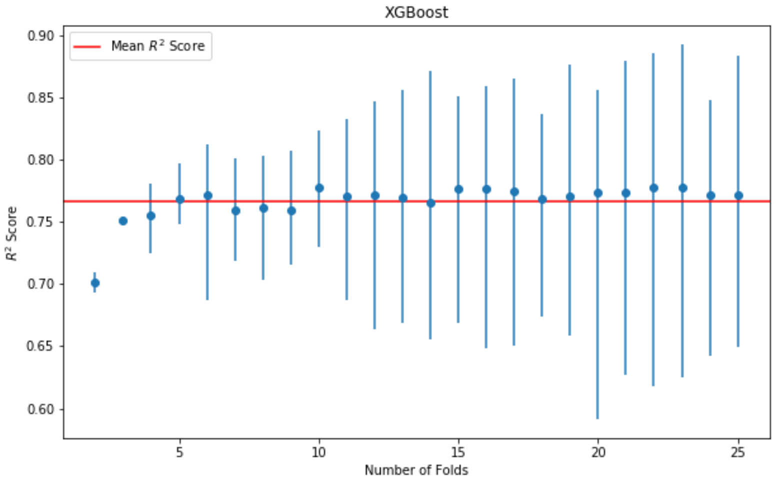
As can be seen in Figure 4.2, irrespective of the dataset used, DNN gave the greatest predictive performance and KNN the poorest, and most of the other algorithms produced similar results. The NN models showed differential performance, with NN models of the *Tetrahymena* dataset demonstrating strong performance for all subsets (see Table 4.3) irrespective of whether shallow or deep networks were created in comparison to other algorithms. However, this trend was not seen for the LD50 dataset, where the performance of the shallow NN was as poor as the worst performing algorithm, KNN, while DNN still remained as the optimal ML algorithm. Due to the additional hidden layer and nodes present in the DNN, it is possible that complex and more variable data, such as the LD50 dataset, can undergo further combinations and transformations as a result of the depth provided; thus, translating to greater performance in comparison to shallow networks (Winkler and Le, 2017).

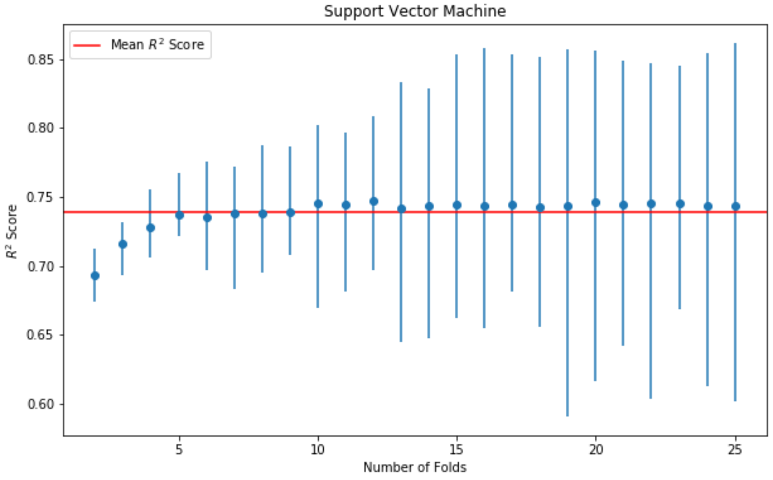
The ability of ML algorithms to handle large amounts of data with little feature selection is evident. For instance, the results demonstrated within the subsets where no removal of collinear descriptors has occurred (i.e., TH_Full and LD_Full), similar performance to those models with feature selection. However, inclusion of redundant descriptors in the sets that provide no contribution to model performance is impractical and may even serve to hinder interpretability. A certain degree of feature selection is therefore likely to be beneficial, often leading to improvements on prediction accuracy, although rigorous selection procedures inevitably will introduce errors – where irrelevant descriptors are selected while omitting descriptors that are relevant (Khan and Roy, 2018; Hawkins, 2004). In order to draw more detailed conclusions about the modelling approaches the TH_90 dataset was selected for model optimisation. This dataset has excluded features with greater than 0.9 collinearity and demonstrated strong performance with all ML methods.

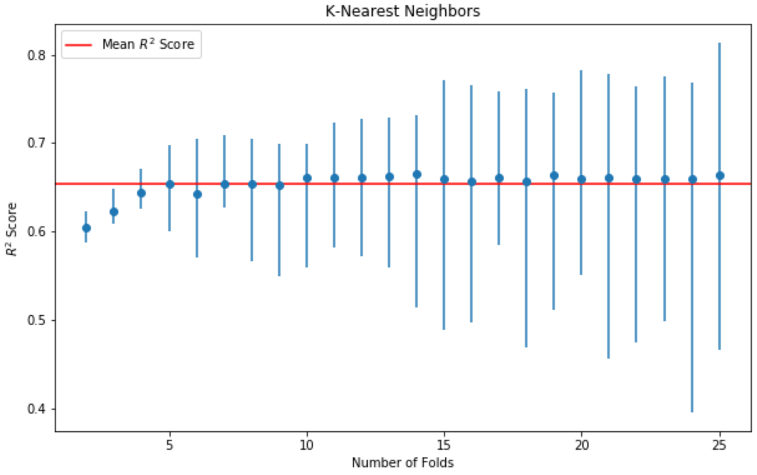
4.3.3. Evaluation of cross-validation approaches

Cross-validation is an essential tool in the development of all QSAR models for toxicity prediction (Gramatica, 2007), and for ML modelling in particular. As part of the cross-validation of ML models, an analysis of the number of folds (i.e., how many smaller sets the original dataset has been split into) was also undertaken. Folds ranging from 2 to 25 were investigated with each ML algorithm. Figure 4.3 shows the R^2 against the number of folds. For all ML methods, cross-validation demonstrated that the performance of all models was poorer with a low number of folds i.e., up to five. When more than five folds were utilised, the performance of all algorithms improved and approached the average observed. Since the initial folds are considered in the mean R^2 score, the score is skewed slightly low, hence the latter results generally performed slightly better than the mean. The largest difference that can be noted from increasing the number of folds is the variation, as denoted by the blue bars in Figure 4.3, between each rising significantly.


Random Forest
Mean R^2 Score
R^2 Score
0.85
0.80
0.75
0.70
0.65
0.60
0.55
5
10
15
20
25
Number of Folds


XGBoost
Mean R^2 Score
R^2 Score
0.90
0.85
0.80
0.75
0.70
0.65
0.60
5
10
15
20
25
Number of Folds


Support Vector Machine
Mean R^2 Score
R^2 Score
0.85
0.80
0.75
0.70
0.65
0.60
5
10
15
20
25
Number of Folds


K-Nearest Neighbors
Mean R^2 Score
R^2 Score
0.8
0.7
0.6
0.5
0.4
5
10
15
20
25
Number of Folds

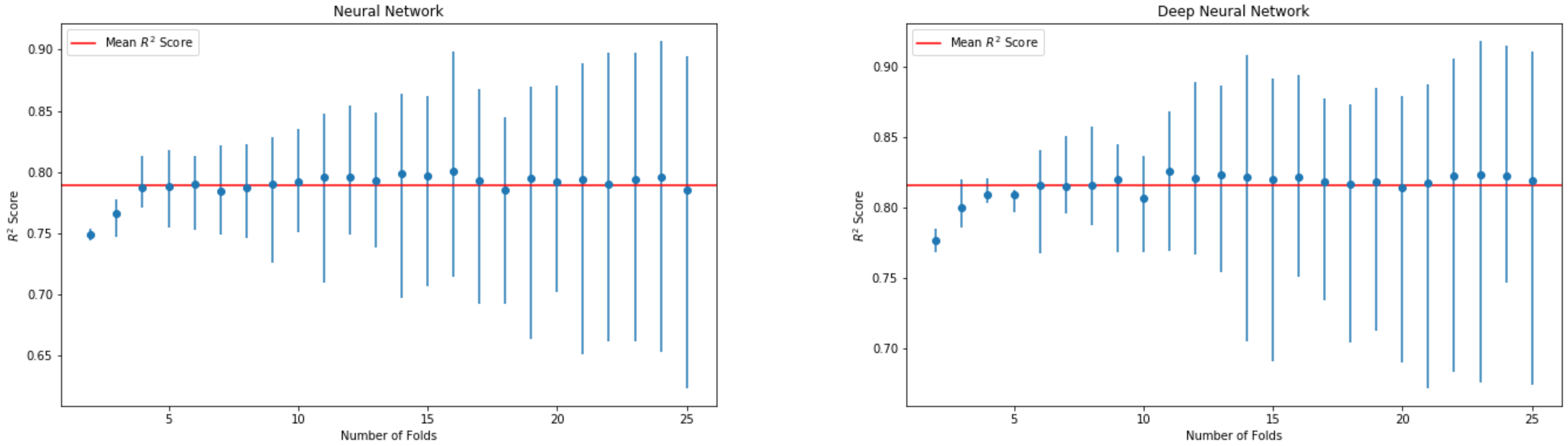


Figure 4.3. Sensitivity analysis of cross-validation for each algorithm dependent upon number of folds. Error bars are indicated in blue.

Splitting a dataset into folds needs to satisfy two essential criteria, these being that the evaluation set is large enough so that randomness in the prediction assessment is accounted for and that the diversity of the full set is reflected in the reduced sample size. Achieving this requires careful balancing of the conflicting directions (Zhang and Yang, 2015). The results from this study indicate that 10-fold (i.e., k = 10) validation is optimal for the assessment of the performance of all ML models, as shown by the plateauing of performance and relatively low variance in Figure 4.3. Ten-fold cross-validation is known to provide a strong middle ground, with not only demonstrating low variance across all algorithms, but also being commonly employed in literature due to its traditionally statistically unbiased results (Vakharia and Gujar, 2019). Thus, confirming that the results conform to the trends of existing knowledge, ten folds were selected to be used throughout the study as the means of cross-validation.

4.3.4. Parameter optimisation

Optimisation of model hyperparameters was undertaken using three different methods, with default values providing a baseline for performance. Firstly, the individual parameters were explored manually, followed by randomised search, and concluded with a Bayesian approach. The complexity of each method increases in comparison to the previous, although the time required, and expert judgment reduces. The performance of each algorithm with hyperparameters identified from the various approaches are reported in Table 4.4. Worthy of note, even though precautionary efforts to limit the effects of overfitting were employed (i.e., cross-validation) results reported within Table 4.4 demonstrate that the majority of models almost perfectly replicated the training set suggesting overfitting has occurred. Whilst further efforts to reduce overfitting could be employed, such as decreasing model complexity, the main focus of this investigation is upon the parameter optimisation procedures.

Table 4.4. Cross-validated statistical performance (k = 10) of each algorithm using optimal parameters identified from each approach upon the TH_90 subset.

Approach	**Statistical Performance**	**Model**					
		RF	**SVM**	**KNN**	**XGBoost**	**NN**	**DNN**
Default	R^2 Train	0.964	0.902	0.782	1.000	0.953	0.968
	R^2 Test	0.750	0.746	0.660	0.778	0.792	0.806
	MSE	0.271	0.276	0.368	0.241	0.225	0.209
	RMSE	0.521	0.526	0.606	0.490	0.475	0.458

	MAE	0.378	0.363	0.441	0.354	0.339	0.317
Manual	R^2 Train	0.962	0.974	0.787	0.974	0.963	0.967
	R^2 Test	0.753	0.794	0.694	0.800	0.785	0.816
	MSE	0.268	0.223	0.332	0.216	0.234	0.199
	RMSE	0.518	0.472	0.576	0.465	0.483	0.446
	MAE	0.376	0.326	0.416	0.335	0.329	0.312
Random Search	R^2 Train	0.966	0.973	0.845	0.986	0.981	0.987
	R^2 Test	0.753	0.804	0.696	0.808	0.800	0.822
	MSE	0.268	0.213	0.328	0.208	0.217	0.193
	RMSE	0.517	0.461	0.573	0.456	0.466	0.440
	MAE	0.375	0.319	0.410	0.326	0.320	0.306
Optuna	R^2 Train	0.966	0.977	0.845	0.995	0.968	0.992
	R^2 Test	0.753	0.804	0.696	0.811	0.809	0.829
	MSE	0.268	0.213	0.328	0.205	0.208	0.186
	RMSE	0.517	0.461	0.573	0.453	0.456	0.431
	MAE	0.375	0.319	0.410	0.324	0.314	0.301

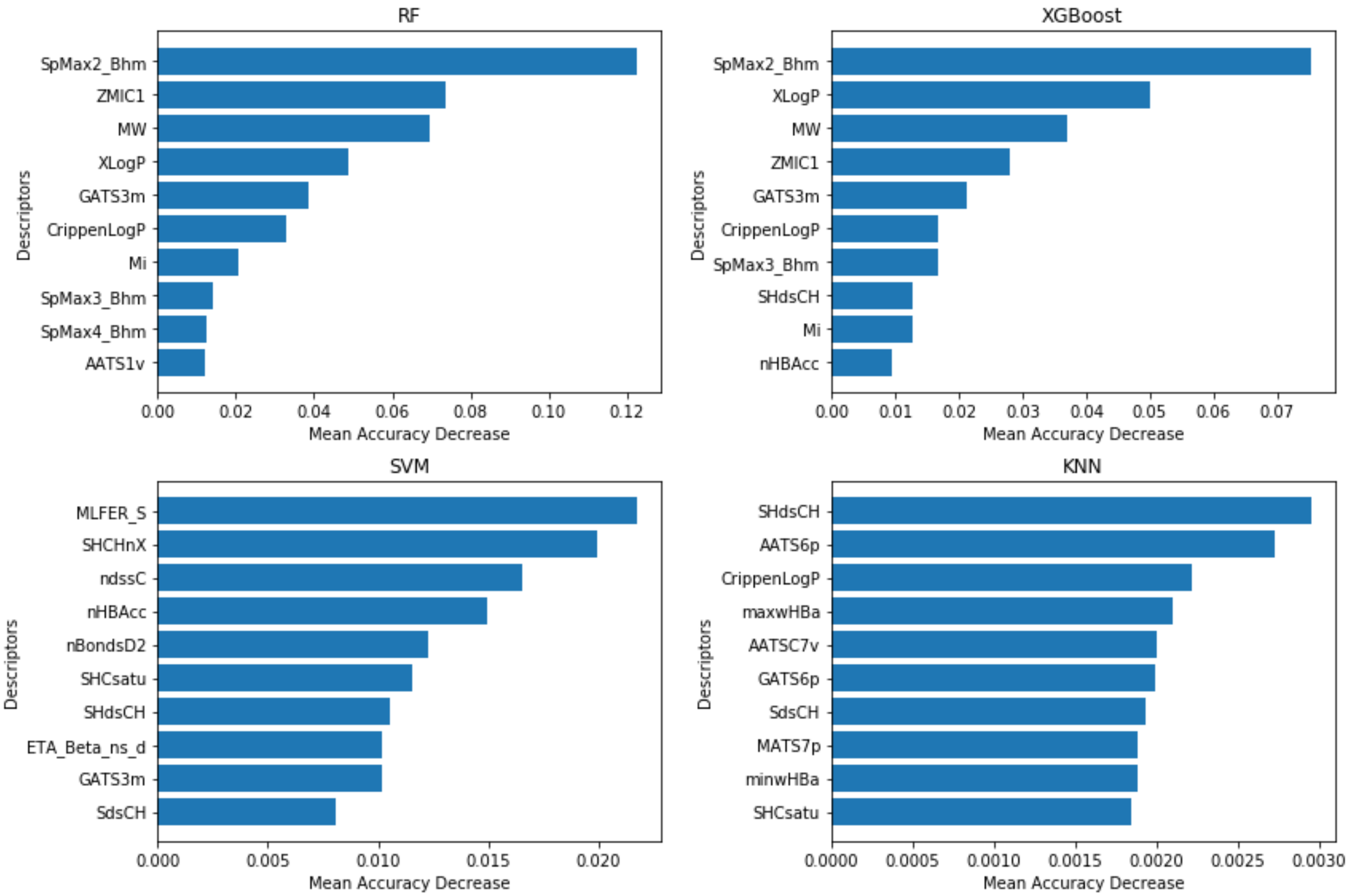
The results of the hyperparameter optimisation demonstrate that approaches which utilise more computational dependencies reported notably stronger performances. Although, it can be observed that algorithms which were tuned on a lower number of hyperparameters, such as RF and KNN, benefitted less than others due to the lower quantity of parameter combinations. On the other hand, algorithms that require a larger quantity of hyperparameter tuning can be seen as benefiting from such mathematically-informed approaches such as Bayesian, where the number of combinations increases exponentially. Overall, the general result for all algorithms demonstrates that Bayesian optimisation approaches reported the strongest performance, whilst also enabling reduced computational times and greater interpretability; as such, input values obtained through this procedure were selected to represent the optimal values for all ML algorithms. Further information regarding the optimisation procedure can be found in *Appendix IV*.

4.3.5. Feature importance

Providing some understanding of the mechanistic basis of a QSAR model for predictive toxicology is crucial to not only provide confidence, but additionally demonstrate quality through interpretability. Descriptors utilised within the model should therefore reflect the mechanisms by which toxicity is brought about. Although this may be a relatively simple process where there are few descriptors, such as in linear regression, current ML algorithms are able to incorporate a large number of features. Therefore, without integrating features known to support mechanistic justification, reporting potential mechanistic drivers may be unfeasible. Identification of potential mechanistic relevance therefore requires understanding of which features are providing the greatest value to each algorithm. Hence, gathering this information requires the calculation of the importance of features.

4.3.5.1. Permutation feature importance

Permutation feature importance randomly shuffles the values of a single descriptor whilst monitoring the difference in model performance. Thus, the importance of the feature can be determined dependent upon the change in predictive accuracy (Breiman, 2001). Due to the feature importance rankings provided being sensitive to model parametrisation, identification of descriptors of greatest importance was conducted post-development. Figure 4.4 shows the ten highest scoring descriptors for each algorithm on the TH_90 dataset as identified through implementation of the permutation feature importance function in *sci-kit learn*. The findings demonstrate that both ensemble methods equally reported the Burden Modified Eigenvalue, SpMax2_Bhm (largest absolute eigenvalue of burden modified matrix – n 2/weighted by relative mass), to be the most influential descriptors during the modelling process. Additionally, the five highest ranking descriptors ZMC1, MW, XLogP, and GATS3m, are additionally present in both RF and XGBoost, with the ordering remaining nearly identical. Investigations into the remaining four models show that electrotopological state indices are routinely present within the top rankings. Accordingly, electronic and topological information regarding each chemical is therefore crucial to model performance. Although the plots reported in Figure 4.4 are useful to provide an insight into the behaviour of models, little is known about how each feature affects toxicity. Furthermore, a recent report by Hooker and Mentch (2019) advocated against traditional permutation importance methods, finding that they can give rise to misleading results particularly while dealing with correlated features. Therefore, to unearth stronger results, associated with greater confidence for interpretability, an additional approach defined as Shapley Additive exPlanations (SHAP) was undertaken.


RF
Descriptors
SpMax2_Bhm
ZMIC1
MW
XLogP
GATS3m
CrippenLogP
Mi
SpMax3_Bhm
SpMax4_Bhm
AATS1v
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Mean Accuracy Decrease
XGBoost
Descriptors
SpMax2_Bhm
XLogP
MW
ZMIC1
GATS3m
CrippenLogP
SpMax3_Bhm
SHdsCH
Mi
nHBAcc
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Mean Accuracy Decrease
SVM
Descriptors
MLFER_S
SHCHnX
ndssC
nHBAcc
nBondsD2
SHCsatu
SHdsCH
ETA_Beta_ns_d
GATS3m
SdsCH
0.000
0.005
0.010
0.015
0.020
Mean Accuracy Decrease
KNN
Descriptors
SHdsCH
AATS6p
CrippenLogP
maxwHBa
AATSC7v
GATS6p
SdsCH
MATS7p
minwHBa
SHCsatu
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
Mean Accuracy Decrease

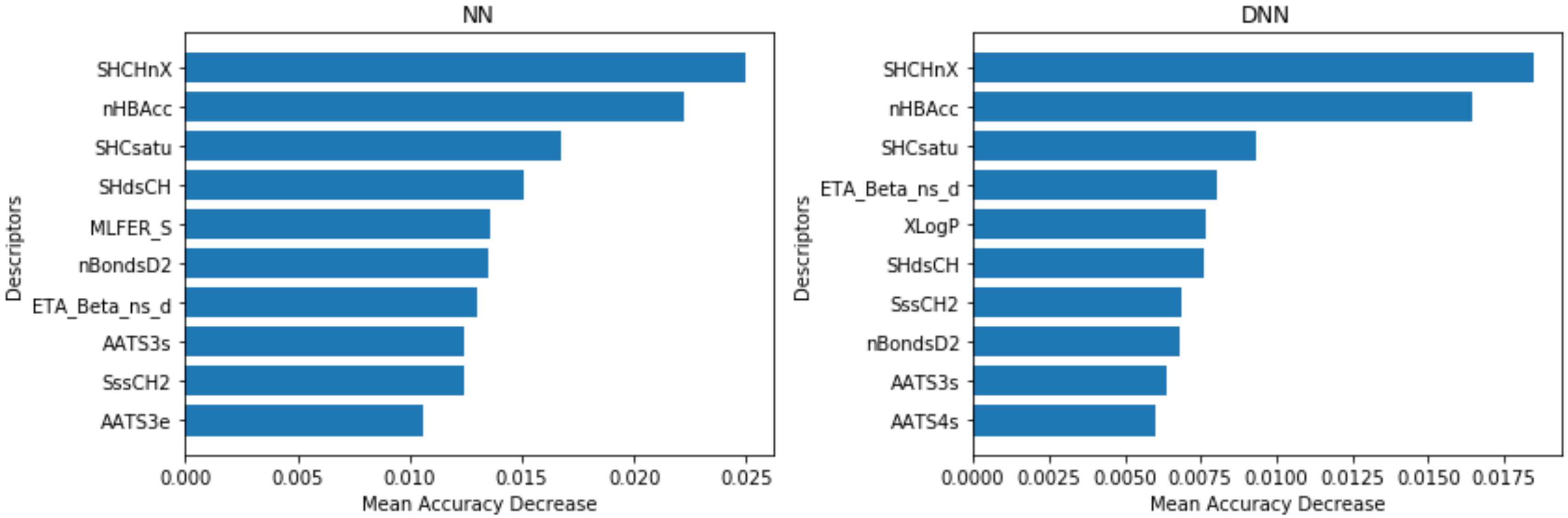
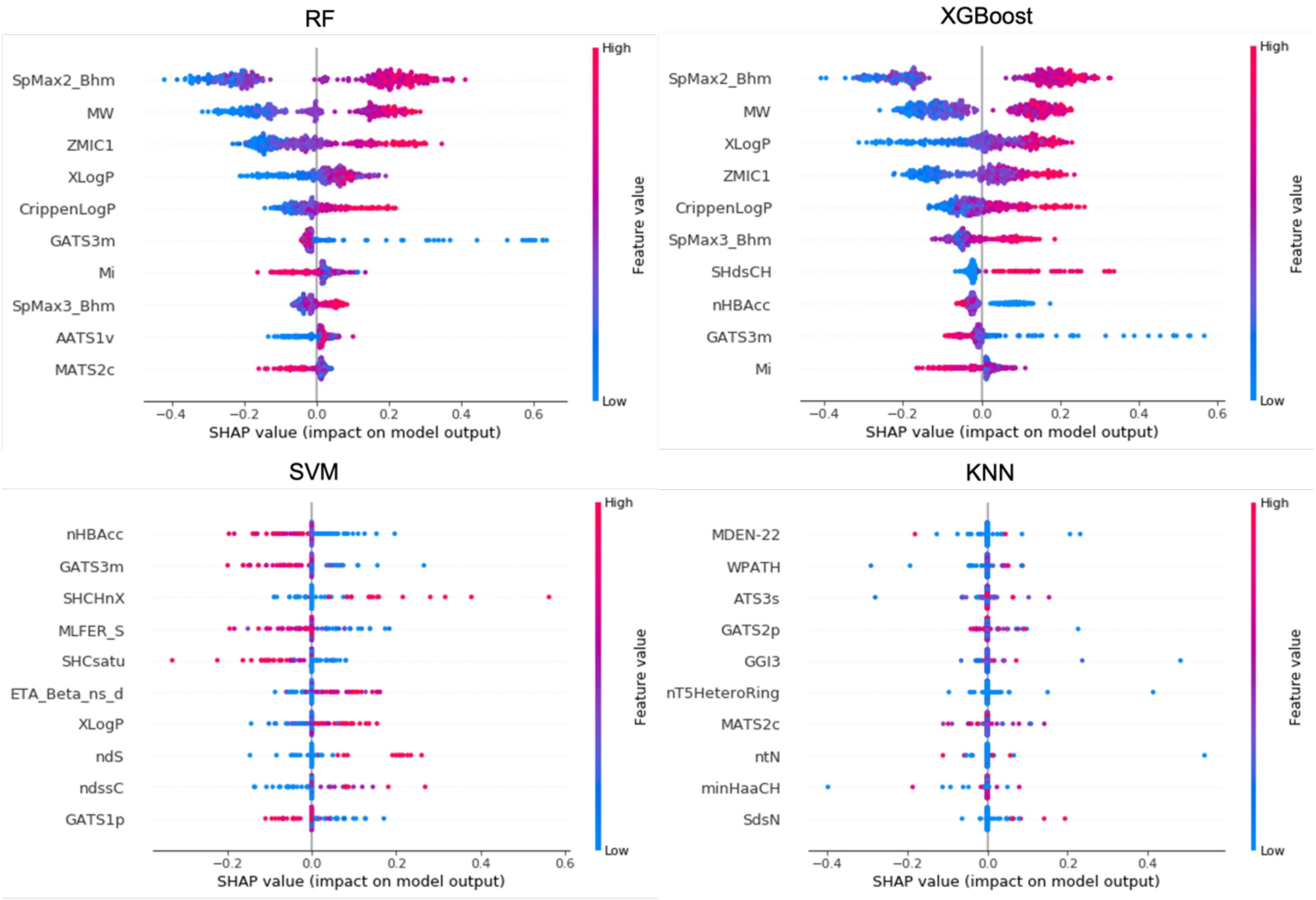


Figure 4.4. Plots of the ten greatest mean decreases in accuracy (measured by mean squared error) for the descriptors in each optimised ML model on the TH_90 subset as identified by their permutation feature importance score.

4.3.5.2. Shapley Additive exPlanations

Given the inherent issues with permutation feature importance scores when dealing with correlated features, an alternative approach was undertaken to provide results with greater confidence for the interpretation of descriptor importance. Shapley Additive exPlanations (SHAP) is a unified theory, where several algorithms (defined as Local Interpretable Model-agnostic Explanations (LIME), DeepLIFT, layer-wise relevance propagation, classic Shapley value estimation, Shapley sampling values, and Quantitative input influence) that have previously been used to interpret ML models have been combined (Lundberg and Lee, 2017). Individual predictions can be examined through SHAP, where impacts from each feature on the predicted value are processed as an additive combination (Carlsson et al., 2020). Calculated SHAP values are established from Shapley values that originate from coalitional game theory. This method enables the pay-out (i.e., the prediction) to be fairly distributed among the players (descriptors); thus, allowing the contribution each presents to be quantified (Molnar, 2019).


RF
SpMax2_Bhm
MW
ZMIC1
XLogP
CrippenLogP
GATS3m
Mi
SpMax3_Bhm
AATS1v
MATS2c
−0.4
−0.2
0.0
0.2
0.4
0.6
SHAP value (impact on model output)
High
Low
Feature value
XGBoost
SpMax2_Bhm
MW
XLogP
ZMIC1
CrippenLogP
SpMax3_Bhm
SHdsCH
nHBAcc
GATS3m
Mi
−0.4
−0.2
0.0
0.2
0.4
0.6
SHAP value (impact on model output)
High
Low
Feature value
SVM
nHBAcc
GATS3m
SHCHnX
MLFER_S
SHCsatu
ETA_Beta_ns_d
XLogP
ndS
ndssC
GATS1p
−0.2
0.0
0.2
0.4
0.6
SHAP value (impact on model output)
High
Low
Feature value
KNN
MDEN-22
WPATH
ATS3s
GATS2p
GGI3
nT5HeteroRing
MATS2c
ntN
minHaaCH
SdsN
−0.4
−0.2
0.0
0.2
0.4
SHAP value (impact on model output)
High
Low
Feature value

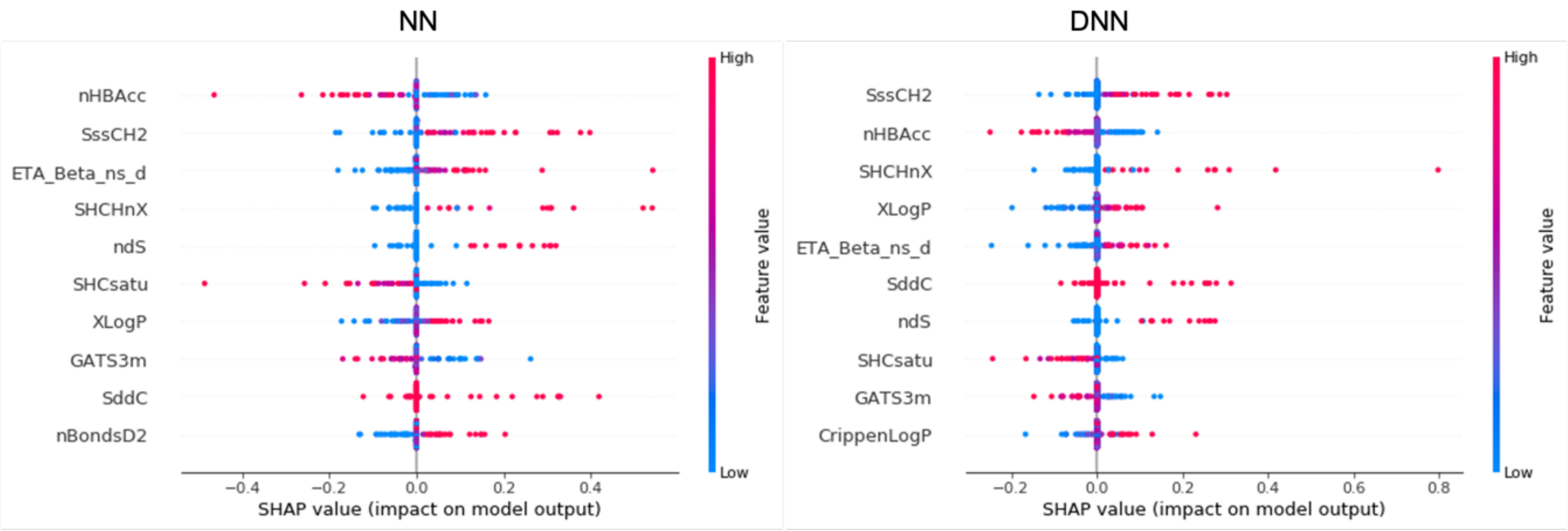
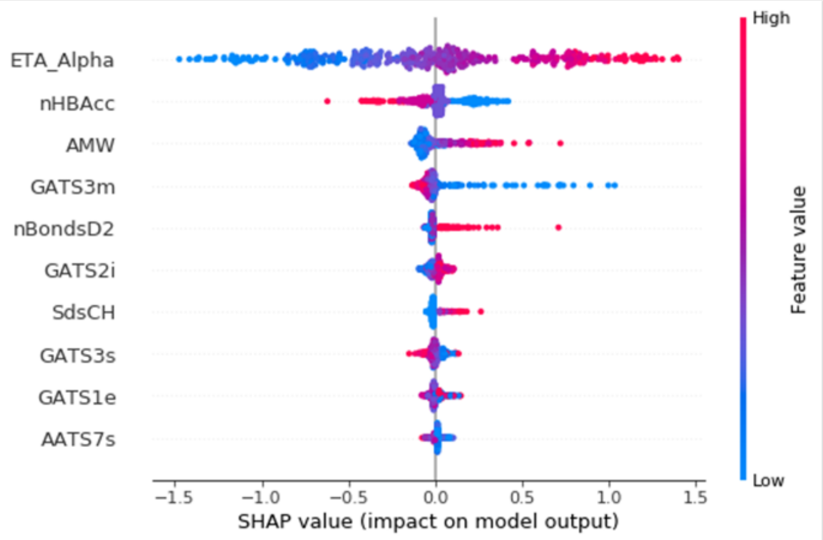


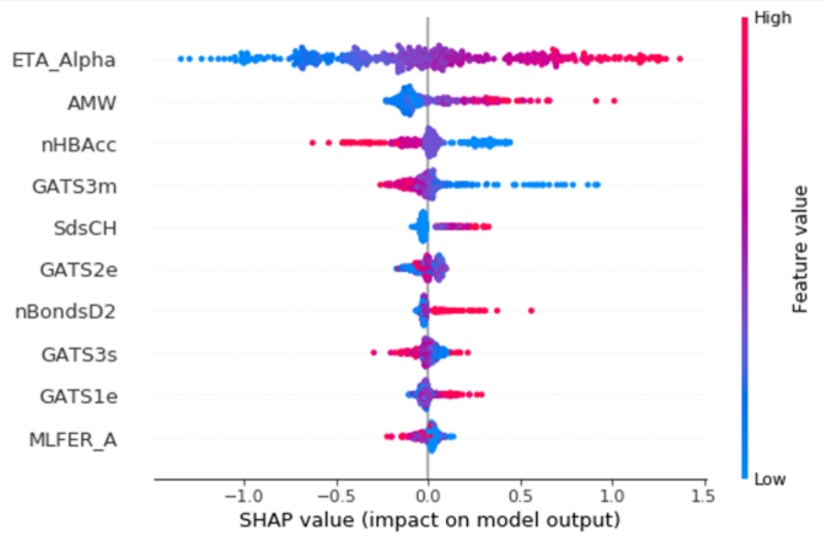
Figure 4.5. Beeswarm plots of the ten highest ranked descriptors and their impact distributions for each ML algorithm based upon their importance as determined via SHAP values.

Ranking of the highest performing descriptors for each model as determined by SHAP is illustrated in Figure 4.5. Each individual point within the plots corresponds to a single prediction and the impact that feature had upon the model's prediction based upon the SHAP value; thus, the relationship between the feature and output can also be determined. Rankings of each descriptor can be determined by the order, where the highest descriptor refers to the most impactful on the model. Both ensemble models, and to a certain extent the DNN model, reported similar results to that obtained through permutation feature importance, whilst all other algorithms discovered alternative descriptors to be of the highest importance. As seen in Figure 4.5, features of greatest importance for both ensemble methods are clearly defined with each descriptor impacting the outcome. On the other hand, although being identified as of the greatest importance for non-ensemble algorithms (i.e., SVM, KNN, NN, and DNN), features were only typically utilised in a handful of predictions with the majority having no impact. Due to descriptors not always being employed in predictions for non-ensemble methods, it was hypothesised that modelling on a reduced set of descriptors would inherently provide greater clarity.

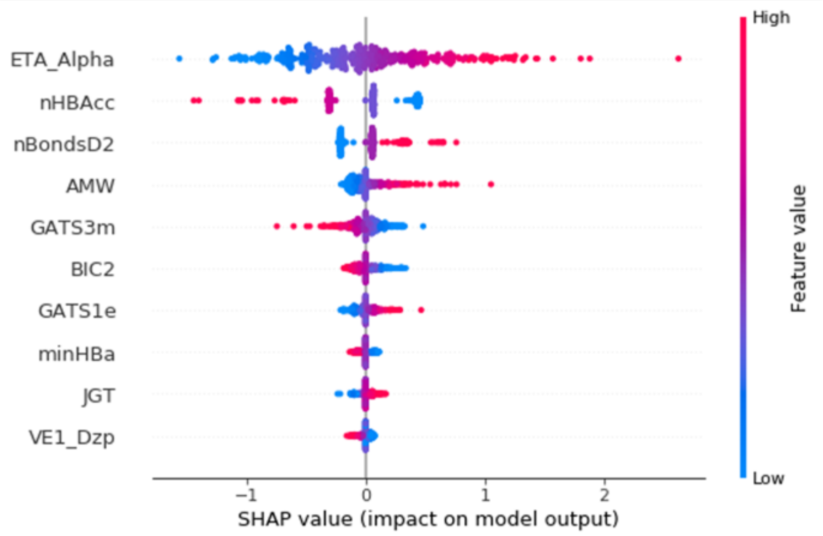
RF



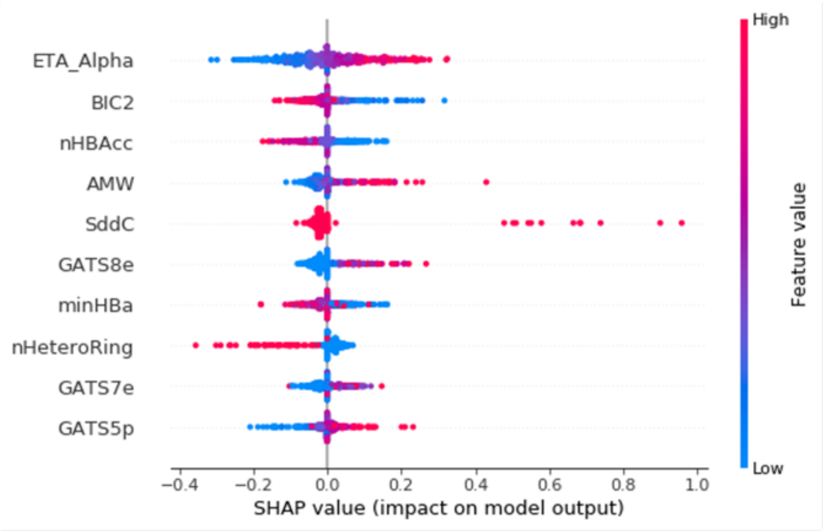
XGBoost



SVM



KNN



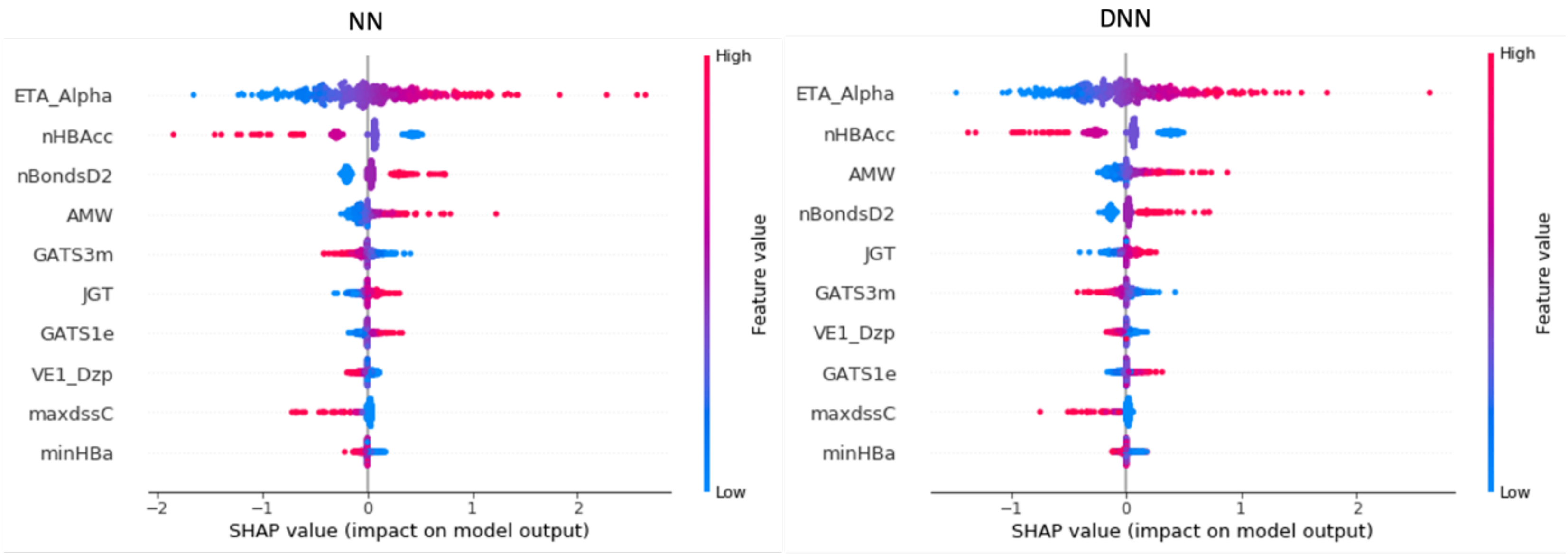


Figure 4.6. Beeswarm plots of the ten highest ranked descriptors and their impact on distributions for each ML algorithm based upon their importance as determined via SHAP values using the dataset TH_50.

Figure 4.6 illustrates the ranking of the top ten features as determined by SHAP, although in this scenario the subset TH_50 that contains only 69 descriptors in comparison to the original 447 was modelled. By reducing the number of descriptors, aiding in the limitation of overfitting, clear distributions of each descriptor can be observed – with features demonstrating a greater engagement in all predictions. As such, as a trade-off for prediction accuracy (see *Section 4.3.1*), reducing the quantity of descriptors used to generate models has been shown to improve the interpretability of the model. Contrary to Figure 4.5, results from Figure 4.6 illustrate that each ML model's most impactful features agree with one another. Particularly, the extended topochemical atom descriptor *ETA_Alpha* (sum of alpha values of all non-hydrogen vertices of a molecule) in all cases was found to have the greatest impact, where increasing this feature was continually found to be related to increasing toxicity. Similarly, *nHBAcc* (number of hydrogen bond acceptors) which measures hydrogen bonding capacity, was continually identified as a top contributing descriptor. Mechanistically, *ETA_Alpha* is associated with hydrophobicity, specifically characterising the average molecular polarisability, which historically has been demonstrated to be fundamental in the prediction of toxicities (Zhu et al., 2020; Zhao et al., 2010; Cronin and Dearden, 1995). Likewise, *nHBAcc* can also be shown to reflect the polarisability of compounds, although most notably it describes the hydrogen bonding ability, with greater number of acceptors present resulting in an increase in toxicity (Wang et al., 2019b).

The dichotomy between the reporting of feature importance and mechanistic interpretation is very striking in this example. The techniques applied allow for significant descriptors to be identified, but mechanistic relevance requires knowledge of the mechanisms and why such descriptors are related to mechanisms of action. The descriptors noted above are likely to be associated with hydrophobicity, which is widely acknowledged as being a key determinant in toxic potency to *Tetrahymena* (Cronin, 2006; Enoch et al., 2008). However, none of these descriptors are likely to capture excess toxicity brought about by electrophilic reactivity (Schultz et al., 2002).

4.3.6. Assessment of the uncertainty of ML models

Identifying and characterising the uncertainty associated with QSARs for toxicity prediction will assist demonstrating their acceptability for a particular purpose (Belfield et al., 2021;

Sahlin, 2013). The uncertainty criteria developed by Cronin et al. (2019) were applied to ML models. The uncertainty criteria were grouped into ten components that summarise the main characteristics of a QSAR relating to its creation, characterisation, and application. Using knowledge gained throughout the study and development of six ML models, the current state of the ten components and their relevance to ML models can therefore be addressed. ML modelling presents a range of challenges that may potentially impact each of the phases of QSAR development, which may not currently be fully considered by these criteria. Notably, throughout the development of models, three distinctive areas which required careful attention were encountered, these being: reproducibility, interpretability, and generalisation. Each of these aspects is likely to affect multiple components within the current criteria and need to be addressed to ensure validity.

4.3.6.1. Reproducibility

At the heart of the validity and reliability of any experimental process is the assurance that the entire experimental procedure can be repeated, with both results and conclusions replicable (Pineau et al., 2020). However, detailed reporting of methods and results is often ignored within ML and AI, with such issues only recently gaining attention in broader uses (Gundersen, 2020). ML presents its own set of unique challenges that need to be fulfilled to achieve reproducibility. By their nature, ML models contain a large number of parameters that are learnt or manually decided upon by the modeller, and that even if left at default for each algorithm may vary between users dependent upon versions of software libraries being employed (Beam et al., 2021). In addition, intrinsic to many ML models is the use of randomness during training, especially for neural networks where weights are assigned stochastically (Scardapane and Wang, 2017), which without being controlled through the use of a pseudorandom number generator will result in models that are impossible to replicate. Thus, to develop a ML QSAR model more information is required to be reported to ensure reproducibility which, without incorporation into consideration by the uncertainty scheme, may lead to an incorrect evaluation of uncertainty.

To ensure reproducibility, developed models need to have been sufficiently documented and reported, requiring the definition of all the components that made up the QSAR to be stated. For non-ML QSARs, provision of descriptors, statistical values, and algorithms utilised is sufficient, although specifically with respect to the models currently developed hyperparameter values and ranges would be also required. Withholding such information will undoubtedly result in models that are irreproducible by another user (Sugimura and Hartl, 2018), and therefore this needs to be assessed within this criterion.

Confirming that all ML models can be reproduced effectively, provision of the original source code, software (including version), and computational hardware is required. Assessment of uncertainty related to reproducing predictions was not scored in this assessment – due to no attempt to reproduce the models being made within the current work, yet the inclusion of relevant information present in the manuscript would certainly enable this to be done, for instance the recalculation of the predictions for the training and test sets. The models and predictions could be replicated by implementing the random seed to initialise the random number generator. Due to the randomness associated with many ML algorithms, neglecting to include a random seed would undoubtedly make it impossible to replicate results (Sugimura and Hartl, 2018).

4.3.6.2. Interpretability

Sound interpretation of results obtained from QSAR models is by no means a novel concept, with mechanistic interpretations making up one of the five original OECD Principles for the Validation of QSARs for Regulatory Use (OECD, 2007). The term mechanistic interpretation refers to directly defining the causality between the chemicals and endpoint (Thoreau, 2016). Explainable performance is essential for model trustworthiness, where the behaviour of ML algorithms may be accepted and understood by humans (Wu et al., 2021). Certain ML methods, such as decision trees and KNN, which researchers have used exhaustively, may already be classified as interpretable, where the logical algorithm structure enables feature importance to be deduced (Molnar, 2019). However, many ML methods are inherently labelled “black box”, where the inner workings are hidden to the user resulting in an opaqueness in the understanding on how the system makes predictions (Carvalho et al., 2019). Interpretation of ML techniques may be differentiated into two types of categories being either of global or local interpretability. Global interpretability ensures that a user can understand how a model works through inspection of the layout and parameters, thereby illuminating the inner workings and so increasing transparency. Whereas local interpretability considers the impact each feature has on a specific prediction, thus in a toxicological assessment chemical features can be causally related towards the outcome (Du et al., 2019).

Hence, for a QSAR model that has been developed using a ML technique, both concepts of interpretability need to be addressed.

In the current study, a range of algorithms was employed which inherently have varying levels of interpretability – for example RF is far easier to interpret than NNs. However, to enable all ML models to be interpreted, feature importance methods have been employed. Through this, descriptors have been scored where the greatest contributors towards the endpoints are defined, with these being causally related towards the outputs. However, worthy of note was the difficultly of interpreting non-ensemble algorithms, where the inclusion of a large descriptor pool overshadowed the relative importance of descriptors leading to greater complexity in interpreting the results. For these specific algorithms, only when the descriptor pool had been sufficiently reduced, could clear interpretations be gathered. As such, this criterion was only scored as to contain moderate uncertainty, as each ML model can be reasonably interpreted externally by another human. However, a clear drawback in interpreting models following this methodology is that only the highest scoring features are being related to the endpoints, whereas descriptors of lower importance are ignored.

Unlike a traditional QSAR model, vast numbers of descriptors may be used within a ML algorithm, therefore relating all of these to the potential mechanisms of action would be impractical. As such, through the usage of feature importance methods, descriptors that demonstrated the greatest importance to the model were mechanistically related towards the endpoint. Thus, in this sense, these descriptors can be thought of as mechanistic drivers. Although, as previously mentioned, this does not account for all the descriptors used throughout the study, and instead only the most impactful features. As a result of only a fraction of the features employed in the model reporting some mechanistic rationale, only moderate uncertainty can be accredited to this criterion. In general, it must be noted that these methods of identifying descriptor importance specifically enable the model interpretability to be defined, which in turn can be used to relate to a predefined mechanistic rationale and support mechanistic understandings. Yet a strong advantage of such methods is in dealing with scenarios where no mechanistic knowledge is available, for instance global datasets, and so enabling mechanistic rationales to be postulated through the information on descriptors' importance.

4.3.6.3. Generalisation

The final aspect of ML that requires greater consideration from the uncertainty criteria is the ability of the model to generalise well to unseen data, i.e., how well the model can adapt and predict unseen data outside of the training set. A fundamental flaw within supervised ML is overfitting, where the model has been trained too well on training data resulting in noise and specifics of the set being memorised (Jabbar and Khan, 2014). Therefore, overfitting limits the model's ability to generalise well on both the observed data within the training set and unseen data in the testing set. Large deviations of the predictive scorings between the training and testing set are a common indication that a model suffers from weak generalisation, leading to the validity of external predictions to be questioned (Dexter et al., 2020). Although the causes of overfitting may be complex, the sources of the phenomenon were classified into three types by Ying (2019). Firstly, through the learning of noise, or irrelevant information, within the training set, whereby specific trends within the training data later act as a basis for predictions. Next, dependent upon complexity of the hypothesis (i.e., the compromise between variance and bias) such that when a model contains too many features, the accuracy may increase at the sacrifice of lower average consistency due to the increase in model complexity. Lastly from the usage of multiple comparison procedures, which are ubiquitous to induction and AI algorithms, where scores from an evaluation function are compared for multiple items with the maximum score being selected. However, items that achieve the highest scoring are not guaranteed to improve model performance and may even reduce accuracy. The complex issue of capturing all areas of a ML algorithm that result in overfitting models may not be accurately reported, therefore issues of uncertainty within performance will undoubtedly be raised.

4.3.6.4. Extending the uncertainty assessment criteria to better evaluate ML models

Development of QSAR models through the use of ML techniques inherently presents its own set of additional issues that affect the validity and uncertainty of predictions. The assessment of these challenges has demonstrated that the current uncertainty scheme and related criteria require further development to ensure that ML models can be evaluated accurately. As shown in Figure 4.7, these concerns have widespread implications upon all phases of QSAR development, affecting a multitude of components. Therefore, to ensure that the assessment criteria are suitable for a variety of modelling approaches, extension, and improvement of the relevant assessment points, as reported in Table 4.5, are suggested. To achieve this, knowledge gained throughout development of ML algorithms and literature has been incorporated into the supplementary information of the criteria.

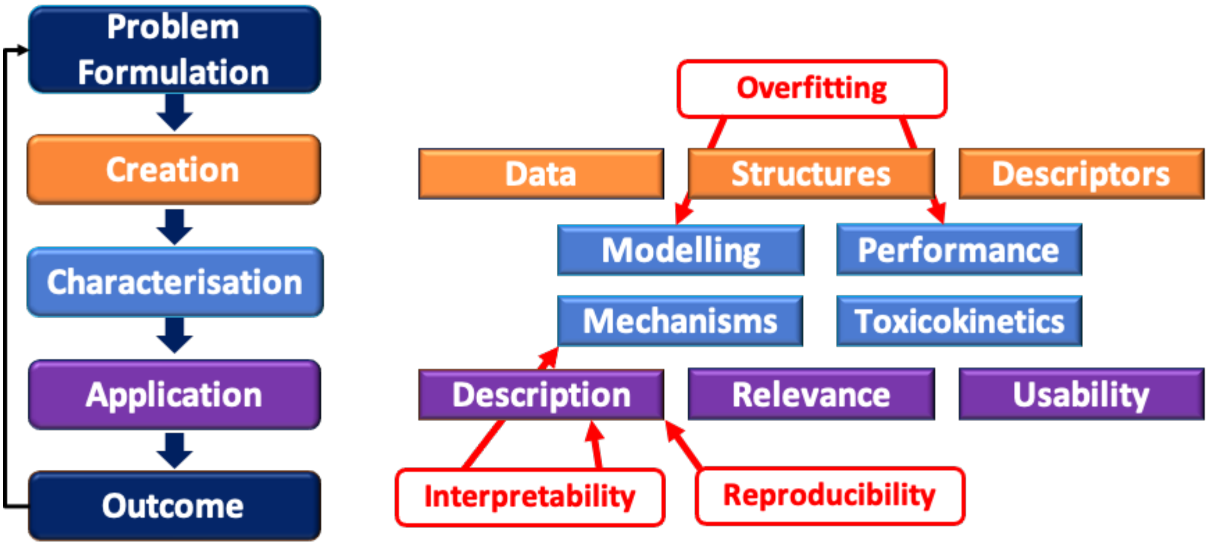


Figure 4.7. Summary of the additional considerations for ML and AI (shown in red text) and their respective components of QSAR uncertainty that they affect.

Concerns of reproducibility and generalisation are well considered within the current criteria and may only require small amendments to enable all potential areas of uncertainty to be captured. With regards to reproducible ML, much research has been conducted into this topic, leading to the development of reporting schemes from many disciplines (Pineau et al., 2020; Heil et al., 2021; McDermott et al., 2019). The knowledge provided from these reporting formats is universally relevant, hence has been incorporated within the applicable criteria to fill gaps that have been overlooked. Similarly, the occurrence of overfitting is a well-documented drawback of ML methods and, as such, a breadth of information to avoid such phenomena occurring have been suggested (Ying, 2019; Dietterich, 1995; Ghojogh and Crowley, 2019). Declaration and employment of the various methods (i.e., cross-validation, regularisation, and early-stopping) that are globally or locally available to ML models should therefore be encouraged, not only as a means of good modelling practice but additionally to reduce uncertainty.

Interpretability is the final aspect that has been updated, although considerable discussion is still required with varying opinions regarding the quality of interpretation techniques. Comparatively, the interpretability of ML and traditional QSAR modelling practices are undoubtedly equivocal. Linear regression models are by their nature model-based techniques, which are heavily reliant upon *a priori* statistical statements (Gao et al., 2018). Specifically, the underlying process that results in the observations is modelled, and so explanatory power is inherently possessed as well as predictive ability (Guha, 2008). Whereas ML algorithms can be defined as model-free methods, which react to the intrinsic data characteristics without being restricted to prior knowledge and are limited to fewer assumptions (Gao et al., 2018). A plethora of explanation methods and techniques are available for ML interpretability, with these initially being separated into two categories: model-specific and model-agnostic. Model-specific interpretations are limited to specific classes and can be interpreted from the inner workings of the algorithm. Whereas model-agnostic techniques are applicable to any model and are applied *post hoc* (Molnar, 2019). Interpretability methods can also be separated dependent upon the results that they provided, being either: feature summary, model internals, data point, or surrogate intrinsically interpretable model (Carvalho et al., 2019). However, for QSAR models these approaches are typically separated into either feature-based or structural interpretation. Whereby definition, feature-based strategies achieve interpretability through the importance of individual descriptors, in comparison to structural interpretations that directly outline particular chemical motifs (Matveieva and Polishchuk, 2021). As can be seen, the field of interpretation techniques and methodology within ML is vast, and the identification and application of all such methods within the realms of QSARs is out of the scope of the current work. However, it is worthy of note that such interpretation strategies are yet to demonstrate their applicability to interpretation of QSAR models with no suitable benchmarks currently being available (Matveieva and Polishchuk, 2021). In addition, feature-based approaches may only provide an overview without sufficiently detailing the structure-activity relationship that has been encoded by the model (Guha, 2008). Despite this, with the inevitable rise of ML approaches within the field of QSAR research into the improvement of interpretability is certainly expected to follow. Thus, it is essential that such techniques are to be appreciated within the uncertainty criteria that enable ML models to be explained from which mechanistic rationales can be derived.

Table 4.5. List of those assessment criteria for individual areas of uncertainty, variability or bias within toxicity-prediction QSAR (as presented by Cronin et al., 2019) updated in light of

consideration of concerns specific to application of ML. Each is grouped in accordance with its relevance either to the reproducibility, interpretability or generalisability of models. Updates to text under heading “comment or other information” are displayed in italics. Please refer to *Appendix V* for presentation in context of unabridged scheme.

ID	Assessment criteria*	Comment or other information
	Reproducibility	
2.1a	Definition and description of model (related to assessment criterion 3.1a)	All terms e.g., descriptors, statistical values, *hyperparameters and ranges*, algorithms should be defined. The QMRF is a possible reporting format.
2.1c	Transparency of the model	*A transparent model can be reproduced, and the model output is (reasonably) interpretable, i.e., user can understand the causation of a prediction.*
3.1a	Reproducibility of the model or QSAR (related to assessment criterion 2.1a)	To determine reproducibility, the model is assumed to be transparent (see assessment criterion 2.1c). *Source code should be provided, with computational infrastructure detailed.*
3.1b	Reproducibility of the QSAR prediction	To obtain reproducible predictions, all parameters (descriptors) need to be available and controllable. *Seeds to control randomisation for certain algorithms need to be specified.*
	Interpretability	
2.1c	As above	As above
2.4c	Relevance of descriptors to mechanism of action/AOP	*Feature importance techniques should be used for algorithms that employ large quantities of descriptors, relating highest scoring descriptors to the mechanism.*
	Generalisability	
1.5a	How appropriate is the modelling approach for the endpoint and to deal with the complexity/non-linearity of the data	This requires a pragmatic and subjective assessment, e.g., a data set based on one mechanism with a single overriding descriptor can be modelled more simply than a more complex scenario. *If applicable, both the optimisation procedure and the sufficiency of resulting approach complexity should also be considered.*
2.2a	Statement of statistical fit, performance and predictivity	The use of appropriate validation methods, *resampling techniques*, and/or external test sets should be demonstrated, different metrics may be required for different models.
2.2b	Interpretation of statistical fit etc with respect to biological measurement error and variability	*The use of strategies to limit overfitting (e.g., early-stopping, pruning, regularisation) may be required for certain algorithms.*

4.4. Good practice in the ML modelling of toxicity

The evaluation of ML for toxicity prediction, and their associated uncertainties, has enabled the identification of areas of good practice that are required in order to improve the acceptability of ML models, particularly to support chemical safety assessment.

- The biological data to be modelled should be evaluated in terms of their quality, consistency, coverage of mechanisms etc.
- The outcome of the evaluation of the biological data to be modelled should be used to assist in problem formulation, particularly to provide realistic (and not-overly optimistic) performance targets.
- Well performed feature selection is required to reduce noise and collinearity. Fewer descriptors are also likely to assist in interpretability. If feature selection is not included, then some rationale should be stated.
- Descriptors must be appropriate to model the effect, i.e., they must relate in some way to the putative mechanisms of action. It is accepted for large datasets, full definition of mechanisms of action is unlikely, but the model and descriptors utilised should be justified and interpreted as best possible.
- Once modelling is complete, use all approaches to evaluate models including model performance, interpretability and uncertainties.
- 10-fold splitting, or thereabouts, is optimal for cross validation. Beneath this, model performance tends to be understated – a greater number, by contrast, adds little value.
- Relate model performance to data quality i.e., to ensure the model does not overfit the data beyond its limitations.
- Hyperparameters tuned during the optimisation procedure should be declared, with the approach undertaken being sufficient for the quantity of hyperparameters.
- Appropriate algorithm selection may be based upon performance metrics, although complexity and interpretability should be considered depending upon the intended purpose.
- Interpretability of the model is crucial, important descriptors can be identified – SHAP is a useful approach for doing this.

- Identification of important descriptors is not the same as mechanistic interpretability which requires the direct relationship between a descriptor and how the molecule causes toxicity to be demonstrated.
- Provide full documentation of the model and demonstrate the good practice described above.

Chapter 5. Making *in silico* predictive models for toxicology FAIR

Preface:

This work has been published in: Cronin et al., (2023). Making *in silico* predictive models for toxicology FAIR. Regul. Toxicol. Pharmacol. 140: 105385. doi: 10.1016/j.yrtph.2023.105385

This was a multi-author paper. Belfield co-authored the work and contributed to the analysis in this study as recognised in the CRediT authorship contribution statement: Conceptualization; Writing - Review & Editing.

Belfield carried out further analysis to extend the study described in the paper; this additional work is presented in Section 5.5.

5.1. Introduction

The FAIR (Findable, Accessible, Interoperable, Reusable) principles have been universally accepted for sharing data and have become fundamental to data storage since their publication in 2016 (Wilkinson et al., 2016). They are based around good practice for data management and stewardship relating to scientific data, such that data may be discovered and re-used for downstream investigations. The aim is to enshrine good practice of data capture, curation and storage such that they may be available for future researchers thus saving time and resources (Briggs et al., 2021). Regarding chemical safety assessment, access to data relating to the intrinsic hazards of a chemical, as well as its exposure, is highly desirable. As such, areas such as toxicology are increasingly investigating the FAIR principles to make historic and newly determined data more readily available. There are numerous reasons to capture all these data, not only to avoid unnecessary repetition of animal tests and support the implementation of the 3Rs principles (Russell and Burch, 1959), but also due to the cost of testing and possible legal reasons for the avoidance of testing (e.g., including, but not limited to, EU Regulation, EC N°1223/2009 (European Commission, 2009b)).

Chemical safety assessment also relies increasingly on computational modelling. Predictive models in computational toxicology are applied for a variety of purposes in approaches such as Next Generation Risk Assessment (NGRA) and Integrated Approaches to Testing and Assessment (IATA). The models are frequently used to fill data gaps where information may be missing, i.e., a test has not been performed, as well as to provide lines of evidence to support an overall weight of evidence for a particular decision (Mahony et al., 2020). There are a great variety of endpoints and properties that may be predicted, ranging from physico-chemical properties to the prediction of toxicological effects themselves (e.g., regulatory endpoints) or mechanistic information (e.g., binding to a receptor) as well as properties relating to internal exposure such as Absorption, Distribution, Metabolism and Excretion (ADME).

There are a very broad range of predictive models that require consideration. These are often based around a form of quantitative structure-activity relationship (QSAR) model to predict physico-chemical and ADME properties and toxicological effects. More detailed physiologically-based kinetic (PBK) and related models are also available to describe internal exposure. Whilst QSAR was founded in transparent regression analysis models in the 1960s, there is now an enormous diversity to the modelling approaches applied (Madden et al., 2020). This study will focus on "knowledge-based" methods that may support chemical safety assessment. In this context, this implies that the methods are characterised by the fact that they start from a defined piece of knowledge (for example a series of compounds of known biological properties) from which an empirical model (a set of rules that describe a regularity between the properties of the objects) is derived. Such methods have common elements (e.g., a training set of compounds, a computational algorithm, predictive quality parameters) and may be used in QSAR or PBK modelling. These may incorporate a variety of computational algorithms from regression analysis to machine learning approaches. Thus, for the purposes of this study and defining the FAIR principles in the toxicological context, the term "*in silico* predictive model" is used; this is assumed to be any knowledge-based computational algorithm that will assist with the prediction of properties relating to chemical safety assessment. Further detail on the components of predictive models for toxicology is given in Section 5.1.1.

The total number of published, or publicly available, QSARs, PBK and other computational models that could support chemical safety assessment is unknown; a conservative estimate would be 10,000+ models. Likewise, the vast majority of endpoints and chemistries for which QSARs have been developed are currently only sparsely and heterogeneously documented, and not easily searchable. This makes the task of finding a usable model for a particular purpose very difficult. There has been a concomitant growth in the use of software programmes which are freely or commercially available. The reality is that we may be missing out on the opportunity to use potentially valid and useful models, simply due to their lack of accessibility and findability (Worth, 2020). In addition, there is often very poor documentation of existing models, and the existing documentation often contains errors, such that even when a QSAR may be found, it may not be possible to reproduce it (Patel et al., 2018; Piir et al., 2018), a problem being particularly noted in the artificial intelligence community (Knight, 2022).

The aim of this chapter was to set out a vision for the full diversity of *in silico* toxicology models that may be suitable for chemical risk assessment to be FAIR. This was done by assessing the requirements for making predictive models FAIR in *in silico* toxicology, considering the current initiatives to share such models, and how the FAIR principles that are currently aligned for data sharing could be adapted for predictive models. Application of the FAIR principles to previously developed models was also investigated. However, this study did not intend to provide an in-depth methodology of *how* FAIRification of models may be achieved, but to highlight the topic and make recommendations for the steps forward to be made to increase the availability and sharing of predictive models.

5.1.1. Anatomy of an *in silico* predictive model for toxicology

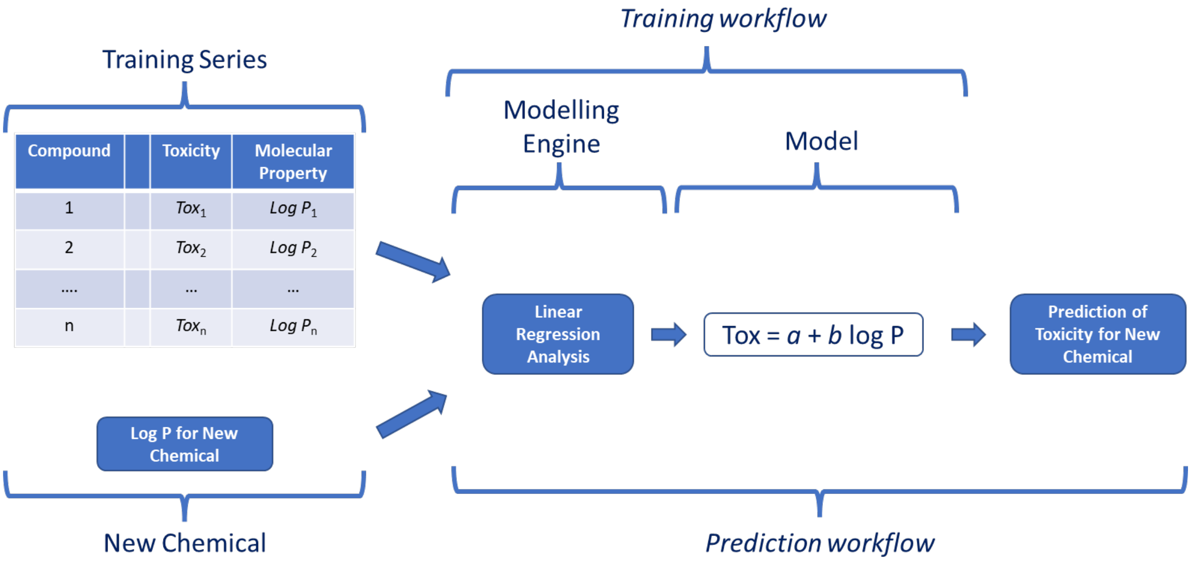
For the purposes of this study, a more detailed description of what we understand by a “model” is provided in this sub-section. In particular, it is important to identify the model components and analyse how they are generated, in addition to whom may own their intellectual property. Once this is established, it becomes easier to determine which components of a model can be shared and how this may be achieved.

Knowledge-based, predictive models result from training a certain “modelling engine” with a collection of objects, often called a “training series” in the QSAR field. The training results in the identification of regularities between properties and annotations of the training series, which are captured in a collection of rules, mathematical functions, or a mixture of both. The outputs are analysed to interpret and understand the relationships between the object properties and the annotations. Characteristic of the models is the expectation of their ability to be applied to new objects so that they can predict annotations from the object properties.

In this generic description of models, the modelling engine describes a component of a predictive workflow, including all the algorithms required to reduce the object properties and annotations to a collection of mathematical variables (descriptors), normalise and scale them appropriately and apply machine learning algorithms. This workflow should have a software implementation to be functional and thus be able to build a model from a training series and predict object annotations for new objects, starting from a previously built model. In this description, we therefore identify the constitutive elements of the models which must be considered in this study:

- The training series
- The modelling engine
- The model

This general description is shown schematically in Figure 5.1 using a simple illustration. In Figure 5.1 a toxicity value is related by regression analysis to a single molecular property, namely the logarithm of the octanol-water partition coefficient (log P), a property that is strongly related to toxicity (Cronin, 2006). In reality, the types of models that may be created could comprise one of many different “modelling engines” with potentially very high dimensionality in property space. The derived model can be used to predict an unknown toxicity for a new compound providing the property value(s) are provided. The latter function, i.e., use of the model, is utilised by the end-user, as noted below this is now often wrapped in a workflow for ease of application. Figure 5.1 also confirms that the modelling engine cannot produce predictions on its own before it is applied to a training series to produce a model. Moreover, the same modelling engine can be used to train an unlimited number of models.



Compound		Toxicity	Molecular Property
1		Tox_1	$Log\ P_1$
2		Tox_2	$Log\ P_2$
….		…	…
n		Tox_n	$Log\ P_n$

Figure 5.1. A schematic representation of a simple *in silico* predictive model for toxicity, namely a regression analysis on one descriptor (logarithm of the octanol-water partition coefficient (log P)), showing the interrelationship between the components of the model and the workflows for training the model and making predictions (the data for the new chemical may flow either into the analysis, e.g. for normalisation, or the model itself).

As a consequence of the complexity of what comprises a model, the model can be shared in different ways. For example, a modelling engine connected to a collection of models can be made available online, thus allowing users to predict the annotations of new compounds. This shared model does not require any access to be given to the model itself, which is only visible via the modelling engine. Moreover, access to the modelling engine can be limited to using pre-built models for prediction or allowing other functionalities, such as retraining existing models or developing new ones. Examples of this method are online modelling servers including oCHEM (Sushko et al., 2011) or the QSAR DB (Ruusmann et al., 2015).

Other means of model sharing include the distribution of the pre-built models in computational formats that locally installed instances of modelling engines can use (the so-called workflow in Figure 5.1). This method requires access to the modelling engines, ideally as open source. Examples of this method are models distributed as KNIME workflows (Steinmetz et al., 2015) or models developed using Flame (Pastor et al., 2021).

Regarding ownership of the model and intellectual property rights, it is also essential to consider the model components. Model developers own the results of the modelling, i.e. the model itself. When sharing models using an online server, the model owner can limit access to the prediction functionality on a per-model basis. When a proprietary modelling engine is used for model building, the modeller owns the resulting model even if the use of these models for carrying out a prediction could require access rights to the prediction functionality of the modelling engine.

5.2. Need for FAIR *in silico* predictive models for toxicology

In silico predictive models in toxicology are typically built on data for chemicals (with defined structure) adding value by creation of predictive capability. The data may represent any aspect of chemical safety assessment, but mainly are based on the endpoints needed to make a safety assessment decision, e.g., the endpoint required for a regulatory submission. The numbers of compounds used to train the model may vary from as few as 5-10, up to the 1000s or even more. As such, a number of different types of modelling algorithms have been applied, with machine learning approaches being seen as the solution to the largest data matrices. The models are based on the properties, or calculated structural descriptors, of molecules that should, in theory at least, be responsible for the biological effect and, where assessed, potency (Madden et al., 2020; Cronin et al., 2022). As noted above, this study concentrated on knowledge-based models.

There are many uses for *in silico* models in chemical safety assessment, ranging from the rapid screening of toxicity in chemical libraries through to acting as surrogates for tests in regulatory submissions. For the latter, protocols have been established to provide means to evaluate a model with a view to making predictions from them acceptable for a particular purpose, e.g., the OECD Principles for the Validation of (Q)SARs (OECD, 2007) and criteria for the characterisation of uncertainties (Cronin et al., 2019). These principles have enabled frameworks to capture QSAR models – notable being the QSAR Model Reporting Format (QMRF) (Worth, 2010). However, there are no standardised means or requirement to share the models. The current lack of model sharing policies constitutes a clear argument for advancing towards the definition of a FAIR models' policy.

It is clear that making models FAIR will assist in the capture, discovery and sharing of QSAR and PBK models and numerous other approaches. It also provides an opportunity to develop and standardise the documentation of models. In addition, making models FAIR will support the independent verification of models which will, in turn, improve trust in models. This will allow for greater use of models to make predictions and encourage global harmonisation of models and modelling approaches. It will also ensure greater reproducibility of models, the lack of which has been highlighted as a fundamental issue (Patel et al., 2018; Piir et al., 2018), enabling the replication or re-use of data. Progress in toxicology is already underway with efforts to standardise approaches and improve collaboration (Martens et al., 2021). Likewise, there has been recent progress in the FAIR Principles for Research Software, the so-called FAIR4RS principles (Chue Hong et al., 2022). There will be a mutual benefit in aligning the FAIR principles for *in silico* models for toxicology with the FAIR4RS principles.

It is not only essential that researchers can find models easily and efficiently, but also to support regulatory submissions from modellers. With regard to regulatory submission, the IMI2 eTRANSAFE (Enhancing TRANslational SAFEty Assessment through Integrative Knowledge Management) project, building on the foundations of the IMI1 eTOX (Integrating bioinformatics and chemoinformatics approaches for the development of expert systems allowing the *in silico* prediction of toxicity) project, has developed a variety of *in silico* models to support the safety assessment of pharmaceuticals (Pognan et al., 2021), including a framework for a cooperative development of predictive models and their usage (Pastor et al., 2021). Previous work in these projects has developed a scheme to demonstrate verification of models and reproducibility of predictions (Hewitt et al., 2015). Such a scheme, to provide evidence that a model is FAIR, will subsequently increase confidence in the models and their predictions, and in particular regarding the use of predictions in regulatory submission.

5.3. Current initiatives to share *in silico* toxicology models

There have been several prior attempts to support the sharing of *in silico* models for toxicology. A non-exhaustive selection of these resources is summarised in Table 5.1. It is noted that not all the resources listed in Table 5.1 are for sharing models directly - it also includes protocols and general information resources. The resources offered in Table 5.1 represent a wide variety of approaches ranging from commercial to publicly available, those offering a predictive capability (i.e., a chemical structure can be entered to obtain a prediction) and those without this capability, as well as formats and approaches to capture models and other resources. Of the resources identified in Table 5.1, it is arguable that the

QSAR DB goes the furthest to achieving FAIR principles for the sharing of models, with reference to making QSAR FAIR made on their website. There also exists a huge number of databases containing information that may support the generation of *in silico* models (Pawar et al., 2019), with these acknowledged but not summarised in this section.

Table 5.1. A selection of resources available to assist in the sharing of *in silico* models for toxicology

Resource	Description	Source	Reference(s) and / or URL
Databases and other compilations of models, with predictive capability			
C-QSAR	A licensable collection of over 18,000 regression based QSARs for a large number of endpoints	BioByte Corp., Covina CA, USA	http://www.biobyte.com/bb/prod/cqsarad.html; Kurup (2003)
COSMOS NG	A freely-available knowledge hub with predictive capability and links to *in silico* models and profilers	MN-AM, Nürnberg, Germany; Columbus OH, USA	https://www.ng.cosmosdb.eu/; Yang et al., (2021)
Danish QSAR Database	A freely-available on-line repository of QSAR model estimates for more than 600,000 substances including physico-chemical properties, environmental fate, bioaccumulation, eco-toxicity, absorption, metabolism and toxicity	Danish Technical University, National Food Institute, Copenhagen, Denmark	https://qsar.food.dtu.dk/; Chinen et al. (2020)
eTRANSAFE	A collaborative project aiming at collecting and sharing drug safety related data and developing *in silico* predictive models based on the data	The eTRANSAFE Consortium	https://etransafe.eu/; https://www.imi.europa.eu/projects-results/project-factsheets/etransafe
oCHEM	A freely-available on-line resource that allows for the creation, storage, dissemination and use of QSARs	Helmholtz Zentrum München, Neuherberg, Germany	https://ochem.eu; Sushko et al. (2011)
QSAR DataBase (DB)	An open on-line platform for the organisation, storage and use of QSARs, searchable by a number of criteria. Contains over 500 QSARs which have each been given a unique identifier (DOI).	Institute of Chemistry, University of Tartu, Estonia	https://qsardb.org/; Ruusmann et al. (2015)
Models reporting formats			

In silico protocols	Guidelines on performing expert review of *in silico* models for a variety of toxicological endpoints	Consortium led by Instem, Columbus OH, USA	A large number of articles including Myatt et al. (2018), Ruiz et al., (2018)
OECD Guidance Document on the characterisation, validation and reporting of PBK models for regulatory purposes	A harmonised template to record all relevant information regarding a PBK model	OECD	https://www.oecd.org/chemicalsafety/risk-assessment/guidance-document-on-the-characterisation-validation-and-reporting-of-physiologically-based-kinetic-models-for-regulatory-purposes.pdf
QSAR-ML	An open XML format for the exchange of QSAR datasets		Spjuth et al., (2010)
QSAR Model Reporting Format (QMRF)	A harmonised template to summarise and report the key information of QSAR models		https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm; Worth (2010)
Model repositories, without predictive capability			
GitHub	Free-to-use provision of repositories for the distribution of QSARs, documentation etc., as well as R code, KNIME Workflows and similar tools	GitHub Inc.	https://github.com/
JRC QSAR Model Database	An historical archive of some 150 QMRFs that had been submitted to EURL ECVAM. The archive is no longer updated but may be downloaded free-of-charge.	European Commission's Joint Research Centre, EU Reference Laboratory for Alternatives to Animal Testing (EURL ECVAM), Ispra, Italy	http://data.europa.eu/89h/e4ef8d13-d743-4524-a6eb-80e18b58cba4; EC JRC (2020)
PBK database	A freely available collection of key information for over 7,500 PBK models for	School of Pharmacy and Biomolecular Sciences, Liverpool	Thompson et al. (2021)

	1,150 chemicals with details of modelling software used, species, chemicals etc.	John Moores University, UK	
Other initiatives relevant to the sharing of models for chemical safety assessment			
BioModels	A freely available repository of mathematical models representing biological systems. Whilst most models in BioModels are not relevant to *in silico* toxicology, there are some examples of PBK models. Models generally do not have predictive capability.	European Bioinformatics Institute, European Molecular Biology Laboratory, UK	https://www.ebi.ac.uk/biomodels/; Glont et al., (2018); Malik-Sheriff et al., (2020); Tiwari et al., (2021)
FAIRsharing	A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies encompassing a collection of registries – including some that are applicable to toxicology. The ELIXIR Toxicology Community is making use of this service to collate toxicology standards.	FAIRsharing team	https://fairsharing.org/
Research Data Management toolkit for Life Sciences (RDMkit)	An online guide which contains guidance for data management with a specific page for toxicology data	ELIXIR	https://rdmkit.elixir-europe.org/toxicology_data
RO-crate	A freely available resource which allows packaging of research data with their metadata	The University of Manchester, UK	https://w3id.org/ro/crate
The FAIRcookbook	An online, open and live resource for the Life Sciences to make and keep data FAIR. It contains recipes for FAIRification – some of which are directly applicable to toxicology or model inputs.	ELIXIR	https://faircookbook.elixir-europe.org/content/home.html

5.4. Development of FAIR principles for *in silico* models

The FAIR principles, originally devised for data sharing, are herein adapted to the needs of computational modelling. It is important to understand the context of the FAIR principles related to data sharing, which aimed to "*define characteristics that contemporary data resources, tools, vocabularies and infrastructures should exhibit to assist discovery and reuse by third-parties*" (Wilkinson et al., 2016). With regards to sharing *in silico* models, all of these concepts are valid, especially with the overall concept of facilitating "*discovery and reuse"* in addition to the other benefits, such as verification and trust, noted above, which will improve the utility and acceptance of models. Whilst the FAIR principles for data sharing do not specifically include verification and trust, they do indeed go further in other areas, emphasising the requirement "*to improve knowledge discovery through assisting both humans, and their computational agents, in the discovery of, access to, and integration and analysis of, task-appropriate scientific data and other scholarly digital objects*" (Wilkinson et al., 2016). Within the context of *in silico* predictive models, this is taken to mean that the model itself should be shared, in a usable form either directly (by sharing an accessible prediction service) or indirectly (by sharing the components and precise instructions to reproduce the model).

Following the spirit of the FAIR principles for data sharing, the FAIR requirements were adapted in the context of *in silico* predictive models. Specifically, these requirements intend to ensure that a model can be located, i.e., it is *Findable*; that once located, the model and appropriate meta-data are retrievable, i.e., it is *Accessible*; the model is defined in a manner that it can be integrated with other software, i.e., it is *Interoperable*; and that predictions can be made by a robust, well-annotated version of the model, that will make the same predictions regardless of the platform and software used, i.e., it is *Reusable*.

The FAIR principles for the sharing of *in silico* predictive models are summarised below (principles marked with an asterisk are the same, or adapted from, those for data sharing):

To be *Findable*:

F1*. Each model is assigned a globally unique and persistent identifier and different versions are assigned distinct identifiers

F2. Models are described with rich meta data covering all aspects of the model, for example:

F2.1 Models are associated with searchable meta data for the property or endpoint to be predicted

F2.2. Models are associated with searchable meta data or descriptions of the chemicals (e.g. InCHI or SMILES), or chemical class(es), within the model, or a description of its applicability domain

F3. Models are registered or indexed in a searchable resource

F3.1 Models' identifiers should be optimised to allow for use in multiple search engines

F4*. Models' (meta)data clearly and explicitly include the identifier of the model they describe and are registered or indexed in a searchable resource

To be *Accessible*:

A1*. Models are retrievable by their identifier using a standardised communications protocol

A1.1. The model (and any associated protocol) is openly accessible or reimplementable

A1.2. The model (and any associated protocol) allows for an authentication and authorisation procedure, where necessary

A2. Model (meta)data are accessible even when the model is no longer available, unless restricted for commercial, ethical or data protection reasons (e.g., blinding of confidential chemical structures)

To be *Interoperable*:

I1. The models and their (meta)data are described in a standardised manner, i.e., standards to define chemical structures, endpoints, molecular descriptors and modelling algorithms

I2. The model reads, writes and exchanges data in a way that meets domain-relevant community standards

I3. The model must be able to integrate with other software, e.g., with a clearly defined input / output i.e., with an appropriate Application Programming Interface (API) for shared web services

I4*. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

I5*. (Meta)data use vocabularies that follow FAIR principles

I6. The model includes qualified references to other objects, such as molecular descriptors

To be *Reusable*:

R1. The model is available for its use in some format (e.g., source code, executable, library or service)

R2. The usage license of the model should be clearly defined and appropriate to encourage its use

R3. The storage of the model and (meta)data should be done on a sustainable and future-proofed platform, anticipating the impact on the availability of software changes over time

R4. Software includes qualified references to other software, e.g., so that the correct molecular descriptors can be obtained, either as part of the model or storage of the molecular descriptors software or experimental protocol

R5*. (Meta)data are richly described with a plurality of accurate and relevant attributes

R5.1.* The model and its (meta)data are associated with detailed provenance

R6*. The model and its (meta)data meet domain-relevant community standards for documentation

5.5. Model assessment utilising the FAIR principles for *in silico* models

In total, 18 principles have been developed and adapted for *in silico* models that cover all aspects of the FAIR ideology. Each of the individual principles provides guidance and considerations for developers that once adhered to will foster a model that has been produced, labelled, and stored in a manner that fully promotes shareability and can be categorised as FAIR. Akin to the work produced in Chapter 2, evaluation of models through application of the FAIR principles can highlight issues within a given workflow, which in turn may be hindering shareability. Demonstrating the ability of the principles to be utilised in this manner, Table 5.2 evaluates the ability of the six previously developed ML models (see Chapter 4) to satisfy the FAIR criteria. Due to the development of each model being identical, and only differing dependent upon algorithm utilised, the models were evaluated as a collective. To this end, the ability of the collective models to satisfy each individual principle was determined by a lead researcher and subsequently verified by another researcher. Each principle was assigned a classification being either 'Yes' (the principle was fully satisfied), 'Partially' (the principle was somewhat satisfied), or 'No' (the principle was not satisfied). Each score was provided with an accompanying reasoning for transparency. In addition, scenarios where 'Partially' or 'No' were recorded a potential improvement strategy was also provided. As seen in Table 5.2, the majority of principles are reported to be sufficiently satisfied, however there exists a handful that are not. Whilst the ratio of successes outweighs that of failures it is essential that for a model to be considered FAIR all principles are sufficiently satisfied, as such mitigation strategies are required.

Table 5.2. Results of the assessment from the models (considered as a whole) towards each FAIR principle and respective improvement strategies as required.

FAIR Principle ID	Verdict and reasoning	Improvement strategies (where applicable)
Findable		
F1	**No.** Models have only been assigned local identifiers that are associated throughout development.	There is a clear need for the models to be assigned a unique global identifier such as a Digital Object Identifier (DOI).
F2.1	**Yes**. Models are developed with searchable meta data for the endpoint of interest which is publicly available.	
F2.2	**Partially**. Unique chemical identifiers are provided for the meta data used within the model. However, applicability domains were not distinguished.	Models need to be associated with a clearly defined applicability domain.
F3.1	**No**. Model identifiers are minimal, providing only barebones information regarding algorithm and data utilised.	Models should be provided with further identifiers that encapsulate the key characteristics including meta data information that would optimise use in multiple search engines.
F4	**No**. Meta data from the models are yet to be made available.	Once meta data are produced, information needs to be stored within a searchable resource with identifiers explicitly related towards the model.
Accessible		
A1.1	**Yes**. Models are openly accessible and stored within a public repository on GitHub.	
A1.2	**Yes**. Models’ full development are publicly available enabling them to be authenticated.	

A2	**Yes**. Meta data is openly accessible and stored within a public repository on GitHub.	
	Interoperable	
I1	**Yes**. Models and their meta data are described in a standardised manner that can be located within the associated documentation.	
I2	**No.** Models exchanging of information cannot follow domain-relevant community standards until these have been proposed.	Once the community standards have been proposed, evidence that the models follow them should be provided.
I3	**Partially**. Clearly defined inputs and outputs for the models are outlined, however no API exists for them currently.	Models need to be further developed with clear consideration for how the API must be presented to be appropriate for shared web services.
I4	**Yes**. Meta data are all described using accepted identifiers and ontologies for knowledge representation.	
I5	**Yes**. Meta data adheres to the FAIR principles.	
I6	**Yes**. Objects outside of the original meta data that have been produced are appropriately referenced to original sources with the information additionally being publicly available.	
	Reusable	
R1	**Yes**.	

	The models are available for their intended use within an executable source code.	
R2	**No**. No usage license for the models have been provided.	Models need to be accompanied by a usage licence that actively encourages their usage.
R3	**Yes**. Models and meta data are stored within a public repository on GitHub, which is globally accepted as a sustainable platform.	
R4	**Yes**. All software used throughout the development of the models and production of descriptors for the meta data are accurately referenced in the associated publication.	
R5.1	**Yes**. The origins of the meta data are clearly provided within the associated publication.	
R6	**No.** Domain-relevant community standards for data documentation are unavailable.	Once community standards for data documentation are proposed such procedures must be adhered to.

Whilst failures of individual principles may be viewed as potentially minor and may simply be overlooked, it is essential that all are upheld. To this end, following individual assessment the identification of appropriate mitigation strategies (as required) should be promoted. Examples of which are further demonstrated within Table 5.2. Specifically, in relation to the models being *Findable*, the assessment revealed issues regarding model identifiers as well as a lack of applicability domain and predictive meta data. Whilst the latter of the two issues reflects the infancy of the models, and may be resolved through further development, addressing the model identifiers requires consideration of improved labelling strategies that captures greater information regarding each model. In particular, a decision on selection of an appropriate global identifier that is assigned to each model needs to be considered, with a possible solution being a unique DOI. With regard to models being *Interoperable* as well as *Reusable,* a lack of consideration as to how models can be incorporated into other web services, with an appropriate license, is identified. To move this forward, improved consideration regarding the models' implementation within other web services must be considered. Additionally, the licence associated with the model should ensure that it is actively promoting usage with only essential restrictions where required. Additional issues within both *Interoperable* and *Reusable* principles is the lack of the data, and the fact that the workflow(s) do not follow community standards – the reason for this being that no such standards have yet been proposed. As compared to the uncertainty assessment criteria, which can be used as a tool to demonstrate fitness-for-purpose with varying levels of uncertainty being accepted, it may be expected that all FAIR principles should be satisfied. Therefore, reflecting upon issues identified through application of the principles, and taking remedial action where possible, is vital to the success of models being considered FAIR.

As demonstrated, these principles may be used *post hoc*, however it is reasonable to assume that the most effective usage is as guiding principles that are continually considered throughout the model lifecycle. Enforcement of the principles during the development phase will ultimately ensure the production of harmonised models that can be easily shared. Examples of considering the FAIR principles at the core of data development can already be observed within other industry standards, such as the bioinformatics communities, with it widely being accepted that for data sharing to flourish the FAIR principles must be upheld. For this reason, national level infrastructure is being updated to encapsulate the FAIR principles, with initiatives also being undertaken to converge industry standards (Mayer et al., 2021; Vesteghem et al., 2020). Whilst the bioinformatics researchers have been considering FAIR principles for the past few years, such workings provide a view towards the future which has the potential to be adapted for the field of *in silico* modelling.

5.6. Priorities to make models FAIR

The benefits of the FAIRification of *in silico* predictive models go beyond the simple advantages of being able to share models successfully. The benefits include making a usable resource that can assist chemical safety assessment, as well as being interrogated to understand the applicability domain of models, and to determine where data gaps exist in the domain. There is also a societal responsibility to enable access to models created and to record the outputs of research efforts. The modelling community must be challenged to make harmonised and usable models. This will reinforce the credibility of models and demonstrate responsible, ethical, transparent and efficient science. The acceptance of the understanding and promotion of the FAIR principles for modelling globally is proposed as a starting point, even if the finer details still need to be resolved.

There will inevitably be a number of issues that require further development and acceptance, beyond the current state-of-the-art. Widescale sharing of models will need appropriate investment in the repository(ies) and resources to maintain the platform on which any repository is based. Comparable efforts to store models do exist e.g., BioModels, and remain active and on-line due to the creation of an appropriate business model. Relying on free storage resources is one way forward, but will be extremely limited in terms of the search capabilities and practical use.

The FAIR principles on accessibility do not preclude restrictions on access but they do require metadata longevity and for the access protocols and access authorisation used to adhere to open standards and be clearly defined (Wise et al., 2019). However, in the case of training and test datasets used to build and validate the model there may be legal (e.g., IPR protection) and ethical (e.g., patient confidentiality) reasons, as well as commercial ones, that would preclude open access.

Key to the development of any data resource to be used in predictive modelling is the harmonisation of the terminology for reporting models. This could start with harmonised ontologies for endpoints, for which much work has already been undertaken. It will also require harmonised ontologies to describe the models e.g., for statistical and machine learning methods, definitions of molecular descriptors and chemical identifiers. In addition, harmonisation will be required in the definition of the methods for analysis of model performance, such as provided by Walsh et al. (2021). Much of this could be adapted from that already used for QMRF and elsewhere for ontologies for statistics (Zheng et al., 2016).

Lastly, and probably most importantly and urgently, an internationally agreed vision for the future with an associated roadmap is required. Only when stakeholders, including potential funders, agree will progress be achieved.

5.6. Conclusions

There is an undoubted, and urgent, need to make *in silico* predictive models for toxicology FAIR. This is an achievable goal and, given appropriate resources, much progress could be made in the short to medium term. There are numerous reasons and benefits to the FAIRification of *in silico* models, most fundamental is to make models available and accessible to all enabling and supporting the 3Rs. It is highly probable that chemical risk assessors are missing out on opportunities to use *in silico* models simply as they may not know of their existence. Similarly, due to poor documentation, *in silico* models may be used inappropriately, e.g., out of applicability domain or for the incorrect endpoint. The ultimate sustainability of *in silico* models is also a key advantage. It is unacceptable that research efforts should be placed into modelling, often from public funding, that are unfindable or unusable. Finally, having open and transparent models, easily accessible, will increase trust for all users. This will be especially important for regulatory submissions where agencies can re-run models to check predictions for the target and similar compounds.

In order to achieve the goal of making *in silico* models in toxicology FAIR, the priorities and an overall strategy should be devised. This will need agreement at multiple levels, across industrial sectors, stakeholders and geographical regions. The intention is that the FAIR principles described in this study will act as a template for FAIR principles to be applied to all models of biology.

Chapter 6. Discussion

Presented in this final chapter is an overall summary of the research undertaken throughout the thesis, and how such work addresses the issue of "Increasing the Confidence of *In Silico* Modelling in Toxicology" and associates issues as highlighted within Chapter 1. Whilst full discussion of the work undertaken may be found within the respective chapters, key findings are summarised here. Contained within the summary of each chapter is identification of the strengths and limitations of the research conducted. Chapter 6 concludes with a view to the future, outlining the potential work that may be undertaken following the research conducted in this thesis. The aim of discussing the future work was to provide a vision towards the implementation of the frameworks, gathered throughout the thesis, that can be structured in a manner that would promote regulatory acceptance of *in silico* models.

6.1. Summary of work

Throughout the thesis, research has been conducted with an overarching theme to improve the acceptance of alternative methods to the use of animals for toxicological risk assessment, with a specific focus on *in silico* approaches and QSARs. The motivation for the research in the thesis arose from the vision that outlined the need for NAMs in toxicological risk assessment (US NRC, 2007; Krewski et al., 2020). Whilst there is a clear desire to incorporate such methods, in reality, their acceptance has not kept pace with scientific development. The issue of acceptance has undoubtedly severely hindered the use of NAMs in chemical safety assessment. As presented in Chapter 1, a recent study by Mahony et al. (2020) highlighted some of the key challenges that are hindering the implementation of such methods, some of which were addressed, in part at least, by the research conducted throughout the thesis. At the core of the issues faced by *in silico* methods is a lack of model understanding, requiring a clear strategy to appropriately validate them, enabling the strengths and weaknesses of the model, in terms of uncertainty, to be fully acknowledged. These crucial challenges were addressed throughout Chapters 2-4 of the thesis.

Chapter 2 brought attention to the recently devised QSAR uncertainty assessment criteria produced by Cronin et al. (2019). These criteria were expanded upon throughout the thesis. To this end, the objective of Chapter 2 was to demonstrate the ability of the criteria to determine fitness-for-purpose. To achieve this, the original 49 uncertainty assessment criteria (published by Cronin et al, 2019) were rationalised and organised to form ten components, each of which relates to a key phase of QSAR modelling – creation, characterisation, and application. Consolidation of the original criteria into the ten general assessment components provided a clear benefit of enabling a comprehensible overview of uncertainty for an individual QSAR model to be established. As such, particular areas of uncertainty relating to a given model could be defined. In addition to being able to pinpoint areas of uncertainty using the components, levels of acceptable uncertainty for a particular purpose were proposed; in turn, enabling a verdict for fitness-for-purpose for an intended use to be easily deduced. Demonstrating this capability of the assessment components, a case study of twelve recently published QSAR models was evaluated. Following the evaluation of each model, common areas of high uncertainty were reported, with these issues relating to data quality, descriptor transparency, consideration of the mechanism of action, and endpoint relevancy for regulatory use. Evidently, these issues reduce the applicability to regulatory use, as such the assessment components supported the improvement of the QSAR models to gain acceptability through targeted mitigation strategies. Whilst the research conducted demonstrated how the uncertainty assessment criteria could be utilised to address fitness-for-purpose, limitations within the methodology exists. In particular, the assessment of the twelve case studies assigned scores were only validated by a singular internal researcher. As such, without further external validation, the scorings assigned may be influenced by biases, leading to inaccurate classifications.

Chapter 2 outlined the value of the uncertainty assessment criteria in supporting the acceptance of QSAR models for regulatory purposes. Whilst the study demonstrated use cases for traditional single chemical issues, Chapter 3 further investigated how the criteria could be employed in relation to a topic of key interest in recent years – mixture assessment. To this end, research in Chapter 3 commenced with a detailed analysis of the different approaches that have been used throughout the development of QSAR models to predict the effects of mixture toxicity. In total, 40 toxicologically-based studies were collected, with each being categorised based on the key characteristics of QSAR models for mixtures. These characteristics summarised information regarding chemical classification, mixture composition, testing species or system, endpoint modelled, formulation of molecular descriptors, and modelling approach. Analysis of the characteristics within the literature identified recurring trends that were present throughout, for instance, there were many examples of binary mixtures at single concentration ratios modelled in an additive manner. Collection of the literature in this manner additionally enabled a general appraisal of the current state of QSAR mixture modelling. Alongside a call for further modelling efforts and data availability, the standout issues presented throughout included a greater emphasis on potential interaction effects, with an improved effort to investigate realistic exposure scenarios.

Lessons learned from the mixture review in Chapter 3 additionally enabled the opportunity to bolster the original QSAR uncertainty assessment criteria so that mixture-specific considerations could be evaluated effectively. To this end, additional guidance relating to the existing criteria was suggested including providing further direction on what information is to be expected of mixture data, as well as outlining the need for mixture-specific validation approaches. Lastly, an additional criterion was proposed, with this specifically assessing the uncertainty associated with the usage of mixture descriptors. Whilst the research undertaken in Chapter 3 presented the theoretical landscape of QSAR mixture assessment, limitations in the methods undertaken were observed. In particular, the literature review only accounted for toxicity-based studies; whereas non-toxicological-based-studies and research upon essential oils and nanoparticles were excluded during the screening procedures. Characteristics obtained from these studies may have contributed to categories with information deficits, such as the investigations into multi-component mixtures.

Building upon the work in Chapter 3, Chapter 4 addressed the most active field of research currently being undertaken within QSAR studies – ML. ML methods have come into focus recently due to their exceptional predictive performance, with such approaches likely to become even more commonplace in the coming years. Whilst the statistical benefit compared to traditional modelling approaches is evident, comprehension of the model and confidence in the methodology have hindered acceptance. As outlined in Chapter 1, there is a clear desire to expand modelling methodologies used within QSARs, thus improving confidence in ML technologies is crucial.

In relation to the acceptance of ML methods, Chapter 4 updated the knowledge contained within the uncertainty assessment criteria to bolster ML-specific considerations.

Identification of such considerations was achieved through the development of six QSAR models, each of which had been established utilising the most commonly employed ML algorithms observed within the QSAR literature. Throughout this process three distinctive themes that require careful consideration were encountered, these being: reproducibility, interpretability, and generalisation. Consideration of such problematic issues within the uncertainty assessment criteria was achieved through the proposal of further guidance within the current criteria. Working towards the improved assessment of reproducibility, a greater acknowledgement of the vast number of parameters needed to develop models was identified; this requires thorough documentation. Interpretability considerations of the ML approaches were addressed through the use of appropriate interpretation techniques, providing insight into the relevance of descriptors employed as well as improved transparency. Lastly, concerns regarding generalisation could be reduced through the application of appropriate resampling procedures, as well as a reflection on the complexity of the approach required.

Accompanying the improved uncertainty assessment of ML models, the identification of good practice for ML approaches was also determined. Contained within these practices was guidance regarding data collection and cleaning, algorithm optimisation and interpretability, validation procedures and, lastly, reporting and documentation. The identification of best practices for ML approaches, as well as extending the scope of the uncertainty assessment criteria to better capture ML-specific issues, provides an improved understanding and assigned credibility to such methods, ultimately working towards the goal of improving acceptance. Whilst the research conducted provided an outline of best practices for ML approaches and improved the uncertainty assessment of such, limitations within the work exist. Most notably, the work focused on the development of models for regression-based outcomes, with classification problems not being investigated here. Whilst it is most probable that the practices and concerns apply to both assessment types, the potential for classification-specific issues may still be present.

Following the work in the previous chapters that expanded upon the original uncertainty assessment criteria, Chapter 5 addressed the need to ensure the shareability and reproducibility of *in silico* models. Motivation for the work once again originated from the challenges identified by Mahony et al. (2020), identifying a desire to apply FAIR principles to the sharing of predictive models. Conversion of the principles was additionally considered to improve the regulatory acceptance of prediction from such models. In total, eighteen principles were developed that covered all aspects relating to models being Findable, Accessible, Interoperable, and Reusable. Similar to the work conducted in Chapter 2, the application of the principles to the models developed in Chapter 4 was undertaken. Following this evaluation, the results demonstrated that most of the models satisfied the majority of principles. However, unlike the uncertainty assessment criteria where varying levels of uncertainty could be deemed acceptable dependent on usage, ensuring that models are fully shareable, requires all principles to be satisfied. Therefore, the principles additionally demonstrated the ability to serve as guidance for the development of improvement strategies. Whilst the principles described throughout the research demonstrate the possibility to serve as a template to ensure predictive models can be successfully shared, limitations within the principles exist. In particular, whilst translating the FAIR principles was successful, further collaboration and engagement within the community are warranted to devise agreed-upon community standards. Such community standards would need to ensure the harmonisation of previous benchmarks that can be employed throughout the various *in silico* fields.

6.2. Main contributions of thesis

The most notable contributions towards research and knowledge from the thesis include:

- A greater understanding of how uncertainty within QSARs can be assessed throughout the stages of model development. This was demonstrated by utilising the uncertainty assessment criteria. The assessment criteria were grouped into ten components, showcasing the utilisation of uncertainty assessment criteria as a tool for determining fitness-for-purpose. Furthermore, it was established that the uncertainty criteria provide essential support in mitigating areas of high uncertainty.

- The identification of the current state-of-the-art of practice for QSAR for mixtures highlighted a greater need for methodologies to better capture multi-component mixtures and differing interaction effects. Specifically, the potential use of AOPs and GNNs was found to offer the greatest potential for the future of the modelling of mixture toxicity. Additionally, utilising information gathered throughout the review enabled the further development of uncertainty assessment criteria, allowing for a more comprehensive assessment of uncertainties associated with QSAR models for mixtures.

- Improved understandings of best practices in ML-based QSAR methodologies were defined. It was deduced that enhanced considerations of data quality, interpretability, and documentation improved the acceptance of ML approaches to support chemical safety assessment. Additionally, this knowledge facilitated the definition of further ML-specific guidance for the uncertainty criteria, outlining considerations to address reproducibility, interpretability, and generalisability.

- Addressing the need for improved reporting strategies for QSAR models to enable efficient finding and sharing, FAIR principles that have streamlined data sharing were adapted for predictive models. This work detailed the information required for a QSAR model to be labelled as FAIR, thereby supporting the ability to efficiently find and utilise QSAR models.

6.3. Future work

Towards the overarching goal of acceptance and implementation of *in silico* models, and in particular QSARs, research throughout this thesis aimed to address the openly acknowledged challenges in this area. Whilst such work has devised improved reporting strategies, to promote a greater appreciation and understanding of such approaches, as well as presented knowledge on state-of-the-art research focuses, obstacles must still be overcome. Throughout the following section, guided by the limitations and outcomes of the thesis, identification of areas for research focus in future are presented and discussed.

6.3.1. QSAR mixture assessment

Chapter 3 presented the current understandings regarding QSAR mixture assessment, specifically considering toxicological studies. Throughout the initial literature harvest, broad searching criteria were employed to capture all potential studies within the field which was later reduced through a screening procedure (see Chapter 3). Omitted from the review were studies based upon non-toxicological studies, as well as research into nanoparticles and essential oils. There now exists an opportunity to perform the same reviewing procedures upon these studies, characterising them in a similar manner to enable comparisons. Whilst the review into toxicological-based mixture assessments observed a clear bias towards binary equitoxic studies, one can expect with the inherent multi-component nature of nanoparticles and essential oils a more diverse result. Findings from such work may provide further guidance upon how current toxicological mixture assessments can be undertaken.

In addition to this, specific challenges within QSAR mixture studies that were identified could be addressed. In particular, one crucial limitation of QSAR mixture assessment was the way interaction effects between components are modelled. Although an additive manner was typically observed throughout most binary mixtures, scaling to a realistic exposure containing an abundance of individual chemicals at varying concentrations may result in underestimations of toxicity. Towards a more complete approach to address interaction effects, a better understanding of when it is appropriate to deviate from a typical additive approach needs to be deduced. One such source of potential knowledge exists within drug combination studies, with such research observing a greater uptake of *in silico* approaches in recent years (Sidorov et al., 2019). In particular a key area of interest is within the employment of ML approaches to deduce synergism, with recent work demonstrating outstanding predictive performance, providing a promising route for consideration within toxicological studies (Preuer et al., 2018). Furthering this, research into the utilisation of GNNs provides an abundance of potential for modelling mixtures, where interaction effects can be defined and incorporated into the architecture of the model itself (Qin et al., 2023). Therefore, an extensive review of the information contained in drug interaction literature should be investigated. Employing the synergistic assessments that are being undertaken in drug interaction studies would enable greater confidence in the assigned interaction effect to be attributed, improving the predictive capability of models.

6.3.2. Machine learning confidence

The use of ML approaches throughout all industrial and informatic sectors has become common place, with no signs of diminishing. The use of ML technologies for the development of QSARs can only be expected to increase, with more complex algorithms being employed. Whilst such approaches are the future for our field, lessons that have guided traditional approaches must not be slackened, nor ignored. Research conducted throughout the thesis has enabled an understanding of ML associated uncertainties to be deduced, however further efforts into ensuring ML approaches align to methodologies previously devised needs to be encouraged. Particularly challenging, within this regard, is the identification of mechanistic rationale. Demonstrated throughout Chapter 4, efforts were specifically undertaken to provide a degree of interpretability for the ML algorithms. Towards this, understandings into feature importance were obtained through the usage of the package SHAP. Explanations from the package have started to gain appreciation within QSAR studies for model understandings, yet now exists an opportunity to relate such information to a mechanistic rationale (Jaganathan et al., 2022; Zhong et al., 2021). Towards this effort, a better understanding of what features are being identified as important to the model and their relevancy to the endpoint needs to be determined. Such work could be undertaken through the investigation of ML approaches, developed upon descriptor pools, which contain strategically incorporated mechanistically relevant descriptors. Once a greater understanding of the association between model and mechanistic interpretability is better understood, an improved level of confidence can be accredited to ML approaches; thus, improving acceptance.

6.3.3. Harmonisation of reporting procedures

Throughout the research undertaken within the thesis, an overarching theme of improving reporting strategies with the implementation of best practices have been developed. Such work has resulted in the improved reporting format for the QSAR uncertainty assessment criteria, as well as the proposal of FAIR principles towards *in silico* models. Whilst such research enables a sound and thorough means of improving the understanding of predictive models, an effort to encourage the uptake of such knowledge needs to be undertaken. Most obviously, this could be completed through the incorporation of the improved criteria and principles into pre-existing reporting frameworks. However, as shown within Chapter 5, a plethora of potential reporting frameworks for QSAR models exist, presenting the issue of which, if not all, must be considered. Evidently, a more conscientious approach needs to be employed. Whilst great amounts of collaborative research efforts have gone into the development of each unique reporting framework, there now exists the opportunity to harmonise such information.

Towards this goal, an opportunity to draw inspiration from the AOP community exists. An integral part of the AOP knowledge base is the AOPWiki (https://aopwiki.org/), which enables AOP information to be effectively captured, shared, and reviewed. Furthermore, this public repository was devised in a manner to actively facilitate collaborative development and engagement within the AOP communities, with information being stored in living documents (Martens et al., 2022). Such living documents provide extreme flexibility to the reporting of AOP frameworks, enabling models to be incorporated at varying levels of completeness, where information deficit queries can be supplemented through external crowd sourcing efforts. Further to this, key information of the networks is also stored whereby they can be actively merged and engage with the existing knowledgebase. Motivated by this, an attractive opportunity exists to develop a similar framework for QSAR models, guided through the harmonisation of previously proposed approaches. Ideally, the development of such a framework would encapsulate the entirety of the QSAR workflow, including assessment schemes whereby a complete understanding of the model is presented. In order to achieve this, proposals outlined throughout the uncertainty assessment criteria could form the basis of guidance required of model development, enabling an evolving level of confidence to be assigned. Furthermore, to ensure that models could be seamlessly shared throughout the framework, information to be stored and collected should be informed by the FAIR principles.

As a community, we need to encourage the development of robust, understandable, and transparent models to gain the credibility needed for acceptance within regulatory use. Evidently, requiring vast amounts of individual documentation that needs to be satisfied will only impede the process, instead calling for a need to pool together our collective resources to form a cohesive and inclusive framework. Ultimately, the development of such a framework may only foster a more inclusive and collaborative environment, moving towards the production of clearly defined models that satisfy regulatory needs.

6.4. Conclusions

Research undertaken throughout the entirety of the thesis has aimed to address some of the key challenges affecting the acceptance of *in silico* approaches in regulatory settings, with a key focus on QSARs. The usability of the QSAR uncertainty assessment criteria was demonstrated to prove fitness-for-purpose through the conversion into ten components each of which related to a key phase of QSAR development. Bolstering the coverage of the criteria, a review into QSAR mixture assessment was conducted, identifying the characteristics and further considerations needed to be acknowledged. Similarly, ML concerns were additionally determined through the development of six commonly employed algorithms. Finally, the recently devised FAIR principles were translated for use in *in silico* modelling, promoting a greater appreciation of information required to effectively share models. Knowledge and reporting formats devised throughout the thesis are envisioned to be implemented into a harmonised framework that would improve acceptance.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Available from: https://www.tensorflow.org/

Agarap AF., 2018. Deep Learning using Rectified Linear Units (ReLU). arXiv:1803.08375 [Preprint]. Available from: https://doi.org/10.48550/arXiv.1803.08375

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

Akiba T, Sano S, Yanase T, Ohta T, Koyama M, 2019. Optuna: A Next-generation Hyperparameter Opimization Framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD ’19). New York: Association for Computing Machinery. pp. 2623-2631.

Altenburger R, Nendza M, Schüürmann G, 2003. Mixture toxicity and its modeling by quantitative structure-activity relationships. Environ. Toxicol. Chem. 22: 1900-15. doi:10.1897/01-386

Ashford JR, 1981. General models for the joint action of mixtures of drugs. Biometrics. 37: 457-74 .

Barros RP, Sousa NF, Scotti L, Scotti MT, 2020. Use of Machine Learning and Classical QSAR Methods in Computational Ecotoxicology. In: Roy K, editors. Ecotoxicological QSARs, Methods in Pharmacology and Toxicology. New York: Humana. pp. 151-175.

Beam AL, Manrai AK, Ghassemi M, 2021. Challenges to the Reproducibility of Machine Learning Models in Health Care. JAMA. 323: 305-306.

Belden JB, Gilliom RJ, Lydy MJ, 2007. How well can we predict the toxicity of pesticide mixtures to aquatic life? Integr. Environ. Assess. Manag. 3: 364-72.

Belfield SJ, Enoch SJ, Firman JW, Madden JC, Schultz TW, Cronin MTD, 2021. Determination of “fitness-for-purpose” of quantitative structure-activity relationship (QSAR) models to predict (eco-)toxicological endpoints for regulatory use. Regul. Toxicol. Pharmacol. 123: 104956.

Bliss CI, 1939. The Toxicity of Poisons Applied Jointly. Ann. Appl. Biol. 26: 585-615. doi:10.1111/j.1744-7348.1939.tb06990.x

Bopp S, Berggren E, Kienzler A, Van Der Linden S, Worth A, 2015. Scientific methodologies for the assessment of combined effects of chemicals - a survey and literature review. JRC Technical Report. EUR 27471 EN: doi:10.2788/093511.

Bopp SK, Barouki R, Brack W, et al., 2018. Current EU research activities on combined exposure to multiple chemicals. Environ. Int. 120: 544-562. doi:10.1016/j.envint.2018.07.037

Breiman L, 2001. Random Forests. Mach. Learn. 45: 5-32.

Briggs K, Bosc N, Camara T, Diaz C, Drew P, Drewe W, Kors J, van Mulligen E, Pastor M, Pognan F, Ramon Quintana J, Sarntivijai S, Steger-Hartmann T, 2021. Guidelines for FAIR sharing of preclinical safety and off-target pharmacology data. ALTEX. 38: 187–197.

Brockmeier EK, Hodges G, Hutchinson TH, Butler E, Hecker M, Tollefsen KE, Garcia-Reyero N, Kille P, Becker D, Chipman K, Colbourne J, Collette TW, Cossins A, Cronin M, Graystock P, Gutsell S, Knapen D, Katsiadaki I, Lange A, Marshall S, Owen SF, Perkins EJ, 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.

Carlsson LS, Samuelsson PB, Jönsson PG, 2020. Interpretable Machine Learning – Tools to interpret the Predictions of a Machine Learning Model Predicting the Electrical Energy Consumption of an Electric Arc Furnace. Steel. Res. Int. 91: 2000053.

Carnesecchi E, Toropov AA, Toropova AP, et al., 2020. Predicting acute contact toxicity of organic binary mixtures in honey bees (*A. mellifera*) through innovative QSAR models. Sci. Total Environ. 704: 135302. doi:10.1016/j.scitotenv.2019.135302

Carvalho DV, Pereira EM, Cardoso JS, 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. Electron. 8: 832.

Cedergreen N, 2014. Quantifying synergy: a systematic review of mixture toxicity studies within environmental toxicology. PLoS One. 9: e96580. doi:10.1371/journal.pone.0096580

Chandrasekaran B, Abed SN, Al-Attraqchi O, Kuche K, Tekade RK, 2018. Chapter 21 - Computer-Aided Prediction of Pharmacokinetic (ADMET) Properties. In: Tekade RK, editor. Dosage Form Design Parameters. Academic Press. p. 731-55.

Chang CM, Ou YH, Liu TC, Lu SY, Wang MK, 2016. A quantitative structure-activity relationship approach for assessing toxicity of mixture of organic compounds. SAR QSAR Environ. Res. 27: 441-53. doi:10.1080/1062936x.2016.1207204

Chatterjee M, Roy K, 2022. Computational Modeling of Mixture Toxicity. In: Benfenati, E. (eds) *In Silico* Methods for Predicting Drug Toxicity. Methods in Molecular Biology. vol 2425. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-1960-5_22

Chen T, Guestrin C, 2016. XGBoost: A Scalable Tree Boosting System. arXiv:1603.02754 [Preprint]. Available from: https://doi.org/10.48550/arXiv.1603.02754

Chen Y-H, Qin L-T, Mo L-Y, Zhao D-N, Zeng H-H, Liang Y-P, 2019. Synergetic effects of novel aromatic brominated and chlorinated disinfection byproducts on *Vibrio qinghaiensis* sp.-Q67. Environ. Pollut. 250: 375-385. doi:10.1016/j.envpol.2019.04.009

Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al., 2014. QSAR Modeling: Where Have You Been? Where Are You Going To?. J. Med. Chem. 57: 4977-5010.

Chinen KK, Klimenko K, Taxvig C, Nikolov NG, Wedebye EB, 2020. QSAR modeling of different minimum potency levels for *in vitro* human CAR activation and inhibition and screening of 80,086 REACH and 54,971 U.S. substances. Comput. Toxicol. 14: 100121.

Chirico N, Gramatica P, 2011. Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. J. Chem. Inf. Model. 51: 2320-35.

Chollet F, 2015. Keras. Available from: https://keras.io

Chue Hong NP, Katz DS, Barker M, Lamprecht A-L, Martinez C, Psomopoulos FE, Harrow J, Castro, LJ, Gruenpeter M, Martinez PA, Honeyman T, Struck A, Lee A, Loewe A, van Werkhoven B, Jones C, Garijo D, Plomp E, Genova F, et al., 2022. FAIR Principles for Research Software (FAIR4RS Principles) (1.0). https://doi.org/10.15497/RDA00068.

Cocu A, Dumitriu L, Craciun M, Segal C, 2008. A Hybrid Approach for Data Preprocessing in the QSAR Problem. In: Lovrek I, Howlett RJ, Jain LC, editors. Knowledge-Based Intelligent Information and Engineering Systems. Lecture Notes in Computer Science: pp. 565-572.

Conley JM, Lambright CS, Evans N, Cardon M, Medlock-Kakaley E, Wilson VS, Gray LE, 2021. A mixture of 15 phthalates and pesticides below individual chemical no observed adverse effect levels (NOAELs) produces reproductive tract malformations in the male rat. Environ. Int. 156: 106615.

Cortes C, Vapnik V, 1995. Support-vector networks. Mach. Learn. 20: 273-297.

Cronin MT, Dearden JC, 1995. QSAR in toxicology. 1. Prediction of aquatic toxicity. Quant. Struct. -Act. Rel. 14: 1-7.

Cronin MT, Richarz A-N, Schultz TW, 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.

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

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

Cronin MTD, Dearden JC, Duffy JC, Edwards R, Manga N, Worth AP, Worgan ADP, 2002. The importance of hydrophobicity and electrophilicity descriptors in mechanistically-based QSARs for toxicological endpoints. SAR QSAR Environ. Res. 13: 167-176.

Cronin M, Doe J, Pereira M, Willett C, 2021. Re: A call for action on the development and implementation of new methodologies for safety assessment of chemical-based products in the EU - A short communication. Regul. Toxicol. Pharmacol. 122: 104911.

Cronin MTD, Enoch SJ, Madden JC, Rathman JF, Richarz A-N, Yang C, 2022. A review of *in silico* toxicology approaches to support the safety assessment of cosmetics-related materials. Comput. Toxicol. 21: 100213.

Cronin MTD, Enoch SJ, Mellor CL, Przybylak KR, Richarz A-N, Madden JC, 2017. *In silico* prediction of organ level toxicity: Linking chemistry to adverse effects. Toxicol. Res. 33: 173-182.

Cronin MTD, Richarz A-N, 2017. Relationship Between Adverse Outcome Pathways and Chemistry-Based *In Silico* Models to Predict Toxicity. Appl. Vitro Toxicol. 3: 286-297. doi:10.1089/aivt.2017.0021

Cronin MTD, Jaworska JS, Walker JD, Comber MHI, Watts CD, 2003. Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. Environ. Health. Perspect. 111, 1391-1401.

Cronin MTD, Yoon M, 2019. Computational methods to predict toxicity. In: Balls M, Combes R, Worth A. The History of Alternative Test Methods in Toxicology. Academic Press. 287-300.

Dahl GE, Jaitly N, Salakhutdinov R, 2014. Multi-task Neural Networks for QSAR Predictions. arXiv. 1406.1231 [Preprint]. Available from: https://doi.org/10.48550/arXiv.1406.1231

Danishuddin, Khan AU, 2016. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug Discov. Today. 21: 1291-302.

Darnag R, Minaoui B, Fakir M, 2017. QSAR models for prediction study of HIV protease inhibitors using support vector machines, neural networks and multiple linear regression. Arab. J. Chem. 10: S600-S608.

Date MS, O'Brien D, Botelho DJ, Schultz TW, Liebler DC, Penning TM, Salvito DT, 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 P, Kar S, Ambure P, Roy K, 2022. Prediction reliability of QSAR models: an overview of various validation tools. Arch. Toxicol. 96: 1279-1295.

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.

Dexter GP, Grannis SJ, Dixon BE, Kasthurirathne SN, 2020. Generalization of Machine Learning Approaches to Identify Notifiable Conditions from a Statewide Health Information Exchange. AMIA Jt. Summits Transl. Sci. Proc. v.2020: 152-161.

Dietterich T, 1995. Overfitting and Undercomputing in Machine Learning. ACM Comput. Surv. 27: 326-327.

Ding Y, Chen M, Guo C, Zhang P, Wang J, 2021. Molecular fingerprint-based machine learning assisted QSAR model development for prediction of ionic liquid properties. J. Mol. Liq. 326: 115212.

Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, et al., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography. 36: 27-46.

Drakvik E, Altenburger R, Aoki Y, et al., 2020. Statement on advancing the assessment of chemical mixtures and their risks for human health and the environment. Environ. Int. 134: 105267. doi:10.1016/j.envint.2019.105267

Du M, Liu N, Hu X, 2019. Techniques for Interpretable Machine Learning. arXiv:1808.00033 [Preprint]. Available from: https://doi.org/10.48550/arXiv.1808.00033

Duchowicz PR, Vitale MG, Castro EA, 2008. Partial Order Ranking for the aqueous toxicity of aromatic mixtures. J. Math. Chem. 44: 541-549. doi:10.1007/s10910-007-9327-6

EC JRC (European Commission, Joint Research Centre), 2020. JRC QSAR Model Database. European Commission, Joint Research Centre (JRC) [Dataset] PID: http://data.europa.eu/89h/e4ef8d13-d743-4524-a6eb-80e18b58cba4

European Chemical Agency (ECHA), 2017a. 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

European Chemical Agency (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.

European Chemicals Agency (ECHA), 2008. Guidance on information requirements and chemical safety assessment Chapter R.6: QSARs and grouping of chemicals. Available at: https://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf/77f49f81-b76d-40ab-8513-4f3a533b6ac9

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

Ekins S, Mestres J, Testa B, 2007. *In silico* pharmacology for drug discovery: methods for virtual ligand screening and profiling. Br. J. Pharmacol. 152: 9-20. https://doi.org/10.1038/sj.bjp.0707305

Enoch SJ, Cronin MT, Schultz TW, Madden JC, 2008. An evaluation of global QSAR models for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*. Chemosphere. 71: 1225-1232.

Erhirhie EO, Ihekwereme CP, Ilodigwe EE, 2018. Advances in acute toxicity testing: strengths, weaknesses and regulatory acceptance. Interdiscip. Toxicol. 11: 5-12. doi:10.2478/intox-2018-0001

Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P, 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ. Health Perspect. 111: 1361-75.

European Chemical Agency (ECHA), 2017b. Guidance for identification and naming of substances under REACH and CLP.

European Commission, 2009b. Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. *Off. J. Eur. Union* L342: 59-209.

European Commission, 2009a. State of the Art Report on Mixture Toxicity – Final Report, Executive Summary.

European Commission, 2012a. Toxicity and Assessment of Chemical Mixtures. doi:10.2772/21444

European Commission, 2012b. The Combination Effects of Chemicals - Chemical mixtures.

European Commission, 2020. Chemicals Strategy for Sustainability - Towards a Toxic-free Environment.

European Parliament and Council, 2010. Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes. Official Journal of the European Union. L 276/33.

Eskes C, Whelan M, 2016. Validation of Alternative Methods for Toxicity Testing. Springer International Publishing. 856: 1-9.

Fang S, Wang D, Zhang X, et al., 2016. Similarities and differences in combined toxicity of sulfonamides and other antibiotics towards bacteria for environmental risk assessment. Environ. Monit. Assess. 188: 429. doi:10.1007/s10661-016-5422-0

Fenton SE, Ducatman A, Boobis A, DeWitt JC, Lau C, Ng C, Smith JS, Roberts SM, 2021., Per- and Polyfluoroalkyl Substance Toxicity and Human Health Review: Current State of Knowledge and Strategies for Informing Future Research. Environ. Toxicol. Chem. 40: 606-630. https://doi.org/10.1002/etc.4890

Freeman JA, Skapura DM, 1991. Neural Networks Algorithms, Applications, and Programming Techniques. Addition Wesley Publishing Company, Reading.

Fulladosa E, Murat JC, Villaescusa I, 2005. Study on the toxicity of binary equitoxic mixtures of metals using the luminescent bacteria *Vibrio fischeri* as a biological target. Chemosphere. 58: 551-7. doi:10.1016/j.chemosphere.2004.08.007

Gadaleta D, Vuković K, Toma C, Lavado GJ, Karmaus AL, Mansouri K, et al., 2019. SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data. J. Cheminform. 11: 58.

Gao C, Sun H, Wang T, Tang M, Bohnen NI, Müller ML, et al., 2018. Model-based and Model-free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease. Sci. Rep. 8: 7129.

Gaskill SJ, Bruce ED, 2016. Binary Mixtures of Polycyclic Aromatic Hydrocarbons Display Nonadditive Mixture Interactions in an *In Vitro* Liver Cell Model. Risk Anal. 36: 968-91. doi:10.1111/risa.12475

Ghafourian T, Cronin MT, 2005. The impact of variable selection on the modelling of oestrogenicity. SAR QSAR Environ. Res. 16: 171-190.

Ghojogh B, Crowley M, 2019. The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial. arXiv:1905.12787 [Preprint]. Available from: https://doi.org/10.48550/arXiv.1905.12787

Gini G, Zanoli F, 2020. Machine Learning and Deep Learning Methods in Ecotoxicological QSAR Modeling. In: Roy K, editors. Ecotoxicological QSARs, Methods in Pharmacology and Toxicology. Humana, New York. pp. 111-149.

Girotti S, Ferri EN, Fumo MG, Maiolini E, 2008. Monitoring of environmental pollutants by bioluminescent bacteria. Analytica. Chimica. Acta. 608: 2-29. doi:10.1016/j.aca.2007.12.008

Glont M, Nguyen TVN, Graesslin M, Hälke R, Ali R, Schramm J, Wimalaratne SM, Kothamachu VB, Rodriguez N, Swat MJ, Eils J, Eils R, Laibe C, Malik-Sheriff RS, Chelliah V, Le Novère N, Hermjakob H, 2018. BioModels: expanding horizons to include more modelling approaches and formats. Nucl. Acids Res. 46: D1248–D1253.

Gramatica P, 2007. Principles of QSAR models validation: internal and external. QSAR Comb. Sci. 26: 694-701.

Guha R, 2008. On the interpretation and interpretability of quantitative structure-activity relationship models. J. Comput. Aided. Mol. Des. 22: 857-871.

Gundersen OE, 2020. The Reproducibility Crisis is Real. AI. Mag. 41: 103-106.

Gunturi SB, Archana K, Khandelwal A, Narayanan R, 2008. Prediction of hERG Potassium Channel Bloackage Using kNN-QSAR and Local Lazy Regression Methods. QSAR Comb. Sci. 27: 1305-1317.

Gupta VK, Rana PS, 2019. Toxicity prediction of small drug molecules of androgen receptor using multilevel ensemble model. J. Bioinf. Comput. Biol. 17: 1950033.

Hansch C, Maloney P, Fujita T, et al., 1962. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. Nature. 194: 178–180. https://doi.org/10.1038/194178b0

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.

Hartung T, 2009. Toxicology for the twenty-first century. Nature. 460: 208–212. https://doi.org/10.1038/460208a

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

Hassold E, Galert W, Schulze J, 2021. Options for an environmental risk assessment of intentional and unintentional chemical mixtures under REACH: the status and ways forward. Environ. Sci. Eur. 33. doi:10.1186/s12302-021-00565-0

Hawkins DM, 2004. The Problem of Overfitting. J. Chem. Inf. Comput. Sci. 44: 1-12.

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.

Heil BJ, Hoffman MM, Markowetz F, Lee S, Greene CS, Hicks SC, 2021. Reproducibility standards for machine learning in the life sciences. Nat. Methods. 18: 1132-1135.

Herbst AL, Ulfelder H, Poskanzer DC, 1971. Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women. N. Engl. J. Med. 284: 878-81. doi: 10.1056/NEJM197104222841604.

Hernández AF, Gil F, Lacasaña M, 2017. Toxicological interactions of pesticide mixtures: an update. Arch. Toxicol. 91: 3211-3223. doi:10.1007/s00204-017-2043-5

Hewitt M, Cronin MT, Rowe PH, Schultz TW, 2011. Repeatability analysis of the *Tetrahymena pyriformis* population growth impairment assay. SAR QSAR Environ. Res. 22: 621-637.

Hewitt M, Ellison CM, Cronin MTD, Pastor M, Steger-Hartmann T, Munoz-Muriendas J, Pognan F, Madden JC, 2015. Ensuring confidence in predictions: A scheme to assess the scientific validity of *in silico* models. Adv. Drug Delivery Rev. 86: 101-111.

Hochreiter S, Klambauer G, Rarey M, 2018. Machine Learning in Drug Discovery. J. Chem. Inf. Model. 58: 1723-1724.

Hodges G, Roberts DW, Marshall SJ, Dearden JC, 2006. Defining the toxic mode of action of ester sulphonates using the joint toxicity of mixtures. Chemosphere. 64: 17-25. doi:10.1016/j.chemosphere.2005.12.021

Hooker G, Mentch L, Zhou S, 2019. Unrestricted Permutation forces Extrapolation: Variable Importance Requires at least One More Model, or There Is No Free Variable Importance. arXiv:1905.03151 [Preprint]. Available from: https://doi.org/10.48550/arXiv.1905.03151

Hoover G, Kar S, Guffey S, Leszczynski J, Sepúlveda MS, 2019. *In vitro* and *in silico* modeling of perfluoroalkyl substances mixture toxicity in an amphibian fibroblast cell line. Chemosphere. 233: 25-33. doi:10.1016/j.chemosphere.2019.05.065

Huang H, Wang X, Shao Y, Chen D, Dai X, Wang L, 2003. QSAR for prediction of joint toxicity of substituted phenols to tadpoles (*Rana japonica*). Bull. Environ. Contam. Toxicol. 71: 1124-30. doi:10.1007/s00128-003-8790-4

Ibrahim IT, 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

Ivanciuc O, 2007. Applications of Support Vector Machines in Chemistry. In: Lipkowitz KB, Cundari TR, editors. Rev. Comput. Chem. 2007. pp. 291-400.

Jabbar HK, Khan RZ, 2014. Methods to Avoid Over-fitting and Under-fitting in Supervised Machine Learning (Comparative Study). In: Stephen J, Rohil H, Vasavi S, editors. Computer Science, Communication & Instrumentation Devices. 2014.

Jaganathan K, Tayara H, Chong KT, 2022. An Explainable Supervised Machine Learning Model for Predicting Respiratory Toxicity of Chemicals Using Optimal Molecular Descriptors. Pharmaceutics. 14: 832. doi: 10.3390/pharmaceutics14040832

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.

Jin H, Wang C, Shi J, Chen L, 2014. Evaluation on joint toxicity of chlorinated anilines and cadmium to *Photobacterium phosphoreum* and QSAR analysis. J. Hazard. Mater. 279: 156-62. doi:10.1016/j.jhazmat.2014.06.068

Johnson C, Ahlberg E, Anger LT, Beilke L, Benigni R, Bercu J, Bobst S, Bower D, Brigo A, Campbell S, Cronin MTD, Crooks I, Cross KP, Doktorova T, Exner T, Faulkner D, Fearon IM, Fehr M, Gad SC, 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 AT, Roberts DW, Schlecker H, Stidl R, Suarez-Rodrigez D, Szabo DT, Tice RR, Urbisch D, Vuorinen A, Wall B, Weiler T, White AT, Whritenour J, Wichard J, Woolley D, Zwickl C, Myatt GJ, 2020. Skin sensitization *in silico* protocol. Regul. Toxicol. Pharmacol. 116: e104688.

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

Kar S, Ghosh S, Leszczynski J, 2018. Single or mixture halogenated chemicals? Risk assessment and developmental toxicity prediction on zebrafish embryos based on weighted descriptors approach. Chemosphere. 210: 588-596. doi:10.1016/j.chemosphere.2018.07.051

Karelson M, Lobanov VS, Katritzky AR, 1996. Quantum-Chemical Descriptors in QSAR/QSPR Studies. Chem. Rev. 96: 1027-1044. doi:10.1021/cr950202r

Karmaus AL, Mansouri K, To KT, Blake B, Fitzpatrick J, Strickland J, et al., 2022. Evaluation of Variability Across Rat Acute Oral Systemic Toxicity Studies. Toxicol. Sci. 188: 34-47.

Khan PM, Roy K, 2018. Current approaches for choosing feature selection and learning algorithms in quantitative structure-activity relationships (QSAR). Expert Opin. Drug Discov. 13: 1075-1089.

Kienzler A, Bopp SK, van der Linden S, Berggren E, Worth A, 2016. Regulatory assessment of chemical mixtures: Requirements, current approaches and future perspectives. Regul. Toxicol. Pharmacol. 80: 321-334. doi:10.1016/j.yrtph.2016.05.020

Kim J, Kim S, 2015. State of the art in the application of QSAR techniques for predicting mixture toxicity in environmental risk assessment. SAR QSAR Environ. Res. 26: 41-59. doi:10.1080/1062936X.2014.984627

Kim J, Kim S, Schaumann GE, 2013a. Reliable predictive computational toxicology methods for mixture toxicity: toward the development of innovative integrated models for environmental risk assessment. Rev. Environ. Sci. Biotechnol. 12: 235-256. doi:10.1007/s11157-012-9286-7

Kim J, Kim S, Schaumann GE, 2013b. Development of QSAR-based two-stage prediction model for estimating mixture toxicity. SAR QSAR Environ. Res. 24: 841-861. doi:10.1080/1062936X.2013.815654

Kim K-W, Won YL, Hong MK, Jo J, Lee SK, 2014. Prediction of the Toxicity of Dimethyl formamide, Methyl Ethyl Ketone, and Toluene Mixtures by QSAR Modeling. Bull. Korean Chem. Soc. 35: 3637-3641.

Kingma DP, Ba J, 2014. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [Preprint]. Available from: https://doi.org/10.48550/arXiv.1412.6980

Knight DJ, Deluyker H, Chaudhry Q, Vidal J-M, Boer A, 2021. A call for action on the development and implementation of new methodologies for safety assessment of chemical-based products in the EU - A short communication. Regul. Toxicol. Pharmacol. 119: 104837.

Knight W, 2021. Sloppy use of machine learning is causing a ‘Reproducibility Crisis’ in science. *Wired* 10 August 2022, available at: https://www.wired.com/story/machine-learning-reproducibility-crisis/

Könemann H, 1981b. Fish toxicity tests with mixtures of more than two chemicals: A proposal for a quantitative approach and experimental results. Toxicology. 19: 229-238. doi:10.1016/0300-483X(81)90132-3

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

Krewski D, Andersen ME, Tyshenko MG, Krishnan K, Hartung T, Boekelheide K, et al., 2020. Toxicity testing in the 21st century: progress in the past decade and future perspectives. Arch. Toxicol. 94: 1-58.

Kulkarni SA, Benfenati E, Barton-Maclaren TS, 2016. Improving confidence in (Q)SAR predictions under Canada’s Chemicals Management Plan – a chemical space approach. SAR QSAR Environ. Res. 27: 851-863.

Kurup A, 2003. C-QSAR: a database of 18,000 QSARs and associated biological and physical data. J. Comput. Aided. Mol. Des. 17: 187-196.

Lambert JC, 2023. Adverse Outcome Pathway 'Footprinting': A Novel Approach to the Integration of 21st Century Toxicology Information into Chemical Mixtures Risk Assessment. Toxics. 11: 37. doi:10.3390/toxics11010037

Lapenna S, Fuart-Gatnik M, Worth A, 2011. Review of QSAR models and software tools for predicting acute and chronic systemic toxicity. doi:10.2788/60766

Laroche C, Annys E, Bender H, Botelho D, Botham P, Brendler-Schwaab S, Clayton R, Corvaro M, Dal Negro G, Delannois F, Dent M, Desaintes C, Desprez B, Dhalluin S, Hartmann A, Hoffmann-Doerr S, Hubesch B, Irizar A, Manou I, Müller BP, Nadzialek S, Prieto P, Rasenberg M, Roggeband R, Rowan TG, Schutte K, van de Water B, Westmoreland C, Whelan M, Wilshut A, Zvonimir Z, Cronin MTD, 2019. Finding synergies for the 3Rs - Repeated Dose Toxicity testing: Report from an EPAA Partners' Forum. Regul. Toxicol. Pharmacol. 108: 104470.

LeCun Y, Bengio Y, Hinton G, 2015. Deep learning. Nature. 521: 436-444.

Lin Z, Yu H, Wei D, Wang G, Feng J, Wang L, 2002. Prediction of mixture toxicity with its total hydrophobicity. Chemosphere. 46: 305-310. doi:10.1016/S0045-6535(01)00083-2

Lin Z, Zhong P, Yin K, Wang L, Yu H, 2003. Quantification of joint effect for hydrogen bond and development of QSARs for predicting mixture toxicity. Chemosphere. 52: 1199-208. doi:10.1016/s0045-6535(03)00329-1

Lo Y-C, Rensi SE, Torng W, Altman RB, 2018. Machine learning in chemoinformatics and drug discovery. Drug Discov. Today. 23: 1538-1546. doi:10.1016/j.drudis.2018.05.010

Loewe S, Muischnek H, 1926. Über Kombinationswirkungen. Naunyn-Schmiedebergs Archiv für experimentelle Pathologie und Pharmakologie. 114: 313-326. doi:10.1007/BF01952257

Long X, Wang D, Lin Z, Qin M, Song C, Liu Y, 2016. The mixture toxicity of environmental contaminants containing sulfonamides and other antibiotics in Escherichia coli: Differences in both the special target proteins of individual chemicals and their effective combined concentration. Chemosphere. 158: 193-203. doi:10.1016/j.chemosphere.2016.05.048

Lu GH, Wang C, Wang PF, Chen ZY, 2009. Joint toxicity evaluation and QSAR modeling of aromatic amines and phenols to bacteria. Bull. Environ. Contam. Toxicol. 83: 8-14. doi:10.1007/s00128-009-9694-8

Luan F, Xu X, Liu H, Cordeiro MNDS, 2013. Prediction of the baseline toxicity of non-polar narcotic chemical mixtures by QSAR approach. Chemosphere. 90: 1980-1986. doi:10.1016/j.chemosphere.2012.10.065

Luan X, Liu X, Fang C, Chu W, Xu Z, 2020. Ecotoxicological effects of disinfected wastewater effluents: a short review of *in vivo* toxicity bioassays on aquatic organisms. Environ. Sci. Water Res. 6: 2275-2286. doi:10.1039/D0EW00290A

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

Lundberg S, Lee S-I, 2017. A Unified Approach to Interpreting Model Predictions. arXiv. 1705.07874 [Preprint]. Available from: https://doi.org/10.48550/arXiv.1705.07874

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2: 56-67.

Madden JC, Enoch SJ, Paini A, Cronin MTD, 2020. A Review of *In Silico* Tools as Alternatives to Animal Testing: Principles, Resources and Applications. Altern. Lab. Anim. 48: 146-172. doi:10.1177/0261192920965977

Mahony C, Ashton RS, Birk B, Boobis AR, Cull T, Daston GP, Ewart L, Knudsen TB, Manou I, Maurer-Stroh S, Margiotta-Casaluci L, Muller BP, Nordlund P, Roberts RA, Steger-Hartmann T, Vandenbossche E, Viant MR, Vinken M, Whelan M, Zvonimir Z, Cronin MTD, 2020. New ideas for non-animal approaches to predict repeated-dose systemic toxicity: Report from an EPAA Blue Sky Workshop. Regul. Toxicol. Pharmacol. 114: 104668.

Malik-Sheriff RS, Glont M, Nguyen TVN, Tiwari K, Roberts MG, Xavier A, Vu MT, Men J, Maire M, Kananathan S, Fairbanks EL, Meyer JP, Arankalle C, Varusai TM, Knight-Schrijver V, Li L, Dueñas-Roca C, Dass G, Keating SM, Park YM, Buso N, Rodriguez N, Hucka M, Hermjakob H, 2020. BioModels — 15 years of sharing computational models in life science. Nucl. Acids Res. 48: D407–D415.

Mansouri K, Cariello N F, Korotcov A, Tkachenko V, Grulke C M, Sprankle CS, et al., 2019. Open-source QSAR models for pKa prediction using multiple machine learning approaches. J. Cheminform. 11: 60.

Martens M, Evelo CT, Willighagen EL, 2022. Providing Adverse Outcome Pathways from the AOP-Wiki in a Semantic Web Format to Increase Usability and Accessibility of the Content. Appl. Vitro Toxicol. 8: 2-13. doi: 10.1089/aivt.2021.0010.

Martens M, Stierum R, Schymanski EL, Evelo CT, Aalizadeh R, Aladjov H, Arturi K, Audouze K, Babica P, Berka K, Bessems J, Blaha L, Bolton EE, Cases M, Damalas DE, Dave K, Dilger M, Exner T, Geerke DP, Grafström R, Gray A, Hancock JM, Hollert H, Jeliazkova N, Jennen D, Jourdan F, Kahlem P, Klanova J, Kleinjans J, Kondic T, Kone B, Lynch I, Maran U, Martinez Cuesta S, Ménager H, Neumann S, Nymark P, Oberacher H, Ramirez N, Remy S, Rocca-Serra P, Salek RM, Sallach B, Sansone S-A, Sanz F, Sarimveis H, Sarntivijai S, Schulze T, Slobodnik J, Spjuth O, Tedds J, Thomaidis N, Weber RJM, van Westen GJP, Wheelock CE, Williams AJ, Witters H, Zdrazil B, Županič A, Willighagen EL, 2021. ELIXIR and Toxicology: a community in development. *F1000Research 2021* 10(ELIXIR): 1129. (https://doi.org/10.12688/f1000research.74502.1)

Martin O, Scholze M, Ermler S, McPhie J, Bopp SK, Kienzler A, Parissis N, Kortenkamp A, 2021. Ten years of research on synergisms and antagonisms in chemical mixtures: A systematic review and quantitative reappraisal of mixture studies. Environ. Int. 146: 106206. doi: 10.1016/j.envint.2020.106206

Mater AC, Coote ML, 2019. Deep Learning in Chemistry. J. Chem. Inf. Model. 59: 2545-2559.

Matveieva M, Polishchuk P, 2021. Benchmarks for interpretation of QSAR models. J. Cheminform. 13: 41.

Mayer G, Müller W, Schork K, Uszkoreit J, Weidemann A, Wittig U, Rey M, Quast C, Felden J, Glöckner FO, Lange M, Arend D, Beier S, Junker A, Scholz U, Schüler D, Kestler HA, Wibberg D, Pühler A, Twardziok S, Eils J, Eils R, Hoffmann S, Eisenacher M, Turewicz M, 2021. Implementing FAIR data management within the German Network for Bioinformatics Infrastructure (de.NBI) exemplified by selected use cases. Brief. Bioinform. 22: bbab010.

McDermott MB, Wang S, Marinsek N, Ranganath R, Ghassemi M, Foschini L, 2019. Reproducibility in Machine Learning for Health. arXiv:1907.01463 [Preprint]. Available from: https://doi.org/10.48550/arXiv.1907.01463

McKim JM, Bradbury SP, Niemi GJ, 1987. Fish acute toxicity syndromes and their use in the QSAR approach to hazard assessment. Environ. Health Perspect. 71: 171-186.

Merck, 2012. Kaggle Merck Molecular Activity Challenge. Available from: https://www.kaggle.com/c/MerckActivity

Molnar C, 2019. Interpretable machine learning. A Guide for Making Black Box Models Explainable. Available from: https://christophm.github.io/interpretable-ml-book/.

Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, et al., 2020. QSAR without borders. Chem. Soc. Rev. 49: 3525.

Muratov EN, Varlamova EV, Artemenko AG, Polishchuk PG, Kuz'min VE, 2012. Existing and Developing Approaches for QSAR Analysis of Mixtures. Mol. Inform. 31: 202-221. doi:10.1002/minf.201100129

Muratov EN, Varlamova EV, Artemenko AG, et al., 2013. QSAR analysis of poliovirus inhibition by dual combinations of antivirals. Struct. Chem. 24: 1665-1679. doi:10.1007/s11224-012-0195-8

Muratov EN, Varlamova EV, Kuzmin VE, et al., 2014. “Everything Out” Validation Approach for Qsar Models of Chemical Mixtures. J. Clin. Pharm. 1: 1005.

Mwense M, Wang XZ, Buontempo FV, Horan N, Young A, Osborn D, 2004. Prediction of Noninteractive Mixture Toxicity of Organic Compounds Based on a Fuzzy Set Method. J. Chem. Inf. Comput. Sci. 44: 1763-1773. doi:10.1021/ci0499368

Mwense M, Wang XZ, Buontempo FV, Horan N, Young A, Osborn D, 2006. QSAR approach for mixture toxicity prediction using independent latent descriptors and fuzzy membership functions. SAR QSAR Environ. Res. 17: 53-73. doi:10.1080/10659360600562202

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

Nelms MD, Simmons JE, Edwards SW, 2018. Adverse Outcome Pathways to Support the Assessment of Chemical Mixtures. In: Rider C, Simmons J, editors. Chemical Mixtures and Combined Chemical and Nonchemical Stressors. Springer Cham. pp 177-201.

O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR, 2011. Open Babel: An open chemical toolbox. J. Cheminform. 3: 33.

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.

Oprisiu I, Varlamova E, Muratov E, et al., 2012. QSPR Approach to Predict Nonadditive Properties of Mixtures. Application to Bubble Point Temperatures of Binary Mixtures of Liquids. Mol. Inform. 31: 491-502. doi:10.1002/minf.201200006

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

Palleria C, Di Paolo A, Giofrè C, et al., 2013. Pharmacokinetic drug-drug interaction and their implication in clinical management. J. Res. Med. Sci. 18: 601-10.

Pastor M, Gómez-Tamayo JC, Sanz F, 2021. Flame: an open source framework for model development, hosting, and usage in production environments. J. Cheminform. 13: 31.

Patel M, Chilton ML, Sartini A, Gibson L, Barber C, Covey-Crump L, Przybylak KR, Cronin MTD, Madden JC, 2018. Assessment and reproducibility of quantitative structure–activity relationship models by the nonexpert. J. Chem. Inform. Model. 58: 673-682.

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 EA, Whelan MP, 2017. A framework to establish credibility of computational models in biology. Prog. Biophys. Mol. Biol. 129: 13-19.

Patterson EA, Whelan MP, Worth AP, 2021. The role of validation in establishing the scientific credibility of predictive toxicology approaches intended for regulatory application. Comp. Toxicol. 17: e100144.

Pawar G, Madden JC, Ebbrell D, Firman JW, Cronin MTD, 2019. *In silico* toxicology data resources to support read-across and (Q)SAR. Front. Pharmacol. 10: 561.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12: 2825-2830.

Pestana CB, Firman JW, Cronin MTD, 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 AT, 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.

Pineau J, Vincent-Lamarre P, Sinha K, Larivière V, Beygelzimer A, d'Alché-Buc F, et al., 2020. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). arXiv:2003.12206 [Preprint]. Available from: https://doi.org/10.48550/arXiv.2003.12206

Pirhadi S, Shiri F, Ghasemi JB, 2015. Multivariate statistical analysis methods in QSAR. RSC Advances. 5: 104635-65.

Pognan F, Steger-Hartmann T, Díaz C, Blomberg N, Bringezu F, Briggs K, Callegaro G, Capella-Gutierrez S, Centeno E, Corvi J, Drew P, Drewe WC, Fernández JM, Furlong LI, Guney E, Kors JA, Mayer MA, Pastor M, Piñero J, Ramírez-Anguita JM, Ronzano F, Rowell P, Saüch-Pitarch J, Valencia A, van de Water B, van der Lei J, van Mulligen E, Sanz F, 2021. The eTRANSAFE Project on Translational Safety Assessment through Integrative Knowledge Management: Achievements and Perspectives. Pharmaceuticals. 14: 237.

Polishchuk PG, Muratov EN, Artemenko AG, Kolumbin OG, Muratov NN, Kuz'min VE, 2009. Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity. J. Chem. Inf. Model. 49: 2481-2488.

Preuer K, Lewis RPI, Hochreiter S, Bender A, Bulusu KC, Klambauer G, 2018. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. Bioinform. 34: 1538–1546. https://doi.org/10.1093/bioinformatics/btx806

Qin L-T, Chen Y-H, Zhang X, Mo L-Y, Zeng H-H, Liang Y-P, 2018. QSAR prediction of additive and non-additive mixture toxicities of antibiotics and pesticide. Chemosphere. 198: 122-129. doi:10.1016/j.chemosphere.2018.01.142

Qin S, Jiang S, Li J, Balaprakash P, Van Lehn RC, Zavala VM, 2023. Capturing molecular interactions in graph neural networks: a case study in multi-component phase equilibrium. Digit. Discov. 2: 138-151. doi:10.1039/D2DD00045H

Qu R, Wang X, Liu Z, Yan Z, Wang Z, 2013. Development of a model to predict the effect of water chemistry on the acute toxicity of cadmium to *Photobacterium phosphoreum*. J. Hazard. Mater. 262: 288-296. doi:10.1016/j.jhazmat.2013.08.039

Rabinowitz JR, Goldsmith MR, Little SB, Pasquinelli MA, 2008. Computational molecular modeling for evaluating the toxicity of environmental chemicals: prioritizing bioassay requirements. Environ. Health Perspect. 116: 573-7. doi:10.1289/ehp.11077

Ram RN, Gadaleta D, Allen THE, 2022. The role of 'big data' and '*in silico*' New Approach Methodologies (NAMs) in ending animal use – A commentary on progress. Comput. Toxicol. 23: 100232.

Ribeiro MT, Singh S, Guestrin C, 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery; 2016. pp. 1135-1144.

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.

Roberts DW, 1991. QSAR issues in aquatic toxicity of surfactants. Sci. Total Environ. 109-110: 557-568. doi:10.1016/0048-9697(91)90209-W

Robinson KG, Akins RE, 2021. Machine learning in epigenetic diseases. In: Tollefsbol TO, editors. Medical Epigenetics. 2021. pp. 513-525.

Rodea-Palomares I, González-Pleiter M, Martín-Betancor K, Rosal R, Fernández-Piñas F, 2015. Additivity and Interactions in Ecotoxicity of Pollutant Mixtures: Some Patterns, Conclusions, and Open Questions. Toxics. 3: 342-369. doi:10.3390/toxics3040342

Rodríguez-Pérez R, Bajorath J, 2020. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. J. Med. Chem. 63: 8761-8777.

Rose VS, Croall IF, Macfie HJ, 1991. An Application of Unsupervised Neural Network Methodology Kohonen Topology-Preserving Mapping to QSAR Analysis. Mol. Inform. 10: 6-15.

Roy K, Kar S, Das RN, 2015. Statistical Methods in QSAR/QSPR. In: Roy K, Kar S, Das RN, editors. A Primer on QSAR/QSPR Modeling: Fundamental Concepts. Cham. Springer International Publishing. p. 37-59.

Roy K, Kar S, 2015. Chapter 3 - How to Judge Predictive Quality of Classification and Regression Based QSAR Models? In: Ul-Haq Z, Madura JD, editors. Front. Comput. Chem.: Bentham Science Publishers. p. 71-120.

Ruiz P, Schilter B, Serafimova R, Simpson W, Stavitskaya L, Stidl R, Suarez-Rodriguez D, Szabo DT, Teasdale A, Trejo-Martin A, Valentin J-P, Vuorinen A, Wall BA, Watts P, White AT, Wichard J, Witt KL, Woolley A, Woolley D, Zwickl C, Hasselgren C, 2018. *In silico* toxicology protocols. Regul. Toxicol. Pharmacol. 96: 1-17.

Russell WMS, Burch RL, 1959. The Principles of Humane Experimental Technique. Methuen.

Russom CL, Bradbury SP, Broderius SJ, Hammermeister DE, Drummond RA, 1997. Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas). Environ. Toxicol. Chem. 16: 948-967.

Ruusmann V, Maran U, 2013. From data point timelines to a well curated data set, data mining of experimental data and chemical structure data from scientific articles, problems and possible solutions. J. Comput. Aided Mol. Des. 27: 583-603.

Ruusmann V, Sild S, Maran U, 2015. QSAR DataBank repository: open and linked qualitative and quantitative structure–activity relationship models. J. Cheminform. 7: 32.

Sahlin U, 2013. Uncertainty in QSAR predictions. Altern. Lab. Anim. 41: 111-125.

Scardapane S, Wang D, 2017. Randomness in neural networks: an overview. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 7: e1200.

Schultz TW, Cronin MT, Netzeva TI, Aptula AO, 2002. Structure-toxicity relationships for aliphatic chemicals evaluated with *Tetrahymena pyriformis*. Chem. Res. Toxicol. 15: 1602-1609.

Schultz TW, 1997. Tetratox: *Tetrahymena pyriformis* population growth impairment endpoint – a surrogate for fish lethality. Toxicol. Mech. Methods. 7: 289-309.

Schultz TW, Cronin MTD, 2017. Lessons learned from read-across case studies for repeated-dose toxicity. Regul. Toxicol. Pharmacol. 88: 185-191.

Schüürmann G, 2004. Quantum chemical descriptors in structure-activity relationships - calculation, interpretation, and comparison of methods. In: Cronin MTD, Livingstone DJ (eds) Predicting Chemical Toxicity and Fate. CRC Press, Boca Raton, Florida, USA, p 85-149

Sheridan RP, Wang WM, Liaw A, Ma J, Gifford EM, 2016. Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. J. Chem. Inf. Model. 56: 2353-2360.

Sidorov P, Naulaerts S, Ariey-Bonnet J, Pasquier E, Ballester PJ, 2019. Predicting Synergism of Cancer Drug Combinations Using NCI-ALMANAC Data. Front. Chem. 7. doi: 10.3389/fchem.2019.00509

Spjuth O, Willighagen EL, Guha R, Eklund M, Wikberg JES, 2010. Towards interoperable and reproducible QSAR analyses: Exchange of datasets. J. Cheminform. 2: 5.

Steinmetz FP, Mellor CL, Meinl T, Cronin MTD, 2015. Screening chemicals for receptor-mediated toxicological and pharmacological endpoints: using public data to build screening tools within a KNIME workflow. Mol. Inform. 34: 1710–178.

Su L, Zhang X, Yuan X, Zhao Y, Zhang D, Qin W, 2012. Evaluation of joint toxicity of nitroaromatic compounds and copper to *Photobacterium phosphoreum* and QSAR analysis. J. Hazard. Mater. 241-242: 450-455. doi:10.1016/j.jhazmat.2012.09.065

Su LM, Zhao YH, Yuan X, Mu CF, Wang N, Yan JC, 2010. Evaluation of combined toxicity of phenols and lead to *Photobacterium phosphoreum* and quantitative structure-activity relationships. Bull. Environ. Contam. Toxicol. 84: 311-4. doi:10.1007/s00128-009-9665-0

Sugimura P, Hartl F, 2018. Building a Reproducible Machine Learning Pipeline. arXiv:1810.04570 [Preprint]. Available from: https://doi.org/10.48550/arXiv.1810.04570

Sushko I, Novotarskyi S, Körner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin II, Palyulin VA, Radchenko EV, Welsh WJ, Kholodovych V, Chekmarev D, Cherkasov A, Aires-de-Sousa J, Zhang QY, Bender A, Nigsch F, Patiny L, Williams A, Tkachenko V, Tetko IV, 2011. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J. Comput. Aided Mol. Des. 25: 533-554.

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 RS, Bahadori T, Buckley TJ, Cowden J, Deisenroth C, Dionisio KL, Frithsen JB, Grulke CM, Gwinn MR, Harrill JA, Higuchi M, Houck KA, Hughes MF, Hunter ES, Isaacs KK, Judson RS, Knudsen TB, Lambert JC, Linnenbrink M, Martin TM, Newton SR, Padilla S, Patlewicz G, Paul-Friedman K, Phillips KA, Richard AM, Sams R, Shafer TJ, Setzer RW, Shah I, Simmons JE, Simmons SO, Singh A, Sobus JR, Strynar M, Swank A, Tornero-Valez R, Ulrich EM, Villeneuve DL, Wambaugh JF, Wetmore BA, Williams AJ, 2019. The next generation blueprint of computational toxicology at the U.S. Environmental Protection Agency. Toxicol. Sci. 169: 317–332.

Thompson CV, Firman JW, Goldsmith MR, Grulke CM, Tan Y-M, Paini A, Penson PE, Sayre RR, Webb S, Madden JC, 2021. A systematic review of published physiologically-based kinetic models and an assessment of their chemical space coverage. Altern. Lab. Anim. 49: 197-208.

Thoreau F, 2016. ‘A mechanistic interpretation, if possible’: How does predictive modelling causality affect the regulation of chemicals?. Big Data Soc. 3: 1-11.

Tiwari K, Kananathan S, Roberts MG, Meyer JP, Shohan MUS, Xavier A, Maire M, Zyoud A, Men J, Ng S, Nguyen TVN, Glont M, Hermjakob H, Malik-Sheriff RS, 2021. Reproducibility in systems biology modelling. Mol. Syst. Biol. 17: e9982.

Toropova AP, Toropov AA, Benfenati E, Gini G, Leszczynska D, Leszczynski J, 2012. CORAL: Models of toxicity of binary mixtures. Chemometr. Intell. Lab. Syst. 119: 39-43. doi:10.1016/j.chemolab.2012.10.001

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

United States Environmental Protection Agency (EPA), 2018. Strategic Plan to Promote the Development and Implementation of Alternative Test Methods Within the TSCA Program. EPA Document# EPA-740-R1-8004. Available at: https://www.epa.gov/sites/default/files/2018-06/documents/epa_alt_strat_plan_6-20-18_clean_final.pdf

United States National Research Council (US NRC), 2007. Applications of toxicogenomic technologies to predictive toxicology and risk assessment. US National Research Council, Washington (DC)

Vakharia V, Gujar R, 2019. Prediction of compressive strength and Portland cement composition using cross-validation and feature ranking techniques. Constr. Build. Mater. 225: 292-301.

van der Zalm AJ, Barroso J, Browne P, Casey W, Gordon J, Henry TR, Kleinstreuer NC, Lowit AB, Perron M, Clippinger AJ, 2022. A framework for establishing scientific confidence in new approach methodologies. Arch. Toxicol. 96: 2865-2879. doi: 10.1007/s00204-022-03365-4.

Varnek A, Baskin I, 2012. Machine Learning Methods for Property Prediction in Chemoinformatics: *Quo Vadis*?. J. Chem. Inf. Model. 52: 1413-1437.

Verhaar HJM, Busser FJM, Hermens JLM, 1995. Surrogate Parameter for the Baseline Toxicity Content of Contaminated Water: Simulating the Bioconcentration of Mixtures of Pollutants and Counting Molecules. Environ. Sci. Technol. 29: 726-734. doi:10.1021/es00003a021

Vesteghem C, Brøndum RF, Sønderkær M, Sommer M, Schmitz A, Bødker JS, Dybkær K, El-Galaly TC, Bøgsted M, 2020. Implementing the FAIR Data Principles in precision oncology: review of supporting initiatives. Brief. Bioinform. 21: 936-945.

Walker JD, Carlsen L, Jaworska J, 2003. Improving Opportunities for Regulatory Acceptance of QSARs: The Importance of Model Domain, Uncertainty, Validity and Predictability. QSAR Comb. Sci. 22: 346-50.

Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, ELIXIR Machine Learning Focus Group, Harrow J, Psomopoulos FE, Tosatto SCE, 2021. DOME: recommendations for supervised machine learning validation in biology. Nat. Methods. 18: 1122–1127.

Wang B, Yu G, Zhang Z, Hu H, Wang L, 2006. Quantitative structure-activity relationship and prediction of mixture toxicity of alkanols. Chin. Sci. Bull. 51: 2717-2723. doi:10.1007/s11434-006-2168-z

Wang C, Lu G, Tang Z, Guo X, 2008. Quantitative structure-activity relationships for joint toxicity of substituted phenols and anilines to Scenedesmus obliquus. J. Environ. Sci. (China). 20: 115-9. doi:10.1016/s1001-0742(08)60018-2

Wang D, Gu Y, Zheng M, Zhang W, Lin Z, Liu Y, 2017. A Mechanism-based QSTR Model for Acute to Chronic Toxicity Extrapolation: A Case Study of Antibiotics on Luminous Bacteria. Sci. Rep. 7: 6022. doi:10.1038/s41598-017-06384-9

Wang D, Shi J, Xiong Y, et al., 2018c. A QSAR-based mechanistic study on the combined toxicity of antibiotics and quorum sensing inhibitors against Escherichia coli. J. Hazard. Mater. 341: 438-447. doi:10.1016/j.jhazmat.2017.07.059

Wang D, Wu X, Lin Z, Ding Y, 2018b. A comparative study on the binary and ternary mixture toxicity of antibiotics towards three bacteria based on QSAR investigation. Environ. Res. 162: 127-134. doi:10.1016/j.envres.2017.12.015

Wang H, Li Y, Huang H, Xu X, Wang Y, 2011b. Toxicity evaluation of single and mixed antifouling biocides using the Strongylocentrotus intermedius sea urchin embryo test. Environ. Toxicol. Chem. 30: 692-703. doi:10.1002/etc.440

Wang H, Liu W, Chen J, 2023. Chapter 11 - QSAR modelling based on graph neural networks. In: Hong H editor. QSAR in Safety Evaluation and Risk Assessment. Academic Press. p. 139-151.

Wang S, Yan LC, Zheng SS, Li TT, Fan LY, Huang T, et al., 2019b. Toxicity of some prevalent organic chemicals to tadpoles and comparison with toxicity to fish based on mode of toxic action. Ecotoxicol. Environ. Saf. 167: 138-145.

Wang T, Lin Z, Yin D, Tian D, Zhang Y, Kong D, 2011a. Hydrophobicity-dependent QSARs to predict the toxicity of perfluorinated carboxylic acids and their mixtures. Environ. Toxicol. Pharmacol. 32: 259-65. doi:10.1016/j.etap.2011.05.011

Wang T, Tang L, Luan F, Cordeiro M, 2018a. Prediction of the Toxicity of Binary Mixtures by QSAR Approach Using the Hypothetical Descriptors. Int. J. Mol. Sci. 19. doi:10.3390/ijms19113423

Wang T, Zhou X, Wang D, Yin D, Lin Z, 2012. Using molecular docking between organic chemicals and lipid membrane to revise the well known octanol-water partition coefficient of the mixture. Environ. Toxicol. Pharmacol. 34: 59-66. doi:10.1016/j.etap.2012.02.008

Wang L, Xing P, Wang C, Zhou X, Dai Z, Bai L, 2019a. 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.

Warne MS, 2003. A Review of the Ecotoxicity of Mixtures, Approaches to, and Recommendations for, their Management. In: A L, M G, B K (eds) Proceedings of the Fifth National Workshop on the Assessment of Site Contamination, Adelaide, Australia, 2003.

Warne MS, Hawker DW, 1995. The number of components in a mixture determines whether synergistic and antagonistic or additive toxicity predominate: the funnel hypothesis. Ecotoxicol. Environ. Saf. 31: 23-28. doi:10.1006/eesa.1995.1039

Wei DB, Zhai LH, Hu HY, 2004. QSAR-based toxicity classification and prediction for single and mixed aromatic compounds. SAR QSAR Environ. Res. 15: 207-216. doi:10.1080/10629360410001697762

Westmoreland C, Bender HJ, Doe JE, Jacobs MN, Kass GEN, Madia F, et al., 2022. Use of New Approach Methodologies (NAMs) in regulatory decisions for chemical safety: Report from an EPAA Deep Dive Workshop. Regul. Toxicol. Pharmacol. 135: 105261.

Wilkinson MD, Dumontier M, Aalbersberg I, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, ’t Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B, 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data. 3: 160018.

Winkler DA, Le TC, 2017. Performance of Deep and Shallow Neural Networks, the Universal Approximation Theorem, Activity Cliffs, and QSAR. Mol. Inform. 37: 1600118.

Wise J, de Barron AG, Splendiani A, Balali-Mood B, Vasant D, Little E, Mellino G, Harrow I, Smith I, Taubert J, van Bochove K, Romacker M, Walgemoed P, Jimenez RC, Winnenburg R, Plasterer T, Gupta V, Hedley V, 2019. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discov. *Today.* 24: 933-938.

Wittwehr C, Blomstedt P, Gosling JP, 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,

Wojtuch A, Jankowski R, Podlewska S, 2021. How can SHAP values help to shape metabolic stability of chemical compounds?. J. Cheminform. 13: 74.

Wold S, Dunn WJ, 1982. Multivariate Quantitative Structure-Activity Relationships (QSAR): Conditions for Their Applicability. J. Chem. Inf. Comput. Sci. 23: 6-13.

World Health Organization (WHO), 2017. Chemical mixtures in source water and drinking-water.

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

Worth AP, 2020. Computational modelling for the sustainable management of chemicals. Comput. Toxicol. 14: 100122.

Wu L, Huang R, Tetko IV, Xia Z, Xu J, Tong W, 2021. Trade-off Predictivity and Explainability for Machine-Learning Powered Predictive Toxicology: An in-Depth Investigation with Tox21 Data Sets. Chem. Res. Toxicol. 34: 541-549.

Xia LY, Wang QY, Cao Z, Liang Y, 2019. Descriptor Selection Improvements for Quantitative Structure-Activity Relationships. Int. J. Neural Syst. 29: 1950016.

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.

Yang C, Cronin MTD, Arvidson KB, Bienfait B, Enoch SJ, Heldreth B, Hobocienski B, Muldoon-Jacobs K, Lan Y, Madden JC, Magdziarz T, Marusczyk J, Mostrag A, Nelms M, Neagu D, Przybylak K, Rathman JF, Park J, Richarz A-N, Richard AM, Ribeiro JV, Sacher O, Schwab C, Vitcheva V, Volarath P, Worth AP, 2021. COSMOS Database and Next Generation: A database and knowledge hub to leverage biological data from public resources in collaboration with regulatory offices for cosmetics and food ingredients. Comput. Toxicol. 19: e100175.

Yang RS, Thomas RS, Gustafson DL, et al., 1998. Approaches to developing alternative and predictive toxicology based on PBPK/PD and QSAR modeling. Environ. Health Perspect. 106 Suppl 6: 1385-93. doi:10.1289/ehp.98106s61385

Yap CW, 2011. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32: 1466-1474.

Ying X, 2019. An Overview of Overfitting and its Solutions. J. Phys. Conf. Ser. 1168: 022022.

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

Yu HX, Lin ZF, Feng JF, Xu TL, Wang LS, 2001. Development of quantitative structure activity relationships in toxicity prediction of complex mixtures. Acta. Pharmacol. Sin. 22: 45-9.

Yuan X, Lu G, Zhao J, 2002. QSAR study on the joint toxicity of 2,4-dinitrotoluene with aromatic compounds to *Vibrio fischeri*. J. Environ. Sci. Health. A. Tox. Hazard. Subst. Environ. Eng. 37: 573-8. doi:10.1081/ese-120003238

Zeng M, Lin Z, Yin D, Yin K, 2008. QSAR for predicting joint toxicity of halogenated benzenes to *Dicrateria zhanjiangensis*. Bull. Environ. Contam. Toxicol. 81: 525-30. doi:10.1007/s00128-008-9570-y

Zhang L, Zhou PJ, Yang F, Wang ZD, 2007. Computer-based QSARs for predicting mixture toxicity of benzene and its derivatives. Chemosphere. 67: 396-401. doi:10.1016/j.chemosphere.2006.09.018

Zhang S, Su L, Zhang X, et al., 2019. Combined Toxicity of Nitro-Substituted Benzenes and Zinc to *Photobacterium Phosphoreum*: Evaluation and QSAR Analysis. Int. J. Environ. Res. Public Health. 16. doi:10.3390/ijerph16061041

Zhang Y, Yang Y, 2015. Cross-validation for selecting a model selection procedure. J. Econom. 187: 95-112.

Zhao YH, Zhang XJ, Wen Y, Sun FT, Guo Z, Qin WC, et al, 2010. Toxicity of organic chemicals to *Tetrahymena pyriformis*: Effect of polarity and ionization on toxicity. Chemosphere.79: 72-77.

Zheng J, Harris MR, Masci AM, Lin Y, Hero A, Smith B, He Y, 2016. The Ontology of Biological and Clinical Statistics (OBCS) for standardized and reproducible statistical analysis. J. Biomed. Semantics. 14: 53.

Zheng W, Tropsha A, 2000. Novel Variable Selection Quantitative Structure-Property Relationship Approach Based on the *k*-Nearest Neighbor Principle. J. Chem. Inf. Comput. Sci. 40: 185-194.

Zhong S, Zhang K, Wang D, Zhang H, 2021. Shedding light on “Black Box” machine learning models for predicting the reactivity of HO radicals toward organic compounds. J. Chem. Eng. 405: 126627. doi: https://doi.org/10.1016/j.cej.2020.126627

Zhu T, Jiang Y, Cheng H, Singh RP, Yan B, 2020. Development of pp-LFER and QSPR models for predicting the diffusion coefficients of hydrophobic organic compounds in LDPE. Ecotoxicol. Environ. Saf. 190: 110179.

Zou J, Han Y, So SS, 2008. Overview of Artificial Neural Networks. In: Livingstone DJ, editors. Artificial Neural Networks. Humana Press. pp. 12-22.

Zou X, Lin Z, Deng Z, Yin D, Zhang Y, 2012. The joint effects of sulfonamides and their potentiator on *Photobacterium phosphoreum*: differences between the acute and chronic mixture toxicity mechanisms. Chemosphere. 86: 30-5. doi:10.1016/j.chemosphere.2011.08.046

Zou X, Zhou X, Lin Z, Deng Z, Yin D, 2013. A docking-based receptor library of antibiotics and its novel application in predicting chronic mixture toxicity for environmental risk assessment. Environ. Monit. Assess. 185: 4513-4527. doi:10.1007/s10661-012-2885-5

Zucco F, De Angelis I, Stammati A, 1998. Cellular Models for *In Vitro* Toxicity Testing. In: Clynes M (ed) Animal Cell Culture Techniques. Springer Berlin Heidelberg, Berlin, Heidelberg, p 395-422

Appendices

Appendix I. Supplementary material associated with Chapter 2

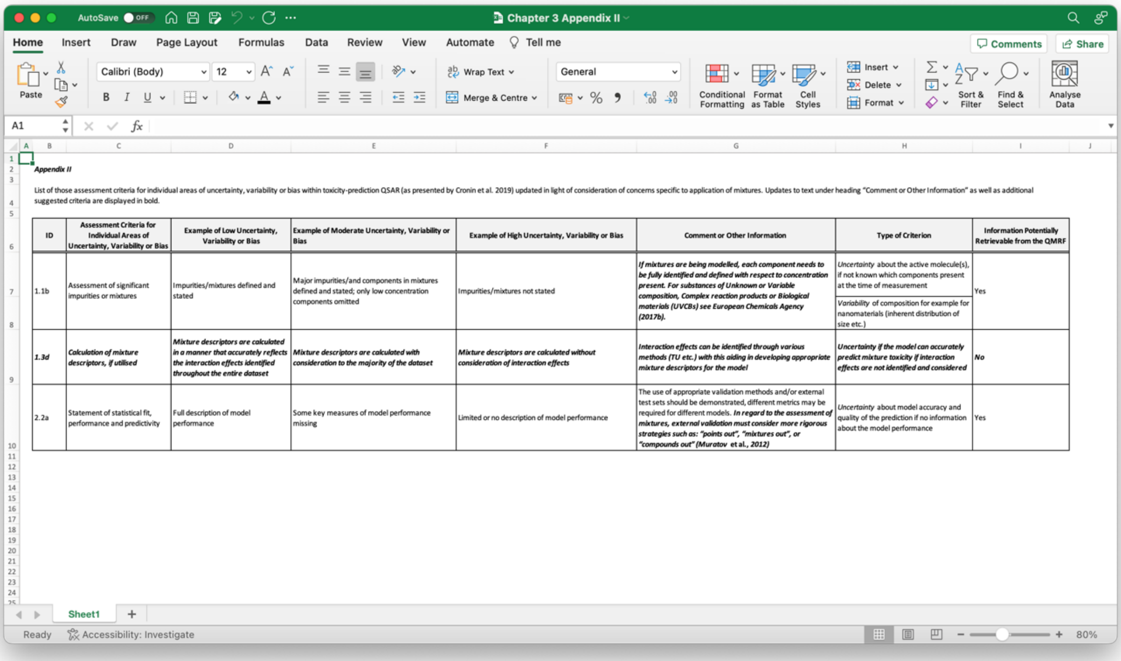
This includes the criteria associated with each component, full evaluations of the 12 QSAR studies against the uncertainty assessment criteria, and proposed mitigation strategies. Tables S1, S2, and S4 can be accessed from the following link: https://github.com/SamBelfield/PhD_Thesis

Table S3. Summary of uncertainty for each of the QSARs considered according to the QSAR components: yellow low uncertainty; green moderate uncertainty; blue high uncertainty.

	Creation			Characteristics				Application		
	Data	**Structures**	**Descriptors**	**Modelling**	**Performance**	**Mechanisms**	**Toxicokinetics**	**Description**	**Usability**	**Relevance**
Luan et al	🟦	🟨	🟩	🟦	🟩	🟦	🟦	🟨	🟨	🟦
Pal et al	🟦	🟨	🟩	🟩	🟩	🟦	🟦	🟨	🟨	🟩
de Morais e Silva et al	🟦	🟨	🟩	🟩	🟩	🟨	🟨	🟨	🟨	🟨
Toropova and Toropov	🟦	🟨	🟩	🟩	🟩	🟦	🟦	🟨	🟩	🟨
Wang et al	🟦	🟦	🟩	🟩	🟩	🟦	🟦	🟨	🟩	🟦
Yang et al	🟦	🟦	🟩	🟩	🟩	🟦	🟦	🟩	🟨	🟩
He et al	🟦	🟩	🟩	🟨	🟩	🟦	🟦	🟦	🟩	🟦
Jiang et al	🟦	🟩	🟩	🟨	🟩	🟦	🟦	🟦	🟩	🟦
Gupta and Rana	🟦	🟦	🟩	🟨	🟩	🟦	🟦	🟦	🟩	🟨
Ibrahim et al	🟦	🟨	🟨	🟩	🟩	🟦	🟦	🟦	🟨	🟨
Hao et al	🟨	🟨	🟩	🟩	🟩	🟩	🟦	🟨	🟨	🟩
Ahmadi	🟦	🟦	🟩	🟨	🟩	🟦	🟦	🟨	🟩	🟦

Appendix II. Unabridged scheme for QSAR mixture uncertainty assessment

This is a snapshot of the suggested supplementary information for the assessment of QSAR mixtures from Chapter 3. The full document can be accessed from the following link: https://github.com/SamBelfield/PhD_Thesis



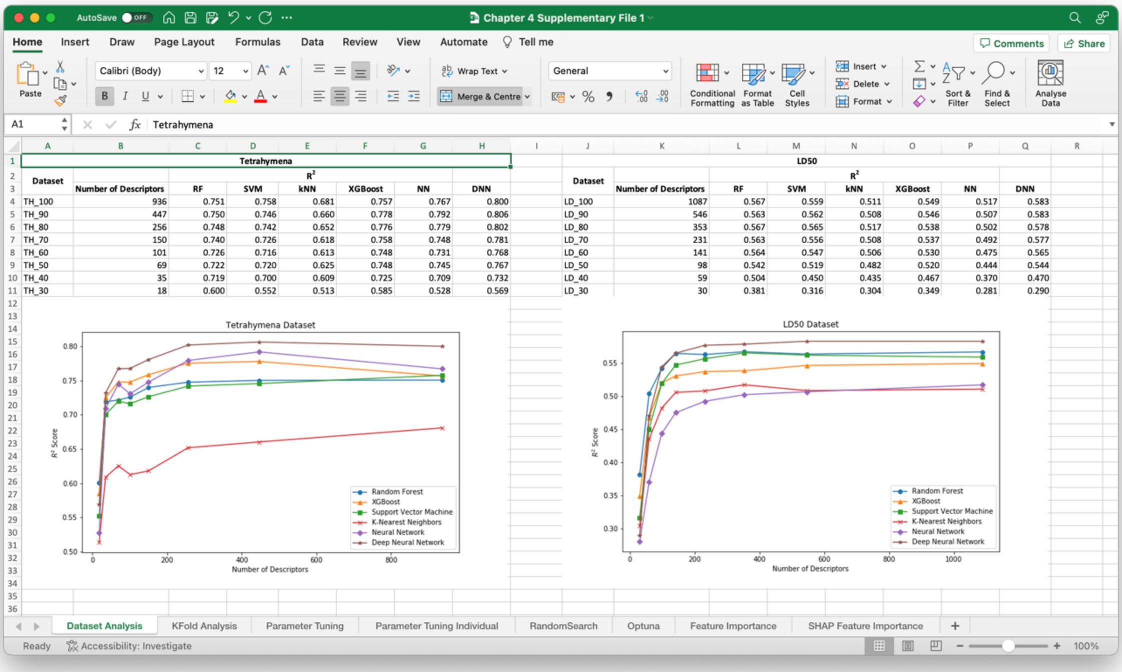
Appendix II

List of those assessment criteria for individual areas of uncertainty, variability or bias within toxicity-prediction QSAR (as presented by Cronin et al. 2019) updated in light of consideration of concerns specific to application of mixtures. Updates to text under heading "Comment or Other Information" as well as additional suggested criteria are displayed in bold.

ID	Assessment Criteria for Individual Areas of Uncertainty, Variability or Bias	Example of Low Uncertainty, Variability or Bias	Example of Moderate Uncertainty, Variability or Bias	Example of High Uncertainty, Variability or Bias	Comment or Other Information	Type of Criterion	Information Potentially Retrievable from the QMRF
1.1b	Assessment of significant impurities or mixtures	Impurities/mixtures defined and stated	Major impurities/and components in mixtures defined and stated; only low concentration components omitted	Impurities/mixtures not stated	***If mixtures are being modelled, each component needs to be fully identified and defined with respect to concentration present. For substances of Unknown or Variable composition, Complex reaction products or Biological materials (UVCBs) see European Chemicals Agency (2017b).***	*Uncertainty* about the active molecule(s), if not known which components present at the time of measurement *Variability* of composition for example for nanomaterials (inherent distribution of size etc.)	Yes
1.3d	***Calculation of mixture descriptors, if utilised***	***Mixture descriptors are calculated in a manner that accurately reflects the interaction effects identified throughout the entire dataset***	***Mixture descriptors are calculated with consideration to the majority of the dataset***	***Mixture descriptors are calculated without consideration of interaction effects***	***Interaction effects can be identified through various methods (TU etc.) with this aiding in developing appropriate mixture descriptors for the model***	***Uncertainty if the model can accurately predict mixture toxicity if interaction effects are not identified and considered***	***No***
2.2a	Statement of statistical fit, performance and predictivity	Full description of model performance	Some key measures of model performance missing	Limited or no description of model performance	The use of appropriate validation methods and/or external test sets should be demonstrated, different metrics may be required for different models. ***In regard to the assessment of mixtures, external validation must consider more rigorous strategies such as: "points out", "mixtures out", or "compounds out" (Muratov et al., 2012)***	*Uncertainty* about model accuracy and quality of the prediction if no information about the model performance	Yes

Appendix III. ML extended results

This is a snapshot of the results from the initial data analysis, hyperparameter optimisation, and feature importance of the six ML methods employed in Chapter 4. The full document can be accessed from the following link: https://github.com/SamBelfield/PhD_Thesis



Tetrahymena

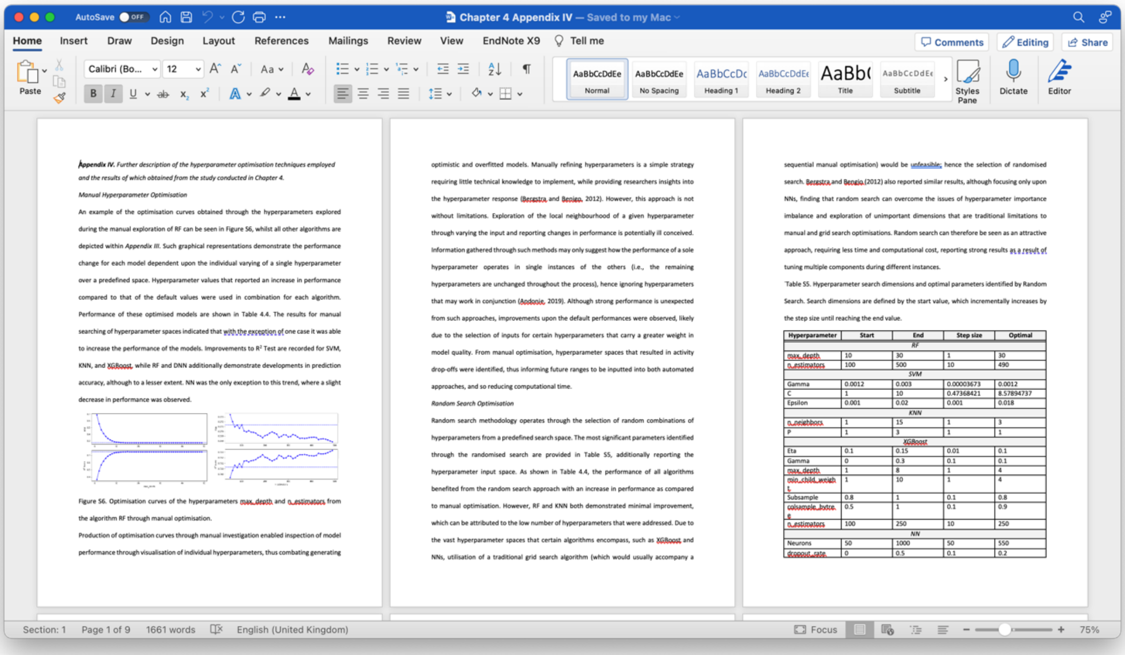
Dataset	Number of Descriptors	R^2					
		RF	SVM	kNN	XGBoost	NN	DNN
TH_100	936	0.751	0.758	0.681	0.757	0.767	0.800
TH_90	447	0.750	0.746	0.660	0.778	0.792	0.806
TH_80	256	0.748	0.742	0.652	0.776	0.779	0.802
TH_70	150	0.740	0.726	0.618	0.758	0.748	0.781
TH_60	101	0.726	0.716	0.613	0.748	0.731	0.768
TH_50	69	0.722	0.720	0.625	0.748	0.745	0.767
TH_40	35	0.719	0.700	0.609	0.725	0.709	0.732
TH_30	18	0.600	0.552	0.513	0.585	0.528	0.569

LD50

Dataset	Number of Descriptors	R^2					
		RF	SVM	kNN	XGBoost	NN	DNN
LD_100	1087	0.567	0.559	0.511	0.549	0.517	0.583
LD_90	546	0.563	0.562	0.508	0.546	0.507	0.583
LD_80	353	0.567	0.565	0.517	0.538	0.502	0.578
LD_70	231	0.563	0.556	0.508	0.537	0.492	0.577
LD_60	141	0.564	0.547	0.506	0.530	0.475	0.565
LD_50	98	0.542	0.519	0.482	0.520	0.444	0.544
LD_40	59	0.504	0.450	0.435	0.467	0.370	0.470
LD_30	30	0.381	0.316	0.304	0.349	0.281	0.290

Appendix IV. ML hyperparameter optimisation extended results

This is a snapshot of the further description of the hyperparameter optimisation techniques employed and the results of which obtained from the study conducted in Chapter 4. The full document can be accessed from the following link: https://github.com/SamBelfield/PhD_Thesis



Appendix IV. *Further description of the hyperparameter optimisation techniques employed and the results of which obtained from the study conducted in Chapter 4.*

Manual Hyperparameter Optimisation

An example of the optimisation curves obtained through the hyperparameters explored during the manual exploration of RF can be seen in Figure S6, whilst all other algorithms are depicted within *Appendix III*. Such graphical representations demonstrate the performance change for each model dependent upon the individual varying of a single hyperparameter over a predefined space. Hyperparameter values that reported an increase in performance compared to that of the default values were used in combination for each algorithm. Performance of these optimised models are shown in Table 4.4. The results for manual searching of hyperparameter spaces indicated that with the exception of one case it was able to increase the performance of the models. Improvements to R^2 Test are recorded for SVM, KNN, and XGBoost, while RF and DNN additionally demonstrate developments in prediction accuracy, although to a lesser extent. NN was the only exception to this trend, where a slight decrease in performance was observed.

Figure S6. Optimisation curves of the hyperparameters max_depth and n_estimators from the algorithm RF through manual optimisation.

Production of optimisation curves through manual investigation enabled inspection of model performance through visualisation of individual hyperparameters, thus combating generating optimistic and overfitted models. Manually refining hyperparameters is a simple strategy requiring little technical knowledge to implement, while providing researchers insights into the hyperparameter response (Bergstra and Benigo, 2012). However, this approach is not without limitations. Exploration of the local neighbourhood of a given hyperparameter through varying the input and reporting changes in performance is potentially ill conceived. Information gathered through such methods may only suggest how the performance of a sole hyperparameter operates in single instances of the others (i.e., the remaining hyperparameters are unchanged throughout the process), hence ignoring hyperparameters that may work in conjunction (Andonie, 2019). Although strong performance is unexpected from such approaches, improvements upon the default performances were observed, likely due to the selection of inputs for certain hyperparameters that carry a greater weight in model quality. From manual optimisation, hyperparameter spaces that resulted in activity drop-offs were identified, thus informing future ranges to be inputted into both automated approaches, and so reducing computational time.

Random Search Optimisation

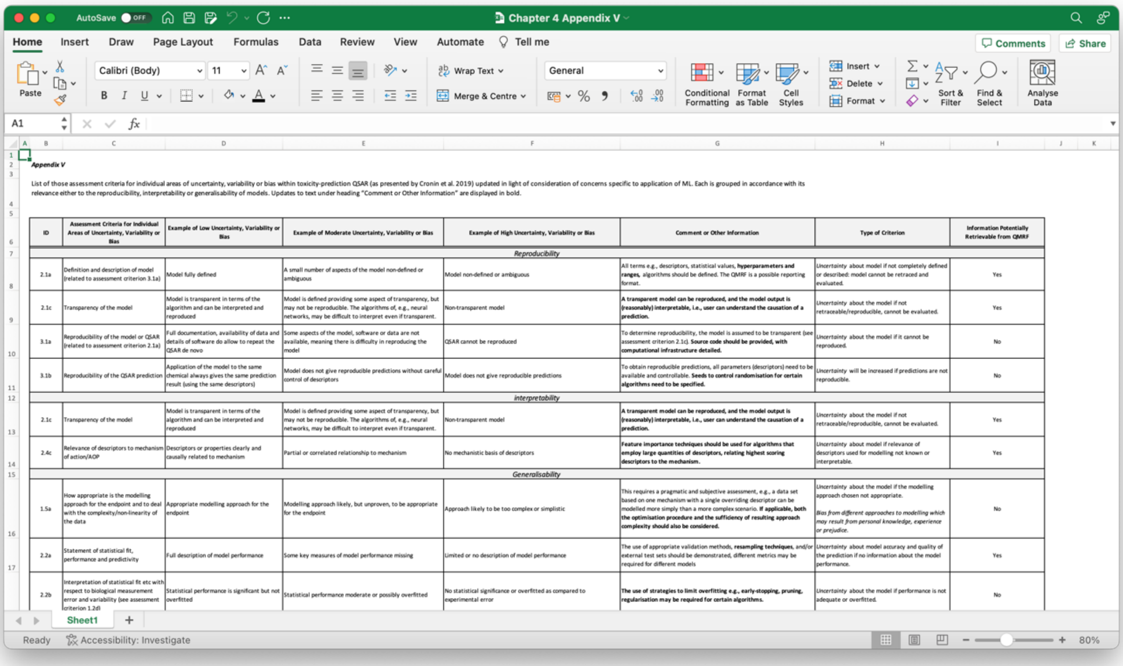
Random search methodology operates through the selection of random combinations of hyperparameters from a predefined search space. The most significant parameters identified through the randomised search are provided in Table S5, additionally reporting the hyperparameter input space. As shown in Table 4.4, the performance of all algorithms benefited from the random search approach with an increase in performance as compared to manual optimisation. However, RF and KNN both demonstrated minimal improvement, which can be attributed to the low number of hyperparameters that were addressed. Due to the vast hyperparameter spaces that certain algorithms encompass, such as XGBoost and NNs, utilisation of a traditional grid search algorithm (which would usually accompany a sequential manual optimisation) would be unfeasible; hence the selection of randomised search. Bergstra and Bengio (2012) also reported similar results, although focusing only upon NNs, finding that random search can overcome the issues of hyperparameter importance imbalance and exploration of unimportant dimensions that are traditional limitations to manual and grid search optimisations. Random search can therefore be seen as an attractive approach, requiring less time and computational cost, reporting strong results as a result of tuning multiple components during different instances.

Table S5. Hyperparameter search dimensions and optimal parameters identified by Random Search. Search dimensions are defined by the start value, which incrementally increases by the step size until reaching the end value.

Hyperparameter	Start	End	Step size	Optimal
RF				
max_depth	10	30	1	30
n_estimators	100	500	10	490
SVM				
Gamma	0.0012	0.003	0.00003673	0.0012
C	1	10	0.47368421	8.57894737
Epsilon	0.001	0.02	0.001	0.018
KNN				
n_neighbors	1	15	1	3
P	1	3	1	1
XGBoost				
Eta	0.1	0.15	0.01	0.1
Gamma	0	0.3	0.1	0.1
max_depth	1	8	1	4
min_child_weight	1	10	1	4
Subsample	0.8	1	0.1	0.8
colsample_bytree	0.5	1	0.1	0.9
n_estimators	100	250	10	250
NN				
Neurons	50	1000	50	550
dropout_rate	0	0.5	0.1	0.2

Appendix V. Unabridged scheme for ML-specific QSAR uncertainty assessment

This is a Snapshot of the suggested supplementary information for the assessment of QSAR models developed with ML algorithms from Chapter 4. The full document can be accessed from the following link: https://github.com/SamBelfield/PhD_Thesis



Appendix V

List of those assessment criteria for individual areas of uncertainty, variability or bias within toxicity-prediction QSAR (as presented by Cronin et al. 2019) updated in light of consideration of concerns specific to application of ML. Each is grouped in accordance with its relevance either to the reproducibility, interpretability or generalisability of models. Updates to text under heading "Comment or Other Information" are displayed in bold.

ID	Assessment Criteria for Individual Areas of Uncertainty, Variability or Bias	Example of Low Uncertainty, Variability or Bias	Example of Moderate Uncertainty, Variability or Bias	Example of High Uncertainty, Variability or Bias	Comment or Other Information	Type of Criterion	Information Potentially Retrievable from QMRF
				Reproducibility			
2.1a	Definition and description of model (related to assessment criterion 3.1a)	Model fully defined	A small number of aspects of the model non-defined or ambiguous	Model non-defined or ambiguous	All terms e.g., descriptors, statistical values, **hyperparameters and ranges,** algorithms should be defined. The QMRF is a possible reporting format.	*Uncertainty* about model if not completely defined or described: model cannot be retraced and evaluated.	Yes
2.1c	Transparency of the model	Model is transparent in terms of the algorithm and can be interpreted and reproduced	Model is defined providing some aspect of transparency, but may not be reproducible. The algorithms of, e.g., neural networks, may be difficult to interpret even if transparent.	Non-transparent model	**A transparent model can be reproduced, and the model output is (reasonably) interpretable, i.e., user can understand the causation of a prediction.**	*Uncertainty* about the model if not retraceable/reproducible, cannot be evaluated.	Yes
3.1a	Reproducibility of the model or QSAR (related to assessment criterion 2.1a)	Full documentation, availability of data and details of software do allow to repeat the QSAR de novo	Some aspects of the model, software or data are not available, meaning there is difficulty in reproducing the model	QSAR cannot be reproduced	To determine reproducibility, the model is assumed to be transparent (see assessment criterion 2.1c). **Source code should be provided, with computational infrastructure detailed.**	*Uncertainty* about the model if it cannot be reproduced.	No
3.1b	Reproducibility of the QSAR prediction	Application of the model to the same chemical always gives the same prediction result (using the same descriptors)	Model does not give reproducible predictions without careful control of descriptors	Model does not give reproducible predictions	To obtain reproducible predictions, all parameters (descriptors) need to be available and controllable. **Seeds to control randomisation for certain algorithms need to be specified.**	*Uncertainty* will be increased if predictions are not reproducible.	No
				interpretability			
2.1c	Transparency of the model	Model is transparent in terms of the algorithm and can be interpreted and reproduced	Model is defined providing some aspect of transparency, but may not be reproducible. The algorithms of, e.g., neural networks, may be difficult to interpret even if transparent.	Non-transparent model	**A transparent model can be reproduced, and the model output is (reasonably) interpretable, i.e., user can understand the causation of a prediction.**	*Uncertainty* about the model if not retraceable/reproducible, cannot be evaluated.	Yes
2.4c	Relevance of descriptors to mechanism of action/AOP	Descriptors or properties clearly and causally related to mechanism	Partial or correlated relationship to mechanism	No mechanistic basis of descriptors	**Feature importance techniques should be used for algorithms that employ large quantities of descriptors, relating highest scoring descriptors to the mechanism.**	*Uncertainty* about model if relevance of descriptors used for modelling not known or interpretable.	Yes
				Generalisability			
1.5a	How appropriate is the modelling approach for the endpoint and to deal with the complexity/non-linearity of the data	Appropriate modelling approach for the endpoint	Modelling approach likely, but unproven, to be appropriate for the endpoint	Approach likely to be too complex or simplistic	This requires a pragmatic and subjective assessment, e.g., a data set based on one mechanism with a single overriding descriptor can be modelled more simply than a more complex scenario. **If applicable, both the optimisation procedure and the sufficiency of resulting approach complexity should also be considered.**	*Uncertainty* about the model if the modelling approach chosen not appropriate. *Bias from different approaches to modelling which may result from personal knowledge, experience or prejudice.*	No
2.2a	Statement of statistical fit, performance and predictivity	Full description of model performance	Some key measures of model performance missing	Limited or no description of model performance	The use of appropriate validation methods, **resampling techniques**, and/or external test sets should be demonstrated, different metrics may be required for different models	*Uncertainty* about model accuracy and quality of the prediction if no information about the model performance.	Yes
2.2b	Interpretation of statistical fit etc with respect to biological measurement error and variability (see assessment criterion 1.2d)	Statistical performance is significant but not overfitted	Statistical performance moderate or possibly overfitted	No statistical significance or overfitted as compared to experimental error	**The use of strategies to limit overfitting e.g., early-stopping, pruning, regularisation may be required for certain algorithms.**	*Uncertainty* about the model if performance is not adequate or overfitted.	No