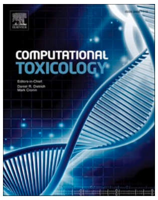
Computational Toxicology 33 (2025) 100338

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.sciencedirect.com/journal/computational-toxicology



Full Length Article

The predictivity of QSARs for toxicity: Recommendations for improving model performance

Mark T.D. Cronin *, Homa Basiri, Georgios Chrysochoou, Steven J. Enoch, James W. Firman, Nicoleta Spînu, Judith C. Madden

School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, United Kingdom

ARTICLE INFO

Keywords:
Quantitative structure–activity relationship (QSAR)
Toxicity
Molecular descriptors
Predictive performance

ABSTRACT

Quantitative structure–activity relationships (QSARs) are invaluable computational tools for the prediction of the biological effects and physico-chemical properties of molecules. For chemical safety assessment they are used frequently to make predictions of toxic or adverse effects, as well as other activities related to toxicokinetics. QSARs and their predictions can be assessed against a number of criteria for their potential use as surrogates for animal, or other, tests. A recent exercise by the Division of Genetics and Mutagenesis, National Institute of Health Sciences, Japan, assessed QSARs to predict the outcome of the Ames test. The predictive performance of models was scrutinised with full disclosure of results. The authors of this publication developed one such model, which had disappointing performance in this predictive exercise. In order to understand why the QSAR had poor performance metrics, this paper reflects on factors that affect a QSAR model. There is no one reason for poor performance of a QSAR model, rather it is likely to be a combination of factors. Reasons for poor performance included inadequate consideration of the underlying data quality, consistency and relevance; lack of appropriate descriptors relating to the endpoint and mechanism of action; not selecting a model correctly in terms of its structure (i.e., complexity) and number of descriptors; not addressing metabolism adequately in the modelling process; ill-defined assessment of the uncertainties within a model; and not ensuring predictions are within the applicability domain of the model. Whilst this paper draws on examples for the prediction of mutagenicity, the findings are applicable to all toxicological activities and physico-chemical properties.

1. Introduction

“To err is human; to really foul things up requires a computer.” Attributed to Paul Ehrlich and others….

A quantitative structure–activity relationship (QSAR) attempts to model how changes in chemical structure affect the biological activities, or physico-chemical properties, that a molecule may elicit or possess. The types of activities and properties that QSARs have attempted to model are very broad and represent almost every measured effect including, but not limited to, pharmacological activities, toxicological effects, biokinetic properties and physico-chemical properties. Mostly, QSARs attempt to produce a generalist model across a number of molecules (which may range from a small number e.g. 5 or 10, to many thousands), the model is then applied to a specific situation, such as hazard identification in chemical safety assessment [1]. The number of molecules reflects the type of model, e.g. localised in chemical and mechanistic space or globally applied across a wide range of chemistries and mechanisms [2].

The modelling of toxicity and biokinetics data has been attempted across the broad range of human health and environmental effects.

Abbreviations: DGM/NIHS, Division of Genetics and Mutagenesis, National Institute of Health Sciences, Japan; DL, Deep Learning; ECHA, European Chemicals Agency; EPA, Environmental Protection Agency; EU, European Union; FAIR, Findability, Accessibility, Interoperability, and Reuse; FL, Federated Learning; GLP, Good Laboratory Practice; HTS, High Throughput Screening; LJMU, Liverpool John Moores University; log P, logarithm of the octanol-water partition coefficient; MIE, Molecular Initiating Event; MHLW, Ministry of Health, Labour and Welfare; NGRA, Next Generation Risk Assessment; NTP, National Toxicology Program; OECD, Organisation for Economic Cooperation and Development; PCA, Principal Component Analysis; QAF, QSAR Assessment Framework; QSAR, Quantitative Structure-Activity Relationship; RAAF, Read-Across Assessment Framework; REACH, Registration, Evaluation, Authorisation and restriction of Chemicals; RF, Random Forest; SHAP, SHapley Additive exPlanations; US, United States.

* Corresponding author.
E-mail address: m.t.cronin@ljmu.ac.uk (M.T.D. Cronin).

https://doi.org/10.1016/j.comtox.2024.100338
Received 6 October 2024; Received in revised form 25 November 2024; Accepted 5 December 2024

There are various objectives to this activity which generally aim to support chemical safety assessment. The specific purposes of creating predictive models range from the detailed assessment of hazard and exposure to support risk assessment of a single compound – which may include the possibility to replace animal testing – through to the assessment of a large inventory of compounds to prioritise those for further assessment [3]. Whilst founded in the 1960s using linear regression, the modelling and statistical approaches applied to QSAR have expanded to be more multivariate and non-linear, with neural networks being applied since the 1990s and the opportunity of artificial intelligence recently becoming a reality [4].

Providing certain criteria are met, the expectation is that predictions from QSARs will be justifiable and correct. With regard to predicting the toxic effects, fate and properties of a substance, particularly with a regulatory focus such as the European Union (EU) Registration, Evaluation, Authorisation and restriction of Chemicals (REACH) regulation, we have many decades of experience to call upon [5–7]. The general principles are that a prediction from a QSAR model may be considered acceptable for regulatory purposes providing i) the scientific validity of the model can be demonstrated and ii) the compound for which the prediction is made is within the applicability domain of the model [8]. In order to demonstrate "validity" we can apply the Organisation for Economic Cooperation and Development (OECD) QSAR Principles [9] to QSAR models; for read-across and grouping the European Chemicals Agency (ECHA) Read-Across Assessment Framework (RAAF) [10], assessment of uncertainties [11], or other methods can be applied. Copious guidance is provided [8,12] to support these approaches, and recently the overall assessment of predictions has been updated through the OECD QSAR Assessment Framework (QAF) [8,13].

In addition to the formal guidance on using QSARs to predict toxicity, there is considerable anecdotal knowledge on good practice in toxicity prediction [14–17]. Many aspects of modelling, and the use and application of models, have been stressed including, but not limited to, a clear definition of the problem to be addressed using a modelling approach, understanding the quality and type of data to be modelled, the role and value of mechanistic interpretation, appropriate goals for the statistical modelling and its assessment etc. Many of these aspects are captured under the auspices of the OECD QSAR Principles and QAF, but are reliant on expert knowledge (in both toxicology and modelling) and are often open to interpretation.

The practical assessment of the performance of QSARs to predict toxicity has been undertaken through a number of exploratory exercises. They include the United States (US) National Toxicology Program (NTP) comparative exercises on the prediction of a relative small number of compounds for rodent carcinogenicity, which went some way to illustrating the issues behind the *in silico* prediction of complex endpoints [18,19]. Worth et al. [20] assessed a variety of QSAR software to predict the genotoxicity and carcinogenicity of pesticides. More recently, the Tox21 Challenge addressed a much larger data set of high throughput screening (HTS) data for nuclear receptor and stress response pathways. Typically, predictive models were provided, with a strong emphasis on machine learning approaches [21]. There have also been collaborative modelling projects where the purpose was less to assess predictive performance, but rather to create a selection of models, based on different approaches, to develop consensus models. These latter approaches were co-ordinated by the US Environmental Protection Agency (EPA) and the NTP Interagency Centre for the Evaluation of Alternative Toxicological Methods (NICEATM), amongst others, and included approaches to predict oestrogen receptor activity (CERAPP – Mansouri et al. [22]), androgen receptor activity (CoMPARA – Mansouri et al. [23]) and acute oral rodent toxicity (Mansouri et al. [24]). The EPA/NICEATM approaches ensured that all data and models were freely available.

Another "blind trial" was undertaken recently by the Division of Genetics and Mutagenesis, National Institute of Health Sciences, Japan (DGM/NIHS). From 2020 to 2022 the DGM/NIHS conducted the Second Ames/QSAR International Challenge Project [25], extending the findings of the First Ames/QSAR International Challenge Project [26]. The purpose of the Second Challenge Project was to extend the First Challenge Project with more model developers, as well as the implementation of machine learning. The Second Challenge Project had an improved test set which allowed QSAR developers to make predictions for a dataset of approximately 1,600 chemicals, with the opportunity to develop models from a training set of approximately 12,000 chemicals. The endpoint selected for model development and predictions was the Ames test with data being made available that had been submitted to the Ministry of Health, Labour and Welfare (MHLW) in Japan since 1979. Full details of the data are provided by Honma et al. [26] and Furuhama et al. [25]. In total, predictions were submitted from more than 50 models created by 21 model developers from eleven countries. Furuhama et al. [25] describe the breadth of modelling approaches which represent a variety of rule-based and statistical approaches. Predictions for the test set were submitted to DGM/NIHS with the performance assessed by Furuhama et al. [25], who also made general observations regarding model performance.

The overall predictive performance metrics of selected models in the DGM/NIHS exercise are detailed in two tables (Tables 9 and 10) published in Furuhama et al. [25] and the supplementary information of Uesawa [27], with balanced accuracy of the models ranging from 49.6 % to 78.5 %. The authors of this publication (from Liverpool John Moores University (LJMU), UK) submitted predictions from two modelling approaches: Deep Learning (DL) and Random Forest (RF). The performance of the models was variable with, perhaps surprisingly, the DL model having the lowest balanced accuracy of any model reported (49.6 %). Thus, whilst the performance could be considered disappointing, it provided an opportunity to reflect on reasons why the QSARs may have performed badly and what could be learned. Thus, the aim of this manuscript was to identify areas that are vital (and often overlooked) in QSAR development, and so provide guidance for the further development and application of QSARs. The findings are illustrated, where possible, with reference to the LJMU DL model reported in Furuhama et al. [25] in particular. Further, new technologies which may support QSAR modelling are identified.

2. Methods

2.1. Details of the LJMU models published in Furuhama et al. [25]

2.1.1. Ames test data provided and assessment of the performance of models

Ames test data in the training and test sets were derived from submissions to the MHLW in Japan. Data were used as provided, namely the chemical name and SMILES string, and Ames Class for the training set: induction of more than 1,000 revertant colonies per milligram, in at least one Ames test strain, with or without of metabolic activation, is termed Class A (strong positive). Induction of a minimum of a 2-fold increase in revertant colonies, (fewer than class A) in at least one Ames strain, with or without metabolic activation, is termed Class B. 'Negative' or Class C was determined by a less than 2-fold increase in revertant colonies. No further information on the test was provided and the data were used as provided. In total, information on 12,134 chemicals was provided for the training set and 1,589 for the test set. No molecular descriptors or advice on modelling was provided with Ames test data.

LJMU proceeded to develop two models, described in Sections 2.1.2 and 2.1.3, from the training set data provided, no additional Ames test data being added. The models were subsequently utilised to make predictions for the test set following calculation of descriptors from the SMILES strings in a similar manner to that for the training set. The predictions for the test set from the models were provided to the Challenge co-ordinators without knowledge of the activity of the compounds. The co-ordinators assessed the performance of the models according to set of metrics commonly utilised to assess predictivity (summarised in Table 1). This investigation draws upon the performance of the LJMU DL model, in particular, which was provided initially to the modellers and subsequently published in Furuhama et al. [25].

2.1.2. LJMU deep learning (DL) model

For the LJMU DL model, the complete training set as provided by Furuhama et al. [25] was utilised. For each compound in the training and test sets, 1,343 one-dimensional and two-dimensional molecular descriptors were generated using the PaDEL software (version 2.2.1) [28]. Following the exclusion of those features possessing either no internal variance (i.e., complete uniformity in output), or excessive collinearity (dropping one from each pair with correlation greater than 0.80), a pool of 283 PaDEL descriptors remained. The PaDEL descriptors were supplemented by 1,227 PubChem Substructure Fingerprints (version 1.3) [29] and ToxPrint Fingerprints (version 2.0 r1520) [30] and the complete data set was utilised as the input. Classifications of mutagenicity activity were provided as the output. Chemical descriptors or features with a unique value (0 or 1) in the training set were excluded. The list of training chemicals was split into training and test sets to train and evaluate the performance of the models, prior to the application to the Furuhama et al. [25] test set. Numerical variables were normalised. Pre-processing layers were set for categorical and numerical input data. Given the imbalanced list of training chemicals, the classes of positive and strong positive were weighted. Classes utilised were: strong positive (A), positive (B), and negative (C).

Three approaches were investigated to evaluate which performed best on the training set. The models included:

- A deep learning model for the problem of allocating to three classes.
- A wide and deep learning model for the three-classification problem. This model is a combination of the deep neural network developed above and an addition of a wide linear model to capture the feature pair correlations and to generalise better for unseen features combinations [31].
- A wide and deep learning model for the two-classification problem. The model predicts if a compound is mutagenic or non-mutagenic compared to the previous two models.

The optimizer used was Adam. One dense layer with one dropout layer was applied. Three hidden layers were included. Rectified linear activation function was used to provide non-linearity to the model for fast convergence. Keras and TensorFlow in Python 3 were used to construct and train the models. Even though the two-classification DL model showed better performance, the three-classification model was submitted to allow for the external validation to be performed, and is summarised in Table 1. Further details of the LJMU DL methodology are provided in the Supplementary Material of Furuhama et al. [25].

Table 1
Summary of the performance of the LJMU models on the external test set as described in Furuhama et al. [25].

Performance metric	Average (min–max)	LJMU DL Model	LJMU RF Model 1	LJMU RF Model 2
A-Sensitivity (%)	65.1 (5–95.7)	19.0	79.7	75.9
Sensitivity (%)	49.5 (2.5–80.5)	20.0	74.5	64.7
Specificity (%)	84.9 (55.3–99.6)	79.3	55.3	72.7
Accuracy (%)	79.7 (58.1–86.8)	70.5	58.1	71.5
Balanced Accuracy (%)	67.2 (49.6–78.5)	49.6	64.9	68.7
Positive Prediction Value (%)	40.3 (14.4–60.3)	14.4	22.5	29.2
Negative Prediction Value (%)	90.9 (85–95)	85.0	92.5	92.2
Mathews Correlation Coefficient	0.32 (−0.01–0.45)	−0.01	0.21	0.28
Coverage (%)	95.4 (35.9–100)	99.5	99.5	99.5
A-F1 score (%)	46.9 (9.1–63.1)	16.4	35.1	42.2
F1 score (%)	41.3 (4.8–53.8)	16.8	34.6	40.3

2.1.3. LJMU random forest (RF) model

Using the 283 PaDEL descriptors as defined in Section 2.1.2, the RandomForest package (present within RStudio) was applied. Two distinct random forest models were trained: one employing the classification system as provided with the training data (categories A, B and C), and another adopting a simplified binary scheme (with A and B pooled as “active”, C as “inactive”). These shall henceforth be referred to, respectively, as Models RF1 and RF2.

In each instance, it was necessary to address the imbalance in distribution of categories present within the dataset (5.35 % of compounds labelled A, 9.02 % B and 85.6 % C). For Model RF1, this was achieved through means of the random undersampling of chemicals falling within categories B and C, so that their numbers were each equal to those within category A. The result was a working list comprising 1,935 compounds. A similar methodology was applied in the construction of Model RF2, where the quantity of “inactives” (C) were reduced to match the combined sum of A and B (“actives”). A training set of 3,466 chemicals resulted.

Both RF1 and RF2 were optimised by way of manual tuning, with adjustment of the hyperparameters ntree (i.e., quantity of decision trees constituting the “forest”) and mtry (i.e., number of descriptors drawn upon for the purposes of tree splitting). Performances were evaluated following the 70:30 division of the aforementioned working lists into training and test sets. It was observed that, within each, combination of ntree = 250 and mtry = 35 provided superior outcomes. Further details of the LJMU RF methodology are provided in the Supplementary Material of Furuhama et al. [25].

2.2. Reflection on the performance of the models

The DL and RF QSAR models developed by LJMU for the prediction of Ames test data, as described in Section 2.1 and Furuhama et al. [25] were critically evaluated with their performance. The performance of the LJMU DL and RF models is summarised in Table 1. Since the performance of the LJMU DL model was poor, with a balanced accuracy of 49.6 % (less than chance), this was used as the subject of the evaluation and reflection. The evaluation was not a formal process, i.e. the OECD Principles were not applied, rather the evaluation attempted to consider the models in the light of the performance of the blind predictions reported in Furuhama et al. [25]. In particular, the consequence of the data type, model structure and descriptors, as well as the predictions themselves were considered. In addition to the reflection on why the models performed sub-optimally, other aspects were considered relating to the use of the predictions.

3. Reasons for poor model performance

The QSARs for the prediction of Ames test results, developed by LJMU, as reported in Furuhama et al. [25], were evaluated with reference to their performance in predicting a blinded test set and are summarised in Table 1. The purpose of this exercise was not to validate the QSARs, but rather to identify reasons for potential poor performance, such that they could inform the QSAR community for the creation of better models – regardless of endpoint. The evaluation identified various reasons for poor model performance (Sections 3.1 – 3.3) as well as more general issues related to the use of the models in (non-animal) chemical safety assessment (Section 3.4).

There is no doubt Ames test data are amenable to QSAR modelling [32], with many modelling approaches reported in Furuhama et al. [25]. A reflection on the performance of the LJMU DL model should also consider the other models in the investigation. There is a very broad range of modelling approaches including knowledge-based approaches, i.e., the application of structural alerts in expert systems through to multiple types of statistical analyses including machine learning, with complexity up to the deep learning technologies. A wide range of molecular and structural descriptors are applied in the models. Nearly all, if not all, descriptors are calculated directly from 2D chemical structure using freely available software or the model developers' own software. Lastly, and potentially most significant, are the data that are modelled. The LJMU models were based solely on the data provided by NIHS, with no attempt to supplement the data. A number of model developers utilised their own existing models and enriched them with the NIHS data, thus providing a potentially more robust dataset and model. Further assessment of these factors could be achieved by greater analysis of the findings provided to NIHS and reported in Furuhama et al. [25], regrettably it is not possible in terms of this investigation.

Full statistical analysis of the findings, in terms of predictivity of the blind test set is reported by Furuhama et al. [25], and summarised with regard to the LJMU DL model in Table 1. The analysis was latterly extended by Uesawa [26]. The assessment as to what may be considered the best model is complex and is based not only on balanced accuracy alone. Furuhama et al. [25] state that high sensitivity, low false-negative rate and wide coverage of chemical space are needed for QSAR models in the regulatory setting. Thus, these criteria should also be borne in mind when assessing overall model performance.

A number of reasons why a model may perform poorly were identified and are summarised below. These are a compilation of many issues previously identified as pitfalls and good practice [14,16,33], associated with known uncertainties [11,34] as well as being part of the accepted processes for the assessment of QSARs [9] and their predictions [8,13] for regulatory purposes. The factors discussed below reflect the three main elements of a QSAR, namely the data modelled (Section 3.1), the independent variables describing molecular structure and properties (Section 3.2), the modelling approach (Section 3.3) as well as elements of how the model is developed and can be applied (Section 3.4). The assessment of the impact of these factors on the performance of the LJMU DL model for mutagenicity is provided where appropriate.

3.1. Reasons for poor QSAR model performance: Endpoint data being modelled

The data to be modelled, as well as their description and curation, are fundamental to the modelling process and the performance of the model [7]. There are numerous criteria regarding the data that will affect model performance. The overriding principle is that the quality of the data (both the values and their description/curation) will reflect the quality of the model. Section 3.1 reflects on some issues that, in the first instance, should be borne in mind by the model developer. Whilst there is no strict definition of "high quality" data, with regard to toxicological evaluation, e.g. for regulatory purposes, this would normally be associated with a test being performed to a standard Test Guideline and under Good Laboratory Practice (GLP) conditions. However, other useful and usable QSARs may be developed from non-standardised data, such as those from receptor binding assays. It is noted that models from non-standardised (non-validated) data have to overcome greater scrutiny before acceptance for regulatory purposes.

In this Section, it is assumed that the data to be modelled are from an experimental measurement, usually from a standardised and recognised toxicological assay (the same criteria will apply to other endpoints modelled, such as toxicokinetics and physico-chemical properties however). The data modelled in the LJMU DL neural network for mutagenicity were from the Ames test and supplied by DGM/NIHS. Therefore, they were equivalent for all models/modellers so will not explain poor inter-model performance.

3.1.1. Inadequate/incomplete or non-existent data curation

The data on which a model should be developed should be curated. Specifically, a biological activity should be associated with an identifier for a defined chemical structure, or substance. A number of reports (for instance Tropsha [33]; Alves et al. [35]) have emphasised the need for unique and unambiguous identifiers, as well as consistency in addressing the representation of entities such as salts or ionised molecules, stereoisomers and tautomers.

The fundamental need to capture substance identity correctly, as the starting place for data curation and association of biological data with chemical structure, is vital. Consistency is required in the recording of chemical structures for model development. There is a need to ensure accurate mapping from chemical name (especially if not in IUPAC nomenclature), CAS number through to SMILES strings or InChi/InChiKeys. Automated processes of structure generation should be checked and verified manually, where possible, against known standards. An example of where issues from incorrect generation of SMILES strings or InChiKeys may arise is with descriptor-generating software which is liable to vary in its handling of alternative substance forms (for example, a free carboxylic acid, its carboxylate anion and its sodium salt). Although these entries might well exhibit near-identical toxic potencies, it is the case that substantially different feature values may nevertheless be attributed to each. Conversely, such calculations typically draw no distinctions between stereoisomers – the activities of which have the potential to contrast greatly. Combined, these factors serve to confound the establishment of statistical association between structure and activity. For this reason, salts (and similar) should ideally be represented as neutral, mono-constituent organic compounds. Duplicate molecules or substances should be removed altogether, alongside all non-discrete or ill-defined entries such as polymers and mixtures. A further issue with the correct reporting of biological data is the accuracy in the transfer of the data from the original source. Errors have been observed in data transference, with the ideal requirement being checking against the original study report [15].

3.1.2. Variability in biological data

The data on which the model is developed are fundamental to the quality of the model, and hence should temper what may be expected, or feasible, in model performance. Poor quality data should, inevitably, lead to lower expectations in the overall model. Due to the variability in the biological measurements on which they are based, no QSAR model will ever achieve consistent 100% accuracy. Whilst not explaining poor performance of models, it should be borne in mind that even low quality models may provide useful information within a consensus approach (see Section 3.4.4).

In the study reported by Furuhama et al. [25], the data have been provided from regulatory submissions to DGM/NIHS. During the study and model development, no attempt was made to evaluate data quality. This is not to be considered as a criticism, but rather an uncertainty within the modelling process (refer to Section 3.4.2 below). Whilst there are no definitive statistics for the reproducibility of the Ames test, a recent study [36] indicated that intra-laboratory reproducibility of negative and positive results (with and without S9 mixes) was high – often over 90%. It is likely that strong positives and negatives will be highly reproducible and it was noted that there is lower reproducibility for equivocal results [36]. Whilst Zeiger et al. [36] focussed on within laboratory reproducibility, much less is known about between laboratory reproducibility, which is likely to be lower. There are many potential reasons for poorer inter-laboratory variability including, but not limited to, local variations in methodology and differences in doses tested. The doses tested are of particular relevance, should only low doses be tested then a weak mutagen may not be identified. In addition, the Ames test is valid only up to the level of cytotoxicity, hence a weak mutagen that may promote cytotoxicity will not be identified. The interpretation of the findings by the toxicologist should also be borne in mind. Whilst the Ames test is highly standardised, there are always elements of interpretation of the results of any experimental findings. Since the Ames test data prepared by DGM/NIHS were from multiple sources, and potentially methodologies, it is difficult to assign any statistic for reliability. What may be concluded is that the DGM/NIHS data set will have lower reproducibility than for the 90% inter-laboratory reproducibility reported by Zeiger et al. [36]. As such, an educated estimate of reliability of the data set, in terms of reproducibility, is likely to be in the region of 70%, with a realisation that it may be 10% higher or lower. Very broadly speaking, this is reflected in the balanced accuracy of the predictions reported for all the models in Furuhama et al. [25], with a range of 49.6–78.5%. As such, it may be that the models have reached, or possibly surpassed, the acceptable experimental limits of data.

The reproducibility of the Ames test should not be confused with its performance to predict carcinogenicity – a positive Ames test is highly indicative of carcinogenicity but a negative result less predictive (in part due to the issue of non-genotoxic carcinogens) [37]. The lack of relevance of the concordance of the Ames test with carcinogenicity is an issue for the application of the data in risk assessment, but not the modelling process.

It is noted that not knowing data quality does not explain poor model performance in this exercise. However, it will guide what may be considered to be acceptable model performance. It is our opinion that a model is not capable of making predictions better than the data on which it is built. Therefore, low quality data should only be considered to create a low quality model. Since there is variability in all biological measurements, however well standardised the test, this must be factored in to how they are used.

3.1.3. Lack of toxicological mechanistic insight and relevance

The ability to relate or develop a model in an appropriate manner to account for mechanism of action is fundamental to modelling of toxicological activity. This is also an essential aspect of the validation of QSAR models, being enshrined in the fifth OECD principle, which implies a QSAR should be associated with "*mechanistic interpretation, if possible*" [9]. In the development of the models for Ames test data, the chemical structures and activities were provided, but no mechanistic information was available or assumed. It should, of course, be hypothesised that the mechanism in the Ames test relates to the ability of chemicals to induce mutations in DNA [38].

The molecular initiating event (MIE) associated with the Ames test, which would promote the reverse mutations, is undoubtedly the interaction of the chemical with bacterial DNA. However, it is probable, although not documented for this data set, that there may be a variety of mechanisms that underpin the interaction with DNA, e.g. various electrophilic mechanisms, formation of reactive oxygen species, production of reactive metabolites etc. [39]. This information would be highly valuable to direct modelling in terms of the appropriate molecular descriptors (Section 3.2.1), accounting for metabolism (Section 3.2.2) and the appropriate structure for the model (Section 3.3.1).

As is common with the use of machine learning in QSAR, the LJMU DL Model was created without implicit reference to mechanism of action. Therefore, greater insight into the mechanism of action would have been likely to improve the models for Ames test data. This could have been improved by restricting the descriptor pool to only those of mechanistic relevance, or adding in more relevant descriptors.

3.1.4. QSAR models are simplifications of biological data which are based on complex processes or outcomes

It is trivial to state that all biological processes are immensely complex, especially at the *in vivo* level. By implication, QSAR attempts to create a simplified generic model of biology, encoding physiology, biochemistry, toxicodynamics and toxicokinetics into a small number of chemistry-based descriptors. Again, using a simplistic analogy, machine learning has the potential to utilise non-linear models to obtain the maximal fit to the data. It is therefore essential that there is an appreciation of the simplification of biology within the model. Such oversimplification may, inevitably, lead to the model not capturing all elements of biology.

The data reported by Furuhama et al. [25] are for the Ames test, or *in vitro*, in nature. Whilst *in vitro* tests may be considered themselves as a simplification of the more complex *in vivo* system, there remains complexity in the process. With regard to the Ames test, any model of the molecule should accommodate complexity of uptake of the molecule into the cell, reactivity and potentially metabolism. Not capturing these issues will inevitably lead to poor performance of the model.

3.2. Reasons for poor QSAR model performance: Descriptors of chemical structure and/or properties

Following the collection and curation of experimental data for modelling, the subsequent task in QSAR development is to obtain appropriate descriptors (the so-called independent variables) of molecular structure and properties. Descriptors may, on occasions, be derived experimentally, e.g., the logarithm of the octanol–water partition coefficient (log P), but are usually calculated directly from a representation of molecular structure. The exact number of molecular descriptors calculable is not known, but is likely to be in the region of 5,000 – 10,000 for each molecule, more if molecular fingerprints are included. The reader is referred to excellent reviews of molecular descriptors for more details; the fundamentals are described by Dearden [40] with detailed analysis by Todeschini and Consonni [41]. The task of the modeller is to obtain and utilise molecular properties and descriptors that are appropriate to the chemicals in the data set as well as the complexity and variability of the endpoint being modelled. Inappropriate or inadequate descriptors will result in poor model performance as they do not account for the influence of the variability of chemical structure on activity. With the advent of machine learning there has, therefore, been a trend to calculate large numbers of descriptors allowing the model to select those appropriate to fit the relationship between biological activity and chemical structure/properties. There are many advantages to this process, such as it being rapid and not informed by modeller bias. However, there are also many issues with the inappropriate or inadequate use of molecular descriptors, some of which are accounted for in this Section.

Recently, a combination of various molecular representation methods based on chemical sequence e.g., as defined by a SMILES string, molecular images and graphs have been explored to capture the chemical space potentially better [42]. This aligns perfectly with the need for larger datasets for deep learning methods. However, more efforts are still required for the causal interpretation of these multimodal data integration approaches.

3.2.1. Models need relevant molecular descriptors

It is obvious that the inputs into a model need to be relevant and of high, or known, quality. With regard to the descriptors of chemical structure and properties, these must be related to the activity that is being modelled. This is best achieved by an understanding of the mechanism of action of the effect being modelled. This implies that the descriptors in a model are related directly to biological outcome, whether this is the ability to get to the site of action, or the interaction of the molecule with the target site (sometimes referred to as the MIE) [43]. It is well established that fundamental high quality QSARs are likely to be developed for smaller datasets, where a single mechanism is known and only one, or a small number of descriptors may be required [44]. Modelling datasets with multiple, or unknown, mechanisms greatly increases the required complexity of the model [2].

A positive outcome in the Ames test is predominantly a result of the covalent interaction of the molecule, or its reactive metabolite, with the DNA within the bacterial cells. Early QSAR studies of mutagenicity demonstrated some reliance of the potency in the Ames test with hydrophobicity for restricted groupings of amines [45] and nitro compounds [46]. However, the overwhelming property that needs to be incorporated into the model is electrophilic reactivity.

Modelling of electrophilic reactivity is best achieved using quantum chemical descriptors, such as the energy of the lowest unoccupied molecular orbital or atomic superdelocalisability [47,48]. These require calculation of an appropriate 3-dimensional structure followed by often complex, time-consuming and costly quantum chemical (or higher level) properties [49]. Given the challenges of a training set comprising over 12,000 molecules, quantum chemical calculations were not feasible. Therefore, modelling of mutagenicity utilised 2-dimensional descriptors from the PaDEL software. These are based on various aspects of chemical structure and, whilst some descriptors may be representative of electrophilic functional groups, these descriptors do not implicitly encode electrophilic reactivity.

Thus, to improve the performance of QSARs, descriptors relevant to the endpoint being modelled, as well as the mechanism(s) of action, should be included. In order to achieve this, greater cognisance of mechanisms of action is required, which is discussed in Section 3.1.3. Descriptor selection should, if possible, also be related to the chemical space and characteristics of the data set which is being modelled. Thus, if there is only one reactive feature in a data set, as may be found, for instance, with a group of nitrosamines, then descriptors need not focus on alerts for mutagenicity but rather factors that modulate the reactivity of the nitrosamine group and its bioavailability. The data modelled in Furuhama et al. [25] were chemically heterogeneous, thus needing a broader range of descriptors to capture intrinsic reactivity with DNA.

3.2.2. QSAR models do not explicitly account for metabolism (or other toxicokinetic factors)

Metabolism of xenobiotics is a fundamental physiological and biochemical process that assists in their detoxification and clearance. Conversely, some of the same metabolic process may elicit a more active molecule, or a more reactive intermediate. This is a crucial mechanistic step in many adverse outcomes and is dependent on the chemistry and the metabolic capability of the species and/or individual [50,51].

For some chemicals, metabolism is an essential process to promote mutagenicity and is captured through the inclusion of an S9 mix in the Ames test. Since the activity data were modelled as positive and negative, the effect of metabolism was not known, but captured in the overall biological activity. This will cause problems as the descriptors are based solely on the parent molecule and not the metabolite, hence this could lower performance.

In terms of modelling, a number of approaches to resolving the issues of metabolites can be taken. It may be assumed (or hoped) that metabolism is captured in the model in some way, i.e., some descriptors incorporate this. For instance, descriptors for a primary aromatic amine group may be important for predicting mutagenicity, however, it is not the amine, rather it is the nitrenium ion formed through metabolism [52,53] that is responsible for the effects. Other approaches, which were not adopted in the DGM/NIHS Second Challenge Project, would be to identify molecules capable of being metabolised and model them separately, or predict metabolites computationally and perform separate assessments. The concern with the latter approach of predicting metabolites, is the explosion in the numbers of potential metabolites, without knowing which are necessarily relevant. With regard to the prediction of metabolites, further work is required to identify only those that are relevant and of concern from a toxicological perspective.

3.2.3. Ensuring "accessibility" of descriptors

Individual descriptors for use in a QSAR model should be calculated using the same software, or in the case of experimental descriptors measured in the same manner. This has clear implications for the use of historic models. As such, prospective model performance may be adversely affected if descriptors generated using different approaches are utilised. This is an important aspect of the sustainability of a model as well as its adequate description. Thus, it is fundamental that when software is used to calculate descriptors the version of the software must be recorded and, where possible, this version must be used to calculate future descriptors for molecules of interest for which predictions are made. Where this is not possible, descriptors should be re-generated for all molecules of interest, or this must be acknowledged as a limitation to the modelling approach. Other aspects related to the accessibility of descriptors include: the availability of software, its cost, its accuracy, reliability and reproducibility in calculating descriptors, and the time and computational resources that may be required to obtain the descriptors.

3.2.4. Modelling of excessively large sets of molecular descriptors

As noted above, it is possible to calculate 1,000s of descriptors very rapidly. Whilst this may be useful for machine learning techniques, there are a number of pitfalls, potentially contributing towards poor model performance, that need to be recognised. The reality is that it is now possible to have "oversquare" data matrices where there are significantly more descriptors than there are chemicals in the data set. As such, these data matrices may have considerable redundancy of information.

Given a sufficient quantity of features, it is possible to develop models which are almost perfectly predictive with respect to their training sets. These are, however, likely to be overfit to a significant degree. Rather than accounting exclusively for key structure–activity relationships, output is informed instead by chance correlations arising between "excess" descriptors and the target variable. In essence, such models are so acutely attuned to their training data that their ability to generalise towards unseen substances is impaired. The probability of encountering interrelatedness, or collinearity, between descriptors within a large feature pool is high. For example, molecular weight and number of heavy atoms, as representatives of general compound size, are likely to correlate closely. Although any minor variation present between the two may reasonably be anticipated as incidental (i.e., noise), it will nevertheless be drawn upon by the algorithm in its search to improve statistical fit.

Ahead of training, the extent of collinearity between all descriptors should, ideally, be determined. Feature reduction can thus be enacted, whereby one from each descriptor-pair exceeding a defined correlation threshold (e.g., Pearson correlation coefficient = 0.7) is removed. Whilst smaller feature sets may produce models less liable to overfitting, the general loss of information may result in predictivity being reduced. In order to identify the presence and extent of overfitting, it is common to apply techniques such as cross-validation (typically ten-fold) – noting variation between the training set and validation set performance metrics. Other methods for descriptor reduction can include selection of significant descriptors, or elimination of less significant descriptors – this is performed in stepwise regression analysis and, to a certain extent, in RF. Other approaches include the reduction of descriptor dimensionality through principal component analysis (PCA). The potential disadvantage of approaches such PCA is that the new descriptors, the principal components, may be difficult to assign a physico-chemical or structural meaning to and will therefore limit interpretation of the model.

3.3. Reasons for poor QSAR model performance: Statistical relationship or model applied

The statistical model that provides the bridge between biological data and descriptors is the third element of a QSAR. Starting in the 1960s with linear regression analysis, a multitude of statistical methods have been attempted incorporating multivariate statistical approaches and now embracing machine learning and generative artificial intelligence approaches. There is no formal guidance or recommendation as to which method to apply. As such, there is a trade-off between the desire for simplicity and transparency, as compared the ability of machine learning approaches to utilise multiple streams of descriptors in a non-linear manner. There is no doubt that predictive performance of

QSARs will be affected by the model used, this Section investigates how and why different modelling approaches may be responsible for different effects on predictions.

3.3.1. Ensuring the modelling approach is appropriate to both the question being asked and the type and extent of data available

Modellers should not treat toxicological data as merely as numerical inputs – unless the purpose of the model is simply to search for patterns in the data. Toxicological data are the outcome of an often complex biological process and should not be generalised. The modelling should take into account these complexities and therefore the model should have an appropriate structure. Many structure–toxicity relationships are complex and non-linear, however some are relatively simple. For instance, some endpoints, such as cytotoxicity, may be modelled successfully by a single descriptor in a linear relationship [54]. Since mutagenicity is a categorical endpoint, the modelling approach is implicitly aiming to find descriptors that can separate mutagenic (or reactive) molecules from non-reactive molecules.

3.3.2. Bigger doesn't always mean better…

As a modeller, there is always the temptation to want as large a data set as possible. The DGM/NIHS data set provided such an opportunity and, as we move into an era of greater data accessibility, further datasets will also materialise. However, modelling of large toxicological data sets may not be intuitive as it may hide (or mask) significant areas of chemical space where the model is inappropriate. This does not, of course, preclude the development of models from large datasets, but the limitations must be borne in mind.

The concept of activity cliffs within data, where a small change in structure may bring about large change in activity, is well established [55]. This is likely to be highly relevant for the Ames test, where small changes in structure will be highly influential in determining whether a molecule may be reactive or metabolised to a reactive intermediate. As noted in Section 3.2.1, such molecular subtleties are difficult to capture with the types of 2-dimensional descriptor being commonly utilised in model development. As such, small, but highly relevant, changes in molecular structure may not be captured in models for large, chemically diverse datasets and the possible learning from it lost within the statistics.

As part of the process of applying models, we should focus on areas of chemical space where we can assess the prediction, i.e., where there is homogeneity in the accuracy of predictions. Likewise, if we can understand chemical space, or parts of the model where there is poor predictivity we may be able to focus on those, either by improving the model or recognising that area as having higher uncertainty. Global models and statistical analyses hide such distinctions. Assessment of nearest-neighbours is one means of achieving this, albeit in a rather artificial manner. It would be beneficial to provide a better assessment of predictivity within chemical space. In other words, knowledge of the prediction quality of the space the target chemical finds itself in.

3.3.3. Setting realistic performance criteria

Setting performance standards as criteria for acceptance of the predictions for a model is often unhelpful, if not unwise, as it may eliminate potentially pertinent lines of evidence from an overall weight of evidence. There is a temptation for the model developer or user to rely on predetermined measures of statistical fit, i.e. a valid model is associated with a certain statistical fit. In fact, few, or no, legislations make reference to pre-defined statistical criteria and it is the authors' opinion that there is a danger in so doing. It is possible to manipulate a model, even if done unwittingly, to increase statistical fit, e.g. by rationalising selection of data or removal of outliers.

A more realistic option is to ensure that performance criteria of a model are realistic with regard to the biological (or other) data being modelled – as noted in Section 3.1. Whilst performance criteria may be less than objective, it is important to benchmark, or baseline, a model against the type of data it is intended to predict. This will assist the model developer or user in understanding whether the model is over- or under-fit. Thus, the model performance itself should not be considered alone, although an important diagnostic of the model, rather its performance against the baseline for the data should be considered.

3.4. Reasons for poor QSAR model performance: Other factors related to the development and application of the model

There are a number of other factors that may affect the performance of a QSAR model adversely. These are not necessarily related to the biological and descriptor data, or directly to the method used to create the model. Rather they relate to the motivation for the creation of a model, its description and storage. In particular, the purpose for which the model is developed should be borne in mind, and whether its subsequent use is commensurate with that purpose.

3.4.1. Establishing the need for, or application of, the QSAR – Problem formation statement

QSAR models are usually developed for a specific purpose(s), that purpose should be conceived before modelling begins. There are many specific purposes for the creation which could include hazard identification (i.e. toxicity effect prediction), through to exploration of a data set, training and education etc. A clear statement of the need, or purpose, of a model will assist the user in applying it correctly. For instance, a model developed to investigate the chemical space of a large inventory of compounds may not be suitable for the direct prediction of toxicity of compounds with known mechanisms. Whilst a model may be used for an alternative purpose in future, this deviation from the intended use of the model should be acknowledged when assessing its performance and utility for that purpose. The purpose of creating models for the dataset provided by Furuhama et al. [25] was to make predictions of the Ames test for a large number of compounds and enable superior model development. Thus, a large global model was required and, indeed, provided by the contributors to Furuhama et al. [25]. However, especially in the case of the LJMU DL, the model cannot be considered to have been superior to those already available.

3.4.2. Uncharacterised uncertainties

In addition to full documentation and description of a QSAR model, it can be considered essential to characterise the uncertainties within it. Consideration of uncertainties is a fundamental component of ECHA's QAF [8]. Separate to the QAF, a framework to characterise uncertainties has been developed [11] and its application to assess QSARs demonstrated [3]. No real analysis of the uncertainties of any of the models for mutagenicity reported by Furuhama et al. [25] was undertaken, therefore users of these models have little insight into key factors that may influence the performance of the models.

It would appear that there are uncertainties associated with various aspects of the models. The data on which the models are built have been curated by DGM/NIHS, however access to original study reports and precise protocols is not available. Similarly, there is incomplete access to the descriptor data and model architecture in the models reported in Furuhama et al. [25]. There is also no definitive mechanistic information, other than a positive outcome in the Ames test is indicative of disruption of bacterial cells.

Not characterising uncertainties does not invalidate the model, nor does it directly affect the performance of a model to make predictions. However, a better characterisation of uncertainties would enable an overall evaluation of the model, with a particular emphasis on how it may be used. Characterisation of the uncertainties will also pinpoint areas where they may be reduced, which in itself could improve, or allow for the better understanding of, model performance.

3.4.3. Applicability domains poorly, or not at all, described

The applicability domain of a QSAR should be defined, and is explicit in the third of the OECD principles for the validation of (Q)SARs [9]. Whilst much effort has been placed into the definition of domains, it remains a difficult and complex process to conceptualise as well as achieve [56]. The applicability domain of a model is highly context dependent and should not be generalised. Dimitrov et al. [57] recommended the domain be considered in four stages, namely physicochemical properties, structural characteristics of the molecules, toxicology and potential metabolites. Most current methods of QSAR applicability domain definition still do not account for all these stages adequately, thus the relevance of whether a molecule is stated to be in or out of the domain for some models should be verified manually, e.g. checking for the validity in the four domains noted by Dimitrov et al. [57], assessment of close analogues etc. With regard to the data considered in the Furuhama et al. [25] study, no effort was made by LJMU in developing the DL model to ensure that the test set compounds were in the applicability domain of the model. This oversight could be corrected to determine if a valid domain can be defined for the training set whether this affected the performance of the DL model for the test set.

3.4.4. *Models in isolation do not build a weight of evidence*

As noted in Section 3.3.3, good and successful modelling does not necessarily mean high statistical fit or performance. The output from a model, such as a prediction for a chemical without data, should be viewed as a piece of evidence to inform a decision to be made. Only in very limited circumstances does a prediction from a model imply that it has the same scientific value or standing as an experimental test, or that it may replace the need for a test [58]. The acceptance of predictions to adapt REACH requirements is one such instance, where strict criteria for the acceptance of predictions from QSARs are put in place. In addition, due to the limitations of QSARs – such as those outlined in this paper – there may be a preference for read-across (data gap filling from a similar substance with adequate data) with regard to more complex health effects [59].

As safety assessment moves towards a process whereby information is compiled and probabilities of safety judged, such as Next Generation Risk Assessment (NGRA) [60], it may be more illustrative to consider the prediction of mutagenicity through the combination or, or formation of a consensus of, predictions from different models (which may implicitly be based on different modelling approaches and represent different areas of chemical space). It should be noted, however, that multiple predictions from models with low predictive capacity is unlikely to be improve significantly on the outcome. Despite these caveats, an example of where the assessment of different modelling approaches has found acceptance is the ICH M7 assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk [61]. ICH M7 does not rely on a single prediction of mutagenicity, but requires two complimentary methods, one knowledge-based and the other a statistical (QSAR) model, to make an assessment. Expert judgement (or review) is also a crucial factor in making a prediction [62].

4. Opportunities for progress in improving how QSAR models predict toxicity

In addition to the identification of areas where current modelling techniques can be improved, the reflection of the poor performance of the LJMU DL model reported by Furuhama et al. [25] also allowed for the consideration of some new technologies, which are described in this section. Several of these technologies relate to better applying and interpreting machine learning models, Section 4.5 focuses on ensuring models are available for use. This is not a complete or comprehensive overview of the new technologies that are relevant to the *in silico* prediction of toxicity, but gives a selection based on their relevance to the models reported by Furuhama et al. [25].

4.1. *Federated learning (FL)*

With regard to increasing the availability of data for modelling, there are also further opportunities to store data and allow for greater compilation of data even when there may be concerns over confidentiality. In this regard, federated learning (FL) has recently been proposed for local training and storage of data at the source. FL is a privacy-preserving decentralised collaborative approach, which enables training a single model on data from multiple sources without jeopardising the privacy and security of those data [63]. More specifically, each client, e.g., private institutions, academic laboratories and business working in FL, has their own local training data that are not communicated to the other clients or the central server. Each client performs training on its local training data, and then the local updates are sent to the central server for aggregation to obtain a global model. The global model is then distributed back to the clients for the next iteration. Thus, FL is a way to build global models while preventing the dissemination of chemical data and avoiding data leaks or breaches compared to a centralised approach where a single central server stores all data locally and manages every step of the training procedure. Collaboration by FL can help to increase predictive performance, particularly when the client holds a small set of private data, and enables the models to capture patterns in datasets outside their own resources, leading to broader chemical space in the model [64]. At present, there are two real-world implementations of FL in drug discovery, namely MELLODY [65–68] and Effiris [69,70], with other ongoing research, for example, on molecular generation using FL graph neural networks, as summarised in Hanser [71]. All these examples have shown that the FL model outperforms each individual model regardless of the data size with an increase in the applicability domain.

4.2. *Data augmentation to improve data availability*

Whilst care should be taken in selecting data for modelling, it is recognised that larger data sets have been shown to improve the performance of DL models. The dataset provided for modelling as part of the challenge described by Furuhama et al. [25] is one of the largest available for QSAR modelling of a toxicological endpoint. For many endpoints, much smaller datasets have to be relied upon. Improving the size of data sets for use in machine learning, especially DL, can be achieved through employing various strategies such as data augmentation (e.g., generating multiple and different instances of the same molecule to be used as input), multi-stage training (e.g., transfer learning, active learning, reinforcement learning), and context-enriched training (e.g., additional context is provided to the model through different inputs), as summarised in van Tilborg et al. [72]. These low-data learning approaches are still in their infancy and require exploration for toxicity prediction tasks.

4.3. *Probabilistic modelling*

The issue of the variability of data, and the difficulty this creates in modelling, are described in Section 3.1.2. One solution to this could be the greater use of modelling approaches that provide a probability of an activity. Probabilistic programming of an endpoint allows for the capture of data variability and quantification of uncertainty about the predictions in the form of a probability distribution. DL methods have the ability to model non-linearity. However, they are not appropriate for small datasets, which can lead to overfitting, neither can they quantify the uncertainty of experimentally measured or predicted values. Bayesian neural networks can address these pitfalls. For example, Bayesian parametric approaches were proposed for modelling the severity of drug-induced liver injury of 237 labelled compounds in preclinical safety assessment using a hierarchical prior instead of hyperparameter optimisation [73]. Bayesian neural networks were also shown to be useful in NGRA by modelling the biochemical activity of 20 molecular initiating events and how the statistical metrics such as standard deviation and credible intervals can inform decision-making and risk quantification [74].

4.4. Improving the interpretability of complex models

The importance of making models interpretable is well-established (see Section 3.1.3). There are distinct advantages to model interpretability, especially with regard to validation and acceptance of predictions. However, even for a simple model, this is a skilled task requiring not only the explanation of the model but also it being put in context of the toxicological mechanism of action. For complex models, such as the LJMU DL model, the problem of interpretation is multiplied, due, in part, to the potential "black box" nature of the model but also the large number of descriptors and non-linear nature of the relationship. There are some solutions, for instance recently, the focus of QSAR models has shifted to the use of methods and strategies that help to understand model results such as assessment of feature importance, e.g., SHapley Additive exPlanations (SHAP) [75]. Other methods explore how representations of compounds are transformed in hidden layers, which is how neural networks learn. For example, Walter et al. [76] developed an interpretation method to identify and explain possible associations between the activation of hidden neurons and known toxicophores for mutagenicity by validating the approach with the structural alerts from the Derek Nexus expert system.

4.5. Making models and their underlying data FAIR

Users of model should be able to find, reproduce and implement the model easily. With regard to the models reported in Furuhama et al. [25], it is not immediately clear how this could be achieved or all models. Thus, whilst not impacting on the performance of models, the adherence of the models, and the data on which they are developed, to the Findability, Accessibility, Interoperability, and Reuse (FAIR) principles of digital assets [77] will make them a truly useful and useable resource. Cronin et al. [78] revised the FAIR principles to make them applicable to *in silico* models for toxicity, and QSARs in particular. Whilst no formal analysis has been undertaken, none of the models described by Furuhama et al. [25] are likely to be compliant with the FAIR principles – although that was never the intention of the study. Notably, lacking from the models are an unique identifier, full description of the model and clear description of the meta data. The advantage of making models FAIR is that this will promote not only use of a model, but also that it is used consistently by any researcher. Likewise, since the biological data are not fully published in Furuhama et al. [25] (due to reasons of commercial sensitivity), they may not be considered FAIR and the models and performance statistics cannot be reproduced, especially with regard to the FAIR principles established by Wilkinson et al. [77]. Thus, not having access to the data and model will hinder the capability to recreate or recalibrate a model and thus will undoubtedly affect reliability and use, and may be a cause of poor model performance i.e., the same model is not recreated.

5. Conclusions on assessing and improving QSAR model performance

Assessment and evaluation of QSAR models for toxicity prediction is a worthwhile exercise. Few truly blinded trials have been attempted in the past, with the study from Furuhama et al. [25] being one of more recent. The output from Furuhama et al. [25] shows a wide variety in the overall performance of models, as measured by predictivity of a large external test set. It is gratifying that so many models were submitted to the DGM/NIHS Second Challenge Project, with a diverse selection of modelling approaches.

Our assessment of our own models is that there is no single reason for the poor performance of a model, but it is likely to be due to a number of factors that need to be put in the context of the endpoint being modelled. Performance is related to a number of issues which should be considered not only by the model developer, but also the user of a model and any third party called upon to evaluate the predictions from a model. These include, but are not limited to:

- Underlying data quality, consistency and relevance.
- Appropriate descriptors relating to the endpoint and mechanism of action.
- Appropriate selection of model in terms of the structure (i.e., complexity) of the model and number of descriptors utilised.
- Addressing metabolism adequately in the modelling process.
- Assessing (quantitatively, where possible) the uncertainties within a model.
- Ensuring predictions are within the applicability domain of the model.

With regard to the LJMU DL model presented in Furuhama et al. [25] and the comparison of its performance with other models, all approaches used the same data so that does not explain poor performance. Therefore, it is probable that poor performance was a result of descriptors in the model not addressing reactivity adequately, an inability of the model to parameterise metabolism and model structures not being appropriate to model the endpoint.

The use of the predictions from the models will depend on the purpose of the assessment. As the move towards "artificial intelligence" grows, it is worthwhile for model developers and users to step back to consider why and when a model will work, and the probability of success. A number of new technologies will assist in the better development of models for toxicity prediction. As a final consideration, we call on all model developers in the Second Ames/QSAR International Challenge Project to reflect critically on their models and associated performance. The opportunity within this exercise is not only to stimulate model development, but also to identify when a model did and did not perform to expectations and, importantly, the reasons why this was the case.

CRediT authorship contribution statement

Mark T.D. Cronin: Conceptualization, Writing – original draft. **Homa Basiri:** Writing – review & editing. **Georgios Chrysochoou:** Writing – review & editing. **Steven J. Enoch:** Writing – review & editing. **James W. Firman:** Writing – review & editing. **Nicoleta Spînu:** Writing – review & editing. **Judith C. Madden:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the project RISK-HUNT3R: RISK assessment of chemicals integrating HUman centric Next generation Testing strategies promoting the 3Rs. RISK-HUNT3R has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 964537 and is part of the ASPIS cluster.

Disclaimer

This work and the opinions expressed in this article reflect only the authors' views. The European Commission is not responsible for any use that may be made of the information it contains.

Data availability

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

References

[1] J.C. Madden, S.J. Enoch, A. Paini, M.T.D. Cronin, A review of *in silico* tools as alternatives to animal testing: principles, resources and applications, *Altern. Lab. Anim.* 48 (2020) 146–172, https://doi.org/10.1177/0261192920965977.

[2] S.J. Enoch, M.T.D. Cronin, T.W. Schultz, J.C. Madden, An evaluation of global QSAR models for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*, *Chemosphere* 71 (2008) 1225–1232, https://doi.org/10.1016/j.chemosphere.2007.12.011.

[3] S.J. Belfield, S.J. Enoch, J.W. Firman, J.C. Madden, T.W. Schultz, M.T.D. Cronin, Determination of "fitness-for-purpose" of quantitative structure-activity relationship (QSAR) models to predict (eco-)toxicological endpoints for regulatory use, *Regul. Toxicol. Pharmacol.* 123 (2021) 104956, https://doi.org/10.1016/j.yrtph.2021.104956.

[4] N. Kleinstreuer, T. Hartung, Artificial intelligence (AI) - it's the end of the tox as we know it (and I feel fine), *Arch. Toxicol.* 98 (2024) 735–754, https://doi.org/10.1007/s00204-023-03666-2.

[5] M.T.D. Cronin, J.S. Jaworska, J.D. Walker, M.H.I. Comber, C.D. Watts, A.P. Worth, Use of QSARs in international decision-making frameworks to predict health effects of chemical substances, *Environ. Health Perspect.* 111 (2003) 1391–1401, https://doi.org/10.1289/ehp.5760.

[6] M.T.D. Cronin, J.D. Walker, J.S. Jaworska, M.H.I. Comber, C.D. Watts, A.P. Worth, Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances, *Environ. Health Perspect.* 111 (2003) 1376–1390, https://doi.org/10.1289/ehp.5759.

[7] G.J. Myatt, E. Ahlberg, Y. Akahori, D. Allen, A. Amberg, L.T. Anger, A. Aptula, S. Auerbach, L. Beilke, P. Bellion, R. Benigni, J. Bercu, E.D. Booth, D. Bower, A. Brigo, N. Burden, Z. Cammerer, M.T.D. Cronin, K.P. Cross, L. Custer, M. Dettwiler, K. Dobo, K.A. Ford, M.C. Fortin, S.E. Gad-McDonald, N. Gellatly, V. Gervais, K.P. Glover, S. Glowienke, J. Van Gompel, S. Gutsell, B. Hardy, J. S. Harvey, J. Hillegass, M. Honma, J.H. Hsieh, C.W. Hsu, K. Hughes, C. Johnson, R. Jolly, D. Jones, R. Kemper, M.O. Kenyon, M.T. Kim, N.L. Kruhlak, S.A. Kulkarni, K. Kümmerer, P. Leavitt, B. Majer, S. Masten, S. Miller, J. Moser, M. Mumtaz, W. Muster, L. Neilson, T.I. Oprea, G. Patlewicz, A. Paulino, E. Lo Piparo, M. Powley, D.P. Quigley, M.V. Reddy, A.N. Richarz, P. Ruiz, B. Schilter, R. Serafimova, W. Simpson, L. Stavitskaya, R. Stidl, D. Suarez-Rodriguez, D. T. Szabo, A. Teasdale, A. Trejo-Martin, J.P. Valentin, A. Vuorinen, B.A. Wall, P. Watts, A.T. White, J. Wichard, K.L. Witt, A. Woolley, D. Woolley, C. Zwickl, C. Hasselgren, *In silico* toxicology protocols, *Regul. Toxicol. Pharmacol.* 96 (2018) 1–17, https://doi.org/10.1016/j.yrtph.2018.04.014.

[8] A. Gissi, O. Tcheremenskaia, C. Bossa, C.L. Battistelli, P. Browne, The OECD (Q)SAR Assessment Framework: A tool for increasing regulatory uptake of computational approaches, *Comput. Toxicol.* 31 (2024) 100326, https://doi.org/10.1016/j.comtox.2024.100326.

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

[10] ECHA (European Chemicals Agency), 2017. *Read-Aacross Assessment Framework (RAAF).* https://echa.europa.eu/documents/10162/13628/raaf_en.pdf.

[11] M.T.D. Cronin, A.N. Richarz, T.W. Schultz, Identification and description of the uncertainty, variability, bias and influence in quantitative structure-activity relationships (QSARs) for toxicity prediction, *Regul. Toxicol. Pharmacol.* 106 (2019) 90–104, https://doi.org/10.1016/j.yrtph.2019.04.007.

[12] ECHA (European Chemicals Agency), 2016. *Practical guide. How to use and report (Q)SARs*. https://echa.europa.eu/documents/10162/13655/pg_report_qsars_en.pdf.

[13] OECD (Organisation for Economic Cooperation and Development) (2023) *(Q)SAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure Activity Relationship models and predictions*. Organisation for Economic Co-operation and Development. https://www.oecd-ilibrary.org/environment/q-sar-assessment-framework-guidance-for-the-regulatory-assessment-of-quantitative-structure-activity-relationship-models-and-predictions_d96118f6-en.

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

[15] J.C. Dearden, M.T. Cronin, K.L. Kaiser, How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR), *SAR QSAR Environ. Res.* 20 (2009) 241–266, https://doi.org/10.1080/10629360902949567.

[16] S.J. Belfield, M.T.D. Cronin, S.J. Enoch, J.W. Firman, Guidance for good practice in the application of machine learning in development of toxicological quantitative structure-activity relationships (QSARs), *PLoS One* 18 (5) (2023), https://doi.org/10.1371/journal.pone.0282924 e0282924.

[17] L.D. Burgoon, F.M. Kluxen, A. Hüser, M. Frericks, The database makes the poison: How the selection of datasets in QSAR models impacts toxicant prediction of higher tier endpoints, *Regul. Toxicol. Pharmacol.* 151 (2024) 105663, https://doi.org/10.1016/j.yrtph.2024.105663.

[18] Benigni R (1997) The first US National Toxicology Program exercise on the prediction of rodent carcinogenicity: definitive results. *Mutat. Res.* 387: 35-45. DOR: 10.1016/s1383-5742(97)00021-5.

[19] R. Benigni, R. Zito, The second National Toxicology Program comparative exercise on the prediction of rodent carcinogenicity: definitive results, *Mutat. Res.* 566 (2004) 49–63, https://doi.org/10.1016/s1383-5742(03)00051-6.

[20] A. Worth, S. Lapenna, E. Lo Piparo, A. Mostrag-Szlichtyng, R. Serafimova, The Applicability of Software Tools for Genotoxicity and Carcinogenicity Prediction: Case Studies relevant to the Assessment of Pesticides, European Commission Joint Research Centre, Ispra, Italy. (2010), https://doi.org/10.2788/61485.

[21] R. Huang, M. Xia, D.-T. Nguyen, T. Zhao, S. Sakamuru, J. Zhao, S.A. Shahane, A. Rossoshek, A. Simeonov, Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs, *Front. Environ. Sci.* 3 (2016) 85, https://doi.org/10.3389/fenvs.2015.00085.

[22] K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek, A. Zakharov, A. Worth, A.M. Richard, C.M. Grulke, D. Trisciuzzi, D. Fourches, D. Horvath, E. Benfenati, E. Muratov, E.B. Wedebye, F. Grisoni, G.F. Mangiatordi, G.M. Incisivo, H. Hong, H.W. Ng, I.V. Tetko, I. Balabin, J. Kancherla, J. Shen, J. Burton, M. Nicklaus, M. Cassotti, N.G. Nikolov, O. Nicolotti, P.L. Andersson, Q. Zang, R. Politi, R.D. Beger, R. Todeschini, R. Huang, S. Farag, S.A. Rosenberg, S. Slavov, X. Hu, R.S. Judson, CERAPP: Collaborative Estrogen Receptor Activity Prediction Project, *Environ. Health Perspect.* 124 (2016) 1023–1033, https://doi.org/10.1289/ehp.1510267.

[23] K. Mansouri, N. Kleinstreuer, A.M. Abdelaziz, D. Alberga, V.M. Alves, P. L. Andersson, C.H. Andrade, F. Bai, I. Balabin, D. Ballabio, E. Benfenati, B. Bhhatarai, S. Boyer, J. Chen, V. Consonni, S. Farag, D. Fourches, A.T. García-Sosa, P. Gramatica, F. Grisoni, C.M. Grulke, H. Hong, D. Horvath, X. Hu, R. Huang, N. Jeliazkova, J. Li, X. Li, H. Liu, S. Manganelli, G.F. Mangiatordi, U. Maran, G. Marcou, T. Martin, E. Muratov, D.T. Nguyen, O. Nicolotti, N.G. Nikolov, U. Norinder, E. Papa, M. Petitjean, G. Piir, P. Pogodin, V. Poroikov, X. Qiao, A. M. Richard, A. Roncaglioni, P. Ruiz, C. Rupakheti, S. Sakkiah, A. Sangion, K. W. Schramm, C. Selvaraj, I. Shah, S. Sild, L. Sun, O. Taboureau, Y. Tang, I.V. Tetko, R. Todeschini, W. Tong, D. Trisciuzzi, A. Tropsha, G. Van Den Driessche, A. Varnek, Z. Wang, E.B. Wedebye, A.J. Williams, H. Xie, A.V. Zakharov, Z. Zheng, R. S. Judson, CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity, *Environ. Health Perspect.* 128 (2020) 27002, https://doi.org/10.1289/EHP5580.

[24] K. Mansouri, A.L. Karmaus, J. Fitzpatrick, G. Patlewicz, P. Pradeep, D. Alberga, N. Alepee, T.E.H. Allen, D. Allen, V.M. Alves, C.H. Andrade, T.R. Auernhammer, D. Ballabio, S. Bell, E. Benfenati, S. Bhattacharya, J.V. Bastos, S. Boyd, J.B. Brown, S.J. Capuzzi, Y. Chushak, H. Ciallella, A.M. Clark, V. Consonni, P.R. Daga, S. Ekins, S. Farag, M. Fedorov, D. Fourches, D. Gadaleta, F. Gao, J.M. Gearhart, G. Goh, J. M. Goodman, F. Grisoni, C.M. Grulke, T. Hartung, M. Hirn, P. Karpov, A. Korotcov, G.J. Lavado, M. Lawless, X. Li, T. Luechtefeld, F. Lunghini, G.F. Mangiatordi, G. Marcou, D. Marsh, T. Martin, A. Mauri, E.N. Muratov, G.J. Myatt, D.T. Nguyen, O. Nicolotti, R. Note, P. Pande, A.K. Parks, T. Peryea, A.H. Polash, R. Rallo, A. Roncaglioni, C. Rowlands, P. Ruiz, D.P. Russo, A. Sayed, R. Sayre, T. Sheils, C. Siegel, A.C. Silva, A. Simeonov, S. Sosnin, N. Southall, J. Strickland, Y. Tang, B. Teppen, I.V. Tetko, D. Thomas, V. Tkachenko, R. Todeschini, C. Toma, I. Tripodi, D. Trisciuzzi, A. Tropsha, A. Varnek, K. Vukovic, Z. Wang, L. Wang, K.M. Waters, A. J. Wedlake, S.J. Wijeyesakere, D. Wilson, Z. Xiao, H. Yang, G. Zahoranszky-Kohalmi, A.V. Zakharov, F.F. Zhang, Z. Zhang, T. Zhao, H. Zhu, K.M. Zorn, W. Casey, N.C. Kleinstreuer, CATMoS: Collaborative Acute Toxicity Modeling Suite, *Environ. Health Perspect.* 129 (2021) 47013, https://doi.org/10.1289/EHP8495.

[25] A. Furuhama, A. Kitazawa, J. Yao, C.E. Matos Dos Santos, J. Rathman, C. Yang, J. V. Ribeiro, K. Cross, G. Myatt, G. Raitano, E. Benfenati, N. Jeliazkova, R. Saiakhov, S. Chakravarti, R.S. Foster, C. Bossa, C.L. Battistelli, R. Benigni, T. Sawada, H. Wasada, T. Hashimoto, M. Wu, R. Barzilay, P.R. Daga, R.D. Clark, J. Mestres, A. Montero, E. Gregori-Puigjané, P. Petkov, H. Ivanova, O. Mekenyan, S. Matthews, D. Guan, J. Spicer, R. Lui, Y. Uesawa, K. Kurosaki, Y. Matsuzaka, S. Sasaki, M.T. D. Cronin, S.J. Belfield, J.W. Firman, N. Spînu, M. Qiu, J.M. Keca, G. Gini, T. Li, W. Tong, H. Hong, Z. Liu, Y. Igarashi, H. Yamada, K.I. Sugiyama, M. Honma, Evaluation of QSAR models for predicting mutagenicity: outcome of the Second Ames/QSAR international challenge project, *SAR QSAR Environ. Res.* 34 (2023) 983–1001, https://doi.org/10.1080/1062936X.2023.2284902.

[26] M. Honma, A. Kitazawa, A. Cayley, R.V. Williams, C. Barber, T. Hanser, R. Saiakhov, S. Chakravarti, G.J. Myatt, K.P. Cross, E. Benfenati, G. Raitano, O. Mekenyan, P. Petkov, C. Bossa, R. Benigni, C.L. Battistelli, A. Giuliani, O. Tcheremenskaia, C. DeMeo, U. Norinder, H. Koga, C. Jose, N. Jeliazkova, N. Kochev, V. Paskaleva, C. Yang, P.R. Daga, R.D. Clark, J. Rathman, Improvement of quantitative structure-activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project, *Mutagenesis* 34 (2019) 3–16, https://doi.org/10.1093/mutage/gey031.

[27] Y. Uesawa, Progress in predicting Ames test outcomes from chemical structures: An in-depth re-evaluation of models from the 1st and 2nd Ames/QSAR International Challenge Projects, *Int. J. Mol. Sci.* 25 (2024) 1373, https://doi.org/10.3390/ijms25031373.

[28] C.W. Yap, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, *J. Comput. Chem.* 32 (2011) 1466–1474, https://doi.org/10.1002/jcc.21707.

[29] PubChem Substructure Fingerprint (2009) National Center for Biotechnology Information. PubChem Subgraph Fingerprint [Internet]. Bethesda, MD: National Institutes of Health. Available from: https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf. Accessed on 25 November 2024.

[30] C. Yang, A. Tarkhov, J. Marusczyk, B. Bienfait, J. Gasteiger, T. Kleinoeder, T. Magdziarz, O. Sacher, C.H. Schwab, J. Schwoebel, L. Terfloth, K. Arvidson, A. Richard, A. Worth, J. Rathman, New publicly available chemical query

language, CSRML, to support chemotype representations for application to data mining and modelling, *J. Chem. Inf. Mod.* 55 (2015) 510–528, https://doi.org/10.1021/ci500667v Available from: github.com/mn-am/toxprint. Accessed on 25 November 2024.

[31] Cheng H-T, Koc L, Harmsen J, Shaked T, Chandra T, Aradhye H, Anderson G, Corrado G, Chai W, Ispir M, Anil R, Haque Z, Hong L, Jain V, Liu X, Shah H (2016) Wide & Deep Learning for Recommender Systems. *arXiv:* 1606.07792. DOI: arxiv.org/abs/1606.07792.

[32] Wichard JD (2017) *In silico* prediction of genotoxicity. *Fd Chem. Toxicol.* 106, Part B: 595-599. DOI : 10.1016/j.fct.2016.12.013.

[33] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, Mol. Inf. 29 (2010) 476–488, https://doi.org/10.1002/minf.201000061.

[34] J. Achar, M.T.D. Cronin, J.W. Firman, G. Öberg, A problem formulation framework for the application of in silico toxicology methods in chemical risk assessment, *Arch. Toxicol.* 98 (2024) 1727–1740, https://doi.org/10.1007/s00204-024-03721-6.

[35] V.M. Alves, S.S. Auerbach, N. Kleinstreuer, J.P. Rooney, E.N. Muratov, I. Rusyn, A. Tropsha, C. Schmitt, Curated data in - trustworthy *in silico* models out: The impact of data quality on the reliability of artificial intelligence models as alternatives to animal testing, *Altern. Lab. Anim.* 49 (2021) 73–82, https://doi.org/10.1177/02611929211029635.

[36] E. Zeiger, C.A. Mitchell, S. Pfuhler, Y. Liao, K.L. Witt, Within-laboratory reproducibility of Ames test results: Are repeat tests necessary? *Environ. Mol. Mutagen.* 65 (2024) 116–120, https://doi.org/10.1002/em.22597.

[37] E. Zeiger, The test that changed the world: The Ames test and the regulation of chemicals, *Mutat. Res. Genet. Toxicol. Environ. Mutagen.* 841 (2019) 43–48, https://doi.org/10.1016/j.mrgentox.2019.05.007.

[38] K. Mortelmans, E. Zeiger, The Ames *Salmonella*/microsome mutagenicity assay. *Mut. Res./Fund, Mol. Mech. Mutag.* 455 (2000) 29–60, https://doi.org/10.1016/S0027-5107(00)00064-6.

[39] C. Hasselgren, E. Ahlberg, Y. Akahori, A. Amberg, L.T. Anger, F. Atienzar, S. Auerbach, L. Beilke, P. Bellion, R. Benigni, J. Bercu, E.D. Booth, D. Bower, A. Brigo, Z. Cammerer, M.T.D. Cronin, I. Crooks, K.P. Cross, L. Custer, K. Dobo, T. Doktorova, D. Faulkner, K.A. Ford, M.C. Fortin, M. Frericks, S.E. Gad-McDonald, N. Gellatly, H. Gerets, V. Gervais, S. Glowienke, J. Van Gompel, J.S. Harvey, J. Hillegass, M. Honma, J.H. Hsieh, C.W. Hsu, T.S. Barton-Maclaren, C. Johnson, R. Jolly, D. Jones, R. Kemper, M.O. Kenyon, N.L. Kruhlak, S.A. Kulkarni, K. Kümmerer, P. Leavitt, S. Masten, S. Miller, C. Moudgal, W. Muster, A. Paulino, E. Lo Piparo, M. Powley, D.P. Quigley, M.V. Reddy, A.N. Richarz, B. Schilter, R. D. Snyder, L. Stavitskaya, R. Stidl, D.T. Szabo, A. Teasdale, R.R. Tice, A. Trejo-Martin, A. Vuorinen, B.A. Wall, P. Watts, A.T. White, J. Wichard, K.L. Witt, A. Woolley, D. Woolley, C. Zwickl, G.J. Myatt, Genetic toxicology *in silico* protocol, *Regul. Toxicol. Pharmacol.* 107 (2019) 104403, https://doi.org/10.1016/j.yrtph.2019.104403.

[40] J.C. Dearden, Physico-chemical descriptors, in: W. Karcher, J. Devillers (Eds.), *Practical Applications of Quantitative Structure-activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, Springer, Dordrecht, The Netherlands, 1990, pp. 25–59.

[41] R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics*, 2, Wiley-VCH, Weinheim, 2009.

[42] C.-X. Lin, Y. Guan, H.-D. Li, Artificial intelligence approaches for molecular representation in drug response prediction, *Curr. Opin. Struct. Biol.* 84 (2024) 102747, https://doi.org/10.1016/j.sbi.2023.102747.

[43] M.T.D. Cronin, A.-N. Richarz, Relationship between Adverse Outcome Pathways and chemistry-based *in silico* models to predict toxicity, *Appl. in Vitro Toxicol.* 3 (2017) 286–297, https://doi.org/10.1089/aivt.2017.0021.

[44] M.T.D. Cronin, J.C. Dearden, J.C. Duffy, R. Edwards, N. Manga, A.P. Worth, A. D. Worgan, The importance of hydrophobicity and electrophilicity descriptors in mechanistically-based QSARs for toxicological endpoints, *SAR QSAR Environ. Res.* 13 (2002) 167–176, https://doi.org/10.1080/10629360290002316.

[45] A.K. Debnath, G. Debnath, A.J. Shusterman, C. Hansch, A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in *Salmonella typhimurium* TA98 and TA100, *Environ. Mol. Mutagen.* 19 (1992) 37–52, https://doi.org/10.1002/em.2850190107.

[46] A.K. Debnath, R.L. Lopez de Compadre, A.J. Shusterman, C. Hansch, Quantitative structure-activity relationship investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 2. Mutagenicity of aromatic and heteroaromatic nitro compounds in *Salmonella typhimurium* TA100, *Environ Mol Mutagen.* 19 (1992) 53–70, https://doi.org/10.1002/em.2850190108.

[47] M.T.D. Cronin, N. Manga, J.R. Seward, G.D. Sinks, T.W. Schultz, Parametrization of electrophilicity for the prediction of the toxicity of aromatic compounds, *Chem. Res. Toxicol.* 14 (2001) 1498–1505, https://doi.org/10.1021/tx015502k.

[48] T.W. Schultz, R.E. Carlson, M.T.D. Cronin, J.L. Hermens, R. Johnson, P.J. O'Brien, D.W. Roberts, A. Siraki, K.B. Wallace, G.D. Veith, A conceptual framework for predicting the toxicity of reactive chemicals: modeling soft electrophilicity, *SAR QSAR Environ. Res.* 17 (2006) 413–428, https://doi.org/10.1080/10629360600884371.

[49] G. Schüürmann, Quantum Chemical Descriptors in Structure-Activity Relationships – Calculation, Interpretation and Comparison of Methods, in: M.T.D. Cronin, D. J. Livingstone (Eds.), *Predicting Chemical Toxicity and Fate*, CRC Press, Boca Raton FL, USA, 2004, pp. 85–150.

[50] T.A. Baillie, A.E. Rettie, Role of biotransformation in drug-induced toxicity: influence of intra- and inter-species differences in drug metabolism, *Drug Metab. Pharmacokinet.* 26 (2011) 15–29, https://doi.org/10.2133/dmpk.dmpk-10-rv-089.

[51] A.S. Kalgutkar, I. Gardner, R.S. Obach, C.L. Shaffer, E. Callegari, K.R. Henne, A. E. Mutlib, D.K. Dalvie, J.S. Lee, Y. Nakai, J.P. O'Donnell, J. Boer, S.P. Harriman, A comprehensive listing of bioactivation pathways of organic functional groups, *Curr. Drug Metab.* 6 (2005) 161–225, https://doi.org/10.2174/1389200054021799.

[52] S.J. Enoch, M.T.D. Cronin, A review of the electrophilic reaction chemistry involved in covalent DNA binding, *Crit. Rev. Toxicol.* 40 (2010) 728–748, https://doi.org/10.3109/10408444.2010.494175.

[53] S.J. Enoch, M.T.D. Cronin, C.M. Ellison, The use of a chemistry-based profiler for covalent DNA binding in the development of chemical categories for read-across for genotoxicity, *Altern. Lab. Anim.* 39 (2011) 131–145, https://doi.org/10.1177/026119291103900206.

[54] M.T.D. Cronin, The role of hydrophobicity in toxicity prediction. *Curr. Comput. – Aid, Drug Des.* 2 (2006) 405–413, https://doi.org/10.2174/157340906778992346.

[55] R.P. Vivek-Ananth, A.K. Sahoo, S.P. Baskaran, J. Ravichandran, A. Samal, Identification of activity cliffs in structure-activity landscape of androgen receptor binding chemicals, *Sci. Total Environ.* 873 (2023) 162263, https://doi.org/10.1016/j.scitotenv.2023.162263.

[56] T.I. Netzeva, A. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J. S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T. Schultz, D.W. Stanton, J. J. van de Sandt, W. Tong, G. Veith, C. Yang, Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52, *Altern. Lab. Anim.* 33 (2005) 155–173, https://doi.org/10.1177/026119290503300209.

[57] S. Dimitrov, G. Dimitrova, T. Pavlov, N. Dimitrova, G. Patlewicz, J. Niemela, O. Mekenyan, A stepwise approach for defining the applicability domain of SAR and QSAR models, *J. Chem. Inf. Model.* 45 (2005) 839–849, https://doi.org/10.1021/ci0500381.

[58] K. Taylor, Ten years of REACH - An animal protection perspective, *Altern. Lab. Anim.* 46 (2018) 347–373, https://doi.org/10.1177/026119291804600610.

[59] C. Pestana, S.J. Enoch, J.W. Firman, J.C. Madden, N. Spînu, M.T.D. Cronin, A strategy to define applicability domains for read-across, *Comput. Toxicol.* 22 (2022) 100220, https://doi.org/10.1016/j.comtox.2022.100220.

[60] M.P. Dent, E. Vaillancourt, R.S. Thomas, P.L. Carmichael, G. Ouedraogo, H. Kojima, J. Barroso, J. Ansell, T.S. Barton-Maclaren, S.H. Bennekou, K. Boekelheide, J. Ezendam, J. Field, S. Fitzpatrick, M. Hatao, R. Kreiling, M. Lorencini, C. Mahony, B. Montemayor, R. Mazaro-Costa, J. Oliveira, V. Rogiers, D. Smegal, R. Taalman, Y. Tokura, R. Verma, C. Willett, C. Yang, Paving the way for application of next generation risk assessment to safety decision-making for cosmetic ingredients, *Regul. Toxicol. Pharmacol.* 125 (2021) 105026, https://doi.org/10.1016/j.yrtph.2021.105026.

[61] A. Amberg, L. Beilke, J. Bercu, D. Bower, A. Brigo, K.P. Cross, L. Custer, K. Dobo, E. Dowdy, K.A. Ford, S. Glowienke, J. Van Gompel, J. Harvey, C. Hasselgren, M. Honma, R. Jolly, R. Kemper, M. Kenyon, N. Kruhlak, P. Leavitt, S. Miller, W. Muster, J. Nicolette, A. Plaper, M. Powley, D.P. Quigley, M.V. Reddy, H. P. Spirkl, L. Stavitskaya, A. Teasdale, S. Weiner, D.S. Welch, A. White, J. Wichard, G.J. Myatt, Principles and procedures for implementation of ICH M7 recommended (Q)SAR analyses, *Regul. Toxicol. Pharmacol.* 77 (2016) 13–24, https://doi.org/10.1016/j.yrtph.2016.02.004.

[62] C. Barber, A. Amberg, L. Custer, K.L. Dobo, S. Glowienke, J. Van Gompel, S. Gutsell, J. Harvey, M. Honma, M.O. Kenyon, N. Kruhlak, W. Muster, L. Stavitskaya, A. Teasdale, J. Vessey, J. Wichard, Establishing best practise in the application of expert review of mutagenicity under ICH M7, *Regul. Toxicol. Pharmacol.* 73 (2015) 367–377, https://doi.org/10.1016/j.yrtph.2015.07.018.

[63] P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, M. Bennis, A.N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R.G.L. D'Oliveira, H. Eichner, S. El Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P.B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, H. Qi, D. Ramage, R. Raskar, M. Raykova, D. Song, W. Song, S.U. Stich, Z. Sun, A. Theertha Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F.X. Yu, H. Yu, S. Zhao, Advances and open problems in federated learning, Found. Trends Mach. Learn. 14 (2021) 1–210, https://doi.org/10.1561/2200000083.

[64] A. Smajic, M. Grandits, G.F. Ecker, Privacy-preserving techniques for decentralized and secure machine learning in drug discovery, *Drug Discov. Today* 28 (2023) 103820, https://doi.org/10.1016/j.drudis.2023.103820.

[65] L. Humbeck, T. Morawietz, N. Sturm, A. Zalewski, S. Harnqvist, W. Heyndrickx, M. Holmes, B. Beck, Don't overweight weights: Evaluation of weighting strategies for multi-task bioactivity classification models, *Molecules* 26 (2021) 6959, https://doi.org/10.3390/molecules26226959.

[66] J. Simm, L. Humbeck, A. Zalewski, N. Sturm, W. Heyndrickx, Y. Moreau, B. Beck, A. Schuffenhauer, Splitting chemical structure data sets for federated privacy-preserving machine learning, *J. Cheminform.* 13 (2021) 96, https://doi.org/10.1186/s13321-021-00576-2.

[67] W. Heyndrickx, A. Arany, J. Simm, A. Pentina, N. Sturm, L. Humbeck, L. Mervin, A. Zalewski, M. Oldenhof, P. Schmidtke, L. Friedrich, R. Loeb, A. Afanasyeva, Y. Moreau, H. Ceulemans, Conformal efficiency as a metric for comparative model assessment befitting federated learning, *Artific. Intell. Life Sci.* 3 (2023) 100070, https://doi.org/10.1016/j.ailsci.2023.100070.

[68] W. Heyndrickx, L. Mervin, T. Morawietz, N. Sturm, L. Friedrich, A. Zalewski, A. Pentina, L. Humbeck, M. Oldenhof, R. Niwayama, P. Schmidtke, N. Fechner, J. Simm, A. Arany, N. Drizard, R. Jabal, A. Afanasyeva, R. Loeb, S. Verma, S. Harnqvist, M. Holmes, B. Pejo, M. Telenczuk, N. Holway, A. Dieckmann, N. Rieke, F. Zumsande, D.A. Clevert, M. Krug, C. Luscombe, D. Green, P. Ertl,

P. Antal, D. Marcus, N. Do Huu, H. Fuji, S. Pickett, G. Acs, E. Boniface, B. Beck, Y. Sun, A. Gohier, F. Rippmann, O. Engkvist, A.H. Göller, Y. Moreau, M.N. Galtier, A. Schuffenhauer, H. Ceulemans, MELLODDY: Cross-pharma Federated Learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information, *J. Chem. Inf. Model.* 64 (2024) 2331–2344, https://doi.org/10.1021/acs.jcim.3c00799.

[69] Hanser T, Bastogne D, Basu A, Davies R, Delaunois A, Fowkes A, Harding A, Johnston LA, Korlowski C, Kotsampasakou E, Plante J, Rosenbrier-Ribeiro L, Rowell P, Sabnis Y, Sartini A, Sibony A, Werner AL, White A, Yukawa T (2022) Using privacy-preserving federated learning to enable pre-competitive cross-industry knowledge sharing and improve QSAR models. Available at: https://www.lhasalimited.org/publications/using-privacy-preserving-federated-learning-to-enable-pre-competitive-cross-industry-knowledge-sharing-and-improve-qsar-models/. Accessed 29 September 2024.

[70] D. Bassani, A. Brigo, A. Andrews-Morger, Federated Learning in computational toxicology: An industrial perspective on the Effiris Hackathon, *Chem. Res. Toxicol.* 36 (2023) 1503–1517, https://doi.org/10.1021/acs.chemrestox.3c00137.

[71] T. Hanser, Federated learning for molecular discovery, *Curr. Opin. Struct. Biol.* 79 (2023) 102545, https://doi.org/10.1016/j.sbi.2023.102545.

[72] D. van Tilborg, H. Brinkmann, E. Criscuolo, L. Rossen, R. Özçelik, F. Grisoni, Deep learning for low-data drug discovery: Hurdles and opportunities, *Curr. Opin. Struct. Biol.* 86 (2024) 102818, https://doi.org/10.1016/j.sbi.2024.102818.

[73] E. Semenova, D.P. Williams, A.M. Afzal, S.E. Lazic, A Bayesian neural network for toxicity prediction, *Comput. Toxicol.* 16 (2020) 100133, https://doi.org/10.1016/j.comtox.2020.100133.

[74] T.E.H. Allen, A.M. Middleton, J.M. Goodman, P.J. Russell, P. Kukic, S. Gutsell, Towards quantifying the uncertainty in *in silico* predictions using Bayesian learning, *Comput. Toxicol.* 23 (2022) 100228, https://doi.org/10.1016/j.comtox.2022.100228.

[75] R. Rodríguez-Pérez, J. Bajorath, Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions, *J. Comput. Aided Mol. Des.* 34 (2020) 1013–1026, https://doi.org/10.1007/s10822-020-00314-0.

[76] M. Walter, S.J. Webb, V.J. Gillet, Interpreting neural network models for toxicity prediction by extracting learned chemical features, *J. Chem. Inf. Model.* 64 (2024) 3670–3688, https://doi.org/10.1021/acs.jcim.4c00127.

[77] M.D. Wilkinson, M. Dumontier, I. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M. E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, *Sci. Data* 3 (2016) 160018, https://doi.org/10.1038/sdata.2016.18.

[78] M.T.D. Cronin, S.J. Belfield, K.A. Briggs, S.J. Enoch, J.W. Firman, M. Frericks, C. Garrard, P.H. Maccallum, J.C. Madden, M. Pastor, F. Sanz, I. Soininen, D. Sousoni, Making *in silico* predictive models for toxicology FAIR, *Regul. Toxicol. Pharmacol.* 140 (2023) 105385, https://doi.org/10.1016/j.yrtph.2023.105385.