LJMU Research Online

Patel, M, Chilton, M, Sartini, A, Gibson, L, Barber, C, Covey-Crump, L, Przybylak, KR, Cronin, MTD and Madden, JC

 Assessment and Reproducibility of Quantitative Structure-Activity Relationship Models by the Nonexpert

http://researchonline.ljmu.ac.uk/id/eprint/7978/

Article

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

Assessment and Reproducibility of QSAR Models by the Non-Expert

Mukesh Patel,*§ Martyn Chilton,§ Andrea Sartini,§ Laura Gibson,§ Chris Barber,§ Liz Covey-Crump,§ Katarzyna R. Przybylak,‡ Mark T. D. Cronin,‡ and Judith C. Madden‡

§Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS, England

‡School of Pharmacy and Chemistry, Byrom Street, Liverpool John Moores University, Liverpool L3 3AF, England

*Email: mukesh.patel@lhasalimited.org

KEYWORDS: QSAR, ADME, Model Reproducibility, OECD Principles

ABSTRACT: Model reliability is generally assessed and reported as an intrinsic component of QSAR publications; it can be evaluated using defined quality criteria such as the Organisation for Economic Cooperation and Development (OECD) Principles for the validation of QSARs. However, less emphasis is afforded to the assessment of model reproducibility, particularly by users who may wish to use model outcomes for decision making, but who are not QSAR experts. In this study we identified a range of QSARs in the area of absorption, distribution, metabolism and elimination (ADME) prediction and assessed their adherence to the OECD Principles, as well as investigating their reproducibility by scientists without expertise in QSAR. 85 papers were

reviewed, reporting over 80 models for 31 ADME-related endpoints. Of these, 12 models were identified that fulfilled at least four of the five OECD Principles and three of these 12 could be readily reproduced. Published QSAR models should aim to meet a standard level of quality and be clearly communicated, ensuring their reproducibility, to progress the uptake of the models in both research and regulatory landscapes. A pragmatic workflow for implementing published QSAR models and recommendations to modellers, for publishing models with greater usability, are presented herein.

Introduction

In the chemical, pharmaceutical, pesticide and personal care product industries there is an ever increasing pressure to deliver ingredients that are fit for purpose, whilst at the same time avoiding any adverse effects to either humans or the environment. The traditional use of animal models to identify potential effects of chemicals has been criticised in terms of scientific relevance, time and cost - driving the quest for suitable alternatives.[1] *In silico* and *in vitro* methods have been proposed to address these concerns as well as the commercial, regulatory and ethical requirements for alternatives.[2-4] Quantitative structure-activity relationship (QSAR) modelling is a widely-used approach for predicting properties of chemicals, predicated on a mathematical relationship being derived between a chemical's structure and its physicochemical or biological (toxicological) properties. Acceptance of QSAR methods is increasing with the European Chemicals Agency (ECHA)[5] reporting that, for information on endpoints concerning vertebrate animals, 34% contain one or more QSAR predictions (6,290 substances analysed).

2

Over a period of more than fifty years a significant number of QSAR models have been published covering an extensive range of properties and biological endpoints. A wide variety of modelling techniques have been employed ranging from simple, transparent (multivariate) linear regression to more complex (potentially "black box"), machine learning algorithms. Cherkasov et al.[6] provide a comprehensive overview of the development of the field of QSAR including its history, methodologies used, advantages, limitations, applications to real world challenges and examples of both good and bad practice in model development. The importance of correctly recording all relevant information so that other users can realistically evaluate and reproduce published QSAR models is acknowledged as well as identifying the need for model developers and users to communicate effectively.[6]

Guidelines to assist in the development of robust *in silico* models for regulatory purposes were captured in the OECD Principles for the validation of QSARs[7] published in 2004. According to these guiding principles a model should have (i) a defined endpoint, (ii) an unambiguous algorithm, (iii) a defined domain of applicability, (iv) appropriate measures of goodness-of-fit, robustness and predictivity and (v) a mechanistic interpretation where possible (a fuller description is provided in the Supporting Information). Other publications have highlighted both good practice[8] and common errors in the development and reporting of QSARs[9, 10] with recent emphasis on more transparent recording of *in silico* models to ensure their reproducibility. Judson et al.[11] proposed Good Computer Modelling Practice (GCMP) guidelines which identify best practice in conducting and recording modelling procedures. The QSAR Model Reporting Format (QMRF),[12] provides a template for recording key information regarding a given QSAR model and associated validation studies, where relevant. The Journal of Chemical Information and Modelling has published required standards[13] for papers in the area of QSAR/QSPR. The QSAR databank[14]

provides a repository for storing QSAR models along with their associated metadata. Similarly, Ball et al.[15] report on best practice for read-across summarising the state-of-the art, lessons learned from REACH submissions to ECHA, how data can be used to support a read-across prediction and consideration of uncertainty. This body of literature emphasises the importance of ensuring that methodology is transparent and reproducible and all associated information is fully documented; this increases the acceptability of models and predictions derived therefrom. ECHA is active in promoting reliable and understandable QSAR models, which can be useful within the context of REACH regulation.

A similar case can be made for all models that are published which require some form of acceptance from other scientists, whether or not they are intended for a regulatory outcome. Cronin[16] reported a process by which QSAR models can be described and evaluated and demonstrates application of the method using case studies in which models are evaluated according to the OECD Principles. Hewitt et al.[17] proposed a scheme for the verification of *in silico* models in accordance with GCMP; the authors describe a Standard Operating Procedure (SOP) for the verification of *in silico* models along with templates (based on the QMRF format) to ensure sufficient information concerning the model is available to evaluate and reproduce the model. In a wider context Hanser et al.[18] discussed confidence in models with respect to the validity of the model for a given compound (applicability), the quality and quantity of the information supporting the prediction (reliability) and finally the likelihood of the outcome of this prediction (decidability).

Despite extensive guidance being available to ensure the validity and appropriate recording of QSAR models, little information is available on how widely these recommendations are implemented and the ramifications when attempting to reproduce literature models. Although *in*

silico techniques, (including QSAR modelling) have been widely applied, they can be perceived as the preserve of specialists. However, the end-user of such a model may well be an industrial, pharmaceutical or regulatory scientist who, whilst not a QSAR expert, may have responsibility for deciding whether or not a chemical progresses to the next stage (or is prioritised for testing) based, in part, on information obtained from QSAR predictions. Whilst literature on best practice in QSAR modelling is readily available much of this is aimed at those with a chemoinformatic/ mathematical understanding of the subject. An end-user with a background in biological sciences/ toxicology needs to have sufficient confidence in the model to justify its application in informing decisions on safety assessment or prioritisation. From a practical perspective guidance needs to be available on how to go about repeating a QSAR analysis, determining its suitability and applying the model to other compounds of interest.

The first step in investigating the reliability and reproducibility of published models is the selection of an appropriate set of models on which to perform such an analysis. For the purposes of the present study, models for a range of absorption, distribution, metabolism and elimination (ADME) properties were selected. These four characteristics determine the internal exposure of an organism to a chemical, which in addition to external exposure and intrinsic efficacy, determine the overall effect that the chemical may elicit.[19] Historically, much effort has been devoted to the assessment of toxicity models. However, whilst toxicokinetic considerations are increasingly recognised as an important component of read-across predictions, assessment of ADME is often neglected.[20] As models for ADME are increasingly being applied and represent a coherent collection of related endpoints these were chosen for investigation. ADME models are essential to predicting chemical activity but have generally been neglected in terms of assessment of suitability.

Detailed reviews have been published concerning datasets, QSAR models and freely available or commercial software tools to predict ADME endpoints.[19-21] The aim of this study was to investigate the reliability and reproducibility of QSAR models for ADME endpoints and to create a workflow to enable evaluation of *in silico* models by others. The specific objectives included investigating a range of ADME models available in the literature, assessment of a representative sample of these models for their adherence to the OECD Principles and determination of model reproducibility using a workflow designed to be appropriate for end-users who are not experts in QSAR. Whilst there are many ways to assess model quality, adherence to the OECD Principles was selected as a pragmatic approach as the Principles are well-known and can provide a consistent approach to assessing (traditional) QSAR models. Recommendations for improvements to the development and reporting of QSAR models, in order to make them more useful to others, are also included.

It is recognised that there is an increasing number of web-based tools that are available to predict ADME (and other) endpoints. These models can be applied by naïve and expert users, enabling the underlying model to be readily "reproduced" by the user.[14, 22, 23] Whilst these are a valuable and increasingly popular resource, the investigation here was focussed on published models, rather than assessment of such tools.

Materials and Methods

Identifying QSAR models for ADME endpoints available in the literature

Models available in the literature for ADME endpoints were identified using information provided in relevant reviews[19-21] in addition to searching on-line literature databases (PubMed, Google). The search terms that were used were taken from a previously identified list of ADME

parameters,[24] which are provided here as Supporting Information. Where multiple publications were identified for a particular ADME parameter, a sample of up to six papers was selected for further analysis; in these instances more recent publications (e.g. from 2010 onwards) were favoured, where available.

Evaluation of the models for their adherence to the OECD Principles for the validation of QSARs

Details from the publications were used to evaluate each model against the OECD Principles for the validation of QSARs.[7] The review of the models was carried out by experienced scientists with a background in chemistry and/or toxicology, who have a basic understanding of QSAR methods, but are considered naive rather than expert QSAR users. This involved a pragmatic application of the OECD Principles from the viewpoint of these stakeholders. Where a model was deemed to have fully met a specific OECD Principle (in the view of an individual scientist using their interpretation) this was indicated as 'Yes'. Where a model met an OECD Principle only in part this was indicated as 'Yes with limitations'. Models which failed to meet an OECD Principle were classed as 'No' (not meeting the Principle) or 'N/A' (not applicable in this case). The results of the evaluation of all of the models were tabulated in an Excel spreadsheet and are provided as a heat-map in the Supporting Information.

It is worth noting that for the number of publications being considered, the assessment exercise was time-bound and as such was conducted at a fairly high level. Furthermore, this analysis is likely to be subjective, as it is based upon the individual assessment of the scientists involved. Examples of how the criteria were used to assess a model's adherence to the principles are given below:

For the first principle, 'a model must have a defined endpoint', a methodology was considered to have fulfilled this principle fully if it clearly described the modelling of one of the previously enumerated ADME parameters.[24]

For the second principle, 'a model must have an unambiguous algorithm', if a clear description of both the QSAR modelling procedure and the final equation were provided, then this principle was considered to have been met in full.

For the third principle, 'a model must have a defined domain of applicability', if a publication describes the general class of chemicals for which a model is suitable (e.g. Volatile Organic Compounds), this would be considered partial information towards fulfilling the Principle and a verdict of "Yes with limitations" would be recorded. However, in order to fully satisfy the Principle the publication would ideally go into much more detail describing and defining the applicability domain, for example by looking at the range of descriptor and response values used or looking at how well particular sub-classes of chemicals within the training set are predicted.

For the fourth principle, 'a model must have appropriate measures of goodness-of-fit, robustness and predictivity', the presence of performance measures (such as R^2 and Q^2) were considered sufficient evidence that the principle had been fulfilled, regardless of the magnitude of these parameters. This principle was also considered to have been fully met even in the absence of a predictivity measure from an external test set, as often experimental data were scarce and the removal of a test set of chemicals was deemed to be impractical.

For the fifth principle, 'a model must have a mechanistic interpretation, if possible', the principle was considered to be fulfilled if the authors had commented on how each of the descriptors used in the algorithm were mechanistically related to the ADME parameter in question.

Assessing the reproducibility of the selected QSAR models using a newly derived workflow

It was initially anticipated that models meeting all five OECD Principles would be taken forward to the next stage of the investigation. However, it was determined that only two out of 86 models met all five Principles, hence the threshold was lowered such that all models meeting four out of the five Principles were considered; this identified a further 10 models. Therefore, 12 models in total were assessed, using a decision tree workflow to ascertain the ease of model reproducibility. As stated above, adherence to the OECD Principles was selected as a pragmatic assessment tool. Scientific judgement was based entirely on interpretation of the information presented in the publications and may not reflect the quality assurance processes undertaken by the authors during model development. Whilst a more experienced QSAR practitioner may have greater success in interpreting or reproducing the models, the aim here was to more accurately replicate the real-life scenario where a non-expert may need to use the results of a model for business-appropriate decision making. The decision tree workflow (illustrated in Figure 1) was designed as a step-by-step process to guide a non-expert in QSAR through the stages required in reproducing and evaluating a model.
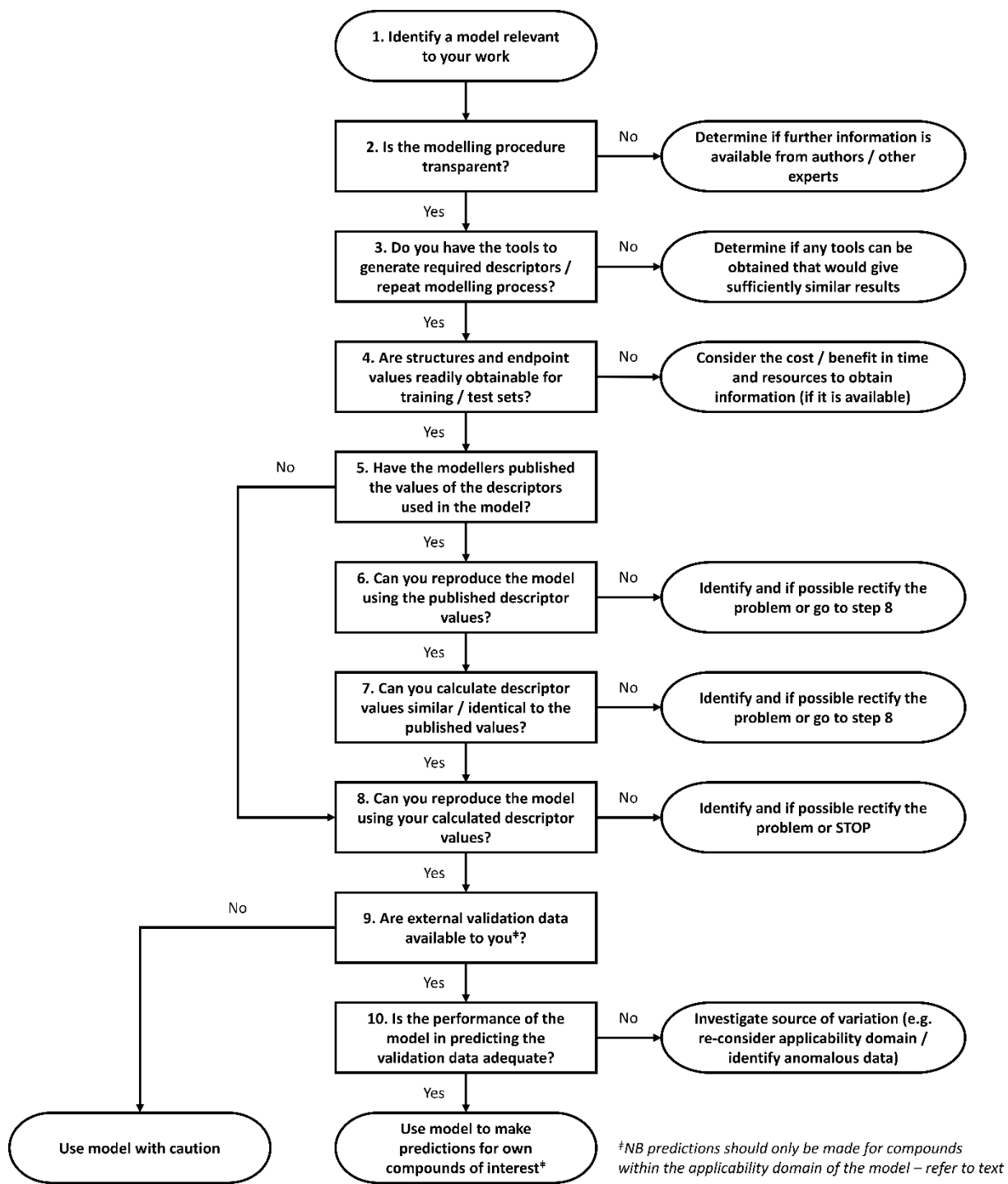
1. Identify a model relevant to your work

2. Is the modelling procedure transparent? — No → Determine if further information is available from authors / other experts

Yes

3. Do you have the tools to generate required descriptors / repeat modelling process? — No → Determine if any tools can be obtained that would give sufficiently similar results

Yes

4. Are structures and endpoint values readily obtainable for training / test sets? — No → Consider the cost / benefit in time and resources to obtain information (if it is available)

Yes

5. Have the modellers published the values of the descriptors used in the model? — No

Yes

6. Can you reproduce the model using the published descriptor values? — No → Identify and if possible rectify the problem or go to step 8

Yes

7. Can you calculate descriptor values similar / identical to the published values? — No → Identify and if possible rectify the problem or go to step 8

Yes

8. Can you reproduce the model using your calculated descriptor values? — No → Identify and if possible rectify the problem or STOP

Yes

9. Are external validation data available to you‡? — No

Yes

10. Is the performance of the model in predicting the validation data adequate? — No → Investigate source of variation (e.g. re-consider applicability domain / identify anomalous data)

Yes

Use model with caution

Use model to make predictions for own compounds of interest‡

‡NB predictions should only be made for compounds within the applicability domain of the model – refer to text

Figure 1. Workflow used to determine the reproducibility of a published QSAR model.

Results

Availability of QSAR models for ADME endpoints in the literature

Many published QSAR models for ADME endpoints were retrieved from searches of relevant publications/review articles and on-line literature resources. In all, 85 papers were identified that reported one or more QSAR models (Supporting Information). Note that one paper contained models for two endpoints whilst other papers contained two or more models for the same endpoint; as such, the 85 papers reviewed contained 86 (sets of) QSAR models in total. The number of models for each endpoint varied significantly depending on the nature of the endpoint. Features associated with absorption, such as human intestinal absorption and skin permeability, were well represented within the publications identified, however, few or no publications were found for endpoints such as distribution to specific tissues and overall extent of metabolism.

Adherence of the models to the OECD Principles for the validation of QSARs

Each set of QSAR models was assessed for its adherence to the OECD Principles. For each of the five Principles the model was deemed to completely fulfil (Yes), partially fulfil (Yes with limitations) or not fulfil the criteria for that Principle (No). The results of this analysis are presented as a heat map in the Supporting Information, where the result for each Principle for every model is recorded. N/A (not applicable) indicates that a Principle could not be readily applied to that model. For example if the required information was not readily accessible or if the model in question was proprietary and no suitable alternative was available. The data are summarised in Table 1 below for all studies (86 models). Results for models published 2006 onwards have been shown separately (68 models). The cut-off year of 2006 was selected as the OECD Principles were published in 2004 and therefore any differences in the proportion of models fulfilling or not fulfilling the Principles before and after they were widely publicised could be investigated. Note

that the Principles themselves were a formalisation of good modelling practices which had been widely accepted for many years, therefore it is not unreasonable to expect models published earlier to adhere to the Principles. Initial analysis determined that only 12 of the 86 models assessed adhered to at least four of the five OECD Principles, with two of these adhering to all five Principles as shown in Table 2.

Table 1. Summary of the number of models adhering to each of the OECD Principles (total number and those published 2006 onwards).

		Percentage of models meeting OECD Principle				
		(i) model based on a defined endpoint	(ii) model uses an unambiguous algorithm	(iii) model has a defined domain of applicability	(iv) appropriate measures of goodness-of-fit, robustness and predictivity used	(v) model is mechanistically interpretable
All studies (86)	Yes	65	28	30	50	26
Yes / Yes with limitations		84	53	60	76	57
Studies 2006 onwards (68)	Yes	68	26	26	54	25
Yes / Yes with limitations		87	51	56	76	60

Table 2. ADME models which were judged to meet at least four of the five OECD Principles.

			Adherence to OECD Principle nos 1-5				
Model	**ADME parameter(s) modelled**	**Reference**	**1**	**2**	**3**	**4**	**5**
1	Skin permeability/retention (PAMPA, K_p)	25	Yes	No	Yes	Yes	Yes
2	Skin absorption (K_p, J_{max})	26	Yes	Yes	Yes	Ywl	Yes
3	Oral absorption (PAMPA, HIA)	27	Yes	Yes	Yes	Yes	Ywl
4	Volume of distribution (V_d)	28	Yes	Yes	Yes	Yes	Yes
5	Blood-brain barrier permeability (log BB)	29	Yes	No	Yes	Yes	Yes
6	Plasma protein binding (log K'_{HSA})	30	Yes	Yes	Yes	Yes	No
7	Tissue:unbound plasma distribution ratio (K_{tb})	31	Yes	Ywl	Yes	Yes	Yes
8	Placental transfer (CI, TI)	32	Yes	Yes	Yes	Yes	Yes
9	Brain:blood/plasma partition coefficient (K_{tb})	33	Yes	Yes	Ywl	Yes	Yes
10	Volume of distribution (V_d)	34	Yes	Yes	Ywl	Yes	Yes
11	Tissue:blood/plasma partition coefficient (K_{tb})	35	Yes	Yes	Ywl	Yes	Yes
12	Biliary excretion (BE%)	36	Yes	Ywl	Yes	Yes	Yes

Ywl = Yes with limitations.

Reproducibility of the selected QSAR models for ADME endpoints

The 12 models identified as meeting at least four of the five OECD Principles were assessed for their reproducibility using the workflow described above. Table 3 shows the results of this analysis indicating whether the model could be readily reproduced or at which step of the process the decision was made not to proceed further with attempts to re-create the model. Where a model could be successfully reproduced, a KNIME workflow was developed to enable other researchers to implement the model more readily.

Table 3. Results of the assessment of reproducibility of the 12 selected models.

Model	Algorithm(s)	Outcome of workflow	Comment
1	MLR, ANN, PLS	Stopped at step 3	This set of models requires the use of commercially available descriptors for which no suitable alternative was available for this study.
2	MLR	Reproduced (to step 9)	KNIME workflow generated.
3	Bi-linear QSAR model	Stopped at step 4	Lack of explicit designation of the training and test set compounds.
4	MLR	Stopped at step 3	Descriptors based on experimentally derived parameters, requires *in vitro* experiments.
5	GBT, BDT	Stopped at step 2	Algorithm and settings are complex and difficult to implement.
6	NLR, SLR	Stopped at step 3	Software and settings for calculating descriptors were not clearly reported for the model. Descriptors not adequately explained.
7	PBPK model, SLR	Reproduced (to step 9)	KNIME workflow generated.
8	MLR, SLR	Reproduced (to step 9)	KNIME workflow generated.
9	MLR	Stopped at step 3	The required descriptors are either experimental values or have been calculated using commercial software not available for the present study.
10	MLR, NLR	Stopped at step 3	Required software not available in the present study, full list of 250 descriptors not provided in the paper.
11	NLR	Stopped at step 3	Descriptors were calculated using proprietary software not available for the present study.
12	MLR, CART, BT, RF	Stopped at step 3	The majority of the descriptors in the paper are calculated with commercial software (not available for the present study) including calculations from a docking experiment where settings were not reported.

MLR – Multiple Linear Regression (i.e. containing 2 or more independent variables), ANN – Artificial Neural Networks, PLS – Partial Least Squares regression, SLR – Simple Linear Regression (i.e. containing only 1 independent variable), NLR – Nonlinear Regression, CART – classification and regression trees, BT – boosted trees, RF – random forest, GBT – gradient boosted trees, BDT – bagging decision trees.

Discussion

This study described herein included a brief review of available QSAR models for ADME endpoints, their adherence to the OECD Principles for the validation of QSARs and an investigation into how readily a non-expert in QSAR could reproduce a selection of the models. The role of internal exposure in predicting effects of chemicals is now well recognised.[37] Models are needed for all ADME endpoints, however, review of the literature demonstrated that there are significant differences in numbers of models available depending on the endpoint in question. This reflects the availability of data on which to build models and the complexity of the processes involved. Future priorities lie in building models for poorly represented endpoints to reduce the knowledge gaps in this area. Data sharing initiatives, such as those undertaken within European Union funded projects, may be useful in developing datasets for modelling purposes.

Availability of QSAR models for ADME endpoints

It would have been advantageous to have equal numbers of models for all endpoints to make a fair comparison, however there was a heterogeneous distribution of models across endpoints. Where many models were available, a maximum of six models for an individual ADME endpoint were selected for investigation. This was a pragmatic approach to ensure a variety of models were selected without placing too much emphasis on well-studied endpoints. Where several models were available priority was given to more recent models.

Assessment of adherence of the models to the OECD Principles

86 models were assessed for their adherence to the OECD Principles. It is recognised that there are many methods by which the reliability or robustness of models can be assessed. The OECD Principles formalised factors generally associated with good modelling practice and therefore, even if models were developed prior to the publication of the Principles, such factors should have

been considered. However, where models are identified as not meeting the OECD criteria (according to our interpretation) this does not necessarily imply poor model quality. Conversely, it is possible that a model that meets most or all criteria could still be considered of poor quality for other reasons. For example, training and test sets may be well-defined but may not have been appropriately selected. Models built or tested using small or restricted datasets will not provide reliable results across diverse chemical space - a model user would need to check the appropriateness of the model for their purposes.

The Excel spreadsheet available in the Supporting Information gives the results for adherence to the Principles for the individual models and these data are summarised in Table 1. Strictly applying the criteria (i.e. 'Yes' fully meets OECD Principle) 65% of all models were considered to satisfy Principle 1 and 50% satisfied Principle 4, with only 28%, 30% and 26% meeting Principles 2, 3 and 5 respectively. With a more generous interpretation of the criteria (i.e. allowing both 'Yes' and 'Yes with limitations') overall scoring improved with 84% of all models satisfying Principle 1 and 76% meeting criteria 4. 53%, 60% and 57% of models adhered to Principles 2, 3 and 5 respectively. No significant differences were seen for models analysed here between those published before and after 2006. This is disappointing as evidencing the validity of a model is associated with greater confidence in their use and wider application. Modellers should be encouraged to ensure published models explicitly satisfy criteria for the validation of QSARs or provide a sound rationale for why there may be exceptions. In this analysis, only two models explicitly met all five criteria as strictly applied, with 12 meeting the criteria when more generous allowance was made. In selecting models to forward to the next stage (i.e. assessment of reproducibility) an assumption was made that all five Principles carry equal weight, such that if any four criteria were met the models were taken forward. This ensured a consistent approach was

applied when selecting models to reproduce, however, it does not imply that these models were ideal. It is also acknowledged that some of the other models which met fewer OECD Principles may have been more readily reproduced; however, this may have given rise to unfair selection bias.

Assessment of model reproducibility

The workflow represented in Figure 1 was used in a step-wise manner to reproduce the models and/or identify at which stage of the process reproduction of the model failed. This workflow serves as a pragmatic tool for other researchers interested in evaluating QSAR models available in the literature.

Using the workflow

The first step is to identify a relevant model for an endpoint of interest e.g. from a literature databases, on-line repositories of models or websites. For certain endpoints a plethora of models are available, whilst for other endpoints models are scarce. Where a selection of models are available from which to choose, those based on large, diverse training sets, employing transparent methodology are useful starting points.

The second step relates to the "transparency" of the model. Is it clear how the model was generated and how the results should be interpreted? Simple (multiple) linear regression models are generally considered transparent, i.e. it is apparent how descriptors for chemical structures are mathematically correlated with the property or activity of interest. In other cases, e.g. neural network-based models, how such a relationship was derived may not be as obvious. As a general rule, a more complex modelling algorithm will at the very least require much more detailed explanation of model architecture/settings etc. in order to be sufficiently transparent to ensure reproducibility and interpretability. If not apparent from the publication itself, it may be possible

to obtain further information from the model developers directly or from other users (e.g. via a user forum).

Step three in the workflow refers to the availability, to the model user, of tools used by the model developer for generating the descriptors and/or creating the model. Ideally, the same tools should be available to the model user, but this may not be practically possible due to costs or accessibility. Problems can also arise where for example, different versions of the same software may give different results. There are also instances where commonly used, freely available software is no longer supported by the software developers and becomes obsolete. Whilst it is not possible to prevent all such eventualities, use of well-established, free tools is preferable. However, it is recognised that specific very high quality software may only be available on a commercial basis. If the same software is not available to the model user then suitable alternatives should be sought.

Step four is fundamental - in order to repeat a model for the training set used the correct structures, unambiguously associated with endpoint values, must be available. Journals now commonly encourage the publication of all relevant data, either within the text or as Supporting Information and therefore the information can often be readily obtained in a given format. Przybylak et al.[24] discuss the importance of data format in determining modellability i.e. how suitable they are for QSAR (and other) modelling in terms of accessibility of structures, relevance, number of data etc. Formats such as Excel spreadsheets are convenient, where two or more concordant identifiers (e.g. SMILES string, CAS number, InChI key, pictorial representation) are associated with each chemical. SMILES strings have the advantage of being interpretable and recognisable by a wide range of software, although there are specific issues with this form of representation (e.g. limited ability to deal with stereochemistry or tautomers). Where chemical structures are not provided in a readily portable format (e.g. use of GIF images or non-standard

names) then the cost: benefit ratio must be considered. It may be unreasonably time consuming and potentially highly error-prone to derive structures from non-standard data formats, therefore the user must determine if the potential benefit of the model out-weighs this cost. The accuracy of the endpoint data is another consideration – a good model cannot be based on poor quality data. Methods for assessing the quality of data and processes of data curation have been extensively reviewed elsewhere and is not the focus here (for further information see Przybylak et al.[38]; Nendza et al.[39]; Williams et al.[40]; Williams and Ekins[41]; Fourches[42]; Tropsha[43]; Young[44]). Where a model is reported by an experienced QSAR practitioner, the model developer should have considered issues of data quality when generating the model.

Steps five to eight cover the process of reproducing the model itself, either using descriptor values as recorded in the publication or generating the descriptor values *de novo*. If descriptor values are given in the publication, these can be compared to those generated by the model user and any discrepancies identified (e.g. differences arising from software versioning or recording errors) and where possible rectified. It is within these steps where model documentation is of paramount importance. Sufficient detail regarding software version, settings, protocols, assumptions or constraints must be provided to enable another user to achieve the same output. In reality it may require repeated attempts and investigation into discrepancies before all issues can be resolved and concordant results reached. Note that results (from generating descriptors, or output from the final model) should be concordant although not necessarily identical. For example different versions of software may result in insignificant differences in values, or a recording error may be identified which when corrected results in non-identical output from the model. The model may still be considered as reproducible and of benefit to the end user as it is pragmatic application of the model in a real-world scenario that is important rather than a direct repetition.

Success at stage eight indicates a reproducible model has been identified; steps nine and ten relate to the confidence which may be placed on the output from the model. Model validation includes an assessment of the applicability of the model to compounds beyond the training set. A model that cannot reliably make predictions beyond the data on which it was trained has limited use (albeit the model may elucidate underlying mechanisms relevant to the training set). Assessing model predictivity using a validation dataset gives a true indication of the usefulness of the model. If reliable results are obtained for validation data then the model can be used with greater confidence for compounds of interest.

Note that all models can only make predictions reliably within the applicability domain of the model i.e. the chemicals for which predictions are to be made must be sufficiently "similar" to the training set compounds. Whilst an applicability domain aims to provide support to allow a decision to be made based upon the model's output, this alone is generally insufficient. A user primarily requires an understanding of the expected certainty (or confidence) associated with an individual prediction and some analysis as to how robust this measure is in order to undertake expert review. Defining an applicability domain can help if there is an explanation of how the boundary is defined, which properties are considered important for the endpoint, how the thresholds have been defined, what level of performance should be expected within the boundary and evidence that this is true across the chemical space it encompasses. There are many ways in which applicability domain may be determined, for example ensuring the descriptor values for the test chemicals do not exceed the values for the training set chemicals or ensuring that models derived from simple structures are not inappropriately extrapolated to complex chemicals with multiple functional groups. Methods for assessing applicability domain are discussed elsewhere.[45-50] This then moves some way towards recognising the difference between the average performance of a model for a test set

(typically the modeller's measure of success) and the specific performance for a particular compound of interest (normally the user's measure of success). Successful navigation of the complete workflow illustrated in Figure 1 indicates a model has been shown to be useful, reproducible and predictive.

Outcome of assessment of reproducibility for 12 selected models

Table 3 summarises the results of the assessment of the 12 selected models for their reproducibility. Full details for each model i.e. workflow forms, detailing the process and decisions taken at each step are available in the Supporting Information. Whilst individual reasons for failing varied on a case by case basis, a summary of the overall trends observed across all 12 of the models considered is presented. Three models were successfully recreated (models 2, 7 and 8 as indicated in Table 3); KNIME[51] workflows, containing the training data, for these models are also provided in the Supporting Information. For the small number of reproducible models the relevant chemical and biological data could be extracted from the publication relatively easily and required a minimal amount of further data curation. The models relied on non-commercial software, or could be recreated using freely available versions of the descriptors used. Furthermore, all three (sets of) models were examples of linear regression, highlighting the fact that this modelling technique is generally both easy to comprehend and to replicate.

In the process of attempting to recreate the 12 ADME QSAR models, a number of useful, free tools were identified, such as the KOWWIN software,[52] produced by the US Environment Protection Agency for calculating the logarithm of the octanol water partition coefficient (log P). The modelling itself was carried out within KNIME, and the RDKit and CDK nodes provided therein were used to calculate a number of different descriptors. Furthermore within KNIME it was possible to access the Royal Society of Chemistry's ChemSpider web service;[53] this was used

in cases where there were no explicit chemical structures provided in the publication in order to generate such information from other chemical identifiers (CAS number and chemical name). Whilst this service proved invaluable in being able to reproduce some of the models, it is worth noting that the structures generated in this manner did require further manual curation, as errors were occasionally produced. One useful way of checking the validity of the structures produced was to compare the generated molecular weight to the molecular weight value reported by the original authors, and to further scrutinise any compounds where there was a large discrepancy between the two values (this, of course, is only possible where the original publication contains the relevant molecular weight information). Issues with data quality, such as ensuring correctness of chemical structure, are well recognised and have been extensively reviewed previously. A review of methods for assessing data quality is beyond the scope of this paper, but can be found elsewhere.[43,44] In assessing model reproducibility here, absolute repetition of the original model was not considered the main criterion. The aim here was to determine if the algorithm could be used, and precise reproducibility is not as important as providing an output that is sufficiently similar for the decision to be made from it. For example if the original model required use of specific software to calculate a molecular property and the same software (or version of the software) was not available when assessing the model, alternative software/versions (giving minor differences in property calculation / endpoint prediction) could be considered suitable. Where errors were identified in the original publication (e.g. repetition of compounds in the training set, or ambiguous structural identity for one or two compounds within a large dataset) these were corrected, although this resulted in non-identical values to the original publication (hence strict duplication of the model failed) the model was still deemed reproducible. In this manner the process reflects how an end-user is likely to utilise models in real-life scenarios.

Table 3 shows that 9 out of the 12 selected models were not successfully reproduced in this analysis. The most common reason for being unable to replicate a model was not having access to the relevant software to generate descriptors, or a suitable alternative. For this reason seven of the models did not proceed beyond step three of the workflow. It is recognised that this limitation will not be an issue for all researchers wishing to replicate the model; however, to ensure greater uptake of a model use of software that is more widely available is preferred. As certain endpoints may only be amenable to modelling techniques requiring proprietary software, this issue may not be resolvable in all cases; however, this highlights the need to use freely available software whenever possible. Another reason for failure (models 5 and 6) was lack of sufficient information in the publication i.e. in relation to software and settings for descriptor generation; this issue is resolvable. Modellers should ensure all details are available and this can be tested prior to publication (e.g. by requesting a third party, in-house or external, to attempt to reproduce the model). This is comparable to chemists publishing in Organic Synthesis where all experimental procedures are reproduced before publication.[54] In this way any lack of detail or possible ambiguities could be identified in advance. Similarly, model 3 failed as the designation of training and test set compounds was not clear and again this is not a difficult problem to overcome. Within the EU eTOX project, templates were devised (based in large part on the QMRF format) for modellers to guide appropriate documentation of the modelling process such that another user could verify or repeat the model.[17] Using such templates should help to resolve some of these miscommunication issues which could lead to wider uptake of published models. Other issues encountered included the use of experimentally-derived descriptors in the model which limits reproducibility to those with comparable laboratory facilities.

Overall it was shown that, for the models studied here, ranking on the basis of adherence to the OECD Principles was not successful in identifying models that were readily reproducible. As alluded to previously, adherence to the Principles does not necessarily imply a good model, an end user would need to determine the appropriateness of the model for a given query. Documentation of rationale, methods and data for published QSAR models were often lacking; descriptors and relevant software tools were not freely available for many of these models. Some models would appear to only be usable by other experienced QSAR practitioners or those with facilities to generate descriptors from experiment; this limits their application by others interested in the outcomes of such models (e.g. risk assessors / regulators). Models may be either unusable or provide results in which the end user has less confidence. This means that models are unlikely to be used, or are indefensible when making decisions based on predictions from such models. These findings confirm those from D'Onofrio[55] who assessed the reproducibility of linear regression models in the European Commission's Joint Research Centre QSAR Database.[22] Despite being recorded in a well documented manner, there were still difficulties in reproducing a number of the models in the database.

Since the publication of the OECD Principles in 2004, there have been many innovations in the area of *in silico* prediction. Whilst application of the Principles to traditionally developed QSAR models is still valid, it is recognised that the Principles do not provide an appropriate means to assess some models generated using newer techniques, such as Deep Learning. Assessment of the usefulness of such models is driven more by their success in modelling endpoints accurately, rather than their transparency and mechanistic interpretability. Newer approaches to model development will similarly require a new approach to model assessment to ensure user confidence in applying the approach and promote greater uptake of the models.

Conclusions

Literature concerning QSAR models can be difficult to understand for a non-expert in the area, however using such models may be a key requirement for their decision making. QSAR models are published and promoted on the basis of cost, utility, scientific validity and reproducibility. Whilst models are developed and published by domain experts, the target users are often experimental or regulatory scientists with diverse backgrounds outside the QSAR modelling arena. If QSAR practitioners wish to promote wider usage of their models, comprehensive reporting of methodology and results to enable assessment of validity, reproducibility and translation into biological relevance is essential. Traditionally-derived QSAR models should follow the OECD Principles and practical guidance for submission to journals is available to encourage transparency and reproducibility. The OECD Principles provide a benchmark for the assessment of QSAR models; however, there is the potential for disparity between assessors as to whether the Principles are interpreted strictly or more flexibly. A lack of adequate documentation has been shown here to be responsible for many instances of models failing to be reproducible. This means the model is of no value to other users and this needs to be addressed by the modelling community if greater uptake of models is to be achieved. The decision tree workflow presented in Figure 1 is designed to assist the non-expert in QSAR in assessing the usefulness, reproducibility and practical application of a QSAR model of interest. The methodology applies to QSAR models in general although the emphasis here was on QSAR models for ADME endpoints. From the above analysis, several areas where improvements can be made have been identified and these are summarised here as recommendations for the future development and reporting of QSAR models.

Recommendations:

(1) More models are required for the ADME endpoints for which there are currently few or no models. Lack of models may be a reflection of lack of data on which to the build models, or inherent complexity of the endpoint. Data sharing initiatives may assist in building suitable datasets for modelling where data are lacking.[56] Developments in computational methods are being used to generate models for more complex endpoints.

(2) The OECD Principles provide a consistent approach for assessing models, however, this does not necessarily equate with reproducibility. Adherence to OECD Principles should be assured explicitly as part of the modelling process (or a rationale provided where adherence to a Principle is not deemed relevant for a specific case). Note that the Principles relate to specific aspects of model quality, it is possible for a model to adhere to the Principles but to be considered "poor quality" for other reasons. The end user may need to ascertain appropriateness of the model for their purpose.

(3) The "simplest" modelling technique, that is suitable for purpose, should be applied, i.e. the most easily reproducible and transparent approach. For certain endpoints more complex methodologies may be required to provide adequate predictivity, however, the philosophy of applying the "simplest" approach to achieve the required outcome is still applicable.

(4) Ideally the same software should be available for model developers and subsequent users. This means that wherever possible freely available tools should be used in model development. Where commercial tools offer the best solution, this can provide a useful model for some researchers but will limit the general uptake of the model. Computational pipeline tools such as KNIME allow for a transparent working model to be provided alongside the publication which is highly beneficial for other users. Where appropriate, providing models as web-based tools for other

researchers is another possibility to ensure reproducibility and promote further uptake and assessment of the model.

(5) Full documentation / appropriate metadata should accompany the published model so that it can be understood, assessed for validity and readily reproduced. Templates for ensuring adequate documentation of models are readily available, for example the QMRF documentation available from the JRC[15] or those published by Hewitt et al. under the auspices of the EU eTOX project[20] which in turn were largely based on the QMRF.

ASSOCIATED CONTENT

Supporting Information.

The following files are available free of charge.

Data summary and heatmap (xlsx)

Workflow_reports- Table of the OECD criteria used in evaluating the models and summaries of using the workflow on each model; (docx)

KNIME workflows prepared in KNIME 3.3 (zip)

AUTHOR INFORMATION

Corresponding Author

*Email: mukesh.patel@lhasalimited.org

Notes

The authors declare no competing financial interest.

REFERENCES

(1) National Research Council. *Toxicity Testing in the 21st Century: A Vision and a Strategy*. The National Academies Press, Washington DC, 2007. https://doi.org/10.17226/11970 (accessed November 21, 2017).

(2) European Commission. Regulation (EC) No. 1907/2006 of the European Parliament and of the Council of 18 December 2006 Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Establishing a European Chemicals Agency, Amending Directive 1999/45/ EC and Repealing Council Regulation (EEC) No. 793/93 and Commission Regulation (EC) No. 1488/94 as well as Council Directive 76/769/ EEC and Commission Directives 91/155/EEC, 93/67/ EEC, 93/105/EC and 2000/21/EC. *Off. J. Eur. Union.* **2006**, *L 396*, 1–849.

(3) European Commission. Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on Cosmetic Products. *Off. J. Eur. Union.* **2009**, *L 342,* 59-209.

(4) European Food Safety Authority. Scientific report of EFSA. Modern Methodologies and Tools for Human Hazard Assessment of Chemicals. *EFSA J.* **2014**, *12(4), 1-87.* http://dx.doi.org/10.2903/j.efsa.2014.3638 (accessed November 21, 2017).

(5) European Chemicals Agency. The use of alternatives to testing on animals for the REACH Regulation Third report under Article 117(3) of the REACH Regulation, *ECHA-17-R-02-EN*. 2014, ISBN: 978-92-9495-760-3.

(6) Cherkasov, A.; Muratov, E.N.; Fourches, D.; Varnek, A.; Baskin, I.I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V.E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modelling: where have you been? Where are you going to? *J. Med. Chem*. **2014**, *57*, 4977-5010.

(7) Organisation for Economic Co-operation and Development. OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models. http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf (accessed November 21, 2017).

(8) Walker, J.D.; Jaworska, J.; Comber, M.H.; Schultz, T.W.; Dearden, J.C. Guidelines for Developing and Using Quantitative Structure-Activity Relationships. *Environ. Toxicol. Chem.* **2003**, *22*, 1653–1665.

(9) Cronin, M.T.D.; Schultz, T.W. Pitfalls in QSAR. *J. Mol. Struct.: THEOCHEM*. **2003,** *622*, 39–51.

(10) Dearden, J.C.; Cronin, M.T.; Kaiser, K.L. How Not to Develop a Quantitative Structure-Activity or Structure-Property Relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **2009**; *20,* 241–266.

(11) Judson, P.N.; Barber, C.; Canipa, S.J.; Poignant, G.; Williams, R. Establishing Good Computer Modelling Practice (GCMP) in the Prediction of Chemical Toxicity. *Mol. Inf*. **2015**, *34*, 276– 283.

(12) European Commission. QSAR Model Reporting Format (QMRF). https://ec.europa.eu/jrc/en/scientific-tool/qsar-model-reporting-format-qmrf (accessed November 21, 2017).

(13) Jorgensen, W.L. QSAR/QSPR and Proprietary Data. *J. Chem. Inf. Model*. **2006**, *46*, 937.

(14) Ruusmann, V.; Sild, S.; Maran, U. QSAR DataBank Repository: Open and Linked Qualitative and Quantitative Structure–Activity Relationship Models. *J. Cheminf.* **2015**, *7, 1-11*. http://doi.org/10.1186/s13321-015-0082-6 (accessed November 17, 2017).

(15) Ball, N.; Cronin, M. T. D.; Shen, J.; Blackburn, K.; Booth, E.D.; Bouhifd, M.; Donley, E.; Egnash, L.; Hastings, C.; Juberg, D.R.; Kleensang, A.; Kleinstreuer, N.; Kroese, D.; Lee, A.C.; Luechtefeld, T,. Maertens, A., Marty, S.; Naciff, J.M.; Palmer, J.; Pamies, D.; Penman, M.; Richarz, A. N.; Russo, D.P.; Stuard, S.B.; Patlewicz, G.; van Ravenzwaay, B.; Wu, S.; Zhu, H.; Hartung, T. Toward Good Read-Across Practice (GRAP) Guidance. *ALTEX*. **2016**, *33*, 149-166.

(16) Cronin, M.T.D. Characterisation, Evaluation and Possible Validation of *In Silico* Models for Toxicity: Determining if a Prediction is Valid. In *In Silico Toxicology: Principles and Applications*; Cronin, M.T.D., Madden, J.C.; Eds.; RSC: Cambridge, UK, 2010; Chapter 11, pp 275-300.

(17) Hewitt, M.; Ellison, C.; Cronin, M. T. D.; Pastor, M.; Steger-Hartmann, T.; Munoz-Muriendas, J.; Pognan, F.; Madden, J.C. Ensuring Confidence in Predictions: A Scheme to Assess the Scientific Validity of In Silico Models. *Adv. Drug Delivery Rev*. **2015**, *86*, 101-111.

(18) Hanser, T.; Barber, C.; Marchaland, J. F.; Werner, S. Applicability Domain: Towards a more Formal Definition. *SAR QSAR Environ. Res.* **2016**, *86*, 865-811.

(19) Madden, J. Toxicokinetic Considerations in Predicting Toxicity. In *In Silico Toxicology: Principles and Applications*; Cronin, M.T.D., Madden, J.C.; Eds.; RSC: Cambridge, UK, 2010; Chapter 21, pp 531–557.

(20) Schultz, T. W.; Cronin, M. T. D. Lessons Learned from Read-Across Case Studies for Repeated-Dose Toxicity. *Regul. Toxicol. Pharmacol.* **2017**, *88*, 185–191.

(21) Mostrag-Szlichtyng, A.; Worth, *A. Review of QSAR Models and Software Tools for Predicting Biokinetic Properties*. JRC Technical Report EUR 24377 EN, Publications Office of the European Union, Luxemburg, 2010.

(22) JRC QSAR Model Database. https://eurl-ecvam.jrc.ec.europa.eu/databases/jrc-qsar-model-database (accessed February 05, 2018).

(23) Online Chemical Database with Modelling Environment. https://ochem.eu/home/show.do (accessed February 05, 2018).

(24) Przybylak, K.; Madden, J. C.; Covey-Crump, E.; Gibson, L.; Barber, C.; Patel, M.; Cronin, M.T.D. Characterisation of Data Resources for *In Silico* Modelling: Benchmark Datasets for ADME Properties. *Expert Opin. Drug Metab. Toxicol.* **2017**, 1-13.

(25) Dobričić, V.; Marković, B.; Nikolic, K.; Savić, V.; Vladimirov, S.; Čudina, O. 17β-Carboxamide Steroids – *In Vitro* Prediction of Human Skin Permeability and Retention using PAMPA Technique. *Eur. J. Pharm. Sci*. **2014**, *52*, 95-108.

(26) Shen, J.; Kromidas, L.; Schultz, T.; Bhatia, S. An *In Silico* Skin Absorption Model for Fragrance Materials. *Food Chem. Toxicol*. **2014**, *74*, 164-176.

(27) Akamatsu, M. Importance of Physicochemical Properties for the Design of New Pesticides. *J. Agric. Food Chem.* **2011**, *59*, 2909-2917.

(28) Lombardo, F.; Obach, R.S.; Shalaeva, M. Y.; Gao, F. Prediction of Human Volume of Distribution Values for Neutral and Basic Drugs. 2. Extended Data Set and Leave-Class-Out Statistics. *J. Med. Chem*. **2004**, *47*, 1242-1250.

(29) Gupta, S.; Basant, N.; Singh, K. P. Qualitative and Quantitative Structure–Activity Relationship Modelling for Predicting Blood Brain Barrier Permeability of Structurally Diverse Chemicals, *SAR QSAR Environ. Res.* **2015**, *26*, 95-124.

(30) Colmenarejo, G. *In Silico* Prediction of Drug-Binding Strengths to Human Serum Albumin. *Med. Res. Rev*. **2003**, *23*, 275-301.

(31) Nestorov, I.; Aarons, L.; Rowland, M. Quantitative Structure-Pharmacokinetics Relationships: II. A Mechanistically Based Model to Evaluate the Relationship Between Tissue Distribution Parameters and Compound Lipophilicity. *J.Pharmacokinet. Biopharm*. **1998**, *26*, 521-545.

(32) Hewitt, M.; Madden, J. C.; Rowe, P. H.; Cronin, M. T. D. Structure-Based Modelling in Reproductive Toxicology: (Q)SARs for the Placental Barrier. *SAR QSAR Environ. Res.* **2007**, *18*, 57-76.

(33) Abraham, M. H.; Ibrahim, A.; Acree Jr., W. E. Air to Brain, Blood to Brain and Plasma to Brain Distribution of Volatile Organic Compounds: Linear Free Energy Analyses. *Eur. J. Med. Chem.* **2006**, *41*, 494-502.

(34) Ghafourian, T.; Barzegar-Jalali, M.; Dastmalchi, S.; Khavari-Khorasani, T.; Hakimiha, N.; Nokhodchi, A. QSPR Models for the Prediction of Apparent Volume of Distribution. *Int. J. Pharm.* **2006**, *319*, 82-97.

(35) Zhang, H. A New Approach for the Tissue−Blood Partition Coefficients of Neutral and Ionized Compounds. *J. Chem. Inf. Model.* **2005**, *45*, 121-127.

(36) Sharifi, M.; Ghafourian, T. Estimation of Biliary Excretion of Foreign Compounds Using Properties of Molecular Structure. *AAPS J.* **2014**, *16*, 65-78.

(37) Bessems, J.; Coecke, S.; Gouliarmou, V.; Whelan, M.; Worth, A. EURL ECVAM Strategy for Achieving 3Rs Impact in the Assessment of Toxicokinetics and Systemic Toxicity. JRC Science and Policy Report EUR 27315 EN. 2015.

(38) Przybylak, K.; R.; Madden, J. C.; Cronin, M. T. D.; Hewitt, M. Assessing Toxicological Data Quality: Basic Principles, Existing Schemes and Current Limitations. *SAR QSAR Environ. Res.* **2012**, *23,* 435-459.

(39) Nendza, M.; Aldenberg, T.; Benfenati, E.; Benigni, R.; Cronin, M.; Escher, S.; Fernandez, A.; Gabbert, S.; Giralt, F.; Hewitt, M.; Hrovat, M.; Jeram, S.; Kroese, D.; Madden, J.;

Mangelsdorf, I.; Rallo, R.; Roncaglioni, A.; Rorije, E.; Segner, H.; Simon-Hettich, B.; Vemeire, T. Data Quality Assessment for *In Silico* Methods: a Survey of Approaches and Needs. In *In Silico Toxicology: Principles and Applications*; Cronin, M.T.D., Madden, J.C.; Eds.; RSC: Cambridge, UK, 2010; Chapter 4, pp 59–118.

(40) Williams, A. J.; Ekins, S.; Tkachenko, V. Towards a Gold Standard: Regarding Quality in Public Domain Chemistry Databases and Approaches to Improving the Situation. *Drug Discovery Today*. **2012**, *17*, 685 – 701.

(41) Williams, A. J.; Ekins, S. A Quality Alert and Call for Improved Curation of Public Chemistry Databases. *Drug Discovery Today.* **2011**, *16*, 747-750.

(42) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: on the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50*, 1189-1204.

(43) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476-488.

(44) Young, D.; Martin, T.; Venkatapathy, R.; Harten, P. Are the Chemical Structures in Your QSAR Correct? *QSAR Comb. Sci.* **2008**, *27*, 1337-1345.

(45) Aniceto, N.; Freitas, A. A.; Bender, A.; Ghafourian, T. A Novel Applicability Domain Technique for Mapping Predictive Reliability across the Chemical Space of a QSAR: Reliability‑Density Neighbourhood. *J. Cheminf.* **2016**, *8*, 69.

(46) Sahlin, U.; Jeliazkova, N.; Öberg, T. Applicability Domain Dependent Predictive Uncertainty in QSAR Regressions. *Mol. Inf.* **2014**, *33*, 26–35.

(47) Hewitt, M.; Ellison, C. M. Developing the Applicability Domain of *In Silico* Models: Relevance, Importance and Methods. In *In Silico Toxicology: Principles and Applications*; Cronin, M.T.D., Madden, J.C.; Eds.; RSC: Cambridge, UK, 2010; Chapter 2, pp. 59–118.

(48) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T..; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D.; Schultz, T.; Stanton, D. W.; van de Sandt, J. J.; Tong, W.; Veith, G.; Yang, C. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure—Activity Relationships. The Report and Recommendations of ECVAM Workshop 52. *ATLA, Altern. Lab. Anim.* **2005**, *32*, 155–173.

(49) Sahigara. F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules*. **2012**, *17*, 4791-4810.

(50) Aniceto, N.; Freitas, A. A.; Bender, A.; Ghafourian, T. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. *J. Cheminf.* **2005**, *45*, 839-849.

(51) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization*; Springer, 2007; pp 319-326.

(52) US EPA. [2017]. Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.11. United States Environmental Protection Agency, Washington, DC, USA.

(53) ChemSpider. http://www.chemspider.com/Default.aspx (accessed November 21, 2017).

(54) Instructions for Authors. *Org. Synth.* **2017**. http://www.orgsyn.org/instructions.aspx (accessed November 21, 2017).

(55) D'Onofrio, E. Reproducibility of Linear QSAR Models in the JRC Database. Marie Curie Initial Training Network Environmental Chemoinformatics (ECO), The Helmholtz Zentrum, München, (2012) available from http://www.eco-itn.eu/sites/eco-itn.eu/files/reports/report_DONOFRIO_final.pdf (accessed 2 February 2018).

(56) Covey-Crump, E.; Elder, D. P.; Harvey, J.S.; Teasdale, A.; White, A.; Williams, R. V. Mutagenic Impurities: Precompetitive/Competitive Collaborative and Data Sharing Initiatives. *Org. Process Res. Dev.* **2015**, *19*, 1486–1494.

For Table of Contents use only