Table S4. Strategies to reduce uncertainties in the QSARs considered.

*Study 1: QSARs for inhibition of growth of aromatic compounds to T. pyriformis*

In Study 1 Luan et al. (2018) developed a series of QSARs for aromatic (single ring with multiple substituents including hydroxy, nitro and halogen groups, amongst others) compounds using historical data for the inhibition of growth to the ciliated protozoan *Tetrahymena pyriformis* extracted from Schultz et al (2003). The data set has been further split into three subsets and Multiple Linear Regression (MLR) equations and Radial Basis Function Neural Networks (RBFNN) models developed. Overall, RBFNN performed better than MLR. The assessment of uncertainties considered the QSARs reported as a whole, although comments could also be addressed to specific models.

Application of the assessment criteria and the resulting scores assigned is available in Table S2a. Analysis of these findings shows that high uncertainty was found mainly to be associated with:

i) The toxicity data modelled. Information on the toxicity endpoint is not provided. This could be obtained from the source publication as well as other publications by the authors who measured the toxicity information.

ii) The data set and chemical selection. Notably, only 160 chemicals have been utilised from the original source of the data (Schultz et al., 2003), which contained a richer set of 385 compounds. The rationale for the selection of this smaller sample has been omitted from the publication, as well as a clear outline as for how the groups have been formed.

iii) Mechanistic understanding. The model also demonstrated high uncertainty in areas relating to the mechanistic understanding regarding substituted aromatic compounds towards *Tetrahymena pyriformis*. Mechanistic understanding allows for a greater understanding of why the approaches and methods used by the authors have been applied, therefore, lack of results in a high uncertainty being accredited.

iv) Model transparency. Full, clear and detailed explanations as to how the model was produced, especially with regard to RBFNN have not been provided. Algorithms for the models are not also not provided within the publication.

v) Reporting of the models. Full reporting of all aspects used to curate the model have not been given, specifically being the calculated descriptor values. Without the descriptor values neither the model nor predictions can be reproduced, resulting in a high uncertainty.

vi) Relevance of the endpoint. Inhibition of growth to *Tetrahymena pyriformis* is not an endpoint required for the regulatory assessment of chemicals.

A strategy to reduce the uncertainties could include:

i) Better assessment and description of the toxicity data utilised. This could include an evaluation of the methodology and possible shortcomings in terms of the effect this may have on QSAR modelling. Biological data used has been shown to be historically of high quality, with measurements being performed by a singular group, in the same laboratory and therefore being unaffected by the inter-laboratory variance component (Ruusmann and Maran, 2013). Alongside this, reproducibility of the data prior has demonstrated definable and excellent results (Hewitt et al., 2011). Uncertainty associated with nominal concentrations and internal exposure may be unimportant if the predictions from the model are being utilised within a weigh-of-evidence study, although it must be noted that this may be significant for other uses (Cronin et al., 2019).

ii) A full justification of the data set selected and why compounds were omitted from the data provided by Schultz et al (2003).

iii) Consideration of the mechanisms of toxic action could be provided, or at least postulated. There is much information in Schultz et al (2003) and related publications. For instance, an assumption of polar narcosis should be made for a data set consisting of phenols and anilines (Enoch et al., 2008). It is highly likely that nitro and other substituents will alter the mechanisms of action to being electrophilic. A useful starting place for analysis is a plot of toxicity against log P as shown in Figure S1:

Chart, scatter chart

Description automatically generatedFigure S1. Relationship between the inhibition of 160 substituted aromatic compounds to *Tetrahymena pyriformis* and log P, with inclusion of both polar and non-polar narcosis baselines. Toxicological data, polar narcosis baseline, and non-polar narcosis baseline have been provided from Luan et al. (2018), Enoch et al. (2008) and Ellison et al. (2008) respectively.

Figure S1 indicates that for some compounds in the data set toxicity relates to the baseline, these have been defined previously by Ellison et al. (2008). There is also a good relationship between toxicity and hydrophobicity for the polar narcotics, as defined by Enoch et al. (2008). Reactive compounds i.e. those capable of electrophilic interactions with biological proteins have toxicity in excess of polar narcosis. An understanding of the mechanisms of toxic action will assist in the better interpretation and justification of the models.

iv) The models are not reported in a transparent fashion e.g. RBFNN could be reported as such so they may be examined in detail and reproduced. In addition, the role of descriptors in a mechanistic context could be defined.

v) The models could be reported more fully (relating to transparency). All aspects used to curate the model specifically with regard to the calculated descriptor values should be provided within the publication or supplementary materials. All the information should be available for the reader allowing for models and predictions to be reproduced.

vi) A discussion of the role of the inhibition of growth to *Tetrahymena pyriformis* in the regulatory assessment of chemicals e.g. as part of a weight of evidence for assessing acute ecotoxicity, could be provided.

*Study 2: QSARs for lethal concentration of substituted benzenes to fathead minnow*

In Study 2 the QSARs developed by Pal et al. (2018) were assessed. These were developed multiple QSARs for the acute toxicity data to the fathead minnow (*Pimephales* promelas) of 15 alkyl and halogen substituted benzenes harvested from Bertinetto et al (2013). The toxicity data were split into three sets (of five compounds) to eliminate bias and both MLR and a backpropagation neural network (NN) used to develop the models. QSARs were developed using either hydrophobicity or electrophilicity descriptors. In general, models developed using electrophilicity were shown to perform better than that those utilising hydrophobicity. The assessment of uncertainties considered the QSARs reported as a whole, however comments could also be addressed to specific models.

Application of the assessment criteria and the resulting scores assigned is available in Table S2b. Analysis of these findings shows that high uncertainty was found mainly to be associated with:

i) The toxicity data modelled. Information regarding how the toxicity tests have been performed and consistency between them has not been included within the publication. Specific knowledge regarding the toxicity values is also unknown.

ii) The data set and chemical selection. In total, 15 chemicals have been harvested from the original source (Bertinetto et al., 2013), although this is only a very small subset of the toxicity data available for this endpoint from the original publication from Russom et al. (1997). The justification for selection of the reduced data set has not been provided.

iii) Mechanistic understanding. Description of the models have not defined any mechanisms of action for the compounds.

iv) Reporting of the models. Not all aspects of model development have been described. Specifically, there are no details relating to calculated descriptors. Due to the lack of descriptor values the models and predictions are unable to be reproduced, which lead to a high uncertainty.

A strategy to reduce the uncertainties could include:

i) Greater assessment and reporting of the toxicity data used. Approaches could include assessing the quality of the data, with respect to the methodology and potential limitations that this may have on the QSARs. The fathead minnow dataset was created in the 1980s and described fully by Russom et al. (1997), the data themselves can be considered to be reliable and robust. Nominal concentration and unknown internal exposure can affect the potential usage of the models, and so this must be taken into consideration when using predictions from this model. Although, the resulting uncertainty may be acceptable for certain uses, such as weight-of-evidence studies.

ii) Inclusion of a rationale as to why certain compounds have been omitted and so resulting in the smaller data set that has been used from Bertinetto et al (2013) and the much larger dataset from Russom et al, (1997).

iii) Mechanisms of toxic action could be provided, or at least hypothesised. The compounds in the dataset are known to act by the non-polar narcosis mechanism of action (Russom et al., 1997; Könemann, 1981). This mechanism is strongly associated with the ability to reach the site of action (i.e. the cellular membrane) and hence hydrophobicity is seen as a controlling. As such, as strong relationship is observed for the complete dataset between toxicity and log P (Figure S2). Whilst the descriptor data are not shown, there is a strong probability of a high correlation between log P and the calculated electrophilicity parameter within this particular group of compounds.

Chart, scatter chart

Description automatically generatedFigure S2. Relationship between the acute toxicity of 15 selected substituted benzenes to the fathead minnow and log P (data from Pal et al., 2018).

iv) All aspects that have been utilised to develop the model could be reported. Inclusion of this information, specifically with respect to calculated descriptors, increases the transparency allowing for models to reproduced and predictions to be made.

*Study 3: QSARs for the acute toxicity of organic compounds to fish*

Study 3 is a QSAR developed for hydroxy and halogen substituted alkyl and aryl compounds considered to act by “MOA 1”, which is a mechanism of action broadly analogous to non-polar narcosis (de Morais e Silva et al., 2018). The model was derived using fish acute toxicity data obtained from Thomas et al (2015). The model has been developed from theoretical molecular descriptors acquired from the Volsurf+ program v. 1.0.7 (<http://www.moldiscovery.com>). In general, the Volsurf descriptors used provided an understanding of the spatial contribution of the organic compounds physicochemical properties that result in fish toxicity, whilst remaining simple to interpret. PLS was used to develop the individual QSARs

Application of the assessment criteria and the resulting scores assigned is available in Table S2c. Analysis of these findings shows that high uncertainty was found mainly to be associated with:

i) The toxicity data modelled. Evidence relaying the reliability and quality of the toxicological tests and the resulting data has not been provided.

ii) The data set and division of compounds into training and test set. Replicate toxicity results for individual compounds have been utilised within the data set. Replicate toxicological values may not only hinder the training set homogeneity, but also the suitability of the training and test sets. Examination of the sets used within the study has identified duplicate chemicals within both, being a direct result of using multiple toxicity values for a single chemical. Presence of replicate structures in the training set is known to overly effect the resulting algorithm. Along with this, the resulting statistics seen with having replicates in both the training and testing set can imply a stronger predictive power of the model that is accurate (Dearden et al., 2009). In addition, the use of replicates has resulted in errors within the reporting. A suggested 61 chemicals were reported, when in fact inspection of the data set revealed a total of 58.

iii) Reporting of the models. All the information and values that have been used to develop the model have not been provided. In particular, the calculated descriptor values are not available, thus neither the model nor predicted values can be replicated resulting in a high uncertainty.

iv) The endpoint of toxicity to fish is ambiguous and therefore would not be applicable for the regulatory assessment of chemicals.

A strategy to reduce the uncertainties could include:

i) A more detailed analysis of the strengths and quality of the toxicity data used. This could be achieved through providing an evaluation of the methodology used to produce the results, and the effects this may have on the model. The origins of the toxicological data can be traced back to the ECETOC (2013), in which many compounds were extracted from the ECHA database (<http://www.echa.europa.eu>) to create the data set. For the curation, all the toxicological data underwent vigorous screening to ensure that there was a strong quality of data. This data set was then utilised by Thomas et al (2015), whom also performed various assessments on the data, removing compounds that were error prone and contained high uncertainty. Acknowledging this, the quality of the toxicological data can be considered to be high, thus eluding to this information within the report can aid in reducing the uncertainty. Attention must also be drawn to the unknown internal exposure, although an understanding of this may not be required for certain uses of the model.

ii) Removal of duplicate values within the data set, and provision of either a singular value, or an average, for one compound. Failing to do this could result in reduced quantity of data within the study, and so it must be reviewed if the quantity is suffice to curate a strong QSAR.

iii) Models could be reported fully (relating to transparency). All the information and values that have been used to curate the model could be provided within either the publication or supplementary materials. Within this case specifically the calculated descriptors should be included, which in turn will allow replication of the models and predictions.

iv) The endpoint could be better defined (in terms of species) and the regulatory relevance explained.

*Study 4: QSARs for the eco-toxicity of nonreactive organic pollutants*

Study 4 (Toropova and Toropov, 2018) involved the development of models for the eco-toxicity of non-reactive organic pollutants. The aims of the study were to estimate the potential use of the software CORAL in building predictive models for eco-toxicity and then also examine the use index of ideality (IIC) as a criterion of predictive potential for the models built.

Application of the assessment criteria and the resulting scores assigned is available in Table S2d. Analysis of these findings shows that high uncertainty was found mainly to be associated with:

i) The toxicity data modelled. Information assessing the reliability of the toxicological data has not been provided, as well as other important aspects relating to how to data was produced originally.

ii) The data set and division of compounds into training and test set. A variety of issues have arisen due to the inclusion of replicate values for individual compounds. Replicate values may not only affect the homogeneity, but also the suitability of the training and test sets. Alongside this, the strength of the predictions may be over exaggerated as a result of the replicate values (Dearden et al., 2009).

iii) Mechanistic understanding. A rationale for how the compounds produce toxicological effects has not been provided.

iv) Reporting of the models. All the values and information used to develop the QSARs have not been listed within the publication, nor supplementary materials. This has resulted in issues as well with transparency, due to neither the models nor predictions can be replicated.

v) The endpoint of toxicity to fish is ambiguous and therefore would not be applicable for the regulatory assessment of chemicals.

A strategy to reduce the uncertainties could include:

i) Greater assessment of the toxicity data utilised and further information provided about how they were produced. The data set utilised within this study has been harvested from the previous study assessed, that being de Morais e Silva et al (2018). Therefore, the strategy for reduction of uncertainties outlined previously remains valid for this study as well.

ii) Ensuring only single toxicological values are provided for each compound used within the dataset. This could be either through selection of a specific value which is known to provide the lowest uncertainty, or through taking an average of all the values provided for a single compound.

iii) Information regarding the mechanisms of toxic action that the compounds may utilise could be provided. As stated for Case Study 3, the compounds in this data set are likely to be acting by non-polar narcosis and a significant and robust relationship can be developed using log P alone (Figure S3).

Chart, scatter chart

Description automatically generatedFigure S3. Relationship between the effective concentration of 58 organic compounds to fish and log P (data from Toropova and Toropov, 2018).

iv) Full and complete reporting of the models could be provided. Specifically, the calculated descriptors could be included within either the publication or the supplementary materials. Inclusion of all aspects that have been used to develop the model can ensure that the model and prediction can be replicated enabling a greater transparency.

v) The endpoint could be better defined (in terms of species) and the regulatory relevance explained.

*Study 5: QSAR for the inhibition of growth of organic molecules to Rana temporaria*

Study 5 (Wang et al 2019) developed a QSAR model for the narcotic potency (analogous to acute toxicity) of small organic molecules to the frog (*Rana temporaria*). To achieve this, the authors employed PCLIENT, a molecule descriptor calculation software, which calculated many physiochemical parameters. A pipeline approach containing internal searching strategies was then employed to identify optimum descriptors. Once selected these were then used to build the model using MLR, PLS and support vector machines (SVM).

Application of the assessment criteria and the resulting scores assigned is available in Table S2e. Analysis of these findings shows that high uncertainty was found mainly to be associated with:

i) Definition of chemical structures. Unambiguous chemical structures have not been reported within the publication leading to uncertainty about the active molecule and isomerism.

ii) The toxicity data modelled. Information that can provide an understanding about the quality of the data that has been modelled has not been provided. Knowledge of the tests that have been conducted, and how consistent these are has also been withheld. There is considerable uncertainty about the source of the data; Wang et al. (2019) state that “*The data studied in this paper are extracted from literature*” citing Reference 2 (Budi et al., 2016). However, Budi et al. (2016) do not present any toxicity data to *R. temporaria*. Instead, Reference 3 in the paper, Abraham and Rafols (1995) does contain a dataset, which have been extracted from the original studies from Overton (1901) which were compiled by Lipnick (1989).

iii) Mechanistic understanding. Information regarding the mechanism of toxic action for the compounds has not been provided within the publication.

iv) Reporting of the models. The full list of information that has been used throughout the study to develop the model was not provided within the publication. Neither the models nor the predictions can be replicated due to this, and thus a high uncertainty is assigned.

v) Relevance of the endpoint. Inhibition of growth to *Rana temporaria* is not an endpoint required for the regulatory assessment of chemicals.

A strategy to reduce the uncertainties could include:

i) Inclusion of unambiguous chemical structures. The chemical structures could be included within the publication and presented within the same format that have been used for modelling (Piir et al., 2018).

ii) Increased assessment and reporting of the toxicity data used. The source of the data set has been referenced as Abraham and Rafols (1995), which in turn has originated from Overton (1901) and Lipnick (1989). Reanalysis of the original source in which the data has been harvested from, with greater reporting as to the quality, as well as the consistency of the results in the publication could reduce the uncertainty.

iii) Reporting or hypothesis of the mechanism(s) of toxic action the compounds are acting by. It is noted that the data set represents classic non-polar narcotic compounds (variously termed narcotics or anaesthetics etc). Figure S4 shows there to be a significant relationship with log P alone, which in itself provides a robust and meaningful model for the toxicity of these compounds.

Chart, scatter chart

Description automatically generatedFigure S4. Relationship between the inhibition of 30 alcohol organic small molecule compounds to tadpoles of *Rana temporaria* and log P (data from Wang *et al*., 2019).

iv) Reporting fully all the aspects and information that have been used to develop the model. Specifically, including the calculated descriptors within either the publication or supplementary material, enabling the models and predictions to be repeated.

v) A discussion of the role of the inhibition of growth to *Rana temporaria* in the regulatory assessment of chemicals e.g. as part of a weight of evidence for assessing ecotoxicity, could be provided.

*Study 6: QSARs for the toxicity of substituted phenols and anilines to alga*

Study 6 (Yan et al., 2019) developed a number of QSARs for different toxicological endpoints, with the toxicity data of phenols and anilines being obtained from Tugcu and Sacan et al (2018). Uniform norm-index descriptors were used within a multiple toxicity endpoint-structure model with the toxicity endpoints for the model being 96-h algae growth inhibition concentrations (IC20, IC50, NOEC, and LOEC).

Application of the assessment criteria and the resulting scores assigned is available in Table S2f. Analysis of these findings shows that high uncertainty was found mainly to be associated with:

i) Definition of chemical structures. Unambiguous chemical structures have not been provided within the publication leading to uncertainty about the active molecule and isomerism.

ii) The toxicity data modelled. A lack of information is found when attempting to identify the quality of both the toxicological tests conducted, as well as the individual toxicity data obtained.

iii) Mechanistic understanding. Throughout the publication a clear mechanism of toxic action has not been defined for the compounds.

A strategy to reduce the uncertainty could include:

i) Inclusion of unambiguous chemical structures. The chemicals should be provided either within the publication or the supplementary material in the format that has been used throughout the study by the authors.

ii) Greater assessment and reporting of the toxicity data used. The information provided in the publication is not sufficient to gauge the quality of the data. Referring back to the original source of the data this could provide a greater insight into the tests conducted and potential pitfalls. The toxicological data have been harvested from Tugcu and coworkers (2017; 2018), in which they performed the algal toxicity tests fulfilling the criteria proposed from the OECD guidelines (OECD, 2006). Understanding that the tests conducted follow these guidelines allows for a greater understanding of the quality of the test and aids in reducing uncertainty. Alongside this, through reanalysis of the original source allows for an understanding of the consistency of the data to be identified. In total, the toxicological data are, for the most part, of low uncertainty. Only the nominal concentrations used and internal exposure reduce the uncertainty, although for certain uses of the prediction this may be unimportant.

iii) Provision or suggestion for some rationale as to what the potential mechanism of toxic action could be provided. There is considerable knowledge of the action of phenols to aquatic species and they are likely to be associated with a range of mechanisms of action from polar narcosis to reactivity (Aptula et al., 2005; Cronin and Schultz, 1996; Enoch et al., 2008). In addition, tests with *Chlorella vulgaris* have shown this algal species to demonstrate a baseline effect, analogous to narcosis, as well as reactive toxicity (Cronin et al., 2004; Worgan et al., 2003). A plot of toxicity against log P for the data considered by Yang et al., (2019) is shown in Figure S5. This demonstrates a clear narcotic effect (represented by minimal toxicity) with compounds with considerable excess toxicity above the baseline. This may form a usable hypothesis for the mechanistic interpretation of the models published by Yang et al (2019).

Chart, scatter chart

Description automatically generatedFigure S5. Relationship between the inhibition of 67 substituted phenols and anilines to *Chlorella vulgaris* and log P (data from **Set A**, Yan *et al*., 2019).

*Study 7: QSAR for the hepatotoxicity of potential drug candidates*

Study 7 (He et al., 2019) developed a model to identify the hepatotoxicity of drug candidates, making use of an extensive collection of datasets retrieved from the PubMed database. Topological and physicochemical properties were then calculated for the combined dataset of 1,254 unique compounds. Eight machine learning methods were employed to develop the models, and it was statistically identified that the QSAR developed by Random Forest algorithm provided the strongest predictions.

Application of the assessment criteria and the resulting scores assigned as available in Table S2g. Analysis of these findings shows that high uncertainty was found mainly to be associated with:

i) The toxicity data modelled. Information regarding the quality of the toxicological data that has been curated from the PubMed database has not been supplied, especially with respect to the consistency.

ii) Model transparency. The resulting algorithms for each model produced from the various machine learning methods have not been included within the publication.

iii) Mechanistic understanding. A rationale suggesting or hypothesising potential mechanisms of actions for the model has not been provided.

iv) Reporting of the models. Inclusion of all the information that has been used to develop the models has not been supplied. In particular, the values for the descriptors utilised are not given. As a result, high uncertainty in the reproducibility of the models and predictions can be observed.

v) Relevance of the endpoint. Hepatotoxicity is not a regulatory endpoint for pharmaceuticals or other chemicals.

A strategy to reduce the uncertainty could include:

i) Greater assessment and reporting of the toxicity data that have been used. In general, the collection, curation and standardisation of the data have been well reported within the publication. However, information assessing the quality of the individual studies from which the dataset has been built from is not specified. Data from many studies have been gathered and the overall level of certainty in the data (as relevant to big data sets) could be provided.

ii) The algorithms for each model produced by the different construction approaches could be supplied within the supplementary material, as well as the optimal model being provided within the main body of the publication.

iii) Provision of knowledge alluding to the potential mechanisms of action of the various chemicals that have been used in the study could be provided. For instance, it may be possible to identify the prominent mechanisms from well-known groups of compounds, whilst acknowledging this may not be possible for all compounds.

iv) Full reporting of all the information that has been used to develop the model should be supplied within the publication or supplementary materials. Specifically for this study, the values for the calculated descriptors must be included. This in turn would allow for greater transparency, as well as allow for the models and predictions to be reproduced.

v) A discussion of the role of hepatotoxicity in the regulatory assessment of chemicals e.g. as part of a weight of evidence for assessing chronic toxicity, could be provided.

*Study 8: QSARs for the reproductive toxicity of potential drug candidates*

Jiang et al. (2018) developed a number of models utilising reproductive toxicological information of 1,823 organic compounds obtained from the public databases ECHA C&L Inventory (<https://echa.europa.eu/information-on-chemicals/cl-inventory-database/>) and OECD eChemPortal (<https://www.echemportal.org/echemportal/index.action>). Binary classification models were built from a total of nine chemical fingerprints in conjunction with six machine learning methods. Statistical analysis demonstrated that the strongest model was produced using the support vector machine (SVM) method with the MACCS fingerprint. The present assessment of uncertainties considered all models reported as a whole, however comments could also be addressed to specific models.

Application of the assessment criteria and the resulting scores assigned as available in Table S2h. Analysis of these findings shows that high uncertainty was found mainly to be associated with:

i) The toxicity data modelled. Information regarding the quality and consistency of the data that have been gathered is missing. Reporting that quality has been verified is stated, but knowledge on how this was performed is omitted.

ii) Model transparency. Algorithms for the models produced by the various machine learning methods are not provided within the publication.

iii) Mechanistic understanding. Knowledge of the mechanisms of action in which the compounds are producing toxicological effects is unknown.

iv) Reporting of the models. Information regarding the particular version of the descriptor software and parameters used is lacking. In addition, SMILES are only given in the tables for AP2DFP, with this information being absent for the other fingerprint calculations.

v) Relevance of the endpoint. Reproductive toxicity is a general description of many potential tests and effects.

A strategy to reduce the uncertainties could include:

i) Thorough discussion of the quality of the toxicological data could be provided. Within the publication it is reported that the data has originated from ECHA and OECD databases (<http://www.echa.europa.eu>; <https://www.oecd.org/>) with acceptance of chemicals having to adhere to strict and vigorous screening processes. Therefore, it can be understood that high quality data have been used, although information of this could be included within the publication.

ii) Inclusion of the resulting algorithms produced by the various machine learning methods could be supplied within either the supplementary material or report.

iii) Appropriate mechanisms of actions could be supplied, or at a minimum hypothesised.

iv) All information and knowledge that has been used to develop the models could be reported. The molecular fingerprint software could be better described allowing for knowledge on version and parameters to be understood. In addition, SMILES strings could be provided within all supplementary tables with fingerprint calculations, allowing for external confirmation of results.

v) A fuller definition of reproductive toxicity in terms of the underlying tests it represents could be provided.

*Study 9: QSAR-based model for the binding of small (drug-like) molecules to the androgen receptor*

Gupta and Rana (2019) produced a multilevel ensemble model for chemicals within the early stages of drug development and their respective toxicological effects. The data used were obtained from PubChem (<https://pubchem.ncbi.nlm.nih.gov/bioassay/743040>) for androgen receptor activities of 10,273 drug-like molecules, with various physicochemical properties being calculated for the set. A multilevel ensemble model was developed, with the first level performing ensemble-based classification, and the second level executing ensemble-based regression of the active drug molecules.

Application of the assessment criteria and the resulting scores assigned is available in Table S2i. Analysis of these findings shows that high uncertainty was found mainly to be associated with:

i) Definition of chemical structures. Unambiguous chemical structures have been excluded from the publication as well as an assessment of impurities or mixtures. Thus, exact structures and active molecules are unknown.

ii) The data modelled. There is little consideration of the quality of toxicological information. Data were harvested from PubChem database without clear regard to either the quality of the source or individual assessment of the assays used.

iii) Creation of the data set for QSAR modelling. A complete data set has not been included within the publication. In addition, of the 10,273 chemicals used only 461 are active, with the rest being inactive resulting in a highly imbalanced data set.

iv) Mechanistic understanding. Explanations as to the potential mechanisms of actions and / or relevance to toxicological mechanisms of the small drug-like molecules are not provided.

v) Reporting of the models. The data set used including calculations for the physicochemical properties of all molecules has been omitted from the publication. As a result, the models and predictions produced cannot be replicated, in turn leading to high uncertainty being accredited.

vi) Relevance of endpoint. Androgen receptor binding is not a regulatory endpoint.

A strategy to reduce the uncertainties could include:

i) Addition of unambiguously defined structures supplied within either the publication or supplementary material. Furthermore, impurities and/or mixtures that are present within the data could be addressed and removed if required.

ii) Greater assessment of the data that have been used. As stated within the report, the origins of the data can be accredited to the PubChem bioassays database, which contains the resulting androgen receptor activities for all chemicals used. Therefore, the quality of the data used is solely dependent upon the individuals contributing towards the database, with PubChem offering no independent quality control mechanisms (Kim et al., 2016). Acknowledgement of these issues could be addressed within the publication.

iii) Creation of a supplementary material, which includes all the relevant data, such as, structures, toxicological and physicochemical properties that have been used throughout the study. Sampling methods to achieve balanced data sets could be considered.

iv) Definitions of relevant mechanisms of actions following androgen receptor binding could be included.

v) The data set could be made available. The full reporting of all information utilised throughout the study will enable for greater transparency to be gained, as well as the models and predictions to be reproduced.

v) A discussion of how androgen receptor binding could support regulatory assessment of reproductive toxicity or endocrine disruption, e.g. as part of weight of evidence, could be provided.

*Study 10: QSAR for the estrogenic activities of persistent organic pollutants*

The aim of Study 10 was to develop a QSAR model with the ability to predict the disruptive capabilities of persistent organic pollutants (Ibrahim et al., 2019). Quantum chemical descriptors were calculated through use of the PaDEL software, and then modelled by a Genetic Function Algorithm technique producing five QSAR models. The assessment of uncertainties considered the QSARs as a whole, although comments could also be addressed to specific models.

Application of the assessment criteria and the resulting scores assigned is available in the supplementary material Table S2j. Analysis of these findings shows that high uncertainty was found mainly to be associated with:

i) The data modelled. The quality of the biological information that has been utilised to develop the models has been not been addressed. High uncertainty has been accredited throughout the assessment of the biological data as a result.

ii) Calculation of molecular descriptors. Difficulties arise when attempting to follow the process used to calculate molecular descriptors. Software that have been used and the respective methods performed during the calculations are also difficult to obtain. Consequently, accuracy and reproducibility of the calculated values is unknown.

iii) Mechanistic understanding. Sufficient supporting information for the mechanistic rationale is lacking, particular for the selected descriptors.

iv) Reporting of the models. Full description and provision of a completed data set used to develop the models are not available. In particular, values of the calculated descriptors that have been used within the modelling process are not present. Therefore, neither the models nor the predictions can be reproduced.

vi) Relevance of endpoint. Oestrogen receptor binding is not a regulatory endpoint.

A strategy to reduce the uncertainties could include:

i) Inclusion of a description and evaluation regarding the quality of the toxicological information that has been utilised. Accreditation to the original provenance of the data has been omitted from the present publication, with no clear description about the respective data also lacking. Thus, inclusion of these aspects including sufficient assessment of the quality of testing utilised can aid with reducing uncertainty.

ii) Reporting of all methods and software that has been used in the generation of molecular descriptors. Within the materials and methods, only reporting of the optimisation approaches used on the molecular structures has been stated. Inclusion of the additional methods and software that has been used within the study to generate the calculated descriptors can reduce the uncertainty.

iii) Reference to knowledge regarding the descriptors and how they relate to the mechanisms of action.

iv) Full description and inclusion of all the aspects that have been used within the development of the model. In particular, the values for the descriptors used could be included, which in turn can allow for reproducibility of the models and predictions reported.

v) A discussion of how oestrogen receptor binding could support regulatory assessment of reproductive toxicity or endocrine disruption, e.g. as part of weight of evidence, could be provided.

*Study 11: QSAR for mutagenicity of nitroaromatic compounds (NACs) to Salmonella typhimurium TA100 strain*

Study 11 has been conducted by Hao et al. (2019), in which QSAR and classification models have been developed using a set of NACs and their respective mutagenicity towards *Salmonella typhimurium* TA100 strain. Both DRAGON and quantum chemical descriptors were used to characterise the NACs, with genetic algorithm and multiple linear regression analyses then employed to screen and develop models. Qualitative classification models were also established using several machine learning methods and six molecular fingerprints. The predictive performance of the optimal model was shown to significantly outperform previously reported models in which quantum chemical descriptors are not used.

Application of the assessment criteria and the resulting scores assigned is available in Table S2k. Analysis of these findings show that high uncertainty was found mainly to be associated with:

i) The toxicity data modelled. Units of concentration associated with the toxicological data are omitted.

ii) Reporting of the models. In general, a well-described data set has been provided, although the values of calculated descriptors are not provided. Thus, neither the models nor predictions can be reproduced.

A strategy to reduce the uncertainties could include:

i) Reporting the units of the toxicological data that has been utilised.

ii) Inclusion of all aspects that have been used throughout the study in the production of the QSAR and classification models. In particular, the calculated descriptor values could be included within the publication or supplementary materials. Presence of all information that has been used throughout the development of models will enable greater transparency.

*Study 12: QFAR for cell viability of metal oxide nanoparticles (MO-NPs) to a variety of human and rat cell lines*

Study 12 was a quasi-QFAR model to predict the cell viability of various cell lines exposed to 8 MO-NPs (Ahmadi, 2020). Models were constructed using quasi-SMILES (SMILES strings that also encapsulate physicochemical properties and/or experimental conditions), and modelled through the CORAL software based upon Monte Carlo optimisation through two target functions.

Application of the assessment criteria and the resulting scores assigned is available in Table S2l. Analysis of these findings shows that high uncertainty was found mainly to be associated with:

i) Definition of chemical structures. For reporting of nanoparticles details describing chemical composition, size, shape and aspect ratio, and surface chemistry should be provided (ECHA, 2017). In addition, an assessment of either impurities or mixtures is not given.

ii) The toxicity data modelled. Sufficient information evaluating the quality of the toxicological tests conducted, and the data acquired has not been specified.

iii) Mechanistic understanding. Recognition or suggestions as to the mechanisms of actions that are being relevant to the MO-NPs are not provided.

iv) Reporting of the models. Complete description of all the information that has been used throughout the construction of the models has not been defined. Specifically, only calculated values for descriptors for the validation set are given, with other sets missing these values. Thus, neither models nor predictions can be reproduced.

v) Relevance of endpoint. Cytotoxicity is not an endpoint required for regulatory risk assessment of nanoparticles.

A strategy to reduce the uncertainties could include:

i) Full description of all aspects suggested by ECHA (2017) of each nanoparticle allowing for an unambiguous definition. Furthermore, presence of any potential impurities or mixtures within the data could be highlighted and addressed.

ii) Thorough reporting of the quality of toxicity data. Toxicological information has been gathered from five different cell lines, four of which are human and one rat. Description of the quality for each could be provided, reporting the testing procedures as well as consistency throughout. Furthermore, the use of many different assays should be discussed with respect towards the uncertainty this contributes and whether this is adequate for the intended purpose.

iii) Description of potential mechanisms of actions could be provided within the publication for the nanoparticles used. In general, MO-NPs have been reported to cause cytotoxicity through the generation of reactive oxygen and nitrogen species, as well as the formation of free ions (Fröhlich, 2013). Therefore, supplementation of this information with relation to the methodology used can aid in reducing the uncertainty.

iv) Inclusion of all information used throughout the study within the development of the models. Values for the descriptors used within the model are only provided for the validation set, therefore, addition of this information for all other sets can allow for a reduction in uncertainty. Application of this strategy will also improve confidence within transparency, due to both the models and predictions being able to be reproduced.

v) The potential role of predictions of the cytotoxicity of nanoparticles for regulatory assessment could be provided. **References *(Appendix I)***

Abraham, M.H., Rafols, C., 1995. Factors that influence tadpole narcosis. An LFER analysis. J. Chem. Soc., Perk. Trans. 2. 1995, 1843-1851.

Ahmadi, S., 2020. Mathematical modeling of cytotoxicity of metal oxide nanoparticles using the index of ideality correlation criteria. Chemosphere, 242, 125192.

Aptula, A.O., Roberts, D.W., Cronin, M.T.D., Schultz, T.W., 2005. Chemistry-toxicity relationships for the effects of di-and tri-hydroxybenzenes to *Tetrahymena pyriformis*. Chem. Res. Toxicol. 18, 844-854.

Bertinetto, C., Duce, C., Solaro, R., Héberger, K., 2013. Modeling of the acute toxicity of benzene derivatives by complementary QSAR methods. MATCH Commun. Math Comput. Chem., 70, 1005–1021.

Budi, S., Suliasih, B.A., Othman, M.S., Heng, L.Y., Surif, S., 2016. Toxicity identification evaluation of landfill leachate using fish, prawn and seed plant. Waste Manag. 55, 231-237.

Cronin, M.T.D., Netzeva, T.I., Dearden, J.C., Edwards, R., Worgan, A.D.P., 2004. Assessment and modeling of the toxicity of organic chemicals to *Chlorella vulgaris*: Development of a novel database. Chem. Res. Toxicol. 17, 545-554.

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

Cronin, M.T.D., Schultz, T.W., 1996. Structure-toxicity relationships for phenols to *Tetrahymena pyriformis*. Chemosphere 32, 1453-1468.

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

Dearden, J.C., Cronin, M.T.D., Kaiser, K.L.E., 2009. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). SAR QSAR Environ. Res. 20, 241-266.

ECETOC. 2013. Technical Report no.120. Activity-Based Relationships for Aquatic Ecotoxicology Data: Use of the Activity Approach to Strengthen MoA Predictions. ECETOC, Brussels, Belgium. Available from: https://www.ecetoc.org/publication/tr-120-activity-based-relationships-for-aquatic-ecotoxicology-data-use-of-the-activity-approach-to-strengthen-moa-predictions/

ECHA. 2017. Appendix R7-1 for nanomaterials applicable to Chapter R7a (Endpoint specific guidance). Available at: https://echa.europa.eu/documents/10162/13632/appendix\_r7a\_nanomaterials\_en.pdf

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

Ellison, C.M., Cronin, M.T.D., Madden, J.C., Schultz, T.W., 2008. Definition of the structural domain of the baseline non-polar narcosis model for *Tetrahymena pyriformis*. SAR QSAR Environ. Res. 19, 751-783.

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

Fröhlich, E., 2013. Cellular targets and mechanisms in the cytotoxic action of non-biodegradable engineered nanoparticles. Curr. Drug Metab. 14, 976-988.

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

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

He, S., Ye, T., Wang, R., Zhang, C., Zhang, X., Sun, G., Sun, X., 2019. An *in silico* model for predicting drug-induced hepatotoxicity. Int. J. Mol. Sci. 20, 1897.

Hewitt, M., Cronin, M.T.D., Rowe, P.H., Schultz, T.W., 2011. Repeatability analysis of the *Tetrahymena pyriformis* population growth impairment assay. SAR QSAR Environ. Res. 22, 621-637.

Ibrahim, I.T., Uzairu, A., Sagagi, B. 2019., QSAR, molecular docking approach on the estrogenic activites of persistent organic pollutants using quantum chemical disruptors. SN Appl. Sci. 1, 1599.

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

Kim, S., Thiessen, P.A., Cheng, T., Yu, B., Shoemaker, B.A., Wang, J., Bolton, E.E., Wang, Y., Bryant, S.H., 2016. Literature information in PubChem: associations between PubChem records and scientific articles. J. Cheminf. 8, 32.

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

Lipnick, R.L., 1989. A Quantitative Structure-Activity Relationship study of Overton’s Data on the narcosis and toxicity of organic compounds to the tadpole, *Rana temporaria*. In: Suter II, G.W., Lewis M.A. (Eds.), Aquatic Toxicity and Environmental Fate: Eleventh Volume, ASTM STP 1007, American Society for Testing and Materials, Philadelphia, pp. 468-489.

Luan, F., Wang, T., Tang, L., Zhang, S., Dias Soeiro Cordeiro, N.M., 2018. Estimation of the toxicity of different substituted aromatic compounds to the aquatic ciliate *Tetrahymena pyriformis* by QSAR approach. Molecules 23, 1002.

Organisation for Economic Co-operation and Development Guideline 201: Freshwater Alga and Cyanobacteria Growth Inhibition Test (2006) (Paris, France)

Overton E., 1901. Studien uber die Narkose, Fischer, Jena, Germany.

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

Piir, G., Kahn, I., García-Sosa, A.T., Sild, S., Ahte, P., Maran U., 2018. Best practices for QSAR model reporting: Physical and chemical properties, ecotoxicity, environmental fate, human health, and toxicokinetics endpoints. Environ. Health Persp. 126, 126001.

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

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

Schultz, T.W., Netzeva, T.I., Cronin, M.T.D., 2003, Selection of data sets for QSARs: Analyses of *Tetrahymena* toxicity from aromatic compounds. SAR QSAR Environ. Res. 14, 59–81.

Thomas, P., Dawick, J., Lampi, M., Lemaire, P., Presow, P., van Egmond, R., Arnot, J.A., Mackay, D., Mayer, P., Galay Burgos, M., 2015. Application of the activity framework for assessing aquatic ecotoxicology data for organic chemicals. Environ. Sci. Technol. 49, 12289-12296.

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

Tugu and Sacan 2017 case study 6

Tugu and Sacan 2018 case study 6

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

Worgan, A.D.P., Dearden, J.C., Edwards, R., Netzeva, T.I., Cronin, M.T.D., 2003. Evaluation of a novel short-term algal toxicity assay by the development of QSARs and inter-species relationships for narcotic chemicals. QSAR Comb. Sci. 22, 204-209.

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