Regulatory Toxicology and Pharmacology 120 (2021) 104843

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

A cross-industry collaboration to assess if acute oral toxicity (Q)SAR models are fit-for-purpose for GHS classification and labelling

Joel Bercu [a], Melisa J. Masuda-Herrera [a], Alejandra Trejo-Martin [a], Catrin Hasselgren [b], Jean Lord [n], Jessica Graham [c], Matthew Schmitz [d], Lawrence Milchak [e], Colin Owens [e], Surya Hari Lal [f], Richard Marchese Robinson [f], Sarah Whalley [f], Phillip Bellion [g], Anna Vuorinen [g], Kamila Gromek [h], William A. Hawkins [i], Iris van de Gevel [j], Kathleen Vriens [j], Raymond Kemper [k], Russell Naven [k], Pierre Ferrer [l], Glenn J. Myatt [m,*]

[a] Gilead Sciences, 333 Lakeside Drive, Foster City, CA, USA
[b] Genentech, Inc., 1 DNA Way, South San Francisco, CA, 94080, USA
[c] Bristol Myers Squibb, 1 Squibb Dr, New Brunswick, NJ, 08903, USA
[d] AbbVie Inc., North Chicago, IL, USA
[e] 3M Company, 3M Center, St. Paul, MN, 55144-1000, USA
[f] Syngenta Crop Protection, Product Safety Department, Jealott's Hill International Research Centre, Bracknell, Berkshire, RG42 6EY, UK[1]
[g] DSM Nutritional Products, Kaiseraugst, Switzerland
[h] Galapagos SASU, 102 Avenue Gaston Roussel, 93230, Romainville, France
[i] GlaxoSmithKline, Park Road, Ware, Hertfordshire, SG12 0DP, United Kingdom
[j] Janssen Pharmaceutical Companies of Johnson & Johnson, 2340, Beerse, Belgium
[k] Vertex Pharmaceuticals Inc., Discovery and Investigative Toxicology, 50 Northern Ave, Boston, MA, USA
[l] Department of Veterinary Physiology and Pharmacology, Interdisciplinary Faculty of Toxicology Program, Texas A&M University, 4466 TAMU, College Station, TX, 77843-4466, USA
[m] Leadscope (an Instem company), 1393 Dublin Rd, Columbus, OH, 43215, USA
[n] Ultragenyx, 60 Leveroni Court, Novato, CA, 94949, USA

ARTICLE INFO

Keywords:
Acute oral Toxicity (Q)SAR
In silico 3Rs Expert review
Expert rule-based Statistical-based model
Classification and labelling
CLP/GHS GHS

ABSTRACT

This study assesses whether currently available acute oral toxicity (AOT) *in silico* models, provided by the widely employed Leadscope software, are fit-for-purpose for categorization and labelling of chemicals. As part of this study, a large data set of proprietary and marketed compounds from multiple companies (pharmaceutical, plant protection products, and other chemical industries) was assembled to assess the models' performance. The absolute percentage of correct or more conservative predictions, based on a comparison of experimental and predicted GHS categories, was approximately 95%, after excluding a small percentage of inconclusive (indeterminate or out of domain) predictions. Since the frequency distribution across the experimental categories is skewed towards low toxicity chemicals, a balanced assessment was also performed. Across all compounds which could be assigned to a well-defined experimental category, the average percentage of correct or more conservative predictions was around 80%. These results indicate the potential for reliable and broad application of these models across different industrial sectors. This manuscript describes the evaluation of these models, highlights the importance of an expert review, and provides guidance on the use of AOT models to fulfill testing requirements, GHS classification/labelling, and transportation needs.

1. Introduction

The purpose of the acute oral toxicity (AOT) study is to characterize general degrees of toxicity and understand the potential for a compound to cause life-threating effects from an acute exposure. Regulatory authorities often require the AOT testing of substances in order to

* Corresponding author. Present address: Leadscope, Inc., 1393 Dublin Road, Columbus, OH, 43,215, USA.
E-mail address: gmyatt@leadscope.com (G.J. Myatt).
[1] Any example workflow or guidance outlined in this paper is not currently endorsed or approved by Syngenta.

https://doi.org/10.1016/j.yrtph.2020.104843
Received 24 September 2020; Received in revised form 19 November 2020; Accepted 14 December 2020
Available online 17 December 2020

characterize their toxicity and assign hazard categories, which informs the labelling of products to indicate appropriate restrictions or precautions to be taken during their handling, transportation, or use (Hamm et al., 2017). While the exact requirements for the content and formatting of labelling may vary by the product type, regulatory agency, and use context, there have been numerous international efforts to harmonize hazard identification, and classification and labelling over the last several decades (Strickland et al., 2018). Examples of frameworks include the United Nations (UN) Recommendations on the Transport of Dangerous Goods and the Globally Harmonized System (GHS) of Classification and Labelling of Chemicals (UN 2019a; UN 2019b). Each framework is regularly revised and updated to reflect national, regional and international experiences in implementing their requirements into laws, as well as the experiences of users who perform the classification and labelling (UN (2019a)).

AOT studies are required for the majority of compounds as part of the European Union's (EU's) legislation on the registration, evaluation, authorization and restriction of chemicals (REACH) produced at ≥ 1 tons per year and manufactured or imported in the EU or European Economic Area (EEA) (EU 2006; ECHA 2015) as well as other international compound registrations. AOT information is also utilized to define labeling information for safety data sheets (SDS) and containers as defined by the UN's GHS for classification and labelling of chemicals (i.e., the purple book, EU's Classification, Labelling and Packaging (CLP)) (UN GHS 2005; EU 2017). Finally, AOT information guides how a chemical should be packaged, labeled and, or transported (49 CFR, Part 178; 16 CFR 1500.3; IATA 2020). The well-established practice and widespread use of AOT studies for these intended purposes, as well as an overall lack of non-animal alternatives, results in the mandated necessity to continue to conduct these tests.

The median lethal dose, LD_{50}, is a general indicator of a chemical substance's acute systemic toxicity. The LD_{50} values from acute toxicity tests in rodents serve as the basis for the toxicological classification. The most commonly performed tests for acute toxicity are described in the OECD guidelines (OECD 2008) and are essentially identical to those called for under the Toxic Substances Control Act (TSCA) (TSCA 2016), Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) (FIFRA 1996), and REACH regulations. The AOT tests, including the limit test, fixed-dose procedure, toxic class method, and up-and-down methods (OECD 2002a; OECD 2002b; OECD 2008, respectively), each represent a more simplified study design compared to the original animal test method (OECD 401, which was deleted in 2002) as a means of minimizing animal use.

GHS provides an internationally compatible system to classify and communicate physical, health, and environmental hazards of a substance for the protection of humans and the environment. Several toxicological endpoints are presented in the GHS regulation to enable proper hazard classification, including acute toxicity by the oral, dermal, and/or inhalation (gases, vapors, dusts & mists) route. There are five GHS categories for acute toxicity (Category 1–5), which are banded based on the dose or concentration required to produce a severe toxic effect or death in 50% of the exposed population (i.e., LD_{50}), with Category 1 chemicals being the most toxic (see Table 1). These five acute toxicity classification categories have corresponding pictograms, signal words, and hazard statements, which are used for hazard communication on safety data sheets and chemical labels (UN GHS 2005). It should be noted that not all classification categories are adopted in all regions in the world. Regulation (EC) 1272/2008 on classification, labelling and packaging of substances and mixtures (CLP Regulation, EU 2008) has adopted Categories 1–4, whereas category 5 substances, with a low toxicity, are designated as "not classified" according to the CLP regulation.

There is a balance in toxicology research for understanding the hazards of chemicals versus the need for animal testing (OECD 2001). The "3Rs" is a global initiative geared toward reducing animal use in research and stands for (1) Replacing animal-dependent study methods with reliable/comparable alternative methods, (2) Reducing the number of animals in a study, and/or (3) Refining studies to improve animal welfare (Russel and Burch, 1959). Industry implements the 3Rs to accelerate scientific discovery, support innovation and technological developments, and address societal concerns about animal research. There are ongoing national and international efforts to employ the 3Rs across toxicology testing and gain regulatory endorsement (NC3Rs (2020); EFPIA (2019); AnimalResearch.Info (2018); Lautenberg Chemical Safety Act (2016); Tox21 (2008)). Additionally, the EU Directive 2010/63/EU mandated the application of reduction, refinement and replacement across the EU (EU Directive, 2010/63/EU).

There have been efforts to reduce the number of laboratory animals needed for the existing *in vivo* methodologies utilized for determining the AOT of compounds. The new OECD guidelines for AOT studies reduced the number of animals needed to define a point estimate while also enabling a more harmonized approach to classifying compounds based on their AOT hazard (UN GHS 2005). Introduction of a limit dose (2000 mg/kg) and a maximum tested dose (5000 mg/kg) to define "not acutely toxic", also reduced the number of animals required for compounds of low toxicity as there was no need for excessive dosing (OECD 2002a; OECD 2002b; OECD 2008; UN GHS 2005). The approval of the Fixed Dose Procedure (OECD TG 420), Acute Toxic Class (OECD TG 423) and Up and down procedure (OECD TG 425) were also considerable advances as historical studies utilized ~100 animals per study and these newer test guidelines utilize 2–15 animals per study (Erhirhie et al., 2017). In addition, the fixed dose procedure relies on clear signs of toxicity at fixed dose levels versus lethality, which reduces animals and offers a refinement that improves animal welfare (OECD 2002a).

At the time of preparing this paper, there are no validated (e.g. OECD test guidelines), internationally accepted, animal-free alternatives to the acute oral toxicity animal study that regulatory bodies accept. Based on their common use in cytotoxicity assessments, the 3T3 (mouse fibroblasts) neutral red uptake (NRU) and the NHK (human keratinocytes) NRU *in vitro* methods have been evaluated as potential alternatives to AOT testing (Creton et al., 2010; Schrage et al., 2011; OECD 2010). However, these methods were found to not be sufficiently accurate as stand-alone test methods but recommended to be incorporated as part of a weight of evidence approach for the selection of starting doses for rodent AOT tests (Creton et al., 2010). (Quantitative) structure activity relationship – or (Q)SAR – models have also not been sufficiently developed or validated to enable them to be used as stand-alone alternatives to animal testing or to classify and waive/not test in the case of REACH. However, (Q)SAR information can be used to supplement experimental test data as part of a weight of evidence or an Intelligent Testing Strategy (ITS) approach (ECHA, 2008; Creton et al., 2010).

AOT *in silico* model development is aligned with the 3Rs mission to replace existing methods that require laboratory animals. An AOT *in silico* model offers an animal-free way to elucidate a compound's acute hazards to fulfill testing requirements, classification/labelling, or transportation purposes. Fundamental to the success of a global AOT *in silico* model is a sufficiently representative, large and high-quality database and algorithms which have the capability to make reliable predictions for a broad range of chemical structures. (In the case of a statistical QSAR, the model itself would be derived from the database

Table 1
GHS classification criteria for AOT.

Acute Toxicity	Category 1	Category 2	Category 3	Category 4	Category 5	Not classified (NC)
Oral (mg/kg)	$LD_{50} \leq 5$	$5 < LD_{50} \leq 50$	$50 < LD_{50} \leq 300$	$300 < LD_{50} \leq 2000$	$2000 < LD_{50} \leq 5000$	$5000 < LD_{50}$

using an algorithm, but the manner in which any (Q)SAR makes predictions of chemical hazard may be considered an algorithm, with data not seen during the model development procedure required for external validation of the final model.) A reliable AOT *in silico* model could complement an existing laboratory study to further reduce animals or refine existing procedures. For example, an *in silico* AOT model can assist in predicting the starting dose for the OECD 420 AOT test (the only AOT test with a non-lethal endpoint), enabling the minimum number of animals to be used and avoid lethality. Another example is if the LD_{50} is predicted to be $>$ 2000 mg/kg, the limit dose can be utilized as the starting dose with greater confidence, eliminating the need for lower doses to be tested and reducing the number of animals used. In addition to use in regulatory requirements, classification and labelling, and transportation needs, a reliable AOT *in silico* tool has potential utility in early stages of research and development as an alternative to *in vivo* testing for assessing the likelihood of acute oral toxicity for a given chemical series to guide subsequent testing strategies and compound design.

If an alternative model predicts AOT as reliably as an *in vivo* study, the alternative method should be preferred and supported. When evaluating an alternative method, it should also be understood that the *in vivo* AOT test itself has a variable response (Pham et al., 2020). Variability, i.e. differences in the GHS class observed for the same chemical, has been observed in animal studies with 18%–25% of studies (depending on the route of exposure) on the same compound resulting in a different GHS category (Allen et al., 2019) and even more-so (25–27% variability; Karmaus 2018) in test sets currently under investigation as alternatives to the AOT test. *In silico* models should not show variability for the same compound, but their accuracy or apparent accuracy will necessarily be limited by the variability in the experimental data used for training and/or testing. Still, if experimental endpoint values used for training or testing were derived from multiple test results per chemical, the variability in the endpoint data could be reduced from the variability in single test results, potentially allowing *in silico* predictions to be more reliable than *individual* test results, but not more reliable than the endpoint values seen during training. Therefore, it is expected that there will be an acceptable limit on the accuracy of *in silico* predictions as has been observed with AOT responses in animal studies.

(Q)SAR[2] *in silico* models are increasingly being considered to predict specific toxicological endpoints, such as LD_{50}, based on the chemical structure alone (Lapenna et al., 2010; Drwal et al., 2014; NASEM 2015; Kleinstreuer et al., 2018). The purpose of this paper is to explore the use of *in silico* models to advance the 3Rs for AOT. This paper will assess *in silico* models against chemicals such as pharmaceuticals, pharmaceutical intermediates, plant protection products, plant protection product intermediates, metabolites, and starting materials, along with specialty chemicals submitted by manufacturers to determine their performance compared to animal models. The results will guide the use and application of *in silico* models within the framework of existing regulations such as REACH, GHS, and transportation. Specifically, the following paper outlines a cross-industry collaboration where each organization collected historical AOT experimental data and ran AOT models over these chemicals. Each collaborator shared the experimental and predicted results and an analysis of all results was performed to understand the AOT model's performance across different methodologies, across different chemical sectors and of the consensus results. In addition, an expert review of experimentally classified category 1 and category 2 results was performed to understand how such a review would support the overall workflow.

[2] The term "(Q)SAR" is as an acronym for computational models that predict a biological response (such as acute toxicity) based on the chemical structure of the test molecule. It refers to both quantitative and non-quantitative structure-activity relationships by placing the "Q" in brackets.

2. Methodology

2.1. (Q)SAR models

There are two commonly used (Q)SAR methodologies referred to as expert rule-based and statistical-based (Myatt et al., 2017). Leadscope (an Instem company) has recently developed and made available a first generation of (Q)SAR models covering both methodologies to predict GHS categories for rat acute oral toxicity (Leadscope 2020). Both methodologies use a database of over 15,000 chemicals with rat AOT results from a number of sources including the Registry of Toxic Effects in Chemical Substances, ECHA, EU's Joint Research Council's AcutoxBase, National Library Medicines (NLM) Hazardous Substances Data Bank, OECD (eChemPortal), PAI (NICEATM) and TEST (NLM ChemIDplus) (RTECS 2011; Kleinstreuer et al., 2018).

A series of individual models have been developed from this combined dataset and used to predict GHS categories (1–5 and NC). These individual statistical models or sets of expert alerts predict whether a chemical is below a specified LD_{50} threshold corresponding to the GHS cut-off values. The statistical-based models use a Partial Logistic Regression algorithm that incorporates structural features and calculated physico-chemical properties. Whilst the models have undergone subsequent development, the models build upon the approach previously reported in the literature (Yang 2005). For the expert rule-based models, a set of 2867 structural alerts were encoded that will predict whether a chemical is below a specified GHS threshold. These models are then used within a decision tree to compute a GHS category (Myatt et al., 2019).

This decision tree approach is outlined in Fig. 1 where for each individual methodology a GHS category is predicted, as well as an overall GHS category prediction derived from the individual methodologies. In Fig. 4, a chemical is predicted to be GHS category 3 using the expert rule-based approach and GHS category 4 using the statistical-based methodology. For the expert rule-based method, a set of alerts predicts whether the chemical's LD_{50} is below the 5 mg/kg threshold. Since it was not predicted to be below this threshold, a second alert set is used to determine whether the chemical is below the 50 mg/kg threshold. Again, the prediction was negative; however, a third set of alerts predicted the chemical was below the 300 mg/kg threshold. Therefore, it was predicted to be between 50 and 300 mg/kg and hence assigned to GHS category 3. A similar process was performed using a series of statistical-based models as shown in Fig. 1. In this case, the overall prediction was category 4 (LD_{50} in the range of 300–2000 mg/kg). The most conservative value (GHS category 3) was used as the final consensus model from the two methodologies.

The models allow for inspection of the underlying model information, such as feature weightings, to support an expert review. In addition, it is possible to review analogs in the database to provide additional supportive evidence, as shown in Fig. 2.

Collaborators were given access to the acute toxicity (Q)SAR models from Leadscope (Leadscope acute rat oral QSAR (v1) and alerts (v1) [System: Leadscope Model Applier v2.4]) to use in this exercise. Each collaborator collected historical information on chemicals where a rat AOT had been performed, with a Klimisch score of 1 or 2 (Klimisch et al., 1997) where possible, along with information on the study protocol, study parameters and results (for the chemicals from the plant protection product sector, 24% of compounds were retrieved from the Pesticide Properties Database (Lewis et al., 2016)). In some cases, a GHS category was derived and in other cases an LD_{50} value or range was identified. The chemicals were then loaded into the (Q)SAR software and prediction results were generated. The software calculated one of the following 8 values for each test chemical: Category 1, Category 2, Category 3, Category 4, Category 5, Not Classified (NC), Out-of-Domain, or Indeterminate. The software may generate an out-of-domain result where a chemical is sufficiently different from the training set examples to make a reliable prediction or where the model's features do not

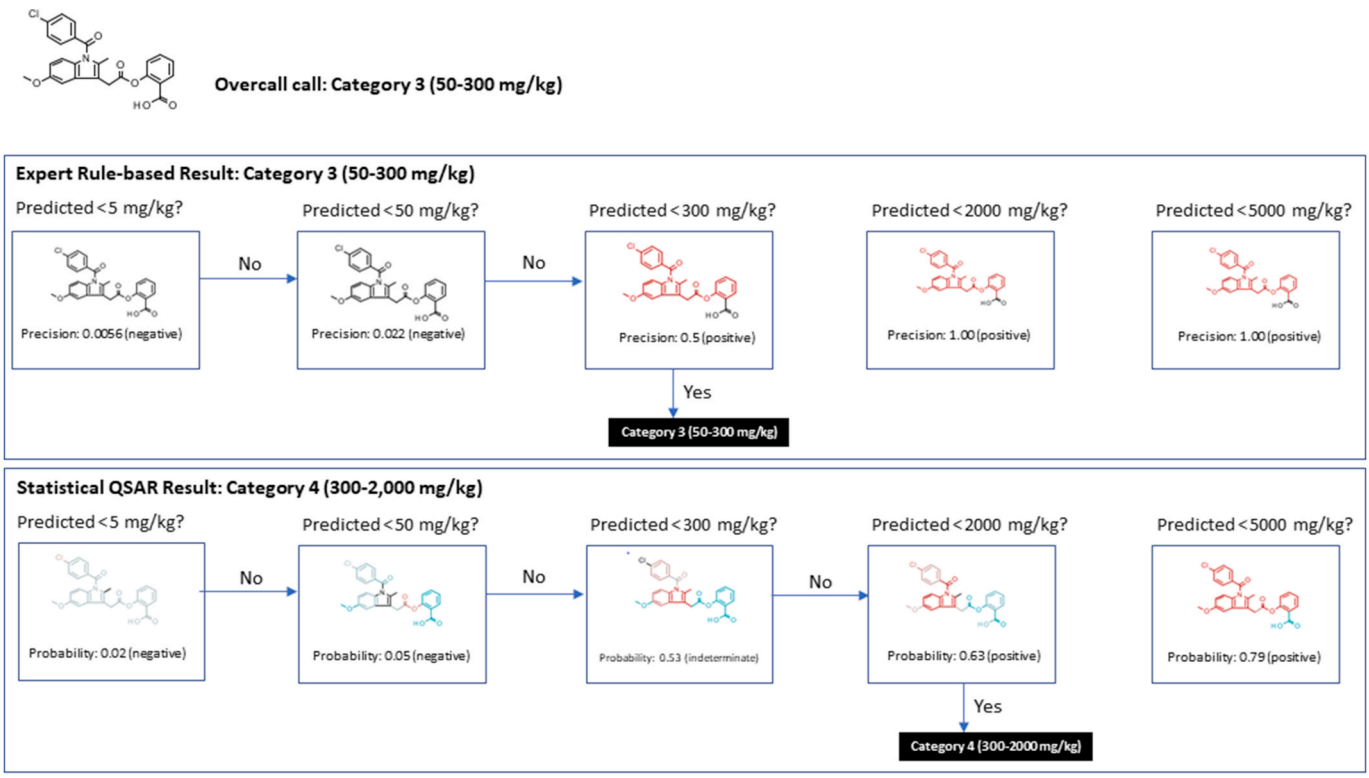

Fig. 1. Illustration of how a prediction, based on two methodologies, are computed.

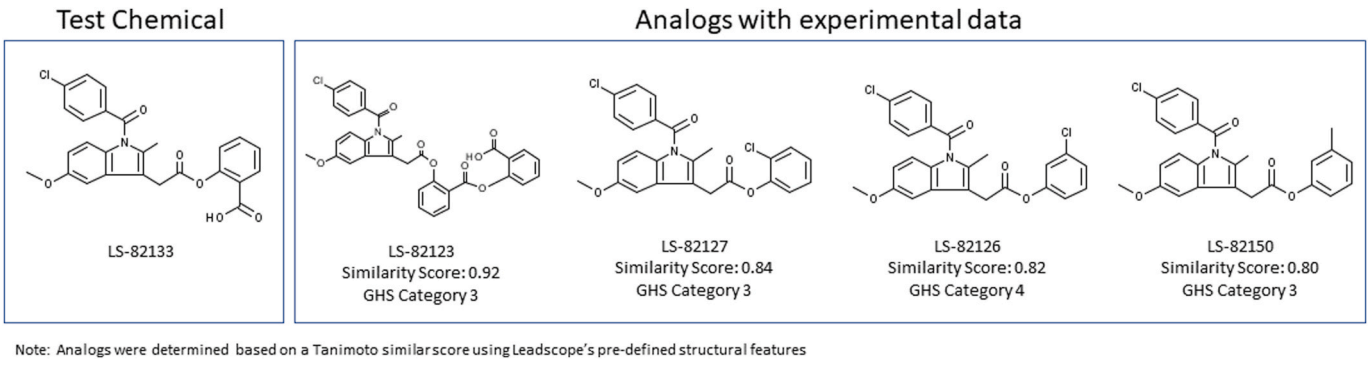

Fig. 2. Analogs of the test chemical with known GHS categories derived from *in vivo* data.

overlap with features in the test chemical. The software may also generate an indeterminate prediction where there is conflicting information, such as where the influence of substituents around a chemical class is not fully understood. Any chemical where it was determined to be part of the training set was removed. This information was then transferred to Excel spreadsheets along with relevant supporting information on the studies. To avoid sharing any potentially confidential information on the individual chemicals, all information that could provide any chemical identification was removed. However, a reference identifier was requested for each chemical in case questions needed to be resolved later.

2.2. Curating and combining the results

Each collaborator shared their *in vivo* results and predictions, as shown in Fig. 3. Initially, the individual results were analyzed to remove entries that could not be used in this exercise, based on the following rules:

- When an *in vivo* LD_{50} range was provided that spans multiple GHS categories (except for >2000 mg/kg since the 5000 mg/kg dose is often only used when it can be justified)
- In cases where it was possible to identify whether a chemical was present in the underlying model's database from the software output

In some cases, the individual collaborators provided both LD_{50} and GHS category results, in others only LD_{50} values or ranges were provided. The following rules were adopted to consistently process the data:

- When only LD_{50} values were provided, a GHS category corresponding to the LD_{50} value or range was computed
- When both an LD_{50} and GHS category were provided then the GHS category was used when justified by the collaborator

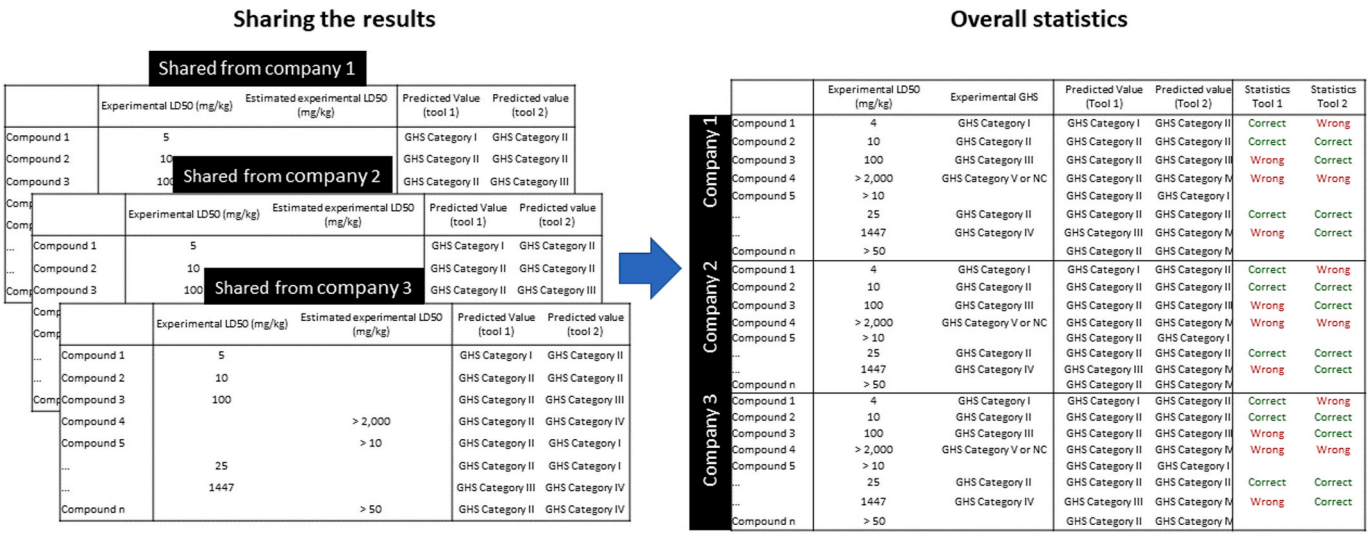

Fig. 3. Combining the results from multiple companies.

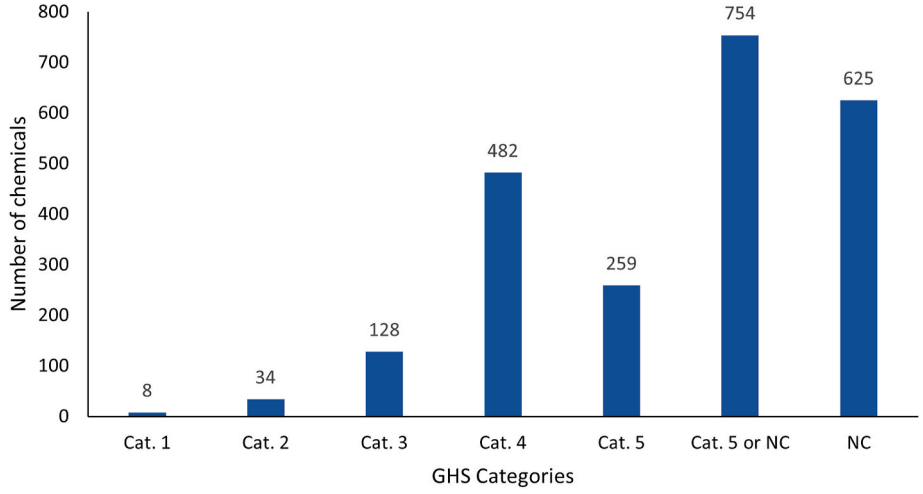

Fig. 4. Number of chemicals for each experimental *in vivo* GHS category.

- When an experimental value of >2000 mg/kg was used, a "Category 5 or Not Classified" entry was used

2.3. Generating summary statistics

The results were consolidated (as shown in Fig. 3), and a series of summary statistics were generated for the entire dataset as well as subsets including collections from the pharmaceutical industry, plant protection product industry and other chemical industries. These summary statistics use an assessment of whether the experimental *in vivo* GHS category exactly matched the predicted GHS category. In cases where the experimental category was assigned to the category "Category 5 or Not Classified", a correct match was recorded if the prediction was Category 5 or Not Classified.

A series of summary statistics were calculated to support an assessment of whether the (Q)SAR test is fit-for-purpose for classification and labeling, that is it predicts either the correct or a more potent category. This analysis was performed on both the entire data set as well as subsets of the data as explained below.

(1) The proportion of compounds correctly or more conservatively classified (for example, if the *in vivo* GHS category was 3, then a prediction of GHS 1, 2 or 3 would be a match)

(2) The proportion of compounds correctly predicted or one category more conservative (for example, if the *in vivo* GHS category was 3, then a prediction of GHS 2 or 3 would be a match)

Two additional summary statistics were computed to assess the accuracy of the models.

(3) The proportion of compounds correctly predicted (for example, if the *in vivo* GHS category was 3, then only a prediction of GHS 3 would be a match)

(4) The proportion of compounds correctly predicted or one category higher/lower (for example, if the *in vivo* GHS category was 3, then a prediction of GHS 2, 3 or 4 would be a match)

For each of these statistics, an overall assessment (i.e., the proportion across all test compounds) as well as a balanced assessment (based on the average proportion for each experimental *in vivo* GHS category) was calculated. Whilst the values derived from the overall assessment are more intuitive, the fact that the dataset was skewed towards a higher proportion of low toxicity chemicals (see below) makes the latter values more appropriate to consider.

In addition, a baseline was computed using a random model (i.e., a random uniformly distributed assignment to category 1 through 5 and

not classified) and the same balanced summary statistics generated. This was used for comparison purposes.

2.4. Expert review

An additional manual assessment of experimentally determined category 1 or 2 chemicals that were predicted by the (Q)SAR models to be in a less potent category was performed. This assessment used both information generated by the software (e.g., analogs, feature weightings) and any other information that would have been generated, including any *in vitro* assay results indicating a chemical's mechanism/mode of action (MoA). The analysis was then revised based on any modified results from this expert review.

3. Results

Results were provided from 3M, Abbvie, Bristol Myers Squibb (BMS), DSM, Genentech, Gilead Sciences, GlaxoSmithKline (GSK), Johnson and Johnson (J&J), Syngenta and Vertex. Information on 2568 chemicals was provided and, after processing the results, 2290 chemicals were used in the analysis. Given that the identities of the chemicals were not shared, it is not possible to determine whether any of the chemicals provided were duplicates; however, since these chemicals represent proprietary lead compounds, candidate active ingredients, intermediates, etc. from different companies, as well as additional marketed plant protection products and metabolites from a single database (Lewis et al., 2016), we can reasonably assume there is limited overlap because of the diverse proprietary chemical space being assessed. Any chemical where it was determined to be part of the training set was removed. Fig. 4 visually shows the number of chemicals in each of the experimental *in vivo* GHS categories. As previously noted, a category "Cat. 5 or NC" was created for chemicals where the experimental LD_{50} result was specified as $>$ 2000 mg/kg.

A summary of how the Leadscope consensus model predicted the experimental *in vivo* GHS categories is shown in Table 2. The seven experimental categories used in this analysis are listed vertically along with the six predicted categories (cat. 1–5 and NC), shown horizontally. Counts of the number of chemicals are shown in the table. To illustrate, there were 8 chemicals that had experimental *in vivo* values placing them in category 1. Five of these 8 were predicted by the consensus model as category 1, 2 were predicted as category 2 and the remaining 1 was predicted as category 5. The total value of 2181 results is less than the 2290 chemicals analyzed since 109 predictions were inconclusive (approximately 5% were either out-of-domain or indeterminate predictions). From this table, it can be seen that 95% of chemicals were either correctly predicted or were assigned to a more conservative category. However, the skewed nature of this dataset, i.e. the higher percentage of low toxic compounds, means that a balanced assessment was also required (see below).

An assessment of the performance of the consensus model for each experimental *in vivo* GHS category is shown in Table 3. Two summary statistics that help to understand whether the model is fit-for-purpose for classification and labelling are presented: (1) the percentage of correctly predicted chemicals or chemicals predicted to be in a more conservative GHS category and (2) the percentage of correctly predicted chemicals or chemicals predicted in an adjacent more conservative category. Two additional summary statistics were calculated to help understand the accuracy of the model: (1) the percentage of correctly predicted chemicals and (2) the percentage of correctly predicted chemicals or chemicals predicted in an adjacent category. The inconclusive results were not used in calculating the summary statistics.

The data collected reflects the typical distribution of GHS categories within corporate collections and as such it is highly imbalanced and weighted towards the less toxic compounds. Therefore, an overall balanced assessment of the 4 summary statistics was calculated alongside a baseline (represented by a random model). The balanced summary statistics were computed by averaging the values for each category, shown in Table 3, apart from the "Cat 5. or NC values", with the averages reported in Table 4. This information was not used in this assessment since this category spans two experimental categories.

The supplemental material contains analogous information to Tables 2–4 for the assessment of statistical-based and the expert rule-based methodologies (supplemental tables S1-S6) as well as the three industrial sectors analyzed: pharmaceutical, plant protection products and other chemicals (Supplemental tables S8-S18). As previously discussed for analysis of the consensus model on the combined dataset, due to the skewed nature of the datasets towards low toxicity chemicals, the balanced statistics presented therein provide valuable insight into the predictive performance of the different types of models on different kinds of chemicals. Table S7 summarizes the results for different (Q)SAR methodologies, statistical-based and expert rule-based, along with the consensus from the two methodologies. The same summary statistics were calculated over all the data (i.e., these values are not balanced). Table S19 summarizes the performance of the consensus model across the different sectors: pharmaceutical sector, plant protection products sector and other chemical sectors.

Supplemental tables S20, S21, and S22 show a series of experimental *in vivo* category 1 or 2 chemicals from the pharmaceutical industry, plant protection product industry and broader chemical industry that are predicted as a less conservative category. For example, a chemical whose experimental *in vivo* result is GHS category 1 yet the prediction is either category 2, 3, 4, 5 or NC. An assessment of other information that would be available for these chemicals is also provided, including other test results, information on chemical analogs as well as other information from within the deployed models. Based on this information a determination was made as to whether the chemical would have been correctly categorized based on an expert review of the totality of the information available. Using this information, Tables 5 and 6 illustrate how a combination of using the (Q)SAR models in addition to an expert review would modify the prediction results for experimental *in vivo* GHS

Table 2
Table showing counts of how the consensus model predicts for the different GHS categories.

		Predicted[c]						
Experimental		Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	NC	Total
	Cat. 1	5[a]	2	0	0	1	0	8
	Cat. 2	5	18[a]	5	2	2	1	33
	Cat. 3	1	29	52[a]	40	2	2	126
	Cat. 4	4	43	115	260[a]	38	8	468
	Cat. 5	1	15	54	106	59[a]	12	247
	Cat. 5 or NC[b]	3	48	164	343	128[a]	23[a]	709
	NC	9	32	119	227	116	87[a]	590
Total		28	187	509	978	346	133	2181

[a] Indicates where a correct prediction is made.
[b] Where chemicals were identified as $>$ 2000 mg/kg they were place in category "Cat. 5 or NC" and not in Cat.5 or NC.
[c] Not including inconclusive predictions.

Table 3
Breakdown of the results across different categories.

			Fit-for-purpose[b]		Accuracy[c]	
Experimental value	Count	Number of inconclusive predictions[a]	Percentage correct or more conservative	Percentage correct or one more conservative	Percentage correct	Percentage correct (+/− one category)
Cat. 1	8	0	62.5%	62.5%	62.5%	87.5%
Cat. 2	34	1	69.7%	69.7%	54.6%	84.9%
Cat. 3	128	2	65.1%	64.3%	41.3%	96.0%
Cat. 4	482	14	90.2%	80.1%	55.6%	88.3%
Cat. 5	259	12	95.1%	66.8%	23.9%	71.7%
Cat. 5 or NC	754	45	100.0%	69.7%	21.3%	73.2%
NC	625	35	100.0%	34.4%	14.8%	34.4%

[a] Not included in the statistics (out of domain or indeterminate).
[b] An assessment of whether the (Q)SAR test is fit-for-purpose for classification and labeling, that is it predicts either the correct or a more potent/conservative category (or predicts one category more potent/conservative).
[c] An assessment of the accuracy of the (Q)SAR test, that is the proportion of correctly predicted or±one GHS category.

Table 4
Balanced summary statistics result.

	Fit-for-purpose		Accuracy	
	Average[a] percentage correct or more conservative	Average[a] percentage correct or one more conservative	Average[a] percentage correct	Average[a] percentage correct (+/− one category)
Consensus model	80.4%	63.0%	42.1%	77.1%
Random model	54.0%	24.5%	13.1%	38.6%

[a] Averages across all experimental classes, excluding compounds in the "Cat. 5 or NC" class.

Table 5
Table showing the results of an expert review of the consensus prediction, with the original consensus prediction results shown in parentheses.

	Predicted						
Experimental	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	NC	Total
Cat. 1	7 (5)	1 (2)	0 (0)	0 (0)	0 (1)	0 (0)	8 (8)
Cat. 2	6 (5)	23 (18)	2 (5)	1 (2)	0 (2)	0 (1)	32 (33)[a]

[a] There is one compound less in the total column for GHS Cat.2 (i.e., 32 (33)) since one result (ID 703) was assigned to inconclusive after an expert review.

category 1 and 2 chemicals. Table 5 shows a table of counts for these modified results and Table 6 displays the performance metrics for these modified results. In both tables the original results (based on only the (Q)SAR models) are shown in parentheses.

4. Discussions

4.1. Expert review

An expert review of the supporting information is considered best practice to improve the overall reliability of any prediction (Myatt et al., 2018). Such a review supports an assessment of the reliability of the information as well as potentially modifying the result with sufficient supportive evidence. In most situations (as shown in Tables 5 and 6), these predictions would have been corrected based on an expert review using the following information:

- related *in vitro* assay results or information on the chemical's MoA for therapeutic or pesticidal activity
- other hazardous properties such as corrosivity
- a search for chemical analogs (e.g., structural similarity, nearest neighbors)
- chemical class considerations with known uncertainties (e.g., reactive fluorinated substances)
- examination of the additional information from the deployed model results and the underlying data
- potential downstream metabolism

These are items to consider as part of an expert review. In addition, an expert review of inconclusive results may provide additional supportive evidence to support an assignment to a GHS category.

A formalized procedure is being developed describing what specific *in silico* model results and/or other experimental data to consider as part of an acute toxicity hazard assessment. This includes recommendations for how such information should be reviewed and consolidated as part of a weight-of-evidence assessment, alongside guidelines for an expert review of this information. The protocol is being developed as part of the *in silico* toxicology protocol consortium (Myatt et al., 2018). This procedure will help ensure future predictions are performed in a consistent, documented and repeatable manner.

4.2. Performance of (Q)SAR models

The performance of the consensus model for different *in vivo* GHS categories was assessed (see Table 3). For *in vivo* GHS category 1 or 2 chemicals, the proportion of correct or a more conservative prediction was over 60%; however, when an expert review was taken into consideration this number increases to approximately 90% (see Table 6). For category 3, although 65.1% were predicted correct or more conservative, 96% were predicted to be in a correct or adjacent category

Table 6
Performance metrics showing the results of an expert review of the consensus prediction, with the original consensus performance metrics without expert review shown in parentheses.

			Fit-for-purpose		Accuracy	
Experimental value	Count	Number of inconclusive results	Percentage correct or more conservative	Percentage correct or one more conservative	Percentage correct	Percentage correct (+/− one category)
Cat. 1	8	0	87.5% (62.5%)	87.5% (62.5%)	87.5% (62.5%)	100% (87.5%)
Cat. 2	34	1	90.6% (69.7%)	90.6% (69.7%)	71.9% (54.6%)	96.9% (84.9%)

(i.e., either category 2, 3, or 4). For all other categories, the percentage of correct or more conservative predictions was greater than 90%.

Experimental *in vivo* GHS 1 and 2 categories had a low number of compounds compared to the other classes which indicates that chemicals do not generally fit in the higher potency classes with most pharmaceutical, plant protection product, and other industrial chemicals typically falling in GHS category 3–5 or NC. Since there were fewer chemicals within the higher potency categories, a series of balanced summary statistics were computed to assess whether the consensus model was fit-for-purpose (i.e., predicting the *in vivo* GHS category or a more conservative category) as shown in Table 4.

Both statistical-based and expert rule-based methodologies were individually assessed and able to predict either the correct category or a more conservative category for over 90% of the chemicals (where a prediction was made) (see Table S7), with a balanced statistic of over 73% (see supplemental tables S3 and S6). A consensus prediction from both methodologies was also calculated and this prediction had the highest score for correct or more conservative. The statistical-based model was more accurate with approximately 80% of the chemicals being correctly predicted or predicted to be in an adjacent class (either higher or lower), with the same balanced statistic value of 80% (see supplemental table S3). Therefore, all three results could be used in different ways for classification and labelling. For example, the consensus prediction may be used, in a regulatory context, to assess what GHS category to use based on the model results (since this is most conservative); however, the statistical-based model may provide more weight to determine whether additional testing is warranted (since this is the more accurate model). Hence, these models could be utilized in different manners depending on the final intended use of the prediction: screening or classification. Although predicting a more conservative value is protective of public health, other considerations (such as the cost of supporting a category 1 assignment) may also influence whether additional testing is needed for those chemicals predicted in the most toxic categories. The summary statistics include an assessment of the prediction of the correct or one more conservative category to support these decisions.

The consensus model predictions were investigated across different industries, i.e., pharmaceutical, plant protection product and other industrial chemicals, including specialty chemicals (see Supplemental tables S8-S19). The proportion of correct or more conservative predictions across all three sources of data was greater than 93% (see supplemental Table S19) indicating a high reliability for the (Q)SAR models (with balanced statistic greater than 75%, after excluding an unreliable statistic for pharmaceutical category 1 compounds based on only two datapoints - see supplemental tables S10, S15 and S18).

Several AOT *in silico* models have been assessed as part of a publication by Graham and co-authors (Graham et al., 2020). This paper illustrates the accuracy, reliability, and applicability of these models in the pharmaceutical chemical space. Graham et al. also elucidates how to utilize these models to fill in data gaps, inform decisions regarding Dangerous Goods classifications and to reduce animal use and reliance on animal test methods for acute oral toxicity GHS categorization.

4.3. Regulatory experience of using (Q)SARs

Other research and development as well as regulatory use cases have successfully incorporated (Q)SAR model results in place of *in vivo* and *in vitro* studies. For example, the ICH M7 regulatory guideline (ICH M7 2017) and the EFSA guidance (EFSA 2016) recommend, for certain kinds of chemical species, the use of two complementary (Q)SAR models, one statistical-based and one expert rule-based. (Q)SAR model results alongside an expert review are accepted as part of regulatory submissions as per these guidelines. Where a mutagenic (Q)SAR prediction is made, it is possible to follow-up this finding with an Ames test and a negative result of this *in vitro* test would then override any positive (Q)SAR prediction. This mirrors the findings in this paper.

Regulatory acceptance has also provided impetus for the development of improved models for predicting bacterial mutation. Landry et al. (2019) shows how, based on a larger training set and improved knowledge of mutagenicity SAR, improvements to both the sensitivity and specificity of the models have been made. A series of papers have been written outlining best practices in the application of the models alongside guidelines for performing an expert review of the results (Powley 2015; Barber 2015; Amberg 2016, 2019). In addition, predictions within specific classes, such as nitrosamines (Bercu et al., 2020) and aromatic primary amines (Ahlberg et al., 2016) are still challenging and are the focus of active R&D developments. These classes also require expert review. This situation parallels the findings of this paper where specific classes, such as reactive fluorinated substances, were singled out for a more in-depth expert review in supplemental Table S22.

4.4. Use cases and workflows

A (Q)SAR assessment of AOT could be utilized to support both transportation and worker safety assessments as well as emergency overdose situations (e.g. poison control) or health hazard assessments for large-scale spills of chemicals that lack AOT data. An example flow-chart for making an AOT GHS assessment based, in part, on (Q)SAR models is shown in Fig. 5.

The first step for any test chemical is to identify whether AOT data are available, either within proprietary databases or through a search of publicly available information. Such a search should return chemicals that match the chemical exactly (including different salt forms, tautomers, etc.). In many situations, the test chemical cannot be submitted to an online service because of intellectual property concerns and so issuing such a query behind a company's firewall is often important. Ideally, such a search will return an adequate AOT study, including information on species tested, route of exposure, and LD_{50} value. Studies not considered adequate or performed via other routes of exposure may be considered as part of the weight of the evidence in the expert review, discussed later. Further assessment of any available studies should focus on whether they are reliable including consideration of whether the Klimisch score (Klimisch et al., 1997) is 1 or 2 (i.e., a well-documented and accepted/sufficient study or data from the literature that is performed according to or partially compliant with valid and/or accepted test guidelines, and preferably performed according to good laboratory practices). Regarding the species tested, the rat is the preferred test species (because of the similarity of the genome between rats and humans); however, if AOT data are available in other animal species, expert scientific judgment should be used to select the most appropriate LD_{50}. Such LD_{50} data could be used directly to assign the chemical to a GHS category.

In situations where there are no AOT studies available, then it may be possible to use other repeated dose studies to derive an estimate for LD_{50} or to separate non-classified chemicals from those requiring follow-up (Bulgheroni et al., 2009; Graepel et al., 2016).

In the absence of reliable AOT experimental data or the ability to derive an LD_{50} value from other information, a (Q)SAR assessment provides an alternative approach to estimate the GHS category. In this paper, we observed that accepting the prediction with the most toxic outcome from two complementary (Q)SAR methodologies provided the most conservative overall results, which is desirable to protect health. Further, the importance of conducting an expert review for any assessment is recognized. Such a review may take into consideration other information on the chemical's MoA, other hazardous properties, chemical analogs (i.e., read-across), inspection of the individual model's results, and any mechanistic information including the potential of the chemical to metabolize. In addition, AOT data not deemed to be sufficiently reliable, on their own, may be included in a weight-of-evidence assessment. The expert review process should generate a documented assessment of the assigned GHS category. It may be possible to generate an assessment even if the (Q)SAR models are unable to predict a GHS

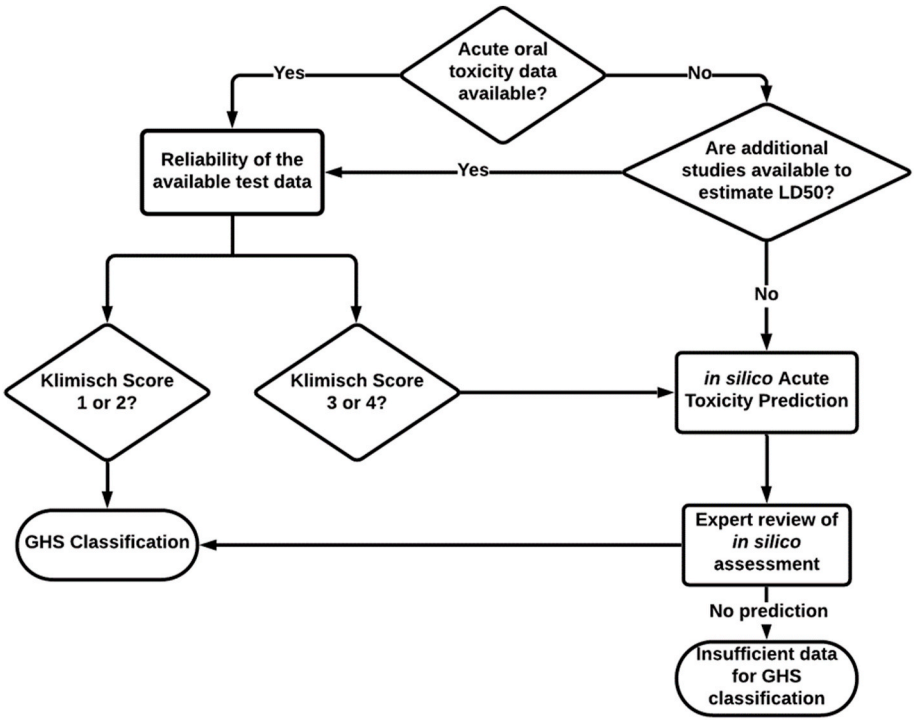

Fig. 5. Establishing a GHS classification based on available data and (Q)SAR model results.

category, such as when a chemical is out-of-domain (i.e., the prediction reliability is expected to be lower) or indeterminate (i.e., there is conflicting information) based on sufficient additional information.

Finally, there may be situations when the (Q)SAR results and expert review does not result in a GHS classification, primarily due to insufficient information. In this situation, another option for hazard identification should be considered.

In addition to their use in establishing a GHS classification, these types of models also have utility for other applications. For example, in early stage research and development (R&D) they could be used as a guide for relative acute toxicity risk and used to help design testing strategies as well as to inform compound design and selection. Different use cases for acute toxicity computational models are also outlined as part of the *in silico* protocol for acute toxicity.

5. Conclusions

As the current standard for acute oral toxicity hazard identification is a test conducted using animals, an AOT *in silico* model potentially offers a rapid and cost-effective alternative approach. *In silico* models have the potential to effectively reduce or eliminate the use of *in vivo* testing, thereby reducing the reliance of industry on these models for AOT hazard identification. Given that *in silico* models have been developed based on the wealth of publicly available AOT data, it is promising to note that the Leadscope AOT suite was capable of making typically reliable AOT hazard predictions for a broad range of chemical structures, spanning numerous industries. The evaluation presented in this manuscript also points out the importance of an expert review to enable a weight of evidence approach. Guidance is also provided on the use of such models to fulfill regulatory requirements, classification and labelling, and transportation needs. In addition, other uses for such models include prioritization and screening of chemicals in early R&D. It can be concluded that for predicting acute toxicity, the use of qualified and transparent (Q)SAR models, such as the Leadscope AOT suite, coupled with an expert review, provides a scientifically rational, reasonable and conservative approach to hazard identification.

Declaration of competing interest

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

Acknowledgements

Research reported in this publication was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award Number R44ES026909. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Pierre Ferrer was supported, in part, through the NIH training grant T32 ES026568.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yrtph.2020.104843.

References

16 CFR 1500.3. Code of Federal Regulations16 CFR 1500.3. Code of Federal Regulations. Federal hazardous substances Act (FHSA) requirements. https://www.cpsc.gov/Business–Manufacturing/Business-Education/Business-Guidance/FHSA-Requirements.

49 CFR, Part 178. Code of Federal Regulations. Title: part 178 - specifications for packagings. https://www.govinfo.gov/content/pkg/CFR-2019-title49-vol3/xml/CFR-2019-title49-vol3-part178.xml.

Ahlberg, E., Amberg, A., Beilke, L.D., Bower, D., Cross, K.P., Custer, L., Ford, K.A., Gompel, J.V., Harvey, J., Honma, M., Jolly, R., Joossens, E., Kemper, R.A., Kenyon, M., Kruhlak, N., Kuhnke, L., Leavitt, P., Naven, R., Neilan, C., Quigley, D.P., Shuey, D., Spirkl, H.-P., Stavitskaya, L., Teasdale, A., White, A., Wichard, J., Zwickl, C., Myatt, G.J., 2016. Extending (Q)SARs to incorporate proprietary knowledge for regulatory purposes: a case study using aromatic amine mutagenicity. Regul. Toxicol. Pharmacol. 77, 1–12. https://doi.org/10.1016/j.yrtph.2016.02.003.

Allen, C.H.G., Mervin, L.H., Mahmoud, S.Y., Bender, A., 2019. Leveraging heterogeneous data from GHS toxicity annotations, molecular and protein target descriptors and Tox21 assay readouts to predict and rationalise acute toxicity. J. Cheminf. 11, 36. https://doi.org/10.1186/s13321-019-0356-5.

Amberg, A., Beilke, L., Bercu, J., Bower, D., Brigo, A., Cross, K.P., Custer, L., Dobo, K., Dowdy, E., Ford, K.A., Glowienke, S., Gompel, J.V., Harvey, J., Hasselgren, C.,

Honma, M., Jolly, R., Kemper, R., Kenyon, M., Kruhlak, N., Leavitt, P., Miller, S., Muster, W., Nicolette, J., Plaper, A., Powley, M., Quigley, D.P., Reddy, M.V., Spirkl, H.-P., Stavitskaya, L., Teasdale, A., Weiner, S., Welch, D.S., White, A., Wichard, J., Myatt, G.J., 2016. Principles and procedures for implementation of ICH M7 recommended (Q)SAR analyses. Regul. Toxicol. Pharmacol. 77, 13–24. https://doi.org/10.1016/j.yrtph.2016.02.004.

Amberg, A., Andaya, R.V., Anger, L.T., Barber, C., Beilke, L., Bercu, J., Bower, D., Brigo, A., Cammerer, Z., Cross, K.P., Custer, L., Dobo, K., Gerets, H., Gervais, V., Glowienke, S., Gomez, S., Gompel, J.V., Harvey, J., Hasselgren, C., Honma, M., Johnson, C., Jolly, R., Kemper, R., Kenyon, M., Kruhlak, N., Leavitt, P., Miller, S., Muster, W., Naven, R., Nicolette, J., Parenty, A., Powley, M., Quigley, D.P., Reddy, M.V., Sasaki, J.C., Stavitskaya, L., Teasdale, A., Trejo-Martin, A., Weiner, S., Welch, D.S., White, A., Wichard, J., Woolley, D., Myatt, G.J., 2019. Principles and procedures for handling out-of-domain and indeterminate results as part of ICH M7 recommended (Q)SAR analyses. Regul. Toxicol. Pharmacol. 102, 53–64. https://doi.org/10.1016/j.yrtph.2018.12.007.

AnimalResearch.Info, 2018. http://www.animalresearch.info/en/designing-research/alternatives-and-3rs/.

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

Bercu et al., Compound- and class-specific limits for common impurities in pharmaceuticals, currently being prepared.

Creton, S., Dewhurst, I.C., Earl, L.K., Gehen, S.C., Guest, R.L., Hotchkiss, J.A., Indans, I., Woolhiser, M.R., Billington, R., 2009. Acute toxicity testing of chemicals—opportunities to avoid redundant testing and use alternative approaches. Crit. Rev. Toxicol. 40, 50–83. https://doi.org/10.3109/10408440903401511.

Drwal, M.N., Banerjee, P., Dunkel, M., Wettig, M.R., Preissner, R., 2014. ProTox: a web server for the in silico prediction of rodent oral toxicity. Nucleic Acids Res. 42 https://doi.org/10.1093/nar/gku401.

ECHA, 2008. Guidance on Information Requirements and Chemical Safety Assessment Chapter R.6: QSARs and Grouping of Chemicals. https://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf/77f49f81-b76d-40ab-8513-4f3a533b6ac9.

ECHA, 2015. Guidance on the Application of the CLP Criteria Guidance to Regulation (EC) No 1272/2008 on Classification, Labelling and Packaging (CLP) of Substances and Mixtures. https://echa.europa.eu/documents/10162/23036412/clp_en.pdf/58b5dc6d-ac2a-4910-9702-e9e1f5051cc5.

EFPIA, 2019. Putting Animal Welfare Principles and 3Rs into Action - European Pharmaceutical Industry Report. 2019 Update. European Federation of Pharmaceutical Industries and Associations (EFPIA).

EFSA, 2016. Guidance on the Establishment of the Residue Definition for Dietary Risk Assessment. https://doi.org/10.2903/j.efsa.2016.4549.

Erhirhie, E.O., Ihekwereme, C.P., Ilodigwe, E.E., 2018. Advances in acute toxicity testing: strengths, weaknesses and regulatory acceptance. Interdiscipl. Toxicol. 11, 5–12. https://doi.org/10.2478/intox-2018-0001.

EU, 2006. 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). http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02006R1907-20161011&from=EN.

EU, 2008. Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on Classification, Labelling and Packaging of Substances and Mixtures, Amending and Repealing Directives 67/548/EEC and 1999/45/EC, and Amending Regulation. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32008R1272.

EU, 2017. Guidance on the Application of the CLP Criteria Guidance to Regulation (EC) No 1272/2008 on Classification, Labelling and Packaging (CLP) of Substances and Mixtures Version 5.0 July 2017. https://echa.europa.eu/documents/10162/23036412/clp_en.pdf/58b5dc6d-ac2a-4910-9702-e9e1f5051cc5.

EU Directive 2010/63/EU. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:276:0033:0079:EN:PDF.

FIFRA, 1996. Federal Insecticide, Fungicide, and Rodenticide Act. https://www.epa.gov/laws-regulations/summary-federal-insecticide-fungicide-and-rodenticide-act.

Graham, J., Rodas, M., Hillegass, J., Schulze, G., 2020. The performance, reliability and potential application of in silico models for predicting the acute oral toxicity of pharmaceutical compounds. Regul. Toxicol. Pharmacol., 104816

Hamm, J., Sullivan, K., Clippinger, A.J., Strickland, J., Bell, S., Bhhatarai, B., Blaauboer, B., Casey, W., Dorman, D., Forsby, A., Garcia-Reyero, N., Gehen, S., Graepel, R., Hotchkiss, J., Lowit, A., Matheson, J., Reaves, E., Scarano, L., Sprankle, C., Tunkel, J., Wilson, D., Xia, M., Zhu, H., Allen, D., 2017. Alternative approaches for identifying acute systemic toxicity: moving from research to regulatory testing. Toxicol. Vitro 41, 245–259. https://doi.org/10.1016/j.tiv.2017.01.004.

IATA 202. https://www.iata.org/.

ICH, M.7, 2017. Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk, p. R1. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Multidisciplinary/M7/M7_R1_Addendum_Step_4_31Mar2017.pdf.

Karmaus, A.L., April 11, 2018. (National Toxicology Program). Rat Oral Acute Toxicity Database and Evaluation of Variability. Predictive Models for Acute Oral System Toxicity Workshop.

Kleinstreuer, N.C., Karmaus, A.L., Mansouri, K., Allen, D.G., Fitzpatrick, J.M., Patlewicz, G., 2018. Predictive models for acute oral systemic toxicity: a workshop to bridge the gap from research to regulation. Computational Toxicology 8, 21–24. https://doi.org/10.1016/j.comtox.2018.08.002.

Klimisch, H.J., Andreae, M., Tillmann, U., 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul. Toxicol. Pharmacol. 25, 1–5. https://doi.org/10.1006/rtph.1996.1076.

Landry, C., Kim, M.T., Kruhlak, N.L., Cross, K.P., Saiakhov, R., Chakravarti, S., Stavitskaya, L., 2019. Transitioning to composite bacterial mutagenicity models in ICH M7 (Q)SAR analyses. Regul. Toxicol. Pharmacol. 109, 104488. https://doi.org/10.1016/j.yrtph.2019.104488.

Lapenna, S., Fuart-Gatnik, M., Worth, A., 2010. Review of QSAR Models and Software Tools for Predicting Acute and Chronic Systemic Toxicity. Office of the European Union, Luxembourg (JRC Technical Report EUR 24639 EN.

Lautenberg Chemical Safety Act (2016). https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/frank-r-lautenberg-chemical-safety-21st-century-act.

Leadscope 2020. https://www.leadscope.com/index.php.

Lewis, K.A., Tzilivakis, J., Warner, D., Green, A., 2016. An international database for pesticide risk assessments and management. Hum. Ecol. Risk Assess. 22 (4), 1050–1064. https://doi.org/10.1080/10807039.2015.1133242.

Myatt, G., Beilke, L., Cross, K., 2017. In silico tools and their application. Comprehensive Medicinal Chemistry III 156–176. https://doi.org/10.1016/b978-0-12-409547-2.12379-0.

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

Myatt, G.J., Bower, D., Cross, K., Johnson, C., Quigley, D.Q., Tice, R., Zwickl, C., 8-11 September 2019. *In Silico* Acute Toxicity Protocols and Models (Poster #747), 55th Congress of the European Societies of Toxicology. Finland, Helsinki.

NASEM, 2015. Application of Modern Toxicology Approaches for Predicting Acute Toxicity for Chemical Defense. National Academies Press, Washington, D.C. https://www.ncbi.nlm.nih.gov/books/NBK321419/#sec_000055.

NC3Rs, 2020. National centre for the replacement refinement & reduction of animals in research (NC3Rs). https://www.nc3rs.org.uk/3rs-toxicology-and-regulatory-sciences.

OECD, 2001. OECD Series on Testing and Assessment Number 24 Guidance Document on Acute Oral Toxicity Testing. https://ntp.niehs.nih.gov/iccvam/suppdocs/feddocs/oecd/oecd-gd24.pdf.

OECD, 2002a. Test No. 420. In: Acute Oral Toxicity - Fixed Dose Procedure, OECD Guidelines for the Testing of Chemicals, vol. 4. OECD Publishing, Paris. https://doi.org/10.1787/9789264070943-en.

OECD, 2002b. Test No. 423. In: Acute Oral Toxicity - Acute Toxic Class Method, OECD Guidelines for the Testing of Chemicals, Section, vol. 4. OECD Publishing, Paris. https://doi.org/10.1787/9789264071001-en.

OECD, 2008. Test No. 425. In: Acute Oral Toxicity: Up-And-Down Procedure, OECD Guidelines for the Testing of Chemicals, Section, vol. 4. OECD Publishing, Paris. https://doi.org/10.1787/9789264071049-en.

OECD, 2010. Guidance Document on Using Cytotoxicity Tests to Estimate Starting Doses for Acute Oral Systemic Toxicity Tests. 20-Jul-2010. Series on Testing and Assessment No. 129. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2010)20&doclanguage=en.

Pham, L.L., Watford, S.M., Pradeep, P., Martin, M., Thomas, R., Judson, R.S., Setzer, R. W., Friedman, K.P., 2020. Variability in in vivo studies: defining the upper limit of performance for predictions of systemic effect levels. Computational Toxicology 15, 100126. https://doi.org/10.1016/j.comtox.2020.100126.

Powley, M.W., 2015. (Q)SAR assessments of potentially mutagenic impurities: a regulatory perspective on the utility of expert knowledge and data submission. Regul. Toxicol. Pharmacol. 71, 295–300. https://doi.org/10.1016/j.yrtph.2014.12.012.

RTECS, 2011. http://www.cdc.gov/niosh/rtecs/default.html.

Russell, W.M.S., Burch, R.L., 1959. The Principles of Humane Experimental Technique. Universities Federation for Animal Welfare, Wheathampstead.

Schrage, A., Hempel, K., Schulz, M., Kolle, S.N., van Ravenzwaay, B., Landsiedel, R., 2011 Jul. Refinement and reduction of acute oral toxicity testing: a critical review of the use of cytotoxicity data. Altern Lab Anim 39 (3), 273–295. https://doi.org/10.1177/026119291103900311.

Strickland, J., Clippinger, A.J., Brown, J., Allen, D., Jacobs, A., Matheson, J., Lowit, A., Reinke, E.N., Johnson, M.S., Quinn, M.J., Mattie, D., Fitzpatrick, S.C., Ahir, S., Kleinstreuer, N., Casey, W., 2018. Status of acute systemic toxicity testing requirements and data uses by U.S. regulatory agencies. Regul. Toxicol. Pharmacol. 94, 183–196. https://doi.org/10.1016/j.yrtph.2018.01.022.

Tox21, 2008. https://www.epa.gov/chemical-research/toxicology-testing-21st-century-tox21.
TSCA, 2016. Toxic Substances Control Act (TSCA). https://www.congress.gov/bill/114th-congress/senate-bill/697/all-info.
UN, 2019a. United Nations Globally Harmonized System of Classification and Labelling of Chemicals Eighth, Revised Edition.
UN, 2019b. United Nations Recommendations on the Transport of Dangerous Goods, Revised Edition.
UN GHS, 2005. Globally Harmonized System of Classification and Labelling of Chemicals (GHS) ("The Purple Book"). United Nations, 2005 First Revised Edition. www.unece.org/trans/danger/publi/ghs/ghs_rev01/01files_e.html or from United Nations Publications (publications@un.org.
Yang, C., Cross, K., Myatt, G.J., Paul, E., Blower, P.E., Rathman, J.F., 2004. Building predictive models for protein tyrosine phosphatase 1B inhibitors based on discriminating structural features by reassembling medicinal chemistry building blocks. J. Med. Chem. 47, 5984–5994.