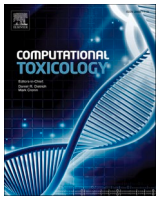
Computational Toxicology 35 (2025) 100374

Contents lists available at ScienceDirect

Computational Toxicology

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



Full Length Article

Conservative consensus QSAR approach for the prediction of rat acute oral toxicity

Jerry Achar [a,*], James W. Firman [b], Mark T.D. Cronin [b]

[a] Institute for Resources Environment, and Sustainability, The University of British Columbia, 2202 Main Mall, Vancouver, BC V6T 1Z4, Canada
[b] School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK

ARTICLE INFO

Keywords:
Quantitative structure–activity relationship (QSAR)
Rat oral acute toxicity
LD_{50}
Consensus model
Conservative prediction
Human-health protective

ABSTRACT

Consensus approaches are applied in different quantitative structure–activity relationship (QSAR) modeling contexts based on the assumption that combining individual model predictions will improve prediction reliability. This study evaluated the performance of TEST, CATMoS and VEGA models for prediction of oral rat LD_{50}, both individually and in consensus, across a dataset of 6,229 organic compounds. Predicted LD_{50} values from the models were compared for each compound, and the lowest value was assigned as the output of the conservative consensus model (CCM). Predictive accuracy was then evaluated based on the agreement of predicted LD_{50}-based GHS category assignments with those derived experimentally. The aim was to allow for the most conservative value to be identified. Results showed that CCM had the highest over-prediction rate at 37 %, compared to TEST (24 %), CATMoS (25 %) and VEGA (8 %). Meanwhile, its under-prediction rate was lowest at 2 %, relative to TEST (20 %), CATMoS (10 %) and VEGA (5 %). Due to the method applied, CCM was the most conservative across all GHS categories. Further, structural analysis demonstrated that no specific chemical classes or functional groups were consistently underpredicted or overpredicted. The utility of CCM lies in its ability to establish a foundation for contextualizing the general use of consensus modeling, in order to derive health-protective oral rat LD_{50} estimates under conditions of uncertainty, especially where experimental data are limited or absent.

1. Introduction

With advances in toxicology aimed towards the replacement of animal testing with suitable alternatives, quantitative structure–activity relationship (QSAR) models have been developed in order to support the prediction of a variety of relevant endpoints. Acute oral toxicity (measured by the lethal dose that kills 50 % (LD_{50}) of test animals) is one such metric for which a number of models have been developed [1–3]. Currently, rat LD_{50} data are commonly used as the primary benchmark to, for example, establish acceptable human exposure limits, to guide the classification of chemical hazards, to assess the potential risk of accidental ingestion of chemical toxicants, or else to set appropriate doses for repeat dose toxicity assessments [4,5].

QSARs for rat acute toxicity are commercially and publicly available [6,7]. In this investigation, three of these models – Toxicity Estimated Software (TEST), Collaborative Acute Toxicity Modeling Suite (CATMoS) and Virtual models for property Evaluation of chemicals within a Global Architecture (VEGA) – were considered for reasons including their availability without cost or licence restriction. They have a history of use in regulatory contexts within frameworks such as the US Toxic Substances Control Act (TSCA), the European Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) regulation, and the Canadian Chemical Management Plan [6–8]. Specifically, CATMoS has been proposed within the 2023 (and ongoing) review of the REACH Annex VII as a potential replacement to the acute oral toxicity test [9]. The models are trained on large and diverse chemical datasets, making them applicable to a wide variety of chemical compounds. Moreover, they employ consensus approaches by integrating multiple QSAR techniques, thereby potentially enhancing their predictive reliability.

TEST (containing Hierarchical clustering, Nearest neighbor, and Consensus methods) was developed by the United States Environmental Protection Agency (US EPA) through a combination of different techniques – for example, in the case of Hierarchical clustering method, using a genetic algorithm-based technique for generating models training clusters or, in the case of Consensus method, taking an average of the individual model predictions (provided the predictions are within each model's applicability domains) [8]. CATMoS (a consensus-based tool consisting of binary, categorical, and continuous models),

* Corresponding author.
E-mail address: jerry.achar@ubc.ca (J. Achar).

https://doi.org/10.1016/j.comtox.2025.100374
Received 24 October 2024; Received in revised form 12 August 2025; Accepted 18 August 2025
Available online 19 August 2025

developed by the United States National Toxicology Program (US NTP), is an ensemble machine learning tool that combines more than 40 models (e.g., random forests and artificial neural networks), for which the output is the weighted average of these model predictions [10]. Similarly, VEGA (available through VEGAHUB as part of software encoding more than 90 *in silico* toxicology tools), developed by the Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy, employs different machine learning techniques (e.g., decision trees, k-NN and logistic regression), which are combined into a consensus output [11].

A small number of studies have assessed the reliability of the TEST, CATMoS, and VEGA models in the prediction of oral rat LD_{50}. For example, Nelms et al. [12] analyzed the predictive performance of the TEST Consensus predictions, Firman et al. [13] assessed the predictive performance of the TEST Hierarchical clustering model, while Bishop et al. [14] estimated the accuracy and reliability of CATMoS predictions and Pampalakis [15] assessed the ability of VEGA to predict oral rat LD_{50} of toxic nerve agents. Weyrich et al. [16] assessed the predicitivity of CATMoS, in combination with expert opinion, for GHS classification data from regulatory submissions. The general conclusion from these works was that the predictive abilities of such models, in terms of accuracy and hazard classification sensitivity, were inherently hindered by associated uncertainty. Consequently, when used to assign health-protective (conservative) LD_{50} values to compounds, it is advised that one should focus on the extent to which the models can accurately predict and differentiate between low and high hazard within chemicals. In other words, under such conditions of uncertainty, it is most appropriate to apply the model with lower tendency towards predicting chemicals as less toxic than the corresponding experimental data imply [17–19]. This is particularly the case where the principle of being "conservative" is applied to account for uncertainty, by giving precaution to the predicted data [20].

In order to use TEST, CATMoS and VEGA conservatively, the challenge remains of determining which model is most reliable for such a purpose. Each has unique attributes which might influence the accuracy of its prediction output. Among these may be errors, limitations and biases stemming from factors including its parameters and its structure, or alternatively, its training data or training process [7,8,11]). One possible way to address this challenge is to use a consensus approach that combines individual model outputs into a single prediction [18,19]. The underlying premise of consensus QSAR modeling is that individual models, because of their reductionist nature, only account for limited structure–activity information within chemicals (as encoded in their structures and in the molecular descriptors used). Consequently, combining these predictions will potentially improve the overall reliability against the same data [21]. In QSAR literature, consensus modeling has been established through combinatorial approaches that apply multiple statistical methods within a model software, or else adopt several descriptors [3,7,22,23]. However, to our knowledge, little (if anything) has been done in order to derive conservative consensus model predictions based on a simple comparison of TEST, CATMoS and VEGA outputs, following then with the selection of the more conservative (more toxic) chemical-specific predictions as representative of health-protective values. Such a conservative approach may prove valuable for the replacement of the acute oral toxicity test, whereas a single prediction will be more prone to error and underestimate.

The aim of this study was, therefore, to assess the performance of a consensus approach of TEST, CATMoS and VEGA, against the individual models, in prediction of a conservative oral rat LD_{50} in a large, diverse selection of organic compounds. To this end, prediction accuracy for hazard classification using the consensus method was used in order to evaluate performance. The utility of this approach lies in its ability to establish a foundation to contextualize the use of consensus modeling in deriving health-protective oral rat LD_{50} estimates under conditions of uncertainty, especially where experimental data are limited or absent.

2. Methods

The flowchart depicted in Fig. 1 illustrates the step-by-step summary of the method applied in the study, from data sourcing to the estimation of model prediction accuracy, as further described below.

2.1. Data sourcing

Oral rat LD_{50} data relating to 8,186 organic compounds, each with Chemical Abstract Service Registration Numbers (CASRN), were obtained from Firman et al. [13]. These data were originally collated from different sources via the efforts of the US EPA and National Toxicity Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), and consisted, in the majority of instances, of two or three distinct point estimates of experimentally derived LD_{50} (expressed in mg/kg) [12]. The data were processed as in the previous analyses, removing duplicates and compounds without defined structures and inorganic compounds, correcting transcription errors, and retrieving SMILES and CASRN either from the US EPA's CompTox Chemicals Dashboard or from other public resources (details about the processing steps can be found in Nelms et al. [12]). In this study, a further processing step was implemented, removing organometallic substances and entries with multiple CASRN identifiers. Additionally, we retained only compounds that could be predicted using each of the TEST Consensus, CATMoS, and VEGA tools. The final dataset consisted of 6,229 organic molecules (see Supplementary information, Table S1, for the raw data).

Although some of these data were used to develop the TEST, CATMoS and VEGA models, the empirical LD_{50} values present within their respective training sets do not exactly match experimental LD_{50} data sourced from Firman et al. [13]. Furthermore, as noted by Bishop et al. [14], when making predictions for compounds already in these models, consensus tools do not take exact experimental values as the predictions. Instead, they generate their predictions based upon consensus of the individual models within. That is, the consensus predictions are generated through multistep mathematical simulations that are not based upon any specific empirical value in the model datasets, thus ensuring that the overlapping compounds within the data from Firman et al. [13] and within the model training sets do not necessarily affect interpretation of prediction results [14,24].

2.2. Prediction of the oral rat acute toxicity

Oral rat LD_{50} values for each of the 6,229 compounds were predicted in TEST software (v5.1.2), using the CASRN identifiers as input. The compounds were first split into batches, each containing approximately500 entries. Our preliminary analysis indicated that, in not exceeding this quantity per prediction exercise, memory issues within TEST, CATMoS and VEGA were avoided. The prediction options in TEST were set as: endpoint – oral rat LD_{50}, method – consensus, and fragment constrain – relaxed. The TEST Consensus method (average of predictions generated by Hierarchical clustering and Nearest neighbor methods) is considered the most reliable (US EPA, 2015); thus, only TEST Consensus (henceforth simply called TEST) predictions (expressed in mg/kg) were downloaded and saved. The compound batches described above were used in CATMoS (available within the OPERA App, v2.9) and VEGA (available within VEGAHUB, v 1.2.4). Accordingly, the CASRN identifiers were imported into CATMoS as inputs, whereas SMILES were imported into VEGA.

2.3. Deriving conservative consensus model (CCM) predictions

The minimum prediction concept for conservativeness was applied to select the lowest (i.e., the most toxic) LD_{50} value for each compound, from across its TEST, CATMoS and VEGA predictions [25]. That is, by "consensus", we mean most conservative [26]. This LD_{50} was then

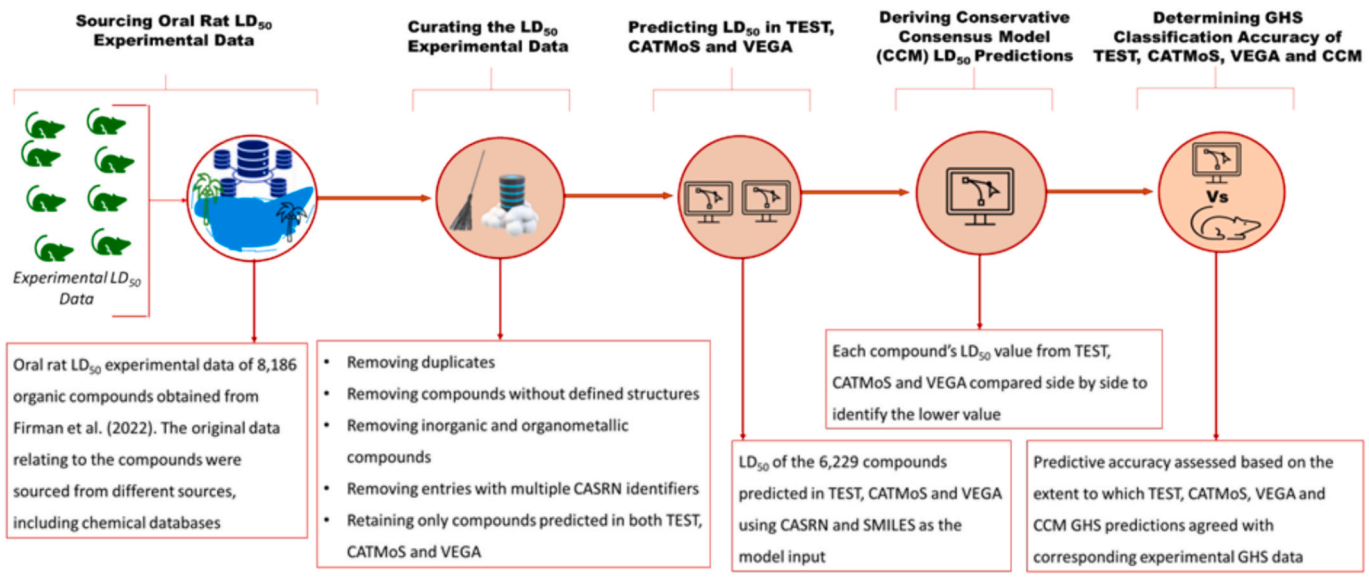


Fig. 1. Summary workflow illustrating the methodology used within this study, outlining the process from data sourcing to estimation of model prediction accuracy.

assigned to the compound as the "conservative consensus model (CCM) prediction". The CCM approach, as applied within this study, i.e., the use of more than one single QSAR model, is inherently more conservative than use of a single QSAR model alone. As proposed by the European Commission in the recent REACH review [9] and scholars [17–19], such an approach is recommended for consideration when single models yield varying data points, even for the same compound.

2.4. Model predictive accuracy for hazard classification

Predictive accuracy was assessed based on the extent to which TEST, CATMoS, VEGA and CCM predictions agreed with corresponding experimental data [27]. This was performed by first assigning compounds into Globally Harmonized System (GHS) categories (see Table 1) [28], based upon their experimental, TEST, CATMoS, VEGA and CCM LD_{50} outputs. Subsequently, predicted GHS vs. experimental GHS categories were compared. An accurate prediction was considered to have occurred when a model GSH classification matched the experimental classification. Under-prediction was stated to occur when a compound was predicted less toxic (i.e., have lower hazard classification) than the corresponding experimental data indicate, while over-prediction – synonymously called a "conservative" prediction – arose when a compound was predicted more toxic (i.e., have higher hazard classification) than suggested experimentally [18,29]).

2.5. Performance of predictions by structural class and functional group

In order to determine whether any chemical classes, or chemicals containing specific functional groups, were liable to be systematically under- or over-predicted by the adopted tools, structures within the dataset were profiled using 722 of the 729 ToxPrint chemotypes (MN-AM, Version 2.0 r711 (2014-06-11), github.com/mn-am/toxprint, accessed on 15 June 2025), seven ToxPrint chemotypes were omitted as they related to "elements" and were not considered further. This was achieved through use of the Chemotyper application (MN-AM, Version 1.3 r14761, github.com/mn-am/chemotyper, accessed on 15 June 2015) [30]. The distribution of under- and over-prediction was investigated with reference to ToxPrint chemotypes matched with compounds in the dataset.

Table 1
GHS classification criteria and associated hazard statements for acute oral toxicity.

GHS Category		Hazard statement
1	($LD_{50} \leq 5$ mg/kg)	Fatal if swallowed
2	($5 < LD_{50} \leq 50$ mg/kg)	Fatal if swallowed
3	($50 < LD_{50} \leq 300$ mg/kg)	Toxic if swallowed
4	($300 < LD_{50} \leq 2000$ mg/kg)	Harmful if swallowed
5	($LD_{50} > 2000 \leq 5000$ mg/kg)	May be harmful if swallowed
NC	$LD_{50} > 5000$ mg/kg	Not classified

NC: Not classified.

3. Results

3.1. Applicability domain

The OECD principles for the validation of (Q)SARs require model predictions to fall within a defined applicability domain, in order for them to be considered reliable [31]. CATMoS automatically checks the applicability domain of entries, where only compounds lying within are returned values [7]. Analysis indicated that all 6,229 compounds appeared within the model's applicability domain. In VEGA, applicability domain can be determined by the degree of similarity between molecules within training and predicted sets, where a similarity score $\geq$ 0.75 is generally considered to indicate reliability [11]. In our analysis, a score of $\geq$ 0.85 was achieved for all compounds, suggesting full domain coverage. TEST Consensus predictions are derived by averaging the outputs from Hierarchical clustering and Nearest neighbor methods, where only compounds predicted in both are considered most reliable and in-domain [8,32]). The full data set of 6,229 compounds were found to match these criteria. Thus, for the CCM, we defined a compound to be in its applicability domain if this was the case for TEST Consensus, CATMoS and VEGA models. As such, the complete dataset lies within the CCM applicability domain.

3.2. Comparing model predictive accuracy for hazard classification

3.2.1. Agreement with experimental data

Fig. 1 shows the distribution of the GHS categories across the 6,229 compounds, based upon experimental LD_{50} data. This distribution indicates that the majority (~39 %; 2,449/6,229) fall within category 4, with the smallest number found within category 1 (~3 %; 208/6,229). As explained in Section 2.4, model predictive accuracy was evaluated based upon the agreement of predicted LD_{50}-based GHS categories with

the corresponding assignment derived from experimental LD_{50}-based, depicted in Fig. 2.

Three accuracy parameters were defined: match (denotes accurate prediction), under-prediction, and over-prediction. A summary of the overall prediction accuracy of the three models is shown in Fig. 3a. Approximately 57 % of compounds (3,548/6,229) predicted in TEST were seen to match (i.e., were in agreement with) the experimental GSH categories, with the overall under-and over-prediction incidences being ~ 20 % (1,204/6,229) and ~ 24 % (1,477/6,229), respectively. As shown in Fig. 3b, most of the matched and over-predictions were distributed within categories 3, 4, and 5 (i.e., $\geq$300 $LD_{50} \leq$ 5000 mg/kg), suggesting that TEST was mostly reliable or conservative within this range.

CATMoS exhibited approximately 65 % (4,042/6,229) agreement with the experimental data, indicating a more accurate hazard category prediction rate than TEST (Fig. 3a). The under-and over-prediction rates were ~ 10 % and 25 %, respectively, which indicates CATMoS also to be more conservative than TEST. Similar to TEST, the majority of these accurate and over-predicted incidences occurred within GSH categories 3, 4, and 5 (Fig. 3b), suggesting that this model was also mostly reliable or conservative in predicting this range. Compared to TEST and CATMoS, VEGA produced the highest proportion of accurate predictions (~87 %; 5,436/6,229), with the lowest under- and over-prediction rates, at ~ 5 % (300/6,229) and ~ 8 % (493/6,229), respectively. These results suggest that VEGA was the least conservative among the three models. Owing to the nature of the kNN methodology employed within the software, this is to be anticipated. When faced with a substance present as part of the model training set, the position of the algorithm is to default towards returning the associated experimental value (as opposed to offering SAR-grounded prediction). This was relevant to 4,814 from out of the 6,299 dataset members.

For CCM, the GHS classification in ~ 56 % (3,520/6,229) of compounds was accurately predicted. This was lower than the accuracy rates recorded in TEST, CATMoS or VEGA alone (Fig. 3a). Graham et al. [18] reported similar findings, where 89 % (290/326) and 91 % (320/353) accurate predictions in CATMoS and Leadscope, relating to pharmaceuticals, were reduced to 77 % (286/370) in a conservative consensus of the two. However, relative to TEST, CATMoS and VEGA, the number of under-predictions in CCM was lowest, at ~ 2 % (130/6,229), while the number of over-predictions was highest at ~ 37 % (2,285/6,229). These results indicate that, by design, CCM was the most conservative. As with the individual models, most over-prediction incidences in CCM appeared within GSH categories 3, 4, and 5 (Fig. 3b). This similarly suggested that CCM was mostly conservative in predicting these categories. Given this concordance, it was considered more informative to characterize the level of conservativeness of each model across all GHS categories (see the discussion below in Section 3.2.2).

3.2.2. *Level of conservativeness of the model predictions*

Maximizing the number of over-predictions and minimizing the number of under-predictions are each important in ensuring model conservativeness [27]. Fig. 3a shows that CCM resulted in roughly 1.6-, 1.5-, and 5-fold more over-predictions relative to TEST, CATMoS and VEGA, respectively, and approximately 9.3-, 4.7- and 2.3-fold fewer under-predictions compared to TEST, CATMoS and VEGA, respectively. To further understand the level of conservativeness of CCM, the extent of coverage of its over-predictions within each GHS category was compared to those of TEST, CATMoS and VEGA, as further discussed below.

Generally speaking, over-predictions in each model systematically increased from GHS category 2 to NC, while under-predictions systematically decreased from category 1 to NC (Fig. 3b) (recognizing that Category 1 could not be overpredicted). Building upon these results, we assessed the distribution of each model's under-and over-predictions within GSH "toxic" ($LD_{50} \leq$ 2000 mg/kg; categories 1, 2, 3, and 4) and "non-toxic" ($LD_{50} >$ 2000 mg/kg; category 5 and NC) classes [28]. This classification may be used in regulatory contexts to communicate hazards associated with chemicals [6]. Similar to Alberga et al. [33] and Bercu et al. [1], we interpreted the degree of conservativeness to increase when a model's over-predictions were more frequently distributed towards "non-toxic" class compared to "toxic" class, or when under-predictions were more frequently distributed towards "toxic" class than to "non-toxic" class.

Table 2 shows the distributions of each model's over-predictions (numbers within the green-shaded cells) and under-predictions (numbers within the orange-shaded cells). For example, 8/6,229 compounds over-predicted in TEST were determined to fall under experimental category 1, whereas 99/6,229 compounds under-predicted in the same model were determined to fall under experimental category 2. The under-and over-predictions were demarcated within GHS "toxic" (outlined in the red rectangles) and "non-toxic" (outlined in the blue rectangles) classes.

Within the "toxic" class, CCM consistently recorded the highest total number of over-predictions at each category: category 2 (TEST = 8, CATMoS = 90, VEGA = 8, CCM = 93), category 3 (TEST = 97, CATMoS = 125, VEGA = 49, CCM = 173), and category 4 (TEST = 241, CATMoS = 596, VEGA = 99, CCM = 689). A similar observation was made in the "non-toxic" categories – i.e., category 5 (TEST = 655, CATMoS = 399, VEGA = 177, CCM = 790) and NC (TEST = 476, CATMoS = 368, VEGA = 160, CCM = 540) (Table 2). In contrast, the number of compounds under-predicted in the models decreased with increasing GHS category, with CCM generally recording the lowest total number of under-predictions in each: category 1 (TEST = 143, CATMoS = 82, VEGA = 27, CCM = 20), category 2 (TEST = 228, CATMoS = 138, VEGA = 49, CCM = 23), category 3 (TEST = 429, CATMoS = 156, VEGA = 99, CCM = 50), category 4 (TEST = 330, CATMoS = 186, VEGA = 98, CCM = 33), and category 5 (TEST = 74, CATMoS = 47, VEGA = 27, CCM = 4) (Table 2).

Overall, the distribution results in Table 2 indicate that CCM, relative to TEST, CATMoS and VEGA, provided the highest number of over-

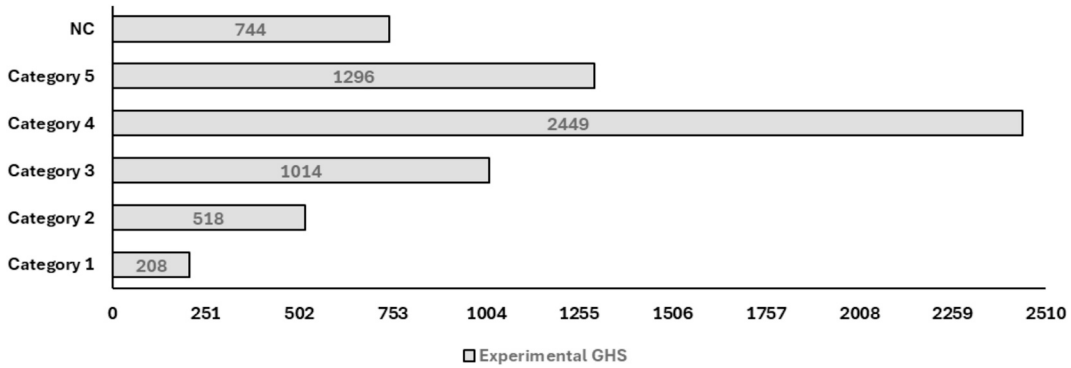


Fig. 2. GHS categories for the 6,229 dataset compounds, derived from experimental LD_{50} data.

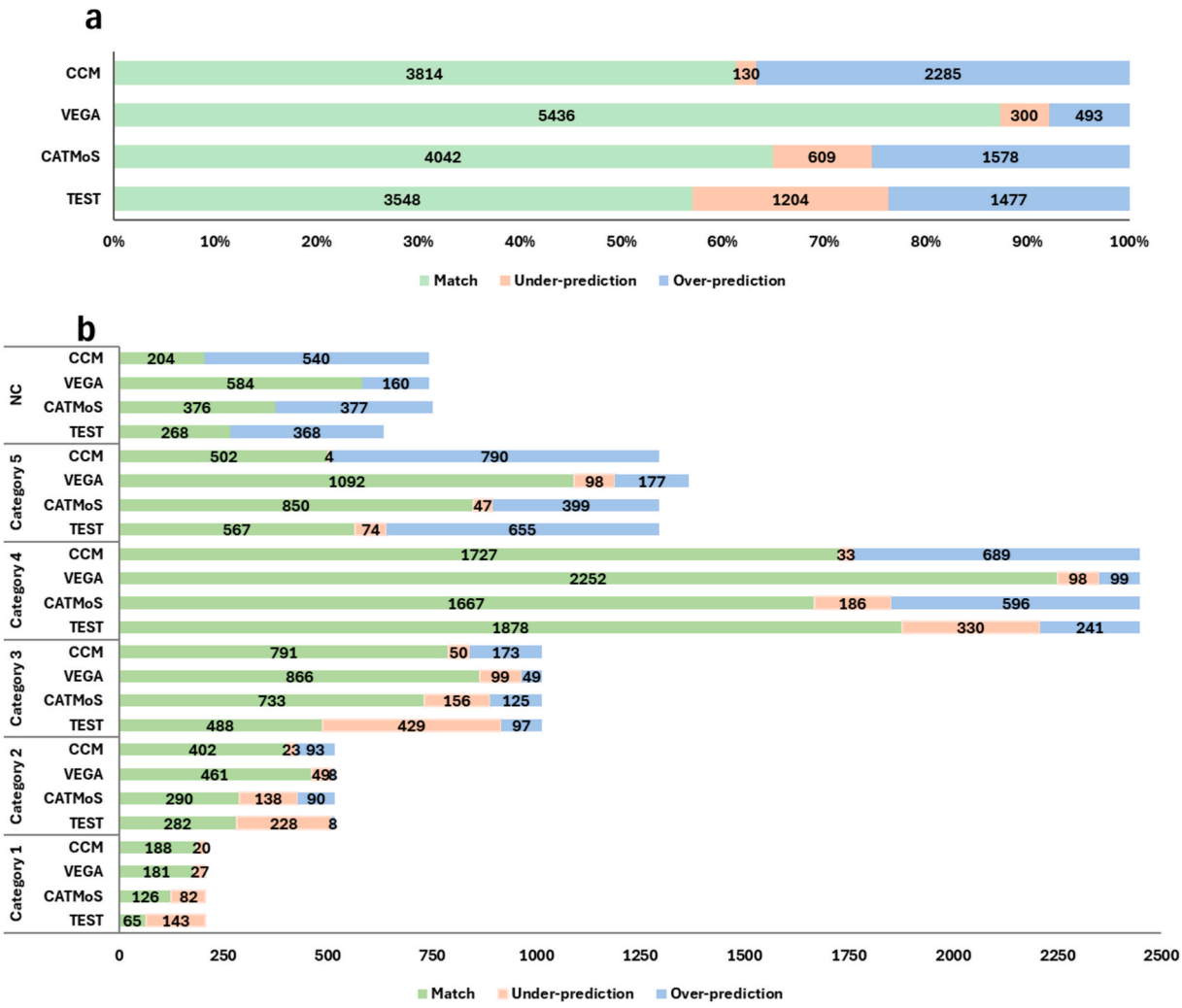


Fig. 3. Evaluation of model predictive accuracy for GHS classification. (a) The match, under-and over-prediction frequencies for each model, as well as (b) how far off each prediction sat from the experimental-based GHS category.

predictions (which were more frequently distributed towards less toxic/non-toxic GSH categories), and also the fewest under-predictions (more frequently distributed towards more toxic/toxic GHS categories). As such, it is reasonable to conclude that CCM possesses the highest level of conservativeness [1,33]. Studies have reported that an approach such as CCM, with its relatively high level of conservativeness, can be utilized for purposes such as providing a more reliable means of establishing safety recommendations for a chemical based on worst-case scenario considerations, for prioritizing compounds for further assessment, or for gaining assurance on non-toxic compounds before releasing them to the market [14,21,22,34].

3.2.3. *Impact of overlapping compounds on CCM prediction*

Using TEST and CATMoS, we analyzed whether overlap between the 6,229 compounds and the model training sets had an influence upon CCM prediction (please note that a similar evaluation was not possible with respect to VEGA output, owing to its tendency to adopt experimental figures in place of predictions whenever such overlap arose). The analysis showed that approximately75% of the compounds were present within each training set, while the remaining 25 % were absent. As shown in Fig. 4, the proportion of matching GHS classifications for TEST was about 58 % for those substances (1,588) absent in its training set and 57 % for those present (4,641). For CATMoS, this was about 56 % for compounds (1,484) absent and 68 % for those (4,745) present. However, the GHS over-prediction rates – the central focus of this study – likewise remained relatively consistent, regardless of whether the compounds were present or absent in the training sets – i.e., at 23 % (present) and 25 % (absent) for TEST and 25 % (present) and 27 % (absent) for CATMoS (see Supplementary materials S2 and S3 for the full analysis). These results reaffirm our earlier note in Section 2.1 that the presence of overlapping compounds in the training sets of the three models may not necessarily affect CCM's tendency to conservatively over-predict the LD_{50}-based GHS of the dataset members.

3.3. *Analysis of predictions by structural class and functional group*

All compounds within the dataset were found to match at least one ToxPrint chemotype. In total, 515 of the 722 ToxPrint Chemotypes were hit – with full analysis reported in Supplementary Table S4 data file. In order to further evaluate Chemotypes, the 218 identified as present in 50 or more molecules were retained. Those Chemotypes associated with underprediction and overprediction of toxicity are reported in Tables 3a and 3b, respectively. The rates of underprediction are low, with maximum occurrence being 6.3 %. No clear or obvious chemical classes or functional groups were identified. Conversely, as required by a conservative approach, rates of overprediction of toxicity were high – reaching up to 52 % – with benzimidazoles identified as the chemical class most often subject to this.

Table 2
Confusion matrices showing the distribution of 6,229 compounds within GHS categories, organized by in accordance with experimental-based GHS assignment.

		Experimental					
		1	**2**	**3**	**4**	**5**	**NC**
TEST predicted	**1**	65	99	25	15	4	0
	2	8	282	182	40	5	1
	3	1	96	488	401	22	6
	4	0	14	227	1878	308	22
	5	1	8	27	619	567	74
	NC	0	3	10	175	288	268
CATMoS predicted	**1**	126	66	15	1	0	0
	2	90	290	127	11	0	0
	3	2	123	733	145	10	1
	4	1	9	586	1667	177	9
	5	0	3	37	359	850	47
	NC	0	2	9	100	257	376
VEGA predicted	**1**	181	18	3	5	1	0
	2	8	461	28	19	1	1
	3	0	49	866	84	11	4
	4	0	8	91	2252	81	17
	5	0	1	3	173	1092	27
	NC	0	0	4	38	118	584
CCM predicted	**1**	188	19	1	0	0	0
	2	93	402	23	0	0	0
	3	2	171	791	50	0	0
	4	1	27	661	1727	33	0
	5	1	10	50	729	502	4
	NC	0	4	17	218	301	204

4. Discussion and conclusion

The aim of this study was to assess the performance of CCM against individual TEST, CATMoS and VEGA methods in the prediction of conservative (health-protective) oral rat LD_{50}, by leveraging a large and diverse dataset (6,229 compounds). Existing QSAR studies on this endpoint (e.g., Firman et al. [13] and Graham et al. [18]) have mainly focused on model performance in general, based on statistical correlation analysis of experimental vs. predicted LD_{50}, or else on hazard classification sensitivity of specific models [16]. In this study, we went beyond the areas addressed in these works by primarily focusing on model conservative predictions – determining the extent to which consensus predictions of TEST, CATMoS and VEGA can provide health-protective (i.e., conservative) predictions. To our knowledge, no study has applied these three models for this purpose. Additionally, we used, arguably, the largest dataset (i.e., 6,229 compounds) assembled for such an exercise. As further discussed below, we argue that the outcome of this study lays a strong foundation to contextualize the use of consensus modeling for deriving health-protective predictions under conditions of uncertainty, particularly where experimental data are scarce.

4.1. Consideration of CCM under conditions of uncertainty

Uncertainty in model predictions can be addressed through conservative approaches, e.g., by applying uncertainty factors or by selecting conservative model estimates [18,29,35]. It has previously been considered for the prediction of acute toxicity [36]. The current study proposes the need to err on the side of over-prediction, accounting for potential uncertainty arising from individual model outputs. This

Table 3a
Ten most significant ToxPrints Chemotypes associated with underpredictions from the CCM.

ToxPrint Chemotype	No. matching compounds	Underpredicted by one class
bond:CC(=O)C_ketone_alkene_generic	111	6.3 %
ring:hetero_[6_6]_O_benzopyran	81	6.2 %
ring:hetero_[5_5]_Z_generic	53	5.7 %
bond:C=N_imine_N(connect_noZ)	53	5.7 %
bond:C=O_carbonyl_ab-unsaturated_aliphatic_(michael_acceptors)	60	5.0 %
bond:COC_ether_alkenyl	60	5.0 %
bond:N[!C]_amino	81	4.9 %
chain:oxy-alkaneLinear_ethyleneOxide_EO2	64	4.7 %
ring:hetero_[3]_Z_generic	87	4.6 %
ring:hetero_[6]_N_pyrimidine	88	4.6 %

Table 3b
Ten most significant ToxPrints Chemotypes associated with overpredictions from the CCM.

ToxPrint Chemotype	No. matching compounds	Underpredicted by one class
ring:hetero_[5_6]_N_benzimidazole	144	52.1 %
bond:CX_halide_alkyl-X_secondary	69	40.6 %
bond:CX_halide_alkenyl-X_dihalo_(1_2-)	52	40.4 %
bond:CX_halide_aromatic-X_trihalo_benzene_(1_2_3-)	67	40.3 %
ring:aromatic_biphenyl	85	40.0 %
bond:CX_halide_alkyl-F_trifluoro_(1_1_1-)	268	39.9 %
ring:hetero_[5]_N_imidazole	229	39.7 %
ring:hetero_[3]_O_epoxide	72	38.9 %
bond:C=N_carboxamidine_generic	103	38.8 %
bond:CN_amine_sec-NH_aromatic	150	38.7 %

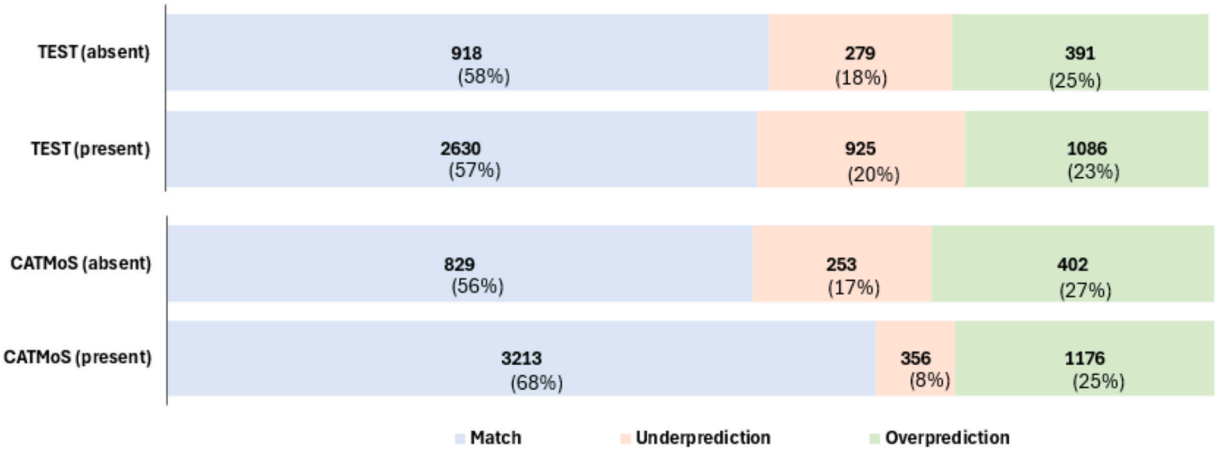


Fig. 4. Summary of TEST and CATMoS GHS Classifications of the compounds present and absent within their training sets.

proposal (established as CCM) combines predictions from individual models. Valsecchi et al. [21] and Zhu et al. [3] explain that such a method can mitigate the potential influence of outliers from sole model predictions. This mitigation potential of CCM could be particularly useful when applying the TEST, CATMoS and VEGA tools for regulatory assessment of chemicals in the absence of experimental data, where it might otherwise become challenging to determine the degree to which any of these models alone can reliably predict oral rat LD_{50} [12,18,19]. In other words, the proposed consensus approach has the potential to reduce or eliminate conflicting predictions between models; consequently, addressing uncertainty due to discrepancies or inconsistency in a compound's estimated LD_{50} values.

As mentioned above, a key challenge in QSAR modeling is the lack of reproducibility of chemical-specific predictions between models. This is attributable to factors such as: different models using differing training sets with varying degrees of experimental error, the use of different model parameters and algorithms, the presence of systematic biases within a model, or else the presence of random errors embedded within a model's structure [3,12,18,19]. TEST and VEGA are built upon raw LD_{50} values [37], whereas CATMoS is built upon both these and also GHS categorical assignments [7]. While this may not be the sole reason for the observed differences in the predictions between the two (Fig. 3 (a and b) and Table 2), studies have shown that point and categorical data inputs may contribute to QSARs yielding varying outcomes [24,38]. This means that if conservative predictions from either of the models are used to set health-protective values, even if one model estimate is more accurate than the other, uncertainty may still exist to the extent that both models may produce conflicting predictions of the same chemical [39]. Moreover, uncertainty might arise regarding potential disagreement as to which model is the more reliable in conducting a prediction [39]. Drawing on the harmonization potential of CCM, it, therefore, becomes clear that this approach could be relied upon to reduce the effect of such uncertainties or minimize statistical type II errors (i.e., incorrectly concluding that a chemical has no adverse health effects) in the face of the uncertainties [40]. It should be noted, again, that predictions from VEGA are biased by its offering of experimental data in place of predicted quantities (where possible), hence improving the apparent accuracy of its score.

The issue of data quality in the development of QSARs has long been appreciated [41]. The consideration of data quality and curation is of great importance as it underpins the QSAR models, and it is understood that predictive accuracy cannot surpass the accuracy of the data themselves being modeled [42–46]. Another aspect is how the original data were manipulated to create the datum point that enters the QSAR model. In this study the original data for modeling were largely taken from a well curated source, namely the Integrated Chemical Environment (ICE) from the US National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) [47]. Thus the underlying toxicological data and chemical structures can be considered to have a good level of confidence. Where individual modeling approaches vary, however, is with the manipulation of the toxicity data when multiple values are available for the same compound. 1-Chloro-4-nitrobenzene (CAS 100-00-5) is an example of a substance for which differing LD_{50} values are noted within each of the three model training sets: 353, 420 and 460 mg/kg in VEGA, TEST and CATMoS (respectively). The precise reasoning for this unclear. In the majority of instances, however, it is likely to be related to how multiple data points are unified into a single point for modeling. There are no hard and fast rules for the manipulation of multiple data points, although various approaches may be attempted and different approaches will affect the outcomes of a QSAR [48]. As the data will vary, even for the same compound, within models, the predictivity will also vary, as will the statistics arising from the evaluation study.

Structural analysis of the compounds with regard to ToxPrint Chemotypes in Tables 3a and 3b (complete analysis provided within the Supplementary Table S5 data file) demonstrates no significant trends, other than confirming conservative nature of the CCM approach. The analysis indicates fewer underpredictions than overpredictions. There is little commonality in the structural classes that are underpredicted, with a maximum underprediction rate of 6.3 % likely to be indicative of experimental error or bias. Overpredictions are more common in the CCM approach (by design). The types of compounds with significant overprediction (by approximately 50 %) were the benzimidazoles and various halide-incorporating compounds. Using the predictions, as presented, would be conservative for such classes in these cases.

4.2. Prioritizing chemicals based on CCM predictions

While it remains for the end user of the model predictions to decide on the best practices to prioritize chemicals either for further assessment or for safety-based decisions, from a health protection point of view, it is not uncommon to use GHS classifications as a factor in prioritizing compounds for further assessment [19,28]. An example is the Canadian Identification of Risk Assessment Priorities (IRAP) scheme, which highlights the importance of using chemical hazard information and new scientific techniques (e.g., QSAR) to guide decisions regarding whether further (or else no) risk assessment, or additional data generation, is required for a substance [49]. The United States Toxic Substances Control Act (TSCA) also permits the ranking of substances to the extent that their health hazards are aligned with GHS categories [50]. As illustrated with the example of six compounds within Table 4, we argue that CCM can support these prioritization needs.

Compounds may be prioritized in the rank order shown in Table 4, established based on CCM GHS categories. For example, 3-pentenenitrile might be considered the highest priority for assessment, or for stringent safety measures and handling protocols, due to its high potential to cause harm (the compound has the lowest LD_{50} value of 2 (mg/kg) and falls under GHS category 1). In contrast, the potential of the non-toxic lactulose (LD_{50} = 9,511 (mg/kg); GHS category NC) to cause harm is very minimal, meaning that safety measures against it may be the most relaxed. The foundation of our argument with this illustration is that CCM could hold significant value in supporting interim or internal decision-making during the initial stages of compound development, where studies typically incorporate rapid screening in order to detect potential toxic hazard that may render a compound unsuitable for examination [51]. In such scenarios, predicted oral rat LD_{50} values are often considered non-conclusive; hence, they should be conservative in order to account for potential uncertainty in the prediction [51].

Overall, the results obtained from this study can bolster confidence that decisions based upon CCM predictions can be the most health-protective, especially within chemical screening-level assessments performed in the absence of experimental data. Alternatively, in demonstrating that CCM can improve the level of conservativeness in the predictions, we argue that it can be adapted for use in regulatory contexts to contribute to a weight-of-evidence approach that justifies the need to prioritize particular chemicals for further assessment. Additionally, regulators can use the information from CCM to balance options or to set precautionary measures aimed at reducing potential human health risks by restricting the use of particular chemicals.

Importantly, we wish to reiterate that CCM is proposed here with the objective of maximizing conservative predictions. Based on our results in Fig. 3 (a and b), CCM significantly expands the number of conservative predictions, underscoring its potential to enhance conservative screening of large, structurally diverse chemical inventories. Of great importance is its potential application in regulatory contexts, where *in silico* (such as QSAR) modeling is often used to address hazard data gaps across broad chemical inventories. For instance, as noted by Collins et al. [52], more than 16,000 industrial chemicals listed on Canada's Domestic Substances List (DSL) are amenable to consensus QSAR assessment. We argue that CCM is a promising tool that could be applied within DSL to conservatively screen and classify these chemicals. Elsewhere, CCM could be well-suited for screening data-poor and high-

Table 4
Ranking of six chemicals based upon CCM GHS category. The compounds are ordered: highly toxic (categories 1 and 2; orange-shaded), toxic (categories 3 and 4; blue-shaded), and non-toxic (category 5 and NC; grey-shaded).

Compound	CASRN	Model predictions								Ranking order
		TEST		CATMoS		VEGA		CCM		
		LD_{50} (mg/kg)	GHS category	LD_{50} (mg/kg)	GHS category	LD_{50} (mg/kg)	GHS category	LD_{50} (mg/kg)	GHS category	
3-Pentenenitrile	4635–87-4	269	3	2	1	47	2	2	1	1
Spinosyn A	131929–60-7	34	2	3755	5	3741	5	34	2	2
Allyl methacrylate	96–05-9	6167	NC	260	4	69	3	69	3	3
Pindolol	13523–86-9	1144	4	622	4	860	4	622	4	4
Heptanoic acid	111–14-8	2055	5	5827	NC	7005	NC	2055	5	5
Lactulose	4618–18-2	9511	NC	12,095	NC	17,968	NC	9511	NC	6

production volume chemicals, such as those identified as potential endocrine-disrupting substances on the US EPA's Tier 1 Screening List [53].

Put together, the above examples highlight the potential utility of CCM within a broader, multi-tiered chemical assessment framework that supports prioritization and advancement toward non-animal testing approaches across jurisdictions. The need for non-animal data is already considered a priority within jurisdictions such as Canada, the EU and the US, where, for example, non-experimental rodent data for acute toxicity endpoints are submitted for new and existing chemicals, for manufacturing concentrates, for metabolites or degradation products, or during chemical premanufacture notice [54]. In other words, CCM presents an opportunity to fully replace the use of the *in vivo* oral rat tests for determining LD_{50}-based GHS hazard classification and labeling of chemicals within regulatory frameworks.

There is increasing interest in the use of QSAR models for the prediction of acute oral toxicity, with the possibility of the CATMoS model being proposed by the European Commission as a replacement [9]. Finally, while acknowledging the importance of CCM as a tool for regulatory applications, several considerations remain for future studies. For example, the performance of CCM could be validated using external datasets to assess its generalizability across broader chemical spaces, particularly for under-predicted chemical groups. A recent example of how this may be achieved is provided by Weyrich et al. [16]. Additionally, machine learning methods could be incorporated into CCM in order to assess its predictive accuracy based on statistical confidence scores, quantifying uncertainty in the predictions to judge their reliability for informed regulatory decision-making. Lastly, as noted earlier, the CCM, i.e., the use of more than one single QSAR model, will be inherently more conservative than use of a single QSAR model alone, as is proposed by the European Commission in the recent REACH review [9]. As such, it is recommended that consideration be given to a CCM approach for the replacement of acute oral toxicity testing.

CRediT authorship contribution statement

Jerry Achar: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **James W. Firman:** Writing – review & editing, Formal analysis. **Mark T.D. Cronin:** Writing – review & editing.

Declaration of competing interest

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.comtox.2025.100374.

Data availability

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

References

[1] J. Bercu, M.J. Masuda-Herrera, A. Trejo-Martin, et al., A cross-industry collaboration to assess if acute oral toxicity (Q)SAR models are fit-for-purpose for GHS classification and labelling, Regul. Toxicol. Pharmacol. 120 (2021) 104843.
[2] C.M. Zwickl, J.C. Graham, R.A. Jolly, et al., Principles and procedures for assessment of acute toxicity incorporating *in silico* methods, Comput. Toxicol. 24 (2022) 100237.
[3] H. Zhu, T.M. Martin, L. Ye, et al., QSAR Modeling of rat acute toxicity by oral exposure, Chem. Res. Toxicol. 22 (2009) 1913–1921.
[4] J. Strickland, A.J. Clippinger, J. Brown, et al., Status of acute systemic toxicity testing requirements and data uses by U.S. regulatory agencies, Regul. Toxicol. Pharmacol. 94 (2018) 183–196.
[5] E. Walum, Acute oral toxicity, Environ. Health Perspect. 106 (1998) 497–503.
[6] R. Gonella Diaza, S. Manganelli, A. Esposito, et al., Comparison of in silico tools for evaluating rat oral acute toxicity, SAR QSAR Environ. Res. 26 (2015) 1–27.
[7] K. Mansouri, A.L. Karmaus, J. Fitzpatrick, et al., CATMoS: collaborative acute toxicity modeling suite, Environ. Health Perspect. 129 (2021) 47013.
[8] US EPA. Toxicity Estimation Software Tool (TEST), https://www.epa.gov/comptox-tools/toxicity-estimation-software-tool-test (2015, accessed 10 March 2024).
[9] European Commission. AP 2 Introduction of new approach methods – human health, https://circabc.europa.eu/ui/group/a0b483a2-4c05-4058-addf-2a4de71b9a98/library/77cfd2b6-ecf5-4fbc-92bf-9ca340237d40/details (2023, accessed 8 August 2025).
[10] K. Mansouri, V. Consonni, M.K. Durjava, et al., Assessing bioaccumulation of polybrominated diphenyl ethers for aquatic species by QSAR modeling, Chemosphere 89 (2012) 433–444.
[11] A. Roncaglioni, A. Lombardo, E. Benfenati, The VEGAHUB platform: the philosophy and the tools, Altern. Lab. Anim. 50 (2022) 121–135.
[12] M.D. Nelms, A.L. Karmaus, G. Patlewicz, An evaluation of the performance of selected (Q)SARs/expert systems for predicting acute oral toxicity, Comput. Toxicol. 16 (2020) 100135.
[13] J.W. Firman, M.T.D. Cronin, P.H. Rowe, et al., The use of Bayesian methodology in the development and validation of a tiered assessment approach towards prediction of rat acute oral toxicity, Arch. Toxicol. 96 (2022) 817–830.
[14] P.L. Bishop, K. Mansouri, W.P. Eckel, et al., Evaluation of *in silico* model predictions for mammalian acute oral toxicity and regulatory application in pesticide hazard and risk assessment, Regul. Toxicol. Pharmacol. 149 (2024) 105614.
[15] G. Pampalakis, Underestimations in the in silico-predicted toxicities of V-agents, J. Xenobiotics 13 (2023) 615–624.
[16] A. Weyrich, N. Peter, N. Watzek, et al., Can acute oral *in vivo* toxicity testing for EU REACH be fully replaced by a QSAR method? Evaluation of the CATMoS model using chemical industry data, Regul. Toxicol. Pharmacol. 162 (2025) 105861.
[17] World Health Organization & International Programme on Chemical Safety. Guidance document on evaluating and expressing uncertainty in hazard characterization, 2nd edition, https://www.who.int/publications/i/item/9789241513548 (2018, accessed 22 July 2024).
[18] J.C. Graham, M. Rodas, J. Hillegass, et al., The performance, reliability and potential application of *in silico* models for predicting the acute oral toxicity of pharmaceutical compounds, Regul. Toxicol. Pharmacol. 119 (2021) 104816.
[19] C. Moudgal, L.T. Anger, W. Muster, et al., The application of acute oral toxicity computational models in dangerous goods classification, Toxicol. Ind. Health 39 (2023) 687–699.
[20] EFSA, D. Benford, T. Halldorsson, et al., The principles and methods behind EFSA's Guidance on uncertainty analysis in scientific assessment, EFSA J. 16 (2018) e05122.

[21] C. Valsecchi, F. Grisoni, V. Consonni, et al., Consensus versus individual QSARs in classification: comparison on a large-scale case study, J. Chem. Inf. Model. 60 (2020) 1215–1223.
[22] F. Lunghini, G. Marcou, P. Azam, et al., Consensus models to predict oral rat acute toxicity and validation on a dataset coming from the industrial context, SAR QSAR Environ. Res. 30 (2019) 879–897.
[23] S. Schieferdecker, F. Rottach, E. Vock, In silico prediction of oral acute rodent toxicity using consensus machine learning, J. Chem. Inf. Model. 64 (2024) 3114–3122.
[24] A.L. Karmaus, K. Mansouri, K.T. To, et al., Evaluation of variability across rat acute oral systemic toxicity studies, Toxicol. Sci. 188 (2022) 34–47.
[25] K. Gromek, W. Hawkins, Z. Dunn, et al., Evaluation of the predictivity of Acute Oral Toxicity (AOT) structure-activity relationship models, Regul. Toxicol. Pharmacol. 129 (2022) 105109.
[26] K. Khan, E. Benfenati, K. Roy, Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: Ranking and prioritization of the DrugBank database compounds, Ecotoxicol. Environ. Saf. 168 (2019) 287–297.
[27] S. Hoffmann, A. Kinsner-Ovaskainen, P. Prieto, et al., Acute oral toxicity: variability, reliability, relevance and interspecies comparison of rodent LD50 data from literature surveyed for the ACuteTox project, Regul. Toxicol. Pharmacol. 58 (2010) 395–407.
[28] United Nations. GHS classification criteria for acute toxicity, https://webapps.ilo.org/static/english/protection/safework/ghs/ghsfinal/ghsc05.pdf (2021, accessed 8 July 2024).
[29] EFSA, Benford D, Halldorsson T, et al. Guidance on Uncertainty Analysis in Scientific Assessments. *EFSA J* 2018; 16: e05123.
[30] C. Yang, A. Tarkhov, J. Marusczyk, et al., New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling, J. Chem. Inf. Model. 55 (2015) 510–528.
[31] OECD. Guidance document on the validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] models, https://doi.org/10.1787/9789264085442-en (2007, accessed 5 April 2024).
[32] M. Noga, A. Michalska, K. Jurowski, Application of toxicology in silico methods for prediction of acute toxicity (LD50) for Novichoks, Arch. Toxicol. 97 (2023) 1691–1700.
[33] D. Alberga, D. Trisciuzzi, K. Mansouri, et al., Prediction of acute oral systemic toxicity using a multifingerprint similarity approach, Toxicol. Sci. 167 (2019) 484–495.
[34] C.R. García-Jacas, Y. Marrero-Ponce, F. Cortés-Guzmán, et al., Enhancing acute oral toxicity predictions by using consensus modeling and algebraic form-based 0D-to-2D molecular encodes, Chem. Res. Toxicol. 32 (2019) 1178–1192.
[35] D. Wikoff, L. Haws, C. Ring, et al., Application of qualitative and quantitative uncertainty assessment tools in developing ranges of plausible toxicity values for 2,3,7,8-tetrachlorodibenzo-p-dioxin, J. Appl. Toxicol. JAT 39 (2019) 1293–1310.
[36] T.W. Schultz, A. Chapkanov, S. Kutsarova, et al., Assessment of uncertainty and credibility of predictions by the OECD QSAR Toolbox automated read-across workflow for predicting acute oral toxicity, Comput. Toxicol. 22 (2022) 100219.
[37] A. Danieli, E. Colombo, G. Raitano, et al., The VEGA tool to check the applicability domain gives greater confidence in the prediction of in silico models, Int. J. Mol. Sci. 24 (2023) 9894.
[38] S.S. Kolmar, C.M. Grulke, The effect of noise on the predictive limit of QSAR models, J. Cheminformatics 13 (2021) 92.
[39] M. Kirchner, H. Mitter, U.A. Schneider, et al., Uncertainty concepts for integrated modeling - review and application for identifying uncertainties and uncertainty propagation pathways, Environ. Model Softw. 135 (2021) 104905.
[40] P.M. Chapman, A. Fairbrother, D. Brown, A critical evaluation of safety (uncertainty) factors for ecological risk assessment, Environ. Toxicol. Chem. 17 (1998) 99–108.
[41] M.T.D. Cronin, T.W. Schultz, Pitfalls in QSAR, J. Mol. Struct. THEOCHEM 622 (2003) 39–51.
[42] S.J. Belfield, M.T.D. Cronin, S.J. Enoch, et al., Guidance for good practice in the application of machine learning in development of toxicological quantitative structure-activity relationships (QSARs), PLoS One 18 (2023) e0282924.
[43] J. Achar, J.W. Firman, M.T.D. Cronin, et al., A framework for categorizing sources of uncertainty in *in silico* toxicology methods: Considerations for chemical toxicity predictions, Regul. Toxicol. Pharmacol. 154 (2024) 105737.
[44] J. Achar, M.T.D. Cronin, J.W. Firman, et al., A problem formulation framework for the application of in silico toxicology methods in chemical risk assessment, in: *Arch Toxicol.* Epub Ahead of Print 30 March, 2024, https://doi.org/10.1007/s00204-024-03721-6.
[45] J. Achar, J.W. Firman, C. Tran, et al., Analysis of implicit and explicit uncertainties in QSAR prediction of chemical toxicity: a case study of neurotoxicity, Regul. Toxicol. Pharmacol. 154 (2024) 105716.
[46] V.M. Alves, S.S. Auerbach, N. Kleinstreuer, et al., Curated data in — trustworthy in silico models out: the impact of data quality on the reliability of artificial intelligence models as alternatives to animal testing, Altern. Lab. Anim. 49 (2021) 73–82.
[47] A.B. Daniel, N. Choksi, J. Abedini, et al., Data curation to support toxicity assessments using the integrated chemical environment, Front. Toxicol. 4 (2022) 987848.
[48] F.P. Steinmetz, S.J. Enoch, J.C. Madden, et al., Methods for assigning confidence to toxicity data with multiple values — Identifying experimental outliers, Sci. Total Environ. 482–483 (2014) 358–365.
[49] Health Canada. The identification of risk assessment priorities, https://www.canada.ca/en/health-canada/services/chemical-substances/fact-sheets/identification-risk-assessment-priorities.html (2017, accessed 12 July 2024).
[50] US EPA. Identifying existing chemicals for prioritization under TSCA, https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/identifying-existing-chemicals-prioritization-under (2018, accessed 12 October 2024).
[51] M.S. Marty, A.K. Andrus, K.A. Groff, Animal metrics: tracking contributions of new approach methods to reduced animal use, ALTEX – Altern. Anim. Exp. 39 (2022) 95–112.
[52] S.P. Collins, B. Mailloux, S. Kulkarni, et al., Development and application of consensus in silico models for advancing high-throughput toxicological predictions, Front. Pharmacol. 15 (2024) 1307905.
[53] US EPA. Final Second List of Chemicals for Tier 1 under the Endocrine Disruptor Screening Program, https://www.epa.gov/endocrine-disruption/final-second-list-chemicals-tier-1-under-endocrine-disruptor-screening-program (2015, accessed 28 June 2025).
[54] J. Strickland, E. Haugabrooks, D.G. Allen, et al., International regulatory uses of acute systemic toxicity data and integration of new approach methodologies, Crit. Rev. Toxicol. 53 (2023) 385–411.