

DOI: 10.1002/minf.201800124

Integrated QSAR Models to Predict Acute Oral Systemic Toxicity

Davide Ballabio,^{*,[a]} Francesca Grisoni,^[a] Viviana Consonni,^[a] and Roberto Todeschini^[a]

This paper is dedicated to Prof. Paola Gramatica on the occasion of her retirement.

Abstract: The ICCVAM Acute Toxicity Workgroup (U.S. Department of Health and Human Services), in collaboration with the U.S. Environmental Protection Agency (U.S. EPA, National Center for Computational Toxicology), coordinated the “Predictive Models for Acute Oral Systemic Toxicity” collaborative project to develop *in silico* models to predict acute oral systemic toxicity for filling regulatory needs. In this framework, new Quantitative Structure-Activity Relationship (QSAR) models for the prediction of very toxic (LD_{50} lower than 50 mg/kg) and nontoxic (LD_{50} greater than or equal to 2,000 mg/kg) endpoints were developed, as described in this study. Models were developed on a large set of chemicals (8992), provided by the project coordinators, considering the five OCED principles for QSAR applicability to regulatory endpoints. A Bayesian consensus approach integrating three different classification QSAR

Keywords: oral toxicity · QSAR · consensus · ICCVAM

algorithms was applied as modelling method. For both the considered endpoints, the proposed approach demonstrated to be robust and predictive, as determined by a blind validation on a set of external molecules provided in a later stage by the coordinators of the collaborative project. Finally, the integration of predictions obtained for the very toxic and nontoxic endpoints allowed the identification of compounds associated to medium toxicity, as well as the analysis of consistency between the predictions obtained for the two endpoints on the same molecules. Predictions of the proposed consensus approach will be integrated with those originated from models proposed by the participants of the collaborative project to facilitate the regulatory acceptance of *in-silico* predictions and thus reduce or replace experimental tests for acute toxicity.


1 Introduction

Acute toxicity testing aims to assess the generic toxic effects of a chemical or product, caused by the exposure through a pre-defined route (oral, dermal, inhalation) and that occur during a subsequent 21-day observation period.^[1] Acute toxicity assessment is a requirement under many regulatory frameworks for the classification and labelling of several types of substances, such as industrial chemicals (REACH, Annexes VII and VIII), biocides (EU 528/2012), pesticides (EC 1107/2009) and cosmetic ingredients (SCCS/1564/15). Acute oral toxicity is usually determined *in vivo*, through the Lethal Dose 50 (LD_{50}) testing, which aims to find the single lethal dose of a substance killing half of the animals in a test group, with ethical and cost/time drawbacks due to the utilization of living animals. In this context, alternatives to animal testing play a central role for the reduction, replacement and rationalization of the tests performed on animals.^[2]

In 2017, the Acute Toxicity Workgroup (ATWG) of the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) organized a collaborative project to develop *in silico* models of acute oral systemic toxicity that predict several acute-toxicity-related endpoints, based on a large dataset of rodent studies and targeted

towards regulatory needs identified across federal agencies.^[3] These endpoints included: (a) identification of “very toxic” chemicals ($LD_{50} < 50$ mg/kg), (b) identification of “nontoxic” chemicals ($LD_{50} \geq 2,000$ mg/kg), (c) point estimates for rodent LD_{50} , (d) categorization of toxicity hazard using the U.S. Environmental Protection Agency (EPA) as well as the United Nations Globally Harmonized System of Classification and Labelling (GHS) classification schemes. The rationale behind the project was to involve several international research groups and develop multiple *in silico* models to be aggregated for regulatory applications. In fact, while regulatory acceptance of computational predictions is unlikely to be achieved with a unique model developed by a single lab, the aggregation of multiple and diverse models allows to leverage the models’ strengths, thereby overcoming the limitations of the individual approaches.

[a] D. Ballabio, F. Grisoni, V. Consonni, R. Todeschini
Department of Earth and Environmental Sciences
University of Milano-Bicocca
Pza della Scienza 1, 20126, Milano, Italy
phone: (+39) 02 64482820
E-mail: davide.ballabio@unimib.it

 Supporting information for this article is available on the WWW under <https://doi.org/10.1002/minf.201800124>

For the scopes of such an international project, the NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), in collaboration with the U.S. Environmental Protection Agency's National Center for Computational Toxicology (U.S. EPA NCCT), has compiled a large body of rat acute oral lethality data that can be used for the development of the *in silico* models of acute oral systemic toxicity. The organizing committee of the ICCVAM collaborative project invited then international research groups to develop *in silico* models that predict any or all of these endpoints. The participants were provided with part of the collected experimental data, while the remaining part of the dataset was used for the model evaluation by the organizing committee. Models developed for the project that met criteria defined by the project organizing committee will be used to generate consensus predictions for the acute oral toxicity endpoints of interest to regulatory agencies.^[3]

In the framework of the ICCVAM collaborative project, this study presents the models developed for the very toxic and nontoxic endpoints by the Milano Chemometrics & QSAR Research Group at the University of Milano-Bicocca. The developed models are based on novel as well as benchmark machine learning techniques that were merged in a consensus approach, thereby increasing their reliability for animal testing replacement and regulatory applications. Models were developed on a large set of chemicals, considering the five OCED principles for QSAR applicability to regulatory endpoints. In particular, a double validation protocol was carried out and models were assessed through a blind validation on a set of external molecules provided in a later stage by the coordinators of the collaborative project. Finally, the identification of hazardous chemicals was undertaken by integrating predictions obtained for the very toxic and nontoxic endpoints.

2 Materials and Methods

2.1 Experimental Toxicity Data

NICEATM and NCCT collected and curated a rat acute oral toxicity database of systemic toxicity Lethal Dose 50 (LD₅₀) values, which represent the concentration needed to cause lethality in 50% of the utilized animals.^[4] A total of five different modelling endpoints related to acute oral systemic toxicity were provided, on the basis of regulatory criteria and decision contexts used by ICCVAM agencies. In this study, two endpoints were considered to calibrate qualitative QSAR models: a) very toxic (VT), defining molecules as positive (very toxic) if their experimental LD₅₀ was lower than 50 mg/kg; b) nontoxic (NT), defining molecules positive (nontoxic) if their experimental LD₅₀ was greater than or equal to 2,000 mg/kg.

A detailed description of data processing to establish the experimental endpoint values is provided in the

collaborative project website^[4] and here summarized. When chemicals were associated to multiple experimental LD₅₀ values, the median of the lower quantile was computed to report a single LD₅₀ value. For the purpose of class identification for test LD₅₀ values on the limit of hazard categories, any experimental value associated to a greater than (>) or less than (<) symbol was adjusted, that is, >2,000 mg/kg was retained as 2,001 mg/kg and <2,000 mg/kg was retained as 1,999 mg/kg.^[4]

The database included 11,992 chemicals associated to experimental LD₅₀ values and was initially semi-randomly divided into a calibration set (75%) and validation set (25%), since equivalent coverage with respect to LD₅₀ distribution was considered for the splitting. The calibration set, including 8,994 chemicals, was thus distributed to the project participants for training and internal validation of QSAR models, together with the QSAR-ready structures in the form of SMILES, CASRN, and one value per molecule corresponding to each of the provided endpoints.

Molecular structures of the calibration set were initially curated to detect potential incorrect SMILES using the RDKit normalizer node of KNIME v. 3.6.1 (default settings). No molecules were removed from the original set of 8,994 chemicals; however, the SMILES of 50 chemicals were modified due to incorrect aromaticity definition. The list of these molecules and their SMILES are provided in Table S1 of the supplementary material of this manuscript.

The validation set (2,894 chemicals) was embedded within a large prediction set to carry out blinded validation and evaluation of models by the coordinators of the collaborative project. The prediction set, including 48,137 chemicals described by CASRN and chemical structures in machine-readable format, was distributed in a later stage after the model calibration had been completed. Characteristics of the calibration and validation sets are summarized in Table 1 for the two modelled endpoints. Note that the number of molecules included in the VT and NT calibration and validation sets differs: 2 and 13 of the 8,994 chemicals included in the calibration set had no experimental data for VT and NT endpoint, respectively; 4 chemicals of the

Table 1. Description of the calibration and validation sets. The total number of chemicals, as well as the number of positive and negative molecules, are reported for the very toxic (VT) and nontoxic (NT) endpoints.

Endpoint	Set	Molecules	Positive	Negative
VT (LD ₅₀ < 50 mg/Kg)	Calibration	8992	741	8251
	▶ Training	6744	556	6188
	▶ Test	2248	185	2063
	Validation	2894	243	2651
NT (LD ₅₀ ≥ 2000 mg/Kg)	Calibration	8981	3848	5133
	▶ Training	6735	2886	3849
	▶ Test	2246	962	1284
	Validation	2890	1235	1655

validation set were not associated to any label for the NT endpoint.

For the sake of simplicity, positive molecules are hereinafter intended as those whose experimental LD₅₀ value respects the defined limit of hazard categories; for the VT endpoint, molecules are labelled as positive if their LD₅₀ value is lower than 50 mg/kg, otherwise as negative; analogously, for the NT endpoint, molecules are defined as positive when associated to experimental LD₅₀ greater than or equal to 2,000 mg/kg, otherwise as negative. Thus, the calibration and validation sets of the VT endpoint resulted to be less balanced than the NT endpoint, the positive class being less represented.

The chemicals provided for model calibration were further divided into a training and a test set, containing respectively 75% and 25% of the molecules included in the calibration set (Table 1). When randomly splitting the calibration set, the same proportion between positive and negative molecules was maintained in the test and training sets (stratified splitting). Thus, 2,248 and 2,246 chemicals were selected as test compounds for the VT and NT endpoints, while the remaining 6,744 and 6,735 chemicals were included in the training sets, respectively.

The calibration, validation and prediction sets are available on-line at the project website.^[4]

2.2 Molecular Descriptors

Binary extended connectivity fingerprints were used as molecular descriptors for model calibration. Fingerprints (FPs) allow the complete representation of the molecular structural fragments in a series of binary digits (bits) encoding the presence or absence (as 1 and 0, respectively) of particular fragments and substructures in the molecule.^[5] There are several different algorithms for the calculation of binary FPs; among them, the most frequently used hashing algorithms produce compact FPs but with a "collision" of multiple molecular fragments in the same bit(s), and a consequent loss of one-to-one correspondence with molecular fragments.

In particular, Extended Connectivity Fingerprints (ECFPs) are obtained by exhaustively enumerating all circular fragments grown radially from each non-hydrogen atom of the molecule up to a specific radius and then hashing these fragments into a fixed-length bit vector.^[6] Thus, ECFPs are calculated through the generation of atom-centred fragments in an iterative procedure, which initially considers only atoms (radius equal to 0), then bonded atoms (radius equal to 1), and so on, until the pre-defined maximum radius. Additional to the fingerprint length and radius considered, the number of bits associated to each radial fragment can be selected. In this study, different types of ECFPs were calculated, considering the occurrences of any fragment identified in the molecule and the options listed in Table 2.

Table 2. Options used for the fingerprint calculation. For each type of ECFPs, the identification code, the minimum and maximum length, the number of bits and the bits per pattern are provided. The latest column specifies if atoms are distinguished according to the number of SSSR rings to which they belong.

Code	min length	max length	no. bits	bits/ pattern	SSSR
ECFP_1024_02R	0	2	1024	2	yes
ECFP_1024_02	0	2	1024	2	no
ECFP_1024_05	0	5	1024	3	no

As an additional set of descriptors, we used the fragments generated for calculating the ECFP: each molecule was characterized by a binary vector indicating the presence (1) or absence (0) of the molecular fragments derived from the calculation of ECFP_1024_05 (Table 2). Only fragments present in at least 10 chemicals were retained, leading to a total number of fragments equal to 3152.

Molecular descriptors calculated in this study for the training, test and validation molecules are available for download at the Milano Chemometrics and QSAR Research Group website: <http://www.michem.unimib.it/download/data/>.

2.3 Classification Models

Since two qualitative endpoints were considered in this study, QSAR models were calibrated through mathematical methods able to predict chemicals in a qualitative class, that is, positive or negative. Three different supervised classification algorithms were used in this framework: N-Nearest Neighbours (N3), Binned-Nearest Neighbours (BNN) and Naïve Bayes (NB), as described below:

- N-Nearest Neighbours (N3) is a local classification method, which, similarly to K-Nearest Neighbour (KNN), uses the experimental class of similar compounds to predict a target chemical.^[7] N3 considers all the available compounds as neighbours and, through an optimized α exponent, tunes their contribution as exponentially decreasing with decreasing their similarity to the target compound. Similarities between compounds were calculated with the Jaccard-Tanimoto similarity index,^[8] which is a benchmark similarity measure between binary vectors.
- Binned-Nearest Neighbours (BNN) is another variant of local classifiers.^[7] BNN predicts the class of a target chemical by means of a variable number (k) of similar compounds, according to the criterion of majority vote. To select the k compounds that have the largest similarity to the target and to be used for the prediction, similarity intervals (i.e., bins) are defined and only chemicals falling into the first non-empty interval (according to their similarity to the target) are considered for prediction.

Similarity bins are determined by optimization of a tuning parameter α , which defines the bin width. As for N3, the Jaccard-Tanimoto similarity index was used to estimate similarities between compounds.^[8]

- Naïve Bayes (NB) is a probabilistic method, based on Bayesian posterior probabilities derived from the maximum likelihood probabilities.^[9] Maximum likelihood probabilities are calculated for each bit of the binary molecular representation on the basis of the frequency of positive (or negative) compounds for the specific bit. A target compound is then predicted in the class associated to the highest posterior probability.

2.4 Model Validation and Assessment

Validation of QSAR models was carried out at different stages of the modelling phase, taking advantage of the data set splitting (see Table 1). Only the training molecules were used to calculate QSAR models, while chemicals included in the test and validation sets never participated in model optimisation and tuning of model parameters. In particular, chemicals included in the test sets were used to initially validate the models calibrated on the training samples, since at the first modelling stage experimental information for molecules included in the validation set was not yet available. As previously explained, the validation set was embedded within the prediction set and used to carry out blinded validation of models by the organizing committee of the ICCVAM collaborative project.^[3] Thus, validation chemicals were used to further validate the proposed QSAR models in a second stage. Finally, during the modelling phase, internal cross validation was used to tune and optimise model parameters, such as the α value for the N3 and BNN approaches. Cross validation was performed with training molecules divided in 5 cancellation groups defined by a venetian blind sampling protocol.^[10]

The classification performance of models was assessed through Sensitivity (S_n) and Specificity (S_p), which are expressed as follows:^[11]

$$S_n = 100 \cdot \frac{TP}{TP + FN}$$

$$S_p = 100 \cdot \frac{TN}{TN + FP}$$

where TP, TN, FP and FN are the number of true positives (positive compounds correctly classified), true negatives (negative compounds correctly classified), false positives (negative compounds classified as positive) and false negatives (positive compounds classified as negative), respectively. The Non-Error Rate (NER) was finally calculated as the average of specificity and sensitivity.^[11]

2.5 Applicability Domain

The applicability domain (AD) defines the structural domain where a given QSAR model can be applied and consequently only the predictions falling in the AD can be considered reliable. One of the primary bases in defining such a domain can be the structural space of the training molecules, as the reliable predictions from that model will be limited to the test samples belonging to this restricted structural domain.^[12]

Specific approaches for defining the model applicability domain were developed for each QSAR modelling method, as follows:

- N3: a target molecule was considered outside the model applicability domain if its highest similarity with respect to one of the defined experimental class spaces (defined in terms of N3 weights) was lower than a defined threshold. Details on N3 weights can be found in literature.^[7] The AD threshold was set as the 95% percentile of N3 weight values for the training chemicals.
- BNN: a target molecule was considered outside the model applicability domain if its average Jaccard-Tanimoto similarity to the most similar molecules of the first non-empty similarity bin was lower than a defined threshold. This threshold was set to 0.6 and 0.5 for the very toxic (VT) and nontoxic (NT) endpoints, respectively.
- NB: a target molecule was considered outside the model applicability domain if its average Jaccard-Tanimoto similarity to the most similar 100 molecules was lower than a defined threshold. This threshold was set as the 95% percentile of average Jaccard-Tanimoto similarity to the nearest 100 molecules for the training chemicals.

2.6 Consensus Analysis

For both the very toxic and nontoxic endpoints, three modelling methods, based on different mathematical approaches, were carried out to classify chemicals as either positive or negative. It has been demonstrated that data integration can increase reliability and reduce the effects of contradictory data by averaging results.^[13] In particular, qualitative consensus methods have been shown to reduce the effects of contradictory data by averaging predictions of models,^[14] when dealing with the combination of several QSAR models. Consensus analysis was thus applied to combine information and predictions obtained by the three different modelling techniques.

Bayesian consensus with discrete probability distributions was adopted for each modelled endpoint. Bayesian rules evaluate the a-priori probability that molecules belong to a specific class for each information source and then combine them to provide a joint probability.^[15] Only a brief introduction to this method is provided, since further details can be found in the literature.^[15,16]

The Bayesian consensus initially considers the first evidence “*e*”, e.g. the class predicted by the first-QSAR classification model. Then, the posterior probabilities $p(h_g | e)$ that hypothesis h_g is true given evidence e is calculated and used as new prior probability for the second step, where the class predicted by the second QSAR classification model is the new evidence e . Thus, the Bayes consensus proceeds with this iterative procedure, until predictions of all information sources have been entered in the process and the fused posterior probabilities are obtained. Finally, the Bayesian consensus predicts the class with the maximum posterior probability. This probability can also be used to accept or discard the predicted class depending on a predefined threshold,^[15] thus providing the so-called “protective” fusion approach. In this way, the Bayesian approach can detect situations where uncertainty is high and therefore a low reliability is associated to the prediction.

2.7 Software

Extended Connectivity binary fingerprints (ECFPs) and molecular fragments were calculated by means of Dragon 7.^[17] N3 and BNN QSAR models were calculated by means of N3-BNN Toolbox for MATLAB.^[7] Naïve Bayes and the consensus strategy were calculated with MATLAB routines written by the authors. Data and MATLAB code to calculate the models proposed in this study are available for download at the Milano Chemometrics and QSAR Research Group website: <http://www.michem.unimib.it/download/data/>.

3 Results

3.1 Individual QSAR Models and Consensus Analysis

QSAR models were developed according to the OECD principles, by ensuring unambiguity and transparency. In fact, five guiding principles have to be fulfilled to foster the applicability of QSARs,^[18] as proposed by the Organization for Economic Collaboration and Development (OECD): (1) a defined end point, (2) an unambiguous algorithm, (3) a defined domain of applicability, (4) appropriate measures

for goodness-of-fit, robustness and predictivity and (5) a mechanistic interpretation, if possible.

QSAR classification models based on N3, BNN and NB algorithms were initially calibrated with the 6744 training molecules for the VT endpoint and 6735 training molecules for the NT endpoint. Internal cross validation was performed on the training molecules to a) optimize the α value for the N3 and BNN approaches and b) select the best set of molecular descriptors for each classification method and for each endpoint. ECFPs with maximum path length equal to 2 were used as descriptors for models to predict the VT endpoint (ECFP_1024_02 in Table 2). The Naïve Bayes model (NB) was trained on ECFPs with the same characteristics as the other models, but the atoms were also distinguished according to the number of SSSR rings to which they belong (ECFP_1024_02R in Table 2). For the NT endpoint, models were calculated with descriptors based on ECFPs with maximum length equal to 5 for the BNN and N3 approaches (ECFP_1024_05 in Table 2), while the 3152 molecular fragments derived from the calculation of ECFP_1024_05 were used for the calibration of the Naïve Bayes classifier.

Afterwards, the calibrated classification models were used to predict molecules included in the test set (2248 and 2246 for the VT and NT endpoints, respectively) and the validation set (2894 and 2890 for the VT and NT endpoints, respectively). Applicability domain was then evaluated on the test and validation chemicals for each classification model and each modelled endpoint. Finally, the Bayesian consensus was applied to integrate the predictions achieved by the three individual QSAR models. A protective approach was adopted, as it is demonstrated that it can usually provide satisfactory results and improve those obtained by individual models.^[16] A protective threshold of 0.90 for the Bayesian approach was selected, which means that a consensus prediction was provided only if the final posterior probability was higher than 0.90. Only single model predictions evaluated inside the model applicability domain were considered for the consensus analysis.

Table 3 and Table 4 collect the classification statistics of the three individual QSAR models and their consensus for both the VT and NT endpoints. In particular, sensitivity (Sn) and specificity (Sp) are reported along with their arithmetic mean (non-error rate, NER). As previously defined, positive

Table 3. Classification performances for the very toxic (VT) endpoint of the three QSAR models (N3, BNN and NB) and their consensus approach estimated on the training, test and validation molecules. Type of molecular descriptors, sensitivity (Sn) and specificity (Sp) values, non-error rate (NER) and the percentage of non-predicted molecules (n.p.) are reported for each model.

Models	Descriptors	Training			Test			Validation				
		NER	Sn	Sp	NER	Sn	Sp	n.p.	NER	Sn	Sp	n.p.
BNN	ECFP_1024_02	80%	64%	97%	79%	61%	98%	27%	83%	69%	97%	28%
NB	ECFP_1024_02R	78%	73%	82%	77%	69%	84%	6%	76%	69%	83%	5%
N3	ECFP_1024_02	83%	80%	86%	85%	83%	87%	6%	86%	87%	85%	6%
Consensus		88%	81%	94%	88%	80%	96%	24%	90%	85%	94%	23%

Table 4. Classification performances for the nontoxic (NT) endpoint of the three QSAR models (N3, BNN and NB) and their consensus approach estimated on the training, test and validation molecules. Type of molecular descriptors (as reported in Table 2, sensitivity (Sn) and specificity (Sp) values, non error rate (NER) and the percentage of not predicted molecules (n.p.) are reported for each model.

Models	Descriptors	Training			Test				Validation			
		NER	Sn	Sp	NER	Sn	Sp	n.p.	NER	Sn	Sp	n.p.
BNN	ECFP_1024_05	79%	75%	82%	79%	75%	82%	30%	79%	74%	84%	30%
NB	Fragments_05	74%	72%	77%	73%	72%	74%	6%	72%	68%	77%	5%
N3	ECFP_1024_05	77%	79%	74%	77%	79%	76%	5%	77%	78%	77%	5%
Consensus		83%	82%	84%	82%	81%	84%	26%	82%	80%	84%	27%

molecules are intended as those whose experimental LD₅₀ value respects the defined limit of hazard categories. Sensitivity and specificity of the models for the VT endpoint (Table 3) refer to the percentage of very toxic molecules (positive, experimental LD₅₀ lower than 50 mg/kg) that are correctly predicted and the percentage of non-very toxic molecules (negative, experimental LD₅₀ equal or greater than 50 mg/kg) that are correctly predicted, respectively. Sensitivity and specificity of the models for the NT endpoint (Table 4) refer to the percentage of nontoxic molecules (positive, experimental LD₅₀ greater than or equal to 2,000 mg/kg) correctly predicted and the percentage of not nontoxic molecules (negative, experimental LD₅₀ lower than 2,000 mg/kg) correctly predicted, respectively. Finally, Table 3 and Table 4 also collect the percentages of non-predicted chemicals (n.p.), which are the molecules found outside the applicability domain of the individual QSAR models. In the case of consensus, a molecule may not be predicted due to a) posterior probability values lower than the defined protective threshold (0.90) in the Bayes protective approach, or b) if it is found outside the applicability domain of all three considered QSAR models.

These statistical measures should be considered all together to assess and compare the predictive classification ability of models, because they account for the different types of error and thus highlight different model behaviours,^[19] such as the ability to limit false positives (corresponding to high specificity values) and avoid false negatives (corresponding to high sensitivity values).

3.2 Results for the very Toxic Endpoint

When considering the results of the three individual QSAR models developed for the VT endpoint (Table 3), balanced performance was achieved, considering that the NER values obtained on the training, test and validation sets were similar. We can thus conclude that stable and non-overfitted models were obtained. Specific predictive behaviours can be analysed looking at the classification results achieved on test and validation molecules. N3 demonstrates slightly better classification performance (NER equal to 85% and 86% on test and validation set, respectively), while BNN is characterised by the highest rates of specificity (98% and

97% on test and validation set, respectively) but to the detriment of a significantly higher percentage of molecules considered outside the applicability domain (27% and 28%). Like BNN, the NB model is characterised by higher specificity than sensitivity, while N3 has more balanced performance on positive and negative molecules.

However, consensus approaches are expected to benefit from the different behaviours of QSAR models (characterised by both balanced and unbalanced values of sensitivity and specificity) participating in the prediction integration. Results obtained by the protective Bayesian approach of N3, BNN and NB models are reported in Table 3. The consensus analysis provided optimal classification performance, NER values being always higher than those provided by individual models for both training, test and validation sets. For the test and validation chemicals, specificity values were higher than 94%, meaning that most of the non-very toxic molecules (negative) were correctly classified, while keeping quite satisfactory sensitivity values (equal or higher than 80%). Another advantage of the protective Bayes approach is that its predictions are associated with posterior probabilities higher than 0.90 and they can, therefore, be evaluated with a greater level of confidence than those provided by the individual QSAR models.

Finally, when dealing with consensus approaches, the number of non-predicted molecules is likely to be higher due to potential disagreements among predictions of different individual models. As expected, the Bayes consensus was characterised by higher percentages of unpredicated chemicals than those of N3 and NB models. However, the percentages of non-predicted molecules achieved on the test and validations set (24% and 23%, respectively) did not represent any issue in the framework of the collaborative project. In fact, the predictions of the consensus approach will be integrated with those originated from models proposed by the participants of the project and thus chemicals unpredicated by the Bayesian consensus are expected to be predicted by the models calibrated by other participants.

3.3 Results for the Nontoxic Endpoint

Balanced classification performances were achieved also for the NT endpoint (Table 4), with comparable NER values obtained on the training, test and validation sets. Again, this indicates stable and non-overfitted behaviours for the proposed individual QSAR models.

BNN shows the best overall classification performance (NER equal to 79% on both test and validation sets), but with high percentage of molecules considered outside its applicability domain (30% for both test and validation sets). N3 is characterised by the highest rates of sensitivity (79% and 78% on test and validation set, respectively) and provides more balanced performances on positive and negative molecules. As for the VT endpoint, the Bayesian consensus approach provides optimal classification results, associated to higher values of both sensitivity and NER than individual QSAR models.

Finally, the Bayes consensus was again characterised by higher percentages of unpredicted chemicals than single N3 and NB models for both the test set (26%) and validation set (27%).

3.4 Combination of very Toxic and Nontoxic Endpoints

At this stage, independent results were obtained on the two considered endpoints of acute oral toxicity; models and predictions were in fact carried out separately for the nontoxic (NT) and very toxic (VT) endpoints. However, experimental and predicted values of these two endpoints can be combined, as they are both based on a categorisation of LD_{50} values according to the thresholds of hazard, that is, 50 mg/kg and 2,000 mg/kg for the VT and NT endpoints, respectively. Integration of VT and NT values can enhance the identification of hazardous chemicals, provide higher content of information associated to QSAR predictions and support and assist the analysis of consistency between the predictions obtained for the two endpoints on the same molecules. In fact, it may happen that the same chemical is predicted as positive for both endpoints, that is, it is predicted both as very toxic and nontoxic. In this case, aggregating the two predictions would allow to identify a less reliable prediction.

The experimental classes of NT and VT allow to identify a third experimental class, that of moderately toxic compounds, that is, compounds having LD_{50} greater than 50 mg/kg and smaller than 2,000 mg/kg. Thus, given a chemical associated to both NT and VT experimental class labels, the following toxicity categories can be derived from the combination of the two and molecules associated to intermediate (medium) toxicity can be defined (Figure 1):

- the molecule is labelled as very toxic (VT) if associated to a positive VT label and negative NT label, i.e. its experimental LD_{50} is lower than 50 mg/kg;

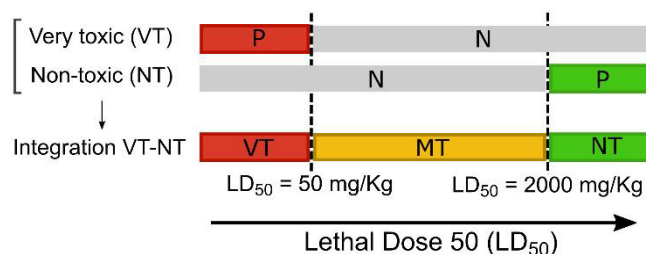


Figure 1. Combination of very toxic (VT) and nontoxic (NT) experimental labels. Compounds that were neither nontoxic ($LD_{50} \geq 2,000$ mg/Kg) nor very toxic ($LD_{50} < 50$ mg/Kg) were labelled as moderately toxic (MT), that is, compounds having 50 mg/Kg $\leq LD_{50} < 2,000$ mg/Kg. For each endpoint, P and N represent positive and negative according to the threshold, respectively.

- the molecule is labelled as nontoxic (NT) if associated to a negative VT label and positive NT label, i.e. its experimental LD_{50} is higher than or equal to 2,000 mg/kg;
- the molecule is labelled as medium toxic (MT) if associated to both negative VT and NT labels, i.e. its experimental LD_{50} ranges between 50 and 2,000 mg/kg;

This integration scheme was applied by combining experimental class labels of VT and NT datasets, both for training, test and validation molecules, leading to the creation of the integrated dataset. Table 5 reports the total number of chemicals and the number of molecules for each of the three experimental toxicity classes (very toxic [VT], moderately toxic [MT], nontoxic [NT]) for the training, test and validation sets. Note that the numbers of training and test chemicals in the integrated dataset differ from those included in the VT and NT datasets (Table 1). This is due to the previous independent splitting performed on the two datasets to select the test molecules. Thus, only chemicals present in both VT and NT training and test sets were retained and considered for the combination of the two endpoints.

Similarly, predictions obtained by means of the Bayesian consensus for the NT and VT endpoints were combined for each chemical, with the following scheme:

- the molecule is predicted as very toxic (VT) if associated to a positive VT prediction and negative NT prediction;
- the molecule is predicted as nontoxic (NT) if associated to a negative VT prediction and positive NT prediction;

Table 5. Characteristics of the integrated dataset. The total number of chemicals, as well as the number of very toxic (VT), medium toxic (MT) and nontoxic (NT) molecules are reported for the training, test and validation sets.

Set	Molecules	VT	MT	NT
Training	5043	421	2483	2139
Test	555	50	277	228
Validation	2890	243	1412	1235

- the molecule is predicted as medium toxic (MT) if associated to both negative VT and NT predictions;
- the molecule is unpredicted if associated to not consistent VT and NT predictions, that is, both VT and NT predictions are positive;
- the molecule is unpredicted if it is not assigned by at least one of the two Bayesian consensus approaches.

Experimental and predicted values for the integrated dataset were thus matched and classification performances were evaluated in terms of a) sensitivity of each toxicity category, that is, the percentage of correctly predicted molecules experimentally belonging to the class, b) non-error rate (NER), the average of class sensitivities and c) the percentage of non-predicted molecules. Classification results are reported in Table 6 for the training, test and validation sets of the integrated dataset.

Stable classification performances were achieved, with comparable NER and sensitivity values obtained on the three sets. These results confirmed the consistency and the agreement when integrating consensus predictions of the NT and VT endpoints, as they are associated to high NER values (around 80%, Table 6). When looking at sensitivity of classes, the intermediate toxicity category (MT) resulted to be characterised by lower values and, thus, a higher degree of overlap with respect to the two extreme classes (very toxic and nontoxic), as expected. However, acceptable percentages of correctly predicted chemicals for the MT class were obtained, sensitivity values being higher than 65% for the training, test and validation sets. On the opposite, sensitivity of VT and NT categories were characterised by higher values (greater than 80%).

4 Conclusions

In this work, new QSAR models for the prediction of acute oral toxicity are proposed. These models were calibrated in the framework of the "Predictive Models for Acute Oral Systemic Toxicity" project, coordinated by the ICCVAM Acute Toxicity Workgroup (U.S. Department of Health and Human Services), in collaboration with EPA's National Center for Computational Toxicology (NCCT). This is a collaborative project to develop in silico models of acute oral systemic

toxicity that predict specific endpoints needed by regulatory agencies.

In this study, the very toxic (LD₅₀ lower than 50 mg/kg) and nontoxic (LD₅₀ greater than or equal to 2,000 mg/kg) endpoints were considered to calibrate qualitative QSAR models. The modelling phase was undertaken by considering the five OCED principles for QSAR validity; thus, models were based on simple molecular descriptors (binary extended connectivity fingerprints and molecular fragments) and a Bayesian consensus approach integrating three classification algorithms (N3, BNN and Naïve Bayes). The applicability domain of each model was defined, and the validation was performed on two sets of molecules.

A blind validation on a set of external molecules provided in a later stage by the coordinators of the collaborative project was performed and the proposed models and their consensus proved to be robust and predictive for both the very toxic and nontoxic endpoints. The integration of predictions obtained for the two endpoints demonstrated absence of conflict and thus a good agreement between the models predicting the very toxic and nontoxic endpoints.

The results of this work will be used in combination with the efforts of the other members of the project consortium. Predictions of the proposed consensus approach will thus be merged with those originated from models proposed by the participants of the project. This integration is expected to facilitate the regulatory acceptance of in-silico predictions and will contribute to reduce or replace experimental tests for acute toxicity as required by regulatory authorities.

Data and MATLAB code to calculate the models proposed in this study are available for download at the Milano Chemometrics and QSAR Research Group website: <http://www.michem.unimib.it/download/data/>

Conflict of interest

None declared.

References

- [1] E. Walum, *Environ. Health Perspect.* **1998**, *106*, 497–503.
- [2] a) E. Schlede, E. Genschow, H. Spielmann, G. Stropp, D. Kayser, *Regul. Toxicol. Pharmacol.* **2005**, *42*, 15–23; b) P. A. Botham, *Toxicol. in Vitro* **2004**, *18*, 227–230.
- [3] a) N. C. Kleinstreuer, A. L. Karmaus, K. Mansouri, D. G. Allen, J. M. Fitzpatrick, G. Patlewicz, *Comput Toxicol.* **2018**, *8*, 21–24; b) D. Alberga, D. Trisciuzzi, K. Mansouri, G. F. Mangiatordi, O. Nicolotti, *Toxicol. Sci.* **2018**, *kfy255*, <https://doi.org/10.1093/toxsci/kfy255>.
- [4] Website of the "Predictive Models for Acute Oral Systemic Toxicity" collaborative project (accessed September 24, 2018): <https://ntp.niehs.nih.gov/go/tox-models>.
- [5] R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics*, Wiley-VCH, **2009**.

Table 6. Classification performances on training, test and validation molecules for the integrated dataset. Sensitivity of each toxicity class (Sn, percentage of correctly predicted molecules experimentally belonging to the class), average of class sensitivities (NER, non-error rate) and the percentage of non-predicted molecules (n.p) are reported for each set. VT: very toxic, MT: medium toxic, NT: nontoxic.

Set	NER	Sn VT	Sn MT	Sn NT	n.p.
Training	77%	81%	66%	84%	33%
Test	81%	91%	70%	81%	39%
Validation	79%	88%	66%	84%	36%

- [6] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- [7] R. Todeschini, D. Ballabio, M. Cassotti, V. Consonni, *J. Chem. Inf. Model.* **2015**, *55*, 2365–2374.
- [8] D. J. Rogers, T. T. Tanimoto, *Science (New York, N.Y.)* **1960**, *132*, 1115–1118.
- [9] T. Hastie, R. Tibshirani, J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Vol. second edition*, Springer Science, **2009**.
- [10] K. Varmuza, P. Filzmoser, *Introduction to multivariate statistical analysis in chemometrics*, CRC press, Boca Raton, FL, **2009**.
- [11] D. Ballabio, F. Grisoni, R. Todeschini, *Chemom. Intell. Lab. Sys.* **2018**, *174*, 33–44.
- [12] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, *Molecules* **2012**, *17*, 4791–4810.
- [13] J. Jaworska, S. Hoffmann, *Altex* **2010**, *27*, 231–242.
- [14] a) N. Baurin, J.-C. Mozziconacci, E. Arnoult, P. Chavatte, C. Marot, L. Morin-Allory, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 276–285; b) M. Hewitt, M. T. D. Cronin, J. C. Madden, P. H. Rowe, C. Johnson, A. Obi, S. J. Enoch, *J. Chem. Inf. Model.* **2007**, *47*, 1460–1468; c) J. R. Votano, M. Parham, L. H. Hall, L. B. Kier, S. Oloff, A. Tropsha, Q. Xie, W. Tong, *Mutagenesis* **2004**, *19*, 365–377.
- [15] a) E. Billoir, M. L. Delignette-Muller, A. R. R. Pery, S. Charles, *Environ. Sci. Technol.* **2008**, *42*, 8978–8984; b) A. Fernandez, A. Lombardo, R. Rallo, A. Roncaglioni, F. Giralt, E. Benfenati, *Environ. Int.* **2012**, *45*, 51–58.
- [16] D. Ballabio, F. Biganzoli, R. Todeschini, V. Consonni, *Environ. Toxicol. Chem.* **2017**, *99*, 1193–1216.
- [17] Kode SRL, Dragon (software for molecular descriptor calculation) version 7.0, **2016**.
- [18] OECD, Guidance document on the validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] models, **2007**.
- [19] ECHA European Chemicals Agency (ECHA). Guidance on information requirements and chemical safety assessment, Chapter R.6: QSARs and grouping of chemicals, **2008**.

Received: September 27, 2018

Accepted: November 26, 2018

Published online on December 14, 2018