# Consensus models to predict oral rat acute toxicity and validation on a dataset coming from the industrial context

F. Lunghini, G. Marcou, P. Azam, Dragos Horvath, R. Patoux, E. van Miert, A. Varnek

SAR AND QSAR IN ENVIRONMENTAL RESEARCH
https://doi.org/10.1080/1062936X.2019.1672089

# Consensus models to predict oral rat acute toxicity and validation on a dataset coming from the industrial context

F. Lunghini[a,b], G. Marcou[a], P. Azam[b], D. Horvath[a], R. Patoux[b], E. Van Miert[b] and A. Varnek[a]

[a]Laboratory of Chemoinformatics, University of Strasbourg, Strasbourg, France; [b]Toxicological and Environmental Risk Assessment unit, Solvay S.A., St. Fons, France

ABSTRACT

We report predictive models of acute oral systemic toxicity representing a follow-up of our previous work in the framework of the NICEATM project. It includes the update of original models through the addition of new data and an external validation of the models using a dataset relevant for the chemical industry context. A regression model for $LD_{50}$ and multi-class classification model for toxicity classes according to the Global Harmonized System categories were prepared. ISIDA descriptors were used to encode molecular structures. Machine learning algorithms included support vector machine (SVM), random forest (RF) and naïve Bayesian. Selected individual models were combined in consensus. The different datasets were compared using the generative topographic mapping approach. It appeared that the NICEATM datasets were lacking some relevant chemotypes for chemical industry. The new models trained on enlarged data sets have applicability domains (AD) sufficiently large to accommodate industrial compounds. The fraction of compounds inside the models' AD increased from 58% (NICEATM model) to 94% (new model). The increase of training sets improved models' prediction performance: RMSE values decreased from 0.56 to 0.47 and balanced accuracies increased from 0.69 to 0.71 for NICEATM and new models, respectively.

ARTICLE HISTORY
Received 12 July 2019
Accepted 21 September 2019

KEYWORDS
QSAR/QSPR; generative topographic mapping (GTM); oral rat acute toxicity; OECD principles; REACH

## Introduction

The estimation of the acute oral toxicity is a mandatory requirement under the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH, EC No. 1907/2006) legislation for substances manufactured or imported in quantities of 1 ton or more per year [1]. In most cases, this information is generated by performing an animal test according to the Organisation for Economic Co-operation and Development (OECD) guidelines. Until 2002, the reference guideline was OECD 401, however it was abolished for animal welfare reasons. Nowadays, more advanced guidelines are available which demand much less testing on animals and are likely to produce more reliable results [2]. Currently used guidelines are: OECD 420 (fixed dose procedure), OECD 423 (acute toxic class method), OECD 425 (up and down procedure) [3]. These guidelines are designed to classify the substances according to the

CONTACT G. Marcou g.marcou@unistra.fr; A. Varnek varnek@unistra.fr
Supplementary material for this article can be accessed at: https://doi.org/10.1080/1062936X.2019.1672089.

Global Harmonized System (GHS) categories and $LD_{50}$ values are only roughly estimated, at best.

To reduce animal testing, REACH encourages the use of non-testing methodologies, such as weight of evidence approaches, read across and QSAR modelling. In the past years, several QSAR models have already been developed to predict Acute Oral toxicity [4–7]. Some models are nowadays implemented in both commercial and free software (Table 1).

By the beginning of 2018, the National Toxicology Programme Interagency Centre for the Evaluation of Alternative Toxicological Methods (NICEATM) [8], as part of the effort to support the use of alternative methods, organized a worldwide workgroup to develop in silico models of acute oral toxicity. In particular, five relevant endpoints needed by regulatory agencies were targeted. These endpoints included (i) identification of 'very toxic' chemicals ($LD_{50}$ less than 50 mg/kg) and (ii) 'non-toxic' chemicals ($LD_{50}$ greater than or equal to 2000 mg/kg), (iii) point estimates for $LD_{50}$s, (iv) categorization of toxicity hazard using the U.S. Environmental Protection Agency (EPA) [9] and (v) the GHS [10] classification schemes. The NICEATM collected rat oral $LD_{50}$ data on over 15,000 substances from different publicly available databases and resources. The curated dataset was split into training and validation set. In the first stage, only the former was provided to the participants. The validation set was later used to externally validate the submitted models. The committee evaluated each model qualitatively with respect to the OECD principles [11] and quantitatively based on the predictive performance against the test set. Models were then employed to screen a large prediction set of ≈40 k chemicals of interest to different agencies and finally were also included into a consensus model, which leverages the strengths and compensate for the weaknesses of each individual approach [12]. More information about data preparation can be found on the workgroup website [8] and described by Ballabio et al. [13].

As participants, we submitted a regression model for $LD_{50}$ estimation. In this manuscript we present our modelling approach and a continuation of our work, including:

(1) Generation of a new multi-classification model based on GHS categories;
(2) Collection of additional acute oral toxicity data from several sources to extend the model's training set;
(3) External validation against a dataset relevant for the context of the chemical industry (hereafter named 'Industrial set'), provided by Solvay.

Finally, all public data was merged to constitute a 'Global set' (counting 11981 compounds) and models were updated. To the best of our knowledge, this is the largest reported dataset used for the development of QSARs predicting acute toxicity (Table 1).

Table 1. Tools for acute oral $LD_{50}$ estimation.

| Model | Tr. size | Employed descriptors | Algorithm | Ref. |
|---|---|---|---|---|
| TEST$^F$ | 7420 | Chemistry Development Kit (CDK) [15] | Consensus on five methods | [16] |
| ADMET$^C$ | 7150 | 2D, 3D molecular descriptors | Artificial neural network | [17] |
| ACD/Labs$^C$ | 8631 | Expert knowledge and structural descriptors | Expert knowledge and classification-SAR | [18] |
| TerraBase$^C$ | ≈ 10000 | Molecular structure descriptors | Probabilistic Neural Network | [19] |
| Accelrys$^C$ | ≈ 4000 | Molecular structure descriptors | Consensus on several models | [20] |

$^F$ = freely available; $^C$ = commercial

Our models are available through the online ISIDA/Predictor platform [14], available at the Laboratory of Chemoinformatics webpage: http://infochim.u-strasbg.fr/cgi-bin/predictor.cgi.

# Methods

## Modelling workflow

A graphical representation of the general workflow is shown in Figure 1; its main steps will be detailed in the present chapter.

## Data collection

Curated experimental data was distributed by the NICEATM workgroup. The original continuous $LD_{50}$ training and validation set counted respectively 6734 and 2174 compounds; analogously, for GHS classes 8960 and 2885 compounds were available. Additional oral rat $LD_{50}$ data was collected from the database of the European Chemicals Agency (ECHA) through the eChem portal [21], the relevant databases from the QSAR Toolbox software (SI, Section 1) [22] and the Toxicity Estimation Software Tool (TEST) training set [16]. Furthermore, a dataset on $LD_{50}$ (Industrial set) was provided by the industrial partner Solvay. This naming has been chosen in order to underline the existing structural differences between the compounds coming from an industrial context, which may represent new trends in large-scale production, from those available in public databases. To support

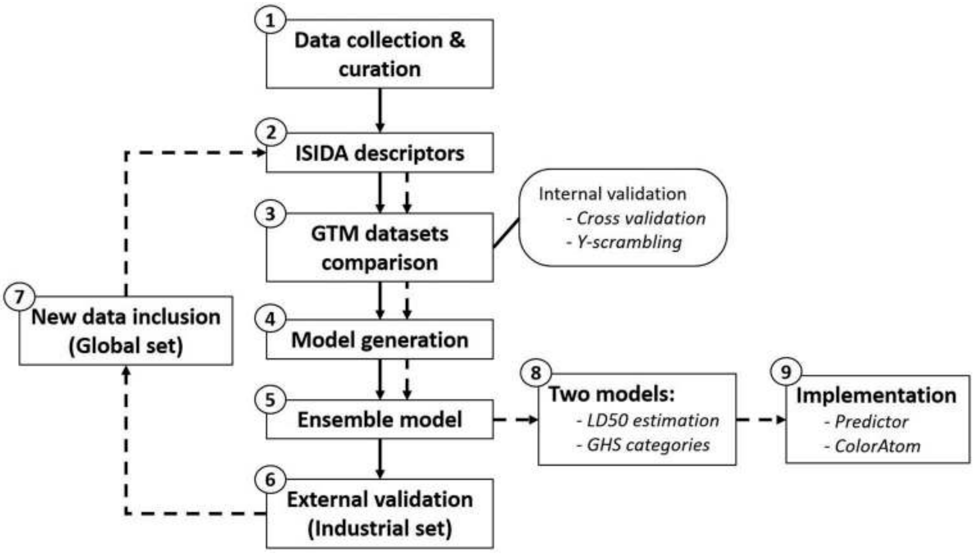


Figure 1. General workflow. (1) data is collected from different sources; (2) ISIDA descriptors encoding; (3) GTM is employed to compare the structural space of the datasets; (4), (5) individual models are trained and combined in consensus; (6) the Industrial set is used for external validation (7) the 'Global set' is issued by the merging of all public data and (8) models are updated; (9) models are published on the online platform.

this statement, collected databases were analysed through GTM (results section) and pairwise comparison of the Tanimoto similarities (SI, Section 2). Both approaches highlighted structural differences between their chemical spaces and the presence of unique chemotypes. Finally, an additional dataset of 462 compounds, not overlapping with the collected data, was provided by Solvay afterwards. This dataset (Blind set) was thus used to externally validate the last model version built on all collected data (public + industrial).

All collected public data (i.e. a total of 13682 unique compounds after the curation procedure) is available on Zenodo (DOI 10.5281/zenodo.3300664) with the respective $LD_{50}$ and/or GHS property; the industrial compounds cannot be provided due to confidentiality reasons.

## Data curation and standardization

To avoid additional sources of variability, data was limited to rat-only assays. Mixtures, polymers and UVCBs (Unknown or Variable composition, Complex reaction products or Biological materials) were discarded. Chemical standardization included: removal of salts/solvents, neutralization, removal of explicit hydrogens, aromatic representation for benzene rings, removal of stereo information, standardization of -nitro and -sulpho containing groups. This step was performed with a standardization workflow implemented in the Konstanz Information Miner (KNIME) [23]. In case of duplicates only one structure was kept and their $LD_{50}$ median value was selected (computed according to norm ISO16269-7). Multiple $LD_{50}$ values available the same compound were used to estimate the experimental error of the measurements. For each compound with at least 2 data points, a $LD_{50}$ range (maximum – minimum over reported values) was calculated, and the average of these range widths over concerned compounds was interpreted as the experimental error. GHS classes [10] were assigned based on the continuous $LD_{50}$ value, using the following thresholds (in mg/kg): ≤5, class 1; >5 and ≤50, class 2; >50 and ≤300, class 3; >300 and ≤2000, class 4; >2000, class 5. In order to maintain the same NICEATM classification system, the GHS ‘not classified’ category (i.e. > 5000 mg/kg) and GHS Category 5 (i.e. > 2000 mg/kg) were merged together in one unique class. For the regression model, $LD_{50}$ values originally expressed in mg/kg body weight were transformed to the inverse log of the molar dose ($pLD_{50}$ in mmol/kg body weight).

## Encoding of chemical structures

ISIDA property-label molecular descriptors [24] were employed. This led to the generation of dozens of different descriptor spaces which corresponds to different fragment sizes, topologies and encoded chemical information, called ‘colouration’ (elements labels, physical properties mapped on the atoms explicit or implicit chemical bonds, atom pairs). The number of fragments of the given descriptor space depends on selected fragmentation scheme. It varied from 387 (IIAB(2–2), atom centred fragments with radius 1) to 31623 (IIAB (2–5), atom centred fragments with radius 5), with an average of 7974 (SI, section 1).

## Generative topographic mapping

The chemical space of the collected databases was compared by means of the GTM approach [25], a dimensionality reduction method allowing the visualization of data distribution on a 2-dimensional map. A data property can be added as a 3rd axis forming such called activity landscape. Each landscape 'spot' on the 2D map is coloured according to the property value (either continuous or categorical); this value is the average property of the data subset concerned by that position on the landscape [26–28]. Two types of analysis were carried out: (i) the NICEATM dataset set was pairwise compared with the other databases (i.e. QSAR Toolbox, TEST, etc.); (ii) a map was generated on the Global set and the $LD_{50}$ value was used as property landscape. For the former, the goal was to identify which chemotypes were unique to the industrial context and under-represented in public available data. For the latter, the goal was to visualize how toxic and non-toxic compounds are distributed in the chemical space. The ISIDA descriptor space IIB(2–2) [24] associated to the best support vector machine (SVM) model (in terms of balanced accuracy) was chosen. These descriptors are based on molecular fragments consisting in an atom and information on the corresponding chemical bonds. The manifold [21] was built on the whole available chemical space (i.e. the Global set).

## Model generation

Employed machine learning approaches included: SVM with linear and radial basis function kernels, random forest (RF) and multinomial naïve Bayesian (NB). SVM models were generated with libSVM (v. 3.22) [29]; WEKA (v. 3.9.3) [30] was used for RF and for NB models. The SVM parameters (Cost and Gamma) corresponding to minimal RMSE in 3-fold CV were found by genetic algorithm driven optimization. The RMSE was estimated using a dedicated 3-fold CV, isolated from the cross-validation procedure used to evaluate the final models, mentioned below. Concerning RF, default parameters of WEKA were selected, with the number of generated trees equal to 100. No strategy was used to compensate the class imbalance in the dataset.

The modelling workflow is depicted in Figure 2: (1) dozens of ISIDA descriptor spaces (DSs) were generated (different fragment sizes and topologies); (2) for each DS, SVM and RF models were trained (individual models); (3) individual models were ranked according to their root mean squared error (RMSE) in 3-fold CV; (4) the best performing individual model

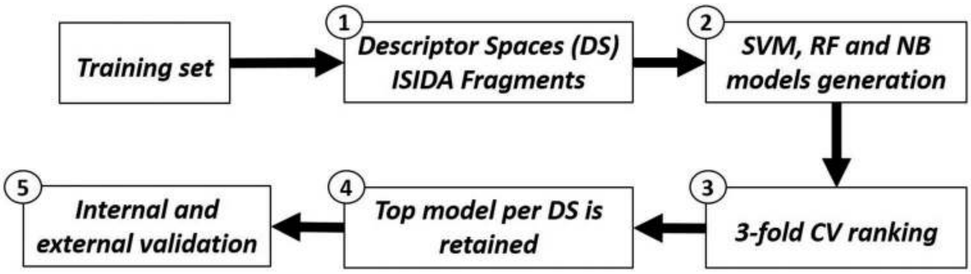

Training set
1
Descriptor Spaces (DS) ISIDA Fragments
2
SVM, RF and NB models generation
3
3-fold CV ranking
4
Top model per DS is retained
5
Internal and external validation

Figure 2. Model generation workflow.

for the given DS was retained; (5) models are internally and externally validated. Internal validation was carried out by random splitting 3-fold CV. This procedure was repeated 5 times after reshuffling (i.e. the property for each molecule is predicted 5 times). The Model quality criteria (see Figure 2) were assessed for each repetition followed by their averaging. During CV no further optimization of SVM parameters was performed. The absence of chance correlation was checked through the Y-scrambling procedure (repeated 100 times).

The Industrial set was used in external validation. In addition, it was predicted by the model TEST (Table 1). To evaluate the performance of regression models, the $r^2$ determination coefficient and the RMSE parameters are reported. For multi-class classification models, the sensitivity (Sn), specificity (Sp) and balanced accuracy (BA) are instead used. Dealing with multi-classes, the overall values for Sn, Sp and BA were computed as the weighted average among the classes based on the number of instances of the given class, following the same approach implemented in WEKA (v. 3.9.3) [30].

The following terminology is adopted:

- 'NICEATM original': the regression $LD_{50}$ model generated for the workgroup. Its training set is based solely on the NICEATM training set.
- 'NICEATM full': regression and multi-class classification models generated on all NICEATM data (i.e. training plus validation set).
- 'Global': regression and classification models generated on all collected data, externally validated on the Industrial set.

## Applicability domain

The applicability domain was evaluated trough the so-called 'fragment control' assessment (Figure 3, step 2): if a test molecule is found to have one fragment (i.e. a determined sequence of atoms and/or bonds) which is not present in the individual model, that molecule is marked to be outside the applicability domain since it is uncertain whether the model's predictions can be extrapolated to this not yet chartered chemical space zone [24].

## Consensus modelling

To derive the consensus decision, the following strategy was implemented (Figure 3). The ensemble decision is taken either by computing the median (regression model) or by a majority vote (classification model) from the individual models of the different algorithms considered together (step 1). All out-of-AD predictions (based on the fragment control) are excluded (step 2) and the consensus is computed (step 3). Finally, a 4-grade reliability scale is associated to the output (step 4), based on a combined score of (i) the concordance of the predictions and (ii) the % of individual models, out of the total, for which the compound was inside the AD. The former was estimated by the median absolute deviation for regression models or the entropy value for classification models (SI, section 2).

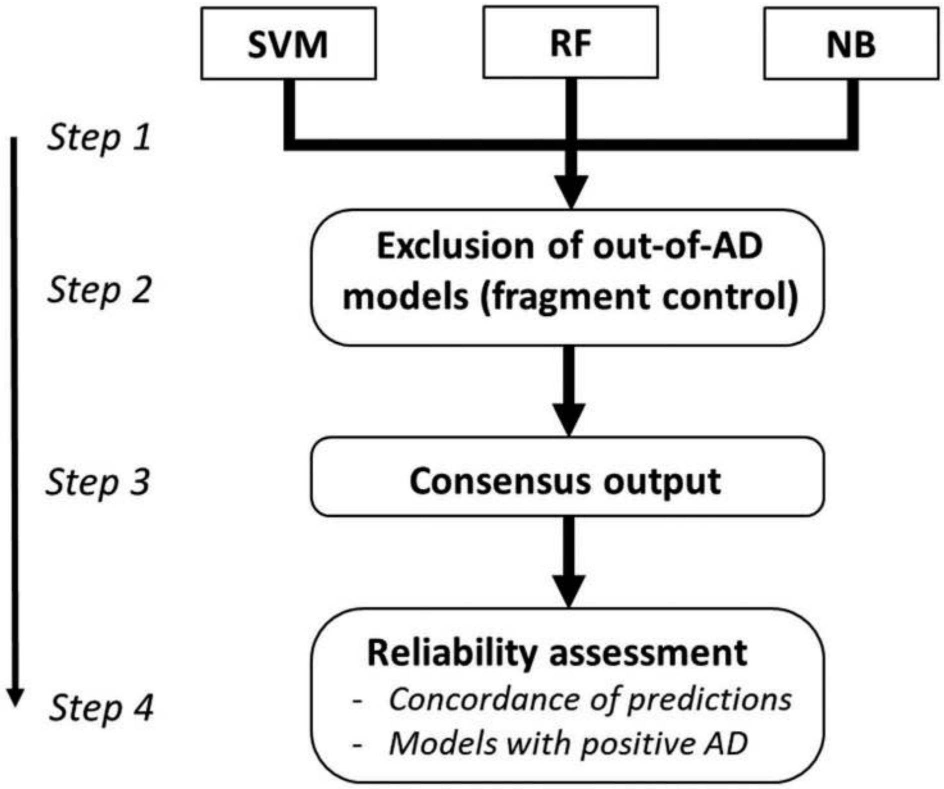

SVM
RF
NB
Step 1
Step 2
Step 3
Step 4
Exclusion of out-of-AD models (fragment control)
Consensus output
Reliability assessment
- Concordance of predictions
- Models with positive AD

Figure 3. Consensus model workflow. Step 1: predictions for each algorithm (SVM, RF and NB) are merged together; Step 2 & 3: the consensus is the average of the predictions, excluding those models identifying the compound as out of applicability domain; Step 4: reliability assessment is associated to the output.

## Graphical interpretation of predictions: coloratom

ISIDA ColorAtom [14] analyses local gradients of descriptors as reflecting their contributions to the variation of the modelled property [31]. A colour is assigned to each atom of the predicted molecule reflecting its positive or negative increment to the modelled property. This is a graphical representation of how the model interpreted the molecule for calculating the predicted value, not a mechanistic statement of the role played by each atom.

# Results

## Curated datasets

Table 2 reports summary statistics of the collected datasets; Figure 4 shows the distribution in the five GHS classes (SI, Section 2). The distribution pattern is the same for NICEATM, QSAR Toolbox and TEST datasets, for which the most populated class is the GHS class 4; on the other hand, for ECHA and the Industrial set the GHS class 5 is the

Table 2. Statistics of the curated datasets.

| | | Numerical $pLD_{50}$ statistics | | | GHS class repartition | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Curated datasets | Total no. | Min | Max | Mean | 1 | 2 | 3 | 4 | 5 |
| NICEATM[a] | 10863 | −2.71 | 4.6 | −0.48 | 180 | 650 | 1395 | 3359 | 2643 |
| QSAR Toolbox | 10531 | −3.34 | 4.21 | −0.53 | 276 | 760 | 1628 | 3987 | 3880 |
| TEST | 7315 | −2.71 | 4.21 | −0.45 | 237 | 661 | 1250 | 2819 | 2348 |
| ECHA | 1717 | −2.79 | 2.42 | −0.97 | 4 | 20 | 131 | 694 | 868 |
| Industrial set | 1563 | −4.57 | 1.31 | −0.95 | 1 | 9 | 121 | 437 | 995 |
| Blind set | 462 | −2.31 | 0.89 | 0.58 | 0 | 0 | 16 | 96 | 211 |
| Global set[b] | 11981 | −4.57 | 4.21 | −0.54 | 317 | 851 | 1773 | 4350 | 4690 |

[a]dataset used to build the 'NICEATM full' model; [b]dataset used to build the 'Global model'. The Global set[b] was issued by merging of the whole public data.

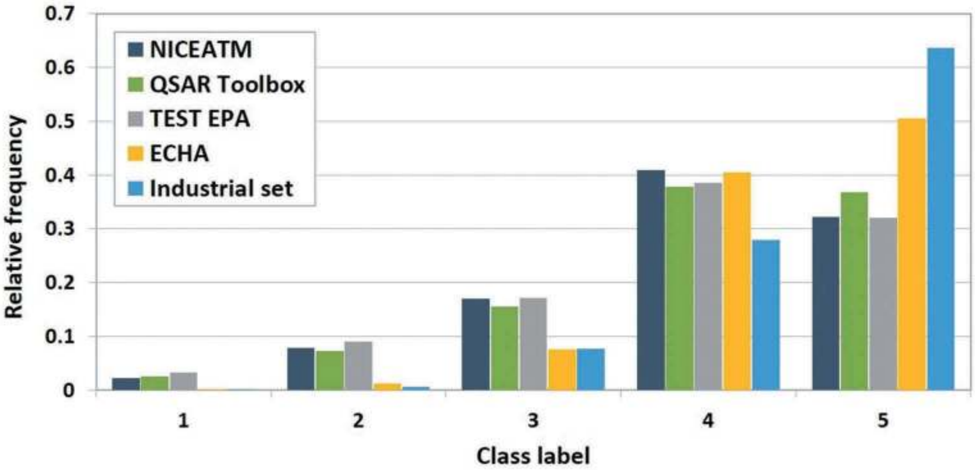


Figure 4. Class frequency distribution for the classification model.

most abundant. The experimental variability, when multiple values for the same compound were available, was calculated to be 0.40 log unit.

## Database comparison by GTM

Once the molecules are projected, landscapes are generated according to the envisaged property, and colours are assigned to the nodes of the map. In this context, two different landscapes were used: (i) the compound's database affiliation (i.e. NICEATM, QSAR Toolbox, etc.) and (ii) the $LD_{50}$ value.

### Database affiliation maps

With this analysis, the NICEATM was pairwise compared against all the remaining datasets. The goal was to verify if its set of compounds was sufficiently diverse to cover most of the chemical space, especially when confronted to the industrial context (i.e. the REACH registration dossiers on the ECHA database and the data provided by Solvay). Figure 5 shows all the pairwise comparison. Red areas are uniquely populated by the NICEATM dataset and blue by the others; intermediate colours are mixed populated areas. As visible from the first and the second landscape, NICEATM is almost

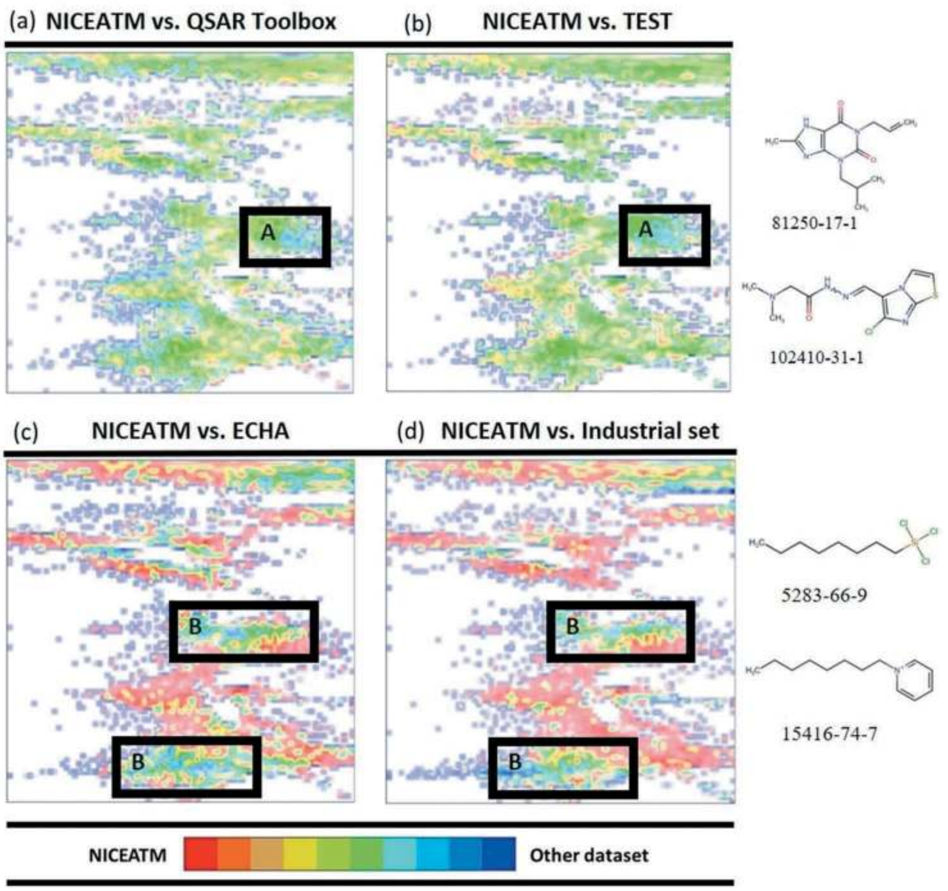


Figure 5. GTM database comparison. Each map compares the NICEATM vs. the other dataset: (a) QSAR Toolbox, (b) TEST, (c) ECHA and (d) Industrial set. Red regions are mainly populated by the NICEATM compounds and blue ones by the dataset it is compared to. White areas are empty regions of the map.

completely overlapping with QSAR Toolbox and TEST datasets. Some exceptions are two areas marked by the black rectangles ‘A’, indicative of some chemotypes under-sampled in the NICEATM dataset. For example, molecules with methylxanthine (CAS 81250-17-1; 66172-75-6) or imidazothiazole (CAS 102410-20-8; 102410-31-1) as substructures are almost unique to the QSAR Toolbox and TEST datasets.

For the third and fourth landscape, the situation is quite different: even though the chemical space is mainly dominated by NICEATM compounds (since its size is almost four times ECHA and the Industrial dataset), there are several spots dominated by ECHA or Industrial compounds (black rectangles ‘B’). Interestingly, these areas are localized on a similar X, Y position of the map, suggesting that the NICEATM dataset is missing some chemotypes which are, however, shared between the Industrial set and ECHA. To provide few examples, the chemotype containing a sequence of Halogen-Silicium-Halogen atoms (e.g. CAS 5283-66-9) and long aliphatic chains terminating with

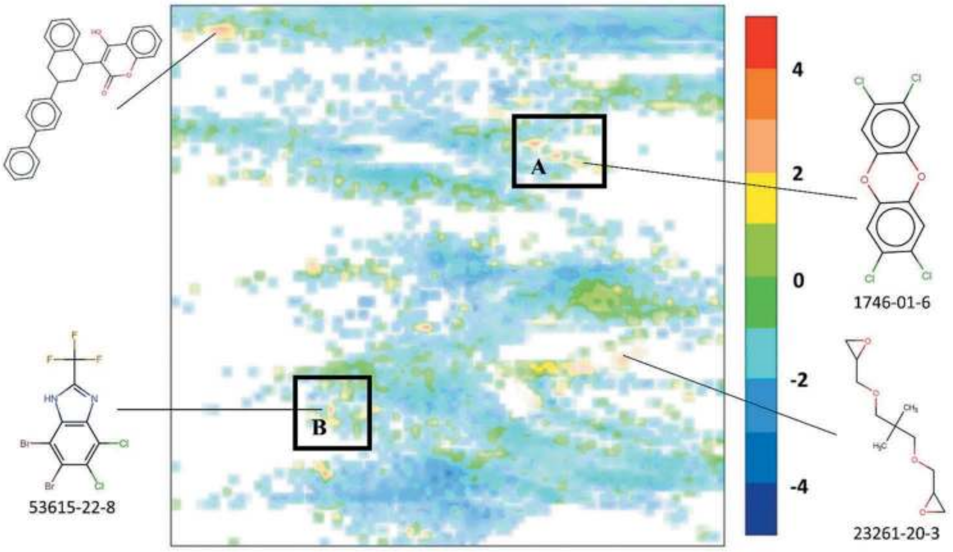


Figure 6. ‘Global map’ for $LD_{50}$. The map is built by merging all the available sources of data. Very toxic compounds are identified by red zones while less toxic compounds by blue ones.

a positively charged nitrogen-containing functional group (e.g. CAS 15416-74-7) are unknown or under-sampled to the other databases.

## $LD_{50}$ property map

Figure 5 reports the Global map coloured according to the $LD_{50}$ value. There are several spots of very highly toxic chemicals (indicated by black rectangles). For example, the area delimited by rectangle ‘A’ is populated by members of the dioxine and furane family (such as TCDD and TCDF); while in the area of the rectangle ‘B’ there is a collection of chemicals with the benzimidazole as substructure (e.g. CAS 89427-34-9) (Figure 6).

Table 3. Model performances.

| | | Internal validation (3-fold CV)[a] | | | External validation | |
|---|---|---|---|---|---|---|
| Regression | Model | $r^2$ | RMSE | $r^2$ Y-scrb | RMSE | Data coverage (%)[b] |
| | NICEATM original | 0.79 (0.050) | 0.55 (0.051) | 0.13 | 0.56 | 58 (287/479) |
| | NICEATM full | 0.77 (0.045) | 0.56 (0.053) | 0.15 | 0.51 | 87 (205/235) |
| | Global model | 0.78 (0.047) | 0.55 (0.055) | 0.12 | 0.47 | 94 (186/197) |
| | TEST [c] | – | – | – | 0.61 | 90 (293/322) |

| | | Internal validation (3-fold CV)[a] | | External validation | | | |
|---|---|---|---|---|---|---|---|
| Classification | Model | BA | BA Y-scrb | BA | Sn | Sp | Data coverage (%)[b] |
| | NICEATM full | 0.70 (0.031) | 0.30 | 0.69 | 0.74 | 0.63 | 82 (669/811) |
| | Global model | 0.70 (0.029) | 0.32 | 0.72 | 0.76 | 0.69 | 85 (635/744) |

Regression $LD_{50}$ model (upper part) and classification model (bottom part). [a]In brackets, the standard deviation computed in the 3-fold CV is reported for the $r^2$ and RMSE values averaged over the number of repetitions. External validation is based on the Industrial set. BA = balanced accuracy, Sn = sensitivity, Sp = specificity. [b]The first number is the data coverage in %; the number between the parentheses is a ratio of the number of compounds inside AD and the total number of compounds. [c]results from the TEST model.

Table 4. Performance of selected machine learning methods.

| Regression | Method | External validation<br>RMSE | | | Data coverage (%) |
|---|---|---|---|---|---|
| | Random forest | 0.47 | | | 94 |
| | SVM linear kernel | 0.51 | | | 82 |
| | SVM RBF kernel | 0.50 | | | 97 |
| | Global model | 0.47 | | | 94 |
| | | External validation | | | |
| Classification | Method | BA | Sn | Sp | Data coverage (%) |
| | Random forest | 0.74 | 0.82 | 0.66 | 81 |
| | SVM linear kernel | 0.69 | 0.81 | 0.56 | 87 |
| | SVM RBF kernel | 0.73 | 0.81 | 0.66 | 85 |
| | Naïve Bayesian | 0.64 | 0.60 | 0.68 | 80 |
| | Global model | 0.72 | 0.76 | 0.69 | 85 |

Regression $LD_{50}$ model (upper part) and classification model (bottom part). External validation is based on the Industrial set. BA = balanced accuracy, Sn = sensitivity, Sp = specificity.

## Model performances

Table 3 reports performances of the generated models: regression $LD_{50}$ model (top) and classification model (bottom). In addition, the performances of the TEST tool are reported for $LD_{50}$. Individual machine learning algorithms performances are reported in Table 4. Overall, all the models scored a good prediction accuracy on the Industrial set, with RMSE values ranging from 0.47 to 0.56 and BA values from 0.69 to 0.72. TEST showed a good data coverage, being able to predict the 90% of the Industrial set. However, its prediction accuracy is worse (0.61 RMSE). The addition of new data is directly correlated to both an increase of prediction accuracy and data coverage. The latter increased from 58% for the NICEATM original model to 94% for the Global model (regression) and from 82 to 85% (classification models). This reflects that the NICEATM data are more comprehensive regarding GHS data. The contamination of models by chance correlations is limited as monitored by Y-scrambling: the maximum observed $r^2$ and BA metrics had very low values ($r^2 < 0.2$ and BA <0.5). Overall, all the models are robust and well generalizable: performances in external validation are comparable to those in cross-validation and the data coverage reaches very high levels.

## Performances on blind set

Finally, the last version of the model (built on all collected data, i.e. public + industrial) was challenged to predict a new list of 462 unique compounds made available afterwards. Of them, 224 had a precise estimation of $LD_{50}$; while 347 had only the categorical statement. Thus, both the regression and classification models were used.

For confidentiality reasons, this dataset cannot be disclosed, and only some general information can be provided. It comprises quite heterogeneous chemical structures,

Table 5. Performances of public and industrial data ensemble models on the blind set.

| | Regression | | | Classification | | | |
|---|---|---|---|---|---|---|---|
| Blind set | $r^2$ | RMSE | Data coverage (%) | Sn | Sp | BA | Data coverage (%) |
| | 0.3 | 0.48 | 92 (207/224) | 0.77 | 0.97 | 0.87 | 93 (303/323) |

BA = balanced accuracy, Sn = sensitivity, Sp = specificity.

from high molecular weight compounds such as long chain aliphatic surfactants and halogenated biphenyls to much smaller ones such as phenol derivates and simple amides. A good number of compounds are organofluorine derivatives. The molecular weight ranges from 41 to 1094 with an experimental $pLD_{50}$ from −2.31 to 0.89 log unit. This dataset is mainly 'non-toxic', as almost 60% of the compounds are not classified under the GHS system (i.e. $LD_{50}$ > 2000 mg/kg).

Performances for the regression model are similar to the previous external validation ($RMSE_{blind}$ = 0.48 vs. $RMSE_{ext}$ = 0.47; Tables 3 and 5). In both instances, the prediction accuracy is better than the one estimated through cross-validation ($RMSE_{cv}$ = 0.55). On the other hand, the classification model performed better ($BA_{blind}$ = 0.87 vs. $BA_{ext}$ = 0.72, Tables 3 and 5). This is probably due to the unbalanced nature of the Blind set, as the majority of the compounds belong to GHS class 5.

The Blind set $r^2$ value may appear disappointing at first sight. However, it must be noticed that its $pLD_{50}$ property range is considerably smaller than the Global model's one (−4.57–4.21). Figure 7 depicts experimental/predicted scatterplot of the Global model's training set (evaluated in 3-fold CV) overlapped with the Blind set. As expected, the Blind set covers only a fraction of the entire property range: this explains the low determination coefficient value.

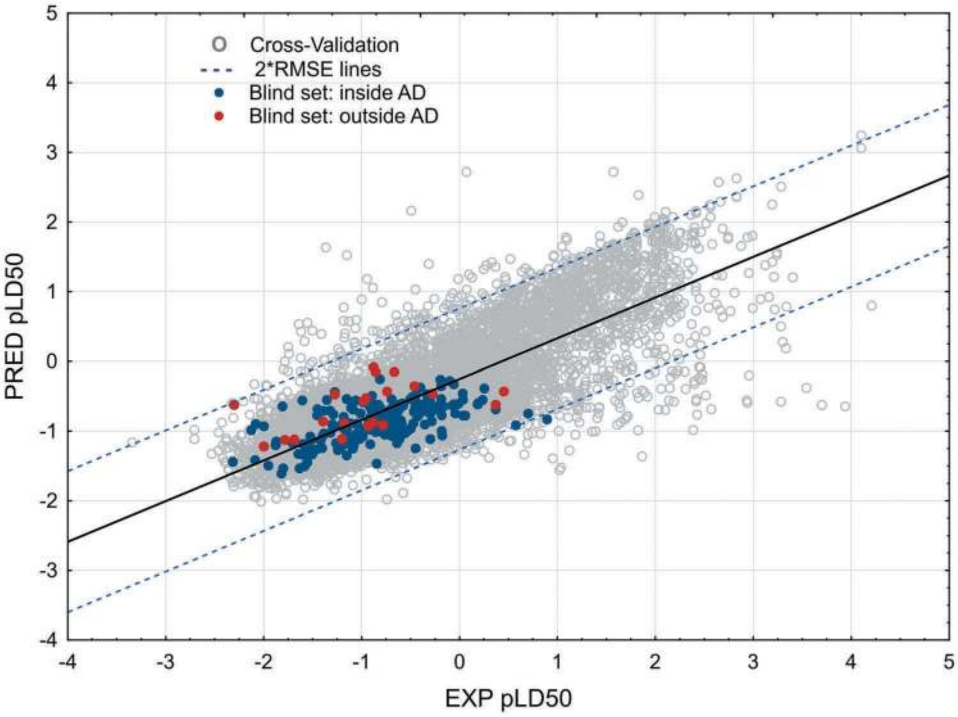
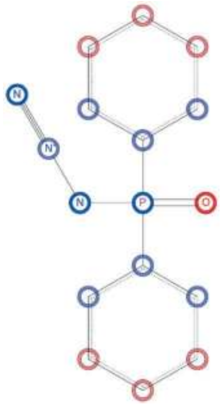
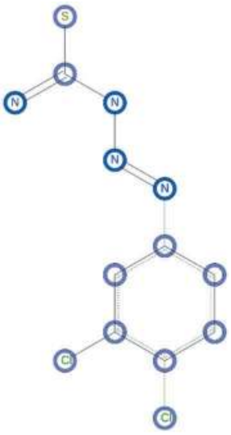
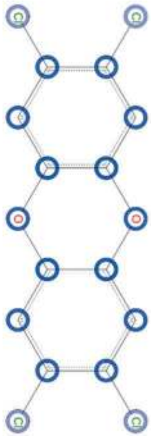


Figure 7. Blind set scatterplot. Grey points represent the training set evaluated in 3-fold CV; red and blue points indicate Blind set molecules outside and inside the AD, respectively. Blue dashed lines mark the ± 2 $RMSE_{cv}$. limits.

Table 6. ColorAtom output.

| Diphenylphosphinyl azide | Chloropromurite | TCDD |
|---|---|---|
| CAS 4129-17-3<br>$LD_{50}$ = 240 mg/Kg | CAS 5836-73-7<br>$LD_{50}$ = 1.0 mg/Kg | CAS 1746-01-6<br>$LD_{50}$ = 0.02 mg/kg |
|  |  |  |

Colours refer to atomic contribution to the predicted value of the property (i.e. $pLD_{50}$ values). Red colour means that the atom contributes to decrease its value (lowering the toxicity); while blue means an increase of its value (i.e. increasing the toxicity).

## Model interpretation with coloratom

For in-depth structure-activity dependence analysis, Table 6 reports three molecules chosen as examples for the ColorAtom: diphenylphosphinyl azide, chloropromurite and TCDD. As expected, as the compounds become more toxic, 'blue-coloured' atoms become dominant. For the first compound, the 'triazo-' substructure is the main driver for its correct prediction as an acute toxic. Similarly, chloropromurite presents two functional groups which are associated with enhanced toxicity: the 'diazo' (CNN) and the 'thiocyanate' (SCN). Finally, all the atoms of TCDD are represented as promoters of toxicity. In these cases, the colouration patterns are actually in agreement with the mechanistical interpretation of the analysed functional groups [32,33]. SI, Section 3 reports additional examples of compounds with the same functional groups that showed the same colouration scheme.

## Discussion

Among the QSAR tools for the estimation of the oral rat acute toxicity reported in Table 1, only one is freely available (TEST). The collaborative NICEATM workshop aimed at filling this gap, by proposing a set of new models which will be freely available [12], implemented in the open source platform OPERA [34]. On the Industrial set, the predictive power of the models (regression and classification) was found to be reasonably high, with RMSE values of 0.47–0.56 and BA values of 0.69–0.71 (5 five classes) for the NICEATM and the Global models, respectively. Data coverage was quite unsatisfactory with the original NICEATM model (58% on the Industrial set), but after the addition of new data from several databases (QSAR

Toolbox, TEST, ECHA) it significantly improved: reaching 85 and 94% for the classification and the regression model, respectively. New data improved models' predictive power as well, with the biggest improvement for the regression model, where the RMSE decreased from 0.56 to 0.47. Finally, new models were built on the ensemble of public data and Industrial data. Cross-validation performances for the regression model were $r^2 = 0.78$ and RMSE = 0.53; while for the classification model BA = 0.69. These models were also externally validated on the Blind set (Table 5), showing good prediction accuracy and data coverage: RMSE = 0.47 with 92% inside AD (regression); and BA = 0.87 with 93% inside AD (classification).

GTM was employed to show positions of 109 'out-of-AD' compounds (Table 3, bottom part) in the public data chemical space, which constitutes the training sets of the models (Figure 8). As expected, the majority of them are located in the regions mainly populated by external set compounds (blue areas), indicating that their chemotypes are quite unique and non-overlapping with those in the models' training sets. For example, compound CAS 34762-90-8 presents the unique chemotype – $N^+BCl_3$. Some compounds are singletons far away from the occupied chemical space, such as CAS 24108–89, a pigment characterized by a very complex and diverse chemical structure. On the other hand, there are some out-of-AD compounds projected in areas of the public data chemical space. This happens when the given molecule both shares several functional groups with the training set compound and contains new chemotype. For example, drometrizole trisiloxane (CAS 155633-54-8), contains trisiloxane

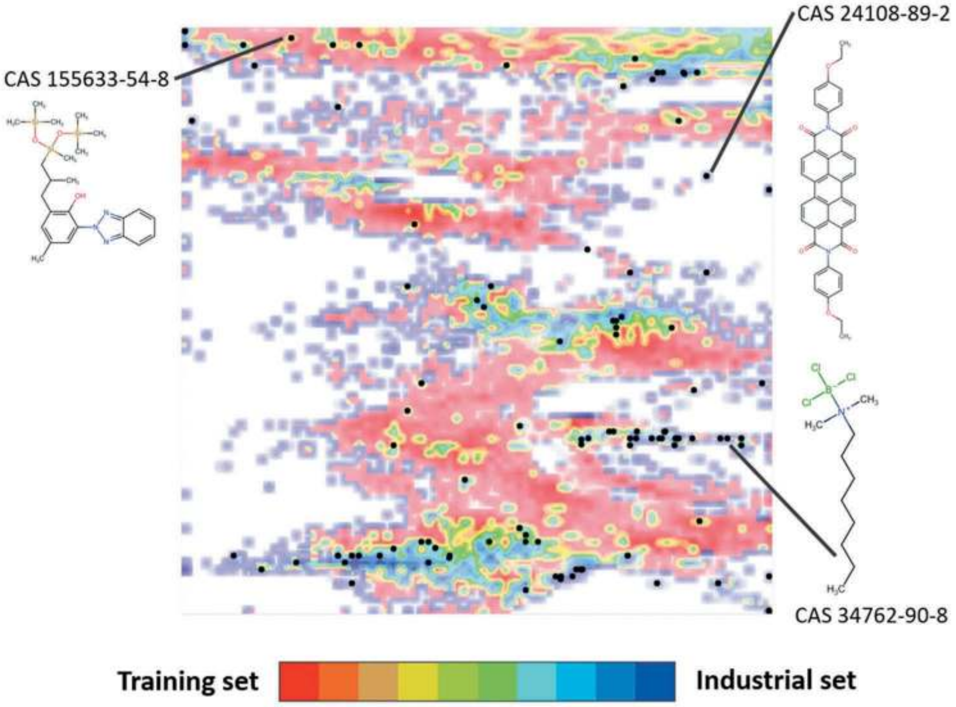


Figure 8. Tracking the out-of-AD compounds in the chemical space. Zones populated by the training set and Industrial set compounds are highlighted in colour. Black points represent the projections of 109 out of AD compounds.

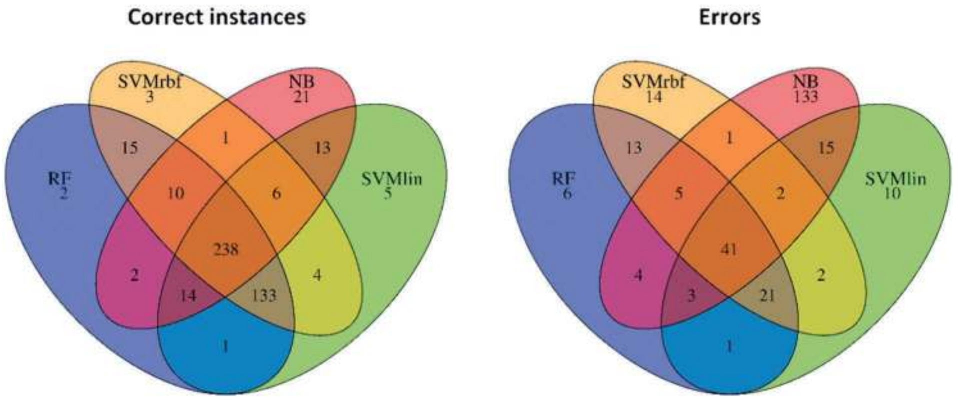


Figure 9. Venn diagrams comparing individual multi-class classification models performances in external validation on Industrial set. Left and right diagrams correspond to correct and erroneous predictions, respectively.

motif absent in the training set, and drometrizole motif present in several training set compounds.

Performances of individual models for the multi-class classification in external validation on Industrial set are represented by means of Venn diagrams (Figure 9). Comparison is performed for both the correct (left) and for erroneous (right) predictions. These results support the conclusion about the robustness of consensus model, since great majority of instances (238) were simultaneously correctly predicted by all four machine-learning algorithms.

Our developed models follow the OECD principles [11]. The endpoint ($LD_{50}$) is well defined. Goodness-of-fit, robustness and predictivity were evaluated using internal and external 3-fold Cross-Validation (CV), Y-scrambling, and external validation [35–37]. The AD of the models was defined using a fragment control assessment [24] together with a reliability scoring function.

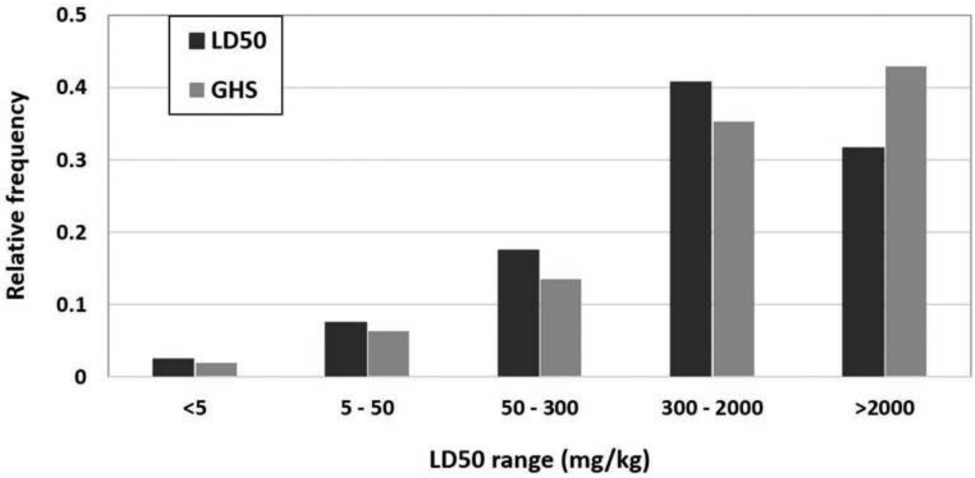


Figure 10. Continuous-$LD_{50}$ and the GHS-classes distribution comparison.

Figure 10 depicts the relative frequency distribution for the continuous-$LD_{50}$ and the GHS-classes for the full NICEATM dataset (training and evaluation set). It is interesting to notice that $LD_{50}$ data is always more frequent than categorical assays for more toxic compounds, with the biggest difference (+15%) for the medium toxicity class (GHS class 4, i.e. 300–2000 mg/kg). On the other hand, for low toxicity values (GHS category 5, i.e. >2000 mg/kg,) categorical data becomes much more frequent. This is related to the current regulatory requirements: in case the substance shows high toxicity, it could be more advantageous for the registrant to have the precise $LD_{50}$, in order to avoid a potential overestimation of the compound's toxicity, leading to a less desirable GHS classification. On the other hand, when the substance is far from GHS thresholds, a looser toxicity estimation could be enough. This bias of the data is also reflected in the model's learned rules: we noticed that the regression $LD_{50}$ model tends to overestimate the toxicity of some very low toxic compounds. Furthermore, as mentioned in the introduction, current guidelines do not foresee anymore the precise estimation of $LD_{50}$. Instead, the goal is to perform limit tests (OECD 420, 423, 425) for estimating the GHS categories, which allows the use of fewer animals. For this reason, new $LD_{50}$ data is unlikely to be generated, and future in-silico models will have to be updated based on the new categorical data.

## Conclusions

In this work we report predictive models of acute oral toxicity obtained in the context of the National Toxicology Programme Interagency Centre for the Evaluation of Alternative Toxicological Methods (NICEATM) workgroup [14,18].

The datasets including 11211 and 13680 compounds for 'Global' regression and classification models respectively, were collected from the publicly available sources. To our knowledge, these are the biggest datasets ever used for the modelling of oral acute toxicity in rodent.

The models were obtained using ISIDA fragment descriptors [24] and support vector machine, random forest and naïve Bayes machine learning methods. Compared to our contribution to the NICEATM project in this paper (i) a new classification model based on GHS toxicity categories was generated (ii) Global models were generated by collecting new data.

The predictive performance of the models was assessed on independent Industrial set provided by Solvay. It has been demonstrated that both regression and classification Global models obtained in this work (RMSE = 0.47 and BA = 0.72) perform better than the previously reported NICEATM models (RMSE = 0.56 and BA = 0.69). Moreover, the Global models have much larger applicability domain: the data coverage on the Industrial set is 85% and 82% (classification) and 94% and 58% (regression) for Global and NICEATM models, respectively. Finally, new models built on the ensemble of public data and Industrial dataset were validated on a set of 462 new structures provided by Solvay. This blind test proved reasonably high predictive power of the models: RMSE = 0.48 and BA = 0.87 for regression and classification, respectively.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## ORCID

G. Marcou http://orcid.org/0000-0003-1676-6708
D. Horvath http://orcid.org/0000-0003-0173-5714

## References

[1] European Commission, Regulation (EC) no 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the registration, evaluation, authorisation and restriction of chemicals (REACH), establishing a European Chemicals Agency, amending directive 1999/45/ECC and repealing Council regulation (EEC) No 793/93 and commission regulation (EC) No 1488/94 as well as Council directive 76/769/EEC and commission directives 91/155/EEC, 93/67/EEC, 93/105/EEC and 2000/21/EC, Off. J. Eur. Union. 50 (2007), pp. 1–281.

[2] A. Gissi, K. Louekari, L. Hoffstadt, N. Bornatowicz, and A.M. Aparicio, Alternative acute oral toxicity assessment under REACH based on sub-acute toxicity values, ALTEX 34 (2017), pp. 353–361. doi:10.14573/altex.1609121.

[3] OECD Guidelines for the Testing of Chemicals, Section 4: Health Effects, Organisation for Economic Cooperation and Development (OECD), Paris, FR, 2019. Available at http://www.oecd.org/env/ehs/testing/oecdguidelinesforthetestingofchemicals.htm.

[4] I. Tsakovska, I. Lessigiarska, T. Netzeva, and A.P. Worth, A mini review of mammalian toxicity (Q)SAR models, QSAR Comb. Sci. 27 (2008), pp. 41–48. doi:10.1002/qsar.200710107.

[5] J.X. Guo, J.J.-Q. Wu, J.B. Wright, and G.H. Lushington, Mechanistic insight into acetylcholinesterase inhibition and acute toxicity of organophosphorus compounds: A molecular modeling study, Chem. Res. Toxicol. 19 (2006), pp. 209–216. doi:10.1021/tx050090r.

[6] A.P. Freidig, S. Dekkers, M. Verwei, E. Zvinavashe, J.G.M. Bessems, and J.J.M. van de Sandt, Development of a QSAR for worst case estimates of acute toxicity of chemically reactive compounds, Toxicol. Lett. 170 (2007), pp. 214–222. doi:10.1016/j.toxlet.2007.03.008.

[7] A.A. Toropov, B.F. Rasulev, and J. Leszczynski, QSAR modeling of acute toxicity for nitrobenzene derivatives towards rats: Comparative analysis by MLRA and optimal descriptors, QSAR Comb. Sci. 26 (2007), pp. 686–693. doi:10.1002/qsar.200610135.

[8] Legal Information Institute, Predictive Models for Acute Oral Systemic Toxicity, National Toxicology Program, US Department of Health and Human Services, Bethesda, Maryland, US, 2019. Available at https://ntp.niehs.nih.gov/pubhealth/evalatm/test-method-evaluations/acute-systemic-tox/models/index.html,

[9] Electronic Code of Federal Regulations (40 Cfr Part 156), United States Government Publishing Office, Washington DC, US, 2019. Available at https://www.law.cornell.edu/cfr/text/40/part-156.html.

[10] The European Parliament and the Council of the European Union, Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006, Off. J. Eur. Union 353 (2008), pp. 1–1389.

[11] OECD, Guidance document on the validation of (quantitative) Structure Activity Relationship [(Q)SAR] models, Tech. Rep. ENV/JM/MONO(2007)2, Organisation for Economic Cooperation and Development, Paris, FR, 2007.

[12] N.C. Kleinstreuer, A.L. Karmaus, K. Mansouri, D.G. Allen, J.M. Fitzpatrickc, and G. Patlewicz, Predictive models for acute oral systemic toxicity: A workshop to bridge the gap from research to regulation, Comput. Tox. 201 (2018), pp. 489–492.

[13] D. Ballabio, F. Grisoni, V. Consonni, and R. Todeschini, Integrated QSAR models to predict acute oral systemic toxicity, preprint (2019), submitted for publication. Available at https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201800124.

[14] G. Marcou, D. Horvath, F. Bonachera and A. Varnek, Laboratoire De Chemoinformatique UMR 7140 CNRS, University of Strasbourg, Strasbourg, FR, 2019. Available at http://infochim.u-strasbg.fr/cgi-bin/predictor.cgi.

[15] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen, The chemistry development kit (CDK): An open-source java library for chemo- and bioinformatics, J. Chem. Inf. Comput. Sci. 43 (2003), pp. 493–500. doi:10.1021/ci025584y.

[16] T. Martin, P. Harten, and D. Young, (TEST) Toxicity Estimation Software Tool V 4.1, US Environmental Protection Agency, 2012; software available at https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test.

[17] Simulations Plus Inc, ADMET Predictor, Simulations Plus Inc., Lancaster, US, 2019; software available at http://www.simulations-plus.com/.

[18] ACD/Labs, ACD/Percepta Platform V 2018.1, Advanced Chemistry Development, Inc. (ACD/Labs), 2019; software available at http://www.acdlabs.com/.

[19] TerraBase, TerraQSAR Biological Effect Programs, TerraBase Inc., 2006; software available at http://www.terrabase-inc.com/.

[20] Accelrys, (TOPKAT) TOxicity Prediction by Komputer Assisted Technology V 3.1, Accelrys software Inc., San Diego, CA, US, 2019; software available at http://www.3dsbiovia.com/.

[21] OECD, Data from: EChemPortal: Global portal to information on chemical substances, Organisation for Economic Co-operation Development, dataset available at https://www.echemportal.org/echemportal/index.action.

[22] OASIS, The OECD QSAR toolbox v 4.1, OASIS Laboratory of Mathematical Chemistry, 2017; software available at http://www.oecd.org/chemicalsafety/risk-assessment.

[23] M. Berthold, N. Cebron, F. Dill, T. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel, KNIME - the konstanz information miner: Version 2.0 and beyond, SIGKDD Explor. 11 (2009), pp. 26–31. doi:10.1145/1656274.1656280.

[24] F. Ruggiu, G. Marcou, A. Varnek, and D. Horvath, ISIDA property-labelled fragment descriptors, Mol. Inform. 29 (2010), pp. 855–868. doi:10.1002/minf.201000099.

[25] C.M. Bishop, M. Svensén, C.K.I. Williams, and M. Svens, The generative topographic mapping, Neural Comput. 10 (1998), pp. 215–234. doi:10.1162/089976698300017953.

[26] N. Kireeva, I.I. Baskin, H.A. Gaspar, D. Horvath, G. Marcou, and A. Varnek, Generative topographic mapping (GTM): Universal tool for data visualization, structure-activity modeling and dataset comparison, Mol. Inform. 31 (2012), pp. 301–312. doi:10.1002/minf.201100163.

[27] H.A. Gaspar, I.I. Baskin, G. Marcou, D. Horvath, and A. Varnek, Chemical data visualization and analysis with incremental generative topographic mapping: Big data challenge, J. Chem. Inf. Model. 55 (2015), pp. 84–94. doi:10.1021/ci500575y.

[28] D. Horvath, I. Baskin, G. Marcou, and A. Varnek, Generative topographic mapping of conformational space, Mol. Inform. 36 (2017), pp. 24–36. doi:10.1002/minf.201700036.

[29] C. Chih-Chung and L. Chih-Jen, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011), pp. 1–27. doi:10.1145/1961189.1961199.

[30] I.H. Witten and E. Frank, The Weka Workbench, in Data Mining: Practical Machine Learning Tools and Techniques, I.H. Witte, eds., Morgan Kaufman Publishers, San Fransisco, 2005, pp. 363–449.

[31] G. Marcou, D. Hor Vath, V. Solov'Ev, A. Arrault, P. Vayer, and A. Varnek, Interpretability of SAR/QSAR models of any complexity by atomic contributions, Mol. Inform. 31 (2012), pp. 639–642. doi:10.1002/minf.201100136.

[32] G.M. Cramer, R.A. Ford, and R.L. Hall, Estimation of toxic hazard-a decision tree approach, Food Cosmet. Toxicol. 16 (1976), pp. 255–276. doi:10.1016/S0015-6264(76)80522-6.

[33] S. Bhatia, T. Schultz, D. Roberts, J. Shen, L. Kromidas, and A. Marie Api, Comparison of Cramer classification between Toxtree, the OECD QSAR Toolbox and expert judgment, Regul. Toxicol. Pharmacol. 71 (2015), pp. 52–62. doi:10.1016/j.yrtph.2014.11.005.

[34] K. Mansouri, C.M. Grulke, R.S. Judson, and A.J. Williams, OPERA models for predicting physicochemical properties and environmental fate endpoints, J. Cheminform. 10 (2018), pp. 1–19. doi:10.1186/s13321-018-0263-1.

[35] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, Mol. Inform. 29 (2010), pp. 476–488. doi:10.1002/minf.201000061.

[36] D. Fourches, E. Muratov, and A. Tropsha, Trust but verify: On the importance of chemical structure curation in chemoinformatics and qsar modeling research, J. Chem. Inf. Model. 50 (2010), pp. 1189–1204. doi:10.1021/ci100176x.

[37] A. Tropsha, P. Gramatica, and V.K. Gombar, The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003), pp. 69–77. doi:10.1002/qsar.200390007.