	***QMRF identifier (JRC Inventory):*** **To be entered by JRC**
	QMRF Title: **Acute Toxicity (LD50) model (KNN) - v. 1.0.0**
	Printing Date: **October 2022**

1.QSAR identifier

1.1.QSAR identifier (title):

Acute Toxicity (LD50) model (KNN) - v. 1.0.0

1.2.Other related models:

NA

1.3.Software coding the model:

VEGA (https://www.vegahub.eu/)

The VEGA software provides QSAR models to predict tox, ecotox, environ, phys-chem and toxicokinetic properties of chemical substances.

emilio.benfenati@marionegri.it

2.General information

2.1.Date of QMRF:

October 2022

2.2.QMRF author(s) and contact details:

[1] Erika Colombo Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy erika.colombo@marionegri.it https://www.marionegri.it/

[2] Alessandra Roncaglioni Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy alessandra.roncaglioni@marionegri.it https://www.marionegri.it/

2.3.Date of QMRF update(s):

No update

2.4.QMRF update(s):

No update

2.5.Model developer(s) and contact details:

[1] Alessandra Roncaglioni Manganaro Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy alessandra.roncaglioni@marionegri.it https://www.marionegri.it/

[2] Alberto Manganaro Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy alberto.manganaro@marionegri.it https://www.marionegri.it/

2.6.Date of model development and/or publication:

2019

2.7.Reference(s) to main scientific papers and/or software package:

[1] D. Gadaleta et al., 'SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data', Journal of Cheminformatics, vol. 11, no. 1, p. 58, Aug. 2019, doi: 10.1186/s13321-019-0383-2.

[2] K. Mansouri et al., 'CATMoS: Collaborative Acute Toxicity Modeling Suite', Environ Health Perspect, vol. 129, no. 4, p. 47013, Apr. 2021, doi: 10.1289/EHP8495

[3] Benfenati E, Roncaglioni A, Lombardo A, Manganaro A. Integrating QSAR, Read-Across, and Screening Tools: The VEGAHUB Platform as an Example. Advances in Computational Toxicology; Springer; 2019. p. 365-81.

2.8.Availability of information about the model:

The model is non-proprietary and the training set is available.

2.9.Availability of another QMRF for exactly the same model:

Another QMRF is not available.

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Rat.

3.2.Endpoint:

Rat acute oral systemic toxicity tests

3.3.Comment on endpoint:

NA

3.4.Endpoint units:

LD50 (mmol/kg)

3.5.Dependent variable:

logLD50 (mmol/kg)

3.6.Experimental protocol:

OECD, Test No. 420: Acute Oral Toxicity - Fixed Dose Procedure, OECD, Test No. 423: Acute Oral toxicity - Acute Toxic Class Method, OECD, Test No. 425: Acute Oral Toxicity: Up-and-Down Procedure

3.7.Endpoint data quality and variability:

The dataset provided by NICEATM and NCCT has been curated as described below.

LD50 values (mg/kg) were converted to logLD50 (mmol/kg) in order to have a distribution of data more suitable for modelling. The MW used for conversion was calculated as the sum of MWs of the main molecule and of its counterion, if present.

There were data referring to the same chemical (i.e., InChi code) in the “Training Dataset” and “Complete LD50 inventory”. They have been defined as follows in this document:

Duplicates: two or more records sharing the same InChI code in the “Training Dataset” (https://ntp.niehs.nih.gov/iccvam/methods/acutetox/model/trainingset_171130.txt) but having different CAS number and/or SMILES. They are the same chemical but they may differ for the syntax of the SMILES or for the associated counterion. Different records may have different LD50 values and/or different classifications.

Replicates: two or more records sharing the same InChI, CAS, SMILES in the “Complete LD50 inventory” (https://ntp.niehs.nih.gov/iccvam/methods/acutetox/model/trainingset_full_ld50.txt) but having different experimental values. They are results of different experiments on the same chemical.

All the values referring to replicates and duplicates were put together and aggregated. Counterions were removed and not considered during the aggregation. Only “point estimate” values were considered from the “Complete LD50 inventory” while “limit tests” were rejected.

For each compound (single InChI), the experimental value was recalculated as the median of the first quantile and the standard deviation of the distribution of values of replicates/duplicates was calculated. The median of the first quantile was used as continuous endpoint value.

Chemicals having standard deviation ≥ 0.50 log units among experimental logLD50 values were excluded for modelling of continuous values. A total of 89 chemicals were excluded in this way.

The final dataset was composed of 6280 substances, 5029 as training set (TS) and 1251 as validation set (VS).

After the implementation in VEGA, the final dataset was composed of 6280 substances (training set).

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The Acute Toxicity (LD50) model is a Regression model (kNN) based on 6280 substances retrieved from several sources. It is specific for the acute oral systemic toxicity tests in Rats.

4.2.Explicit algorithm:

Regression model (kNN).

This in house program, in Java language, is based on the CDK and VEGA libraries, for building, evaluating and applying k-NN models. At the basis of a k-NN approach there is the idea of similarity, which leads to the selection of a certain (k) number of neighbors for the target compound that are used to provide a prediction. The similarity index developed in the VEGA project is used in istKnn (Floris et al., 2014). This index was developed to provide a global measurement of similarity, i.e. a “generic” chemical similarity as it could be conceived by the human expert. While usual modeling approaches require the selection of a set of descriptors, in istKnn the VEGA similarity index is used without modification.

The software istKNN takes as input a training set of chemical structures expressed as SMILES and their related experimental value, and assesses the predictive power by calculating the predictions on each molecule from the training set itself in a leave-one-out (LOO) approach (the molecule to be predicted is extracted from the training set, and the k-NN model is run on the remaining compounds). The software can provide predictions for qualitative datasets (where the prediction is a classification label) and for quantitative datasets (where the prediction is a continuous value).

Here the software was used to model the point estimate in terms of logLD50 (mmol/kg).

The predicted value is calculated with the following algorithm:

1. The first k molecules with the closest similarity to the target compound are extracted.

2. Molecules with a similarity index lower than a given threshold are excluded.

3. If no molecules are left, no prediction is provided (missing value).

4. If only one molecule is left, it is used as prediction only if its similarity value is equal to or higher than a given threshold, otherwise no prediction is provided (missing value).

5. In all other cases, the prediction is calculated as a weighted consensus of the experimental values among the remaining molecules. The similarity value of each molecule is used as its weight. Optionally, the weights (similarity values) can be raised to the power of a given value E, called the enhance factor, as for integer larger than 1 the result is to enhance the role of molecules with higher similarity values in the prediction. The range of variation in the experimental values of the similar compounds used for the prediction can be also analyzed and if too variation is observed the prediction is not provided.

A k-NN model is defined just by its training set and by the settings chosen for the parameters:

- Number of neighbours (K) – 2 to 5
- Similarity threshold: 0.7 to 0.9, step 0.05
- Similarity threshold for single molecules: 0.85 to 0.9, step 0.05
- Enhance factor for weights: 1 to 3 and
- Experimental range: 1 to 2, step 0.5

Building a proper model involves selecting the most suitable parameters. There is no single best approach, as the changing of parameters affects the model's accuracy and number of non-predicted compounds. Usually, higher thresholds for similarities increases the accuracy of the model, but also leads to a larger number of non-predicted compounds, and vice versa.

Performance of derived models were evaluated both in internal and external validation. The output of the batch process (360 models) was analyzed, and 5 models were selected on the basis of the R2 obtained in the LOO procedure on the TS and different amount of compounds left as not predicted.

Then, the best model (nr. 135) in terms of balance between R2 and coverage was selected and evaluated also on the VS.

Finally, once the favorite setting was selected (K = 3; Min Similarity = 0.8; Min Similarity for single molecule = 0.85; Enhance factor =3; Exper. Range = 2), the model was re-trained on the entire dataset (TS+VS)

4.3. Descriptors in the model:

NA

4.4. Descriptor selection:

NA

4.5. Algorithm and descriptor generation:

NA

4.6. Software name and version for descriptor generation:

istKnn v.0.9.3 software was used (https://chm.kode-solutions.net/)d

4.7.Chemicals/Descriptors ratio:

NA

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The Applicability Domain (AD) is assessed using the original algorithm implemented within VEGA. An overall AD index is calculated, based on a number of parameters, which relate to the results obtained on similar chemicals within the training and test sets.

Indices are calculated on the first k = the number of k compounds used in the KNN model for the prediction most similar molecules, each having Sk similarity value with the target molecule.

5.2.Method used to assess the applicability domain:

The Applicability domain chemical similarity is measured with the algorithm developed for VEGA. Full details are in the VEGA website (www.vegahub.eu), including the open access paper describing it [6]. The VEGA AD also evaluates the correctness of the prediction on similar compounds (accuracy), the consistency between the predicted value for the target compound and the experimental values of the similar compounds, the range of the descriptors, and the presence of unusual fragments, using atom centred fragments.

5.3.Software name and version for applicability domain assessment:

VEGA (www.vegahub.eu)

5.4.Limits of applicability:

The model is not applicable to inorganic chemicals and substances containing unusual elements (i.e., different from C, O, N, S, P, Cl, Br, F, I). Salts can be predicted only if converted to the neutralized form.

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:

Yes

6.2.Available information for the training set:

CAS RN: Yes

Chemical Name: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: No

6.3.Data for each descriptor variable for the training set:

NA

6.4.Data for the dependent variable for the training set:

NA

6.5.Other information about the training set:

Training set: n = 6280

6.6.Pre-processing of data before modelling:

The dataset was analyzed by mean of a principal component analysis (PCA) based on CDK descriptors. Chemicals being outlier based on values on the first two principal components were removed from the dataset (8 compounds, highlighted with red circles in Figure 1).

Inorganic chemicals and chemicals including unusual chemical elements (i.e., B and Se) were removed (23 compounds). Silicates were kept.

6.7.Statistics for goodness-of-fit:

For the original model:

Model No.	TS	R2 (predictions)	RMSE (predictions)	Valid predictions	Coverage (% of valid predictions on VS)
135	5029	0.562	0.592	4363	86.8%

LOO statistics on TS+VS

Model No.	R2 (predictions)	RMSE (predictions)	Valid predictions	Coverage (% of valid predictions on VS)
135	0.558	0.586	1096	87.6%

6.8.Robustness - Statistics obtained by leave-one-out cross-validation:

After the implementation in VEGA, all data (TS +VS) are implemented ad training set:

LOO statistics: not predicted 747, RMSE 0.57, R2 0.58, coverage 88,1%

6.9.Robustness - Statistics obtained by leave-many-out cross-validation:

NA

6.10.Robustness - Statistics obtained by Y-scrambling:

NA

6.11.Robustness - Statistics obtained by bootstrap:

NA

6.12.Robustness - Statistics obtained by other methods:

NA

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

No, the original model has been externally validated but it is not available in VEGA

7.2.Available information for the external validation set:

NA

7.3.Data for each descriptor variable for the external validation set:

NA

7.4.Data for the dependent variable for the external validation set:

NA

7.5.Other information about the external validation set:

NA

7.6.Experimental design of test set:

NA

7.7.Predictivity - Statistics obtained by external validation:

NA

7.8.Predictivity - Assessment of the external validation set:

NA

7.9.Comments on the external validation of the model:

NA

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

NA

8.2.A priori or a posteriori mechanistic interpretation:

NA

8.3.Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1.Comments:

NA

9.2.Bibliography:

[1] OECD, Test No. 420: Acute Oral Toxicity - Fixed Dose Procedure. Paris: Organisation for Economic Co-operation and Development, 2002. Accessed: Nov. 02, 2022. [Online]. Available: https://www.oecd-ilibrary.org/environment/test-no-420-acute-oral-toxicity-fixed-dose-procedure_9789264070943-en

[2] OECD, Test No. 423: Acute Oral toxicity - Acute Toxic Class Method. Paris: Organisation for Economic Co-operation and Development, 2002. Accessed: Nov. 02, 2022. [Online]. Available: https://www.oecd-ilibrary.org/environment/test-no-423-acute-oral-toxicity-acute-toxic-class-method_9789264071001-en

[3] OECD, Test No. 425: Acute Oral Toxicity: Up-and-Down Procedure. Paris: Organisation for Economic Co-operation and Development, 2022. Accessed: Nov. 02, 2022. [Online]. Available: https://www.oecd-ilibrary.org/environment/test-no-425-acute-oral-toxicity-up-and-down-procedure_9789264071049-en

[4] “NICEATM: Alternative Methods.” https://ntp.niehs.nih.gov/whatwestudy/niceatm/index.html (accessed Nov. 02, 2022).

[5] O. US EPA, “About the National Center for Computational Toxicology (NCCT).” https://19january2017snapshot.epa.gov/aboutepa/about-national-center-computational-toxicology-ncct (accessed Nov. 02, 2022).

[6] Floris, M., Manganaro, A., Nicolotti, O. et al. A generalizable definition of chemical similarity for read-across. J Cheminform 6, 39 (2014). https://doi.org/10.1186/s13321-014-0039-1

9.3.Supporting information:

Training set(s)Test set(s)Supporting information:

All available datasets are present in the model inside the VEGA software.

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

To be entered by JRC

10.2.Publication date:

To be entered by JRC

10.3.Keywords:

To be entered by JRC

10.4.Comments:

To be entered by JRC