	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Daphnia Magna Acute (EC50) Toxicity model (IRFMN/Combase) (version 1.0.1)**
	Printing Date: **June 14 2022**

1.QSAR identifier

1.1. QSAR identifier (title):

Daphnia Magna Acute (EC50) Toxicity model (IRFMN/Combase) (version 1.0.01

1.2. Other related models:

NA

1.2. 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:

June 14, 2022

2.2. QMRF author(s) and contact details:

[1]Andrey Toropov Istituto di Ricerche Farmacologiche Mario Negri IRCCSandrey.toropov@marionegri.it https://www.marionegri.it/

[2]Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri IRCCSemilio.benfenati@marionegri.it https://www.marionegri.it/

[3]Giovanna J. Lavado Istituto di Ricerche Farmacologiche Mario Negri IRCCSgiovanna.lavado@marionegr.it https://www.marionegri.it/

[4]Domenico Gadaleta Istituto di Ricerche Farmacologiche Mario Negri IRCCSdomenico.gadaleta@marionegri.it https://www.marionegri.it/

2.3. Date of QMRF update(s):

NA

2.4. QMRF update(s):

NA

2.5. Model developer(s) and contact details:

[1] Andrey Toropov Istituto di Ricerche Farmacologiche Mario Negri IRCCSandrey.toropov@marionegri.it

[2]Alla Toropova Istituto di Ricerche Farmacologiche Mario Negri IRCCSalla.toropova@marionegri.it

2.6. Date of model development and/or publication:

September 2018

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

[1] Kabiruddin Khan, Pathan Mohsin Khan, Giovanna Lavado, Cecile Valsecchi, Julia Pasqualini, Diego Baderna, Marco Marzo, Anna Lombardo, Kunal Roy, Emilio Benfenati Corrigendum to “QSAR modeling of Daphnia magna and fish toxicities of biocides using 2D descriptors” [Chemosphere 229 (2019) 8–17] Chemosphere, Volume 237, December 2019, Pages 124397

[2] Benfenati E, Manganaro A, Gini G

VEGA-QSAR: AI inside a platform for predictive toxicology

Proceedings of the workshop "Popularize Artificial Intelligence 2013", December 5th 2013, Turin, Italy

Published on CEUR Workshop Proceedings Vol-1107

[3] Toropov AA, Toropova AP, Roncaglioni A, Benfenati E. Prediction of Biochemical Endpoints by the CORAL Software: Prejudices, Paradoxes, and Results. Methods Mol Biol. 2018; 1800:573-583. doi: 10.1007/978-1-4939-7899-1_27.

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:

The species name is Daphnia magna

3.2. Endpoint:

ECOTOX 6.1.3. Short-term toxicity to aquatic invertebrates. invertebrates (Daphnia immobilization). OECD TG 202 Daphnia sp., Acute Immobilisation Test []

3.3. Comment on endpoint:

48 hour Daphnia magna EC50 according to OECD 202

3.4. Endpoint units:

EC50 expressed in mmol/L

3.5. Dependent variable:

Log EC50

3.6. Experimental protocol:

OCDE Test No. 202: It measures the percentage of immobilized daphnids after a 48h exposure to a substance

3.7. Endpoint data quality and variability:

To build the datasets of biocides with acute toxicity data towards Daphnia and fish, we downloaded the list of biocides from the ECHA website. We manually retrieved the structures and we compared them with the chemical names, and the Chemical Abstracts Service (CAS) numbers using online databases (ChemCell (https://github.com/cdd/chemcell); MarvinView (Marvin 17.28.0, 2017, ChemAxon (http://www.chemaxon.com)); ChemIDplus Advanced (https://chem.nlm.nih.gov/chemidplus/); PubChem (https://pubchem.ncbi.nlm.nih.gov/); ChemSpider (http://www.chemspider.com/)). For modeling purposes, we removed the compounds with a chemical structure not clearly identified, the inorganic compounds, the metal complexes, the salts containing organic polyatomic counterions, the mixtures and the substances of Unknown or Variable composition (UVCB). In addition, we neutralized the structure of the salts. The final list was of 143 compounds.

We searched for the toxicity data on several public sources: the OECD QSAR toolbox v. 4.2 (www.qsartoolbox.org), the Pesticide Properties Database (PPDB) database (https://sitem.herts.ac.uk/aeru/ppdb/), the Office of Pesticide Programs (OPP) Pesticides Ecotoxicity Database (http://www.ipmcenters.org/ecotox/)), the European Food Safety Authority (EFSA) (http://www.efsa.europa.eu/) database and the ECOTOX (https://cfpub.epa.gov/ecotox/) database.

In case of multiple data for the same compound, we used the threshold established by the European Commission (SANCO/10597/2003) (European, 2012) as the ratio between maximum value and the minimum experimental value (x/y) (compounds with difference of >3 were removed).

Experimental data were carefully screened for specific endpoints and identical exposure time in order to get reliable predictions from standardized data. For the ease of interpretation, the half maximal effective

concentration (EC50) values were converted in to a molar unit (EC50 in molL−1) followed by transformation into a negative logarithmic scale, i.e., pEC50 as customary in ecotoxicological QSAR analysis.

For Daphnia 133 compounds for immobilization at 48 h (EC50–48 h) were collected by strictly following OECD guideline 203. The drawn structures were cleaned and explicit hydrogens were added using MarvinSketch (version 5.11, calculation module developed by ChemAxon, http://www.chemaxon.com/products/marvin/marvinsketch/2013) and saved as MDL.mol, a recommended format for PaDEL-Descriptor (Yap, 2011) and Dragon (Todeschini et al., 2004; Mauri et al., 2006) software.

4.Defining the algorithm - OECD Principle 2

4.1. Type of model:

Regression

4.2. Explicit algorithm:

Endpoint = C0 + C1 DCW (T, N)

CORAL (http://www.insilico.eu/coral) was used to develop regression QSAR model from SMILES-based optimal descriptors. A CORAL mathematical model describes the relationship between an endpoint (dependent variable) and relevant SMILES attributes (independent variable), as shown in the equation:

Endpoint = C0 + C1 DCW (T, N)

where C0 and C1 are the intercept and slope for the relationship, and DCW (T, N) is the combination of SMILES-based attributes, each associated with a correlation weight (CW).

CWs are determined with the Monte Carlo algorithm in an iterative procedure that aims to optimize a target function (TF). The TF is calculated as shown in the equation:

TF = R + R' – |R-R'| 0.01

where R and R' are the correlation coefficients between DCW(T,N) and the endpoints for TS and ITS.

This procedure is defined as a balance of correlations (BC).

The TF is a function of the CWs and is optimized by iteratively modifying them. In the first part of the optimization, CWs are incremented by a value D start. This increment is repeated as long as there was a corresponding improvement of the TF. When no further improvement is observed, the D start value is modified to D start,1 = -0.5 (D start) for subsequent iterations. D start is iteratively modified each time that an increment of CWs fails to correspond to an increment of TF, until | D start | is lower than a threshold value (D precession)

4.3. Descriptors in the model:

No descriptors were used but SMILES-based attributes

These SMILES-based attributes can be described as in the following Equation:

DCW (T*,N*) = CW(Sk) + CW (SSk)

where Sk and SSk are SMILES attributes defined by a sequence of atoms and bonds present in the SMILES string. Sk represents single elements, and SSk two elements combined.

Attributes with a positive CW are considered promoters of an increase of the endpoint value, while attributes with a negative correlation weights are considered promoters of a decrease

4.4. Descriptor selection:

N is the number of epochs of Monte Carlo for optimization of the TF, and T is a threshold used to classify SMILES attributes as rare or not rare. An attribute is defined as rare if it is found in the SMILES of the CS less than T times. Rare SMILES attribute values were set to zero so they were not involved in the modeling. T and N are set to optimize the statistical performance for the CS. For this model, parameters were set as follows:

T = 1; N = 35; Dstart= 0.5; Dprecession = 0.1

4.5. Algorithm and descriptor generation:

Simplified Molecular Input Line Entry System (SMILES) notation describes the structure of a chemical using linear strings in place of the classical bi- or tri- dimensional representation. CORAL breaks the SMILES

strings of the TS compounds into small components (SMILES-based attributes). Each SMILES-based attributes check the presence of particular characters (or combinations of characters) within the SMILES

4.6. Software name and version for descriptor generation:

CORAL-2017

4.7. Chemicals/Descriptors ratio:

133 chemicals/95 SMILES attributes (NB, there are not descriptors)

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.

ADI is defined in this way for this QSAR model´s predictions:

If 1 ≥ AD index > 0.85, the predicted substance is regarded in the Applicability Domain of the model. It corresponds to “good reliability” of prediction.

If 0.85 ≥ AD index ≥ 0.7, the predicted substance could be out of the Applicability Domain of the model. It corresponds to “moderate reliability” of prediction.

If AD index < 0.7, the predicted substance is regarded out of the Applicability Domain of the model. It corresponds to “low reliability” of prediction.

5.2. Method used to assess the applicability domain:

The Applicability domain and chemical similarity are measured with the algorithm developed for VEGA. Full details are in the VEGA website (www.vegahub.eu), including the open access paper describing it [5]. 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.

These indices are defined in this way for this QSAR model:

Similar molecules with known experimental value:

This index takes into account how similar are the first two most similar compounds found. Values near 1 mean that the predicted compound is well represented in the dataset used to build the model, otherwise the prediction could be an extrapolation. Defined intervals are:

If 1 ≥ index > 0.85, strongly similar compounds with known experimental value in the training set have been found

If 0.85 ≥ index > 0.7, only moderately similar compounds with known experimental value in the training set have been found

If index ≤ 0.7, no similar compounds with known experimental value in the training set have been found

Accuracy (average error) of prediction for similar molecules:

This index takes into account the classification accuracy in prediction for the two most similar compounds found. Values near 1 mean that the predicted compounds fall in an area of the model's space where the model gives reliable predictions (no misclassifications), otherwise the lower is the value, the worse the model behaves. Defined intervals are:

If index < 0.8, accuracy of prediction for similar molecules found in the training set is good

If 1.2 > index ≥ 0.8, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≥ 1.2, accuracy of prediction for similar molecules found in the training set is not adequate

Concordance for similar molecules:

This index takes into account the difference between the predicted value and the experimental values of the two most similar compounds. Values near 0 mean that the prediction made disagrees with the values found in the model's space, thus the prediction could be unreliable. Defined intervals are:

If index < 0.8, molecules found in the training set have experimental values that agree with the target compound predicted value

If 1.2 > index ≥ 0.8, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index ≥ 1.2, similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value

Maximum error of prediction between similar molecules:

This index takes into account the maximum error in prediction between the two most similar compounds. Values near 0 means that the predicted compounds fall in an area of the model's space where the model gives reliable predictions without any outlier value. Defined intervals are:

If index < 0.8, the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability

If 1.2 > index ≥ 0.8, the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

If index ≥ 1.2, the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability

Model descriptors range check:

This index checks if the descriptors calculated for the predicted compound are inside the range of descriptors of the training and test set. The index has value 1 if all descriptors are inside the range, 0 if at least one descriptor is out of the range. Defined intervals are:

If index = True, descriptors for this compound have values inside the descriptor range of the compounds of the training set

If index= False, the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

Atom Centered Fragments similarity check:

This index takes into account the presence of one or more fragments that aren't found in the training set, or that are rare fragments. First order atom centered fragments from all molecules in the training set are calculated, then compared with the first order atom centered fragments from the predicted compound; then the index is calculated as following: a first index RARE takes into account rare fragments (those who occur

less than three times in the training set), having value of 1 if no such fragments are found, 0.85 if up to 2 fragments are found, 0.7 if more than 2 fragments are found; a second index NOTFOUND takes into account not found fragments, having value of 1 if no such fragments are found, 0.6 if a fragments is found, 0.4 if more than 1 fragment is found. Then, the final index is given as the product RARE * NOTFOUND. Defined intervals are:

If index = 1, all atom centered fragment of the compound have been found in the compounds of the training set

If 1 > index ≥ 0.7, some atom centered fragment of the compound have not been found in the compounds of the training set or are rare fragments

If index < 0.7, a prominent number of atoms centered fragments of the compound have not been found in the compounds of the training set or are rare fragments

5.3. Software name and version for applicability domain assessment:

VEGA (www.vegahub.eu)

5.4. Limits of applicability:

VEGA provides a quantitative value for the prediction of each substance. This helps the user to identify potential critical aspects, which are indicated. Similar compounds are shown.

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:

No

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

All

6.5. Other information about the training set:

NA

6.6. Pre-processing of data before modelling:

For modelling purposes, these compounds have been removed: with a chemical structure not clearly identified, the inorganics compounds, the metal complexes, the salts containing organic polyatomic counterions, the mixtures and the substances of Unknown or Variable composition (UVCB). In addition, the structure of the salts has been neutralized and duplicates have been removed. The final list was of 133 compounds

6.7. Statistics for goodness-of-fit:

Statistic in the development phase of model:

Training set (TS): R^2 0.7522, RMSE 1.16

Invisible training set (ITS): R^2 0.7578, RMSE 1.26

Calibration set (CS): R^2 0.7005, RMSE 1.35

Statistics of the implemented version of the model for the training set:

R^2 0.72, RMSE 1.24, n of compounds 99

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

NA

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:

Training set (TS): Q^2 0.7163

Invisible training set (ITS): Q^2 0.7333

Calibration set (CS): Q^2 0.6651

7.External validation - OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Available information for the external validation set:

CAS RN: Yes

Chemical Name: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: No

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

No

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

All

7.5. Other information about the external validation set:

NA

7.6. Experimental design of test set:

The initial dataset was randomly divided into a training set (TS) of 33compounds, an invisible training set (ITS) of 34 compounds, a calibration set (CS) of 32 compounds, and a validation set (VS) of 34compounds. The CORAL software was unable to process the SMILES for one chemical which was filtered out. The TS was used for regression model’s derivation. The ITS was used as “inspector” during model derivation, to confirm (or reject) predictivity of the model for substances which were not involved directly to the optimization process. The CS detected the beginning of overfitting by verifying the increase of the correlation between descriptors and endpoint during the optimization process, until improvements were no longer observed. The VS is the final estimator of the predictive potential of the model

7.7. Predictivity - Statistics obtained by external validation:

Statistic in the development phase of model:

Validation set (VS): R2 0.7506

Statistics of the implemented version of the model for the test set:

R^2 0.60, RMSE 1.08, n of compounds 34

Test in AD: R2, 0.42 RMSE 1.18

Test set “could be out of AD”: R2, 0.75 RMSE 1.05

Test set out of AD: R2, 0.55 RMSE 1.01

Or

Test set in AD: n = 0

Test set Could be out of AD: n = 4, 0.42 RMSE 1.18

Test set out of AD: n = 30, R2, 0.62 RMSE 1.03

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:

A posteriori

8.3. Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1. Comments:

NA

9.2. Bibliography:

[1] Toropov, A.A., Toropova, A.P., Benfenati, E. et al. Use of quasi-SMILES to model biological activity of “micelle–polymer” samples. Struct Chem 29, 1213–1223 (2018). https://doi.org/10.1007/s11224-018-1115-3

[2] Alla P. Toropova, Andrey A. Toropov, Emilio Benfenati, Sara Castiglioni, Renzo Bagnati, Alice Passoni, Ettore Zuccato, Roberto Fanelli. Quasi-SMILES as a tool to predict removal rates of pharmaceuticals and dyes in sewage, Process Safety and Environmental Protection, Volume 118, 2018, Pages 227-233, https://doi.org/10.1016/j.psep.2018.07.003.

[3] Alla P. Toropova, Andrey A. Toropov, Emilio Benfenati, Danuta Leszczynska, Jerzy Leszczynski. Prediction of antimicrobial activity of large pool of peptides using quasi-SMILES, Biosystems,Volumes 169–170,2018,Pages 5-12, https://doi.org/10.1016/j.biosystems.2018.05.003.

[4] Toropova AP, Toropov AA, Marzo M, Escher SE, Dorne JL, Georgiadis N, Benfenati E. The application of new HARD-descriptor available from the CORAL software to building up NOAEL models. Food Chem Toxicol. 2018 Feb; 112:544-550. doi: 10.1016/j.fct.2017.03.060

[5] Toropov AA, Toropova AP, Roncaglioni A, Benfenati E. Prediction of Biochemical Endpoints by the CORAL Software: Prejudices, Paradoxes, and Results. Methods Mol Biol. 2018;1800:573-583. doi: 10.1007/978-1-4939-7899-1_27

[6] OECD, Test No. 202: Daphnia Sp. Acute Immobilisation Test (Paris: Organisation for Economic Co-operation and Development, 2004), https://www.oecd-ilibrary.org/environment/test-no-202-daphnia-sp-acute-immobilisation-test_9789264069947-en.

[7] 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