

	QMRf identifier (JRC Inventory): To be entered by JRC
	QMRf Title: VEGA Daphnia Acute (EC50) toxicity model (IRFMN) v 1.0.1
	Printing Date: Apr 30, 2019

1. QSAR identifier

1.1. QSAR identifier (title):

VEGA Daphnia Acute (EC50) toxicity model (IRFMN) v1.0.1

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:

11/04/2019

2.2. QMRf author(s) and contact details:

Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri - IRCCS Via Mario Negri 2, 20156 Milano, Italy emilio.benfenati@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] Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri IRCCS emilio.benfenati@marionegri.it

[2] Cosimo Toma Istituto di Ricerche Farmacologiche Mario Negri IRCCS cosimo.toma@marionegri.it

[3] Claudia Ileana Cappelli Institut national de l'environnement industriel et des risques Claudia.CAPPELLI@ineris.fr

[4] Alberto Manganaro Istituto di Ricerche Farmacologiche Mario Negri IRCCS & KODE s.r.l alberto.manganaro@marionegri.it

2.6. Date of model development and/or publication:

2017

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

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

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 invertebrate, OECD TG 202 *Daphnia* sp., Acute Immobilisation Test (2004 & 1984) [2]

3.3. Comment on endpoint:

Number and percentage of daphnids that were immobilised (including abnormal behaviour) in the controls and in each treatment group.

3.4. Endpoint units:

The model provides a quantitative prediction for *Daphnia magna* EC50 (48 hour), given in $-\log(\text{mol/l})$ and its conversion to mg/L

3.5. Dependent variable:

$-\log_{10}(\text{LC}_{50})$ in $-\log(\text{mol/l})$

3.6. Experimental protocol:

OECD 202 Test. It measures the immobilized daphnids after 48h of exposure to a substance.

3.7. Endpoint data quality and variability:

445 experimental data retrieved from the Japanese Ministry of Environment (http://www.env.go.jp/en/chemi/sesaku/aquatic_Mar_2016.pdf) and selected according to the OECD TG 202 requirements. For other information see [6.6]. The dataset was split randomly in Training set: $n = 312$ Test set: $n = 133$

All tests were performed according to OECD 202 and GLP but some tests performed before 2002 employed dispersants.

Commented [h1]: The whole dataset was hence 45 substances. In Toma et al 2021 the total number of substances was 428. Please explain whether Toma et al 2021 is relevant – and if so add this ref. to the list of ref.s and explain the small difference in number of data used.

Commented [e2R1]: Toma et al 2021 is not the correct reference for this model

4. Defining the algorithm - OECD Principle 2

4.1. Type of model:

Tree Ensemble Random Forest using 12 molecular descriptors.

4.2. Explicit algorithm:

Tree Ensemble Random Forest

Tree ensemble builds a series of regression trees with different rows and different variables (according to certain parameters) and then the results are aggregated as an ensemble of models. The parameters for the variables of each tree and the number of compounds are chosen evaluating the performance of several models (Hyperparameter tuning Research) using as metric R2 of a Bootstrap (100 iterations) cross-validation on training set

4.3. Descriptors in the model:

[1]S.106 R-SH

[2]Me mean atomic Sanderson electronegativity (scaled on Carbon atom)

[3]MATS5e Moran autocorrelation of lag 5 weighted by Sanderson electronegativity

[4]MATS4p Moran autocorrelation of lag 4 weighted by polarizability

[5]GATS1m Geary autocorrelation of lag 1 weighted by mass

[6]EEig15bo eigenvalue n. 15 from edge adjacency mat. weighted by bond order

[7]EEig8dm eigenvalue n. 8 from edge adjacency mat. weighted by dipole moment

[8]B2.C..O. Presence/absence of C - O at topological distance 2

[9]B10.C..N. Presence/absence of C - N at topological distance 10

[10]F4.Cl..Cl. Frequency of Cl - Cl at topological distance 4

[11]F10.O..O. Frequency of O - O at topological distance 10

[12]ALogP Ghose-Crippen octanol-water partition coeff. (logP)

4.4. Descriptor selection:

Descriptors have been filtered according to the following procedure:

Descriptors with constant values ($\text{var}(X) = 0$) or which correlate over 0.95 (Pearson) with at least one another descriptor have been removed. A genetic algorithm (R package *gaselect*) have been used to select the best pool of descriptors.

4.5. Algorithm and descriptor generation:

CDK and VEGA

4.6. Software name and version for descriptor generation:

CDK

The Chemistry Development Kit

The CDK developers

<https://github.com/cdk>

VEGA

Virtual models for property Evaluation of chemicals within a Global Architecture

Istituto di Ricerche Farmacologiche Mario Negri Milano, Laboratory of Environmental Chemistry and Toxicology <https://www.vegahub.eu/>

4.7. Chemicals/Descriptors ratio:

312 (training)/12 (descriptors) = 26

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 and is defined in this way for this QSAR model's predictions :

If $1 \geq \text{AD index} > 0.85$, the predicted substance is regarded in the Applicability Domain of the model. It corresponds to "good reliability" of prediction.

If $0.85 \geq \text{AD index} > 0.7$, the predicted substance could be out of the Applicability Domain of the model. It corresponds to "moderate reliability" of prediction.

If $\text{AD index} \leq 0.7$, the predicted substance is regarded out of the Applicability Domain of the model. It corresponds to "low reliability" of prediction.

Indices are calculated on the first $k = 2$ most similar molecules, each having S_k similarity value with the target molecule.

Similarity index (*IdxSimilarity*) is calculated as:

$$\frac{\sum_k S_k}{k} \times (1 - \text{Diam}^2)$$

where *Diam* is the difference in similarity values between the most similar molecule and the k -th molecule.

Accuracy index (*IdxAccuracy*) is calculated as:

$$\frac{\sum_c^k |exp_c - pred_c|}{k}$$

where exp_c is the experimental value of the c -th molecule in the training set and $pred_c$ is the c -th molecule predicted value by the model.

Concordance index (*IdxConcordance*) is calculated as:

$$\frac{\sum_c^k |exp_c - pred_{target}|}{k}$$

where exp_c is the experimental value of the c -th molecule in the training set and $pred_{target}$ is the predicted value for the input target molecule.

Max Error index (*IdxMaxError*) is calculated as:

$$\max(|exp_c - pred_c|)$$

where exp_c is the experimental value of the c -th molecule in the training set and $pred_{target}$ is the predicted value for the input target molecule, evaluated over the k molecules.

ACF contribution (*IdxACF*) index is calculated as

$$ACF = rare \times missing$$

where: *rare* is calculated on the number of fragments found in the molecule and found in the training set in less than 3 occurrences as following: if the number is 0, *rare* is set to 1.0; if the number is 1, *rare* is set to 0.6; otherwise *rare* is set to 0.4

missing is calculated on the number of fragments found in the molecule and never found in the training set as following: if the number is 0, *missing* is set to 1.0; if the number is 1, *missing* is set to 0.6; otherwise *missing* is set to 0.4

Descriptors Range (*IdxDescRange*) index is calculated as 1.0 if all molecular descriptors used in the prediction fall within the range of descriptors used in the whole training set, 0.0 otherwise.

AD final index is calculated as following:

$$ADI = IdxSimilarity \times IdxACF \times IdxDescRange$$

The initial ADI index is the used together with the other sub-indices to calculate the final ADI, on the basis of the assessment class in which each sub-index falls:

IdxAccuracy \geq	IdxConcordance \geq	IdxMaxError \geq	InitialADI \geq	ADI
1.2	1.2	1.2	0.85	1.0
0.8	0.8	0.8	0.7	0.85
All other cases				0.7

5.2. Method used to assess the applicability domain:

The Applicability Domain and 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 [1]. The 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 $1 \geq \text{index} > 0.85$, strongly similar compounds with known experimental value in the training set have been found

If $0.85 \geq \text{index} > 0.7$, only moderately similar compounds with known experimental value in the training set have been found

If $\text{index} \leq 0.7$, 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 $\text{index} < 0.8$, accuracy of prediction for similar molecules found in the training set is good

If $1.2 > \text{index} \geq 0.8$, accuracy of prediction for similar molecules found in the training set is not optimal

If $\text{index} \geq 1.2$, 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 $\text{index} < 0.8$, molecules found in the training set have experimental values that agree with the target compound predicted value

If $1.2 > \text{index} \geq 0.8$, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If $\text{index} \geq 1.2$, similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value

Maximum error of prediction among similar molecules:

This index takes into account the maximum error in prediction among 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 $\text{index} < 0.8$, the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability

If $1.2 > \text{index} \geq 0.8$, 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

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 $1 > \text{index} \geq 0.7$, 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 atom centered fragments of the compound have not been found in the compounds of the training set or are rare fragments

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:

Index = TRUE, descriptors for this compound have values inside the descriptor range of the compounds of the training set

Index = FALSE, descriptors for this compound have values outside the descriptor range of the compounds of the training set

5.3. Software name and version for applicability domain assessment:

VEGA (www.vegahub.eu)

5.4. Limits of applicability:

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: Yes

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:

No

6.5. Other information about the training set:

NA

6.6. Pre-processing of data before modelling:

We generated the SMILES structures from the chemical name and CAS RN for each substance using ChemCell and Marvin View.

We manually checked among several websites and public database (ChemIDplus Advanced, PubChem, ChemSpider, DSSTox, ...) the correspondence and correctness among the obtained structures, chemical name and CAS RN. We also added the structures not automatically generated.

Then, we pruned the initial dataset as described below. We excluded from the initial dataset metal complexes, inorganics, mixtures of structural isomers, ambiguous structures, non-ionic surfactant mixtures, complex disconnected structures (e.g. polymers), chemicals whose correspondence name-CAS was not found, UVCB. We excluded salts, keeping the acid form of the compounds only.

We selected continuous experimental values and we excluded those reported as a range, as greater/less than a certain threshold, or as approximate values. We kept toxicity values deriving from experimental conditions of the assays as they are defined in the OECD. We also eliminated pH adjusted toxicity values. We calculated the molecular weight from each chemical structure to change the experimental toxicity value from mg/l to mmol/l.

We checked the multiple values: the range between the maximum and the minimum values has to be less or equal to one log unit when the experimental conditions and the reliability of the studies are the same (as reported the ECHA guidance R.10 for the ecotoxicological continuous endpoints). If possible, we found the outlier, otherwise we eliminated the data. We also checked if the experimental toxicity values were higher than the water solubility values. If it was so, we removed the chemical.

6.7. Statistics for goodness-of-fit:

Training set (312 chemicals): $R^2=0.68$, $RMSE=0.62$

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:

NA

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: Yes

Commented [h3]: $R^2=0.94$ $RMSE=0.64$ in Toma et al 2021. Please explain why these results are different

Commented [e4R3]: See above

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:

Data were randomly split in training and test set with the ratio of 80:20. In order to obtain a uniform distribution of the endpoint values between the two subsets it was applied an activity and descriptors sampling method.

7.7. Predictivity - Statistics obtained by external validation:

Calibration Test: n = 133 R² = 0.60, RMSE = 0.70

Test set in AD: n = 44; R² = 0.66; RMSE = 0.60

Test set could be out of AD: n = 51; R² = 0.51; RMSE = 0.66

Test set out of AD: n = 38; R² = 0.57; RMSE = 0.85

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:

The mechanistic approach of the present model is supported by the use of the specific atoms, bonds, and molecular fragments involved in the model descriptors

8.2. A priori or a posteriori mechanistic interpretation:

8.3. Other information about the mechanistic interpretation:

NA

9. Miscellaneous information

9.1. Comments:

NA

9.2. Bibliography:

[1] Floris M, Manganaro A, Nicolotti O, Medda R, Mangiatordi GF, Benfenati E. A generalizable definition of chemical similarity for read-across. J Cheminform. 2014 Oct 18;6(1):39. doi: 10.1186/s13321-014-0039-1. PMID: 25383097; PMCID: PMC4212147.

[2] OECD TG 202 (1984 & 2004) Daphnia sp., Acute Immobilisation Test. Organisation for Economic Co-operation and Development: Paris

9.3. Supporting information:

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

Commented [h5]: R² = 0.56 & 0.69 RMSE = 1.31 & 1.09 for not taking ADI into account and requiring ADI > 0.85, respectively in Toma et al 2021. Please explain why these values differ.

Commented [e6R5]: See previous comment

All available dataset 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