	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: *VEGA Daphnia Acute (EC50) toxicity model (IRFMN)*
	Printing Date: **30-apr-2019**

1.QSAR identifier

1.1.QSAR identifier (title):

VEGA Daphnia Acute (EC50) toxicity model (IRFMN)

1.2.Other related models:

1.3.Software coding the model:

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/

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

emilio.benfenati@marionegri.it emilio.benfenati@marionegri.it

http://www.marionegri.it/en_US/home/research_en/dipartimenti_en/environmental_health_sciences/environmental_chemistry_and_toxicology

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1]Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri IRCCS

emilio.benfenati@marionegri.it emilio.benfenati@marionegri.it

http://www.marionegri.it/en_US/home/research_en/dipartimenti_en/environmental_health_sciences/environmental_chemistry_and_toxicology

[2]Cosimo Toma Istituto di Ricerche Farmacologiche Mario Negri IRCCS

cosimo.toma@marionegri.it cosimo.toma@marionegri.it

http://www.marionegri.it/en_US/home/research_en/dipartimenti_en/environmental_health_sciences/environmental_chemistry_and_toxicology

[3]Claudia Ileana Cappelli Institut national de l'environnement industriel et des risques

Claudia.CAPPELLI@ineris.fr Claudia.CAPPELLI@ineris.fr https://www.ineris.fr/fr

[4]Alberto Manganaro Istituto di Ricerche Farmacologiche Mario Negri IRCCS & KODE s.r.l

alberto.manganaro@marionegri.it alberto.manganaro@marionegri.it

http://www.marionegri.it/en_US/home/research_en/dipartimenti_en/environmental_health_sciences/environmental_chemistry_and_toxicology

2.6.Date of model development and/or publication:

09/2017

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

Benfenati, Emilio, Alberto Manganaro, and Giuseppina C. Gini. "VEGA-QSAR: AI Inside a Platform for Predictive Toxicology."

2.8.Availability of information about the model:

Freely available through the VEGA web site together with a file describing the models. The training, test and validation sets are also available (see 9.3)

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

Other QMRF for this model are not available

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Daphnia magna

3.2.Endpoint:

ECOTOX 6.1.3. Short-term toxicity to aquatic inverterbrates

3.3.Comment on endpoint:

Number and percentage of daphnids that were immobilised or showed any adverse effects (including abnormal behaviour) in the controls and in each treatment group

3.4.Endpoint units:

The model provides a quantitative prediction for Daphnia Magna LC50 (48 hour), given in -log(mol/l) and its conversion in mg/L

3.5.Dependent variable:

The model is a Tree Ensemble Random Forest made on 12 molecular descriptors associated to Daphnia sp. toxicity.

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]. Training set: n = 312 Test set: n = 133

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

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 (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.

5.2.Method used to assess the applicability domain:

The chemical similarity is measured with the algorithm developed for VEGA. Full details in the VEGA website (www.vegahub.eu), including the open access paper describing it. 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. Full details in the VEGA website.

5.3.Software name and version for applicability domain assessment:

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/

5.4.Limits of applicability:

The model is not applicable on inorganic chemicals and those including unusual elements (i.e., different from C, O, N, S, Cl, Br, F, I). Salts can be predicted only if stripped of the counterion and 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: 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:

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)

R2 = 0.68

RMSE = 0.62

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

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

6.10.Robustness - Statistics obtained by Y-scrambling:

6.11.Robustness - Statistics obtained by bootstrap:

6.12.Robustness - Statistics obtained by other methods:

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

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:

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:

Test set: n = 133

R2 = 0.7

RMSE = 0.61

7.8.Predictivity - Assessment of the external validation set:

7.9.Comments on the external validation of the model:

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:

The mechanistic interpretation was determined a posteriori by interpreting and modifying the final set of descriptors which contributed to the best fit.

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

9.3.Supporting information:

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

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