	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **48 hour Daphnia Magna LC50 Model version EPA v1.0.8**
	Printing Date: Sept 2022

1.QSAR identifier

1.1. QSAR identifier (title):

48 hour Daphnia Magna LC50 Model version EPA

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:

September 2022

2.2. QMRF author(s) and contact details:

[1] Emilio Benfenati IRCCS-Istituto di Ricerche Farmacologiche Mario Negri Via La Masa 19, 20156Milano, Italy emilio.benfenati@marionegri.it https://www.marionegri.it/

[2] Anna Lombardo IRCCS-Istituto di Ricerche Farmacologiche Mario Negri Via La Masa 19, 20156Milano, Italy anna.lombardo@marionegri.it https://www.marionegri.it/

[3] Chayawan IRCCS-Istituto di Ricerche Farmacologiche Mario Negri Via La Masa 19, 20156Milano, Italy chayawan.chayawan@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] Alberto Manganaro RCCS-Istituto di Ricerche Farmacologiche Mario Negri Via La Masa 19,20156 Milano, Italy alberto.manganaro@marionegri.it

[2] Emilio Benfenati IRCCS - Istituto di Ricerche Farmacologiche Mario Negri Via La Masa 19, 20156Milano, Italy emilio.benfenatii@marionegri.it https://www.marionegri.it/

2.6. Date of model development and/or publication:

NA

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 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 in mg/l and -log(mol/l)

3.5. Dependent variable:

-log(mol/l)

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:

The regression coefficient has been calculated on the T.E.S.T. original dataset, that contains 337 compounds extracted from the ECOTOX aquatict oxicity database (http://cfpub.epa.gov/ecotox/), split in 269 compounds for the training set and 68 for the test set. More details on the model can be found in the T.E.S.T. documentation:http://www.epa.gov/nrmrl/std/cppb/qsar/testuserguide.pdf

4.Defining the algorithm - OECD Principle 2

4.1. Type of model:

The model is developed on 269 substances.

Linear regression equation built using descriptors entirely calculated by an in-house software module in which they are implemented as described in: R. Todeschini and V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, 2009. The original code for descriptors calculation has been kindly provided by Todd Martin

4.2. Explicit algorithm:

NA

4.3. Descriptors in the model:

[1]xc4: Simple 4th order cluster chi index

[2]StN: Sum of (tN) E-States (StN)

[3]SsSH: Sum of (-SH) E-States (SsSH)

[4]SsOH_acnt: Count of (-OH) (SsOH_acnt)

[5]Hmax: Maximum hydrogen E-State value in molecule

[6]MDEN33: Molecular distance edge between all tertiary nitrogens

[7]BEHm1: Highest eigenvalue n. 1 of Burden matrix / weighted by atomic masses

[8]BEHp1: Highest eigenvalue n. 1 of Burden matrix / weighted by atomic polarizabilities

[9]Mv: Mean atomic van der Waals volume (scaled on Carbon atom)

[10]MATS1m: Moran autocorrelation - lag 1 / weighted by atomic masses

[11]MATS1e: Moran autocorrelation - lag 1 / weighted by atomic Sanderson electronegativities

[12]GATS3m: Geary autocorrelation - lag 3 / weighted by atomic masses

[13]AMR: Ghose-Crippen molar refractivity

[14]C(=S)- [2 nitrogen attach]: -C(=S)- [2 nitrogen attach] fragment count

[15]AN: AN fragment count

[16]N< [attached to P]: -N< [attached to P] fragment count

[17]S(=O)(=O)- [aromatic attach]: -S(=O)(=O)- [aromatic attach] fragment count

4.4. Descriptor selection:

A genetic algorithm was used to select descriptors, maintaining rule of maximum 1 descriptor for 5 molecules (topliss ratio). The genetic algorithm is used to maximize the adjusted 5 fold leave many out cross validation coefficient. If a chemical is determined to be an outlier, the chemical is deleted from the cluster and the genetic algorithm descriptor selection is repeated. The process of model building via the genetic algorithm and outlier removal is repeated until no outliers are detected in the optimized mode

4.5. Algorithm and descriptor generation:

The descriptors are entirely calculated by an in-house software module in which they are implemented as described in: R. Todeschini and V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley- VCH, 2009. The original code for descriptors calculation has been kindly provided by Todd Martin

4.6. Software name and version for descriptor generation:

NA

4.7. Chemicals/Descriptors ratio:

269 chemicals/17 descriptors = 15.82

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 ≥ 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 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 ≥ 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 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 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

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 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:

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:

Experimental value [-log(mol/l)]

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:

All the chemical structures were manually checked deleting doubtful compounds, mixture, inorganic compounds and tautomers

6.7. Statistics for goodness-of-fit:

Training set (269 chemicals): R^2 0.71, RMSE 0.96

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: No

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:

To test the model a test set of 68 compounds, obtained from the same source of the training set was used

7.6. Experimental design of test set:

The dataset were randomly split into training and test set with respectively the 80% and the 20% of the compounds

7.7. Predictivity - Statistics obtained by external validation:

Test set (68 chemicals) R2 = 0.47 RMSE = 1.7

Test set in AD: n 12, R2 0.80, RMSE 0.51

Test set could be out AD: n 27, R2 0.55, RMSE 0.91

Test set out AD: n 29, R2 0.36, RMSE 1.52

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]R. Todeschini and V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, 2009

[2]Topliss, J. G., and Edwards, R. P. 1979. Chance factors in Studies of Quantitative Structure-Activity Relationships. Journal of Medicinal Chemistry 22 (10):1238-1244.

[3]User’s Guide for T.E.S.T. (version 4.2) (Toxicity Estimation Software Tool), 2012, U.S.Environmental Protection Agency

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