	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **48 hour Daphnia Magna LC50 Model version**
	Printing Date: **Feb 14, 2020**

1.QSAR identifier

1.1. QSAR identifier (title):

48 hour Daphnia Magna LC50 Model version

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:

14 February 2019

2.2. QMRF author(s) and contact details:

[1] Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri – IRCCS emilio.benfenati@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]Nicolas Amaury BioChemics Consulting SAS, Orléans, France

[3]Elena Boriani ebor@food.dtu.dk

[4]Mosè Casalegno

[5]Antonio Chana

[6]Qasim Chaudhry qasim.chaundhry@fera.gsi.gov.uk

[7]Jacques R. Chrétien

[8]Jane Cotterill

[9]Frank Lemke

[10]Nadège Piclin

[11]Marco Pintore

[12]Chiara Porcelli

[13]Nicholas Price

[14]Alessandra Roncaglioni Istituto di Ricerche Farmacologiche Mario Negri – IRCCS alessandra.roncaglioni@marionegri.it

[15]Andrey Toropov Istituto di Ricerche Farmacologiche Mario Negri - IRCCS andrey.toropov@marionegri.i

2.6. Date of model development and/or publication:

September 2007

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

[1] Amaury, N.; Benfenati, E.; Boriani, E.; Casalegno, M.; Chana, A.; Chaudhry, Q.; Chrétien, J. R.; Cotterill, J.; Lemke, F.; Piclin, N.; Pintore, M.; Porcelli, C.; Price, N.; Roncaglioni, A.; Toropov, A. Chapter 7 - Results of DEMETRA Models. In Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes; Benfenati, E., Ed.; Elsevier: Amsterdam, 2007; pp 201–281. https://doi.org/10.1016/B978-044452710-3/50009-4.

[2] Chiara Porcelli, Elena Boriani, Alessandra Roncaglioni, Antonio Chana, and Emilio Benfenati

Environmental Science & Technology 2008 42 (2), 491-496

DOI: 10.1021/es071430t

[3] 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 (2004 & 1984)

3.3. Comment on endpoint:

Number and percentage of daphnids that were immobilised in the controls and in each treatment group

3.4. Endpoint units:

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

3.5. Dependent variable:

-Log48h LC50 in -log(mol/l)

3.6. Experimental protocol:

OECD Test No. 202: It measures the percentage of immobilized daphnids after a 48h exposure to a substance All tests were performed according to OECD 202 and GLP but some tests performed before 2002 employed dispersants.

3.7. Endpoint data quality and variability:

The regression coefficients have been calculated on the DEMETRA project original dataset,[1] that contains 263 compounds extracted from various databases, split in 220 compounds for the training and 43 for the test set

4.Defining the algorithm - OECD Principle 2

4.1. Type of model:

Hybrid model based on multiple linear regressions, using on 16 molecular descriptors.

4.2. Explicit algorithm:

NA

4.3. Descriptors in the model:

[1] BEHm1 highest eigenvalue n. 1 of Burden matrix / weighted by atomic masses

[2] Eig1p Leading eigenvalue from polarizability weighted distance matrix

[3] IC2 information content index (neighborhood symmetry of 2-order)

[4] IDE mean information content on the distance equality

[5] MLOGP Moriguchi octanol-water partition coeff. (logP)

[6] Mp mean atomic polarizability (scaled on Carbon atom)

[7] MW molecular weight

[8] nHAcc number of acceptor atoms for H-bonds (N O F)

[9] nNR2Ph number of tertiary amines (aromatic)

[10] nP number of Phosphorous atoms

[11] O-057 phenol / enol / carboxyl OH

[12] O-060 Al-O-Ar / Ar-O-Ar / R..O..R / R-O-C=X

[13] S-107 R2S / RS-SR; Class: atom-centred fragments

[14] SRW05 self-returning walk count of order 05

[15] T(F..Cl) sum of topological distances between F..Cl

[16] WA mean Wiener index

4.4. Descriptor selection:

See [2] in bibliography

4.5. Algorithm and descriptor generation:

See [2] and [3]

4.6. Software name and version for descriptor generation:

NA

4.7. Chemicals/Descriptors ratio:

220 chemicals/16 descriptors = 14

5.Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model:

The 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 > 0.8, the predicted substance is regarded in the Applicability Domain of the model. It corresponds to “good reliability” of prediction.

If 0.8 ≥ AD index > 0.65, the predicted substance could be out of the Applicability Domain of the model. It corresponds to “moderate reliability” of prediction.

If AD index ≤ 0.65, 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:

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

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 occurences 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 initialADI 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 ≥	IdxConcordance ≥	IdxMaxError ≥	InitialADI ≥	ADI
1.0	1.0	1.0	0.85	1.0
0.7	0.7	0.7	0.65	0.8
All other cases				0.65

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 in the VEGA website (www.vegahub.eu), including the open access paper describing it [5]. 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 > 0.75, strongly similar compounds with known experimental value in the training set have been found

If 0.75 ≥ index > 0.65, only moderately similar compounds with known experimental value in the training set have been found

If index ≤ 0.65, 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 < 0.7, accuracy of prediction for similar molecules found in the training set is good

If 1 ≥ index ≥ 0.7, accuracy of prediction for similar molecules found in the training set is not optimal

If index > 1, 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 < 0.7, molecules found in the training set have experimental values that agree with the target compound predicted value

If 1 ≥ index ≥ 0.7, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index > 1, 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 > index ≥ 0.7, the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

If index ≥ 1, 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:

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:

All

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 (220 chemicals): R^2 = 0.65, RMSE = 1.37

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: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: No

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

All

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 dataset were randomly split into training and test set with respectively the 84% and the 16% of the compounds

7.7. Predictivity - Statistics obtained by external validation:

Validation set (43chemicals) R^2 = 0.57 RMSE = 1.49

Test set in AD: No molecules are “in AD”

Test set could be out of AD: No molecules are “could be out of AD”

Test set out of AD: n = 43; R^2 = 0.57 RMSE = 1.49

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] Emilio Benfenati, Elena Boriani, Marian Craciun, Ladan Malazizi, Daniel Neagu, Alessandra Roncaglioni, Chapter 2 - Databases for pesticide ecotoxicity, Editor(s): EMILIO BENFENATI, Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes, Elsevier, 2007, Pages 59-81, https://doi.org/10.1016/B978-044452710-3/50004-5.

[2] Nicolas Amaury, Emilio Benfenati, Severin Bumbaru, Antonio Chana, Marian Craciun, Jacques R. Chrétien, Giuseppina Gini, Gongde Guo, Frank Lemke, Viorel Minzu, Johann-Adolf Müller, Daniel Neagu, Marco Pintore, Silviu Augustin Stroia, Paul Trundle. Chapter 5 – Hybrid systems, Editor(s): EMILIO BENFENATI, Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes, Elsevier, 2007 Pages 149-183, https://doi.org/10.1016/B978-044452710-3/50007-0.

[3] Nicolas Amaury, Emilio Benfenati, Elena Boriani, Mosé Casalegno, Antonio Chana, Qasim Chaudhry, Jacques R. Chrétien, Jane Cotterill, Frank Lemke, Nadége Piclin, Marco Pintore, Chiara Porcelli, Nicholas Price, Alessandra Roncaglioni, Andrey Toropov, Chapter 7 - Results of DEMETRA models, Editor(s): EMILIO BENFENATI, Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes, Elsevier, 2007, Pages 201-281, https://doi.org/10.1016/B978-044452710-3/50009-4.

[4] 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. https://pubs.acs.org/doi/abs/10.1021/jm00196a017

[5] 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.

[6] OECD. Test No. 202 (2004 & 1984): 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:

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