*QMRF identifier (JRC Inventory):*To be entered by JRC

*QMRF Title:*48 hour Daphnia Magna EC50 Model version

*Printing Date:*9-nov-2020

1.QSAR identifier

1.1.QSAR identifier (title):

48 hour Daphnia Magna EC50 Model version

1.2.Other related models:

1.3.Software coding the model:

VEGA platform

The VEGA software provides QSAR models to predict tox, ecotox, environ, and phys-chem properties of chemical substances

emilio.benfenati@marionegri.it

https://www.vegahub.eu/

2.General information

2.1.Date of QMRF:

24/02/2018

2.2.QMRF author(s) and contact details:

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

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

[3]Chayawan IRCCS-Istituto di Ricerche Farmacologiche Mario Negri Via La Masa 19, 20156 Milano, Italy chayawan.chayawan@marionegri.it https://www.marionegri.it/

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

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, 20156 Milano, Italy emilio.benfenatii@marionegri.it https://www.marionegri.it/

2.6.Date of model development and/or publication:

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

2.8.Availability of information about the model:

The files describing the parameters are freely available through the VEGA web site. The training and test sets are also available

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

No

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Daphnia magna

3.2.Endpoint:

ECOTOX 3.1 Short-term toxicity to Daphnia (immobilization)

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:

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

3.7.Endpoint data quality and variability:

The regression coefficient have been calculated on the T.E.S.T. original dataset, that contains 337 compounds extracted from the ECOTOX aquatic toxicity 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:

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:

N/A

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 model

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:

4.7.Chemicals/Descriptors ratio:

269 chemicals (training set)/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 of predictions is assessed using an Applicability Domain Index (ADI) as implemented in VEGA that has values from 0 (worst case) to 1 (best case). The ADI is calculated by grouping several other indices, each one taking into account a particular issue of the applicability domain. Most of the indices are based on the calculation of the most similar compounds found in the training and test set of the model, calculated by a similarity index that consider molecule's fingerprint and structural aspects (count of atoms, rings and relevant fragments).

5.2.Method used to assess the applicability domain:

For each index, including the final ADI, three intervals for its values are defined, such that the first interval corresponds to an evaluation within the applicability domain, the second one corresponds to a borderline evaluation and the last one corresponds to an evaluation outside the applicability domain.
The ingredients of the final ADI are the following:
- Similar molecules with known experimental value. This index takes into account how similar are the first two most similar compounds found.
The Applicability Domain of the model is defined by considering several parameters as described below:
1. 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:

1 >= index > 0.85 strongly similar compounds with known experimental value in the training set have been found, 0.85 >= index > 0.7 only moderately similar compounds with known experimental value in the training set have been found, index <= 0.7 no similar compounds with known experimental value in the training set have been found.

2. Accuracy (average error) of prediction for similar molecules. This index takes into account the error in prediction for the two most similar compounds found. Values near 0 mean that the predicted compounds falls in an area of the model's space where the model gives reliable predictions, otherwise the greater is the value, the worse the model behaves. Defined intervals are: index < 0.6 accuracy of prediction for similar molecules found in the training set is good, 0.6 <= index <= 1.2 accuracy of prediction for similar molecules found in the training set is not optimal, index > 1.2 accuracy of prediction for similar molecules found in the training set is not adequate.

3. Concordance with similar molecules (average difference between target compound prediction and experimental values of 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 agrees with the experimental values found in the model's space, thus the prediction is reliable. Defined intervals are: index < 0.6 similar molecules found in the training set have experimental values that agree with the target compound predicted value, 0.6 <= index<= 1.2 similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value, index > 1.2 similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value.4. 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 falls in an area of the model's space where the model gives reliable predictions without any outlier value. Defined intervals are: index < 0.6 the maximum error in prediction of similar molecules found in the training sethas a low value, considering the experimental variability, 0.6 <= index < 1.2 the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability, index >= 1.2 the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability.5. 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: index = 1 all atom centered fragment of the compound have been found in the compounds of the training set, 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, 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.6. 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 set7. Global AD Index. The final global index takes into account all the
previous indices, in order to give a general global assessment on the applicability domain for the predicted compound. Defined intervals are:1 >= index > 0.85 predicted substance is into the Applicability Domain
of the model, 0.85 => index > 0.7 predicted substance could be out of the Applicability Domain of the model, index <= 0.7 predicted substance is out of the the Applicability Domain of the model

5.3.Software name and version for applicability domain assessment:

VEGA
Included in the VEGA software and automatically displayed when running the model
emilio.benfenati@marionegri.it
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 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:

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:

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:

The statistics for linear regression model are the following:Training set (269 chemicals)R2 = 0.71RMSE = 0.96

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: 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:

The statistic for linear regression model are the following:

Training set (68 chemicals)

R2 = 0.49

RMSE = 1.02

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:

N/A

8.2.A priori or a posteriori mechanistic interpretation:

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

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

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