	QMRF identifier (JRC Inventory):**To be entered by JRC**
	QMRF Title:**Daphnia Magna Acute (EC50) Toxicity model (IRFMN/Combase) (version 1.0.0)**
	Printing Date:**13-feb-2020**

1.QSAR identifier

1.1.QSAR identifier (title):

Daphnia Magna Acute (EC50) Toxicity model (IRFMN/Combase) (version 1.0.0)

1.2.Other related models:

None

1.3.Software coding the model:

CORAL

http://www.insilico.eu/coral

2.General information

2.1.Date of QMRF:

15 April 2019

2.2.QMRF author(s) and contact details:

[1]Andrey Toropov Istituto di Ricerche Farmacologiche Mario Negri IRCCS

andrey.toropov@marionegri.it https://www.marionegri.it/

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

emilio.benfenati@marionegri.it https://www.marionegri.it/

[3]Giovanna J. Lavado Istituto di Ricerche Farmacologiche Mario Negri IRCCS

giovanna.lavado@marionegr.it https://www.marionegri.it/

[4]Domenico Gadaleta Istituto di Ricerche Farmacologiche Mario Negri IRCCS

domenico.gadaleta@marionegri.it https://www.marionegri.it/

2.3.Date of QMRF update(s):

13/02/2020

2.4.QMRF update(s):

Updates made by Giuseppa Raitano, email: giuseppa.raitano@marionegri.it

Updates in the fields:

-2.1 date

-2.2 addition of affiliation's url

-2.7 addition of DOI

2.5.Model developer(s) and contact details:

[1]Andrey Toropov Istituto di Ricerche Farmacologiche Mario Negri IRCCS

andrey.toropov@marionegri.it

[2]Alla Toropova Istituto di Ricerche Farmacologiche Mario Negri IRCCS

alla.toropova@marionegri.it

2.6.Date of model development and/or publication:

September, 2018

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

Prediction of biochemical endpoints by the CORAL software: Prejudices, Paradoxes, and Results

2.8.Availability of information about the model:

Model’s guide is available for download from VEGA platform

(www.vegahub.eu)

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

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Daphnia magna

3.2.Endpoint:

[1][2]ECOTOX 6.1.3. Short-term toxicity to aquatic inverterbrates

3.3.Comment on endpoint:

48 hour Daphnia magna EC50 according to OECD 202

3.4.Endpoint units:

EC50 in mmol/L

3.5.Dependent variable:

Log EC50

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:

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Regression

4.2.Explicit algorithm:

Endpoint = C0 + C1 DCW (T, N)

CORAL (http://www.insilico.eu/coral) was used to develop regression QSAR model from SMILES-based optimal descriptors. A CORAL mathematical model describes the relationship between an endpoint (dependent variable) and relevant SMILES attributes (independent variable), as shown in the equation: Endpoint = C0 + C1 DCW (T, N) where C0 and C1 are the intercept and slope for the relationship, and DCW (T, N) is the combination of SMILES-based attributes, each associated with a correlation weight (CW). CWs are determined with the Monte Carlo algorithm in an iterative procedure that aims to optimize a target function (TF). The TF is calculated as shown in the equation: TF = R + R' – |R-R'| 0.01 where R and R' are the correlation coefficients between DCW(T, N) and the endpoints for TS and ITS. This procedure is defined as a balance of correlations (BC). The TF is a function of the CWs and is optimized by iteratively modifying them. In the first part of the optimization, CWs are incremented by a value Dstart. This increment is repeated as long as these was a corresponding improvement of the TF. When no further improvement is observed, the Dstart value is modified to Dstart,1 = -0.5 (Dstart) for subsequent iterations. Dstart is iteratively modified each time that an increment of CWs fails to correspond to an increment of TF, until | Dstart | is lower than a threshold value (Dprecession).

4.3.Descriptors in the model:

SMILES-based attributes These SMILES-based attributes can be described as in the following Equation: DCW (T*,N*) = CW(Sk) + CW (SSk) where Sk and SSk are SMILES attributes defined by a sequence of atoms and bonds present in the SMILES string. Sk represents single elements, and SSk two elements combined. Attributes with a positive CW are considered promoters of an increase of the endpoint value, while attributes with a negative correlation weights are considered promoters of a decrease.

4.4.Descriptor selection:

N is the number of epochs of Monte Carlo for optimization of the TF, and T is a threshold used to classify SMILES attributes as rare or not rare. An attribute is defined as rare if it is found in the SMILES of the CS less than T times. Rare SMILES attribute values were set to zero so they were not involved in the modeling. T and N are set to optimize the statistical performance for the CS. For this model, parameters were set as follows:

T = 1; N = 35; Dstart= 0.5; Dprecession = 0.1

4.5.Algorithm and descriptor generation:

Simplified Molecular Input Line Entry System (SMILES) notation describes the structure of a chemical using linear strings in place of the classical bi- or tri- dimensional representation. CORAL breaks the SMILES strings of the TS compounds into small components (SMILES-based attributes). Each SMILES-based attributes check the presence of particular characters (or combinations of characters) within the SMILES.

4.6.Software name and version for descriptor generation:

CORAL-2017

4.7.Chemicals/Descriptors ratio:

133 chemicals /95 SMILES attributes (NB, these are not descriptors)

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.

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:

CORAL-2017

5.4.Limits of applicability:

Inorganic compounds are not predicted; rare SMILES attributes are identified by the software.

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:

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:

For modelling purposes, these compounds have been removed: with a chemical structure not clearly identified, the inorganics compounds, the metal complexes, the salts containing organic polyatomic counterions, the mixtures and the substances of Unknown or Variable composition (UVCB). In addition, the structure of the salts has been neutralized and duplicates have been removed. The final list was of 133 compounds.

6.7.Statistics for goodness-of-fit:

Statistics in the development phase of the model:

Training set (TS) R2=0.7522 RMSE=1.16

Invisible training set (ITS) R2=0.7578 RMSE=1.26

Calibration set (CS) R2=0.7005 RMSE=1.35

Statistics of the implemented version of the model for the training set:

RMSE=1.24, R2=0.7237, n. of compounds=99

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:

Training set (TS)

Q2=0.7163

Invisible training set (ITS)

Q2=0.7333

Calibration set (CS)

Q2=0.6651

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:

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:

The initial dataset was randomly divided into a training set (TS) of 33 compounds, an invisible training set (ITS) of 34 compounds, a calibration set (CS) of 32 compounds, and a validation set (VS) of 34 compounds. The CORAL software was unable to process the SMILES for one chemical which was filtered out. The TS was used for regression model's derivation. The ITS was used as "inspector" during model derivation, to confirm (or reject) predictivity of the model for substances which were not involved directly to the optimization process. The CS detected the beginning of overfitting by verifying the increase of the correlation between descriptors and endpoint during the optimization process, until improvements were no longer observed. The VS is the final estimator of the predictive potential of the model.

7.7.Predictivity - Statistics obtained by external validation:

Statistics in the development phase of the model: Validation set (VS) R2=0.7506 Statistics of the implemented version of the model for the

test set:RMSE 1.08, R2 0.5970, n. of compounds= 34

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:

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1]Use of quasi-SMILES to model biological activity of "micelle–polymer" samples

[2]Quasi-SMILES as a tool to predict removal rates of pharmaceuticals and dyes in sewage https://doi.org/10.1016/j.psep.2018.07.003

[3]Prediction of antimicrobial activity of large pool of peptides using quasi-SMILES https://doi.org/10.1016/j.biosystems.2018.05.003

[4]The application of new HARD-descriptor available from the CORAL software to building up

NOAEL models https://doi.org/10.1016/j.fct.2017.03.060

[5]Prediction of biochemical endpoints by the coral software: Prejudices, paradoxes, and Results

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