	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Activated sludge toxicity (ProtoQSAR/COMBASE) (Version 1.0.0)**
	Printing Date: **2019-apr-16**

1.QSAR identifier

1.1.QSAR identifier (title):

Activated sludge toxicity (ProtoQSAR/COMBASE) (Version 1.0.0)

1.2.Other related models:

None

1.3.Software coding the model:

Activated sludge toxicity (ProtoQSAR/COMBASE) V 1.0.0

The model is based on the OECD 209, Activated Sludge, Respiration Inhibition Test. Test data provides a qualitative prediction when non-toxic and a quantitative prediction when toxic.

http://www.vega-qsar.eu/

2.General information

2.1.Date of QMRF:

April 2019

2.2.QMRF author(s) and contact details:

Sergi Gómez ProtoQSAR SL +34 960880658 sgomez@protoqsar.com https://protoqsar.com/

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1]Sergi Gómez-Ganau ProtoQSAR SL +34 960880658 sgomez@protoqsar.com https://protoqsar.com/

[2]Rafael Gozalbes ProtoQSAR SL +34 960880658 rgozalbes@protoqsar.com https://protoqsar.com/

2.6.Date of model development and/or publication:

February 2019

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

VEGA-QSAR: AI inside a platform for predictive toxicology. Proceedings of the workshop "Popularize Artificial Intelligence 2013" Benfenati E, Manganaro A, Gini G., December 5th 2013, Turin, Italy Published on CEUR Workshop Proceedings Vol-1107 http://ceur-ws.org/Vol1107/paper8.pdf

2.8.Availability of information about the model:

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

None to date

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Activated sludge

3.2.Endpoint:

Ecotoxicological properties: Activated Sludge, Respiration Inhibition Test (OECD 209)

3.3.Comment on endpoint:

This method to determine the effects of a substance on microorganisms from activated sludge (largely bacteria) by measuring their respiration rate (carbon and/or ammonium oxidation) under defined conditions in the presence of different concentrations of the test substance. Thanks to this test, a rapid screening can be performed to assess the effects of chemical compounds on the microorganisms of the activated sludge.

The respiration rates of samples of activated sludge with test substance and without (blank controls) is incubated with synthetic sewage and measured in an enclosed cell containing an oxygen electrode after a contact time of 3 hours. The sensitivity of each batch of activated sludge is also tested with a suitable reference substance (i.e. 3,5-dichlorophenol). The test is typically used to determine the ECx (e.g. EC_{50}) of the test substance and/or the no-observed effect concentration (NOEC).

3.4.Endpoint units:

EC 50 in mg/L

3.5.Dependent variable:

Log EC50

3.6.Experimental protocol:

3.7.Endpoint data quality and variability:

Experimental data for EC_{50} after 3 hours on activated sludge, respiratory inhibition test, was retrieved from the COMBASE dataset and the different databases available within the OECD QSAR Toolbox, v. 4.2. (www.qsartoolbox.org).

95 biocide-like compounds were found by application of the biocide-like filters. Biocide-like filters were previously defined as those properties featuring the structural chemical space of most of biocides. To do this, and in the context of the LIFE-EU COMBASE project (http://www.life-combase.com), different cut-off values for a list of physicochemical parameters were determined by comparing databases of biocides and generic chemicals, and served to identify a set of common biocide properties

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

An integrated model to predict the respiratory inhibition in activated sludge was arranged cascading a qualitative QSAR model and a quantitative QSAR model. Firstly, the quantitative model discriminates between a toxic or non-toxic compound. A compound is considered toxic when EC50 < 100 mg/L. If the compound is considered non-toxic, a qualitative output is given. When the prediction for the compound is toxic, the quantitative model base on a MLR is applied and a value of toxicity is given.

4.2.Explicit algorithm:

Boosted trees

The algorithm for Boosting Trees evolved from the application of boosting methods to regression

trees. The general idea is to compute a sequence of (very) simple trees, where each successive tree is built for the prediction residuals of the preceding tree. Thus, at each step of the boosting (boosting trees algorithm), a simple (best) partitioning of the data is determined, and the deviations of the observed values from the respective means (residuals for each partition) are computed. The next 3-node tree will then be fitted to those residuals, to find another partition that will further reduce the residual (error) variance for the data, given the preceding sequence of trees.

Multiple linear regression

Multiple linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical (dummy coded as appropriate).

4.3.Descriptors in the model:

[1]Qualitative model descriptors:

[2]MaxHother Maximum atom-type H E-State: H on aaCH, dCH2 or dsCH

[3]MinwHBa Minimum E-States for weak Hydrogen Bond acceptors

[4]ETA_BetaP_ns_d A measure of lone electrons entering into resonance relative to molecular size

[5]Gats3c Geary autocorrelation - lag 3 / weighted by charges

[6]MinsCH3 Minimum atom-type E-State: -CH3

[7]ATSC4p Centered Broto-Moreau autocorrelation - lag 4 / weighted by polarizabilities

[8]SpMax1_Bhm Largest absolute eigenvalue of Burden modified matrix - n 1 / weighted by relative mass

[9]GATS1i Geary autocorrelation - lag 1 / weighted by first ionization potential

[10]Quantitative model descriptors

[11]ATSC7v Centered Broto-Moreau autocorrelation - lag 7 / weighted by van der Waals volumes

[12]MinHBint Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 2

[13]VE3_DzZ Logarithmic coefficient sum of the last eigenvector from Barysz matrix / weighted by atomic number

[14]AATSC4e Average centered Broto-Moreau autocorrelation - lag 4 / weighted by Sanderson electronegativities

[15]BCUTp-1l nhigh lowest polarizability weighted BCUTS

4.4.Descriptor selection:

Molecular descriptors were calculated using CDK, Padel descriptor and E-Dragon software. Constant variables, near-constant variables and 0.95 pair-correlation variables were discarded. A sensitivity analysis for the qualitative model and a forward stepwise for the quantitative model was used for variable selection.

4.5.Algorithm and descriptor generation:

Molecular descriptors were calculated using CDK, Padel descriptor and E-Dragon software.

4.6.Software name and version for descriptor generation:

4.7.Chemicals/Descriptors ratio:

Qualitative model: 95 / 8 = 11.87

Quantitative model: 35 / 5 = 7

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. On 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:

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: null

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:

6.6.Pre-processing of data before modelling:

The model is based on a dataexperimental data for EC_{50} after 3 hours on activated sludge, respiratory inhibition test. Data was retrieved from the COMBASE dataset and the different databases available within the OECD QSAR Toolbox, v. 4.2. (www.qsartoolbox.org).

95 biocide-like compounds were found by application of the biocide-chemical space, and were used to develop the quantitative model. Molecular

descriptors were calculated using CDK, Padel descriptor and E-Dragon software. Constant variables, near-constant variables and 0.95 pair-correlation variables were discarded. Once the variables were calculated, STATISTICA and MINITAB packages were used to carry out the model. First the whole dataset was randomly divided in training set (70%) and validation set (15%). After, a sensitivity analysis approach was used for variable selection and boosted trees analysis was performed. Once the qualitative model was developed, a quantitative model was performed by using the 35 toxic compounds (EC50 < 100 mg/L).

6.7.Statistics for goodness-of-fit:

Qualitative model:

Training (70% dataset): Accuracy 87.79%, Specificity 97.29%, Sensitivity 78.26%

Quantitative model:

R2 = 0.70

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

Q2 LOO = 0.69

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

Q2 L5O = 0.66

6.10.Robustness - Statistics obtained by Y-scrambling:

N/A

6.11.Robustness - Statistics obtained by bootstrap:

N/A

6.12.Robustness - Statistics obtained by other methods:

N/A

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: null

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:

7.6.Experimental design of test set:

A validation was performed in the qualitative model randomly selecting 15% of the data collection.

7.7.Predictivity - Statistics obtained by external validation:

Qualitative model:

External validation set (15% dataset): Accuracy 80.30%, Specificity 77.27%, Sensitivity 83.33%

7.8.Predictivity - Assessment of the external validation set:

N/A

7.9.Comments on the external validation of the model:

N/A

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:

N/A

9.Miscellaneous information

9.1.Comments:

N/A

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