	***QMRF identifier (JRC Inventory):*** **To be entered by JRC**
	QMRF Title: **Sludge Classification Toxicity model (ProtoQSAR/COMBASE) (Version 1.0.1)**
	Printing Date: Apr 16, 2022

1.QSAR identifier

1.1.QSAR identifier (title):

Sludge Classification Toxicity model (ProtoQSAR/COMBASE) (Version 1.0.1)

1.2.Other related models:

NA

1.3.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:

April 2022

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):

NA

2.4.QMRF update(s):

NA

2.5.Model developer(s) and contact details:

[1] Alberto Manganaro Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2,20156 Milano, Italy alberto.manganaro@marionegri.it https://www.marionegri.it/

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

[3]Rafael Gozalbes ProtoQSAR SL +34 960880658 rgozalbes@protoqsar.comhttps://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:

Gomez-Ganau S, Marzo M, Gozalbes R, Benfenati E Computational approaches to evaluate ecotoxicity of biocides: cases from the project COMBASE In: Ecotoxicological QSARs Springer Nature Switzerland AG, Cham 2020

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:

NA

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Activated sludge

3.2.Endpoint:

Ecotoxicological properties: Activated Sludge, Respiration Inhibition Test (OECD 209) (2010)[1]

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:

Two classes based on 3h EC50 (threshold: EC50 (3hrs) < 100 mg/L for toxic substances)

3.5.Dependent variable:

Binary classification: Toxic, NON-Toxic

3.6.Experimental protocol:

According to the OECD 209 test guideline (2010)[1]

3.7.Endpoint data quality and variability:

Experimental data on mono-constituent organic substances for EC 50 after 3 hours on activated sludge, respiratory inhibition test, was retrieved from the COMBASE dataset [2] and the different databases available within the OECD QSAR Toolbox, v. 4.2. (www.qsartoolbox.org). 94 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) [2], 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. Particularly, to identify a set of common biocide-like relevant properties, a comparison was carried out, using a set of physical parameters and cut-off values between the data collected for biocides and the dataset of generic organic compounds. Biocide-like filters were defined as those properties able to maximize the difference between the retention and the rejection of compounds from both datasets. The different parameters to characterize the structures from both databases, were calculated using different software: Padel descriptor, CDK and FAF-drugs4 [5-7]. Based on the statistical distribution, different cut-off values for the physicochemical parameters were applied to both databases trying 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 qualitative model discriminates between a toxic or non-toxic compound. A compound is considered toxic when EC50 (3hrs) < 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 may be applied and a value of toxicity predicted

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.

4.3.Descriptors in the model:

Qualitative model descriptors:

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

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

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

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

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

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

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

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

4.4.Descriptor selection:

Molecular descriptors were calculated using CDK, Padel descriptor and E-Dragon software. Constant variables, near-constant variables and 0.95pair-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:

NA

4.7.Chemicals/Descriptors ratio:

Qualitative model: 95/8 = 12

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 (c.f. point 5.2 below)

ADI 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.

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:

$$\frac{\sum_c log\ (1 + S_c)}{\sum_k log\ (1 + S_k)}$$

where the molecules with *c* index are the subset of the *k* molecules where the prediction of the model matches with the experimental value of the molecule.

Concordance index (*IdxConcordance*) is calculated as:

$$\frac{\sum_c log\ (1 + S_c)}{\sum_k log\ (1 + S_k)}$$

where the molecules with *c* index are the subset of the *k* molecules where the experimental value of the molecule matches with the prediction made for the target molecule.

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:

$$ADI = (IdxSimilarity^{0.5} \times IdxAccuracy^{0.25} \times IdxConcordance^{0.25}) \times IdxACF \times IdxDescRange$$

5.2. Method used to assess the applicability domain:

The Applicability Domain and the 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 [3]. The VEGA 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 1.5 ≥ index ≥ 0.8, accuracy of prediction for similar molecules found in the training set is not optimal

If index > 1.5, 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 1.5 ≥ index ≥ 0.8, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index > 1.5, 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 1.5 ≥ 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 > 1.5, 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 descriptor 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:

If index = TRUE, descriptors for this compound have values inside the descriptor range of the compounds of the training set

If 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

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 to inorganic chemicals and substances containing unusual elements (i.e., different from C, O, N, S, P, Cl, Br, F, I). Salts can be predicted only if converted to the neutralized form.

Model is applicable for the Molecular weight between 18<MW<750 and octanol water partition coefficient between (-3 to +6)

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:

NA

6.6.Pre-processing of data before modelling:

First the whole dataset was randomly divided in training set and validation set.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 n = 60 (23 Toxic, 37 NON-Toxic), Sensitivity = 78%; Specificity = 97. %; Accuracy = 88%

TP = 18, TN = 36, FP = 1, FN = 5

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:

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:

NA

7.6.Experimental design of test set:

A validation was performed in the qualitative model randomly selecting.

7.7.Predictivity - Statistics obtained by external validation:

Qualitative model:

Validation set n = 34 (12 Toxic, 22 NON-Toxic), Sensitivity = 83%; Specificity = 77%; Accuracy = 80.%

TP=10, TN = 17, FP = 5, FN = 2

Validation set in AD: n = 6; Sensitivity = 1; Specificity = 0.67; Accuracy = 0.83 MCC = 0.71 TP:3 TN:2 FP:1 FN:0

Validation set could be out of AD: n = 8, Sensitivity = 1; Specificity = 0.67; Accuracy = 0.75; MCC = 0.58 TP:2, TN:4, FP:2, FN=0

Validation set out of AD: n = 20; Sensitivity = 0.71; Specificity = 0.85; Accuracy = 0.80; MCC = 0.56 TP:5, TN:11, FP:2, FN:2

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:

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:

NA

9.Miscellaneous information

9.1.Comments:

NA

9.2.Bibliography:

[1] OECD. Test Guideline No. 209: Activated Sludge, Respiration Inhibition Test (Carbon and Ammonium Oxidation). Paris: Organisation for Economic Co-operation and Development, 2010. https://www.oecd-ilibrary.org/environment/test-no-209-activated-sludge-respiration-inhibition-test_9789264070080-en.

[2] 'COMBASE'. Accessed 2 March 2022. https://www.life-combase.com/index.php/it/.

[3] Floris, Matteo, Alberto Manganaro, Orazio Nicolotti, Ricardo Medda, Giuseppe Felice Mangiatordi, and Emilio Benfenati. 'A Generalizable Definition of Chemical Similarity for Read-Across'. Journal of Cheminformatics 6, no. 1 (18 October 2014): 39. https://doi.org/10.1186/s13321-014-0039-1.

[4] 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

[5] Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry, 32(7), 1466–1474. https://doi.org/10.1002/jcc.21707

[6] Steinbeck, C., Hoppe, C., Kuhn, S., Floris, M., Guha, R., & Willighagen, E. L. (2006). Recent developments of the chemistry development kit (CDK)—An open-source java library for chemo- and bioinformatics. Current Pharmaceutical Design, 12(17), 2111–2120. https://doi.org/10.2174/138161206777585274

[7] FAFDrugs4 Home. (n.d.). Retrieved April 22, 2022, from https://fafdrugs4.rpbs.univ-paris-diderot.fr/

9.3.Supporting information:

Training set(s)Test set(s)Supporting information:

All available dataset 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