	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Fish Acute Toxicity (LC50) toxicity model (IRFMN-Combase) v 1.0.1**
	Printing Date: **14-07-2022**

1.QSAR identifier

1.1. QSAR identifier (title):

Fish Acute Toxicity (LC50) toxicity model (IRFMN-Combase) v 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

CORAL

A CORAL mathematical model describes the relationship between an endpoint and relevant SMILES attributes. andrey.toropov@marionegri.it

http://www.insilico.eu/coral

CDK

The Chemistry Development Kit

The CDK developers

https://github.com/cdk

2.General information

2.1. Date of QMRF:

10-07-2022

2.2. QMRF author(s) and contact details:

[1]Alessio Gamba Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2,20156 Milano, Italy alessio.gamba@marionegri.it https://www.marionegri.it/

[2]Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2,20156 Milano, Italy emilio.benfenati@marionegri.it https://www.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] 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]Andrey Toropov Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, Italy andrey.toropov@marionegri.it https://www.marionegri.it/

2.6. Date of model development and/or publication:

The model was developed in 2018.

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

[1] Khan, K., Khan, P. M., Lavado, G., Valsecchi, C., Pasqualini, J., Baderna, D., Marzo, M., Lombardo, A., Roy, K., & Benfenati, E. (2019). QSAR modeling of Daphnia magna and fish toxicities of biocides using 2D descriptors. Chemosphere, 229, 8–17. https://doi.org/10.1016/j.chemosphere.2019.04.204

[2] Toropov, A.A., Toropova, A.P., Roncaglioni, A., Benfenati, E. "Prediction of biochemical endpoints by the CORAL software: Prejudices, Paradoxes, and Results" (2018) Methods in Molecular Biology,1800, pp. 573-583. doi: 10.1007/978-1-4939-7899-1_27

.

[3]Benfenati E, Roncaglioni A, Lombardo A, Manganaro A. Integrating QSAR, Read-Across, andScreening Tools: The VEGAHUB Platform as an Example. Advances in Computational Toxicology;Springer; 2019. p. 365-81.

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:

Fish

3.2. Endpoint:

ECOTOX 6.1.1. Short-term toxicity to fish. OECD Test Guideline No. 203: Fish, Acute Toxicity Test [4]

3.3. Comment on endpoint:

The test evaluates the effects of the tested chemical on fish. LC50 96h, acute (mortality)

3.4. Endpoint units:

-Log(mmol/L) (-Log = negative logarithm with base 10) then it is converted in mg/L

3.5. Dependent variable:

-Log(96h EC50)

3.6. Experimental protocol:

OECD TG 203 [4]

3.7. Endpoint data quality and variability:

To build the datasets of biocides with acute toxicity data, we downloaded the list of biocides from the ECHA website. We manually retrieved the structures and we compared them with the chemical names, and the Chemical Abstracts Service (CAS) numbers using online databases:

-ChemCell (https://github.com/cdd/chemcell);

-MarvinView (Marvin v17.28.0, 2017,ChemAxon (http://www.chemaxon.com));

-ChemIDplus Advanced (https://chem.nlm.nih.gov/chemidplus/);

-PubChem (https://pubchem.ncbi.nlm.nih.gov/);

-ChemSpider (http://www.chemspider.com/)).

For modeling purposes, we removed the compounds with a chemical structure not clearly identified, the inorganic compounds, the metal complexes, the salts containing organic polyatomic counterions, the mixtures and the substances of Unknown or Variable composition (UVCB). In addition, we neutralized the structure of the salts. Hence the substances used as training set for our model exclusively consisted of mono-constituent organic substances. We searched for the toxicity data on several public sources: the OECD QSAR toolbox v. 4.2 (www.qsartoolbox.org), the Pesticide Properties Database (PPDB) database (https://sitem.herts.ac.uk/aeru/ppdb/), the Office of Pesticide Programs (OPP) Pesticides Ecotoxicity

Database (http://www.ipmcenters.org/ecotox/)), the European Food Safety Authority (EFSA) (http://www.efsa.europa.eu/) database and the ECOTOX (https://cfpub.epa.gov/ecotox/) database. For fish, we also used the AMBIT (http://cefic-lri.org/toolbox/ambit/) database. We only accepted test data generated with the relevant Organization of Economic Cooperation and Development (OECD) guidelines (OECD TG 203 Fish acute toxicity (4). In case of multiple LC50 value for the same compound, we used the threshold established by the European Commission (SANCO/10597/2003) (European, 2012) as the ratio between maximum value and the minimum experimental value (x/y) (compounds with values of difference >3 were removed).

We compared the final LC50 in mg/l with the experimental water solubility retrieved in the OECD QSAR toolbox. We eliminated the compounds (11) with the LC50 greater than the water solubility.

Experimental data were carefully screened for specific endpoints and identical exposure time (96 hrs) in order to get reliable predictions from standardized data. For the ease of interpretation, the half maximal lethal concentration (LC50) or the half maximal effective concentration (EC50) values were converted into a molar unit (LC50/EC50 in molL−1) followed by transformation into a negative logarithmic scale, i.e., pEC50 or pLC50 as customary in ecotoxicological QSAR analysis. In contrast to half maximal effective/lethal concentration, a higher value of pEC50 or pLC50 corresponds to higher toxicity and vice versa. The final QSAR analysis was performed using toxicity data of 88 compounds for mortality at 96 h (LC50–96 h) of fish (i.e., in this case we had data for Brachdanio rerio, Pimephales promelas, Cyprinus carpio, Oryzias latipes, Poecilia reticulata, Lepomis macrochirus, Oncorhyncus mykiss).

In case of multiple data, the smallest LC50-96h value is kept..

The final dataset was composed of 88 mono-constituent organic chemicals. The data set was distributed in the active training set (≈25%), invisible training set (≈25%), calibration set (≈25%), and external validation set (≈25%) which are categories which are sued in the specific CORAL modeling approach (http://www.insilico.eu/coral/) c.f. also point 6.5

4.Defining the algorithm - OECD Principle 2

4.1. Type of model:

The regression QSAR model provides a quantitative prediction of acute toxicity in Fish (LC50), given in -log(mmol/L) and converted in mg/L on 88 biocides. It has been developed inside the EU LIFE COMBASE project (LIFE15 ENV/ES/416). It is implemented inside the VEGA online platform, accessible at: http://www.vegahub.eu.

4.2. Explicit algorithm:

Regression model from optimal descriptors based on SMILES.The CORAL mathematical model describes the relationship between an endpoint (dependent variable) and relevant SMILES attributes (independent variable)

The final model formula obtained with CORAL is:

Endpoint = -0.9887356 (± 0.1138834) + 0.0454373 (± 0.0013499) * DCW(1,15)

where DCW is represented by the following formula

$$DCW(T^*, N^*) = \sum_{k=1}^{NA} CW(S_k) + \sum_{k=1}^{NA-1} CW(SS_k) + \sum_{k=1}^{NA-2} CW(SSS_k)$$

where Sk is SMILES atom; SSk is pairs of SMILES atoms; SSSk is compositions of three SMILES atoms; the CW (Sk), CW(SSk), and CW(SSSk) are correlation weights of the above SMILES attributes; the T is the threshold to separate SMILES attributes in two classes, i.e. rare and non-rare. The N is the number of

epochs of the Monte Carlo optimisation. The T* and N* are values which provide better statistics for the calibration set. The NA is the number of Sk in the SMILES.

4.3. Descriptors in the model:

The complete list of 368 fragments (descriptors) selected is provided here:

SAk	CW(SAk)
#.....	3.74739
(...(.....	-1.12485
(.....	0.00307
(...C...#...	0.0
(...C...(...	0.62320
(...F...(...	0.93749
(...Br..(...	0.0
(...I...(...	0.0
(...Cl..(...	0.43955
(...N...#...	0.0
(...N...(...	2.06503
(...O...(...	-0.68385
(...c...(...	0.87382
(...n...(...	0.0
+.....	0.0
+...[...(...	0.0
-.....	0.0
-...[...(...	0.0
1...(...(...	0.0
1...(.....	1.56523
1.....	2.62983
1...2...(...	0.0
1...C...(...	2.37940
1...Br..(...	0.0
1...Cl..(...	0.0
1...N...(...	0.0
1...S...(...	0.0
1...c...(...	1.99726
2...(...(...	0.0
2...(.....	3.37471
2.....	0.25095
2...1.....	0.0
2...C...(...	-0.87306
2...Br..(...	0.0
2...Cl..(...	0.0
2...N...(...	-0.00233
2...O...(...	0.0
2...c...(...	0.25214
2...c...1...	8.50023
2...s...1...	3.44182
3...(.....	-0.31557
3.....	-0.31599
3...4...(...	0.0
3...C...(...	0.99653
3...C...2...	0.0
3...Cl..(...	0.0
3...c...(...	2.56065
4...(.....	-1.12389
4.....	0.19204

4...3.....	0.0
4...c...(...	0.18812
4...c...1...	0.43449
5...(.....	0.00219
5.....	0.12045
5...C...(...	0.49691
5...C...1...	0.81447
5...c...1...	0.0
5...c...2...	0.18678
5...c...3...	0.0
6...(.....	0.68812
6.....	-0.00208
=...(...(...	0.0
=...(.....	1.19152
=.....	1.74963
=...C...(...	0.50413
=...C...1...	0.31270
=...C...2...	0.25325
=...C...3...	0.0
=...N...(...	0.0
=...O...(...	-1.24952
=...S...(...	-0.30936
C...#.....	3.81210
C...#...C...	1.37415
C...(...(...	1.12305
C...(.....	0.43854
C...(...1...	2.74838
C...(...2...	-0.25445
C...(...3...	0.50328
C...(...4...	0.55937
C...(...6...	0.69136
C...(...=...	2.50366
C...(...C...	0.31547
C.....	1.12376
C...1...(...	4.74948
C...1.....	0.49732
C...1...C...	-1.43557
C...2...(...	0.0
C...2.....	0.12906
C...2...C...	0.0
C...3...(...	1.56273
C...3.....	0.50275
C...3...4...	0.0
C...3...C...	0.68545
C...4.....	0.0
C...4...C...	0.0
C...5...(...	0.06655
C...5.....	0.06459
C...=...(...	2.00332
C...=.....	4.50464
C...=...C...	6.37113
C...C...#...	4.37473
C...C...(...	0.50490
C...C.....	1.99597
C...C...1...	0.0

C...C...2...	3.00161
C...C...3...	0.43875
C...C...=...	4.49812
C...C...C...	2.12749
C...Br..(...	0.0
C...I...(...	1.75238
C...N...(...	1.44158
C...N...1...	0.0
C...N...=...	0.49761
C...O...(...	0.62532
C...O...1...	0.0
C...O...2...	0.0
C...O...C...	0.0
C...S...(...	0.0
C...S...1...	0.0
C...S...2...	-0.62596
C...S...C...	1.99814
C...c...1...	0.75021
C...c...2...	2.93655
C...c...3...	-3.81195
C...n...2...	-0.00284
F...(...(...	0.05913
F...(.....	0.74675
F...(...C...	0.37817
F...(...F...	1.43815
F.....	0.25451
H.....	0.62691
H...[...1...	0.50476
Br..(...(...	0.0
Br..(.....	0.0
Br..(...C...	0.0
Br..(...Br..	0.0
Br.....	0.0
Br..1.....	0.0
Br..2.....	0.0
Br..C...(...	0.0
Br..C.....	0.0
I...(.....	4.12339
I...(...C...	0.0
I...(...I...	0.0
I.....	3.43458
I...C...#...	0.87854
I...C.....	2.30955
Cl..(...(...	-0.18798
Cl..(.....	1.18498
Cl..(...1...	-0.06480
Cl..(...2...	0.0
Cl..(...3...	0.94074
Cl..(...C...	2.25004
Cl..(...F...	-0.81204
Cl..(...Cl..	1.99563
Cl.....	1.62338
Cl..1.....	0.0
Cl..2.....	0.0
Cl..3.....	0.0

Cl..N...(...	0.12207
N...#.....	3.93884
N...#...C...	5.37277
N...(...(...	3.18380
N...(.....	2.06538
N...(...1...	0.0
N...(...C...	1.25217
N...(...I...	2.05953
N...(...Cl..	-3.31302
N...(...N...	0.0
N...+.....	0.0
N.....	1.18729
N...1.....	0.0
N...1...C...	0.0
N...1...Cl..	0.0
N...2.....	-0.06498
N...2...C...	0.37214
N...=...(...	0.0
N...=.....	-0.19158
N...=...C...	-4.30784
N...C...#...	-0.06567
N...C...(...	0.0
N...C.....	1.24620
N...C...1...	0.0
N...C...2...	0.0
N...C...C...	0.93825
N...Cl..(...	-0.18672
N...Cl.....	1.50282
N...S...(...	0.0
N...S...1...	1.87828
N...S...C...	0.0
N...[...=...	0.0
N...c...1...	0.37815
O...(...(...	-1.12256
O...(.....	-0.12432
O...(...1...	0.0
O...(...2...	0.0
O...(...3...	0.0
O...(...5...	0.0
O...(...=...	0.81172
O...(...C...	-1.75248
O...(...Br..	0.0
O...(...N...	1.99795
O...(...O...	0.0
O...-.....	0.0
O.....	0.49868
O...1.....	0.06169
O...1...C...	0.18672
O...2...(...	0.0
O...2.....	0.0
O...=...(...	-0.74566
O...=.....	0.05866
O...=...C...	0.37693
O...C...(...	1.93818
O...C.....	1.12973

O...C...4...	0.0
O...C...=...	0.0
O...C...C...	0.49669
O...O...(...	3.68709
O...O.....	5.75390
O...[...(...	0.0
O...c...1...	2.55832
O...c...2...	3.37636
O...c...6...	-0.24621
S...(.....	2.24615
S...(...=...	-0.62606
S...(...C...	0.0
S...(...Cl..	0.0
S.....	3.99638
S...1...(...	0.0
S...1.....	2.24512
S...1...N...	0.0
S...2.....	0.12217
S...=...(...	0.0
S...=.....	4.49794
S...=...O...	0.18947
S...C...#...	2.12163
S...C...(...	0.0
S...C.....	2.25077
S...C...=...	0.0
S...C...C...	0.12585
S...C...S...	1.62136
S...N.....	10.12430
S...N...1...	0.0
S...N...2...	2.74929
S...c...2...	1.93918
[...(.....	1.74813
[...(...2...	0.00372
[...(...C...	0.0
[...(...N...	0.0
[...(...[...	0.0
[...+.....	0.0
[...+...N...	0.0
[...-.....	0.0
[...-...O...	0.0
[.....	-0.18726
[...1...(...	0.74817
[...1.....	1.44004
[...=.....	0.0
[...=...O...	0.0
[...H.....	-0.19176
[...N...+...	0.0
[...N.....	0.0
[...O...-...	0.0
[...O.....	0.0
[...c...(...	0.0
[...c...1...	0.00263
[...n...H...	0.12081
c...(...(...	1.75077
c...(.....	0.43394

c...(...1...	0.44031
c...(...2...	0.56472
c...(...C...	-0.24907
c...(...F...	-0.25078
c...(...Cl..	1.00485
c...(...N...	0.0
c...(...O...	2.50327
c...(...[...	0.0
c...(...c...	0.12232
c.....	1.00089
c...1...(...	0.12634
c...1.....	0.37362
c...1...2...	0.0
c...1...C...	-0.19168
c...1...Br..	0.0
c...1...N...	0.0
c...1...O...	0.0
c...1...S...	0.37525
c...1...c...	3.93614
c...2...(...	1.56416
c...2.....	0.93464
c...2...C...	0.37341
c...2...Br..	0.0
c...2...Cl..	0.0
c...2...O...	0.0
c...2...c...	0.30867
c...3...(...	-0.56068
c...3.....	0.81391
c...3...C...	0.0
c...3...Cl..	0.0
c...3...c...	0.49858
c...4...(...	0.24632
c...4.....	0.30893
c...4...c...	-0.06070
c...5...(...	0.0
c...5.....	0.25183
c...5...C...	0.43617
c...5...c...	0.0
c...6...(...	-0.00196
c...6.....	-0.00285
c...6...c...	-0.18818
c...C...#...	0.0
c...C...(...	0.56144
c...C.....	0.93356
c...C...1...	0.0
c...C...O...	-0.31442
c...N...(...	-0.37783
c...N.....	-0.24619
c...O...(...	2.12292
c...O.....	1.99593
c...O...1...	-0.05816
c...O...C...	0.0
c...S.....	2.18393
c...S...C...	2.87066
c...[.....	0.24847

c...[...H...	0.12219
c...c...(...	0.87399
c...c.....	0.75071
c...c...1...	0.87296
c...c...2...	0.62662
c...c...3...	-0.99567
c...c...4...	0.24507
c...c...5...	0.0
c...c...6...	0.00172
c...c...[...	0.44135
c...c...c...	0.62641
c...n...(...	0.0
c...n...1...	0.62021
c...n...2...	0.06451
c...n...c...	-2.93387
n...(.....	0.0
n...(...Cl..	0.0
n...(...c...	0.0
n.....	0.00205
n...1.....	1.69058
n...1...c...	0.93915
n...2.....	0.99865
n...2...c...	9.93463
n...2...n...	-0.12718
n...C...(...	-0.62980
n...C.....	-1.18674
n...H.....	0.44235
n...H...[...	0.18689
n...[...(...	8.50378
n...[.....	0.12223
n...[...c...	0.31723
n...c...(...	0.87989
n...c.....	-2.12188
n...c...1...	2.06282
n...c...2...	0.68608
n...c...n...	-4.00202
n...n...(...	0.0
n...n.....	0.0
n...n...1...	0.0
n...n...c...	0.0
s.....	2.31655
s...1.....	2.56089
s...1...c...	1.74884
s...2.....	1.69083

4.4. Descriptor selection:

See 4.2

4.5. Algorithm and descriptor generation:

See section 4.2

4.6. Software name and version for descriptor generation:

CORAL

The CORAL mathematical model describes the relationship between an endpoint and relevant SMILES attributes.andrey.toropov@marionegri.ithttp://www.insilico.eu/cora

4.7. Chemicals/Descriptors ratio:

66/368 = 0.179

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.

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.

5.2. Method used to assess the applicability domain:

The Applicability domain and chemical similarity are measured with the algorithm developed for VEGA. Full details are in the VEGA website (www.vegahub.eu), including the open access paper describing it [5]. 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 1.2 ≥ index ≥ 0.8, accuracy of prediction for similar molecules found in the training set is not optimal

If index > 1.2, 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 1.2 ≥ index ≥ 0.6, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index > 1.2, similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value

Maximum error of prediction between similar molecules:

This index takes into account the maximum error in prediction between 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 1.2 ≥ index ≥ 0.6, the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

If index > 1.2, the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability

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:

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

If index= False, the maximum error in prediction of similar molecules found in the training set has a moderate 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 atoms centered fragments of the compound have not been found in the compounds of the training set or are rare fragments

5.3. Software name and version for applicability domain assessment:

VEGA

The VEGA software provides QSAR models to predict tox, ecotox, environ, and phys-chemproperties of chemical substances. 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 stripped of the counter ion and 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:

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:

Dataset N=88, Training set N=66, Validation set N=22.

Training set is divided in: Active Training set (N=22), Passive Training set(N=22), Calibration set (N=22)

The training set is composed of “Training”, “Invisible Training” and “Calibration”. Invisible training and calibration sets were used during the model development for tuning model's parameters".

The training represents an *active* training set and is used to build up optimal correlation weights for the optimal descriptor; in other word it represents the set on which the model is built on. The passive (or invisible) training set is for the purpose of checkup whether current correlation weights (and the optimal descriptor) are satisfactory for chemicals, which are not involved in the calculation of the correlation weights

The task for the calibration set is to detect the moment when overtraining begins (http://www.insilico.eu/coral/)

6.6. Pre-processing of data before modelling:

NA

6.7. Statistics for goodness-of-fit:

Active Training set: R^2 = 0.79, Q^2 = 0.74, RMSE = 0.89

Passive training set: R^2 = 0.83, Q^2 = 0.81, RMSE = 1.40

Calibration set: R^2 = 0.84, Q^2 = 0.80, RMSE = 0.71

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:

No

7.2. Available information for the external validation set:

NA

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

NA

7.4. Data for the dependent variable for the external validation set:

NA

7.5. Other information about the external validation set:

NA

7.6. Experimental design of test set:

NA

7.7. Predictivity - Statistics obtained by external validation:

Validation set: R^2 = 0.73, RMSE = 0.72

Validation set in AD: n = 0

Validation set could be out of AD: n = 2,

Validation set out of AD: n = 20, RMSE 0.56, R2 0.57

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:

Statistical model CORAL (http://www.insilico.eu/coral) was used to develop a 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 explained in: Toropov, A.A., Toropova, A.P.,Roncaglioni, A., Benfenati, E. "Prediction of biochemical endpoints by the CORAL software: Prejudices, Paradoxes, and Results" (2018) Methods in Molecular Biology, 1800, pp. 573-583

8.2. A priori or a posteriori mechanistic interpretation:

NA

8.3. Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1. Comments:

NA

9.2. Bibliography:

[1] Khan, K., Khan, P. M., Lavado, G., Valsecchi, C., Pasqualini, J., Baderna, D., Marzo, M., Lombardo, A., Roy, K., & Benfenati, E. (2019). QSAR modeling of Daphnia magna and fish toxicities of biocides using 2D descriptors. Chemosphere, 229, 8–17. https://doi.org/10.1016/j.chemosphere.2019.04.204

[2] Toropov, A.A., Toropova, A.P., Roncaglioni, A., Benfenati, E. "Prediction of biochemical endpoints by the CORAL software: Prejudices, Paradoxes, and Results" (2018) Methods in Molecular Biology,1800, pp. 573-583. doi: 10.1007/978-1-4939-7899-1_27

[3]Benfenati E, Roncaglioni A, Lombardo A, Manganaro A. Integrating QSAR, Read-Across, andScreening Tools: The VEGAHUB Platform as an Example. Advances in Computational Toxicology;Springer; 2019. p. 365-81

[4] OECD. (1992, 2019). Test Guideline No. 203: Fish, Acute Toxicity Test. Organisation for Economic Co-operation and Development. https://www.oecd-ilibrary.org/environment/test-no-203-fish-acute-toxicity-test_9789264069961-en

[5] Floris, M., Manganaro, A., Nicolotti, O. et al. A generalizable definition of chemical similarity for read-across. J Cheminform 6, 39 (2014). https://doi.org/10.1186/s13321-014-0039-1

[6] Gadaleta D, Lombardo A, Toma C, Benfenati E. A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications. J Cheminform. 2018 Dec 10;10(1):60. doi: 10.1186/s13321-018-0315-6. Erratum in: J Cheminform. 2019 Apr 25;11(1):31. PMID: 30536051; PMCID: PMC6503381.

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