	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Fish Toxicity classification Model (SARpy/IRFMN) v-1.0.3**
	Printing Date: **June 10, 2022**

1.QSAR identifier

1.1. QSAR identifier (title):

Fish Toxicity classification Model (SARpy/IRFMN) v-1.0.3

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:

June 10, 2022

2.2. QMRF author(s) and contact details:

[1] Anna Lombardo Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2,20156 Milano, Italy anna.lombardo@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/

[3] Erika Colombo Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2,20156 Milano, Italy erika.colombo@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] Anna Lombardo Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2,20156 Milano, Italy anna.lombardoi@marionegri.it https://www.marionegri.it/

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

2.6. Date of model development and/or publication:

The model was developed in 2019.

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

[1] Gini G, Ferrari T, Lombardo A, Cassano A, Benfenati E (2019) A New QSAR Model for Acute Fish Toxicity based on Mined Structural Alerts. J Toxicol Risk Assess 5:016. doi.org/10.23937/2572-4061.1510016

[2] 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:

Another QMRF is not available.

3.Defining the endpoint - OECD Principle 1

3.1. Species:

Fathead minnow (Pimephales promelas)

3.2. Endpoint:

ECOTOX 6.1.1. Short-term toxicity to fish. OECD TG 203 “Fish, Acute Toxicity Test” [3]

3.3. Comment on endpoint:

The endpoint is the acute toxicity (LC50 96h) toward fish expressed in four classes:

< 1 mg/L Very toxic to aquatic organisms (class 1)

1 - 10 mg/L Toxic to aquatic organisms (class 2)

10 - 100 mg/L Harmful to aquatic organisms (class 3)

> 100 mg/L May cause long term adverse effects to aquatic organisms (class 4)

3.4. Endpoint units:

Adimensional

3.5. Dependent variable:

The initial value in Log 1/LC50 were converted in mg/L and then in the classes listed at the point 3.3

3.6. Experimental protocol:

The data used to develop the model (training set) has been collected from [2]. They refer to LC50 96h for Pimephales promelas. The data used for the test set (LC50 96 h for Oncorhynchus mykiss) come from two sources: the beta version of the OECD toolbox selecting the OECD-HPV inventory, and the pesticide dataset used to develop the Demetrad model for fish. The data were checked to eliminate duplicates, mixture andinorganic compounds. Salt were considered in their neutral form

3.7. Endpoint data quality and variability:

Data quality check was described in [1].

The data set contains 568 different compounds. Data are quite representative for most industrial chemicals, but they are still a very small percentage of the commercialized chemicals, and an even minor part of all possible chemicals humans can be exposed to. Since these values are widely spread and to take into account the regulations, the results were also transformed into the classification for toxicity to fish as provided by Directive 92/32/EEC of the EU for dangerous substances [5].

4.Defining the algorithm - OECD Principle 2

4.1. Type of model:

Classification

4.2. Explicit algorithm:

NA

4.3. Descriptors in the model:

SARpy fragments Model was build using the fragments extracted from the SARpy.

4.4. Descriptor selection:

Following, the list of the alerts, as SMARTS strings, related to class 1 toxicity:

[1] C(OCCCC)c1cccc(c1)

[2] c1cc(c(O)c(c1)C(C)C)

[3] Oc1ccc(cc1C)Cl

[4] O=Cc1cccc(Oc2ccc(cc2))c1

[5] Nc1ccc(cc1)CCCCCCCC

[6] O=[N+]([O-])c1c(cc(c(c1)[N+](=O)[O-])Cl)Cl

[7] NCCCCCCCCCCCC

[8] O(CC)P(=S)(OCC)SCS

[9] CC[Sn](CC)(CC)CC

[10] c1cc(ccc1CCl)

[11] O=CC=CC(=O)

[12] N#CCC#N

[13] [c]([Cl,Br,F])[c]([Cl,Br,F])[c]([Cl,Br,F])

[14] CCC(=O)OC[a]

Following, the list of the alerts, as SMARTS strings, related to class 2 toxicity:

[15] Oc1ccc(cc1)Cl

[16] C(C)CCCCCCCCC

[17] c1ccc(Oc2ccccc2)cc1

[18] C(=O)Oc1ccccc1

[19] C(OCC=C)CC

[20] c1c(cccc1CC)CC

[21] O(CC)P(=S)(O)

[22] C=CC=C

[23] C(=O)OCCCC

[24] O=[N+]([O-])c1cc(cc(c1)C)

[25] c1cc(c(cc1C))C

[26] SCC

[27] O(c1ccccc1)CCCC

[28] c1cccc2ccccc12

[29] c1cc(ccc1O)Br

[30] c1ccccc1c2ccccc2

[31] O=Cc1c(F)cccc1

[32] I

[33] O=C(OC)CC

[34] [*;D1]#C[C;!D4][!C;D1]

[35] *[C;D2][C;D2][C;D2][C;D2][C;D2][C;D2]*

[36] [s;R]

[37] [$([c]([Cl,Br,F])[c]([Cl,Br,F])),$([c]([Cl,Br,F])[c][c]([Cl,Br,F]))]

Following, the list of the alerts, as SMARTS strings, related to class 3 toxicity:

[38] C(OC)c1ccc(cc1)

[39] NCCCCCC

[40] c1cc(c(cc1)C)C

[41] Fc1ccccc1

[42] O=C(OCCC)C

[43] c1ccc(cc1)Br

[44] O=[N+]([O-])c1cc(N)ccc1

[45] c1ccc(cc1)CCCC

[46] C=CCCC

[47] c1ccc(cc1)Cl

[48] Oc1ccc(cc1)C

[49] c1cc(ccc1N(C)C)

[50] c1cc(N)ccc1O

[51] CCCCCCCC

[52] C(C(Cl))Cl

[53] O=P(OCC)(OCC)OCC

[54] N(CCC)(CCC)C

[55] o1c(ccc1)

[56] C#CCC(O)

[57] C(CCCl)C

[58] OCC#CC

[59] O=[C;D2][C;D2]

[60] C=C

[61] c1ccccc1

4.5. Algorithm and descriptor generation:

The model has been built as three sets of rules, extracted with Sarpy software, related to different toxicity classes. Each of the three set contains a list of relevant fragment (expressed in SMARTS notation)related to the first three toxicity classes, defined on the basis of the classification for toxicity to fish provided by Directive 92/32/EEC of the EU for dangerous substances:

Class =1-> LC50=1 mg/L -> Damage: Very toxic to aquatic organisms;

Class=2 -> LC50=1–10 mg/L -> Damage: Toxic to aquatic organisms;

Class=3 -> LC50=10–100 mg/L -> Damage: Harmful to aquatic organisms;

Class=4 -> LC50>100 mg/L -> Damage: May cause long term adverse effects to aquatic organisms.

If one or more rules are verified for the given compound, the model will assign the compound to the most toxic class available among the verified rules. If no rules apply to the given compound, the prediction will be class 4. The extraction of the rules has been performed on a training set consisting of 567 compounds, where experimental data, originally expressed as acute toxicity database to fathead minnow (Pimephales promels), has been classified on the basis of the reported scheme. The experimental data have been retrieved from [2]and described in [1].

4.6. Software name and version for descriptor generation:

SARpySARpy software can be used to extract the fragments which are responsible for the any binary class activity. emilio.benfenati@marionegri.ithttps://www.vegahub.eu/portfolio-item/sarpy/

4.7. Chemicals/Descriptors ratio:

Not applicable

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 and 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 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 [9]. 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.
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 > 0.8, 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 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 set

5.3. Software name and version for applicability domain assessment:

VEGA (www.vegahub.eu)

5.4. Limits of applicability:

6.Internal validation - OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Available information for the training set:

CAS RN: No

Chemical Name: No

Smiles: Yes

Formula: No

INChI: No

MOL file: Yes

NanoMaterial: Null

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:

training set n 567

6.6. Pre-processing of data before modelling:

See [1, 2]. 1 compound were eliminated because considered not of sufficient quality

6.7. Statistics for goodness-of-fit:

[1] reports an accuracy for the three thresholds of 100, 10, and 1 mg/L of 87.5%, 85%, and 92% respectively.

After the implementation in VEGA, the balanced accuracy for the three thresholds of 1, 10, 100 mg/L and no-Tox are 0.74, 0.78, 0.76, 0.83 respectively

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:

NO

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:

See [1]

7.6. Experimental design of test set:

See [1]

7.7. Predictivity - Statistics obtained by external validation:

Paper [1] reports an accuracy for the three thresholds of 100, 10, and 1 mg/L of79%, 63%, and 65% respectively for an external dataset.

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 model includes SAs to identify the activity class of the compounds. The VEGA system provides, in the final PDF report for the prediction, a set built with the most similar compounds found in the training and test set of the model. An expert-based analysis of these compounds like the predicted one, which are provided with their experimental activity, can lead to a further mechanistic interpretation of the results given by the model

8.2. A priori or a posteriori mechanistic interpretation:

A posteriori: the fragments identified as statistically associated tothe toxicity class can be investigated to explore the mechanistic basisof the model

8.3. Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1. Comments:

The sensitivity and specificity can not be calculated since the model is not a binary classifier.

9.2. Bibliography:

[1]Gini et al. A New QSAR Model for Acute Fish Toxicity based on Mined Structural Alerts. J ToxicolRisk Assess 2019, 5:016 DOI: 10.23937/2572-4061.1510016

[2]C.L. Russom, S.P. Bradbury, D.E. Hammermeister, S.J. Drummond “Predicting modes of toxicaction from chemical structure: acute toxicity in the fathead minnow (Pimephales promelas)”Environmental Toxicology Chemistry 16, 1997, pp. 948-967.

[3] OECD (1981, 1984 and 1992), Test No. 203: Fish, Acute Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, https://www.oecd.org/env/ehs/testing/section2-effects-on-biotic-systems-replaced-and-cancelled-test-guidelines.htm

[4] Floris M, Manganaro A, Nicolotti O, Medda R, Mangiatordi GF, Benfenati E. A generalizable definition of chemical similarity for read-across. J Cheminform. 2014 Oct 18;6(1):39. doi: 10.1186/s13321-014-0039-1. PMID: 25383097; PMCID: PMC4212147.

[5] Directive 92/32/ECC, the seventh amendment to Directive 67/548/EEC, UJL 154 of 5.VI.92, 1992, www.eurunion.org1legislat1chemical.htm.

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