	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Fish Acute (LC50) Toxicity model (NIC) (version 1.0.1)**
	Printing Date: **08-07-2022**

1.QSAR identifier

1.1. QSAR identifier (title):

Fish Acute (LC50) Toxicity model (NIC) (version 1.0.1)

1.2. Other related models:

Fish Acute Toxicity Read-Across version 1.0.1

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

Counter Propagation Artificial Neural Network (CP ANN)

CPANNatNIC Software for developing self-organizing maps (SOM) and counter-propagation artificial neural network models https://www.ki.si/en/departments/d01-theory-department/laboratory-for-cheminformatics/software/

CPANNatNIC software for counter-propagation neural network to assist in read-across

Article describing the software (available in Additional files) that implements the CP ANN algorithm used to build the model.

https://jcheminf.biomedcentral.com/articles/10.1186/s13321-017-0218-y

2.General information

2.1. Date of QMRF:

July 2022

2.2. QMRF author(s) and contact details:

[1]Viktor Drgan National Institute of Chemistry, Hajdrihova 19, 1001 Ljubljana, Slovenia (NIC)viktor.drgan@ki.si

[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] Viktor Drgan National Institute of Chemistry, Hajdrihova 19, 1001 Ljubljana, Sloveniaviktor.drgan@ki.si https://www.ki.si/

[3]Marjana Novic National Institute of Chemistry, Hajdrihova 19, 1001 Ljubljana, Sloveniamarjana.novic@ki.si https://www.ki.si/

2.6. Date of model development and/or publication:

The model was developed in 2016.

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

[1] Benfenati, E.; Lombardo, A.; Drgan, V.; Novič, M.; Manganaro, A. The Tools for Aquatic Toxicology within the VEGAHUB System. In Chemometrics and Cheminformatics in Aquatic Toxicology; John Wiley & Sons, Ltd, 2021; pp 493–511. https://doi.org/10.1002/9781119681397.ch25.

[2] LM. Su, X. Liu, Y. Wang, J.J. Li, X.H. Wang, L.X. Sheng, Y.H. Zhao, “The discrimination of excesstoxicity from baseline effect: Effect of bioconcentration”, Science of the Total Environment 484(2014), 137–145 https://www.sciencedirect.com/science/article/pii/S0048969714003702

[3] V. Drgan, Š. Župerl, M. Vracko, C.I. Cappelli, M. Novic. "CPANNatNIC software for counter-propagation neural network to assist in read-across", Journal of Cheminformatics 9 (2017), 30QMRF identifier (JRC Inventory):To be entered by JRCQMRF Title:Fish Acute (LC50) Toxicity model (NIC) (version 1.0.0)Printing Date:08-Jul-20201.QSAR identifier2.General information

https://jcheminf.biomedcentral.com/articles/10.1186/s13321-017-0218-y

[4]Benfenati E, Roncaglioni A, Lombardo A, Manganaro A. Integrating QSAR, Read-Across, and Screening 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:

Poecilia reticulata (guppy), Oncorhynchus mykiss (rainbow trout), Pimephales promelas (fathead minnow) and Oryzias latipes (japanese medaka)

3.2. Endpoint:

ECOTOX 6.1.1. Short-term toxicity to fish

3.3. Comment on endpoint:

The acute toxicity data expressed as LC50 (mmol/L), the concentration required to kill 50% of fish within 96 h.

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(LC50)

3.6. Experimental protocol:

OECD TG 203 “Fish, Acute Toxicity Test” (1981, 1984 & 1992)

3.7. Endpoint data quality and variability:

The data used to build the model were obtained from the paper [2]. Data collection includes data from tests done with 4 test species recommended by the OECD guideline 203.

In the paper, different sources were used to compile the data representing average value for four fish species: rainbow trout (Oncorhynchus mykiss), medaka (Oryzias latipes), fathead minnow (Pimephales promelas) and guppy (Poecilia reticulata)). For each compound, average value of -log(LC50) was calculated, depending on the availability of the data, from the values for all four fish species. The averages of -log(LC50) were used to build the model.

4.Defining the algorithm - OECD Principle 2

4.1. Type of model:

The model has been built as a Counter Propagation Artificial NeuralNetwork (CP ANN) by the National Institute of Chemistry, Slovenia (NIC)

4.2. Explicit algorithm:

Counter Propagation Artificial Neural Network (CP ANN)

A detailed description of the algorithm is given in the paper [3]

4.3. Descriptors in the model:

The complete list of 368 fragments (descriptors) selected is provided in the pdf guide of the model.

[1]nCIR number of circuits

[2]nTB number of Triple Bonds

[3]nP number of P atoms

[4]nCL number of Cl atoms

[5]nR10 number of 10-membered rings

[6]TI1 first Mohar index from Laplace matrix

[7]S2K 2-path Kier alpha-modified shape index

[8]T(N..P) sum of topological distances between N..P

[9]T(P..Cl) sum of topological distances between P..Cl

[10]T(Cl..Cl) sum of topological distances between Cl..Cl

[11]piPC09 molecular multiple path count of order 9

[12]PCR ratio of multiple path count over path count

[13]Xindex Balaban X index

[14]MATS1e Moran autocorrelation - lag 1/weighted by Sanderson electronegativity

[15]GATS7m Geary autocorrelation - lag 7/weighted by mass

[16]EEig14x eigenvalue n. 14 from edge adjacency matrix weighted by edge degree

[17]EEig14d eigenvalue n. 14 from edge adjacency matrix weighted by dipole moment

[18]ESpm01d spectral moment of order 1 rom edge adjacency matrix weighted by dipole moment

[19]GGI4 topological charge index of order 4

[20]SEigv spectral moment of order 1 from Brysz matrix weighted by van der Waals volume

4.4. Descriptor selection:

The calculation of the descriptors was done with Dragon. Calculation of descriptors was based on 2D molecular structures of compounds (for calculation of the descriptors the “Exclude 3D” option in Dragon was used). Initially, 472 descriptors were calculated for all compounds. The number of descriptors was reduced using maximum allowed pair correlation coefficient of 0.95 and minimum accepted variance of descriptor value of 0.005. In this way, 162 descriptors were obtained. Then a 25x25 Kohonen network was used to select training, test, and validation set based on distribution of compounds on Kohonen top-map. 20% of compounds were selected for external validation set, then 20% of compounds were selected for test set and the remaining compounds were put in training set. All sets contained compounds from the entire top-map so that test and external validation set contained compounds which were structurally similar to the compounds in the training set. Then, using only the data for training and test set and transposed data matrix, descriptors were grouped on a Kohonen neural network with 8x8neurons. The descriptors with the minimum and maximum Euclidean distance to each excited neuron were selected which resulted in 95 selected descriptors. These 95 descriptors were used in optimization of neural network parameters and selection of descriptors using genetic algorithm. The compounds in the training set were used to correct weight of the model and the compounds in the test set aided in finding optimal conditions. External validation set was used to evaluate the model with selected optimal conditions.

4.5. Algorithm and descriptor generation:

The descriptors were calculated, in the original model, by means of Dragon software and are now entirely calculated by an in-house software module in which they are implemented as described in [3]

4.6. Software name and version for descriptor generation:

Dragon softwarehttps://chm.kode-solutions.net/products_dragon.php

4.7. Chemicals/Descriptors ratio:

564/20= 28

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

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:

$$\frac{\sum_c^k |exp_c - pred_c|}{k}$$

where exp_c is the experimental value of the c-*th* molecule in the training set and $pred_c$ is the c-*th* molecule predicted value by the model.

Concordance index (*IdxConcordance*) is calculated as:

$$\frac{\sum_c^k |exp_c - pred_{target}|}{k}$$

where exp_c is the experimental value of the c-*th* molecule in the training set and $pred_{target}$ is the predicted value for the input target molecule.

Max Error index (*IdxMaxError*) is calculated as:

$$max(|exp_c - pred_c|)$$

where exp_c is the experimental value of the c-*th* molecule in the training set and $pred_{target}$ is the predicted value for the input target molecule, evaluated over the k molecules.

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:

$$ADI = IdxSimilarity \times IdxACF \times IdxDescRange$$

The initialADI index is the used together with the other sub-indices to calculate the final ADI, on the basis of the assessment class in which each sub-index falls:

IdxAccuracy ≥	IdxConcordance ≥	IdxMaxError ≥	InitialADI ≥	ADI
1.2	1.2	1.2	0.85	1.0
0.6	0.6	0.6	0.7	0.85
All other cases				0.7

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 [4]. 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 > 0.85, 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 < 0.6, accuracy of prediction for similar molecules found in the training set is good

If 1.2 ≥ index ≥ 0.6, 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 < 0.6, 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 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 < 0.6, 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

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

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:

Training set N=564

6.6. Pre-processing of data before modelling:

Normalized (standardized) descriptor values were used for modelling. Training set was used to calculate normalization factors (mean and standard deviation). All sets were normalized using the normalization factors. Normalization factors are available

6.7. Statistics for goodness-of-fit:

RMSE = 0.42, R2 = 0.89

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

RMSE = 1.00 when using training and test set which were used for optimization of CP ANN parameters and selection of descriptors, without 15 (~2%) worst predictions RMSE=0.91. RMSE = 1.08 when using only training set

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

Observed cross-validation RMSE values were between 1.08 and 1.2 when leave-many-out cross-validation was made using training set and 1, 4, 8, 16, 20, 24, 28, 40 and 50 objects left out during cross-validation

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:

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:

All

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

All

7.5. Other information about the external validation set:

Sum of the number of the test and external validation set compounds: 382

7.6. Experimental design of test set:

Compounds in the external validation set were selected from Kohonen top-map. See *section* 4.4

7.7. Predictivity - Statistics obtained by external validation:

For the external validation set the following statistics was obtained: R^2=0.49, RMSE=0.90. When using predictions where absolute values of standardized residuals were not above 2.0, the following statistics was obtained: R^2=0.65, RMSE=0.70

In VEGA external and test set are merged in one test set. The statistics on the sum of the test and external validation set compounds are:

Test set: R2 0.53, RMSE 0.90Test set in AD: No molecules are “in AD”

Test set could be out of AD: No molecules are “could be out of AD”

Test set out of AD: n = 380; R2= 0.53; RMSE = 0.90

7.8. Predictivity - Assessment of the external validation set:

External validation set was selected from entire Kohonen top-map so that structurally similar compounds were in training set. See Section 4.4

7.9. Comments on the external validation of the model:

For the test set compounds which were used for optimization of CP ANN parameters and selection of descriptors the following statistics was obtained: R^2=0.59, RMSE=0.90.

When using predictions where absolute values of standardized residuals were not above 2.0, the following statistics was obtained for the test set: R^2=0.68, RMSE=0.79.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

The model is based on the assumption that structurally similar compounds have similar properties. Compunds exciting the same neuron in CP ANN are considered as structurally similar

8.2. A priori or a posteriori mechanistic interpretation:

8.3. Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1. Comments:

NA

9.2. Bibliography:

[1] LM. Su, X. Liu, Y. Wang, J.J. Li, X.H. Wang, L.X. Sheng, Y.H. Zhao, “The discrimination of excess toxicity from baseline effect: Effect of bioconcentration”, Science of the Total Environment 484(2014) 137–145 https://www.sciencedirect.com/science/article/pii/S0048969714003702

[2] J. Zupan, M. Novic, J. Gasteiger,"Neural networks with counter-propagation learning strategy used for modelling", Chemometrics and Intelligent Laboratory Systems, 27 (1995), 175-187.

[3] R. Todeschini and V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, 2009

[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.

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