	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Fish Acute Toxicity Read-Across version 1.0.1**
	Printing Date: **June 10, 2022**

1.QSAR identifier

1.1. QSAR identifier (title):

Fish Acute Toxicity Read-Across version 1.0.1

1.2. Other related models:

Fish Acute Toxicity Model version 1.0.1 (NIC)

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

istKNN

application KNN read across

http://chm.kode-solutions.net

2.General information

2.1. Date of QMRF:

June 10, 2022

2.2. QMRF author(s) and contact details:

[1] 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] 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:

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:

NA

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

[1] Su LM, Liu X, Wang Y, Li JJ, Wang XH, Sheng LX, Zhao YH. The discrimination of excess toxicity from baseline effect: effect of bioconcentration. Sci Total Environ. 2014 Jun 15;484:137-45. doi:10.1016/j.scitotenv.2014.03.040

[2] 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:

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. OECD TG 203 “Fish, Acute Toxicity Test” [2]

3.3. Comment on endpoint:

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

3.4. Endpoint units:

mg/L

3.5. Dependent variable:

-Log (LC50) [mg/L]

3.6. Experimental protocol:

OECD TG 203 “Fish, Acute Toxicity Test” [2]

3.7. Endpoint data quality and variability:

Data collection includes data from tests done with 4 test species recommended by the OECD guideline nr 203. This dataset has been made by Istituto di Ricerche Farmacologiche Mario Negri, merging experimental data from several reliable sources: the database compiled by the MED-Duluth group, the OECD Toolbox, the DEMETRA Project (Rainbow Trout toxicity model) and the work of Su et al. (“The discrimination of excess toxicity from baseline effect: Effect of bioconcentration”, Science of the Total Environment, 2014, 484, 137-145)

The data was cleaned excluding metal complexes, inorganics , mixtures of structural isomers, ambiguous structures, non-ionic surfactant mixtures, complex disconnected structures, UVCB. Duplicates were removed. The final dataset is composed of 972 chemicals.

4.Defining the algorithm - OECD Principle 2

4.1. Type of model:

KNN read across

The model performs a read-across on a dataset of 972 chemicals. This dataset has been made by Istituto di Ricerche Farmacologiche Mario Negri, merging experimental data from several reliable sources (see 3.7)

4.2. Explicit algorithm:

Read-across model has been built with the istKNN application (developed by Kode srl, http://chm.kode-solutions.net) and it is based on the similarity index developed inside the VEGA platform; the index takes into account several structural aspects of the compounds, such as their fingerprint, the number of atoms, of cycles, of heteroatoms, of halogen atoms, and of particular fragments (such as nitro groups). The index value ranges from 1 (maximum similarity) to 0.

On the basis of the structural similarity index, the four compounds from the dataset resulting most similar to the chemical to be predicted are taken into account; compounds with a similarity value lower than 0.7 are discarded, and if only one compound remains available for prediction, it is kept only if it has a similarity value higher than 0.75. If no compounds fall under these conditions, no prediction is provided. Furthermore, if the range of experimental values observed in the chosen molecules is higher than 3.5 log units, no prediction is provided. The estimated toxicity value is calculated as the weighted average value of the

experimental values of the chosen compounds, using their similarity values as weight. Their similarity values are raised to the power of 3 in order to enhance the weight of the most similar compounds in the calculated prediction

4.3. Descriptors in the model:

NA

4.4. Descriptor selection:

NA

4.5. Algorithm and descriptor generation:

NA

4.6. Software name and version for descriptor generation:

istKNN application emilio benfenati http://chm.kode-solutions.ne

4.7. Chemicals/Descriptors ratio:

NA

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 ≥ 0.9, the predicted substance is regarded in the Applicability Domain of the model, It corresponds to “good reliability of prediction

If 0.9 > 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 [3]. 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 > 0.75, strongly similar compounds with known experimental value in the training set have been found

If 0.75 ≥ 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 ≥ 1.5, 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

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: Yes

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=936

6.6. Pre-processing of data before modelling:

NA

6.7. Statistics for goodness-of-fit:

Training set: n = 927, RMSE = 0.75; R^2 = 0.60; MAE 0.54

Not predicted: 45

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:

NA

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:

NA

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] Su LM, Liu X, Wang Y, Li JJ, Wang XH, Sheng LX, Zhao YH. The discrimination of excess toxicity from baseline effect: effect of bioconcentration. Sci Total Environ. 2014 Jun 15;484:137-45. doi:10.1016/j.scitotenv.2014.03.040

[2] Test No. 203: Fish, Acute Toxicity Test | OECD Guidelines for the Testing of Chemicals, Section 2: Effects on Biotic Systems | OECD iLibrary. (n.d.). Retrieved June 10, 2022, from https://www.oecd-ilibrary.org/environment/test-no-203-fish-acute-toxicity-test_9789264069961-en

[3] 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

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