	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Guppy LC50 model (KNN/IRFMN) - v. 1.1.1**
	Printing Date: **20-10-2022**

1.QSAR identifier

1.1.QSAR identifier (title):

Guppy LC50 model (KNN/IRFMN) - v. 1.1.1

1.2.Other related models:

A similar approach is used for Fathead Minnow LC50 model (KNN/IRFM) - v. 1.1.0

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:

20-07-2022

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):

No update

2.4.QMRF update(s):

No update

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:

2015

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

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:

The fish Guppy (Poecilia reticulata).

3.2.Endpoint:

ECOTOX 6.1.1. Short-term toxicity to fish

3.3.Comment on endpoint:

NA

3.4.Endpoint units:

LC50 in mmol/L

3.5.Dependent variable:

log 1/LC50 (mmol/L)

3.6.Experimental protocol:

OECD, Test No. 203: Fish, Acute Toxicity Test

3.7.Endpoint data quality and variability:

The model is based on 207 chemicals collected as decribed in: 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

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The model provides a quantitative evaluation for acute lethal toxicity in Guppy (Poecilia reticulata). It has been developed by Istituto di Ricerche Farmacologiche Mario Negri, using a k nearest neighbor (KNN) modelling approach based on the similarity index developed inside the VEGA platform (Floris et al. "A generalizable definition of chemical similarity for read-across." Journal of cheminformatics 6.1 (2014): 39). The first 2 similar compounds in the training set according to the similarity index are used for the prediction. If the mean similarity value of these compound is lower than 0.85 prediction is not provided, as also if only one neighbor is found and it has a similarity value lower than 0.8. The prediction is calculated as the weighted average value of the experimental values of the compounds selected with the above mentioned procedure, with an enhance factor of 2. If the range of activity between the selected compounds is above 3fold (minimum activity is less than 1/3 than the maximum) the prediction is discarded.

4.2.Explicit algorithm:

The k nearest neighbor (KNN) algorithm

KNN with k=2 (the model selects the 2 most similar molecules in dataset to predict the target compound).

4.3.Descriptors in the model:

Descriptors are not used in this model.

4.4.Descriptor selection:

NA

4.5.Algorithm and descriptor generation:

NA

4.6.Software name and version for descriptor generation:

NA

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 > 0.85, 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 [2]. 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 < 0.8, 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 ≥ 0.8, 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 < 0.8, 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 ≥ 0.8, the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

If index ≥ 1.2, 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 (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.

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:

NA

6.4.Data for the dependent variable for the training set:

NA

6.5.Other information about the training set:

Training set: n = 207

6.6.Pre-processing of data before modelling:

The statistics are obtained applying the read-across prediction to its original dataset, with a leave-one-out approach (read-across for each compound has been performed on the whole dataset without the compound itself).

6.7.Statistics for goodness-of-fit:

Statistics on Test set are not available, we used only the leave-one-out approach.

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

R2 = 0.87; RMSE = 0.47

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:

External validation set is not available.

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:

The model is a KNN model. No assumption on the mechanism is done.

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]L.M. 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.

[2]Floris, et al. "A generalizable definition of chemical similarity for read-across." Journal of cheminformatics 6.1 (2014): 39.

[3]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.

[4]Manganaro A, Pizzo F, Lombardo A, Pogliaghi A, Benfenati E, "Predicting persistence in the sediment compartment with a new automatic software based on the k-Nearest Neighbor (k-NN) algorithm", Chemosphere 2016; 144 : 1624-1630.

[5] OECD, Test No. 203: Fish, Acute Toxicity Test. Paris: Organisation for Economic Co-operation and Development, 2019. Accessed: Mar. 14, 2022. [Online]. Available: https://www.oecd-ilibrary.org/environment/test-no-203-fish-acute-toxicity-test_9789264069961-en

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