QMRF identifier (JRC Inventory): **To be entered by JRC**
QMRF Title: **VEGA bee acute toxicity model (kNN/IRFMN) - v.1.0.1**
Printing Date: **October 2022**

1.QSAR identifier

1.1. QSAR identifier (title):

VEGA bee acute toxicity model (kNN/IRFMN) - 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

2.General information

2.1. Date of QMRF:

October 2022

2.2. QMRF author(s) and contact details:

[1]Edoardo Carnesecchi Istituto di Ricerche Farmacologiche Mario Negri IRCCS Via Mario Negri 2,20156 Milano, Italy edoardo.carnesecchi@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] Alberto Manganaro Kode srl info@kode-solutions.net

2.3. Date of QMRF update(s):

NA

2.4. QMRF update(s):

NA

2.5. Model developer(s) and contact details:

Alberto Manganaro Kode srl info@kode-solutions.net

2.6. Date of model development and/or publication:

The model was developed in 2017.

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

[1] Como, F., Carnesecchi, E., Volani, S., Dorne, J. L., Richardson, J., Bassan, A., ... & Benfenati, E.(2017). Predicting acute contact toxicity of pesticides in honeybees (Apis mellifera) through a k-nearest neighbor model. Chemosphere, 166, 438-444

[2] 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:

Honebee (Apis mellifera)

3.2. Endpoint:

ECOTOX 6.3.2. Toxicity to terrestrial arthropods

3.3. Comment on endpoint:

Median lethal dose (LD50) after 48 hours of exposure to a substance by acute contact -according to OECD guideline 214

3.4. Endpoint units:

µg/bee

3.5. Dependent variable:

48h LD50

3.6. Experimental protocol:

According to OECD TG 214, adult worker honeybees are exposed to a range of doses of the test substance dissolved in an appropriate carrier, by direct application to the thorax (droplets). The test duration is 48h. Mortality is recorded daily and compared with control values. The results are analyzed in order to calculate the LD50 at 24 and 48h, and in case the study is prolonged at 72h and 96h

3.7. Endpoint data quality and variability:

This dataset has been made by Istituto di Ricerche Farmacologiche Mario Negri, merging experimental data from several reliable sources such:

- EFSA's OpenFoodTox database, https://www.efsa.europa.eu/en/microstrategy/openfoodtox
- DEMETRA database https://doi.org/10.1016/B978-044452710-3/50004-5
- the terrestrial US-EPA ECOTOX database. https://qsartoolbox.org/resources/databases/

Data pruning according to OECD 214 TG

4.Defining the algorithm - OECD Principle 2

4.1. Type of model:

The bee Acute toxicity model is developed is k-nearest neighbor (k-NN) model developed on 256 substances.

4.2. Explicit algorithm:

k-nearest neighbor (kNN) model

This model has been built with the istKNN application (developed by Kode srl, http://chm.kodesolutions.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 from1 (maximum similarity) to 0

The algorithm of this software is a similarity-based approach that predicts the property of a substance in relation to the experimental data for the most similar compounds “nearest neighbours” on the training set. The algorithm of istKNN software is based on the first “k” compounds of the training set able to satisfy the similarity requirement related to the target compound, indicated by threshold “S1”. If only one molecule satisfies this requirement then another, stricter threshold, indicated by threshold “S2”, is applied. If no molecule satisfies this similarity requirement, no prediction is provided (missing value)

4.3. Descriptors in the model:

Similarity index - Descriptors are only used to identify the similar compounds

4.4. Descriptor selection:

The present model does not require any descriptor as it is built on k-NN algorithm

4.5. Algorithm and descriptor generation:

The algorithm is an extension of kNN as described in section 4.2. The descriptors are only used for the similarity, as described in section4.3

4.6. Software name and version for descriptor generation:

IstKNN 0.9

in-house software for kNN modelling

alberto.manganaro@kode-solutions.ne

4.7. Chemicals/Descriptors ratio:

No descriptors (descriptors are used only to identify the similar compounds).

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 ≥ 0.80, the predicted substance is regarded in the Applicability Domain of the model. It corresponds to “good reliability” of prediction.

If 0.80 > AD index ≥ 0.6, the predicted substance could be out of the Applicability Domain of the model. It corresponds to “moderate reliability” of prediction.

If AD index < 0.6, 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 AD 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 [5]. 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 > 0.80, strongly similar compounds with known experimental value in the training set have been found

If 0.80 ≥ index > 0.6, only moderately similar compounds with known experimental value in the training set have been found

If index ≤ 0.6, 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 1 ≥ index > 0.80, accuracy of prediction for similar molecules found in the training set is good

If 0.8 ≥ index > 0.6, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≤ 0.6, 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

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:

No

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

All

6.5. Other information about the training set:

The dataset used to develop the model implemented on VEGA was not splitted in training and test sets. The training set consists in the overall dataset, n= 256

6.6. Pre-processing of data before modelling:

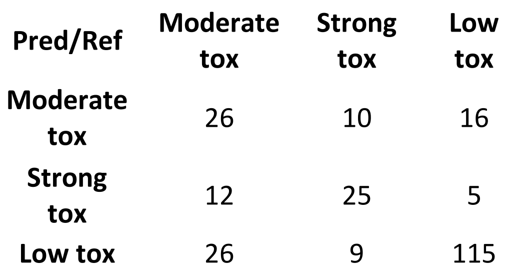
NA

6.7. Statistics for goodness-of-fit:

Statistics for the original model (Como et al. 2017): Training set, n= 192 Accuracy 0.88 Sensitivity 0.68 Specificity 0.93 MCC 0.59

Statistics for model implemented on VEGA: n = 256

Predicted compounds (LOO) = n 244, Not predicted = 12, Accuracy 0.68



Pred/Ref	Moderate tox	Strong tox	Low tox
Moderate tox	26	10	16
Strong tox	12	25	5
Low tox	26	9	115

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:

Yes

7.2. Available information for the external validation set:

CAS RN: Yes

Chemical Name: No

Smiles: Yes

Formula: Yes

INChI: yes

MOL file: Yes

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:

External validation has been done in the original model and it is not available for the implemented version in VEGA

7.6. Experimental design of test set:

NA

7.7. Predictivity - Statistics obtained by external validation:

Statistics for external validation are only available for the original model (Como et al. 2017):

Test set, n= 50 Accuracy 0.86 Sensitivity 0.75 Specificity 0.88 MCC 0.55

7.8. Predictivity - Assessment of the external validation set:

The predicitvity of the model is better when the compounds fall within the applicability domain of the model

7.9. Comments on the external validation of the model:

The use of the applicability domain index improves the robustness of the model

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] Predicting acute contact toxicity of pesticides in honeybees (Apis mellifera) through a k-nearestneighbor model. Como, F., Carnesecchi, E., Volani, S., Dorne, J. L., Richardson, J., Bassan, A., ... &Benfenati, E. (2017). Chemosphere, 166, 438-444

1770

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