	QMRF identifier (JRC Inventory):**Q17-24a-0011**
	QMRF Title:**VEGA BCF model (kNN/Read-Across)**
	Printing Date:**Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

VEGA BCF model (kNN/Read-Across)

1.2.Other related models:

1.3.Software coding the model:

BCF model (kNN/Read-Across) v.1.1.0

The model performs a read-across and provides a quantitative prediction of bioconcentration factor (BCF) in fish.

http://www.vega-qsar.eu/

2.General information

2.1.Date of QMRF:

01/06/2016

2.2.QMRF author(s) and contact details:

Emilio Benfenati IRCCS - Istituto di Ricerche Farmacologiche Mario Negri

emilio.benfenati@marionegri.it

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

Alberto Manganaro Kode srl info@kode-solutions.net www.kode-solutions. net

2.6.Date of model development and/or publication:

The model was developed on April 2015

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

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

[2]M. Floris, A. Manganaro, O. Nicolotti, R. Medda, G. F. Mangiatordi, E. Benfenati, “A generalizable definition of chemical similarity for read-across”, Journal of Cheminformatics (2014), vol. 6, 39

2.8.Availability of information about the model:

The model has been released open source and is freely available through the portal of the VEGA platform (www.vega-qsar.eu). The training and test set are available (see 9.3).

2.9.Availability of another QMRF for exactly the same model:

Other QMRF for this model are not available

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Fish

3.2.Endpoint:

2.Environmental fate parameters 2.4.a.Bioconcentration . BCF fish

3.3.Comment on endpoint:

The bioconcentration factor (BCF) is the concentration of test substance in the fish or specified tissues thereof divided by the concentration of the chemical in the surrounding medium at steady state.

3.4.Endpoint units:

Continuous value expressed in Log(L/kg)

3.5.Dependent variable:

3.6.Experimental protocol:

Bioconcentration factor (BCF) data are provided from tests that are conducted with respect to the OECD Test No. 305: Bioaccumulation in Fish. A procedure for characterising the bioconcentration potential of substances in fish.

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, including the original dataset of the CAESAR BCF model (note that experimental values may differ from the ones in the CAESAR BCF dataset, as this new dataset has been built including more sources). Data have been compared in case of multiple values and the mean value was calculated for the compounds with more than one value.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The read-across model has been built with the istKNN application (developed by Kode srl, www.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.

4.2.Explicit algorithm:

kNN

kNN prediction is based on the k most similar compounds retrieved with the similarity index developed in VEGA. Explanation of the kNN approach is available in A. Manganaro, F. Pizzo, A. Lombardo, A. Pogliaghi, E. Benfenati, “Predicting persistence in the sediment compartment with a new automatic software based on the k-Nearest Neighbor (k-NN) algorithm”, Chemosphere (2016), vol. 144, 1624-1630

The predicted value is calculated with the following algorithm:

1. The first *k* molecules with the closest similarity to the target compound are extracted.
2. Molecules with a similarity index lower than a given threshold S1 are excluded.
3. If no molecules are left, no prediction is provided (missing value).
4. If only one molecule is left, it is used as prediction only if its similarity value is equal to or higher than a given threshold S2, otherwise no prediction is provided (missing value).
5. In all other cases, the prediction is calculated as the

weighted average value of the k most similar compounds experimental values, where for each compounds the weight is given by its similarity value. The weights (similarity values) can be raised to the power of a given value E, called the enhance factor, as for integer larger than 1 the result is to enhance the role of molecules with higher similarity values in the prediction. Furthermore, a threshold for experimental values can be provided, so that the range of experimental values found in the chosen k most similar molecules is higher than the given threshold, no prediction is provided.

The k-NN model has the following settings:

K (neighbours number): 4S1 (Similarity threshold:) 0.7

S2 (Similarity threshold for single molecules): 0.75

E (Enhance factor): 3

Allowed experimental range: 3.5

4.3.Descriptors in the model:

Similarity index - Descriptors are only used to identify the similar compounds. Index for generic similarity as described in M. Floris, A. Manganaro, O. Nicolotti, R. Medda, G. F. Mangiatordi, E. Benfenati, “A generalizable definition of chemical similarity for read-across”, Journal of Cheminformatics (2014), vol. 6, 39

4.4.Descriptor selection:

No selection

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 section 4.3.

4.6.Software name and version for descriptor generation:

istKNN 0.9

in-house software for kNN modelling

alberto.manganaro@kode-solutions.net

http://chm.kode-solutions.net

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 of predictions is assessed using an Applicability Domain Index (ADI) that has values from 0 (worst case) to 1 (best case).

5.2.Method used to assess the applicability domain:

The ADI is calculated by grouping several other indices, each one taking into account a particular issue of the applicability domain. For each index, including the final ADI, two intervals for its values are defined, such that the first interval corresponds to a positive evaluation, and the second one corresponds to a negative evaluation.

5.3.Software name and version for applicability domain assessment:

BCF model (kNN/Read-Across) v.1.1.0

This is included in the stand alone version of VEGA: VEGANic v 1.1.1

http://www.vega-qsar.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: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: 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 model performs a read-across on a dataset of 860 chemicals. This dataset has been made by Istituto di Ricerche Farmacologiche Mario Negri, merging experimental data from several reliable sources, including the original dataset of the CAESAR BCF model (note that experimental values may differ from the ones in the CAESAR BCF dataset, as this new dataset has been built including more sources).

6.6.Pre-processing of data before modelling:

No

6.7.Statistics for goodness-of-fit:

Training set: n = 836; R2 = 0.67; RMSE = 0.76

Non predicted compounds: n = 24

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

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

6.10.Robustness - Statistics obtained by Y-scrambling:

6.11.Robustness - Statistics obtained by bootstrap:

6.12.Robustness - Statistics obtained by other methods:

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: No

Chemical Name: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

7.3.Data for each descriptor variable for the external validation set:

No

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

No

7.5.Other information about the external validation set:

-

7.6.Experimental design of test set:

Not applicable

7.7.Predictivity - Statistics obtained by external validation:

External validation set: n = 148; R2 = 0.491; RMSE = 0.783 Data with ADI > 0.85 : n = 95; R2 = 0.778; RMSE = 0.471 M. I.Petoumenou, F. Pizzo, J. Cester, A. Fernández, E. Benfenati, "Comparison between bioconcentration factor(BCF) data provided by industry to the European Chemicals Agency (ECHA) and data derived from QSAR models", Environmental Research 142(2015) pages: 529–534.

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 robusteness of the model.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

No: the model adopts the read-across approach based on chemical similarity.

8.2.A priori or a posteriori mechanistic interpretation:

-

8.3.Other information about the mechanistic interpretation:

-

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

M. I.Petoumenou, F. Pizzo, J. Cester, A. Fernández, E. Benfenati, "Comparison between bioconcentration factor(BCF) data provided by industry to the European Chemicals Agency (ECHA) and data derived from QSAR models", Environmental Research 142(2015) pages: 529–534.

9.3.Supporting information:

dataset_BCF_KNN.txt	http://qsardb.jrc.ec.europa.eu/qmrffile:///C:\Users\MPetoumenou\Desktop\dataset_BCF_KNN.txt

Test set(s)Supporting information

10.Summary (JRC Inventory)

10.1.QMRF number:

Q17-24a-0011

10.2.Publication date:

2017-09-21

10.3.Keywords:

To be entered by JRC;

10.4.Comments:

To be entered by JRC