	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **BCF model (Arnot-Gobas) - v. 1.0.1**
	Printing Date: **February 3, 2022**

1.QSAR identifier

1.1.QSAR identifier (title):

BCF model (Arnot-Gobas) - v. 1.0.1

1.2.Other related models:

This model is an implementation of Arnot-Gobas BAF-BCF model of EPISUITE.

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:

16-11-2020

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/

[3] 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):

No update

2.4.QMRF update(s):

No update

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/

2.6.Date of model development and/or publication:

2003

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

[1] 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] Arnot, Jon & Gobas, Frank. (2003). A Generic QSAR for Assessing the Bioaccumulation Potential of Organic Chemicals in Aquatic Food Webs. QSAR & Combinatorial Science. 22. 337 - 345. 10.1002/qsar.200390023.

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:

Rainbow trout (Oncorhynchus mykiss)

3.2.Endpoint:

QMRF 2. E FATE parameters - QMRF 2. 4.a Bioconcentration.BCF fish

3.3.Comment on endpoint:

Bioconcentration is the process where the chemical concentration in an aquatic organism achieves a level that exceeds that in the water as a result of the exposure of an organism to a chemical in the water but does not include exposure via the diet. Bioconcentration refers to a situation, typically derived under controlled laboratory conditions, wherein the chemical is absorbed from the water via the respiratory surface (e.g. gills) and/or the skin only. Standard protocols for conducting bioconcentration tests have been developed. The extent of chemical bioconcentration is usually expressed in the form of a bioconcentration factor (BCF), which is the ratio of the chemical concentration in the organism (CB) and the water (CW) BCF=CB/CW at steady-state (i.e when the chemical concentration in the organism does not change in respect to time)

3.4.Endpoint units:

Bioconcentration = $L*Kg^{-1}$

3.5.Dependent variable:

BCF log(L/kg)

3.6.Experimental protocol:

Bioconcentration: Flow-through Fish Test, 305, OECD Guidelines for Testing Chemicals, 1996 [3]

3.7.Endpoint data quality and variability:

See [2] in section 2.7

episuite

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Regression model

4.2.Explicit algorithm:

$BCF = (1 - L_B) * (K_1 \phi/(K_2 + K_e + K_g + K_m))$

4.3.Descriptors in the model:

L_B Lipid content of organism

K_1 Uptake rate constant $1/((0.01 +1/K_{OW})*W^{0.4})$

K_2 Elimination rate constant K_1/L_B*K_{OW}

K_E Fecal egestion rate constant $0.125 * K_D$

K_G Growth rate constant $0.0005 * W^{-0.2}$

K_M Metabolic transformation rate constant (0 day^{-1} (default))

ϕ: For non-ionizing hydrophobic organic substances, the fraction of freely dissolved chemical in the water can be estimated from the concentrations of particulate and dissolved organic carbon as: $\phi = 1/(1+\chi_{DOC}* 0.35 * K_{OW} + \chi_{POC} * 0.1 * 0.35 * K_{OW})$

where:

$\chi_{:POC}$ Concentration of particulate organic carbon 5 * 10-7 g/ml

χ_{DOC} Concentration of dissolved organic carbon 5 * 10-7 g/ml

4.4. Descriptor selection:

A standard descriptor selection wasn't performed: the descriptors reported in the section 4.3 were selected according to biological mechanism.

.

4.5. Algorithm and descriptor generation:

See [2] in section 2.7

4.6. Software name and version for descriptor generation:

See [2] in section 2.7

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 ≥ index > 0.75, the predicted substance could be out of the Applicability Domain of the model. It corresponds to "moderate reliability" of prediction.

If AD index ≤ 0.75, the predicted substance is regarded out of the Applicability Domain of the model. It corresponds to "low reliability" of prediction.

No ADI threshold was used to provide performance calculations of the validation set

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 > 0.75, only moderately similar compounds with known experimental value in the training set have been found

If index ≤ 0.75, 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 < 0.5, accuracy of prediction for similar molecules found in the training set is good

If 1 > index ≥ 0.5, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≥ 1, 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 < 0.5, molecules found in the training set have experimental values that agree with the target compound predicted value

If 1 > index ≥ 0.5, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index ≥ 1, 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 < 0.5, the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability

If 1 > index ≥ 05, the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

If index ≥ 1, the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability

Reliability of logP prediction.

This model strongly relies on logP. This parameter verifies the reliability of the logP prediction. The index takes into account the maximum error in prediction between 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:

Index = 1, reliability of logP value used by the model is good

Index = 0.7, reliability of logP value used by the model is not optimal

Index = 0, reliability of logP value used by the model is not adequate

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 (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:

To include the metabolism in fish in the bioaccumulation assessment, the models developed by Arnot and Gobas (2003) was reimplemented and made available through the BCF BAF module in the EPISuite v4.1

platform (US EPA, 2019). The BCF equation for the upper trophic levels was reimplemented in the model. For the implementation, a dataset of 692 compounds, retrieved from EPISuite platform, has been used, processed and cleaned. We excluded from the initial dataset metal complexes, inorganics, mixtures of structural isomers, ambiguous structures, non-ionic surfactant mixtures, complex disconnected structures (e.g. polymers), chemicals whose correspondence name-CAS was not found, UVCB.

We generated the SMILES structures from the chemical name and CAS RN for each substance

6.6.Pre-processing of data before modelling:

NA

6.7.Statistics for goodness-of-fit:

N = 692, R^2 = 0.65; RMSE = 0.83

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:

NA

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] Arnot, J.A., Gobas, F.A.P.C. (2003): A Generic QSAR for Assessing the Bioaccumulation Potential of Organic Chemicals in Aquatic Food Webs. QSAR & Combinatorial Science 22, 337-345.

[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.

[3] OECD (1996), Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure, OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris
https://www.oecd.org/chemicalsafety/testing/section3-degradation-and-accumulation-replaced-and-cancelled-test-guidelines

https://www.oecd.org/env/ehs/testing/E305_Fish%20Bioaccumulation.pdf

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

9.3.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