

	QMRF identifier (JRC Inventory): To be entered by JRC
	QMRF Title: Fish Bioconcentration Factor (BCF) MLR
	Printing Date: 3-ott-2022

1. QSAR identifier

1.1. QSAR identifier (title):

Fish_Bioconcentration_Factor_(BCF)_MLR

1.2. Other related models:

Lunghini, F.; Marcou, G.; Azam, P.; Patoux, R.; Enrici, M.H.; Bonachera, F.; Horvath, D.; Varnek, A.

A QSPR Models for Bioconcentration Factor (BCF): Are They Able to Predict Data of Industrial Interest? SAR QSAR Environ. Res. 2019, 30, 507–524, doi:<https://doi.org/10.1080/1062936X.2019.1626278>.

1.3. Software coding the model:

QSAR-ME Profiler

Software for QSAR models predictions

nicola.chirico@uninsubria.it; ester.papa@uninsubria.it

<http://dunant.dista.uninsubria.it/qsar/>

2. General information

2.1. Date of QMRF:

29/09/2022

2.2. QMRF author(s) and contact details:

Linda Bertato University of Insubria Linda Bertato; Ester Papa l.bertato@uninsubria.it; ester.papa@uninsubria.it <http://dunant.dista.uninsubria.it/qsar/>

2.3. Date of QMRF update(s):

-

2.4. QMRF update(s):

-

2.5. Model developer(s) and contact details:

Linda Bertato; Ester Papa University of Insubria Linda Bertato; Ester Papa l.bertato@uninsubria.it; ester.papa@uninsubria.it <http://dunant.dista.uninsubria.it/qsar/>

2.6. Date of model development and/or publication:

Date of publication: 30 September 2022

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

[1]R (version 4.0.2) <https://cran.r-project.org/>

[2]PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints v 2.21 <http://www.yapcsoft.com/dd/padeldescriptor/>

[3]Predicting the Bioconcentration Factor in Fish from Molecular Structures <https://doi.org/10.3390/toxics10100581>

[4][5]QSPR Models for Bioconcentration Factor (BCF): Are They Able to Predict Data of Industrial Interest? <https://doi.org/10.1080/1062936X.2019.1626278>.

2.8. Availability of information about the model:

Model developed as output of a PhD Program in Chemical and Environmental Sciences (DiSCA, University of Insubria) PhD scholarship

to Linda Bertato and of a post-doc grant to Dr. Nicola Chirico.

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

no

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Fish

3.2.Endpoint:

QMRF 2. Environmental fate parameters QMRF 2. 4.a. Bioconcentration . BCF fish

3.3.Comment on endpoint:

This QSAR has been developed to model the bioconcentration factor (BCF) in fish. Bioconcentration describes the process by which a chemical substance is absorbed by an organism from the environment only through its respiratory and dermal surfaces, i.e., chemical exposure in the diet is not included. The BCF can be calculated as the ratio of the chemical concentration in the organism and the chemical concentration in the water at steady state, i.e., $BCF = CB/CWD$.

3.4.Endpoint units:

L/kg bdwt

3.5.Dependent variable:

Log_BCF

3.6.Experimental protocol:

Reference Protocol OECD 305 - Bioaccumulation in Fish: Aqueous and Dietary Exposure

3.7.Endpoint data quality and variability:

Experimental values for Log BCF (L/kg bdwt) were taken from the raw dataset published by Lunghini et al. in 2019 which were collected from multiple sources including public-available databases and literature research.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Multiple Linear Regression by means of Ordinary Least Squares

4.2.Explicit algorithm:

MLR - QSAR model

Multiple Linear Regression by means of Ordinary Least Squares

$$\text{Log BCF} = -1.44 (\pm 0.62)^{***} + 0.80 (\pm 0.09)^{***} \text{MWC4} + 0.24 (\pm 0.04)^{***} \text{SubFPC171} - 0.10 (\pm 0.02)^{***} \text{SubFPC295} - 1.19 (\pm 0.19)^{***} \text{maxHBd} - 0.06 (\pm 0.01)^{***} \text{maxdO} - 0.51 (\pm 0.21)^{***} \text{IC0}$$

Significance (P values): ^{***}, 0.001; ^{**}, 0.01; ^{*}, 0.0

4.3.Descriptors in the model:

- [1]MWC4 Molecular walk count of order 4 ($\ln(1+x)$)
- [2]SubFPC171 Counts of Arylchloride [Cl][c]
- [3]SubFPC295 Counts of: C ONS bond
- [4]maxHBd Maximum E-States for (strong) Hydrogen Bond donors
- [5]maxdO Maximum atom-type E-State: =O
- [6]IC0 Information content index (neighborhood symmetry of 0-order)

4.4.Descriptor selection:

An input file including more than 7000 molecular descriptors of different types (0D, 1D, 2D) were calculated in PaDEL-Descriptor v. 2.21. Constant and nearly constant descriptors as well as descriptors found to be correlated pairwise more than 80% and 95 % respectively were excluded in a pre-reduction step prior to modelling. The remaining molecular descriptors were then used as input for the Variable Subset Selection (VSS) procedure which led to a population of models up to 10 variables using the step-up procedure, selected by choosing R^2 as the fitness function, for a total of 500 models.

4.5.Algorithm and descriptor generation:

Molecular descriptors were calculated using the software Padel-Descriptor v. 2.21 using canonicalized SMILES as input. SMILES were canonicalized using the software OpenBabel v. 2.3.2.

4.6.Software name and version for descriptor generation:

Padel Descriptor v. 2.21
Software to Calculate Molecular Descriptors and Fingerprints
-
<http://www.yapcsoft.com/dd/padeldescriptor/>

Open Babel v. 2.3.2
Open Babel: The Open Source Chemistry Toolbox
-
<http://openbabel.org>

4.7.Chemicals/Descriptors ratio:

150

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Statistical AD:

The applicability domain of the model was verified by the leverage approach and fixed thresholds has been used to define both structural and response outliers (see section 5.4). The plot of leverages (HAT diagonals) versus standardised residuals, i.e. the Williams plot, verified the presence of response outliers (i.e. compounds with cross-validated standardized residuals greater than 2.5 standard deviation units) and chemicals structurally very influential in determining the model's coefficients (i.e. compounds with a leverage

value (h) greater than $3p'/n$ (h^*), where p' is the number of model variables plus one, and n is the number of the objects used to calculate the model). For new compounds without experimental data, leverage can be used as a quantitative measure for evaluating the degree of extrapolation: for compounds with a high leverage value ($h > h^*$), which are structural outliers, predictions should be considered less reliable.

Mechanistic AD: The applicability domain of the model is related to the most probable site of reaction and the related reactivity, identified by the Toxtree module SMARTCyp.

5.2. Method used to assess the applicability domain:

The structural applicability domain of the model was assessed by the leverage approach, on the bases of a cut-off hat value $h^*=0.023$. HAT values for each compound are calculated as the diagonal elements of the HAT matrix ($H = X(X^T X)^{-1} X^T$).

The response applicability domain can be verified by the standardized residuals (cut off values 2.5 standard units).

5.3. Software name and version for applicability domain assessment:

R (version 4.0.2)

MLR-OLS models were generated using in-house developed R scripts

nicola.chirico@uninsubria.it

<https://cran.r-project.org/>

5.4. Limits of applicability:

HAT i/i ($h^*=0.023$)

6. Internal validation - OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Available information for the training set:

CAS RN: No

Chemical Name: No

Smiles: Yes

Formula: No

INChI: No

MOL file: 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:

920 chemicals were included in the training set.

6.6. Pre-processing of data before modelling:

The endpoint was already log transformed prior to modelling.

6.7. Statistics for goodness-of-fit:

R^2 : 0.62 RMSE_{TR}:0.80

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

Q^2_{LOO} : 0.61

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

Q^2_{LMO} : 0.615-fold RMSE_{cv}: 0.81

6.10. Robustness - Statistics obtained by Y-scrambling:

R^2_{YSCR} : 0.02

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:

All

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

All

7.5. Other information about the external validation set:

To verify the predictive capability of the proposed model, the dataset was split, before model development, into a training set used for model development and a prediction set used for external validation.

7.6. Experimental design of test set:

Data were ordered according to increasing experimental response and one every two chemicals were put in the prediction set, keeping the first and the last chemicals in the training set. (459 chemicals in the predictions set).

7.7. Predictivity - Statistics obtained by external validation:

RMSE: 0.78 R^2 : 0.64

7.8. Predictivity - Assessment of the external validation set:

Training and prediction sets are balanced according to both structure and the response.

7.9. Comments on the external validation of the model:

The full model, calibrated on the complete dataset (thus ensuring a wider applicability domain), is implemented in the software QSAR-ME Profiler for predictive purposes. The model equation used for the external validation (reported also in section 4.2) and the statistics are the following: $\text{Log BCF} = -1.44 (\pm 0.62)^{***} + 0.80 (\pm 0.09)^{***} \text{MWC4} + 0.24 (\pm 0.04)^{***} \text{SubFPC171} - 0.10 (\pm 0.02)^{***} \text{SubFPC295} - 1.19 (\pm 0.19)^{***} \text{maxHBd} - 0.06 (\pm 0.01)^{***} \text{maxdO} - 0.51 (\pm 0.21)^{***} \text{IC0}$ Significance (P values): ***, 0.001; **, 0.01; *, 0.0

Domain of applicability: $h^* = 0.023$

8. Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

The model was developed by statistical selection of the molecular descriptors. The interpretation of these descriptors, listed in section 4.3, is provided *a posteriori* and described in details in the published article.

8.2. A priori or a posteriori mechanistic interpretation:

a priori and *a posteriori*.

8.3. Other information about the mechanistic interpretation:

9. Miscellaneous information

9.1. Comments:

9.2. Bibliography:

Predicting the Bioconcentration Factor in Fish from Molecular Structures
<https://doi.org/10.3390/toxics10100581>

9.3. Supporting information:

Training set(s) Test set(s) Supporting information

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