

	<b>QMRF identifier (JRC Inventory): To be entered by JRC</b>
	<b>QMRF Title: Fish Bioconcentration Factor (BCF) classification</b>
	<b>Printing Date: 3-ott-2022</b>

## 1. QSAR identifier

### 1.1. QSAR identifier (title):

Fish\_Bioconcentration\_Factor\_(BCF)\_classification

### 1.2. Other related models:

A QSPR Models for Bioconcentration Factor (BCF): Are They Able to Predict Data of Industrial Interest?

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

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](mailto:nicola.chirico@uninsubria.it); [ester.papa@uninsubria.it](mailto: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@uninsubria.it](mailto:ester.papa@uninsubria.it); [l.bertato@uninsubria.it](mailto:l.bertato@uninsubria.it);

[ester.papa@uninsubria.it](mailto: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](mailto:l.bertato@uninsubria.it);

[ester.papa@uninsubria.it](mailto: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]QSAR-Co v.1.1.0 <https://sites.google.com/view/qsar-co>

[2]QSAR-Co: An Open Source Software for Developing Robust Multitasking or Multitarget Classification-Based QSAR Models <https://doi.org/10.1021/acs.jcim.9b00295>

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

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

[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:**

The model is non-proprietary and training and prediction sets are available.

## **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.,  $BCF = CB/CWD$ . According to the regulatory cut off value used to discretize the log BCF values into two a priori classes was 2000 (i.e.  $\text{Log}_{10} 2000 = 3.30$ ).

## **3.4.Endpoint units:**

L/kg bdwt

## **3.5.Dependent variable:**

Class B = bioaccumulable ( $BCF > 2000$ )

Class not-B = not bioaccumulable ( $BCF < 2000$ )

## **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. In addition, correspondence between SMILES and CAS, and their correctness were checked and data corresponding to wrong or uncertain SMILES were excluded. Multiple data were averaged.

# **4.Defining the algorithm - OECD Principle 2**

## **4.1.Type of model:**

Linear Discriminant Analysis (LDA)

## **4.2.Explicit algorithm:**

LDA-QSAR model

Linear Discriminant Analysis

LDA Linear Discriminant Analysis Linear Score Function, re-calculated in R software: class B =  $-25.33 + 13.63 \times \text{IC}_2 + 69.82 \times \text{MWC}_4 - 16.50 \times \text{TopoPSA} - 16.58 \times \text{MAXDP} + \log(0.55)$  class notB =  $-18.33 + 20.79 \times \text{IC}_2 + 50.84 \times \text{MWC}_4 - 7.26 \times \text{TopoPSA} - 11.37 \times \text{MAXDP} + \log(0.45)$

#### 4.3.Descriptors in the model:

- [1]IC2 Information content index (neighborhood symmetry of 2-order)
- [2]TopoPSA Topological polar surface area
- [3]MAXDP Maximum positive intrinsic state difference in the molecule (related to the electrophilicity of the molecule). Using  $\Delta V = (Z_v - \max \text{BondedHydrogens}) / (\text{atomicNumber} - Z_v - 1)$ . Gramatica, P., Corradi, M., and Consonni, V. (2000). Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors.
- [4]MWC4 Molecular walk count of order 4 ( $\ln(1+x)$ )

#### 4.4.Descriptor selection:

Descriptors constant or nearly constant for more than 80% of the values, as well as descriptors with a pairwise correlation greater than 95% were excluded in a pre-reduction step prior to modelling. About 400 descriptors were then used as input for the Variable Subset Selection (VSS) procedure by a Genetic Algorithm (GA) used to generate the classification 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.yapcwsoft.com/dd/padeldescriptor/>

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

#### 4.7.Chemicals/Descriptors ratio:

73

### 5.Defining the applicability domain - OECD Principle 3

#### 5.1.Description of the applicability domain of the model:

The dataset included Log BCF values for heterogeneous chemicals of environmental and toxicological interest aromatic organohalogen compounds (e.g., PAH, PCBs, dioxins and furans) as well as some perfluorinated compounds. Few molecules fall outside the applicability domain of the QSAR according to the method used to assess it.

#### 5.2.Method used to assess the applicability domain:

Structural applicability domain of the best combination of modelling variables was calculated for LDA using the Confidence Estimation (CE) approach (threshold = 0.5) and the standardization approach available in QSAR-Co.

#### 5.3.Software name and version for applicability domain assessment:

-

<https://sites.google.com/view/qsar-co>

#### **5.4.Limits of applicability:**

The applicability domain calculated for the LDA model show that a few chemicals fall outside the applicability domain of the here proposed QSAR.

### **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

#### **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:**

292 chemicals were included in the training set.

#### **6.6.Pre-processing of data before modelling:**

The endpoint was already log transformed prior to modelling. Chemicals were classified using the threshold proposed in the literature of log BCF>3.3.

#### **6.7.Statistics for goodness-of-fit:**

Accuracy (in %; train): 86.9863; Precision (in %; train): 88.2716;

Sensitivity (in %; train): 88.2716; Specificity (in %; train): 85.3846;

F-measure (train): 0.8827; MCC (train): 0.7366. True Positive(train): 143;

False Positive(train): 19; True Negative(train): 111; False

Negative(train): 19

#### **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:**

Area Under the Curve cv test : 0.89

#### **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: Yes

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:**

Chemicals were split using the euclidean distance and keeping 70% of the compounds in the training set. This partitioning led to a training set composed of 292 chemicals (i.e. 162 B and 130 not B) and a prediction set composed of 125 substances (i.e. 63 B and 62 not B).

### **7.7.Predictivity - Statistics obtained by external validation:**

Accuracy (in %; test): 84 Precision (in %; test): 80.2817 Sensitivity (in %; test): 90.4762 Specificity (in %; test): 77.4194 F-measure (test): 0.8507 MCC (test): 0.6853 True Positive(test): 57 False Positive(test): 14 True Negative(test): 48 False Negative(test): 6

### **7.8.Predictivity - Assessment of the external validation set:**

The splitting performed in the software QSAR-Co allowed for the selection of meaningful training sets and representative prediction sets on the basis of euclidean distance taking into account structural similarity.

### **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 QSARINS-Chem for predictive purposes. The model equation used for the external validation (reported also in section 4.2) and the statistics are the following: class B =  $-25.33 + 13.63 \times \text{IC2} + 69.82 \times \text{MWC4} - 16.50 \times \text{TopoPSA} - 16.58 \times \text{MAXDP} + \log(0.55)$  class notB =  $-18.33 + 20.79 \times \text{IC2} + 50.84 \times \text{MWC4} - 7.26 \times \text{TopoPSA} - 11.37 \times \text{MAXDP} + \log(0.45)$

## **8.Providing a mechanistic interpretation - OECD Principle 5**

### **8.1.Mechanistic basis of the model:**

The model was developed using a statistical approach of selection of the molecular descriptors with no mechanistic assumption.

## 8.2.A priori or a posteriori mechanistic interpretation:

*A posteriori.*

the descriptors of the model are: IC2, i.e., the information content index for neighborhood symmetry of second order; TopoPSA, i.e., the topological polar surface area and MAXDP maximum positive intrinsic state difference in the molecule (related to the electrophilicity of the molecule), and MWC4, a molecular walk count representing self-returning counts at length four within the molecule.

## 8.3.Other information about the mechanistic interpretation:

-

## 9.Miscellaneous information

### 9.1.Comments:

-

### 9.2.Bibliography:

- [1]Predicting the Bioconcentration Factor in Fish from Molecular Structures  
<https://doi.org/10.3390/toxics10100581>
- [2]QSAR-Co v.1.1.0 <https://sites.google.com/view/qsar-co>
- [3]QSAR-Co: An Open Source Software for Developing Robust Multitasking or Multitarget Classification-Based QSAR Models <https://doi.org/10.1021/acs.jcim.9b00295>
- [4]QSPR Models for Bioconcentration Factor (BCF): Are They Able to Predict Data of Industrial Interest? <https://doi.org/10.1080/1062936X.2019.1626278>.

### 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