

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

## 1. QSAR identifier

### 1.1. QSAR identifier (title):

Fish\_Dietary\_Biomagnification\_Factor

### 1.2. Other related models:

Acceptable-by-Design QSARs to Predict the Dietary Biomagnification of Organic Chemicals in Fish, Grisoni et al. 2018;

Detecting the bioaccumulation patterns of chemicals through datadriven approaches, Grisoni et al. 2018

### 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:

16/06/2022

### 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: May 2022

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

[1] Classification-based QSARs for predicting dietary biomagnification in fish doi: 10.1080/1062936X.2022.2066174

[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] QSAR-Co v.1.1.0 <https://sites.google.com/view/qsar-co>

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

### 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 (19 species, primarily trout and carp)

#### 3.2. Endpoint:

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

#### 3.3. Comment on endpoint:

Dietary Biomagnification Factor in Fish

#### 3.4. Endpoint units:

kg lipid/kg lipid

#### 3.5. Dependent variable:

Class BM = biomagnifiable (BMF>1)

Class not-BM = not biomagnifiable (BMF<1)

#### 3.6. Experimental protocol:

OECD 305-III: Dietary Exposure Bioaccumulation Fish Test

#### 3.7. Endpoint data quality and variability:

Data were initially curated by Arnot and Quinn 2015 according to the compliance with the OECD TG 305. In addition, BMF data associated to conflicting structures were excluded as well as multiple records available for stereoisomers. Standard deviation and Peirce's method were used to evaluate suspect records.

### 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:  $BM = -36.84 + 25.56 \text{ BCUTw-1I} + 13.49 \text{ SpMin3\_Bhp} + 4.50$

$\text{PubchemFP738} + 60.44 \text{ SpMin2\_Bhi} - 29.79 \text{ GGI7}$

$\text{not-BM} = -26.30 + 18.40 \text{ BCUTw-1I} + 24.66 \text{ SpMin3\_Bhp} + 2.10 \text{ PubchemFP738} + 45.25 \text{ SpMin2\_Bhi} - 15.17 \text{ GGI7}$

#### 4.3. Descriptors in the model:

[1]BCUTw-1I highest lowest atom weighted BCUTS

[2]SpMin3\_Bhp Smallest absolute eigenvalue of Burden modified matrix - n 3 / weighted by relative polarizabilities

[3]PubchemFP738 Cc1cc(Cl)ccc1; These bits test for the presence of complex SMARTS patterns, regardless of count, but where bond orders and bond aromaticity are specific.

[4]SpMin2\_Bhi Smallest absolute eigenvalue of Burden modified matrix - n 2 / weighted by relative first ionization potential

[5]GGI7 Topological charge index of order 7

#### 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.

The remaining 535 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:**

44.8

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

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

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. In addition, AD was graphically inspected by Principal Component Analysis (PCA) performed on the best molecular descriptors selected by GA-LDA. Further analysis was performed evaluating ranges of the descriptors calculated for the compounds in the training and in the prediction set.

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

QSAR-Co v.1.1.0

A software for Developing Robust Multitasking or Multitarget Classification-Based QSAR Models

-

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

Minitab Statistical Software 2022

Software for statistical analysis

#### 5.4.Limits of applicability:

The applicability domain calculated for the LDA model and by PCA show that a few chemicals (mainly siloxanes) 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:

224 chemicals were included in the training set.

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

The endpoint was log transformed prior to modelling and chemicals were classified using the threshold proposed in the literature of  $\log \text{BMF} > 0$ .

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

Accuracy (in %; train): 94.196

Precision (in %; train): 91.089

Sensitivity (in %; train): 95.833

Specificity (in %; train): 92.969

F-measure (train): 0.934

MCC (train): 0.883

AUC: 0.95

True Positive(train): 92

False Positive(train): 9

True Negative(train): 119

False Negative(train): 4

#### 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:

AUCcv: 0.87

## 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:

Euclidean Distance (70% of compounds in training) was used to split the dataset and create the prediction set. The automatic procedure was available in QSAR-Co. (96 chemicals in the test set).

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

Accuracy (in %; test): 86.458

Precision (in %; test): 87.931

Sensitivity (in %; test): 89.474

Specificity (in %; test): 82.051

F-measure (test): 0.887

MCC (test): 0.718

True Positive(test): 51

False Positive(test): 7

True Negative(test): 32

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 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:  $BM = -36.84 + 25.56 \text{ BCUTw-1l} + 13.49 \text{ SpMin3\_Bhp} + 4.50 \text{ PubchemFP738} + 60.44$

SpMin2\_Bhi - 29.79 GGI7 not-BM = -26.30 + 18.40 BCUTw-1l + 24.66  
SpMin3\_Bhp + 2.10 PubchemFP738 + 45.25 SpMin2\_Bhi - 15.17 GGI7

## 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:

BCUTw-1l is a BCUT descriptor weighted over atom weights, SpMin2\_Bhi represents the smallest absolute eigenvalue of Burden modified matrix - n<sup>2</sup> / weighted by relative first ionization potential, SpMin3\_Bhp is the smallest absolute eigenvalue of Burden modified matrix - n<sup>3</sup> / weighted by relative polarizabilities, GGI7 describes the topological charge index of order 7 and the fragment PubchemFP738 counts the presence of the fragment Cc1cc(Cl)ccc1 (chlorophenyl group) within the molecular structure.

### 8.3. Other information about the mechanistic interpretation:

-

## 9. Miscellaneous information

### 9.1. Comments:

-

### 9.2. Bibliography:

- [1] Classification-based QSARs for predicting dietary biomagnification in fish  
<https://doi.org/10.1080/1062936X.2022.2066174>
- [2] Detecting the bioaccumulation patterns of chemicals through datadriven approaches  
<https://doi.org/10.1016/j.chemosphere.2018.05.157>
- [3] Acceptable-by-Design QSARs to Predict the Dietary Biomagnification of Organic Chemicals in Fish DOI: 10.1002/ieam.4106
- [4] Development and evaluation of a database of dietary bioaccumulation test data for organic chemicals in fish <https://doi.org/10.1021/es506251q>

### 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