	QMRF identifier (JRC Inventory):**Q13-24a-0012**
	QMRF Title:**QSAR for bioconcentration factor in fish**
	Printing Date:**Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

QSAR for bioconcentration factor in fish

1.2.Other related models:

Gramatica P & Papa E (2003). QSAR Modeling of Bioconcentration Factor by theoretical molecular descriptors. QSAR & Combinatorial Science, 22, 374-385.

1.3.Software coding the model:

DRAGON

Software for molecular descriptors calculations - 2001, version 5.2 for Windows

Roberto Todeschini, Talete srl, Milan, Italy

2.General information

2.1.Date of QMRF:

29/05/2009

2.2.QMRF author(s) and contact details:

[1]Paola Gramatica University of Insubria paola.gramatica@uninsubria.it

[2]Ester Papa University of Insubria ester.papa@uninsubria.it

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1]Ester Papa University of Insubria ester.papa@uninsubria.it

[2]Paola Gramatica University of Insubria paola.gramatica@uninsubria.it

2.6.Date of model development and/or publication:

2005

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

Gramatica P & Papa E (2005). An Update of the BCF QSAR model based on theoretical molecular descriprors. QSAR & Combinatorial Science 24, 953-960.

2.8.Availability of information about the model:

This model is not proprietary. Training and test sets are available

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

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:

Bioconcentration Factor (BCF) data were taken from studies where it appeared that steady states of bioconcentration had been reached. Both equilibrium and kinetic methods were included. Most of the collected BCF

data were obtained under flow conditions, while a few were determined from semi-static conditions. Test species related to several fish such as rainbow trout, guppies, fathead minnows, bluegill sunfish, golden ide, etc. There were differences in the lipid content and weight of the fish and exposure concentration and duration. The BCFs used in this study were measured basing on the total weight of fish.

3.4.Endpoint units:

Dimensionless

3.5.Dependent variable:

Logarithm of the Bioconcentration Factor (logBCF)

3.6.Experimental protocol:

3.7.Endpoint data quality and variability:

Experimentally determined BCF values for 238 non-ionic organic compounds were collected from an extensive literature review.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR

4.2.Explicit algorithm:

multilinear regression QSAR

LogBCF= -1.01 (?0.35) + 2.53 (?0.13)$^{v}I^{M}_{D,Deg}$
- 1.02 (?0.12)HIC - 0.46 (?0.04)nHAcc - 1.13 (?0.18)GATS1e - 1.70 (?0.34)MATS1p

4.3.Descriptors in the model:

[1]IDDM 2D descriptor, the mean information content of the distance degree magnitude of Bonchev
[2]nHAcc number of atom acceptors in the hydrogen bonds
[3]MATS1p Moran 2D-autocorrelation
[4]GATS1e Geary 2D-autocorrelation
[5]HIC 3D GETAWAY - mean information content on leverage magnitude

4.4.Descriptor selection:

A total of 1202 molecular descriptors of different kinds were calculated to describe the chemical diversity of the compounds.When there was more than 97% pair-wise correlation, one variable was deleted. The Genetic Algorithm (GA) was applied to a final set of 512 descriptors for variable selection

4.5.Algorithm and descriptor generation:

Multiple linear regression (OLS) was applied to generate the QSAR model. Descriptors were generated by DRAGON software. The input files for descriptor calculation contain information on atom and bond types, connectivity, partial charges and atomic spatial coordinates, relative to the minimum energy conformation of the molecule, and were obtained by the Molecular Mechanics method of Allinger (MM+) using the package HYPERCHEM.

4.6.Software name and version for descriptor generation:

DRAGON

Software for molecular descriptors calculations - 2001, version 5.2 for Windows

Roberto Todeschini, Talete srl, Milan, Italy

4.7.Chemicals/Descriptors ratio:

179 chemicals / 5 descriptors = 35.8

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

High leverage compounds (training set): tetrachloromethane, 2-methyl-4,6-dinitrophenol, hexachloroethane, tetrachloroethylene, ethyl acetate

Structural Outliers: octachlorodibenzofuran.

No chemicals in the validation set were identified as being outside the structural-response domain of the model.

The AD of the model covers different chemical classes: halogenated benzenes, chlorinated biphenyls, chlorinated naphthalenes, chlorinated aliphatic hydrocarbons, monocyclic aromatic hydrocarbons, polycyclic aromatic hydrocarbons, chlorinated dibenzo-dioxins, chlorinated dibenzofurans, phenols, anilines, phthalates, carbonates, phosphates, esters, ethers and nitroaromatics.

5.2.Method used to assess the applicability domain:

The structural AD of the model was checked by leverage approach (chemicals structurally very influential in determining model parameters have leverage value (hat) greater than 3p'/n (h*), where p' is the number of the model variables plus one, and n is the number of the objects used to calculate the model). Outliers for the response were identified by analysis of the standardised residuals (values higher than 2.5 standard deviation units).

5.3.Software name and version for applicability domain assessment:

MOBY-DIGS MOdels by Descriptors In Genetic Selection - ver. 1 - beta for windows, Talete srl, Milan, Italy.

Calculation of the hat values (leverages)

Roberto Todeschini

Excel

Calculation of standardised residuals

5.4.Limits of applicability:

High leverage compounds: hat value > 0.1

Outliers for the response: standardised residuals > 2.5 standard deviation units

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: No

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:

238 compounds

6.6.Pre-processing of data before modelling:

BCF data were transformed into logarithmic units (logBCF)

6.7.Statistics for goodness-of-fit:

$n_{training}$=179, R^2=80.7%, s=0.57, F_{173}=146.9,
SDEP=0.58, SDEC=0.56,

RMSE (training. set)=0.56 RMSE (cross-val. set)=0.58

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

Q^2_{LOO}=79.6%,

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

$Q^2_{LMO25\%}$=79%,$Q^2_{LMO50\%}$=78.2%,

6.10.Robustness - Statistics obtained by Y-scrambling:

R^2_{yscr} = 0.03

6.11.Robustness - Statistics obtained by bootstrap:

Q^2_{BOOT}=79.1%,

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: Yes

Smiles: No

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:

59 compounds

7.6.Experimental design of test set:

The splitting of the original dataset into a training set of 179 chemicals representative of the entire data set and a validation set of 59 chemicals was obtained by applying the Experimental Design procedure D-Optimal Design (DOLPHIN Software).This procedure provides a strategy for selecting the most dissimilar molecular structures in a data set, thus, the chemicals in the training set represent the breadth, or variety, of all existing chemicals within that domain (i.e. diversity) and the test set is within the training set domain.

7.7.Predictivity - Statistics obtained by external validation:

Q^2_{EXT}=86.4%; R^2_{EXT}=90.5%;

RMSE (validation set)=0.57

7.8.Predictivity - Assessment of the external validation set:

The splitting was done on the basis of (complete) structural similarity information that is obtained from the molecular descriptors used in the GA and also taking into account the BCF response. The external validation set was sufficiently large and representative of the applicability domain.

7.9.Comments on the external validation of the model:

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The model was developed by a statistical approach. No mechanistic basis was defined a priori.

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori mechanistic interpretation: the variables most frequently selected by Genetic Algorithm as the most informative to model the BCF were related to the molecular size of the chemical, to the chemical ability to form hydrogen bonds and to the distribution of polar atoms in the molecule.

The most important descriptors in the proposed model were IDDM, HIC and nHAcc. This highlights the relevant role of the aspects of molecular size, molecular complexity and hydrogen bonding capacity in BCF modelling. As expected, bigger molecules have a higher tendency to bioconcentrate, while those with a marked tendency towards hydrogen bonding tend to remain in the aqueous phase.

Note that hydrogen bonding descriptors were demonstrated to be related to BCF and included in QSAR models for the first time by Gramatica and Papa (2003).

The two autocorrelation descriptors that are the less important in the regression equation (MATS1p and GATS1e) bring information related to electronic properties which highlights the significance of the molecular polarity or polarizability for BCF.

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1]Lu X, Tao S, Hu H & Dawson RW (2000.) Estimation of Bioconcentration Factors of Nonionic Organic Compounds in Fish by Molecular Connectivity Indices and Polarity Correction Factors. Chemosphere 41, 1675-1688.

[2]Gramatica P & Papa E (2003). QSAR Modeling of Bioconcentration Factor by theoretical molecular descriptors. QSAR & Combinatorial Science 22, 374 –385.

[3]Tropsha A, Gramatica P & Gombar VK (2003). The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. QSAR & Combinatorial Science 22, 69-76.

9.3.Supporting information:

BCF fish Training_179.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q13-24a-0012/attachment/A646
BCF fish Test_59.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q13-24a-0012/attachment/A647

Test set(s)

10.Summary (JRC QSAR Model Database)

Supporting information

10.1.QMRF number:

Q13-24a-0012

10.2.Publication date:

2013-06-21

10.3.Keywords:

BCF;bioconcentration factor;fish;

10.4.Comments:

formerQ2-17-16-140