	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Insubria QSAR PaDEL-Descriptor model for prediction of metabolic biotranformation half-life in fish. (Split 3)** **Keywords: Biotransformation rate; metabolic half-life; QSAR; consensus modelling; risk assessment; chemical prioritization.**
	Printing Date: **30-gen-2018**

1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSAR PaDEL-Descriptor model for prediction of metabolic biotranformation half-life in fish. (Split 3)

Keywords: Biotransformation rate; metabolic half-life; QSAR; consensus modelling; risk assessment; chemical prioritization.

1.2.Other related models:

T.Brown, J.A. Arnot, F. Wania, Iterative fragment selection: a group contribution approach to predictiong fish biotrasformation half-lives. Environ Sci Technol. 2012; 46:8253-60 [1]

E. Papa, L. van der Wal, J.A. Arnot, P. Gramatica, Metabolic biotransformation half-lives in fish: QSAR modelling and consensus analysis. STOTEN. 2014;470-471:1040-1046 [2]

1.3.Software coding the model:

PaDEL-Descriptor 2.18

A software to calculate molecular descriptors and fingerprints [3]

http://padel.nus.edu.sg/software/padeldescriptor/index.html

QSARINS 2.0

Software for the development, analysis and validation of QSAR MLR models [4,5]

paola.gramatica@uninsubria.it

www.qsar.it

HyperChem ver.7.3

Software used to design, check and optimize chemical structures.

Instant JChem 5.5.0

software for calculating acid and basic pKas

Http://www.chemaxon.com

ACD Labs 12.5

software for calculating acid and basic pKas

2.General information

2.1.Date of QMRF:

02/10/17

2.2.QMRF author(s) and contact details:

[1]Alessandro Sangion DiSTA, University of Insubria (Varese - Italy) alessandro.sangion@uninsubria.it www.qsar.it

[2]Lucrezia Motta DiSTA, University of Insubria (Varese - Italy)

[3]Ester Papa DiSTA, University of Insubria (Varese - Italy) ester.papa@uninsubria.it www.qsar.it

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1]Ester Papa DiSTA, University of Insubria (Varese - Italy) ester.papa@uninsubria.it www.qsar.it

[2]Paola Gramatica DiSTA, University of Insubria (Varese - Italy) paola.gramatica@uninsubria.it www.qsar.it

[3]Leon van der Wal REACH Mastery, Como, Italy

[4]Jon A. Arnot ARC Arnot Research & Consulting, Toronto, ON, Canada ; Department of Physical and Environmental Science, University of Toronto, ON, Canada

2.6.Date of model development and/or publication:

2013/2014

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

[1]Gramatica P., et al. QSARINS: A new software for the development, analysis and validation of QSAR MLR models, J. Comput. Chem. (Software News and Updates), 2013, 34 (24), 2121-2132 [4]

[2]Gramatica P., et al. QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS, submitted to J. Comput. Chem. (Software News and Updates), 2013. [5]

[3]Papa E., et al. Metabolic biotransformation half-lives in fish: QSAR modelling and consensus analysis, STOTEN. 2014;470-471:1040-1046 [2]

[4]Yap, C.W. PaDEL descriptor: an open source software to calculate molecular descriptors and fingerprints., J. Comput. Chem. 2011 32, 1466-1474 [3]

2.8.Availability of information about the model:

Non-proprietary. Defined algorithm, available in QSARINS [4, 5]. Training and prediction sets are available in the attached sdf file of this QMRF (section 9) and in the QSARINS-Chem database [5].

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

No other information available.

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Range of fish species, body sizes and temperatures were normalized.

3.2.Endpoint:

Bioaccumulation Metabolic biotransformation in fish

3.3.Comment on endpoint:

The biotransformation rate constants from a range of fish species, body sizes and temperatures were normalized to rate constants for fish with a body weight of 0.01kg at 15°C

3.4.Endpoint units:

km (day^{-1}) rate was converted to normalized biotransformation half-life value (HL_ndays), and then expressed in base 10 log

units $LogHL_n$

3.5.Dependent variable:

Log (HL_n)

3.6.Experimental protocol:

No information available

3.7.Endpoint data quality and variability:

A mass balance method was developed to estimate in vivo whole body metabolic biotransformation rate costants in fish from laboratory bioaccumulation data. The method includes a screening level uncertainty analysis for the km estimates and was applied to a database of evaluated laboratory bioaccumulation test data to derive an in vivo km database.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR - Multiple linear regression model (OLS - Ordinary Least Square)

4.2.Explicit algorithm:

LogHLn (fish biotransformation half-life)_Split-3

OLS-MLR method. Model developed on a training set of 421 compounds

LogHLn (fish biotransformation half-life)_Full Model

OLS-MLR method. Model developed on a training set of 632 compounds

Split-3 model equation: $LogHL_n$ = - 4.19 + 1.058 VAdjMat - 0.254 MAXDP + 0.112 nX - 0.154 nHBAcc + 0.154 MDEC-11 + 0.464 FP362 - 0.141 naaaC - 0.804 minHBd - 0.311 FP376

Full model Equation: $LogHL_n$ = - 4.24 + 1.067 VAdjMat - 0.27 MAXDP + 0.12 nX - 0.14 nHBAcc + 0.44 FP362 + 0.144 MDEC-11 - 0.14 naaaC - 0.74 minHBd - 0.29 FP376

4.3.Descriptors in the model:

[1]VAdjMat Vertex adjacency index

[2]nX Number of halogens

[3]minHBd Minimum E-States for (strong) Hydrogen Bond donors

[4]MAXDP maximum electrotopological positive variation

[5]naaaC count of atom-type E-state

[6]nHBAcc number of hydrogen bond acceptor (using CDK algorithm)

[7]MDEC-11 malecular distance edge between all primary carbons

[8]FP362 simple atoms nearest neighborg C(Cl)(:C)

[9]FP376 simple atom nearest neighbors C(N)(:C)

4.4.Descriptor selection:

A total of 1567 molecular descriptors of differing types (0D, 1D, 2D) and fingerprints were calculated in PaDEL-Descriptor 2.18. Constant and semi-constant values and descriptors found to be correlated pairwise were excluded in a pre-reduction step (one of any two descriptors with a correlation greater than 0.98 was removed to reduce redundant information), and a final set of 520 molecular descriptors were used as

input variables for variable subset selection. The models were initially developed by the all-subset-procedure, and then GA was applied to obtain the final population of models (nine variables). The optimized parameter used was Q^2_{LOO}(leave-one-out).

4.5.Algorithm and descriptor generation:

Multiple linear regression (Ordinary Least Square method) was applied to generate the model.

Molecular descriptors were generated by PaDEL-Descriptor software. The input files for descriptor calculation contain information on atom and bond types, presence of halogens, E-state energy, electrotopological state,molecular dimension and hydrofobicity.

4.6.Software name and version for descriptor generation:

PaDEL-Descriptor 2.18

A software to calculate molecular descriptors and fingerprints [3]

Yap Chun Wei, Department of Pharmacy, National University of Singapore.

http://padel.nus.edu.sg/software/padeldescriptor/index.html

HYPERCHEM - ver. 7.03

Software for molecular drawing and conformational energy optimization

4.7.Chemicals/Descriptors ratio:

Split: 421 chemicals / 9 descriptors = 46.8

Full model: 632 chemicals / 9 descriptors = 70.2

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The applicability domain of the model was verified by the identification of response outliers (i.e.compounds with cross-validated standardized residuals greater than 2.5 standard deviation units) and structural outliers with 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). The applicability domain was also graphically investigated through the William plot of hat value versus standardized residuals.

Response and descriptor space:

Range of experimental $LogHL_n$values: -1.56 / 3

Range of descriptor values: VAdjMat 2.58 / 6.75 ; nX 0 /12 ; minHBd 0 / 0.91 ; MAXDP 0 / 7.76 ; naaaC 0 / 8 ; MDEC-11 0 / 17.11 ; FP362 0 / 1 ; FP376 0 / 1 ; nHBAcc 0 / 11 .

5.2.Method used to assess the applicability domain:

As it has been stated in section 5.1, the structural applicability domain of the model was assessed by the leverage approach, providing a cut-off hat value (h*=0.0475). HAT values are calculated as the diagonal elements of the HAT matrix:

$H = X(X^TX)^{-1}X^T$

The response applicability domain can be verified by the standardized residuals in cross-validation greater than 2.5 standard deviation units.

5.3.Software name and version for applicability domain assessment:

QSARINS 2.0

Software for the development, analysis and validation of QSAR MLR models

paola.gramatica@uninsubria.it

www.qsar.it

5.4.Limits of applicability:

FULL model domain: Outliers for structure, hat>3p'/n(h*):
Octaethylene glycol monotridecyl ether, Benzo[a]pyrene, Dibenzo[a,h]anthracene, Tetradecamethylcycloheptasiloxane (D7), PeryleneBenzo(k)fluoranthene, Benzo[b]chrysene, Dodecamethylcyclohexasiloxane (D6), Decamethylcyclopentasiloxane (D5), Cyclotetrasiloxane, 175 Factor L, 175 Factor J, Spinosad Factor D, Spinosad Factor A, 2,5-dimethyl-2,5-bis(tert-butylperoxy)hexane, Phenol-2,4,6-tris(1,1-dimethylethyl)-, 1,3,5-tri-tert-butylbenzene, 2,2,4,4,6,8,8-heptamethyl nonane, Fipronil sulfone.
Outliers for response, standardised residuals > 2.5 standard deviation units: Benzene-1,1,1-(chloromethylidyne)tris-, Cyclohexane-1,2,3,4,5-pentabromo-6-chloro-, Dibenzofuran, 1,3,5-trimethyl cyclohexane, 'Oxirane,[(dibromomethylphenoxy)methyl]-', Cis 1,1,3,5 tetramethyl cyclohexane, Thioperoxydicarbonic diamide-([(H2N)C(S)]2S2)-tetramethyl-.

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: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: null

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:

To verify the predictive capability of the proposed models, the dataset (n=632) was split, before model development, into a training set used for model development and a prediction set used later for external validation; the splitting scheme was the same used by Brown [1] and took into account both the range of the reponse and the structural similarity (n training=421 , n prediction=211). The range of LogHLn are: -1.57 / 3

6.6.Pre-processing of data before modelling:

Transformation of km (day^{-1}) into $LogHL_n$(day)

6.7.Statistics for goodness-of-fit:

Ordered response split model:

R^2= 0.76 ; CCC_{tr}[6,7]=0.86; $RMSE_{tr}$= 0.57

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

Ordered response Split model:

Q^2_{LOO}= 0.74 ; CCC_{cv}= 0.85 ; $RMSE_{cv}$= 0.59

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

Q^2_{LMO}= 0.74

6.10.Robustness - Statistics obtained by Y-scrambling:

R^2_{yscr}= 0.02

6.11.Robustness - Statistics obtained by bootstrap:

No information available

6.12.Robustness - Statistics obtained by other methods:

No information available

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: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: null

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 models, the dataset (n=632) was split, before model development, into a training set used for model development and a prediction set used later for external validation; the splitting scheme was the same used by Brown [1] and took into account both the range of the reponse and the structural similarity (n training=421 , n prediction=211); the range of LogHLn are: -1.40 / 2.86

7.6.Experimental design of test set:

The splitting was the same as the one used previously by Brown (Brown et al., 2012)[1]

7.7.Predictivity - Statistics obtained by external validation:

Ordered response split model:

Q^2_{extF1}[8]= 0.75; Q^2_{extF2}[9]=

0.75 ; Q^2_{extF3}[10]= 0.75; CCC_{ex}=0.86;
$RMSE_{ex}$= 0.58

7.8.Predictivity - Assessment of the external validation set:

The splitting methodology based on similarity analysis allowed for the selection of meaningful training sets and representative prediction sets. Training and prediction sets are balanced according to structure . In particular, for response the range of $LogHL_n$values are [-1.57 /3] and [-1.40 / 2.86] respectively for training and prediction sets.

As much as concern structural representativity, the range of descriptors values is:

nX: training set (0 / 12), prediction set (0 / 10);
naaaC: training set (0 / 8), prediction set (0 / 8);
minHBd: training set (0 / 0.89), prediction set (0 / 0.915);
nHBAcc: training set (0 / 11), prediction set (0 / 11);
MAXDP: training set (0 / 7.7), prediction set (0 / 7.76);
MDEC-11: training set (0 / 17.11), prediction set (0 / 7.98);
VAdjMAT: training set (2.58 / 6.73), prediction set (2.58 / 6.75);
FP362: training set (0 / 1), prediction set (0 / 1);
FP376: training set (0 / 1), prediction set (0 / 1);

7.9.Comments on the external validation of the model:

no other information available

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The model was developed by statistical approach. No mechanistic basis for this physico-chemical property was set a priori, but a mechanistic interpretation of molecular descriptors was provided a posteriori (see 8.2).

8.2.A priori or a posteriori mechanistic interpretation:

The most relevant descriptors for the modeling of the selected response are the vertex adjacemcy index (VAdjMAT), the number of haogens (nX) and the minumun E-state energy for hydrogen bond donors (minHBd). The first two descriptors gave information abut molecular dimension, hydrophobicity and presence of halogen atoms; the correlation among VAdjmat e LogKow is 62%, and the correlation among nX and VAdjMAT and molecular weight is 61% and 75% respectively. The last descriptor describes the ability of the chemical to participate in intramoleular interactions. Two other important descriptors are the maximum electrotopologicalpositive variation (MAXDP) and the count of atom-type E-state (naaaC). Also there are other variables selected in the proposed model encode for specific information related to structural domains, atoms fingerprints, functional groups and bonds. Moving from slowly biotransformated chemicals to chemicals that are relatively quickly biotransformated, is observed an increase in MAXDP and minHBd values while VAdjMAT and nX values decrease; in particular most of the slowly metabolized compounds have VAdjMAT values between 4.5 and 6,

MAXDP values between 0 and 1, and minHBd = 0. This means that in the current dataset, slower biotransformation is associated to chemicals with no, or limited, ability to participate in non-covalent intramolecuar interactions. These chemicals are characterized by large hydrophobic, halogenated structures, with few or no ramifications, and one or more aromatic rings. The increasing presence of polar and ionizable groups as well as the number and variety of reactive functional groups simultaneously present in the molecule is generally associated with faster biotransformation rates (shorter HL_n)

8.3.Other information about the mechanistic interpretation:

no other information available

9.Miscellaneous information

9.1.Comments:

Given the good results of the external validation, this model has a good applicability domain and therefore unsuccessful applications are probably very reduced. Anyhow, the check of outliers by the Williams plot and the Insubria graph for chemicals without experimental data will allow to verify the model applicability. To predict LogHLn for new chemicals without experimental data, it is suggested to apply the equation of the full model, developed on all the available chemicals (n training = 632)

$LogHL_n$ = - 4.24 + 1.067 VAdjMat - 0.27 MAXDP + 0.12 nX - 0.14 nHBAcc + 0.44 FP362 + 0.144 MDEC-11 - 0.14 naaaC - 0.74 minHBd - 0.29 FP376

n training set =632 ; R^2= 0.76 ; Q^2_{LOO}= 0.75 ; $Q^2_{lmo30\%}$= 0.73 ; CCC_{tr}= 0.86 ; CCC_{cv}= 0.86 ; $RMSE_{tr}$= 0.57 ; $RMSE_{cv}$= 0.58

9.2.Bibliography:

[1]T.Brown, J.A. Arnot, F. Wania, Iterative fragment selection: a group contribution approach to predictiong fish biotrasformation half-lives. Environ Sci Technol. 2012; 46:8253-60

[2]E. Papa, L. van der Wal, J.A. Arnot, P. Gramatica, Metabolic biotransformation half-lives in fish: QSAR modelling and consensus analysis. STOTEN. 2014;470-471:1040-1046

[3]Yap C. PaDEL-Descriptor: an open source software to calculate molecular desctriptors and fingerprints. J Comput Chem. 2011;32:1466-74

[4]Gramatica P., et al. QSARINS: A new software for the development, analysis and validation of QSAR MLR models, J Comput Chem (Software News and Updates). 2013, 34 (24), 2121-2132

[5]Gramatica P., et al. QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS, submitted to J Comput Chem (Software News and Updates). 2013.

[6]Chirico N., Gramatica P., Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model. 2011;51:2320-2335

[7]Chirico N., Gramatica P., Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. J Chem Inf Model.

2012;52:2044-2058
[8]Shi L.M. et al. QSAR models using a large diverse set of estrogens, J Chem Inf Comput Sci. 2001;41:186-195
[9]Schuurman G. et al. External validation and prediction employing the predictive squared correlation coefficient - Test set activity mean vs training set activity mean, J Chem Inf Model. 2008;48:2140-2145
[10]Consonni V., Ballabio D., Todeschini R., Comments on the definition of the Q2 parameter for QSAR validation. J Chem Inf Model. 2009;49:1669-1678

9.3.Supporting information:

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

10.Summary (JRC Inventory)

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