	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Development of human biotransformation QSARs and application for PBT assessment refinement (B3)** **Keywords: In vivo biotransformation; biotransformation half-life; QSAR; hazard assessment; refined PBT assessment.**
	Printing Date: **29-gen-2018**

1.QSAR identifier

1.1.QSAR identifier (title):

Development of human biotransformation QSARs and application for PBT assessment refinement (B3)

Keywords: In vivo biotransformation; biotransformation half-life; QSAR; hazard assessment; refined PBT assessment.

1.2.Other related models:

S. Cassani, P. Gramatica, Identification of potential PBT behavior of personal care products by structural approaches. Sustain Chem Pharm, 2015;1:19-27 [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]

A. Sangion, P. Gramatica, PBT assessment and prioritization of contaminants of emerging concern: pharmaceuticals. Environ Res, 2016a;147:297-306 [3]

1.3.Software coding the model:

PaDEL-Descriptor 2.18

A software to calculate molecular descriptors and fingerprints [4]

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

QSARINS 2.0

Software for the development, analysis and validation of QSAR MLR models [5,6]

paola.gramatica@uninsubria.it

www.qsar.it

2.General information

2.1.Date of QMRF:

10/10/17

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]Alessandro Sangion DiSTA, University of Insubria (Varese - Italy)

alessandro.sangion@uninsubria.it www.qsar.it
[3]Jon A. Arnot ARC Arnot Research & Consulting, Toronto, ON, Canada ; Department of Physical and Environmental Science, University of Toronto, ON, Canada
[4]Paola Gramatica DiSTA, University of Insubria (Varese - Italy) paola.gramatica@uninsubria.it www.qsar.it

2.6.Date of model development and/or publication:

2016/2017

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

[1]S. Cassani, P. Gramatica, Identification of potential PBT behavior of personal care products by structural approaches. Sustain Chem Pharm, 2015;1:19-27 [1]
[2]Papa E., et al. Metabolic biotransformation half-lives in fish: QSAR modelling and consensus analysis, STOTEN. 2014;470-471:1040-1046 [2]
[3]A. Sangion, P. Gramatica, PBT assessment and prioritization of contaminants of emerging concern: pharmaceuticals. Environ Res, 2016a;147:297-306 [3]
[4]Yap, C.W. PaDEL descriptor: an open source software to calculate molecular descriptors and fingerprints., J. Comput. Chem. 2011 32, 1466-1474 [4]
[5]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]
[6]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]
[7]J.A. Arnot, T.N. Brown, F. Wania, Estimating screening-level organic chemical half-lives in humans. Environ Sci Technol, 2014; 48:723-730 [7]

2.8.Availability of information about the model:

Non-proprietary. Defined algorithm, available in QSARINS [5, 6]. Training and prediction sets are available in the attached sdf file of this QMRF (section 9) and in the QSARINS-Chem database [6].

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:

Human

3.2.Endpoint:

Bioaccumulation Metabolic biotransformation in human

3.3.Comment on endpoint:

This study addresses the development of QSAR models for the prediction of the whole body biotransformation half-lives (HL_B). The first aim of this work is the creation of statistically valid and predictive models for the prediction of half-lives in human; the second aim is to show how QSAR predictions can be used for the refinement of chemical screening procedures for hazard assessment.

3.4.Endpoint units:

kB (h^{-1}) rate was converted to normalized biotransformation half-life value (HL_B, h), and then expressed in base 10 log units $LogHL_B$

3.5.Dependent variable:

Log (HL_B)

3.6.Experimental protocol:

No information available

3.7.Endpoint data quality and variability:

The dataset was taken from literature [7]

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:

LogHLB3 (biotransformation half-life in human)

OLS-MLR method. Model developed on a training set of 467 compounds

LogHLB3 (biotransformation half-life in human)_Full model

OLS-MLR method. Model developed on a training set of 935 compounds

Split model equation: $LogHL_{B3}$ = - 1.0587 + 0.662
AATS7p + 0.139 nX + 0.1311 SsCl - 0.7018 maxHsOH + 1.5218 JGT + 0.6609
GATS1s + 0.7009 MATS1c + 0.8135 FMF

Full model Equation: $LogHL_{B3}$ = - 1.0393 + 0.5304
AATS7p + 0.1522 nX + 0.1459 SsCl - 0.6435 maxHsOH + 1.6538 JGT + 0.693
GATS1s + 0.8358 FMF + 0.6658 MATS1c

4.3.Descriptors in the model:

[1]nX Number of halogen atoms

[2]MATS1c Moran autocorrelation lag 1 weighted by charges

[3]GATS1s geary autocorrelation of lag 1 weighted by the instrinsic-state

[4]SsCl sum of atom-type E-state -Cl

[5]AATS7p average Broto-Moreau autocorrelation of lag 7 weighted by polarizabilities

[6]maxHsOH maximum atom-type E-state -OH

[7]JGT global topological charge index

[8]FMF complexity of a molecule

4.4.Descriptor selection:

SMILES notation were used to encode for 2Dstructural information for all the molecules in the dataset; canonical smiles were derived by OpenBabel [8]. The smiles string were used to calculate mono- and bidimensional descriptors by the software PaDEL-Descriptor[4]. Constatnt descriptors and descriptors with a correlation greater than 0.98 were excluded from the total amout of descriptiors, using QSARINS software [5,6]. The models were initially developed by the all-subset-procedure, and then GA was applied to obtain the final population of models (eight 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 [4]

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

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

OpenBabel 2.3.2

Open Babel: the open source chemistry toolbox. [8]

4.7.Chemicals/Descriptors ratio:

Split: 467 chemicals / 8 descriptors = 58.375

Full model: 935 chemicals / 8 descriptors = 116.875

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 $LogHL_{B3}$values: -1.30 / 6.31

Range of descriptor value: ScCl -0.22 / 10.86 ; AATS7p 0 / 5.97 ; MATS1c -1 / 0.015 ; nX 0 / 17 ; GATS1s 0 / 1.70 ; FMF 0 / 0.71 ; maxHsOH 0 / 0.98 ; JTG 0 / 0.97

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*=0.0289). 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:

000006-01-7, 000050-00-0, 000051-48-9, 000051-75-2, 000052-24-4, 000057-74-9, 000058-14-0, 000058-89-9, 000067-66-3, 000071-43-2, 000075-01-4, 000075-69-4, 000075-71-8, 000079-01-6, 000087-86-5, 000104-31-4, 000113-00-8, 000118-74-1, 000123-91-1, 000154-93-8, 000156-60-5, 000307-24-4, 000335-67-1, 000355-46-4, 000396-01-0, 000461-78-9, 000657-24-9, 001163-19-5, 001744-22-5, 001763-23-1, 002051-24-3, 003194-55-6, 004428-95-9, 006893-02-3, 019982-08-2, 023288-49-5, 028523-86-6, 052485-79-7, 053230-10-7, 054143-55-4, 059080-40-9, 059933-66-3, 060348-60-9, 129453-61-8, 164656-23-9, 187523-35-9, 189084-64-8, 207122-15-4, 207122-16-5, 486460-32-6.

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

000051-75-2, 001163-19-5, 129453-61-8, 000052-01-7, 000054-05-7, 000059-66-5, 000067-97-0, 000072-55-9, 000098-95-3, 000100-00-5, 000846-48-0, 001746-01-6, 005436-43-1, 041318-75-6, 054350-48-0, 067227-57-0, 112953-11-4, 115956-12-2.

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

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=935) 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 as the one used prevously by Arnot [7] (n training=467 , n prediction=468). The range of $Log_{HLB3}$are: -1.30 / 6.31

6.6.Pre-processing of data before modelling:

Transformation of kB (h^{-1}) into $LogHL_B(h)$

6.7.Statistics for goodness-of-fit:

Ordered response split model:

R^2= 0.79 ; CCC_{tr}[9,10]= 0.88 ; $RMSE_{tr}$=

0.62

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

Ordered response Split model:

Q^2_{LOO}= 0.78 ; CCC_{cv}= 0.88; $RMSE_{cv}$= 0.64

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

Q^2_{LMO}= 0.78

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: No

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=935) 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 as the one used prevously by Arnot [7] (n training=467 , n prediction=468). The range of $LogHL_{B3}$are: -1.30 / 6.31

7.6.Experimental design of test set:

The splitting was the same as the one used previously by Arnot (Arnot et al., 2014) [7]

7.7.Predictivity - Statistics obtained by external validation:

Ordered response split model:

Q^2_{extF1}[11]= 0.76 ; Q^2_{extF2}[12]= 0.76 ; Q^2_{extF3}[13]= 0.75 ; CCC_{ex}=0.87 ; $RMSE_{ex}$= 0.68

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 $LogHL_{B3}$values are [-1.30 / 6.31] and [-1.08 / 5.67] respectively for training and prediction sets.
As much as concern structural representativity, the range of descriptors values is:
nX: training set (0 / 8), prediction set (0 / 17);
SsCl: training set (0 / 8.48), prediction set (-0.22 / 10.86);
GATS1s: training set (0 / 1.67), prediction set (0.39 / 1.70);
AATS7p: training set (0 / 4.24), prediction set (0 / 5.97);
MATS1c: training set (-0.98 / -0.03), prediction set (-0.99 / 0.015);
maxHsOH: training set (0 / 0.9), prediction set (0 / 0.98);
FMF: training set (0 / 0.71), prediction set (0 / 0.67);
JGT: training set (0.167 / 0.83), prediction set (0 / 0.97);

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 sum of atom-type electrotopological state -Cl (SsCl), the number of halogen atoms (nX) and the average Broto-Moreau autocorrelation of lag 7, weighted by polarizabilities (AATS7p). SsCl encodes for for information about the electrotopological state of atom bounded to chlorine atoms. nO gives informations about the presence and the number of halogen atoms, large nX values will increase the predicted biotransformation half-life. AATS7p is an autocorrelation descriptor that account for the intramolecular variation of the polarizability all over the molecular structure, encodes for the presence of polar atoms in large molecules, it has direct effect on biopersistence. Another important descriptor is geary autocorrelation of lag 1 weighted by the Intrinsic-state (GATS1s), it accounts for the distribution of the I-state , the ratio of and lone pair electrons over the count of the bonds in the molecular graph for the considered atoms; high values for this descriptor reflect stable molecule with long half-lives predicted by the model.

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 LogHLB3 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 = 935)

$LogHL_{B3}$ = - 1.0393 + 0.5304 AATS7p + 0.1522 nX + 0.1459 SsCl - 0.6435 maxHsOH + 1.6538 JGT + 0.693 GATS1s + 0.8358 FMF + 0.6658 MATS1c

n training set = 935 ; R^2= 0.78 ; Q^2_{LOO}= 0.77 ; $Q^2_{lmo30\%}$= 0.77 ; CCC_{tr}= 0.87 ; CCC_{cv}= 0.87 ; $RMSE_{tr}$= 0.65 ; $RMSE_{cv}$= 0.66

9.2.Bibliography:

[1]S. Cassani, P. Gramatica, Identification of potential PBT behavior of personal care products by structural approaches. Sustain Chem Pharm, 2015;1:19-27

[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]A. Sangion, P. Gramatica, PBT assessment and prioritization of contaminants of emerging concern: pharmaceuticals. Environ Res, 2016a;147:297-306

[4]Yap C. PaDEL-Descriptor: an open source software to calculate molecular desctriptors and fingerprints. J Comput Chem. 2011

[5]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

[6]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.

[7]J.A. Arnot, T.N. Brown, F. Wania, Estimating screening-level organic chemical half-lives in human. Environ Sci Technol. 2014; 48:723-730

[8]N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchinson, Open Babel: an open chemical toolbox. J Cheminform, 2011;3:33

[9]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

[10]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

[11]Shi L.M. et al. QSAR models using a large diverse set of estrogens, J Chem Inf Comput Sci. 2001;41:186-195

[12]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

[13]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