QMRF identifier (JRC Inventory):	*To be entered by JRC*
QMRF Title:	*Insubria QSPR PaDEL-Descriptor model for octanol-air partition coefficient (logKoa) prediction of Polybrominated Diphenyl Ethers.*
Printing Date:	*Jan 20, 2014*

1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSPR PaDEL-Descriptor model for octanol-air partition coefficient (logKoa) prediction of Polybrominated Diphenyl Ethers.

1.2.Other related models:

E. Papa, S. Kovarich, P. Gramatica, 2009, Validation and Inspection of the Applicability Domain of QSPR Models for Physicochemical Properties of Polybrominated Diphenyl Ethers, QSAR & Comb.Sci. 28, 790-796 [9].

1.3.Software coding the model:

[1]PaDEL-Descriptor 2.18 A software to calculate molecular descriptors and fingerprints http://padel.nus.edu.sg/software/padeldescriptor/index.html
[2]QSARINS 1.2 Software for the development, analysis and validation of QSAR MLR models paola.gramatica@uninsubria.it www.qsar.it

2.General information

2.1.Date of QMRF:

20/11/2013

2.2.QMRF author(s) and contact details:

[1]Stefano Cassani DiSTA, University of Insubria (Varese - Italy) +390332421439 stefano.cassani@uninsubria.it www.qsar.it
[2]Alessandro Sangion DiSTA, University of Insubria (Varese - Italy) +390332421439 a.sangion@hotmail.it www.qsar.it
[3]Paola Gramatica DiSTA, University of Insubria (Varese - Italy) paola.gramatica@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]Stefano Cassani DiSTA, University of Insubria (Varese - Italy) +390332421439 stefano.cassani@uninsubria.it www.qsar.it
[2]Paola Gramatica DiSTA, University of Insubria (Varese - Italy) paola.gramatica@uninsubria.it www.qsar.it

2.6.Date of model development and/or publication:

July 2013

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 [1]
[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.

2.8.Availability of information about the model:

The model is non-proprietary and published in a scientific peerreviewed journal. All information in full details are available (e.g.training and prediction set, algorithm, ecc...).

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:

No information available

3.2.Endpoint:

1.Physicochemical effects 1.8.Octanol-air partition coefficient (Koa)

3.3.Comment on endpoint:

The octanol-air partition coefficient (Koa) is the ratio of the solute concentration in air versus octanol when the octanol-air system is at equilibrium. Koa values are reported in literature in Log units (LogKoa).

3.4.Endpoint units:

Dimensionless

3.5.Dependent variable:

LogKoa

3.6.Experimental protocol:

Experimentally measured LogKoa for 30 PBDEs (Polybrominated Diphenyl Ethers) were collected from 3 different sources: Harner and Shoeib, 2002 (data for 13 PBDEs) [2], Wania et al., 2002 (data for 22 PBDEs) [3], Gouin and Harner, 2003 (data for 5 PBDEs) [4]. When more than one experimental value was available for a single compound, the average value was calculated and used as input data for the development of the QSPR model.

3.7.Endpoint data quality and variability:

The availability of experimental data from different sources made it possible to verify the data quality and the variability between different laboratories (data reproducibility). When more than one experimental value was available for a single compound, the variation of data was quantified by calculation of standard deviation (smax = 0.382) and coefficient of variation (CV% = 0.2-3.6% - 7 comp. with multiple exp val.).

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSPR - Multiple linear Regression Model (OLS - Ordinary least-squares)

4.2.Explicit algorithm:

LogKoa (split model)

OLS-MLR method. Model developed on a training set of 24 compounds

LogKoa (full model)
OLS-MLR method. Model developed on all the available experimental data (training set of 30 compounds).

Split Model: LogKoa = 5.20 + 25.45 VCH-7

Full Model: LogKoa = 5.34 + 24.87 VCH-7

4.3.Descriptors in the model:

VCH-7 Valence chain, order 7

4.4.Descriptor selection:

A total of 682 molecular descriptors of differing types (0D, 1D, 2D) 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 130 molecular descriptors were used as input variables for variable subset selection. The models were developed by the all-subset-procedure with only one variable. The optimized parameter used was Q2LOO (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, connectivity, partial charges and atomic spatial coordinates, relative to the minimum energy conformation of the molecule, and were firstly obtained by the semi empirical AM1 method using the package HYPERCHEM. Then, these files were converted by OpenBabel into MDL-MOL format and used as input for the calculation of descriptors in PaDEL-Descriptor.

4.6.Software name and version for descriptor generation:

PaDEL-Descriptor 2.18
A software to calculate molecular descriptors and fingerprints
Yap Chun Wei, Department of Pharmacy, National University of Singapore.
http://padel.nus.edu.sg/software/padeldescriptor/index.html

4.7.Chemicals/Descriptors ratio:

Split Model: 24 chemicals / 1 descriptor = 24

Full Model: 30 chemicals / 1 descriptor = 30

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 leverage approach and fixed thresholds has been used to define both structural and response outliers (see section 5.4). The plot of leverages (hat diagonals) versus standardised residuals, i.e. the Williams plot, verified the presence of response outliers (i.e.compounds with cross-validated standardized residuals greater than 2.5 standard deviation units) and chemicals very structurally influential in determining model parameters

parameters (i.e. compounds with a 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). For new compounds without experimental data, leverage can be used as a quantitative measure for evaluating the degree of extrapolation: for compounds with a high leverage value (h > h*), that are structural outliers, predictions should be considered less reliable.

Response and descriptor space:
Range of experimental LogKoa values: 7.34 / 11.97
Range of descriptor values: VCH-7: 0.08 / 0.28

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*=0.200). HAT values are calculated as the diagonal elements of the HAT matrix:

H = X(XTX)-1XT

The response applicability domain can be verified by the standardized residuals, calculated as: r'i = ri / s√(1-hii), where ri = Yi-Ŷi.

5.3.Software name and version for applicability domain assessment:

QSARINS 1.2
Software for the development, analysis and validation of QSAR MLR models
paola.gramatica@uninsubria.it
www.qsar.it

5.4.Limits of applicability:

Split model domain: outliers for structure, hat>0.250 (h*): 1-Bromo-3-phenoxybenzene (6876-00-2), 1-bromo-2-phenoxybenzene (36563-47-0). Outliers for response, standardised residuals > 2.5 standard deviation units: 3,3',4,4',5-PentaBDE (366791-32-4). **FULL model domain**: outliers for structure, hat>0.200 (h*): no. Outliers for response, standardised residuals > 2.5 standard deviation units: 3,3',4,4',5-PentaBDE (366791-32-4).

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:Yes
INChI:No
MOL file:Yes

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:
The training set of the Split Model consists of 24 PBDEs, from di- to hepta-BDEs; training and test set are structurally belanced, being the splitting based on the structural similarity analysis.
6.6.Pre-processing of data before modelling:
Raw data, collected from 3 different sources [2-4], have been combined before modelling (if more than one experimental value was available for a single compound, the average value was used).
6.7.Statistics for goodness-of-fit:
R^2= 0.96; CCCtr [5]=0.98; RMSE= 0.26
6.8.Robustness - Statistics obtained by leave-one-out cross-validation:
Q^2LOO= 0.95; CCCcv=0.97; RMSEcv= 0.29
6.9.Robustness - Statistics obtained by leave-many-out cross-validation:
Q^2LMO= 0.95.
6.10.Robustness - Statistics obtained by Y-scrambling:
R^2y-sc= 0.04
6.11.Robustness - Statistics obtained by bootstrap:
No information available (since we have calculated Q^2LMO)
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:Yes
INChI:No
MOL file:Yes
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:
The external validation set of the Split Model consists of 6 compounds (mono- to tetra-BDEs), with a range of logKoa: 7.34 / 10.66.
7.6.Experimental design of test set:
The splitting of the original data set (30 compounds) into a training set of 24 compounds (representative of the entire data set) and a validation set of 6 compounds (splitting 20%) was realized by applying Self Organized Maps Kohonen Artificial Neural Networks (SOM

K-ANN).

7.7.Predictivity - Statistics obtained by external validation:

Q^2extF1 [6]= 0.99; Q^2extF2 [7]= 0.96; Q^2extF3 [8]= 0.96; CCCex=0.98; RMSE= 0.24

7.8.Predictivity - Assessment of the external validation set:

The splitting methodology based on similarity analysis (performed by the application of the Kohonen maps Artificial Neural Networks - KANN) allowed for the selection of a meaningful training set and a representative prediction set.

Training and prediction set are balanced according to both structure and response. In particular, for response the range of logKoa values are [7.34 / 11.97] and [7.34 / 10.66] respectively for training and prediction set. As much as concern structural representativity, the range of descriptors values is:

VCH-7: training set (0.12 / 0.28), prediction set (0.08 / 0.21)

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 model equation published in Papa et al. [9] was:

LogKoa: 6.654 + 0.222 T(O...Br)

where T(O..Br) is the sum of topological distances between O..Br.

This descriptor gives a double structural information: its values increases according to both the number and the distance of bromine substituents, on each phenyl ring, from the oxygen ether.Thus, T(O...Br) takes also into account the information related to the position of the bromine atoms on the phenyl rings.

The equation of the new PaDEL-descriptor model included in QSARINS is : LogKoa = 5.34 + 24.87 VCH-7

where VCH-7 is the Valence chain of order 7, which values increases with the number of Bromine atoms.

The correlation between T(O..Br) and VCH-7 is high, 0.98: both the descriptors are similarly able to model the tendency of the Koa increase with the number of bromine substituents.

8.3.Other information about the mechanistic interpretation:

no other information available

9.Miscellaneous information

9.1.Comments:

To predict logKoa for new chemicals without experimental data, it is suggested to apply the equation of the Full Model, developed on all the available chemicals (N=30), thus ensuring a wider applicability domain.

The equation (reported also in section 4.2) and the statistical parameters of the full model are:

logKoa= 5.340 + 24.869 VCH-7

N = 30; R2 = 0.97; Q2 = 0.96; Q2LMO = 0.96; CCC = 0.98; CCCcv = 0.98; RMSE= 0.253; RMSEcv = 0.271

9.2.Bibliography:

[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.
[2]T. Harner, M. Shoeib, J. Chem. Eng. Data, 2002, 47, 228-232.
[3]F. Wania, Y.D. Lei, T. Harner, Anal. Chem., 2002, 74, 3476-3483.
[4]T. Gouin, T. Harner, Environ. Int., 2003, 29, 717-724.
[5]Chirico N. and 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, pp 2044– 2058
[6]Shi L.M. et al. QSAR Models Using a Large Diverse Set of Estrogens, J. Chem. Inf. Comput. Sci. 41 (2001) 186–195.
[7]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. 48 (2008) 2140-2145.
[8]Consonni V. et al. Comments on the Definition of the Q2 Parameter for QSAR Validation, J. Chem. Inf. Model. 49 (2009) 1669-1678
[9]E. Papa, S. Kovarich, P. Gramatica, 2009, Validation and Inspection of the Applicability Domain of QSPR Models for Physicochemical Properties of Polybrominated Diphenyl Ethers, QSAR & Comb.Sci. 28, 790-796 [9].

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