
	QMRF identifier (JRC Inventory): To be entered by JRC	
	QMRF Title: Insubria QSPR PaDEL-Descriptor model for Vapor Pressure prediction of Polybrominated Diphenyl Ethers.	
	Printing Date: Jan 20, 2014	

1. QSAR identifier

1.1. QSAR identifier (title):

Insubria QSPR PaDEL-Descriptor model for Vapor Pressure 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:

22/11/2013

2.2. QMRF author(s) and contact details:

Stefano Cassani DiSTA, University of Insubria (Varese - Italy)
+390332421439 stefano.cassani@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 peer-reviewed 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 1.4. Vapour pressure

3.3. Comment on endpoint:

The vapor pressure (P_v) describes the equilibrium partitioning of a chemical between its pure condensed state and the gaseous state at a specified temperature. In particular, the modelled endpoint refers to subcooled/supercooled liquid vapour pressure (VP/PL) determined with the gas chromatographic (GC) retention time technique.

3.4. Endpoint units:

Experimentally measured vapor pressures (Pa, 25 °C) were transformed into logarithmic units (Log_1/P_v). The modelled endpoint is thus dimensionless.

3.5. Dependent variable:

Log_1/P_v

3.6. Experimental protocol:

Experimentally measured vapor pressures for 34 PBDEs (Polybrominated Diphenyl Ethers) were collected from 3 different sources: Wong et al., 2001 (data for 23 PBDEs), Wania and Dugani, 2003 (data for 6 PBDEs), Palm et al., 2002 (data for 14 PBDEs) [2-4]. The homogeneous experimental data by Wong et al. (2001) were used and were integrated (but not averaged) with those collected from different sources by Wania and Dugani (2003) and Palm et al. (2002).

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). The values reported by different authors are comparable.

4. Defining the algorithm - OECD Principle 2

4.1. Type of model:

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

4.2. Explicit algorithm:

log_1/VP (split model)

OLS-MLR method. Model developed on a training set of 28 compounds

log_1/VP (Full model)

OLS-MLR method. Model developed on a training set of 34 compounds

Split Model: $\log_1/VP = 0.58 + 1.72 \text{ SsBr}$

Full Model: $\log_1/VP = 0.54 + 1.75 \text{ SsBr}$

4.3.Descriptors in the model:

SsBr Sum of atom-type E-State: -Br.

4.4.Descriptor selection:

A total of 672 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>

HYPERCHEM - ver. 7.03

Software for molecular drawing and conformational energy optimization

OpenBabel ver.2.3.2

Open Babel: The Open Source Chemistry Toolbox. Used for conversion between HYPERCHEM files (hin) and MDL-MOL files.

<http://openbabel.org>

4.7.Chemicals/Descriptors ratio:

Split Model: 28 chemicals / 1 descriptor = 28

Full Model: 34 chemicals / 1 descriptor = 34

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 (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 \log_1/VP values: 0.79 / 6.77

Range of descriptor values: SsBr: 0.31 / 3.64

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 ($h^*=0.176$). HAT values are calculated as the diagonal elements of the HAT matrix:

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

The response applicability domain can be verified by the standardized residuals, calculated as: $r'_i = r_i / \sqrt{s^2(1-h_{ii})}$, where $r_i = Y_i - \hat{Y}_i$.

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, $hat > 0.214$ (h^*): 2,2',3,3',4,4',6,6'-octaBDE (no CAS). Outliers for response, standardised residuals > 2.5 standard deviation units: no. **FULL**

model domain: outliers for structure, $hat > 0.176$ (h^*): 2,2',3,3',4,4',6,6'-octaBDE (no CAS). Outliers for response, standardised residuals > 2.5 standard deviation units: no.

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 28 PBDEs; training and test set are structurally balanced, being the splitting based on the structural similarity analysis.

6.6.Pre-processing of data before modelling:

Raw data, collected from 3 different references (Wong et al. (2001), Wania and Dugani (2003), Palm et al. (2002)), have been combined (data by Wong et al. (2001) were used and were integrated with those collected from the other two sources), transformed into logarithmic units and multiplied by -1 before modelling.

6.7.Statistics for goodness-of-fit:

$R^2 = 0.99$; $CC_{tr}[5] = 0.99$; $RMSE = 0.16$

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

$Q^2_{LOO} = 0.98$; $CCC_{cv} = 0.99$; $RMSE_{cv} = 0.18$

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

$Q^2_{LMO} = 0.98$.

6.10.Robustness - Statistics obtained by Y-scrambling:

$R^2_{y-sc} = 0.04$

6.11.Robustness - Statistics obtained by bootstrap:

No information available (since we have calculated Q^2_{LMO})

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, with a range of log₁/VP: 0.89 / 3.81.

7.6. Experimental design of test set:

The splitting of the original data set (34 compounds) into a training set of 28 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:

$Q^2_{\text{extF1}} [6] = 0.98$; $Q^2_{\text{extF2}} [7] = 0.95$; $Q^2_{\text{extF3}} [8] = 0.98$;
CC_{Cex} = 0.97; RMSE = 0.19

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 log₁/VP values are [0.79 / 6.77] and [0.89 / 3.81] respectively for training and prediction set. As much as concern structural representativity, the range of descriptors values is:

SsBr: training set (0.31 / 3.64), prediction set (0.38 / 1.79)

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

$$\log_1/\text{VP} = 0.115 + 0.213 T(\text{O..Br})$$

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.

In the PaDEL-descriptor model included in QSARINS the equation is:
 $\log_1/VP = 0.54 + 1.75 SsBr$

where SsBr is the Sum of atom-type E-State: -Br, which values increase with the number of bromine atoms.

The correlation between (O..Br) and SsBr is very high: >0.99.

8.3. Other information about the mechanistic interpretation:

no other information available

9. Miscellaneous information

9.1. Comments:

To predict \log_1/VP for new BFR chemicals without experimental data, it is suggested to apply the equation of the Full Model, developed on all the available chemicals (N=34), thus ensuring a wider applicability domain.

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

$$\log_1/VP = 0.54 + 1.75 SsBr$$

N = 34; R² = 0.99; Q² = 0.98; Q²LMO = 0.98; CCC = 0.99; CCC_{cv} = 0.99; RMSE = 0.17; RMSE_{cv} = 0.18

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 [1]
- [2] A. Wong, Y. D. Lei, M. Alaei, F. Wania, J. Chem. Eng. Data 2001, 46, 239 – 242.
- [3] F. Wania, C. B. Dugani, Environ. Toxicol. Chem. 2003, 22, 1252 – 1261.
- [4] A. Palm, I. T. Cousins, D. Mackay, M. Tysklynnd, M. Alaei, Environ. Pollut. 2002, 117, 195 – 213.
- [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.

9.3.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