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

1. QSAR identifier

1.1. QSAR identifier (title):

Insubria QSPR PaDEL-Descriptor model for Vapor Pressure prediction of (Benzo-)Triazoles

1.2. Other related models:

B. Bhatarai and P. Gramatica, 2011. Modelling physico-chemical properties of (benzo)triazoles, and screening for environmental partitioning. *Water Res.* 45, 1463-1471. [8]

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:

3/12/2013

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:

Vapor Pressure (VP) is the pressure exerted by a vapor in equilibrium with the solid or liquid phase of the same substance.

3.4. Endpoint units:

mmHg

3.5. Dependent variable:

LogVP

3.6. Experimental protocol:

Experimentally measured VP for 33 (B)TAZs ((benzo-)triazoles) were collected from the ChemID plus database [2], compiled by the Syracuse Research Center (SRC) [3].

3.7. Endpoint data quality and variability:

No information about the data quality were available.

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:

LogVP (SOM split model)

OLS-MLR method. Model developed on a training set of 25 compounds

LogVP (Full model)

OLS-MLR method. Model developed on a training set of 33 compounds

SOM Split model equation: $\log VP = -6.43 + 12.75 \text{ ATSc5} - 1.43 \text{ n6Ring} + 0.84 \text{ minwHBa}$

Full model equation: $\log VP = -6.78 + 12.00 \text{ ATSc5} - 1.24 \text{ n6Ring} + 0.89 \text{ minwHBa}$

4.3. Descriptors in the model:

[1]ATSc5 ATS autocorrelation descriptor, weighted by charges

[2]n6Ring Number of 6-membered rings

[3]minwHBa Minimum E-States for weak Hydrogen Bond acceptors

4.4. Descriptor selection:

A total of 717 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 241 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 (three variables). 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:

SOM Split Model: 25 chemicals / 3 descriptor = 8.33

Full Model: 33 chemicals / 3 descriptor = 11

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 logVP values: -11.56 / -4.06

Range of descriptor values: ATSc5: -0.20 / 0.26; n6Ring: 0 / 2;
minwHBa: -1.11 / 1.53

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 ($h^*=0.364$). 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: $r'_i = r_i / s\sqrt{(1-h_{ii})}$, 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:

SOM Split model domain: outliers for structure, $hat > 0.480 (h^*)$: isazophos (42509-80-8), Flumetsulam (98967-40-9). Outliers for response, standardised residuals > 2.5 standard deviation units: cafenstrole (125306-83-4). **FULL model domain:** outliers for structure, $hat > 0.364 (h^*)$: no. Outliers for response, standardised residuals > 2.5 standard deviation units: cafenstrole (125306-83-4).

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:

To verify the predictive capability of the proposed models, the dataset (n=33) was split, before model development, into a training set used for model development and a prediction set used later for external validation. The procedure was made by structural similarity analysis (SOM, n training=25).

6.6. Pre-processing of data before modelling:

The data was used as LogVP mmHg

6.7. Statistics for goodness-of-fit:

$R^2 = 0.82$; $CC_{tr} [4] = 0.90$; $RMSE = 0.73$

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

$Q^2_{LOO} = 0.71$; $CCC_{cv} = 0.85$; $RMSE_{cv} = 0.92$

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

$Q^2_{LMO} = 0.70$.

6.10. Robustness - Statistics obtained by Y-scrambling:

$R^2_{y-sc} = 0.13$

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 Split Model consists of 8 compounds, with a range of logVP: -10.27 / -4.06

7.6. Experimental design of test set:

The splitting of the original data set (33 compounds) into a training set of 25 compounds (representative of the entire data set) and a validation set of 8 compounds was realized by applying Self Organized Maps Kohonen Artificial Neural Networks (SOM).

7.7. Predictivity - Statistics obtained by external validation:

$Q^2_{\text{extF1}} [5] = 0.71$; $Q^2_{\text{extF2}} [6] = 0.71$; $Q^2_{\text{extF3}} [7] = 0.64$;
CCCEX=0.87; RMSE= 1.02

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 meaningful training set and representative prediction set.

Training and prediction sets are balanced according to both structure and response. In particular, for response the range of logVP values are [-11.56 / 4.85] [-10.27 / -5.54], respectively for SOM training and prediction sets.

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

ATSc5: SOM Split training set (-0.20 / 0.25), prediction set (-0.09 / 0.26);

n6Ring: SOM Split training set (0 / 2), prediction set (0 / 2);

minwHBa: SOM Split training set (-1.11 / 1.53), prediction set (-0.94 / 1.28);

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 Bhatarai and Gramatica [8] was :

$$\log VP = 17.30 - 15.67 \text{BELp2} + 0.44 \text{RBN} + 1.38 \text{B09(N-Cl)}$$

where

BELp2: lowest eigenvalue n. 2 of Burden matrix / weighted by atomic polarizabilities

RBN: number of rotatable bonds

B09(N-Cl): presence/absence of N-Cl at topological distance 09

The polar properties of the (benzo)triazoles reduce their VP, whereas rotatable bonds, which provide flexibility, and chemicals with N-Cl linkage increase it.

In the new PaDEL-descriptor model equation: $\log VP = -6.78 + 12.00 \text{ATSc5} - 1.24 \text{n6Ring} + 0.89 \text{minwHBa}$

ATSc5 is ATS autocorrelation descriptor, weighted by charges

n6Ring is the Number of 6-membered rings

minwHBa is Minimum E-States for weak Hydrogen Bond acceptors

8.3. Other information about the mechanistic interpretation:

no other information available

9. Miscellaneous information

9.1. Comments:

To predict VP for new (B)TAZs chemicals without experimental data, it is suggested to apply the equation of the Full Model, developed on all the available chemicals (N=33), thus ensuring a wider applicability domain.

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

Full model equation: $\log VP = -6.78 + 12.00 \text{ ATSc5} - 1.24 \text{ n6Ring} + 0.89 \text{ minwHBa}$

$N = 33$; $R^2 = 0.80$; $Q^2 = 0.73$; $Q^2_{LMO} = 0.71$; $CCC = 0.89$; $CCC_{cv} = 0.86$; $RMSE = 0.79$; $RMSE_{cv} = 0.91$.

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] ChemID plus database <http://chem.sis.nlm.nih.gov/chemidplus/>

[3] Syracuse Research Center (SRC)

[4] 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

[5] Shi L.M. et al. QSAR Models Using a Large Diverse Set of Estrogens, *J. Chem. Inf. Comput. Sci.* 41 (2001) 186–195.

[6] 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.

[7] Consonni V. et al. Comments on the Definition of the Q2 Parameter for QSAR Validation, *J. Chem. Inf. Model.* 49 (2009) 1669-1678.

[8] B. Bhattarai and P. Gramatica, 2011. Modelling physico-chemical properties of (benzo)triazoles, and screening for environmental partitioning. *Water Res.* 45, 1463-1471.

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