



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

1. QSAR identifier

1.1. QSAR identifier (title):

Insubria QSPR PaDEL-Descriptor model for MP 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:

2/12/2013

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 1.1. Melting point

3.3. Comment on endpoint:

Melting Point (MP) is the temperature at which the solid melts to become a liquid.

3.4. Endpoint units:

°C

3.5. Dependent variable:

MP

3.6. Experimental protocol:

Experimentally measured MP for 56 (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:

MP (SOM split model)

OLS-MLR method. Model developed on a training set of 35 compounds

MP (Ordered response split model)

OLS-MLR method. Model developed on a training set of 38 compounds

MP (Full model)

OLS-MLR method. Model developed on a training set of 56 compounds

SOM Split Model: $MP = -59.09 + 207.03 \text{ minHBd} + 29.69 \text{ nTRing} + 68.79 \text{ Pubchem FP572} + 15.05 \text{ minssssC}$

Ordered Response Split Model: $MP = -49.65 + 195.82 \text{ minHBd} + 27.57 \text{ nTRing} + 71.57 \text{ Pubchem FP572} + 15.37 \text{ minssssC}$

Full Model: $MP = -41.2 + 201.41 \text{ minHBd} + 27.43 \text{ nTRing} + 63.11 \text{ Pubchem FP572} + 16.60 \text{ minssssC}$

4.3. Descriptors in the model:

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

[2]nTRing Number of rings (includes counts from fused rings)

[3]minssssC Minimum atom-type E-State: >C<

[4]Pubchem FP572 N:C:N:C; These bits test for the presence of simple SMARTS patterns, regardless of count, but where bond orders are specific and bond aromaticity matches both single and double bonds.

4.4.Descriptor selection:

A total of 1649 molecular descriptors of differing types (0D, 1D, 2D, 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 206 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 (four 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: 35 chemicals / 4 descriptor = 8.75

Ordered Response Split Model: 38 chemicals / 4 descriptor = 9.5

Full Model: 56 chemicals / 4 descriptor = 14

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 MP values: 3.5 / 300

Range of descriptor values: minHBd: 0 / 0.68; nTRing: 1 / 7; Pubchem FP572: 0 / 1; minssssC: -5.68 / 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 ($h^*=0.268$). HAT values are calculated as the diagonal elements of the HAT matrix:

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

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:

SOM Split model domain: outliers for structure, $hat > 0.429$ (h^*): tetraconazole (112281-77-3), Flupoxam (119126-15-7). Outliers for response, standardised residuals > 2.5 standard deviation units: triazophos (24017-47-8), 8-azaadenine (1123-54-2). **Ordered Response Split model domain:** outliers for structure, $hat > 0.395$ (h^*): tetraconazole (112281-77-3), Flupoxam (119126-15-7). Outliers for response, standardised residuals > 2.5 standard deviation units: triazophos (24017-47-8), 8-azaadenine (1123-54-2). **FULL model domain:** outliers for structure, $hat > 0.268$ (h^*): tetraconazole (112281-77-3), Flupoxam (119126-15-7). Outliers for response, standardised residuals $>$

2.5 standard deviation units: triazophos (24017-47-8), 8-azaadenine (1123-54-2), 1H-Benzotriazole (95-14-7).

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=56) was split, before model development, into a training set used for model development and a prediction set used later for external validation. Two different splitting techniques were applied: by SOM (n training=35) and by Ordered response (n training=38).

6.6. Pre-processing of data before modelling:

The data was taken and used as MP °C.

6.7. Statistics for goodness-of-fit:

SOM Split model:

$R^2 = 0.82$; CCCTr [4]=0.90; RMSE= 29.68

Ordered response split model:

$R^2 = 0.80$; CCCTr=0.89; RMSE= 28.99

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

SOM Split model:

$Q^2_{LOO} = 0.76$; CCCcv=0.87; RMSEcv= 34.20

Ordered response Split model:

$Q^2_{LOO} = 0.74$; CCCcv=0.86; RMSEcv= 33.02

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

SOM Split model:

$Q^2_{LMO} = 0.76$.

Ordered response split model:

$Q^2_{LMO} = 0.73$.

6.10. Robustness - Statistics obtained by Y-scrambling:

SOM Split model:

$R^2_{y-sc} = 0.12$

Ordered response split model:

$R^2_{y-sc} = 0.11$

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 Models consists of 21 compounds in the case of SOM, with a range of MP: 25 / 217, and 18 compounds in the case of Ordered Response, with the same range of MP: 25 / 239

7.6. Experimental design of test set:

The splitting of the original data set (56 compounds) into two training sets of 35 and 38 compounds (representative of the entire data set) and a validation set of 21 and 18 compounds was realized by applying Self Organized Maps Kohonen Artificial Neural Networks (SOM) and by Sorted response (Ordered response).

7.7. Predictivity - Statistics obtained by external validation:

SOM Split model:

Q^2_{extF1} [5] = 0.77; Q^2_{extF2} [6] = 0.70; Q^2_{extF3} [7] = 0.88; CCC_{ex} = 0.84; RMSE = 24.46

Ordered response split model:

Q^2_{extF1} = 0.81; Q^2_{extF2} = 0.81; Q^2_{extF3} = 0.86; CCC_{ex} = 0.90; RMSE = 24.15

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) and by Ordered response allowed for the selection of meaningful training sets and representative prediction sets.

Training and prediction sets are balanced according to both structure and response. In particular, for response the range of MP values are [3.5 / 300][25 / 217] and [3.5 / 300][25 / 239] respectively for SOM and Ordered Response training and prediction sets.

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

minHBd: **SOM Split** training set (0 / 0.68), prediction set (0 / 0.67);
Ordered response split training set (0 / 0.68), prediction set (0 / 0.67)

nTRing: **SOM Split** training set (1 / 7), prediction set (1 / 6); **Ordered response split** training set (1 / 7), prediction set (1 / 7)

Pubchem FP572: **SOM Split** training set (0 / 1), prediction set (0 / 1); **Ordered response split** training set (0 / 1), prediction set (0 / 1)

minssssC: **SOM Split** training set (-4.45 / 0), prediction set (-5.69 / 0.28); **Ordered response split** training set (-4.45 / 0), prediction set (-5.69 / 0.28)

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:

$$\text{MP} = 109.25 - 162.83 \text{R2e} + 53.22 \text{GGI4} + 26.82 \text{F03(N-N)} - 1693.0 \text{X1A}$$

Where R2e is R autocorrelation of lag 2 / weighted by atomic Sanderson Electronegativities.

It is a 3D GETAWAY descriptor that encodes geometrical information given by the influence matrix, topological information given by the molecular graph, and chemical information from electronegativity.

GGI4 is a topological charge index-based 2D descriptor, used to evaluate charge transfer between atom pairs and, therefore, the global charge transfer in the molecule.

F03[N-N] is a 2D frequency fingerprint descriptor that corresponds to the frequency of N-N at a topological distance 3.

X1A is a 2D topological descriptor encoding the average Randic connectivity index c1, which describes molecular branching and complexity.

The electronegativity and the branching and complexity of the molecule thus seem to have a negative influence on the MP of (benzo) triazoles, while other properties like charge transfer and the presence of N-N atoms are positively correlated to MP.

The equation of the new PaDEL-descriptor model included in QSARINS is

:

$$MP = -41.2 + 201.41 \text{ minHBd} + 27.43 \text{ nTRing} + 63.11 \text{ Pubchem FP572} + 16.60 \text{ minssssC}$$

where minHBd is Minimum E-States for (strong) Hydrogen Bond donors
nTRing is the Number of rings (includes counts from fused rings)
minssssC is Minimum atom-type E-State: >C<
Pubchem FP572 is a fingerprint that test for the presence of N:C:N:C

The correlation among DRAGON X1A and PaDEL descriptor nTRing is -0.72.

8.3. Other information about the mechanistic interpretation:

no other information available

9. Miscellaneous information

9.1. Comments:

To predict MP 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=56), thus ensuring a wider applicability domain.

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

$$MP = -41.2 + 201.41 \text{ minHBd} + 27.43 \text{ nTRing} + 63.11 \text{ Pubchem FP572} + 16.60 \text{ minssssC}$$

$$N = 56; R^2 = 0.81; Q^2 = 0.77; Q^2_{LMO} = 0.77; CCC = 0.90; CCC_{cv} = 0.88; RMSE = 27.14; RMSE_{cv} = 29.59$$

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. Bhatarai 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