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

1.QSAR identifier

1.1.QSAR identifier (title):

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

1.2.Other related models:

B. Bhhatarai 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 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 1.6.Octanol-water partition coefficient (Kow)

3.3.Comment on endpoint:

The octanol-water partition coefficient (Kow) is definied as the ratio of a chemical's concentration in the octanol phase to its concentration in the aqueous phase of a two-phase octanol/water system. Kow values are reported in literature in Log units (LogKow).

3.4.Endpoint units:

Dimensionless

3.5.Dependent variable:

LogKow

3.6.Experimental protocol:

Experimentally measured LogKow for 64 (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:

logKow (SOM split model)
OLS-MLR method. Model developed on a training set of 48 compounds

logKow (Ordered Response split model)
OLS-MLR method. Model developed on a training set of 48 compounds

logKow (Full model)
OLS-MLR method. Model developed on a training set of 64 compounds

SOM Split Model: logKow= 1.21 + 0.02 MW - 0.55 nHBAcc + 0.25 MDEC-12 - 0.25 nN

Ordered Response Split Model: logKow= 1.33 + 0.01 MW - 0.48 nHBAcc + 0.22 MDEC-12 - 0.29 nN

Full Model: logKow= 1.45 + 0.01 MW - 0.54 nHBAcc + 0.24 MDEC-12 - 0.26 nN

4.3.Descriptors in the model:

[1]MW Molecular Weight

[2]nHBAcc Number of hydrogen bond acceptors (using CDK HBondAcceptorCountDescriptor algorithm)

[3]MDEC-12 Molecular distance edge between all primary and secondary carbons

[4]nN Number of nitrogens

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 288 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 Q^2LOO (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: 48 chemicals / 4 descriptor = 12
Ordered Response Split Model: 48 chemicals / 4 descriptor = 12
Full Model: 64 chemicals / 4 descriptor = 16

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 logKow values: -1.97 / 5.3
Range of descriptor values: MW: 69.03 / 437.19; nHBAcc: 3 / 9; MDEC-12: 0 / 5.41; nN: 3 / 6

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*=0.234). 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:

SOM Split model domain: outliers for structure, hat>0.313 (h*): Ribavirin (36791-04-5). Outliers for response, standardised residuals > 2.5 standard deviation units: 3-amino-5-(2-(ethylamino)-4-pyridyl)-1,2,4-triazole (77314-77-3); Drometrizole (2440-22-4). **Ordered Response Split model domain**: outliers for structure, hat>0.313 (h*): Ribavirin (36791-04-5). Outliers for response, standardised residuals > 2.5 standard deviation units: 3-amino-5-(2-(ethylamino)-4-pyridyl)-

1,2,4-triazole (77314-77-3); Drometrizole (2440-22-4); N-(2-Hydroxyethyl)-2-(3-nitro-1,2,4-triazol-1-yl)acetamide (104958-85-2). **FULL model domain**: outliers for structure, hat>0.234 (h*): Ribavirin (36791-04-5). Outliers for response, standardised residuals > 2.5 standard deviation units: 3-amino-5-(2-(ethylamino)-4-pyridyl)-1,2,4-triazole (77314-77-3); Drometrizole (2440-22-4).

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=64) 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=48) and by Ordered response (n training=48).

6.6.Pre-processing of data before modelling:

The data was taken and used as LogKow.

6.7.Statistics for goodness-of-fit:

SOM Split model:

R^2= 0.87; CCCtr [4]=0.93; RMSE= 0.69

Ordered response split model:

R^2= 0.85; CCCtr=0.92; RMSE= 0.68

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

SOM Split model:

Q^2LOO= 0.83; CCCcv=0.91; RMSEcv= 0.78

Ordered response Split model:

Q^2LOO= 0.79; CCCcv=0.89; RMSEcv= 0.80

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

SOM Split model:

Q^2LMO= 0.83.

Ordered response split model:

Q^2LMO= 0.81.

6.10.Robustness - Statistics obtained by Y-scrambling:

SOM Split model:

R^2y-sc= 0.09

Ordered response split model:

R^2y-sc= 0.09

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 both Split Models consists of 16 compounds, with a range of logKow: -0.96 / 3.7 (SOM) and -1.97 / 4.3 (Ordered response)

7.6.Experimental design of test set:

The splitting of the original data set (64 compounds) into two training sets of 48 compounds (representative of the entire data set) and a validation set of 16 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^2extF1 [5]= 0.79; Q^2extF2 [6]= 0.78; Q^2extF3 [7]= 0.89; CCCex=0.92; RMSE= 0.61

Ordered response split model:

Q^2extF1= 0.88; Q^2extF2= 0.88; Q^2extF3= 0.88; CCCex=0.93; RMSE= 0.61

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 logKow values are [-1.97 / 5.3][-0.96 / 3.7] and [-1.85 / 5.3][-1.97 / 4.03] respectively for SOM and Ordered Response training and prediction sets.

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

MW: **SOM Split** training set (69.03 / 437.18), prediction set (69.03 / 391.95); **Ordered response split** training set (69.03 / 437.18), prediction set (69.03 / 385.98)

nHBAcc: **SOM Split** training set (3 / 9), prediction set (3 / 6); **Ordered response split** training set (3 / 9), prediction set (3 / 7)

MDEC-12: **SOM Split** training set (0 / 5.17), prediction set (0 / 5.41); **Ordered response split** training set (0 / 5.41), prediction set (0 / 3.14)

nN: **SOM Split** training set (3 / 6), prediction set (3 / 6); **Ordered response split** training set (3 / 6), prediction set (3 / 6)

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 Bhhatarai and Gramatica [8] was :

logKow= 1.71 + 2.80 B08(C-C) - 0.69 nN +1.16 GATS3m + 1.53 MATS1v

where: B08(C-C) is a binary fingerprint based 2D descriptor that takes into account the presence of C-C (carbon carbon single bond) at a topological distance 8

nN is the number of nitrogen atoms

GATS3m is Geary autocorrelation - lag 3 / weighted by atomic masses that describes the spatial autocorrelation of atomic masses

MATS1v is Moran autocorrelation - lag 1 / weighted by atomic van der Waals volumes.

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

logKow= 1.45 + 0.01 MW - 0.54 nHBAcc + 0.24 MDEC-12 - 0.26 nN

where: MW is Molecular weight

nHBAcc is Number of hydrogen bond acceptors (using CDK

HBondAcceptorCountDescriptor algorithm)

nN is number of nitrogens MDEC-12 is Molecular distance edge between all primary and secondary carbons

The DRAGON descriptor B08[C-C] shows an high correlation with MW (0.87). The bigger chemicals (higher MW and MDEC-12 values) and those with lower tendency to form hydrogen bonds with water (represented by nN and nHAcc), show higher tendency to partitioning in octanol than in water.

8.3.Other information about the mechanistic interpretation:

no other information available

9.Miscellaneous information

9.1.Comments:

To predict logKow 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=64), thus ensuring a wider applicability domain.

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

$$logKow = 1.45 + 0.01\ MW - 0.54\ nHBAcc + 0.24\ MDEC\text{-}12 - 0.26\ nN$$

$N = 64$; $R^2 = 0.86$; $Q^2 = 0.83$; $Q^2LMO = 0.83$; $CCC = 0.92$; $CCCcv = 0.91$; $RMSE = 0.66$; $RMSEcv = 0.73$

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. Bhhatarai 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