	QMRF identifier (JRC Inventory): *To be entered by JRC*
	QMRF Title: *Insubria QSAR PaDEL-Descriptor model for prediction of (Benzo-)Triazoles toxicity in O.mykiss*
	Printing Date: *Jan 20, 2014*

1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSAR PaDEL-Descriptor model for prediction of (Benzo-)Triazoles toxicity in *O.mykiss*

1.2.Other related models:

S.Cassani, S. Kovarich, E.Papa, P.P. Roy, L.v.d.Wal, P. Gramatica, 2013. Daphnia and fish toxicity of (benzo)triazoles: Validated QSAR models,and interspecies quantitative activity–activity modelling, J.Haz. Mater. 258-259, 50-60. [7]

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:

4/12/2013

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
[3]Alessandro Sangion DiSTA, University of Insubria (Varese - Italy) a.sangion@hotmail.it www.qsar.it

2.6.Date of model development and/or publication:

November 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:

Oncorhynchus mykiss

3.2.Endpoint:

3.Ecotoxic effects 3.3.Acute toxicity to fish (lethality)

3.3.Comment on endpoint:

A selected set of experimental 96h LC50 data was taken from FOOTPRINT PPDB (Pesticide Properties DataBase) online database [2].

3.4.Endpoint units:

The median lethal concentrations are reported as the logarithm of the inverse molar concentration: log(1/LC50)

3.5.Dependent variable:

log (1/LC50) or pLC50

3.6.Experimental protocol:

OECD 203 test protocol

3.7.Endpoint data quality and variability:

The data classified as "verified data", "verified data used for regulatory purposes" and "unverified data from known source" were included in model development. The classification of "verified data", "verified data used for regulatory purposes" and "unverified data from known source" was given directly by the FOOTPRINT PPDB database.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR - Multiple linear regression model (OLS - Ordinary Least Square)

4.2.Explicit algorithm:

Log 1/LC50 O.mykiss SOM Split model
OLS-MLR method. Model developed on a training set of 52 compounds

Log 1/LC50 O.mykiss Ordered response split model
OLS-MLR method. Model developed on a training set of 52 compounds

Log 1/LC50 O.mykiss FULL model
OLS-MLR method. Model developed on a training set of 75 compounds

SOM Split model equation: pLC50= 3.49 + 0.32 VP-1 - 0.18 nHBAcc - 1.23 minHBd

Ordered Response Split model equation: pLC50: 3.48 + 0.32 VP-1 - 0.18 nHBAcc - 1.11 minHBd

Full model equation: pLC50= 3.62 + 0.30 VP-1 – 0.17 nHBAcc – 1.20 minHBd

4.3.Descriptors in the model:

[1]VP-1 Valence path, order 1

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

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

4.4.Descriptor selection:

A total of 729 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 227 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:

Split by SOM model: 52 chemicals / 3 descriptors = 17.33

Split by Ordered response model: 52 chemicals / 3 descriptors= 17.33

Full model: 75 chemicals / 3 descriptors = 25

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 pLC50 O.mykiss values: 1.92 / 6.7

Range of descriptor values: VP-1: 1.51 / 10; nHBAcc: 1 / 13; minHBd: 0 / 0.80.

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*=0.160). 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>0.231 (h*): azimsulfuron (120162-55-2). Outliers for response, standardised residuals > 2.5 standard deviation units: fenchlorazole-ethyl (103112-35-2), fluroxypyr methylheptyl ester (81406-77-3). **Ordered**

Response Split model domain: outliers for structure, hat>0.231 (h*):

azimsulfuron (120162-55-2). Outliers for response, standardised residuals > 2.5 standard deviation units: fenchlorazole-ethyl (103112-35-2), fluroxypyr methylheptyl ester (81406-77-3), mepanipyrim (110235-47-7). **FULL model domain**: outliers for structure, hat>0.160 (h*): no. Outliers for response, standardised residuals > 2.5 standard deviation units: fenchlorazole-ethyl (103112-35-2), fluroxypyr methylheptyl ester (81406-77-3), mepanipyrim (110235-47-7).

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=75) 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 structural similarity (SOM) and by ordered response (n training=52 in both cases).

6.6.Pre-processing of data before modelling:

Transformation of LC50 (mg/L) into Log1/LC50 (mol/L)

6.7.Statistics for goodness-of-fit:

SOM Split model:

R^2= 0.76; CCCtr [3]=0.86; RMSE= 0.54

Ordered response split model:

R^2= 0.76; CCCtr=0.87; RMSE= 0.52

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

SOM Split model:

Q^2LOO= 0.71; CCCcv=0.84; RMSEcv= 0.59

Ordered response Split model:

Q^2LOO= 0.72; CCCcv=0.84; RMSEcv= 0.56

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

SOM Split model:

Q^2LMO= 0.73.

Ordered response split model:

Q^2LMO= 0.73.

6.10.Robustness - Statistics obtained by Y-scrambling:

SOM Split model:

R^2y-sc= 0.06

Ordered response split model:

R^2y-sc= 0.06

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:

To verify the predictive capability of the proposed models, the dataset (n=75) 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 structural similarity (SOM) and by sorted response (n external validation set =23 in both cases); the range of pLC50 are: 2.85 / 5.99 for SOM prediction set, 2.85 / 6.26 for Ordered Response prediction set.

7.6.Experimental design of test set:

In the case of split by sorted response model, chemicals were ordered according to their increasing activity, and one out of every three chemicals was put in the prediction set (always including the most and the least active compounds in the training set). The splitting based on structural similarity (SOM) takes advantages of the clustering capabilities of Kohonen Artifical Neural Network, allowing the selection of a structurally meaningful training set and an equally representative prediction set.

7.7.Predictivity - Statistics obtained by external validation:

SOM Split model:

Q^2extF1 [4]= 0.76; Q^2extF2 [5]= 0.75; Q^2extF3 [6]= 0.83; CCCex=0.88; RMSE= 0.46

Ordered response split model:

Q^2extF1= 0.74; Q^2extF2= 0.74; Q^2extF3= 0.77; CCCex=0.86; RMSE= 0.51

7.8.Predictivity - Assessment of the external validation set:

The splitting methodology based on similarity analysis 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 pLC50 values are [1.92 / 6.7][2.85 / 5.99] and [1.92 / 5.99][2.85 / 6.7] respectively for SOM and Ordered Response training and prediction sets.

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

VP-1: SOM Split training set (1.51 / 10), prediction set (2.74 / 9.61); Ordered response split training set (1.51 / 10), prediction set (2.74 / 9.61)

nHBAcc: SOM Split training set (3 / 12), prediction set (1 / 13); Ordered response split training set (2 / 13), prediction set (1 / 10)

minHBd: SOM Split training set (0 / 0.78), prediction set (0 / 0.80); Ordered response split training set (0 / 0.80), prediction set (0 / 0.78)

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 PaDEL-descriptor model equation published in Cassani et al. [7] is: $pLC50 = 2.325 + 0.392VP\text{-}1 - 0.049SHBint2 - 0.335maxHaaCH$

where

VP-1= Valence path, order 1

SHBint2= Sum of E-State descriptors of strength for potential Hydrogen Bonds of path length 2

maxHaaCH=Maximum atom-type H E-State: :CH:

VP-1 is related to molecular dimension and branching, and increasing values of this variable are related to an increase in the observed toxicity values. Additionally, being slightly correlated with LogP, VP-1 also encodes for hydrophobic properties of molecules. SHBint2 representing the hydrogen bond acceptor or donor strength of the molecule and is related to the chemicals tendency to form hydrogen bonds with water. maxHaaCH describes the maximum energy associated to the

presence of the CH aromatic fragment in the molecule.

In the new PaDEL-descriptor model equation: pLC50= 3.62 + 0.30 VP-1 – 0.17 nHBAcc – 1.20 minHBd

VP-1 is Valence path, order 1, which is related to molecular dimension and branching; generally high values of this descriptor are connected to high/medium toxicity (correlation with pLC50 0.64).

nHBAcc is Number of hydrogen bond acceptors (using CDK HBondAcceptorCountDescriptor algorithm)

minHBd is Minimum E-States for (strong) Hydrogen Bond donors

SHBint2 in the previous model has been now substituted by nHBAcc (correlation: 0.81)

8.3.Other information about the mechanistic interpretation:

no other information available

9.Miscellaneous information

9.1.Comments:

To predict pLC50 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=75), thus ensuring a wider applicability domain. The full model was then applied to a further Evaluation Set composed of 18 (B)TAZs, became available in the literature. The equation (reported also in section 4.2) and the statistical parameters of the full model, also when applied to the EV, are the following:

pLC50= 3.62 + 0.30 VP-1 - 0.17 nHBAcc - 1.20 minHBd

N = 75 (*full model applied to the EV, N=18); R2 = 0.76; Q2 = 0.73; Q2LMO = 0.74; CCC = 0.86; CCCcv = 0.85; RMSE= 0.51; RMSEcv = 0.54; RMSEex*= 0.51; CCCex*= 0.86; Q2extF1=0.78; Q2extF2=0.78; Q2extF3=0.77.

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]FOOTPRINT PPDB (Pesticide Properties DataBase), 2009 http://sitem.herts.ac.uk/aeru/footprint/en/

[3]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

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

[5]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.
[6]Consonni V. et al. Comments on the Definition of the Q2 Parameter for QSAR Validation, J. Chem. Inf. Model. 49 (2009) 1669-1678
[7]S.Cassani, S. Kovarich, E.Papa, P.P. Roy, L.v.d.Wal, P. Gramatica, 2013. Daphnia and fish toxicity of (benzo)triazoles: Validated QSAR models,and interspecies quantitative activity–activity modelling, J.Haz. Mater. 258-259, 50-60.

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