1.QSAR identifier

1.1.QSAR identifier (title):

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

1.2.Other related models:

P. Gramatica, S.Cassani, P.P. Roy, S. Kovarich, C.W.Yap, E.Papa, 2012. QSAR Modeling is not "Push a Button and Find a Correlation": A Case Study of Toxicity of (Benzo-)triazoles on Algae, Mol. Inf. 31, 817-835. [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:

04/12/2013

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:

September 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:

Pseudokirchneriella subcapitata

3.2.Endpoint:

3.Ecotoxic effects 3.2.Short-term toxicity to algae (inhibition of the exponential growth rate)

3.3.Comment on endpoint:

A selected set of experimental 72h EC50 data was taken from FOOTPRINT PPDB (Pesticide Properties DataBase) online database [2].

3.4.Endpoint units:

The median effect concentrations are reported as the logarithm of the inverse molar concentration: log(1/EC50)

3.5.Dependent variable:

log (1/EC50) or pEC50

3.6.Experimental protocol:

OECD 201 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/EC50 P.subcapitata SOM Split model
OLS-MLR method. Model developed on a training set of 22 compounds

Log 1/EC50 P.subcapitata Ordered response split model
OLS-MLR method. Model developed on a training set of 24 compounds

Log 1/EC50 P.subcapitata FULL model
OLS-MLR method. Model developed on a training set of 35 compounds

SOM Split model equation: pEC50= 2.38 + 0.07 SwHBa + 0.44 MDEN-22 + 0.04 WPOL

Ordered Response Split model equation: pEC50= 2.29 + 0.06 SwHBa + 0.05 WPOL + 0.46 MDEN-22

Full model equation: pEC50= 2.43 + 0.07 SwHBa + 0.04 WPOL + 0.45 MDEN-22

4.3.Descriptors in the model:

[1]SwHBa Sum of E-States for weak Hydrogen Bond acceptors
[2]WPOL Weiner polarity number
[3]MDEN-22 Molecular distance edge between all secondary nitrogens

4.4.Descriptor selection:

A total of 721 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 222 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: 22 chemicals / 3 descriptors = 7.33
Split by Ordered response model: 24 chemicals / 3 descriptors= 8

Full model: 35 chemicals / 3 descriptors = 11.67

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 pEC50 P.subcapitata values: 3.09 / 6.72
Range of descriptor values: SwHBa: -4.98 / 28.21; WPOL: 0 / 51; MDEN-22: 0 / 4.06.

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*=0.343). 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>0.545 (h*): no. Outliers for response, standardised residuals > 2.5 standard deviation units: Pyroxsulam (422556-08-9), chlorsulfuron (64902-72-3). **Ordered**

Response Split model domain: outliers for structure, hat>0.500 (h*): no. Outliers for response, standardised residuals > 2.5 standard deviation units: Pyroxsulam (422556-08-9), chlorsulfuron (64902-72-3). **FULL**

model domain: outliers for structure, hat>0.343 (h*):

Hydroxyterbuthylazine (66753-07-9). 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:

To verify the predictive capability of the proposed models, the dataset (n=35) 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, n training= 24) and by ordered response (n training=22).

6.6.Pre-processing of data before modelling:

Transformation of EC50 (mg/L) into Log1/EC50 (mol/L)

6.7.Statistics for goodness-of-fit:

SOM Split model:

R^2= 0.83; CCCtr [3]=0.91; RMSE= 0.45

Ordered response split model:

R^2= 0.84; CCCtr=0.91; RMSE= 0.42

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

SOM Split model:

Q^2LOO= 0.73; CCCcv=0.85; RMSEcv= 0.56

Ordered response Split model:

Q^2LOO= 0.74; CCCcv=0.86; RMSEcv= 0.52

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

SOM Split model:

Q^2LMO= 0.72.

Ordered response split model:

Q^2LMO= 0.74.

6.10.Robustness - Statistics obtained by Y-scrambling:

SOM Split model:

R^2y-sc= 0.15

Ordered response split model:

R^2y-sc= 0.13

6.11.Robustness - Statistics obtained by bootstrap:

No information available (since we have calculated Q2LMO)

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=35) 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 (n external validation set =13) and sorted response (n external validation set =11); the range of pEC50 are: 3.1 / 6 for SOM prediction set, 3.36 / 6.12 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]= 0.79; Q^2extF2 [5]= 0.79; Q^2extF3 [6]= 0.88; CCCex=0.89; RMSE= 0.37

Ordered response split model:

Q^2extF1= 0.76; Q^2extF2= 0.76; Q^2extF3= 0.81; CCCex=0.88; RMSE= 0.44

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 pEC50 values are [3.09 / 6.72] [3.1 / 6] and [3.09 / 6.72][3.36 / 6.12] respectively for SOM and Ordered Response training and prediction sets.

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

SwHBa: SOM Split training set (-4.98 / 21.37), prediction set (0.125 / 28.21); Ordered response split training set (-2.94 / 28.21), prediction set (-4.98 / 18.6)

WPOL: SOM Split training set (9 / 51), prediction set (0 / 34); Ordered response split training set (9 / 50), prediction set (0 / 51)

MDEN-22: SOM Split training set (0 / 4.06), prediction set (0.25 / 1.89); Ordered response split training set (0 / 4.06), prediction set (0.25 / 3.15)

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 Gramatica et al. [7] was

$$pEC50= 1.5051 + 0.0269\ AMR + 0.4322\ MDEN\text{-}22 +0.472\ maxwHBa$$

where

AMR is Molar refractivity

MDEN-22 is Molecular distance edge between all secondary nitrogens

maxwHBa is Maximum E-States for weak Hydrogen Bond acceptors

IThe equation of the new PaDEL-descriptor model included in QSARINS is: $pEC50= 2.43 + 0.07\ SwHBa + 0.04\ WPOL + 0.45\ MDEN\text{-}22$

where

SwHBa is Sum of E-States for weak Hydrogen Bond acceptors

WPOL is Weiner polarity number MDEN-22 is Molecular distance edge between all secondary nitrogens

The correlation between AMR and WPOL is 0.90, therefore these descriptors encode for the same structural feature related to the modeled toxicity; the correlation between maxwHBa and SwHBa is 0.81.

8.3.Other information about the mechanistic interpretation:

no other information available

9.Miscellaneous information

9.1.Comments:

To predict pEC50 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=35), thus ensuring a wider applicability domain. The full model equation (reported also in section 4.2) and the statistical parameters are the following:

pEC50= 2.43 + 0.07 SwHBa + 0.04 WPOL + 0.45 MDEN-22

N = 35; R2 = 0.82; Q2 = 0.76; Q2LMO = 0.76; CCC = 0.90; CCCcv = 0.87; RMSE= 0.42; RMSEcv = 0.49.

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]P. Gramatica, S.Cassani, P.P. Roy, S. Kovarich, C.W.Yap, E.Papa, 2012. QSAR Modeling is not “Push a Button and Find a Correlation”: A Case Study of Toxicity of (Benzo-)triazoles on Algae, Mol. Inf. 31, 817-835.

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