

	QMRF identifier (JRC Inventory): To be entered by JRC
	QMRF Title: Insubria QSAR PaDEL-Descriptor model for prediction of (Benzo-)Triazoles toxicity in <i>P.subcapitata</i>
	Keywords: PaDEL-Descriptor; GA-OLS; Algae toxicity EC50; (Benzo-)Triazoles; QSARINS; INSUBRIA
	Printing Date: 9-mar-2015

1. QSAR identifier

1.1. QSAR identifier (title):

Insubria QSAR PaDEL-Descriptor model for prediction of (Benzo-)Triazoles toxicity in *P.subcapitata*
 Keywords: PaDEL-Descriptor; GA-OLS; Algae toxicity EC50; (Benzo-)Triazoles; QSARINS; INSUBRIA

1.2. Other related models:

Gramatica, P., Cassani, S., Roy, PP., Kovarich, S. Yap, CW., Papa, E.
 QSAR Modeling is not "Push a Button and Find a Correlation": A Case Study of Toxicity of (Benzo-)triazoles on Algae, *Mol. Inf.* 2012, 31, 817-835. [1]

1.3. Software coding the model:

PaDEL-Descriptor
 A software to calculate molecular descriptors and fingerprints, version 2.18 [2]
 Yap Chun Wei, email: phayapc@nus.edu.sg
<http://padel.nus.edu.sg/software/padeldescriptor/index.html>

QSARINS

Software for the development, analysis and validation of QSAR MLR models [3,4], version 1.2 (verified also with 2.2, 2015)
 Prof. Paola Gramatica, paola.gramatica@uninsubria.it
<http://www.qsar.it/>

2. General information

2.1. Date of QMRF:

2/2/2015

2.2. QMRF author(s) and contact details:

[1] Stefano Cassani Insubria University, Department of Theoretical and Applied Sciences (DiSTA), via J.H. Dunant 3, 21100 Varese (Italy) +390332421439 stefano.cassani@uninsubria.it
<http://www.qsar.it/>
 [2] Paola Gramatica Insubria University, Department of Theoretical and Applied Sciences (DiSTA), via J.H. Dunant 3, 21100 Varese (Italy) +390332421573 paola.gramatica@uninsubria.it
<http://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 Insubria University, Department of Theoretical and Applied Sciences (DiSTA), via J.H. Dunant 3, 21100 Varese (Italy) +390332421439 stefano.cassani@uninsubria.it
<http://www.qsar.it/>

[2]Paola Gramatica Insubria University, Department of Theoretical and Applied Sciences (DiSTA), via J.H. Dunant 3, 21100 Varese (Italy) +390332421573 paola.gramatica@uninsubria.it
<http://www.qsar.it/>

2.6.Date of model development and/or publication:

Developed in 2013, Published in 2014 [4]

2.7.Reference(s) to main scientific papers and/or software package:

- [1]Gramatica, P., Cassani, S., Roy, PP., Kovarich, S. Yap, CW., Papa, E. QSAR Modeling is not "Push a Button and Find a Correlation": A Case Study of Toxicity of (Benzo-)triazoles on Algae, Mol. Inf. 2012, 31, 817-835. DOI: 10.1002/minf.201200075
- [2]Yap, C.W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. 2011, J.Comput.Chem. 32, 1466-1474 DOI: 10.1002/jcc.21707
- [3]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. DOI: 10.1002/jcc.23361
- [4]Gramatica P., et al. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS, J. Comput. Chem. (Software News and Updates), 2014, 35 (13), 1036-1044. DOI: 10.1002/jcc.23576

2.8.Availability of information about the model:

Non-proprietary. Defined algorithm, available in QSARINS [3,4]. Training and prediction sets are available in the attached sdf files of this QMRF (section 9) and in the QSARINS-Chem database [4].

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 EC50 (72h) data for the selected species of green algae was taken from FOOTPRINT PPDB (Pesticide Properties DataBase) online database [5].

3.4.Endpoint units:

The median effect concentrations are reported as the logarithm of the inverse molar concentration: $\log(1/EC50)$ or pEC50 M

3.5.Dependent variable:

$\log(1/EC50)$ or pEC50

3.6.Experimental protocol:

OECD 201 test protocol

3.7.Endpoint data quality and variability:

In FOOTPRINT PPDB, 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

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 (N Training=22): $pEC50 = 2.38 + 0.07 SwHBa + 0.44 MDEN-22 + 0.04 WPOL$

Ordered Response Split model equation (N Training=24): $pEC50 = 2.29 + 0.06 SwHBa + 0.05 WPOL + 0.46 MDEN-22$

Full model equation (N Training=35): $pEC50 = 2.43 + 0.07 SwHBa + 0.04 WPOL + 0.45 MDEN-22$

The modeling descriptors, calculated in PaDEL-Descriptor 2.18, are:

SwHBa, WPOL, MDEN-22. See section 4.3 for a more detailed description of the three molecular descriptors.

4.3. Descriptors in the model:

[1]SwHBa dimensionless Sum of E-States for weak Hydrogen Bond acceptors, descriptor for weak hydrogen bond acceptors. Aromatic and otherwise unsaturated carbons are considered to be weak acceptors. The hydrogen binding ability is therefore represented here by SwHBa, with a positive contribution on toxicity in algae: the higher SwHBa values tend to be toxic molecules

[2]WPOL dimensionless Weiner polarity number, degree of polarity of chemicals: here, WPOL contributes positively towards the toxicity.

[3]MDEN-22 dimensionless Molecular distance edge between all secondary nitrogens, designed to capture the structure and the interaction between the atoms (secondary nitrogens here) in a molecular graph.

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). The GA-VSS, by Ordinary Least Squares regression (OLS), included in QSARINS, was applied to select only the best combination of descriptors from input pool: 3 modeling descriptors selected from 222.

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 7.03. Then, these files were converted by OpenBabel 2.3.2 into MDL-MOL format and used as input for the calculation of descriptors in PaDEL-Descriptor. Any user can re-derive the model calculating the molecular descriptors by PaDEL-Descriptor 2.18 software (included in QSARINS 2.2) and applying the given equation (automatically done by QSARINS 2.2).

4.6. Software name and version for descriptor generation:

PaDEL-Descriptor

A software to calculate molecular descriptors and fingerprints, version 2.18.

Yap Chun Wei, Department of Pharmacy, National University of Singapore.

<http://padel.nus.edu.sg/software/padeldescriptor/index.html>

HYPERCHEM

Software for molecular drawing and conformational energy optimization, version 7.03, 2002.

Phone: (352)371-7744

<http://www.hyper.com/>

OpenBabel

Open Babel: The Open Source Chemistry Toolbox. Used for conversion between HYPERCHEM files (hin) and MDL-MOL files, version 2.3.2.

Contact not available

http://openbabel.org/wiki/Main_Page

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 (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. In QSARINS the Insubria graph allows to identify for which chemicals the predictions are inter- or extrapolated by the model.

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:

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

The response applicability domain can be verified by the standardized residuals in cross-validation greater than 2.5 standard deviation units

5.3. Software name and version for applicability domain assessment:

QSARINS

Software for the development, analysis and validation of QSAR MLR models, version 1.2 (verified also with 2.2, 2015)

Prof. Paola Gramatica; paola.gramatica@uninsubria.it

<http://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: No

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 whole dataset (n=35) was split, before model development, into training sets used for model development and prediction sets used later for external validation. Two different splitting techniques were applied: by structural similarity (Self Organizing Maps, SOM, n training= 24) and by ordered response (n training=22). In the SOM splitting, training and prediction set are structurally balanced, being the splitting based on the structural similarity analysis (performed with Kohonen artificial neural network, K-ANN or SOM method included in KOALA software [11]). In the Ordered response splitting chemicals have been ordered according to their increasing toxicity and one out of every three chemicals has been assigned to the prediction set (always including the most and the least persistent compound in the training set, i.e. the lowest and the highest pEC50). This splitting guarantees that the training set covers the entire range of the modeled response.

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 [6,7]=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:**

$Q^2_{LOO} = 0.73$; CCCcv=0.85; RMSEcv= 0.56

Ordered response Split model:

$Q^2_{LOO} = 0.74$; CCCcv=0.86; RMSEcv= 0.52

6.9.Robustness - Statistics obtained by leave-many-out cross-validation:**SOM Split model:**

$Q^2_{LMO30\%} = 0.72$.

Ordered response split model:

$Q^2_{LMO30\%} = 0.74$. High value of Q^2_{LMO} (average

value for 2000 iterations, with 30% of chemicals put out at every iteration) means that the model is robust and stable.

6.10. Robustness - Statistics obtained by Y-scrambling:

SOM Split model:

$R^2_{y-sc} = 0.15$

Ordered response split model:

$R^2_{y-sc} = 0.13$. Low value of scrambled R^2 (average value for 2000 iterations, in where the Y-responses are randomly scrambled), means that the model is not given by chance-correlation.

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: No

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 (SOM, see section 6.5; n prediction set=13) and ordered response (see section 6.5; n prediction 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 Artificial Neural Network, allowing the selection of a structurally meaningful training set and an equally representative prediction set. Through its clustering capabilities, SOM ensures that both sets are

homogeneously distributed within the entire area of the descriptor space; in this case the chemicals in both sets, selected to maximize the coverage of the descriptor space (i.e. representativity), represent the structural variety of the studied data set in a balanced way. The selected training chemicals are those with the minimal distance from the centroid of each cell in the top map. In this case, the representative points of the prediction set are close (in the same cell of the top map) to representative points of the training set in the multidimensional structural descriptor.

7.7. Predictivity - Statistics obtained by external validation:

SOM Split model:

Q^2_{extF1} [8]= 0.79; Q^2_{extF2} [9]= 0.79; Q^2_{extF3} [10]= 0.88; CCC_{ex}=0.89; RMSE= 0.37

Ordered response split model:

Q^2_{extF1} = 0.76; Q^2_{extF2} = 0.76; Q^2_{extF3} = 0.81; CCC_{ex}=0.88; RMSE= 0.44

The high values of external Q^2 and concordance correlation coefficient-CCC (threshold for accepting the external $Q^2_{\text{F1-F2-F3}}$ is 0.70, threshold for CCC is 0.85, [7]), show that the proposed model is predictive, when applied to chemicals never seen during the model development (prediction sets).

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)

The applicability domain of the model on the prediction set has been verified by the Williams plot: in both split models, no compounds of the prediction sets are outliers for the response or for structure. Even if the prediction sets are small (11 and 14 chemicals), these results are a proof of the good applicability domain of the proposed model.

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:

A posteriori mechanistic interpretation:

The equation of the PaDEL-descriptor full model included in QSARINS 2.2 is: $pEC_{50} = 2.43 + 0.07 SwHBa + 0.04 WPOL + 0.45 MDEN-22$

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 most important descriptor in the proposed model equation is

SwHBA, a descriptor for weak hydrogen bond acceptors (sum of E-State values for all weak hydrogen bond acceptors). Aromatic and otherwise unsaturated carbons are considered to be weak acceptors. The hydrogen binding ability is therefore represented here by SwHBa, with a positive contribution on toxicity in algae: the higher SwHBa values tend to be toxic molecules. The contribution of the descriptor WPOL indicates that the algae toxicity of studied BTAZs is also influenced by the degree of polarity of chemicals: WPOL contributes positively towards the toxicity. MDEN-22 is the less important descriptor, with a positive contribute to the studied toxicity, and belongs to molecular graph-theoretical indices. These descriptors are designed to capture the structure and the interaction between the atoms (secondary nitrogens here) in a molecular graph.

8.3. Other information about the mechanistic interpretation:

no other information available

9. Miscellaneous information

9.1. Comments:

Given the good results of the external validation, this model has a good applicability domain and therefore unsuccessful applications are probably very reduced. Anyhow, the check of outliers by the Williams plot and the Insubria graph for chemicals without experimental data (see section 5.1) will allow to verify the model applicability.

To predict pEC₅₀ for new (Benzo-)Triazoles without experimental data, it is suggested to apply the equation of the **Full Model**,

developed on all the available chemicals (N Training=35).

$$pEC50 = 2.43 + 0.07 \text{ SwHBa} + 0.04 \text{ WPOL} + 0.45 \text{ MDEN-22}$$

N Training set= 35; $R^2 = 0.82$; $Q^2_{LOO} = 0.76$; $Q^2_{LMO30\%} = 0.76$; CCC = 0.90; CCCcv = 0.87; RMSE = 0.42; RMSEcv = 0.49.

9.2. Bibliography:

- [1] Gramatica, P., Cassani, S., Roy, P.P., Kovarich, S. Yap, C.W., Papa, E. QSAR Modeling is not "Push a Button and Find a Correlation": A Case Study of Toxicity of (Benzo-)triazoles on Algae, Mol. Inf. 2012, 31, 817-835 DOI: 10.1002/minf.201200075
- [2] Yap, C.W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 2011, 32, 1466-1474. DOI: 10.1002/jcc.21707
- [3] 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. DOI: 10.1002/jcc.23361
- [4] Gramatica P., et al. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS, J. Comput. Chem. (Software News and Updates), 2014, 35 (13), 1036-1044. DOI: 10.1002/jcc.23576
- [5] FOOTPRINT PPDB (Pesticide Properties DataBase), 2009 <http://sitem.herts.ac.uk/aeru/ppdb/en/> (accessed 30, Jan 2015)
- [6] Chirico N. and Gramatica P., Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, J. Chem. Inf. Model. 2011, 51, 2320-2335. DOI: 10.1021/ci200211n
- [7] 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, 2044-2058 DOI: 10.1021/ci300084j
- [8] Shi L.M. et al. QSAR Models Using a Large Diverse Set of Estrogens, J. Chem. Inf. Comput. Sci. 41 (2001) 186-195. DOI: 10.1021/ci000066d
- [9] 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. doi: 10.1021/ci800253u
- [10] Consonni V. et al. Comments on the Definition of the Q2 Parameter for QSAR Validation, J. Chem. Inf. Model. 49 (2009) 1669-1678 DOI: 10.1021/ci900115y
- [11] KOALA Rel. 1.0 for Windows, 2001. R. Todeschini, V. Consonni, A. Mauri, Milan, Italy no url available

9.3. Supporting information:

Training set(s)

BTaz P. subcapitata tox_ord resp_train.sdf	file:///C:/Documents and Settings/lab-qsar/Desktop/PaDEL QMRF da mandare 2015/BTaz tox\Algae\BTaz P. subcapitata tox_ord resp_train.sdf
BTaz P. subcapitata tox_SOM_train.sdf	file:///C:/Documents and Settings/lab-qsar/Desktop/PaDEL QMRF da mandare 2015/BTaz tox\Algae\BTaz P. subcapitata tox_SOM_train.sdf

Test set(s)

BTAZ P. subcapitata tox_ord resp_pred.sdf	file:///C:/Documents and Settings/lab-qsar/Desktop/PaDEL QMRF da mandare 2015/BTAZ tox\Algae/BTAZ P. subcapitata tox_ord resp_pred.sdf
BTAZ P. subcapitata tox_SOM_pred.sdf	file:///C:/Documents and Settings/lab-qsar/Desktop/PaDEL QMRF da mandare 2015/BTAZ tox\Algae/BTAZ P. subcapitata tox_SOM_pred.sdf

Supporting information

BTAZ P. subcapitata tox_FULL.sdf	file:///C:/Documents and Settings/lab-qsar/Desktop/PaDEL QMRF da mandare 2015/BTAZ tox\Algae/BTAZ P. subcapitata tox_FULL.sdf
----------------------------------	---

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