	QMRF identifier (JRC Inventory):**Q15-31-0011**
	QMRF Title:**QSARINS model for (benzo-)triazole toxicity in** *Daphnia magna*
	Printing Date:**Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

QSARINS model for (benzo-)triazole toxicity in *Daphnia magna*

1.2.Other related models:

1.3.Software coding the model:

PaDEL-Descriptor

A software to calculate molecular descriptors and fingerprints, version 2.18 [ref 2; sect 9.2]

Yap Chun Wei, 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, version 1.2 (also verified with version 2.2, 2015)

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

http://www.qsar.it/

2.General information

2.1.Date of QMRF:

03/02/2015

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 [ref 4; sect 9.2]

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

[1]Cassani S et al (2013). Daphnia and fish toxicity of (benzo)triazoles: Validated QSAR models,and interspecies quantitative activity–activity modelling. Journal of Hazardous Material 258-259, 50-60.

DOI: 10.1016/j.jhazmat.2013.04.025
[2]Yap CW (2011). PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry 32 (7), 1466-1474. DOI: 10.1002/jcc.21707
[3]Gramatica P et al (2013). QSARINS: A new software for the development, analysis and validation of QSAR MLR models. Journal of Computational Chemistry (Software News and Updates) 34 (24), 2121-2132. DOI: 10.1002/jcc.23361
[4]Gramatica P et al (2014). QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. Journal of Computational Chemistry (Software News and Updates) 35 (13), 1036-1044. DOI: 10.1002/jcc.23576

2.8.Availability of information about the model:

Non-proprietary. Defined algorithm, available in QSARINS [ref 3,4; sect 9.2]. Training and prediction sets are available in the Supporting Information of the related paper [ref 1; sect 9.2], in the attached sdf files of this QMRF (section 9)and in the QSARINS-Chem database [ref 4; sect 9.2].

2.9.Availability of another QMRF for exactly the same model:

None to date.

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Daphnia magna

3.2.Endpoint:

3.Ecotoxic effects 3.1.Short-term toxicity to Daphnia (immobilisation)

3.3.Comment on endpoint:

A selected set of experimental EC 50(48h) data was taken from FOOTPRINT PPDB (Pesticide Properties DataBase) online database [ref 5; sect 9.2].

3.4.Endpoint units:

The median lethal 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 202 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 [ref 5; sect 9.2].

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 D.magna SOM Split model
OLS-MLR method. Model developed on a training set of 65 compounds

Log 1/EC50 D.magna Ordered response split model
OLS-MLR method. Model developed on a training set of 65 compounds

Log 1/EC50 D.magna FULL model
OLS-MLR method. Model developed on a training set of 97 compounds
SOM Split model equation: pEC50= 4.44 - 0.02 TopoPSA + 0.0007 WPATH +0.09 C2SP2 - 0.67 nT9Ring - 0.59 maxHaaCH
Ordered Response Split model equation: pEC50: 4.55 - 0.02 TopoPSA + 0.0008 WPATH +0.08 C2SP2 - 0.69 maxHaaCH - 0.46 nT9Ring
Full model equation: pEC50= 4.58 ? 0.02 TopoPSA + 0.0008 WPATH + 0.07 C2SP2 ?0.60 maxHaaCH ?0.47 nT9Ring

The modeling descriptors, calculated in PaDEL-Descriptor 2.18, are:
TopoPSA, WPATH, C2SP2, maxHaaCH, nT9Ring. See section 4.3 for a more detailed description of the five descriptors.

4.3.Descriptors in the model:

[1]TopoPSA dimensionless Topological polar surface area, related to the electronic distribution and polarizability of the molecule and negatively influence toxicity in the multiple linear regression equation
[2]WPATH dimensionless Weiner path number, related to molecular dimension and complexity and in particular, for a constant number of atoms, it reach a maximum value for the most linear structures and a minimum for the most branched. For this reason, it could be suggested as a measure of molecular branching
[3]C2SP2 dimensionless Doubly bound carbon bound to two other carbons, detects the presence of aromaticity in the molecule
[4]maxHaaCH dimensionless Maximum atom-type H E-State: :CH:, encodes for the CH aromatic fragment
[5]nT9Ring dimensionless Number of 9-membered rings (includes counts from fused rings)

4.4.Descriptor selection:

A total of 729 molecular descriptors of differing types (0D, 1D, 2D)
were calculated in PaDEL-Descriptor 2.18 [ref 2; sect 9.2]. 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 230 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 (five variables). The optimized parameter used was
Q 2LOO (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: 5

modeling descriptors
selected from 230.

4.5.Algorithm and descriptor generation:

Multiple linear regression (Ordinary Least Square method) was applied to generate the model.

Molecular descriptors were generated with the PaDEL-Descriptor software 2.18 [ref 2; sect 9.2]..

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 the PaDEL-Descriptor software. Any user can re-derive the model calculating the molecular descriptors with the PaDEL-Descriptor software 2.18 (recently 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, ver. 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, 2012.
Contact not available
http://openbabel.org/wiki/Main_Page

4.7.Chemicals/Descriptors ratio:

Split by SOM model: 65 chemicals / 5 descriptors = 13
Split by Ordered response model: 65 chemicals / 5 descriptors= 13
Full model: 97 chemicals / 5 descriptors = 19.4

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 were 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 *D.magna*values: 2.84 / 6.22

Range of descriptor values: TopoPSA: 29.1 / 208.74 ; WPATH: 15 / 3386;C2SP2: 0 / 15; maxHaaCH: 0 / 0.88 ; nT9Ring: 0 / 1.

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*=0.186). HAT values are calculated as the diagonal elements of the HAT matrix:

H = X(X TX) -1X 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>0.277 (h*): no. Outliers for response, standardised residuals > 2.5 standard deviation units: amitrole (61-82-5), prothioconazole (178928-70-6). Ordered Response Split model domain: outliers for structure, hat>0.277 (h*): tricyclazole (41814-78-2), Flumetsulam (98967-40-9), N-(2,6-difluorophenyl)-8-fluoro-5-hydroxy[1,2,4]triazolo[1,5-c]pyrimidine-2-sulfonamide (NO CAS). Outliers for response, standardised residuals > 2.5 standard deviation units: prothioconazole (178928-70-6), 4(1H)-Pyrimidinone (2814-20-2). FULL model domain:outliers for structure, hat>0.186 (h*): no. Outliers for response, standardised residuals > 2.5 standard deviation units: amitrole (61-82-5), prothioconazole (178928-70-6), 4(1H)-Pyrimidinone (2814-20-2).

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=97) 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= 65) and by ordered response (n training=65). In the SOM splitting, the training and prediction sets are structurally balanced, since the splitting was based on the structural similarity analysis (performed with Kohonen artificial neural network, K-ANN or SOM method included in KOALA software [ref 11; sect 9.2]). In the Ordered response splitting chemicals were ordered according to their increasing toxicity and one out of every three chemicals was 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= 0.72; CCCtr [ref 6,7; sect 9.2]=0.84; RMSE= 0.43

Ordered response split model:

R^2= 0.69; CCCtr=0.82; RMSE= 0.44

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

SOM Split model:

Q^2_{LOO}= 0.67; CCCcv=0.81; RMSEcv= 0.47

Ordered response Split model:

Q ^{2}LOO= 0.63; CCCcv=0.78; RMSEcv= 0.48

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

SOM Split model:

Q ^{2}LMO30%= 0.68

Ordered response split model:

Q ^{2}LMO30%= 0.64

Values of Q 2LMO (average value for 2000 iterations, with 30% of chemicals put out at every iteration) that are close to the original Q ^{2}LOO, mean that the model is robust and stable, even if in this case the parameters are barely acceptable.

6.10.Robustness - Statistics obtained by Y-scrambling:

SOM Split model:

R ^{2}y-sc= 0.08

Ordered response split model:

R ^{2}y-sc= 0.08

Low value of scrambled R 2(average value for 2000iterations, 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:

None to date. (since we have calculated Q ^{2}LMO)

6.12.Robustness - Statistics obtained by other methods:

None to date.

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 models, the dataset (n=97) 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 =32 in both cases); the range of pEC50 are: 3.40 / 5.92 for SOM prediction set, 2.92 / 5.92 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 advantage of the clustering capabilities of Kohonen Artifical 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:

Q^2extF1 [ref 8; sect 9.2]= 0.71; Q^2extF2 [ref 9; sect 9.2]= 0.68; Q^2extF3 [ref 10; sect 9.2]= 0.75; CCCex=0.85; RMSE= 0.42

Ordered response split model:

Q^2extF1= 0.78; Q^2extF2= 0.78; Q^2extF3=0.77; CCCex=0.89; RMSE= 0.38

The high values of external $Q^2$and concordance correlation coefficient-CCC (threshold for accepting the external Q^2F1-F2-F3 is 0.70, threshold for CCC is 0.85, [ref 7; sect 9.2]), show that the 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 [2.84 / 6.22][3.40 / 5.92] and [2.84 / 6.22][2.92 / 5.92] respectively for SOM and ordered response training and prediction sets.

As concerns structural representativity, the range of descriptors values is:

TopoPSA: SOM Split training set (29.1 / 189.31), prediction set (30.71 / 208.74); Ordered response split training set (30.71 / 189.31),
prediction set 29.1 / 208.74)
WPATH: SOM Split training set (15 / 2820), prediction set (213 / 3386); Ordered response split training set (15 / 2820), prediction set (26 / 3386)
C2SP2: SOM Split training set (0/ 13), prediction set (0/ 15); Ordered
response split training set (0 / 13), prediction set (0 / 15)
maxHaaCH: SOM Split training set (0/ 0.88), prediction set (0 / 0.79); Ordered response split training set (0 / 0.83), prediction set (0 / 0.88) nT9Ring: SOM Split training set (0 / 1), prediction set (0 / 1); Ordered response split training set (0 / 1), prediction set (0 / 1)
The applicability domain of the model on the prediction set was verified by the Williams plot: in SOM splitting, there are no compounds of the prediction set that are recognised as outliers, both for response and for structure; in the ordered by response splitting, only 2 chemicals of the prediction set are outliers for structure (no outliers for response). These results demonstrate the broad applicability domain of the 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 a 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 PaDEL-Descriptor model equation published in Cassani et al. [ref 7; sect 9.2] and included in QSARINS is:

pEC50= 4.58 - 0.02 TopoPSA + 0.0008 WPATH +0.07 C2SP2 - 0.60
maxHaaCH - 0.47 nT9Ring

where
TopoPSA: Topological polar surface area WPATH:Weiner path number
C2SP2:Doubly bound carbon bound to two other carbons maxHaaCH: Maximum atom-type H E-State: :CH:
nT9Ring: Number of 9-membered rings (includes counts from fused rings)
TopoPSA is related to the electronic distribution and polarizability of the molecule and negatively influence toxicity in the multiple linear
regression equation; WPATH, with a positive influence towards Daphnia toxicity, is related to molecular dimension and complexity and in particular, for a constant number of atoms, it reach a maximum value for the most linear structures and a minimum for the most branched. For this

reason, it could be suggested as a measure of molecular branching. C2SP2 positively contributes to the studied toxicity and detects the presence
of aromaticity in the molecules of the studied dataset, which is mainly composed of aromatic compounds. maxHaaCH encodes for the CH aromatic
fragment, and nT9Ring for the number of 9-membered rings (includes counts from fused rings).

8.3.Other information about the mechanistic interpretation:

No other information available

9.Miscellaneous information

9.1.Comments:

Given the results of the external validation, this model has a broad 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=97).

pEC50= 4.58 - 0.02 TopoPSA + 0.0008 WPATH +0.07 C2SP2 - 0.60 maxHaaCH -
0.47 nT9Ring

N Training set= 97; R^2= 0.73; Q^2LOO = 0.70; Q^2LMO 30%= 0.69; CCC = 0.84; CCCcv = 0.83;RMSE= 0.42; RMSEcv = 0.44.

9.2.Bibliography:

[1]Cassani S et al (2013). Daphnia and fish toxicity of (benzo)triazoles: Validated QSAR models,and interspecies quantitative activity–activity modelling. Journal of Hazardous Material 258-259, 50-60. DOI: 10.1016/j.jhazmat.2013.04.025

[2]Yap CW (2011). PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry, 32 (7), 1466-1474. DOI: 10.1002/jcc.21707

[3]Gramatica P et al (2013). QSARINS: A new software for the development, analysis and validation of QSAR MLR models. Journal of Computational Chemistry (Software News and Updates). 34 (24), 2121-2132. DOI: 10.1002/jcc.23361

[4]Gramatica P et al (2014). QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS, Journal of Computational Chemistry (Software News and Updates) 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 & Gramatica P (2011). Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, Journal of Chemical Information &.Modeling., 51, 2320-2335. DOI: 10.1021/ci200211n

[7]Chirico N & Gramatica P (2012). Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection, Journal of Chemical Information &.Modeling 52 (8), 2044–2058. DOI: 10.1021/ci300084j

[8]Shi LM et al (2001). QSAR Models Using a Large Diverse Set of Estrogens. Journal of Chemical

Information & Computer Science. 41 186–195. DOI: 10.1021/ci000066d

[9]Schuurman G et al (2008). External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean vs Training Set Activity Mean. Journal of Chemical Information &.Modeling..48 (11), 2140-2145. DOI: 10.1021/ci800253u

9.3.Supporting information:

Training set(s)Test set(s)Supporting information

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q15-31-0011

10.2.Publication date:

2015-06-11

10.3.Keywords:

PaDEL-Descriptor;Daphnia magna;EC50;benzotriazole;triazole;QSARINS;

10.4.Comments: