	QMRF identifier (JRC Inventory): *To be entered by JRC*
	QMRF Title: *Insubria QSAR PaDEL-Descriptor model for prediction of Pharmaceuticals toxicity in Pimephales promelas.* *Keywords: QSARINS; PaDEL-Descriptor; GA-OLS; Pimephales promelas toxicity EC50; Pharmaceuticals; INSUBRIA*
	Printing Date: *Jan 25, 2017*

1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSAR PaDEL-Descriptor model for prediction of Pharmaceuticals toxicity in Pimephales promelas.

Keywords: QSARINS; PaDEL-Descriptor; GA-OLS; *Pimephales promelas* toxicity EC50; Pharmaceuticals; INSUBRIA

1.2.Other related models:

A. Sangion, P. Gramatica, Hazard of pharmaceuticals for aquatic environment: Prioritization by structural approaches and prediction of ecotoxicity, Environ. Int. 95 (2016) 131–143. doi:10.1016/j.envint.2016.08.008 [1]

1.3.Software coding the model:

[1]PaDEL-Descriptor A software to calculate molecular descriptors and fingerprints, version 2.21 [2] Yap Chun Wei, phayapc@nus.edu.sg http://padel.nus.edu.sg/software/padeldescriptor/index.html

[2]QSARINS Software for the development, analysis and validation of QSAR MLR models, version 2.2.1 [3,4] Prof. Paola Gramatica, paola.gramatica@uninsubria.it http://www.qsar.it/

2.General information

2.1.Date of QMRF:

17/01/2017

2.2.QMRF author(s) and contact details:

[1]Alessandro Sangion Insubria University, Department of Theoretical and Applied Sciences (DiSTA), via J.H. Dunant 3, 21100 Varese (Italy) +390332421439 alessandro.sangion@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]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]Alessandro Sangion Insubria University, Department of Theoretical and Applied Sciences (DiSTA), via J.H. Dunant 3, 21100 Varese (Italy)

+390332421439 alessandro.sangion@uninsubria.it http://www.qsar.it/

2.6.Date of model development and/or publication:

Developed and Published in 2016.

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

[1]A. Sangion, P. Gramatica, Hazard of pharmaceuticals for aquatic environment: Prioritization by structural approaches and prediction of ecotoxicity, Environ. Int. 95 (2016) 131–143 doi:10.1016/j.envint.2016.08.008

[2]C.W. Yap, PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints., JComput Chem. 32 (2011) 1466–1474. doi:10.1002/jcc.21707

[3]P. Gramatica, N. Chirico, E. Papa, S. Cassani, S. Kovarich, QSARINS: A new software for the development, analysis and validation of QSAR MLR models, J Comput Chem. 34 (2013) 2121–2132 doi:10.1002/jcc.23361.

[4]P. Gramatica, S. Cassani, N. Chirico, QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS, J.Comput.Chem. 35 (2014) 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 Supporting Information of the related paper [1], 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:

Pimephales promelas

3.2.Endpoint:

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

3.3.Comment on endpoint:

A selected set of experimental EC50 (96h) data was taken from ECOTOX online database [5].

3.4.Endpoint units:

The median lethal concentrations are reported as the minus logarithm of the millimolar concentration: -log(mmol/l)

3.5.Dependent variable:

pEC50

3.6.Experimental protocol:

OECD test 203

3.7.Endpoint data quality and variability:

The experimental data were specifically filtered for the species, defined time of exposure, endpoints and measured effects, trying to ensure the highest degree of homogeneity in experimental measures. If different and multiple values were found for a specific chemical, the minimum LC/EC50 value was taken and modelled, considering the "worst

case scenario"(i.e.the most toxic value). Once these experimental values were selected and filtered, the data were additionally carefully checked, removing the duplicates and measures reported as "nominal concentration".

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:

pEC50 P.promelas Random split
MLR-OLS method. Model developed on a training set of 44 compounds

pEC50 P.promelas Ordered Response split
MLR-OLS method. Model developed on a training set of 42 compounds

pEC50 P.promelas Structural Similarity split
MLR-OLS method. Model developed on a training set of 42 compounds.

pEC50 P.promelas Full Model
MLR-OLS method. Model developed on 62 compounds.

Random split equation: $pLC50\ (96h)_{P.promelas}$ = -3.48 +0.37Kier2 −0.36nHBAcc +0.02AATS3v +1.22SpMin7_Bhp

Ordered Response split equation: $pLC50\ (96h)_{P.promelas}$ = -3.42 +0.02AATS3v +0.34Kier2 −0.38nHBAcc +1.09SpMin7_Bhp

Structural Similarity split equation: $pLC50\ (96h)_{P.promelas}$ = -3.45 +0.02AATS3v −0.38nHBAcc +0.31Kier2 +1.19SpMin7_Bhp

Full model equation: $pLC50\ (96h)_{P.promelas}$ = −3.5 + 0.35Kier2 + 0.02AATS3v −0.39nHBAcc +1.11SpMin7_Bhp

4.3.Descriptors in the model:

[1]Kier2 Second kappa shape index
[2]nHBAcc Number of hydrogen bond acceptors (using CDK HBondAcceptorCountDescriptor algorithm)
[3]AATS3v Average Broto-Moreau autocorrelation - lag 3 / weighted by van der Waals volumes
[4]SpMin7_Bhp Smallest absolute eigenvalue of Burden modified matrix - n 7 / weighted by relative polarizabilities

4.4.Descriptor selection:

A total of 1444 molecular descriptors of differing types (0D, 1D, 2D) were calculated by PaDEL-Descriptor 2.21 [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 547 molecular descriptors were used as input variables for variable subset selection. All the possible combinations of two descriptors were investigated by the all-subset procedure, then, the Genetic Algorithm (GAVSS) was applied to explore new combinations with additional descriptors using Q^2_{LOO} (leave one out) as fitness function to be optimized during the variable selection procedure.

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 2.21. 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 [6]. Then, these files were converted by

OpenBabel 2.3.2 [7] into MDL-MOL format and used as input for the calculation of descriptors in PaDEL-Descriptor. Any user can re-derives the model calculating the molecular descriptors by PaDEL-Descriptor 2.21 software (included in QSARINS 2.2.1) and applying the given equation (automatically done by QSARINS 2.2.1).

4.6.Software name and version for descriptor generation:

PaDEL-Descriptor
A software to calculate molecular descriptors and fingerprints, version 2.21
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
Hypercube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA, 2002
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.
http://openbabel.org/wiki/Main_Page

4.7.Chemicals/Descriptors ratio:

Random split equation: 44/4=11
Ordered Response split equation: 42/4=10.5
Structural Similarity split equation: 42/4=10.5
Full model equation: 62/4=15.5

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 standardized residuals versus leverages (hat diagonals), 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 new chemicals the predictions are inter- or extrapolated by the model.

Response and descriptor space:

Range of experimental pEC50 *P.promelas* values: -2.64 / 5.11 [-log(mmol/L)]

Range of descriptor values: AATS3v: 50.79 / 314.11; Kier2: 1 / 10.71; SpMin7_Bhp: 0.03 / 2.34; nHBAcc: 0 / 10

5.2.Method used to assess the applicability domain:

As it has been stated in section 5.1, the response applicability domain can be verified by the standardized residuals in cross-validation; chemicals with a standardized residual greater than 2.5 deviation units were considerate outliers. The structural applicability domain of the full model was assessed by the leverage approach, providing a cut-off hat value (h*=0.242 for the full model). HAT values are calculated as the diagonal elements of the HAT matrix: $H = X(X^TX)^{-1}X^T$.

5.3.Software name and version for applicability domain assessment:

QSARINS

Software for the development, analysis and validation of QSAR MLR models, version 2.2.1

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

http://www.qsar.it/

5.4.Limits of applicability:

Random split model domain: Outliers for respons (std residual > 2.5): 58-24-5, 52645-53-1; Outliers for structure (hat>0.341): 60-00-4, 64-19-7, 107-19-7. Ordered response split model domain: Outliers for respons (std residual > 2.5): 52645-53-1; Outliers for structure (hat>0.357): 60-00-4, 64-19-7, 107-19-7. Structural Similarity split model domain: Outliers for respons (std residual > 2.5): 52645-53-1; Outliers for structure (hat>0.357): 64-19-7, 107-19-7. Full model

domain: Outliers for respons (std residual > 2.5): 58-27-5; Outlier for structure (hat>0.242): 60-00-4, 64-19-7, 107-19-7.

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:No
Smiles:Yes
Formula:No
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 whole dataset (n=62) was split, before model development, into training sets used for model development and prediction sets used later for external validation. Three different splitting techniques were applied: Random (n training = 44), by ordered response (n training =42) and by structural similarity (n training= 42).

In the Random splitting chemicals are randomly assigned to the training set.

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.

In the structural similarity splitting, training and prediction set are structurally balanced, being the splitting based on the structural similarity analysis (performed with Principal Component Analysis of molecular descriptors).

6.6.Pre-processing of data before modelling:

Transformation of EC50 (mg/L) into EC50 (mmol/L) and than in the logarithmical form Log(1/EC50)

6.7.Statistics for goodness-of-fit:

Random split:

R^2= 0.80; RMSE= 0.69

Ordered Response split:

R^2= 0.79; RMSE= 0.76

Structural Similarity split:

R^2= 0.79; RMSE= 0.77

Full model:
R^2= 0.80; RMSE= 0.72

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

Random split:
Q^2_{loo}=0.75
Ordered Response split:
Q^2_{loo}=0.73
Structural Similarity split:
Q^2_{loo}=0.73

Full model: Q^2_{loo}=0.76

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

Random split:
$Q^2_{LMO30\%}$=0.73
Ordered Response split:
$Q^2_{LMO30\%}$=0.71
Structural Similarity split:
$Q^2_{LMO30\%}$=0.71

Full model: $Q^2_{LMO30\%}$=0.75

6.10.Robustness - Statistics obtained by Y-scrambling:

Random split:
R^2_{Yscr}=0.09
Ordered Response split:
R^2_{Yscr}=0.10
Structural Similarity split:
R^2_{Yscr}=0.10
Full model:
R^2_{Yscr}=0.07

6.11.Robustness - Statistics obtained by bootstrap:

No information available (since only Q^2_{LMO} was calculated)

6.12.Robustness - Statistics obtained by other methods:

Random split:
$RMSE_{CV}$= 0.78; CCC_{CV}=0.86
Ordered Response split:
$RMSE_{CV}$= 0.87; CCC_{CV}=0.85
Structural Similarity split:
$RMSE_{CV}$= 0.88; CCC_{CV}=0.85
Full model: $RMSE_{CV}$=0.79; CCC_{CV}=0.87

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:No
Smiles:Yes
Formula:No
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:

As said in section 6.5, to verify the predictive capability of the proposed models, the dataset (n=62) was split, before model development, into a training set used for model development and a prediction set used later for external validation.

7.6.Experimental design of test set:

As said in section 6.5, in the Random splitting chemicals are randomly assigned to the training set.

In the case of split by Ordered 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, allowed the selection of a structurally meaningful training set and an equally representative prediction set. The selection is based on Principal Component Analysis that is able to project chemicals in the multivariate descriptors space. This method ensures that both sets are homogeneously distributed within the entire area of the descriptors space.

7.7.Predictivity - Statistics obtained by external validation:

Random split model:

Q^2_{extF1} [8]= 0.78; Q^2_{extF2} [9]= 0.77; Q^2_{extF3}[10]= 0.70; CCC_{ext}[11]=0.87; RMSE= 0.84

Ordered response split model:

Q^2_{extF1}= 0.82; Q^2_{extF2}= 0.82; Q^2_{extF3}=0.85; CCC_{ext}=0.89; RMSE= 0.64

Structural Similarity split model:

Q^2_{extF1}= 0.81; Q^2_{extF2}= 0.81; Q^2_{extF3}=0.86; CCC_{ext}=0.89; RMSE= 0.64

The high values of external Q^2 and concordance correlation coefficient-CCC (threshold for accepting the external $Q^2_{F1-F2-F3}$ is 0.70, threshold for CCC is 0.85,[11,12]), show that the proposed model is predictive, when applied to chemicals never seen during the model development (prediction sets)[13].

7.8.Predictivity - Assessment of the external validation set:

The splitting methodology based on similarity analysis, by ordered response and random allowed for the selection of meaningful training sets and representative prediction sets.

Training and prediction sets were balanced according to both structure and response and the predictivity was assessed even on random training and prediction set. In particular:

Random split model:

-response range of pEC50 values: training [-2.55 / 5.11] prediction [-2.64 / 3.82]

-descriptor range:

AATS3v: training [77.50 / 314.11] prediction [50.79 / 253.01]

Kier2: training [1 / 10.71] prediction [1.33 / 9.83]

SpMin7_Bhp: training [0.04 / 2.07] prediction [0.03 / 2.34]

nHBAcc: training [0 / 10] prediction [0 / 9]

Ordered response split model:

-response range of pEC50 values: training [-2.64 / 5.11] prediction [-2.55 / 3.53]

-descriptor range: AATS3v: training [50.79 / 314.11] prediction [90.81 / 232.06] Kier2: training [1 / 10.71] prediction [1.63 / 8.74]

SpMin7_Bhp: training [0.03 / 2.07] prediction [0.05 / 2.34] nHBAcc: training [0 / 10] prediction [0 / 9]

Structural Similarity split model:

-response range of pEC50 values: training [-2.55 / 5.11] prediction [-2.64 / 3.53] -descriptor range: AATS3v: training [50.79 / 314.11] prediction [94.36 / 232.06] Kier2: training [1 / 10.71] prediction [1.33 / 10.17]

SpMin7_Bhp: training [0.03 / 2.07] prediction [0.04 / 2.34]

nHBAcc: training [0 / 10] prediction [0 / 6]

The applicability domain of the model on the prediction set has been verified by the Williams plot.

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 biological activity 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:

Kier2 (std coefficient 0.56) is a topological shape descriptor related to spatial density of atoms in a molecule; in other words this descriptor gives information about the degree of the linearity of the molecular graph. It has a positive influence in increasing toxicity. AATS3v (std coefficient 0.53) refers to the average spatial autocorrelation of the topological structure and describes how the Van der Waals volumes are distributed along the chemical graph. It is a dimensional descriptor with an increasing influence in the toxicity. nHBAcc (std coefficient− 0.52), inversely related to the toxicity, encodes for the number of hydrogen bond acceptor atoms and thus it is related to the intermolecular contacts and interaction of the molecule. Increasing intermolecular interactions and particularly increasing hydrophilicity have a negative influence on the toxicity of the studied compounds. [1]

8.3.Other information about the mechanistic interpretation:

no other information available

9.Miscellaneous information

9.1.Comments:

To predict pEC50 for new Pharmaceuticals without experimental data, it is suggested to apply the equation of the Full Model, developed on all the available chemicals (N Training=62).

Full model equation: pEC50 (96h)*P.promelas* = −3.5 + 0.35Kier2 + 0.02AATS3v −0.39nHBAcc +1.11SpMin7_Bhp

N Training set= 62; R^2= 0.80; Q^2_{LOO} = 0.76; $Q^2_{LMO30\%}$= 0.75; CCC_{CV} = 0.87; RMSE= 0.72; $RMSE_{CV}$ = 0.79.

End-point, descriptors and splitting status for each chemical are reported in the supporting information.

9.2.Bibliography:

[1]A. Sangion, P. Gramatica, Hazard of pharmaceuticals for aquatic environment: Prioritization by structural approaches and prediction of ecotoxicity, Environ. Int. 95 (2016) 131–143 doi:10.1016/j.envint.2016.08.008

[2]C.W. Yap, PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints., JComput Chem. 32 (2011) 1466–1474 doi:10.1002/jcc.21707.

[3]P. Gramatica, N. Chirico, E. Papa, S. Cassani, S. Kovarich, QSARINS: A new software for the development, analysis and validation of QSAR MLR models, J Comput Chem. 34 (2013) 2121–2132. doi:10.1002/jcc.23361.

[4]P. Gramatica, S. Cassani, N. Chirico, QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS, J.Comput.Chem. 35 (2014) 1036–1044. doi:10.1002/jcc.23576.

[5]US EPA, ECOTOX User Guide: ECOTOXicology Database System. Version 4.0. http:/www.epa.gov/ecotox/,

[6]Hypercube, inc, HyperChem(TM), Hypercube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA, 2002 http://www.hyper.com/
[7]N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: An open chemical toolbox, J. Cheminformatics. 3 (2011) 33 doi:10.1186/1758-2946-3-33
[8]L.M. Shi, H. Fang, W.D. Tong, J. Wu, R. Perkins, R.M. Blair, W.S. Branham, S.L. Dial, C.I. Moland, D.M. Sheehan, QSAR models using a large diverse set of estrogens, J. Chem. Inf. Comput. Sci. 41 (2001) 186–195. doi:10.1021/ci000066d.
[9]G. Schüurmann, R.-U. Ebert, J. Chen, B. Wang, R. Kuehne, 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]V. Consonni, D. Ballabio, R. Todeschini, Comments on the Definition of the Q(2) Parameter for QSAR Validation, J. Chem. Inf. Model. 49 (2009) 1669–1678. doi:10.1021/ci900115y.
[11]Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient, 51 (2011) 2320-2335 10.1021/ci200211n
[12]N. Chirico, P. Gramatica, 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. 52 (2012) 2044–2058 doi:10.1021/ci300084j.
[13]P. Gramatica, A. Sangion, A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology, J. Chem. Inf. Model. 56 (2016) 1127–1131. doi:10.1021/acs.jcim.6b00088.

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