	QMRF identifier (JRC Inventory):**To be entered by JRC**
	QMRF Title:**Insubria QSAR PaDEL-Descriptor model for Modeling Aquatic Toxicity of Organic Chemicals in*****Pimephales promelas*****(Fathead minnow)** **Keywords: PaDEL-Descriptor; GA-OLS;*****Pimephales promelas*** **LC50; QSARINS; INSUBRIA**
	Printing Date:**9-mar-2015**

1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSAR PaDEL-Descriptor model for Modeling Aquatic Toxicity of Organic Chemicals in*Pimephales promelas*(Fathead minnow)

Keywords: PaDEL-Descriptor; GA-OLS;*Pimephales promelas* LC50; QSARINS; INSUBRIA

1.2.Other related models:

Papa E., Villa F., Gramatica P., Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales promelas (Fathead Minnow), J. Chem. Inf. Model., 2005,. 45, 1256-1266.[1]

1.3.Software coding the model:

PaDEL-Descriptor

A software to calculate molecular descriptors and fingerprints, version 2.18 [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 [3,4], version 1.2 (also verified with 2.2, 2015)

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

http://www.qsar.it/

2.General information

2.1.Date of QMRF:

06/02/2015

2.2.QMRF author(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]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/

2.3.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]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/

2.6.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]Papa E., Villa F., Gramatica P., Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales promelas (Fathead Minnow), J. Chem. Inf. Model., 2005,. 45, 1256-1266 DOI: 10.1021/ci050212l

[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 information available

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Pimephales promelas(Fathead minnow)

3.2.Endpoint:

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

3.3.Comment on endpoint:

A selected set of experimental LC50 (96h) data (from the original U.S.-E.P.A. Duluth Fathead minnow Database) was taken from Russom et al. (1997) [5]. It consists of flow-through bioassays, conducted with juvenile fathead minnows, on chemicals selected from a cross section of the Toxic Substances Control Act Inventory of industrial organic chemicals.

3.4.Endpoint units:

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

3.5.Dependent variable:

log(1/LC50)

3.6.Experimental protocol:

Experimentally determined LC50 values for 468 industrial organic chemicals were collected from Russom et al. (1997) [5] (original source: U.S.-E.P.A. Duluth Fathead minnow Database)

3.7.Endpoint data quality and variability:

A detailed analysis of the quality of the data reported in Duluth Fathead minnow database was made by Russom et al. (1997) [5]

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 split model

MLR-OLS method. Model developed on a training set of 249 compounds.

Log 1/LC50 full model

MLR-OLS method. Model developed on a training set of 449 compounds.

Split model equation (N Training set: 249): log(1/LC50) 96h = 2.25 + 0.57 VP-1 - 1.09 MLFER_BH + 0.13 nAtomLAC - 0.88 HybRatio + 0.18 naasC - 0.25 nN

Full model equation (N Training set: 449): log(1/LC50) 96h = 2.31 + 0.56 VP-1 - 1.09 MLFER_BH + 0.12 nAtomLAC - 0.86 HybRatio + 0.17 naasC - 0.23 nN

The modeling descriptors (0-2D), calculated in PaDEL-Descriptor 2.18, are: VP-1, MLFER_BH, nAtomLAC, HybRatio, naasC and nN. See section 4.3 for a more detailed description of the six descriptors.

4.3.Descriptors in the model:

[1]VP-1 dimensionless Valence path, order 1. The information related to dimensional features is condensed in this descriptor, positively correlated with the toxicity in Fathead minnow

[2]MLFER_BH dimensionless Overall or summation solute hydrogen bond basicity, negatively correlated with toxicity.

[3]HybRatio dimensionless Fraction of sp3 carbons to sp2 carbons, gives a negative contribution towards fathead minnow toxicity.

[4]nAtomLAC dimensionless Number of atoms in the longest aliphatic chain. This counter is mainly needed to model some particular chemicals in the data set

[5]naasC dimensionless Count of atom-type E-State: :C:- This counter is mainly needed to model some particular chemicals in the data set

[6]nN dimensionless number of nitrogens. This counter is mainly needed to model some particular chemicals in the data set

4.4.Descriptor selection:

A total of 681 molecular descriptors of differing types (0D, 1D, 2D) were calculated in PaDEL-Descriptor 2.18 [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 166 molecular descriptors were used as input variables for variable subset selection by genetic algorithm (GA-VSS).The models were initially developed by the all-subset-procedure until two variable. Then

the GA was applied in order to explore new combinations of variables, selecting the variables by a mechanism of reproduction/mutation. The optimized parameter used was Q^2LOO (leave-one-out). The GA-VSS, by Ordinary Least Squares regression (OLS), included in QSARINS [3,4], was applied to select only the best combination of descriptors from input pool: 6 modeling descriptors selected from 166.

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-derives the model calculating the molecular descriptors by PaDEL-Descriptor 2.18 software (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, 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 (2012)
Contact not available
http://openbabel.org/wiki/Main_Page

4.7.Chemicals/Descriptors ratio:

Split Model: 249 chemicals / 6 descriptors = 41.5
Full Model: 449 chemicals / 6 descriptors = 74.8

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. 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 log(1/LC50) values: 0.04 / 8.45

Range of descriptor values: nN (0 / 4), VP-1 (0.45 / 11.11), naasC (0 / 10), HybRatio (0 / 1), nAtomLAC (0 / 13), MLFER_BH (-0.41 / 3.38)

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*=0.047). HAT values are calculated as the diagonal elements of the HAT matrix:

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

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:

Split model domain: outliers for structure, hat>0.084 (h*):

Nicotine Sulfate (65-30-5), Hexachlorophene (70-30-4), 1,2,4-Triazin-3-amine (17584-12-2), Rotenone (83-79-4), Caffeine (58-08-2), Fensulfothion (115-90-2), Diazinon (333-41-5), Malathion (121-75-5), carbophenothion (786-19-6), 2,4-dinitroaniline (97-02-9).

Outliers for response, standardised residuals > 2.5 standard deviation units: chloroacetonitrile (107-14-2), hexylene glycol (107-41-5), propylene glycol monoacrylate (999-61-1), 2-hydroxyethyl acrylate (818-61-1), hexachloroethane (67-72-1), 1-Octyne-3-ol (818-72-4), But-2-yn-1-ol (764-01-2), isovalerylaldehyde (590-86-3), 3-butyn-2-ol

(65337-13-5).**FULL model domain**: outliers for structure, hat>0.047 (h*): Nicotine Sulfate (65-30-5), Hexachlorophene (70-30-4), 1,2,4-Triazin-3-amine (17584-12-2), Rotenone (83-79-4), Caffeine (58-08-2), Fensulfothion (115-90-2), Diazinon (333-41-5), Malathion (121-75-5), carbophenothion (786-19-6), Chlorpyrifos (2921-88-2). Outliers for response, standardised residuals > 2.5 standard deviation units: chloroacetonitrile (107-14-2), hexylene glycol (107-41-5), propylene glycol monoacrylate (999-61-1), 2-hydroxyethyl acrylate (818-61-1), hexachloroethane (67-72-1), 1-Octyne-3-ol (818-72-4), 2-butyn-1-ol(764-01-2).

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=449) was split, before model development, into training set used for model development and prediction set used later for external validation. The splitting was made by structural similarity (Self Organizing Maps, SOM, n training= 249). 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]).

The training set of the Split Model consists of 249 heterogeneous organic compounds (including almost all the principal functional groups present mainly in pesticides) with a range of log(1/LC50) values from 0.04 to 8.45.

6.6.Pre-processing of data before modelling:

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

6.7.Statistics for goodness-of-fit:

R^2= 0.77; CCCtr [6,7]=0.87; RMSE= 0.62

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

Q^2LOO= 0.76; CCCcv=0.86; RMSEcv= 0.64

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

$Q^2LMO_{30\%}$= 0.76.

Values of Q^2LMO (average value for 2000 iterations, with 30% of chemicals put out at every iteration) that are high and close to the original Q^2LOO, mean that the model is robust and stable.

6.10.Robustness - Statistics obtained by Y-scrambling:

$R^2y\text{-}sc$= 0.02.

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^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: 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=449) was split, before model development, into a training set used for model development and a prediction set used later for external validation.

The external prediction set consists of 200 heterogeneous organic compounds with a range of log(1/LC50) values from 0.84 to 6.72. Training and prediction set are structurally balanced, being the splitting based on the structural similarity analysis (SOM, see Section 6.5)

7.6.Experimental design of test set:

The splitting of the original data set (449 compounds) into a training set of 249 compounds and a prediction set of 200 compounds was realized by Kohonen artificial neural network (K-ANN or SOM). The splitting based on structural similarity (SOM) takes advantages of the clustering capabilities of K-ANN, 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 [8]= 0.71; Q^2extF2 [9]= 0.71; Q^2extF3 [10]= 0.77; CCCex=0.84; RMSE= 0.63

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, [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 allowed for the selection of a meaningful training set and a representative prediction set.

Training and prediction set are balanced according to both structure and response. In particular, for response the range of log(1/LC50) values are [0.04 / 8.45] and [0.84 - 6.72] respectively for training and prediction set. As much as concern structural representativity, the range of descriptors values are:

nN: training set (0 / 4), prediction set (0 / 4)

VP-1: training set (0.45 / 11.11), prediction set (0.81 / 8.71)

naaasC: training set (0 / 10), prediction set (0 / 6)

HybRatio: training set (0 / 1), prediction set (0 / 1)

nAtomLAC: training set (0 / 13), prediction set (0 / 12)

MLFER_BH: training set (-0.41 / 3.38), prediction set (-0.16 / 2.29)

The applicability domain of the model on the prediction set has been verified by the Williams plot: only two compounds of the prediction set (n=200) is recognised as outlier for structure, and four as ouliers for response. These results are a proof of the large applicability domain of the proposed model.

7.9.Comments on the external validation of the model:

No 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, even if in this case is particularly difficult:

The equation of the PaDEL-Descriptor model included in QSARINS 2.2 is : $\log(1/LC50)_{96h}$ = 2.31 + 0.56 VP-1 - 1.09 MLFER_BH+ 0.12 nAtomLAC - 0.86 HybRatio+ 0.17 naasC - 0.23 nN

Where:

VP-1: Valence path, order 1

MLFER_BH: Overall or summation solute hydrogen bond basicity

HybRatio: Fraction of sp3 carbons to sp2 carbons

nAtomLAC: Number of atoms in the longest aliphatic chain

naasC: Count of atom-type E-State: :C:- nN: number of nitrogensThe six theoretical descriptors selected in this model (see Section 4.3 for a short explanation) are a combination of global structural features, able to represent the high structural heterogeneity of the training and prediction set: VP-1, MLFER_BH, HybRatio, nAtomLAC, naasC, nN. The information related to dimensional features is condensed in the most important descriptor VP-1 (positively correlated with the toxicity in fish), while some counters (nAtomLAC, naaSC, nN) are mainly needed to model some particular chemicals in the data set.

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 pLC50 (Fathead minnow) for new organic chemicals without experimental data, it is suggested to apply the equation of the **Full Model**, developed on all the available chemicals (N Training=75).

log(1/LC50) 96h= 2.31 + 0.56 VP-1 - 1.09 MFLER_BH + 0.12 nAtomLAC - 0.86 HybRatio + 0.17 naasC - 0.23 nN

N Training Set= 449; R^2= 0.75; Q^2LOO = 0.74; $Q^2LMO_{30\%}$= 0.74; CCC = 0.86; CCCcv = 0.85 ;RMSE= 0.626; RMSEcv = 0.637

9.2.Bibliography:

[1]Papa E. et al. Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales promelas (Fathead Minnow), J. Chem. Inf. Model., 2005,. 45, 1256-1266 DOI: 10.1021/ci050212l

[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]Russom, C. L. et al. Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas). EnViron. Toxicol. Chem. 1997, 16, 948-967. DOI: 10.1002/etc.5620160514

[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 url not available

9.3.Supporting information:

Training set(s)

P. promelas_train.sdf	file:///C:\Documents and Settings\lab-qsar\Desktop\PaDEL QMRF da mandare 2015\Pimephales\P. promelas_train.sdf

Test set(s)

P. promelas_pred.sdf	file:///C:\Documents and Settings\lab-qsar\Desktop\PaDEL QMRF da mandare 2015\Pimephales\P. promelas_pred.sdf

Supporting information

P. promelas FULL.sdf	file:///C:\Documents and Settings\lab-qsar\Desktop\PaDEL QMRF da mandare 2015\Pimephales\P. promelas 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