	QMRF identifier (JRC Inventory):*To be entered by JRC*	
QMRF	QMRF Title: Insubria QSAR PaDEL-Descriptor model for Modeling PFC inhalation toxicity in Rat	QMRF
	Printing Date:*Jan 20, 2014*	

1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSAR PaDEL-Descriptor model for Modeling PFC inhalation toxicity in Rat

1.2.Other related models:

Bhhatarai B., Gramatica P., Per- and Polyfluoro Toxicity (LC50 Inhalation) Study in Rat and Mouse Using QSAR Modeling, Chem. Res. Toxicol., 2010, 23, 528–539 [7]

1.3.Software coding the model:

[1]PaDEL-Descriptor 2.18 A software to calculate molecular descriptors and fingerprints http://padel.nus.edu.sg/software/padeldescriptor/index.html

[2]QSARINS 1.2 Software for the development, analysis and validation of QSAR MLR models paola.gramatica@uninsubria.it www.qsar.it

2.General information

2.1.Date of QMRF:

01/12/2013

2.2.QMRF author(s) and contact details:

Alessandro Sangion DiSTA, University of Insubria (Varese - Italy) a.sangion@hotmail.it www.qsar.it

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1]Paola Gramatica DiSTA, University of Insubria (Varese - Italy) paola.gramatica@uninsubria.it www.qsar.it

[2]Stefano Cassani DiSTA, University of Insubria (Varese - Italy) stefano.cassani@uninsubria.it www.qsar.it

2.6.Date of model development and/or publication:

July 2013

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

[1]Gramatica P., et al. QSARINS: A new software for the development, analysis and validation of QSAR MLR models, J. Comput. Chem. (Software News and Updates), 2013, 34 (24), 2121-2132 [1]

[2]QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS, submitted to J. Comput. Chem. (Software News and Updates)

2.8.Availability of information about the model:

The model is non-proprietary and published in a scientific peerreviewed journal. All information in full details are available (e.g.training and prediction set, algorithm, ecc...).

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

No

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Rat (Rattus)

3.2.Endpoint:

4.Human health effects 4.1.Acute inhalation toxicity

3.3.Comment on endpoint:

lethal concentration 50 (LC50)

Standard measure of the toxicity of the surrounding medium that will kill half of the sample population of a specific test-animal in a specified period through exposure via inhalation (respiration). LC50 is measured in micrograms (or milligrams) of the material per liter, or parts per million (ppm), of air or water

3.4.Endpoint units:

The median lethal concentrations are reported as the inverse log of the molar concentration: pLC50 rat (mmol/m^3)

3.5.Dependent variable:

pLC50

3.6.Experimental protocol:

The experimental data on rat LC50 inhalation toxicities were collected from ChemID plus [2]

3.7.Endpoint data quality and variability:

The ChemID plus data was verified as much as possible and filtered by performing principle component analysis (PCA) and by omitting the spurious compounds which could badly influence the regression models.

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:

pLC50 PaDEL-Descriptor full model for PFC rat inhalation Toxicity
OLS - Multiple linear Regression Model developed on a training set of 52 chemicals

pLC50 PaDEL-Descriptor split model (Ordered Response) for PFC rat inhalationToxicity
OLS - Multiple linear Regression Model developed on a training set of 42 chemicals

Full model equation: pLC50= 0.72 + 0.52 C2SP2 + 0.05 TopoPSA + 0.52 nAtomLAC - 0.92 minHCsats

Split by Ordered Response model equation: pLC50= 0.41 + 0.58

C2SP2 + 0.06 TopoPSA + 0.58 nAtomLAC -0.58 minHCsats

4.3.Descriptors in the model:

[1]C2SP2 Doubly bound carbon bound to two other carbons

[2]TopoPSA Topological polar surface area

[3]nAtomLAC Number of atoms in the longest aliphatic chain

[4]minHCsats Minimum atom-type H E-State: H bonded to B, Si, P, Ge, As, Se, Sn or Pb

4.4.Descriptor selection:

A total of 734 molecular descriptors of different kinds (0D, 1D, 2D) were calculated by PaDEL-Descriptor software to describe the chemical diversity of the compounds. Constant and semi-constant (at least 20% compounds must have values different from zero or from the values of other chemicals) values and descriptors found to be pair-wise correlated more than 0.98 were excluded in a prereduction step. The Genetic Algorithm (GA) was applied to a final set of 144 descriptors for variable selection.

4.5.Algorithm and descriptor generation:

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

Molecular descriptors were generated by PaDEL-Descriptor software. The input files for descriptor calculation contain information on atom and bond types, connectivity, partial charges and atomic spatial coordinates, relative to the minimum energy conformation of the molecule, and were firstly obtained by the semi empirical AM1 method using the package HYPERCHEM. Then, these files were converted by OpenBabel into MDL-MOL format and used as input for the calculation of descriptors in PaDEL-Descriptor.

4.6.Software name and version for descriptor generation:

PaDEL-Descriptor

An open source software to calculate molecular descriptors and fingerprints, ver. 2.13, 2012.

Yap C.W, National University of Singapore

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

HYPERCHEM - ver. 7.03

Software for molecular drawing and conformational energy optimization

OpenBabel ver.2.3.0, 2010

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

http://openbabel.org

4.7.Chemicals/Descriptors ratio:

Full model: 52 chemicals / 4 descriptros = 13

Split by Ordered response: 42 chemicals / 4 descriptors = 10.5

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 leverages (hat diagonals) versus standardised residuals, i.e. the Williams plot, verified the presence of response outliers (i.e.compounds with cross-validated standardized residuals greater than 2.5 standard deviation units) and chemicals very structurally influential in determining model parameters parameters (i.e. compounds with a leverage value (h) greater than $3p'/n$ (h*), where p' is the number of model variables plus one, and n is the number of the objects used to calculate the model). For new compounds without experimental data, leverage can be used as a quantitative measure for evaluating the degree of extrapolation: for compounds with a high leverage value (h > h*), that are structural outliers, predictions should be considered less reliable.

Response and descriptor space:

Range of experimental pLC50 values: -0.81 / 6.61.

Range of descriptor values:C2SP2 (0 / 6) minHCsats (0 / 1.544) nAtomLAC (0 / 6) TopoPSA (0 / 52.6)

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*=0.288). 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, calculated as: $r'_i = r_i / s\sqrt{(1-h_{ii})}$, where $r_i = Y_i - \hat{Y}_i$.

5.3.Software name and version for applicability domain assessment:

QSARINS 1.2
Software for the development, analysis and validation of QSAR MLR models
paola.gramatica@uninsubria.it
www.qsar.it

5.4.Limits of applicability:

Full model domain:outliers for structure, hat>0.288 (h*):no; Outliers for response, standardised residuals > 2.5 standard deviation units: cyanuric fluoride (675-14-9)

Split by Ordered Response model domain:outliers for structure, hat>0.357 (h*): no; Outliers for response, standardised residuals > 2.5 standard deviation units: cyanuric fluoride (675-14-9)

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:

Yes

6.2.Available information for the training set:
CAS RN:Yes
Chemical Name:Yes
Smiles:Yes
Formula:Yes
INChI:No
MOL file:Yes
6.3.Data for each descriptor variable for the training set:
All
6.4.Data for the dependent variable for the training set:
All
6.5.Other information about the training set:
   The training set of the **Split by Ordered Response Model** consists of 42 perfluorinated compounds with a range of pLC50 values from -0.808 to 6.609
6.6.Pre-processing of data before modelling:
   The original g/m^3 data were converted into the $mmol/m^3$ and expressed in inverse log unit for modeling which are represented as pLC50
6.7.Statistics for goodness-of-fit:
   Split by Ordered Response Model:
   R^2: 0.79; CCCtr[3]: 0.88; RMSEtr: 0.79
6.8.Robustness - Statistics obtained by leave-one-out cross-validation:
   Split by Ordered Response Model:
   Q^2loo: 0.73; CCCcv: 0.85; RMSEcv: 0.91
6.9.Robustness - Statistics obtained by leave-many-out cross-validation:
   Split by Ordered Response Model: Q^2LMO: 0.72
6.10.Robustness - Statistics obtained by Y-scrambling:
   Split by Ordered Response Model: R^2Yscr: 0.095
6.11.Robustness - Statistics obtained by bootstrap:
   No information available (since we have calculated Q2LMO)
6.12.Robustness - Statistics obtained by other methods:
   No information available

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:
Yes
7.2.Available information for the external validation set:
CAS RN:Yes
Chemical Name:Yes
Smiles:Yes
Formula:Yes
INChI:No
MOL file:Yes
7.3.Data for each descriptor variable for the external validation set:

All

7.4.Data for the dependent variable for the external validation set:
All

7.5.Other information about the external validation set:
To verify the predictive capability of the proposed models, the dataset (n=52) was split, before model development, into a training set used for model development and a prediction set used later for external validation by Ordered Response (n external validation set =10) technique.

7.6.Experimental design of test set:
In the split by **Ordered Response model**, chemicals were ordered according to their increasing activity, and one out of every five chemicals was put in the prediction set (always including the most and the least active compounds in the training set)

7.7.Predictivity - Statistics obtained by external validation:
Split by Oredered Response model: n prediction= 10; R^2ext = 0.75; Q^2ext F1[4]= 0.68; Q^2 ext F2[5] = 0.68; Q^2 ext F3[6] = 0.77; CCCex = 0.86; RMSEex = 0.84; MAEex =0.74

7.8.Predictivity - Assessment of the external validation set:
Range of response for prediction set **(Ordered Response split**, n=10) compounds:

log(1/LC50) mmol/m^3: 0.6793 / 5.3576 (range of corrispondig training set: -0.8079 / 6.6091)

Range of modeling descriptors for prediction set (**Ordered Response split,** n=10) compounds:

C2SP2: 0 / 5 (range of corrispondig training set: 0 / 6)

minHCsats: 0 / 1.54375 (range of corrispondig training set: 0 / 1.45)

nAtomLAC: 0 / 5 (range of corrispondig training set: 0 / 6)

TopoPSA: 0 / 52.6 (range of corrispondig training set:0 / 47.58)

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 was defined a priori.

8.2.A priori or a posteriori mechanistic interpretation:
The DRAGON model published in Bhhatarai B., Gramatica P. [7] is:

pLC50= - 12.76 + 1.87 Jhetv + 11.43 PCR- 0.60 MLOGP - 1.41 B02[Cl-Cl]

where Jhetv: Balaban-type index for Van der Waals weighted distance matrix

PCR: Ratio of multiple path count over path count

MLOGP: Moriguchi octanol-water partition coeff.
B02[Cl-Cl]: presence/absence of Cl-Cl at topological distance 02

The two most influential 2D descriptors Jhetv and PCR were with positive signs indicating that an increase in their value is more favorable for the increase in toxic activity.

Jhetv exhibits the bond multiplicity, the heteroatoms, and the number of atoms present in a compound. When the number of heteroatoms in a molecule increases, the molecule hydrates better in water and becomes more soluble.

PCR represents the ratio of the conventional bond-order ID number (piID) and the total path count TPC. In simple terms, the piID number accounts for multiple bonds in the molecule, and for saturated molecules, each edge weight is equal to one; therefore, piID coincides with the total path count TPC, which gives a ratio (PCR) piID/TPC) equal to one.

The next important descriptor was MlogP, which was with a negative coefficient, similar to the mouse LC50 model. It signifies that more toxic compounds are less hydrophobic or more polar in nature. Perfluorinated chemicals, however, have surfactant-like properties with a long alkyl tail bearing multiple fluorines, giving hydrophobic properties and an end group which is mostly polar in nature. The fluorine in turn is also capable of forming hydrogen bonds. Thus, the increase in hydrophobicity was found to be negative in increasing the toxicity or positive in decreasing it. But increase in the polar character will aid in increasing the toxicity of perfluorinated chemicals.

The least significant descriptor was B02[Cl-Cl], the negative coefficient of which confirms that compounds with a higher value of B02[Cl-Cl] are less toxic for the rat LC50 inhalation end point.

The equation of the new PaDEL-descriptor model included in QSARINS is :   pLC50= 0.72 + 0.52 C2SP2 + 0.05 TopoPSA + 0.52 nAtomLAC - 0.92 minHCsats

where C2SP2= Doubly bound carbon bound to two other carbons

TopoPSA= Topological polar surface area

nAtomLAC= Number of atoms in the longest aliphatic chain

minHCsats= Minimum atom-type H E-State: H bonded to B, Si, P, Ge, As, Se, Sn or Pb

The PaDEL-Descriptor model is logPfree and the performances of the two models are similar and comparable, even if the lipophilicity information carried by MlogP in the DRAGON model is encoded by other dimensional descriptors, such as TopoPSA and nAtomLAC. Out of modeling descriptors a high correlation was found only for the couple PCR – C2SP2 (0.82). It is interesting to note that GA selected the descriptor TopoPSA, which was present in the model for mouse inhalation toxicity of PFCs also for the corresponding rat toxicity.

8.3.Other information about the mechanistic interpretation:

no other information available

9.Miscellaneous information

9.1.Comments:

To predict inhalation toxicity in rat for new PFC chemicals without experimental data, it is suggested to apply the equation of the Full Model, developed on all the available chemicals (N=52), thus ensuring a wider applicability domain.

The equation (reported also in section 4.2) and the statistical parameters of the full model are:

Full model equation: pLC50= 0.72 + 0.52 C2SP2 + 0.05 TopoPSA + 0.52 nAtomLAC - 0.92 minHCsats

N = 52; R^2 = 0.79; Q^2 = 0.73; Q^2LMO =   0.74; CCC = 0.88; CCCcv = 0.85; RMSE= 0.78; RMSEcv = 0.86

9.2.Bibliography:

[1]Gramatica P., et al. QSARINS: A new software for the development, analysis and validation of QSAR MLR models, J. Comput. Chem. (Software News and Updates), 2013, 34 (24), 2121-2132

[2]ChemID Plus http://chem.sis.nlm.nih.gov/chemidplus/

[3]Chirico N. and Gramatica P., Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection, J. Chem. Inf. Model. 2012, 52, pp 2044– 2058

[4]Shi L.M. et al. QSAR Models Using a Large Diverse Set of Estrogens, J. Chem. Inf. Comput. Sci. 41 (2001) 186–195.

[5]Schuurman G. et al. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean vs Training Set Activity Mean, J. Chem. Inf. Model. 48 (2008) 2140-2145.

[6]Consonni V. et al. Comments on the Definition of the Q2 Parameter for QSAR Validation, J. Chem. Inf. Model. 49 (2009) 1669-1678

[7]Bhhatarai B., Gramatica P., Per- and Polyfluoro Toxicity (LC50 Inhalation) Study in Rat and Mouse Using QSAR Modeling, Chem. Res. Toxicol., 2010, 23, 528–539

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