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

1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSAR PaDEL-Descriptor model for PFC Oral toxicity in Rat

1.2.Other related models:

Bhhatarai B., Gramatica P., Oral LD50 toxicity modeling and prediction of per- and polyfluorinated chemicals on rat and mouse, Mol. Divers., 2011, 15, 467-476 [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:

14/11/2013

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

3.2.Endpoint:

4.Human health effects 4.2.Acute oral toxicity

3.3.Comment on endpoint:

lethal dose 50 (LD50)

Standard measure of the toxicity of a material that will kill half of the sample population of a specific test animal in a specified period through exposure via ingestion, skin contact, or injection. LD50 is measured in micrograms (or milligrams) of the material per kilogram of the test-animal's body weight.

3.4.Endpoint units:

The median lethal doses are reported as the inverse log of the molar dose: pLD50 rat (mmol/Kg)

3.5.Dependent variable:

pLD50

3.6.Experimental protocol:

The experimental data on rat LD50 oral toxicities were collected from ChemID plus[2]

3.7.Endpoint data quality and variability:

No information available

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:

pLD50 PaDEL-Descriptor full model for PFC Rat oral Toxicity
OLS - Multiple linear Regression Model developed on a training set of 50 chemicals

pLD50 PaDEL-Descriptor split model (SOM) for PFC Rat oral Toxicity
OLS - Multiple linear Regression Model developed on a training set of 36 chemicals

pLD50 PaDEL-Descriptor split model (Ordered Response) for PFC Rat oral Toxicity

OLS - Multiple linear Regression Model developed on a training set of 37 chemicals

Full model equation: pLD50= 1.93 + 22.71 SCH-5 + 0.03 SHBint3 + 0.07 maxdO -0.25 SHCsats

Split by SOM model equation: pLD50= 2.07 + 19.36 SCH-5 + 0.03 SHBint3 -0.31 SHCsats + 0.06 maxdO

Split by Ordered Response model equation: pLD50= 1.97 + 21.64 SCH-5 + 0.03 SHBint3 + 0.07 maxdO -0.28 SHCsats

4.3.Descriptors in the model:

[1]SCH-5 Simple chain, order 5

[2]SHBint3 Sum of E-State descriptors of strength for potential Hydrogen Bonds of path length 3

[3]maxdO Maximum atom-type E-State: =O

[4]SHCsats Sum of atom-type H E-State: H on C sp3 bonded to saturated C

4.4.Descriptor selection:

A total of 1565 molecular descriptors of different kinds (0D, 1D, 2D, fingerprints) 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 220 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: 50 chemicals / 4 descriptros = 12.5
Split by SOM: 36 chemicals / 4 descriptors = 9
Split by Ordered response: 37 chemicals / 4 descriptors = 9.25

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 pLD50 values: 0.984 / 5.24.

Range of descriptor values: SCH-5 (0 / 0.096), SHBint3 (0 / 99.94), maxdO (0 / 11.03), SHCsats (0 / 3.25)

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*=0.300). HAT values are calculated as the diagonal elements of the HAT matrix:

H = X(XTX)-1XT

The response applicability domain can be verified by the standardized residuals, calculated as: r'i = ri / s√(1-hii), where ri = Yi-Ŷi.

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>0.300 (h*): 3-Penten-1,5-diol, 3-methyl-1,1,5,5-tetrakis(trifluoromethyl) (756-91-2).Outliers for response, standardised residuals > 2.5 standard

deviation units: no

Split by SOM model domain: outliers for structure, hat>0.417 (h*): 3-Penten-1,5-diol, 3-methyl-1,1,5,5-tetrakis(trifluoromethyl) (756-91-2); Outliers for response, standardised residuals > 2.5 standard deviation units: 1,3-dichlorotetrafluoroacetone (127-21-9), 1,2,2-Trichloropentafluoropropane (1599-41-3).

Split by Ordered Response model domain: outliers for structure, hat>0.405 (h*): 3-Penten-1,5-diol, 3-methyl-1,1,5,5-tetrakis(trifluoromethyl) (756-91-2); Outliers for response, standardised residuals > 2.5 standard deviation units:3-Penten-1,5-diol, 3-methyl-1,1,5,5-tetrakis(trifluoromethyl) (756-91-2), 1,2,2-Trichloropentafluoropropane (1599-41-3), perfluoropentane (138495-42-8).

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 SOM Model** consists of 36 perfluorinated compounds with a range of pLD50 values from 1.268 to 5.02.

The training set of the **Split by Ordered Response Model** consists of 37 perfluorinated compounds with a range of pLD50 values from 0.984 to 5.24.

6.6.Pre-processing of data before modelling:

The original mg/kg data were converted into the mmol/kg and expressed in inverse log unit for modeling which are represented as pLD50

6.7.Statistics for goodness-of-fit:

Split by SOM Model:

R^2: 0.87; CCCtr[3]: 0.93; RMSEtr: 0.39

Split by Ordered Response Model:

R^2: 0.89; CCCtr: 0.94; RMSEtr: 0.39

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

Split by SOM Model:

Q^2loo: 0.82; CCCcv: 0.90; RMSEcv: 0.46

Split by Ordered Response Model:

Q^2loo: 0.85; CCCcv: 0.92; RMSEcv: 0.46

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

Split by SOM Model: Q^2LMO: 0.76

Split by Ordered Response Model: Q^2LMO: 0.83

6.10.Robustness - Statistics obtained by Y-scrambling:

Split by SOM Model: R^2Yscr: 0.11

Split by Ordered Response Model: R^2Yscr: 0.11

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=50) 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 **Ordered Response** (n external validation set =13) and by **structural similarity (SOM)** (n external validation set =14).

7.6.Experimental design of test set:

In the case of **split by Ordered Response model**, chemicals were ordered according to their increasing activity, and one out of every four chemicals was put in the prediction set (always including the most and the least active compounds in the training set). The **splitting by SOM model** takes advantages of the clustering capabilities of Kohonen Artifical Neural Network (K-ANN), allowing the selection of a structurally meaningful training set and an equally representative prediction set.

7.7.Predictivity - Statistics obtained by external validation:

Split by SOM model: n prediction= 14; R^2ext = 0.94; Q^2ext F1[4] = 0.90; Q^2 ext F2[5] = 0.89; Q^2 ext F3[6] = 0.80; CCCex = 0.93; RMSEex = 0.49; MAEex = 0.38.

Split by Oredered Response model: n prediction= 13; R^2ext = 0.88; Q^2ext F1= 0.89; Q^2 ext F2 = 0.89; Q^2 ext F3 = 0.86; CCCex = 0.94; RMSEex = 0.44; MAEex = 38 .

7.8.Predictivity - Assessment of the external validation set:

Range of response for prediction set (**SOM split**, n=14) compounds:

log(1/LD50) mmol/Kg: 0.984 / 5.24 (range of corrispondig training set: 1.268 / 5.02)

Range of modeling descriptors for prediction set (**SOM split**, n=14) compounds:

SCH-5: 0 / 0.096 (range of corrispondig training set: 0 / 0.096)
SHBint3: 0 / 39.08 (range of corrispondig training set: 0 / 99.94)
maxdO: 0 / 11.03 (range of corrispondig training set: 0 / 10.82)
SHCsats : 0 / 1.54 (range of corrispondig training set:0 / 3.25)

Range of response for prediction set (**Ordered Response** split, n=13) compounds:

log(1/LD50) mmol/Kg: 1.348 / 5.24 (range of corrispondig training set: 0.984 / 5.24)

Range of modeling descriptors for prediction set (**Ordered Response split**, n=13) compounds:

SCH-5: 0 / 0.096 (range of corrispondig training set: 0 / 0.096)
SHBint3: 0 / 37.99 (range of corrispondig training set: 0 / 99.94)
maxdO: 0 / 10.82 (range of corrispondig training set: 0 / 11.04)
SHCsats : 0 / 3.18 (range of corrispondig training set:0 / 3.25)

The distribution of response values of the chemicals in the two different training sets is comparable to the distribution of the response values of the two prediction set.

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. and Gramatica P.[7] is: pLD50= - 2.277 + 0.041 D/Dr09 + 2.943 MATS1e + 8.838 E1u + 1.166 H8m whereD/Dr09: distance/detour ring index of order 9 MATS1e: Moran autocorrelation - lag 1 / weighted by atomic Sanderson eletronegativities E1u: 1st component accessibility directional WHIM index / unweighted; (3D

representing information regarding the quantity of unfilled space per projected atom) H8m: H autocorrelation of lag 8 / weighted by atomic masses (3D) The increase in molecular mass increases the value of H8m descriptor, and an increase in polycyclic rings increases the value of D/Dr09 The equation of the new PaDEL-descriptor model included in QSARINS is : pLD50= 1.93 + 22.71 SCH-5 + 0.03 SHBint3 + 0.07 maxdO - 0.25 SHCsats where SCH-5= Simple chain, order 5 SHBint3= Sum of E-State descriptors of strength for potential Hydrogen Bonds of path length 3 maxdO= Maximum atom-type E-State: =O SHCsats= Sum of atom-type H E-State: H on C sp3 bonded to saturated C The PaDEL-Descriptor model is based only on 2D-descriptors, while the DRAGON model was based on two 3D-descriptors, and consequently the PaDEL model is simpler and independent on the molecular conformation. Only D/Dr09 and SCH-5 are highly correlated (0.98), bringing very similar structural information in the modeling.

8.3.Other information about the mechanistic interpretation:

no other information available

9.Miscellaneous information

9.1.Comments:

To predict oral 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=50), 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: pLD50= 1.93 + 22.71 SCH-5 + 0.03 SHBint3 + 0.07 maxdO -0.25 SHCsats

N = 50; R^2 = 0.89; Q^2 = 0.86; Q^2LMO = 0.86; CCC = 0.94; CCCcv = 0.93 ;RMSE= 0.41; RMSEcv = 0.45

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., Oral LD50 toxicity modeling and prediction of per- and polyfluorinated chemicals on rat and mouse, Mol. Divers., 2011, 15, 467-476

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