| | |
|---|---|
| | *QMRF identifier (JRC Inventory):* **Q15-41-0014** |
| | *QMRF Title:* **QSARINS model for inhalation toxicity of polyfluorinated compounds in mouse** |
| | *Printing Date:* **Dec 11, 2019** |
| | |

## 1.QSAR identifier

### 1.1.QSAR identifier (title):

QSARINS model for inhalation toxicity of polyfluorinated compounds in mouse

### 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 [ref 3,4; sect 9.2], 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:

05/02/2015

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

http://www.qsar.it/

[3]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]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 [ref 4; sect 9.2]

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

[1]Bhhatarai B & Gramatica P (2010). Per- and Polyfluoro Toxicity (LC50 Inhalation) Study in Rat and Mouse Using QSAR Modeling. Chemical Research in Toxicology 23, 528–539 DOI: 10.1021/tx900252h

[2]Yap CW (2011). PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry. 32, 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 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:**

mouse (*Mus musculus*)

**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 mouse (mmol/$m^3$)

**3.5.Dependent variable:**

pLC50

**3.6.Experimental protocol:**

**3.7.Endpoint data quality and variability:**

The experimental data on mouse LC50 inhalation toxicities were collected from ChemID plus [ref 5; sect 9.2]

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 Mouse inhalation Toxicity
OLS - Multiple linear Regression Model developed on a training set of 56 chemicals

pLC50 PaDEL-Descriptor split model (SOM) for PFC Mouse inhalation Toxicity
OLS - Multiple linear Regression Model developed on a training set of 40 chemicals

pLC50 PaDEL-Descriptor split model (Ordered Response) for PFC Mouse inhalationToxicity
OLS - Multiple linear Regression Model developed on a training set of 44 chemicals
**Full model equation**: pLC50= 2.95 + 1.36 VP-3 + 0.05 TopoPSA -
1.03 nsssCH - 0.42 XlogP
**Split by SOM model equation**: pLC50= 3.28 + 1.27 VP-3 -1.06 nsssCH
-0.44 XLogP + 0.04 TopoPSA
**Split by Ordered Response model equation:**pLC50= 2.98 + 1.35 VP-3
+ 0.04 TopoPSA -1.10 nsssCH -0.42 XLogP
The modeling descriptors, calculated in PaDEL-Descriptor 2.18, are:
VP-3, TopoPSA, nsssCH, XlogP. See section 4.3 for a more detailed
description of the four descriptors.

### 4.3.Descriptors in the model:

[1]VP-3 dimensionless Valence path, order 3. It has a positive influence on mouse toxicity, and accounts for the presence of the heteroatom and double and triple bonds present in the compound
[2]nsssCH dimensionless Count of atom-type E-State: >CH-, with a negative influence on mouse toxicity
[3]XlogP dimensionless A logP calculated in PaDEL-Descriptor, with a negative influence on studied endpoint; for fluorinated chemicals studied here, the contribution of hydrophobicity, within this combination of descriptors, demonstrates a decreasing trend for mouse inhalation toxicity.
[4]TopoPSA dimensionless Topological polar surface area based on fragment contributions, has a slightly positive contribution on mouse inhalation toxicity

### 4.4.Descriptor selection:

A total of 1609 molecular descriptors of different kinds (0D, 1D, 2D, fingerprints) were calculated by the 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. 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: 4 modeling descriptors selected from 144.

### 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 [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-derives the model calculating the molecular descriptors with the PaDEL-Descriptor 2.18 software (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

An open source software to calculate molecular descriptors and fingerprints, ver. 2.18

Yap Chun Wei, email: phayapc@nus.edu.sg

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, version 2.3.2, 2012. Used for conversion between HYPERCHEM files (hin) and MDL-MOL files.

openbabel-discuss@lists.sf.net

http://openbabel.org/wiki/Main_Page

### 4.7.Chemicals/Descriptors ratio:

**Full model**: 56 chemicals / 4 descriptors = 14

**Split by SOM**: 40 chemicals / 4 descriptors = 10

**Split by Ordered response**: 44 chemicals / 4 descriptors = 11

## 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 pLC50 values**: 0.269 / 6.542.
**Range of descriptor values**:VP-3 (0 / 2.88) XLogP (0.619 / 7.81) TopoPSA ( 0 / 47.58) nsssCH ( 0 / 3)

### 5.2.Method used to assess the applicability domain:

As 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*=0.268). 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-i.

### 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 (also verified with 2.2, 2015)
Prof. Paola Gramatica; paola.gramatica@uninsubria.it
http://www.qsar.it/

### 5.4.Limits of applicability:

**Full model domain**:outliers for structure, hat>0.268 (h*): Propane, 2-chloro-1,1,3,3-tetrafluoro- (19041-02-2); Perfluorodibutyl ether (308-48-5); Outliers for response, standardised residuals > 2.5 standard deviation units: no
**Split by SOM model domain:**outliers for structure, hat>0.375 (h*): Propane, 2-chloro-1,1,3,3-tetrafluoro- (19041-02-2). Outliers for response, standardised residuals > 2.5 standard deviation units: no
**Split by Ordered Response model domain:**outliers for structure, hat>0.341 (h*): Propane, 2-chloro-1,1,3,3-tetrafluoro- (19041-02-2); Perfluorodibutyl ether (308-48-5); Pentadecafluorotriethylamine (359-70-6).Outliers for response, standardised residuals > 2.5 standard deviation units: no

## 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=56) 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= 40) and by ordered response (n training=44). In the SOM splitting, 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 [11]). 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.

**The training set of the Split by SOM Model** consists of 40 perfluorinated compounds with a range of pLC50 values from 0.269 to 6.542.

The training set of the Split by Ordered Response Model consists of 44 perfluorinated compounds with a range of pLC50 values from 0.315 to 6.255.

**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 SOM Model**:

$R^2$: 0.79; CCCtr[6,7]: 0.88; RMSEtr: 0.68

**Split by Ordered Response Model**:

$R^2$: 0.72; CCCtr: 0.84; RMSEtr: 0.77

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

**Split by SOM Model**:

$Q^2$loo: 0.74; CCCcv: 0.85; RMSEcv: 0.76

**Split by Ordered Response Model:**

$Q^2$loo: 0.66; CCCcv: 0.81; RMSEcv: 0.85

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

**Split by SOM Model**: $Q^2LMO_{30\%}$: 0.73

**Split by Ordered Response Model**: $Q^2LMO_{30\%}$:

0.67

High and/or acceptable value of $Q^2$LMO (average value for 2000 iterations, with 30% of chemicals put out at every iteration) means that the model is robust and stable.

**6.10.Robustness - Statistics obtained by Y-scrambling:**

**Split by SOM Model**: $R^2$Yscr: 0.10

**Split by Ordered Response Model**: $R^2$Yscr: 0.09

Very 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 $Q^2$LMO)

**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=56) was split, before model development, into training set(s) used for model development and prediction set(s) used later for external validation. Two different splitting techniques were applied: by**Ordered Response**(n external validation set =12) and by**structural similarity (SOM)**(n external validation set =16).

**7.6.Experimental design of test set:**

In the case of split by the**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). The splitting by the**SOM model** takes advantage 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 (see section 6.5)

**7.7.Predictivity - Statistics obtained by external validation:**

**Split by SOM model:**n prediction= 16; R 2ext = 0.78; Q 2ext F1[ref 8; sect 9.2] = 0.77; Q 2ext F2 [ref 9; sect 9.2]= 0.71; Q 2ext F3 [ref 10; sect 9.2]= 0.70; CCCex = 0.81; RMSEex = 0.80; MAEex = 0.60.

**Split by Ordered Response model:**n prediction= 12; R 2ext = 0.95; Q 2ext F1= 0.95; Q 2ext F2 = 0.95; Q 2ext F3 = 0.93; CCCex = 0.97; RMSEex = 0.40; MAEex = 0.35.

The high values of external Q 2and 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]), showthat the model is highly predictive when applied to 550 chemicals not used during the model development.

**7.8.Predictivity - Assessment of the external validation set:**

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 sets. The applicability domain of the model on the prediction set was verified by the Williams plot: 3 compounds on 12 of the prediction set, ordered by response splitting, are outliers for structure (no outliers for response); no compounds of the prediction set, in SOM splitting, are outliers (for response and for structure). 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 statistical approach. No mechanistic basis was defined a priori.

**8.2.A priori or a posteriori mechanistic interpretation:**

A posteriori mechanistic interpretation:

The equation of the PaDEL-descriptor model included in QSARINS 2.2 is the following :

**pLC50**= 2.95 + 1.36 VP-3 + 0.05 TopoPSA - 1.03 nsssCH - 0.42 XlogP

where VP-3= Valence path, order 3
nsssCH= Count of atom-type E-State: >CH-
XlogP= a calculated logP value
TopoPSA= Topological polar surface area

The most influential descriptor is VP-3, with a positive influence on mouse toxicity. The VP-3 (valence path, order 3) accounts for the presence of the heteroatom and double and triple bonds present in the compound. TopoPSA, the topological polar surface area based on fragment contributions, has a slightly positive contribution on mouse inhalation toxicity. The E-State nsssCH has a negative coefficient in the equation, as well as XlogP; therefore for fluorinated chemicals studied here, the contribution of hydrophobicity, within this combination of descriptors, demonstrates a decreasing trend for mouse inhalation toxicity.

### 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 verifying the model applicability.

To predict pLC50 for new PFCs without experimental data, it is suggested to apply the equation of the **Full Model**, developed on all the available chemicals (N Training=56).

**Full model equation**: pLC50= 2.95 + 1.36 VP-3 + 0.05 TopoPSA - 1.03 nsssCH - 0.42 XlogP

N Training set= 56; $R^2$= 0.79; $Q^2$LOO = 0.75; $Q^2LMO_{30\%}$= 0.75; CCC = 0.88; CCCcv = 0.86; RMSE= 0.70; RMSEcv = 0.76

### 9.2.Bibliography:

[1]Bhhatarai B & Gramatica P (2010). Per- and Polyfluoro Toxicity (LC50 Inhalation) Study in Rat and Mouse Using QSAR Modeling. Chemical Research in Toxicology. 23, 528–539. DOI: 10.1021/tx900252h

[2]Yap CW (2011). PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry. 32, 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]ChemID Plus http://chem.sis.nlm.nih.gov/chemidplus/
[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 and 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 and Modeling. 52, 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 and Computer Sciences. 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 and Modeling. 48, 2140-2145. DOI: 10.1021/ci800253u
[10]Consonni V et al (2009). Comments on the Definition of the Q2 Parameter for QSAR Validation, Journal of Chemical Information and Modeling. 49, 1669-1678 DOI: 10.1021/ci900115y 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)Test set(s)Supporting information**

## 10.Summary (JRC QSAR Model Database)

**10.1.QMRF number:**

Q15-41-0014

**10.2.Publication date:**

2015-06-12

**10.3.Keywords:**

PaDEL-Descriptor;polyfluorinated;mouse;inhalation toxicity;QSARINS; ;

**10.4.Comments:**