
	QMRF identifier (JRC Inventory): To be entered by JRC	
	QMRF Title: Insubria QSPR PaDEL-Descriptor model for PFC Water Solubility (Sw)	
	Printing Date: Feb 11, 2014	

1. QSAR identifier

1.1. QSAR identifier (title):

Insubria QSPR PaDEL-Descriptor model for PFC Water Solubility (Sw)

1.2. Other related models:

Bhatarai B., Gramatica P., Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals, Environ. Sci. Technol., 2011, 45, 8120–8128 [8]

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:

16/01/2014

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 peer-reviewed 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:

No information available

3.2. Endpoint:

1. Physicochemical effects 1.3. Water solubility

3.3. Comment on endpoint:

The solubility of a chemical in water (S_w) may be defined as the maximum amount of the chemical that will dissolve in pure water at a specified temperature. Above this concentration, two phases will exist if the organic chemical is a solid or a liquid at the system temperature: a saturated aqueous solution and a solid or liquid organic phase. Aqueous concentrations are usually stated in terms of weight per weight (ppm, ppb, g/kg, etc.) or weight per volume (mg/L, moles/L, etc.).

3.4. Endpoint units:

mg/L

3.5. Dependent variable:

$\log S_w$

3.6. Experimental protocol:

The experimental data were collected from SRC PhysProp database[2] and the data compiled in EU-FP6 PERFORCE report[3]. For S_w (in mg/L), data reported (20 compounds) at temperature range of 293-298K in PERFORCE report were used. Shake flask method was used for most of the data, method incorporated as standard OECD tests on solubility.

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-squares)

4.2. Explicit algorithm:

$\log S_w$ (Full model)

OLS-MLR method. Model developed on a training set of 20 compounds

$\log S_w$ (Split by SOM model)

OLS-MLR method. Model developed on a training set of 15 compounds

$\log S_w$ (Split by Ordered Response model)

OLS-MLR method. Model developed on a training set of 15 compounds

Full model equation: $\log Sw = 4.02 - 0.86 \text{ XLogP} + 1.89$
PubchemFP344

Split by SOM model equation: $\log Sw = 4.22 - 0.92 \text{ XLogP} + 1.98$
PubchemFP344

Split by Ordered Response model equation: $\log Sw = 3.93 - 0.84 \text{ XLogP} + 1.80$
PubchemFP344

4.3.Descriptors in the model:

[1]XlogP XlogP

[2]PubchemFP344 C(~C)(~H). Simple atom nearest neighbors - These bits test for the presence of atom nearest neighbor patterns, regardless of bond order (denoted by "~") or count, but where bond aromaticity (denoted by ":") is significant.

4.4.Descriptor selection:

A total of 1571 molecular descriptors of differing types (0D, 1D, 2D, fingerprints) were calculated in PaDEL-Descriptor 2.18. 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 110 molecular descriptors were used as input variables for variable subset selection. The models were developed by the all-subset-procedure. The optimized parameter used was Q2LOO (leave-one-out).

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 2.18

A software to calculate molecular descriptors and fingerprints

Yap Chun Wei, Department of Pharmacy, 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.2.3.2

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: 20 chemicals / 2 descriptors = 10

Split by SOM: 15 chemicals / 2 descriptors = 7.5

Split by Ordered response: 15 chemicals / 2 descriptors = 7.5

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 (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: logSw values: -2.29 / 3.74

Range of descriptor values: XLogP (1.18 / 9.21), PubchemFP344 (0 / 1)

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 ($h^*=0.450$). HAT values are calculated as the diagonal elements of the HAT matrix:

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

The response applicability domain can be verified by the standardized residuals, calculated as: $r'_i = r_i / \sqrt{s^2(1-h_{ii})}$, 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, $hat > 0.450$ (h^*): no;
Outliers for response, standardised residuals > 2.5 standard deviation units: no

Split by SOM model domain: outliers for structure, $hat > 0.600$ (h^*):

no; Outliers for response, standardised residuals > 2.5 standard deviation units: 2-Decenoic acid, 3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,10-hexadecafluoro (CAS 70887-84-2)

Split by Ordered Response model domain: outliers for structure, $\hat{h} > 0.600$ (h^*): no; 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: 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 15 perfluorinated compounds with a range of logSw values from -2.29 to 3.74.

The training set of the **Split by Ordered Response Model** consists of 15 perfluorinated compounds with a range of logSw values from -2.29 to 3.74.

6.6. Pre-processing of data before modelling:

The original Sw data were expressed in log unit logSw (mg/L)

6.7. Statistics for goodness-of-fit:

Split by SOM Model:

R^2 : 0.88 ; CCCtr[4]: 0.94 ; RMSEtr: 0.61

Split by Ordered Response Model:

R^2 : 0.84 ; CCCtr: 0.91 ; RMSEtr: 0.73

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

Split by SOM Model:

Q^2_{loo} : 0.82 ; CCCcv: 0.90 ; RMSEcv: 0.76

Split by Ordered Response Model:

Q^2_{loo} : 0.77 ; CCCcv: 0.88 ; RMSEcv: 0.88

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

Split by SOM Model: Q^2_{LMO} : 0.70

Split by Ordered Response Model: Q^2_{LMO} : 0.69

6.10. Robustness - Statistics obtained by Y-scrambling:

Split by SOM Model: R^2_{Yscr} : 0.15

Split by Ordered Response Model: R^2_{Yscr} : 0.14

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: 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=20) 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 =5) and by **structural similarity (SOM)** (n external validation set =5).

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 Artificial 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= 5; $R^2_{ext} = 0.79$; $Q^2_{ext} F1[5] = 0.73$; $Q^2_{ext} F2[6] = 0.72$; $Q^2_{ext} F3[7] = 0.78$; $CCC_{ex} = 0.88$; $RMSE_{ex} = 0.84$; $MAE_{ex} = 0.80$.

Split by Ordered Response model: n prediction= 5; $R^2_{ext} = 0.93$; $Q^2_{ext} F1 = 0.91$; $Q^2_{ext} F2 = 0.89$; $Q^2_{ext} F3 = 0.94$; $CCC_{ex} = 0.93$; $RMSE_{ex} = 0.43$; $MAE_{ex} = 0.42$.

7.8. Predictivity - Assessment of the external validation set:

Range of response for prediction set (**SOM split**, n=5) compounds:

$\log Sw$ (mg/L): -1.96 / 2.20 (range of corresponding training set: -2.29 /

3.74)

Range of modeling descriptors for prediction set (**SOM split**, n=5) compounds:

XLogP: 1.18 / 9.21 (range of corresponding training set: 2.68 / 7.45)

PubchemFP344: 0 / 1 (range of corresponding training set: 0 / 1)

Range of response for prediction set (**Ordered Response split**, n=5) compounds:

logSw: -0.83 / 2.99 (range of corresponding training set: -2.29 / 3.74)

Range of modeling descriptors for prediction set (**Ordered Response split**, n=5) compounds: XLogP 3.06 / 7.45 (range of corresponding training set: 1.18 / 9.21)

PubchemFP344: 0 / 1 (range of corresponding training set: 0 / 1)

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 Bhatarai B. and Gramatica P. [8] is:

$$\log Sw = -0.418 - 0.003 T(F..F) + 5.185 SIC1$$

T(F..F): sum of topological distances between pair of fluorine atoms (it increases with the number and the distance between two fluorine atoms in a molecule)

SIC1: structural information content (neighborhood symmetry of 1-order). (is a descriptor, based on neighbor degrees and edge multiplicity, that gives information mainly on the structural symmetry in the molecule.)

the distance of fluorine atoms in the structure is a dominant factor.

PaDEL equation: $\log Sw = 4.02 - 0.86 XLogP + 1.89 PubchemFP344$

XLogP = XLogP

PubchemFP344 = fingerprint that test for the presence of: C(~C)(~H)

High correlation between T(F..F) and XLogP (0.94). The most important

factor in modeling the solubility in water of PFCs is XlogP, with a negative sign in the equation; it is also correlated (0.92) with the number of Fluorine atoms.

8.3. Other information about the mechanistic interpretation:

no other information available

9. Miscellaneous information

9.1. Comments:

To predict Sw 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=20), 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: $\log Sw = 4.02 - 0.86 \text{ XLogP} + 1.89 \text{ PubchemFP344}$

$N = 20$; $R^2 = 0.85$; $Q^2 = 0.80$; $Q^2_{LMO} = 0.73$; $CCC = 0.92$; $CCC_{cv} = 0.89$; $RMSE = 0.67$; $RMSE_{cv} = 0.77$

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] SRC PhysProp database. <http://www.syrres.com>

[3] Krop, H.; de Voogt, P. EU-FP6 PERFORCE (PERFluorinated ORganic Chemicals in the European environment) 2, IBED-ESPM, 2008

[4] 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

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

[6] 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.

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

[8] Bhatarai B., Gramatica P., Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals, Environ. Sci. Technol., 2011, 45, 8120-8128

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