Electronic Supplementary Material (ESI) for Green Chemistry.
This journal is © The Royal Society of Chemistry 2023

Supplementary Information 2

Towards higher scientific validity and regulatory acceptance of predictive models for PFAS

Authors: Anita Sosnowska[1*], Natalia Bulawska[1], Dominika Kowalska[1], Tomasz Puzyn[1,2*]

1. QSAR Lab, ul. Trzy Lipy 3, Gdańsk, Poland

2. University of Gdansk, Faculty of Chemistry, Wita Stwosza 63, 80-308 Gdansk, Poland

Contents

VP1. Prediction of Vapor Pressure for Aquatic Partitioning of Perfluorinated Chemicals

1. QSAR identifier

1.1. QSAR identifier (title):

VPrediction of Vapor Pressure for Aquatic Partitioning of Perfluorinated Chemicals

1.2. Other related models:

Gramatica, P.; Cassani, S.; Chirico, N. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. J. Comput. Chem. 2014, 35, 1036–1044.

QDB archive DOI: 10.15152/QDB.177

Property M20.logVP: Vapor pressure as log(VP) [log(mm Hg)]

1.3. Software coding the model:

Mobydigs sofware

Todeschini,R.;Consonni,V.;Pavan,M.MOBYDIGS—Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows, Talete srl: Milan, Italy, 2002

2. General information

2.1. Date of QMRF:

6/12/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.
Trzy Lipy 3 Street, Building B,
80-172 Gdansk, Poland
+48 795 160 760
contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

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/

2.6. Date of model development and/or publication:

Published in 2011

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

1. Bhhatarai, B., & Gramatica, P. (2011). Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals. Environmental Science & Technology, 45(19), 8120–8128 https://pubs.acs.org/doi/10.1021/es101181g

2. DRAGON software https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_ molecular_descriptor_calculations

2.8. Availability of information about the model:

The three descriptors based MLR model on 35 compounds was obtained based on theoretical molecular descriptors selected by GA.

Supporting info: https://doi.org/10.1021/es101181g

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Vapor pressure

3.3. Comment on the endpoint:

Vapor pressure: log P_L (mmHg)

3.4. Endpoint units:

mmHg

3.5. Dependent variable:

logVP

3.6. Experimental protocol:

The data given in 20°C temperature were extrapolated for 25°C using the Wagner and Antoine equation.

3.7. Endpoint data quality and variability:

In the case of VP (in mmHg), 24 compounds from SRC data reported at 25°C and 11 additional compounds from PERFORCE reported at 20 or 25°C for liquid or subcooled liquid were added (total 35 compounds). The data given in 20°C temperature were extrapolated for 25°C using the Wagner and Antoine equation as reported in the original articles.

The data from different lab sources were used in the current study only if the experimental settings and conditions are identical.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSPR - Multiple linear regression model (OLS - Ordinary Least Square)

4.2. Explicit algorithm:

Full model equation:

$\log VP = 7.97(\pm 1.26) - 0.16(\pm 0.02)\,F03[C - F] - 3.16(\pm 0.92)AAC - 0.64(\pm 0.40)nDB$

4.3. Descriptors in the model:

1. F03[C-F] - 2D frequency fingerprint descriptor, frequency of C-F at topological distance
2. AAC - 2D information indices
3. nDB - constitutional descriptor, number of double bonds

4.4. Descriptor selection:

The reduced sets of input descriptors were subjected to variable selection method using Genetic Algorithm (GA). Genetic algorithm was applied of the set for approximately 400 molecular descriptors for each compound.GA was applied to choose the best set of few descriptors, which have the most relevant variables, in combination, in modeling studied property.

4.5. Algorithm and descriptor generation:

The input files for descriptor calculation were obtain by the semi empirical AM1 method (minimized their lowest energy conformation) using HYPERCHEM software.

Molecular descriptors were generated using DRAGON software.

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

4.6. Software name and version for descriptor generation:

1. DRAGON software

A software to calculate theoretical molecular descriptors

https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_molecular_descriptor_calculations

2. HYPERCHEM

Software used for molecular drawing and conformational energy optimization AM1 www.hyper.com

4.7. Chemicals/ Descriptors ratio:

Full model: 35 chemicals / 3 descriptors = 11.67

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The Williams plot was used to verify the presence of response outliers (i.e., compounds with cross-validated standardized residuals greater than 2.5 standard deviation units, and chemicals very influential for their structure in determining model parameters (i.e., compounds with high leverage value (h) ($h > h^*$, the critical value being $h^* = 3p' /n$, where p' is the number of model variables plus one, and n is the number of the objects used to calculate the model). The leverage approach was applied also for the definition of the structural chemical domain of each model for chemicals without experimental data by plotting Y-predicted versus hat value. The predictions for compounds having high leverage value are extrapolated and should be considered less reliable, but those interpolated within the training domain should be predicted with similar accuracy as for training chemicals.

5.2. Method used to assess the applicability domain:

To obtain structural AD the leverage approach providing a cut-off value was used ($h^* = 0.343$).

5.3. Software name and version for applicability domain assessment:

Mobydigs sofware

Todeschini,R.; Consonni,V.;Pavan, M.MOBYDIGS—Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows, Talete srl: Milan, Italy, 2002

5.4. Limits of applicability:

Full model: Only 14 compounds were out of the AD plot with the cutoff being the average hat value (vertical line) of 0.343. Thus, the coverage of our model on the analyzed PFCs is 93.7%. The compounds with very high leverage value which are found outside the AD were perfluoro-perhydro-fluoranthene (CAS 662-28-2), tricyclic perfluorinated compound (CAS 306-91-2) followed by a C-17 perfluoro iodide (CAS 29809- 35-6), in terms of decreasing distance from the average hat value. Apart from those compounds, other out of AD compounds were mostly saturated bicyclic or tricyclic chemicals, acids (CAS 16517-11-6 and CAS 18122-53-7) or C-13 and C-15 long chain esters and iodides. The inapplicability of model to these compounds is justified by the chemicals used in the training set that are mainly linear and short chain PFCs.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: No

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable (response) 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=35) was split - before model development - into training and prediction sets. Training set: used to model development Prediction set: used to external validation. Two different splitting criteria were used: (1) splitting by structural similarity SOM - 31% (n train = 24) (2) random by activity - 37% (n train = 22)

6.6. Pre-processing of data before modelling:

The original VP data were expressed in log unit log VP(mmHg)

6.7. Statistics for goodness-of-fit:

a) Split by SOM: $R^2 = 91.07$, $RMSE_{TR} = 0.83$

b) Random by activity: $R^2 = 93.75$, $RMSE_{TR} = 0.64$

c) Full model: $R^2 = 90.93$, $RMSE_{TR} = 0.83$

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

a) Split by SOM: $Q^2_{LOO} = 84.33$

b) Random response activity: $Q^2_{LOO} = 91.23$

c) Full model: $Q^2_{LOO} = 88.21$

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

a) Split by SOM: $R^2Y_{SCR} = 12.69$

b) Random response activity: $R^2Y_{SCR} = 14.08$

c) Full model: $R^2Y_{SCR} = 8.95$

6.11. Robustness – Statistics obtained by bootstrap:

a) Split by SOM: $Q^2_{BOOT} = 81.63$

b) Random response activity: $Q^2_{BOOT} = 82.13$

c) Full model: $Q^2_{BOOT} = 86.06$

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Availability information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: No

Formula: No

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 (response) for the external validation set:

All

7.5. Other information about the external validation set:

Approximately 25-30% of the whole data set was used for prediction set based on the availability of data and the distribution of chemical classes. Set was used after model development in order to external validation.

7.6. Experimental design of test set:

To verify the predictive capability of the proposed models, the whole dataset (n=35) was split - before model development - into training and prediction sets. Training set: used to model development Prediction set: used to external validation. Two different splitting criteria were used: (1) splitting by structural similarity SOM - 31% (n validation = 11) (2) random by activity - 37% (n validation = 13)

7.7. Predictivity – Statistics obtained by external validation:

a) Split by SOM:

$Q^2_{F1} = 87.00$, $Q^2_{F3} = 87.78$, $RMSE_{EXT} = 0.97$

b) Random response activity:

$Q^2_{F1} = 86.26$, $Q^2_{F3} = 80.36$, $RMSE_{EXT} = 1.14$

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

An external validation was performed for the developed model after splitting the compounds into a validation and training set. Despite access to data for all compounds used in the model, they were not added to the corresponding sets.

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

The model was developed by statistical approach. No mechanistic was defined a priori.

8.2. A priori or a posteriori mechanistic interpretation:

$$\log VP = 7.97(\pm 1.26) - 0.16(\pm 0.02)F03[C - F] - 3.16(\pm 0.92)AAC - 0.64(\pm 0.40)nDB$$

where:

F03[C-F] - 2D frequency fingerprint descriptor, frequency of C-F at topological distance

AAC - 2D information indices

nDB - constitutional descriptor, number of double bonds

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Bhhatarai, B., & Gramatica, P. (2011). Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals. Environmental Science & Technology, 45(19), 8120–8128 https://pubs.acs.org/doi/10.1021/es101181g

2. Shi, L. M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; Blair, R. M.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. QSAR models using a large diverse set of estrogens. J. Chem. Inf. Comput. Sci. 2001, 41, 186–195.

3. Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the definition of the Q2 parameter for QSAR validation. J. Chem. Inf. Model 2009, 49, 1669–1678.

9.3. Supporting information:

Supporting information: https://doi.org/10.1021/es101181g

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSPR PaDEL-Descriptor model for Vapor Pressure prediction of PFC

1.2.Other related models:

Bhhatarai B., Gramatica P., 2011, 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:

05/12/2013

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:

No information available

3.2.Endpoint:

1.Physicochemical effects 1.4.Vapour pressure

3.3.Comment on endpoint:

Vapor Pressure (VP) is the pressure exerted by a vapor in equilibrium with the solid or liquid phase of the same substance.

3.4.Endpoint units:

mmHg

3.5.Dependent variable:

LogVP

3.6.Experimental protocol:

24 compounds from SRC PhysProp database[2] reported at 25 °C and 11 additional compounds from EU-FP6 PERFORCE report[3] at 20° or 25 °C for liquid or subcooled liquid were added (total 35 compounds). The data given in 20 °C temperature were extrapolated for 25 °C using the Wagner and Antoine equation

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:

LogVP (Full model)
OLS-MLR method. Model developed on a training set of 35 compounds

LogVP (Split by SOM modell)
OLS-MLR method. Model developed on a training set of 24 compounds

LogVP (Split by Ordered Response model)
OLS-MLR method. Model developed on a training set of 22 compounds

Full model equation: logVP= 4.47 - 0.46 nH - 0.20 nSF - 2.41 nHBint2

Split by SOM model equation: logVP= 4.47 - 2.71 nHBint2 - 0.19 nsF - 0.43 nH

Split by Ordered Response model equation: logVP= 4.64 - 0.24 nsF - 0.45 nH - 2.07 nHBint2

4.3.Descriptors in the model:

[1]nH Number of hydrogen atoms
[2]nSF Count of atom-type E-State: -F
[3]nHBint2 Count of E-State descriptors of strength for potential Hydrogen Bonds of path length 2

4.4.Descriptor selection:

A total of 717 molecular descriptors of differing types (0D, 1D, 2D) 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 123 molecular descriptors were used as input variables for variable subset selection. The models were initially developed by the all-subset-procedure, and then GA was applied to obtain the final population of models (three variables). 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: 35 chemicals / 3 descriptros = 11.67
Split by SOM: 24 chemicals / 3 descriptors = 8
Split by Ordered response: 22 chemicals / 3 descriptors = 7.34

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: logVP values: -6.37 / 4.39
Range of descriptor values: nH: 0 / 15; nHBint2: 0 / 2; nsF: 3 / 27

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*=0.343). 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>0.343 (h*): 2-{ethyl[(heptadecafluorooctyl)sulfonyl]amino}ethyl acrylate (423-82-5), Sulfluramid [ISO] (4151-50-2); Outliers for response, standardised residuals > 2.5 standard deviation units: no

Split by SOM model domain: outliers for structure, hat>0.500 (h*): 2-{ethyl[(heptadecafluorooctyl)sulfonyl]amino}ethyl acrylate (423-82-5), Sulfluramid [ISO] (4151-50-2), N-ethyl-

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 24 perfluorinated compounds with a range of logVP values from -6.37 to 4.39.

The training set of the **Split by Ordered Response Model** consists of 22 perfluorinated compounds with a range of logVP values from -4.82 to 3.82.

6.6.Pre-processing of data before modelling:

The data was used as LogVP mmHg; The data given in 20 °C temperature were extrapolated for 25 °C using the Wagner and Antoine equation

6.7.Statistics for goodness-of-fit:

Split by SOM Model:

R^2:0.95 ; CCCtr[4]:0.97 ; RMSEtr: 0.65

Split by Ordered Response Model:

R^2: 0.92 ; CCCtr: 0.96 ; RMSEtr:0.71

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

Split by SOM Model:

Q^2loo:0.92 ; CCCcv: 0.96 ; RMSEcv: 0.81

Split by Ordered Response Model:

Q^2loo: 0.89 ; CCCcv: 0.94 ; RMSEcv: 0.83

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

Split by SOM Model: Q^2LMO: 0.90

Split by Ordered Response Model: Q^2LMO: 0.86

6.10.Robustness - Statistics obtained by Y-scrambling:

Split by SOM Model: R^2Yscr:0.13

Split by Ordered Response Model: R^2Yscr: 0.14

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: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=35) 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 =11).

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 three chemicals was put in the prediction 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= 11 ; R^2ext = 0.98; Q^2ext F1[5] = 0.88; Q^2 ext F2[6] = 0.88; Q^2 ext F3[7] = 0.89; CCCex = 0.94; RMSEex = 0.94 ; MAEex =0.81 .

Split by Oredered Response model: n prediction= 13 ; R^2ext = 0.94 ; Q^2ext F1= 0.92; Q^2 ext F2 = 0.92; Q^2 ext F3 = 0.89; CCCex = 0.96; RMSEex = 0.86; MAEex = 0.71.

7.8.Predictivity - Assessment of the external validation set:

Range of response for prediction set (**SOM split**, n=11) compounds:

logVP (mmHg): -4.82 / 3.82 (range of corrispondig training set: -6.37 / 4.39)

Range of modeling descriptors for prediction set (**SOM split**, n=11) compounds:

nH: 0 / 12 (range of corrispondig training set: 0 / 10)
nHBint2: 0 / 1 (range of corrispondig training set: 0 / 2)
nsF: 3 / 21 (range of corrispondig training set: 4 / 27)

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

logVP: -6.37 / 4.39 (range of corrispondig training set: -4.82 / 3.82)

Range of modeling descriptors for prediction set (**Ordered Response split**, n=13) compounds: nH: 0 / 10 (range of corrispondig training set: 0 / 12)

nHBint2: 0 / 2 (range of corrispondig training set: 0 / 1) nsF: 3 / 27 (range of corrispondig training set: 3 / 23) 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 [8] is:

logVP= 7.97 - 0.16 F03[C-F] - 3.16 AAC - 0.64 nDB

where F03[C-F]: 2D frequency fingerprint descriptor, meaning the frequency of C-F at topological distance 03 (these values are higher for branched and cyclic compounds)

AAC: 2Dinformation indices, particularly mean information index on atomic composition (increases with higher atomic weight atoms or larger molecule)

nDB: 0D constitutional descriptor is the number of double bonds

All these descriptors being inversely related to VP have an influence in decreasing the vapor pressure.

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

logVP= 4.47 - 0.46 nH - 0.20 nSF - 2.41 nHBint2

where nH= Number of hydrogen atoms

nSF= Count of atom-type E-State: -F nHBint2= Count of E-State descriptors of strength for potential Hydrogen Bonds of path length 2 In the two models there is an high correlation between F03[C-F] and nSF (0.96), which encode for the same structural information related to fluorine atoms, and acceptable correlation also among AAC and nH (0.73).

8.3.Other information about the mechanistic interpretation:

no other information available

9.Miscellaneous information

9.1.Comments:

To predict VP 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=35), 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: logVP= 4.47 - 0.46 nH - 0.20 nSF - 2.41 nHBint2

N = 35; R^2 = 0.93 ; Q^2 = 0.91 ; Q^2LMO = 0.90; CCC = 0.96; CCCcv = 0.95; RMSE= 0.72; RMSEcv = 0.82 .

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]Bhhatarai B., Gramatica P., Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals, Environ. Sci. Technol., 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

VP3. Predictive model for physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors

1. QSAR identifier

1.1. QSAR identifier (title):

Predictive model for physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors.

1.2. Other related models:

N/A

1.3. Software coding the model:

Linear regression was performed using **Sigmaplot 10 (Systat Software Inc, Point Rich)**

2. General information

2.1. Date of QMRF:

09/11/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com

https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Minhee Kim Department of Civil Engineering, University of British Columbia, 6250 Applied Science Lane, Vancouver, BC, Canada V6T 1Z4

2. Loretta Y. Li Department of Civil Engineering, University of British Columbia, 6250 Applied Science Lane, Vancouver, BC, Canada V6T 1Z4

3. John R. Grace Department of Chemical and Biological Engineering, University of British Columbia, 2360 East Mall, Vancouver, BC, Canada V6T 1Z3

4. Chaoyang Yue Department of Chemical and Biological Engineering, University of British Columbia, 2360 East Mall, Vancouver, BC, Canada V6T 1Z3

2.6. Date of model development and/or publication:

3 November 2014

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

Minhee Kim, Loretta Y. Li, John R. Grace, Chaoyang Yue, Selecting reliable physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors, Environmental Pollution, Volume 196, 2015, Pages 462-472, ISSN 0269-7491. https://doi.org/10.1016/j.envpol.2014.11.008.

2.8. Availability of information about the model:

Molecular structures, correlation equations and statistical results of 23 PFASs encompassing 10 PFCAs, 3 PFSAs, 6 FASAs and 4 FTOHs are represented in Supplementary Information.

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Vapor pressure

3.3. Comment on the endpoint:

Vapor pressure determines the exchange rate of compounds across an air/water interface, volatilization from water, soil or plants, and transport of trace organics throughout the global environment. Selected experimental log P_L values were collected from measured data based on indirect and direct measurement methods

3.4. Endpoint units:

P_L pressure units in Pa

3.5. Dependent variable:

log P_L

3.6. Experimental protocol:

Steele et al. (2002a, b) determined the vapor pressures of per fluoroheptanoic acid (PFHpA) and perfluorobutanoic acid (PFBA) using a twin ebulliometric apparatus. Kaiser et al. (2005) obtained the vapor pressures of perfluorooctanoic (PFOA), -nonanoic (PFNA), -decanoic (PFDA), -undecanoic (PFUnA) and -dodecanoic (PFDoA) acids using a dynamic method developed by Scott (1986). Lei et al. (2004) estimated the liquid-phase vapor pressures of FTOHs and ethyl perfluorooctane sulfonamidoethanol (EtFOSE) as a function of temperature, based on a gas chromatographic retention time (GC RT) technique which utilizes the consistent retention time of a compound in a GC column at a given temperature.

3.7. Endpoint data quality and variability:

Calculated properties of PFASs were obtained from U.S. EPA's, EPI Suite. Selected experimental log P_L values were collected from measured data based on indirect and direct measurement methods. The above methods were successfully employed for reference compounds. The resulting data sets are used for QSPR analysis of PFAS properties.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSPR

Simple linear regression model

4.2. Explicit algorithm:

log P_L= -0.0081 x V_M + 3.4361

4.3. Descriptors in the model:

Molar volume (V_M) is the volume occupied by one mole of a chemical element or a chemical compound. Units (cm3 mol-1). Geometric descriptor derived from 3-D structures.

4.4. Descriptor selection:

Vapor pressure was correlated to F_N, V_M and TSA. Vapor pressure is calculated from single descriptor model selected by statistical results and checked for internal consistency. The best model was selected based on statistical parameters.

4.5. Algorithm and descriptor generation:

Molar volume is a geometric descriptor derived from 3-D structure.

4.6. Software name and version for descriptor generation:

ACD/Labs' ACD/PhysChem Suite

https://www.acdlabs.com/solutions/academia/?gclid=CjwKCAiA1aiMBhAUEiwACw25MUw_ZleJmBKmPqJ5ALizlwu6tzK5G8Sk8LycZpNlbT2UgRwAp20uthoCE9EQAvD_BwE

4.7. Chemicals/ Descriptors ratio:

10 chemicals/1 descriptor

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

N/A

5.2. Method used to assess the applicability domain:

N/A

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

N/A

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN:Yes

Chemical Name:Yes

Smiles:Yes

Formula:No

INChI:No

MOL file:No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

N/A

6.6. Pre-processing of data before modelling:

Data was converted to a log scale for modeling.

6.7. Statistics for goodness-of-fit:

$r^2 = 0.8981$

F =70.48

S.E.= 0.201

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

Cross-validated coefficients using the leave-one-out technique (LOO-Rcv^2) were calculated from the equation: LOO-Rcv^2=1 - PRESS/TSS, where PRESS is the prediction error sum of squares and TSS is the total sum of squares.

Q^2_{LOO}= 0.8597

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

N/A

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

N/A

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

The model was developed by statistical approach. No mechanistic was defined a priori.

8.2. A priori or a posteriori mechanistic interpretation:

The volatility parameter, log P_L, is correlated with F_N, V_M and TSA of PFASs. In the statistical results, the P values of three single descriptor models were significant at the <0.05 level of significance for their respective regression models. However, the model based on V_M was statistically better than those based on F_N and TSA, because it has the highest Q^2 (0.860), largest r^2 (0.898) and smallest S.E. (0.201).

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Steele, W.V., Chirico, R.D., Knipmeyer, S.E., Nguyen, A., 2002a. Vapor pressure, heat capacity, and density along the saturation line: measurements for benzenamine, butylbenzene, secbutylbenzene, tert-butylbenzene, 2,2- dimethylbutanoic acid, tridecafluoroheptanoic acid, 2-butyl-2-ethyl-1,3- propanediol, 2,2,4-trimethyl-1,3-pentanediol, and 1-chloro-2-propanol. J. Chem. Eng. Data 47, 648e666

2. Steele, W.V., Chirico, R.D., Knipmeyer, S.E., Nguyen, A., 2002b. Measurements of vapor pressure, heat capacity, and density along the saturation line for cyclopropane carboxylic acid, N,Ndiethylethanolamine,

2,3-dihydrofuran, 5-hexen- 2-one, perfluorobutanoic acid, and 2-phenylpropionaldehyde. J. Chem. Eng. Data 47, 715e724

3. Kaiser, M.A., Larsen, B.S., Kao, C.P.C., Buck, R.C., 2005. Vapor pressures of perfluorooctanoic, -nonanoic, -decanoic, -undecanoic, and -dodecanoic acids. J. Chem. Eng. Data 50, 1841e1843

4. Scott, L.S., 1986. Determination of activity coefficients by accurate measurement of boiling point diagram. Fluid Phase Equilibria 26, 149e163.

5. Lei, Y.D., Wania, F., Mathers, D., Mabury, S.A., 2004. Determination of vapor pressures, octanol/air, and water/air partition coefficients for polyfluorinated sulfonamide, sulfonamidoethanols, and telomer alcohols. J. Chem. Eng. Data 49, 1013e1022.

9.3. Supporting information:

Supplementary data can be found at: *http://dx.doi.org/10.1016/j.envpol.2014.11.008.*

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

VP4. Predictive models for estimating the vapor pressure of poly- and perfluorinated compounds at different temperatures

1. QSAR identifier

1.1. QSAR identifier (title):

Predictive models for estimating the vapor pressure of poly- and perfluorinated compounds at different temperatures.

1.2. Other related models:

N/A

1.3. Software coding the model:

Simca-S software (Version 6.0, Umetri AB & Erisoft AB

2. General information

2.1. Date of QMRF:

04/01/2022

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.
Trzy Lipy 3 Street, Building B,
80-172 Gdansk, Poland
+48 795 160 760
contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

Willie J.G.M. Peijnenburg Laboratory for Ecological Risk Assessment, RIVM, National Institute for Public Health and the Environment

Tel.: +31 30 2743129

willie.peijnenburg@rivm.nl

2.6. Date of model development and/or publication:

Published in 2013

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

Ding, G., Shao, M., Zhang, J., Tang, J., & Peijnenburg, W. J. G. M. (2013). Predictive models for estimating the vapor pressure of poly- and perfluorinated compounds at different temperatures. Atmospheric Environment, 75, 147–152.
https://doi.org/10.1016/j.atmosenv.2013.04.042

2.8. Availability of information about the model:

Temperature-dependent predictive models for vapor pressure of PFCs were developed based on experimental data. More information about training/validation sets and descriptors used in model attached in supplementary material: https://doi.org/10.1016/j.atmosenv.2013.04.042.

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Vapor pressure

3.3. Comment on the endpoint:

Vapor pressure (P) is a key physicochemical property determining the partitioning behavior of a chemical between the gaseous phase and the liquid/solid phase. Therefore, P is often used to assess the ability of a chemical to partition into the gas phase and to assess the potential of a chemical for long-range transport to remote locations via the atmosphere. Accurate assessment of P is especially important for precursors of perfluoroalkyl acids as they have been reported to be the source of PFOA/PFOS in remote locations, such as the Artic area. Furthermore, P values at different temperatures, not only 25°C, are required for the fate assessment of these pollutants, as vapor pressures have strong temperature dependence, and as environmentally important temperatures vary considerably with latitudinal and seasonal variations.

3.4. Endpoint units:

Pa

3.5. Dependent variable:

logP

3.6. Experimental protocol:

N/A

3.7. Endpoint data quality and variability:

P values (Pa) of PFCs at different temperatures were taken from various publications (Kauck and Diesslin, 1951; Crowder et al., 1967; Mousa, 1978; Weber and Defibaugh, 1996; Steele et al., 1997a, 1997b; Shi et al., 1999; Kul et al., 2001; Steele et al., 2002a, 2002b; Dias et al., 2004; Duan et al., 2004; Kaiser et al., 2004; Lei et al., 2004; Shoeib et al., 2004; Stock et al., 2004; Dias et al., 2005; Kaiser et al., 2005; Krusic et al., 2005; Washburn et al., 2005; Cobranchi et al., 2006; Barton et al., 2008, 2009; Feng et al., 2010; Costa et al., 2012). Only experimentally measured values were selected, and a total of 1103 data points for 42 PFCs at different temperatures were collected. Before the data were used for statistical analyses, a logarithm transformation was performed. For QSPR modeling, the data set was split into a training set with 883 data points for model development and an external validation set with 220 data points. A complete list of these vapor pressure data can be found in the electronic Supplementary material.

Supplementary materials: https://doi.org/10.1016/j.atmosenv.2013.04.042.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSPR - partial least-squares (PLS) regression

4.2. Explicit algorithm:

$\log P = c_0 + c_1/T + c_2 logT + c_3 T + c_4 S_1 + c_5 S_2 + c_6 S_3 + \ldots$

4.3. Descriptors in the model:

1. - average molecular polarizability

2. - dipole moment

3. Hf - standard heat of formation [kJ]

4. TE - total energy [eV]

5. EE - electronic energy [eV]

6. CCR - core-core repulsion energy [eV]

7. E_{HOMO} - the energy of the highest occupied molecular orbital [eV]

8. E_{LUMO} - the energy of the highest unoccupied molecular orbital [eV]

9. q_C^- - the most negative net atomic charges on a carbon atom

10. q_C^+ - the most positive net atomic charges on a carbon atom

11. q_H^+ - the most positive net atomic charges on a hydrogen atom

12. q_F^- - the most negative net atomic charges on a fluorine atom

13. q^+ - the most positive net atomic charges on a atom

14. q^- - the most negative net atomic charges on a atom

15. Mw - molecular weight

16. A_{cosmo} - COSMO area [A^2]

17. V_{cosmo} - COSMO volume [A^3]

18. CAA - connolly accessible area [A^2]

19. CMA - connolly molecular area [A^2]

20. CSEV - connolly solvent-excluded volume [A^3]

21. Ov - Ovality - the ratio of the Molecular Surface Area to the Minimum Surface Area

4.4. Descriptor selection:

As various intermolecular interactions, such as dispersion interactions, dipole-dipole interactions, dipole-induced dipole interactions, electrostatic interactions and hydrogen bonding, may govern the magnitude of the vapor pressure, 21 molecular structural descriptors were selected to develop the predictive QSPR model.

4.5. Algorithm and descriptor generation:

The PM6 method was performed with MOPAC2012 program (MOPAC2012, 2012). Molecular structural descriptors: a, m, DHf, TE, EE, CCR, E_{HOMO}, E_{LUMO}, q_C^-, q_C^+, q_H^+, q_F^-, q^+, q^-, Mw, A_{cosmo} and V_{cosmo}, were taken from output files of MOPAC2012. CAA, CMA, CSEV and Ov were calculated by CS Chem3D Ultra, while S0K and PHI were obtained using Dragon (Version 2.1) software package.

4.6. Software name and version for descriptor generation:

1. MOPAC2012

Stewart Computational Chemistry. Stewart J.J.P., Colorado Springs, CO, USA http://OpenMOPAC.net

2. DRAGON

Calculation of several sets of molecular descriptors from molecular geometries (topological, geometrical, WHIM, 3D-MoRSE, molecular profiles, etc.)

Prof. R.Todeschini - distributed by Talete srl, via Pisani 13, 20124 Milano, Italy http://www.disat.unimib.it/chm

4.7. Chemicals/ Descriptors ratio:

1103 data points for 42 PFCs / 21 molecular descriptors

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model

The Williams plot was used to visualize the AD of selected model (Model (4)). The h* value was calculated to be 0.0442. It was found that the data points numbered 69-73, 583-593, 644-

646, 652-659, 660-667 in the training set and 900, 1028-1030, 1044, 1046-1049 in the validation set had hi values great than h*. Therefore, those data in the training set were influential points that could affect the model development, whereas the predictions of those data points in the validation set might be unreliable. From the standardized residuals, the points 69-80, 583, 616-619, 652-659 and 860 in the training set and 900-902, 1028, 1037, 1046, 1047, 1097 in the validation set were identified as response outliers, the log P values of which were not well predicted. Therefore, the data points 69-73, 583 and 652-659 in the training set were determined as outliers of Model (4), as they had hi values great than h*, and standardized residuals greater than 3 or less than 3. Authors of the model removed outliers from model (4) which resulting final model (6). Removed compounds have been pointed in Supplementary material.

5.2. Method used to assess the applicability domain:

The Williams plot was used to visualize the AD of selected model. The h* value was calculated to be 0.0442. It was found that the data points numbered 69-73, 583-593, 644-646, 652-659, 660-667 in the training set and 900, 1028-1030, 1044, 1046-1049 in the validation set had h values great than h*. Therefore, those data in the training set were influential points that could affect the model development, whereas the predictions of those data points in the validation set might be unreliable. From the standardized residuals, the points 69-80, 583, 616-619, 652-659 and 860 in the training set and 900-902, 1028, 1037, 1046, 1047, 1097 in the validation set were identified as response outliers, the log P values of which were not well predicted. Therefore, the data points 69-73, 583 and 652-659 in the training set were determined as outliers of Model (4), as they had hi values great than h*, and standardized residuals greater than 3 or less than 3. Authors of the model removed outliers from model (4) which resulting final model (6). Removed compounds have been pointed in Supplementary material.

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

From the standardized residuals, the points 69-80, 583, 616-619, 652-659 and 860 in the training set and 900-902, 1028, 1037, 1046, 1047, 1097 in the validation set were identified as response outliers, the log P values of which were not well predicted. Therefore, the data points 69-73, 583 and 652-659 in the training set were determined as outliers of Model (4), as they had hi values great than h*, and standardized residuals greater than 3 or less than 3.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN: Yes

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

Number of data points in training set: 831

6.6. Pre-processing of data before modelling:

Before the data were used for statistical analyses, a logarithm transformation was performed.

6.7. Statistics for goodness-of-fit:

$R^2_{X(adj)(cum)} = 0.979$

$R^2_{Y(adj)(cum)} = 0.976$

$r = 0.988$

RMSE = 0.157

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

N/A

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Availability information for the external validation set:

CAS RN: Yes

Chemical Name: No

Smiles: No

Formula: No

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 (response) for the external validation set:

All

7.5. Other information about the external validation set:

Number of data points in the external validation set: 220

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

$Q^2_{EXT} = 0.978$

$Q^2_{cum} = 0.976$

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

The main factors governing the values of log P of PFCs, are intermolecular dispersive interactions, hydrogen bonding, temperature, intermolecular dipole-induced dipole interactions and dipolee dipole interactions.

8.2. A priori or a posteriori mechanistic interpretation:

$\log P = c_0 + c_1/T + c_2 \log T + c_3 T + c_4 S_1 + c_5 S_2 + c_6 S_3 + \ldots$

where:
$c_0, c_1, c_2, c_3, c_4, \ldots, c_n$ – regression coefficients
$S_1, S_2, S_3, S_4, \ldots, S_n$ – stand for various molecular structural descriptors

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

Ding, G., Shao, M., Zhang, J., Tang, J., & Peijnenburg, W. J. G. M. (2013). Predictive models for estimating the vapor pressure of poly- and perfluorinated compounds at different temperatures. Atmospheric Environment, 75, 147–152.
https://doi.org/10.1016/j.atmosenv.2013.04.042

9.3. Supporting information:

Supplementary material: https://doi.org/10.1016/j.atmosenv.2013.04.042.

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

S1. Prediction of Aqueous Solubility for Aquatic Partitioning of Perfluorinated Chemicals

1. QSAR identifier

1.1. QSAR identifier (title):

Prediction of Aqueous Solubility for Aquatic Partitioning of Perfluorinated Chemicals

1.2. Other related models:

Gramatica, P.; Cassani, S.; Chirico, N. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. J. Comput. Chem. 2014, 35, 1036–1044.

QDB archive DOI: 10.15152/QDB.177

Property M15.logSw: Water solubility as log(Sw) [log(mg/L)]

1.3. Software coding the model:

Mobydigs sofware

Todeschini,R.;Consonni,V.;Pavan,M.MOBYDIGS—Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows, Talete srl: Milan, Italy, 2002

2. General information

2.1. Date of QMRF:

6/12/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.
Trzy Lipy 3 Street, Building B,
80-172 Gdansk, Poland
+48 795 160 760
contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

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/

2.6. Date of model development and/or publication:

Published in 2011

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

1. Bhhatarai, B., & Gramatica, P. (2011). Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals. Environmental Science & Technology, 45(19), 8120–8128 https://pubs.acs.org/doi/10.1021/es101181g

2. DRAGON software https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_ molecular_descriptor_calculations

2.8. Availability of information about the model:

The two descriptors based MLR model on 20 compounds was obtained based on theoretical molecular descriptors selected by GA.

Supporting info: https://doi.org/10.1021/es101181g

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Water Solubility

3.3. Comment on the endpoint:

Water Solubility: logAqS (mg/L)

3.4. Endpoint units:

Mg/L

3.5. Dependent variable:

logAqS

3.6. Experimental protocol:

Shake flask method was used for most of the data unless indicated in the reference cited. For a strongly hydrophobic compound (CAS 1691-99-2) the data reported using “generator columns” experiment was used. Both methods are incorporated as standard OECD tests on solubility. For FTOH (CAS 865-86-1), solubility was determined with the log-linear cosolvency approach.

3.7. Endpoint data quality and variability:

The experimental data were collected from SRC PhysProp database, literature and the datacompiled in EU-FP6 PERFORCE report.

For AqS (in mg/L), data reported (20 compounds) at temperature range of 293-298K in PERFORCE report were used.

The data from different lab sources were used in the current study only if the experimental settings and conditions are identical.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSPR - Multiple linear regression model (OLS - Ordinary Least Square)

4.2. Explicit algorithm:

Full model equation:

logAqS = -0.418(±1.940) - 0.003(±0.001)T(F..F) + 5.185(±3.849)SIC1

4.3. Descriptors in the model:

1. T(F.F.) - topological descriptor, represents the sum of distance between pair of fluorine atoms

2. SIC1 - descriptor based on neighbor degrees and edge multiplicity, mainly give information on the structural symmetry in the molecule

4.4. Descriptor selection:

The reduced sets of input descriptors were subjected to variable selection method using Genetic Algorithm (GA). Genetic algorithm was applied of the set for approximately 400 molecular descriptors for each compound.GA was applied to choose the best set of few descriptors, which have the most relevant variables, in combination, in modeling studied property.

4.5. Algorithm and descriptor generation:

The input files for descriptor calculation were obtain by the semi empirical AM1 method (minimized their lowest energy conformation) using HYPERCHEM software.

Molecular descriptors were generated using DRAGON software.

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

4.6. Software name and version for descriptor generation:

1. DRAGON software

A software to calculate theoretical molecular descriptors

https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_molecular_descriptor_calculations

2. HYPERCHEM

Software used for molecular drawing and conformational energy optimization AM1 www.hyper.com

4.7. Chemicals/ Descriptors ratio:

Full model: 20 chemicals / 2 descriptors = 10

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The Williams plot was used to verify the presence of response outliers (i.e., compounds with cross-validated standardized residuals greater than 2.5 standard deviation units, and chemicals very influential for their structure in determining model parameters (i.e., compounds with high leverage value (h) (h > h*, the critical value being h* = 3p' /n, where p' is the number of model variables plus one, and n is the number of the objects used to calculate the model). The leverage approach was applied also for the definition of the structural chemical domain of each model for chemicals without experimental data by plotting Y-predicted versus hat value. The predictions for compounds having high leverage value are extrapolated and should be considered less reliable, but those interpolated within the training domain should be predicted with similar accuracy as for training chemicals.

5.2. Method used to assess the applicability domain:

To obtain structural AD the leverage approach providing a cut-off value was used (h* = 0.45).

5.3. Software name and version for applicability domain assessment:

Mobydigs sofware

Todeschini,R.;Consonni,V.;Pavan,M.MOBYDIGS—Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows, Talete srl: Milan, Italy, 2002

5.4. Limits of applicability:

For structural AD, 201 extra compounds were used along with 20 of the experimental data, the average hat value obtained was 0.450 and only 27 compounds were found outside the structural applicability domain (87.8% coverage). Out of these CAS 29809-35-6, CAS 24151-81-3, CAS 16517-11-6, and CAS 355-49-7 have very high hat value corresponding to long chain and bulky structure of PFCs.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: No

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable (response) 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=20) was split - before model development - into training and prediction sets. Training set: used to model development. Prediction set: used to external validation. Two different splitting criteria were used: (1) splitting by structural similarity SOM - 25% (n train = 15) (2) random by activity - 25% (n train = 15).

6.6. Pre-processing of data before modelling:

The original AqS data were expressed in log unit logAqS (mg/L)

6.7. Statistics for goodness-of-fit:

a) Split by SOM: $R^2 = 79.93$, $RMSE_{TR} = 0.90$

b) Random by activity: $R^2 = 74.12$, $RMSE_{TR} = 0.93$

c) Full model: $R^2 = 76.31$, $RMSE_{TR} = 0.84$

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

a) Split by SOM: $Q^2_{LOO} = 62.60$

b) Random response activity: $Q^2_{LOO} = 63.16$

c) Full model: $Q^2_{LOO} = 69.13$

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

a) Split by SOM: $R^2Y_{SCR} = 13.80$

b) Random response activity: $R^2Y_{SCR} = 14.45$

c) Full model: $R^2Y_{SCR} = 10.39$

6.11. Robustness – Statistics obtained by bootstrap:

a) Split by SOM: Q^2_{BOOT} = 64.85

b) Random response activity: Q^2_{BOOT} = 63.01

c) Full model: Q^2_{BOOT} = 65.31

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Availability information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: No

Formula: No

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 (response) 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 whole dataset (n=20) was split - before model development - into training and prediction sets. Training set: used to model development. Prediction set: used to external validation. Two different splitting criteria were used: (1) splitting by structural similarity SOM - 25% (n validation = 5) (2) random by activity - 25% (n validation = 5).

7.6. Experimental design of test set:

Data set for every endpoint was split into training and prediction set. Training for model development, prediction for validation.

Two different approaches were used:

(i) 'self-organizing map' (SOM) by using k-NN

(ii) random selection through property sammpling

7.7. Predictivity – Statistics obtained by external validation:

a) Split by SOM:

Q^2_{F1} = 79.21, Q^2_{F3} = 82.73, $RMSE_{EXT}$ = 0.74

b) Random response activity:

Q^2_{F1} = 87.94, Q^2_{F3} = 92.76, $RMSE_{EXT}$ = 0.49

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

An external validation was performed for the developed model after splitting the compounds into a validation and training set. Despite access to data for all compounds used in the model, they were not added to the corresponding sets.

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

The model was developed by statistical approach. No mechanistic was defined a priori.

8.2. A priori or a posteriori mechanistic interpretation:

logAqS = -0.418(±1.940) - 0.003(±0.001)T(F..F) + 5.185(±3.849)SIC1

where:
T(F.F.) - topological descriptor, represents the sum of distance between pair of fluorine atoms
SIC1 - descriptor based on neighbour degrees and edge multipilicity, mailny give information on the structural symmetry in the molecule

The two descriptors which appear in the model are both bidimensional. The regression coefficient of T(F..F) is very low compared to that of SIC1, the standardized coefficient of T(F..F) is higher (0.741) than that of SIC1 (-0.342) indicating that T(F..F) is more important descriptor than SIC1. Thus, the distance of fluorine atoms in the structure is a dominant factor. For SOM split, only 10:2 FTOH (CAS 865-86-1) came out as structural outlier while no response outlier was seen. For the response split and the full model none of the compounds are seen as response or structural outlier.

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Bhhatarai, B., & Gramatica, P. (2011). Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals. Environmental Science & Technology, 45(19), 8120–8128 https://pubs.acs.org/doi/10.1021/es101181g

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

3. SRC PhysProp database. http://www.syrres.com (accessed October 7, 2010).

9.3. Supporting information:

Supporting information: https://doi.org/10.1021/es101181g

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

1.QSAR identifier

1.1.QSAR identifier (title):

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

1.2.Other related models:

Bhhatarai 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 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:

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 (Sw) 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:

logSw

3.6.Experimental protocol:

The experimental data were collected from SRC PhysProp database[2] and the datacompiled in EU-FP6 PERFORCE report[3]. For Sw (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:

LogSw (Full model)

OLS-MLR method. Model developed on a training set of20 compounds

LogSw (Split by SOM modell)

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

LogSw (Split by Ordered Response model)

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

Full model equation: logSw= 4.02 - 0.86 XLogP + 1.89 PubchemFP344

Split by SOM model equation: logSw= 4.22 - 0.92 XLogP + 1.98 PubchemFP344

Split by Ordered Response model equation:logSw= 3.93 - 0.84 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 descriptros = 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 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(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>0.450 (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>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^2loo: 0.82; CCCcv:0.90 ; RMSEcv: 0.76

Split by Ordered Response Model:

Q^2loo: 0.77; CCCcv: 0.88; RMSEcv: 0.88

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

Split by SOM Model: Q^2LMO: 0.70

Split by Ordered Response Model: Q^2LMO: 0.69

6.10.Robustness - Statistics obtained by Y-scrambling:

Split by SOM Model: R^2Yscr: 0.15

Split by Ordered Response Model: R^2Yscr: 0.14

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: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 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= 5; R^2ext = 0.79 ; Q^2ext F1[5] = 0.73; Q^2 ext F2[6] = 0.72; Q^2 ext F3[7] = 0.78; CCCex = 0.88; RMSEex = 0.84; MAEex = 0.80.

Split by Ordered Response model: n prediction= 5; R^2ext = 0.93 ; Q^2ext F1= 0.91; Q^2 ext F2 = 0.89; Q^2 ext F3 = 0.94; CCCex = 0.93; RMSEex = 0.43; MAEex = 0.42.

7.8.Predictivity - Assessment of the external validation set:

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

logSw (mg/L): -1.96 / 2.20 (range of corrispondig 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 corrispondig training set: 2.68 / 7.45)

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

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

logSw: -0.83 / 2.99 (range of corrispondig 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 corrispondig training set: 1.18 / 9.21)

PubchemFP344: 0 / 1 (range of corrispondig 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 Bhhatarai B. and Gramatica P. [8] is:

logSw= -0.418 - 0.003 T(F..F) + 5.185 SIC1

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

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

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

PaDEL equation: logSw= 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: logSw= 4.02 - 0.86 XLogP + 1.89 PubchemFP344

N = 20; R^2 = 0.85; Q^2= 0.80 ; Q^2LMO = 0.73; CCC = 0.92; CCCcv = 0.89 ;RMSE= 0.67; RMSEcv = 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]Bhhatarai 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

S3. Predictive model for physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors.

1. QSAR identifier

1.1. QSAR identifier (title):

Predictive model for physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors.

1.2. Other related models:

N/A

1.3. Software coding the model:

Linear regression was performed using **Sigmaplot 10 (Systat Software Inc, Point Rich)**

2. General information

2.1. Date of QMRF:

09/11/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com

https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Minhee Kim Department of Civil Engineering, University of British Columbia, 6250 Applied Science Lane, Vancouver, BC, Canada V6T 1Z4

2. Loretta Y. Li Department of Civil Engineering, University of British Columbia, 6250 Applied Science Lane, Vancouver, BC, Canada V6T 1Z4

3. John R. Grace Department of Chemical and Biological Engineering, University of British Columbia, 2360 East Mall, Vancouver, BC, Canada V6T 1Z3

4. Chaoyang Yue Department of Chemical and Biological Engineering, University of British Columbia, 2360 East Mall, Vancouver, BC, Canada V6T 1Z3

2.6. Date of model development and/or publication:

3 November 2014

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

Minhee Kim, Loretta Y. Li, John R. Grace, Chaoyang Yue, Selecting reliable physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors, Environmental Pollution, Volume 196, 2015, Pages 462-472, ISSN 0269-7491. https://doi.org/10.1016/j.envpol.2014.11.008.

2.8. Availability of information about the model:

Molecular structures, correlation equations and statistical results of 23 PFASs encompassing 10 PFCAs, 3 PFSAs, 6 FASAs and 4 FTOHs are represented in Supplementary Information.

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Water solubility

3.3. Comment on the endpoint:

Water solubility is a measure of the amount of chemical substance that can dissolve in water at a specific temperature.

3.4. Endpoint units:

S_L- units in mol L^{-1}

3.5. Dependent variable:

log S_L

3.6. Experimental protocol:

Experimental data from U.S. EPA (2003), Kaiser et al. (2006) and Liu and Lee (2007) were selected to predict the solubility of all PFASs in this study. The solubility is affected by the intermolecular forces between solute and solvent, as well as the melting point and enthalpy of the solute. Subcooled liquid solubility is a function of both molecular size and melting point, while solid solubility is a function of only molecular size (Yue and Li, 2013). Therefore, to overcome the difference between congeners due to the melting-point effect, aqueous solubility for solid substances have to be converted into the subcooled liquid using the fugacity ratio (F).

3.7. Endpoint data quality and variability:

The solubility data reported in the references were determined by the method incorporated as a standard by the Organization for Eco nomic Co-operation and Development. Calculated properties of PFASs were obtained from U.S. EPA's, EPI Suite.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSPR

Simple linear regression model

4.2. Explicit algorithm:

log S_L= - 0.0221 x TSA + 4.4391

4.3. Descriptors in the model:

TSA-total surface area [Å2]. The total surface area is the sum of the area of all the surfaces of the solid.

4.4. Descriptor selection:

Water solubility was correlated to F_N, V_M and TSA. Water solubility is calculated from single descriptor model selected by statistical results and checked for internal consistency. The best model was selected based on statistical parameters.

4.5. Algorithm and descriptor generation:

Total surface area is a geometric descriptor derived from 3-D structure.

4.6. Software name and version for descriptor generation:

Chem Axon

https://chemaxon.com/tag/cheminformatics?gclid=CjwKCAiA1aiMBhAUEiwACw25MWrDbB7WFmcLjf4Nrc-Qi03n_RiDVZ9q1BRj9ZuqbZG4B8RhSrJLhxoCJSkQAvD_BwE

4.7. Chemicals/ Descriptors ratio:

7 chemicals/1 descriptor

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

N/A

5.2. Method used to assess the applicability domain:

N/A

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

N/A

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

N/A

6.6. Pre-processing of data before modelling:

Data was converted to a log scale for modeling.

6.7. Statistics for goodness-of-fit:

$r^2 = 0.7313$

F =13.607

S.E.= 1.41

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

Cross-validated coefficients using the leave-one-out technique (LOO-Rcv^2) were calculated from the equation: LOO-Rcv^2=1 - PRESS/TSS, where PRESS is the prediction error sum of squares and TSS is the total sum of squares.

Q^2_{LOO}=0.931

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN:No

Chemical Name: No

Smiles:No

Formula:No

INChI:No

MOL file:No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

N/A

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

N/A

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

The model was developed by statistical approach. No mechanistic was defined a priori.

8.2. A priori or a posteriori mechanistic interpretation:

The water solubility parameter, log S_L, is correlated with F_N, V_M and TSA of PFASs. In the statistical results, the P values of three single descriptor models were significant at the <0.05 level of significance for their respective regression models. However, the model based on TSA was statistically better than those based on F_N and V_M.

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. U.S. Environmental Protection Agency, 2003. Preliminary Risk Assessment of the Developmental Toxicity Associated with the Exposure to Perfluorooctanoic Acid and its Salts

2. Kaiser, M.A., Barton, C.A., Botelho, M., Buck, R.C., Buxton, L.W., Gannon, J., Kao, C.P.C., Larsen, B.S., Russell, M.H., Wang, N., Waterland, R.L., 2006. Understanding the transport of anthropogenic fluorinated compounds in the environment. Organohalogen Compd. 68, 675e678.

3. Liu, J., Lee, L.S., 2007. Effect of fluorotelomer alcohol chain length on aqueous solubility and sorption by soils. Environ. Sci. Technol. 41, 5357e5362.

4. Yue, C., Li, L.Y., 2013. Filling the gap: estimating physicochemical properties of the full array of polybrominated diphenyl ethers (PBDEs). Environ. Pollut. 180, 312e323.

9.3. Supporting information:

Supplementary data can be found at: *http://dx.doi.org/10.1016/j.envpol.2014.11.008.*

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

Kow. Predictive model for physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors

1. QSAR identifier

1.1. QSAR identifier (title):

Predictive model for physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors.

1.2. Other related models:

N/A

1.3. Software coding the model:

Linear regression was performed using **Sigmaplot 10 (Systat Software Inc, Point Rich)**

2. General information

2.1. Date of QMRF:

09/11/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com

https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Minhee Kim Department of Civil Engineering, University of British Columbia, 6250 Applied Science Lane, Vancouver, BC, Canada V6T 1Z4

2. Loretta Y. Li Department of Civil Engineering, University of British Columbia, 6250 Applied Science Lane, Vancouver, BC, Canada V6T 1Z4

3. John R. Grace Department of Chemical and Biological Engineering, University of British Columbia, 2360 East Mall, Vancouver, BC, Canada V6T 1Z3

4. Chaoyang Yue Department of Chemical and Biological Engineering, University of British Columbia, 2360 East Mall, Vancouver, BC, Canada V6T 1Z3

2.6. Date of model development and/or publication:

3 November 2014

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

Minhee Kim, Loretta Y. Li, John R. Grace, Chaoyang Yue, Selecting reliable physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors, Environmental Pollution, Volume 196, 2015, Pages 462-472, ISSN 0269-7491. https://doi.org/10.1016/j.envpol.2014.11.008.

2.8. Availability of information about the model:

Molecular structures, correlation equations and statistical results of 23 PFASs encompassing 10 PFCAs, 3 PFSAs, 6 FASAs and 4 FTOHs are represented in Supplementary Information.

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Octanol-water partition coefficient (K_{OW})

3.3. Comment on the endpoint:

The octanol/water partition coefficient is the ratio of the concentrations of organic compounds in n-octanol and water phases at equilibrium. The octanol/water partition coefficient is the best known environmental parameter used to characterize the lipophilicity of compounds to evaluate the tendency of compounds to partition between an organic phase and an aqueous phase.

3.4. Endpoint units:

-

3.5. Dependent variable:

log K_{OW}

3.6. Experimental protocol:

PFASs tend to form three immiscible layers when they are added to an octanol/water mixture because PFASs are simultaneously hydrophobic and lipophobic. Therefore, it is impossible to determine directly their K_{OW} values using ‘regular’ methods that are common for organic chemicals. In this study, the selected experimental octanol/water partition coefficient data (Ext-Sel. log K_{OW}) were obtained from Jing et al. (2009). Jing et al. (2009) determined K_{OW} of a homologous series of perfluoroalkyl and alkyl carboxylates using ion-transfer cyclic voltammetry. Jing et al. (2009) reported the first experimental K_{OW} for several representative straight-chain PFCA and PFSA congeners. This data provided the opportunity to assess the predictive abilities of various software programs (e.g., SPARC and ALOGPS 2.1) using the SMILES molecular formula language to predict K_{OW} of PFASs.

3.7. Endpoint data quality and variability:

Est'd-Sel. properties of PFASs were obtained from U.S. EPA's, EPI Suite.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSPR

Simple linear regression model

4.2. Explicit algorithm:

log $K_{OW} = 0.3047 \text{ x } F_N - 2.7258$

4.3. Descriptors in the model:

Fluorine number (F_N)

4.4. Descriptor selection:

Octanol-water partition coefficient (K_{OW}) was correlated to F_N, V_M and TSA. K_{OW} is calculated from single descriptor model selected by statistical results and checked for internal consistency. The best model was selected based on statistical parameters.

4.5. Algorithm and descriptor generation:

F_N is a constitutional descriptor derived from atomic composition.

4.6. Software name and version for descriptor generation:

Chem Axon

https://chemaxon.com/tag/cheminformatics?gclid=CjwKCAiA1aiMBhAUEiwACw25MWrDbB7WFmcLjf4Nrc-Qi03n_RiDVZ9q1BRj9ZuqbZG4B8RhSrJLhxoCJSkQAvD_BwE

4.7. Chemicals/ Descriptors ratio:

9 chemicals/1 descriptor

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

N/A

5.2. Method used to assess the applicability domain:

N/A

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

N/A

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

N/A

6.6. Pre-processing of data before modelling:

Data was converted to a log scale for modeling.

6.7. Statistics for goodness-of-fit:

$r^2 = 0.9876$

$F = 557.76$

S.E.= 0.1548

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

Cross-validated coefficients using the leave-one-out technique (LOO-Rcv^2) were calculated from the equation: LOO-Rcv^2=1 - PRESS/TSS, where PRESS is the prediction error sum of squares and TSS is the total sum of squares.

Q^2_{LOO}= 0.980

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN:No

Chemical Name: No

Smiles:No

Formula:No

INChI:No

MOL file:No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

N/A

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

N/A

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

The model was developed by statistical approach. No mechanistic was defined a priori.

8.2. A priori or a posteriori mechanistic interpretation:

Based on the experimental data collected by Jing et al. (2009) equations were obtained for three molecular descriptors: F_N, V_M and TSA of PFASs. In the statistical results, the P values of three single descriptor models were significant at the <0.05 level of significance for their respective regression models. However, the model based on F_N was statistically better than those based on TSA and V_M.

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Jing, P., Rodgers, P.J., Amemiya, S., 2009. High lipophilicity of perfluoroalkyl carboxylate and sulfonate: implications for their membrane permeability. J. Am. Chem. Soc. 131, 2290e2296.

2. Arp, H.P.H., Niederer, C., Goss, K.-U., 2006. Predicting the partitioning behavior of various highly fluorinated compounds. Environ. Sci. Technol. 40, 7298e7304.

9.3. Supporting information:

Supplementary data can be found at: *http://dx.doi.org/10.1016/j.envpol.2014.11.008.*

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

Kaw. Predictive model for physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors.

1. QSAR identifier

1.1. QSAR identifier (title):

Predictive model for physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors.

1.2. Other related models:

N/A

1.3. Software coding the model:

Linear regression was performed using **Sigmaplot 10 (Systat Software Inc, Point Rich)**

2. General information

2.1. Date of QMRF:

09/11/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com

https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Minhee Kim Department of Civil Engineering, University of British Columbia, 6250 Applied Science Lane, Vancouver, BC, Canada V6T 1Z4

2. Loretta Y. Li Department of Civil Engineering, University of British Columbia, 6250 Applied Science Lane, Vancouver, BC, Canada V6T 1Z4

3. John R. Grace Department of Chemical and Biological Engineering, University of British Columbia, 2360 East Mall, Vancouver, BC, Canada V6T 1Z3

4. Chaoyang Yue Department of Chemical and Biological Engineering, University of British Columbia, 2360 East Mall, Vancouver, BC, Canada V6T 1Z3

2.6. Date of model development and/or publication:

3 November 2014

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

Minhee Kim, Loretta Y. Li, John R. Grace, Chaoyang Yue, Selecting reliable physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors, Environmental Pollution, Volume 196, 2015, Pages 462-472, ISSN 0269-7491. https://doi.org/10.1016/j.envpol.2014.11.008.

2.8. Availability of information about the model:

Molecular structures, correlation equations and statistical results of 23 PFASs encompassing 10 PFCAs, 3 PFSAs, 6 FASAs and 4 FTOHs are represented in Supplementary Information.

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Air- water partition coefficient (K_{AW})

3.3. Comment on the endpoint:

The air/water partition coefficient (K_{AW}) is the concentration ratio of a substance in equilibrium between air and water phases. Can be expressed either in dimensionless form or with the same units as the Henry's law constant (H).

3.4. Endpoint units:

-

3.5. Dependent variable:

log K_{AW}

3.6. Experimental protocol:

The air/water partition coefficient is useful for predicting vapor exchange rates between air and water interfaces (Burkhard et al., 1985). Lei et al. (2004) and Goss et al. (2006) measured K_{AW} for the FTOHs (4:2 FTOH, 6:2 FTOH and 8:2 FTOH) using static head space. However, Lei et al. (2004) was unable to account fully for the intermolecular interaction between target molecules and the water phase because of the limited number of experimental temperatures. Goss et al. (2006) found that the experimental values from Lei et al. (2004) contradicted thermodynamic theory, and the 8:2 FTOH result raised some questions. Therefore, because there is little information on different K_{AW} values for FTOHs and none for PFCAs, PFSAs and FASAs, Ext-Sel. log K_{AW} values were calculated from Ext-Sel. log PL and Ext-Sel. log S_L in this study.

3.7. Endpoint data quality and variability:

Est'd-Sel. properties of PFASs were obtained from U.S. EPA's, EPI Suite.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSPR

Simple linear regression model

4.2. Explicit algorithm:

log K_{AW} = 0.2804 x F_N - 1.5251

4.3. Descriptors in the model:

Fluorine number (F_N)

4.4. Descriptor selection:

Air- water partition coefficient (Henrys law constant, H) was correlated to FN, VM and TSA. This property was calculated from single descriptor models selected by statistical results and checked for internal consistency.

4.5. Algorithm and descriptor generation:

F_N is a constitutional descriptor derived from atomic composition.

4.6. Software name and version for descriptor generation:

Chem Axon

https://chemaxon.com/tag/cheminformatics?gclid=CjwKCAiA1aiMBhAUEiwACw25MWrDbB7WFmcLjf4Nrc-Qi03n_RiDVZ9q1BRj9ZuqbZG4B8RhSrJLhxoCJSkQAvD_BwE

4.7. Chemicals/ Descriptors ratio:

7 chemicals/1 descriptor

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

N/A

5.2. Method used to assess the applicability domain:

N/A

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

N/A

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN:Yes

Chemical Name:Yes

Smiles:Yes

Formula:No

INChI:No

MOL file:No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

N/A

6.6. Pre-processing of data before modelling:

Data was converted to a log scale for modeling.

6.7. Statistics for goodness-of-fit:

$r^2 = 0.8692$

F = 33.23

S.E.= 0.6284

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

Cross-validated coefficients using the leave-one-out technique (LOO-Rcv^2) were calculated from the equation: LOO-Rcv^2=1 - PRESS/TSS, where PRESS is the prediction error sum of squares and TSS is the total sum of squares.

Q^2_{LOO}= 0.8295

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN:No

Chemical Name: No

Smiles:No

Formula:No

INChI:No

MOL file:No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

N/A

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

N/A

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

The model was developed by statistical approach. No mechanistic was defined a priori.

8.2. A priori or a posteriori mechanistic interpretation:

Endpoint equations were obtained for three molecular descriptors: F_N, V_M and TSA of PFASs. In the statistical results, the P values of three single descriptor models were significant at the <0.05 level of significance for their respective regression models. However, the model based on F_N was statistically better than those based on TSA and V_M.

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Ding, G., Peijnenburg, W.J.G.M., 2013. Physicochemical properties and aquatic toxicity of poly and perfluorinated compounds. Environ. Sci. Technol. 43, 598e678.

2. Shiu,W.Y., Ma, K.C., 2000. Temperature dependence of physical-chemical properties of selected chemicals of environmental interest. II. chlorobenzenes, polychlorinated biphenyls, polychlorinated dibenzo-p-dioxins and dibenzofurans. J. Phys. Chem. Ref. Data 29, 1e76.

3. Lei, Y.D., Wania, F., Mathers, D., Mabury, S.A., 2004. Determination of vapor pressures, octanol/air, and water/air partition coefficients for polyfluorinated sulfonamide, sulfonamidoethanols, and telomer alcohols. J. Chem. Eng. Data 49, 1013e1022.

4. Goss, K.-U., Bronner, G., Harner, T., Hertel, M., Schmidt, T.C., 2006. The partition behavior of fluorotelomer alcohols and olefins. Environ. Sci. Technol. 40, 3572e3577.

9.3. Supporting information:

Supplementary data can be found at: *http://dx.doi.org/10.1016/j.envpol.2014.11.008.*

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

Koa. Predictive model for physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors.

1. QSAR identifier

1.1. QSAR identifier (title):

Predictive model for physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors.

1.2. Other related models:

N/A

1.3. Software coding the model:

Linear regression was performed using **Sigmaplot 10 (Systat Software Inc, Point Rich)**

2. General information

2.1. Date of QMRF:

09/11/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com

https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Minhee Kim Department of Civil Engineering, University of British Columbia, 6250 Applied Science Lane, Vancouver, BC, Canada V6T 1Z4

2. Loretta Y. Li Department of Civil Engineering, University of British Columbia, 6250 Applied Science Lane, Vancouver, BC, Canada V6T 1Z4

3. John R. Grace Department of Chemical and Biological Engineering, University of British Columbia, 2360 East Mall, Vancouver, BC, Canada V6T 1Z3

4. Chaoyang Yue Department of Chemical and Biological Engineering, University of British Columbia, 2360 East Mall, Vancouver, BC, Canada V6T 1Z3

2.6. Date of model development and/or publication:

3 November 2014

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

Minhee Kim, Loretta Y. Li, John R. Grace, Chaoyang Yue, Selecting reliable physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors, Environmental Pollution, Volume 196, 2015, Pages 462-472, ISSN 0269-7491. https://doi.org/10.1016/j.envpol.2014.11.008.

2.8. Availability of information about the model:

Molecular structures, correlation equations and statistical results of 23 PFASs encompassing 10 PFCAs, 3 PFSAs, 6 FASAs and 4 FTOHs are represented in Supplementary Information.

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Octanol-air partition coefficient (K_{OA})

3.3. Comment on the endpoint:

The octanol/air partition coefficient is a key property when assessing partitioning between the atmosphere and organic phases, such as for organic films on aerosols, organic carbon in soil, the waxy cuticle and lipid portions of vegetation. However, application of KOA is based on the assumption that interactions between organic chemicals and environmental organic phases closely resemble the interaction between the chemical and octanol.

3.4. Endpoint units:

-

3.5. Dependent variable:

log K_{OA}

3.6. Experimental protocol:

Experimental log K_{OA} (Ext-Sel. log K_{OA}) data for PFASs were selected from Shoeib et al. (2004), Goss et al. (2006) and Dreyer et al. (2009). Shoeib et al. (2004) and Dreyer et al. (2009) measured K_{OA} for perfluoroalkyl sulfonamides (FASAs) and per fluoroalkyl sulfonamidoethanols (FASEs) using a modified generator column. Goss et al. (2006) determined K_{OA} of FTOHs based on the method described by Shoeib et al. (2004). These results were supported by data reported by Thuens et al. (2008).

3.7. Endpoint data quality and variability:

Est'd-Sel. properties of PFASs were obtained from U.S. EPA's, EPI Suite.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSPR

Simple linear regression model

4.2. Explicit algorithm:

log K_{OA} = 0.0129 x V_M +2.4847

4.3. Descriptors in the model:

Molar volume (V_M) is the volume occupied by one mole of a chemical element or a chemical compound. cm^3 mol^{-1}. Geometric descriptor derived from 3-D structure.

4.4. Descriptor selection:

Octanol-air partition coefficient (K_{OA}) was correlated to F_N, V_M and TSA. This property was calculated from single descriptor models selected by statistical results and checked for internal consistency.

4.5. Algorithm and descriptor generation:

V_M is geometric descriptor derived from 3-D structure.

4.6. Software name and version for descriptor generation:

ACD/Labs' ACD/PhysChem Suite

https://www.acdlabs.com/solutions/academia/?gclid=CjwKCAiA1aiMBhAUEiwACw25MUw_ZleJmBKmPqJ5ALizlwu6tzK5G8Sk8LycZpNlbT2UgRwAp20uthoCE9EQAvD_BwE

4.7. Chemicals/ Descriptors ratio:

8 chemicals/1 descriptor

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

N/A

5.2. Method used to assess the applicability domain:

N/A

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

N/A

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN:Yes

Chemical Name:Yes

Smiles:Yes

Formula:No

INChI:No

MOL file:No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

N/A

6.6. Pre-processing of data before modelling:

Data was converted to a log scale for modeling.

6.7. Statistics for goodness-of-fit:

$r^2 = 0.8989$

F = 53.36

S.E.= 0.3068

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

Cross-validated coefficients using the leave-one-out technique (LOO-Rcv^2) were calculated from the equation: LOO-Rcv^2=1 - PRESS/TSS, where PRESS is the prediction error sum of squares and TSS is the total sum of squares.

Q^2_{LOO}= 0.8716

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN:No

Chemical Name: No

Smiles:No

Formula:No

INChI:No

MOL file:No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

N/A

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

N/A

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

The model was developed by statistical approach. No mechanistic was defined a priori.

8.2. A priori or a posteriori mechanistic interpretation:

Endpoint equations were obtained for three molecular descriptors: F_N, V_M and TSA of PFASs. In the statistical results, the P values of three single descriptor models were significant at the <0.05 level of significance for their respective regression models. However, the model based on V_M was statistically better than those based on TSA and F_N.

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Goss, K.-U., Bronner, G., Harner, T., Hertel, M., Schmidt, T.C., 2006. The partition behavior of fluorotelomer alcohols and olefins. Environ. Sci. Technol. 40, 3572e3577.

2. Shoeib, M., Harner, T., Ikonomou, M., Kannan, K., 2004. Indoor and outdoor air concentrations and phase partitioning of perfluoroalkyl sulfonamides and polybrominated diphenyl ethers. Environ. Sci. Technol. 38, 1313e1320.

3. Dreyer, A., Langer, V., Ebinghaus, R., 2009. Determination of octanol/air partition coefficients (KOA) of fluorotelomer acrylates, perfluoroalkyl sulfonamides, and perfluoroalkylsulfonamido ethanols. J. Chem. Eng. Data 54, 3022e3025.

4. Thuens, S., Dreyer, A., Sturm, R., Temme, C., Ebinghaus, R., 2008. Determination of the octanol/air partition coefficient (KOA) of fluorotelomer alcohols. J. Chem. Eng. Data 53, 223e227.

9.3. Supporting information:

Supplementary data can be found at: *http://dx.doi.org/10.1016/j.envpol.2014.11.008.*

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

Ki. QSPR model to predict interfacial adsorption coefficients.

1. QSAR identifier

1.1. QSAR identifier (title):

QSPR model to predict interfacial adsorption coefficients.

1.2. Other related models:

N/A

1.3. Software coding the model:

Statistical analyses were conducted using Microsoft excel.

2. General information

2.1. Date of QMRF:

09/11/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com

https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

Mark L Brusseau Soil, Water, and Environmental Science Department, School of Earth and Environmental Sciences, University of Arizona Tucson, AZ 85721 Hydrology and Atmospheric Sciences Department, School of Earth and Environmental Sciences, University of Arizona Tucson, AZ 85721

2.6. Date of model development and/or publication:

11 January 2019

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

Mark L. Brusseau, The influence of molecular structure on the adsorption of PFAS to fluid-fluid interfaces: Using QSPR to predict interfacial adsorption coefficients, Water Research, Volume 152, 2019, Pages 148-158, ISSN 0043-1354. https://doi.org/10.1016/j.watres.2018.12.057

2.8. Availability of information about the model:

Surface-tension and interfacial-tension data sets collected from the literature were used to determine interfacial adsorption coefficients for 42 individual PFAS. The PFAS evaluated comprise homologous series of perfluorocarboxylates and perfluorosulfonates, branched perfluoroalkyls, polyfluoroalkyls, alcohol PFAS, and nonionic PFAS.

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Air-water interfacial adsorption coefficient (Ki)

3.3. Comment on the endpoint:

The Ki values vary across eight orders of magnitude and are a function of molecular structure. Given the surface-active nature of PFAS, it is likely that air-water interfacial adsorption is an important factor in the atmospheric transport of PFAS.

3.4. Endpoint units:

cm

3.5. Dependent variable:

Log Ki

3.6. Experimental protocol:

Different methods were used in the studies to measure surface/interfacial tension, including Du Nouy ring, Wilhelmy plate, pendant drop, and drop-weight methods. Therefore, the data sets were collated to generate the largest data set for one consistent set of conditions. The great majority of the reported surface-tension measurements were conducted using deionized water as the aqueous phase. Also, salt forms of PFAS rather than acid forms comprised the majority of the data sets. Measured Vm values are readily available for many common hydrocarbon surfactants.

3.7. Endpoint data quality and variability:

Approximately 65 journal articles were retrieved that reported measured surface-tension or interfacial-tension data sets for PFAS. Some studies reported data for multiple compounds or conditions. Hence, a total of ~160 individual data sets for PFAS were obtained from the literature search. In keeping with the theme of simplicity, Vms for this analysis were calculated as MW/rc (e.g., Kaliszan, 1987; Baum, 1998; Reinhard and Drefahl, 1999). Densities were obtained from ChemSpider, PubChem, and product manufacturer's websites.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSPR

Simple linear regression model

4.2. Explicit algorithm:

log Ki = 0.021 (±0.002) x Vm - 8.56 (±0.42)

4.3. Descriptors in the model:

The four primary descriptors to be tested are MW, V_M, CN, and number of fluorine atoms (FN). However, it is observed that molar volume is the only descriptor that provides very good representation of the training and validation data set.

4.4. Descriptor selection:

A wide variety of molecular descriptors are available for QSPR analysis, ranging from simple size-based descriptors such as molecular weight (MW) and molar volume (V_M), simple constitutional descriptors based on numbers of a specific type of atom or bond (e.g., carbon number, CN), descriptors characterizing molecular structure (such as the molecular connectivity index, X), to complex 3-D geometrical descriptors (e.g., Todeschini and Consonni, 2009). V_M based model has the best statistical parameters.

4.5. Algorithm and descriptor generation :

Molar volumes (cm^3/mol) can be determined in three ways, measured by experiment, calculated using molecular-analysis approaches, and calculated as the quotient of MW and compound density (rc). Measured V_M values are readily available for many common hydrocarbon surfactants. Conversely, there are minimal reported values for PFAS. In keeping with the theme of simplicity, V_M values for this analysis were calculated as MW/rc (e.g., Kaliszan, 1987; Baum, 1998; Reinhard and Drefahl, 1999).

4.6. Software name and version for descriptor generation:

V_M values for several PFA were calculated using molecular analysis software from ACD/Labs (employing an additive atomic increment approach).

4.7. Chemicals/ Descriptors ratio:

The ratio of number of training set compounds to the number of descriptors is 15:1, much greater than the recommended minimum of 5:1 (Dearden et al., 2009).

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The results demonstrate that the molar-volume based QSPR model can provide robust predictions of air-water interfacial adsorption coefficients for a wide variety of PFAS with greatly varying structures.

5.2. Method used to assess the applicability domain:

N/A

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

Significant prediction errors are observed for 3 of the 54 data, specifically FTOH, FC8diol, and UDFOS, all with MSEs >1. The large errors observed for FTOH and FC8diol may be related to the differing behavior of alcohols, wherein the OH group confers significant hydrogen-bonding potential, and the resultant impact on interactions in solution and at the interface. There is no readily discernible explanation for the greater error observed for UDFOS, as the other polyfluoroalkyl data are well predicted.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN: No

Chemical Name: Yes

Smiles: No

Formula: Yes

INChI: No

MOL file:No

6.3. Data for each descriptor variable for the training set:

N/A

6.4. Data for the dependent variable (response) for the training set:

Yes

6.5. Other information about the training set:

The training set consists of the homologous series of PFCAs and PFSAs, which represent the "standard" linear anionic PFAS. The PFCAs and PFSAs comprise 10 and 5 compounds, respectively, for a total of 15 data points for the training set. Compounds comprising the training set represent the simplest molecular structures among the data set.

6.6. Pre-processing of data before modelling:

logarithmic transformation

6.7. Statistics for goodness-of-fit:

Full model (54 compounds):

$r^2 = 0.94$

MSE = 0.17.

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

N/A

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: Yes

Smiles: No

Formula: Yes

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

Yes

7.5. Other information about the external validation set:

39 data points for the validation set. The validation set comprises much greater molecular-structure complexity, including 2 branched PFAS, 15 poly-PFAS (of which three are cationic rather than anionic, and another three are branched), 8 nonionic PFAS, and 2 alcohols

7.6. Experimental design of test set:

39 data points for the validation set, which is more than double the number of data comprising the training set.

7.7. Predictivity – Statistics obtained by external validation:

N/A

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

The model was developed by statistical approach.

8.2. A priori or a posteriori mechanistic interpretation:

Prior research has shown that the surface activity of these PFAS is a function of chain length (e.g., Hendricks, 1953; Shinoda et al., 1972; Tamaki et al., 1989; Kissa, 2001; Lunkenheimer et al., 2015). The strong dependence of surface activity on molecular structure, in this case chain length (CNt), indicates high probability for developing effective QSPR models for predicting fluid-fluid interfacial adsorption coefficients

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Dearden, J.C., Cronin, M.T.D., Kaiser, K.L.E., 2009. How not to develop a quantitative

structureeactivity or structureeproperty relationship (QSAR/QSPR). SAR QSAR Environ. Res. 20, 241e266.

2. Kim, M., Li, L.Y., Grace, J.R., Yue, C., 2015. Selecting reliable physicochemical properties of perfluoroalkyl and polyfluoroalkyl substances (PFASs) based on molecular descriptors. Environ. Pollut. 196, 462e472.

3. Lyu, Y., Brusseau, M.L., Chen, W., Yan, N., Fu, X., Lin, X., 2018. Adsorption of PFOA at the airwater interface during transport in unsaturated porous media. Environ. Sci. Technol. 52, 7745e7753.

[4]Kissa, E., 2001. Fluorinated Surfactants and Repellents, second ed. Marcel Dekker, Inc, New York, NY.

5. Goodwin, J.W., 2004. Colloids and Interfaces with Surfactants and Polymers e an Introduction. John Wiley and Sons, Ltd, West Sussex, England.

6. Saripalli, K.P., Kim, H., Rao, P.S.C., Annable, M.D., 1997. Measurement of specific fluidefluid interfacial areas of immiscible fluids in porous media. Environ. Sci. Technol. 31 (3), 932e936.

7. Cho, J., Annable, M.D., 2005. Characterization of pore scale NAPL morphology in homogeneous sands as a function of grain size and NAPL dissolution. Chemosphere 61, 899e908.

9.3. Supporting information:

N/A

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments

MP1. CADASTER QSPR Models for Predictions of Melting Point of Perfluorinated Chemicals (University of Insubria).

1. QSAR identifier

1.1. QSAR identifier (title):

CADASTER QSPR Models for Predictions of Melting Point of Perfluorinated Chemicals (University of Insubria)

1.2. Other related models:

CADASTER QSPR Models for Predictions of Melting Point of Perfluorinated Chemicals (developed by LNU, IDEA and HMGU)

1.3. Software coding the model:

N/A

2. General information

2.1. Date of QMRF:

6/12/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.
Trzy Lipy 3 Street, Building B,
80-172 Gdansk, Poland
+48 795 160 760
contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

B. Bhhatarai, S. Kovarich, E. Papa, P. Gramatica, QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria (UI)

paola.gramatica@uninsubria.it http://www.qsar.it/

2.6. Date of model development and/or publication:

Published in 2011

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

1. B. Bhhatarai, W. Teetz, T. Liu, T. Öberg, N. Jeliazkova, N. Kochev, O. Pukalov, I. V. Tetko, S. Kovarich, E. Papa, P. Gramatica. (2011). CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204. https://doi.org/10.1002/minf.201000133

2. DRAGON https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_ molecular_descriptor_calculations

2.8. Availability of information about the model:

For MP 94 compounds containing mainly long chain alkylates, saturated cyclic and few aromatic PFCs were split into training and prediction set close to 50% and 15 more compounds from PERFORCE were used as EV- set.

More information about model attached in supplementary materials: doi:10.1002/minf.201000133

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Physical Chemical Properties: Melting Point

3.3. Comment on the endpoint:

Melting Point (MP) is an important endpoint as it influence the transport, solubility, distribution and environmental fate.

3.4. Endpoint units:

°C

3.5. Dependent variable:

MP

3.6. Experimental protocol:

N/A

3.7. Endpoint data quality and variability:

- 55 data points of physiochemical property were collected mainly from SRC-PhysProp
- 19 data points from Trepka et al. [2]
- 7 data points from Gajewski et al. [3]

- 13 data points from Platonov et al. [4]

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

Multiple linear regression model (OLS - Ordinary Least Square)

4.2. Explicit algorithm:

MP = 132.95(±22.63) AAC + 3.04(±0.97) F02[C-F] - 18.97(±7.98)C-013 + 209.04(±139.9) RBF - 243.16

4.3. Descriptors in the model:

1. AAC - 'information indices' which characterizes mean information index on atomic composition
2. F02[C-F] - frequency of C-F at topological distance 2
3. C-013 - carbon connected to at least three electronegative atoms (X) as CRX3
4. RBF - rotable bond fraction

4.4. Descriptor selection:

Only 1D and 2D descriptors were selected for modeling MP, easy and simply calculated from the SMILES.

- highly correlated descriptor were excluded (more than 90% pairwise correlation)
- filtered descriptors were subject to variable selection method using genetic algorithm (GA)
- using selected descriptors, a multiple linear regression (MLR) model by using the ordinary-least-squares (OLS) techniques were build

4.5. Algorithm and descriptor generation:

The input files for descriptor calculation were obtain by the semi empirical AM1 method (minimized their lowest energy conformation) using HYPERCHEM software. Molecular descriptors were generated using DRAGON software.

4.6. Software name and version for descriptor generation:

1. DRAGON sofware

A software to calculate theoretical molecular descriptors https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_ molecular_descriptor_calculations

2. HYPERCHEM

Software used for molecular drawing and conformational energy optimalization AM1 www. hyper.com

4.7. Chemicals/ Descriptors ratio:

SOM splitting: 53 chemicals / 4 descriptors = 13.25

Response splitting: 48 chemicals / 4 descriptors = 12

Full model: 94 chemicals / 4 descriptors = 23.5

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The Williams plot was used to verify the presence of response outliers (i.e. compounds with cross-validated standardized residuals greater than three standard deviation units, ±3σ) and chemicals very influential for their structure in determining model parameters (i.e. compounds with high leverage value (h) (h>h*, the critical value being h*=3p'/n, where p' is the number of model variables plus one, and n the number of the objects used to calculate the model). The hat (high leverage h) value is different for each model. The leverage approach was applied for the definition of the structural chemical domain model for new chemicals without experimental data.

5.2. Method used to assess the applicability domain:

Leverage (Williams) and standardization approach

5.3. Software name and version for applicability domain assessment:

OCHEM (on-line CHEMical database and Modeling environment)

https://www.ochem.eu/home/show.do

5.4. Limits of applicability:

Full model: The leverage approach based AD study has a hat cut-off values for the MP MLR model which is 0.16. Based on this cut-off, there were four training compounds (CAS 306-91-2, CAS 311-89-7, CAS 338-83-0 and CAS 359-70-6) which were outside the structural applicability domain of MP model that why they were highly influential in selecting modeling variables. In addition, there were 16 other new compounds which were out of the structural domain (high leverage or high hat values). They were bulky polyfluorinated chemicals, mostly linear PFCs with 13 or 15 carbon atom long including 2 acrylates and nitrogen centered PFCs. These long chain PFCs were probably extrapolated as the longest compound in the training model were with 7 carbon atoms only. Altogether only 20/397 compounds are found to be out of AD, thus the model has 94.9% coverage.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability 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:

N/A

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

94 compounds were split into training and prediction set using to methods: by SOM and Response Splitting on the data. The two training sets - of 53 chemicals (SOM) and 48 chemicals (Response) were used to develop a population of models. Additional full model.

6.6. Pre-processing of data before modelling:

N/A

6.7. Statistics for goodness-of-fit:

a) Split by SOM: R^2=0.83, $RMSE_{TR}$=33.95

b) Random response activity: R^2=0.84, $RMSE_{TR}$=37.16

c) Full model: R^2= 0.81, $RMSE_{TR}$=40.24

d) Full model (same data excluding compound CAS 426-65-3): R^2= 0.82, $RMSE_{TR}$=39.39

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

a) Split by SOM: Q^2_{LOO} = 0.79

b) Random response activity: Q^2_{LOO} = 0.79

c) Full model: Q^2_{LOO} = 0.78

d) Full model (same data excluding compound CAS 426-65-3): Q^2_{LOO} = 0.79

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

a) Split by SOM: R^2Y_{SCR} = 0.07

b) Random response activity: R^2Y_{SCR} = 0.08

c) Full model: R^2Y_{SCR} = 0.04

6.11. Robustness – Statistics obtained by bootstrap:

a) Split by SOM: Q^2_{BOOT} = 0.78

b) Random response activity: Q^2_{BOOT} = 0.78

c) Full model: Q^2_{BOOT} = 0.77

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

94 compounds were split into training and prediction set using to methods: by SOM and Response Splitting on the data. The two training sets - of 53 chemicals (SOM) and 48 chemicals (Response) were used to develop a population of models. Additional full model. 15 compounds from PERFORCE were used for external validation as EV-set.

7.6. Experimental design of test set:

EV-set contains long chain perfluoroalkylated chemicals which are prevalent in the environment and are of higher interest in terms of physico-chemical properties and environmental distribution.

7.7. Predictivity – Statistics obtained by external validation:

a) Split by SOM:

Q^2_{F1} = 0.76, Q^2_{F3} = 0.62*, $RMSE_{EXT}$ = 51.69

EV-set: Q^2_{F1} = 0.72, Q^2_{F3} = 0.91, $RMSE_{EXT}$ = 24.92,

b) Random response activity:

$Q^2_{F1} = 0.73$, $Q^2_{F3} = 0.76$, $RMSE_{EXT} = 45.47$

EV-set: $Q^2_{F1} = 0.73$, $Q^2_{F3} = 0.87$, $RMSE_{EXT} = 32.08$

c) Full model:

EV-set: $Q^2_{F1} = 0.83$, $Q^2_{F3} = 0.91$, $RMSE_{EXT} = 27.19$

d) Full model (same data excluding compound CAS 426-65-3):

EV-set: $Q^2_{F1} = 0.81$, $Q^2_{F3} = 0.89$

$^*Q^2_{F3}$ was equal to 0.72 after removing 2 compounds (#28 CAS 354-32-5 and #14 CAS 309-91-2)

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

An external validation was performed for the developed model after splitting the compounds into a validation and training set. Despite access to data for all compounds used in the model, they were not added to the corresponding sets.

For this model, additional external validation was performed using compounds for which the authors had collected experimental data. The EV-set was chosen to help as an additional validation of models that have demonstrated their validity during earlier splits.

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

N/A

8.2. A priori or a posteriori mechanistic interpretation:

$$MP= 132.95(\pm 22.63)AAC + 3.04(\pm 0.97)F02[C\text{-}F] - 18.97(\pm 7.98)C\text{-}013 + 209.04(\pm 139.9)RBF - 243.16$$

where:

AAC – ‘information indices’ which characterizes mean information index on
atomic composition

F02[C-F] - Frequency of C-F at topological distance 2

C-013 - carbon connected to at least three electronegative atoms (X) CRX3

RBF - rotable bond fraction

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

Finally, a consensus model was developed for MP data by a simple average of the results predicted by all the models developed by project partners, such that each sub-model has equal weight. The correct experimental data were used for the calculation of statistical parameters. The consensus models were better than the individual submodels.

Statistics of consensus model:

$R^2 = 0.88$,

$RMSE_{TR} = 31.81$

9.2. Bibliography:

1. B. Bhhatarai, W. Teetz, T. Liu, T. Öberg, N. Jeliazkova, N. Kochev, O. Pukalov, I. V. Tetko, S. Kovarich, E. Papa, P. Gramatica. (2011). CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204. doi:10.1002/minf.201000133

2. R. D. Trepka, J. K. Harrington, J. W. McConville, K. T. McGurran, A. Mendel, D.R. Pauly, J.E. Robertson, J.T. Waddington, J. Agric. Food Chem. 1974, 22, 1111–1119. https://pubs.acs.org/doi/10.1021/jf60196a044

3. R. P. Gajewski, G. D. Thompson, E. H. Chio, J. Agric. Food Chem. 1988, 36, 174–177.

4.V. E. Platonov, A. Haas, M. Schelvis, M. Lieb, K. V. Dvornikova, O. I. Osina, Y. V. Gatilov, J. Fluorine Chem. 2001, 109, 131 – 139.

9.3. Supporting information:

More information about the model attached in supplementary materials:

doi:10.1002/minf.201000133

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

MP2. CADASTER QSPR Models for Predictions of Melting Point of Perfluorinated Chemicals (Linnaeus University)

1. QSAR identifier

1.1. QSAR identifier (title):

CADASTER QSPR Models for Predictions of Melting Point of Perfluorinated Chemicals (Linnaeus University)

1.2. Other related models:

CADASTER QSPR Models for Predictions of Melting Point of Perfluorinated Chemicals (developed by UI, IDEA and HMGU)

1.3. Software coding the model:

SIMCA-P, v.11.0 Umetrics AB

2. General information

2.1. Date of QMRF:

6/12/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.
Trzy Lipy 3 Street, Building B,
80-172 Gdansk, Poland
+48 795 160 760
contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

T. Liu, T. Öberg

School of Natural Sciences, Linnaeus University (LNU)SE-39182, Kalmar, Sweden

2.6. Date of model development and/or publication:

Published in 2011

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

1. Bhhatarai, B., Teetz, W., Liu, T., Öberg, T., Jeliazkova, N., Kochev, N., Pukalov, O., Tetko, I.V., Kovarich, S., Papa, E., Gramatica, P. (2011). CADASTER QSPR Models for Predictions

of Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204. https://doi.org/10.1002/minf.201000133

2. DRAGON https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_ molecular_descriptor_calculations

2.8. Availability of information about the model:

For MP 94 compounds containing mainly long chain alkylates, saturated cyclic and few aromatic PFCs were split into training and prediction set close to 50% and 15 more compounds from PERFORCE were used as EV- set.

More information about model attached in supplementary materials: doi:10.1002/minf.201000133

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Physical Chemical Properties: Melting Point

3.3. Comment on the endpoint:

Melting Point (MP) is an important endpoint as it influence the transport, solubility, distribution and environmental fate.

3.4. Endpoint units:

°C

3.5. Dependent variable:

MP

3.6. Experimental protocol:

N/A

3.7. Endpoint data quality and variability:

- 55 data points of physiochemical property were collected mainly from SRC-PhysProp
- 19 data points from Trepka et al. [2]
- 7 data points from Gajewski et al. [3]
- 13 data points from Platonov et al. [4]

4. Defining the algorithm – OECD Principle 2

4.1. Type of mod

Partial least squares regression (PLSR)

4.2. Explicit algorithm:

PLSR is based on a linear transformation of the theoretical molecular descriptors to a limited number of orthogonal factors, attempting to maximize the covariance between the descriptors and the response variable.

Normalised equation:

MP = 9.62 + 11xBELm3 + 7.61xBEHv5 + 7.56xO-057 + 6.53xnROH - 6.35xMe + 5.93xAMR + 5.77xnRCOOH + 5.68xBELe4 + 5.68xBELm4 + 5.33xpiPC08 - 5.27xF-083 + 5.21xX1sol - 5.14xC-013 + 5xX0sol + 4.43xC-040 + 4.02xUi + 3.35xB02[O-O] - 3.27xnCRX3 + 3.1xIAC - 2.51xnCp + 2.44xX3sol + 2.4xTIC1 + 2.38xnHDon + 2.11xTPSA(NO) + 1.95xX2sol + 1.84xnO + 1.62xX4sol + 1.18xTPSA(Tot) + 0.458xIC1 + 0.403xAAC - 0.255xISH - 0.181xBELm5 + 0.0739xX5sol - 0.0104xHy

4.3. Descriptors in the model:

Descriptors belonging to various types (e.g., constitutional, topological, walk and path counts, connectivity index, information index, 2D autocorrelations, edge adjacency indices, BCUT descriptors, topological charge index, eigenvalue-based index, Randic molecular profiles, geometrical, RDF descriptors, 3D-MoRSE descriptors, WHIM, GETAWAY, functional group counts, atom-centered fragments, charge, molecular properties, 2D Binary and 2D frequency fingerprints) were calculated by using Dragon. These molecular descriptors were used as input variables for modeling in order to capture maximum of the relevant structural features related to the response.

Descriptors: Me, IAC, AAC, IC0, TIC0, IC1, TIC1, ISH, nCp, Hy, AMR, nO, X0sol, X1sol, X2sol, BEHv5, nROOC, NROH, nCRX3, nHDon, C-013, C-040, H-050, O-057, F-083, Ui, TPSA(NO), TPSA(Tot), B02[O-O], X3sol, BELm3, BELm4, X4sol, X5sol, BELm5, BELe4, piPC08

4.4. Descriptor selection:

- PLSR was used for data analysis and modeling
- the data analysis and multivariate calibrations were carried out with the software SIMCA-P
- the 37 largest values of molecular descriptors at 3 components using VIP approach were selected

4.5. Algorithm and descriptor generation:

The input files for descriptor calculation were obtain by the semi empirical AM1 method (minimized their lowest energy conformation) using HYPERCHEM software. Molecular descriptors were generated using DRAGON software.

4.6. Software name and version for descriptor generation:

1. DRAGON software

A software to calculate theoretical molecular descriptors https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_ molecular_descriptor_calculations

2. HYPERCHEM

Software used for molecular drawing and conformational energy optimalization AM1

www.hyper.com

4.7. Chemicals / Descriptors ratio:

94 chemicals / 37 descriptors at 3 components

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The valid applicability domain for the PLSR model was assessed by the residual standard deviation (the Euclidean distance to the PLSR model) and the leverage (the Mahalanobis distance within the PLSR model space) to that of the calibration objects. From these distances the SIMCA software estimates the probability that a new compound belongs to the model (PModXPS and PModXPS +). Here we used a probability of belonging to the model at the 99% significant level (PModXPS < 0.01 and PModXPS + < 0.01) as the cut-off criterion to decide if a compound was outside of the model AD. The 1% clarification limit is based on the training set properties, meaning that using a higher cut-off would put several of the calibration compounds outside of the applicability domain, which does not seem appropriate.

5.2. Method used to assess the applicability domain:

Residual standard deviation (the Euclidean distance to the PLSR model) and the leverage (the Mahalanobis distance within the PLSR model space) approach.

5.3. Software name and version for applicability domain assessment:

SIMCA-P, v.11.0 Umetrics AB

5.4. Limits of applicability:

94 compounds were used with MP as a calibration set to study applicability domain, 37 descriptors with the highest values selected using VIP were used to domain study. 20 compounds with a probability of belonging of less than 1 % (PModXPS < 0.01) were

considered to be outliers, in which PModXPS represents the probability of the observation belonging to the model in the X space. After excluding outliers in the calibration model, 83.69% variance in Y is described, and 83.2% variance is predicted by the model at 1 component. Then 303 compounds without any MP values were used further to determine their domain, the distance to the model may be augmented with PModXPS+ measuring how far outside the acceptable model domain the projection of the observation falls. 145 compounds were found to be outside the domain using PModXPS + approach that combines PMdodXPS plus Hotelling's T^2. The strong outliers were determined by Hotelling's T^2 diagnosis at 99% confidence interval.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: No

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

N/A

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

94 compounds were split into training and prediction set using to methods: by SOM and Response Splitting on the data. The two training sets - of 53 chemicals (SOM) and 48 chemicals (Response) were used to develop a population of models. Additional full model.

6.6. Pre-processing of data before modelling:

All descriptor variables were preprocessed by auto-scaling to zero mean and unit variance.

6.7. Statistics for goodness-of-fit:

a) Split by SOM: R2=0.89, $RMSE_{TR}$=28.17

b) Random response activity: R^2=0.89, $RMSE_{TR}$=32.42

c) Full model: R^2=0.82, $RMSE_{TR}$=38.64

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

a) Split by SOM: Q^2_{LOO} =0.84

b) Random response activity: Q^2_{LOO} =0.81

c) Full model: Q^2_{LOO} =0.82

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

94 compounds were split into training and prediction set using to methods: by SOM and Response Splitting on the data. The two training sets - of 53 chemicals (SOM) and 48 chemicals (Response) were used to develop a population of models. Additional full model. 15 compounds from PERFORCE were used for external validation as EV-set.

7.6. Experimental design of test set:

EV-set contains long chain perfluoroalkylated chemicals which are prevalent in the environment and are of higher interest in terms of physico-chemical properties and environmental distribution.

7.7. Predictivity – Statistics obtained by external validation:

a) Split by SOM:

$RMSE_{EXT}$=42.05

EV-set: $RMSE_{EXT}$=16.47,

b) Random response activity:

$RMSE_{EXT}$= 37.82

EV-set: $RMSE_{EXT}$= 30.67

c) Full model:

EV-set: $RMSE_{EXT}$= 25.96

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

An external validation was performed for the developed model after splitting the compounds into a validation and training set. Despite access to data for all compounds used in the model, they were not added to the corresponding sets.

For this model, additional external validation was performed using compounds for which the authors had collected experimental data. The EV-set was chosen to help as an additional validation of models that have demonstrated their validity during earlier splits.

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

N/A

8.2. A priori or a posteriori mechanistic interpretation:

In the model, the descriptors in first component were dominated by the weak intermolecular interactions, and the descriptors in second component were mainly related to polar interactions contributed by the numbers of hydroxyl groups, the topological polar surface area, and the numbers of donor atoms for hydrogen bonding. The free energy transfer of MP corresponding to contributions from van der Waals and polar interactions was explained by these two components.

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

Finally, a consensus model was developed for MP data by a simple average of the results predicted by all the models developed by project partners, such that each sub-model has equal weight. The correct experimental data were used for the calculation of statistical parameters. The consensus models were better than the individual submodels.

Statistics of consensus model:

$R^2 = 0.88$,

$RMSE_{TR} = 31.81$

9.2. Bibliography:

1. B. Bhhatarai, W. Teetz, T. Liu, T. Öberg, N. Jeliazkova, N. Kochev, O. Pukalov, I. V. Tetko, S. Kovarich, E. Papa, P. Gramatica. (2011). CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204. doi:10.1002/minf.201000133

2. R. D. Trepka, J. K. Harrington, J. W. McConville, K. T. McGurran, A. Mendel, D.R. Pauly, J.E. Robertson, J.T. Waddington, J. Agric. Food Chem. 1974, 22, 1111–1119. https://pubs.acs.org/doi/10.1021/jf60196a044

3. R. P. Gajewski, G. D. Thompson, E. H. Chio, J. Agric. Food Chem. 1988, 36, 174–177.

4.V. E. Platonov, A. Haas, M. Schelvis, M. Lieb, K. V. Dvornikova, O. I. Osina, Y. V. Gatilov, J. Fluorine Chem. 2001, 109, 131 – 139.

9.3. Supporting information:

doi:10.1002/minf.201000133

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

MP3. CADASTER QSPR Models for Predictions of Melting Point of Perfluorinated Chemicals (Ideaconsult Ltd.)

1. QSAR identifier

1.1. QSAR identifier (title):

CADASTER QSPR Models for Predictions of Melting Point of Perfluorinated Chemicals (Ideaconsult Ltd.)

1.2. Other related models:

CADASTER QSPR Models for Predictions of Melting Point of Perfluorinated Chemicals (developed by UI, LNU and HMGU)

1.3. Software coding the model:

JBSMM software

N. Kochev, O. Pukalov , Software system for molecular modeling–JBSMM; Scientific Researches of the Union of Scientists in Bulgaria – Plovdiv, Series C: Technics and Technologies, Vol. V., Balkan Conference of Young Scientists; Plovdiv, 16–18 June, 2005, pp. 315 – 320.

2. General information

2.1. Date of QMRF:

6/12/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.
Trzy Lipy 3 Street, Building B,
80-172 Gdansk, Poland
+48 795 160 760
contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

N. Jeliazkova Ideaconsult Ltd. (IDEA) 4A. Kanchev str., Sofia 1000, Bulgaria

2.6. Date of model development and/or publication:

Published in 2011

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

1. Bhhatarai, B., Teetz, W., Liu, T., Öberg, T., Jeliazkova, N., Kochev, N., Pukalov, O., Tetko, I.V., Kovarich, S., Papa, E., Gramatica, P. (2011). CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204.

https://doi.org/10.1002/minf.201000133

2. JBSMM software

N. Kochev, O. Pukalov, Software system for molecular modeling–JBSMM; Scientific Researches of the Union of Scientists in Bulgaria – Plovdiv, Series C: Technics and Technologies, Vol. V., Balkan Conference of Young Scientists; Plovdiv, 16–18 June, 2005, pp. 315 – 320.

2.8. Availability of information about the model:

For MP 94 compounds containing mainly long chain alkylates, saturated cyclic and few aromatic PFCs were split into training and prediction set close to 50% and 15 more compounds from PERFORCE were used as EV- set.

More information about model attached in supplementary materials: doi:10.1002/minf.201000133

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Physical Chemical Properties: Melting Point

3.3. Comment on the endpoint:

Melting Point (MP) is an important endpoint as it influences the transport, solubility, distribution and environmental fate.

3.4. Endpoint units:

°C

3.5. Dependent variable:

MP

3.6. Experimental protocol:

N/A

3.7. Endpoint data quality and variability:

- 55 data points of physiochemical property were collected mainly from SRC-PhysProp
- 19 data points from Trepka et al. [2]
- 7 data points from Gajewski et al. [3]
- 13 data points from Platonov et al. [4]

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

Multiple linear regression model (OLS - Ordinary Least Square)

4.2. Explicit algorithm:

$\mathrm{MP} = 2.34(\pm 0.28)\mathrm{In}^2 W - 164.76(\pm 16.51)\mathrm{W}_{\mathrm{rel}}^{\mathrm{F}} + 33.14(\pm 8.90) N_{\mathrm{HBDON}} + 79.38$

4.3. Descriptors in the model:

1. $\mathrm{In}^2 W$ - general degree of branching through the square of the logarithm of Wiener index
2. $\mathrm{W}_{\mathrm{rel}}^{\mathrm{F}}$ - the number of fluorine atoms and their relative connectivity (this is the sum of all paths to fluorine atoms divided by the Wiener index W)
3. N_{HBDON} - denotes the number of H-bond donors in the molecule

4.4. Descriptor selection:

- descriptors were chosen manualy in a step-wise manner base on expert knowledge
- choice of descriptors combination was guided by three practicaltules: (1) each coeff in the MLRA must be statistically significant based on RSD (RSD<25%); (2) the choice of extra descriptors obeys a chemical logic in order to fix compounds with bad model values trying to handle the missing structural fragments; (3) successively selected descriptors must decrease RMSE value and increase Q^2 and R^2 values

4.5. Algorithm and descriptor generation:

Set of topological, constitutional and group based descriptors were calculated using JBSMM software. The descriptor combinations were chosen manually in a step-wise manner applying chemical expert knowledge.

4.6. Software name and version for descriptor generation:

1. JBSMM software

N. Kochev, O. Pukalov , Software system for molecular modeling–JBSMM; Scientific Researches of the Union of Scientists in Bulgaria – Plovdiv, Series C: Technics and Technologies, Vol. V., Balkan Conference of Young Scientists; Plovdiv, 16–18 June, 2005, pp. 315 – 320.

4.7. Chemicals/ Descriptors ratio:

SOM splitting: 53 chemicals / 3 descriptors = 17.67

Random response activity: 48 chemicals / 3 descriptors = 16

Full model: 94 chemicals / 3 descriptors = 31.33

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The Williams plot was used to verify the presence of response outliers (i.e. compounds with cross-validated standardized residuals greater than three standard deviation units, ±3σ) and chemicals very influential for their structure in determining model parameters (i.e. compounds with high leverage value (h) (h>h*, the critical value being h*=3p'/n, where p' is the number of model variables plus one, and n the number of the objects used to calculate the model). The hat (high leverage h) value is different for each model. The leverage approach was applied for the definition of the structural chemical domain model for new chemicals without experimental data.

5.2. Method used to assess the applicability domain:

Leverage (Williams) and standardization approach.

5.3. Software name and version for applicability domain assessment:

JBSMM software

5.4. Limits of applicability:

Full model: Applicability domain for melting point shows that there are only five compounds out of 303 compounds that are not satisfying the criteria used leverage value smaller than the warning leverage value.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: No

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

N/A

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

94 compounds were split into training and prediction set using to methods: by SOM and Response Splitting on the data. The two training sets - of 53 chemicals (SOM) and 48 chemicals (Response) were used to develop a population of models. Additional full model.

6.6. Pre-processing of data before modelling:

N/A

6.7. Statistics for goodness-of-fit:

a) Split by SOM: R^2=0.89, $RMSE_{TR}$=28.17

b) Random response activity: R^2=0.84, $RMSE_{TR}$=37.15

c) Full model: R^2=0.82, $RMSE_{TR}$=38.64

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

a) Split by SOM: Q^2_{LOO} = 0.84

b) Random response activity: Q^2_{LOO} = 0.81

c) Full model: Q^2_{LOO} = 0.77

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

94 compounds were split into training and prediction set using to methods: by SOM and Response Splitting on the data. The two training sets - of 53 chemicals (SOM) and 48 chemicals (Response) were used to develop a population of models. Additional full model. 15 compounds from PERFORCE were used for external validation as EV-set.

7.6. Experimental design of test set:

EV-set contains long chain perfluoroalkylated chemicals which are prevalent in the environment and are of higher interest in terms of physico-chemical properties and environmental distribution.

7.7. Predictivity – Statistics obtained by external validation:

a) Split by SOM:

$Q^2_{F1} = 0.76$, $Q^2_{F3} = 0.61$, $RMSE_{EXT} = 21.71$

EV-set: $Q^2_{F1} = 0.39$, $Q^2_{F3} = 0.80$, $RMSE_{EXT} = 36.88$,

b) Random response activity:

$Q^2_{F1} = 0.72$, $Q^2_{F3} = 0.76$, $RMSE_{EXT} = 46.06$

EV-set: $Q^2_{F1} = 0.56$, $Q^2_{F3} = 0.79$, $RMSE_{EXT} = 42.02$

c) Full model:

EV-set: $Q^2_{F1} = 0.65$, $Q^2_{F3} = 0.81$, $RMSE_{EXT} = 38.82$

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

N/A

8.2. A priori or a posteriori mechanistic interpretation:

$$MP = 2.34(\pm 0.28)\mathrm{In}^2 W - 164.76(\pm 16.51)W_{rel}^{F} + 33.14(\pm 8.90)N_{HBDON} + 79.38$$

where:

$\ln^2 W$ - general degree of branching through the square of the logarithm of Wiener index

W_{rel}^F - the number of fluorine atoms and their relative connectivity (this is the sum of all paths to fluorine atoms divided by the Wiener index W)

N_{HBDON} - denotes the number of H-bond donors in the molecule

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

Finally, a consensus model was developed for MP data by a simple average of the results predicted by all the models developed by project partners, such that each sub-model has equal weight. The correct experimental data were used for the calculation of statistical parameters. The consensus models were better than the individual submodels.

Statistics of consensus model:

$R^2 = 0.88$,

$RMSE_{TR} = 31.81$

9.2. Bibliography:

1. B. Bhhatarai, W. Teetz, T.Liu, T. Öberg, N. Jeliazkova, N. Kochev, O. Pukalov, I. V. Tetko, S. Kovarich, E. Papa, P. Gramatica. (2011). CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204. doi:10.1002/minf.201000133

2. R. D. Trepka, J. K. Harrington, J. W. McConville, K. T. McGurran, A. Mendel, D.R. Pauly, J.E. Robertson, J.T. Waddington, J. Agric. Food Chem. 1974, 22, 1111–1119. https://pubs.acs.org/doi/10.1021/jf60196a044

3. R. P. Gajewski, G. D. Thompson, E. H. Chio, J. Agric. Food Chem. 1988, 36, 174–177.

4.V. E. Platonov, A. Haas, M. Schelvis, M. Lieb, K. V. Dvornikova, O. I. Osina, Y. V. Gatilov, J. Fluorine Chem. 2001, 109, 131 – 139.

9.3. Supporting information:

doi:10.1002/minf.201000133

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

MP4. CADASTER QSPR Models for Predictions of Melting Point of Perfluorinated Chemicals (German Research Center for Environmental Health)

1. QSAR identifier

1.1. QSAR identifier (title):

CADASTER QSPR Models for Predictions of Melting Point of Perfluorinated Chemicals (German Research Center for Environmental Health)

1.2. Other related models:

CADASTER QSPR Models for Predictions of Melting Point of Perfluorinated Chemicals (developed by UI, LNU and IDEA)

1.3. Software coding the model:

OCHEM (on-line CHEMical database and Modeling environment)

https://www.ochem.eu/home/show.do

2. General information

2.1. Date of QMRF:

6/12/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.
Trzy Lipy 3 Street, Building B,
80-172 Gdansk, Poland
+48 795 160 760
contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

W. Teetz, I. V. Tetko Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum Muenchen – German Research Center for Environmental Health (HMGU), Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany.

2.6. Date of model development and/or publication:

Published in 2011

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

1. B. Bhhatarai, W. Teetz, T. Liu, T. Öberg, N. Jeliazkova, N. Kochev, O. Pukalov, I. V. Tetko, S. Kovarich, E. Papa, P. Gramatica. (2011). CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204. https://doi.org/10.1002/minf.201000133

2. OCHEM (on-line CHEMical database and Modeling environment) https://www.ochem.eu/home/show.do

2.8. Availability of information about the model:

For MP 93 compounds containing mainly long chain alkylates, saturated cyclic and few aromatic PFCs were split into training and prediction set close to 50% and 15 more compounds from PERFORCE were used as EV- set.

More information about model attached in supplementary materials: doi:10.1002/minf.201000133

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Physical Chemical Properties: Melting Point

3.3. Comment on the endpoint:

Melting Point (MP) is an important endpoint as it influence the transport, solubility, distribution and environmental fate.

3.4. Endpoint units:

°C

3.5. Dependent variable:

MP

3.6. Experimental protocol:

N/A

3.7. Endpoint data quality and variability:

- 54 data points of physiochemical property were collected mainly from SRC-PhysProp
- 19 data points from Trepka et al. [2]
- 7 data points from Gajewski et al. [3]
- 13 data points from Platonov et al. [4]

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

Associative neural networks (ASNN)

4.2. Explicit algorithm:

Associative neural networks (ASNN) represent a combination of an ensemble of feed-forward neural networks and kNN (k-Nearest Neighbors). The neural net- works ensemble of 100 networks with one hidden layer was used. The neural networks had 3 hidden neurons.

4.3. Descriptors in the model:

42 pre-filtered descriptors: SssssC, SsF, Se1C3O1ds, Se2C3O1s, SdssC, Se1C3C4ds, SsH, Se1H1O1s, SdO, SsCl, Se1C2H1s, Se1C3F1s, Se1C3H1s, Se1C2H1a, SaaCH, Se1C3C4as, SeaC2C3aa, SaasC, Se2C3C3ss, Se1C3F1d SsBr, Se1C4N3ss, Se1C3Cl1d, Se1Br1C3a, SsI, Se1C4I1s, SstC, SddssS, Se1H1N2s, Se1C1H1s, SsCH3, SssNH, Se1N2S4sd, Se1C3N3ad, SddsN, Se1C3N2as, Se1C3N2ds, Se1C3Cl1a, Se1C1N3ss, Se1C3N3as, Se1C3C3aa, Se1C1N2ss

Supersab, 1000 iterations, 3 neurons ensemble=100 k=64

4.4. Descriptor selection:

Constant values and descriptors that had Pearson pairwise correlation greater than 95% were excluded and all remaining indices were used to build models. For the final MP model, 87 filtered descriptors remained.

4.5. Algorithm and descriptor generation:

The electrotopological state (E-state) indices introduced by Hall and Kier combine together both electronic and topological characteristics of the analyzed molecules. For each atom type in a molecule the E-state indices are summed and are used in a group contribution manner.

4.6. Software name and version for descriptor generation:

OCHEM (on-line CHEMical database and Modeling anvironment)

https://www.ochem.eu/home/show.do

4.7. Chemicals/ Descriptors ratio:

93 chemicals / 87 descriptors

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The distance to model based on standard deviation of ensemble prediction was used. The DM was calibrated using 5-fold cross-validation values (i.e., for each DM value the corresponding prediction error was estimated). Using this calibrated DM the accuracy of prediction for any new molecule can be estimated. The standard deviations, which covered 95% of compounds from the training set, were used as ADs of the HMGU model. These ADs were used to compare results with other approaches reported in this study.

5.2. Method used to assess the applicability domain:

The AD of models was estimated using standard deviation (STD) of ensembles of neural network models.

5.3. Software name and version for applicability domain assessment:

OCHEM (on-line CHEMical database and Modeling environment)

https://www.ochem.eu/home/show.do

5.4. Limits of applicability:

The AD of models was estimated using standard deviation (STD) of ensembles of neural network models. As aforementioned, the STD covering 95% of molecules within the training set as the AD of the model was selected. For melting point, this cutoff STD value is 418, corresponding to a predicted RMSE of 50°C. Out of 303 new compounds, 212 were inside of AD of the model, i.e. we expect that their average predicted errors are below 50°C. The remaining 91 compounds were outside of AD and thus we expect that their average errors are above 50 K. Several compounds had extremely large STD values > 100°. These compounds are all cyclic carbonic acid esters of 2- (hydroxymethyl)-1,3-propanediol. They all have bridged ring structures with three ether bonds. The estimation of contributions of these structural features for MP is difficult since there are no similar structures in the training set.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability 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:

N/A

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

93 compounds were split into training and prediction set using to methods: by SOM and Response Splitting on the data. The two training sets - of 53 chemicals (SOM) and 47 chemicals (Response) were used to develop a population of models. Additional full model.

6.6. Pre-processing of data before modelling:

N/A

6.7. Statistics for goodness-of-fit:

a) Split by SOM: $R^2 = 0.80$, $RMSE_{TR} = 39.00$

b) Random response activity: $R^2 = 0.81$, $RMSE_{TR} = 43.00$

c) Full model: $R^2 = 0.85$, $RMSE_{TR} = 37.00$

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

N/A

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

93 compounds were split into training and prediction set using to methods: by SOM and Response Splitting on the data. The two training sets - of 53 chemicals (SOM) and 47 chemicals (Response) were used to develop a population of models. Additional full model. 15 compounds from PERFORCE were used for external validation as EV-set.

7.6. Experimental design of test set:

EV-set contains long chain perfluoroalkylated chemicals which are prevalent in the environment and are of higher interest in terms of physico-chemical properties and environmental distribution.

7.7. Predictivity – Statistics obtained by external validation:

a) Split by SOM:

$RMSE_{EXT} = 46.00$

EV-set: $RMSE_{EXT} = 39.00$,

b) Random response activity:

$RMSE_{EXT} = 47.00$

EV-set: $RMSE_{EXT} = 40.00$

c) Full model:

EV-set: $RMSE_{EXT} = 34.00$

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

An external validation was performed for the developed model after splitting the compounds into a validation and training set. Despite access to data for all compounds used in the model, they were not added to the corresponding sets.

For this model, additional external validation was performed using compounds for which the authors had collected experimental data. The EV-set was chosen to help as an additional validation of models that have demonstrated their validity during earlier splits.

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

N/A

8.2. A priori or a posteriori mechanistic interpretation:

N/A

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

Finally, a consensus model was developed for MP data by a simple average of the results predicted by all the models developed by project partners, such that each sub-model has equal weight. The correct experimental data were used for the calculation of statistical parameters. The consensus models were better than the individual submodels.

Statistics of consensus model:

$R^2 = 0.88$,

$RMSE_{TR} = 31.81$

9.2. Bibliography:

1. B. Bhhatarai, W. Teetz, T. Liu, T. Öberg, N. Jeliazkova, N. Kochev, O. Pukalov, I. V. Tetko, S. Kovarich, E. Papa, P. Gramatica. (2011). CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204. doi:10.1002/minf.201000133

2. R. D. Trepka, J. K. Harrington, J. W. McConville, K. T. McGurran, A. Mendel, D.R. Pauly, J.E. Robertson, J.T. Waddington, J. Agric. Food Chem. 1974, 22, 1111–1119. https://pubs.acs.org/doi/10.1021/jf60196a044

3. R. P. Gajewski, G. D. Thompson, E. H. Chio, J. Agric. Food Chem. 1988, 36, 174–177.

4.V. E. Platonov, A. Haas, M. Schelvis, M. Lieb, K. V. Dvornikova, O. I. Osina, Y. V. Gatilov, J. Fluorine Chem. 2001, 109, 131 – 139.

9.3. Supporting information:

More information about the model attached in supplementary materials: doi:10.1002/minf.201000133

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

BP1. CADASTER QSPR Models for Predictions of Melting Point of Perfluorinated Chemicals (University of Insubria)

1. QSAR identifier

1.1. QSAR identifier (title):

CADASTER QSPR Models for Predictions of Melting Point of Perfluorinated Chemicals (University of Insubria)

1.2. Other related models:

CADASTER QSPR Models for Predictions of Melting Point of Perfluorinated Chemicals (developed by LNU, IDEA and HMGU)

1.3. Software coding the model:

N/A

2. General information

2.1. Date of QMRF:

6/12/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.
Trzy Lipy 3 Street, Building B,
80-172 Gdansk, Poland
+48 795 160 760
contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

B. Bhhatarai, S. Kovarich, E. Papa, P. Gramatica, QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria (UI)

paola.gramatica@uninsubria.it http://www.qsar.it/

2.6. Date of model development and/or publication:

Published in 2011

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

1. B. Bhhatarai, W. Teetz, T. Liu, T. Öberg, N. Jeliazkova, N. Kochev, O. Pukalov, I. V. Tetko, S. Kovarich, E. Papa, P. Gramatica. (2011). CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204. https://doi.org/10.1002/minf.201000133

2. DRAGON https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_ molecular_descriptor_calculations

2.8. Availability of information about the model:

For MP 94 compounds containing mainly long chain alkylates, saturated cyclic and few aromatic PFCs were split into training and prediction set close to 50% and 15 more compounds from PERFORCE were used as EV- set.

More information about model attached in supplementary materials: doi:10.1002/minf.201000133

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Physical Chemical Properties: Melting Point

3.3. Comment on the endpoint:

Melting Point (MP) is an important endpoint as it influence the transport, solubility, distribution and environmental fate.

3.4. Endpoint units:

°C

3.5. Dependent variable:

MP

3.6. Experimental protocol:

N/A

3.7. Endpoint data quality and variability:

- 55 data points of physiochemical property were collected mainly from SRC-PhysProp
- 19 data points from Trepka et al. [2]
- 7 data points from Gajewski et al. [3]

- 13 data points from Platonov et al. [4]

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

Multiple linear regression model (OLS - Ordinary Least Square)

4.2. Explicit algorithm:

MP = 132.95(±22.63) AAC + 3.04(±0.97) F02[C-F] - 18.97(±7.98)C-013 + 209.04(±139.9) RBF - 243.16

4.3. Descriptors in the model:

1. AAC - 'information indices' which characterizes mean information index on atomic composition
2. F02[C-F] - frequency of C-F at topological distance 2
3. C-013 - carbon connected to at least three electronegative atoms (X) as CRX3
4. RBF - rotable bond fraction

4.4. Descriptor selection:

Only 1D and 2D descriptors were selected for modeling MP, easy and simply calculated from the SMILES.

- highly correlated descriptor were excluded (more than 90% pairwise correlation)
- filtered descriptors were subject to variable selection method using genetic algorithm (GA)
- using selected descriptors, a multiple linear regression (MLR) model by using the ordinary-least-squares (OLS) techniques were build

4.5. Algorithm and descriptor generation:

The input files for descriptor calculation were obtain by the semi empirical AM1 method (minimized their lowest energy conformation) using HYPERCHEM software. Molecular descriptors were generated using DRAGON software.

4.6. Software name and version for descriptor generation:

1. DRAGON sofware

A software to calculate theoretical molecular descriptors https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_ molecular_descriptor_calculations

2. HYPERCHEM

Software used for molecular drawing and conformational energy optimalization AM1 www. hyper.com

4.7. Chemicals/ Descriptors ratio:

SOM splitting: 53 chemicals / 4 descriptors = 13.25

Response splitting: 48 chemicals / 4 descriptors = 12

Full model: 94 chemicals / 4 descriptors = 23.5

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The Williams plot was used to verify the presence of response outliers (i.e. compounds with cross-validated standardized residuals greater than three standard deviation units, ±3σ) and chemicals very influential for their structure in determining model parameters (i.e. compounds with high leverage value (h) (h>h*, the critical value being h*=3p'/n, where p' is the number of model variables plus one, and n the number of the objects used to calculate the model). The hat (high leverage h) value is different for each model. The leverage approach was applied for the definition of the structural chemical domain model for new chemicals without experimental data.

5.2. Method used to assess the applicability domain:

Leverage (Williams) and standardization approach

5.3. Software name and version for applicability domain assessment:

OCHEM (on-line CHEMical database and Modeling environment)

https://www.ochem.eu/home/show.do

5.4. Limits of applicability:

Full model: The leverage approach based AD study has a hat cut-off values for the MP MLR model which is 0.16. Based on this cut-off, there were four training compounds (CAS 306-91-2, CAS 311-89-7, CAS 338-83-0 and CAS 359-70-6) which were outside the structural applicability domain of MP model that why they were highly influential in selecting modeling variables. In addition, there were 16 other new compounds which were out of the structural domain (high leverage or high hat values). They were bulky polyfluorinated chemicals, mostly linear PFCs with 13 or 15 carbon atom long including 2 acrylates and nitrogen centered PFCs. These long chain PFCs were probably extrapolated as the longest compound in the training model were with 7 carbon atoms only. Altogether only 20/397 compounds are found to be out of AD, thus the model has 94.9% coverage.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability 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:

N/A

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

94 compounds were split into training and prediction set using to methods: by SOM and Response Splitting on the data. The two training sets - of 53 chemicals (SOM) and 48 chemicals (Response) were used to develop a population of models. Additional full model.

6.6. Pre-processing of data before modelling:

N/A

6.7. Statistics for goodness-of-fit:

a) Split by SOM: R^2=0.83, $RMSE_{TR}$=33.95

b) Random response activity: R^2=0.84, $RMSE_{TR}$=37.16

c) Full model: R^2= 0.81, $RMSE_{TR}$=40.24

d) Full model (same data excluding compound CAS 426-65-3): R^2= 0.82, $RMSE_{TR}$=39.39

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

a) Split by SOM: Q^2_{LOO} = 0.79

b) Random response activity: Q^2_{LOO} = 0.79

c) Full model: Q^2_{LOO} = 0.78

d) Full model (same data excluding compound CAS 426-65-3): Q^2_{LOO} = 0.79

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

a) Split by SOM: R^2Y_{SCR} = 0.07

b) Random response activity: R^2Y_{SCR} = 0.08

c) Full model: R^2Y_{SCR} = 0.04

6.11. Robustness – Statistics obtained by bootstrap:

a) Split by SOM: $Q^2_{BOOT} = 0.78$

b) Random response activity: $Q^2_{BOOT} = 0.78$

c) Full model: $Q^2_{BOOT} = 0.77$

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

94 compounds were split into training and prediction set using to methods: by SOM and Response Splitting on the data. The two training sets - of 53 chemicals (SOM) and 48 chemicals (Response) were used to develop a population of models. Additional full model. 15 compounds from PERFORCE were used for external validation as EV-set.

7.6. Experimental design of test set:

EV-set contains long chain perfluoroalkylated chemicals which are prevalent in the environment and are of higher interest in terms of physico-chemical properties and environmental distribution.

7.7. Predictivity – Statistics obtained by external validation:

a) Split by SOM:

$Q^2_{F1} = 0.76$, $Q^2_{F3} = 0.62^*$, $RMSE_{EXT} = 51.69$

EV-set: $Q^2_{F1} = 0.72$, $Q^2_{F3} = 0.91$, $RMSE_{EXT} = 24.92$,

b) Random response activity:

$Q^2_{F1} = 0.73$, $Q^2_{F3} = 0.76$, $RMSE_{EXT} = 45.47$

EV-set: $Q^2_{F1} = 0.73$, $Q^2_{F3} = 0.87$, $RMSE_{EXT} = 32.08$

c) Full model:

EV-set: $Q^2_{F1} = 0.83$, $Q^2_{F3} = 0.91$, $RMSE_{EXT} = 27.19$

d) Full model (same data excluding compound CAS 426-65-3):

EV-set: $Q^2_{F1} = 0.81$, $Q^2_{F3} = 0.89$

$^*Q^2_{F3}$ was equal to 0.72 after removing 2 compounds (#28 CAS 354-32-5 and #14 CAS 309-91-2)

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

An external validation was performed for the developed model after splitting the compounds into a validation and training set. Despite access to data for all compounds used in the model, they were not added to the corresponding sets.

For this model, additional external validation was performed using compounds for which the authors had collected experimental data. The EV-set was chosen to help as an additional validation of models that have demonstrated their validity during earlier splits.

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

N/A

8.2. A priori or a posteriori mechanistic interpretation:

MP= 132.95(±22.63)AAC + 3.04(±0.97)F02[C-F] - 18.97(±7.98)C-013 + 209.04(±139.9)RBF - 243.16

where:

AAC – ‘information indices’ which characterizes mean information index on
atomic composition

F02[C-F] - Frequency of C-F at topological distance 2

C-013 - carbon connected to at least three electronegative atoms (X) CRX3

RBF - rotable bond fraction

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

Finally, a consensus model was developed for MP data by a simple average of the results predicted by all the models developed by project partners, such that each sub-model has equal weight. The correct experimental data were used for the calculation of statistical parameters. The consensus models were better than the individual submodels.

Statistics of consensus model:

$R^2 = 0.88$,

$RMSE_{TR} = 31.81$

9.2. Bibliography:

1. B. Bhhatarai, W. Teetz, T. Liu, T. Öberg, N. Jeliazkova, N. Kochev, O. Pukalov, I. V. Tetko, S. Kovarich, E. Papa, P. Gramatica. (2011). CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204. doi:10.1002/minf.201000133

2. R. D. Trepka, J. K. Harrington, J. W. McConville, K. T. McGurran, A. Mendel, D.R. Pauly, J.E. Robertson, J.T. Waddington, J. Agric. Food Chem. 1974, 22, 1111–1119. https://pubs.acs.org/doi/10.1021/jf60196a044

3. R. P. Gajewski, G. D. Thompson, E. H. Chio, J. Agric. Food Chem. 1988, 36, 174–177.

4.V. E. Platonov, A. Haas, M. Schelvis, M. Lieb, K. V. Dvornikova, O. I. Osina, Y. V. Gatilov, J. Fluorine Chem. 2001, 109, 131 – 139.

9.3. Supporting information:

More information about the model attached in supplementary materials:

doi:10.1002/minf.201000133

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

BP2. CADASTER QSPR Models for Predictions Boiling Point of Perfluorinated Chemicals (Linnaeus University)

1. QSAR identifier

1.1. QSAR identifier (title):

CADASTER QSPR Models for Predictions Boiling Point of Perfluorinated Chemicals (Linnaeus University)

1.2. Other related models:

CADASTER QSPR Models for Predictions of Boiling Point of Perfluorinated Chemicals (developed by UI, IDEA and HMGU)

1.3. Software coding the model:

SIMCA-P, v.11.0 Umetrics AB

2. General information

2.1. Date of QMRF:

6/12/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.
Trzy Lipy 3 Street, Building B,
80-172 Gdansk, Poland
+48 795 160 760
contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

T. Liu, T. Öberg

School of Natural Sciences, Linnaeus University (LNU)SE-39182, Kalmar, Sweden

2.6. Date of model development and/or publication:

Published in 2011

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

1. B. Bhhatarai, W. Teetz, T. Liu, T. Öberg, N. Jeliazkova, N. Kochev, O. Pukalov, I. V. Tetko, S. Kovarich, E. Papa, P. Gramatica. (2011). CADASTER QSPR Models for Predictions of

Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204. https://doi.org/10.1002/minf.201000133

2. DRAGON

https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_ molecular_descriptor_calculations

2.8. Availability of information about the model:

The 93 compounds compiled were split into training and prediction set by 50%, based on two different splitting criteria, random by response activity and Self Organizing Map (kANN). In addition, 25 compounds for PERFORCE were used for external validation as EV-set.

More information about model attached in supplementary materials: doi:10.1002/minf.201000133.

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Physical Chemical Properties: Boiling Point

3.3. Comment on the endpoint:

Boiling Point (BP) is an important endpoint as it influences the transport, solubility, distribution and environmental fate. The BP of a liquid normally depends on the atom type, molecular mass and the intermolecular force. The typical character of fluorine, small size and high electronegativity, and a polar C–F bond may complicate the phenomenon of boiling.

3.4. Endpoint units:

°C

3.5. Dependent variable:

BP

3.6. Experimental protocol:

N/A

3.7. Endpoint data quality and variability:

-77 data points of physiochemical property were collected mainly from SRC-PhysProp

-16 data points from Hendricks et al.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

Partial least squares regression (PLSR)

4.2. Explicit algorithm:

PLSR is based on a linear transformation of the theoretical molecular descriptors to a limited number of orthogonal factors, attempting to maximize the covariance between the descriptors and the response variable

4.3. Descriptors in the model:

Descriptors belonging to various types (e.g., constitutional, topological, walk and path counts, connectivity index, information index, 2D autocorrelations, edge adjacency indices, BCUT descriptors, topological charge index, eigenvalue-based index, Randic molecular profiles, geometrical, RDF descriptors, 3D-MoRSE descriptors, WHIM, GETAWAY, functional group counts, atom-centered fragments, charge, molecular properties, 2D Binary and 2D frequency fingerprints) were calculated by using Dragon. These molecular descriptors were used as input variables for modeling in order to capture maximum of the relevant structural features related to the response.

4.4. Descriptor selection:

- PLSR was used for data analysis and modeling
- the data analysis and multivariate calibrations were carried out with the software SIMCA-P
- the 149 descriptors at 4 components using VIP approach were selected

4.5. Algorithm and descriptor generation:

The input files for descriptor calculation were obtain by the semi empirical AM1 method (minimized their lowest energy conformation) using HYPERCHEM software. Molecular descriptors were generated using DRAGON software

4.6. Software name and version for descriptor generation:

1. DRAGON sofware

A software to calculate theoretical molecular descriptors https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_ molecular_descriptor_calculations

2. HYPERCHEM

Software used for molecular drawing and conformational energy optimalization AM1 www. hyper.com

4.7. Chemicals/ Descriptors ratio:

93 chemicals / 149 descriptors at 4 components

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The valid applicability domain for the PLSR model was assessed by the residual standard deviation (the Euclidean distance to the PLSR model) and the leverage (the Mahalanobis distance within the PLSR model space) to that of the calibration objects. From these distances the SIMCA software estimates the probability that a new compound belongs to the model (PModXPS and PModXPS +). Here we used a probability of belonging to the model at the 99% significant level (PModXPS < 0.01 and PModXPS + < 0.01) as the cut-off criterion to decide if a compound was outside of the model AD. The 1% clarification limit is based on the training set properties, meaning that using a higher cut-off would put several of the calibration compounds outside of the applicability domain, which does not seem appropriate.

5.2. Method used to assess the applicability domain:

Residual standard deviation (the Euclidean distance to the PLSR model) and the leverage (the Mahalanobis distance within the PLSR model space) approach.

5.3. Software name and version for applicability domain assessment:

SIMCA-P, v.11.0 Umetrics AB

5.4. Limits of applicability:

93 compounds with BP were used as calibration set to study AD, which selected the 149 largest descriptors based on VIP at 4 components. 20 compounds above the 1% critical limit were considered to be outside the domain. After excluding the outliers, 96.84 % variance in Y is described,and 94.70% variance is predicted by the model at 4 components. Then 271 compounds without any BP values as prediction set were used to determine their domain, 171 compounds were tested to be outside the domain using PModXPS+ approach, coverage of 36.9%.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability 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:

N/A

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

93 compounds were split into training and prediction set using to methods: by SOM and Random response activity. The two training sets - of 50 chemicals (SOM) and 47 chemicals (Random) were used to develop a population of models

6.6. Pre-processing of data before modelling:

All descriptor variables were preprocessed by auto-scaling to zero mean and unit variance.

6.7. Statistics for goodness-of-fit:

a) Split by SOM: $R^2 = 0.97$, $RMSE_{TR} = 11.73$

b) Random response activity: $R^2 = 0.97$, $RMSE_{TR} = 13.69$

c) Full model: $R^2= 0.97$, $RMSE_{TR} = 14.11$

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

a) Split by SOM: $Q^2_{LOO} = 0.94$

b) Random response activity: $Q^2_{LOO} = 0.92$

c) Full model: $Q^2_{LOO} = 0.94$

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

25 compounds for PERFORCE were used for external validation as EV-set.

7.6. Experimental design of test set:

EV-set contains long chain perfluoroalkylated chemicals which are prevalent in the environment and are of higher interest in terms of physico-chemical properties and environmental distribution.

7.7. Predictivity – Statistics obtained by external validation:

a) Split by SOM:

$RMSE_{EXT} = 19.36$

EV-set: $RMSE_{EXT} = 21.69$

b) Random response activity:

$RMSE_{EXT} = 20.10$

EV-set: $RMSE_{EXT} = 16.50$

c) Full model:

EV-set: $RMSE_{EXT} = 21.92$

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

An external validation was performed for the developed model after splitting the compounds into a validation and training set. Despite access to data for all compounds used in the model, they were not added to the corresponding sets.

For this model, additional external validation was performed using compounds for which the authors had collected experimental data. The EV-set was chosen to help as an additional validation of models that have demonstrated their validity during earlier splits

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

N/A

8.2. A priori or a posteriori mechanistic interpretation:

In the PLSR model, the first latent variables explained 83.50 % of the variance in BP and 30.80% of the variance in the calculated descriptors; these descriptors were related to size and thus dominated by van der Waals' interactions. The second latent variables explained an additional 9.86 % of the variance in BP and an additional 10.60 % variance in the calculated descriptors, and these descriptors were mainly related to polar interactions. The third latent variables described only 3.16 % of the variance in BP and 7.36 % in calculated descriptors.

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

Finally, a consensus model was developed for BP data by a simple average of the results predicted by all the models developed by project partners, such that each sub-model has equal weight. The correct experimental data were used for the calculation of statistical parameters. The consensus models were better than the individual submodels.

Statistics of consensus model:

$R^2 = 0.96$,

$RMSE_{TR} = 14.92$

9.2. Bibliography:

1. B. Bhhatarai, W. Teetz, T. Liu, T. Öberg, N. Jeliazkova, N. Kochev, O. Pukalov, I. V. Tetko, S. Kovarich, E. Papa, P. Gramatica. (2011). CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204. doi:10.1002/minf.201000133

2. J. O. Hendricks, Ind. Eng. Chem. 1953, 45, 99 – 105

9.3. Supporting information:

More information about the model attached in supplementary materials: doi:10.1002/minf.201000133

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

BP3. CADASTER QSPR Models for Predictions Boiling Point of Perfluorinated Chemicals (Ideaconsult Ltd.)

1. QSAR identifier

1.1. QSAR identifier (title):

CADASTER QSPR Models for Predictions Boiling Point of Perfluorinated Chemicals (Ideaconsult Ltd.)

1.2. Other related models:

CADASTER QSPR Models for Predictions of Boiling Point of Perfluorinated Chemicals (developed by UI, LNU and HMGU)

1.3. Software coding the model:

JBSMM software

N. Kochev, O. Pukalov, Software system for molecular modeling–JBSMM; Scientific Researches of the Union of Scientists in Bulgaria – Plovdiv, Series C: Technics and Technologies, Vol. V., Balkan Conference of Young Scientists; Plovdiv, 16–18 June, 2005, pp. 315 – 320.

2. General information

2.1. Date of QMRF:

6/12/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

N. Jeliazkova Ideaconsult Ltd. (IDEA) 4A. Kanchev str., Sofia 1000, Bulgaria

2.6. Date of model development and/or publication:

Published in 2011

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

1. Bhhatarai, B., Teetz, W., Liu, T., Öberg, T., Jeliazkova, N., Kochev, N., Pukalov, O., Tetko, I.V., Kovarich, S., Papa, E., Gramatica, P. (2011). CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204.

https://doi.org/10.1002/minf.201000133

2. JBSMM software

N. Kochev, O. Pukalov, Software system for molecular modeling–JBSMM; Scientific Researches of the Union of Scientists in Bulgaria – Plovdiv, Series C: Technics and Technologies, Vol. V., Balkan Conference of Young Scientists; Plovdiv, 16–18 June, 2005, pp. 315 – 320.

2.8. Availability of information about the model:

The 93 compounds compiled were split into training and prediction set by 50%, based on two different splitting criteria, random by response activity and Self Organizing Map (kANN). In addition, 25 compounds for PERFORCE were used for external validation as EV-set.

More information about model attached in supplementary materials: doi:10.1002/minf.201000133.

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Physical Chemical Properties: Boiling Point

3.3. Comment on the endpoint:

Boiling Point (BP) is an important endpoint as it influence the transport, solubility, distribution and environmental fate. The BP of a liquid normally depends on the atom type, molecular mass and the intermolecular force. The typical character of fluorine, small size and high electronegativity, and a polar C–F bond may complicate the phenomenon of boiling.

3.4. Endpoint units:

°C

3.5. Dependent variable:

BP

3.6. Experimental protocol:

N/A

3.7. Endpoint data quality and variability:

-77 data points of physiochemical property were collected mainly from SRC-PhysProp

-16 data points from Hendricks et al.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

Multiple linear regression model (OLS - Ordinary Least Square)

4.2. Explicit algorithm:

$BP = 138.44(\pm4.24)MW^{1/4} - 220.98(\pm11.25)W_{rel}^{F} + 42.83(\pm6.64)N_{HBDON} - 24.52(\pm4.32) N_{[Cl,Br,I]} - 32.67(\pm5.68)N_{COC} - 50.08(\pm9.85)N_{C(=O)([!O])[!O]} - 18.58(\pm6.71)N_{N} - 188.36$

4.3. Descriptors in the model:

1. $MW^{1/4}$ - molecular weight of power 0.25 accounts for rhe antire skeleton of the molecule
2. W_{rel}^{F} - the relative connectivity of all fluorine atoms
3. N_{HBDON} - denotes the number of H-bond donors in the molecule
4. $N_{[Cl,Br,I]}$ - accounts for the halogen atoms, except fluorine
5. N_{COC} - corresponds to the ether groups
6. $N_{C(=O)([!O])[!O]}$ - accounts the carbonyl groups
7. N_{N} – numer of nitrogen atoms

4.4. Descriptor selection:

- descriptors were chosen manualy in a step-wise manner base on expert knowledge
- choice of descriptors combination was guided by three practicaltules: (1) each coeff in the MLRA must be statistically significant based on RSD (RSD<25%); (2) the choice of extra descriptors obeys a chemical logic in order to fix compounds with bad model values trying to handle the missing structural fragments; (3) successively selected descriptors must decrease RMSE value and increase Q2 and R2 values

4.5. Algorithm and descriptor generation:

Set of topological, constitutional and group based descriptors were calculated using JBSMM software. The descriptor combinations were chosen manually in a step-wise manner applying chemical expert knowledge.

4.6. Software name and version for descriptor generation:

1. JBSMM software

N. Kochev, O. Pukalov , Software system for molecular modeling–JBSMM; Scientific Researches of the Union of Scientists in Bulgaria – Plovdiv, Series C: Technics and Technologies, Vol. V., Balkan Conference of Young Scientists; Plovdiv, 16–18 June, 2005, pp. 315 – 320.

4.7. Chemicals/ Descriptors ratio:

SOM splitting: 50 chemicals / 7 descriptors = 7.14

Random response activity: 47 chemicals / 7 descriptors = 6.71

Full model: 93 chemicals / 7 descriptors = 13.29

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The Williams plot was used to verify the presence of response outliers (i.e. compounds with cross-validated standardized residuals greater than three standard deviation units, ±3σ) and chemicals very influential for their structure in determining model parameters (i.e. compounds with high leverage value (h) (h>h*, the critical value being h*=3p'/n, where p' is the number of model variables plus one, and n the number of the objects used to calculate the model). The hat (high leverage h) value is different for each model. The leverage approach was applied for the definition of the structural chemical domain model for new chemicals without experimental data

5.2. Method used to assess the applicability domain:

Leverage (Williams) and standardization approach.

5.3. Software name and version for applicability domain assessment:

JBSMM software

5.4. Limits of applicability:

Full model: 115 compounds out of 271 testing compounds fall outside AD of the BP model, a coverage of 57.5%. Applicability domain includes all compounds that have leverage value smaller than the critical value h*. Most of the compounds that fall outside AD have one or more fragments that are not fully described with the current model since they were not represented in the training data set. Predominantly these are groups containing nitrogen atoms e.g. nitriles and nitro groups. The model contains only one descriptor taking into account the presence of nitrogen atoms. Primary, secondary and tertiary amines are not distinguished as well as nitrogen atoms in cyclic and aromatic systems. Also were observed another group of compounds falling outside AD – compounds containing oxygen atoms in cyclic systems (or both oxygen and nitrogen atoms together).

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability 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:

N/A

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

93 compounds were split into training and prediction set using to methods: by SOM and Random response activity. The two training sets - of 50 chemicals (SOM) and 47 chemicals (Random) were used to develop a population of models.

6.6. Pre-processing of data before modelling:

N/A

6.7. Statistics for goodness-of-fit:

a) Split by SOM: $R^2 = 0.94$, $RMSE_{TR} = 16.74$

b) Random response activity: $R^2 = 0.96$, $RMSE_{TR} = 16.51$

c) Full model: $R^2 = 0.95$, $RMSE_{TR} = 17.25$

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

Full model: $Q^2_{LOO} = 0.94$

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

Full model: $R^2Y_{SCR} = 0.07$

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

25 compounds for PERFORCE were used for external validation as EV-set.

7.6. Experimental design of test set:

EV-set contains long chain perfluoroalkylated chemicals which are prevalent in the environment and are of higher interest in terms of physico-chemical properties and environmental distribution.

7.7. Predictivity – Statistics obtained by external validation:

a) Split by SOM:

$Q^2_{F1} = 0.94$, $Q^2_{F3} = 0.92$, $RMSE_{EXT} = 20.31$

EV-set: $Q^2_{F1} = 0.96$, $Q^2_{F3} = 0.92$, $RMSE_{EXT} = 20.21$

b) Random response activity:

$Q^2_{F1} = 0.92$, $Q^2_{F3} = 0.93$, $RMSE_{EXT} = 20.70$

EV-set: $Q^2_{F1} = 0.97$, $Q^2_{F3} = 0.94$, $RMSE_{EXT} = 18.90$

c) Full model:

EV-set: $Q^2_{F1} = 0.96$, $Q^2_{F3} = 0.92$, $RMSE_{EXT} = 22.56$

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

An external validation was performed for the developed model after splitting the compounds into a validation and training set. Despite access to data for all compounds used in the model, they were not added to the corresponding sets.

For this model, additional external validation was performed using compounds for which the authors had collected experimental data. The EV-set was chosen to help as an additional validation of models that have demonstrated their validity during earlier splits

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

N/A

8.2. A priori or a posteriori mechanistic interpretation:

$BP = 138.44(\pm4.24)MW^{1/4} - 220.98(\pm11.25)W_{rel}^{F} + 42.83(\pm6.64)N_{HBDON} - 24.52(\pm4.32) N_{[Cl,Br,I]} - 32.67(\pm5.68)N_{COC} - 50.08(\pm9.85)N_{C(=O)([!O])[!O]} - 18.58(\pm6.71)N_{N} - 188.36$

where:

1. $MW^{1/4}$ - molecular weight of power 0.25 accounts for rhe antire skeleton of the molecule
2. W_{rel}^{F} - the relative connectivity of all fluorine atoms
3. N_{HBDON} - denotes the number of H-bond donors in the molecule
4. $N_{[Cl,Br,I]}$ - accounts for the halogen atoms, except fluorine
5. N_{COC} - corresponds to the ether groups
6. $N_{C(=O)([!O])[!O]}$ - accounts the carbonyl groups
7. N_{N} – number of nitrogen atoms

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

Finally, a consensus model was developed for BP data by a simple average of the results predicted by all the models developed by project partners, such that each sub-model has equal weight. The correct experimental data were used for the calculation of statistical parameters. The consensus models were better than the individual submodels.

Statistics of consensus model:

$R^2 = 0.96$,

$RMSE_{TR} = 14.92$

9.2. Bibliography:

1. B. Bhhatarai, W. Teetz, T. Liu, T. Öberg, N. Jeliazkova, N. Kochev, O. Pukalov, I. V. Tetko, S. Kovarich, E. Papa, P. Gramatica. (2011). CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204. doi:10.1002/minf.201000133

2. J. O. Hendricks, Ind. Eng. Chem. 1953, 45, 99 – 105

9.3. Supporting information:

More information about model attached in supplementary materials: doi:10.1002/minf.201000133.

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

BP4. CADASTER QSPR Models for Predictions Boiling Point of Perfluorinated Chemicals (German Research Center for Environmental Health)

1. QSAR identifier

1.1. QSAR identifier (title):

CADASTER QSPR Models for Predictions Boiling Point of Perfluorinated Chemicals (German Research Center for Environmental Health)

1.2. Other related models:

CADASTER QSPR Models for Predictions of Boiling Point of Perfluorinated Chemicals (developed by UI, LNU and IDEA)

1.3. Software coding the model:

OCHEM (on-line CHEMical database and Modeling environment)

https://www.ochem.eu/home/show.do

2. General information

2.1. Date of QMRF:

6/12/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.
Trzy Lipy 3 Street, Building B,
80-172 Gdansk, Poland
+48 795 160 760
contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

W. Teetz, I. V. Tetko Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum Muenchen – German Research Center for Environmental Health (HMGU), Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany.

2.6. Date of model development and/or publication:

Published in 2011

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

1. B. Bhhatarai, W. Teetz, T. Liu, T. Öberg, N. Jeliazkova, N. Kochev, O. Pukalov, I. V. Tetko, S. Kovarich, E. Papa, P. Gramatica. (2011). CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204. https://doi.org/10.1002/minf.201000133

2. OCHEM (on-line CHEMical database and Modeling environment) https://www.ochem.eu/home/show.do

2.8. Availability of information about the model:

The 93 compounds compiled were split into training and prediction set by 50%, based on two different splitting criteria, random by response activity and Self Organizing Map (kANN). In addition, 25 compounds for PERFORCE were used for external validation as EV-set.

More information about model attached in supplementary materials: doi:10.1002/minf.201000133.

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Physical Chemical Properties: Boiling Point

3.3. Comment on the endpoint:

Boiling Point (BP) is an important endpoint as it influence the transport, solubility, distribution and environmental fate. The BP of a liquid normally depends on the atom type, molecular mass and the intermolecular force. The typical character of fluorine, small size and high electronegativity, and a polar C–F bond may complicate the phenomenon of boiling.

3.4. Endpoint units:

°C

3.5. Dependent variable:

BP

3.6. Experimental protocol:

N/A

3.7. Endpoint data quality and variability:

-77 data points of physiochemical property were collected mainly from SRC-PhysProp

-16 data points from Hendricks et al.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

Associative neural networks (ASNN)

4.2. Explicit algorithm:

Associative neural networks (ASNN) represent a combination of an ensemble of feed-forward neural net- works and kNN (k-Nearest Neighbors). [The neural net- works ensemble of 100 networks with one hidden layer was used. The neural networks had 3 hidden neurons.

4.3. Descriptors in the model:

N/A

4.4. Descriptor selection:

Constant values and descriptors that had Pearson pairwise correlation greater than 95 % were excluded and all remaining indices were used to build models. For the final MP model, 66 filtered descriptors remained.

4.5. Algorithm and descriptor generation:

The electrotopological state (E-state) indices introduced by Hall and Kier combine together both electronic and topological characteristics of the analyzed molecules. For each atom type in a molecule the E-state indices are summed and are used in a group contribution manner.

4.6. Software name and version for descriptor generation:

OCHEM (on-line CHEMical database and Modeling anvironment)

https://www.ochem.eu/home/show.do

4.7. Chemicals/ Descriptors ratio:

N/A

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The distance to model based on standard deviation of ensemble prediction was used. The DM was calibrated using 5-fold cross-validation values (i.e., for each DM value the corresponding prediction error was estimated). Using this calibrated DM the accuracy of prediction for any new molecule can be estimated. The standard deviations, which covered 95% of compounds from the training set, were used as ADs of the HMGU model. These ADs were used to compare results with other approaches reported in this study.

5.2. Method used to assess the applicability domain:

The AD of models was estimated using standard deviation (STD) of ensembles of neural network models.

5.3. Software name and version for applicability domain assessment:

OCHEM (on-line CHEMical database and Modeling anvironment)

https://www.ochem.eu/home/show.do

5.4. Limits of applicability:

The STD covering 95% of molecules within the training set (STD = 55) was used as the AD of the model for BP. This cutoff value corresponded to a predicted RMSE of 42K. For the boiling point model, 191 out of 271 compounds were within the AD of the model. Seven compounds had a very large standard deviation above 100°C, so their prediction accuracy can be expected to be very low. Six of these are cyclic amines out of which 5 have two amino groups in meta position in the ring while the other has one amino group in a ring of size 8. The seventh compound contains a hydrazine group. Apparently, there is considerable variance between the neural network outputs for compounds containing multiple amino groups in rings. The training data contains multiple tertiary amines, cyanide and an aromatic primary amine. The latter was already mentioned as a model outlier for BP, as the amino group strongly increases its dipole moment and BP. For all these cases, replacing the nitrogen atom with a carbon atom lowers the BP by around 50°C. On the other hand, faced with differently connected nitrogen atoms, this prediction can be expected to be very imprecise. Inserting nitrogen atoms into an aromatic ring usually has a much smaller effect on BP, e.g. the BP for benzene is 80°C, for pyridine 115°C and for pyrimidine 124°C, so the second nitrogen accounts for only 98°C change. It is promising that the ASNN predictions for compounds containing multiple such nitrogen atoms are diverging and these compounds are thus rightfully and decisively excluded from the AD.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability 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:

N/A

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

93 compounds were split into training and prediction set using to methods: by SOM and Random response activity. The two training sets - of 50 chemicals (SOM) and 47 chemicals (Random) were used to develop a population of models.

6.6. Pre-processing of data before modelling:

N/A

6.7. Statistics for goodness-of-fit:

a) Split by SOM: $R^2 = 0.79$, $RMSE_{TR} = 36.00$

b) Random response activity: $R^2 = 0.78$, $RMSE_{TR} = 41.00$

c) Full model: $R^2 = 0.85$, $RMSE_{TR} = 32.00$

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

N/A

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

25 compounds for PERFORCE were used for external validation as EV-set.

7.6. Experimental design of test set:

EV-set contains long chain perfluoroalkylated chemicals which are prevalent in the environment and are of higher interest in terms of physico-chemical properties and environmental distribution.

7.7. Predictivity – Statistics obtained by external validation:

a) Split by SOM:

$RMSE_{EXT} = 56.00$

EV-set: $RMSE_{EXT} = 37.00$

b) Random response activity:

$RMSE_{EXT} = 59.00$

EV-set: $RMSE_{EXT} = 27.00$

c) Full model:

EV-set: $RMSE_{EXT} = 22.00$

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

An external validation was performed for the developed model after splitting the compounds into a validation and training set. Despite access to data for all compounds used in the model, they were not added to the corresponding sets.

For this model, additional external validation was performed using compounds for which the authors had collected experimental data. The EV-set was chosen to help as an additional validation of models that have demonstrated their validity during earlier splits

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

N/A

8.2. A priori or a posteriori mechanistic interpretation:

N/A

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

Finally, a consensus model was developed for BP data by a simple average of the results predicted by all the models developed by project partners, such that each sub-model has equal weight. The correct experimental data were used for the calculation of statistical parameters. The consensus models were better than the individual submodels.

Statistics of consensus model:

$R^2 = 0.96$,

$RMSE_{TR} = 14.92$

9.2. Bibliography:

1. B. Bhhatarai, W. Teetz, T. Liu, T. Öberg, N. Jeliazkova, N. Kochev, O. Pukalov, I. V. Tetko, S. Kovarich, E. Papa, P. Gramatica. (2011). CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. Molecular Informatics, 30(2-3), 189–204. doi:10.1002/minf.201000133

2. J. O. Hendricks, Ind. Eng. Chem. 1953, 45, 99 – 105

9.3. Supporting information:

More information about model attached in supplementary materials: doi:10.1002/minf.201000133.

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

CMC. Prediction of Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals

1. QSAR identifier

1.1. QSAR identifier (title):

Prediction of Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals

1.2. Other related models:

N/A

1.3. Software coding the model:

Mobydigs software

Todeschini,R.;Consonni,V.;Pavan,M.MOBYDIGS—Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows, Talete srl: Milan, Italy, 2002

2. General information

2.1. Date of QMRF:

3/01/2022

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.
Trzy Lipy 3 Street, Building B,
80-172 Gdansk, Poland
+48 795 160 760
contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

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/

2.6. Date of model development and/or publication:

Published in 2011

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

1. Bhhatarai, B., & Gramatica, P. (2011). Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals. Environmental Science & Technology, 45(19), 8120–8128 https://pubs.acs.org/doi/10.1021/es101181g

2. DRAGON software https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_ molecular_descriptor_calculations

2.8. Availability of information about the model:

A total of 10 CMC experimental values for PFCs were collected and due to its limited size this small data set was not split into training and prediction set. The existing data, even if limited, could be a source of useful information as it contains diverse structural information.Supporting info: https://doi.org/10.1021/es101181g

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Critical Micelle Concentration

3.3. Comment on the endpoint:

Critical Micelle Concentration: logCMC

3.4. Endpoint units:

mol/L

3.5. Dependent variable:

logCMC

3.6. Experimental protocol:

The experimental data generated using the surface tension method collected at 298 K were considered except for PFOSH where the conductivity method was used.

3.7. Endpoint data quality and variability:

For CMC (in M, i.e., mol/L), 10 compounds were used for QSPR modeling compiled mainly from PERFORCE which has collected data from the literature.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSPR - Multiple linear regression model (OLS - Ordinary Least Square)

4.2. Explicit algorithm:

Full model equation:

logCMC = 1.351(± 0.183) - 0.30(± 0.018)X3

4.3. Descriptors in the model:

X3 - bidimensional descriptor, connectivity indices which represents the molecular complexity and branching

4.4. Descriptor selection:

The reduced sets of input descriptors were subjected to variable selection method using Genetic Algorithm (GA). Genetic algorithm was applied of the set for approximately 400 molecular descriptors for each compound.GA was applied to choose the best set of few descriptors, which have the most relevant variables, in combination, in modeling studied property.

4.5. Algorithm and descriptor generation:

The input files for descriptor calculation were obtain by the semi empirical AM1 method (minimized their lowest energy conformation) using HYPERCHEM software.

Molecular descriptors were generated using DRAGON software.

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

4.6. Software name and version for descriptor generation:

1. DRAGON software

A software to calculate theoretical molecular descriptors

https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_molecular_descriptor_calculations

2. HYPERCHEM

Software used for molecular drawing and conformational energy optimization AM1

www.hyper.com

4.7. Chemicals/ Descriptors ratio:

Full model: 10 chemicals / 1 descriptors = 10

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The Williams plot was used to verify the presence of response outliers (i.e., compounds with cross-validated standardized residuals greater than 2.5 standard deviation units, and chemicals very influential for their structure in determining model parameters (i.e., compounds with high leverage value (h) (h > h*, the critical value being h* = 3p' /n, where p’ is the number of model variables plus one, and n is the number of the objects used to calculate the model). The leverage approach was applied also for the definition of the structural chemical domain of each model for chemicals without experimental data by plotting Y-predicted versus hat value. The predictions for compounds having high leverage value are extrapolated and should be considered less reliable, but those interpolated within the training domain should be predicted with similar accuracy as for training chemicals.

5.2. Method used to assess the applicability domain:

To obtain structural AD the leverage approach providing a cut-off value was used (h* = 0.60).

5.3. Software name and version for applicability domain assessment:

Mobydigs software

Todeschini,R.; Consonni,V.;Pavan, M.MOBYDIGS—Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows, Talete srl: Milan, Italy, 2002

5.4. Limits of applicability:

For structural AD, a data set of 211 extra compounds was used, the average hat value obtained was 0.60 and 51 compounds (76.9% coverage) were found outside the structural applicability domain.

The compounds with high hat values belong to 18C long acrylates (CAS 59778-97-1, CAS 65150-93-8), acid (CAS 16517-11-6), and iodides (CAS 29809-35-6, CAS 65150-94-9). The broader insight to bigger set based on information from 10 compounds might be an overshoot but the structural diversity of the small data set motivated such AD study. Besides, the lack of public experimental data and a need to estimate the property calls for this analysis. It is evident that the presence of more experimental data for modeling will definitely help to encode key structural features which will govern the CMC behavior typical of PFCs, but for most of them the model based on a simple descriptor as 3 can be used in order to estimate CMC.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: No

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

N/A

6.6. Pre-processing of data before modelling:

The original CMC data were expressed in log unit logCMC (mol/L)

6.7. Statistics for goodness-of-fit:

R^2= 97.35

$RMSE_{TR}$=0.15

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

$Q2_{LOO}$=95.93

$RMSE_{CV}$=0.18

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

R^2Y_{SCR}=10.94

6.11. Robustness – Statistics obtained by bootstrap:

$Q2_{BOOT}$ = 95.96

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

N/A

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

N/A

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

The model was developed by statistical approach. No mechanistic was defined a priori.

8.2. A priori or a posteriori mechanistic interpretation:

$\log CMC = 1.351(\pm 0.183) - 0.30(\pm 0.018)X3$

X3 - bidimensional descriptor, connectivity indices which represents the molecular complexity and branching

The model was based on a bidimensional descriptor, belonging to the connectivity indices which represents the molecular complexity and branching. As CMC is reported to be independent of branching, thus the good inverse correlation with descriptor is able to differentiate the complexity of PFCs in terms of structure. It has higher values for longer PFCs and it could differentiate between different functional groups as sulfonates and carboxylates.

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Bhhatarai, B., & Gramatica, P. (2011). Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals. Environmental Science & Technology, 45(19), 8120–8128 https://pubs.acs.org/doi/10.1021/es101181g

2. Shinoda, K.; Hato, M.; Hayashi, T. The physico-chemical properties of aqueous solutions of fluorinated surfactants. J. Phys. Chem. 1972, 76, 909–914 https://pubs.acs.org/doi/10.1021/j100650a021

3. Kunieda, H.; Shinoda, K. Krafft points, critical micelle concentrations, surface tension, and solubilizing power of aqueous solutions of fluorinated surfactants. J. Phys. Chem. 1976, 80, 2468–2470 https://pubs.acs.org/doi/10.1021/j100563a007

4. Guo, W. T.; Fung, B. M. Micelles and aggregates of fluorinated surfactants. J. Phys. Chem. 1991, 95, 1829–1836. https://pubs.acs.org/doi/10.1021/j100157a060

9.3. Supporting information:

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

DF. Relationships between quantum chemical parameters and the reported overall defluorination ratio (deF%) of multiple PFAS structural categories with quantitative structure-activity relationship (QSAR) models

1. QSAR identifier

1.1. QSAR identifier (title):

Relationships between quantum chemical parameters and the reported overall defluorination ratio (deF%) of multiple PFAS structural categories with quantitative structure-activity relationship (QSAR) models.

1.2. Other related models:

N/A

1.3. Software coding the model:

SPSS 20.0. Statistical Package for the Science

2. General information

2.1. Date of QMRF:

09/11/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com

https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Zhemin Shen School of Environmental Science and Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, P.R. China; State Environmental Protection Key Laboratory of Environmental Health Impact Assessment of Emerging Contaminants, Shanghai 200240, P.R. China; Shanghai Engineering Research Center of Solid Waste Treatment and Resource Recovery, Shanghai 200240, P.R. China zmshen@sjtu.edu.cn

2.6. Date of model development and/or publication:

Published online: 2021

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

1. Gaussian 09 Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; et al. Gaussian 09, rev. C.01; Gaussian Inc.: Wallingford, CT, 2009.
2. Material Studio 7.0 Module, C. C. Material Studio, ver. 7.0; Accelrys Inc.: San Diego, 2013.
3. Zhiwen Cheng, Qincheng Chen, Zekun Liu, Jinyong Liu, Yawei Liu, Shiqiang Liu, Xiaoping Gao, Yujia Tan, and Zhemin Shen, ,, Interpretation of Reductive PFAS Defluorination with Quantum Chemical Parameters'',Environmental Science & Technology Letters 2021 8 (8), 645-650
DOI: 10.1021/acs.estlett.1c00403

2.8. Availability of information about the model:

An optimal QSAR model that included 38 PFAS to find an universal relationship between deF% and quantum chemical parameters and reveal the intrinsic factors that influence the deF%. However, validation, which plays an important role in judging the reliability of predictions, was performed to select the optimal QSAR model. In this study, squared correlation coefficient (R2), Fisher test (F value), T-test (t value), significance test (sig. value), standard deviation (SD value), root-mean-square error (RMSE value), variation inflation factors (VIF value), leave-one-out internal validation (q2), external validation, and Y- randomization tests were applied to evaluate the stability, reliability, and prediction ability of the developed QSAR models.

More information included in supplementary materials: https://doi.org/10.1021/acs.estlett.1c00403

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

defluorination ratio (deF%)

3.3. Comment on the endpoint:

N/A

3.4. Endpoint units:

%

3.5. Dependent variable:

$\log_{10}$(deF%)

3.6. Experimental protocol:

Therefore, in this study, we collected the deF% values of 38 PFAS with different headgroups and chain lengths from the previous studies of Bentel et al. and calculated their quantum chemical parameters through DFT by using Gaussian 0934 and Material Studio 7.0.35

3.7. Endpoint data quality and variability:

N/A

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSAR

Multiple linear regression model (MLR)

4.2. Explicit algorithm:

$\log_{10}(\text{deF\%}) = - 4.333 - 68.710\ E_{\text{LUMO}} - 5.873\ f(+)x$

4.3. Descriptors in the model:

E_{LUMO} energy of the lowest unoccupied molecular orbital

f(+)x The Fukui index wuth respect to nucleophiliv attack

4.4. Descriptor selection:

It was found that the toxicity profile of species tested was similar and had good relations with the fluorinated carbon chain length of the PFCs investigated. nC - fluorinated carbon-chain length is simple to calculate.

4.5. Algorithm and descriptor generation:

N/A

4.6. Software name and version for descriptor generation:

Gaussian 09 program

Material Studio 7.0

4.7. Chemicals/ Descriptors ratio:

7 chemicals/1 descriptor

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

All of the points lie within || < 3 and hi < h* (the AD area), indicating that the classified QSAR models are appropriate for predicting log10(deF%) of PFCA, PFdiCA, FTCA, and PFECA. As for the PFAS model, all of the points lie within the APD except potassium nonafluorobutanesulfonate (PFBS), but the small hi value of PFBS reveals that the PFAS model has good generalizability. Therefore, all of the developed QSAR models exhibit good performance for the training and test sets.

5.2. Method used to assess the applicability domain:

A Williams plot was exploited to visualize the applicability domain of the QSAR models, and this principle is essential for determining whether the model can be reliably applied in accordance with the guideline proposed by the Organization for Economic Co-operation and Development (OECD). Standardized residuals (σ) versus leverage (*hi*) approach.

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

Potassium nonafluorobutanesulfonate (PFBS), but the small hi value of PFBS reveals that the PFAS model has good generalizability

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

No

6.2. Availability information for the training set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula:No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

N/A

6.4. Data for the dependent variable (response) for the training set:

N/A

6.5. Other information about the training set:

Division into training set and test set in a ratio of approximately 4:1. There are no information which compounds are in the training set.

6.6. Pre-processing of data before modelling:

The original data deF% was expressed as logdeF%

6.7. Statistics for goodness-of-fit:

R^2=0.815

RMSE=0.385

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

Q^2=0.726

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

The results of Y randomization tests indicate that the models have no possibility of chance correlation and exhibit good robustness.

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

In this study, squared correlation coefficient (R2), Fisher test (F value), T-test (t value), significance test (sig. value), standard deviation (SD value), root-mean-square error (RMSE value), variation inflation factors (VIF value), leave-one-out internal validation (q2), external validation (Q2ext , Equation 7), and Y-randomization tests were applied to evaluate the stability, reliability, and prediction ability of the developed QSAR models.

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

Division into training set and test set in a ratio of approximately 4:1. There are no information which compounds are in the test set.

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

Q2ext=0.707

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

An external validation was performed for the developed model after splitting the compounds into a validation and training set. Despite access to data for all compounds used in the model, they were not added to the corresponding sets.

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

The model was developed by statistical approach.

8.2. A priori or a posteriori mechanistic interpretation:

As for the PFAS model, the structural characteristics for determining log10(deF%) are f(+)x and E_{LUMO}. E_{LUMO} is always associated with f(+)x and related to the ability to accept electrons; Because the Fukui indices have been widely used to predict the molecular degradation mechanism,the f(+) contour surfaces can be used to enhance understanding of the PFAS degradation pathway.

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Cheng, Z., Chen, Q., Liu, Z., Liu, J., Liu, Y., Liu, S., … Shen, Z. (2021). Interpretation of Reductive PFAS Defluorination with Quantum Chemical Parameters. Environmental Science & Technology Letters, 8(8), 645–650. https://doi.org/10.1021/acs.estlett.1c00403

3. Bentel, M. J.; Yu, Y.; Xu, L.; Kwon, H.; Li, Z.; Wong, B. M.; Men, Y.; Liu, J. Degradation of perfluoroalkyl ether carboxylic acids with hydrated electrons: Structure-reactivity relationships and environmental implications. Environ. Sci. Technol. 2020, 54 (4), 24892499.

3. Bentel, M. J.; Yu, Y.; Xu, L.; Li, Z.; Wong, B. M.; Men, Y.; Liu, J. Defluorination of per- and polyfluoroalkyl substances (PFASs) with hydrated electrons: structural dependence and implications to PFAS remediation and management. Environ. Sci. Technol. 2019, 53 (7), 37183728.

9.3. Supporting information:

N/A

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

CDFE. A Machine Learning Approach for Predicting Defluorination of Per- and Polyfluoroalkyl Substances (PFAS) for Their Efficient Treatment and Removal

1. QSAR identifier

1.1. QSAR identifier (title):

A Machine Learning Approach for Predicting Defluorination of Per- and Polyfluoroalkyl Substances (PFAS) for Their Efficient Treatment and Removal

1.2. Other related models:

N/A

1.3. Software coding the model:

The Random Forest calculations for the various PFAS bond dissociation energies were carried out with the RandomForestRegressor module within the scikit-learn Python package.

The LASSO Regression calculations for the various PFAS bond dissociation energies were carried out with the Lasso module within the scikit-learn Python package.

The FNN model was trained with the Keras module that runs on top of the Tensorflow software package (available as a standalone Python code)

2. General information

2.1. Date of QMRF:

09/11/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com

https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Akber Raza Department of Electrical & Computer Engineering,

2. Sharmistha Bardhan Department of Chemical & Environmental Engineering;

3. Lihua Xu Department of Chemical & Environmental Engineering

4. Sharma S. R. K. C. Yamijala Department of Chemical & Environmental Engineering, §Materials Science & Engineering Program, and

5. Chao Lian Department of Chemical & Environmental Engineering, §Materials Science &

Engineering Program, and Department of Physics & Astronomy, University of California, Riverside, Riverside, California 92521, United States

6. Hyuna Kwon Department of Chemical & Environmental Engineering

7. Bryan M. Wong Department of Electrical & Computer Engineering, Department of Chemical & Environmental Engineering, Materials Science & Engineering Program, and Department of Physics & Astronomy, University of California, Riverside, Riverside, California 92521, United States

2.6. Date of model development and/or publication:

Published: September 9, 2019

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

A Machine Learning Approach for Predicting Defluorination of Per- and Polyfluoroalkyl Substances (PFAS) for Their Efficient Treatment and Removal Akber Raza, Sharmistha Bardhan, Lihua Xu, Sharma S. R. K. C. Yamijala, Chao Lian, Hyuna Kwon, and Bryan M. Wong Environmental Science & Technology Letters 2019 6 (10), 624-629 DOI: 10.1021/acs.estlett.9b00476

2.8. Availability of information about the model:

To predict and understand C−F bond dissociation energies in various PFAS structures of environmental importance, a variety of machine learning techniques that include Random Forest, Least Absolute Shrinkage and Selection Operator (LASSO) Regression, Feed-forward Neural Network (FNN), and t-distributed Stochastic Neighbor Embedding (t- SNE) algorithms were utilized.

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

C-F bond dissociation energies

3.3. Comment on the endpoint:

N/A

3.4. Endpoint units:

kcal/mol

3.5. Dependent variable:

N/A

3.6. Experimental protocol:

N/A

3.7. Endpoint data quality and variability:

N/A

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

Random Forest, Least Absolute Shrinkage and Selection Operator Regression, Feed-forward Neural Networks models

4.2. Explicit algorithm:

Equations for Random Forest, Least Absolute Shrinkage and Selection Operator Regression, Feed-forward Neural Networks models are available in the Supporting Information at : https://pubs.acs.org/doi/10.1021/acs.estlett.9b00476

4.3. Descriptors in the model:

The four-sphere bond descriptor based on the PFAS structures

4.4. Descriptor selection:

The chosen descriptors should not be expensive to compute and, therefore, should satisfy the following four requirements for describing PFAS molecular structures: the desired chemical descriptor should (1) use a simple algorithm, (2) not rely on a quantum chemistry calculation to be carried out, (3) not require an optimized 3D geometry of the molecule, and (4) not explicitly use bond orders (i.e., single or double bonds).

4.5. Algorithm and descriptor generation:

The chemically intuitive bond descriptor scheme by Qu et al., which utilizes the Chemistry Development Kit libraries. Although the prior work by Qu et al. originally focused on organic molecules containing H, C, N, O, and S atoms, we have modified their open-source Java-based source code (available for download at http://www.bmwong-group.com/software) to include various C-F bond descriptors for the PFAS structures examined in this study.

4.6. Software name and version for descriptor generation:

http://www.bmwong-group.com/software

4.7. Chemicals/ Descriptors ratio:

N/A

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

N/A

5.2. Method used to assess the applicability domain:

N/A

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

N/A

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

No

6.2. Availability information for the training set:

CAS RN:No

Chemical Name: No

Smiles: No

Formula:No

INChI:No

MOL file:No

6.3. Data for each descriptor variable for the training set:

N/A

6.4. Data for the dependent variable (response) for the training set:

N/A

6.5. Other information about the training set:

The entire dataset (comprised of 564 unique C–F bond dissociation energies) was divided into 414 randomly chosen values that were used for training (~74% of the full dataset), with the remainder utilized for the test set.

6.6. Pre-processing of data before modelling:

N/A

6.7. Statistics for goodness-of-fit:

The FNN approach yields predictions that are in excellent agreement with the reference DFT data with an impressive (R^2= 0.93, MAD (kcal/mol)=0.70 (0.51), RMSE (kcal/mol)= 1.22 (0.89)).

The Random Forest approach (R^2= 0.79, MAD (kcal/mol)=2.42 (1.77), RMSE (kcal/mol)= 2.65 (1.94)).

The LASSO Regression approach (R^2= 0.89, MAD (kcal/mol)=1.96 (1.44), RMSE (kcal/mol)= 1.87 (1.37)).

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

N/A

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

N/A

7.2. Availability information for the external validation set:

CAS RN:No

Chemical Name: No

Smiles:No

Formula:No

INChI:No

MOL file:No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

The entire dataset (comprised of 564 unique C–F bond dissociation energies) was divided into 414 randomly chosen values that were used for training (~74% of the full dataset), with the remainder utilized for the test set.

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

N/A

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

N/A

8.2. A priori or a posteriori mechanistic interpretation:

It has been shown that described machine-learned model only requires knowledge of the simple chemical connectivity in a PFAS structure (i.e., neither a 3D geometry nor even a rough estimate of bond lengths/ orientations are required) to yield reliable results.

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Qu, X.; Latino, D. A. R. S.; Aires-de-Sousa, J. A Big Data Approach to the Ultra-Fast Prediction of DFT-Calculated Bond Energies. J. Cheminf. 2013, 5, 34.

9.3. Supporting information:

Supporting information available at: 10.1021/acs.estlett.9b00476

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

*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^2loo: 0.74; CCCcv: 0.85; RMSEcv: 0.76

Split by Ordered Response Model:

Q^2loo: 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^2LMO (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^2Yscr: 0.10

Split by Ordered Response Model: R^2Yscr: 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^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=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^2LOO = 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:

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(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>0.288 (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^{2}loo: 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^{2}LMO: 0.72

6.10.Robustness - Statistics obtained by Y-scrambling:

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

LC50(3). Per- and Polyfluoro Toxicity (LC_{50} Inhalation) Study in Mouse Using QSAR Modeling

1. QSAR identifier

1.1. QSAR identifier (title):

Per- and Polyfluoro Toxicity (LC_{50} Inhalation) Study in Mouse Using QSAR Modeling

1.2. Other related models:

Gramatica, P.; Cassani, S.; Chirico, N. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. J. Comput. Chem. 2014, 35, 1036–1044.

QDB archive DOI: 10.15152/QDB.177

Property M12.pLC50: Acute inhalation toxicity in mice as log(1/LC50)
[-log(mmol/m^3)]

1.3. Software coding the model:

Moby Digs software

Software used to performed GA-VSS (genetic algorithm-variable subset selection) method for the variable selection and multiple linear regression (MLR) analysis using the ordinary least squares regression (OLS) method

Todeschini, R.;Consonni, V.; Pavan, M.MOBYDIGS - Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows, Talete srl: Milan, Italy, 2002

2. General information

2.1. Date of QMRF:

24/02/2022

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Paola Gramatica University of Insubria (Varese - Italy)

paola.gramatica@uninsubria.it

www.qsar.it

2. Barun Bhhatarai University of Insubria (Varese - Italy)

barun.bhhatarai@uninsubria.it

2.6. Date of model development and/or publication:

Published in 2010

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

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 https://pubs.acs.org/doi/10.1021/tx900252h

2.8. Availability of information about the model:

More information about training/validation sets and descriptors used in model attached in supplementary materials: https://doi.org/10.1021/tx900252h

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

mouse (*mus musculus*)

3.2. Endpoint:

Human Health Effects: Acute Inhalation toxicity

3.3. Comment on the endpoint:

LC_{50} - lethal concentration 50

Standard measure of the toxicity of the surrounding medium that will kill 50% of the sample population of a specific test-animal in a specified period through exposure via inhalation (respiration).

3.4. Endpoint units:

The Inhalation data are expressed in log1/LC_{50} unit where LC_{50} was expressed in mmol

3.5. Dependent variable:

log 1/LC_{50}

3.6. Experimental protocol:

The experimental data on rat and mouse LC_{50} inhalation studies were collected from ChemID plus, which has compiled the data from various literature and patents.

3.7. Endpoint data quality and variability:

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:

OLS- ordinary least squares regression

$\log 1/LC_{50} = 4.21 - 1.27(\pm 0.31)\text{MlogP} + 1.43(\pm 0.46)\text{X3v} + 0.38(\pm 0.13)\text{F01[C-C]}$

$- 1.14(\pm 0.37)\text{H-048}$

4.3. Descriptors in the model:

1. MlogP - moriguchi octanol-water partition coefficient

2. F01[C-C] - frequency of C-C at topological distance 01 (2D frequency fingerprint)

3. X3v - the valence connectivity index chi-3

4. H-048 - descriptor representing H attached to the C2(sp3)/C1(sp2)/C0(sp) carbon belonging to atom centered fragments

4.4. Descriptor selection:

The reduced set of 537 descriptors for mouse inhalation were subjected to the variable selection method using the genetic algorithm (GA)

4.5. Algorithm and descriptor generation:

Multiple linear regression (Ordinary Least Square method) was applied to generate the model. The collected compounds were drawn by using the HYPERCHEM software, and they were minimized to their lowest energy conformation first by using molecular mechanics MM+ and then by using the semiempirical AM1 method. After that, to provide energy information and their electronic contribution quantum chemical descriptors, highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies were computed by the semiempirical molecular orbital program MOPAC (AM1 semiempirical for geometry optimization). From HOMO and LUMO values, HOMO-LUMO gap, absolute electronegativity [-1/2(HOMO+ LUMO)], and absolute hardness

[1/2(LUMO-HOMO)] were also calculated. The theoretical molecular descriptors for these compounds were then calculated using DRAGON software. 3224 molecular descriptors of different types were calculated.

4.6. Software name and version for descriptor generation:

1. DRAGON software

A software to calculate theoretical molecular descriptors

https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_ molecular_descriptor_calculations

2. MOPAC

AM1 semiempirical for geometry optimization Stewart Computational Chemistry. Stewart J.J.P., Colorado Springs, CO, USA

http://OpenMOPAC.net

3. HYPERCHEM

Software for molecular drawing and conformational energy optimization

www.hyper.com

4.7. Chemicals/ Descriptors ratio:

compound / descriptor ratio below 5

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The Williams plot was used to verify the presence of response outliers (i.e., compounds with cross-validated standardized residuals greater than 2.5 standard deviation units, (σ2.5) and chemicals very influential for their structure in determining model parameters (i.e., compounds with high leverage value (h) (h > h*, the critical value being h*) 3p'/n, where p' is the number of model variables plus one and n the number of the objects used to calculate the model). The leverage approach was applied for the definition of the structural chemical domain of each model.

5.2. Method used to assess the applicability domain:

Full model: The HAT cutoff values for the mouse model was at 0.267. The reason for which some chemicals were beyond the optimum leverage value could be (a) that they have some structural anomalies that were not well modeled by the selected descriptors of our models or (b) that these chemicals were too structurally particular and thus were extrapolated (with leverage higher than the h* value). To find the extreme high leverage compounds, an arbitrary new cutoff value of 0.5 was chosen for mouse set.

5.3. Software name and version for applicability domain assessment:

Mobydigs sofware

Todeschini, R.;Consonni,V.; Pavan, M.MOBYDIGS—Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows, Talete srl: Milan, Italy, 2002

5.4. Limits of applicability:

On the basis of this cutoff, there were 61 compounds which were out of the structural domain of the training (higher leverage) of the mouse model (75.6% coverage). Mostly linear long chain compounds were extremely out of AD. These long chain perfluorinated chemicals greater than 15 carbons were probably extrapolated as the longest compound in the training model were with 7 carbon atoms only. Thus, to find the extreme high leverage compounds, an arbitrary new cutoff value of 0.5 was chosen for mouse set, which gave 34 compounds as outliers.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN:Yes

Chemical Name:Yes

Smiles:No

Formula:No

INChI:No

MOL file:No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

In order to obtain compounds for external validation, the available set of compounds with experimental data were split into training sets and external prediction sets. Two different splitting methods were applied: (1) splitting realized by Kohonen map-artificial neural network or self-organizing maps (SOM 28.5%) using the package KOALA (2) splitting carried out by random selection through activity sampling (20%). Additionally full model was prepared.

6.6. Pre-processing of data before modelling:

The inhalation data are expressed in log 1/LC50 units where LC50 values expressed in gm/m3 or ppb were converted into mmol first before converting into the inverse log scale.

6.7. Statistics for goodness-of-fit:

a) Split by SOM: $R^2 = 82.99$, $RMSE_{TR} = 0.61$

b) Split by Random by Activity: $R^2 = 77.07$, $RMSE_{TR} = 0.70$

c) Full model: $R^2 = 79.83$, $RMSE_{TR} = 0.68$

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

a) Split by SOM: $Q^2_{LOO} = 78.09$

b) Split by Random by Activity: $Q^2_{LOO} = 71.73$

c) Full model: $Q^2_{LOO} = 76.31$

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

a) Split by SOM: $R^2Y_{SCR} = 10.32$

b) Split by Random by Activity: $R^2Y_{SCR} = 8.99$

c) Full model: $R^2Y_{SCR} = 7.05$

6.11. Robustness – Statistics obtained by bootstrap:

a) Split by SOM: $Q^2_{BOOT} = 75.46$

b) Split by Random by Activity: $Q^2_{BOOT} = 69.89$

c) Full model: $Q^2_{BOOT} = 75.38$

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Availability information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: No

Formula: No

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 (response) 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: SOM (n external validation set = 16) and random selection through activity sampling (n external validation set = 12)

7.6. Experimental design of test set:

In order to obtain compounds for external validation, the available set of compounds with experimental data were split into training sets and external prediction sets. Two different splitting methods were applied: (1) splitting realized by Kohonen map-artificial neural network or self-organizing maps (SOM 28.5%) using the package KOALA (2) splitting carried out by random selection through activity sampling (20%).

7.7. Predictivity – Statistics obtained by external validation:

a) Split by SOM:

$Q^2_{F1} = 71.62$, $Q^2_{F3} = 63.41$, $RMSE_{EXT} = 0.89$

b) Split by Random by Activity:

$Q^2_{F1} = 85.11$, $Q^2_{F3} = 78.86$, $RMSE_{EXT} = 0.67$

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

An external validation was performed for the developed model after splitting the compounds into a validation and training set. Despite access to data for all compounds used in the model, they were not added to the corresponding sets.

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:

$\log 1/LC_{50} = 4.21 - 1.27(\pm 0.31)MlogP + 1.43(\pm 0.46)X3v + 0.38(\pm 0.13)F01[C\text{-}C] - 1.14(\pm 0.37)H\text{-}048$
where:

MlogP - moriguchi octanol-water partition coefficient

F01[C-C] - frequency of C-C at topological distance 01 (2D frequency fingerprint)

X3v - the valence connectivity index chi-3

H-048 - descriptor representing H attached to the C2(sp3)/C1(sp2)/C0(sp) carbon belonging to atom centered fragments

The best descriptor is the Moriguchi octanol-water partition coefficient (MlogP), followed by the valence connectivity index chi-3 (3V), then frequency of C-C at topological distance 01 (F01[C-C]) belonging to the 2D frequency fingerprint, and finally (H-048), a descriptor representing H attached to the C2(sp3)/C1(sp2)/C0(sp) carbon belonging to atom centered fragments (the superscript represents the formal oxidation number of a carbon atom). The importance of observed descriptors was accessed by their standardized coefficient values, which are given adjacent to each descriptor symbol in parentheses. The more influential descriptor was MlogP (-0.81) with the negative coefficient, which shows that for fluorinated chemicals, the contribution of hydrophobicity, within this combination of descriptors, demonstrates a decreasing trend for mouse inhalation toxicity. Moreover, it is important to note that the correlation between the logarithm of mouse LC50 data with MlogP was very low (R = 0.020) and that the single model with MlogP was statistically invalid. The other two successive descriptors X3v (0.60) and F01 [C-C] (0.48) had a positive influence on mouse toxicity. The X3v (valence connectivity index) accounts for the presence of the heteroatom and double and triple bonds present in the compound. Increase in one or the other features increases the value of X3v in total. The F01[C-C] represents the total number of C-C bonds. As the alkyl chain length increases, C-C increases, giving high values to the longer chain. Thus, increasing chain length and increase in bond order, as well as the presence of the heteroatom contributes to increase in mouse inhalation toxicity. The least significant H-048 (-0.42) and its negative coefficient shows the decrease in toxicity value for compounds which have higher values of this descriptor. In the mouse inhalation data set, compounds with carbon atom in an alkyl chain connected to one or more electronegative atoms (O, F, Cl, and Br) have been encoded by H-048, which equals the oxidation number of a carbon

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Gramatica, P.; Cassani, S.; Chirico, N. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. J. Comput. Chem. 2014, 35, 1036–1044.

DOI: 10.15152/QDB.177

2. http://chem.sis.nlm.nih.gov/chemidplus/

9.3. Supporting information:

More information about training/validation sets and descriptors used in model attached in supplementary materials: https://doi.org/10.1021/tx900252h

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

LC50(4). Per- and Polyfluoro Toxicity (LC50 Inhalation) Study in Rat Using QSAR Modeling

1. QSAR identifier

1.1. QSAR identifier (title):

Per- and Polyfluoro Toxicity (LC50 Inhalation) Study in Rat Using QSAR Modeling

1.2. Other related models:

Gramatica, P.; Cassani, S.; Chirico, N. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. J. Comput. Chem. 2014, 35, 1036–1044.

QDB archive DOI: 10.15152/QDB.177

Property M13.pLC50: Acute inhalation toxicity in rat as log(1/LC50) [-log(mmol/m^3)]

1.3. Software coding the model:

Moby Digs software

Software used to performed GA-VSS (genetic algorithm-variable subset selection) method for the variable selection and multiple linear regression (MLR) analysis using the ordinary least squares regression (OLS) method

Todeschini, R.; Consonni, V.; Pavan, M.MOBYDIGS - Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows, Talete srl: Milan, Italy, 2002

2. General information

2.1. Date of QMRF:

24/02/2022

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Paola Gramatica University of Insubria (Varese - Italy)

paola.gramatica@uninsubria.it

www.qsar.it

2. Barun Bhhatarai University of Insubria (Varese - Italy)

barun.bhhatarai@uninsubria.it

2.6. Date of model development and/or publication:

Published in 2010

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

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 https://pubs.acs.org/doi/10.1021/tx900252h

2.8. Availability of information about the model:

More information about training/validation sets and descriptors used in model attached in supplementary materials: https://doi.org/10.1021/tx900252h

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

Rat (*Rattus*)

3.2. Endpoint:

Human Health Effects: Acute Inhalation toxicity

3.3. Comment on the endpoint:

LC_{50} - lethal concentration 50

Standard measure of the toxicity of the surrounding medium that will kill 50% of the sample population of a specific test-animal in a specified period through exposure via inhalation (respiration).

3.4. Endpoint units:

The Inhalation data are expressed in log1/LC_{50} unit where LC_{50} was expressed in mmol

3.5. Dependent variable:

log 1/LC_{50}

3.6. Experimental protocol:

The experimental data on rat and mouse LC50 inhalation studies were collected from ChemID plus, which has compiled the data from various literature and patents.

3.7. Endpoint data quality and variability:

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:

OLS- ordinary least squares regression

$\log 1/LC_{50} = -12.76 + 1.87(\pm 0.20)\mathrm{Jhetv} + 11.43(\pm 1.27)\mathrm{PCR} - 0.60(\pm 0.12)\mathrm{MlogP} - 1.41(\pm 0.40)\mathrm{B02[Cl\text{-}Cl]}$

4.3. Descriptors in the model:

1. Jhetv - the Balaban type index from van der Waals weighted distance matrix encoding the topological properties
2. PCR - ratio of multiple path count over path count
3. MlogP - Moriguchi octanol-water partition coefficient
4. B02[Cl-Cl] - 2D binary fingerprint descriptor which determines the presence/absence of Cl-Cl at a topological distance of 02

4.4. Descriptor selection:

The reduced set of 360 descriptors for rat inhalation were subjected to the variable selection method using the genetic algorithm (GA).

4.5. Algorithm and descriptor generation:

Multiple linear regression (Ordinary Least Square method) was applied to generate the model. The collected compounds were drawn by using the HYPERCHEM software, and they were minimized to their lowest energy conformation first by using molecular mechanics MM+ and then by using the semiempirical AM1 method. After that, to provide energy information and their electronic contribution quantum chemical descriptors, highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies were computed by the semiempirical molecular orbital program MOPAC (AM1 semiempirical for geometry optimization). From HOMO and LUMO values, HOMO-LUMO gap, absolute electronegativity [-1/2(HOMO+ LUMO)], and absolute hardness [1/2(LUMO-HOMO)] were also calculated. The theoretical molecular descriptors for these compounds were then calculated using DRAGON software. 3224 molecular descriptors of different types were calculated.

4.6. Software name and version for descriptor generation:

1. DRAGON software

Software for descriptor calculation

https://www.researchgate.net/publication/216208341_DRAGON_software_An_easy_approach_to_molecular_descriptor_calculations

www.talete.it.

2. MOPAC

AM1 semiempirical for geometry optimization Stewart Computational Chemistry. Stewart J.J.P., Colorado Springs, CO, USA

http://OpenMOPAC.net

3. HYPERCHEM

Software for molecular drawing and conformational energy optimization

www.hyper.com

4.7. Chemicals/ Descriptors ratio:

compound/descriptor ratio below 5

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The Williams plot was used to verify the presence of response outliers (i.e., compounds with cross-validated standardized residuals greater than 2.5 standard deviation units, (σ2.5) and chemicals very influential for their structure in determining model parameters (i.e., compounds with high leverage value (h) (h > h*, the critical value being h*) 3p'/n, where p' is the number of model variables plus one and n the number of the objects used to calculate the model). The leverage approach was applied for the definition of the structural chemical domain of each model.

5.2. Method used to assess the applicability domain:

The HAT cutoff values for the rat model was at 0.273. The reason for which some chemicals were beyond the optimum leverage value could be (a) that they have some structural anomalies that were not well modeled by the selected descriptors of our models or (b) that these chemicals were too structurally particular and thus were extrapolated (with leverage higher than the h* value). To find the extreme high leverage compounds, an arbitrary new cutoff value of 0.5 was chosen for rat set.

5.3. Software name and version for applicability domain assessment:

Mobydigs sofware

Todeschini,R.;Consonni,V.;Pavan,M.MOBYDIGS—Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows, Talete srl: Milan, Italy, 2002

5.4. Limits of applicability:

On the basis of this cutoff, there were 58 compounds which were out of the structural domain of the training (higher leverage) of the rat model (76.8% coverage). Mostly linear long chain compounds were extremely out of AD. These long chain perfluorinated chemicals greater than 15 carbons were probably extrapolated as the longest compound in the training model were with 7 carbon atoms only. Thus, to find the extreme high leverage compounds, an arbitrary new cutoff value of 0.5 was chosen for rat set, which gave 27 compounds as outliers.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN:Yes

Chemical Name:Yes

Smiles:No

Formula:No

INChI:No

MOL file:No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

In order to obtain compounds for external validation, the available set of compounds with experimental data were split into training sets and external prediction sets. Two different splitting methods were applied: (1) splitting realized by Kohonen map-artificial neural network or self-organizing maps (SOM 28.5%) using the package KOALA (2) splitting carried out by random selection through activity sampling (20%). Additionally full model was prepared.

6.6. Pre-processing of data before modelling:

The inhalation data are expressed in log 1/LC50 units where LC50 values expressed in gm/m3 or ppb were converted into mmol first before converting into the inverse log scale.

6.7. Statistics for goodness-of-fit:

a) Split by SOM: $R^2 = 78.36$, $RMSE_{TR} = 0.80$

b) Split by Random by Activity: $R^2 = 80.01$, $RMSE_{TR} = 0.77$

c) Full model: $R^2 = 78.14$, $RMSE_{TR} = 0.79$

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

a) Split by SOM: $Q^2_{LOO} = 72.99$

b) Split by Random by Activity: $Q^2_{LOO} = 80.01$

c) Full model: $Q^2_{LOO} = 78.14$

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

a) Split by SOM: $R^2Y_{SCR} = 8.75$

b) Split by Random by Activity: $R^2Y_{SCR} = 9.91$

c) Full model: $R^2Y_{SCR} = 7.64$

6.11. Robustness – Statistics obtained by bootstrap:

a) Split by SOM: $Q^2_{BOOT} = 71.95$

b) Split by Random by Activity: $Q^2_{BOOT} = 74.12$

c) Full model: $Q^2_{BOOT} = 75.26$

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Availability information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: No

Formula: No

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 (response) 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 training set(s) used for model development and prediction set(s) used later for external validation. Two different splitting techniques were applied: SOM (n exttternal validation set = 10) and random selection through activity sampling (n external validation set = 10).

7.6. Experimental design of test set:

In order to obtain compounds for external validation, the available set of compounds with experimental data were split into training sets and external prediction sets. Two different splitting methods were applied: (1) splitting realized by Kohonen map-artificial neural network or self-organizing maps (SOM) using the package KOALA (2) splitting carried out by random selection through activity sampling

7.7. Predictivity – Statistics obtained by external validation:

a) Split by SOM:

$Q^2_{F1} = 75.47$, $Q^2_{F3} = 79.67$, $RMSE_{EXT} = 0.77$

b) Split by Random by Activity:

$Q^2_{F1} = 66.70$, $Q^2_{F3} = 75.41$, $RMSE_{EXT} = 0.86$

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

An external validation was performed for the developed model after splitting the compounds into a validation and training set. Despite access to data for all compounds used in the model, they were not added to the corresponding sets.

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:

$\log 1/LC_{50} = -12.76 + 1.87(\pm0.20)Jhetv + 11.43(\pm1.27)PCR -0.60(\pm0.12)MlogP - 1.41(\pm0.40)B02[Cl\text{-}Cl]$

where:
Jhetv - the Balaban type index from van der Waals weighted distance matrix encoding the topological properties PCR - ratio of multiple path count over path count MlogP - Moriguchi octanol-water partition coefficient B02[Cl-Cl] - 2D binary fingerprint descriptor which determines the presence/absence of Cl-Cl at a topological distance of 02. The best descriptor was Jhetv, the Balaban type index from van der Waals weighted distance matrix encoding the topological properties, followed by PCR, the walk and path count descriptor which is the ratio of multiple path count over path count, then MlogP, Moriguchi octanol-water partition coefficient, and finally (B02[Cl-Cl], a 2D binary fingerprint descriptor which determines the presence/absence of Cl-Cl at a topological distance of 02. The two most influential 2D descriptors Jhetv (0.694) and PCR (0.692) 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 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. In the data set studied, PCR encoded 22 compounds with a single value of 1. The next important descriptor was MlogP (-0.38), 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. As mentioned above, the correlation between the logarithm of rat LC50 data with MlogP was very low (R =0.078), and the single model with MlogP was statistically invalid. The least significant descriptor was B02[Cl-Cl], (-0.24) 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. It was found only for 5 compounds in the modeled set of 52 compounds and for 11 of all the 250

total compounds in the set. Thus, it becomes clear that it was present in the model mainly as a fitting descriptor to encompass all of the freons (with fluorine and chlorine). Freons were included in the data set as their available LC50 activity could possibly help to increase the applicability domain of the model

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Gramatica, P.; Cassani, S.; Chirico, N. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. J. Comput. Chem. 2014, 35, 1036–1044.

DOI: 10.15152/QDB.177

2. http://chem.sis.nlm.nih.gov/chemidplus/

9.3. Supporting information:

More information about training/validation sets and descriptors used in model attached in supplementary materials: https://doi.org/10.1021/tx900252h

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

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

LD50(2). Oral LD_{50} toxicity modeling and prediction of per- and polyfluorinated chemicals on rat

1. QSAR identifier

1.1. QSAR identifier (title):

Oral LD_{50} toxicity modeling and prediction of per- and polyfluorinated chemicals on rat.

1.2. Other related models:

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

1.3. Software coding the model:

Moby Digs software. Todeschini R, Consonni V, Pavan M (2002) MOBY DIGS - Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows. Talete srl, Milan, Italy

2. General information

2.1. Date of QMRF:

09/11/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com

https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Paola Gramatica QSAR Research Unit in Environmental Chemistry and Ecotoxicology,

Department of Structural and Functional Biology (DBSF), University of Insubria, via J.H. Dunant 3, 21100 Varese, Italy e-mail: bhhataba@gmail.com

2. Barun Bhhatarai QSAR Research Unit in Environmental Chemistry and Ecotoxicology,

Department of Structural and Functional Biology (DBSF), University of Insubria, via J.H. Dunant 3, 21100 Varese, Italy e-mail: bhhataba@gmail.com

2.6. Date of model development and/or publication:

28 August 2010

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

Bhhatarai B, Gramatica P (2011) Oral LD50 toxicity modeling and prediction of per- and polyfluorinated chemicals on rat and mouse. Molecular Diversity 15:467–476. https://doi.org/10.1007/s11030-010-9268-z

2.8. Availability of information about the model:

Model was published in a scientific journal and is available online : DOI10.1007/s11030-010-9268-z

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

Rat

3.2. Endpoint:

Acute oral toxicity

3.3. Comment on the endpoint:

Lethal dose 50 (LD_{50}). 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. LD_{50} is measured in micrograms (or milligrams) of the material per kilogram of the test-animal's body weight.

3.4. Endpoint units:

mmol/kg

3.5. Dependent variable:

pLD_{50}

3.6. Experimental protocol:

N/A

3.7. Endpoint data quality and variability:

The experimental data on rat LD50 oral toxicities were collected from ChemID plus.

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:

Full model equation: pLD50 = -2.277 + 0.041 (±0.003) D/Dr09 +2.943 (±0.580) MATS1e +8.838 (±1.712) E1u +1.166 (±0.211) H8m

4.3. Descriptors in the model:

1. D/Dr09 - distance/detour ring index of order 9

2. MATS1e - 2D Moran autocorrelation descriptor

3. E1u H8m - 3D descriptor corresponding to first component accessibility directional WHIM index

4. H8m - 3D GETWAY descriptor representing H autocorrelation of lag 8 (weighted by atomic mass).

4.4. Descriptor selection:

Constant values and descriptors, found to be pairwise correlated by greater than 95%, were excluded, minimizing the redundant information. The reduced set of more than 600 descriptors in each set was subjected to variable selection method using Genetic Algorithm (GA).

4.5. Algorithm and descriptor generation:

The 0D-3D theoretical molecular descriptors were then calculated from the 3D structures using DRAGON software. The collected compounds (input for descriptor calculation) were drawn in 2D using SMILES and minimized to their lowest energy conformation using HYPERCHEM first by using molecular mechanics MM+ and then by using semiempirical AM1 methods.

4.6. Software name and version for descriptor generation:

DRAGON Software - Todeschini R, Consonni V, Mauri A, Pavan M (2007) DRAGON v.5 Talete srl. Milan, Italy. http://www.talete.it. Hyperchem, 7.03, (2002) Hypercube Inc., Florida USA. http:// www.hyper.com

4.7. Chemicals/ Descriptors ratio:

50 PFCs/ 4 descriptors

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The structural AD study of rat oral data was studied on 376 common compounds with or without data. The PCA plot of chemicals, represented by their cumulative toxicity data on rat and mouse, helps to identify the most common toxic compounds within the AD of the QSAR models. The compounds beyond the arbitrary cutoff of PC1 = 1.25 were predicted to be most toxic, based on the descriptors observed in the QSAR models developed from the available experimental data. The 48 compounds, most of them linear PFCs, were found beyond the cutoff, including fluorinated benzimidazole and dinitro-benzenamine. Out of them 30 long-chain PFCs (Supporting Information, Fig. S3), which are of major interest in the CADASTER project, are proposed for further experimental design on toxicity studies.

5.2. Method used to assess the applicability domain:

The structural applicability domain of the model was assessed by the leverage approach (Willliams plot), 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.

5.3. Software name and version for applicability domain assessment:

Mobydigs sofware

Todeschini,R.;Consonni,V.;Pavan,M.MOBYDIGS—Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows, Talete srl: Milan, Italy, 2002

5.4. Limits of applicability:

Response outliers were checked by using Williams plot and for the response split set CAS 376-18-1 was found beyond ±3 and for SOM split set no response outlier was found. Instead, two structural outliers that were beyond the average HAT cut-off of 0.42 were present. They were CAS 311-89-7 of prediction set and CAS 335-76-2 of training set.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN:Yes

Chemical Name:No

Smiles:No

Formula:No

INChI:No

MOL file:No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

The experimental dataset was split a priori into training and prediction set by using (a) Kohonen map-artificial neural network or self-organizing maps (SOM) and (b) by Random selection through activity sampling.

Train compounds:

Split by SOM 28.0 % Model: 36

Random split by activity 26% Model:37

Full model: 50

6.6. Pre-processing of data before modelling:

The reported mg/kg data were converted into the mmol/kg and expressed in inverse log unit for modeling which are represented as pLD_{50}.

6.7. Statistics for goodness-of-fit:

Split by SOM 28.0 % Model:

R^2=85.49, $RMSE_{TR}$=0.41

Random split by activity 26% Model:

R^2=90.69, $RMSE_{TR}$ =0,36

Full model:

R^2=88.28, $RMSE_{TR}$ =0,42

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

Split by SOM 28.0 % Model:

$Q^2{}_{LOO}$=80,32

Random split by activity 26% Model:

$Q^2{}_{LOO}$ =87,46

Full model:

$Q^2{}_{LOO}$ =85.5

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

Split by SOM 28.0 % Model:

$R^2{}_{YS}$=12.21

Random split by activity 26% Model:

R^2_{YS} =11.28

Full model:

R^2_{YS} =8.03

6.11. Robustness – Statistics obtained by bootstrap:

Split by SOM 28.0 % Model:

Q^2_{BOOT}=70,9

Random split by activity 26% Model:

Q^2_{BOOT} =85,59

Full model:

Q^2_{BOOT} =82,2

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Availability information for the external validation set:

CAS RN: Yes

Chemical Name: No

Smiles:No

Formula:No

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 (response) for the external validation set:

All

7.5. Other information about the external validation set:

The experimental dataset was split a priori into training and prediction set by using (a) Kohonen map-artificial neural network or self-organizing maps (SOM) and (b) by Random selection through activity sampling.

Test compounds:

Split by SOM 28.0 % Model: 14

Random split by activity 26% Model: 13

7.6. Experimental design of test set:

The principal components, based on the molecular descriptors, were used to develop a Kohonen map, and the clustering capability of SOM was used for selection of a meaningful training and representative prediction set. This splitting is used to capture the difference in the structure of the molecule and to guarantee that the chemical domains in the two sets are not too dissimilar. A parallel splitting was carried out by random selection through activity sampling, by orderingthe chemicals according to their descending experimental values, selecting the most and the least active in the training set, and taking every nth chemical from the set to be used as a prediction set. This splitting is guided by the response of the molecule. These two splittings were used to develop and identify a statistically robust final model with common set of descriptors, based on both the training sets, which will be used to check the model predictivity of both the prediction sets. The prediction set was thus used only after model development for external validation.

7.7. Predictivity – Statistics obtained by external validation:

Split by SOM 28.0 % Model:

Q^2_{F1}=91.07, Q^2_{F3}=81.38, $RMSE_{ext}$=0.46

Random split by activity 26% Model:

Q^2_{F1}=80.69, Q^2_{F3}=75.05, $RMSE_{ext}$=0.59

Full model:

$RMSE_{ext}$=0.47 (cv)

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

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:

For the rat LD50 oral model, topological 2D descriptor D/Dr09 (0.765), distance/detour ring index of order 9, was found to be most important, followed by 2D Moran autocorrelation descriptor MATS1e (0.331) representing electronegativity. The third most important descriptor was E1u (0.316) a 3D descriptor corresponding to first component accessibility directional WHIM index (unweighted), representing information regarding the quantity of unfilled space per projected atom and ultimately by H8m (0.294) 3D GETWAY descriptor representing H autocorrelation of lag 8 (weighted by atomic mass). The increase in molecular mass increases the value of H8m descriptor, and an increase in polycyclic rings increases the value of D/Dr09.

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Hyperchem, 7.03, (2002) Hypercube Inc., Florida USA. http:// www.hyper.com

2. Todeschini R, Consonni V, Pavan M (2002) MOBY DIGS - Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows. Talete srl, Milan, Italy

3. ChemID Plus. http://chem.sis.nlm.nih.gov/chemidplus. Accessed 26 Apr 2010

4. Gramatica, P., Cassani, S. & Chirico, N. QSARINSchem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. J Comput Chem 35, 1036–1044 (2014).

9.3. Supporting information:

DOI 10.1007/s11030-010-9268-z

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

LD50(3). Oral LD_{50} toxicity modeling and prediction of per- and polyfluorinated chemicals on mouse

1. QSAR identifier

1.1. QSAR identifier (title):

Oral LD_{50} toxicity modeling and prediction of per- and polyfluorinated chemicals on mouse.

1.2. Other related models:

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

1.3. Software coding the model:

Moby Digs software. Todeschini R, Consonni V, Pavan M (2002) MOBY DIGS - Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows. Talete srl, Milan, Italy

2. General information

2.1. Date of QMRF:

09/11/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com

https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Paola Gramatica QSAR Research Unit in Environmental Chemistry and Ecotoxicology,

Department of Structural and Functional Biology (DBSF), University of Insubria, via J.H. Dunant 3, 21100 Varese, Italy e-mail: bhhataba@gmail.com

2. Barun Bhhatarai QSAR Research Unit in Environmental Chemistry and Ecotoxicology,

Department of Structural and Functional Biology (DBSF), University of Insubria, via J.H. Dunant 3, 21100 Varese, Italy e-mail: bhhataba@gmail.com

2.6. Date of model development and/or publication:

28 August 2010

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

Bhhatarai B, Gramatica P (2011) Oral LD50 toxicity modeling and prediction of per- and polyfluorinated chemicals on rat and mouse. Molecular Diversity 15:467–476. https://doi.org/10.1007/s11030-010-9268-z

2.8. Availability of information about the model:

Model was published in a scientific journal and is available online : DOI10.1007/s11030-010-9268-z

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

Mouse

3.2. Endpoint:

Acute oral toxicity

3.3. Comment on the endpoint:

Lethal dose 50 (LD_{50}). 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. LD_{50} is measured in micrograms (or milligrams) of the material per kilogram of the test-animal's body weight.

3.4. Endpoint units:

mmol/kg

3.5. Dependent variable:

pLD_{50}

3.6. Experimental protocol:

N/A

3.7. Endpoint data quality and variability:

The experimental data on mouse LD_{50} oral toxicities were collected from ChemID plus.

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:

Full model equation: $pLD_{50} = 4.543 - 2.450\ (\pm 0.312)\ \mathrm{HATS2u} + 1.362\ (\pm 0.203)\ \mathrm{B09}\ [\mathrm{C–O}] - 0.142\ (\pm 0.032)\ \mathrm{F01}\ [\mathrm{C–O}] - 0.486\ (\pm 0.174)\ \mathrm{B04}\ [\mathrm{C–F}]$

4.3. Descriptors in the model:

1.HATS2u - 3D GETAWAY descriptor encoding leverage-weighted autocorrelation of lag 2/ unweighted

2. B09[C–O] - binary fingerprint descriptor

3. F01[C–O] - frequency finger-print descriptor

4. B04[C–F] - binary fingerprint descriptor

4.4. Descriptor selection:

Constant values and descriptors, found to be pairwise correlated by greater than 95%, were excluded, minimizing the redundant information. The reduced set of more than 600 descriptors in each set was subjected to variable selection method using Genetic Algorithm (GA).

4.5. Algorithm and descriptor generation:

The 0D-3D theoretical molecular descriptors were then calculated from the 3D structures using DRAGON software. The collected compounds (input for descriptor calculation) were drawn in 2D using SMILES and minimized to their lowest energy conformation using HYPERCHEM first by using molecular mechanics MM+ and then by using semiempirical AM1 methods.

4.6. Software name and version for descriptor generation:

DRAGON Software - Todeschini R, Consonni V, Mauri A, Pavan M (2007) DRAGON v.5 Talete srl. Milan, Italy. http://www.talete.it. Hyperchem, 7.03, (2002) Hypercube Inc., Florida USA. http:// www.hyper.com

4.7. Chemicals/ Descriptors ratio:

58 PFCs/ 4 descriptors

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The structural AD study of rat oral data was studied on 376 common compounds with or without data. The PCA plot of chemicals, represented by their cumulative toxicity data on rat and mouse, helps to identify the most common toxic compounds within the AD of the QSAR models. The compounds beyond the arbitrary cutoff of PC1 = 1.25 were predicted to be most toxic, based on the descriptors observed in the QSAR models developed from the available experimental data. The 48 compounds, most of them linear PFCs, were found beyond the cutoff, including fluorinated benzimidazole and dinitro-benzenamine. Out of them 30 long-chain PFCs (Supporting Information, Fig. S3), which are of major interest in the CADASTER project, are proposed for further experimental design on toxicity studies.

5.2. Method used to assess the applicability domain:

The structural applicability domain of the model was assessed by the leverage approach (Willliams plot), 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.

5.3. Software name and version for applicability domain assessment:

Mobydigs sofware

Todeschini,R.;Consonni,V.;Pavan,M.MOBYDIGS—Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows, Talete srl: Milan, Italy, 2002

5.4. Limits of applicability:

Response outliers were checked for both splitting by using Williams plot, and none of them were found outside for a standardized residuals greater than three standard deviation units, ±3. None of the compounds were observed as structural outliers (high leverage compounds).

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN: Yes

Chemical Name:No

Smiles:No

Formula:No

INChI:No

MOL file:No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

The experimental dataset was split a priori into training and prediction set by using (a) Kohonen map-artificial neural network or self-organizing maps (SOM) and (b) by Random selection through activity sampling.

Train compounds:

Split by SOM 28.0 % Model: 36

Random split by activity 26% Model: 39

Full model: 58

6.6. Pre-processing of data before modelling:

The reported mg/kg data were converted into the mmol/kg and expressed in inverse log unit for modeling which are represented as pLD_{50}.

6.7. Statistics for goodness-of-fit:

Split by SOM 28.0 % Model:

R^2=82.77, $RMSE_{TR}$=0.32

Random split by activity 26% Model:

R^2=81.44, $RMSE_{TR}$ =0.35

Full model:

R^2=75.93, $RMSE_{TR}$ =0.39

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

Split by SOM 28.0 % Model:

Q^2_{LOO}=75.68

Random split by activity 26% Model:

Q^2_{LOO} =75.45

Full model:

Q^2_{LOO} =71.89

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

Split by SOM 28.0 % Model:

R^2_{YS}=11.54

Random split by activity 26% Model:

R^2_{YS} =12.23

Full model:

R^2_{YS} =6.97

6.11. Robustness – Statistics obtained by bootstrap:

Split by SOM 28.0 % Model:

Q^2_{BOOT}=74.15

Random split by activity 26% Model:

Q^2_{BOOT} =74.05

Full model:

Q^2_{BOOT} =67.91

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Availability information for the external validation set:

CAS RN:Yes

Chemical Name: No

Smiles:No

Formula:No

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 (response) for the external validation set:

All

7.5. Other information about the external validation set:

The experimental dataset was split a priori into training and prediction set by using (a) Kohonen map-artificial neural network or self-organizing maps (SOM) and (b) by Random selection through activity sampling.

Test compounds:

Split by SOM 28.0 % Model: 22

Random split by activity 26% Model: 19

7.6. Experimental design of test set:

The principal components, based on the molecular descriptors, were used to develop a Kohonen map, and the clustering capability of SOM was used for selection of a meaningful training and representative prediction set. This splitting is used to capture the difference in the structure of the molecule and to guarantee that the chemical domains in the two sets are not too dissimilar. A parallel splitting was carried out by random selection through activity sampling, by orderingthe chemicals according to their descending experimental values, selecting the most and the least active in the training set, and taking every nth chemical from the set to be used as a prediction set. This splitting is guided by the response of the molecule. These two splittings were used to develop and identify a statistically robust final model with common set of descriptors, based on both the training sets, which will be used to check the model predictivity of both the prediction sets. The prediction set was thus used only after model development for external validation.

7.7. Predictivity – Statistics obtained by external validation:

Split by SOM 28.0 % Model:

Q^2_{F1}=66.57, Q^2_{F3}=55.58, $RMSE_{ext}$=0.51

Random split by activity 26% Model:

Q^2_{F1}=62.97, Q^2_{F3}=64.49, $RMSE_{ext}$=0.48

Full model:

$RMSE_{ext}$=0.42 (cv)

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

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 most important descriptor for mouse LD50 oral model was HATS2u (−0.589)—a 3D GETAWAY descriptor encoding leverage-weighted autocorrelation of lag 2/ unweighted—and it appeared negative. This descriptor encodes information on the effective position of substituents and fragments in the molecular space [51]. In addition, 2D binary and frequency finger-print descriptor representing the presence and the frequency of a defined atom pair at a given topological distance ‘n’ were observed. These descriptors give information about the constitution of compounds of which the descriptor representing the presence of [C–O] at n = 9 viz. B09[C–O] (0.478) was positive while its frequency at n = 1 viz. F01[C–O] (−0.303) was negative. Moreover, the descriptor encoding the presence of atom pair [C–F] at n = 4 viz. B04[C–F] (−0.198) was negative for mouse LD50 activity.

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Hyperchem, 7.03, (2002) Hypercube Inc., Florida USA. http:// www.hyper.com

2. Todeschini R, Consonni V, Pavan M (2002) MOBY DIGS - Software for multilinear regression analysis and variable subset selection by genetic algorithm. Ver. 1.2 for Windows. Talete srl, Milan, Italy

3. ChemID Plus. http://chem.sis.nlm.nih.gov/chemidplus. Accessed 26 Apr 2010

4. Gramatica, P., Cassani, S. & Chirico, N. QSARINSchem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. J Comput Chem 35, 1036–1044 (2014).

9.3. Supporting information:

DOI 10.1007/s11030-010-9268-z

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

IC50(1). In silico modeling of perfluoroalkyl substances mixture toxicity in amphibian fibroblast cell line

1. QSAR identifier

1.1. QSAR identifier (title):

In silico modeling of perfluoroalkyl substances mixture toxicity in amphibian fibroblast cell line.

1.2. Other related models:

N/A

1.3. Software coding the model:

Models from each of the divisions were developed using two stepwise multiple linear regression (MLR) and genetic algorithms (GA) based MLR using Stepwise MLR v1.2, Genetic Algorithm v4.1 and MLR Plus Validation GUI 1.3, respectively.

R 3.4.3 and Rstudio with the package "drc" were used to create dose-response curves and obtain the equation for the resultant sigmoidal curve of the model.

2. General information

2.1. Date of QMRF:

09/11/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com

https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

Gary Hoover; Supratik Kar; Samuel Guffey; Jerzy Leszczynski; Maria S. Sepúlveda Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN 47907, USA; Interdisciplinary Center for Nanotoxicity, Department of Chemistry, Physics and Atmospheric Sciences, Jackson State University, Jackson, MS 39217, USA mssepulv@purdue.edu https://doi.org/10.1016/j.chemosphere.2019.05.065

2.6. Date of model development and/or publication:

Published in 2019

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

G. Hoover, S. Kar, S. Guffey, J. Leszczynski, M. S. Sepúlveda, 2019. In vitro and in silico modeling of perfluoroalkyl substances mixture toxicity in an amphibian fibroblast cell line; Chemospere 233 (2019) 25-33 https://www.sciencedirect.com/science/article/abs/pii/S0045653519309634?via%3Dihub

2.8. Availability of information about the model:

The cytotoxicity of PFAS singly and in binary mixtures has been examined using an amphibian fibroblast cell line. Second, we used this experimental data to develop quantitative structure-activity relationship (QSAR) models for single and binary mixtures. The cytotoxicity of four common PFAS: perfluorooctane sulfonate (PFOS); perfluorooctanoic acid (PFOA); perfluorohexane sulfonate (PFHxS); and perfluorohexanoic acid (PFHxA). Using this data, QSAR modeling was used for predicting the toxicity of 24 single and 1380 binary mixtures (theoretically generated).

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

Amphibian - Xenopus tropicalis (Western clawed frog)

3.2. Endpoint:

Cytotoxicity-Cell Growth Assay (IC50)

3.3. Comment on the endpoint:

IC50 - Inhibitory concentrations (ppm) resulting in 50% reduction of absorbance at 560 nm following MTT assay (IC50) with 95% confidence interval (CI) limits and concentration range tested (ppm) for each chemical.

3.4. Endpoint units:

ppm

3.5. Dependent variable:

pIC50

3.6. Experimental protocol:

Single chemical exposures were initially tested in a dose range-finding fashion, testing from very low concentrations (high dilutions) up to the maximum (stock) concentration (for range of concentrations tested) within a single plate. Subsequent plates were run on a more narrowly focused scale to determine a more precise IC50. Binary mixtures were tested by exposing all cells to a single concentration of one chemical while varying the concentration of a second chemical; the concentrations used were based on fractions of the single-chemical IC50.

3.7. Endpoint data quality and variability:

The cytotoxic effects of four common PFAS (PFOS, PFOA, PFHxS, and PFHxA) was examined, alone and in binary mixtures, on an amphibian fibroblast cell line. Cells exposed to single chemicals responded with sigmoidal dose-response curves, which then were used to determine inhibition curves and dictate the concentrations used in binary mixture experiments.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSAR

Multiple linear regression model (MLR)

4.2. Explicit algorithm:

$pIC50 = 1.85 (\pm 0.39) + 1.33 (\pm 0.11) piPC06$

4.3. Descriptors in the model:

piPC06: molecular multiple path count of order 6

4.4. Descriptor selection:

Genetic algorithm (GA) was used for descriptor selection.

4.5. Algorithm and descriptor generation:

Single PFAS were drawn, and followed by optimization employing the semi-empirical AM1 method into their lowest energy conformation. Output files were then used to compute a total of 195 1D and 2D descriptors (constitutional indices, topological indices, walk-path counts, connectivity indices, functional group counts, Atom-type E-state indices and Atom-centered fragments).

4.6. Software name and version for descriptor generation:

HyperChem 8.07 (Hypercube Inc., USA)

DRAGON 6 (2011)

4.7. Chemicals/ Descriptors ratio:

12 compounds in the training set / 1 descriptor

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

All single PFAS and mixtures tested fell within the AD created by the model. QSAR Model was also used to predict the external dataset of 24 single and 1380 binary mixture of PFAS (1404 total data points).

5.2. Method used to assess the applicability domain:

Applicability domain (AD) test calculated using both a standardization technique (STD-AD) and a Euclidean distance approach (ED-AD). Test compounds passed the AD (applicability domain) test by both algorithm and prediction of all test compounds are 'Good' according to Prediction Reliability Indicator' tool.

5.3. Software name and version for applicability domain assessment:

Prediction Reliability Indicator' tool.

5.4. Limits of applicability:

Out of the 1404 datapoints from the external dataset, 120 PFAS (three single and 117 binary mixtures) fell outside of the AD.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN: No

Chemical Name: Yes

Smiles: No

Formula:No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

3 single PFAS and 9 binary mixtures

6.6. Pre-processing of data before modelling:

The originally generated IC_{50} values were converted into a negative logarithmic molar scale (pIC_{50}) for the purposes of modeling

6.7. Statistics for goodness-of-fit:

n_{Train} =12

SEE=0.09

R^2= 0.94

R^2a=0.93

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

Q^2_{LOO}=0.88

$r^2_{m(LOO)}$ =0.85

$\Delta r^2_{m(LOO)}$=0.09

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

R^2=0.1

Q^2=-0.36

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Availability information for the external validation set:

CAS RN:No

Chemical Name: No

Smiles:No

Formula:No

INChI:No

MOL file:No

7.3. Data for each descriptor variable for the external validation set:

Yes

7.4. Data for the dependent variable (response) for the external validation set:

Yes

7.5. Other information about the external validation set:

1 single PFAS and 5 binary mixtures

7.6. Experimental design of test set:

Chemicals were ordered according to their increasing endpoint value, and one out of every three chemicals was put in the prediction set (always including chemical with the largest and lowest value in the training set).

7.7. Predictivity – Statistics obtained by external validation:

n_{Test}=6

Q^2_{F1}=0.9

Q^2_{F2}=0.9

$r^2_{m(test)}$=0.81

$\Delta r^2_{m(test)}$ =0.08

RMSEp=0.12

7.8. Predictivity – Assessment of the external validation set:

Range of response for prediction set (n=6) compounds:
Exp. pIC50 (molar): 2.16 - 3.34 (range of corresponding training set: 2.15 - 3.37
Range of modeling descriptors for prediction set (n=6) compounds:
piPC06: 3.117 - 3.866 (range of corresponding training set: 2.994 - 3.951)

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

The model was developed by statistical approach.

8.2. A priori or a posteriori mechanistic interpretation:

Mechanistic interpretations : The molecular multiple path counts (piPCk) are defined as path counts weighted by the bond order:

$piPCk = kp_{ij}*w_{ij}$

where kp_{ij} denotes a path of length k from vertex i to vertex j and wij is the path weight given by the product of the bond orders in the path:

$w_{ij} = k_{b=1}*b$

A piPC06 suggests that carbon chain length of six or more is a crucial component of PFAS toxicity which is supported by the in vitro experimental data as well as by our previous study and that of others (Kar et al., 2017; Kleszczyski et al., 2007; Liu et al., 2014). Specifically, the toxicity of PFAS (with the same functional group) increases as chain length increases from six to eight. Among the modeled compounds, PFHxA was the least toxic and had the lowest descriptor value (2.944), compared to PFOS which was the most toxic and produced the highest descriptor value (3.951). For PFAS with the same chain length, sulfonates had a higher descriptor value compared to acids (i.e., PFHxS was more toxic than PFHxA).

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Kar, S., Sepúlveda, M.S., Roy, K., Leszczynski, J., 2017. Endocrine-disrupting activity of per- and polyfluoroalkyl substances: exploring combined approaches of ligand and structure based modeling. Chemosphere 184, 514e523. https://doi.org/10.1016/j.chemosphere.2017.06.024.

2. Hoover, G.M., Chislock, M.C., Tornabene, B.J., Guffey, S.C., Choi, Y.J., De Perre, C., Hoverman, J.T., Lee, L.S., Sepúlveda, M.S., 2017. Uptake and depuration of four per/polyfluoroalkyl substances (PFASs) in northern leopard frog Rana pipiens tadpoles. Environ. Sci. Technol. Lett. 4 (10), 399e403. https://doi.org/10.1021/ acs.estlett.7b00339.

3. Roy, K., Kar, S., 2014. The rm2 metrics and regression through origin approach: reliable and useful validation tools for predictive QSAR models (Commentary on 'Is regression through origin useful in external validation of QSAR models?'). Eur. J. Pharm. Sci. 62, 111e114. https://doi.org/10.1016/j.ejps.2014.05.019.

4. Kleszczy nski, K., Gardzielewski, P., Mulkiewicz, E., Stepnowski, P., Sklandanowski, A.C., 2007. Analysis of structureecytotoxicity in vitro relationship (SAR) for perfluorinated carboxylic acids. Toxicol. In Vitro 21 (6), 1206e1211. https://doi.org/10.1016/j.tiv.2007.04.020.

5. Liu, C., Chang, V.W.C., Gin, K.Y.H., Nguyen, V.T., 2014. Genotoxicity of perfluorinated chemicals (PFCs) to the green mussel (Pernaviridis). Sci. Total Environ. 487, 117e122. https://doi.org/10.1016/j.scitotenv.2014.04.017.

9.3. Supporting information:

N/A

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

IC50(2). Risk assessment and developmental toxicity prediction on zebrafish embryos based on weighted descriptors approach

1. QSAR identifier

1.1. QSAR identifier (title):

Risk assessment and developmental toxicity prediction on zebrafish embryos based on weighted descriptors approach.

1.2. Other related models:

N/A

1.3. Software coding the model:

DTC Lab Software Tools

http://teqip.jdvu.ac.in/QSAR_Tools/

2. General information

2.1. Date of QMRF:

20/12/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com

https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Supratik Kar, Interdisciplinary Nanotoxicity Center, Department of Chemistry, Physics and Atmospheric Sciences, Jackson State University, Jackson, MS, USA.

2. Shinjita Ghosh, School of Public Health, Jackson State University, Jackson, MS, USA.

3. Jerzy Leszczynski, Interdisciplinary Nanotoxicity Center, Department of Chemistry, Physics and Atmospheric Sciences, Jackson State University, Jackson, MS, USA jerzy@icnanotox.org.

2.6. Date of model development and/or publication:

Published in 2018

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

S. Kar; S. Ghosh; J. Leszczynski, 2019. Single or mixture halogenated chemicals? Risk assessment and developmental toxicity prediction on zebrafish embryos based on weighted descriptors approach; Chemosphere 210 (2018) 588-596. https://www.sciencedirect.com/science/article/pii/S0045653518313092?via%3Dihub

2.8. Availability of information about the model:

9 halogenated compounds consist of 5 single (TBBPA (E1), TDCPP (E2), PFOA (E3), DOPO (E4), and PFBA (E5)) and 4 tertiary mixtures tested on adult zebrafish (AB wild-type) were considered from literature (Godfrey et al., 2017). More information about model attached in supplementary materials: https://doi.org/10.1016/j.chemosphere.2018.07.051

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

Zebrafish -Danio rerio

3.2. Endpoint:

Developmental toxicity IC_{50}.

3.3. Comment on the endpoint:

Inhibitory concentrations (ppm) resulting in 50% reduction of absorbance at 560 nm following MTT assay (IC50) with 95% confidence interval (CI) limits and concentration range tested (ppm) for each chemical.

3.4. Endpoint units:

-

3.5. Dependent variable:

pIC50 (molar scale)

3.6. Experimental protocol:

Experimental toxicity data was considered from literature (Godfrey et al., 2017). The toxicity of the following halogenated compounds was tested: TBBPA (CAS No. 79-94-7), TDCPP (CAS No. 13674-87-8), PFOA (CAS No. 335-67-1), DOPO (CAS No. 35948-25-5), and PFBA (CAS No. 206-786-3). A stock solution of each chemical was prepared by dissolving the appropriate amount of powdered chemical in 1 L Reverse Osmosis (RO) water containing 12.5 mL Replenish (Seachem Laboratories Inc.) and adjusted pH to neutral ($7^{7.5}$). Because of the low solubility of some of the compounds tested the following solvents were used: 0.1 M NaOH (<0.003% and <0.06%) for TBBPA and DOPO, respectively; and DMSO (<0.0001%) for TDCPP.

3.7. Endpoint data quality and variability:

Experimental toxicity data for a set of 9 halogenated compounds consist of 5 single (TBBPA (E1), TDCPP (E2), PFOA (E3), DOPO (E4), and PFBA (E5)) and 4 tertiary mixtures tested on adult zebrafish (AB wild-type) were considered from literature (Godfrey et al., 2017).

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSTR model

Multiple Linear Regression approach (MLR)

4.2. Explicit algorithm:

Equation (1) developed employing training set generated by activity sorted response division:

pIC_{50} (molar) = -6.677(±1,869) + 11,398(±2,221) x (PJI2)

Equation (2) developed employing training set generated with Euclidean distance based approach:

pIC_{50} (molar) = -0,323(±1,133) + 0,658(±0,203) x (2Xv)

Equation (3) generated employing training set generated by Kennard-Stone based method:

pIC_{50} (molar) = -0,825(±0,765) + 0,345(±0,065) x (0Xv)

Equation (4) computed employing training set generated with modified-k-medoid approach:

pIC_{50} (molar) = -2,196(±1,653) + 0,437(±0,127) x (0Xv)

4.3. Descriptors in the model:

PJI2: a topological index, defines 2D Petitjean shape index that has a positive contribution towards toxicity.

2Xv a connectivity index, signifies valence connectivity index of order 2 and has a positive contribution towards toxicity.

0Xv: a connectivity index, defines valence connectivity index of order 0. This descriptor has a positive contribution to toxicity.

4.4. Descriptor selection:

Descriptors are selected through genetic function approximation (GFA) and final models are developed by multiple linear regressions (MLR) approach. For both steps, open access tools Genetic Algorithm v4.1 and MLR Plus Validation GUI 1.3 have been employed, respectively (DTC Lab Software Tools).

4.5. Algorithm and descriptor generation:

A total of 95 1D and 2D descriptors were generated from constitutional indices, connectivity indices, topological indices, functional group counts and Atom-centered fragments for reproducibility purpose by any user. Once the descriptors are calculated for single molecules, we have used weighted descriptor generation approach for mixtures (Muratov et al., 2012).

4.6. Software name and version for descriptor generation:

HyperChem 8.07 (Hypercube Inc., USA)

software package used to draw structures

www.talete.mi.it

DRAGON 6 software

Version 6.0, 2011;

http://www.talete.mi.it/

4.7. Chemicals/ Descriptors ratio:

6 chemicals in the training set/1 descriptor

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The applicability domain (AD) was also verified employing two different approaches: a) the Euclidean distance approach (ED-AD) and b) the standardization technique (STD-AD) (Roy et al., 2015).

5.2. Method used to assess the applicability domain:

The applicability domain studies are performed to check influential observations under training set and outliers for the test set with two different methods: Euclidean-based AD and Standardization-based AD. In case of the STD-based AD, the output will provide the 'outlier' details. On the contrary, for ED-based approach, any test compounds with more than 1 normalized mean distance denoted as 'outlier'. The predictions of outliers are not reliable. In that case, the best option is to go for consensus prediction for higher reliability.

5.3. Software name and version for applicability domain assessment:

Euclidean Applicability domain 1.0

http://dtclab.webs.com/software-tools

Applicability Domain v1.0

http://dtclab.webs.com/software-tools

5.4. Limits of applicability:

The cumulative AD study suggested that no compound is an outlier for equations (1) and (2), and one compound each outlier for equation (3) (Compound E1) and 4 (Compound E5).

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN: No

Chemical Name: Yes

Smiles: No

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

N/A

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

To improve accuracy of the applied computational technique we have divided the dataset employing four different approaches maintaining each molecule present in the test set at least once. Thus, modeling will consider the effect of all molecules when consensus model will come into the prediction. The applied techniques are: 1. Activity sorted response, 2. Euclidean distance approach, 3. Kennard stone technique and 4. modified k-Medoid clustering. First three divisions were performed employing ‘Dataset division GUI 1.2’ and the last one was done by ‘Modified K-Medoid GUI 1.3’ tool (DTC Lab Software Tools).

Training set includes always 6 compounds, but which compounds they are depends on Equation selected.

6.6. Pre-processing of data before modelling:

Data was converted to a log scale for modeling.

6.7. Statistics for goodness-of-fit:

(1) $R^2 = 0.87$; MAE95%(Train)= 0.59

(2) $R^2 = 0.94$; MAE95%(Train)= 0.78

(3) $R^2 = 0.87$; MAE95%(Train)= 0.42

(4) $R^2 = 0.94$; MAE95%(Train)= 0.75

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

(1) Q^2 = 0.65 , MAE95%(Test)= 0.70

(2) Q^2 = 0.54 , MAE95%(Test)= 0.43

(3) Q^2 = 0.66 , MAE95%(Test)= 0.64

(4) Q^2 = 0.56 , MAE95%(Test)= 0.38

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

(1) ${}^{C}R^2_P$=0.79

(2) ${}^{C}R^2_P$=0.64

(3) ${}^{C}R^2_P$=0.79

(4) ${}^{C}R^2_P$=0.66

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: Yes

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

All

7.5. Other information about the external validation set:

Always 3 compounds are in the training set

7.6. Experimental design of test set:

* Test compounds for Eqn. (1),

+ Test compounds for Eqn. (2),

\$ Test compounds for Eqn. (3)

^ Test compounds for Eqn. (4).

E1 $^{*\$}$, E2, E3 * , E4 $^{+}$, E5 $^{\wedge}$, E6 $^{\wedge}$, E7 $^{*\$}$, E8 $^{+\wedge .}$,E9 $^{+\$}$

7.7. Predictivity – Statistics obtained by external validation:

(1) R^2pred = 0.79;

(2) R^2pred = 0.64;

(3) R^2pred = 0.79;

(4) R^2pred = 0.66;

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

An external validation was performed for the developed model after splitting the compounds into a validation and training set. Despite access to data for all compounds used in the model, they were not added to the corresponding sets.

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

N/A

8.2. A priori or a posteriori mechanistic interpretation:

Equation developed employing training set generated by activity sorted response division: pIC50(molar) = -6.677(±1,869) + 11,398(±2,221) x (PJI2) PJI2: a topological index, defines 2D Petitjean shape index that has a positive contribution towards toxicity. Equation developed employing training set generated with Euclidean distance based approach: pIC50(molar) = -0,323(±1,133) + 0,658(±0,203) x (2Xv) 2Xv: a connectivity index, signifies valence connectivity index of order 2 and has a positive contribution towards toxicity. Equation generated employing training set generated by Kennard stone based method: pIC50(molar) = -0,825(±0,765) + 0,345(±0,065) x (0Xv) 0Xv: a connectivity index, defines valence connectivity index of order 0. This descriptor has a positive contribution to toxicity. Equation computed employing training set generated with modified-k-medoid approach: pIC50(molar) = -2,196(±1,653) + 0,437(±0,127) x (0Xv) Though each equation consists of single descriptor, but consensus model considers the greater chemical and structural space due to the contribution of each descriptor for its prediction quality. equation (1) consists of PJI2, a topological index, defines 2D Petitjean shape index that has a positive contribution towards toxicity. equation (2) modeled with 2cv, a connectivity index, signifies valence connectivity index of order 2 and has a positive contribution towards toxicity. equations (3) and (4) model with 0cv, a connectivity index, defines valence connectivity index of order 0. This descriptor has a positive contribution to toxicity. All three descriptors have positive contributions which suggest that the higher value of these descriptors increases the toxicity of a compound and vice versa.

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. S. Kar; S. Ghosh; J. Leszczynski, 2019. Single or mixture halogenated chemicals? Risk assessment and developmental toxicity prediction on zebrafish embryos based on weighted descriptors approach; Chemosphere 210 (2018) 588-596. https://www.sciencedirect.com/science/article/pii/S0045653518313092?via%3Dihub

2. Godfrey, A., Abdel-moneim, A., & Sepúlveda, M. S. (2017). Acute mixture toxicity of halogenated chemicals and their next generation counterparts on zebrafish embryos. Chemosphere, 181, 710–712. https://doi.org/10.1016/j.chemosphere.2017.04.146

3. Roy, K., Kar, S., Ambure, P., 2015. On a simple approach for determining applicability domain of QSAR models. Chemometr. Intell. Lab. Syst. 145, 22e29.

4. Muratov, E.N., Varlamova, E.V., Artemenko, A.G., Polishchuk, P.G., Kuz'min, V.E., 2012. Existing and developing approaches for QSAR analysis of mixtures. Mol. Inform 31, 202e221.

9.3. Supporting information:

Supporting information available at: https://doi.org/10.1016/j.chemosphere.2018.07.051

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

IC50(3). Predictive model for competitive binding of poly- and perfluorinated compounds to the thyroid hormone transport protein transthyretin.

1. QSAR identifier

1.1. QSAR identifier (title):

Predictive model for competitive binding of poly- and perfluorinated compounds to the thyroid hormone transport protein transthyretin.

1.2. Other related models:

N/A

1.3. Software coding the model:

The PCA and PLS calculations were performed on a PC using SIMCA-P version 11.0 software 2005 (Umetrics AB, Umea, Sweden).

https://www.sartorius.com/en/products/process-analytical-technology/data-analytics-software/mvdasoftware/simca

Molecular Operating Environment (Chemical Computing Group)

https://www.chemcomp.com/Products.htmL

2. General information

2.1. Date of QMRF:

09/11/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com

https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Jana M. Weiss Institute for Environmental Studies, Department of Chemistry and Biology, VU University, 1081 HV Amsterdam, The Netherlands; and †Department of Chemistry, Umea° University, 90187 Umea° , Sweden

2. Patrik L. Andersson Institute for Environmental Studies, Department of Chemistry and Biology, VU University, 1081 HV Amsterdam, The Netherlands; and †Department of Chemistry, Umea° University, 90187 Umea° , Sweden

3. Marja H. Lamoree Institute for Environmental Studies, Department of Chemistry and Biology, VU University, 1081 HV Amsterdam, The Netherlands; and †Department of Chemistry, Umea° University, 90187 Umea° , Sweden

4. Pim E. G. Leonards Institute for Environmental Studies, Department of Chemistry and Biology, VU University, 1081 HV Amsterdam, The Netherlands; and †Department of Chemistry, Umea° University, 90187 Umea° , Sweden

5. Stefan P. J. van Leeuwen Institute for Environmental Studies, Department of Chemistry and Biology, VU University, 1081 HV Amsterdam, The Netherlands; and †Department of Chemistry, Umea° University, 90187 Umea° , Sweden

6. Timo Hamers Institute for Environmental Studies, Department of Chemistry and Biology, VU University, 1081 HV Amsterdam, The Netherlands; and †Department of Chemistry, University, 90187 Umea° , Sweden

2.6. Date of model development and/or publication:

March 17, 2009

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

Jana M. Weiss, Patrik L. Andersson, Marja H. Lamoree, Pim E. G. Leonards, Stefan P. J. van Leeuwen, Timo Hamers, Competitive Binding of Poly- and Perfluorinated Compounds to the Thyroid Hormone Transport Protein Transthyretin, Toxicological Sciences, Volume 109, Issue 2, June 2009, Pages 206–216

2.8. Availability of information about the model:

TTR binding potency was searched using partial least squares regression (PLS). In brief, the systematic variation is searched in an X-matrix, here including the calculated chemical descriptors, and correlated with the systematic variation in the Y-matrix (here IC50 values) including the observed response. Latent variables are formed of the matrices and their relationship is maximized through a weight vector. Competitive binding of PFCs to TTR, as observed for human TTR in the present study, may explain altered thyroid hormone levels described for PFC-exposed rats and monkeys.

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

human thyroid hormone transport protein transthyretin (TTR)

3.2. Endpoint:

TTR binding potencies (IC50)

3.3. Comment on the endpoint:

The resulting PLS model developed on the 56 chemical descriptors and the IC50 values for all 15 active PFCs.

3.4. Endpoint units:

nM

3.5. Dependent variable:

Log IC_{50}

3.6. Experimental protocol:

For all compounds, 10mM stock solutions were prepared in dimethyl sulfoxide (Acros Organics, Geel, Belgium; 99.9%), except for perfluoroundecanoic acid, perfluorododecanoic acid (PFDoA), and perfluorotetradecanoic acid (PFTdA),where 5mM stock solutions were prepared due to limited solubility at higher concentrations. The TTR-binding assay has been described in detail earlier (Hamers et al., 2006; Lans et al., 1993) and is modified from a method first described by Somack et al. (1982).

3.7. Endpoint data quality and variability:

N/A

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

partial least squares regression (PLS) model

4.2. Explicit algorithm:

N/A

4.3. Descriptors in the model:

56 descriptors were used comprising, e.g., molecular volume, surface area, weight, diameter, radius, number of hydrogen bond donors and acceptors, topological indices such as Zagreb, Balaban, Kier, and Hall chi indices, and partial charges. The partial charge descriptors include 16 parameters reflecting total and fractional positive and negative partial charges and partial charges related to total surface area and hydrophobic and polar surface area. HPLC retention times as described above were also included in the set of chemical descriptors.

4.4. Descriptor selection:

In order to study the chemical diversity of the PFCs, as described by the 56 calculated molecular descriptors, principal component analysis (PCA) was applied (Jackson, 1991). The most significant chemical property that separates the chemicals in the first dimension (t1) of the PCA is the molecular size, here reflected by, e.g., molecular volume and area, connectivity indices, retention times, and total positive partial charge. The acids and telomer alcohols are split from the sulfonamides and sulfonates based merely on their partial charge characteristics (second dimension [t2]).

4.5. Algorithm and descriptor generation :

N/A

4.6. Software name and version for descriptor generation:

Molecular Operating Environment (Chemical Computing Group)

https://www.chemcomp.com/Products.htm

4.7. Chemicals/ Descriptors ratio:

15 chemicals with data/ 56 descriptors

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

The chemical domain of the most active substances covers average size PFCs and mainly the acids.

5.2. Method used to assess the applicability domain:

In order to discuss outlying properties of the compounds' TTR binding potency, DModY was used. DModY is the distance of an observation to the model in the Y space (Eriksson et al., 1999).

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

The PLS model diagnosed PFBS as an outlier according to its distance to the X- and the Y-block (DModX and -Y), i.e., according to its chemical properties and its TTR binding potency.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: No

Formula: Yes

INChI: No

MOL file:No

6.3. Data for each descriptor variable for the training set:

Yes

6.4. Data for the dependent variable (response) for the training set:

Yes

6.5. Other information about the training set:

N/A

6.6. Pre-processing of data before modelling:

Data were preprocessed by auto-scaling and mean centering. The responses were normalized by the logarithm prior modeling

6.7. Statistics for goodness-of-fit:

N/A

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

N/A

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN:No

Chemical Name: No

Smiles:No

Formula:No

INChI:No

MOL file:No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

N/A

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

N/A

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

An alternative mechanism of binding to TTR has been shown, where the hydroxyl group of the potent TTR-binding compounds can be directed to the outer binding site, emphasizing that the presence of hydroxyl groups is not essential to be able to bind to TTR (Baures et al., 1998; Gosh et al., 2000). This ‘‘reversed’’ orientation was the exclusive binding mode for penta- and tri-bromophenols (Gosh et al., 2000).

8.2. A priori or a posteriori mechanistic interpretation:

The present study clearly demonstrates that TTR binding of PFCs can also explain the mechanism of PFC retention in human blood. The hydroxyl group of the natural TTR ligand T4, which is accompanied by two adjacent iodine atoms, is oriented to the inner part of the

TTR-binding site (Gosh et al., 2000). An alternative mechanism of binding to TTR has been shown, where the hydroxyl group of the potent TTR-binding compounds can be directed to the outer binding site, emphasizing that the presence of hydroxyl groups is not essential to be able to bind to TTR (Baures et al., 1998; Gosh et al., 2000). This ‘‘reversed’’ orientation was the exclusive binding mode for penta- and tri-bromophenols (Gosh et al., 2000). For PFCs with a carbon chain length of four to eight, TTR binding potencies were significantly higher for compounds containing a sulfonate functional group than for those containing a carboxylic acid functional group (paired t-test of mean IC50 value in each functional group category, p=0.01).

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Gosh, M., Meerts, I. A. T. M., Cook, A., Bergman, A° ., Brouwer, A., and Johnson, L. N. (2000). Structure of human transthyretin complexed with bromophenols: New mode of binding. Acta Crystallogr. Sect. D Biol. Crystallogr. D56, 1085–1095.

2. Baures, P. W., Peterson, S. A., and Kelly, J. W. (1998). Discovering transthyretin amyloid fibril inhibitors by limited screening. Bioorg. Med. Chem. 6, 1389–1401.

3. Jackson, J. E. (1991). In A User’s Guide to Principal Components. John Wiley, New York.

4. Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal component analysis. Technometrics 20, 397–405.

9.3. Supporting information:

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

EC50(1). QSAR model for the aquatic toxicity of seven PFCs to lettuce (L. sativa) seeds

1. QSAR identifier

1.1. QSAR identifier (title):

QSAR model for the aquatic toxicity of seven PFCs to lettuce (L. sativa) seeds.

1.2. Other related models:

N/A

1.3. Software coding the model:

N/A

2. General information

2.1. Date of QMRF:

09/11/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com

https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Guanghui Ding College of Environmental Science and Engineering, Dalian Maritime University, Dalian 116026, People’s Republic of China; Laboratory for Ecological Risk Assessment, National Institute of Public Health and the Environment, 3720 BA Bilthoven, The Netherlands

2. Marja Wouterse Laboratory for Ecological Risk Assessment, National Institute of Public Health and the Environment, 3720 BA Bilthoven, The Netherlands

3. Rob Baerselman Laboratory for Ecological Risk Assessment, National Institute of Public Health and the Environment, 3720 BA Bilthoven, The Netherlands

4. Willie J. G. M. Peijnenburg Laboratory for Ecological Risk Assessment, National Institute of Public Health and the Environment, 3720 BA Bilthoven, The Netherlands; Institute of Environmental Sciences, Faculty of Science, Leiden University, Leiden, The Netherlands

2.6. Date of model development and/or publication:

Published online: 31 May 2011

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

Ding G, Wouterse M, Baerselman R, Peijnenburg WJ. Toxicity of polyfluorinated and perfluorinated compounds to lettuce (Lactuca sativa) and green algae (Pseudokirchneriella subcapitata). Arch Environ Contam Toxicol. 2012 Jan;62(1):49-55. Epub 2011 May 31. PMID: 21626016. doi: 10.1007/s00244-011-9684-9

2.8. Availability of information about the model:

The aquatic toxicity of seven PFCs to lettuce (L. sativa) seeds was first tested to provide more toxicity information for PFCs. Subsequently, QSAR model for the toxicity were developed and toxicity profiles across the species investigated. QSAR model was developed based on linear PFCAs and 5H 4:1 FTOH.

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

lettuce (Lactuca sativa)

3.2. Endpoint:

root elongation (EC_{50})

3.3. Comment on the endpoint:

Root elongation is the result of the enlargement of new cells constantly being formed by cell divisions in the general region of the apical meristem. Subsequent differentiation of these cellular components of the root gives rise. to the complex tissue systems which comprise the mature structure.

3.4. Endpoint units:

mM

3.5. Dependent variable:

logEC$_{50}$

3.6. Experimental protocol:

Root Elongation Test on Lettuce (L. sativa): US EPA OPPTS 850.4200 and OECD test guideline 208

3.7. Endpoint data quality and variability:

The EC$_{50}$s and NOECs derived for each of the PFCs studied are listed in publication. The endpoint of interest, root elongation of lettuce (L. sativa) after a total exposure time of 5 days, includes possible effects of PFCs on germination. Limited by the solubility and aggregation of PFUnA and PFDoA, the EC$_{50}$ values for these two chemicals could not be experimentally determined. In addition, enhanced aggregation and possible binding to filter article might affect their bioavailability and subsequently the extent of toxicity. The measured EC50values of the additional PFCs studied were in the range of 0.266 to 4.186 mM.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSAR

Simple linear regression model

4.2. Explicit algorithm:

Log EC$_{50}$= -0.170 (± 0.0409) x nC + 1.197 (± 0.271)

4.3. Descriptors in the model:

nC - fluorinated carbon-chain length

4.4. Descriptor selection:

It was found that the toxicity profile of species tested was similar and had good relations with the fluorinated carbon chain length of the PFCs investigated. nC - fluorinated carbon-chain length is simple to calculate.

4.5. Algorithm and descriptor generation:

N/A

4.6. Software name and version for descriptor generation:

N/A

4.7. Chemicals/ Descriptors ratio:

7 chemicals/1 descriptor

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

QSAR model was developed based on linear PFCAs and 5H 4:1 FTOH, so it is suitable to predict the EC50values of PFCAs with similar structure. PFUnA and PFDoA are two similar PFCAs with longer fluorinated carbon-chain lengths, so the QSAR was used to extrapolate the EC_{50}.

5.2. Method used to assess the applicability domain:

N/A

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

QSAR model is inadequate for branched PFCAs and perfluoroalkyl sulfonate acids. Compared with the EC_{50} values, the NOEC values also decreased with nC, except that the NOEC of 5H 4:1 FTOH was an outlier compared with the NOECs of the PFCAs. The NOEC of 5H 4:1 FTOH was found to be just 0.1 mM. This is similar to the values obtained for PFNA and PFDA, both being PFCs with a greater fluorinated carbon-chain length.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: No

Formula:No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable (response) for the training set:

All

6.5. Other information about the training set:

N/A

6.6. Pre-processing of data before modelling:

The original data EC_{50} (95% CL [mM]) was expressed as logEC50 (mM).

6.7. Statistics for goodness-of-fit:

R^2=0.853

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

N/A

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

N/A

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

N/A

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

The model was developed by statistical approach.

8.2. A priori or a posteriori mechanistic interpretation:

The toxic effects on lettuce seeds were found to be similar in a relative sense and were shown to have a good relationship with the fluorinated carbon-chain length. The toxicity of these chemicals increased with increasing fluorinated carbon chain length.

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Verweij W, Durand, AM, Maas JL, Van der Grinten E (2009) PAM test: acute effects on photosynthesis in algae. In protocols belonging to the report ‘‘Toxicity measurements in concentrated water samples.’’ National Institute for Public Health and the Environment report 607013011/2009, pp 41–51.

2. Organisation for Economic Co-operation and Development (2006) OECD guidelines for the testing of chemicals / section 2: effects on biotic systems test no. 208: Terrestrial plant test: Seedling emergence and seedling growth test.

3. United States Environmental Protection Agency (1996) Ecological effects test guidelines (OPPTS 850.4200): seed germination/root elongation toxicity test. EPA 712-C-96-154. Washington, DC.

9.3. Supporting information:

N/A

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

EC50(2). QSAR model for the aquatic toxicity of seven PFCs to green algae (Pseudokirchneriella subcapitata)

1. QSAR identifier

1.1. QSAR identifier (title):

QSAR model for the aquatic toxicity of seven PFCs to green algae (Pseudokirchneriella subcapitata).

1.2. Other related models:

N/A

1.3. Software coding the model:

N/A

2. General information

2.1. Date of QMRF:

09/11/2021

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com

https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Guanghui Ding College of Environmental Science and Engineering, Dalian Maritime University, Dalian 116026, People’s Republic of China; Laboratory for Ecological Risk

Assessment, National Institute of Public Health and the Environment, 3720 BA Bilthoven, The Netherlands

2. Marja Wouterse Laboratory for Ecological Risk Assessment, National Institute of Public Health and the Environment, 3720 BA Bilthoven, The Netherlands

3. Rob Baerselman Laboratory for Ecological Risk Assessment, National Institute of Public Health and the Environment, 3720 BA Bilthoven, The Netherlands

4. Willie J. G. M. Peijnenburg Laboratory for Ecological Risk Assessment, National Institute of Public Health and the Environment, 3720 BA Bilthoven, The Netherlands; Institute of Environmental Sciences, Faculty of Science, Leiden University, Leiden, The Netherlands

2.6. Date of model development and/or publication:

Published online: 31 May 2011

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

Ding G, Wouterse M, Baerselman R, Peijnenburg WJ. Toxicity of polyfluorinated and perfluorinated compounds to lettuce (Lactuca sativa) and green algae (Pseudokirchneriella subcapitata). Arch Environ Contam Toxicol. 2012 Jan;62(1):49-55. Epub 2011 May 31. PMID: 21626016. doi: 10.1007/s00244-011-9684-9

2.8. Availability of information about the model:

The aquatic toxicity of seven PFCs to lettuce (L. sativa) seeds was first tested to provide more toxicity information for PFCs. Subsequently, QSAR model for the toxicity were developed and toxicity profiles across the species investigated. QSAR model was developed based on linear PFCAs and 5H 4:1 FTOH.

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

green algae (P. subcapitata)

3.2. Endpoint:

Inhibition of photosynthetic efficiency (EC50)

3.3. Comment on the endpoint:

Photosynthesis efficiency can be calculated from the recorded signals.

3.4. Endpoint units:

mM

3.5. Dependent variable:

logEC50

3.6. Experimental protocol:

PAM Test: Acute Effects on Photosynthesis in Algae developed in the RIVM

(Verweij et al. 2009).

3.7. Endpoint data quality and variability:

The EC50s and NOECs derived for each of the PFCs studied are listed in publication. The endpoint of interest, green algae after a total exposure time of 5 days, includes possible effects of PFCs. Limited by the solubility and aggregation of PFUnA and PFDoA, the EC50 values for these two chemicals could not be experimentally determined. In addition, enhanced aggregation and possible binding to filter article might affect their bioavailability and subsequently the extent of toxicity.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSAR

Simple linear regression model

4.2. Explicit algorithm:

Log EC50= -0.156 (± 0.0122) x nC + 1.313 (± 0.0881)

4.3. Descriptors in the model:

nC - fluorinated carbon-chain length

4.4. Descriptor selection:

It was found that the toxicity profile of species tested was similar and had good relations with the fluorinated carbon chain length of the PFCs investigated. nC - fluorinated carbon-chain length is simple to calculate.

4.5. Algorithm and descriptor generation:

N/A

4.6. Software name and version for descriptor generation:

N/A

4.7. Chemicals/ Descriptors ratio:

7 chemicals/1 descriptor

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

QSAR model was developed based on linear PFCAs and 5H 4:1 FTOH, so it is suitable to predict the EC50values of PFCAs with similar structure. PFUnA and PFDoA are two similar PFCAs with longer fluorinated carbon-chain lengths, so the QSAR was used to extrapolate the EC50.

5.2. Method used to assess the applicability domain:

N/A

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

QSAR model is inadequate for branched PFCAs and perfluoroalkyl sulfonate acids.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: No

Formula:No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

Yes

6.4. Data for the dependent variable (response) for the training set:

Yes

6.5. Other information about the training set:

N/A

6.6. Pre-processing of data before modelling:

The original data EC50 (95% CL [mM]) was expressed as logEC50 (mM).

6.7. Statistics for goodness-of-fit:

R^2=0.988

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

N/A

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

No

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

N/A

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

N/A

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

The model was developed by statistical approach.

8.2. A priori or a posteriori mechanistic interpretation:

The toxic effects on green algae were found to be similar in a relative sense and were shown to have a good relationship with the fluorinated carbon-chain length. The toxicity of these chemicals increased with increasing fluorinated carbon chain length.

8.3. Other information about the mechanistic interpretation:

N/A

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Verweij W, Durand, AM, Maas JL, Van der Grinten E (2009) PAM test: acute effects on photosynthesis in algae. In protocols belonging to the report ''Toxicity measurements in concentrated water samples.'' National Institute for Public Health and the Environment report 607013011/2009, pp 41–51.

2. Organisation for Economic Co-operation and Development (2006) OECD guidelines for the testing of chemicals / section 2: effects on biotic systems test no. 208: Terrestrial plant test: Seedling emergence and seedling growth test.

3. United States Environmental Protection Agency (1996) Ecological effects test guidelines (OPPTS 850.4200): seed germination/root elongation toxicity test. EPA 712-C-96-154. Washington, DC.

9.3. Supporting information:

N/A

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

EC50(3). A QSAR Model Prediction of the acute toxicity for Per- and Polyfluoroalkyl Substances (*Pseudokirchneriella subcapitata*)

1. QSAR identifier

1.1. QSAR identifier (title):

A QSAR Model Prediction of the acute toxicity for Per- and Polyfluoroalkyl Substances (*Pseudokirchneriella subcapitata*)

1.2. Other related models:

N/A

1.3. Software coding the model:

SIMCA software

https://www.sartorius.com/

2. General information

2.1. Date of QMRF:

10/01/2022

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Jianghong Shi State Environmental Protection Key Laboratory of Integrated Surface Water- Groundwater Pollution Control, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China shijh@sustech.edu.cn

2. Hiu Ge State Environmental Protection Key Laboratory of Integrated Surface Water-Groundwater Pollution Control, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China geh@sustech.edu.cn

2.6. Date of model development and/or publication:

Published in 2021

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

Zhang, J.;Zhang, M.; Tao, H.; Qi, G.; Guo, W.; Ge, H.; Shi, J. A QSAR–ICE–SSD Model Prediction of the PNECs for Per- and Polyfluoroalkyl Substances and Their Ecological Risks in an Area of Electroplating Factories. Molecules 2021,26,6574.
https://doi.org/ 10.3390/molecules26216574

2.8. Availability of information about the model:

More information about model attached in supplementary materials:
https://doi.org/10.3390/molecules26216574

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

Pseudokirchneriella subcapitata

3.2. Endpoint:

Acute toxicity - *Pseudokirchneriella subcapitata*

3.3. Comment on the endpoint:

EC_{50} - median effect concentration

3.4. Endpoint units:

mg/L

3.5. Dependent variable:

$\log EC_{50}$

3.6. Experimental protocol:

The test methods were in accordance with standard test methods (e.g., the methods of the Organization for Economic Cooperation and Development).

3.7. Endpoint data quality and variability:

The acute toxicity data were mainly obtained from the US EPA ECOTOX database, the literature, and relevant government documents.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSAR - Stepwise multiple linear regression model

4.2. Explicit algorithm:

$\log EC50 = -\log K_{ow} \times 8.82 + TE \times 47.8 + \log E_{LUMO} \times 1.47 - E_{CCR} \times 39.7 + 50.3$

4.3. Descriptors in the model:

1. K_{ow} - Octanol-water partition coefficient

2. TE - total energy [EV]

3. E_{LUMO} - Lowest unoccupied molecule orbital energy [EV]

4. E_{CCR} - core-core repulsion energy [EV]

4.4. Descriptor selection:

The stepwise regression method in SPSS was used to establish the multiple regression statistical models between the logarithmic values of the toxicity data (i.e., log LC(EC)50) and the molecular descriptors (including their logarithmic values)).

4.5. Algorithm and descriptor generation:

The molecular energy was optimized using the GAMESS Interface method in ChemBio3D (https://perkinelmerinformatics.com/). A total of 12 semi-empirical molecular descriptors were then calculated using the AM1 method in MOPAC 2016 (http://openmopac.net/) and the K_{ow} values were calculated using EPI Suite software (https://www.epa.gov/)

4.6. Software name and version for descriptor generation:

1. MOPAC 2016

total of 12 semi-empirical molecular descriptors were then calculated using the AM1 method

http://openmopac.net/

2. EPI Suite software

were used to calculate K_{ow} values

https://www.epa.gov/

4.7. Chemicals/ Descriptors ratio:

14 chemicals / 4 descriptors = 3.5

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

N/A

5.2. Method used to assess the applicability domain:

N/A

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

ICE models were used after verifying the application domain.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability 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 (response) for the training set:

All

6.5. Other information about the training set:

Training set consist of 14 compounds

6.6. Pre-processing of data before modelling:

The original EC_{50} data were expressed in log unit, also descriptors values were expressed in logarithmic values

6.7. Statistics for goodness-of-fit:

R^2= 0.770

r^2=0.742

* R^2 – coefficient of determination of the multiple regression, r^2 – conventional correlation coefficient or non-validation correlation coefficient

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

Q^2_{LOO} = 0.701

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

N/A

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI:No

MOL file:No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

N/A

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

N/A

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

N/A

8.2. A priori or a posteriori mechanistic interpretation:

$$\log EC50 = -\log K_{ow} \times 8.82 + TE \times 47.8 + \log E_{LUMO} \times 1.47 - E_{CCR} \times 39.7 + 50.3$$

There was a positive correlation between the $\log EC_{50}$ and the total energy (TE), which is a molecular descriptor related to the molecular energies and stabilities of PFASs. These include molecular internal energy, translational kinetic energy, the energy of electrons in a molecule, the vibration energy between atoms in a molecule, and the energy of a molecule rotating around the center of a mass. A higher TE value indicates that the molecule is not easily polarized or absorbed by cells, thus resulting in a lower toxicity. There was a positive correlation between the $\log EC_{50}$ and the lowest unoccupied molecule orbital energy (E_{LUMO}). As the electronegativity of the F atom is the strongest, the PFASs reacted with the action site of the target organism as the electron acceptor. The $\log EC_{50}$ was negatively correlated with the octanol–water partition coefficient (K_{ow}), which is related to the lipophilicity of PFASs. With an increase in the K_{ow} value, PFASs accumulate more easily in an organism, thus corresponding with a higher toxicity. K_{ow} is a key physico-chemical parameter serving as a classic molecular descriptor in QSAR modeling.

8.3. Other information about the mechanistic interpretation:

The molecular descriptors of the established model practicably explained the mechanism of acute toxicity (MOA).

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Zhang, J.;Zhang,M.; Tao, H.; Qi, G.; Guo, W.; Ge, H.; Shi, J. A QSAR–ICE–SSD Model Prediction of the PNECs for Per- and Polyfluoroalkyl Substances and Their Ecological Risks in an Area of Electroplating Factories. Molecules 2021,26,6574. https://doi.org/ 10.3390/molecules26216574

2. US EPA ECOTOX database

https://cfpub.epa.gov/ecotox/

9.3. Supporting information:

More information about model attached in supplementary materials:

https://doi.org/10.3390/molecules26216574

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

EC50(4). A QSAR Model Prediction of the acute toxicity for Per- and Polyfluoroalkyl Substances (*Chlorella vulgaris*)

1. QSAR identifier

1.1. QSAR identifier (title):

A QSAR Model Prediction of the acute toxicity for Per- and Polyfluoroalkyl Substances (*Chlorella vulgaris*)

1.2. Other related models:

N/A

1.3. Software coding the model:

SIMCA software

https://www.sartorius.com/

2. General information

2.1. Date of QMRF:

10/01/2022

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.
Trzy Lipy 3 Street, Building B,
80-172 Gdansk, Poland
+48 795 160 760
contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Jianghong Shi State Environmental Protection Key Laboratory of Integrated Surface Water- Groundwater Pollution Control, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China shijh@sustech.edu.cn

2. Hiu Ge State Environmental Protection Key Laboratory of Integrated Surface Water-Groundwater Pollution Control, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China geh@sustech.edu.cn

2.6. Date of model development and/or publication:

Published in 2021

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

Zhang, J.;Zhang, M.; Tao, H.; Qi, G.; Guo, W.; Ge, H.; Shi, J. A QSAR–ICE–SSD Model Prediction of the PNECs for Per- and Polyfluoroalkyl Substances and Their Ecological Risks in an Area of Electroplating Factories. Molecules 2021,26,6574.

https://doi.org/ 10.3390/molecules26216574

2.8. Availability of information about the model:

More information about model attached in supplementary materials:

https://doi.org/10.3390/molecules26216574

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

Chlorella vulgaris

3.2. Endpoint:

Acute toxicity - *Chlorella vulgaris*

3.3. Comment on the endpoint:

EC_{50} - median effect concentration

3.4. Endpoint units:

mg/L

3.5. Dependent variable:

$\log EC_{50}$

3.6. Experimental protocol:

The test methods were in accordance with standard test methods (e.g., the methods of the Organization for Economic Cooperation and Development).

3.7. Endpoint data quality and variability:

The acute toxicity data were mainly obtained from the US EPA ECOTOX database, the literature, and relevant government documents.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSAR - Stepwise multiple linear regression model

4.2. Explicit algorithm:

log EC_{50} = - 4.18 × K_{ow} - 0.332 × E_{CCR} - 4.29

4.3. Descriptors in the model:

1. E_{CCR} - core-core repulsion energy [EV]

2. K_{ow} - Octanol-water partition coefficient

4.4. Descriptor selection:

The stepwise regression method in SPSS was used to establish the multiple regression statistical models between the logarithmic values of the toxicity data (i.e., log LC(EC)50) and the molecular descriptors (including their logarithmic values)).

4.5. Algorithm and descriptor generation:

The molecular energy was optimized using the GAMESS Interface method in ChemBio3D (https://perkinelmerinformatics.com/). A total of 12 semi-empirical molecular descriptors were then calculated using the AM1 method in MOPAC 2016 (http://openmopac.net/) and the K_{ow} values were calculated using EPI Suite software (https://www.epa.gov/)

4.6. Software name and version for descriptor generation:

1. MOPAC 2016

total of 12 semi-empirical molecular descriptors were then calculated using the AM1 method

http://openmopac.net/

2. EPI Suite software

were used to calculate K_{ow} values

https://www.epa.gov/

4.7. Chemicals/ Descriptors ratio:

10 chemicals / 2 descriptors = 5

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

N/A

5.2. Method used to assess the applicability domain:

N/A

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

ICE models were used after verifying the application domain.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability 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 (response) for the training set:

All

6.5. Other information about the training set:

Training set consist of 10 compounds

6.6. Pre-processing of data before modelling:

The original EC_{50} data were expressed in log unit, also descriptors values were expressed in logarithmic values

6.7. Statistics for goodness-of-fit:

$R^2= 0.592$

$r^2=0.751$

* R^2 – coefficient of determination of the multiple regression, r^2 – conventional correlation coefficient or non-validation correlation coefficient

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

$Q^2_{LOO} = 0.673$

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

N/A

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI:No

MOL file:No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

N/A

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

N/A

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

N/A

8.2. A priori or a posteriori mechanistic interpretation:

$\log EC_{50} = - 4.18 \times K_{ow} - 0.332 \times E_{CCR} - 4.29$

The logEC50 was negatively correlated with the octanol–water partition coefficient (K_{ow}), which is related to the lipophilicity of PFASs. With an increase in the K_{ow} value, PFASs accumulate more easily in an organism, thus corresponding with a higher toxicity. K_{ow} is a key physico-chemical parameter serving as a classic molecular descriptor in QSAR modeling. The electron cloud of atoms in a molecule is deformed more easily with an increase in the ECCR value, which makes PFASs more likely to polarize and enter a cell.

8.3. Other information about the mechanistic interpretation:

The molecular descriptors of the established model practicably explained the mechanism of acute toxicity (MOA).

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Zhang, J.;Zhang,M.; Tao, H.; Qi, G.; Guo, W.; Ge, H.; Shi, J. A QSAR–ICE–SSD Model Prediction of the PNECs for Per- and Polyfluoroalkyl Substances and Their Ecological Risks in an Area of Electroplating Factories. Molecules 2021,26,6574.
https://doi.org/ 10.3390/molecules26216574

2. US EPA ECOTOX database
https://cfpub.epa.gov/ecotox/

9.3. Supporting information:

More information about model attached in supplementary materials:
https://doi.org/10.3390/molecules26216574

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

LC50(5). A QSAR Model Prediction of the acute toxicity for Per- and Polyfluoroalkyl Substances (*Daphnia magna*)

1. QSAR identifier

1.1. QSAR identifier (title):

A QSAR Model Prediction of the acute toxicity for Per- and Polyfluoroalkyl Substances (*Daphnia magna*)

1.2. Other related models:

N/A

1.3. Software coding the model:

SIMCA software

https://www.sartorius.com/

2. General information

2.1. Date of QMRF:

10/02/2022

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.

Trzy Lipy 3 Street, Building B,

80-172 Gdansk, Poland

+48 795 160 760

contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

2.5. Model developer(s) and contact details:

1. Jianghong Shi State Environmental Protection Key Laboratory of Integrated Surface Water- Groundwater Pollution Control, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China shijh@sustech.edu.cn

2. Hiu Ge State Environmental Protection Key Laboratory of Integrated Surface Water-Groundwater Pollution Control, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China geh@sustech.edu.cn

2.6. Date of model development and/or publication:

Published in 2021

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

Zhang, J.;Zhang, M.; Tao, H.; Qi, G.; Guo, W.; Ge, H.; Shi, J. A QSAR–ICE–SSD Model Prediction of the PNECs for Per- and Polyfluoroalkyl Substances and Their Ecological Risks in an Area of Electroplating Factories. Molecules 2021,26,6574.
https://doi.org/ 10.3390/molecules26216574

2.8. Availability of information about the model:

More information about model attached in supplementary materials:
https://doi.org/10.3390/molecules26216574

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

Daphnia magna

3.2. Endpoint:

Acute toxicity - *Daphnia magna*

3.3. Comment on the endpoint:

LC_{50} - median lethal concentration

3.4. Endpoint units:

mg/L

3.5. Dependent variable:

$\log LC_{50}$

3.6. Experimental protocol:

The test methods were in accordance with standard test methods (e.g., the methods of the Organization for Economic Cooperation and Development).

3.7. Endpoint data quality and variability:

The acute toxicity data were mainly obtained from the US EPA ECOTOX database, the literature, and relevant government documents.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSAR - Stepwise multiple linear regression model

4.2. Explicit algorithm:

$\log LC_{50} = - K_{ow} \times 4.09 + \log TE \times 9.75 - E_{CCR} \times 7.03 + \log E_{LUMO} \times 1.63 + 1.95$

4.3. Descriptors in the model:

1. K_{ow} - Octanol-water partition coefficient

2. TE - total energy [EV]

3. E_{LUMO} - Lowest unoccupied molecule orbital energy [EV]

4. E_{CCR} - core-core repulsion energy [EV]

4.4. Descriptor selection:

The stepwise regression method in SPSS was used to establish the multiple regression statistical models between the logarithmic values of the toxicity data (i.e., log LC(EC)50) and the molecular descriptors (including their logarithmic values)).

4.5. Algorithm and descriptor generation:

The molecular energy was optimized using the GAMESS Interface method in ChemBio3D (https://perkinelmerinformatics.com/). A total of 12 semi-empirical molecular descriptors were then calculated using the AM1 method in MOPAC 2016 (http://openmopac.net/) and the K_{ow} values were calculated using EPI Suite software (https://www.epa.gov/)

4.6. Software name and version for descriptor generation:

1. MOPAC 2016

total of 12 semi-empirical molecular descriptors were then calculated using the AM1 method

http://openmopac.net/

2. EPI Suite software

were used to calculate K_{ow} values

https://www.epa.gov/

4.7. Chemicals/ Descriptors ratio:

10 chemicals / 4 descriptors = 2.5

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

N/A

5.2. Method used to assess the applicability domain:

N/A

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

ICE models were used after verifying the application domain.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability 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 (response) for the training set:

All

6.5. Other information about the training set:

Training set consist of 10 compounds

6.6. Pre-processing of data before modelling:

The original LC_{50} data were expressed in log unit, also descriptors values were expressed in logarithmic values

6.7. Statistics for goodness-of-fit:

R^2= 0.370

r^2=0.605

* R^2 – coefficient of determination of the multiple regression, r^2 – conventional correlation coefficient or non-validation correlation coefficient

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

Q^2_{LOO} = 0.580

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

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

N/A

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI:No

MOL file:No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

N/A

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

N/A

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

N/A

8.2. A priori or a posteriori mechanistic interpretation:

$$\log LC_{50} = - K_{ow} \times 4.09 + \log TE \times 9.75 - E_{CCR} \times 7.03 + \log E_{LUMO} \times 1.63 + 1.95$$

According to the frontier orbital theory, the occurrence of the reaction is related to the difference between the highest occupied orbital energy (E_{HOMO}) of the electron donor and the E_{LUMO} of the electron acceptor; that is, E_{HOMO}–E_{LUMO} (also known as the energy band gap). The larger the band gap, the easier the reaction and the stronger the binding force between the electron donor and the electron acceptor. Hence, the larger the band gap, the more obvious the toxicity and the lower the $\log LC_{50}$ value. There was a negative correlation between the log LC_{50} and the nuclear–nuclear repulsive energy (E_{CCR}). The electron cloud of atoms in a molecule is deformed more easily with an increase in the E_{CCR} value, which makes PFASs more likely to polarize and enter a cell. With an increase in the K_{ow} value, PFASs accumulate more easily in an organism, thus corresponding with a higher toxicity. K_{ow} is a key physico-chemical parameter serving as a classic molecular descriptor in QSAR modeling.

8.3. Other information about the mechanistic interpretation:

The molecular descriptors of the established model practicably explained the mechanism of acute toxicity (MOA).

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Zhang, J.;Zhang,M.; Tao, H.; Qi, G.; Guo, W.; Ge, H.; Shi, J. A QSAR–ICE–SSD Model Prediction of the PNECs for Per- and Polyfluoroalkyl Substances and Their Ecological Risks in an Area of Electroplating Factories. Molecules 2021,26,6574. https://doi.org/ 10.3390/molecules26216574

2. US EPA ECOTOX database https://cfpub.epa.gov/ecotox/

9.3. Supporting information:

More information about model attached in supplementary materials: https://doi.org/10.3390/molecules26216574

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments:

LC50(6). A QSAR Model Prediction of the acute toxicity for Per- and Polyfluoroalkyl Substances (*Danio rerio)*

1. QSAR identifier

1.1. QSAR identifier (title):

A QSAR Model Prediction of the acute toxicity for Per- and Polyfluoroalkyl Substances (*Danio rerio)*

1.2. Other related models:

N/A

1.3. Software coding the model:

SIMCA software

https://www.sartorius.com/

2. General information

2.1. Date of QMRF:

10/01/2022

2.2. QMRF author(s) and contact details:

QSAR Lab Sp. z o.o.
Trzy Lipy 3 Street, Building B,
80-172 Gdansk, Poland
+48 795 160 760
contact@qsarlab.com https://www.qsarlab.com

2.3. Date of QMRF update(s):

N/A

2.4. QMRF update(s):

N/A

Model developer(s) and contact details:

1. Jianghong Shi State Environmental Protection Key Laboratory of Integrated Surface Water- Groundwater Pollution Control, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China shijh@sustech.edu.cn

2. Hiu Ge State Environmental Protection Key Laboratory of Integrated Surface Water-Groundwater Pollution Control, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China geh@sustech.edu.cn

2.5. Date of model development and/or publication:

Published in 2021

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

Zhang, J.;Zhang, M.; Tao, H.; Qi, G.; Guo, W.; Ge, H.; Shi, J. A QSAR–ICE–SSD Model Prediction of the PNECs for Per- and Polyfluoroalkyl Substances and Their Ecological Risks in an Area of Electroplating Factories. Molecules 2021,26,6574.

https://doi.org/ 10.3390/molecules26216574

2.7. Availability of information about the model:

More information about model attached in supplementary materials:

https://doi.org/10.3390/molecules26216574

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

N/A

3. Defining the endpoint – OECD Principle 1

3.1. Species:

Danio rerio

3.2. Endpoint:

Acute toxicity - *Danio rerio*

3.3. Comment on the endpoint:

LC_{50} - median lethal concentration

3.4. Endpoint units:

mg/L

3.5. Dependent variable:

$\log LC_{50}$

3.6. Experimental protocol:

The test methods were in accordance with standard test methods (e.g., the methods of the Organization for Economic Cooperation and Development).

3.7. Endpoint data quality and variability:

The acute toxicity data were mainly obtained from the US EPA ECOTOX database, the literature, and relevant government documents.

4. Defining the algorithm – OECD Principle 2

4.1. Type of model:

QSAR - Stepwise multiple linear regression model

4.2. Explicit algorithm:

$\log LC_{50} = -K_{ow} \times 1.03 - E_{CCR} \times 1.04 + E_{LUMO} \times 0.318 + 2.94$

4.3. Descriptors in the model:

1. K_{ow} - Octanol-water partition coefficient

2. E_{CCR} - core-core repulsion Energy [EV]

3. E_{LUMO} - Lowest unoccupied molecule orbital energy [EV]

4.4. Descriptor selection:

The stepwise regression method in SPSS was used to establish the multiple regression statistical models between the logarithmic values of the toxicity data (i.e., log LC(EC)50) and the molecular descriptors (including their logarithmic values)).

4.5. Algorithm and descriptor generation:

The molecular energy was optimized using the GAMESS Interface method in ChemBio3D (https://perkinelmerinformatics.com/). A total of 12 semi-empirical molecular descriptors were then calculated using the AM1 method in MOPAC 2016 (http://openmopac.net/) and the K_{ow} values were calculated using EPI Suite software (https://www.epa.gov/)

4.6. Software name and version for descriptor generation:

1. MOPAC 2016

total of 12 semi-empirical molecular descriptors were then calculated using the AM1 method

http://openmopac.net/

2. EPI Suite software

were used to calculate K_{ow} values

https://www.epa.gov/

4.7. Chemicals/ Descriptors ratio:

12 chemicals / 3 descriptors = 4

5. Defining the applicability domain – OECD Principle 3

5.1. Description of the applicability domain of the model:

N/A

5.2. Method used to assess the applicability domain:

N/A

5.3. Software name and version for applicability domain assessment:

N/A

5.4. Limits of applicability:

ICE models were used after verifying the application domain.

6. Defining goodness-of-fit and robustness – OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Availability 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 (response) for the training set:

All

6.5. Other information about the training set:

Training set consist of 12 compounds

6.6. Pre-processing of data before modelling:

The original LC_{50} data were expressed in log unit, also descriptors values were expressed in logarithmic values

6.7. Statistics for goodness-of-fit:

R^2= 0.558

r^2=0.722

* R^2 – coefficient of determination of the multiple regression, r^2 – conventional correlation coefficient or non-validation correlation coefficient

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

Q^2_{LOO} = 0.630

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

N/A

6.10. Robustness – Statistics obtained by Y-scrambling:

N/A

6.11. Robustness – Statistics obtained by bootstrap:

N/A

6.12. Robustness – Statistics obtained by other methods:

N/A

7. Defining predictivity – OECD Principle 4

7.1. Availability of the external validation set:

N/A

7.2. Availability information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI:No

MOL file:No

7.3. Data for each descriptor variable for the external validation set:

N/A

7.4. Data for the dependent variable (response) for the external validation set:

N/A

7.5. Other information about the external validation set:

N/A

7.6. Experimental design of test set:

N/A

7.7. Predictivity – Statistics obtained by external validation:

N/A

7.8. Predictivity – Assessment of the external validation set:

N/A

7.9. Comments on the external validation of the model:

N/A

8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model:

N/A

8.2. A priori or a posteriori mechanistic interpretation:

$\log LC_{50} = -K_{ow} \times 1.03 - E_{CCR} \times 1.04 + E_{LUMO} \times 0.318 + 2.94$

According to the frontier orbital theory, the occurrence of the reaction is related to the difference between the highest occupied orbital energy (E_{HOMO}) of the electron donor and the E_{LUMO} of the electron acceptor; that is, E_{HOMO}–E_{LUMO} (also known as the energy band gap). The larger the band gap, the easier the reaction and the stronger the binding force between the electron donor and the electron acceptor. Hence, the larger the band gap, the more obvious the toxicity and the lower the log LC_{50} value. There was a negative correlation between the log LC_{50} and the nuclear–nuclear repulsive energy (E_{CCR}). The electron cloud of atoms in a molecule is deformed more easily with an increase in the E_{CCR} value, which makes PFASs more likely to polarize and enter a cell. With an increase in the K_{ow} value, PFASs accumulate more easily in an organism, thus corresponding with a higher toxicity. K_{ow} is a key physico-chemical parameter serving as a classic molecular descriptor in QSAR modeling.

8.3. Other information about the mechanistic interpretation:

The molecular descriptors of the established model practicably explained the mechanism of acute toxicity (MOA).

9. Miscellaneous information

9.1. Comments:

N/A

9.2. Bibliography:

1. Zhang, J.;Zhang,M.; Tao, H.; Qi, G.; Guo, W.; Ge, H.; Shi, J. A QSAR–ICE–SSD Model Prediction of the PNECs for Per- and Polyfluoroalkyl Substances and Their Ecological Risks in an Area of Electroplating Factories. Molecules 2021,26,6574.
https://doi.org/ 10.3390/molecules26216574

2. US EPA ECOTOX database
https://cfpub.epa.gov/ecotox/

9.3. Supporting information:

More information about model attached in supplementary materials:
https://doi.org/10.3390/molecules26216574

10. Summary for the JRC QSAR Model Database (compiled by JRC)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments: