	*QMRF identifier (JRC Inventory):***Q13-22b-0015**
	*QMRF Title:***QSAR for the Global Half-Life Index of Persistent Organic Pollutants**
	*Printing Date:***Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

QSAR for the Global Half-Life Index of Persistent Organic Pollutants

1.2.Other related models:

1.3.Software coding the model:

MOBY DIGS

Software for multilinear regression analysis and variable subset selection by Genetic Algorithm, ver. 1.0 beta for Windows, 2004

Todeschini Roberto, Talete srl, Milan (Italy)

2.General information

2.1.Date of QMRF:

29/05/2009

2.2.QMRF author(s) and contact details:

[1]Paola Gramatica University of Insubria paola.gramatica@uninsubria.it

[2]Ester Papa University of Insubria ester.papa@uninsubria.it

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1]Ester Papa University of Insubri ester.papa@uninsubria.it

[2]Paola Gramatica University of Insubria paola.gramatica@uninsubria.it

2.6.Date of model development and/or publication:

2007

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

Gramatica P & Papa E (2007). Screening and Ranking of POPs for Global Half-Life: QSAR Approaches for Prioritization Based on Molecular Structure. Environmental Science & Technology 41, 2833-2839.

2.8.Availability of information about the model:

This model is non proprietary. Training and test sets are available

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

3.Defining the endpoint - OECD Principle 1

3.1.Species:

3.2.Endpoint:

2.Environmental fate parameters 2.2.b.Persistence: Abiotic degradation in air (Phototransformation). Indirect photolysis (OH-radical reaction, ozone-radical reaction, other)

3.3.Comment on endpoint:

The Global Half-Life Index (GHLI) is a macro-variable which condenses the chemical tendency to environmental persistence. It is derived by Principal Component Analysis (PCA) from half-life data for transformation in air, water, sediment and soil for a set of 250 organic

POP-type chemicals (Mackay et al, 2000). The scores of the compounds along PC1, which provides alone the largest part (78%) of the total information, defined the Global Half Life Index (GHLI). GHLI ranks the compounds according to their cumulative half-life and discriminates between them with regard to persistence.

3.4.Endpoint units:

The logarithm of half-life values (hours) in the four studied environmental media (air, water, sediment and soil) were combined by Principal Components Analysis. The GHLI, obtained by PCA (PC1 values), is a dimensionless endpoint.

3.5.Dependent variable:

GHLI

3.6.Experimental protocol:

3.7.Endpoint data quality and variability:

For the development of the GHLI semiquantitative degradation half lives in air, soil, water and sediment have been taken from Mackay et al (2000). These half lives are organized in nine half-life categories. In the present study the respective category averages have been taken as reference data based on experimental information, even though some of these handbook data can be based on expert judgement.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR

4.2.Explicit algorithm:

Multilinear regression model

GHLI = -3.12 (±0.77) + 0.33 (±0.045) X0v + 5.1 (±0.99) Mv - 0.32 (±0.0613) MAXDP - 0.61 (±0.10) nHDon - 0.5 (±1.15) CIC0 - 0.61 (±0.13) O-060

4.3.Descriptors in the model:

[1]X0v Randic connectivity index of 0 order

[2]Mv constitutional descriptor, mean atomic van der Waals volume

[3]MAXDP topological descriptor, maximal electrotopological positive variation

[4]nHDon number of donor atoms for H bonds

[5]CIC0 information index related to the complementary information content

[6]O-060 fragment related to the presence of oxygen atoms in ethers and esters

4.4.Descriptor selection:

A total of 662 molecular descriptors of different kinds were calculated to describe the chemical diversity of the compounds. Constant values and descriptors found to be correlated pairwise were excluded in a prereduction step (when there was more than 98% pairwise correlation, one variable was deleted). The Genetic Algorithm (GA) was applied to a final set of 474 descriptors for variable selection.

4.5.Algorithm and descriptor generation:

Multiple linear regression (OLS) was applied to generate the model. Descriptors were generated by DRAGON 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 obtained by the Molecular Mechanics method of Allinger (MM+) using the package HYPERCHEM.

4.6.Software name and version for descriptor generation:

DRAGON

Software for Molecular Descriptor Calculations - 2006, version 5.4 for Windows

Roberto Todeschini, Talete srl, Milan, Italy.

4.7.Chemicals/Descriptors ratio:

125 chemicals / 6 descriptors = 20.8

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

No outliers in the structural domain.

Outliers for the response (training set):*o*-cresol;
1,1'-biphenyl- 4,4'-diamine

Outliers for the response (validation set): 1,2-dichloropropane ;
1,2-dichloroethane; tetramethyl thioperoxydicarbonic diamide ; malathion
; 2-propenal ; bromodichloromethane

5.2.Method used to assess the applicability domain:

The structural AD of the model was checked by Leverage approach. The presence of outliers (i.e. compounds with crossvalidated standardized residuals greater than 2.5 standard deviation units) and chemicals very structurally influential in determining model parameters (i.e. compounds with a high*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)
was verified by the Williams plot.

5.3.Software name and version for applicability domain assessment:

MOBY DIGS - Professionals Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm, Version 1.0 beta for Windows; Talete srl: Milan, 2004.

Calculation of hat values; calculated - predicted values of the response

Excel

Calculation of standardised residuals (compounds with crossvalidated standardized residuals greater than 2.5 standard deviation units)

5.4.Limits of applicability:

High leverage compounds: hat value > 0.168

Outliers for the response: standardised residuals > 2.5 standard deviation units

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: 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 for the training set:

All

6.5.Other information about the training set:

125 compounds

6.6.Pre-processing of data before modelling:

Half-life (hours) data in 4 environmental compartments were transformed into logarithmic units and then combined by PCA to obtain GHLIndex (modelled endpoint). The PC1 score values were multiplied by -1 to obtain increasing positive values of the GHLI Index (high positive GHLI values = High persistence). Raw Half Life (hours) data are given as well as the GHL Index.

6.7.Statistics for goodness-of-fit:

n = 125, R^2 = 0.85

RMSE = 0.76, $RMSE_{CV}$= 0.70

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

Q^2_{LOO}= 0.83

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

$Q^2_{LMO(50\%)}$ = 0.82

6.10.Robustness - Statistics obtained by Y-scrambling:

R^2_{Y-SC}= 0.05

6.11.Robustness - Statistics obtained by bootstrap:

Q^2_{BOOT}= 0.83

6.12.Robustness - Statistics obtained by other methods:

no other data 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: 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 for the external validation set:

All

7.5.Other information about the external validation set:

125 compounds

7.6.Experimental design of test set:

The splitting of the original was carried out by random selection through activity sampling: the whole range of property was sorted through ascending order, and every second compound was assigned to the training set.

7.7.Predictivity - Statistics obtained by external validation:

$R^2{}_{EXT}$ = 0.79; RMSE (validation set) = 0.78

7.8.Predictivity - Assessment of the external validation set:

The external validation set was representative of the response and of the structural domain of the training set. All the compounds of the validation set fell into the structural AD of the training set

7.9.Comments on the external validation of the model:

The application of the GHLI-QSPR model to 10 reference chemicals gave prediction in agreement with those obtained by applying multimedia models by Klasmeier et al. (2006). Moreover, the results of the GHLI-QSPR model is in satisfactory agreement with those of the U.S. EPA PBT-Profiler.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

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

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori mechanistic interpretation: the variables take account of the different environmental persistence tendency, such as chemical size (X0v and Mv, as more complex chemicals are generally expected to be more persistent than simpler) and electronic features (CIC0, MAXDP, nHDon, and O-060, more related to a compound’s ability to form electrostatic and dipole-dipole interactions in the surrounding media). These features can directly influence the bioavailability and partitioning of chemicals into different environmental compartments and can indirectly determine their availability for different degradation pathways.
It is also interesting to note that both the descriptor MAXDP, related to molecule electrophilicity, and the topological CIC0 had already demonstrated the ability to model important environmental partition properties such as the soil sorption partition parameter Koc and tropospheric degradability by OH radicals. There is an evident relation between sorption and persistence as the more sorbed chemicals are the most recalcitrant to biotic and abiotic degradation, thus rendering them more persistent.

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1]Klasmeier J, Matthies M, MacLeod M, Fenner K, Scheringer M, Stroebe M, Le Gall AC, McKone T, Van De Meent D & Wania F (2006). Application of Multimedia Models for Screening Assessment of Long-Range Transport Potential and Overall Persistence. Environmental Science & Technology 40, 53-60.

[2]Mackay D, Shiu WY & Ma KC (2000). Physical-Chemical Properties and Environmental Fate Handbook, CRCnet-BASE CD-ROM; Chapman and Hall/CRC: Boca Raton, FL, USA.

[3]U.S. EPA PBT-Profiler; U.S. EPA: Washington, DC (accessed December 13, 2006).

[4]Gramatica P, Pilutti P & Papa E (2004). Validated QSAR Prediction of OH Tropospheric degradability: splitting into training-prediction set and consensus modeling. Journal of Chemical Information and Computer Sciences 44, 1794-1802.

[5]Gramatica P, Corradi M & Consonni V (2000). Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by different sets of molecular descriptors. Chemosphere 41, 763-777.

[6]Gramatica P, Giani E & Papa E (2007). Statistical External Validation and Consensus Modeling: a QSPR Case Study for Koc Prediction. Journal of Molecular Graphics and Modelling 25, 755-766.

[7]Papa E & Gramatica P (2008). Screening of persistent organic pollutants by QSPR classification models: a comparative study. Journal of Molecular Graphics and Modelling 27, 59-65.

[8]Gramatica, P. Principles of QSAR models validation: internal and external. QSAR Comb. Sci. 2007, 26(5), 694-701.

9.3.Supporting information:

GHLI Training_125.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q13-22b-0015/attachment/A654
GHLI test_125.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q13-22b-0015/attachment/A655

Test set(s)

Supporting information

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q13-22b-0015

10.2.Publication date:

2013-06-24

10.3.Keywords:

Global Half-Life Index;GHLI;Persistent Organic Pollutant;POP;MOBYDIGS;persistence;

10.4.Comments:

former Q7-17-11-112