1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSPR PaDEL-Descriptor model for Global Half-Life Index (GHLI)

1.2.Other related models:

P. Gramatica, E. Papa, Screening and Ranking of POPs for Global Half-Life: QSAR Approaches for Prioritization Based on Molecular Structure. Environ. Sci. Technol., 2007, 41, 2833-2839. [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:

20/11/2013

2.2.QMRF author(s) and contact details:

Stefano Cassani DiSTA, University of Insubria (Varese - Italy) +390332421439 stefano.cassani@uninsubria.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) +390332421439 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]Gramatica P., et al. QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS, submitted to J. Comput. Chem. (Software News and Updates), 2013.

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 information available

3.Defining the endpoint - OECD Principle 1

3.1.Species:

No information available

3.2.Endpoint:

Environmental Fate parameters Persistence Global Half-Life Index (GHLI)

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 Analisys (PCA) from half-life data for transformation in air, water, sediment and soil for a set of 250 organic POP-type chemicals. The scores of the compounds along PC1, which provides alone the largest part (78%) of the total information, defined the Global HalfLife 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 , were combined by Principal Component Analisys. The GHLIndex, obtained by PCA (PC1 values), is an adimensional endpoint.

3.5.Dependent variable:

GHLI

3.6.Experimental protocol:

The data set includes 250 organic compounds of known half-lives for transformation into four environmental media: air, water, soil, and sediment. (original source: Mackay, D.; Shiu, W. Y.; Ma, K. C. Physical-Chemical Properties and Environmental Fate Handbook, CRCnet-BASE CD-ROM; Chapman and Hall/CRC: Boca Raton, FL, 2000 [2])

3.7.Endpoint data quality and variability:

For the development of the GHLIndex semiquantitative degradation half lives in air, soil, water and sediment have been taken from: Mackay, D.; Shiu, W. Y.; Ma, K. C. Physical-Chemical Properties and Environmental Fate Handbook, CRCnet-BASE CD-ROM; Chapman and Hall/CRC: Boca Raton, FL, 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 - Multiple linear regression model (OLS - Ordinary Least Square)

4.2.Explicit algorithm:

GHLI Split model

MLR-OLS method. Model developed on a training set of 125 compounds.

GHLI Full model

MLR-OLS method. Model developed on a training set of 250 compounds.

Split model equation: GHLI= -0.57 + 0.01 MW - 0.15 maxHBa - 0.43 nHBDon_Lipinski - 0.05 nBondsS2 + 0.60 minsCl

Full model equation: GHLI= -0.57 + 0.01 MW - 0.15 maxHBa + 0.74 minsCl - 0.05 nBondsS2 - 0.43 nHBDon_Lipinski

4.3.Descriptors in the model:

[1]MW Molecular Weight
[2]maxHBa Maximum E-States for (strong) Hydrogen Bond acceptors
[3]minsCl Minimum atom-type E-State: -Cl
[4]nBondsS2 Total number of single bonds (including bonds to hydrogens, excluding aromatic bonds)
[5]nHBDon_Lipinski Number of hydrogen bond donors (using Lipinski's definition: Any OH or NH. Each available hydrogen atom is counted as one hydrogen bond donor)

4.4.Descriptor selection:

A total of 1561 molecular descriptors of differing types (0D, 1D, 2D) and 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 121 molecular descriptors were used as input variables for variable subset selection by genetic algorithm (GA-VSS).The models were initially developed by the all-subset-procedure until two variables. Then the GA was applied in order to explore new combinations of variables, selecting the variables (five) by a mechanism of reproduction/mutation. 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:

Split Model: 125 chemicals / 5 descriptors = 25
Full Model: 250 chemicals / 5 descriptors = 50

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 GHLI values: -3.13 / 4.98
Range of descriptor values: MW (32.03 / 493.69), maxHBa (0 / 12.84), nBondsS2 (3 / 37), nHBDon_Lipinski (0 / 4), minsCl (0 / 1.38)

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*=0.072). 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:

Split model domain: outliers for structure, hat>0.144 (h*): benzidine (92-87-5), n-dodecane (112-40-3). Outliers for response, standardised residuals > 2.5 standard deviation units: Methoxychlor (72-43-5), Malathion (121-75-5), Aldicarb (116-06-3), dalapon (75-99-0), Hexachlorobenzene (118-74-1). **FULL model domain**: outliers for structure, hat>0.072 (h*): benzidine (92-87-5), n-dodecane (112-40-3), decachlorobiphenyl (2051-24-3). Outliers for response, standardised residuals > 2.5 standard deviation units: Methoxychlor (72-43-5), Malathion (121-75-5),Aldicarb (116-06-3), dalapon (75-99-0).

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 Model consists of 125 compounds with a range of GHLI values from -3.13 to 4.98. The splitting is based on the ordered response.

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).

6.7.Statistics for goodness-of-fit:

R^2= 0.86; CCCtr [3]=0.93; RMSE= 0.67

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

Q^2LOO= 0.85; CCCcv=0.92; RMSEcv= 0.70

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

Q^2LMO= 0.85

6.10.Robustness - Statistics obtained by Y-scrambling:

R^2y-sc= 0.04

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:

The external prediction set consists of 125 compounds with a range of GHLI values from -2.79 to 4.73.

7.6.Experimental design of test set:

The splitting of the original data set (250 compounds) into a training set of 125 compounds and a prediction set of 125 compounds was realized by Ordered response.

7.7.Predictivity - Statistics obtained by external validation:

Q^2extF1 [4]= 0.83; Q^2extF2 [5]= 0.83; Q^2extF3 [6]= 0.84; CCCex=0.90; RMSE= 0.71

7.8.Predictivity - Assessment of the external validation set:

The splitting methodology based on ordered response allowed for the selection of a meaningful training set and a representative prediction set.

Training and prediction set are balanced according to both response and structure. In particular, the range of GHLI values are [-3.13 / 4.98] and [-2.79 / 4.73] respectively for training and prediction set. As much as concern structural representativity, the range of descriptors values are:

MW: training set (45.06 / 493.69), prediction set (32.03 / 455.74)
maxHBa: training set (0 / 12.36), prediction set (0 / 12.84)
nBondsS2: training set (3 / 37), prediction set (4 / 37)

nHBDon_Lipinski: training set (0 / 4), prediction set (0 / 3)
minsCl: training set (0 / 1.38), prediction set (0 / 1.38)

7.9.Comments on the external validation of the model:

No 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 for this physico-chemical property was set a priori, but a mechanistic interpretation of molecular descriptors was provided a posteriori (see 8.2).

8.2.A priori or a posteriori mechanistic interpretation:

The DRAGON model published in Gramatica and Papa [7] is:

GHLI= -3.12 + 0.33 X0v + 5.06 Mv - 0.32 MAXDP - 0.61 nHDon - 0.5 CIC0 - 0.61 O-060 (split)

where
X0v=valence connectivity index chi-0
Mv=mean atomic van der waals volume (scaled on carbon atom)
MAXDP=maximal electrotopological positive variation
nHDon=number of donor atoms for H-bonds (N and O)
CIC0=complementary information content (neighborhood symmetry of 0-order)
O-060=Al-O-Ar/ Ar-O-Ar / R-O-R / R-O-C=X

The equation of the new PaDEL-descriptor model included in QSARINS is : GHLI= -0.57 + 0.01 MW - 0.15 maxHBa + 0.74 minsCl - 0.05 nBondsS2 - 0.43 nHBDon_Lipinski where
MW=Molecular Weight maxHBa=Maximum E-States for (strong) Hydrogen Bond acceptors
minsCl= Minimum atom-type E-State: -Cl
nBondsS2= Total number of single bonds (including bonds to hydrogens, excluding aromatic bonds) nHBDon_Lipinski=Number of hydrogen bond donors (using Lipinski's definition: Any OH or NH. Each available hydrogen atom is counted as one hydrogen bond donor)

The modeling variables take account the different structural properties involved in defining environmental persistence tendency, such as chemical size (MW, as more complex chemicals are generally expected to be more persistent than simpler) and electronic features (maxHBa, nHBDon_Lipinski). 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. VThe modeling variables take account the different structural properties involved in defining environmental persistence tendency, such as chemical size (MW, as more complex chemicals are generally expected to be more persistent than simpler) and electronic features (maxHBa, nHBDon_Lipinski). 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.

The new PaDEL model has five descriptors, instead of six of published DRAGON model. Despite this, the PaDEL model shows slighly better performances, both in fitting and internal/external validation. Three couples of descriptors in DRAGON and PaDEL-Descriptor models shows high correlation: X0v/MW (0.98), nHdon/nHDon_Lipinski (0.96) and MAXDP/maxHBa (0.88).

8.3.Other information about the mechanistic interpretation:

No other information available

9.Miscellaneous information

9.1.Comments:

To predict GHLI for new chemicals without experimental data, it is suggested to apply the equation of the Full Model, developed on all the available chemicals (N=250), thus ensuring a wider applicability domain.

The equation (reported also in section 4.2) and the statistical parameters of the full model are:

GHLI= -0.57 + 0.01 MW - 0.15 maxHBa + 0.74 minsCl - 0.05 nBondsS2 - 0.43 nHBDon_Lipinski

N = 250; R2 = 0.85; Q2 = 0.85; Q2LMO = 0.84; CCC = 0.92; CCCcv = 0.91 ;RMSE= 0.687; RMSEcv = 0.704

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]Mackay, D.; Shiu, W. Y.; Ma, K. C. Physical-Chemical Properties and Environmental Fate Handbook, CRCnet-BASE CD-ROM; Chapman and Hall/CRC: Boca Raton, FL, 2000.

[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]P. Gramatica, E. Papa, Screening and Ranking of POPs for Global Half-Life: QSAR Approaches for Prioritization Based on Molecular Structure. Environ. Sci. Technol., 2007, 41, 2833-2839.

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