	QMRF identifier (JRC Inventory):**Q15-66-0018**
	QMRF Title:**QSARINS model for Global Half-Life Index**
	Printing Date:**Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

QSARINS model for Global Half-Life Index

1.2.Other related models:

1.3.Software coding the model:

PaDEL-Descriptor

A software to calculate molecular descriptors and fingerprints [ref 2; sect 9.2], version 2.18

Yap Chun Wei, email: 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 (verified also with 2.2, 2015)

Paola Gramatica, email: paola.gramatica@uninsubria.it

www.qsar.it

2.General information

2.1.Date of QMRF:

23/01/2015

2.2.QMRF author(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/

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(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]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:

Developed in 2013, Published in 2014

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

[1]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. DOI: 10.1021/es061773b
[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 (sect 9.3) and in 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:

No information available

3.2.Endpoint:

6.Other 6.6.Other

3.3.Comment on endpoint:

The Global Half-Life Index is a macro-variable which condenses the chemical tendency to environmental persistence. It is derived by Principal Components Analysis (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 as 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 Analysis. The GHLIndex, obtained by PCA (PC1 values), is thus an adimensional endpoint.

3.5.Dependent variable:

GHLI

3.6.Experimental protocol:

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 WY, Ma KC (2000). Physical-Chemical Properties and Environmental Fate Handbook, CRCnet-BASE CD-ROM; Chapman and Hall/CRC: Boca Raton, FL [ref 5; sect 9.2]. 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 judgment.

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:

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 (N Training: 125):**GHLI**= -0.57 + 0.01 MW - 0.15 maxHBa - 0.43 nHBDon_Lipinski - 0.05 nBondsS2 + 0.60 minsCl

Full model equation (N Training:250):**GHLI**= -0.57 + 0.01 MW - 0.15 maxHBa + 0.74 minsCl - 0.05 nBondsS2 - 0.43 nHBDon_Lipinski

The five modeling descriptors, calculated with PaDEL-Descriptor 2.18, are the following(see section 4.3 for a more detailed explanation):

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)

4.3.Descriptors in the model:

[1]MW g/mol Molecular Weight. Encodes for molecular size, as more complex chemicals are generally expected to be more persistent than simpler.

[2]maxHBa dimensionless Maximum E-States for (strong) Hydrogen Bond acceptors. Encodes for electronic features; 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.

[3]minsCl dimensionless Minimum atom-type E-State: -Cl. It is related to the presence of chlorine atoms in the molecule. It has a positive sign in the equation, as, usually, the substitution with chlorine confers greater resistance to biological and photolytic degradation, resulting in a higher persistence of the chemical.

[4]nBondsS2 dimensionless Total number of single bonds (including bonds to hydrogens, excluding aromatic bonds)

[5]nHBDon_Lipinski dimensionless Number of hydrogen bond donors (using Lipinski's definition: Any OH or NH. Each available hydrogen atom is counted as one hydrogen bond donor). Encodes for electronic features; 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.

4.4.Descriptor selection:

A total of 1561 molecular descriptors of differing types (0D, 1D, 2D) and fingerprints were calculated with PaDEL-Descriptor 2.18 [ref 2; sect 9.2]. 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 five variables by a mechanism of reproduction/mutation. The optimized parameter used was Q^2LOO (leave-one-out). 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: 5 modeling descriptors selected from 121.

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 (open source). 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 [6]. Then, these files were converted by OpenBabel 2.3.2 into MDL-MOL format and used as input for the calculation of descriptors in PaDEL-Descriptor. Any user can re-derive the model calculating the molecular descriptors with PaDEL-Descriptor 2.18 software (now included in QSARINS 2.2) and applying the given equation (automatically done in QSARINS 2.2).

4.6.Software name and version for descriptor generation:

PaDEL-Descriptor
A software to calculate molecular descriptors and fingerprints, version 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. Used for conversion between HYPERCHEM files (hin) and MDL-MOL files, version 2.3.2

http://openbabel.org/wiki/Main_Page

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 have been used to define both structural and response outliers (see section 5.4). The plot of leverages (hat diagonals) versus standardized residuals, i.e. the Williams plot, verified the presence of response outliers (i.e. compounds with cross-validated standardized residuals greater than 2.5 standard deviation units) and chemicals very structurally influential in determining model parameters (i.e. compounds with a leverage value (h) greater than 3p'/n (h*), where p' is the number of model variables plus one, and n is the number of the objects used to calculate the model).

For new compounds without experimental data, leverage can be used as a quantitative measure for evaluating the degree of extrapolation: for compounds with a high leverage value (h > h*), that are structural outliers, predictions should be considered less reliable.

Response and descriptor space:

Range of 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 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:

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

The response applicability domain can be verified by the standardized residuals in cross-validation greater than 2.5 standard deviation units

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 version 2.2, 2015)

Paola Gramatica, email: 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: 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:

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 (50% of chemicals in the prediction set) is based on the Ordered Response method. Chemicals have been ordered according to their increasing GHLI and one out of every two chemicals has been 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 GHLI). This splitting guarantees that the training set covers the entire range of the modeled 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 the GHL(modeled 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 [8,9]=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:

$Q^2LMO_{30\%}$= 0.85. High 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:

R^2y-sc= 0.04. 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:

The prediction set (50% of the whole set of chemicals, a strong splitting) 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 (50%) was realized by Ordered response. Chemicals have been ordered according to their increasing GHLI and one out of every two chemicals has been assigned to the prediction set. (see section 6.5)

7.7.Predictivity - Statistics obtained by external validation:

Q^2extF1 [10]= 0.83; Q^2extF2 [11]= 0.83; Q^2extF3 [12]= 0.84; CCCex=0.90; RMSE= 0.71

7.8.Predictivity - Assessment of the external validation set:

The prediction set, representing the half of the whole initial set (250 chemicals) is sufficiently large for a reliable external validation, containing 125 compounds. The splitting methodology based on ordered response (explained in section 6.5) 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.
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)

The applicability domain of the model on the prediction set has been verified by the Williams plot: 3 compounds on 125 of the prediction set are outliers for the response (not well predicted) and only 1 is a structural outlier (extrapolated, even if, in this case, verified as a good prediction). These results support the large applicability domain of the proposed model.

7.9.Comments on the external validation of the model:

The high values of external $Q^2$and CCC (threshold for accepting the external Q^2F1-F2-F3 is 0.70, threshold for CCC is 0.85, [ref 9; sect 9.2]), show that the proposed model is predictive, when applied to 125 chemicals not seen during the model development.

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 for this physical-chemical property was set a priori, but a mechanistic interpretation of molecular descriptors was provided a posteriori (see sect 8.2).

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori mechanistic interpretation:

The equation of the full model 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 model variables take account of 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. minsCl it is related to the presence of chlorine atoms in the molecule; it has a positive sign in the equation, as, usually, the substitution with chlorine confers greater resistance to biological and photolytic degradation, resulting in a higher persistence of the chemical.

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 large 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 GHLI for new chemicals without experimental data for persistence, it is suggested to apply the equation of the **Full Model**, developed on all the available chemicals (N=250).

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 Training set= 250; R^2= 0.85; Q^2LOO = 0.85; $Q^2LMO_{30\%}$= 0.84; CCC = 0.92; CCCcv = 0.91 ;RMSE= 0.687; RMSEcv = 0.704

9.2.Bibliography:

[1]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. DOI: 10.1021/es061773b

[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]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. url not available

[6]HyperChem 7.03, 2002 http://www.hyper.com/

[7]OpenBabel 2.3.2, 2012 http://openbabel.org

[8]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

[9]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

[10]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

[11]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

[12]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

9.3.Supporting information:

Training_GHLI _125.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q15-66-0018/attachment/A978
Prediction_GHLI_125.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q15-66-0018/attachment/A979
GHLI.sdf	http://qsardb.jrc.ec.europa.eu/qmrffile:///C:\Documents and Settings\lab-qsar\Desktop\QMRF da mandare 2015\GHLI padel\GHLI.sdf

Supporting information

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q15-66-0018

10.2.Publication date:

2015-06-12

10.3.Keywords:

PaDEL-Descriptor;Global Half-Life Index;GHLI;Persistence;QSARINS;INSUBRIA;

10.4.Comments: