QMRF identifier (JRC Inventory):	*To be entered by JRC*
QMRF Title:	*Insubria QSPR PaDEL-Descriptor model for organic carbon-sorption partition coefficient (logKoc) prediction.*
Printing Date:	*Jan 20, 2014*

1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSPR PaDEL-Descriptor model for organic carbon-sorption partition coefficient (logKoc) prediction.

1.2.Other related models:

Gramatica P., Giani E., Papa E., Statistical external validation and consensus modeling: A QSPR case study for Koc prediction, Journal of Molecular Graphics and Modeling, 2007, 25, 755-766. [9]

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:

8/10/2013

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]Stefano Cassani DiSTA, University of Insubria (Varese - Italy) +390332421439 stefano.cassani@uninsubria.it www.qsar.it
[2]Paola Gramatica DiSTA, University of Insubria (Varese - Italy) paola.gramatica@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

3.Defining the endpoint - OECD Principle 1

3.1.Species:

No information available

3.2.Endpoint:

2.Environmental fate parameters 2.6.Organic carbon-sorption partition coefficient (organic carbon; Koc)

3.3.Comment on endpoint:

The soil sorption partition coefficient is expressed as the ratio between chemical concentration in soil and in water, normalized on organic carbon (Koc). This parameter is an indicator of the sorption of chemicals by soils and sediments, thus providing an estimation of compound mobility and persistence in these compartments.The Koc experimental data for 643 heterogeneous organic compounds were collected from literature [2-4] and compiled into a single dataset. These three references were not the primary source of the experimental data, but a collection of previous literature data, which were already used to develope published and good-quality QSPR models.

3.4.Endpoint units:

Dimensionless

3.5.Dependent variable:

log Koc

3.6.Experimental protocol:

No information available

3.7.Endpoint data quality and variability:

As stated in section 3.3, we used data already satisfactorily modelled [2-4]: previous results are a proof of data quality. If more than one Koc value was available for a single compound, the average of the values was used.

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:

LogKoc model (Split Model)
MLR-OLS method. Model developed on a training set of 93 compounds.

LogKoc model (Full Model)
MLR-OLS method. Model developed on all the available experimental data (training set of 643 compounds).

Split model equation: logKoc= 0.63 + 0.28 VP-0 - 0.26 nHBAcc + 0.09 nAromBond - 0.19 MAXDP

Full model equation: logKoc= 0.87 + 0.26 VP-0 - 0.23 nHBAcc + 0.08 nAromBond - 0.19 MAXDP

4.3.Descriptors in the model:

[1]VP-0 Valence path, order 0

[2]nHBAcc Number of hydrogen bond acceptors (using CDK HBondAcceptorCountDescriptor algorithm)

[3]nAromBond Number of aromatic bonds

[4]MAXDP Maximum positive intrinsic state difference in the molecule (related to the electrophilicity of the molecule). Using deltaV = (Zv-maxBondedHydrogens) /(atomicNumber-Zv-1).

4.4.Descriptor selection:

A total of 681 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 169 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 variable. Then the GA was applied in order to explore new combinations of variables, selecting the variables 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.2.3.0, 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:

Split Model: 93 chemicals / 4 descriptors = 23.25
Full Model: 643 chemicals / 4 descriptors = 160.75

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 3 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 LogKoc values: -0.31 - 6.33.

Range of descriptor values: VP-0 (1.11 / 22.53), nHBAcc (0 / 10), nAromBond (0 / 29), MAXDP (0 / 6.82)

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*=0.023). 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, calculated as: $r'_i = r_i / s\sqrt{(1-h_{ii})}$, where $r_i = Y_i-\hat{Y}_i$.

5.3.Software name and version for applicability domain assessment:

QSARINS 1.2
Software for the development, analysis and validation of QSAR MLR models
paola.gramatica@uninsubria.it
www.qsar.it

5.4.Limits of applicability:

Split model domain: outliers for structure, hat>0.1613 (h*): Dibenzo(aj)anthracene (224-41-9), 1,2,5,6-dibenzanthracene (53-70-3), Hexabromobiphenyl (36355-01-8), Camphechlor (8001-35-2), Thifensulfuron-methyl (79277-27-3), Metsulfuron-methyl (74223-64-6), Indeno(1,2,3-cd)pyrene (193-39-5), Benzo(ghi)perylene (191-24-2), Tralomethrin (66841-25-6), C.I. Vat Yellow 4 (128-66-5), Chlordecone

(143-50-0), Dibenzo(a,i)pyrene (189-55-9), Mirex (2385-85-5). Outliers for response, standardised residuals > 3 standard deviation units: Methylurea (598-50-5), Endothal (145-73-3), Pentachloroaniline (527-20-8), Esfenvalerate (66230-04-4), Quizalofop-ethyl (76578-14-8), Camphechlor (8001-35-2), Isoxaben (82558-50-7), Hexabromobiphenyl (36355-01-8).
FULL

model domain: outliers for structure, hat>0.023 (h*): Cyromazine (66215-27-8), Thiophanate-methyl (23564-05-8), 1,2,5,6-dibenzanthracene (53-70-3), Dibenzo(aj)anthracene (224-41-9), Chlorsulfuron (64902-72-3), Indeno(1,2,3-cd)pyrene (193-39-5), Benzo(ghi)perylene (191-24-2), Hexabromobiphenyl (36355-01-8), Tralomethrin (66841-25-6), C.I. Vat Yellow 4 (128-66-5), Chlordecone (143-50-0), Camphechlor (8001-35-2), Chlorimuron-ethyl (90982-32-4), Dibenzo(a,i)pyrene (189-55-9), Sulfometuron-methyl (74222-97-2), Thifensulfuron-methyl (79277-27-3), Metsulfuron-methyl (74223-64-6), Mirex (2385-85-5). Outliers for response, standardised residuals > 3 standard deviation units: Isoxaben (82558-50-7).

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 93 heterogeneous organic compounds (including almost all the principal functional groups present mainly in pesticides and various organic pollutants) with a range of logKoc values from -0.31 to 6.02. Training and prediction set are structurally balanced, being the splitting based on the structural similarity analysis.

6.6.Pre-processing of data before modelling:

Transformation of Koc into logarithmic units (log Koc). If more than one value was available for a single compound, the average of the values was used. Only processed data are given.

6.7.Statistics for goodness-of-fit:

R^2= 0.84; CCCtr [5]=0.91; RMSE= 0.49

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

Q^2LOO= 0.82; CCCcv=0.90; RMSEcv= 0.52

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

Q^2LMO= 0.82

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 550 heterogeneous organic compounds with a range of logKoc values from 0 to 6.33. Training and prediction set are structurally balanced, being the splitting based on the structural similarity analysis.

7.6.Experimental design of test set:

The splitting of the original data set (643 compounds) into a training set of 93 compounds and a prediction set of 550 compounds was realized by Kohonen artificial neural network.

7.7.Predictivity - Statistics obtained by external validation:

Q^2extF1 [6]= 0.78; Q^2extF2 [7]= 0.78; Q^2extF3 [8]= 0.79; CCCex=0.89; RMSE= 0.57

7.8.Predictivity - Assessment of the external validation set:

The splitting methodology based on similarity analysis (performed by the application of the Kohonen maps Artificial Neural Networks - KANN) allowed for the selection of a meaningful training set and a representative prediction set.

Training and prediction set are balanced according to both structure and response. In particular, for response the range of logKoc values are [-0.31 - 6.02] and [0 - 6.33] respectively for training and prediction set. As much as concern structural representativity, the range of descriptors

values are:

VP-0: training set (2.27 / 18.17), prediction set (1.16 / 22.53)
nAromBonb: training set (0 / 26), prediction set (0 / 29)
MAXDP: training set (0.008 / 5.66), prediction set (0 / 6.82)
nHBAcc: training set (0 / 10), prediction set (0 / 10)

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 et al.[9] is:

log Koc= -1.92 + 2.07VED1 - 0.31nHAcc - 0.31MAXDP - 0.39CIC0

VED1= eigenvector coefficient sum for distance matrix
nHAcc= number of acceptor atoms for H-bonds (N,O,F)
MAXDP= maximal electrotopological positive variation
CIC0= complementary information content (neighborhood symmetry of 0-order)

The nHAcc descriptor, that is related to electronegative atoms of molecules, and MAXDP, related to molecule electrophilicity, represent different ways of taking into account the probability of bond formation between chemicals and groundwater: as expected, these descriptors are negative in sign as high affinity for water precludes soil sorption of the chemicals. The other two descriptors are related to molecular size, but their relevance is very different: the more important VED1 has a positive sign, highlighting that the bigger compounds are more sorbed than leached, the less relevant descriptor CIC0, added as the last variable in the nested models, is probably useful only to improve model quality in order to adapt some particular chemicals.

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

logKoc= 0.87 + 0.26 VP-0 - 0.23 nHBAcc + 0.08 nAromBond - 0.19 MAXDP

where VP-0: Valence path, order 0 nHBAcc: Number of hydrogen -bond acceptors

nAromBond: Number of aromatic bonds

MAXDP: Maximum positive intrinsic state

difference in the molecule (related to the electrophilicity of the

molecule).

Two descriptors are the same in both DRAGON and PaDEL models: nHBAcc descriptor and MAXDP (see above for meaning)

The other two descriptors (VP-0, the most important in the QSAR equation, and nAromBond) are related to molecular size and have positive signs: VP-0 highlights that the bigger compounds are more sorbed than leached, while the less relevant descriptor nAromBond is a counter for aromatic bonds.

8.3.Other information about the mechanistic interpretation:

No other information available

9.Miscellaneous information

9.1.Comments:

To predict logKoc for new chemicals without experimental data, it is suggested to apply the equation of the Full Model, developed on all the available chemicals (N=643), thus ensuring a wider applicability domain.

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

logKoc= 0.87 + 0.26 VP-0 - 0.23 nHBAcc + 0.08 nAromBond - 0.19 MAXDP

N = 643; R2 = 0.79; Q2 = 0.79; Q2LMO = 0.79; CCC = 0.89; CCCcv = 0.88 ;RMSE= 0.543; RMSEcv = 0.547

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]Sabljic A., et al. QSAR modeling of soil sorption. improvements and systematics of log Koc vs. log Kow correlations, Chemosphere 31 (1995) 4489–4514.

[3]Tao S., et al. Estimation of organic carbon normalized sorption coefficient (KOC) for soils using the fragment constant method, Environ. Sci. Technol. 33 (1999) 2719–2725.

[4]Huuskonen J., Prediction of soil sorption coefficient of a diverse set of organic chemicals from molecular structure, J. Chem. Inf. Comput. Sci. 43 (2003) 1457–1462.

[5]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

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

[7]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.

[8]Consonni V. et al. Comments on the Definition of the Q2 Parameter for QSAR Validation, J. Chem. Inf. Model. 49 (2009) 1669-1678
[9]Gramatica P., Giani E., Papa E., Statistical external validation and consensus modeling: A QSPR case study for Koc prediction, Journal of Molecular Graphics and Modeling, 2007, 25, 755-766.

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