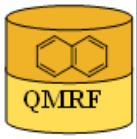	*QMRF identifier (ECB Inventory):* *To be entered by JRC* *QMRF Title:* *QSAR model for Lipophilicity (Octanol-water partition coefficient) for diverse organics* *Printing Date:* *May 3, 2010*	

1.QSAR identifier

1.1.QSAR identifier (title):

QSAR model for Lipophilicity (Octanol-water partition coefficient) for diverse organics

1.2.Other related models:

-

1.3.Software coding the model:

QSARModel 4.0.4 Molcode Ltd., Turu 2, Tartu, 51014, Estonia http://www.molcode.com

2.General information

2.1.Date of QMRF:

18.04.2010

2.2.QMRF author(s) and contact details:

Dana Martin, Indrek Tulp, Tarmo Tamm, Dimitar Dobchev, Gunnar Karelson, Jaak Jänes, Kaido Tämm, Eneli Härk, Deniss Savchenko, Andres Kreegipuu, Mati Karelson. Molcode model development team Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

Dana Martin, Indrek Tulp, Tarmo Tamm, Dimitar Dobchev, Gunnar Karelson, Jaak Jänes, Kaido Tämm, Eneli Härk, Deniss Savchenko, Andres Kreegipuu, Mati Karelson. Molcode model development team Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

2.6.Date of model development and/or publication:

18.04.2010

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

[1]M. Karelson, D. Dobchev, T. Tamm, I. Tulp, J. Jänes, K. Tämm, A. Lomaka, D. Savchenko, G. Karelson, Correlation of blood-brain penetration and human serum albumin binding with theoretical descriptors, ARKIVOC 16, 38-60 (2008).

[2]M. Karelson, G. Karelson, T. Tamm, I. Tulp, J. Jänes, K. Tämm, A. Lomaka, D. Savchenko, and D. Dobchev, QSAR study of pharmacological permeabilities, ARKIVOC 2, 218 – 238 (2009)

2.8.Availability of information about the model:

All information in full detail is available.

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

No other QMRF available for the same model

3.Defining the endpoint - OECD Principle 1

3.1.Species:

n/a

3.2.Endpoint:

1.Physicochemical effects 1.6.Octanol-water partition coefficient (Kow)

3.3.Comment on endpoint:

Annex (REACH) VII, test number 7.8 Partition coefficient n-octanol/water, flask shake method

3.4.Endpoint units:

n/a

3.5.Dependent variable:

logP

partition coefficient, the ratio of concentrations of a compound in two phases (water and octanol)

3.6.Experimental protocol:

Partition coefficient was determined using the OECD Test Guideline 107 (shake flask method). The partition coefficient (P or Kow) is defined as the ratio of the equilibrium concentrations (C) of a dissolved substance in a two-phase system consisting of two largely immiscible solvents. In the case n-octanol and water:

P = Kow = Cn-octanol/Cwater

The partition coefficient therefore is the quotient of two concentrations and is usually given in the form of its logarithm to base 10 (log P or log Kow).

The measuring range of the method is determined by the limit of detection of the analytical procedure. This should permit the assessment of values of log P in the range of -2 to 4 (occasionally when conditions apply, this range may be extended to log P up to 5) when the concentration of the solute in either phase is not more than 0,01 mol per litre [1-2]. The partitioning of organic compounds between aqueous and lipophilic phase is an important endpoint used extensively in medicinal chemistry, drug design, pharmacy, and environmental toxicity in predicting biological and hazardous effects of chemicals. No other physicochemical property has attracted as much interest in quantitative structure–activity relation (QSAR) studies as partition coefficient - lipophilicity (synonymously called hydrophobicity). The lipophilicity of the heterogeneous set of 223 compounds as used in [3] was chosen from the original compilation made by Hansch et al [4].

3.7.Endpoint data quality and variability:

Experimental data from different sources has been used. The previous successful modelling [3] supports consistency of the data.

Statistics:

max value: 6.63

min value: -3.21

standard deviation: 1.89

skewness: -0.06

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

2D and 3D regression-based QSAR Multilinear regression model based on 3-D quantum chemical descriptors.

4.2.Explicit algorithm:

multilinear regression QSAR

LogP= 0.83 -3.01* HASA-2/SQRT(TMSA) (AM1) +0.81* Kier&Hall index (order 1) -0.12* count of H-donor sites (AM1) (all)

4.3.Descriptors in the model:

[1]HASA-2/SQRT(TMSA) (AM1)

[2]Kier&Hall index (order 1)

[3]count of H-donor sites (AM1) (all)

4.4.Descriptor selection:

Initial pool of ~1000 descriptors. Stepwise descriptor selection based on a set of statistical selection rules (1-parameter equations: Fisher criterion and R2 over threshold, variance and t-test value over threshold, intercorrelation with another descriptor not over threshold), (2 parameter equations: intercorrelation coefficient bellow threshold, significant correlation with endpoint in terms of correlation coefficient and t-test). Stepwise trial of additional descriptors not significantly correlated to any already in the model.

4.5.Algorithm and descriptor generation:

1D, 2D, and 3D theoretical calculations

quantum chemical descriptors derived from AM1 calculation. Model developed by using multilinear regression.

4.6.Software name and version for descriptor generation:

QSARModel 4.0.4

Molcode Ltd., Turu 2, Tartu, 51014, Estonia

http://www.molcode.com

4.7.Chemicals/Descriptors ratio:

58.6(6) (176 chemicals / 3 descriptors)

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

By chemical identity: diverse set of organic compounds (alcohols, ketones, carboxylic acids, nitriles, halogeno-compounds, amines, aliphatic, aromatic and heteroaromatic compounds, etc)

By descriptor value range: this model is suitable for compounds that have the descriptors in the following range: HASA-2/SQRT(TMSA) (AM1) (min: 0.00, max:1.65) Kier&Hall index (order 1) (min: 0.00 , max: 10.04) count of H-donor sites (AM1) (all) (min: 0.00, max: 26.00).

5.2.Method used to assess the applicability domain:

presence of functional groups in structures

Range of descriptor values in training set with ±30% confidence

Descriptor values must fall between maximal and minimal descriptor values of training set ± 30%

5.3.Software name and version for applicability domain assessment:

QSARModel 4.0.4

Molcode Ltd., Turu 2, Tartu, 51014, Estonia

http://www.molcode.com

5.4.Limits of applicability:

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: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:

data points: 176, negative: 22, positive values: 154

6.6.Pre-processing of data before modelling:

6.7.Statistics for goodness-of-fit:

R2 = 0.92 (Correlation coefficient);
s = 0.55 (Standard error of the estimate);
F = 635.73 (Fisher function);

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

R2cv = 0.91 LOO;

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

R2cv = 0.91 LMO;

6.10.Robustness - Statistics obtained by Y-scrambling:

6.11.Robustness - Statistics obtained by bootstrap:

6.12.Robustness - Statistics obtained by other methods:

ABC analysis (2:1 training : prediction) on sorted data divided into 3 subsets (A;B;C). .Training set formed with 2/3 of the compounds (set A+B, A+C, B+C) and validation set consisted of 1/3 of the compounds (C, B, A)
average R2 (fitting) = 0.92
average R2 (prediction) = 0.915

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: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:

data points:19 , negative: 2, positive values: 17

7.6.Experimental design of test set:

The full experimental dataset was sorted according to increasing values of logP and each tenth compound was assigned to the test set.

7.7.Predictivity - Statistics obtained by external validation:

R2=0.84

7.8.Predictivity - Assessment of the external validation set:

The descriptors for the test set are in the limit of applicability

7.9.Comments on the external validation of the model:

The validation R2 for the test set is good.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The solubility in octanol depends on the H bonding capability of the molecule and on the shape of the molecule. The descriptors HASA-2/SQRT(TMSA) (AM1) and count of H-donor sites (AM1) (all) reflect the H bonding capability of the molecule. They appear in the model with negative sign which means that the solubility in octanol decreases with increasing of the H bonding capability of the molecule. This is expected, because the H bonding capability render the molecule more soluble in polar solvents like water than in non polar solvents like octanol. The presence of the descriptor Kier&Hall index (order 1) in the model indicates that the solubility depends also on the shape and size of the molecule.

8.2.A priori or a posteriori mechanistic interpretation:

a posteriori mechanistic interpretation, consistent with published scientific interpretations of experiments [1,3-4]

8.3.Other information about the mechanistic interpretation:

Kier&Hall index (order 1) is a topological index that describes the atomic connectivity in the molecule and gives information about the shape and the size of the molecule.

The descriptors HASA-2/SQRT(TMSA) (AM1) and count of H-donor sites (AM1) (all) reflect the H bonding capability of the molecule.

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1]Agrawal VK, Gupta M, Singh J, Khadikar PV., A novel method of estimation of lipophilicity using distance-based topological indices: dominating role of equalized electronegativity, Bioorg Med Chem. 2005 Mar 15;13(6):2109-20. http://dx.doi.org/10.1016/j.bmc.2005.01.003

[2]JRC Testing methods http://ecb.jrc.ec.europa.eu/testing-methods/

[3]Agrawal V. K., Gupta M., Singh J. and Khadikar P. V., 2005. A novel method of estimation of lipophilicity using distance-based topological indices: dominating role of equalized electronegativity. Bioorganic & Medicinal Chemistry 13, 2109–2120.
[4]Hansch, C., Leo, A. and Hockman, D., 1995. Exploring QSAR: Hydrolysis and Steric Constants; American Chemical Society: Washington DC.

9.3.Supporting information:

Training set(s)

Liphophilicity_trainingset.sdf	http://qsardb.jrc.ec.europa.eu:80/qmrf/download_attachment.jsp?name=qmrf205_Liphophilicity_trainingset.sdf

Test set(s)

Liphophilicity_testset.sdf	http://qsardb.jrc.ec.europa.eu:80/qmrf/download_attachment.jsp?name=qmrf205_Liphophilicity_testset.sdf

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