1.QSAR identifier

1.1.QSAR identifier (title):

QSAR ANN model for Persistence: Biotic degradation in water

1.2.Other related models:

QSAR for Persistence: Biotic degradation in water (submitted 05.05.2010)

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:

31.05.2012

2.2.QMRF author(s) and contact details:

[1]Tarmo Tamm Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[2]Gunnar Karelson Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[3]Dimitar Dobchev Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[4]Kaido Tämm Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[5]Jaak Jänes Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[6]Mati Karelson 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:

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:

22.05.2012

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

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

[2]Karelson M, Karelson G, Tamm T, Tulp I, Jänes J, Tämm K, Lomaka A,

Savchenko D & Dobchev D (2009). QSAR study of pharmacological permeabilities. ARKIVOC 2, 218–238.

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:

05.05.2010

3.Defining the endpoint - OECD Principle 1

3.1.Species:

n/a

3.2.Endpoint:

2.Environmental fate parameters 3.Persistence: Biodegradation 2.3.a.Ready/not ready biodegradability

3.3.Comment on endpoint:

The half-life is the time required for the concentration of a substance to halve its original value in a particular environmental medium. The half-lives of organic compounds are among the most commonly used criteria for studying persistence [1]. The semiquantitative data based on expert judgment and actual experimental values have already been suggested by Webster et al. [2] as preferable for half life identification, and are commonly used to develop the widely applied multimedia models [3,4]. In addition, a simple QSPR regression model has been demonstrated to be an useful tool for the identification and prioritization of existing or not yet synthesized potential persistent organic pollutants [5].

3.4.Endpoint units:

The half-life values (in h) were transformed into logarhitmic form for modelling

3.5.Dependent variable:

log T(0.5)

3.6.Experimental protocol:

The dataset of structurally heterogeneous and highly representative of many classes of already defined problematic chemicals includes 206 organic compounds of known half-lives for transformation into air [6].

3.7.Endpoint data quality and variability:

min value training set: 1.23
max value training set: 4.74
avrg training set: 2.63

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

2D and 3D regression-based QSAR

4.2.Explicit algorithm:

nonlinear regression QSAR
artificial neural networks model with architecture 6-5-1 trained with back

propagation of the error

4.3.Descriptors in the model:

[1]FPSA3 Fractional PPSA (PPSA-3/TMSA) (Zefirov) unitless fractional positive surface area
[2]Lowest resonance energy (AM1) for C - Cl bonds eV
[3]Molecular weight amu
[4]Negatively Charged Surface Area (Zefirov) C. A2
[5]Number of halogenide groups unitless
[6]Partial Charged (Zefirov) Surface Area of H atoms au.A2

4.4.Descriptor selection:

Initial consisted of ~1000 descriptors per structure. All descriptors were correlated with the property and then the first 20 descriptors were selected which indicated the largest correlation coeficient. Further, 11 networks with different structures and iputs form the preselected 20 descriptors were tested in order to find the best ANN with lowest RMS (root-mean-squared error) for training, selection and test sets. The best ANN model had 6 inputs. Then 206 epochs were used to train the final network with architecture depicted in 4.2. Optimization of the weights was performed with Levenberg-Marquardt algorithm using linear(inputs) and hyperbolic(hidden) and logistic(output) activation functions.

4.5.Algorithm and descriptor generation:

1D, 2D, and 3D theoretical calculations. Quantum chemical descriptors derived from AM1 calculation. Model developed by using the ANN depicted in 4.2

4.6.Software name and version for descriptor generation:

QSARModel
QSAR/QSPR package that will compute chemically meaningful descriptors and includes statistical tools for regression modeling
Molcode Ltd, Turu 2, Tartu, 51014, Estonia
www.molcode.com

4.7.Chemicals/Descriptors ratio:

21 (training set 126 chemicals / 6 descriptors)

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Applicability domain based on training set:

a) by chemical identity: diverse set of organic pollutants (aromatic, alphatic and cyclic amines, ketones, alcohols, esters, etc)

b) by descriptor value range: The model is suitable for compounds that have the descriptors
in the following minimal-maximal range:

Partial Charged (Zefirov) Surface Area of H atomsLowest resonance energy (AM1) for C - Cl bondsFPSA3 Fractional PPSA (PPSA-3/TMSA) (Zefirov)Number of halogenide groupsNegatively Charged Surface Area (Zefirov)Molecular weight min0.000-13.6490.0000.0009.38344.053

max0.0370.0000.0418.000440.931443.752

5.2.Method used to assess the applicability domain:

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

QSAR/QSPR package that will compute chemically meaningful descriptors and includes statistical tools for regression modeling

Molcode Ltd, Turu 2, Tartu, 51014, Estonia

http://www.molcode.com

5.4.Limits of applicability:

See 5.1

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:

126 data points

6.6.Pre-processing of data before modelling:

normalization by the min max values

6.7.Statistics for goodness-of-fit:

Data Mean2.58

Data S.D.0.77

Error S.D.0.39

S.D. Ratio0.50

Correlation0.87

RMS0.11

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

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

6.10.Robustness - Statistics obtained by Y-scrambling:

n/a

6.11.Robustness - Statistics obtained by bootstrap:

n/a

6.12.Robustness - Statistics obtained by other methods:

Validation test statistics

Data Mean 2.43
Data S.D. 0.70
Error S.D. 0.53
S.D. Ratio 0.76
Correlation 0.70
RMS 0.16

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:

Validation test(selection test) - 20 data points, this set is used to control and monitor the training of the ANN model

External tes set - 20 datapoints

7.6.Experimental design of test set:

From sorted source data, 20 data points were subjected to the test set using selection according to the data distribution

7.7.Predictivity - Statistics obtained by external validation:

Data Mean2.54
Data S.D.0.81
Error S.D.0.67
S.D. Ratio0.83
Correlation0.57
RMS0.19

7.8.Predictivity - Assessment of the external validation set:

7.9.Comments on the external validation of the model:

In addition to the training set the ANN uses also validation (or selection) set to control and monitor the RMS error during training and external test set

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

Because of the nonlinear nature of the ANNs deeper analysis of the descriptor is difficult compared to the normal multilinear analysis. The ANN model descriptors are mainly related to the charge ditributions of the compounds e.g. FPSA3 Fractional PPSA (PPSA-3/TMSA) (Zefirov), Negatively Charged Surface Area (Zefirov), Partial Charged (Zefirov) Surface Area of H atoms.

In addition the halogens reactivity plays also important role for the water degradation. Generally the halogens have larger LogT values.

8.2.A priori or a posteriori mechanistic interpretation:

a posteriori mechanistic interpretation,

8.3.Other information about the mechanistic interpretation:

Similar interpretation can be found in scientific literature [5]

9.Miscellaneous information

9.1.Comments:

The data are gathered from handbook (Physical-Chemical Properties and Environmental Fate Handbook) which includes data from different sources. Therefore the experimental protocol cannot be prvided. The data were also semiquantitatively classified as proposed by Mackay [1]. [1] Webster, E.; Mackay, D.; Wania, F. Evaluating Environmental Persistence. Environ. Toxicol. Chem. 1998, 17, 2148-2158.

9.2.Bibliography:

[1]UNEP, Stockholm Convention on Persistent Organic Pollutants, United Nations Environment Program , Geneva , Switzerland , 200 http://www.pops.int

[2]Webster, E.; Mackay, D.; Wania, F. Evaluating Environmental Persistence, Environ. Toxicol. Chem. 1998, 17, 2148–2158

[3]Klasmeier, J.; Matthies, M.; MacLeod, M.; Fenner, K.; Scheringer, M.; Stroebe, M.; Le Gall, A. C.; McKone, T.; Van De Meent, D.; Wania, F. Application of Multimedia Models for Screening Assessment of Long-Range Transport Potential and Overall Persistence, Environ. Sci. Technol. 2006, 40, 53-60.

[4]Fenner, K; Scheringer, M.; Macleod, M.; Matthies, M.; McKone, T.; Stroebe, M.; Beyer, A.; Bonnell, M.; Le Gall, A. C.; Klasmeier, J.; Mackay, D.; Van de Meent, D.; Pennington, D.; Scharenberg, B.; Suzuki, N.; Wania, F. Comparing Estimates of Persistence And Long-Range Transport Potential among Multimedia Models, Environ. Sci. Technol. 2005, 39, 1932-1942

[5]Gramatica, P.; Papa, E. Screening and ranking of POPs for global halflife: QSAR approaches for prioritization based on molecular structure, Environ. Sci. Technol. 2007, 41, 2833–2839

[6]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

9.3.Supporting information:

Training set(s)Test set(s)Supporting information

10.Summary (ECB Inventory)

10.1.QMRF number:

10.2.Publication date:

10.3.Keywords:

10.4.Comments: