1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSPR PaDEL-Descriptor model for Melting Point prediction of Polybrominated Diphenyl Ethers.

1.2.Other related models:

E. Papa, S. Kovarich, P. Gramatica, 2009, Validation and Inspection of the Applicability Domain of QSPR Models for Physicochemical Properties of Polybrominated Diphenyl Ethers, QSAR & Comb.Sci. 28, 790-796. [10]

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:

21/11/2013

2.2.QMRF author(s) and contact details:

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

[2]Alessandro Sangion DiSTA, University of Insubria (Varese - Italy) +390332421439 a.sangion@hotmail.it www.qsar.it

[3]Paola Gramatica DiSTA, University of Insubria (Varese - Italy) paola.gramatica@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 other information available

3.Defining the endpoint - OECD Principle 1

3.1.Species:

No information available

3.2.Endpoint:

1.Physicochemical effects 1.1.Melting point

3.3.Comment on endpoint:

The melting point (MP) of a solid is the temperature range at which it changes state from solid to liquid. At the melting point the solid and liquid phase exist in equilibrium.

3.4.Endpoint units:

Dimensionless

3.5.Dependent variable:

MP

3.6.Experimental protocol:

Experimentally measured melting temperatures for 25 PBDEs (Polybrominated Diphenyl Ethers) were collected from 4 different sources: Tittlemier et al, 2002 (data for 11 PBDEs), Marsh et al., 1999 (data for 18 PBDEs), Palm et al., 2002 (data for 5 PBDEs), Kuramochi et al., 2007 (data for 4 PBDEs) [2-5]. When more than one experimental value was available for a single compound, the average value was used as input data for the development of the QSPR model.

3.7.Endpoint data quality and variability:

The availability of experimental data from different sources made it possible to verify the data quality and the variability between different laboratories (data reproducibility). When more than one experimental value was available for a single compound, the variation of data was quantified by calculation of standard deviation (smax = 7.601) and coefficient of variation (CV% = 0.9-6.7%).

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR - Multiple linear Regression Model (OLS - Ordinary least-squares)

4.2.Explicit algorithm:

LogKoa (split model)

OLS-MLR method. Model developed on a training set of 20 compounds

LogKoa (full model)

OLS-MLR method. Model developed on all the available experimental data (training set of 25 compounds).

Split Model: MP= 44.07 + 195.20 SC-5

Full Model: MP= 45.02 + 194.95 SC-5

4.3.Descriptors in the model:

SC-5 Simple cluster, order 5

4.4.Descriptor selection:

A total of 672 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 131 molecular descriptors were used as input variables for variable subset selection. The models were developed by the all-subset-procedure with only one variable. 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: 20 chemicals / 1 descriptor = 20

Full Model: 25 chemicals / 1 descriptor = 25

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 experimental MP values: 48.5 / 206
Range of descriptor values: SC-5: 0 / 0.84

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*=0.240). 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>0.300 (h*): no. Outliers for response, standardised residuals > 2.5 standard deviation units: 2,4,4',6-TetraBDE (189084-63-7). **FULL model domain**: outliers for structure, hat>0.240 (h*): no. Outliers for response, standardised residuals > 2.5 standard deviation units: 2,4,4',6-TetraBDE (189084-63-7)

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 20 PBDEs; training and test set are structurally belanced, being the splitting based on the structural similarity analysis.

6.6.Pre-processing of data before modelling:

Raw data, collected from 4 different references (Tittlemier et al. (2002), Marsh et al. (1999), Palm et al. (2002) and Kuramochi et al. (2007) [2-5], have been combined and mediated (if more than one value was available for the same compound) before modelling.

6.7.Statistics for goodness-of-fit:

R^2= 0.79; CCCtr [6]=0.88; RMSE= 21.94

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

Q^2LOO= 0.75; CCCcv=0.86; RMSEcv= 24.15

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

Q^2LMO= 0.70.

6.10.Robustness - Statistics obtained by Y-scrambling:

R^2y-sc= 0.06

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 validation set of the Split Model consists of 5 compounds , with a range of MP: 57.5 / 184.

7.6.Experimental design of test set:

The splitting of the original data set (25 compounds) into a training set of 20 compounds (representative of the entire data set) and a validation set of 5 compounds (splitting 20%) was realized by applying Self Organized Maps Kohonen Artificial Neural Networks (SOM K-ANN).

7.7.Predictivity - Statistics obtained by external validation:

Q^2extF1 [7]= 0.91; Q^2extF2 [8]= 0.89; Q^2extF3 [9]= 0.92; CCCex=0.94; RMSE= 13.92

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 MP values are [48.5 / 206] and [57.5 / 184] respectively for training and prediction set. As much as concern structural representativity, the range of descriptors values is:

SC-5: training set (0.12 / 0.84), prediction set (0 / 0.60)

7.9.Comments on the external validation of the model:

no other 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 equation published in Papa et al. [10] was:

MP= 1968.06 - 6227.09 X2A

where X2A is the average connectivity index chi-2 and brings information related to molecular dimension and was found to be inversely related to the increase of melting points.

IThe equation of the new PaDEL-descriptor model included in QSARINS is : MP= 45.02 + 194.95 SC-5 where SC-5 is simple cluster descriptor of order 5, which values increases with the number of Bromine atoms, thus indicating that the MP increases with the increasing of the number of bromine substituents.

The correlation between X2A and SC-5 is negative: -0.97.

8.3.Other information about the mechanistic interpretation:

no other information available

9.Miscellaneous information

9.1.Comments:

To predict MP for new chemicals without experimental data, it is suggested to apply the equation of the Full Model, developed on all the available chemicals (N=25), thus ensuring a wider applicability domain.

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

$$MP= 45.02 + 194.95\ SC\text{-}5$$

N = 25; R2 = 0.82; Q2 = 0.79; Q2LMO = 0.75; CCC = 0.90; CCCcv = 0.88; RMSE= 20.57; RMSEcv = 22.16

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]S. A. Tittlemier, T. Halldorson, G. A. Stern, G. T. Tomy, Environ. Toxicol. Chem. 2002, 21, 1804 – 1810.

[3]G. Marsh, J. Hu, E. Jakobsson, S. Rahm, A. Bergman, Environ. Sci. Technol. 1999, 33, 3033 – 3037.

[4]A. Palm, I. T. Cousins, D. Mackay, M. Tysklinnd, M. Alaee, Environ. Pollut. 2002, 117, 195 – 213.

[5]H. Kuramochi, K. Maeda, K. Kawamoto, Chemosphere 2007, 67, 1858 – 1865.

[6]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

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

[8]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.

[9]Consonni V. et al. Comments on the Definition of the Q2 Parameter for QSAR Validation, J. Chem. Inf. Model. 49 (2009) 1669-1678

[10]E. Papa, S. Kovarich, P. Gramatica, 2009, Validation and Inspection of

the Applicability Domain of QSPR Models for Physicochemical Properties of Polybrominated Diphenyl Ethers, QSAR & Comb.Sci. 28, 790-796.

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