1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSAR PaDEL-Descriptor model for the prediction of chemicals PBT behaviour (PBT Index)

1.2.Other related models:

E.Papa and P.Gramatica, 2010. QSPR as a support for the EU REACH regulation and rational design of environmentally safer chemicals: PBT identification from molecular structure, Green Chem. 12, pp 836-843 (Hot Article) [9]

1.3.Software coding the model:

[1]PaDEL-Descriptor 2.18 A software to calculate molecular descriptors and fingerprints Chun Wei Yap 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:

10/12/2013

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:

January 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:

Environmental Fate parameters Persistence - Bioaccumulation - Toxicity PBT Index

3.3.Comment on endpoint:

The PBT Index is a macro-variable which condenses the chemical tendency to environmental persistency, bioaccumulation and (eco)toxicity. It is derived by Principal Component Analisys (PCA) from half-life, BCF and *P.promelas* toxicity experimental and reliable predicted data for a set of 180 organic chemicals. The scores of the compounds along PC1, which provides alone the largest part (77.1%) of the total information, defined the cumulative PBT Index; this index ranks the compounds according to their cumulative Persisten, Bioaccumulative and Toxic behaviour.

3.4.Endpoint units:

GHLI [2], log BCF(experimental and predicted, [3]) and *Pimephales promelas* pLC_{50} values [4] were combined by Principal Component Analisys. The PBT Index, obtained by PCA (PC1 values), is an adimensional endpoint.

3.5.Dependent variable:

PBT Index

3.6.Experimental protocol:

The whole training set includes 180 organic compounds; experimental values for 54 chemicals were taken from literature [2-4], while the rest of the dataset was composed of reliable predicted data.

3.7.Endpoint data quality and variability:

No other information available

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR - Multiple linear regression model (OLS - Ordinary Least Square)

4.2.Explicit algorithm:

PBT Index Split model

MLR-OLS method. Model developed on a training set of 92 compounds.

PBT Index Full model

MLR-OLS method. Model developed on a training set of 180 compounds.

Split model equation: PBT Index = -1.42 + 0.65 nX + 0.22 nBondsM - 0.41 nHBDon_Lipinksi - 0.09 MAXDP2

Full model equation: PBT Index = -1.46 + 0.64 nX + 0.22 nBondsM - 0.39 nHBDon_Lipinksi - 0.06 MAXDP2

4.3.Descriptors in the model:

[1]nX Number of halogen atoms (F, Cl, Br, I, At, Uus)

[2]nBondsM Total number of bonds that have bond order greater than one (aromatic bonds have bond order 1.5).

[3]nHBDon_Lipinksi 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]MAXDP2 Maximum positive intrinsic state difference in the molecule (related to the electrophilicity of the molecule). Using deltaV = Zv-maxBondedHydrogens.

4.4.Descriptor selection:

Hundreds of molecular descriptors were calculated with PaDEL-Descriptor 2.18. Taking into account the DRAGON descriptors involved in the original PBT Index model, we then manually selected the same four variables (called in PaDEL-Descriptor with slighly different names) encoding the PBT Index: nX (same name in DRAGON), nBondsM (nBM in DRAGON), nHBDon_Lipinski (nDon in DRAGON) and MAXDP2 (MAXDP in DRAGON).

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. For two chemicals (79-10-7, 80-62-6), SMILES notation (taken from PubChem) were used instead of MDL-MOL format for the calculation of descriptors; this step was necessary to avoid some problems arising from the inaccurate conversion between HyperChem and MDL-MOL files for these two molecules.

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: 92 chemicals / 4 descriptors = 23
Full Model: 180 chemicals / 4 descriptors = 45

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 PBT-Index values: -3.08 / 5.02
Range of descriptor values: nX (0 / 6), nBondsM (0 / 16), nHBDon_Lipinski (0 / 2), MAXDP2 (0 / 5.24)

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*=0.083). 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: $r'i = ri / s\sqrt{}(1-hii)$, where $ri = Yi-\hat{Y}i$.

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>0.163 (h*): no. Outliers for response, standardised residuals > 2.5 standard deviation units: quinoline (91-22-5), N-nitrosodiphenylamine (86-30-6), benzophenone (119-61-9). **FULL model domain**: outliers for structure, hat>0.083 (h*): no. Outliers for response, standardised residuals > 2.5 standard deviation units: quinoline (91-22-5), N-nitrosodiphenylamine (86-30-6), benzophenone (119-61-9).

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 92 compounds with a range of PBT Index from -3.08 to 5.02. The splitting was based structural similarity (ordering the chemicals on PC1 Score).

6.6.Pre-processing of data before modelling:

GHLI, log BCF(experimental and predicted) and Pimephales promelas pLC50 values were combined by Principal Component Analisys. The PBT Index, obtained by PCA (PC1 values), is an adimensional endpoint.

6.7.Statistics for goodness-of-fit:

R^2= 0.89; CCCtr [5]=0.94; RMSE= 0.52

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

Q^2LOO= 0.88; CCCcv=0.93; RMSEcv= 0.55

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

Q^2LMO= 0.87

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 88 compounds with a range of PBT Index from -2.94 to 3.87

7.6.Experimental design of test set:

The splitting of the original data set (180 compounds) into a training set of 92 compounds and a prediction set of 88 compounds was realized by ordering PC1 Score (after PCA analysis of the modeling descriptors)

7.7.Predictivity - Statistics obtained by external validation:

Q^2extF1 [6]= 0.89; Q^2extF2 [7]= 0.89; Q^2extF3 [8]= 0.90; CCCex=0.94; RMSE= 0.49

7.8.Predictivity - Assessment of the external validation set:

The splitting methodology based on ordered PC1 score 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 PBT Index are [-3.08 / 5.02] and [-2.94 / 3.87] respectively for training and prediction set. As much as concern structural representativity, the range of descriptors values are:

nX: training set (0 / 6), prediction set (0 / 6)
nBondsM: training set (0 / 15), prediction set (0 / 16)
nHBDon_Lipinski: training set (0 / 2), prediction set (0 / 2)
MAXDP2: training set (0.04 / 5.19), prediction set (0 / 5.24)

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 Papa and Gramatica [9] is: PBT Index = -1.44 + 0.65 nX + 0.22 nBM - 0.39 nHDon - 0.07 MAXDP Where nX: number of halogen atoms nBM: number of multiple bonds nHDon: number of donor atoms forHbonds MAXDP: maximal electrotopological positive variation The equation of the new PaDEL-descriptor model included in QSARINS is : PBT Index = -1.46 + 0.64 nX + 0.22 nBondsM - 0.39 nHBDon_Lipinksi - 0.06 MAXDP2 Where nX: Number of halogen atoms (F, Cl, Br, I, At, Uus) nBondsM: Total number of bonds that have bond order greater than one (aromatic bonds have bond order 1.5). 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)MAXDP2: Maximum positive intrinsic state difference in the molecule (related to the electrophilicity of the molecule). Using deltaV = Zv-maxBondedHydrogens. The two most important descriptors, nX and nBondsM, which encode for substitution with halogens and unsaturation, are known to increase the PBT behaviour of chemicals. On the contrary, MAXDP2 and nHBDon_Lipinski are inversely related to the PBT Index. These last two descriptors are related to a compound’s ability to form electrostatic and dipole–dipole interactions, as well as hydrogen bonds in the surrounding media.

8.3.Other information about the mechanistic interpretation:

No other information available

9.Miscellaneous information

9.1.Comments:

To predict PBT Index for new chemicals, it is strongly suggested to apply the equation of the Full Model, developed on all the available chemicals (N=180), thus ensuring a wider applicability domain.

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

PBT Index = -1.46 + 0.64 nX + 0.22 nBondsM - 0.39 nHBDon_Lipinksi - 0.06 MAXDP2

N = 180; R2 = 0.89; Q2 = 0.88; Q2LMO = 0.88; CCC = 0.94; CCCcv = 0.94 ;RMSE= 0.51; RMSEcv = 0.52

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]P. Gramatica and E. Papa, Screening and Ranking of POPs for Global

Half-Life: QSAR Approaches for Prioritization Based on Molecular Structure, Environ. Sci. Technol., 2007, 41, 2833.

[3]P. Gramatica and E. Papa, An Update of the BCF QSAR Model Based on Theoretical Molecular Descriptors, QSAR Comb. Sci., 2005, 24, 953.

[4]E. Papa, F. Villa and P. Gramatica, Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales promelas (Fathead Minnow), J. Chem. Inf. Model., 2005, 45, 1256.

[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]E.Papa and P.Gramatica, 2010. QSPR as a support for the EU REACH regulation and rational design of environmentally safer chemicals: PBT identification from molecular structure, Green Chem. 12, pp 836-843 (Hot Article)

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