*QMRF identifier (JRC Inventory):*To be entered by JRC

*QMRF Title:*QSARINS (QSAR-INSUBRIA) model of PBT Index by PaDEL descriptors
Keywords: PaDEL-Descriptor; QSAR; PBT; Prioritization; Benign by design; QSARINS; INSUBRIA

*Printing Date:*9-mar-2015

1.QSAR identifier

1.1.QSAR identifier (title):

QSARINS (QSAR-INSUBRIA) model of PBT Index by PaDEL descriptors
Keywords: PaDEL-Descriptor; QSAR; PBT; Prioritization; Benign by design;
QSARINS; INSUBRIA

1.2.Other related models:

E.Papa and P.Gramatica, QSPR as a support for the EU REACH regulation and rational design of environmentally safer chemicals: PBT identification from molecular structure, Green Chem. 2010, 12, 836-843 (selected as Hot Article) [1]

1.3.Software coding the model:

PaDEL-Descriptor
A software to calculate molecular descriptors and fingerprints [2], version 2.18
Yap Chun Wei, email: phayapc@nus.edu.sg
http://padel.nus.edu.sg/software/padeldescriptor/index.html

QSARINS
Software for the development, analysis and validation of QSAR MLR models [3,4]. Version 1.2 (verified also with version 2.2, 2015)
Paola Gramatica, email: paola.gramatica@uninsubria.it
www.qsar.it

2.General information

2.1.Date of QMRF:

30/01/2015

2.2.QMRF author(s) and contact details:

[1]Paola Gramatica Insubria University, Department of Theoretical and Applied Sciences (DiSTA), via J.H. Dunant 3, 21100 Varese (Italy) +390332421573 paola.gramatica@uninsubria.it http://www.qsar.it/

[2]Stefano Cassani Insubria University, Department of Theoretical and Applied Sciences (DiSTA), via J.H. Dunant 3, 21100 Varese (Italy) +390332421439 stefano.cassani@uninsubria.it http://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 Insubria University, Department of Theoretical and Applied Sciences (DiSTA), via J.H. Dunant 3, 21100 Varese (Italy) +390332421439 stefano.cassani@uninsubria.it http://www.qsar.it/

[2]Paola Gramatica Insubria University, Department of Theoretical and Applied Sciences (DiSTA),

via J.H. Dunant 3, 21100 Varese (Italy) +390332421573 paola.gramatica@uninsubria.it
http://www.qsar.it/

2.6.Date of model development and/or publication:

Developed in 2012, Published in 2014

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

[1]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, 2010, 836-843 (selected as Hot Article) DOI: 10.1039/B923843C

[2]Yap, C.W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. 2011, J.Comput.Chem. 32, 1466-1474 doi: 10.1002/jcc.21707

[3]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. DOI: 10.1002/jcc.23361

[4]Gramatica P., et al. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS, J. Comput. Chem. (Software News and Updates), 2014, 35 (13), 1036-1044. DOI: 10.1002/jcc.23576

2.8.Availability of information about the model:

Non-proprietary. Defined algorithm, available in QSARINS [3,4]. Training and prediction sets are available in the attached sdf files of this QMRF (see section 9).

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 PBT Index

3.3.Comment on endpoint:

The PBT Index is a macro-variable which condenses the chemical cumulative 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 heterogeneous organic chemicals. The scores of the compounds along PC1, which provides alone the largest part (77.1%) of the total information, defined the PBT Index; this index ranks the compounds according to their cumulative Persistent, Bioaccumulative and Toxic behavior.

3.4.Endpoint units:

GHLI [5], log BCF(experimental and predicted, [6]) and*Pimephales promelas*pLC_{50}values [7] were combined by Principal Component Analisys. The final endpoint, the PBT Index obtained by PCA (PC1 values), is thus adimensional.

3.5.Dependent variable:

PBT Index (PC1 values)

3.6.Experimental protocol:

The whole training set includes 180 organic compounds; experimental values for 54 chemicals were taken from literature (our previous papers [5-7], see section 3.4) while the rest of the dataset was composed of reliable predicted data (interpolated predictions, within the Applicability Domain of the models).

3.7.Endpoint data quality and variability:

As stated in section 3.6, we took data already used and modeled, verified for their goodness (data curation) [5-7]. Previous results are also a proof of data quality.

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 (N Training: 92) : PBT Index = -1.42 + 0.65 nX + 0.22 nBondsM - 0.41 nHBDon_Lipinksi - 0.09 MAXDP2

Full model equation (N Training: 180): PBT Index = -1.46 + 0.64 nX + 0.22 nBondsM - 0.39 nHBDon_Lipinksi - 0.06 MAXDP2

The four modeling descriptors, calculated with the open source PaDEL-Descriptor software, are: nX (number of halogen atoms), nBondsM (number of bonds that have bond order greater than one, where aromatic bonds have bond order 1.5), nHBdon_Lipinski (number of hydrogen bond donors using Lipinski's definition, see section 4.3) and MAXDP2 (Maximum positive intrinsic state difference in the molecule). See section 4.3 for a more detailed explanation of the descriptors.

4.3.Descriptors in the model:

[1]nX dimensionless Number of halogen atoms (F, Cl, Br, I, At, Uus), encodes for substitution with halogens and it is known to increase the PBT behaviour of chemicals.

[2]nBondsM dimensionless Total number of bonds that have bond order greater than one (aromatic bonds have bond order 1.5). Encodes for unsaturation and it is known to increase the PBT behaviour of chemicals.

[3]nHBDon_Lipinksi dimensionless Number of hydrogen bond donors (using Lipinski's definition: Any OH or NH. Each available hydrogen atom is counted as one hydrogen bond donor). It is inversely related to the PBT Index and encodes for a compound’s ability to form hydrogen bonds in the surrounding media

[4]MAXDP2 dimensionless Maximum positive intrinsic state difference in the molecule, using deltaV = Zv-maxBondedHydrogens. It takes into account the electronic distribution in the topological graph and is related to molecule electrophilicity. It is inversely related to the PBT Index and encodes for a compound’s ability to form electrostatic and dipole–dipole interactions.

4.4.Descriptor selection:

Hundreds of molecular descriptors were calculated with PaDEL-Descriptor 2.18 [2]. Taking into account the DRAGON [8] descriptors involved in the original PBT Index model [1], we then decided to manually selected the same four variables (included 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 7.03 [9]. Then, these files were converted by OpenBabel 2.3.2 [10] 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
A software to calculate molecular descriptors and fingerprints, version 2.18
Yap Chun Wei, email: phayapc@nus.edu.sg
http://padel.nus.edu.sg/software/padeldescriptor/index.html

HyperChem
Software for molecular drawing and conformational energy optimization, version 7.03
Phone: (352)371-7744
http://www.hyper.com/

OpenBabel
Open Babel: The Open Source Chemistry Toolbox. Used for conversion between HYPERCHEM files (hin) and MDL-MOL files. Version 2.3.2
http://openbabel.org/wiki/THANKS
http://openbabel.org/wiki/Main_Page

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 (with the Insubria graph, included in QSARINS): 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:

$H = X(X^TX)^{-1}X^T$

The response applicability domain can be verified by the standardized residuals in cross-validation greater than 2.5 standard deviation units

5.3.Software name and version for applicability domain assessment:

QSARINS

Software for the development, analysis and validation of QSAR MLR models. Versiion 1.2 (verified also with version 2.2, 2015)

Paola Gramatica, email: paola.gramatica@uninsubria.it

http://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: No

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: after a PCA analysis, in the space of descriptors calculated in PaDEL-Descriptor 2.18, we ordered the PC1 score and selected, out of every two chemicals, a compound for its inclusion in the prediction set. After this, we can sat yhat training and prediction set are structurally balanced, being the splitting based on the structural similarity analysis (PC1 score information).

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 [11,12]=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:

$Q^2LMO_{30\%}$= 0.87. High value of Q^2LMO (average value for 2000 iterations, with 30% of chemicals put out at every iteration) means that the model is robust and stable.

6.10.Robustness - Statistics obtained by Y-scrambling:

R^2y-sc= 0.04. Low value of scrambled R^2(average value for 2000 iterations, in where the Y-responses are randomly scrambled), means that the model is not given by chance-correlation.

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: No

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 a PCA analysis of the descriptors, see section 6.5).

7.7.Predictivity - Statistics obtained by external validation:

Q^2extF1 [13]= 0.89; Q^2extF2 [14]= 0.89; Q^2extF3 [15]= 0.90; CCCex=0.94; RMSE= 0.49.

The high values of external $Q^2$and concordance correlation coefficient-CCC (threshold for accepting the external Q^2F1-F2-F3 is 0.70, threshold for CCC is 0.85, [11]), show that the proposed model is predictive, when applied to 88 chemicals never seen during the model development.

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.The prediction set is sufficiently large, consisting of 88 compounds (92 in training set) and thus representing the half of the whole initial set (180 chemicals).

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)

The applicability domain of the model on the prediction set has been verified by the Williams plot: only 1compound on 88 of the prediction set is outlier for the response (not well predicted) and no structural outliers are present. These results are a prrof of the large applicability domain of the proposed PBT Index model.

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 PBT cumulative property was set a priori, but a mechanistic interpretation of the four molecular descriptors was provided a posteriori (see 8.2).

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori mechanistic interpretation:

The equation of the full model, included in QSARINS 2.2, for the prediction of the cumulative PBT behavior of chemicals, is the following:

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:

Given the good results of the external validation, this model has a large applicability domain and therefore unsuccessful applications are probably very reduced. Anyhow, the check of outliers by the Williams plot and the Insubria graph for chemicals without experimental data (see section 5.1) will allow to verify the model applicability.

To predict the cumulative PBT Index for new chemicals without experimental data for P, B and T, it is suggested to apply the equation of the **Full Model**, developed on all the available chemicals (N=180).

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 Training set= 180; R^2= 0.89; Q^2LOO = 0.88; Q^2LMO$_{30\%}$= 0.88; CCC = 0.94; CCCcv = 0.94

;RMSE= 0.51; RMSEcv = 0.52

9.2.Bibliography:

[1]Papa E. and Gramatica P., QSPR as a support for the EU REACH regulation and rational design of environmentally safer chemicals: PBT identification from molecular structure, Green Chem. 2010, 12, 836-843 (Hot Article) DOI: 10.1039/B923843C
[2]Yap, C.W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J.Comput.Chem. 2011, 32, 1466-1474. doi: 10.1002/jcc.21707
[3]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. DOI: 10.1002/jcc.23361
[4]Gramatica P., et al. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS, J. Comput. Chem. (Software News and Updates), 2014, 35 (13), 1036-1044. DOI: 10.1002/jcc.23576
[5]Gramatica P. and Papa E., Screening and Ranking of POPs for Global Half-Life: QSAR Approaches for Prioritization Based on Molecular Structure, Environ.Sci.Technol. 2007, 41, 2833-2839 DOI: 10.1021/es061773b
[6]Gramatica P. and Papa E., An Update of the BCF QSAR Model Based on Theoretical Molecular Descriptors, QSAR Comb. Sci., 2005, 24, 953. DOI: 10.1002/qsar.200530123
[7]Papa E., Villa F. and Gramatica P., 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. DOI: 10.1021/ci050212l
[8]DRAGON for Windows (Software for molecular descriptors calculation) ver.5.5, Talete srl, Milano, Italy, 2007 http://www.talete.mi.it/
[9]HyperChem 7.03, 2002 http://www.hyper.com/
[10]OpenBabel 2.3.2, 2012 http://openbabel.org
[11]Chirico N. and Gramatica P., Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, J. Chem. Inf. Model. 2011, 51, 2320-2335. doi: 10.1021/ci200211n
[12]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, 2044–2058 DOI: 10.1021/ci300084j
[13]Shi L.M. et al. QSAR Models Using a Large Diverse Set of Estrogens, J. Chem. Inf. Comput. Sci. 2001, 41, 186–195. DOI: 10.1021/ci000066d
[14]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. 2008, 48, 2140-2145. doi: 10.1021/ci800253u
[15]Consonni V. et al. Comments on the Definition of the Q2 Parameter for QSAR Validation, J. Chem. Inf. Model. 2009, 49, 1669-1678 DOI: 10.1021/ci900115y

9.3.Supporting information:

Training set(s)

PBT Index training set.sdf	file:///C:\Documents and Settings\lab-qsar\Desktop\QMRF to send 2015\PBT Index PaDEL\PBT Index training set.sdf

Test set(s)

PBT Index prediction set.sdf	file:///C:\Documents and Settings\lab-qsar\Desktop\QMRF to send 2015\PBT Index PaDEL\PBT Index prediction set.sdf

Supporting information

PBT Index full.sdf	file:///C:\Documents and Settings\lab-qsar\Desktop\QMRF to send 2015\PBT Index PaDEL\PBT Index full.sdf

10.Summary (JRC QSAR Model Database)

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