QSPR as a support for the EU REACH regulation and rational design of environmentally safer chemicals: PBT identification from molecular structure†

Ester Papa and Paola Gramatica*

Received 13th November 2009, Accepted 27th January 2010
First published as an Advance Article on the web 5th March 2010
DOI: 10.1039/b923843c

The chemicals that are jointly Persistent, Bioaccumulative and Toxic (PBT) are substances of very high concern (SVHC) and subject to an authorization step in the new European REACH regulation, which includes plans for safer substitutions of recognized hazardous compounds. The limited availability of experimental data necessary for the hazard/risk assessment of chemicals and the expected high costs have increased the interest, also in REACH, for alternative predictive *in silico* methods, such as Quantitative Structure–Activity (Property) Relationships (QSA(P)Rs). A structurally-based approach is proposed here for a holistic screening of potential PBTs in the environment. Persistence, bioconcentration and toxicity data available for a set of 180 organic chemicals, some of which are known PBTs, have been combined in a multivariate approach by Principal Component Analysis. This method is applied to rank the studied compounds according to their cumulative PBT behaviour; this ranking can be defined as a PBT Index. A simple, robust and externally predictive QSPR multiple linear regression model (MLR), which is based on four molecular descriptors, has been developed for the PBT Index. This QSPR model is proposed as a hazard screening tool, applicable also by regulators, for the early identification and prioritization of not yet known PBTs, only on the basis of the knowledge of their molecular structure. New, safer chemicals can be designed as alternatives to hazardous PBT chemicals by applying the proposed QSPR model, according to the green chemistry philosophy of "benign by design". A consensus approach is also proposed from the comparison of the results obtained by different screening methods.

Introduction

Thousands of new chemicals are being developed each year (almost 38 million commercially available chemicals are reported in CAS[1]), but despite the fact that there is a higher degree of knowledge on physico-chemical properties, environmental reactivity and biological activities for new chemicals, the same cannot be said for the majority of "existing" chemicals in commercial use, not even for High Production Volume (HPV) compounds. In Europe, the new regulation REACH (Registration, Evaluation, Authorization and Restriction of Chemicals)[2] has created a single system to obtain relevant information on the properties and activities of all "existing" and "new" commercialized substances and plans to use such data for safe chemical management. According to this new regulation, in the next ten years, in order to evaluate any risk connected to a chemical's production and use, about 30000 existing substances will be processed on the basis of physico-chemical and toxicity data. Additionally, an authorization will be required to use, and commercialize, specific groups of substances considered to cause serious adverse effects to humans and the environment (*i.e.* Substances of Very High Concern—SVHC), such as chemicals that are classified as Carcinogenic, Mutagenic or toxic for Reproduction (CMR), Endocrine Disruptors (ED), Persistent (P), Bioaccumulative (B) and Toxic (T) (PBT), or Very Persistent and Very Bioaccumulative (vPvBs).

The REACH regulation requires the PBT/vPvB assessment of a substance to be carried out on the basis of the criteria defined in Annex XIII.[3] However, it has been observed[4–6] that PBT screening, according to defined criteria, is still a challenging process because of the limited amount and low quality of available P, B and T data.

Since the ratification of the Stockholm Convention on POPs,[7] different screening approaches have been proposed and applied, mainly by regulatory agencies, for the preliminary identification of potential PBT chemicals. These approaches have similar basic assumptions, which include the use of empirical criteria with defined cut-off values for single P, B and T properties,

QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria, via Dunant 3, 21100 Varese, Italy.
E-mail: paola.gramatica@uninsubria.it; Fax: +39-0332-421554; Tel: +39-0332-421573

† Electronic supplementary information (ESI) available: Table S1 reports the list of the 250 chemicals investigated, the experimental and predicted values of the studied endpoints, the values of the molecular descriptors selected for the proposed QSAR Model, information on the structural Applicability Domain, and the results of the PBT screening by U.S EPA PBT profiler and by our approach. Table S2 reports the results of the comparative screening of 45 chemicals by different approaches. Tables S1 and S2 are available as PDF and SDF files. SMILES for structures reported in SDF files were generated by EPI Suite 4.0. See DOI: 10.1039/b923843c

multimedia partitioning models, and Quantitative Structure–Activity (Property) Relationships–(QSA(P)R) models.[8–14] However, these approaches screen chemicals on the basis of P, B, and T properties taken singularly and compared to cut off values for each endpoint. Arnot and Mackay recently proposed[5] a holistic approach, which screens the risk related to PBT chemicals and takes into account environmental partitioning and quantities. Nevertheless, this innovative approach also requires the application of empirical data.

In order to minimize the consequences related to the unavailability of experimental data, which are necessary to perform realistic PBT assessment and screening, we decided to apply an alternative approach that combines multivariate analysis and QSAR/QSPR. In this study, we describe the development and validation of a model that (i) takes the input information from the available experimental data or commonly used persistence indicators, (ii) is based on structural molecular descriptors, and (iii) is able to identify PBT-like chemicals, with a combined approach, on the basis of only chemical structure.

At the hazard level, the influence of molecular structure is the dominant factor in determining chemical properties and behaviour, thus it is at this level that QSAR/QSPR-based approaches can be usefully applied. The development and use of QSAR to assess the hazard of substances is expressly promoted and included in the REACH regulation,[2] not only to fill the data gaps but also for a progressive substitution of dangerous substances with suitable, safer substances, as required by the authorization step.

High interconnectivity with Green (or Sustainable) Chemistry[15] is evident. Indeed, the Green Chemistry Principles 2 and 10 (2: design safer chemicals and products, 10: design chemicals and products to degrade after use) are in perfect agreement with the "benign by design" concept. The design of a safe molecule is, in fact, the earliest phase of the long process of placement of new safe substances on the market. Like drug design, molecular projecting modeling approaches such as QSAR can be successfully applied in "safe chemical design". Kummerer, in a very interesting paper,[16] has recently focused on the necessity of prevention of chemical hazards from the very beginning in the rational design of molecules.

The approach presented in our work was developed bearing in mind the Green Chemistry concepts reported above, with the aim to propose a QSAR model, based only on structural descriptors, for the screening of the cumulative PBT behaviour of existing chemicals, of new, not yet synthesized chemicals, or possible transformation products.

Methods

Data set

A structural set of 250 heterogeneous compounds was considered to study the global PBT-like behaviour of chemicals (Table S1, ESI†). This set was representative of many classes of pollutants of various chemical structures, such as dioxins, PCBs, PAHs, pesticides, and also various industrial chemicals, with different PBT behaviour. From these representative structures, two data sets were used to perform the calculation of the PBT Index and to validate the basic assumptions of our approach.

The first dataset (A) included only the 54 chemicals with data available for all three properties of persistence (P), bioconcentration (B) and toxicity (T). These data were collected from the literature[17–19] for a total of 162 experimental values.

The second dataset (B) included dataset A (54 compounds) and in addition another 126 compounds with experimental or predicted P, B, T data. About 70% of the values were experimental or extrapolated from experimental data (363) and 30% were predicted by QSAR (177) for a total of 540 data. Since only experimental data and reliable predictions were used in the study, values predicted exclusively inside the applicability domain of the respective QSAR model were included in dataset B. The 250 chemicals of the complete dataset, in which datasets A and B are included, as well as values of the P, B and T properties, are listed as Electronic Supplementary Information (ESI) in Table S1.† Details of the different P, B and T endpoints are given below:

Persistence (P)

Values of the Global Half-Life Index—GHLI[17] were used as a quantitative measure of a compound's persistence in different environmental media. GHLI is a holistic index of persistence, which was derived from the combination of overall half-life data[20] for transformation in air, water, soil and sediment, of 250 compounds by Principal Component Analysis (PCA). These data have been widely used in a lot of modeling approaches (fugacity models, for instance). Empirical GHLI values were available for all of the studied chemicals, and ranged from −3.134 to 4.255. Values >1 are associated with POP-like behaviour.

Bioconcentration (B)

Experimental log BCF data were taken from the literature.[18] Predicted log BCF data were estimated applying the QSPR equation proposed by Gramatica and Papa.[18]

Toxicity (T)

Experimental 96 h LC_{50} data (Duluth fathead minnow database) were taken from the literature.[19] Predicted toxicity data were estimated applying the QSAR equation proposed by Papa *et al.*[19] for Direct Toxicity Prediction (DTP-LogP-based). In order to have a positive trend of the response, all of the lethal concentrations were transformed into the logarithm of the inverse molar concentration $\log(1/LC_{50})_{96h}$.

Principal component analysis

PCA is used as an explorative multivariate technique that condenses, by linear combination, the relevant information of a group of variables that describes a system; the result is a smaller number of new and highly informative variables called Principal Components (PCs). Principal components are calculated according to the maximum variance criterion *i.e.* each successive component covers the maximum of the variance not accounted for by the previous components. The scores of the objects define their ranking along each PC.[21,22]

In this study, two PC Analyses were performed on P, B and T data available for dataset A and dataset B.

Molecular descriptors

The files for descriptor calculation, which contain information on atom and bond types, connectivity, and atomic spatial coordinates, were obtained by the software HYPERCHEM.[23] A set of 597 theoretical molecular descriptors (zero-, mono-, and bi-dimensional) was computed for all of the studied chemicals by the software DRAGON.[24] One of the advantages of using simple descriptors, that do not require the application of quantum mechanical methods for their calculation or conformational studies, is that they can be derived from the 2D structures of the chemicals or even from their SMILES code. Constant and near-constant descriptors were excluded in a pre-reduction step, thus 410 molecular descriptors underwent subsequent selection for the best modeling variables.

QSAR modeling

Multiple linear regression (MLR) and variable selection were performed by Ordinary Least Squares regression (OLS).[25] The Genetic Algorithm-Variable Subset Selection (GA-VSS) approach was applied to the input set of 410 descriptors to select those most relevant to obtain models with the highest predictive power in modeling the PBT Index, defined by the PCA scores. The coefficient of determination (R^2) was reported as a measure of the total variance of the response explained by the regression models (fitting). All of the models were internally validated by the leave-one-out procedure (Q^2_{LOO}), and the robustness of the models was further evaluated by bootstrap (Q^2_{BOOT}). Evidence that the proposed models were well founded, and not just the result of chance correlation, was provided by Y scrambling permutation testing: new models were recalculated for a randomly reordered response, which resulted in a significantly lower R^2 than the originally proposed models. The averaged scrambled R^2 (R^2_{YS}) was calculated after 500 scrambling iterations.

External validation of the QSAR models

External validation was performed to verify the real predictive power of the models.[26–28] The model developed for the PBT Index calculated on experimental data only (Dataset A–54 compounds) was externally validated on PBT Index values calculated from experimental plus predicted P, B and T data for the remaining 126 compounds from Dataset B.

Finally, an additional measure of the accuracy of the proposed QSPRs is the Root Mean Squared of Errors (RMSE) that summarizes the overall error of the model. It is calculated as the square root of the sum of squared errors in prediction divided by their total number. This parameter was used to compare the accuracy and the stability of our models in the training ($RMSE_T$) and in the prediction ($RMSE_P$) sets.

Chemical applicability domain

QSAR models are developed on a defined domain of compounds with known properties and structures (training set). For this reason they cannot be applied for predictive purposes to every new chemical. Quantitative measures of a model applicability domain (AD) are needed to evaluate the degree of extrapolation and for the identification of problematic compounds.[26–28] In this study, the AD was defined by the leverage approach (for the structural domain), and by the identification of response outliers (compounds with cross-validated standardized residuals greater than 2.5 standard deviation units). Graphically, the plot of hat values (h) *versus* standardized residuals, *i.e.* the Williams graph, verified the presence of response outliers and chemicals in the training set structurally very influential in determining model parameters (compounds with leverage value (h) greater than $3p'/n$ (h^*), where p' is the number of the model variables plus one, and n is the number of the objects used to calculate the model).[29] The data predicted for high leverage chemicals in the prediction set are extrapolated and could be unreliable.

Results and discussion

Our work is based on two different steps: (a) the development of a multivariate tool for screening chemicals according to their cumulative PBT properties (definition of a PBT Index), and (b) the development of a QSAR model of the PBT Index.

Chemical ranking according to PBT behaviour and definition of the PBT Index

In this study, PCA was applied to rank chemicals according to their potential cumulative PBT behaviour. Due to the unavailability of complete P, B, T experimental data for all of the studied compounds, the strength of our approach was validated in two sequential steps. In the first step, PCA was performed on the available experimental dataset A, thus 54 chemicals were ranked along PC1 according to their experimental P, B, T values (PCA-A).

In the second step, PCA was performed on all of the PBT data available for dataset B (dataset A plus an additional 126 compounds = PCA-B), then the results from PCA-A and PCA-B were compared.

Fig. 1 shows the graph of the first and second components from PCA-A: the position of the compounds is defined by the

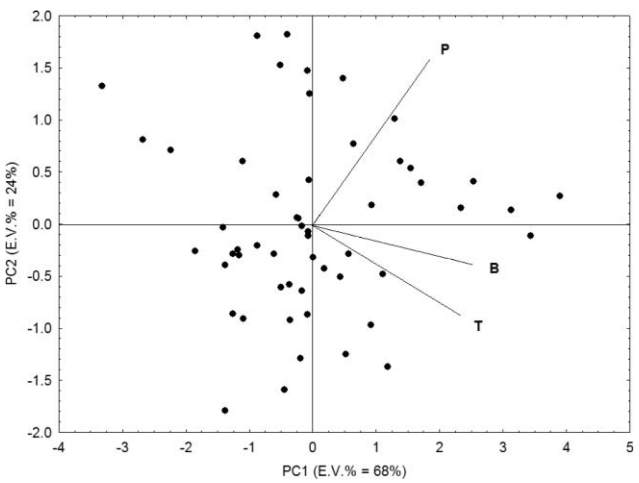

Fig. 1 Principal Component Analysis on experimental P, B and T data for 54 organic compounds (Dataset A). PC1–PC2: Explained Variance = 92%. PC1 values (PBT Index) shown in this picture (and considered in this paper) were previously obtained by multiplying the original PC1 score values by −1. This was done to obtain a left to right increasing ranking of the chemicals along PC1, which defines the PBT Index.

coordinates (scores) along the new PC1 and PC2 axes. The cumulative explained variance of the first two PCs is 92%, whereas the PC1 alone provides close to 70% of the total information. The loading lines show the importance of each P, B and T property in the new PC1–PC2 space.

It is interesting to note that all of the PBT properties (loading lines) are oriented in the same direction along the first Principal Component (PC1), so that the compounds are ranked from left to right according to increasing PBT behaviour potential. Therefore, this PC1 is now considered as a new macro-variable that condenses the PBT potential of chemicals, and it is defined here as the PBT Index. Even though this ranking successfully condenses the PBT potential of chemicals, the structural and response domain of this analysis is limited to 54 compounds, compared to the larger structural domain included in Dataset B. Therefore, in order to enlarge the structural and the properties domain of our analysis, a second PCA (PCA-B) was performed based on Dataset B, which is composed of an integration of experimental and reliable predicted data for the P, B, T properties (70% experimental data) for a total of 180 chemicals (Fig. 2).

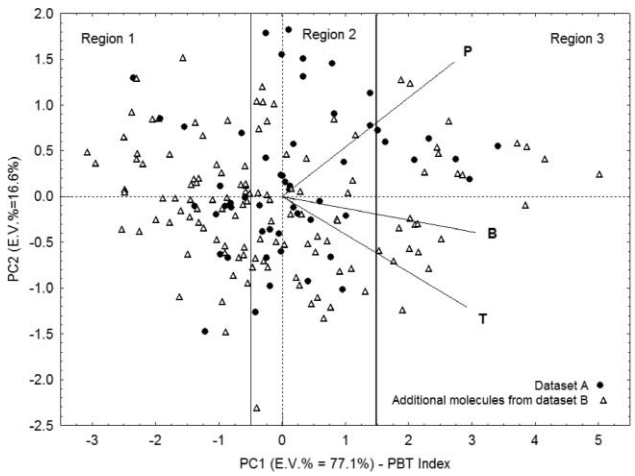

Fig. 2 Principal Component Analysis on experimental and predicted PBT data for 180 organic compounds. The 54 compounds belonging to dataset A are reported as black dots. PC1–PC2: Explained Variance = 93.7%. PC1 values (PBT Iindex) shown in this picture (and considered in this paper) were previously obtained by multiplying the original PC1 score values by −1. This was done to obtain a left to right increasing ranking of the chemicals along PC1, which defines the PBT Index (PBT and vPvB compounds = high values of PBT Index). The full vertical lines are related to the cut off values commented on in the text, which identify PBT and vPvB chemicals (PBT Index value > 1.5—Region 3), chemicals with medium PBT behaviour (−0.5 < PBT Index value < 1.5—Region 2) and not PBT chemicals (PBT Index value < −0.5—Region 1).

The cumulative explained variance of the first two PCs in PCA-2 is 93.7% (PC1 explained variance = 77.1%). As expected, the relative distribution of chemicals along PC1 is very similar to that found in PCA-A. In particular, the correlation between PBT Index values obtained from PCA-A and PCA-B for the 54 compounds in common is 99.4%. This fact is a proof of the consistency of the PBT Index as calculated only on experimental data or on a larger domain that also includes 30% of predicted values. Arbitrary threshold values are then defined, which can help in the identification of PBT-like compounds. A first threshold value is set at PC1 score ≥ 1.5 (Fig. 2—Region 3) to highlight the PBT and vPvB chemicals. A second threshold value is set at PC1 score < −0.5 to separate compounds with no-PBT behaviour (Fig. 2—Region 1) from compounds with intermediate categorized PBT behaviour (−0.5 < PC1 score < 1.5; Fig. 2—Region 2). We were able to verify that, according to the screening criteria for PBT, reported in Annex XIII of the REACH regulation,[3] at least 60% of the compounds in Region 3 are PBTs and all of the compounds in this region are vPvBs (average values for the P, B, T properties of the compounds grouped in region 3: GHLI index = 2.19, BCF = 17702, Fish Acute Toxicity = 2.32 mg L^{-1}). The P, B, and T values used to perform the PCA analysis are reported in the ESI, Table S1;† detailed information about the average values of the studied properties in the three regions of Fig. 2 are reported in Table 1.

Table 1 Average values for the P, B, T properties of the compounds grouped in Regions 1, 2 and 3 (Fig. 2) according to cut off values

	P[a]	B	T/mg L^{-1}
Region 3	2.12	16774.09	2.27[b]
Region 2	0.01	99.05	36.97
Region 1	−1.36	11.11	947.08

[a] Value for the GHL-index (Gramatica and Papa, 2007). Values > 1 are expected for very Persistent (vP) chemicals. [b] 60% of T values in 1 were <1 mg L^{-1}

PC2, which is less informative than PC1 (PC2 Explained Variance = 16.6%), separates the compounds vertically, mainly on the basis of their higher persistence (positive values of PC2) or bioconcentration and toxicity (negative values of PC2). The B and T properties, which are correlated to hydrophobicity, are here separated from the Persistence property, which is chemically/biologically quite independent from hydrophobicity.

QSPR modeling and validation of the PBT Index

A QSPR model was developed for the PBT Index with the aim of predicting the position of new chemicals following the PBT trend, thus also predicting their potential PBT behaviour from information concerning only their molecular structure. According to the OECD Principles for the Validation of (Q)SAR Models for Regulatory Purposes,[28] a QSAR model must be externally validated to be considered predictive.[26,27] For this reason, the PBT Index from PCA-A (PC1 scores) was used as training information to develop the QSAR model, while the external predictivity was tested on the values of PBT Index calculated for the remaining 126 compounds in PCA-B (prediction set).

A population of MLR models was developed by applying the Genetic Algorithm procedure to the calculated theoretical molecular descriptors. Only those variables whose combination successfully modeled the PBT Index were included in the MLR population of high performance models. Among these, the best model was chosen as the one with the best balance of high values of fitting (R^2 = 80.72%, $RMSE_T$ = 0.62), robustness (Q^2_{LOO} = 75.70%; Q^2_{boot} = 75.0%), absence of chance correlation (R^2_{YS} = 0.07), external predictivity (Q^2_{ext} = 80.72%; R^2_{ext} = 89.27%; $RMSE_P$ = 0.72), and minimum complexity, in terms of number and better interpretability of the structural modeling descriptors.

The descriptors selected for the best model by the Genetic Algorithm procedure are (in descending order of importance): *n*X (number of halogen atoms), *n*BM (number of multiple bonds), *n*HDon (number of donor atoms for H bonds), MAXDP (maximal electrotopological positive variation).[30,31] All of these parameters are mono- or bi-dimensional and independent of chemical conformation, thus easily calculable from the topological graph (2D sketch) or even from the SMILES code. These variables take into account different chemical properties. The most important descriptors, *n*X and *n*BM, which encode for substitution with halogens and unsaturation, are known to increase the PBT behaviour of chemicals. On the contrary, MAXDP and *n*HDon 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. In particular, the descriptor MAXDP is calculated from the hydrogen depleted molecular 2D graph:[30,31]:

$$\mathrm{MAXDP} = \max|\Delta I_i| \text{ if } \Delta I_i > 0,\ i = 1\ldots A$$

where ΔI_i is the field effect on the *i*th atom due to the perturbation of all other atoms, as defined by Kier and Hall. The intrinsic state of an atom, which is used to define the parameter ΔI_i, is calculated as the ratio between Kier–Hall atomic electronegativity and the vertex degree, *i.e.* the number of bonds of the atom; it encodes 2D information related to both partial charges of atoms and their topological position relative to the whole molecule.[30,31]

The fact that these variables were selected in the model for the PBT Index is not surprising; we have already demonstrated that their contribution is relevant, in combination with other descriptors, to model also single endpoints related to soil sorption,[30,32,33] persistence,[17] bioaccumulation[18] and toxicity.[19] The results shown above are a proof of the high external predictive power of the model developed for the experimental PBT Index; however, the analysis of the applicability domain highlighted that the predicted data for some chemicals in the prediction set (*p*,*p*′–DDT (1); ethanethioamide (14); *p*,*p*′–DDE (32); phenanthrene (53); 1-butanamine (130); anthracene (157); *o*,*p*′-DDE (228); PCB 122 (235)) were extrapolated (high leverage). This fact is due to the limited structural domain represented in the training set of 54 chemicals.

Therefore, in order to enlarge the domain of the model and to generalize its applicability, the best combination of the previously selected modelling variables, was used to model the PBT Index values calculated for all of the 180 chemicals (Full Model). The new modelled endpoint consists of the union of the PBT Index values from PCA-A, with the additional 126 PBT Index values from PCA-B.

The Full QSPR model, proposed for its application in REACH for the screening of PBT-like compounds, has the following equation and statistical parameters:

$$\text{PBT Index} = -1.44\ (\pm 0.10) + 0.65\ (\pm 0.03)\ n\mathrm{X} + 0.22\ (\pm 0.01)\ n\mathrm{BM} - 0.39\ (\pm 0.06)\ n\mathrm{HDon} - 0.07\ (\pm 0.03)\ \mathrm{MAXDP} \quad (1)$$

$$n = 180,\ R^2 = 88.40\%;\ Q^2 = 87.72\%;\ Q^2_{boot} = 87.58\%;\ R^2_{YS} = 0.02;\ \mathrm{RMSE} = 0.52$$

This model is based on the whole structural and response domain of this heterogeneous dataset, and it is more adequate than the model developed on only 54 chemicals for the prediction of the PBT Index and the screening of many new compounds. In fact, no extrapolated predictions were obtained, since no structurally influential chemicals were detected from the analysis of the applicability domain of this model. The model showed a broad applicability domain also when applied to all of the 250 chemicals of the complete dataset. In fact, only 17 of the 70 compounds with unknown PBT behaviour (Table S1, ESI†) fell outside of the structural AD of the model.

Moreover, only three compounds were found to be response outliers with standardized residuals slightly higher than 2.5σ: n-nitroso-n-phenyl benzeneamine (55), quinoline (65), and benzophenone (156). However, it should be noted that these compounds were not detected as PBTs. In fact, the PBT Index for each of these compounds was always below the cut off value 1.5, which excluded their PBT behaviour.

Values of the variables included in eqn (1), calculated for 250 compounds, as well as predicted and experimental values of the PBT Index, are reported in the ESI, Table S1†. The plot of the experimental *vs.* predicted values for the 180 compounds in the training set and the predictions for the 70 chemicals with unknown PBT behaviour (seen distributed on the straight-line), included in this study, are shown in Fig. 3.

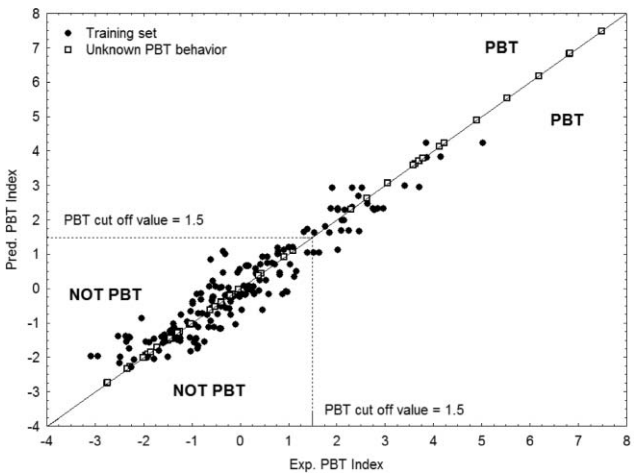

Fig. 3 Scatter plot of experimental PBT Index values (calculated by PCA) *vs.* values predicted by eqn (1) for all of the 250 chemicals included in the dataset. Training set chemicals and compounds with unknown behaviour are labelled differently. Vertical and horizontal dotted lines identify the cut off value of the PBT Index for PBT and vPvB compounds = 1.5.

As an additional analysis, we also verified that the modeling power of *n*BM, *n*X, MAXDP and *n*HDon, singularly applied to the individual P, B, T data (dataset B), depends on the modeled response, and that the most powerful descriptors were always *n*X and *n*BM. This is in agreement with eqn (1). *n*BM had the highest single-modeling power for the endpoints LogBCF and Log1/LC_{50} (R^2 = 38 and 47% respectively), followed by *n*X (R^2 = 36 and 23%, respectively, for logBCF and Log1/LC_{50}). *n*X had the highest single-modeling power for the endpoint GHLI (R^2 = 55%). Therefore, the number of halogens is clearly related to persistency, but it also plays an important role to model BCF and

toxicity. *n*HDon and MAXDP were singularly low correlated to all of the responses (R^2 ranges from about 3 to 9%).

Therefore, the PBT Index model takes into account different structural features whose relevance can be similarly ranked in respect of P, B and T endpoints.

Moreover, the combination of the 4 variables was successful in modeling the three properties individually with ranges of R^2 from 72 to 76%. These results are a demonstration that only the multivariate combination of all of these descriptors is able to model and codify complex mechanisms, such as partitioning in environmental/biological phases, and availability for degradation and metabolisation, which determine the persistence, bioaccumulation and toxicity of a chemical.

Therefore, on the basis of the performances of the proposed model, the strong validation and the analysis of applicability domain commented on above, we can conclude that eqn (1) is a robust and predictive model, which can be applied by regulators and QSAR users to screen chemicals with unknown PBT-like behaviour, simply by calculating the modeling structural descriptors. This calculation can be performed for each chemical from the SMILES string by using the appropriate software,[24] which is also freely available online at http://www.vcclab.org/lab/edragon. It is noteworthy that the synthesis of chemicals with predicted PBT Index values > 1.5 must be discouraged, while possible alternatives to unsafe compounds should be searched for among existing or newly designed chemicals with predicted PBT Index values below the proposed cut off.

The intrinsic advantage of the application of this holistic model for predictive/screening purposes lies in the fact that, differently from other existing approaches,[4–14] no experimental data for the single P, B and T properties, or knowledge about environmental partitioning, are needed by users to predict the potential PBT behaviour of a chemical.

Comparison with other methods to screen PBTs

In order to verify the quality of the output of our approach, the accuracy has to be strongly verified with increasing levels of external validation. This can also be performed by comparison of our results with other PBT screening approaches, such as the US-EPA PBT Profiler,[11] the Canadian DSL List,[34] and HAF values proposed by Arnot and Mackay.[5]

A first comparison was done between the results obtained by applying our model of PBT Index to all of the 250 chemicals originally included in our dataset, with results obtained by the US-EPA PBT Profiler.[11] Further information about this comparison is reported in Fig. 4 (for the 180 compounds included in PCA-B) and in the ESI, Table S1† (for all of the 250 compounds included in this study).

It is interesting to note that the screening of chemicals as PBTs or not PBTs performed by the two approaches (PBT Index > 1.5; P, B and T cut off exceeded in PBT Profiler) gave results with 75% correlation. Only two chemicals among those identified as PBTs by the PBT Profiler were not predicted as PBTs by our QSPR model: octane (145) and isopropalin (240) (Table S1, ESI†). Octane was detected as a possible PBT by the US-EPA software, since it exceeds the lowest criteria for persistence in sediment (2 months), as well as the B and T criteria considered

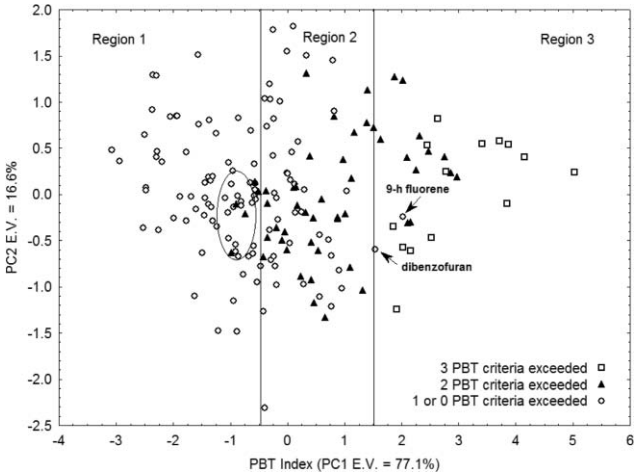

Fig. 4 Principal Component Analysis on PBT data for 180 organic compounds (PC1–PC2: Explained Variance = 93.7%). Chemicals are labelled according to the results obtained by the US-EPA PBT Profiler (14): empty squares identify PBTs, full triangles identify compounds with two out of three PBT criteria, circles identify compounds with one or no PBT features. The full vertical lines are related to our cut off values commented on in the text, which identify PBT and vPvB chemicals (PBT Index value > 1.5—Region 3), chemicals with medium PBT behaviour (–0.5 < PBT Index value < 1.5—Region 2) and not PBT chemicals (PBT Index value < –0.5—Region 1). The full triangles circled in zone 1 identify PT compounds according to the US-EPA PBT Profiler (4-chloroaniline; 114 bis(2-chloroisopropyl)ether; 162 *N*,*N*-dimethyaniline; 214 EPTC). The underestimated compounds commented on in the text, octane and isopropalin, are not included in the 180 compounds used to develop this PCA, thus, they don't appear in Fig. 4. Results of their screening are reported in the ESI, Table S1.†

by the EPA PBT Profiler. However, it should be noted that none of the other alkanes included in the dataset (pentane, hexane, decane, dodecane) were predicted as PBT-like by the US-EPA PBT Profiler or by our model, thus, in our opinion, the identification of octane as PBT could be an overestimation of the PBT profiler.

Differently, the experimentally verified PBT behaviour of the herbicide isopropalin (145) (http://sitem.herts.ac.uk/aeru/footprint/it/Reports/407.htm) was not fully detected by our model (predicted PBT Index value = 0.379). In the studied dataset only one other chemical (trifluralin (217), correctly detected as PBT) is similar to isopropalin; however, trifluralin's structure also includes three fluorine atoms. The absence of halogen atoms from the isopropalin structure, and the lack of structurally similar chemicals in our studied dataset, was probably the reason for the underestimation of the PBT behaviour of this herbicide. In addition, the potential behaviour of four compounds (4-chloroaniline; 114 bis(2-chloroisopropyl)ether; 162 *N*,*N*-dimethyaniline; 214 EPTC), which were screened as PT according to the US-EPA PBT Profiler, was slightly underestimated by the PBT Index.

On the contrary, considering the chemicals screened as PBTs by our approach, as they exceeded the 1.5 cut off of our PBT Index (Fig. 2 and 4), it is interesting to note that 30% of them were not recognized as PBTs by the US-EPA PBT Profiler, in particular, 9h-fluorene (56) and dibenzofuran (175) were screened as not P, not B (Fig. 4). On the basis of

these observations we can conclude that our approach gives comparable results to the US-EPA PBT Profiler, but it allows for a more conservative identification of PBT-like compounds. Thus, it is in agreement with the precautionary principle.

A second comparison was made with the results of the PBT screening method recently proposed by Arnot and Mackay.[5] They developed a holistic approach for the priority setting of PBTs, which was tested on chemicals included in the Canadian DSL list.[34] In order to further validate our approach, the PBT Index values were compared to the Hazard Assessment Factor (HAF), which is a combined function of P, B, and T properties. The results of this comparison are listed as ESI in Table S2.† This table reports the PBT screenings performed for 45 reference compounds included in the DSL list, but not present in the training set used to develop our QSAR model. These profiles were defined by applying four approaches based on different assumptions. Approach A—US-EPA PBT-Profiler—is based on chemical partitioning in environmental media and single P, B and T cut off criteria;[11] approach B—DSL classification—is based on single P, B and T cut off criteria;[33] approach C—holistic function HAF[5]—is based on partitioning and physico chemical properties; Approach D—QSPR model for the PBT Index—is based on molecular structure and one cut off value (this paper).

The reason for this comparison is to highlight that, due to the number of criteria and the different applicable methods, it is not possible to assign chemicals as potential PBTs following a single approach. In fact, Table S2 (ESI†) shows that despite the general agreement of the different approaches (A, B, C and D), 10 compounds were screened differently, probably due to different basic assumptions existing in each screening method.

It is interesting to highlight that the EPA PBT Profiler was the least precautionary approach of the four.

Considering all of the 45 reference chemicals, the predictions of our PBT Index were in good agreement with all the other approaches: the correlations were 73, 76 and 87% respectively with the Arnot–Mackay approach,[5] the DSL List[34] and the PBT Profiler.[11] In particular, our predictions were more conservative than the PBT Profiler, but less restrictive than the DSL categorization.

The effect of these different results can be minimized without losing the relevant information associated with each single screening method. In fact, to overcome this problem, the authors suggest the use of a consensus approach that allows for a final detection of PBT compounds on the basis of all of the results from different screenings. The consensus is the most frequent result shared among the different approaches. In the case of equity, the precautionary principle is applied and chemicals are defined as PBTs.

Table S2 (ESI†) shows the result of the screening by consensus for the 45 reference chemicals. It can be seen that the combination of all of the approaches, even with different results, was useful to define the possible PBT profile of a compound.

The strength of this consensus approach is that the final result takes into account the different assumptions characterizing each method, *i.e.* from chemical structure to partitioning and cut off criteria; therefore, it allows for a more reliable judgment in a complex situation. Considering the lack of information existing for the PBT properties, in our opinion it is important to combine and compare results from as many reliable screening tools as possible; this will minimize the limitations of each screening tool and it will give a more realistic assessment of the potential PBT behaviour of chemicals from different perspectives.

Conclusions

The presented structure-based approach is a response to two levels of actions in relation to the management, according to REACH regulations, of chemicals of highest concern: (a) the need of tools for identification and prioritization, and (b) incentives for the discovery and production of safer alternatives.

The core of our QSAR approach, and its characteristic in comparison to other methods for PBT assessment, is that it is based only on structural information. The application by regulators of the herewith proposed QSPR model of PBT Index (eqn (1), which will be implemented as a web tool into a National Project framework) allows for fast screening and ranking of heterogeneous PBT-like compounds just starting from their molecular structure, represented by four very simple descriptors. The PBT screening based on this QSPR model can be used as a preliminary screening tool and as a support for other existing methods to highlight potential PBTs among existing and new chemicals, hypothetical metabolites or even not yet synthesized products with no available data for their P, B, and T properties (screening *a priori*), as it does not require any knowledge of the persistence, bioaccumulation or toxicity data of the chemicals of interest.

Therefore, we strongly believe that this QSAR approach is particularly useful for the environmentally benign design of safer replacement solutions for recognized PBTs. No method other than QSAR is applicable to chemical design and to detect *a priori*, from the drawn structures, the potential PBT behaviour of completely new compounds.

Acknowledgements

Financial support by MIUR (Italian Ministry for Instruction, University and Research) through the National project "Development of chemoinformatics tools for screening and identification of Persistent Bioaccumulable and Toxic (PBTs) compounds and Endocrine Disruptors (EDs) for REACH regulation" (PRIN-2007R57KT7) is gratefully acknowledged. We also thank Dr Barun Bhhatarai for his critical comments on the draft of the manuscript.

References

1 *CAS Registry*, available at: http://www.cas.org/cgi-bin/cas/regreport.pl.
2 REACH Regulation (EC), No 1907/2006, available at: http://eur-lex.europa.eu/LexUriServ/site/en/oj/2006/l_396/l_39620061230-en00010849.pdf.
3 *Guidance on Information Requirements and Chemical Safety Assessment*, European Chemicals Agency, 2008, available at: http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_r11_en.pdf?vers=20_08_08.
4 J. Arnot and F. A. P. C. Gobas, *Environ. Rev.*, 2006, **14**, 257.
5 J. Arnot and D. Mackay, *Environ. Sci. Technol.*, 2008, **42**, 4648.
6 A. V. Weisbrod, L. P. Burkhard, J. Arnot, O. Mekenyan, P. H. Howard, C. Russom, R. Boethling, Y. Sakuratani, T. Traas, T.

Bridges, C. Lutz, M. Bonnell, K. Woodburn and T. Parkerton, *Environ. Health Perspect*, 2007, **15**, 255.
7 *Stockholm Convention on Persistent Organic Pollutants*, United Nations Environment Program, Geneva, Switzerland, 2001, available at: http://www.pops.int.
8 D. Muir and P. H. Howard, *Environ. Sci. Technol.*, 2006, **40**, 7157.
9 *Toxic Substances Management Policy. Persistence and Bioaccumulation Criteria*, Environment Canada, Ottawa, ON, 1995, En 40-499./1, 2.E,.
10 U.S. EPA, *Fed. Regist.*199863, (192), 53417.
11 *PBT-Profiler*, U.S. EPA, available at: http://www.pbtprofiler.net.
12 *Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.00*, United States Environmental Protection Agency, Washington, DC, USA, 2009.
13 L. Carlsen and J. D. Walker, *QSAR Comb. Sci.*, 2003, **22**, 49.
14 O. G. Mekenyan, S. D. Dimitrov, T. S. Pavlov and G. D. Veith, *SAR QSAR Environ. Res.*, 2005, **16**, 103.
15 P. Anastas and J. Warner, *Green Chemistry: Theory and Practice*, Oxford University Press, New York, 1998.
16 K. Kummerer, *Green Chem.*, 2007, **9**, 899.
17 P. Gramatica and E. Papa, *Environ. Sci. Technol.*, 2007, **41**, 2833.
18 P. Gramatica and E. Papa, *QSAR Comb. Sci.*, 2005, **24**, 953.
19 E. Papa, F. Villa and P. Gramatica, *J. Chem. Inf. Model.*, 2005, **45**, 1256.
20 D. Mackay, M. Y. Shiu and K. C. Ma, *Physical-Chemical properties and Environmental Fate Handbook*, CRCnet-BASE CD-ROM, Chapman and Hall/CRC, Boca Raton, FL (USA), 2000.
21 J. E. Jackson, *A User's Guide to Principal Components*, Wiley, New York, 1991.
22 *SCAN Software for Chemometric Analysis, ver. 1.1 for Windows*, Minitab (USA), 1995.
23 *HyperChem, rel. 7.03 for Windows*, Autodesk, Inc., Sausalito, CA (USA), 2002.
24 *DRAGON for Windows (Software for molecular descriptors calculation) ver.5.5*, Talete srl, Milano, Italy, 2007.
25 *MOBY DIGS Professional for Windows (Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm) ver. 1.0 beta*, , Talete srl, Milano, Italy2004.
26 A. Tropsha, P. Gramatica and V. K. Gombar, *QSAR Comb. Sci.*, 2003, **22**, 69.
27 P. Gramatica, *QSAR Comb. Sci.*, 2007, **26**, 694.
28 *OECD Principles for the Validation, for Regulatory Purposes, of (Q)SAR Models*, Organisation for Economic Co-operation and Development, available at: http://www.oecd.org/dataoecd/33/37/37849783.pdf.
29 A. C. Atkinson, *Plots, transformations and regression*, Clarendon Press, Oxford, UK, 1985.
30 P. Gramatica, M. Corradi and V. Consonni, *Chemosphere*, 2000, **41**, 763.
31 L. B. Kier, L. H. Hall and J. W. Frazer, *J. Math. Chem.*, 1991, **7**, 229.
32 P. Gramatica, E. Giani and E. Papa, *J. Mol. Graphics Modell.*, 2007, **25**, 755.
33 E. Papa and P. Gramatica, *J. Mol. Graphics Modell.*, 2008, **27**, 59.
34 *Domestic Substances List Categorization and Screening Program*, Environment Canada, available at: http://www.ec.gc.ca/substances/ese/eng/dsl/dslprog.cfm.