Prioritization of Chemicals Based on Chemoinformatic Analysis

55

Paola Gramatica

Contents

Abstract

Several different chemical properties/activities must be contemporaneously taken into account to prioritize compounds for their hazardous behavior. Examples of application of chemoinformatic methods, such as principal component analysis for obtaining ranking indexes and hierarchical cluster analysis for grouping chemicals with similar properties, are summarized for various classes of compounds of environmental concern. These cumulative endpoints are then modeled

P. Gramatica (✉)
QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Theoretical and Applied Sciences, University of Insubria, Varese, Italy
e-mail: paola.gramatica@uninsubria.it

J. Leszczynski et al. (eds.), *Handbook of Computational Chemistry*,
DOI 10.1007/978-3-319-27282-5_58

by validated quantitative structure–activity relationships, based on theoretical molecular descriptors, to predict the potential hazard of new chemicals.

Introduction

The chemical universe is huge and is rapidly enlarging every day: the number of chemicals registered in the Chemical Abstract Service (CAS) registry (www.cas.org) gets nowadays over 100 million of chemicals, the majority of them are commercially available, and almost 345,000 are regulated and listed in various inventories (for instance, EU-EINECS, US-EPA TSCA, Canada-DSL). While many new chemicals are being developed continuously (many thousands each year), with increasing possibility to interact with humans and wildlife, information on physicochemical properties, reactivity, and biological activities are more slowly produced. The problem of lack of data and slow assessment procedures is highly significant: in fact, so far, we have extensive information about only a few chemicals, but the majority of compounds (>95 %), even high production volume (HPV) compounds, have been not sufficiently well characterized for their environmental behavior and potential to cause human or ecologic toxicity (Judson et al. 2009; Arnot et al. 2012). The filling of this data gap, in order to assess and control the chemicals effectively, is one of the main aims of several legislations worldwide, in particular of the recent European legislation REACH (Registration Evaluation Authorization and restriction of Chemicals) (EC Regulation 2006).

However, it is clearly impossible to measure all chemicals in all media to which humans and ecological receptors are exposed, as well as to test a plethora of endpoints. In order to reduce costs, time, and sacrificed animals, there is an urgent need to prioritize the use of testing resources toward those chemicals and endpoints that present the greatest potential of risk to human health and environment. Therefore, it is highly evident that the prioritization of chemicals is nowadays a big challenge, mainly for the identification of new emerging pollutants.

Chemoinformatic analysis has a clear and fundamental role in dealing with this issue. By chemoinformatic methods, it is possible to analyze and model the experimental information, which is already available for tested chemicals, and to exploit this information applying the developed tools to chemicals without experimental data or even before their synthesis. The main aim is to better use the existing knowledge for preventing, as soon as possible, potential dangerous properties of not yet tested compounds and also for planning *a priori* the synthesis of safe chemicals.

In recent years, many scientists have faced this important problem, studying various endpoints by different chemoinformatic approaches, but with the same common aim: to highlight the most hazardous chemicals by screening data sets of several compounds (here some representative of more recent examples: Gramatica and Di Guardo 2002; Salvito et al. 2002; Sanderson et al. 2003, 2004; Schmieder et al. 2003; Tong et al. 2003; Gramatica et al. 2004a, b, 2015, 2016a, b; Knekta et al. 2004; Öberg 2004, 2005, 2006; Muir and Howard 2006; Klasmeier et al. 2006; Hansson and Rudén 2006; Liu et al. 2006, 2007; Dix et al. 2007; Gramatica and

Papa 2007; Brown and Wania 2008; Papa and Gramatica 2008, 2010; Wegmann et al. 2009; Judson et al. 2009; Stenberg et al. 2009; Kavlock and Dix 2010; Li and Gramatica 2010a, b; Bhhatarai and Gramatica 2010, 2011a; Howard and Muir 2010; Kovarich et al. 2011, 2012; Öberg and Iqbal 2012; Strempel et al. 2012; Roos et al. 2012; Sanderson 2012; Zarfl et al. 2012; Scheringer et al. 2012; Arnot et al. 2012; Guillen et al. 2012; Singh et al. 2014; Cassani and Gramatica 2015; Wedebye et al. 2015; Sangion and Gramatica 2016a, b).

In this chapter, the crucial topic of screening chemicals of heterogeneous molecular structure is faced, studying some specific endpoints and focusing on compounds of environmental concern. These screening studies have two different aims: (1) to rank, highlight, and prioritize the most hazardous compounds among the already used chemicals, also those without experimental data, and (2) to predict the potential dangerous behavior of not yet synthesized compounds, in an *a priori* approach of the "benign by design" strategy of green chemistry. The focus here will be on the potential hazard intrinsically related to the chemical structure, thus on the utility of QSAR (quantitative structure–activity relationship) modeling, in particular based on a preliminary chemoinformatic analysis. A chemometric method of explorative analysis, such as principal component analysis (PCA), is applied for defining trends and ranking indices, as well as for the *a priori* data set splitting for external validation of QSAR models. Hierarchical cluster analysis (HCA) is used for grouping chemicals, according to some properties, and for defining the *a priori* classes for a subsequent classification by various classification methods. Particular emphasis is here devoted to multivariate linear regression (MLR) models, in particular ordinary least squares (OLS) models, based on genetic algorithm (GA) for variable selection and developed by the software QSARINS (QSAR-INSubria) for QSAR model development and validation (Gramatica et al. 2013). Some of the QSAR models of cumulative ranking endpoints here presented, applied to several classes of chemicals of emerging concern (CEC), are implemented in the module QSARINS-Chem (Gramatica et al. 2014) for easy applicability. Moreover, results of prioritization of endocrine disruptors (EDs), performed by various classification models, are also here commented.

A previous review of these chemoinformatic approaches, presenting the basis of externally validated QSAR modeling, illustrated according to the OECD principles for QSAR in regulation (OECD 2004) was published in Chapter 12 of the book *Recent Advances in QSAR Studies* (Gramatica 2009), edited by the same editors of the present book. Therefore, this chapter is mostly an updating of the previous one on the approaches widely applied in the Insubria QSAR research group, with special emphasis on the QSAR modeling of ranking indices obtained by PCA.

QSAR Modeling for Prioritization

QSAR (quantitative structure–activity relationship) modeling is based on the assumption that the molecular structure of a chemical (i.e., its geometric, steric, and electronic properties) contains the features responsible for its physical, chemical, and biological properties. Such modeling techniques are the best chemoinformatic approaches for finding and exploiting the information inherent in the molecular structure related to the intrinsic hazard of any chemical. In fact, by QSAR models, based on theoretical molecular descriptors and validated chemometric methods, both of regression and classification, the biological activity (or property, reactivity, etc.) of new or untested chemicals can be inferred from the molecular structure of compounds whose activities (properties, reactivities, etc.) have already been assessed.

There is no need, in this chapter, to enter into details on QSAR modeling, commented elsewhere in this book; this has been also the topic of my chapter of the above-cited book (Gramatica 2009) and of all my papers in these last 20 years, some cited also here.

However, it is important to stress the point that, particularly for prioritization aim in screening big data sets, it is not sufficient to be able to reproduce well the available data and it should not be of primary importance to understand underlying mechanisms. To know the real predictivity of QSAR models and to which chemicals the model could be more reliably applied is of crucial and fundamental relevance. For this reason, all my QSAR works, which are mainly devoted to screening and prioritization, are focused on external validation on chemicals never used for model development and on model applicability domain (AD) check (Gramatica 2007, 2014; Gramatica et al. 2012). The proposal of concordance correlation coefficient (CCC) as validation parameter for QSAR models and its comparison with other used statistical parameters (Chirico and Gramatica 2011, 2012), highlighting some drawbacks and proposing intercomparable thresholds for real predictivity, were also done for this purpose. Recently, all these compared external validation parameters, and the Insubria graph for checking the applicability domain of QSAR models to chemicals without experimental data, have been implemented in our software QSARINS, freely available on request (www.qsar.it) for academia and research centers.

Ranking Indexes: PC Scores as New Cumulative Endpoints for QSAR Models

The behavior of chemicals in the environment and their impact on humans and wildlife are dependent on many different variables such as physicochemical properties, chemical reactivity, biological activity, etc. of the compounds. Since many parameters could be of contemporary importance, it is crucial to understand, rationalize, and interpret the covariance, which is inherent in this environmental complexity. Explorative methods of multivariate analysis, applied to various topics of environmental concern, give a combined view that generates ranking of the studied chemicals and highlight variable relationships. Then, based on these chemoinformatic tools, a more focused investigation can be made into chemicals of higher concern, guiding experimental tests on the prioritized compounds.

A multivariate explorative technique, such as principal component analysis (PCA), is used by several researchers (e.g., Knekta et al. 2004; Öberg and Iqbal

2012) to visualize the distribution of chemicals, represented by structural descriptors or environmental properties, with the aim to select representative compounds for experimental testing or to highlight the structural properties more related to specific hazardous behavior.

Some of my researches, summarized in this chapter, led to the proposal of a cumulative index, based on the outcome of the application of PCA for chemical screening and ranking. This index (the PC1 score) condenses the main information related to the studied properties. If it explains a reasonably significant variance of the studied variables, it can be usefully modeled as a new aggregate endpoint by QSAR approaches. The developed QSAR models exploit the already available information and can be used to predict the potential behavior of chemicals without experimental values or of new chemicals even before their synthesis. In fact, the QSAR approach, which is based on theoretical molecular descriptors that can be calculated for whatever drawn chemicals, can be applied without knowledge of any experimental parameter.

Similarly, another chemoinformatic method, such as hierarchical cluster analysis, can be applied for grouping chemicals according to their similarity based on several properties. The obtained groups can then be modeled by QSAR classification methods.

Many studies and published papers of the author 20-year researches at Insubria University are focused on this combined approach of chemoinformatic analysis (ranking indexes from PCA plus regression QSAR models of these cumulative indexes; groups modeled as classes by classification QSAR models) for prioritization aims. These will be summarized in the following paragraphs, organized according to the studied endpoints.

Multivariate Explorative Methods: PCA and HCA

The multivariate explorative techniques have the principal aim to condense the information, present in any multivariate data set, into a more easily interpretable view.

Principal component analysis (PCA) (Jackson 1995; Jolliffe 2002) is probably the most widely known and used explorative multivariate method. In PCA, the studied variables are linearly combined so that the obtained combinations (the principal components, PCs) explain the variation in the original data with decreasing explained variance. The first principal component (PC1) condenses the maximum amount of possible data variance in a single variable, while the following orthogonal PCs account for successively smaller quantities of the original variance.

To be useful and be considered sufficiently representative of the main information included in the data, it is desirable that the first two PCs account for a substantial proportion of the variance in the original data, while the remaining PCs condense irrelevant information and noise and could be disregarded. The more common representations of PCA are score plot, loading plot, and biplot, defined as the joint representation of the rows and columns of a data matrix: the points (scores) represent the chemicals, while the vectors or lines represent the variables (loadings). The length of each vector indicates the information associated with that specific variable, while the angle between the vectors reflects their correlation.

Hierarchical cluster analysis (HCA) (Kaufman and Rousseeuw 1990) is a clustering technique with the purpose to build a binary tree of the data that successively merges similar groups of points. HCA creates clusters according to the measure of distance or similarity between data points, based on measured characteristics, connecting, in an iterative process, the nearest groups of objects. It is based on the idea that objects are more related to nearby objects than to objects farther away. The main output of the HCA is the dendrogram that summarizes the relationships between objects in a visual binary tree. Clusters can be identified cutting the dendrogram at different similarity levels. Hierarchical cluster analysis is one of the best ways to observe how homogeneous groups of objects with similar properties are formed and to identify classes.

QSAR Modeling of Ranking Indexes and Classes

In our environmental chemistry studies, PCA for obtaining ranking indexes and regression QSAR modeling of these ranking indexes, as new cumulative endpoints, and HCA for obtaining *a priori* classes and classification QSAR models have been widely used in my group for screening, ranking, and priority setting in many contexts. I'll cite here and comment below only some of the most significant and/or recent: (a) environmental partitioning and leaching of pesticides (Gramatica and Di Guardo 2002; Gramatica et al. 2004a) and benzotriazoles (Bhhatarai and Gramatica 2011b), (b) degradability of volatile organic compounds (VOCs) by tropospheric oxidants (Gramatica et al. 2004b), (c) persistence of POPs by global half-life index (GHLI) (Gramatica and Papa 2007; Papa and Gramatica 2008), (d) rat/mouse toxicity of perfluorinated compounds (PFCs) (Bhhatarai and Gramatica 2010, 2011a), (e) aquatic toxicity of personal care products (PCPs) (Gramatica et al. 2016b) and pharmaceuticals (Sangion et al. 2015; Sangion and Gramatica 2016b), and (f) PBT screening of various compounds by PBT Index (Papa and Gramatica 2010; Gramatica et al. 2015, 2016a; Cassani and Gramatica 2015; Sangion and Gramatica 2016a).

Environmental Behavior

Leaching of Pesticides

The tendency of pesticides to pollute groundwaters and, in general, the partitioning of pesticides into different environmental compartments depend mainly on physicochemical properties, such as soil organic carbon partition coefficient (Koc), n-octanol/water partition coefficient (Kow), water solubility (Sw), vapor pressure

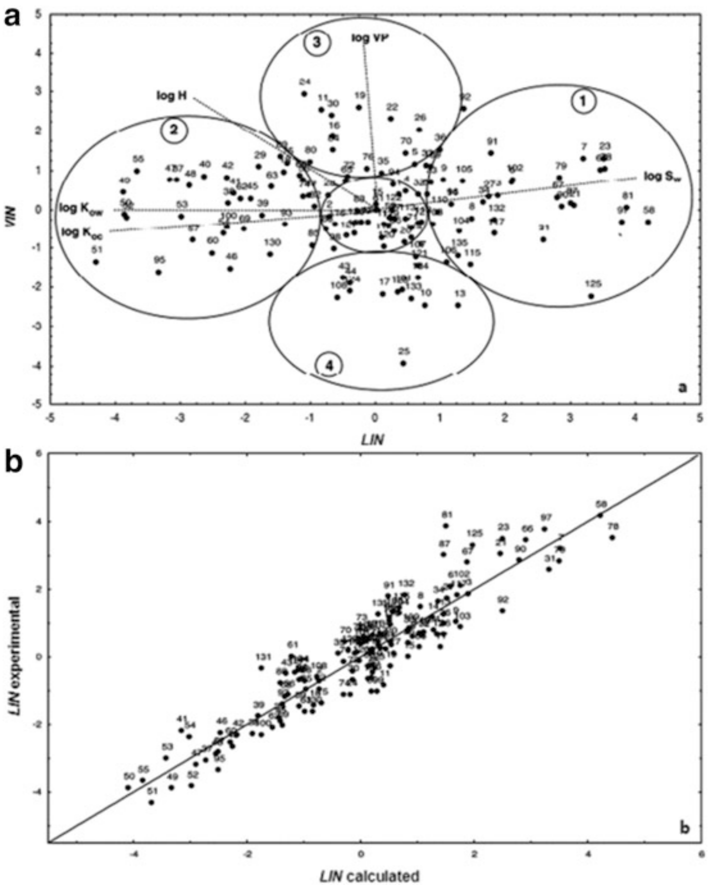

Fig. 1 (**a**) PCA of environmental physicochemical properties of 135 pesticides and definition of leaching index (*LIN*) and volatility index (*VIN*). (**b**) Scatter plot of the OLS model of LIN (Permission from Gramatica and Di Guardo, *Chemosphere*, 2002)

(VP), and Henry's law constant (H). We have applied PCA on these various environmental partitioning properties of a heterogeneous and highly representative data set of 135 pesticides of different chemical classes (acetanilides, carbamates, dinitroanilines, organochlorines, organophosphates, phenylureas, triazines, triazoles) to study the tendency to leach from soil into the surface and subsurface waters (Gramatica and Di Guardo 2002) (Fig. 1a).

The resultant macrovariables, the PC1 and PC2 scores, were called leaching index (LIN) and volatility index (VIN) and were proposed as cumulative partitioning indexes in different environmental media. The component LIN tends to discriminate between the relatively more sorbed/less soluble (on the left of Fig. 1a) and the less sorbed/more soluble pesticides (on the right), while VIN appears to differentiate between volatile (upper part of Fig. 1a) and nonvolatile compounds.

Both indexes were modeled by OLS QSAR model, using theoretical DRAGON molecular descriptors (Talete 2007) selected by genetic algorithms: they are mainly atom and group count parameters as a number of halogens, nitro groups and sulfur, plus Ms the mean electrotopological state of the molecule related to the polarizability, and a topological descriptor ICR (the radial centric information index) (Bonchev and Rouvray 1991). The model robustness and internal predictive power are satisfactory; below the QSAR model equation for LIN, the cumulative index which explains 65 % of the data variance and the corresponding plot in Fig. 1b are reported:

$$\mathrm{LIN} = -3.04 - 0.96\mathrm{nX} - 2.28\mathrm{nNO2} + 3.42\mathrm{Ms} - 1.74\mathrm{ICR} - 0.45\mathrm{nS}$$

$$n = 135,\ R^2 = 87.0\ \%,\ \mathrm{Q}^2{}_{\mathrm{LOO}} = 85.8\ \%,\ \mathrm{Q}^2{}_{\mathrm{LMO}} = 85.7\ \%,$$

$$\mathrm{s} = 0.66,\ \mathrm{F} = 172.22,\ \mathrm{SDEP} = 0.68\ \text{ and }\ \mathrm{SDEC} = 0.65.$$

A combination of two chemoinformatic methods, principal component analysis for ranking and hierarchical cluster analysis for the definition of four *a priori* classes (Fig. 2a), according to the environmental behavior as soluble, sorbed, volatile, or nonvolatile/medium class, was applied to the environmental physicochemical properties of 54 pesticides of various chemical categories (Gramatica et al. 2004b). The pesticides were finally assigned to the defined four classes by three different classification methods (classification and regression tree (CART, the classification tree in Fig. 2b), k-nearest neighbors (k-NN), and regularized discriminant analysis (RDA)) with misclassification risk in cross validation ranging from 17 % to 18 %. The discriminant variables were simple theoretical molecular descriptors, such as MW, nHDon, and topological Balaban Index (Balaban 1983) (named J in DRAGON). MW, which encodes information on molecule dimension, is able to discriminate the chemicals that are contemporaneously most sorbed in organic soils and least soluble in water; in fact it is well known that biggest molecules have the greatest tendency to bind, by van der Waals forces, to the organic component of the soil (mainly humic acids). The more soluble pesticides, which have the higher possibility to form hydrogen bonds with water molecules, are discriminated by nHDon, the number of groups able to donate hydrogen in the hydrogen bonds. Furthermore, the chemicals with fewer intramolecular hydrogen bonds are the most volatile.

Similar PCA ranking was also useful to highlight which benzotriazoles could be most dangerous for the aquatic compartment (Bhhatarai and Gramatica 2011a): these chemicals, used in the past mainly as pesticides, are now recognized as new contaminants of emerging concern (CEC) for the environment. In fact they are nowadays used as deicing additives and are a major source of pollution predominantly of aquatic resources near the airports of major cities.

The presented chemoinformatic approaches allow the screening of the environmental distribution of pesticides and a rapid predetermination of their possibility to

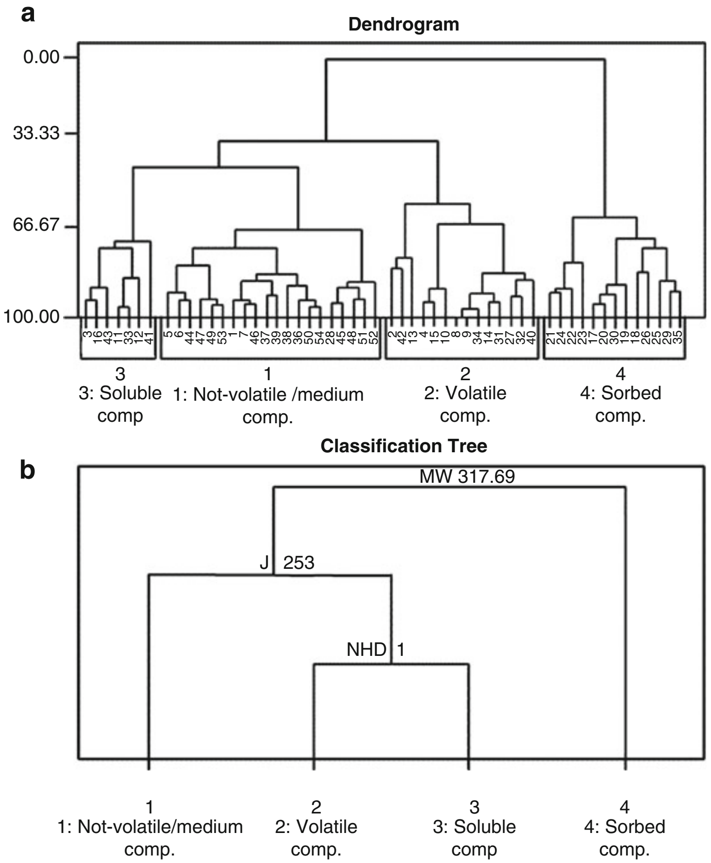

Fig. 2 (**a**) Hierarchical cluster analysis of the environmental physicochemical properties of 54 pesticides for the definition of four *a priori* classes. (**b**) Classification tree of CART model of the four classes (Permission from Gramatica et al. *Int. J. Environ. Anal. Chem*, 2004b)

pollute both surface and groundwaters, starting only from the molecular structure without any *a priori* knowledge of the physicochemical properties.

Persistence

Persistence in the environment is an important criterion in prioritizing hazardous chemicals and in identifying new persistent organic pollutants (POPs).

Various studies, based on various approaches, have been performed on this topic (Öberg 2005, 2006; Muir and Howard 2006; Klasmeier et al. 2006; Wegmann et al. 2009; Howard and Muir 2010; Puzyn et al. 2011; Scheringer et al. 2012). Here some of my studies regarding the screening and prioritization using cumulative indexes are summarized.

Degradability of Volatile Organic Compounds (VOCs) by Tropospheric Oxidants

An indirect measure of the persistence of volatile organic compounds (VOCs) in the atmosphere, and therefore a necessary preliminary parameter in environmental exposure assessment, is the degradability, measured by the reaction rates with the main tropospheric oxidants: hydroxyl radical and ozone during the daytime and nitrate radical at night. The contemporaneous variation and influence of the rate constants of the degradation by three oxidants (kOH, kNO_3, and kO_3) of several VOCs, in determining their inherent tendency to degradability, were explored by principal component analysis (Fig. 3a). The first component, along which the variables are grouped in the same direction, discriminates between the less degradable compounds, so the relatively more persistent (chemicals to the right in the PCA graph) and the more degradable chemicals (chemicals to the left).

Thus, the PC1 score, which explained 81 % of the data variance for 399 chemicals, was defined as ATDIN, an atmospheric degradability index, and was modeled by OLS, based on theoretical molecular descriptors and externally validated (Gramatica et al. 2004a) (scatter plot in Fig. 3b).

It is also interesting to note that the PC2 is able to highlight the different reactivity of chemicals with ozone and with OH and NO_3 radicals, respectively.

$$\mathrm{ATDIN} = -17.59–1.80\mathrm{HOMO} + 2.87\mathrm{nBnz}–0.51\mathrm{BEHe4}$$

$$n_{\mathrm{training}} = 227,\ R^2 = 93.9\ \%,\ \mathrm{Q^2}_{\mathrm{LOO}} = 93.7\ \%,\ \mathrm{Q^2}_{\mathrm{LMO(50\%)}} = 93.5\ \%;$$

$$n_{\mathrm{prediction}} = 172; \mathrm{Q^2}_{\mathrm{EXT}} = 92.3\ \%;$$

$$\mathrm{s} = 0.387; \mathrm{SDEC} = 0.384; \mathrm{SDEP} = 0.391$$

The molecular descriptors of this model are informative of different aspects of the studied reaction. The best descriptor is the energy of the highest occupied molecular orbital (HOMO), as a measure of the molecular reactivity. The number of aromatic rings (nBnz) is probably selected in the model to encode for the different reactivity of aliphatic and aromatic chemicals in relation to the attack sites for the three different oxidants. 2D-BCUT descriptors, BEHe4, weighted by the atomic electronegativity of Sanderson (Burden 1989) encode for charge distribution factors.

This modelcan be useful in avoiding the release into the environment of potential persistent volatile compounds which are not inherently degradable in troposphere, causing risk to humans and wildlife for their persistence at that atmospheric level. Moreover, these chemicals could reach the stratosphere with potential dangerous behavior on the ozone layer.

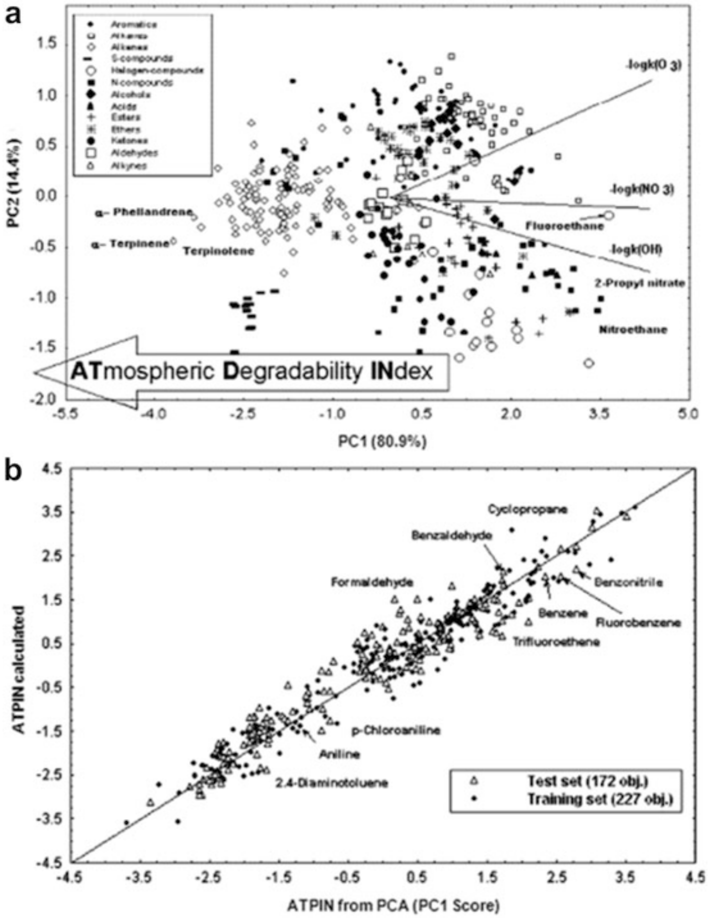

Fig. 3 (**a**) PCA of the kinetic rate constants of degradation by tropospheric oxidants of 399 VOCs and definition of ATDIN. (**b**) Scatter plot of QSAR model of ATDIN (Permission from Gramatica et al. *Atmos Environ*2004a)

Screening of POPs by Environmental Half-Life (HL)

The degradation half-lives in various compartments are among the more commonly used criteria for studying environmental persistence. Available half-life data for degradation in air, water, sediment, and soil, for a set of 250 organic chemicals, were combined in multivariate approach by principal component analysis. A ranking of the studied organic pollutants according to their relative overall half-life is obtained in this way: we named this global half-life index (GHLI) (Gramatica and Papa 2007).

The biplot relative to the first and second components is reported in Fig. 4a, where the chemicals (points) are distributed according to their environmental persistence, represented by the linear combination of their half-lives in the four selected

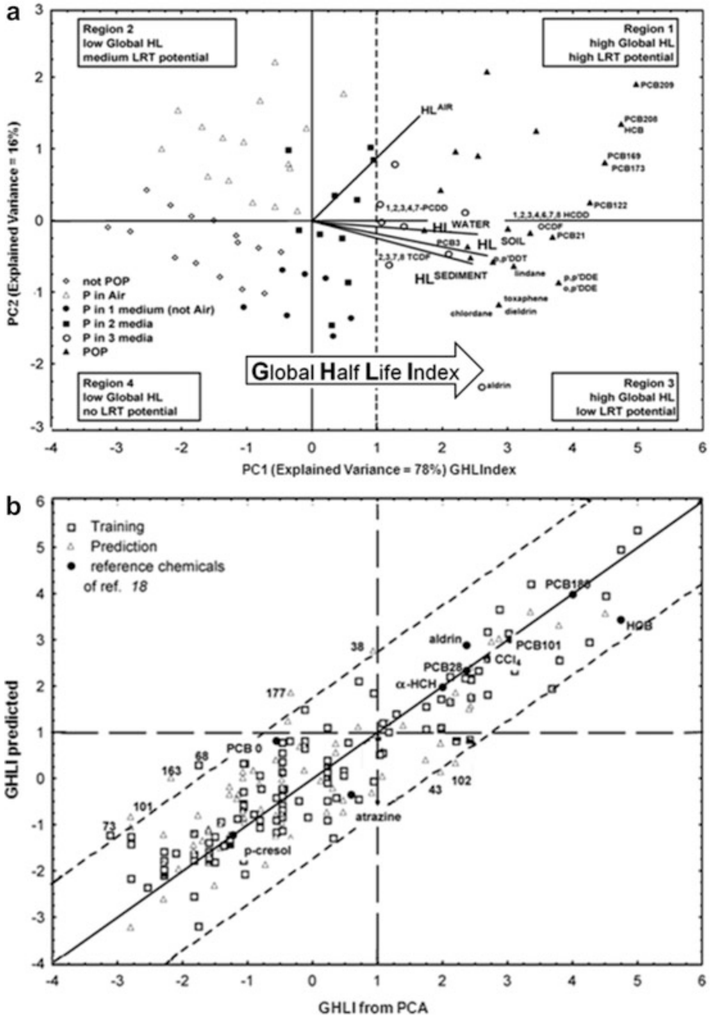

Fig. 4 (**a**) PCA of the half-lives in four environmental compartments of 250 heterogeneous chemicals and definition of global half-life index (*GHLI*) for POPs. (**b**) Scatter plot of the QSAR model of GHLI (Permission from Gramatica and Papa, *Environ. Sci. Technol*., 2007)

media (the loading lines show the importance of each variable in the first two PCs). The cumulative explained variance of the first two PCs is 94 %; the PC1 alone providing a very significant part of the total information is 78 %. Since all the half-lives in different media (the lines) are oriented in the same direction along the first principal component, PC1 is a new macro-variable representing cumulative half-life and condensing chemical tendency to environmental persistence. Therefore, PC1 is useful to discriminate chemicals with regard to persistence: chemicals with high half-life values in all the media are located to the right of the PCA plot (Fig. 4a), in the zone of global higher persistence (very persistent chemicals anywhere); chemicals with a lower global half-life fall to the left of the graph, not being persistent in any medium.

PC2, although less informative (E.V. 16 %), is also interesting: it separates the compounds more persistent in air (upper parts in Fig. 4a, region 1), i.e., those with higher long-range transfer (LRT) potential, from those more persistent in water, soil, and sediment (region 3 in Fig. 4a).

A deeper analysis of the distribution of the studied chemicals confirms experimental evidences: to the right, among the very persistent chemicals in all the compartments (full triangles in Fig. 4a), most of the compounds recognized as POPs by the Stockholm Convention (UNEP 2014) are located. Highly chlorinated PCBs and hexachlorobenzene are among the most persistent compounds in this reference scenario; they are grouped in region 1 owing to their global high persistence, especially in air. The less chlorinated PCBs (PCB 3 and PCB 21), p,p'DDT, p,p'-DDE and o,p'-DDE, highly chlorinated dioxins and some dioxin-like compounds, as well as pesticides toxaphene, lindane, chlordane, dieldrin, and aldrin fall in region 3 of highly persistent chemicals, mainly in compartments different from air.

This global index GHLI was then modeled as a cumulative endpoint using a QSAR approach based on theoretical DRAGON molecular descriptors. The original set of available data was first randomly split into training and prediction sets: 50 % of the compounds (125 compounds) were used for OLS model development, while the other 50 % was put into the prediction set to validate the QSPR model. Given below is the best model, selected by statistical approaches, and its statistical parameters, which confirm model robustness and real external predictivity. Figure 4b shows the plot of GHLI values from PCA versus the predicted GHLI values:

$$\begin{aligned}\text{GHL Index} = {} & -3.12 + 0.33\text{X0v} + 5.1\text{Mv} - 0.32\text{MAXDP} - 0.61\text{nHDon} \\ & - 0.5\text{CIC0} - 0.61\text{O} - 060\end{aligned}$$

$$n_{\text{training}} = 125;\ R^2 = 0.85;\ \text{Q}^2{}_{\text{LOO}} = 0.83;\ \text{Q}^2{}_{\text{BOOT}} = 0.83,$$

$$\text{RMSE} = 0.76;\ \text{RMSE cv} = 0.70;$$

$$n_{\text{prediction}} = 125,\ \ R^2{}_{\text{EXT}} = 0.79;\ \text{RMSEP} = 0.78.$$

A similar highly predictive model for GHLI, based on PaDEL-Descriptors (Yap 2011), has been recently implemented in the module QSARINS-Chem of the software QSARINS (www.qsar.it) for easy applicability on new compounds, in order to help in avoiding the synthesis of chemicals that have, inherent in their molecular structure, the potentiality to be POPs.

The chemical environmental half-lives were also used for POP pre-screening (Papa and Gramatica 2008) developing predictive classification models, based on k-NN, CART, and counter-propagation artificial neural networks (CP-ANN). In this approach, the three *a priori* classes of different degrees of persistence (high, medium, and low) were determined by hierarchical cluster analysis (Fig. 5a) applied

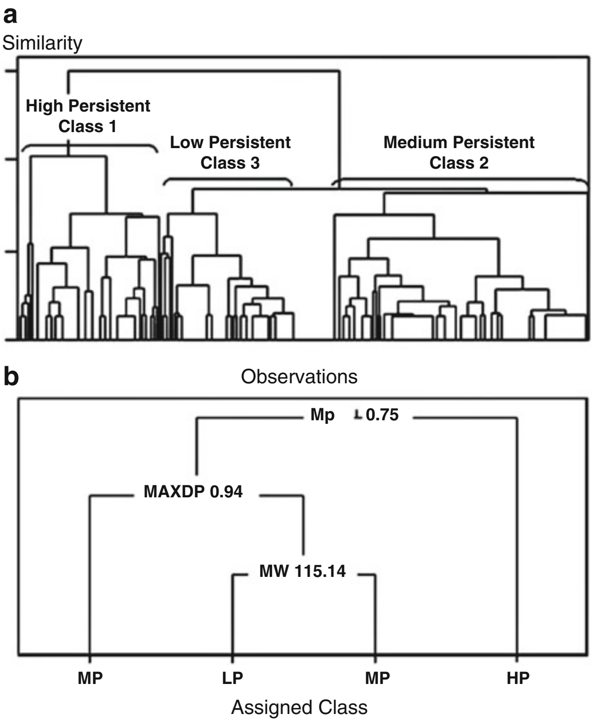

Fig. 5 (**a**) Hierarchical cluster analysis of environmental half-lives for defining three classes of persistence; (**b**) CART tree of three classes for POP ranking (Permission from Papa and Gramatica *J. Mol. Graph. Mod.*, 2008)

to environmental half-lives. The range of overall external predictivity of the three classification models was high, 75–85 %. The three discriminant structural variables selected in this study (mean polarizability, Mp; maximum electrotopological variation, MAXDP; and molecular weight, MW, in the decision nodes of the CART tree in Fig. 5b) are all bidimensional descriptors independent of chemical conformation, thus easily calculable from the bidimensional structural graph of a compound. The calculations of these variables can be performed starting from the simple SMILES string of a chemical.

The application of both kind of models, regression of GHLI and classification of persistence classes, using only a few structural descriptors, could allow a fast preliminary identification and prioritization of not yet known POPs, just from the knowledge of their molecular structure. The proposed multivariate approach is particularly useful not only to screen and to make an early prioritization of environmental persistence for pollutants already on the market, but also for not yet synthesized compounds, which could represent safer alternative and replacement solutions for recognized POPs. No method other than QSAR is applicable to detect the potential persistence of completely new compounds.

Toxicity

Endocrine Disruptors (EDs)

A large number of environmental chemicals are suspected to disrupt endocrine systems by mimicking or antagonizing natural hormones. Such chemicals, named endocrine disruptors (EDs), may have dangerous effects on the health of humans and wildlife. Under REACH, the chemicals with demonstrated endocrine disruption activity require authorization to be produced and used; in addition, safer alternatives should be proposed. However, it is practically impossible to perform a variety of toxicological tests on all potential EDs; thus, QSAR modeling for prioritization has been applied by many authors in these last years (e.g., Shi et al. 2001; Hong et al. 2002; Fang et al. 2003; Schmieder et al. 2003; Tong et al. 2003; Roncaglioni et al. 2004; Asikainen et al. 2004, 2006; Saliner et al. 2006; Devillers et al. 2007, 2015; Dybdahl et al. 2012; Vuorinen et al. 2013; Browne et al. 2015) providing promising methods for the screening of a set of chemicals for potential estrogenic activity.

QSAR models of the estrogen receptor binding affinity of a data set of 128 NCTR heterogeneous compounds (Ding et al. 2010) were built also in our laboratory by OLS method using theoretical DRAGON descriptors (Liu et al. 2006), giving full consideration to the OECD principles regulating QSAR acceptability (OECD 2004) during model construction and assessment.

The results of several validation paths using different splitting methods (D-optimal design, self-organizing maps (SOM), random on activity sampling) give proof that the proposed QSAR model is robust and satisfactory (Q^2_{pred} range, 0.76–0.81), thus providing a feasible and practical tool for the rapid screening of the estrogen activity of organic compounds, supposed EDs. A similar regression model, based on the same training set, and on PaDEL-Descriptors (Yap 2011), is now available in the QSARINS-Chem module of the QSARINS software for easy application on new chemicals (www.qsar.it).

On the same topic, satisfactory predictive models for EDs' classification, based on different classification methods, were proposed (Liu et al. 2007). In this study, QSAR models, based on 232 structurally diverse chemicals from the NCTR database as training set and on theoretical structural descriptors, were developed to quickly and effectively identify possible estrogen-like chemicals by using nonlinear classification methodologies (least squares support vector machine (LS-SVM), counter-propagation artificial neural network (CP-ANN), and k-nearest neighbor (k-NN)). The main descriptor, able to classify alone, with a concordance percentage near 80 %, was nArOH, the number of phenolic groups. Therefore, chemicals with phenolic groups have a great potentiality to be endocrine disruptors. The models were externally validated on 87 chemicals (prediction set) not included in the training set. All three methods gave satisfactory prediction results both for training and prediction sets (accuracy range, 82.8–89.7 %); the most accurate model was obtained by the LS-SVM approach. In addition, our models were also applied for screening a big data set from US-EPA (more than 58,000 discrete organic chemicals)

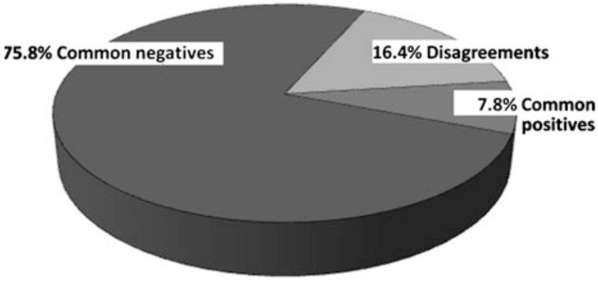

Fig. 6 Classification obtained by consensus of three methods on the screening of more than 50,000 chemicals as potential EDs

to verify the predictions by consensus: about 76 % of the screened chemicals were predicted not to bind to an estrogen receptor (Fig. 6). The obtained results indicate that the proposed QSAR models could provide a feasible and practical tool for the rapid screening of huge data sets and a very useful prioritization approach for focusing experimental tests only on potential estrogens. In fact, the common 40,300 negative compounds could be excluded from the potential ED list without experiments, because the models have high accuracy and low false-negative rate (3–9 %), while costly and lengthy experimental tests should be concentrated on the common positives (7.8 %, meaning >4000) and eventually on the disagreements, which are predicted as EDs by one classification method and not by the others.

An additional screening work was done to classify a big data set of EDs, both estrogen receptor (ER) binders and androgen receptor (AR) antagonists, mainly aiming to improve the external sensitivity in comparison to the literature models (Vinggaard et al. 2008) and to screen for potential AR binders (Li and Gramatica 2010a). The k-NN, the local lazy method and alternating decision tree methods, and the consensus approach were used to build different models, which improved the sensitivity on external chemicals from 57.1 % (Vinggaard et al. 2008) to 76.4 %. The models' predictive abilities were further validated on a blind data set: the sensitivity was even higher >85 %. Then the proposed classifiers were used: (i) to distinguish a set of AR binders into antagonists and agonists, (ii) to screen a combined estrogen receptor binder database to find out possible chemicals that can bind to both AR and ER, and (iii) to virtually screen our in-house environmental chemical database. The *in silico* screening results suggested that: (i) some compounds can affect the normal endocrine system through a complex mechanism because they could bind both to ER and AR; (ii) new EDs, which are non-ER binders, are recognized *in silico* as binders to AR; and (iii) about 20 % of compounds in a big data set of environmental chemicals are predicted as new AR antagonists. Therefore, the priority should be given to them for experimentally testing their binding activities with AR. The results of this complex screening study are summarized in Fig. 7.

Other QSAR models for screening and prioritization of potential EDs were also developed in Insubria group on brominated flame retardants (Kovarich et al. 2011), on perfluorinated chemicals (Kovarich et al. 2012), as well as on other heterogeneous chemicals (Li and Gramatica 2010b, c).

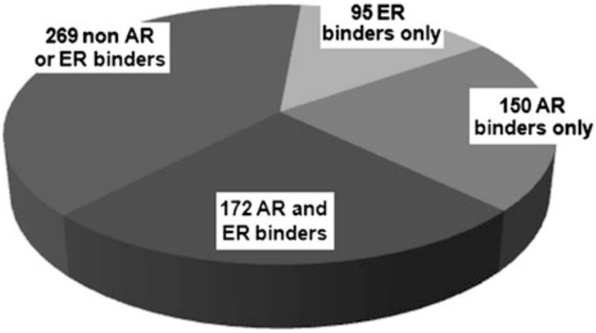

Fig. 7 Results of classification models for endocrine disruptors and prediction of potential receptor binding (estrogen or androgen) (Permission from Li and Gramatica, *J. Chem. Inf. Mod.*, 2010a)

Ecotoxicity

A lot of chemicals that enter into the environment could have dangerous toxic impact on different wild species. Aquatic organisms, such as algae, crustacean, and fish, are normally used as test organism to determine ecotoxicological data. However, also in relation to ecotoxicity, there is a relevant lack of experimental data; therefore, several chemoinformatic approaches have been, and are continuously, applied to model the existing data and to predict ecotoxicological endpoints for various classes of chemicals of high concern (e.g., Vighi et al. 2001; Salvito et al. 2002; Sanderson et al. 2003, 2004; Öberg 2004; Lo Piparo et al. 2006; Roy 2006; Mazzatorta et al. 2006; Sanderson and Thomsen 2009; Gramatica et al. 2012; Sanderson 2012; Kar and Roy 2012; Cassani et al. 2013, 2014; Singh et al. 2014; Roy et al. 2015). The majority of chemical regulators apply ECOSAR models (US EPA 2012), but these models are not always applicable with satisfactory reliability to all kind of chemicals. This problem, related to applicability domain, has been recently highlighted for an important class of emerging environmental pollutants, the pharmaceuticals (Madden et al. 2009), but it is relevant also for other chemicals, such as personal care products (PCPs). Below, some examples of new ecotoxicity studies, recently performed by the Insubria group on these contaminants of emerging concern for the environment (PCPs and pharmaceuticals), are reported.

Aquatic Toxicity of Personal Care Products

PCP ingredients, widely used all over the world, during the last years became chemicals of increasing environmental concern, mainly because they are detected in water and may harm wildlife. Due to their high structural heterogeneity, to the big number of endpoints and the huge lack of experimental data, *in silico* tools, as QSAR models based on structural molecular descriptors, are highly useful to quickly highlight the most hazardous and toxic compounds, prioritizing existing or even not yet synthesized chemicals. In a recent study (Gramatica et al. 2016a), new externally validated QSAR models, specific to predict acute PCP toxicity in three key organisms of aquatic trophic level, i.e., algae (*Pseudokirchneriella subcapitata*), crustacean (*Daphnia magna*), and fish (*Pimephales promelas*), were developed according to the OECD principles for the validation of QSARs, using the QSARINS software (Gramatica et al. 2013). These OLS models, based on theoretical molecular descriptors calculated by free PaDEL-Descriptor, selected by genetic algorithm, are statistically robust and externally predictive (CCC range:89–95 %). They were applied to predict the three modeled acute toxicities for 534 PCPs without experimental data, verifying the wide structural applicability domain of each model by the Insubria graphs (more than 95 % of 534 screened PCPs were into the AD). The root mean squared error (RMSE) of each model for chemicals in the prediction sets, not used for model development, was compared with the corresponding RMSE of the ECOSAR models (US EPA 2012): the RMSE of the Insubria models is always around half logarithmic units, while ECOSAR models exhibit always RMSE values higher than one logarithmic unit. Then, according to the consolidated approach in our group, a trend of cumulative acute aquatic toxicity was highlighted by PCA of the three endpoints of ecotoxicity, allowing the ranking of the overall most toxic for all three trophic levels of the aquatic compartment with more than 90 % of data explained variance (Fig. 8). In the biplot of Fig. 8, the most dangerous PCPs, in the right zone, are highlighted as filled squares, using the multi-criteria decision making (MCDM), included in QSARINS-Chem (Gramatica 2014). MCDM is a technique that summarizes the performances of a certain number of criteria simultaneously, as a single number (score) between 0 and 1. This is done associating to every criteria, in our case different predictions for the studied endpoints, a desirability function which values range from 0 to 1 (where 0 represents the less toxic compound and 1 the most toxic), and giving different weights to the selected criteria. The sum of the weights of the criteria must be 1, and in our case, we used the same weight for each criterion: 0.333, which is 1/3 (total/number

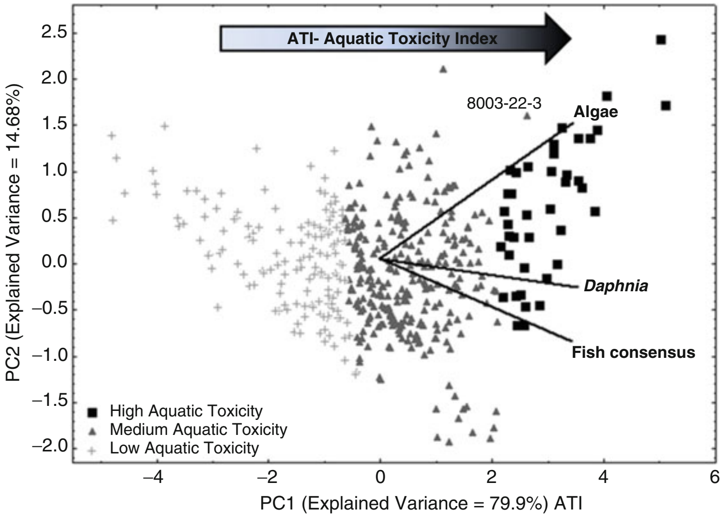

Fig. 8 PCA of experimental and predicted aquatic toxicity data for 484 PCPs and definition of the aquatic toxicity index (*ATI*) (Permission from RCS: Gramatica et al., Green Chem. 2016b)

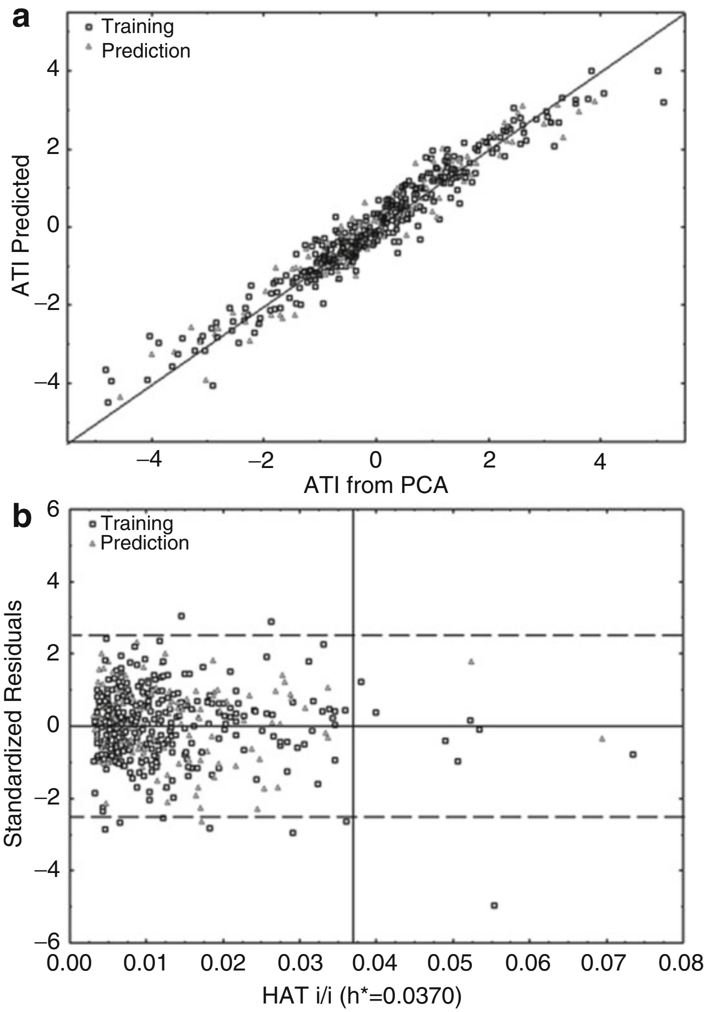

Fig. 9 (**a**) Scatter plot of the OLS model of ATI; (**b**) Williams plot for the AD of ATI model (Permission from RCS: Gramatica et al., Green Chem. 2016b)

of criteria). The geometric average of all the values obtained from the desirability functions gives the MCDM value (i.e., the ranking).

A priority list of 40 most hazardous PCPs was then proposed: it includes mainly UV filters (in particular benzotriazoles), some phthalates, and also some fragrances.

Finally, the trend of PC1, which explains about 80 % of the data variance, was proposed as an aquatic toxicity index (ATI). A QSAR model for the prediction of ATI was developed, by OLS on 484 data and using the PaDEL descriptors (scatter plot in Fig. 9a), to be applicable in QSARINS (www.qsar.it) for the *a priori* detection of not yet tested PCPs and also for the preliminary chemical design of possibly not environmentally hazardous PCPs. The model, with the following equation, is highly predictive and has a good applicability domain verified by the Williams plot of Fig. 9b:

$$\mathrm{ATI} = -14.27 + 0.33\ \mathrm{XlogP} + 18.33\ \mathrm{Mp} + 0.02\ \mathrm{TIC1}$$

$$n_{\mathrm{training}} = 324;\ R^2 = 0.94;\ \mathrm{Q}^2{}_{\mathrm{LOO}} = 0.93; \mathrm{Q}^2{}_{\mathrm{LMO}} = 0.93,\ \mathrm{RMSE} = 0.40;$$
$$\mathrm{RMSE\ cv} = 0.40;$$

$$n_{\mathrm{prediction}} = 160;\ \mathrm{Q}^2{}_{\mathrm{Fn}} = 093 - 0.94;\ \mathrm{CCC}_{\mathrm{EXT}} = 0.97;\ \mathrm{RMSE}_{\mathrm{EXT}} = 0.39.$$

The model is mainly driven by LogP descriptor, here represented by XlogP, which has, as expected, a positive sign in the equation, increasing the toxicity of a chemical. The remaining two descriptors, Mp and TIC1, both with a positive sign in the model equation and thus with a positive impact on the cumulative toxicity trend of PCPs, encode respectively for mean atomic polarizability and total information content index (neighborhood symmetry of 1 order). These three descriptors are mainly related to the complexity and the dimension of the molecule, but also to the presence of electronegative atoms, giving higher values for big and halo-substituted compounds.

Aquatic Toxicity of Pharmaceuticals

Similarly to the modeling and screening developed for PCPs, active pharmaceutical ingredients (APIs) were also studied. Pharmaceuticals have become ubiquitary present in the environment; for this reason in 2006, the European Medicines Evaluation Agency (EMEA) published guidelines for the environmental risk assessment of human pharmaceuticals. Every environmental risk assessment (ERA) requires large amounts of data for each chemical, but, unfortunately, information on ecotoxicological data is available only for a little percentage of APIs in literature and databases. From literature, we collected and carefully curated data sets for the limited quantity of ecotoxicity data of species at different trophic levels (algae, *Daphnia*, fish). For each species, we developed *ad hoc* QSAR acute toxicity models, based on PaDEL molecular descriptors, using the OLS method and genetic algorithm for variable subset selection in QSARINS software (www.qsar.it). All models are robust ($R^2 > 0.75$, $\mathrm{Q}^2{}_{\mathrm{LOO}} > 0.70$) and externally validated (CCC > 0.85) on different splitting schemes, thus allowing reliable application to new pharmaceuticals. The structural applicability domain (AD) of the proposed models to pharmaceuticals without experimental data was verified and demonstrated to be very high with 74 % of chemicals inside the AD for all the toxicity models. The predictions from Insubria models were always better than those obtained by ECOSAR (US EPA 2012) (average RMSE of 0.5 for Insubria models against an average RMSE of 1.3 for ECOSAR). Moreover, reliable predictions from the different models, applied on a set of more than 1000 pharmaceuticals, were combined by PCA to find an ecotoxicity trend representative of the pharmaceuticals' toxicity in the whole aquatic ecosystem. This trend, called overall aquatic toxicity index (ATI) for pharmaceuticals, was then modeled by molecular descriptors to obtain a QSAR model useful to highlight, directly from the chemical structure, the pharmaceuticals potentially most hazardous for the environment. This index, and the predictions obtained by it, could be used to refine procedures of input prevention and control at consumer level as well as *a priori* in the rational design of environmentally safer pharmaceuticals (Sangion et al. 2015; Sangion and Gramatica 2016b; Fig. 10).

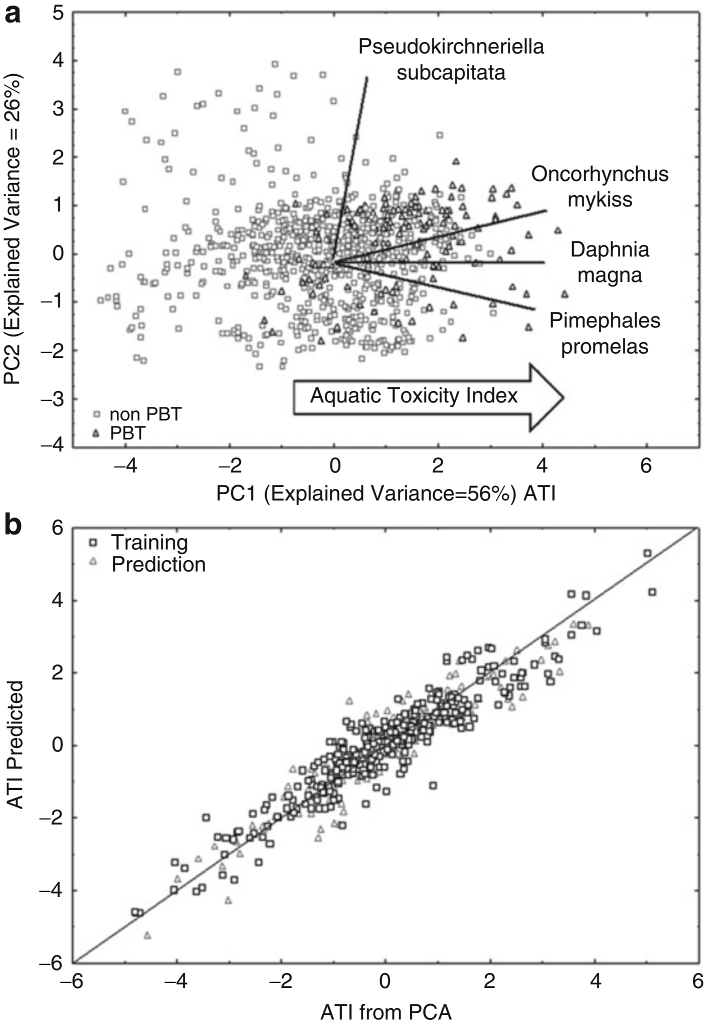

Fig. 10 (**a**) PCA of experimental and predicted aquatic toxicity data for 984 pharmaceuticals and definition of the aquatic toxicity index (*ATI*); (**b**) scatter plot of the OLS model of ATI for pharmaceuticals

Mammalian Toxicity on Rodents of Fluorinated Chemicals

Fully or highly fluorinated compounds, known as per- and polyfluorinated chemicals (PFCs), are widely distributed and released in the environment, because of their use in different household and industrial products. Some long-chain PFCs are classified as emerging pollutants; in fact, they are found undegraded worldwide and even in polar regions, but their environmental and toxicological effects are mostly not well known. In our lab, under the CADASTER FP7 EU Project (www.cadaster.eu), QSAR models of the mammalian toxicity of PFCs on two species of rodents, rat and mouse, were developed, on two endpoints: LC50 data for inhalation (Bhhatarai and Gramatica 2010) and LD50 oral toxicity (Bhhatarai and Gramatica 2011b).

The OLS models, based on DRAGON molecular descriptors selected by genetic algorithm, were always developed on data sets split in different ways (random, by structural similarity using self-organizing maps (SOM)) to verify the satisfactory external predictivity.

Furthermore, to understand the contribution of each toxicity endpoint and the mutual effect of rodent toxicity, the four models were applied individually to a combined data set of 376 compounds, obtaining predictions for those without experimental data. The compounds were checked for their belonging to the structural AD of each model by Insubria graphs, and 204 compounds were found to be within AD. The PCA of experimental and predicted inhalation and oral toxicity on both rodents (rat and mouse) of PFCs, within the AD of all the four models, allowed to rank PFCs according to their cumulated toxicity on rodents. In Fig. 11 the biplot

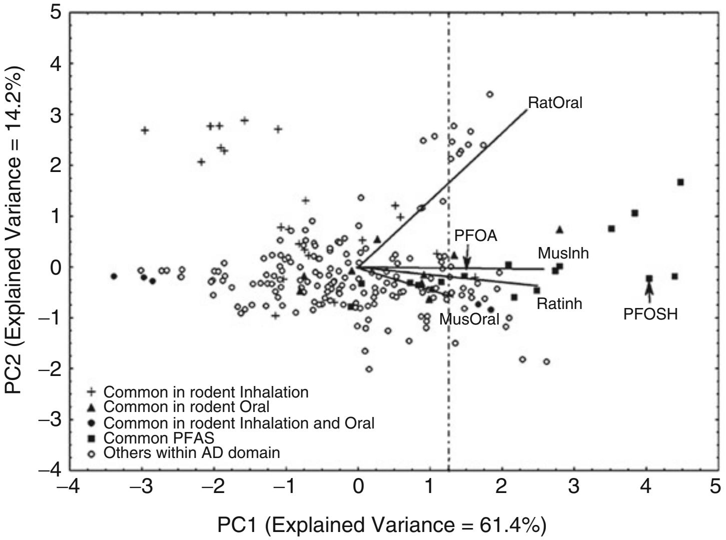

Fig. 11 PCA of inhalation and oral toxicity for two rodent species of 204 PFCs (Permission from Bhhatarai and Gramatica *Mol. Divers.* 2011b)

of this PCA, which explains 75.6 % of data variance, is reported. The compounds with experimental values of rodent inhalation (cross), rodent oral (filled triangles), both rodent inhalation and oral (filled circles), and most common alkylated PFCs (filled squares) are tagged. PC1 ranks the PFCs according to their cumulative rodent toxicity, while PC2 differentiates the PFCs more toxic for rat oral toxicity from those more toxic on the other endpoints. In the right zone of the PCA, the most dangerous and already banned PFCs, as perfluorooctanoic acid (PFOA) and perfluorosulfonic acid (PFSOA), are correctly located: this is a proof of the reliability of the obtained screening and ranking results. Therefore, some other PFCs, which are located in the same zone of concern, should be considered of potential toxicity and prioritized for experimental tests.

Persistence Bioaccumulation Toxicity (PBT)

The chemicals that are contemporaneously persistent, bioaccumulative, and toxic (PBT) are priority chemicals due to the potential risk they pose to humans and ecosystems; therefore, they are considered substances of very high concern (SVHC), which require authorization for use and plan for safer alternatives by REACH. Therefore, in accordance to the precautionary principle, chemicals have to be screened and evaluated for their overall PBT behavior. Unfortunately, for many of the existing substances, even for high production volume (HPV), it is not known, currently, if they could have a potential PBT-like behavior.

Several screening works (Muir and Howard 2006; Howard and Muir 2010; Öberg and Iqbal 2012; Strempel et al. 2012) have highlighted that, among commercial chemicals, many might be PBTs. Therefore, it is evident that not only different approaches for priority setting should be compared to identify the existing potential PBTs but also the "benign by design" approach of green chemistry should be applied for planning the synthesis of safer alternatives to these dangerous compounds (Papa and Gramatica 2010; Strempel et al. 2012; Gramatica et al. 2015, 2016a; Cassani and Gramatica 2015; Sangion and Gramatica 2016a).

Currently, identifying substances as potential PBT or POP candidates relies mainly on determining if specific properties of a chemical exceed threshold values for each property related to PBT behavior (commonly, half-life in various compartments for P, BCF for B, and a number of toxicity evidences for T) (Muir and Howard 2006; Howard and Muir 2010). The most widely used tool for PBT assessment is the US-EPA PBT Profiler (2006), because it is easily applicable from the web (US EPA 2006). The PBT Profiler screens chemicals on the basis of individual P, B, and T properties, calculated by QSAR models and compared to cut off values for each endpoint.

As an alternative approach for PBT prioritization, we developed and proposed a new tool for the screening of chemicals for their potential cumulative PBT behavior (Papa and Gramatica 2010), as an inherent property of a compound that makes it potentially hazardous. PCA of overall half-life in various compartments (the global half-life index, commented above (Gramatica and Papa 2007)), for taking into

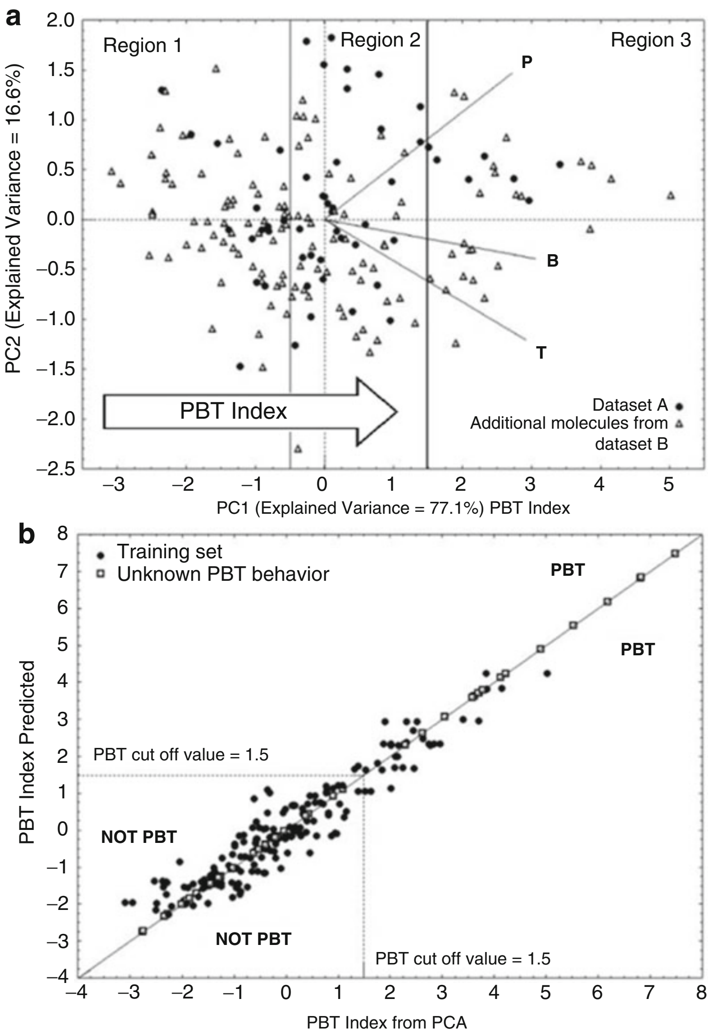

Fig. 12 (**a**) PCA of persistence, bioaccumulation, and toxicity for 180 heterogeneous chemicals and definition of the PBT Index; (**b**) scatter plot of the QSAR model of the PBT Index (Permission from Papa and Gramatica, *Green Chem*. 2010)

account the chemical persistence (P), bioconcentration factor (BCF) (Gramatica and Papa 2005) for bioaccumulation (B), and aquatic toxicity on *Pimephales promelas* (Papa et al. 2005) for toxicity (T), allowed to rank 180 representative non-PBT and PBT chemicals according to their cumulative PBT behavior. In fact, since the loadings of the three properties (P, B, and T) are oriented in the same direction in the PCA biplot (Fig. 12a), the PBTs are ranked on the right of the plot. Therefore, the

PC1 score, which explains more than 77 % of the data variance, can be considered as a PBT Index. The cutoff for PBTs was arbitrary fixed at 1.5 score value, by comparison with the thresholds of the criteria normally applied for PBTs and very P very B (vPvB). It is interesting to note that this Index is precautionary; in fact, also some chemicals with only two thresholds of criteria exceeded are located in the right zone (>1.5 of PC1 score): these compounds would be not recognized as PBT by the normally applied criteria (as those in the US-PBT Profiler), while all the chemicals located in region 3 are considered PBTs by the PBT Index.

This Insubria PBT Index was then modeled by OLS QSAR model, using four simple molecular descriptors, with high verified external predictivity ($Q^2{}_{EXT}$ =80 %). Thus, our work was based on two different steps according to our consolidated approach: (a) the application of a multivariate tool, such as PCA, for screening chemicals according to their cumulative PBT properties and for the definition of the PC1 score as a PBT Index (biplot in Fig. 12a) and (b) the development of a QSAR model of the PBT Index (scatter plot in Fig. 12b).

Our PBT Index model was recently redeveloped using the PaDEL-Descriptor (Yap 2011), freely calculable online: therefore, the model is now easily applicable in the module QSARINS-Chem of the software QSARINS (Gramatica et al. 2013, 2014) (www.qsar.it).

$$\text{PBT Index} = -1.42 + 0.65\ \text{nX} + 0.22\ \text{nBondsM} - 0.41\ \text{nHBDon\_Lipinksi} - 0.09\ \text{MAXDP2}$$

$$n_{\text{training}} = 92,\ R^2 = 0.89,\ \text{RMSE} = 0.52,\ \text{Q}^2{}_{\text{LOO}} = 0.88,\ \text{Q}^2{}_{\text{LMO30\%}} = 0.87,\ \text{RMSE}_{\text{CV}} = 0.55,$$

$$n_{\text{prediction}} = 88,\ \text{Q}^2{}_{\text{F1}} = 0.89; \text{CCC}_{\text{EXT}} = 0.94; \text{RMSE}_{\text{EXT}} = 0.49.$$

The descriptors, selected for the best model by the genetic algorithm procedure, are (in descending order of importance) nX (number of halogen atoms), nBondsM (number of multiple bonds), nHDon_Lipinski (number of donor atoms for H bonds), and MAXDP2 (maximal electrotopological positive variation). All of these parameters are mono- or bidimensional 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, nX and nBM, which encode for substitution with halogens and unsaturation, are known to increase the PBT behavior of chemicals. On the contrary, MAXDP and nHDon 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.

Recently, we have extensively applied our PBT Index model for the screening of large data sets of hundreds of heterogeneous chemicals (part 1 of the series

"Early PBT assessment and prioritization of emerging environmental contaminants" in Gramatica et al. 2015), to personal care product (PCP) ingredients (part 2 in Cassani and Gramatica 2015), to flame retardants (part 3 in Gramatica et al. 2016a), and to pharmaceuticals (part 4 in Sangion and Gramatica 2016a). In all these screening studies, we have compared the PBT Index results with those obtained by the US-EPA PBT Profiler, and we have proposed to consider as highly reliable the predictions obtained in agreement by two methods (plot for flame retardants in Fig. 13a).

It is interesting to note that in the screening of flame retardants (FRs), some supposed "safer alternatives" to the banned FRs, which are already in commerce, are detected as intrinsically hazardous for their PBT properties. They are "regrettable substitutions." If reliable predictive models, based on the chemical structure, would be more often applied *a priori* in a green chemistry approach, from the very beginning of the product development process, it would be possible to avoid the continuous placing on the market, and consequently in the environment, of compounds that will be identified as PBTs, only several years after their use.

Regarding the Insubria PBT Index, we have verified that it is, in the majority of the screenings, more precautionary in highlighting more compounds as PBTs, and the analysis of the disagreements, based on experimental evidences, supports our results for the majority of the cases. It is also important to highlight that the prediction by PBT Index for new chemicals can be verified for the model applicability domain by the Insubria graph (Fig. 13b).

From these screening and prioritization studies, we have verified the reliability of the Insubria PBT Index and, in general, the satisfactory agreement with the PBT Profiler (always >70 %). Main interesting results are:

(a) In the screening of the Insubria data set of 2780 chemicals of environmental concern (Gramatica et al. 2015), included in the QSARINS-Chem module of the software QSARINS (Gramatica et al. 2013, 2014) (www.qsar.it), the compounds predicted as PBTs by consensus (agreement >82 %) are more than 300.

(b) In the screening of 534 PCP ingredients, only eight are prioritized as PBTs by consensus of two methods: they are mainly UV filters as benzotriazoles (Cassani and Gramatica 2015).

(c) In the screening of 128 flame retardants (FRs) (Fig. 13), some already banned and some on the market as substitutes, 30 FRs, which are supposed "safer alternatives," are predicted as PBTs by both modeling tools in agreement (Gramatica et al. 2016a).

(d) In the screening of 1267 pharmaceutical ingredients, only 35, of various therapeutic categories, were included in a priority list of potential PBTs, while 83 % of the screened pharmaceuticals are predicted as non-PBTs by consensus (Sangion and Gramatica 2016a).

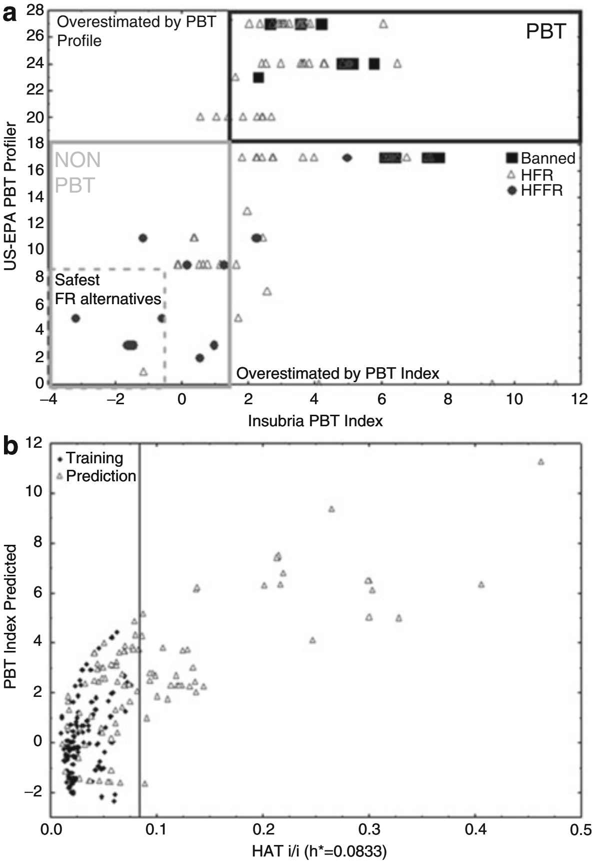

Fig. 13 (**a**) Graph of the agreement between Insubria PBT Index and US-EPA PBT Profiler for the flame retardants study (in the graph are labeled banned flame retardant, halogen flame retardants (*HFRs*), and halogen-free flame retardants (*HFFRs*)) (Permission from Gramatica et al., *J. Hazard. Mater*. 2016a), (**b**) Insubria graph for the AD of Flame Retardant model

Conclusions

In this chapter, a review of the most recent and significant studies performed for prioritization of chemicals in the Insubria QSAR research group is presented. These studies are based on the fundamental assumption that the hazard of any chemical is an inherent property of the molecular structure; therefore, that QSAR models are an incomparable tool to extract and exploit the information related to any physicochemical or biological property of compounds with available experimental data. The application of externally validated predictive QSAR models to chemicals without experimental data, which are into the model applicability domain, is useful for highlighting the potentially most hazardous compounds in the screening and priority setting of big data sets of chemicals. This prioritization can allow to concentrate experiments on prioritized chemicals, thus reducing time, costs, and animal test, but also to avoid the synthesis, and introduction to the market and into the environment, of harmful compounds, which could be recognized dangerous only after evidence of human health concerns has been manifested. This is the basis of the “benign by design” approach of green chemistry.

In these studies, we have demonstrated that, taking into account that the chemical behavior derives from a contemporaneous combination of variables, the application of various chemoinformatic methods, such as explorative analysis by PCA and HCA, is useful in ranking and grouping chemicals according to their cumulative behavior, based on more than a single property or activity. In this way, ranking indexes can be proposed for priority-setting purposes and can be also modeled by QSAR to exploit the fundamental information inherent in the chemical structure. The possibility to continuously contaminate the environment with “regrettable substitutions” could be highly reduced if *a priori* screenings, by combining QSAR models and other chemometric approaches, will be more widely applied. Unfortunately, so far, we do not learn enough from the past knowledge, and we do not take advantage, in an extensive way, from the existing information on the inherent hazard that is included in the chemical structure. This should be done, but more expertise on chemoinformatic method for prioritization should be present also at chemical regulation level.

Acknowledgments All my collaborators in the researches performed at the QSAR Research Unit of Insubria University are greatly acknowledged, in particular now Alessandro Sangion for his help in the preparation of the manuscript.

Bibliography

Arnot, J. A., Brown, T. N., Wania, F., et al. (2012). Prioritizing chemicals and data requirements for screening-level exposure and risk assessment. *Environmental Health Perspectives, 120*, 1565–1570. doi:10.1289/ehp.1205355.

Asikainen, A. H., Ruuskanen, J., & Tuppurainen, K. A. (2004). Consensus kNN QSAR: A versatile method for predicting the estrogenic activity of organic compounds in silico. A comparative study with five estrogen receptors and a large, diverse set of ligands. *Environmental Science and Technology, 38*, 6724–6729. doi:10.1021/es049665h.

Asikainen, A., Kolehmainen, M., Ruuskanen, J., & Tuppurainen, K. (2006). Structure-based classification of active and inactive estrogenic compounds by decision tree, LVQ and kNN methods. *Chemosphere, 62*, 658–673. doi:10.1016/j.chemosphere.2005.04.115.

Balaban, A. T. (1983). Topological indices based on topological distances in molecular graphs. *Pure and Applied Chemistry, 55*, 199–206.

Bhhatarai, B., & Gramatica, P. (2010). Per- and polyfluoro toxicity (LC50 Inhalation) study in rat and mouse using QSAR modeling. *Chemical Research in Toxicology, 23*, 528–539. doi:10.1021/tx900252h.

Bhhatarai, B., & Gramatica, P. (2011a). Modelling physico-chemical properties of (benzo)triazoles, and screening for environmental partitioning. *Water Research, 45*, 1463–1471. doi:10.1016/j.watres.2010.11.006.

Bhhatarai, B., & Gramatica, P. (2011b). Oral LD50 toxicity modeling and prediction of per- and polyfluorinated chemicals on rat and mouse. *Molecular Diversity, 15*, 467–476. doi:10.1007/s11030-010-9268-z.

Bonchev, D., & Rouvray, D. H. (1991). *Chemical graph theory*. New York: Gordon & Breach.

Brown, T. N., & Wania, F. (2008). Screening chemicals for the potential to the persistent organic pollutants: A case study of Arctic contaminants. *Environmental Science and Technology, 42*, 5202–5209. doi:10.1021/es8004514.

Browne, P., Judson, R. S., Casey, W. M., et al. (2015). Screening chemicals for estrogen receptor bioactivity using a computational model. *Environmental Science and Technology, 49*, 8804–8814. doi:10.1021/acs.est.5b02641.

Burden, F. (1989). Molecular-identification number for substructure searches. *Journal of Chemical Information and Computer Science, 29*, 225–227. doi:10.1021/ci00063a011.

Cassani, S., & Gramatica, P. (2015). Identification of potential PBT behavior of personal care products by structural approaches. *Sustainable Chemistry and Pharmacy, 1*, 19–27. doi:10.1016/j.scp.2015.10.002.

Cassani, S., Kovarich, S., Papa, E., et al. (2013). Daphnia and fish toxicity of (benzo)triazoles: Validated QSAR models, and interspecies quantitative activity–activity modelling. *Journal of Hazardous Materials, 258–259*, 50–60. doi:10.1016/j.jhazmat.2013.04.025.

Cassotti, M., Ballabio, D., Consonni, V., et al. (2014). Prediction of acute aquatic toxicity toward daphnia magna by using the GA-kNN method. *Alternatives to Laboratory Animals, 42*, 31–41.

Chirico, N., & Gramatica, P. (2011). Real external predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *Journal of Chemical Information and Modeling, 51*, 2320–2335. doi:10.1021/ci200211n.

Chirico, N., & Gramatica, P. (2012). Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *Journal of Chemical Information and Modeling, 52*, 2044–2058. doi:10.1021/ci300084j.

Devillers, J., Marchand-Geneste, N., Dore, J. C., et al. (2007). Endocrine disruption profile analysis of 11,416 chemicals from chemometrical tools? *SAR and QSAR in Environmental Research, 18*, 181–193. doi:10.1080/10629360701303669.

Devillers, J., Bro, E., & Millot, F. (2015). Prediction of the endocrine disruption profile of pesticides. *SAR and QSAR in Environmental Research, 26*, 831–852. doi:10.1080/1062936X.2015.1104809.

Ding, D., Xu, L., Fang, H., et al. (2010). The EDKB: An established knowledge base for endocrine disrupting chemicals. *BMC Bioinformatics, 11*, S5. doi:10.1186/1471-2105-11-S6-S5.

Dix, D. J., Houck, K. A., Martin, M. T., et al. (2007). The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicological Sciences, 95*, 5–12. doi:10.1093/toxsci/kfl103.

Dybdahl, M., Nikolov, N. G., Wedebye, E. B., et al. (2012). QSAR model for human pregnane X receptor (PXR) binding: Screening of environmental chemicals and correlations with genotoxicity, endocrine disruption and teratogenicity. *Toxicology and Applied Pharmacology, 262*, 301–309. doi:10.1016/j.taap.2012.05.008.

EC Regulation. (2006). Registration, evaluation, authorisation and restriction of chemicals (REACH). Regulation (EC) No. 1907/2006 of the European Parliament and of the Council.

Fang, H., Tong, W. D., Branham, W. S., et al. (2003). Study of 202 natural, synthetic, and environmental chemicals for binding to the androgen receptor. *Chemical Research in Toxicology, 16*, 1338–1358. doi:10.1021/tx030011g.

Gramatica, P. (2007). Principles of QSAR models validation: Internal and external. *QSAR and Combinatorial Science, 26*, 694–701. doi:10.1002/qsar.200610151.

Gramatica, P. (2009). Chemometric methods and theoretical molecular descriptors in predictive QSAR Modeling of the environmental behavior of organic pollutants, Chapter 12. In T. Puzyn, J. Leszczynski, & M. T. Cronin (Eds.), *Recent advances in QSAR studies* (pp. 327–366). New York: Springer.

Gramatica, P. (2014). External evaluation of QSAR models, in addition to cross validation: Verification of predictive capability on totally new chemicals. *Molecular Informatics, 33*, 311–314. doi:10.1002/minf.201400030.

Gramatica, P., & Di Guardo, A. (2002). Screening of pesticides for environmental partitioning tendency. *Chemosphere, 47*, 947–956. doi:10.1016/S0045-6535(02)00007-3.

Gramatica, P., & Papa, E. (2005). An update of the BCF QSAR model based on theoretical molecular descriptors. *QSAR & Combinatorial Science, 24*, 953–960. doi:10.1002/qsar.200530123.

Gramatica, P., & Papa, E. (2007). Screening and ranking of POPs for global half-life: QSAR approaches for prioritization based on molecular structure. *Environmental Science and Technology, 41*, 2833–2839. doi:10.1021/es061773b.

Gramatica, P., Pilutti, P., & Papa, E. (2004a). A tool for the assessment of VOC degradability by tropospheric oxidants starting from chemical structure. *Atmospheric Environment, 38*, 6167–6175. doi:10.1016/j.atmosenv.2004.07.026.

Gramatica, P., Papa, E., & Battaini, B. (2004b). Ranking and classification of non-ionic organic pesticides for environmental distribution: A QSAR approach. *International Journal of Environmental Analytical Chemistry, 84*, 65–74. doi:10.1080/0306731031000149732.

Gramatica, P., Cassani, S., Roy, P. P., et al. (2012). QSAR modeling is not “push a button and find a correlation”: A case study of toxicity of (benzo-)triazoles on algae. *Molecular Informatics, 31*, 817–835. doi:10.1002/minf.201200075.

Gramatica, P., Chirico, N., Papa, E., et al. (2013). QSARINS: A new software for the development, analysis and validation of QSAR MLR models. *Journal of Computational Chemistry, 34*, 2121–2132. doi:10.1002/jcc.23361.

Gramatica, P., Cassani, S., & Chirico, N. (2014). QSARINS-Chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. *Journal of Computational Chemistry, 35*, 1036–1044. doi:10.1002/jcc.23576.

Gramatica, P., Cassani, S., & Sangion, A. (2015). PBT assessment and prioritization by PBT Index and consensus modeling: Comparison of screening results from structural models. *Environmental International, 77*, 25–34. doi:10.1016/j.envint.2014.12.012.

Gramatica, P., Cassani, S., & Sangion, A. (2016a). Are some “safer alternatives” hazardous as PBTs? The case study of new flame retardants. *Journal of Hazardous Materials, 306*, 237–246. doi:10.1016/j.jhazmat.2015.12.017.

Gramatica, P., Cassani, S., & Sangion, A. (2016b). Aquatic Ecotoxicity of Personal Care Products: QSAR models and ranking for prioritization and safer alternatives’ design. (*Green Chemistry, Advance Article, online.* doi: 10.1039/C5GC02818C.).

Guillen, D., Ginebreda, A., Farre, M., et al. (2012). Prioritization of chemicals in the aquatic environment based on risk assessment: Analytical, modeling and regulatory perspective. *Science of the Total Environment, 440*, 236–252. doi:10.1016/j.scitotenv.2012.06.064.

Hansson, S. O., & Rudén, C. (2006). Priority setting in the REACH system. *Toxicological Sciences: An Official Journal of the Society of Toxicology, 90*, 304–308. doi:10.1093/toxsci/kfj071.

Hong, H. X., Tong, W. D., Fang, H., et al. (2002). Prediction of estrogen receptor binding for 58,000 chemicals using an integrated system of a tree-based model with structural alerts. *Environmental Health Perspectives, 110*, 29–36. doi:10.1289/ehp.0211029.

Howard, P. H., & Muir, D. C. G. (2010). Identifying new persistent and bioaccumulative organics among chemicals in commerce. *Environmental Science and Technology, 44*, 2277–2285. doi:10.1021/es903383a.

Jackson, J. E. (1995). Review of a user’s guide to principal components. *Journal of Educational and Behavioral Statistics, 20*, 105–107. doi:10.2307/1165392.

Jolliffe, I. T. (2002). *Principal component analysis*. New York: Springer.

Judson, R., Richard, A., Dix, D. J., et al. (2009). The toxicity data landscape for environmental chemicals. *Environmental Health Perspectives, 117*, 685–695. doi:10.1289/ehp.0800168.

Kar, S., & Roy, K. (2012). Risk assessment for ecotoxicity of pharmaceuticals – An emerging issue. *Expert Opinion on Drug Safety, 11*, 235–274. doi:10.1517/14740338.2012.644272.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Hoboken: Wiley.

Kavlock, R., & Dix, D. (2010). Computational toxicology as implemented by the Us Epa: Providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk. *Journal of Toxicology and Environmental Health, Part B: Critical Reviews, 13*, 197–217. doi:10.1080/10937404.2010.483935.

Klasmeier, J., Matthies, M., Macleod, M., et al. (2006). Application of multimedia models for screening assessment of long-range transport potential and overall persistence. *Environmental Science and Technology, 40*, 53–60. doi:10.1021/es0512024.

Knekta, E., Andersson, P. L., Johansson, M., & Tysklind, M. (2004). An overview of OSPAR priority compounds and selection of a representative training set. *Chemosphere, 57*, 1495–1503. doi:10.1016/j.chemosphere.2004.07.056.

Kovarich, S., Papa, E., & Gramatica, P. (2011). QSAR classification models for the prediction of endocrine disrupting activity of brominated flame retardants. *Journal of Hazardous Materials, 190*, 106–112. doi:10.1016/j.jhazmat.2011.03.008.

Kovarich, S., Papa, E., Li, J., & Gramatica, P. (2012). QSAR classification models for the screening of the endocrine-disrupting activity of perfluorinated compounds. *SAR and QSAR in Environmental Research, 23*, 207–220. doi:10.1080/1062936X.2012.657235.

Li, J., & Gramatica, P. (2010a). Classification and virtual screening of androgen receptor antagonists. *Journal of Chemical Information and Modeling, 50*, 861–874. doi:10.1021/ci100078u.

Li, J., & Gramatica, P. (2010b). The importance of molecular structures, endpoints' values, and predictivity parameters in QSAR research: QSAR analysis of a series of estrogen receptor binders. *Molecular Diversity, 14*, 687–696. doi:10.1007/s11030-009-9212-2.

Li, J., & Gramatica, P. (2010c). QSAR classification of estrogen receptor binders and identification of pleiotropic EDCs. *SAR and QSAR in Environmental Research, 21*, 657–669. doi:10.1080/1062936X.2010.528254.

Liu, H., Papa, E., & Gramatica, P. (2006). QSAR prediction of estrogen activity for a large set of diverse chemicals under the guidance of OECD principles. *Chemical Research in Toxicology, 19*, 1540–1548. doi:10.1021/tx0601509.

Liu, H., Papa, E., Walker, J. D., & Gramatica, P. (2007). In silico screening of estrogen-like chemicals based on different nonlinear classification models. *Journal of Molecular Graphics and Modelling, 26*, 135–144. doi:10.1016/j.jmgm.2007.01.003.

Lo Piparo, E., Smiesko, M., Mazzatorta, P., et al. (2006). Preliminary analysis of toxicity of benzoxazinones and their metabolites for Folsomia candida. *Journal of Agricultural and Food Chemistry, 54*, 1099–1104. doi:10.1021/jf.050916v.

Madden, J. C., Enoch, S. J., Hewitt, M., & Cronin, M. T. D. (2009). Pharmaceuticals in the environment: Good practice in predicting acute ecotoxicological effects. *Toxicology Letters, 185*, 85–101. doi:10.1016/j.toxlet.2008.12.005.

Mazzatorta, P., Cronin, M. T. D., & Benfenati, E. (2006). A QSAR study of avian oral toxicity using support vector machines and genetic algorithms. *QSAR and Combinatorial Science, 25*, 616–628. doi:10.1002/qsar.200530189.

Muir, D. C., & Howard, P. H. (2006). Are there other persistent organic pollutants? A challenge for environmental chemists. *Environmental Science and Technology, 40*, 7157–7166. doi:10.1021/es061677a.

Öberg, T. (2004). A QSAR for baseline toxicity: Validation, domain of application, and prediction. *Chemical Research in Toxicology, 17*, 1630–1637. doi:10.1021/tx0498253.

Öberg, T. (2005). A QSAR for the hydroxyl radical reaction rate constant: Validation, domain of application, and prediction. *Atmospheric Environment, 39*, 2189–2200. doi:10.1016/j.atmosenv.2005.01.007.

Öberg, T. (2006). Virtual screening for environmental pollutants: Structure-activity relationships applied to a database of industrial chemicals. *Environmental Toxicology and Chemistry, 25*, 1178–1183. doi:10.1897/05-326R.1.

Öberg, T., & Iqbal, M. S. (2012). The chemical and environmental property space of REACH chemicals. *Chemosphere, 87*, 975–981. doi:10.1016/j.chemosphere.2012.02.034.

OECD. (2004). Principles for the validation, for regulatory purposes, of (Quantitative) Structure-Activity Relationship Models. http://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm

Papa, E., & Gramatica, P. (2008). Screening of persistent organic pollutants by QSPR classification models: A comparative study. *Journal of Molecular Graphics and Modelling, 27*, 59–65. doi:10.1016/j.jmgm.2008.02.004.

Papa, E., & Gramatica, P. (2010). QSPR as a support for the EU REACH regulation and rational design of environmentally safer chemicals: PBT identification from molecular structure. *Green Chemistry, 12*, 836–843. doi:10.1039/B923843C.

Papa, E., Villa, F., & Gramatica, P. (2005). Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow). *Journal of Chemical Information and Modeling, 45*, 1256–1266. doi:10.1021/ci050212l.

Puzyn, T., Gajewicz, A., Rybacka, A., & Haranczyk, M. (2011). Global versus local QSPR models for persistent organic pollutants: Balancing between predictivity and economy. *Structural Chemistry, 22*, 873–884. doi:10.1007/s11224-011-9764-5.

Roncaglioni, A., Novic, M., Vracko, M., & Benfenati, E. (2004). Classification of potential endocrine disrupters on the basis of molecular structure using a nonlinear modeling method. *Journal of Chemical Information and Computer Science, 44*, 300–309. doi:10.1021/ci030421a.

Roos, V., Gunnarsson, L., Fick, J., et al. (2012). Prioritising pharmaceuticals for environmental risk assessment: Towards adequate and feasible first-tier selection. *Science of the Total Environment, 421*, 102–110. doi:10.1016/j.scitotenv.2012.01.039.

Roy, K. (2006). Ecotoxicological modeling and risk assessment using chemometric tools. *Molecular Diversity, 10*, 93–94. doi:10.1007/s11030-006-9025-5.

Roy, K., Kar, S., & Das, R. (2015). Understanding the basics of QSAR for applications. In *Pharmaceutical sciences and risk assessment* (1st ed.). Amsterdam/Boston: Academic.

Saliner, A. G., Netzeva, T. I., & Worth, A. P. (2006). Prediction of estrogenicity: Validation of a classification model. *SAR and QSAR in Environmental Research, 17*, 195–223. doi:10.1080/10659360600636022.

Salvito, D. T., Senna, R. J., & Federle, T. W. (2002). A framework for prioritizing fragrance materials for aquatic risk assessment. *Environmental Toxicology and Chemistry, 21*, 1301–1308. doi:10.1897/1551-5028(2002)021<1301:AFFPFM>2.0.CO;2.

Sanderson, H. (2012). Challenges and directions for regulatory use of QSARs for predicting active pharmaceutical ingredients environmental toxicity. *Current Drug Safety, 7*, 309–312.

Sanderson, H., & Thomsen, M. (2009). Comparative analysis of pharmaceuticals versus industrial chemicals acute aquatic toxicity classification according to the United Nations classification system for chemicals. Assessment of the (Q)SAR predictability of pharmaceuticals acute aquatic toxicity and their predominant acute toxic mode-of-action. *Toxicology Letters, 187*, 84–93. doi:10.1016/j.toxlet.2009.02.003.

Sanderson, H., Johnson, D. J., Wilson, C. J., et al. (2003). Probabilistic hazard assessment of environmentally occurring pharmaceuticals toxicity to fish, daphnids and algae by ECOSAR screening. *Toxicology Letters, 144*, 383–395. doi:10.1016/S0378-4274(03)00257-1.

Sanderson, H., Johnson, D. J., Reitsma, T., et al. (2004). Ranking and prioritization of environmental risks of pharmaceuticals in surface waters. *Regulatory Toxicology and Pharmacology, 39*, 158–183. doi:10.1016/j.yrtph.2003.12.006.

Sangion, A., & Gramatica, P. (2016a). PBT assessment and prioritization of selected Pharmaceuticals. *Environmental Research, 147*, 297–306. DOI: 10.1016/j.envres.2016.02.021.

Sangion, A. & Gramatica, P. (2016b). Prioritization of Pharmaceuticals of higher concern for water pollution: structural approaches for modeling and prediction of ecotoxicity (Under revision on Environment International).

Sangion, A., Cassani, S., Papa, E., & Gramatica, P. (2015). *Identification of potential environmentally hazardous pharmaceuticals by QSAR modeling*. In SETAC Europe 25th annual meeting Barcelona (Spain).

Scheringer, M., Strempel, S., Hukari, S., et al. (2012). How many persistent organic pollutants should we expect? *Atmospheric Pollution Research, 3*, 383–391. doi:10.5094/APR.2012.044.

Schmieder, P., Mekenyan, O., Bradbury, S., & Veith, G. (2003). QSAR prioritization of chemical inventories for endocrine disruptor testing. *Pure and Applied Chemistry, 75*, 2389–2396. doi:10.1351/pac200375112389.

Shi, L. M., Fang, H., Tong, W. D., et al. (2001). QSAR models using a large diverse set of estrogens. *Journal of Chemical Information and Computer Science, 41*, 186–195. doi:10.1021/ci000066d.

Singh, K. P., Gupta, S., Kumar, A., & Mohan, D. (2014). Multispecies QSAR modeling for predicting the aquatic toxicity of diverse organic chemicals for regulatory toxicology. *Chemical Research in Toxicology, 27*, 741–753. doi:10.1021/tx400371w.

Stenberg, M., Linusson, A., Tysklind, M., & Andersson, P. L. (2009). A multivariate chemical map of industrial chemicals – Assessment of various protocols for identification of chemicals of potential concern. *Chemosphere, 76*, 878–884. doi:10.1016/j.chemosphere.2009.05.011.

Strempel, S., Scheringer, M., Ng, C. A., & Hungerbühler, K. (2012). Screening for PBT Chemicals among the "Existing" and "New" Chemicals of the EU. *Environmental Science & Technology, 46*, 5680–5687.

Talete. (2007). DRAGON for Windows (Software for Molecular Descriptor Calculations). Talete srl. Milano, Italy. www.talete.mi.it

Tong, W. D., Fang, H., Hong, H. X., et al. (2003). Regulatory application of SAR/QSAR for priority setting of endocrine disruptors: A perspective. *Pure and Applied Chemistry, 75*, 2375–2388. doi:10.1351/pac200375112375.

UNEP (2014). Stockholm convention on persistent organic pollutants (POPs). Stockholm, Sweden.

US EPA. (2006). PBT profiler; persistent, bioaccumulative, and toxic profiles estimated for organic chemicals on-line. http://www.pbtprofiler.net/. Accessed 14 Jan 2016.

US EPA. (2012). The ECOSAR (ECOlogical Structure Activity Relationship) class program. www.epa.gov/tsca-screening-tools/ecological-structure-activity-relationships-ecosar-predictive-model. Accessed 14 Jan 2016.

Vighi, M., Gramatica, P., Consolaro, F., & Todeschini, R. (2001). QSAR and chemometric approaches for setting water quality objectives for dangerous chemicals. *Ecotoxicology and Environmental Safety, 49*, 206–220. doi:10.1006/eesa.2001.2064.

Vinggaard, A. M., Niemela, J., Wedebye, E. B., & Jensen, G. E. (2008). Screening of 397 chemicals and development of a quantitative structure-activity relationship model for androgen receptor antagonism. *Chemical Research in Toxicology, 21*, 813–823. doi:10.1021/tx7002382.

Vuorinen, A., Odermatt, A., & Schuster, D. (2013). In silico methods in the discovery of endocrine disrupting chemicals. *Journal of Steroid Biochemistry and Molecular Biology, 137*, 18–26. doi:10.1016/j.jsbmb.2013.04.009.

Wedebye, E. B., Dybdahl, M., Nikolov, N. G., et al. (2015). QSAR screening of 70,983 REACH substances for genotoxic carcinogenicity, mutagenicity and developmental toxicity in the ChemScreen project. *Reproductive Toxicology, 55*, 64–72. doi:10.1016/j.reprotox.2015.03.002.

Wegmann, F., Cavin, L., MacLeod, M., et al. (2009). The OECD software tool for screening chemicals for persistence and long-range transport Potential. *Environmental Modelling and Software, 24*, 228–237. doi:10.1016/j.envsoft.2008.06.014.

Yap, C. W. (2011). PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry, 32*, 1466–1474.

Zarfl, C., Hotopp, I., Kehrein, N., & Matthies, M. (2012). Identification of substances with potential for long-range transport as possible substances of very high concern. *Environmental Science and Pollution Research, 19*, 3152–3161. doi:10.1007/s11356-012-1046-2.