Screening and Ranking of POPs for Global Half-Life: QSAR Approaches for Prioritization Based on Molecular Structure

ARTICLE *in* ENVIRONMENTAL SCIENCE AND TECHNOLOGY · MAY 2007

Impact Factor: 5.33 · DOI: 10.1021/es061773b · Source: PubMed

2 AUTHORS:

Paola Gramatica
Università degli Studi dell'Insubria
176 PUBLICATIONS **7,571** CITATIONS
SEE PROFILE

Ester Papa
Università degli Studi dell'Insubria
60 PUBLICATIONS **1,954** CITATIONS
SEE PROFILE

Environ. Sci. Technol. **2007**, *41,* 2833−2839

Screening and Ranking of POPs for Global Half-Life: QSAR Approaches for Prioritization Based on Molecular Structure

PAOLA GRAMATICA* AND ESTER PAPA

QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria, via Dunant 3, 21100 Varese, Italy

Persistence in the environment is an important criterion in prioritizing hazardous chemicals and in identifying new persistent organic pollutants (POPs). Degradation half-life in various compartments is among the more commonly used criteria for studying environmental persistence, but the limited availability of experimental data or reliable estimates is a serious problem. Available half-life data for degradation in air, water, sediment, and soil, for a set of 250 organic POP-type chemicals, were combined in a multivariate approach by principal component analysis to obtain a ranking of the studied organic pollutants according to their relative overall half-life. A global half-life index (GHLI) applicable for POP screening purposes is proposed. The reliability of this index was verified in comparison with multimedia model results. This global index was then modeled as a cumulative end-point using a QSAR approach based on few theoretical molecular descriptors, and a simple and robust regression model externally validated for its predictive ability was derived. The application of this model could allow a fast preliminary identification and prioritization of not yet known POPs, just from the knowledge of their molecular structure. This model can be applied a priori also in the chemical design of safer and alternative non-POP compounds.

Introduction

High persistence of chemicals in the environment is a recognized dangerous and undesired property, and there is marked concern regarding such environmental behavior in persistent organic pollutants (POPs). Such chemicals persist in the environment, as their degradation by physical, chemical, and biological processes is slow; thus, there is more time for them to accumulate in the environment and biota and to cause chronic effects. Given the evidence of the long-range transport (LRT) of these substances to regions where they have never been used or produced, and the ensuing worldwide pollution, the international community has called for urgent global action to identify, control, and even ban these chemicals. In 1998, the United Nations Environment Program (UNEP) developed the Stockholm Convention on POPs to protect human health and the environment from these kinds of chemicals. Screening criteria for the identification and priority ranking of substances needing more detailed assessment are continuously requested so as to add new possible POPs to a preliminary list of 12 already identified POPs (aldrin, chlordane, DDT, dieldrin, endrin, heptachlor, hexachlorobenzene, mirex, polychlorinated biphenyls, polychlorinated dibenzo-*p*-dioxins, polychlorinated dibenzo-*p*-furans, and toxaphene) (*1*, *2*).

The need for a scientific foundation of the criteria used to evaluate the persistence and long-range transport of such organic chemicals in the environment was recently highlighted in several papers where different approaches with varying levels of complexity were developed and applied (*3*−*21*). The majority of these models, including the OECD recommended multimedia model approaches (*18*−*20*, *22*) employs half-life data from handbooks (*23*, *24*). In fact, the half-life of organic pollutants in various compartments (i.e., the time required for the concentration of a substance to halve its original value in a particular environmental medium), although not easily determined, is among the most commonly used criteria for studying persistence (*5*). Indeed, it is regarded as a key parameter in the assessment of environmental fate, and ecological and human health hazards, but due to the serious difficulties in determining effective decay rates, few organic compounds have experimental data available. According to Klasmeier et al. (*18*), the UNEP approach (*2*) to persistence based on the evaluation of the single-media half-life and defining a substance as persistent if any one of its half-lives exceeds specific criteria in four environmental media (air >2 days, surface water >60 days, and soil or sediment >180 days), is a simplistic approach to screening chemicals in the environment for POP-like behavior. Instead, multimedia models, with the same half-life data input, are quantitative models that compute numerical indicators for overall persistence (*7*, *18*, *19*) and consider chemical partitioning and environmental variables (*9*). Recent papers discuss and compare different multimedia models (*20*), which have been recommended (*18*) as being a more efficient alternative to the single-media half-life approach proposed by the UNEP Stockholm Convention (*1*, *2*).

Multimedia models are indeed invaluable tools for chemicals with known experimental data (half-life data, partitioning properties, and environmental conditions), but there is still a need for complementary methods for simpler and earlier identification of new POPs, even without experimental data, and for directing the synthesis of safer alternatives to POPs.

In relation to these topics, our group has proposed different QSAR approaches, based only on structural information, for the ranking of POPs according to their atmospheric persistence and their inherent tendency toward global mobility or long-range transport (LRT) potential (*25*−*27*). The influence of partitioning properties on the global mobility and LRT potential has been studied, and QSPR (quantitative structure−property relationships) models of regression (*25*, *27*) and classification (*26*, *27*) have been proposed for the preliminary detection of those hazards, also for new chemicals. Furthermore, externally validated predictive QSPR models for the prediction of atmospheric degradation rate constants with OH (*28*, *29*), NO_3 radicals (*28*, *30*), and ozone (*31*) have been developed. Frameworks based on the PCA-QSAR approach to rank VOCs according to their overall tropospheric atmospheric degradability, by applying simple regression models based only on molecular structure descriptors, were recently proposed (*32*, *33*). The present paper proposes a method of general applicability in the POP screening context that can be particularly useful for ranking

* Corresponding author phone: +39-0332-421573; fax: +39-0332-421554; e-mail: paola.gramatica@uninsubria.it.

10.1021/es061773b CCC: $37.00
Published on Web 03/20/2007

and prioritization purposes.

Chemical degradation half-life is defined by Mackay (*11*) as a quasi-intensive property since it is independent of quantity and is a characteristic of a molecule within a defined environmental medium. In the present study, we aim to extract and model only the intensive aspect (the characteristics of the molecule) of the chemical half-life, catching the structural features of a chemical that are related to its intrinsic ability or tendency to persist in the environment, independent of its partitioning properties and environmental conditions. This approach will allow a simpler and preliminary identification (labeling) of a chemical as possibly POP-like, even when the only information available is the molecular structure; thus, identification can take place even before the synthesis of the chemical. This last application is particularly useful as the Stockholm Convention (*2*) stimulates the discovery of new, cheap, and effective alternatives to the world's most dangerous POPs.

The first aim of this paper is to verify whether the combination of half-life data available in various compartments by a multivariate method like the principal component analysis (PCA), will rank chemicals in a reasonable and effective order according to their global environmental persistence. This would result in an index for a global half-life (GHLI) starting from already available knowledge (half-life data). The second and fundamental step highlights the relevance of molecular structure in determining the intrinsic tendency of a molecule to be persistent, independent of quantity, partitioning, and environmental properties: a QSPR (quantitative structure−persistence relationship) model for the prediction of this index, as a global half-life tendency, is thus developed. The proposed QSPR model, rigorously verified for external prediction ability (*34*) on chemicals not participating in model development, could be applied, not only for not yet recognized POPs already present in the environment, but also a priori even before the synthesis of a new chemical. Thus, by exploiting the limited knowledge available, the synthesis can be directed toward safer non-POPs chemicals as effective alternatives to recognized POPs.

Materials and Methods

Half-Life Data. This study dealt with an overall data set proposed by the U.S. EPA; it included 250 organic compounds of known half-lives for transformation into four environmental media (air, water, soil, and sediment) (*24*). Because of the wide half-life range (5−55 000 h), the data were transformed into logarithmic values to linearize the range of variation. This semiquantitative classification of compound half-life was proposed by Mackay (*5*, *24*) as a preferable approach for screening purposes.

This approach, although less precise than estimating a specific numeric value, reduces the variability effect in environmental persistence of a compound by assigning the half-life to one of nine classes on a semi-decade logarithmic scale. This assignment scale is the result of careful analysis of the reaction rates of the chemicals in each medium, considering all relevant degradation processes. The studied data set (Supporting Information Table SI1) is structurally heterogeneous and highly representative of many classes of already defined problematic chemicals. These include some of the most relevant environmental pollutants such as polychlorobiphenyls (PCBs), various pesticides, chlorobenzenes, polychlorodibenzodioxins (PCDD), polychlorodibenzofurans (PCDF), polycyclic aromatic hydrocarbons (PAHs), and also heterogeneous industrial chemicals. Some of the compounds included for screening purposes are among the 12 UNEP POPs, while other POP-like and non-POPs were added to ensure a wider and more representative set of persistence potential in our reference scenario. Half-life data of 10 reference chemicals and three hypothetical chemicals used by Klasmeier et al. (*18*), included here for model comparison, are also listed in Table SI1.

Molecular Descriptors. A set of 662 theoretical molecular descriptors for QSPR modeling was computed using the software DRAGON (*35*). The input files for descriptor calculation containing information on atom and bond types, connectivity, partial charge, and atomic spatial coordinates relative to the minimum energy conformation of the molecule were obtained by the molecular mechanics method of Allinger (MM+) using the package HyperChem (*36*). Constant and near constant descriptors were excluded in a prereduction step; thus, 474 molecular descriptors underwent subsequent selection for the best modeling variables.

The typology of the calculated descriptors is summarized in Table SI2 of the Supporting Information; further information regarding the molecular descriptors can be found in ref *37*.

Explorative Analysis. PCA, in which linear combinations of the input variables (half-life data here) are created, is used as an explorative multivariate method; the new variables, the principal components (PCs), condense the relevant information in the data and allow a powerful global visualization (*38*). The PCs are derived from standardized data (via the data correlation matrix), and the combined biplot of scores (coordinates of chemicals on the new combined variables) and loadings (weights of original variables on which the linear combination PCs are built) is presented. PCA was performed by the software SCAN (*39*).

QSAR Modeling. Multiple linear regression and variable selection were performed by the package MOBY DIGS (*40*) using ordinary least-squares regression (OLS).

Given the possible correlation among the numerous calculated descriptors, and the impossibility of performing multilinear regression on them, a variable selection procedure was necessary: GA-VSS (genetic algorithm-variable subset selection) (*41*) was applied to the input set of 474 descriptors to select those most relevant to obtain models with the highest predictive power in modeling the global half-life index. All the calculations were performed maximizing the cross-validated Q^2 leave-one-out (LOO) (i.e. leaving out one molecule at a time from the training set and predicting it). Moreover, the bootstrap approach (*42*), repeated 5000 times for each validated model, was applied to avoid overestimation of model predictive power and to verify its robustness and internal prediction more thoroughly on a wider set of chemicals, which were randomly and iteratively put out of the training set and predicted. The model obtained on the first selected chemicals is used to predict the values for the excluded compounds, and then Q^2 is calculated for each model (Q^2_{BOOT}). The models were checked for reliability by Y scrambling to verify that the proposed models are not obtained by chance correlation (*34*).

Chemical Applicability Domain. Prediction reliability was checked by diagnostic procedures for applicability domain. Each model is based on a limited training set of compounds of known experimental half-life (literature data); therefore, the model cannot be expected to be applicable to every chemical. Thus, a quantitative measure of the model applicability domain is needed to evaluate the degree of extrapolation. The presence of outliers (i.e., compounds with residuals greater than 2.5 standard deviation units) and chemicals very influential in determining model parameters was verified by plotting standardized residuals versus hat values (Williams plot, Supporting Information Figure SI2). To check influential chemical leverage, hat values were used (*43*): they represent the compound distance from the model experimental space.

External Validation. To develop, and propose, predictive QSPR models for chemicals not participating in model development, the available set of chemicals was preliminarily

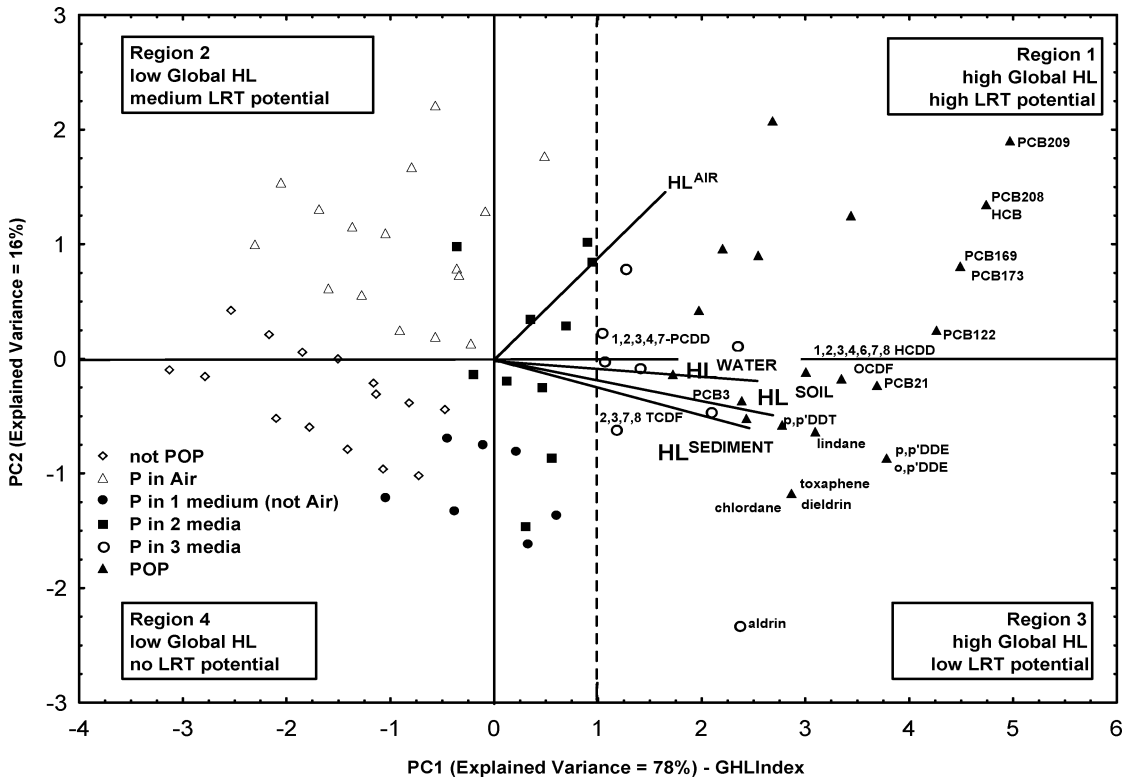

FIGURE 1. Principal component analysis on half-life data for 250 organic compounds in the various compartments (air, water, sediment, and soil) (PC1-PC2: explained variance = 94%). P = persistent.

split into a training set of 125 chemicals, on which the models were developed, and an external prediction set of 125 compounds, with which the models were verified only after their development. The splitting was carried out by random selection through activity sampling: the whole range of property was sorted through ascending order, and every second compound was assigned to the training set.

The coefficient of determination calculated for the prediction set (R^2) is reported, as well as the root mean squared error (RMSE) calculated for the training set on fitted values, RMSEcv calculated for the training set on LOO cross-validated values, and RMSEp calculated for the prediction set.

Results and Discussion

Chemical Ranking According to Cumulative Half-Life and Definition of a Global Half-Life Index (GHLI). As stated by Klasmeier et al. (*18*), the UNEP single-media half-life criteria and multimedia models could give qualitatively different classifications: different models delivering diverse and perhaps complementary information. Our study proposes a complementary modeling approach for screening persistence, applicable also when partitioning and environmental properties are not available. With the aim of obtaining a cumulative ranking index of global half-life, our first step was to employ PCA to combine all the available information on the half-life of the chemicals in air, water, soil, and sediment. Our combination by PCA of the available half-life data has been verified to give an effective ranking of chemicals according to their persistence, thus deriving the new index GHLI.

PCA on the available half-life data for 250 organic compounds in various compartments (air, water, sediment, and soil) was performed as an explorative multivariate method, taking into account the simultaneous change of all the half-lives. This PCA generates a distribution of the studied compounds according to their cumulative, or global, half-life and relative persistence in different media, without any preliminary bias of chemical partitioning and environmental conditions (temperature, wind, etc.).

The biplot relative to the first and second components is reported in Figure 1, where the chemicals (points or scores) are distributed according to their environmental persistence, represented by the linear combination of their half-lives in the four selected media (loadings shown as diagonal lines). The cumulative explained variance of the first two PCs is 94%, and the PC1 alone provides the largest part, 78%, of the total information. The loading lines show the importance of each variable in the first two PCs.

It is interesting to note that all the half-life values (loading lines) are oriented in the same direction along the first PC, being highly correlated to it (PC1-HL^{AIR} correlation = 0.70; PC1-HL^{WATER} correlation = 0.91; PC1-HL^{SOIL} correlation = 0.96; and PC1-$HL^{SEDIMENT}$ correlation = 0.94). Thus, the PC1, derived from a linear combination of half-life in different media, is a new macro-variable condensing chemical tendency to environmental persistence. PC1 ranks the compounds according to their cumulative half-life and discriminates between them with regard to persistence (different label in Figure 1): chemicals with high half-life values in all the media are located to the right of the plot, 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 or persistent in only one medium; and chemicals persistent in two or three media are located in the intermediate zone of Figure 1.

PC2, although less informative (explained variance 16%), is also interesting: it separates the compounds more persistent in air (upper parts in Figure 1, regions 1 and 2), thus those with higher LRT potential, from those more persistent, in water, soil, and sediment (lower parts in Figure 1, regions 3 and 4, where chemicals with low or no LRT potential in air are located).

This plot, based only on half-life data, can be compared with a classification plot proposed and commented on by

Klasmeier et al. (*18*), identifying four regions (A−D) for overall persistence (P_{ov}) and LRTP obtained from multimedia models. The same chemical classification was found in our plot (Figure 1), identifying four analogous regions (1−4) which boundaries were defined by the PCA axes. Region 1 (top right corner), containing the more persistent compounds in all the media and mainly in air, is the zone of POPs with the higher LRT potential (most dangerous chemicals); region 2 (top left corner) includes lower persistence compounds with a relatively high half-life in air, thus with medium LRT potential; region 3 (bottom right corner) holds chemicals persistent mainly in water, soil, and sediment, thus without a tendency to LRT in the atmosphere; and region 4 (bottom left corner) shows the chemicals less persistent in each medium and also in air (less dangerous chemicals with no LRT tendency).

A deeper analysis of the distribution of the studied chemicals in Figure 1 reveals some interesting results and gives confirmation of the experimental evidence: to the right, among the very persistent chemicals in all the compartments (labeled in Figure 1), we find most of the compounds recognized as POPs by the Stockholm Convention (*1*, *2*).

Highly chlorinated PCBs (PCB-122, PCB-169, PCB-173, PCB-208, and PCB-209) and hexachlorobenzene are among the most persistent compounds in our reference scenario. All these compounds are grouped in region 1 owing to their global high persistence, especially in air (decachlorobiphenyl PCB-209, in particular, is the most persistent in the air medium, in the studied dataset). The less chlorinated PCBs (PCB-3 and PCB 21) fall in region 3 near the horizontal boundaries in the zone of very persistent chemicals, due to their lower persistence in air as compared with highly chlorinated congeners. p,p′-DDT, p,p′-DDE, and o,p′-DDE, highly chlorinated dioxins and dioxin-like compounds, as well as some pesticides (toxaphene, lindane, chlordane, dieldrin, and aldrin), fall in region 3 of the highly persistent chemicals mainly in compartments different from air.

To confirm the reliability of our approach, we verified, in our PCA plot, the position of 10 reference chemicals (listed in Supporting Information Table SI1 and highlighted in Figure SI1) and three hypothetical chemicals (HypoA, -B, -C in Supporting Information Table SI3) used by Klasmeier et al. (*18*) to illustrate the multimedia-based classification as compared with UNEP criteria.

In our studied scenario, CCl_4 is located in region 1 (high persistence, mainly in air), and this led us to conclude, in agreement with Klasmeier et al., that it must be included in chemicals harmful for persistence and LRT, although still not identified as a POP by UNEP. HCB and a-HCH, both in region 1, are confirmed as persistent chemicals and with a high LRT potential. PCBs 180, 101, and 28 fall, in order of decreasing persistence, near the boundary between regions 1−3, and thus, they are very persistent but do not exhibit strong LRT potential.

Aldrin is clearly a highly persistent compound; its position at the bottom of region 3 reveals its resistance to degradation in water, soil, and sediment, but it has high degradability in air and thus a low potential for transport in the atmosphere. Atrazine in region 3, close to the boundary with region 4, can be classified as an intermediate persistent compound. Note also its clear position below the LRT boundary, where chemicals more persistent in compartments other than air lie. In fact, atrazine is a recognized problem of groundwater pollution. Biphenyl and mainly *p*-cresol are in region 2 of non-POP compounds falling close to the LRT boundary (medium LRT potential).

Hypo-A, which exceeds UNEP criteria for both persistence and LRT but is classified non-POP and non-LRT by multimedia, falls in the central area of our PCA plot, close to the boundaries. Our intermediate result supports the difficulty in classifying this chemical (opposite classification by UNEP and multimedia approaches). Hypo-B has a strong LRT potential in both UNEP and multimedia approaches, but the two approaches give opposite classification for persistence. This compound, falling in region 1 of the plot, is relatively persistent but has a high resistance to degradation in the atmosphere and LRT potential. Hypo-C falls in region 3, being confirmed as a highly persistent chemical in all the compartments but air and with no LTR potential in air. A detailed comparison of the classification obtained by UNEP, Klasmeier et al. (*18*), and our approach for the three hypothetical chemicals is reported in Supporting Information Table SI3.

Our PCA results are in good agreement with the classification based on eight different multimedia models and with the empirical evidence (*18*). This confirms the reliability of our proposed approach to correctly rank chemicals for persistence. The method can be viewed as complementary to UNEP criteria and multimedia models, always bearing in mind that an arbitrary definition of the boundaries is included in all these approaches.

A threshold value to highlight very persistent chemicals (vP) can be set in Figure 1 as PC1 score ≥ 1, which identifies chemicals persistent in three or four media at the same time. Supporting Information Table SI4 lists the 69 vP compounds in our dataset. The compounds ranking along the PC1 axis (PC1 score) according to global persistence in the environment, verified as stated previously for reliability, can be defined as a new combined end-point for persistence: the GHLI.

QSPR Modeling of the GHLI. This GHLI, obtained from existing knowledge of generalized chemical persistence over a wide scenario of 250 chemicals, was modeled as a condensed end-point in a QSPR approach.

To propose really predictive QSPR models, also for chemicals not participating in model calibration, we provided an external validation of the model (*34*): for this purpose, the original set of available data was first randomly split into training and prediction sets; 50% of the compounds was put into the prediction set (125 compounds), while the other 50% was used to build the QSPR model. Thus, the model was developed on a limited number of representative chemicals, and the compounds in the prediction set are actually new for the model and used only to verify its predictive power. A population of MLR models (OLS method) was developed by selecting the few descriptors that are the most relevant in their combination for modeling the response, by applying the genetic algorithm procedure on 474 structural theoretical molecular descriptors and using them as input variables. Given next is the best QSPR model, selected by statistical approaches, and its statistical parameters (Figure 2 shows the plot of GHLI values obtained from PCA vs QSPR predicted GHLI values):

$$\begin{aligned}\mathrm{GHLI} = {} & -3.12(\pm 0.77) + 0.33(\pm 4.5 \times 10^{-2})\mathrm{X0v} + \\ & 5.1(\pm 0.99)\mathrm{Mv} - 0.32(\pm 6.13 \times 10^{-2})\mathrm{MAXDP} \\ & -0.61(\pm 0.10)\mathrm{nHDon} - 0.5(\pm 1.15)\mathrm{CIC0} - \\ & 0.61(\pm 0.13)\mathrm{O\text{-}060}\end{aligned}$$

In this equation, $n_{training} = 125$, $R^2 = 0.85$, $Q^2_{LOO} = 0.83$, $Q^2_{BOOT} = 0.83$, RMSE = 0.76, and RMSE cv = 0.70; $n_{prediction} = 125$, $R^2_{EXT} = 0.79$, and RMSEp = 0.78.

This model presents good internal and external predictive power, a result that must be highlighted also as proof of the model robustness.

Indeed, the model shows comparable internal and external predictive ability ($R^2_{EXT} = 0.79$ in comparison with $R^2 = 0.85$) as also demonstrated by the RMSEp value, just slightly higher than the RMSE. An analysis of the model applicability domain did not reveal any outliers in the structural domain, while some slight outliers for the response were identified in the

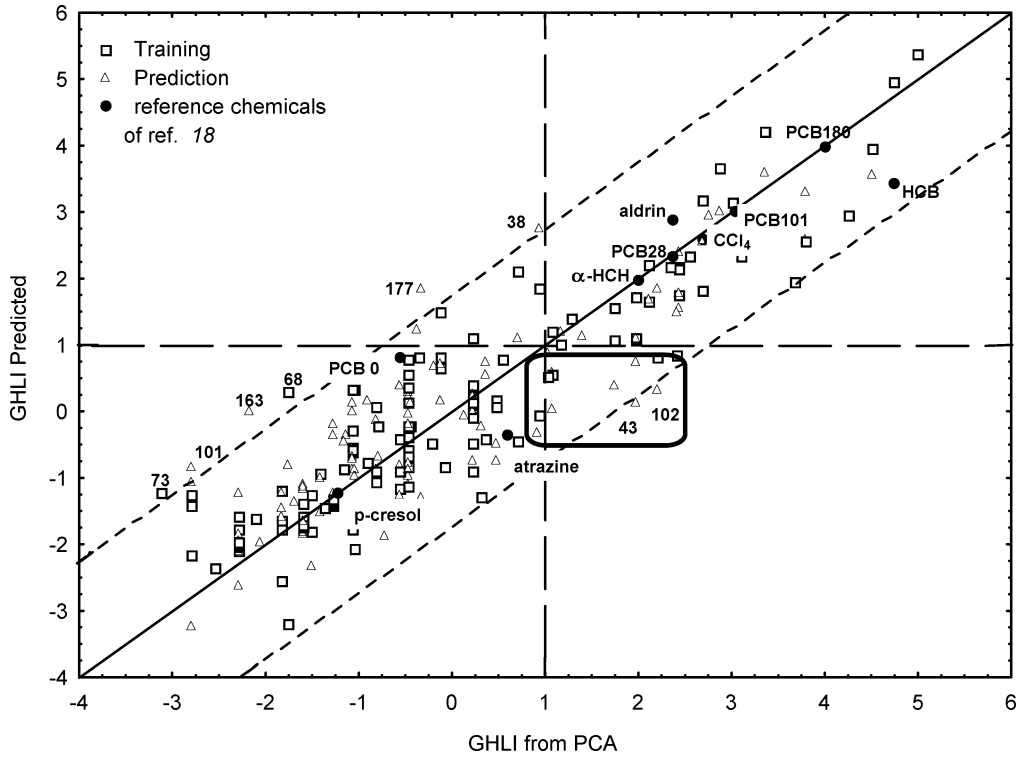

FIGURE 2. Scatter plot of the GHLI values calculated by PCA vs predicted values by the model. GHLI values for the training and prediction set chemicals are labeled differently. Diagonal dotted lines indicate that the 2.5σ interval and response outliers are numbered as in the text and in Tables SI1 and SI5. Vertical and horizontal dotted lines identify the cutoff value of GHLI = 1.

training (two chemicals: *o*-cresol (73) and 1,1′-biphenyl-4,4′-diamine (68)) and in the prediction set (six chemicals: 1,2-dichloropropane (43), 1,2-dichloroethane (102), tetramethyl thioperoxydicarbonic diamide (177), malathion (163), 2-propenal (101), and bromodichloromethane (38)) (Williams plot of model domain in Supporting Information Figure SI2).

An analysis of overestimated and underestimated chemicals (model uncertainty) provides additional information on the outcome of the proposed model for the studied chemicals. In fact, even if the statistical quality of the model is high, the chemicals lying above the line of Figure 2 are predicted as more persistent than the calculated GHLI. However, for the precautionary approach, this overestimation can be considered less problematic than underestimation. In fact, the Stockholm Convention (*2*) for POP identification clearly states that the required standard of evidence will be based on the need for precaution in considering additional candidates for the POPs list.

Among the underestimated chemicals, the only really dangerous zone in the proposed model is circled in Figure 2. Here, a few very persistent chemicals (cutoff value of GHLI = 1), mainly small and chlorinated aliphatic hydrocarbons and aromatics with a few chlorine substituents, were predicted as medium persistent. Our preliminary findings suggest the need for a deeper analysis of these chemicals.

Finally, it should be noted that the application of this QSPR model to the 10 reference chemicals of ref *18* (named in Figure 2) gave reliable GHLI predictions, in satisfactory agreement with the position expected for these chemicals along the *x*-axis in the PCA of Figure 1 and Supporting Information Figure SI1.

In this model, the descriptors selected by the genetic algorithm procedure are, in descending order of importance in the modeling: X0v (Randic connectivity index of 0 order), Mv (mean atomic van der Waals volumes), MAXDP (maximal electrotopological positive variation), nHDon (number of donor atoms for H bonds), CIC0 (information index related to the complementary information content), and O-060 (fragment related to the presence of oxygen atoms in ethers and esters). All are bi-dimensional parameters independent of chemical conformation and easily calculable starting only from the topological graph. The calculations can be performed by the appropriate software (*35*) even starting from the SMILES string for each chemical. These descriptors can be mechanistically interpreted: the variables take account of the different structural properties involved in defining environmental persistence tendency, such as chemical size (X0v and Mv, as more complex chemicals are generally expected to be more persistent than simpler) and electronic features (CIC0, MAXDP, nHDon, and O-060, more related to a compound's ability to form electrostatic and dipole−dipole interactions in the surrounding media). These features can directly influence the bioavailability and partitioning of chemicals into different environmental compartments and can indirectly determine their availability for different degradation pathways. It is also interesting to note that both the descriptor MAXDP, related to molecule electrophilicity, and the topological CIC0 had already demonstrated the ability to model important environmental partition properties such as the soil sorption partition parameter Koc (*44*, *45*) and tropospheric degradability by OH radicals (*29*). There is an evident relation between sorption and persistence as the more sorbed chemicals are the most recalcitrant to biotic and abiotic degradation, thus rendering them more persistent.

It is not possible to make a too precise comparison with other pollutant persistence studies from the literature as the models are based on different approaches with varying levels of complexity, and their purposes and levels of applicability are very different. Furthermore, any screening approach is also clearly dependent on the studied chemicals and the data used as input. However, the results of our approach appear in satisfactory agreement with those of the literature (*3*−*20*, *25*, *26*) and the U.S. EPA PBT-Profiler (*46*). Moreover, the reliability of our approach was validated by comparing

our results, obtained on a wide set of 250 chemicals, with those achieved by Klasmeier et al. (*18*) using multimedia models on 10 reference chemicals.

Obviously, our developed model could be improved by the input of more accurate experimental half-life data or more and different compounds to enlarge the chemical domain. Nevertheless, it must be borne in mind that the model was developed starting from revised data (the same kind of data commonly used to run and develop the widely applied multimedia models) of a heterogeneous set of chemicals (wide reference scenario), and the external prediction ability of the model was rigorously evaluated.

Obviously, alternative multivariate methods for the analysis and the modeling (i.e., the partial least-squares (PLS) approach), could be applied to this topic to compare the results. However, in this paper, we decided to use a simple and portable modeling approach such as the combination of PCA and MLR.

This model is easy to apply, also by non-specialists, for a fast screening of very persistent chemicals. The equation proposed in the text can be applied after calculating the six modeling descriptors (by DRAGON (*35*) also freely available online at http://www.vcclab.org/lab/edragon) to obtain the GHLI value for any new chemical. Chemicals with a GHLI value >1 can be viewed as persistent: the higher the GHLI value, the higher the risk of persistence. Chemicals detected as very persistent by this model, using only the molecular structure information included in the equation, should be banned or not synthesized; whereas, for a more detailed assessment, an additional check is suggested for those chemicals predicted as less persistent and less dangerous. The method, applicable before any other modeling approach, is preliminary and complementary to multimedia models and is much simpler, being independent of other experimental data and/or environmental assumptions.

It is important to remember that the primary goal of any screening assessment based on QSAR is to avoid past mistakes by using previously obtained information (the already existing knowledge of half-life) as an invaluable input to be exploited to highlight the most highly prioritized compounds or to suggest the synthesis of harmless chemicals.

Thus, the proposed multivariate approach is particularly useful not only to screen and to make an early prioritisation 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. To date, no other QSPR (Quantitative Structure-Persistence Relationships) models have been developed, those here proposed could be also applied usefully for PBT screening and in the new European Legislation (REACH − Registration, Evaluation and Authorization of Chemicals) (*47*).

Acknowledgments

We thank the University of Insubria for a Ph.D. fellowship for E.P. Anonymous reviewers are gratefully acknowledged for their helpful suggestions that improved the paper. We are also thankful to Prof. Don Mackay for his encouragement in pursuing the application of our multivariate QSAR-based approach to persistence.

Supporting Information Available

PCA of half-life values for our data set including chemicals from ref *18* (Figure SI1); applicability domain (Williams Plot) of the proposed QSPR model (Figure SI2); studied data set (Table SI1); typology of theoretical molecular descriptors used as input in the study (Table SI2); comparison of classification for persistence by UNEP criteria, multimedia approach, and GHLI approach of the hypothetical compounds of Klasmeier et al. (Table SI3); list of the very persistent chemicals (Table SI4); list of the molecular descriptors and calculated and predicted GHLI values for the studied chemicals (Table SI3). This material is available free of charge via the Internet at http://pubs.acs.org.

Literature Cited

(1) UNEP. *Report of the First Session of the INC for an International Legally Binding Instrument for Implementing International Action on Certain Persistent Organic Pollutants (POPs)*; International Institute for Sustainable Development: Winnipeg, Canada, 1998; p 10 or http://www.pops.int/documents/meetings/inc1/RITTER-En.html (accessed December 13, 2006).

(2) UNEP. *Stockholm Convention on Persistent Organic Pollutants*; United Nations Environment Program: Geneva, Switzerland, 2001; http://www.pops.int (accessed December 13, 2006).

(3) Muller-Herold, U.; Caderas, D; Funck, P. Validity of Global Lifetime Estimates by a Simple General Limiting Law for the Decay of Organic Compounds with a Long-Range Pollution Potential. *Environ. Sci. Technol.* **1997**, *31*, 3511−3515.

(4) Scheringer, M. Characterization of the Environmental Distribution Behavior of Organic Chemicals by Means of Persistence and Spatial Range. *Environ. Sci. Technol.* **1997**, *31*, 2891−2897.

(5) Webster, E.; Mackay, D.; Wania, F. Evaluating Environmental Persistence. *Environ. Toxicol. Chem.* **1998**, *17*, 2148−2158.

(6) Klecka, G.; Boethling, B.; Franklin, J.; Grady, L.; Graham, D.; Howard, P. H.; Kannan, K.; Larson, R. L.; Mackay, D.; Muir, D.; van de Meent, D. M. *Evaluation of Persistence and Long-Range Transport of Organic Chemicals in the Environment*; SETAC Press: Pensacola, FL, 2000.

(7) Rodan, B. D.; Pennington, D. W.; Eckley, N.; Boethling, R. S. Screening of Persistent Organic Pollutants: Techniques to Provide a Scientific Basis for POPs Criteria in International Negotiations. *Environ. Sci. Technol.* **1999**, *33*, 3482−3488.

(8) Beyer, A.; Mackay, D.; Matthies, M.; Wania, F.; Webster, E. Assessing Long-Range Transport Potential of Persistent Organic Pollutants. *Environ. Sci. Technol.* **2000**, *34*, 699−703.

(9) Gouin, T.; Mackay, D.; Webster, E.; Wania, F. Screening Chemicals for Persistence in the Environment. *Environ. Sci. Technol.* **2000**, *34*, 881−884.

(10) Pennington, D. W. An Evaluation of Chemical Persistence Screening Approaches. *Chemosphere* **2001**, *44*, 1589−1601.

(11) Mackay, D.; McCarty, L. S.; MacLeod, M. On the Validity of Classifying Chemicals for Persistence, Bioaccumulation, Toxicity, and Potential for Long-Range Transport. *Environ. Toxicol. Chem.* **2001**, *20*, 1491−1498.

(12) Wania, F.; Daly, G. L. Estimating the Contribution of Degradation in Air and Deposition to the Deep Sea of the Global Loss of PCBs. *Atmos. Environ.* **2002**, *36*, 5581−5593.

(13) Gouin, T.; Cousin, I.; Mackay, D. Comparison of Two Methods for Obtaining Degradation Half-Lives. *Chemosphere* **2004**, *56*, 531−535.

(14) Jones, K. C.; Sweetman, A.; Mackay, D. Editorial, Special Issue. *Environ. Pollut.* **2004**, *128*, 1−2.

(15) Stroebe, M.; Scheringer, M.; Hungerbuhler, K. Measures of Overall Persistence and the Temporal Remote State. *Environ. Sci. Technol.* **2004**, *38*, 5665−5673.

(16) Arnot, J. A.; Mackay, D.; Webster, E.; Southwood, J. M. Screening Level Risk Assessment Model for Chemical Fate and Effects in the Environment. *Environ. Sci. Technol.* **2006**, *40*, 2316−2323.

(17) Aronson, D.; Boethling, R.; Howard, P.; Stiteler, W. Estimating Biodegradation Half-Lives for Use in Chemical Screening. *Chemosphere* **2006**, *63*, 1953−1960.

(18) Klasmeier, J.; Matthies, M.; MacLeod, M.; Fenner, K.; Scheringer, M.; Stroebe, M.; Le Gall, A. C.; McKone, T.; Van De Meent, D.; Wania, F. Application of Multimedia Models for Screening Assessment of Long-Range Transport Potential and Overall Persistence. *Environ. Sci. Technol.* **2006**, *40*, 53−60.

(19) Bennet, H. D.; Kastenberg, W. E.; McKone, T. E. General Formulation on Characteristic Time for Persistent Chemicals in a Multimedia Environment. *Environ. Sci. Technol.* **1999**, *33*, 503−509.

(20) Fenner, K; Scheringer, M.; Macleod, M.; Matthies, M.; McKone, T.; Stroebe, M.; Beyer, A.; Bonnell, M.; Le Gall, A. C.; Klasmeier, J.; Mackay, D.; Van de Meent, D.; Pennington, D.; Scharenberg, B.; Suzuki, N.; Wania, F. Comparing Estimates of Persistence and Long-Range Transport Potential among Multimedia Models. *Environ. Sci. Technol.* **2005**, *39*, 1932−1942.

(21) Oberg, T. Virtual Screening for Environmental Pollutants: Structure−Activity Relationships Applied to a Database of Industrial Chemicals. *Environ. Toxicol. Chem*. **2006**, *25*, 1178−1183.
(22) OECD. *Guidance Document on the Use of Multimedia Models for Estimating Overall Environmental Persistence and Long-Range Transport; No. 45 in Series of Testing and Assessment*; OECD Environment, Health, and Safety Publications: Paris, France, 2004.
(23) Howard, P. H.; Boethling, R. S.; Jarvis, W. F.; Meylan, W. M.; Michalenko, E. M. *Handbook of Environmental Degradation Rates*; Lewis: Chelsea, MI, 1991.
(24) Mackay, D.; Shiu, W. Y.; Ma, K. C. *Physical−Chemical Properties and Environmental Fate Handbook*, CRCnet-BASE CD-ROM; Chapman and Hall/CRC: Boca Raton, FL, 2000.
(25) Gramatica, P.; Consolaro, F.; Pozzi, S. QSAR Approach to POPs Screening for Atmospheric Persistence. *Chemosphere* **2001**, *43*, 655−664.
(26) Gramatica, P.; Pozzi, S.; Consonni, V.; Di Guardo, A. Classification of Environmental Pollutants for Global Mobility Potential. *SAR QSAR Environ. Res.* **2002**, *13*, 205−217.
(27) Gramatica, P.; Papa, E.; Pozzi, S. Prediction of POP Environmental Persistence and Long-Range Transport by QSAR and Chemometric Approaches. *Fresenius Environ. Bull.* **2004**, *13*, 1204−1209.
(28) Gramatica, P.; Consonni, V.; Todeschini, R. QSAR Study of the Tropospheric Degradation of Organic Compounds. *Chemosphere* **1999**, *38*, 1371−1378.
(29) Gramatica, P.; Pilutti, P.; Papa, E. Validated QSAR Prediction of OH Tropospheric Degradability: Splitting into Training−Prediction Set and Consensus Modeling. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1794−1802.
(30) Gramatica, P.; Pilutti, P.; Papa, E. Predicting the NO_3 Tropospheric Degradability of Organic Pollutants by Theoretical Molecular Descriptors *Atmos. Environ.* **2003**, *37*, 3115−3124.
(31) Gramatica, P.; Pilutti, P.; Papa, E. QSAR Prediction of Ozone Tropospheric Degradation. *QSAR Comb. Sci.* **2003**, *22*, 364−373.
(32) Gramatica, P.; Pilutti, P.; Papa, E. Ranking of Volatile Organic Compounds for Tropospheric Degradability by Oxidants: A QSPR Approach. *SAR QSAR Environ. Res.* **2002**, *13*, 743−753.
(33) Gramatica, P.; Pilutti, P.; Papa, E. A Tool for the Assessment of VOC Degradability by Tropospheric Oxidants Starting from Chemical Structure. *Atmos. Environ.* **2004**, *38*, 6167−6175.
(34) Tropsha, A.; Gramatica, P.; Gombar, V. J. K. The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.
(35) DRAGON for Windows (Software for Molecular Descriptor Calculations), Version 5.4; Talete s.r.l.: Milan, 2006; http://www.talete.mi.it (accessed July 26th, 2006).
(36) *HyperChem*, version 7.03 for Windows; Autodesk, Inc.: Sausalito, CA, 2002.
(37) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley: New York, 2000; p 667.
(38) Jackson, J. E. *A User's Guide to Principal Components*; Wiley: New York, 1991.
(39) *SCAN−Software for Chemometric Analysis*, Version 1.1 for Windows; Minitab: State College, PA, 1995.
(40) *MOBY DIGS Professional−Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm*, Version 1.0 beta for Windows; Talete srl: Milan, 2004.
(41) Leardi, R.; Boggia, R.; Terrile, M. Genetic Algorithms as a Strategy for Feature Selection. *J. Chemom.* **1992**, *6*, 267−281.
(42) Wehrens, R.; Putter, H.; Buydens, L. M. C. The Bootstrap: A Tutorial. *Chemom. Intell. Lab. Syst.* **2002**, *54*, 35−52.
(43) Atkinson, A. C. *Plots, Transformations, and Regression*; Clarendon Press: Oxford, 1985.
(44) Gramatica, P.; Corradi, M.; Consonni, V. Modeling and prediction of soil sorption coefficients of non-ionic organic pesticides by different sets of molecular descriptors. *Chemosphere* **2000**, *41*, 763−777.
(45) Gramatica, P.; Giani, E.; Papa, E. Statistical External Validation and Consensus Modeling: A QSPR Case Study for Koc Prediction. *J. Mol. Graph. Model.* **2007**, *25*, 755−766.
(46) U.S. EPA PBT-Profiler; U.S. EPA: Washington, DC; www.pbt-profiler.net (accessed December 13, 2006).
(47) http://europa.eu.int/comm/environment/chemicals/reach.htm (accessed December 13, 2006).

Received for review July 26, 2006. Revised manuscript received February 6, 2007. Accepted February 6, 2007.

ES061773B