CHAPTER 12

CHEMOMETRIC METHODS AND THEORETICAL MOLECULAR DESCRIPTORS IN PREDICTIVE QSAR MODELING OF THE ENVIRONMENTAL BEHAVIOR OF ORGANIC POLLUTANTS

PAOLA GRAMATICA

QSAR Research Unit in Environmental Chemistry and Ecotoxicology,
Department of Structural and Functional Biology, University of Insubria, Varese, Italy,
e-mail: paola.gramatica@uninsubria.it; http://www.qsar.it

Abstract: This chapter surveys the QSAR modeling approaches (developed by the author's research group) for the validated prediction of environmental properties of organic pollutants. Various chemometric methods, based on different theoretical molecular descriptors, have been applied: explorative techniques (such as PCA for ranking, SOM for similarity analysis), modeling approaches by multiple-linear regression (MLR, in particular OLS), and classification methods (mainly k-NN, CART, CP-ANN). The focus of this review is on the main topics of environmental chemistry and ecotoxicology, related to the physico-chemical properties, the reactivity, and biological activity of chemicals of high environmental concern. Thus, the review deals with atmospheric degradation reactions of VOCs by tropospheric oxidants, persistence and long-range transport of POPs, sorption behavior of pesticides (K_{oc} and leaching), bioconcentration, toxicity (acute aquatic toxicity, mutagenicity of PAHs, estrogen binding activity for endocrine disruptors compounds (EDCs)), and finally persistent bioaccumulative and toxic (PBT) behavior for the screening and prioritization of organic pollutants. Common to all the proposed models is the attention paid to model validation for predictive ability (not only internal, but also external for chemicals not participating in the model development) and checking of the chemical domain of applicability. Adherence to such a policy, requested also by the OECD principles, ensures the production of reliable predicted data, useful also in the new European regulation of chemicals, REACH.

Keywords: QSAR, Chemometric methods, Theoretical molecular descriptors, MLR, Classification, Environmental pollutants, Ranking

12.1. INTRODUCTION

The QSAR world has undergone profound changes since the pioneering work of Corwin Hansch, considered the founder of modern QSAR modeling [1, 2]. The main change is reflected in the growth of a parallel and quite different conceptual

T. Puzyn et al. (eds.), Recent Advances in QSAR Studies, 327–366.
DOI 10.1007/978-1-4020-9783-6_12,

approach to the modeling of the relationships among a chemical's structure and its activity/properties.

In the Hansch approach, still applied widely and followed by many QSAR modelers (for instance, [3–5]), molecular structure is represented by only a few molecular descriptors (typically log K_{ow},[1] Hammett constants, HOMO/LUMO, some steric parameters) selected personally by the modeler and inserted in the QSAR equation to model a studied endpoint. Alternatively, in a different approach chemical structure is represented, in the first preliminary step, by a large number of theoretical molecular descriptors which are then, in a second step, selected by different chemometric methods as the best correlated with the response and, finally, included in the QSAR model (the algorithm), the fundamental aim being the optimization of model performance for prediction.

According to the Hansch approach, descriptor selection is guided by the modeler's conviction to have *a priori* knowledge of the mechanism of the studied activity/property. The modeler's presumption is to assign mechanistic meaning to any used molecular descriptor selected by the modeler from among a limited pool of potential modeling variables. These descriptors are normally well known and used repeatedly (for instance, log K_{ow} is a universal parameter mimicking cell membrane permeation, thus it is used in models for toxicity, but it is also related to various partition coefficients such as bioconcentration/bioaccumulation, soil sorption coefficient; HOMO/LUMO energies are often selected for modeling chemical reactivity, etc.).

On the other hand, the "statistical" approach, an approach parallel to the previous so-called "mechanistic" one, is based on the fundamental conviction that the QSAR modeler should not influence, *a priori* and personally, the descriptor selection through mechanistic assumptions. Instead they should apply unbiased mathematical tools to select, from a wide pool of input descriptors, those descriptors most correlated to the studied response. The number and typology of the available input descriptors must be as wide and different as possible in order to guarantee the possibility of representing any aspect of the molecular structure. Different descriptors are different ways or perspectives to view a molecule. Descriptor selection should be performed by applying mathematical approaches to maximize, as an optimization parameter, the predictive power of the QSAR model, as the real utility of any model considered is its predictivity.

The first aim of any modeler should be the validation for predictive purposes of the QSAR model, for both the mechanistic and statistical approaches; in fact, a QSAR model must, first of all, be a real model, robust and predictive, to be considered a reliable model; only a stable and predictive model can be usefully interpreted for its mechanistic meaning, even so this is not always easy or feasible [6]. However, this is a second step in the statistical QSAR modeling.

---

[1] The symbol refers to the same property as log P (namely to the n-octanol/water partition coefficient). However, in many environmental studies this partition coefficient is abbreviated by "log K_{ow}" to be consistent with the other environmentally relevant coefficients, e.g., n-octanol/air partition coefficient (K_{oa}), air/water partition coefficient (K_{aw}).

QSAR model validation has been recognized by specific OECD expert groups as a crucial and urgent requirement in recent years, and this has led to the development, for regulatory purposes, of the "OECD principles for the validation of (Q)SAR models" (http://www.oecd.org/document/23/0,3343,fr_2649_34365_33957015_1_1_1_1,00.html).

The need for this important action was mainly due to the recent new chemicals policy of the European Commission (REACH: *R*egistration, *E*valuation, *A*uthorization and restriction of *Ch*emicals) (http://europa.eu.int/comm/environment/chemicals/reach.htm) that explicitly states the need to use (Q)SAR models to reduce experimental testing (including animal testing). Obviously, to meet the requirements of the REACH legislation (see also Chapter 13) it is essential to use (Q)SAR models that produce reliable estimates, i.e., validated (Q)SAR models. Thus, reliable QSAR model must be associated with the following information: (1) a defined endpoint; (2) an unambiguous algorithm; (3) a defined domain of applicability; (4) appropriate measures of goodness-of-fit, robustness and predictivity; (5) a mechanistic interpretation, if possible.

Some crucial points of the statistical approach of QSAR modeling, applied by the author's group, are put into context, according to the guidelines of the OECD principles, which are the chemometric approach steps.

12.2. A DEFINED ENDPOINT (OECD PRINCIPLE 1)

The most common regulatory endpoints, associated with OECD test guidelines, are related to (a) physico-chemical properties (such as melting and boiling points, vapor pressure, K_{ow}, K_{oc}, water solubility); (b) environmental fate (such as biodegradation, hydrolysis, atmospheric oxidation, bioaccumulation); (c) human health (acute oral, acute inhalation, acute dermal, skin irritation, eye irritation, skin sensitization, genotoxicity, reproductive and developmental toxicity, carcinogenicity, specific organ toxicity (e.g., hepatotoxicity, cardiotoxicity)); and (d) ecological effects (acute fish, acute daphnid, alga, long-term aquatic, and terrestrial toxicity) of chemicals.

The various experimental endpoints that have been modelled by the QSAR Research Unit of Insubria University are described in the following sections, after the discussion on the main methodological topics. A distinction will be made between single endpoints and cumulative endpoints, which take into account a contemporaneous contribution of different properties or activities.

12.3. AN UNAMBIGUOUS ALGORITHM (OECD PRINCIPLE 2)

The algorithms used in (Q)SAR modeling should be described thoroughly so that the user will understand exactly how the estimated value was produced and can reproduce exactly the calculations also for new chemicals, if desired.

When the studied endpoint needs to be modelled using more than one descriptor (selected by different approaches) multivariate techniques are applied. As there can be multiple steps in estimating the endpoint of a chemical, it is important that the nature of the used algorithms be unambiguous, as required by OECD Principle 2.

12.3.1. Chemometric Methods

12.3.1.1. Regression Models

Regression analysis is the use of statistical methods for modeling a dependent variable Y, a quantitative measures of response (e.g., boiling point, LD_{50}), in terms of predictors X (independent variables or molecular descriptors).

There are many different multivariate methods for regression analysis, more or less widely applied in QSAR studies: multiple linear regression (MLR), principal component regression (PCR), partial least squares (PLS), artificial neural networks (ANNs), fuzzy clustering and regression are among more commonly used approaches for regression modeling.

Although all QSAR models (linear and not linear) are based on algorithms, the most common regression method, which describes models by completely transparent and easily reproducible mathematical equations, is multiple linear regression (MLR), in particular ordinary least squares (OLS) method. This method has been applied by the author in her QSAR studies; to cite some most recent papers, see [7–28] and Chapter 6. Some of these models are commented on in the following paragraphs.

The correlation of the variables in the modeling must be controlled carefully (for instance, by applying the QUIK rule [29]) and the problem of possible overfitting [30], common also to other modeling methods, must also be checked by statistical validation methods to verify robustness and predictivity. The selection of descriptors in MLR can be performed either *a priori* by the model developer on a mechanistic basis or by evolutionary techniques such as genetic algorithms. In this second approach, the model's developer should try to interpret mechanistically the descriptors selected, but only after model development and statistical validation for predictivity.

12.3.1.2. Classification Models

Another common problem in QSAR analysis is prediction of the group membership from molecular descriptors. In the simplest case, chemicals are categorized into one, two, or more groups depending on their activity, indicated by the same value of a categorical variable: active/inactive or, for instance, toxic/non-toxic.

Classification models are quantitative models based on relationships between independent variables X (in this case molecular descriptors) and a categorical response variable of integer numerical values, each representing the class of the corresponding sample.

The term "quantitative" is referred to the numerical value of the variables necessary to classify the chemicals in the qualitative classes (a categorical response) and it specifies the quantitative meaning of a QSAR-based classification process.

Such classification, also called supervised pattern recognition, is the assignment, on the basis of a classification rule, of chemicals to one of the classes defined *a priori* (or of groups of chemicals in the training set). Thus, the goal of a classification method is to develop a classification rule (by selecting the predictor variables) based

on a training set of chemicals of known classes so that the rule can be applied to a test set of compounds of unknown classes. A wide range of classification methods exists, including discriminant analysis (DA; linear quadratic, and regularized DA), soft independent modeling of class analogy (SIMCA), k-nearest neighbors (k-NN), classification and regression tree (CART), artificial neural network, support vector machine, etc.

The QSAR Research Unit of Insubria University has developed some satisfactory, validated, and usable classification models (for instance, among the more recent [16, 31–35]) by applying different classification methods, mainly classification and regression tree (CART) [36, 37], k-nearest neighbor (k-NN) [38], and artificial neural networks (in particular, Kohonen maps or self-organizing maps (SOM) [39–41]).

CART is a non-parametric unbiased classification strategy to classify chemicals with automatic stepwise variable selection. As the final output, CART displays a binary, immediately applicable, classification tree; each non-terminal node corresponds to a discriminant variable (with the threshold value of that molecular descriptor) and each terminal node corresponds to a single class. To classify a chemical, at each binary node, the tree branch, matching the values of the chemical on the corresponding splitting descriptor, must be followed.

The k-NN method is a non-parametric unbiased classification method that searches for the k-nearest neighbors of each chemical in a data set. The compound under study is classified by considering the majority of classes to which the k^{th} nearest chemicals belong. k-NN is applied to autoscaled data with *a priori* probability proportional to the size of the classes; the predictive power of the model is checked for k nearest neighbors between 1 and 10.

Counter-propagation artificial neural networks (CP-ANNs), particularly Kohonen maps, are supervised classification methods. Input variables (molecular descriptors) calculated for the studied chemicals provide the input for the net or the Kohonen layer. The architecture of the net is constituted by $N \times N \times p$, where p is the number of input variables and each p-dimensional vector is a neuron (N). Thus, the neurons are vectors of weights, corresponding to the input variables. During the learning, n chemicals are presented to the net – one at a time – a fixed number of times (epochs); each chemical is then assigned to the cell for which the distance between the chemical vector and the neuron is minimum. The target values (i.e., the classes to be modelled) are given to the output layer (the top-map: a two-dimensional plane of response), which has the same topological arrangement of neurons as the Kohonen layer. The position of the chemicals is projected to the output layer and the weights are corrected in such a way that they fit the output values (classes) of corresponding chemicals. The Kohonen-ANN automatically adapts itself in such a way that similar input objects are associated with topologically close neurons in the top-map. The chemical similarity decreases with increasing of the topological distance.

The trained network can be used for predictions; a new object in the Kohonen layer will lie on the neuron with the most similar weights. This position is then projected to the top-map, which provides a predicted output value. It is important

to remember that the Kohonen top-map has toroid geometry; each neuron has the same number of neighbors, including the neurons on the borders of the top-map.

According to the OECD principles, for a QSAR model to be acceptable for use to make regulatory decisions it must be clearly defined, easily and continuously applicable in such a way that the calculations for the prediction of the endpoint can be reproduced by everyone, and applicable to new chemicals. The unambiguous algorithm is characterized not only by the mathematical method of calculation used, but also by the specific molecular descriptors required in the model mathematical equation. Thus, the exact procedure used to calculate the descriptors, including compound pre-treatment (e.g., energy minimization, partial charge calculation), the software employed, and the variable selection method for QSAR model development should be considered integrative parts of the overall definition of an unambiguous algorithm.

12.3.2. Theoretical Molecular Descriptors

It has become quite common to use a wide set of molecular descriptors of different kinds (experimental and/or theoretical) that are able to capture all the structural aspects of a chemical to translate the molecular structure into numbers. The various descriptors are different ways or perspectives to view a molecule, taking into account the various features of its chemical structure, not only one-dimensional (e.g., the simple counts of atoms and groups), but also two-dimensional from a topological graph or three-dimensional from a minimum energy conformation. Livingstone has published a survey of these approaches [42]. Much of the software calculates broad sets of different theoretical descriptors, from SMILES, 2D-graphs to 3D-x,y,z-coordinates. Some of the frequently used descriptor calculation software includes ADAPT [43], OASIS [44], CODESSA [45], DRAGON [46], and MolConnZ [47]. It has been estimated that more than 3000 molecular descriptors are now available, and most of them have been summarized and explained [48–50]. The great advantage of theoretical descriptors is that they can be calculated homogeneously by a defined software for all chemicals, even those not yet synthesized, the only need being a hypothesized chemical structure. This peculiarity explains their wide and successful use in QSAR modeling. The DRAGON software has always been used in models developed by the author's group. In the version more frequently used by the author (5.4), 1664 molecular descriptors of the following different typologies were calculated: (a) 0D-48 constitutional (atom and group counts); (b) 1D-14 charge descriptors; (c) 1D-29 molecular properties; (d) 2D-119 topological; (e) 2D-47 walk and path counts, (f) 2D-33 connectivity index; (g) 2D-47 information index; (h) 2D-96 various auto-correlations from the molecular graph; (i) 2D-107 edge adjacency indices; (j) 2D-64 descriptors of Burden (BCUTs eigenvalues); (k) 2D-21 topological charge indices; (l) 2D-44 eigenvalue-based indices; (m) 3D-41 Randic molecular profiles; (n) 3D-74 geometrical descriptors; (o) 3D-150 radial distribution function; (p) 3D-160 Morse; (q) 3D-99 weighted holistic invariant molecular descriptors (WHIMs) [51–53]; (r) 3D-197 geometry, topology and atom-weights assembly (GETAWAY) descriptors [54, 55]; (s) 154 functional

groups; (t) 120 atom-centered fragments. The list and meaning of the molecular descriptors are provided by the DRAGON package and the calculation procedure is explained in detail, with related literature references, in the Handbook of Molecular Descriptors from Todeschini and Consonni [50] and in Chapter 3. The DRAGON software is continuously implemented with new descriptors.

12.3.3. Variable Selection and Reduction. The Genetic Algorithm Strategy for Variable Selection

The existence of a huge number of different molecular descriptors, experimental or theoretical, to describe chemical structure is a great resource as it allows QSAR modelers (particularly those working with the statistical approach) to have different X-variables available that take into account each structural feature in various ways. In principle, all the different possible combinations of the X-variables should be investigated to find the most predictive QSAR model. However, this can be quite taxing, mainly for reasons of time.

Sometimes molecular descriptors, which are only different views of the same molecular aspect, are highly correlated. Thus, when dealing with a large number of highly correlated descriptors, variable selection is necessary to find a simple and predictive QSAR model, which must be based on the minimum number of descriptors, and the least correlated, as possible. First, objective selection is applied using only independent variables (X): descriptors to discard are identified by tests of identical values and pairwise correlations, looking for descriptors less correlated to one another.

Secondly, modeling variable selection methods, which additionally use dependent variable values (Y), are applied to this pre-reduced set of descriptors to further reduce it to the true modeling set, not only in fitting but, most importantly, in prediction. Such selection is performed by alternative variable selection methods.

Several strategies for variable subset selection have been applied in QSAR (stepwise regressions, forward selection, backward elimination, simulated annealing, evolutionary and genetic algorithms, among those most widely applied). A comparison of these methods [56] has demonstrated the advantages, and the success, of genetic algorithms (GAs) as a variable selection procedure for QSAR studies.

GAs are a particular kind of evolutionary algorithms (EAs), shown to be able to solve complex optimization problems in a number of fields, including chemistry [57–59]. The natural principles of the evolution of species in the biological world are applied, i.e., the assumption that conditions leading to better results will prevail over poorer ones, and that improvement can be obtained by different kinds of recombination of independent variables, i.e., reproduction, mutation, and crossover. The goodness-of-fit of the selected solution is measured by a function that has to be optimized.

Genetic algorithms, first proposed as a strategy for variable subset selection in multivariate analysis by Leardi et al. [60] and applied to QSAR modeling by Rogers and Hopfinger [61], are a very effective tool with many merits compared to other methods. GAs are now widely and successfully applied in QSAR approaches,

where there is quite a number of molecular descriptors, in various modified versions, depending on the way of performing reproduction, crossover, mutation, etc. [62–66].

In variable selection for QSAR studies, a bit equal to 1 denotes a variable (molecular descriptor) present in the regression model or equal to 0 if excluded. A population, constituted by a number of 0/1 bit strings (each of length equal to the total number of variables in the model), is evolved following genetic algorithm rules, maximizing the predictive power of the models (verified by the explained variance in prediction, Q^2_{cv} or by the root mean squared error of prediction, RMSEcv). Only models producing the highest predictive power are finally retained and further analyzed with additional validation techniques.

Whereas EAs search for the global optimum and end up with only one or very few results [64, 65, 67], GAs simultaneously create many different results of comparable quality in larger populations of models with more or less the same predictive power. Within a given population the selected models can differ in the number and kind of variables. Similar descriptors, which are able to capture some specific aspects of chemical structure, can be selected by GA in alternative combinations for modeling the response. Thus, similarly performing models can be considered as different perspectives to arrive at essentially the same conclusion. Owing to this, the GA-based approach has no single "best" set of descriptors related to the Y-dependent variable; there is a population of good models of similar performance that could be also combined in consensus modeling approaches [18, 19] to obtain averaged predictions.

Different rules can be adopted to select the final preferred "best" models. In the author's researches the QUIK (Q under influence of K) rule [29] is always applied as the first filter to avoid multi-collinearity in model descriptors without prediction power or with "apparent" prediction power (chance correlation). According to this rule, only models with a K multivariate correlation calculated on the X+Y block, at least 5% greater than the K correlation of the X-block, are considered statistically significant and checked for predictivity (both internally by different cross-validations and externally on chemicals which do not participate in model development).

Another important parameter that must be considered is the root mean squared error (RMSE) that summarizes the overall error of the model; it is calculated as the root square of the sum of squared errors in calculation (RMSE) or prediction (RMSEcv and RMSEp) divided by their total number. The best model has the smallest RMSE and very similar RMSE values for training and external prediction chemicals, highlighting the model's generalizability [68].

12.4. APPLICABILITY DOMAIN (OECD PRINCIPLE 3)

The third OECD Principle takes into consideration another crucial problem: the definition of the applicability domain (AD) of a QSAR model. Even a robust, significant, and validated QSAR model cannot be expected to reliably predict the property

modelled for the entire universe of chemicals. In fact, only predictions for chemicals falling within the domain of the developed model can be considered reliable and not model extrapolations. This topic was dealt with at a recent workshop where several different approaches for linear and non-linear models were proposed [69], in relation to different model types.

The AD is a theoretical spatial region defined by the model descriptors and the response modelled, and is thus defined by the nature of the chemicals in the training set, represented in each model by specific molecular descriptors. To clarify recent doubts [70], it is important to note that each QSAR model has its own specific AD based on the training set chemicals, not just on the kind of included chemicals but also on the values of the specific descriptors used in the model itself; such descriptors are dependent on the type of the training chemicals.

As was explained above, a population of MLR models of similar good quality, developed by variable selection performed with a genetic algorithm [66] can include a 100 different models developed on the same training set but based on different descriptors: even if developed on the same chemicals, the AD for new chemicals can differ from model to model, depending on the specific descriptors. Through the leverage approach [71] (shown below) it is possible to verify whether a new chemical will lie within the model domain (in this case predicted data can be considered as interpolated and with reduced uncertainty, at least similar to that of training chemicals, thus more reliable) or outside the domain (thus, predicted data are extrapolated by the model and must be considered of increased uncertainty, thus less reliable). If it is outside the model domain a warning must be given. Leverage is used as a quantitative measure of the model applicability domain and is suitable for evaluating the degree of extrapolation, which represents a sort of compound "distance" from the model experimental space (the structural centroid of the training set). It is a measure of the influence a particular chemical's structure has on the model: chemicals close to the centroid are less influential in model building than extreme points. A compound with high leverage in a QSAR model would reinforce the model if the compound is in the training set, but such a compound in the test set could have unreliable predicted data, the result of substantial extrapolation of the model.

The prediction should be considered unreliable for compounds in the test set with high-leverage values ($h > h^*$, the critical value being $h^* = 3p'/n$, where p' is the number of model variables plus one and n is the number of the objects used to calculate the model). When the leverage value of a compound is lower than the critical value, the probability of accordance between predicted and actual values is as high as that for the training set chemicals. Conversely, a high-leverage chemical is structurally distant from the training chemicals, thus it can be considered outside the AD of the model. To visualize the AD of a QSAR model, the plot of standardized cross-validated residuals (R) vs. leverage (Hat diagonal) values (h) (the Williams plot) can be used for an immediate and simple graphical detection of both the response outliers (i.e., compounds with cross-validated standardized residuals greater than three standard deviation units, $>3\sigma$) and structurally influential chemicals in a model ($h > h^*$).

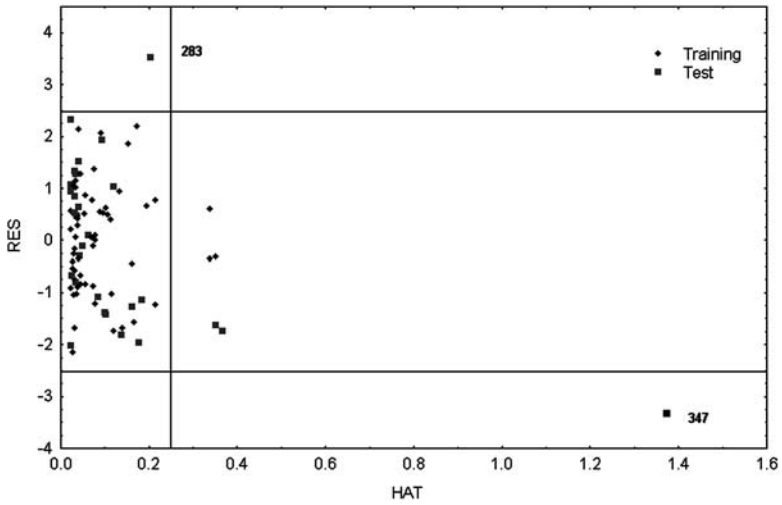

Figure 12-1. Williams plot for an externally validated model for the toxicity to *Pimephales promelas* of polar narcotics. Cut-off value: 2.5 h* (with copyright permission from [26])

It is important to note that the AD of a model cannot be verified by studying only a few chemicals, as in such cases [72] it is impossible to obtain conclusions that can be generalized on the applicability of the model itself.

Figure 12-1 shows the Williams plot of a model for compounds that act as polar narcotics to *Pimephales promelas* [26]; as an example, here the toxicity of chemical no. 347 is incorrectly predicted (>3σ) and it is also a test chemical completely outside the AD of the model, as defined by the Hat vertical line (high h leverage value), thus it is both a response outlier and a high-leverage chemical. Two other chemicals (squares at 0.35 h) slightly exceed the critical hat value (vertical line) but are close to three chemicals of the training set (rhombus), slightly influential in the model development: the predictions for these test chemicals can be considered as reliable as those of the training chemicals. The toxicity of chemical no. 283 is incorrectly predicted (>3σ), but in this case it belongs to the model AD, being within the cut-off value of Hat. This erroneous prediction could probably be attributed to error or variability in the experimental data rather than to molecular structure or model.

12.5. MODEL VALIDATION FOR PREDICTIVITY (OECD PRINCIPLE 4)

Model validation must always be used to avoid the possibility of "overfitted" models, i.e., models where too many variables, useful only for fitting the training data, have been selected, and to avoid the selection of variables randomly correlated (by chance) with the dependent response. Particular care must be taken against overfitting [30], thus subsets with the fewest variables are favored, as the chance of finding

"apparently acceptable" models increases with increasing X-variables. The proportion of random variables selected by chance correlation could also increase [73]. The ratio of chemicals to variables should always be higher than five for a small data set, but the number of descriptors must be the lowest as possible for bigger data sets too (according to the Ockham's Razor: "avoid complexity if not necessary").

Therefore, a set of models of similar performance, verified by leave-one-out model validation, need to be further validated by leave-more-out cross-validation or bootstrap [74, 75]. This is done to avoid overestimation of the model's predictive power by Q^2_{LOO} [76, 77] and to verify the stability of model predictivity (robustness). Response permutation testing (Y scrambling) [6] or other resampling techniques are also applied for excluding that the developed model is based on descriptors that could be related to the response only by chance. Finally, for the most stringent evaluation of model applicability for prediction of new chemicals, external validation (verified by Q^2_{EXT} or R^2_{EXT}) of all models is recommended as the last step after model development, and for the assessment of true predictive ability [6, 10, 78].

The preferred model will be that with the highest prediction parameter values and the most balanced results between the cross-validation parameters on the training chemicals (Q^2_{cv}, Q^2_{LMO}, Q^2_{BOOT}), verified during descriptor selection, and the predictive power (Q^2_{EXT} or R^2_{EXT}), verified later on the external prediction chemicals.

The limiting problem for efficient external validation of a QSAR model is, obviously, data availability. Given the availability of a sufficiently large number (never less than five or 20% of training set) of really new and reliable experimental data, the best proof of an already developed model accuracy is to test model performance on these additional data, at the same time checking the chemical AD. However, it is usually difficult to have data available for new experimentally tested compounds (in useful quantity and quality) for external validation purposes, thus, in the absence of additional data, external validation by *a priori* splitting the available data can be usefully applied to define the actual predictive power of the model more precisely.

12.5.1. Splitting of the Data Set for the Construction of an External Prediction Set

In the absence of new additional data, we assume that there is less data than is actually available; this is the reason for splitting the data in a reasonable way (commented on below) into a training set and a prediction set of "momentarily forgotten chemicals."

Thus, before model development, the available input data set can be split adequately by different procedures into the training set (for model development) and the prediction set (never used for variable selection and model development, but used exclusively once for model predictive assessment, performed only after model development). At this point the underlying goal is to ensure that both the training and prediction sets separately span the whole descriptor space occupied by the entire data set, and that the chemical domain in the two data sets is not too dissimilar [77, 79–81] as it is impossible for a model to be applied outside its chemical domain and obtain reliable predictions. The composition of the training and prediction sets

is of crucial importance. The best splitting must guarantee that the training and prediction sets are scattered over the whole area occupied by representative points in the descriptor space (representativity), and that the training set is distributed over an area occupied by representative points for the whole data set (diversity). The more widely applied splitting methodologies are based on structural similarity analysis (for instance, Kennard Stone, duplex, D-optimal distance [11–13, 17, 18, 20, 21, 81, 82], self-organizing map (SOM) or Kohonen-map ANN [17, 18, 20, 21, 26, 27, 35, 39, 41, 80]. Alternatively, to split the available data without any bias for structure, random selection through activity sampling can be applied. Random splitting is highly useful if applied iteratively in splitting for CV internal validation and can be considered quite similar to real-life situations, but it can give very variable results when applied in this external validation, depending greatly on set dimension and representativity [80, 83, 84]. In addition, in this last case there is a greater probability of selecting chemicals outside the model structural AD in the prediction set; thus, the predictions for these chemicals could be unreliable, simply as they are extrapolated by the model.

12.5.2. Internal and External Validation

External validation should be applied to any proposed QSAR model to determine both its generalizability for new chemicals that, obviously, must belong to the model AD and the "realistic" predictive power of the model [6, 83–85]. The model must be tested on a sufficiently large number of chemicals not used during its development, at least 20% of the complete data set is recommended, but the most stable models (of easily modelled endpoints) can also be checked on a prediction set larger than the training set [19, 85]; this will avoid "supposed" external validation based on too few chemicals [72]. In fact, it has been demonstrated that if the test set consists only of a small number of compounds, there is increased possibility of chance correlation between the predicted and observed response of the compounds [79].

It is not unusual for models with high internal predictivity, verified by internal validation methods (LOO, LMO, Bootstrap), but externally less predictive or even absolutely unpredictive, to be present in populations of models developed using evolutionary techniques to select the descriptors. The statistical approach to QSAR modeling always carefully checks this possibility by externally validating any model, stable in cross-validation, before its proposal. In fact, cross-validation is necessary but is not a sufficient validation approach for really predictive models [6, 77–79]. In relation to this crucial point of QSAR model validation, there is a wide debate and discordant opinions in the QSAR community concerning the different outcomes of internal and external validation on QSAR models. A mini-review dealing with this problem has been recently published by the author [84], where an examination is made of the OECD Principles 2, 3, and 4, and particular attention has been paid to the differences in internal and external validation. The theoretical constructs are illustrated with examples taken from both the literature and personal experience, derived also from a recent report for the European Centre for Validation

of Alternative Methods (ECVAM) on "Evaluation of different statistical approaches to the validation of Quantitative Structure–Activity Relationships" [83].

Since GAs simultaneously create many different, similarly acceptable models in a population, the user can choose the "best model" according to need: the possibility of having reliable predictions for some chemicals rather than others, the interpretability of the selected molecular descriptors, the presence of different outliers, etc.

In the statistical approach the best model is selected by maximizing all the CV internal validation parameters, by applying CV in the proper way and step. Then, only the good models (Q^2_{LOO}>0.7), stable and internal predictive (with similar values of all the different CV-Q^2), are subjected to external validation on the *a priori* split prediction set.

In our works we always select, from among the best externally predictive models, those with the smallest number of response outliers and structurally influential chemicals, especially those in the prediction set.

12.5.3. Validation of Classification Models

To assess the predictive ability of classification models, the percentage of misclassified chemicals, as error rate (ER%) and error rate in prediction (ER_{cv}%), are calculated by the leave-one-out method (where each chemical is taken out of the training set once and predicted by the model). Comparison with the no-model error rate (NoMER) is used to evaluate model performance. NoMER represents the object distribution in the defined classes before applying any classification method, and is calculated as an error rate by considering all the objects as misclassified into the greatest class. This provides a reference classification parameter to evaluate the actual efficiency of a classifier: the greater the difference between NoMER and the actual ER, the better the model performance.

The outputs of a classification model are the class assignments and the misclassification matrix, which shows how well the classes are separated. The goodness of the classification models is also assessed by the following parameters: accuracy or concordance (the proportion of correctly classified chemicals), sensitivity (the proportion of active chemicals predicted to be active), specificity (the proportion of non-active chemicals predicted to be non-active), false negatives (the proportion of active chemicals falsely predicted as non-active) and false positives (the proportion of non-active chemicals falsely predicted as active). Depending on the intended application of the predictive tool, the classification model can be optimized in either direction. In drug design the objective is to obtain a high specificity as a false positive prediction could result in the loss of a valuable candidate. In the regulatory environment, for safety assessment and consumer protection, the precautionary principle must be applied, so an optimization of sensitivity would be desirable, as every false negative compound could result in a lack of protection and consequently pose a risk for the user.

12.6. MOLECULAR DESCRIPTOR INTERPRETATION, IF POSSIBLE (OECD PRINCIPLE 5)

Regarding the interpretability of the descriptors it is important to take into account that the response modelled is frequently the result of a series of complex biological or physico-chemical mechanisms, thus it is very difficult and reductionist to ascribe too much importance to the mechanistic meaning of the molecular descriptors used in a QSAR model. Moreover, it must also be highlighted that in multivariate models such as MLR models, even though the interpretation of the singular molecular descriptor can certainly be useful, it is only the combination of the selected set of descriptors that is able to model the studied endpoint. If the main aim of QSAR modeling is to fill the gaps in available data, the modeler's attention should be focused on model quality. In relation to this point, Livingstone, in an interesting perspective paper [42] states: "The need for interpretability depends on the application, since a validated mathematical model relating a target property to chemical features may, in some cases, be all that is necessary, though it is obviously desirable to attempt some explanation of the 'mechanism' in chemical terms, but it is often not necessary, per se." Zefirov and Palyulin [78] took the same position, differentiating predictive QSARs, where attention essentially concerns the best prediction quality, from descriptive QSARs where the major attention is paid to descriptor interpretability.

The author's approach to QSAR modeling will be illustrated in the following sections of this chapter through the modeling of environmental endpoints. The approach starts with a statistical validation for predictivity and continues on through further interpretation for the mechanistic meaning of the selected descriptors, but only if possible, as set down by the fifth OECD principle [6]. Therefore, the application domain of this approach (the "statistical approach") is mainly related to the production of predicted data (predictive QSAR), strongly verified for their reliability; such data can be more usefully applied to screen and rank chemicals providing priority lists.

12.7. ENVIRONMENTAL SINGLE ENDPOINTS

12.7.1. Physico-chemical Properties

Organic chemicals now need to be characterized by many parameters, either because of the registration policy required to chemical industries (see for example, the new European REACH policy) or for an understanding of the environmental behavior of chemicals present as pollutants in various compartments. Unfortunately there is an enormous lack of knowledge for many important endpoints, such as various physico-chemical properties (for instance, melting point, boiling point, aqueous solubility, volatility, hydrophobicity, various partition coefficients), environmental reactivity and derived persistence, toxicity, mutagenicity. This lack of knowledge calls for a predictive approach to the assessment of chemicals, such as by QSAR modeling.

A set of various physico-chemical properties for important classes of chemicals present in the environment, pollutant compounds such as PAHs [86] haloaromatics [87], PCBs [88], chemicals of EEC Priority List 1 [89] have been modelled using the weighted holistic invariant molecular (WHIM) descriptors [51–53, 90, 91]. WHIM descriptors are theoretical three-dimensional molecular indices that contain information, in terms of size, shape, symmetry, and atom distribution, on the whole molecular structure. These indices are calculated from the (x, y, z) coordinates of a molecule within different weighting schemes by principal component analysis and represent a very general approach to describe molecules in a unitary conceptual framework, independent from the molecular alignment. Their meaning is defined by the same mathematical properties of the algorithm used for their calculation, and their application in QSAR modeling was very successful. A recent paper [92] again highlighted that, contrary to erroneous statements in the literature [93, 94], one set of WHIM descriptors, the k descriptors, are very useful in discriminating the shape of chemicals and can thus be used to study structural similarity.

Since then other physico-chemical properties have been modelled successfully by combining different kinds of theoretical molecular descriptors (mono-dimensional, bi-dimensional, and three-dimensional) calculated by the DRAGON software [46]: the basic physico-chemical properties of organic solvents [95], esters [15] and brominated flame retardants, mainly polybromodiphenyl ethers (PBDE) [24], the soil sorption coefficient (K_{oc}) for pesticides [19, 96] (discussed below in Section 12.7.1.1).

A general classification of 152 organic solvents has been proposed [95] by applying the k-nearest neighbor method and counter propagation artificial neural networks (CP-ANN), in particular Kohonen-maps. A good separation for five classes was obtained by the net architecture ($20\times20\times4$, 200 iterations), based on simple molecular descriptors (unsaturation index – UI, hydrophilicity factor – Hy, average atomic composition – AAC, and the number of nitrogen atoms in the molecular structure – nN). The performances were very satisfactory: ER (%)=4.4 and ER_{cv} (%)=11.4 (to be compared with the error rate without the model NoMER (%)=69.5.)

12.7.1.1. Soil Sorption of Pesticides

Sorption processes play a major role in determining the environmental fate, distribution, and persistence of chemicals. An important parameter when studying soil mobility and environmental distribution of chemicals is the soil sorption coefficient, expressed as the ratio between chemical concentration in soil and in water, normalized to organic carbon (K_{oc}).

Many QSAR papers on soil sorption coefficient prediction have been published and reviewed by some authors [85, 96–104].

The proposed models were mainly based on the correlation with octanol/water partition coefficients (K_{ow}) and water solubility (S_w), others on theoretical molecular structure descriptors. A recent paper by the author dealt with log K_{oc} of a heterogeneous set of 643 organic non-ionic compounds [19]; the response range

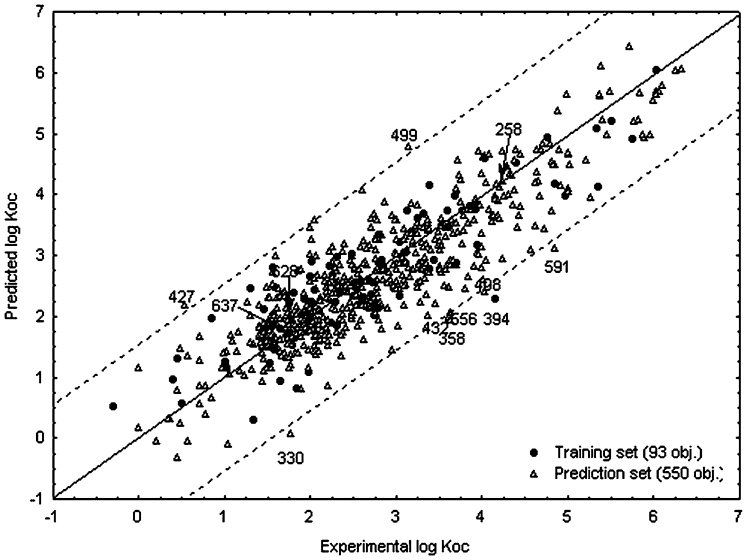

Figure 12-2. Plot of experimental vs. predicted log K_{oc} for the Eq. (12-1). The values for the training and prediction set chemicals are labeled differently, the outliers are numbered. The *dotted lines* indicate the 3σ interval (with copyright permission from [19])

was more than six log units, and prediction was made by a statistically validated QSAR modeling approach based on MLR and theoretical molecular descriptors, selected by GA from DRAGON (see Eq. 12-1). The high generalizability of one of the proposed models (scatter plot in Figure 12-2) was verified on external chemicals, performed by adequately splitting, by SOM and also randomly, the available set of experimental data into a very reduced representative training set (even less than 15% of the original data set) for model development and a large prediction set (more than 85% of the original data) used only for model performance inspection.

$$\log K_{oc} = -2.19(\pm 0.30) + 2.10(\pm 0.14)\text{VED1} - 0.34(\pm 0.04)\text{nHAcc} - 0.31 (\pm 0.05)\text{MAXDP} - 0.33(\pm 0.12)\text{CIC0}$$

$$\text{n(training)} = 93\ R^2 = 0.82\ Q^2_{cv} = 0.80\ Q^2_{BOOT} = 0.79\ \text{RMSE} = 0.523\ \text{RMSEp}_{LOO} = 0.523$$

$$\text{n(prediction set)} = 550\ Q^2_{EXT} = 0.78\ \text{RMSEp}_{EXT} = 0.560 \qquad (12\text{-}1)$$

The proposed models have good stability, robustness, and predictivity when verified by internal validation (cross-validation by LOO and Bootstrap) and also by external validation on a much greater data set. The stability of RMSE/RMSEp for both the training and prediction sets is further proof of model predictivity.

The chemical applicability domain is verified by the Williams graph: nine outliers for response and three structurally influential chemicals have been highlighted (numbered in Figure 12-2).

The selected molecular descriptors have a clear mechanistic meaning; they are related to both the molecular size of the chemical and its electronic features relevant to soil partitioning, as well as to the chemical's ability to form hydrogen bonds with water. A combination of different models from the GA-model population also allowed the proposal of predictions obtained by the better consensus model that, compared with published models and EPISuite predictions [105], are always among the best. The proposed models fulfill the fundamental points set down by OECD principles for the regulatory acceptability of a QSAR and could be reliably used as scientifically valid models in the REACH program.

The application of a single and general QSAR model, based on theoretical molecular descriptors for a large set of heterogeneous compounds, could be very useful for the screening of big data sets and for designing new chemicals, environmentally friendly as safer alternatives to dangerous chemicals.

12.7.2. Tropospheric Reactivity of Volatile Organic Compounds with Oxidants

The troposphere is the principal recipient of volatile organic compounds (VOCs) of both anthropogenic and biogenic origin. An indirect measure of the persistence of organic compounds in the atmosphere, and therefore a necessary parameter in environmental exposure assessment, is the rate at which these compounds react. The tropospheric lifetime of most organic chemicals, deriving from terrestrial emissions, is controlled by their degradation reaction with the OH radical and ozone during the daytime and NO_3 radicals at night.

In recent years, several QSAR/QSPR models predicting oxidation rate constants with tropospheric oxidants have been published and the different approaches to molecular description and the adopted methodology have been compared [13, 14, 18, 23, 106–117].

The most used method, implemented in AOPWIN of EPISUITE [118] for estimating tropospheric degradation by hydroxyl radicals is Atkinson's fragment contribution method [107]. New general MLR models of the OH radical reaction rate for a wide and heterogeneous data set of 460 volatile organic compounds (VOCs) were developed by the author's group [18]. The special feature of these models, in comparison to others, is the selection of theoretical molecular descriptors by a genetic algorithm as a variable subset selection procedure, their applicability to heterogeneous chemicals, and their validation for predictive purposes by both internal and external validation. External validation was performed by splitting the original data set by two different methods: the statistical experimental design procedure (D-optimal distance) and the Kohonen self-organizing map (SOM); this was performed to verify the impact that the structural heterogeneity (in chemicals' split into training and prediction sets) has on model performance. The consequences on

the model predictivity are also compared. D-optimal design, where the most dissimilar chemicals are always selected for the training set, leads to models with better predictive performance than models developed on the training set selected by SOM. The chemical applicability domain of the models and the reliability of the predictions are always verified by the leverage approach. The best proposed predictive model is based on four molecular descriptors and has the following equation (12-2):

$$\log\ \mathrm{k(OH)} = 5.15(\pm 0.35) - 0.66(\pm 0.03)\mathrm{HOMO} + 0.33(\pm 0.03)\mathrm{nX} - 0.37 \\ (\pm 0.04)\mathrm{CIC0} + (\pm 0.02)0.13\ \mathrm{nCaH}$$

$$\mathrm{n(training)} = 234\ R^2 = 0.83\ Q^2 = 0.82\ Q^2\mathrm{LMO}(50\%) = 0.81\ \mathrm{RMSE} = 0.473$$

$$\mathrm{n(test)} = 226\ Q^2_{\mathrm{EXT}} = 0.81\ \mathrm{RMSEp} = 0.484\ K_{xx} = 33.8\%\ K_{xy} = 44.6\% \quad (12\text{-}2)$$

It is evident from the statistical parameters that the proposed model has good stability, robustness, and predictivity verified by internal (cross-validation by LOO and LMO) and also external validation. The influential chemicals are mainly the highly fluorinated chemicals, which have a strong structural peculiarity that the model is not able to capture. In Figure 12-3 the experimental values vs. those predicted by Eq. (12-2) are plotted.

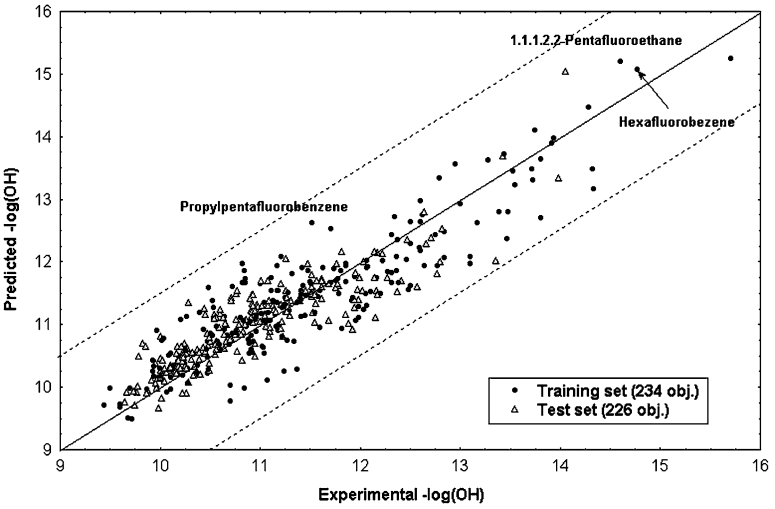

Figure 12-3. Plot of experimental and predicted log k(OH) values for the externally validated model by experimental design splitting. The training and test set chemicals are labeled differently, the outliers and influential chemicals are highlighted. The *dotted lines* indicate the 3σ interval (with copyright permission from [18])

The availability in the GA population of several possible models, similarly reliable for response prediction, also allowed the proposal of a consensus model which provides better predicted data than the majority of individual models, taking into account the more unique aspects of a particular structure.

While good models for OH rate constants are proposed in the literature for various chemical classes [107, 110–113, 115, 117], the modeling of reactivity with NO_3 radicals is more problematic. Most published QSAR models were obtained from separate training sets for aliphatic and aromatic compounds and the rate constants of aliphatic chemicals with NO_3 radicals were successfully predicted [106, 108, 109]; however, the models for aromatic compounds do not appear to be so satisfactory, often being only local models built on very small training sets and, consequently, without any reasonable applicability for data prediction.

New general QSAR models for predicting oxidation rate constants (kNO_3) for heterogeneous sets containing both aliphatic and aromatic compounds, based on few theoretical molecular descriptors (for instance, HOMO, number of aromatic rings, and an autocorrelation descriptor, MATS1m), were recently developed by the author's group [13, 23]. The models have high predictivity even on external chemicals, obtained by splitting the available data using different methods. The possibility of having molecular descriptors available for all chemicals (even those not yet synthesized), the good prediction performance of models applicable to a wide variety of aromatic and aliphatic chemicals, and the possibility of verifying the chemical domain of applicability by the leverage approach makes these useful models for producing reliable estimated NO_3 radical rate constants, when experimental parameters are not available.

The author has also proposed a predictive QSAR model of reaction rate with ozone for 125 heterogeneous chemicals [14]. The model, based on molecular descriptors, always selected by GA (HOMO–LUMO gap plus four molecular descriptors from DRAGON), has good predictive performance, also verified by statistical external validation on 42 chemicals not used for model development (Q^2_{EXT}=0.904, average RMS=0.77 log units). This model appears more predictive than the model previously proposed by Pompe and Veber [114], a six-parameter MLR model developed on 116 heterogeneous chemicals and based on molecular descriptors, calculated by the CODESSA software, selected by a stepwise selection procedure. The predictive performance of this model was verified only internally by cross-validation with 10 groups of validation (Q^2=0.83) and had an average RMS of 0.99 log units.

12.7.3. Biological Endpoints

12.7.3.1. Bioconcentration Factor

The bioconcentration factor (BCF) is an important parameter in environmental assessment as it is an estimate of the tendency of a chemical to concentrate and, consequently, to accumulate in an organism. The most common QSAR method, and the oldest, for estimating chemical bioconcentration is to establish correlations between

BCF and chemical hydrophobicity using K_{ow}, i.e., the n-octanol/water partition coefficient. A comparative study of BCF models based on log K_{ow} was performed by Devillers et al. [119]. Different models for BCF using theoretical molecular descriptors have been developed, among others: [120–124] and also by the author's group [8, 9, 27], with particular attention, as usual, to the external predictivity and the chemical applicability domain.

An example is the model reported by the following equation (12-3):

$$\begin{aligned}\log \mathrm{BCF} = {} & -0.74(\pm 0.35) + 2.55(\pm 0.13)\,{}^{\mathrm{V}}\mathrm{I}^{\mathrm{M}}_{\mathrm{D,deg}} - 1.09(\pm 0.11)\mathrm{HIC} \\ & -0.42(\pm 0.03)\mathrm{nHAcc} - 1.22(\pm 0.17)\mathrm{GATS1e} - 1.55(\pm 0.34)\mathrm{MATS1p}\end{aligned}$$

$$\begin{aligned}& n_{(\mathrm{training})} = 179\ R^2 = 0.81\ Q^2_{\mathrm{LOO}} = 0.79\ Q^2_{\mathrm{BOOT}} = 0.79 \\ & \mathrm{RMSE}_{(\mathrm{train\ set})} = 0.56\ \mathrm{RMSE}_{(\mathrm{cross-val.\ set})} = 0.58\end{aligned}$$

$$n_{(\mathrm{prediction})} = 59\ Q^2_{\mathrm{EXT}} = 0.87\ \mathrm{RMSE}_{(\mathrm{prediction\ set})} = 0.57 \qquad (12\text{-}3)$$

12.7.3.2. Toxicity

Acute aquatic toxicity. The European Union's so-called "List 1" of priority chemicals dangerous for the aquatic environment (more than 100 heterogeneous chemicals) was modelled for ecotoxicological endpoints (aquatic toxicity on bacteria, algae, *Daphnia*, fish, mammals) [89] by different theoretical descriptors, mainly WHIM. In addition, WHIM descriptors were also satisfactory in the modeling of a more reduced set of toxicity data on *Daphnia* (49 compounds including amines, chlorobenzenes, organotin and organophosphorous pesticides) [125].

An innovative strategy for the selection of compounds with a similar toxicological mode of action was proposed as a key problem in the study of chemical mixtures (PREDICT European Research Project) [126]. A complete representation of chemical structures for phenylureas and triazines by different molecular descriptors (1D-structural, 2D-topological, 3D-WHIM) allowed a preliminary exploration of structural similarity based on principal components analysis (PCA), multidimensional scaling (MDS), and hierarchical cluster. The use of a genetic algorithm to select the most relevant molecular descriptors in modeling toxicity data makes it possible both to develop good predictive toxicity models and select the most similar phenylureas and triazines. The way of doing this is to apply chemometric approaches based only on molecular similarity related to toxicological mode of action.

The Duluth data set of toxicity data to *P. promelas* was recently studied by the author group [26] and new statistically validated MLR models were developed to predict the aquatic toxicity of chemicals classified according to their mode of action (MOA). Also, a unique general model for direct toxicity prediction (DTP model) was developed to propose a predictive tool with a wide applicability domain, applicable independently of *a priori* knowledge of the MOA of chemicals.

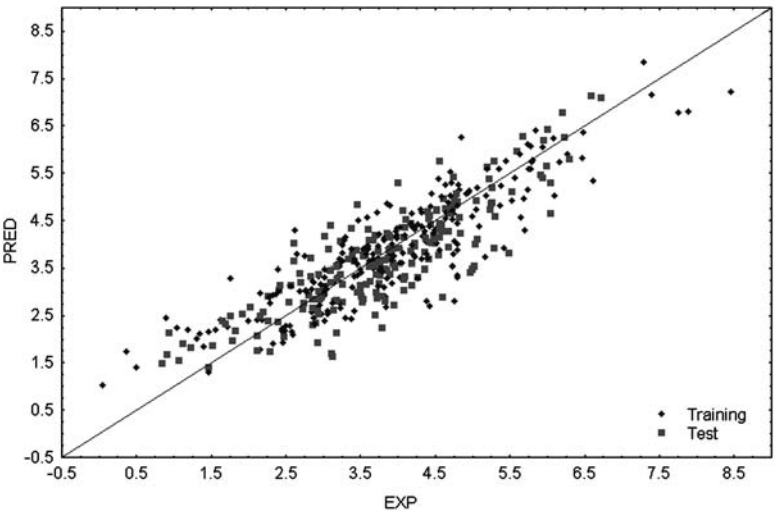

Figure 12-4. Plot of experimental and predicted toxicity values (*Pimephales promelas*) of the externally validated general-DTP log P-free model developed on a training set of 249 compounds (with copyright permission from [26])

The externally validated general-DTP log P-free model, reported below (Eq. 12-4) with statistical parameters, was developed on a training set of 249 compounds and applied for the prediction of the toxicity of 200 external chemicals, obtained by splitting the data by SOM (scatter plot in Figure 12-4):

$$\log(1/LC_{50})_{96h} = -2.54 + 0.91WA + 6.2Mv + 0.21nCb^{-} + 0.08H - 046 - 0.19MAXDP - 0.33nN$$

$$n_{training} = 249\ R^2 = 0.79\ Q^2_{LOO} = 0.78\ Q^2_{BOOT} = 0.78\ RMSE = 0.595$$

$$n_{test} = 200\ Q^2_{EXT} = 0.71\ RMSEcv = 0.613\ RMSEp = 0.64 \qquad (12\text{-}4)$$

Chronic toxicity: mutagenicity. The potential for mutagenicity of chemicals of environmental concern, such as aromatic amines and PAHs, is of high relevance; many QSAR models, based on the mechanistic approach, have been published on this topic and reviewed by Benigni [5, 127].

With regard to this important topic, our group has published useful MLR models, always verified for their external predictivity on new chemicals, for the Ames test results on amines [12] and nitro-PAHs [20]. Externally validated classification models, by k-NN and CART, were also developed for the mutagenicity of benzocyclopentaphenanthrenes and chrysenes, determined by the Ames test [128], and PAH mutagenicity, determined on human B-lymphoblastoid [35].

Endocrine Disruption. A large number of environmental chemicals, known as endocrine disruptor chemicals (EDCs), are suspected of disrupting endocrine functions by mimicking or antagonizing natural hormones. Such chemicals may pose a serious threat to the health of humans and wildlife; they are thought to act through a variety of mechanisms, mainly estrogen receptor-mediated mechanisms of toxicity. Under the new European legislation REACH (http://europa.eu.int/comm/environment/chemicals/reach.htm) EDCs will require an authorization to be produced and used, if safer alternative are not available. However, it is practically impossible to perform thorough toxicological tests on all potential xenoestrogens, thus QSAR modeling has been applied by many other authors in these last years [129–142] providing promising methods for the estimation of a compound's estrogenic activity.

QSAR models of the estrogen receptor binding affinity of a large data set of heterogeneous chemicals have been built also in our laboratory using theoretical molecular descriptors [21, 33] giving full consideration, during model construction and assessment, to the new OECD principles for the regulatory acceptance of QSARs. A data set of 128 NCTR compounds (EDKB, http://edkb.fda.gov/databasedoor.html) including several different chemical categories, such as steroidal estrogens, synthetic estrogens, antiestrogens, phytoestrogens, other miscellaneous steroids, alkylphenols, diphenyl derivatives, organochlorines, pesticides, alkylhydroxybenzoate preservatives (parabens), phthalates, and a number of other miscellaneous chemicals, was studied. An unambiguous multiple linear regression (MLR) algorithm was used to build the models by selecting the modeling descriptors by a genetic algorithm. (Table 12-1 presents the statistical parameters of the best-selected model.) The predictive ability of the model was validated, as usually, by both internal and external validation, and the applicability domain was checked by the leverage approach to verify prediction reliability.

Twenty-one chemicals of the Kuiper data set [143] were used for external validation, with the following highly satisfying results: R^2_{pred}=0.778, Q^2_{EXT}=0.754, RMSE of prediction of 0.559 (Figure 12-5).

The results of several validation paths using different splitting methods performed in parallel (D-optimal design, SOM, random on activity sampling) give additional proof that the proposed QSAR model is robust and satisfactory (R^2_{pred} range: 0.761–0.807), thus providing a feasible and practical tool for the rapid screening of the estrogen activity of organic compounds, supposed endocrine disruptors chemicals.

On the same topic, satisfactory predictive models for the EDC classification based on different classification methods have been developed and recently proposed [33]. In this study, QSAR models were developed to quickly and effectively identify possible estrogen-like chemicals based on 232 structurally diverse chemicals from the NCTR database (training set) by using several non-linear classification methodologies (least square support vector machine (LS-SVM), counter propagation artificial neural network (CP-ANN), and k-nearest neighbor (kNN)) based on molecular structural descriptors. The models were validated externally with 87 chemicals (prediction set) not included in the training set. All three methods gave

Table 12-1. The MLR model between the structural descriptor and the log RBA of estrogens

Variable	Full name of variable	*Reg.coeff.*	*Err.coeff.*	*Std.coeff.*
Intercept		15.83	2.20	0.00
X2A	Average connectivity index chi-2	−43.75	5.28	−0.49
TIC1	Total information index (neighborhood symmetry first-order)	0.04	0.00	0.89
EEig02d	Eigenvalue 2 from edge adjacency matrix weighted dipole moments	−2.67	0.31	−0.56
JGI10	Mean topological charge index of order 10	79.92	10.85	0.32
SPH	Spherosity index	2.60	0.56	0.24
E1u	The first component accessibility directional WHIM index/unweighted	−7.12	1.57	−0.25
RTm+	R maximal index weighed by atomic masses	4.78	0.74	0.28
nArOR	The number of aromatic ether groups	−1.25	0.15	−0.39

Model parameters: n=128, R^2=0.824, R^2_{adj}=0.812, Q^2_{LOO}=0.793, Q^2_{BOOT}=0.780, RMSEcv=0.7484, RMSEp=0.8105, K_x=35.13, K_{xy}=37.89, and s=0.7762.

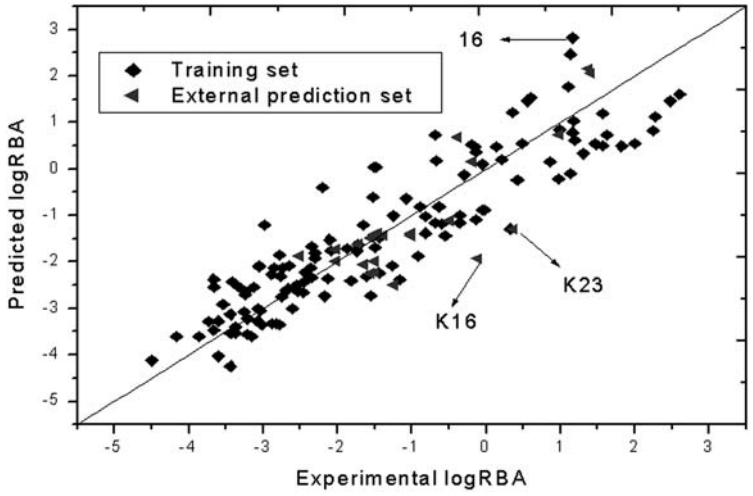

Figure 12-5. Predicted Log RBA values vs. experimental values for the original data set of estrogens (NCTR data set) and external prediction set (Kuiper's data set) (with copyright permission from [21])

satisfactory prediction results both for training and prediction sets; the most accurate model was obtained by the LS-SVM approach. The highly important feature of all these models is their low false negative percentage, useful in a precautionary approach. Our models were also applied to about 58,000 discrete organic chemicals from US-EPA; about 76% were predicted, by each model, not to bind to an estrogen receptor.

The obtained results indicate that the proposed QSAR models are robust, widely applicable, and could provide a feasible and practical tool for the rapid screening of potential estrogens. It is very useful information to prioritize chemicals for more expensive assays. In fact, the common 40,300 negative compounds could be excluded from the potential estrogens without experiments and a high accuracy (low false negative value).

A review on the applications of machine learning algorithms in the modeling of estrogen-like chemicals has been recently published [144].

12.8. MODELING MORE THAN A SINGLE ENDPOINT

12.8.1. PC Scores as New Endpoints: Ranking Indexes

The environment is a highly complex system in which many parameters are of contemporaneous relevance: the understanding, rationalization, and interpretation of their covariance are the principal pursuit of any environmental researcher. Indeed, environmental chemistry deals with the behavior of chemicals in the environment, behavior which is regulated by many different variables such as physico-chemical properties, chemical reactivity, biological activity.

The application of explorative methods of multivariate analysis to various topics of environmental concern allows a combined view that generates ordination and grouping of the studied chemicals, in addition to the discovering of variable relationships. Any problem related to chemical behavior in the environment can be analyzed by multivariate explorative techniques, the outcome being to obtain chemical screening and ranking according to the studied properties, reactivities, or activities and, finally, the proposal of an index.

This was the starting point, and also the central core, of most of the author 15-year research of QSAR modeling at Insubria University.

The significant combination of variables from multivariate analysis can be used as a score value (a cumulative index), and modelled as a new endpoint by the QSAR approach to exploit already available information concerning chemical behavior, and to propose models able to predict such behavior for chemicals for which the same information is not yet known, or even for new chemicals before their synthesis. In fact, our QSAR approach, both for modeling quantitative response by regression methods and qualitative response by classification methods, is based on theoretical molecular descriptors that can be calculated for any drawn chemicals starting from the atomic coordinates, thus without the knowledge of any experimental parameter.

12.8.2. Multivariate Explorative Methods

The principal aim of any explorative technique is to capture the information available in any multivariate context and condense it into a more easily interpretable view (a score value or a graph). Thus, from these exploratory tools a more focused investigation can be made into chemicals of higher concern, directing the next investigative

steps or suggesting others. Some of the more commonly used exploratory techniques are commented on here and applied in environmental chemistry and ecotoxicology.

12.8.2.1. Principal Component Analysis

Probably the most widely known and used explorative multivariate method is principal component analysis (PCA) [145, 146] (Chapter 6). In PCA, linear combinations of the studied variables are created, and these combinations explain, to the greatest possible degree, the variation in the original data. The first principal component (PC1) accounts for the maximum amount of possible data variance in a single variable, while subsequent PCs account for successively smaller quantities of the original variance. Principal components are derived in such a way that they are orthogonal. Indeed, it is good practice, especially when the original variables have different ranges of scales, to derive the principal components from the standardized data (mean of 0 and standard deviation of 1), i.e., via the correlation matrix. In this way all the variables are treated as if they are of equal importance, regardless of their scale of measurement. To be useful, it is desirable that the first two PCs account for a substantial proportion of the variance in the original data, thus they can be considered sufficiently representative of the main information included in the data, while the remaining PCs condense irrelevant information or even experimental noise. It is quite common for a PCA to be represented by a score plot, loading plot, or biplot, defined as the joint representation of the rows and columns of a data matrix; points (scores) represent the chemicals and vectors or lines represent the variables (loadings). The lengths of the vectors indicate the information associated with the variable, while the cosine of the angle between the vectors reflects their correlation. In our environmental chemistry studies, PCA has been widely used for screening and ranking purposes in many contexts: (a) tropospheric degradability of volatile organic compounds (VOCs) [11, 17, 106]; (b) mobility in the atmosphere or long-range transport of persistent organic pollutants (POPs) [16, 31, 147]; (c) environmental partitioning tendency of pesticides [7, 32]; (d) POP and PBT screening [10, 24, 34, 147–149].

In addition, this multivariate approach was adopted to study aquatic toxicity of EU-priority listed chemicals on different endpoints [150] and esters [25], the endocrine disrupting activity based on three different endpoints [33] and the abiotic oxidation of phenols in an aqueous environment [9].

12.8.2.2. QSAR Modeling of Ranking Indexes

Tropospheric Persistence/Degradability of Volatile Organic Compounds (VOCs). Studies has been made of the screening/ranking of volatile organic chemicals according to their tendency to degrade in the troposphere. Indeed, as the atmospheric persistence of a chemical is mainly dependent on the degradation rates of its reaction with oxidants, the contemporaneous variation and influence of the rate constants for their degradation by OH, NO_3 radicals, and ozone (kOH, kNO_3, and kO_3), in determining the inherent tendency to degradability, was explored by principal component analysis (PCA).

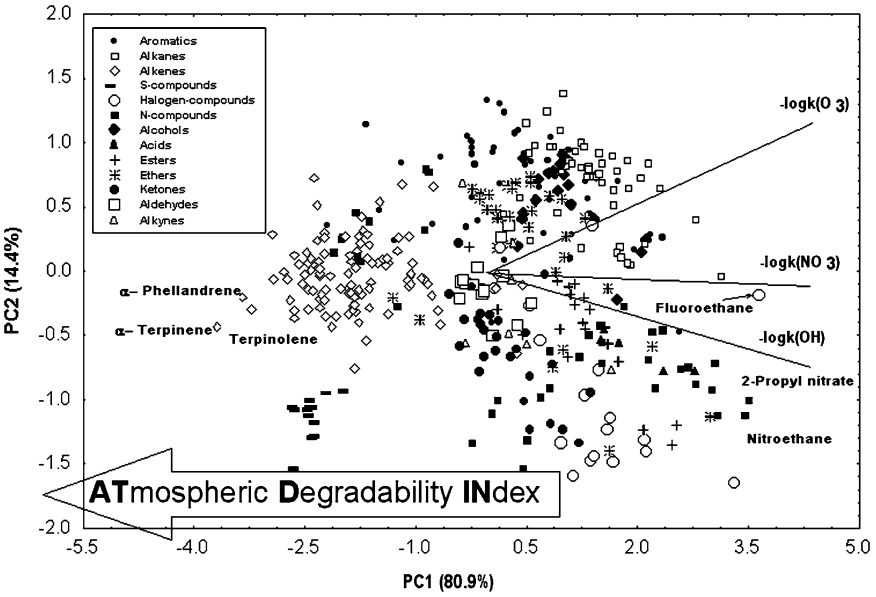

Figure 12-6. Score plot and loading plot of the two principal component analysis of three rate constants (kOH, kNO_3, kO_3) for 399 chemicals (labeled according to chemical classes). ATDIN: ATmospheric Degradability INdex. Cumulative explained variance: 95.3%. Explained Variance of PC1 (ATDINdex)=80.9% (with copyright permission from [17])

In a preliminary study, the experimental data allowed the ranking of a set of 65 heterogeneous VOCs, for which all the degradation rate constants were known; an atmospheric persistence index (ATPIN) had been defined and modelled by theoretical molecular descriptors [11]. Later, the application of our MLR models, developed for each studied degradation rate constant (kNO_3, kO_3, and kOH) [13, 14, 18], allowed a similar PC analysis (Figure 12-6) of a much larger set of 399 chemicals.

This new more informative index (PC1 score of Figure 12-6, 80.9% of explained variance, newly defined ATDIN – atmospheric degradability index), based on a wider set of more structurally heterogeneous chemicals, was also satisfactorily modelled by MLR based on theoretical molecular descriptors and externally validated (Q^2 0.94; Q^2_{EXT} 0.92) (scatter plot in Figure 12-7) [17].

Mobility in Atmosphere and Long-Range Transport of Persistent Organic Pollutants (POPs). The intrinsic tendency of compounds toward global mobility in the atmosphere has been studied, since it is a necessary property for the evaluation of the long-range transport (LRT) of POPs [16, 31]. As the mobility potential of a chemical depends on the various physico-chemical properties of a compound, principal component analysis was used to explore the contemporaneous variation and influence of all the properties selected as being the most relevant to LRT potential (such as vapor pressure, water solubility, boiling point, melting point, temperature of condensation, various partition coefficients among different compartments;

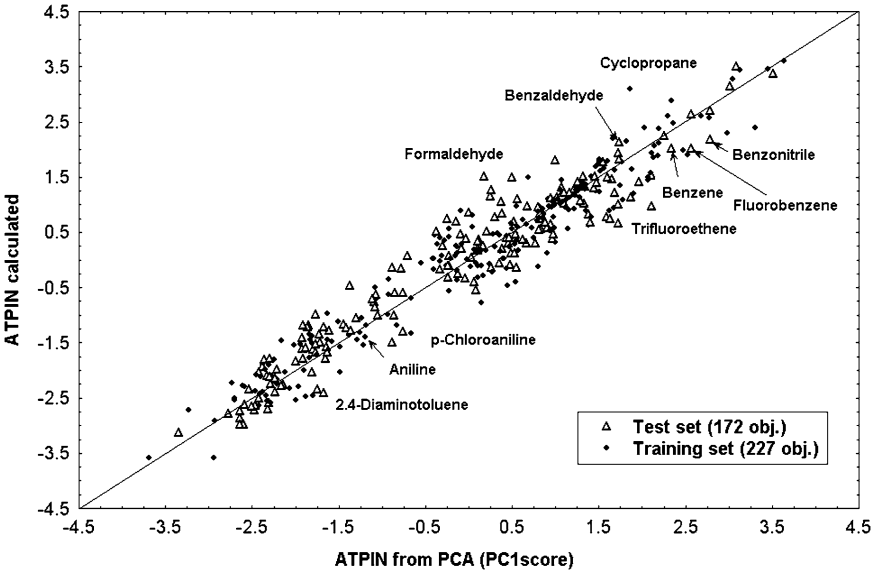

Figure 12-7. Regression line for the externally validated model of ATPIN (ATmospheric Persistence Index: the opposite of ATDIN). The training and test set chemicals are differently highlighted, the outliers and influential chemicals are named (with copyright permission from [17])

for instance, Henry's law constant, octanol/water partition coefficient, soil sorption coefficient, octanol/air partition coefficient).

A simple interpretation of the obtained PC1 is as a scoring function of intrinsic tendency toward global mobility. We have proposed this PC1 scoring as the ranking score for the 82 possible POPs in four *a priori* classes: high, relatively high, relatively low, and low mobility.

These classes have been successfully modelled by the CART method, based on four theoretical molecular descriptors (two Kier and Hall connectivity indexes, molecular weight, and sum of electronegativities) with only 6% of errors in cross-validation. The main aim was to develop a simple and rapid framework to screen, rank, and classify also new organic chemicals according to their intrinsic global mobility tendency, just from the knowledge of their chemical structure.

An analogous approach was previously applied to a subset of 52 POPs to define a long-range transport (LRT) index derived from the PC1 score, on the basis of physico-chemical properties and additionally taking into account atmospheric half-life data [147].

Environmental partitioning tendency of pesticides. The partitioning of pesticides into different environmental compartments depends mainly on the physico-chemical properties of the studied chemical, such as the organic carbon partition coefficient (K_{oc}), the n-octanol/water partition coefficient (K_{ow}), water solubility (S_w), vapor pressure (V_p), and Henry's law constant (H). To rank and classify the 54 studied pesticides, belonging to various chemical categories, according to their distribution

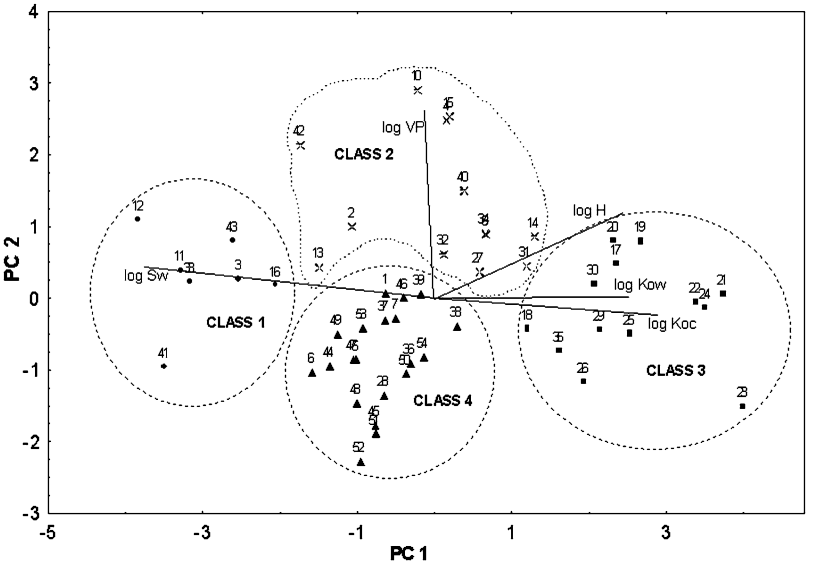

Figure 12-8. Score plot and loading plot of the two first principal components of PCA of five physicochemical properties (K_{oc}, K_{ow}, S_w, V_p, and Henry's law constant) for 54 pesticides. Cumulative explained variance: 94.6%; explained variance of PC1: 70.1% (with copyright permission from [32])

tendency in various media, we applied [32] a combination of two multivariate approaches: principal component analysis (Figure 12-8) for ranking and hierarchical cluster analysis for the definition of the four *a priori* classes, according to their environmental behavior (1. soluble, 2. volatile, 3. sorbed, and 4. non-volatile/medium class) (*circles* in Figure 12-8).

The pesticides were finally assigned to the defined four classes by different classification methods (CART, k-NN, RDA) using theoretical molecular descriptors (for example, the CART tree is reported in Figure 12-9). Two of the selected molecular descriptors are quite easily interpretable, in particular (a) MW encodes information on molecule dimension; it is well known that big molecules have the greatest tendency to bind, by van der Waals forces, to the organic component of the soil, becoming the most sorbed in organic soils but the least soluble in water (Class 3) and (b) the possibility of a chemical to link by hydrogen bonds to water molecules (encoded in the molecular descriptor nHDon) results in the higher solubility of the Class 1 pesticides; furthermore, chemicals with fewer intramolecular hydrogen bonds are the most volatile (Class 2). The last topological descriptor J, that discriminates Class 4 of the medium-behavior pesticides, is not easily interpretable.

A wider, heterogeneous, and quite representative data set of pesticides of different chemical classes (acetanilides, carbamates, dinitroanilines, organochlorides, organophosphates, phenylureas, triazines, triazoles), already studied for their K_{oc}

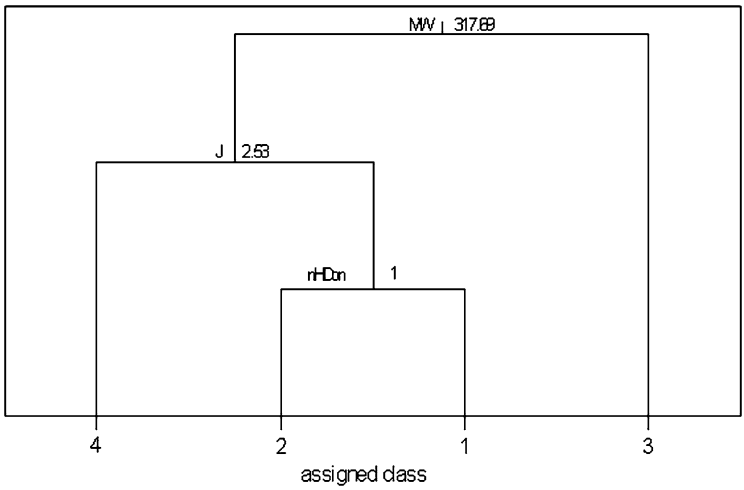

Figure 12-9. Classification tree by classification and regression tree (CART) of mobility classes for 54 pesticides. Error rate (ER) 11.11%; ER in prediction: 18.53%; NoMER: 62.96% (with copyright permission from [32])

modeling [96] has also undergone PC analysis of various environmental partitioning properties (solubility, volatility, partition coefficients, etc.) to study leaching tendency [7]. The resultant macrovariables, PC1 and PC2 scores, called the leaching index (LIN) and volatility index (VIN), have been proposed as cumulative environmental partitioning indexes in different media. These two indexes were modelled by theoretical molecular descriptors with satisfactory predictive power (Q^2 leave-30%-out=0.85 for LIN). Such an approach allows a rapid pre-determination and the screening of the environmental distribution of pesticides, starting only from the molecular structure of the pesticide without any *a priori* knowledge of the physico-chemical properties.

The proposed index LIN was used in a comparative analysis with GUS and LEACH index for highlighting the pesticides most dangerous to the aquatic compartment among those widely used in Uzbekistan, in the Amu-Darya river basin [151].

POPs and PBTs. QSAR approaches, based on molecular structure for the prioritization of chemicals for persistence, particularly persistent organic pollutants (POPs) screening and ranking method for global half-life, have recently been proposed [10, 24, 148, 149].

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, have

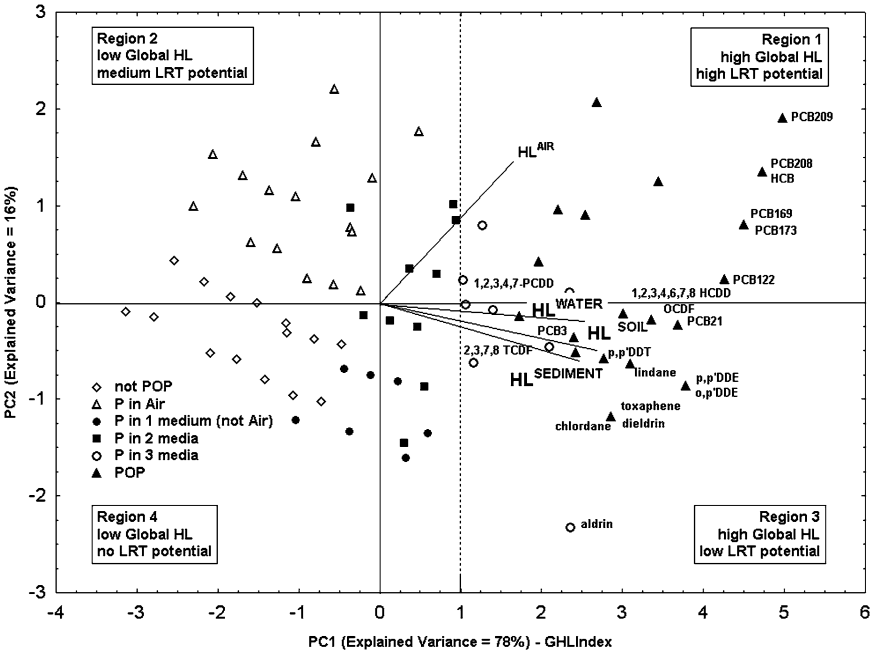

Figure 12-10. Principal component analysis on half-life data for 250 organic compounds in the various compartments (air, water, sediment, and soil) (PC1–PC2: explained variance=94%). P=persistent (with copyright permission from [10])

been combined in a multivariate approach by principal component analysis. This PCA distributes the studied compounds according to their cumulative, or global, half-life and relative persistence in different media, to obtain a ranking of the studied organic pollutants according to their relative overall half-life.

The biplot relative to the first and second components is reported in Figure 12-10, 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 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 (lines) are oriented in the same direction along the first principal component, thus 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; chemicals with high half-life values in all the media (highlighted in the PCA graph) 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 (labeled in Figure 12-10) or persistent in only one medium; chemicals persistent in 2 or 3 media are located in the intermediate zone of Figure12-10.

PC2, although less informative (E.V. 16%), is also interesting; it separates the compounds more persistent in air (upper parts in Figure 12-10, regions 1 and 2), i.e., those with higher LRT potential from those more persistent in water, soil, and sediment (lower parts in Figure 12-10, regions 3 and 4).

A deeper analysis of the distribution of the studied chemicals gives some interesting results and confirms experimental evidence: to the right, among the very persistent chemicals in all the compartments (*full triangles* in Figure 12-10), we find most of the compounds recognized as POPs by the Stockholm Convention [152]. Highly chlorinated PCBs 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. The less chlorinated PCBs (PCB-3 and PCB 21) fall in the zone of very persistent chemicals, but not in the upper part of Region 1, due to their lower persistence in air compared with highly chlorinated congeners. *p*,*p*′-DDT, *p*,*p*′-DDE, *o*,*p*′-DDE, highly chlorinated dioxins and dioxin-like compounds, as well as pesticides toxaphene, lindane, chlordane, dieldrin, and aldrin fall in Region 3 (highly persistent chemicals mainly in compartments different from air).

A global half-life index (GHLI) obtained from existing knowledge of generalized chemical persistence over a wide scenario of 250 chemicals, which reliability was verified through comparison with multimedia model results and empirical evidence, was proposed from this PC analysis [10]. This global index, the PC1 score, was then modelled as a cumulative endpoint using a QSAR approach based on theoretical molecular descriptors; a simple and robust regression model externally validated for its predictive ability [6, 84] has been derived. The original set of available data was first randomly split into training and prediction sets; 50% of the compounds were put into the prediction set (125 compounds) while the other 50% was used to build the QSPR model by MLR. Given below (Eq. 12-5) is the best QSPR model, selected by statistical approaches and its statistical parameters (Figure 12-11 shows the plot of GHLI values from PCA vs. predicted GHLI values):

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

$$\begin{aligned}&n_{\text{training}} = 125\ R^2 = 0.85\ Q^2_{\text{LOO}} = 0.83\ Q^2_{\text{BOOT}} = 0.83 \\ &\text{RMSE} = 0.76\ \text{RMSEcv} = 0.70;\end{aligned}$$

$$n_{\text{prediction}} = 125\ R^2_{\text{EXT}} = 0.79\ \text{RMSEp} = 0.78 \qquad (12\text{-}5)$$

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

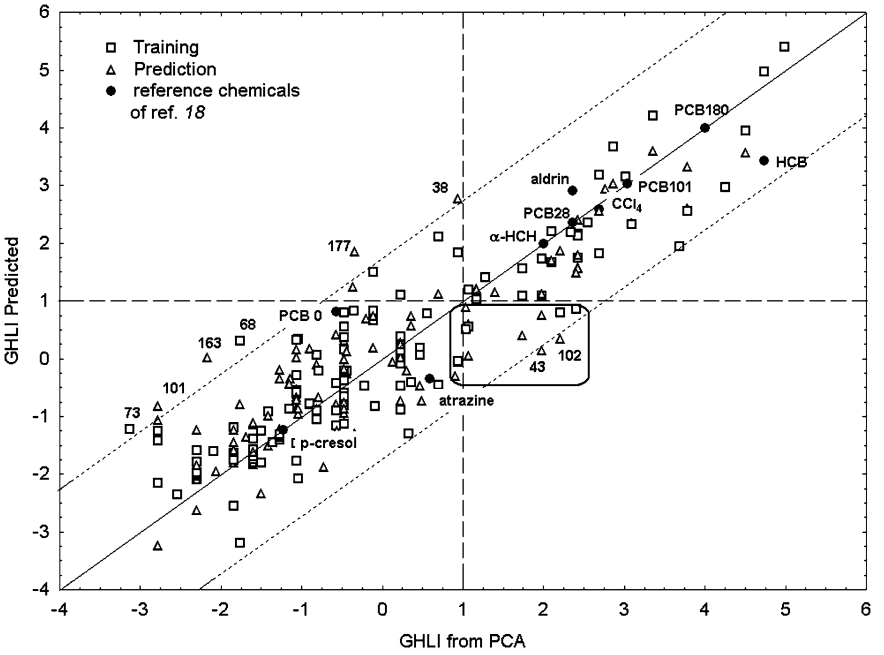

Figure 12-11. Scatter plot of the GHLI values calculated by PCA vs. predicted values by the model. The GHLI values for the training and prediction set chemicals are labeled differently. The *diagonal dotted lines* indicate the 2.5σ interval and response outliers are numbered. *Vertical and horizontal dotted lines* identify the cut-off value of GHLI=1 for high-persistent chemicals (with copyright permission from [10])

The only really dangerous zone in the proposed model is the underestimation zone (*circled* in Figure 12-11).

The application of this model, 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 compounds not yet synthesized, which could represent safer alternative and replacement solutions for recognized POPs. No method other than QSAR is applicable to detect the potential persistence of new compounds.

Similarly, highly predictive classification models, based on k-NN, CART, and CP-ANN, have been developed and can be usefully applied for POP pre-screening. The *a priori* classes have been defined by applying hierarchical cluster analysis to the half-life data [34].

An approach analogous to GHLI has been successfully applied to the PCA-combination of data obtained from the above cumulative half-lives for persistence GHLI, bioconcentration data of fish, and acute toxicity data of *P. promelas* in order to propose, and then model by QSPR approach, a combined index of PBT behavior

[24, 148, 149]. A simple model, based on easy calculable molecular descriptors and with high external predictivity (Q^2_{EXT}>0.8), has been developed and will be published. This PBT index can be applied also to chemicals without any experimental data and even to not yet synthesized compounds.

These QSAR-based tools, validated for their predictivity on new chemicals, could help in highlighting the POP and PBT behavior also of chemicals not yet synthesized, and could be usefully applied for the new European Regulation REACH, which requires most demanding authorization steps for PBTs and the design of safer alternatives. The results of our predictions were comparable with those from the US-EPA PBT profiler (http://www.epa.gov/pbt/tools/toolbox.htm).

12.9. CONCLUSIONS

A statistical approach to QSAR modeling, based on heterogeneous theoretical molecular descriptors and chemometric methods and developed with the fundamental aim of predictive applications, has been introduced and discussed in this review. Several applications to environmentally relevant topics related to organic pollutants, performed by the Insubria QSAR Research Unit in last 15 years, have been presented. Different endpoints related to physico-chemical properties, persistence, bioaccumulation, and toxicity have been modelled, not only singularly, but also as combined endpoints, obtained by multivariate analysis; the approach is innovative and highly useful for ranking and prioritizing purposes. All the proposed models characteristically check the predictive performance and applicability domain of the chemicals, even new chemicals that never participated in the model development. The fulfillment of the "OECD principles for QSAR validation" is a guarantee for the reliability of the predicted data obtained by our models and their possible applicability in the context of REACH.

ACKNOWLEDGEMENT

Many thanks to my collaborators who participated in the research, reviewed here, carried out over the past 15 years, particularly Ester Papa and Pamela Pilutti. Thanks are also due to Roberto Todeschini who was my teacher of chemometric QSAR.

REFERENCES

1. Hansch C, Fujita T (1964) *p*-s-p analysis. A method for the correlation of biological activity and chemical structure. J Am Chem Soc 86:1616–1626
2. Hansch C, Leo A (1995) Exploring QSAR: Fundamentals and applications in chemistry and biology. American Chemical Society, Washington, DC 490–496
3. Schultz TW, Cronin MTD, Netzeva TI et al. (2002) Structure–toxicity relationships for aliphatic chemicals evaluated with Tetrahymena pyriformis. Chem Res Toxicol 15:1602–1609
4. Veith GD, Mekenyan O (1993) A QSAR approach for estimating the aquatic toxicity of soft electrophiles (QSAR for soft electrophiles). Quant Struct -Act Relat 12:349–356
5. Benigni R (2005) Structure–activity relationship studies of chemical mutagens and carcinogens: Mechanistic investigations and prediction approaches. Chem Rev 105:1767–1800

6. Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 22:69–77
7. Gramatica P, Di Guardo A (2002) Screening of pesticides for environmental partitioning tendency. Chemosphere 47:947–956
8. Gramatica P, Papa E (2003) QSAR modeling of bioconcentration factor by theoretical molecular descriptors. QSAR Comb Sci 22:374–385
9. Gramatica P, Papa E (2005) An update of the BCF QSAR model based on theoretical molecular descriptors. QSAR Comb Sci 24:953–960
10. Gramatica P, Papa E (2007) Screening and ranking of POPs for global half-life: QSAR approaches for prioritization based on molecular structure. Environ Sci Technol 41:2833–2839
11. Gramatica P, Pilutti P, Papa E (2002) Ranking of volatile organic compounds for tropospheric degradability by oxidants: A QSPR approach. SAR QSAR Environ Res 13:743–753
12. Gramatica P, Consonni V, Pavan M (2003) Prediction of aromatic amines mutagenicity from theoretical molecular descriptors. SAR QSAR Environ Res 14:237–250
13. Gramatica P, Pilutti P, Papa E (2003) Predicting the NO3 radical tropospheric degradability of organic pollutants by theoretical molecular descriptors. Atmos Environ 37:3115–3124
14. Gramatica P, Pilutti P, Papa E (2003) QSAR prediction of ozone tropospheric degradation. QSAR Comb Sci 22:364–373
15. Gramatica P, Battaini F, Papa E (2004) QSAR prediction of physico-chemical properties of esters. Fresenius Environ Bull 13:1258–1262
16. Gramatica P, Papa E, Pozzi S (2004) Prediction of POP environmental persistence and long range transport by QSAR and chemometric approaches. Fresenius Environ Bull 13:1204–1209
17. Gramatica P, Pilutti P, Papa E (2004) A tool for the assessment of VOC degradability by tropospheric oxidants starting from chemical structure. Atmos Environ 38:6167–6175
18. Gramatica P, Pilutti P, Papa E (2004) Validated QSAR prediction of OH tropospheric degradation of VOCs: Splitting into training-test sets and consensus modeling. J Chem Inf Comput Sci 44:1794–1802
19. Gramatica P, Giani E, Papa E (2007) Statistical external validation and consensus modeling: A QSPR case study for K_{oc} prediction. J Mol Graph Model 25:755–766
20. Gramatica P, Pilutti P, Papa E (2007) Approaches for externally validated QSAR modelling of nitrated polycyclic aromatic hydrocarbon mutagenicity. SAR QSAR Environ Res 18:169–178
21. 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. Chem Res Toxicol 19:1540–1548
22. Liu H, Papa E, Gramatica P (2008) Evaluation and QSAR modeling on multiple endpoints of estrogen activity based on different bioassays. Chemosphere 70:1889–1897
23. Papa E, Gramatica P (2008) Externally validated QSPR modelling of VOC tropospheric oxidation by NO_3 radicals. SAR QSAR Environ Res 19:655–668
24. Papa E, Gramatica P (2009) QSPR as a support to the EU REACH legislation: PBTs identification by molecular structure. Environ Sci Technol (in press)
25. Papa E, Battaini F, Gramatica P (2005) Ranking of aquatic toxicity of esters modelled by QSAR. Chemosphere 58:559–570
26. 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). J Chem Inf Model 45:1256–1266
27. Papa E, Dearden JC, Gramatica P (2007) Linear QSAR regression models for the prediction of bioconcentration factors by physicochemical properties and structural theoretical molecular descriptors. Chemosphere 67:351–358

28. Papa E, Kovarich S, Gramatica P (2009) Development, validation and inspection of the applicability domain of QSPR models for physico-chemical properties of polybrominated diphenyl ethers. QSAR Comb Sci. doi: 10.1002/qsar.200860183
29. Todeschini R, Maiocchi A, Consonni V (1999) The *K* correlation index: Theory development and its application in chemometrics. Chemom Int Lab Syst 46:13–29
30. Hawkins DM (2004) The problem of overfitting. J Chem Inf Comput Sci 44:1–12
31. Gramatica P, Pozzi S, Consonni V et al. (2002) Classification of environmental pollutants for global mobility potential. SAR QSAR Environ Res 13:205–217
32. Gramatica P, Papa E, Battaini F (2004) Ranking and classification of non-ionic organic pesticides for environmental distribution: A QSAR approach. Int J Environ Anal Chem 84:65–74
33. Liu H, Papa E, Walker JD et al. (2007) *In silico* screening of estrogen-like chemicals based on different nonlinear classification models. J Mol Graph Model 26:135–144
34. Papa E, Gramatica P (2008) Screening of persistent organic pollutants by QSPR classification models: A comparative study. J Mol Graph Model 27:59–65
35. Papa E, Pilutti P, Gramatica P (2008) Prediction of PAH mutagenicity in human cells by QSAR classification. SAR QSAR Environ Res 19:115–127
36. Breiman L, Friedman JH, Olshen RA et al. (1998) Classification and regression trees. Chapman & Hall/CRC, Boca Raton, FL
37. Frank JE, Friedman JH (1989) Classification: Oldtimers and newcomers. J Chemom 3:463–475
38. Sharaf MA, Illman DL, Kowalski BR (1986) Chemometrics. Wiley, New York
39. Gasteiger J, Zupan J (1993) Neural networks in chemistry. Angew Chem Int Ed Engl 32:503–527
40. Hecht-Nielsen R (1988) Applications of counter-propagation networks. Neural Netw 1:131–139
41. Zupan J, Novic M, Ruisanchez I (1997) Kohonen and counter-propagation artificial neural networks in analytical chemistry. Chemom Int Lab Syst 38:1–23
42. Livingstone DJ (2000) The characterization of chemical structures using molecular properties. A survey. J Chem Inf Comput Sci 40:195–209
43. Stuper AJ, Jurs PC (1976) ADAPT: A computer system for automated data analysis using pattern recognition techniques. J Chem Inf Comput Sci 16:99–105
44. Mekenyan O, Bonchev D (1986) OASIS method for predicting biological activity of chemical compounds. Acta Pharm Jugosl 36:225–237
45. Katritzky AR, Lobanov VS (1994) CODESSA. Ver. 5.3, University of Florida, Gainesville
46. Todeschini R, Consonni V, Mauri A et al. (2006) DRAGON – software for the calculation of molecular descriptors. Ver. 5.4 for Windows, Talete srl, Milan, Italy
47. MollConnZ (2003) Ver. 4.05. Hall Ass. Consult., Quincy, MA
48. Devillers J, Balaban AT (1999) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach Science Publishers, Amsterdam
49. Karelson M (2000) Molecular descriptors in QSAR/QSPR. Wiley-InterScience, New York
50. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Weinheim, Germany
51. Todeschini R, Lasagni M (1994) New molecular descriptors for 2D and 3D structures. J Chemom 8:263–272
52. Todeschini R, Gramatica P (1997) 3D-modelling and prediction by WHIM descriptors.5. Theory development and chemical meaning of WHIM descriptors. Quant Struct Act Relat 16:113–119
53. Todeschini R, Gramatica P (1997) 3D-modelling and prediction by WHIM descriptors.6. Application of WHIM descriptors in QSAR studies. Quant Struct Act Relat 16:120–125
54. Consonni V, Todeschini R, Pavan M et al. (2002) Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. J Chem Inf Comput Sci 42:682–692.

55. Consonni V, Todeschini R, Pavan M et al. (2002) Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. J Chem Inf Comput Sci 42:693–705
56. Xu L, Zhang WJ (2001) Comparison of different methods for variable selection. Anal Chim Acta 446:477–483
57. Davis L (1991) Handbook of genetic algorithms. Van Nostrand Reinhold, New York
58. Hibbert DB (1993) Genetic algorithms in chemistry. Chemom Int Lab Syst 19:277–293
59. Wehrens R, Buydens LMC (1998) Evolutionary optimisation: A tutorial. TRAC 17:193–203
60. Leardi R, Boggia R, Terrile M (1992) Genetic algorithms as a strategy for feature-selection. J Chemom 6:267–281
61. Rogers D, Hopfinger AJ (1994) Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships. J Chem Inf Comput Sci 34:854–866
62. Devillers J (1996) Genetic algorithms in computer-aided molecular design. In: Devillers J (ed) Genetic algorithms in molecular modeling. Academic Press Ltd, London
63. Leardi R (1994) Application of a genetic algorithm to feature-selection under full validation conditions and to outlier detection. J Chemom 8:65–79
64. Kubinyi H (1994) Variable selection in QSAR studies. 1. An evolutionary algorithm. Quant Struct Act Relat 13:285–294
65. Kubinyi H (1994) Variable selection in QSAR studies. 2. A highly efficient combination of systematic search and evolution. Quant Struct Act Relat 13:393–401
66. Todeschini R, Consonni V, Pavan M et al. (2002) MOBY DIGS. Ver. 1.2 for Windows, Talete srl, Milan, Italy
67. Kubinyi H (1996) Evolutionary variable selection in regression and PLS analyses. J Chemom 10:119–133
68. Guha R, Serra JR, Jurs PC (2004) Generation of QSAR sets with a self-organizing map. J Mol Graph Model 23:1–14
69. Netzeva TI, Worth AP, Aldenberg T et al. (2005) Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships – the report and recommendations of ECVAM Workshop 52. ATLA 33:155–173
70. Tunkel J, Mayo K, Austin C et al. (2005) Practical considerations on the use of predictive models for regulatory purposes. Environ Sci Technol 39:2188–2199
71. Atkinson AC (1985) Plots, transformations and regression. Clarendon Press, Oxford
72. Hulzebos EM, Posthumus R (2003) (Q)SARs: Gatekeepers against risk on chemicals? SAR QSAR Environ Res 14:285–316
73. Jouan-Rimbaud D, Massart DL, deNoord OE (1996) Random correlation in variable selection for multivariate calibration with a genetic algorithm. Chemom Int Lab Syst 35:213–220
74. Efron B, Gong G (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. Am Stat 37:36–48
75. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, London
76. Shao J (1993) Linear-model selection by cross-validation. J Am Stat Assoc 88:486–494
77. Golbraikh A, Tropsha A (2002) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. J Comput Aid Mol Des 16:357–369
78. Zefirov NS, Palyulin VA (2001) QSAR for boiling points of "small" sulfides. Are the "high-quality structure-property-activity regressions" the real high quality QSAR models? J Chem Inf Comput Sci 41:1022–1027
79. Golbraikh A, Shen M, Xiao ZY et al. (2003) Rational selection of training and test sets for the development of validated QSAR models. J Comput Aid Mol Des 17:241–253
80. Leonard JT, Roy K (2006) On selection of training and test sets for the development of predictive QSAR models. QSAR Comb Sci 25:235–251

81. Sjostrom M, Eriksson L (1995) Chemometric methods in molecular design. van de Waterbeend H (ed) Vol. 2. VCH, New York, p 63
82. Marengo E, Todeschini R (1992) A new algorithm for optimal, distance-based experimental-design. Chemom Int Lab Syst 16:37–44
83. Gramatica P (2004) Evaluation of different statistical approaches to the validation of quantitative structure-activity relationships. http://ecb.jrc.it/DOCUMENTS/QSAR/Report_on_QSAR_validation_methods.pdf Accessed April 2008
84. Gramatica P (2007) Principles of QSAR models validation: Internal and external. QSAR Comb Sci 26:694–701
85. Kahn I, Fara D, Karelson M et al. (2005) QSPR treatment of the soil sorption coefficients of organic pollutants. J Chem Inf Model 45:94–105
86. Todeschini R, Gramatica P, Provenzani R et al. (1995) Weighted holistic invariant molecular descriptors. 2. Theory development and applications on modeling physicochemical properties of polyaromatic hydrocarbons. Chemom Int Lab Syst 27:221–229
87. Chiorboli C, Gramatica P, Piazza R et al. (1997) 3D-modelling and prediction by WHIM descriptors. Part 7. Physico-chemical properties of haloaromatics: Comparison between WHIM and topological descriptors. SAR QSAR Environ Res 7:133–150
88. Gramatica P, Navas N, Todeschini R (1998) 3D-modelling and prediction by WHIM descriptors. Part 9. Chromatographic relative retention time and physico-chemical properties of polychlorinated biphenyls (PCBs). Chemom Int Lab Syst 40:53–63
89. Todeschini R, Vighi M, Finizio A et al. (1997) 3D-modelling and prediction by WHIM descriptors. Part 8. Toxicity and physico-chemical properties of environmental priority chemicals by 2D-TI and 3D-WHIM descriptors. SAR QSAR Environ Res 7:173–193
90. Todeschini R, Gramatica P (1997) The WHIM theory: New 3D-molecular descriptors for QSAR in environmental modelling. SAR QSAR Environ Res 7:89–115
91. Todeschini R, Gramatica P (1998) 3D-QSAR in drug design. Kubiny H, Folkers G, Martin YC (eds) vol. 2. KLUWER/ESCOM, Dordrecht, p 355
92. Gramatica P (2006) WHIM descriptors of shape. QSAR Comb Sci 25:327–332
93. Patel H, Cronin MTD (2001) A novel index for the description of molecular linearity. J Chem Inf Comput Sci 41:1228–1236
94. Nikolova N, Jaworska J (2003) Approaches to measure chemical similarity – a review. QSAR Comb Sci 22:1006–1026
95. Gramatica P, Navas N, Todeschini R (1999) Classification of organic solvents and modelling of their physico-chemical properties by chemometric methods using different sets of molecular descriptors. TRAC 18:461–471
96. Gramatica P, Corradi M, Consonni V (2000) Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors. Chemosphere 41:763–777
97. Sabljic A, Gusten H, Verhaar H et al. (1995) QSAR modeling of soil sorption – improvements and systematics of Log K_{oc} vs Log K_{ow} correlations. Chemosphere 31:4489–4514
98. Gawlik BM, Sotiriou N, Feicht EA et al. (1997) Alternatives for the determination of the soil adsorption coefficient, K_{oc}, of non-ionicorganic compounds – a review. Chemosphere 34: 2525–2551
99. Doucette WJ (2003) Quantitative structure–activity relationships for predicting soil-sediment sorption coefficients for organic chemicals. Environ Toxicol Chem 22:1771–1788
100. Tao S, Piao HS, Dawson R et al. (1999) Estimation of organic carbon normalized sorption coefficient (K_{oc}) for soils using the fragment constant method. Environ Sci Technol 33:2719–2725
101. Huuskonen J (2003) Prediction of soil sorption coefficient of a diverse set of organic chemicals from molecular structure. J Chem Inf Comput Sci 43:1457–1462
102. Huuskonen J (2003) Prediction of soil sorption coefficient of organic pesticides from the atom-type electrotopological state indices. Environ Toxicol Chem 22:816–820

103. Andersson PL, Maran U, Fara D et al. (2002) General and class specific models for prediction of soil sorption using various physicochemical descriptors. J Chem Inf Comput Sci 42:1450–1459
104. Delgrado EJ, Alderete JB, Gonzalo AJ (2003) A simple QSPR model for predicting soil sorption coefficients of polar and nonpolar organic compounds from molecular formula. J Chem Inf Comput Sci 43:1928–1932
105. EPI Suite. Ver. 3.12 (2000) Environmental Protection Agency, USA http://www.epa.gov/opptintr/exposure/docs/EPISuitedl.htm. Accessed 9 February 2007
106. Gramatica P, Consonni V, Todeschini R (1999) QSAR study on the tropospheric degradation of organic compounds. Chemosphere 38:1371–1378
107. Atkinson R (1987) A structure-activity relationship for the estimation of rate constants for the gas-phase reactions of OH radicals with organic compounds. Int J Chem Kinet 19:799–828
108. Sabljic A, Gusten H (1990) Predicting the nighttime NO_3 radical reactivity in the troposphere. Atmos Environ A-General Topics 24:73–78
109. Müller M, Klein W (1991) Estimating atmospheric degradation processes by SARS. Sci Total Environ 109:261–273
110. Medven Z, Gusten H, Sabljic A (1996) Comparative QSAR study on hydroxyl radical reactivity with unsaturated hydrocarbons: PLS versus MLR. J Chemom 10:135–147
111. Klamt A (1996) Estimation of gas-phase hydroxyl radical rate constants of oxygenated compounds based on molecular orbital calculations. Chemosphere 32:717–726
112. Bakken G, Jurs PC (1999) Prediction of hydroxyl radical rate constants from molecular structure. J Chem Inf Comput Sci 39:1064–1075
113. Güsten H (1999) Predicting the abiotic degradability of organic pollutants in the troposphere. Chemosphere 38:1361–1370
114. Pompe M, Veber M (2001) Prediction of rate constants for the reaction of O-3 with different organic compounds. Atmos Environ 35:3781–3788
115. Meylan WM, Howard PH (2003) A review of quantitative structure–activity relationship methods for the prediction of atmospheric oxidation of organic chemicals. Environ Toxicol Chem 22: 1724–1732
116. Pompe M, Veber M, Randic M et al. (2004) Using variable and fixed topological indices for the prediction of reaction rate constants of volatile unsaturated hydrocarbons with OH radicals. Molecules 9:1160–1176
117. Oberg T (2005) A QSAR for the hydroxyl radical reaction rate constant: Validation, domain of application, and prediction. Atmos Environ 39:2189–2200
118. AOPWIN. Ver. 1.90 (2000) Environmental Protection Agency, USA
119. Devillers J, Bintein S, Domine D (1996) Comparison of BCF models based on log P. Chemosphere 33:1047–1065
120. Meylan WM, Howard PH, Boethling RS et al. (1999) Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient. Environ Toxicol Chem 18:664–672
121. Lu XX, Tao S, Hu HY et al. (2000) Estimation of bioconcentration factors of nonionic organic compounds in fish by molecular connectivity indices and polarity correction factors. Chemosphere 41:1675–1688
122. Dearden JC, Shinnawei NM (2004) Improved prediction of fish bioconcentration factor of hydrophobic chemicals. SAR QSAR Environ Res 15:449–455
123. Dimitrov S, Dimitrova N, Parkerton T et al. (2005) Base-line model for identifying the bioaccumulation potential of chemicals. SAR QSAR Environ Res 16:531–554
124. Zhao CY, Boriani E, Chana A et al. (2008) A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). Chemosphere 73:1701–1707

125. Todeschini R, Vighi M, Provenzani R et al. (1996) Modeling and prediction by using WHIM descriptors in QSAR studies: Toxicity of heterogeneous chemicals on Daphnia magna. Chemosphere 32:1527–1545
126. Gramatica P, Vighi M, Consolaro F et al. (2001) QSAR approach for the selection of congeneric compounds with a similar toxicological mode of action. Chemosphere 42:873–883
127. Benigni R, Giuliani A, Franke R et al. (2000) Quantitative structure–activity relationships of mutagenic and carcinogenic aromatic amines. Chem Rev 100:3697–3714
128. Gramatica P, Papa E, Marrocchi A et al. (2007) Quantitative structure–activity relationship modeling of polycyclic aromatic hydrocarbon mutagenicity by classification methods based on holistic theoretical molecular descriptors. Ecotoxicol Environ Saf 66:353–361
129. Shi LM, Fang H, Tong W et al. (2001) QSAR models using a large diverse set of estrogens. J Chem Inf Comput Sci 41:186–195
130. Hong H, Tong W, 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. Environ Health Perspect 110:29–36
131. Tong W, Fang H, Hong H et al. (2003) Regulatory application of SAR/QSAR for priority setting of endocrine disruptors: A perspective. Pure Appl Chem 75:2375–2388
132. Tong W, Welsh WJ, Shi LM et al. (2003) Structure–activity relationship approaches and applications. Environ Toxicol Chem 22:1680–1695
133. Fang H, Tong W, Sheehan DM (2003) QSAR models in receptor-mediated effects: the nuclear receptor superfamily. J Mol Struct (THEOCHEM) 622:113–125
134. Saliner AG, Amat L, Carbo-Dorca R et al. (2003) Molecular quantum similarity analysis of estrogenic activity. J Chem Inf Comput Sci 43:1166–1176
135. Saliner AG, Netzeva TI, Worth AP (2006) Prediction of estrogenicity: Validation of a classification model. SAR QSAR Environ Res 17:195–223
136. Coleman KP, Toscano WA, Wiese TE (2003) QSAR models of the in vitro estrogen activity of bisphenol A analogs. QSAR Comb Sci 22:78–88
137. Roncaglioni A, Novic M, Vracko M et al. (2004) Classification of potential endocrine disrupters on the basis of molecular structure using a nonlinear modeling method. J Chem Inf Comput Sci 44:300–309
138. Asikainen A, Ruuskanen J, Tuppurainen K (2003) Spectroscopic QSAR methods and self-organizing molecular field analysis for relating molecular structure and estrogenic activity. J Chem Inf Comput Sci 43:1974–1981
139. Asikainen A, Kolehmainen M, Ruuskanen J et al. (2006) Structure-based classification of active and inactive estrogenic compounds by decision tree, LVQ and kNN methods. Chemosphere 62:658–673
140. Devillers D, Marchand-Geneste N, Carpy A et al. (2006) SAR and QSAR modeling of endocrine disruptors. SAR QSAR Environ Res 17:393–412
141. Roncaglioni A, Benfenati E (2008) In silico-aided prediction of biological properties of chemicals: Oestrogen receptor-mediated effects. Chem Soc Rev 37:441–450
142. Roncaglioni A, Piclin N, Pintore M et al. (2008) Binary classification models for endocrine disrupter effects mediated through the estrogen receptor. SAR QSAR Environ Res 19:697–733
143. Kuiper GG, Lemmen JG, Carlsson B et al. (1998) Interaction of estrogenic chemicals and phytoestrogens with estrogen receptor ß. Endocrinology 139:4252–4263
144. Liu H, Yao X, Gramatica P (2009) The applications of machine learning algorithms in the modeling of estrogen-like chemicals. Comb Chem High Throughput Screen (special issue on "Machine learning for virtual screening) 12(5) (in press)
145. Joliffe IT (1986) Principal component analysis. Springer-Verlag, New York 490–496
146. Jackson JE (1991) A user's guide to principal components. John Wiley & Sons, New York

147. Gramatica P, Consolaro F, Pozzi S (2001) QSAR approach to POPs screening for atmospheric persistence. Chemosphere 43:655–664
148. Papa E, Gramatica P (2005) PBTs screening by multivariate analysis and QSAR modeling platform presented at 10th EuCheMS-DLE Intern. Conf., Rimini, Italy
149. Papa E, Gramatica P (2006) Structurally-based PBT profiler: The PBT index from molecular structure. Presented at 16th Annual Meeting SETAC-Europe, The Hague, Holland.
150. Vighi M, Gramatica P, Consolaro F et al. (2001) QSAR and chemometric approaches for setting water quality objectives for dangerous chemicals. Ecotoxicol Environ Saf 49:206–220
151. Papa E, Castiglioni S, Gramatica P et al. (2004) Screening the leaching tendency of pesticides applied in the Amu Darya Basin (Uzbekistan). Water Res 38:3485–3494
152. UNEP (2001) Stockholm convention on persistent organic pollutants. United Nations Environmental Program, Geneva, Switzerland. http://www.pops.int. Accessed 9 February 2007