*Environmental Health Perspectives*
*Vol. 61, pp. 257-268, 1985*

# Toxicity Modeling and Prediction with Pattern Recognition

## by Svante Wold*, William J. Dunn† and Sven Hellberg*

**Empirical models can be constructed relating the change in toxicity to the change in chemical structure for series of similar compounds or mixtures. The first step is to translate the variation in structure to quantitative numbers. This gives a data table, a data matrix denoted by $X$, which then is analyzed. The same type of the models can be used to relate the variation of *in vivo* data to the variation of a battery of *in vitro* tests.**

**A single data analytical model cannot be applied to a set of compounds of diverse chemical structure. For such data sets, separate models must be developed for each subgroup of compounds. The data analytical problem then partly is one of classification, pattern recognition (PARC). The assumption of structural and biological similarity within each subset of modeled compounds is then essential for empirical models to apply.**

**PARC is often used to classify compounds as active (toxic) or inactive. The data structure is then often asymmetric which puts special demands on the data analysis, making the traditional PARC methods inapplicable.**

**Depending on the desired information from the data analysis and on the type of available data, four levels of PARC can be distinguished: (I) the data $X$ are used to develop rules for classifying future compounds into one of the classes represented in X; (II) same as I, but the possibility of future compounds belonging to "unknown" classes not represented in X is taken into account; (III) same as II, plus the quantitative prediction of one activity variable (here toxicity) in some classes; (IV) same as III, but several quantitative activity (toxicity) variables are predicted.**

## Introduction

Pattern recognition (henceforth briefly PARC) and related multivariate data analytical methods have recently been applied to the problem of predicting the biological activity of chemical compounds from their chemical structure. Structure–toxicity models can be seen as a special, albeit important, case. Unfortunately, several of the published applications are merely examples of spurious correlations because simple chemical and statistical rules have not been obeyed (*1,2*).

Recently an understanding has emerged of how and when empirical mathematical models can be used in the modeling and prediction of the biological activity (*1–3*). These models usually involve several variables describing the chemical structure of the compounds. Moreover, although it is sometimes stated otherwise (*4–6*), the models are usually only locally applicable within series of congeneric, structurally similar, compounds. Hence, for sets of structurally diverse compounds, one model must be formulated for each biologically and chemically similar subset. Consequently, the data analytical problem then is partly one of classification, PARC.

We shall here discuss the application of PARC in the modeling and prediction of toxicity. Some emphasis will be given to pitfalls and commonly made mistakes. This is because the area is of such importance. Mistakes and overstated results may directly affect human safety.

## Form of Pattern Recognition Problems and Data Sets

PARC applies when multiple data (variables) are used to characterize a set of "objects," here compounds or mixtures. The objects are divided into two sets, the training set and the test set. The training set is often further divided into several subsets, classes of objects with inherent similarity. Thus the available data have the form shown in Figure 1.

We here emphasize that data do not appear like this automatically. As the first phase, a given problem must be translated to one that can be handled by PARC. Classes, variables, and objects must be specified or selected according to available knowledge. The subse-

*Chemistry Institute, Umeå University, S-901 87 Umeå, Sweden.
†Medicinal Chemistry and Pharmacognocy, College of Pharmacy, University of Illinois at Chicago Health Center, P.O. Box 6998, Chicago, IL 60680.

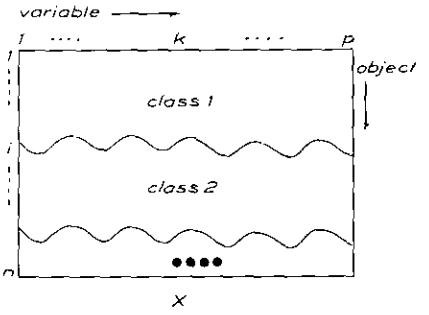


FIGURE 1. The data $X$ with elements $x_{ik}$ (compound $i$ and variable $k$) for PARC levels I and II.

quent data analysis is the simplest part of the problem, but we must be aware of the possibilities to analyze various types of data to be able to translate our problem efficiently in the first phase.

The scope of PARC data analysis is partly one of classification. The training set data are used to develop mathematical rules which then can be used to assign new objects to one of the classes on the basis of the same type of data measured on these new objects.

In case the training set consists of just a single class, the classification of new objects corresponds to finding out if they are similar to the training class or not.

In the analysis of structure–activity relationships, the simplest level of PARC outlined above—henceforth called PARC level I—is rarely sufficient. One cannot be certain that all compounds, neither in the training set nor in the test set, indeed belong to one of the given classes. The possibility of new unexpected classes must be taken into account, i. e., the possibility that a compound is an "outlier" to the given classes must be considered in the data analysis. Such analysis is PARC level II.

In many applications one desires also to get a quantitative model of the relation between the potency of the compounds and their structure. If this potency is measured by a single variable—here denoted $y$—we have PARC level III. The structure data $X$ are then used to (a) classify compounds into one of the given classes or as outliers and (b) to predict the value of the activity variable $y$. When the activity is quantitatively measured by several variables, giving the activity matrix $Y_g$ for class $g$ in the training set, we have, finally, PARC level IV. The scope is the same as level III except that all the multiple $y$-variables enter the models simultaneously. Thus we then have the data shown in Figure 2.

The natural level for the analysis of structure–activity relationships is, we believe, PARC level IV. This is because a single structural variable rarely is able to capture the complex effects of modifying chemical structure: hence, multivariate data $X$. Analogously, a single activity measurement can rarely describe the state of a biological system and how this state is affected by chemical compounds. Indeed, in pharmacological and toxicological investigations, multiple activity data are usually measured, but for some strange reasons, rarely analyzed properly as a set of multivariate $Y$ data. This

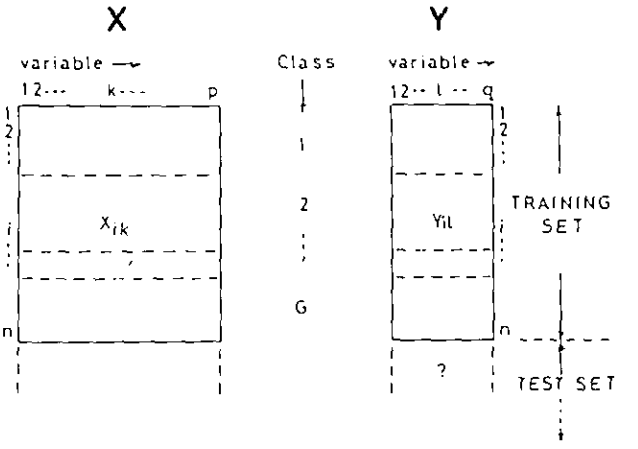


FIGURE 2. The data set of PARC levels III and IV consists of two matrices, $X$ and $Y$, both divided into a number of class training sets plus a test set. Usually the test set daya of the $Y$ matrix are initially undefined. The number of y variables q is one for PARC III and two or larger for PARC IV. The scope of the PARC analysis is to develop rules for classification on the basis of the $X$ data plus models for the quantitative prediction of $Y$ from $X$. These models are usually different for the different classes.

may be due to a lack of knowledge of the availability of appropriate data analytic methodology.

## Recent Developments in Data Analysis and the Effect on QSAR

Applied mathematics has lately provided new tools of data analysis. In particular, it is now possible simultaneously to analyze the values of many variables and their joint influence on a set of other variables. This can be done even for a limited number of cases—here compounds or mixtures—still keeping the risk for spurious results small and under control.

Hence more realistic data sets can now be analyzed in applications of quantitative structure activity relationships (QSAR) including structure–toxicity models. The toxicity of a set of compounds can be measured in several different ways giving multivariate activity data. The "block" of toxicity variables can be modeled in terms of a "block" of a great number of structure descriptor variables. And this for a number of compounds which may be small compared to the number of variables in any or both blocks.

"Biological activity," e. g., toxicity, is usually the result of a complicated system of "fundamental processes." Therefore, several biological measurements and tests are usually needed to "capture" the nature of this activity. Analogously, the modification of chemical structure and its influence on the biological activity needs a substantial number of variables, since the number of possible types of chemical–biological interactions is large. Thus the change of any part of a molecule induces changes in lipophilicity, steric factors, the electron distribution, hydrogen bonds, and so on.

The traditional PARC methods such as linear dis-

criminant analysis and the linear learning machine are, like multiple regression, severely restricted in chemical applications because they need the number of cases (compounds) to substantially exceed the number of variables in the analyzed data set. Efforts to get around this condition based on the selection of variables—often stepwise—usually does not solve the problem. On the contrary, while stepwise variable selection often leads to apparently nice classifications or correlations, the risk for the results to be spurious is embarrassingly high (*1*,*2*).

However, projection methods such as principal components analysis (*7*,*8*) and partial least-squares modeling in latent variables (PLS) (*8*,*9*) can deal with large numbers of both dependent variables ($y$) and predictor variables ($x$) and their relationships without increasing the risk for spurious results. This is because the data in each block ($Y$ and $X$) are separately contracted to a few "latent" variables which then are related to each other.

Since the contraction is made in a statistically controlled way separately for each block, and since only so few latent variables are extracted that are far fewer than the number of cases, the resulting relations between the blocks of variables are significant and the risk for spurious correlations plaguing the traditional methods is small and kept under statistical control.

With these projection methods, QSAR data can now be effectively and appropriately analyzed (*10*,*11*). Multivariate activity data can be related to multivariate structure descriptor data. When chemically diverse sets of compounds are investigated, they can be divided into homogeneous subsets, for which separate models can be constructed even if the number of compounds in the subsets becomes small.

## Models Can be Constructed Only for Sets of Similiar Compounds

The fact that only (sub)sets of chemically and pharmacologically similar compounds can be modeled has previously been an obstacle. Since the data analytic methods could not deal with the small subsets and large numbers of variables, the problem was ignored and models constructed for larger sets of structurally diverse compounds. However, this substantially increases the risk for spurious results because the cases (compounds) are no longer statistically independent. If this is not realized, the statistical significance is evaluated with the incorrect number of degrees of freedom and the resulting apparent probability levels are grossly inflated (*2*).

This, in turn, has led to the erroneous belief that in fact PARC and sometimes MR can be applied to structurally diverse sets of compounds. When one scrutinizes the mathematical foundation of empirical structure–activity models, one immediately realizes that this is an impossibility. If, for instance, some compounds are active according to one biological mechanism and some others are active according to another mechanism, any chemist knows that these two types of compounds cannot simultaneously be entered into the same model because a change in chemical structure affects the two mechanisms in two different ways.

Analogously, structurally diverse compounds cannot be entered into the same model because a small change in structure in one type of compounds has different chemical and biological effects than the same small change in another type of compounds.

This problem has now been sorted out and it has been shown that all claims that structure–activity models can be formulated for structurally diverse compounds are erroneous and based on incorrect data analysis and incorrect evaluation of statistical significance levels (*2*). The only problem remaining is one of psychology. We must accept that the construction of structure–activity models is difficult, demanding profound biological knowledge about the modeled system. Such knowledge often does not exist and therefore the modeling often fails. There is no fast and simple way to obtain information.

# Valid Ways To Construct Multivariate Models

We shall discuss below what one can do and cannot do in the context of structure–activity modeling, in particular when the biological activity relates to "toxicity." To facilitate a rational discussion, we divide the QSAR problem into a number of separate subproblems. Though these subproblems are only loosely connected, the data analysis has a strong influence on the earlier parts of a QSAR investigation. In particular, the design of an investigation is strongly affected by the knowledge that data with several toxicity measures ($Y$) are much more informative than a single measurement and that such multivariate activity data indeed can be related to a multitude of structural data ($X$).

## Modeling in General

The idea of modeling is closely related to the analogy concept. The idea is to construct models which behave analogously to the system we really wish to study and predict. Models are practical because they are simpler to study and thereby even easier to "understand." They are also cheaper and ethically preferable to manipulate and observe.

In toxicity studies, a number of model types can be seen. All of them can with advantage be used "multivariately." The relation between the "measurements" made on the model's system and the "real system" can be qualitative (classification, PARC levels 1 and 2, see above) or quantitative or both. Chain models can also be envisioned as exemplified by type 6. The pertinent system for which we wish to draw conclusions—the "real system" below—usually is man, but often other systems

are of interest, e.g., ecological systems, with the implicit assumption that detrimental effects of chemical agents on these systems indicate that they are harmful also for man.

(1) Model: A mathematical description of the chemical structure. Real system: The measured biological activity in one or several tests (in man or in an animal systems or in a cell test system).

(2) Model: Chemical and physical measurements on chemical compounds. Real system: as in 1.

(3) Model: *In vitro* tests. Real system: An animal, say rabbit, or man.

(4) Model: A combination of *in vitro* tests and chemical structure description. Real system: as in 3.

(5) Model: One or several animal system. Real system: Man or an ecological system.

(6) Models: Chemical structure (block 1), *in vitro* tests (block 2), animals (block 3). Real system: as in 5. A multiblock chain model can be organized as, for instance, block 1 → block 2 → block 3 → block 4.

We shall here discuss mainly models of type 1 and 2, relations between chemical structure and biological activity. However, the general approach and the data analytical methods is equally applicable to models of types 3–6. The PLS models discussed below can handle chains of blocks with any number from two to about a hundred blocks.

***Hard And Soft Models.*** We distinguish between hard (fundamental) and soft (empirical) models and argue that the latter at present are more suitable for toxicity models because of the general lack of knowledge of the fundamental mechanisms involved.

Hard models incorporate assumptions about the mathematical form of the model, derived from the "fundamental knowledge" about the studied system. Kinetic models in the form of systems of differential equations are typical examples of hard models.

Soft models, in contrast, incorporate as few assumptions as possible, usually only assumptions about homogeneity and continuity of the data. The homogeneity assumption in the present context corresponds to the assumption that the biological data measure "the same mechanism" of toxicity over the set of studied compounds. Hence, empirical models are necessarily local in their nature. Each model is valid only for a set of biologically and thereby also chemically similar compounds. Claims that empirical models can be constructed for sets of compounds with diverse structures are wrong as discussed above and based on an incorrect statistical evaluation of sampling artifacts.

The continuity assumption means that a small change in chemical structure shall also cause a small change in the measured toxicity. Though seemingly trivial, this assumption is essential in any kind of empirical modeling.

The lack of assumptions in empirical models is compensated by the use of multivariate data. Several measurements are used to characterize both the biological activity and the structure. These data inserted into the soft model give "patterns" specific for each type of systems. These "patterns" can then in retrospect be used to check if the "fundamental knowledge" is consistent with the actual data. Additionally, the patterns can be used to predict the biological activity ($Y$) of new compounds from their structural descriptors ($X$).

Soft multivariate models, below exemplified with the PC and PLS models, can be used to analyze experimental data and make predictions long before sufficient knowledge is available to apply fundamental, hard, models. The knowledge about a system is never complete, in particular not in a research situation. Hence, soft models are the best choice in early stages of an investigation and for new subproblems in older investigations. When the empirical knowledge increases, this can be incorporated into the models, making them less soft and more hard.

***The Analogy Assumption.*** In the present context, one further crucial assumption is made, namely, that of analogy: that the variation in structure can be described by numbers which are derived from chemical standard reactions. Thus, one assumes that the change of a substituent in the studied set of compounds, say from methyl to chloro, is causing a change in toxicity that can be modeled in terms of the same "effects" as in an ensemble of chemical standard reactions. The change from methyl to chloro induces a certain change in (*a*) $pK_a$ of *para*-substituted benzoic acids, (*b*) the distribution between octanol and water of *para*-substituted phenols, (*c*) the rate of ester hydrolysis of alpha-substituted acetic acid ethyl esters, and so on. These changes can be described as numerical values in different variables corresponding to (*a*), (*b*), and (*c*), etc. By selecting a sufficient range of chemical standard reactions, we hope to "capture" all effects by which a change in a substituent can change chemical reactivity.

We then assume that the toxicity is affected by the same effects, but in an unknown combination. Thus, toxicity is seen as a chemical reaction, albeit rather complicated. As will be discussed below, the set of standard reaction data can then be combined in a PLS model which well predicts the change in toxicity which takes place when a change is made in the chemical structure. Again, since chemical effects combined differently in different reactions, the models apply only for chemically similar compounds causing toxicity by the same biological mechanism.

A greater variation in chemical structure, and thereby in the way structure influences toxicity, must therefore be handled by an ensemble of disjoint models, one for each subset of similar compounds. For neighboring subsets, the models may be joined by soft relations between the latent variables.

***Design, Selection of Compounds.*** With any type of models—hard or soft—the selection of what and where to measure is of utmost importance for the later application of the models for predictions. With any model, extrapolations are imprecise far outside the domain of

the data on which the model was "calibrated." Therefore each structural "factor" must be varied to "span" the domain in which predictions are sought.

As shown by statistical investigations of design strategies, it is extremely inefficient to change one factor at a time in empirical modeling (*12*). In the present context of structure–activity models, this means that one must not construct the set of studied compounds by changing one structural element, one substituent, at a time. If, for instance, we have three substituent sites and we can at each site put the four substituents a, b, c or d, an inefficiently selected set would be: aaa, baa, caa, daa, aba, aca, ada, aab, aac and aad. This set gives no information about the joint influence of the three sites on the biological activity. A much better set derived by fractional factorials (*12,13*) would be aaa, acd, bbd, bda, cbb, cdc, dac and dcb.

***The Number of Compounds.*** With the projection methods discussed here, the relation between the number of objects (here compounds) and variables is unimportant. To span the abstract space of chemical structural variation for a given class of compounds, a certain minimal number of carefully selected compounds is necessary. In case we can divide the chemical structure into a fixed "backbone" plus a number ($m$) of substituent sites, the minimal set of compunds is $8 \times 2^{m-1}$, i.e., 8 for one substituent site, 16 for two, 32 for three, 64 for four sites, etc. The selection of these compounds is made by fractional factorial designs (*12,13*) in the variables used as structural descriptors, four to five per site (see below). These numbers may seem large, but we must remember the number of possible compounds which is at least $100^m$, i.e., 100 millions for the case with four sites.

## Biological Data

The characterization of the toxicity or other biological effect of a chemical compound is best made by a multitude of variables. Such multivariate data allow the independent judgment of the quality of the biological variables and also the resolution of these data into different "factors" as discussed below.

In traditional science, there is a strong tendency to try to characterize the state of a system by a single variable. The chemical potential is a typical example from physical chemistry, and $LD_{50}$ an example from toxicology. The more complicated an investigated system is, however, the more unlikely it is that this single variable is sufficient in a given problem. Thus, in the typical study of relationships between chemical structure and toxicity, one must measure toxicity in several different ways, preferably in several different test systems, to capture most of the different ways the chemical compounds can affect biological systems.

The limitation of the number of different toxicological measures one should include in an investigation is mainly economical. The larger the number of different measurements, the more information one obtains, but at a greater cost. At some point the marginal price of the next increment of information is too high and the practical limit is reached.

## Chemical Structure and Its Quantification

To apply PARC data analysis, the variation in chemical structure between the investigated compounds must first be translated to values of variables. The most direct and usually most informative way is to use chemical and physical properties measured on the chemical compounds. Lipophilicity (log $P$ octanol/water), acid-base properties ($pK_a$), solubility in water, IR and NMR spectra, and reactivities in model reactions are often relevant in structure–activity and toxicity studies. We note that to get these measurements, the compounds must actually exist in the real world (models type 2).

In models of type 1, one wishes to construct theoretical models so that predictions can be obtained for new compounds before they are actually synthesized and available for real measurements. If the compounds can be divided into a fixed structural backbone plus substituent sites, this task is reduced to the simpler problem of describing the substituents sitting at the different sites. The description of the electronic type of substituent demands two variables (*14,15*), another variable is needed for lipophilicity (*16*), another one or two for steric size (*17,18*); in total at least four or five structural variables per substituent site are required.

In the more difficult case, when a common backbone cannot be distinguished, there is presently no general way to describe chemical structure. The occurrence of structural fragments has frequently been used in PARC QSAR applications, but this is not recommended, since such variables lack the continuity properties required for PARC. Moreover, their use seems to increase the risk for spurious correlations (*2*). Finally, predictions are difficult to make for compounds with new structural fragments not represented in the training set.

Wise and Cramer (*19*) and Marshall (*20*) have promising approaches to deal with the quantification of the structural variation of flexible molecules, but as yet the experience is not sufficient to recommend these approaches for routine applications.

Quantum mechanical indices and energy levels may be useful both for the simpler case with backbone and substituents and for the case with flexible molecules. Molecular mechanics is possibly useful to calculate conformations of such molecules, but also these latter approaches are still far from routine.

As discussed above, the larger the number of relevant variables that are used to describe the chemical structure, the better. The data analysis is not complicated by a multitude of chemical descriptors.

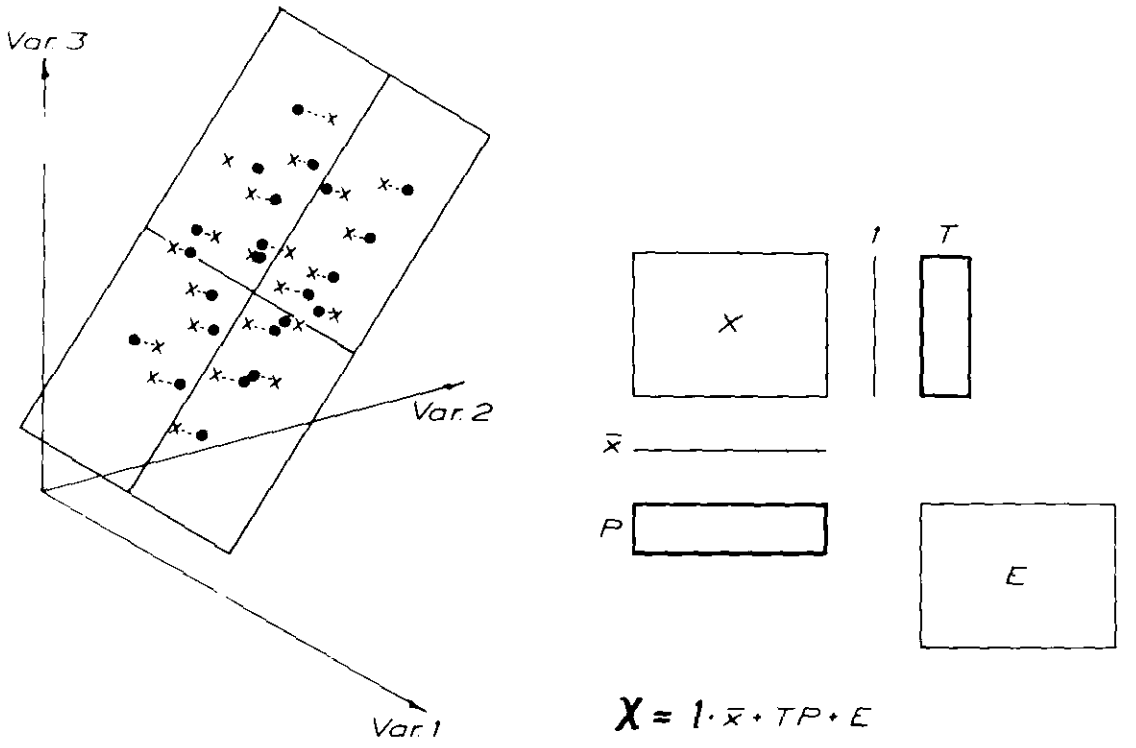


FIGURE 3. By PC analysis, the objects—rows in $X$ represented as points in a $p$-dimensional space—are projected down on a few-dimensional hyperplane (left). Algebraically, this corresponds to the decomposition of $X$ into a mean vector plus the product of two matrices $T$ and $P$ plus residuals $E$ (right).

## Data Analysis

The PARC methods used in structure–activity and toxicity studies usually are level-one methods; the linear learning machine (LLM), linear discriminant analysis (LDA), or $K$ nearest-neighbor methods (KNN). As discussed above and in the literature (*2,3,10,11*), these methods are not suitable for these applications. First, they cannot in a statistically appropriate way cope with outliers. Second, LLM and LDA are not applicable when the number of structure descriptor variables ($X$) exceeds the number of compounds in the training set. Methods of variable selection are then applied which greatly increase the risk for spurious results (*1,2*). In addition, the data are often scaled to enhance the class separation, which also increases the risk for spurious results even with the otherwise robust KNN method.

To formulate quantitative models of structure–activity relations, multiple regression (MR) is commonly used. Since MR has the same unfortunate limitation as LDA and LLM with respect to the number of $X$ variables, stepwise variable selection is also often used here. The resulting correlations are then often not statistically valid because the effect of the variable selection on the statistical significance levels is not taken into proper account (*1*). MR cannot simultaneously model several activity variables, which is another reason why it is less useful in the present context. Finally, MR gives unpredictable and little useful results when the $X$ matrix is collinear, which is often the case in QSAR due to the difficulties in applying traditional experimental design (*11*).

***Projection Methods, Principal Components, and PLS.*** An efficient way to analyze one or several data tables is to project the tables down on smaller tables with orthogonal columns. With PARC level I and II, the $X$ matrix of each class (denoted by $X_g$ for class $g$) is projected down on a few column matrix $T_g$ by means of the projection matrix $P_g$. Geometrically, this is equivalent to representing each object vector as a point in a $p$-dimensional space ($p$ is the number of $x$ variables) and then modeling the point swarm of each class as a few dimensional hyperplane. This is indicated in Figure 3 for a single class. In statistics this is called principal components (PC) analysis of the class data matrix.

This gives a model of the class in terms of the "class middle"—the vector $\bar{x}$)—and the direction coefficients of the hyperplane—the matrix $P$. New objects can be classified as similar to the class or not in terms of their calculated distance in $p$-space to the class hyperplane. Mathematically, this distance is calculated by a simple multiple regression with the new object data as the "dependent" variable and the rows in the $P$ matrix as predictor variables. Objects far from all class models are labeled as outliers.

We realize that the projections can be calculated regardless of the number of $x$ variables ($p$) and its relation to the number of objects in a class ($n_g$) or in the total

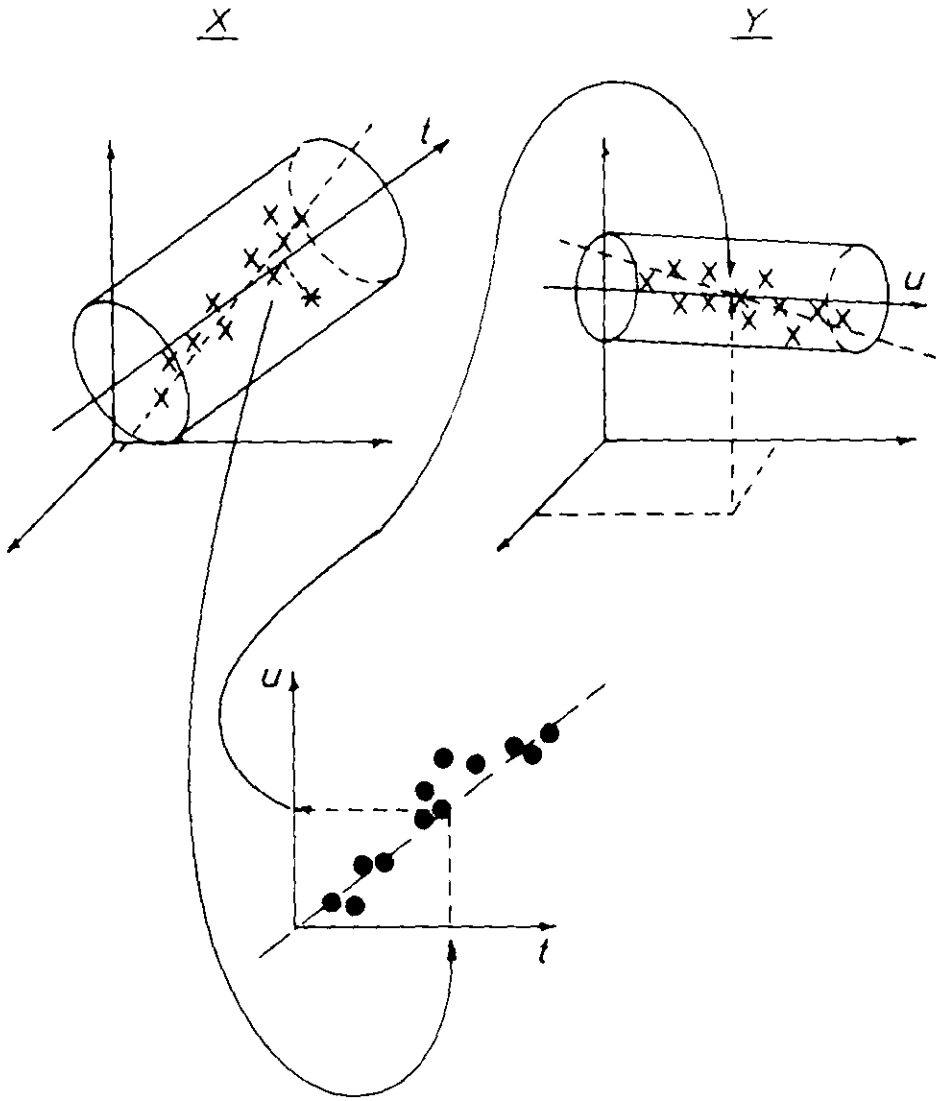


FIGURE 4. In PARC level IV, the data of one class are represented as points in two spaces, one for $X$ and one for $Y$. The PLS method projects these points down on hyperplanes in the two spaces. Here the simplest hyperplanes are shown, i.e., lines. These projections are made by least-squares so that the $t$ and $u$ coordinates of the objects correlate and so that hyperplanes describe much of the data variation in the $X$ and $Y$ spaces, respectively. Hypercylindrical tolerance intervals are constructed around each class model on the basis of the scatter of the training set points from the model. New objects falling inside the $X$ tolerance interval are assigned to the class and predictions of its activity values $Y$ are obtained from the chain $t$–$u$–$Y$

training set ($n$). In fact, the projections are more stable the larger the number of relevant $x$ variables.

This makes this PARC method—the SIMCA method (*2,3,8,10,11*)—well suited for structure activity and toxicity relationships. Separate PC models are calculated for each separate class—the subset of chemically similar compounds. The fact that these subsets consist of a limited number of compounds in comparison to the number of structure descriptor variables is no limitation for the SIMCA method.

With PARC levels III and IV we have also a $Y$ matrix. This can be represented in a separate space—the $Y$-space—and the $Y$ matrix of each class can separately be projected down on a hyperplane by means of the projection matrix $Q$. The object coordinates in this hyperplane comprise the matrix $U$. With the PLS method (partial least-squares models in latent variables) the projections of the $X$ and $Y$ matrices can be made simultaneously so that also the correlations between the $t$ and $u$ vectors (columns in $T$ and $U$) are optimized, thereby creating a connection between the two spaces (*9,11*). This gives a model for each class which can be used (a) to classify new objects (compounds) as similar to the class or not and (b) for compounds similar to the class, to obtain predictions of the $y$ data in terms of the connections between $t$ and $u$ (Fig. 4).

When only a single $Y$ variable is available (PARC III), the $Y$-space is one-dimensional. Then no projection

is calculated in $Y$-space, but this $y$ variable is used directly as the $u$-vector in each model dimension. In this case, PLS converges towards multiple regression (MR) with increasing model dimensionality. Thus, if indeed MR is appropriate for the given data, PLS will give the same results. In practice, however, PLS usually gives far fewer dimensions because of collinearities in $X$ and/or the presence of irrelevant variables in $X$.

The number of model dimensions in the PC and PLS models is determined by cross-validation. Part of the class data set is kept out from the computation of the class model and then predicted by the resulting model. The predictions are compared with the actual values and then another part of the data is kept out, a new model calculated, etc., until each data element has been kept out once and only once. This is made for each dimension, and that model dimensionality is selected which gives the best predictions of the kept out data.

***Estimation of the Reliability of Results.*** The use of cross-validation with PC and PLS models gives a reliable but somewhat conservative estimate of the amount of information in the data relevant for the current problem. Moreover, since the class models are developed separately for the different classes without any enhancement of class differences, the resulting class separation is not overestimated. This is in contrast to LLM and LDA, which find the maximal separation of the classes. This is much overestimated, particularly when the number of $x$ variables is large.

Similarly, multiple regression (MR) overestimates the amount of variance of $y$ explained by $X$. This because MR combines the $x$ variables to obtain the maximum correlation with $y$. Again, this overestimation is severe when the number of $x$ variables is large, and when the number used in the final model is reduced by stepwise selection.

The use of cross-validation to evaluate the results of any PARC or MR analysis would greatly reduce the frequency of spurious results publications. Such cross-validation must then be used so that all scaling, variable selection, and data analysis is made independently for each round of keeping part of the data set out of the model development. The erroneous practice to use only the final set of variables and objects and keep one out at a time is not a proper evaluation, since it does not evaluate the effects of scaling and variable selection on the results. These effects are often large (*1,2*).

## Psychological Problems

With a proper data analysis and a proper validation of the models, one often gets rather disappointing results in structure–activity investigations. This is natural because the studied systems are very complex, and the knowledge about how to describe chemical structure and its effect on biological activity is not well developed. Nevertheless, it is always disappointing to get diffuse results with very little of the biological activity explained.

One obvious way out is to use nonvalidated results of data anayltic methods such as LLM and MR known to give overoptimistic correlations. Since man has a strong desire to find positive results, there is a psychological resistance against a proper statistical validation of scientific models, including those used in toxicology predictions. This allows the publication of many dubious correlations in the present field and even commercial operations based on spurious models advertising that the toxicity, mutagenicity, and carcinogenicity of arbitrary untested compounds can be predicted from just the structural formula of the compounds.

## Asymmetric Nature of the Active-Inactive Classification

In structure–toxicity investigations, the classification problem is often formulated as the discrimination between active (toxic) and inactive (nontoxic) compounds. If one then applies level one PARC, the analysis is likely to fail or give spurious results. This because the problem is not symmetric (*21*).

The active class may be well defined structurally and toxicologically and thereby occupies a small regular volume in $p$-space. The lack of toxicity in the given biological test system, however, is not a class-defining property. Any compound which lacks the proper structural elements needed to trigger the mechanism of toxicity will be nontoxic. Hence, the majority of all billions

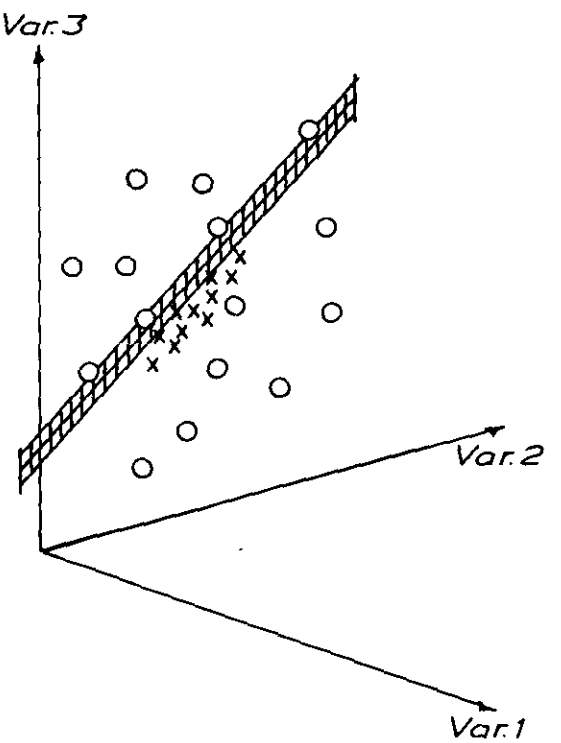


FIGURE 5. An asymmetric data structure with the class of active compounds occupying a small regular volume in $p$-space, while the nonactive compounds are randomly scattered in the same space. A discriminant hyperplane (LLM or LDA) cannot in a meaningful way separate the two classes.

of compounds will belong to the "nonactive" class. It is clear that in any conceivable space, these nonactive compounds will be randomly distributed. The active class can be modeled, while the "nonactive" class cannot. Level-one PARC methods such as LLM and LDA that try to separate the two classes by a hyperplane will, of course, fail (Fig. 5). This is one further reason why these methods cannot be recommended for structure–toxicity relationships.

Level two PARC methods such as SIMCA apply also in the asymmetric problem. Provided that the structural descriptors are relevant to the problem, the proper class of active compounds with similar structure can be well approximated by a PC model. The inactive compounds are likely to be situated outside the tolerance interval of this model. Hence new objects can be classified as similar to the active class (toxic) or not (Fig. 6).

We note one complication with the asymmetric problem formulation. A compound correctly classified as nonsimilar to the active class may be either inactive or active (toxic) according to another mechanism than the training set actives. Thus not-in-active-class does not necessarily mean inactive. This complication follows from the problem formulation and not from the way the data are analyzed. Hence, the problem is the same also in biological model systems. Lack of activity in a cell test battery need not necessarily correspond to lack of activity *in vivo*, just that the model system does not apply in the same way as before to the presently tested compound.

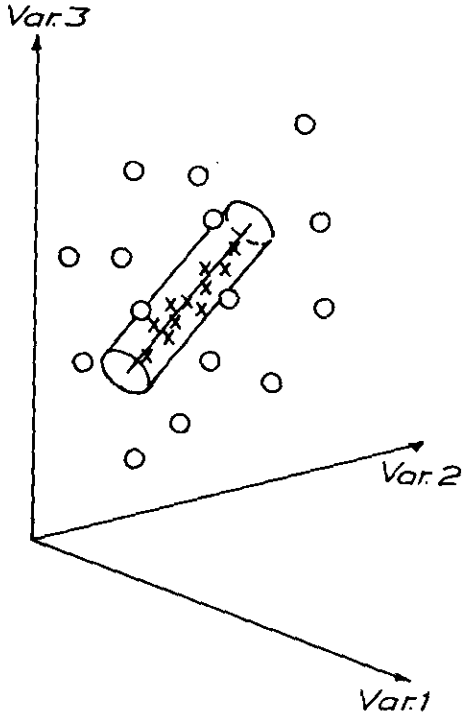


FIGURE 6. Level II PARC methods such as SIMCA work in the asymmetric case by enclosing the "active" class in a closed volume. New compounds are classified as active or nonactive, respectively, according to their position inside or outside this volume.

## Summary of Statistical Requirements

Empirical modeling is based on certain assumptions about the problem formulation and the properties of the data, notably those of homogeneity and continuity. In structure–activity and toxicity modeling, these assumptions can be translated to certain chemical and pharmacological requirements on the problem and the data as listed below. In addition, to give maximum information, the data should be well designed and relevant, both the structural and the biological activity variation should be described multivariately.

Each model should be applied only to a set of structurally and toxicologically similar compounds. If the data set contains structurally and toxicologically diverse compounds, separate models must be formulated for each homogeneous subgroup—proper class. We note that this is no loss of information, since it is very easy to distinguish between structural classes of compounds, this distinction is not the objective of the modeling.

If traditional statistical methods are used for the data analysis, the initial set of variables, before any selection and reduction, must have fewer members than a fourth of the number of compounds in the training set. When the number of variables is larger than this limit, or if there are collinearities in the $X$ matrix, linear discriminant analysis, linear learning machine or multiple regression methods cannot be used unless the data set is reduced by independent means, e.g., by principal components (factor) analysis.

Projection methods such as SIMCA and PLS work also with many variables and collinear data matrices.

If the problem is formulated as active–inactive, a method able to handle the asymmetric problem should be used, e.g., SIMCA.

When the data set is incomplete, i.e., there are missing observations in the $X$ and $Y$ matrices, most data analytical methods except projection methods (SIMCA, PLS), fail.

The variables used for describing the structural variation should be continuous, preferably directly derived from measured data on model systems (pi, sigma and the like).

The biological data should be relevant and precisely measured. A lack of precision in a single variable can be compensated by including several different variables measured on the same system or related systems.

Each class of compounds should have a certain minimum size. If a single substituent site is varied, eight compounds is minimum, 16 for two sites, 32 for three sites, and so on. These compounds should be selected according to an appropriate design, e.g., fractional factorial. It is not correct to change one structural factor at a time in the compound set.

We emphasize that these demands refer to models of type 1 relating theoretical structural descriptors to toxicity data. In models of types 2–6, where both data blocks $X$ and $Y$ consist of measured data, 15 compounds or thereabout in each class is sufficient to develop models

in most instances. If several structural fragments can be varied in a class, this variation must still not be made so that one fragment at a time is varied.

## Weak Points of Structure–Activity and Toxicity Models

With the availability of modern computerized data analytical methods the weak point in this field is not the data analytical methodology. Provided that the problem is correctly formulated, classes appropriately chosen, variables correctly selected and measured, and compounds correctly chosen to span the structural variation in the classes, the data analysis will extract all available information. Hence, when a QSAR does not give desired results, there is usually something wrong either in the problem formulation, in the design, in the chemical structure description, or in the biological data.

The most common error is the one of design, i.e., of compound selection. One structural factor is varied at a time or structurally diverse compounds are included into the same single model.

We still have much to learn about which effects a structural variation can have on chemical reactivity. If we see biological activity as a special case of chemical reactivity, we realize that the description of structure in QSAR is still far from appropriate. Only for rigid compounds with well-defined sites of substitution can we deal fairly routinely with this problem, but even here much development of substituent descriptors relevant for biological applications is needed. With flexible molecules, structure–reactivity and activity modeling is still in its infancy, and validated successful applications are very rare.

The greatest problem, however, is the specification, selection, and measurement of the biological data. To begin with, a good pharmacological model must be developed, a cell or an animal system that models the effects on the system of interest, often man.

Since a model is never an exact copy of a system, there are always imperfections in the predictions based on the model. One way to decrease the magnitude of these imperfections is to combine a multitude of pharmacological and toxicological measurements into the model, i.e., to use a battery of tests. As indicated above, a suitable data analysis with a projection method will twist the battery results in a way optimal for the given problem. We realize that this multivariate approach is difficult to accept because we are all brought up to the belief that the scientific approach is equivalent to search for a single "crucial variable" and to measure that with the highest precision. With the advent of computers and projection methods, however, the latter approach is no longer informationally optimal, but it will take a long time to change the scientific dogma in this respect.

A great difficulty in pharmacological and toxicological modeling is caused by the demand of homogeneity—chemical and pharmacological similarity in the mechanism of action—of the studied set of compounds. The level of knowledge is rarely such that this homogeneity can be assured in advance. Hence, the collected data are grouped and clustered, and one must be prepared to analyze the data accordingly. Often, however, the data do not directly contain information about inhomogeneities among the compounds, in particular if only a single activity variable ($y$) is available. Then only biological and chemical insight helps, but it may be insufficient in new and complicated problems.

If the data are grouped, one should not try to model them by a single relationship. If one does anyway, a good criterion of success is that the model not only connects the mean values of each group but that it also predicts the variation in biological activity better than chance within each subgroup. If not, each subgroup acts just like a point, and the model is just an elaboration of a line through two points.

## Examples

We shall not discuss here the applications of traditional PARC methods (LLM, LDA, KNN) to the classification of compounds as active (toxic) or inactive, because most of these applications have been made in such a way that the risk is high for the results to be just spurious.

However, a number of PARC applications based on projection methods have recently been published (*10,11,22–26*). In these studies, at least the risk for spurious results was kept under statistical control by avoiding stepwise variable selection and class-separation enhancing data scaling.

Most of the data sets in these applications were to some extent asymmetric; the "nonactive" class usually did not have any systematic data structure that could be modeled. This indicates that an asymmetric data structure is a common consequence of this type of active-inactive problem formulation.

To illustrate the projection methodology, we use a graph from a paper recently published (*11*). Callen et al. (*27*) published mutagenicity and toxicity data for seven halogenated hydrocarbons. We contracted the activity data to five $y$ variables, four measuring mutagenicity and one toxicity. The structural variation among the seven compounds was described by eleven $x$ variables, including traditional variables such as molecular refractivity, MR, and lipophilicity ($\log P$) and four quantum-chemically calculated variables (charges and electronegativities of carbon and chlorine atoms, respectively).

The data set ($n=7$, $p=11x$ variables, $q=5y$ variables) is a typical one that cannot be analyzed with traditional data analytical methods. A PLS-analysis gives two highly significant model dimensions, of which the first is shown in Figure 7. The first structural PLS dimension consists mainly of the traditional variables, MR, $\log P$, etc. Hence, most of the variation in toxicity

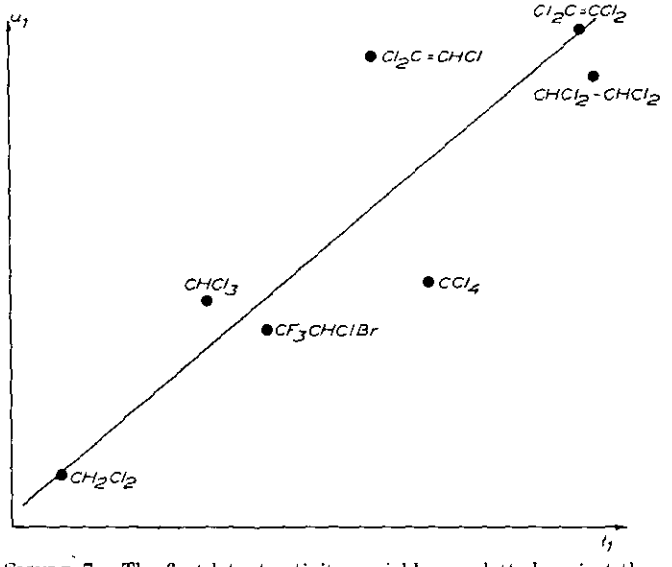


FIGURE 7. The first latent activity variable, $u$, plotted against the first latent structure variable $t$, showing a good connection between the $Y$ and $X$ space.

and mutagenicity seems to be connected to nonspecific interactions between compound and the cell system.

A separate principal components analysis of the $Y$ matrix gives two significant dimensions describing 95% of the variance in $Y$. This shows that the quality of the biological data is high. The PLS model explains 85% of this systematic part of $Y$, showing that there is still some precision to gain by a better structural description, i.e. better $X$. Thus a combination of a PC analysis of $Y$ and a PLS analysis of $X$ and $Y$ indicates where, i.e., in which part of the model, improvements can be made. It is interesting that in the present example the model deficiencies lie in $X$ and not, as is usually automatically assumed, in the biological data, $Y$.

The resulting PLS model can be used also for predictions. As an illustration, vinyl chloride was described by the same eleven $x$ variables and entered into the PLS model. The $X$ vector fits the model rather poorly, which says that vinyl chloride is not very similar to the seven-compound "training set" and thereby that the predicted biological activity is not very reliable. Vinyl chloride is predicted to be about as active as methylene chloride, the least active of the training set compounds.

## Discussion

In this article we have discussed the problem of modeling toxicity from the data-analytical and information theoretic point of view. This has made us rather critical to how "pattern recognition" and other data analysis—notably multiple regression—usually is applied in this problem area. We have the strong opinion that if not data analytical methods are used which correspond to the actual problem and data structures, little more than trivial and, more seriously, misleading results will be obtained.

A lack of understanding of the necessity to include only homogeneous data into a single model leads to severely grouped data, and the number of real degrees of freedom is grossly overestimated. The consequence is that statistically insignificant results may be thought to be significant and spurious correlations are taken to be real.

We know of only one exception in which structurally diverse compounds may be modeled by a single equation, namely, nonspecific toxicity. This is often related to the total lipophilicity of the compound (log $P$). This total log $P$ may be calculated approximately from fragment constants ($\pi$) (*16*). Thus, a model may be constructed which predicts the nonspecific toxicity from the total log $P$ which, in turn, is predicted from the sum of fragment contributions $\pi$.

The projection methods of principal components analysis and PLS have the advantage of being applicable also in cases when the number of analyzed variables exceeds the number of cases, systems, objects, compounds. Hence, there is no more need for data sets with large numbers of compounds. The data can be divided into homogeneous subsets without any loss of information and properly analyzed by separate models for each subset.

The fact that also several activity (toxicity) variables can be modeled simultaneously makes the projection models useful for data sets that are multivariate both in $X$ and $Y$. Since the complexity of toxicity more or less demands such multivariate characterization, we now have data analytical methods that don't have to mutilate the problem.

The use of multivariate $X$ and $Y$ data also allows the data analysis to separate the variation in the $Y$ data into a systematic part and "noise" and then further divide the systematic part into one part modeled by $X$ and one nonmodeled part. This allows the investigator to track deficiencies in the model and thereby to improve it.

The projection methods can handle problems on different "ambition" levels from mere classification (level I) to the combination of classification and quantitative modeling of several activity variables (level IV). Hence the same data analytical framework can be used for most or all data analytical problems in toxicological modeling.

The data analytical nomenclature may in this context be somewhat confusing. Many chemists use pattern recognition only for problems on level I and II, and see the quantitative modeling (levels III and IV) as totally different. With the projection methods we get all the levels into one and the same statistical and philosophical framework, thereby letting the problem guide the data analysis and not the reverse.

In conclusion, we can now concentrate on the important parts of structure–toxicity modeling, namely to get relevant biological data, to use an informative way to describe the variation in chemical structure, and, not the least, the design problem to select representative sets of compounds which well map the complicated abstract spaces in which predictions are sought.

We are grateful for financial support from the Swedish Natural Science Research Council (NFR), the Swedish Council for Planning and Coordination of Research (FRN) and the National Swedish Board for Technical Development (STU).

## REFERENCES

1. Topliss, J. G., and Edwards, R. P. Chance factors in studies of quantitative structure-activity relationships J. Med. Chem. 22: 1238–1244 (1979).
2. Wold, S., and Dunn, W. J. III. Multivariate quantitative structure activity relationships (QSAR): conditions for their applicability. J. Chem. Inf. Comput. Sci. 23: 6–13 (1983).
3. Wold, S., Hellberg, S., and Dunn, W. J. III. Computer methods for the assessment of acute toxicity. Proc. 1st CFN Symposium on LD50 and Possible Alternatives (P. Lindgren, Ed.), Acta Pharm. Toxicol. 52 (Suppl. II): 158–189 (1983).
4. Enslein, K., Lander, T. R., and Stranage, J. R. Teratogenesis: a statistical structure-activity model. Teratog. Carcinog. Mutag. 3: 289–309 (1983).
5. Jurs, P. C., and Yuan, M. Computer-assisted structure-activity studies of chemical carcinogens. A heterogeneous data set. J. Med. Chem. 22: 476–483 (1979).
6. Tinker, J. Relating mutagenicity to chemical structure. J. Chem. Inf Comput. Sci. 21: 3–7 (1981).
7. Mardia, K. V., Kent, J. T., and Bibby, J. M. Multivariate Analysis. Academic Press, New York, 1979.
8. Wold, S., Albano, C., Dunn, W. J., III, Esbensen, K., Hellberg, S., Johansson, E., and Sjöström, M. Pattern recognition: finding and using regularities in multivariate data. Proc. IUFOST Conf. Food Research and Data Analysis (H. Martens and H. Russwurm, Jr., Eds.), Applied Science Publ., London, 1983.
9. Wold, H. Soft modeling. The basic design and some extensions. In: Systems under Indirect Observation, Vol. II (K. G. Jöreskog and H. Wold, Eds.), North-Holland Amsterdam, 1982.
10. Dunn, W. J., III and Wold, S. Relationships between chemical structure and biological activity modelled by SIMCA pattern recognition. Bioorg. Chem. 9: 505–523 (1980).
11. Dunn, W. J., III, Wold, S., Edlund, U., Hellberg, S., and Gasteiger, J. Multivariate structure-activity relationships between data from a battery of biological tests and an ensemble of chemical descriptors: the PLS method. QSAR 3: 131–137 (1984).
12. Box, G. E. P., Hunter, W. G., and Hunter, J. S. Statistics for Experimenters. Wiley, New York, 1978.
13. Austel, V. 2ⁿ-Factorial schemes in drug design. Extensions increasing versatility. QSAR 2: 59–65 (1983).
14. Alunni, S., Clementi, S., Edlund, U., Johnels, D., Hellberg, S., Sjöström, M., and Wold, S. Multivariate data analysis of substituent descriptors. Acta Chem. Scand. B37: 47–53 (1983).
15. Johnels, D., Clementi, S., Dunn, W. J., III, Edlund, U., Grahn, H., Hellberg, S., Sjöström, M., and Wold, S. Clustering of aryl carbon-13 NMR substituent chemical shifts. A multivariate data analysis using principal components. J. Chem. Soc. Perkin II 1983: 863–871.
16. Hansch, C., and Leo, A. J. Substituent Constants for Correlation Analysis in Chemistry and Biology. Wiley, New York, 1979.
17. Verloop, A., Hoogenstraaten, W., and Tipker, J. Development and application of new steric substituent parameters in drug design. In: Drug Design (E. J. Ariens, Ed.), Academic Press, New York, Vol. 7, 1976.
18. Taft, R. W. Separation of polar, steric, and resonance effects in reactivity. In: Steric Effects in Organic Chemistry (M. S. Newman, Ed.), Wiley, New York, 1956.
19. Wise, M., Cramer, R. D., Smith, D., and Exman, I. Progress in three dimensional drug design: the use of real time colour graphics and computer postulation of bioactive molecules. In: Quantitative Approaches to Drug Design (J. C. Dearden, Ed.), Pharmacochemistry Library Vol. 6, Elsevier, Amsterdam, 1983.
20. Marshall, G. R. Computer graphics and receptor modelling. In: Quantitative Approaches to Drug Design (J. C. Dearden, Ed.), Pharmacochemistry Library Vol. 6, Elsevier, Amsterdam, 1983.
21. Dunn, W. J., III, and Wold, S. Structure–activity analyzed by pattern recognition: the asymmetric case. J. Med. Chem. 23: 595–599 (1980).
22. Nordén, B., Edlund, U., and Wold, S. Carcinogenicity of polycyclic hydrocarbons studied by SIMCA pattern recognition. Acta Chem. Scand. B32: 602 (1978).
23. Nordén, B., Edlund, U., Johnels, D., and Wold, S. Simplified C-13 NMR parameters related to the carcinogenic potency of PAH. QSAR 2: 73–76 (1983).
24. Dunn, W. J., III, and Wold, S. A structure-carcinogenicity study of 4-nitroquinoline 1-oxides using the SIMCA method of pattern recognition. J. Med. Chem. 21: 1001 (1978).
25. Dunn, W. J., III and Wold, S. The carcinogenicity of N-nitroso compounds. A SIMCA pattern recognition study. Bioorg. Chem. 10: 29–45 (1981).
26. Hellberg, S., Wold, S., Dunn, W. J., III, Gasteiger, J., and Hutchings, M. G. The anaesthetic activity and toxicity of halogenated ethyl methyl ethers, a multivariate QSAR modelled by PLS. QSAR 4: 1–11 (1985).
27. Callen, D. F., Wolf, C. R., and Philpot, R. M. Cytochrome P-450 mediated genetic activity and cytotoxicity of seven halogenated hydrocarbons. Mutat. Res. 77: 55–63 (1980).