# QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure

**Alan R. Katritzky, Victor S. Lobanov**
*Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, P.O. Box 117200, Gainesville, FL 32611-7200, U.S.A.*
**Mati Karelson**
*Department of Chemistry, University of Tartu, 2 Jakobi Str., Tartu, EE 2400, Estonia*



## 1 Aims and Objectives of QSPR Research

The structural formula of an organic compound, in principle, contains coded within it all of the information which predetermines the chemical, biological, and physical properties of that compound. The molecular formula defines precisely all of the molecular properties and features, including, for example, the compound's rate of oxidation, the equilibrium constant and rate of absorption on any defined surface, the degree to which it will inhibit rust formation in sea water under any defined set of conditions, and so on. If we could only read the code, such properties could be elucidated simply from a knowledge of the molecular formula.

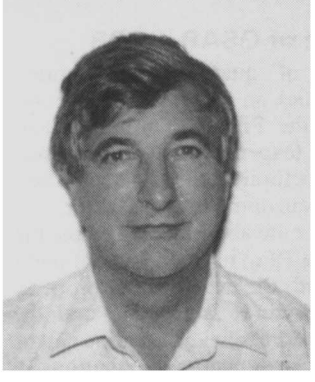
There are two main alternative approaches to Edisonian random testing to find compounds with superior properties. One consists of theoretical calculations using quantum and statistical mechanics. Solution of the Schrödinger equation would allow prediction of all of these factors in stationary states of molecules. However, although much progress has been made, particularly with semi-empirical methods, the practical application of quantum theory to complex systems still remains a distant possibility.

The other alternative is QSAR/QSPR. A major goal of Quantitative Structure–Activity Relationship (QSAR) or Quantitative Structure Property Relationship (QSPR) studies is to find a mathematical relationship between the activity or property under investigation (*e.g.* $LD_{50}$, $pK_a$, *etc.*), and one or more descriptive parameters (descriptors) related to the structure of the molecule. While such descriptors can themselves be experimental properties of the molecule, it is generally more useful to use descriptors derived mathematically from either the 2D or the 3D molecular structure, since this allows any relationship so derived to be extended to the prediction of the property or activity for unavailable compounds. If an acceptable model of this type can be found, it can guide the synthetic chemist in the choice between alternative hypothetical structures. More fundamentally, such studies can illuminate, or even elucidate, the 'mechanism' by which the property or activity in question is related to the chemical structure.

## 2 Background

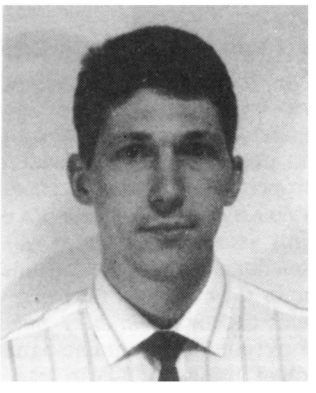
Some years ago, in a major ongoing project in one of our laboratories,[1–3] we were faced with the analysis of complex mixtures containing up to 30 different compounds derived from a single precursor at high temperature. We found that such mixtures could satisfactorily be analysed by the GC/MS method. The peaks were resolved, and from their fragmentation patterns, knowledge of the precursor, and some idea of the mechanism, we were able to assign chemical structures to all of our products. However, before we could carry out a quantitative analysis, it was necessary to know the response factors for all of the compounds, *i.e.*, the relationship between the integrated area under the GC curve for the individual peaks, and the compound structure. Response factors vary considerably. Many of the compounds formed as products were unavailable, and in some cases, even unknown; it was quite impracticable to measure the response factors experimentally. Thus, we began an investigation of QSPR relationships between GC analytical response factors and chemical structures.

*Alan R. Katritzky (b. 1928, London, U.K.) is Kenan Professor of Chemistry and Director of the Institute for Heterocyclic Compounds at the University of Florida. A light-hearted account of his life appeared in* J. Het. Chem., 1994, **31**, pp.569—602, *and an overview of his scientific work in* Heterocycles, 1994, **37**, pp.3—130. *He is actively engaged in research and teaching, editing, industrial consulting, international travel, and windsurfing.*



Alan R. Katritzky

*Victor S. Lobanov, born in 1966 in Yekaterinburg, Russia, received his M.S. in chemistry at the Moscow State University, Russia, in 1988. He is now working for his Ph.D. with Professors Karelson and Katritzky.*

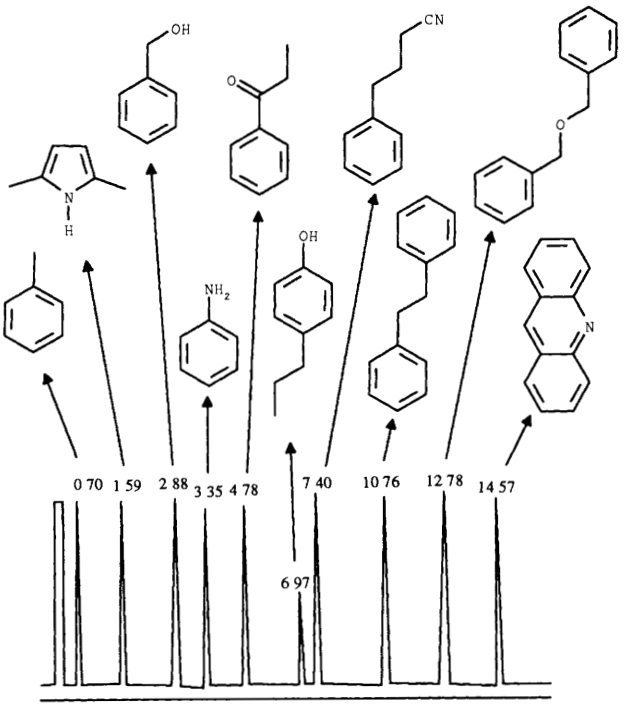


Victor S. Lobanov

*Mati Karelson (b. 1948) is the Professor and Head of Theoretical Chemistry at the University of Tartu, Estonia. He received his Ph.D. in physical organic chemistry in 1975. His research has dealt with the theory of solvent effects, the foundations of QSAR/QSPR, and the development of the respective computer software. He is a member of the International Society of Quantum Biology and Pharmacology and the New York Academy of Sciences. In 1994, he was nominated as a Courtesy Professor in Chemistry at the University of Florida.*

Mati Karelson

(a)




(b)

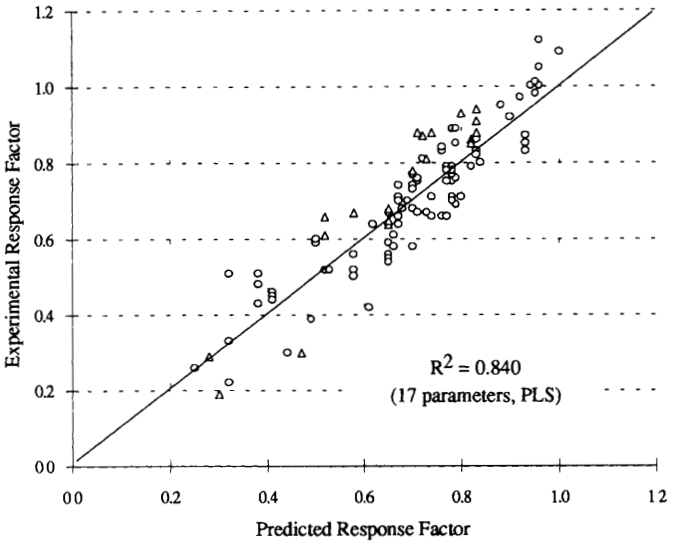



**Figure 1** (a) GC trace of a mixture of ten compounds (b) Plot of the experimental *vs.* calculated (by the PLS model) response factor values for compounds from the reference set (○) and from the test set (△) [4]

An illustration of the problem is demonstrated by the analysis of the GC trace of the mixture shown in Figure 1a. We used Partial Least Squares Analysis to search for a reliable dependence of the RF on the structural descriptors in terms of atomic groupings.[4] We measured the response factors for a set of 120 different compounds under standard conditions. We used these response factors as the dependent variable in the Partial Least Squares Analysis and used seventeen descriptors as the explanatory variables. All of these descriptors were simply determined from the chemical structure: (in order from 1 to 17) molecular weight, numbers of C, H, O, N, and S atoms, numbers of C=C, C=O, and C=N double bonds, number of C≡N triple bonds, number of rings, and numbers of carboxyl, hydroxyl, aldehyde, ester, amino, and ether or sulfide groups. The results of the Partial Least Squares Analysis led to equation 1, which then enabled the prediction of response factors for hypothetical compounds. The success of the method is shown graphically in Figure 1b. Although the correlation is not perfect, it is certainly much better than the alternative, which was to have set all of the RF values as equal to unity.

$$\begin{aligned} RF = {} & 0\,991 - 0.000908x_1 + 0.00234x_2 + 0.00276x_3 \\ & - 0.112x_4 - 0\,0711x_5 - 0.160x_6 + 0.00337x_7 \\ & - 0.0434x_8 - 0.0479x_9 - 0\,0777x_{10} + 0.00481x_{11} \\ & - 0.323x_{12} - 0.0536x_{13} - 0\,206x_{14} - 0.0459x_{15} \\ & - 0\,166x_{16} - 0.0675x_{17} \end{aligned} \quad (1)$$

This initial problem, and its interim solution, led to our interest in the search for regularities in the manner in which various molecular properties change, and how such variations depend upon molecular structure – the main focus of QSAR/QSPR investigations. The general solution of this problem is of great importance. Even for a narrow class of compounds, any regularities disclosed can be used to rationalize the behaviour of molecules in the set, and especially to forecast the properties of other (sometimes hypothetical) compounds belonging to the given class. Moreover, the relationships thus revealed between structures and properties (or biological activities) could be important for the development of a new theory, which could in turn both explain the observed phenomena and illuminate the mechanisms of physical or chemical phenomena or of biological activities. Thus QSAR/QSPR studies can bring us nearer to the ultimate goal of targeted molecular design.

QSAR/QSPR investigations in the past two decades have made significant progress in the search for quantitative relations between structure and property. Various mathematical modelling techniques have been employed and a whole new area involving the application of computers to chemistry has opened up. Extensive software for the determination of 'structure–activity/property' relationships has been created, including modules for structure input and for the calculation of empirical and also non-empirical descriptors of the given structures. In most cases, a statistical treatment of these results is also incorporated into the package. To mention a few representative programs, there are universal packages, such as ADAPT distributed by MDL and CODESSA distributed by Semichem. Other QSAR packages have been designed to treat congeneric structures only: DARC/PELCO by University of Paris, and OASIS by PI Burgas, Bulgaria. Additionally, some programs provide QSAR for specific compounds or data: TOPKAT by HDI, MEDCHEM by Pomona College, and POLLY by University of MN.

## 3 Theoretical Foundation of QSAR/QSPR

The mathematical foundation of quantitative structure–activity/property relationships relies on the principle of polylinearity (PPL).[5] According to the PPL, a continuous and singular dependence between the (experimentally measurable) property $P_i$ and some intrinsic structural factor of the molecule, $X_j$, is assumed to be linear in a certain domain of this factor, $\{x_j\}$. This assumption may be valid or invalid, depending on the functional form of the dependence $P_i(x_j)$ in the vicinity of point $x_j$. In the event of PPL being valid, the experimental property may depend upon only one structural factor, $X_1$, and the corresponding linear one-parameter regression equation can be found using the linear least squares method (equation 2). Alternatively, the property may depend linearly on several structural factors $X_j$, and the corresponding multi-parameter

regression is found using the multilinear least squares method (equation 3)

$$P_i = a_{0i} + a_{1i}X_1 \quad (2)$$

$$P_i = a_{0i} + a_{1i}X_1 + a_{2i}X_2 + a_{3i}X_3 + \quad (3)$$

In the event that the PPL approximation is invalid, some form of nonlinear equation has to be applied for the description of the dependence of the experimental property on structural factors If a nonlinear transformation of the structural factors is used, *e g* square, logarithm, or cross-term (equation 4), then the formal functional dependence of $P_i$ on $X'_j$ is still linear and the corresponding correlation equation can be found using the same multilinear least squares method (equation 5)

$$\begin{aligned}&(\text{square})X'_j = X_j^2 \text{ or } (\text{logarithm})X'_j = \ln X_j \\ &\text{or } (\text{cross-term})X'_j = X_jX_k\end{aligned} \quad (4)$$

$$P_i = a'_{0i} + a'_{1i}X'_1 + a'_{2i}X'_2 + a'_{3i}X'_3 + \quad (5)$$

In some cases – notable examples in chemistry include the exponential relationships between experimentally measurable quantities and intrinsic structural factors in chemical or absorption kinetics – the nonlinear least squares method has to be applied in order to find the regression parameters $a_{ji}$, predetermining the functional dependence $P_i(X_j)$ For instance, in the case of the simple parallel reaction A→B or C with two rate constants $k_1$ and $k_2$, the observable rate constant by initial compound A is given by equation 6 and the regression coefficients $a_{ij}$ can be found using nonlinear least squares techniques

$$\begin{aligned}&k_{obs}([A]) = k_1 + k_2 = \exp(-\Delta G^{\ddagger}_1/RT) + \exp(-\Delta G^{\ddagger}_2/RT) \\ &\text{where } \Delta G^{\ddagger}_1 = a_{01} + a_{11}X_1 + a_{21}X_2 + a_{31}X_3 + \\ &\text{and } \quad \Delta G^{\ddagger}_2 = a_{02} + a_{12}X_1 + a_{22}X_2 + a_{32}X_3 +\end{aligned} \quad (6)$$

However, most of the QSAR/QSPR applications are based on the PPL and therefore we limit ourselves in the following discussion to linear relationships only, keeping in mind that their extension to nonlinear dependenceis is straightforward

Historically, the first applications of quantitative structure–property relationships in chemistry were related to chemical reactivity in solution Hammett[6] defined the $\sigma$-constant of a substituent in a phenyl ring as the logarithmic ratio of the acidic dissociation constants of the substituted to unsubstituted benzoic acid in aqueous solution (equation 7), he demonstrated its applicability for the description of rate and equilibrium constants of various chemical reactions involving substituted benzenes

$$\sigma = \log(K_X/K_H) \quad (7)$$

Taft[7] extended these so-called linear free energy relationships to aliphatic structures by defining $\sigma^*$ constants of substituents as the differences in the logarithmic ratios of the rate constants for acid- and base-catalysed hydrolysis reactions of substituted and unsubstituted ethyl acetates (equation 8)

$$\sigma^* = 1/2\ 48[\log(k/k_0)_B - \log(k/k_0)_A] \quad (8)$$

To date, numerous empirical scales have been proposed to describe the inductive, resonance, and steric substituent effects in the molecules as well as for the description of solvent effects on the chemical and physical properties of compounds [8–10]

A major disadvantage in the use of these empirical molecular structural factors as descriptors results from their definition on the basis of some experimental information Consequently, certain standard compounds used to define the substituent or solvent parameters have to be synthesized and the corresponding standard properties measured Purely experimental problems (instability of the compound, insolubility in a given medium, *etc* ) may substantially restrict the selection of compounds whose properties can be predicted using these descriptors Notably, because of the insufficient solubility of many substituted benzoic acids in aqueous solution, Hammett had to use $pK_a$ values in 50% ethanol–water mixtures (another standard series) for the extension of his $\sigma$-scale

## 4 Main Types of Descriptors used in a QSAR/QSPR Program

We now review the main types of theoretical descriptors derived from molecular structure We believe that it is advantageous to subdivide descriptors into various subsets according to the molecular peculiarities which they reflect Thus, we distinguish constitutional, topological, geometric, electrostatic, quantum-chemical, thermodynamic, and solvation descriptors It should be emphasized, however, that many descriptors are, in fact, simultaneously sensitive to a number of molecular features, and the classification given below is therefore somewhat approximate and provisional Examples of commonly used descriptors are given in Table 1

Constitutional descriptors depend fundamentally on the composition of the molecule rather than on the topology, geometry, or electronic structure The counts of atoms of different elements and the molecular weight reflect the composition only, however, numbers of rings or double bonds are also sensitive to the molecular topology Constitutional descriptors, whilst very simple in nature, should be included in QSAR/QSPR studies If not, the possibility exists that a simple dependency (additivity) upon number of atoms or molecular weight would be represented by other more complex (and thus more difficult-to-comprehend) descriptors, or may even be overlooked altogether

Topological descriptors are probably the most widely used class of descriptors and include such well-known classical molecular parameters as the Wiener index,[11] the Randic index,[12] and the Kier & Hall Molecular Connectivity index [13] These descriptors are obviously most sensitive to the molecular topology (*i e* molecular connectivity), and in particular to the branching of the molecule Again, some topological descriptors (*e g* the Kier & Hall index) also reflect molecular composition, although to a lesser extent than the constitutional descriptors

The electrostatic descriptors essentially reflect the electrostatic structure of the molecule (*e g* the partial charge distribution or the electronegativities of the atoms), although in many

**Table 1** Main types of molecular descriptors

**Constitutional descriptors**
molecular weight
counts of atoms and bonds
counts of rings

**Topological descriptors**
Weiner index
Randic indices
Kier & Hall indices
information contents

**Electrostatic descriptors**
partial charges
polarity indices
charged partial surface areas

**Geometrical descriptors**
principal moments of inertia
molecular volume
solvent-accessible molecular surface
shadow indices

**Quantum-chemical descriptors**
net atomic charges
dipole moment
polarizability
$\sigma,\pi$ bond orders
HOMO, LUMO energies
FMO reactivity indices

cases they are also related to the molecular topology and composition.[14] For example, the descriptors of the 'Charged Partial Surface Areas' family reflect, in comparable proportions, the electrostatic, geometric, and topological features of a molecule.[15,16] The distribution of partial charges can be calculated by one or more non-empirical procedures within the QSAR/QSPR program, or independently of the program by any desired method, for instance, by a quantum-chemical program.

Geometric descriptors represent the three-dimensional characteristics of the molecular structure, *i.e.* molecular size and molecular shape.[17]

Quantum-chemical descriptors are a relatively new and rapidly developing class of molecular descriptors. As *ab initio* and semi-empirical quantum chemical calculations become increasingly available and routine, these descriptors have become more widely used. Quantum-chemical calculations can provide vast amounts of varied information about chemical structure including geometric and electrostatic data. But most importantly such calculations can provide information about the internal electronic properties of molecules which is not available by other means. Thus, quantum-chemical descriptors extend the areas of application of QSAR/QSPR techniques. The most frequently used quantum-chemical descriptors include the energy of the highest occupied and lowest unoccupied molecular orbitals, frontier orbital electron densities, Mulliken population charge distribution, and dipole moments.[18–24]

In addition to using these descriptors in the QSAR/QSPR in their traditional forms, it is frequently possible to include additional 'modified' descriptors. For example, most descriptors may also be normalized by dividing by the number of atoms, therefore giving an 'average' value of the descriptor. Advanced QSAR/QSPR programs can calculate not only the standard descriptor values, but also several modifications of them.

## 5 General Flow-Chart of a QSAR/QSPR Program

We proceed now with a more detailed description of the overall procedure of computer-assisted QSAR/QSPR research which is summarized by the following steps (Figure 2).

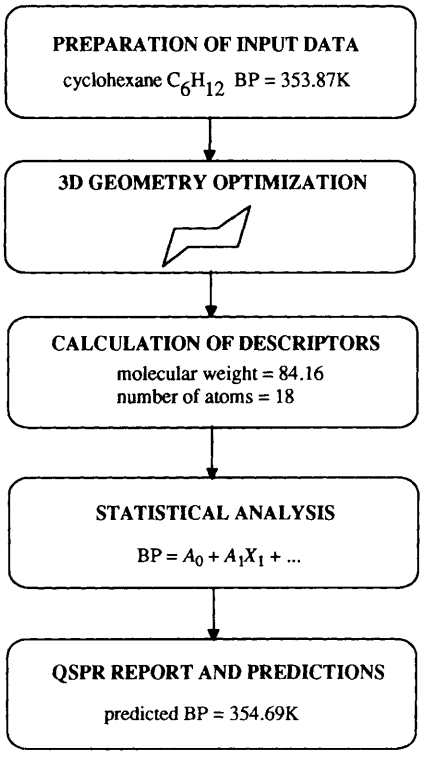



**Figure 2** Flow chart of a QSPR study.

### 5.1 Preparation of Input Data

First, the set of experimental data and the set of corresponding structures each needs to be prepared in computer-acceptable format. The experimental data, being numerical values, are usually easily represented in computer-readable format. The computer-compatible representation of structural formulae is an important step, which will be discussed in more detail below. Briefly, a specialized molecular editor is usually needed for the conversion of drawings into the corresponding connectivity tables.

### 5.2 3D-Geometry Optimization

Molecular shapes and conformations are often of great importance for the prediction and description of biological activities and molecular properties. Simple molecular editors create connectivity tables which do not contain any geometrical information. In such cases a 2D→3D converter is required. For example, the MOLGEO program[25] converts the 2D-connectivity tables into a 3D-representation of the molecule with concurrent geometry optimization. Programs exist which combine a molecular editor with a geometry optimization routine that is usually based on molecular mechanics or a semi-empirical method, *e.g.* PCMODEL, SYBYL, HyperChem. If more precise 3D Cartesian coordinates are required an additional refinement, by quantum-chemical *ab initio* or semi-empirical methods such as AM1[26] (*e.g.* MOPAC[27] and AMPAC[28] programs), is suitable.

### 5.3 Calculation of Molecular Descriptors

After the set of 3D-optimized structures is prepared, the QSAR/QSPR program carries out the calculation of molecular descriptors. The molecular descriptors applicable in QSAR/QSPR research were discussed in more detail above. In order to calculate quantum-chemical descriptors, preliminary calculations of the molecular electronic structure by the appropriate method (*e.g.* AM1[26]) are required.

### 5.4 Statistical Treatment of Data

The set of molecular descriptors calculated by the QSAR/QSPR package is further treated *via* comparison with the experimental data by (multi)parameter linear regression analysis.[29] Two major problems arise which are related to the use of a large number of molecular descriptors as potential independent variables in the QSAR/QSPR equations under development. First, in the case of hundreds or even thousands of descriptor scales, the number of possible (multi)linear equations to be verified for the best correlation is astronomically large and the calculation of all relevant regression equations takes an impractically long time even using the fastest of modern computers. Secondly, the calculated descriptor scales are, in general, non-orthogonal, *i.e.* the corresponding intercorrelation coefficients deviate significantly from zero. It is well-known from basic mathematics that an arbitrary scale (property $P$) cannot be singularly represented in the space of such descriptor scales. In other words, two or more different correlation equations exist in this space with exactly the same statistical fit characteristics. This obviously complicates the physical interpretation of any of these equations immensely. Even worse, the collinearity of descriptor scales may often lead to statistically meaningless regression equations without any predictive power. Therefore, special precautions have to be taken to avoid or alleviate these problems. For instance, factor analysis methods like principal component analysis (PCA) or nonlinear partial least-squares (NIPALS) can be applied to transform the non-orthogonal descriptor scales into orthogonal formal principal factors, describing within a given statistical tolerance all of these scales.[30] The dimensionality of the corresponding vector space is usually dramatically reduced making the search for the best correlation quite

straightforward However, it is often very difficult or even impossible to give a physically meaningful interpretation either to the regression coefficients preceding the formal scales, or to these scales themselves In order to overcome this difficulty, numerous target transformation techniques are available which aim to find (nearly) orthogonal subsets of original descriptor scales with the maximum coverage of the full descriptor space [31] Complex logical or heuristic procedures, described in more detail below, can also be efficiently used to find the best (multi)linear representation of the given property in the (nearly) orthogonal basis of natural descriptor sets

### 5.5 QSAR/QSPR Report

The typical output of a QSAR/QSPR program includes a number of correlation equations including correlation coefficients ($R$ or $R^2$), statistical significance tests ($F$-test, $t$-test), involved descriptors ($x_i$) and their corresponding regression coefficients ($a_i$) Selection of the best correlation model is usually done by validation of each model either by cross-validation techniques or by prediction response values for the test set

## 6 QSAR/QSPR Software

A statistical procedure is always necessary for the statistical treatment of data obtained from a QSAR/QSPR program There are several commercially available statistical packages, such as SAS or STATGRAPHICS [32] However, such modules are not always suitable for handling very large data sets or for producing an unambiguous and easily understandable selection of the parameters involved in the required multiparameter regression Numerous QSAR/QSPR packages are available both commercially and in the academic environment The DARC and OASIS programs are configured to treat so-called 'congeneric' sets of molecules In other words, all structures in the set under consideration should possess the same basic molecular fragment Such packages utilize the most traditional approach in QSAR, established by Hansch in the late 1960's [33 34] The Hansch approach assumes that (*i*) there is a basic fragment in the structure responsible for a given kind of property/activity, (*ii*) the variation of the property/activity value can be explained by the influence of different substituents attached to the basic fragment, and (*iii*) the property/activity value can be calculated as an additive value, the sum of terms being derived from the physicochemical constants of the substituents (similar to the Hammett–Taft equations 7,8)

However, many failures of the Hansch equation have been observed, and significant limitations of this approach have been recognized over the past 30 years A further serious drawback is evident, in the real world the series of compounds that most interest chemists are rarely congeneric! It is impossible for the Hansch approach to be applied to such non-congeneric series Even when a series does comprise 'related' compounds, it is often difficult to assign a physicochemical constant to an unusual substituent Hansch analysis proved to be a good beginning, but the development of other approaches became essential

There are many programs currently available which have been formulated to work on specific data sets A good example is the TOPKAT program, designed exclusively to predict toxicity endpoints from molecular structure The MEDCHEM program is structured to calculate log$P$ values The POLLY program operates on a restricted set of molecular descriptors By contrast, the ADAPT and CODESSA programs are universal packages able, in principle, to treat any set of compounds and their associated experimental data, although in some individual cases the use of one of the specific programs might lead to a more precise regression model

The ADAPT program offers numerous options and possibilities such as intellectual cluster analysis within a given set of compounds, complex statistical treatment of the experimental data together with the calculated descriptors, combination of the empirical parameters with those obtained by the modern non-empirical quantum chemical calculations, expert system features, databases, highly developed user interfaces with graphical input–output, and other menus or windows

CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis) is a chemical multi-purpose statistical analysis and prediction program CODESSA operates in a Microsoft Windows environment for personal computers and also provides an easy-to-use user interface with menus and windows for textual and graphical output Within the framework of the program up to one thousand varied molecular descriptors can be calculated Moreover, CODESSA provides the facility to construct numerous new descriptors using any previously calculated descriptors and standard mathematical operations and functions The correlation techniques available include multi-linear regression analysis, principal component analysis, and the heuristic method Elaborate techniques have been developed in CODESSA for a fast and adequate search for the best correlation equations for a given property and structure set In the following section one of these techniques is described in more detail

## 7 Statistical Treatment of Data in QSAR/QSPR

As discussed above, the rigorous selection of the uniquely best parameters for a correlation equation from a very large descriptor set remains an unsolved statistical problem As frequently pointed out in statistical texts and papers, the existing heuristic selection of descriptors in a step-wise multi-parameter regression analysis does not necessarily guarantee the best selection of descriptors

In the CODESSA program the search for the multi-parameter regression with the maximum predictive ability is performed using the following strategy [35]

1 First, the intercorrelations between all descriptors are calculated and all orthogonal pairs of descriptors i and j (with $R_{ij}^2 < R_{min}^2$) are found

2 The statistical analysis of the property starts with the calculation of the two-parameter regression with each orthogonal pair of descriptors, obtained in step 1 The descriptor pairs with high regression correlation coefficients are then selected for higher-order regression treatments

3 For each descriptor pair selected in the previous step a non-collinear descriptor scale, $k$ (with $R_{ik}^2 < R_{nc}^2$ and $R_{kj}^2 < R_{nc}^2$), is added, and the respective three-parameter regression is calculated If the Fisher criterion at a given probability level, $F$, is smaller than that for the best two-parameter correlation, the latter is chosen as the final result and the program proceeds to the printout section (step 5) Otherwise, the descriptor triplets with the highest regression correlation coefficients are selected for the next step

4 In a similar fashion, an additional non-collinear descriptor scale is added to each descriptor subset selected in the previous step, and the respective $(n + 1)$-parameter regression treatment is performed If the Fisher criterion at the given probability level $F$, is now smaller than for the best $n$-parameter correlation found in the previous step, the latter is chosen as the final result and the program proceeds to the printout section (step 5) Otherwise, the descriptor subsets with the highest regression correlation coefficients are selected, and the procedure described in Step 4 is repeated with $n = n + 1$

5 The final equation, used in the following prediction section of the program, is selected on the basis of the maximum Fisher criterion and the highest cross-validated correlation coefficient The cross-validation technique is carried out as follows (*i*) for each experimental data point multi-linear regression is recalculated with the same descriptors for the data set without this point, (*ii*) the obtained regression equation is then used to predict the value of this data point, and (*iii*) finally, the obtained array of predicted data points is linearly correlated with the array of experimental data points providing a cross-validated

correlation coefficient. Thus, the cross-validation technique provides an estimation of the stability of the obtained regression model, *i.e.* the sensitivity of the model to the elimination of any single data point.

All criteria involved in the statistical treatment can be changed if so desired, in order to maximize the effectiveness of the search.

## 8 Overview of Some QSPR Results Obtained

We now present some results concerning the correlation of structure and property in organic compounds. These findings were recently obtained using the CODESSA program. The goal of these studies is to find both the best regression model for the prediction of properties, and also to determine which types of descriptors are the most sensitive to each of the various properties.

### 8.1 Response Factors

Returning to the calculation of response factor (RF) values, Figure 3a illustrates the results obtained from a QSPR treatment of a set of 152 diverse organic compounds.[36] Use of the extended set of descriptors developed in the CODESSA program allowed us to improve the correlation significantly over the PLS analysis.[4] The regression presented in Figure 3a with only 6 parameters (one third of the number used in the PLS treatment) has a better correlation coefficient ($R^2 = 0.892$ *vs.* $R^2 = 0.840$). 98% of the observed values were found to be within a 95% confidence interval for values predicted using the six-parameter equation (Table 2a).

The data set for gas chromatographic response factors and retention times contained structures belonging to classes of organic compounds very diverse in their chemical nature: alkanes, aromatic compounds, ethers, carbonyl compounds, carboxylic acids, alcohols, aldehydes, *etc.*

The measured response factor corresponds to the 'response' of the compound in a flame ionization detector (FID). Flame ionization is a multi-step process involving thermal decomposition of a compound with subsequent 'chemi-ionization'. Therefore, the yield of this process depends on the chemical nature of the molecules and the atoms from which they are constructed. As expected from this qualitative reasoning, both types of descriptors were involved in obtaining the correlation. The most important descriptor is the relative weight of 'effective' carbon atoms in the molecule, which has precedence given that only non-oxidized carbon atoms effectively produce a response in the FID. In this study an 'effective' carbon atom was defined as one connected only to other carbon or hydrogen atoms. The relative number of 'effective' carbon atoms is also important for the same reasoning. The thermal cracking of a compound inside the flame of an FID begins with the weakest C–X bond (where X is any atom). This being the case, the minimum total bond order of a carbon atom is obviously an important descriptor. The total molecular one-centre electron–electron repulsion also proved to be an important descriptor. This quantity summarizes the repulsion of electrons in the atoms constituting the molecule, and is probably related to the tendency of thermally cracked products to undergo 'chemi-ionization'.

### 8.2 Retention Times

A QSPR study of retention time data for the same set of 152 diverse organic compounds yielded a highly successful correlation (Figure 3b).[36] The best six-parameter equation obtained ($R^2 = 0.959$, Table 2b) is stable (cross-validated correlation coefficient $R^2_{cv} = 0.955$) and could be used, with considerable confidence, for the prediction of a retention time for an unknown compound. The few outliers (hexamethylbenzene, fluorene, 2-isopropoxyphenol, 1-methyl-2-pyridone, and methyl phenyl sulfoxide) do not belong to any recognizable class of compounds and seem to be random.

The most important descriptors in this correlation are the *α-polarizability* of the molecule and the *minimum valency of an H atom* in the compound. These quantum chemical indices can be considered to be related to the intermolecular interaction between the molecule studied and the gas chromatographic medium. The *α-polarizablity* of the compound characterizes the effectiveness of its intermolecular induction and dispersion interaction with the medium. The positive value of the respective regression coefficient is in accordance with physical considerations – compounds with higher polarizabilities have stronger interactions with the medium and thus higher retention values. The *minimum valency of an H atom* characterizes the compound

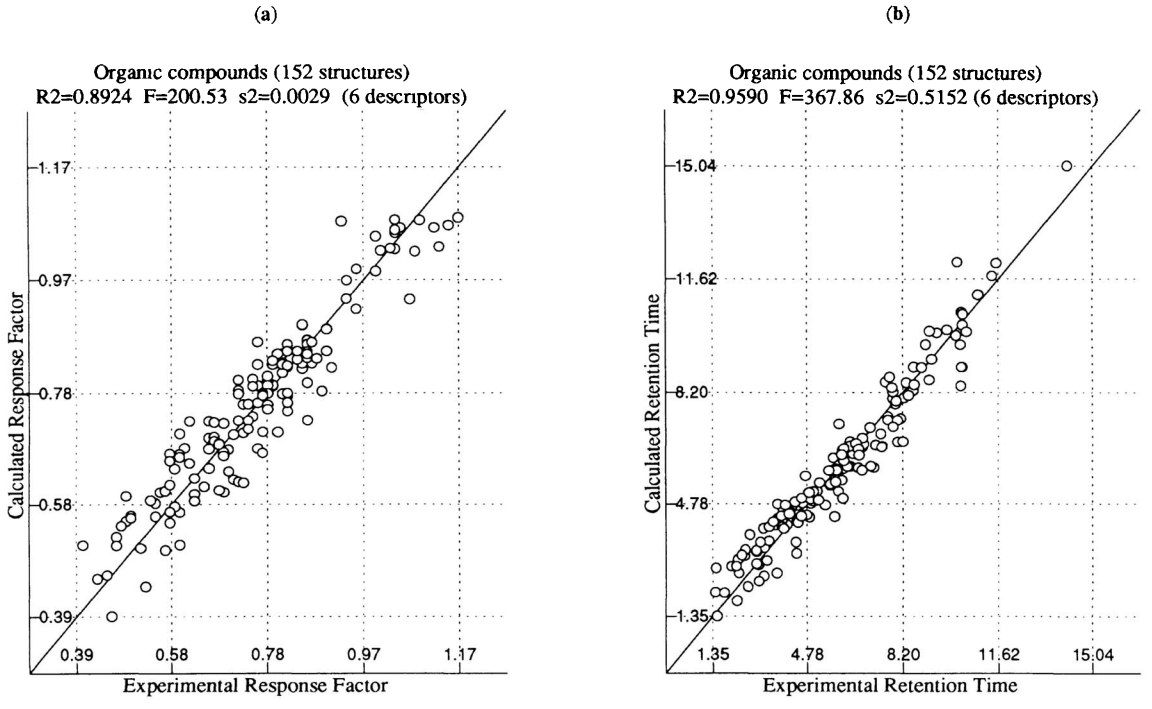



**Figure 3** Plots of the calculated *vs.* experimental response factor (a) and retention time (b) values for the diverse set of 152 organic compounds

**Table 2a** Details of the correlation developed for the response factor data

| Descriptor | Regression coefficient | *t*-criterion | $R^2$ |
|---|---|---|---|
| Intercept | − 2 327 | − 6 41 | |
| (1) Relative weight of 'effective' C atoms | − 0 9581 | 20 76 | 0 4655 |
| (2) Total molecular one-centre electron–electron repulsion energy | − 0 002889 | 11 34 | 0 7522 |
| (3) Relative number of 'effective' C atoms | − 1 160 | 11 34 | 0 7776 |
| (4) Minimum total bond order (> 0 1) of a C atom | − 0 2060 | 10 60 | 0 8245 |
| (5) Minimum valency of an H atom | 3 316 | 8 83 | 0 8869 |
| (6) Total hybridization component of the molecular dipole | − 0 0304 | 2 73 | 0 8924 |

**Table 2b** Details of the correlation developed for the retention time data

| Descriptor | Regression coefficient | *t*-criterion | $R^2$ |
|---|---|---|---|
| Intercept | 26 50 | 6 942 | |
| (1) Relative number of C−H bonds | − 6 91 | 9 269 | 0 1229 |
| (2) Total entropy of the molecule at 300 K divided by a number of atoms | − 0 871 | 8 543 | 0 6969 |
| (3) α-Polarizability | 0 04624 | 8 389 | 0 9064 |
| (4) Molecular weight | 0 01873 | 5 869 | 0 9397 |
| (5) Minimum valency of an H atom | − 21 55 | 5 362 | 0 9539 |
| (6) Maximum atomic orbital electronic population | 0 929 | 4 256 | 0 9590 |



as a hydrogen-bonding donor Therefore, the presence of this term in the correlation indicates the importance of hydrogen bond formation between the compound studied and the GC medium The negative value of the respective regression coefficient is expected (compounds with a lower value for minimum valency have stronger hydrogen bonds and, correspondingly, longer retention times)

### 8.3 Boiling Points

We studied the boiling point data for a set of 85 substituted pyridines provided by Reilly Industries, Inc [37] Among the substituents were methyl-, ethyl-, amino-, carboxamido-, cyano-, chloro-, carbonitrile-, and hydroxy-groups Preliminary component analysis of the data set revealed clustering of the compounds into two distinctive groups An examination of the compound distribution suggested hydrogen bonding as the most probable reason for such clustering Indeed, all compounds containing hydroxy-, amino-, and carboxamido-substituents fell into one group and the remainder into the other Hydrogen bonding is expected to lead to associated liquids, and therefore the boiling points of these structures predicted by the equation derived for non-associated compounds should be significantly lower than the corresponding experimental values This was demonstrated by our treatment, the reduced set of 63 non-associated compounds produced a good correlation ($R^2$ = 0 927) with only two descriptors the *gravitation index* calculated for all bonds and the *total point-charge component of the molecular dipole moment* As expected, the boiling points predicted for the remaining 22 structures were lower than the experimental values

A good multilinear regression model was obtained for all 85 compounds for which data were available using six structural parameters ($R^2$ = 0 948, Figure 4a) In the final model (Table 3a) two descriptors were the same as previously mentioned the *gravitation index (all bonds)* and *total point-charge component of the molecular dipole* which characterize the general relationshp between the chemical structure and the boiling point Two descriptors, the *hydrogen acceptors surface area* and the *relative-negative charged surface area*, are solely related to hydrogen bonding and adjust the model to describe associated structures The two remaining descriptors, the *minimum total bond order of an N atom* and the *average atomic nucleophilic reactivity index for an N atom*, describe the availability of a nitrogen lone electron pair for intermolecular hydrogen bonding

Considering that high boiling points are difficult to measure directly and are usually measured under reduced pressure then corrected to atmospheric pressure, the correlation equation obtained provides a successful fit to the data

**Table 3a** Details of the correlation developed for the boiling point data

$R^2$ = 0 948, $R^2_{cv}$ = 0 876, $F$ = 238 7, $s$ = 13 8, $n$ = 85

| Descriptor | Regression coefficient | *t*-criterion |
|---|---|---|
| Intercept | − 309 3 | − 11 8 |
| (1) Hydrogen acceptors surface area | 6 463 | 22 8 |
| (2) Gravitation index (all bonds) | 0 265 | 13 9 |
| (3) Minimum total bond order (> 0 1) of an N atom | 186 0 | 14 7 |
| (4) Total point-charge component of the molecular dipole | 20 69 | 10 4 |
| (5) Relative negative charged surface area | − 27 82 | − 9 8 |
| (6) Average atomic nucleophilic reactivity index for an N atom | 2680 2 | 6 4 |

**Table 3b** Details of the correlation developed for melting point data

$R^2$ = 0 857, $R^2_{cv}$ = 0 843, $F$ = 133 6, $s$ = 36 1, $n$ = 141

| Descriptor | Regression coefficient | *t*-criterion |
|---|---|---|
| Intercept | − 61 40 | − 0 6 |
| (1) Fractional hydrogen acceptors surface area | 525 8 | 19 9 |
| (2) Maximum atomic force constant | 14 43 | 8 2 |
| (3) Maximum atomic orbital electronic population | 244 7 | 6 3 |
| (4) Average structural information contents of the second order | − 61 61 | − 5 2 |
| (5) HOMO–LUMO energy gap | − 38 10 | − 4 7 |
| (6) Total hybridization component of the molecular dipole | 37 02 | 4 2 |

### 8.4 Melting Points

We also searched for correlations of melting points of substituted pyridines [37] The melting point is a difficult property to

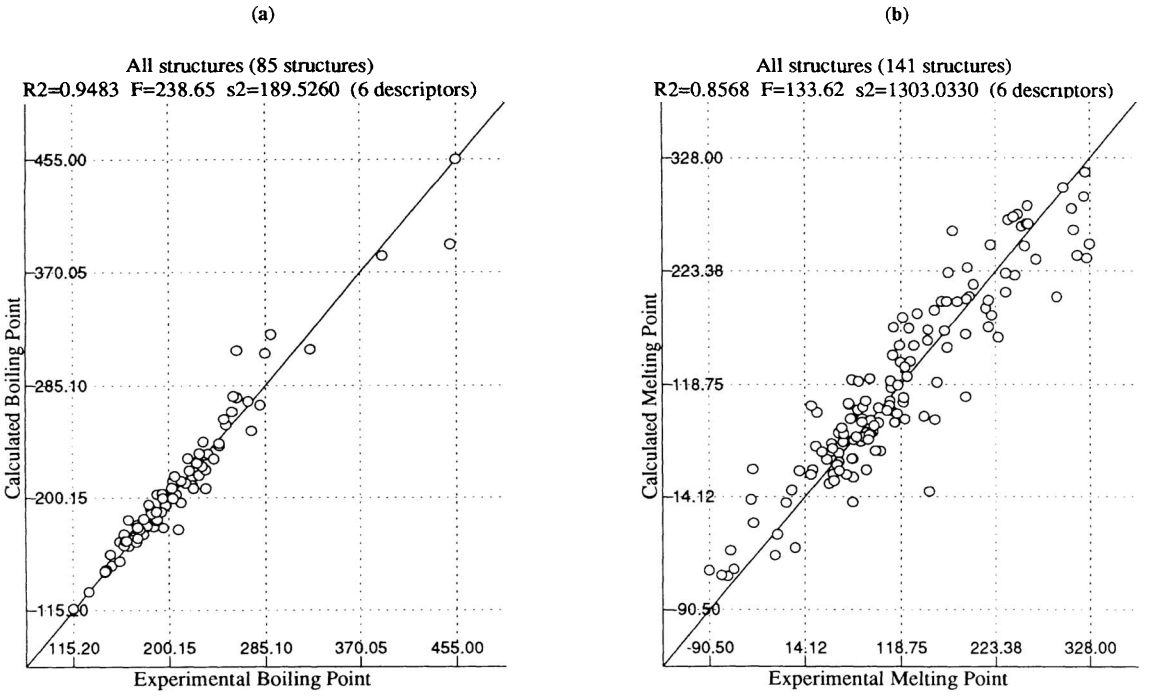



**Figure 4** Plots of the calculated *vs.* experimental boiling points (a) and melting points (b) for the two sets of substituted pyridines

correlate because singular molecular descriptors do not satisfactorily describe many-body crystal packing effects and intermolecular forces in condensed media. However, the multiple linear regression analysis of the melting points for a set of 141 compounds for which data were available resulted in a satisfactory six parameter correlation equation ($R^2$ = 0.857, Figure 4b). Again, one of the most important descriptors in the correlation equation (Table 3b) was a hydrogen bonding specific descriptor: *fractional hydrogen acceptors surface area*. This parameter directly represents the ratio of the surface area of the hydrogen acceptor atoms to the total surface area of the molecule. The other descriptors involved in this correlation were physically more diverse than those in the correlation of boiling points. This is not surprising as the melting point is expected to depend on more subtle intermolecular interactions in condensed phases than the boiling point. However, three descriptors in the six-parameter equation obtained for the melting points (the *maximum atomic force constant*, the *maximum atomic orbital electronic population* and the *total hybridization component of the molecular dipole*) can be related to the intermolecular interactions in condensed media (charge-transfer and dipole–dipole interactions). The *average structural information contents of the second order* reflects the number of different structural fragments in the molecule and may therefore be related to the details of the crystal lattice packing. Notably, one of the main factors in this correlation was the *HOMO–LUMO energy gap*. For insulators, such as solid state substituted pyridines, this quantity can be related to the energy gap between the valence band and the empty band. The negative sign of the respective regression coefficient implies that solids with a smaller band gap are more resistant to disordering (melting).

Hydrogen bonding obviously has a significant effect on the melting point, although the distinctive clustering of compounds according to association capability (as in the case of the boiling point) was not observed.

### 8.5 Partition Coefficients

A good correlation ($R^2$ = 0.943) was obtained with the octanol–water partition coefficient of 71 substituted pyridines (Figure 5).[37] The equation obtained (Table 4) is quite successful, bearing in mind the great variety of functionality: amino, alkyl, amidio,

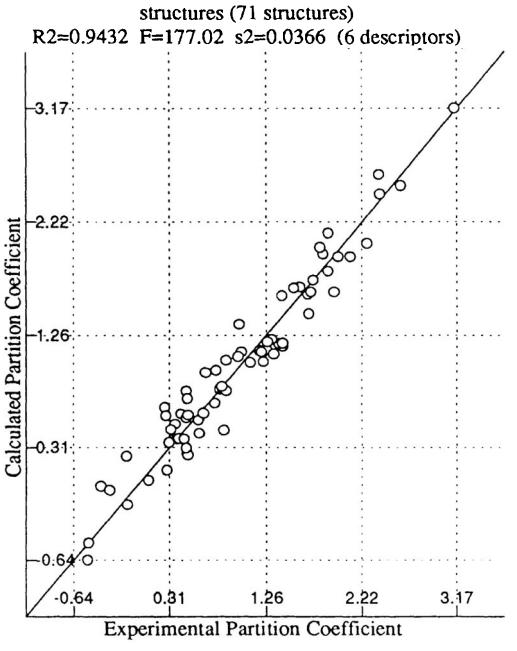



**Figure 5** Plot of the calculated *vs* experimental octanol–water partition coefficient values for the set of substituted pyridines

nitro, hydroxy, cyano, and thio groups, halogen atoms, ethers, esters, and aromatic rings. Moreover, both solids and liquids (at room temperature) were represented in the set. *Kier and Hall valence connectivity index of the zeroth order* and the *number of double bonds* proved to be the most significant descriptors for the set of structures under investigation. The fact that similar descriptors have been reported to correlate with partition coefficients of different compounds[38] suggests that this correlation model has wider applications. The other descriptors are directly related to the properties of the N atom in the pyridines. This atom is obviously acting as a hydrogen bonding acceptor and the appearance of the respective descriptors in the best correlation reflects the variance of the hydrogen-bond accepting ability of different pyridines in water and in octanol. The correlation

**Table 4** Details of the correlation developed for octanol–water partition coefficient data

$R^2 = 0\ 943$, $R^2_{cv} = 0\ 929$, $F = 177\ 0$, $s = 0\ 19$, $n = 71$

| Descriptor | Regression coefficient | *t*-criterion |
|---|---|---|
| Intercept | − 20 23 | − 10 1 |
| (1) Kier & Hall valence connectivity index of zeroth order | 0 567 | 22 4 |
| (2) Number of double bonds | − 1 046 | − 18 4 |
| (3) Minimum resonance energy for a C−N bond | 0 225 | 9 7 |
| (4) Maximum valency of an N atom | 3 033 | 9 2 |
| (5) Number of N atoms | − 0 454 | − 6 9 |
| (6) Minimum electron–electron repulsion for an N atom | 0 0433 | 6 5 |

obtained for log*P* can be of practical significance, as this quantity is of major importance in predicting the biological activity of chemical compounds

## 9 Future Perspectives

We believe that advanced software for QSAR/QSPR should include all of the various types of molecular descriptors since there is no evidence that one particular class of descriptors should necessarily predominate in regression models The development of new descriptors will obviously continue in new areas of QSPR applications, such as the treatment of polymers and mixtures, as well as in attempts to describe temperature-dependencies With the involvement of quantum-chemical calculations providing a vast amount of information regarding structure, the development of descriptor definition language so that new descriptors can be generated instantly appears feasible

The search for effective procedures to find the best correlation between molecular descriptors and target performance dominates QSAR/QSPR research New methods include principal component analysis and nonlinear regression analysis A promising alternative to the correlation equation is the use of neural networks Properly built and trained neural networks provide effective predictive power

The most challenging problem in QSAR/QSPR research is the so called 'Inverse Problem' (targeted search for the compound(s) with a prescribed value of molecular property or biological activity), which is attracting more and more attention from computational chemists around the world Prerequisites to significant progress in this problem are a good data set, powerful QSAR/QSPR software, a reliable regression model, and descriptors which can elucidate structural features endowing targeted property/activity One of the first phases in solving this problem would be to create descriptor databases and allow computers to search for structures with a targeted property value, once a correlation model has been established

## 10 References


1 M Siskin and A R Katritzky, *Science*, 1991, **254**, 231
2 A R Katritzky, A R Lapucha, R Murugan, F J Luxem, M Siskin, and G Brons, *Energy Fuels*, 1990, **4**, 493
3 A R Katritzky, R A Barcock, M Siskin, and W N Olmstead, *Energy Fuels*, 1994, **8**, 990
4 G Musumarra, D Pisano, A R Katritzky, A R Lapucha, F J Luxem, R Murugan, M Siskin, and G Brons, *Tetrahedron Comp Method*, 1989, **2**, 17
5 V A Palm, 'Fundamentals in Quantitative Theory of Organic Chemistry', Khimiya, Leningrad, 1967 (in Russian)
6 L P Hammett, *Chem Rev*, 1935, **17**, 125
7 R W Taft, in 'Steric Effects in Organic Chemistry', ed M S Newman Wiley, New York, 1956, p 556
8 A Leo, C Hansch, and C Church, *J Med Chem*, 1969, **12**, 766
9 C Hansch and E Coats, *J Pharm Sci*, 1970, **59**, 731
10 C Hansch, in 'Drug Design', ed E J Ariens, Academic Press, New York, 1971, Vol 1, Chapt 2
11 H Wiener, *J Am Chem Soc*, 1947, **69**, 17
12 M Randic, *J Am Chem Soc*, 1975, **97**, 6609
13 L B Kier and L H Hall, 'Molecular Connectivity in Chemistry and Drug Research', Academic Press, New York, 1976
14 K Osmialowski, J Halkiewicz, and R Kaliszan, *J Chromatogr*, 1986, **361**, 63
15 D T Stanton and P C Jurs, *Anal Chem*, 1990, **62**, 2323
16 D T Stanton, L M Egolf, P C Jurs, and M G Hicks, *J Chem Inf Comput Sci*, 1992, **32**, 306
17 R H Rohrbaugh and P C Jurs, *Anal Chim Acta*, 1987, **199**, 99
18 H Sklenar and J Jager, *Int J Quantum Chem*, 1979, **16**, 467
19 R E Brown and A M Simas, *Theoret Chim Acta (Berl)*, 1982, **62**, 1
20 L Buydens, D L Massart, and P Geerlings, *Anal Chem*, 1983, **55**, 738
21 A Cartier and J -L Rivail, *Chemom Intell Lab Sys*, 1987, **1**, 335
22 O Kikuchi, *Quant Struct -Act Relat*, 1987, **6**, 179
23 M Cocchi, M C Menziani, P G De Benedetti, and G Cruciani, *Chemom Intell Lab Sys*, 1992, **14**, 209
24 B W Clare, *Theoret Chim Acta*, 1994, **87**, 415
25 E V Gordeeva, A R Katritzky, V V Shcherbukhin, and N S Zefirov, *J Chem Inf Comput Sci*, 1993, **33**, 102
26 M J S Dewar, E G Zoebisch, E F Healy, and J J P Stewart, *J Am Chem Soc*, 1985, **107**, 3902
27 J J P Stewart, MOPAC 6 0, QCPE No 445, 1990
28 AMPAC 5 0, Semichem, 1994, Summit, Shawnee, KS 66216
29 R H Mayers, 'Classical and Modern Regression with Applications', Duxbury Press, Boston, 1986
30 'Chemometrics, Mathematics, and Statistics in Chemistry', ed B R Kowalski, D Reidel, Dordrecht, 1984
31 I T Jolliffe, 'Principal Component Analysis', Springer-Verlag, New York, 1986
32 M Meloun, J Militky, and M Forina, 'Chemometrics for Analytical Chemistry', Vol 1, Ellis Horwood, New York, 1992
33 C Hansch and T Fujita, *J Am Chem Soc*, 1964, **86**, 1616
34 T Fujita and C Hansch, *J Med Chem*, 1967, **10**, 991
35 A R Katritzky, V S Lobanov, M Karelson, R Murugan, M P Grendze, and J E Toomey, Jr, *Chemom Intell Lab Sys*, 1995, in press
36 A R Katritzky, E S Ignatchenko, R A Barcock, V S Lobanov, and M Karelson, *Anal Chem*, 1994, **66**, 1799
37 R Murugan, M P Grendze, J E Toomey, Jr, A R Katritzky, M Karelson, V S Lobanov, and P Rachwal, *CHEMTECH*, 1994, **24**, 17
38 M St J Warne, D W Connell, D W Hawker, and G Schuurmann, *Chemosphere*, 1990, **21**, 877