**PERSPECTIVE**

# The Characterization of Chemical Structures Using Molecular Properties. A Survey

David J. Livingstone†

ChemQuest, Delamere House, 1 Royal Crescent, Sandown, Isle of Wight PO36 8LZ, U.K., and Centre for Molecular Design, University of Portsmouth, Portsmouth, Hants PO1 2DY, U.K.

Received October 26, 1999

## INTRODUCTION

In a review of structure−property correlations in molecular design,[1] the relevance today of what is arguably one of the earliest publications[2] on QSAR was pointed out. In this paper Crum Brown and Frazer wrote that "physiological action", $\phi$, was a function of "chemical constitution", $C$, using the equation shown below:

$$\Phi = f(C) \tag{1}$$

The major problem in obtaining an accurate definition of $f$ in the equation was attributed to the difficulty of expressing changes in $\phi$ and $C$ with sufficient "definiteness". This situation may have changed now with regard to biological responses, but the problem of the characterization of changes in chemical structure remains. Despite 130 years of research into the experimental and theoretical investigation of chemical structure and physicochemical properties, there is still no agreement on what constitutes the "best" set of descriptors for molecular design. This, of course, may be the wrong question since different classes of descriptor encode different classes of information, and although their champions may argue otherwise it is likely that the "best" set will be comprised of parameters from a number of different sources.

This review sets out to survey the main classes of physicochemical properties that are used to characterize chemical structures. The primary purpose of this characterization is "molecular design", that is to say, the discovery of "performance" chemicals for use as pharmaceuticals, agrochemicals, fragrance ingredients, dyestuffs, flavorings, and so on. There are other applications of such information, such as the investigation of chemical and biochemical reaction mechanisms, but the important features of these properties should be just as significant to these uses. There are also other ways in which differences in chemical structure may be expressed, the presence or absence of substructural fragments, for example, but these approaches are not considered here.[3] The division of properties into the following sections is somewhat arbitrary but has been made in an attempt to organize them into a sensible number of subdivisions.

## PARAMETERS FROM PHYSICAL ORGANIC CHEMISTRY

These are the properties used in the method that became the mainstream technique in quantitative drug design—The Hansch approach.[4] The QSAR technique that became known as the "Free and Wilson"[5] method had been published some years previously in 1956[6] although this describes chemical structure in terms of the contributions that fragments make to activity, rather than physicochemical descriptors. While never as popular as the Hansch approach, the Free and Wilson method has a place in the QSAR toolbox as reviewed by Kubinyi.[7] Relationships involving theoretical computed properties were also published at around the same time as Hansch's seminal work: a correlation between antibacterial potency and superdelocalizability[8] and a relationship between hallucinogenic activity and Ehomo,[9] for example. The general scarcity of adequate computing resources, their low performance, and the lack of easily operated software combined to restrict the widespread use of theoretical descriptors until the 1980s.

The rationale behind the use of physical organic chemistry parameters in drug design is the same as that for the study of chemical reactions in general. A simple chemical model system is designed so that measurements of some "feature" of that model, an equilibrium constant for example, are expected to yield information concerning a specific physicochemical characteristic. Values of this "feature" can then be used as descriptors of that physicochemical characteristic when attempting to explain or understand how a more complicated chemical process is controlled.

The standard physicochemical descriptors can be divided into three main categories with a "miscellaneous" group encompassing quite a wide variety of properties and chemical model systems. The following subsections describe some of the more important features of these parameters.

**Hydrophobic Parameters.** Although the electronic and steric descriptors appear to predate hydrophobic parameters ($\sigma$, 1935[10] and $E_s$, 1952,[11] see below) in fact this is not true as some of the earliest reports of correlations between biological activity and structure involved hydrophobic descriptors. Overton[12] and Meyer[13] independently described local anaesthetics in terms of oil/water partition coefficients. The selection of octanol/water partitioning as a chemical model system for hydrophobicity was one of Hansch's major contributions to the field, and hydrophobic descriptors continue to play a major part in successful QSAR studies. A survey of all the QSAR models reported in the QSAR journal in 1988,[14] both papers and abstracts, showed that more than 40% involved some hydrophobicity descriptor. A similar survey on the 1998 journal showed that this had risen to over 50%.

The archetypal hydrophobicity parameter is $\pi$, a substituent constant derived from measurements of octanol water parti-

† E-mail: davel@chmqst.demon.co.uk. Web: www.chmqst.demon.co.uk.

10.1021/ci990162i CCC: $19.00 
Published on Web 12/31/1999

**Table 1.** Some Commercially Available Computer Programs for log *P* Calculation[a]

| program[b] | platform | calculation method[c] | supplier[d] |
|---|---|---|---|
| ClogP | Mac, PC | fragmental-HL | Biobyte |
| PCMODELS | Unix | fragmental-HL | Daylight |
| PROLOGP | PC | fragmental-R | Compudrug |
| SYBYL* | Unix | fragmental-R | Tripos |
| KLOGP | PC | fragmental-C | Biosoft |
| LOGKOW | PC | fragmental-A/F | Syracuse |
| ACD/LogP | PC | fragmental-A/F | Advanced Chemistry Development Inc. |
| AutologP | PC | properties | CTIS |
| VLOGP* | Vax, PC | properties | Health Designs |
| SCILOGP | PC | properties | SciVision |
| CLIP | Unix | properties | University of Lausanne |
| HINT | VAX | properties | EduSoft |
| CERIUS[2]* | Unix | atomic values | MSI |
| ATOMIC5 | PC | atomic values | Compudrug |
| CHEMICALC-2 | Vax, PC | atomic values | QCPE |
| TSAR* | Unix | atomic values | Oxford Molecular |

[a] For a recent comparison of different calculation approaches, see: Mannhold, R.; Dross, K. *Quant. Struct.-Act. Relat.* **1996**, *15*, 403−9. [b] These are stand-alone programs except those marked with *. [c] The fragmental methods refer to the system of Hansch and Leo (HL), Rekker (R), computer identified (C), and atom/fragment contributions (A/F). Properties means that various molecular properties are used in the calculations. Atomic values means that tables of atom-based values are used. [d] Supplier addresses are not given but can be easily obtained by Web search.

tion coefficients (log *P*) for a parent compound and substituted derivatives:

$$\pi_X = \log P_X - \log P_H \quad (2)$$

In eq 2 the parent is indicated by the subscript H and the substituent by X. The first series for which this parameter was derived was a set of monosubstituted phenoxyacetic acids,[15] but it soon became clear that $\pi$ values were not strictly additive across different parent series, due principally to electronic interactions, and it became necessary to measure $\pi$ values in other series such as substituted phenols, benzoic acids, anilines, and so on.[16] This proliferation of different $\pi$ series is analogous to the many different $\sigma$ scales that are used for various parent systems and reaction types (see below), and thus one of the major problems in the application of QSAR in the early 1970s was the selection of the "correct" scale for the description of both hydrophobic and electronic effects. This can be a difficult task even for a simple chemical reaction when, for example, the rate-determining step is not known. In the QSAR situation where the system is a complex mixture of biochemicals and enzymes (in vitro) or biochemicals, transport, and metabolism processes (in vivo) the choice is often arbitrary. A further complication arises in the choice of substitution position or even parent structure. Most real drug molecules are much more complex than the simple compounds used in the chemical model systems to characterize particular effects. These compounds are selected so that the assignment of effects to particular chemical fragments is unambiguous, but it is rarely obvious how these values should be assigned to bioactive structures.

The problem of the selection of an appropriate $\pi$ scale has been resolved by the increasing use of either whole molecule log *P* values or of "fragmental" values to describe portions of molecules. The fragmental values were produced as byproducts of the process of creating expert systems that could be used to predict log *P* values. The fragmental system of Nys and Rekker[17−19] was one of the earliest such systems to be developed and was based on a statistical analysis of a large number of measured partition coefficient values to give the "best" values for particular molecular fragments. This "reductionist" approach resulted in a large number of fragments with a small number of correction factors. The Rekker system was quickly followed by a "constructionist" approach due to Leo and Hansch.[20,21] This method was based on the use of a small number of fundamental fragments, derived from very precise log *P* measurements, with a correspondingly larger set of correction factors. Comparisons of the two techniques have concluded, perhaps not surprisingly, that each method gives better results for some sets of compounds than for others.[22−25] At first, the two methods were used to calculate log *P* values manually, and this had major drawbacks since not only was it labor intensive but it was also difficult to achieve consistency since even a relatively simple molecule may be broken into fragments in a number of ways. The Leo and Hansch technique was the first to be made available as an automated system in the program CLOGP, which used the elegantly simple SMILES notation for chemical structures.[26,27] Since then the SMILES system has been widely adopted as the basis of database systems and as a chemical structure entry system for a variety of chemical modeling programs. One other form of fragmental value for log *P* calculation deserves mention here, and that is the atom-based system of Ghose and Crippen.[28] These fragment values were devised specifically for use in three-dimensional (3D) QSAR studies and thus differ from the fragmental values of the Rekker or Hansch and Leo systems. Atom-based fragmental constants such as these and others are often employed in molecular modeling packages for the calculation of log *P* (see Table 1).

Partition coefficients and $\pi$ values have been shown to correlate with measures of biological activity in a very wide variety of experimental systems, ranging from simple protein binding to animal and human in vivo effects. This is presumably because hydrophobic effects are important not only in the intermolecular interactions that occur between a drug and its target site but also in the distribution of a compound within a biosystem and its interaction with competing binding sites. It may be questioned, however, whether octanol/water is the "right" model system for hydrophobic effects. That it has been successful is without question, but might not another model system be more

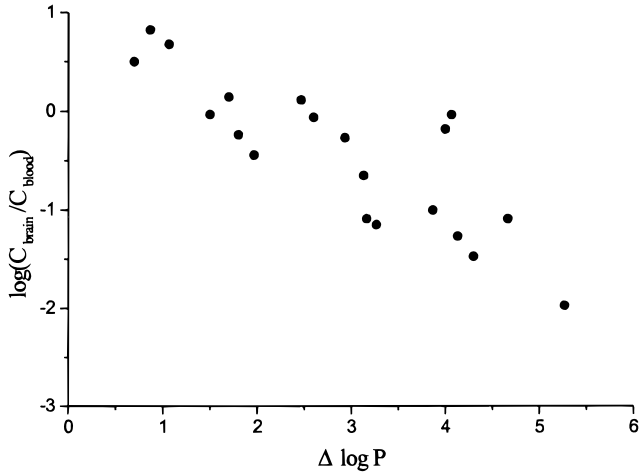


**Figure 1.** Relationship between brain penetration and Δ log *P* for 20 structurally diverse compounds (reprinted with permission from *J. Med. Chem.* **1988**, *31*, 656−71).

successful? Part of the answer to this is the fact that partition coefficients from different model systems may be modeled by the use of Collander[29] equations:

$$\log P_2 = a \log P_1 + c \tag{3}$$

This might explain how a single partition coefficient could be applicable to so many different mechanisms since the weighting coefficient implicit in the Collander relationship would be estimated as part of the mathematical modeling process. It has been shown, however, that partition coefficients from other systems do contain extra information, which is useful in the description of biological properties. Young and co-workers demonstrated[30] that the difference between octanol/water and cyclohexane/water log *P* values (Δ log *P*) could be used to explain brain penetration as shown in Figure 1. It was suggested that this parameter, first introduced by Seiler,[31] might be a useful general descriptor for brain penetration.[32] Extending this concept of the utility of other partitioning systems, Leahy and co-workers suggested that four model systems might be required in order to describe the properties of real membranes.[33] These models consist of water combined with the following:

(1) an amphiprotic solvent (e.g., octanol)
(2) an inert solvent (e.g., any alkane)
(3) a pure proton donor (e.g., chloroform)
(4) a pure proton acceptor

Propyleneglycol dipelargonate (PGDP) was proposed as a suitable compound for the pure proton acceptor, and 216 partition coefficient values were reported along with a calculation scheme for other values. It remains to be seen how useful such log *P* scales will be.

Perhaps because of the widespread acceptance of the value of log *P* as a predictor of biological activity, numerous investigations of this property have been reported and a variety of computer programs are available for log *P* calculation as shown in Table 1. There is extensive literature on the topic including several monographs[18,19,34,35] and a number of reviews dealing with its measurement,[36,37] calculation,[38] and general uses[39−42] including recent contributions from Rekker and co-workers that almost filled a complete issue of the QSAR journal.[43−45]

**Steric Properties.** At first sight the description of steric properties might appear to be the simplest task of all through the use of easily calculated descriptors such as molecular weight or molar volume. Unfortunately life, or at least chemistry, is not as simple as that since the term "steric" applies to bulk or size descriptors and also shape. It is also debatable, as will be seen later, as to what exactly constitutes "bulk". The first steric descriptor to be used in QSAR studies was the $E_s$ parameter due to Taft:[11,46]

$$E_s = \log(k/k_0)_A \tag{4}$$

Equation 4 compares the rate of acid-catalyzed hydrolysis of RCOOR′ ($k$) with that of the methyl-substituted parent $CH_3COOR'$ ($k_0$). Hancock modified $E_s$ values by correcting[47] for the number of α-hydrogen atoms, and Fujita expressed $E_s$ values for complex substituents as a weighted sum of their constituent fragments.[48] As in the case of $\pi$ values, $E_s$ is used by consulting look-up tables of substituent values and herein lies the problem in the use of all such tables, and that is missing values. This is particularly important in the case of $E_s$ since there are a number of common substituents that themselves are unstable under the conditions of acid hydrolysis and thus cannot be experimentally determined. Charton addressed this problem by demonstrating[49] that $E_s$ is related to the van der Waals' radii of substituents, and Kutter and Hansch used this to formulate[50] $E_s$ values for a variety of new substituents. $E_s$ values have been criticized on the grounds that they contain electronic information as well as measuring bulk.

A more commonly used descriptor of steric properties, both for substituents and whole molecules, is molar refractivity, $M_R$, as defined by the Lorentz−Lorenz equation:[51]

$$M_R = \frac{n^2 - 1}{n^2 + 2}\frac{M_W}{d} \tag{5}$$

In eq 5 $n$ is the refractive index, $d$ is the density, and $M_W$ is the molecular weight of a compound, normally a liquid. Molar refraction is an additive−constitutive property of molecules, and thus fragment values have been calculated for many common groups of atoms. $M_R$ values can be automatically calculated for molecules using a variety of computer programs (e.g., CLOGP), and it has found numerous applications in reported QSAR studies. Since eq 5 contains molar volume ($M_W/d$) $M_R$ is clearly related to volume and size, but the refractive-index-related correction term in eq 5 accounts for polarizability. Thus, $M_R$ is not a "pure" size or bulk descriptor although it has been shown to correlate with van der Waals' volume, parachor (a surface-tension-weighted molar volume[52]), and fragmental volume constants.[23] Atom-based fragmental constants have been devised[53] for the calculation of $M_R$ values, in much the same way as the atom-based fragmental constants for log *P*, and these are often used in molecular modeling packages.

One of the problems with using single descriptors of size such as $E_s$ or $M_R$ is that there is no discrimination between substituents that have different shapes; the $M_R$ values of butyl and *tert*-butyl are 19.61 and 19.62 in the Hansch and Leo compilation,[21] for example. Recognizing this, Verloop and co-workers devised[54,55] the STERIMOL parameters based on models of compounds or substituents using standard bond

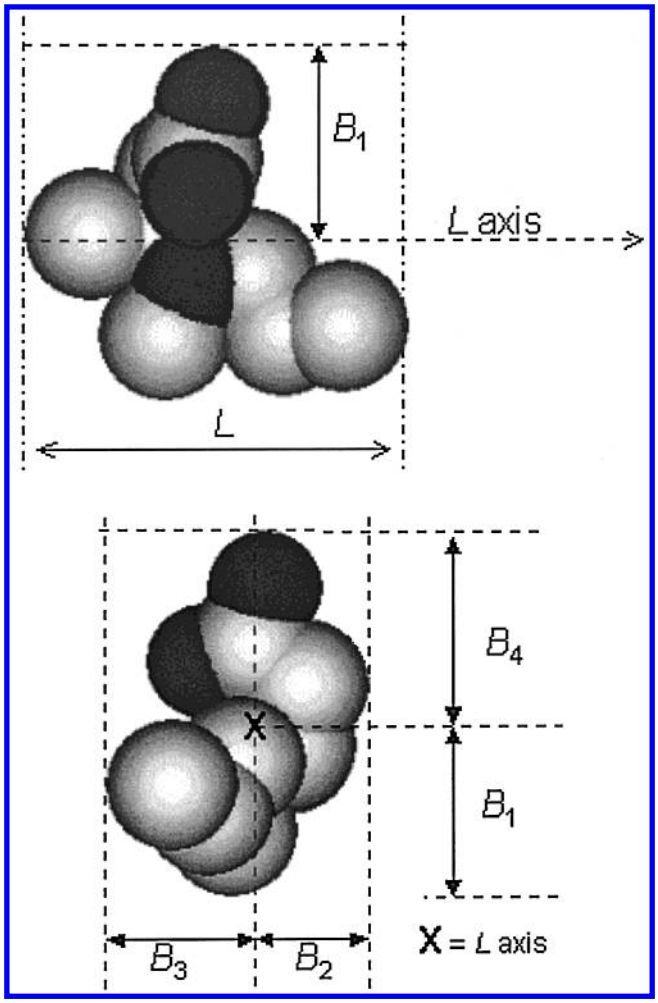


**Figure 2.** Definition of the original STERIMOL parameters. (A, top) Viewed orthogonal to the point of attachment of the substituent. A box is placed around the substituent, and *L* is defined as the length of the box along the axis of attachment. (B, bottom) Viewed perpendicular to the axis of attachment (marked X). Four dimensions of the box, B1−B4 are defined as shown.

lengths and angles. Figure 2 shows how the original five STERIMOL parameters were defined. The length parameter, *L*, is defined as the length of the substituent along the axis of its substitution with the parent while the width parameters, *B*, are orthogonal to *L* and have angles of 90° with respect to one another. Originally the width parameters were ordered smallest to largest, $B_1$ to $B_4$, with a slightly different ordering scheme suggested by Skagerberg and co-workers.[56] Because of the large number of parameters required to characterize each substituent position, and the corresponding requirement for a large number of training set compounds, it was later suggested that the width parameters could be reduced to just two.[57] These are $B_1$, the smallest width parameter, and $B_5$ the largest width orthogonal to *L* but independent of its angle to $B_1$.

This and other ingenious approaches to the problem of the treatment of shape in QSAR went some way toward improving correlations with biological responses, but the next major step forward involved the elaboration of 3D QSAR techniques as described in a later section.

**Electronic Effects.** The Hammett equation defines[58] the electronic effect of a substituent on a reaction using the simple equation shown below:

$$\log K_X - \log K_H = \rho\sigma_X \quad (6)$$

In eq 6 $K_X$ represents an equilibrium (or rate) constant for a process involving an X-substituted compound with $K_H$ the corresponding constant for the unsubstituted parent. The two parameters on the right-hand side of the equation are the reaction constant, $\rho$, and the substituent constant, $\sigma$. The ionization of benzoic acids was taken as a reference system for which the reaction constant, $\rho$, was defined as 1.0. Substituent $\sigma$ values were then computed from measured p$K_a$ values of substituted benzoic acids.

These original $\sigma$ values were shown to explain many different types of chemical rates and equilibria, using a variety of $\rho$ values, but there were of course always failures. These failures were variously ascribed to particular interactions between the substituent and the reaction center and led to the establishment of alternative $\sigma$ scales for particular substituents or reaction conditions.[59] Part of the problem in the description of the electronic effects of substituents lies in the fact that these effects may be transmitted through field (inductive) effects and resonance effects. There has been constant debate among physical organic chemists ever since, leading not only to a considerable number of different $\sigma$ scales but also to many extensions and enhancements of the original Hammett treatment.[60] This multiplicity of $\sigma$ scales makes the task of choosing the right one for a biochemical or other complex chemical application very difficult indeed.

In an attempt to stop the proliferation of different $\sigma$ scales, Swain and Lupton proposed[61] that any set of $\sigma$ values could be expressed by a weighted linear combination of two components, $\mathcal{T}$ and $R$, which they termed the field and resonance components, respectively. Equations 7 and 8 show their correlations for $\sigma_m$ and $\sigma_p$:

$$\sigma_m = 0.60(\pm 0.00)\mathcal{T} + 0.27(\pm 0.00)R + 0.00(\pm 0.00) \quad (7)$$

$$(n = 42, \quad r = 1.00, \quad s = 0.00)$$

$$\sigma_p = 0.56(\pm 0.00)\mathcal{T} + 1.0(\pm 0.00)R + 0.00(\pm 0.00) \quad (8)$$

$$(n = 42, \quad r = 1.00, \quad s = 0.00)$$

Like much of the work in this field, these $\mathcal{T}$ and $R$ scales also prompted much debate including a mini-symposium in print in the *Journal of Organic Chemistry*[62−65] following a revised treatment published in 1983.[66] Whatever the merits of these arguments, the field and resonance components of $\sigma$, now restyled *F* and *R*, provided a useful simplification of the various $\sigma$ scales for use in QSAR. A compilation of electronic substituent constant data, including redefined $\mathcal{T}$ and $R$ values, has been published for 530 substituents.[67]

A major problem in the use of electronic substituent constants, indeed in the use of any substituent constants, lies in the definition of "parent", substituents, and even substituent positions. This is particularly so for electronic parameters where the transmission of effects can be heavily dependent on substituent position. As a result of this and because of the increasing need to consider noncongeneric series (see later) there has been an increasing trend to use whole molecule descriptors, either measured, calculated, or theoretical.

**Miscellaneous. (a) Experimental Properties.** There are many obvious properties that could be used in attempts to explain biological activity. Solubility, for example, was one of the earliest successful descriptors[68] where an inverse relationship was shown between toxicity and water solubility. Other experimentally determined properties have been shown to be useful, such as melting point,[69] NMR chemical shift,[70] infrared and Raman stretching frequency,[71] reactivity with 4-nitrophenol,[72] and even rate of hydrolysis in mouse plasma![73] There are problems, of course, in the use of experimental properties since these are usually not predictable and the purpose of most modeling studies of biological (and other) properties is prediction. There are ways in which experimental properties can be estimated; see, for example, the book by Lyman, Reehl, and Rosenblatt,[74] and some methods such as those for log *P* work very well indeed, but in general the lack of predictability is a distinct disadvantage.

**(b) Hydrogen Bonding.** The importance of hydrogen bonding in drug receptor interactions has long been known, and attempts to produce H bond donor[75,76] and acceptor[77] scales began in the late 1960s. Fujita and co-workers proposed[78] the use of indicator variables to account for hydrogen bonding, one for acceptors and one for donors, and Hansch and Leo included such indicator values in their 1979 substituent constants book.[21] Investigation of the formation of H bonds by diazepines using $^1$H NMR showed[79] correlations between the free energy of H bond interactions and the pharmacological activity of the compounds. The success of $\Delta\log P$ in explaining blood−brain barrier penetration has been ascribed to H-bonding capability, and it has been suggested that this can be estimated using measured log *P* values and calculated molar volumes.[80]

Recent work has resulted in a set of substituent scales for H bond donors and acceptors based[81,82] on experimental measurements, and a large database of measured H bond complexes along with a computer program for the estimation of unknown compounds is available.[83]

Unfortunately, despite the obvious recognition of the importance of hydrogen bonding, there has been no general acceptance of a set of H bond scales, either donor or acceptor, for use in QSAR. Perhaps this is because these properties are already accounted for by the use of experimental descriptors such as log *P*, as suggested by van de Waterbeemd and Kansy,[80] or because the relationships already obtained are of sufficient quality. The increased use of 3D QSAR methods, where such interactions may already be encoded, might also explain their lack of general use.

**(c) Linear Solvation Energy Relationships.** Biological properties are controlled to some extent by solute−solvent interactions. One approach to describing such interactions is a method known as linear solvation energy relationships (LSER) developed by Kamlet, Taft, Abraham, and co-workers.[84−86] The LSER technique describes a solute−solvent property (SSP) by means of the following equation:

$$\log \mathrm{SSP} = \log \mathrm{SSP}_0 + \text{cavity term} + \text{polarizability/dipolarity term(s)} + \text{hydrogen bonding terms} \quad (9)$$

A set of parameters has been developed for this equation based on spectral shifts, and they have thus been given the name solvatochromic parameters. Many different solute−solvent-related properties have been successfully correlated[87] with these descriptors, and an improved set measured by gas chromatography has also been reported.[88] Correlations with biological properties have been reported,[89] and there is little doubt that these descriptors offer a useful way of describing chemical structure in physicochemical terms. Like all approaches based on experimental measurements, however, problems arise when tabulated values are missing. Famini and co-workers have described[90,91] an interesting theoretical extension of this approach known as theoretical linear solvation energy relationships (TLSER). The TLSER equation for multiple solutes and a single solvent has the following form:

$$\log \mathrm{SSP} = \log \mathrm{SSP}_0 + aV_{\mathrm{mc}} + b\pi_{\mathrm{I}} + c\epsilon_{\mathrm{B}} + dq_- + e\epsilon_{\mathrm{A}} + fq_+ \quad (10)$$

The TLSER parameters are a cavity term ($V_{\mathrm{mc}}$), a polarizability term ($\pi_{\mathrm{I}}$), "covalent" hydrogen bond donor and acceptor basicities ($\epsilon_{\mathrm{B}}$ and $\epsilon_{\mathrm{A}}$), and their "electrostatic" equivalents($q_-$ and $q_+$). These quantities are calculated from volumes, charges, orbital energies, and so on.[92] Like the LSER approach, TLSER parameters have been shown to correlate various experimental properties as shown[93] in Table 2. It is interesting to see from this table how the different terms in the TLSER are important;[94] the equation for p$K_a$, for example, only involves the two H bond acceptor basicities. TLSER descriptors have also been shown to be useful in the description of biological data.[92,95]

**Table 2.** TLSER Correlations[a] for Experimental Systems (From Ref 93)

| property | *a* | *b* | *c* | *d* | *e* | *f* | R |
|---|---|---|---|---|---|---|---|
| charcoal absorption | 2.46 | 5.78 | n/s | −4.66 | n/s | n/s | 0.955 |
| HPLC retention | 1.83 | n/s | n/s | −3.05 | n/s | −5.52 | 0.980 |
| log *P* | 3.14 | n/s | n/s | −5.92 | n/s | n/s | 0.974 |
| hydrolysis rate constant | n/s | −13.1 | n/s | n/s | −6.11 | n/s | 0.990 |
| p$K_a$ | n/s | n/s | n/s | n/s | 127 | −41.1 | 0.941 |
| electronic absorption | n/s | −37 617 | n/s | 1945 | n/s | 7613 | 0.972 |

[a] The columns *a*−*f* represent the values of the regression coefficients shown in the TLSER equation, n/s means nonsignificant, and *R* is the correlation coefficient.

## CALCULATED PROPERTIES

The move away from the use of "traditional" substituent constants has been prompted by the increasing application of QSAR techniques to noncongeneric series of molecules. In the early applications of QSAR, it was found necessary to consider essentially only congeners in a QSAR model for two reasons. First, drug design was seen (quite reasonably) as an extension of physical organic chemistry, and so the procedures applied to the study of chemical properties, i.e., the examination of the properties of congeners, were also applied to these problems. Second, the way in which drug molecules were characterized was by means of tables of substituent constant data. One notable exception to this was the tables of atom contributions to molar refractivity, but apart from this all characterization was by means of "look up". The tables were generally designed to apply to series of congeners, most often aromatic series at that, and so QSAR was restricted to congeneric series.

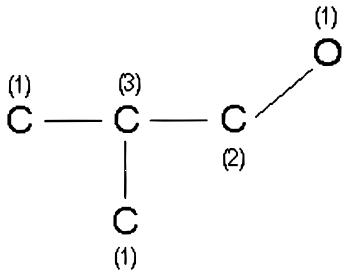


**Figure 3.** Example of degree of connectivity for isobutyl alcohol.

Another disadvantage of tabulated data sources is the question of missing data. In this situation there are really only two options, estimate the missing value or make the necessary experimental measurements to calculate it. Neither of these options are particularly satisfactory, and so the ability to calculate properties for whole molecules, fragments, and atoms for just about any compound has obvious attractions. As mentioned earlier, the rapid improvements in computing coupled with the development of easily operated molecular modeling packages has contributed greatly to the increased use of calculated physicochemical properties. The starting point for most of these calculations, except topological properties as described below, is the 3D representation of chemical structures. While there will always be debate about the best way to estimate 3D structure, discussion of this is outside the scope of this review, and here it will be assumed that reasonable 3D structures can be readily computed for all compounds. There are techniques that specifically aim to produce 3D QSAR models, and these are discussed in a separate section.

**Topological Descriptors.** These are perhaps the ultimate calculated descriptors being based solely on the standard 2D representation of a chemical structure. The most well-known topological parameters are the molecular connectivity indices first described by Randic[96] and extensively investigated by Hall and Kier and co-workers.[97−99] Connectivity indices in their simplest form are computed from the hydrogen-suppressed skeleton of a compound by the assignment of a degree of connectivity, $\delta_i$, to each atom ($i$) representing the number of atoms connected to it. Figure 3 shows the degree of connectivity for each of the four heavy atoms in isobutyl alcohol. For each bond in the structure, a bond connectivity, $C_k$, can be calculated by taking the reciprocal of the square root of the product of the connectivities of the atoms at either end of the bond. For example, the bond connectivity for the first carbon−carbon bond (from the left) in the structure is

$$C_1 = \frac{1}{\sqrt{(1 \times 3)}}$$

More generally the bond connectivity of the $k$th bond is given by

$$C_k = \frac{1}{\sqrt{(\delta_i \delta_j)}}$$

where the subscripts $i$ and $j$ refer to the atoms at either end of the bond. The molecular connectivity index, $\chi$, for a molecule is found by summation of the bond connectivities over all of its $N$ bonds.

$$\chi = \sum_{k=1}^{N} C_k$$

For the butanol shown in the figure, the four bond connectivities are the reciprocal square roots of (1 × 3), (1 × 3), (2 × 3), and (2 × 1), which gives a molecular connectivity value of 2.269. This simple connectivity index is known as the first-order index because it considers only individual bonds, in other words, paths of two atoms in the structure. Higher order indices may be generated by the consideration of longer paths in a molecule, and other refinements have been considered, such as valence connectivity values, path, cluster, and chain connectivities.[100]

Connectivity indices have been shown to correlate[97] with cavity surface area (CSA), and since CSA is a fundamental property it is perhaps not surprising that they have also been shown to correlate with molecular polarizability,[97] water solubility,[101] boiling point,[101] partition coefficient,[102] and van de Waals' volume.[103] There seems little doubt that molecular connectivity descriptors contain information that relates primarily to molecular shape.[104] A principal component analysis of 108 molecular connectivity descriptors for a set of *n*-alkanes and polychlorinated biphenyls showed[105] that three principal components accounted for 98% of the variance in the data set. These three PC's were associated with

(1) degree of branching
(2) molecular size or bulk
(3) structural flexibility.

Molecular connectivity descriptors have been shown to explain a variety of different types of biological properties including a number of applications in the environmental area.[106−109] This is perhaps due to the fact that these data sets often contain quite diverse structures, and connectivity indices, unlike many other descriptors, may be readily calculated for compounds without a common parent. Caution should be used in the indiscriminant use of connectivity indices for such sets, however, as it has been pointed[110] out that there are dangers in their application to nonhomologous series.

Connectivity descriptors still continue to find application in the generation of QSAR models,[111−115] and work continues on the elaboration of other connectivity[116−121] and topological descriptors.[122−125]

Some of the reported applications of the use of connectivity descriptors have been criticized because of the use of large numbers of these parameters, with the inherent danger of chance effects. Other comments have involved the physicochemical interpretation of the meaning of these descriptors, and attempts have been made to justify their significance,[126,127] leading to a dialogue[128,129] somewhat reminiscent of the debate over the $\mathcal{T}$ and $R$ parameters of Swain and Lupton as described earlier.

**Computational Chemistry.** As discussed in the introduction to the physical organic chemistry section, theoretical quantities were involved[8,9] in QSAR models in the early 1960s, but the widespread use of computational chemistry to produce quantitative descriptions of chemical structure did not happen until the 1980s.[130,131] Early applications involved the use of just one or two descriptors that were presumably selected on mechanistic grounds. Lewis, for example,

**Table 3.** An Example of a Set of Calculated Properties (Reproduced with Permission from Ref 136)

| Calculated Property Set (81 Parameters, 79 Compounds) | |
|---|---|
| | Whole Molecule Properties |
| "bulk" descriptors | M.Wt., van der Waals' volume, dead space volume, collision diameter, approach diameter, surface area, molar refraction |
| "shape" descriptors | moment of inertia in *X*, *Y*, and *Z* axes, principal ellipsoid axes in *X*, *Y*, and *Z* directions |
| electronic and energy descriptors | dipole moment, *X*, *Y*, and *Z* components of dipole moment, energies (total, core−core repulsion, and electronic) |
| hydrophobicity descriptors | log *P* |
| | Substituent Properties |
| for 2 substituents | coordinates (*X*, *Y*, and *Z*) of the center, ellipsoid axes (*X*, *Y*, and *Z*) of the substituent |
| | Atom-centered Properties |
| electronic | atom charges, nucleophilic and electrophilic superdelocalizability, for atoms 1−14 |
| shape | interatomic distances between six pairs of heteroatoms |

reported[132] the following correlation between tumor inhibition and electron density on a nitrogen atom in the HOMO for a set of aniline mustards:

$$\log 1/E_D = -5.05(\pm 0.87)Q(1)\mathrm{HOMO} + 1.63(\pm 0.64) \quad (11)$$

$$n = 11, \quad r = 0.89, \quad s = 0.35, \quad F = 33.5$$

The description of an in-house routine for the calculation of computationally derived parameters[133] and the recognition that statistical methods could be used for their selection[134,135] has led to the use[136] of a much wider range of parameters calculated by computational chemistry programs, as shown in Table 3. The properties in regular use now include atomic charges, HOMO and LUMO energies, orbital electron densities, superdelocalizabilities, atom−atom polarizabilities, molecular polarizabilities, dipole moments, polarity indices, and energies along with quantities derived from them.[137,138] Indeed, current molecular modeling packages such as TSAR[139] and Cerius[140] allow the routine calculation of many such descriptors for modeled structures.

In addition to these properties, there are a number of other descriptors that may be computed for the construction of QSAR models. Molecular electrostatic potentials (MEP) have been shown[141] to correlate with experimentally based quantities such as $\sigma_I$, $\sigma_R$, p$K_a$, $\alpha$, and $\beta$ (solvatochromic H bond donor and acceptor parameters). Thus, it is probably not surprising that MEP values, often minima or maxima, have been shown to relate to biological properties such as inotropes,[142] the methylation of DNA bases,[143] anti-HIV activity,[144] and several other responses.[131,145] Charged partial surface areas have been computed and applied to the prediction of a variety of experimental physicochemical properties,[146−149] and although they have not been applied extensively to biological properties it has been suggested that they have much potential in this area.[150] Blaney and co-workers have reported a system for the comparison of molecular surface properties and have shown[151] the application of electrostatic potential and shape properties to the study of $\beta$-lactam antibiotics and phosphodiesterase inhibitors. Redox properties have been computed for benzoquinones in order to give some insight into the redox cycling mechanisms of the anthracycline antibiotics for which they serve as simple models.[152] A computed molecular lipophilicity potential, called a heuristic molecular lipophilicity potential (HMLP), has been shown to reproduce experimental log *P* values quite well.[153] The difference between HMLP and other attempts to produce molecular lipophilicity potential (MLP) models is that the former makes use of molecular geometries and surfaces whereas MLP models generally use an empirical parameter set. The HMLP approach may provide a useful source of data in the generation of 3D QSAR models, which are discussed in the next section.

**"Spectroscopic" Descriptors.** The use of experimental spectroscopic properties, such as NMR chemical shifts and infrared and Raman stretching frequencies, has already been briefly mentioned and the major problem with the employment of experimental quantities, their predictability, pointed out. Spectroscopic properties can be calculated, with varying degrees of reliability, from ab initio and semiempirical quantum mechanics packages, and these may be used as a replacement for experimentally determined values. A recently published[154,155] approach, EVA,[156] makes use of quantities derived from quantum chemical calculations based on 3D molecular coordinates which bear some similarity to spectroscopic data. The EVA method involves the following steps:

(1) 3D coordinates are generated for each compound in the data set.

(2) Normal coordinate frequencies are calculated using a semiempirical program such as MOPAC.

(3) The normal coordinate eigenvalues are projected onto a bounded frequency scale, typically over the range 0−4000 $cm^{-1}$ as shown in Figure 4.

(4) A Gaussian smoothing function of "width" $\sigma$ is applied to each vibration (Figure 4b).

(5) The resulting descriptor, which looks like a spectrum (Figure 4c), is sampled at fixed intervals (*L*) to provide a set of values to describe each molecule.

The procedure followed above was chosen to overcome a problem associated with the direct use of normal coordinate frequencies. The number of normal modes varies with the number of atoms in a molecule, and thus generally, unless each molecule contained the same number of atoms, there would be different numbers of descriptors for each compound. A disadvantage of this approach, however, is that there are two adjustable parameters, $\sigma$ and *L*, which potentially need to be optimized for each data set, although a recent report has shown that $\sigma$ may be used to control conformational sensitivity.[157]

The EVA descriptor has been successfully applied to the calculation of octanol/water partition coefficients[154] and to a number[158] of biological data sets that had also been modeled using the 3D QSAR method CoMFA (see later section). The EVA technique produced some comparable models to CoMFA but without the need for a specific structural

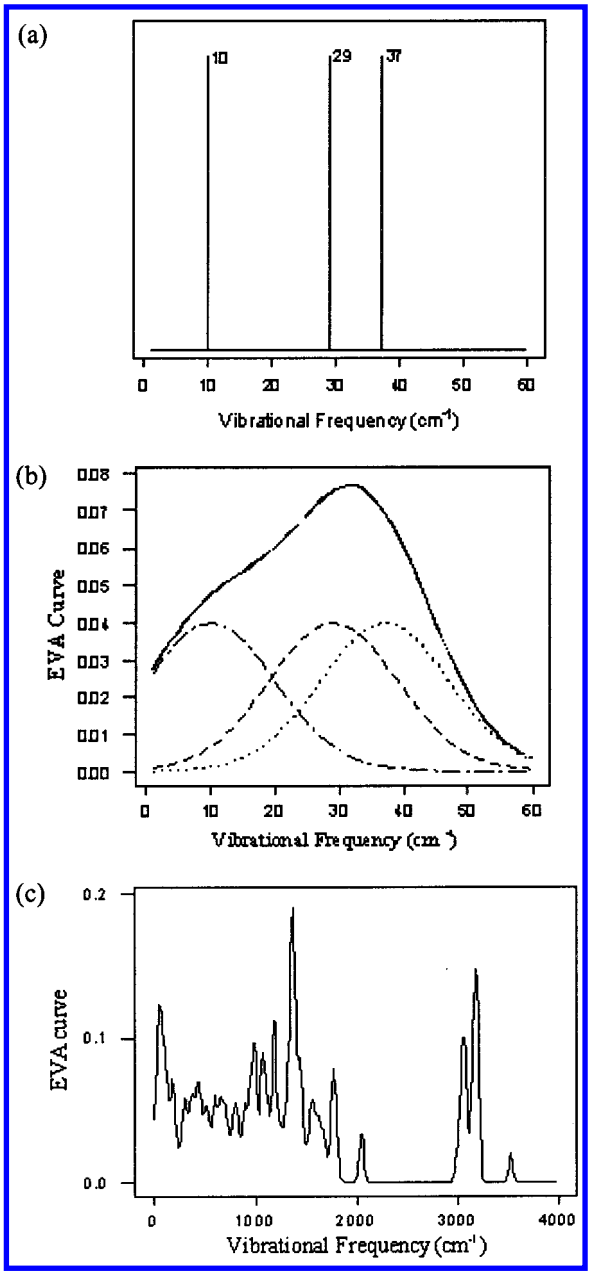


**Figure 4.** Projection of eigenvalues onto a bounded frequency scale (a) and application of a Gaussian smoothing function (b) to produce an EVA descriptor (c).

alignment that is a requirement for CoMFA and a source of potential difficulty for the technique. Benigni and co-workers have compared EVA with experimental infrared spectra, along with three other descriptor sets, and conclude that the experimental data contains extra information compared with the calculated properties.[159] Tuppurainen has described an interesting modification of the EVA approach based on electronic eigenvalues (called EEVA) and has shown that it can be used to generate quite good models for a number of biological data sets.[160] Simulated and experimental spectral data have also been employed in a technique called comparative spectra analysis as described in the CoMFA subsection of 3D QSAR methods.

**Fingerprints.** As mentioned in the Introduction, there are techniques for characterizing chemical structure based on the presence or absence of molecular fragments or functional groups, but these will not be discussed further here. There are related approaches, however, based on molecular "fingerprints", which should be considered in this review. A substructure fragment description system typically uses a fragment description bit-string where the positions in the string denote particular fragments as defined by a fragment dictionary. An alternative approach is the generation of all fragments of a particular type, e.g., all linear sequences of atoms up to a fixed number of bonds, and their presence in the molecule denoted by a bit-string generated using a superimposed coding procedure.[161] These fingerprints are then often subject to a hashing procedure.[162] There are advantages and disadvantages in the use of fingerprints,[163] but one of their principal advantages is that they are based[164] on a 2D representation of molecules and thus can be rapidly computed even for large collections of compounds such as combinatorial libraries. The fact that fingerprints ignore 3D information might appear to be a disadvantage, but a report comparing 2D and 3D descriptors concluded, perhaps surprisingly, that the 2D parameters performed better than 3D.[165]

Molecular fingerprints are used in several chemical information handling systems and are thus readily available for manipulation in order to create new descriptors. For example, Martin and co-workers used Daylight[166] fingerprints combined with log *P*, topological indices, and five "receptor recognition" descriptors to produce a means of characterizing large compound collections.[167] The Unity[168] hashed fingerprint has formed the basis of a new descriptor, termed a molecular hologram, which has been shown to yield quite reasonable PLS models for a set of benzodiazepines binding to the $GABA_A$ receptor.[169] The process of producing the molecular hologram descriptors involves the choice of fragment lengths, and this, of course, provides an adjustable parameter that may be altered as the analysis is performed. Seel and co-workers have examined the effects of changes in the type and size of fragments used in the hologram QSAR approach.[170]

## 3D QSAR

Recognition of the importance of the 3D structure of drugs to their biological properties, increasing knowledge of the 3D structure of biological macromolecules such as proteins, acceptance of the importance of stereochemistry, and dissatisfaction with the limitations of "traditional" approaches led to the evolution of many attempts to generate 3D QSAR models. There is not space here to consider all of the different approaches, and, since many of them involve novel statistical modeling methodology, not properties, they are probably not relevant to a review of properties. What will be covered in this section are the two most commonly used approaches, CoMFA and GRID, along with some other methods in which novel descriptors are involved. A 3D approach, EVA, has already been described, and the use of 3D fingerprints is presumably a 3D method. There are a number of comprehensive reviews of other approaches to 3D QSAR modeling.[171−174]

**CoMFA and GRID.** CoMFA (comparative molecular field analysis) is arguably the older of these two methods since it was first presented at an American Chemical Society meeting in 1979.[175] CoMFA, originally called DYLOMMS[176] (dynamic lattice oriented molecular modeling system), was designed at the outset to characterize sets of small molecules.

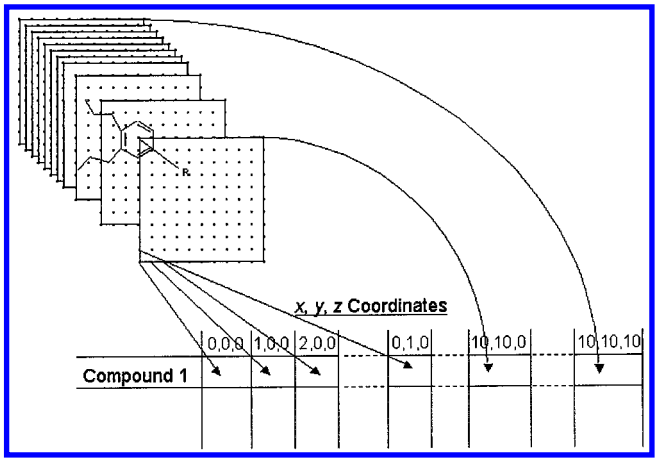


**Figure 5.** Illustration of the CoMFA and GRID procedures for a $10 \times 10 \times 10$ matrix.

GRID, on the other hand, was initially described as "a computational procedure for determining energetically favorable binding sites on biologically important macromolecules".[177]

The two techniques are, in fact, quite closely related as they both involve the calculation of interaction energies between a probe and a small molecule, or part of a macromolecule, at points in space defined by a grid placed around the target structure. Figure 5 shows a molecule surrounded by a $10 \times 10 \times 10$ grid of points, and it can be seen from this how a data matrix consisting of interaction energies can be built up. In the case of CoMFA, the steps involved in the procedure can be summarized as follows:

(1) Obtain a suitable 3D structure for each molecule in the training set.

(2) Derive partial atomic charges so that an electrostatic field can be generated.

(3) Align the molecules using some suitable alignment strategy (after conformational analysis if required).

(4) Create a cubic lattice of points around the molecules (usually larger than the largest member of the set).

(5) Compute interaction energies using a probe such as a pseudo methyl group with a unit positive charge. This generates a steric interaction energy based on a Lennard-Jones potential and an electrostatic interaction energy based on a Coulombic potential.

(6) Fit a PLS model to the biological response and the interaction energies.

(7) Make predictions for a test set, and visualize the results as contour plots on displays of the individual molecules in the set.

This is obviously a gross simplification! An enormous amount of work has been involved in many of the steps listed above, and debate, often quite vigorous, continues over a number of them. Norinder, in his review[178] of progress in CoMFA methodology, lists four main areas where advances have been made:

(1) protocols for the alignment of compounds

(2) introduction of new fields

(3) variable selection techniques

(4) statistical developments

Of these, perhaps the most problematic is the selection of a suitable alignment strategy. It might even be argued that a common alignment is one reason some CoMFA models give poor results since it has been shown that alternative binding modes exist for similar ligands.[179] Perhaps what is required is an alignment of each molecule in the set with a receptor model based on some docking[180] algorithm. Of course this requires knowledge of the receptor, but with the present progress in uncovering the human genome it is likely that any given receptor sequence will soon be available, although the construction of models from sequence is not a trivial task.[181]

A single technical advance was probably responsible for the development of the CoMFA technique into the popular method that it is today, and that was the application of the partial least-squares[182,183] (PLS) statistical modeling method to building relationships[184] between the grid data and the biological response. As Kubinyi points[185] out "the book *Quantitative Drug Design*,[186] comprising 766 pages and published in 1990, contains less than one page(!) on 3D QSAR methods related to CoMFA", and yet three years later he was the editor of an entire book devoted to the approach.[187] Progress since that book has been remarkable; there have been hundreds of reported applications of the technique, and those published between 1993 and 1997 are nicely collated in a chapter by Kim.[188]

So, what are the principal differences between CoMFA and GRID? A major difference is that GRID uses more probes for the determination of interaction energies, presumably as a result of the original design aim of GRID to investigate macromolecular binding sites. Although GRID was originally intended for the examination of such sites, and of course still is,[189] it can obviously also be used as a replacement for the interaction fields in a CoMFA type approach. In addition to probes such as $CH_3$, $NH_2$, $NH_3^+$, and $O^-$, considerable effort has been concentrated on the parametrization of hydrogen bonding in the GRID force field.[190,191] Goodford has recently reviewed[192] the GRID force field, and it would seem, from the number of reported CoMFA type applications that use GRID, that this multiprobe approach is popular. The current version of GRID offers 56 probes from the regular menu with 9 multiatom probes available from an auxiliary menu. Other types of field have been added to the standard CoMFA approach, the first being the hydropathic interaction technique (HINT) of Kellog and co-workers,[193,194] with others including HOMO[195] and LUMO[196] fields. No doubt the development of new forms of interaction fields will continue to drive CoMFA and related techniques forward.

The other principal difference between GRID and CoMFA is the use of PLS, which is an integral part of the CoMFA method. GRID, as originally devised, simply computed interaction fields that could then be visualized using a molecular graphics program. Later applications of GRID in CoMFA-type approaches used PLS to build regression models, but this is not a requirement of the GRID method. The research group of Clementi has been responsible[197−199] for an improved PLS technique called GOLPE (generalizing optimal PLS estimations), and of course other multivariate modeling techniques could be employed.

**(a) Related Techniques.** Two related techniques to CoMFA are CoMSIA[200] (comparative molecular similarity index analysis) and CoMMA[201] (comparative molecular moment analysis). The first of these makes use of similarity measures calculated between a common probe and each

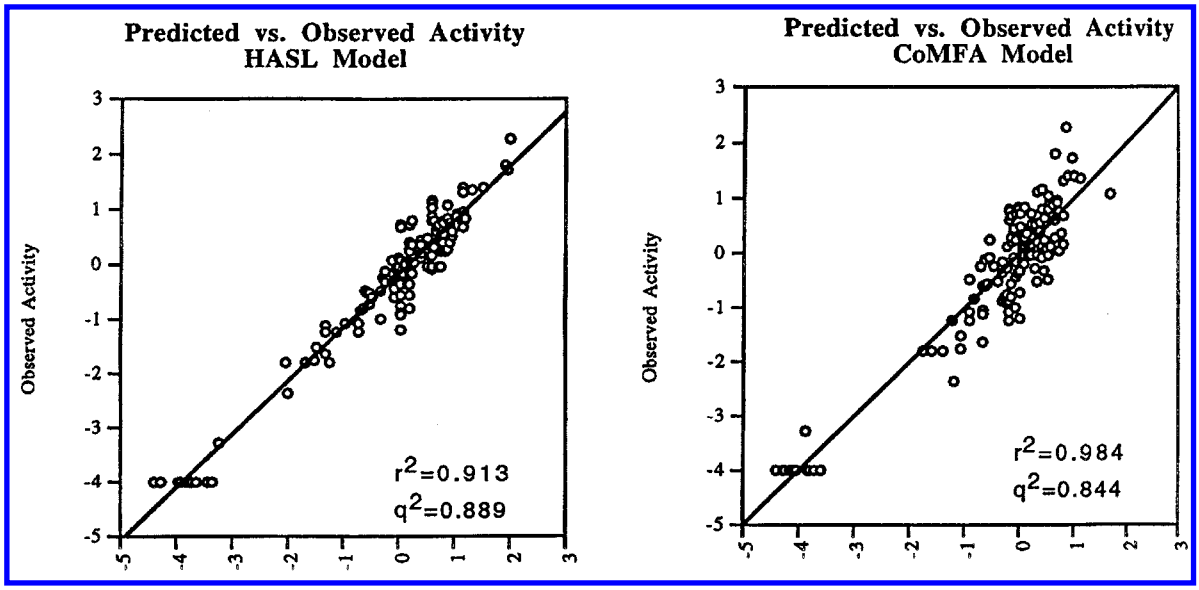


**Figure 6.** Comparison of HASL and CoMFA models for 154 artemisinin analogues (reprinted with permission from *J. Comput.-Aided. Mol. Des.* **1998**, *12*, 165−81).

molecule in the set over a regularly spaced grid. The distance dependence of the similarities are computed using a Gaussian function, which it is claimed avoids singularities near the atomic positions, unlike the Lennard-Jones and Coulomb potentials used in a standard CoMFA treatment. In an analysis of thermolysin inhibitors, using five different property fields, it was claimed that the CoMSIA diagrams were easier to interpret than standard CoMFA results.[202] The CoMMA method makes use of molecular moments such as the principal moments of inertia and properties derived from the dipole and quadrupole moments to characterize[203] compounds in a CoMFA-like study. One of the interesting claims for this technique is that it requires no superimposition step, a common problem in most 3D QSAR approaches, and a recent report examines this in some detail.[204]

In an investigation of multiconformational composite molecular potential fields, it has been shown that an overlay of the extrema of molecular electrostatic potential and surface interactions can be used to identify important conformations.[205] This technique uses a CoMFA-like grid as a starting point but then allows the interaction energy points to converge onto a small number of extrema. This has the advantages that no overlay is required, the method also considers multiple conformers, and the resulting properties may be used as molecular descriptors. Application to the docking of a set of complexes that bind through aromatic stacking and hydrogen bonding yielded results that agreed well with experimental oservations.[206] Another recently published method that does not require an overlay strategy is a technique based on measured and simulated spectral information, comparative spectra analysis (CoSA).[207] In a comparison of the performance of this technique with CoMFA on the well-known steroid set, comparable or better results were obtained using CoSA. Interestingly, a combination of spectral descriptors with CoMFA fields gave enhanced performance in most cases (see also Conclusions).

The final method that will be mentioned here is HASL[208] (hypothetical active site lattice). This technique involves the use of a grid to describe the properties of molecules in a set but, in a sense, operates in the opposite direction to CoMFA. In the construction of a CoMFA grid, the grid is generated to enclose the molecule but grid points that lie very close to, or within, the van der Waals' radii of the molecules are discarded. The HASL method also uses a grid of points, but here the "molecular lattice" is chosen to be those points that lie within the van der Waals' radii. The lattice points are assigned properties (a HASL-type value) according to the properties of the nearest atom, and these molecular lattices then form descriptors for each compound in the set. Molecular alignment in a HASL treatment can be achieved by alignment of the lattices or by alignment of the molecules just as for a regular CoMFA study. Successful applications of the HASL approach have been reported,[209,210] and a recent study comparing HASL and CoMFA for a set of antimalarial artemisinin analogues shows[211] quite comparable quality of fitting as shown in Figure 6. This report concludes that the HASL and CoMFA models represent complementary approaches.

**WHIM Descriptors.** The WHIM (weighted holistic invariant molecular) descriptors method originates from the group of Todeschini in Milan.[212−214] This intriguing approach to 3D QSAR is based on a principal components treatment of atomic coordinates that is claimed to be invariant to rotation and translation (see † below) and thus needs no superimposition strategy. The original WHIM procedure consisted of the following steps:

(1) A weighting scheme (one of six) is applied to the 3D molecular coordinates.

(2) The coordinates are centered†, and a weighted covariance matrix is calculated.

(3) The covariance matrix is used as the basis of a principal components† analysis (PCA).

(4) The scores matrix of the atoms is used to extract statistical parameters from which the WHIM descriptors are derived.

†Because the coordinates are centered the resulting descriptors are invariant to translation. The PCA step gives a unique solution that is invariant to rotation. Thus, the WHIM descriptors for a set are unaffected by the relative orientation of the compounds in the set.

The weighting schemes consist of unit weighting (unweighted), atomic mass, van der Waals' volume, Mulliken atomic electronegativity, atomic polarizabilities, and the electrotopological indices of Kier and Hall.[215] For each weighting scheme, two kinds of WHIM descriptors have been defined, "directional" and "nondirectional". These parameters are computed from various results of the PCA; the eigenvalues $\lambda_1$, $\lambda_2$, and $\lambda_3$, for example, are used as "size" descriptors, and the eigenvalue proportions are used as "shape" descriptors. A total of 66 directional WHIM descriptors may be computed (48 if only planar molecules are involved) and 37 nondirectional.[216] Initial applications of the WHIM descriptors showed that they could be used to describe a variety of physicochemical properties such as melting point, boiling point, solubility, $M_R$, polarizability, and so on.[213,214,217,218] It is thus not surprising that they have also been shown to be successful in the establishment of models of biological responses ranging from toxicity[219] to *Daphnia magna* to antagonists[217] to epinephrine.

Since the WHIM approach is based on a set of 3D coordinates, it can be applied to representations of molecules other than the coordinates of the atoms. Recent reports have shown an extension of WHIM to consider molecular surfaces[220] and a grid-based treatment of interaction fields.[221] There is little doubt that the WHIM descriptors contain useful chemical information. They have the advantage that they may be computed for any compound, given that a "reasonable" 3D structure is available, and the fact that no superimposition strategy is required is a distinct benefit compared with other 3D QSAR approaches. A disadvantage lies in their ease of interpretation, compared with other more familiar properties. Perhaps the future for these and other new descriptors lies in their combination with other sets of properties (see Conclusions).

**The Consideration of Conformation.** One important property of molecules that has been virtually neglected in all of the commonly used descriptors, measured, calculated, or theoretical, is their ability to change conformation. This has important consequences in many aspects of the interaction of compounds with a biological system; conformational changes may enhance binding at an active site, for example, or may alter the way a compound distributes within the system. Various individual studies have considered a particular aspect of the conformational freedom of the molecules involved and have shown that a certain feature, such as a bond or torsion angle, may be used to help explain activity. An example of this can be seen for a set of rigid analogues of nifedipine where their calcium channel affinities (log *KI*) were correlated[222] with the deviation from 90° of the torsion angle between two rings ($\Delta\alpha$):

$$\log KI = 0.067(\pm 0.017)\Delta\alpha + 0.19(\pm 0.34) \quad (12)$$

$$n = 7, \quad r = 0.88, \quad F = 16.5$$

Although the parameter $\Delta\alpha$ is clearly useful in this case, it is only applicable to this set and other examples of such descriptors are similarly restricted. Some more general attempts to account for conformational flexibility include molecular shape analysis,[223−225] the use of conformational energy and entropy values,[226−228] a substituent entropy constant,[229] and flexibility indices derived from molecular shape descriptors.[230] Ford and co-workers used molecular dynamics to derive "time-averaged" descriptors for a set of pyrethroid analogues.[231] In a previous study[232] of this set, it was shown that a number of computed properties, from a total set of 70, were correlated with the biological responses. A factor analysis of the descriptors gave eight "significant" factors, and these were shown[233] to correlate with two in vivo activities, knockdown (KDA) and kill (KA), and an in vitro response, neurotoxicity (NT), by equations of the following form (coefficients omitted for clarity):

$$\mathrm{KA} = -F1 - F2 + F4 - F7 \quad (13)$$

$$\mathrm{KDA} = -F1 - F2 + F4 - F5 \quad (14)$$

$$\mathrm{NT} = -F1 - F2 + F4 + F6 - F7 \quad (15)$$

It would seem that factors 1, 2, and 4 are associated with binding at the receptor (eq 15) and the whole insect knockdown and insecticidal activity. Factor analysis of the time-averaged set revealed an extra "significant" factor compared with the static set, and this was shown to relate to the lifetime of the "extended" conformation versus the "folded". Rognan has recently reviewed the use of molecular dynamics simulations in drug design.[234]

There are many other examples of studies that consider the effects of conformational changes and conformational flexibility on the construction of models of biological response, but most of these do not involve the explicit parametrization of this molecular property. The final method that will be discussed here is a technique, known as 4D-QSAR, which considers both conformational and alignment freedom in the construction of 3D QSAR models.[235] This technique sets out to address the following three problems in a 3D QSAR approach:

(1) The identification of *active* conformations/molecular shapes of flexible compounds.

(2) The choice of a *molecular alignment*.

(3) The identification of an *interaction pharmacophore*.

The active conformation/shape is defined as that which optimizes the 3D QSAR. The molecular alignment chosen by the 4D QSAR technique is similarly that alignment which maximizes the 3D QSAR, while the interaction pharmacophore is the partitioning of the different parts of each molecule in terms of their type of interaction with a receptor or other binding site. 4D QSAR is based on a reference grid placed around 3D structures of the molecules, and the descriptors involved are derived from grid cell occupancies. The reference grid and grid cell occupancies are analogous to a standard CoMFA approach, but where 4D QSAR differs is in the performance of conformational ensemble sampling by molecular dynamics, the fourth dimension, and by the selection of trial alignments. The choice of conformations is part of the modeling process that uses PLS and a genetic algorithm, whereas the choice of alignment is made by the operator although this could presumably also be made a part of the modeling process. Impressive results have been obtained for antiarrhythmics[236] and thromboxane A2 receptor antagonists[237] with the resulting models containing quite small numbers of descriptors compared with a standard CoMFA treatment.

## CONCLUSIONS

As the preceding sections have shown, there is a very large variety of different types of molecular descriptors available

**Table 4.** Summary of Discriminant Analysis Results for a Set of 90 Mutagens (Adapted with Permission from Ref 243)

| ID analysis | pooled data set[a] | no. of variables[b] | classification function (% correct) | jackknifed validation (% correct) |
|---|---|---|---|---|
| pool 1 | EVA 342 SMC + WHIM SMC | 6 | 92.2 | 90.0 |
| pool 2 | EVA 342 PDRv2.0 + WHIM SMC | 9 | 96.7 | 96.7 |
| pool 3 | EVA 288 CORCHOP + WHIM SMC | 9 | 97.8 | 95.6 |
| pool 4 | EVA 342 SMC + TSAR 58 | 9 | 92.2 | 90.0 |
| pool 5 | EVA 342 PDRv2.0 + TSAR 58 | 9 | 93.3 | 91.1 |
| pool 6 | EVA 288 CORCHOP + TSAR 58 | 5 | 90.0 | 86.7 |
| pool 7 | EVA 342 SMC + TSAR stand. | 9 | 92.2 | 90.0 |
| pool 8 | EVA 342 PDRv2.0 + TSAR stand. | 8 | 93.3 | 91.1 |
| pool 9 | EVA 288 CORCHOP + TSAR stand. | 5 | 90.0 | 86.7 |
| pool 10 | WHIM SMC + TSAR 58 | 8 | 97.8 | 94.4 |
| pool 11 | WHIM SMC + TSAR stand. | 8 | 94.4 | 92.2 |
| pool 12 | EVA 342 SMC + WHIM SMC + TSAR 58 | 9 | 97.8 | 94.4 |
| pool 13 | EVA 342 PDRv2.0 + WHIM SMC + TSAR 58 | 9 | 97.8 | 94.4 |
| pool 14 | EVA 288 CORCHOP + WHIM SMC + TSAR 58 | 9 | 96.7 | 93.3 |
| pool 15 | EVA 342 SMC + WHIM SMC + TSAR stand. | 10 | 96.7 | 92.2 |
| pool 16 | EVA 342 PDRv2.0 + WHIM SMC + TSAR stand. | 10 | 96.7 | 92.2 |
| pool 17 | EVA 288 CORCHOP + WHIM SMC + TSAR stand. | 8 | 95.6 | 92.2 |

[a] EVA, WHIM, and TSAR refer to variables computed using the EVA and WHIM methods and the structure activity program TSAR.[139] CORCHOP, SMC, and PDRv2.0 refer to variable selection methods; see refs 134 and 243. [b] The number of variables included in the discriminant function; see ref 245 for details.

and all have shown utility in the explanation of measured physicochemical properties and/or biological activities. The range of applications varies from set to set, and the ease of interpretation, in other words the chemical "meaning", of the different types of parameters ranges from the obvious to the virtually obscure. The need for interpretability depends on the application, of course, since a mathematical model relating some target property to chemical features may be all that is required. It is obviously desirable to be able to attempt some explanation of "mechanism" in chemical terms, but it is often not necessary, per se.

So how should a set of descriptors be chosen for a particular study? The answer is clearly dictated to some extent by the practical consideration of what descriptors are accessible to the investigator, but a major determinant lies in the nature of the compounds in the set. A collection of "fairly" rigid, "mostly" similar, molecules lends itself to characterization by almost all of the methods described here, whereas a very diverse or very large set precludes the use of some descriptors because there is not a "parent", or an obvious overlay strategy, or because of the computational overhead involved. These considerations aside, how else can a descriptor set be chosen? The information content of a set is clearly of importance, and Procrustes[238] analysis followed by principal coordinates analysis has been applied[239] to the comparison of different property sets used to characterize a diverse compound collection. This method was first used by Greenwood to compare the information in a set of spectral descriptors with that in a set of substituent constants.[240] Benigni and co-workers employed canonical correlation analysis and cluster analysis to compare two different descriptor sets for a highly noncongeneric series of 293 molecules.[241] This study concluded that the two descriptor sets were similar at the "global" level but showed different characteristics on a detailed scale. This has important consequences for the description of different sized sets of molecules; large collections may be characterized using almost any set of properties, whereas the descriptor set used for a congeneric series will determine the quality of the models that may be constructed. Another important criterion

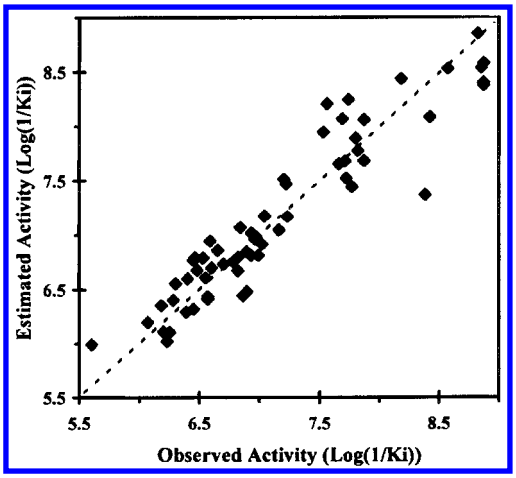


**Figure 7.** Plot of estimated vs observed activity for a set of DHFR inhibitors (reprinted with permission from *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11−20).

is that of success, in other words, how often have these descriptors been shown to be useful, but care must be taken in the assessment of success since descriptor sets may be used because of habit or fashion, rather than any particular choice for a purpose.

As noted in the Introduction, the "best" set of descriptors may in fact be achieved by careful selection of properties of several different types. This is not often done since most studies use only a single class of descriptor, but reports are beginning to appear where a subset of parameters is selected from a combination of two or more descriptor types.[242−244] A study of 90 diverse mutagens reported[245] that superior classification functions could be built using descriptors from three different sets as shown in Table 4. The classification results shown here are impressive (mutagen activity was taken as active/inactive), and it can be seen that different combinations of all three types of descriptor produced useful classification functions. Stanton has evaluated a descriptor called BCUT, which is an extension by Pearlman[246] of a set of parameters originally reported by Burden.[247] In this study a set of 105 topological, geometric, and electronic descriptors

were combined with 85 BCUT parameters.[248] An objective feature selection algorithm, reminiscent of the CORCHOP[134] procedure, was used to remove variables, and models were generated using generalized simulated annealing. Figure 7 shows a plot of calculated and measured log $1/K_i$ values for a set of 62 DHFR inhibitors using a six-term regression model. A separate examination of the original Burden descriptors comparing them with classical descriptors, molecular connectivities, 2D distances, EVA, and WHIM concluded[249] that many of the descriptor sets contain overlapping information and that the "Burden numbers" were remarkably versatile and interchangeable with other parameter sets.

The construction of appropriate subsets is not a trivial problem, and one that needs to be addressed before a general approach to the construction of QSAR models consists of the creation and combination of many different types of chemical descriptors.

## REFERENCES AND NOTES


(1) Livingstone, D. J. Structure−Property Correlations in Molecular Design. In *Structure−Property Correlations in Drug Research*; van de Waterbeemd, H., Ed.; R. G. Landes: Austin, TX, 1996; pp 81−110.
(2) Crum Brown, A., Frazer T. *Trans. R. Soc. Edinburgh* **1868−9**, *25*, 151−203.
(3) For a recent review of these descriptors, see: Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−96.
(4) Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; VCH: Weinheim, 1993.
(5) S. M., Free, Jr.; Wilson, J. W. *J. Med. Chem.* **1964**, *7*, 395−399.
(6) Bruice, T. C.; Kharasch, N.; Winzler, R. J. *Arch. Biochem. Biophys.* **1956**, *62*, 305−317.
(7) Kubinyi, H. *Quant. Struct.-Act. Relat.* **1988**, *7*, 121−33.
(8) Yoneda, F.; Nitta, Y. *Chem. Pharm. Bull. Jpn.* **1964**, *12*, 1264.
(9) Snyder, S. H.; Merril, C. R. *Proc. Nat. Acad. Sci. U.S.A.* **1965**, *54*, 258.
(10) Hammett, L. P. *Chem. Rev.* **1935**, *17*, 125.
(11) Taft, R. W. *J. Am. Chem. Soc.* **1952**, *74*, 3120.
(12) Overton, E. *Z. Phys. Chem.* **1897**, *22*, 189−209.
(13) Meyer, H. *Arch. Exp. Path. Pharm.* **1899**, *42*, 109−18.
(14) Livingstone, D. J. Quantitative Structure−Activity Relationships. In *Similarity Models in Organic Chemistry, Biochemistry and Related Fields*; Zalewski, R. I., Krygowski, T. M., Shorter, J., Eds.; Elsevier: Amsterdam, 1991; pp 557−627.
(15) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. *Nature* **1962**, *194*, 178−80.
(16) Fujita, T.; Iwasa, J.; Hansch, C. *J. Am. Chem. Soc.* **1964**, *86*, 5175−80.
(17) Nys, G. G.; Rekker, R. F. *Chim. Ther.* **1973**, *8*, 521−35.
(18) Rekker, R. F. *The hydrophobic fragmental constant*; Elsevier: Amsterdam, 1977.
(19) Rekker, R. F.; Mannhold, R. *Calculation of drug lipophilicity. The hydrophobic fragmental constant approach*; VCH: Weinheim, 1992.
(20) Leo, A.; Jow, P. Y. C.; Silipo, C.; Hansch, C. *J. Med. Chem.* **1979**, *14*, 865−8.
(21) Hansch, C.; Leo, A. *Substituent constants for correlation analysis in chemistry and biology*; Wiley: New York, 1979.
(22) Mayer, J. M.; van de Waterbeemd, H.; Testa, B. *Eur. J. Med. Chem.* **1982**, *17*, 17−25.
(23) Van de Waterbeemd, H.; Testa, B. *Adv. Drug Res.* **1987**, *16*, 85−225.
(24) Mannhold, R.; Dross, K. P.; Rekker, R. F. *Quant. Struct.-Act. Relat.* **1990**, *9*, 21−28.
(25) Rekker, R. F.; ter Laak, A. M.; Mannhold, M. *Quant. Struct.-Act. Relat.* **1993**, *12*, 152−7.
(26) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−6.
(27) Weininger, D.; Weininger, A.; Weininger, J. L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(28) Ghose, A. K.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1986**, *7*, 565.
(29) Collander, R. *Acta Chem. Scand.* **1951**, *5*, 774−80.
(30) Young, R. C.; Mitchell, R. C.; Brown, T. H.; Ganellin, C. R.; Griffiths, R.; Jones, M.; Rana, K. K.; Saunders, D.; Smith, I. R.; Sore, N. E.; Wilks, T. J. *J. Med. Chem.* **1988**, *31*, 656−71.
(31) Seiler, P. *Eur. J. Med. Chem.* **1974**, *9*, 473−9.
(32) See also the section on H bonding descriptors in the Miscellaneous section.
(33) Leahy, D. E.; Taylor, P. J.; Wait, A. R. *Quant. Struct.-Act. Relat.* **1989**, *8*, 17−31.
(34) Dunn, W. J. III, Block, J. H., Pearlman, R. S., Eds.; *Partition coefficient: Determination and estimation*; Pergamon: New York, 1986.
(35) Pliska, V., Testa, B., van de Waterbeemd, H., Eds.; *Lipophilicity in drug action and toxicology*; VCH: Weinheim, 1996.
(36) Dearden, J. C.; Bresnen, G. M. *Quant. Struct.-Act. Relat.* **1988**, *7*, 133−44.
(37) Leo, A. J. Hydrophobic parameter: Measurement and calculation. *Methods Enzymol.* **1991**, *202*, 544−91.
(38) Leo, A. J. *Chem. Rev.* **1993**, *93*, 1281−1306.
(39) Leo, A.; Hansch, C.; Elkins, D. *Chem. Rev.* **1971**, *71*, 525−616.
(40) Kubinyi, H. *Prog. Drug. Res.* **1979**, *23*, 97−198.
(41) Dearden, J. C. *Environ. Health Perspect.* **1985**, *61*, 203−228.
(42) Taylor, P. J. Hydrophobic properties of drugs. In: *Quantitative Drug Design*, Vol. 4 of *Comprehensive Medicinal Chemistry. The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds*; Hansch, C., Sammes, P. G., Taylor, J. B., Eds.; Pergamon Press: Oxford, 1991; pp 241−294.
(43) Mannhold, R.; Rekker, R. F.; Dross, K.; Bijloo, G.; de Vries, G. *Quant. Struct.-Act. Relat.* **1998**, *17*, 517−36.
(44) Rekker, R. F.; Mannhold, R.; Bijloo, G.; de Vries, G.; Dross, K. *Quant. Struct.-Act. Relat.* **1998**, *17*, 537−48.
(45) Dross, K.; Rekker, R. F.; de Vries, G.; Mannhold, R. *Quant. Struct.-Act. Relat.* **1998**, *17*, 549−57.
(46) Taft, R. W., Jr. *Steric effects in organic chemistry*; Newman, M. S., Ed.; Wiley: New York, 1956; pp 556−675.
(47) Hancock, C. K.; Meyers, E. A.; Yager, B. J. *J. Am. Chem. Soc.* **1961**, *83*, 4211−13.
(48) Fujita, T.; Iwamura, H. *Top. Curr. Chem.* **1983**, *114*, 119−57.
(49) Charton, M. *J. Am. Chem. Soc.* **1969**, *91*, 615−8.
(50) Kutter, E.; Hansch, C. *J. Med. Chem.* **1969**, *12*, 647−52.
(51) See ref 21, pp 44−47.
(52) Ahmad, P.; Fyfe, C. A.; Mellors, A. *Biochem. Pharmacol.* **1975**, *24*, 1103−9.
(53) Ghose, A. R.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21−35.
(54) Verloop, A.; Hoogenstraaten, W.; Tipker, J. *Drug Design*; Ariens, E. J., Ed.; Academic Press: New York, 1976; Vol. VII, pp 165−207.
(55) Verloop, A. *The STERIMOL approach to drug design*; Marcel Dekker: New York, 1987.
(56) Skagerberg, B.; Bonelli, D.; Clementi, S.; Cruciani, G.; Ebert, C. *Quant. Struct.-Act. Relat.* **1989**, *8*, 32−38.
(57) Verloop, A. The contribution of the STERIMOL approach to receptor mapping. In *QSAR and strategies in the design of bioactive compounds*; Seydel, J. K., Ed.; VCH: Weinheim, 1985; pp 98−104.
(58) Hammett, L. P. *J. Am. Chem. Soc.* **1937**, *59*, 96−103.
(59) A similar situation arose some 25 years later for $\pi$ scales as discussed in the section on hydrophobicity.
(60) Shorter, J. Substituent effect parameters and models applied in organic chemistry. In *Similarity Models in Organic Chemistry, Biochemistry and Related Fields*; Zalewski, R. I., Krygowski, T. M., Shorter, J., Eds.; Elsevier: Amsterdam, 1991; pp 77−147.
(61) Swain, C. G.; Lupton, E. C. *J. Am. Chem. Soc.* **1968**, *90*, 4328−37.
(62) Reynolds, W. F.; Topsom, R. D. *J. Org. Chem.* **1984**, *49*, 1989−92.
(63) Hoefnagel, A. J.; Oosterbeck, W.; Wepster, B. M. *J. Org. Chem.* **1984**, *49*, 1993−97.
(64) Charton, M. *J. Org. Chem.* **1984**, *49*, 1997−2001.
(65) Swain, C. G. *J. Org. Chem.* **1984**, *49*, 2005−10.
(66) Swain, C. G.; Unger, S. H.; Rosenquist, N. R.; Swain, M. S. *J. Am. Chem. Soc.* **1983**, *105*, 492−502.
(67) Hansch, C.; Leo, A. J.; Taft, R. W. *Chem. Rev.* **1991**, *91*, 165−95.
(68) Richet, M. C. *C. R. Soc. Biol.* (*Paris*) **1893**, *45*, 775−6.
(69) Selwood, D. L.; Livingstone, D. J.; Comley, J. C. W.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. *J. Med. Chem.* **1990**, *33*, 136−42.
(70) Seydel, J. K.; Ahrens, H.; Losert, W. *J. Med. Chem.* **1975**, *18*, 234.
(71) Rastelli, A.; De Benedetti, P. G.; Battistuzzi, G. G. *J. Med. Chem.* **1975**, *18*, 963.
(72) Turner, N. J.; Battershell, R. D. *Contrib. Boyce Thompson Inst.* **1969**, *24*, 139.
(73) Wechter, W. J.; Johnson, M. A.; Hall, C. M.; Warner, D. T.; Berger, A. E.; Wenzel, A. H.; Gish, D. T.; Neil, G. L. *J. Med. Chem.* **1975**, *18*, 339−44.
(74) Lyman, W.; Reehl, W.; Rosenblatt, D. *Handbook of Chemical Property Estimation Methods*; McGraw-Hill: New York, 1982.
(75) Higuchi, T.; Richards, J. H.; Davis, S. S.; Kamata, A.; Hou, J. P.; Nakano, M.; Nakano, N. I.; Pitman, I. H. *J. Pharm. Sci.* **1969**, *58*, 661.

(76) Clotman, D.; van Lerberghe, D.; Zeegars-Huyskens, Th. *Spectrochim. Acta, Part A* **1970**, *26*, 1621.
(77) Taft, R. W.; Gurka, D.; Joris, L.; v. R. Schleyer, P.; Rakshys, J. W. *J. Am. Chem. Soc.* **1969**, *91*, 4801.
(78) Fujita, T.; Nishioka, T.; Nakajima, M. *J. Med. Chem.* **1977**, *20*, 1971−81.
(79) Paul, H.-H.; Sapper, H.; Lohmann, W. *Biochem. Pharmacol.* **1980**, *29*, 137−40.
(80) van de Waterbeemd, H.; Kansy, M. *Chimia* **1992**, *46*, 299−303.
(81) Abraham, M. H.; Grellier, P. L.; Prior, D. V.; Taft, R. W.; Morris, J. J.; Taylor, P. J.; Lourence, C.; Berthelot, M.; Doherty, R. M.; Kamlet, M. J.; Abboud, J.-L.M.; Sraidi, K.; Guiheneuf, G. *J. Am. Chem. Soc.* **1988**, *110*, 8534.
(82) Abraham, M. H. *Chem. Soc. Rev.* **1993**, *22*, 73.
(83) Raevsky, O. A. Hydrogen bond strength estimation by means of the HYBOT program package. In *Computer-Assisted lead finding and optimization*; van de Waterbeemd, H., Testa, B., Folkers, G., Eds.; Wiley-VCH: Weinheim, 1997; pp 367−78.
(84) Kamlet, M. J.; Abboud, J.-L.M.; Abraham, M. H.; Taft, R. W. *J. Org. Chem.* **1983**, *48*, 2877−87.
(85) Taft, R. W.; Abboud, J.-L.M.; Kamlet, M. J.; Abraham, M. H. *J. Solution Chem.* **1985**, *14*, 153−86.
(86) Abraham, M. H.; Doherty, R. M.; Kamlet, M. J.; Taft, R. W. *Chem. Br.* **1986**, *22*, 551−4.
(87) Kamlet, M. J.; Taft, R. W.; Famini, G. R.; Doherty, R. M. *Acta Chem. Scand.* **1987**, *41*, 589.
(88) Abraham, M. H.; Whiting, G. S.; Fuchs, R.; Chambers, E. J. *J. Chem. Soc., Perkin Trans. 2* **1990**, 291.
(89) Kamlet, M. J.; Doherty, R. M.; Abraham, M. H.; Taft, R. W. *Quant. Struct.-Act. Relat.* **1988**, *7*, 71−78.
(90) Wilson, L. Y.; Famini, G. R. *J. Med. Chem.* **1991**, *34*, 1668.
(91) Famini, G. R.; Kassel, R. J.; King, J. W.; Wilson, L. Y. *Quant. Struct.-Act. Relat.* **1991**, *10*, 344.
(92) Famini, G. R.; Loumbev, V. P.; Frykman, E. K.; Wilson, L. Y. *Quant. Struct.-Act. Relat.* **1998**, *17*, 558−64.
(93) Famini, G. R.; Penski, C. A.; Wilson, L. Y. *Phys. Org. Chem.* **1992**, *5*, 395−408.
(94) It is tempting to draw general conclusions from this, but it should be borne in mind that these relationships are specific for the compound sets involved.
(95) Famini, G. R.; Ashman, W. P.; Mickiewicz, A. P.; Wilson, L. Y. *Quant. Struct.-Act.* **1992**, *11*, 162−70.
(96) Randic, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609−15.
(97) Kier, L. B.; Hall, L. H.; Murray, W. J.; Randic, M. *J. Pharm. Sci.* **1975**, *64*, 1971−74.
(98) Kier, L. B.; Murray, W. J.; Hall, L. H. *J. Med. Chem.* **1975**, *18*, 1272−1274.
(99) Kier, L. B.; Hall, L. H. *Molecular connectivity in chemistry and drug research*; Academic Press: New York, 1976.
(100) Kier, L. B.; Hall, L. H. *Molecular connectivity in structure−activity analysis*; Wiley: New York, 1986.
(101) Hall, L. H.; Kier, L. B.; Murray, W. J. *J. Pharm. Sci.* **1975**, *64*, 1974−7.
(102) Murray, W. J.; Hall, L. H.; Kier, L. B. *J. Pharm. Sci.* **1975**, *64*, 1978−81.
(103) Hall, L. H.; Kier, L. B. *Eur. J. Med. Chem.* **1981**, *16*, 399−407.
(104) Dearden, J. C.; Mays, P. K. *J. Pharm. Pharmacol.* **1985**, *37*, 70P.
(105) Burkhard, L. P.; Andrew, A. W.; Armstrong, D. E. *Chemosphere* **1983**, *12*, 935−43.
(106) Basak, S. C.; Magnuson, V. R. *Arzneim.-Forsch.* **1983**, *33*, 501−3.
(107) Niemi, G. J.; Regal, R. R.; Veith, G. D. *ACS Symp. Ser.* **1985**, *292*, 148−59.
(108) Nirmalakhandan, N.; Speece, R. E. *Environ. Sci. Technol.* **1988**, *22*, 606−15.
(109) Okey, R. W.; Stensel, H. D. *Water Environ. Res.* **1993**, *65*, 772−80.
(110) Lopez de Compadre, R. L.; Compadre, C. M.; Castillo, R.; Dunn, W. J., III, *Eur. J. Med. Chem.* **1983**, *18*, 569−71.
(111) Osmialowski, K.; Kaliszan, R. *Quant. Struct.-Act. Relat.* **1991**, *10*, 125−34.
(112) Cash, G. G.; Breen, J. J. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 275−9.
(113) Llorente, B.; Rivero, N.; Carrasco, R.; Martinez, R. S.; Estrada, E. *Quant. Struct.-Act. Relat.* **1994**, *13*, 419−25.
(114) Shapiro, S.; Guggenheim, B. *Quant. Struct.-Act. Relat.* **1998**, *17*, 327−37.
(115) Gozalbes, R.; Galvez, J.; Garcia-Domenech, R.; Derouin, F. *SAR QSAR Environ. Res.* **1999**, *10*, 47−60.
(116) Kier, L. B. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109−116.
(117) Kier, L. B. *Quant. Struct.-Act. Relat.* **1986**, *5*, 7−12.
(118) Kier, L. B. *Med. Res. Rev.* **1987**, *7*, 417−40.
(119) Kier, L. B.; Hall, L. H. *Quant. Struct.-Act. Relat.* **1991**, *10*, 134−40.
(120) Bonchev, D. *SAR QSAR Environ. Res.* **1997**, *7*, 23−43.
(121) Estrada, E.; Guevara, N.; Gutman, I.; Rodriguez, L. *SAR QSAR Environ. Res.* **1998**, *9*, 229−40.
(122) Hall, L. H.; Mohney, B.; Kier, L. B. *Quant. Struct.-Act. Relat.* **1991**, *10*, 43−51.
(123) Burden, R. *Quant. Struct.-Act. Relat.* **1997**, *16*, 309−14.
(124) Lobato, M.; Amat, L.; Besalu, E.; Carbo-Dorca, R. *Quant. Struct.-Act. Relat.* **1997**, *16*, 465−72.
(125) Winkler, D. A.; Burden, F. R.; Watkins, A. J. R. *Quant. Struct.-Act. Relat.* **1998**, *17*, 14−9.
(126) Saxena, A. K. *Quant. Struct.-Act. Relat.* **1995**, *14*, 31−8.
(127) Saxena, A. K. *Quant. Struct.-Act. Relat.* **1995**, *14*, 142−8.
(128) Kubinyi, H. *Quant. Struct.-Act. Relat.* **1995**, *14*, 149−50.
(129) Saxena, A. K. *Quant. Struct.-Act. Relat.* **1995**, *14*, 150.
(130) Kikuchi, O. *Quant. Struct.-Act. Relat.* **1987**, *6*, 179−84.
(131) Saunders, M. R.; Livingstone, D. J. Electronic structure calculations in quantitative structure−property relationships. In *Advances in quantitative structure−property relationships*; Charton, M., Ed.; JAI Press: Greenwich, CT, 1996; pp 53−79.
(132) Lewis, D. F. V. *Xenobiotica* **1989**, *19*, 243−51.
(133) Glen, R. C.; Rose, V. S. *J. Mol. Graphics* **1987**, *5*, 79−86.
(134) Livingstone, D. J.; Rahr, E. *Quant. Struct.-Act. Relat.* **1989**, *8*, 103−8.
(135) Ford, M. G.; Livingstone, D. J. *Quant. Struct.-Act. Relat.* **1990**, *9*, 107−114.
(136) Hyde, R. M.; Livingstone, D. J. *J. Comput.-Aided Mol. Des.* **1988**, *2*, 145−55.
(137) Cartier, A.; Rivail, J.-L. *Chemom. Intell. Lab. Syst.* **1987**, *1*, 335−47.
(138) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. *Chem. Rev.* **1996**, *96*, 1027−43.
(139) Oxford Molecular, The Medawar centre, Oxford Science park, Sandford-on-Thames, Oxford OX4 4GA, U.K.
(140) Molecular Simulations, Inc., 9685 Scranton Rd., San Diego, CA 92121-3752.
(141) Murray, J. S.; Brinck, T.; Grice, M. E.; Politzer, P. *J. Mol. Struct.* **1992**, *256*, 29−45.
(142) Bohl, M.; Ponsold, K.; Reck, G. *J. Steroid Biochem.* **1984**, *21*, 373−9.
(143) Lewis, D. F. V.; Griffiths, V. S. *Xenobiotica* **1987**, *17*, 769−76.
(144) Kumar, A.; Mishra, P. C. *J. Mol. Struct.* **1992**, *277*, 299−312.
(145) Pepe, G.; Siri, D.; Reboul, J.-P. *J. Mol. Struct.* **1992**, *256*, 175−85.
(146) Stanton, D. T.; Jurs, P. C. *Anal. Chem.* **1990**, *62*, 2323−29.
(147) Stanton, D. T.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 109−15.
(148) Egolf, L. M.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 616−25.
(149) Egolf, L. M.; Jurs, P. C. *Ind. Eng. Chem. Res.* **1992**, *31*, 1798−1807.
(150) Jurs, P. C.; Dixon, S. L.; Egolf, L. M. Representations of molecules. In *Chemometric methods in molecular design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; pp 15−38.
(151) Blaney, F.; Finn, P.; Phippen, R.; Wyatt, M. *J. Mol. Graphics* **1993**, *11*, 98−105.
(152) Mariam, Y. H.; Sawyer, A. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 441−60.
(153) Du, Q.; Arteca, G. A.; Mezey, P. G. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 503−15.
(154) Ferguson, A. M.; Heritage, T.; Jonathon, P.; Pack, S. E.; Phillips, L.; Rogan, J.; Snaith, P. J. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 143−7.
(155) Ginn, C. M. R.; Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 23−37.
(156) The name EVA comes from normal coordinate *E*igen*V*Alues.
(157) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 271−96.
(158) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 409−22.
(159) Benigni, R.; Passerini, L.; Livingstone, D. J.; Johnson, M. A.; Giuliani, A. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 558−62.
(160) Tuppurainen, K. *SAR QSAR Environ. Res.* **1999**, *10*, 39−46.
(161) Willett, P. *Perspect. Drug Discov. Des.* **1997**, *7/8*, 1−11.
(162) Ihlenfeldt, W. D.; Gasteiger, J. *J. Comput. Chem.* **1994**, *15*, 793−813.
(163) Brown, R. D. *Perspect. Drug Discov. Des.* **1997**, *7/8*, 31−49.
(164) 3D fingerprint descriptors have also now been reported.
(165) Brown, R. D.; Martin, Y. C. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(166) Daylight Chemical Information Systems, INC., 2740 Los altos, Suite #360, Mission Viejo, CA 92691.
(167) Martin E. J.; Blaney, J. M.; Siani, M. S.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. *J. Med. Chem.* **1995**, *38*, 1431−6.
(168) Tripos Associates Inc., 1699 South Hanley Rd., St Louis, MO 63144.

(169) Winkler, D. A.; Burden, F. R. *Quant. Struct.-Act. Relat.* **1998**, *17*, 224−31.
(170) Seel, M.; Turner, D. B.; Willett, P. *Quant. Struct.-Act. Relat.* **1999**, *18*, 245−52.
(171) See ref 4, pp 149−159.
(172) Blankley, C. J. Recent Developments in 3D QSAR. In *Structure−Property Correlations in Drug Research*; van de Waterbeemd, H., Ed.; R. G. Landes: Austin, TX, 1996; pp 111−258.
(173) Martin, Y. C. *Perspect. Drug Discov. Des.* **1998**, *12/13/14*, 3−23.
(174) Hahn, M.; Rogers, D. *Perspect. Drug Discov. Des.* **1998**, *12/13/14*, 117−33.
(175) Cramer, R. D., III; Milne, M. *Abstracts of Papers*; Meeting of the American Chemical Society April 1979; American Chemical Society: Washington, D.C., 1979; Computer Chemistry Section, No. 44.
(176) Cramer, R. D., III; Bunce, J. D. The DYLOMS method: Initial results from a comparative study of approaches to 3D QSAR. In *QSAR in Drug Design and Toxicology*; Hadzi, D., Jerman-Blazic, B., Eds.; Elsevier: Amsterdam, 1987; pp 3−12.
(177) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849−57.
(178) Norinder, U. *Perspect. Drug Discov. Des.* **1998**, *12/13/14*, 25−39.
(179) Mattos, C.; Ringe, D. Multiple binding modes. In *3D QSAR in Drug Design*: *Theory Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 226−54.
(180) Jones, G.; Willett, P. *Curr. Opin. Biotechnol.* **1995**, *6*, 652−6.
(181) Blaney, F. From sequence to structure. In *GENOMICS*: *Commercial opportunities from a scientific revolution*; Dixon, G. K., Copping, L. G., Livingstone, D., Eds.; Bios: Oxford, 1998; pp 77−104.
(182) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III *SIAM J. Sci. Stat. Comput.* **1984**, 5, 735−43.
(183) Dunn, W. J., III; Wold, S.; Edlund, U.; Hellberg, S.; Gasteiger, J. *Quant. Struct.-Act. Relat.* **1984**, *3*, 131−7.
(184) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959−67.
(185) See ref 4, p 159.
(186) Ramsden, C. A., Ed. *Quantitative Drug Design*, Vol. 4 of *Comprehensive Medicinal Chemistry. The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds*; Hansch, C., Sammes, P. G., Taylor, J. B., Eds.; Pergamon Press: Oxford, 1990.
(187) Kubinyi, H., Ed. *3D QSAR in Drug Design*: *Theory Methods and Applications*; ESCOM: Leiden, 1993
(188) Kim, K. H. *Perspect. Drug Discov. Des.* **1998**, *12/13/14*, 317−33.
(189) Wade, R. C. Molecular Interaction Fields. In *3D QSAR in Drug Design*: *Theory Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 486-505.
(190) Wade, R. C.; Clark, K.; Goodford, P. J. *J. Med. Chem.* **1993**, *36*, 140−7.
(191) Wade, R. C.; Clark, K.; Goodford, P. J. *J. Med. Chem.* **1993**, *36*, 148−56.
(192) Goodford, P. J. *J. Chemom.* **1996**, *10*, 107−11.
(193) Kellog, G. E.; Semus, S. F.; Abraham, D. J. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 545−52.
(194) Kellog, G. E.; Joshi, G. S.; Abraham, D. J. *Med. Chem. Res.* **1992**, *1*, 444−53.
(195) Waller, C. L.; Marshall, G. R. *J. Med. Chem.* **1993**, *36*, 2390−2403.
(196) Poso, A.; Tuppurainen, K.; Gynther, J. *J. Mol. Struct.* **1994**, *304*, 255−60.
(197) Baroni, M.; Clementi, S.; Cruciani, G.; Costantino, G.; Riganelli, D.; Oberrauch, E. *J. Chemom.* **1992**, *6*, 347−56.
(198) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9−20.
(199) Cruciani, G.; Clementi, S. GOLPE: philosophy and applications in 3D QSAR. In *Advanced Computer-Assisted Techniques in Drug Discovery*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; pp 61−88.
(200) Klebe, G.; Abraham, U.; Metzner, T. *J. Med. Chem.* **1994**, *37*, 4130−46.
(201) Silverman, B. D.; Platt, D. E. *J. Med. Chem.* **1996**, *39*, 2129−40.
(202) Klebe, G.; Abraham, U. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 1−10.
(203) Silverman, B.; Platt, D. E.; Pitman, M.; Rigoutsos, I. *Perspect. Drug Discov. Des.* **1998**, *12/13/14*, 183−96.
(204) Siverman, B. D.; Pitman, M. C.; Platt, D. E.; Rigoutsos, I. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 525−32.
(205) Vinter, J. G.; Trollope, K. I. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 297−307.
(206) Vinter, J. G. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 417−26.
(207) Bursi, R.; Dao, T.; van Wijk, T.; de Gooyer, M.; Kellenbach, E.; Verwer, P. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 861−7.
(208) Doweyko, A. M. *J. Med. Chem.* **1988**, *31*, 1396−406.
(209) Wiese, M. The hypothetical active-site lattice. In *3D QSAR in Drug Design*: *Theory Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 431−42.
(210) Doweyko, A. M. *J. Med. Chem.* **1994**, *37*, 1769−78.
(211) Woolfrey, J. R.; Avery, M. A.; Doweyko, A. M. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 165−81.
(212) Todeschini, R.; Lasagni, M.; Marengo, E. *J. Chemom.* **1994**, *8*, 263−72.
(213) Todeschini, R.; Gramatica, P.; Provenzani, R. *Chemom. Intell. Lab. Syst.* **1995**, *27*, 221−9.
(214) Todeschini, R.; Gramatica, P. *Quant. Struct.-Act. Relat.* **1997**, *16*, 113−9.
(215) Kier, L. B.; Hall, L. H.; Frazer, J. W. *J. Math. Chem.* **1991**, *7*, 229−41.
(216) Todeschini, R.; Gramatica, P. *SAR QSAR Environ. Res.* **1997**, *7*, 89−115.
(217) Todeschini, R.; Gramatica, P. *Quant. Struct.-Act. Relat.* **1997**, *16*, 120−5.
(218) Chiorboli, C.; Gramatica, P.; Piazza, R.; Pino, A.; Todeschini, R. *SAR QSAR Environ. Res.* **1997**, *7*, 133−50.
(219) Todeschini, R.; Vighi, M.; Provenzani, R.; Finizio, A.; Gramatica, P. *Chemosphere* **1996**, *32*, 1527−45.
(220) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 79−92.
(221) Todeschini, R.; Moro, G.; Boggia, R.; Bonati, L.; Cosantino, U.; Lasagni, M.; Pitea, D. *Chemom. Intell. Lab. Syst.* **1997**, *36*, 65−73.
(222) Seidel, W.; Meyer, H.; Born, L.; Kazda, K.; Dompert, W. Rigid calcium antagonists of the nifedipine type: Geometrical requirements for the dihydropyridine receptor. In *QSAR and strategies in the design of bioactive compounds*; Seydel, J. K., Ed.; VCH: Weinheim, 1985; pp 366−9.
(223) Hopfinger, A. J. *J. Am. Chem. Soc.* **1980**, *102*, 7196−206.
(224) Hopfinger, A. J.; Potenzone, R., Jr. *Mol. Pharmacol.* **1982**, *21*, 187−95.
(225) Mabilia, M.; Pearlstein, R. A.; Hopfinger, A. J. *Eur. J. Med. Chem.* **1985**, *20*, 163−74.
(226) Avbelj, F.; Hadzi, D. *Mol. Pharmacol.* **1985**, *27*, 466−70.
(227) Lopez de Compadre, R. L.; Pearlstein, R. A.; Hopfinger, A. J.; Seydel, J. K. *J. Med. Chem.* **1987**, *30*, 900−6.
(228) Hopfinger, A. J.; Lopez de Compadre, R. L.; Koehler, M. G.; Emery, S.; Seydel, J. K. *Quant. Struct.-Act. Relat.* **1987**, *6*, 111−7.
(229) Sasaki, Y.; Takagi, T.; Yamazato, Y.; Iwata, A.; Kawaki, H. *Chem. Pharm. Bull.* **1981**, *29*, 3073−5.
(230) Kier, L. B. *Quant. Struct.-Act. Relat.* **1989**, *8*, 221−5.
(231) Hudson B. D.; George A.; Ford M. G.; Livingstone D. J. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 191−201.
(232) Livingstone D. J.; Ford M. G.; Buckley D. S. A multivariate QSAR study of pyrethroid neurotoxicity based upon molecular parameters derived by computer chemistry. In *Neurotox '88*: *Molecular basis of drug and pesticide action*; Lunt, G. G., Ed.; Elsevier: Amsterdam, 1988; pp 483−495.
(233) Ford M. G.; Greenwood R.; Turner C. H.; Hudson B.; Livingstone D. J. *Pestic. Sci.* **1989**, *27*, 305−326.
(234) Rognan, D. *Perspect. Drug Discov. Des.* **1998**, *9/10/11*, 181−209.
(235) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, J.; Duraiswan, C. *J. Am. Chem. Soc.* **1997**, *119*, 10509−24.
(236) Klein, C. D. P.; Hopfinger, A. J. *Pharm. Res.* **1998**, *15*, 303−11.
(237) Albuquerque, M. G.; Hopfinger, A. J.; Barreiro, E. J.; Alencastro, R. B. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 925−38.
(238) Sibson, R. J. R. *Statistic. Soc. B* **1978**, *40*, 234−8.
(239) Rose, V. S.; Rahr, E.; Hudson, B. D. *Quant. Struct.-Act. Relat.* **1994**, *13*, 152−8.
(240) Greenwood, R.; Ford, M. G.; Higgins, B. R. The comparison of multivariate ordinations in QSAR studies. In *QSAR*: *Rational approaches to the design of bioactive compounds*; Silipo, C., Vittoria, A., Eds.; Elsevier: Amsterdam, 1991; pp 227−230.
(241) Benigni, R.; Gallo, G.; Giorgi, F.; Giuliani, A. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 575−8.
(242) Todeschini, R.; Vighi, M.; Finizio, A.; Gramatica, P. *SAR QSAR Environ. Res.* **1997**, *7*, 173−93.
(243) Todeschini, R.; Gramatica, P. *Perspect. Drug Discov. Des.* **1998**, *9/10/11*, 355−80.
(244) Gute, B. D.; Grunwald, G. D.; Basak, S. C. *SAR QSAR Environ. Res.* **1999**, *10*, 1−15.
(245) Smith, M. A Quantitative Structure Activity Relationship Study of the Ames Mutagenicity Assay. Ph.D. Thesis, University of Portsmouth, 1999.
(246) Pearlman, R. S.; Smith, K. M. *Perspect. Drug Discov. Des.* **1998**, *9/10/11*, 339−53.
(247) Burden, F. R. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−7.
(248) Stanton, D. T. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11−20.
(249) Benigni, R.; Passerini, L.; Pino, A.; Giuliani, A. *Quant. Struct.-Act. Relat.* **1999**, *18*, 449−55.

CI990162I