# On Computation of Optimal Parameters for Multivariate Analysis of Structure-Property Relationship

## Milan Randić

*Department of Mathematics and Computer Science Drake University, Des Moines, Iowa 50311*

We outline a search for optimal parameters involving heteroatoms for use in multivariate regression analysis in structure-property and structure-activity studies. The problem consists of determining optimal numerical values for the diagonal elements of the adjacency matrix in graphs with atoms of different kind. In particular we consider weighted paths as the basic molecular descriptors and search for optimal parameters for carbon atom and oxygen atom in a correlation of molecular structure with isomeric variations in the boiling points of hexanols. Standard error is taken as the criterion for the selection of the optimal parameters. The weighting algorithm restricts the diagonal entries to values greater than $-1$. The selection of positive diagonal values leads to reducing the path numbers and the negative values lead to enlarging the role of path numbers relative to the zero diagonal values implied by simple graphs in which heteroatoms are not discriminated. A systematic search for optimal parameters for alcohols gave for carbon atom and oxygen atom diagonal entries: $x = 1.50$ and $y = -0.85$ respectively when a single path number is used as a descriptor and $x = -0.15$ and $y = -0.94$ when two path numbers are used. The parameters derived for 17 hexanols have been successively applied to 37 heptanols demonstrating thus transferability of the parameters.

*Modeling is a ubiquitous and often misunderstood enterprise.* M.C. Kohn

## INTRODUCTION

A model, whether crude (qualitative) or computational (quantitative), is a set of hypotheses about the properties of the actual system being studied.[1] The test consists in comparison of the behavior of the model with observations of the properties of the actual system. From this point of view we can in chemical applications conveniently differentiate the following:

1. Property-property modeling
2. Structure-property modeling
3. Structure-structure modeling.

In each case a different hypothesis is involved. In the first case one seeks relationship between apparently unrelated properties, hoping to describe more complex behavior in terms of simpler (and presumably already better understood) property. An illustration of property-property relationship is that of a correlation between the chromatographic retention volumes and the molecular boiling points, observed first by Kovats.[2] Property-property modeling is often the major ingredient in the traditional QSAR (quantitative structure-activity relationship) studies as developed by Hansch and Fujita,[3] a work

that followed the pioneering studies on free energy relationship of Hammett, Bronstedt, and a whole school of physical organic chemistry.[4] There, as "simpler" properties, one uses as descriptors among others, log $P$ (octanol/water partition coefficient), hydrophobic constants, and molecular refractions.

Structure-property model involves a description of a molecule by "simple" structural invariants, i.e., invariants that have a direct structural interpretation and thus offer some insights into the factors that critically govern the property of interest. Simple structural descriptors include the number of atoms, the number of bonds, the number of paths of different length, the number of self-returning walks. The meaning of some structural invariants is less apparent, as already illustrated by the Wiener number $W$, the first nontrivial graph theoretical descriptor proposed already in late 1940s.[5] The $W$ represents the sum of all the distances between the vertices (measured in terms of the number of intervening edges[6]) within a graph. It is not immediately clear why such an apparently ad hoc descriptor is so successful in correlating various physicochemical properties of compounds like, alkanes, alcohols, amines, etc. Platt made an effort to interpret the apparently unfamiliar descriptor and found that it qualitatively parallels molecular volumes.[7] The list of viable graph theoretical invariants, referred to as topological indices, is lengthy (Table I).

Structure-structure relationship is more common than perhaps anticipated, even though frequently it

**Table I.** A selection of graph theoretical invariants (topological indices) used in structure-property and structure-activity studies

| Descriptor | Invariant | Ref. |
|---|---|---|
| $^1\chi$ | Connectivity index | 14 |
| $Z$ | Hosoya's topological index | 29 |
| $W$ | Wiener number | 5 |
| $^1\chi^v$ | Valence connectivity | 13 |
| $^i\chi$ | Higher connectivity indices | 14 |
| $^i\pi$ | Weighted path numbers | 23 |
| ID | Identification numbers | 25 |
| $J$ | Balaban's $J$ index | a |
| $\kappa$ | Shape index | b |
| TT | Topological torsion | c |
| $C$ | Centric index | d |
| $\varepsilon$ | Electropy index | e |
| $P_D^w$ | Wiener Information index | f |

[a] A.T. Balaban, *Chem. Phys. Lett.*, **89**, 399 (1982).

[b] L.B. Kier, *Quant. Struct.-Act. Relat.*, **4**, 109 (1985).

[c] R. Nilakantan, N. Bauman, J.S. Dixon, and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.*, **27**, 82 (1987).

[d] A.T. Balaban, *Theor. Chim. Acta.*, **53**, 355 (1979).

[e] W.T. Yee, K. Sakamoto and Y.J. I'Haya, *Rept. Univ. Electro-Comm.*, **27**, 53 (1976).

[f] D. Bonchev and N. Trinajstic, *J. Chem. Phys.*, **67**, 4517 (1977).

is not recognized as such. The idea here is to express complicated, often cumbersome, theoretical quantities in terms of less convoluted structural concepts. Illustrations are Pauling[8] and Coulson's bond orders,[9] quantities derived from Kekule valence structures and from the coefficients of occupied molecular orbitals, respectively.

Structure-structure relationships may illustrate subtle aspects of distinct theoretical molecular models and because they are explicitly confined to comparisons of alternative theoretical constructions they are likely to offer structural insights and even lead to novel conceptual advances.

## CHALLENGES TO MODEL DESIGNERS

Designing models often suffers from employment of mathematical representations of the system that are more detailed than necessary to achieve the goals of the enterprise.[1] Yates[10] included among the goals of modeling: The identification of the underlying structure of the system; the recognition of interactions among components; and the prediction of unobserved behavior. Attempts to accomplish some of the goals lead to a hierarchy of models:

1. Minimal models (ad hoc approach that mimics observations)
2. Empirical models (many parametric representations of a system)
3. Analytical Models (equations of wide applicability)

4. Mechanistic models (derived from first principles)

Clearly multivariate regression analysis is at the lower end of the above hierarchy of models and as such does not contribute to the cause-effect relationship. However it has capability to clarify a number of important goals in modeling, particularly when descriptors selected to characterize a structure have a direct structural interpretation. In applications most often the emphasis is placed on "prediction of unobserved behavior," which clearly dominates many pragmatic interests, particularly those of medicinal chemistry.

Important aspects of modeling, however, include considerations such as underlying complexity of a model, numerical inconsistencies, recognition of constraints, error propagation:

> It is unusual for only one model to be compatible with experimental observations. Often the data are not sufficiently extensive to discriminate among rival models and new experiments must be designed to answer outstanding questions. The statistical, graph theoretical, and sensitivity analysis methods (outlined above) can identify the areas for further investigation that are likely to produce significant new results.[1]

Hence, in view of the above, to be able to discriminate among alternative models remains of considerable interest in studies of structure-property relationship.

Modeling leads to many challenges. It is not only important to have a capability to derive sufficiently accurate predictions, to resolve ambiguities of the model (such as choice of alternative parameters) but also be able to identify dominant components and interpret them in terms of model components and in this way illuminate their role in a particular study. In practice this often translates into efforts on construction of models with as few as possible empirical parameters.

## MULTIVARIATE REGRESSION ANALYSIS

Multivariate regression analysis is an important tool in data reduction. In applications of multivariate analysis to structure-property relationship, the problem consists in finding a suitable expression of the form

$$P = P(x_1, x_2, x_3, \ldots x_k)$$

where $P$ stands for a property and $x_i$ are various molecular descriptors. The challenge here is that of arriving at an acceptable regression equation using as few as possible descriptors (variables). If such descriptors have a direct structural interpretation even better. The pool of descriptors is generally un-

limited. The descriptors may be selected from simpler molecular properties, from various graph theoretical indices and other structural descriptors, and/or from quantum chemically computed quantities. The corresponding SAR and QSAR analyses have been classified accordingly as structure-cryptic, structure-explicit, and structure-implicit, respectively.[11]

The nature of the descriptors used in multivariate analysis of structure-property relationship is either empirical or nonempirical, depending on whether the parametrization of the descriptors is based on observed quantities or whether they are based on assumed (idealized) theoretical concepts and computed without use of outside experimental data on the molecules considered. Such nonempirical approaches do not imply mechanistic considerations which, for example, typify *ad initio* quantum chemical calculations. The attribute merely indicates a lack of adjustable parameters. In QSAR, for instance, the scheme of Hansch[12] illustrates a highly empirical approach to the structure-activity, while the approach of Wiener[5] represents the other extreme, use of nonempirical descriptors. The number of parameters used in a typical Hansch-type approach is generally difficult to assess because the constants used to describe various atomic groups have often been predetermined. That is, the parameters have been selected for a general use, not for the particular single study. On the other hand nonempirical schemes derive their descriptors solely from the information on the molecular structure. In Table I we listed a selection of nonempirical molecular descriptors of interest in SAR and QSAR. It is easy to verify that only information that can be deduced from a structure suffices to evaluate the derived descriptors, and that no experimental data is involved in their construction.

## COLORING OF GRAPHS—
## CHARACTERIZATION OF HETEROATOMS

Nonempirical models of SAR are often restricted to hydrocarbon compounds and to compounds in which heteroatoms occur generally within a same local environment. This limitation is similar to the restriction of simple (topological) HMO method which, when one wishes to consider heterocyclic conjugated compounds, requires modifications. We will consider here the problem of "selecting" colors, i.e., selecting parameters for the characterization of atoms of different kind in a molecule. One can approach the problem of determination of the parameters for heteroatoms either empirically or using structural data only. For example, Kier and Hall[13] generalized the connectivity index[14] and higher connectivity indices[15] to molecules with heteroatoms in a nonempirical

fashion. They assigned to each atom its "valence-delta" parameter, which is not based on empirical data but derived directly from the information on the kind and the number of valence electrons of the atom considered. The advantage of nonempirical parameters is that they are generally available, however they are not likely to lead to the optimal description of the problem.

Here we will examine an alternative route to characterization of heteroatoms that leads to empirical molecular descriptors that are nevertheless closely related to molecular connectivity indices. Rather then assigning to a heteroatom a weighting factor ("valence delta") to be combined with the weighting algorithm $1/\sqrt{mn}$, for bonds of type $(m, n)$, one considers the diagonal entries of the adjacency matrix and attributes to those distinct values for different kinds of atoms. Such an approach is analogous to the well-known generalization of HMO to heteroconjugated systems. The preliminary results using so modified adjacency matrices for heteroatoms have shown promise.[16] A comparison was made between multivariate analysis of antihypertensive activity of clonidine-type compounds of Timmermans and Zwieten,[17] who used the traditional Hansch-type empirical analysis and a graph theoretical alternative in which weighted path numbers are used as descriptors. A tentatively assigned diagonal entry of $-0.20$ to chlorine atoms sufficed to discriminate among carbon atoms and the chlorine atom and produced a slightly better regression with three descriptors than the best five-parameters equation based on the traditional empirical QSAR descriptors. Moreover, upon "compacting" the information from three path numbers, by using the orthogonalization procedure for molecular descriptors recently introduced,[18] it was possible to satisfy expectations based on the principal component analysis and obtain a regression of a satisfactory quality based on only two graph theoretical descriptors.[19]

Augmented adjacency matrix obtained by assigning nonzero diagonal entries to atoms of different kind has opened an avenue for construction of novel invariants of interest in QSAR studies. Finding *optimal* empirically determined descriptors, however, has not been previously undertaken. One expects nonzero diagonal parameters characteristic of individual heteroatoms eventually to be found and moreover, that they may reflect some structural traits of the heteroatoms involved. But first, before such explorations can be initiated one has to accumulate data, at least for most common heteroatoms: oxygen, nitrogen, chlorine, etc., from a series of studies along the lines here outlined. We initiated recently such work,[20] but efforts of many such studies are needed before formulating rules for construction of semiempirical set of parameters for different atoms.

## SPECIFIC TREATMENT OF OXYGEN ATOMS IN ALCOHOLS

We will describe a search for optimal parameters for heteroatoms by considering oxygen atoms in alcohols for which sufficient and more accurate experimental data is available. Smeeks and Jurs[21] recently reported on prediction of the boiling points of alcohols from molecular structure. They computed normal boiling points (i.e., the boiling points at which the liquid's vapor pressure equals 1 atm,[22]) for some 120 $C_5$–$C_8$ aliphatic alcohols using multiple linear regression analysis employing among other descriptors numerous topological indices calculated directly from the molecular structure. Alcohols offer not only sufficient experimental data of presumably higher accuracy, but since they have been previously studied by alternative approaches, their reexamination allows one to compare regression analyses based on different approaches.

By considering variation of boiling points within isomers we eliminate the dominant role of the molecular size in a regression, which tends to obscure a finer dependence of properties on the mode of skeletal branching. We will therefore first examine only hexanols. There are still numerous additional alcohols (e.g., data on 37 heptanols) which can be used to test the derived optimal parameters on compounds not used in their derivation.

As molecular descriptors we will take weighted paths,[23] in particular the weighted paths of length one and two, $^1\pi$, $^2\pi$ respectively. The weighting algorithm is the same one already used in the construction of the connectivity index, which thus makes $^1\pi$ identical to $^1\chi$. Higher weighted paths $^2\pi$, $^3\pi$, $^4\pi$ correspond to but are numerically somewhat different from the higher connectivity indices $^2\chi$, $^3\chi$, $^4\chi$. Because of a close parallelism with the connectivity indices we may refer to the weighted paths as connectivity indices, in particular when no confusion with the latter can arise. All the path numbers used here were computed from a modified ALLPATH program.[24]

In Table II we show the augmented adjacency matrix for 1-hexanol in which we assigned to all carbon atoms the diagonal element $x = 1.50$ and to the oxygen atom the diagonal entry $y = -0.85$. The lower part of Table II lists in each row the weighted paths and the atomic path sums for the individual atoms. The path numbers for a molecule as a whole is shown in the last row, and finally the total number of weighted paths in the molecule, the molecular ID number,[25] is also given. Let us briefly outline the origin of the numerical values for selected path numbers. The first column, the $p0$ paths, are in fact the atomic weighting factors. Instead of using $1/\sqrt{m}$, where $m$ is the formal valency of a vertex, as one does in the case of ordinary graphs (i.e., when all the diagonal entries are zero), now one uses factors $1/\sqrt{(m + x)}$, where $x$ is the corresponding diagonal entry in the adjacency matrix. Consequently for bond weights (the values of $p1$) instead of $1/\sqrt{mn}$ we have $1/\sqrt{(m + x)(n + x')}$, where $x$, $x'$ are the diagonal entries for the vertices forming the bond.

As the criterion for determining the optimal descriptors we use the standard error $S$ in a multivariate regression of paths numbers against the normal boiling points. Therefore in searching $(x, y)$ space first we choose a pair of values of $x$, $y$ and from ALLPATH program derive the corresponding molecular path numbers for each of 17 hexanols considered for which the experimental data was available (Table III). Subsequently a regression of path numbers against normal boiling points is derived and standard error recorded. The process is repeated till a pair $x$, $y$ is found that is associated with the min-

**Table II.** Augmented adjacency matrix for 1-hexanol with diagonal entries for carbon atoms $x = 1.50$ and oxygen atom $y = -0.85$

Adjacency matrix

| 1.5 | 1 | 0 | 0 | 0 | 0 | 0 |
|-----|-----|-----|-----|-----|-----|-----|
| 1 | 1.5 | 1 | 0 | 0 | 0 | 0 |
| 0 | 1 | 1.5 | 1 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1.5 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 1.5 | 1 | 0 |
| 0 | 0 | 0 | 0 | 1 | 1.5 | 1 |
| 0 | 0 | 0 | 0 | 0 | 1 | -0.85 |

Atomic path numbers:

| Atom | $p0$ | $p1$ | $p2$ | $p3$ | $p4$ | $p5$ | $p6$ | Atomic ID |
|------|------|------|------|------|------|------|------|-----------|
| 1 | 0.5345 | 1.6658 | 0.0816 | 0.0233 | 0.0066 | 0.0023 | | 2.3142 |
| 2 | 0.5345 | 0.5714 | 0.4760 | 0.0233 | 0.0079 | | | 1.6131 |
| 3 | 0.5345 | 0.5714 | 0.1633 | 0.1403 | | | | 1.4095 |
| 4 | 0.5345 | 0.5714 | 0.1782 | 0.0233 | 0.0322 | | | 1.3397 |
| 5 | 0.5345 | 0.6238 | 0.0816 | 0.0233 | 0.0067 | 0.0092 | | 1.2791 |
| 6 | 0.6325 | 0.3381 | 0.0966 | 0.0276 | 0.0079 | 0.0023 | 0.0031 | 1.1079 |
| 7 | 2.5820 | 1.3801 | 0.3943 | 0.1127 | 0.0322 | 0.0092 | 0.0031 | 4.5136 |

| Molecular path numbers | | | | | | | Molecular ID |
|------|------|------|------|------|------|------|-----------|
| 5.8871 | 2.8610 | 0.7358 | 0.1869 | 0.0467 | 0.0114 | 0.0031 | 9.7321 |

**Table III.**  The weighted paths numbers, experimental boiling points, and calculated boiling points for hexanols assuming $x = 1.50$ and $y = -0.85$

|    | Molecule              | $^1\pi$  | BP(exp) | BP(calc) | Diff   |
|----|-----------------------|----------|---------|----------|--------|
| 1  | 1-Hexanol             | 2.86105  | 157.00  | 155.15   | +1.85  |
| 2  | 2-Hexanol             | 2.67677  | 139.90  | 137.67   | +2.23  |
| 3  | 3-Hexanol             | 2.68295  | 135.40  | 138.26   | -2.86  |
| 4  | 2-Methyl-1-pentanol   | 2.80600  | 148.00  | 149.93   | -1.93  |
| 5  | 3-Methyl-1-pentanol   | 2.80600  | 152.40  | 149.93   | +2.47  |
| 6  | 4-Methyl-1-pentanol   | 2.79982  | 151.80  | 149.93   | +1.93  |
| 7  | 2-Methyl-2-pentanol   | 2.49202  | 121.40  | 120.15   | +1.25  |
| 8  | 3-Methyl-2-pentanol   | 2.62571  | 134.20  | 132.83   | +1.37  |
| 9  | 4-Methyl-2-pentanol   | 2.61554  | 131.70  | 131.86   | -0.16  |
| 10 | 2-Methyl-3-pentanol   | 2.62570  | 126.50  | 132.83   | -6.23  |
| 11 | 3-Methyl-3-pentanol   | 2.50261  | 121.50  | 121.15   | +0.35  |
| 12 | 2-Ethyl-1-butanol     | 2.81218  | 146.50  | 150.51   | -4.01  |
| 13 | 2,2-Dimethyl-1-butanol | 2.71340 | 136.80  | 141.15   | -4.35  |
| 14 | 2,3-Dimethyl-1-butanol | 2.74876 | 149.00  | 144.50   | +4.50  |
| 15 | 3,3-Dimethyl-1-butanol | 2.70280 | 143.00  | 140.14   | +3.86  |
| 16 | 2,3-Dimethyl-2-butanol | 2.43762 | 118.60  | 114.99   | +3.61  |
| 17 | 3,3-Dimethyl-2-butanol | 2.52535 | 120.00  | 123.31   | -3.31  |

imal $S$. One should add that the derived optimal parameters will critically depend on the property studied. For different properties different numerical parameters for molecules and same atoms will follow. Moreover a same atom such as oxygen may lead to different optimal diagonal entries even for a same property but in different types of compounds (e.g., alcohols, ethers, aldehydes, etc). This, as is well known, occurs also in semiempirical MO calculations.

From preliminary calculations on clonidine and related compounds[16] containing chlorine as a heteroatom we have seen that negative diagonal entries increase the relative role of heteroatoms, while positive diagonal entries decrease the role of such atoms. This is illustrated in Table IV and Table V where we show the leading path numbers and the molecular ID number for 1-hexanol as a function of the diagonal entry describing oxygen atom and carbon atom respectively. We see a dramatic increase in the magnitudes of the path numbers as the diagonal element approaches the singularity at $y = -1$ (Table

**Table IV.**  Variation of path numbers for 1-hexanol as a function of $y$ (the diagonal entry for oxygen atom)

| $y$   | $^1\Pi$ | $^2\Pi$ | $^3\Pi$ | ID       |
|-------|---------|---------|---------|----------|
| -0.99 | 9.7782  | 4.6391  | 2.1945  | 32.98075 |
| -0.97 | 6.7896  | 3.1448  | 1.4474  | 22.83179 |
| -0.96 | 6.2426  | 2.8713  | 1.3107  | 20.97440 |
| -0.95 | 5.8694  | 2.6847  | 1.2174  | 19.79686 |
| -0.94 | 5.5939  | 2.5469  | 1.1485  | 18.77119 |
| -0.93 | 5.3797  | 2.4399  | 1.0949  | 18.04400 |
| -0.90 | 4.9432  | 2.2216  | 0.9858  | 16.56153 |
| -0.85 | 4.5329  | 2.0164  | 0.8832  | 15.16810 |
| -0.80 | 4.2883  | 1.8941  | 0.8221  | 14.33745 |
| -0.75 | 4.1213  | 1.8107  | 0.7803  | 13.77059 |
| -0.50 | 3.7071  | 1.6036  | 0.6768  | 12.36396 |
| -0.25 | 3.5236  | 1.5118  | 0.6309  | 11.74080 |
| 0     | 3.4142  | 1.4571  | 0.6036  | 11.36932 |
| +0.25 | 3.3396  | 1.4198  | 0.5849  | 11.11581 |
| +0.50 | 3.2845  | 1.3922  | 0.5711  | 10.92868 |

IV). Similarly, as the diagonal entries corresponding to carbon atoms increase the relative magnitudes of the path number decrease (Table V). Observe that with the variation in the diagonal entries the ratio of $^1\pi/^2\pi$ and $^2\pi/^3\pi$ alters. Path numbers are sensitive on the assumed values of the diagonal entries in adjacency matrix as is reflected in the changes not only for $^1\pi$, $^2\pi$, $^3\pi$ shown in Table IV and Table V, but also for their ratios.

The optimal $x$ (carbon atom) and $y$ (oxygen atom) values found when a single molecular descriptor $^1\pi$ (or $^1\chi$) is used are 1.50 and $-0.85$ respectively. The corresponding standard error $S = 3.30$ is less than half of the standard error corresponding to the case when carbon atoms and oxygen atom are not discriminated ($S = 7.86$). In Table VI we give the variations of the standard error $S$ as a function of $x$, $y$. The top row in Table VI represents the diagonal matrix element of the carbon atoms and the entries in the first column indicate the values for diagonal matrix elements of the oxygen atom. Without any prior information on approximate values for the parameters we examined about 50 individual regressions, each time changing the values of $x$ and $y$. Table VI displays a regular variation of $S$ on $x$ and $y$ from which one can locate the minimum for $S(x, y)$ at

**Table V.**  Variation of path numbers for 1-hexanol as a function of $x$ (the diagonal entry for carbon atoms)

| $x$  | $^1\pi$ | $^2\pi$ | $^3\pi$ | ID      |
|------|---------|---------|---------|---------|
| 1.00 | 2.3189  | 0.6619  | 0.1836  | 7.82217 |
| 1.10 | 2.2502  | 0.6218  | 0.1670  | 7.62478 |
| 1.15 | 2.2175  | 0.6032  | 0.1595  | 7.53171 |
| 1.20 | 2.1859  | 0.5854  | 0.1524  | 7.44210 |
| 1.25 | 2.1553  | 0.5685  | 0.1458  | 7.35573 |
| 1.50 | 2.0154  | 0.4942  | 0.1179  | 6.96663 |
| 1.75 | 1.8945  | 0.4341  | 0.0968  | 6.63610 |
| 2.00 | 1.7887  | 0.3847  | 0.0805  | 6.35111 |
| 3.00 | 1.4708  | 0.2542  | 0.0428  | 5.51209 |
| 4.00 | 1.2575  | 0.1818  | 0.0257  | 4.95744 |

**Table VI.** Variations of the standard error $S$ as a function of $x$ and $y$ when a single path descriptors is used

|  | 4 | 3 | 2.5 | 2 | 1.75 | 1.50 | 1.20 | 1.15 | 1.10 | 1.00 | 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| −.99 |  |  |  |  |  |  |  | 4.818 |  |  | 4.744 |
| −.95 |  |  |  |  |  |  |  | 3.770 |  |  | 3.748 |
| −.90 |  |  |  | 3.450 | 3.420 | 3.394 | 3.369 | 3.366 | 3.364 | 3.360 | 3.596 |
| −.85 | 3.474 | 3.380 | 3.339 | 3.309 | 3.302 | 3.301 | 3.315 | 3.319 | 3.324 | 3.336 | 3.832 |
| −.80 |  |  |  | 3.326 | 3.343 | 3.370 | 3.422 | 3.433 | 3.445 | 3.477 | 4.190 |
| −.75 |  |  |  |  |  |  |  | 4.288 |  |  | 4.524 |
| −.50 |  |  |  |  |  |  |  | 4.676 |  |  | 6.138 |
| 0 | 4.772 | 5.092 |  | 5.544 | 5.705 | 5.852 | 6.081 | 6.123 |  | 6.259 | 7.863 |

$x = 1.50$ and $y = -0.85$.

The origin, $x = 0, y = 0$, corresponds to a model in which alcohols are represented by simple graphs, i.e., as objects in which carbon and oxygen are not discriminated. Such a model is not expected to be very good, because among the 17 hexanols, there are compounds that are described by the same simple graph yet differ considerably in their boiling points. When higher path numbers are added as descriptors the relatively large standard error ($S = 7.86$) did not decrease. This is understandable in view of the mentioned inability of such model to discriminate carbon and oxygen atoms. The most left column in Table VI shows how the standard error $S$ decreases when the oxygen atom is discriminated by introduction of a negative diagonal entry in the adjacency matrices of the molecular graphs representing alcohols. The effect of changing diagonal entries for carbon (while leaving oxygen at $y = 0$) is shown in the bottom row of Table VI. Here the effect is less dramatic, though also a significant reduction in $S$ follows with the variation of the diagonal entries. Both alternatives result in a similar effect on the relative magnitudes for $^1\pi$ for oxygen and carbon atom. Apparently $S$ is more sensitive to changes in $y$ then $x$, which in part is a consequence of the form for the weighting function $(1/\sqrt{mn})$, which is steep in the interval (0, 1) and rather shallow for values larger than 1. The case $y = -1$ (and the same would be the case for $x = -1$) represents a singularity, hence the domain for $S(x, y)$ is restricted to $x, y$ values greater than $-1$. Thus as one approaches $-1$ (see the top part of Table IV) the magnitudes of the path numbers increase the computations are more sensitive to smaller changes in $y$. Hence the number of significant digits specifying the optimal value for the parameter increase. In contrast, as $x$—the diagonal entry for carbon atoms—increases, the induced changes in the magnitudes of the path numbers are small (the bottom part of Table V), hence the increments in $x$ are relatively large.

Reducing the standard error $S$ by more than a half represents a considerable improvement of a multivariate regression, usually achieved by introduction of additional descriptors. That this was possible by a "natural" generalization of path numbers is quite pleasing even though the value of $S = 3.30$ may still be somewhat high for practical needs. Looking at

Table III we see that 2-methyl-3-pentanol for no apparent reasons shows a large deviation (6.23°C). A *single* connectivity index (with or without modification of corresponding diagonal entries) cannot differentiate between 2-methyl-3-pentanol and 3-methyl-2-pentanol, both having the same bond types, and therefore both having the same $^1\pi = 2.6257$, yet differing in their boiling point by 8°C. Hence there are intrinsic limitations of a single-descriptor approach. Nevertheless we expect that the derived regression equation for hexanols

$$BP = 94.849\ ^1\pi - 116.22 \qquad (1)$$

will be hard to rival by other *single* variable regressions. Thus it presents a challenge to competing alternative approaches. In particular, the methods for estimating boiling points are generally considered poor.[26] The difficulties are reflected by limitations of schemes using group contributions to reproduce data for isomers.[27]

## USE OF TWO DESCRIPTORS

With use of two descriptors the standard error is expected to decrease. It is of interest to see how standard error $S$ varies with $x, y$ in the case of several descriptors. If we start with the case $x = 0, y = 0$ (i.e., we do not differentiate carbon and oxygen atoms) we witness no important improvement in the regression statistics whether using two descriptors ($S = 7.89$ as compared to $S = 7.86$) or a single descriptor. While disappointing, this result could also be misleading in suggesting no improvement to be possible with the added descriptor. But when the restriction for the values of $x$ and $y$ are lifted and different atoms discriminated we find significant improvements in the regression when using two, rather than a single molecular descriptor. When we use two descriptors $^1\pi$ and $^2\pi$ and the same optimal values for $x$ and $y$ based on a single descriptor we find a reduction of about 10% in $S$, (from the value of 3.30 to $S = 3.07$). The improvement is significant, and of course, not necessarily optimal, since with two descriptors the optimal values for $x$ and $y$ may shift. Hence a fresh search for the optimal $x, y$ when two path numbers are used has to be undertaken.

In Table VII we show at a greater resolution an $x$,

**Table VII.**  Variations of the standard error $S$ as a function of $x$ and $y$ when two path numbers are used as descriptors

|       | .50   | .25   | 0     | −.15  | −.25  | −.35  | −.50  |
|-------|-------|-------|-------|-------|-------|-------|-------|
| −.99  |       |       | 3.473 |       |       |       |       |
| −.97  |       |       | 2.578 |       |       |       |       |
| −.96  |       |       | 2.416 |       |       |       |       |
| −.95  | 2.355 | 2.330 | 2.319 | 2.326 | 2.348 | 2.405 | 2.659 |
| −.94  |       | 2.343 | 2.309 | 2.292 | 2.296 | 2.319 |       |
| −.93  |       |       | 2.352 | 2.325 | 2.309 | 2.305 |       |
| −.90  |       |       | 2.643 |       |       |       |       |

$y$ neighborhood in which one can locate the optimal values for $x$ and $y$. They now occur at $x = -0.15$ and $y = -0.94$ with the standard error $S = 2.29$ (and the regression coefficient $R = 0.985$). The improvement in the standard error when one compares the regressions using optimal parameters for a single and optimal parameters for two descriptors is dramatic. The $S$ value is now cut almost by a factor of ⅓. The optimal $x$ and $y$ obtained for all the diagonal entries (for carbon atoms and oxygen atoms) are now negative, although the value for carbon is very close to zero. The optimal weighted path numbers and the predicted boiling points for the 17 alcohols are listed in Table VIII. The two-descriptor regression equation is

$$BP = 49.221\ {}^1\pi + 12.425\ {}^2\pi - 173.044 \quad (2)$$

The standard error associated with the optimal $x$, $y$ parameters is reasonably small. The predicted boiling points using ${}^1\pi$ and ${}^2\pi$ are practically within the limits of the computed standard error for all alcohols except for a single case, 2,3-dimethyl-1-butanol, the boiling point of which was predicted some 6.5°C too low. If this single structure is eliminated from the regression we obtain

$$BP = 50.414\ {}^1\pi + 13.780\ {}^2\pi - 184.532 \quad (3)$$

with the correlation coefficient of 0.992 and a remarkably low standard error of only 1.70°C. More-

over, *if* we were justified (*vide infra*) to eliminate 2,3-dimethyl-1-butanol as an outlier we could expect further improvement (even if slight) in the regression as a consequence of the revised optimization of the $x$, $y$ values based now on the set of 16 alcohols (that exclude the outlier). In order to appreciate the quality of the above result we give in Table IX a summary of the available results of multivariate analyses of alcohols from other sources. Since a different set of compounds were used in different studies the results cannot be directly compared, but they illustrate the effort (i.e., an increase in the number of descriptors) needed to reduce the standard error by factor of 2 and more. We have to emphasize also that the results of Table IX were based on larger samples: $n = 28$ and $n = 48$ in the case of Kier and Hall study,[13,28] and even $n = 123$ in the study of Smeeks and Jurs.[21] A larger number of alcohols in those studies justified use of the additional descriptors. The outlier 2,3-dimethyl-1-butanol is among half a dozen "the worst cases" among the 123 examined in the predictions of the study of Smeeks and Jurs, with deviation of 4.13°C. On the other hand Kier and Hall show no deviation for calculated $BP$ of this compound, but they adopted as the experimental value of $BP$ for that compound the value 145°C, which, if correct would agree with the predictions in our study or that of Smeeks and Jurs.

It appears that our approach to isomeric variations

**Table VIII.**  The weighted paths numbers, experimental boiling points, and calculated boiling points for hexanols assuming $x = -0.15$ and $y = -0.94$

| Molecule    | ${}^1\pi$ | ${}^2\pi$ | $BP$ (exp) | $BP$ (calc.) | Diff.  |
|-------------|-----------|-----------|------------|--------------|--------|
| 1-OH        | 5.96112   | 2.93004   | 157.00     | 156.77       | +0.23  |
| 2-OH        | 5.37478   | 3.84532   | 139.90     | 139.28       | +0.62  |
| 3-OH        | 5.42471   | 3.30974   | 135.40     | 135.09       | +0.31  |
| 2-M-1-OH    | 5.85299   | 2.72291   | 148.00     | 148.88       | −0.88  |
| 3-M-1-OH    | 5.85299   | 2.95441   | 152.40     | 151.75       | +0.65  |
| 4-M-1-OH    | 5.80307   | 3.12244   | 151.80     | 151.38       | +0.42  |
| 2-M-2-OH    | 4.89890   | 4.43334   | 121.40     | 123.17       | −1.77  |
| 3-M-2-OH    | 5.28707   | 3.63300   | 134.20     | 132.33       | +1.87  |
| 4-M-2-OH    | 5.21674   | 4.04875   | 131.70     | 134.03       | −2.33  |
| 2-M-3-OH    | 5.28707   | 3.26544   | 126.50     | 127.76       | −1.26  |
| 3-M-3-OH    | 4.97772   | 3.86164   | 121.50     | 119.94       | +1.56  |
| 2-E-1-OH    | 5.90292   | 2.57074   | 146.50     | 149.44       | −2.96  |
| 2,2-MM-1-OH | 5.65393   | 2.69797   | 136.80     | 138.77       | −1.93  |
| 2,3-MM-1-OH | 5.71536   | 2.82889   | 149.00     | 143.42       | +6.58  |
| 3,3-MM-1-OH | 5.57511   | 3.36310   | 143.00     | 143.15       | −0.15  |
| 2,3-MM-1-OH | 4.77308   | 4.36848   | 118.60     | 116.17       | +2.43  |
| 3,3-MM-2-OH | 5.02101   | 3.89509   | 120.00     | 122.49       | −2.49  |

**Table IX.** A summary of multivariate regressions on aliphatic alcohols

| Sample size | Number of descriptors | Standard error | Regression coefficient | Ref. |
|---|---|---|---|---|
| $n = 37$ | 1 | 10.3 | 0.956 | a |
| | 2 | 4.4 | 0.992 | |
| | 3 | 1.44 | 0.996 | |
| $n = 48$ | 1 | 8.39 | 0.963 | 13 |
| | 2 | 3.68 | 0.993 | |
| $n = 62$ | 1 | 9.66 | 0.9459 | |
| | 2 | 3.12 | 0.9943 | |
| | 3 | 2.05 | 0.9976 | |
| | 4 | 1.28 | 0.9991 | |
| | 5 | 0.86 | 0.9996 | |
| $n = 123$ | 2 | 4.05 | 0.975 | 21 |
| | 3 | 3.49 | 0.982 | |
| | 4 | 2.79 | 0.989 | |
| | 5 | 2.666 | 0.990 | |
| | 6 | 2.46 | 0.991 | |
| | 7 | 2.24 | 0.993 | |

[a]P.G. Seybold, M. May, and U.A. Bagal, *J. Chem. Educ.*, **64**, 575 (1987).

in boiling points of hexanols, in which diagonal entries of adjacency matrices are used to discriminate among carbon atoms and oxygens, offers a promising route to quantitative relationship of boiling points and molecular structure for heteromolecules in general, and for alcohols in particular. The approach can be extended to three, four, and more descriptors (weighted path numbers), as needed. However, in view of the relatively small sample here considered ($n = 17$) we will not pursue such extensions of the analysis but will rather consider extending the analysis to a larger pool of alcohols.

## HEPTANOLS

To test the outlined approach, and in particular to see if the optimal parameters found for hexanols can be transferred to other alkyl alcohols we will consider data on 37 heptanols, not used in determining the optimal $x$, $y$ values for discrimination of heteroatoms. As expected the case $x = 0$ and $y = 0$ is of no particular interest. Of interest is to see if a use

**Table X.** The weighted paths numbers, experimental boiling points, and calculated boiling points for heptanols assuming $x = 1.50$ and $y = -0.85$

| | Molecule | $^1\pi$ | BP (exp) | BP (calc.) | Diff. |
|---|---|---|---|---|---|
| 1 | 1-Heptanol | 3.14676 | 176.25 | 172.98 | +3.27 |
| 2 | 2-Heptanol | 2.96248 | 159.20 | 156.64 | +2.56 |
| 3 | 3-Heptanol | 2.96867 | 156.80 | 157.18 | -0.38 |
| 4 | 4-Heptanol | 2.96867 | 155.00 | 157.18 | -2.18 |
| 5 | 2-Methyl-1-hexanol | 3.09172 | 164.00 | 168.10 | -4.10 |
| 6 | 3-Methyl-1-hexanol | 3.09172 | 172.00 | 168.10 | +3.90 |
| 7 | 4-Methyl-1-hexanol | 3.09172 | 173.00 | 168.10 | +4.90 |
| 8 | 5-Methyl-1-hexanol | 3.08554 | 172.00 | 167.55 | +4.45 |
| 9 | 2-Methyl-2-hexanol | 2.77774 | 142.50 | 140.25 | +2.25 |
| 10 | 3-Methyl-2-hexanol | 2.91142 | 152.00 | 152.11 | -0.11 |
| 11 | 4-Methyl-2-hexanol | 2.90744 | 151.00 | 151.75 | -0.75 |
| 12 | 5-Methyl-2-hexanol | 2.90126 | 151.00 | 151.21 | -0.21 |
| 13 | 2-Methyl-3-hexanol | 2.91142 | 148.00 | 152.11 | -4.11 |
| 14 | 3-Methyl-3-hexanol | 2.78832 | 142.80 | 141.19 | +1.61 |
| 15 | 4-Methyl-3-hexanol | 2.91760 | 149.00 | 152.65 | -3.65 |
| 16 | 5-Methyl-3-hexanol | 2.90744 | 148.00 | 151.75 | -3.75 |
| 17 | 2-Ethyl-1-pentanol | 3.09790 | 166.00 | 168.64 | -2.64 |
| 18 | 2,2-Dimethyl-1-pentanol | 2.99911 | 153.00 | 159.88 | -6.88 |
| 19 | 2,3-Dimethyl-1-pentanol | 3.04065 | 164.00 | 163.57 | +0.43 |
| 20 | 2,4-Dimethyl-1-pentanol | 3.03049 | 159.00 | 162.67 | -3.67 |
| 21 | 3,3-Dimethyl-1-pentanol | 2.99911 | 165.00 | 159.88 | +5.12 |
| 22 | 3,4-Dimethyl-1-pentanol | 3.03447 | 165.00 | 163.02 | +1.98 |
| 23 | 4,4-Dimethyl-1-pentanol | 2.98852 | 160.00 | 158.94 | +1.06 |
| 24 | 3-Ethyl-2-pentanol | 2.91760 | 152.00 | 152.65 | -0.65 |
| 25 | 2,3-Dimethyl-2-pentanol | 2.72951 | 140.00 | 135.97 | +4.03 |
| 26 | 2,4-Dimethyl-2-pentanol | 2.71650 | 133.00 | 134.82 | -1.82 |
| 27 | 3,3-Dimethyl-2-pentanol | 2.82165 | 147.00 | 144.14 | +2.86 |
| 28 | 3,4-Dimethyl-2-pentanol | 2.85418 | 153.00 | 147.03 | +5.97 |
| 29 | 4,4-Dimethyl-2-pentanol | 2.80424 | 138.00 | 142.60 | -3.40 |
| 30 | 3-Ethyl-3-pentanol | 2.79891 | 142.40 | 142.13 | +0.27 |
| 31 | 2,2-Dimethyl-3-pentanol | 2.81725 | 136.00 | 143.75 | -7.75 |
| 32 | 2,3-Dimethyl-3-pentanol | 2.73392 | 140.00 | 136.36 | +3.64 |
| 33 | 2,4-Dimethyl-3-pentanol | 2.85418 | 138.80 | 147.03 | -8.23 |
| 34 | 2,2-Methylethyl-1-butanol | 3.00970 | 157.00 | 160.82 | -3.82 |
| 35 | 2,2,3-Trimethyl-2-butanol | 2.94470 | 157.00 | 155.06 | +1.94 |
| 36 | 2,3,3-Trimethyl-1-butanol | 2.94030 | 160.00 | 154.67 | +5.33 |
| 37 | 2,3,3-Trimethyl-2-butanol | 2.63118 | 131.00 | 127.25 | +3.75 |

**Table XI.** The weighted paths numbers, experimental boiling points and calculated boiling points for heptanols assuming $x = -0.15$ and $y = -0.94$

| | Molecule | $^1\pi$ | $^2\pi$ | BP (exp) | BP (calc.) | Diff. |
|---|---|---|---|---|---|---|
| 1 | 1-OH | 6.50166 | 3.22222 | 176.25 | 173.96 | +2.29 |
| 2 | 2-OH | 2.91532 | 4.13751 | 159.20 | 157.49 | +1.71 |
| 3 | 3-OH | 5.96525 | 3.60192 | 156.80 | 154.66 | +2.14 |
| 4 | 4-OH | 5.96525 | 3.62891 | 155.00 | 154.91 | +0.09 |
| 5 | 2-M-1-OH | 6.39353 | 3.01509 | 164.00 | 167.44 | −3.44 |
| 6 | 3-M-1-OH | 6.39353 | 3.27358 | 172.00 | 169.83 | +2.17 |
| 7 | 4-M-1-OH | 6.39353 | 3.24660 | 173.00 | 169.58 | +3.42 |
| 8 | 5-M-1-OH | 6.34361 | 3.41462 | 172.00 | 169.02 | +2.98 |
| 9 | 2-M-2-OH | 5.43944 | 4.72553 | 142.50 | 142.70 | −0.20 |
| 10 | 3-M-2-OH | 5.82761 | 2.91142 | 152.00 | 152.05 | −0.05 |
| 11 | 4-M-2-OH | 5.80720 | 4.17291 | 151.00 | 153.22 | −2.22 |
| 12 | 5-M-2-OH | 5.75728 | 2.90126 | 151.00 | 152.55 | −1.55 |
| 13 | 2-M-3-OH | 5.82761 | 3.58461 | 148.00 | 148.64 | −0.64 |
| 14 | 3-M-3-OH | 5.51826 | 4.19643 | 142.80 | 141.15 | +1.65 |
| 15 | 4-M-3-OH | 5.87754 | 3.40712 | 149.00 | 149.13 | −0.13 |
| 16 | 5-M-3-OH | 5.80720 | 3.80535 | 148.00 | 149.82 | −1.82 |
| 17 | 2-E-1-OH | 6.44346 | 2.88991 | 166.00 | 168.40 | −2.40 |
| 18 | 2,2-MM-1-OH | 6.19447 | 3.03276 | 153.00 | 159.14 | −6.14 |
| 19 | 2,3-MM-1-OH | 6.30582 | 2.97056 | 164.00 | 163.30 | +0.70 |
| 20 | 2,4-MM-1-OH | 6.23549 | 3.21852 | 159.00 | 162.60 | −3.60 |
| 21 | 3,3-MM-1-OH | 6.19447 | 3.39828 | 165.00 | 162.52 | +2.48 |
| 22 | 3,4-MM-1-OH | 6.25590 | 3.37956 | 165.00 | 164.96 | +0.04 |
| 23 | 4,4-MM-1-OH | 6.11562 | 3.65528 | 160.00 | 161.55 | −1.55 |
| 24 | 3-E-2-OH | 5.87754 | 3.81752 | 152.00 | 152.92 | −0.92 |
| 25 | 2,3-MM-2-OH | 5.36354 | 4.52029 | 140.00 | 137.57 | +2.43 |
| 26 | 2,4-MM-2-OH | 5.28139 | 4.93534 | 133.00 | 137.92 | −4.92 |
| 27 | 3,3-MM-2-OH | 5.64037 | 3.94324 | 147.00 | 144.00 | +3.00 |
| 28 | 3,4-MM-2-OH | 5.68998 | 4.06531 | 153.00 | 147.24 | +5.76 |
| 29 | 4,4-MM-2-OH | 5.52932 | 4.58798 | 138.00 | 145.25 | −7.25 |
| 30 | 3-E-3-OH | 5.59708 | 3.65645 | 142.40 | 139.51 | +2.91 |
| 31 | 2,2-MM-3-OH | 5.61147 | 3.67934 | 136.00 | 140.33 | −4.33 |
| 32 | 2,3-MM-3-OH | 5.39244 | 4.14453 | 140.00 | 135.31 | +4.69 |
| 33 | 2,4-MM-3-OH | 5.68998 | 3.54747 | 138.80 | 142.44 | −3.64 |
| 34 | 2,2-ME-1-OH | 6.27330 | 2.76487 | 157.00 | 160.01 | −3.01 |
| 35 | 2,2,3-M$_3$-2-OH | 6.06866 | 3.09210 | 157.00 | | |
| 36 | 2,3,3-M$_3$-1-OH | 6.03976 | 3.32978 | 160.00 | 155.31 | +4.69 |
| 37 | 2,3,-M$_3$-2-OH | 5.10432 | 4.78094 | 131.00 | 128.96 | +2.04 |

of the optimal $x$, $y$ derived for hexanols will lead to relatively small standard errors in the case of the heptanols. Also it would be of interest to find if use of additional descriptors (with the revised optimal values of $x$ and $y$) would lead to a significantly reduced standard error when compared with the regressions based on fewer descriptors.

In Table X we listed 37 aliphatic heptanols, the weighted path numbers (using $x = 1.50$ and $y = -0.85$), and the predicted boiling points. The regression equation

$$BP = 88.695\ ^1\pi - 106.124 \qquad (4)$$

with the regression coefficient $R = 0.944$ and the standard error $S = 3.91$ may be viewed as comparable to the similar results for hexanols. Thus the first question raised is answered affirmatively. Except for 2,4-Dimethyl-3-pentanol none of the predicted $BP$ deviate by more than two standard errors from the experimental value, though about a dozen show deviations exceeding the standard error. More critical for a success of the approach is the analysis

of the regression data based on two descriptors $^1\pi$ and $^2\pi$, shown in Table XI. The regression equation:

$$BP = 42.528\ ^1\pi + 9.257\ ^2\pi - 132.372 \qquad (5)$$

gives $R = 0.963$ and $S = 3.25$ clearly indicates that the parametrization derived from hexanols *extends* to heptanols as well. The drop in the standard error $S$ from 3.91 to 3.25 is appreciable, even if not dramatic. Except for the predicted $BP$ of 4,4-dimethyl-2-pentanol (145.25°C), which is too large (and which was similarly predicted in reference 18 at a somewhat larger vlaue of 142.17°C) other predictions are within twice the standard error. Thus we conclude that optimal parameters are *transferable*, at least among alcohols of different size, i.e., similar molecular groupings are apparently well characterized by the same heteroatom parameters.

In order to better illustrate the transferability of the optimal parameters from hexanols to heptanols in Figure 1 we show the computed boiling points from eqs. (1) and (4) respectively, which suggests almost parallel regression lines. There is a small dif-
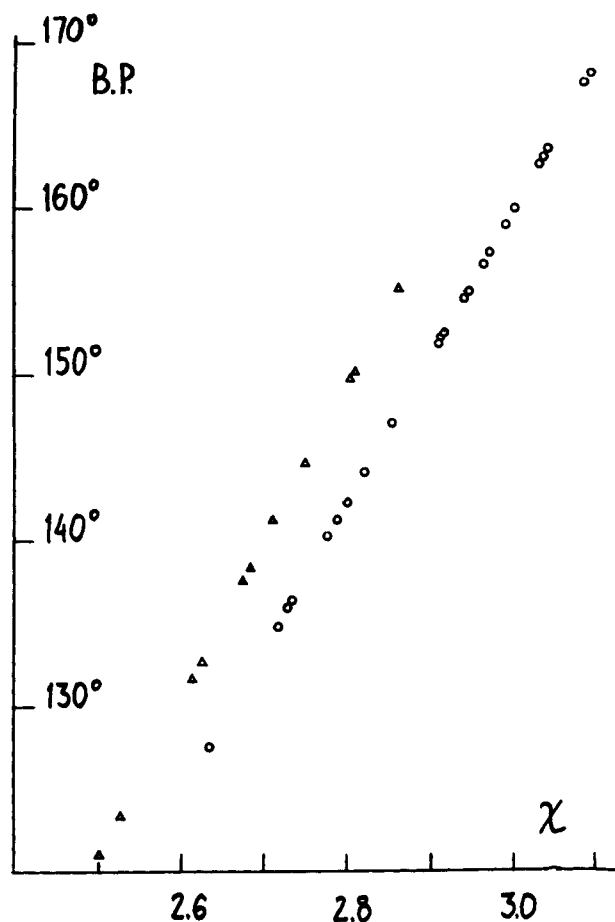
**Figure 1.** The plot of calculated boiling points for hexanols (triangles) and heptanols (circles), eqs. {1} and {4} respectively.

ference in the slope of the two regression lines (barely noticeable in the figure, but reflected in eqs. (1) and (4) in the different coefficients of $^1\pi$), which suggests a small quadratic term. More apparent is a shift of the two lines which clearly indicates that the path numbers (the connectivity indices) $^1\pi$, while adequately characterizing the molecular branching, require a "correction" (of about 0.80) for the size difference among hexanes and heptanes. We see from Figure 1 that there is a price to be paid for using a single descriptor when correlating properties of molecules of different size. Since the molecular size is a well understood property (and not difficult to characterize) one can combine the regressions on hexanols and heptanols into a single regression in which size can be taken into account by introducing an appropriate "size-correction" of the connectivity indices.

## DISCUSSION

Structure-property studies and QSAR suffer from lack of *comparative* studies, i.e., studies on a same set of compounds and a same property but based on different descriptors or different methodology. Such

comparative studies could point to advantages of alternative molecular descriptors. Typically, individual structure-property-activity studies are based on use of different descriptors and apply to a selection of different compounds. The time has arrived to challenge chemometricians, medicinal chemists, and graph theorists to test their designs on a same set of standard compounds. In this way we could expect answers to questions, such as

- What is the best *single* descriptor for a correlation of a selected property (e.g., the boiling points, molar refraction, entropy, etc.) for a selected class of compounds (such as alkanes, alcohols, amines, etc)?
- What is the best *two-parameter* regression for the same properties and same compounds?
- What is the best *three-parameter* regression, etc.?

Kier and Hall in their book[28] briefly considered the boiling points of $C_2$-$C_7$ alkanes and reviewed the connectivity index,[14] the Wiener number,[5] and the Hosoya's $Z$ index[29] as *single* molecular descriptors. The following standard deviations for the three descriptors were reported: 6.74°C, 16.4°C, and 21.0°C respectively. This shows clearly for this particular property, the superiority of the connectivity index $\chi$ over $W$ and $Z$, if used alone. Seybold and co-workers recently[30] extended such comparative study by considering half a dozen physicochemical properties of alkanes and reported regression equations involving from one to three best descriptors. In the case of alcohols Smeeks and Jurs[21] tested over 70 various molecular descriptors that fell into several categories, numerical (e.g., counting terminal methyl groups), characterizing substructures, environment (e.g., atomic charges) and atom-atom connectivity. The best regressions using from two to seven descriptors were reported and the standard errors found varied from 4.05°C to 2.24°C. Incidentally most of the "finalists," i.e., the best descriptors found, were of graph theoretical origin. Reporting such results is very important as they offer a clear challenge to model designers to develop better alternative characterizations, if possible. Neither the results of Kier and Hall, nor those of Smeeks and Jurs, despite being rather ambitious, are necessarily optimal, but merely the best currently available.

One should not necessarily only be concerned to arrive at a better regression equation, measured by $R$ and particularly $S$, i.e., regression accompanied with a smaller standard error or use fewer descriptors. Rather the emphasis ought to be to arrive at better structural descriptors, i.e., descriptors that cannot only better "accommodate" the variability of data but also provide a satisfactory structural *interpretation* of a multivariate regression. Such descriptors are likely to be of interest in characterization of other molecular properties. One may furthermore heighten the focus of such studies by considering

only regressions involving molecules of a same size. For many physicochemical properties the size is often the dominant structural factor. It can, however, be well represented by simple molecular descriptors such as the number of atoms in a molecule or molecular weight. On the other hand it usually obscures the role of other structural factors, such as those that govern isomeric variations in properties. Hence, in order to better understand the role of other structural components that influence the variations in molecular properties with variations in the molecular shape, skeletal branching, conformation, etc., we need descriptors that could give satisfactory characterizations of these subtle molecular features. In this way we would better understand the role of different structural components and eventually be more successful in designs of molecules of prescribed properties.

## References

1. M.C. Kohn, *Bull. Math. Biol.*, **48**, 417 (1986).
2. E. Kovats, *Zeit. Anal. Chem.*, **181**, 351 (1961).
3. C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, **86**, 1616 (1964).
4. L.P. Hammett, *Physical Organic Chemistry*, McGraw Hill, New York, 1940.
5. H. Wiener, *J. Am. Chem. Soc.*, **69**, 17 (1947).
6. F. Harary, *Graph Theory*, Addison-Wesley, Reading, Mass., 1969.
7. J.R. Platt, *J. Phys. Chem.*, **15**, 419 (1947).
8. L. Pauling, *The Nature of the Chemical Bond*, Cornell Univ. Press, Ithaca, NY, 1960.
9. C.A. Coulson, *Valence*, Oxford Univ. Press, London, 1952.
10. F.E. Yates, *Amer. J. Physiol.*, **234**, R159 (1978).
11. N. Trinajstić, M. Randić, and D.J. Klein, *Acta Pharm. Jugosl.*, **36**, 267 (1986).
12. C. Hansch, *Acc. Chem. Res.*, **2**, 232 (1969).
13. L.B. Kier and L.H. Hall, *J. Pharm. Sci.*, **65**, 1806 (1976).
14. M. Randić, *J. Am. Chem. Soc.*, **97**, 6609 (1975).
15. L.B. Kier, W.J. Murray, M. Randić, and L.H. Hall, *J. Pharm. Sci.*, **65**, 1226 (1976).
16. M. Randić, *Chemometrics & Intel. Lab. Syst.*, **10**, 213 (1991).
17. B.M.W.M. Timmermans and P.A. van Zwieten, *J. Med. Chem.*, **20**, 1636 (1977).
18. M. Randić, New J. Chem. (in press). M. Randić, *J. Chem. Inf. & Comput. Sci.* **31**, 311 (1991). M. Randić, Croat. Chem. Acta **64**, 43 (1991).
19. M. Randić, New J. Chem. (submitted). M. Randić and S. Basak, *Math. Comput. Modelling* (to be submitted).
20. M. Randić, *Quant. Struct.-Act. Relat.* (submitted).
21. F.C. Smeeks and P.C. Jurs, *Anal. Chim. Acta*, **233**, 111 (1990).
22. I.N. Levine, *Physical Chemistry* (2nd ed.), McGraw-Hill, New York, 1982.
23. M. Randić, *Int. J. Quant. Chem: Quant. Biol. Symp.*, **11**, 137 (1984).
24. M. Randić, G.M. Brissey, R.B. Spencer, and C.L. Wilkins, *Comput. Chem.*, **3**, 5 (1979). M. Randić, G.M. Brissey, R.B. Spencer, and C.L. Wilkins, *Comput. Chem.*, **4**, 27 (1980).
25. M. Randić, *J. Chem. Inf. Comput. Sci.*, **24**, 164 (1984).
26. R.C. Reid, J.M. Prausnitz, and T.K. Sherwood, *Properties of Gases and Liquids* (3rd ed.), McGraw-Hill, New York, 1977.
27. W.J. Lyman, W.F. Reehl, and D.H. Rosenblatt, *Handbook of Chemical Property Estimation Methods*, McGraw-Hill, New York, 1982.
28. L.B. Kier and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic, New York, 1976.
29. H. Hosoya, *Bull. Chem. Soc. Jpn*, **44**, 2332 (1971).
30. D.E. Needham, I-C. Wei, and P.G. Seybold, *J. Am. Chem. Soc.*, **110**, 4186 (1988).