Search for Useful Graph Theoretical Invariants of Molecular Structure

MILAN RANDIĆ,† PETER J. HANSEN,‡ and PETER C. JURS*

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802

Received July 10, 1987

We have reexamined several graph theoretical invariants that have been used in the past in discussions of structure–property and structure–activity correlations in the search for optimal single-variable descriptors for molecular structures. We found that judicious selection of the functional form for a correlation can lead to significant improvements in the correlations. The approach is illustrated with the connectivity index, the molecular ID number, the Wiener number, and the Hosoya topological index. We review the most suitable forms for the considered indices for correlations with boiling point. These findings have been further supported by considering alternative (polynomial expansion) models. The best results are about 2–3 times better (when measured by the magnitude of the standard deviation) than the best considered previously.

INTRODUCTION

The starting point in many theoretical considerations concerns the choice of the basis for the description of the system under examination. In quantum chemical computations the basis means the selection of the orbitals (atomic) from which one builds molecular wave functions, and the orbitals considered in the past included Slater-type atomic orbitals, double-zeta STO, Hermite-type extensions of STO, Gaussian atomic orbitals and contracted Gaussian sets, Hylaraas orbitals, geminals, etc. In various problems different bases offer different advantages, and it is up to the practitioner to select the best basis for the problem considered. In mathematical and physical computations (including among others the evaluation of molecular integrals) one similarly considers alternative coordinate systems, and in addition to the more common Cartesian and polar coordinates, one may in different applications choose between confocal ellipsoidal, confocal paraboloidal, conical cylindrical, ellipsoidal, elliptic cylindrical, oblate spheroidal, parabolic, parabolic cylindrical, prolate spheroidal, or toroidal and of course the even more general curvilinear coordinates if the problem warrants. It is generally considered advantageous to have such a variety, and one does not expect objections if for a selected problem one considers yet another coordinate system as more suitable. In contrast, when it comes to the problem of describing molecular structures and novel molecular descriptors are introduced, one is defensive and under pressure not to proliferate yet another index or parameter. When a new coordinate system is proposed, it is done so in connection with solving a particular problem, and one expects that the new approach is either simpler, adds to the problem some insights, or solves a problem that was not solvable with alternative schemes. One does not invent a coordinate system for coordinate system's sake, however, and the same prudence must accompany the introduction of novel structural (and graph theoretical) invariants. In Table I we list several graph theoretical invariants, indicating their origin and structural foundation. Not all the invariants, graph theoretical indices, and more general mathematical objects (polynomials, sequences) have been used as extensively as others. Some, like the connectivity indices, have been discussed and applied in hundreds of papers, while others have been used very little. The number of uses of a single index is not, of course, the best criterion for its evaluation, but on the other hand the evaluation of indices that have not been used is hardly warranted. In this paper we critically examine a selection of graph theoretical indices to see if they offer optimal molecular descriptions, and if not, in which way they ought to be modified. The purpose of this paper is not to develop a recommendation for any specific index but to outline a way of testing indices. By use of the language of coordinate systems, this amounts to the outline of an approach that can guide one in selecting one or a very few suitable coordinate systems for a problem of interest. In analytical calculus and its applications, the problem of selection of the coordinate system usually does not arise, because all the coordinate systems (listed above) have simple symmetry properties and this usually suffices to make the choice obvious. But in structure–property or structure–activity studies it is by no

Table I. Selection of Graph Theoretical Indices, Their Origin, and Their Structural Foundation

graph theoretical invariant	name/symbol	comments and ref
path numbers	P_i	ref 25
nonadjacent bonds (matching)	$p(G,k)$	ref 3 and 7
	$Z = \sum_k p(G,k)$	
sum of path lengths	W	ref 2
sum over weighted bonds	connectivity index χ	ref 1; based on discrimination of bond types (m,n)
sum over atom weighted paths	higher connectivity indices,^{m}	ref 26
weighed average distances between atoms	J	ref 12; defined similarly to χ but distance matrix is used
sum over weighted path lengths	identification number ID	ref 18; becomes same as W if bond weighting is eliminated
count of bonds relative to extremes	shape attribute κ	ref 27
generalized weighted bonds sum	$\chi(k)$	ref 21
index (largest) eigenvalue		ref 14
pruning sequence of terminal vertices	centric index	ref 13
count of paths of length 2 relative to extreme isomers	higher shape attributes $^m\chi$	ref 28
modified connectivity	χ'	this work; exponent in $(mn)^k$ varied

† On sabbatical leave from the Department of Mathematics and Computer Science, Drake University, Des Moines, IA 50311, and Ames Laboratory—DOE, Iowa State University, Ames, IA 50011.

‡ On sabbatical leave from the Department of Chemistry, Northwestern College, Orange City, IA 51041.

0095-2338/88/1628-0060$01.50/0

Table II. Experimental Boiling Points for Lower Alkanes and Predicted Values Based on Simple Linear Regression for the Connectivity Index χ, Wiener Number W, and Hosoya Topological Index Z (Revised from Ref 4).

compound	bp (obsd), °C	Randić index (χ)	bp (calcd), °C	Wiener no. (W)	bp (calcd)	Hosoya index (Z)	bp (calcd), °C
ethane	−88.63	1.000	−70.51	1	−36.27	2	−26.16
propane	−42.07	1.414	−40.24	4	−27.97	3	−18.31
2-methylpropane	−11.73	1.732	−16.99	9	−14.13	4	−10.46
n-butane	−0.50	1.914	−3.68	10	−11.36	5	−2.61
2,2-dimethylpropane	9.50	2.000	2.61	16	5.24	5	−2.61
2-methylbutane	27.85	2.270	22.35	18	10.78	7	13.09
n-pentane	36.07	2.414	32.88	20	16.37	8	20.94
2,2-dimethylbutane	49.74	2.561	43.63	28	38.46	9	28.79
2,3-dimethylbutane	57.99	2.643	49.62	29	41.22	10	36.64
2-methylpentane	60.27	2.770	58.91	32	49.53	11	44.49
3-methylpentane	63.28	2.808	61.69	31	46.76	12	52.34
n-hexane	68.74	2.914	69.44	35	57.83	13	60.19
2,2-dimethylpentane	79.20	3.061	80.19	46	86.57	14	68.04
2,4-dimethylpentane	80.50	3.126	84.94	48	93.81	15	75.89
2,2,3-trimethylbutane	80.88	2.943	71.56	42	77.20	13	60.19
3,3-dimethylpentane	86.03	3.121	84.58	44	82.74	16	83.74
2,3-dimethylpentane	89.78	3.181	88.96	46	88.28	17	91.59
2-methylhexane	90.05	3.270	95.47	52	104.88	18	99.44
3-methylhexane	91.85	3.308	98.25	50	99.35	19	107.30
3-ethylpentane	93.48	3.346	101.03	48	93.81	20	115.15
n-heptane	98.42	3.414	106.00	56	115.92	21	123.00
correlation coefficient			0.9914		0.9432		0.9128
estimated standard deviation			6.746		17.09		21.01

means obvious which index or indices may be best for a particular study. The question then is, how should one choose among graph descriptors?

We will consider the boiling points of alkanes, specifically all alkanes from C_2 to C_7, a total of 21 structures. The particular property of boiling point is used for illustrative purposes only. Our aim is not to suggest a better correlation for the boiling points of alkanes (for this it would be advisable to enlarge the sample by including octanes, possibly nonanes, decanes, etc.). Our aim is to outline the strategy for selecting graph theoretical indices (invariants), to examine more closely why some give better answers than others, and to seek conditions that may improve the correlations.

We start with the connectivity index, χ,[1] the Wiener number, W,[2] and the Hosoya topological index,[3] Z, and show in Table II a comparison of the direct use of these three indices for predicting the boiling points of alkanes. Table II was originally reported in the book by Kier and Hall.[4] These authors reported the Wiener index—defined as the sum of the Wiener number (W) and the number of carbon atoms separated by three bonds (polarity number p)—but we have used the Wiener number alone (W). We have ordered the structures in ascending order with respect to their boiling points, and inspection of the predicted values for the boiling points on the basis of the three descriptors—the connectivity index χ, W, and Z—shows few discrepancies as to order. Hence, if one takes as a criterion the prediction of the order of the boiling points of the compounds, all three indices perform similarly (with slightly better performance for the connectivity index and Z). This is an interesting result in view of the fact that the three descriptors have widely different and unrelated structural origins.

The connectivity index, χ, assumes bond additivity, with the proviso that bonds have different weights so that bonds whose terminal atoms have more adjacent neighbors make a smaller contribution. The weights are given by $(mn)^{-1/2}$ where m,n indicate the number of neighbors (neglecting hydrogens) for the terminal carbon atoms. This particular form was adopted as it represents a simple solution of a set of inequalities constructed to reproduce the correct ordering of the structures.[1]

The Wiener number, W, is one of two parameters introduced by Wiener to give correlations with thermodynamical properties of alkanes (including boiling points) and other homologous compounds (e.g., fatty acids). Hence, its use here is somewhat improper since it was not meant to be used alone. Nevertheless, it is a well-defined parameter, representing the sum of the lengths of all paths in a molecule, where paths are counted in their graph theoretical sense.[5] Platt[6] tried to interpret W and suggested that W is a measure of molecular volume. The number can be derived from the distance matrix[5] simply as the sum of all distances in a structure, that is, the sum of the distances between all pairs of carbon atoms (for acyclic structures, such as the alkanes considered here).

Finally, Hosoya's Z number is based on a count of nonadjacent bonds. If $p(G,k)$ represents the number of different ways of selecting k nonadjacent bonds in graph G (e.g., a structure), then Z is the sum of $p(G,k)$ for k equal to zero, up to half the number of atoms [by definition, $p(G,0) = 1$ and $p(G,1)$ = number of bonds]. This was the first so-called topological index (more correctly referred to as a graph theoretical index) that was designed for representing structures. Hence, Hosoya was the first to consider the nontrivial problem of representing a structure by a single number of structural origin. Although originally motivated by the needs of chemical documentation, it was immediately recognized for its potential in structure–property studies. The numbers $p(G,k)$ emerged in the polynomials that Heilmann and Lieb[7] constructed in statistical mechanics. When k is a maximum, $p(G,k)$ becomes the number of arrangements of dimers on a grid, considered explicitly by Fowler and Rushbrooke as early as 1937;[8] they also represent the number of Kekulé valence structures if G is a molecular skeleton of a conjugated hydrocarbon.

It is remarkable that three indices that are so different in content and form can reproduce the ordering of the same physicochemical property for a group of small alkanes so well. It is also remarkable that a single number appears to capture some essential structural features, in fact different features in each of the three cases, and offers a basis for further quantitative correlations. Because the three parameters give the same ordering (or nearly the same), one expects that they will be mutually correlated. In fact, as Heilbronner and Schmelzer[9] have pointed out, even random numbers give relatively high correlations if one has ordered them prior to use. Hence, the correlation between various indices is bound to be high for any indices that offer an acceptable ordering of structures. There have been several comparative studies

of the selection of topological indices showing a high degree of mutual correlation,[10] and we will not duplicate this work. In passing, we should mention that high intercorrelation does not mean that individual indices are related beyond their having a dominant component in common. What makes indices individual is in how they differ and, in particular, whether the differences characterize distinct structural traits. It is commonly accepted that "...the presence of serious multicollinearity often does not affect the usefulness of the fitted model for making inferences about mean responses or making predictions, provided that the values of the independent variables for which inferences are to be made follow the same multicollinearity patterns as the data on which the regression model is based".[11]

We will now examine Table II more closely. Both the correlation coefficient and the standard deviation for the simple regression with the boiling points are superior for the connectivity index. But is this particular comparison fully unbiased? Clearly, comparing χ, W, and Z reveals that the first index is associated with a 2–3 times smaller standard deviation. However, this by no means disqualifies W and Z because each can be expanded to give better nonlinear or two or three parameter expressions. Alternatively, a topological index can be combined with other indices, such as was the case in the pioneering work of Wiener when he also considered P, the number of paths of length three, a measure of crowdedness. Hence, all that Table II says is that when simple linear regression is employed, the connectivity index performs best. In a way none of the standard deviations of Table II (6.746, 17.09, and 21.01 for χ, W, and Z, respectively) are very good values, and many of the individual predicted boiling points are far too large or too small to be of much use. Can these indices be so poor and yet be useful in selected studies? Figure 1 illustrates the relationships of all three indices to boiling point. One can immediately see that all three indices are very reasonable structural descriptors. By choosing a simple linear regression, one distorts the correlative value considerably, and one should not pass judgment on the suitability of a particular index or even about the preference for one or another. Of interest here is the discovery of suitable criteria for the evaluation of various indices and, second, the best single-parameter expressions. We will first consider the problem of the best single-index expression for the correlation with alkane boiling points, but with the same restriction imposed in Table II; that is, we are looking for the best index for a simple linear regression model.

LINEAR REGRESSION MODEL

Other graph invariants besides χ, W, and Z have been reported. Several reviews considered alternative indices, such as a generalization of the connectivity index suggested by Balaban[12] called the distance sum connectivity index, which is defined similarly to the connectivity index, except that instead of m,n (the number of neighbors for the terminal carbon atoms of each bond) one considers the extended connectivity as derived from the distance matrix. Another index, also due to Balaban, is referred to as the centric index and has been found to give the best correlation when octane numbers are considered.[13] The maximal eigenvalue (of the characteristic polynomial) is also an index encoding different structural traits,[14] and average path number provides yet another variable of interest for describing molecular structure.[15] This list is by no means exhaustive, and the number of such indices can be doubled by simply using the Shannon[16] equation and deriving the information content of each index.[17] Rather than following the study of indices that have been reviewed in the past, we decided to look for some novel alternatives and modifications.

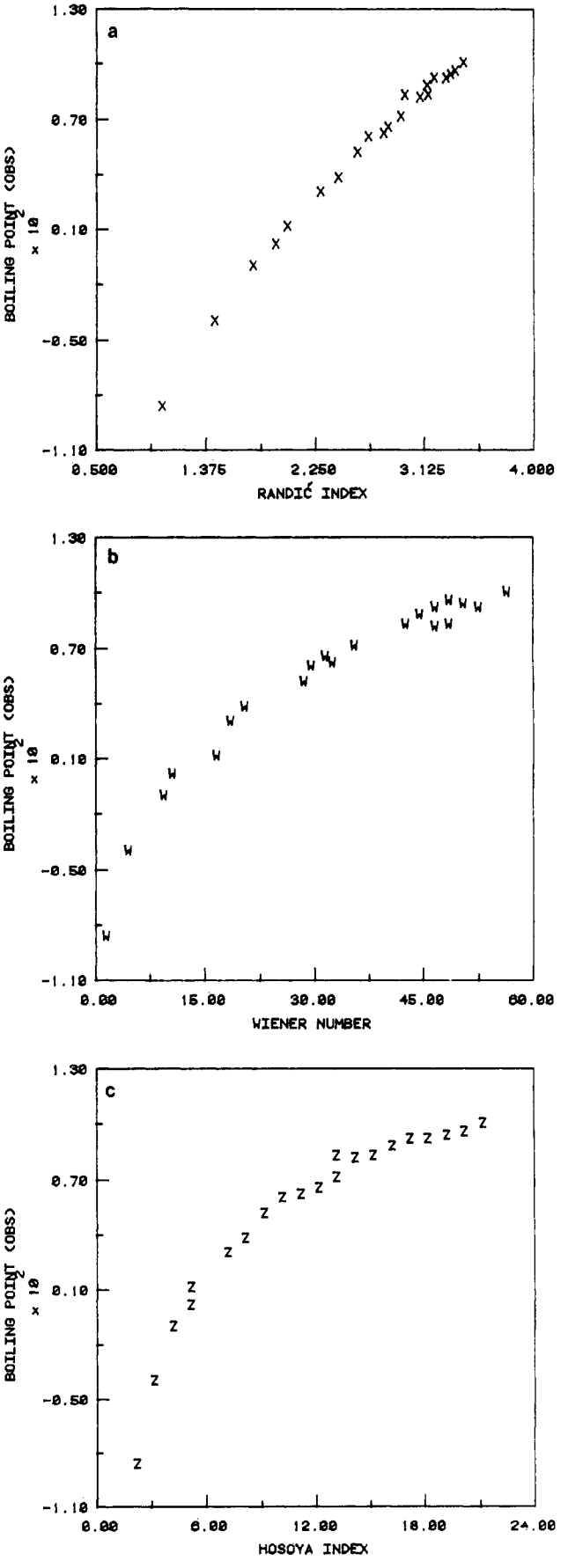

Figure 1. Correlations between the observed boiling points of the C_2–C_7 alkanes and (a) the connectivity index, (b) the Wiener number, and (c) the Hosoya topological index. Observe the increasing curvature for the three correlations that explains the large differences in the correlation coefficients (r) and the standard deviations (s) shown in Table II.

Table III. Correlation between the Experimental Boiling Points for C_2–C_7 Alkanes of Table II with (a) the Molecular ID Numbers, (b) the Connectivity Indices Based on Shifted Values for the Valencies, and (c) the Connectivity Indices Based on Variations in k, the Exponent in $(mn)^k$ [a]

compound	ID	$m - {}^1/_2$	$m + {}^1/_2$	$m + 1$	$k = -{}^1/_3$	$k = -{}^1/_4$
ethane	3.0000	2.0000	0.6667	0.5000	1.0000	1.0000
propane	4.9142	2.3094	1.0328	0.8164	1.5874	1.6817
2-methylpropane	6.7321	2.6832	1.3092	1.0605	2.0801	2.2795
n-butane	6.8713	2.9761	1.4328	1.1497	2.2174	2.3889
2,2-dimethylpropane	8.5000	3.0236	1.5396	1.2648	2.5198	2.8284
2-methylbutane	8.6968	3.4599	1.7273	1.4039	2.7307	2.9995
n-pentane	8.8499	3.6428	1.8328	1.4830	2.8473	3.0960
2,2-dimethylbutane	10.4660	3.8588	1.9692	1.6150	3.1836	3.5568
2,3-dimethylbutane	10.5236	3.9776	2.0313	1.6640	3.2542	3.6167
2-methylpentane	10.6792	4.1266	2.1273	1.7372	3.3607	3.7066
3-methylpentane	10.6759	4.2366	2.1454	1.7473	3.3814	3.7195
n-hexane	10.8391	4.3095	2.2328	1.8163	3.4773	3.8031
2,2-dimethylpentane	12.4490	4.5255	2.3692	1.9483	3.8135	4.2639
2,4-dimethylpentane	12.5092	4.6104	2.4218	1.9914	3.8741	4.3172
2,2,3-trimethylbutane	12.2931	4.3946	2.2795	1.8792	3.7134	4.1782
3,3-dimethylpentane	12.4427	4.6892	2.3988	1.9652	3.8473	4.2852
2,3-dimethylpentane	12.5052	4.7543	2.4494	2.0074	3.9049	4.3367
2-methylhexane	12.6704	4.7933	2.5273	2.0705	3.9907	4.4137
3-methylhexane	12.6600	4.9033	2.5454	2.0806	4.0114	4.4266
3-ethylpentane	12.6692	5.0133	2.5635	2.0907	4.0321	4.4395
n-heptane	12.8338	4.9762	2.6328	2.1496	4.1072	4.5102
correlation coefficient	0.9920	0.9826	0.9937	0.9948	0.9953	0.9948
standard deviation	6.496	9.532	5.748	5.201	5.003	5.217

[a] Observe some improvement in r (correlation coefficient) and s (the standard deviation) for selected case if compared to Table II.

First, we consider the molecular ID (identification) number,[18] introduced not only as a discriminating number that can be assigned to a structure[19] but also as number of interest in structure–activity work.[20] The molecular ID is defined as the sum of weighted path numbers, with the same weights used for the connectivity index. If one restricts the summation to paths of length 1, one would obtain the connectivity index χ. If all the weights were 1.0000, the sum would give the Wiener number W. Hence, molecular ID numbers combine and bridge two of the graph invariants already considered, and it is of interest to see what happens when these numbers are used. The correlation between molecular ID and boiling point is depicted in Figure 2 and should be compared with the correlation between χ and boiling point shown in Figure 1. A visual inspection may suggest that χ gives a better correlation than ID number, but the estimated standard deviations (6.496 and 6.746) and the correlation coefficients (0.9920 and 0.9914) are nearly identical. (From Figure 2 it is apparent that the ID number is deficient: it groups isomers, but each group is associated with too steep a slope if considered separately.) These results indicate what perhaps should have been anticipated: the examples in Table II are not necessarily optimal, and there may be considerable room for improvement. In this particular case the improvement is not necessarily a practical advantage, however, since the evaluation of paths is more involved and the simple additivity of bonds is lost. There is one somewhat arbitrary step in the construction of ID numbers that merits some scrutiny here. ID numbers are defined as the sum of all paths [with each bond m,n weighted by $(mn)^{-1/2}$] plus the number of atoms, the latter viewed formally as paths of length zero. If we do not include this direct size factor and examine the regression of boiling point against (ID − n), where n is the number of carbon atoms, we obtain a result that is significantly less useful (the standard deviation of 13.25 and correlation coefficient of 0.949 are only marginally better than those of W in Table II).

Both the connectivity index and the ID number have similar correlations, suggesting that perhaps the $(mn)^{-1/2}$ weights are the essential common ingredient. From the empirical point of view one can question the form of these weights, which in fact arose from the multiplication of the atomic factors $1/m^{1/2}$ and $1/n^{1/2}$, and ask if perhaps some other (related) forms may

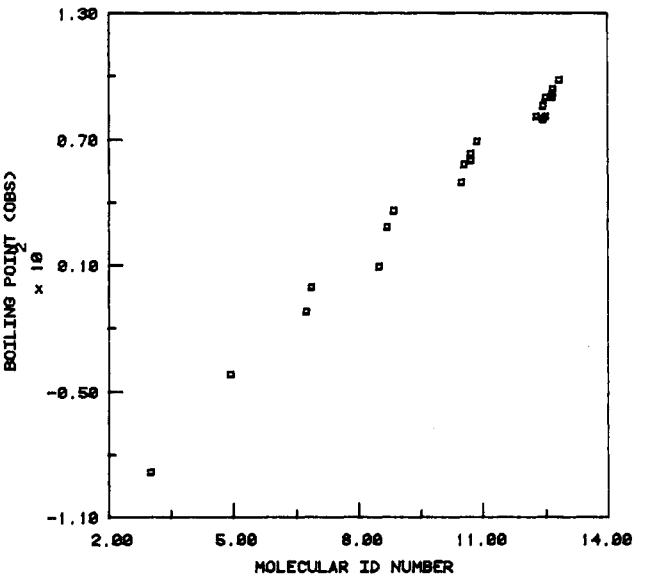

Figure 2. Correlation between the observed boiling points of the C_2–C_7 alkanes and the molecular ID numbers.

not yield better regression results. One could either utilize different m,n values from those representing molecular graph valencies or use exponents other than $-{}^1/_2$. We considered both approaches in order to examine the improvement or worsening of the regression that follows. We have not considered simultaneous changes of both factors as such effects can be deduced, at least semiquantitatively, from the observations for the cases considered. In Table III are the results obtained by replacing m and n with $m - {}^1/_2$, m (Table II), $m + {}^1/_2$, and $m + 1$, revealing the following trend for the standard deviations: 9.532, 6.746, 5.748, and 5.201 (correlation coefficients parallel this trend). The improvement is rather substantial, which speaks for the flexibility of m,n values. For each of the above cases, one must first construct the bond weights (m,n), shown in Table IV, and then derive the corresponding connectivity indices, which are shown in Table III. From Table IV we observe that bond types (1,4) and (2,2), which are degenerate for χ giving in both cases bond weights of 0.2500, now produce different weights for each of

Table IV. Bond Types and New Bond Weights for the Case in Table III

bond type	$(m - {}^1/_2)(n - {}^1/_2)$	$(m + {}^1/_2)(n + {}^1/_2)$	$(m + 1)(n + 1)$	$k = -{}^1/_3$	$k = -{}^1/_4$
(1, 1)	2.0000	0.6667	0.5000	1.0000	1.0000
(1, 2)	1.1547	0.5164	0.4082	0.7937	0.8410
(1, 3)	0.8944	0.4364	0.3536	0.6934	0.7598
(1, 4)	0.7559	0.3849	0.3162	0.6300	0.7071
(2, 2)	0.6667	0.4000	0.3333	0.6300	0.7071
(2, 3)	0.5164	0.3381	0.2887	0.5503	0.6389
(2, 4)	0.4364	0.2981	0.2582	0.5000	0.5946
(3, 3)	0.4000	0.2857	0.2500	0.4807	0.5774
(3, 4)	0.3381	0.2520	0.2236	0.4368	0.5373
(4, 4)	0.2857	0.2222	0.2000	0.3969	0.5000

the three cases considered. Moreover, the relative magnitudes of the (1,4) and (2,2) bond weights differ for the cases of decreasing m, for which (1,4) is greater than (2,2), and increasing m, for which (2,2) is greater than (1,4). Because the latter two cases give better agreement, we have to conclude that the accidental degeneracy of (1,4) and (2,2) when using χ gives the (1,4) bond type a somewhat greater role than the experimental data would justify. From the above it is clear that changing m and n involves the recalculation of weights and hence the connectivity indices for the alkanes. We are interested in the best values possible, and the gradual change of m,n values (in steps of ${}^1/_2$) is a tedious route to finding optimal values. Hence we used the above four values as four (m,s) coordinates, s being the standard deviation [(−0.5, 9.532); (0.0, 6.746); (0.5, 5.748); (1.0, 5.201)] and fitted a quadratic equation from which the minimum would suggest the optimal increment for m. The results indicated that the values of $m + 1$ (and $n + 1$) are optimal. Extrapolation to $m + {}^3/_2$ gave a value of $s = 6.112$ for the standard deviation, and even small changes around 1.0 produced increases in the standard deviation: 0.9 gave $s = 5.237$ and 1.1 gave $s = 5.349$. Hence, the optimal modification was to increase m (the valency of the graph vertex) by 1. Speculation suggests that obtaining the optimal result by using an integer may have some structural significance. However, when setting the rules for defining the weights, one is at liberty to introduce weights based on m or $m + 1$, from the a priori position both are equally arbitrary, and one is preferred to the other solely on the merits of the application. We do not wish at this stage to recommend one over the other; all that we claim here is that an additional degree of freedom exists, which, if used judiciously, may lead to improvements in structure–property correlations.

The other flexibility in considering weights for (m,n) bond types is in the selection of the exponent k in $(mn)^k$; this need not have the value $-{}^1/_2$ as in the Randić index. In fact, Altenburg[21] has already pointed out that χ can be viewed as a special case of more general quantities that describe molecular branching

$$\chi(k) = \sum_i (\epsilon_{mn})_i^k \quad k \neq 0$$

where i is the label for a bond ϵ between atoms m and n and the summation is carried over all bonds. The value $k = -{}^1/_2$ gives the connectivity index, while the value $k = 1$ gives an index used to approximate the π-electron energy of conjugated hydrocarbons.[22] Altenburg has listed cases $k = 1, +{}^1/_2, -{}^1/_2$, and -1 and found a relationship with the quadratic mean radius for small alkanes. We examined the cases $k = -{}^1/_3$ and $k = -{}^1/_4$. In Tables III and IV are listed the new connectivity numbers and the new bond weights, respectively. Here the degeneracy of bond weights (1,4) and (2,2) remains (of necessity), but we find that the standard deviation in both cases improves in comparison with the $k = -{}^1/_2$ case. In fact, the case of $k = -{}^1/_3$ with a standard deviation of 5.003, illustrated in Figure 3, represents the best result of all the alternatives of Tables II and III. Combining the variations in m,n (i.e., to $m + 1, n + 1$) and simultaneously changing k to $-{}^1/_3$, one could expect a further reduction in the standard deviation, but it appears that it is hard to reduce the standard deviation below 5.000 with single parameter descriptors. Or is it possible to considerably reduce the standard deviation, without increasing the number of parameters?

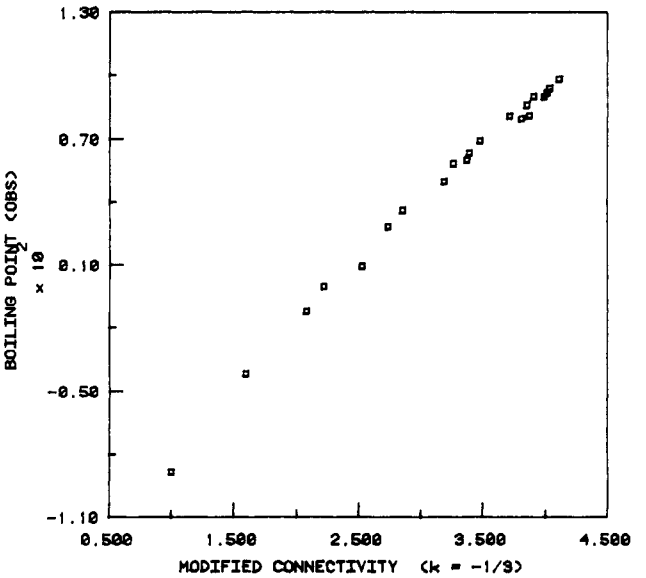

Figure 3. Correlation between the observed boiling points of the C_2–C_7 alkanes and the modified connectivity index (based on $k = -{}^1/_3$, i.e., bond weights $(mn)^{-1/3}$).

EMPIRICAL APPROACH

As stated above, the connectivity index is based on bond additivity with bond weights $(mn)^{-1/2}$, m and n being the valencies of the bond's terminal atoms. The approaches considered so far consisted of altering these weights either by changing the exponent or by adding an integer or simple fraction to the valencies. We also undertook an empirical approach in which the bond weights were assumed to equal $a_m a_n$; that is, for each of the four possible atom types (valencies) a constant is assumed to exist, but no a priori constraints are placed upon their values. Nonlinear regression was then applied to the set of 21 equations

$$\mathrm{bp}_{\mathrm{obsd}} = \sum a_m a_n N_{mn} + \text{constant}$$

to obtain optimal values for the a_m. Table V gives the decomposition of bond types N_{mn} for the 21 alkanes included in this study as well as the results of the regression analysis. The estimated standard deviation of 5.134 is comparable to those obtained in the previous section; the improved fit is of course offset by the reduction in the number of degrees of freedom. We conclude that, within the assumed restrictions on the form of the descriptors and regression, we cannot expect to achieve an estimated standard deviation smaller than about 5.

To compare the results of this analysis with those obtained in the previous section, the linear correlations between the a_m shown in Table V and the corresponding values of $m^{-1/2}$, $(m$

Table V. Decomposition of the C_2–C_7 Alkanes into Their (m,n) Bond Types and the Best Fit of Their Boiling Points Obtained by Using Empirically Derived Atom Factors a_m

compound	bp, °C			bond-type count (N_{mn})								
	obsd	calcd	residual	1, 1	1, 2	1, 3	1, 4	2, 2	2, 3	2, 4	3, 3	3, 4
ethane	−88.63	−78.02	−10.61	1	0	0	0	0	0	0	0	0
propane	−42.07	−42.10	0.03	0	2	0	0	0	0	0	0	0
2-methylpropane	−11.73	−12.83	1.10	0	0	3	0	0	0	0	0	0
n-butane	−0.50	−6.14	5.64	0	2	0	0	1	0	0	0	0
2,2-dimethylpropane	9.50	11.78	−2.28	0	0	0	4	0	0	0	0	0
2-methylbutane	27.85	23.21	4.64	0	1	2	0	0	1	0	0	0
n-pentane	36.07	29.82	6.25	0	2	0	0	2	0	0	0	0
2,2-dimethylbutane	49.74	47.90	1.84	0	1	0	3	0	0	1	0	0
2,3-dimethylbutane	57.99	52.74	5.25	0	0	4	0	0	0	0	1	0
2-methylpentane	60.27	59.17	1.10	0	1	2	0	1	1	0	0	0
3-methylpentane	63.28	59.26	4.02	0	2	1	0	0	2	0	0	0
n-hexane	68.74	65.78	2.96	0	2	0	0	3	0	0	0	0
2,2-dimethylpentane	79.20	83.86	−4.66	0	1	0	3	1	0	1	0	0
2,4-dimethylpentane	80.50	88.52	−8.02	0	0	4	0	0	2	0	0	0
2,2,3-trimethylbutane	80.88	77.60	3.28	0	0	2	3	0	0	0	0	1
3,3-dimethylpentane	86.03	84.02	2.01	0	2	0	2	0	0	2	0	0
2,3-dimethylpentane	89.78	88.79	0.99	0	1	3	0	0	1	0	1	0
2-methylhexane	90.05	95.13	−5.08	0	1	2	0	2	1	0	0	0
3-methylhexane	91.85	95.22	−3.37	0	2	1	0	1	2	0	0	0
3-ethylpentane	93.48	95.30	−1.82	0	3	0	0	0	3	0	0	0
n-heptane	98.42	101.74	−3.32	0	2	0	0	4	0	0	0	0
correlation coefficient		0.9958										
estimated standard deviation		5.134										
	a_1 = 6.192	a_2 = 5.997	a_3 = 5.573	a_4 = 5.174	constant = −116.359							

$+ 1)^{-1/2}$ and $m^{-1/3}$ were computed, yielding linear correlation coefficients of 0.9049, 0.9412, and 0.9177, respectively. (It should be noted that if ethane was excluded from the set, in view of its unique status, these same correlations yielded the values 0.9991, 0.9906, and 0.9974, respectively, and an estimated standard deviation of 3.624, the latter demonstrating the major contribution of ethane to the total error.) Before, however, trying to seek possible structural factors that may be responsible for larger deviations [such as has been possible in the case of the correlation of the chromatographic retention indices[23] with the connectivity, when the presence of large numbers of paths of length 3 between terminal (primary) carbon atoms could account for the discrepancies] we will seek additional single-variable descriptors but relax the constraint that the regression be linear. Figure 1 points to limitations of the linearity assumption, as one can see that all three initial descriptors, χ, W, and Z, give useful correlations but with increasing curvature.

SEARCH FOR FUNCTIONAL FORM

There are essentially two alternative routes in the search for a nonlinear functional form for a correlation: (1) one can apply various simple transformations, or (2) one can use polynomial expansions. We will examine both approaches, as they supplement one another and reinforce our conclusions. We start with selecting various simple nonlinear forms for the same descriptors as already considered. Table VI summarizes these results. For both the connectivity index χ and the Wiener number W, we examined the square, cube, fourth, and fifth roots using simple linear regression, and substantially improved correlations were obtained. As measured by the standard deviation of 2.825, the $\chi^{1/3}$ model was the best single-descriptor model developed in this study, and the $\chi^{1/4}$ model gave nearly indistinguishable results. The high standard deviation of 17.09 (Table II) yielded by W itself dropped successively to 8.12 for $W^{1/2}$, to 5.08 for $W^{1/3}$, and finally to 4.42 for $W^{1/4}$ before climbing to 4.60 for $W^{1/5}$. Hence for both χ and W the standard deviation dropped significantly below 5.000, the previous limit approached by both the empirical fit and the modified connectivity based on $k = -1/3$. Observe that up to this point the best correlations have been associated with the cube and quartic roots. This may or may not be significant. In the case of W, one should recall the work of Platt,[6] who interpreted W as a measure of volume, suggesting that the cube root of W can be considered a measure of linear dimension. By linear, of course, we do not mean to imply simple molecular length but what might correspond to (in a very loose sense) an average length of a three-dimensional molecule.

Table VI. Summary of the Results: the Correlation Coefficients (r) and Standard Deviations (s) for Various Single-Parameter Linear Regressions Based on Variations of the Connectivity Index, ID Number, Wiener W Number, and Hosoya Z Topological Index

descriptor	correlation coefficient	standard deviation
$\chi^{1/2}$	0.9979	3.364
$\chi^{1/3}$	0.9985	2.825
$\chi^{1/4}$	0.9985	2.835
$\chi^{1/5}$	0.9983	2.939
$W^{1/2}$	0.9875	8.119
$W^{1/3}$	0.9951	5.082
$W^{1/4}$	0.9963	4.417
$W^{1/5}$	0.9960	4.602
Z^2	0.8092	30.22
$Z^{1/2}$	0.9591	14.56
$Z^{1/3}$	0.9717	12.15
Z^{-1}	0.9758	11.23
$Z^{-1/2}$	0.9967	4.192
$Z^{-1/3}$	0.9976	3.547
$Z^{-1/4}$	0.9969	4.040
ID^2	0.9677	12.96
$ID^{1/2}$	0.9939	5.658
$ID^{1/3}$	0.9922	6.406
$(Z - 1)^{-1/2}$	0.9810	9.955
$(Z + 1)^{-1/2}$	0.9971	3.930
$(Z + 2)^{-1/2}$	0.9942	5.485
$(Z + 3)^{-1/2}$	0.9907	6.986

In a similar manner, we also examined Hosoya's Z number. Numerous works of Hosoya and collaborators[24] demonstrate the usefulness of this index in the considerations of properties, suggesting that the index is probably much better than evidenced by the standard deviation shown in Table II. We considered Z^2, $Z^{1/2}$, $Z^{1/3}$, Z^{-1}, $Z^{-1/2}$, $Z^{-1/3}$, and $Z^{-1/4}$. The results are collected in Table VI. Clearly, those forms with positive exponents are not nearly as successful as those computed by using negative roots. The standard deviation of 3.55

yielded by $Z^{-1/3}$ is the lowest value, representing a drop of over 80% from the original value of 21.0, and an even better result than obtained by using $W^{1/4}$. Figure 4 testifies to the good linearity of the $Z^{-1/3}$ correlation.

With the molecular ID number a similar approach reduced the standard deviation from the original 6.496 to 5.658 by using $ID^{1/2}$. Although a significant decrease, the end result did not approach those obtained for the other indices when this strategy was applied. This is perhaps not surprising, however, considering the nature of the correlation of boiling point with ID number as revealed by Figure 2.

In summary, we found that all the indices considered here, the connectivity χ, ID, modified connectivity with $k = -1/3$, $W^{1/3}$, $W^{1/4}$, and finally $Z^{-1/3}$ give very good standard deviations for the sample examined. We have not exhausted all simple forms, and from what we have seen further possibilities emerge as worthy of testing. The end of Table VI shows one such possibility based on the Hosoya Z index: by increasing (or decreasing) Z by integers, we can test the form $1/(Z + q)^{1/2}$, and we find that the standard deviation drops to 3.930 when $q = 1$, which reminds us of the case of the variation in m, where again $m + 1$ gave the best result. We conclude, therefore, that all the above-mentioned indices deserve full attention in structure–property and structure–activity studies and may be viewed as extensions of our coordinate systems to additional forms.

CONCLUDING THE SEARCH

The number of functional forms is unlimited, and one may argue that sooner or later one is going to hit the best scheme and for this reason one may tend to devaluate our results. In defending our approach, we wish to emphasize that we restricted our attention to simple transformations, involving integer and fractional powers, and translations (i.e., linear increments) that were limited to integers or simple fractions. The search for additional forms is legitimate, but in our view the first matter of business, now that we have opened the pool of invariants, would be to see how they work on larger samples and different properties. In order to strengthen our findings, we examined the graph theoretical indices and their best transformations by the alternative route of using a power expansion. If the functional form is optimal, then the power series will converge quickly, and the higher polynomial terms will result in little improvement. Conversely, if the functional form is poor, the addition of higher power terms will make significant improvements in the correlation coefficient and the standard deviation. Along these lines, we tested χ, ID, W, and Z, the latter two in particular displaying appreciable departure from linearity (see Figures 1 and 2). Table VII summarizes our results. For χ the addition of a single term (χ^2) reduced the standard deviation by more than half to 2.93 (one of the lowest values obtained in this study), but further expansion resulted in no better fit. For both W and Z the addition of quadratic and cubic terms resulted in substantial improvements in the standard deviations to 6.55 and 6.64, respectively, from their original values of 17.1 and 21.0 (Table II). Attempts to expand these polynomials further resulted in computational problems arising from matrix near-singularity caused by the high intercorrelation of the polynomial terms. In contrast, expansion of the molecular ID number expression resulted in a reduction of the standard deviation of less than 15%. This is perhaps not surprising in view of the very slight curvature observed in the boiling point versus ID number plot of Figure 2. Hence the molecular ID number appears to possess an optimal form without expansion or transformation.

To further test this method, we also considered the polynomial expansions of the optimal forms of χ, W, and Z (as shown in Table VI), that is, $\chi^{1/3}$, $W^{1/4}$, and $Z^{-1/3}$. In each of

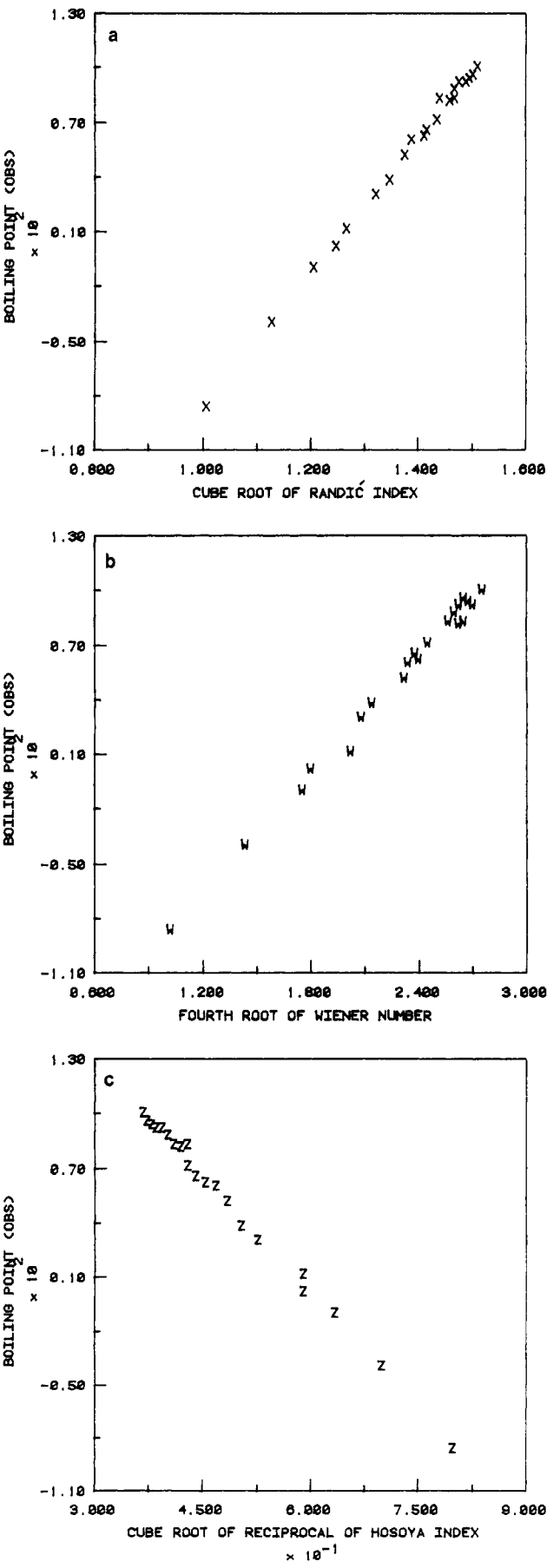

Figure 4. Correlation between the observed boiling points of the C_2–C_7 alkanes and the optimum forms of the topological indices: (a) $\chi^{1/3}$ (χ being the Randić index, also called the molecular connectivity); (b) $W^{1/4}$ (W being the Wiener number); and (c) $Z^{-1/3}$ (Z being Hosoya's index).

Table VII. Correlation Coefficients and Standard Deviations for Polynomial Regressions for Selected Graph Invariants versus Boiling Points of Small Alkanes[a]

leading term	correlation coefficient	standard deviation
	Randić Index (χ)	
χ	0.9914	6.746
χ^2	0.9985	2.929
χ^3	0.9985	2.962
$\chi^{1/3}$	0.9985	2.825
$\chi^{2/3}$	0.9985	2.879
	Wiener Number (W)	
W	0.9432	17.09
W^2	0.9866	8.615
W^3	0.9927	6.548
$W^{1/4}$	0.9963	4.417
$W^{2/4}$	0.9963	4.533
	Hosoya Index (Z)	
Z	0.9128	21.01
Z^2	0.9816	10.10
Z^3	0.9925	6.644
$Z^{-1/3}$	0.9976	3.547
$Z^{-2/3}$	0.9977	3.591
	Molecular ID Number	
ID	0.9918	6.565
ID^2	0.994	5.644
ID^3	0.994	5.760
	Modified Connectivity (χ') ($k = -1/3$)	
χ'	0.9953	5.003
χ'^2	0.9987	2.721
χ'^3	0.9987	2.762

[a] The "leading term" entries in the table refer to the highest power used in each polynomial regression. For example, Z^3 refers to the fitting of a cubic equation.

the three expansions the addition of a quadratic term resulted in no improvement in correlation, evidenced by slight increases in the standard deviations. In view of the linearity observed in Figure 4, one might have expected these results. Finally, we tested the modified connectivity index based on $k = -1/3$. The addition of a quadratic term to this model yielded the best correlation obtained in this study (see Figure 5). The initial standard deviation of 5.003 (obtained from simple linear regression) was further reduced to 2.721.

CONCLUDING REMARKS

Structural invariants are essential for structure–property or structure–activity studies, and they are of potential interest in many empirical studies, such as pattern recognition and clustering. It is perhaps not too difficult to introduce additional structural invariants, but what may be more difficult than anticipated is to arrive at new invariants that have novel different structural bases and cannot be simply (if not trivially) related to those already existing. In this paper we reconsidered several important structural invariants, in particular χ, W, Z, and ID numbers, and examined some derivatives of these by relaxing the constraints on their functional form. Visible improvements have been reported for all the invariants considered, suggesting further testing for $\chi^{1/3}$, $W^{1/4}$, and $Z^{-1/3}$, and also for modified connectivity indices using $m + 1$ and $n + 1$ instead of m,n in $(mn)^{-1/2}$ as well as $(mn)^{-1/3}$. The last case produced, in a quadratic polynomial, the overall best fitting, while $\chi^{1/3}$ gave the best single-term, linear regression.

The emphasis in this paper is on the strategies to be used in the searches for new invariants, not on proposing alternatives to the existing indices. Such replacements may follow should the present results survive more critical tests of time and wider applications beyond the sample considered here. However, we feel that even when one is limited to small subsets of structures, modifications of the parameters and invariants are legitimate, in particular since they can reduce the number of

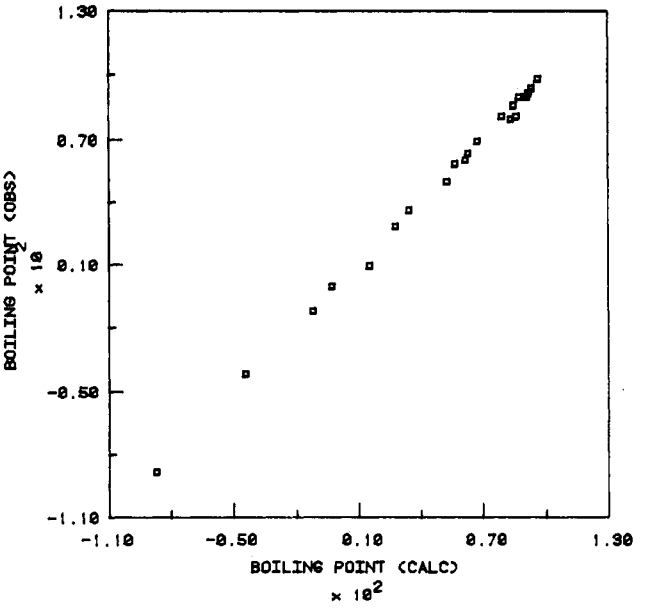

Figure 5. Best correlation of this study based on a quadratic relationship of the boiling points of the C_2–C_7 alkanes and the modified molecular connectivity with $k = -1/3$ (e.g., bp = $a\chi'^2 + b\chi' + c$).

indices or expansion terms used. The connectivity index, which has seen such broad application, has been frequently combined with its reciprocal, $1/\chi$, and higher connectivity indices, ${}^m\chi$, and the present study suggests an additional approach for testing published correlations and regressions to see if they may also suggest other fewer term representations and reveal important modifications that are currently hidden in the restrictive use of these topological indices.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation under Grant CHE-8503542. The PRIME 750 computer was purchased with partial financial support of the National Science Foundation. We thank E. P. Jaeger for his contribution to the improvement of the presentation of the material.

REFERENCES

(1) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
(2) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.
(3) Hosoya, H. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332.
(4) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic: New York, 1976.
(5) (a) Trinajstić, N. *Chemical Graph Theory*, Vol. I, II; CRC: Boca Raton, FL, 1983. (b) Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA, 1969.
(6) Platt, J. R. *J. Phys. Chem.* **1952**, *56*, 328.
(7) Heilman, O. J.; Lieb, E. H. *Commun. Math. Phys.* **1972**, *25*, 190.
(8) Fowler, R. H.; Rushbrooke, G. S. *Trans. Faraday Soc.* **1937**, *33*, 1272.
(9) Heilbronner, E.; Schmelzer, A. *Nouv. J. Chim.* **1980**, *4*, 23.
(10) (a) Motoc, I.; Balaban, A. T. *Rev. Roum. Chim.* **1981**, *26*, 593. (b) Razinger, M.; Chretian, J. R.; Dubois, J. E. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 23.
(11) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Methods*, 2nd ed.; Richard D. Irwin: Homewood, IL, 1985; p 393.
(12) Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399.
(13) Balaban, A. T. *Theor. Chim. Acta* **1979**, *53*, 355.
(14) (a) Lovasz, L.; Pelikan, A. *J. Period. Math. Hung.* **1973**, *3*, 175. (b) Cvetkovic, D.; Gutman, I. *Croat. Chem. Acta* **1977**, *49*, 115.
(15) Jurs, P. C. *Computer Software Applications in Chemistry*; Wiley-Interscience: New York, 1976; Chapter 11.
(16) Shannon, C.; Weaver, W. *Mathematical Theory of Communication*; University of Illinois: Urbana, IL, 1949.
(17) Bonchev, D.; Trianjstić, N. *J. Chem. Phys.* **1977**, *67*, 4517.
(18) Randić, M. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164.
(19) (a) Randić, M. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 134. (b) Szymanski, K.; Mueller, W. R.; Knop, J. V.; Trinajstić, N. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 413.
(20) Randić, M. *Int. J. Quantum Chem., Quantum Biol. Symp.* **1984**, *11*, 137.
(21) Altenburg, K. *Z. Phys. Chem. (Leipzig)* **1980**, *261*, 389.

(22) (a) Gutman, I.; Trinajstić, N. *Chem. Phys. Lett.* **1972**, *17*, 535. (b) Gutman, I.; Ruščić, B.; Trinajstić, N.; Wilcox, C. F., Jr. *J. Chem. Phys.* **1975**, *62*, 3399.
(23) Randić, M. *J. Chromatogr.* **1978**, *161*, 1.
(24) Hosoya, H.; Kawasaki, K.; Mizutani, K. *Bull. Chem. Soc. Jpn.* **1972**, *45*, 3415.
(25) Platt, J. R. *J. Chem. Phys.* **1947**, *15*, 419.
(26) Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. *J. Pharm. Sci.* **1976**, *65*, 1226.
(27) Kier, L. B. *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1986**, *5*, 7.
(28) Kier, L. B. *Acta Pharm. Jugosl.* **1986**, *36*, 171.

End-User Searching of CAS ONLINE. Results of a Cooperative Experiment between Imperial Chemical Industries and Chemical Abstracts Service

WENDY A. WARR*,† and ANGELA R. HAYGARTH JACKSON‡

Information Services Section, Imperial Chemical Industries PLC, Pharmaceuticals Division, Alderley Park, Macclesfield, Cheshire, SK10 4TG, United Kingdom

Received September 24, 1987

Chemical Abstracts Service staff trained 88 Imperial Chemical Industries chemists to use CAS ONLINE over a 2-year period in a cooperative experiment between the two organizations. The effectiveness of the training, the problems encountered by end-users, the usage made of CAS ONLINE, the impact of end-user searching on information scientists, and the attitudes of management were studied.

BACKGROUND

Late in 1982, after 10-years' experience of online interactive searching by ICI information scientists and librarians, one of us gave a paper that stated categorically that the end-user will search online.[1] This opinion was just beginning to obtain credence in the literature.[2-5] However, the ICI view was mainly based on developments within the company in which a policy was being introduced to allow scientists, and others, to access internal company databases themselves, interactively, without the intervention of an intermediary who would interrupt the flow of scientific thought.[6] It was recognized that given suitable terminal and network facilities, most scientists would also wish for direct access to the external databases, representing the published literature.

However, at that time, the online search services were even less "user-friendly" and the telecommunication links less reliable than today. It was recognized that mastery of the intricacies of successful searching took time, training, and experience to acquire. There was top management concern relating to the costs involved in both the end-user scientist's time spent searching and the actual online charges. In addition, management was concerned about the mechanisms for managing this new online information resource, in that a free-for-all approach, in effect an open check, was not an acceptable policy. The information scientists were worried that their knowledge and skill in accessing online services would no longer be required.

Other companies have considered the same issues.[7,8]

INITIATION OF THE ICI CAS ONLINE COOPERATIVE EXPERIMENT

Informal discussions were started with Chemical Abstracts Service (CAS) in 1984. ICI felt that chemists were the most suitable end-users to be trained first because they already had significant experience of using an in-house, interactive system,[6] and their needs for external information could be substantially met by access to the online databases provided by CAS. ICI information scientists were accessing these databases under DARC[9] and CAS ONLINE[10] in the ratio of about 1 to 2 in cost terms, but the company wished to train chemists in one system only. The likely development of both search services was an important issue, but the offer by CAS to train groups of ICI chemists to use CAS ONLINE helped to clinch the matter.

JOINT CAS AND ICI OBJECTIVES FOR THE COOPERATIVE EXPERIMENT

The aim of the cooperative experiment between CAS and ICI was to gather information on end-user chemists' needs and usage of CAS ONLINE. The results would, it was hoped, assist CAS in their developments and marketing of CAS end-user chemists' to industry and would assist ICI in the provision and management within the company of CAS ONLINE as a resource to chemical innovation and chemistry in general. The joint specific objectives of CAS and ICI were as follows:

(1) To determine how effective CAS training was for both end-users and the information scientists who supported them and to assess the appropriateness of related documentation and user manuals.

(2) To find what problems end-users encounter.

(3) To determine the number and type of searches done by an average end-user.

(4) To study the effectiveness of end-user searching.

(5) To determine the types of searching problems that cause an end-user to seek help from an information scientist.

(6) To study the reaction of information scientists to end-user searching.

(7) To study ICI management reaction, including the opinions of both chemistry and information managers on the cost effectiveness of end-user searching and the developments needed from CAS if further progress were to be made.

(8) To learn how CAS might better serve the needs of end-user scientists.

MAIN TERMS OF THE COOPERATIVE EXPERIMENT

The experimental period was 25 months from December 1984 to December 1986. The main terms were as follows:

(1) ICI was to make two advance lump-sum payments to the Royal Society of Chemistry (U.K. marketing agent for CAS ONLINE) against the use of CAS ONLINE by ICI in the U.K.

† Manager.
‡ Former Manager.

0095-2338/88/1628-0068$01.50/0