CROATICA CHEMICA ACTA CCACAA **66** (2) 289–312 (1993)

ISSN 0011-1643
*CCA–2139* *Original Scientific Paper*

# Comparative Regression Analysis. Regressions Based on a Single Descriptor*

*Milan Randić*

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

Received March 3, 1992

In this article we report the results on over 750 linear regressions based on a single descriptor, using some twenty molecular properties of alkanes and some forty distinct molecular descriptors. It is shown that, among the numerous descriptors, less than a dozen descriptors outperform all the others. Hence, while the search for graph invariants has been fruitful, few of the constructed descriptors show sufficiently novel characteristics and have displayed sufficiently different behavior in correlations with physico-chemical properties. At the same time, we find that some properties are more susceptible to a successful single-descriptor regression analysis, while few properties remain difficult to characterize by a single descriptor. At the end of this report, we have listed a few challenges to be considered by those involved in structure-property-activity studies. We have drawn attention to some properties of alkanes (octanes in particular) for which better regression results are warranted. We also recommend that the performance of novel indices be compared with the known results, like those reported here. If a novel index shows a better behavior with respect to any of the properties already reported in the literature, they certainly deserve publicity. If an index shows a performance comparable to some existing descriptors, they ought to have other advantages in order to replace the existing descriptors. The burden of »proof« ought to be on the »inventor« of a novel topological index. Vague statements that an index may show promise in structure-property studies ought to be replaced by a comparative study, such as shown here for octanes. While the present paper answers the questions which are the best single simple descriptors for correlations of octane physico-chemical properties, our restriction to the use of a single descriptor in structure-property correlations neither signifies that we expect all properties to be well represented by simple regressions, nor that the best single descriptors will necessarily remain the dominant descriptor in multiple regression analysis.

* This paper is dedicated to Professor L. B. Kier for his untiring explorations of the use of the connectivity indices in structure-property-activity studies.

## INTRODUCTION

Recently, we initiated a systematic comparative study of structure-property regression analysis.[1] We reported use of the connectivity indices as a *basis* in structure-property analysis, rather than an *ad hoc* combination of indices. The distinction is important. In the first case, same indices are always used in all regression while in the second case, optimal indices (descriptors) are selected. In the first case, the descriptors play a role analogous to those of basis functions (vectors) and comparison between different regressions is simpler, particularly when the descriptors, prior to their direct use, have been made first orthogonal.[2] In an effort to continue developing a useful alternative basis for the connectivity index and higher connectivities for structure-property studies, using multiple regression analysis, we will here examine regressions based on a single descriptor. Clearly, if the first descriptor to be used can account for the major part of a correlation, such an index may be of interest in construction of alternative bases.

By restricting attention to simple single-variable regressions, we have neither assumed that models based on a single topological index are necessarily able to describe a different (if any) molecular property to a desired precision (comparable to the accuracy determined by the experimental errors), nor that such »the best« single-variable descriptor will necessarily remain the dominant descriptor in multiple regression analysis. Our position is pragmatical: Search for the best single descriptor is a mathematically well defined process which will facilitate a comparative study of structure-property relationships. In a way, such regressions will often point to the first steps in finding more comprehensive multiple-variable regressions. The pool of mathematical descriptors for representing a structure is inexhaustible, but one should impose restrictions on graph invariants and prune such pools of mathematical descriptors to chemically *useful* descriptors. These can be qualified as descriptors that show a superior correlation with at least one molecular property. Such descriptors may appear promising for the study of other molecular properties or, alternatively, may contribute to refinement of a molecular model used in structure-property studies. Hence our interest in a systematic search for such promising molecular descriptors.

In order to proceed, we have to select structures to be considered, properties to be considered and descriptors to be considered. We decided to select the 18 octane isomers as the structures to be considered. Octane isomers show sufficient structural variations (the degree of variations in bond types, in the number of primary, secondary, and tertiary carbon atoms, the lengths of the longest chains *etc.*) and, at the same time, for these molecules a large number of properties are known (with a lesser or greater accuracy). By restricting attention to isomers of octane, we have deliberately eliminated the dominant role of the molecular size, which tends to obscure minor variations of properties with shape (the pattern of branching). Equally, we have minimized uncertainty associated with the parametrization of heteroatoms, such as oxygen in alcohols, or nitrogen in amines. The role of heteroatom ought to be carefully studied, but this is outside the scope to the present work. Recently, a general approach to finding the empirical parameters characterizing heteroatom was outlined.[3] We hope that studies similar to the one reported here for octanes will be followed on families of other structurally related compounds, including also compounds having heteroatoms. While it is true that the results found here, valid for alkanes, need not extend to molecules having heteroatoms, we feel also that it is unlikely that more general descriptors which fail to show correlations with the properties of alkanes (when reduced for such compounds) will be found useful in more general situations.

## SELECTION OF PROPERTIES

Next we have to select properties to be examined. In Table I, we collected available data for some twenty properties of octanes used in this study. We indicate the source (not necessarily the original papers describing the data, but more recent publications). Interested reader should consult earlier work in order to assess the quality of experimental data. Our aim here is not to propose a particular regression equation as *optimal*. Our aim is to compare molecular descriptors and find which among many are of continuing interest and which show lesser promise, which of the descriptors appear as dominant descriptors for a single property. Hence, for our purpose, detailed information on the accuracy of the experimental data is not so essential. In one case (heats of formation), we even adopted two sources and duplicated the work to illustrate the robust character of the regressions, which are not so sensitive to minor variations in the data.

## SELECTION OF DESCRIPTORS

In Table II, we list over 40 molecular descriptors that we examined. In the Appendix, we list the values of the descriptors for octane isomers, which have not yet been reported in the literature. The numerical values of other descriptors can be found in the literature (as indicated in Table II). Equally, one can find in the literature detailed definitions for most of the known descriptors. Here, we will only briefly sum-

TABLE I

*List of properties of the octanes investigated. Data has been taken as reported in the references shown. Alternative sources could be used to assess the sensitivity of the regressions to smaller variation in experimental values, as reported by different sources.*

| Symbol | Property | Source | Alternative |
|---|---|---|---|
| BP | Boiling points | 4 | |
| $S$ | Entropy | 5 | |
| $\Delta H$ | Heat of vaporization | 6 | |
| HF | Heat of formation | 6 | 5 |
| HA | Heat of atomization | 6 | |
| $\Delta_f H$ | Heat of formation | 7 | |
| $\Delta_l^g H$ | Heat of vaporization (liquid) | 7 | |
| $\Delta_v H$ | Heat of vaporization (vapor) | 7 | HV |
| HV | Heats of vaporization | 4 | $\Delta H_v$ |
| $T_c$ | Critical temperature | 4 | 8 |
| $p_c$ | Critical pressure | 4 | |
| $V_c$ | Critical volume | 8 | |
| $\sigma$ | Surface tension | 4 | |
| $R$ | Quadratic mean radius | 9 | |
| AC | Pitzer acentric factor | 10 | |
| $N$ | Octane number | 11 | |
| $V_m$ | Molecular Volume | 4 | |
| $R_m$ | Molar refraction | 4 | 6 |
| $\rho$ | Density | 6 | |
| CS | Carbon-13 chemical shift sum | 12 | |
| $C_p$ or $C_v$ | Heat capacity | 5 | |
| RT | Retention time | 13 | |

marize the description of the descriptors and group them according to their origin into one of the several types:

*Descriptors derived or related to the adjacency matrix:*

The connectivity index $\chi$[14] is bond additive and uses the weighting scheme $(m\ n)^{-1/2}$, where $m$, $n$ are the valencies of the vertices forming the bond. The weighting can be considered as a special case of $(m\ n)^p$, where the exponent $p = -1/2$. If $p = 1$, we have the so called Zagreb group index,[24,25] which we will denote as $\chi$.[1] The cases $p = -1$ and $p = 1/2$ have been considered by Altenburg,[9] while more recently cases $p = -1/3$ and

TABLE II

*List of descriptors used in the comparative study or single variable linear regressions*

| Symbol | Descriptor | Listed in reference: |
|---|---|---|
| $^1\chi$ | Connectivity index | 14 19 |
| $^2\chi$,$^3\chi$ | Higher connectivity indices | 6 |
| $W$ | Weiner numbers | 15 19 |
| $Z$ | Hosoya index | 15 |
| ID(P) | Path ID numbers | 16 |
| X(**A**) | The first eigenvalue | 17 |
| EC | Eccentricity | 18 |
| PO | Ponderal index | 18 |
| $J$ | Balaban's $J$ index | 19 |
| MTI | Topological index of Schulz | 19 |
| TI | Topological index of Schulzes | 19 |
| WW | Expanded Wiener number | 10 |
| P′/P | Path bond order | 20 |
| $^1\chi'/^1\chi$ | Connectivity bond order ratio | 20 |
| $U$ | Balaban's $U$ index | 8 |
| $V$ | Balaban's $V$ index | 8 |
| $X$ | Balaban's $X$ index | 8 |
| $Y$ | Balaban's $Y$ index | 8 |
| AZH | Balaban's AZV index | 8 |
| $\chi^{[1]}$ | Altenburg's $p=1$ index | 9 |
| $\chi^{[1/2]}$ | Altenburg's $p=1/2$ index | 9 |
| $\chi^{[-1]}$ | Altenburg's $p=-1$ index | 9 |
| IED | Information edge/distance | 17 |
| IWD | Information (Wiener/distance) | 17 |
| $D3$ | Graph disection index | 21 |
| HD | Hybrid matrix determinant | 22 |
| $W'$ | Wiener bond order | Appendix |
| $W'/W$ | Wiener bond order ratio | Appendix |
| $Z'/Z$ | Hosoya bond order ratio | Appendix |
| ID($^1\chi$) | Connectivity ID numbers | Appendix |
| $^1\chi$(***D***) | Distance matrix connectivity | Appendix |
| ID(***D***) | Distance matrix ID numbers | Appendix |
| X(***D***) | Distance matrix first eigenvalue | Appendix |
| $^1\chi$(***W***) | Wiener matrix connectivity | Appendix |
| ID(***V***) | Wiener matrix ID numbers | Appendix |
| X(***V***) | Wiener matrix first eigenvalue | Appendix |
| WW | Wiener row sums | Appendix |

even $p = -1/4$ (and some other functional forms applied also to a few selected indices not based on the connectivity) have received attention.[26] Adjacency matrix will produce weighted paths (when ALL PATH program is used[27-29]) and, as a sum of all weighted paths, we obtain the molecular ID number.[16] Finally, the first eigenvalue of the adjacency matrix represents an index that, according to Lovasz and Pelikan,[30] reflects well the degree of the skeletal branching in acyclic structures. The coefficients of the characteristic polynomials led Bonchev and Trinajstić[17] to compose an information theoretic index using the associated partitioning in the Shannon equation for evaluating the information content of a partition.[31]

### *Descriptors derived or related to the distance matrix*

Balaban extended the notion of the connectivity index to distance matrix and used the weighting $(d_i\, d_j)^{-1/2}$, where $d_i$ and $d_j$ are the row sums of the entries in rows $i$ and $j$ of the distance matrix, respectively. This resulted in an index designated as the $J$ index, which has shown a considerable discrimination between isomers.[19,32] The smallest trees with the same $J$ index have $n = 12$ vertices. In comparison, we already have among octanes isomers having the same connectivity index. The size of the smallest graphs for which »duplication« occurs is a measure of the deficiency of the »basis« of the descriptors.[33] However, one can use the distance matrix and apply the WEIGHTED PATH program[34] which will evaluate weighted paths that form a sequence of »distance connectivities« indices and »distance ID numbers«. Finally, the Wiener index $W$ can be viewed as closely related to the distance matrix, since it can be evaluated numerically by adding all the entries in the distance matrix above the main diagonal.[35]

### *Descriptors derived or related to the Wiener matrix*

Recently, a novel matrix associated with trees has been suggested. The construction of some of its elements may be viewed as a generalization of the Wiener procedure for construction of the Wiener number.[36] Hence, it was named the Wiener matrix.[37] The $(i,j)$ entry in the Wiener matrix enumerates all the paths in a tree in which the path $(i,j)$ between vertex $i$ and $j$ occurs, *i.e.* all paths of which the path between $(i,j)$ is a subgraph. Once a matrix is constructed, it generates other graphs invariants, such as the weighted paths, the ID numbers, the first eigenvalue, *etc.*, all of which can be used as novel topological indices.

### *Combinations of indices and matrices*

In addition to the indices directly related to a graph matrix, one can combine various indices in simple arithmetical and algebraic combinations, like reciprocals, ratios, or differences. In this way, for example, we obtained the descriptors: $1/{}^1\chi$, $1/{}^2\chi$, $1/{}^3\chi$, $\mathrm{ID}/{}^1\chi$, ${}^1\chi - {}^2\chi$, ${}^2\chi - {}^3\chi$. There are additional simple combinations that one might consider. Already Kier and Hall reported some correlations using the reciprocal connectivity index $1/{}^1\chi$,[6] and the difference in the connectivity indices ${}^1\chi - {}^1\chi^v$.[38] The differences ${}^1\chi - {}^2\chi$ and ${}^2\chi - {}^3\chi$ remind one of the differences in path numbers $p_1 - p_2$, $p_2 - p_3$, which have been found useful in ordering isomers and recognizing regularities in their properties.[39]

Another way of arriving at novel invariants is to modify or combine graph matrices. Thus, Schultz[19] combined $\boldsymbol{A}$ and $\boldsymbol{D}$ matrix to arrive at the topological index

MTI, which we labeled $\boldsymbol{A} + \boldsymbol{D}$, Mihalić and coworkers[19] combined $\boldsymbol{A} + 2\boldsymbol{D}$ and Tratch and coworkers[10] combined $\boldsymbol{A}$ and $\boldsymbol{E}$, an extended distance matrix that they introduced.

Finally, once can consider nonalgebraic operations on graphs and generate novel graphical invariants. In particular, we consider indices constructed by considering for each edge the residual graph G–e, for which selected invariants are sought. In our collection of the descriptors analyzed, such indices are indicated as $d'/d$, where descriptor $d$ can represent the connectivity index ${}^1\chi$, Wiener number $W$, ID numbers, Hosoya $Z$ index, Balaban $J$ index, the first eigenvalues, and even the reciprocal $1/{}^1\chi$. Here, $d'$ represents the sum of the values of the invariant considered for graph G–e, which is obtained from graph G by erasing one edge e at a time.

Another general approach to introducing an additional invariant is to apply the Shannon information theoretic formula to available partitions. We considered only a few of such information theoretic indices, in order to assess their overall behavior.

For more details on other indices used here, readers should consult the source. Several indices included in our study do not have any structural relationship to other indices. They are based on definitions in which a particular quality of molecular graphs is used. For example, the Hosoya $Z$ index[35] is based on enumeration of numbers $p(G,k)$ which signify the number of nonadjacent edges in a graph. The Wiener number[36] $E$ enumerates all pairs of carbon atoms at different sides of each bond, summed over all bonds. The centric index[23] was introduced by counting the steps in the pruning of terminal edges in a tree. Such *ad hoc* descriptors, because they are apparently unrelated to other descriptors, are more likely to show a distinct behavior in different regressions and, as such, are desirable. On the other hand, an index that is structurally related to the existing indices is less likely to have additional qualities. For this reason, we have not included combinations of different indices, like the super-index.[40] Since here we have an averaging process which, while possibly resulting in a distinctive index for different structures, is likely to reduce the signal-to-noise ratio of the components that may be dominant in a regression. We decided not to consider indices based on topographic and distance-sensitive matrices[41] and indices derived from combining topological and topographic features of a structure.[42] Such descriptors represent 3-dimensional structures and will be important in extending structure-property studies to structure-activity studies. However, for most of the available experimental information on octanes it is not clear to which conformation, or what a mixture of conformations, the data correspond. Hence, such 3-dimensional descriptors will have to wait their applications when most of physico-chemical properties are considered. They may be of interest in correlations of molecular magnetic (NMR) data, and certainly will find use in structure-activity studies where the 3-dimensional geometry of receptors will dictate critically many properties of drugs. The preliminary results of Trinajstić and collaborators,[19] who considered a generalized Wiener index associated with a 3-dimensional structures, illustrate the difficulties associated with the selection of a single conformer to represent a structure (isomer). Once the difficulties associated with flexible molecules and mixtures of conformers are better understood, one will be in a position to select properly combinations of 3-dimensional structural invariants, and these are likely to be useful since they will contain more information than the graphical model of a molecule in which only the connectivities are registered.

Finally, we decided not to include in this study, indices that show considerable degeneracy for the 18 isomers of octane. This eliminates, for example, simple indices based on the valences of vertices. For this same reason, we have not considered Kier's

kappa (shape) indices[43] since their variation among isomers is the same as described by the connectivity indices. However, the exclusion of such indices does not mean that they may not be useful when the pool of the structures considered is widened.

## RESULTS

We collected the statistical information, that is the standard error $S$ and the coefficient of regression $R$, for over 750 regressions using some 20 properties of octanes and about 40 descriptors. Before we analyze the wealth of the results we will make a few general remarks. By restricting the analysis to octane isomers, we have eliminated the dominant role of the molecular size in correlations. As a consequence, there is a parallelism between the standard errors and the correlation coefficients for every single property. In Figure 1, we illustrate the relationship between $S$ and $R$ for the heats of formation $\Delta H$. The dependence of $S$ on $R$ shown in Figure 1 is typical of all the properties. The relationship between $S$ and $R$ allows one to compare correlations of different properties, since we can use the value of $R$ as a basis for the relative quality of individual regressions. In order to facilitate comparison and qualify various results, we have adopted the following, somewhat arbitrary, but conservative, scale:

| Regression coefficient | Quality |
|---|---|
| 0.990 (and higher) | Outstanding |
| 0.975 (and higher) | Excellent |
| 0.950 (and higher) | Very good |
| 0.925 (and higher) | Good |
| 0.900 (and higher) | Fair |

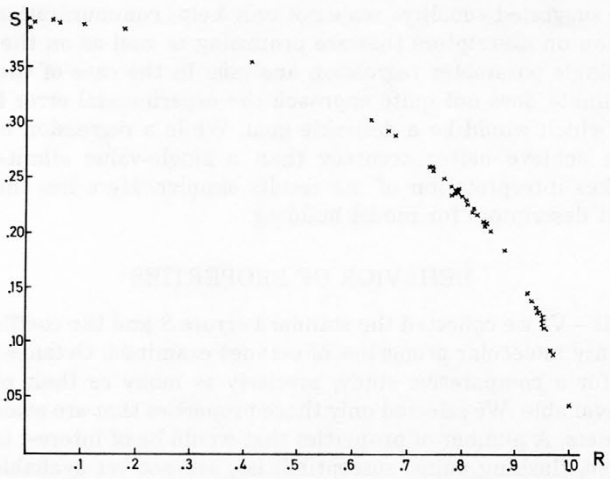


Figure 1. Dependence of the standard error $S$ on the regression coefficient $R$ for the case of the heats of formation regression using the data in Table IV – Table VIII. The figure illustrates how even unacceptable results (too low $R$ values) can, when combined with other results, offer useful insights.

According to the above scale, in the case of octane boiling points, the following standard errors are expected to qualify as:

| | |
|---|---|
| limit | $S$ = 1.29 °C and smaller |
| outstanding | $S$ = 1.46 °C and smaller |
| excellent | $S$ = 1.71 °C and smaller |
| very good | $S$ = 2.10 °C and smaller |
| good | $S$ = 2.45 °C and smaller |
| fair | $S$ = 2.76 °C and smaller |

The »limit« and the other values shown refer to simple single- variable regressions (of the boiling points in octanes). This particular »quality« scale holds for regressions when molecules of the *same size* are considered. When alkanes of different size are considered, the dominant effect of the molecular size will, as a rule, considerably improve $R$, the coefficient of regression, however, without improving $S$, the standard error. Since the dependence of molecular properties on the number of atoms in a molecule (or molecular weight) can be described by many simple descriptors, it seems only prudent to eliminate their dominance when investigating variations in properties due to *shape, branching, cyclicity*. The »limit« is the extrapolated value for $R$ = 1.000 when the curve such as that in Figure 1 (summarizing $R/S$ relationship for regressions for the boiling points in octanes) was fitted with the 5-th order polynomial (which produced the smallest standard error for a polynomial fitting: $S$ = 0.00723) for polynomials from degree two to degree twelve.[44]

If we now examine the first row in Tables III – VI, which lists the standard errors and the coefficients of regression for boiling points of the 18 isomers of octane we see that the best results (the smallest $S$) derived are still below desirable quality. On the other hand, we obtained very good and even excellent correlations for several other properties. The suggested »quality« scale not only helps communication but allows one to focus attention on descriptors that are promising as well as on the properties that are elusive to single parameter regression analysis. In the case of the boiling points, the projected »limit« does not quite approach the experimental error for the reported boiling points, which would be a desirable goal. While a regression based on several descriptors can achieve better accuracy than a single-value »limit«, use of fewer descriptors makes interpretation of the results simpler. Here lies the importance of single dominant descriptors for model building.

## BEHAVIOR OF PROPERTIES

In Tables III – VI we collected the standard errors $S$ and the coefficients of regressions $R$ for twenty molecular properties of octanes examined. Octanes offer a good set of compounds for a comparative study, precisely as many as their physico-chemical properties are available. We selected only those properties that are available for at least 15 or more isomers. A number of properties that would be of interest to examine, such as soot threshold, flashing point, susceptibilities, are not yet available but for half a dozen compounds. Hence, they were excluded. Equally, we did not consider melting points, since these depend on the crystal packing, and that is beyond the scope of molecular descriptors which best simulate the properties of individual molecules.

In Tables III – VI, we have indicated in bold type all the $R$ values above 0.900. This will help readers to identify »fair«, »good«, »very good« and »excellent« regres-

TABLE III

*Standard errors* S *(top lines) and the coefficients of regressions* R *(bottom lines) for regressions using connectivity index and closely related descriptors*

| Property | Descriptors | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $^1\chi$ | $^2\chi$ | $^3\chi$ | $^1\chi-{}^2\chi$ | $^2\chi-{}^3\chi$ | $^1\chi/{}^2\chi$ | $1/{}^1\chi$ | $1/{}^2\chi$ | $1/{}^3\chi$ | $^{ID}\chi$ | ID |
| BP | 3.60 | 2.98 | 6.03 | 3.09 | 3.11 | 2.93 | 3.78 | 2.91 | 5.74 | 4.01 | 5.19 |
| | .821 | .882 | .295 | .872 | .870 | .886 | .801 | .887 | .416 | .772 | .569 |
| $S$ | 1.97 | 2.61 | 4.11 | 2.45 | 4.32 | 2.59 | 1.99 | 2.78 | 4.39 | 1.61 | 1.76 |
| | **.906** | .829 | .468 | .850 | .371 | .831 | **.905** | .803 | .337 | **.938** | **.926** |
| $\Delta H$ | .113 | .133 | .393 | .124 | .291 | .118 | .136 | .131 | .395 | .123 | .206 |
| | **.958** | **.941** | .108 | **.949** | .676 | **.955** | **.932** | .943 | .054 | **.951** | .854 |
| HA | .678 | .528 | 1.26 | .556 | .697 | .513 | .786 | 725 | .471 | .786 | 1.05 |
| | .851 | **.912** | .206 | **.902** | .841 | **.917** | .827 | **.931** | .306 | .793 | .578 |
| $\Delta_f H$ | 2.85 | 2.21 | 5.24 | 2.33 | 2.86 | 2.16 | 3.04 | 1.97 | 5.09 | 3.30 | 4.40 |
| | .847 | **.911** | .215 | **.901** | .846 | **.916** | .824 | **.930** | .314 | .789 | .571 |
| $\Delta_l^g H$ | 2.78 | 2.25 | 3.36 | 2.36 | 2.11 | 2.32 | 2.87 | 2.17 | 3.23 | 3.09 | 3.69 |
| | .686 | .810 | .477 | .787 | .835 | .797 | .663 | .825 | .538 | .590 | .269 |
| $\Delta_v H$ | .662 | .824 | 1.50 | .777 | 1.30 | .772 | .713 | .804 | 1.50 | .627 | .911 |
| | .898 | .837 | .095 | .856 | .505 | .859 | .881 | .846 | .074 | **.909** | .796 |
| HV | .446 | .561 | 1.50 | .514 | 1.16 | .536 | .478 | .621 | 1.47 | .445 | .891 |
| | **.955** | **.927** | .013 | **.939** | .631 | **.934** | **.948** | **.910** | .190 | **.955** | .803 |
| $\sigma$ | .751 | .609 | .584 | .643 | .241 | .668 | .751 | .650 | .521 | .807 | .904 |
| | .559 | .740 | .765 | .704 | **.964** | .674 | .559 | .696 | .818 | .454 | .061 |
| $N$ | 16.0 | 20.3 | 23.2 | 19.4 | 25.3 | 19.3 | 16.6 | 20.1 | 24.7 | 13.1 | 9.96 |
| | .778 | .609 | .419 | .650 | .131 | .656 | .759 | .619 | .257 | .860 | **.921** |
| AC | .0162 | .0221 | .0271 | .0208 | .0307 | .0203 | .0170 | .0213 | .0296 | .0120 | .0058 |
| | .855 | .706 | .493 | .744 | .174 | .759 | .837 | .729 | .309 | **.923** | **.983** |
| $R$ | .119 | .141 | .168 | .136 | .180 | .129 | .121 | .140 | .179 | .116 | .108 |
| | .770 | .655 | .436 | .684 | .256 | .723 | .761 | .663 | .278 | .786 | .814 |
| $T_c$ | 8.70 | 7.66 | 7.75 | 7.91 | 4.59 | 8.34 | 8.60 | 8.24 | 7.55 | 9.03 | 9.97 |
| | .499 | .647 | .635 | .616 | .889 | .556 | .516 | .571 | .659 | .437 | .115 |
| $p_c$ | 1.47 | 1.48 | .998 | 1.48 | 1.31 | 1.47 | 1.48 | 1.48 | 1.12 | 1.45. | 1.29 |
| | .141 | .036 | .739 | .004 | .472 | .104 | .094 | .082 | .652 | .197 | .497 |
| $V_c$ | 15.2 | | 12.6 | 15.9 | 15.8 | | 15.5 | | 13.1 | 14.2 | 10.7 |
| | .376 | | .640 | .236 | .259 | | .332 | | .603 | .499 | .760 |
| $V_m$ | 2.63 | 2.55 | .554 | 2.59 | 1.71 | 2.61 | 2.63 | 2.59 | .921 | 2.61 | 2.24 |
| | .013 | .246 | **.978** | .189 | .762 | .142 | .000 | .183 | **.937** | .143 | .529 |
| $R_m$ | 188 | .186 | .046 | .187 | .132 | .188 | .188 | .187 | .077 | .184 | .151 |
| | .085 | .179 | **.970** | .121 | .712 | .071 | .071 | .116 | **.913** | .218 | .601 |
| $\rho$ | .0122 | .0116 | | .0119 | .0120 | | .0123 | | .0044 | .0123 | .0103 |
| | .057 | .323 | | .767 | .238 | | .064 | | **.936** | .090 | .550 |
| CS | 15.6 | 12.9 | 14.2 | 13.6 | 14.2 | | 15.6 | | 12.8 | 18.9 | 19.0 |
| | .581 | .739 | .672 | .706 | .676 | | .583 | | .743 | .860 | **.921** |

TABLE IV

*Standard errors* S *(top lines) and coefficients of regressions* R *(bottom lines) for regressions using invariants derived from adjacency, distance and vicinal matrices and their combinations.*

| Property | Descriptors | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | TI | **A+D** | ID(**A**) | $X$(**A**) | $\chi(W)$ | ID($W$) | $X(W)$ | $J$ | $\chi$(**D**) | ID(**D**) |
| BP | 4.37 | 5.43 | 4.31 | | 3.70 | 4.18 | 5.66 | 5.50 | 4.40 | 5.28 |
| | .722 | .509 | .731 | | .809 | .750 | .442 | .489 | .717 | .547 |
| $S$ | 1.71 | 2.41 | 1.64 | | 2.25 | 2.02 | 2.63 | 2.07 | 3.48 | 1.98 |
| | **.931** | .856 | **.936** | | .875 | **.901** | .826 | .896 | .665 | **.906** |
| $\Delta H$ | .141 | .240 | .136 | .215 | .127 | .137 | .264 | .239 | .263 | .221 |
| | **.934** | .794 | **.939** | .840 | **.947** | **.938** | .743 | .797 | .746 | .829 |
| HA | .867 | 1.14 | .864 | 1.06 | .759 | | 1.19 | 1.13 | .877 | 1.10 |
| | .741 | .467 | .742 | .571 | .808 | | .393 | .481 | .733 | .522 |
| $\Delta_f H$ | 3.63 | 4.77 | 3.63 | 4.43 | 3.19 | 3.67 | 4.95 | 4.73 | 3.64 | 4.60 |
| | .735 | .457 | .737 | .564 | .803 | .728 | .383 | .472 | .733 | .514 |
| $\Delta_1^g H$ | 3.30 | 3.79 | 3.31 | 3.69 | 3.02 | 3.32 | 3.83 | 3.79 | 3.05 | 3.77 |
| | .507 | .148 | .503 | .271 | .615 | .498 | .063 | .147 | .605 | .186 |
| $\Delta_v H$ | .689 | .961 | .649 | | .675 | .660 | 1.04 | .972 | 1.09 | .852 |
| | .889 | .556 | **.903** | | .894 | .899 | .721 | .764 | .689 | .825 |
| HV | .624 | .996 | .583 | | .548 | .589 | 1.09 | 1.01 | 1.08 | .933 |
| | **.909** | .746 | **.921** | | **.931** | **.919** | .687 | .737 | .690 | .782 |
| $\sigma$ | .859 | .905 | .855 | | .793 | .851 | .900 | .904 | .739 | .905 |
| | .315 | .029 | .330 | | .483 | .341 | .110 | .058 | .578 | .023 |
| $N$ | 10.6 | 8.18 | 10.2 | | 14.3 | 9.31 | 10.7 | 9.27 | 23.0 | |
| | **.910** | **.947** | **.917** | | .828 | **.931** | **.909** | **.932** | .437 | |
| AC | .0090 | .0096 | .0092 | .0069 | .0157 | .0109 | .0116 | .0076 | .0272 | .0087 |
| | **.957** | **.951** | **.956** | **.975** | .864 | **.937** | **.928** | **.970** | .489 | **.960** |
| $R$ | .093 | .085 | .093 | | .095 | .080 | .095 | .096 | .142 | .080 |
| | .868 | .890 | .866 | | .861 | **.904** | .861 | .858 | .646 | **.902** |
| $T_c$ | 9.65 | 10.0 | 9.53 | | 9.13 | 9.58 | 10.0 | 10.0 | 8.44 | 9.99 |
| | .274 | .003 | .313 | | .415 | .298 | .059 | .007 | .541 | .100 |
| $p_c$ | 1.36 | 1.18 | 1.39 | | 1.43 | 1.37 | 1.14 | 1.21 | 1.48 | 1.29 |
| | .390 | .607 | .347 | | .258 | .386 | .637 | .581 | .071 | .496 |
| $V_c$ | 13.6 | 9.97 | 13.5 | 10.8 | 14.8 | 13.4 | 9.24 | 9.99 | 16.4 | 11.3 |
| | .557 | .794 | .565 | .753 | .429 | .572 | .826 | .793 | .022 | .722 |
| $V_m$ | 2.52 | 2.06 | 2.52 | | 2.62 | 2.53 | 1.93 | 2.03 | 2.57 | 2.17 |
| | .297 | .621 | .287 | | .113 | .283 | .677 | .639 | .223 | .567 |
| $R_m$ | .175 | .135 | .176 | | .185 | .176 | .124 | .134 | .186 | .144 |
| | .370 | .698 | .364 | | .192 | .367 | .752 | .706 | .169 | .644 |
| $\rho$ | .0119 | .0090 | .0119 | .0101 | .0123 | .0119 | .0082 | .0091 | .0119 | .0100 |
| | .267 | .679 | .256 | .572 | .042 | .251 | .743 | .675 | .261 | .585 |
| CS | 17.9 | 19.2 | 17.8 | 19.1 | 16.7 | 17.9 | 19.2 | 19.2 | | 19.1 |
| | .365 | .010 | .378 | .130 | .498 | .360 | .069 | .031 | | .109 |

TABLE V

*Standard errors* S *(top lines) and the coefficients of regressions* R *(bottom lines) for regressions using miscellaneous simple structural invariants*

| Property | Descriptors | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\chi^{[1]}$ | $\chi^{[1/2]}$ | $\chi^{z[-1]}$ | P′/P | $\chi'/\chi$ | $W'/W$ | WW | $W'$ | $Z$ | $W$ | W/Z |
| BP | 5.46 | 5.01 | 3.25 | 5.12 | 5.91 | 5.52 | 5.31 | 6.11 | 2.90 | 5.31 | 3.09 |
| | .500 | .609 | .857 | .584 | .350 | .483 | .540 | .248 | .888 | .539 | .872 |
| $S$ | 1.57 | 1.40 | 2.60 | 2.13 | 4.61 | 2.49 | 2.66 | 4.38 | 4.86 | 2.23 | 3.79 |
| | **.942** | **.954** | .829 | .888 | .135 | .845 | .821 | .338 | .580 | .878 | .582 |
| $\Delta H$ | .231 | .187 | .150 | .213 | .389 | .248 | .238 | .360 | .095 | .226 | .243 |
| | .812 | .881 | **.925** | .842 | .180 | .779 | .798 | .414 | **.971** | .820 | .840 |
| HA | 1.08 | .991 | .587 | 1.08 | 1.27 | 1.16 | 1.13 | 1.28 | .574 | 1.11 | .586 |
| | .542 | .640 | .891 | .550 | .182 | .441 | .476 | .095 | .896 | .506 | .891 |
| $\Delta_f H$ | 4.53 | 4.15 | 2.45 | 4.51 | 5.27 | 4.84 | 4.73 | 5.34 | 2.41 | 4.65 | 4.43 |
| | .535 | .633 | .889 | .541 | .188 | .432 | .466 | .083 | .893 | .497 | .564 |
| $\Delta_l^g H$ | 3.75 | 3.59 | 2.40 | 3.72 | 3.65 | 3.81 | 3.77 | 3.83 | 2.52 | 3.76 | 2.02 |
| | .211 | .350 | .780 | .240 | .303 | .117 | .179 | .033 | .753 | .189 | .850 |
| $\Delta_v H$ | .972 | .822 | .828 | .887 | 1.40 | .981 | .940 | 1.25 | .686 | .921 | 1.19 |
| | .764 | .838 | .836 | .808 | .375 | .759 | .781 | .556 | .890 | .791 | .611 |
| HV | .995 | .808 | .627 | .904 | 1.40 | 1.02 | .964 | 1.27 | .429 | .949 | 1.04 |
| | .747 | .842 | **.908** | .797 | .355 | .730 | .764 | .524 | **.958** | .773 | .721 |
| $\sigma$ | .905 | .898 | .643 | .903 | .769 | .904 | .905 | .901 | .677 | .905 | .363 |
| | 0.33 | .130 | .704 | .078 | .527 | .057 | .021 | .093 | .663 | .008 | **.916** |
| $N$ | 10.2 | 8.89 | 20.1 | 8.56 | 25.5 | 9.98 | 7.40 | 20.9 | 18.1 | 7.42 | 24.6 |
| | **.917** | **.938** | .617 | **.942** | .089 | **.921** | **.957** | .576 | .704 | **.957** | .268 |
| AC | .0063 | .0039 | .0222 | .0102 | .0309 | .0100 | 0.112 | .0230 | .0190 | .0083 | .0294 |
| | **.978** | **.992** | .702 | **.945** | .123 | **.947** | **.933** | .673 | .791 | **.964** | .328 |
| $R$ | .109 | .102 | .136 | .091 | .186 | .090 | .080 | .143 | .124 | .085 | .171 |
| | .813 | .837 | .684 | .873 | .058 | .875 | **.903** | .642 | .747 | .890 | .395 |
| $T_c$ | 10.0 | 9.88 | 7.89 | 9.96 | 9.63 | 10.0 | 10.0 | 9.77 | 8.28 | 10.0 | 5.51 |
| | .072 | .117 | .618 | .121 | .281 | .023 | .023 | .231 | .565 | .041 | .836 |
| $p_c$ | 1.29 | 1.33 | 1.48 | 1.28 | 1.47 | 1.16 | 1.18 | 1.10 | 1.47 | 1.21 | 1.38 |
| | .493 | .438 | .019 | .509 | 1.47 | .620 | .608 | .668 | .099 | .579 | .367 |
| $V_c$ | 11.0 | 12.0 | 16.1 | 9.89 | 16.4 | 10.2 | 10.2 | 12.5 | 15.6 | 10.3 | 16.1 |
| | .740 | .683 | .178 | .798 | .018 | .781 | .785 | .646 | .308 | .780 | .752 |
| $V_m$ | 2.08 | 2.32 | 2.58 | 2.24 | 2.37 | 2.03 | 2.15 | 2.19 | 2.62 | 2.12 | 2.03 |
| | .612 | .475 | .206 | .524 | .433 | .639 | .576 | .553 | .112 | .592 | .636 |
| $R_m$ | .140 | .156 | .187 | .150 | .171 | .132 | .141 | .145 | .189 | .140 | .153 |
| | .669 | .542 | .142 | .605 | .421 | .715 | .662 | .641 | .036 | .669 | .588 |
| $\rho$ | .0095 | .0108 | .0118 | .0102 | .0111 | .0088 | .0093 | .0092 | .0120 | .0095 | .0091 |
| | .633 | .479 | .280 | .561 | .430 | .697 | .651 | .659 | .206 | .641 | .670 |
| CS | 19.1 | 18.7 | 13.6 | 19.1 | 16.3 | 19.2 | 19.2 | 18.4 | 14.4 | 19.2 | 19.1 |
| | .105 | .227 | .709 | .112 | .534 | .017 | .020 | .283 | .663 | .056 | **.929** |

TABLE VI

*Standard errors* S *(top lines) and the coefficients of regressions* R *(bottom lines) for regressions using information theoretic and matrix multiplicative invariants.*

| Property | Descriptors | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | IED | IWD | EC | PO | $U$ | $V$ | $X$ | $Y$ | AZV |
| BP | 5.34 | 5.19 | 5.61 | 6.16 | 6.04 | 5.58 | 5.43 | 5.77 | 3.53 |
| | .533 | .569 | .456 | .214 | .286 | .467 | .510 | .403 | .823 |
| $S$ | 1.88 | 2.43 | 2.02 | 4.64 | 2.91 | 1.91 | 1.90 | 1.84 | 1.71 |
| | **.915** | .853 | **.901** | .092 | .780 | **.912** | **.913** | **.919** | .874 |
| $\Delta H$ | .224 | .222 | .253 | .394 | .307 | .244 | .230 | .264 | .095 |
| | .824 | .827 | .768 | .061 | .628 | .787 | .814 | .745 | **.965** |
| HA | 1.11 | | 1.16 | 1.27 | 1.25 | 1.14 | 1.11 | 1.16 | .702 |
| | .507 | | .437 | .178 | .240 | .475 | .510 | .440 | .811 |
| $\Delta_f H$ | 4.65 | 4.58 | 4.84 | 5.27 | 5.22 | 4.74 | 4.64 | 4.83 | 2.95 |
| | .498 | .519 | .429 | .178 | .230 | .466 | .501 | .432 | .809 |
| $\Delta_l^g H$ | 3.79 | 3.73 | 3.82 | 3.75 | 3.80 | 3.81 | 3.78 | 3.82 | 2.83 |
| | .153 | .229 | .060 | .206 | .131 | .119 | .172 | .051 | .675 |
| $\Delta_v H$ | .914 | .902 | 1.02 | 1.51 | 1.19 | 1.00 | .952 | 1.08 | .688 |
| | .795 | .801 | .737 | .014 | .610 | .747 | .775 | .699 | .890 |
| HV | .943 | .922 | 1.06 | 1.49 | 1.26 | 1.04 | .983 | 1.13 | .439 |
| | .776 | .787 | .704 | .039 | .542 | .715 | .754 | .654 | **.956** |
| $\sigma$ | .905 | .904 | .901 | .866 | .865 | .902 | .905 | .891 | .770 |
| | .006 | .060 | .094 | .291 | .297 | .092 | .032 | .178 | .527 |
| $N$ | 9.01 | 7.27 | | 25.5 | 14.7 | 10.2 | 8.92 | 12.6 | 15.6 |
| | **.936** | **.959** | | .025 | .819 | **.916** | **.937** | .871 | .792 |
| AC | .0081 | .0105 | .0116 | .0308 | .0156 | .0077 | .0067 | .0095 | .0144 |
| | **.965** | **.942** | **.928** | .146 | .865 | **.969** | **.977** | **.952** | .887 |
| $R$ | .090 | .084 | .101 | .187 | .113 | .105 | .099 | .120 | .120 |
| | .876 | .894 | .841 | .030 | .796 | .826 | .848 | .768 | .769 |
| $T_c$ | 9.98 | 10.0 | 10.0 | 9.46 | 9.88 | 10.0 | 10.0 | 10.0 | 8.75 |
| | .103 | .070 | .049 | .334 | .175 | .020 | .048 | .009 | .263 |
| $p_c$ | 1.30 | 1.22 | 1.30 | 1.40 | 1.12 | 1.25 | 1.25 | 1.29 | 1.26 |
| | .483 | .567 | .484 | .312 | .658 | .540 | .541 | .496 | .484 |
| $V_c$ | 11.0 | 10.5 | 11.3 | 16.4 | 8.67 | 10.0 | 10.3 | 10.2 | 14.8 |
| | .734 | .766 | .727 | .069 | .849 | .791 | .780 | .783 | .436 |
| $V_m$ | 2.13 | 2.20 | 2.04 | 2.50 | 1.59 | 1.97 | 2.07 | 1.84 | 2.63 |
| | .590 | .546 | .634 | .314 | .797 | .663 | .619 | .716 | .045 |
| $R_m$ | .140 | .147 | .134 | .180 | .097 | .130 | .137 | .121 | .187 |
| | .669 | .630 | .703 | .304 | .857 | .725 | .686 | .768 | .117 |
| $\rho$ | .0097 | .0095 | .0093 | .0117 | .0067 | .0089 | .0093 | .0083 | .2476 |
| | .616 | .633 | .652 | .304 | .839 | .694 | .653 | .739 | .269 |
| CS | 19.1 | 19.2 | 19.2 | 18.1 | 18.8 | 19.2 | 19.2 | 19.2 | 16.4 |
| | .091 | .074 | .029 | .337 | .212 | .034 | .069 | .007 | .507 |

sions among the several hundreds reported. We may also add that only one (in several hundreds!) of the reported regressions qualifies as »outstanding«. It has to be seen if this exceptional value for a single descriptor regression of physico-chemical properties (among molecules of the same size) can be matched for other molecular properties, and other molecules.

A glance at Tables III – VI immediately reveals that some of the properties considered can be classified as »difficult« and some as »easy« if one judges difficult/easy as the ratio between successful regressions (*i.e.* those with $R$ above 0.900) and not so successful ones (regressions with $R$ close to but less than 0.900 and greater than 0.800 may be referred to as mediocre), or uninteresting ($R$ below 0.800) and outright non-existent ($R$ below 0.100).

## »DIFFICULT« MOLECULAR PROPERTIES

### *Boiling point*

We have already mentioned regressions of the boiling points in octane that are hard to represent successfully by a single-variable regression. The following descriptors produce $R$ greater than 0.800 (but less than desirable $R$ = 0.900 or better):

| | $R$ | $S$ |
|---|---|---|
| *Hosoya Z* | 0.888 | 2.90 |
| Reciprocal $1/{}^2\chi$ | 0.887 | 2.91 |
| Ratio ${}^1\chi/{}^2\chi$ | 0.886 | 2.93 |
| Connectivity ${}^2\chi$ | 0.882 | 2.98 |
| Difference ${}^1\chi-{}^2\chi$ | 0.872 | 3.09 |
| Difference ${}^2\chi-{}^3\chi$ | 0.870 | 3.11 |
| AZV | 0.823 | 3.53 |
| Connectivity ${}^1\chi$ | 0.821 | 3.60 |
| Wiener matrix ${}^1\chi$ | 0.809 | 3.71 |
| Reciprocal $1/{}^1\chi$ | 0.801 | 3.78 |

Note that, though the connectivity index ${}^1\chi$ does not give the best regression, various simple combinations of ${}^1\chi$ give improved correlations, which also includes connectivity-type weighted paths of length based on the Wiener matrix of a graph.

### *Critical temperature*

The rows in Tables III – VI referring to the critical temperatures show not only that this property is hard to represent by a single descriptor, but that most descriptors do not point to any correlation. Here we find only two descriptors giving $R$ above 0.800:

| | | |
|---|---|---|
| Difference ${}^1\chi - {}^2\chi$ | 0.889 | 4.59 |
| Ratio $W/Z$ | 0.836 | 5.51 |

Ten descriptors show no correlation at all ($R$ less than 0.100), including also the Wiener index $W$ ($R$ = 0.041), yet in a combination with the Hosoya $Z$ index (which yields $R$ = 0.565) we obtain appreciable improvement: $R$ = 0.836. This illustrates well

that simple (arithmetic, algebraic and other) combinations of indices may offer acceptable solutions when the component indices fail. Interpretation of such combinations may lead to novel structural concepts. Plat[45] has argued that $W$ can be viewed as a measure of the molecular volume. On the other hand, the Hosoya $Z$ index correlated well with the properties expected to depend on the molecular »surface« (such as BP, for instance). The ratio $W/Z$, thus, corresponds to a measure of an effective »radius« of a molecule. Clearly, this is not the only index that can be so interpreted, the individual cases corresponding to different measures for volume and surface.

*Critical pressure*

This property appears even harder to correlate with a single descriptor, judging by the descriptors investigated here. The best result

$$^3\chi \qquad R = 0.739 \qquad S = 0.998$$

does not even reach the $R = .800$ value. If we look at the column of Table III corresponding to $^3\chi$, we see that this descriptor is rather unique (among those considered) in that it is the only descriptor (besides its reciprocal $1/^3\chi$) that is successful in correlating *MV*, *MR*, and liquid densities in octane isomers. The correlations qualify as excellent »almost« excellent, and are among the »best« results in Tables III – VI.

*Critical volume*

In contrast to the regressions involving the critical pressure, almost half of the descriptors (twenty) have the $R$ value above 0.700, while none did so in the case of critical pressure, except the indicated $^3\chi$. Among the more successful indices, we find Balaban's $U$ ($R = .849$), the ratio $ID/^1\chi$ based on the distance matrix ($R = 0.832$) and the first eigenvalue of the vicinal matrix ($R = 0.826$). Interpretation of these results, while possibly better, awaits a more successful single-descriptor regression, which may not only produce a simpler interpretation but may somewhat illuminate some of the descriptors that are more convoluted.

*Heat of formation (liquid)*

Again, no descriptor yields a regression with $R$ above 0.900, the following being the best results:

| Descriptor | $R$ | $S$ |
|---|---|---|
| $W/Z$ | 0.850 | 2.02 |
| $^2\chi - {}^3\chi$ | 0.835 | 2.11 |
| $1/^2\chi$ | 0.825 | 2.17 |
| $^2\chi$ | 0.810 | 2.25 |
| $^1\chi/^2\chi$ | 0.797 | 2.32 |
| $^1\chi - {}^2\chi$ | 0.787 | 2.36 |
| $\chi^{[-1]}$ | 0.780 | 2.40 |
| $Z$ | 0.753 | 2.52 |

Again, we see $W/Z$ to be a useful descriptor, even though $W$ alone has $R = 0.189$ ! Note also again how simple combinations of the connectivity indices emerge as useful

descriptors. Among better descriptors, we also find Altenburg's index, which we denoted as $\chi^{[-1]}$ and which is related to the connectivity index $^1\chi$.

As additional difficult properties we may include those properties from Tables III – VI in which only 1–2 descriptors offer regression with $R$ above 0.900. This includes the chemical shift sums (or the average carbon–13 chemical shift for a molecule) with the best results with $W/Z$ and the distant second choice $^2\chi$

| | | |
|---|---|---|
| $W/Z$ | $R$ = 0.929 | $S$ = 7.12 |
| $^2\chi$ | $R$ = 0.739 | $S$ = 12.95 |

as well as previously mentioned densities, molecular volumes and molar refractions.

## »EASY« MOLECULAR PROPERTIES

As easy properties (*i.e.* easy to find a descriptor that will give a fair or better regression) we consider those that lead to acceptable regressions with a dozen and more descriptors (among the 42 considered). These include:

| Property | Correlations | | | | | |
|---|---|---|---|---|---|---|
| | $R$ > 0.990 Outstanding | $R$ > 0.975 Excellent | $R$ > 0.950 Very good | $R$ > 0.925 Good | $R$ > 0.900 Fair | Total |
| AC | 1 | 4 | 9 | 7 | 1 | 22 |
| $\Delta H$ | | | 5 | 9 | 0 | 14 |
| $N$ | | | 3 | 6 | 7 | 16 |
| HV | | | 4 | 5 | 5 | 14 |
| $S$ | | | 1 | 5 | 9 | 15 |

We see from the above table that almost half of the descriptors can well correlate with the acentric character (AC). Hence, this property should not be taken as an indicator for the »success« of a descriptor. The designer of a novel descriptor for this particular property ought to get results better than the best available, or at least as good, *i.e.* better than:

| Descriptor | $R$ | $S$ |
|---|---|---|
| $\chi^{[+1/2]}$ | 0.992 | .0039 |
| ID | 0.983 | .0058 |
| $\chi^{[+1]}$ | 0.980 | .0062 |

## CHALLENGES SEARCH FOR BETTER REGRESSIONS

In Table VII, we summarized the best results so far obtained for the 18 isomers of hexane and the twenty properties examined. A challenge to those interested in structure-property studies is to design or discover better *ad hoc* descriptors for the properties listed. It is possible that yet another combination of the topological indices here used may produce still better results, and that also has to be investigated. However, highly convoluted *ad hoc* constructions, while possibly producing a better single-variable regression may become less transparent to interpretation. The object of the

TABLE VII

*The best results: the best standard errors* S *and the accompanying best coefficients of regressions* R *together with an indication of the descriptors used. The descriptors marked by asterisks have been tested in this work for the first time*

| Symbol | Property | $S$ | $R$ | Descriptor | |
|---|---|---|---|---|---|
| BP | Boiling points | 2.90 | .888 | $Z$ | |
| $S$ | Entropy | 1.40 | **.954** | $m_{1/2}$ | |
| $\Delta H$ | Heat of vaporization | .113 | **.942** | $^1\chi$ | |
| HF | Heat of formation | .471 | **.931** | $1/^2\chi$ | |
| HA | Heat of atomization | .725 | **.931** | $1/^2\chi$ | |
| $\Delta_f H$ | Heat of formation | .197 | **.930** | $1/^2\chi$ | |
| $\Delta_l^g H$ | Heat of vap. (liquid) | 2.02 | .850 | $W/Z$ | * |
| $\Delta_v H$ | Heats of vap.(vapor) | .627 | **.909** | ID($\chi$) | * |
| HV | Heats of vap. | .429 | **.958** | $Z$ | |
| $T_c$ | Critical temperature | 4.59 | .889 | $^1\chi - {}^2\chi$ | |
| $p_c$ | Critical pressure | 1.10 | .668 | $1/^2\chi$ | |
| $V_c$ | Critical volume | 8.67 | .849 | $\chi(V)$ | * |
| $\sigma$ | Surface tension | .241 | **.964** | $^2\chi - {}^3\chi$ | |
| $R$ | Quadratic mean radius | .080 | **.904** | ID($V$) | * |
| AC | Pitzer acentric factor | .0039 | **.992** | $^2\chi$ | |
| $N$ | Octane number | 7.27 | **.959** | $I_{WD}$ | |
| $V_m$ | Molecular Volume | .554 | **.978** | $^3\chi$ | |
| $R_m$ | Molar refraction | .046 | **.970** | $^3\chi$ | |
| $\rho$ | Density | .0025 | **.979** | $^3\chi$ | |
| CS | C–13 chem. shift sum | 19.1 | **.929** | $W/Z$ | * |
| $C_p$ or $C_v$ | Heat capacity | | | | |
| RT | Retention time | | | | |

regression analysis, besides offering a predictive tool, is also to offer some insights into the structure-property relationship and molecular modeling. While most of the descriptors employed here gave a straightforward structural interpretation, such an interpretation becomes less and less apparent as the underlying constructional steps are compounded. For example, regularities in the first eigenvalues of a matrix (adjacency, distance or vicinal) are not so apparent, and neither is the structural relationship of the information theoretic indices so direct. The same, to some extent, applies to several of Balaban's indices derived by matrix-vector multiplications. But the ease of interpretation is, to a great extent, in the eye of the beholder! So one should not be discouraged if an interpretation appears convoluted, as long as one can recognize the regularities associated with a particular descriptor.

A closer look at Table VII immediately shows that the connectivity indices (in various combinations, such as reciprocals, differences, ID numbers) emerged among the leading indices. An asterisk (*) indicates indices that have not been previously considered. They include the ID numbers based on the connectivity indices (rather than based on weighted paths). Let us also point out the simple ratio $W/Z$, of the two oldest topological descriptors which, as shown here for the first time, appears as a useful descriptor. Hence, a new source of indices to be explored are simple combinations of

the existing indices. The topological state matrix $\boldsymbol{T}$ of Kier and Hall, the elements $t_{ij}$ of which are given by the product of the inverse square root of valences for all the vertices in the unique path between $i$ and $j$, is also a potential source of additional graph invariants.

In concluding this section on challenges, it is worth observing that even the best regression for the critical pressure ($R$ = 0.668) is far from satisfactory. Is there some significance in the fact that the best regressions for the critical temperatures and the critical volumes are not very satisfactory?

## BEHAVIOR OF DESCRIPTORS

If we closely examine Tables III – VII by focussing attention on individual columns rather than rows, we again see a considerable difference in the performance of individual descriptors. The connectivity indices and their combinations appear to show good correlation with several properties. On the other hand, several of the descriptors examined failed to lead to a single regression exceeding $R$ = 0.900. This is the case of the ponderal index PI, the distance matrix based $\chi$, the ratio of ID numbers and $\chi$ (both based on the distance matrix), index $\chi'/\chi$, $W'$, and Balaban's $U$ index. Among these, the ponderal index produced no promising correlation (the highest $R$ for the twenty properties being about 0.300) but, of course, there are other properties and other molecules, and the possibility of combining this index with other that may eventually show some use for this descriptor. However, while indices like ponderal, and possibly a few others, have yet to prove themselves, we see that connectivity indices and combinations and a few others, like $W$ and $Z$, have already proved themselves ! Of particular significance is the fact that, while ${}^1\chi$ and ${}^2\chi$ are somewhat parallel to one another, ${}^3\chi$ is complementary in that it shows good correlations where the other two fail and *vice versa*. This signifies that the indices have captured distinctive structural features and this is the underlying reason why the regressions based on the connectivity indices have been so successful.

## THE BEST DESCRIPTORS

From the abundance of regression data it is hard to single out the best descriptor, or even the best descriptors since, for different properties, different descriptors appear as optimal. It is even difficult to suggest which indices are irrelevant since, by considering various combinations of apparently »uninteresting« descriptors a promising new descriptor may emerge. We have seen this in the case of $W/Z$ (and the same may be true of other cases) that even when one of the components performs poorly, if considered isolated (for the property under review), when combined with another descriptor, we obtain a novel variable which can outperform other descriptors. Our motivation for considering $W/Z$ and related combinations like ID/$\chi$ came from the attempts to interpret such variables as »diameter« of a molecule. Similarly, the difference ${}^1\chi - {}^2\chi$ was motivated by the well established properties of the differences in path numbers $P_2$–$P_3$.

It is possible that systematic explorations of several simple combinations of a selection of indices already considered here may produce even better regressions and better single variables. Rather than following along such a direction (which deserves attention), we would like to point to two distinct routes for the construction of novel descriptors that may be promising.

(1) Search for an optimal functional form,

(2) Search for optimal compacted descriptors,

(3) Search for novel graph matrices as a source of new descriptors.

*Search for an optimal functional form:*

To illustrate the search for optimal descriptors, consider several of the regression results based on $\chi^{[1]}$, $\chi^{[1/2]}$, $\chi^{[-1/2]}$, and $\chi^{[-1]}$, shown in Table VIII. We have grouped these into five classes, depending on the way the standard error behaves under a change in the exponential parameter $p$, which takes values 1, 1/2, −1/2 and −1. The case $p=1$ corresponds to an earlier index of the Zagreb group, the value −1/2 represents the connectivity index, and the other two values were introduced by Altenburg in a discussion of the molecular mean radius.[9] The first class (the most popular) shows a steady decrease in $S$ as $p$ decreases, so the domain of the exponents has to be extended in order to find the optimal value. In the second class, the trend is just opposite and the optimal exponent $p$ is to be sought at positive values of $p$. The next class of compounds show a minimal standard error about $p=1/2$ (although the precise value of $m$ has to be determined for each). Finally, the last class in the upper part of the table includes properties that show a minimum for $p = -1/2$ (or the value to be yet more precisely determined). In the lower part of Table VIII, we have a class for itself which is characterized by exhibiting a maximum (not minimum) at $p=1/2$. Optimal $p$ in this case ought to be sought at either side of the range of the $m$ values considered.

TABLE VIII

*Comparison of the statistical data for the functionally tested indices suggestive that the optimal functional form has yet to be established. The trend of* S *(standard error) allows one to classify the properties accordingly*

| Property | $\chi^{[1]}$ | $\chi^{[1/2]}$ | $\chi^{[-1/2]}$ | $\chi^{[-1]}$ |
|---|---|---|---|---|
| BP | 5.46 | 5.01 | 3.60 | 3.25 |
| HF | 1.08 | 0.99 | 0.68 | 0.59 |
| $\Delta_f H$ | 4.53 | 4.14 | 2.85 | 2.45 |
| CT | 10.01 | 9.88 | 8.70 | 7.89 |
| $\Delta_1^g H$ | 3.75 | 3.59 | 2.78 | 2.40 |
| $\sigma$ | .905 | .898 | .751 | .643 |
| CS | 9.11 | 18.72 | 15.65 | 13.55 |
| $p_c$ | 1.29 | 1.33 | 1.47 | 1.48 |
| $V_c$ | 11.03 | 11.97 | 15.19 | 16.13 |
| $S$ | 1.57 | 1.40 | 1.97 | 2.60 |
| AC | .0062 | .0039 | .0162 | .0222 |
| $R^2$ | .109 | .102 | .119 | .136 |
| $N$ | 10.22 | 8.89 | 16.05 | 20.11 |
| $\Delta H$ | .231 | .187 | .113 | .150 |
| $\Delta_v H$ | .972 | .822 | .662 | .828 |
| HV | .995 | .808 | .446 | .627 |
| $V_m$ | 2.08 | 2.32 | 2.63 | 2.58 |
| $R_m$ | 1.40 | .159 | .188 | .187 |
| $\rho$ | .0095 | .0108 | .0123 | .0118 |

Clearly, Table VIII is suggestive of a novel generalization of the connectivity type indices. In another study, Randić, Hansen and Jurs[26] investigated the role of exponent $p$ on the regression in smaller alkanes and observed that linearity of regressions can be improved by relaxing exponent $p$ somewhat from the preselected value of $-1/2$, which defines the connectivity index. Even if the optimal values of $p$ obtained in such search are still less successful than some of the alternative descriptors of Tables III – VI, we see how less than »perfect« descriptors can, nevertheless, be of use, as illustrated here on the outlined classification of molecular properties.

*Search for optimal compacted descriptors:*

Another procedure for deriving better descriptors is to construct novel descriptors by combining the existing descriptors in an orthogonalization process in which one compacts information from two (and more) descriptors into a single descriptor. In this respect, of potential interest are descriptors which did not perform well, including those that show almost zero correlation coefficients. If such a descriptor ($Y$) correlates with another, a promising descriptor ($X$), then one can extract from $X$ the »undesirable« component (the part that parallels $Y$) and obtain a descriptor that exceeds the original descriptor $X$. We illustrate this on $W$ and $Z$ which are correlated and the residuals shown in Table IX. The residual of $W$ against $Z$ are the parts of $W$ that do not correlate with $Z$. Even though the regression between $Z$ and $W$ is not particularly high, we are able, by using the derived residuals as a new descriptor, to compact properties of two descriptors into a single descriptor. Only that part of $W$ is retained which has no parallel with $Z$. When we apply this new descriptor, we find few acceptable regressions (the first columns in Table X). For comparison, we show $R$ and $S$ values for $W/Z$, $W$, and $Z$, respectively, in the remaining columns. An impressive improvement was achieved by the orthogonalization and by extraction of »undesirable«

TABLE IX

*Regression of* W *against* Z *and the residuals of such regression viewed as novel (»compacted«) descriptors*

| Isomer | $W$ | $Z$ | $W/Z$ | Residual |
|---|---|---|---|---|
| $n$ | 84 | 34 | 2.4706 | +0.2199 |
| 2M | 79 | 29 | 2.7241 | –2.1639 |
| 3M | 76 | 31 | 2.4516 | +1.4058 |
| 4M | 75 | 30 | 2.5000 | +0.9290 |
| 3E | 72 | 32 | 2.2500 | +4.4988 |
| 22MM | 71 | 23 | 3.0870 | –3.9780 |
| 23MM | 70 | 27 | 2.5926 | +0.5452 |
| 24MM | 71 | 26 | 2.7308 | –0.9780 |
| 25MM | 74 | 25 | 2.9600 | –3.5477 |
| 33MM | 67 | 25 | 2.6800 | +0.1149 |
| 34MM | 68 | 29 | 2.3448 | +3.5917 |
| 23ME | 67 | 28 | 2.3929 | +3.1149 |
| 33ME | 64 | 28 | 2.2867 | +4.6847 |
| 223MMM | 63 | 22 | 2.8636 | –0.7921 |
| 224MMM | 66 | 19 | 3.4737 | –5.3618 |
| 233MMM | 62 | 23 | 2.6957 | +0.7311 |
| 2234MMM | 65 | 24 | 2.7083 | +0.1614 |
| 2233MMMM | 58 | 17 | 3.4118 | –3.1759 |

TABLE X

*Comparison of the statistical data for single variable regressions based on the Wiener number* W, *Hosoya topological index* Z *and indices derived from them by simple division and by orthogonalization*

| Descriptor | $W\,Z$ | | $W/Z$ | |
|---|---|---|---|---|
| Property | $R$ | $S$ | $R$ | $S$ |
| $V_m$ | 0.931 | 0.958 | 0.636 | 2.032 |
| $R_m$ | 0.963 | 0.0512 | 0.588 | 0.1527 |
| $\rho$ | 0.949 | 0.0039 | 0.670 | 0.0091 |
| Descriptor | $W$ | | $Z$ | |
| Property | $R$ | $S$ | $R$ | $S$ |
| $V_m b$ | 0.592 | 2.12 | 0.112 | 2.032 |
| $R_m$ | 0.669 | 0.140 | 0.036 | 0.189 |
| $\rho$ | 0.641 | 0.0094 | 0.206 | 0.0120 |

components of $Z$ in $W$. The standard error was reduced two to three times, and arrived at acceptable regressions termed »good« and »very good«. We refer to the above procedure as »compacting« of descriptors. Generally, in this way, by »discarding« parts of a descriptor that parallel (highly correlate with) descriptors that are not suitable for the property considered, one can arrive at better descriptors. The procedure illustrates how use is made of descriptors that have not been found suitable in a particular regression. For more details, the reader is directed to a detailed illustration of the construction of compacted descriptors for clonidine-like compounds.[46]

It is beyond the scope of the present paper to pursue both of the two mentioned strategies and to try to anticipate the outcomes of such constructions, but it seems desirable to examine these and other routes of improving the performance of single descriptors in an effort to arrive at optimal single descriptors for regression analysis.

## *Search for novel graph matrices as a source of new descriptors*

Besides the widely known »old« graph matrices, the adjacency and the distance matrix, several novel matrices have been recently introduced in the chemical graph theory. Kier and Hall's electrotopological state matrix has already been mentioned.[38] If all elements except those adjacent in such a matrix are assumed zero, we obtain a weighted adjacency matrix, the higher powers of which can generate an infinite sequence of graph invariants.[47] We have also mentioned the Wiener matrix.[37,48] For cyclic graphs matrix using (uniform electrical) resistances appears of considerable interest.[49] Another recently introduced matrix for trees (but equally applies to cyclic graphs) is based on restricted random walks over graphs.[50] From such matrices, novel invariants can be constructed using the already familiar procedures. Thus, in analogy to Balabans's $J$ index, the Weiner matrix yields the $K$ index by using row sums and the reciprocal square root procedure. The $K$ index appears even more discriminatory than the $J$ index.[51] Alternatively, one may combine matrix elements corresponding to paths of equal length and, in this way, arrive at sequences that may be viewed as generalized path numbers, and a single number that is obtained by adding all entries in a matrix, or all elements above the main diagonal (analogous to $W$ which can thus

be derived from the distance matrix). Some properties of such sequences for the Wiener matrix have been reported.[37]

## CONCLUDING REMARKS

An important step in arriving at multiple regressions with a high correlation coefficient and small relative error is a search for the best one-descriptor regressions. We have examined several hundreds of regressions, all for the same set of 18 isomers of octanes, using some 40 molecular descriptors. It appears that less than half a dozen descriptors again and again dominate simple regressions for the molecular properties. This finding has been confirmed in other studies too and, in particular, Katritzky and Gordeeva[52] have recently shown that »for the estimation of physico-chemical properties, the best small regression models with 1–4 parameters are mainly comprised of »classical« topological indices, such as the Randić index, Wiener index and Molecular Connectivity indices. For the correlation of biological activity, combinations of topological indices with geometrical descriptors have produced regression models of the best quality«. This recent work thus supports our expectation that for structure-activity, 3-dimensional aspects of the molecular structure will also play an important role.

For the first time, it has been shown that upper bounds to the standard error can be derived for physico-chemical properties using statistical information on many regressions (including also regressions which, if considered isolated, are of no practical importance). We have seen that some properties are »difficult« and offer a challenge to structure-property regression analysis, while some other properties appear »promiscuous«, *i.e.*, almost any reasonable descriptor may be used with considerable success to obtain a reasonable regression. This is important to know in order to curb proliferation of unwarranted descriptors, and eliminate descriptors which only apply to »easy« properties. A novel descriptor should be tested on »difficult« properties, properties that evade successful regressions. Only such selective descriptors are likely to tell us something specific about a particular property. It is an »open« question if, for any physico-chemical molecular property, single dominant descriptor could be found, a descriptor with a simple and direct structural interpretation. It is also questionable if such a descriptor is unique, *i.e.*, well separated (by the evaluated standard error $S$) from other descriptors. There may be such single-variables for all properties but we have not been clever enough to find them. On the other hand, there is no guarantee that a single simple (as defined above) descriptor ought to exist for every bulk molecular property.

We suggested three directions to improve the present pool of descriptors. One is based on the search for an optimal mathematical form for a descriptor, using the minimal standard error as the criterion in analogy with the similar procedure recently outlined in the search for optimal descriptors for heteroatom.[3] The second route is that using the orthogonalization process to eliminate parts of a descriptor that are less important in a particular correlation by making the descriptor orthogonal to the descriptors expected or known to be of no interest in such a case. This has been illustrated on the $W$ and $Z$ indices and volume-dependent molecular properties. The third route in a way requires imagination, and as such may be very productive. This is well illustrated by the restricted random walks matrices from which a bond additive index $P_1$ was derived (by adding all entries in the matrix corresponding to adjacent vertices, that is bonds). This novel index was tested on molecular entropy $S^o$ and a very good single variable regression was obtained[48] with $R = 0.964$ (and $S = 1.265$), which is

better than any of the entropy correlations shown in Tables III – VI ! Indeed, by this single demonstration of a superior regression, matrices and descriptors based on restricted random walks deserve serious attention.

Last but not least, an important advantage of comparative studies of regressions is the possibility to classify properties (and descriptors), as it has been illustrated in Table VIII of this paper.

## APPENDIX

In the following tables, we list the numerical values for a dozen descriptors (topological indices) which have not been previously reported in the literature. Here,

APPENDIX

| Isomer | $\chi(\boldsymbol{D})$ | ID($\boldsymbol{D}$) | $\chi(\boldsymbol{W})$ | $\chi(\boldsymbol{W})$ | $X(W)$ | ID($\chi$) |
|---|---|---|---|---|---|---|
| | 3 + | 19 + | 3 + | 17 + | | |
| octane | .94464 | .44445 | .88567 | 2.21975 | 57.1698 | 9.49262 |
| 2M | .94940 | .49068 | .81037 | 1.83728 | 52.6122 | 9.56895 |
| 3M | .94511 | .50645 | .81100 | 1.78681 | 48.4059 | 9.60493 |
| 4M | .94199 | .50944 | .81660 | 1.79520 | 46.6606 | 9.61645 |
| 3E | .94022 | .53958 | .83376 | 1.79155 | 42.2041 | 9.68228 |
| 22MM | .94797 | .54054 | .72688 | 1.38827 | 44.4713 | 9.74313 |
| 23MM | .94716 | .56752 | .74978 | 1.45920 | 42.0581 | 9.69478 |
| 24MM | .95197 | .52466 | .74701 | 1.46078 | 43.4185 | 9.68228 |
| 25MM | .95634 | .55429 | .74196 | 1.48887 | 47.7238 | 9.64563 |
| 33MM | .94353 | .58086 | .74591 | 1.38667 | 38.5332 | 9.80516 |
| 34MM | .94728 | .59479 | .75434 | 1.43762 | 39.2901 | 9.72050 |
| 23ME | .94767 | .61769 | .77287 | 1.48284 | 37.4277 | 9.76237 |
| 33ME | .94444 | .63249 | .76877 | 1.40393 | 34.1415 | 9.86396 |
| 223MMM | .95699 | .66226 | .68266 | 1.09841 | 34.9935 | 9.85984 |
| 224MMM | .95737 | .62957 | .67109 | 1.10112 | 39.1411 | 9.82049 |
| 233MMM | .95144 | .66756 | .69353 | 1.11824 | 33.4679 | 9.88327 |
| 234MMM | .95505 | .64875 | .69288 | 1.16965 | 37.0246 | 9.77343 |
| 2233MMM | .96052 | .73834 | .62117 | 0.80549 | 30.3305 | 10.0000 |
| | P′/P | $W'/W$ | $W'$ | WW | $\chi'/\chi$ | $\chi'$ |
| octane | 4.0000 | 3.00000 | 252 | 210 | 6.67876 | 26.1420 |
| 2M | 4.1786 | 3.40506 | 269 | 184 | 6.77107 | 25.5276 |
| 3M | 4.2857 | 3.52632 | 268 | 170 | 6.62467 | 25.2274 |
| 4M | 4.3214 | 3.60000 | 270 | 165 | 6.62467 | 25.2274 |
| 3E | 4.4286 | 3.83333 | 276 | 150 | 6.61116 | 25.4272 |
| 22MM | 4.4643 | 3.80282 | 270 | 149 | 6.59764 | 23.4922 |
| 23MM | 4.5000 | 3.91429 | 274 | 143 | 6.58092 | 24.2224 |
| 24MM | 4.4643 | 3.95775 | 281 | 147 | 6.58121 | 24.1129 |
| 25MM | 4.3571 | 3.54054 | 262 | 161 | 6.59511 | 23.9132 |
| 33MM | 4.6071 | 4.08955 | 274 | 131 | 6.57590 | 23.8133 |
| 34MM | 4.4514 | 4.05882 | 276 | 134 | 6.58784 | 24.4982 |
| 23ME | 4.6071 | 4.14925 | 278 | 129 | 6.56740 | 24.4222 |
| 33ME | 4.7143 | 4.40625 | 282 | 118 | 6.55467 | 24.1343 |
| 223MMM | 4.7500 | 4.34921 | 274 | 115 | 6.92158 | 24.2532 |
| 224MMM | 4.6429 | 4.15152 | 274 | 127 | 7.02224 | 23.9915 |
| 233MMM | 4.6429 | 4.48287 | 278 | 111 | 6.52808 | 22.8744 |
| 234MMM | 4.6786 | 4.24615 | 276 | 122 | 6.53383 | 23.2173 |
| 2233MMM | 4.9286 | 4.65517 | 270 | 97 | 6.49988 | 21.1246 |

***D*** represents the distance matrix and ***W*** the Wiener matrix of a graph, from which invariants $\chi$ and ID were derived using the WEIGHTED PATH program. The hyper-Wiener number is derived by adding up all the entries in the Wiener matrix, the index being analogous to the Wiener number, which represents the sum of all entries of the distance matrix. Kier and Hall's[38] total topological index $\tau$ is similarly constructed from the so called topological state matrix. The topological state value $S_i$ for vertex $i$ is analogous to the atomic sums of *W*.

## REFERENCES


1. M. Randić and N. Trinajstić, *J. Mol. Struct. (Theochem)* in press; M. Randić, *Theor. Chim. Acta* (submitted).
2. M. Randić, *New J. Chem.* **15** (1991) 517; M. Randić, *J. Chem. Inf. Comput. Sci.* **31** (1991) 311; M. Randić, *Croat. Chem. Acta* **64** (1991) 43; M. Randić, *J. Mol. Struct. (Theochem)* **223** (1991) 45.
3. M. Randić, *J. Comput. Chem.* **12** (1991) 970; M. Randić and J. Cz. Dobrowolski, *J. Math. Chem.* (submitted); M. Randić and S. Basak, *ibid.* (submitted).
4. D. E. Needham, I-C. Wei, and P. G. Seybold, *J. Amer. Chem. Soc.* **110** (1988) 4186.
5. D. W. Scott, *J. Chem. Phys.* **60** (1974) 3144.
6. L. B. Kier and L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic press, New York, 1976.
7. M. Garbalena and W. C. Herndon, *J. Chem. Inf. Comput. Sci.* **32** (1992) 37.
8. A. T. Balaban and V. Feroiu, *Rep. Mol. Theory* **1** (1990) 133.
9. K. Altenburg, *Z. Phys. Chem. (Leipzig)* **261** (1980) 389.
10. S. S. Tratch, M. I. Stankevitch, and N. S. Zefirov, *J. Comput. Chem.* **11** (1990) 899.
11. A. T. Balaban and I. Motoc, *MATCH* **5** (1979) 197.
12. M. Randić, *J. Magn. Res.* **39** (1980) 431.
13. G. Castello and G. D'Amato, *J. Chrom.* **107** (1975) 1; E. sz. Kovats, *Z. Anal. Chem.* **181** (1961) 351.
14. M. Randić, *J. Amer. Chem. Soc.* **97** (1975) 6609.
15. H. Narumi and H. Hosoya, *Bull. Chem. Soc. Japan* **58** (1985) 1778.
16. M. Randić, *J. Chem. Inf. Comput. Sci.* **24** (1984) 164.
17. D. Bonchev and N. Trinajstić, *J. Chem. Phys.* **67** (1977) 4517.
18. A. T. Balaban, *Phys. Theor. Chem.* **51** (1987) 159.
19. Z. Mihalić, S. Nikolić, and N. Trinajstić, *J. Chem. Inf. Comput. Sci.* **32** (1992) 28.
20. M. Randić, *J. Math. Chem.* **7** (1991) 155.
21. M. Randić and W. L. Woodworth, *MATCH* **13** (1982) 291.
22. Lu Xu, Preprint, (Changchun Institute of Applied Chemistry, 1992).
23. A. T. Balaban, *Theor. Chim. Acta* **53** (1979) 355.
24. I. Gutman and N. Trinajstić, *Chem. Phys. Lett.* **17** (1972) 535; I. Gutman, B. Ruščić, N. Trinajstić, and C. F. Wilcox, Jr., *J. Chem. Phys.* **62** (1975) 3339.
25. A. T. Balaban, I. Motoc, D. Bonchev, and Ov. Mekenyan, *Topics Curr. Chem.* **114** (1983) 21.
26. M. Randić, P. J. Hansen, and PO. C. Jurs, *J. Chem. Inf. Comput. Sci.* **28** (1988) 60.
27. P. C. Jurs, *Computer Software Applications in Chemistry*, Wiley, New York, 1986.
28. M. Randić, G. M. Brissey, R. B. Spencer, and C. L. Wilkins, *Computers & Chemistry*, **3** (1979) 5.
29. M. Randić, G. M. Brissey, R. B. Spencer, and C. L. Wilkins, *ibid.* **4** (1980) 27.
30. L. Lovasz and J. Pelikan, *Period. Math. Hung.* **3** (1973) 175.
31. C. Shannon and W. Weaver, *Mathematical Theory of Communication*, University of Illinois Press, Urbana, 1949.
32. A. T. Balaban, *Chem. Phys. Lett.* **89** (1982) 399.
33. M. Randić, *J. Chem. Inf. Comput. Sci.* **32** (1992) 57,

34. M. Randić, WEIGHTED PATHS are computed using (unpublished) modification of ALL PATH program (Ref. 27, 28) in which paths are weighted by $(m\ n)^{-1/2}$ (same as the connectivity index).
35. H. Hosoya, *Bull. Chem. Soc. Japan* **44** (1971) 2332.
36. H. Weiner, *J. Amer. Chem. Soc.* **69** (1947) 17.
37. M. Randić, T,. Oxley, Xiaofeng Guo and P. Krishnapriyan, *J. Math. Chem.* (to be submitted).
38. L. H. Hall, *Computational Chemical Graph Theory* D. H. Rouvray, Ed., Nova Sci. Publ., Commack, NY, 1990, p. 201.
39. M. Randić and N. Trinajstić, *Theor. Chim. Acta* **73** (1988) 233; Y. Miyashita, H. Ohsako, T. Okuyama, S. Sasaki, and M. Randić, *Magn. Res. Chem.* **29** (1991) 362.
40. D. Bonchev, Ov. Mekenyan, and N. Trinajstić, *J. Comput. Chem.* **2** (1981) 127.
41. M. Randić, *Stud. Phys. Theor. Chem.* **54** (1988) 101; M. Randić, *Int. J. Quant. Chem., Quant. Biol. Symp.* **15** (1988) 201; M. Randić, B. Jerman-Blažić, and N. Trinajstić, *Comput. Chem.* **14** (1990) 237; K. Balasubramanian, *Chem. Phys. Lett.* **169** (1990) 224.
42. M. Randić, A. F. Kleiner, and L. de Alba, *Properties of novel matrices for graphs embedded on regular 2-dimensional and 3-dimensional lattices* (to be reported at the 5-th Int. Conference of Mathematical Chemistry, Kansas City, (Missouri), May 1993).
43. L. B. Kier, *Computational Chemical Graph Theory,* D. H. Rouvray, Ed., Nova Sci. Publ., Commack, NY, 1990, p. 151.
44. M. Randić, *J. Comput. Chem.*
45. J. R. Platt, *J. Phys. Chem.* **56** (1952) 328.
46. M. Randić, *Croat. Chem. Acta* (submitted).
47. M. Randić, *J. Chem. Inf. Comput. Sci.* **32** (1992) 686.
48. M. Randić, *Chem. Phys. Lett.*, (submitted).
49. D. J. Klein and M. Randić, *J. Math. Chem.* **12** (1993) 81.
50. M. Randić, *Chem. Phys. Lett.* (submitted).
51. M. Randić and Xiaofeng, *J. Math. Chem.* (to be submitted)
52. E. V Gordeeva, A. R. Katritzky, V. V. Shcherbukhin, and N. S. Zefirov, *J. Chem. Inf. Comput. Sci.* **33** (1993) 102.

## SAŽETAK

### Komparativna regresijska analiza. Regresije temeljene na jednom deskriptoru

*Milan Randić*

Prikazani su rezultati dobiveni za 750 linearnih regresija koje se temelje na jednom deskriptoru koristeći dvadeset molekularnih svojstava za alkane i četrdeset različitih molekularnih deskriptora. Budući da je traženje grafičkih invarijanata dalo dobre rezultate, dobiveno je dosta novih karakteristika za izvedene deskriptore koji su pokazali različita ponašanja u korelacijama s fizičko-kemijskim svojstvima. Istovremeno, pronađeno je da se neka svojstva dobro slažu s jednim deskriptorom u regresijskoj analizi, ali još uvijek postoje svojstva koja je teško karakterizirati deskriptorom. Na kraju članka dan je popis problema koje bi trebalo razmotriti pri proučavanju odnosa strukture, svojstava i aktivnosti.