# Traditional Topological Indices *vs* Electronic, Geometrical, and Combined Molecular Descriptors in QSAR/QSPR Research

Alan R. Katritzky* and Ekaterina V. Gordeeva[‡]

Department of Chemistry, University of Florida, Gainesville, Florida 32611-2046

Received January 27, 1993[®]

A comparison of the performance of molecular descriptors of different types was conducted. The study was concentrated on determining which descriptors are included in the best linear multivariate regression models constructed for modeling various physicochemical properties (melting point, boiling point, refractive index, molar volume, and density) and biological activities (anaesthetic activity, narcotic activity, sweetness intensity). A total of 84 molecular descriptors published over the past 2 decades was included in this study, plus some of their normalized and squared forms. It was shown that, for the estimation of physicochemical properties, the best small regression models with one to four 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 produced regression models of the best quality. In-house developed software was used for generation of the molecular descriptors (the GROUND program) and for the statistical QSAR/QSPR analysis (the GROUNDSTAT program).

## INTRODUCTION

The search for regularities in the manner in which various molecular properties change, depending upon their molecular structures, is the main subject of investigations in the field of quantitative structure–activity (structure–property) relationships (QSAR/QSPR). The solution of this problem even for a narrow class of compounds is of great importance, since the regularities disclosed could be available for use in the systematization of the behavior of molecules in this set and especially in forecasting the properties of other (sometimes hypothetical) compounds belonging to the given class. Moreover, the relations between structures and properties (or biological activities) thus revealed could be important in developing a new theory, in explaining the phenomena observed, or in illuminating the mechanisms of physical or chemical properties or biological activities of the compounds.

To find quantitative relations between structure and properties, various methods of mathematical modeling have been applied. In this way, QSAR/QSPR investigations have opened a new page in the history of computer applications in chemistry and created extensive software dealing with the search for structure–activity/property relationships during the past 2 decades.[1,2]

The major goal of any QSAR/QSPR research is to assign to the structure a number or a set of numbers which (i) must correlate well with the property (activity) value measured experimentally and, if possible, (ii) should provide some physical insight to the molecular behavior. This numerical representation of the structure which describes the structure as a whole is usually called a molecular descriptor. Molecular descriptors can be derived in either empirical or nonempirical ways.

The empirical molecular descriptors are usually those experimental data which are more easily accessible than the values of the property being studied. For example, such data as octanol–water partition coefficient,[3–5] molecular refraction,[6] ionization potential, and molecular polarizability[7] were used as empirical descriptors of the whole molecule. In other words, the results of more expensive or time-consuming tests can be sometimes predicted from such less expensive and fast measurements. Although descriptors of this type are not derived from the structure directly, they may still be useful and complementary in combination with other types of descriptors.

Next to the famous and widely used Hansch approach,[8] the most developed QSAR/QSPR techniques now involve the calculation of various nonempirical molecular descriptors. In many cases, the use of nonempirical molecular descriptors proved to possess advantages over the Hansch approach.[9] The majority of structure sets are noncongeneric (do not have a common basic skeleton), and for these compounds the Hansch approach is not applicable at all. For such data sets the use of techniques based on calculation of nonempirical molecular descriptors, with subsequent correlation of those descriptors vs experimental data, is the only feasible way to establish the structure–property (activity) relationships.

Attempts to describe chemical compounds by a single numerical value arose concurrently with the structural theory of organic compounds. The relationship between the number of atoms and boiling points observed in the series of *n*-alkanes[10] can be perhaps considered as one of the first QSPR in organic chemistry. Correlations on the set of *n*-alkanes and some other sets of simple compounds[11] were sufficiently successful to encourage various fascinating but more complicated approaches to the problem of structure to number conversion. For several years, attention was mainly concentrated on the so-called topological indices which are molecular descriptors derived from information on connectivity and composition of a molecule. The creation, testing and theoretical explanation of new topological indices brought with it a considerable intervention of graph theory and discrete mathematics into organic chemistry.[9,12]

With the rapid development of both instrumental and theoretical methods for the analysis and description of molecular structure, additional geometrical and electronic parameters became available for inclusion into new and more complicated molecular descriptors. Intuitively, a purely topological representation of structure seemed to be insufficient in many cases.

[‡] Present address: Molecular Design Ltd., 2132 Farallon Dr., San Leandro, CA 94577.

[®] Abstract published in *Advance ACS Abstracts*, September 1, 1993.

0095-2338/93/1633-0835$04.00/0

So far, over a hundred different types of nonempirical molecular descriptors have been developed; if one counts their numerous modifications, then the total number of molecular descriptors now available to chemists amounts to a few hundred. If the purpose of the QSAR/QSPR research is not solely to obtain a high $R^2$ (modern statistical manipulation makes this possible for almost all data sets) but also to obtain a close insight into the molecular behavior and property performance, then one needs guidelines and navigation aids through this ocean of existing, developing, and continuously emerging molecular descriptors. One attempt to investigate and compare a performance of different molecular descriptors has been reported.[13] However, the study conducted by Basak and co-workers[13] was related exclusively to the prediction of the log *P* value within a diverse set of 382 compounds. Also, in that study the topological descriptors were mixed with a specific hydrogen bonding parameter[14] $HB_1$ which makes the analysis less clear.

While making a choice of descriptors to be included in the regression model, the chemist has to answer fundamental questions. When dealing with the large set of molecular descriptors available, one has to decide whether to select manually certain types of descriptors or to rely on a selection using statistical principles. Also, it is hard to predict if a correlation could be improved with some new descriptors added or with modifications (squared, logarithmic, inverse forms) of descriptors from the current set.

We believe this is an appropriate time to consider more closely the actual performance of those numerous descriptors in QSAR/QSPR research. We hope that such an analysis will provide a guideline for chemists who intend to apply QSAR/QSPR techniques in their own research.

In this work, we present a comparison of the performance of geometrical and electronic descriptors vs the traditional topological indices. The literature conclusions regarding the usefulness of these two types of descriptors are not clear. For example, Kaliszan and Osmialowski expressed a strong preference in favor of electronic structure parameters used as molecular descriptors but against topological indices[15,16] for the correlations between structure and chromatographic retention time. Kier and Hall impressed the scientific world with amazingly good correlations between their connectivity indices and various molecular properties.[17,18] After 15 years, the strong single-parameter correlation between ${}^1\chi$ and the minimum blocking concentration (log MBC) on a very diverse set of compounds[17] still attracts the attention of QSAR researchers.[19] In a series of papers, Jurs and co-workers showed that geometrical and electronic descriptors, separately or in combination with topological indices, provide very good correlations between structures and different biological and physical–chemical properties, such as sweetness,[20] chromatographic retention indices,[21] surface tension,[22] and normal boiling points.[23] In contrast, Balaban and co-workers[24] predicted normal boiling points even more precisely, with fewer parameters and on a larger set of compounds, using only topological indices and counts of atoms.

Our goal was to compare in the most objective manner possible the correlations obtained on the same data sets using topological, geometrical, electronic, and mixed (combined) descriptors—separately and in combination. Over the past several years we have had significant experience in the application of QSAR/QSPR methods to industrial research, and sharing some of the generalizations resulting from this experience should help chemists to judge better the value and usefulness of the various kinds of nonempirical molecular descriptors. Many of the descriptors published over the past 2 decades have been incorporated into our program GROUND (Generation and Recollection Of Updated Nonempirical Descriptors) and thoroughly tested on various data sets, both diverse and homogeneous. In the present paper, we hope to provide an unbiased comparison of the statistical performance of different kinds of descriptors, both separately and in combination. To illustrate the behavior of descriptors, we use many different (all previously published) descriptors calculated for several diverse data sets from the literature.

## DESCRIPTORS INCLUDED IN STUDY

We divided the whole world of molecular descriptors into two basic sets: topological indices vs electronic, geometrical, and combined descriptors. Descriptors derived solely from connectivity and composition of the structure we designated as topological indices. Descriptors derived purely from 3D molecular geometry or partial charge distribution were designated respectively as geometrical or electronic ones. Those descriptors which can be calculated only by using simultaneously information on connectivity and electronic structure, or information on 3D molecular geometry and electronic structure, were designated as combined descriptors. From the numerous descriptors published in the literature only those which have previously been shown to participate successfully in QSAR/QSPR correlations were selected. We next present a list of descriptors included in our study, along with the descriptor designations used hereafter in this paper, and brief comments for those descriptors which were not previously reviewed thoroughly.

**Topological Indices.** These types of molecular descriptors have been highly developed from a theoretical point of view during the past 4 decades since Wiener's pioneering paper in 1947, and they have been used extensively in various QSAR studies (see the reviews of successful applications of topological indices in refs 17 and 25–28). The most popular and widely used topological indices are as follows: W, Wiener index;[29] ${}_1\chi$, Randić index;[30] ${}^1\chi^v$, Kier and Hall valence connectivity index;[17,18] ${}^1\kappa$, Kier shape index;[31,32] $\Phi$, Kier flexibility index;[32] *J*, Balaban index of average distance sum connectivity.[33]

For the connectivity index $\chi$, the valence connectivity index $\chi^v$, and the shape index $\kappa$ we include the modifications[17,18,34] for paths of lengths 2 and 3, namely, indices ${}^2\chi$, ${}^3\chi$, ${}^2\chi^v$, and ${}^3\chi^v$ and ${}^2\kappa$, ${}^3\kappa$.[31,32] Connectivity indices ${}^0\chi$ and ${}^0\chi^v$ of zero order,[17,18] although more closely related to molecular composition than to molecular connectivity, are also included in the present study.

Balaban's index of average distance sum connectivity was included in its traditional form,[33] without the corrections for the presence of heteroatoms suggested in ref 24. This choice was made because we tried to include the molecular descriptors in their most generic forms, so that our set of descriptors should be applicable to QSAR/QSPR research on various and diverse data sets.

Information content indices, based on the Shannon information theory,[35] can also be designated as topological indices since they can be derived solely from the connectivity and composition of the molecule. These indices have been applied successfully in chemistry (for a review see ref 26). In our study, we have included the so-called entropy of probability distribution (EPD) and information content (IC) along with their modifications designated as structural information content (SIC) and complementary information content (CIC) formulated in papers by Basak and co-workers.[13,36] From our own experience, we found that analogous modifications applied to the entropy of probability distribution, namely, the structural entropy of probability distribution (SEPD) and the complementary entropy of probability distribution (CEPD), can be also useful for QSAR studies. All these indices are related to the diversity of the atoms and the bonds in the molecule. The selection and partition of the atoms into the equivalency classes can be done in different ways. In the present study we include information content indices of four levels of detail.

Zero order: Atoms are distinguished, if and only if, the corresponding atomic symbols are different.

*i*th order ($i = 1$–$3$): Atoms are distinguished if their nearest neighbors are different on the basis of the $(i - 1)$th order partition of atoms (cf. ref 37).

**Electronic Descriptors.** These descriptors reflect the electronic structure of the molecule and, basically, some overall characteristics of the partial charge distribution. The following parameters[38] of the electronic molecular structure are included as descriptors in this study: MNC, the most negative charge ($q^-_{max}$); MPC, the most positive charge ($q^+_{max}$); SPP, submolecular polarity parameter $|q^+_{max} - q^-_{max}|$.

From our experience, we add to this set a sum of absolute values of partial charges (SAPC) as a useful descriptor.

**Geometrical Descriptors.** These descriptors require access to the 3D coordinates of all the atoms in the given molecule. In this work we use three shadow indices[20] designated as $S_1$, $S_2$, and $S_3$ (areas of the molecular shadow projected on the $XY$, $YZ$, and $XZ$ planes, respectively). Each area is normalized by dividing the index by the area of the rectangle defined by the maximum dimensions of the projection on the plane ($S_4 = S_1/X_{max}Y_{max}$, $S_5 = S_2/Y_{max}Z_{max}$, and $S_6 = S_3/X_{max}Z_{max}$).

In the set of geometrical descriptors we have included the so-called shape parameter ($\eta$) defined as the ratio of the largest to smallest of the three dimensions of the box built on the shadows of the given molecule.[39]

The molecular volume (MV) is calculated according to the procedure suggested in ref 40.

Total solvent-accessible surface area (TSASA) is calculated as described in ref 41.

**Combined Descriptors.** These descriptors take into account simultaneously electronic and geometrical, or electronic and topological, parameters of the molecule.

Electronic–topological descriptor ($E^T$) was suggested in ref 42 as an electronic analogy to the valence connectivity index $\chi^v$. The difference between $E^T$ and $\chi^v$ is that the absolute values of partial charges are used instead of $\delta$-values[17,18] in the formulas for ${}^n\chi^v$, and the summation is carried out on the set of all the bonds, including hydrogens. Similarly to $\chi^v$ indices of orders 0–3, index $E^T$ is also calculated for separated atoms (${}^0E^T$) and bonds (${}^1E^T$) and for paths of lengths 2 (${}^2E^T$) and 3 (${}^3E^T$).

The topographic–electronic index $T^E$, connecting submolecular polarity parameters with molecular topography expressed by interatomic distances, was suggested in ref 43. This index is calculated as a sum of ratios $|q_i - q_j|/r_{ij}^2$ over all the pairs of atoms, both connected and disconnected ($q_i$ and $q_j$ are the corresponding partial charges on atoms $i$ and $j$ in the pair, whereas $r_{ij}$ is the interatomic distance). On the basis of our experience, we added to the set of combined descriptors the "connected" modification of the $T^E$ descriptor (${}^cT^E$), which is calculated very similarly to $T^E$ but only takes into account bonded atoms in the summing procedure.

The distance between the atoms bearing the maximum positive and the maximum negative partial charges (DIST) and the descriptor PD calculated as the ratio of submolecular polarity parameter to the squared value of DIST (PD = SPP/DIST$^2$) we also added to the set of combined descriptors in the present study.

The majority of the combined molecular descriptors is formed by 25 descriptors based on the conception of the charged partial surface areas in the molecule.[21] These descriptors showed good correlations with normal boiling points[23] and with surface tension for relatively large sets of compounds.[24] In our study we use the abbreviated names for all 25 descriptors, just as suggested in ref 21.

## METHOD

The fundamental topological and information content indices were collated in the descriptor subset $A$. Subset $A$ includes 38 descriptors: $W$, ${}^0\chi - {}^3\chi$, ${}^0\chi^v - {}^3\chi^v$, ${}^1\kappa - {}^3\kappa$, $\Phi$, ${}^0$IC – ${}^3$IC, ${}^0$EPD – ${}^3$EPD, ${}^0$SIC – ${}^3$SIC, ${}^0$CIC – ${}^3$CIC, ${}^0$SEPD – ${}^3$SEPD, ${}^0$CEPD – ${}^3$CEPD, and $J$.

The fundamental electronic, geometrical, and combined descriptors were collated in the subset $B$ of 46 descriptors: MNC, MPC, SPP, SAPC, $S_1 - S_6$, MV, TSASA, $\eta$, ${}^0E^T - {}^3E^T$, $T^E$, ${}^cT^E$, DIST, PD, PPSA1 – PPSA3, PNSA1 – PNSA3, DPSA1 – DPSA3, FPSA1 – FPSA3, FNSA1 – FNSA3, WPSA1 – WPSA3, WNSA1 – WNSA3, RPCG, RNCG, RPCS, and RNCS.[21]

Combination of subsets $A$ and $B$ makes the joint set of descriptors $C$, comprising 84 descriptors in total. Subset $B$ is significantly larger than subset $A$ with regard to the total number of descriptors included (46 vs 38). To provide a more objective comparison, we created two additional subsets ($D$ and $E$) of modified topological indices. Subset $D$ includes all the descriptors from subset $A$ plus the normalized values of such additive descriptors as $\chi$ and $W$. Descriptors $W$, ${}^0\chi$, and ${}^0\chi^v$ were normalized by dividing the descriptor value by the number of atoms ($N_a$) in the hydrogen-reduced skeleton. Descriptors ${}^1\chi - {}^3\chi$ and ${}^1\chi^v - {}^3\chi^v$ were normalized by dividing on number of bonds ($N_b$) in the hydrogen-reduced skeleton. In total, subset $D$ comprises 47 descriptors. The importance of normalized topological indices in QSAR studies was demonstrated by Jurs and co-workers who found good correlations which included descriptors such as "molecular ID/no. of atoms in molecule"[22,23] and "count of paths 0–45/no. of atoms".[22] Also, as Kaliszan and Osmialowski explicitly pointed out, topological indices are often highly correlated with the molecular weight of the compound or with the total number of atoms (bonds) in the molecule.[15] In our experience, we have also noticed that topological indices in their normalized form frequently help to attain a good and stable regression model.

Subset $E$ comprises 76 topological descriptors, including those from subset $A$ plus each of the descriptors in squared form. In their book,[17] in a chapter entitled "The Use of a Quadratic Expression in Chi", Kier and Hall have demonstrated that the $\chi$ index can be very useful in its squared form, particularly for the prediction of biological activity.[17,18]

A comparative study of the performance of the different types of descriptors was conducted for each data set as follows. Firstly, for a given set of compounds the descriptors from sets $A$–$C$ were calculated by means of the GROUND program. Since the electronic, geometrical, and combined descriptors are dependent on the source of partial charge distribution and on the 3D molecular geometry, the 3D geometry of all structures was optimized using the MOPAC package (AM1 method). Then for each structure 3D Cartesian coordinates and the corresponding partial charge distributions from the MOPAC output were used as the basis for the calculation of the electronic, geometrical, and combined descriptors in the GROUND program.

Secondly, exhaustive searches for the best one-, two-, three- and four-parameter correlations between each set of descriptors (*A*, *B*, *C*, *D*, or *E*) and the experimental property values were conducted using a specific option incorporated into our GROUNDSTAT program (STATistical treatment of the GROUND output vs experimental data sets). The $R^2$ value was used as a criterion for the estimation of correlation quality. Each correlation was subject to a cross-validation check known as the jacknifing procedure (see brief description in refs 18 and 23). Apart from the $R^2$ value, we also estimated, for each of the regression models obtained, the $R^2_{\text{crossval}}$ which is $R^2$ for correlation "observed–predicted" experimental values. The difference between $R^2$ and $R^2_{\text{crossval}}$ is that in the latter case each predicted value is calculated from the regression model obtained on a training set which does not contain that particular observation. Therefore $R^2_{\text{crossval}}$ allows us to estimate the stability of a regression model which should not depend strongly on an individual data point. In other words, the higher and closer the values $R^2$ and $R^2_{\text{crossval}}$ are, the better the regression model obtained.

An exhaustive search for the best five-parameter regression model was impossible because of the enormous requirement for CPU time. Therefore, the "best" five-parameter regression model was generated according to a "4 + 1" scheme. In other words, the best four-parameter model was exhaustively tried with the fifth parameter being taken from the corresponding subset of descriptors (*A*, *B*, *C*, *D*, or *E*), and the resulting five-parameter model with the best $R^2$ value was thus selected.

Although a regression model of only five parameters may appear to be too small to provide the best correlation equation, we believe that consideration of larger regression models could result in confusing or even misleading conclusions, whereas the general regularities in the behavior of descriptors should still be observed in regressions with few parameters. Also, we rationalized that an exhaustive search through all possible one- to five-parameter regressions should give us a more objective picture rather than the application of stepwise regression analysis, where the result sometimes depends on the sequence in which parameters are added or to deleted from the regression model.

In order to improve understanding of the behavior of different subsets of descriptors, we tried to estimate the nonredundancy of descriptors in each subset. We conducted pairwise correlations between all descriptors in each subset, and from each pair with an intercorrelation $R^2$ above 0.9, we dropped off one descriptor (in the GROUNDSTAT program that descriptor is dropped off which provides the lower $R^2$ when correlating with the property in a single parameter model). The descriptors remaining in the set after this selection are hereafter designated as the set of nonredundant descriptors (NRD). In our view, NRD value represents more adequately the choice of descriptors than simply the number of descriptors included in each set, because all highly intercorrelated and therefore somewhat redundant descriptors are excluded. However, the NRD values were used only for general comparison of the descriptor subsets. To generate a correlation equation, an exhaustive search was conducted throughout each whole subset of descriptors; i.e. even highly redundant descriptors were not skipped while searching for the best one- to five-parameter regression models.

## PHYSICOCHEMICAL PROPERTIES

We began our investigation and comparison of the behavior of different types of molecular descriptors used in QSPR research by investigating the success of predictions of five different physicochemical properties, each measured in three different sets of organic structures: aldehydes, amines, and ketones. Data for aldehydes and ketones are taken from ref 44 and for amines from ref 45.

**Table I.** Intercorrelations between Properties in the Sets of Aldehydes, Amines, and Ketones

| | | aldehydes | | amines | | ketones | |
|---|---|---|---|---|---|---|---|
| property 1 | property 2 | *n* | $R^2$ | *n* | $R^2$ | *n* | $R^2$ |
| boiling point | melting point | 27 | 0.645 | 48 | 0.501 | 30 | 0.712 |
| boiling point | density | 36 | 0.386 | 96 | 0.426 | 45 | 0.333 |
| boiling point | molar volume | 36 | 0.466 | 96 | 0.425 | 45 | 0.522 |
| boiling point | refractive index | 35 | 0.765 | 96 | 0.529 | 44 | 0.592 |
| melting point | density | 23 | 0.530 | 43 | 0.289 | 33 | 0.267 |
| melting point | molar volume | 23 | 0.214 | 43 | 0.191 | 33 | 0.332 |
| melting point | refractive index | 23 | 0.638 | 42 | 0.224 | 32 | 0.481 |
| density | molar volume | 59 | 0.025 | 110 | 0.011 | 60 | 0.000 |
| density | refractive index | 58 | 0.593 | 107 | 0.863 | 57 | 0.783 |
| molar volume | refractive index | 58 | 0.040 | 107 | 0.026 | 57 | 0.122 |

The set of aldehydes is quite diverse. It comprises aromatic and aliphatic aldehydes, $\alpha,\beta$-conjugated aldehydes, aldehydes containing chlorine and bromine substituents, and cyano and nitro groups. The set of amines is also very diverse. It combines primary, secondary, and tertiary aliphatic amines with substituted pyridines, pyrazines, piperidines, piperazines, and morpholines. By contrast, the set of ketones is more homogeneous: basically it includes aryl, alkyl, and cycloalkyl ketones.

All five physicochemical properties studied in this work, namely, melting point, boiling point, molar volume, density, and refractive index, are of great importance in academic and industrial research. As shown below, considerable previous effort to model these properties using both QSPR technique and incremental methods has been reported.

The three aforementioned structure sets attracted our attention for three reasons. Firstly, the number of nonmissing data points in each data set and for each property lies in the range 43–72 (aldehydes), 54–110 (amines), and 49–60 (ketones). These are reasonable numbers of data points for modeling a property for a fairly diverse set of structures, using small regression models with up to five parameters.

Secondly, a pairwise correlation between the properties, showed us that no pair within the five properties is highly intercorrelated. The results of intercorrelation analysis are presented in Table I, where each row corresponds to a particular pair of properties (property 1 vs property 2). As seen from the Table, for some pairs of properties, such as for boiling point vs melting point and melting point vs density, the number of common nonmissing values is quite low. This means that the property values available correspond to the subsets of structures which overlap only to a moderate extent. In general, little intercorrelation was observed between the properties, with the exception of density vs refractive index for amines ($R^2$ = 0.863, *n* = 107). In the set of ketones, fair intercorrelations ($R^2$ = 0.712 and $R^2$ = 0.783) exist between boiling point and melting point, and between density and refractive index, but only for a subset of structures (*n* = 30 and *n* = 57, respectively). For the set of aldehydes, some fair correlation ($R^2$ = 0.765) was noticed between boiling point and refractive index, but again within a portion of data (*n* = 35).

Thirdly, these particular data sets were previously treated with the PCA technique,[44,45] and no quantitative model for estimation of property values was presented. Therefore, it seemed interesting to apply the more traditional QSPR technique, based on multivariate regression analysis, in order

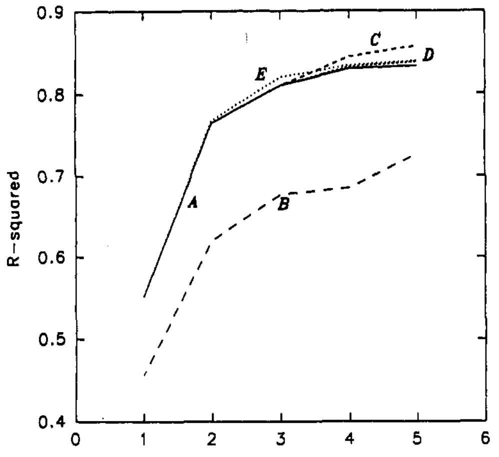

**Figure 1.** Dependence of $R^2$ on model size (structures, aldehydes;[44] property, melting point).

to see how well these five properties could be modeled with a linear regression model which includes only a few molecular descriptors.

**Melting Point.** Melting points have proved to be a challenging property for QSPR research. Even for alkanes, which usually present very "friendly" or "well-behaved" data sets for correlations with their physical–chemical properties (boiling points, densities, molar refractions, heats of evaporation, surface tensions, etc.), it was shown that melting points are not well modeled with topological indices.[11] In both of the books[17,18] authored by Kier and Hall, the numerous examples of successful correlations with $\chi$ indices did not include any regression models derived for the prediction of melting points. Apparently, crystal packing effects and the associated intermolecular attractive forces cannot be adequately described with the existing topological and ad hoc descriptors.[11] Therefore, it was particularly interesting to see if melting points can be treated more satisfactorily with electronic, geometrical, and combined descriptors used as parameters of regression models.

Here we consider the regression models of one to five parameters obtained for each subset of descriptors from *A* to *E* on three data sets: aldehydes ($n$ = 72), amines ($n$ = 54), and ketones ($n$ = 52).

*Aldehydes* (*Figure 1*). Subset *A* of topological descriptors (NRD = 23) provides a poor one-parameter correlation, a fair two-parameter correlation, and reasonably good three-, four-, and five-parameter regression models. Increasing the number of independent variables from 3 to 5 gradually improves both fit and stability of the regression models. The descriptor showing the best, although quite poor, correlation ($R^2$ = 0.552 and $R^2_{\text{crossval}}$ = 0.524) in the single parameter regression model is structural information content of first order ($^1$SIC). A combination of two other information indices ($^1$EPD and $^2$SEPD) gives the best possible two-parameter correlation in subset *A* ($R^2$ = 0.765, $R^2_{\text{crossval}}$ = 0.745). A significant jump of $R^2$ is observed when the correlation was allowed to combine the information indices with connectivity descriptors. The best three-parameter equation ($R^2$ = 0.810, $R^2_{\text{crossval}}$ = 0.781) is based on a combination of Randić index $^0\chi$ with two information indices ($^1$SIC and $^2$IC). The best four-parameter model ($R^2$ = 0.830, $R^2_{\text{crossval}}$ = 0.801) includes, in addition, another information index $^0$SEPD. An attachment of the fifth independent variable $^1\chi^v$ to the model improves the fit but at the same time decreases the stability: $R^2$ = 0.833, $R^2_{\text{crossval}}$ = 0.795.

Subset *B* of electronic, geometrical, and combined descriptors shows only poor correlations for regression models with up to four descriptors included: $R^2$ values belong to the range 0.457–0.685. Even addition of a fifth descriptor provides only fair correlation with $R^2$ of 0.727 (cf.: subset *A* provided a correlation with $R^2$ = 0.765 even for the two-parameter regression model). The best regression models obtained for subset *B* included CPSA descriptors in combination with shadow indices and volume and shape descriptors (molecular volume, total solvent-accessible surface area, shape parameter). It is noteworthy that NRD for subset *B* is 35 vs NRD = 23 for subset *A*; i.e. in spite of a wider choice of more or less nonredundant descriptors, subset *B* still gives a worse correlation than subset *A*.

The combination set of descriptors *C* was expected to give much better correlations than either one of sets *A* and *B*, because of the variety and quantity of descriptors included (NRD = 58). It was obvious that the best one-parameter correlation must be the same as for subset *A*, but for two- and three-parameter regression models one could expect a major boost in the $R^2$ value. However, for both the two- and three-parameter regression models, an exhaustive search failed to find any correlation which outperformed the best two- and three-parameter correlations generated from subset *A* alone. The influence of the nontopological descriptors showed up only for the four- and five-parameter regression models, where addition of descriptors RPCG ("relative partial charge") and then shape parameter $\eta$ brings, not enormous but still significant, increases of $R^2$ from 0.810 to 0.844 and then to 0.857. It is noteworthy that four- and five-parameter correlations obtained for the combination set *C* have not only the best $R^2$ values but also the best $R^2_{\text{crossval}}$ (0.819 and 0.826, respectively).

Subset *D* of topological plus normalized topological descriptors (NRD = 30) showed no significant advantage over the basic subset *A*. The best one-, two-, and three-parameter correlations are exactly the same as for subset *A*, whereas the regression models with 4 and 5 independent variables are just slightly improved. The best four-parameter model ($R^2$ = 0.832, $R^2_{\text{crossval}}$ = 0.805) includes a normalized descriptor $^0\chi/N_a$, in addition to the combination of Randić index $^0\chi$ with two information indices $^1$IC and $^2$IC. An attachment of the Wiener index as the fifth parameter improves the fit and stability, but to a very minor extent: $R^2$ = 0.838, $R^2_{\text{crossval}}$ = 0.808.

The results obtained for subset *E* of topological plus squared topological descriptors are very close in quality to those generated for subset *D*. Only very slight improvement in comparison with subset *A* was observed for the two-parameter correlation ($R^2$ = 0.768, $R^2_{\text{crossval}}$ = 0.748) which includes descriptors $^0\chi$ and $^3$IC (squared). The best three-parameter model ($R^2$ = 0.821, $R^2_{\text{crossval}}$ = 0.798) includes two information indices ($^1$SIC and $^2$IC) in squared form and the Randić index $^0\chi$ in linear form. Notably, this equation is the best among all three-parameter models generated throughout subsets *A*–*E*. The best four-parameter correlation, which is built on the combination of descriptors $^0\chi$, $^1$SIC, and $^2$IC with squared $^0$SEPD, is, however, slightly worse than that in subset *C*: $R^2$ = 0.834, $R^2_{\text{crossval}}$ = 0.805. At the level of the five-parameter regression model which includes an additional descriptor $^0$SEPD in linear form, we again observe a decrease of the stability of the model, although the fit is certainly improved: $R^2$ = 0.839, $R^2_{\text{crossval}}$ = 0.804. It is noteworthy that subset *E* contains many highly intercorrelated descriptors: although

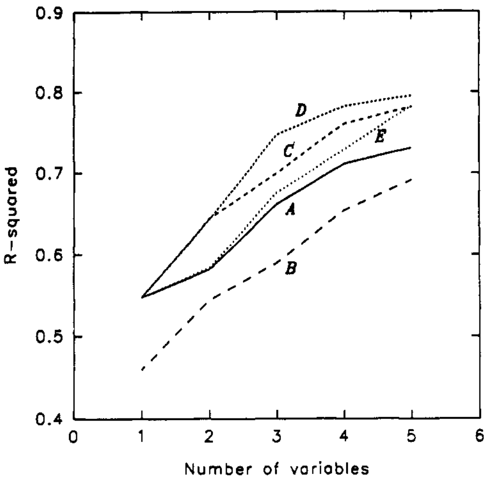

**Figure 2.** Dependence of $R^2$ on model size (structures, amines;[45] property, melting point).

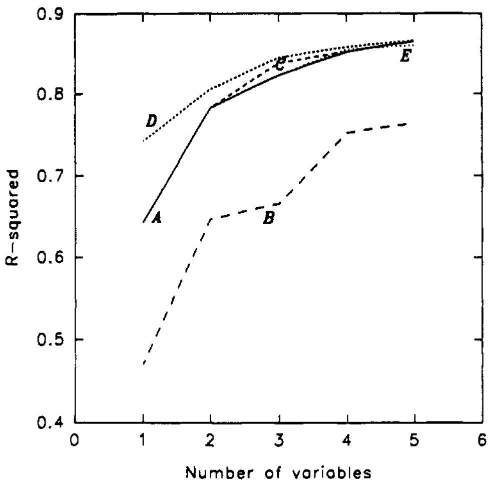

**Figure 3.** Dependence of $R^2$ on model size (structures, ketones;[44] property, melting point).

the total number of descriptors in subset $D$ is 47 vs 76 in subset $E$, the NRD values for both are close: 30 vs 33.

*Amines* (*Figure 2*). For the set of amines, only fair correlations with the melting points were found. As for the aldehydes, the best five-parameter correlation in subset $A$ (NRD = 22) is a combination of connectivity descriptors with information indices, but the $R^2$ value is significantly lower for the amines: 0.730 vs 0.833 for the aldehydes.

Subset $B$ (NRD = 26) also showed no advantages over subset $A$. Even the best five-parameter regression model, which includes CPSA descriptor, shadow index, submolecular polarity parameter, and descriptors PD and DIST, has a poor $R^2$ of 0.691.

The combination set of descriptors $C$ brings some, although not very much, improvement in the quality of correlation: combination of connectivity index $^3\chi$ with electronic–geometrical descriptor PD gives the best, although still poor ($R^2 = 0.644$, $R^2_{\text{crossval}} = 0.608$), two-parameter correlation among all five subsets of descriptors. Some further improvement of the correlations in subset $C$ was achieved when the combination of connectivity descriptors with information indices was extended by the electronic–topological descriptors. The best four-parameter model, which includes descriptors $^3\chi$, $^0$IC, $^1E^T$, and $^2E^T$ is of fairly reasonable quality and stability: $R^2 = 0.760$, $R^2_{\text{crossval}} = 0.696$. Further possible improvement of the model is questionable, since the attachment of the descriptor $^3$SEPD decreases the stability of the model to a significant extent: $R^2 = 0.781$, $R^2_{\text{crossval}} = 0.663$.

However, the regression models obtained for subsets $D$ and $E$ showed that, for this data set, purely topological descriptors can still correlate fairly well with melting points, if used in normalized or squared form. Indeed, the best two-parameter correlation ($R^2 = 0.643$, $R^2_{\text{crossval}} = 0.605$) which includes normalized descriptor $^0\chi^v/N_a$ and information index $^2$CEPD, is of approximately the same quality as the best two-parameter regression model obtained for the combination set $C$. The optimum correlations with three, four, and five independent variables obtained for the subset $D$ (NRD = 30) are superior to correlations with the same number of parameters obtained for subsets $A$–$C$ and $E$. The best three-parameter regression model ($R^2 = 0.747$, $R^2_{\text{crossval}} = 0.695$) includes descriptors $^1\chi^v/N_b$, $^1$EPD, and $^2$SEPD. An attachment of another normalized connectivity descriptor $^2\chi^v/N_b$ gives the best four-parameter equation ($R^2 = 0.782$, $R^2_{\text{crossval}} = 0.720$). Unlike the five-parameter equations generated in subsets $A$–$C$, the model which includes, in addition, descriptor $^3\chi^v/N_b$ has both better fit and higher stability: $R^2 = 0.795$, $R^2_{\text{crossval}} = 0.727$.

It is noteworthy that subset $E$ (NRD = 25) also showed fairly good performance: better than subset $A$ and comparable with the combination set $C$.

Throughout the subsets $A$–$E$, the best single parameter correlation found is based on the descriptor $^3\chi$ ("Randić index of the 3rd order"). However, this descriptor was not included in the majority of the best multiparameter regression models obtained. Apparently, the information indices calculated as modifications of the "entropy of probability distribution" provided the major boost in $R^2$ and $R^2_{\text{crossval}}$ values, when combined with molecular connectivity descriptors, especially if the latter are included in the normalized form.

*Ketones* (*Figure 3*). The melting points for the set of ketones are correlated much better than those of the amines and slightly better than those of the aldehydes. Subset $A$ of topological descriptors (NRD = 21) correlates quite well, and gives three-, four-, and five-parameter regressions with $R^2$ ($R^2_{\text{crossval}}$) of 0.823 (0.759), 0.852 (0.731), and 0.865 (0.739), respectively. Subset $D$ shows very good correlations with melting points: even the single parameter regression which includes the normalized descriptor $^0\chi/N_a$ has the very reasonable $R^2$ of 0.742 ($R^2_{\text{crossval}} = 0.713$). The addition of a second variable, namely, the descriptor $^3\chi^v$ ("molecular connectivity index of 3rd order") increases $R^2$ to 0.806 ($R^2_{\text{crossval}} = 0.768$). Combination of normalized connectivity descriptors of the first and third order with $^3\kappa$ ("Kier shape index of 3rd order") gives the best three-parameter correlation with $R^2$ of 0.845 ($R^2_{\text{crossval}} = 0.743$), which is better than any other three-parameter correlation obtained for sets $A$–$C$ and $E$. However, four- and five-parameter regression models obtained for the sets of topological descriptors are of almost the same quality, with $R^2$ in the range of 0.852–0.858 for four-parameter regression and in the range of 0.860–0.866 for five-parameter regression. The four-parameter regression model obtained for subset $E$ ($R^2 = 0.855$) should be mentioned particularly as a correlation with the relatively high $R^2_{\text{crossval}}$ value of 0.781. This correlation includes two topological indices in the traditional form (Randić index and molecular connectivity index, both of the 1st order) in combination with the information index ("complementary information content of zero order") and the squared shape index of the 2nd order.

With regard to topological descriptors, the best correlations were obtained by using connectivity indices suggested by Randić and Kier and Hall, especially if the regressions allowed inclusion of those indices in a normalized form. Further improvement of regression models was achieved using the shape indices. Information indices were still useful, but to a lesser extent: $^0$IC was included, in combination with $^1\chi^v$, in the best two-parameter correlations for subsets *A*, *C*, and *E*. The $^0$CIC ("complementary information content of zero order") was included in the best five-parameter regression models for subsets *A* and *C*, whereas $^0$CEPD and squared $^1$SEPD were found to be the best additional fifth independent variable in correlations for subsets *D* and *E*.

Notably, subset *B* of electronic, geometrical, and combined descriptors showed a significantly worse performance than any of the subsets comprising topological indices. The best five-parameter regression obtained for subset *B* has an $R^2$ value of only 0.764, which is worse than the best two-parameter correlations obtained for any of the subsets *A*, *D*, and *E*. Also, this correlation proved to be not very stable: $R^2_{\text{crossval}} = 0.600$.

As seen in Figure 3, the five-parameter regression model found for *E* is slightly worse than the five-parameter correlation obtained for *A*. At first glance, this seems incorrect because set *E* includes the whole subset *A*. Here we should recall that the search for five-parameter correlation is conducted according to the so-called "4 + 1" scheme rather than exhaustively (see the section *Method*). In this case, the best four-parameter regression model obtained for set *A* happened to be a better basis for "attachment" of the fifth independent variable than the best four-parameter correlation obtained for set *E*. In most cases, however, a better four-parameter correlation usually leads to a better five-parameter regression model.

The most surprising result obtained for the set of ketones is that the exhaustive search conducted in the combination set *C* reproduced absolutely the same one- to five-parameter regression models as those obtained for subset *A*. This means that addition of 46 electronic, geometrical, and combined descriptors to the set of 38 topological indices did not bring any improvement in the one- to five-parameter correlations. The NRD value for the combination set *C* is 44; in other words, the combined set has twice as many nonredundant descriptors as subset *A*. This excludes the possibility that, for this particular data set, all nontopological descriptors happened, by chance, to be highly intercorrelated and therefore redundant to topological indices.

**Boiling Point.** Currently, there is considerable interest in exploring relationships between boiling points and organic structures. Here we first mention two papers by Jurs and co-workers devoted to QSPR research on sets of furans, tetrahydrofurans, thiophenes, pyrans, and pyrroles.[23,46] Balaban and co-workers published research on the correlations between chemical structure and normal boiling points for the set of halogenated alkanes $C_1$–$C_4$ and for a set of acyclic ethers, peroxides, acetals, and their sulfur analogues.[24,47]

Jurs' work concentrated on the regression models which combine molecular descriptors of various types. For example, in the best regression model obtained for a set of 209 furans and tetrahydrofurans ($R^2$ = 0.969, $s$ = 11.2 °C), 11 independent variables were included; among them 8 descriptors used in the present study: *J*, $^1\chi$, $^3\chi^v$, PPSA1, PPSA3, FNSA3, WPSA3, RPCG. The simple count of single bonds, the square root of $^3\kappa$, and the LUMO energy obtained from the simple Hückel method were also included in the Jurs model.

The best seven-parameter regression model ($R^2$ = 0.974, $s$ = 7.9 °C) obtained on a set of 134 thiophenes also included a set of various molecular descriptors, including both simple descriptors such as molecular weight and a count of single bonds, CPSA descriptors, dipole moment, one of topological indices (square root of the ratio of the paths of all lengths to the number of atoms in the molecule), and the radius of gyration.

For a set of pyrans and pyrroles, approximately the same assortment of molecular descriptors was used. Most of the regression models reported in ref 46 included some of the CPSA descriptors in combination with the dipole moment, radius of gyration, count of single bonds, count of atoms, plus some topological indices such as $^1\chi^v$, $^3\chi^v$, and *J*.

The following seven-parameter correlations were obtained:[46] set of 146 pyrans ($R^2$ = 0.978, $s$ = 10.2 °C), set of 278 pyrroles ($R^2$ = 0.962, $s$ = 12.3 °C).

In Jurs' work, the descriptors were generated by the ADAPT program (also developed in the Jurs group), and the statistical analysis was held in both ADAPT and Minitab packages.[23,46]

In contrast, the QSPR study conducted by Balaban and co-workers was concentrated primarily on the use of topological indices (including many descriptors used in this study: $^1\chi$, $^0\chi^v$, $^1\chi^v$, $^2\chi^v$, $\kappa$ and $\Phi$), along with various simple and fractional counts of non-hydrogen atoms, halogens, and geminal halogens. The best six-parameter regression model obtained in Balaban's work for a set of 532 haloalkanes has $R^2$ = 0.97 and $s$ = 10.94. For a homogeneous subset of 44 halomethanes[24] with two or more halogen atoms, the best four-parameter correlation obtained has $R^2$ of 0.99 and $s$ = 5.26.

For a more diverse set of 185 acyclic ethers, peroxides, acetals, and their sulfur analogues, the best two-parameter regression model obtained ($R^2$ = 0.964, $s$ = 9.0 °C) included the Randić index of the 1st order $^1\chi$ and the sulfur atom count.[47] In Balaban's work, the molecular descriptors were generated using in-house developed software, while the statistical analysis was performed with the SAS package. Jurs' research mentioned that the regression models presented were checked by the jackknifing technique and proved to be stable, although no quantitative data were provided. In Balaban's study, some information on the intercorrelation between descriptors included in the same correlation equation, was provided, but no cross-validation or jackknifing technique was applied.

This brief analysis of the literature shows that the results obtained for the correlation between structures and boiling points are somewhat contradictory. Although both research groups (Balaban's and Jurs') obtained satisfactory statistical evidence that boiling points can be predicted from molecular descriptors, the question arises: Which types of descriptors (only topological, only nontopological, or in combination) are better to use in QSPR research conducted on sets of structures vs boiling points?

Therefore, in this study we investigated relationships between the boiling points and the structures of aldehydes ($n$ = 43), amines ($n$ = 104), and ketones ($n$ = 49), when structures are represented with different subsets of molecular descriptors (*A*–*E*). Just as for our research on melting points, the three data sets were taken from refs 44 and 45.

*Aldehydes (Figure 4).* The best single-parameter correlation equation ($R^2$ = 0.932, $R^2_{\text{crossval}}$ = 0.926) is based on the descriptor $^3\chi$ ("Randić index of the 3rd order"). However, for two-parameter regression models a combination of another Randić index descriptor $^1\chi$ with the information index $^0$CIC provides a better equation ($R^2$ = 0.960, $R^2_{\text{crossval}}$ = 0.948). These two descriptors showed up together in the best two-parameter correlations found for subsets *A*, *C*, *D*, and *E*. In the last case, the $^0$CIC appeared in the squared form, and this

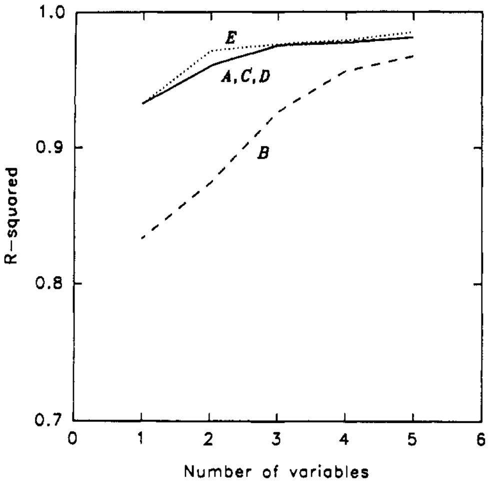

**Figure 4.** Dependence of $R^2$ on model size (structures, aldehydes;[44] property, boiling point).

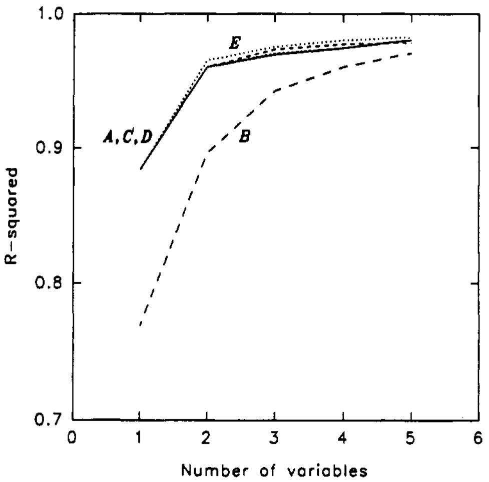

**Figure 5.** Dependence of $R^2$ on model size (structures, amines;[45] property, boiling point).

improved both the fit and stability of the model: $R^2 = 0.971$ and $R^2_{\mathrm{crossval}} = 0.966$.

In subset *A* (NRD = 20), the best three-parameter regression ($R^2 = 0.975$, $R^2_{\mathrm{crossval}} = 0.967$) includes the Wiener index in combination with $^0$CIC and $^1\chi$. Addition of the Balaban index *J* gives the best four-parameter equation ($R^2 = 0.977$, $R^2_{\mathrm{crossval}} = 0.968$). Further improvement of the model was achieved with the attachment of connectivity descriptor $^0\chi^v$: $R^2 = 0.981$, $R^2_{\mathrm{crossval}} = 0.972$.

The best single-parameter equation generated for subset *B* is based on the shadow index $S_4$ ($R^2 = 0.833$, $R^2_{\mathrm{crossval}} = 0.818$). Addition of descriptor WNSA3 improves the model, but to a small extent: $R^2 = 0.874$, $R^2_{\mathrm{crossval}} = 0.835$. The best three-parameter regression model is based on a combination of two CPSA descriptors (PPSA2 and DPSA3) with one geometrical descriptor (MV). This model has a good fit ($R^2 = 0.926$), but lacks in stability ($R^2_{\mathrm{crossval}} = 0.848$). The stability of the model is improved significantly ($R^2 = 0.956$, $R^2_{\mathrm{crossval}} = 0.929$) at the level of the four-parameter equation, which includes topological–electronic index $T^E$, submolecular polarity parameter (SPP), relative partial charge (RPCG), and descriptor MV ("molecular volume). The addition of a fifth independent variable (PPSA3) improves both the fit and stability: $R^2 = 0.967$, $R^2_{\mathrm{crossval}} = 0.949$. In general, subset *B* (NRD = 33) showed the worst performance for all the numbers of variables (1–5) used to create a correlation equation. The best five-parameter regression model found for subset *B* is worse than the three-parameter correlations obtained for subsets *A*, *C*, and *D* and worse than a two-parameter equation generated for subset *E*.

Subset *C* (NRD = 52) gives absolutely the same one- to three-parameter correlations as those found for subset *A*. Therefore, the addition of nontopological descriptors did not bring any improvement for small regression models. The four-parameter correlation obtained for subset *C* (with FNSA1 descriptor instead of *J*) is just slightly better ($R^2 = 0.978$, $R^2_{\mathrm{crossval}} = 0.970$) than that in subset *A*, but the addition of the fifth descriptor (RPCS) makes the five-parameter regression model worse ($R^2 = 0.981$, $R^2_{\mathrm{crossval}} = 0.966$) than the five-parameter equation in subset *A*.

Subset *D* (NRD = 27) provides exactly the same correlation equations as subset *A*. Therefore, the normalized forms of topological descriptors appeared not to be very important for this data set.

By contrast, the squared form of topological indices in subset *E* (NRD = 24) proved to be very useful for the correlation between aldehyde structures and boiling points. As mentioned above, the improvement of the regression model was already remarkable for the two-parameter equation. The three-parameter equations obtained for subsets *A*, *C*, *D*, and *E* are very close in quality and stability, but the four- and five-parameter regression models obtained for *E* are again significantly better than those found for subsets *A*, *C*, and *D*. The best four-parameter equation ($R^2 = 0.979$, $R^2_{\mathrm{crossval}} = 0.976$) includes the Randić index $^1\chi$ in linear form and the three descriptors ($^0\chi$, $^0\chi^v$, and $^0$CIC) in squared forms. An addition of connectivity descriptor $^1\chi^v$ improves the fit, but slightly decreases the stability of the model (five-parameter equation; $R^2 = 0.985$, $R^2_{\mathrm{crossval}} = 0.976$).

*Amines* (*Figure 5*). For the set of amines, the best single-parameter correlation equation obtained ($R^2 = 0.884$, $R^2_{\mathrm{crossval}} = 0.856$) is worse than that obtained for the aldehydes. Throughout subsets *A*, *C*, *D*, and *E*, the Randić index $^1\chi$ was included in the best one-parameter regression model, rather than descriptor $^3\chi$ as for aldehydes.

In subset *A* (NRD = 22), the addition of the second independent variable, namely, information index $^1$CIC improves the correlation very significantly: ($R^2 = 0.960$, $R^2_{\mathrm{crossval}} = 0.948$), providing a correlation of approximately the same quality as that for the set of aldehydes. The same descriptors $^1\chi$ and $^1$CIC were included in the best two-parameter regression models in subsets *C* and *D*.

The best three-parameter regression model ($R^2 = 0.969$, $R^2_{\mathrm{crossval}} = 0.964$) for subset *A* is based on the combination of Wiener index (*W*) with Randić index $^1\chi$ and information index $^1$CEPD. A combination of Wiener and Randić indices with descriptors $^0\chi^v$ and $^1$SIC improves the model to a very minor extent: $R^2 = 0.974$, $R^2_{\mathrm{crossval}} = 0.966$. Addition of the fifth descriptor ($^2$SIC), however, improves both the fit and stability more significantly: $R^2 = 0.980$, $R^2_{\mathrm{crossval}} = 0.976$.

The best one- and two-parameter correlation equations obtained for subset *B* are not very impressive. Descriptor DPSA3, included in the best monoparameter model, provides only fair correlation: $R^2 = 0.769$, $R^2_{\mathrm{crossval}} = 0.761$. An addition of descriptor RNCG improves the model, but not greatly ($R^2 = 0.896$, $R^2_{\mathrm{crossval}} = 0.890$). More significant improvement of the model ($R^2 = 0.942$, $R^2_{\mathrm{crossval}} = 0.938$) was achieved at the level of the three-parameter regression which includes descriptors WPSA1, WPSA3, and MV. The best four-parameter equation ($R^2 = 0.960$, $R^2_{\mathrm{crossval}} = 0.949$) is based on descriptors $^cT^E$, PPSA3, WPSA1, and MV. In general, use of subset *B* of nontopological descriptors (NRD

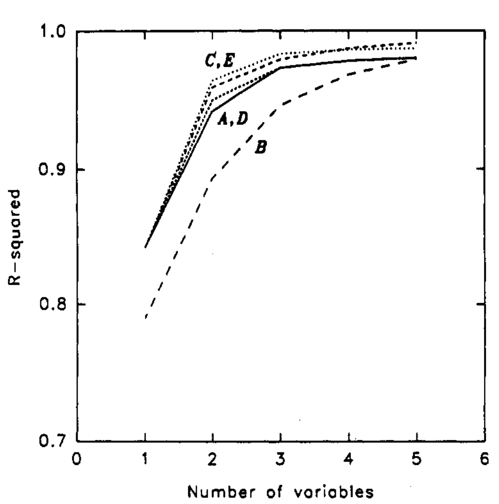

**Figure 6.** Dependence of $R^2$ on model size (structures, ketones;[44] property, boiling point).

= 29) is again substantially inferior to the subsets of topological indices: the quality of the best five-parameter regression, which includes an additional descriptor $S_2$, is comparable with the three-parameter correlation equations found for subsets *A*, *C*, *D*, and *E*.

At the level of one- and two-parameter regression models, subset *C* (NRD = 49) indicates no improvement in comparison with subset *A*. The influence of nontopological descriptors shows up in the three-parameter correlation equation ($R^2$ = 0.973, $R^2_{\rm crossval}$ = 0.970), which is based on the Randić index $^1\chi$ in combination with the CPSA descriptors (PPSA3 and WPSA1). The best four-parameter regression ($R^2$ = 0.977, $R^2_{\rm crossval}$ = 0.974) includes the combination of two connectivity descriptors ($^1\chi$ and $^3\chi^v$) with two CPSA descriptors (DPSA2 and DPSA3). This correlation is the best throughout all five subsets *A*–*E*. Some further small improvement of this regression model ($R^2$ = 0.978, $R^2_{\rm crossval}$ = 0.975) was achieved with the addition of the electronic–topological descriptor $^1E^{\rm T}$.

Subset *D* (NRD = 30) provided absolutely the same correlations as subset *A*, with one exception: the three-parameter regression model ($R^2$ = 0.970, $R^2_{\rm crossval}$ = 0.962) which includes the Wiener index, Randić index $^1\chi$, and $^0\chi^v/N_a$, is slightly better than the best three-parameter correlation obtained for subset *A*.

Although subset *E* (NRD = 33) provides better correlation equations for the two- to five-parameter regression models than subset *A*, the difference in quality between correlations obtained for these two subsets is not as significant as in the case of aldehydes (cf. Figures 4 and 5). All two- to five-parameter equations include the combination of $^1\chi$ with the squared form of $^0\chi^v$. The best three-parameter equation ($R^2$ = 0.975, $R^2_{\rm crossval}$ = 0.973), which includes, in addition, the information index $^1$EPD, provides the best fit and stability in comparison with all the best three-parameter models in subsets *A*–*E*. However, an attachment of the fourth descriptor ($^2$CEPD squared) to the model decreases the stability of the equation: $R^2$ = 0.980, $R^2_{\rm crossval}$ = 0.972. No significant improvement of the model was observed at the level of the five-parameter regression with the additional descriptor *W* (squared): $R^2$ = 0.982, $R^2_{\rm crossval}$ = 0.972.

*Ketones* (*Figure 6*). Similarly to the subset of aldehydes, the Randić index $^3\chi$ is again the best for a single-parameter correlation equation. However, the quality of correlation thus obtained is significantly lower for ketones than for aldehydes ($R^2$ = 0.842, $R^2_{\rm crossval}$ = 0.831). The much improved two-parameter correlation equation ($R^2$ = 0.942, $R^2_{\rm crossval}$ = 0.890) was obtained using the combination of the Wiener index *W* and the Randić index $^1\chi$. Nevertheless, this correlation was still worse than the two-parameter equations found for the sets of aldehydes and amines. At the level of the three-parameter regression model, when the descriptor $^0$SIC was attached, the correlation was significantly improved, especially from the point of stability ($R^2$ = 0.973, $R^2_{\rm crossval}$ = 0.938). This regression model is already comparable to other three-parameter equations obtained for the sets of aldehydes and amines. Further, although not dramatic, improvement of the regression model for subset *A* (NRD = 24) was achieved, when descriptors *W* and $^1\chi$ were combined with information indices ($^2$SIC and $^3$CIC): $R^2$ = 0.978, $R^2_{\rm crossval}$ = 0.945. An attachment of the shape index ($^3\kappa$) improves the model to a very minor extent: $R^2$ = 0.980, $R^2_{\rm crossval}$ = 0.946.

The best monoparameter correlation equation (descriptor WNSA2) in subset *B* is only of a fair quality: $R^2$ = 0.790, $R^2_{\rm crossval}$ = 0.774). The best two-parameter correlation which includes descriptors PPSA2 and MV is of a much better quality: $R^2$ = 0.893, $R^2_{\rm crossval}$ = 0.874. The three-parameter regression (descriptors $^1E^{\rm T}$, PPSA3, and WPSA1) provides significant improvement in both the fit and stability of the model: $R^2$ = 0.946, $R^2_{\rm crossval}$ = 0.922. The combination of descriptors $^0E^{\rm T}$, $^cT^{\rm E}$, DPSA3, and WPSA2 gives a four-parameter model of very good quality: $R^2$ = 0.968, $R^2_{\rm crossval}$ = 0.944. Furthermore, significant improvement of the correlation equation was achieved when descriptor FPSA3 was added to the model: $R^2$ = 0.979, $R^2_{\rm crossval}$ = 0.970. Thus, the one- to four-parameter correlations obtained for subset *B* are again worse than those found in other subsets. However, for this particular data set, subset *B* (NRD = 24) provides a surprisingly good five-parameter regression model, which not only gives almost the same fit as the five-parameter equation in subset *A* ($R^2$ = 0.979 vs 0.980) but is also more stable ($R^2_{\rm crossval}$ = 0.970 vs 0.946).

Given the better than usual performance of subset *B*, we expected that the combination set *C* (NRD = 44) would bring a real improvement to the regression models generated. Indeed, even at the level of the two-parameter correlation, we obtained a very good ($R^2$ = 0.959, $R^2_{\rm crossval}$ = 0.924) equation which includes both topological ($^1\chi$) and nontopological (WPSA2) descriptors. At the level of the three-parameter regression model, with a DPSA3 descriptor attached, subset *C* provided an excellent and stable correlation ($R^2$ = 0.979, $R^2_{\rm crossval}$ = 0.976). The best four-parameter equation ($R^2$ = 0.987, $R^2_{\rm crossval}$ = 0.979) was obtained as a combination of $^1\chi$ with three CPSA descriptors: PNSA2, WPSA2, and WNSA3. Addition of the fifth descriptor (information index $^0$SEPD) improves the fit ($R^2$ = 0.991), but slightly decreases the stability of the model ($R^2_{\rm crossval}$ = 0.974).

The correlations obtained for subset *D* (NRD = 31) again very much overlap those found in subset *A*. A small improvement, in comparison with subset *A*, was achieved with the two-parameter regression model ($R^2$ = 0.950, $R^2_{\rm crossval}$ = 0.909), which included the normalized Wiener index ($W/N_a$) and $^1\chi$. Correlations with three and four parameters are exactly the same as for subset *A*. The five-parameter equation again includes the normalized Wiener index and gives a slightly better fit ($R^2$ = 0.980), but lower stability ($R^2_{\rm crossval}$ = 0.942). Notably, the best five-parameter regression model for subset *A* has the same fit but is slightly more stable.

For the set of ketones, the subset of descriptors *E* (NRD = 31) appeared to be the best source of two- and three-parameter regression models. The best two-parameter equation ($R^2$ = 0.964, $R^2_{\rm crossval}$ = 0.934) includes the Randić index

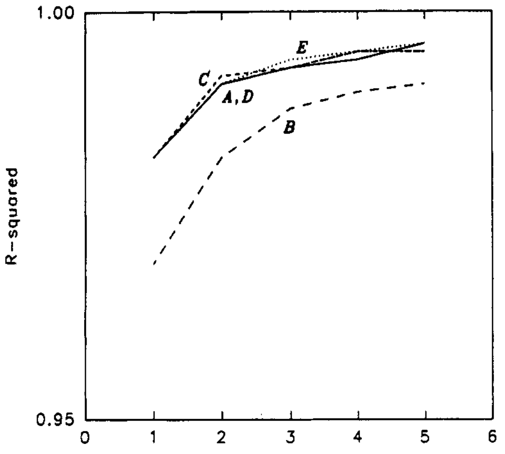

**Figure 7.** Dependence of $R^2$ on model size (structures, aldehydes;[44] property, molar volume).

($^1\chi$) in combination with the squared information index ($^0$CIC). The same combination of descriptors showed up as the best two-parameter correlation for the set of aldehydes. The best three-parameter correlation equation ($R^2 = 0.983$, $R^2_{\text{crossval}} = 0.947$) includes the squared Wiener index in combination with descriptors $^1\chi$ and $^1\kappa$. However, this equation is less stable than the three-parameter model obtained for subset *C*. The best four-parameter equation, which in addition includes the information index $^1$IC, has almost the same fit ($R^2 = 0.986$) as the four-parameter equation in subset *C*, but is somewhat less stable ($R^2_{\text{crossval}} = 0.960$). Addition of the Wiener index in linear form brings no significant improvement to the model: $R^2 = 0.987$, $R^2_{\text{crossval}} = 0.960$.

**Molar Volume.** The calculation and modeling of molar volume, which is defined as the ratio of molecular weight to density ($V_m = \text{MW}/d$), has attracted attention (i) because of the direct access to the estimation of density thus provided[48] and (ii) as a suitable example for development of additive schemes.[48,49]

Molar volume was shown to correlate well with connectivity descriptors for the set of alkanes[11] and for the set of alkylbenzenes.[50]

We especially wished to investigate the structure–molar volume relationship using our five different subsets of descriptors (*A–E*), since it seemed logical that such geometrical descriptors as total molecular volume (MV) and total solvent-accessible molecular surface (TSASA) should correlate better with molar volume than topological indices.

For this study we again used the three data sets from refs 44 and 45: aldehydes ($n = 59$), amines ($n = 109$), and ketones ($n = 60$).

*Aldehydes* (*Figure 7*). All one- to five-parameter correlations obtained for the set of aldehydes are of excellent quality, with a minimum $R^2$ value of 0.969 (single-parameter correlation in subset *B* of nontopological descriptors) and a maximum $R^2$ value of 0.996 (five-parameter regression model in subset *A* of topological indices). All regression models are very stable, with $R^2_{\text{crossval}}$ lying in the interval 0.966–0.995. For the subsets *A*, *C*, *D*, and *E* the best single-parameter correlation ($R^2 = 0.982$, $R^2_{\text{crossval}} = 0.980$) was obtained using the descriptor shape index $^1\kappa$. For subset *A* (NRD = 19), the best two-parameter correlation equation ($R^2 = 0.991$, $R^2_{\text{crossval}} = 0.989$) is based on a combination of connectivity descriptor $^0\chi^v$ and information index $^0$CIC. Notably, both indices participate in the "zero-order" form; in other words, connectivity in this case is reflected only slightly, and the basic role is played by the composition of the aldehyde. At the level of the three-parameter regression model, another information index ($^0$SEPD) joins the combination of $^0\chi^v$ and $^0$CIC, giving the optimum three-parameter correlation equation with $R^2 = 0.993$ and $R^2_{\text{crossval}} = 0.989$. However, none of these descriptors is involved in the best four-parameter regression model ($R^2 = 0.994$, $R^2_{\text{crossval}} = 0.993$). Instead, the best four- and five-parameter correlation equations include descriptors such as *W*, $^2\chi^v$, $^2\kappa$, and $^0$CEPD (four-parameter model) plus again $^1\kappa$ (five-parameter model).

The results obtained for subset *B* (NRD = 31) indeed supported our initial hypothesis that either descriptor MV or descriptor TSASA need show up in the best correlations obtained. In fact, the single-parameter regression model based on TSASA gives a very good fit ($R^2 = 0.969$, $R^2_{\text{crossval}} = 0.966$), but still inferior the equation based on descriptor $^1\kappa$ in subset *A*. The combination of TSASA with the shape parameter ($\eta$) improves the correlation significantly ($R^2 = 0.982$, $R^2_{\text{crossval}} = 0.980$). The best three- to five-parameter regression models all include descriptor molecular volume (MV), calculated on the basis of 3D molecular geometry and the values of atomic radii. However, apart from the purely geometrical descriptor MV, all the best three- to five-parameter regression models obtained for subset *B* include, correspondingly, two, three, and four combined descriptors, namely, CPSA descriptors (three- and four-parameter regression models) plus the topological–electronic descriptor ($T^E$) in the five-parameter regression model. All correlations obtained for subset *B* possess lower stability in comparison with those for subset *A*: $R^2_{\text{crossval}}$ belongs to the range 0.966–0.985.

The combination subset *C* (NRD = 52) showed better correlations at the level of two- to four-parameter regression models. For example, an excellent two-parameter correlation ($R^2 = 0.992$, $R^2_{\text{crossval}} = 0.990$) was obtained by using the combination of Randić index $^1\chi$ with descriptor MV. Further improvement of the model was achieved when, for the three-parameter regression, the Wiener index *W* was included ($R^2 = 0.993$, $R^2_{\text{crossval}} = 0.991$) and, for the four-parameter regression, the information index $^0$SEPD ($R^2 = 0.995$, $R^2_{\text{crossval}} = 0.994$). An addition of the Balaban index *J* to the model brings virtually no improvement ($R^2 = 0.995$, $R^2_{\text{crossval}} = 0.994$).

As expected, descriptor MV appeared to be very significant for correlations with molar volume: this descriptor showed up in all the best two- to five-parameter correlations.

At the level of one- and two-parameter correlations, subset *D* (NRD = 31) showed no superiority over subset *A*: exactly the same correlation equations as for *A* were found. Addition of the normalized connectivity index $^2\chi^v/N_b$ only improved slightly the three- and four-parameter regression models.

Subset *E* (NRD = 27) showed once again that the squared forms of topological indices can play a significant role in the best correlation equations. Indeed, the best three-parameter regression model ($R^2 = 0.994$, $R^2_{\text{crossval}} = 0.992$) obtained throughout all five subsets of descriptors (*A–E*), includes descriptors $^0\chi^v$, $^0$CEPD, and $^1\kappa$ (squared). The best four-parameter regression ($R^2 = 0.995$, $R^2_{\text{crossval}} = 0.994$) obtained for subset *E* is also the best throughout all five subsets, and this model includes topological descriptors in both linear ($^0\chi^v$, $^3\kappa$) and squared form (*W* and $^0$CEPD). The attachment of the fifth descriptor (squared Balaban index *J*) improves the correlation ($R^2 = 0.996$, $R^2_{\text{crossval}} = 0.995$), to the same quality as the five-parameter equation obtained for subset *A*.

*Amines* (*Figure 8*). The set of amines also showed excellent correlations between molecular descriptors and molar volumes,

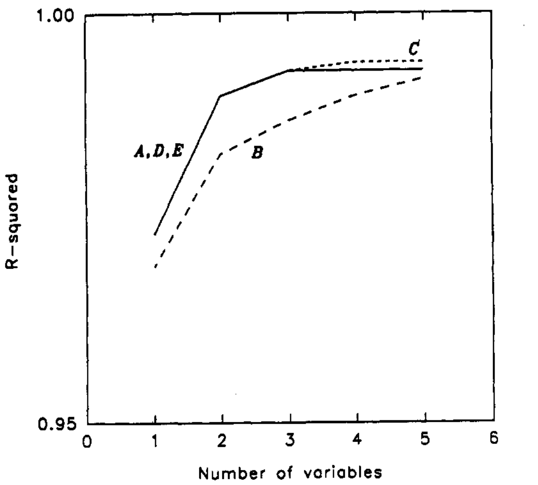

**Figure 8.** Dependence of $R^2$ on model size (structures, amines;[45] property, molar volume).

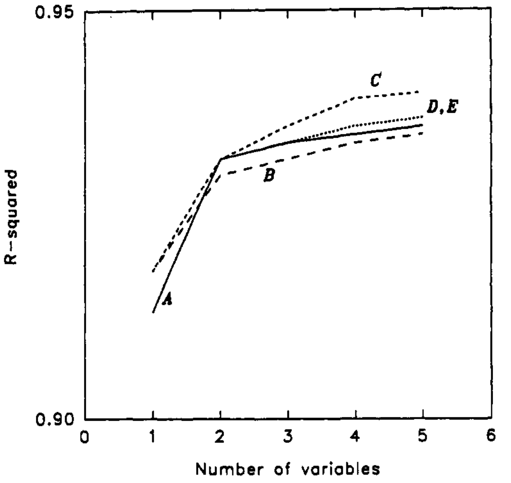

**Figure 9.** Dependence of $R^2$ on model size (structures, ketones;[44] property, molar volume).

although the regression models obtained for the set of aldehydes were slightly better.

Similarly to the set of aldehydes, the best single-parameter correlation equation ($R^2 = 0.973$, $R^2_{\text{crossval}} = 0.972$) obtained in subset $A$ (NRD = 23) is based on the shape index ${}^{1}\kappa$. The best two-parameter equation ($R^2 = 0.990$, $R^2_{\text{crossval}} = 0.989$) includes a combination of two connectivity descriptors (${}^{3}\chi$ and ${}^{0}\chi^{v}$). Further, this combination is joined with the "flexibility index" $\Phi$ (best three-parameter regression: $R^2 = 0.993$, $R^2_{\text{crossval}} = 0.992$) and then with the shape index ${}^{2}\kappa$ (best four-parameter regression: $R^2 = 0.993$, $R^2_{\text{crossval}} = 0.992$). Addition of another connectivity descriptor (${}^{1}\chi^{v}$) gives us the best five-parameter correlation equation with $R^2 = 0.993$ and $R^2_{\text{crossval}} = 0.992$. As seen from these data, no improvement was observed for all regression models larger than the two-parameter model. Notably, none of the best one- to five-parameter regression models includes any information indices.

The best single-parameter correlation ($R^2 = 0.969$, $R^2_{\text{crossval}} = 0.966$) obtained for subset $B$ (NRD = 29) is based on the descriptor molecular volume (MV). This descriptor is also included in all the best two- to five-parameter correlation equations found for subset $B$. The best two-parameter equation ($R^2 = 0.983$, $R^2_{\text{crossval}} = 0.982$) is based on a combination of the descriptors MV and PPSA1. The best three-parameter regression model ($R^2 = 0.987$, $R^2_{\text{crossval}} = 0.985$) includes topological–electronic index (${}^{c}T^{E}$) combined with FPSA3 and MV. Further improvement of the model was achieved with the addition of the WPSA1 descriptor ($R^2 = 0.990$, $R^2_{\text{crossval}} = 0.989$) and then with the addition of the submolecular polarity parameter (SPP). Thus, the best five-parameter correlation equation in subset $B$ has $R^2 = 0.992$ and $R^2_{\text{crossval}} = 0.990$, which is, however, still worse than the three-parameter regression model obtained for subset $A$.

The results obtained for the combination set $C$ (NRD = 52) repeat those from subset $A$ at the level of regression models with one to two parameters. Only the correlations with three to five parameters include geometrical descriptors, along with topological indices. The descriptor of total molecular surface (TSASA), in combination with two connectivity descriptors (${}^{1}\chi$ and ${}^{0}\chi^{v}$), appeared in the best three-parameter correlation ($R^2 = 0.993$, $R^2_{\text{crossval}} = 0.992$). Addition of the shadow index $S_3$ to this combination gives the best four-parameter model ($R^2 = 0.994$, $R^2_{\text{crossval}} = 0.993$), and this model is indeed found to be the best throughout all five subsets of descriptors. A further, albeit very slight, improvement ($R^2 = 0.994$, $R^2_{\text{crossval}} = 0.993$) was achieved for the five-parameter correlation equation, with the addition of another shadow index ($S_6$).

The best correlations generated for subset $D$ (NRD = 31) are exactly the same as those found for subset $A$, with the exception of the five-parameter regression model. This model ($R^2 = 0.993$, $R^2_{\text{crossval}} = 0.992$) includes the Wiener index normalized by the number of skeletal atoms ($W/N_a$) and is hardly better than the five-parameter equation obtained for subset $A$.

The best one- to three-parameter regression models found for subset $E$ (NRD = 32) are exactly the same as those generated for subsets $A$ and $D$. Moreover, even the four- and five-parameter correlations generated for subsets $A$, $D$, and $E$ possess exactly the same values of $R^2$ and $R^2_{\text{crossval}}$, although the descriptors involved in the equations are different in the case of subset $E$.

*Ketones* (*Figure 9*). By contrast with the results obtained for the sets of aldehydes and amines, the best one-parameter regression model ($R^2 = 0.913$, $R^2_{\text{crossval}} = 0.908$) found for ketones in subset $A$ (NRD = 23) is based on the information index ${}^{0}$IC. Apparently, the topological indices of zero order are of particular importance for the adequate description of molar volume for the sets of aldehydes, amines, and ketones. This means that differences in the composition of the compounds in the set prevails over the structural and geometrical differences.

Combination of two connectivity descriptors (${}^{2}\chi$ and ${}^{0}\chi^{v}$) gives the best two-parameter regression model generated for subset $A$ ($R^2 = 0.932$, $R^2_{\text{crossval}} = 0.926$). Furthermore, the best three-parameter correlation equation ($R^2 = 0.934$, $R^2_{\text{crossval}} = 0.926$). Furthermore, the best three-parameter correlation equation ($R^2 = 0.934$, $R^2_{\text{crossval}} = 0.928$) also includes only connectivity descriptors (${}^{0}\chi$, ${}^{2}\chi$, and ${}^{2}\chi^{v}$). At the level of the four-parameter regression model ($R^2 = 0.935$, $R^2_{\text{crossval}} = 0.929$), the following descriptions showed up in the best correlation obtained: $W$, ${}^{1}\chi^{v}$, ${}^{1}$IC, and $J$. Further, although insignificant, improvement ($R^2 = 0.936$, $R^2_{\text{crossval}} = 0.929$) of the regression model was achieved with the attachment of information index ${}^{0}$CIC as the fifth independent variable.

An exhaustive search for the best monoparameter correlation throughout subset $B$ brought somewhat unexpected results. The best single-parameter regression model ($R^2 = 0.918$, $R^2_{\text{crossval}} = 0.914$) appeared to be based not on the volume- or shape-related descriptors as in the case of aldehydes and amines but on the purely electronic descriptor SAPC (sum of absolute values of partial charges). In our view, this may be another reflection of the important role played by the composition of a ketone in the structure–molar volume relationship. However, starting at the level of the two-parameter regression, the descriptor MV takes a strong position in all the best correlation equations found throughout subset *B*. Combination of the MV descriptor with topological-electronic index (${}^{c}T^{E}$) gives a two-parameter regression model ($R^2 = 0.930$, $R^2_{\text{crossval}} = 0.924$) which is even better than the two-parameter equation found for subset *A*. Addition of the third independent variable (WNSA3) very slightly improves the fit but decreases the stability of the model ($R^2 = 0.932$, $R^2_{\text{crossval}} = 0.921$). Combination of the MV descriptor with three CPSA descriptors (PPSA3, WPSA2, and RNCS) brings a small improvement in both the fit and stability of the model: $R^2 = 0.934$, $R^2_{\text{crossval}} = 0.920$. However, attachment of the fifth parameter (shadow index $S_4$), although it improves the fit to a small extent, decreases the stability more significantly ($R^2 = 0.935$, $R^2_{\text{crossval}} = 0.920$). Notably, all the three- to five-parameter regression models found in subset *B* (NRD = 24) are worse than those obtained in subset *A*.

After analysis of the correlations generated for subsets *A* and *B*, the results obtained for the combination set *C* (NRD = 43) are not surprising. The best single-parameter equation is based on the SAPC descriptor (the same equation as discussed for subset *B*), whereas the best two- and three-parameter correlations include a connectivity descriptor of zero order in combination with CPSA descriptors. The best two-parameter equation includes descriptors ${}^0\chi^v$ and PPSA2 ($R^2 = 0.932$, $R^2_{\text{crossval}} = 0.927$). The best three-parameter equation includes descriptors ${}^0\chi$, PNSA3, and WPSA2 ($R^2 = 0.936$, $R^2_{\text{crossval}} = 0.928$). At the level of the four-parameter regression model ($R^2 = 0.939$, $R^2_{\text{crossval}} = 0.928$), the PNSA3 descriptor is replaced with PNSA1 descriptor, and, in addition, the electronic–topological descriptor ${}^0E^T$ is involved, again in the form of zero order. As well as for subset *B*, the attachment of the geometrical descriptor (shadow index $S_4$), although it increases the fit, decreases the stability significantly ($R^2 = 0.940$, $R^2_{\text{crossval}} = 0.920$).

The results obtained for subset *D* (NRD = 29) differ from those found for subset *A* only at the level of four- and five-parameter regression model. Addition of the normalized connectivity descriptor (${}^2\chi^v/N_b$) to the combination of three other topological indices ($W$, ${}^2\chi$, and ${}^0\chi^v$) gives the best four-parameter correlation equation ($R^2 = 0.936$, $R^2_{\text{crossval}} = 0.930$) which is slightly better than that obtained for subset *A*, but worse than the four-parameter equation found for the combination set *C*. Attachment of the fifth independent variable ${}^1$EPD improves the correlation insignificantly ($R^2 = 0.937$, $R^2_{\text{crossval}} = 0.930$).

The results for subset *E* (NRD = 29) are of almost the same quality as those generated for subsets *A* and *D*. Again, two connectivity indices (${}^2\chi$ and ${}^0\chi^v$) are the major descriptors involved in the best correlation equations obtained. At the level of the three-parameter regression model, the information index ${}^0$CIC (squared form) showed up in the best correlation ($R^2 = 0.934$, $R^2_{\text{crossval}} = 0.928$). The best four-parameter correlation equation ($R^2 = 0.936$, $R^2_{\text{crossval}} = 0.926$) includes a combination of two connectivity descriptors ${}^2\chi$ and ${}^0\chi^v$ with shape index ${}^2\kappa$, in both its traditional and squared form. Attachment of the fifth parameter (${}^1\chi$ in squared form) brings only marginal improvement of the fit, while at the same time slightly decreasing the stability of the model ($R^2 = 0.937$, $R^2_{\text{crossval}} = 0.922$).

**Density.** In the previous section, we demonstrated that molar volume can be satisfactorily estimated and predicted from correlation equations based on molecular descriptors of different types. Density values can be easily derived from the molar volume values by the simple formula $d = \text{MW}/V_m$. In their book[17] Kier and Hall also showed that descriptor ${}^1\chi$ correlates with density values much better if taken in the reciprocal form $1/{}^1\chi$ (set of alkanes, $n = 82$, $R^2 = 0.815$).

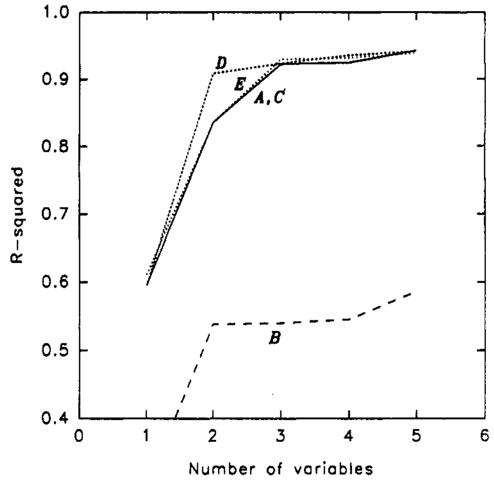

**Figure 10.** Dependence of $R^2$ on model size (structures, aldehydes;[44] property, density).

It seemed of interest to analyze if there is a linear, not reciprocal, dependence of the density value on molecular structure. For this purpose, the five subsets of descriptors (*A*–*E*) derived from the structures of aldehydes ($n = 59$), amines ($n = 109$), and ketones ($n = 60$) taken from refs 44 and 45 were correlated with density values.

Our study showed that density values usually correlate well with molecular structures if multiparameter correlation equations are considered. In most cases, for the best monoparameter correlation equations $R^2$ falls in the range 0.5–0.6. However, already at the level of the two-parameter regression model we obtained a good linear equation, and at the level of the five-parameter regression model we sometimes obtained even better correlation for density than for molar volume.

*Aldehydes* (*Figure 10*). No single descriptor from subset *A* (NRD = 19) correlates well with density: the best $R^2$ value is only 0.595 ($R^2_{\text{crossval}} = 0.555$) for the correlation equation based on the information index ${}^0$EPD. Apparently, for one-parameter estimation of density values the shape index ${}^1\kappa$ should be used in reciprocal form $1/{}^1\kappa$ or, even better, MW/${}^1\kappa$. However, a combination of information index ${}^0$EPD with connectivity descriptor ${}^3\chi^v$ much improves the correlation ($R^2 = 0.834$, $R^2_{\text{crossval}} = 0.803$). The best three-parameter regression model ($R^2 = 0.922$, $R^2_{\text{crossval}} = 0.896$) includes a combination of connectivity descriptor ${}^1\chi^v$ with two information indices (${}^0$IC and ${}^0$SIC). Progressing to the level of a four-parameter regression model, the correlation equation was not improved significantly (descriptors ${}^1\chi^v$, ${}^0$EPD, ${}^0$SEPD, and ${}^0$CIC; $R^2 = 0.924$, $R^2_{\text{crossval}} = 0.854$). Even the best five-parameter regression ($R^2 = 0.942$, $R^2_{\text{crossval}} = 0.899$), which also includes the Balaban index $J$, is inferior to the single-parameter equation obtained for the molar volume in the set of aldehydes.

Subset *B* (NRD = 30) correlates only very poorly with the density values, and the $R^2$ values for the best one- to five-parameter regression models obtained belong to the range 0.296–0.586. Therefore, it was not surprising that the combination set *C* (NRD = 48) showed no advantage over subset *A* and provided exactly the same correlations as subset *A*.

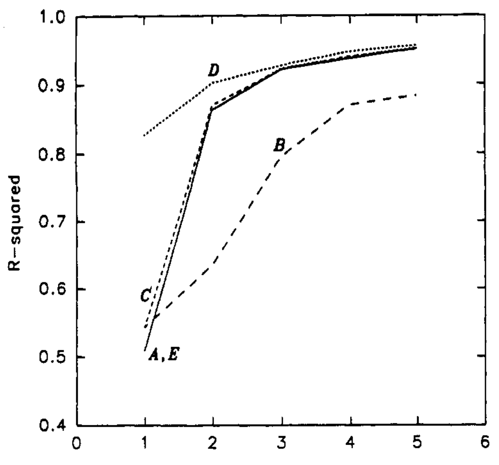

**Figure 11.** Dependence of $R^2$ on model size (structures, amines;[45] property, density).

An exhaustive search for the best single-parameter regression model in subset *D* (NRD = 26) showed that $^{0}$EPD is still the best, yet poorly correlated, descriptor. However, the combination of this information index with the connectivity descriptor $^{3}\chi^{v}$ normalized by the number of skeletal bonds ($^{3}\chi^{v}/N_b$) gives a two-parameter correlation equation ($R^2$ = 0.908, $R^2_{\text{crossval}}$ = 0.885) which is the best of all the two-parameter equations obtained for subsets *A–E*. Addition of another normalized connectivity descriptor $^{2}\chi/N_b$ to this combination provides the best three-parameter regression model in subset *D* ($R^2$ = 0.923, $R^2_{\text{crossval}}$ = 0.894). The best four-parameter correlation equation ($R^2$ = 0.935, $R^2_{\text{crossval}}$ = 0.901) includes the three connectivity descriptors ($^{0}\chi$, $^{2}\chi$, $^{2}\chi^{v}$—all in normalized form) plus the information index $^{0}$EPD. Further slight improvement of the model was achieved with addition of the shape index $^{3}\kappa$ (five-parameter equation: $R^2$ = 0.941, $R^2_{\text{crossval}}$ = 0.915).

The results obtained for subset *E* (NRD = 27) confirm that the information index $^{0}$EPD is important for structure–density correlation in this data set. This index, now in its squared form, is again the best for the single-parameter regression model ($R^2$ = 0.611, $R^2_{\text{crossval}}$ = 0.572). The two-parameter correlation equation generated for subset *E* is the same as that for subset *A*; i.e. at the level of the two-parameter regression model addition of descriptors in squared form brought no advantage. By contrast, at the level of the three-parameter regression model, a combination of two forms (traditional and squared) of the same connectivity descriptor $^{0}\chi^{v}$ with the information index $^{0}$CEPD, gives the best and the most stable three-parameter correlation equation ($R^2$ = 0.929, $R^2_{\text{crossval}}$ = 0.914). The optimum four-parameter regression model, which includes the information indices $^{0}$IC (in squared form), $^{0}$EPD, shape index $^{3}\kappa$, and connectivity descriptor $^{1}\chi^{v}$, is less stable, although it gives a slightly better fit ($R^2$ = 0.931, $R^2_{\text{crossval}}$ = 0.903). Addition of the fifth descriptor $^{0}$SEPD (in its squared form) decreases the stability even further ($R^2$ = 0.938, $R^2_{\text{crossval}}$ = 0.854).

*Amines* (*Figure 11*). The best single-parameter correlation (descriptor $^{3}\chi$) obtained for subset *A* (NRD = 22) is even worse than that obtained for the set of aldehydes ($R^2$ = 0.509, $R^2_{\text{crossval}}$ = 0.471). However, at the level of the two-parameter regression model ($R^2$ = 0.863, $R^2_{\text{crossval}}$ = 0.856), the correlation equation obtained for the set of amines is better and more stable than that for the aldehydes. This equation is based on a combination of the Randić index $^{1}\chi$ with the shape index $^{1}\kappa$. Addition of the information index $^{0}$EPD improves both the fit and stability of the regression model (three-parameter equation: $R^2$ = 0.921, $R^2_{\text{crossval}}$ = 0.902). Further improvement of the model was achieved with the inclusion of the Balaban index *J* (four-parameter equation: $R^2$ = 0.937, $R^2_{\text{crossval}}$ = 0.919). This four-parameter equation is better than that obtained for the set of aldehydes. Addition of the fifth independent variable (information index $^{1}$SIC), which improves both fit and stability ($R^2$ = 0.951, $R^2_{\text{crossval}}$ = 0.940), also gives a better five-parameter regression model, in comparison with the equations obtained for the set of aldehydes.

Although for amines subset *B* (NRD = 30) provides fairly good correlations at the level of four- and five-parameter regression models ($R^2$ = 0.870–0.883, $R^2_{\text{crossval}}$ = 0.830–0.844), the results are still much inferior to those for subset *A*. The descriptors which appear in the best correlations obtained for subset *B* include MV, $^{c}T^{E}$, FPSA3, and WPSA3. A single-parameter regression model, based on the descriptor FNSA2, is slightly better ($R^2$ = 0.543, $R^2_{\text{crossval}}$ = 0.520) than an equation based on the connectivity descriptor $^{3}\chi$ (see above).

Consequently, the combination set *C* (NRD = 51) provided the best single-parameter regression model based on the descriptor FNSA2. The best two-parameter correlation equation ($R^2$ = 0.870, $R^2_{\text{crossval}}$ = 0.859) combines descriptors $^{1}\chi$ and MV and is marginally preferable to the two-parameter equation found for subset *A*. At the level of a three-parameter regression model, the combination set showed no advantage over subset *A*: exactly the same regression model was found. The combination of the three connectivity descriptors ($^{1}\chi$, shape index $^{1}\kappa$, and information index $^{0}$SIC) with the nontopological descriptor FNSA1 gives the best four-parameter correlation equation ($R^2$ = 0.940, $R^2_{\text{crossval}}$ = 0.931). Addition of a fifth descriptor (FPSA3) gives a five-parameter regression model which is slightly better than that for subset *A* ($R^2$ = 0.950, $R^2_{\text{crossval}}$ = 0.941).

Subset *D* (NRD = 30) provides a surprisingly good single-parameter regression model ($R^2$ = 0.827, $R^2_{\text{crossval}}$ = 0.820) with the connectivity descriptor $^{0}\chi^{v}$ in normalized form. The combination of this descriptor with $^{1}\chi$, also in normalized form, gives the best two-parameter correlation equation ($R^2$ = 0.902, $R^2_{\text{crossval}}$ = 0.892). The best three-parameter regression model ($R^2$ = 0.927, $R^2_{\text{crossval}}$ = 0.918) includes descriptors $^{3}\chi$, $^{0}\chi^{v}/N_a$, and $\Phi$ and is the best three-parameter equation throughout all five subsets of descriptors. The combination of three connectivity descriptors ($^{1}\chi$, $^{1}\chi/N_b$, and $^{0}\chi^{v}$) with information index $^{0}$EPD gives the best four-parameter regression model ($R^2$ = 0.947, $R^2_{\text{crossval}}$ = 0.935), again throughout all five subsets of descriptors. Addition of the normalized Wiener Index ($W/N_a$) improves the fit but slightly decreases the stability of the model ($R^2$ = 0.956, $R^2_{\text{crossval}}$ = 0.932).

The results obtained for subset *E* (NRD = 32) are almost the same as for subsets *A* and *C*. The one- and two-parameter regression models are exactly the same as for subset *A*. Very little improvement, in comparison with subset *A*, was achieved at the level of the three-parameter regression model ($R^2$ = 0.923, $R^2_{\text{crossval}}$ = 0.905), where the information index $^{0}$EPD was included in its squared form, in combination with descriptors $^{1}\chi$ and $^{1}\kappa$ (cf.: the two-parameter equation in subset *A* also contains $^{1}\chi$ and $^{1}\kappa$, but $^{0}$EPD is included in the traditional, nonsquared form). The best four- and five-parameter equations generated in subset *E* are again very similar to those in subset *A*, with the exception that descriptor $^{0}$EPD is included, in the former case, in the squared form. The excellent fit and high stability of the five-parameter regression

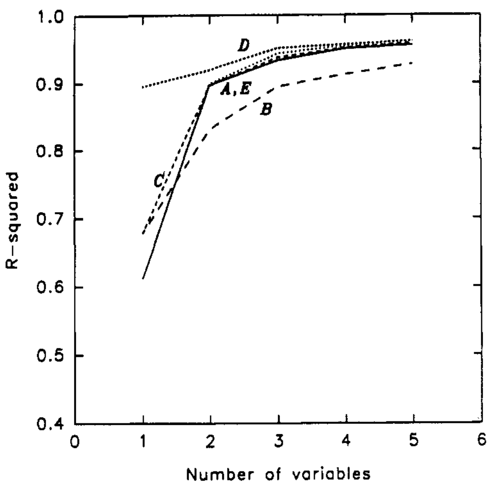

**Figure 12.** Dependence of $R^2$ on model size (structures, ketones;[44] property, density).

model should be pointed out: $R^2 = 0.953$ and $R^2_{\text{crossval}} = 0.943$.

*Ketones* (*Figure 12*). Similarly to the sets of aldehydes and amines, subset $A$ (NRD = 24) shows only a poor one-parameter correlation ($R^2 = 0.612$, $R^2_{\text{crossval}} = 0.586$) with a topological descriptor (Balaban index $J$), whereas the two-parameter regression model (descriptors ${}^{3}\chi$ and ${}^{0}$EPD) brings significant improvement: $R^2 = 0.897$, $R^2_{\text{crossval}} = 0.883$. The combination of the connectivity descriptor ${}^{1}\chi^{v}$ with the two information indices ${}^{0}$IC and ${}^{0}$SIC gives the best three-parameter correlation equation ($R^2 = 0.934$, $R^2_{\text{crossval}} = 0.923$). The combination of the three information indices (${}^{0}$EPD, ${}^{0}$SEPD, and ${}^{1}$SEPD) with connectivity descriptor ${}^{3}\chi$ improves both the fit and stability of the regression model ($R^2 = 0.951$, $R^2_{\text{crossval}} = 0.939$), whereas addition of fifth independent variable (${}^{1}$CIC) does not improve the model significantly ($R^2 = 0.956$, $R^2_{\text{crossval}} = 0.948$).

Subset $B$ (NRD = 26) gives a slightly better single-parameter correlation ($R^2 = 0.678$, $R^2_{\text{crossval}} = 0.658$) with the descriptor MPC ("maximum positive charge") than any one-parameter regression model obtained for subset $A$. In general, the homogeneous set of ketones correlates better with subset $B$ of nontopological descriptors than do the more diverse sets of aldehydes and amines. Nevertheless, even the best five-parameter regression model obtained in subset $B$ performs significantly less satisfactorily than three-parameter equations found for subsets $A$, $D$, and $E$.

The results obtained for the combination set $C$ (NRD = 48) at the level of the one- and two-parameter regression model correspond to the best one-parameter equation in subset $B$ and the best two-parameter equation in subset $A$. The combination of the Randić index ${}^{1}\chi$ with the Wiener index $W$ and RPCG gives the best three-parameter regression model ($R^2 = 0.938$, $R^2_{\text{crossval}} = 0.928$), which is almost of the same quality as the three-parameter equation found for subset $A$. The same applies to the four-parameter equation: although it includes descriptors (${}^{1}\chi$, ${}^{0}$SIC, FPSA1, and MV) which are different from those in the four-parameter model for subset $A$, the quality of the model is almost the same: $R^2 = 0.952$, $R^2_{\text{crossval}} = 0.935$. Addition of the Balaban index $J$ gives a better fit, but a lower stability of the five-parameter equation ($R^2 = 0.959$, $R^2_{\text{crossval}} = 0.944$), compared with the five-parameter model obtained in subset $A$.

Just as for the amines, subset $D$ (NRD = 31) of topological descriptors in normalized form gives a very good ($R^2 = 0.895$, $R^2_{\text{crossval}} = 0.886$) single-parameter correlation with the normalized connectivity descriptor ${}^{0}\chi^{v}/N_a$. The two-parameter regression model (descriptors ${}^{1}\chi/N_b$ and ${}^{0}\chi^{v}/N_a$; $R^2 = 0.920$, $R^2_{\text{crossval}} = 0.908$) and the three-parameter equation (descriptors $W/N_a$, ${}^{1}\chi$, and ${}^{0}$EPD; $R^2 = 0.952$, $R^2_{\text{crossval}} = 0.944$) are also the best throughout all five subsets $A$–$E$. Addition of the fourth independent variable ${}^{0}\chi^{v}/N_a$ slightly improves both the fit and the stability (four-parameter regression model; $R^2 = 0.957$, $R^2_{\text{crossval}} = 0.945$). Further improvement of the model was achieved with attachment of a fifth parameter, namely, the Balaban index $J$ (five-parameter model; $R^2 = 0.962$, $R^2_{\text{crossval}} = 0.952$). It is noteworthy that, for the set of ketones, the four- and five-parameter regression models obtained for the density estimation are better than those generated for the estimation of the molar volume. Hence, in the multiparameter equation, density can sometimes be expressed better as a sum of linear terms than as a reciprocal.

Subset $E$ (NRD = 30) provides slightly better correlations than subset $A$ at the level of two-, three-, and four-parameter regression models, but at the level of a five-parameter model the quality of the best equations generated in subsets $A$ and $E$ is almost the same. The best two-parameter correlation equation ($R^2 = 0.899$, $R^2_{\text{crossval}} = 0.889$) includes the connectivity descriptor ${}^{3}\chi$ and the information index ${}^{0}$EPD (the latter in squared form). The best three-parameter regression model is based on a combination of the connectivity descriptor ${}^{1}\chi$, with the squared shape index ${}^{1}\kappa$ and the squared information index ${}^{0}$EPD. The four-parameter correlation equation shows both excellent fit and high stability ($R^2 = 0.955$, $R^2_{\text{crossval}} = 0.947$). In this equation, two descriptors are included in their squared forms (${}^{3}\chi$ and ${}^{1}$SEPD), in combination with two information indices of zero order: ${}^{0}$EPD and ${}^{0}$SEPD. Addition of a fifth descriptor (${}^{1}$SIC squared) brought no significant improvement to the model (five-parameter equation: $R^2 = 0.957$, $R^2_{\text{crossval}} = 0.946$).

**Refractive Index.** Unlike its derivative property, molar refraction, the refractive index itself so far has not been extensively studied by QSAR/QSPR. Therefore, in this part of our investigation, our purpose was 2-fold: both to study the behavior of the different types of descriptors and to find out if the refractive index can be estimated and/or predicted from molecular structure. In general, we found that refractive index correlates in a manner somewhat similar to density, but different from molecular volume and boiling point. Apparently, the refractive index cannot be modeled with a single-parameter model, but it can be satisfactorily or even very well predicted using two-, three-, and four-parameter correlation equations. The data sets used in this study are as follows: aldehydes ($n = 60$), amines ($n = 110$), and ketones ($n = 59$); all data were taken from refs 44 and 45.

*Aldehydes* (*Figure 13*). Subset $A$ (NRD = 19) provides a very poor correlation ($R^2 = 0.597$, $R^2_{\text{crossval}} = 0.564$) at the level of a single-parameter regression model with connectivity descriptor ${}^{3}\chi$. However, progressing to the two-parameter model, the correlation improved markedly. The best two-parameter equation, which is based on the combination of Randić index ${}^{1}\chi$ with the information index ${}^{0}$CIC, has $R^2 = 0.882$ and $R^2_{\text{crossval}} = 0.862$. Combination of the connectivity descriptor ${}^{3}\chi$ with two information indices ${}^{2}$IC and ${}^{3}$EPD improves both the fit and stability, but to a smaller extent ($R^2 = 0.903$, $R^2_{\text{crossval}} = 0.882$). The best four-parameter regression model ($R^2 = 0.919$, $R^2_{\text{crossval}} = 0.894$) includes two connectivity descriptors (${}^{1}\chi$ and ${}^{3}\chi$) and two information indices of the first order (${}^{1}$IC and ${}^{1}$SIC). Attachment of the fifth descriptor (${}^{2}$IC) gives an excellent five-parameter regression model with $R^2 = 0.931$ and $R^2_{\text{crossval}} = 0.908$.

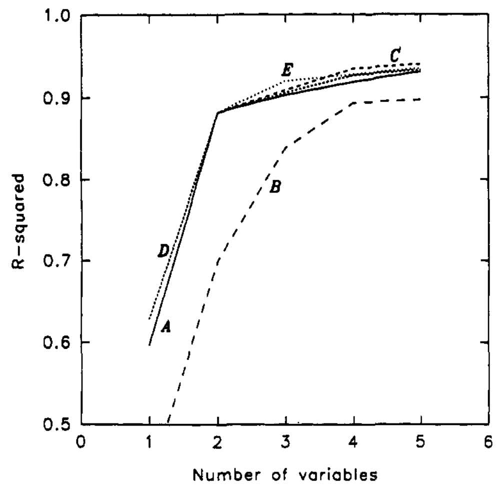

**Figure 13.** Dependence of $R^2$ on model size (structures, aldehydes;[44] property, refractive index).

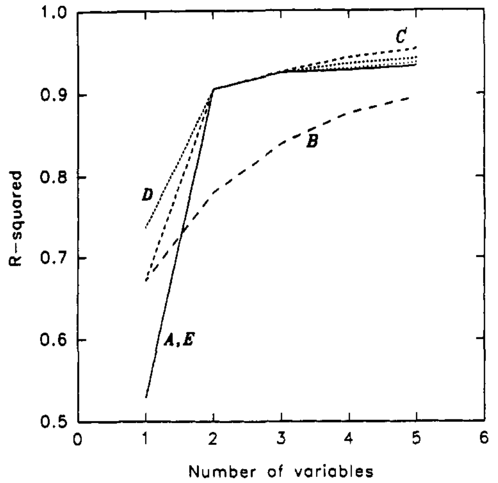

**Figure 14.** Dependence of $R^2$ on model size (structures, amines;[45] property, refractive index).

The best single-parameter correlation obtained with the descriptor MPC (maximum positive charge) in subset $B$ (NRD = 31) is even worse than that for subset $A$: $R^2 = 0.431$, $R^2_{\text{crossval}} = 0.396$. In other words, no relationship was detected at the level of a monoparameter model. The combination of two shadow indices $S_1$ and $S_2$ gives a fair regression model with $R^2 = 0.701$ and $R^2_{\text{crossval}} = 0.669$. Only at the level of a three-parameter regression model, a combination of two CPSA descriptors (WPSA1 and WPSA3) with the shadow index $S_1$ was a fairly good correlation equation obtained: $R^2 = 0.839$ and $R^2_{\text{crossval}} = 0.808$. Further improvement of this regression model was achieved with attachment of the second shadow index $S_2$ (four-parameter equation: $R^2 = 0.893$, $R^2_{\text{crossval}} = 0.871$) and then with addition of a fifth independent variable FNSA1 (five-parameter equation: $R^2 = 0.897$, $R^2_{\text{crossval}} = 0.860$).

The combination set $C$ (NRD = 50) shows no advantage over subset $A$ at the level of one- and two-parameter regression models: exactly the same correlation equation was generated as for subset $A$. At the level of a three-parameter regression model, a very minor improvement was achieved, when the model included the connectivity descriptor $^3\chi$, the information index $^0$CIC, and the shadow index $S_1$ ($R^2 = 0.909$, $R^2_{\text{crossval}} = 0.890$). However, the combination of the Randić index $^3\chi$ with two information indices ($^1$IC and $^2$CIC) and shadow index $S_1$ boosted the quality of correlation significantly: $R^2 = 0.935$ and $R^2_{\text{crossval}} = 0.916$. This four-parameter equation is the best throughout all five subsets $A$–$E$. Further minor improvement of correlation ($R^2 = 0.940$, $R^2_{\text{crossval}} = 0.921$) was achieved with addition of a fifth parameter (descriptor DIST).

The subset $D$ (NRD = 27) provides a slightly better single-parameter regression model ($R^2 = 0.629$, $R^2_{\text{crossval}} = 0.577$) with the normalized descriptor $^3\chi/N_b$ than those obtained in subsets $A$ and $B$. Nevertheless, the best two-parameter correlation equation is exactly the same as for subset $A$, with no improvement from addition of normalized descriptors. The best three-parameter regression model ($R^2 = 0.906$, $R^2_{\text{crossval}} = 0.887$) which includes the normalized descriptor $^3\chi^v/N_b$, in addition to a combination of $^1\chi$ with $^0$CIC, is of almost the same quality as the three-parameter models generated in subsets $A$ and $C$. The best four-parameter correlation equation (descriptors $^1\chi$, $^1\chi^v$, $^3\chi^v/N_b$, and $^1$IC; $R^2 = 0.927$ and $R^2_{\text{crossval}} = 0.905$) is substantially better than the four-parameter model found in subset $A$, but not so good as that generated from the combination set $C$. Addition of the fifth descriptor (information index $^1$SIC) yields a five-parameter correlation equation ($R^2 = 0.933$, $R^2_{\text{crossval}} = 0.909$) of approximately the same quality as the five-parameter model in subset $A$.

Subset $E$ (NRD = 27) shows its advantage over subsets $A$ and $C$ at the level of a three-parameter regression model. The single-parameter equation with descriptor $^3\chi$ is exactly the same as that for subsets $A$ and $C$. The best two-parameter model, although slightly different from that in subset $A$ (information index $^0$CIC is included in the squared form), still has the same values of $R^2$ and $R^2_{\text{crossval}}$. However, the three-parameter correlation equation ($R^2 = 0.920$, $R^2_{\text{crossval}} = 0.902$), generated in subset $E$, is the best throughout all five subsets $A$–$E$. This equation is based on the Randić index $^1\chi$ in linear form and the two information indices ($^1$CEPD, $^1$IC) in squared forms. Addition of another connectivity descriptor $^3\chi$ gives the four-parameter equation ($R^2 = 0.928$, $R^2_{\text{crossval}} = 0.908$) which is slightly inferior to the four-parameter regression model obtained in the combination set $C$. Further addition of another information index ($^3$CEPD) improves both the fit and the stability of the model so that it becomes of almost the same quality ($R^2 = 0.935$, $R^2_{\text{crossval}} = 0.914$) as that generated from the combination set $C$.

*Amines* (*Figure 14*). Similarly to the set of aldehydes, it was shown for the set of amines that the Randić index $^3\chi$ is again the most suitable variable to be included in a monoparameter regression. However, the correlation obtained in subset $A$ (NRD = 22) is still very poor: $R^2 = 0.529$ and $R^2_{\text{crossval}} = 0.486$. The best two-parameter equation obtained for the set of amines is of much higher quality ($R^2 = 0.906$, $R^2_{\text{crossval}} = 0.894$). Notably, this equation is based on the same two descriptors ($^1\chi$ and $^0$CIC) as the best two-parameter model generated for aldehydes. The combination of connectivity descriptor $^1\chi$, shape index $^1\kappa$, and information index $^1$CIC makes the best three-parameter regression model generated in subset $A$ ($R^2 = 0.926$, $R^2_{\text{crossval}} = 0.914$). The best four-parameter equation which includes two connectivity descriptors ($^1\chi$ and $^3\chi$) and two information indices ($^0$CIC and $^1$SEPD) does not show any significant improvement of model quality ($R^2 = 0.929$, $R^2_{\text{crossval}} = 0.920$). Addition of a fifth independent variable (descriptor $^1\kappa$) brings a notable improvement in both the fit and stability of the model: $R^2 = 0.934$ and $R^2_{\text{crossval}} = 0.924$.

Although subset $B$ (NRD = 31) shows a slightly better single-parameter correlation with descriptor WNSA1 ($R^2 = 0.671$, $R^2_{\text{crossval}} = 0.643$), all multiparameter regression models generated in this subset are significantly worse than those obtained in subsets *A*, *D*, and *E*. Values of $R^2$ and $R^2_{\mathrm{crossval}}$ fall in the ranges 0.780–0.896 and 0.758–0.856, respectively. The best correlations obtained are based almost exclusively on the CPSA descriptors; from other types of descriptors, only one geometrical descriptor MV is included in the best four- and five-parameter regression models. The best five-parameter correlation ($R^2 = 0.896$, $R^2_{\mathrm{crossval}} = 0.696$), although providing a fairly good fit, is insufficiently stable. The stability of this model decreased after the fifth variable (DPSA2) was added to the best four-parameter model (descriptors PNSA3, DPSA1, DPSA3, and MV; $R^2 = 0.877$, $R^2_{\mathrm{crossval}} = 0.856$).

As one should expect, the combination set *C* (NRD = 49) provided the best single-parameter model identical to that from subset *B*, whereas the best two-parameter correlation equation repeats that from subset *A*. At the level of a three-parameter regression model, the nontopological descriptor PNSA1 showed up in a model which includes also connectivity descriptor $^{1}\chi$ and shape index $^{1}\kappa$. However, the quality of this equation ($R^2 = 0.927$, $R^2_{\mathrm{crossval}} = 0.914$) is virtually the same as that of the three-parameter regression generated in subset *A*. Significant improvement of the model ($R^2 = 0.944$, $R^2_{\mathrm{crossval}} = 0.933$) was achieved when the Randić index $^{3}\chi$ was combined with the three CPSA descriptors (FPSA3, WNSA1, and WNSA2). This correlation is the best four-parameter equation throughout all five subsets *A*–*E*. Attachment of the fifth independent variable (electronic–topological descriptor $^{2}E^{T}$) makes a five-parameter regression model of excellent quality and stability: $R^2 = 0.954$ and $R^2_{\mathrm{crossval}} = 0.945$.

As we observed in the correlations with density, for the set of amines, the normalized descriptor $^{0}\chi^{v}/N_{b}$ alone provides a fair correlation ($R^2 = 0.736$, $R^2_{\mathrm{crossval}} = 0.722$) with refractive index too. However, at the level of two- and three-parameter regression models, normalized topological descriptors bring no advantage over topological descriptors in the traditional form (subset *A*). The best correlation equations with two and three independent variables, generated for subset *D* (NRD = 30), are exactly the same as those obtained in subset *A*. Notably, the best four-parameter regression model, which includes descriptors $^{1}\chi$, $^{3}\chi$, $^{0}\chi^{v}/N_{a}$, and $^{1}$CIC, is of substantially higher quality ($R^2 = 0.937$, $R^2_{\mathrm{crossval}} = 0.928$) than the best four-parameter equation generated in subset *A*. Further significant improvement of the model quality ($R^2 = 0.943$, $R^2_{\mathrm{crossval}} = 0.935$) was achieved at the level of a five-parameter regression, when the information index $^{0}$IC was included.

The results obtained for subset *E* (NRD = 33) are of almost the same quality as those generated for subset *A*. The best one- and three-parameter correlation equations are exactly the same as those generated in subset *A*. The best two-parameter equation includes Randić index $^{1}\chi$ and information index $^{0}$CIC (the latter in squared form), but the quality of the model ($R^2 = 0.906$, $R^2_{\mathrm{crossval}} = 0.894$) is the same as that for the two-parameter equation in subset *A*. The combination of the two connectivity descriptors ($^{1}\chi$ and $^{3}\chi$) with two information indices ($^{1}$CEPD and $^{1}$IC, both in squared form) makes the best four-parameter equation which, again, is of the almost same quality ($R^2 = 0.931$, $R^2_{\mathrm{crossval}} = 0.918$) as the four-parameter model in subset *A*. Addition of a fifth parameter (squared connectivity descriptor $^{0}\chi$) improves both the fit and stability of the model ($R^2 = 0.938$, $R^2_{\mathrm{crossval}} = 0.930$), although the superiority of this equation over the one from subset *A* is marginal.

*Ketones* (*Figure 15*). The set of ketones correlates with the refractive index better than the sets of aldehydes and amines. Already at the level of a single-parameter regression model, a good correlation ($R^2 = 0.832$, $R^2_{\mathrm{crossval}} = 0.821$) with the connectivity descriptor $^{3}\chi$ was obtained in subset *A* (NRD = 24). At the level of a two-parameter regression model, the combination of descriptors $^{1}\chi$ and $^{0}$CIC again makes the best ($R^2 = 0.951$, $R^2_{\mathrm{crossval}}$) 2-parameter equation. In our study, this particular combination of descriptors showed up in the best two-parameter equations in all three different sets of structures: aldehydes, amines, and ketones. Apparently, the combination of descriptors $^{1}\chi$ and $^{0}$CIC indeed reflects some regularity in the relationship between the structure and refractive index. Attachment of the Wiener index *W* to this combination gives the best three-parameter regression model ($R^2 = 0.972$, $R^2_{\mathrm{crossval}} = 0.968$). Further significant improvement of the model was achieved by a combination of descriptors *W*, $^{1}\chi$, $^{1}$CIC, and $^{2}$IC in the correlation equation ($R^2 = 0.980$, $R^2_{\mathrm{crossval}} = 0.976$). Addition of a fifth independent variable (information index $^{1}$CEPD) left the quality of the model almost unchanged ($R^2 = 0.982$, $R^2_{\mathrm{crossval}} = 0.976$).

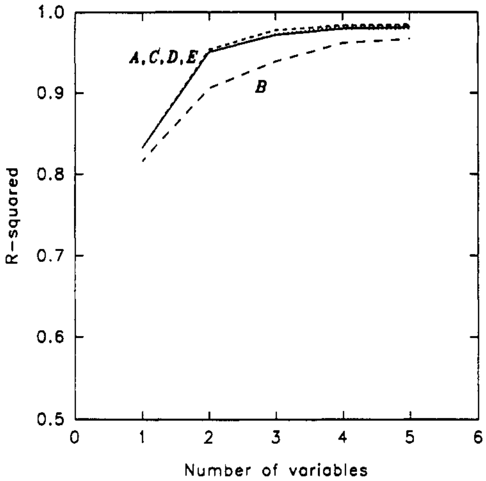

**Figure 15.** Dependence of $R^2$ on model size (structures, ketones;[44] property, refractive index).

Subset *B* (NRD = 23) also provides a single-parameter correlation of reasonable quality (descriptor FNSA2; $R^2 = 0.816$, $R^2_{\mathrm{crossval}} = 0.805$), although inferior to the monoparameter model obtained in subset *A*. Just as for subset *A*, the quality of the regression model was markedly improved by increasing the number of parameters from 1 to 4 ($R^2$ changed from 0.816 to 0.962 and $R^2_{\mathrm{crossval}}$ changed from 0.805 to 0.956). Then, again similarly to subset *A*, almost no change was observed when the regression model was upgraded from four to five parameters (five-parameter equation: $R^2 = 0.967$, $R^2_{\mathrm{crossval}} = 0.956$).

The combination set *C* (NRD = 45) gives, as expected, the same best monoparameter equation as subset *A*. At the level of the two-parameter regression model, the combination of the Randić index $^{1}\chi$ with PPSA2 gives a very good correlation equation ($R^2 = 0.954$, $R^2_{\mathrm{crossval}} = 0.947$) which is slightly better than the one generated in subset *A*. The best three-parameter model ($R^2 = 0.978$, $R^2_{\mathrm{crossval}} = 0.974$) includes the Wiener index *W*, the Randić index $^{1}\chi$, and the combined descriptor FPSA2. This model is indeed the best of all the three-parameter equations obtained for the five subsets *A*–*E*. The best four-parameter regression model ($R^2 = 0.984$, $R^2_{\mathrm{crossval}} = 0.980$) includes descriptors $^{1}\chi$, WNSA1, WNSA2, and TSASA. Only a very minor improvement of this model ($R^2 = 0.985$, $R^2_{\mathrm{crossval}} = 0.981$) was achieved when the shape index $^{1}\kappa$ was included in the correlation equation as the fifth independent variable.

The normalized topological descriptors did not show up in any of the best correlation equations generated for subset *D*

(NRD = 31). In other words, all the best one- to five-parameter regression models obtained in subset *D* are exactly the same as those found for subset *A*.

As we observed for the sets of aldehydes and amines, the same descriptors ($^1\chi$ and $^0$CIC) are included in the best two-parameter regression model in both subsets *A* and *E*. Although for subset *E* (NRD = 32) the information index is included in the squared form, the quality of both correlation equations is the same. For the set of ketones, the same descriptors are included in the best three-parameter models obtained in subsets *A* and *E*. Again, in subset *E* two descriptors (*W* and $^0$CIC) are squared, and this brings a very minor improvement of the model quality ($R^2 = 0.973$, $R^2_{\text{crossval}} = 0.969$). The superiority of subset *E* over subset *A* is pronounced at the level of the four-parameter regression model ($R^2 = 0.983$, $R^2_{\text{crossval}} = 0.979$) which includes descriptors *W*, $^1\chi$, $^2$IC, and $^1$CEPD (squared). This excellent correlation is of almost the same quality as the four-parameter model obtained in the combination set *C*. As well as for all subsets *A*–*D*, in subset *E*, the exhaustive search for the best five-parameter correlation equation gives virtually no better model ($R^2 = 0.984$, $R^2_{\text{crossval}} = 0.980$) in comparison with the four-parameter equation.

## BIOLOGICAL ACTIVITY

So far we have considered in detail the behavior of different subsets of descriptors used in the correlations between structures and five physicochemical properties (melting point, boiling point, molar volume, density, and refractive index). The potential to estimate and/or predict physicochemical properties from molecular structure is of great interest for chemists, especially in industrial research. The ability to derive from the structure the expected value of biological activity is of primary importance for chemists in the pharmaceutical industry and for researchers in medicinal chemistry. Therefore, in this section we consider the behavior of different classes of descriptors in correlations between structure and biological activity.

For our study of biological activity data, which is of primary interest for the majority of QSAR/QSPR researchers, we tried to select from the literature those data sets where good correlations with molecular descriptor(s) had already been found. We were particularly interested in correlations based exclusively on topological or nontopological descriptors which are available in our GROUND program, so that we could compare our results with those presented in the literature.

For these reasons, we selected three data sets from the a book[17] in which the authors had presented impressive correlations between properties and connectivity descriptors. Another data set, which seemed of great interest to us, was treated and described independently in two sources,[3,20] and both sources showed that the activity in the study can be well modeled with geometrical descriptors only. This prompted us to apply the aforementioned method of comparative analysis of different types of descriptors, to see if our conclusions match those in the literature.

**Nonspecific Local Aaesthetic Activity.** In their book,[17] Kier and Hall showed that the Randić index $^1\chi$ is in good correlation ($R^2 = 0.964$, $s = 0.41$, $n = 36$) with the nerve-blocking concentration of organic compounds of various structures. However, in a review,[19] Gupta noted that the original data set[7] contains 39, not 36 structures. In our study we took the original data set of 39 structures to investigate if the topological descriptors are indeed the best molecular descriptors for estimating biological activity of this type. The results were surprising: geometrical descriptors were demonstrated to correlate significantly better than topological indices (see Figure 16).

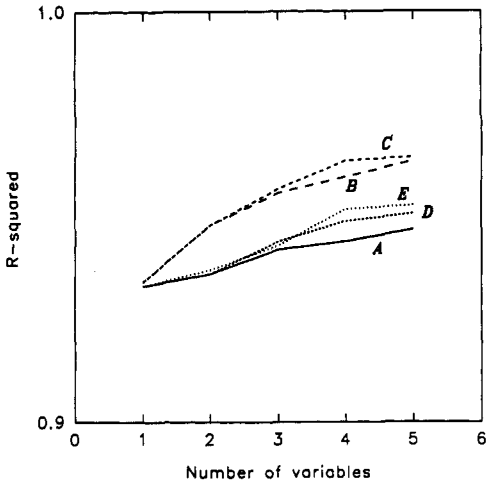

Figure 16. Dependence of $R^2$ on model size (structures,[17] property: nonspecific local anaesthetic activity).

Subset *A* (NRD = 17) indeed gives very good single-parameter correlation with the connectivity descriptor $^1\chi^v$ ($R^2 = 0.933$, $R^2_{\text{crossval}} = 0.926$). The best two-parameter regression model also included the Wiener index *W*, but no significant improvement of the model was observed ($R^2 = 0.936$, $R^2_{\text{crossval}} = 0.926$). The combination of the connectivity descriptor $^2\chi^v$ with two information indices ($^2$CIC and $^3$SEPD) brings only a minor improvement ($R^2 = 0.942$, $R^2_{\text{crossval}} = 0.934$). The same applies to the combination of descriptor $^2\chi^v$ with three information indices ($^2$CEPD, $^3$CEPD, and $^3$CIC): the quality of this four-parameter model ($R^2 = 0.944$, $R^2_{\text{crossval}} = 0.936$) is almost the same as that of the three-parameter equation. Attachment of the fifth parameter (information index $^1$SIC) brings virtually no change: $R^2 = 0.947$ and $R^2_{\text{crossval}} = 0.937$.

The best monoparameter correlation equation obtained for subset *B* (NRD = 29) is based on the descriptor MV (molecular volume). The quality of this regression model ($R^2 = 0.934$, $R^2_{\text{crossval}} = 0.928$) is slightly better than that of the one-parameter equation generated in subset *A*. However, already at the level of a two-parameter regression model, the correlation obtained in subset *B* outperforms ($R^2 = 0.948$, $R^2_{\text{crossval}} = 0.938$) the best five-parameter regression found for subset *A*. The best two-parameter equation generated for subset *B* includes two geometrical descriptors: molecular volume (MV) and the shadow index $S_6$. A combination of these two descriptors with DPSA3 produces the best three-parameter regression model ($R^2 = 0.956$, $R^2_{\text{crossval}} = 0.941$). The best four-parameter equation ($R^2 = 0.960$, $R^2_{\text{crossval}} = 0.943$) includes three geometrical descriptors (MV, $S_5$, and $S_6$) in combination with PPSA3. Attachment of the fifth independent variable (descriptor RNCS) improves both the fit and the stability of the model, but to a minor extent: $R^2 = 0.964$ and $R^2_{\text{crossval}} = 0.946$.

As expected, the combination set *C* (NRD = 45) shows that the best single-parameter equation is based on the descriptor MV. The best two-parameter correlation includes both a topological ($^2\chi^v$) and a geometrical descriptor ($S_1$), but the quality of the regression model ($R^2 = 0.948$, $R^2_{\text{crossval}} = 0.938$) is the same as that for the best two-parameter equation in subset *B*. The best three-parameter model ($R^2 = 0.957$, $R^2_{\text{crossval}} = 0.942$), which includes the connectivity descriptor $^2\chi^v$, information index $^0$CIC, and geometrical descriptor $S_6$, is marginally superior to the three-parameter

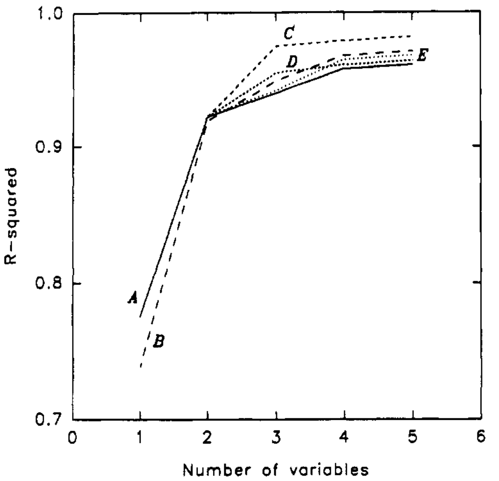

**Figure 17.** Dependence of $R^2$ on model size (structures,[17] property: narcosis of *Arenicola* larvae).

equation in subset *B*. The combination of two connectivity descriptors ($^1\chi$ and $^2\chi^v$) with two geometrical ones (MV and $S_6$) yields the best four-parameter regression model which is now significantly better ($R^2 = 0.946$, $R^2_{\text{crossval}} = 0.946$) than that in subset *B*. Addition of another geometrical descriptor ($S_5$) to the correlation equation leaves the model almost unchanged ($R^2 = 0.965$, $R^2_{\text{crossval}} = 0.946$).

Subset *D* (NRD = 23) shows no advantage over subset *A* at the level of the one- and two-parameter regression models: the same correlations are obtained as in subset *A*. The best three-parameter correlation equation (descriptors $^2\chi^v$, $^0\chi^v/N_a$, and $^2\kappa$; $R^2 = 0.944$, $R^2_{\text{crossval}} = 0.932$), although including two descriptors different from those in the three-parameter equation in subset *A*, is of the same quality. The best four-parameter regression model, which includes descriptors $^2\chi^v$, $^0\chi^v/N_a$, $^1\kappa$, and $^0$SIC, has a slightly better fit but a lower stability ($R^2 = 0.949$, $R^2_{\text{crossval}} = 0.919$) than the four-parameter model in subset *A*. The same applies to the five-parameter correlation equation (additional descriptor $^2\chi$; $R^2 = 0.951$, $R^2_{\text{crossval}} = 0.918$).

The best single-parameter equation generated in subset *E* (NRD = 21) is the same as for subset *A*. Here, the addition of the Wiener index in squared form improves the correlation, although to a minor extent ($R^2 = 0.937$, $R^2_{\text{crossval}} = 0.929$). The best three-parameter regression model, which is based exclusively on connectivity descriptors $^1\chi$, $^2\chi$, and $^2\chi^v$ (the latter two both in squared forms) provides a slightly better fit but significantly lower stability in comparison with the three-parameter model found for subset *A*. A minor advantage of subset *E* over subset *A* is noticed only at the level of a four-parameter equation ($R^2 = 0.952$, $R^2_{\text{crossval}} = 0.929$), which is again based exclusively on connectivity descriptors ($^1\chi$ in linear form; $^2\chi$, $^2\chi^v$, and $^3\chi^v$ in squared form). However, this model is also less stable than the four-parameter equation in subset *A*. An addition of the linear term $^3\chi^v$ decreases the stability further, whereas the fit remains almost the same (five-parameter equation: $R^2 = 0.953$, $R^2_{\text{crossval}} = 0.912$).

**Narcosis of *Arenicola* Larvae.** The results obtained in the previous section showed that geometrical descriptors can play a more important role in the estimation/prediction of biological activity than for physicochemical properties. Therefore, we extended our study of correlation between structures and nonspecific narcotic activity to the varied set of 20 compounds from ref 17 (see Figure 17). For this data set, Kier and Hall reported a fairly good correlation ($R^2 = 0.797$, $s = 0.48$, $n = 20$) between the connectivity descriptor $^1\chi^v$ and narcotic concentration of these compounds against the *Arenicola* larvae.

The results obtained for subset *A* of topological descriptors (NRD = 27) showed that the combination of connectivity descriptor $^1\chi^v$ with information index $^1$IC provides a correlation of significantly higher quality ($R^2 = 0.922$, $R^2_{\text{crossval}} = 0.890$) than that previously reported.[17] The combination of the connectivity descriptor $^3\chi^v$ with the information index $^1$EPD and the Balaban index *J* improves the fit but decreases the stability of the correlation (three-parameter equation: $R^2 = 0.940$, $R^2_{\text{crossval}} = 0.889$). A significant improvement in both the fit and stability of the model was achieved at the level of a four-parameter regression model, which includes descriptors $^3\chi^v$, $^1$SEPD, $^1$CEPD, and *J* ($R^2 = 0.958$, $R^2_{\text{crossval}} = 0.930$). Attachment of a fifth descriptor $^2\chi$ leaves the model almost unchanged ($R^2 = 0.961$, $R^2_{\text{crossval}} = 0.928$).

Subset *B* (NRD = 32) provides only a fair correlation with the shadow index $S_3$ ($R^2 = 0.738$, $R^2_{\text{crossval}} = 0.700$) although this is the best single-parameter regression model obtained in this subset. Similarly to subset *A*, (i) the two-parameter equation (descriptors DIST and $S_3$; $R^2 = 0.919$, $R^2_{\text{crossval}} = 0.900$) is much improved, whereas (ii) the three-parameter model has a better fit but lower stability (descriptors $^cT^E$, SAPC, and $S_3$; $R^2 = 0.949$, $R^2_{\text{crossval}} = 0.895$). At the level of a four-parameter regression model, subset *B* shows significant superiority over subset *A*. The correlation based on descriptors $^cT^E$, DPSA3, MV, and $S_3$ has both a better fit and stability: $R^2 = 0.968$ and $R^2_{\text{crossval}} = 0.943$. Just as for subset *A*, addition of a fifth variable (WPSA2) does not significantly improve the correlation ($R^2 = 0.971$, $R^2_{\text{crossval}} = 0.944$).

At the level of the one- and two-parameter regression models, the results obtained in the combination set *C* (NRD = 55) are exactly the same as those in subset *A*. However, the combination of the Randić index $^3\chi$ with the information index $^3$IC and the geometrical descriptor $S_3$ considerably boosts the quality of the model: $R^2 = 0.975$ and $R^2_{\text{crossval}} = 0.964$. Further, although not so marked, improvement of the correlation equation ($R^2 = 0.979$, $R^2_{\text{crossval}} = 0.967$) was achieved with attachment of the connectivity descriptor $^1\chi^v$ to the model. At the level of the five-parameter correlation equation with the additional descriptor $^0$EPD, one observes a decrease of the stability, whereas the fit was improved to a minor extent: $R^2 = 0.982$, $R^2_{\text{crossval}} = 0.966$.

The one- and two-parameter equations generated for subset *D* (NRD = 34) are the same as for subset *A*. At the level of three-, four-, and five-parameter regression models, the results obtained for subset *D* are similar to those generated for subset *C*, although the quality of the correlations is lower for subset *D* in comparison with the combination set *C*. The best three-parameter model (descriptors $^3\chi^v/N_b$, $^1\kappa$, and $^1$SIC; $R^2 = 0.955$, $R^2_{\text{crossval}} = 0.937$) is significantly, but not drastically, better than the two-parameter equation. As for subset *C*, attachment of the descriptor $^1\chi^v$ (four-parameter model: $R^2 = 0.961$, $R^2_{\text{crossval}} = 0.944$) and then of the descriptor *J* (five-parameter model: $R^2 = 0.964$, $R^2_{\text{crossval}} = 0.944$) brings only minor improvement. The best four- and five-parameter equations generated in subset *D* are worse than those found for subset *B*.

The best single-parameter regression model generated in subset *E* (NRD = 32) is the same as for subsets *A*, *C*, and *D*. The best two-parameter equation ($R^2 = 0.923$, $R^2_{\text{crossval}} = 0.895$) includes the connectivity descriptor $^1\chi^v$ in linear form and the information index $^1$IC in squared form (the same descriptors, both in linear form, produced the optimum two-parameter model in subsets *A* and *C*). The quality of the best three-parameter correlation, which includes descriptors $^3\chi^v$, $^1$EPD, and *J* (squared), is approximately the same ($R^2$

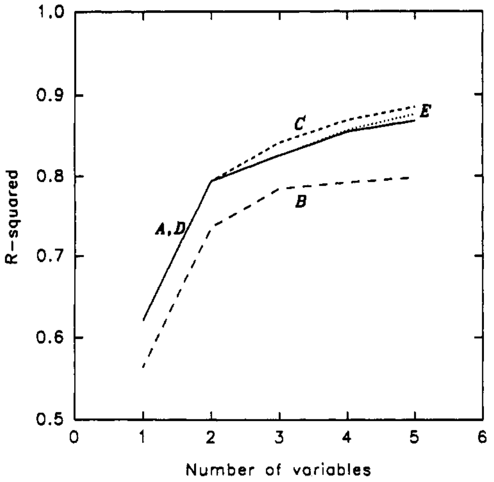

**Figure 18.** Dependence of $R^2$ on model size (structures,[17] property: frog tadpole narcosis).

$= 0.942$, $R^2_{\text{crossval}} = 0.896$) as the best three-parameter equation in subset $A$. At the level of four- and five-parameter regression models, the results obtained in subset $E$ are close in quality to the results generated for subset $B$. The best four-parameter correlation equation ($R^2 = 0.965$, $R^2_{\text{crossval}} = 0.940$) includes descriptor $^2$SIC in linear form and descriptors $^1\chi^v$, $^0$CIC, and $^1$CEPD in squared forms. Attachment of the Wiener index (squared) to the model gives an excellent five-parameter equation ($R^2 = 0.968$, $R^2_{\text{crossval}} = 0.950$), which is also the most stable regression model throughout all five subsets $A$–$E$.

**Frog Tadpole Narcosis.** The study conducted for the data sets in two preceding sections showed that nontopological descriptors may correlate more successfully than topological ones. In such a case, the combination set $C$ correlates amazingly well with the property. However, the following study demonstrates that topological descriptors still can appear as superior to nontopological ones, even when the property is a biological (in our research—narcotic) activity (see Figure 18).

We now consider another example taken from the book[17] authored by Kier and Hall. They found that the connectivity descriptors $^1\chi^v$ derived from the structures of 52 diverse organic compounds correlate reasonably well ($R^2 = 0.784$, $s = 0.54$, $n = 52$) with the effective narcotic concentration, measured in terms of the narcosis on frog tadpoles.

The best one- and two-parameter equations obtained from subset $A$ (NRD = 26) included the same descriptors ($^1\chi^v$ and $^1$IC) as the one- and two-parameter models generated for the previous data set. For this data set, however, the difference in quality between regressions with one and two parameters is not so pronounced (two-parameter model: $R^2 = 0.794$, $R^2_{\text{crossval}} = 0.763$). The best three- and four-parameter models retain the combination of topological indices $^1\chi^v$ and $^1$IC. Apart from these two descriptors, the best three-parameter equation includes the Randić index $^1\chi$ ($R^2 = 0.826$, $R^2_{\text{crossval}} = 0.795$). The best four-parameter model includes, in addition, the information index $^2$SEPD, which brings a substantial improvement to the model quality ($R^2 = 0.855$, $R^2_{\text{crossval}} = 0.819$). Attachment of a fifth independent variable (Randić index $^3\chi$) also improves both the fit and stability of the model, but to a lesser extent ($R^2 = 0.868$, $R^2_{\text{crossval}} = 0.823$).

Set $B$ (NRD = 33) of nontopological descriptors provides correlations of much lower quality than those from subset $A$. The best one-parameter regression model ($R^2 = 0.564$, $R^2_{\text{crossval}} = 0.533$) includes the shape parameter $\eta$. The optimum two-parameter equation ($R^2 = 0.736$, $R^2_{\text{crossval}} = 0.699$) is based on descriptors $T^E$ and TSASA. The combination of descriptors $^cT^E$, DPSA2, and RNCS provides the best three-parameter regression model ($R^2 = 0.784$, $R^2_{\text{crossval}} = 0.743$), which is also the most stable one in subset $B$. Further extensions of the model with the shadow index $S_4$ (four-parameter equation: $R^2 = 0.792$, $R^2_{\text{crossval}} = 0.727$) and then with descriptor RPCG (five-parameter equation: $R^2 = 0.798$, $R^2_{\text{crossval}} = 0.720$), although improving the fit, at the same time decrease the stability of the regression.

Apparently, descriptors $^1\chi^v$ and $^1$IC are of great importance for the correlation between structures and nonspecific narcotic activity. Again, one can observe that this combination of descriptors shows up in all the best two-, three-, four-, and five-parameter correlation equations generated in the combination set $C$ (NRD = 56). In addition to these two descriptors, the best three-parameter regression model ($R^2 = 0.842$, $R^2_{\text{crossval}} = 0.811$) includes also shape parameter $\eta$, which was the best single-correlated descriptor in subset $B$. The best four-parameter model ($R^2 = 0.869$, $R^2_{\text{crossval}} = 0.836$) is based on descriptors $^1\chi^v$, $^1$IC, $^2$CIC, and WPSA3. Attachment of a fifth descriptor (DIST) to this model improves both the fit and stability of the regression ($R^2 = 0.885$, $R^2_{\text{crossval}} = 0.842$). Thus, at the level of three-, four-, and five-parameter models, the combination set $C$ provides the equations of the highest quality, in comparison with those in subsets $A$, $B$, $D$, and $E$. However, the difference in quality of regressions obtained for subsets $A$ and $C$ is not as significant as that for the two preceding data sets.

The results obtained for subset $D$ (NRD = 33) are exactly the same as for subset $A$. The descriptors in normalized forms appeared to have no influence at the level of the one- to five-parameter regression models.

Subset $E$ (NRD = 32) provides exactly the same one-, two-, and three-parameter correlation equations as those obtained in subset $A$. However, the influence of the topological indices in squared form shows up in the best four-parameter regression model ($R^2 = 0.857$, $R^2_{\text{crossval}} = 0.822$) which includes descriptors $^1\chi$, $^1\chi^v$, $^1$IC, and $^2$SEPD (squared). Addition of Randić index $^3\chi$ (squared) to the model gives the five-parameter correlation equation which is almost as good ($R^2 = 0.876$, $R^2_{\text{crossval}} = 0.830$) as that obtained in the combination set $C$.

**Sweet Taste Intensity.** Iwamura[3] and later Rohrbaugh and Jurs[20] investigated a relationship between geometrical descriptors and sweet taste intensity of 20 compounds (derivatives of 3-nitroaniline and 3-cyanoaniline). The best regression model ($R^2 = 0.810$, $s = 0.32$, $n = 20$), obtained by Iwamura, included two of Verloop's STERIMOL parameters.[51] Rohrbaugh and Jurs showed that a three-parameter correlation of higher quality ($R^2 = 0.941$, $s = 0.18$, $n = 20$) can be built using shadow indices $S_1$, $S_4$, and $S_3$. On the other hand, Kier and Hall[17] succeeded in estimating the sweet taste potency of nine substituted nitroanilines with a two-parameter regression model ($R^2 = 0.908$, $s = 0.22$, $n = 9$) based exclusively on topological indices ($^1\chi$ and $^1\chi^v$). Therefore, it seemed worthwhile to compare the behavior of various types of descriptors for the estimation and prediction of the sweet taste intensity using the same set of 20 compounds as that used in refs 3 and 20.

The best monoparameter correlation equation ($R^2 = 0.752$, $R^2_{\text{crossval}} = 0.662$) generated for subset $A$ (NRD = 16), is based on the connectivity descriptor $^0\chi^v$. This correlation is inferior to those previously obtained,[3,20] although one should keep in mind that this equation includes only one independent variable. However, the best two-parameter regression model, which includes the information index $^2$EPD and connectivity descriptor $^3\chi^v$, is of higher quality ($R^2 = 0.850$, $R^2_{\text{crossval}} = 0.778$) than the model with Verloop's parameters. If we take into account that STERIMOL parameters require quite precise 3D coordinates and, to some extent, depend on conformation, then it becomes clear that the use of topological indices for modeling this property may be indeed advantageous. The combination of connectivity descriptor $^2\chi^v$ with two information indices $^2$SIC and $^2$IC gives the best three-parameter model, which is only slightly inferior ($R^2 = 0.894$, $R^2_{\text{crossval}} = 0.812$) to the correlation based on the three shadow indices.[20] The best four-parameter regression (descriptors $W$, $^1\kappa$, $^0$EPD, and $^3$CIC; $R^2 = 0.934$ and $R^2_{\text{crossval}} = 0.882$) is of high quality. Attachment of a fifth parameter (shape index $^3\kappa$), however, decreases the stability of the model ($R^2 = 0.942$, $R^2_{\text{crossval}} = 0.866$), which is perhaps to be expected when a multiparameter equation is applied to a small data set.

The results obtained for subset $B$ (NRD = 24) were surprising. We expected that geometrical descriptors would show up in the best correlations generated. At the level of the single-parameter regression model this assumption was confirmed, and the best, although quite poor ($R^2 = 0.620$, $R^2_{\text{crossval}} = 0.499$), correlation equation indeed involved geometrical descriptor TSASA. However, the best, but still poor ($R^2 = 0.695$, $R^2_{\text{crossval}} = 0.582$), two-parameter model is based solely on CPSA descriptors FNSA1 and PNSA1. The best three-parameter regression (descriptors MPC, DPSA2, and TSASA) is of significantly improved quality ($R^2 = 0.825$, $R^2_{\text{crossval}} = 0.737$), but by far not as good as the best three-parameter equation from subset $A$. The best four-parameter correlation equation is based again exclusively on CPSA descriptors (PNSA3, FNSA1, WNSA1, and WNSA2), and the performance ($R^2 = 0.896$, $R^2_{\text{crossval}} = 0.840$) again lags that of the four-parameter regression obtained in subset $A$. Addition of the DPSA3 descriptor, although improving the fit to a minor extent, reduces the stability of the model: $R^2 = 0.904$ and $R^2_{\text{crossval}} = 0.814$.

The best one- and two-parameter regression models obtained for the combination set $C$ (NRD = 40) are exactly the same as those in subset $A$. However, the fit and stability of the model were both substantially improved ($R^2 = 0.902$, $R^2_{\text{crossval}} = 0.831$) when the model included a combination of topological ($^1\chi^v$ and $^3$CIC) and geometrical ($S_3$) descriptors. Nevertheless, the best four-parameter correlation equation ($R^2 = 0.951$, $R^2_{\text{crossval}} = 0.904$) was constructed without the participation of geometrical descriptors: it included three topological indices ($^1\chi^v$, $^2$IC, and $^2$SIC) along with descriptor RNCS. Just as in subsets $A$ and $B$, the use of a five-parameter model (by extension of a four-parameter equation with descriptor $S_6$) seems unjustifiable for this small data set because of its lower stability, although it provides the best fit: $R^2 = 0.963$, $R^2_{\text{crossval}} = 0.897$.

The results obtained for subset $D$ (NRD = 19) are close in quality to those obtained for subset $A$. The one-, three-, and four-parameter correlation equations are actually the same as those in subset $A$. The best two-parameter regression model (descriptors $^2\chi^v/N_b$ and $^2$EPD; $R^2 = 0.854$, $R^2_{\text{crossval}} = 0.789$) marginally outperforms that from subset $A$. However, the optimum five-parameter equation built from the four-parameter model plus descriptor $^2\chi/N_b$ is of excellent quality and high stability: $R^2 = 0.946$ and $R^2_{\text{crossval}} = 0.890$.

Our study has revealed that squared forms of topological indices are of greater importance for correlation in this data set than either geometrical or CPSA descriptors. Although for the single-parameter equation the descriptor $^0\chi^v$ in linear form is still the most suitable variable, the combination of $^1\chi^v$ and $^1\chi^v$ (squared) provides the best two-parameter regression ($R^2 = 0.886$, $R^2_{\text{crossval}} = 0.853$) throughout all five subsets $A$–$E$. The three-parameter correlation which includes descriptors in both linear ($^2\chi^v$) and squared ($^3$IC and $^3$SIC) forms is also the best ($R^2 = 0.930$, $R^2_{\text{crossval}} = 0.897$) that was generated in our study for this data set. The same applies to the four-parameter regression model (descriptors $^2\chi^v$, $^1$CIC, $^2$IC, and squared $^1$CEPD; $R^2 = 0.954$, $R^2_{\text{crossval}} = 0.918$). By contrast with all other subsets of descriptors, subset $E$ (NRD = 20) also gives an excellent five-parameter equation. In fact, the five-parameter model obtained in subset $E$ as an extension of the four-parameter equation with descriptor $^2\chi$ (squared) combines a very high fit with stability: $R^2 = 0.968$ and $R^2_{\text{crossval}} = 0.940$.

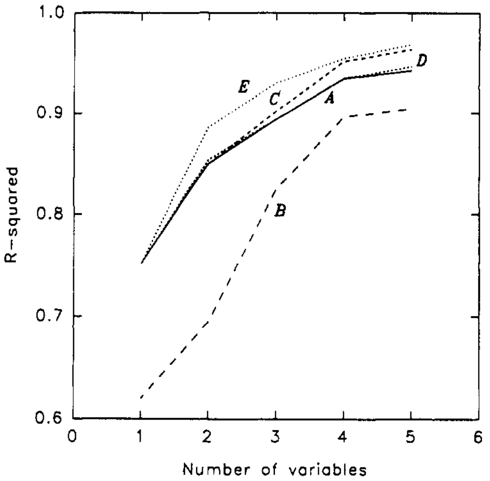

**Figure 19.** Dependence of $R^2$ on model size (structures,[3] property: sweet taste intensity).

We could not reproduce the three-parameter correlation equation obtained by Rohrbaugh and Jurs, a fact which is perhaps due to the different method of structure alignment used in the ADAPT and GROUND programs. In the GROUND program, prior to calculation of geometrical descriptors, all 3D structures are reoriented in the space so that the longest molecule axis is aligned with the abscissa $X$, and the ordinate $Y$ is assigned to the next longest molecule axis perpendicular to the $X$ axis. To obtain the correlation with three shadow indices, an orientation was used such that the aromatic ring was in the $XY$ plane.[20] The discrepancy between our results (Figure 19) and those presented by Rohrbach and Jurs underscores a weak point of some geometrical descriptors which, as shown, are strongly dependent not only on conformation of the molecule but also on assumptions regarding the orientation of the structure in 3D space.

## DESCRIPTORS INCLUDED IN THE BEST CORRELATIONS

Since this study was designed to examine the performance and usefulness of different types of descriptors, it seemed worthwhile to analyze which descriptors participate most frequently in the optimum correlation obtained.

To answer this question, we considered all the best one-, two-, three-, and four-parameter correlations obtained for the aforementioned data sets, using the combination set $C$. We counted how many times each particular descriptor participated in those correlations. We also counted descriptors which were added as the fifth independent variable at the level of the five-parameter regression models. This study is, of course, very approximate: some descriptors which performed excellently in subsets *A*, *B*, *D*, and *E*, did not show up in the best correlations generated for subset *C*. On the other hand, if we counted descriptors participating in the best correlations overall all five subsets (*A*–*E*), then we could expect an unjustified prevalence of the topological indices, presented in three subsets vs one subset of nontopological descriptors. In the combination set *C*, all descriptors have approximately the same opportunity to be included in the best regression model, and comparison of their performance there seemed to be an objective approach.

The analysis thus conducted revealed a clear "champion" among the 84 descriptors included in the combination set *C*. Surprisingly, the classical Randić index $^1\chi$, which participated in 26 of the best regression models, showed a performance which is far better than the "second place winner" $^3\chi$ which appeared in 18 best correlations. The "third place" was awarded to the molecular connectivity descriptor $^1\chi^v$ included in 14 of the best equations generated.

Within the group of connectivity descriptors, indices $^0\chi$, $^0\chi^v$, $^2\chi^v$, and $^3\chi^v$ showed up in 4, 5, 4, and 3 of the best correlations, respectively. Another classical topological index *W* (developed by Wiener) showed good performance too: it was included in 6 of the best regression models. The shape index $^1\kappa$, which participated in 7 of the best correlation equations, also appeared to be very useful.

The leaders in the group of information indices are clearly descriptors $^0$CIC, $^1$IC, and $^0$EPD which showed up in 9, 7, and 6 of the best regression models.

Notably, in the study conducted by Basak and co-workers,[13] descriptors $^3\chi^v$, $^0\chi^v$, $^0$IC, and $^1$CIC showed up in the best one- to five-parameter regression models derived for estimation of the log *P* property.

Among all the nontopological descriptors, we should stress the success of the group of geometrical descriptors. In particular, the MV descriptor was included in 8 of the best correlation equations. Shadow indices $S_3$, $S_6$, and $S_1$ showed up, respectively, in 4, 4, and 3 of the best regression models generated. The TSASA descriptor related to the total solvent-accessible surface area participated in 3 of the best correlation equations.

The CPSA descriptors which comprise the largest group of 25 are almost evenly spread between the best regression models generated, so that most of the individual descriptors are included in 1 or 2 models only. In this group, we can emphasize the descriptors WPSA2, WNSA1, and FPSA3 which participate, respectively, in 5, 3, and 3 of the best correlation equations.

## CONCLUSION

This study aimed to clarify the roles and priorities which different types of descriptors play in linear multivariate regression models. In total, 84 global (i.e. related to the whole molecule) molecular descriptors published in the literature over the past 2 decades were included in our study with some of them also in their normalized and squared forms. It should be particularly emphasized that this research dealt exclusively with the global molecular descriptors. Therefore, the conclusions thus derived may not be applicable to the local descriptors of various nature, such as the length of a specified bond, charge on a particular atom, etc.

The global molecular descriptors collated from the literature were divided into 5 subsets (topological, nontopological, combination set of 84 descriptors, normalized modifications, and squared modifications). An exhaustive search for the best one-, two-, three-, and four-parameter regression models was conducted on several different data sets taken from literature. The best five-parameter regression model was constructed from the best combination of the optimum four-parameter model with the fifth descriptor (the so-called "4 + 1" scheme).

As seen from the statistical plots presented, the optimum number of parameters for the correlation equations was 2 or 3. Further extension of the model to four or five parameters did not improve the correlation significantly in the majority of cases. Notably, for some properties, there is a tendency for different subsets of descriptors to approach the same level of correlation quality with an extension of the regression model over the limit of five parameters.

It was shown that, for physicochemical properties, the topological descriptors seem to be more advantageous than nontopological ones for linear regression analysis. Furthermore, the combination set of descriptors which included all different kinds of descriptors (in total 84), showed little, if any, superiority over the subset of 38 traditional topological indices. The subset of 46 nontopological (namely, electronic, geometrical, and combined) descriptors produced, in most cases, regression models of significantly lower quality than the subset of topological descriptors.

It was shown that modifications of topological descriptors, such as normalized and squared descriptors, can sometimes improve the regression model at the level of the one- or two-parameter equation, but the larger models are usually less sensitive to these modifications. However, in many cases, a combination of topological indices in both linear and squared form, or in traditional and normalized form, can give correlations of higher quality than a combination of topological indices with various electronic, geometrical, and combined descriptors.

We have demonstrated that the more homogeneous data sets tend to correlate at approximately the same level of quality with descriptors of different types. Nevertheless, almost all correlations with topological indices still outperformed those with nontopological descriptors.

For biological activity, the results obtained in our study are not so one-sided as those for the physicochemical properties. Evidently, more extensive research needs to be conducted, and additional various data sets should be tested, before any final conclusion can be drawn. For two data sets, the geometrical descriptors appeared to be of greater importance than the topological ones. However, in both cases, the difference in correlation quality was relatively small. By contrast, in other examples the topological descriptors showed significantly better performance than electronic, geometrical, and combined descriptors. In other words, if there is a good correlation with nontopological descriptors, the topological indices correlate at approximately the same level of quality, but not vice versa. As far as the combination set of descriptors is concerned, we can conclude from our study that a combination of topological and geometrical descriptors seems to be the most promising choice.

When we started this study, we could not clearly forecast which type of descriptors—topological or nontopological—would correlate more satisfactorily. However, we expected that electronic, geometrical, and combined descriptors should produce regression models of higher quality because these descriptors reflect more important structural features than simple composition and topology. We were surprised to see that the results presented in the previous sections are almost consistently in favor of the topological descriptors vs electronic, geometrical, and combined ones. Of course, the capacity of powerful modern workstations and mainframe computers already allow the user to conduct a simultaneous statistical treatment of the significant number of descriptors of different types. In other words, an exhaustive search for the best correlation now can be conducted on a quite large (70–90 entries) "combination set" of descriptors.

Although it is clear that there is no way to define any standard and "ideal" combination set of descriptors, this study was intended to derive some general guidelines, which the user might take into account, selecting the molecular descriptors to be included in his/her QSAR/QSPR research. To a larger extent, this study appeals to the numerous users of less powerful personal computers who are restricted in both speed of calculations and memory requirements. These researchers have no choice other than to run the QSAR/QSPR investigation on a very restricted set of descriptors. We hope that this study demonstrates to them some priorities among the descriptors currently available from the literature.

## EXPERIMENTAL SECTION

In-house developed software was used both for the generation of molecular descriptors (the GROUND program) and for the statistical QSAR/QSPR analysis (the GROUNDSTAT program). Statistical plots were prepared using the SigmaPlot package.[52] Input information on molecular structures was prepared in the form of MOLfiles, according to the standard set by Molecular Design, Ltd. In-house developed molecular editor GRAPHIN, or, alternatively, the commercially available ChemText package[53] (Molecular Design, Ltd.), was used for structure input. The optimized 3D molecular geometry was obtained in two steps: firstly, the 2D MOLfiles were pre-optimized and converted into the MOPAC format by the in-house developed MOLGEO program.[54] Secondly, the 3D coordinates of the pre-optimized molecular structure were loaded as input information to run the MOPAC 6.0 program.[55] The UNIX version of the MOPAC package (AM1 method) was run on an IBM RISC/6000 workstation. The output MOPAC files (*.ARC and *.SYBYL files) were then converted back to MOLfiles, now with the refined 3D Cartesian coordinates. The atomic partial charges were extracted from the *.ARC MOPAC files and transferred (in ASCII format of the *.PCD files[56]) to an IBM PC. The MOLfiles were used as input information for the GROUND program. The partial charges (*.PCD files) were also used for GROUND when descriptors from subsets *B* and *C* were calculated.

The GROUND/GROUNDSTAT package is written in FORTRAN77 code and runs under MS DOS, VAX/VMS and UNIX operating systems.

All QSAR/QSPR results presented in this study were obtained utilizing the MS DOS version of GROUND/GROUNDSTAT on an IBM PC 486/50 Mhz.

## ACKNOWLEDGMENT

We are thankful to Mr. Peter Rachwal for his assistance in the procedures for optimizing the 3D geometry of the structures studied.

## REFERENCES AND NOTES

(1) *Comprehensive Medicinal Chemistry*; Hansch, C., Ed.; Quantitative Drug Design, Pergamon Press: New York, 1990; Vol. 4.

(2) *QSAR: Quantitative Structure-Activity Relationships in Drug Design*; Fauchère, J. L., Ed.; Proceedings of the 7th European Symposium on QSAR held in Interlaken, Switzerland, Sept 5–9, 1988; Alan R. Liss, Inc.: New York, 1989.

(3) Iwamura, H. Structure–Taste Relationship of Perillartrine and Nitro- and Cyanoaniline Derivatives. *J. Med. Chem.* **1980**, *23*, 308–312.

(4) Naito, Y.; Sugiura, M.; Yamaura, Y.; Fukaya, C.; Yokoyama, K.; Nakagawa, Y.; Ikeda, T.; Senda, M.; Fugita, T. Quantitative Structure-Activity Relationship of Catechol Derivatives Inhibiting 5-Lipoxygenase. *Chem. Pharm. Bull.* **1991**, *39*, 1736–1745.

(5) Hansch, C.; Kerley, R. Role of the Benzyl Moiety in Biochemical and Pharmacological Processes. *J. Med. Chem.* **1970**, *13*, 957–964.

(6) Magee, P. S. A New Approach to Active-Site Binding Analysis. Inhibitor of Acetylcholinesterase. *Quant. Struct.-Act. Relat.* **1990**, *9*, 202–215.

(7) Agin, F.; Hersh, L.; Holtzman, D. *Proc. Natl. Acad. Sci. U.S.A.* **1965**, *53*, 952.

(8) Hansch, C. A Quantitative Approach to Biochemical Structure–Activity Relationships. *Acc. Chem. Res.* **1969**, *2*, 232–239.

(9) Tute, M. S. History and Objectives of Quantitative Drug Design. In *Comprehensive Medicinal Chemistry*; Hansch, C., Ed.; Pergamon Press: New York, 1990; pp 1–31.

(10) Whitmore, F. C. *Organic Chemistry*; D. van Nostrand Co.: New York, 1937.

(11) Needham, D. E.; Wei, I.-C.; Seybold, P. G. Molecular Modeling of the Physical Properties of the Alkanes. *J. Am. Chem. Soc.* **1988**, *110*, 4186–4194.

(12) *Computational Graph Theory*; Rouvray, D. H., Ed.; Nova Science Publishers, Inc.: New York, 1990.

(13) Basak, S. C.; Niemi, G. J.; Veith, G. D. Recent Developments in the Characterization of Chemical Structure Using Graph-Theoretic Indices. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science Publishers: New York, 1989; pp 235–277.

(14) Ou, X.-C.; Quang, Y.; Lien, E. J. Examination of Quantitative Relationship of Partition Coefficient (log P) and Molecular Weight, Dipole Moment and Hydrogen Bond Capability of Miscellaneous Compounds. *J. Mol. Sci.* (*Int. Ed.*) **1986**, *4*, 89–95 (CAS, 1986, 105; 105:197863z).

(15) Osmialowski, K.; Kaliszan, R. Studies of Performance of Graph Theoretical Indexes in QSAR Analysis. *Quant. Struct.-Act. Relat.* **1991**, *10*, 125–134 (CAS, 1991, 115:182253h).

(16) Kaliszan, R.; Osmialowski, K. Correlation Between Chemical Structure of Non-Congeneric Solutes and Their Retention on Polybutadiene-Coated Alumina. *J. Chromatogr.* **1990**, *506*, 3–16.

(17) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(18) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis, Chemometrics Series*, Research Studies Press Ltd.: Wiley: New York, 1986; Vol. 9.

(19) Gupta, S. P. Quantitative Structure–Activity Relationship Studies on Local Anesthetics. *Chem. Rev.* **1991**, *91*, 1109–1119.

(20) Rohrbaugh, R.; Jurs, P. C. Descriptions of Molecular Shape Applied in Studies of Structure/Activity and Structure/Property Relationships. *Anal. Chim. Acta* **1987**, *199*, 99–109.

(21) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure–Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323–2329.

(22) Stanton, D. T.; Jurs, P. C. Computer-Assisted Study of the Relationship between Molecular Structure and Surface Tension of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 109–115.

(23) Stanton, D. T.; Jurs, P. C.; Hicks, M. G. Computer-Assisted Prediction of Normal Boiling Points of Furans, Tetrahydrofurans, and Thiophenes. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 301–310.

(24) Balaban, A. T.; Joshi, N.; Kier, L. B.; Hall, L. H. Correlations between Chemical Structure and Normal Boiling Points of Halogenated Alkanes $C_1$–$C_4$. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 233–237.

(25) Rouvray, D. H. Should We Have Designs on Topological Indexes? In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier Science Publishers: Amsterdam, 1983; pp 159–177 (CAS, 1984, 100:102407p).

(26) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; Research Studies Press: Chichester, England, 1983 (CAS, 1984, 100:12959r).

(27) (a) Balaban, A. T. Applications of Graph Theory in Chemistry. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334–343. (b) Balaban, A. T. *Top. Curr. Chem.* **1983**, *114*, 21–55.

(28) Stankevich, M. I.; Stankevich, I. V.; Zefirov, N. S. Topological Indices in Organic Chemistry. *Russ. Chem. Rev.* (*Engl. Transl.*) **1988**, *57*, 191–208.

(29) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.

(30) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6614.

(31) Kier, L. B. Indexes of Molecular Shape From Chemical Graphs. *Acta Pharm. Jugosl.* **1986**, *36*, 171–188.

(32) Kier, L. B. Indexes of Molecular Shape from Chemical Graphs. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science Publishers: New York, 1990; pp 151–174.

(33) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399–404.

(34) Hall, L. H. Computational Aspects of Molecular Connectivity and its Role in Structure-Property Modeling. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science Publishers: New York, 1990; pp 201–233.

(35) Shannon, C.; Weaver, W. *Mathematical Theory of Communication*; The University of Illinois Press: Urbana, IL, 1962.

(36) Basak, S. C.; Magnuson, V. R. Molecular Topology and Narcosis. *Arzneim.-Forsch.* **1983**, *33*, 501–503.

(37) Sarkar, R.; Roy, A. B.; Sarkar, P. K. Topological Information Content of Genetic Molecules-1. *Math. Biosci.* **1978**, *39*, 299–312.

(38) Osmialowski, K.; Halkiewicz, J.; Radecki, A.; Kaliszan, R. Quantum Chemical Parameters in Correlation Analysis of Gas Liquid Chromatographic Retention Indices of Amines. *J. Chromatogr.* **1985**, *346*, 53–60.

(39) Radecki, A.; Lamparczyk, H.; Kaliszan, R. A Relationship Between the Retention Indices on Nematic and Isotropic Phases and the Shape of Polycyclic Aromatic Hydrocarbons. *Chromatographia* **1979**, *12*, 595–599.

(40) Higo, J.; Gõ, N. Algorithm for Rapid Calculation of Excluded Volume of Large Molecules. *J. Comput. Chem.* **1989**, *10*, 376–379.

(41) Pearlman, R. S. Molecular Surface Areas and Volumes and Their Use in Structure/Activity Relationships. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, 1980; pp 321–347.

(42) Harriss, D. K.; Gordeeva, E. V.; Trofimov, M. I.; Zefirov, N. S. Topological-Electronic Index Based Upon Molecular Connectivity and Partial Atomic Charges. *Proceedings of the Second World Congress of Theoretical Organic Chemists WATOC-90*, July 8–14, 1990, University of Toronto, Toronto, Canada.

(43) Osmialowski, K.; Halkiewicz, J.; Kaliszan, R. Quantum Chemical Parameters in Correlation Analysis of Gas-Liquid Chromatographic Retention Indices of Amines. II. Topological Electronic Index. *J. Chromatogr.* **1986**, *361*, 63–69.

(44) Carlson, R.; Prochazka, M. P.; Lundstedt, T. Principal Properties for Synthetic Screening: Ketones and Aldehydes. *Acta Chem. Scand.* **1988**, *B42*, 145–156.

(45) Carlson, R.; Prochazka, M. P.; Lundstedt, T. Principal Properties for Synthetic Screening: Amines. *Acta Chem. Scand.* **1988**, *B42*, 157–165.

(46) Stanton, D. T.; Egolf, L. M.; Jurs, P. C.; Hicks, M. G. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 306–316.

(47) Balaban, A. T.; Kier, L. B.; Joshi, N. Correlations between Chemical Structure and Normal Boiling Points of Acyclic Ethers, Peroxides, Acetals, and Their Sulfur Analogues. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 237–244.

(48) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of Chemical Property Estimation Methods*; McGraw-Hill: New York, 1982.

(49) Radhakrishnan, T. P.; Herndon, W. C. Molar Volume of Alkanes and Topological Indexes. *J. Math. Chem.* **1988**, *2*, 391–329.

(50) Jain, D. V. S.; Singh, S.; Combar, V. K. Correlation between Topological Features and Physicochemical Properties of Alkylbenzenes. *Indian J. Chem.* **1988**, *27A*, 923–931.

(51) Verloop, A.; Hoogenstraaten, W.; Tipker, J. Development and Application of New Steric Substituent Parameters in Drug Design. *Drug. Des.* **1976**, *7*, 165–207.

(52) *SigmaPlot, version 4.0. User's Manual*; Jandel Scientific: 1989.

(53) *ChemText, Chemist's Personal Software Series*, User's Guide, Version 1.5; Molecular Design Ltd.: 1992.

(54) Gordeeva, E. V.; Katritzky, A. R.; Shcherbukhin, V. V.; Zefirov, N. S. Rapid Conversion of Molecular Graphs to Three-Dimensional Representation Using the MOLGEO Program. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 102–111.

(55) *MOPAC Manual*, 6th ed. *QCPE*, 1990.

(56) Katritzky, A. R.; Gordeeva, E. V. GROUND/GROUNDSTAT: Program Package for QSAR/QSPR Research. Theoretical Introduction and User's Manual. Department of Chemistry, University of Florida, 1991.