

## Graphical Bond Orders: Novel Structural Descriptors†

Milan Randić\*

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

Zlatko Mihalić

Faculty of Science, The University of Zagreb, HR-41000 Zagreb, The Republic of Croatia

Sonja Nikolić and Nenad Trinajstić

The Rugjer Bošković Institute, HR-41001 Zagreb, The Republic of Croatia

Received May 18, 1993\*

We outline an algorithm for construction of novel molecular descriptors from known structural invariants or molecular properties viewed as descriptors. The novel descriptors are bond-additive quantities derived by assigning to each bond a contribution  $x'$  obtained by evaluating invariant  $X$  for graph  $G'-e$ , which is attained from graph  $G$  (representing a given molecule) by deleting edge  $e$ . The molecular descriptor  $X'/X$  is obtained as a normalized sum of bond contributions, where  $X'$  is equal to the sum of bond orders  $x'$ . The approach is illustrated by presenting  $X'/X$  descriptors for smaller alkanes for several well-known topological indices, including the connectivity index, Hosoya's  $Z$  index, the Wiener index, and others. The algorithm is quite general and allows one to include molecular properties as source data for the construction of novel descriptors. This is particularly important in view of a limited number of *properties* used as descriptors in traditional quantitative structure–activity studies. The new algorithm literally doubles the number of descriptors available to traditional chemometricians in their quest for novel property–activity relationships.

## 1. INTRODUCTION

Difficulties in studies of structure–property relationships and extension of such studies to problems involving bioactivities center on the characterization of chemical structures.<sup>1,2</sup> What is a chemical structure? How does one construct a useful model of a structure? How does one describe structure quantitatively? How does one measure structural similarity and dissimilarity?

Difficulties associated with the study of these and related questions originate with difficulties in arriving at a complete representation of a chemical structure—as there may be none. A complete representation ought to allow one to *extract* and *emphasize*, as models do, any aspect of a structure of interest for a particular study. In this light the statement by Norbert Wiener, “The best model of a cat is another cat or, better, the cat itself”<sup>3</sup> can be fully appreciated and is tantamount to our suggestion of the nonexistence of *complete* representation of complex systems and structures, cats included! Hence, as a pragmatic resolution of the dilemma, we build models that include structural features of interest for a particular study. In the case of a chemical structure, the building of models which vary with their computational complexity has received considerable attention. As extremes these include the Schrödinger equation as one of the more complex models and bond additivities as one of the simplest models. The former relates to *the nature of the chemical bond*,<sup>4</sup> while the latter acknowledges the presence of bonds and tries to extract regularities in selected molecular properties that follow. Bond additivities illustrate one aspect of the algebraic approach to the chemical structure well-represented within chemical graph theory,<sup>5</sup> the prime concern of which may be summarized as *the nature of the chemical structure*.<sup>1</sup> Thus, rather than being concerned with details resulting in the stability of a particular constellation of electrons and nuclei, governed by the laws of

quantum mechanics, chemical graph theory is concerned with *consequences* of a particular bonding pattern, i.e., how various properties of chemical structure relate, if they do, to structural constituents. Difficulties of the “art” of chemical graph theory are in perceiving relevant *components*, in considering useful *partitioning* of a structure, and in arriving at novel *concepts* that may eventually lead to data reduction. According to Max Planck,<sup>6</sup> “The chief problem in every science is that of endeavouring to arrange and collate the numerous individual observations and details which present themselves, in order that they may become a part of one comprehensive picture.” Observe that the Schrödinger equation achieves this reduction of data at electron–electron, electron–nuclei, and nuclei–nuclei levels, but it generates “observation” at a molecular level. Chemical graph theory strives to achieve the data reduction at the molecular level, i.e., by trying to “arrange and collate the numerous” molecular properties, and in doing so generates “raw” material for studies at a higher conceptual or organizational level, i.e., a level of lower resolution. An important distinction, however, between the quantum chemistry and chemical graph theories is that the basic laws that define the former are well-established, while chemical graph theory lacks clear-cut analogous laws as guiding criteria. As is known with the increase of organizational complexity, novel underlying principles may emerge as illustrated by such statements as Aristotle's “*Natura non facit saltus*” (i.e., nature does not make jumps), Carl von Linné's “*All living from living*”, Ernst Haeckel's “*Ontogenesis is short recapitulation of phylogenesis*”, and so on. The laws of thermodynamics can also be viewed as a reflection of the complexity of systems, even though in this case it is merely the size, i.e., the number of molecules (or atoms) involved which allow one to go from discrete to continuum and formulate new *laws* (of a statistical nature). Chemical graph theory is too “close” to the fundamental basis that makes generalizations (such as the above mentioned laws and principles) difficult. Similar difficulties are apparent,

† Dedicated to John A. Pople, a builder of chemical models.

• Abstract published in *Advance ACS Abstracts*, February 15, 1994.

for example, in the "gray" area of differentiating large clusters and bulk. Nevertheless, some generalizations in chemical graph theory appear possible, such as "the principle of graduality",<sup>7</sup> which essentially says that in a comparison of discrete objects, such as chemical structures, nature does not make abrupt changes. Of course, the term "abrupt" is to be taken relative to the discrete nature of objects considered. The above principle may be viewed as a generalization of Emil Fisher's<sup>8</sup> notion of a "lock and key" model for drug receptor interaction, which is also today one of the guiding criteria in searching for novel compounds of desired properties.

Having limited guiding criteria and given the elusive nature of the representation of structures, chemical graph theory faces a difficult task in data reduction. It is, therefore, natural that much effort has been put into development of tools for characterization of chemical structures. The basic philosophy behind such efforts is to "translate" structure-property problems into "property"-property problems in which one set of properties (i.e., "properties") is *mathematical* properties,<sup>9</sup> while the other is physical, chemical, or biological, including, of course, medicinal properties such as those that are of interest in drug design and the therapeutic use of chemicals. Such an approach focuses attention on the extraction of the mathematical properties of chemical structure, and this requires a mathematical *description* of a structure, hence the use of matrices as representations of chemical structure and matrix invariants as descriptors [e.g., see ref 5].

## 2. MATRICES AS REPRESENTATION OF CHEMICAL STRUCTURE

One of the simplest structural matrices is the *adjacency* matrix<sup>5</sup> in which only the presence or the absence of bonding between a pair of atoms is recorded. Thus, the elements of the adjacency matrix are either  $a_{ij} = 1$  if sites  $i$  and  $j$  are connected by a bond or  $a_{ij} = 0$  if sites  $i$  and  $j$  are not adjacent (i.e., connected). The resulting mathematical object is called a *graph*<sup>10</sup> or molecular graph [e.g., see ref 5]. Graphs, thus, capture the most essential structural feature of a molecule—bonding between atoms. It took some time for workers to recognize that the Hückel molecular orbital (HMO) method<sup>11</sup> is essentially a graph-theoretical model [e.g., see refs 12–15]. The underlying basic assumption, adopted by Hückel in formulating the HMO method, is that of Felix Bloch's "tight" binding or nearest neighbor interactions,<sup>16</sup> which is essentially a binary relation on atomic centers of molecules. Hence, the Hückel matrix is a binary matrix, which nowadays most textbooks illustrate by introducing a variable substitution, unfortunately without properly emphasizing the graph-theoretical structure of the derived matrix. Many simple quantum-chemical models are "disguised" graph-theoretical models, as is well illustrated, for example, in Hameka's demonstration of the additivity for magnetic susceptibilities in alkanes.<sup>17</sup> The basic assumption in Hameka's model is that all C–C bonds (and all C–H bonds, respectively) are *equal* among themselves within a molecule and in different molecules. This is a plausible presumption when one focuses attention on differences in the properties of different molecules viewed as built from the same building blocks. But the presumption is a disguise for stating explicitly that molecules are viewed as graphs. Assuming all C–C bonds to be equal is equivalent to a statement that variations in bond lengths do not matter; what matters is being bonded—which is precisely what defines graphs. It should not therefore be surprising to see that the same additivity of the magnetic susceptibilities in alkanes can be cast in an equivalent and somewhat simpler graph-theoretical formulation.<sup>18</sup>

Another matrix of interest in chemical graph theory is the *distance* matrix.<sup>5,19,20</sup> The elements of the distance matrix are  $d_{ij}$  = the (smallest) number of bonds between atoms  $i$  and  $j$  and  $d_{ii} = 0$ . The distance matrix is a representative of geodesics and implies a metric, which has already been defined for graphs by Cayley.<sup>21</sup> This singular structural feature adds combinatorial content to graphs and thus distinguishes them from purely topological objects for which only the concept of neighborhood, but not metric, exists. In fact before the 1920s, graph theory in mathematics was known as "combinatorial and topological" theory,<sup>22</sup> although the term "graph" had been introduced by Sylvester much earlier (1878).<sup>23</sup>

Adjacency and distance matrices are devoid of information on spatial architecture of a chemical structure. Hence, invariants derived from such matrices refer to graphs rather than the three-dimensional structures. There are advantages and disadvantages to simple models.<sup>24–27</sup> For example, graphs lead to some "hidden" symmetries; i.e., the mathematical description of a molecule in such models (including, for instance, the HMO model) may involve "higher" symmetries than a structure apparently shows, as is reflected in "excessive" degeneracies within the HMO model.<sup>28</sup> Topological indices and bond additivities that are derived from molecular graphs are equally devoid of three-dimensional information. Successes of many regressions of molecular properties or activities against graph invariants only point to the fact that the corresponding properties are equally insensitive to spatial characteristics of a molecule. However, some properties very much depend on molecular three-dimensional structure and require molecules to be represented by more generalized matrices to allow adequate discussions of their properties [e.g., see refs 5 and 29].

Topographic (or geometric-distance) matrices record three-dimensional distances between pairs and atoms.<sup>5,20,30–33</sup> By involving all pairwise distances, such matrices preserve information on spatial molecular form and allow one to reconstruct the molecule in three-dimensional space, if required. Matrix invariants based on such matrices reflect features that depend on molecular geometry [e.g., see ref 34]. This allows one to discriminate between *cis* and *trans* isomers, *chair* and *boat* conformations, etc. The pool of matrices for representing various aspects of a chemical structure can further be broadened by construction of matrices whose entries represent selected molecular features, such as bond orders, bond overlaps, bond charges, interatomic potentials, and so on, and include any property that can be partitioned and represented as a matrix.<sup>35</sup> The elements of such matrices need not even correspond to atoms and bonds, but could correspond to molecular descriptors, such as orbitals, hybrids, path numbers, etc.<sup>36</sup>

## 3. CONSTRUCTION OF STRUCTURAL INVARIANTS

Many important structural invariants were introduced in an apparent *ad hoc* manner, including the first nontrivial descriptor, the Wiener index  $W$ .<sup>37</sup> Although most of the molecular descriptors (some known also as topological indices<sup>38</sup>) have a simple direct structural interpretation, their relationship to known molecular properties is not always equally transparent. Platt<sup>39</sup> tried to interpret the somewhat mysterious but successful descriptor  $W$ , the Wiener index, in terms of molecular volume. Recently Labanowski and co-workers<sup>40</sup> discussed the physical meaning of several topological indices. Such efforts are not essential for use of the descriptors, but it seems worthwhile to continue to elaborate on structural interpretations of graph-theoretical invariants in order to

facilitate interpretation of the results when such descriptors are used.

Because of the 1:1 correspondence between graphs and matrices,<sup>10,41</sup> all graph-theoretical indices can formally be derived from corresponding graph matrices, the adjacency matrix and distance matrix in particular.<sup>42</sup> Topological indices can further be generalized by constructing additional graph matrices or by performing some algebraic computations with them. This can lead to a still larger pool of invariants, as illustrated by recent construction of fragment descriptors by Mekenyan et al.<sup>43</sup> and by Balaban and co-workers.<sup>44</sup> While at first one may get an impression of uncontrolled proliferation of molecular descriptors, it is important to recognize that these most recent advances do not introduce *de novo* descriptors, but rather "digest" existing descriptors and extract novel information from the already available graph-theoretical invariants. In a way, one can say that these more recent contributions illustrate *algorithmic* approaches to construction of topological indices. Once such approaches are outlined, they apply equally to already available qualified molecular descriptors. We will introduce here yet another algorithmic approach to graph indices which bears some conceptual relationship to the "external" fragment topological indices of Mekenyan et al.<sup>43</sup>

#### 4. GRAPHICAL BOND ORDER

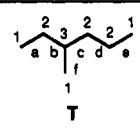
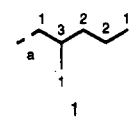
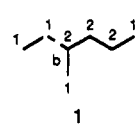
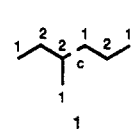
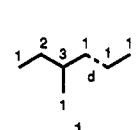
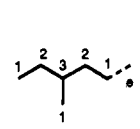
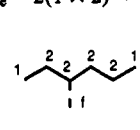
The novel algorithm assigns to each edge  $e$  in a graph  $G$  a value  $x'$  which is obtained by evaluating invariant  $X$  for a graph  $G-e$ . The same algorithm can be applied to any fragment  $f$  in a graph  $G$  by evaluating value  $y'$  which is obtained by evaluating invariant  $Y$  for a graph  $G-f$ . We will, however, here restrict our attention to bonds such as elementary fragments and will continue with examination of various  $x'$  quantities, assigned to individual edges in a graph. The restriction to an edge as a fragment would appear to limit applications of so-derived quantities to properties of bonds, such as the characterization of molecular bond-additive features. But in fact the derived bond orders may be viewed as weights of the adjacency matrix and in this way they will generate novel matrices and accompanying invariants, such as path numbers, etc. Hence, novel quantities  $x'$  have considerable potential in modeling structure-property relationships. They can be applied to comparative studies of molecules, such as are implicit in the quantification of molecular similarity, where we are now in a position to emphasize selected features of bonds, as exemplified by different molecular descriptors  $X$ , and to seek similarities with respect to diverse molecular features.

The distinction between our approach and that of Mekenyan et al.<sup>43</sup> is in how the components relating to  $G$  and  $G'$  are combined. Mekenyan et al. were primarily interested in a fragment description which they derive as a *difference*  $X(G) - [X(f) + X(G')]$ ,  $G'$  being in our case  $G-e$ , which is a special case of the more general subgraphs  $G' = G-f$  considered by Mekenyan et al. In contrast we normalize  $x' = X(G')$  by dividing it by  $X$  and use the quotients  $x'/X$ , to which we refer as *graphical bond orders*, to construct an overall molecular index  $X'/X$ , where  $X'/X$  is obtained as a bond-additive molecular characteristic. Alternatively, as already mentioned, we can use  $x'/X$  as weights for the adjacency matrix.

#### 5. ILLUSTRATIVE EXAMPLE

To illustrate the construction of graphical bond orders we will consider molecular connectivity  $\chi$ <sup>45</sup> as the invariant  $X$  to

**Table I.** Computation of Graphical Bond Orders for 3-Methylhexane<sup>a</sup>

 <p style="text-align: center;">T</p>	
(a) computation of the connectivity index for T	
$\chi = 2(1 \times 2)^{-1/2} + (1 \times 3)^{-1/2} + (2 \times 2)^{-1/2} + 2(2 \times 3)^{-1/2} = 3.308\ 06$	
(b) computation of graphical bond orders	
(b.1)	 $\chi_a = (1 \times 2)^{-1/2} + 2(1 \times 3)^{-1/2} + (2 \times 2)^{-1/2} + (2 \times 3)^{-1/2} = 2.770\ 06$ $p_a = \chi_a / \chi = 0.837\ 37$
(b.2)	 $\chi_b = 1 + 2(1 \times 2)^{-1/2} + 2(2 \times 2)^{-1/2} = 3.414\ 21$ $p_b = \chi_b / \chi = 1.032\ 09$
(b.3)	 $\chi_c = 4(1 \times 2)^{-1/2} + (2 \times 2)^{-1/2} = 3.328\ 42$ $p_c = \chi_c / \chi = 1.006\ 16$
(b.4)	 $\chi_d = 1 + (1 \times 2)^{-1/2} + 2(1 \times 3)^{-1/2} + (2 \times 3)^{-1/2} = 3.272\ 36$ $p_d = \chi_d / \chi = 0.989\ 21$
(b.5)	 $\chi_e = 2(1 \times 2)^{-1/2} + (1 \times 3)^{-1/2} + 2(2 \times 3)^{-1/2} = 2.808\ 06$ $p_e = \chi_e / \chi = 0.848\ 85$
(b.6)	 $\chi_f = 2(1 \times 2)^{-1/2} + 3(2 \times 2)^{-1/2} = 2.914\ 21$ $p_f = \chi_f / \chi = 0.880\ 94$
(c) invariant based on graphical bond orders	
$\chi' / \chi = \sum_{i=a}^f p_i = 5.594\ 62$	

<sup>a</sup> The carbon skeleton of this molecule is represented by tree T. Numbers at each site represent the corresponding graph-theoretical valencies, while letters are bond labels.

which our algorithm will apply. We selected the molecular graph of 3-methylhexane (considering, as is usual, only carbon atoms) to which we apply the algorithm as illustrated in Table 1.

Each time one bond is deleted and the connectivity indices for the remaining part (which when deleted bonds are internal makes two components) is found and divided by  $\chi$  for 3-methylhexane (3.308 06). By adding all bond orders  $x'/X$ , we obtain the novel invariant as 5.594 62, while the individual

Table II. Several Graph-Theoretical Descriptors  $X'/X$  for Nonanes<sup>a</sup>

alkane	$\chi'/\chi$	$W'/W$	$Z'/Z$	ID'/ID	$J'/J$	$P'/P$
2,2,3,3-tetramethylpentane	7.5237	5.3659	5.4000	7.2406	7.6715	5.7222
2,2,3,4-tetramethylpentane	7.5288	5.1860	5.2903	7.2427	7.9154	5.6111 <sup>a</sup>
2,2,3-trimethylhexane	7.5724	4.8696	4.8571	7.2826	8.3886	5.4444 <sup>b</sup>
2,2-dimethyl-3-ethylpentane	7.5600	5.1136	4.9167	7.2848	7.9450	5.5556 <sup>c</sup>
3,3,4-trimethylhexane	7.5533	5.0909	4.6667	7.2876	7.9710	5.5556 <sup>c</sup>
2,3,3,4-tetramethylpentane	7.5222	5.2857	5.2121	7.2438	7.7271	5.6667
2,3,3-trimethylhexane	7.5656	4.9778	4.8333	7.2843	8.1642	5.5000 <sup>d</sup>
2,3-dimethyl-3-ethylpentane	7.5466	5.2093	4.7000	7.2877	7.8112	5.6111 <sup>a</sup>
2,2,4,4-tetramethylpentane	7.5431	5.0682	5.7500	7.2365	8.2822	5.5556 <sup>c</sup>
2,2,4-trimethylhexane	7.5726	4.7872	4.9091	7.2823	8.5986	5.3889 <sup>i</sup>
2,4,4-trimethylhexane	7.5657	4.8913	4.8824	7.2840	8.3739	5.4444 <sup>b</sup>
2,2,5-trimethylhexane	7.5854	4.5714 <sup>a</sup>	4.9375	7.2780	9.1502	5.2778 <sup>e</sup>
2,2-dimethylheptane	7.6261	4.2500	4.4324	7.3144	9.6740	5.1111 <sup>f</sup>
3,3-dimethylheptane	7.6065 <sup>a</sup>	4.5714 <sup>a</sup>	4.2927 <sup>a</sup>	7.3212	8.8778	5.2778 <sup>e</sup>
4,4-dimethylheptane	7.6065 <sup>a</sup>	4.6875	4.4615	7.3233	8.5733	5.3333 <sup>g</sup>
3-ethyl-3-methylhexane	7.5874	4.9130 <sup>b</sup>	4.2955	7.3260 <sup>a</sup>	8.2423	5.4444 <sup>b</sup>
3,3-diethylpentane	7.5689	5.1364	4.2500 <sup>b</sup>	7.3287	7.9228	5.5556 <sup>c</sup>
2,3,4-trimethylhexane	7.5580 <sup>b</sup>	4.9130 <sup>b</sup>	4.5366	7.2891	8.2119	5.4444 <sup>b</sup>
2,4-dimethyl-3-ethylpentane	7.5580 <sup>b</sup>	5.0444	4.7692	7.2878	7.9827	5.5000 <sup>d</sup>
2,3,5-trimethylhexane	7.5704	4.7083	4.7027	7.2851	8.6528	5.3333 <sup>g</sup>
2,3-dimethylheptane	7.6104	4.3922	4.1364	7.3218	9.1491	5.1667 <sup>h</sup>
3-ethyl-2-methylhexane	7.5983 <sup>c</sup>	4.7500	4.2667	7.3247	8.4248	5.3333 <sup>g</sup>
3,4-dimethylheptane	7.5983 <sup>c</sup>	4.6122	4.1087	7.3272	8.6543	5.2778 <sup>e</sup>
3-ethyl-4-methylhexane	7.5863	4.8511	4.1458	7.3290	8.2399	5.3889 <sup>i</sup>
2,4-dimethylheptane	7.6108 <sup>d</sup>	4.4118	4.2927 <sup>a</sup>	7.3229	9.0891	5.1667 <sup>h</sup>
4-ethyl-2-methylhexane	7.5986 <sup>e</sup>	4.6531	4.2500 <sup>b</sup>	7.3242	8.6487	5.2778 <sup>e</sup>
3,5-dimethylheptane	7.5986 <sup>e</sup>	4.5200	4.0889	7.3260 <sup>a</sup>	8.8779	5.2222
2,5-dimethylheptane	7.6108 <sup>d</sup>	4.3077	4.1395	7.3213	9.4128	5.1111 <sup>f</sup>
2,6-dimethylheptane	7.6233	4.0926	4.2500 <sup>b</sup>	7.3164	9.9876	5.0000 <sup>j</sup>
2-methyloctane	7.6607	3.7895	3.7021	7.3510	10.4545	4.8333
3-methyloctane	7.6484 <sup>f</sup>	4.0000	3.6200	7.3562	9.8915	4.9444
4-methyloctane	7.6484 <sup>f</sup>	4.1111	3.7143	7.3583	9.5503	5.0000 <sup>j</sup>
3-ethylheptane	7.6362 <sup>g</sup>	4.3462	3.7115	7.3591	9.1320	5.1111 <sup>f</sup>
4-ethylheptane	7.6362 <sup>g</sup>	4.4706	3.8039	7.3599	8.8569	5.1667 <sup>h</sup>
n-nonane	7.6957	3.5000	3.1636	7.3845	10.8782	4.6667

<sup>a</sup> Superscript letters denote degenerate sets of descriptors.

bond orders vary from 0.837 37 to 1.032 09. One can observe some regularities in derived bond orders  $\chi'/X$  in that bonds that are now terminal make a lesser contribution compared to contributions of such bonds to the connectivity index, while internal bonds make a larger contribution. Observe also that the "degeneracy" associated with classification of bonds in various  $(m,n)$  types is now removed and bond orders  $\chi'/X$  belonging to a same  $(m,n)$  type are now different. This suggests it is worthwhile to investigate the "degeneracy" in molecular descriptors  $X'/X$  based on the connectivity indices and other graph invariants.

## 6. RESULTS

In Table II we have collected the nonanes novel graph-theoretical descriptors  $X'/X$  for a selection of well-known graph invariants. They include the following:  $\chi$ , the connectivity index;<sup>45</sup> the Wiener index  $W$ , which sums all paths of all lengths in a graph;<sup>37</sup> Hosoya's  $Z$  index, which counts nonadjacent bonds in a molecular graph;<sup>38</sup> Balaban's index  $J$ , which is a distance matrix analogue of the connectivity index;<sup>46</sup> the total path number in graphs;<sup>47,48</sup> and the molecular ID number.<sup>49</sup>

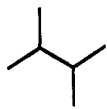


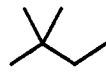


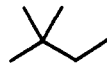
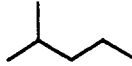

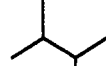
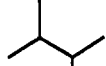
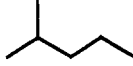
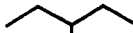
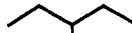
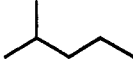
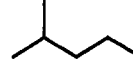


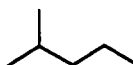



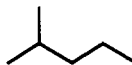







It is of particular interest to observe a diversity of relative values for the descriptors  $X'/X$  among isomers as well as some apparent "duplication" of such relative trends in  $X'/X$ . All  $X'/X$  descriptors for nonanes, except  $J'/J$ , show degeneracies. However, in view of the recently discussed procedure for construction of orthogonal molecular descriptors,<sup>50</sup> one should not be concerned with apparent parallelisms between different descriptors, unless they are strictly collinear (i.e., the regression coefficient in their mutual correlation is exactly 1). The orthogonalization procedure shows that even "almost" collinear

descriptors differ in some important structural aspects. Though such differences tend to be obscured by nonorthogonality, they can be fully exposed in orthogonalized combinations of descriptors, and therefore highly intercorrelated descriptors continue to be of interest, despite their apparent redundancy.

## 7. DISCUSSION

We will discuss some of the derived  $X'/X$  descriptors, in particular their variations with molecular shape as depicted by various isomers considered. Such discussion may suggest which descriptor may be better suited for characterization of which molecular property. Ideally one would like to have a *single* descriptor with a simple and direct structural interpretation to characterize a *single* property. In such situations one anticipates a possibility for reliable prediction of the same properties for unknown structures. It is this goal of single descriptor for single property that justifies the search for novel structural invariants and development of strategies for searches for such descriptors. The possibility of constructing orthogonal molecular descriptors and evaluating contributions of individual descriptors in a multivariate regressions makes possible for the first time progress in such a direction.<sup>50-53</sup> We demonstrated recently that descriptors  $P'/P$  and  $J'/J$  give the best single-variable correlations for octane numbers and molecular cavity areas, respectively.<sup>54,55</sup> In the latter case a regression based on  $J'/J$  for the nine heptane isomers was characterized by  $r = 0.9925$  and  $s = 1.7027$  ( $r$  is the regression coefficient and  $s$  is the standard error), which *exceeded* regression based on use of even four connectivity indices (irrespective of the high risk for chance correlation when four

Table III. Ordering of Hexane Trees According to their  $X'/X$  Descriptors

$x'/x$	$w'/w$	$z'/z$	ID'/ID	$J'/J$	$P'/P$
 4.5062	 2.0000	 2.9231	 4.2349	 4.4078	 2.6667
 4.5072	 2.3750	 3.2500	 4.2396	 4.4844	 2.8667
 4.5522	 2.5161	 3.4545	III:  4.3034	 4.9970	 2.9333
 4.5699	 2.7586	 3.8000	 4.3075	 5.1425	 3.0667
 4.6274	 2.8571	 4.0000	 4.3611	 5.8128	 3.1333

descriptors are used on only nine data points!) characterized by  $r = 0.9924$  and  $s = 2.2747$ . Even if from some independent validation testing we could be convinced that a regression based on four connectivity indices is not enhanced by chance factors, the interpretation of a correlation based on a single descriptor is much simpler than interpretation on several descriptors, regardless of how straightforward a structural meaning they may have. The situation is further aggravated by nonorthogonality of most descriptors, and while that can now be removed, such orthogonal counterparts will still have many components that must be fully identified. It is the simplicity of interpretations based on a single descriptor, if well designed to emphasize specific features of a structure, that appears desirable. Hence, whenever possible, one should strive to reduce the number of parameters in multiple regressions to a minimum, ideally to a single variable. Novel descriptors may play an important role here. In our opinion, one should avoid composite descriptors, such as the so-called superindex [e.g., see ref 56], unless the quantities considered combine similar structural qualities, as is the case with descriptors  $p_2 - p_3$ , and  $p_0 + p_1 - p_2 - p_3$ , which were recently found to describe carbon-13 chemical shifts in alkanes adequately.<sup>57</sup>

In Table III we give the ordering of the hexane isomers according to their  $X'/X$  descriptors. Similarly, in Table IV we give the ordering of the hexane isomers according to their selected thermodynamic properties.

These tables reveal that often different descriptors as well as different properties show visibly distinct behavior. Hence, if we would like to seek regressions using a single descriptor, we need structurally different descriptors. Even if we have to resort to multiple regressions using several descriptors, some information on relative magnitudes of descriptors and relative ordering of properties may facilitate considerably the search for an optimal regression.



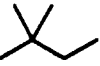

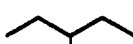

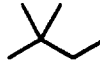

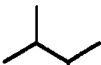
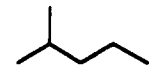
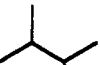
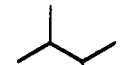
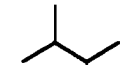
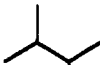
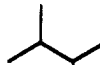
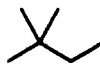
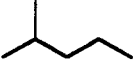
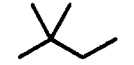
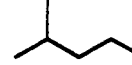


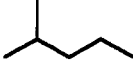
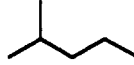

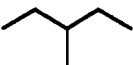
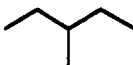

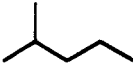
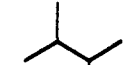
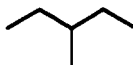
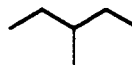
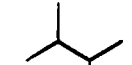

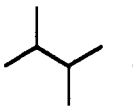

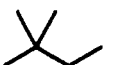
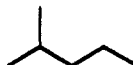


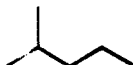
**7.1. Connectivity  $x'/x$  Descriptor.** Terminal bonds apparently have the smallest bond orders  $x'/X$ , and the values

increase with the valency of the nonterminal vertex. Bond orders of terminal bonds show also some increase with an increase in the size of the molecules. Inner bonds have larger connectivity bond orders  $x'/X$ , and the values increase with the valencies of the vertices involved. Some bond orders in different graphs show a simple relationship. For example, the bond order of 1.000 00 appears regularly in linear chains for the bond next to terminal bonds. The molecular descriptor  $x'/x$  shows a regular size dependence; it increases with an increase of the number of vertices in a molecular graph and may therefore be expected to correlate well with size-dependent molecular properties. In the case of hexanes their ordering according to the increase of the  $x'/x$  values does not follow any pattern of the ordering of hexanes according to their selected physical properties.

**7.2. Wiener  $w'/w$  Descriptor.** Here terminal bonds have considerably higher bond orders  $w'/w$  than inner bonds, often by a factor of 2. The molecular descriptor  $w'/w$  increases with the size of a graph, but there is some overlap in  $w'/w$  values for graphs of different sizes, which makes this descriptor suitable for correlation with many bond-additive properties, as these tend on average to increase monotonically with the size of a molecule. The relative order of  $w'/w$  among heptane isomers parallels ordering of several molecular properties, such as octane numbers and threshold soot index.<sup>55</sup> In the case of hexanes the  $w'/w$  descriptor approximates inversely the ordering of pattern of boiling points, critical temperatures, and heats of vaporization.

**7.3. Hosoya's  $z'/z$  Descriptor.** Here again the bond orders  $z'/z$  for terminal bonds make lesser contributions to the molecular  $z'/z$  than the bond orders of inner bonds. The overall variation in bond orders within a molecule, as well as within the family of isomers, is not dramatic, and there is no systematic increase in bond orders with an increase of the size of molecular graphs. As a consequence,  $z'/z$  values for molecules of different sizes overlap. For example,  $z'/z$  for dimethylpropane is larger than  $z'/z$  for *n*-hexane, 2-meth-

**Table IV.** Ordering of Hexane Isomers (Depicted by the Corresponding Trees) According to the Increase in the Values of Their Selected Physical Properties<sup>a</sup>

bp/°C	CP/atm	CT/°C	MV/(cm <sup>3</sup> mol <sup>-1</sup> )	MR/(cm <sup>3</sup> mol <sup>-1</sup> )	HV/(kJ mol <sup>-1</sup> )	ST/(dyn cm <sup>-1</sup> )	mp/°C
 49.74	 29.92	 27.69	 129.72	 29.80	 27.69	 16.30	 -95.35
 57.99	 29.95	 29.12	 130.24	 29.81	 29.12	 17.37	 -99.87
 60.27	 30.67	 29.86	 130.69	 29.91	 29.86	 17.38	 -118.00
 63.28	 30.83	 30.27	 131.93	 29.93	 30.27	 18.12	 -128.54
 68.74	 30.99	 31.55	 132.74	 29.95	 31.55	 18.42	 -153.67

<sup>a</sup> Physical properties of hexanes are taken from ref 58. The abbreviations for the physical properties considered are as follows: bp = boiling points; CP = critical pressures; CT = critical temperatures; MV = molar volumes (at 20 °C); MR = molar refractions (at 20 °C); HV = heats of vaporization (at 25 °C); ST = surface tensions (at 20 °C); mp = melting points.

ylpentane, and 3-methylpentane. Clearly, therefore,  $Z'/Z$  cannot be a useful descriptor for size-dependent properties, but it may be of interest for study of isomeric variations. However, the  $Z'/Z$  index inversely, but correctly, orders heptane isomers with respect to their boiling points, critical temperatures, and heats of vaporization; i.e., the smallest bp, CT, or HV value corresponds to the largest  $Z'/Z$  value. Because of this result, we carried out the structure-property analysis between  $Z'/Z$  and boiling points, critical temperatures, and heats of vaporization for nonanes using as the source of the experimental data work by Needham et al.,<sup>58</sup> but only in the case of the linear relationship between HV and  $Z'/Z$  have reasonable statistical parameters ( $r = 0.93$ ;  $s = 0.60$ ) been obtained.

**7.4. Identification ID'/ID Number.** Bond orders based on ID numbers more than those based on the Hosoya  $Z$  index show smaller variations within individual molecules. As the molecular size increases, the ID'/ID descriptor shows a slight increase in its magnitude. Apparently this suffices to make the molecular ID'/ID descriptor increase monotonously with molecular size and keeps molecules of different sizes apart. Hence the ID'/ID descriptor may be of interest in correlations of properties that depend strongly on molecular size.

The ordering of the hexane trees by the ID'/ID descriptor follows exactly the ordering of hexane isomers by boiling points, critical temperatures, heats of vaporization, and surface tensions. However, only the linear structure-property relationships between bp and ID'/ID, and HV and ID'/ID show reasonable statistical parameters; i.e., for the former case  $s = 3.3$  and  $r = 0.91$ , while for the latter case  $s = 0.55$  and  $r = 0.94$ . In the case of nonanes only the relationship between

HV and ID'/ID descriptor is still of some statistical significance:  $s = 0.73$  and  $r = 0.89$ .

**7.5. Balaban's  $J'/J$  Descriptor.** Bond orders based on Balaban's  $J'/J$  descriptor show considerable variations within individual molecules, with terminal bonds making smaller contributions. On average  $J'/J$  bond orders increase somewhat with the size of the molecules considered, but the overall increase of the molecular  $J'/J$  descriptor as the number of carbon atoms increases is not sufficient to separate molecules of different sizes, even though this appears to be the case for smaller alkanes. The relative ordering of hexane isomers parallels that of  $W'/W$  except for the reversal; i.e.,  $n$ -hexane has the largest  $J'/J$  but the smallest  $W'/W$ . It appears therefore that  $J'/J$  will compete with  $W'/W$  in some correlations, and this has already been observed in correlations of molecular cavity areas in heptanes.<sup>55</sup> The two best single descriptors found were  $W'/W$  with  $r = 0.988$  and  $s = 2.119$  and  $J'/J$  with  $r = 0.993$  and  $s = 1.703$ . In contrast, the correlation of the cavity area against  $W$  and  $J$  as descriptors shows less impressive statistics:  $r = 0.979$ ,  $s = 2.823$  and  $r = 0.980$ ,  $s = 3.290$ , respectively. Hence a switch from  $J$  to  $J'/J$  almost halves the standard error, which is an impressive improvement in the regression. Typically the standard error is halved by introduction of an additional descriptor in multiple regressions.

In the case of the hexanes the  $J'/J$  descriptor follows the pattern of boiling points, critical temperatures, and heats of vaporization, except that 2-methylpentane and 3-methylpentane are given in the reverse order.

**7.6. Path Number  $P'/P$ .** The path orders  $p'/p$  and molecular  $P'/P$  indices were in fact the first graphical bond

orders considered by one of the present authors.<sup>54,55</sup> They give the same relative magnitudes for isomeric variations in hexanes as does  $W'/W$  and thus offer alternative descriptors for similar correlations. Because all trees (acyclic graphs) with the same number of vertices have the same total number of paths  $P$ , the index  $P'/P$  may show degeneracies (i.e., two or more different structures may have the same molecular descriptor  $P'/P$ ), because the denominator in  $P'/P$  is the same for all isomers and this reduces the possibility for variations in  $P'/P$ . Apparently this is not the case for hexanes, but abundant degeneracies are found in the case of nonanes (see Table II). The pattern of ordering the hexane trees by the  $P'/P$  descriptor is not followed by any physical property, the closest being the ordering by melting points. However, there is disagreement in the positions of 2,2-dimethylbutane and 2-methylpentane

## 8. CONCLUDING REMARKS

We have outlined in this paper a general approach to the construction of novel descriptors  $X'/X$  derived from bond components  $x'/x$ , referred to as graphical bond orders. These quantities are among the simplest of the external fragment indices, and other related quantities can be obtained by considering other fragments instead individual bonds. We have adopted the terminology of Mekenyan, Bonchev, and Balaban,<sup>43</sup> who refer to a fragment index as external, when it is derived by excising the fragment from a structure, in contrast to the internal (or inherent) fragment index when the fragment is considered as a whole unit. A same internal index characterizes a same fragment regardless of where in a molecule it is located, but the external index is sensitive to the fragment environment.

Graphical bond orders  $x'/x$  give rise to weighted matrices, which will lead to additional graph invariants, such as those obtained from a count of the corresponding path numbers, as was discussed recently elsewhere.<sup>54</sup> We argued for novel molecular descriptors as a way to reduce many multiple regressions to simple regressions or at least to reduce the number of variables in multiple regressions. A comparison of relative magnitudes of descriptors with relative magnitudes of selected properties may suggest the best descriptor or at least the best initial descriptor, i.e., the descriptor which will account for the major part of a regression. Available procedures based on orthogonalization of descriptors<sup>50</sup> may then be used to find additional descriptors that will produce better and still statistically significant regressions.

## ACKNOWLEDGMENT

S.N., N.T., and Z.M. were supported by the Ministry of Science of the Republic of Croatia via Grants 1-07-159 and 1-07-185.

## REFERENCES AND NOTES

- Randić, M. *J. Math. Chem.* **1990**, *4*, 157.
- Randić, M. *J. Chem. Educ.* **1992**, *69*, 713.
- Quoted by: Krishna Murty, E. V. In *Computer Modeling of Complex Biological Systems*, 3rd printing; Iyengar, S. S., Ed.; CRC Press: Boca Raton, FL, 1985; p 77.
- Pauling, L. *The Nature of the Chemical Bond*, 2nd ed., 12th printing; Cornell University Press: Ithaca, NY, 1948.
- Trinajstić, N. *Chemical Graph Theory*, 2nd revised ed.; CRC Press: Boca Raton, FL, 1991.
- Planck, M. *Survey of Physical Theory*; Dover: New York, 1960.
- Trinajstić, N.; Klein, D. J.; Randić, M. *Int. J. Quantum Chem.: Quantum Chem. Symp.* **1986**, *20*, 699.
- Fischer, E. *Chem. Ber.* **1894**, *27*, 2985.
- Randić, M. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley-Interscience: New York, 1990; p 77.
- Harary, F. *Graph Theory*, 2nd printing; Addison-Wesley: Reading, MA, 1971.
- Hückel, E. *Z. Phys.* **1931**, *60*, 204.
- Günthard, H. H.; Primas, H. *Helv. Chim. Acta* **1956**, *39*, 1645.
- Ruedenberg, K. *J. Chem. Phys.* **1961**, *34*, 1861.
- Marcus, R. A. *J. Chem. Phys.* **1965**, *43*, 2643.
- Trinajstić, N. In *Semiempirical Methods of Electronic Structure Calculation. Part A: Techniques*; Segal, G. A., Plenum Press: New York, 1977; p 1.
- Bloch, F. *Z. Phys.* **1929**, *52*, 555; **1930**, *61*, 206.
- Hameka, H. F. *J. Chem. Phys.* **1961**, *34*, 1966.
- Randić, M. *Chem. Phys. Lett.* **1978**, *53*, 602.
- Rouvray, D. H. In *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: London, 1976; p 175.
- Mihalić, Z.; Veljan, D.; Amić, D.; Nikolić, S.; Plavšić, D.; Trinajstić, N. *J. Math. Chem.* **1992**, *11*, 223.
- Cayley, A. *Cambridge Math. J.* **1841**, *2*, 267.
- Biggs, N. L.; Lloyd, E. K.; Wilson, B. J. *Graph Theory 1736-1936*; Clarendon Press: Oxford, U.K., 1976.
- Sylvester, J. J. *Nature* **1877-1878**, *17*, 284.
- Trindle, C. *Croat. Chem. Acta* **1984**, *57*, 1231.
- Turro, N. J. *Angew. Chem., Int. Ed. Engl.* **1986**, *25*, 882.
- Trinajstić, N. In *MATH/CHEM/COMP 1987*; Lacher, R. C., Ed.; Elsevier: Amsterdam, 1988; p 83.
- Maksić, Z. B. In *Atomic Hypothesis and the Concept of Molecular Structure*; Maksić, Z. B., Ed.; Springer-Verlag: Berlin, 1990; p XIII.
- Wild, U.; Keller, J.; Günthard, H. H. *Theor. Chim. Acta* **1969**, *14*, 383.
- Crippen, G. M.; Havel, T. F. *Distance Geometry and Molecular Conformation*; Research Studies Press/Wiley: Taunton, Somerset, England, 1988.
- Randić, M. In *MATH/CHEM/COMP 1987*; Lacher, R. C., Ed.; Elsevier: Amsterdam, 1988; p 101.
- Balasubramanian, K. *Chem. Phys. Lett.* **1990**, *169*, 224.
- Balasubramanian, K. *J. Comput. Chem.* **1990**, *11*, 829.
- Nikolić, S.; Trinajstić, N.; Mihalić, Z.; Carter, S. *Chem. Phys. Lett.* **1991**, *179*, 21.
- Mihalić, Z.; Nikolić, S.; Trinajstić, N. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 28.
- Amić, D.; Trinajstić, N. Manuscript in preparation.
- Randić, M. *J. Math. Chem.* **1992**, *9*, 97.
- Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.
- Hosoya, H. *Bull. Chem. Soc.* **1971**, *44*, 2332.
- Platt, J. R. *J. Phys. Chem.* **1952**, *56*, 328.
- Labanowski, J. K.; Motoc, I.; Dammkoehler, R. A. *Comput. Chem.* **1991**, *15*, 47.
- Malkovitch, J.; Meyer, W. *Graphs, Models and Finite Mathematics*; Prentice-Hall: Englewood Cliffs, NJ, 1974; p 146.
- Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. *Topics Curr. Chem.* **1983**, *114*, 21.
- Mekenyan, O.; Bonchev, D.; Balaban, A. T. *J. Math. Chem.* **1988**, *2*, 347.
- Filip, P. A.; Balaban, T. S.; Balaban, A. T. *J. Math. Chem.* **1987**, *1*, 61.
- Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
- Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399.
- Randić, M.; Wilkins, C. L. *Chem. Phys. Lett.* **1979**, *63*, 332.
- Randić, M.; Wilkins, C. L. *J. Phys. Chem.* **1979**, *83*, 1525.
- Randić, M. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164. Szymanski, K.; Müller, W. R.; Knop, J. V.; Trinajstić, N. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 413.
- Randić, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 57.
- Randić, M. *New J. Chem.* **1991**, *15*, 517.
- Randić, M. *J. Mol. Struct. (THEOCHEM)* **1991**, *233*, 45.
- Randić, M.; Trinajstić, N. *J. Mol. Struct. (THEOCHEM)* **1993**, *284*, 209.
- Randić, M. *J. Math. Chem.* **1991**, *7*, 155.
- Randić, M. *Croat. Chem. Acta* **1991**, *64*, 43.
- Bonchev, D.; Mekenyan, O.; Trinajstić, N. *J. Comput. Chem.* **1981**, *2*, 127.
- Randić, M. *J. Magn. Reson.* **1980**, *39*, 931.
- Needham, D. E.; Wei, I.-C.; Seybold, P. G. *J. Am. Chem. Soc.* **1988**, *110*, 4186.