Novel Shape Descriptors for Molecular Graphs

Milan Randić[†]

Department of Chemistry and Computer Science, Drake University, Des Moines, Iowa 50311, and
National Institute of Chemistry, SI-1001 Ljubljana, Hajdrihova 19, P. O. B. 3430, Slovenia

We report on novel graph theoretical indices which are sensitive to the shapes of molecular graphs. In contrast to the Kier's kappa shape indices which were based on a comparison of a molecular graph with graphs representing the extreme shapes, the linear graph and the "star" graph, the new shape indices are obtained by considering for all atoms the number of paths and the number of walks within a graph and then making the quotients of the number of paths and the number of walks the same length. The new shape indices show much higher discrimination among isomers when compared to the kappa shape indices. We report the new shape indices for smaller alkanes and several cyclic structures and illustrate their use in structure−property correlations. The new indices offer regressions of high quality for diverse physicochemical properties of octanes. They also have lead to a novel classification of physicochemical properties of alkanes.

INTRODUCTION

To a great extent molecular properties reflect various structural features of a molecule, such as the molecular size, the pattern of bonding, branching and cyclicity, the molecular shape, the degree of flexibility, and, of course, the nature of various functional groups contained in the structure. To get an insight into the structure−property relationship we need molecular descriptors that can adequately characterize molecular size, molecular branching, the variations in molecular shapes, and influence of heteroatoms on the structure and its properties. The graph theoretical invariants (known popularly even if not quite correctly as topological indices) offer a useful pool of molecular descriptors when the molecular properties are dominated by atom−atom connectivity.[1−8] If 3-D molecular structure is critical for an understanding of a particular structure−property relationship, then one should consider structural invariants derived from 3-D molecular structure.[9−20] A number of graph theoretical invariants have been generalized so that they produce geometry-dependent descriptors.[21−24] Similarly a number of graph theoretical indices have been extended to discriminate among atoms of different kinds.[25−29] In such a generalization, strictly speaking, we go beyond the graph theory. Nevertheless, the combinatorial properties, and the associated metric of graphs, have some common characteristics whether we treat 2-D or 3-D molecular models or whether we treat hydrocarbons or molecules having heteroatoms. Often useful descriptors for a 3-D model or a structure having heteroatoms is derived by a modification of characterizations of the structure−property relationships derived for a 2-D molecular graph and suitable modification of descriptors initially developed for hydrocarbons only. Illustration of the former are the 3-D connectivity indices[20−24] and the 3-D Wiener index.[12−15] The illustration of the latter are the valence connectivity indices of Kier and Hall,[25,26] Balaban's J index

for heteroatomic systems,[27,28] modifications made by Kupchik,[29] and variable molecular descriptors.[30−42]

Most molecular properties are dominated by molecular size, which is well characterized by most of the topological indices. The difficulties arise when one wants to describe smaller variations in molecular properties for molecules of the same size. Many molecular descriptors fail to adequately characterize smaller variations in properties among molecules of the same size (isomers). In such situations molecular size is constant, and its role has been effectively neutralized. Now molecular shape becomes the prime factor in dictating the variations in the properties of the molecules considered.

Within a set of compounds of a same size the relative magnitudes for selected molecular properties may vary widely. Hence, a single molecular descriptor may not suffice for an adequate characterization. In complex situations, such as structure-biological activity studies, even several indices need not describe some properties adequately. The challenge in quantitative structure−property relationship (QSPR), quantitative structure−activity relationship (QSAR), and quantitative structure-shape relationship (QShAR) is to find adequate mathematical descriptors that can characterize diverse molecules and diverse molecular shapes.

ON SELECTION OF MOLECULAR DESCRIPTORS

Multivariate regression analysis (MRA) imposes no restrictions on the type and the number of descriptors to be used in a structure−property-activity study. To maintain a valid statistical significance of the MRA one should restrict the maximal number of descriptors to be used, which will depend on N, the number of compounds investigated. As a rule one would like to use as few descriptors as possible, ideally a single descriptor. In such a case there will be no ambiguities in the interpretation of the regression. However, structure−property-activity relationships are generally more involved and require characterization using several descriptors. Most molecular descriptors are interrelated, sometimes highly interrelated.

[†] Corresponding author fax: (515)292-8629. Current address: 3225 Kingman Road, Ames, IA 50014.

Moreover, when a sufficiently large pool of descriptors is available there is a possibility of multiple solutions to a particular structure−property study. This topic has apparently not received enough attention in the literature on MRA. In recent studies this problem has become more pronounced because of the availability of computational packages which generate hundreds of topological indices and other molecular descriptors. One way to reduce the number of plausible solutions is to impose restrictions on the selection of descriptors. For example, one can restrict the selection of descriptors to families of structurally related descriptors. The first such family of descriptors, the path numbers, was suggested over 50 years ago as potentially useful descriptors in structure−property studies by Platt.[43] One could similarly view as a family of structurally related descriptors with walks of different lengths.[44−49] The significance of the work of Platt was overlooked for a while, and his work only resurrected with the revival of interest in the chemical graph theory during the mid 1970s.[50−57]

The connectivity indices[58,59] and the valence connectivity indices[60,61] are families of molecular descriptors widely used in QSAR during past 25 years. More recently several other topological indices, including Wiener number W^{62} and the Hosoya Z topological index,[63] have been generalized to yield the "higher order" indices of the same kind.[64−69] Kier's kappa shape indices,[70−75] which in the case of trees are reduced in essence to squared reciprocal paths, also present one such useful basis of structurally related indices.

KIER'S KAPPA SHAPE INDICES

Molecular shape is not easy to define, and when a molecule is modeled by a graph the notion of shape becomes even more diffused. Kier[70−75] constructed his kappa indices (κ), by examining the extreme shapes for trees: the graph representing a chain and the graph representing a "star". The shape index was defined as a measure of a relative "position" of a graph with respect to the two extreme shapes. The derived indices, which are proportional to the inverse square of molecular path numbers, are sensitive to the form of a graph more than its size. Hence, the kappa shape indices ought to correlate with properties that are independent of molecular size. A good correlation with such properties would be a justification for calling the indices the "shape" descriptors.

Kier's kappa shape descriptors were tested in a correlation on Pitzer's acentric factor. According to Pitzer and collaborators[76,77] the acentric factor, which is defined in terms of the reduced vapor pressure, measures a deviation of a molecule from an ideal spherical shape. Hence, the acentric factors are expected to critically depend on molecular shape. Kier obtained an excellent regression for the Pitzer's acentric factors for a set of alkanes. Over 50 structures having between 5 and 17 carbon atoms were considered, using two kappa descriptors and gave regression characterized by satisfactory statistical parameters:[75] $r = 0.995$, $s = 0.014$, and $F = 2277$. Here r is the coefficient of the regression, s is the standard error, and F is the Fisher ratio. However, the hypothesis that the acentric factor mirrored solely the molecular shape needs to be examined critically. A close look of the data shows that the Pitzer acentric factors depend to *some extent* on the molecular size. A correlation of the Pitzer's acentric factor against n, the number of carbon atoms, gives the following statistical parameters: $r = 0.9482$, $s = 0.0425$, and $F = 437$.[78] Clearly the kappa indices, which show a very good regression with the Pitzer's acentric factors must have some size dependence, otherwise they could not lead to such a good correlation with a property that is shape and size dependent.

Whether an index could be called a shape descriptor or not may be a matter of semantics. Despite, however, possible semantic ambiguity the fact is that the Kier's κ indices even if not "pure" shape descriptors gave a very satisfactory correlation of a property that other descriptors failed to characterize. This is an important factor that should not be overlooked.

Kier's κ indices involve some arbitrary steps. For instance $^2\kappa$ in comparison with $^1\kappa$ is somewhat smaller, so Kier decided to scale the computed $^2\kappa$ by $^2\kappa/4$ as a new true value for this kappa index. This, of course, does not influence its performance in MRA and such modifications are legitimate. Nevertheless, they reflect on a human intervention at this fundamental level of the definition of the descriptors. Can we arrive at descriptors that will have fewer of such ad hoc interventions?

ATOMIC PATH/WALKS DESCRIPTORS

We will define shape index for a molecular graph by considering paths and walks of different length within a graph. Paths and walks have been considered in chemical graph theory for quite some time but always separately. In fact most, if not all, hitherto considered molecular descriptors were defined independently of each other. Only relatively recently several descriptors (and matrices) based on the quotient of two quantities have been proposed[79−83] and even extended for characterization of DNA[84−87] and proteomics.[88,89] However, descriptors defined by quotients of two descriptors have shown some interesting properties. In the case of graphs embedded in 2-D or 3-D space, the leading eigenvalue derived from the D/D matrices (the elements of which are given as the ratio of Euclidean and the graph theoretic distance between atoms in a molecule) have been interpreted as a measure of "folding" of a chain in space.[79] Similarly, the cyclic character of a graph has been characterized using the leading eigenvalue of D/DD matrices, where D and DD are the topological distance matrix (D) and the detour distance matrix (DD).[90−93] The elements of D/DD matrices are given for each pair of vertices as the quotients of the shortest and the longest distances in a graph.

In Table 1 we have listed paths of different lengths (p_i) and the walks of different lengths (w_i) for the individual carbon atoms as well as the quotients of paths and walks of the same length (p_i/w_i) for 2-methylpentane illustrated in Figure 1. First, as we see, the entries in sequence of path/walk numbers for individual atoms are not size-dependent. The first entry for any atom is always equal to one, because the number of paths of length one and the number of walks of length one is the same (equal to the number of bonds). While the number of paths in a molecule is bounded, and determined by the size of the longest path, the number of walks in a graph is unbounded. However, because we are interested in the quotients of the number of paths and the number of walks of the same length we terminate the count

Table 1. Starting from Each Atom the Count of Paths, the Count of Walks, and the Quotients Path/Walks for Carbon Atoms of 2-Methylpentane[a]

	paths				walks			
	p_1	p_2	p_3	p_4	w_1	w_2	w_3	w_4
1	1	2	1	1	1	3	4	11
2	3	1	1		3	4	11	
3	2	3			2	5		
4	2	1	2		2	3	7	
5	1	1	1	2	1	2	3	7
6	1	2	1	1	1	3	4	11
total	10	10	6	4	10	20	29	29

	p_1/w_1	p_2/w_2	p_3/w_3	p_4/w_4
1	1/1	2/3	1/4	1/11
2	3/3	1/4	1/11	
3	2/2	3/5		
4	2/2	1/3	2/7	
5	1/1	1/2	1/3	2/7
6	1/1	2/3	1/4	1/11
		Molecule		
p/w	1	0.50278	0.20166	0.07792
P/W	1	0.50000	0.20690	0.13793

[a] Numbering of atoms is shown in Figure 1.

6

2

1 3 4

5
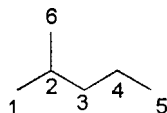
Figure 1. 2-Methylpentane: numbering of carbon atoms.

Table 2. Quotients of the Count of Path and the Count of Walks for Symmetry Nonequivalent Carbon Atoms of Smaller Benzenoid Hydrocarbons[a]

molecule	atom	p_2/w_2	p_3/w_3	p_4/w_4	p_5/w_5
benzene	a	2/4	2/8	2/16	2/32
naphthalene	a	4/7	4/17	4/39	4/91
	b	3/5	4/11	4/26	5/59
	c	2/4	3/9	5/20	6/46
anthracene	a	4/6	6/14	6/36	8/86
	b	4/7	5/18	7/43	8/166
	c	3/5	4/11	5/27	6/63
	d	2/4	3/9	5/20	7/47
phenanthrene	a	5/8	6/20	6/50	7/122
	b	3/5	5/12	6/29	7/71
	c	2/4	3/9	6/21	8/49
	d	2/4	3/9	5/20	8/48
	e	3/5	4/11	5/27	8/63
	f	4/7	5/18	7/43	8/107
	g	3/5	5/12	7/30	9/73
pyrene	a	2/4	4/10	6/22	10/58
	b	3/5	4/11	7/29	11/68
	c	4/7	6/19	9/46	9/121
	d	3/5	5/12	8/31	12/77
	e	6/9	8/23	8/61	10/153
tetracene	a	2/4	3/9	5/20	7/47
	b	3/5	4/11	5/27	8/63
	c	4/7	6/18	7/43	9/107
	d	4/6	6/14	7/37	11/90
	e	4/7	6/19	8/47	10/121
triphenylene	a	2/4	3/9	6/21	10/51
	b	3/5	5/12	7/30	10/75
	c	5/8	7/21	9/54	11/138

[a] Atomic labels are shown in Figure 2.

of walks when they exceed the length of the corresponding paths. Table 2 illustrates new shape indices for carbon atoms of smaller benzenoids illustrated in Figure 2. The enumera-
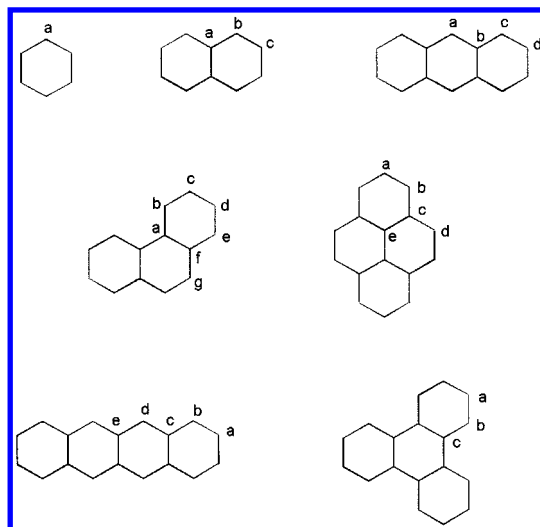
Figure 2. Smaller benzenoids of Table 2.

Table 3. Quotient of the Count of Path and the Count of Walks for Octane Isomers Obtained by Summing Atomic Contributions

	molecule	p_2/w_2	p_3/w_3	p_4/w_4	p_5/w_5
1	*n*-octane	0.45833	0.22321	0.10119	0.04750
2	2-M	0.50208	0.21151	0.08679	0.04301
3	3-M	0.50000	0.26230	0.09393	0.04691
4	4-M	0.49167	0.25524	0.12589	0.03199
5	3-E	0.49167	0.30227	0.13375	0.03554
6	2,2-MM	0.55625	0.20000	0.07813	0.04688
7	2,3-MM	0.54167	0.28769	0.09243	0.03170
8	2,4-MM	0.53958	0.24681	0.10081	0.02941
9	2,5-MM	0.54583	0.20328	0.07487	0.05882
10	3,3-MM	0.55417	0.28297	0.09128	0.01563
11	3,4-MM	0.54167	0.33173	0.09644	0.01471
12	2-M,3-E	0.53810	0.32756	0.12179	0
13	3-M,3-E	0.55982	0.36191	0.08333	0
14	2,2,3-MMM	0.60268	0.30717	0.06140	0
15	2,2,4-MMM	0.59345	0.17950	0.09717	0
16	2,3,3-MMM	0.60774	0.34454	0.04701	0
17	2,3,4-MMM	0.58810	0.30989	0.07692	0
18	2,2,3,3-MMMM	0.66964	0.32143	0	0

tion of paths[94,95] and walks[46-49] has been described in the literature.

MOLECULAR SHAPE INDICES

From the derived atomic path/walk indices we can construct molecular path/walk indices either by adding the corresponding quotients p_i/w_i for all atoms (line p/w at the bottom of Table 1) or by first finding the total number of paths (P_i) in a molecule and the total number of walks (W_i) in a molecule of the corresponding length and then taking the quotient P/W,[96] the line P/W at the bottom of Table 1. We consider both alternatives. We will refer to the derived normalized quotients as p_i/w_i and P_i/W_i shape descriptors. The normalization is obtained by dividing the sum of atomic contributions by n, the number of atoms. In Table 3 we listed for the 18 octane isomers the shape indices p_i/w_i, and in Table 4 we listed for the same for octane isomers the indices P_i/W_i. As one can see by comparing the two tables the corresponding numbers are similar but are not quite equal. Even though the differences are small, they are significant, and they do lead to a visibly different regressions in a some situations (as will be seen in Tables 5 and 6). The descriptor P_2/W_2 shows a relatively high degree of degeneracy (which can also be observed for Kier's kappa indices).

610 *J. Chem. Inf. Comput. Sci., Vol. 41, No. 3, 2001*

RANDIC

Table 4. Quotient of the Count of Path and the Count of Walks for Octane Isomers Obtained as a Quotient of the Total Number of Paths and the Total Number of Walks for a Molecule as a Whole

	molecule	P_2/W_2	P_3/W_3	P_4/W_4	P_5/W_5
1	*n*-octane	0.46154	0.20833	0.08889	0.03571
2	2-M	0.50000	0.19231	0.07692	0.03061
3	3-M	0.50000	0.22222	0.07407	0.02830
4	4-M	0.50000	0.22222	0.09091	0.01869
5	3-E	0.50000	0.24138	0.08621	0.01709
6	2,2-MM	0.56250	0.16667	0.05797	0.02273
7	2,3-MM	0.53333	0.23333	0.06349	0.01550
8	2,4-MM	0.53333	0.20690	0.08065	0.01653
9	2,5-MM	0.53333	0.17857	0.06780	0.03540
10	3,3-MM	0.56250	0.21875	0.05479	0.00676
11	3,4-MM	0.53333	0.25806	0.07273	0.00735
12	2-M,3-E	0.53333	0.25806	0.07576	0
13	3-M,3-E	0.56250	0.26471	0.04000	0
14	2,2,3-MMM	0.58824	0.22857	0.03704	
15	2,2,4-MMM	0.58824	0.15625	0.07692	0
16	2,3,3-MMM	0.58824	0.25000	0.02439	0
17	2,3,4-MMM	0.56250	0.24242	0.05556	0
18	2,2,3,3-MMMM	0.63158	0.22500	0	0

Table 5. Statistical Information for Regression of Selection of Properties of Octane Isomers Using Two Descriptors p_2/w_2 or p_3/w_3 and P_2/W_2, P_3/W_3[a]

	descriptors					
	$p_2/w_2, p_3/w_3$			$P_2/W_2, P_3/W_3$		
property	r	s	F	r	s	F
DH$_f$	0.9405	0.449	57.4	0.9619	0.3612	92.8
steric	0.9781	0.394	165.5	0.9729	0.4378	133.0
BP	0.9141	2.64	38.1	0.9340	2.33	51.3
CT	0.8508	5.45	19.7	0.8384	5.6513	17.7
CP	0.7400	1.03	9.1	0.7572	1.00	10.1
S	0.9503	1.50	69.8	0.9541	1.57	62.8
DH	0.9542	0.123	76.3	0.9899	0.078	201.4
R^2	0.8538	0.100	20.2	0.8547	0.1000	20.3
r	0.9908	0.0017	350.0	0.9887	0.0018	283.4
H_v	0.9520	0.660	72.6	0.9705	0.520	121.7
H_f	0.9387	0.459	55.7	0.9608	0.369	90.2
DF	0.8177	0.519	15.1	0.8429	0.485	18.4
H_c	0.9395	0.460	56.5	0.9606	0.373	89.5
V_c	0.7031	0.0112	6.8	0.2807	0.0197	0.6
T_c	0.9098	3.88	33.7	0.9039	4.00	31.2
P_c	0.9868	0.208	279.3	0.9840	0.229	228.9
C^{13}	0.9115	8.17	36.8	0.9685	4.95	113.5
acentric	0.9197	0.0127	38.4	0.9879	0.0050	283.9
octane #	0.8050	0.114	13.8			
MR	0.9948	0.0201	568.7	0.9927	0.024	440.2

[a] r = the coefficient of regression, s = the standard deviation, and F = Fisher ratio.

In Table 5 we show the statistical parameters for numerous regression of octane physicochemical properties using the shape descriptors p_2/w_2, p_3/w_3 as well as P_2/W_2, P_3/W_3. The large number of physicochemical properties of octanes were previously studied using a pool of over 40 different molecular descriptors.[97] A glance at Table 5 immediately reveals that both pairs of descriptors, p_2/w_2, p_3/w_3 and P_2/W_2, P_3/W_3, appear quite successful topological indices for characterization of isomeric variations of properties among octanes. In both cases well over half of the regressions have the correlation coefficient over 0.9000. In view of the fact that we consider isomers, i.e., molecules having the same size, this is quite an impressive result. Clearly a descriptor that can correlate well the isomeric variations deserves the label "shape index". Notice also a good parallelism between the shape descriptors p/w and P/W. However, the small differ-

Table 6. Statistical Information for Regression of a Selection of Properties of Octanes Using a Single Descriptor p_2/w_2 or p_3/w_3[a]

	descriptor					
	p_2/w_2			p_3/w_3		
property	r	s	F	r	s	F
DH$_f$	0.7353	0.867	18.8	0.3019	1.22	1.6
steric	0.8951	0.818	64.4	0.6748	1.35	13.4
BP	0.7006	4.50	15.4	0.3145	5.99	1.75
CT	0.3075	9.55	1.7	0.6418	7.70	11.2
CP	0.3210	1.40	1.8	0.7361	1.00	18.9
S	0.9466	1.50	137.8	0.4002	4.27	3.1
DH	0.9118	0.164	78.9	0.3988	0.399	0.3
R^2	0.8402	0.101	38.4			
r*	0.3056	0.0112	1.4	0.9898	0.016	725
H_v	0.8735	1.02	51.5	0.0596	2.08	0.05
H_f	0.7395	0.868	19.3	0.2929	1.23	1.5
DF	0.2640	0.843	1.2	0.8176	0.503	32.3
H_c	0.7433	0.870	19.8	0.2882	1.245	1.4
V_c	0.4195	0.014	3.2	0.6624	0.011	11.7
T_c	0.0161	9.03	0.004	0.8650	4.53	46.6
P_c	0.7308	0.850	18.3	0.8719	0.610	50.7
C^{13}	0.1510	19.0	0.4	0.5532	2.05	7.1
acentric	0.9166	0.013	78.8	0.4614	0.028	4.1
octane #	0.7908	15.6	23.4	0.4081	23.3	2.8
MR	0.5173	0.162	4.8	0.9902	0.0016	706.3

[a] r = the coefficient of regression, s = the standard deviation, and F = Fisher ratio.

ences between the two are sufficient to make visible variations in their correlation abilities which are reflected differently for different properties.

The Pitzer's acentric factor mirrors molecular shape in that it is defined to measure a deviation of a molecule from an ideal spherical shape. Hence, if path/walk indices describe molecular shapes, then they ought to give satisfactory regression for the Pitzer's acentric factor as a property which may critically depend on molecular shape. We found for a correlation of the Pitzer's acentric factor of octane using p_2/w_2, p_3/w_3 and P_2/W_2, P_3/W_3 respectively $r = 0.9650$ and $r = 0.9879$.

The best two-variable regression statistics were obtained for the density (ρ), the critical pressure (P_c), and the molar refraction (MR). In all cases the descriptors p_2/w_2, p_3/w_3 have been marginally better than the molecular shape descriptors P_2/W_2, P_3/W_3, but in other cases the opposite was the case. All correlation, with the exception of the critical temperature (CT), the mean radius squared (R^2), and the critical volume (V_c), have the coefficient of regression above 0.9000. Recall that the physicochemical properties considered include numerous properties viewed as dependent on the molecular surface (e.g., BP, H$_f$) and several properties viewed as depending on the molecular volume (e.g., density, MR). It is remarkable that the same pair of topological indices were capable of giving good correlations with so diverse physicochemical properties. However, few properties apparently "refused" to be "conquered". Why these exceptions? Could the experimental data be wrong? We will consider this question in the next section.

We should also add that novel indices listed in Tables 3 and 4 for octane isomers offer a possibility for construction of additional molecular descriptors, to be mentioned briefly later, one of which in particular appears very promising.

ON CLASSIFICATION OF MOLECULAR PROPERTIES

Regressions using a single descriptor have clear advantage when one tries to interpret regression results, because the

Table 7. Classification of Physicochemical Properties of Octanes According the Role of the Descriptors p_2/w_2 and p_3/w_3

dominant p_2/w_2	important p_2/w_2	50:50	important p_3/w_3	dominant p_3/w_3
entropy S	DH_f	steric	CT	density r
DH	BP	P_c	CP	MR
acentric	r^2		DF	
	H_v		T_c	
	H_f		C^{13}	
	H_c			

problem associated with mutual relatedness of descriptors is avoided. Is it possible that a single descriptor, such as p_2/w_2 or p_3/w_3, and alternatively P_2/W_2 or P_3/W_3, would produce a good regression? In Table 6 we show the statistical parameters for simple regressions for the same properties of the octane isomers using a single descriptor, p_2/w_2 or p_3/w_3. As we see from Table 6 p_2/w_2 is a very good descriptor for the entropy S ($r = 0.9476$), the heat of formation DH ($r = 0.9130$), and the Pitzer's acentric factor ($r = 0.9353$). These results are comparable to, even though slightly lower than, the best single descriptors derived by screening over 40 topological indices. The fact that among the properties that are well correlated with a single p_i/w_i index we find Pitzer's acentric factor supports the suggestion that the novel indices reflect the molecular shape.

If we look at the right half of Table 6 we see that p_3/w_3 is an excellent descriptor for the density ($r = 0.9898$). The corresponding Fisher F value is the largest value among all the regressions of Tables 5 and 6 ($F = 725$). This result is visibly better than one previously reported as the best correlation for density of octanes based on the descriptor $^3\chi$ (yielding $r = 0.979$). Observe also that both descriptors p_3/w_3 and $^3\chi$ are intimately related to the number of paths of length three which Wiener[62] introduced already over 50 years ago as an index that characterizes the "crowdedness" of close atoms in a molecule.

There is no rigorous classification of physicochemical properties of alkanes, except for a somewhat vague division of properties as "surface" properties or "volume" properties. The different behavior of physicochemical properties when two descriptors p_2/w_2 and p_3/w_3 are used in correlations offers a novel classification of physicochemical properties. We are now in a position to make a finer distinction among the properties within each of this group. In Table 7 we have classified properties of octanes into five groups, based on the value or the regression coefficients for regressions using p_2/w_2 and p_3/w_3 as molecular descriptors. If the correlation coefficient r is above 0.9000 we refer to the descriptor as dominant, and if r is above 0.7000 the descriptor has been classified as important unless a similar r value appears for the other descriptor. If both descriptors play a similar role we have a "mixed case" (a 50:50 classification). One can accompany such a classification of properties by associating with each property a numerical index given by the quotient of the two regression coefficients. This would allow a finer distinction between properties currently having the same classification; however, at this stage of the development of the quantitative measure such a finer distinction within each classification group appears to not be warranted.

It seems more important to try to understand why properties that show some parallelism also display visible differences. The outlined classification of physicochemical properties depends on the descriptors used and to some extent may be sensitive on the choice of the descriptors. The mathematical descriptors that are structurally related and are not highly interrelated will have advantages for classification of properties as compared to highly interrelated descriptors, whether ad hoc or structurally related. Interrelation obscures the role of individual descriptors in regression analysis[98−104] as well as in similarity studies.[105] When descriptors are highly interrelated (as is the case with the connectivity indices $^1\chi$ and $^2\chi$) they would classify all properties (to which regression applies) as "mixed", 50:50. This would be useless, as it makes no differentiation between properties. In contrast p_2/w_2 and p_3/w_3 have lead to important insight in differentiation among physicochemical properties that has been hitherto overlooked.

DETECTION OF IRREGULARITIES IN DATA AND IMPROVED REGRESSIONS

The next question to consider is why Tables 4 and 5 show considerable variations in the statistical parameters for some properties that are perceived to be related or similar. Could it be that some experimental data are less precise or even erroneous? Or could it be that a structural factor that is relevant for some of the considered physicochemical properties has not been adequately represented by the new descriptors? Which is the case?

In an ideal case the most direct way to detect an experimental error is to make a property−property correlation if the properties are highly interrelated. If a single point in such regressions is an outlier, then we have a strong suspicion that one of the two reported experimental findings is inaccurate, even though it need not be apparent which of the properties is in error. However, in practice, most physicochemical properties that mutually correlate do not show such a high correlation well enough, and the scatter of points around the least-squares line will obscure minor errors that become undetected, yet an error even if not excessive can influence the correlations parameters significantly. If, however, two experimental properties show a high correlation, such as is the case of the boiling points and chromatographic retention indices or between various solubilities, experimental errors can be detected.[34]

Hunting down erroneous data (including misprints), outliers, and, in general, not bona fide members of the population to which the other observations belong need not be so straightforward as it may appear, unless, of course, the outliers are apparent, i.e., they exceed three or four times the standard deviations.[106] We have reexamined closely all the input data on octane isomers and were able to detect several outliers in the data that produced the regressions of Tables 4 and 5. In Table 8 we list the outliers and the revised statistical parameters. By comparing Table 8 with Tables 5 and 6 we see that the regression involving V_c has been dramatically improved, even though apparently V_c remains an elusive property for correlating with most topological indices. In other cases we have minor but significant improvements, if judged by the reduction of the standard deviations.

It is interesting to observe also that by removing an outlier both r and F can worsen, rather than improve! Hence one

Table 8. Improved Statistical Parameters of the Regressions Using p_2/w_2, p_3/w_3, or Both Descriptors for Selected Properties of Octanes by Excluding Outliers

property	outlier	descriptor		
		p_2/w_2	p_3/w_3	p_2/w_2 and p_3/w_3
steric	#15	$r = 0.8998$	$r = 0.8013$	$r = 0.9941$
	#17	$s = 0.79$	$s = 1.09$	$s = 0.2045$
		$F = 59.5$	$F = 25.0$	$F = 546.1$
CP	#18	$r = 0.6557$	$r = 0.8886$	$r = 0.9821$
		$s = 1.09$	$s = 0.66$	$s = 0.28$
		$F = 11.3$	$F = 56.3$	$F = 190.2$
CT	#18	$r = 0.0666$	$r = 0.8650$	$r = 0.9253$
		$s = 9.05$	$s = 4.55$	$s = 3.56$
		$F = 0.07$	$F = 44.6$	$F = 41.7$
H_v	#17	$r = 0.9076$	$r = 0.2434$	$r = 0.9394$
		$s = 0.77$	$s = 1.78$	$s = 0.65$
		$F = 70.1$	$F = 0.9$	$F = 52.6$
r	#18		$r = 0.9934$	
	#10		$s = 0.0014$	
			$F = 1053.3$	
DH	#17	$r = 0.9330$		
		$s = 0.132$		
		$F = 100.8$		
S	#18	$r = 0.9145$	$r = 0.3592$	$r = 0.9382$
		$s = 1.42$	$s = 3.23$	$s = 1.26$
		$F = 76.6$	$F = 2.8$	$F = 51.4$
P_c	#18	$r = 0.6150$	$r = 0.9165$	$r = 0.9874$
		$s = 0.89$	$s = 0.45$	$s = 0.185$
		$F = 9.1$	$F = 78.8$	$F = 273.2$
V_c	#5	$r = 0.4355$	$r = 0.6530$	$r = 0.9874$
		$s = 0.014$	$s = 0.0115$	$s = 0.185$
		$F = 3.5$	$F = 11.1$	$F = 273.2$

should base the decision when eliminating an outlier on its influence on the standard error. The regressions when the outliers have been discarded call for a reclassification of P_c as a property with the dominant p_3/w_3 descriptor and H_v as a property with the predominant p_2/w_2.

ACKNOWLEDGMENT

REFERENCES AND NOTES

(1) Randić, M. Topological Indices. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III., Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, 1998; pp 3018−3032.

(2) Randić, M.; Novic, M.; Vracko, M. *Molecular Descriptors, New and Old*; Lecture Notes in Chemistry; Submitted for publication.

(3) Randić, M. On characterization of chemical structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672−687.

(4) Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. Topological indices for structure−activity correlations. In Steric effects on drug design; Charton, M., Motoc, I., Eds.; *Topics Curr. Chem.* **1983**, *114*, 21−55.

(5) Balaban, A. T. *Historical developments of topological indices.* In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, The Netherlands, 1999; pp 403−453.

(6) Balaban, A. T. Topological indices and their uses: a new approach for coding alkanes. *J. Mol. Struct. (THEOCHEM)* **1988**, *165*, 243−253.

(7) Basak, S. C. Information theoretic indices of neighborhood complexity and their applications. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, The Netherlands, 1999; pp 563−593.

(8) Balaban, A. T. From Chemical Graphs to 3D Molecular Modeling. In *From Chemical Topology to Three-dimensional Geometry*; Plenum Press: New York, 1977; pp 1−24.

(9) Trinajstić, N. In *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1992; Chapter 10, pp 225−273.

(10) Basak, S. C.; Grunwald, G. D.; Niemi, G. J. Use of graph-theoretic geometrical molecular descriptors in structure−activity relationships. In *From Chemical Topology to Three-dimensional Geometry*; Plenum Press: New York, 1977; pp 73−116.

(11) Randić, M.; Razinger, M. On characterization of 3D molecular structure. In *From Chemical Topology to Three-dimensional Geometry*; Plenum Press: New York, 1977; pp 159−236.

(12) Mihalić, Z.; Trinajstić, N. The algebraic modeling of chemical structures: On the development of three-dimensional molecular descriptors. *J. Mol. Struct. (THEOCHEM)* **1991**, *232*, 65−78.

(13) Bogdanov, B.; Nikolić, S.; Trinajstić, N. On the three-dimensional Wiener number. *J. Math. Chem.* **1988**, *3*, 299−309.

(14) Bogdanov, B.; Nikolić, S.; Trinajstić, N. On the three-dimensional Wiener number. A comment. *J. Math. Chem.* **1990**, *5*, 305−306.

(15) Bogdanov, B. On the three-dimensional Wiener number. *J. Math. Chem.* **1989**, *3*, 299−309.

(16) Diudea, V.; Horvath, D.; Graovac, A. Molecular Topology. 15. 3D distance matrices and related topological indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 129−135.

(17) Estrada, E. Three dimension molecular descriptors based on charge density weighted graphs. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 708−713.

(18) Zhu, H.-Y.; Klein, D. J. Graph-geometric invariants for molecular structure. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1067−1075.

(19) Randić, M.; Razinger, M. Molecular topographic indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 140−147.

(20) Randić, M. Molecular topographic descriptors. *Stud. Phys. Theor. Chem.* **1988**, *54*, 101−108.

(21) Randić, M. On the characterization of three-dimensional structures. *Int. J. Quantum Chem: Quantum Biol. Symp.* **1988**, *15*, 201−208.

(22) Randić, M.; Jerman-Blazić, B.; Trinajstić, N. Development of 3-dimensional molecular descriptors. *Computer Chem.* **1990**, *14*, 237−246.

(23) Randić, M. Molecular shape profiles. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 373−382.

(24) Randić, M. Molecular profiles. Novel geometry-dependent molecular descriptors. *New J. Chem.* **1995**, *196*, 781−791.

(25) Kier, L. B.; Hall, L. H. Derivation and significance of valence molecular connectivity. *J. Pharm. Sci.* **1981**, *70*, 583−589.

(26) Kier, L. B.; Hall, L. H. Molecular connectivity VII: Specific treatment of heteroatoms. *J. Pharm. Sci.* **1976**, *65*, 1806−1809.

(27) Balaban, A. T. Chemical graphs. 48. Topological index J for heteroatom-containing molecules taking into account periodicities of element properties. *MATCH* **1986**, *21*, 115−122.

(28) Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. Design of topological indices. Part 10. Parameters based on electronegativity and covalent radius for the computation of molecular graph descriptors for heteroatom-containing molecules. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 395−401.

(29) Kupchik, E. J. General treatment of heteroatoms with the Randić Molecular connectivity index. *Quant. Struct.-Act. Relat.* **1989**, *8*, 98−103.

(30) Randić, M. Novel Graph theoretical approach to heteroatom in quantitative structure−activity relationship. *Chemometrics Intel. Lab. Systems* **1991**, *10*, 213−227.

(31) Randić, M. On computation of optimal parameters for multivariate analysis of structure−property relationship. *J. Comput. Chem.* **1991**, *12*, 970−980.

(32) Randić, M.; Dobrowolski, J. Cz. Optimal molecular connectivity descriptors for nitrogen containing molecules. *Int. J. Quantum Chem.* **1998**, *70*, 1209−1215.

(33) Randić, M.; Basak, S. C. Optimal molecular descriptors based on weighted path numbers. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 261−266.

(34) Randić, M.; Basak, S. C. Multiple regression analysis with optimal molecular descriptors. *SAR QSAR Environ. Res.* **2000**, *11*, 1−23.

(35) Randić, M.; Pompe, M. On characterization of CC double bond n alkanes. *SAR QSAR Environ. Res.* **1999**, *10*, 451−471.

(36) Randić, M. High quality structure−property regressions. Boiling points of smaller alkanes. *New J. Chem.* **2000**, *24*, 328−336.

(37) Randić, M.; Pompe, M. The variable molecular descriptors based on distance related matrices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 575−581.

(38) Randić, M.; Plavšić, D.; Nerš, N. Variable connectivity index for cycle−containing structures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 657−662.

(39) Randić, M.; Pompe, M. The variable connectivity index $^1\chi^f$ versus the traditional molecular descriptors: A comparative study of $^1\chi^f$ against descriptors of CODESSA. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 631−638.

(40) Randić, M.; Basak, S. C. On use of the variable connectivity index in QSAR: toxicity of aliphatic ethers. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 614−618.

(41) Pompe, M.; Veber, M.; Randić, M. Prediction of reaction rate constants for the reaction of different organic compounds with OH radical using variable connectivity indices. *J. Chem. Inf. Comput. Sci.* Submitted for publication.

(42) Randić, M.; Basak, S. C.; Pompe, M.; Novic, M. Prediction of Gas Chromatographic Retention Indices Using Variable Connectivity Index. *Acta Chim. Slovenica* In press.

(43) Platt, J. R. Prediction of isomeric differences in paraffin properties. *J. Phys. Chem.* **1952**, *56*, 328−336.

(44) Razinger, M. Discrimination and ordering of chemical structures by the number of walks. *Theor. Chim. Acta* **1986**, *70*, 365−378.

(45) Razinger, M. Extended connectivity in chemical graphs. *Theor. Chim. Acta* **1982**, *61*, 581−586.

(46) Rücker, G.; Rücker, C. Counts of all walks as atomic and molecular descriptors. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 683−695.

(47) Randić, M. Random walks and their diagnostic value for characterization of atomic environments. *J. Comput. Chem.* **1980**, *1*, 386−399.

(48) Ivanciuc, O.; Balaban, A. T. The characterization of chemical structures by the atomic counts of self-returning walks: on the construction of isocodal graphs. *Croat. Chem. Acta* **1996**, *69*, 66−74.

(49) Rücker, G.; Rücker, C. On using adjacency matrix power method for perception of symmetry and for isomorphism testing of highly intricate graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 123−126.

(50) Randić, M. Characterization of atoms, molecules, and classes of molecules based on path enumerations. *MATCH* **1979**, *7*, 3−60.

(51) Randić, M.; Wilkins, C. L. On graph theoretical basis for ordering of structures. *Chem. Phys. Lett.* **1979**, *63*, 332−336.

(52) Randić, M.; Wilkins, C. L. Graph theoretical ordering of structures as a basis for systematic searches for regularities in molecular data. *J. Phys. Chem.* **1979**, *83*, 1525−1540.

(53) Randić, M.; Wilkins, C. L. Graph-based fragment search in polycyclic structures. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 23−31.

(54) Randić M.; Wilkins, C L. Graph theoretical approach to recognition of structural similarity in molecules. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 31−37.

(55) Randić, M.; Wilkins, C. L. Graph theoretical study of structural similarity in benzomorphans. *Int. J. Quantum Chem: Quantum Biol. Symp.* **1979**, *6*, 55−71.

(56) Randić, M.; Wilkins, C. L. A procedure for characterization of the rings of a molecule. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 36−46.

(57) Wilkins, C. L.; Randić, M. A graph theoretical approach to structure-property and structure−activity correlations. *Theor. Chim. Acta* **1980**, *58*, 45−68.

(58) Randić, M. On characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

(59) Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. Molecular Connectivity concept applied to density. *J. Pharm. Sci.* **1976**, *65*, 1226−1230

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

(61) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press: Letchworth, England, 1986.

(62) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(63) Hosoya, H. Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.

(64) Randić, M. Novel molecular descriptor for structure−property studies. *Chem. Phys. Lett.* **1993**, *211*, 478−483.

(65) Randić, M.; Guo, X.; Oxley, T.; Krishnapriyan, H. Wiener matrix: Source of novel graph invariants. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 709−716.

(66) Randić, M.; Guo, X.; Oxley, T.; Krishnapriyan, H.; Naylor, L. C. Wiener matrix invaraints. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 361−367.

(67) Hermann, A.; Zinn, P. List operations on chemical graphs. 6. Comparative study of combinatorial topological indexes of the Hosoya type. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 551−560.

(68) Randić, M.; Morales, D. A.; Araujo, O. Higher order Fibonacci numbers. *J. Math. Chem.* **1996**, *20*, 79−94.

(69) Plavšić, D.; Šoškić, M.; Landeka, I.; Gutman. I.; Graovac, A. On the relation of the path numbers ^1Z, ^2Z and the Hosoya Z index. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1118−1122.

(70) Kier, L. B. Shape indexes of orders one and three from molecular graphs. *Quant. Struct.-Act. Relat.* **1986**, *5*, 1−7.

(71) Kier, L. B. Inclusion of symmetry as shape attributes in kappa index analysis. *Quant. Struct.-Act. Relat.* **1987**, *6*, 8−12.

(72) Kier, L. B. Indexes of molecular shape from chemical graphs. *Acta Pharm. Jugosl.* **1986**, *36*, 171−188.

(73) Kier, L. B. A structure based approach to molecular shape. *QSAR in Drug Design and Toxicology*; Hadzi, D., Jerman-Blazic, B., Eds.; Elsevier: Amsterdam, 1987.

(74) Kier, L. B. Indexes of Molecular shape from Chemical Graphs. *Medicinal Res. Rev.* **1987**, *70*, 417−440.

(75) Kier, L. B. Indexes of Molecular shape from Chemical Graphs. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Sci. Publ.: New York, 1990; pp 151−174.

(76) Pitzer, K. S. The volumetric and thermodynamic properties of fluids. I. Theoretical basis and virial coefficients. *J. Am. Chem. Soc.* **1955**, *97*, 3427−3433.

(77) Pitzer, K. S.; Lippmann, D. Z.; Curl, R. F.; Huggins, C. M.; Peterson, D. E. The volumetric and thermodynamic properties of fluids. II. Compressibility factor, vapor pressure and entropy of vaporization. *J. Am. Chem. Soc.* **1955**, *97*, 3433−3440.

(78) Randić, M. On characterization of shape of molecular graphs. *SAR QSAR Environ. Res.* Submitted for publication.

(79) Randić, M.; Kleiner, A. F.; DeAlba, L. M. Distance/distance matrices. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 277−286.

(80) Randić, M.; Krilov, G. On characterization of folding of proteins. *Int. J. Quantum Chem.* **1999**, *75*, 1017−1026.

(81) Randić, M. On Characterization of Cyclic Structures. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1063−1071.

(82) Randić, M.; DeAlba, M. L.; Harris, F. E. Graphs with the same detour matrix. *Croat. Chem. Acta* **1998**, *71*, 53−68.

(83) Bytautas, L.; Klein, D. J.; Randić, M.; Pisanski, T. Foldedness in linear polymers: A difference between graphical and Euclidean distances. *DIMACS Ser. Discrete Math., Theor. Comput. Sci.* **2000**, *51*, 39−61.

(84) Randić, M. On characterization of DNA primary sequences by condensed matrix. *Chem. Phys. Lett.* **2000**, *317*, 29−34.

(85) Randić, M.; Vračko, M. On similarity of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 599−606.

(86) Randić, M.; Nandy, A.; Basak, S. C. On numerical characterization of DNA primary sequences. *J. Math. Chem.* In press.

(87) Randić, M.; Vracko, M.; Nandy, A.; Basak, S. C. On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235−1244.

(88) Randić, M. On the graphical representation of proteomics and their numerical characterization. *J. Chem. Inf. Comput. Sci.* In press.

(89) Randić, M.; Zupan, J.; Novic, M. On 3-D Graphical representation of proteomics maps. *J. Chem. Inf. Comput. Sci.* In press.

(90) Lukovits, I.; Razinger, M. On calculation of the detour index. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 283−286.

(91) Trinajstić, N.; Nikolić, S.; Mihalić, Z. On computing the molecular detour matrix. *Int. J. Quantum Chem.* **1997**, *65*, 415−419.

(92) Amic, D.; Trinajstić, N. On the detour matrix. *Croat. Chem. Acta* **1997**, *37*, 283−286.

(93) Nikolić, S.; Trinajstić, N.; Juric, A.; Mihalić, Z. The detour matrix and the detour index of weighted graphs. *Croat. Chem. Acta* **1996**, *69*, 1577−1591.

(94) Randić, M.; Brissey, G. M.; Spencer, R. E.; Wilkins, C. L. Search for all self-avoiding paths for molecular graphs. *Comput. Chem.* **1979**, *3*, 65−13.

(95) Jurs, P. C. *Computer Software applications in Chemistry*; Wiley: New York, 1986; pp 162−169.

(96) Randić, M. On characterization of molecular attributes. *Acta Chim. Slovenica* **1998**, *45*, 239−252.

(97) Randić, M. Comparative regression analysis. Regressions based on a single descriptor. *Croat. Chem. Acta* **1993**, *66*, 289−312.

(98) Randić, M. Resolution of ambiguities in structure−property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311−320.

(99) Randić, M. Orthogonal molecular descriptors. *New J. Chem.* **1991**, *15*, 5179−525.

(100) Randić, M. Search for optimal molecular descriptors. *Croat. Chem. Acta* **1991**, *64*, 43−54.

(101) Randić, M. Correlation of enthalpy of octanes with orthogonal connectivity indices. *J. Mol. Struct. (THEOCHEM)* **1991**, *2334*, 453−59.

(102) Randić, M. Curve-fitting paradox. *Int. J. Quantum Chem: Quantum Biol. Symp.* **1994**, *21*, 215−225.

(103) Randić, M.; Seybold, P. G. Molecular shape as a critical factor in structure−property-activity studies. *SAR QSAR Environ. Res.* **1993**, *1*, 77−85.

(104) Šoškić, M.; Plavšić, D.; Trinajstić, N. Link between orthogonal and standard multiple regression models. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 829−832.

(105) Randić, M. Orthosimilarity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 6, 1092−1097.

(106) Taylor, R.; Allen, F. H. Statistical and numerical methods of data analysis. In *Structure Correlation*; Burgi, H.-B., Dunitz, J. D., Eds.; VCH: Weinheim, 1994; Vol. 1, pp 111−161.