Overall Connectivities/Topological Complexities: A New Powerful Tool for QSPR/QSAR[†]

Danail Bonchev[‡]

Program for Theory of Complex Systems, Texas A&M University, Galveston, Texas 77553-1675

Received September 9, 1999

Earlier attempts to assess the complexity of molecules are analyzed and summarized in a number of definitions of general and topological complexity. A concept which specifies topological complexity as *overall connectivity*, and generalizes the idea of molecular connectivities of Randić, Kier, and Hall, is presented. Two overall connectivity indices, TC and TC1, are defined as the connectivity (the sum of the vertex degrees) of *all* connected subgraphs in the molecular graph. The contributions to TC and TC1, which originate from all subgraphs having the same number of edges *e*, form two sets of *e*th-order overall connectivities, ^{e}TC and ^{e}TC1. The total number of subgraphs *K* is also analyzed as a complexity measure, and the vector of its *e*th-order components, ^{e}K, is examined as well. The TC, TC1, and *K* indices match very well the increase in molecular complexity with the increase in the number of atoms and, at a constant number of atoms, with the increased degree of branching and cyclicity of the molecular skeleton, as well as with the multiplicity of bonds and the presence of heteroatoms. The potential of the three sets of *e*th-order complexities for applications to QSPR was tested by the modeling of 10 alkane properties (boiling point, critical temperature, critical pressure, critical volume, molar volume, molecular refraction, heat of formation, heat of vaporization, heat of atomization, and surface tension), in parallel with Kier and Hall's molecular connectivity indices $^{k}\chi$. The topological complexity indices were shown to outperform molecular connectivity indices in 44 out of the 50 pairs of models compared, including all models with four and five parameters.

DEVELOPMENT OF THE CONCEPT OF MOLECULAR CONNECTIVITY

The idea of chemical structure lies at the very heart of chemical science. Introduced more than a century ago by Butlerov, this idea is best expressed at the threshold of the 21st century by applying the mathematical tools of topology and graph theory.[1−3] Molecular graph *G* provides the basis for a quantitative characterization of chemical structure. The simplest number that can be associated with chemical structure is the *graph adjacency, A(G)*, which is the sum of all entries of the adjacency matrix of the graph. However, this simplest topological index is extremely degenerate; it has the same numerical value for all graphs having the same number of edges *E*.

Various attempts have been reported to express the connectivity of atoms in the molecule by more discriminating graph invariants. Morgan[4] in 1965 introduced for the purposes of chemical documentation the concept of *extended connectivity*, according to which the vertex degrees *i* are recalculated in successive steps as sums of the vertex degrees of the neighboring vertexes *j* until the same vertex ordering results in two consecutive steps. The sum of the extended vertex degrees $^{k}a_i$ in step *k* is called *k*th-order extended connectivity, ^{k}EC(*G*):

$$^{k}\mathrm{EC}(G) = \sum_{i=1}^{n} {}^{k}a_i = \sum_{i=1}^{n}\sum_{j=1}^{{}^{0}a_i} {}^{k-1}a_{ji} \qquad (1)$$

Razinger et al.[5] showed that the ^{k}EC values are connected to the respective *k*th powers of the adjacency matrix. Later, Rücker and Rücker[6] rigorously derived this relationship by proving that the vertex (or atom) walk count, ^{k}awc(*i*), and the graph (or molecule) walk count, ^{k}mwc(*G*), of length *k* are identical to Morgan's $^{k}a_i$ and ^{k}EC(*G*), respectively. Extended connectivity indices found application to structure−property studies.[8]

Another line of development has been initiated by Gutman and Trinajstić, et al.[8] Their first Zagreb group index M1 was defined as the sum of the squared vertex degrees (rather than a simple sum), whereas the second Zagreb group index M2 is the sum over all edges of the product of the vertex degrees of the pairs of neighboring vertexes:

$$\mathrm{M1} = \sum_{i=1}^{n} a_i^{\,2}; \quad \mathrm{M2} = \sum_{\text{all edges}} a_i a_j \qquad (2)$$

Randić[9] transformed M2 into an inverse square-root function χ:

$$\chi = \sum_{\text{all edges}} (a_i a_j)^{-1/2} \qquad (3)$$

Kier and Hall[10−12] extended this idea from edges (paths of length one) to paths of length two, three, etc. Molecular

[†] Dedicated to Milan Randić, Lemont B. Kier, and Lowell H. Hall.
[†] E-mail: bonchevd@aol.com.

10.1021/ci990120u CCC: $19.00
Published on Web 06/01/2000

descriptors thus constructed were termed *molecular connectivities* of first, second, third, etc. order, respectively. A zero-order index was defined for completeness:

$$^{0}\chi = \sum_{i=1}^{n}(a_i)^{-1/2}; \quad ^{1}\chi = \sum_{\text{all edges}}(a_i a_j)^{-1/2}; \quad ^{2}\chi = \sum_{\text{all 2-paths}}(a_i a_j a_k)^{-1/2}; \quad ... \quad (4)$$

Kier and Hall have also made an important extension of their approach to molecules with heteroatoms by introducing an analogous set of *valence connectivity* indices. The latter are defined by a modified formula (eq 4), the vertex degrees in which are substituted by their respective heteroatomic equivalent, the atomic valence connectivities δ_i^{v}. The latter (for main group elements) are equal to the number of valence electrons diminished by the number of adjacent hydrogen atoms. Molecular connectivity indices have found a large application to QSPR/QSAR studies, and to the design of new drugs and chemical products. The concept of molecular connectivity prompted the construction of a variety of novel molecular descriptors. We shall briefly mention Balaban's J index[13] (substituting vertex degrees in eq 4 by the vertex distance degrees), the electrotopological index of Kier and Hall,[14] Toropov and collaborators' 3D-molecular connectivity,[15] the edge connectivity/line graph connectivity indices of Estrada and Gutman,[16] and Randić's very recent augmented valence connectivity.[17]

MOLECULAR COMPLEXITY

Another line of development in theoretical chemistry is related to the idea of molecular complexity.[18,19] The term "*molecular complexity*" was used in its quantitative context for the first time by Bertz[20,21] in 1981. However, quantitative assessments of chemical structures can be traced back to the 1950s,[22] followed in the 1960s by a rigorous information-theoretic analysis of the *complexity of graphs* by Mowshowitz.[23] Information theory is widely used in other sciences as a measure of complexity.[24,25] However, it has been shown[26−28] that the criterion of equivalence of structural elements used in this approach is not a good choice for measuring *topological* or *structural* complexity, because symmetry is a *simplifying* factor. The complexity of graphs has been evaluated by different graph invariants by Minoli,[29] whose "combinatorial complexity" includes the total number of vertexes n, edges e, and paths p in the graph

$$\mathrm{CC}(G) = nep/(n + e) \quad (5)$$

Unfortunately, this index is highly degenerate; it cannot distinguish the complexity of isomeric acyclic compounds for which it is a constant number: $\mathrm{CC(acyclics)} = n^2(n - 1)^2/(4n - 2)$. Equation 5 was later modified by replacing the number of paths by their total length.[30] A further improvement is proposed here by substituting the total number of paths p by the count of self-returning walks, SRW, or that of all walks, mwc, graph invariants with a high sensitivity[6] toward the patterns of branching and cyclicity.[31−37]

$$\mathrm{CC1}(G) = ne(\mathrm{SRW})/(n + e); \quad \mathrm{CC2}(G) = ne(\mathrm{mwc})/(n + e) \quad (6)$$

The resulting complexity indices CC1 and, particularly, CC2 have very low degeneracy (as shown by Rücker and Rücker,[6] there is no degeneracy in the values of mwc at least up to $n = 12$).

Another complexity measure, applicable to cyclic graphs only, is the count of spanning trees (a spanning tree of a graph contains all the vertexes of the graph but no cycles) proposed by Nikolić et al.[38] The number of spanning trees in chemically relevant graphs was evaluated earlier by Gutman et al.[39] Also earlier, the number of spanning trees was used by Bonchev and Temkin et al.[40−42] for the evaluation of the complexity of the cyclic graphs used in chemical kinetics, and for the classification of chemical reaction networks.

An analysis of the quantitative concepts of structural complexity in chemistry is presented in the next section.

CONCEPTS AND DEFINITIONS OF COMPLEXITY

The views of the pioneers of an information-theoretic approach to complexity (Rashevsky,[22] Kolmogorov,[43] and others) may be summarized in definition 1.

Definition 1: The higher the information content of a system, the more complex the system.

As has already been shown,[26,27] this definition is appropriate when the complexity of the elemental composition of a molecule is assessed; however, it has very serious flaws when applied to the evaluation of the *structural* complexity. The latter was approached in a better way by Minoli.[29] One may paraphrase his complexity criteria for graphs in definition 2.

Definition 2: The larger the size of the system, and the higher the degree of its connectedness, the more structurally complex the system.

Conceptually correct, Minoli's views on the complexity of graphs were expressed quantitatively by a formula (see the previous section) which was not sensitive enough to the enormous variety of molecular architectures, and was entirely useless for acyclic structures.

In 1983, Bertz[21] proposed the first *hierarchical* concept of molecular complexity. According to this concept, the basics of complexity are determined by topology and elemental composition, followed by size, symmetry, and functionality. Topology itself is represented in two levels: level 1 with rings and multiple bonds, and level 2 with branching. Bertz's $C(\eta)$ complexity index can be decomposed into terms for size, symmetry, and elemental composition. In addition, his size term is in a state to match the number of rings, branches, and multiple bonds. One might try to convert Bertz's hierarchical scheme into definition 3.

Definition 3: Molecular complexity increases with the number of rings, multiple bonds, and branches, as well as with molecular size, and with the variety in elemental composition and functional groups, and it decreases with symmetry.

With definitions of complexity becoming more complex themselves, a necessity arose for a separate, more detailed concept of *topological complexity*. This was done in 1987 by Bonchev and Polansky[27] in a hierarchical scheme which includes levels of connectedness, adjacency, connectivity patterns (linearity, bridging, branching, and cyclicity), symmetry, metrics (distances, paths, walks, etc.), and subgraphs (two-edge subgraphs, three-edge subgraphs, etc.). On its turn, topological complexity was regarded as a part of the *general complexity* in a hierarchical scheme that starts with the size, and continues with topology, physical nature (e.g., the nature of atoms and bonds), a specific metric (e.g., 3D geometry), and specific symmetry (e.g., point groups of symmetry). The mathematical model of complexity was regarded as a vector with components representing each of the hierarchical levels. Separate definitions of general and topological complexity may be extracted from the concept of Bonchev and Polansky.[27]

Definition 4A: The general complexity of any system increases with the increase in the complexity of its size, topology, specific nature, and metric, and with the decrease in its symmetry.

Definition 4B: The topological complexity of any system increases with the increase in its connectedness and adjacency, with the presence of more complex connectivity patterns (cycles and branches), with the larger number of distances, paths, and walks, with the larger number of subgraphs of increasing size, and with the lower topological symmetry.

Recent development in this area will be discussed in the next section.

TOPOLOGICAL COMPLEXITY AS OVERALL CONNECTIVITY

Albeit the topological complexity approach of Bonchev and Polansky[27] seems conceptually correct, the quantitative measure suggested has not been discriminating enough for isomeric molecules. A new approach was developed by me,[26,44,45] with the first results being reported at the 12th Southwest Theoretical Chemistry Conference in Arlington, TX, in 1996. The main idea is to extend the concept of molecular connectivity, so as to better characterize the complexity of the molecular graph. The new concept may be formulated as another definition of topological complexity.

Definition 5: The higher the connectivity of the molecular graph *and its connected subgraphs*, the more topologically complex the molecule.

We have already mentioned that the total adjacency $A(G)$ of the graph G is a highly degenerate quantity. A much more sensitive measure, termed the *topological complexity index* $TC(G)$ of the graph G, may be defined as the sum of the total adjacencies of all ${}^eK$ connected subgraphs having e edges and n_t vertexes of degree a_i, including the graph itself which has E edges and K connected subgraphs; summarizing the information on the connectivity of vertexes in all subgraphs, the new index has the meaning of the *overall connectivity* of G:

$$\mathrm{TC}(G) = \sum_{e=0}^{E} {}^e\mathrm{TC}(G) = \sum_{e=0}^{E} \sum_{t=1}^{{}^eK} \sum_{i=1}^{n_t} a_i(G) \qquad (7)$$

The ${}^e\mathrm{TC}(G)$ is the topological complexity of order e, e.g., ^{0}TC is the zero-order complexity (complexity of vertexes), ^{1}TC is the first-order complexity (complexity of edges), ^{2}-TC is the second-order complexity (complexity of two-edge subgraphs), etc. Ordered in a sequence they form the vector of topological complexity (overall connectivity) $\mathrm{TC}'(G)$:

$$\mathrm{TC}'(G) = \mathrm{TC}({}^0\mathrm{TC},\ {}^1\mathrm{TC},\ {}^2\mathrm{TC},\ ...,\ {}^e\mathrm{TC}) \qquad (8)$$

The *e*th-order overall connectivities may be regarded as a generalization of Kier and Hall's vertex molecular connectivities[10−12] (eq 4) and the edge connectivities of Estrada[16] which have been defined (with the Randić[9] inverse-square root function) for the entire molecular graph but not for all subgraphs of it. The use of $\mathrm{TC}'(G)$ in parallel with $\mathrm{TC}(G)$ further increases the discrimination ability (vide infra) of the overall connectivity index for which the summation in eq 7 could by chance produce two nonisomorphic graphs having the same TC value.

The $\mathrm{TC}(G)$ index is defined in two quantitatively different versions. In the basic version, the vertex degrees a_i are those in the entire graph G; in the second version, the $\mathrm{TC1}(G)$ index is calculated with the vertex degrees taken from the corresponding subgraph G_t.

Essential for the definition of our topological complexities/overall connectivities is the description of the molecular graph in full detail by counting all connected subgraphs K. The index K itself can be used as a measure of the topological complexity although one may expect it to be more degenerate, due to the simple summation procedure. In our analysis, we shall present for comparison the K values as well, along with its components ${}^eK$ which will also be ordered in a vector form:

$$K(G) = \sum_{e=0}^{E} {}^eK;\quad K'(G) = K({}^0K,\ {}^1K,\ {}^2\mathrm{K},\ ...,\ {}^EK) \qquad (9)$$

Formulas for the *K*, *TC1*, and *TC*, as well as for their *e*th-order components, are given elsewhere[45] for the classes of chain graphs, star graphs, and monocyclic graphs.

The problem for finding all subgraphs of a molecular graph has been of interest mainly related to drug design,[46−48] as well as to predictions of physicochemical[49] and spectral properties.[50] Until recently, however, the only publication using explicitly the total number of subgraphs $K(G)$ as a measure of molecular similarity and complexity was that of Bertz and Herndon[51] published in 1986. (We thus correct our previous reference[44,45] to Gordon and Kennedy[49] as the first authors using the total number of subgraphs as a molecular descriptor. In fact, these authors dealt with selected components of $K(G)$ only in a scheme for calculating molecular properties. On the other hand, it was Matula[52] who in 1970 first calculated the total number of subgraphs in a class of acyclic graphs, which however were of little relevance to chemistry.) During the last several years, simultaneously with our studies,[26,44] other publications on $K(G)$ appeared. Bone and Villar[53] presented an extensive enumeration of molecular substructures in some chemically important classes of compounds. They found that "the number of constituent substructures of each size are related to the molecular topology, in particular the degree of branching". Analytical expressions for $K(G)$ have been provided for several simple classes of molecules, along with the substructure distribution shapes. Bertz and Sommer[54] proposed to use the total number of subgraphs and the number of kinds of subgraphs as two new complexity indices and analyzed briefly their application to strategic bonds and synthetic analysis. In more detail, these ideas were presented by Bertz and Wright[28] in a paper published in 1998. Thus,

Table 1. Complexity Index *K* and Topological Complexity Indices TC and TC1 of Graphs **1**−**5**

index	1	2	3	4	5
K	15	17	20	26	54
TC1	40	50	64	110	310
TC	60	76	100	160	482

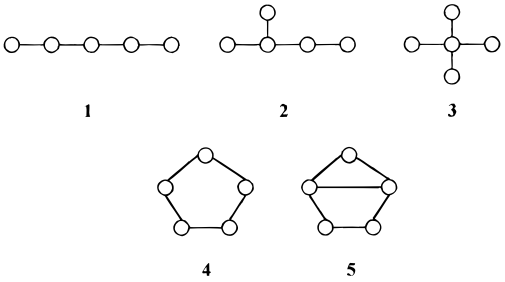


Figure 1. Five five-vertex graphs ordered according to their increasing structural complexity.

the idea of Bertz and Herndon has been further developed, and the number of substructures was recognized as a very convenient basis for assessments of molecular complexity, whereas the sets of *e*th-order complexity introduced by us have shown potential for structure/property and structure/activity studies.

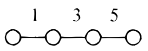
The two indices of topological complexity TC and TC1 and their predecessor, the *K* index, are illustrated in Table 1 for graphs **1**−**5** from Figure 1. All three indices increase in a regular way on going from the linear structure to the monobranched structure, and then to the dibranched, the monocyclic, and the bicyclic structures. Thus, they all show the potential to be used as measures of molecular branching and cyclicity, the two major features of topological complexity. The synchronous change of the *K*, TC1, and TC indices for graphs **1**−**5** occurs for isomers with a relatively small number of vertexes. With the increase of the number of isomers, e.g., in the sets of 35 C9 and 75 C10 alkane isomers, some typical cases of reordering appear, due to the increasing role of the centrality factor (the positioning of branches with respect to the graph center). These sets are analyzed, along with a series of cyclic isomers, in a forthcoming paper,[55] as well as in a chapter in preparation for the special volume of the *Mathematical Chemistry Series* devoted to complexity in chemistry.[56] Here, we show in Table 2 the ordering of the C8 alkane isomers according to the three complexity measures.

One of the criteria which a complexity measure should satisfy[27] is to distinguish well isomeric structures. Our analysis shows that the discriminatory power of the three complexity measures is the lowest in the *K* index, and the highest in the TC index. Thus, for the sets of 18 C8, 35 C9, and 75 C10 alkanes, a total of 128 structures, the *K* index has 103 different values; i.e., its degree of degeneracy[57] is 1.24. For the TC1 and TC indices the numbers of different values are 125 and 127, respectively (degrees of degeneracy 1.02 and 1.008, respectively). The only pair of isomers for which our basic index of topological complexity, TC, cannot produce a different value is 2,3,4-trimethylpentane and 3-methyl-3-ethylpentane (TC = 532). However, this degeneracy results from a *chance* compensation of the summands which differ considerably. (This can certainly be avoided by constructing a squared sum of the *e*th-order terms, in the style of the first Zagreb group index[8] M1). The two vectorial indices TC′ (eq 8) of the two isomers are, respectively, 532 (14, 32, 61, 98, 128, 120, 65, 14) and 532 (14, 32, 66, 115, 134, 105, 52, 14). Thus, the use of TC′ discriminates all alkane isomers up to C10. It remains a challenge to find the first pair of alkane isomers for which the TC′ values of the two isomers are the same. One may then compare the performance of TC′ with that of the molecular walk count of Rücker and Rücker,[6] which not only shows no degeneracy for isomeric alkanes having up to C12, but also seems to satisfy the most important criteria for a good complexity measure.[58]

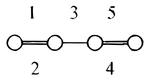
Our analysis has also shown that the topological complexity index TC can be regarded as an adequate mathematical model for the hierarchical concept of topological complexity of Bonchev and Polansky.[27] Thus, TC increases with the general connectedness of the system (the first level of topological complexity) on going from disconnected graphs to connected planar graphs to connected nonplanar graphs. It also increases, on the second hierarchical level, from directed graphs to nondirected graphs and to multigraphs. It also increases, on the third hierarchical level, from linear molecular structures to structures having an increasing number of branches and cycles, as well as with more subtle topological factors such as the more central location of branches and cycles and their clustering, size, etc.

MULTIPLE BONDS AND HETEROATOMS

As a rule, the topological descriptors of molecular structure are defined proceeding from the basic topology of the molecular skeleton, as encoded in the simple molecular graph, devoid of multiple edges and loops. When the number of substructures is counted, however, multiple bonds must be accounted for, since they give a significant contribution to this count. An illustration is presented below by the comparison of the subgraph counts of *n*-butane (graph **6**) and butadiene (graph **7**) molecules. The *K*′ vector of *n*-butane



6



7

is 10 (4, 3, 2, 1). From the labeled multigraph of butadiene one finds that two more edges (2 and 4) are added, along with four more two-edge subgraphs (12, 23, 34, 45), five more three-edge subgraphs (123, 134, 234, 235, 345), four four-edge subgraphs (1234, 1235, 1345, 2345), and one five-edge subgraph (the entire multigraph). The resulting *K*′ vector of butadiene, 26 (4, 5, 6, 6, 4, 1), is considerably larger than that of *n*-butane. Even larger is the difference in the TC1′ and TC′ vectors of these molecules: TC1′(butadiene) = 112 (0, 10, 24, 36, 32, 10) and TC′(butadiene) = 184 (10, 26, 42, 56, 40, 10).

Being defined as overall connectivities, our topological complexities TC and TC1 can easily be modified to account for the presence of heteroatoms in molecules by adopting Kier and Hall's valence connectivity concept,[59] which proved its value in numerous QSAR/QSPR studies. Replacing the vertex degrees a_i in our basic formula (eq 7) by the valence connectivity δ^v, we obtain the formula for the valence overall

Table 2. Substructure Count and Overall Connectivity of Isomeric Octanes

no. molecule	index *X*	${}^{0}X$	${}^{1}X$	${}^{2}X$	${}^{3}X$	${}^{4}X$	${}^{5}X$	${}^{6}X$	${}^{7}X$	X_{tot}
1 nC8	*K*	8	7	6	5	4	3	2	1	36
	TC1	0	14	24	30	32	30	24	14	168
	TC	14	26	34	38	38	34	26	14	224
2 2MC7	*K*	8	7	7	6	5	4	3	1	41
	TC1	0	14	28	36	40	40	36	14	208
	TC	14	28	41	47	49	47	39	14	279
3 3MC7	*K*	8	7	7	7	6	5	3	1	44
	TC	0	14	28	42	48	50	36	14	232
	TC1	14	28	43	56	60	58	39	14	312
4 4MC7	*K*	8	7	7	7	7	5	3	1	45
	TC1	0	14	28	42	56	50	36	14	240
	TC	14	28	43	58	69	58	39	14	323
5 25MMC6	*K*	8	7	8	7	6	6	4	1	47
	TC1	0	14	32	42	48	60	48	14	258
	TC	14	30	48	56	62	72	52	14	348
6 3EC6	*K*	8	7	7	8	8	6	3	1	48
	TC1	0	14	28	48	64	60	36	14	264
	TC	14	28	45	67	80	69	39	14	356
7 24MMC6	*K*	8	7	8	8	8	7	4	1	51
	TC1	0	14	32	48	64	70	48	14	290
	TC	14	30	50	67	83	83	52	14	393
8 22MMC6	*K*	8	7	9	9	8	7	4	1	53
	TC1	0	14	36	54	64	70	48	14	300
	TC	14	32	58	75	83	83	52	14	411
9 23MMC6	*K*	8	7	8	9	9	7	4	1	53
	TC1	0	14	32	54	72	70	48	14	304
	TC	14	30	52	77	92	83	52	14	414
10 34MMC6	*K*	8	7	8	10	10	8	4	1	56
	TC1	0	14	32	60	80	80	48	14	328
	TC	14	30	54	86	104	94	52	14	448
11 23MEC5	*K*	8	7	8	10	11	8	4	1	57
	TC1	0	14	32	60	88	80	48	14	336
	TC	14	30	54	88	113	94	52	14	459
12 33MMC6	*K*	8	7	9	11	11	8	4	1	59
	TC1	0	14	36	66	88	80	48	14	346
	TC	14	32	62	96	113	94	52	14	477
13 224MMMC5	*K*	8	7	10	10	11	10	5	1	62
	TC1	0	14	40	60	88	100	60	14	376
	TC	14	34	65	88	119	120	65	14	519
14 234MMMC5	*K*	8	7	9	11	12	10	5	1	63
	TC1	0	14	36	66	96	100	60	14	386
	TC	14	32	61	98	128	120	65	14	532
15 33MEC5	*K*	8	7	9	13	13	9	4	1	64
	TC1	0	14	36	78	104	90	48	14	384
	TC	14	32	66	115	134	105	52	14	532
16 223MMMC5	*K*	8	7	10	13	14	11	5	1	69
	TC1	0	14	40	78	112	110	60	14	428
	TC	14	34	71	118	150	131	65	14	597
17 233MMMC5	*K*	8	7	10	14	15	11	5	1	71
	TC1	0	14	40	84	120	110	60	14	442
	TC	14	34	73	128	159	131	65	14	618
18 2233MMMMC4	*K*	8	7	12	17	20	15	6	1	86
	TC1	0	14	48	102	160	150	72	14	560
	TC	14	38	90	164	220	180	78	14	798

connectivity $TC^{v}(G)$ as a measure of complexity in heteroatomic compounds:

$$TC^{v}(G) = \sum_{e=0}^{E} {}^{e}TC^{v}(G) = \sum_{e=0}^{E} \sum_{t=1}^{{}^{e}K} \sum_{i=1}^{n_t} \delta^{v}{}_{i}(G)$$

As known from the seminal paper of Kier and Hall,[59] for the main group elements δ^{v} is equal to the number of valence electrons diminished by the number of adjacent H atoms. Thus, for the nitrogen atom in $-NH_2$, $\delta^{v} = 3$, in $-NH-$ and $=NH$ it is 4 and in $=N-$ it is 5. For the oxygen atom in $-OH$, $\delta^{v} = 5$, whereas in $=O$ and in $-O-$ it is 6, etc. The difference between the valence connectivity of an atom and the vertex degree representing the atom in a hydrogen-depleted graph is in the accounting for the lone-pair electrons in the atoms (provided the π-electronic compounds are presented by a system of localized single and double bonds as in the valence bond theory). This turns the zero values of the ${}^{0}TC1$ terms, reflecting the zero complexity of isolated atoms, into nonzero ${}^{0}TC1^{v}$ terms; e.g., ${}^{0}TC1^{v}(F) = 6$, ${}^{0}TC1^{v}(O) = 4$, and ${}^{0}TC1^{v}(N) = 2$.

As an example, consider the acetic acid molecule, CH_3-COOH (Figure 2). The vertex degrees of the four non-hydrogen atoms are 1, 4, 2, and 1, respectively. The valence connectivities of the two carbon atoms coincide with the vertex degrees; however, the two oxygen atoms have valence connectivities 6 and 5, respectively. Thus, ${}^{0}TC1^{v}$(acetic acid) $= 0 + 0 + 4 + 4 = 8$, whereas ${}^{0}TC^{v}$(acetic acid) $= 1 + 4$

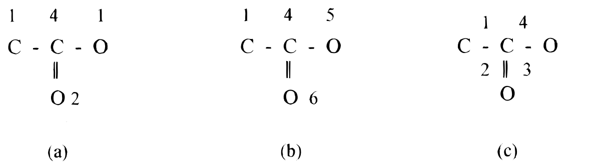


Figure 2. Molecule of acetic acid and its hydrogen-depleted graph: (a) vertex degrees a_i; (b) valence connectivities δ_i^v; (c) edge labels.

Table 3. Complexity Index *K*, Overall Connectivities TC1 and TC, and Valence Overall Connectivities TC1^v and TC^v of the Acetic Acid Molecule

indicex *X*	$^eX = 0$	$^eX = 1$	$^eX = 2$	$^eX = 3$	$^eX = 4$	X_{total}
K	4	4	6	4	1	19
TC1	0	8	24	24	8	64
TC	8	22	40	30	8	108
TC1^v	8	20	55	52	16	151
TC^v	16	34	72	58	16	196

+ 5 + 6 = 16. When a heteroatom forms a double bond, such as the oxygen in the acetic acid, its valence connectivity in calculating TC1^v is diminished by one in each subgraph containing only one of these two bonds (the TC1 scheme of calculation counts the valencies in the subgraph but not in the entire graph). The complete calculation for the acetic acid molecule is shown in Table 3. In calculating the four overall connectivity indices, all *K* subgraphs of the multigraph are taken into account: edges, 1, 2, 3, 4; two-edge subgraphs, 14, 23, 12, 13, 24, 34; three-edge subgraphs, 123, 124, 134, 234; four-edge subgraph (the entire graph), 1234.

APPLICATION OF THE TOPOLOGICAL COMPLEXITY/ OVERALL CONNECTIVITY TO QUANTITATIVE STRUCTURE–PROPERTY RELATIONSHIPS

The potential of the three overall indices of topological complexity, and that of their respective *e*th-order partial complexities, for applications in QSPR/QSAR can be verified by a systematic comparison with their predecessors, the Kier and Hall molecular connectivity. A study in progress deals with a QSAR application. Here, we restrict our goal to structure–property analysis.

Ten physicochemical properties of the C3–C8 alkanes were used in this comparative study. These are the boiling

Table 4. Statistics of the Best Overall Connectivity Models vs the Best Molecular Connectivity Models for Nine Properies of C3–C8 Alkanes

property	*r*	*s*	*F*	property	*r*	*s*	*F*
			Three-Variable Models				
ΔH_v	0.993/0.989	0.76/0.96	816/509	V_m	0.9997/0.9992	0.39/0.62	14300/5548
ΔH_f(g)	0.999/0.997	1.29/2.44	6551/1832	T_c	0.991/0.993	6.9/6.3	654/795
T_B	0.996/0.994	3.8/4.6	1457/993	P_c	0.973/0.957	0.90/1.13	199/123
R_m	0.9999/0.9999	0.048/0.062	84800/51700	V_c	0.993/0.989	0.0088/0.0113	811/487
ΔH_{at}	1.000/0.999	0.37/13.5	13000000/9915	ST	0.9879/0.9876	0.29/0.30	393/383
			Four-Variable Models				
ΔH_v	0.995/0.990	0.66/0.93	816/416	V_m	0.9997/0.9995	0.35/0.48	13200/7000
ΔH_f(g)	0.999/0.998	1.23/2.18	5385/1716	T_c	0.996/0.994	4.8/5.7	1034/724
T_B	0.999/0.996	2.4/4.0	2804/976	P_c	0.989/0.972	0.59/0.93	360/139
R_m	0.9999/0.9999	0.046/0.057	69200/44600	V_c	0.994/0.993	0.0084/0.0090	666/587
ΔH_{at}	1.0000/0.9997	0.33/9.18	12000000/16100	ST	0.994/0.989	0.21/0.29	598/307
			Five-Variable Models				
ΔH_v	0.995/0.991	0.66/0.91	657/346	V_m	0.9998/0.9995	0.32/0.47	12100/5723
ΔH_f(g)	0.999/0.998	1.23/2.18	4312/1375	T_c	0.9963/0.9956	4.7/5.1	858/720
T_B	0.999/0.997	1.7/3.3	4120/1142	P_c	0.990/0.981	0.56/0.78	317/161
R_m	1.0000/0.9999	0.046/0.049	56600/48600	V_c	0.9944/0.9936	0.0082/0.0087	570/497
ΔH_{at}	1.0000/0.9998	0.31/8.60	11000000/14600	ST	0.996/0.989	0.17/0.29	709/238

Table 5. Experimental[60] vs Calculated by Eq 18 Boiling Points of the C3–C8 Alkanes

no.[a]	compd	exptl	calcd	diff	no.[a]	compd	exptl	calcd	diff
1	*n*-C3	−42.07	−42.50	0.43	20	223MMMC4	80.88	81.00	−0.12
2	*n*-C4	−0.50	0.27	−0.77	21	*n*-C8	125.66	124.97	0.69
3	2MC3	−11.73	−12.30	0.57	22	2MC7	117.65	117.72	−0.07
4	*n*-C5	36.07	36.69	−0.62	23	3MC7	118.93	118.77	0.16
5	2MC4	27.85	28.3	−0.48	24	4MC7	117.71	117.27	0.44
6	22MMC3	9.50	9.22	0.28	25	3EC6	118.53	118.4	0.11
7	*n*-C6	68.70	68.80	−0.10	26	22MMC6	106.84	106.55	0.29
8	2MC5	60.27	60.83	−0.56	27	23MMC6	115.61	115.84	−0.23
9	3MC5	63.28	63.33	−0.05	28	24MMC6	109.43	111.2	−1.86
10	22MMC4	49.74	49.93	−0.19	29	25MMC6	109.10	109.29	−1.19
11	23MMC4	57.99	57.70	0.29	30	33MMC6	111.97	111.22	0.75
12	*n*-C7	98.43	98.00	0.43	31	34MMC6	117.73	117.93	−0.20
13	2MC6	90.05	90.31	−0.26	32	2M3EC5	115.65	116.31	−0.66
14	3MC6	91.85	91.74	0.11	33	3M3EC5	118.26	117.92	0.34
15	3EC5	93.50	93.18	0.32	34	223MMMC5	109.84	110.05	−0.21
16	22MMC5	79.20	78.99	0.21	35	224MMMC5	99.24	98.84	0.40
17	23MMC5	89.78	89.45	0.33	36	233MMMC5	114.76	113.90	0.86
18	24MMC5	80.50	79.88	0.62	37	234MMMC5	113.47	112.85	0.62
19	33MMC5	86.06	85.60	0.46	38	2233MMMMC4	106.47	107.63	−1.16

[a] This numbering corrects some numbering misprints in the corresponding table in refs 44 and 45, where the boiling points were calculated by eq 17.

point,[60] T_B (°C); critical temperature,[60] T_c (°C); critical pressure,[60] P_c (atm);[60] critical volume,[61] V_c (L/mol); molar volume, V_m (cm^3/mol);[11] molecular refraction, R_m (cm^3/mol);[11] surface tension,[60] ST (dyn/cm); heat of formation in the gaseous state, $\Delta H_f(g)$ (kJ/mol);[62] heat of vaporization, ΔH_v (kJ/mol);[62] and heat of atomization, ΔH_a (kcal/mol).[11] The values of the molecular connectivity indices are taken from Kier and Hall's monograph.[11] For each property the best models obtained by overall connectivity indices were compared to the best model produced by molecular connectivity indices. The comparison involved consecutively models with one, two, three, four, and five parameters. The best model statistics are shown in part in Table 4. Topological complexity indices outperform molecular connectivity indices in 44 out of 50 compared cases, including all models with four and five parameters. The better performance can be traced in all three statistical estimates given; particularly impressive is the improvement obtained in reducing the standard deviation of the models. Thus, the decrease in the standard deviation is less than 10% only for R_m, T_c, and V_c; for ΔH_v, P_c, V_m, and ST, the descrease is within the 20−40% range, whereas for T_B, ΔH_f(g), and ΔH_a, it is over 40%. (Separate models were derived for ΔH_f(g) and ΔH_a because the linear dependence between the two quantities stays within an isomeric series only. This frequently forgotten fact may well be illustrated by the statistics of their intercorrelation: for C8 isomers $r = 0.9998$, $s = 0.025$, and $F = 42100$; for C7 $r = 0.9992$, $s = 0.022$, and $F = 4240$; for C7 + C8, $r = 0.8634$, $s = 69.2$, and $F = 73$.)

It was not the purpose of this study to search for the best models ever for the examined properties. However, it is worth mentioning that, when conveniently combined with other molecular descriptors, the topological complexities can provide very high quality models. As an illustration, we show below our best five-parameter topological complexity model of the boiling points of the examined 38 alkanes (eq 11), and the additional improvement obtained after using the new indices jointly with those included in the OASIS software package[63] (eq 12).

$$T_B = 27.17\ {}^0K - 5.417({}^1\mathrm{TC}) - 1.044(\mathrm{TC1}) + 0.7441(\mathrm{TC}) + 0.4078({}^3\mathrm{TC1}) - 116.16 \quad (11)$$

$$r = 0.9992, \quad s = 1.75, \quad F = 4120$$

$$T_B = 44.17 I_{WG} + 60.69\,{}^0\chi - 14.73({}^1\mathrm{TC}) + 27.29(\mathrm{D2}) + 2.606({}^2\mathrm{TC}) - 422.67 \quad (12)$$

$$r = 0.9999, \quad s = 0.64, \quad F = 30700$$

The improved model, with a standard deviation almost 3 times lower than that of eq 11, incorporates the first- and second-order topological complexities ^{1}TC and ^{2}TC, along with the zero-order molecular connectivity $^0\chi$, the information index for the distribution of the metric (3D) distances I_{WG}, and the Balaban D2 index[64] (the mean square graph distance). The calculated boiling points are compared to the experimental ones in Table 5.

One may conclude that the concept of topological complexity, as the overall connectivity of the molecular graph, quantifies the structural complexity of molecules by a measure that is of very low degeneracy and mirrors the basic complexity patterns in molecular architecture. The verification of the concept by the models obtained for the physicochemical properties of alkanes indicates the high potential for application in QSPR/QSAR studies. The main idea of the new concept, namely, to characterize the molecular graph in full detail by accounting for all subgraphs, can serve as a basis for an entire class of new topological indices. Work is in progress[65] to investigate the properties of the similarly defined *overall distance* of a graph, which is a generalization of the Wiener number.

ACKNOWLEDGMENT

I am very much indebted to C. Rücker (Freiburg) for the careful examination of all my work in topological complexity, to S. Bertz (Mendham) for useful discussions, and to M. Randić (Ames) for the discussion on the correlation between the heats of formation and atomization.

REFERENCES AND NOTES

(1) Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA, 1969.

(2) *Chemical Applications of Graph Theory*; Balaban A. T., Ed.; Academic Press: London, 1976.

(3) Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.

(4) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstract Service. *J. Chem. Doc.* **1965**, *5*, 107−113.

(5) Razinger, M. Extended Connectivity in Chemical Graphs. *Theor. Chim. Acta* **1982**, *61*, 581−586. Razinger, M. Discrimination and Ordering of Chemical Structures by the Number of Walks. *Theor. Chim. Acta* **1986**, *70*, 365−378. Razinger, M.; Chretien, J. R.; Dubois, J.-E. Structural Selectivity of Topological Indexes in Alkane Series. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 23−27.

(6) Rücker, G.; Rücker, C. Counts of All Walks as Atomic and Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 683−695.

(7) Toropov, A. A.; Toropova, A. P.; Ismailov, T. T.; Voropaeva, N. L.; Ruban, I. N. Extended Molecular Connectivity: Prediction of the Boiling Points of Alkanes. *J. Strukt. Khim.* **1997**, *38*, 1154−1159.

(8) Gutman, I.; Ruščić, B.; Trinajstić, N.; Wilcox, C. W., Jr. Graph Theory and Molecular Orbitals. 12. Acyclic Polyenes. *J. Chem. Phys.* **1975**, *62*, 3399−3405.

(9) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

(10) Kier, L. B.; Hall, L. H.; Murray, W. J.; Randić, M. Molecular Connectivity I: Relationship to Nonspecific Local Anesthesia. *J. Pharm. Sci.* **1975,** *64*, 1971−1974.

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

(12) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis;* Research Studies Press: Chichester, U.K., 1986.

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

(14) Kier, L. B.; Hall, L. H.; Frazer, J. W. An Index of Electrotopological State for Atoms in Molecules. *J. Math. Chem.* **1991**, *7*, 229−241.

(15) Toropov, A. A.; Toropova, A. P.; Ismailov, T. T.; Bonchev, D. 3D Weighting of Molecular Descriptors for QSPR/QSAR by the Method of Ideal Symmetry(MIS). 1. Application to Boiling Points of Alkanes. *THEOCHEM* **1998**, *424*, 237−247.

(16) Estrada, E. Graph Theoretical Invariant of Randic Randić Revisited. *J. Chem. Inf. Comput.* Sci. **1995**, *35*, 1022−1025. Estrada, E.; Guevara, N.; Gutman, I. Extension of Edge Connectivity Index. Relationship to Line Graph Indices and QSPR Applications. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 428−431.

(17) Randić, M. Augmented Valence—Novel Graph Theoretical Index with an Apparent Interpretation. *Int. J. Quantum Chem.*, in press.

(18) Bonchev, D.; Seitz, W. A. The Concept of Complexity in Chemistry. In *Concepts in Chemistry: Contemporary Challeng*; Rouvray, D. H., Ed.; Research Studies Press: Taunton, U.K., 1996; pp 353−381.

(19) Kier, L. B.; Testa, B. Complexity and Emergence in Drug Research. *Adv. Drug Res.* **1995**, *26*, 1−43.

(20) Bertz, S. H. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599−3601.

(21) Bertz, S. H. A Mathematical Model of Molecular Complexity. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier: New York, 1983; pp 206−221.

(22) Rashevsky, N. Life, Information Theory, and Topology. *Bull. Math. Biophys.* **1955**, *17*, 229−235.

(23) Mowshovitz, A. Entropy and the Complexity of Graphs. 1. An Index of the Relative Complexity of a Graph. *Bull. Math. Biophys.* **1968**, *30*, 175−204.

(24) Shannon, C., and Weaver, W. *Mathematical Theory of Communications*; University of Illinois Press: Urbana, IL, 1949.

(25) Bonchev, D. *Information-Theoretic Indices for Characterization of Chemical Structures*; Research Studies Press: Chichester, U.K., 1983; 249 pp.

(26) Bonchev D. Kolmogorov's Information, Shannon's Entropy, and Topological Complexity of Molecules. *Bulg. Chem. Commun.* **1995**, *28*, 567−582.

(27) Bonchev, D.; Polansky, O. E. On the Topological Complexity of Chemical Systems. In *Graph Theory and Topology in Chemistry*; King, R. B., Rouvray, D. H., Eds.; Elsevier: Amsterdam, 1987; pp 126−158.

(28) Bertz, S. H.; Wright, W. F. The Graph Theory Approach to Synthetic Analysis: Definition and Application of Molecular Complexity and Synthetic Complexity. *Graph Theory Notes N. Y. Acad. Sci.* **1998**, 32−48.

(29) Minoli, D. Combinatorial Graph Complexity. *Atti. Acad. Naz. Lincei Rend.* **1976**, *59*, 651−661.

(30) Bonchev, D. The Problems of Computing Molecular Complexity. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Publications: New York, 1990; pp 34−67.

(31) Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix, and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517−4533.

(32) Ruch, E.; Gutman, I. The Branching Extent of Graphs. *J. Comb. Inf. Syst. Sci.* **1979**, *4*, 285−295.

(33) Bonchev, D.; Mekenyan, O.; Trinajstić, N. Topological Characterization of Cyclic Structures. *Int. J. Quantum Chem.* **1980**, *17*, 845−893.

(34) Bertz, S. H. Branching in Graphs & Molecules. *Discr. Appl. Math.* **1988**, *19*, 65−83.

(35) Bonchev, D. Topological order in molecules. 1. Molecular Branching Revisited. *THEOCHEM* **1995**, *336*, 137−156.

(36) Randić, M. On Molecular Branching. *Acta Chim. Slov.* **1997**, *44*, 57−77.

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

(38) Nikolić, S.; Trinajstić, N.; Jurić, A.; Mihalić, Z.; Krilov, G. Complexity of Some Interesting (Chemical) Graphs. *Croat. Chem. Acta* **1996**, *69*, 883−897.

(39) Gutman, I.; Mallion, R. B.; Essam, J. W. Counting the Spanning Trees of a Labeled Molecular Graph. *Mol. Phys.* **1983**, *50*, 859−877.

(40) Bonchev, D.; Kamensky, D.; Temkin, O. N. Complexity Index for the Linear Mechanisms of Chemical Reactions. *J. Math. Chem.* **1987**, *1*, 345−388.

(41) Gordeeva, K.; Bonchev, D.; Kamenski, D.; Temkin, O. N. Enumeration, Coding, and Complexity of Linear Reaction Mechanisms, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 244−247.

(42) Temkin, O. N.; Zeigarnik, A. V.; Bonchev, D. *Chemical Reaction Networks. A Graph Theoretical Approach;* CRC Press: Boca Raton, FL, 1996; 286 pp.

(43) Kolmogorov, A. N. Three Approaches to the Quantitative Definition of Information. *Probl. Inf. Transm. (Engl. Transl.)* **1965**, *1*, 1−7.

(44) Bonchev, D. Novel Indices for the Topological Complexity of Molecules. *SAR QSAR Environ. Res.* **1997**, *7*, 23−43.

(45) Bonchev, D. Overall Connectivity and Molecular Complexity. In *Topological Indices and Related Descriptors;* Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Reading, U.K., 1999; pp 361−401.

(46) Dubois, J.-E. Ordered Chromatic Graphs and Limited Environment Concept. In *Chemical Applications of Graph Theory*; Balaban A. T., Ed.; Academic Press: London, 1976.

(47) Willet, P. Some Heuristics for Nearest-Neighbor Searching in Chemistry-Structure Files. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 3, 22−25.

(48) Klopman, G. Artificial Intelligence Approach to Structure−Activity Studies. Computer Automated Structure Evaluation of Biological Activities of Organic Molecules. *J. Am. Chem. Soc*. **1984**, *106*, 7315−7321.

(49) Gordon, M.; Kennedy, J. W. The Graph-Like State of Matter. Part 2. TCGI Schemes for the Thermodynamics of Alkanes and the Theory of Inductive Inference. *J. Chem. Soc., Faraday Trans. 2* **1973**, *69*, 484−504.

(50) Panaye, A.; Doucet, J.-P.; Fan, B. T. Topological Approach of C-13 NMR Spectral Simulation. Application to Fuzzy Substructures. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 3, 258−265.

(51) Bertz, S.; Herndon, W. C. Similarity of Graphs and Molecules. In *Artificial Intelligence Applications in Chemistry;* American Chenical Society: Washington, DC, 1986; pp 169−175.

(52) Matula, D. W. On the Number of Subtrees in a Symmetric n-ary Trees. *SIAM J. Appl. Math*. **1970**, *18*, 688−703.

(53) Bone, R. G. A.; Villar, H. O. Exhaustive Enumeration of Molecular Substructures. *J. Comput. Chem*. **1997**, *18*, 86−107.

(54) Bertz, S. H.; Sommer, T. J. Rigorous Mathematical Approaches to Strategic Bonds and Synthetic Analysis Based on Conceptually simple new complexity indices. *Chem. Commun.* **1997**, 2409−2410.

(55) Bonchev, D.; Gordeeva, E. Partial Orderings Induced by Topological Complexity. *MATCH*, in press.

(56) Bonchev, D. The Concept of Topological Complexity as Overall Connectivity. In *Mathematical Chemistry Series. Volume 7, Complexity in Chemistry*; Bonchev, D., Rouvray, D. H., Eds.; Gordon & Breach: Reading, U.K., in preparation.

(57) Bonchev, D.; Mekenyan, O.; Trinajstić, N. Isomer Discrimination by Topological Information Approach. *J. Comput. Chem*. **1981**, *2*, 127−148.

(58) Rücker, G.; Rücker, C. Walk Counts, Labirinthicity, and Complexity of Acyclic and Cyclic Graphs and Molecules. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 99−106.

(59) Kier, L. B.; Hall, L. H. Derivation and Significance of Valence Molecular Connectivity. *J. Pharm. Sci.* **1981**, *70*, 583−589.

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

(61) *Selected Values of Properties of Hydrocarbons and Related Compounds*; Research Project 44; American Petroleum Institute: Pittsburgh, PA, 1977.

(62) Garbalena, M.; Herndon, W. C. Optimum Graph-Theoretical Models for Enthalpic Properties of Alkanes. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 37−42.

(63) Mekenyan, O.; Karabunarliev, S.; Bonchev, D. The OASIS Concept for Predicting Biological Activity of Chemical Compounds. *J. Math. Chem.* **1990**, *4*, 207−215.

(64) Bonchev, D.; Balaban, A. T.; Mekenyan, O. Generalization of the Graph Center Concept, and Derived Topological Indices. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 106−113.

(65) Bonchev, D. Overall Distance of a Graph. A Concept Generalizing the Wiener Number. To be submitted for publication.

CI990120U