Substructure, Subgraph, and Walk Counts as Measures of the Complexity of Graphs and Molecules

Gerta Rücker and Christoph Rücker*

Department of Mathematics, Universität Bayreuth, D-95440 Bayreuth, Germany

Received June 4, 2001

In discussions of unsaturated compounds represented by multigraphs it is necessary to distinguish between the notions of substructure and subgraph. Here the difference is explained and exemplified, and a computer program is introduced which for the first time is able to construct and count all substructures and subgraphs for a colored multigraph (a molecular compound which may contain unsaturation and heteroatoms). Construction of *all* substructures and subgraphs is computationally demanding; therefore, two alternatives are pointed out for the treatment of large sets of compounds: (i) Often it will suffice to consider counts of substructures/subgraphs up to a certain number of edges only, information which is provided by the program much more rapidly. (ii) It is shown that information equivalent to that gained from substructure or subgraph counts is often far more easily available using walk counts. Some problems and their consequences for substructure/subgraph/walk counts are discussed that arise from the models used in organic chemistry for certain compounds such as aromatics and from the necessity to express qualitative features of molecular structures numerically.

INTRODUCTION

In 1986 Bertz and Herndon proposed to count the connected subgraphs of a (saturated) molecular graph *inter alia* as a measure of its complexity,[1] and Bertz worked out this idea and generalized it to molecular graphs containing multiple bonds in a series of papers recently.[2−5] Independently Bonchev since 1997 used the number of all connected subgraphs and several graph invariants derived therefrom for structure−property modeling of saturated hydrocarbons.[6−11] In 1997 also Bone and Villar in the context of molecular diversity considerations pointed to the importance of the number of all (induced) substructures of a saturated molecular structure.[12]

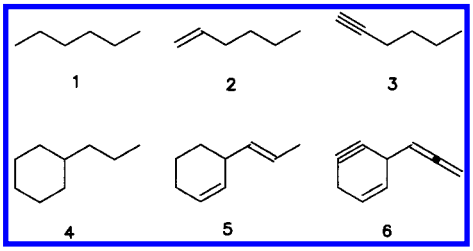
Since these authors did not present sufficiently mighty computer programs for constructing and counting connected subgraphs and substructures, we developed programs to find the connected subgraphs of simple graphs (graphs without multiple edges or loops)[13] and the substructures of chemical structures (which may contain multiple bonds and heteroatoms).[14] In a different context, in 1993 we found the count of walks of different lengths in a simple tree graph to be a useful measure of its complexity,[15] and later we demonstrated the ability of walk counts to numerically express the complexity of general graphs and molecular structures.[16] In the present work we compare the notions and performance of substructure, subgraph, and walk counts for measuring the complexity of molecular graphs and compounds.

RESULTS AND DISCUSSION

Substructures and Subgraphs. A subgraph is obtained from a graph by removing at least one line, and a substructure is obtained from a chemical structure by removing at least one bond.[17] (A *connected* subgraph/substructure is a subgraph/substructure which consists of one piece only, i.e., every pair of vertices/atoms is connected by a sequence of lines/bonds.) For this parallelism several authors including ourselves used the expressions subgraph and substructure interchangeably, one term coming from graph theory, the other from chemistry. As long as only saturated chemical compounds are considered this is quite correct, and in fact most molecular complexity discussions were concerned with saturated structures. For leading references to molecular complexity see a recent book[18] and articles,[14,16,19] numerical values and ordering of structures by several complexity indices were compared recently.[16,19−22]

However, as soon as unsaturated compounds (represented by multigraphs) are considered, it becomes useful or even necessary to distinguish between subgraphs and substructures. In a substructure, all bonds still present have their multiplicities as they are in the parent structure, while a subgraph may correspond to a less unsaturated analogue of a (sub)structure, i.e., may have a single or double line where the parent multigraph has a double or triple edge.

To see the difference and its consequences let us consider as examples *n*-hexane, 1-hexene, and 1-hexyne, compounds **1**−**3**, see also Table 1.



In *n*-hexane the following connected substructures are found: 6 methane substructures (single C atoms), 5 ethane

* Corresponding author phone: +49-921-553386; fax: +49-921-553385; e-mail: Christoph.Ruecker@uni-bayreuth.de.

10.1021/ci0100548 CCC: $20.00
Published on Web 11/06/2001

Table 1. Results of BERTZ Runs for Some Graphs/Structures

	substructures		subgraphs		
structure	N_t	N_s	N_{tB}	N_{sB}	time [s][a]
1	21	6	21	6	0.00
2	21	10	31	11	0.00
3	21	10	51	16	0.00
4	109	25	109	25	0.05
5	109	55	401	87	0.05
6	109	67	2737	267	0.16
7	26	15	90	29	0.00
8	529	260	4243	771	0.39
7+8	555	268	4333	784	0.39
9	30767	12742	205401	35396	51.74
10	10843	3837	37119	7423	9.61
11	15	9	15	9	0.00
12	17	9	17	9	0.00
13	17	10	17	10	0.00
14	20	9	20	9	0.00
15	29	13	29	13	0.00
11−15	98	16	98	16	0.00
16	15	10	27	14	0.00
17	17	12	29	14	0.00
18	17	12	27	15	0.00
19	29	20	53	23	0.00
16−19, 14	98	24	156	28	0.05
20	2441	64	2441	64	0.60
21	2007	1434	18695	4802	5.60
22	6460	5620	60106	19002	68.05
23	44548	16797	7482482	281860	1203.7
24	887784	439266	1668264	613522	1811.8
25					>20000
26	1408347	1087080	8410625	2437765	12866.63
27	17232	4152	37716	5574	12.74
28	16473	4091	32437	5438	11.98
29	12840	3707	23586	4695	10.11
27[b]	580	36	822	36	0.11
28[b]	537	29	679	29	0.11
29[b]	421	28	519	28	0.11
30	100	34	900	114	0.06
31	100	38	892	115	0.05
33	135	67	1291	217	0.11
34	135	65	1291	219	0.11
36	60	33	480	103	0.06
38	37	18	37	18	0.00
39	21	11	21	11	0.00
40	21	13	21	13	0.00
41	21	12	21	12	0.00
42	37	13	37	13	0.00
43	37	21	37	21	0.00
44	37	13	37	13	0.00

[a] The smallest run time displayed by the PC is 0.05 s. The entry 0.00 here means less than 0.05 s. [b] In these runs, the maximal number of edges in substructures to be found was 5, see text.

substructures C−C, 4 propane substructures C−C−C, 3 butane substructures C−C−C−C, 2 pentane substructures C−C−C−C−C, and the hexane structure itself. Thus the total number of connected substructures N_t is 21; the number of distinct (kinds of) connected substructures N_s is 6. The total number of connected subgraphs N_{tB} and the number of distinct connected subgraphs N_{sB} are also 21 and 6, respectively, since *n*-hexane is a saturated compound.

Since we are interested in connected subgraphs/substructures only, in the following we always omit the qualifier *connected*.

In 1-hexene there are 6 single C atoms (methane substructures), 4 ethane substructures C−C, 1 ethene substructure C═C, 3 propane substructures C−C−C, 1 propene substructure C═C−C, 2 butane substructures C−C−C−C, 1 butene substructure C═C−C−C, 1 pentane substructure C−C−C−C−C, 1 pentene substructure C═C−C−C−C, and the hexene structure itself. Thus the total number of substructures N_t is 21, and the number of distinct substructures N_s is 10. N_s(1-hexene) > N_s(*n*-hexane) reflects the higher complexity of 1-hexene compared to *n*-hexane. Now each of the substructures is also a subgraph, but according to the formal subgraph construction procedure, the removal of lines, there exist additional subgraphs for 1-hexene.

According to Bertz[2,3,5] each of the ethene, propene, butene, pentene, and hexene substructures contributes in addition to itself another two isomorphic subgraphs (remove one or the other line from a double edge), which are ethane, propane, butane, pentane, and hexane graphs, increasing the total number of subgraphs N_{tB} for 1-hexene to 31 and the number of distinct subgraphs N_{sB} to 11.[23] Thus the higher complexity of 1-hexene compared to *n*-hexane is also mirrored by the total and distinct subgraph counts.

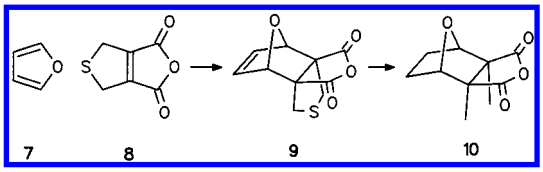
For 1-hexyne the total number of substructures and the number of distinct substructures are exactly as for 1-hexene, $N_t = 21$ and $N_s = 10$. To see this, replace any C═C expression above by C≡C and any -ene suffix by -yne. That is, as measured by the numbers of substructures, 1-hexyne, though more complex than *n*-hexane, is not more complex than 1-hexene. In the logic of substructures, a triple bond is just different from a double bond, and both are different from a single bond, a further numerical differentiation is not made. However, there are additional possibilities to remove lines from a triple edge. From each substructure containing C≡C the first or second or third of the triple bond's three lines can be removed or even two of them (first and second, first and third, second and third). Thus each -yne substructure in addition to the corresponding -yne subgraph gives rise to three isomorphic -ene subgraphs and three isomorphic -ane subgraphs, increasing the total number of subgraphs N_{tB} for 1-hexyne to 51 and the number of distinct subgraphs N_{sB} to 16.

Defining, constructing, and counting subgraphs by this formal approach is justified by this desired result, the higher numbers obtained for compounds of a higher degree of unsaturation, and thus the opportunity to attribute a higher number to the complexity of more highly unsaturated compounds. The drastic effect of multiple unsaturation is seen from examples **4**−**6**,[24] see Table 1.

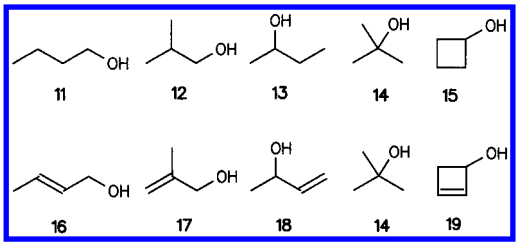
A consequence of the subgraph definition is that the set of subgraphs of an unsaturated compound contains all its less unsaturated analogues, e.g. any aldehyde or ketone has the corresponding alcohol graph among its subgraphs, while the alcohol is not normally considered to be a substructure of an aldehyde or ketone. Similarly, a nitrile has the corresponding imine and primary amine graphs in its set of subgraphs, though the imine and amine are not normally considered to be substructures of a nitrile.[25] Substructures manifest themselves in a compound's reactivity, properties, and spectra and thus play a vital role in organic synthesis and in manual or automatic structure elucidation, while subgraphs are theoretical constructs that found applications in the complexity, similarity, and diversity area.

The Program BERTZ. In the terminology described above our recently presented computer program NIMSG constructs and counts the substructures of a graph or molecular structure, N_t and N_s.[14] We now developed the program so as to construct and count the subgraphs as well,

N_{tB} and N_{sB} (program BERTZ). The total number of subgraphs N_{tB} is easily calculated once the substructures are known: While each saturated substructure contributes exactly one to the total number of subgraphs (and thus each distinct (kind of) saturated substructure contributes its occurrence number), each distinct substructure containing n2 double bonds and n3 triple bonds contributes to N_{tB} exactly $3^{n2}*7^{n3}*$occurrence number. The *maximum* contribution of this same substructure to the number of distinct subgraphs N_{sB} is $2^{n2}*3^{n3}$, but this number is seldom realized, since some of these kinds of subgraphs are usually found elsewhere in the structure.[26] Therefore from each unsaturated substructure all $2^{n2}*3^{n3}-1$ less unsaturated subgraphs have to be constructed and stored; finally all subgraphs constructed have to be checked for isomorphism, for N_{sB} isomorphic ones have to be counted only once. This is done in the program by the methods described recently, making use of well-discriminant graph invariants such as the Balaban index J and the eigenvalues of the distance matrix.[14]

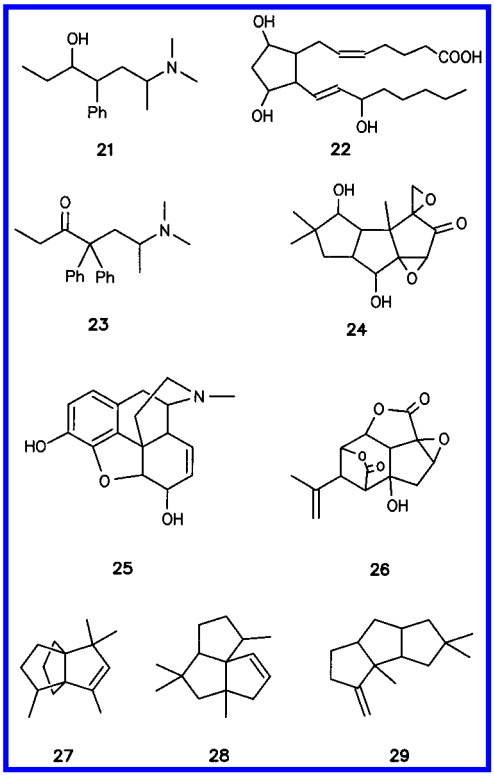


Program BERTZ is very broadly applicable; it accepts as input also nonconnected chemical graphs corresponding to ensembles of molecules, so that the complexity increase associated with e.g. addition reactions can easily be followed, as measured by the quantities N_t, N_s, N_{tB}, and N_{sB}. As an example consider the Diels−Alder reaction of furan (**7**) and the anhydride **8** to yield adduct **9**, the key intermediate in the synthesis of cantharidin (**10**). Similarly the diversity of a set of compounds is semiquantitatively expressed by N_s or N_{sB}, as demonstrated here for two example sets, five saturated butanols (**11**−**15,** example from ref 1), and a set of corresponding butenols (**16**−**19, 14**).



The substructure and subgraph counts of some typical organic compounds are also given in Table 1 (cubane (**20**), prostaglandin $F_{2\alpha}$ (**22**), methadone (**23**), coriolin (**24**), morphin (**25**), picrotoxinin (**26**), modhephene (**27**), silphinene (**28**), hirsutene (**29**)). Substructure counts for most of these were given in an earlier publication already.[14] Note that for aromatic compounds systematic differences in N_s between there and here are found, which are caused by the different models used to describe an arene for the purposes of counting substructures on one hand and subgraphs on the other, see below.

The run times given in the last columns of the tables in this work were obtained on a Windows PC (Pentium III processor, 866 MHz, 192 MB RAM, Windows ME). BERTZ runs include output of a list of all distinct substructures and

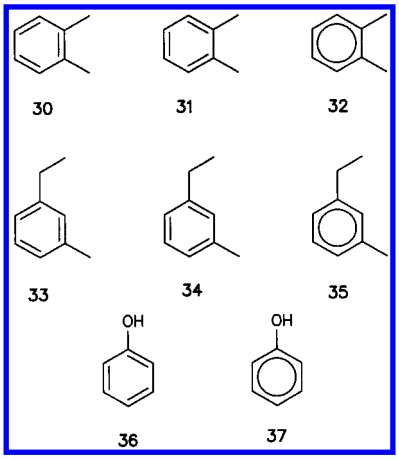


all distinct subgraphs. For molecules of even modest size BERTZ run times may be inconveniently long; they are caused by the demanding (exponentially growing with problem size) procedure to construct and check for isomorphism all substructures and subgraphs.[13,14]

To be able to treat large sets of compounds, e.g. combinatorial libraries, two alternatives may be envisaged: (i) The maximal size of substructures/subgraphs to be constructed may be limited to a certain number of edges. A version of program BERTZ is available doing this, and as a result run times are drastically reduced, see Table 1, compounds **27**−**29**. (ii) Walk counts may be used instead of substructure/subgraph counts, see below.

Substructures and Subgraphs in Aromatic Compounds. To be able to perform its central job, the removal of lines from multiple edges, the program BERTZ needs as input a decent (localized) multigraph. This means that aromatic rings have to be entered as and are treated as localized cyclopolyenes. Consider a simple disubstituted benzene, *o*-xylene. The N_s, N_{tB}, and N_{sB} values will differ depending on whether there is a single or double bond drawn connecting the substituted ring carbon atoms, see **30** and **31** and Table 1. While there is no such problem with *m*- and *p*-xylene, the problem reappears already with *meta*-disubstituted benzenes as soon as the substituents are different, see **33** and **34** and Table 1. Even in monosubstituted benzenes a similar problem is encountered: In "localized" phenol **36** both a substructure O−C−C and a substructure O−C═C together with the corresponding subgraphs are present, whereas in "aromatic" phenol **37** there is but one kind of substructure O−C÷C occurring twice (symbol ÷ here stands for an aromatic bond).

Thus counting subgraphs (or walks, see below) in most substituted aromatics involves some arbitrariness. In contrast, substructures can be counted without such a problem: In a substructure, as in the corresponding parent structure, a bond may be aromatic (i.e. different from single, double, or triple,



e.g. 1.5-fold, without further numerical consequences being drawn from this input). Accordingly, program NIMSG (but not BERTZ or MORGAN) accepts aromatic bonds if denoted as such explicitly, so that in this case (as in reality) there is only one 1,2-dimethylbenzene (**32**), only one 1-ethyl-3-methylbenzene (**35**), and only one kind of substructure O−C÷C in phenol (**37**).

Walks. A walk in a graph or molecular graph is an alternating sequence of vertices and the edges connecting consecutive vertices. Each vertex and edge may appear in a walk more than once. The length of a walk is the number of edges in it. Though a walk can be of any length, it makes sense to count the walks up to a certain length only, i.e., up to length $n-1$, where n is the number of the graph's vertices. The number of all walks in a graph of lengths 1 to $n-1$ is called its total walk count, twc. For details of these definitions see our earlier publications.[15,16,19,27] Index twc increases with increasing size, branching, number of rings, and edge and vertex weights (unsaturation, heteroatoms) and therefore is useful as a complexity index for (molecular) graphs.[16] Numerical values of twc are usually very large; there are many more walks in a graph than paths, substructures, or even subgraphs. Nevertheless walk counts can be obtained easily, either by multiplication of the adjacency matrix by itself or by the Morgan algorithm, an extremely easy calculation involving nothing but addition steps, as executed by the program MORGAN.[15] The MORGAN run times therefore are very attractive, as shown in Table 2. Even the most problematic compounds in Table 1 (in terms of BERTZ run times) required less than 0.05 s in MORGAN runs. It should be noted that the MORGAN run times given include a topological symmetry determination by TOPSYM[28] which is used for a walk complexity[16] calculation. Thus by switching off this latter procedure the MORGAN run times can be reduced even further.

It is suggested here that for most complexity applications walk counts can serve as well as substructure or subgraph counts or other measures, and far faster.

Thus the complexity differences between the isomers modhephene (**27**), silphinene (**28**), and hirsutene (**29**) are mirrored by twc as by the substructure or subgraph counts.

For analysis of syntheses (producing complexity-versus-step plots) Bertz used his complexity indices η (the number of pairs of adjacent edges in a graph) or C(η) or N_{tB}, N_{sB}.[2,29] Other authors were concerned about the difficulty to obtain numerical values of these indices and instead developed rather *ad hoc* methods for the same purpose.[30] These latter methods are not mathematically rigorous; they are problematic in that they use an arbitrary list of complexity features (in which constitution/topology is not included) which are weighted by factors obtained by intuition. For a discussion of such a scheme see ref 5. Obviously, the mathematically more exact substructure, subgraph, or walk counts can serve the same purposes, and at least for walk counts run times should be of no concern.

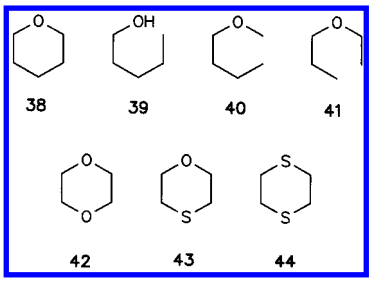
Bertz used N_{tB}, N_{sB} to identify strategic bonds, such bonds that when broken result in a particularly strong decrease in complexity.[2] For the example of a fused six-membered ring system he concluded, based on N_{sB}, that the most and second-most strategic bonds are the fusion bond and the bond adjacent to it. In fact based on twc one arrives at exactly the same conclusion (not shown). For the breaking of one bond in oxacyclohexane (**38**) N_{sB} and twc agree that the C−O bond

Table 2. Results of MORGAN Runs for Some Graphs/Structures

structure[a]	twc	time [s][b]
1	222	0.00
2	500	0.00
3	1498	0.00
4	6792	0.00
5	34550	0.00
6	179826	0.00
7	600	0.00
8	945699	0.00
9	428204678	0.00
10	80062440	0.00
11	119	0.00
12	139	0.00
13	145	0.00
14	185	0.00
15	219	0.00
11−15	807[c]	
16	267	0.00
17	289	0.00
18	257	0.00
19	397	0.00
16−19, 14	1395[c]	
20	26232	0.00
21	1307581050	0.00
22	1763994021748	0.00
23	4869080829024	0.00
24	98827648507	0.00
25	860280420590[d]	0.00
26	449507019165	0.00
27	75436484	0.00
28	47871730	0.00
29	32253602	0.00
30	30582	0.00
31	30178	0.00
33	99010	0.00
34	99010	0.00
36	8769	0.00
38	587	0.00
39	297	0.00
40	388	0.00
41	423	0.00
42	806	0.00
43	1369	0.00
44	1936	0.00

[a] Heteroatoms are arbitrarily colored as follows: O 1, N 2, S 2. [b] The smallest run time displayed by the PC is 0.05 s. The entry 0.00 here means less than 0.05 s. [c] Program MORGAN does not treat disconnected graphs. The twc values given here are the sums of the twc values of the specified compounds. [d] Double bonds in the aromatic ring formally arranged as shown in structure **25**; the alternative arrangement leads to 873008740622.

is the most strategic one, but interestingly, twc orders the three *seco*-oxacyclohexanes (*n*-pentanol (**39**), butyl methyl ether (**40**), ethyl propyl ether (**41**)) in a more natural manner (**39** < **40** < **41**) than does N_{sB} (**39** < **41** < **40**), see Tables 1 and 2. Accordingly, different bonds in **38** qualify as second-most strategic when measured by N_{sB} or by twc.



Edge Weight-Quantitative and Vertex Weight-Quantitative vs -Qualitative Variables. In our opinion the problems resulting from representing multiple bonds and heteroatoms by numbers in chemical graph theory have been discussed far less than they deserve to be. A thorough discussion obviously cannot be given here, suffice it to point out a few problems encountered in the present work.

If in a multigraph the numerical value of a graph invariant depends on the particular numeric input edge weights, we call that graph invariant edge weight-quantitative. Analogously, if in a vertex-colored graph the numerical value of a graph invariant depends on the particular numeric input vertex weights, we call that graph invariant vertex weight-quantitative. Otherwise a variable is edge or vertex weight-qualitative. The difference between quantitative and qualitative invariants is in the processing of numerical input information: In the *n*-hexane/1-hexene/1-hexyne example, substructure counts N_t and N_s treat edges of weights 1, 2, and 3 as simply different (just as they would treat edges of any other three weights). In contrast, subgraph counts N_{tB} (contributions 1, 3, and 7, respectively) and N_{sB} are a numerical function of the particular input values. Thus, the substructure numbers, N_t and N_s, are edge weight-qualitative and vertex weight-qualitative, whereas the subgraph numbers, N_{tB} and N_{sB}, are edge weight-quantitative but vertex weight-qualitative. Walk counts and therefore twc are both edge weight-quantitative and vertex weight-quantitative. The latter is true since a vertex weight may be interpreted as the number of loops at that vertex, and loops result in additional walks. Edge or vertex weight-quantitative variables provide a higher flexibility in modeling and at the same time some arbitrariness. Thus for vertex and edge-weighted graphs (unsaturated heteroatom-containing molecules) comparisons by twc depend on the particular weighting of unsaturation and heteroatoms. Conversely, comparisons by edge and vertex weight-qualitative invariants such as substructure counts N_t and N_s are rather inflexible and at the same time objective.

Further, it is by no means clear what the particular numerical input for unsaturation and heteroatoms should be. Though nothing seems more natural to an organic chemist than to represent an alkene and an alkyne by graphs containing a double and a triple edge, respectively, we should be aware that this is a somewhat arbitrary decision and may lead to unexpected numerical consequences in edge weight-quantitative variables. Thus for the total number of subgraphs a substructure containing one triple bond contributes exactly 7/3 as much as a substructure containing one double bond, while for other complexity indices this ratio is different, e.g. for η it is 3 and for Whitlock's index[30a] S it is 2.

A similar problem with edge weight-quantitative invariants is the following: Any such graph invariant will place an alkene between the corresponding alkane and alkyne, if as usual the olefinic bond is given an edge weight (say 2) between those of an alkane and an alkyne bond (say 1 and 3). However, property values of an alkene are not always between those of the corresponding alkane and alkyne. Thus the boiling points of *n*-hexane, 1-hexene, and 1-hexyne are 68.95, 63.35, and 71.3 °C, respectively.

Similarly problematic is the numerical input for heteroatoms, if vertex weight-quantitative invariants are to be used. Consider as examples the three compounds 1,4-dioxane (**42**), 1,4-oxathiane (**43**), and 1,4-dithiane (**44**). If their complexities are measured by N_s (or N_{sB}), oxathiane is found more complex than dioxane or dithiane (which are equally complex) irrespective of the particular numerical input for an oxygen and a sulfur atom. If a vertex weight-quantitative invariant is used, then the outcome depends on the particular numerical input: If the number of lone-pair electrons at an atom is used for input, then there will be no difference at all between the three compounds in the graph-theoretical model. If, on the other hand, any atom characteristic is used for input which differs for O and S (e.g. the atomic weight or radius, the atom or row number in the periodic system, or even a pair of arbitrary numbers such as 1 and 2), then there will be a considerable difference between dioxane and dithiane, and oxathiane will invariably be ordered in between. This may be reasonable for some properties to be modeled (e.g. the boiling point) but may be unreasonable for others (e.g. complexity).

The computer programs mentioned herein are written in FORTRAN, they are available from the authors.

REFERENCES AND NOTES

(1) Bertz, S. H.; Herndon, W. C. The Similarity of Graphs and Molecules. In *Artificial Intelligence Applications in Chemistry*; Pierce, T. H., Hohne, B. A., Eds.; American Chemical Society: Washington, DC, 1986; pp 169−175.

(2) 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.

(3) 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 of New York* **1998**, *35*, 32−48.

(4) Bertz, S. H.; Zamfirescu, C. M. New Complexity Indices Based on Edge Covers. *MATCH − Commun. Math. Comput. Chem.* **2000**, *42*, 39−70.

(5) Bertz, S. H. *Complexity of Molecules and Their Synthesis.* A chapter in ref 18.

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

(7) (a) Bonchev, D. Overall Connectivity and Topological Complexity: A New Tool for QSPR/QSAR. In *Topological Indices and Related Descriptors in QSAR and QSPR;* Devillers, J., Balaban, A. T., Eds.; Gordon & Breach: Reading, U.K., 1999; pp 361−401. (b) Bonchev, D. Overall Connectivities/Topological Complexities: A New Powerful Tool for QSPR/QSAR. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 934−941.

(8) Bonchev, D.; Gordeeva, E. Hierarchical Partially Ordered Sets Based on Topological Complexity. *MATCH − Commun. Math. Comput. Chem.* **2000**, *42*, 85−117.

(9) Bonchev, D.; Trinajstić, N. Overall Molecular Descriptors. 3. Overall Zagreb Indices. *SAR QSAR Environ. Res.* **2001**, *12*, 213−236.

(10) Bonchev, D. The Overall Wiener Index − A New Tool for Characterization of Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 582−592.
(11) Bonchev, D. Overall Connectivity − A Next Generation Molecular Connectivity. *J. Mol. Graphics Modell*. **2001**, *20*, 65−75.
(12) Bone, R. G. A.; Villar, H. O. Exhaustive Enumeration of Molecular Substructures. *J. Comput. Chem.* **1997**, *18*, 86−107.
(13) Rücker, G.; Rücker, C. Automatic Enumeration of All Connected Subgraphs. *MATCH − Commun. Math. Comput. Chem.* **2000**, *41*, 145−149.
(14) Rücker, G.; Rücker, C. On Finding Nonisomorphic Connected Subgraphs and Distinct Molecular Substructures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 314−320, 865.
(15) Rücker, G.; Rücker, C. Counts of All Walks as Atomic and Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 683−695.
(16) Rücker, G.; Rücker, C. Walk Counts, Labyrinthicity, and Complexity of Acyclic and Cyclic Graphs and Molecules. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 99−106.
(17) Nevertheless, usually the graph itself (the structure itself) is also included in the set of its subgraphs (substructures). We follow this practice here.
(18) *Complexity in Chemistry* (*Mathematical Chemistry Series;* Bonchev, D., Rouvray, D. H., Eds.; Taylor & Francis: Reading, U.K., Vol. 7, in press.
(19) Nikolić, S.; Trinajstić, N.; Tolić, I. M.; Rücker, G.; Rücker, C. *On Molecular Complexity Indices*. A chapter in ref 18.
(20) Nikolić, S.; Tolić, I. M.; Trinajstić, N. On the Complexity of Molecular Graphs. *MATCH − Commun. Math. Comput. Chem.* **1999**, *40*, 187−201.
(21) Nikolić, S.; Tolić, I. M.; Trinajstić, N.; Baucić, I. On the Zagreb Indices as Complexity Indices. *Croat. Chem. Acta* **2000**, *73*, 909−921.
(22) Randić, M. On the Concept of Molecular Complexity. *Croat. Chem. Acta*, in press.
(23) Bonchev, in a slightly different approach, arrives at the same total number of subgraphs N_{tB} but gathers them into groups according to their numbers of lines, not according to isomorphism.[7a,b]
(24) Structure **6** is shown here only to demonstrate the effect; it is not implied that **6** is capable of existence as a real compound.
(25) A so-called substructure search in the CAS Registry File or in Beilstein Crossfire (which could more correctly be called a superstructure search) performed on an alcohol query will *not* retrieve the corresponding carbonyl compound, nor will a nitrile be retrieved as a result of such a search on a primary amine or imine query.
(26) For example, any alkene has as one of its substructures ethene, which gives rise to two kinds of subgraphs, ethene and ethane. The latter is found in most organic compounds anyway.
(27) Gutman, I.; Rücker, C.; Rücker, G. On Walks in Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 739−745.
(28) (a) Rücker, G.; Rücker, C. Computer Perception of Constitutional (Topological) Symmetry: TOPSYM, a Fast Algorithm for Partitioning Atoms and Pairwise Relations among Atoms into Equivalence Classes. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 187−191. (b) Rücker, G.; Rücker, C. Isocodal and Isospectral Points, Edges, and Pairs in Graphs and How to Cope with Them in Computerized Symmetry Recognition. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 422−427.
(29) (a) 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. (b) Bertz, S. H.; Sommer, T. J. Application of Graph Theory to Synthesis Planning: Complexity, Reflexivity, and Vulnerability. In *Organic Synthesis: Theory and Applications*; Hudlicky, T., Ed.; JAI Press: Greenwich, CT, 1993; Vol. 2, pp 67−92.
(30) (a) Whitlock, H. W. On the Structure of Total Synthesis of Complex Natural Products. *J. Org. Chem.* **1998**, *63*, 7982−7989. (b) Barone, R.; Chanon, M. A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 269−272.

CI0100548