Physical Chemistry

Electronegativity of atoms of ring-containing molecules—NMR spectroscopy data correlations: a description within the framework of topological index approach

*M. I. Trofimov and E. A. Smolenskii**

N. D. Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences,
47 Leninsky prosp., 117913 Moscow, Russian Federation.
Fax: +7 (095) 135 5328. E-mail: smolensk@cacr.ioc.ac.ru

A method is proposed for taking into account peculiarities of the topology of ring-containing organic molecules. The use of this method in combination with the previously developed procedure for calculations of the electronegativities of atoms in organic molecules makes it possible to develop a model for calculating the $J_{^{13}C-H}$ spin-spin coupling constant. The approach was evaluated with two sets of reported data. The method can be used for rapid assessment of fundamental characteristics of ring-containing organic compounds (including aromatic compounds) in solving a wide range of computer chemistry problems. The method and mathematical model are implemented as a module of the EDIP (Electrostatic Description of Index-Property Problem) program developed for solving various problems of organic computer chemistry within the framework of a topological index approach using electrostatic models.

Key words: computer chemistry, topological index, electronegativity, NMR, spin-spin coupling constant $J_{^{13}C-H}$, structure—property correlations, EDIP program.

The spin-spin coupling constant $J_{^{13}C-H}$ is a convenient physical parameter that is most often used in studying the peculiarities of the properties of ring-containing organic molecules. Interrelations between this constant and electronegativities have been considered.[1—3]

Different approaches have been proposed to take into account peculiarities of the structure of ring-containing molecules. We will outline some of them that proved to be efficient in solving various problems of organic computer chemistry. Based on analysis of these approaches,[3] we developed an algorithm for consideration of characteristic features of the topology of ring-containing organic molecules (see below). Since this is not a unique algorithm, we critically analyzed the efficiency of the known approaches listed below in the model experiments.

Numerous algorithms for enumerating all cycles of graphs have been proposed to date. The results of comparative analysis of some of these algorithms have been reported (see review[1]). Often, a large number of cycles of a molecular graph is an insurmountable barrier to implementation of an enumeration algorithm. It was reported that the real upper limit for such an algorithm is the molecular graph of ferrocene, which requires 10! or 3628800 iterations.[2]

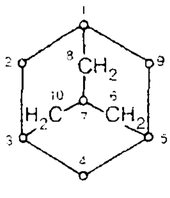
It is well known that there is no need to enumerate all rings constituting ring-containing organic molecules when naming them using systematic names. For in-

Translated from *Izvestiya Akademii Nauk. Seriya Khimicheskaya,* No. 3, pp. 401—406, March, 2000.

1066-5285/00/4903-0402 $25.00 © 2000 Kluwer Academic/Plenum Publishers

stance, only three rings in the molecule of adamantane (**1**) should be considered and the corresponding circuits enumerated following the known rules[3] to name this compound systematically (tricyclo[3.3.1.1^{3,7}]decane).

It follows from this example that the number of enumerated rings should be minimized, which requires enumeration of particular "chemically significant" rings only rather than all rings. In fact, any cycle (R) of a graph can be represented as a linear combination of basic cycles (B_i) of the graph[4]:



1

$$R = (\sum_{(i)}) B_i,$$

where (Σ) denotes the sum modulo 2 (a symmetric difference of sets), which is defined for two arbitrary sets A_1 and A_2 as

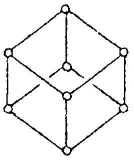
$$(\sum_{i=1}^{2}) A_i = (A_1 \cup A_2) \backslash (A_1 \cap A_2).$$

In the general case, several different bases of a graph can exist. However, most chemical applications require unambiguous choice of a set of "chemically significant" rings. In the early studies of different algorithms of search for the "chemically significant" rings, the definitions of "real" cycle, "pseudocycles," and the smallest covering cycle were proposed.[5,6]

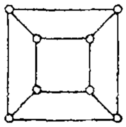
The concept of the "smallest set of smallest rings" (SSSR)[7–9] has become one of the most popular concepts developed for analyzing the ring-containing systems in chemical applications of the graph theory. In this concept, the criterion for small size of a ring has an advantage over the criterion for the size of the set. Specific drawbacks of the approach used, *e.g.*, inapplicability of the algorithm to particular molecular graphs corresponding to real (though rather exotic) chemical substances, have also been reported.[7–9] However, a more grave drawback of practical implementation of the SSSR concept is its low efficiency due to the necessity of enumerating and storing all cycles of a graph in the computer memory.

An efficient algorithm of the search for the SSSR was developed for use in the program that generates the Wiswesser linear code.[8] The SSSR concept was used in the analysis of ring-containing structures for assessing the ring strain energies,[10] as well as in the calculations of electronegativities and atomic charges.[11] A special algorithm of the search for the SSSR,[12] which is free from several drawbacks of the known algorithms,[7,8] was implemented in the latter case. The calculated ring strain energies were compared with analogous data obtained using the molecular mechanics method and with the experimental reaction enthalpies.[10] Correlations between calculated electronegativities, atomic radii, and bond energies for several molecules have been reported.[12]

High correlation coefficients obtained indicate correctness of the models used,[10,11] which are based on the SSSR concept. At the same time, mention of the following fundamentally important fact related to the SSSR concept has also been made in these works. There exist two representations of the molecular graph of cubane, namely, a three-dimensional representation (**2**) as a cube whose faces are formed by six four-membered cycles, and a planar representation (**3**), which can be considered as a structure formed by four-membered cycles (the external contour is not considered in this case).



2

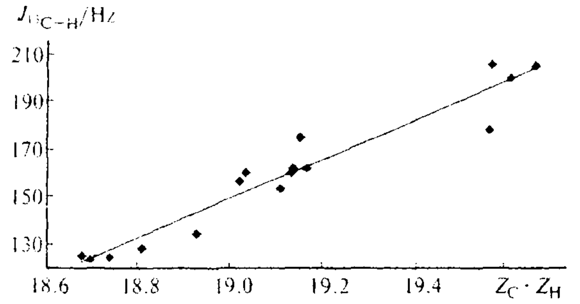


3

The number of linearly independent cycles in the molecular graph of cubane is 5 and the corresponding SSSR contains five four-membered cycles. At the same time, the experimental ring strain energy for cubane is 162.7 kcal mol^{-1} (26.5 kcal mol^{-1} for one four-membered ring). Thus, the closest estimate of the energy (26.5 · 6 = 159.0 kcal mol^{-1}) is obtained by approximating the cubane molecule using six four-membered cycles (instead of five four-membered cycles in accord with the SSSR concept). Critical analysis of the SSSR concept was reported.[9] The concept of an essential set of essential rings (ESER) was proposed,[13] developed to obtain necessary and sufficient information on the changes in the ring-containing structures in the course of reaction. Insufficiency of the SSSR concept for representing information required for solving the problems of molecular design has also been mentioned in this work.

In some cases, the problem of laconic and informative description of peculiarities of the topology of ring-containing systems is solved for particular classes of compounds taking into account characteristic features of the structure of their molecules (see, *e.g.*, Ref. 14, dedicated to description of fused aromatic molecules). Specific solutions have also been proposed such as the topological index approach, which explicitly considers the cyclic structure of molecules.[15]

Despite high popularity of the SSSR concept, it cannot be used in our case. Actually, calculations of electronegativities of carbon atoms in, *e.g.*, the cubane molecule, performed using this concept result in different values, which contradicts equivalency of the atoms. Therefore, in this work the computer search for "chemically essential" rings was carried out using a heuristic algorithm of the formation of a complete set of the smallest cycles described below.



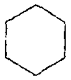
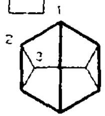
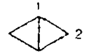
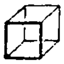
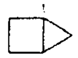
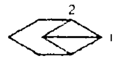
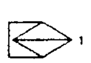
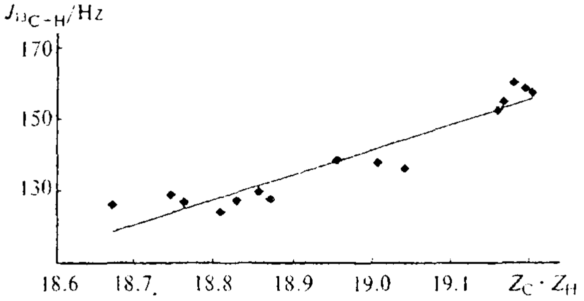
Molecule	Atom	$J_{^{13}C-H}$/Hz	
		Exp.	Calc.
		123	123.62
		124	127.43
		128	133.22
		134	143.06
	1	153	158.37
	2	175	161.68
	3	162	160.27
	1	156	150.84
	2	205	204.36
		160	160.10
		162	162.59
	1	178	195.68
		160	151.91
	1	200	199.65
	2	125	122.05
	1	206	196.17

Fig. 1. Dependence $J_{^{13}C-H} = a \cdot Z(\bar{S}_C) \cdot Z(\bar{S}_H) + b$ for sample of mono- and polycyclolkanes.

In this work, we performed calculations for the test sets of mono- and polycycloalkanes[16,19] (Fig. 1) and aromatic compounds[17] (Fig. 2) using the following formula which relates the $J_{^{13}C-H}$ values to electronegativities of the atoms:

$$J_{^{13}C-H} = a \cdot Z(\bar{S}_C) \cdot Z(\bar{S}_H) + b, \qquad (1)$$

where a and b are constants; $\bar{S}_C$, $\bar{S}_H$ are the electronegativities of C and H atoms constituting the molecule, respectively; and Z is a function of the type:



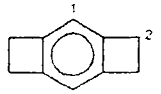
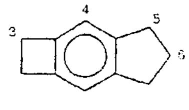
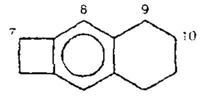
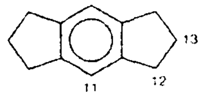
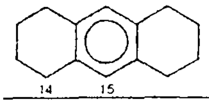
Molecule	Atom	$J_{^{13}C-H}$/Hz	
		Exp.	Calc.
	1	160.2	154.18
	2	138.4	138.30
	3	158.7	155.16
	4	129.4	131.55
	5	128.9	123.82
	6	137.6	141.95
	7	157.3	155.86
	8	124.0	128.18
	9	126.0	118.56
	10	136.0	144.34
	11	155.1	153.33
	12	127.5	132.47
	13	126.6	124.95
	14	152.3	152.77
	15	127.0	129.59

Fig. 2. Dependence $J_{^{13}C-H} = a \cdot Z(\bar{S}_C) \cdot Z(\bar{S}_H) + b$ for sample of aromatic compounds.

$$Z(\bar{S}) = M\sqrt{\bar{S}}, \qquad (2)$$

where M is the scale factor specific to each molecule. To assess the electronegativities of the atoms, we modified a known computational scheme[18] by introducing a correction based on calculations of topological indices of the "chemically significant" cycles.

Generation of complete set of the smallest cycles and the topological indices of "chemically significant" cycles

Let $v_1—v_2—v_3—...—v_{i-1}—v_i—...—v_{n-1}—v_n$ be a subgraph of graph G = {V,E}, where V is the set of all vertices of G, E is the set of all edges of G, and

$$\{v_1, v_2, v_3, ..., v_n\} \in V,$$

$$\{v_1—v_2, v_2—v_3, ..., v_{i-1}—v_i, ..., v_{n-1}—v_n\} \in E;$$

then this subgraph is called the circuit (v_1, v_2, v_3, ..., v_n) and $n - 1$ is the topological length of the circuit.

Let u, v be a pair of vertices connected by circuit l in G. Let G−l denote a graph obtained from G by excluding all edges belonging to circuit l.

The circuits k and l are called equal circuits ($k = l$) if they connect an equal number of vertices.

The circuit p is called a longer circuit if it connects a larger number of vertices than circuit l ($p > l$).

The circuit connecting all vertices and edges of the circuits k and l, which have at least one common vertex, is called a combination (or sum) of circuits ($k + l$). The initial vertex of the sum of circuits ($k + l$) is the initial vertex of circuit k. The final vertex of the sum of circuits ($k + l$) is the final vertex of circuit l.

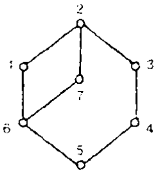
If in graph G_1 for any pair of vertices connected by circuits l_1^1 and l_1^2 of the cycle $l_1^1 + l_1^2$ ($l_1^1 \le l_1^2$) and for any circuit l_1^3 the condition

$$l_1^1 \le l_1^2 \le l_1^3 \qquad (3)$$

is met, the cycle $l_1^1 + l_1^2$ is called the smallest cycle of type I of graph G_1 (hereafter, the subscript in the circuit notation denotes the type of cycle while the superscript in the circuit notation denotes an arbitrary ordinal number of the circuit). If the cycle $l_1^1 + l_1^2$ is not the smallest cycle of type 1, *i.e.*, if there exists a circuit l_1^3 which does not meet condition (3), we will search for the smallest cycles of type I in graphs G_2 defined as follows: $G_2 = G_1 - l_1^1$ and $G_2' = G_1 - l_1^2$. These cycles are called the smallest cycles of type 2.

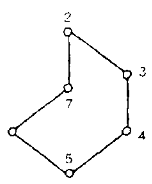
If the cycle $l_2^1 + l_2^2$ is not the smallest cycle of type 2 of graph G_2 (and, correspondingly, of graph G_2'), we will search for the smallest cycles of type I in graphs G_3 defined as follows: $G_3 = G_2 - l_2^1$ and $G_3' = G_2 - l_2^2$ (and, correspondingly, in graphs $G_3 = G_2' - l_2^1$ and $G_3' = G_2' - l_2^2$). These cycles are called the smallest cycles of type 3. Let us make an inductive generalization: all cycles of type 1 for the graphs $G_i = G_{i-1} - l^1_{i-1}$ and $G_i' = G'_{i-1} - l^2_{i-1}$ are called the smallest cycles of type i of graph G.

Let us clarify the definitions introduced using the following examples. For any pair of vertices of the 1—2—7—6 cycle of graph **4** there exists no circuit l_1^3 which simultaneously does not belong to the cycle and does not meet condition (3). Therefore the 1—2—7—6 cycle is the smallest cycle of type I of graph **4**.



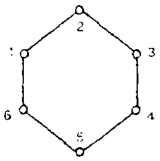
4

For the 1—2—3—4—5—6 cycle of graph **4** there exist a pair of vertices 2, 6, connected by circuit 2—7—6, which does not belong to this cycle and does not meet condition (3). Therefore, this cycle is not the smallest cycle of type I of graph **4**. By excluding the edges (1, 2) and (1, 6) of this graph we get graph **5** containing the 2—3—4—5—6—7 cycle, which is the smallest cycle of type 1 of graph **5**. Hence, this cycle is the smallest cycle of type 2 of graph **4**.

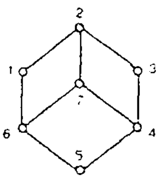


5

By analogy, the 2—3—4—5—6—7 cycle is not the smallest cycle of type I of graph **4**, since exclusion of the (2, 7) and (6, 7) edges generates graph **6** containing the 1—2—3—4—5—6 smallest cycle of type I of graph **5**. The latter cycle is the smallest cycle of type 2 of graph **4**. Thus, the complete set of the smallest cycles of graph **4** contains one four-membered cycle and two six-membered cycles.



6

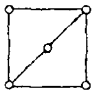


7

Let us consider yet another example. For any pair of the vertices of the cycles 1—2—7—6, 2—3—4—7, and 4—5—6—7 of graph **7** there exist no pairs of vertices which do not meet condition (3). Hence, these cycles are the smallest cycles of graph **7**.

For cycle 1—2—3—4—5—6 of graph **7** there exists a pair of vertices 2, 6, connected by the 2—7—6 circuit, which does not belong to the cycle and does not meet condition (3). By excluding the (1, 2) and (1, 6) edges we get a graph containing the 2—3—4—5—6—7 six-membered cycle, which is not the smallest cycle of the graph, since the 7—4 circuit exists. By excluding the (6, 5), (6, 7), and (4, 5) edges, we get a graph consisting of the above-mentioned smallest cycle 2—3—4—7.

By considering analogously all other cycles of graph **7** we can reveal that none of them, except for the above-mentioned cycles, is the smallest cycle. Thus, the complete set of the smallest cycles of graph **7** contains three four-membered cycles. The last example is graph **8** containing three four-membered smallest cycles of type 2.



8

Let us define the differential topological index (sometimes, it is called the vertex index) as follows:

$$C_i = \sum_{j=1}^{m} \frac{f_{ij}}{N_j},$$

where m is the number of "chemically significant" cycles of a molecular graph (in this work, this is the number of the smallest cycles);

$$f_{ij} = \begin{cases} 1, \text{ if vertex } i \text{ belongs to cycle } j; \\ 0 \text{ otherwise;} \end{cases}$$

and N_j is the size of cycle j.

In this case

$$\sum_{j=1}^{h} C_i = m,$$

where h is the number of all vertices in the cycles.

Calculations of electronegativities of the atoms in a molecule of an organic compound

The calculations of electronegativities are based on the known formula[18]

$$S_i = \sqrt[n_i+1]{S_i^0 \prod_{j=1}^{N} (k_{ij} S_j)},$$

where S_i is the desired electronegativity of the ith atom: S_i^0 are the electronegativities of individual atoms given according to Sanderson[19] (H 2.592, C 2.746, N 3.194, O 3.654, and F 4.000); n_i is the number of atoms bonded to the ith atom; N is the number of atoms in the molecule; and

$$k_{ij} = \begin{cases} 1, \text{ if atoms } i \text{ and } j \text{ are bonded;} \\ 0, \text{ if atoms } i \text{ and } j \text{ are nonbonded or } i = j. \end{cases}$$

By taking a logarithm of this relationship and denoting $X_i = \ln S_i$ and $X_i^0 = \ln S_i^0$ we get

$$X_i = \frac{1}{n_i + 1}\left(X_i^0 + \sum_{j=1}^{N} k_{ij} X_j \right).$$

After writing analogous relationships for all atoms ($i = 1, \ldots, N$) we get a system of N linear equations with N unknowns:

$$\begin{cases} X_1 = \dfrac{1}{n_1 + 1}\left(X_1^0 + \sum\limits_{j=1}^{N} k_{1j} X_j \right), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ X_N = \dfrac{1}{n_N + 1}\left(X_N^0 + \sum\limits_{j=1}^{N} k_{Nj} X_j \right). \end{cases} \tag{4}$$

It can be shown that such a system always has a unique solution.

Calculations of spin-spin coupling constants $J_{^{13}C-H}$

To calculate the spin-spin coupling constant $J_{^{13}C-H}$, let us normalize the C_i indices

$$\overline{C}_i = \frac{C_i}{m} = \frac{1}{m}\sum_{j=1}^{m}\frac{f_{ij}}{N_j}$$

(the notations are the same as above). According to formulas (4), we get

$$\overline{X}_i = \frac{1}{n_i + 1}\left(\sum_{j=1}^{m} \overline{X}_j + \overline{X}_i^0 \right),$$

where $\overline{X}_i^0$ is the normalized standard electronegativity calculated as

$$\overline{X}_i^0 = \begin{cases} X_i^0 - g(\overline{C}_i - 1/q) & \text{for all vertices of the molecular graph, which constitute the cycles taken into account in the calculations of electronegativities,} \\ X_i^0 & \text{for all other vertices;} \end{cases}$$

$\overline{X}_i$ is the "new" electronegativity calculated taking into account the presence of rings in the molecule; q is the number of vertices of the molecular graph, which constitute the cycles taken into account in the calculations of electronegativities; and g is the coefficient of proportionality, whose dimensionality coincides with that of the electronegativity.

To optimize the value of the coefficient of proportionality g, we studied the dependence of the correlation coefficient r^2 between $J_{^{13}C-H}$ and $Z_C \cdot Z_H$ on g and chose a g value of 0.9.

Let us define the scale factor M in relationship (2) as

$$M = S_{ag} \cdot (S_{max}/S_{min})^{1/n},$$

where S_{ag} is the electronegativity of the molecule, calculated as the geometric mean; S_{max} and S_{min} are the greatest and the smallest electronegativities among those of the atoms in the given molecule; and n is the number of atoms in the molecule.

It should be noted that the achievement of higher correlation coefficients requires consideration of molecular characteristics of another nature including the conformational parameters along with electronegativities. However, this goes beyond the scope of this work. The data taken for computer experiments are averaged experimental values that reflect different dynamic contributions of conformations to real systems. At the same time, it is obvious that such an averaging is far from leveling the conformational effects.

Thus, the computational procedure used in this work includes:

1) analysis of the molecular graph in order to reveal "chemically significant" cycles;

2) calculations and normalization of topological indices C_i;

3) calculations of the logarithms of normalized standard electronegativities $\overline{X}_i^0$;

4) formation and solution of a system of linear equations of the type

$$\overline{X}_i = \frac{1}{n_i + 1}\left(\overline{X}_i^0 + \sum_{j=1}^{N} k_{ij} \overline{X}_j \right);$$

5) calculations of the desired values using formulas (1) and (2).

The approach described above has been implemented as a module of the EDIP (Electrostatic Description of Index-Property Problem) program for IBM PC-compatible computers.[20] Detailed description of the software implementation goes beyond the scope of this work. However, it should be noted that no long iterative or recursive procedures are used in the program. This makes the computational cost low. Practically, calculations for samples similar to those presented in this work can be carried out on a personal computer in the interactive mode. The results of comparison of the experimental and calculated values are presented in Figs. 1 and 2. The correlation coefficients are $r = 0.964$ (the root-mean-square (RMS) deviation is 52.591) for the sample of 16 items (see Fig. 1) and $r = 0.945$ (the RMS deviation is 18.674) for the sample of 15 items (see Fig. 2). However, there is no need to place strong emphasis on the values of coefficients obtained for so few samples of so dissimilar chemical compounds.

Thus, a simple procedure proposed in this work for considering the effect of peculiarities of the topology on the properties of ring-containing organic molecules was evaluated by comparing the calculated values of the spin-spin coupling constant $J_{^{13}C-H}$ with the reported data. The results of comparison (see Figs. 1 and 2) can be considered to be satisfactory from the viewpoint of rapid computational methods used for solving different problems of the "structure—property" and "structure—activity" type, which is topical for computer chemistry of organic compounds. The method proposed in this work is based on two approaches that can be called "classical" ones. The first of them includes calculations of formal topological indices of molecular graphs to take into account specific features of the molecular structure rather than the elemental composition of the compound. The second approach includes calculations of more or less formal parameters (*e.g.*, electronegativities) that reflect to a greater extent the elemental composition of the molecule rather than the molecular structure. In this work, we succeeded in combining both these approaches.

References

1. P. Mateti and N. Deo, *SIAM J. Comput.*, 1976, **5**, 90.
2. R. G. Freeland, S. A. Funk, L. J. O'Korn, and G. A. Wilson, *J. Chem. Inf. Comput. Sci.*, 1979, **19**, 94.
3. J. D. Roberts and M. C. Caserio, *Basic Principles of Organic Chemistry. Supplement for Basic Principles of Organic Chemistry*, California Institute Technology, W. A. Benjamin, Inc., New York—Amsterdam, 1964; *A Guide to IUPAC Nomenclature of Organic Compounds*, Blackwell, Oxford, 1993, 182 pp.
4. W. Lipski, *Kombinatoryka dla Programistow* [*Combinatorial Analysis for Programmers*], Wydawnictwa Naukowo-Techniczne, Warszawa, 1982 (in Polish).
5. R. Fugman, U. Dolling, and H. Nickelsen, *Angew. Chem. Int. Ed. Engl.*, 1967, **6**, 723.
6. E. J. Corey and G. A. Petersson, *J. Am. Chem. Soc.*, 1972, **94**, 460.
7. M. Plotkin, *J. Chem. Doc.*, 1971, **11**, 60.
8. A. Zamora, *J. Chem. Inf. Comp. Sci.*, 1976, **16**, 40.
9. S. B. Elk, *J. Chem. Inf. Comp. Sci.*, 1988, **28**, 1.
10. J. Gasteiger and O. Dammer, *Tetrahedron*, 1978, **34**, 2939.
11. M. D. Guillen and J. Gasteiger, *Tetrahedron*, 1983, **39**, 1331.
12. J. Gasteiger and C. Jochum, *J. Chem. Inf. Comp. Sci.*, 1979, **19**, 43.
13. S. Fujita, *J. Chem. Inf. Comp. Sci.*, 1988, **28**, 1.
14. W. C. Herndon and A. J. Bruce, *J. Math. Chem.*, 1988, **2**, 155.
15. M. Randić, *J. Chem. Inf. Comp. Sci.*, 1988, **28**, 142.
16. H. Guenther and W. Herrig, *Chem. Ber.*, 1973, **106**, 3938.
17. R. P. Thummel and W. Nutakul, *J. Org. Chem.*, 1978, **43**, 3170.
18. N. S. Zefirov, M. A. Kirpichenok, F. F. Izmailov, and M. I. Trofimov, *Dokl. Akad. Nauk SSSR*, 1987, **296**, 883 [*Dokl. Chem.*, 1987 (Engl. Transl.)].
19. R. T. Sanderson, *J Am. Chem. Soc.*, 1983, **105**, 2259.
20. M. I. Trofimov and E. A. Smolenskii, *Int. Symp. CACR-96*, Moscow, 1996, 104 [this material can also be accessed *via* the Internet at URL http://www.glasnet.ru/~mtrofimow, "Computer Chemistry" section].

Received June 29, 1999