

Design of Molecules from Quantitative Structure-Activity Relationship Models. 2. Derivation and Proof of Information Transfer Relating Equations

Lowell H. Hall*

Department of Chemistry, Eastern Nazarene College, Quincy, Massachusetts 02170

Lemont B. Kier

School of Pharmacy, Virginia Commonwealth University, Richmond, Virginia 23298

Jack W. Frazer

Sterling-Winthrop Pharmaceutical Research Division, Malvern, Pennsylvania 19355

Received September 29, 1992

The second paper in this series presents the derivation and proof of the interconversion equations between graph path counts and vertex degree counts. These relations are first presented in a conceptual manner based on the inspection of information observed for acyclic graphs. Then these equations are derived in a more formal manner, first for acyclic systems and then for cyclic graphs. General relationships are also developed between edge-type counts and the counts of paths of length one and paths of length two and also between degree counts and edge-type counts. These developments provide the foundation for relationships between graph characteristics, obtained from molecular connectivity χ indexes and κ shape indexes and graph primitives, which are the basis for graph construction.

INTRODUCTION

In paper 1 in this series,¹ interconversion equations were introduced for chemical graphs for relations between the counts of vertex degrees, 1D , and the count of paths of orders 1 and 2, 1p and 2p , and the number of rings, R . The set of equations was introduced without proof.

From a practical point of view, the interconversion equations permit the construction of graphs from data on 1p , 2p , and R because they establish relationships between the low-order path counts and vertex degree counts. It remains to demonstrate the generality and rigor of these equations, which were intuitively developed, by providing a complete derivation and proof. This important development is presented here both for acyclic and cyclic systems. Terms and symbols used in this paper were introduced in paper 1, which outlined the general method. Paper 3 extends the development to path three count.

The interconversion equations are as follows:

$$^1D + ^4D + 3^5D + 6^6D = ^2p - ^1p + 3 - 3R \quad (1)$$

$$^2D - 3^4D - 8^5D - 15^6D = -2^2p + 3^1p - 3 + 3R \quad (2)$$

$$^3D + 3^4D + 6^5D + 10^6D = ^2p - ^1p + 1 - R \quad (3)$$

Here we added quantities for the fifth and sixth degree vertexes which did not appear in paper 1.

We will examine the conceptual basis of certain graph properties to establish the approach to a proof-of-concept for the relating equations. Further, these relations will be generalized in this paper to include the effect of the number of rings as well as branch points of orders higher than 4 and, in a later paper, to relations on path counts of orders higher than 2. In this development, we will examine information on acyclic graphs, such as that given in Figure 1, and then extend


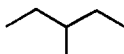
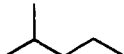
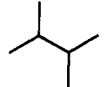

Graph	1p	2p	1D	2D	3D	4D
	5	4	2	4	0	0
	5	5	3	2	1	0
	5	5	3	2	1	0
	5	6	4	0	2	0
	5	7	4	1	0	1

Figure 1. Set of (hydrogen-suppressed) graphs for the isomeric acyclic hexanes, showing the counts of paths of length 1 and 2, along with the counts of vertexes of the four degrees.

to general cyclic and acyclic graphs.

OBSERVATIONS

The strategy we will follow in this development is to consider first the relations which exist in unbranched alkane graphs. Then we examine the impact of branching on path counts and vertex degrees while maintaining the number of atoms constant. Generalizations based on counts will be presented as four observations.²⁻⁴

Observation I. The introduction of a three-way branch ($^3D = 1$) into a molecular graph while maintaining a constant number of atoms, increases 2p by 1; 1D increases by 1; 3D increases by 1; 2D decreases by 2.

Observation II. The introduction of a four-way branch ($^4D = 1$) into a molecular graph, for a constant number of atoms, increases 2p by 3; 1D increases by 2; 4D increases by 1; 2D decreases by 3.

Observation III. The order-2 path count, 2p , increases as the number of branch points increases and as the vertex degrees increase.

* Author to whom correspondence should be addressed.

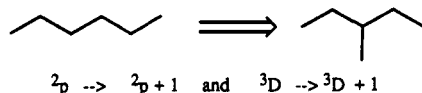


Figure 2. Conversion of the six-vertex unbranched chemical graph into a graph with one three-way branch point, illustrating the change in path count and degree count.

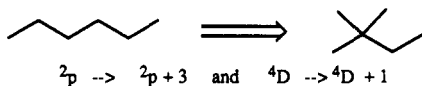


Figure 3. Conversion of the six-vertex unbranched chemical graph into a graph with one four-way branch point, illustrating the change in path count and degree count.

Observation IV. The 2p increase for each type of branch point is independent of the presence of other branch points.

Development of Equation 1. For unbranched, acyclic graphs, the number of paths of length 1 is one less than the number of atoms, A ; and the number of paths of length two (2p) is one less than the number of paths of length one (1p).

$${}^2p = {}^1p - 1 = A - 2 \quad (4)$$

When one three-way branch point is introduced (for example, by removing one terminal point from the graph and making it into a one-atom side chain at a branch point, holding A constant as in Figure 2), 2p increases by 1, while 1p remains constant. To maintain the equality in eq 4, the count of 3D must be added to the right-hand side. The increment of 2p is 1, which is equal to the increment in 3D . Thus, we obtain eq 5, which holds for graphs containing only three-way branch points.

$${}^2p = {}^1p - 1 + {}^3D \quad (5)$$

When one four-way branch point is introduced (for example, by removing two terminal atoms from the graph and making them into two one-atom side chains at a single branch point as in Figure 3), 2p increases by 3. To maintain the equality of either eqs 4 or 5, three times the count of four-way branch points must be added. The increment of 2p is 3, which is equal to three times the increment in 4D . Thus, we obtain eq 6.

$${}^2p = {}^1p - 1 + {}^3D + 3{}^4D \quad (6)$$

This equation holds for graphs containing both three- and four-way branch points. It can be rearranged to eq 1, a form with vertex degrees on the left side and path counts on the right.

$${}^3D + 3{}^4D = {}^2p - {}^1p + 1 \quad (7)$$

Development of Equation 2. Again, consider the basic relation for the unbranched alkane graph:

$${}^2p = {}^1p - 1 \quad (8)$$

Note also for this case that ${}^1D = 2$. We can then add this equality to that of eq 8, as follows:

$${}^2p + 2 = {}^1p - 1 + {}^1D \quad (9)$$

This equation also holds for graphs with three-way branch points since the incorporation of a three-way branch point increases 2p by 1 and 1D also increases by 1. See Observation I.

However, when a four-way branch is introduced (for example, by taking the two terminal vertexes and placing them onto the same vertex, thus, keeping the number of atoms

constant as in Figure 3), 2p increases by an additional count. See Observation II. Therefore, to maintain the equality, 4D must be included and 1D also increases by 2 so that the overall equality is maintained.

$${}^2p + 2 = {}^1p - 1 + {}^1D + {}^4D \quad (10)$$

This equation may be rearranged to eq 2.

$${}^1D + {}^4D = {}^2p - {}^1p - 3 \quad (11)$$

Development of Equation 3. To obtain the remaining relation, eq 3, we will use a simple conservation equation: the sum of the vertex degree counts must be equal to the number of graph vertexes, which for the noncyclic graph is $A = {}^1p + 1$.

$$\sum {}^iD = {}^1D + {}^2D + {}^3D + {}^4D = {}^1p + 1 = A \quad (12)$$

When eqs 1 and 2 are subtracted from eq 12, then eq 3 is obtained.

$${}^2D - 3{}^4D = -2{}^2p + 3{}^1p - 3 \quad (13)$$

Derivation and Proof of Basic Relations. The conceptual or intuitive development given above can be put into a more formal scheme which provides proof of the three relating equations. This scheme is more readily extended to higher order branch points and to cyclic systems. We will develop three basic equations, different in appearance from eqs 1–3 and show that they are equivalent.

The simplest of the basic relations is obtained from the fact that the sum of vertex degree counts is the number of graph vertexes or the count of skeletal atoms, A . Each graph vertex with i neighbors has one vertex degree count, iD . Further, for

$${}^1D + {}^2D + {}^3D + \dots + {}^nD = \sum {}^iD = A \quad (14)$$

acyclic graphs the path-1 count, 1p , is simply related to the atom count, A .

$${}^1p = A - 1 \quad (15)$$

Therefore, we obtain the first basic relation for acyclic graphs:

$${}^1D + {}^2D + {}^3D + \dots + {}^nD = \sum {}^iD = {}^1p + 1 \quad (16)$$

The second basic relation is the well-known Handshake Lemma:⁵ the sum of the vertex degrees is twice the number of edges, 1p . The number of vertex degrees is the number of vertexes of a given degree, iD , times the degree, i . This equation

$$1({}^1D) + 2({}^2D) + 3({}^3D) + \dots + n({}^nD) = \sum (i){}^iD = 2({}^1p) \quad (17)$$

is labeled eq E in the final table of equations.

A third basic relation is obtained by counting paths of length 2, realizing that such a count has to be made only at branch points with no adjacency information necessary: 2p is independent of the arrangement of branch points. This independence arises because each path of order 2 involves only edges adjacent to the branch point. The order-2 path, centered on a given vertex, does not extend beyond adjacent vertexes, and therefore, only their count but not their degree influences the path-2 count. The count of paths of length 2 is essentially a combinatorics relation: the number of ways two edges can be obtained from a set of n edges at a branch point, $N(n,2)$.

$$N(n,2) = n!/2!(n-2)! \quad (18)$$

Based on these considerations, it is possible to write an equation for 2p generated by all the branch points in a graph, including branch points of all orders. For our current con-

$${}^2D + 3({}^3D) + 6({}^4D) + 10({}^5D) + 15({}^6D) + \dots + \frac{n!}{2!(n-2)!}({}^nD) = {}^2p \quad (19)$$

sideration we will include vertexes with degree up to six. This relation is labeled F in the final equations.

It is now possible to derive the interconversion equations (for acyclic graphs) by algebraic manipulation of eqs 15, 17, and 19. The resulting relations are listed as eqs a, b, and c below. These relations are written explicitly for branch points up to and including the sixth order for acyclic graphs.

It can be shown that eq a is obtained as eq 19 + eq 17 - eq 16. In like manner, eq b = eq 17 - eq 16 - eq c. Finally, eq a = eq 16 - eq b - eq c. These equations will be given later, also labeled A, B, and C but with appropriate modification to include cyclic systems.

Interconversion Equations (Acyclic Graphs).

$${}^1D + {}^4D + 3{}^5D + 6{}^6D = {}^2p - {}^1p + 3 \quad (a)$$

$${}^2D - 3{}^4D - 8{}^5D - 15{}^6D = -2{}^2p + 3{}^1p - 3 \quad (b)$$

$${}^3D + 3{}^4D + 6{}^5D + 10{}^6D = {}^2p - {}^1p + 1 \quad (c)$$

Extension to Cyclic Graphs. To extend the interconversion equations to cyclic systems, it is only necessary to consider those basic relations which are directly affected by cyclization of graphs. The relation between the number of vertexes, A , and the number of edges, 1p , referred to as eq 16, is the only relation among the basic relations which is shown to depend upon the number of rings in the graph. Neither the Handshake Lemma (eq 17) nor the relation on the count of 2p by branch-point contribution (eq 19) depends upon the number of rings in a graph.

The relation among the number of atoms, rings, and path-1 count (edges) is well-known as the relation among vertexes, faces, and edges of geometric figures:⁶

$$V + F = E + 1 \quad (20)$$

where E is the number of edges, F is the number of faces, and V is the number of vertexes. In the notation we are using, this relation is written as

$$A + R = {}^1p + 1 \quad (21)$$

For each added ring, both R and 1p increase by 1. To express the number of rings:

$$R = {}^1p - (A - 1) \quad (22)$$

This relation can be readily incorporated into eq 16 as follows:

$${}^1D + {}^2D + {}^3D + \dots + {}^nD = \sum {}^iD = A = {}^1p + 1 - R \quad (23)$$

The quantity R is the count of rings in the graph and is also known in chemical graph theory as the cyclomatic number:⁶ $CN = {}^1p - (A - 1)$. It should be pointed out that this manner of counting rings corresponds to the usual way that organic chemists count rings in a molecular structure. However, it does not correspond to the usual way faces are counted in polyhedral structures. For those systems the usual count is based on the Euler formula:⁷

$$V + F = E + 2 \quad (24)$$

The difference between these two relations is that in the reckoning of the cyclomatic number, one of the faces counted

in eq 24 is taken as an 'infinite' face in a polyhedral figure and is not counted in the cyclomatic number. This difference may be illustrated with the two views of a tetrahedron below: a tetrahedron shown in perspective and the graph of a tetrahedron shown in Figure 4.

Use of eq 21, with $A = 4$ and ${}^1p = 6$, leads to $R = 3$; the three rings may be identified with the three 'internal' triangles in the figure to the right. Use of eq 24 leads to $F = 4$. The 'fourth face' or ring in the left figure becomes the 'infinite face' in the figure to the right. So as long as we use eq 21, no confusion need arise in applying the interconversion equations.

This concept may be further illustrated with a more complex molecular structure. Figure 5 shows the molecular structure of adamantane. The traditional structural formula is shown in Figure 5a along with the hydrogen-suppressed graph in Figure 5b. For adamantane there are 10 skeletal groups (CH_2 or CH) or graph vertexes: $A = 10$; there are 12 skeletal bonds or graph edges: ${}^1p = 12$. Hence, by eq 22, $R = 3$, as can be more clearly seen in Figure 5c, which is the hydrogen-suppressed graph redrawn in a more standard form.

Interconversion Equations (Acyclic and Cyclic). It is now possible to derive the interconversion equations by algebraic manipulation of the basic relations (modified to include R). The final interconversion equations are listed as eqs A, B, and C below. Additional equations are tabulated below as D to I. These relations are written explicitly for branch points up to and including the sixth order.

A useful intermediate relation, G , is obtained from $E - D$. Two of the interconversion equations can now be obtained. C can be obtained as $F - G$ and B as $G - 2C$. Relation H can also be obtained as $B + C$. Finally, A can be obtained as $D - H$. Further, it is possible to obtain a relation among the degree counts which does not explicitly contain path counts, eq I. In this way, the interconversion equations used in these papers are obtained and labeled A, B, and C in the following table.

Interconversion Relations for Acyclic and Cyclic Graphs.

$${}^1D + {}^4D + 3{}^5D + 6{}^6D = {}^2p - {}^1p + 3 - 3R \quad (A)$$

$${}^2D - 3{}^4D - 8{}^5D - 15{}^6D = -2{}^2p + 3{}^1p - 3 + 3R \quad (B)$$

$${}^3D + 3{}^4D + 6{}^5D + 10{}^6D = {}^2p - {}^1p + 1 - R \quad (C)$$

$${}^1D + {}^2D + {}^3D + {}^4D + {}^5D + {}^6D = {}^1p + 1 - R \quad (D)$$

$${}^1D + {}^2D + {}^3D + {}^4D + {}^5D + {}^6D = {}^2p \quad (E)$$

$${}^2D + 3{}^3D + 6{}^4D + 10{}^5D + 15{}^6D = {}^2p \quad (F)$$

$${}^2D + {}^3D + {}^4D + {}^5D + {}^6D = {}^1p - 1 + R \quad (G)$$

$${}^2D + {}^3D - 2{}^5D - 5{}^6D = -2{}^2p + 2{}^1p - 2 + 2R \quad (H)$$

$${}^1D - 3{}^3D - 2{}^4D - 3{}^5D - 4{}^6D = 2 - 2R \quad (I)$$

Relation between Low-Order Path Counts and Edge Type Counts. The development above shows how low-order path counts are related to counts of vertex degrees. It is also possible to develop relations between path counts and counts of edge types. In typical chemical graphs for organic molecules, there are four types of vertexes, corresponding to degrees of one, two, three, and four. (For this present discussion we will restrict our investigations to degrees of four or less. The

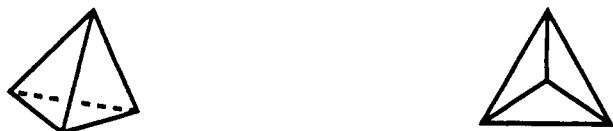


Figure 4. Drawing of a tetrahedron shown in perspective to reveal all four faces along with the graph of a tetrahedron, revealing the three rings (equal to the cyclomatic number).

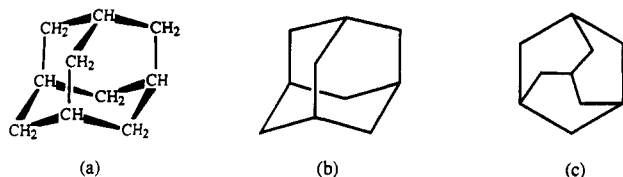


Figure 5. Structure representations of adamantane to illustrate the counting of rings. (a) Structural molecular formula of adamantane; (b) hydrogen-suppressed graph of adamantane, (c) another rendering of the hydrogen-suppressed graph of adamantane which more clearly reveals the ring count as three. See text.

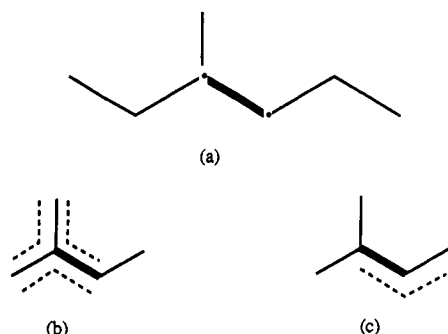


Figure 6. Illustration of the contributions to path-2 count, 2p , from a '23'-type edge. (a) The 3-methylhexane hydrogen-suppressed graph with focus on the '23' edge in boldface; (b) the '23' edge with appended edges, also showing the three paths of length 2 (dashed lines) associated with the vertex of degree three; (c) the '23' edge with appended edges, also showing the one path of length 2 (dashed lines) associated with the vertex of degree two.

extension to higher order degrees is straightforward.) An edge type is symbolized by e_{ij} where i and j are the degrees of the vertexes at each ends of the edge. For the four degree types there are 10 edge types: e_{11} , e_{12} , e_{13} , e_{14} , e_{22} , e_{23} , e_{24} , e_{33} , e_{34} , and e_{44} . To represent the count of each edge type, we use the symbol ne_{ij} .

Since 1p is the count of edges, then it is simply the sum of all the edge-type counts:

$$^1p = ne_{11} + ne_{12} + ne_{13} + ne_{14} + ne_{22} + ne_{23} + ne_{24} + ne_{33} + ne_{34} + ne_{44} \quad (25)$$

The count of paths of length 2 may also be related to edge-type counts. A particular edge may be extracted from a graph along with any appended edges, so as to reveal any path of length 2 associated with that particular edge. Consider Figure 6 and the embedded '23' edge, indicated by the presence of two dots and the boldface line.

From Figure 6a it may be said that the vertex of degree three generates three paths of length 2 and that the vertex of degree two generates two paths of length 2. However, in each case, these same paths of length 2 are also generated by adjacent edges. For example, the vertex of degree three generates the same path-2 three times, once for each of the edges incident at the vertex of degree three. Hence, the edge contributions to 2p are not independent and must be divided by the appropriate multiplicity factor, as shown in Table I. It follows, then, that the relation between 2p and the count

Table I. Contributions from Edge-Types to Counts of Paths of Length 2, 2p

ij	embedded edge	expressed contributions ^a	contribution to 2p
11		$0 + 0(0)$	0
12		$0 + 1(1/2)$	$1/2$
13		$0 + 3(1/3)$	1
14		$0 + 6(1/4)$	$3/2$
22		$1(1/2) + 1(1/2)$	1
23		$1(1/2) + 3(1/3)$	$3/2$
24		$1(1/2) + 6(1/4)$	2
33		$3(1/3) + 3(1/3)$	2
34		$3(1/3) + 6(1/4)$	$5/2$
44		$6(1/4) + 6(1/4)$	3

^a Contribution is expressed in two parts: the first term is the 2p contribution arising from the left vertex in the embedded edge, and the second term arises from the right vertex.

of edge types may be written as follows:

$$^2p = 1/2ne_{12} + ne_{13} + 3/2ne_{14} + ne_{22} + 3/2ne_{23} + 2ne_{24} + 2ne_{33} + 5/2ne_{34} + 3ne_{44} \quad (26)$$

In this manner both 1p and 2p may be related to edge-type counts. These relations may become part of a scheme to determine edge-type counts, or they may serve as restricting relations in the use of degree counts for the construction of graphs. It will be shown in subsequent papers how the higher order path counts 3p and 4p may be related to edge-type counts and related graph quantities.

It is also possible to obtain relations between vertex degree counts and edge-type counts. These relations arise from the fact that the presence of an edge type, e_{ij} , requires the presence of associated vertexes with degrees i and j . A vertex of degree j will be part of j edges of type ij . Thus, the following must hold:

$$^1D = 2ne_{11} + ne_{12} + ne_{13} + ne_{14} \quad (27)$$

$$^2D = 1/2[ne_{12} + 2ne_{22} + ne_{23} + ne_{24}] \quad (28)$$

$$^3D = 1/3[ne_{13} + ne_{23} + 2ne_{33} + ne_{34}] \quad (29)$$

$$^4D = 1/4[ne_{14} + ne_{24} + ne_{34} + 2ne_{44}] \quad (30)$$

These relations are not independent because the sum $\sum(i)^iD = 2^1p$ is the Handshake Lemma, eq 17. These relations are useful in determining valid degree sets as will be discussed in a subsequent paper.

DISCUSSION

A graph may be represented by the number of vertexes A and the number of rings R . These two quantities may be considered as the independent variables which describe the graph at its most primitive level. A and R may be considered the simplest measures of size and complexity. From this perspective, the edge count 1p is a dependent variable,

depending only upon the atom and ring counts. One need not know anything of the branching pattern, adjacency, or details of the ring structure in order to obtain the edge count: ${}^1p = (A - 1) + R$. Relations may be obtained between 1p and other graph primitives such as the vertex degree counts and edge-type counts, as shown above.

Observations on graphs clearly indicate that the count of paths of length 2, 2p , requires a greater level of information than that embodied in atom and ring counts. Elementary adjacency information is required. Specifically, the actual vertex degrees are the most primitive graph information needed. This relation is reported as eq 19 and states that the vertex degree contributions to 2p for each vertex are independent. In order to compute the 2p contribution of a given vertex, one does not need to know the nature of adjacent vertexes; only the number and the degree of vertexes is required.

For higher order path counts, it also clear that the graph information required is of greater complexity than atom and ring counts along with vertex degree counts. A greater measure of adjacency information is required. It will be shown in subsequent papers how this information is obtained from primitives such as edge-type counts and related quantities.

In discussions of chemical graph theory, a question often arises as to the role of three-dimensional molecular geometry information. It is clear that the graph representation of molecular structure does not explicitly contain three-dimensional information.^{2-4,6,8} However, the various indexes of structure which are developed in chemical graph theory clearly contain three-dimensional information in an implicit manner.^{2,3} The quality of QSAR models of physicochemical and biological properties based on graph-theoretical models require that some three-dimensional structure information is implicitly encoded^{4,6,8} because these properties depend in part on three-dimensional geometry of the molecules.

In Paper I in this series, an example showed that QSAR models based on molecular connectivity indexes can generate molecular structures which possess a desired property value. Since properties depend in part upon three-dimensional geometry, this inverse QSAR method must implicitly contain

three-dimensional information. It should also be stated that not all aspects of three-dimensional geometry are encoded in the indexes presently available. Such issues remain as important areas to be investigated.

CONCLUSIONS

It is possible to derive equations which give the relation between path counts and vertex degrees for chemical graphs. The three equations A, B, and C are independent and may be used to develop relations from QSAR equations as a basis for obtaining vertex degrees. Further, basic relations also exist among low-order path counts and edge-type counts. The approach developed here is a foundation for development of additional relationships which may prove useful in the inverse imaging process.

ACKNOWLEDGMENT

We wish to express appreciation for support of this work from Sterling-Winthrop Pharmaceutical Research Division, Inc., Malvern, PA, and to Robert S. Dailey for helpful discussions and preparation of figures and tables.

REFERENCES AND NOTES

- (1) Kier, L. B.; Hall, L. H.; Frazer, J. W. Design of Molecules from Quantitative Structure-Activity Relationship Models. 1. Information Transfer between Path and Vertex Degree Counts. *J. Chem. Inf. Comput. Sci.* **1993**, preceding paper in this issue.
- (2) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
- (3) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; John Wiley: London, 1986.
- (4) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983; Vols. I and II.
- (5) See ref 4, p 10.
- (6) Hall, L. H. Computational Aspects of Molecular Connectivity and its Role in Structure-Property Modeling. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Press: New York, 1990; Chapter 8, p 213.
- (7) See ref 4, pp 12-13.
- (8) Rouvray, D. H. Predicting Chemistry from Topology. *Sci. Am.* **1986**, 255, 40.