On Highly Discriminating Molecular Topological Index

Chang-Yu Hu and Lu Xu*

Applied Spectroscopy Laboratory, Changchun Institute of Applied Chemistry, Academia Sinica, Changchun 130022, Jilin, People's Republic of China

Received September 8, 1995[⊗]

A highly discriminating molecular topological index, EAID, is proposed based on the extended adjacency matrix. A systematic search for degeneracy was performed for 3 807 434 alkane trees, 202 558 complex cyclic or polycyclic graphs, and 430 472 structures containing heteroatoms. No counterexamples (two or more nonisomorphic structures with the same EAID number) were found. This is a hitherto unheard of power of discrimination. Thus EAID might be possibly used as supplementary reference for CAS Registry Numbers for structure documentation.

1. INTRODUCTION

Topological indices are graph invariants which are independent of the particular drawing or numbering of the graph's vertices. They are single numbers for characterization of molecular structures. Two goals in developing topological indices are followed: (i) design of a useful descriptor for QSAR/QSPR[1−12] and (ii) design of a useful descriptor for chemical documentation, thus allowing storage and retrieval.[13−16] Those two goals do not parallel one another. For the latter purpose, many efforts have been made in pursuing the highly discriminatory power to find a mathematical invariant to characterize a molecule uniquely. The Matula numbers are a topological index for alkane trees.[17] Randić's molecular identification number, ID, was defined as the sum of all weighted paths in a molecule.[14] An exhaustive search revealed that the smallest degenerate alkane trees with the same ID numbers occurred for $n = 15$ vertices. By assigning prime number weight values to various bond types, the revised prime ID[15] is unique for alkane trees with up to 19 vertices. Balaban suggested a different weighted approach namely to replace the vertex degree in the original ID formula by the distance sum which resulted in the index BID.[18] According to Muller et al.,[19] BID numbers were unique for alkane trees with up to at least 20 vertices. Muller and collaborators' SID (the self-returning walk ID numbers) were unique for alkane trees with up to 18 vertices.[20] Hall and Kier introduced the δ^v values, molecular connectivity valence, to characterize heteroatoms, and a different scheme for weighting the path between nodes i and j, proposing the topological index τ.[21] The search for counterexamples to the total topological index τ in the field of alkane trees up to 20 atoms produced five degeneracies, two pairs in the nonadecane family, and three pairs in the eicosane family.[22] This topological index τ can be applied to structures containing heteroatoms. The authors used another different formula for weighting the path and developed a new topological index which was unique for alkane trees with up to at least 20 vertices.[16]

In this study, a new scheme for weighting the edge of the molecular graphs was proposed to extend the adjacency matrix. A new topological index was developed based on the power of the extended adjacency matrix, namely EAID, which is a highly discriminating molecular IDentification number.

H 0. 37						
Li 1. 225	Be 0. 889	B 0. 80	C 0. 74	N 0. 74	O 0. 74	F 0. 72
Na 1. 572	Mg 1. 364	Al 1. 248.	Si 1. 173	P 1. 10	S 1. 04	Cl 0. 994
						Br 1. 142
						I 1. 334

Figure 1. The covalent radii of some atoms (in angstroms, Å).[26] Note that Li, Be, Na, and Mg usually form highly polarized covalent bonds (practically ionic bonds).

2. OUTLINE OF THE APPROACH

In this study, a molecule is viewed as a colored graph in which vertices are interpreted as distinct atoms and the edges are colored by multiple connections. In our systematic approach, the distinct atoms are characterized by their covalent radii of the atoms (see Figure 1) and their connectivity valence δ which is similar to Hall and Kier's δ^v [10]

$$\delta = Z - h$$

where Z is the number of valence electrons and h is the number of attached hydrogen atoms. The δ values are listed in Table 1. The multiple bonds are single, double, triple bond, and aromatic bonds which are coded as 1, 2, 3, and 1. 5, respectively. The algorithm can be described by following steps.

(1) Set the weight of atom $S[i]$. As in earlier papers,[23−25] a regressive weight is introduced from layer matrix. From the viewpoint of node i, i.e., the immediate neighbors of the node i form the first layer; the outer neighbors connecting immediately to the nodes of the first layer form the second layer, and so on. Then, a new type layer-matrix, connectivity valence matrix (CVM), is proposed, whose element $(cvm)_{ij}$ is defined as the sum of molecular connectivity valence (the δ value) for all nodes situated in the jth layer. The bond

[⊗] Abstract published in *Advance ACS Abstracts,* January 1, 1996.

0095-2338/96/1636-0082$12.00/0

Table 1. Molecular Connectivity Valence (the δ Values) of Some Functional Groups

no.	symbol	δ	no.	symbol	δ
1	CH_3-	1	20	>N(=)-	5
2	-CH_2-	2	21	-N(=)=	5
3	CH_2=	2	22	=N≡	5
4	>CH-	3	23	HS-	5
5	-CH=	3	24	-S-	6
6	CH≡	3	25	S=	6
7	>C<	4	26	>S=	6
8	>C=	4	27	=S=	6
9	-C≡	4	28	>S(=)=	6
10	=C=	4	29	H_2P-	3
11	HO	5	30	-PH_2=	3
12	-O-	6	31	-PH-	4
13	O=	6	32	>PH=	4
14	NH_2-	3	33	>P-	5
15	-NH-	4	34	=P<-	5
16	NH=	4	35	F-	7
17	>N-	5	36	Cl-	7
18	-N=	5	37	Br-	7
19	N≡	5	38	I-	7

matrix (B) is proposed, whose element b_{ij} is defined as the sum of the codes of the bonds which connected between the nodes of the jth layer and the nodes of the $(j-1)$th layer. The weight of a node is calculated by the following function

$$S[i] = (\mathrm{cvm})_{i1} + \sum_{j=1}^{K} (\mathrm{cmv})_{i(j+1)} b_{ij} \times 10^{-j}$$

where $S[i]$ is the weight of the ith node; $(\mathrm{cvm})_{ij}$ and b_{ij} are the elements of the CVM and B matrices, respectively; and K is the number of layers from the viewpoint of node i.

(2) Set up adjacency matrix $A = \{a_{ij}\}$

$$a_{ij} = \begin{cases} 0 & \text{no connection} \\ 1 & \text{a single bond} \\ 2 & \text{a double bond} \\ 2 & \text{a double bond} \\ 3 & \text{a triple bond} \\ 1.5 & \text{an aromatic bond} \end{cases}$$

(3) Set up the extended adjacency matrix $\mathrm{EA} = \{(\mathrm{ea})_{ij}\}$

$$(\mathrm{ea})_{ij} = \begin{cases} \dfrac{\sqrt{(\mathrm{Radii})_i}}{6} & i = j \\ \dfrac{(\sqrt{a_{ij}})w_{ij}}{6} & i \neq j \end{cases}$$

Where $(\mathrm{Radii})_i$ is the covalent radii (in angstroms, Å) of the ith atom (see Figure 1), and w_{ij} is a weight factor calculated by the following function

$$w_{ij} = \sqrt{\frac{S[i]}{S[j]}} + \sqrt{\frac{S[j]}{S[i]}}$$

where $S[i]$ is the weight of the ith node.

(4) Evaluate a new matrix $\mathrm{EA}^* = \{(\mathrm{ea})_{ij}\}$ (sum of the EA matrix powers)

$$\mathrm{EA}^* = \sum_{k=0}^{N-1} (\mathrm{EA})^k$$

Table 2. Numbers of Some Alkane Isomers (Molecular Formula: C_nH_{2n+2})

n	no. of alkane isomers	total no. of alkane isomers	n	no. of alkane isomers	total no. of alkane isomers
1	1	1	12	355	664
2	1	2	13	802	1466
3	1	3	14	1859	3324
4	2	5	15	4347	7671
5	3	8	16	10359	18030
6	5	13	17	24894	42924
7	9	22	18	60523	103447
8	18	40	19	148284	251731
9	35	75	20	366319	618050
10	75	150	21	910726	1528776
11	159	309	22	2278658	3807434

Table 3. Numbers of Some Cyclic or Polycyclic Isomers of Carbon Molecules (Molecular Formula: C_n)

n	no. of isomers	total no. of isomers
3	1	1
4	3	4
5	6	10
6	19	29
7	50	79
8	204	283
9	832	1115
10	4330	5445
11	25227	30672
12	171886	202558

Table 4. Some Molecular formulas (Having $n = 8$ Vertexes) and Their Isomer Numbers

molecular formula	no. of isomers	total no. of isomers	molecular formula	no. of isomers	total no. of isomers
C_8H_{18}	18	18	C_7H_9N	24627	127723
$C_7H_{16}O$	72	90	C_6H_7NO	61255	188978
$C_7H_{17}N$	89	179	C_8H_8	7437	196415
$C_6H_{15}NO$	405	584	C_7H_6O	15804	212219
C_8H_{16}	139	723	C_7H_7N	34745	246964
$C_7H_{14}O$	596	1319	C_6H_5NO	61974	308938
$C_7H_{15}N$	801	2120	C_8H_6	7982	316920
$C_6H_{13}NO$	3418	5538	C_7H_4O	11332	328252
C_8H_{14}	654	6192	C_7H_5N	31163	359415
$C_7H_{12}O$	2589	8781	C_6H_3NO	33896	393311
$C_7H_{13}N$	3826	12607	C_8H_4	5308	398619
$C_6H_{11}NO$	14410	27017	C_7H_2O	3971	402590
C_8H_{12}	2082	29099	C_7H_3N	15489	418097
$C_7H_{10}O$	7166	36265	C_6H_1NO	7038	425117
$C_7H_{11}N$	11773	48038	C_8H_2	1804	426921
C_6H_9NO	37202	85240	C_7O	356	427277
C_8H_{10}	4679	89919	C_7HN	2991	430268
C_7H_8O	13177	103096	C_8	204	430472

where N is the number of atoms in the molecule. When $k = 0$, $(\mathrm{EA})^0$ is an **I** matrix.

(5) Calculate the topological index EAID

$$\mathrm{EAID} = \sum_{i=1}^{N} (\mathrm{ea}^*)_{ii}$$

Figure 2 traces the algorithm for structure I.

3. GENERATION OF CONSTITUTIONAL ISOMERS

In order to assess the discriminatory power of the proposed topological index in a systematic manner, an exhaustive and irredundant structural isomer generator is useful to produce all members of a given family of chemical structures.

Major requirements in the design of a structural generator include exhaustiveness, irredundancy, and efficiency. The

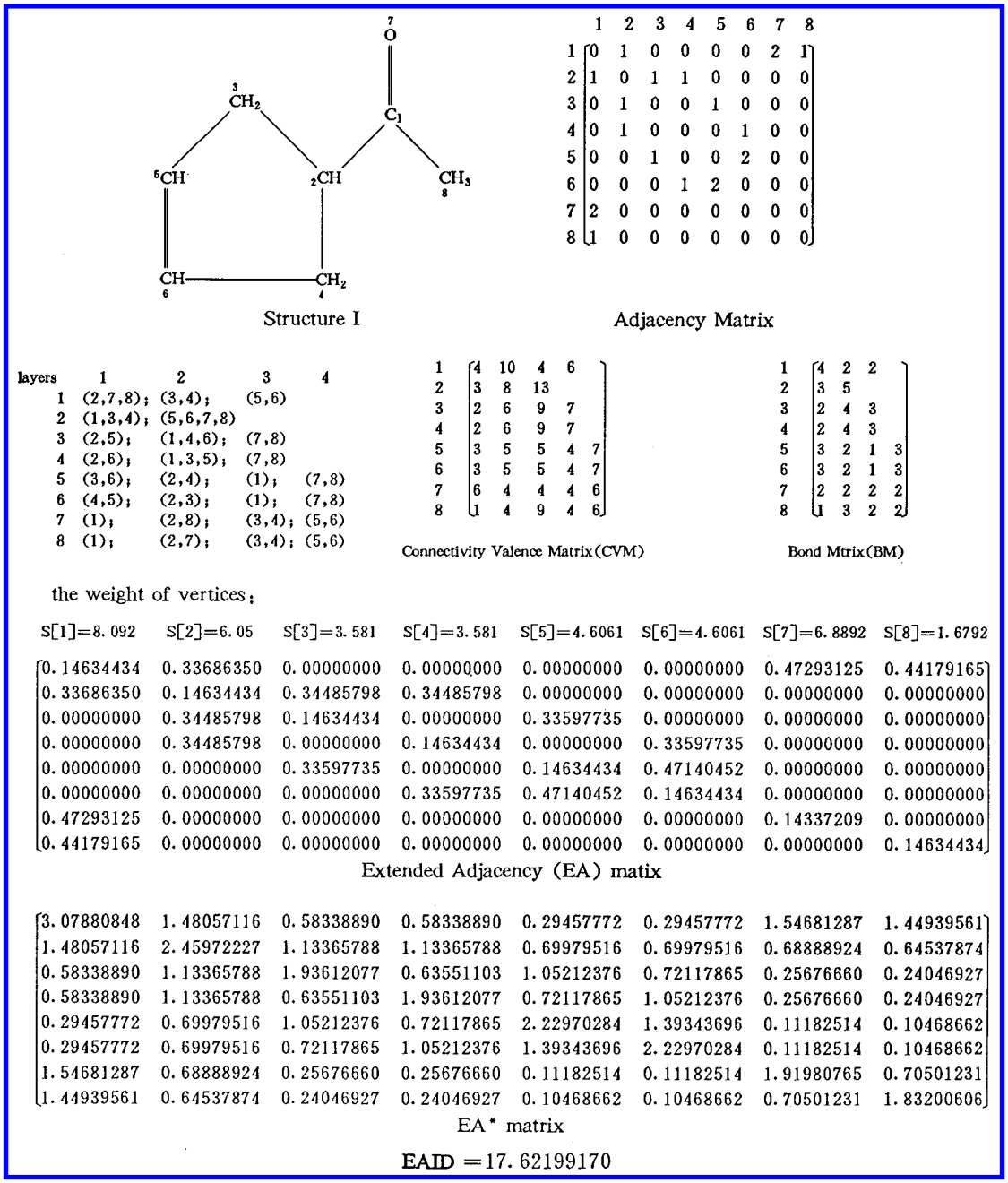


Figure 2. Computation of the EAID number.

exhaustiveness and irredundancy can be reached in many ways. But the efficiency is important for a practicing generator, especially in generating the isomers of large molecules, because the population of isomers is very large (there are total 2 278 658 isomers for the molecular formula $C_{22}H_{46}$), and it is extremely time consuming to consider all the $N!$ permutations for isomorphism check.

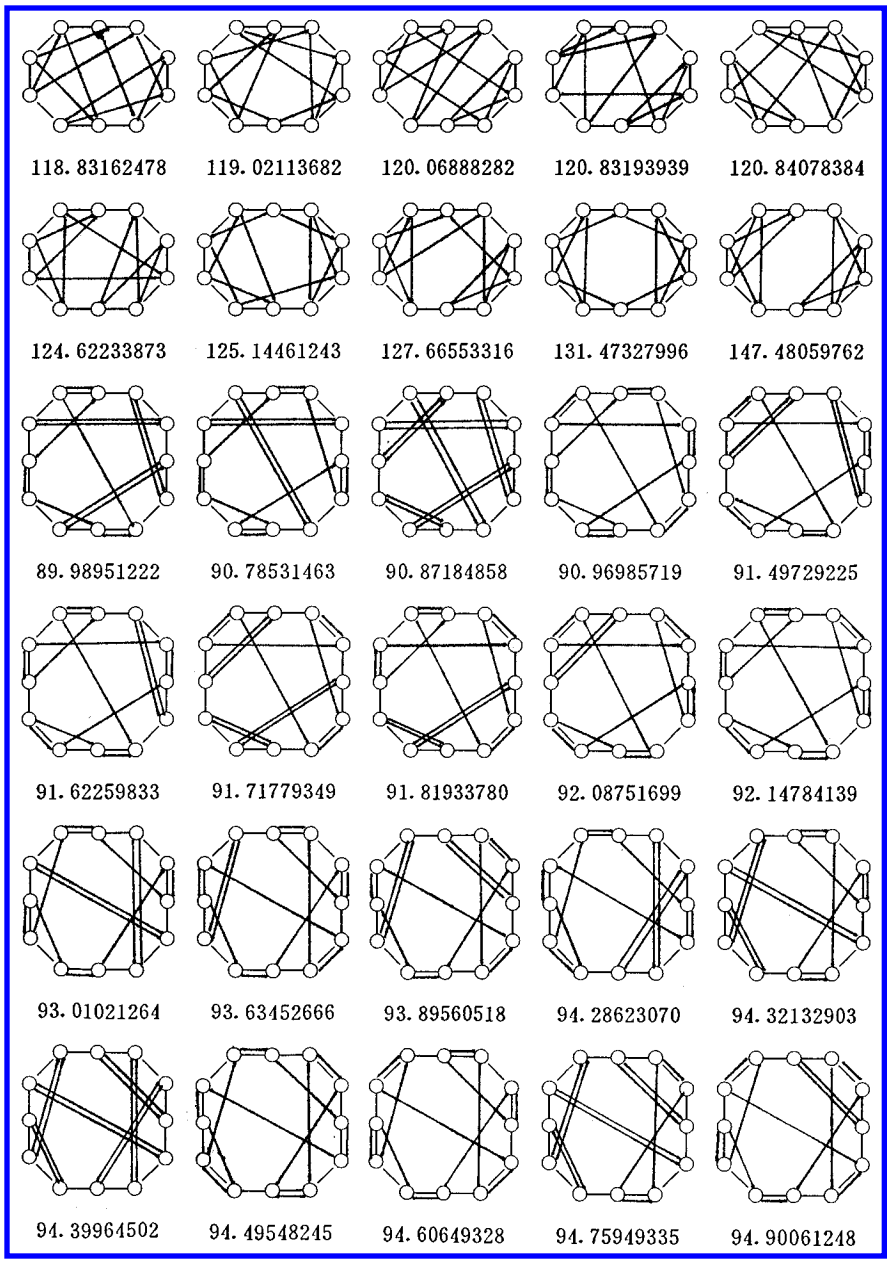
In our laboratory, an algorithm for the exhaustive and irredundant generation of structural isomers from molecular formula has been devised, and its exhaustiveness and irredundancy has been shown.[23,27,28] To substantiate the exhaustive and irredundant character of the algorithm, some empirical formulas were supplied to the structure generator to build up all the isomers for the given empirical formulas. The numbers of structural isomers generated for alkanes (C_nH_{2n+2}), alcohols, and ethers ($C_nH_{2n+2}O$) are the same with the calculating results of graph theory.[29] The numbers of isomers of other molecules containing various elements such as O, N, S, etc., are the same as the results from the DENDRAL system.[30] The correctness of the structure generation algorithm in the DENDRAL system was rigorously proved,[31,32] and the results obtained from the DENDRAL system have been of value in varying the correct implementation of other subsequently developed exhaustive and irredundant algorithms.[33] Therefore, the results of our structure generator are exhaustive and irredundant.

Our algorithm is also very effective for generating various kinds of molecular structures. For example, the CPU times for the generation of the 910 726 isomers of $C_{21}H_{44}$ was about 15 h on a SUN SPARK 1102 workstation, and 2 min and 20 s were needed for generating the 204 isomers of C_8.

4. ARE THE EAID NUMBERS UNIQUE?

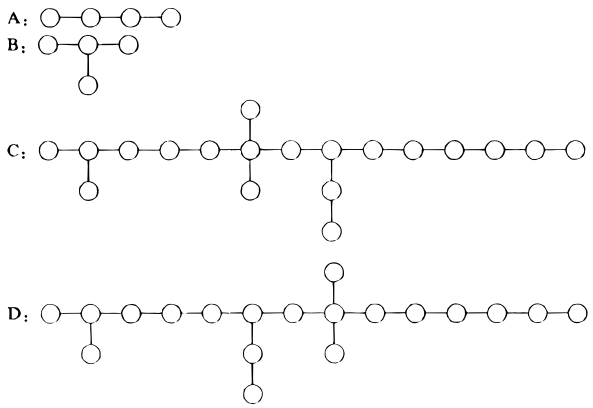
We have examined over 4 000 000 structures and have not detected one pair of isomers with an identical EAID

Chart 1



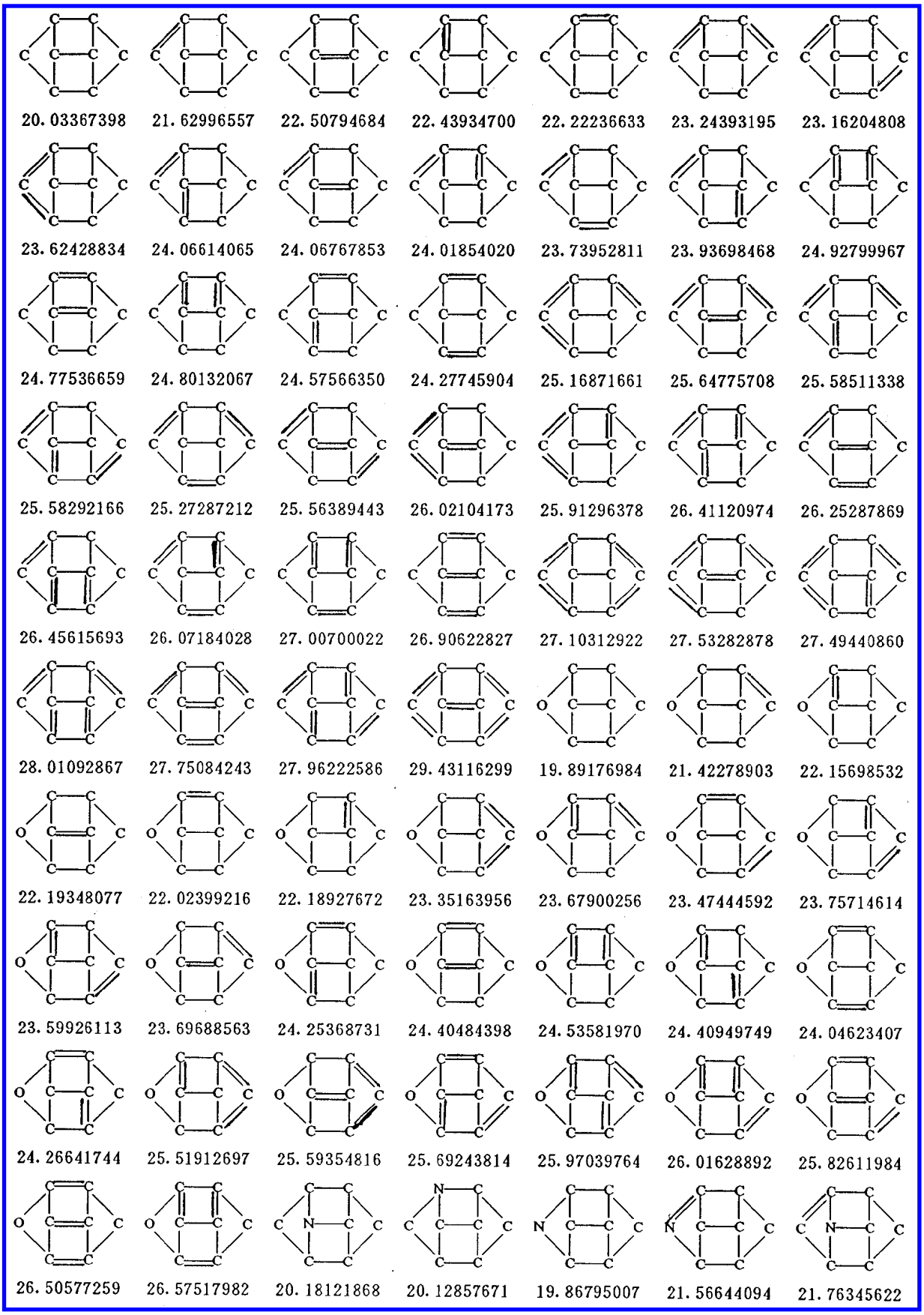
number. The families of those structures are (i) alkane trees, (ii) cyclic and polycyclic structures, and (iii) structures containing heteroatoms.

4.1. Alkane Trees. In this paper, only constitutional alkanes are included, and the steric hindrance is not considered. The acyclic alkanes are trees with no vertex having a degree higher than 4. It is a simple group in the chemical graph family, and the isomers of small molecules can even be listed by hand. But the number of isomers increases rapidly in this series, for example, while there are only 35 isomers of $C_{10}H_{22}$, the number of isomers of $C_{20}H_{42}$ is 366 319, and the number of isomers of $C_{22}H_{46}$ is more than 2 000 000 (see Table 2); the structural difference between the isomers with the same number of carbon atoms become more and more subtle with increasing number of atoms, e.g., the dissimilarity between structures A and B is much greater than that between structures C and D. Thus the early molecular topological indices can only discriminate the alkane trees of small molecules.[34]



To test the uniqueness of the EAID numbers for alkane trees, the molecular formula C_nH_{2n+2} (n = 1, ..., 22) was supplied to our isomer generator to produce all members of the alkane tree family with up to 22 atoms, and a total of 3 807 434 isomers was obtained (Table 2 shows the molecular formulas and their isomer numbers). We have examined all the EAID numbers of those 3 807 434 isomers, and

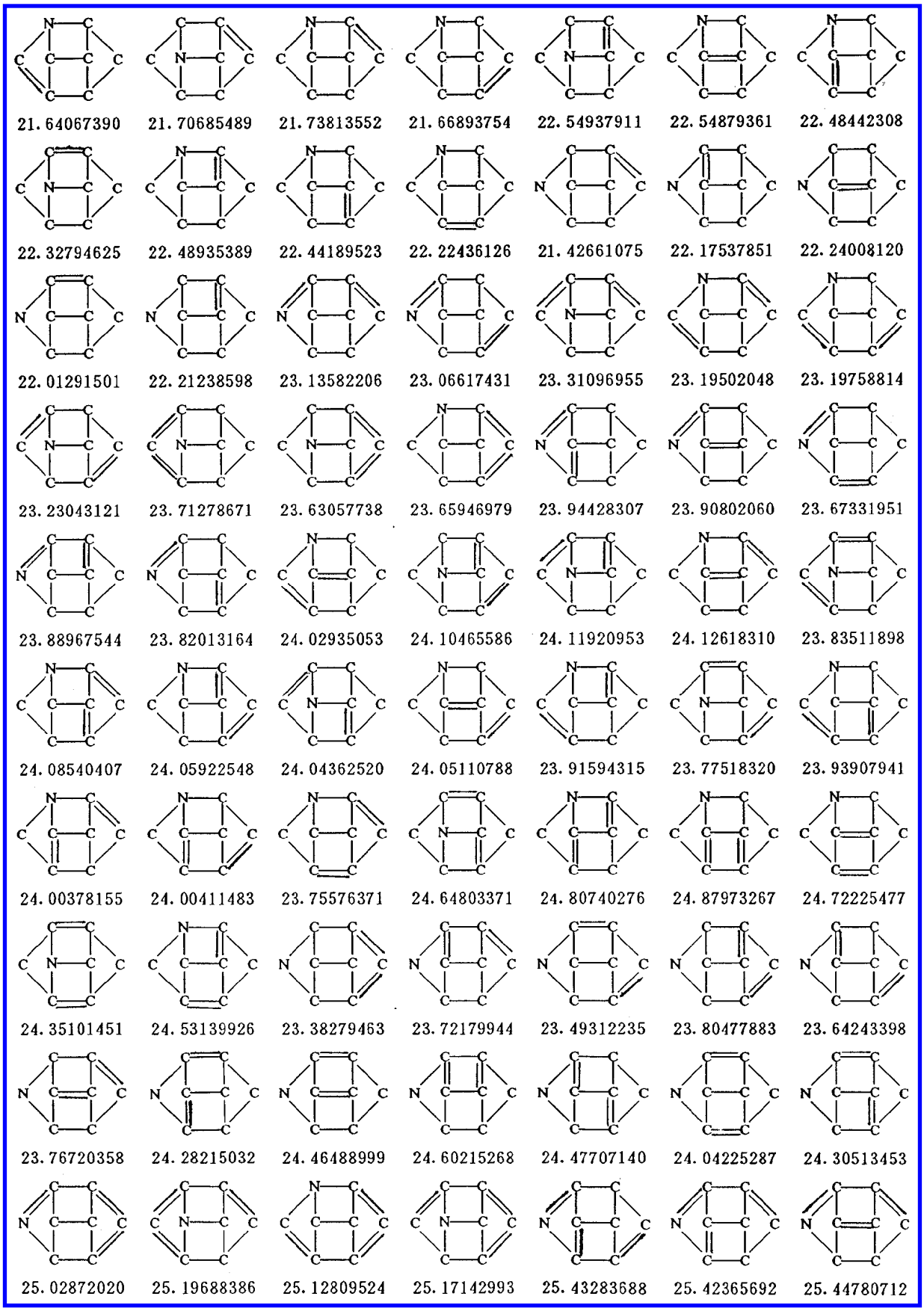
Chart 2



no duplicate with identical EAID number occurred. This is a hitherto unheard of power of discrimination.

4.2. Cyclic or Polycyclic Structures. Cyclic or polycyclic structures represent a more challenging group. To observe the discrimination of the EAID numbers for the cyclic or polycyclic structures, we have collected an exhaustive list of cyclic or polycyclic structures with up 12 atoms whose molecular connectivity valence (δ) value is always

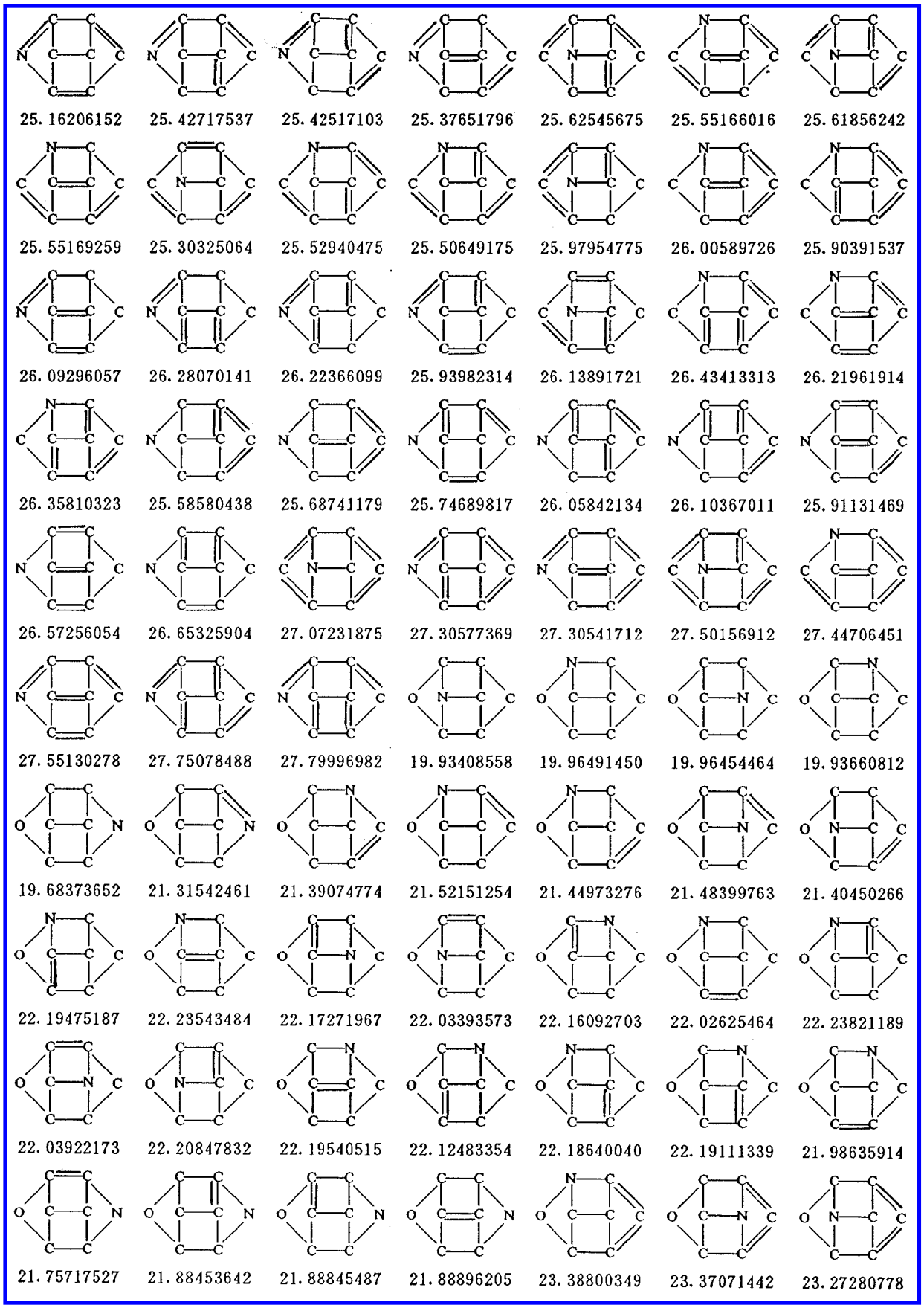
Chart 3



4, e.g., those structures are exhaustively generated from the molecular formulas C_n (n = 3, ..., 12). Table 3 shows the molecular formulas and their isomer numbers.

All those structures are complex cyclic or polycyclic structures, and a great many of them have multiple bonds. Some of those structures are highly regular graphs, for example, there are 1849 structures having the same vertex >C<, and 461 structures having the same vertex >C=. Chart 1 shows some of those regular graphs and their EAID numbers. Some of those structures have the same graph core

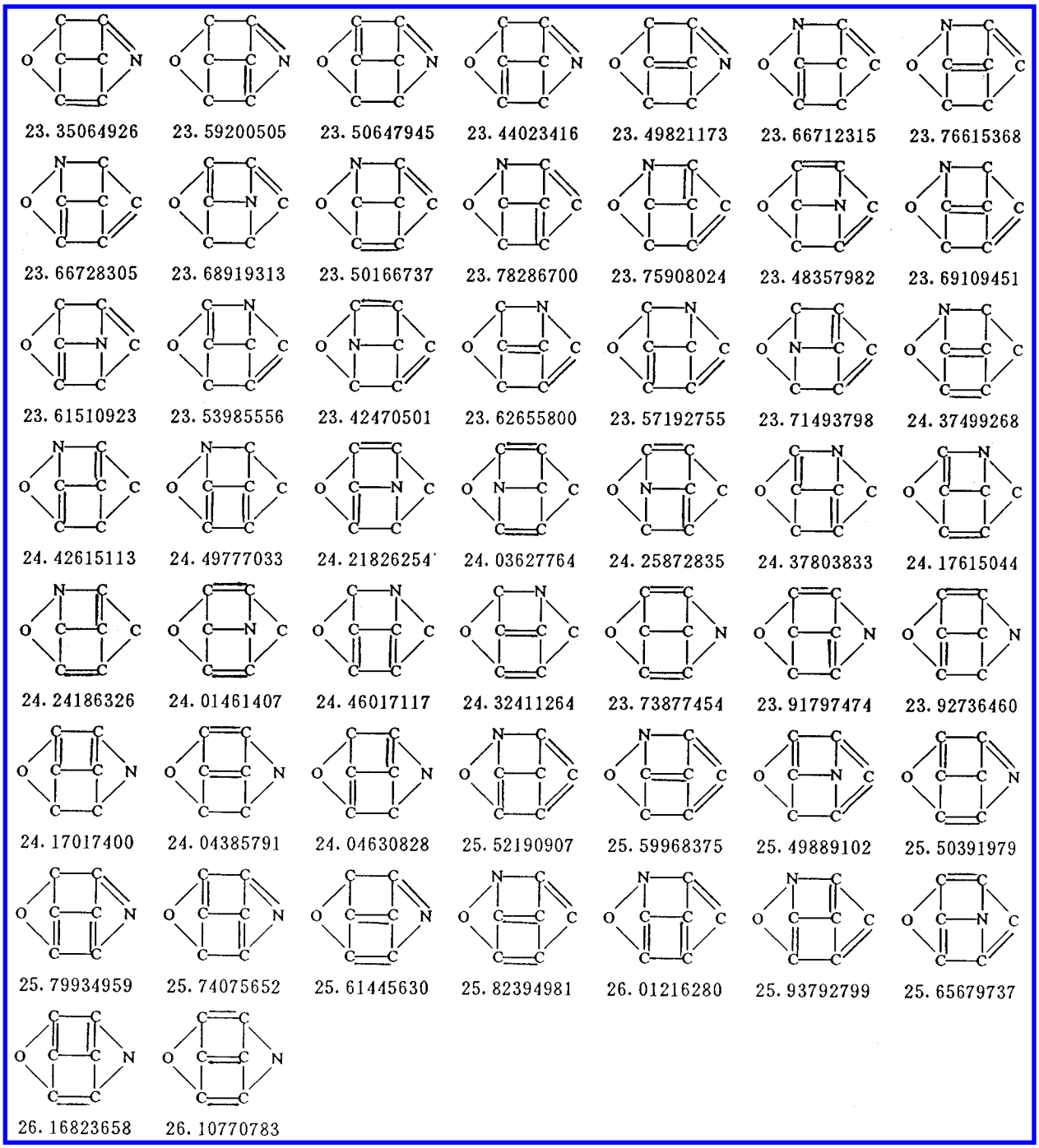
Chart 4



(reduced from colored graphs by ignoring the nodes and edges properties), e.g., the ten structures in lines 2 and 3 of Chart 1 have the same graph core.

We have calculated and compared all the EAID numbers of those structures, and no counterexample was found. It is evident that the EAID is a highly discriminatory topological

Chart 5



index for the cyclic or polycyclic structures.

4.3. Structures Containing Heteroatoms. Because more than 80% of the known molecular structures contain heteroatoms, it is a much more important task to process structural information on heteroatoms for graph theoretical approaches. In order to test the uniqueness of the EAID numbers for the molecules with heteroatoms, an exhaustive set of totally 430 472 structures having $N = 8$ atoms with up to one oxygen or/and one nitrogen heteroatoms were selected. Those structures are exhaustively generated from the following molecular formulas

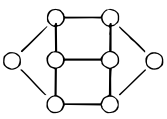
(1) C_8H_n $n = 18,16,14,12,10,8,6,4,2,0$

(2) C_7H_nO $n = 16,14,12,10,8,6,4,2,0$

(3) C_7H_nN $n = 17,15,13,11,9,7,5,3,1$

(4) C_6H_nNO $n = 15,13,11,9,7,5,3,1$

The molecular formulas and their isomer numbers are listed in Table 4. Those structures include various types of structures such as saturated and unsaturated structures or acyclic, cyclic, and polycyclic structures. Many of them have the same graph core with only different locations of heteroatoms or/and multiple bonds. For example, there are 261 structures with the same graph core



(these 261 structures and their EAID numbers are listed in Charts 2−5).

After calculating the EAID numbers of those 430 472 structures, no duplicate with identical EAID number was found. Thus the EAID numbers are highly discriminatory topological indices for structures containing heteroatoms.

All the algorithms are programmed in ANSI C and run on an IBM PC computer and SGI, SUN, SPARC workstations. All calculation are carried out with double precision.

5. CONCLUSION

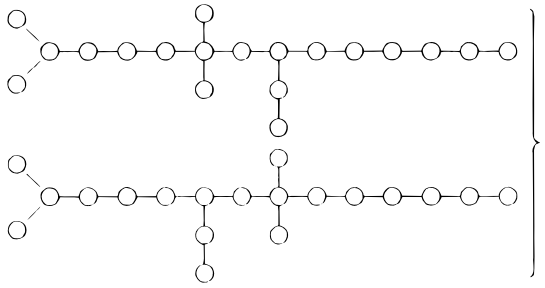
A new topological index, EAID, has been developed based on the power of the extended adjacency matrix. A systematic search for degeneracy was performed for various types of structures such as saturated and unsaturated isomers of acyclic, cyclic, and polycyclic structures, e.g., alkane trees, cyclic or polycyclic graphs, and structures containing heteroatoms. No degeneracies are found. That is a hitherto unheard of power of discrimination. Thus EAID might be possibly used as supplementary reference for CAS Registry Numbers for structure documentation.

ACKNOWLEDGMENT

The authors acknowledge the financial support of the National Natural Science Foundation of China and partial financial support of Yanbian Aodong Pharmaceutical Groups of Jilin province.

REFERENCES AND NOTES

(1) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc*. **1975**, *97*, 6609−6615.
(2) Xu, L.; Wang, H. Y.; Su, Q. A Newly Proposed Topological Index for Discrimination of cis/trans Isomers and for the Studies of QSAR/QSPR. *Comput. Chem*., **1992**, *16*, 187−194.
(3) Xu, L.; Wang, H. Y.; Su, Q. Correlation Analysis in Structure and Chromatographic Data of Organophosphorous Compounds by GAI. *Comput. Chem*. **1992**, *16*, 195−199.
(4) Yao, Y. Y.; Xu, L.; Yang, Y. Q.; Yuan, X. S. Studies on Structure−Activity Relationships of Organic Compounds—Three Topological Indexes and Their Applications. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 590−594.
(5) Yang, Y. Q.; Xu, L.; Hu, C. Y. Extended Adjacency Matrix Indices and Their Applications. *J. Chem. Inf. Comput. Sci.* **1994**, *34*(5), 1149−1145.
(6) Xu, L.; Yao, Y. Y.; Wang, H. M. New Topological Indices and Prediction of Phase Transfer Energy for Protorated Amines and Tetraalkylamines Ions. *J. Chem. Inf. Comput. Sci*. **1995**, *35*(*1*), 45−49.
(7) Xu, L. Molecular Topological Index a_N and its Extension. *J. Serb. Chem. Soc*. **1992**, *57*, 485−495.
(8) Balaban, A. T. Topological Index Based on Topological Distances in Molecular Graphs. *Pure Appl. Chem*. **1983**, *55*, 199−206.
(9) Clerc, J. T.; Terkovics, A. L. Versatile topological Structure Descriptor for quantitative structure/property studies. *Anal. Chim. Acta* **1990**, *235*, 93−102.
(10) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press: Lectchworth, Hertfordshire, England, 1986.
(11) Randić, M. Novel Graph Theoretical Approach to Heteroatoms in Quantitative Structure−Activity Relationships. *Chemom. Intell. Lab. System*. **1991**, *10*, 213−227.
(12) Rucker, G.; Rucker, C. Counts of All Walks and Molecular Descriptors. *J. Chem. Inf. Comput. Sci*. **1983**, *33*, 683−695.
(13) Burden, F. R. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Comput. Sci*. **1989**, *29*, 225−227.
(14) Randić, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci*. **1984**, *24*, 164−175.
(15) Randić, M. Molecular ID Numbers: by Design. *J. Chem. Inf. Comput. Sci*. **1986**, *26*, 134−136.
(16) Hu, C. Y.; Xu, L. A New Topological Index for CIAC-^{13}C NMR Information System. *Anal. Chem. Acta* Accepted for publication.
(17) Elk, S. B.; Gutman, I. Further properties derivable from the Matula numbers of an alkane. *J. Chem. Inf. Comput. Sci*. **1994**, *44*, 331−334.
(18) Balaban, A. T. Numerical Modelling of Chemical Structures: Local Graph Invariants and Topological Indices. in *Graph Theory and Topology*; King, R. B., Rouvray, D. H., Eds; Elsevier: Amsterdam, 1987; pp 159−176.
(19) Muller, W. R.; Szymanski, K.; Knop, J. V.; Mihalic, Z.; Trinajstic, N. The Walk ID Numbers Revisited. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 231−233.
(20) Note: According to our test, there are 32 pairs of counterexamples (two or more nonisomorphic structures with the identical SID number) in the nonadecane family, such as the following structure. So, the author's conclusion that SID is unique for alkane trees up to 20 vertices in ref 18 seems not to be correct.



SID = 19. 0054419987358970

(21) Hall, L. H., Kier, L. B. Determination of Topological Equivalence in Molecular Graphs from the Topological State. *Quant. Struct.-Act. Relat*. **1990**, *9*, 115−131.
(22) Hu, C. Y.; Xu, L. On Hall and Kier's Topological State and Total Topological Index. *J. Chem. Inf. Comput. Sci*. **1994**, *34*(*6*), 1251−1258.
(23) Hu, C. Y.; Xu, L. An New Algorithm for Computer Perception of Topological Symmetry. *Anal. Chim. Acta* **1994**, *295*, 127−134.
(24) Diudea, M. V.; Minailiuc, O.; Balaban, A. T. Molecular Topology. IV. Regressuve Vertex Degrees (New Graph Invariants) and Derived Topological Indices. *J. Comput. Chem*. **1991**, *12*, 527−535.
(25) Balaban, A. T.; Diudea, M. V. Real Number Vertex Invariants: Regressive Distance Sums and Related Topological Indices. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 421−428.
(26) Fessenden, R. J.; Fessenden, J. S. *Organic Chemistry*; Willard Grant Press: Boston, MA, 1979; pp 5−6.
(27) Hu, C. Y.; Xu, L. Expert System fo Elucidation of Structures of Organic Compounds. *Sci. China, Ser. B* **1995**, *38*(*3*), 296−304.
(28) Hu, C. Y.; Xu, L. Principles for Structure Generation of Organic Isomers from Molecular Formula. *Anal. Chim. Acta* **1994**, *298*, 75−85.
(29) Balaban, A. T. *Chemical Applications of Graph Theory*; Academic Press: 1976.
(30) Smith, D. H. *J. Chem. Inf. Comput. Sci*. **1975**, *15*, 203−206.
(31) Brown, H.; Hjelmeland, L., Masinter, L. M. *Discrete Math*. **1974**, *7*, 1−30.
(32) Brown, H.; Masinter, L. M. *Discrete Math*. **1974**, *8*, 227−244.
(33) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; Wiley: New York, 1986.
(34) 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.

CI9501150