JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Extension of a Highly Discriminating Topological Index

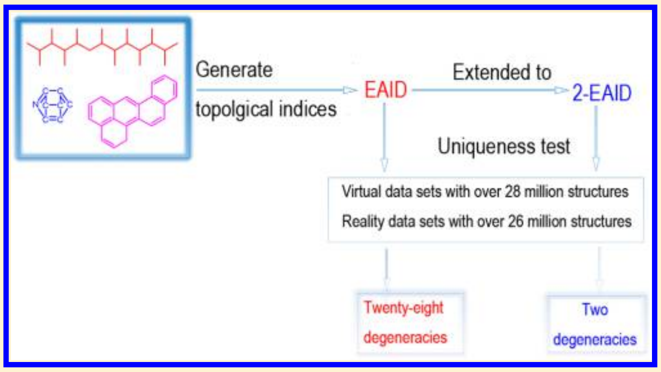
Qingyou Zhang,*[,†] Chengcheng Wu,[†] Fangfang Zheng,[†] Tanfeng Zhao,[†] Yanmei Zhou,[†] and Lu Xu*[,‡]

[†]Institute of Environmental and Analytical Sciences, College of Chemistry and Chemical Engineering, Henan University, Kaifeng 475004, China

[‡]Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, China

S *Supporting Information*

ABSTRACT: A highly discriminating topological index, EAID, is generated in our laboratory. A systematic search for degeneracy was performed on a total of over 14 million structures, and no duplicate occurred. These structures are as follows: over 3.8 million alkane trees with 1−22 carbon atoms; over 0.38 million structures containing heteroatoms; over 4 million benzenoids with 1−13 benzene rings; and over 5.9 million compounds from three reality databases. However, in a search of over 20 million alkane trees with 23 and 24 carbon atoms, five and 13 duplicates occurred, respectively, and for over 20 million compounds from the ZINC database, 10 duplicates occurred. To increase the discriminating power of the index, EAID has been extended, and the resulting index is termed 2-EAID. All of the over 55 million structures mentioned above were uniquely identified by 2-EAID except for two duplicates that occurred for the ZINC database. EAID and 2-EAID are the most highly discriminating indices examined to date. Thus, the two indices possess not only theoretical significance but also potential applications. For example, they could possibly be used as a supplementary reference for CAS Registry Numbers for structure documentation.


Generate
topolgical indices
EAID
Extended to
2-EAID
Uniqueness test
Virtual data sets with over 28 million structures
Reality data sets with over 26 million structures
Twenty-eight degeneracies
Two degeneracies

A graph theoretical (topological) index generated from a molecular structure is a mathematical representation that is not dependent on the sequence number of the nodes in that structure. The main aims for this kind of research are to provide codes (i.e., independent variables x) for quantitative structure−activity/property relationship (QSAR/QSPR) studies[1−4] and for information administration. The two distinct roles are not necessarily compatible, as the emphases of the two roles are different. For example, the Wiener index has been used for the paraffin boiling points to get satisfactory results,[5] but the uniqueness (selectivity) is not sufficient.

It should be noted that in recent years another role for molecular invariants has appeared, as outlined in a recent paper by Randić,[6] which speaks of molecular descriptors for searching combinatorial libraries to find structures that are most similar to a target structure. This is an application in SAR and QSAR but is different from the traditional use of molecular descriptors in multiple regression analysis and artificial neural networks.

For information administration, a great deal of effort has been made to search for a highly selective index to uniquely label compounds such as acyclic alkanes. The acyclic alkanes are trees in graphic theory (termed "alkane trees") with no vertex having a degree higher than 4. The molecular identification (ID) number suggested by Randić is defined as the sum of all weighted paths in a molecule. The uniqueness determination revealed that the smallest degenerate alkane trees with the same ID numbers occurred for n = 15 carbon atoms.[7] By the assignment of prime-number weight values to various bond types, the revised prime ID was uniquely tested for all alkane trees up to 19 vertices.[8] Balaban[9] replaced the vertex degree in the original ID formula by the distance sum, resulting in the Balaban ID (BID) index, for which the selectivity was determined for all alkane trees containing up to 20 vertices. The uniqueness observation for the topological index τ constructed by Hall and Kier[10] showed that five degenerations occurred for n = 20 carbon atoms. Hu and Xu[11] used another different formula to weight the paths and developed a new topological index that was unique for all alkane trees up to at least 20 vertices. Subsequently, Hu and Xu[12] derived a topological index called EAID that is based on weighting of edges in a structure graph and involves layer matrices and powers of an extended adjacency matrix, and 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 (i.e., two or more nonisomorphic structures with the same EAID number) were found.[12] The relevant studies have not been active for nearly two decades because no better topological indices than EAID have been found, i.e., EAID is the most highly discriminating index tested to date.[13]

Investigating the discriminating ability among different kinds of isomers is still a challenging and ongoing problem. As we know, the number of isomers increases rapidly with increasing

Received: January 26, 2015
Published: June 3, 2015

DOI: 10.1021/acs.jcim.5b00044

Table 1. Structure Fragment Set in the Structure Generator

no.	1	2	3	4	5	6	7	8	9	10	11
fragment	CH_3-	$-CH_2-$	$CH_2=$	−CH<	−CH=	CH≡	>C<	>C=	−C≡	=C=	−OH
no.	12	13	14	15	16	17	18	19	20	21	22
fragment	−O−	O=	SH−	−S−	S=	NH_2-	−NH−	NH=	−N<	−N=	N≡
no.	23	24	25	26	27	28	29	30	31	32	33
fragment	PH_2-	−PH−	−P<	SiH_3-	$-SiH_2-$	>SiH−	>Si<	F−	Cl−	Br−	I−

number of carbon atoms for alkane tree systems. For example, the number of isomers of $C_{22}H_{46}$ is 2 278 658, while that for $C_{23}H_{48}$ is 5 731 580, which is ca. 2.5 times larger. If one more carbon atom is added ($C_{24}H_{50}$), the number of isomers increases to 14 490 245, which again is ca. 2.5 times larger. At the same time, the structural differences between the isomers with the same number of carbon atoms become more and more subtle with increasing number of atoms. Thus, uniquely discriminating alkane trees by a single number becomes more and more difficult.

As mentioned above, over 3.8 million alkane trees with 1−22 carbon atoms were tested by us for the uniqueness of EAID (14 significant digits herein) without degeneracy. To face the more huge challenge, the authors individually examined the discriminating powers of EAID for alkane trees with 23 and 24 carbon atoms, and we found that five degeneracies occurred with 23 carbon atoms (over 5.73 million alkane trees) and 13 degeneracies occurred with 24 carbon atoms (over 14.49 million alkane trees). Thus, after that, EAID was extended. The extended EAID is called 2-EAID. The discriminating power of the extended EAID is obviously increased compared with the original EAID.

To examine the discriminating power of 2-EAID, a total of over 55 million structures were uniquely identified except for two degeneracies. These structures are as follows: (1) over 24 million alkane trees with 1−24 carbon atoms; (2) over 0.38 million compounds containing eight non-hydrogen atoms, including one or two heteroatoms, such as C_7N and C_7O; (3) over 4 million benzenoids with 1−13 benzene rings; and (4) over 26.6 million different kinds of compounds from four reality databases. The high discriminating power of a topological index possesses not only theoretical significance but also practical potential applications, such as administration of large compound databases, affirming a proposed new compound to be really a new one, and so forth.

1. METHODS

1.1. Structure Generator.[14,15]

The structures included in item (4) in the above list were obtained from open databases. However, the structures included in items (1) and (2) above were generated using a structure generator developed by the author's laboratory to construct the virtual database in this research. Thus, herein we introduce chiefly our program system.

In general, a structure generator can exhaustively generate the isomers without redundancy once it has accepted a molecular formula. The program should meet needs such as speediness, efficiency, the ability to accept some constraint conditions, and so on. The structures, say all of the alkane trees with 20 carbon atoms, could be generated in a short time using our structure generator. For example, the generation of 366 319 alkanes with 20 carbon atoms took ca. 20 min using an 3.9 GHz AMD A8 6600K processor and 8 GB of RAM on a 64-bit Windows XP operating system. The major parts of this structure generator are as follows.

(1) Structure Fragment Set. The structure fragment set in the structure generator (see Table 1) contains 33 structure fragments generated from the chemical elements in organic compounds. These elements are C, H, O, N, S, P, Si, F, Cl, Br, and I. For example, the valence of carbon is 4. For bonding of hydrogen and carbon atoms, the basic structure units are CH_3, CH_2, CH, and C. Then continuous addition of single, double, and triple bonds with those basic structure units generates the structure fragments >C<, >C=, =C=, −C≡, >CH−, =CH−, CH≡, $-CH_2-$, $=CH_2$, and $-CH_3$. The fragments for other elements are similar to those for carbon.

(2) Exhaustive Generation of Two-Dimensional Isomers. The strategy to generate a structure with the structure fragment set usually adopts depth-first to traverse the spanning tree. The algorithms can be divided roughly into two types: direct extension of substructures or filling the connectivity matrix. A method that involves filling the connectivity matrix was used in this research.

(3) Exhaustive Generation of Stereomers.[16,17] On the basis of the two-dimensional isomers generated in step (2), the determination of stereocenters by using topological equivalence was performed. Then the isomorphic group and the automorphism group were generated, and an analysis for stereomers was carried out. Through the above procedure, the stereomers of a two-dimensional isomer could be obtained. For example, hexachlorocyclohexane is a highly symmetrical molecule possessing nine stereomers, which were obtained using the program of exhaustive generation of stereomers in the structure generator (see Figure 1). The bottom two molecules are a pair of enantiomorphs.

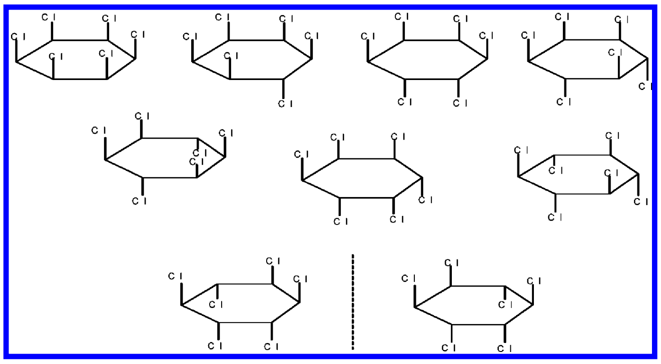


Figure 1. Stereomers of hexachlorocyclohexane.

1.2. The Topological Index EAID.[12]

Although the formula for EAID has been reported previously,[12] for the sake of clarity we briefly introduce the EAID algorithm again in this article.

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. Each distinct atom is

Table 2. Numbers of Alkane Trees (N) of Formula C_nH_{2n+2}

n	N	n	N	n	N	n	N	n	N	n	N
1	1	5	3	9	35	13	802	17	24894	21	910726
2	1	6	5	10	75	14	1858	18	60523	22	2278658
3	1	7	9	11	159	15	4347	19	148284	23	5731580
4	2	8	18	12	355	16	10359	20	366319	24	14490245

characterized by its covalent radius and its connectivity valence δ, which is similar to δ^v used by Kier and Hall.[18] The types of bonds are single, double, triple, and aromatic, whose δ values are coded as 1, 2, 3, and 1.5, respectively. The algorithm can be described by the following steps:

(1) Set the weight $S[i]$ of atom i. From the viewpoint of node i, the immediate neighbors of 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. Thus, a connectivity valence matrix (**CVM**) is proposed, whose ijth element $(\text{cvm})_{ij}$ is defined as the sum of the molecular connectivity valences (the δ values) for all of the nodes situated in the jth layer around node i. The bond matrix (**B**) is proposed, whose ijth element b_{ij} is defined as the sum of the codes of the bonds connecting the nodes of the jth layer and the nodes of the $(j-1)$th layer around node i. The weight $S[i]$ of node i is then calculated using the following function:

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

where K is the number of layers from the viewpoint of node i.

(2) Set up the adjacency matrix **A**, whose elements a_{ij} are obtained from the bond orders:

$$a_{ij} = \begin{cases} 0 & \text{atoms } i \text{ and } j \text{ are not bonded } (d_{ij} > 1) \\ 1 & \text{single bond between atoms } i \text{ and } j \\ 2 & \text{double bond between atoms } i \text{ and } j \\ 3 & \text{triple bond between atoms } i \text{ and } j \\ 1.5 & \text{aromatic bond between atoms } i \text{ and } j \end{cases} \tag{2}$$

where d_{ij} is the topological distance between atoms i and j.

(3) Set up the extended adjacency matrix **EA**, whose elements $(\text{ea})_{ij}$ are computed as follows:

$$(\text{ea})_{ij} = \begin{cases} \dfrac{\sqrt{r_i}}{6} & i = j \\ \dfrac{w_{ij}\sqrt{a_{ij}}}{6} & i \neq j \end{cases} \tag{3}$$

where r_i is the covalent radius (in Å) of the ith atom and w_{ij} is a weight factor calculated as

$$w_{ij} = \sqrt{\frac{S[i]}{S[j]}} + \sqrt{\frac{S[j]}{S[i]}} \tag{4}$$

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

(4) Evaluate a new matrix $\mathbf{EA}^* = \{(\text{ea})_{ij}\}$, defined as the sum of powers of the **EA** matrix:

$$\mathbf{EA}^* = \sum_{k=0}^{N-1} (\mathbf{EA})^k \tag{5}$$

where N is the number of atoms in the molecule. The matrix $(\mathbf{EA})^0$ (obtained when $k = 0$) is the identity matrix.

(5) Calculate the topological index EAID as follows:

$$\text{EAID} = \sum_{i=1}^{N} (\text{ea}^*)_{ii} \tag{6}$$

1.3. Extension of EAID To Obtain 2-EAID. The extention of EAID to obtain the topological index 2-EAID proceeds as follows. When the topological distance between two atoms, d_{ij}, is larger than 1 (i.e., atoms i and j are not directly connected), the value 0 is assigned to the element of adjacency matrix, a_{ij} (eq 2). In this case, if $i \neq j$, the extended adjacency matrix element $(\text{ea})_{ij}$ generated from the two atoms i and j is equal to zero in the EAID algorithm (eq 3). In order to improve the discrimination ability of EAID, these zero-valued matrix elements $(\text{ea})_{ij}$ are replaced by a distance factor derived from d_{ij} according to eq 3a:

$$(\text{ea})_{ij} = \begin{cases} \dfrac{\sqrt{r_i}}{18} & i = j \\ \dfrac{w_{ij}\sqrt{a_{ij}}}{18} & i \text{ and } j \text{ are bonded} \\ \dfrac{2^{-d_{ij}/5} w_{ij}}{21} & i \text{ and } j \text{ are not bonded } (d_{ij} > 1) \end{cases} \tag{3a}$$

The basic principle is that the larger the distance between two atoms i and j is, the smaller is $(\text{ea})_{ij}$.

In addition, in order to incorporate all of the information in the matrix **EA***, instead of the sum of diagonal elements as used for EAID (eq 6), all of the elements of **EA*** are summed to obtain 2-EAID (eq 6a):

$$\text{2-EAID} = \sum_{i=1}^{N} \sum_{j=1}^{N} (\text{ea}^*)_{ij} \tag{6a}$$

The other steps in the calculation of 2-EAID are the same as for EAID. The discriminating power of 2-EAID calculated using the above formulas is obviously increased compared with EAID (see below).

2. DATA SETS

We employed data in two different kinds of data sets: virtual data sets and reality data sets. The data included in the virtual data sets were generated by our structure generator as mentioned above. The data included in the reality data sets were obtained from the NCI, PubChem, AKos, and ZINC databases.

2.1. Virtual Data Sets. *(1). Alkane Trees.* In this research, only constitutional alkane trees (without considering H atoms) are included, wherever steric hindrance is not considered. The degree of a vertex in those trees is 4. Over 24 million alkane trees with 1–24 carbon atoms were generated using our structure generator (see Table 2).

(2). Structures with Eight Non-Hydrogen Atoms, Including One or Two Heteroatoms. This system includes the molecular formulas C_6NO, C_7N, and C_7O, which contain heteroatoms. The isomers of these formulas were also exhaustively generated using our structure generator. The numbers of generated structures for C_6NO, C_7N, and C_7O are 207 632, 123 247, and 55 063, respectively, giving a total of 385 942 compounds containing heteroatoms.

(3). Benzenoids. These structures are composed of benzene rings connected in parallel, having no empty inside.[19] Two examples of the benzenoids composed of 11 benzene rings are shown in Figure 2. Over 4 million isomers with 1−13 benzene rings were generated using the program developed by us.[19,20] The numbers of isomers corresponding to the various numbers of benzene rings are shown in Table 3.

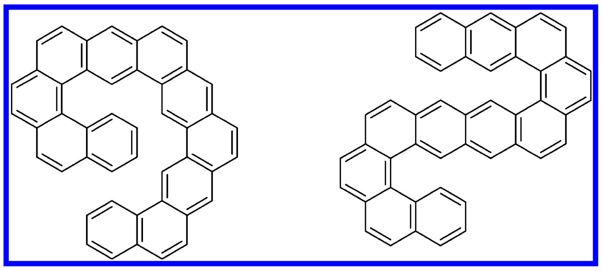


Figure 2. Two benzenoids composed of 11 benzene rings.

Table 3. Numbers of Benzenoids Composed of 1−13 Benzene Rings

no. of benzene rings	no. of benzenoids
1	1
2	1
3	3
4	7
5	22
6	81
7	331
8	1435
9	6505
10	30086
11	141229
12	669584
13	3198256

2.2. Reality Data Sets. *(1). NCI Database Information System.* A total of 265 242 compounds were obtained from the U.S. National Cancer Institute (NCI) database. An entry for a compound in this database contains the NCI and CAS registration numbers, IUPAC nomenclature of organic chemistry, and structure of the compound and a list of some properties. This information system was updated in May 2012.

(2). PubChem Database Information System. A total of 461 937 compounds were taken from the PubChem Database System. This system is a biological activity database of small organic molecules, which was updated in September 2011. It is maintained by the U.S. National Center for Biotechnology Information.

(3). AKos Commercial Database Information System. A total of 5 429 833 compounds were downloaded from commercially available libraries of AKos.[21]

(4). ZINC Database. A total of 47 329 414 compounds were taken from the standard version of the ZINC database in March 2015. It is a free database of commercially available compounds for virtual screening. The standard version is composed of five subsets, namely, “Lead-Like”, “Fragment-Like”, “Drug-Like”, “All Purchasable” and “Shards”. The structures appearing in one subset may possibly exist in another subset, i.e., some compounds are repeated.

The four reality data sets are structurally diverse, including structures that are saturated, unsaturated, chain, cyclic, heterocyclic, polycyclic, bridge-bonded, and so on. Several molecules contained more than 300 atoms.

As the sources of the four data sets were different, prior to use the files included in those data sets were pretreated. The processes in this paper were as follows: (1) SDfile was adopted as the download format; (2) the compound structures were normalized (e.g., by adding H atoms or aromatizing) using the Standardize program supported by JChem (ChemAxon Company); (3) “compounds” including more than one molecule, nonconnected molecules, mixtures, salt compounds, etc., were deleted; (4) for the fourth database (ZINC), repetitions of a compounds in different subsets were deleted, and cis/trans isomers and chiral centers were ignored. Finally, the data sets taken from the NCI, PubChem, AKos, and ZINC databases contained 212 197, 444 513, 5 244 425, and 20 738 250 compounds, respectively, for a total of ca. 26.64 million compounds.

3. RESULTS AND DISCUSSION

The discriminating powers of EAID and 2-EAID were determined for all of the data sets introduced above. The results of uniqueness tests and the potential applications of EAID and 2-EAID will be given in the following sections.

3.1. Uniqueness Test for Alkane Trees. As mentioned above, the structural differences among the isomers with the same number of carbon atoms become more and more subtle as the number of carbon atoms in the alkane tree increases. Figure 3 shows several isomers of $C_{23}H_{48}$. From this figure we can see that the index value decreases with the leftward movement of the isopropyl side chain, but the differences between the four isomers are quite small. Therefore, it is a huge challenge for the selectivity of a new topological index. As previously stated, the discriminating powers of the EAID topological index have been examined for over 3.8 million alkane trees with 1−22 carbons, and no duplicate with an identical EAID number occurred. However, for alkane trees

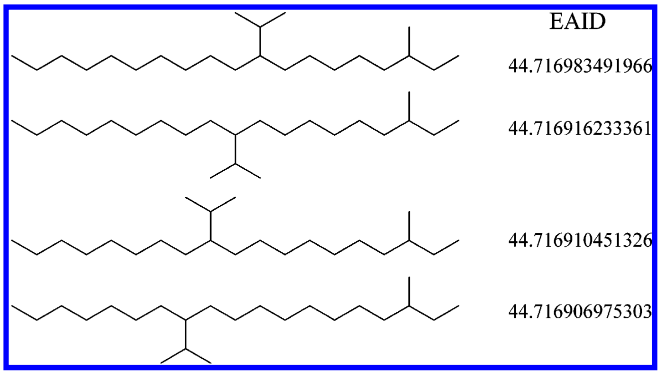


Figure 3. Some isomers of $C_{23}H_{48}$ and their corresponding EAID values.

with 23 and 24 carbon atoms, five and 13 degeneracies appeared, respectively.

To make clear the cause of degeneracy, we examined the diagonal elements of the **EA*** matrix from the calculation of EAID, $(ea^*)_{ii}$, and found that the corresponding elements in the two degenerate isomers were not equal. Moreover, degeneracies occur only when the number of alkane trees is very large. This means that in this case the number of significant digits of the EAID index is too small relative to the number of alkane trees, resulting in chance overlap. Thus, if the number of significant digits of EAID index could be extended, the degeneracy would be avoided. Figure 4 shows the five pairs of $C_{23}H_{48}$ isomers that cannot be distinguished by EAID. It is apparent that the difference between the 2-EAID values for each pair of isomers is obviously larger than the corresponding EAID. That is to say, the discriminating power of 2-EAID has been increased greatly compared with EAID. We have examined the discriminating powers of the index 2-EAID for $C_{23}H_{48}$ (over 5.7 million alkane trees) and $C_{24}H_{50}$ (over 14.4 million alkane trees), and no degeneracies occurred.

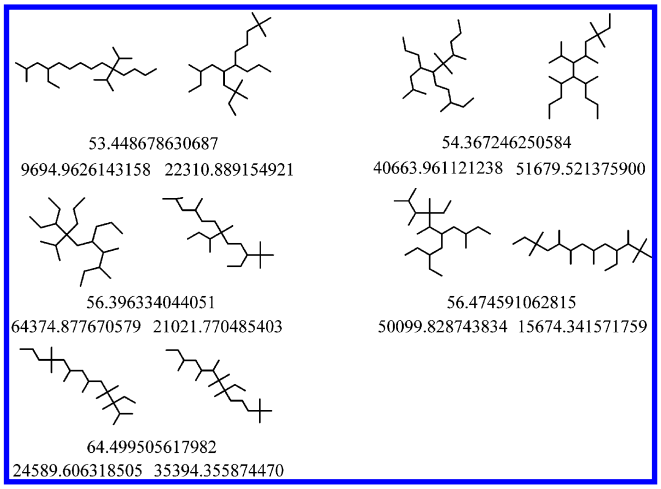


Figure 4. Five pairs of isomers of $C_{23}H_{48}$ that cannot be distinguished by EAID. Under each structure, the first line gives the EAID value and the second line the two 2-EAID values.

3.2. Uniqueness Test for Structures Containing Heteroatoms.

The formulas C_6NO, C_7N, and C_7O contain eight non-hydrogen atoms, and the valences for C, O, and N are 4, 2 and 3, respectively. Hydrogen atoms are added for saturation during the generation of a structure. A total of 385 942 cyclic isomers were generated from these formulas using our structure generator, and the discriminating powers of EAID and 2-EAID for the isomers derived from the different formulas above were examined. All of those isomers were identified uniquely with no degeneracy.

3.3. Uniqueness Test for Benzenoids.

The structures of many benzenoids are similar. For example, we have proved that all 141 229 structures of benzenoids composed of 11 benzene rings can be put into an isosceles trapezoid as shown in Figure 5.[19] The lengths of the bases and legs can be described by numbers of benzene rings. In Figure 5, the lengths of the two bases are 11 and five benzene rings, while the lengths of both legs are seven benzene rings. Three benzenoids composed of 11 rings are also illustrated separately in Figure 5 in green, blue, and red.

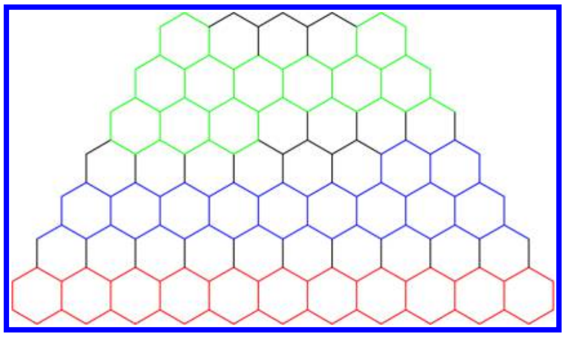


Figure 5. Region in which all of the benzenoids composed of 11 benzene rings can be enumerated.

All of the benzenoids composed of less than or equal to 13 benzene rings were tested using both EAID and 2-EAID, and no duplicate values were found.

3.4. Uniqueness Test for the Reality Data Sets.

As mentioned above, the total of over 5.9 million compounds in the first three data sets taken from reality databases have high species diversity, such as alkanes, aromatic rings, condensed-nucleus molecules, heterocycles, large molecules possessing a few hundreds atoms, and so on. The uniqueness examinations of EAID for those databases were performed, and no degeneracy was found to occur. Examples of EAID uniqueness examination for condensed-nucleus molecules, condensed-nucleus molecules linked with a long-chain alkane, very complicated large rings containing fused rings, and endo compounds are shown in Figures 6, 7, 8 and 9, respectively. Similarly, the discriminating power of 2-EAID for the same >5.9 million compounds was determined, and no degeneracy occurred. Examples of more complicated structures in the reality data sets are given in Figure S1 in the Supporting Information.

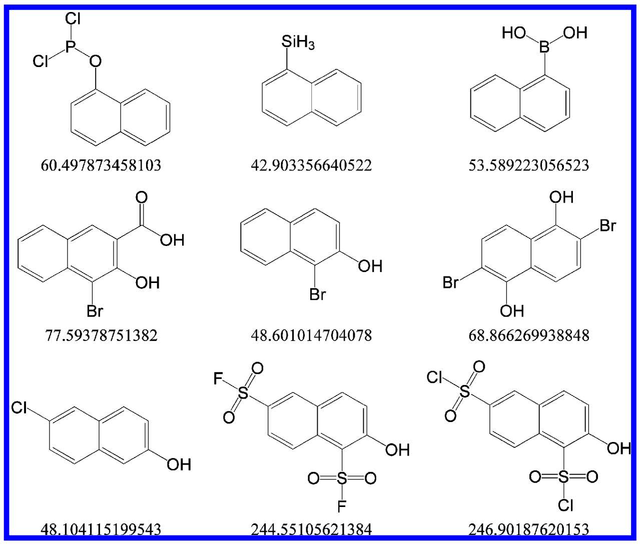


Figure 6. Examples of EAID uniqueness examination for condensed-nucleus compounds.

In addition, the discriminating examinations of 20 738 250 compounds in the fourth reality data set, taken from the ZINC database, were performed, and there were 10 duplications for EAID and two duplications for 2-EAID (see Figure 10). We noticed that again the two duplicated molecules in each case are

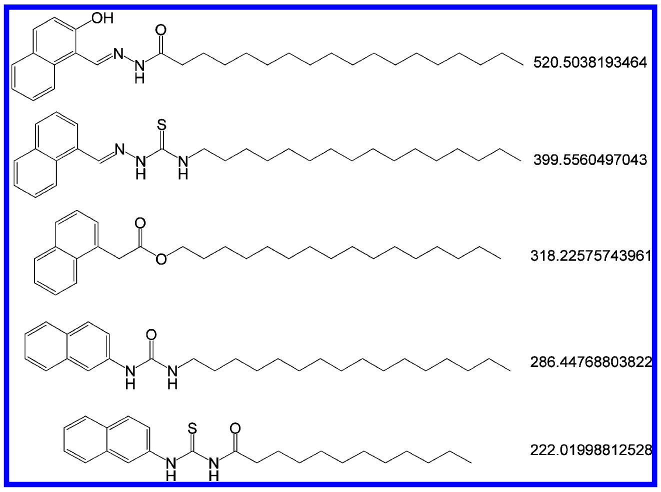


Figure 7. Examples of EAID uniqueness examination for condensed-nucleus compounds linked with a long-chain alkane.

fairly dissimilar, and the degeneracy may be caused by chance overlap as mentioned above.

In summary, the number of structures examined with 2-EAID was more than 55 million, and only two duplications occurred (the number of substances registered by CAS was more than 91 million as of Jan 16, 2015, and the CAS Registry is updated daily with about 15 000 substances).[22] Thus, the high discriminating powers possessed by EAID and 2-EAID have not only theoretical significances but also practical potential applications.

3.5. On Applications. In some research work, such as the administration of a large compound database, expert systems for evaluating structures of organic compounds, and computer-aided synthesis, it is always necessary to uniquely identify a compound that is new or already existing. But this is not easy because a compound's name, molecular formula, molecular weight, and so forth are not unique. For this purpose, the two-dimensional connectivity matrix of a structure could be used. However, the code is dependent on the atomic ordering, i.e., if the atomic ordering is not the same, the code will be different. Because of this problem, Morgan[23] suggested a method called Morgan coding to get canonical numbers of the atoms in a structure, and later this method was extended by the others.[24−26]

The canonical connectivity matrix can be obtained on the basis of the canonical numbers of atoms achieved using Morgan's method, and then a character string is formed by connecting a row and another row from the beginning to the end of that matrix for further use. One could imagine that the generated string, termed the "connectivity stack", is quite long for a large molecule. Compared with the stack, the 2-EAID index is a single number, i.e., the computation to affirm a new compound is of a sequence comparison, which is much simpler. Thus, the 2-EAID topological index could be used for identifying new compounds. Such a function would be very useful for a very large compound database. Even though degeneracies may occur, the numbers of duplicated compounds should not be high, and thus, further treatment would be simple. In the administration of a large compound database, one could use the EAID number as a field to generate the indexed file when constructing a compound database, and searching for a structure would be fast and the result would be unique. In contrast, when the molecular formula or molecular weight is taken as a key word, the result is often not ideal. As the databases become larger and larger, the speed of retrieving compounds should be more and more important. In this case,

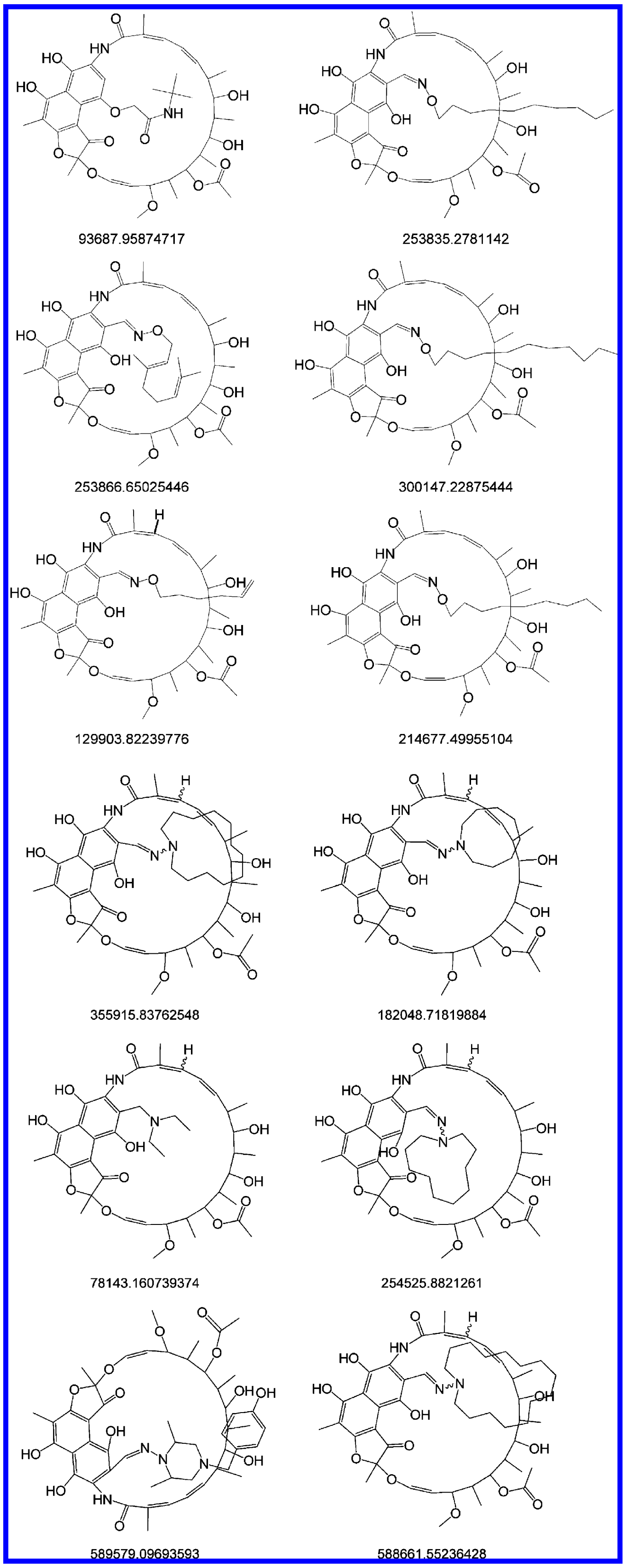


Figure 8. Examples of EAID uniqueness examination for very complicated large ring compounds, each containing a condensed nucleus.

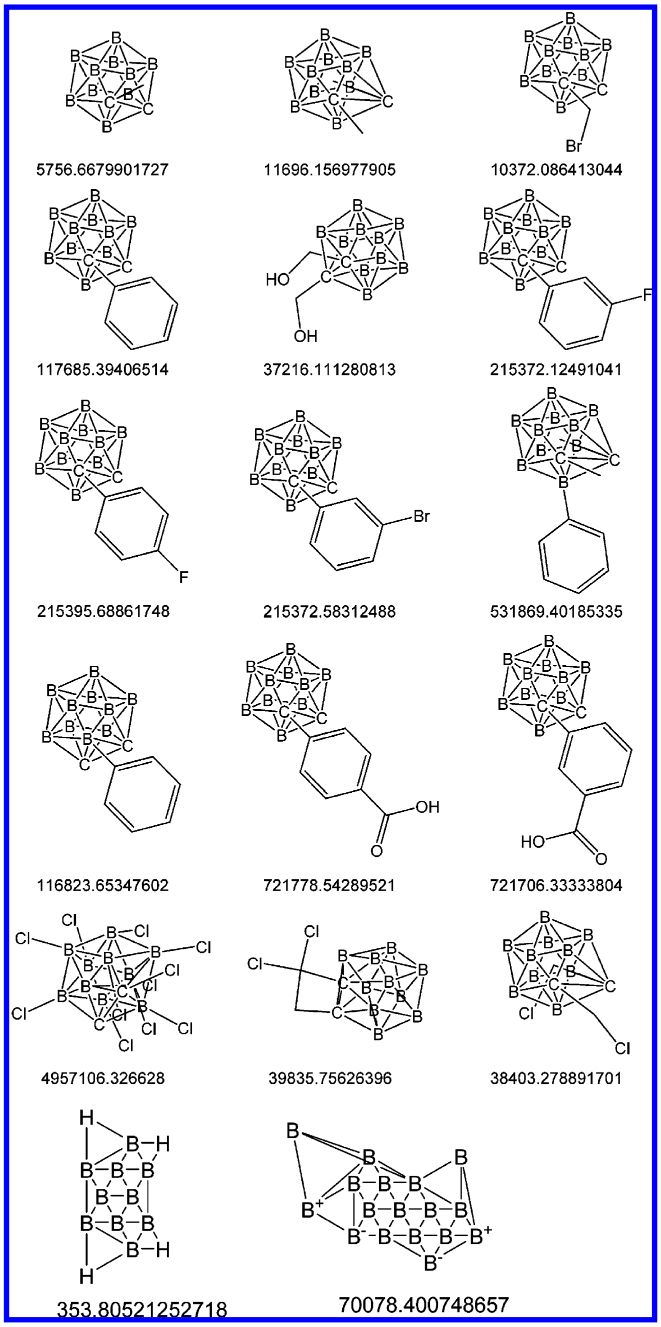


Figure 9. Examples of EAID uniqueness examination for boron compounds.

the use of a single number, such as a highly discriminating topological index, should be faster than the other techniques.

4. CONCLUSIONS

The uniqueness of the topological index EAID for over 3.8 million alkane trees with 1−22 carbon atoms was examined, and no duplicate with an identical EAID number occurred; for over 0.38 million structures containing heteroatoms, no duplicate occurred; and for 5.9 million compounds possessing high diversity from three reality databases, no duplicate occurred. However, for the 5 731 580 alkane trees with 23 carbon atoms, five duplicates occurred, while for the 14 490 245 alkane trees with 24 carbon atoms, 13 duplicates occurred. Also, for 20 738 250 compounds taken from a fourth reality database, 10 duplicates occurred.

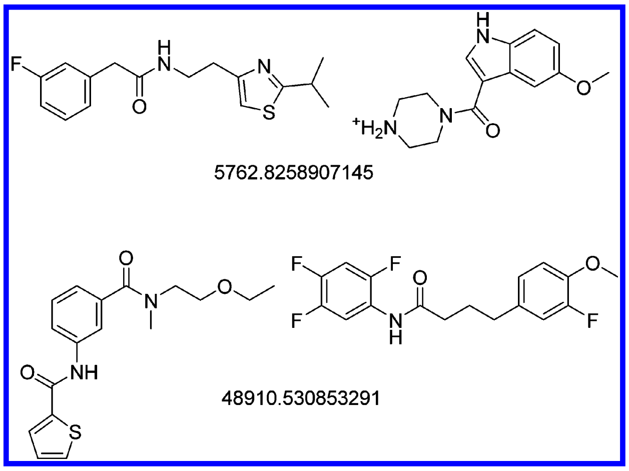


Figure 10. Two degeneracies for 2-EAID in the reality data set taken from the ZINC database.

For the above reason, the EAID index was extended to obtain a new index termed 2-EAID. The discriminating power of 2-EAID is obviously increased compared with EAID. The uniqueness of the topological index 2-EAID for all of the alkane trees with 1−24 carbon atoms (a total of over 24 million compounds) was examined, and no duplicate occurred. Also, for over 0.38 million structures containing heteroatoms, no duplicate occurred; for 5.9 million compounds possessing high diversity taken from the three reality databases, no duplicate occurred; and for the 20 738 250 compounds taken from the fourth reality database, two duplicates occurred. Of course, the calculation of 2-EAID is slightly more complicated than the calculation of EAID. To the best of our knowledge, EAID and 2-EAID are the most highly discriminating indices examined to date.

Because of their high discriminating power, these indices possess not only theoretical significance but also potential applications, such as confirmation that a new compound actually is new, administration of large compound databases, expert systems for organic structure evaluation, computer-aided synthesis of organic compounds, and so forth.

In this research, the uniqueness examinations of compounds did not involve the stereoisomers. If needed, we could do this on the basis of our previous studies, which included exhaustive generation of stereoisomers,[16,17] extraction of characterizations for stereoisomers, and description of stereoisomers for quantitative structure−activity relationship (QSAR) studies representing chiral centers of stereoisomers, 3D QSAR studies for drugs, and so on.[27−32]

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Additional examples of complicated structures in the reality data sets. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00044.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: luxu@ciac.ac.cn.
*E-mail: zhqingyou@henu.edu.cn.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the National Natural Science Foundation of China (20875022) for financial support. The authors acknowledge the International Science and Technology Cooperation of Henan Province (114300510009). The project was also sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (20091001).

■ REFERENCES

(1) Balaban, A. T. Topological Index Based on Topological Distances in Molecular Graphs. *Pure Appl. Chem.* **1983**, *55*, 199−206.

(2) Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: Amsterdam, 1999.

(3) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, 2000.

(4) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Mannhold, R., Kubinyi, H., Folkers, G., Eds.; Wiley-VCH: Weinheim, Germany, 2009.

(5) Wiener, H. Structure Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(6) Randić, M. On of molecular similarity based on a single molecular descriptor. *Chem. Phys. Lett.* **2014**, *599*, 1−6.

(7) Randić, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.

(8) 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.

(9) Balaban, A. T. Numerical Modeling 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.

(10) 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.

(11) Hu, C. Y.; Xu, L. A New Topological Index for the Changchun Institute of Applied Chemistry ^{13}C NMR Information System. *Anal. Chim. Acta* **1995**, *318*, 117−123.

(12) Hu, C. Y.; Xu, L. On Highly Discriminating Molecular Topological Index. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 82−90.

(13) Diudea, M. V.; Ilić, A.; Varmuza, K.; Dehmer, M. Network Analysis Using a Novel Highly Discriminating Topological Index. *Complexity* **2011**, *16*, 32−39.

(14) Hu, C. Y.; Xu, L. Expert System for Elucidation of Structures of Organic Compounds—Structural Generator of ESESOC-II. *Sci. China, Ser. B* **1995**, *3*, 296−304.

(15) Xu, L.; Hu, C. Y. *Applied Chemical Graph Theory*; Science Press: Beijing, 2000; pp 112 − 145.

(16) Hao, J. F.; Xu, L. Automorphism Group Algorihm in Computer-Aided Structure Elucidation. *Chin. J. Anal. Chem.* **2000**, *28*, 1209−1213.

(17) Hao, J. F.; Xu, L.; Hu, C. Y. Expert System for Elucidation of Structures of Organic Compounds (ESESOC)—Algorithm on Stereoisomer Generation. *Sci. China, Ser. B* **2000**, *43*, 503−515.

(18) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press: Letchworth, Hertfordshire, England, 1986; pp 1−100.

(19) Zhao, T. F.; Zhang, Q. Y.; Long, H. L.; Xu, L. Graph Theoretical Representation of Atomic Asymmetry and Molecular Chirality of Benzenoids in Two-Dimensional Space. *PLoS One* **2014**, *9*, No. e102043.

(20) Caporossi, G.; Hansen, P. Enumeration of Polyhex Hydrocarbons to $h = 21$. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 610−619.

(21) AKos. http://www.akosgmbh.de/AKosSamples/download.htm (accessed September 2014).

(22) CAS. http://www.cas.org/content/chemical-substances (accessed Jan 16, 2015).

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

(24) Wipke, W. T.; Dyott, T. M. Stereochemically Unique Naming Algorithm. *J. Am. Chem. Soc.* **1974**, *96*, 4834−4842.

(25) Jochum, C.; Gasteiger, J. Canonical Numbering and Constitutional Symmetry. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 113−117.

(26) Randić, M. On Canonical Numbering of Atoms in a Molecule and Graph Isomorphism. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 171−180.

(27) Zhang, Q. Y.; Carrera, G.; Gomes, M. J. S.; Aires-de-Sousa, J. Automatic Assignment of Absolute Configuration from 1D NMR Data. *J. Org. Chem.* **2005**, *70*, 2120−2130.

(28) Zhang, Q. Y.; Zhang, D. D.; Li, J. Y.; Long, H. L.; Xu, L. Prediction of Enantiomeric Excess in a Catalytic Process: A Chemoinformatics Approach Using Chirality Codes. *MATCH* **2012**, *67*, 773−786.

(29) Zhang, Q. Y.; Aires-de-Sousa, J. Physicochemical Stereodescriptors of Atomic Chiral Centers. *J. Chem. Inf. Model.* **2006**, *46*, 2278−2287.

(30) Suo, J. J.; Zhang, Q. Y. The Derivation of a Chiral Substituent Code for Secondary Alcohols and Its Application to the Prediction of Enantioselectivity. *J. Mol. Graphics Modell.* **2013**, *43*, 11−20.

(31) Suo, J. J.; Zhang, Q. Y. Prediction of Enantioselectivity of Primary Alcohols Involved in Racemic Resolutions Using Chiral Substituent Code. *Chemom. Intell. Lab. Syst.* **2013**, *128*, 118−123.

(32) Zhang, Q. Y.; Xu, L. Z.; Li, J. Y.; Zhang, D. D.; Long, H. L.; Leng, J. Y.; Xu, L. Methods of Studies on Quantitative Structure−Activity Relationships for Chiral Compounds. *J. Chemom.* **2012**, *26*, 497−508.

DOI: 10.1021/acs.jcim.5b00044