Molecular Shape Profiles†

Milan Randić

Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311

Received October 24, 1994⊗

We introduce a scheme that gives a numerical characterization for a molecular shape and shapes in general! A sequence that represents molecular shape is derived from the powers of interatomic distances for all atoms at the atomic periphery. One may view the derived sequence, which we call shape profile, analogous to a power series expansion of a function. The elements of the sequence are given as the average contribution from all atoms at the molecular periphery. We illustrate the shape profiles for smaller benzenoid systems and use them in a discussion of molecular similarity for benzenoidal shapes. The approach can be extended to other molecular forms, including nonplanar structures. The "resolution" of the characterization of the shapes can be increased by increasing the number of points (not necessarily atoms) on the molecular periphery.

INTRODUCTION

> "When you can measure what you are speaking about, and express it in numbers, you know something about it. But when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: It may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science."
>
> Lord Kelvin
> (Popular Lectures & Addresses 1891−1894)

Molecular shape is an important though elusive chemical concept. It has been used mostly in a qualitative way, without trying to quantify it rigorously, or even without trying to define it mathematically. We all have a fair notion what is a shape, and molecular shape in particular, although it remains difficult to describe our notions of the shape. We are aware of inherent difficulties in trying to quantitatively characterize such quantity, the task that almost appears impossible. However, without a numerical characterization it is difficult to use this important concept in quantitative discussions of structure−property−activity relationships. As Lord Kelvin observed, unless we can numerically characterize a quantity, we do not fully know it!

Advances in structure−property and structure−activity studies made a need for a quantification of the concept of molecular shape apparent. The time has come to attempt the "impossible", to define shapes as mathematical objects and represent them, including molecular shapes, numerically. The purpose of this paper is to outline a numerical characterization of molecular shapes. Although we consider a special case, the shapes defined by the periphery of smaller benzenoid systems, the outlined procedure is quite general. It applies not only to arbitrary planar contours that represent two-dimensional shapes but also to arbitrary shapes of three-dimensional objects.

Molecular shapes have received some attention in the literature. For planar benzenoid systems a simple, and a very coarse, description was suggested by using the ratio of the length and the width of a rectangle in which molecule is "packed".[1] Even so rough descriptor has been found useful in discussion of chromatographic data for benzenoid compounds. A more precise representation of molecular shape should be based on details of the molecular geometry. Such information may be supplied by a list of all interatomic distances, which can conveniently be presented as a distance matrix. Alternatively, a molecule can be represented by a set of contours of surfaces of equal electronic density for two-dimensional and three-dimensional models, respectively. The first approach gives a discrete representation of a molecule, the second one a continuous one. The former approach gives exact atomic positions but only approximately suggests the bonding pattern; the latter approach gives the bonding pattern and suggests approximately the atomic coordinates. While both approaches may help one to visualize the shape of a molecule, neither of the two representations of a molecule is conducive to a characterization of the "molecular shape". We should add here that in contrast to shapes of sold objects the notion of molecular shape is more involved. Is shape to be specified by positions of atoms in a molecule, or is it to be given by molecular surface? Molecular surface itself is a fuzzy concept in the sense that it depends on the assumed threshold for electron density around the nuclei. These questions about the fuzzy nature of three-dimensional molecular shapes have been raised and discussed by Mezey.[2,3]

Mezey and co-workers, over the years, have been analyzing molecular shapes, mostly by focusing attention on the topological properties of the molecular electron density surfaces.[4−8] They also considered quantifying molecular shapes by using the shapes of "square animals" to fill the interiors of contours of electron densities. By using "square animals" of different size one can arrive at numerical characterization of different contours.

Another avenue to characterization of molecular shapes has been recently outlined by this author and M. Razinger.[9−11] These authors considered molecular periphery (or contours of electron density) for planar structures and developed a

† This contribution is dedicated to George W. A. Milne, the Editor of the Journal of Chemical Information & Computer Science, whose enthusiasm and professionalism has transformed a "peripheral" chemistry journal into one of the more highly visible chemistry documents.

⊗ Abstract published in *Advance ACS Abstracts,* April 1, 1995.

0095-2338/95/1635-0373$09.00/0

binary code that encodes the shape of a contour embedded on a hexagonal regular grid. Although the approach was illustrated on planar benzenoid forms the method is general and applies to contours of any shape. The binary code for molecular periphery was based on a superposition of the molecule on a regular graphite lattice. By using 0 and 1 as labels for the right and the left turn at each site of the lattice when circling around the molecular periphery one arrives at a binary string that allow shape reconstruction. One could trace such binary codes to Rouse Ball,[12] who used binary code to label Hamiltonian paths on a cubic graph. In chemical context Balaban used similar binary codes to label configurations of annulenes.[13]

By choosing the smallest binary label one obtains unique shape codes for arbitrary shape drawn on a graphite lattice. The so obtained codes have been used to establish the degree of molecular similarity[9] and a measure of chirality for a pair of chiral structures.[10] The similarity measure was based on the count of agreements and disagreements among the binary digits of the corresponding codes when they are shifted such that they differ the least. The magnitude of the (two-dimensional) chirality of planar benzenoids was based on the degree of the overlap of the binary codes for the pair of chiral enantiomers. Alternatively, a set of all cyclically shifted binary codes can be used to form a symmetric matrix. Invariants of such matrices, including the eigenvalues of the characteristic polynomial, may serve as molecular shape descriptors.[11] The outlined methodology not only applies to planar structures of arbitrary shape but also can be extended to three-dimensional objects as was briefly outlined in the literature.

One has to differentiate between codes and invariants. Codes can result in a unique molecular representation that allows structure reconstruction. A code assumes or induces an atomic labeling, hence, it presumes a convention for its construction. In contrast a structural invariant represents a molecular or structural property, hence, it is independent of atomic labeling. To evaluate (or measure) its magnitude does not require canonical labeling of vertices. All this gives an advantage to use of invariants for representation of molecules. Their disadvantage is that there is no guarantee that a finite list of invariants is unique. Even if a list of invariants for a set of compounds is found to be unique, there are no guarantees that a structure can be recovered from such a list.

In other words, a construction of an invariant is associated with a loss of information. On the other hand, by being itself structural property invariants are convenient and logically suitable quantities to discuss structure−property relationships. Such a relationship is then reduced to a property−property relationship, that is, molecular structural property is described and discussed in terms of mathematical properties of the structure. In contrast there is no loss of information associated with construction of molecular codes. Hence, codes are essential for chemical documentation of a structure or for input of a structure into a computer. The disadvantage of codes is that all users have to have the code-book, i.e., have to know how the code is obtained, so that they can use it in reverse to reconstruct the structure. In addition, codes are not convenient for a direct discussion of the structure−property relationship, since they relate quantities of different kind: labels dependent and label independent quantities.

SHAPE INVARIANTS

In this article we will be interested in shape invariants, not shape codes. Are there structural invariants that are sensitive to the molecular shape? Such, if found, can be used to classify and characterize molecular shapes. Each time we come across a descriptor that gives different numerical values for molecules of different shape or gives different numerical values for the same molecule in a different conformation, we can use such a descriptor to differentiate among structurally closely related systems. Individual descriptors "project" but a single molecular (mathematical) property of a structure. Though being useful, and even perhaps sufficient in specific applications, they give but a limited "portrayal" of a structure. Hence, a longer list of invariants is going to better serve in a general situation. Such "longer" lists of structural invariants we refer to as a characterization or signature. In a way one can look at such "longer" lists as a "basis" for a representation of a structure if the same set of invariants can be used for variety of structures.

In the case of molecular graphs such characterizations are given by a list of the so called topological indices.[14] They can be collected in an *ad hoc* manner, or alternatively one can focus attention to a set of structurally closely related invariants. Illustrations of the latter are the connectivity indices,[15−18] the path numbers, augmented paths, walks of different length, or extended connectivities.[19−21] Use of several invariants allows one to make comparative regression analysis and study how different properties of a set of molecules depend on the same structural factors.[22−24] Structurally related invariants have therefore advantages, particularly when they have a clear structural interpretation. Hence our goal is to design structurally related invariants that can characterize molecular shapes.

The geometry-based molecular distance matrix, being sensitive to details of molecular architecture, is a natural source for geometry-dependent structural invariants. Distance matrix is essentially the table of interatomic distances. Clearly such a table fully defines an object and allows reconstruction of a structure. In fact, distance matrices contain redundant information, since knowledge of the relative positions of three consecutive bonds suffices to determine exactly their geometry. After finding the positions of the first four atoms one can ignore the first atom and use the remaining three atoms to fix the coordinates of the next atom relative to the three known atomic positions. We continue in such a manner until the positions of all atoms have been determined, which completes the reconstruction. From the chemist point of view it is important to realize that the distance matrix does not give explicit information on the bonding within a molecule. The bonding has to be inferred by considering the shortest distances and having information on valence electrons. In contrast the distance matrix associated with molecular graphs gives information of chemical bonding (adjacency) but not on the geometry. One may say that the geometrical distance matrices reflect "through space" interactions, while graph theory adjacency and distance matrices reflect "through bonding" interactions. We should add that often "through bond" and "through space" descriptions strongly overlap. This is, for example, reflected in the intercorrelation of the 2-D and 3-D Wiener index, which is very high (ref 13, p 264).

Graph theory adjacency and distance matrices have been a source of many molecular invariants, the so called topological indices.[25-28] While the adjacency matrix has already been used by Cayley, the graph distance matrix was introduced in graph theory relatively recently by F. Harary.[29] One of the first topological indexes constructed from the graph distance matrix is the J index of Balaban.[30-31] He used the row sums of the distance matrix for the construction of the J index in an analogy to the relationship of the row sums of the adjacency matrix and the connectivity index χ.[15] One can view the Wiener index W[32] as derived from the elements of the distance matrix, as has been pointed out long ago by Hosoya,[33] although it was originally defined without a reference to the distance matrix.

The same "techniques" for construction of topological indices have been extended from graph distance matrices to geometry distance matrices. In this way we arrive at the so called "topographic" indices, which can differentiate *cis* and *trans* isomers, *gauche* and *anti*, etc.[34-37] Similarly the Wiener index was generalized for three-dimensional structures by using elements of geometry matrix rather than adjacency matrix.[38,39] Finally, recently the information from both matrices, the geometry and the topology, was combined into a single D/D matrix. Suitably normalized first eigenvalue of such a matrix, it was argued, describes the "degree of molecular folding".[40]

We will here outline yet another general approach to construction of structural invariants from a structural matrix. We refer to this as "shape characterization" or "shape expansion", in an analogy with the power expansions of functions. The procedure gives a characterization to any shape to which this way a single function can be assigned. The significance of this result is that we succeeded in mapping a shape into a well-defined simple function. We will refer to these functions as (molecular) shape profiles. Here is the outline of the method that accomplished the mapping: We start with the (geometry) distance matrix of a structure

$$\mathbf{D} = \begin{pmatrix} d_{1,1} & d_{1,2} & d_{1,3} & d_{1,4} & \dots \\ d_{2,1} & d_{2,2} & d_{2,3} & d_{2,4} & \dots \\ d_{3,1} & d_{3,2} & d_{3,3} & d_{3,4} & \dots \\ d_{4,1} & d_{4,2} & d_{4,3} & d_{4,4} & \dots \\ \dots & \dots & \dots & \dots & \end{pmatrix}$$

From this matrix we extract the row (or columns) sums $R_1, R_2, R_3, R_4, \dots$ The quantity in which we are interested is the average of row sums: $R = (R_1 + R_2 + R_3 + R_4 + \dots)/N$. Here N is the number of atoms in the structure or the number of points taken to represent the structure. In the next step we consider the matrix

$${}^2\mathbf{D} = \begin{pmatrix} d^2_{1,1} & d^2_{1,2} & d^2_{1,3} & d^2_{1,4} & \dots \\ d^2_{2,1} & d^2_{2,2} & d^2_{2,3} & d^2_{2,4} & \dots \\ d^2_{3,1} & d^2_{3,2} & d^2_{3,3} & d^2_{3,4} & \dots \\ d^2_{4,1} & d^2_{4,2} & d^2_{4,3} & d^2_{4,4} & \dots \\ \dots & \dots & \dots & \dots & \end{pmatrix}$$

where $d^2_{i,j}$ is the square of the element $d_{i,j}$. Again we consider the row sums and construct the average row sum: ${}^2R = ({}^2R_1 + {}^2R_2 + {}^2R_3 + {}^2R_4 + \dots)/N$, where ${}^2R_i$ are the sums of the ith row of the ${}^2\mathbf{D}$ matrix.

The process extends to matrices constructed from the elements obtained from the higher powers of $d_{i,j}$. In this way we obtain the sequence: $(R, {}^2R, {}^3R, {}^4R, {}^5R, {}^6R, \dots)$. In order to reduce the role of ever increasing powers we normalize the (averaged) row sums by the factorial given by the power exponent used. Thus we obtain the sequence:

$${}^1R, {}^2R/2!, {}^3R/3!, {}^4R/4!, {}^5R5!, {}^6R/6!, \dots$$

This can, alternatively, be presented as a power series:

$$S = N + {}^1Rx + {}^2R/2!x^2 + {}^3R/3!x^3 + {}^4\mathrm{R}/4!x^4 + {}^5\mathrm{R}/5!x^5 + {}^6R/6!x^6 + \dots$$

or sequence:

$$S = {}^1S, {}^2S, {}^3S, {}^4S, {}^5S, {}^6S, \dots$$

We can formally introduce the constant N, which indicates the size of the system, as the leading term in the sequence:

$$S = {}^0S, {}^1S, {}^2S, {}^3S, {}^4S, {}^5S, {}^6S, \dots$$

The leading term ${}^0S$ can be viewed as derived from the matrix $d^0_{i,j}$.

The shape descriptors ${}^1S, {}^2S, {}^3S, {}^4S, {}^5S, {}^6S$, or the sequence S, represent the molecular profile. Because of the presence of the factorials in the denominators the sequence will always converge. In the following section we will illustrate molecular shapes of planar benzenoid systems.

OUTLINE OF THE MOLECULAR SHAPE PROFILES

For catacondensed benzenoids all carbon atoms are on the molecular periphery. Therefore molecular codes based on taking into account all atoms and shape codes that consider only atoms at the molecular periphery will neceesarily be identical. Difference between the two will emerge when one considers perycondensed benzenoids. Before considering larger benzenoids we will first examine the simplest benzenoids: benzene and naphthalene. In the case of benzene the distance matrix is particularly simple:

$$\begin{pmatrix} 0 & 1 & \sqrt{3} & 2 & \sqrt{3} & 1 \\ 1 & 0 & 1 & \sqrt{3} & 2 & \sqrt{3} \\ \sqrt{3} & 1 & 0 & 1 & \sqrt{3} & 2 \\ 2 & \sqrt{3} & 1 & 0 & 1 & \sqrt{3} \\ \sqrt{3} & 2 & 3 & 1 & 0 & 1 \\ 1 & \sqrt{3} & 2 & \sqrt{3} & 1 & 0 \end{pmatrix}$$

All the row sums are equal to 7.464 101 62, hence no need for averaging. The sequence R for the first six powers of the matrix becomes

$$R = 6, 7.464\ 101\ 62, 12, 20.392\ 304\ 9, 36, 65.176\ 914\ 7, 120, \dots$$

which, when normalized, gives the profile

$$S = 6, 7.464\ 101\ 62, 6, 3.398\ 717\ 48, 1.5, 0.543\ 140\ 956, 0.166\ 666\ 667, \dots$$

We see, at least for this case, that to obtain the molecular shape profile S it suffices to consider the first N powers of

Table 1. The Row Sums for the Three Nonequivalent Carbon Atoms of Naphthalene for Ten Powers and the Corresponding Shape Coefficient (the Last Column)

N	C(1)	C(2)	C(3)	average
1	13.928	17.488	20.180	17.852 505
2	23	38	53	20.5
3	39.785	89.629	154.354	17.591 682 1
4	71	224	479	12.308 333 4
5	129.354	583.049	1545.986	7.312 378 18
6	239	1,562	5,117	3.776 944 47
7	446.061	4272.282	17225.387	1.723 865 05
8	839	11886	58679	0.703 993 061
9	1588.184	33322.786	201639.804	0.259 872 887
10	3023	94418	697493	0.087 458 389 3
1	13.928	17.488	20.180	17.852 505
2	11.5	19	26.5	20.5
3	6.630 833 333	14.938 166 67	25.725 666 67	17.591 682 1
4	2.958 333 333	9.333 3333 33	19.958 333 33	12.308 333 4
5	1.077 95	4.858 741 667	12.883 216 67	7.312 378 18
6	0.331 944 444	2.169 444 444	7.106 944 444	3.776 944 47
7	0.088 107 341	0.847 675	3.417 735 516	1.723 865 05
8	0.020 808 532	0.294 791 667	1.455 332 341	0.703 993 061
9	0.004 376 609	0.091 828 665	0.555 665 245	0.259 872 887
10	0.000 833 058	0.026 019 07	0.192 210 373	0.087 458 389 3

the distance matrix. N is the number of atoms on the molecular periphery, not the number of atoms in a molecule, since the interior atoms do not contribute to the characterization of molecular shape.

In the case of naphthalene there are three nonequivalent atoms (on the periphery), thus it suffices to consider only three rows of the distance matrix, shown below:

$$\begin{array}{cccccccccc} 0 & 1 & \sqrt{3} & 2 & \sqrt{3} & 1 & \sqrt{3} & 2 & \sqrt{3} & 1 \\ 1 & 0 & 1 & \sqrt{3} & 2 & \sqrt{3} & \sqrt{7} & 3 & \sqrt{7} & \sqrt{3} \\ \sqrt{3} & 1 & 0 & 1 & \sqrt{3} & 2 & 3 & \sqrt{13} & \sqrt{12} & 7 \end{array}$$

The corresponding row sums are, respectively,

$$13.928\ 293\ 2,\ 17.487\ 655\ 1,\ 20.179\ 505\ 8$$

which gives as the average 17.852 505.

In Table 1 we give the row sums for the three nonequivalent carbon atoms of naphthalene. The entries are the row sums ${}^{i}R$ for the first ten powers of the matrices d^N of naphthalene from which the coefficients ${}^{i}S$ of the shape profile (shown in the last column) are constructed. To obtain the entries of the last column one first averages the contributions of individual carbon atoms using in the case of naphthalene, $\{2C(1) + 4C(2) + 4C(3)\}/10$, since there are two central atoms and four of each kind of peripheral carbon atoms. The coefficients ${}^{i}S$ are obtained by dividing the ${}^{i}R$ by the corresponding factorial. In the lower part of Table 1 we give normalized atomic contributions for the three nonequivalent carbon atoms of naphthalene. From this part of the table we immediately see that the atomic shape sequences

$$A = {}^{1}A, {}^{2}A, {}^{3}A, {}^{4}A, {}^{5}A, {}^{6}A, \ldots$$

are characteristic for individual atoms of a molecular periphery. The central carbon atoms have a short range, a short profile, while the most exposed terminal carbon atoms have the long range influence, a "longer" profile. The molecular shape profile is obtained by the superposition of all individual atomic profiles. It is clear from Table 1 that the "long" range of the profile is dominated by the atoms at the greatest separation form other atoms, that is by atoms away from the molecular center.

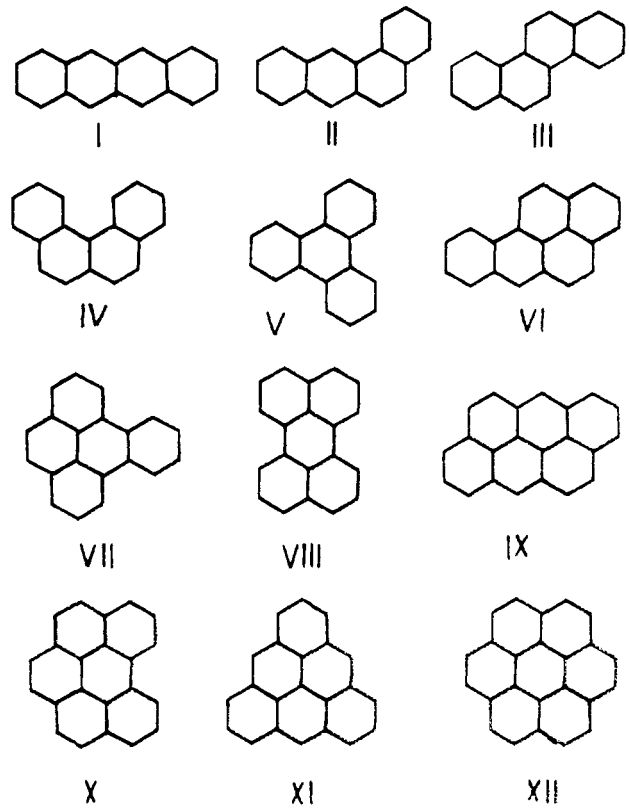


Figure 1. The shapes of smaller benzenoidal systems having periphery of 18 CC bonds.

The atomic profiles seem to offer a very natural resolution to the problem of finding the central vertices in a graph. The topic of the graph center received some attention in the literature.[41-44] The attempts to determine the central vertices had to be supplemented by hierarchical rules in order to discriminate nonequivalent vertices that compete for the "title" of being a graph center. Atomic profiles offer an alternative route to graph center.[45]

PLANAR BENZENOID SHAPES

We will now consider benzenoid systems of Figure 1. These are all possible structures having on the molecular periphery 18 carbon atoms or having 18 peripheral CC bonds. Alternatively they illustrate all possible hexagonal shapes having perimeter $P = 18$. Despite a rather limited size of the set, the shapes of Figure 1 illustrate "long" structure (tetracene I), "zig-zag" structure (chrysene III), a "crescent" (benzphenanthrene IV), a "star" shape (triphenylene V), a "plate" (anthrathene VII), a "box" (perylene VIII), a "triangle" (triangulene XI), and a "spherical" structure (coronene XII). Observe the rather vague descriptors: long, zig-zag, crescent, star, plate, box, triangular, and spherical. In studies of benzenoids often more colorful language was used to describe shapes. Thus one speaks of bay regions, fjords, snow flakes, perforated rectangles, flounders, waffles, etc.[46-48] In neurobiology neurons of different shape have been described as pyramidal cells, spider cells, claws, etc.[49] This alone is a signal of rather unsatisfactory state, i.e., use of very qualitative language for describing molecular shapes and shapes in general.

In Table 2 we list molecular profiles for the benzenoids of Figure 1. The computed profiles are also illustrated in Figure 2 in a form of bar diagrams. There are some

Table 2. Shape Profiles for the Ten Benzenoids of Figure 1

I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
51.671	49.282	48.412	45.651	45.507	51.289	49.130	49.237	54.292	52.664	53.549	56.108
100.500	88.833	85.500	73.000	72.000	93.833	83.333	84.000	103.500	95.333	99.000	108.000
154.089	124.085	117.444	87.806	84.968	130.651	104.352	106.511	147.402	126.197	133.968	150.505
197.375	143.708	134.597	86.000	81.000	149.708	104.778	109.000	170.264	132.778	144.000	165.000
217.115	142.586	132.953	71.388	64.987	146.846	88.186	93.965	166.714	116.286	128.684	149.287
208.979	123.788	115.449	51.508	45.050	126.279	64.022	70.100	142.265	87.372	98.525	115.100
178.592	95.510	89.374	32.886	27.497	96.755	40.913	46.116	107.828	57.540	66.031	77.376
137.119	66.297	62.359	18.832	14.991	66.840	23.365	27.132	73.594	33.749	39.366	46.141
95.512	41.810	39.566	9.776	7.383	42.022	12.066	14.434	45.706	17.850	21.141	24.735
60.854	24.152	23.000	4.640	3.316	24.226	5.668	7.006	26.047	8.599	10.330	12.048
35.711	12.867	12.328	2.028	1.368	12.891	2.467	3.126	13.713	3.803	4.631	5.379
19.417	6.359	6.128	0.822	0.522	6.366	0.990	1.290	6.708	1.555	1.918	2.217
9.832	2.930	2.838	0.310	0.185	2.932	0.370	0.495	3.064	0.591	0.738	0.849
4.658	1.264	1.230	0.109	0.061	1.265	0.130	0.178	1.312	0.210	0.265	0.303
2.072	0.513	0.501	0.036	0.019	0.513	0.043	0.060	0.529	0.070	0.089	0.102
0.869	0.196	0.192	0.011	0.006	0.196	0.013	0.019	0.201	0.022	0.028	0.032
0.345	0.071	0.070	0.003	0.002	0.071	0.004	0.006	0.073	0.007	0.009	0.010
0.130	0.024	0.024	0.001	0.000	0.024	0.001	0.002	0.025	0.001	0.002	0.003

Table 3. Similarity/Dissimilarity Matrix for the 12 Benzenoids of Figure 1

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
I	0	183	201	333	346	177	304	292	146	256	234	204
II		0	19	151	164	11.6	121	109	51	75	58	54
III			0	133	146	30	104	92	70	63	50	61
IV				0	13	160	36	46	199	92	116	157
V					0	173	48	59	212	104	128	169
VI						0	129	117	41	80	60	48
VII							0	11.9	166	56	81	122
VIII								0	154	46	70	111
IX									0	113	90	59
X										0	24	65
XI											0	41
XII												0

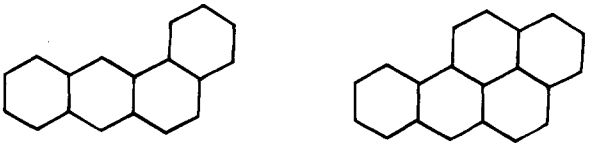
characteristic features of the profile digrams that are worth emphasizing. For example, in the case of the "long" molecules, represented here by tetracene (I), we see that the profile tapes off "slowly". The reason for this is that in such molecules the atoms at large separations dominate the averaging process for large powers of n. On the other hand, in the "star" triphenylene (V) or the "spherical" coronene (XII) the largest interatomic separation is smaller than that in the "long" molecules, hence the corresponding power expansion converges faster. The leading coefficients are also characteristic for the individual shapes. The "long" molecules show a "slow" start, while the spherical systems show a faster start and hence have the maximum of the shape profile at a smaller exponent index n.

SIMILARITY DERIVED FROM SHAPE PROFILES

We can consider shape profiles of Table 2 as vectors in N-dimensional space ($N = 18$) and construct the similarity/dissimilarity matrix using Euclidean distance as a measure of the degree of similarity among the molecules. The smaller the separation between two objects in the N-dimensional space the more similar are the corresponding structures. In Table 3 we give the similarity/dissimilarity matrix for the 12 benzenoids of Figure 1. Here we referred to triangulene as a benzenoid although it is a biradical, a molecule without a single Kekulé valence structure, hence not like aromatic benzene. The resulting "distances" (or dissimilarities) among the 12 structures considered vary by an order of magnitude, from the smallest entry of 11.59 to the largest entry of 345.92. For larger benzenoids one can expect an even larger span of the values between the most similar pair and the least similar pair of structures. In Figures 3 and 4 we depicted the least similar and the most similar pairs of benzenoids as found using the shape profiles to represent the molecules.

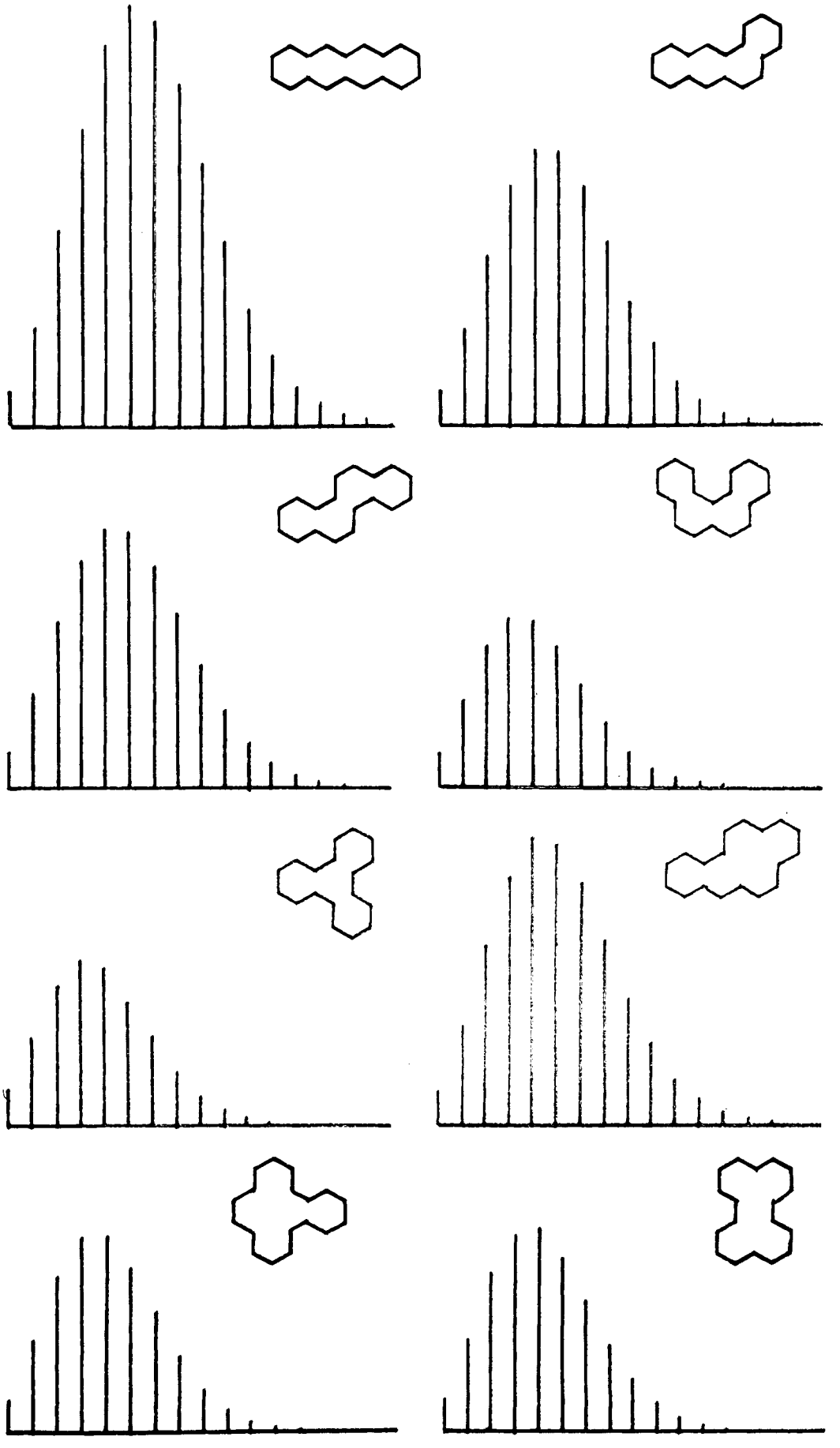
First, observe that by mere inspection of Figure 1 it is difficult and somewhat arbitrary to decide which is the most similar and which is the least similar pair of structures. On the other hand, the results of our analysis do not contradict the intuitive notions of similarity of shapes. We find tetracene (I) and triphenylene (V) as the least similar. Indeed, the former is a linearly fused system, while the latter is a maximally branched system among catacondensed benzenoids having the same perimeter. One can similarly rationalize the great dissimilarity of linear tetracene (I) and highly "bent" benzphenanthrene (IV), in which every benzene ring is fused so to produce the most "bent" structure possible. Among the less similar pairs we also find tetracene (I) and coronene (XII), the "linear" and the "spherical" isomeric structures. Finally, tetracene (I) and chrysene (III) are among the least similar structures, the first being "linear" and the latter having the maximal number of "kinks". As was to be expected, the pair (I, III) is less dissimilar than the pair (I, IV) (tetracene, benzphenanthrene), which has the same number of "kinks". In chrysene the "bending" of one part of the molecule is "corrected" by the opposite bend of the other part of the molecule, making chrysene less dissimilar to a "linear" tetracene.

In contrast to identifying the least similar shapes by inspection the task of determining the most similar pairs by an inspection of molecular forms is more difficult. We found from our analysis as the most similar pair benzanthracene (II) and benzo[*a*]pyrene (VI). At first sight this may look somewhat unexpected. Let us look more closely for this apparently unexpected result. The two molecules



benzanthracene (II) and benzo[a]pyrene (VI)

have all the periphery carbon atoms the same, except for a



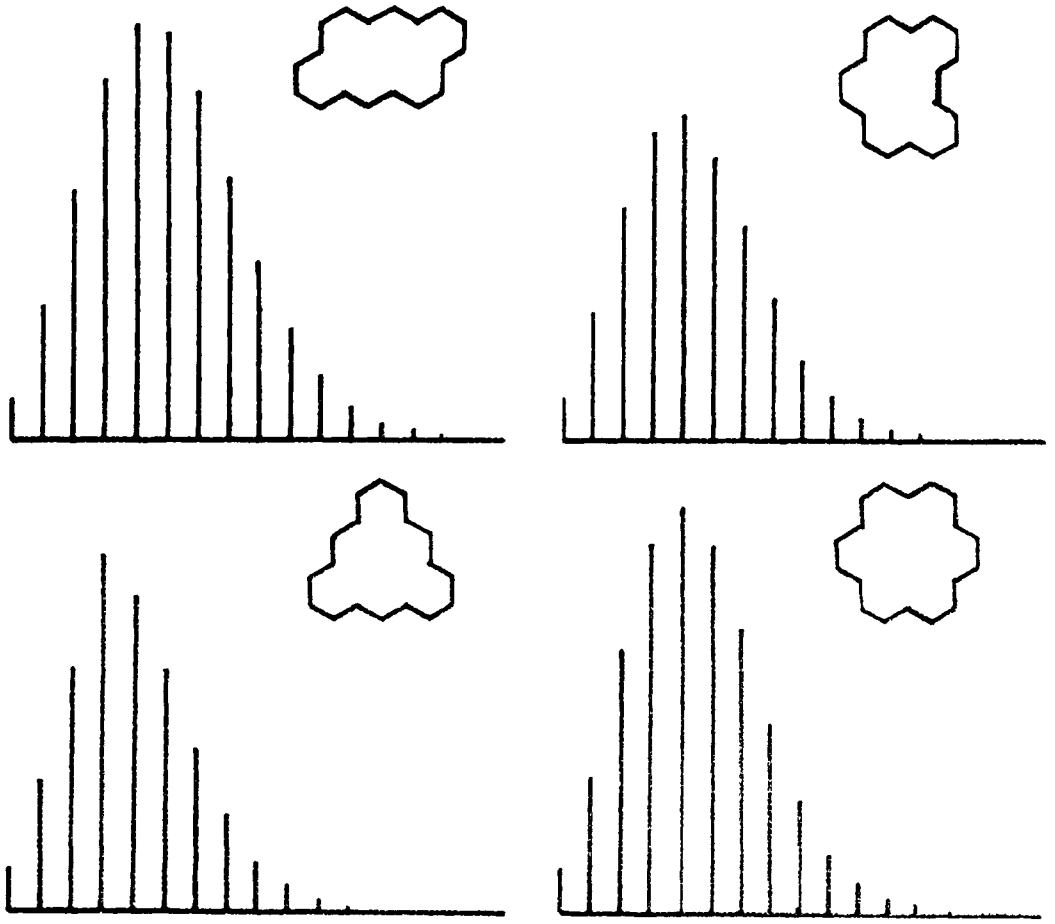


Figure 2. Graphical representation of the shape profiles for the smaller benzenoids of Figure 1.

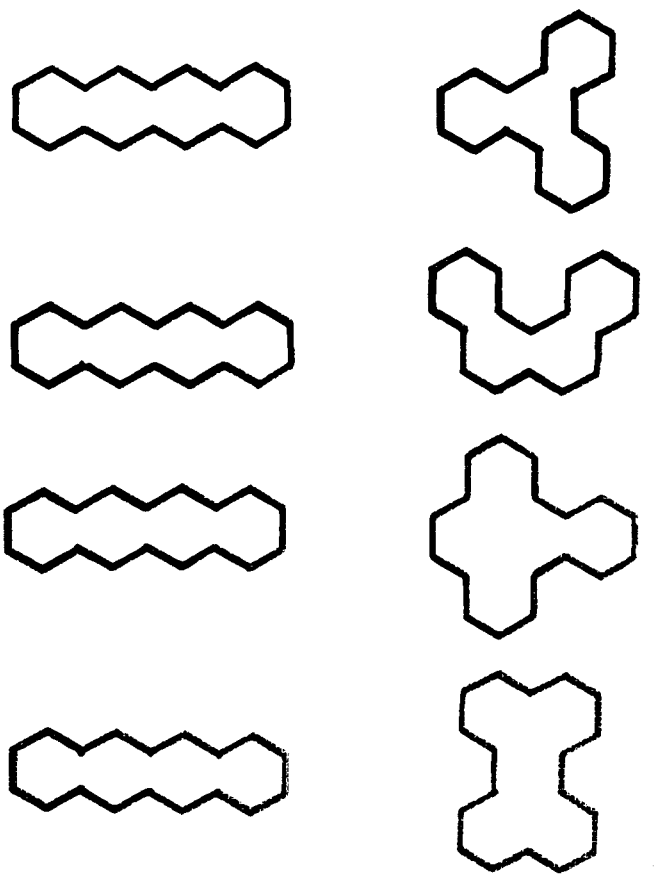


Figure 3. The least similar pairs of the benzenoid shapes of Figure 1.

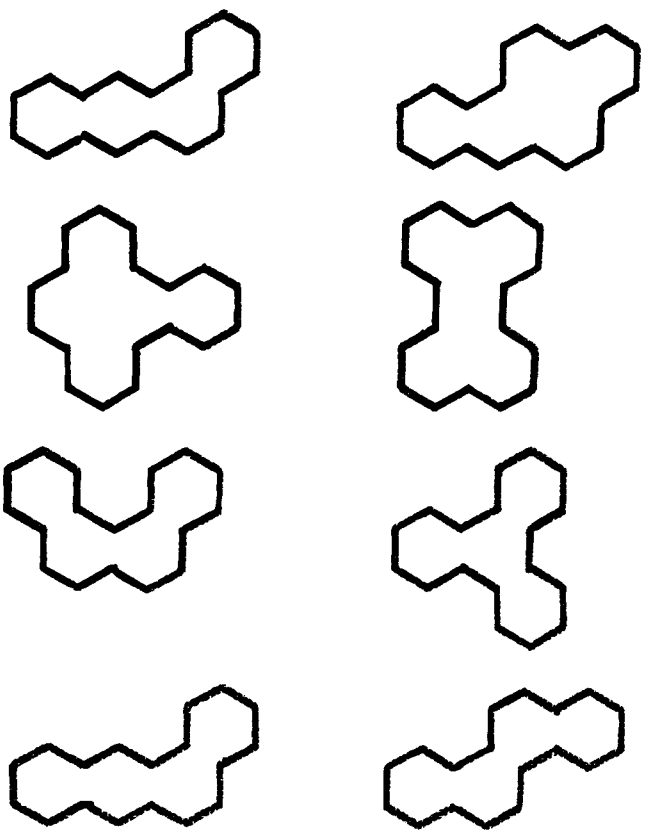


Figure 4. The most similar pairs of the benzenoid shapes of Figure 1.

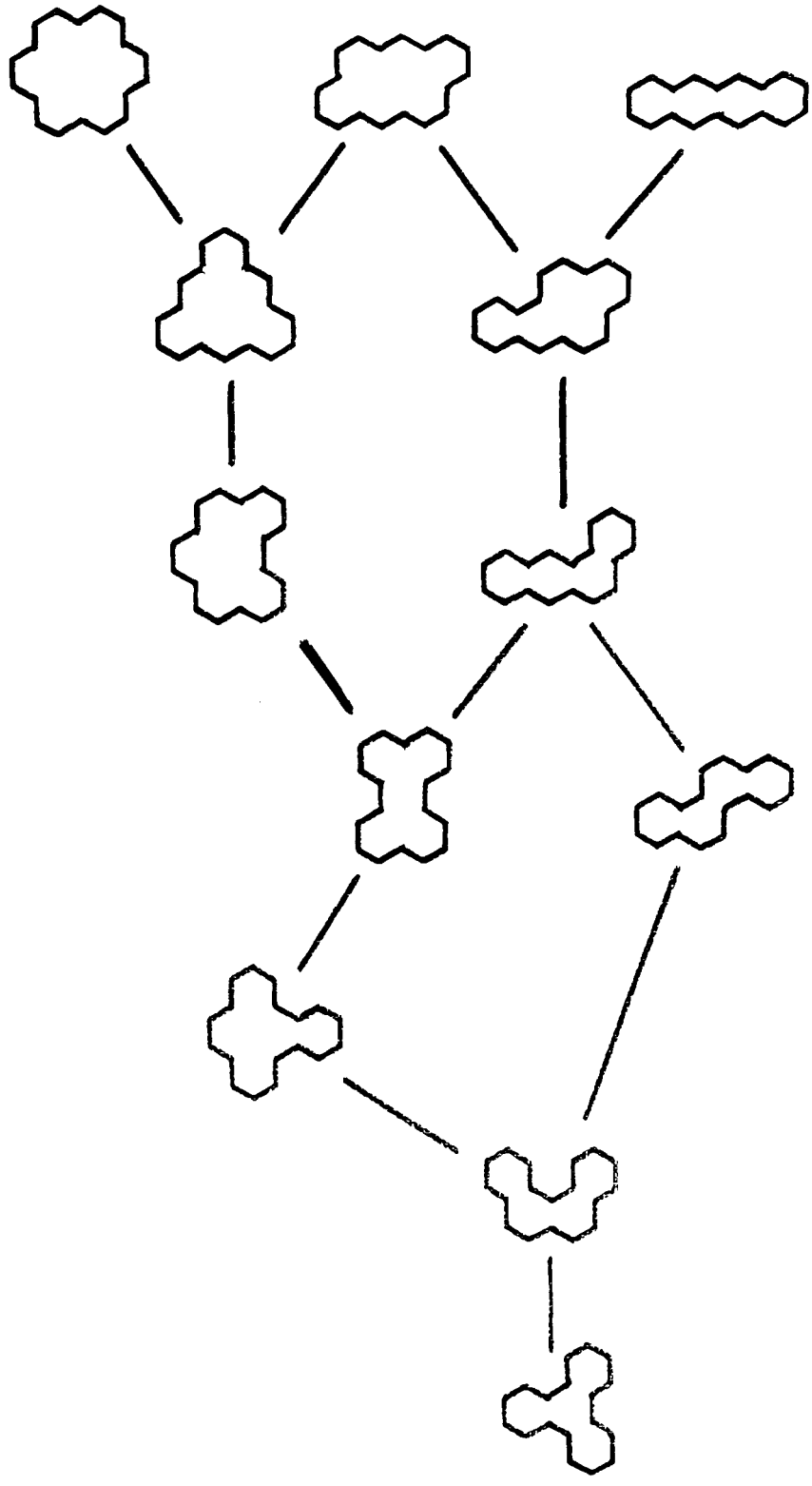


Figure 5. The partial order among benzenoids induced by the molecular profiles. The shape forms at the top dominate those below.

single pair. Moreover, the pair of carbon atoms in which the two shapes differ are within the same benzene ring, i.e., they are relatively close. The consequence is that most of the interatomic distances in the two structures remain identical. The distances arising from the carbon atoms in which the two structures differ will not be drastically different. This explains not only great similarity between (II) and (VI) but also between benzanthracene (II) and chrysene (III), the third most similar pair of Figure 4, and other similar pairs (III) and (VI), (V) and (VII), (VI) and (IX), and (XI) and (XII), and this can also explain a relatively small value for the dissimilarity of (II) and (IX). The similarity between benzpyrene (VII) and pyrene (VIII), the second most similar pair, and between benzphenanthrene (IV) and triphenylene (V), the third "best" pair, is not so obvious. Perhaps one can seek explanation in the fact that both of these molecules are "compact," which will have, as a consequence, shorter average interatomic distances for all the structures. This will result in fewer differences in their respective shape profiles.

ORDERING OF STRUCTURES

By representing each molecule by a sequence we are in a position to order the considered shapes by ordering the corresponding sequences. Generally ordering of sequences leads to a partial order, since initial members of one sequence can dominate the initial members of another sequence, but later the second sequence dominates the corresponding members of the first sequence. This, for example, happens if we compare the sequences for tetracene (I) and coronene (XII). One concludes that these two structures are not comparable, though both of them dominate the sequence representing triphenylene (V). In Figure 5 we show the graph of the partial order as induced by the molecular profiles of Table 2. The hierarchical ordering is assumed to be transitive. Thus for example, because dibenzopyrene (X) dominates perylene (VII), and perylene dominates benzo[*e*]-pyrene (VII) it follows that (X) dominates also (VII). In Figure 5 therefore we have not connected (X) and (VII) since they are connected through (VIII).

One regularity becomes apparent from the chart shown in Figure 4: Benzenoid B always dominates benzenoid B′, that is, its proper subgraph, i.e., B′ can be derived from B by erasing some peripheral CC edges. One can understand this regularity from the definition of the molecular profiles. The erased peripheral carbon atoms can only increase distances since internal carbons, which thus became peripheral, are more centrally located within a molecular structure.

The results of Figure 5 are interesting. We see at the top of the partial order coronene, anthrathene, and tetracene, the shapes that can be viewed as the most convex. The last, at the bottom, is triphenylene, the shape that can be viewed as the most concave or informally, the least convex. This observed regularity is likely to hold also for shapes built from a larger number of fuxed hexagons as well as for arbitrary shapes.

CONCLUDING REMARKS

This contribution outlines a novel approach to a mathematical representation of shapes. Although here we restricted attention to planar forms obtained by fusing hexagonal rings, the approach is general and applies to arbitrary planar shapes. All that one has to do in such a general case is to distribute N points on the perimeter of a shape and construct the corresponding distance matrix **D**, from which the shape coefficients follow. There is no restriction on the number of points. That number will depend on the resolution that one wishes to achieve. The size of an object can be eliminated from the characterization by additional normalization, such that all objects have a perimeter of a unit length.

Extension of this approach to three-dimensional systems is possible in two ways. A straightforward approach simply considers N points distributed, more or less uniformly, over the molecular surface. After selecting such N points we construct $N \times N$ distance matrix. Once we have a distance matrix we continue just as we did in the case of planar systems. Alternatively, one may consider a collection of contours on the surface of a three-dimensional object and represent the system by a list of profiles for the individual contours. Both approaches ought to yield a similar overall characterization. The distinction is in partitioning of the shape profile into contour profiles in one case and no partitioning of shape profile at all in the other case. The choice may depend on whether one is interested in local or global properties of a molecular shape.

This paper is seminal, in the sense that it opened a new direction in characterization of shapes. However, it will take some time before we answer all potentially interesting questions and come up with the shape descriptors that will satisfy everyone's needs. Let us end this by saying that even the normalization of powers of the distance matrices may be achieved in a different way. One possibility is based on factorials as outlined here. Another possibility is to use powers P^n instead of factorials. Recently Bonchev, Liu, and Klein discussed alternative weighting procedures as applying to self-returning walks,[50] but such procedures could be extended also to shape descriptors discussed in this contribution.

REFERENCES AND NOTES

(1) Radecki, A.; Lamparczyk H.; Kaliszan, R. A relationship between the retention indices on nematic and isotropic phases and the shape of polycyclic aromatic hydrocarbons. *Chromatographia* **1979**, *12*, 597−599.
(2) Mezey, P. G. The Shape of Molecular Charge Distributions: Group Theory without Symmetry. *J. Comput. Chem.* **1987**, *8*, 462−469.
(3) Mezey, P. G. *Shape in Chemistry: An Introduction to Molecular Shape and Topology*; VCH Publ.: New York, 1993.
(4) Mezey, P. G. The degree of similarity of three-dimensional bodies: Application to molecular shape analysis. *J. Math. Chem.* **1991**, *7*, 39−49.
(5) Harary, F.; Mezey, P. G. Similarity and Complexity of the shapes of square-cell configurations. *Theor. Chim. Acta* **1991**, *79*, 379−387.
(6) Walker, P. D.; Mezey, P. G. Representation of square-cell configurations in the complex plane: Tools for the characterization of molecular monolayers and cross sections of molecular surfaces. *Int. J. Quant. Chem.* **1992**, *43*, 375−392.
(7) Mezey, P. G. Iterated similarity sequences and shape ID numbers for molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 244−247.
(8) Arteca, G.; Mezey, P. G. Molecular conformation and molecular shape: A discrete characterization of continua of van der Waals surfaces. *Int. J. Quant. Chem.* **1988**, *34*, 517−526.
(9) Randić, M.; Razinger, M. On Characterization of molecular shapes. *J. Chem. Inf. Comput. Sci.*, in press.
(10) Randić, M.; Razinger, M. Molecular shapes and chirality. To be published.
(11) Randić, M.; Razinger, M. Molecular topographic indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 140−147.
(12) Rouse Ball, W. W. *Mathematical Recreations and Essays*; Macmillan: New York, 1967; 5th printing, p 262.
(13) Balaban, A. T. Chemical Graphs. Part 12. Configuration of annulenes. *Tetrahedron* **1971**, *27*, 6115−6131.
(14) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1992.
(15) Randić, M. On characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(16) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(17) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1966.
(18) Kier, L. B.; Hall, L. H. Molecular Connectivity in Structure−Activity Analysis; Research Studies Press: Chichester, England, 1986.
(19) Randic, M.; Wilkins, C. L. Graph theoretical ordering of structures as a basis for systematic searches for regularities in molecular data. *J. Phys. Chem.* **1979**, *83*, 1525−1540.
(20) Randić, M. Random walks and their diagnostic value for characterization of atomic environment. *J. Comput. Chem.* **1980**, *1*, 386−399.
(21) Randić, M. On representation of molecular graphs by basis graphs. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 57−69.
(22) Randić, M. Similarity based on extended basis descriptors. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 686−692.
(23) Randić, M. Comparative structure-property studies: Regressions using a single descriptor. *Croat. Chem. Acta* **1993**, *66*, 289−312.
(24) Randić, M.; Seybold, P. G. Molecular shape as a critical factor in structure−property activity studies. *SAR & QSAR* **1993**, *1*, 77−85.
(25) Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. Topological indices for structure−activity correlations. *Topics Curr. Chem.* **1983**, *114*, 21−55.
(26) Randić, M.; Trinajstić, N. In search for graph invariants of chemical interest. *J. Mol. Struct. (Theochem)* **1993**, *300*, 551−571.

(27) Randić, M. Generalized molecular descriptors. *J. Math. Chem.* **1991**, *7*, 155−168.
(28) Randić, M. In search of graph invariants. *J. Math. Chem.* **1992**, *9*, 97−146.
(29) Haray, F. Graph Theory; Addison-Wesley: Reading, MA, 1969.
(30) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(31) Balaban, A. T. Topological indices based on topological distances in molecular graphs. *Pure Appl. Chem.* **1983**, *55*, 199−206.
(32) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(33) Hosoya, H. Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.
(34) Randić, M. Molecular topographic descriptors. *Studies Phys. Theor. Chem.* **1988**, *54*, 101−108.
(35) Randić, M. On the characterization of three-dimensional structures. *Int. J. Quant. Chem.: Quant. Biol. Symp.* **1988**, *15*, 201−208.
(36) Randić, M.; Jerman-Blazic, B.; Trinajstic, N. Development of 3-dimensional molecular descriptors. *Comput. Chem.* **1990**, *14*, 237−246.
(37) Pogliani, L. On a graph theoretical characterization of cis/trans isomers. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 801−804.
(38) Bogdanov, B.; Nikolić, S.; Trinajstić, N. On the three-dimensional Wiener number. *J. Math. Chem.* **1989**, *3*, 299−309.
(39) Bogdanov, B.; Nikolić, S.; Trinajstić, N. On the three-dimensional Wiener number. A comment. *J. Math. Chem.* **1989**, *5*, 305−306.
(40) Randić, M.; Kleiner, A. F.; DeAlba, L. M. Distance/distance matrices. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 277−286.
(41) Bonchev, D.; Balaban, A. T.; Randić, M. The graph center concept for polycyclic graphs. *Int. J. Quant. Chem.* **1981**, *19*, 61−82.
(42) Bonchev, D. The concept for the center of a chemical; structure and its applications. *J. Mol. Struct. (Theochem)* **1989**, *185*, 155−168.
(43) Balaban, A. T.; Bonchev, D.; Seitz, W. T. Topological/chemical distances and graph centers in molecular graphs with multiple bonds. *J. Mol. Struct. (Theochem)* **1993**, *280*, 253−1260.
(44) Bonchev, D.; Balaban, A. T.; Liu, X.; Klein, D. J. Molecular cyclicity and centricity of polycyclcic graphs. I. Cyclicity based on resistance distances or reciprocal distances. *Int. J. Quant. Chem.* **1994**, *50*, 1020.
(45) Randić, M., work in progress.
(46) Chen, R.-S.; Cyvin, S. J. Enumeration of Kekule structures: Perforated rectangles. *J. Mol. Struct. (Theochem)* **1989**, *200*, 251−260.
(47) Cyvin, S. J.; Brunvoll, J.; Cyvin, B. N.; Tošić, R.; Kovačević, M.; Waffles, *J. Mol. Struct. (Theochem)* **1989**, *200*, 261−275.
(48) Tosic, R.; Cyvin, S. J. Enumeration of Kekule structures in benzenoid hydrocarbons: "Flounders". *J. Math. Chem.* **1989**, *3*, 393−401.
(49) Randić, M. Graph theoretical characterization of the dendritic fields. *Int. J. Quant. Chem.: Quant. Biol. Symp.* **1981**, *8*, 463−479.
(50) Bonchev, D.; Liu, X.; Klein, D. J. Weighted self-returning walks for structure−property correlation. *Croat. Chem. Acta* **1993**, *66*, 141−150.

CI9401175