Exploring Chemical Rings in a Simple Topological-Descriptor Space

Alan H. Lipkus†

Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210-0012

Received October 20, 2000

A new method for organizing chemical rings based on their topology is presented. It uses three simple descriptors that characterize separate aspects of ring topology. These descriptors are integers and can thus be interpreted as the coordinates of discrete cells in a three-dimensional space. The descriptor values of any ring topology correspond to the coordinates of some cell. A database of rings can be distributed in this descriptor space by assigning each of them to the corresponding cell. This approach is applied to a database of 40 182 different ring topologies, derived from a comprehensive collection of chemical rings extracted from the CAS Registry File. This database is distributed among 7387 cells, and the population statistics and spatial distribution of these cells are discussed. An examination of selected cells shows that ring topologies which are similar tend to be close together in descriptor space. Some results of using this space to study ring diversity are presented. It is found that the distribution of the ring-topology database is not highly compact but has many significant voids. It is also found that the distribution of medicinally relevant rings in this space shows the influence of certain structural constraints on drug molecules.

INTRODUCTION

The ring systems of a molecule are key features in determining its shape and properties. Rings play an especially significant role in pharmacological activity. The rings of a drug molecule can serve to properly position and orient functional groups that interact with a receptor and can also help to reduce the unfavorable loss of conformational entropy upon receptor binding. The manipulation of ring structure has long been a fundamental strategy in designing analogues to optimize the potency and selectivity of a lead molecule.[1] With regard to the discovery of lead molecules, rings are closely associated with two issues currently receiving much attention: understanding the structural patterns that distinguish drug from nondrug molecules and enhancing the diversity of molecules to be tested. Recently, for example, ring analysis was used to characterize the most common structural frameworks in a database of known drugs as a guide for future drug discovery.[2] In another study, the ring content of compounds was used as a basis for measuring and comparing the diversity of databases of pharmaceutical interest.[3] The continuing importance of ring systems in drug design suggests that new ways to explore the "space" of chemical rings could have a number of applications, e.g., mining a large database like the CAS Registry File for its drug-relevant ring content.[4]

It is very convenient to treat chemical rings as graphs, and this is often acceptable despite the importance of three-dimensional structure in determining properties such as bioactivity. A simple way to organize graph representations of rings for analysis, browsing, or searching is to use topological indexes. A number of topological indexes have been developed specifically for application to chemical rings.[5−13] Two of these are of particular interest here. Nilakantan et al.[8] proposed as a ring index the quantity $nb^2 - na^2 + na$, where *na* is the number of atoms and *nb* is the number of bonds. This index was applied to ring data extracted from a commercial database. It was shown that one could find ring systems similar in size and complexity to a target ring system by examining those rings grouped under the same index value as the target. It was suggested that this might be used to find novel variants of a ring system. More recently, Wife and co-workers[12] have characterized the topology of rings using an ordered list of the sizes of all cycles (i.e., all closed paths) in the ring system. This index was used to group database compounds into a series of bins. It was proposed that the ring content and diversity of a database is encoded in the distribution pattern of these bins and that such patterns can be used to compare databases.

In this paper, a new method for organizing chemical rings is presented. Like the two indexes just cited, it can group together ring systems whose topologies are different but similar. This method is based on three simple descriptors that characterize separate aspects of ring topology. These three descriptors are integers and can thus be interpreted as the coordinates of discrete cells in a three-dimensional space. Each ring system can be assigned to the cell whose coordinates are the descriptor values for that system. In this way, all of the rings in a database can be distributed among the cells of the descriptor space. This approach has been applied to a comprehensive collection of rings extracted from the CAS Registry File. This ring collection is a good representation of the set of all known organic chemical rings. Hence, the analysis of these rings—besides serving to illustrate the descriptor-space approach—is of special interest for the insights it may provide into the existing ring diversity of organic chemistry.

DEFINITION OF TOPOLOGICAL DESCRIPTORS USED

Consider a ring-system graph where *E* is the number of edges and *N* is the number of nodes. The minimum number

† Phone: (614)447-3600; e-mail: alipkus@cas.org.

10.1021/ci000144x CCC: $20.00
Published on Web 02/21/2001

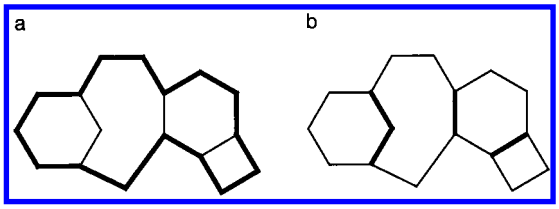


Figure 1. Partitioning of the edges in a ring-system graph as implied by descriptors *P* and *B*: (a) the 16 edges contributing to *P* are highlighted in bold and (b) the four edges contributing to *B* are highlighted.

of rings required for a basis set that describes the ring system is $R = E - N + 1$. The basis set consisting of the *R* smallest rings that are linearly independent is the so-called Smallest Set of Smallest Rings (SSSR). The SSSR is widely used for the description of chemical ring systems.[14] A problem with the SSSR is that the choice of smallest rings is not necessarily unique, and this has led to the definition of several other types of ring sets that incorporate additional rings.[14,15] This problem is of no consequence here because the topological descriptors to be used depend only on the sizes of the SSSR rings.

Let *S* be the sum of the ring sizes in the SSSR. Two new descriptors of ring topology can now be defined:

$$P = 2E - S \tag{1}$$

$$B = S - E \tag{2}$$

By definition, $P + B = E$. For this reason, the values of *P* and *B* for a ring-system graph imply a partitioning of its edges into those that contribute to *P* and those that contribute to *B*. For many ring topologies, this partitioning has a simple structural interpretation. This is illustrated by the graph in Figure 1, for which $P = 16$ and $B = 4$ (for this graph, $E = 20$ and $S = 24$, which is the sum of the SSSR ring sizes 4, 6, 6, 8). The formula for *P* (eq 1) essentially counts every edge twice and then removes each edge as many times as it appears in the SSSR; the edges left are those that are not shared between rings of the SSSR. In Figure 1a, these edges form what can be called the "perimeter" of the ring system. The edges contributing to *B*, on the other hand, are those that are shared between rings of the SSSR. In Figure 1b, these edges form what can be considered the "bridges" of the system.

The values of *P* and *B* for several other known ring topologies are shown in Figure 2. These examples demonstrate that, in contrast to the example of Figure 1, it is not meaningful in general to interpret *P* as the number of perimeter edges and *B* as the number of bridge edges (though the symbols "*P*" and "*B*" were suggested by this association). For instance, it can be seen that the value of $P = 5$ for the polyhedral ring system in Figure 2f (dodecahedrane) arises from the edges of one of its pentagonal faces, but calling this a perimeter would not make intuitive sense to many chemists since the ring system contains other closed paths that are much larger. It is clear that a simple structural interpretation of *P* and *B* has its limitations; however, this does not prevent these descriptors from serving as the basis of a useful parameter space.

The ring systems in Figure 2, examined in order (a−f), show a steady increase in the value of *B*/*E*, i.e., the proportion of edges contributing to *B*. That increase appears to correlate

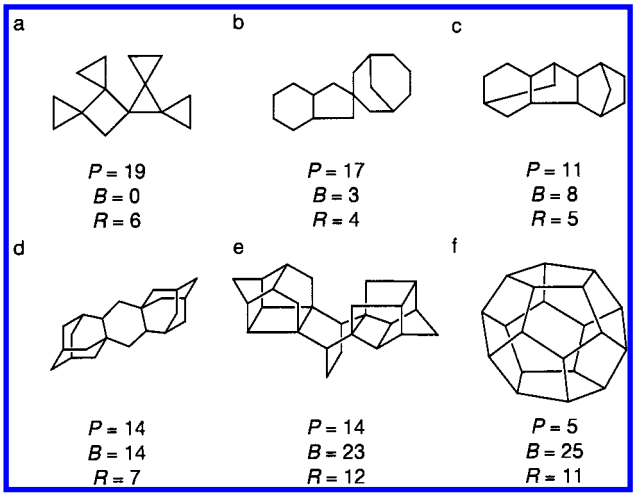
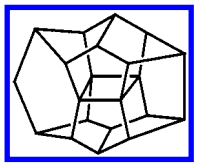


Figure 2. Values of *P*, *B*, and *R* for several ring topologies.

with an increase in the topological complexity of the systems. This correlation is reasonable since the magnitude of *B* relative to *P* is indicative of the extent to which the rings of the SSSR share edges; the greater the degree of edge sharing, the more highly bridged or caged the ring system will be.[16] This relationship between ring complexity and the relative magnitudes of *B* and *P* will prove helpful in understanding distribution patterns within descriptor space.

While *B* equals zero for single-ring and all-spiro systems (e.g., Figure 2a), this quantity can never be negative since every edge must appear in the SSSR no less than once. On the other hand, there is no prohibition against negative values of *P*, at least as applied to abstract graphs that have a sufficiently high degree of connectivity.[17] It is not clear whether chemical graphs, whose connectivity is limited by physical constraints, can have negative or zero values of *P*. In the ring database used here, 17 topologies were assigned nonpositive values of *P*, but an examination of these cases suggests they are most likely due to failures of the SSSR algorithm to properly perceive an SSSR that contains only smallest rings; such a failure can result in a value of *S* that is too large, which could yield a nonpositive value of *P*. The smallest positive value of *P* found in the database is exhibited by the following ring topology[18]



for which $P = 1$, $B = 32$, and $R = 13$. This value of *P* is surprisingly low but appears to be valid and is presumably due to the high degree of edge sharing among the SSSR rings. This example is certainly consistent with the notion that a high value of *B* relative to *P* is an indicator for especially complex rings.

CLASSIFYING RING TOPOLOGIES IN *PBR*-SPACE

A three-dimensional space for classifying ring topologies can be defined by associating the descriptors *P*, *B*, and *R* with its axes. This will be referred to as *PBR*-space. These descriptors are integers, and so a combination of *P*, *B*, *R* values can be viewed as the coordinates of a discrete cell in

this space. The descriptor values of any ring topology correspond to the coordinates of some cell. A collection of chemical rings can be classified by assigning each of them to the corresponding cell. For instance, a chemical ring having the topology shown in Figure 2a would be assigned to the (19, 0, 6) cell. Due to the limited specificity of *P*, *B*, and *R*, it is expected that many cells will be assigned more than one topology.

The classification of rings in *PBR*-space can be refined somewhat by using more fully the information in the SSSR. Only the sum of the ring sizes in the SSSR is used in calculating *P* and *B*, but there is clearly more information in the actual set of ring sizes. Within a cell, those ring systems that have the same set of ring sizes can be grouped together. Each such group will be called a bin of that cell. As an example, consider the ring topology in Figure 1. Since the ring sizes of its SSSR are (in ascending order) 4, 6, 6, and 8, this topology would fall into the {4, 6, 6, 8} bin of the (16, 4, 4) cell.

A Ring-Topology Database. In a database of chemical rings extracted from a structure file, many topologies will be represented by more than one type of chemical ring. Decalin, naphthalene, and quinoline, for instance, all represent the same ring topology. To explore the properties of *PBR*-space, it is better to use a database in which each topology is represented only once. This will ensure that the number of database entries assigned to a cell equals the number of different topologies. Such a ring-topology database has been created from CAS ring data.

CAS maintains a nonredundant file of all ring systems found in the compounds of the Registry File. Every ring system in this file is assigned an ID number for each of three characteristics: the ring-system graph, the pattern of elements in the ring system, and the pattern of bond types. For decalin, these numbers are 591, 49, and 1, respectively (naphthalene and quinoline, which have the same topology as decalin, also have a graph ID number of 591). Those three numbers are combined into a unique Ring Identifier (RID).[19] For decalin, 591.49.1 is the RID. With every ring system in the file is stored information about its SSSR. The algorithms used by CAS for SSSR perception have been described in detail.[20,21]

As of the end of 1998, the CAS ring file contained 474 859 different ring systems. From these ring systems were selected those that contain as ring atoms only the elements B, C, Si, N, P, As, O, S, Se, Te, and the halogens. This step eliminated all metal-containing ring systems as well as systems with bridging H atoms. This left a file of 290 634 chemical rings (these rings are not strictly "organic" since 2973 contain no carbon). A nonredundant set of all the ring-system graphs (i.e., graph ID numbers) in this file was obtained. The result was a database of 40 182 different ring topologies.

Population Statistics of Cells and Bins. Descriptors *P*, *B*, and *R* were calculated for every ring topology in the database, and each topology was put into the appropriate cell in *PBR*-space. The entire database is distributed among 7387 cells. This distribution is quite uneven, as seen in Figure 3. This plot shows the number of cells as a function of cell population (the population of a cell being the number of ring topologies it contains). The leftmost point corresponds to the 4040 cells that have a population of one, i.e., contain a single topology. Most cells contain only a few topologies; 90% of the cells have a population of less than 10. There are, however, a number of highly populated cells; 60 cells have a population of 100 or more, and these cells contain almost 25% of all topologies. In Table 1 are listed the 20 largest cell populations. The *P*, *B*, *R* coordinates of these cells indicate they are fairly close to one another in descriptor space.

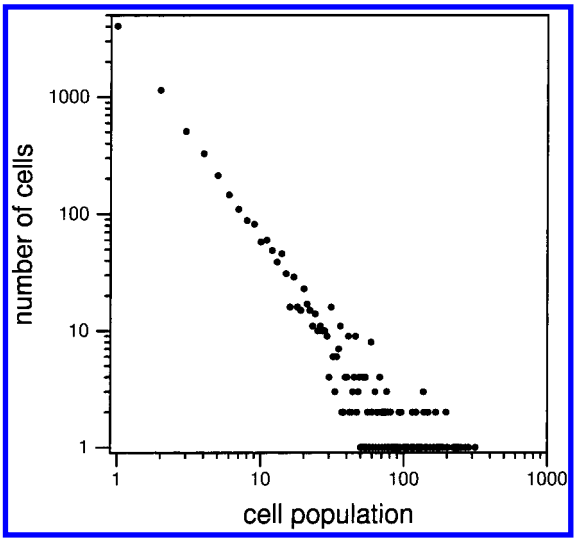


Figure 3. Population statistics for the cells in *PBR*-space occupied by the ring-topology database.

Table 1. Cells Containing the Highest Numbers of Ring Topologies

cell	cell population	cell	cell population
(20, 4, 5)	315	(15, 4, 4)	223
(19, 4, 5)	283	(17, 5, 5)	222
(18, 5, 5)	270	(23, 5, 6)	202
(18, 4, 5)	252	(22, 4, 5)	201
(19, 5, 5)	243	(14, 4, 4)	200
(24, 5, 6)	239	(17, 6, 5)	197
(21, 4, 5)	235	(14, 3, 4)	197
(15, 3, 4)	230	(25, 5, 6)	187
(16, 6, 5)	227	(22, 6, 6)	179
(16, 3, 4)	226	(17, 3, 4)	176

Table 2. Bins Containing the Highest Numbers of Ring Topologies

cell	bin	bin population
(20, 4, 5)	{5, 5, 6, 6, 6}	223
(18, 5, 5)	{5, 5, 6, 6, 6}	189
(24, 5, 6)	{5, 5, 6, 6, 6, 6}	187
(16, 6, 5)	{5, 5, 6, 6, 6}	155
(25, 5, 6)	{5, 6, 6, 6, 6, 6}	154
(19, 5, 5)	{5, 6, 6, 6, 6}	150
(22, 6, 6)	{5, 5, 6, 6, 6, 6}	148
(22, 3, 5)	{5, 5, 6, 6, 6}	135
(17, 6, 5)	{5, 6, 6, 6, 6}	134
(26, 4, 6)	{5, 5, 6, 6, 6, 6}	134

As previously noted, the ring topologies within a cell can be binned according to their sets of SSSR ring sizes. When this is done, the database is found to be distributed among 14 405 bins, 9911 of which have a population of one. There are 22 bins with a population greater than 100, each from a different cell. For these 22 bins, the SSSR consists only of five- or six-membered rings, presumably a result of the prevalence of these ring sizes in the Registry File.[22] The 10 largest bins are listed in Table 2. As this table shows, the same set of SSSR ring sizes may appear in more than one cell (there are only 7711 different sets of ring sizes in the database compared to 14 405 bins). A complete breakdown of the largest cell into bins is given in Table 3. Most cells

Table 3. Distribution of Ring Topologies among the Bins of the (20, 4, 5) Cell

bin	bin population	bin	bin population
{5, 5, 6, 6, 6}	223	{3, 3, 5, 6, 11}	3
{4, 6, 6, 6, 6}	30	{4, 5, 5, 6, 8}	1
{5, 5, 5, 6, 7}	18	{3, 5, 6, 7, 7}	1
{3, 5, 6, 6, 8}	11	{3, 5, 5, 7, 8}	1
{4, 5, 6, 6, 7}	7	{3, 5, 5, 6, 9}	1
{4, 4, 6, 6, 8}	6	{3, 4, 5, 8, 8}	1
{5, 5, 5, 5, 8}	5	{3, 3, 7, 7, 8}	1
{3, 6, 6, 6, 7}	5	{3, 3, 3, 5, 14}	1

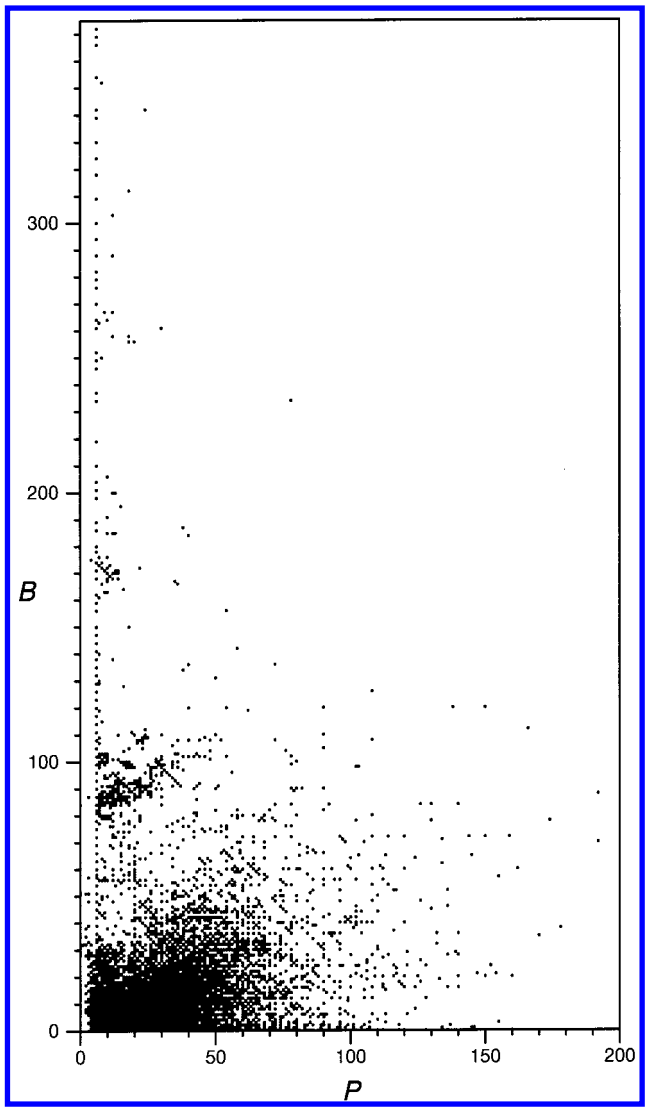


Figure 4. A two-dimensional view of the distribution in *PBR*-space of the cells populated by the ring-topology database. Only the *P* and *B* coordinates of each cell are plotted.

have far fewer bins than this example; 91% of the cells have fewer than five bins.

Spatial Distribution of Cells. One way to obtain a simple overview of the distribution of the populated cells in *PBR*-space is by plotting these cells using their *P* and *B* coordinates only. This is shown in Figure 4. The neglect of the *R* coordinate means that a single point may represent many different cells. The two-dimensional pattern in this plot can be interpreted as the "shadow" cast onto the *PB*-plane by the three-dimensional distribution of populated cells. The highest values of *P* and *B* found in the database are 240 and 372, respectively (Figure 4 does not go out to $P = 240$ as there is only one occupied cell with $P > 200$). The highest value of *R* is 127. These values may seem rather high, but the Registry File contains some very large ring systems, including theoretical ring systems that have appeared in the literature.

The most noticeable feature in Figure 4 is the long column of cells at $P = 6$. This feature appears to be due primarily to the large class of fullerene cage compounds; the topology of buckminsterfullerene (C_{60}) falls into the (6, 84, 31) cell. This feature is only a two-dimensional projection of a line of cells in the three-dimensional space. A mathematical description of this line can be easily derived. In the graphs of fullerene ring systems, every node has a degree of three, i.e., is connected to three other nodes. It can be shown that such graphs obey the relation $R = E/3 + 1$;[23] for the case of $P = 6$, this becomes $R = B/3 + 3$. Hence, the fullerene-associated cells which lie in the $P = 6$ plane of *PBR*-space fall on the line described by the equation $R = B/3 + 3$.

The distribution of populated cells is examined in more detail in Figure 5. This figure focuses on the subset of *PBR*-space for which $P \leq 50$ and $B \leq 50$. As in Figure 4, cells are plotted using their *P* and *B* coordinates only. Each of the three distributions in Figure 5 is for all cells with populations in a particular range. In going from low populations (Figure 5a) to moderately high populations (Figure 5b), the region occupied by the cells is considerably reduced. For very high populations (Figure 5c), the cells are concentrated in a very small region. For these cells, containing 100 or more ring topologies, the value of *R* is confined to the range $4 \leq R \leq 7$. Thus, the most highly populated cells in *PBR*-space are fairly close together (as was suggested by the cell coordinates listed in Table 1).

Why it is that this particular region of *PBR*-space contains the most populated cells is an interesting question, though one that cannot be answered definitively. Synthetic feasibility is probably a factor. The topologies in these cells are reasonably small in overall size; the number of edges, $P + B$, is less than 30 for most of the cells in Figure 5c. Also, these topologies are not extremely complex insofar as $B < P$ for all of the cells in Figure 5c. These attributes enhance the ease of synthesis. On the other hand, the number of rings in these topologies (i.e., four to seven) seems somewhat high with respect to ease of synthesis. The overriding factor here may be simple combinatorics. Ring systems with more rings allow many more structural possibilities to choose from than systems with fewer rings.

TOPOLOGICAL SIMILARITY AND *PBR*-SPACE

A primary motivation for defining and studying *PBR*-space was the expectation that similar ring topologies would tend to be closer together in this space than ring topologies having little similarity. In particular, it was expected that highly similar ring topologies would fall into the same cell or into neighboring cells. One way to test this is to select some ring-system graph as a target and see if similar graphs can be found in the region of *PBR*-space closest to the target. As an example, a target was chosen from the (12, 2, 2) cell. Figure 6a shows this graph surrounded by four other graphs that are highly similar to it. Each of these four graphs was found in one of the cells adjacent to (12, 2, 2) in the $R = 2$ plane. This is certainly a case of very similar ring topologies being in close proximity in *PBR*-space.

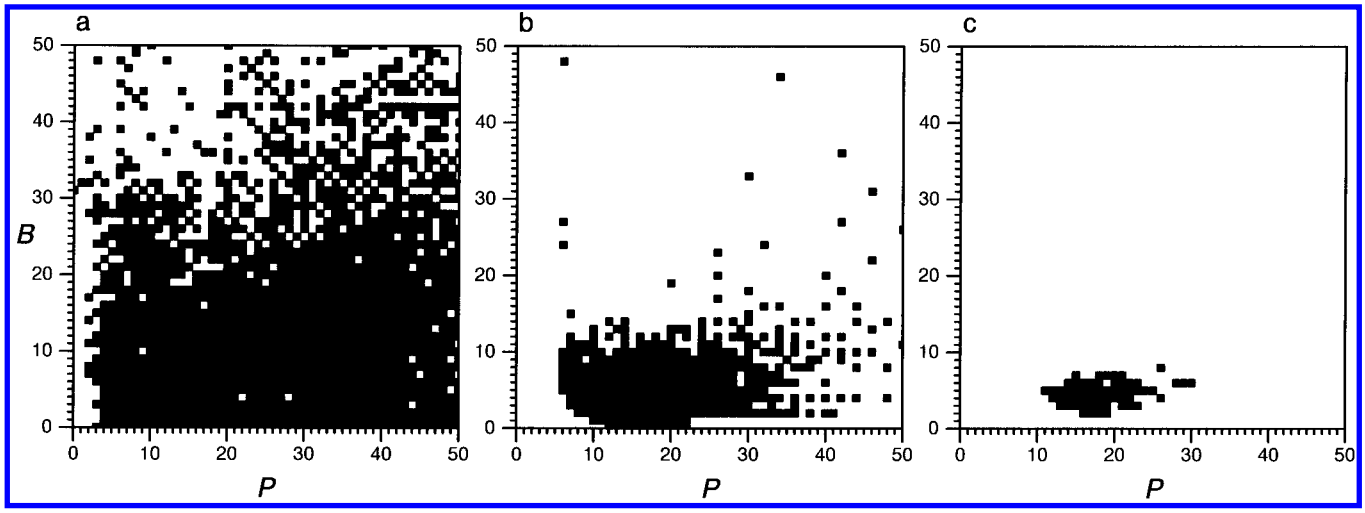


Figure 5. Distribution of populated cells in a subset of *PBR*-space ($P \leq 50$, $B \leq 50$). In each figure are plotted the *P* and *B* coordinates of all cells with populations in a given range: (a) cells containing 1−9 topologies; (b) cells containing 10−99 topologies; and (c) cells containing 100 or more topologies.

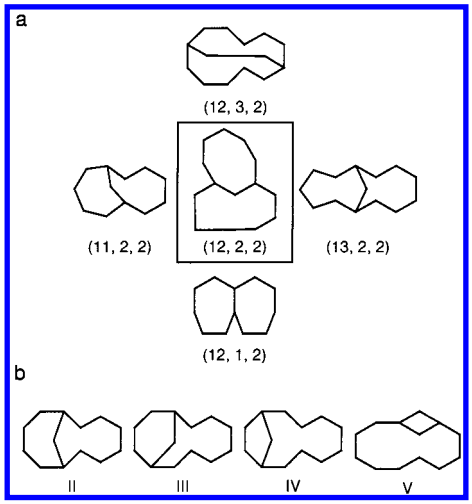


Figure 6. Similarity and proximity in *PBR*-space: (a) a ring topology (in box) from the (12, 2, 2) cell and four similar topologies from adjacent cells and (b) the four other ring topologies in the (12, 2, 2) cell.

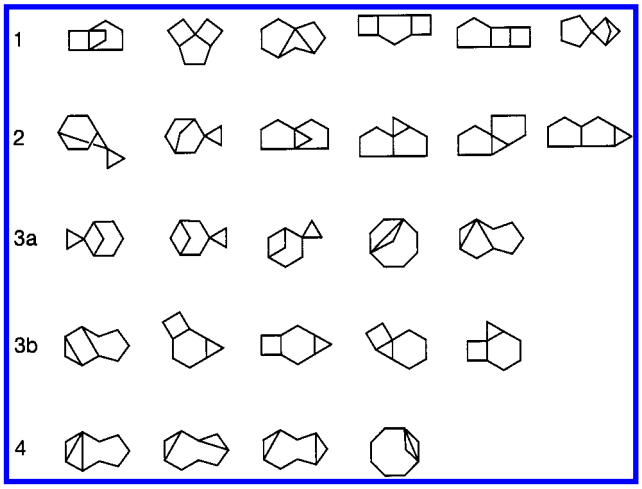
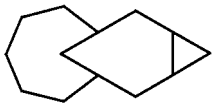


Figure 7. All ring topologies in the (9, 2, 3) cell. Rows 1−4 correspond to the different bins in this cell. These bins are {4, 4, 5} (row 1), {3, 5, 5} (row 2), {3, 4, 6} (rows 3a−b), and {3, 3, 7} (row 4).

The (12, 2, 2) cell also contains four other ring-system graphs besides the target, and these are shown in Figure 6b. These graphs are in equal proximity to the target, being in the same cell, but they are not equally similar to it. One of them, graph II, is a good similarity match to the target, but in comparison, graph V is a poor match. This illustrates the general observation that, while similar ring topologies are relatively close together in *PBR*-space (as in Figure 6a), proximity alone does not imply similarity. This is a consequence of the limited ability of *P*, *B*, and *R* to discriminate between ring topologies.

Ring topologies similar to a target need not come only from nearby cells with the same value of *R* as the target. Two topologies may be similar despite a difference in ring count. For example, in the (12, 3, 3) cell is found the following topology



which is very similar to the target from the (12, 2, 2) cell shown in Figure 6a, the only difference being the one-bond bridge that results in a three-membered ring. It is expected, however, that distance along the *R* direction will not necessarily have the same relationship to similarity as distance along the *P* and *B* directions. As one moves away from the cell of a target, the chance of finding good similars to the target will probably fall off much faster in the *R* direction than in the *P* or *B* directions.

A more populated region of *PBR*-space is examined in Figure 7. This shows all the ring topologies contained in the (9, 2, 3) cell. These topologies have fewer edges than those in the (12, 2, 2) cell, but they display much more diversity as a consequence of having three rings. These topologies are distributed among four bins, as indicated in Figure 7. This offers an opportunity to examine the relationship between bin assignment and similarity for same-cell topologies. It can be seen that each of the bins in this cell contains some sets of topologies that are highly similar to one another. However, ring topologies in the same bin are not always equally similar. There are a number of topologies from different bins that are more similar than some topologies in the same bin. Hence, bin assignment is like cell assignment

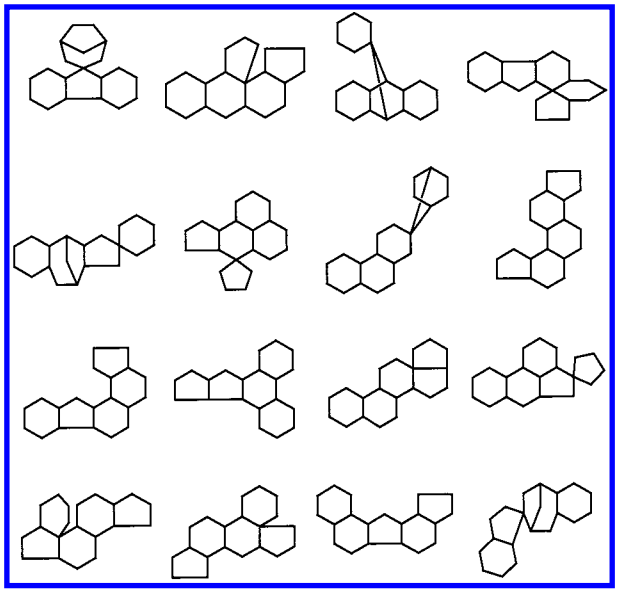


Figure 8. Random sample of the 223 ring topologies in the {5, 5, 6, 6, 6} bin of the (20, 4, 5) cell.

in regard to similarity: very similar topologies may fall into the same bin, but presence in the same bin does not ensure high similarity.

The ability to bring together ring topologies that are similar but which have different sets of SSSR ring sizes, as in Figure 7, is an important aspect of *PBR*-space. It is also important to separate topologies that have the same set of SSSR ring sizes but which are dissimilar. For example, there are 18 topologies in the database with SSSR ring sizes of 4, 4, and 5, and they are distributed among several cells:[24] (5, 4, 3), (7, 3, 3), (9, 2, 3), (11, 1, 3), and (13, 0, 3). The {4, 4, 5} bin of the (5, 4, 3) cell contains the following topologies.

The {4, 4, 5} bin of the (9, 2, 3) cell contains the topologies shown in row 1 of Figure 7. The {4, 4, 5} bin of the (13, 0, 3) cell contains four all-spiro systems. The separation of these three sets of topologies in different cells seems reasonable based on the levels of dissimilarity between them.

A bin with a very high population is illustrated in Figure 8 through a sampling of its contents. The ring topologies displayed were randomly selected from the {5, 5, 6, 6, 6} bin of the (20, 4, 5) cell. This is the largest bin in the database, as shown in Table 2. It is clear that the intrinsic "resolution" of *PBR*-space is not sufficient to separate these topologies. It would be helpful if a very large bin like this could be subdivided into smaller groups. This could be done with the use of additional topological information. For instance, Figure 8 suggests that the presence or absence of a spiro ring fusion might be an appropriate feature for further partitioning of this bin. There are many possible graph descriptors that might be useful for subdividing bins.

The cells that have been examined in Figures 6−8 are from the part of *PBR*-space in which $B < P$. Ring topologies of greater complexity will be found in cells for which $B > P$. A cell from this part of *PBR*-space is shown in Figure 9.

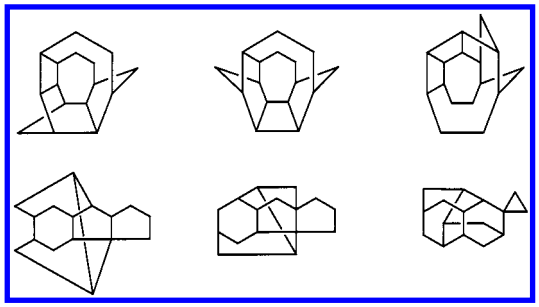


Figure 9. All ring topologies in the (9, 12, 6) cell.

As expected, its topologies are fairly complex. The inspection of complicated ring diagrams such as these for similarity can be quite difficult. The ability to perceive similarity in complex rings may be very dependent on the way these rings are drawn. In Figure 9, similarity perception is made much easier by the fact that similar topologies have been algorithmically drawn using similar templates.

USING *PBR*-SPACE TO STUDY RING DIVERSITY

The basic concern of diversity analysis is analyzing the way in which a given set of chemical structures is distributed over the "space" of all chemical structures. Using *PBR*-space, diversity analysis can be focused specifically on the topologies of chemical rings. In this section, two interesting issues regarding ring diversity are examined. The first is the significance of empty regions, or voids, in *PBR*-space. The second is the distribution pattern in *PBR*-space of rings associated with medicinal applications.

Voids in *PBR*-Space. An important use of diversity analysis is the comparison of compound databases. Such comparisons are often concerned with deciding whether acquisitions can be made from database *A* that will enhance the diversity of database *B*.[25] One way to address this question is by distributing both databases in some mathematical space and determining the extent to which regions occupied by *A* are aligned with regions not occupied by *B*; a cell-based space is well suited to this type of comparison.[26] *PBR*-space could be used in this way to compare the diversity of databases according to their ring content, a characteristic that is especially important for databases of pharmaceutical interest.[3] For instance, the ring content of a compound library used for bioactivity screening could be compared to a large database of chemical rings derived from the CAS Registry File. Acquisitions to enhance the ring diversity of the library might be found by examining those cells (or bins) that contain rings from the Registry File but no rings from the library.

While a comprehensive database of chemical rings derived from the Registry File could be used to fill voids in other ring files, such a database will itself have voids in *PBR*-space. These voids have special significance. Since they cannot be filled by any known rings—assuming the database represents the set of all known chemical rings—these empty regions indicate opportunities for the creation of ring topologies that are new to chemistry.[27] Empty cells that lie near a number of occupied cells are particularly interesting in this respect since it may be possible to relate them to some existing strategies for ring synthesis.

An attempt to visualize voids in the ring-topology database is shown in Figure 10. For the purpose of this plot, a void

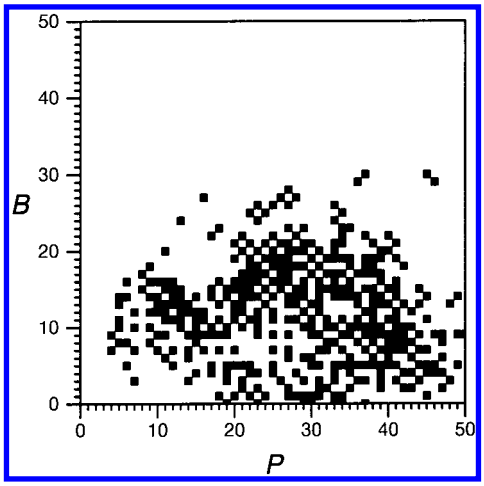


Figure 10. Distribution of voids in a subset ($P \leq 50$, $B \leq 50$, $R \leq 8$) of *PBR*-space, where a void is any empty cell with a minimum number of occupied adjacent cells (see text for details). Only the *P* and *B* coordinates of each void are plotted.

was defined as any empty cell, (*P*, *B*, *R*), for which at least three of the four adjacent cells, ($P - 1$, *B*, *R*), ($P + 1$, *B*, *R*), (*P*, $B - 1$, *R*), (*P*, $B + 1$, *R*), are occupied. This definition is fairly restrictive in that it requires a void to be somewhat surrounded by occupied cells. As a result, not many voids will be found according to this definition if the occupied cells are grouped very compactly. A total of 486 voids was found in the subset of *PBR*-space examined in Figure 10. This number is rather large and indicates that the grouping of occupied cells is not very compact even in this highly populated region of *PBR*-space.

The large number of voids in the distribution of the ring-topology database in *PBR*-space raises questions about why these cells are empty, but such questions may have no obvious answer. The fact that these voids are, by the above definition, close to occupied cells suggests they are not empty strictly as a result of synthetic obstacles. For instance, the (14, 3, 5) cell is empty, but there are ring topologies in adjacent cells: e.g., there are three topologies in (13, 3, 5) and five in (14, 2, 5). Though it is surprising that no topology in the database falls into the (14, 3, 5) cell, this cannot be attributed to any underlying chemical or graph-theoretical constraint.

The detection of voids in *PBR*-space combined with an algorithm for generating topologies to fill them could be a useful tool in the creation of new ring systems. It should be possible to write an algorithm to generate all ring topologies satisfying the *P*, *B*, *R* coordinates of a given cell.[28] In this way, empty cells in *PBR*-space could be filled. Recently, work along similar lines was carried out by Wife and co-workers, who developed a program to generate all topologies having a specific ring index, where this index consists of an ordered list of the sizes of all cycles in the ring system.[29] An algorithm that generates all the ring topologies for any empty cell in *PBR*-space could also be applied to a populated cell, in which case it would generate the known topologies currently found in the cell as well as those which belong in the cell but are unknown to chemistry. This would make it possible to look for new ring topologies even in the most highly populated regions of *PBR*-space.

Medicinally Relevant Rings. To study the distribution in *PBR*-space of medicinally relevant rings, an appropriate

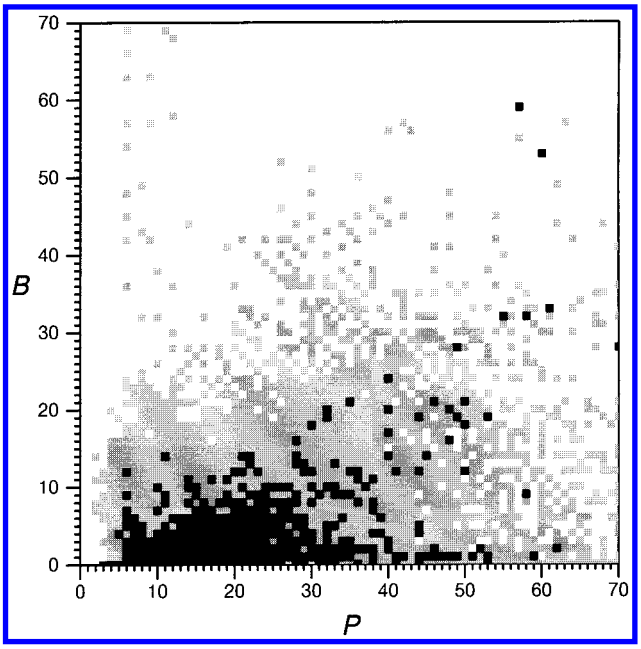


Figure 11. Distribution in *PBR*-space of a medicinally relevant subset of the ring-topology database. The *P* and *B* coordinates of the cells occupied by this subset are shown in black. For comparison, a distribution of ring topologies from the full database is shown in gray.

subset of the ring-topology database was selected. The selection of this subset began with a file of all substances with connection tables in the CAS Registry File that had been assigned the "Therapeutic Use" role at least twice.[30] This file had more than 99 000 substances, containing 3716 different ring systems with 930 different topologies. Some of these topologies came from metal-containing systems and so were not present in the ring-topology database. Of the 851 topologies that were in the database, only those with 10 or fewer rings were retained in order to remove from consideration some large cage systems (boranes, fullerenes), which are not typical of known drugs. The remaining 832 topologies were taken as the medicinally relevant subset of the database.

The distribution of this subset in *PBR*-space—more accurately, the two-dimensional "shadow" of this distribution on the *PB*-plane—is shown in Figure 11 (black boxes). Also shown, for comparison, is the distribution of *all* database topologies with 10 or fewer rings (gray boxes). The fact that the medicinally relevant subset is confined to a relatively small number of cells is expected. What is noteworthy is the nature of this confinement and its relationship to certain structural constraints.

There are two approximate bounds on the two-dimensional pattern produced by the medicinal subset. The first is the line $B = (2/3)P$. More than 95% of the cells occupied by the subset fall below this line. Only a few of the cells are in the region where $B > P$. This implies a limit on the topological complexity of these medicinally relevant rings. It is not due to any intrinsic limit on ring complexity since the gray-colored distribution in Figure 11 indicative of all database topologies shows that many ring systems with more complex topologies do exist. However, complex ring systems can be hard to synthesize and/or derivatize, and so this limit on ring complexity probably reflects the importance of synthetic accessibility in the design of medicinal compounds.

The second approximate bound is the line $P + B = 40$. More than 80% of the cells occupied by the medicinal subset fall below this line; these cells contain ring topologies with fewer than 40 edges. This bound can be interpreted as a rough limit on ring-system size. It does not translate directly into a limit on molecular size since side chains and additional rings are not taken into account. Nevertheless, this falloff around 40 edges is consistent with the proposed rule that compounds having a molecular weight greater than 500 are unlikely to be good drug candidates due to bioavailability problems, i.e., poor solubility and permeability.[31] Thus, the pattern in Figure 11 produced by the medicinal subset appears to show the influence of two important constraints on drug molecules: synthetic accessibility and bioavailability.

CONCLUSIONS

As a tool to explore chemical rings based on their topology, the descriptor space presented here has several desirable attributes. First, *PBR*-space has low dimensionality. With only three dimensions, it is easy to visualize, and a distribution pattern in this space could be unambiguously represented using interactive graphics (of course it will often be more convenient to project the pattern onto a two-dimensional plane with some loss of information, as done in this paper). Second, *PBR*-space uses simple and understandable descriptors. This facilitates browsing by making the correlation between incremental movements in the space and changes in ring topology more comprehensible. Third, *PBR*-space is composed of discrete cells. This makes the space very suitable for diversity analysis, as illustrated in the previous section.

It is instructive to compare the use of *PBR*-space with the alternative of simply indexing ring topologies by their SSSR ring sizes. In comparison to this alternative, the advantages of *PBR*-space are (i) it separates topologies with the same ring sizes but which are dissimilar and (ii) it brings together topologies with different ring sizes but which are similar. With regard to ring sizes and similarity, note that a high degree of similarity between ring topologies is unlikely if their ring sizes are too different. For example, consider the sets of ring sizes listed in Table 3. The {5, 5, 6, 6, 6} and {4, 6, 6, 6, 6} bins are presumably much more likely to contain very similar topologies than the {5, 5, 6, 6, 6} and {3, 3, 3, 5, 14} bins. Since there are ways to compare two "vectors" of ring sizes in a quantitative manner, e.g., the Euclidean distance, it is possible this observation could be used in some way to constrain similarity searching in *PBR*-space.

An important aspect of this study is the considerable size of the ring-topology database used to populate *PBR*-space: more than 40 000 topologies, derived from a comprehensive collection of about 290 000 organic chemical rings. Because this database has so many closely related topologies, the ability of *PBR*-space to bring together similar topologies was clearly demonstrated. Just as important, the fact that ring topologies which are *not* highly similar may fall into the same cell or bin was also demonstrated. This is an intrinsic shortcoming of *PBR*-space and reflects the limited discriminating ability of the *P*, *B*, and *R* descriptors. The use of additional graph descriptors to further differentiate between topologies would help compensate for this.

The database that has been used here is a good representation of the set of all ring topologies known to organic chemistry. The classification of this database in *PBR*-space thus provides an opportunity to investigate the existing topological diversity of organic rings. The ability to conduct such studies was one of the motivations for developing this descriptor space. In this paper, a very simple study of ring diversity has been presented. An interesting finding is that the distribution of the database in *PBR*-space, rather than being highly compact, appears to have many significant voids. The analysis of such voids may have practical application in the search for new and novel ring systems, a topic of some interest in drug discovery. This and other aspects of chemical ring diversity deserve further examination.

ACKNOWLEDGMENT

I thank Steven Layten and Susan Funk for their technical assistance in preparing the database of ring topologies. I also thank Qiong Yuan and William Fisanick for making available the file of "Therapeutic Use" substances.

REFERENCES AND NOTES

(1) Wermuth, C. G. Ring Transformations. In *The Practice of Medicinal Chemistry*; Wermuth, C. G., Ed.; Academic Press: London, 1996; pp 239−260.
(2) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(3) Nilakantan, R.; Bauman, N.; Haraki, K. S. Database Diversity Assessment: New Ideas, Concepts, and Tools. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 447−452.
(4) Lipkus, A. H. Mining a Large Database for Peptidomimetic Ring Structures Using a Topological Index. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 582−586.
(5) Bonchev, D.; Mekenyan, O.; Trinajstić, N. Topological Characterization of Cyclic Structures. *Int. J. Quantum Chem.* **1980**, *17*, 845−893.
(6) Mekenyan, O.; Bonchev, D.; Trinajstić, N. Algebraic Characterization of Bridged Polycyclic Compounds. *Int. J. Quantum Chem.* **1981**, *19*, 929−955.
(7) Randić, M. Ring ID Numbers. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 142−147.
(8) Nilakantan, R.; Bauman, N.; Haraki, K. S.; Venkataraghavan, R. A Ring-Based Chemical Structural Query System: Use of a Novel Ring-Complexity Heuristic. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 65−68.
(9) Bonchev, D.; Balaban, A. T.; Liu, X.; Klein, D. J. Molecular Cyclicity and Centricity of Polycyclic Graphs. I. Cyclicity Based on Resistance Distances or Reciprocal Distances. *Int. J. Quantum Chem.* **1994**, *50*, 1−20.
(10) Lipkus, A. H. A Ring-Imbedding Index and Its Use in Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 92−97.
(11) Randić, M. On Characterization of Cyclic Structures. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1063−1071.
(12) Wife, R. L. Identifying and Filling Holes in Diversity Space. Presented at the CHI meeting on Chemoinformatics, Arlington, VA, May 1997.
(13) Pisanski, T.; Plavšić, D.; Randić, M. On Numerical Characterization of Cyclicity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 520−523.
(14) Downs, G. M. Ring Perception. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: Chichester, 1998; Vol. 4, pp 2509−2515.
(15) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Review of Ring Perception Algorithms for Chemical Graphs. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 172−187.
(16) The extent to which the rings of the SSSR share *nodes* should also correlate with topological complexity. In fact, a measure of ring complexity based on this idea has been suggested. This measure is the sum of the number of atoms in every SSSR ring divided by the number of atoms in the ring system. For details, see: Gasteiger, J.; Jochum, C. An Algorithm for the Perception of Synthetically Important Rings. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 43−48.
(17) For instance, consider a graph of n nodes in which each node is connected to every other node (the so-called complete graph of order n). For this graph, $E = n(n-1)/2$, and so $R = (n-1)(n-2)/2$. Since all the rings in the SSSR are triangles, $S = 3R$. It follows from eq 1 that $P = -(n-1)(n-6)/2$. Thus, $P < 0$ if $n > 6$.

(18) Pinkos, R.; Weiler, A.; Voss, T.; Weber, K.; Wahl, F.; Melder, J.-P.; Fritz, H.; Hunkler, D.; Prinzbach, H. From pagodanes to homologous, nonpentagonal dodecahedranes. *Liebigs Ann./Recl*. **1997**, *10*, 2069−2088.

(19) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. The Chemical Abstracts Service Chemical Registry System. I. General Design. *J. Chem. Inf. Comput. Sci*. **1976**, *16*, 111−121.

(20) Zamora, A. An Algorithm for Finding the Smallest Set of Smallest Rings. *J. Chem. Inf. Comput. Sci*. **1976**, *16*, 40−43.

(21) Qian, C.; Fisanick, W.; Hartzler, D. E.; Chapman, S. W. Enhanced Algorithm for Finding the Smallest Set of Smallest Rings. *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 105−110.

(22) Stobaugh, R. E. Chemical Abstracts Service Chemical Registry System. 11. Substance-Related Statistics: Update and Additions. *J. Chem. Inf. Comput. Sci*. **1988**, *28*, 180−187.

(23) This follows from $R = E - N + 1$ and $3N = 2E$. The latter is simply a statement of the fact that the sum of the degrees of the nodes in a graph equals twice the number of edges.

(24) Note that it is possible to jump from any of these cells to the next by increasing (decreasing) P by two and simultaneously decreasing (increasing) B by one; such jumps do not change the value of S since $S = P + 2B$. In other words, these cells fall on a line of slope $-1/2$ in the $R = 3$ plane. In any plane of constant R, all cells that have the same value of S fall on a line of slope $-1/2$.

(25) Shemetulskis, N. E.; Dunbar, J. B., Jr.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the Diversity of a Corporate Database Using Chemical Database Clustering and Analysis. *J. Comput.-Aided Mol. Des*. **1995**, *9*, 407−416.

(26) Pearlman, R. S.; Smith, K. M. Software for Chemical Diversity in the Context of Accelerated Drug Discovery. *Drugs Future* **1998**, *23*, 885−895.

(27) There are regions of *PBR*-space that are empty because of graph-theoretical constraints. In any plane of constant R, there can be no occupied cells below the line $P + B = \epsilon(R)$, where $\epsilon(R)$ is the minimum number of edges a ring-system graph must contain in order to have R rings.

(28) Such an algorithm might be simplified if all the possible bins in the cell were generated first and the algorithm could then separately generate topologies for each set of ring sizes. All bins could be found using partition generation (e.g., see: Nijenhuis, A.; Wilf, H. S. *Combinatorial Algorithms*; Academic: New York, 1975; pp 63−69) since every possible set of ring sizes for a cell is a partition of S ($= P + 2B$) into R integers not smaller than 3. However, it may not be possible to generate ring topologies for each of these partitions. For a given partition, there may not exist any topology that has both the partition as its ring-size set *and* the required values of P and B.

(29) Information on this program, called SORT&gen, is available at http://www.specs.net.

(30) This role is assigned to substances in the case of "demonstrated, claimed, or clearly intended application or formulation of the substance for medical or veterinary use in therapy, prophylaxis, or diagnosis" [*CAS Roles on STN User Guide* (available at http://www.cas.org/ONLINE/UG/roles.pdf); Chemical Abstracts Service, 1995; p 22].

(31) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev*. **1997**, *23*, 3−25.

CI000144X