CONCORD and CAMBRIDGE: Comparison of Computer-Generated Chemical Structures with X-ray Crystallographic Data

Marissa A. Hendrickson, Marc C. Nicklaus, and George W. A. Milne*

Laboratory of Medicinal Chemistry, DTP, DCT, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892

D. Zaharevitz

PRI-Dyncorp, Inc., NCI/Frederick Cancer Research and Development Center, Frederick, Maryland 21702

Received July 7, 1992

The structures of a number of molecules as determined by X-ray crystallography have been compared with the structures for the same molecules as calculated by the 3D structure generation program, CONCORD. In 41% of the cases, the calculated structures were essentially identical to those measured by X-ray diffraction. In most of the remainder, there were significant differences arising primarily from the inability of the structure generation program to predict the correct torsion angle at one or more rotatable bonds. The implications of these findings on the process of 3D database generation are discussed.

INTRODUCTION

While two-dimensional substructure searching has been an established and widely used technique in chemical data management for over 15 years, it is only in the last 5 years that significant efforts have been made to build databases of three-dimensional atomic coordinate data along with programs that can search them. A major impetus behind these efforts is the realization that a *pharmacophore*—that part of a drug molecule which is responsible for its biological activity[1]—is a three-dimensional entity, and the molecule's biological action is expressed by interaction between the pharmacophore and the active site in an enzyme, which is also a three-dimensional array of atoms.[2]

The study of chemical structures in two dimensions occasionally can account for the observed structure–activity relationships, but this is only when the two- and three-dimensional structures are similar, as, for example, in the polycyclic aromatic hydrocarbons. In most cases, two-dimensional structures fail to account properly for the behavior of compounds with respect to biological systems, and consequently a number of research groups, many of whom are involved in medicinal chemistry,[3-7] have begun to make use of pharmacophore searching in three dimensions.

There are two aspects to the problem of searching of three-dimensional databases. The first is the source of the database, and the second is the need to develop software that can search such databases efficiently. Willett's group[8] has taken the lead in the software development area, and many systems, including some that are commercially available,[9,10] use programs which are based upon those developed by this group. The development of three-dimensional databases, on the other hand, is a matter of active concern to numerous groups, and the validation of programs that are used for this purpose is the subject of this paper.

METHODS

The program CONCORD[11,12] is a popular means of creating a three-dimensional database from the corresponding two-dimensional file. Starting with a file of two-dimensional structures in any of several standard formats such as connection tables, it generates three-dimensional atomic coordinates for each structure. The program relies on a combination of expert system (rule-based) procedures and pseudomolecular mechanics.[11] It uses standard bond angles and bond lengths, but does not attempt energy calculations. The program does however determine bond angles and torsion angles by minimization of an internal strain function.[11] It takes internal hydrogen bonds into account to some extent, but it fails to calculate long-range electrostatic interactions. CONCORD's strength lies in its speed; it is at least 10 times faster than energy-based algorithms[13] and can process 1–10 structures per second on a VAX 9210 for example. This makes it ideal for use with large databases, and it has been used in numerous systems.[14] In 1988, Chemical Abstracts Service used CONCORD in an attempt to convert the 11 million compound Registry to a three-dimensional database which can be searched online. This was not totally successful; for a variety of reasons, which will be discussed in this paper, three-dimensional coordinates could be derived for only about half of the structures in the Registry. This file is available for online access,[15] although atomic coordinate searching is not yet supported.

CONCORD's major weakness is that it is an approximate method. It manages rigid structures very well but, faced with different conformational possibilities, it is often forced into ad hoc decisions. In order to assess the reliability of three-dimensional databases generated by CONCORD, it was decided to use it to convert a large database[16] of compounds of interest to the National Cancer Institute to a three-dimensional version and compare structures from it with the corresponding structures as defined by X-ray crystallographic measurements.

Since 1955, the NCI has examined approximately 450 000 compounds for anticancer activity. Since 1988, some 20 000 of the compounds have also been examined for anti-HIV activity. The yield of clinical antitumor agents discovered by this program has been very low,[17] but an important aspect of this group of compounds is that it contains approximately 4% of all known organic compounds[18] and is thus a large and possibly statistically significant subset of all structures published. It is, therefore, a useful database in which to search for compounds containing pharmacophores that are thought to be involved in some specific antitumor activity.

0095-2338/93/1633-0155$04.00/0

The most widely used source of X-ray crystallographic data is the Cambridge Structural Database (CSD),[19] which is maintained by the Cambridge Crystallographic Data Centre. This database currently contains some 90 000 organic compounds whose structures have been determined by X-ray diffraction methods.

RESULTS

1. CONCORD Processing of DIS Database. The 448 557 structures in the Drug Information Service (DIS) as of January 21, 1992, were processed with CONCORD version 2.9.3.[20] Because CONCORD has access to the necessary parameters only for H, C, N, O, F, Si, P, S, Cl, Br, and I, some 51 234 compounds containing other elements were rejected by the program. A three-dimensional database containing 396 789 structures was thus obtained from the CONCORD processing. The close contact ratio (CCR), defined as the ratio between the actual interatomic distance and a minimum acceptable distance determined by CONCORD, was greater than 1.0 for 69 of the 90 structures in the test set (see next section) and lay between 0.5 and 1.0 for the 21 remaining structures. Of these 21, 48% had >3 rotatable bonds. CONCORD thus regarded all 90 structures as acceptable, but issued a "CCR warning" for 21 of them.

Each of the structures produced by CONCORD was stored in the "PDB" format used by the Protein Data Bank.[21] This format is acceptable as input to QUANTA,[22] a molecular modeling program which can compute the similarity between different structures and which was used to compare the structures produced by CONCORD with those in the Cambridge file.

X-ray-derived structures from the Cambridge Structural Database (CSD) are output by the search program QUEST, which is provided by the Cambridge Crystallographic Data Centre, in an "FDAT" format which cannot be read directly by QUANTA. A program was written in FORTRAN77 to convert the FDAT format to the "fractional coordinate" format read by QUANTA version 3.2.4. [While recent versions of QUANTA read the fractional coordinates format, older versions (3.0 or older) read a "Cambridge" format. Both formats are characterized by the .xr file extension and are essentially identical. The conversion program referred to here is available upon request from the authors.]

2. Overlap between DIS and Cambridge File. Chemical Abstracts Service Registry numbers (CAS RN) are stored for a portion of the compounds in the DIS, and the same is true for the Cambridge Structural Database. In the DIS, 196 441 compounds (43.79% of the total) have CAS RN,[23] while in the Cambridge file, the proportion is about 12%.[24] A total of 194 structures with CAS RN were found in both files. Of these, CONCORD could only process 134 successfully, most of the failures being due to "heavy" atoms. Of the 134 structures, 27, although present in the CSD, had no useful atomic coordinate data. A further 17 compounds could not be compared because CONCORD was not provided with the correct stereochemistry (see Discussion, section 3). This left 90 compounds which were present in both databases and which could be used in the subsequent comparisons. Certainly there must be more compounds common to both files, but given the low proportion of compounds with CAS RN, particularly in the Cambridge file, it is not possible without great difficulty to identify them. The 90-compound subset is probably not large enough to be statistically significant, and so the statistics that we cite, the percentages of different classes of compounds whose CONCORD and X-ray structures match with different RMS values, should be treated with care. Analysis of various cases, as is done below, however, permits some useful general conclusions to be drawn, as will be seen.

Table I. Occurrence of Rotatable Bonds in Different Databases

	0 rotors (%)	1 rotor (%)	2 rotors (%)	3 rotors (%)	>3 rotors (%)
Cambridge	12	(33)	(28)	(10)	17
NCI(1)	4				
NCI(2)	9	10	13	14	54
90-subset	29	10	14	17	27

3. Database Characteristics. Comparison of results from different databases, particularly with small subsets of compounds, should be accompanied by some characterization of the databases. In this paper, we define a rotatable bond, or rotor, as a bond, usually acyclic, which can rotate freely and which has at either end at least one non-hydrogen. Methyl, primary amino groups, and hydroxyl groups are excluded from this definition. It is of interest to know what the proportion of rigid compounds is in these files, how many compounds have just one rotor, and so on. Both the Cambridge file and the NCI database are substructurally searchable in a variety of ways, and these search algorithms were used to establish an approximate distribution within each file of structures with 0, 1, 2, etc. rotors.

The results of this analysis are given in Table I. It can be seen from these data that the Cambridge Structural Database has a slightly higher population of rigid compounds[25]—those with zero rotors. Substructure search techniques applied to the NCI file[26] led to the NCI(1) identification of 18 319 "rigid" structures, about 4% of the database. When 60 000 structures from the NCI database were examined by the model builder in Chem-X,[9] the NCI(2) data were obtained. This suggested that 9% of the compounds had no rotatable bonds. The NCI-(2) data were based upon a sample of the file, and this, together with the fact that the definition of rotor is subtly different in the NCI(1) and NCI(2) cases, is responsible for the discrepancy between the two results. Thus, fewer than 10% of the structures in the NCI file are rigid. That the number is slightly higher in the Cambridge database is intuitively reasonable as that database contains a fairly high proportion of "parent" structures, whose structure were examined by X-ray methods for just that reason. The number of rigid compounds in the 90-compound subset, by actual count, is 26 (29%) (see Table I).

The data presented here depend to some extent upon definitions. The important point they reveal, however, is that the 90-compound subset that was processed by CONCORD is significantly biased toward rigid compounds, which account for 29% of the 90—far in excess of those found in either the CSD or the NCI file. Since rigid compounds present less of a problem to model building programs such as CONCORD, the subset is biased in CONCORD's favor—a point that will be revisited in the discussion of results below.

4. Structure Comparison. Each compound from the 90-member subset was expressed as a CONCORD structure and also as a Cambridge structure. The two structures were then imported into QUANTA, running on a Silicon Graphics 4D/310GTX or 4D/70G workstation, where direct visual and mathematical comparisons could be made. Visual checks were made to determine whether there were any gross differences in torsion angles between the structures. In the mathematical comparison, an attempt was made to superimpose one structure upon the other. In such comparisons, all non-hydrogen atoms were used, and the result was expressed as the mean square

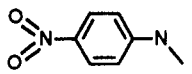
Table II. Results of RMS Calculations

RMS (Å)[a]	no. of rotatable bonds					
	0	1	2	3	>3	Σ
<0.5	18	8	9	2	0	37
0.5–1.0	5	1	2	5	8	21
1.0–1.5	2	0	3	4	7	16
1.5–2.0	1	0	1	4	4	10
>2.0	0	0	1	0	5	6
totals	26	9	16	15	24	90

[a] All non-hydrogens.

(RMS) of the deviations in atomic positions, summed over all the non-hydrogen atoms.

The results of comparison of 90 pairs of structures are summarized in Table II and presented in detail in Table III. In 37 of these cases (41%), the modeled structure was essentially identical to that determined by X-ray measurements. A typical example is given by *N*-methyl-*p*-nitroaniline (**1**). All the non-hydrogen atoms in this compound, the nitro



1

$CH_3COCOCH_3$

2

group, the phenyl carbons, and the methylamino atoms are coplanar. This is seen clearly in both the X-ray structure and the CONCORD-derived structure (Figure 1). The RMS deviation in an 11-atom match of the two structures in Figure 1 is 0.061 Å. A similar result is obtained for biacetyl (**2**). Here, the oxygen atoms are fully trans to one another, as is seen in both structures in Figure 2. The only difference between the two is in the orientation of the methyl groups with respect to the main chain. The RMS in this 6-atom match is 0.035 Å.

4.1. Exact Matches. More of the compounds that give such close matches are shown in Figure 3. Here the CONCORD-derived structure—the lower one of the pair in all cases—is essentially the same as that derived from X-ray measurements. In the 2-hydroxyfluorenone (Figure 3a), the CONCORD-derived structure is identical to that reported from X-ray measurements. The only detectable difference is in the position of the hydroxyl hydrogen, and the RMS difference between the two structures is 0.033 Å. This means that, on average, each nonhydrogen in the calculated structure is within 0.033 Å—2.2% of a bond length—of its correct position as determined by X-ray diffraction. A similar result is obtained with nitromethane (Figure 3b). Here, the positions of the hydrogens were not reported in the X-ray structure, and the non-hydrogen atoms are placed essentially exactly by CONCORD. The same coincidence of non-hydrogen atoms is seen in thioacetamide, tetrahydrofuran, oxetane, and 9-anthracenemethanol, and 12 other relatively simple molecules (Table III). In such cases, where bond rotation is not a major variable, CONCORD performs essentially flawlessly, and the structures it builds are virtually indistinguishable from measured structures.

4.2. Near Matches. Some structures built almost correctly are shown in Figure 4. The case of 4-methylphenanthrene (Figure 4a) is of interest because the X-ray structure reveals some deformation of the central ring, clearly caused by the presence of the 4-methyl group. This effect is not detected by CONCORD, with the result that a somewhat higher RMS of 0.118 is calculated. When bond rotation is possible and different rotameric modifications of a structure become possible, CONCORD has more difficulty but can still perform with credibility. In *meso*-hydrobenzoin (Figure 4b), the program miscalculates the torsion angle of one of the phenyl rings with respect to the central bond. As a result, ten of the carbon atoms are placed correctly, but the remaining four are slightly misplaced, and the overall RMS is 0.133 Å. CONCORD computes the two ring systems of the benzodiazaphosphole (Figure 4c) accurately but assigns a slightly incorrect torsion angle to the bond linking them. As a result, four atoms are misplaced, and the resulting RMS is 0.264 Å. In the octahydroanthracene (Figure 4d), CONCORD applies a twist to both outer rings, but the sense of the twist is the same in both rings. The X-ray determination shows this to be in error and as a result, the RMS in this case is 0.278 Å. The two rings of *p*-hydroxybiphenyl (Figure 4e) seek to be coplanar so as to maximize the electronic delocalization, and this is clearly seen in the X-ray structure. CONCORD fails to recognize this effect and instead attempts to relieve steric crowding by placing the two rings orthogonally to one another. Four of the 13 atoms in the structure are thus misplaced, and the RMS value is 0.347 Å. In diphenylmethane (Figure 4f), a poor choice of two torsion angles leads to the RMS of 0.368 Å.

4.3. Mismatches. The RMS error introduced by a miscalculated torsion angle is often small, particularly if it is the rotation of a small group that is involved. If hydrogens are neglected, as here, then the rotation of a methyl group does not change the RMS. Rotation of a group such as a phenyl or naphthyl however can have a considerable impact on the RMS value. Some cases where this is so are shown in Figure 5. In glyoxime (Figure 5a), the C–N bonds have been assigned rotations that are 176° and 181° in error. Any three atoms of the model of glyoxime can be superimposed exactly upon those of the X-ray structure; if the two errant torsion angles are adjusted in QUANTA, then all six atoms can be superimposed with an RMS of 0.096 Å. The fluorine atom in 2-fluoroacetamide (Figure 5b) is shown by the X-ray structure to be trans to the amide oxygen. CONCORD has this exactly wrong. This seriously misplaces half the atoms in this small molecule, and the resulting RMS (0.817 Å) is quite large. CONCORD produces a correct coplanar nitrophenyl unit in the diphenyl sulfide (Figure 5c) but miscalculates the torsion angles of the C–S bonds, and this leads to an RMS of 0.910 Å. Both the side chains in 2,6-diacetylaminopyridine (Figure 5d) are misrotated by 120°, and this is reflected in an RMS of 0.948 Å. Even more serious torsion angle problems are seen in the remaining examples in Figure 5 and lead to RMS values as high as 2.11 Å. This means that every non-hydrogen is more than one bond length from its correct position.

4.4. Large Rings. Modeling of rings containing 3–7 atoms is fairly straightforward because torsion angles between ring atom are largely controlled by the structure. As rings become larger, however, this control disappears[27] and an appreciable conformational analysis problem appears. This is a challenge that CONCORD fails. Some models of structures with large rings are shown with the X-ray structures in Figure 6. In 1,4,7,10-tetrathiacyclododecane (Figure 6a), CONCORD seeks to minimize internal interactions by producing an open, almost circular ring. The X-ray data in the upper diagram shows this to be an error: the ring in fact is almost square. A similar bad choice of internal torsion angles was made for the thiacyclononane (Figure 6b), and the results were even worse with the phosphacyclononane (Figure 6c) and the thiacyclononane (Figure 6d). These results suggest that

Table III. 90-Compound Set of Compounds Modeled by CONCORD

CAS no.	CSD no.	RMS	rotors[a]	comments	CAS no.	CSD no.	RMS	rotors[a]	comments
53-19-0	DCCPET	0.31	3		6937-59-3	JEDAIC	0.68	5	
54-36-4	BIHYEW10	1.93	3		7597-43-5	JEBFEB01	0.26	1	
56-41-7	LALNIN03	0.15	1		N/A		0.05	0	
57-85-2	ZZZRCG01	0.56	3		13652-13-6	DADLUP	1.68	5	
62-55-5	THACEM01	0.03	0	acyclic	N/A		0.10	0	
70-25-7	NOGUNA02	1.80	4		N/A		1.10	4	
75-52-5	NTROMA08	0.03	1		15718-46-4	PYMSUL10	0.87	3	
92-69-3	BOPSAA01	0.35	1		N/A		0.35	0	unsubstituted ring system
100-15-2	FUXNAN	0.06	2		21416-87-5	ICRFRA10	1.98	3	
101-59-7	FIHBED	0.91	3		N/A		1.88	9	
101-81-5	ZZZMKS01	0.37	2		24584-09-6	ICRFRB10	1.44	3	
109-99-9	BUNJAV01	0.05	0	unsubstituted 5-membered ring	25423-56-7	FOPCAO01	0.78	12	12-membered ring
296-41-3	BOWROU	1.69	18	unsubstituted 18-membered ring	27848-84-6	GICVUJ	2.11	5	
311-03-5	DOFSUM	1.06	0	unsubstituted ring system	N/A		0.09	0	acyclic
431-03-8	CABBIQ01	0.03	1		30868-30-5	PYRZOM01	1.51	3	
503-30-0	CIVXIO01	0.05	0	unsubstituted 4-membered ring	32846-66-5	FIGMAJ	0.49	2	
518-75-2	CITNIN02	0.18	1		N/A		1.27	5	
557-30-2	GLOXIM11	0.67	3		38734-05-3	CEDDUK	0.69	10	10-membered ring
579-43-1	VABVEZ	0.13	3		N/A		0.29	3	
640-19-7	FACETA01	0.82	1		N/A		1.33	3	
N/A[b]		0.20	0	unsubstituted 6-membered ring	64332-37-2	MEGONE	2.60	9	
N/A		0.25	0	unsubstituted ring system	65114-88-7	BEYTOO	1.10	9	9-membered ring
N/A		0.82	4		66054-22-6	ICRFRD01	0.20	2	
791-28-6	TPEPHO04	0.56	3		N/A		0.16	0	monosubstituted ring
832-64-4	FUVGEI	0.12	0	unsubstituted ring system	N/A		0.96	8	
838-41-5	DEMTIY	0.81	2		N/A		0.31	1	
948-44-7	POXTSO	0.43	0	unsubstituted ring system	71138-48-2	TNONDX	0.95	9	9-membered ring
1005-51-2	BULVAL02	0.35	0	unsubstituted ring system	77097-65-5	BACLAS	1.31	7	
N/A		0.10	0	unsubstituted ring system	77762-21-1	DUBPUL	0.14	0	unsubstituted ring
1079-71-6	CEKWEU01	0.28	0	unsubstituted ring system	N/A		1.45	6	
1138-48-3	VEMLUU	0.74	2		N/A		0.11	2	
1226-42-2	CASGEI	1.13	5		80799-73-1	BETRAT10	2.58	5	
1439-41-4	CETPOX01	0.78	9	acyclic	N/A		0.57	4	acyclic
1468-95-7	VAFMUK	0.08	1		84472-85-5	GATHOY	1.49	3	
2589-31-3	TABRIX02	0.21	2		85048-88-0	CAGLEB	0.22	0	unsubstituted ring
N/A		1.40	3		N/A		1.02	4	
4023-53-4	CNETPD	1.41	9	acyclic	88430-84-6	CEVPIC	0.87	9	
N/A		0.85	4		88589-00-8	CEVGIT	0.73	14	14-membered ring
4988-33-4	KAGMUA	0.24	0	unsubstituted ring system	88946-46-7	CEWKUK	0.68	1	
5441-02-1	DOPDAN	0.95	4		91190-12-4	COCZIO	2.35	8	
6067-31-8	FIPKAQ	1.61	2		91296-23-0	BUYTOE10	1.42	2	
6344-60-1	BESGEL	0.03	0	unsubstituted ring system	91296-27-4	BUYTIY10	2.25	2	
6510-63-0	SEYJUB	1.04	2		91384-92-8	COVTUC	0.11	0	unsubstituted ring
6829-31-8	CADZAI	1.28	2		92900-65-7	COYDUP	1.73	7	
N/A		0.05	1		92952-33-5	GAPCUV	1.93	3	

[a] Rotors are defined as rotatable bonds with more than one non-hydrogen at either end. [b] Structures for these compounds are proprietary.

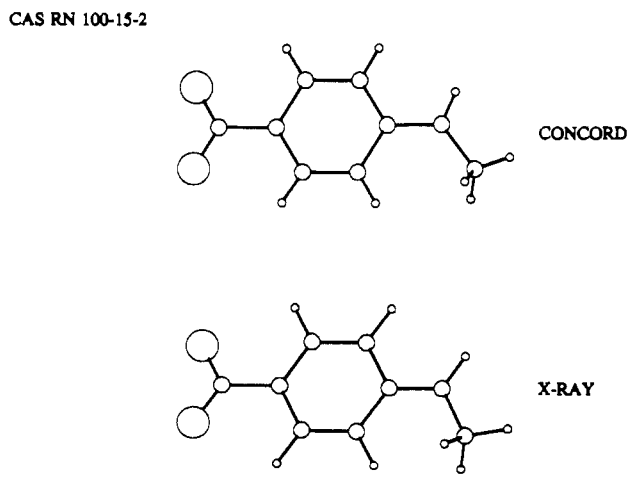


Figure 1. Structures of *N*-methyl-*p*-nitroaniline.

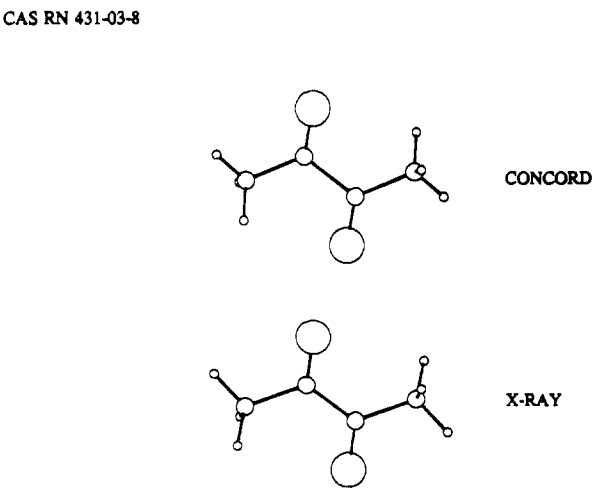


Figure 2. Structures of biacetyl.

models produced by CONCORD for structures containing large rings should be used with care. On the other hand, conformational transitions in large rings are often associated with very small energy differences, and for a molecule in solution, the relationship of its favored conformation to that measured in the crystal is unclear.

4.5. Classification of Matches According to Structural Features. In order to allow an approximate quantification of the accuracy of structure generation delivered by CONCORD, the set of 90 molecules that was compared in this work was subdivided into several classes according to the number of

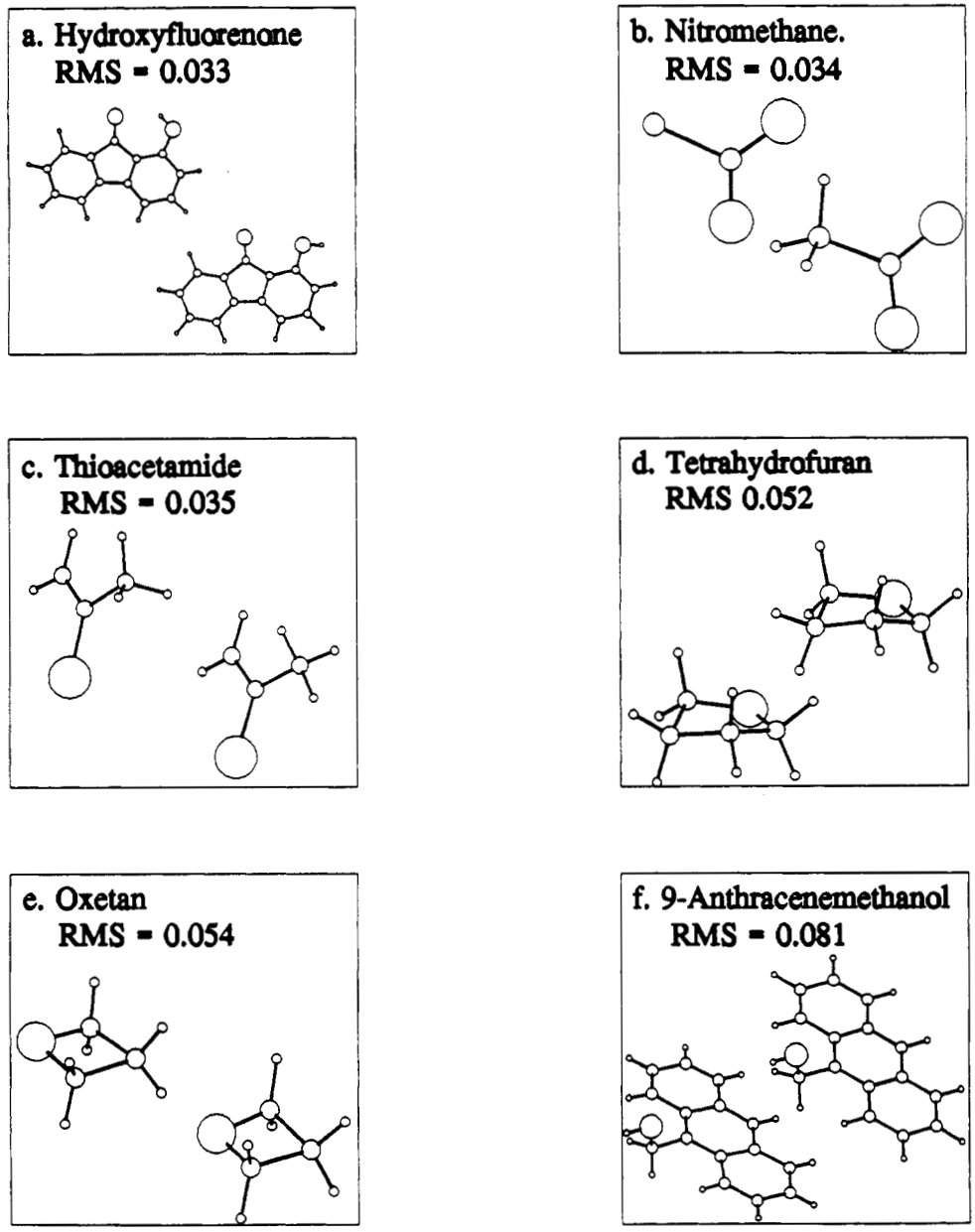


Figure 3. Structures built correctly by CONCORD.

rotatable bonds in each structure. Since hydrogen atoms were disregarded in the RMS calculations reported here, bonds to groups such as methyl or hydroxyl were ignored in counting rotatable bonds. Other methods for denoting flexibility of molecules have been proposed. The ϕ index, described by Kier,[28] is based upon κ indices and reflects molecular size, branching, rings, and heteroatoms. The global simple (GS) index proposed by Fisanick[29] weights the size of the fragments at either end of a rotor; a phenoxy group contributes more to the GS than a methoxy group. Both these measures go beyond the number of rotors but provide more definition than is necessary for the present comparison.

If compounds in the 90-compound set are classified according to the number of rotors they possess, then within each class the number of structures with RMS differences in different ranges are given in Table II. Thus, of the 26 structures with no rotatable bonds, 18 gave an RMS value <0.5 Å, five had an RMS value between 0.5 and 1.0 Å, and so on. A "perfect" match was defined for the sake of this discussion as one in which the RMS for all non-hydrogens in the two structures was <0.5 Å. As can be seen from the table, the likelihood that CONCORD will produce a perfect match drops sharply as the number of rotatable bonds in the structure increases. Of the 26 compounds with no rotatable bonds, only 18 (69%) afforded perfect matches because all but one of the eight remaining compounds contained large rings whose conformation was not calculated well by CONCORD even though they contain no "freely rotatable" bonds.

In terms of RMS values, structures containing a single rotatable bond were handled by CONCORD about as well as those with no rotatable bonds. With two or more rotors, however, the RMS values begin to increase to the point where, with more than three rotors, no perfect matches are obtained and a majority of the RMS values are above 1 Å. Finally, it should be noted that for compounds with greater than three rotatable bonds, the RMS do not deteriorate uniformly. The data in the table suggest that even with very flexible molecules, CONCORD can still sometimes arrive at a reasonably correct structure. This happens when a molecule adopts a predictable conformation—one in which substituents are all in a trans-anti juxtaposition and no subtle long-range nonbonded interactions are present. An example of this is the diphenyl ethylene glycol (3). In this structure, the hydroxyl groups are trans to one another, and the phenyl rings rotate into a plane

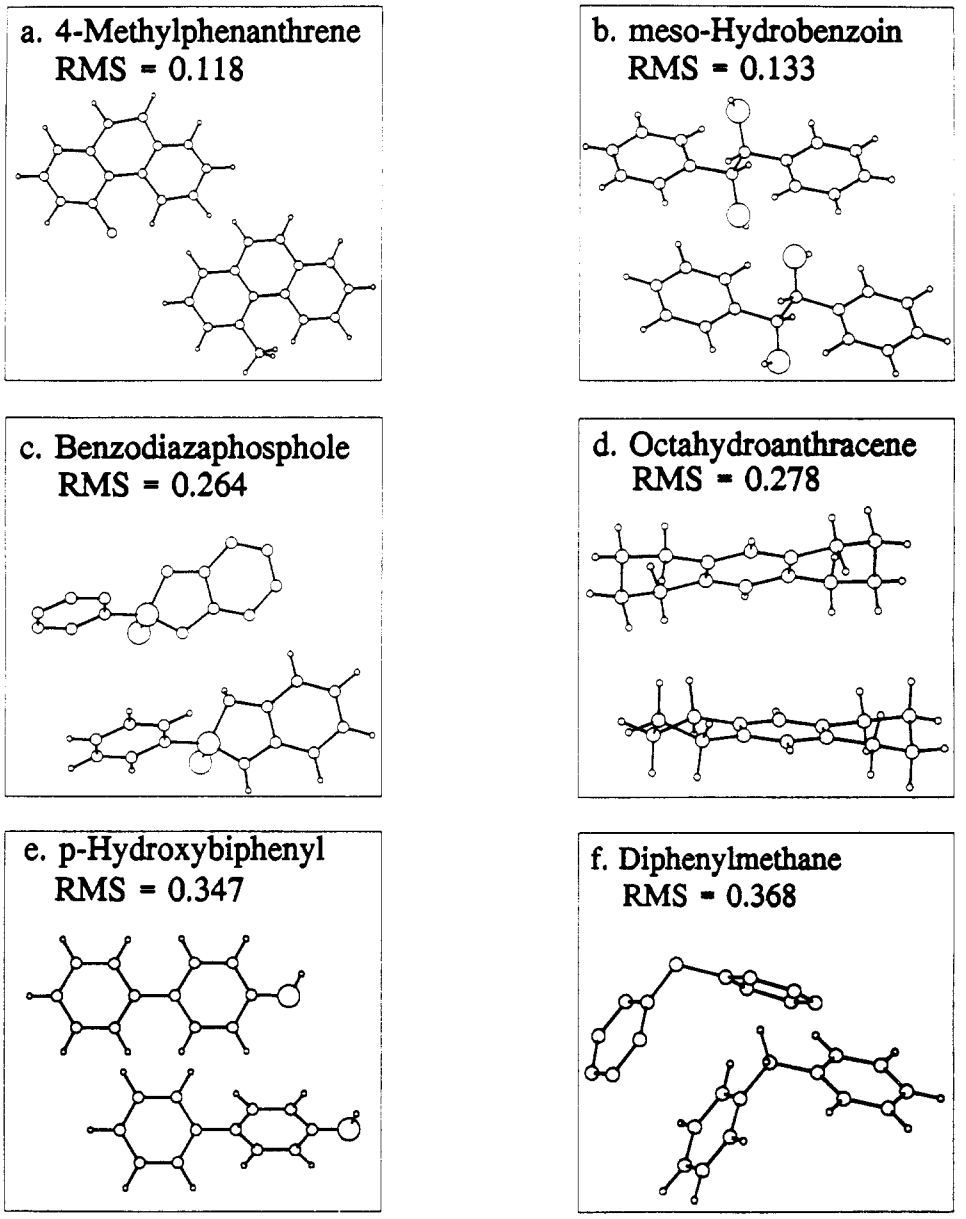
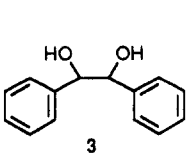
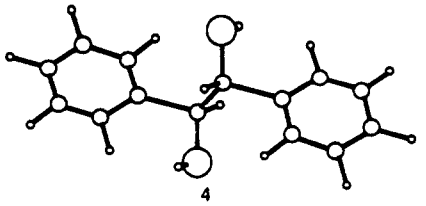


Figure 4. Near-correct structures built by CONCORD.

where interactions with the hydroxyls are minimized. CONCORD based its model (4) upon exactly these rules, and as a result, the model compares well to the X-ray structure (RMS = 0.133 Å). Many of the perfect matches were of rigid molecules, but several flexible structures were found in this group.



3



4

DISCUSSION

1. Rigid Molecules. Modeling of small, largely inflexible molecules presents relatively little difficulty to programs such as CONCORD. Such pseudomolecular mechanics approaches lead to structures which are generally identical to the structures emerging from X-ray measurements and can be used in 3D databases with some confidence. CONCORD generates such structures very rapidly, and for these structures, its use in database generation is quite appropriate.

2. Bond Rotation. When one or more bonds in a structure can rotate more or less freely, structure generation programs such as CONCORD, which do not carry out energy-based conformational searches, can only use a rule-based approach to the determination of the correct torsional angles(s). With CONCORD, this works remarkably well. Realistic torsion angles are incorporated into the models, the driving force for the selection of these angles being the relief of strain. Where the relief of strain is in fact a major influence, CONCORD often comes close to the correct torsion angle, but where other factors, such as the energy associated with bond delocalization intervene, CONCORD often errs. The value of such models in a 3D database is obviously rather low if conformational flexibility cannot be tolerated by the search software. In fact, modern algorithms such as those provided by Chem-Design[9] and BioCad[30] do examine torsional degrees of freedom, and it is clear that this must be done if accurate searching is desired.

A point that should be mentioned here is that, as far as torsion angles are concerned, "correctness" is not an absolute quality. We have used X-ray diffraction data here as a

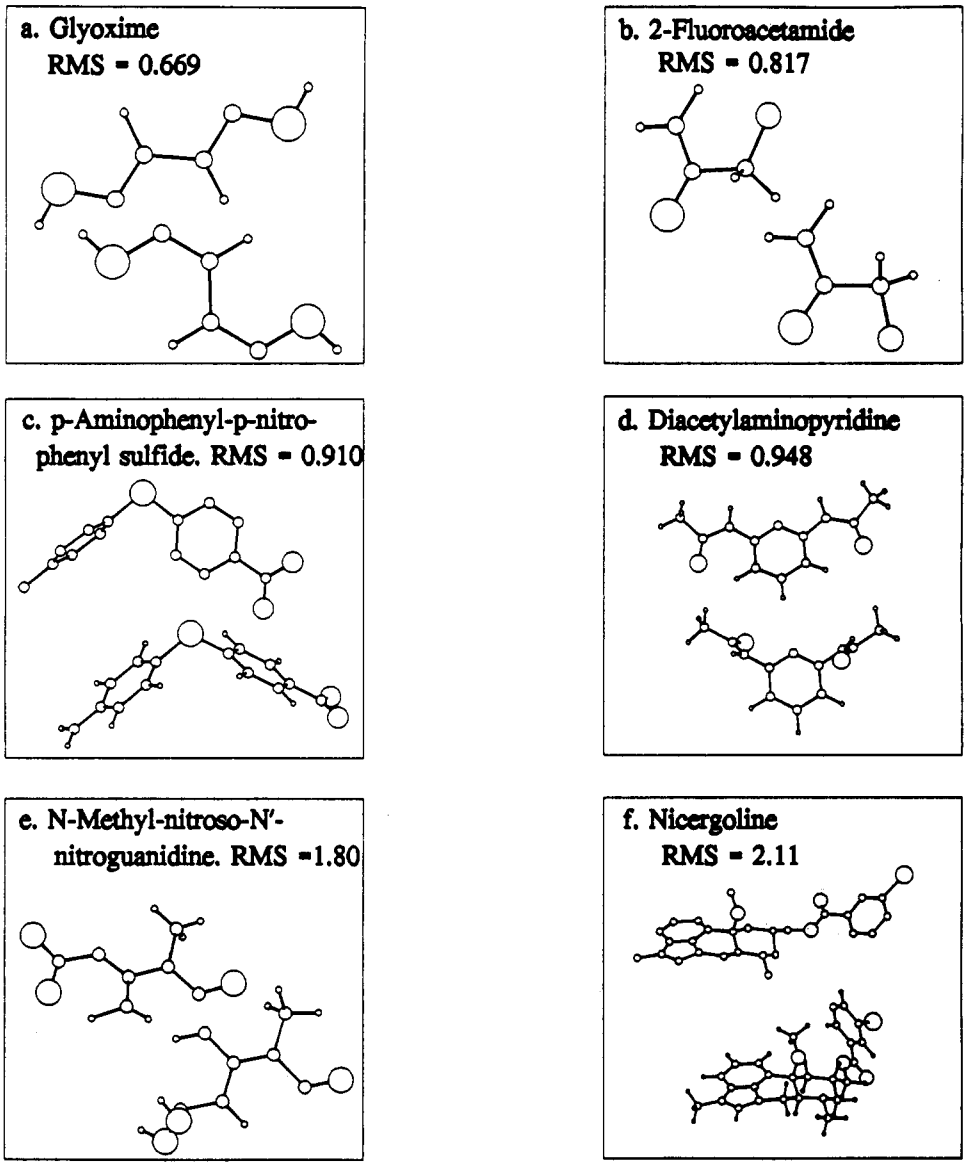


Figure 5. Incorrect structures built by CONCORD.

standard aganist which the modeled structures can be examined, but many torsion angles can change with no energy demand beyond what is available thermally. Torsion angles as measured in the crystal are influenced by forces peculiar to the crystal—e.g., intermolecular packing forces. That the same torsion angles have different values in the modeled structure does not necessarily mean that the modeling method is in error. The problem of freely rotating bonds is best managed either by conformational searching—a very time-consuming approach—or by incorporation into 3D databases of multiple conformational variations.[9]

3. Chiral Centers. Chiral centers pose a different sort of problem to those building 3D databases. If the modeling program is supplied with the correct chirality at every chiral center, it will produce a model which, in terms of optical asymmetry, is correct. Most of the large 2D databases, however, including the NCI database and even the CAS Registry, have incomplete information about chirality. Prior to the mid-1970s, chiral centers were often ignored when building these large 2D files, and this vitiates efforts to model 3D structures from these 2D data. In this work, CONCORD arbitrarily assigns chirality where none is given, and evidently in half the cases the assignment is wrong. In trivial cases with one chiral center, the center was inverted after the CONCORD processing, and then the comparison was carried out. In a number of cases, however, several chiral centers were misassigned, and such compounds were not included in the overall statistics discussed here. The message to those building 3D databases is that special attention must be paid to the reliability of the chirality information that is input to the modeling program.

SUMMARY

As has been noted, CONCORD has become popular as a means of generating 3D databases from the corresponding 2D data. The resulting 3D databases are being used in all manner of endeavors such as pharmacophore searching.

The program cannot handle "heavy" atoms, i.e., atomic number >18, and of course, it must have correct chirality information. Both these problems will be resolved as better data become available to the program, and work is now in progress by a number of groups to deal with them.

In 41% of the compounds we examined, the program produces structures which are correct. In most of the remainder, the torsion angle problem is more or less serious. It is clear from the results described here that 3D model builders and search systems must be able to manage confor-

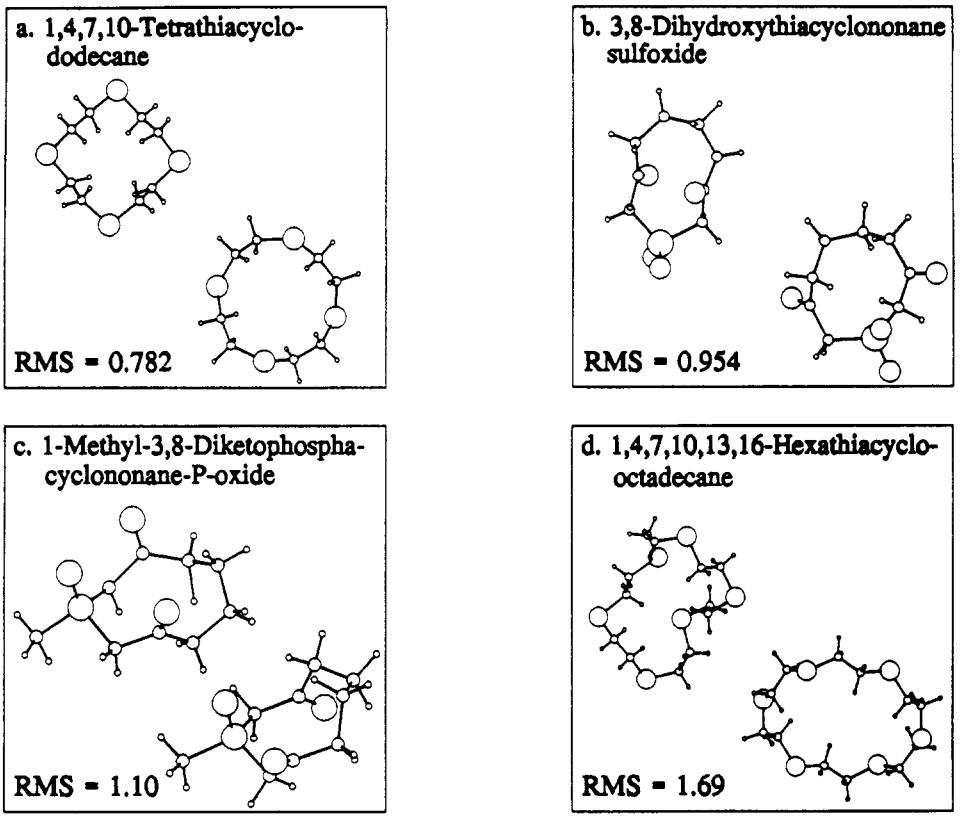


Figure 6. Structures containing large rings.

mational flexibility because such flexibility is present in over 80% of organic structures, and if molecules are assumed to be conformationally rigid, serious errors are inevitable. Methods to deal with flexible bonds are beginning to appear,[9,28] and these will have to be used with modeling programs like CONCORD before the results can be fully trusted.

ACKNOWLEDGMENT

The authors gratefully acknowledge the assistance of K. Paull in the acquisition of the CONCORD software.

REFERENCES AND NOTES

(1) Ehrlich, P. Über den Jetzigen Stand der Chemotherapie. *Chem. Ber.* **1909**, *42*, 17–47.
(2) Martin, Y. C. 3D Database Searching in Drug Design. *J. Med. Chem.* **1992**, *35*, 2145–2154.
(3) Sheridan, R. P.; Nilakantan, R.; Rusinko, A., III; Bauman, N.; Haraki, K. S.; Venkataraghavan, R. 3DSEARCH: A System for Three-Dimensional Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255–260.
(4) Martin, Y. C.; Bures, M. G.; Willet, P. Searching Databases of Three-Dimensional Structures. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. D., Eds.; VCH Publishers: New York, 1990; pp 213–263.
(5) Martin, Y. C.; Danaher, E. B.; May, C. S.; Weininger, D. MENTHOR, a Database System for the Storage and Retrieval of Three-Dimensional Molecular Structures and Associated Data Searchable by Substructural, Biologic, Physical or Geometric Properties. *J. Comput.-Aided Mol. Des.* **1988**, *2*, 15–29.
(6) Güner, O. F.; Hughes, D. W.; Dumont, L. M. An Integrated Approach to Three-Dimensional Information Management with MACCS-3D. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 408–414.
(7) Van Drie, J. H.; Weininger, D.; Martin, Y. C. ALADDIN: An Integrated Tool for Computer-Assisted Molecular Design and Pharmacophore Recognition from Geometric, Steric and Substructure Searching of Three-Dimensional Molecular Structures. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 225–351.
(8) (a) Jakes, S. E.; Watts, N.; Willet, P.; Bawden, D.; Fisher, J. D. Pharmacophoric Pattern Matching in Files of 3D Chemical Structures: Evaluation of Search Performance. *J. Mol. Graphics* **1987**, *5*, 41–48. (b) Poirrette, A. R.; Willett, P.; Allen, F. H. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Characterization and Use of Generalized Valence Angle Screens. *J. Mol. Graphics* **1991**, *9*, 203–217. (c) Clark, D. E.; Willett, P.; Kenny, P. W. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Use of Smoothed Bounded Distances for Incompletely Specified Query Patterns. *J. Mol. Graphics* **1991**, *9*, 157–60.
(9) Murrall, N. W.; Davies, E. K. Conformational Freedom in 3D Databases. 1. Techniques. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312–316.
(10) Christie, B. D.; Henry, D. R.; Güner, O. F.; Mook, T. E. MACCS-3D: A Tool for Three-Dimensional Drug Design. *Proc. 14th Int. Online Inf. Mtg.* **1990**, 137–161.
(11) CONCORD (copyright 1987, 1988, University of Texas, Austin, TX). *CONCORD User's Manual*; TRIPOS Associates: St. Louis, MO, 1988.
(12) (a) Rusinko, A., III. Tools for Computer-Assisted Drug Design. Ph.D. Thesis, University of Texas at Austin, Austin, TX, 1988. (b) Pearlman, R. S. Rapid Generation of High Quality Approximate 3-D Molecular Structures. *Chem. Des. Auto. News* **1987**, *2*, 1–6.
(13) Sprague, J. T.; Tai, J. C.; Yuh, Y.; Alligner, N. L. The MMP2 Calculational Method. *J. Comput. Chem.* **1987**, *8*, 581–603.
(14) Rusinko, A., III; Sheridan, R. P.; Nilakantan, R.; Haraki, K. S.; Bauman, N.; Venkataraghavan, R. Using CONCORD to Construct a Large Database of Three Dimensional Coordinates from Connection Tables. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 251–255.
(15) This and other CAS databases are available and can be searched online via STN, the CAS online service.
(16) This database of ≈ 450 000 chemicals, including many which have been examined for antitumor activity, is maintained under the NCI Drug Information System, which was designed and built by NCI. See: Milne, G. W. A.; Miller, J. A.; et al. The NCI Drug Information System. Parts 1–6. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 154–197.
(17) Over the lifetime of the program, 1 compound from every 3000–5000 tested has shown sufficient activity to enter clinical trials, and about 10% of these have reached the point where they have been marketed as drugs for the treatment of cancer. There are currently 39 drugs in clinical use for treatment of cancer; 30 of these are cytotoxic agents, and 12 of these were developed by NCI, which was also involved at some stage in the development of most of the remaining 18.
(18) Between 1967 and 1992, Chemical Abstracts Service registered 11 million compounds. Estimates indicate that a further 1 million different compounds were published before 1967.
(19) Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. The Cambridge Crystal Data Centre: Computer-Based Search, Retrieval, Analysis, and Display of Information. *Acta Crystallogr.* **1979**, *B35*, 2331–2339.
(20) Version 2.9.3 of CONCORD was used to generate all the structures discussed here. With these structures, there were no significant differences between these results and those obtained with CONCORD version 2.9.1.

(21) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *12*, 535–542.
(22) QUANTA is produced by Molecular Simulations Inc., 200 Fifth Ave., Waltham, MA 02254.
(23) CAS RN in the NCI File were originally assigned by Chemical Abstracts Service.
(24) CAS RN in the Cambridge File are provided by the authors of the primary articles and collected by the CCDC during database assembly.
(25) The QUEST search software provided with the CSD supports searches (bit screens 578–581) for structures with specific numbers of acyclic C–C bonds. These screens were used to estimate the percentage of the whole file with a specific number of rotors.Compounds where the acyclic bond was not single were subtracted from these totals, compounds containing C–X (X = O, N, S, etc.) were added, and compounds in which the only rotor was a bond such as C–CH_3, C–OH, or C–NH_2, which for our purposes are not rotors, were also removed from the totals.
(26) The NCI database was searched for compounds which have an *n*-membered ring (*n* = 2 [double bond], 3, 4, 5, 6, 7, 8, 9, or 10) free or imbedded in a larger ring system, with or without heteroatoms (N, O, S, etc.) in the ring and no ring substitutents or acyclic bonds, other than to CH_3, OH, or NH_2. This conservative definition of a rigid structure was met by only 18 319 compounds (4.06%) in the database.
(27) Saunders, M.; Houk, K. N.; Wu, Y.-D.; Still, W. C.; Lipton, M.; Chang, G.; Guida, W. C. Conformations of Cycloheptadecane. A Comparison of Models for Conformational Searching. *J. Am. Chem. Soc.* **1990**, *112*, 1419–1427.
(28) Kier, L. B. An Index of Molecular Flexibility from Kappa Shape Attributes. *Quant. Struct.–Act. Relat.* **1989**, *8*, 218–221.
(29) Fisanick, W.; Cross, K. P.; Rusinko, A., III. Characteristics of Computer-Generated 3D and Related Molecular Property data for CAS Registry Substances. Private communication.
(30) Sprague, P. W. Catalyst: A Computer-Aided Drug Design System Specifically Designed for Medicinal Chemists. *Proc. Montreux Chem. Inf. Conf.* **1991**, *3*, 107–112.