From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders

Jens Sadowski and Johann Gasteiger*

Organisch-Chemisches Institut, Technische Universität München, Lichtenbergstrasse 4, D-85747 Garching, Germany

Received March 24, 1993 (Revised Manuscript Received September 3, 1993)

Contents

I. Introduction

The three-dimensional structure of a molecule determines to a large extent many physical, chemical, and, particularly, biological properties. The understanding of molecular properties, of chemical reactivity, and of biological activity requires not only information on how atoms are connected in a molecule (the constitution or two-dimensional (2D) connectivity) but also on their three-dimensional (3D) structure. Experimental sources of information on the 3D structure of stable compounds can be obtained from such methods as X-ray crystallography, microwave spectroscopy, electron diffraction, or NMR spectroscopy. For several reasons these sources often are not sufficient: (a) The number of compounds whose 3D structure has been determined is small indeed when compared to the number of known substances. It is just not feasible to experimentally determine the 3D structure of the many millions of known compounds. (b) Computational techniques in organic chemistry as structure elucidation,[1] synthesis planning,[2] QSAR, and drug design[3,4] investigate enormous numbers of hypothetical structures which are not yet known or even not stable. The missing link between the 2D and 3D worlds is a technique capable of generating a 3D model of a chemical structure starting from the 2D connectivity information giving the constitution of a molecule. Quantum mechanical[5] or molecular mechanics[6] calculations can produce 3D molecular models of high quality but need at least some reasonable starting geometries. Because of the basic role of the 3D structure in all these areas 3D structure generation is one of the fundamental problems in computational chemistry.

Jens Sadowski was born on November 8, 1963 in Dresden, Germany. He studied both at the College of Technology of Merseburg, where he received his M.Sc. (1990), and in the group of Danail Bonchev at the Chemical–Technological Institute of Burgas, Bulgaria. He has been a member of Johann Gasteiger's group at the Technical University of Munich since 1990. His research interests lie in the fields of 3D structure prediction and conformational analysis.

Johann Gasteiger was born in Dachau, Germany, on October 27, 1941 and received his Ph.D. (1971) from the University of Munich. He spent a year as NATO postdoctoral fellow at the University of California, Berkeley. Since 1972 he has been at the Technical University of Munich where he obtained his Habilitation in 1979. In 1991 he obtained the Beilstein-Gmelin medal of the German Chemical Society for his achievements in computer chemistry. His research centers on the development of computer programs for synthesis design, reaction and reactivity prediction, analysis and simulation of mass spectra, prediction of molecular properties, the modeling of molecules and organic reactions, and the application of neutral networks for chemical problems.

The need for computer-generated 3D molecular structures has clearly been recognized in drug design. Searching in 3D databases[7] is widely used for finding new lead compounds. This task requires large 3D databases containing high quality structures from a wide variety of organic chemistry. Indeed several

0009-2665/93/0793-2567$12.00/0

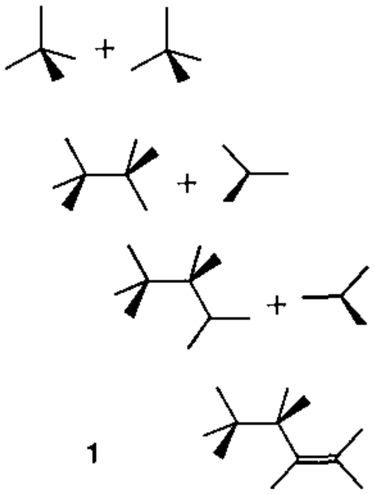

1

Figure 1. Construction of butene-1 (1) from monocentric fragments.

Table I. Commercially Available Databases of 3D Molecular Structures

database	no. of entries	source of coordinates	ref(s)
CSD	100 000	X-ray	8
CAST-3D	370 000	CONCORD	9, 15
CAS-RF	4 500 000	CONCORD	9, 15
MDDR-3D	12 000	CONCORD	10, 14
FCD-3D	57 000	CONCORD	10, 14
CHCD	216 000	Chem-X Builder	11, 16

databases of 3D molecular structures have become commercially available. Table I compares the numbers of entries of some selected databases[8-11] containing experimental or computer-generated 3D structures.

As can be seen, the number of computer-generated models already now is larger than the number of compounds whose structure has experimentally been determined. This is not counting the large in-house 3D databases that are in use at some companies.

II. Description of the Problem

A. Conceptional Problems

Each approach to the automatic generation of 3D molecular models has to solve several general problems.

1. Coordinate System

The first question is the choice of an appropriate representation of the 3D molecular models. The positions of the atomic nuclei of an N-atomic molecule can be described by $3N$-6 coordinates. Commonly either cartesian (xyz) or internal coordinates are used. Internal coordinates can be a nonredundant set of linearly independent bond lengths, bond angles, and dihedral angles.

The strategy for building a molecular model from these internal coordinates can be compared with the use of a mechanical molecular model building kit.[12] Figure 1 demonstrates this with the example of butene-1 (**1**). We will go here through the various processes in mechanically building a molecular model to develop an understanding of the problems that have to be solved in computer generation of a 3D model.

Monocentric fragments which represent different hybridization states of a carbon atom are connected using joins with a length corresponding to the bond lengths. A basic assumption in this process of 3D structure generation is that the geometries of fragments of atoms and bonds in molecules can be represented by standard values for bond lengths and bond angles. This reduction is allowed since bond lengths and bond angles possess only one rigid minimum. Problems arise from dihedral or torsional angles, which describe the twisting of a fragment of four atoms, connected by a sequence of bonds, since the steric energy may have multiple minima around a rotable bond and the values of these minima may be rather similar. This leads to more than one possibility for constructing a model for such molecules.

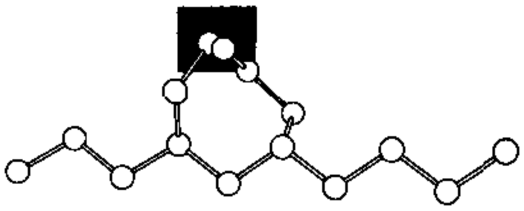


Figure 2. Crowded atoms (gray underlaid) in an unsatisfactory 3D model.

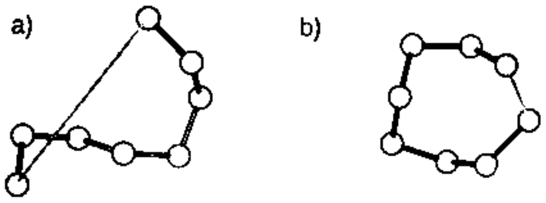


Figure 3. Ring closure in an eight-membered ring: (a) not closed and (b) closed. The ring closure is marked in gray.

2. Acyclic Structures

In open-chain and branched structures or substructures, the preferred torsional angles are those which simultaneously minimize torsional strain and the steric interactions between nonbonded atoms. The relatively large flexibility of such systems gives rise to multiple solutions (conformations) for the process of structure generation which have quite similar energy. Account of this flexibility has to be taken and is important in applications like drug design. However, limitations in storage space and computation times do not allow one to handle all these different geometries in large databases. In any case, the overlap of atoms must strictly be avoided (Figure 2). The flexibility of these structures may result in an ensemble of experimentally observed conformations. Even if only one conformation is preferred in the crystal field or in a specific solvent, the chances that a generated 3D structure corresponds to the experimentally preferred structure becomes rather unlikely. Thus, the generation of all low-energy structures becomes a problem of its own.

3. Cyclic Structures

Ring closure reduces the degrees of freedom particularly for the torsional angles. Figure 3 shows with the example of an eight-membered ring the requirement for appropriately chosen torsional angles within the ring in order to close the ring.

This is also expressed in a reduction in the number of possible conformations compared to those in acyclic systems. In addition, the energy barriers between these conformations can become relatively high, resulting in rather rigid geometries. Rather than trying to close a ring one can use information on possible single ring conformations. These conformations can be stored as 3D coordinate fragments or as lists of torsional angles.

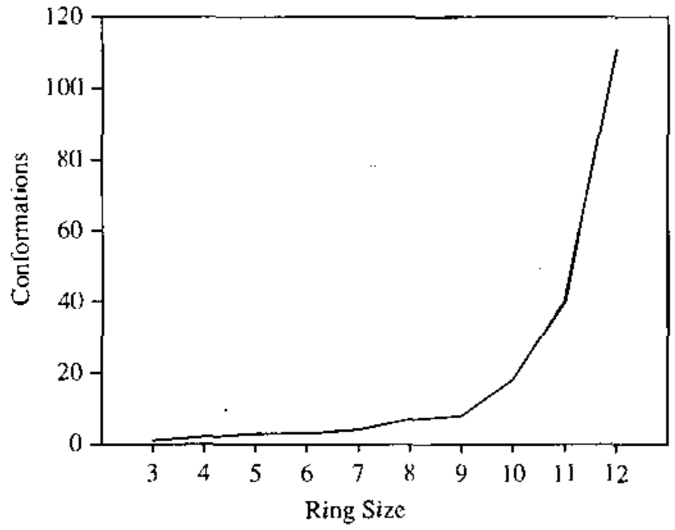


Figure 4. Increase of the number of known conformations of cycloalkanes with increasing ring size.

These so-called ring templates implicitly fulfill the additional condition of ring closure.

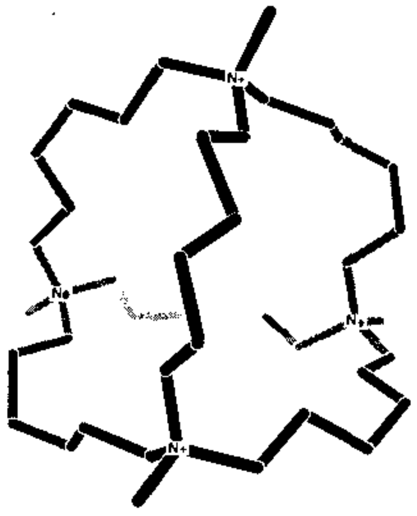
Another problem arises from the requirement to ring closure in cases of small rings where some internal coordinates must be deformed from their standard values. For example, the endocyclic bond angles in cyclopropane have values of about 60° instead of the ideal tetrahedral angle of 109.47°. These deviations from the standard values give rise to strain in such ring systems. Deviations from the optimum values of torsional and bond angles, and, in some cases, even of bond lengths, often have to be made to build models of polycyclic structures like fused or bridged ring systems. This, too, may result in strain energy.

4. Macrocyclic Structures

With increasing ring size the reduction in the flexibility due to the ring closure decreases. Large rings are, apart from the requirement to ring closure, as flexible as acyclic systems. Figure 4 shows the increase in the numbers of known conformations of cycloalkanes with dependence on ring size. Note, that the number of conformations is dependent on the force field used. In Figure 4 the numbers of conformations for rings of size three to eight were taken from ref 13a; the numbers of conformations of nine- to twelve-membered rings were obtained with the MM2 force field and taken from ref 13b.

The conformational flexibility and thus the number of valid 3D molecular models steeply increases from ring size nine upward. At the same time, a decision for one specific preferred conformation becomes more and more questionable since the energy differences between these different conformations decrease. These problems have to be regarded in processing macrocyclic systems. An explicit use of potential ring conformations becomes unreasonable from ring size eight on. A 3D structure generator which is able to handle such systems sufficiently, must, on one hand, be able to choose from a large number of possible conformations the appropriate ones and, on the other hand, be able to handle the flexibility of the systems. Similarly, like in acyclic systems it becomes less and less likely that the generated 3D structure corresponds to the experimental geometry. On the other hand, the flexibility of a large ring is reduced when it is fused to smaller rings or when it is being bridged.

In fact, a particularly severe problem is posed in polymacrocyclic systems. Although the individual macrocyclic systems may each have many conformations of about equal energy only one or a few are valid when several of these macrocycles are merged into a polymacrocyclic system like 2 in Figure 5.



2

Figure 5. Trimacrocyclic bridged system 2.

These conceptional problems suggest a separate handling of acyclic and cyclic systems—a strategy which is indeed used in nearly all present approaches to 3D structure generation.

However, an additional problem then arises when, after generation of 3D structures for the cyclic and acyclic parts, these substructures are assembled to build the entire molecular structure. Care has then to be taken that this process does not introduce too much strain or even results in the overlap of atoms.

B. Computational Requirements

The interest in databases of 3D structures greatly stimulated research and development of 3D structure generation systems. However, the use of these 3D generators for building large databases of 3D structures from 2D connectivity information[14-17] also imposes restrictions on the development of this area.

The decision to use a specific conversion program plays a crucial role since a change to another program will only be made with difficulties. First, the amount of computer resources for the conversion of hundreds of thousands of structures is quite large, and, secondly, a lot of scientific work will be based on such a database and so a change of these data makes a lot of the work already performed questionable or obsolete. Thus, the choice to use a particular 3D structure generation program should be made only after a careful evaluation process. On the other hand, the task of generating 3D structures from connectivity information (the constitution of a molecule) is just too important and the problems to be solved so diversified that it should always be open to new ideas and approaches. 3D database developers at Molecular Design Ltd. formulated the following criteria for a 2D-to-3D-conversion program[14] (the quotes are slightly abbreviated and modified):

Robustness. The program should run with a long mean time to a failure and indicate the actions taken on failure rather than simply crash.

Large Files. The program should be able to handle large numbers of structures contained in a single file in order to minimize the number of conversion jobs.

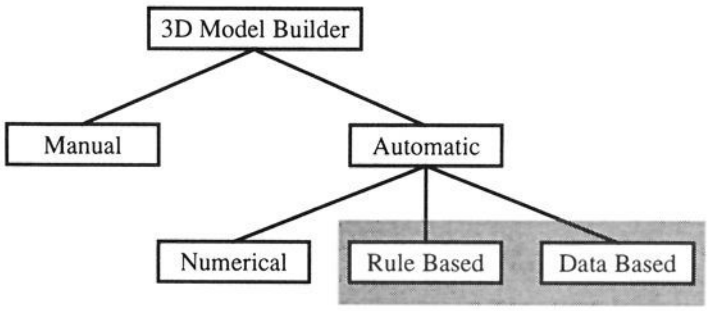


Figure 6. Classification scheme for the concepts in 3D model building. Only the fields underlaid in gray are covered in the review.

Variety of Chemical Types. The program should be able to handle a wide variety of structural types.

Stereochemistry. The stereochemical information contained in the input data must be handled correctly.

Rapid and Automated. The large size of the databases to be processed requires the conversion program to run in batch mode and to work with acceptable speed.

High-Quality Models. The generated models should be of high quality without further energy minimization and should represent at least one low-energy conformation. It should have internal diagnostics to validate the models generated.

High Conversion Rate. As many 2D structures as possible should be converted.

In this review we will cover only those published approaches which fulfill more or less these criteria. In the next section we will classify them into subdivisions and define the borderline to other, related methods. Then, detailed reviews of some of these subdivisions follow. The literature has been surveyed through January 1993. Unfortunately, the concepts inherent in the commercially available programs have not been indicated in detail in the few publications that are available. We therefore have to illustrate in more detail the concepts, solutions, and results of 3D structure generation with that program that we know best, one developed by ourselves and our co-workers (CORINA[18-21]). To our knowledge, up to now, no exhaustive comparison of all of these programs has been published. Comparison has been made of the results of one such program (CONCORD) with X-ray crystallographic data.[22]

III. Classification of Concepts

We attempt here a classification of concepts for 3D structure generation as illustrated in Figure 6. Only the fields underlaid gray will be covered in this review.

In the early beginning of thinking in three dimensions in organic chemistry, 3D molecular models were built by hand, using standard bond length and bond angle units from mechanic molecular model building kits.[12] This technique, useful still today, found in the age of computational chemistry its modern expression in interactive 3D structure building options incorporated into nearly each program package for molecular modeling.[23] The user may construct a 3D molecular geometry interactively, positioning atoms and bonds on a 3D graphics interface using standard bond lengths and angles or connecting predefined fragments. We will summarize all these methods under "manual" methods since all model building steps are performed by hand, irrespective of whether this is done in real space or with computer models.

Distinct from these are "automatic" methods which directly transform 2D input information on atoms, bonds, and the stereochemistry in a molecule as expressed in a connection table into 3D atomic coordinates without any user intervention. We divide the automatic methods into "numerical", "rule-based", and "data-based" methods.

Under numerical methods we cover quantum mechanical calculations[5] (QM), molecular mechanics[6] (MM), and distance geometry[24-26] (DG) since they are based on extensive numerical optimization procedures requiring long computation times (QM $\gg$ MM $>$ DG). All these methods, especially molecular mechanics and distance geometry, often are used in combination with a systematic or stochastic conformational analysis in order to scan the conformational space of a given system. A detailed review of conformational analysis methods is given elsewhere.[27] While quantum mechanical or molecular mechanics programs need a reasonable starting geometry, the distance geometry approach by Crippen[24,25] represents a stand-alone modeling procedure of its own since the so-called embedding procedure generates starting coordinates for further optimization. Several improvements like energy embedding[28,29] allow an effective scanning of the conformational space of a given molecule. Since distance geometry became a standard method in molecular modeling some other developments[30,31] are based on it. The MOLGEO program of Katritzky et al.[31] uses only the geometry optimization part, replacing the embedding procedure, which often results in a rather poor starting geometry, by a systematic depth-first conformational search. However, all these numerical methods need relatively long computation times and are therefore excluded from the construction of large 3D databases following the criteria for 3D model builders given in section II.B.

The next subdivision, the rule-based methods represent an approach based on the knowledge of chemists on geometrical and energy rules and principles for constructing 3D molecular models. This knowledge was originally gained from experimental data and theoretical investigations. It is built into 2D-to-3D-conversion programs in the form of chemical knowledge either in explicit (e.g., rules) or in implicit form (e.g., data on allowed ring conformations). Since this network of rules and data allows a direct building of 3D molecular models these methods are some orders of magnitude faster than numerical ones.

At the far end of rule-based methods are methods based almost exclusively on structural data. We cover these methods under a separate subdivision as data-based methods. These methods follow the concept of constructing molecular models from fragments that are as large and as similar as possible to the molecule to be built. These fragments are taken from a library of 3D structures. These programs make extensive use of the implicit knowledge on model building represented by databases of 3D structures. Of course data-based methods need also explicit rules on the fragmentation of the input structures, on finding closest analogs in the libraries, and on combining fragments to the entire molecular model. However, these methods may also construct 3D models without falling back upon timeconsuming numerical optimization procedures.

Clearly, there is no sharp border between rule-based methods and the data-based ones. The criterion for division is given when fragments larger than one single ring structure are used. These we call data-based methods. In this review only these two groups of automatic 3D model builders are covered since they are the only ones fulfilling the above criteria of automation and speed. In addition methods for conformational analysis are included as far as they fulfill these criteria. They are classified into the same two subdivisions, rule-based and data-based methods. Therefore, conformational search methods based on exhaustive systematic or stochastic scanning of the conformational space and/or numerical minimization methods will be excluded.

The scientific community has already acquired a lot of experience with automatic 3D structure generators. However, we can base our review only on that information that is available in commonly accessible publications. Thus, the performance of some of the programs mentioned might have been improved in newer versions.

IV. Rule-Based Methods

A. Early Precursors

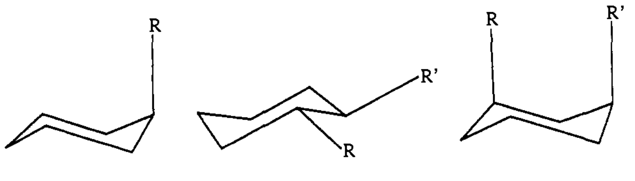
Ring systems represent a special challenge in 3D model building because of the additional constraint imposed by the requirement for ring closure. In this section we will mention some methods developed some time ago for the rapid and automatic conformational search of ring systems. These methods played a pioneering role in the further development of 3D structure generators, although they do not fulfill the above criteria for automatic 3D model builders that nowadays have to be required.

1. PRXBLD

A first step into rapid and automatic 3D structure prediction was made by Wipke and co-workers in 1972. PRXBLD,[32] a module of the SECS synthesis planning program,[33] was the first program able to generate a 3D model rapidly from a 2D drawing with stereochemistry. PRXBLD combined heuristics with a simplified force field to achieve speed and to avoid false minima. However, the program was interactively driven and no further details have been published.

2. Conformational Analysis for Six-Membered Rings in the LHASA Program

Corey and Feiner[34] semiquantitatively assigned conformations of six-membered ring systems during the development of the synthesis design program LHASA. The aim of this work was the prediction of the preferred conformations of synthetically important six-membered ring systems in order to evaluate the steric hindrance of different reaction sites in a molecule. First, several possible geometries are assigned to the single rings (e.g., chair, half-chair, boat) and the flexibility of these rings is evaluated (e.g., the possibility to distort them or to flip them into another conformation) using the 2D connection table and the stereochemical information. Second, the exocyclic substituents of the ring atoms are labeled to be either axial or equatorial. Third, the relative energy differences between several possible conformations of flexible ring systems are calculated using empirical procedures based on energy increment schemes for the single ring conformations, for intraring interactions (e.g., axial substituents, 1,2-diequatorial, or 1,3-diaxial interactions in chair conformations), and inter-ring interactions between different rings of one ring system. Destabilization energies E_D in monoaxial substituted cyclohexane chairs (**2**) are calculated using energy increments for a specific substituent. These values describe the energy difference between the axial and the equatorial configuration of a monosubstituted cyclohexane ring (eq 1). The interactions in 1,2-diequatorial (**3**) and 1,3-diaxial (**4**) substituted rings are described using separate increment schemes (eq 2 and eq 3, Figure 7). The substituent increments A_R, G_R, and U_R are based only on the nature of the atom directly connected to the ring.



3 **4** **5**

Figure 7. Monoaxial (**3**), 1,2-diequatorial (**4**), and 1,3-diaxial (**5**) substituted cyclohexane chairs.

$$E_D = A_R \tag{1}$$

$$E_D = G_R + G_{R'} \tag{2}$$

$$E_D = U_R + U_{R'} \tag{3}$$

The method is completed by some rules for the influence of heteroatoms. Similar computational schemes are used for other six-membered ring conformations. In a series of examples, sufficient agreement was found with energies obtained by molecular mechanics and with geometries obtained by X-ray crystallography. The strength of the method was the use of symbolic logic for the geometry and energy prediction. However, the approach was limited to six-membered ring conformations and no explicit 3D structures were generated.

3. The SCRIPT Program

Cohen, Colin, and Lemoine[35] presented in 1981 the SCRIPT program. A molecule is considered as an assembly of chain and ring fragments, possessing different conformations. The conformations are handled in an abstract form as "conformational diagrams" containing symbolic descriptions of the torsional angles of each bond. Chain fragments are treated as sequential four-atom fragments. Several possible low-energy conformations are given for the torsional angles in such a fragment that only depend on the nature of the central bond. Ring fragments are handled as templates that are joined. Possible conformers of rings of three to eight atoms are taken from a predefined table of templates that depend on the ring size and the distribution of double bonds. These conformers are stored in the form of conformational diagrams as shown

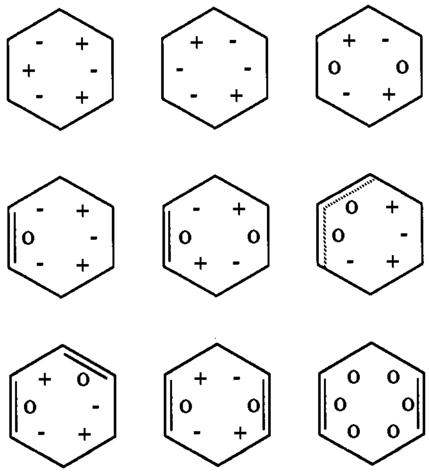


Figure 8. The nine possible conformational diagrams for a six-membered ring in the SCRIPT program. The torsional angles of the ring bonds are only defined by their sign (+/−) or zero (0) for a planar bond.

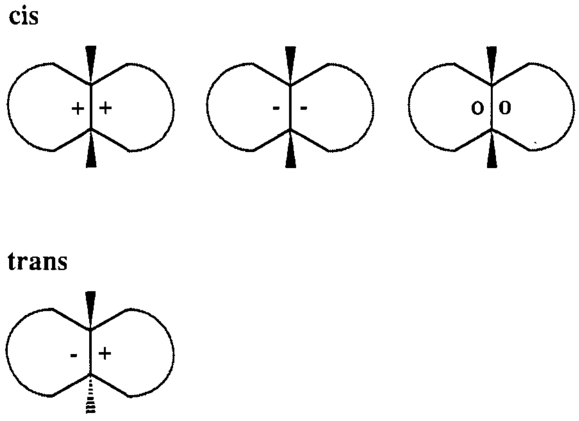


Figure 9. Rules of constraints for fused rings with an sp^3–sp^3 fusion bond. The diagrams show the only allowed combinations of torsional angles for cis and trans fused rings.

in Figure 8 for the six-membered ring. The torsional angles of the ring bonds in these diagrams are represented only by their sign (+/−) for gauche angle types or zero (0) for a cis bond.

For ring fragments consisting of more than one ring, being either fused or bridged, a set of rules is used that restrict the allowed conformations of two adjacent rings. Figure 9 shows the constraints rules for fused rings with an sp^3–sp^3 fusion bond. These rules consist of allowed combinations of torsional angles of the bond of fusion in the two regarded rings that depend on the stereochemistry of the bridgehead atoms. There are three combinations for a cis fusion and one for a trans fusion.

Similar rules exist for other types of fusion bonds (e.g., sp^3–sp^2, sp^2–sp^2 single, or sp^2═sp^2 double bonds). Bridged rings are only checked whether the bridgehead atoms are in a cis or a trans configuration depending on the stereochemistry of the bridgehead atoms and on the torsional angles assigned to the bonds forming the bridge. Trans configurations are not allowed.

In a first step, the possible conformations are generated on a symbolic level of conformational diagrams. The combinatorial product of all conformational diagrams for rings and chains forms the conformational space of the molecule. In a second step, a set of rules and computational schemes allows the direct translation of the conformational diagrams into 3D atomic coordinates by using standard bond lengths, bond angles, and torsional angles calculated from the symbolic descriptions in the diagrams. This is achieved by computational schemes based on ring sizes. The 3D coordinates obtained are regarded to be rather crude. They may be evaluated by the calculation of the conformational energy based on molecular mechanics potentials. However, only the energies obtained after a geometry optimization are useful for a ranking of the conformers. In other words, to obtain a reasonable molecular model a number of force field optimizations of different conformations is necessary.

The major strength of the SCRIPT method is the use of symbolic logic to construct possible ring conformations from a table of single ring templates and the direct translation of these symbolic representations into 3D atomic coordinates which makes these processing stages rather fast. The major weakness of this approach is the generation of rather crude 3D coordinates and the lack of an energy evaluation of the conformations at the symbolic level of conformational diagrams. The program was used with some benefit in reaction design studies.[36]

4. SCA: Systematic Conformational Analysis for Cyclic Systems

De Clercq[37–39] has developed a program called SCA ("Systematic Conformational Analysis") for the construction of conformations of ring systems consisting of three- to seven-membered rings. Like the SCRIPT program[35] it is based on lists of allowed conformations of single rings and a set of rules for determining torsion constraints in fused or bridged systems, i.e., the sign and the magnitude of the torsional angles common to two neighboring rings. The original procedures have been developed for a manual systematic conformational analysis starting from a two-dimensional structure with stereocenters indicated by a wedged/hashed bond notation.

After an interactive structure input via a 2D drawing of the structural formula augmented with stereodescriptors, the SCA program[38] performs the following steps. First, it analyzes the input and assigns possible conformations to all five-, six-, and seven-membered single rings considering the torsion constraints introduced by unsaturated bonds and fused or bridged systems. These single-ring conformations are stored in the form of lists of torsional angles. An energy value is assigned to each conformation (calculated from the conformational energy of the unsubstituted form), the influence of an exocyclic double bond, contributions from exocyclic substituents, and interactions of vicinal substituents. Second, the single-ring conformations are combined and the resulting abstract conformations of the entire ring system are ranked by the sum of the energies of the single-ring conformations. This energy ranking does not contain any information on long-range interactions as, e.g., exerted by substituents of two different rings. Therefore, in a third step, the abstract representations are translated into 3D atomic coordinates using standard values for bond lengths, bond angles, and torsional angles. A special procedure is used to perfectly close the rings of strained systems by deforming some endocyclic bond angles. Then, a new energy ranking is calculated for these 3D structures using the above energy terms with the exception of the contributions of the substituents, which are replaced by separate nonbonded energy terms for interactions between small (S) substituents, i.e., hydrogens and lone pairs, and large substituents (L). This is achieved by eqs 4–6

$$\text{S–S interaction:} \quad E = -10.4d + 24.2 \qquad (4)$$

$$\text{S–L interaction:} \quad E = -18.3d + 49.3 \qquad (5)$$

$$\text{L–L interaction:} \quad E = -35.4d + 107.3 \qquad (6)$$

where d is the nonbonded distance, in angstroms, and E the energy contribution in kilojoules per mole. This fine tuning of the conformational energy by rather simple linear functions of the nonbonded distances was tested by calculating the energy differences between the axial and equatorial forms of the chair–chair conformations of several methyl-*cis*-decalins. The reported results compare rather favorably with the energy differences calculated by molecular mechanics.[38]

The strength of the method is the rapid construction of reasonable 3D geometries of ring systems using symbolic logic and an energy ranking scheme which allows the derivation of best candidate conformations without having to invoke a geometry optimization. The weakness of the approach is the limitation to ring systems with up to seven members, although the handling of exocyclic chains is possible via the input of all necessary acyclic torsional angles.

B. WIZARD and COBRA

Extending an earlier work by Dolata and Carter,[40] Dolata, Leach, and Prout[41–47] developed two programs, WIZARD and COBRA,[48] for the systematic conformational analysis using symbolic logic and techniques of artificial intelligence (AI). The basic idea of this approach is to develop a set of rules for the construction of molecular models derived from the method of a human expert who recognizes conformational units with well-known optimum geometries (e.g., cyclohexane chair) and joins them to an entire system. The following steps are performed.

(1) The molecule is analyzed and "conformational units" are recognized. A conformational unit is a connected substructure for which the AI system has some knowledge on its conformational behavior. Figure 10 shows this fragmentation process for cyclazocine (**6**).[41] The molecule contains four monocyclic and five acyclic conformational units. Cyclic units consist of one or more rings. Acyclic units consist of one to three bonds. Note that neighboring fragments overlap.

(2) An abstract hierarchical representation of the molecule is generated in the form of a so-called "unit graph". The conformational units are the nodes of this graph. The edges of the unit graph are formed by the type of junction between two neighboring units (i.e., acyclic join, fused rings, or bridged rings). Figure 11 shows the unit graph for cyclazocine.[41]

(3) Lists of "conformational templates" are assigned to all conformational units, which are taken from a library. A template contains some knowledge on the fragment conformation, i.e., symbolic description of the conformation, strain energy, flexibility, and coordinates. If no exact expression of a specific unit can be found in the library, similar templates are searched on several levels of generalization.[47] If, for example, no template for a heterocycle can be found, the corresponding carbocycle is taken. The templates are obtained either from molecular mechanics or X-ray crystallography.

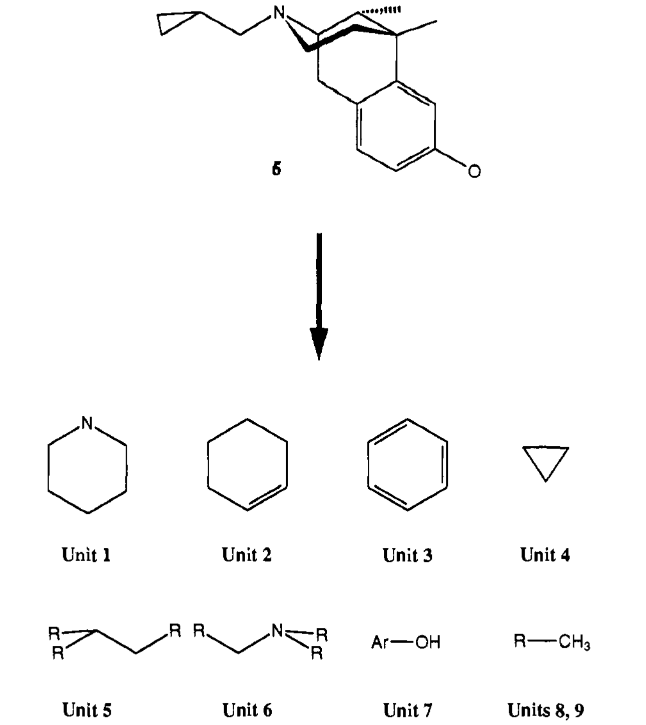


Figure 10. Recognition of conformational units in cyclazocine (**6**).

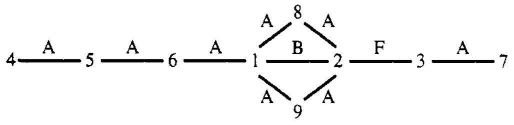
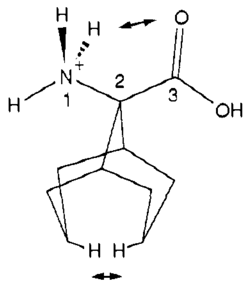


Figure 11. Abstract representation of cyclazocine (**6**). The conformational units are numbered in analogy to Figure 10. The joins are marked by capital letters: A = acyclic join, F = fused rings, B = bridged rings.

(4) "Symbolic suggestions" of conformations are built on the abstract level of the unit graph. The whole conformational space is formed by the combinatorial product of the templates assigned to the conformational units. The conformational space is searched by using a directed strategy: the A* algorithm.[46] The obtained symbolic suggestions are criticized using a set of predefined and self-learned rules. The program looks for connections of units which are historically known to be bad (e.g., gauche$^-$ gauche$^+$ pentane) or which have been found to be bad in an earlier stage of computation.

(5) The symbolic suggestions are translated into coordinate representations combining the template coordinates. Since neighboring templates are overlapping, two templates can be joined by a least squares fit of the coordinates of the common atoms.[42] The program has several weighting schemes for the common atoms. For instance, a substituent atom of a cyclic unit gets a lower weight than the atoms of the cycle itself. Different matching strategies are used for fused rings, for spiro rings, or for bridged systems. The coordinate representations are automatically criticized



7

Figure 12. Schematic representation of a criticized conformation of 9-amino[3.3.1]bicyclononane-9-carboxylic acid (7).

after each combination step. Critics are the quality of the fit and problems from long range interactions. The quality of the fit is characterized by the RMS value of the matching atom positions. Types of long-range interactions are hydrogen bridges or close van der Waals contacts.

(6) If no noncriticized conformation can be found, the least criticized suggestions are chosen for further refinement. Another tree search is performed looking for conformational units which can be deformed in order to solve the problem (i.e., changing one torsional angle in an acyclic unit or assigning a deformed template to a cyclic unit). Figure 12 shows a criticized conformation of 9-amino[3.3.1]bicyclononane-9-carboxylic acid (7). The program detects a close contact between the hydrogens in the 3- and 4-positions and a possible hydrogen bond between one carboxylic oxygen and one amino hydrogen.[45] The problem is solved by using twisted ring templates for the two cyclohexane units and by changing the torsional angles of the C_2–N_1 and C_2–C_3 bonds and closing the N_1–C_2–C_3 angle.

WIZARD and COBRA accept molecules with up to 200 atoms given in a number of various file formats.

The strength of the approach lies in the extensive use of symbolic representations for the suggested conformations and the use of optimum geometries for the coordinate representations of the templates. This makes the algorithm some orders of magnitude faster than numerical methods like distance geometry. It allows the construction of high quality molecular models without further optimization. In addition, when different conformations are possible, no conformation will be overlooked, and a sequence of desired conformations may be produced. Problems may arise when templates are lacking or fit only imperfectly. In other words, the quality of the result for a given problem strongly depends on whether suitable templates are contained in the library or not. On the other hand, the addition of a new template to the library requires database searches on X-ray structures and/or molecular mechanics calculations. In a recent publication[49] Leach and Smellie presented a combined model-building and distance geometry approach which tries to overcome the problem of lacking conformational templates by performing a conformational search by distance geometry calculations. However, a substantial loss of speed in the calculation is the price that has to be paid for this extension.

C. CONCORD

The commercially available program CONCORD of Pearlman[50–52,7b] is the most widely used method for converting large databases of 2D structures to 3D representations.[14,15,17] The program is based on rules and a simplified force field method. It performs the following steps for model building.

(1) The input structure is analyzed and separated into ring systems and acyclic atoms. Two rings are regarded to belong to one and the same ring system if they both have at least two atoms in common with another ring of the same system. Thus, spiro-connected rings are handled separately.

(2) Bond lengths and bond angles are taken from a table. They depend on atom type and bond order. The atom types are rather detailed and consider hybridization state and some first sphere neighbor atoms or small ring size. Thus, for carbon, 21 atom types are considered.

(3) Ring systems are processed by the assignment of a general conformation (e.g., "chair", "boat", etc.) to each ring. These general conformations reflect constraints from the conformations of the other rings of the entire ring system. Then, the rings are ordered according to a certain priority and are optimized in steps in this order by the minimization of a special strain function. The coordinates of rings already previously processed (on a higher level of priority) remain unchanged. Endocyclic bond angles and torsional angles are simultaneously changed in order to get perfectly closed rings and minimal steric energies.

(4) Finally, the torsional angles of the acyclic parts are set to values which minimize the steric interactions of the largest 1,4-interactions.

CONCORD considers the elements H, C, N, O, F, Si, P, S, Cl, Br, and I, and is able to process molecules with up to 200 non-hydrogen atoms. Furthermore, the maximum connectivity (coordination number) of an atom is four. For multifragment compounds CONCORD only models the largest fragment and passes the smaller fragments through. The produced structure is one single low-energy conformation. The program accepts a number of various input file format like SMILES strings[53] or MOLFILE[54] format.

Conversion rates of between 86% and 91%[14] and of 88%[17] have been found in using CONCORD to convert large databases. An average conversion time of 0.56 s per molecule on a VAX 11/8700 was reported.[17]

In a recent study Milne et al.[22] presented a comparison of structures generated by CONCORD with X-ray crystallographic data. They used a dataset of initially 194 crystal structures taken from the Cambridge Structural Database (CSD).[8] CONCORD was able to process 134 structures (69%). Since no appropriate coordinates were stored in the CSD for another 27 structures and the stereochemistry was ambiguous for another 17 there remained 90 structures with both experimental and CONCORD generated structures. The RMS value of the non-hydrogen positions was chosen as criterion for comparing these structures. The results where classified on the basis of the number of free rotable bonds in the molecules as a measure of flexibility. It was found that CONCORD manages rigid structures very well but fails in cases of flexible molecules. Especially for large rings rather poor geometries have been created. In 41% of all cases the RMS value was less than 0.5 Å and for these the two geometries were regarded to be essentially identical.

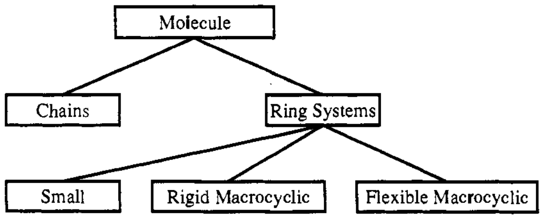


Figure 13. Fragmentation scheme of the program CORINA.

D. CORINA

Extending an earlier approach of Hiller and Gasteiger,[18] Gasteiger, Rudolph, and Sadowski have developed the 3D structure generator CORINA.[19-21] The program was developed for the reaction prediction system EROS[55,56] in order to model the influence of the spatial arrangement of the atoms in a molecule on its reactivity. Therefore, the approach had to be applicable to the entire range of organic chemistry including reactive intermediates, macrocyclic, and organometallic compounds. In order to handle large amounts of hypothetical structures it had to be automatic and rapid. The program performs the following steps in generating a 3D model.

(1) Bond lengths and bond angles are set to standard values taken from a table. Bond lengths depend on the atom types, the atomic hybridization states, and the bond order of the regarded atom pair. For bond types not found in the table reasonable values are calculated from covalent atomic radii and electronegativities. Bond lengths in conjugated systems are relaxed using a Hückel MO scheme. Bond angles only depend on the atom type and the hybridization state of the central atom. Atoms with up to six neighbors can be handled, using one of the following elementary geometry types according to the VSEPR model:[57] terminal, linear, planar, tetrahedral, trigonal bipyramidal, or octahedral. The tables of bond lengths and bond angles are parameterized for the entire periodic table.

(2) The molecule is fragmented into ring systems and acyclic parts. Ring systems contain the ring atoms plus the exocyclic atoms directly bonded to ring atoms. The exocyclic atoms are included since their positions and their long range interactions are strongly influenced by the ring conformation. Two rings belong to the same ring system if they have at least one atom in common with another ring of the same system. The ring systems are further classified into "small-ring systems" that include rings with up to eight atoms, "rigid macrocyclic systems" that include large rings with low flexibility that is limited by bridges or fused rings, and "flexible macrocyclic systems" containing one flexible large ring which may be fused or bridged only to a limited number of small rings. Figure 13 illustrates this process.

Since the conformational flexibility of ring systems cannot be handled simply by varying some torsional angles CORINA is able to produce a list of conformations for ring systems. By default, only the conformation with lowest energy is written to the output, but the other conformations can be obtained, too.

(3) Small-ring systems can be handled by using a table of allowed single-ring conformations since rings of sizes three to eight have a limited number of conformations available. These templates are stored as lists of torsional angles depending on the distribution of unsaturations in the rings. They are characterized and ordered by a strain energy value describing the conformational energy. Figure 14 shows the list of conformations for cyclohexane and cyclohexene.

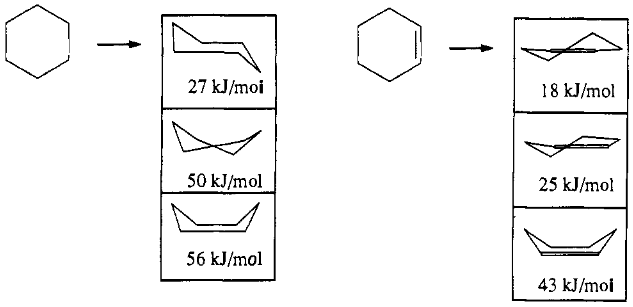


Figure 14. Ring templates for cyclohexane and cyclohexene.

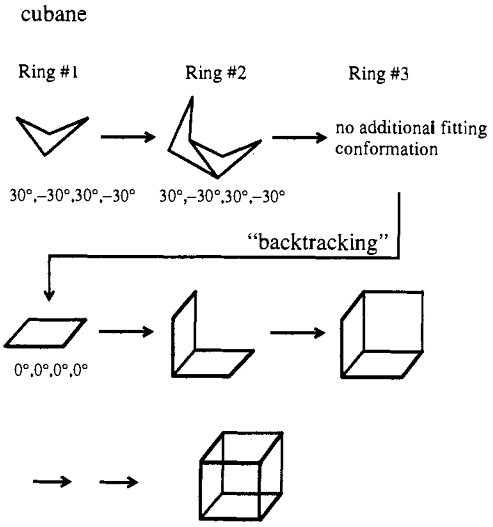
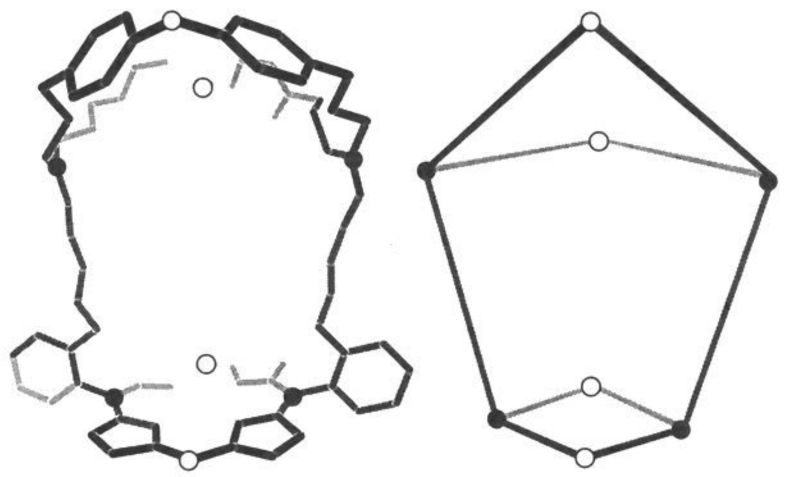


Figure 15. Backtracking procedure for the generation of a 3D model of cubane.

A backtracking algorithm is used for a ring system consisting of more than one ring being fused or bridged to find possible combinations of the conformations of the single rings. First, all rings of the smallest set of smallest rings (SSSR) are ordered according to their priority P following eq 7. The priority P of a single ring

$$P = \frac{\sum_{i}^{n} M_i}{n} \qquad (7)$$

is the sum of the Morgan numbers[58] M_i of the ring atoms i weighted by the ring size n. The priority is a measure of how central the ring lies within the ring system. Central rings are processed first since they have the highest number of interactions with neighboring rings. Second, lists of possible conformations are assigned to all rings. These conformations are ordered by increasing strain energy. Third, possible combinations of ring templates are searched in a backtracking procedure rotating and flipping the torsional angles lists of the ring templates. Figure 15 illustrates this procedure for cubane. The lowest energy conformation for a four-membered ring has the sequence (30, −30, 30, −30) of torsional angles. Two such rings can be fused together, but a third one cannot be fitted to them. This initiates the backtracking algorithm to try the second possible conformation (0, 0, 0, 0). Since this is impossible for



8

Figure 16. A macrocyclic molecule 8 and the corresponding superstructure. The bridgehead atoms in common are marked by filled circles and the anchor atoms by empty circles.

the third ring as well as for the second ring, the planar form is finally assigned to the first ring and the entire cubane skeleton is successively built from planar rings.

In this process, the deviation of the torsional angles of the bonds in common to two neighboring rings is checked. It must be less than a preset tolerance value. This value is set to 7° but can be relaxed to values of 25° and 40° on several levels of generation if no valid combination of ring templates can be found. This limitation of the deviation of torsional angles implies the use of allowed combinations of torsional patterns for cis/trans fused rings and for bridged systems. Continuing the backtracking procedure, a list of symbolic conformations of the ring systems is generated until no more combinations can be found or a maximum number of 10 000 conformations is reached. These conformations are ordered by summing a symbolic energy value E_{Sj} for each ring j (eq 8).

$$E_{S,j} = E_{TA} + E_{P} + 10\sum \Delta TA_i + 1000\Delta C \qquad (8)$$

Where E_{TA} = the strain energy of the ring template used, E_P = the Pitzer strain energy caused by acyclic substituents, ΔTA_i = the torsional angle deviation of the ith bond in common with a neighboring ring, and ΔC = the deviation of the unsaturation patterns of the real ring and the ring template used.

The Pitzer strain energy is calculated by a rather simple increment scheme for monoaxial substituents. For 1,2-diequatorial pairs of substituents half the sum of these increments is added. The last term, ΔC enables the program to use also ring templates with differing unsaturation patterns. The best conformations obtained from eq 8 are then translated into 3D atomic coordinates using the standard bond lengths and bond angles, and the torsional angles of the symbolic conformations. Since heteroatoms and strained systems may cause imperfect ring closure a pseudo-force field calculation (*vide infra*) is performed in order to optimize the ring geometries.

(4) Rigid polymacrocyclic systems cannot be handled with the procedure given above for small-ring systems. No conformations are available from the table of ring templates for rings with a size larger than eight. However, polymacrocyclic structures quite often show an overall general outline, a superstructure. For example, the polymacrocyclic molecule 8 in Figure 16 shows a cage-like superstructure that retains the approximate shape and symmetry of the molecule and is shown at the right-hand side of Figure 16. The procedure for generating a 3D structure for polymacrocyclic systems is based on the so-called "principle of superstructure". The general steps of building a 3D model of rigid polymacrocyclic systems are as follows. First, the polymacrocyclic system is reduced to its essential topological features, the superstructure. This superstructure retains only the number of macrocycles and the bridgehead atoms of the original system. If two bridgehead atoms are directly connected by two bonds, additional so-called anchor atoms are inserted. Small rings like the six-membered rings of 8 are removed as well. Second, a 3D model of the superstructure is generated following the algorithms for small-ring systems, but using long (super-) bonds. The length of these bonds is determined by the number of atoms between the bridgehead or anchor atoms. Third, the atoms originally removed are inserted alternatingly to the left and the right of the long bonds to come up with favorable torsional angles. In addition, small ring systems which have been removed from the superstructure, are added again superimposing them with the 3D superstructure. The resulting approximate 3D model is further refined by a pseudo-force field calculation (*vide infra*).

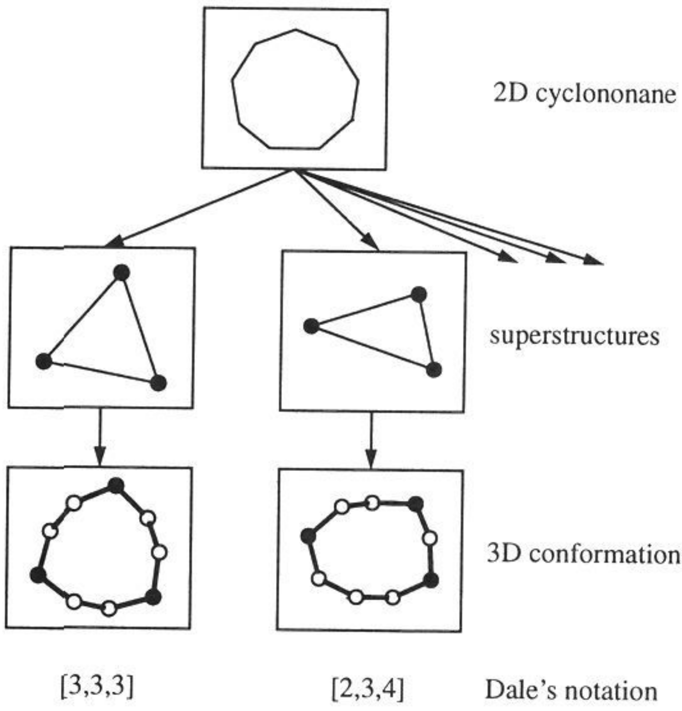


Figure 17. Superstructures for cyclononane resulting in more than one conformation.

(5) For flexible macrocyclic systems the above principle of superstructure cannot be used for a simple reason. For a flexible large ring like, e.g., cyclononane one cannot define a single superstructure since no bridgehead atoms are involved in the system. At least three anchor atoms must be chosen in order to define a cyclic structure. But the choice of these atoms is completely arbitrary and may result in more than on superstructure and furthermore in more than one conformation of the macrocycle as illustrated in Figure 17.

From this fact it becomes clear that such systems can only be handled by generating and evaluating several conformations. A simple conformational analysis procedure for large rings was developed,[21] which is based on Dale's notation[59] for the conformations. This notation is based on the assumption, that low-energy conformations of large rings take a polygon shape like the superstructures in Figure 17. The notation consists

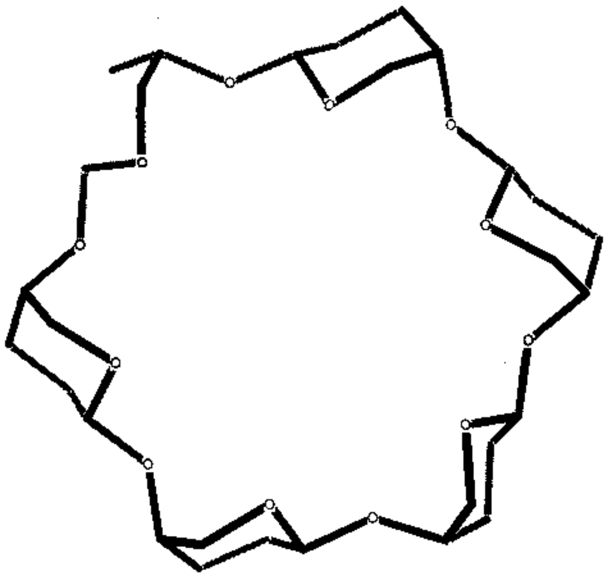


Figure 18. 3D structure generated for an α-cyclodextrin (only the ring atoms are shown).

of linear codes of the number of bonds between the corner (anchor) atoms that define the polygon. Thus, the triangles in Figure 17 obtain the codings [3,3,3] and [2,3,4]. These one-dimensional (1D) symbolic representations can quickly be generated and directly be translated into 3D atomic coordinates, constructing the specified polygons. A simple linear combination of features calculated from the linear notations allows an energy ranking of the 1D conformations (eq 9).

$$E_S = 8.1/N_E - 3.9/(N_A - 8)^6 + 4.7E_1 + 2.7E_2 - 1.1E_{1,2} - 0.9E_{1,>2} \quad (9)$$

Where E_S = the symbolic strain energy [kcal/mol], N_E = the number of edges of the polygon, N_A = the number of ring atoms, E_1,E_2 = the number of edges of lengths 1 and 2 in the polygon, and $E_{1,2}$, $E_{1,>2}$ = the number of neighboring edges of lengths 1,2 and 1,>2.

This energy function was scaled using force field strain energies of 51 conformations of the nine- to twelve-ring hydrocarbons. It was obtained by linear regression with a correlation coefficient of 0.89 and a standard deviation of the strain energy of 1.1 kcal/mol. It was found to be accurate enough to order the 1D conformations within an energy window of 1–3 kcal/mol. Since the 1D representations are directly correlated to the torsional angles within the large ring it is, in an easy way, possible to connect the geometry of a large ring with the geometries of small-ring fragments fused to it. Figure 18 shows as an example the obtained 3D structure for a cyclodextrin. The conformation of the macrocycle was obtained from the linear code [5,5,5,5,5,5].

(6) A pseudo-force field is used to optimize the geometries obtained by the above algorithms for ring systems. Two assumptions are made. First, rather rigid ring systems will be optimized. Thus, torsional energies and nonbonded interactions can be regarded as second-order influences on the geometries. Second, the major aim is geometry optimization instead of energy calculation. Thus, no real energy values have to be computed. These two assumptions lead to a rather simplified so-called pseudo-force field with a reduced number of energy terms and rather general parameters applicable to the entire range of organic chemistry. In addition, the energy functions are directly derived from geometrical considerations instead of physical functions. The energy E is the sum of stretching (E_S) energies, bending (E_B) energies, out-of-plane (E_O) energies, torsional energies of conjugated bonds (E_{TC}) and torsional energies in small four- and five-membered rings ($E_T{}^{4,5}$) (eq 10). The E_S and E_B values are obtained

$$E = E_S + E_B + E_O + E_{TC} + E_T{}^{4,5} \quad (10)$$

from a simple distance criterion like it is used in the error function of the distance geometry methods[25] (eqs 11 and 12).

$$E_{Sij} = 20.0(d_{0ij}{}^2 - d_{ij}{}^2)^2 \quad (11)$$

where d_{0ij} is the standard bond length between atoms i and j and d_{ij} is the actual value.

$$E_{Bijk} = 2.0(d_{0ik}{}^2 - d_{ik}{}^2)^2 \quad (12)$$

where d_{0ik} is the standard distance between atoms i and k, calculated from the standard bond angle i–j–k and the standard bond lengths i–j and j–k.

In order to obtain planarity for sp^2 atoms and conjugated bonds a volume criterion is used (eqs 13 and 14).

$$E_{Oijkl} = 5.0\sigma_{ijkl}{}^2 \quad (13)$$

where σ_{ijkl} is the volume of the parallelepiped, that is formed by the vectors from the central sp^2 atom i to the ligand atoms j, k, and l. In the nonstrained case of an ideal planar configuration σ takes a value of zero. The same volume is used to obtain planar geometries for conjugated bonds. Here, the vectors ij, ik, and il are calculated from the positions of four sequentially connected atoms i–j–k–l defining a planar torsional angle (eq 14).

$$E_{TCijkl} = k_{TC}\sigma_{ijkl}{}^2 \quad (14)$$

where k_{TC} is a force constant specific for the type of the conjugated bond. For single bonds in conjugated systems like the central bond in 1,3-butadiene a value of k_{TC} = 1.0 is taken, for aromatic or double bonds k_{TC} is set to 5.0.

In four- and five-membered rings the stretching and bending forces as formulated in the eqs 11 and 12 tend to flatten the starting geometries of favorable folded or envelope ring geometries. Thus, for the torsional angles i–j–k–l in four- and five-membered rings an additional energy term of the reciprocal volume σ_{ijkl} is used (eq 15).

$$E_{Tijkl}{}^{4,5} = 3.5/(6.0 + \sigma_{ijkl}{}^2) \quad (15)$$

This pseudo-force field calculation is only applied to ring atoms. In this way the adjustment of bond lengths, bond angles, and torsional angles in ring systems is rapidly achieved, converging after few iterations through the minimization procedure.

(7) Acyclic chains are stretched as much as possible following the "principle of longest pathways". The longest pathways or main chains in the acyclic parts of a molecule are defined according to two rules. First, a bond may participate in only one pathway. Second, an atom may participate in more than one pathway. Along these pathways the torsional angles are set to a trans configuration if not a cis double bond was specified. This principle is rather straightforward but it effectively minimizes nonbonded interactions.

(8) A reduced conformational analysis is performed in those rare cases that have inappropriate long-range interactions after all cyclic and acyclic fragments have been combined. This procedure starts with the assumption that problems resulting from nonbonded interactions between two atoms can be solved by changing one torsional angle within the pathway that connects these two atoms. First, for each overlapping atom pair a set of rotable bonds (i.e., bonds not involved in a ring and not conjugated) within the pathway between these atoms is determined. Second, descriptors for these bonds are calculated from the topological distance to the pair of atoms that should be moved apart and from the π-character of the bond. These increments are added if a bond is involved in more than one pathway. Thus, these bonds get a higher descriptor. Third, a minimal subset of strategic rotable bonds with maximum descriptors is chosen. Fourth, a systematic conformational analysis is performed changing the torsional angles at the strategic bonds in steps. The conformations are evaluated using 12-6-Lennard-Jones potentials for the nonbonded interactions and simplified torsional energy terms. This reduced conformational analysis leads in short computation times to a low-energy conformation, which is free of problems from nonbonded atoms interactions.

CORINA accepts molecules given in a variety of file formats such as MDL MOLFILE's.[54] The maximum number of atoms is not explicitly limited by the program. Molecules with about 300 non-hydrogen atoms have been processed without problems. CORINA is applicable to the entire periodic table. The input structures must be expressible in a valence bond description. The maximum connectivity of an atom is six. Multiple fragments are allowed. Intermolecular interactions or hydrogen bonds are not explicitly handled. On request, up to 20 different conformations of the ring systems ordered by the strain energy are written onto the output file.

The program was tested using a dataset of 639 X-ray structures from the Cambridge Structural Database[8] (CSD). The testset was obtained by choosing 19 of the chemical classes of the CSD that comprise organic molecules with at least one ring. These were about 25 000 structures or 28% of the entire CSD. This set was then reduced to 568 structures by selecting only those structures that had an R factor less than 3%. By removing structures with errors in the coordinates or in the valence bond notation and by splitting cases with more than one molecule into single molecular structures, provided a dataset of 639 molecules. Finally, stereodescriptors for molecules containing chiral centers or stereochemistry at double bonds or ring centers were calculated from the 3D coordinates. CORINA was able to process 100% of the test structures—an extremely high conversion rate. The CPU time on a Sun SPARC station 10/20 was 0.61 s per molecule on average. The results obtained were compared to the X-ray geometries. Figure 19 shows histograms of the RMS deviations of the atom positions of the X-ray structures from those generated by CORINA. The positions of all non-hydrogen positions (a) and of the ring atoms only (b) have been compared. The conformations of structures with an RMS deviation of less than 0.3 Å can be regarded as being essentially equal. These where 42% of the test data set comparing all non-hydrogen atoms and 89% comparing the ring atoms only. Thus, CORINA modeled most of the ring structures with a high accuracy. For more than one third of the structures the X-ray geometry was reproduced including also the flexible parts of the molecules.

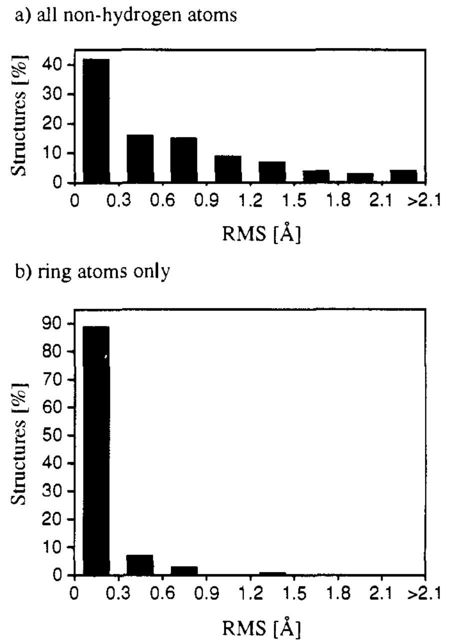


Figure 19. Histograms of the RMS deviations of atom positions of the X-ray geometries compared to those generated by CORINA: (a) all non-hydrogen atoms; (b) ring atoms only.

A study is under way for testing the performance of other available 3D structure generators with this dataset.[60]

V. Data-Based Methods

A. AIMB

Wipke and Hahn[61-64] have developed a unique 3D model building technique that is based on finding near analogies of a molecule or of substructures of it in a database of 3D molecular structures: AIMB (Analogy and Intelligence in Model Building). A human expert is able to construct a 3D model in a very efficient, nonnumerical, and fast manner, reasoning by analogy on the basis of his knowledge on similar problems. The program tries to automate this method with knowledge already captured by crystallography and stored, e.g., in the Cambridge file. The basic idea is that a large and widespread data collection of experimental molecular geometries contains implicitly the "knowledge" of the molecules themselves on model building. The following steps are performed by the different components of the method.

(1) The "knowledge base" (KB) of AIMB was constructed from the Cambridge Crystallographic Database, selecting organic molecules (C, N, O, P, S, Si, B, F, Cl, Br, and I) with less than 65 non-hydrogen atoms. Structures with atoms having a coordination number of more than five, polymer structures, and poor crystal structures were removed. Hydrogens were removed because their positions are normally experimentally not determined. This subset of the Cambridge file was processed to generate abstractions that are hierarchi-

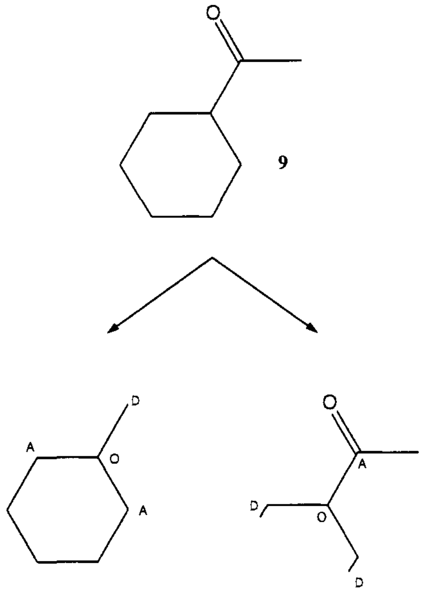


Figure 20. Subproblems of methyl cyclohexyl ketone (**9**). Origin (O), α (A), and dummy (D) atoms are marked.

cally ordered for rapid access. Since there is nothing in the program which requires the above limitations any other kind of KB would be possible.

(2) The "problem analyzer" perceives the target structure to identify rings, chains, aromaticity, and stereochemistry. If the target or a close analogy is not contained in the compound library, the "decomposer" uses a "divide and conquer" strategy to create substructures of the target and to treat them as new problems. The subdivision strategy follows the rule that there is maximum interaction within a unit and minimum interaction between units. First, the target is subdivided into ring assemblies and chains. If the program again fails in finding an analogy in the KB, the subproblems may be divided further. Ring assemblies are subdivided only once more into elemental single ring units which cannot be divided further. A chain can be broken down into elemental chain fragments of simple bonded atom pairs. If an elemental subproblem cannot be solved the model building process is aborted. In addition, the atoms of a subunit are weighted differently. These weights are assigned to atoms in descending order of priority: Origin atoms which form the join to another unit, α-origin atoms which are nonorigin atoms in α-position to an origin atom, real atoms—all remaining atoms of a unit, and in addition, dummy atoms which are attached to origin atoms and contain some information about the chemical environment around the unit (e.g., rings, substituents, etc.). Figure 20 shows the division of methyl cyclohexyl ketone (**9**) into subunits.

(3) The "analogy finder" searches for analogies of the subproblems in the KB. The hierarchical structure of the KB allows one to probe the file at different levels of abstraction. If no exact expressions of a subproblem can be found the matching tolerances are increased until an analogy is found. Typically this search is continued until a maximum search depth of five to 10 analogies is reached. This search strategy on several levels of abstraction guarantees that the best analogies are found first.

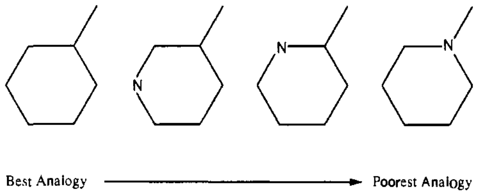


Figure 21. Analogies for the cyclic suproblem of **9** (Figure 20) in decreasing order of similarity: exact match, real atom mismatch, alpha-origin atom mismatch, and origin atom mismatch.

(4) The "analogy evaluator" scores each found analogy in order to select the best analogies. The problem is 2-fold: First, a similarity measure must be defined which describes reasonably the distance between different analogies. Second, the mapping problem of projecting the target atoms onto the analogy atoms has to be solved, i.e., all possible mappings of the target and the analogy are to be explored, and it cannot be assumed that the target and the analogy are isomorphic. Since there are some constraints on atom and bond mapping (e.g., nondummy atoms must always be mapped onto nondummy atoms), not all permutations are to be checked. The similarity score of an analogy is calculated following eq 16:

$$S_I = \sum_{J=1}^{\mathrm{NA}}\sum_{K=1}^{\mathrm{NP}} W_K(A_J)\cdot D_K(A_J,A'_L) \qquad (16)$$

Where S_I = analogy score of the Ith map, NA = number of atoms in the subproblem, NP = number of attributes to be evaluated, A_J = the Jth atom of the subproblem, A'_L = the Lth atom of the analogy where L = map_I (J), W_K = weighting factor of the Kth attribute, and D_K = dissimilarity of the Kth attribute for the mapped pair.

The attributes include atom type, charge, valence, hybridization, and stereochemistry. The weighting factor differentiates between origin, α-origin, real, and dummy atoms and takes values between 40 and 100. The dissimilarity reflects the difference between different attribute values and takes values between 0 and 10. Figure 21 shows several analogies of the cyclic subunit of methyl cyclohexyl ketone (**9**) in descending order of similarity.

(5) The "model assembler" combines the analogies found for the subproblems to a coordinate representation of the original problem. The combination is performed in steps, superimposing the origin and dummy atoms of the subunits. The resulting differences in bond lengths and bond angles between both the welded fragments are calculated as a measure of the quality of the fit.

The described algorithms rapidly build reasonable 3D molecular models which represent minimum-energy conformations. The results are explained to the user using the information on the structures where the analogies are taken from. Since several analogies can be found for the subunits depending on the search depth, it is possible to perform a conformational search. Although the program does not contain any energy evaluation procedure and does not take into account long-range interactions, it was shown in several cases that the fragments used "know" something about these problems. A helical model of pentahelicene was built from single benzene rings since the best analogies found where taken from other helicenes. In this way, implicit knowledge on energy and long-range interactions can be extracted from the KB.

Several investigations have been performed to characterize the program's performance. First, the problem-solving speed was studied as a function of the library size. It was shown for knowledge bases of 500, 1 000, 5 000, and 10 000 of 3D geometries that there is a substantial increase in speed with increasing size of the KB. The larger and more widespread the KB is the earlier AIMB finds good analogies of the subproblems. Second, the model quality was tested against the size of the KB. It was found that the better models were constructed the larger the KB was. Third, the speed versus the search depth was explored. The time needed for the model construction increased linearly with the search depth, i.e., the desired number of analogies for each unit. Finally, the speed as a function of the target complexity was studied. The time per molecule increased linearly with the number of atoms and with the number of subunits in the target.

The strength of the method is its speed and that it is exclusively based on experimental 3D structures and fast database searching techniques. The models built are as accurate as X-ray structures and can be explained by the parent structures where the subunits are taken from. One of the most interesting qualities of AIMB is its ability to build more accurate models more rapidly the more knowledge is present in the KB. The only limitation in the range of chemistry that can be handled is the contents of the KB. Therefore, a possible difficulty of the program is that the quality of the models built strongly depends on the quality of the database of 3D structures available as KB. Problems may also arise from the requirement of a large amount of disk space for the KB. Another problem may be the use of redundant information since a lot of substructures with very similar geometries, e.g., benzene rings, are contained many times in the library.

B. Chem-X Builder

Chemical Design Ltd. have developed a 2D-to-3D builder of their own[16,65,66] which assembles fragments retrieved from a database similar to the AIMB program by Wipke and Hahn.[61–64]

The heart of the 3D builder is a relatively small library of common ring substructure fragments containing specific carbocyclic and heterocyclic groups together with generalized fragments with unspecified atom types. Furthermore, the fragments are characterized by different patterns of unsaturation and by stereochemistry. The default library contains about 100 preoptimized cyclic structures. The model builder first tries to find exact matches of the cyclic substructures in the library. If no exact match can be found, generalized fragments are taken. Ring systems may be handled as whole fragments or as single-ring structures, which are fitted together. Acyclic parts of the molecule are constructed with torsional angles of the main chains in extended form. If more than one hit is found for a fragment, a conformational search can be performed. A special handling of stereochemistry allows the generation of different stereoisomers which is useful in converting databases not containing stereoinformation. The range of validity of the model builder can be extended by updating of the library of ring fragments, but this slows down the program.

The 2D-to-3D builder accepts several file formats including SMILES strings[53] and MOLFILE.[54]

The program was used to convert large databases[16] at Chemical Design Ltd. A conversion rate of 87% and conversion times of 0.5–30 s per molecule on an IBM RS/6000 have been reported.

The program seems to be more general than the AIMB approach since side chains are constructed straightforward instead of taking them from the library. Its major strength is the speed of the coordinate generation. Its major weakness is the rather simple construction scheme for side chains which may result in problems from long-range interactions. The strategies used seem to be simpler than those used in AIMB and the models produced lack the explanation capabilities of AIMB. Especially the library search strategies seem to be less efficient since the addition of new fragments to the knowledge base slows the program down in contrast to the AIMB program where the speed increases with the size of the database.

VI. Concluding Remarks

The representation of chemical structures for computer management has gone through several levels of sophistication in the last 25 years. First, linear notations and fragment codes were introduced. This was followed by connection tables that give explicit account of the atoms and bonds in a molecule and are thus a direct reflection of the constitution of a molecule. Connection tables are now in general use. However, the need for a more detailed representation of chemical structures, for access to the three-dimensional coordinates of the atoms in a molecule, is increasingly felt. Particularly, problems in drug design can only be solved by analysis of the 3D structure of molecules.

Automatic 3D structure generation, the conversion of a connection table into a 3D molecular model has been pursued in the last years in both academic research and commercial software development. Surveying the approaches developed since 1980 it was shown that a number of interesting solutions to this fundamental task of computational chemistry exists. The various methods derive their knowledge to various degrees from data of experimental or computed geometries or from rules about the construction of molecular models. Usually the initial guess of a geometry is further refined by empirical calculations. Great care has been devoted to making these 3D structure generators as rapid as possible in order to apply them to large datasets of molecules.

The users of 3D structure generation programs, especially database developers, will increasingly become interested not only in the speed of the conversion programs but also in their accuracy, their conversion rate, and their explanation capabilities which allow the user to evaluate the way the 3D structure was generated. The explanation capabilities should include both run-time explanation which is specific for a structure processed and full publication of the general algorithms inherent in the programs. Another trend is the development of conformational analysis programs which do their job as fast as programs do which provide only one single low-energy conformation of the input structure.

Three-dimensional structures obtained from these generators can serve in a variety of applications. They can provide reasonable starting geometries for molecular mechanics and quantum mechanical calculations of various degrees of sophistication. They give models for the experimental determination of 3D structures in solution by NMR techniques. Most important, however, is the generation of datasets of 3D structures for use in searching for pharmacophores and new lead structures with desirable biological, chemical, or physical properties.

VII. References

(1) Gray, N. A. B. In *Chemical Structure Systems*; Ash, J. E., Warr, W. A., Willett, P., Eds.; Ellis Horwood: New York, 1991; pp 263–298.

(2) Loftus, F. G. In *Chemical Informations Systems—Beyond the Structure Diagram*; Bawden, D., Mitchel, E. M., Eds.; Ellis Horwood: New York, 1990; pp 93–104.

(3) Seydel, J. K. *QSAR and Strategies in the Design of Bioactive Compounds*; VCH Publishers: New York, 1985.

(4) Franke, R. *Theoretical Drug Design Methods*; Elsevier: Amsterdam, 1984.

(5) (a) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. *Ab Initio Molecular Orbital Theory*; John Wiley and Sons: New York, 1986. (b) Stewart, J. J. P. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1990; Vol. 1, pp 45–82. (c) Dykstra, E. C.; Augspurger, J. D.; Kirtman, B.; Malik, D. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1990; Vol. 1, pp 83–118.

(6) Burkert, U.; Allinger, N. L. *Molecular Mechanics*; ACS Monograph 177; American Chemical Society: Washington, DC, 1982.

(7) (a) Martin, Y. C.; Bures, M. G.; Willett, P. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1990; Vol. 1, pp 213–263. (b) Pearlman, R. S. In *Emerging Technologies and New Directions in Drug Abuse Research*; Rapake, R., Ed.; Row Scientific: Washington, DC, 1991; pp 62–77.

(8) Cambridge Structural Database: 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. *Acta Crystallogr. Sect. B: Struct. Crystallogr. Cryst. Chem.* **1979**, *B35*, 2331.

(9) CAST-3D: CAS 3D Structure Templates File, CAS-RF: CAS Registry File, available from Chemical Abstracts Service, Columbus, OH.

(10) MDDR-3D: MACCS-II Drug Data Report-3D, FCD-3D: Fine Chemicals Directory-3D, available from Molecular Design Ltd., San Leandro, CA.

(11) (a) Chapman & Hall Chemical Databases, available from Chemical Design Ltd., Oxford, England. (b) *Chem. Des. News* **1992**, Spring/Summer.

(12) Dreiding, A. S. *Helv. Chim. Acta* **1959**, *42*, 1339.

(13) (a) Dale, J. *Stereochemistry and Conformational Analysis*; VCH Publishers: New York, 1978; pp 113–170. (b) Saunders, M. *J. Comput. Chem.* **1991**, *12*, 645.

(14) Henry, D. R.; McHale, P. J.; Christie, B. D.; Hillman, D. *Tetrahedron Comput. Method.* **1990**, *3* (6C), 531.

(15) Fisanick, W.; Cross, K. P.; Rusinko, A., III. *Tetrahedron Comput. Method.* **1990**, *3*, 635.

(16) Davies, K.; Upton, R. *Tetrahedron Comput. Method.* **1990**, *3*, 665.

(17) Rusinko, A., III; Sheridan, R. P.; Nilakantan, R.; Haraki, K. F.; Bauman, N.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 251.

(18) Hiller, C.; Gasteiger, J. In *Software-Entwicklung in der Chemie*; Gasteiger, J., Ed.; Springer: Berlin, 1987; Vol. 1, pp 53–66.

(19) Gasteiger, J.; Rudolph, C.; Sadowski, J. *Tetrahedron Comput. Method.* **1990**, *3*, 537.

(20) Sadowski, J.; Rudolph, C.; Gasteiger, J. *Anal. Chim. Acta* **1992**, *265*, 233.

(21) Sadowski, J.; Gasteiger, J. In *Software Development in Chemistry*; Ziessow, D., Ed.; GDCh: Frankfurt/M., 1993; Vol. 7, pp 65–76.

(22) Hendrickson, M. A.; Nicklaus, M. C.; Milne, G. W. A.; Zaharevitz, D. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 155.

(23) See, for example: (a) Potenzone, R., Jr.; Cavicchi, E.; Hopfinger, A. J.; Weintraub, H. J. R. *Comput. Chem.* **1977**, *1*, 187. (b) Gund, P.; Andose, J. D.; Rhodes, J. B.; Smith, G. M. *Science* **1980**, *208*, 1425. (c) Dyott, T. M.; Stuper, A. J.; Zander, G. S. *J. Chem. Inf. Comp. Sci.* **1980**, *20*, 28. (d) Liljefors, T. *J. Mol. Graphics* **1983**, *1*, 111. (e) White, D.; Pearson, J. *J. Mol. Graphics* **1986**, *4*, 134.

(24) Crippen, G. M. *J. Comput. Phys.* **1978**, *26*, 449.

(25) Crippen, G. M.; Havel, T. F. *Distance Geometry and Molecular Conformations, Chemometrics Research Studies 15*; Wiley: New York, 1988.

(26) Blaney, J. M.; Crippen, G. M.; Dearing, A.; Dixon, J. S. *QCPE Program* QCPE-No. 590.

(27) Leach, A. R. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1991; Vol. 2, pp 1–55.

(28) Crippen, G. M.; Havel, T. F. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 222.

(29) Crippen, G. M. *J. Comput. Chem.* **1992**, *13*, 351.

(30) Wenger, J. C.; Smith, D. H. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 29.

(31) Gordeeva, E. V.; Katritzky, A. R.; Shcherbukhin, V. V.; Zefirov, N. S. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 102.

(32) Wipke, W. T.; Verbalis, J.; Dyott, T. Three-Dimensional Interactive Model Building; Presented at the 162nd National Meeting of the American Chemical Society, Los Angeles, August 1972.

(33) (a) Wipke, W. T. In *Computer Representation and Manipulation of Chemical Information*; Wipke, W. T., Heller, S. R., Feldmann, R. J., Hyde, E., Eds.; John Wiley and Sons: New York, 1974; pp 147–174. (b) Wipke, W. T.; Smith, G.; Choplin, F.; Sieber, W. In *SECS—Simulation and Evaluation of Chemical Synthesis: Strategy and Planning*; American Chemical Society: 1977; Vol. 61, pp 97–127.

(34) (a) Corey, E. J.; Feiner, N. F. *J. Org. Chem.* **1980**, *45*, 757. (b) Corey, E. J.; Feiner, N. F. *J. Org. Chem.* **1980**, *45*, 765.

(35) Cohen, N. C.; Colin, P.; Lemoine, G. *Tetrahedron* **1981**, *37*, 1711.

(36) Grierson, D. S.; Vuilhorgne, M.; Husson, H.-P.; Lemoine, G. *J. Org. Chem.* **1982**, *47*, 4439.

(37) (a) De Clerc, P. J. *J. Org. Chem.* **1981**, *46*, 667. (b) *Tetrahedron* **1981**, *37*, 4277. (c) *Tetrahedron* **1984**, *40*, 3717. (d) *Tetrahedron* **1984**, *40*, 3729.

(38) Hoflack, J.; De Clercq, P. J. *Tetrahedron* **1988**, *44*, 6667.

(39) De Clercq, P. J.; Hoflack, J.; Cauwberghs, S. *QCPE Program* QCPE-No. QCMP079.

(40) Dolata, D. P.; Carter, R. E. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 36.

(41) Dolata, D. P.; Leach, A. R.; Prout, K. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 73.

(42) Leach, A. R.; Prout, K.; Dolata, D. P. *J. Comput.-Aided Mol. Des.* **1988**, *2*, 107.

(43) Dolata, D. P.; Leach, A. R.; Prout, K. In *Computer Aided Molecular Design*; Richards, W. G., Ed.; IBC Technical Services Ltd.: London, 1989; pp 67–82.

(44) Leach, A. R.; Prout, K.; Dolata, D. P. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 271.

(45) Leach, A. R.; Prout, K.; Dolata, D. P. *J. Comput. Chem.* **1990**, *11*, 680.

(46) Leach, A. R.; Prout, K. *J. Comput. Chem.* **1990**, *11*, 1193.

(47) Leach, A. R.; Dolata, D. P.; Prout, K. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 316.

(48) COBRA is available from Oxford Molecular Ltd., Oxford, England.

(49) Leach, A. R.; Smellie, A. S. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 379.

(50) Rusinko, A., III. Tools for Computer-Assisted Drug Design; Ph.D. Thesis, University of Texas at Austin, Austin, TX, 1988.

(51) Pearlman, R. S. *Chem. Des. Auto. News* **1987**, *2*, 1.

(52) CONCORD is available from Tripos Ass., St. Louis, MO.

(53) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.

(54) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244.

(55) Gasteiger, J.; Hutchings, M. G.; Christoph, B.; Gann, L.; Hiller, C.; Löw, P.; Marsili, M.; Saller, H.; Yuki, K. *Top. Curr. Chem.* **1987**, *137*, 19.

(56) Gasteiger, J.; Ihlenfeldt, W. D.; Röse, P. *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 270.

(57) Sidgwick, N. V.; Powell, H. M. *Proc. Roy. Soc.* **1940**, *A176*, 153.

(58) Morgan, H. C. *J. Chem. Doc.* **1965**, *5*, 107.

(59) Dale, J. *Acta Chem. Scand.* **1973**, *27*, 1115.

(60) Sadowski, J.; Gasteiger, J.; Klebe, G. Manuscript in preparation.

(61) Wipke, W. T.; Hahn, M. A. In *Applications of Artificial Intelligence in Chemistry*; Pierce, T., Hohne, B., Eds.; ACS Symposium Series 306; American Chemical Society: Washington, DC, 1986; pp 136–146.

(62) Wipke, W. T.; Hahn, M. A. *Tetrahedron Comput. Method.* **1988**, *2*, 141.

(63) Wipke, W. T.; Hahn, M. A. In *Chemical Structures*; Warr, W. E., Ed.; Springer: Berlin, 1988; Vol. 1, pp 267–268.

(64) Hahn, M. A.; Wipke, W. T. In *Chemical Structures*; Warr, W. E., Ed.; Springer: Berlin, 1988; Vol. 1, pp 269–278.

(65) Davies, K.; Dunn, D.; Upton, R. An Algorithm to Generate 3D Structures from 2D Connection Tables; 5th Molecular Modeling Workshop, Darmstadt, 1991; poster.

(66) The Chem-X 2D-to-3D builder is available from Chemical Design Ltd., Oxford, England.