action is an additional subject of further investigations.

## ACKNOWLEDGMENT

We thank the Fonds der Chemischen Industrie and the BAYER AG for financial support. C.B. thanks the Studienstiftung des deutschen Volkes for a scholarship.

## REFERENCES AND NOTES

(1) Kos, A. J.; Grethe, G. Reaktionsdatenbanken—Werkzeuge für den Synthese-Chemiker. *Nachr. Chem., Tech. Lab.* **1987**, *35*, 586–94.

(2) Zass, E.; Müller, S. Neue Möglichkeiten zur Recherche von organisch-chemischen Reaktionen—Ein Vergleich der ≪in-house≫-Datenbanksysteme REACCS, SYNLIB and ORAC. *Chimia* **1986**, *40*, 38–50.

(3) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS ONLINE Search System. 1. General System Design and Selection, Generation and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.

(4) Funatsu, K.; Endo, T.; Kotera, N.; Sasaki, S.-I. Automatic Recognition of Reaction Site in Organic Chemical Reactions. *Tetrahedron Comput. Methodol.* **1988**, *1*, 53–69.

(5) Dugundji, J.; Ugi, I. An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. *Top. Curr. Chem.* **1973**, *39*, 19–64.

(6) For example, see the following articles and further references cited there: (a) Ugi, I., et al. Neue Anwendungsgebiete für Computer in der Chemie. *Angew. Chem.* **1979**, *91*, 99–111. (b) Jochum, C.; Gasteiger, J.; Ugi, I. Das Prinzip der minimalen chemischen Distanz. *Angew. Chem.* **1980**, *92*, 503–13. (c) Jochum, C.; Gasteiger, J.; Ugi, I. The Principle of Minimum Chemical Distance and the Principle of Minimum Structure Change. *Z. Naturforsch.* **1982**, *37B*, 1205–15. (d) Brandt, J.; Bauer, J.; Frank, R. M.; von Scholley, A. Classification of Reactions by Electron Shift Patterns. *Chem. Scr.* **1981**, *18*, 53–60. (e) Brandt, J.; von Scholley, A. An Efficient Algorithm for the Computation of the Canonical Numbering of Reaction Matrices. *Comput. Chem.* **1983**, *7*, 51–9. (f) Wochner, M.; Brandt, J.; von Scholley, A.; Ugi, I. Chemical Similarity, Chemical Distance, and its Exact Determination. *Chimia* **1988**, *42*, 217–25. (g) Fontain, E.; Bauer, J.; Ugi, I. Computer assisted Bilateral Generation of Reaction Networks from Educts and Products. *Chem. Lett.* **1987**, 37–40.

(7) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–13.

(8) Lynch, M. F.; Willett, P. The Automatic Detection of Chemical Reaction Sites. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 154–9.

(9) Hutchins, R. O.; Hoke, D.; Keogh, J.; Koharski, D. Sodium Borohydride in Dimethyl Sulfoxide or Sulfolane. Convenient Systems for Selective Reductions of Primary, Secondary and Certain Tertiary Halides and Tosylates. *Tetrahedron Lett.* **1969**, 3495–8.

(10) Hojo, K.; Yoshino, H.; Mukaiyama, T. New Synthetic Reactions Based on 1-Methyl-2-fluoropyridinium Salts. Facile Conversion of Alcohols to Thioalcohols. *Chem. Lett.* **1977**, 133–6.

(11) Bates, G. S. New $\alpha$-Keto Acid Synthon; Alkylation of the Potassium Dianion of Bis(ethylthio)acetic Acid. *J. Chem. Soc., Chem. Commun.* **1979**, 161–3.

(12) Mukaiyama, T.; Yamaguchi, M.; Narasaka, K. A Regioselective Coupling Reaction of Allyl Pyridyl Ethers with Grignard Reagents. *Chem. Lett.* **1978**, 689–92.

(13) Solas, D.; Wolinsky, J. Total Synthesis of (−)-$\alpha$-Acoradiene and (−)-$\alpha$-Cedrene. *J. Org. Chem.* **1983**, *48*, 670–3.

(14) Hiyama, T.; Wakasa, N. Asymmetric Coupling of Asrylmagnesium Bromides with Allylic Esters. *Tetrahedron Lett.* **1985**, *26*, 3259–62.

(15) Fried, J.; Hallinan, E. A.; Szwedo, M. J. Synthesis and Properties of 7,7-DifluoroDerivatives of the 2,6-Dioxa[3.1.1]bicycloheptane Ring System Present in Thromboxane A2. *J. Am. Chem. Soc.* **1984**, *106*, 3871–2.

(16) Huggins, J. M.; Bergman, R. G. Mechanism, Regiochemistry, and Stereochemistry of the Insertion Reaction of Alkynes with Methyl-(2,4-pentanedionato)(triphenylphosphine)nickel. A Cis Insertion that Leads to Trans Kinetic Products. *J. Am. Chem. Soc.* **1981**, *103*, 3002–11.

(17) Hart, H.; Sasaoka, M. Exocyclic Benzenes. Synthesis and Properties of Benzo[1,2-*c*:3,4-*c*′:5,6-*c*″]trithiophene, a Tristhiahexaradialene. *J. Am. Chem. Soc.* **1978**, *100*, 4326–7.

(18) Cacioli, P.; Reiss, J. A. The Formation and Some Reactions of a Spirocyclic Chroman Derived from a 1-Oxaspiro[2.5]octa-5,7-dien-4-one. *Aust. J. Chem.* **1984**, *37*, 2599–605.

(19) Molecular Design Limited, 2132 Farallon Drive, San Leandro, CA 94577, or Molecular Design MDL AG, Wallstrasse 8, CH-4002 Basel, Switzerland.

(20) Borkent, J. H.; Onkes, F.; Noordik, J. H. Chemical Reaction Searching Compared in REACCS, SYNLIB, and ORAC. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 148–50.

(21) Bebak, H.; et al. The Standard Molecular Data Format (SMD Format) as an Integration Tool in Computer Chemistry. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 1–5.

(22) Shelley, C. A.; Munk, M. E. Computer Perception of Topological Symmetry. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 110–3.

# Smiles. 3. Depict. Graphical Depiction of Chemical Structures

DAVID WEININGER

Daylight Chemical Information Systems, 111 Rue Iberville, No. 610, New Orleans, Louisiana 70130

Received February 21, 1989

The DEPICT program converts SMILES, the linear notation of a chemical structure's molecular graph, into a depiction of molecular structure without user interaction. The resulting two-dimensional output display allows all aspects of SMILES representation of structure to be verified easily, including aromaticity, formal charge, bond order assignment, and hydrogen attachment. DEPICT is particularly well suited for computer-generated structures since it requires no manual or structural input. It is designed for use with SMILES notation, a lexical form of a connection table, and it follows that any other connection table can also be used with DEPICT provided only that its input is first converted to SMILES.

## INTRODUCTION

Computer graphics of chemical structures and formulas are most important for the interaction between chemist and computer, particularly in the fields of organic synthesis and substructure searching. What appears to be well within the scope of modern technology is surprisingly difficult, because different conventional methods of presenting structures pictorially do not always follow the same rules. Ambiguities and exceptions are often encountered. Publications on this subject[1,2] generally deal with computer-interactive graphics by electronic input with a stylus on a tablet, or with light pens on graphical CRT. Originally, at Chemical Abstracts Service, graphical data were processed manually. As computer processing of two-dimensional graphs improved, structural representations were standardized and files were established for molecular fragments as well as for complete structures.[1] In some cases it was possible to store essential information for creating a diagram separately from connection tables. But for complex structures in large databases this was impractical, so the problem of displaying structural diagrams from connection tables had to be addressed. A procedure developed by Shelley[3] involved an initial perception of structural features as a data tree. This was followed by generating atoms, ring structures, and their connecting bonds with graph-invariant codes. They were used

**Table I.** SETXY Algorithm To Compute 2D Coordinates for a Molecular Graph

```
1. PREPST—prepare structure for procedure to follow
  a. partition molecule into N trees and back edge lists
  b. form bonds between parts of disconnected structures
  c. identify bicyclo rings; break one back edge
  d. set all relative angles to 0
  e. mark all angles "not fixed"
2. CYFIX—fix relative angles of ring bonds
  a. mark all rings "not done". Go to step 2d.
  b. find "next" ring based on priority:
     maximum number of atoms in common with a "done" ring
     select largest ring among those tied
     else, pick any among the largest ties
  c. if "next" ring shares atoms with a "done" ring, find possible "from" ring(s) based on priority:
     "done" rings that share a bond with "next ring (fusion)
     "done" rings that share 1 atom with "next" ring (spiro)
     "done" rings that share more than 1 bond with "next" ring (bicyclo)
     "done" rings attached to "next" ring with an acyclic bond (e.g., biphenyl)
  ... else:
     any other new ring system (no "from" selected)
     larger bicyclo rings (no "from" selected)
  d. final selection of "next" ring, if tie from (2c), then
     select smallest ring
     if tied, select among smallest rings with most bridgeheads
  e. fix relative directions of ring bonds
     project from "from" to "next" ring center [or from (0,0) on first pass]
     compute relative directions as perfect polygons
  f. record ring
     mark relative directions "fixed"
     mark ring "done"
     mark ring "bicyclo" if approprite
  g. loop back to (2b) if any rings remain which are not "done"
3. DIRFIX—fix relative angles of "directional" bonds
4. HEADAT—select and fix head atom
  a. pick highly connected noncyclic atom as head
  b. build DFS tree from head atom as a root
  c. fix directions to root at zero
5. SPREAD—spread out unfixed atoms
  a. loop over each unfixed atom
  b. save directions and initial function evaluation
  c. select angle increments and phase based on degree
  d. minimize functions over given increments
  e. if improvement found, replace relative directions
  f. if noncyclic bond to ring atom, "tilt" the ring
  g. loop back to (5a) for next relative direction
  h. if no convergence, reduce increment and iterate
  i. if no convergence in MXITER iterations, give up
6. POSTPR—SETXY postprocessing
  a. rotate direction of root to match aspect ratio.
  b. compute X,Y coordinates from relative angles
  c. abolish bonds joining disconnected parts
  d. rejoin broken back edges
```

to minimize atom crowding and bond overlap and to orient ring systems and generate system coordinates independent of assigned sequence numbers.

The above methods of interactive graphics required file entries and computer codes. This is not the case with DEPICT. It accepts structural data from any database via connection tables which first generate the description of the structure in the linear chemical notation language SMILES. The preceding papers in this series[4,5] describe SMILES, which saves time and space in computer processing and provides the chemist easy accessibility to its linear notation of chemical structure. It is important for the machine to convert linear notation into graphical depiction because the latter is the easiest and most natural way for a chemist to identify a chemical structure. Furthermore, the user can view the graphical depiction of a structure and write down its linear notation. This is the subject of the present paper, which describes and illustrates the DEPICT program.

## DEPICT ALGORITHM

The objective of DEPICT is to convert the structure stored in the computer into a depiction that is easily recognized by the chemist. No manual input or explicit pictorial information is needed to produce a picture. All structures for which SMILES notation can be written, including connection tables that are first converted into SMILES, are represented by a drawing. The primary advantage of such an approach is that the user can benefit from the high degree of accuracy associated with graphical description while using powerful languages and nongraphical entry tools, such as SMILES and other linear computer languages. Another advantage is that structures stored in existing connection tables can be directly interpreted and depicted. No attempt is made to show a three-dimensional representation; rather, the result is a graph of the chemical structure.

Input to the DEPICT process, described in Table I, contains the same sort of information as in a connection table. Although the starting point is SMILES, various data structures are used internally for representation of a molecular graph as are appropriate to different parts of the algorithm. These representations, including edge and bond arrays, father lists, tree assignments, and relative and absolute geometry lists, are used until the final pictorial representation of the molecular structure emerges.

**Table II.** DPICTO Algorithm To Display Molecular Graph from 2D Coordinates

```
1. preliminaries
  a. find minimum, maximum, and midpoints of $X$ and $Y$
  b. scale graphics window to provide 1:1 aspect ratio
  c. computer center of aromatic rings
2. call DRWATM to draw atomic representation of each atom
  a. generate atomic symbols:
     if singly connected to the ring
        if H count > 0, start with H; full size, on base line
        if H count > 1, add H count; 0.6 size, subscript
     if isotope specified, add atomic weight; 0.6 size, superscript
     add standard atomic symbol, full size, on base line
     if not singly connected to the right
        if H count > 0, start with H; full size, on base line
        if H count > 1, add H count; 0.6 size, superscripts
     if charged
        add sign of charge; 0.6 size, superscript
        if abs (charge) is not 1, add charge value; 0.6 size, superscript
  b. determine symbol size:
     0.00 for neutral, aromatic carbon with no isotopic specification
     0.25 for ring heteroatom and unusual carbon
     0.25 for complex symbol (>1 of charge, H count, isotope)
     0.35 for all others
  c. invoke appropriate graphics color from supplied integer vector (INSET)
  d. display sized symbol centered on $X,Y$ from SETXY
3. call DRWBND to draw two-part bond between pairs of connected atoms
  a. for each end, calculate a box 1.2 times symbol size in $X$ and $Y$
  b. the following bonds will be in two colors if the end atom colors differ
  c. for all but double bonds, draw line between intersecting boxes
  d. draw parallel lines interesecting boxes for double (0.09) and triple bonds (0.12)
  e. for aromatic rings only:
     i. select color of ring atoms if all the same, else select neutral color
     ii. draw. a circle inside each aromatic ring of radius $0.3/\tan(\pi/\text{cycle length})$
```

The DEPICT algorithm has two major parts. The first, SETXY, computes the two-dimensional coordinates for a given molecular graph. These coordinates correspond to a good depiction of the structure in terms of simplicity, clarity, and aesthetics. They are converted into a graphical display of the chemical structure by the second algorithm, DPICTO.

In turn, these two major algorithms are divided into separate algorithms or subdivisions to accomplish specific objectives. They are listed in Tables I and II, which identify six individual steps for processing the molecular graph (SETXY) and three steps for converting the coordinates into the depiction of the chemical structure (DPICTO).

## LOCATION OF ATOMS AND BONDS

SETXY determines a location for each atom in the graphical representation of the chemical structure. As far as possible, all bonds have the same length and are maximized within the restrictions of the overall graphic window. Ring structures are treated as single entities for the purpose of projecting the distance from one ring center to another. After some preparatory steps (see Table I, phase 1, PREPST), the other algorithms of SETXY involve fixing the relative angles of bonds, first in cyclic structures, if rings are present (phase 2, CYFIX) and in isomeric structures (phase 3, DIRFIX). The root of the tree is then selected (phase 4, HEADAT, and the final dimensions of the graph are computed, minimizing the distance between atoms within the restrictions of the previously determined bond angles (phase 5, SPREAD). Finally, in the postprocessing stage (phase 6, POSTPR) artifacts introduced in the previous steps are eliminated and $X,Y$ coordinates are calculated from relative angles.

**Phase 1. PREPST—Preparatory Steps.** A series of preparatory operations are performed to allow more efficient processing during later phases. A fundamental axiom of modern software design is that an appropriate data structure is essential for effective implementation of an algorithm. DEPICT uses several subalgorithms in which the same chemical structure is viewed at different times as a set of lists, a graph, a forest, a tree, and graph partitionings. The preparatory steps described here set up the data structures required by subsequent algorithms.

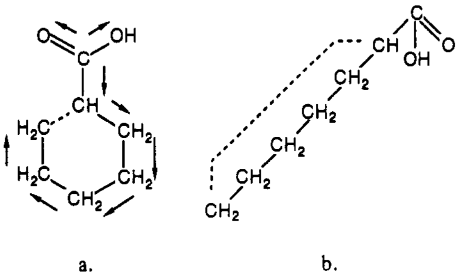

**Figure 1.** Partitioning of cyclohexanecarboxylic acid using the carboxy carbon as the root.

One of the tasks of an algorithm such as DEPICT is to avoid crowding in the display of highly connected structures. To this end, DEPICT repeatedly measures distances between 2D atom locations as "bond angles" are adjusted, which is a slow step in the algorithm. Not all interatomic distances vary as a given bond angle is changed, so one way to improve the algorithm speed is to avoid measuring of distances that do not vary at a given time. However, nothing will be gained if the method used to detect the distances which vary takes longer than computing unneeded distances. Fortunately, representing a structure as a tree data structure known as a "partition" allows this to be done efficiently.

In viewing a structure as a "tree", bonds and atoms are associated with the edges and nodes of a graph.[5] Tree partitionings provide a more appropriate data structure than graphs for much of the DEPICT algorithm. The conversion of a molecular graph into tree representation is done by the process of partitioning.[6] For a given root atom, partitioning an acyclic structure is straightforward because the tree will have no back edges. When cyclic structures are involved, the standard procedure used for describing them in linear SMILES notation is here applied as a first step in assigning coordinates to the atoms. This is illustrated in Figure 1 for cyclo-

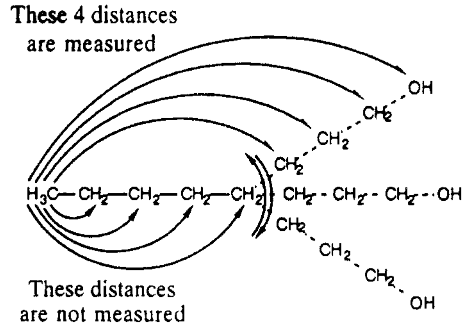
These 4 distances are measured
OH
CH2
CH2
CH2
H3C—CH2—CH2—CH2—CH2
CH2 - CH2 - CH2 - OH
CH2
CH2
CH2
These distances are not measured
OH

Figure 2. Variation of the angle about the δ carbon in *n*-octanol.

hexanecarboxylic acid [unique SMILES: OC(=O)-C1CCCCCC1]. The graph is partitioned into tree edges (bonds in the linear part of the structure) and back edges (bonds that complete a ring), shown in Figure 1b as solid and dotted lines, respectively.

The process of partitioning a tree is similar to developing a bond list for a connection table. Partitioning produces an ordered list of connected atoms and back edges which is stored for future use without affecting the initial structure representation. Each atom in the structure is considered as the tree root, and for each a different partitioning is obtained above and below its location in the chain. Thus, for a structure with $N$ atoms, there will be $N$ different tree partitionings. For example, the graph of cyclohexanecarboxylic acid has nine graph nodes (atoms), so that nine different trees are considered. One of them is shown in Figure 1.

Since bond lengths are equal and fixed initially, the graph is developed by evaluating relative angles between atoms. In step 1d all angles are set to 0°, and those which are not fixed are marked for further evaluation. Figure 2 shows an example of the variation of one angle in a simple, acyclic structure (*n*-octanol). Note that the only distances that need to be measured are those between atoms in different branches in the tree partitioning with the δ-carbon as the root.

Each angle between tree edges is evaluated. Angles are examined at incremental values (two of these are illustrated in Figure 2) with the best value being determined by the criterion of maximizing the distance between atoms. This is calculated by minimizing the function

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}^{-2}$$

where $d$ is the interatomic distance, $n$ is the number of atoms, and $i$ and $j$ are the atoms being examined. This procedure yields angles corresponding to a maximum separation of atoms. However, if it were carried out for each atom of a large molecular structure, e.g., one consisting of 100 atoms, it would require 10 000 multiplications, 5000 divisions, and 5000 additions for each separate function evaluation. The SETXY algorithm is simplified by considering only distances between atoms that move with respect to each other when the angle of interest is changed. These are the atoms on different branches of the tree (members of different partitions). As shown in Figure 2, only four distances from the methyl carbon need to be calculated, since the others are fixed.

Other preparatory operations in phase 1 involve temporarily modifying the structure to accommodate certain classes of structures. Disconnected structures are connected (step 1b) to form a single connected one which simplifies the placement of the components. For bicyclic structures, back edges are broken, leaving a nonbicyclo structure (step 1c), as will be explained in the description of the SPREAD algorithm. Temporary structural modifications are undone before depiction (phase 6).

**Phase 2. CYFIX—Establishing the Relative Angles of Ring Bonds.** In this algorithm all cyclic bond angles are fixed for the subsequent SPREAD routine. For multicyclic structures, the sequence of ring selection follows a series of priorities, specified in Table I, steps 2b–d. A root is selected in the initial ring structure; it is assigned a direction zero, and then, taking account of the back edge in each ring, relative angles are computed by the previously discussed minimization calculation. Except for bicyclo ring elements (which were previously broken in PREPST), all but one bond angle in each ring is fixed by CYFIX. The last bond angle of each ring is never considered; it usually forms a perfect polygon, although some ring systems will not allow that.

A smallest set of smallest ring is enumerated with a ring-finder routine, similar to that proposed by Balducci and Pearlman.[7] It distinguishes fused rings, spiro rings, bicyclo rings, bridgeheads, and other ring features. After the rings have been characterized, the smallest ring is chosen as a starting point for evaluating bond angles.

Table I, phase 2, describes the ring selection process in terms of "done", "next", and "from" rings. The term "done" refers to a ring that has been completely evaluated and from which the algorithm may continue to operate. Whenever possible, new ring positions are based on a projection from a done ring. The "from" ring is the ring from which this projection is made. The term "next" is to be understood in the chronological sense, and not spatially in the sense of "adjacent".

Phase 2 begins with the selection and evaluation of an initial ring (steps 2a and 2d–f). Remaining rings are iteratively selected (according to priorities listed in steps 2b–d) and evaluated (steps 2e,f). The process continues until all rings are done (step 2g).

**Phase 3. DIRFIX—Depiction of Isomeric Structures.** The DIRFIX phase applies to structures with isomeric specifications only. Its purpose is to represent ISOMERIC SMILES notation correctly. The latter, an extension of the linear chemical notation language SMILES, is the subject of a future paper in this series. For example, the SMILES notation for *trans*-dichloroethylene is Cl/C=C/Cl. DIRFIX ensures that the two chlorines will be displayed on opposites sides of the double bond and indicates that their relative position has been specified.

**Phase 4. HEADAT—Selection of Atom as the Root of the Structure.** With bond angles for ring bonds fixed in phase 2 and for isomeric structures in phase 3, the head atom of the overall structure is selected in phase 4. Part 4b refers to a depth-first search (DFS), establishing the graph of the molecular structure in the form of a tree.

**Phase 5. SPREAD—Spread Out Unfixed Atoms.** After CYFIX, there remain unfixed atoms in the noncyclic (or bicyclo) parts of the molecule, for which the SPREAD algorithm establishes coordinates.

Each angle between tree edges is evaluated. Angles are examined at incremental values (two of these are illustrated in Figure 2) with the best value being determined by the criterion of maximizing the distance between atoms. This is calculated as described for phase 1, PREPST (above), by minimizing the function

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}^{-2}$$

where $d$ is the interatomic distance, $n$ is the number of atoms, and $i$ and $j$ are the atoms being examined.

This computation does not take into account bicyclo back edges, which become important for structures such as [2.2.1]bicycloheptane (Figure 3). The algorithm used to establish the locations of ring atoms (CYFIX, phase 2) does not work well for bicyclo structures. For this reason, an algorithm

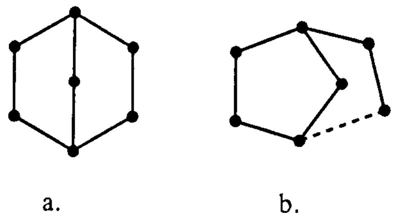

**Figure 3.** Conventional treatment of [2.2.1]bicycloheptane (a) treatment by the method of SETXY (b).

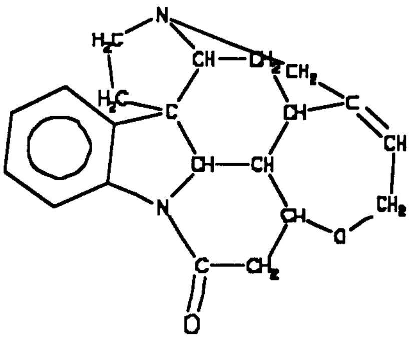

**Figure 4.** Depiction of strychnine.

used to spread out acyclic atoms (SPREAD, phase 7) was designed to handle such cases.

In Figure 3, edges marked by solid lines always have length 1.0. One 5-membered ring is displayed as a perfect pentagon, and two edges are left on a free chain. Provision has to be made only for one back edge per bicyclo bridge (drawn as a broken line). This is accomplished by adding a second term to the function to be minimized in which a large number ($k$) appears as a multiplicative factor applied to the difference between 1.0 and the actual length of the $n'$ bicyclo back edges, vs.

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}^{-2} + k\sum_{i=1}^{n'-1}\sum_{j=i+1}^{n'} (d_{ij} - 1.0)$$

The left-hand term represents all atoms not on the same branch; the right-hand term represents back edges, with a penalty for those with length >1.0.

This approach deals well with a large number of more complex bicyclo ring systems, which is not the case for template-based methods. The DEPICT output for strychnine (Figure 4) shows such an example. Strychnine has a 7-cycle ring system, including one bicyclo element. Five rings are drawn as perfect polygons with bonds of equal length. Because of the ring fusion pattern, one bond in the sixth ring (the N-CO bond) is longer than 1.0. The longest bond is the bicyclo back edge that was originally broken in PREPST. Note that in both the examples shown in Figures 3 and 4, the second term in the SPREAD function causes the display to "tighten up" rather than "spread out", as would be required if just the first term were used. An example of the SPREAD algorithm is the development of the representation of 1-*sec*-butyl-2-isopropylcyclohexane in Figure 5.

All bond angles are initially set arbitrarily to 0°. The drawings of the structure of Figure 5a,b are purely illustrative. If, for the purpose of illustration, all relative angles were set to 0°, the drawing would be one straight line. Instead, the angles are shown in an unresolved state as indicated by dashed lines in Figure 5a. A root is assigned (circled node), it is assigned a direction of zero, and the DFS order of the tree is generated (Figure 5b). CYFIX then computes and fixes the ring angles and bond angles off the ring (Figure 5c). All bonds have length 1.0 in this example. Bond angles are given their initial values when the root is selected. The missing ring bond is the back edge of the ring; note that it does not enter into any further computation. The arcs show the angles that are fixed and will not be changed in the following SPREAD algorithm. Parts d and e of Figure 5 show the fitting process continuing for different, previously unfixed angles in SPREAD until the final depiction, Figure 5f, is reached. SPREAD does not alter the root direction; it is rotated later to match the aspect ratio of the final output window (in phase 6a of SETXY).

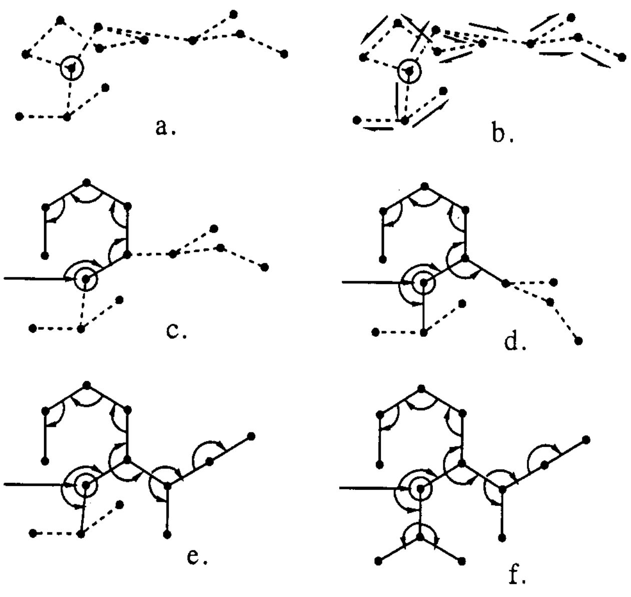

**Figure 5.** Illustration of the SPREAD algorithm. The root node is circled; dotted lines are unfixed bonds; once fixed, relative angles are indicated with arcs. (a) represents the initial state with root identified; arrows indicate the DFS order in (b); SETXY results in (c); bonds on rings fixed in (d); relative angles of acyclic bonds in (e); final result after SPREAD is in (f).

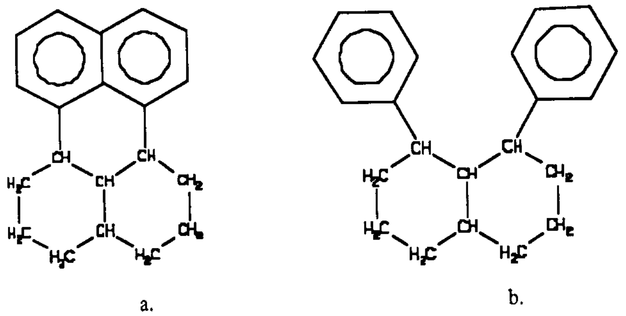

**Figure 6.** Effect of tilting fixed bond angles.

For joined rings, both spiro and fused rings, step 5f provides one more adjustment. When a nonring bond is joined to a ring atom, it may have to be "tilted" to avoid a clumsy representation. For example, in the case of diphenyldecalin (Figure 6), two bond angles are tilted, although they had previously been fixed. The tilt operation of step 5f spreads out the two (non-fused) phenyl groups as shown in Figure 6a,b.

The SETXY function is minimized by iteratively changing the relative angles between each pair of bonds that are connected at an atom. If the function does not meet the specified convergence criterion, the angle increments are reduced, e.g., from 120° to 90°, and the process is repeated (phase 5h). Each iteration is more complex because the number of angles through which the structure is swept increases. It is only rarely that convergence (i.e., obtaining an acceptable value for the minimization function) does not occur when the stipulated maximum number of iterations (MXITER) has taken place (phase 5i). This happens when two atoms are forced on top of each other so the function value is always huge (i.e., $d^2 = 0$). An example of this is the spiral structure of hexahelicene.

**Phase 6. POSTPR-SETXY—Postprocessing.** The principle task of this phase is the calculation of the atoms' $(X,Y)$ co-

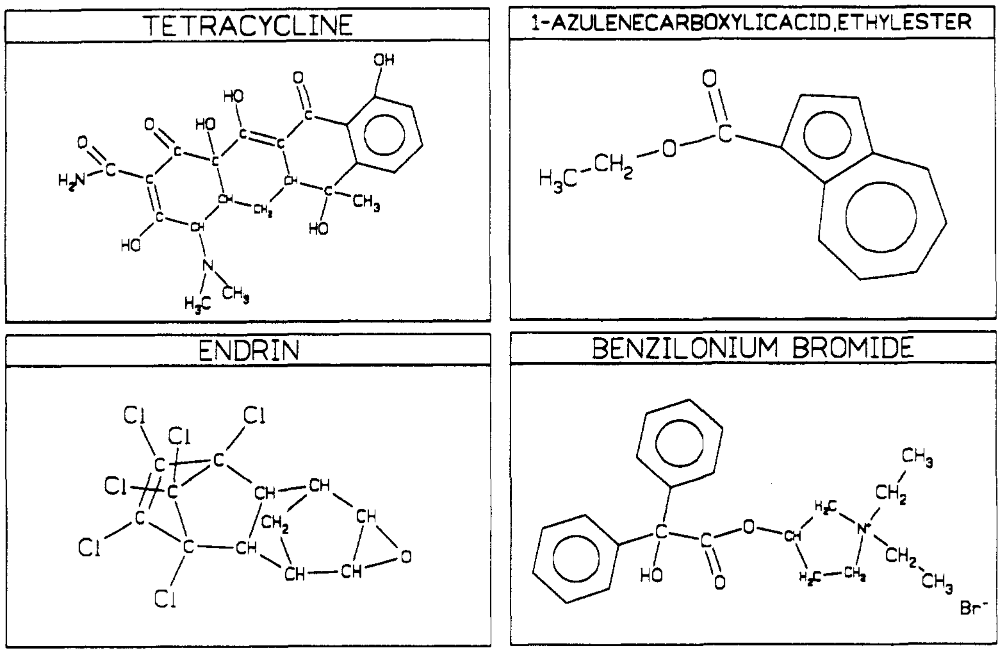

**Figure 7.** Examples of DEPICT output.

ordinates from the relative bond angles (phase 6b). Additionally, phase 6a provides the desired orientation for a given window's aspect ratio. Finally, the temporary bonds that were introduced to handle disconnected structures are removed (phase 6c), and the broken bicyclo edges of rings are rejoined (phase 6d).

## DISPLAY OF MOLECULAR GRAPH

In the second part of DEPICT, the DPICTO algorithm converts the coordinates of the atoms into a graphical display. It consists of three phases (see Table II): preliminaries, drawing the atomic symbols, and drawing bonds.

**Phase 1. Preliminaries.** In this phase the coordinates of the atoms are examined and midpoints of the figure are established (phase 1a) to scale the graphic window with a uniform $X$:$Y$ aspect ratio (phase 1b). For aromatic structures, the centers of the rings are computed for later representation of aromatic rings (phase 1c).

**Phase 2. DRWATM—Atomic Representation of Each Atom.** Atomic labels consist of atomic symbols, hydrogen count, charge, and isotopic mass, if needed. If there is a connection to one or more hydrogen atoms, the hydrogen count specifies the relative location of the symbol H with respect to the atomic symbol. The symbol has a standard, full size against which subscripts, superscripts, and charges are suitably scaled. As an aid to visual recognition of complex structures, atomic symbols in cyclic structures are drawn smaller (80%) than aliphatic, nonring atoms. Aromatic carbons have 0% size, i.e., they are not shown but are presented by the aromatic ring symbol within the ring (phase 2b).

**Phase 3. DRWBND—Representation of Bonds Connecting Atoms.** This phase completes the depiction of the chemical structure. A rectangular box, hidden in the final presentation, is placed around each atomic symbol (phase 3a) which is located at the assigned $X$,$Y$ coordinates. Bonds are then established by drawing one, two, or three lines between centers of adjacent boxes (phases 3b, 3c, and 3d). The representation is completed by drawing a circle inside each aromatic ring (phase 3e).

## DISCUSSION

The DEPICT program consists of the algorithms listed in Tables I and II, which are explained in the text of this paper. It is distinguished by the fact that it requires no manual or explicit structural input to produce a two-dimensional output display of a molecular structure. DEPICT does not require, but is facilitated by, the use of SMILES, a linear notation language that is both user- and machine-friendly. The algorithms can be followed independently of SMILES by any other input method. However, because of the efficiency of SMILES, it will be advantageous in most applications first to convert the input, from whatever language it is initially present, to SMILES.

Initially, the input is converted to a unique SMILES notation,[5] which is internally converted to a minimum spanning tree containing all structural fragments while it simultaneously initializes a number of lists: bond distances, angles between bonds, back edges of rings, temporary bonds between separate components, and others. At the start the relative angles are set at 0°, and bond distances are unity. The angles are then adjusted by maximizing the distances between atoms. Further operations involve the identification and treatment of different types of ring structures. Finally, a simple two-dimensional depiction of a molecular structure, easily understood by chemists, is presented.

DEPICT draws circles inside rings which are designated as aromatic in the computer representation of the structure. The SMILES language establishes aromaticity on the basis of delocalization of $4N + 2$ $\pi$ electrons.[4] This may include other than 6-membered rings, e.g., thiophene, furan, 1*H*-pyrrole, the cyclopentadienyl anion, and azulene (see Figure 7). Other conventions result in different definitions of aromaticity, but the circles drawn by DEPICT always indicate which rings, if any, are assigned the "aromatic" property.

For multicyclic structures DEPICT attempts to retain polygonal representation of as many rings as possible. While it stretches out one bond per multicycle set, it allows atoms in the "loose" ends to reorient themselves away from others in the structure. It favors short over long bond stretching, but will stretch a bond in a multicyclic structure as far as needed. This is shown for strychnine (Figure 4).

Multicyclic structures often require bond crossings for clarity, but since the DEPICT algorithm spreads out atoms, bond crossings are relatively rare, although they are not prohibited. An example of bond crossing is shown in the drawing for endrin (Figure 7), which also illustrates the use of smaller symbols for ring atoms.

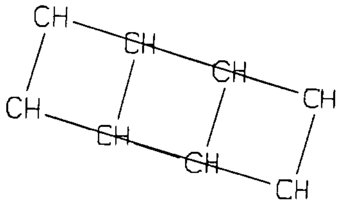

**Figure 8.** Depiction of cubane, illustrating poor coordinate selection for a condensed multicyclic system.

Yet, there are a few types of structures for which DEPICT does not provide adequate graphics. This is the case for condensed multicyclic systems which cannot be dealt with easily without drawing curved lines. An example of this pitfall is cubane (Figure 8). The cubane structure is drawn correctly but the attempt to make "as many perfect polygons as possible" forces two bonds to fall directly on top of others, leading to a confusing picture. A general solution to such problems is currently under investigation.

## SUMMARY

DEPICT is a computer algorithm that produces a graphical display of any chemical structure for which the linear notation language SMILES can be generated. Atomic coordinates are computed by evaluating angles between atoms while treating the structure as a tree with fixed length edges. The graphic representation is then derived from the coordinates.

The most important aspects of DEPICT are illustrated in the four examples in Figure 7. These include chain positioning, ring system representation, aromaticity indication, and display of atomic properties such as charge.

No explicit graphical information is required as input to DEPICT. Since most connection table formats and other linear notations can be converted to SMILES, DEPICT can be used to display structures stored in such formats. Furthermore, DEPICT is ideal for display of structures which are not stored in a database, e.g., novel structures generated by a computer.

## ACKNOWLEDGMENT

I thank Steven Burns and Arthur Weininger for assistance in programming and Joseph L. Weininger for editorial assistance.

## REFERENCES AND NOTES

(1) Goodson, A. L. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 212.
(2) Kalbfleisch, W.; Ohnacker, G. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 176.
(3) Shelley, C. A. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 61.
(4) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
(5) Weininger, D.; Weininger, A.; Weininger, J. L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97.
(6) Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA, 1972; Chapter 6 (Partitions).
(7) Balducci, R.; Pearlman, R. F. Novel Algorithms for the Rapid Perception of the Unique, Optimal Set of Rings. Unpublished results.

# Computerized Retrieval of Information on Biosynthesis and Metabolic Pathways[1]

SANDOR BARCZA,* LAWRENCE A. KELLY, and CHRISTOPHER D. LENZ
Sandoz Research Institute, East Hanover, New Jersey 07936

Received February 22, 1990

Biosynthetic metabolic pathways were analyzed, and a hierarchy of attributes was constructed. Representation of the knowledge base on metabolic conversions was effected in terms of this hierarchy of attributes and the chemical structures of molecules participating in metabolic conversion steps. A prototype database was constructed with the MACCS and DATACCS programs, already in use for storage, searching, and reporting of chemical structures, chemical-biological data, and chemical reactions at Sandoz.[2,3] Key data in the new metabolic conversions database are the enzyme name and classification, effectors, inhibitors, literature reference, etc. Participating molecules, if known and under 255 heavy atoms, are stored and diagrammed as stereostructures. The crucial data on metabolic conversions are represented by "From" and "To" datatypes. All the data are exact match and range searchable for text and numbers. Thus, precursors and progenitors of compounds can be found. Structures are match and substructure searchable. This tool is a useful and very flexible complement to metabolic charts. It in itself can be used to report and graph conversion steps and sequences.

## INTRODUCTION

*Living organisms* are probably the most complex entities of the universe. Man attempts to describe, document, and understand organisms for several reasons: academic knowledge for its own sake; understanding, so that the organism can be influenced, controlled, repaired—e.g., hybridization of corn and curing of diseases; and to transfer the ingenuity found in nature to products of man, e.g., preceptrons, sensory systems, robotics, *biomimetic* organic *syntheses*, etc.

The *description* of organisms occurs at several levels and with different armamentaria: ecosystems, anatomy, physiology, biochemistry, biophysics, quantum biology, etc. An outstandingly important level of description of constituents and processes of living organisms is that of the chemical *transformation of molecules in the body*. These interconversions make up the *metabolic pathways*, *biosynthetic pathways*, replication of genetic material, etc. The main pathways have been classified into broad (and in some cases fuzzy) sets of catabolism—involving degradation and energy release—and anabolism—involving energy enrichment and construction.

The pathways of biosynthesis and metabolism form a highly complex information system. In order to make this information computer storable and retrievable, the issues of *representation* of knowledge must be addressed. The description should be readily understandable or at least serve some utilitarian purposes. It should be preservable on paper and readable, transformable, writable, and usable by computer. As a practical pedagogic matter, it should be easily taught, which requires graphical representations as well.

**The System.** A knowledge base of a biological system is typically too complex for easy comprehension and memory. Further, application of the knowledge often becomes quite