A Versatile, Efficient, and Interactive Program To Build Molecular Structures for Theoretical Calculations and Chemical Information Systems

JAMES KAO,* CHARLES EYERMANN, LORAINE WATT, ROBERT MAHER, and DIANE LEISTER
Research Center, Philip Morris, U.S.A., Richmond, Virginia 23261

Received December 3, 1984

A generalized device-independent computer program, MOLBUL, is described that allows the user to interactively build 3-D or 2-D structures with different display options by using various graphics devices. The 3-D structure can be used for energy minimization, and the 2-D structure is for the common on-line substructure search of our chemical information system. MOLBUL is designed to be very user friendly through the implementation of English-like commands and templates. While MOLBUL has its own standard format, it can also read and save files containing Cartesian or internal coordinates in various formats specific to theoretical tools (molecular orbital, molecular mechanics, or molecular descriptor calculations). It has various powerful transformation capabilities (translation, rotation, and software zooming). It is also equipped with a fragment database that allows the user to select several fragments at a time to facilitate building structures. The Chemical Draftsman module of MOLBUL is appropriate for graphic art applications such as adding text and special symbols to a display for publications and presentations. Many more useful functionalities of MOLBUL are demonstrated.

INTRODUCTION

The advance of computer technology in the past decades has had a revolutionary impact on almost every field of business and science. The manpower expense has increased gradually while the computer-hardware costs have declined sharply. Similar trends are expected to continue for the future. Computerized automation will be a major objective of every organization to increase production efficiency while cutting down the incurred costs. It is now highly justified to encourage chemists to computerize their information processing and to perform molecular modeling. After considering our system environment, we have decided to develop an in-house molecular information and modeling system.[1] Ways to represent molecular structures are the heart of any molecular information and modeling system. Chemists are used to graphics representations of chemical structures, and in fact, they draw them for their daily work with pencils, with stencils and transfer symbols, or, more recently, with computer graphics techniques. Three-dimensional (3-D) structures are important to better understand molecular properties. Chemists therefore use perspective views, wedges, etc. to represent 3-D features in their daily discussions and reports due to the limitation of the common communication media, paper.

A similar 2-D approach for representing structures has been adopted in computer information processing. The practical reason is probably that experimental 3-D structures are not always known. However, it is also possibly due to the lack of adequate techniques to effectively represent and store 3-D structures at comparable computer costs. The common techniques used to represent 2-D structural (canonical) information are Wiswesser line notation (WLN) and connection tables.[2,3] Most of the chemical databases that are commercially available are based on one of these techniques.[2,3] Recently, there has been much progress in utilizing computer graphics to build 2-D structures.

The demand for 3-D molecular structures has become much stronger due to the advance of theoretical chemistry and the decrease of computer-hardware costs. A 3-D structure is the required input for all theoretical calculations such as molecular mechanics and quantum mechanics to derive optimized geometries.[4] It has been proved that molecular structures and properties obtained from sophisticated theoretical methods are as good as experimental ones.[4,5] The 3-D molecular structures are necessary for detailed molecular modeling applications.

One way to obtain initial 3-D structures is to convert 2-D structures, which may be either retrieved from databases or constructed from a 2-D structure computer program. However, it is unfortunate that the conversion from 2-D to 3-D does not always work; in particular, it poses many problems in studying large and nonplanar structures. On the other hand, the conversion from 3-D to 2-D always works. For this reason, there have been efforts to develop programs that will make it easier to build 3-D structures. However, these packages have one or more of the following deficiencies: expensive hardware and software, unfriendly, limited in scope, company proprietary, and machine dependent and/or not integrated. A low-cost and friendly system to manipulate and build 3-D structures is then necessary in molecular information and modeling.

In this paper,[6] we describe MOLBUL (*Mol*ecular *Bul*der), the front end of our system. MOLBUL is a versatile, efficient, integrated, device-independent, and interactive program for chemists to effectively build 3-D or 2-D molecular structures for *graphics applications*, *theoretical calculations*, and *chemical information* systems. To the best of our knowledge, no such *unique*, *integrated*, and *friendly* software is currently available elsewhere in the public domain.[7]

METHODOLOGY

The general program design specifications of MOLBUL are as follows: (a) it is to be executed (with no or minimal modifications) on different machines; (b) it allows the user to promptly build the desired *2-D* or *3-D* structures from any graphics device (including *low-price* graphics terminals); (c) it is equally convenient for use by either nonexperienced or experienced users. In other words, "program portability and maintenance", "on-line interactive", and "user friendly" were major objectives in program design. We have made the following decisions to meet the objectives as closely as possible.

The program was written entirely in ANSI 77 FORTRAN since FORTRAN is probably the most popular language in scientific communities and almost every computer for scientific applications can do the compilation. The textual information was mainly handled with the A1 format to increase machine independence. Since the program was expected to be small, program portability was of more concern to us than the memory usage.

The FORTRAN program has to communicate to a graphics device to display chemical structures. However, there are no common standards for communications between a FORTRAN program and a graphics device.[8] Thus, no fully machine-independent graphics program currently exists, due to the lack of common computer graphics standards. Fortunately, there

0095-2338/85/1625-0400$01.50/0

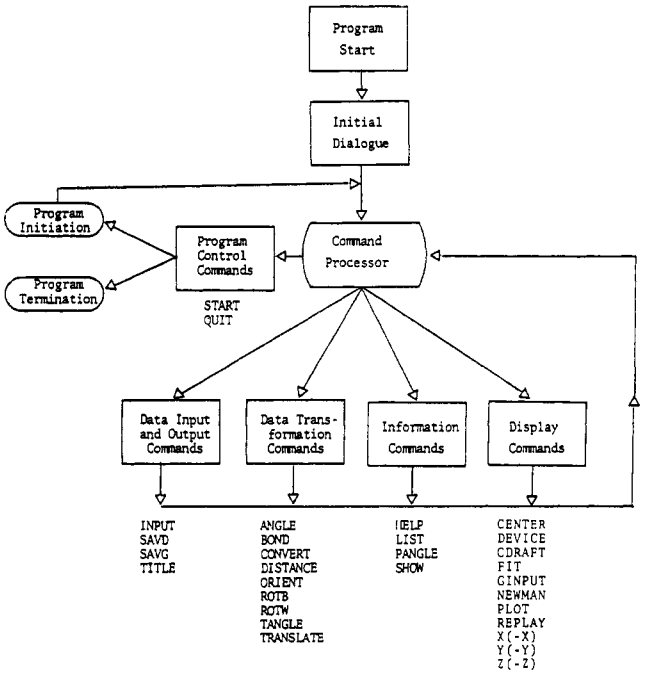


Figure 1. Program flow diagram.

are a few interactive graphic packages available on the market that come with different device drivers that can relieve the user's program of device-dependent features. We have chosen to use the PLOT-10 IGL of Tektronix, since it is one of the widely accepted packages and is currently available on our machine.

To handle two different types of users, minimal and extensive users, the program dialogue with the user has been written as friendly and informative as possible for the former and as concise as possible for the latter. The branch-point technique is used to facilitate this specification. The program is "friendly" interactive since the user commands are plain English words and are quite easy to remember. The program also has a HELP file to provide information about user commands. Information is entered as free format, and the program has certain validating capabilities. Input entered from the keyboard may be upper- or lower-case letters. For instances where lower-case letters must be retained for plotting (such as chemical symbols), no conversion will be carried out.

The program is modularly structured and documented to increase its portability and maintainability. The basic flow diagram of the program is displayed in Figure 1. After the start of the program, the user will be greeted with the initial dialogue message. The user at this point can receive a detailed message about the program. The program then processes user commands until the QUIT command is given. This command terminates the program.

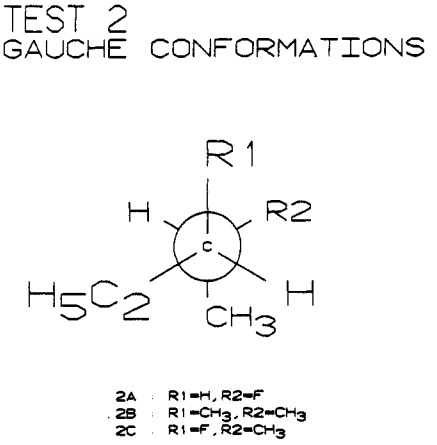
The program maintains two types of structures, namely, the primary structure and fragments. All structures are associated with appropriate connection tables.[9] More detailed definitions of both kinds of structures are given later. The user commands have been divided into five categories according to their functions. These categories are shown in Figure 1 together with the specific commands for each category. We shall briefly describe each category here while discussing the commands in detail under User Commands.

The PROGRAM CONTROL commands allow the user to end the program or restart the program. The DATA INPUT/OUTPUT commands allow the user to input data easily and save structures once they have been built or displayed.

The DATA TRANSFORMATION commands manipulate data for various purposes. These commands allow the user to rotate the primary structure around any axis, rotate part of the primary structure around a specific bond, translate the primary structure along any axis, orient the primary structure to a specific position in space, and modify or calculate bond types, bond lengths, bond angles, or torsional angles of the primary structure. A 2-D structure can also be converted to a 3-D structure with the CONVERT command. Many of the DATA TRANSFORMATION commands are available for the fragments.

The INFORMATION commands provide help information and report structural data about the current primary structure in the program. The structural data provided include the connection table, bond types, bond lengths, bond angles, dihedral angles, and distance matrix. The angle between two planes, where a plane is defined by three unique atoms in the primary structure, can be calculated with the PANGLE command.

The DISPLAY commands control the graphics displays. The structures may be viewed along any of the axes from either the positive or negative direction. The NEWMAN command allows user to draw quality Newman projections. Newman projections, first advocated by Newman in 1952,[10] have become the uniformly accepted and practiced mode of three-dimensional structural information by organic chemists. Indeed, one can hardly find any standard organic chemical text book or current organic chemistry journal without Newman projections. Newman projections are useful in molecular modeling to build molecular structures with chiral centers.[11] A typical Newman projection produced by the NEWMAN module is



The program has two command levels. The first or top level, which has just been described, permits the user to enter the user command key words that direct the program flow. The second level is accessed by the PLOT, GINPUT (*Graphic Input*), FIT, and CDRAFT commands. The PLOT mode displays the primary structure. Many different types of display options are available. Some of them include stick, ball and stick, space filled, dot cloud, color or no color, etc. (see Example 1 and Example 2). These options are fully controlled by the user. The PLOT command uses a menu to direct these options.

The GINPUT mode is the heart of the program where the user can *graphically* build *2-D* and *3-D* molecular structures. The building of a structure is done with the use of templates (see Figure 2a,b). The main template that appears when the user specifies the GINPUT mode uses conventional means for building a molecular structure. Template commands that allow the user to add atoms, replace atoms with another atom or group, change bond types, add or delete bonds, etc. are included. The SAVE DATA template command stores the primary structure and fragments into a file. Once a template command has been picked, it is in effect until another template command is chosen. This makes the use of the template easier and faster since the user, for example, need only select the

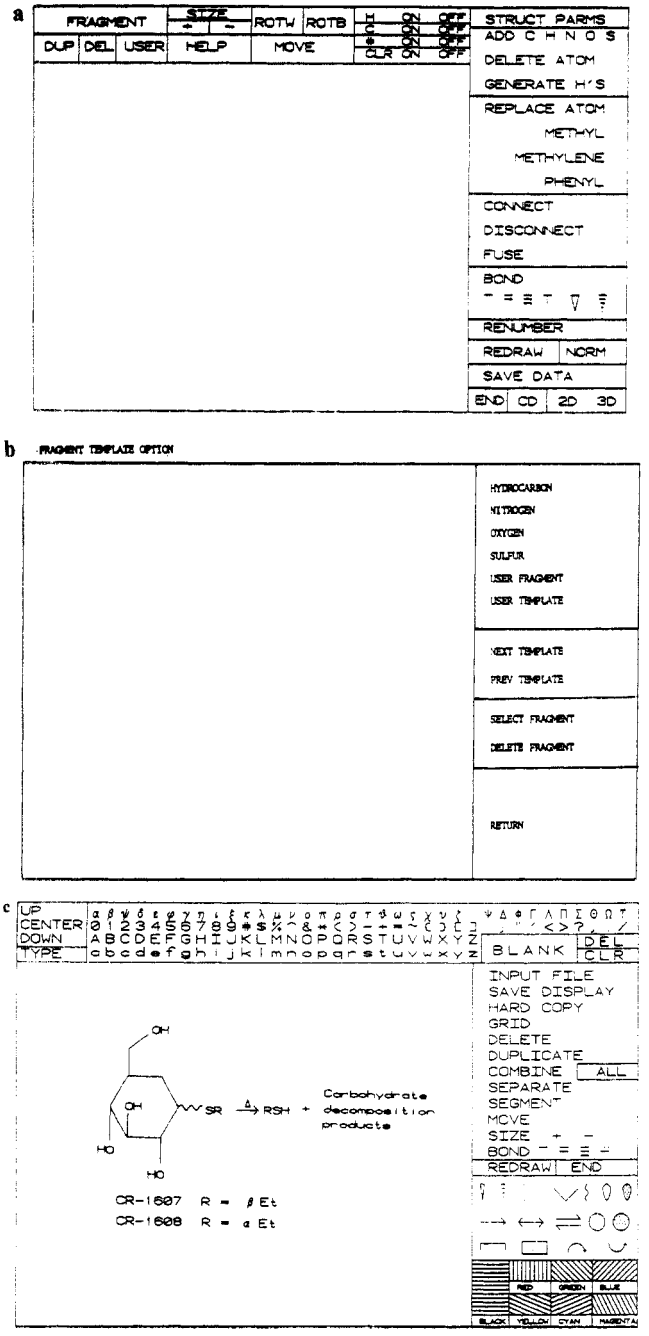


Figure 2. (a) Main template; (b) fragment template; (c) draftsman template.

ADD C command once to add several carbons onto the primary structure.

The program, however, has another means of building structures that is unique. This is through the use of fragments. The FRAGMENT template command erases the main template and provides the user with a fragment template. From this template the user may pick fragments from the many stored fragments already in the program. User-defined fragments that are specific to the user's applications can also be brought into the working area from the main template or fragment template (Figure 2). A fragment can then be connected or fused to the primary structure to build the desired molecular structure.

The program makes a distinction between the primary structure and fragments. Certain template commands can only be applied to the primary structure while others are more general and can also be applied to fragments. There can be only one primary structure, but there may be as many as four fragments. The difference between the primary structure and a fragment depends on how the structure is entered into the program. The primary structure can be entered by two methods. The first method is with the use of the INPUT user command. The data entered through this command automatically becomes the primary structure. If there was a previous primary structure, it is replaced but none of the fragments are affected. The second method is through the use of the GINPUT mode. If there is no current primary structure when the user specifies the GINPUT mode, the first data entered through the use of a template command will become the primary structure. Fragments can only be entered through the GINPUT mode, as described above.

The GINPUT mode will always display the current primary structure and the current fragments. The PLOT mode only displays the primary structure. The program still maintains the current fragments when the program goes from the second level of the GINPUT mode to the top level of the user command mode. The next time the program goes back into the GINPUT mode the current fragments will be plotted.

Many of the commands at the top level are also available on the main template in the GINPUT mode. However, there is a marked distinction between commands on the template and commands at the top level. The commands at the top level are called USER commands and are activated when the user types in (via the keyboard) the letters that comprise a specific command key word followed by a carriage return. The TEMPLATE commands, however, are selected with the use of the graphic input device (e.g., cross hairs, light pen, mouse, etc.). The execution of a particular template command may require the user to enter additional information either through the keyboard or through the graphics input device. The template commands are described in detail in the next section.

The FIT mode performs a least-squares fit between two molecular structures.[12] The graphics result is the display of the two structures overlapped as specified by the least-squares calculations. A menu is provided for the user to guide the comparison. The fit module is very useful for structure–activity correlations.

The CDRAFT (Chemical *Draft*sman) is a user friendly and versatile module[13] that enables chemists to produce various types of displays (such as structures, equations, and reactions) for publications or presentations. The features of CDRAFT have made MOLBUL more complete and more appealing to chemists. Structural graphs can easily be included in a display by using the CDRAFT template (Figure 2c) to build the graph or by entering a file previously saved in one of the formats specific to theoretical tools (molecular orbital, molecular mechanics, or molecular descriptor calculations). Structures can be displayed in a 3-D perspective view or can be displayed showing bond types (single, double, triple, etc.). Many special symbols (such as arrows, wedges, orbital representations, and brackets) that are frequently used in chemical displays are available on the template. Text strings can include upper- and lower-case Roman and Greek characters, punctuation symbols, subscripts, and superscripts.

The DATA TRANSFORMATION process will be further elaborated below. Two major tasks have to be carried out within this process: coordinate transformation and picture projection. The heart of coordinate transformation is to properly derive direction cosines. The basic principle of picture projection is illustrated in Figure 3. In this figure, the user views through the x axis. Any point, $P(X, Y, Z)$ in the Cartesian coordinate system will be projected into the projection plane, $P'(X', Y')$. X' and Y' are proportional to PPL2I (projection plane to eye) and the reciprocal of the difference between ORI2I (origin to eye) and Z. X' and Y' are of course also proportional to X and Y, respectively. The coordinates on the projection plane system have to be appropriately

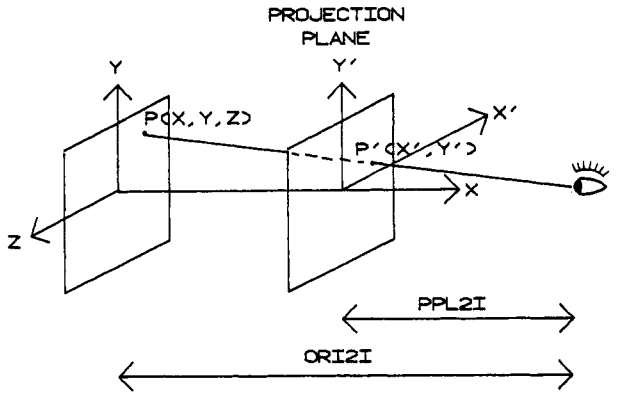


Figure 3. Basic principle of perspective projection.

mapped into the viewport size of the display surface. In our applications, circles are drawn around the point or symbols are written on the point to represent atoms, while bars are shown between atoms to represent chemical bonds.

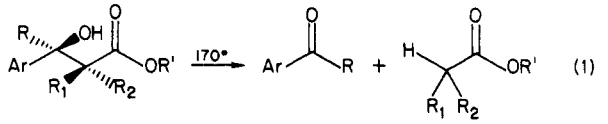
IMPORTANCE OF MOLBUL IN TODAY'S R&D ENVIRONMENT

Due to the advance of computer technology, computerized automation is an objective of major organizations to increase productivity while cutting down the manpower expense. However, the rapid advance of both hardware and software technology has caused certain grievous pain for applications since there are no real standards among operating systems, device drivers, etc. For example, we switched from Xerox's Sigma system to DEC-2060 systems about 3 years ago, and we are acquiring VAX, GOULD, MASSCOMP, and IBM computers for various applications. Conversions from one machine to another would be extremely undesirable if the software is highly device-dependent. Unfortunately, most software packages we have seen are hooked to specific hardware. MOLBUL has been designed and developed by keeping device dependency (both host computer and terminal devices) as its prime goal. Thus, it is fully portable from one machine to another and is ready to grow with the rapid advance of computer technology.

It is our opinion that MOLBUL is the *most unique* system that integrates applications of graphic art (module Chemical Draftsman), 2-D graphs, and 3-D structures. Traditionally, molecular information (2-D intensive) and molecular modeling (3-D intensive) are handled in different groups of a company. MOLBUL is probably the first system trying to *effectively* bring molecular information and molecular modeling together. Apparently, integration of both will increase the usage of a company's resource and increase the productivity of research personnel in R&D efforts.[1] A typical example would be as follows. Chemist A has spent quite an effort to construct a 3-D structure (probably with energy minimization) for a complex molecule. That structure should be stored and later retrieved in the R&D centralized STRUCTURE database. It would be an unnecessary waste for Chemist B to start from ground zero if he is interested in the structure of that molecule or structures of similar compounds. Within MOLBUL, a connection table is maintained for both 2-D and 3-D structures. A standard MOLBUL format that contains this connection table and structure can be saved for an information system.

We started to develop the MOLBUL system of programs about 3 years ago, after we realized that there were no such programs available for us. Today, the Phase I goal of MOLBUL has been completed, and it is the heart of our in-house Molecular Information and Modeling System (MIMS). We should stress that MOLBUL is very friendly and versatile and every function is developed for practical applications. It was our approach that as long as we developed a new functionality for MOLBUL we put it into an application for testing. We shall mention a few state of the art MOLBUL applications in the following.

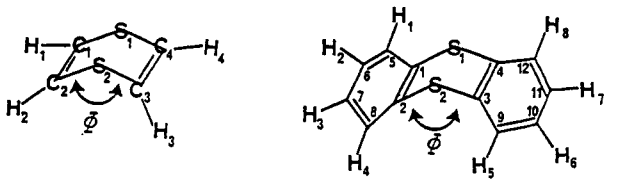
It has been our interest in correlating kinetic data (reaction rate constants) with chemical structures. As a typical example, without the existence of MOLBUL, we would not be able to finish the study "Theoretical Modelling of Pyrolysis Reactions. Thermal Retroaldol Reactions of β-Hydroxyesters"[14] in a very short period (eq 1). The TANGLE command has been found



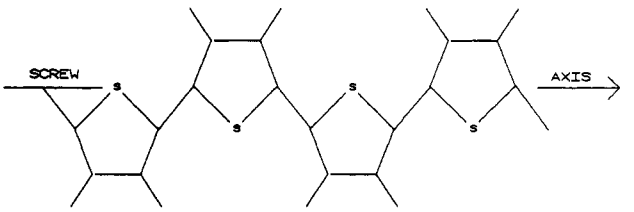
(1)

extremely useful since it allows chemists to effectively build a rotamer and to perform conformational analysis.

The PANGLE command has been found useful in comparing theoretical structures to the reported experimental butterfly angle (ϕ) of dithiin and its derivatives:



Without the existence of MOLBUL, we would not be able to derive as many structures (~100) for testing our new technique, "The Molecular-Orbital-Based Molecular Mechanics".[5] MOLBUL has been found extremely powerful in our polymer work in locating *screw axis* and *cell unit*, which are essential for calculating band gap, band structure, etc:



The NEWMAN and the Chemical Draftsman modules have been successfully used in producing graphs for reports, presentations and publications.[10,11]

In our continued effort in quantitative structure–activity correlation (QSAR), MOLBUL provides either the basic and essential functions (such as dot-cloud display and structure comparison) or the information to other programs (such as MOLPRO for calculating connectivity, surface and volume, as well as MIMS for database applications).[1] The importance of MOLBUL (or similar packages) in today's R&D environment cannot be overlooked, and its future enhancements should be encouraged.

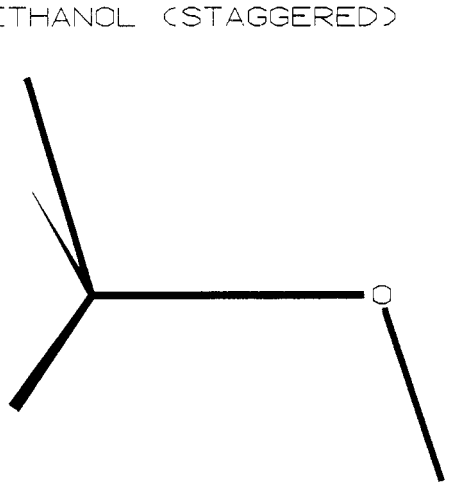
USER COMMANDS

The richness of MOLBUL cannot be fully appreciated without briefly describing the user commands. There are currently 27 commands available for users to carry out specific tasks. After a user command is entered and the carriage return is hit, the system may prompt for additional information depending on the nature of the command. User commands may be abbreviated as long as the abbreviation can uniquely identify a command. The system responds to any illegal command by printing the following message:

UNRECOGNIZED COMMAND, TRY AGAIN.

This allows the user to retype a command. After completing a user command, the program will reach another branch point,

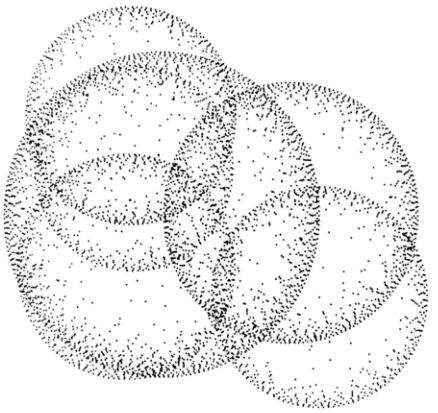
```
@MOLBUL

*****************************************
*                    VERSION 1          *
*                                       *
* WELCOME TO THE MOLBUL COMPUTER PROGRAM *
* PHILIP MORRIS RESEARCH CENTER         *
* RICHMOND, VIRGINIA 23261              *
*                                       *
*****************************************

NEED INFORMATION ON HOW TO USE THIS PROGRAM (Y/N)?
n

USER COMMANDS AND THEIR SHORTEST ALLOWABLE FORM

ANGLE       A          LIST       L          SAVD       SD
BOND        B          NEWMAN     N          SHOW       SH
CONVERT     CO         ORIENT     O          START      ST
CDRAFT      CD         PANGLE     PA         TANGLE     TA
DEVICE      DE         PLOT       PL         TITLE      TI
DISTANCE    DI         QUIT       Q          TRANSLATE  TR
FIT         F          REPLAY     RE         X(-X)      X(-X)
GINPUT      G          ROTB       RB         Y(-Y)      Y(-Y)
HELP        H          ROTW       RW         Z(-Z)      Z(-Z)
INPUT       I          SAVD       SD         ?          ?

ENTER USER COMMAND:
help

USER COMMANDS:
 A   ANGLE     ALLOWS THE USER TO CALCULATE OR MODIFY THE
               ANGLE BETWEEN ANY THREE ATOMS.
 B   BOND      ALLOWS THE USER TO SPECIFY WHICH BONDS ARE TO
               BE DRAWN AND THEIR RESPECTIVE BOND TYPES.  THE
               DEFAULT IS TO AUTOMATICALLY FIND ALL BONDS BASED
               ON INTERATOMIC DISTANCES AND COVALENT RADII. ONCE
               SPECIFIED THEY REMAIN THE SAME UNTIL CHANGED BY
               THE USER.
CO   CONVERT   CONVERTS 2D STRUCTURES TO 3D STRUCTURES.
CD   CDRAFT    ENTERS THE CHEMICAL DRAFTSMAN MODULE.
DE   DEVICE    SETS THE DESIRABLE DEVICE AND OPTION.
               THE DEFAULT IS DEVICE=4027 AND OPTION=1.
DI   DIST      ALLOWS THE USER TO CALCULATE OR MODIFY THE
               INTERNUCLEAR DISTANCE BETWEEN ANY TWO ATOMS.
 F   FIT       ENTERS THE LEAST SQUARES FIT MODULE.
 G   GINPUT    ENTERS THE GRAPHIC INPUT MODULE.
 H   HELP      PRINTS THE DESCRIPTION OF ALL USER COMMANDS.
 I   INPUT     ENTERS THE FILE INPUT MODULE.
 L   LIST      DISPLAYS THE CURRENT DATA FILE.
 N   NEWMAN    DRAWS NEWMAN PROJECTIONS.
 O   ORIENT    ALLOWS THE USER TO PLACE THE MOLECULE IN A
               DESIRED ORIENTATION WITH RESPECT TO THE X, Y,
               AND Z AXIS.
PA   PANGLE    CALCULATES THE ANGLE BETWEEN TWO PLANES.
               EACH PLANE IS DEFINED BY THREE UNIQUE ATOMS.
PL   PLOT      INSTRUCTS PROGRAM TO MAKE A PLOT.
 Q   QUIT      TERMINATES THE PROGRAM EXECUTION.
RE   REPLAY    REQUESTS THE PROGRAM TO PLOT USING A PREVIOUS
               SAVED FILE.
RB   ROTB      PERMITS THE USER TO ROTATE SOME (OR ALL) OF
               THE SPECIFIED ATOMS BY AN ANGLE, THROUGH AN
               AXIS DEFINED BY TWO ATOMS.
RW   ROTW      ROTATES THE WHOLE MOLECULE FOR A SPECIFIED
               ANGLE AROUND AN AXIS.
SD   SAVD      REQUESTS THE PROGRAM TO STORE THE CURRENT
               DATA FOR FUTURE USAGE.
SG   SAVG      PERMITS THE USER TO STORE THE GRAPHIC DATA
               AND ENVIRONMENT FOR FUTURE DISPLAY.
SH   SHOW      OUTPUTS ON THE TERMINAL OR LINE PRINTER A
               SUMMARY OF BONDED ATOMS, BOND TYPES, BOND
               LENGTHS, BOND ANGLES, DIHEDRAL ANGLES, AND A
               DISTANCE MATRIX.
ST   START     INSTRUCTS THE PROGRAM TO START FROM THE
               BEGINNING WITHOUT STOPPING THE EXECUTION.
TA   TANGLE    ALLOWS THE USER TO CALCULATE OR MODIFY THE
               TORSION ANGLE BETWEEN ANY FOUR ATOMS.
TI   TITLE     REQUESTS TO CHANGE THE DATA TITLE.
TR   TRANSLATE TRANSLATES MOLECULE ALONG X, Y, OR Z AXIS.
     X (-X)    VIEWS THROUGH X (-X) AXIS.
     Y (-Y)    VIEWS THROUGH Y (-Y) AXIS.
     Z (-Z)    VIEWS THROUGH Z (-Z) AXIS.
 ?   ?         LISTS THE USER COMMANDS AND THEIR SHORTEST ALLOWABLE FORM.


ENTER USER COMMAND:
input
DO YOU HAVE A SAVED DATA FILE?
n
ENTER THE PRIMARY STRUCTURE NAME:
METHANOL (STAGGERED)
ENTER THE NUMBER OF ATOMS IN THE PRIMARY STRUCTURE:
6
ENTER X,Y, AND Z COORDINATES, AND ATOMIC NO FOR ATOM   1
-0.73 0.0 0.0 6
ENTER THE ATOMIC SYMBOL FOR ATOM   1
C
ENTER X,Y, AND Z COORDINATES, AND ATOMIC NO FOR ATOM   2
OR ENTER "E" AND CR TO REENTER INFORMATION FOR ATOM   1
0.67
INPUT ERROR, TRY AGAIN
ENTER X,Y, AND Z COORDINATES, AND ATOMIC NO FOR ATOM   2
OR ENTER "E" AND CR TO REENTER INFORMATION FOR ATOM   1
0.67 0.0 0.0 8
ENTER THE ATOMIC SYMBOL FOR ATOM   2
O
ENTER X,Y, AND Z COORDINATES, AND ATOMIC NO FOR ATOM   3
OR ENTER "E" AND CR TO REENTER INFORMATION FOR ATOM   2
-1.05 1.05 0.0 1
ENTER THE ATOMIC SYMBOL FOR ATOM   3
H
ENTER X,Y, AND Z COORDINATES, AND ATOMIC NO FOR ATOM   4
OR ENTER "E" AND CR TO REENTER INFORMATION FOR ATOM   3
-1.07 -0.52 0.9 1
ENTER THE ATOMIC SYMBOL FOR ATOM   4
H
ENTER X,Y, AND Z COORDINATES, AND ATOMIC NO FOR ATOM   5
OR ENTER "E" AND CR TO REENTER INFORMATION FOR ATOM   4
1.07 -0.52 -0.9 1
ENTER THE ATOMIC SYMBOL FOR ATOM   5
H
ENTER X,Y, AND Z COORDINATES, AND ATOMIC NO FOR ATOM   6
OR ENTER "E" AND CR TO REENTER INFORMATION FOR ATOM   5
E
ENTER X,Y, AND Z COORDINATES, AND ATOMIC NO FOR ATOM   5
OR ENTER "E" AND CR TO REENTER INFORMATION FOR ATOM   4
-1.07 0.52 -0.9 1
ENTER THE ATOMIC SYMBOL FOR ATOM   5
H
ENTER X,Y, AND Z COORDINATES, AND ATOMIC NO FOR ATOM   6
OR ENTER "E" AND CR TO REENTER INFORMATION FOR ATOM   5
0.97 -0.89 0.0 1
ENTER THE ATOMIC SYMBOL FOR ATOM   6
H
ENTER "E" AND CR TO REENTER THE LAST ATOM, OTHERWISE JUST ENTER CR


ENTER USER COMMAND:
PLOT

DO YOU WANT TO CHANGE THE PLOTTING PARAMETERS (Y/N)?
(CURRENT DEVICE = 4027, FOR OTHER PARAMETERS SEE THE MENU)
N
```




```
ENTER USER COMMAND:
SHOW
REPORT TO BE PRINTED ON CRT (0) OR LINE PRINTER (1) ?
0
METHANOL (STAGGERED)

NUMBER OF ATOMS =   6
                    X          Y          Z
 C ( 1)       -0.7300     0.0000     0.0000
 O ( 2)        0.6700     0.0000     0.0000
 H ( 3)       -1.0500     1.0500     0.0000
 H ( 4)       -1.0700    -0.5200     0.9000
 H ( 5)       -1.0700     0.5200    -0.9000
 H ( 6)        0.9700    -0.8900     0.0000

CONNECTION TABLE (BOND TYPE IN PARENTHESIS)
C ( 1) IS BONDED TO:  2(1),  3(1),  4(1),  5(1),
O ( 2) IS BONDED TO:  1(1),  6(1),
H ( 3) IS BONDED TO:  1(1),
H ( 4) IS BONDED TO:  1(1),
H ( 5) IS BONDED TO:  1(1),
H ( 6) IS BONDED TO:  2(1),

BOND LENGTHS
C ( 1)- O ( 2) =    1.4000   C ( 1)- H ( 3) =    1.0977
C ( 1)- H ( 4) =    1.0936   C ( 1)- H ( 5) =    1.0936
O ( 2)- H ( 6) =    0.9392

BOND ANGLES
 O ( 2)- C ( 1)- H ( 3)   =  106.9493      O ( 2)- C ( 1)- H ( 4)   =  108.1133
 O ( 2)- C ( 1)- H ( 5)   =  108.1133      H ( 3)- C ( 1)- H ( 4)   =  111.3584
 H ( 3)- C ( 1)- H ( 5)   =   56.9435      H ( 4)- C ( 1)- H ( 5)   =  143.7738
 C ( 1)- O ( 2)- H ( 6)   =  108.6280

DIHEDRAL ANGLES
 H ( 3) - C ( 1) - O ( 2) - H ( 6) =  180.0002
 H ( 4) - C ( 1) - O ( 2) - H ( 6) =   59.9817
 H ( 5) - C ( 1) - O ( 2) - H ( 6) = -120.0185

DISTANCE MATRIX
                1          2          3          4          5
 1  C     0.000000
 2  O     1.400000   0.000000
 3  H     1.097679   2.015167   0.000000
 4  H     1.093618   2.026820   1.809779   0.000000
 5  H     1.093618   2.026820   1.044653   2.078846   0.000000
 6  H     1.918880   0.939202   2.800714   2.260199   2.638124

                6
 6  H     0.000000


ENTER USER COMMAND:
SAVD
ENTER FILE NAME.
METH.DAT
ENTER CONVERSION CODE HERE (0=NO CONVERSION, 1=MINDO,
2=MOLECULAR MECHANICS, 3=AB INITIO, AND 4=MNDO).
0
```

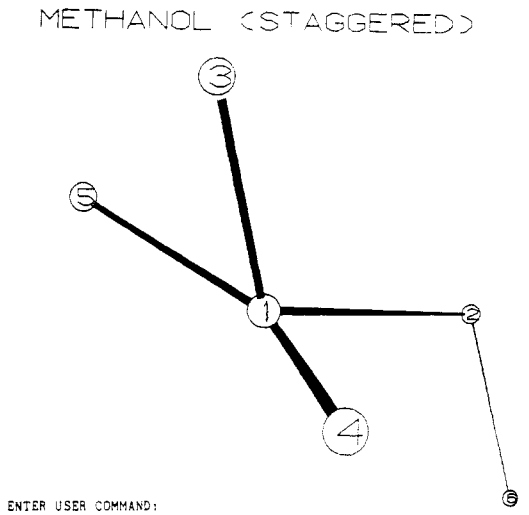
Figure 4

```
ENTER USER COMMAND:
PLOT

DO YOU WANT TO CHANGE THE PLOTTING PARAMETERS (Y/N)?
(CURRENT DEVICE = 4027, FOR OTHER PARAMETERS SEE THE MENU)
Y

ENTER ONE OF THE FOLLOWING:
 1. PLOT
 2. HELP
 3. RETURN TO COMMAND LEVEL
 4. DISPLAY OPTION - CURRENT OPTION: STICK
 5. NUMBER, NO NUMBER - CURRENTLY NO NUMBER
 6. PAINT, NO PAINT - CURRENTLY NO PAINT
 7. TITLE, NO TITLE - CURRENTLY TITLE
 8. HYDROGEN, NO HYDROGEN - CURRENTLY HYDROGEN
 9. CARBON, NO CARBON - CURRENTLY CARBON
10. AXES, NO AXES - CURRENTLY NO AXES
11. CHANGE SIZE PARAMETERS
12. DEVICE - CURRENTLY 4027
13. 2D OR 3D - CURRENTLY 3D
2
  DISPLAY OPTION
        PERMITS THE USER TO DEFINE HOW THE MOLECULE WILL BE
        REPRESENTED.  THE OPTIONS ARE (1) STICK, (2) BALL AND
        STICK, (3) SYMBOL, (4) SPACE FILLED AND (5) DOT
        CLOUD.  THE DEFAULT OPTION IS STICK.
  NUMBER, NO NUMBER
        ALLOWS THE USER TO TURN ON OR OFF THE NUMBERING OF
        THE ATOMS.
  PAINT, NO PAINT
        ALLOWS THE USER TO TURN ON OR OFF THE COLORING OF
        THE ATOMS.  COLOR IS NOT AVAILABLE ON ALL DEVICES.
  TITLE, NO TITLE
        ALLOWS THE USER TO TURN ON OR OFF THE PRINTING OF THE
        TITLE ON THE PLOT.
  HYDROGEN, NO HYDROGEN
        ALLOWS THE USER TO TURN ON OR OFF THE PLOTTING OF THE
        HYDROGEN ATOMS.
  CARBON, NO CARBON
        ALLOWS THE USER TO TURN ON OR OFF THE PLOTTING OF THE
        CARBON ATOMS.
  AXES, NO AXES
        ALLOWS THE USER TO TURN ON OR OFF THE PLOTTING OF THE
         X,Y AND Z COORDINATE AXES.
        IF THERE IS A + SIGN NEAR THE JUNCTION OF THE AXES, THE
        VIEW AXIS COORDINATE VALUE OF ALL THE ATOMS ARE POSITIVE.
        A - SIGN INDICATES ALL THE VALUES ARE NEGATIVE.
  SIZE
        THIS COMMAND ALLOWS THE USER TO SPECIFY THE
        FOLLOWING PARAMETERS:
        1. SIZES OF ATOM SYMBOLS FOR PLOTTING.
        2. BOND THICKNESS FOR CHEMICAL BONDS.
        3. DISTANCE FROM VIEWPOINT TO ORIGIN.
        4. DISTANCE FROM PROJECTION PLANE TO ORIGIN.
        5. THE VIEWPORT SIZE WHICH SETS THE PLOT SIZE.
  DEVICE
        SETS THE DESIRABLE DEVICE AND OPTION CODES.


ENTER ONE OF THE FOLLOWING:
 1. PLOT
 2. HELP
 3. RETURN TO COMMAND LEVEL
 4. DISPLAY OPTION - CURRENT OPTION: STICK
 5. NUMBER, NO NUMBER - CURRENTLY NO NUMBER
 6. PAINT, NO PAINT - CURRENTLY NO PAINT
 7. TITLE, NO TITLE - CURRENTLY TITLE
 8. HYDROGEN, NO HYDROGEN - CURRENTLY HYDROGEN
 9. CARBON, NO CARBON - CURRENTLY CARBON
10. AXES, NO AXES - CURRENTLY NO AXES
11. CHANGE SIZE PARAMETERS
12. DEVICE - CURRENTLY 4027
13. 2D OR 3D - CURRENTLY 3D
4

ENTER DISPLAY OPTION NUMBER:
1. STICK
2. BALL AND STICK
3. SYMBOL
4. SPACE FILLED - ONLY AVAILABLE ON THE AED TERMINAL
5. DOT CLOUD
5

ENTER ONE OF THE FOLLOWING:
 1. PLOT
 2. HELP
 3. RETURN TO COMMAND LEVEL
 4. DISPLAY OPTION - CURRENT OPTION: DOT CLOUD
 5. NUMBER, NO NUMBER - CURRENTLY NO NUMBER
 6. PAINT, NO PAINT - CURRENTLY NO PAINT
 7. TITLE, NO TITLE - CURRENTLY TITLE
 8. HYDROGEN, NO HYDROGEN - CURRENTLY HYDROGEN
 9. CARBON, NO CARBON - CURRENTLY CARBON
10. AXES, NO AXES - CURRENTLY NO AXES
11. CHANGE SIZE PARAMETERS
12. DEVICE - CURRENTLY 4027
13. 2D OR 3D - CURRENTLY 3D
1
```

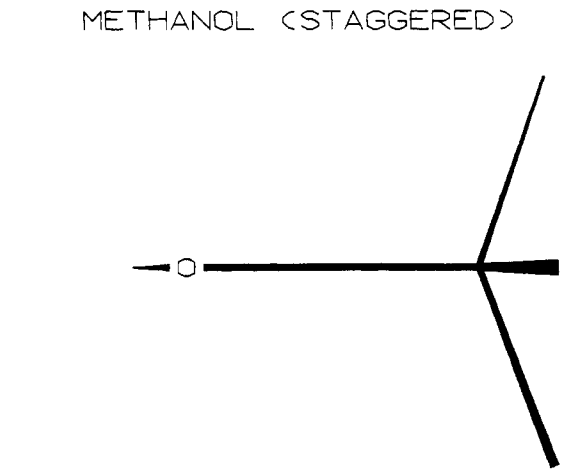


```
ENTER USER COMMAND:
ROTW
ENTER THE ROTATION AXIS (1=X,2=Y,3=Z):
2
ENTER ANGLE OF ROTATION (DEGREES):
45
  45.0 DEG ROTATION ABOUT "Y   " AXIS.

ENTER USER COMMAND:
LIST
METHANOL (STAGGERED)
     6     1
( 1)  C       6   -0.5162    0.0000    0.5162
( 2)  O       8    0.4738    0.0000   -0.4738
( 3)  H       1   -0.7425    1.0500    0.7425
( 4)  H       1   -0.1202   -0.5200    1.3930
( 5)  H       1   -1.3930    0.5200    0.1202
( 6)  H       1    0.6859   -0.8900   -0.6859

( 1)      6    2    3    4    5    0    0    0    0 1 1 1 1 0 0 0 0
( 2)      8    1    6    0    0    0    0    0    0 1 1 0 0 0 0 0 0
( 3)      1    1    0    0    0    0    0    0    0 1 0 0 0 0 0 0 0
( 4)      1    1    0    0    0    0    0    0    0 1 0 0 0 0 0 0 0
( 5)      1    1    0    0    0    0    0    0    0 1 0 0 0 0 0 0 0
( 6)      1    2    0    0    0    0    0    0    0 1 0 0 0 0 0 0 0

ENTER USER COMMAND:
pl

DO YOU WANT TO CHANGE THE PLOTTING PARAMETERS (Y/N)?
(CURRENT DEVICE = 4027, FOR OTHER PARAMETERS SEE THE MENU)
y


ENTER ONE OF THE FOLLOWING:
 1. PLOT
 2. HELP
 3. RETURN TO COMMAND LEVEL
 4. DISPLAY OPTION - CURRENT OPTION: STICK
 5. NUMBER, NO NUMBER - CURRENTLY NO NUMBER
 6. PAINT, NO PAINT - CURRENTLY NO PAINT
 7. TITLE, NO TITLE - CURRENTLY TITLE
 8. HYDROGEN, NO HYDROGEN - CURRENTLY HYDROGEN
 9. CARBON, NO CARBON - CURRENTLY CARBON
10. AXES, NO AXES - CURRENTLY NO AXES
11. CHANGE SIZE PARAMETERS
12. DEVICE - CURRENTLY 4027
13. 2D OR 3D - CURRENTLY 3D
4

ENTER DISPLAY OPTION NUMBER:
1. STICK
2. BALL AND STICK
3. SYMBOL
4. SPACE FILLED - ONLY AVAILABLE ON THE AED TERMINAL
5. DOT CLOUD
2

ENTER ONE OF THE FOLLOWING:
 1. PLOT
 2. HELP
 3. RETURN TO COMMAND LEVEL
 4. DISPLAY OPTION - CURRENT OPTION: BALL AND STICK
 5. NUMBER, NO NUMBER - CURRENTLY NO NUMBER
 6. PAINT, NO PAINT - CURRENTLY NO PAINT
 7. TITLE, NO TITLE - CURRENTLY TITLE
 8. HYDROGEN, NO HYDROGEN - CURRENTLY HYDROGEN
 9. CARBON, NO CARBON - CURRENTLY CARBON
10. AXES, NO AXES - CURRENTLY NO AXES
11. CHANGE SIZE PARAMETERS
12. DEVICE - CURRENTLY 4027
13. 2D OR 3D - CURRENTLY 3D
5

ENTER ONE OF THE FOLLOWING:
 1. PLOT
 2. HELP
 3. RETURN TO COMMAND LEVEL
 4. DISPLAY OPTION - CURRENT OPTION: BALL AND STICK
 5. NUMBER, NO NUMBER - CURRENTLY NUMBER
 6. PAINT, NO PAINT - CURRENTLY NO PAINT
 7. TITLE, NO TITLE - CURRENTLY TITLE
 8. HYDROGEN, NO HYDROGEN - CURRENTLY HYDROGEN
 9. CARBON, NO CARBON - CURRENTLY CARBON
10. AXES, NO AXES - CURRENTLY NO AXES
11. CHANGE SIZE PARAMETERS
12. DEVICE - CURRENTLY 4027
13. 2D OR 3D - CURRENTLY 3D
1
```

Figure 4 continued




```
ENTER USER COMMAND:
START

ENTER USER COMMAND:
INPUT
DO YOU HAVE A SAVED DATA FILE?
Y
ENTER FILE NAME.
METH.DAT
ENTER CONVERSION CODE HERE (0=NO CONVERSION, 1=MINDO,
2=MOLECULAR MECHANICS, 3=AB INITIO, AND 4=MNDO).
0
```




```
ENTER USER COMMAND:
Y
VIEW ALONG THE "Y   " AXIS

ENTER USER COMMAND:
PL

DO YOU WANT TO CHANGE THE PLOTTING PARAMETERS (Y/N)?
(CURRENT DEVICE = 4027, FOR OTHER PARAMETERS SEE THE MENU)
N
```

```
ENTER USER COMMAND:
QUIT
@
```

Figure 4 continued

with the exception of the QUIT command. The function of each command is briefly discussed in the following:

(1) ANGLE
The ANGLE command allows the user to *calculate* or *modify* the (bond) angle defined by three atoms.

(2) BOND
The BOND command allows the user to specify a bond (connection) between two atoms in the primary structure and define its bond type.

(3) CONVERT
The CONVERT command converts 2-D structures to 3-D structures. After the user enters the CONVERT module, the system will perform the task by carrying out structural optimization using a simplified force field method.

(4) DEVICE
The DEVICE command sets the desirable device and option code. This allows the user to specify the type of device he wishes to use for graphic input or output.

(5) DISTANCE
The DISTANCE command allows the user to *calculate* or *modify* the internuclear distance between any two atoms in the primary structure.

(6) CDRAFT
The CDRAFT command enters the Chemical Draftsman module, which allows the user to add text and special symbols to a display for publications and presentations.

(7) FIT
The FIT command enters the least-squares fit module, which performs a comparison of two structures.

(8) GINPUT
The GINPUT or G command enters the graphics input module, which is necessary for building or modifying a structure of interest. Once the command is entered, the system will respond by printing the following message: 2D OR 3D APPLICATIONS? ENTER 3 FOR 3D, OTHERS FOR 2D:. When the dimension is specified, a template (as shown in Figure 2a) will be drawn on the terminal screen along with the commands available and the current primary structure (if present). The primary structure serves as the starting point for building larger structures; furthermore, it may be modified in a variety of ways by choosing a different template command. The template commands will be explained in detail below. Once the template is drawn, the computer will ask the user to HIT CR AND ENTER A TEMPLATE COMMAND. CR stands for carriage return. In the template command mode, the user must employ the graphic input device (GID) to select the template commands. The template commands (Figure 2a illustrates the main template as it appears on the screen) are as follows:

(8.1) FRAGMENT: The FRAGMENT command will erase the main template and draw a FRAGMENT TEMPLATE (as shown in Figure 2b). This template allows the user to select various fragments of choice in a variety of ways. Once the fragments needed have been chosen (up to a total of four), the user may return to the main GINPUT template.

(8.2) DUP: The DUP command allows the primary structure or a fragment to be duplicated.

(8.3) DEL: The DEL command allows a fragment or the primary structure to be deleted.

(8.4) USER: The USER command permits the user to enter an existing file in any of several allowable formats. If there is no primary structure or no fragments present in the working area, the file entered will become the primary structure. It is necessary to have a primary structure before any building or manipulating can be done. If a primary structure already exists in the working area, the entered file will simply become another fragment, provided there are less than four fragments present. If there exists a primary structure and four fragments in the working space, the USER command may not be implemented.

(8.5) SIZE: The SIZE command enables the user to enlarge or shink the display.

(8.6) HELP: The HELP command provides information on the different template commands.

(8.7) ROTW: The ROTW command will rotate the primary structure of a fragment. The coordinates of the structure are changed to reflect the rotation.

(8.8) ROTB: The ROTB command will rotate specified atoms of the primary structure or a fragment through an axis defined by two atoms of the chosen structure.

(8.9) MOVE: The MOVE command will move the primary structure or a fragment any place in the template display areas.

(8.10) H ON/OFF, C ON/OFF, # ON/OFF: The H ON/OFF, C ON/OFF, and # ON/OFF commands allow the user to either show the H (hydrogens), C (carbons), or # (numbers) or to suppress them.

(8.11) CLR ON/OFF: The CLR ON/OFF command is most useful on a color terminal. When implemented, all atoms and bonds are colored. The structure(s) will be drawn as ball and stick regardless of the current display option. On a black and white screen the atoms will be shaded.

(8.12) STRUCT PARMS: The STRUCT PARMS command allows the user to change the default parameter values for bond length, bond angle, and dihedral angle.

(8.13) ADD C H N O S: The ADD C H N O S commands allow the user to add a C (carbon), H (hydrogen), N (nitrogen), O (oxygen), or S (sulfur).

(8.14) DELETE ATOM: The DELETE ATOM command allows the user to delete atoms from the primary structure that appear within the template working area.

(8.15) GENERATE H'S: The GENERATE H'S command will add the correct number of hydrogens needed to complete the primary structure.

(8.16) REPLACE ATOM: The REPLACE ATOM command allows an atom of the primary structure to be replaced by any other atom on the periodic table.

(8.17) METHYL, METHYLENE, PHENYL: The METHYL, METHYLENE, and PHENYL commands will permit the user to replace an atom of the primary structure with any of these groups.

(8.18) CONNECT: The CONNECT command permits a primary structure and fragments to be connected with standard bond lengths and bond angles. The CONNECT command may also be used to form a bond between two atoms of the primary structure.

(8.19) DISCONNECT: This DISCONNECT command will disconnect the bond between two atoms of the primary structure.

(8.20) FUSE: The FUSE command allows the user to fuse a fragment to the primary structure.

(8.21) BOND: This BOND command permits the user to change the bond type of any of the bonds of the primary structure in the working space.

(8.22) RENUMBER: The RENUMBER command will renumber the connected atoms of the primary structure.

(8.23) REDRAW: The REDRAW command will erase and redraw everything within the template working space.

(8.24) NORM: The NORM command adjusts all bond lengths and bond angles according to a simplified two-dimensional force field method.

(8.25) SAVE DATA: The SAVE DATA command permits the user to save all structures within the template working area into a file.

(8.26) END: The END command permits the user to return to the main command level.

(8.27) CD: The CD command is used to exit the GINPUT template module and enter the Chemical Draftsman module. The primary structure and any fragments will be brought into the Chemical Draftsman template working area.

(8.28) 2D: The 2D command will erase the current template and redraw the 2D GINPUT template with the primary structure and fragments shown in 2D.

(8.29) 3D: The 3D command will erase the current template and redraw the 3D GINPUT template with the primary structure and fragments shown in 3D.

(9) HELP

The HELP command prints a brief description of all user commands.

(10) INPUT

The INPUT command allows the user to enter the input module; one must enter this module before any data can be manipulated or displayed.

(11) LIST

The LIST command displays the current data; it is similar to the SHOW command but is less informative. After the command has been entered, the system will print the title and number of atoms in the primary structure. On the following lines, the ID number, atomic symbol, atomic number, and the X, Y, and Z coordinates are listed for each atom. This is followed by a connection table; this table tells which atoms are bonded together and the type of bond between the atoms. The LIST command displays the data file as it is stored in the MOLBUL format.

(12) NEWMAN

The NEWMAN command draws quality Newman projections.

(13) ORIENT

The ORIENT command allows the user to place the primary structure of interest in a desired orientation with respect to the X, Y, and Z axes.

(14) PANGLE

The PANGLE command calculates the angle between two planes, where each plane is defined by three unique atoms of the primary structure.

(15) PLOT

The PLOT command instructs the program to make a plot. A plot menu is implemented for this command.

(16) QUIT

The QUIT command enables the user to terminate the program execution. To save the current working data file, the user must issue the SAVD command before quitting.

(17) REPLAY

The REPLAY command requests the program to PLOT using a previous saved file. The system responds to this command by prompting the user for a file name and then plotting the graphic data in that file.

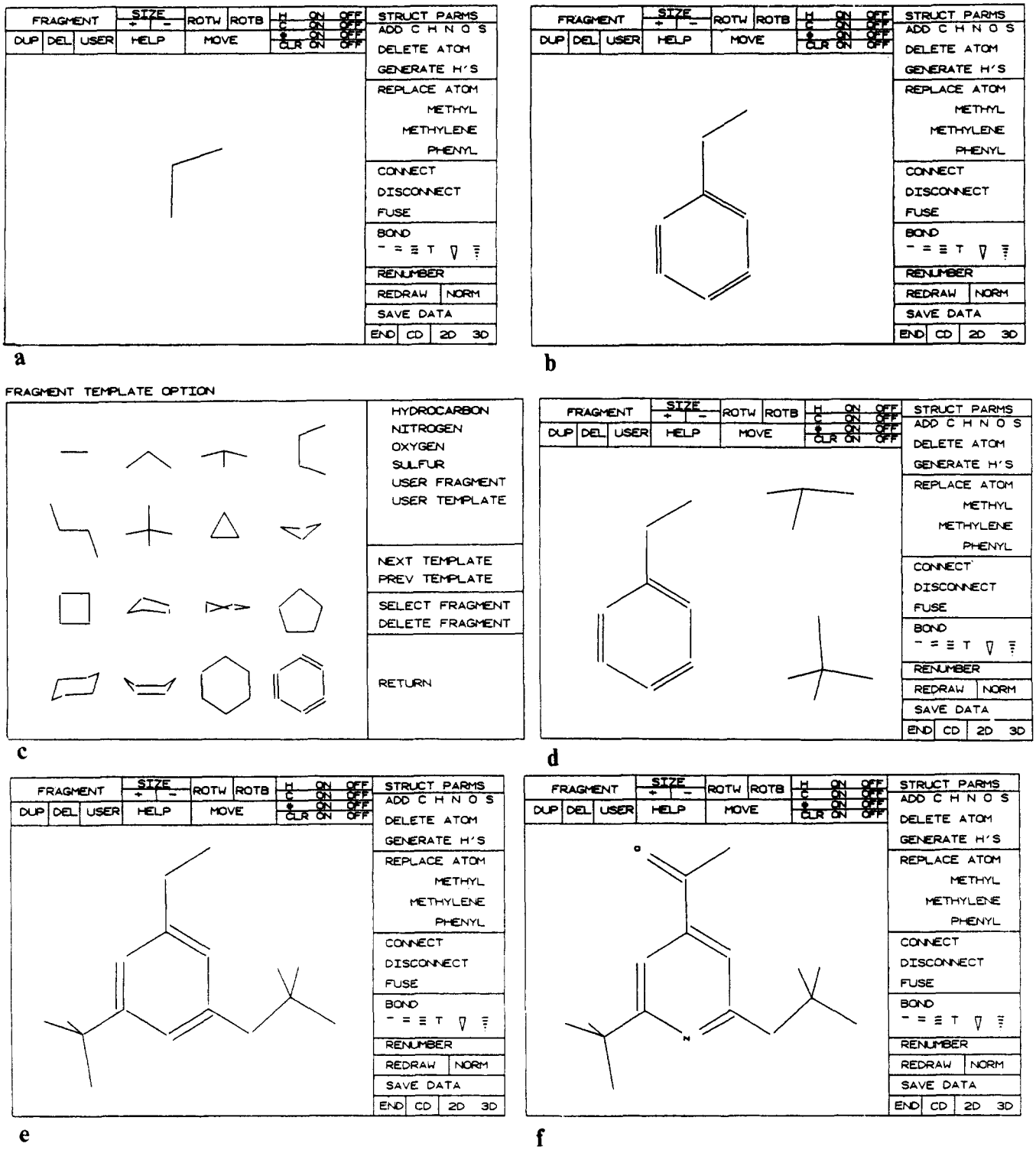


Figure 5. Displays to illustrate the GINPUT operation.

(18) ROTB
The ROTB command permits the user to rotate some (or all) of the specified atoms by an angle, through an axis defined by two atoms of the primary structure. This command is particularly useful for building various conformers or geometric isomers.

(19) ROTW
The ROTW command allows the user to rotate the entire primary structure for a specified angle around an axis.

(20) SAVD
The SAVD command requests the program to store the current data for future usage into a file.

(21) SAVG
The SAVG command permits the user to store the graphic data and environment for future display. The SAVG command is very useful for creating files for demonstration.

(22) SHOW
The SHOW command outputs on the terminal or line printer a summary of bonded atoms, bond types, bond lengths, bond angles, dihedral angles, and a distance matrix of the primary structure.

(23) START
The START command instructs the program to start from the beginning of MOLBUL without stopping the execution. It can save time in logging into and out of the program when the user would like to

process many different data files.

(24) TANGLE
The TANGLE command allows the user to calculate or modify the torsional angle between any four atoms of the primary structure.

(25) TITLE
The TITLE command allows the user to enter a title for a data file or to change the current data file title.

(26) TRANSLATE
The TRANSLATE command allows the user to translate the primary structure along the *X*, *Y*, or *Z* axis.

(27) X(−X), Y(−Y), Z(−Z)
The X(−X), Y(−Y), and Z(−Z) commands permit the user to view the structures along the *X* (−*X*), *Y* (−*Y*), or *Z* (−*Z*) axis, respectively. (The minus sign in front of the *X*, *Y*, and *Z* allows the user to view the structure along a negative axis.) The system will echo the command; furthermore, the user may wish to utilize the PLOT command to view the structure.

PROGRAM EXAMPLES

Included in this section are two examples of using the MOLBUL program. Example 1 demonstrates the use of several top level commands and shows some sample plots. Example 2 illustrates features available in the GINPUT mode for building chemical structures.

Example 1 (See Figure 4). After the program is started, the HELP command is given and then the INPUT mode is entered. This first use of the input module illustrates how data are entered directly into the program. Later in the example the input module is used to enter a file name that contains the necessary data. The PLOT command is given without changing any of the parameters to produce the first display. Next, the SHOW command prints out all the structural information about the primary structure, and the SAVD command stores the information in the file METH.DAT. The plot command is given again, and the structure is displayed with the dot-cloud display feature. The ROTW command is then used to rotate the primary structure 45° about the *Y* axis. The LIST command is used to show that the coordinates are changed to reflect the rotation but the connection table is unchanged. The new primary structure is then displayed with the appropriate parameters changed to produce a ball and stick plot with the atoms numbered.

The START command is given, which initializes the program. The INPUT module is again entered, but this time the previously saved data file (METH.DAT) is used. Since the methanol file was saved in the MOLBUL format, it must be read in with no conversion. However, another file with the format from the MINDO, molecular mechanics, ab initio, or MNDO tools could just as easily be read in through the INPUT module. View along the *Y* axis is specified, and the display is produced with the default plotting parameters. The program is then ended with the QUIT command.

Example 2. Procedures for building the 3-D structure shown in Figure 5f, using the GINPUT mode of MOLBUL, are outlined in this example. There are easier and quicker ways to build this structure than the procedures presented here. However, the purpose of this example is to illustrate some of the more important template commands available in the GINPUT mode. The GINPUT mode is based on the interactive use of a graphic input device, which cannot be fully demonstrated here. Only a few CRT displays are illustrated.

Enter the GINPUT mode by issuing the GINPUT user command. The initial primary structure, a three carbon atom chain, is entered into the working area via the USER template command (Figure 5a). Next, one of the carbon atoms is replaced by a phenyl group with the REPLACE PHENYL template command (Figure 5b). The FRAGMENT template command is then selected to obtain the structures for the two alkyl substituents from the fragment database. This causes the programs to display a fragment template. To display the hydrocarbon fragments, the HYDROCARBON template command is selected (Figure 5c). By use of the SELECT FRAGMENT command the *tert*-butyl and neopentyl fragments may be selected and then brought back to the main GINPUT template with the RETURN command (Figure 5d).

The two fragments are then connected by employing the CONNECT command (Figure 5d). To add the ketone functionality, the ADD O command is used followed by the BOND = command. Finally, the structure is completed by replacing a ring carbon with a nitrogen atom with the REPLACE atom command (Figure 5f).

CONCLUSIONS

Ways to represent molecular structures are the heart of any molecular information and modeling system. In conjunction with our activities in the area of molecular information and molecular modeling, we have developed the MOLBUL (*Mol*bul *Bui*lder) program, the front end of our system. MOLBUL is a versatile, efficient, integrated, and interactive program for chemists to build and display 3-D or 2-D molecular structures for theoretical calculations, graphic art work, and chemical information systems. MOLBUL is a multifunction multipurpose tool. It can be used with various (*low-cost*) graphics terminals. To the best of our knowledge, no such *unique*, *integrated*, and *friendly* software is currently available elsewhere in the public domain.

ACKNOWLEDGMENT

We thank our managers R. Thomson and R. Waugh for their support and Dr. J. Seeman, Dr. E. Southwick, and M. Sito for their participation and various contributions.

REFERENCES AND NOTES

(1) Kao, J.; Day, V.; Watt, L. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 129.
(2) Wiswesser, W. J. *Comput. Autom.* **1970**, *19*, 2. Smith, E. G. "The Wiswesser Line-Formula Chemical Notation"; McGraw-Hill: New York, 1968. Hyde, E.; Matthews, F. W.; Thomson, L. H.; Wiswesser, W. J. *J. Chem. Doc.* **1967**, *7*, 200. Thomson, L. H.; Hyde, E.; Matthew, F. W. *J. Chem. Doc.* **1967**, *7*, 204. Wipke, W. T.; Heller, S. R.; Feldmann, R. J.; Hyde, E., Eds. "Computer Representation and Manipulation of Chemical Information"; Wiley: New York, 1974.
(3) Feldmann, R. J.; Milne, G. W. A.; Heller, S. R.; Fein, A.; Miller, J. A.; Koch, B. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 157. Fujiwara, Y.; Nakayama, T. *Anal. Chim. Acta* **1981**, *133*, 647. Moreau, G.; *Nouv. J. Chim.* **1980**, *4*, 17. Farmer, N. A.; O'Hara, M. P. *Database*, **1980**, *3*, 10. Howe, W. J.; Hagadone, T. R. *ACS Symp. Ser.* **1978**, *No. 84*, 107. Wipke, W. T.; Dyott, T. M. *J. Am. Chem. Soc.* **1974**, *96*, 4825, 4834.
(4) Burkert, U.; Allinger, N. L. "Molecular Mechanics"; American Chemical Society: Washington, DC, 1982. Osawa, E.; Musso, H. *Top. Stereochem.* **1982**, *13*, 117. Kao, J.; Allinger, N. L. *J. Am. Chem. Soc.* **1977**, *99*, 975. Kao, J.; Huang, T. N. *J. Am. Chem. Soc.* **1979**, *101*, 5546.
(5) Kao, J.; Leister, D.; Sito, M. *Tetrahedron Lett.* **1985**, *26*, 2403. Kao, J.; Eyermann, C.; Southwick, E.; Leister, D. *J. Am. Chem. Soc.*, in press. Kao, J., unpublished data.
(6) Presented in part at the 2nd Conference on Computers and Chemistry, Tallahassee, FL, 1983, and the 25th Annual Medicinal Chemistry Symposium, Buffalo, NY, 1984.
(7) This program will be submitted to QCPE for distribution.
(8) For example, Martin, P. *The Economist*, **1982**, *August*, 67.
(9) A common connection table is maintained for both 2-D and 3-D structures. The format of our connection tables is very similar to the one used in the NIH/EPA SANSS database. However, our connection tables are potentially more flexible and are not limited to neutral organic compounds. The details of our connection table and other features will come with the user's manual after we release the software.
(10) Newman, M. S. *Rec. Chem. Prog.* **1952**, *13*, 111; *J. Chem. Educ.* **1955**, *32*, 344.

(11) The detailed algorithms of NEWMAN are described in Kao, J.; Watt, L. *Comput. Chem.*, in press.
(12) Nyburg, S. C. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **1974**, *B30*, 251.
(13) The detailed algorithms of CDRAFT are described in Watt, L.; Kao, J. *Comput. Chem.*, in press.
(14) Houminer, Y.; Kao, J.; Seeman, J. I. *J. Chem. Soc., Chem. Commun.* **1984**, 1608.

Procedures for Sorting Chemical Names for *Chemical Abstracts'* Indexes

ALLEN C. ISENBERG, JOANN T. LEMASTERS, ABE F. MAXWELL, and GERALD G. VANDER STOUW*

Chemical Abstracts Service, Columbus, Ohio 43210

Received January 25, 1985

In the preparation of each *Chemical Substance Index* to *Chemical Abstracts* (CA), nearly three-quarters of a million chemical substance names must be sorted by computer program into an invariant order. This sorting is done on sortkeys that are generated from the character strings in the names and is done in a way that takes advantage of the data elements used by Chemical Abstracts Service (CAS) in preparing these names. The organization of CA index nomenclature and the rules used in sortkey generation are described.

INTRODUCTION

The *Chemical Substance Index* to *Chemical Abstracts* (CA) each year includes index entries that refer to nearly three-quarters of a million different chemical substances. These alphabetical indexes, which are published twice annually, are merged every 5 years into a collective index. The preparation of these volume and collective indexes requires that very large lists of chemical substance names be sorted into a consistent order, so that the user of the printed indexes can locate a substance of interest with confidence that it has been placed at the correct point in the index.

For many years the preparation of the CA indexes required the efforts of a group of clerical staff who devoted their time to sorting thousands of index entries typed on separate cards. Although manual sorting achieved remarkably consistent results, the rapid growth of the indexes during the 1950s and 1960s made maintaining the quality of these efforts increasingly difficult and expensive. Since the early 1970s, Chemical Abstracts Service (CAS) has used computer processing extensively in the preparation of its indexes.[1] These computer systems include programs that carry out, with no human intervention, the sorting of chemical names that was formerly done by hand. Recently published descriptions of two algorithms for sorting chemical names[2,3] prompt us to describe the procedures that CAS uses in sorting names.

DATA ELEMENT STRUCTURE FOR CHEMICAL NAMES

To appreciate the way CAS sorts chemical names, it is necessary to understand two general aspects of the sorting process: first, the way in which CAS constructs a chemical name from individual data elements and ranks these data elements for sorting purposes; second, the way in which data elements that occur at the same ranking level are sorted by the use of sortkeys. The first two sections of this paper discuss data elements and their utility in sorting; the last section describes the use of sortkeys.

CAS uses an extensive and rigorous set of rules for generating chemical names. These rules, which are applied by human nomenclature experts with extensive computer support, ensure that a given chemical substance can be found at a predictable place in the printed *Chemical Substance Index*.[4] The systematic names that result from these rules appear in the *Index* in an "inverted" form; i.e., that portion of the name that refers to a "parent" structure is given before the names of the structural fragments that are attached to that parent structure. Thus, for example, the name

2-Butenedioic acid, 2-butyl-

gives the parent name 2-Butenedioic acid before the name of the attached substituent represented by the string 2-butyl-. The corresponding "uninverted" name would be

2-butyl-2-butenedioic acid

In the inverted form of this name, the characters before the first comma (sometimes referred to as the "comma of inversion") constitute the data element known as the *heading parent*. This data element normally has one of three forms: (a) a molecular skeleton name such as Butene, to which is attached the name of the principal functional group if one is present (dioic acid in this instance); (b) a functional parent compound in which no skeleton is expressed, such as Carbonic acid; (c) a trivially named parent such as Phenol or Urea. The names of the attached substituents, such as 2-butyl-, are included in a separate *substituent* data element.

The *heading parent* and *substituent* are two of the data elements that CAS uses for chemical names; the others are described later in this section. These data elements are assigned by the nomenclature specialist when a name is prepared. As described in the next section of this paper, the data element identifications play an important role in the sorting programs. They are also important in formating names for the printed indexes. The formating programs use the data elements to determine, for example, that two names which sort together have identical heading parents; the heading parent then needs to be printed only in the first name and can be represented by a long dash in the second name. Similarly, if two esters of an acid sort together, the formating process will cause the name of the acid to appear only once, with the two esters identified under it. The data element identifiers do not themselves appear in CAS printed services or online files, however.

Frequently a name contains a character string that describes a derivative of the principal functional group, such as the ester of an acid or the oxime or hydrazone of a ketone. Thus, for example, if the above name were modified to

2-Butenedioic acid, 2-butyl-, dimethyl ester

the string dimethyl ester would constitute the *name modifi-*

0095-2338/85/1625-0410$01.50/0