

Computer Retrieval and Analysis of Molecular Geometry.

I. General Principles and Methods

BY PETER MURRAY-RUST

Department of Chemistry, University of Stirling, Stirling, Scotland

AND SAM MOTHERWELL

Cambridge Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge, England

(Received 17 January 1978; accepted 9 March 1978)

Methods are described for the efficient retrieval of reliable structural information from the Cambridge Crystallographic Data File. Statistical analysis of molecular structure by multivariate techniques is outlined.

Introduction

The Cambridge Crystallographic Data Centre (CCDC) maintains a file of organic and organometallic crystal structures which is now available in many countries and is a unique tool for the study of crystal and molecular geometry (Kennard, Watson, Allen, Motherwell, Town & Rodgers, 1975). The effect of automating the retrieval of structural data may be likened to the introduction of computer control of single-crystal diffractometers. There is a vast amount of structural and chemical information as yet unrevealed in the 19 000+ published organic crystal structures and, because of the growth of structure analysis as a chemical tool, this is rapidly increasing. Exploitation of this information is exemplified by studies on reaction pathways (*e.g.* Bürgi, Dunitz & Shefter, 1973; Murray-Rust, Bürgi & Dunitz, 1975) and the deformation of benzene rings (Domenicano, Vaciano & Coulson, 1975). This type of study has so far been infrequent because it requires a manual search of the literature and transcription of results, but it is now made enormously easier with the aid of the computer and the data file. In this paper we describe the general principles of the computerized study of molecular geometry and in following papers show some of the statistical procedures necessary (Murray-Rust & Bland, 1978; hereafter MB) and give an example of the method (Murray-Rust & Motherwell, 1978; hereafter MM).

The numeric data on the file can be used to investigate a number of effects (*e.g.* crystal packing, disorder, intermolecular interactions) but we confine ourselves here to analysis of molecular geometry. A given molecule, or part of a molecule, may occur in many different crystal structures and show considerable variation in its molecular geometry. This variation

reflects the different crystal environment in each case (crystal packing effects) but taken as a whole is a measure of the intrinsic variability of the geometry of the molecule. Thus phosphate groups in 211 different environments (Baur, 1974) show much variation in bond lengths and angles and this has been used to investigate the potential-energy surface for the PO_4^{3-} ion (Murray-Rust, Bürgi & Dunitz, 1978*a*). Variability of geometry is, however, often due to experimental errors rather than real crystallographic or chemical effects and these errors may invalidate an automatic analysis unless their effect can be eliminated. We shall therefore discuss in some detail the way data are recorded on the file, particularly that information which indicates the probable accuracy of a structure. The causes and treatment of errors are outlined in the following paper (MB).

A comprehensive description of the CCDC file* is available from Cambridge as are programs and procedures for searching the file and calculation of geometry. We describe here only those aspects which affect the accuracy or reliability of a search or can be used to estimate experimental errors in the data. Some selection of structures is usually necessary in analysing molecular geometry and we shall describe the criteria that may be used in decision-making routines. To carry out the analysis we have used the very widely available *SPSS* package (*Statistical Package for the Social Sciences*; Nie, Hull, Jenkins, Steinbrenner & Bent, 1975) since its file-handling system, Fortran-like data transformation and statistical procedures are exactly

* This term is used loosely in this paper to cover all the files produced by Cambridge (bibliographic, chemical connectivity and structural data). The fragmentation is due to historical and technical reasons and the information can be conceptually regarded as being on a single file.

what are needed for analysis of molecular geometry. Several of the tables contain examples of the way data can be selected with *SPSS*, with mnemonic variable names representing quantities retrieved from file or output by the geometry program.

The Crystallographic Data File

The bibliographic and chemical-connectivity files cover the literature from 1935 onwards. The structural data file at present is complete for 1960 onwards and it is hoped that the pre-1960 entries will be added during 1978. At present (1 January 1978) the CCDC file contains 19 113 entries of which 13 980 have atomic coordinates. The CCDC file is being continually updated with new publications and redeterminations of crystal structures at a rate of about 3000 per year.

It is very important that any analysis of molecular geometry from the file should state exactly which updated version was used for the study.

The data on the file are often more reliable than in the original publications. Typographical errors occur in about 10% of crystal structure publications and are corrected whenever possible by the CCDC. These are detected by the *UNIMOL* program (Allen, Kennard, Motherwell, Town, Watson, Scott & Larson, 1974). The file is described in detail in the publications of the Data Centre (Kennard, Watson & Town, 1972; Allen, Isaacs, Kennard, Motherwell, Pettersen, Town & Watson, 1973; Allen, Kennard, Motherwell, Town &

Watson, 1973) but Table 1 shows schematically some of the information which is important for studying molecular geometry.

Numerical data banks

The main aim of a bank of numeric data is to eliminate the tedious and error-prone process of transcribing tables from the original publications. Ideally the user would also wish to consult the publication to evaluate the reliability of the work since data without information about experimental conditions and errors cannot normally be used with total confidence. The crystallographic data file has a considerable amount of information about experimental techniques and errors and this paper shows how these can be used to assess the reliability of any data set. It will be clear that there will be some cases in which the data file cannot be used without referring to at least part of the original literature but we shall show that these are relatively few and that much evaluation can be carried out by computer.

There are two extreme philosophies which can be followed when compiling a data bank. The first is to include only those data sets which achieve a given level of accuracy. This appears attractive but is in fact difficult since it requires careful checking of every paper and evaluation of the techniques used. Moreover, a consensus of the scientific community must exist for determining the level of accuracy. The alternative, which has been followed by the Cambridge Crystallographic Data Centre, is to include every published experiment and to put in as much about the experimental errors and techniques as space will allow. The advantage of this is that many more data are available and even though some are dubious they may still have limited value. (For example, studies on crystal packing might be possible even with very poor atomic coordinates.)

Because the crystallographic data file contains *every* organic and organometallic crystal structure, any analysis which requires data of high accuracy must include an evaluative procedure (*screening*) to reject structures of insufficient reliability. It is the purpose of these papers to suggest automatic and efficient ways of carrying out this screening, for without it studies of the numeric data on files will rightly be attacked as containing poor data sets.

We now describe a procedure of analysis using the Cambridge program package and *SPSS*, in order to illustrate the general principles involved.

Retrieving numeric data

To extract numeric data from the file we must first identify (*retrieve*) those data sets which are appropriate

Table 1. *Some of the information on the structural data file relating to molecular geometry and its analysis*

Field name	Information
REMARK	General remarks.
DISORD	Nature of partial disorder.
ERROR	Nature of errors in publication and whether corrected.
CELL	Unit-cell dimensions, space-group symbol, <i>etc.</i>
ATOM	Atomic coordinates.
RFACT	<i>R</i> factor.
QUAL	Neutron study, low temperature, absolute configuration.
FLAGS	
INTF	Intensity-collection method.
POL	Polymeric structure.
TDS	Two-dimensional study.
TD	Total disorder.
PD	Partial disorder.
PC	Partial connectivity.
AS	Average $\sigma(C-C)$: 0.001–0.005 (AS = 1), 0.006–0.010 (AS = 2), 0.011–0.030 (AS = 3), >0.030 Å (AS = 4).
ERR	Error status after <i>UNIMOL</i> processing.
VAL	Valency check.
SB	Short bonds (bond length much less than sum of covalent radii).
BC	Calculated and published bond lengths do not agree.
RPA	Error – problem referred to author.

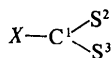
Table 2. Examples of bibliographic search questions for BIBSER, given in order of decreasing efficiency

The compound name field is unfortunately not standardized so names may be missed in some cases.

Question coding	Remarks
(i) Q *CLASS '48'	Very efficient: only amino acids and peptides retrieved.
(ii) Q *COMPND 'TRIPHENYLPHOSPHINE'	Very efficient: no unwanted structures but a few may be missed.
(iii) Q *ELEMENT 'P' AND *COMPND 'ABSOLUTE CONF'	Very efficient.
(iv) Q *COMPND 'PENICILL'	Efficient: but 'ampicillin' would be missed and this is important since only a few structures are on file.
(v) Q *COMPND 'CYCLOHEXADIENE'	Poor: cyclohexa-2,5-diene is missed but 2,5-cyclohexadiene is retrieved.

Table 3. An example of a chemical-connectivity search question for CONNSER, to retrieve XCS₂ fragments (Murray-Rust, 1976)

Coding	Comments
Q CXS2 FRAGMENTS	Title for question.
AT1 C 3 EXACT 0	(Three-coordinate carbon, no H atoms; <i>i.e.</i> normally <i>sp</i> ² .)
AT2 S }	Any sulphur atoms.
AT3 S }	
BO 2 1 }	Any kind of C-S bonds, cyclic or acyclic.
BO 1 3 }	



to the particular problem. Since the process is to be carried out automatically by computer it is important that the efficiency of this *searching* is as high as possible. It is usually far more important to ensure that no unwanted data sets are included by mistake than that a few potentially useful ones are omitted. If, say, 5% of the desired data sets are omitted from a statistical analysis the effect will be hardly noticeable, whereas if 5% of the retrieved data sets are not applicable to the problem, this may be disastrous. The search programs (Tables 2 and 3) available with the file can be very efficient and are discussed briefly later. There are, however, some searches which are not possible with these programs. Most stereochemical features (*e.g.* *cis* and *trans* groups, optical isomers) cannot be coded and indeterminate coordination for metals is often a problem since bonds can be of widely different lengths so that there is no satisfactory structural formula. We present later a method of overcoming most of these problems, by screening derived molecular geometry.

After the bibliographic and connectivity searches are complete we have a file which should contain only structures appropriate to our problem (Fig. 1).

It is often worthwhile to scan the data quickly by eye (*visual screen*) to see how successful the search has

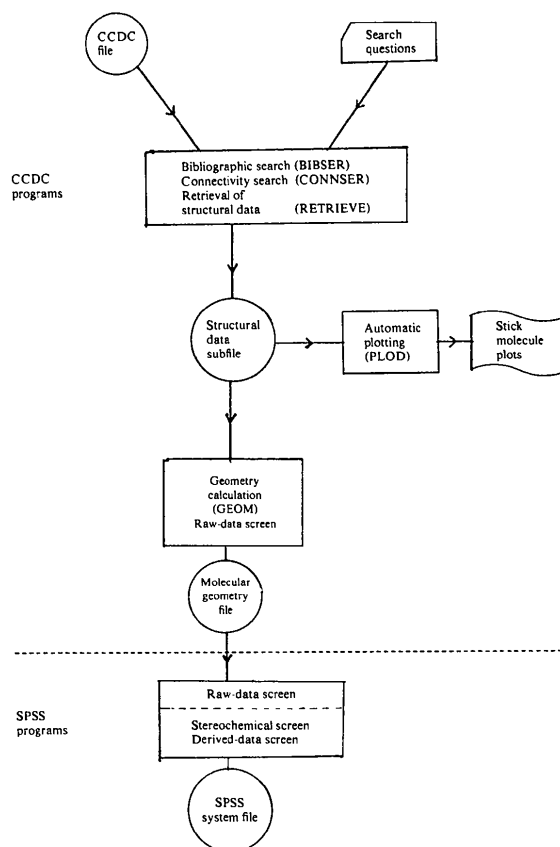


Fig. 1. An overall view of the process for setting up a file of molecular geometry with the CCDC programs and SPSS. Details of CCDC programs are given in Cambridge Crystallographic Data Centre (1976) and Motherwell (1976).

been. The rapid plotting of molecular structures by the computer is particularly useful. This check may often suggest an improvement in the original search question before proceeding further. Some data sets will be found to be inaccurate for the purpose of the study and must be rejected. We shall refer to this rejection as *raw-data screening* and attempt to carry it out with the error indicators on the file. After this screening (whose

position in the procedure will depend on the problem) the file should contain only those data which are accurate enough for our needs. We can now evaluate molecular geometry (*derived data*); routinely bonds, angles and torsion angles will be produced but molecular planes, intermolecular contacts and other quantities can also be calculated.

It may now be advisable to reject further data sets on the basis of their molecular geometry, a process we call *geometry screening* (a particular case of the general process of *derived-data screening*). There are two quite different reasons for this screening. Firstly it may be the only straightforward way of removing structures with a correct chemical connectivity but incorrect stereochemistry (*e.g.* mannose and galactose may be retrieved when only glucose is required). Logically this should be part of the chemical substructure search, but the CCDC chemical-connectivity file does not contain stereochemical information. In practice it is easy to remove unwanted stereoisomers at this stage (*stereochemical screening*). The second and more important reason is that we can screen on the basis of numeric data not explicitly held in the file but calculated from them. This *derived-data screening* is a new concept for data file searching and offers enormous potential to the crystallographer and chemist. Whereas most data files (especially those with inverted files) are set up with particular search questions in mind, the potential of derived-data screening is limited only by the ability of the searcher to express his search question as a numerical relationship. This is one of the reasons why the CCDC file is such a powerful tool because we can now search it to investigate geometrical and stereochemical concepts. Examples of questions will show the power of the derived-data screen:

(i) Retrieve all compounds with a C–S bond longer than 1.80 Å.

(ii) Retrieve compounds with distorted carboxylate anions (*i.e.* unequal C–O bond lengths or C–C–O angles).

(iii) Retrieve compounds with acyclic eclipsed CSSC systems (torsion angles in the range -10° to $+10^\circ$).

(iv) Retrieve compounds with cyclohexane rings in the boat-twist conformation.

The coding for each of these questions is discussed later and shown in Table 6.

Searching the file

(a) Bibliographic searching

The program *BIBSER* retrieves entries which contain certain strings of characters (Table 2). Some searches can be very efficient but for analysing molecular geometry most are only valuable as initial screens before using *CONNSEER*.

(b) Connectivity searching

The program *CONNSEER* retrieves entries which contain a given chemical substructure. A simple example is given in Table 3. The search is very efficient but cannot deal with stereoisomers, valence isomers and compounds with an arbitrary structural formula (such as those with very strong non-bonded interactions or metal compounds with poorly defined coordination spheres). These problems can be overcome by using a somewhat inefficient coding which will include the desired compounds but will also retrieve some unwanted structures. These can then be removed manually or in the automatic stereochemical screen (see later, Table 5).

(c) Raw-data screening

The CCDC file contains several fields in each entry which can be used for raw-data screening, some of which are given in Table 1.

These contain information on the reliability, accuracy and content of the entry. It is important to realize that this information is quite objective and there is no subjective comment by the CCDC on the correctness of the structure, except insofar as the atomic parameters are not consistent with published or normal bond lengths. If data from the file are to be used *automatically* it is essential to exclude all structures which are unreliable or inaccurate as judged by these objective entries. For this reason we shall discuss the various fields for each entry (Table 1).

(i) *Intensity collection method*. If a structure has been determined by X-ray diffraction, the method of intensity measurement is recorded (visual, densitometric or diffractometric). Low-temperature studies, neutron diffraction and determination of absolute configuration are also recorded and can be usefully used in the screening process.

(ii) *R factor*. This is of considerable use and may easily be used to screen out inaccurate structures, although the problem of structures containing heavy atoms should not be overlooked.

(iii) *Estimated standard deviations*. This is a very valuable feature of the file. Although, for reasons of space, individual e.s.d.'s are not held, the average e.s.d. for C–C bonds is indicated as lying within four ranges of accuracy. From these values estimates of the errors in other bonds and angles can be made.

(iv) *Disorder*. Atoms which are disordered have no atomic coordinates listed. The type of disorder is always mentioned in a text field.

(v) *Error set*. This can be due to two causes. If a structure is grossly wrong it is possible that some of the atoms will be in chemically impossible positions giving rise to short bonds or long bonds. More commonly, the structure may be correct but the final publication may contain misprints in the cell dimensions or atomic

Table 4. *Examples of raw-data screens to exclude unwanted structures*

These are coded as instructions for the SPSS package and numeric quantities from the DATA file have been given obvious mnemonic variable names.

- (i) SELECT IF (ERR EQ 0): selects only error-free sets.
- (ii) SELECT IF (ERR EQ 0 AND RFACT LT 0.05 AND AS LE 2): error-free sets with $R < 0.05$ and (C-C) < 0.01 .
- (iii) SELECT IF (TD OR PD EQ 1): selects only disordered structures.
- (iv) COMPUTE CELLVOL = A*B*C; SELECT IF (CELLVOL LT 1000 AND SPGRP EQ 'P212121'): structures in P_212_1 , with smallish cells.

coordinates. Usually this can be corrected at the CCDC by calculation or consulting the author, but otherwise the data are simply flagged as an error set. It is important to realize that an *error-free set* means only that there are no detectable typographical errors; severe crystallographic errors may still be present.

(vi) The inclusion of H atoms in a structure can be deduced from the file and used as a criterion of accuracy. Conversely the proportion of heavy atoms in a structure can be calculated from the molecular formula and if high can be used as a rejection criterion.

Examples of raw-data screens are given in Table 4. The level of raw-data screening will depend on the problem to be studied. An analysis of small variations (~ 0.02 Å) in C-C lengths [e.g. the work of Domenicano, Vaciago & Coulson (1975) on distortions of benzene rings] requires only structures with e.s.d.'s < 0.005 Å in which there is no disorder and H atom positions have been refined. In contrast a study of torsion angles in triphenylphosphine groups may well be possible with structures containing heavy atoms where rigid-group refinement was used for the benzene rings. In the nucleoside study (MM) different criteria were used for analysis of torsion angles and bond angles.

Even if the raw-data screens do not filter out all the unwanted structures we shall show in the following paper how statistical techniques may be used to identify those data sets which do not fit a general trend. These sets must be carefully examined and in many cases the original literature (*document*) must be examined to see if there are known experimental errors.

(d) *Interactive searching*

A system of inverted files and interactive search programs based on the CCDC file has been developed by Feldmann (1974) and the Rutherford Laboratory (Machin, Froud, Mills, Mills & Elder, 1977). The system (CSSR) allows very quick bibliographic, connectivity and raw-data screening and is capable of displaying structures visually. It forms a valuable

adjunct to the Cambridge programs in formulating searches but is not adapted to tabulating molecular geometry for a large number of structures. It cannot be used for stereochemical screening (except manually) and is not adapted for derived-data screening.

(e) *Molecular-geometry calculation*

When the data are retrieved, they are processed by the Cambridge geometry program (*GEOM*) to give the derived data (bond lengths, angles, etc.) for the fragment of interest. Additional calculations such as mean planes through groups of atoms, angles between interatomic vectors and similar derived quantities are also frequently useful. Large numbers of compounds can be processed automatically and the output can be restricted to the desired parameters of the molecular fragment. A standard numbering scheme of the users choice is superimposed on the fragment* so that the bonds and angles are systematically organized for comparison. Very often a structure contains two or more independent examples of the same fragment and all these are output. [The program, moreover, serves as a final check on the correctness of the searches since output will only be generated from any structure if the crystallographic connectivity can be matched. This is important since it avoids the problem of missing values which may result in a non-Gramian correlation matrix (see MB).] It is also possible to output other numeric quantities with the molecular geometry at this stage, e.g. R factors, e.s.d., category. We shall refer to the file of data output at this stage as the *molecular-geometry subfile*; it may now be subjected to stereochemical and derived-data screening before statistical analysis of the molecular geometry.

(f) *Stereochemical screening*

The coding of chemical formulae in the connectivity file is topological and gives no indication of stereochemistry. Since the atomic coordinates represent the whole geometry of the molecule they contain all the relative stereochemical information, but it is not always easy to translate this into chemical terms. There are several aspects to the problem.

(i) *Absolute configuration*. If the structure determination has taken anomalous scattering into account then the coordinates, referred to right-handed axes,† should represent the correct absolute configuration.

In many cases (e.g. light-atom structures) the configuration is not determined, and the authors may publish coordinates representing the enantiomorph of

* The numbering scheme in the file is derived from the original publications and is thus often arbitrary and variable.

† *GEOM* always uses right-handed axes for the calculation of torsion angles and other signed quantities.

Table 5. *Examples of stereochemical screens*

The selection of the molecular structure must be carried out in three stages:

- CONNSE*R or *BIBSE*R searches the CCDC file for all molecules containing the required fragment.
 - GEOM* is used to select all fragments with the correct connectivity (but with no consideration of bond type) and to calculate molecular geometry.
 - This is then input to *SPSS* and logical operations are used to select only the required isomer.
- (i) *meso*-Butane-2,3-diol derivatives

GEOM

FRAGMENT BUTANEDIOL

AT1 C 1

AT2 C 3 E

AT3 C 3 E

AT4 C 1

AT5 O 1

AT6 O 1

BO 2 5

BO 2 3

BO 3 6

BO 2 1

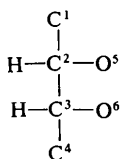
BO 3 4

END OF FRAGMENT

DEF T1234 1 2 3 4

DEF T5236 5 2 3 6

EXCL H

*SPSS*

SELECT IF (COS((T1234 - T5236)*3.1412/180) GT 0.866)

- (ii) To retrieve all *trans* isomers of azo compounds, including severely twisted conformations

GEOM

FRAGMENT AZO

AT1 C 1

AT2 N 2 E

AT3 N 2 E

AT4 C 1

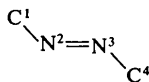
BO 3 2

BO 2 1

BO 3 4

END OF FRAGMENT

DEF T1234 1 2 3 4

*SPSS*

SELECT IF (ABS(T1234) GT 90)

- (iii) Square-planar nickel complexes

GEOM

FRAGMENT NI(X)4

AT1 NI 4 E

AT2 AA

AT3 AA

AT4 AA

AT5 AA

BO 1 2

BO 1 3

BO 1 4

BO 1 5

END OF FRAGMENT

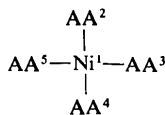
DEF A213 2 1 3

DEF A215 2 1 5

DEF A514 5 1 4

DEF A413 4 1 3

(Note: hydrogen is allowed as a possible ligand.)

*SPSS*

SELECT IF (A213 + A215 + A514 + A413 LT 370)

the correct structure. Thus in 100 studies of nucleosides, about 10% were published with enantiomorphic coordinates, and this would produce serious errors in analyses of torsion angles since the signs are changed. The method of detecting this and correcting it automatically is shown in MM.

(ii) *Relative configuration*. Where more than one chiral centre is present in a molecule, torsion angles can often be used to distinguish diastereoisomers automatically. If the two centres are directly bonded, as in *meso* and racemic tartaric acids, torsion angles about this bond differentiate the isomers [Table 5(i)]. If the two centres are not bonded, but in a ring, then torsion angles may still be useful. To distinguish between *R* and *S* configurations in more difficult situations it may be necessary to calculate for each chiral centre the sign of the product $\mathbf{r}_1 \cdot \mathbf{r}_2 \times \mathbf{r}_3$ (where $\mathbf{r}_1 = \mathbf{X}_1 - \mathbf{X}_0$ etc.) from the coordinates $(\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$ of the tetrahedron, where the numbering obeys the Cahn-Ingold-Prelog rules.

(iii) *Geometrical isomers*. *cis* and *trans* configurations cannot be specified in *CONNSE*R but are easily separated by considering the torsion angle [Table 5(ii)].

(iv) *Other types of isomerism*. For most organic molecules it will be possible to find an effective screen using torsion angles. Some types of coordination to metals will present a problem and can be coded at this point. Table 5(iii) shows how planar (as opposed to tetrahedral) four-coordinate Ni can be selected, although this could also be regarded as a derived-data screen.

(g) *Derived-data screens*

This concept has enormous potential for retrieving molecules with unusual geometry. Its use for bond lengths and angles is relatively straightforward [Table 6(i),(ii),(iii)]. For distortions which are more complicated, the use of symmetry coordinates [for a full discussion see Murray-Rust, Bürgi & Dunitz (1978*b*)] is very effective. Examples (iv) shows how ring distortions can be economically coded in this way.

The whole screening process is shown in Fig. 1. The order of the screens is somewhat variable and may depend on computer efficiency; for example, if only accurate structures were being analysed it might pay to carry out raw-data screening first and use the subfile in all subsequent work.

Computational difficulties can arise with small highly-symmetrical fragments which may occur many times in the same structure. Thus the fragment C-C-C-C will occur in almost all structures on file and many times within each structure, often partially overlapping. Because of this *GEOM* does not output overlapping fragments. The all-carbon analogue of the question in Table 6(iii) would require some modification as torsion angles would only be calculated for

Table 6. *Examples of derived-data screens (see text)*

*CONNSE*R or *BIBSE*R is used to retrieve data for all compounds that contain the required fragment. The molecular geometry is then calculated and input to *SPSS*, where further derived data [for example, see (iv)] may be calculated. The input for *GEOM*, although very similar to that for *CONNSE*R, is not usually identical, and because the structural data file contains no information on bond type additional screens on bond lengths may be necessary.

- (i) Retrieve all compounds with at least one C—S bond longer than 1.80 Å

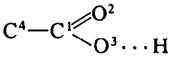
```
GEOM
FRAGMENT CS          C2-S2
AT1 S 1 }
AT2 C 1 }           Selects all compounds containing
BO 1 2 }           a C—S bond of any type or length.
END OF FRAGMENT
DEF D12 1 2
```

```
SPSS
SELECT IF (D12 GT 1.80)
```

(Note: This procedure might miss some C—S bonds in symmetrical fragments such as S—C—S or C—S—C. To retrieve *all* data these possibilities would have to be tested separately.)

- (ii) Retrieve all compounds with an asymmetric carboxylate group (*i.e.* appreciably distorted from C_{2v} symmetry)

```
GEOM
FRAGMENT CARBOXYLATE
AT1 C 3 E
AT2 O 1 E
AT3 O 1 E
AT4 C 1
BO 1 2
BO 1 3
BO 1 3
END OF FRAGMENT
EXCL H
DEF D12 1 2
DEF D13 1 3
DEF A413 4 1 3
DEF A412 4 1 2
DEF A213 2 1 3
```



```
SPSS
SELECT IF (((ABS(D12 - D13) GT 0.05) OR (ABS(A413
- A412) GT 3)) AND ((ABS(A413 + A412 + A213) GT
358))
```

[Note: Carboxylates and carboxylic acids (but not esters or coordination compounds) would be retrieved since hydrogen atoms are not always present in the structural data. The last clause of the screen is to ensure that only planar groups are considered.]

- (iii) Retrieve all acyclic C—S—S—C fragments with an eclipsed conformation

```
GEOM
FRAGMENT CSSC
AT1 C 1
AT2 S 2 E
AT3 S 2 E
AT4 C 1
BO 2 3
BO 2 1
BO 3 4
END OF FRAGMENT
DEF T1234 1 2 3 4
```

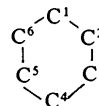
C¹-S²-S³-C⁴

```
SPSS
SELECT IF (ABS(T1234) LT 10)
```

Table 6 (cont.)

- (iv) Retrieve compounds with unfused cyclohexane rings in the boat-twist family of conformations. *CONNSE*R is first run to retrieve structures which contain one or more cyclohexane rings.

```
GEOM
FRAGMENT CYCLOHEXANE
AT1 C 2
AT2 C 2
AT3 C 2
AT4 C 2
AT5 C 2
AT6 C 2
BO 1 2
BO 2 3
BO 3 4
BO 4 5
BO 5 6
BO 6 1
END OF FRAGMENT
DEF D12 1 2
```



```

:
DEF T6123 6 1 2 3
```

```
SPSS
COMPUTE E2UA = (2*T1234 + 2*T4561 - T2345 - T3456
- T5612 - T6123)/(SQRT(12))
COMPUTE E2UB = (T2345 - T3456 + T5612 - T6123)/2
COMPUTE E2U = SQRT (E2UA**2 + E2UB**2)
COMPUTE B2G = (T1234 - T2345 + T3456 - T4561
+ T5612 - T6123)/(SQRT(6))
IF (1.45 LT D12 AND D23 AND D34 AND D45
AND D56 AND D61)CYCLOH = 1
SELECT IF (E2U GT 10 AND B2G LT 10 AND CYCLOH
EQ 1)
```

[Note: The boat-twist family have a non-zero value for the E_{2u} (out-of-plane displacement) symmetry coordinate and a small or zero value for B_{2g} (the chair-like distortion), referred to a reference point group D_{6h} . Some of the structures retrieved by *CONNSE*R may contain benzene or cyclohexene rings as well as cyclohexane. To make sure that screening is only carried out on the latter, bond lengths must be examined.]

some of the C—C—C—C fragments imbedded in a long hydrocarbon chain; a similar problem occurs with rings.

Analysing molecular geometry

No theory is at present able to predict the accurate geometry of molecules in the crystalline state especially when they are large and easily deformed. *Ab initio* molecular-orbital theory can be used to calculate accurate geometries for small molecules in the gas phase but cannot be used for most of the compounds on the data file. Most attempts to explain or predict molecular geometry are therefore usually fairly empirical and highly parameterized. Force-fields involving interatom potential functions have been widely used for studying molecular geometry and crystal packing and have had considerable success when considering conformational problems. Changes in bond angles and

especially bond lengths are not so well catered for. Moreover, most of the force-fields are satisfactorily parameterized only for a small fraction of the type of molecules on the file, usually hydrocarbons and their derivatives. An alternative approach is the bond-valence method, based on Pauling's electrostatic principles for ionic crystals. Quite good predictions of bond lengths, and somewhat poorer ones for angles, can be made for small oxy-cations and anions if the crystal environment is known.

Since there is no general theory capable of predicting accurately the geometry of a given molecule in the crystal, empirical relationships involving geometrical parameters can be very valuable. These are found by studying a number of molecules with features in common, and there are two conceptually different ways of proceeding. The most usual is to postulate a model which suggests a relationship (often in an analytical form) between two or more parameters. Examples involving molecular geometry include: bond length/bond order in delocalized systems; bond length/vibrational frequency; bond angle/NMR chemical shift; bond angle/Mössbauer chemical shift; UV absorption/torsion angle in conjugated systems; bond length/electronegativity; bond angle/electronegativity; and vicinal (^1H - ^1H) coupling constant/torsion angle (Karplus, 1959, 1963). The last example has a particularly well-defined analytical form. Very important are the simple theories of bonding which relate bond lengths to angles; the molecular-orbital, valence-bond and VSEPR theories have all suggested relationships between geometrical parameters at an empirical level.

These relationships usually express one variable as a function of one or more other variables. The parameters in the equations can be determined by linear or non-linear regression methods. The progress of such an analysis follows the sequence of the procedures on the left-hand side of Fig. 2, where the regression produces coefficients which quantify the ideas in the model. Regression techniques are most convincing where there is a high correlation between the independent and dependent variables and residual scores are small. If there are significant effects not included in the original model they will produce scatter and although the regression procedure is still valid the correlation coefficients will be smaller. In favourable cases it may be possible, by examining residual scores as described below, to refine the model and repeat the procedure but this is not commonly done.

A different way of analysing the data, particularly if there are no models and no obvious single linear relationship, is based on *multivariate analysis*. The study of reaction pathways has shown that strong inter-relationships exist between geometrical parameters but that often there are two or more distinct effects operating (Murray-Rust, Bürgi & Dunitz, 1978b). In

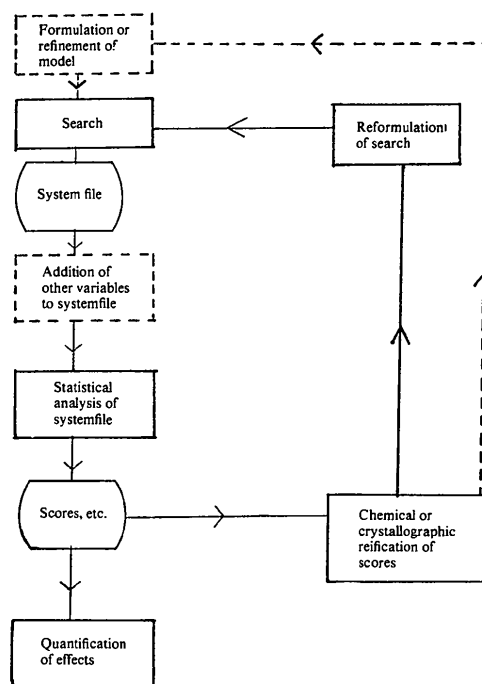


Fig. 2. Schematic relationship of the operations in the statistical analysis of molecular geometry. The procedures and pathways in broken lines need not always be involved. Simple regression involves the procedures on the left-hand side. 'System file' refers to the *SPSS* system file, whose initial creation is described in Fig. 1.

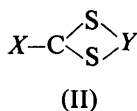
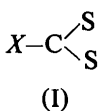
simple cases models may sometimes predict the form of the relationships but in large molecules with many parameters the multi-dimensional function may be somewhat complex. This type of problem is common in the biological and social sciences where models are often lacking and statistical methods have been developed for classification and the discovery of relationships. The analysis of a large file of molecular geometry is similar to the treatment of socio-economic data from censuses or central statistics.

We shall not describe the techniques (bivariate correlation, factor analysis, discriminant analysis and cluster analysis) here, but it is important to realize that they merely involve precise (algorithmic) transformations of the original data. Thus factor analysis (described in MB and MM) finds the linear combinations of parameters which best account for the variation in these parameters. Discriminant analysis could be used to give the best (linear) formula for distinguishing between two groups of compounds. Hopefully, however, the transformed quantities (*scores*) may be subsequently linked to some chemical or crystallographic phenomenon, a process which we call *reification*, following the social-scientific usage. This *chemical* concept may then be used to reformulate the search so that the effect is specifically excluded or included. In this way a number of independent effects

linked to chemical or crystallographic ideas may be discovered for a group of compounds. When this can be done progress will be made towards an empirical description of molecular geometry.

The schematic progress of such an analysis is shown in Fig. 2 and we shall illustrate it by a hypothetical unidimensional example. A study of C—C—C angles will show a wide range (at least 55–130°). We can compute the mean and standard deviation and express each case by the deviation from the mean (*z*-scores, see MB). Very small angles will have large negative scores whereas large angles will have large positive scores. Manual examination of the compounds with negative scores will show them all to have three- and four-membered rings (a somewhat trivial result) whereas positive scores will be associated with unsaturation of the C—C—C system. The screens could now be revised to exclude all cyclopropane and cyclobutane derivatives and to include only saturated carbon. The new *z*-scores would now be computed and high absolute values could be examined. Much of the variation would be seen to be due to experimental error and the raw-data screen on accuracy would have to be made more selective. A third iteration might now reveal some second-order effects; for example, large angles would be found in some medium-sized rings or where there were bulky groups (steric hindrance). We stress that caution is necessary before attributing effects to chemical causes since crystallographic errors may produce a systematic effect. For example, many bond lengths in saturated carbon chains and rings are often shorter than expected because of thermal motion.

A multivariate study based on data from the file is the factor analysis of the geometry of XCS_2 groups (I) in 324 cases (Murray-Rust, 1976).



Three factors were found, the first two of which were interpreted as distortion pathways consistent with VSEPR theory. The third factor was not obviously reifiable, but graphical plotting of the structures showed that large negative scores all corresponded to structures of type (II). This factor thus described the effect of chelation on narrowing the angle SCS, and similar chemical explanations were found for the positive scores.

We have shown in outline how relationships can be unearthed from the huge amount of data on the file. Before any results are accepted as physical effects, however, the question of random and systematic errors must be carefully considered, and this is the subject of the following paper (MB). The third paper (MM) shows

in detail the progress of an analysis for a fragment of biological molecules.

The Cambridge programs described here were written by Sam Motherwell in collaboration with the other staff of the CCDC. We are grateful to the British Science Research Council and to the affiliated data centres of the CCDC in other countries for financial support. We also thank the Cambridge University Computing Service for the use of the IBM 370/165 computer.

References

- ALLEN, F. H., ISAACS, N. W., KENNARD, O., MOTHERWELL, W. D. S., PETERSEN, R. C., TOWN, W. G. & WATSON, D. G. (1973). *J. Chem. Doc.* **13**, 211–218.
- ALLEN, F. H., KENNARD, O., MOTHERWELL, W. D. S., TOWN, W. G. & WATSON, D. G. (1973). *J. Chem. Doc.* **13**, 118–123.
- ALLEN, F. H., KENNARD, O., MOTHERWELL, W. D. S., TOWN, W. G., WATSON, D. G., SCOTT, T. J. & LARSON, A. C. (1974). *J. Appl. Cryst.* **7**, 73–78.
- BAUR, W. H. (1974). *Acta Cryst.* **B30**, 1195–1215.
- BÜRGI, H. B., DUNITZ, J. D. & SHEFTER, E. (1973). *J. Am. Chem. Soc.* **95**, 5065–5066.
- Cambridge Crystallographic Data Centre (1976). User Manual. Univ. of Cambridge, England.
- DOMENICANO, A., VACIAGO, A. & COULSON, C. A. (1975). *Acta Cryst.* **B31**, 221–234.
- FELDMANN, R. J. (1974). *Computer Representation and Manipulation of Chemical Information*, pp. 55–81. New York: John Wiley.
- KARPLUS, M. (1959). *J. Chem. Phys.* **30**, 11–15.
- KARPLUS, M. (1963). *J. Am. Chem. Soc.* **85**, 2870–2871.
- KENNARD, O., WATSON, D. G., ALLEN, F. H., MOTHERWELL, W. D. S., TOWN, W. G. & RODGERS, J. (1975). *Chem. Br.* **11**, 213–216.
- KENNARD, O., WATSON, D. G. & TOWN, W. G. (1972). *J. Chem. Doc.* **12**, 14–19.
- MACHIN, P. A., FROUD, D., MILLS, J. N., MILLS, O. S. & ELDER, M. (1977). *CSSR—Crystal Structure Search Retrieval*. Daresbury Laboratory, Science Research Council.
- MOTHERWELL, W. D. S. (1976). *Crystallographic Computing Techniques*, pp. 481–487. Copenhagen: Munksgaard.
- MURRAY-RUST, P. (1976). Abstracts. Third European Crystallographic Meeting, Zurich, p. 206.
- MURRAY-RUST, P. & BLAND, R. (1978). *Acta Cryst.* **B34**, 2527–2533.
- MURRAY-RUST, P., BÜRGI, H. B. & DUNITZ, J. D. (1975). *J. Am. Chem. Soc.* **97**, 921–922.
- MURRAY-RUST, P., BÜRGI, H. B. & DUNITZ, J. D. (1978a). *Acta Cryst.* **B34**, 1787–1793.
- MURRAY-RUST, P., BÜRGI, H. B. & DUNITZ, J. D. (1978b). *Acta Cryst.* **B34**, 1793–1803.
- MURRAY-RUST, P. & MOTHERWELL, S. (1978). *Acta Cryst.* **B34**, 2534–2546.
- NIE, N. H., HULL, C. H., JENKINS, J. G., STEINBRENNER, K. & BENT, D. H. (1975). *Statistical Package for the Social Sciences*, 2nd edition. New York, London: McGraw-Hill.