The SHELX-97 Manual

Contents

The refinement program **SHELXL** includes many new features to make it easier to use for macromolecules, even at moderate resolution (say better than 2.5Å). It also incoporates a large number of small improvements suggested by small-molecule users of SHELXL-93.

In view of the fact that users were encouraged to adapt the 1993 version of **CIFTAB**, which produces tables from the CIF format files generated by SHELXL, only minor corrections have been made to this program.

An anonymous user has kindly donated the program **SHELXA** that can be used to make an 'absorption correction' by fitting the observed to the calculated intensities (like DIFABS). This is intended for emergency use only, e.g. when it is impossible to apply proper absorption corrections because the world's only crystal has been lost before measurements of crystal faces or azimuthal scans could be made. It would be quite unethical to submit a structure processed in this way for publication, and the anonymous donor does not wish to be cited in this non-existent publication since it would ruin his scientific reputation!

A new feature in SHELX-97 is an interactive interface program **SHELXPRO** that is specific to protein applications; SHELXS and SHELXL are very general and in no way specific to certain types of crystal structure. SHELXPRO handles problems of communication with other widely used protein programs; for example it can convert PDB to SHELX format, adding appropriate restraints etc., and can generate sigma-A

quicker than reading the manual, but all emails asking the questions in Chapter 18 (Frequently Asked Questions) will be ignored! The programs are provided on the understanding that the author is in no way liable for any consequences of errors in the programs or their documentation.

WPDB and LIST 6 should be used. The program

2.2 The *.ins* instruction file

All instructions commence with a four (or fewer) character word (which may be an atom name); numbers and other information follow in free format, separated by one or more spaces. Upper and lower case input may be freely mixed; with the exception of the text string input using TITL, the input is converted to upper case for internal use in SHELXL. The TITL, CELL, ZERR, LATT (if required), SYMM (if required), SFAC, DISP (if required) and UNIT instructions must be given in that order; all remaining instructions, atoms, etc. should come between UNIT and the last instruction, which is always HKLF (to read in reflection data).

A number of instructions allow atom names to be referenced; use of such instructions without any atom names means 'all non-hydrogen atoms' (in the current residue, if one has been defined). A list of atom names may also be abbreviated to the first atom, the symbol '>' (separated by spaces), and then the last atom; this means 'all atoms between and including the two named atoms but excluding hydrogens'.

2.3 The reflection data file *name.hkl*

The *.hkl* file consists of one line per reflection in FORMAT(3I4,2F8.2,I4) for h,k,l,F_o^2, $\sigma(F_o^2)$, and (optionally) a batch number. This file should be terminated by a record with all items zero; individual data sets within the file should NOT be separated from one another - the batch numbers serve to distinguish between groups of reflections for which separate scale factors are to be refined (see the BASF instruction). The reflection order and the batch number order are unimportant. This '*.hkl*' file is read each time the program is run; unlike SHELX-76, there is no facility for intermediate storage of binary data. This enhances computer independence and eliminates several possible sources of confusion. The *.hkl* file is read when the HKLF instruction (which terminates the *.ins* file) is encountered. The HKLF instruction specifies the format of the *.hkl* file, and allows scale factors and a reorientation matrix to be applied. Lorentz, polarization and absorption corrections are assumed to have been applied to the data in the *.hkl* file. Note that there are special extensions to the *.hkl* format for Laue and powder data, as well as for twinned crystals that cannot be handled by a TWIN instruction alone.

In general the *.hkl* file should contain all measured reflections without rejection of systematic absences or merging of equivalents. The systematic absences and R_{int} for equivalents provide an excellent check on the space group assignment and consistency of the input data. Since complex scattering factors are used throughout by SHELXL, Friedel opposites should normally not be averaged in preparing this file; an exception can be made for macromolecules without significant anomalous scatterers. Note that SHELXS always merges Friedel opposites.

2.4 Refinement against F^2

SHELXL always refines against F^2, even when F-values are input. Refinement against ALL F^2-values is demonstrably superior to refinement against F-values greater than some threshold [say $4\sigma(F)$]. More experimental information is incorporated (suitably weighted) and the chance of getting stuck in a local minimum is reduced. In pseudo-symmetry cases it is

very often the weak reflections that can discriminate between alternative potential solutions. It is difficult to refine against ALL *F*-values because of the difficulty of estimating $\sigma(F)$ from $\sigma(F^2)$ when F^2 is zero or (as a result of experimental error) negative.

The diffraction experiment measures intensities and their standard deviations, which after the various corrections give F_o^2 and $\sigma(F_o^2)$. If your data reduction program only outputs F_o and $\sigma(F_o)$, you should correct your data reduction program, not simply write a routine to square the F_o values ! It is also legal to use HKLF 3 to input F_o and $\sigma(F_o)$ to SHELXL. Note that if an F_o

For some refinements of twinned crystals, and for least-squares refinement of batch scale factors, it is necessary to suppress the merging of equivalent reflections with MERG 0.

2.6 Least-squares refinement

Small molecules are almost always refined by full-matrix methods (using the L.S. instruction in SHELXL), which give the best convergence per cycle, and allows esd's to be estimated. The CPU time per cycle required for full-matrix refinement is approximately proportional to the number of reflections times the square of the number of parameters; this is prohibitive for all but the smallest macromolecules. In addition the (single precision) matrix inversion suffers from accumulated rounding errors when the number of parameters becomes very large. An excellent alternative for macromolecules is the conjugate-gradient solution of the normal equations, taking into account only those off-diagonal terms that involve restraints. This method was employed by Konnert & Hendrickson (1980) in the program PROLSQ; except for modifications to accelerate the convergence, exactly the same algorithm is used in SHELXL (instruction CGLS). The CGLS refinement can be also usefully employed in the early stages of refinement of medium and large 'small molecules'; it requires more cycles for convergence, but is fast and robust. The major disadvantage of CGLS is that it does not give esds.

For both L.S. and CGLS options, it is possible to block the refinement so that a different combination of parameters is refined each cycle. For example after a large structure has been refined using CGLS (without BLOC), a final job may be run with L.S. 1, DAMP 0 0 and BLOC 1 (or e.g. BLOC N_1 > LAST for a protein) to obtain esds on all geometric parameters; the anisotropic displacement parameters are held fixed, reducing the number of parameters by a factor of three and the cycle time by an order of magnitude.

2.7 *R*-indices and weights

One cosmetic disadvantage of refinement against F^2 is that *R*-indices based on F^2 are larger than (more than double) those based on *F*. For comparison with older refinements based on *F*

analysis of variance and a GooF close to unity [there was a bug in SHELXL-93 that can occasionally cause the program to abort when trying to estimate the new weighting parameters, though it appeared to happen only with poor quality data or the wrong solution]. If the weights are varied too soon, the convergence may be impaired, because features such as missing atoms are 'weighted down'. For macromolecules it may be advisable to leave the weights at the default settings; and to accept a GooF greater than one as an admission of inadequacies in the model.

When not more than two WGHT parameters are specified, the weighting scheme simplifies to:

$$w = 1 / [\sigma^2(F_o^2) + (aP)^2 + bP]$$

where P is $[2F_c^2 + \mathrm{Max}(F_o^2,0)] / 3$. The use of this combination of F_o^2 and F_c^2 was shown by Wilson (1976) to reduce statistical bias.

It may be desirable to use a scheme that does not give a flat analysis of variance to emphasize particular features in the refinement, for example by weighting up the high angle data to remove bias caused by bonding electron density (Dunitz & Seiler, 1973).

2.8 Fourier syntheses

Fourier syntheses are summarized in the form of peak-lists (which can be edited and re-input for the next refinement job), or as 'lineprinter plots' with an analysis of non-bonded interactions

2.10 Tables

For small structures, bond lengths and angles for the full connectivity array may be tabulated with BOND, and all possible torsion angles with CONF. Although hydrogen atoms are not normally included in the connectivity array, they may be included in the bond lengths and angles tables by BOND $H. Alternatively HTAB produces a convenient way of analysing hydrogen bonds. It is also possible to be selective by naming specific atoms on the BOND and CONF instructions, or by using the RTAB instruction (which was designed with macromolecules in mind). Least-squares planes and distances of (other) atoms from these planes may be generated with MPLA. Symmetry equivalent atoms may be specified on any of these instructions by reference to EQIV symmetry operators. All esds output by SHELXL take the unit-cell esds into account and are calculated using the full covariance matrix. The only exception is the esd in the angle between two least-squares planes, for which an approximate treatment is used. Note that dag112Se(t of)-2ement (see above) leads to underestimates of the esds; in difficult cases a may be performed with DAMP 0 0 (no dag112S, but no shifts applied) to obtain good esds.

The HTAB instruction has been introduced in SHELXL-97 to analyze the hydrogen bonding in the structure. A search is made over all *hydrogen atoms* to
This is a convenient way of t o symmetry operations necessary for the second form of HTAB instructions (needed to obtain esds and CIF output), and also reveals potential misplaced hydrogens, e.g. because they do not make any hydrogen bonds, or because the automatic plac12Seof hydrogen atoms has assigned the hydrogenseof two different O-H or N-H groups to the same hydrogen bond. In the second form of the HTAB instruction, HTAB is followed by the names of the donor atom D and the acceptor atom A; for the latter a symmetry operation may also be specified. The program then most suitable hydrogen atom to

.lst *.cif*f ACTA is present).

3. Examples of Small Molecule Refinements with SHELXL

It is possible to set up special position constraints on the x,y,z-coordinates, occupation factors, and U_{ij} components by hand. However this is totally unnecessary because the program will do this automatically for any special position in any space group, conventional or otherwise. Similarly the program recognizes polar space groups ($P\bar{4}$ is non-polar) and applies appropriate restraints (Flack & Schwarzenbach, 1988), so it is no longer necessary to worry about fixing one or more coordinates to prevent the structure drifting along polar axes. It is not necessary to set the overall scale factor using an FVAR instruction for this initial job, because the program will itself estimate a suitable starting value. Comments may be included in the *.ins* file either as REM instructions or as the rest of a line following '!'; this latter facility has been used to annotate this example.

```
TITL AGS4 in P-4                          ! title of up to 76 characters
CELL 0.71073 8.381 8.381 6.661 90 90 90   ! wavelength and unit-cell
ZERR 1 .002 .002 .001 0 0 0               ! Z (formula-units/cell), cell esd's
LATT -1                                ! non-centrosymmetric primitive lattice
SYMM -X, -Y, Z
SYMM Y, -X, -Z              ! symmetry operators (x,y,z must be left out)
SYMM -Y, X, -Z
SFAC C AG AS F N S          ! define scattering factor numbers
UNIT 4 1 1 6 4 8            ! unit cell contents in same order
L.S. 10                     ! 10 cycles full-matrix least-squares
ACTA                        ! CIF-output, bonds, Fourier, peak search
OMIT -2 3 1                 ! suppress bad reflection
ANIS                        ! convert all (non-H) atoms to anisotropic
WGHT 0.037 0.31             ! weighting scheme
AG  2  .000  .000  .000
AS  3  .500  .500  .000
```

```
S2 - S2_$1 S1
C - N S1
N - C Ag
F1 - As
F2 - As

Operators for generating equivalent atoms:

$1   -x+1, -y+1, z
$2   -x, -y, z
$3   y, -x, -z
$4   -y, x, -z
$5   -x+1, -y+1, z
$6   y, -x+1, -z
$7   -y+1, x, -z
```

Note that in addition to symmetry operations generated by the program, one can also define operations with the EQIV instruction and then refer to the corresponding atoms with _$n in the same way. Thus:

```
EQIV $1 1-x, -y, z
CONF S1 S2 S2_$1 S1_$1
```

could have been included in *ags4.ins* to calculate the S-S-S-S torsion angle. If EQIV instructions are used, the program renumbers the other symmetry operators accordingly.

The next part of the output is concerned with the data reduction:

```
 1475  Reflections read, of which     1  rejected

0 =< h =< 10,     -9 =< k =< 10,      0 =< l =<  8,   Max. 2-theta =   55.00

    0  Systematic absence violations

Inconsistent equivalents etc.

 h   k   l       Fo^2    Sigma(Fo^2)  Esd of mean(Fo^2)

 3   4   0      387.25       8.54       47.78

     1  Inconsistent equivalents
   903  Unique reflections, of which      0  suppressed

R(int) = 0.0165     R(sigma) = 0.0202      Friedel opposites not merged
```

Special position constraints are then generated and the statistics from the first least-squares cycle are listed (the output has been compacted to fit the page). The maximum vector length refers to the number of reflections processed simultaneously in the rate-determining calculations; usually the program utilizes all available memory to make this as large as possible, subject to a maximum of 511. This maximum may be reduced (but not increased) by means of the fourth parameter on the L.S. (or CGLS) instruction; this may be required to prevent unnecessary disk transfers when large structures are refined on virtual memory systems with limited physical memory. The number of parameters refined in the current cycle is followed by the total number of refinable parameters (here both are 55).

```
Special position constraints for Ag
x =  0.0000       y =  0.0000         z =  0.0000         U22 = 1.0 * U11
U23 = 0           U13 = 0             U12 = 0             sof = 0.25000

Special position constraints for As
x =  0.5000       y =  0.5000         z =  0.0000         U22 = 1.0 * U11
U23 = 0           U13 = 0             U12 = 0             sof = 0.25000

Special position constraints for F2
x =  0.5000       y =  0.5000         U23 = 0             U13 = 0
sof = 0.50000


Least-squares cycle 1   Maximum vector length=511  Memory required=1092/82899

wR2 =  0.5042 before cycle   1 for    903 data and   55 /   55 parameters

GooF = S =  8.127;     Restrained GooF =      8.127  for      0 restraints

Weight = 1/[sigma^2(Fo^2)+(0.0370*P)^2+0.31*P] where P=(Max(Fo^2,0)+2*Fc^2)/3

** Shifts scaled down to reduce maximum shift/esd from   17.64  to   15.00 **

  N     value       esd        shift/esd  parameter

  1     2.31065     0.04324     9.042     OSF
  2     0.07314     0.00206    11.250     U11 Ag
 11     0.07309     0.00669     3.453     U33 S1
 47     0.11304     0.01391     4.533     U33 F1

Mean shift/esd =   1.238    Maximum =  11.250 for  OSF

Max. shift = 0.045 A for C      Max. dU = 0.033 for F2
```

Only the largest shift/esd's are printed. More output could have been obtained using 'MORE 2' or 'MORE 3'. The largest correlation matrix elements are printed after the last cycle, in which the mean and maximum shift/esd have been reduced to 0.003 and 0.017 respectively. This is followed by the full table of refined coordinates and U_{ij}'s with esd's (too large to include here, but similar to the corresponding table in SHELX-76 except that U_{eq} and its esd are also printed) and by a final structure factor calculation:

```
Final Structure Factor Calculation for  AGS4 in P-4

Total number of l.s. parameters = 55  Maximum vector length = 511
wR2 =  0.0780 before cycle  11 for    903 data and    2 /   55 parameters

GooF = S =     1.063;     Restrained GooF =      1.063  for      0 restraints
Weight = 1/[sigma^2(Fo^2)+(0.0370*P)^2+0.31*P] where P=(Max(Fo^2,0)+2*Fc^2)/3
R1 =  0.0322 for    818 Fo > 4.sigma(Fo)  and  0.0367 for all    903 data
wR2 =  0.0780,  GooF = S =   1.063,  Restrained GooF =    1.063  for all data

Flack x parameter = 0.0224   with esd  0.0260   (expected values are 0
(within 3 esd's) for correct and +1 for inverted absolute structure)
```

There are some important points to note here. The weighted *R*-index based on F_o^2 is (for compelling statistical reasons) much higher than the conventional *R*-index based on F_o with a threshold of say $F_o > 4\sigma(F_o)$. For comparison with structures refined against *F* the latter is therefore printed as well (as *R*1). Despite the fact that *wR*2 and not *R*1 is the quantity minimized, *R*1 has the advantage that it is relatively insensitive to the weighting scheme, and so is more difficult to manipulate.

Since the structure is non-centrosymmetric, the program has automatically estimated the Flack absolute structure parameter x in the final structure factor summation. In this example x is within one esd of zero, and its esd is also relatively small. This provides strong evidence that the absolute structure has been assigned correctly, so that no further action is required. The program would have printed a warning here if it would have been necessary to 'invert' the structure or to refine it as a racemic twin.

.

This is followed by a list of principal mean square displacements U for all anisotropic atoms. It will be seen that none of the smallest components (in the third column) are in danger of going negative [which would make the atom 'non positive definite' (NPD)] but that the motion of the two unique fluorine atoms is highly anisotropic (not unusual for an AsF_6 anion). The program suggests that the fluorine motion is so extended in one direction that it would be possible to represent each of the two fluorine atoms as disordered over two sites, for which x, y and z coordinates are given; this may safely be ignored here (although there may well be some truth in it). The two suggested new positions for each 'split' atom are placed equidistant from the

```
Resolution(A)  0.77  0.81  0.85  0.90  0.95  1.02  1.10  1.22  1.40  1.74  inf
Number in group    97.   84.   92.   91.   89.   90.   89.   90.   93.   88.
GooF         1.067 0.959 0.935 0.895 1.035 1.040 1.115 1.149 1.161 1.228
K            1.047 1.010 1.009 0.991 1.004 0.996 0.989 1.012 0.997 0.982
R1           0.166 0.100 0.069 0.059 0.051 0.036 0.033 0.027 0.020 0.020

Recommended weighting scheme:  WGHT   0.0314 0.3674


Most Disagreeable Reflections (* if suppressed or used for Rfree)

    h   k   l       Fo^2      Fc^2    Delta(F^2)/esd  Fc/F(max)  Resolution(A)
    4   4   4       18.32    33.30      3.62          0.062       1.11
   -4   1   3       15.79     4.17      3.50          0.022       1.50
    0   2   2       41.60    57.32      3.26          0.082       2.61      etc.
```

After the table of bond lengths and angles (BOND was implied by the ACTA instruction), the data are merged (again) for the Fourier calculation after correcting for dispersion (because the

```
S2 - Distance  Angles
S2_$1 2.0114(0.0028)
S1    2.0633(0.0025) 105.37(0.07)
         S2 -        S2_$1

C -  Distance  Angles
N   1.1472(0.0074)
S1  1.6819(0.0069) 175.67(0.49)
       C -         N

N -  Distance  Angles
C   1.1472(0.0074)
Ag  2.2788(0.0058) 152.38(0.45)
       N -         C
F1 - Distance  Angles
As  1.6724(0.0037)
       F1 -

F2 - Distance  Angles
As  1.6399(0.0075)
       F2 -


FMAP and GRID set by program

FMAP 2   3  18
GRID  -3.333  -2  -1     3.333   2   1

R1 = 0.0370 for 590 unique reflections after merging for Fourier


Electron density synthesis with coefficients Fo-Fc

Highest peak   0.32  at  0.0000  0.0000  0.5000  [2.60 A from N]
Deepest hole  -0.36  at  0.5000  0.5000  0.1863  [0.40 A from F2]
Mean = 0.00, Rms deviation from mean = 0.07 e/A^3 Highest memory used 1133/13851


Fourier peaks appended to .res file

         x       y       z       sof      U   Peak  Dist to nearest atoms
Q1  1  0.0000  0.0000  0.5000  0.25000  0.05  0.32  2.60 N  2.69 C  3.33 AG
Q2  1  0.5690  0.3728  0.1623  1.00000  0.05  0.27  1.20 F1 1.34 F2 1.62 AS
Q3  1  0.5685  0.3851 -0.1621  1.00000  0.05  0.24  1.19 F1 1.25 F2 1.56 AS
Q4  1  0.4075  0.4717  0.2378  1.00000  0.05  0.23  0.81 F2 1.78 AS 1.79 F1
Q5  1  0.5848  0.2667  0.0312  1.00000  0.05  0.23  0.55 F1 2.09 AS 2.47 F1
Q6  1  0.5495  0.3425 -0.1122  1.00000  0.05  0.21  0.83 F1 1.57 AS 1.65 F2
Q7  1  0.2617 -0.1441  0.1446  1.00000  0.05  0.20  1.59 N  2.17 F1 2.40 C
Q8  1  0.7221  0.1898  0.0030  1.00000  0.05  0.20  1.55 F1 2.39 N  2.54 N
Q9  1  0.1997  0.0293  0.1024  1.00000  0.05  0.19  0.75 N  1.79 C  1.82 AG
Q10 1  0.4606 -0.0113  0.8165  1.00000  0.05  0.19  0.91 S2 1.41 S2 2.82 S1
```

```
    0.01862  -0.00372  -0.00330  -0.01185

HKLF 4   ! read intensity data from 'sigi.hkl'; terminates '.ins' file
END
```

The data reduction reports 1904 reflections read (one of which was rejected by OMIT) with indices $-7 \le h \le 7$, $-8 \le k \le 9$ and $-9 \le l \le 9$. Note that these are the limiting index values; in fact only about 1.5 times the unique volume of reciprocal space was measured. The maximum 2θ was 50.00, and there were no systematic absence violations, 34 (not seriously) inconsistent equivalents, and 1296 unique data. R(int) was 0.0196 and R(sigma) 0.0151.

The program uses different default distances to hydrogen for different bonding situations; these may be overridden by the user if desired. These defaults depend on the temperature (set using TEMP) in order to allow for librational effects. The list of default X-H distances is followed by the (squashed) circular difference electron density syntheses to determine the C-OH and C-CH_3 initial torsion angles:

```
Default effective X-H distances for T =   20.0 C

AFIX m =    1    2    3     4   4[N]  3[N]  15[B]  8[O]   9   9[N]   16
d(X-H) =  0.98 0.97 0.96  0.93  0.86  0.89  1.10  0.82  0.93  0.86  0.93

Difference electron density (eA^-3x100) at 15 degree intervals for AFIX 147 group
attached to O2. The center of the range is eclipsed (cis) to C7 and rotation is
clockwise looking down C5 to O2
  2 -2 -6 -9 -8 -5 -1  0  0  0  1  0 -2 -2  0  9 23 39 48 42 29 16  9  5

Difference electron density (eA^-3x100) at 15 degree intervals for AFIX 137 group
attached to C10. The center of the range is eclipsed (cis) to N6 and rotation is
clockwise looking down C9 to C10
  50 47 39 28 19 15 20 30 38 41 39 37 34 29 25 27 33 35 29 19 12 15 29 43

After local symmetry averaging:   40  41  36  28  21  20  24  33
```

It will be seen that the hydroxyl hydrogen is very clearly defined, but that the methyl group is rotating fairly freely (low potential barrier). After three-fold averaging, however, there is a single difference electron density maximum. The (squashed) least-squares refinement output

```
   4    0.01781    0.00946     1.777    EXTI

Mean shift/esd =   0.747    Maximum = -10.702 for FVAR  2

Max. shift = 0.028 A for H10A     Max. dU =-0.020 for H5A

  .......... etc (cycles 2 and 3 omitted) .........

Least-squares cycle 4  Maximum vector length = 511 Memory required =1836/136080

wR2 =  0.1035 before cycle   4 for   1296 data and  105 /  105 parameters

GooF = S =     1.016;     Restrained GooF =      1.016  for      0 restraints

Weight = 1/[sigma^2(Fo^2)+(0.0600*P)^2+0.15*P] where P=(Max(Fo^2,0)+2*Fc^2)/3

   N     value       esd     shift/esd  parameter
   1    0.97902    0.00358    -0.003     OSF
   2    0.03605    0.00176     0.012     FVAR  2
   3    0.07345    0.00376    -0.031     FVAR  3
   4    0.02502    0.01081    -0.010     EXTI

Mean shift/esd =   0.008    Maximum =  -0.244 for tors H10A

Max. shift = 0.004 A for H10A       Max. dU = 0.000 for H2


Largest correlation matrix elements

 0.509 U12 O2 / U22 O2    0.507 U12 O3 / U11 O3
 0.509 U12 O2 / U11 O2    0.500 U12 O3 / U22 O3


Idealized hydrogen atom generation before cycle   5

Name     x       y       z    AFIX  d(X-H)  shift  Bonded   Conformation
                                                    to      determined by
H2   -0.6017  0.2095  0.8832  147   0.820   0.000   O2        C5  H2
H5A  -0.2721  0.0676  0.9001   23   0.970   0.000   C5        O2  C7
H5B  -0.2964  0.1554  1.0576   23   0.970   0.000   C5        O2  C7
H6A   0.3572  0.1389  0.4085   93   0.860   0.000   N6        C9  C8
H6B   0.3073  0.1559  0.2347   93   0.860   0.000   N6        C9  C8
H7   -0.3331  0.4598  0.8575   13   0.980   0.000   C7        O3  C5  C11
H10A -0.0176  0.2947  0.1525  137   0.960   0.000   C10       C9  H10A
H10B -0.2042  0.4192  0.2692  137   0.960   0.000   C10       C9  H10A
H10C -0.1764  0.2036  0.2964  137   0.960   0.000   C10       C9  H10A
H11A -0.3575  0.2948  0.6198   23   0.970   0.000   C11       C8  C7
H11B -0.3198  0.4943  0.5737   23   0.970   0.000   C11       C8  C7
```

The final structure factor calculation, analysis of variance etc. produces the following edited output:

```
Final Structure Factor Calculation for  SIGI in P-1
Total number of l.s. parameters = 105    Maximum vector length =  511
```

wR2 = 0.1035 before cycle 5 for 1296 data and   0 / 105 parameters

factor is a little high for the weakest reflections in this example; this may well be a statistical artifact and may be ignored (selecting the groups on F_c will tend to make F_o^2 greater than F_c^2 for this range). The increase in the GooF at low resolution (the 1.79 to infinity range) is caused in part by systematic errors in the model such as the use of scattering factors based on spherical atoms which ignore bonding effects, and is normal for purely light-atom structures (this interpretation is confirmed by the fact that difference electron density peaks are found in the middle of bonds). In extreme cases the lowest or highest resolution ranges can be conveniently suppressed by means of the SHEL instruction; this is normal practice in macromolecular refinements, but refining a diffuse solvent model with SWAT may be better, inadequate solvent modeling for macromolecules produces similar symptoms to lack of extinction refinement for small molecules.

```
 (* indicates atom used to define plane)

  2.3443 (0.0044) x + 7.4105 (0.0042) y - 0.0155 (0.0053) z = 1.9777 (0.0044)

*    -0.0743 (0.0008)  C7
*     0.0684 (0.0008)  C11
*    -0.0418 (0.0009)  C8
*    -0.0062 (0.0008)  C4
*     0.0538 (0.0008)  O3
     -0.0061 (0.0020)  O1
     -0.0980 (0.0028)  N6
     -0.0562 (0.0023)  C9
     -0.0314 (0.0030)  C10

 Rms deviation of fitted atoms =   0.0546


  2.5438 (0.0040) x + 7.3488 (0.0040) y - 0.1657 (0.0042) z = 1.8626 (0.0026)

 Angle to previous plane (with approximate esd) =  2.45 ( 0.07 )

*     0.0054 (0.0008)  C10
*     0.0082 (0.0008)  N6
*    -0.0052 (0.0012)  C9
*    -0.0337 (0.0012)  C8
*     0.0135 (0.0008)  C11
*     0.0118 (0.0009)  C4
      0.0568 (0.0019)  O1
      0.0214 (0.0018)  O3
     -0.1542 (0.0020)  C7

 Rms deviation of fitted atoms =   0.0162


 Hydrogen bonds with  H..A < r(A) + 2.000 Angstroms  and  <DHA > 110 deg.

 D-H           d(D-H)   d(H..A)   <DHA    d(D..A)   A
 O2-H2          0.820    2.041   174.05    2.858    O1 [ x-1, y, z ]
 N6-H6A         0.860    2.225   129.29    2.849    O1
 N6-H6B         0.860    2.172   155.06    2.974    O2 [ x+1, y, z-1 ]
```

All esds printed by the program are calculated rigorously from the full covariance matrix, except for the esd in the angle between two least-squares planes, which involves some approximations. The contributions to the esds in bond lengths, angles and torsion angles also take the errors in the unit-cell parameters (as input on the ZERR instruction) rigorously into account; an approximate treatment is used to obtain the (rather small) contributions of the cell errors to the esds involving least-squares planes.

There follows the difference electron density synthesis and line printer 'plot' of the structure and peaks. The highest and lowest features are 0.27 and -0.17 eA^{-3} respectively, and the rms difference electron density is 0.04. These values confirm that the treatment of the hydrogen atoms was adequate, and are indeed typical for routine structure analysis of small organic molecules. This output is too voluminous to give here, and indeed users of the Siemens SHELXTL molecular graphics program XP will almost always suppress it by use of the default

option of a positive number on the PLAN instruction, and employ interactive graphics instead for analysis of the peak list.

4. Constraints and Hydrogen Atoms

4.1 *Constraints* versus *restraints*

In crystal structure refinement, there is an important distinction between a *constraint* and a *restraint*. A constraint is an exact mathematical condition that enables one or more least-squares variables to be expressed exactly in terms of other variables or constants, and hence eliminated. An example is the fixing of the x, y and z coordinates of an atom on an inversion center. A *restraint* takes the form of additional information that is not exact but is subject to a probability distribution; for example two chemically but not crystallographically equivalent bonds could be restrained to be approximately equal. A restraint is treated as an extra experimental observation, with an appropriate esd that determines its weight relative to the X-ray data. An excellent account of the use of constraints and restraints to control the refinement of difficult structures has been given by Watkin (1994).

Often there is a choice between constraints and restraints. For example, in a triphenylphosphine complex of a heavy element, the light atoms will be less well determined from the X-ray data than the heavy atoms. In SHELX-76 a rigid group *constraint* was often applied to the phenyl groups in such cases: the phenyl groups were treated as rigid hexagons with C-C bond lengths of 1.39 Å. This introduces a slight bias (e.g. in the P-C bond length), because the *ipso*-angle should be a little smaller than 120º. In SHELXL such rigid group constraints may still be used, but it is more realistic to apply FLAT and SADI (or SAME) *restraints* so that the phenyl groups are planar and have mm2 (C_{2v}) symmetry, subject to suitable esds. In addition, the phenyl groups may be restrained to have similar geometries to one another.

4.2 Free variables, occupancy and isotropic U-constraints

SHELXL employs the concept of *free variables* exactly as in SHELX-76. A free variable is a refinable parameter that can be used to impose a variety of additional linear constraints, e.g. to atomic coordinates, occupancies or displacement parameters. Starting values for all free variables are supplied on the FVAR instruction. Since the first FVAR parameter is the (*F*-relative) overall scale factor, there is no free variable 1. If an atom parameter is given a value greater than 15 or less than -15, it is interpreted as a reference to a free variable. A positive value ($10k+p$) is decoded as p times free variable number k [$fv(k)$], and a negative value (i.e. k and p both negative) is decoded as p times [$fv(-k)-1$]. This appears more complicated than it is in practice: for example to assign a common occupancy parameter to describe a two component disorder, the occupancies of all atoms of one component can be replaced by 21, and the occupancies of all atoms of the second component by -21, where the starting value for the occupancy is the second FVAR parameter. A further disorder, not correlated with the first, would then use free variable number 3 and codes 31 and -31 etc. If there are more than two components of a disordered atom or group, it is necessary to apply a restraint (SUMP) to the free variables used to represent the occupancies.

Free variables may be used to constrain the isotropic U-values of chemically similar hydrogen atoms to be the same; for example one could use the fourth FVAR parameter and code 41 for all methyl hydrogens (which tend to have larger U-values), and the fifth FVAR parameter and code 51 for the rest. An alternative way to constrain hydrogen isotropic displacement

parameters is to replace the U-value on the atom instruction by a code *q* between -0.5 and -5; the U-value is then calculated as |*q*| times the (equivalent) isotropic U of the last atom not treated in this way (usually the carbon or other atom on which the hydrogen rides). Typical q values are -1.5 for methyl and hydroxyl hydrogens and -1.2 for others.

4.3 Special position constraints

Constraints for the coordinates and anisotropic displacement parameters for atoms on special positions are generated automatically by the program for ALL special positions in ALL space groups, in conventional settings or otherwise. For upwards compatibility with SHELX-76, free variables may still be used for this, but it is better to leave it to the

to restrain their distances to atoms not in the same rigid group (this was not allowed in SHELX-76).

A particularly useful constraint for the refinement of hydrogen atoms is the *riding model* (

4.6 Hydrogen atom generation and refinement

It is difficult to locate hydrogen atoms accurately using X-ray data because of their low scattering power, and because the corresponding electron density is smeared out, asymmetrical, and is not centered at the position of the nucleus. In addition hydrogen atoms tend to have larger librational amplitudes than other atoms. For most purposes it is preferable to calculate the hydrogen positions according to well-established geometrical criteria and then to adopt a refinement procedure which ensures that a sensible geometry is retained. The above table summarizes the options for generating hydrogen atoms; the hydrogen coordinates are re-idealized before each cycle. The distances given in this table are the values for room temperature, they are increased by 0.01 or 0.02 Å for low temperatures (specified by the TEMP instruction) to allow for the smaller librational correction at low temperature.

4.7 Special facilities for -CH_3 and -OH groups

Methyl and hydroxyl groups are difficult to position accurately (unless neutron data are available!). If good (low-temperature) x-ray data are available, the method of choice is HFIX 137 for -CH_3 and HFIX 147 for -OH groups; in this approach, a difference electron density synthesis is calculated around the circle which represents the locus of possible hydrogen positions (for a fixed X-H distance and Y-X-H angle). The maximum electron density (in the case of a methyl group after local threefold averaging) is then taken as the starting position for the hydrogen atom(s). In subsequent refinement cycles (and in further least-squares jobs) the hydrogens are re-idealized at the start of each cycle, but the current torsion angle is retained; the torsion angles are allowed to refine whilst keeping the X-H distance and Y-X-H angle fixed (n=7). If unusually high quality data are available, AFIX 138 would allow the refinement of a common C-H distance for a methyl group but not allow the group to tilt; a variable metric rigid group refinement (AFIX 9 for the carbon followed by AFIX 135 before the first hydrogen) would allow it to tilt as well, but still retain tetrahedral H-C-H angles and equal C-H distances within the group.

If the data quality is less good, then the refinement of torsion angles may not converge very well. In such cases the hydrogens can be positioned geometrically and refined using a riding model by HFIX 33 for methyl and HFIX 83 for hydroxyl groups. This staggers the methyl groups, and -OH groups attached to saturated carbons, as well as possible; -OH groups attached to aromatic rings are tested in one of the two positions with one hydrogen in the plane. In both cases the choice of hydrogen position is then determined by best hydrogen bond (to an N, O, Cl or F atom) that can be created. For disordered methyl groups (with two sites rotated by 60 degrees from one another) HFIX 123 is recommended, possibly with refinement of the corresponding site occupation factors via a 'free variable' so that their sum is unity (e.g. 21 and -21).

The choice of a suitable (default) O-H distance is very difficult. O-H internuclear distances for isolated molecules in the gas phase are about 0.96 Å (cf. 1.10 for C-H), but the appropriate distance to use for X-ray diffraction must be appreciably shorter to allow for the displacement of the center of gravity of the electron distribution towards the oxygen atom, and also for librational effects. Although the (temperature dependent) value assumed by the program fits

bonded systems; short H...O distances are always associated with long O-H distances. If there are many such O-H groups and good quality data are available, HFIX 88 (or 148) plus SADI restraints to make all the O-H distances approximately equal (with an esd of say 0.02) is a good approach.

4.8 Further peculiarities involving hydrogen atoms

Hydrogen atoms are identified as such by their scattering factor numbers, which must correspond to a SFAC name H (or $H). The special treatment of hydrogens does not apply if they reference a different SFAC name (e.g. D !). Other elements that need to be specifically identified (e.g. so that HFIX 43 can use different default C-H and N-H distances) are defined similarly. However for the output of the PLAN instruction, hydrogen atoms are identified as those atoms with a radius of less the 0.4 Å. This is not as illogical as it may sound; the PLAN output is concerned with potential hydrogen bonds etc., not with the scattering power of an atom, and SHELXL has to handle neutron as well as X-ray data.

Hydrogen atoms may also 'ride' on atoms in rigid groups (unlike SHELX-76); for example HFIX 43 could reference carbon atoms in a rigid phenyl ring. In such a case further geometrical restraints (SADI, SAME, DFIX, FLAT) are not permitted on the hydrogen atoms; this is the only exception to the general rule that any number of restraints may be applied to any atom, whatever constraints are also being applied to it.

OMIT $H (or OMIT_* $H if residues are employed) combined with L.S. 0, FMAP 2 and PLAN -100 enables an 'omit map' to be calculated, in which the hydrogen atoms are retained but do not contribute to F_c. If a non-zero electron density appears in the 'Peak' column for a hydrogen atom in the Fourier output, then there was an actual peak in the difference electron density synthesis within 0.31 Å of the expected hydrogen position.

Sometimes it is known that the crystal contains a deuterated solvent molecule (e.g. $CDCl_3$) because it was crystallized in an n.m.r. tube. In such a case, an element 'D' may be added after 'H' on the SFAC instruction, and the appropriate numbers of H and D in the cell specified on the UNIT instruction. This enables the formula weight and density to be calculated correctly. The H and D atoms that follow in the *.ins* file should both be given the SFAC number corresponding to H, so that they are both treated as 'hydrogens' for all other purposes.

5. Restraints and Disorder

A *restraint* is incorporated in the least-squares refinement as if it were an additional experimental observation; $w(yt–y)^2$ is added to the quantity $\Sigma w(F_o^2–F_c^2)^2$ to be minimized, where a quantity y (which is a function of the least-squares parameters) is to be restrained to a target value yt, and the weight w (for either a restraint or a reflection) is $1/\sigma^2$. In the case of a reflection, σ^2 is estimated using a weighting scheme; for a restraint σ is simply the effective standard deviation. In SHELXL the restraint weights are multiplied by the mean value of $w(F_o^2–F_c^2)^2$ for the reflection data, which allows for the possibility that the reflection weights may be relative rather than absolute, and also gives the restraints more influence in the early stages of refinement (when the Goodness of Fit is invariably much greater than unity), which improves convergence. It is possible to use Brünger's R_{free} test (Brünger, 1992) to fine-tune the restraint esds. In practice the optimal restraint esds vary little with the quality and resolution of the data, and the standard values (assumed by the program if no other value is specified) are entirely adequate for routine refinements. Default values for the various classes of restraint may be also set with DEFS instructions; there may be several DEFS instructions in the same .ins file: each applies to all restraints encountered before the next DEFS instruction (or the end of the file).

5.1 Floating origin restraints

Floating origin restraints are generated automatically by the program as and when required by the method of Flack & Schwarzenbach (1988), so the user should not attempt to fix the origin in such cases by fixing the coordinates of a heavy atom. These floating origin restraints effectively fix the X-ray 'center of gravity' of the structure in the polar axis direction(s), and lead to smaller correlations than fixing a single atom in structures with no dominant heavy atom. Floating origin restraints are not required (and will not be generated by the program) when CGLS refinement is performed.

5.2 Geometrical restraints

A particularly useful restraint is to make chemically but not crystallographically equivalent distances equal (subject to a given or assumed esd) without having to invent a value for this distance (SADI). The SAME instruction can generate SADI restraints automatically, e.g. when chemically identical molecules or residues are present. This has the same effect as making equivalent bond lengths and angles but not torsion angles equal (see also section 5.5).

The FLAT instruction restrains a group of atoms to lie in a plane (but the plane is free to move and rotate); the program achieves this by treating the restraint as a sum of chiral volume restraints with zero target volumes. Thus the restraint esd has units of $Å^3$. For comparison with other methods, the r.m.s. deviation of the atoms from their restraint planes is also calculated.

DFIX and DANG restrain distances to target values. DANG was introduced so that the default sigma for 1,3-distances could be made twice that for 1,2-distances (the first DEFS parameter). The DANG restraints are applied in exactly the same way as DFIX, but are also listed separately in the restraints summary tables.

5.4 Restraints on anisotropic displacement parameters

switched off, and any number of NCS domains

dummy atoms with zero coordinates. Since the C-C distance is uncertain (there may well be an appreciable librational shortening in such a case) we refine the C_5-ring as a *variable metric rigid group*, i.e. it remains a regular pentagon but the C-C distance is free to vary. In SHELXL this may all be achieved by inserting one instruction (AFIX 59) before the five carbons and one (AFIX 0) after them:

```
AFIX 59                  ! AFIX mn with m = 5 to fit pentagon (default C-C
C1 1 .6755 .2289 .0763   ! 1.42 A) and n = 9 for v-m rigid-group refinement
C2 1 .7004 .2544 .0161
C3 1 0 0 0               ! the coordinates for C3 and C4 are obtained by the
C4 1 0 0 0               ! fit of the other 3 atoms to a regular pentagon
C5 1 .6788 .1610 .0766
AFIX 0                   ! terminates rigid group
```

Since U_{ij} values were not specified, the atoms would refine isotropically starting from U = 0.05. To refine with anisotropic displacement parameters in the same or a subsequent job, the instruction:

```
ANIS C1 > C5
```

should be inserted anywhere before C1 in the '.ins' file. The SIMU and ISOR restraints on the U_{ij} would be inappropriate for such a group, but:

```
DELU C1 > C5
```

could be applied if the anisotropic refinement proved unstable. The five hydrogen atoms could be added and refined with the 'riding model' by means of:

```
HFIX 43 C1 > C5
```

anywhere before C1 in the input file. For good data, in view of possible librational effects, a suitable alternative would be:

```
HFIX 44 C1 > C5
SADI 0.02 C1 H1 C2 H2 C3 H3 C4 H4 C5 H5
```

which retains a riding model but allows the C-H bond lengths to refine, subject to the restraint that they should be equal within about 0.02 Å.

In analogous manner it is possible to generate missing atoms and perform rigid group refinements for phenyl rings (AFIX 66) and Cp* groups (AFIX 109). Very often it is possible and desirable to remove the rigid group constraints (by simply deleting the AFIX instructions) in the final stages of refinement; there is good experimental evidence that the *ipso*-angles of phenyl rings differ systematically from 120° (Jones, 1988; Maetzke & Seebach, 1989; Domenicano, 1992).

As a second example, assume that the structure contains two molecules of poorly defined THF solvent, and that we have managed to identify the oxygen atoms. A rigid pentagon would clearly be inappropriate here, except possibly for placing missing atoms, since THF molecules are not planar. However we can *restrain* the 1,2- and the 1,3-distances in the two molecules to be similar by means of a 'similarity restraint' (SAME). Assume that the molecules are

numbered O11 C12 ... C15 and O21 C22 ... C25, and that the atoms are given in this order in the atom list. Then we can either insert the instruction:

```
SAME O21 > C25
```

atoms of the second molecule. Free variable 2 is then the occupation factor of the first molecule; its starting value must be specified on the FVAR instruction. The possibility of spurious bonds is eliminated by inserting 'PART 1' before the first molecule, 'PART 2' before the second, and 'PART 0' after it. Hydrogen atoms can be inserted in the usual way using the HFIX instruction since the connectivity table is 'correct'; they will automatically be assigned the site occupation factors of the atoms to which they are bonded.

Finally we would like to refine with anisotropic displacement parameters because the thermal motion of such solvent molecules is certainly not isotropic, but the refinement will be unstable unless we restrain the anisotropic displacement parameters to behave 'reasonably' by means of rigid bond restraints (DELU) and 'similar U_{ij}' restraints (SIMU); fortunately the program can set up these restraints automatically. DELU restrains the differences in the components of the displacement parameters of two atoms to zero along the 1,2- and 1,3-vector directions; these restraints are derived automatically with the help of the connectivity table. Since the SIMU restraints are much more approximate, we restrict them here to atoms which, because of the disorder, are almost overlapping (i.e. are within 0.7 Å of each other). Note that the SIMU restraints ignore the connectivity table and are based directly on a distance criterion specifically because the connectivity table does not link the disordered atoms. In order to specify a non-standard distance cut-off which is the third SIMU parameter, we must also give the first two parameters, which are the restraint esds for distances involving non-terminal atoms (0.02) and at least one terminal atom (0.04) respectively. The *.ins* file now contains:

```
HFIX 23 C12 > C15 C22 > C25
ANIS O11 > C25
DELU O11 > C25
SIMU O11 > C25 0.04 0.08 0.7
FVAR ..... 0.75
....
PART 1
SAME O21 > C25
SAME O11 C15 < C12
O11 4 ..... ..... ..... 21
C12 1 ..... ..... ..... 21
C13 1 ..... ..... ..... 21
C14 1 ..... ..... ..... 21
C15 1 ..... ..... ..... 21
PART 2
O21 4 ..... ..... ..... -21
C22 1 ..... ..... ..... -21
C23 1 ..... ..... ..... -21
C24 1 ..... ..... ..... -21
C25 1 ..... ..... ..... -21
PART 0
```

An alternative type of disorder common for THF molecules and proline residues in proteins is when one atom (say C14) can flip between two positions (i.e. it is the flap of an envelope conformation). If we assign C14 to PART 1, C14' to PART 2, and the remaining ring atoms to PART 0, then the program will be able to generate the correct connectivity, and so we can also generate hydrogen atoms for both disordered components (with AFIX, not HFIX):

```
SIMU C14 C14'
ANIS O11 > C14'
```

```
FVAR ..... 0.7
....
SAME O11 C12 C13 C14' C15
O11 4 ..... ..... .....
C12 1 ..... ..... .....
AFIX 23
H12A 2 ..... ..... .....
H12B 2 ..... ..... .....
AFIX 0
C13 1 ..... ..... .....
PART 1
AFIX 23
H13A 2 ..... ..... ..... 21
H13B 2 ..... ..... ..... 21
PART 2
AFIX 23
H13C 2 ..... ..... ..... -21
H13D 2 ..... ..... ..... -21
AFIX 0
PART 1
C14 1 ..... ..... ..... 21
AFIX 23
H14A 2 ..... ..... ..... 21
H14B 2 ..... ..... ..... 21
AFIX 0
PART 0
C15 1 ..... ..... .....
PART 1
AFIX 23
H15A 2 ..... ..... ..... 21
H15B 2 ..... ..... ..... 21
PART 2
AFIX 23
H15C 2 ..... ..... ..... -21
H15D 2 ..... ..... ..... -21
AFIX 0
C14' 1 ..... ..... ..... -21
AFIX 23
H14C 2 ..... ..... ..... -21
H14D 2 ..... ..... ..... -21
AFIX 0
PART 0
```

It will be seen that six hydrogens belong to one conformation, six to the other, and two are common to both. The generation of the idealized hydrogen positions is based on the connectivity table but also takes the PART numbers into account. These procedures should be able to set up the correct hydrogen atoms for all cases of two overlapping disordered groups. In cases of more than two overlapping groups the program will usually still be able to generate the hydrogen atoms correctly by making reasonable assumptions when it finds that an atom is 'bonded' to atoms with different PART numbers, but it is possible that there are rare examples of very complex disorder which can only be handled by using dummy atoms constrained (EXYZ and EADP) to have the same positional and displacement parameters as atoms with different PART numbers (in practice it may be easier - and quite adequate - to ignore hydrogens except on the two components with the highest occupancies !).

When the site symmetry is high, it may be simpler to apply similarity restraints using SADI or DFIX rather than SAME. For example the following three instruction sets would all restrain a perchlorate ion (CL,O1,O2,O3,O4) to be a regular tetrahedron:

```
SAME CL O2 O3 O4 O1
SADI  O1 O2  O1 O3
```

followed immediately by the atoms CL, O1... O4; the SAME restraint makes all the Cl-O bonds equal but introduces only FOUR independent restraints involving the O...O distances, which allows the tetrahedron to distort retaining only one $\bar{4}$ axis, so one further restraint must be added using SADI.

or:

```
SADI  CL O1  CL O2  CL O3  CL O4
SADI  O1 O2  O1 O3  O1 O4  O2 O3  O2 O4  O3 O4
```

or:

```
DFIX 31  CL O1  CL O2  CL O3  CL O4
DFIX 31.6330  O1 O2  O1 O3  O1 O4  O2 O3  O2 O4  O3 O4
```

in the case of DFIX, one extra least-squares variable (free variable 3) is needed, but it is the mean Cl-O bond length and refining it directly means that its esd is also obtained. If the perchlorate ion lies on a three-fold axis through CL and O1, the SADI method would require the use of symmetry equivalent atoms (EQIV $1 y, z, x and O2_$1 etc. for R3 on rhombohedral axes) so DFIX would be simpler (same DFIX instructions as above with distances involving O3 and O4 deleted) [the number 1.6330 in the above example is of course twice the sine of half the tetrahedral angle].

If you wish to test whether you have understood the full implications of these restraints, try the following problems:

(a) A C-O-H group is being refined with AFIX 87 so that the torsion angle about the C-O bond is free. How can we or: e8ed usilet

6. Refinement of Twinned Structures; Absolute Structure

A typical definition of a twinned crystal is the following: "Twins are regular aggregates consisting of crystals of the same species joined together in some definite mutual orientation" (Giacovazzo,1992). So for the description of a twin two things are necessary: a description of the orientation of the different species relative to each other (twin law) and the fractional contribution of each component. The *twin law* can be expressed as a matrix that transforms the *hkl*

6.2 Absolute structure

Even if determination of absolute configuration is not one of the aims of the structure determination, it is important to refine

The offending space groups and corresponding correct MOVE instructions are:

Fdd2	**MOVE .25 .25 1 -1**	**$I4_1cd$**	**MOVE 1 .5 1 -1**
$I4_1$	**MOVE 1 .5 1 -1**	**$I\bar{4}2d$**	**MOVE 1 .5 .25 -1**
$I4_122$	**MOVE 1 .5 .25 -1**	**$F4_1$**	

6.6 Processing of twinned and powder data

The HKLF 5 and 6 instructions force MERG 0, i.e. neither a transformation of reflection indices into a standard form nor a sort-merge is performed before refinement. If twinning is specified using the TWIN instruction, any MERG instruction may be used and the default

(h) There appear to be one or more unusually long axes, but also many absent reflections.

(i) There are problems with the cell refinement.

(j) Some reflections are sharp, others split.

(k) K = mean(F_o^2) / mean(F_c^2) is systematically high for the reflections with low intensity.

(l) For all of the 'most disagreeable' reflections, F_o is much greater than F_c.

6.8 Conclusions

Twinning usually arises for good structural reasons. When the heavy atom positions correspond to a higher symmetry space group it may be difficult or impossible to distinguish between twinning and disorder of the light atoms; see Hoenle & von Schnering (1988). Since refinement as a twin usually requires only two extra instructions and one extra parameter, in such cases it should be attempted first, before investing many hours in a detailed interpretation of the 'disorder'! Indeed, it has been suggested by G.B. Jameson that all structures (including proteins) that are solved in space groups (such as $P3_1$) that could be merohedrally twinned without changing the systematic absences should be tested for such twinning (possible only present to a minor extent) by:

```
TWIN 0 1 0  1 0 0  0 0 -1
BASF tr extent) by:
 ad1.stwino0tw (ı0.00tw ( on1.stwin,nt) 0nging the systematic6(t)-6n usDı0.008 )-7P)8(ne
```

SYMM symmetry operation

Symmetry operators, i.e. coordinates of the general positions as given in International Tables. The operator x, y, z is always assumed, so MUST NOT be input. If the structure is centrosymmetric, the origin MUST lie on a center of symmetry. Lattice centering and the presence of an inversion center should be indicated by LATT, not SYMM. The symmetry operators may be specified using decimal or fractional numbers, e.g. 0.5-x, 0.5+y, -z or Y-X, -X, Z+1/6; the three components are separated by commas.

SFAC elements

Element symbols which define the order of scattering factors to be employed by the program. The first 94 elements of the periodic system are recognized. The element name may be preceded by '$' but this is not obligatory (the '$' character is allowed for logical consistency but is ignored). The program uses the neutral atom scattering factors, f', f" and absorption coefficients from International Tables for Crystallography, Volume C (1992), Ed. A.J.C. Wilson, Kluwer Academic Publishers, Dordrecht: Tables 6.1.1.4(pp. 500-502), 4.2.6.8 (pp. 219-222) and 4.2.4.2 (pp. 193-199) respectively. The covalent radii stored in the program are based on experience rather than taken from a specific source, and are deliberately overestimated for elements which tend to have variable coordination numbers so that 'bonds' are not missed, at the cost of generating the occasional 'non-bond'. The default radii (not those set for individual atoms by CONN) are printed before the connectivity table.

SFAC label a1 b1 a2 b2 a3 b3 a4 b4 c f' f" mu r wt

Scattering factor in the form of an exponential series, followed by real and imaginary dispersion terms, linear absorption coefficient, covalent radius and atomic weight. Except for the 'label' and atomic weight the format is the same as that used in SHELX-76. label consists of up to 4 characters beginning with a letter (e.g. Ca2+) and should be included before a1; for consistency the first label character may be a '$', but this is ignored (note however that the '$', if used, counts as one of the four characters, leaving only three for the rest of the label). The two SFAC formats may be used in the same *.ins* file; the order of the SFAC instructions (and the order of element names in the first type of SFAC instruction) define the scattering factor numbers which are referenced by atom instructions. The units of mu should be barns/atom, as in Table 4.2.4.2 of International Tables, Volume C (see above). For neutrons this format should be used, with a1...b4 set to zero.

Hydrogen atoms are treated specially by SHELXL; they are recognized by having the scattering factor number that corresponds to 'H' on the SFAC instruction. For X-ray structures that contain both D and H, e.g. because the crystals were grown from a deuterated solvent in an n.m.r tube (a common source of good crystals!), both H and D should be included on the SFAC and UNIT instructions, but all the H and D atoms should employ the 'H' scattering factor number. In this way the density will be calculated correctly, but the D atoms may be idealized using HFIX etc.

DISP E f' f" [#] mu [#]

The DISP instruction allows the dispersion and (optionally) the absorption coefficient of a particular element (the name may be optionally prefaced by '$') to be read in without having to use the full form of the SFAC instruction. It will typically be used for synchrotron data where the wavelength does not correspond to the values (for Cu, Mo and Ag radiation) for which these terms are stored in the program. All other terms on the SFAC instruction are independent of the wavelength, so its short form may then be used. DISP instructions, if present, MUST come between the last SFAC and the UNIT instruction.

UNIT n1 n2 ...
Number of atoms of each type in the unit-cell, in SFAC order.

LAUE E
Wavelength-dependent values of f' and f'' may be

If the time t (measured in seconds from the start of the job) is exceeded, SHELXL performs no further least-squares cycles, but goes on to the final structure factor calculation followed by bond lengths, Fourier calculations etc. The default value of t is installation dependent, and is either set to 'infinity' or to a little less than the maximum time allocation for a particular class of job. Usually t is 'CPU time', but some some operating systems (e.g. MSDOS) the elapsed time may have to be used instead.

END

END is used to terminate an 'include' file, and may also be included after HKLF in the *.ins* file (for compatibility with SHELX-76).

7.2 Reflection data input

Before running SHELXL, a reflection data file *name.hkl* must have been prepared. The HKLF command tells the program which format has been chosen for this file, and allows the indices to be transformed using the 3x3 matrix $r_{11}...r_{33}$, so that the new *h*

The remaining options (n > 2) all require FORMAT(3I4,2F8.2,I4); other compatible formats (e.g. F8.0 or even I8) may be used for the floating point numbers provided that eight columns are used in all and a decimal point is present.

n = 3: *h k l* F_o $\sigma(F_o)$ BN [1] (if BN is absent or zero it is set to 1).The use of data corresponding to this format is allowed but is NOT RECOMMENDED, since the generation of F_o and $\sigma(F_o)$ from F_o^2 and $\sigma(F_o^2)$ is a tricky statistical problem and could introduce bias.

n = 4: *h k l* F_o^2 $\sigma(F_o^2)$ BN [1] is the standard reflection data file. Since F_o^2 is obtained as the difference of the experimental peak and background counts, it may be positive or slightly negative. BN may be made negative (e.g. by SHELXPRO) to flag a reflection for inclusion in the R_{free} reference set (see CLGS and L.S. with a second parameter of -1).

n = 5: *h k l* F_o^2 $\sigma(F_o^2)$ m where m is the twin component number. Each measured F_o^2 value is fitted to the sum of $k_{|m|}F_{c|m|}^2$ over all contributing components, multiplied by the overall scale factor. m should be given as positive for the last contributing component and negative for the remaining ones (if any). The values of F_o^2 and $\sigma(F_o^2)$ are taken from the last ('prime') reflection in a group, and may simply be set equal for each component, but the indices *h,k,l* will in general take on different values for each component. The starting values of the twin factors $k_2..k_{max(m)}$ are specified on BASF instruction(s); k_1 is given by one minus the sum of the other twin factors. Note that many simple forms of twinning can also be handled with HKLF 4 and a TWIN instruction to generate the indices of the remaining twin component(s); HKLF 5 is required if the reciprocal space lattices of the components cannot be superimposed exactly. HKLF 5 sets MERG 0, and may not be used with TWIN.

n = 6: *h k l* F_o^2 $\sigma(F_o^2)$ m as for n = 5, there may be one or more sets of reflection indices corresponding to a single F_o^2 value. The last reflection in a group has a positive m value and the previous members of the group have negative m. The values of F_o^2 and $\sigma(F_o^2)$ are taken

sum of k

will then invent suitable starting values. Note that a different formula was employed in SHELXL-93, and so parameter values from SHELXL-93 may well be unsuitable starting values for the new version.

Since both extinction and diffraction from diffuse solvent tend to affect primarily the strong reflections at low diffraction angle, they tend to show the same symptoms in the analysis of variance, and so a combined warning message is printed. It will however be obvious from the type of structural problem which of the two should be applied. The program does not permit the simultaneous refinement of SWAT and EXTI.

HOPE nh [1]
Refines 12 anisotropic scaling parameter as suggested by Parkin, Moezzi & Hope (1995). nh points to the BASF parameter that stores the value of the first HOPE parameter; if nb is negative the 12 parameters are fixed at their current values. These parameters are highly correlated with the individual atomic anisotropic displacement parameters, and so are only useful for structures that are refined isotropically , e.g. macromolecules at moderate resolution. To some extent they can also model absorption errors. If HOPE is given without any parameters and there are no BASF instructions, the program will generate appropriate starting values. If BASF parameters are needed for twin refinement or as scale factors for different batches of data, nh should be given an absolute value greater than one.

MERG n[2]
If n is equal to 2 the reflections are sorted and merged before refinement; if the structure is non-centrosymmetric the Friedel opposites are not combined before refinement (necessary distinction from SHELXS). If n is 1 the indices are converted to a 'standard setting' in which *l* is maximized first, followed by *k*, and then *h*; if n is zero, the data are neither sorted nor converted to a standard setting. n = 3 is the same as n = 2 except that Friedel opposites are also merged (this introduces small systematic errors and should only be used for good reason, e.g. to speed up the early stages of a refinement of a light atom structure before performing the final stages with MERG 2). Note that the reflections are always merged, and Friedel opposites combined, before performing Fourier calculations in SHELXL so that the (difference) electron density is real and correctly scaled. Even with n = 0 the program will change the reflection order within each data block to optimize the vectorization of the structure factor calculations (it is shuffled back into the MERG order for LIST 4 output). Note that MERG may not be used in conjunction with TWIN or HKLF 5 or 6. In SHELX-76, MERG 3 had a totally different meaning, namely the determination of inter-batch scale factors; in SHELXL, these may be included in the refinement using the BASF instruction.

MERG 4 averages all equivalents, including Friedel opposites, and sets all δf'' values to zero; it is often used in refinement of macromolecules.

7.3 Atom lists and least-squares constraints

Atom instructions begin with an atom name (up to 4 characters that do not correspond to any of the SHELXL command names, and terminated by at least one blank) followed by a scattering factor number (which refers to the

in their order of U_{ij} components; SHELXL uses the same order as SHELX-76. The exponential factor takes the form $\exp(-8\pi^2 U[\sin(\theta)/\lambda]^2)$ for an isotropic displacement parameter U and:

$$\exp (-2\pi^2 [h^2(a^*)^2U_{11} + k^2(b^*)^2U_{22} + ... + 2hka^*b^*U_{12}])$$

for anisotropic U_{ij}. An atom is specified as follows in the *.ins* file:

atomname sfac x y z sof [11] U [0.05] or U_{11} U_{22} U_{33} U_{23} U_{13} U_{12}

The atom name must be unique, except that atoms in different residues - see RESI - may have the same names; in contrast to SHELX-76 it is not necessary to pad out the atom name to 4 characters with blanks. To fix any atom parameter, add 10. Thus the site occupation factor is normally given as 11 (i.e. fixed at 1). The site occupation factor for an atom in a special position should be multiplied by the multiplicity of that position (as given in International Tables, Volume A) and divided by the multiplicity of the general position for that space group. This is the same definition as in SHELX-76 and is retained for upwards compatibility; it might have been less confusing to keep the multiplicity and occupation factor separate. An atom on a fourfold axis for example will usually have s.o.f. = 10.25.

If any atom parameter is given as (10_*m+p), where abs(p) is less than 5 and m is an integer, it is interpreted as $p \bullet fv_m$, where fv_m is the mth 'free variable' (see FVAR). Note that there is no fv_1, since this position on an FVAR instruction is occupied by the overall scale factor, and m=1 corresponds to fixing an atom by adding 10. If m is negative, the parameter is interpreted as $p \bullet (fv_{-m}-1)$. Thus to constrain two occupation factors to add up to 0.25 (for two elements occupying the same fourfold special position) they could be given as 20.25 and -20.25, i.e. $0.25 \bullet fv_2$ and $0.25 \bullet (1-fv_2)$, which correspond to p=0.25, m=2 and p=-0.25, m=-2 respectively.

In SHELX-76, it was necessary to use free variables and coordinate fixing in this way to set up the appropriate constraints for refinement of atoms on special positions. In SHELXL, this is allowed (for upwards compatibility) but is NOT NECESSARY: the program will automatically work out and apply the appropriate positional, s.o.f. and U_{ij} constraints for any special position

given the s.o.f.'s 30.25 and -30.25 respectively. In this case it would be desirable to use the EADP instruction to equate the Ca^{2+} and Ba^{2+} (anisotropic) displacement parameters.

If U is given as -T, where T is in the range 0.5 < T < 5, it is fixed at T times the U_{eq} of the previous atom not constrained in this way. The resulting value is not refined independently but is updated after every least-squares cycle.

SPEC del[0.2]
All following atoms (until the next SPEC instruction) are considered to lie on special positions (for the purpose of automatic constraint generation) if they lie within del (Å) of a special position. The coordinates of such an atom are also adjusted so that it lies exactly on the special position.

RESI class[ ] number[0] alias
Until the next RESI instruction, all atoms are considered to be in the specified 'residue', which may be defined by a class (up to four characters, beginning with a letter) or number (up to four digits) or both. The same atom names may be employed in different residues, enabling them to be referenced globally or selectively. The residue number should be unique to a particular residue, but the class may be used to refer to a class of similar residues, e.g. a particular type of amino acid in a polypeptide.

Residues may be referenced by any instruction that allows atom names; the reference takes the form of the character '_' followed by either the residue class or number without intervening spaces. If an instruction codeword is followed immediately by a residue number, all atom names referred to in the instruction are assumed to belong to that residue unless they are themselves immediately followed by '_' and a residue number, which is then used instead. Thus:

```
RTAB_4 Ang N H0 O_11
```

would cause the calculation of an angle N_4 - H0_4 - O_11, where the first two atoms are in residue 4 and the third is in residue 11.

If the instruction codeword is followed immediately by a residue class, the instruction is effectively duplicated for all residues of that class. '_* ' may be used to match all residue classes; this includes the default class ' ' (residue number 0) which applies until the first RESI instruction is encountered. Thus:

```
MPLA_phe CB > CZ
```

would calculate least-squares planes through atoms CB to CZ inclusive of all residues of class 'phe' (phenylalanine). In the special case of HFIX, only the FIRST instruction which applies to a given atom is applied. Thus:

```
HFIX_1 33 N
HFIX_* 43 N
```

would add hydrogens to the N-terminal nitrogen (residue 1) of a polypeptide to generate a (protonated) $-NH_3^+$ group, but all other (amide) nitrogens would become -NH-.

Individual atom names in an instruction may be followed by '_' and a residue number, but not by '_*' or '_' and a residue class. If an atom name is not followed by a residue number, the current residue is assumed (unless overridden by a global residue number or class appended to the instruction codeword). The symbols '_+' meaning 'the next residue' and '_-' meaning 'the preceding residue'(i.e. residues number n+1 and n-1 if the current residue number is n) may be appended to atom names but not to instruction codenames. Thus the instruction:

```
RTAB_* Omeg CA_+ N_+ C CA
```

could be used to calculate all the peptide ω torsion angles in a protein or polypeptide. If (as at the C-terminus in this example) some or all of the named atoms cannot be found for a particular residue, the instruction is simply ignored for that residue.

'_$n' does not refer to a residue; it uses the symmetry operation $n defined by a preceding 'EQIV $n' instruction to generate an equivalent of the named atom (see EQIV). alias specifies

anisotropic of all atoms specified by ANIS until a given number of least-squares cycles has been performed.

AFIX mn d[#] sof[11] U[10.08]

AFIX applies constraints and/or generates idealized coordinates for all atoms until the next AFIX instruction is read. The digits mn of the AFIX code control two logically quite separate operations. Although this is confusing for new users, it has been retained for upwards compatibility with SHELX-76, and because it provides a very concise notation. m refers to geometrical operations which are performed before the first refinement cycle (hydrogen atoms are idealized before every cycle), and n sets up constraints which are applied throughout the least-squares refinement. n is always a single digit; m may be two, one or zero digits (the last corresponds to m = 0).

The options for idealizing hydrogen atom positions depend on the connectivity table that is set up using CONN, BIND, FREE and PART; with experience, this can also be used to generate hydrogen atoms attached to disordered groups and to atoms on special positions. d determines the bond lengths in the idealized groups, and sof and U OVERRIDE the values in the atom list for all atoms until the next AFIX instruction. U is not applied if the atom is already anisotropic, but is used if an isotropic atom is to be made anisotropic using ANIS. Any legal U value may be used, e.g. 31 (a free variable reference) or -1.2 (1.2 times Ueq of the preceding normal atom). Each AFIX instruction must be followed by the required number of hydrogen or other atoms. The individual AFIX options are as follows; the default X-H distances depend on both the chemical environment and the temperature (to allow for librational effects) which is specified by means of the TEMP instruction.

m = 0 No action.

m = 1 Idealized tertiary C-H with all X-C-H angles equal. There must be three and only three other bonds in the connectivity table to the immediately preceding atom, which is assumed to be carbon. m = 1 is often combined with a riding model refinement (n = 3).

m = 2 Idealized secondary CH_2 with all X-C-H and Y-C-H angles equal, and H-C-H determined by X-C-Y (i.e. approximately tetrahedral, but widened if X-C-Y is much less than tetrahedral). This option is also suitable for riding refinement (n = 3).

m = 3 Idealized CH_3 group with tetrahedral angles. The group is staggered with respect to the shortest other bond to the atom to which the $-CH_3$ is attached. If there is no such bond (e.g. an acetonitrile solvent molecule) this method cannot be used (but m = 13 is still viable).

m = 4 Aromatic C-H or amide N-H with the hydrogen atom on the external bisector of the X-C-Y or X-N-Y angle. m = 4 is suitable for a riding model refinement, i.e. AFIX 43 before the H atom.

m = 5 Next five non-hydrogen atoms are fitted to a regular pentagon, default d = 1.42 Å.

m = 6 Next six non-hydrogen atoms are fitted to a regular hexagon, default d = 1.39 Å.

m = 7 Identical to m = 6 (included for upwards compatibility from SHELX-76). In SHELX-76 only the first, third and fifth atoms of the six-membered ring were used as target atoms; in SHELXL this will still be the case if the other three are given zero coordinates, but the procedure is more general because any one, two or three atoms may be left out by giving them zero coordinates.

m = 8 Idealized OH group, with X-O-H angle tetrahedral. If the oxygen is attached to a saturated carbon, all three staggered positions are considered for the hydrogen. If it

angle is set that maximizes the sum of the electron density at the three calculated hydrogen positions. Since even this is not an infallible method of getting the correct torsion angle, it should normally be combined with a rigid or rotating group refinement for the methyl group (e.g. mn = 137 before the first H). In subsequent least-squares cycles the group is re-idealized retaining the current torsion angle

m = 14 Idealized OH group, with X-O-H angle tetrahedral. If the coordinates of the hydrogen atom are non-zero, they are used to define the torsion angle. Otherwise (or if HFIX was used to set up the AFIX instruction) the torsion angle is chosen which maximizes the electron density (see m = 13). Since this torsion angle is unlikely to be very accurate, the use of a rotating group refinement is recommended (i.e. AFIX 147 before the H atom).

m = 15 BH group in which the boron atom is bonded to either four or five other atoms as part of an polyhedral fragment. The hydrogen atom is placed on the vector that represents the negative sum of the unit vectors along the four or five other bonds to the boron atom.

m = 16 Acetylenic C-H, with X-C-H linear. Usually refined with the riding model, i.e. AFIX 163.

m > 16 A group defined in a FRAG...FEND section with code = m is fitted, usually as a preliminary to rigid group refinement. The FRAG...FEND section MUST precede the corresponding AFIX instruction in the '*.ins*' file, but there may be any number of AFIX instructions with the same m corresponding to a single FRAG...FEND section.

When a group is fitted (m = 5, 6, 10 or 11, or m > 16), atoms with non-zero coordinates are used as target atoms with equal weight. Atoms with all three coordinates zero are ignored. Any three or more non-colinear atoms may be used as target atoms.

'Riding' (n = 3, 4) and 'rotating' (n = 7, 8) hydrogen atoms, but not other idealized groups, are re-idealized (if m is 1, 2, 3, 4, 8, 9, 12, 13, 14, 15 or 16) before each refinement cycle (after the first cycle, the coordinates of the first hydrogen of a group are always non-zero, so the torsion angle is retained on re-idealizing). For n = 4 and 8, the angles are re-idealized but the (refined) X-H bond length is retained, unless the hydrogen coordinates are all zero, in which case d (on the AFIX instruction) or (if d is not given) a standard value which depends on the chemical environment and temperature (TEMP) is used instead.

n = 0 No action.

n = 1 The coordinates, s.o.f. and U or U_{ij} are fixed.

n = 2 The s.o.f. and U (or U_{ij}) are fixed, but the coordinates are free to refine.

n = 3 The coordinates, but not the s.o.f. or U (or U_{ij}) 'ride' on the coordinates of the previous atom with n not equal to 3. The same shifts are applied to the coordinates of both atoms, and both contribute to the derivative calculation. The atom on which riding is performed may not itself be a riding atom, but it may be in a rigid group (m = 5, 6 or 9).

n = 4 This constraint is the same as n = 3 except that the X-H distance is free to refine. The X-H vector direction does not change. This constraint requires better quality reflection data than n = 3, but allows for variations in apparent X-H distances caused by libration and bonding effects. If there is more than one equivalent hydrogen, the same shift is applied to each equivalent X-H distance (e.g. to all three C-H bonds in a methyl group). n = 4 may be combined with DFIX or SADI restraints (to restrain chemically equivalent X-H distances to be equal) or embedded inside a rigid (n = 6) group, in which case the next atom (if any) in the same rigid group must follow an explicit AFIX instruction with n = 5. Note that n = 4 had a different effect in SHELX-76.

n = 5 The next atom(s) are 'dependent' atoms in a rigid group. Note that this is automatically generated for the atoms following an n = 6 or n = 9 atom, so does not need to be included specifically unless m has to be changed (e.g. AFIX 35 before the first hydrogen of a rigid methyl group with AFIX 6 or 9 before the preceding carbon).

n = 6 The next atom is the 'pivot atom' of a NEW rigid group, i.e. the other atoms in the rigid group rotate about this atom, and the same translational shifts are applied to all atoms in the rigid group.

n = 7 The following (usually hydrogen) atoms (until the next AFIX with n not equal to 7) are allowed to ride on the immediately preceding atom X and rotate about the Y-X bond; X must be bonded to one and only one

solution may be given zero coordinates, in which case they will be generated from the rigid group fit.

A rigid group or set of dependent hydrogens must ALWAYS be followed by 'AFIX 0' (or another AFIX instruction). Leaving out 'AFIX 0' by mistake is a common cause of error; the program is able to detect and correct some obvious cases, but in many cases this is not logically possible.

HFIX mn U[#] d[#] atomnames

HFIX generates AFIX instructions and dummy hydrogen atoms bonded to the named atoms, the AFIX parameters being as specified on the HFIX instruction. This is exactly equivalent to the corresponding editing of the atom list. The atom names may reference residues (by appending '_n' to the name, where n is the residue number), or SFAC names (preceded by a '$' sign). U may be any legal value for the isotropic temperature factor, e.g. 21 to tie a group of hydrogen U value to free variable 2, or -1.5 to fix U at 1.5 times U(eq) of the preceding normal atom. HFIX MUST precede the atoms to which it is to be applied. If more than one HFIX instruction references a given atom, only the FIRST is applied. 'HFIX 0' is legal, and may be used to switch off following HFIX instructions for a given atom (which is useful if they involve '_*' or a global reference to a residue class).

FRAG code[17] a[1] b[1] c[1] α[90] β[90] γ[90]

Enables a fragment to be input using a cell and coordinates taken from the literature. Orthogonal coordinates may also be input in this way. Such a fragment may be fitted to the set of atoms following an AFIX instruction with m = code (code must be greater than 16); there must be the same number of atoms in this set as there are following FRAG, and they must be in the same order. Only the coordinates of the FRAG fragment are actually used; atom names, sfac numbers, sof and U_{ij} are IGNORED. A FRAG fragment may be given anywhere between UNIT and HKLF or END, and must be terminated by a FEND instruction, but must precede any AFIX instruction which refers to it. ber)0.004t usingFid 1.tl, an-20(thet)10(r)10a0(r)10

```
EADP F11 F14
EADP F12 F15
EADP F13 F16
C1 .......
PART 1
F11 ...... 21 ......
F12 ...... 21 ......
F13 ...... 21 ......
PART 2
F14 ...... -21 ......
F15 ...... -21 ......
F16 ...... -21 ......
PART 0
```

EADP applies an (exact) *constraint*. The SIMU instruction *restrains* the Uij components of neighboring atoms to be approximately equal with an appropriate (usually fairly large) esd.

EQIV $n symmetry operation

Defines symmetry operation $n for referencing symmetry equivalent atoms on any instruction which allows atom names, by appending '_$n' (where n is an integer between 1 and 511 inclusive) to the atom name. Such a symmetry operation must be defined beforeit is used; it does not have to be an allowed operation of the space group, but the same notation is used as on the SYMM instruction. The same $n may not appear on two separate EQIV instructions. Thus:

```
EQIV $2 1-x, y, 1-z
CONF C1 C2 C2_$2 C1_$2
```

could be used to calculate a torsion angle across a crystallographic twofold axis (note that this may be required because CONF with no atom names only generates torsion angles automatically that involve the unique atom list and a one atom deep shell of symmetry equivalents). If the instruction codeword refers to a residue, this is applied to the named atoms before any symmetry operation specified with '_$n'. Thus:

```
RTAB_23 O..O OG_12 O_$3
```

would calculate the (hydrogen bond) distance between OG_12 and (O_23)_$3, i.e. between OG in residue 12 and the equivalent obtained by applying the symmetry operation defined by EQIV $3 to the atom O in residue 23.

OMIT atomnames

The named atoms are retained in the atom list but ignored in the structure factor calculation and least-squares refinement. This instruction may be used, together with L.S. 0 and FMAP 2, to create an 'OMIT map' to get a clearer picture of disordered regions of the structure; this concept will be familiar to macromolecular crystallographers. In particular, 'OMIT $H' can be used to check the hydrogen atom assignment of -OH groups etc. If an actual peak is present within 0.31 A of the calculated hydrogen atom position, the electron density appears in the 'Peak' column of the output created by PLAN with a negative first parameter. OMIT_* $H must be used for this if residues are employed.

7.4 The connectivity list

The connectivity list is a list of 'bonds' that is set up automatically, and may be edited using BIND and FREE. It is used to define idealized hydrogen atom positions, for the BOND and PLAN output of bond lengths and angles, and by the instructions DELU, CHIV, SAME and SIMU. Hydrogen atoms are excluded from the connectivity list (except when introduced by hand using BIND).

CONN bmax[12] r[#] atomnames or CONN bmax[12]

The CONN instruction fine-tunes the generation of the connectivity table and is particularly useful when π-bonded ligands or metal ions are present in the structure. For the purposes of the connectivity table (which is always generated), bonds are all distances between non-hydrogen atoms less than r1 + r2 + 0.5 Å, where r1 and r2 are the covalent radii of the atoms in question (taking PART into consideration as explained below). A shell of symmetry equivalent atoms is also generated, so that all unique bonds are represented at least once in the list. All bonds, including those to symmetry equivalent atoms, may be deleted or added using the FREE or BIND instructions.

Default values of r (identified by the scattering factor type) are stored in the program. These defaults may be changed (for both the connectivity table AND the PLAN -n output) by using the full form of the SFAC instruction. Alternatively the defaults may be overridden for the named atoms by specifying r on a CONN instruction, in which case r is used in the generation of the connectivity list but not by the PLAN instruction. '$' followed by an element name (the same as on a SFAC instruction) may also be employed on a CONN instruction (and also does not apply to PLAN). The second form of the CONN instruction may be used to change the maximum coordination number bmax for all atoms (which defaults to 12 if there is no CONN instruction).

If, after generating bonds as above and editing with FREE and BIND, there are more than bmax bonds to a given atom, the list is pruned so that only the bmax shortest are retained. A harmless side-effect of this pruning of the connectivity list is that symmetry operations may be stored and printed that are never actually used. Note that this option only removes one entry for a bond from the connectivity list, not both, except in the case of 'CONN 0' which ensures that there are no bonds to or from the named atoms. 'CONN 0' is frequently used to prevent the solvent water in macromolecular structures from making additional 'bonds' to the macromolecule which confuse the generation of idealized hydrogen atoms etc. In some cases it will be necessary to use FREE to remove a 'bond' from a light atom to an alkali metal atom (for example) in order to generate hydrogen atoms correctly. Refinements of macromolecules will often include BUMP and 'CONN 0 O_200 > LAST' (where the water happens to begin with residue 200). 'LAST' is used to indicate the last atom in the file, which saves trouble when adding extra waters.

The CONN instruction, like ANIS and HFIX, MUST precede the atoms to which it is to be applied. Repeated CONN instructions are allowed; the LAST relevant CONN preceding a particular atom is the one which is actually applied. CONN without atom names changes the default value of bmax for all following atoms. The following example illustrates the use of CONN:

```
CONN Fe 0
MPLA 5 C11 > C15 Fe
```

```
MPLA 5 C21 > C25 Fe
Fe  .....
C11 .....
.........
C25 .....
```

which would prevent bonds being generated from the iron atom to all 10 carbons in ferrocene. In this example, the distances of the iron atom from the two ring planes would be calculated instead.

PART n sof

The following atoms belong to PART n of a disordered group. The automatic bond generation ignores bonds between atoms with different PART numbers, unless one of them is zero (the value before the first PART instruction). If a site occupation factor (sof) is specified on the PART instruction, it overrides the value on the following atom instructions (even if set via an AFIX instruction) until a further PART instruction, e.g. 'PART 0', is encountered).

If n is negative, the generation of special position constraints is suppressed and bonds to symmetry generated atoms with the same or a different non-zero PART number are excluded; this is suitable for a solvent molecule disordered on a special position of higher symmetry than the molecule can take (e.g. a toluene molecule on an inversion center). A PART instruction remains in force until a further PART instruction is read; 'PART 0' should be used to continue with the non-disordered part of the structure.

Some care is necessary in generating hydrogen atoms where disordered groups are involved. If the hydrogen atoms are assigned a PART number, then even if the atom to which they are attached has no part number (i.e. PART 0) the above rules may be used by the program to work out the correct connectivity for calculating the hydrogen atom positions. HFIX hydrogens are assigned the PART number of the atom to which they are attached. If the hydrogens and the atom to which they are attached belong to PART zero but the latter is bonded to atoms with non-zero PART, the LOWEST of these non-zero PART numbers is assumed to be the major component and is used to calculate the hydrogen positions. In general, if the same residue numbers and names and the same atom names but different PART numbers are used for different disorder components in a macromolecule, HFIX will generate hydrogen atoms correctly without any special action being required. For example the use of HFIX with the following disordered serine residue:

```
HFIX_Ser 33 N
HFIX_Ser 13 CA
HFIX_Ser 23 CB
HFIX_Ser 83 CG
 :
RESI 32 Ser
N  .....
CA .....
C  .....
O  .....
PART 1
CB   1  ...  ...  ...  21  ...
OG   4  ...  ...  ...  21  ...
PART 2
CB   1  ...  ...  ... -21  ...
OG   4  ...  ...  ... -21  ...
```

```
PART 0
```

would set up the AFIX hydrogens as if the following had been input. Note that only one, fully occupied, hydrogen is attached to CA; for this reason, and also to prevent small inconsistencies in the DFIX and DANG restraints, the disorder should be traced back one more atom than can be resolved (i.e. CB should be split even if it does not look as though this would be necessary in an electron density map):

```
RESI 32 Ser
N  .....
AFIX 43
H0   2  ...  ...  ...  11  -1.2
AFIX 0
CA .....
AFIX 13
HA   2  ...  ...  ...  11  -1.2
AFIX 0
C  .....
O  .....
PART 1
CB   1  ...  ...  ...  21  ...
AFIX 23
HB1  2  ...  ...  ...  21  -1.2
HB2  2  ...  ...  ...  21  -1.2
AFIX 0
OG   4  ...  ...  ...  21  ...
AFIX 83
HG   2  ...  ...  ...  21  -1.5
AFIX 0
PART 2
CB   1  ...  ...  ... -21  ...
AFIX 13
HB1  2  ...  ...  ... -21  -1.2
HB2  2  ...  ...  ... -21  -1.2
AFIX 0
OG   4  ...  ...  ... -21  ...
AFIX 83
HG   2  ...  ...  ... -21  -1.5
AFIX 0
PART 0
```

where free variable 2 is the occupation factor for PART 1 (say 0.7) and the occupation factor of the second component is tied to 1-fv(2) (i.e. 0.3). The value for this free variable is set on the FVAR instruction and is free to refine. If there were more than two components, a linear free variable restraint (SUMP) could be used to restrain the sum of occupation factors to unity. The addition of disorder components after including hydrogen atoms will require some hand editing and so is less efficient, but the auxiliary program SHELXPRO can be persuaded to do most of the work

BIND atom1 atom2

The specified 'bond' (which may be of any length) is added to the connectivity list if it is not there already. Only one of the two atoms may be an equivalent atom (i.e. have the extension _$n).

FREE atom1 atom2

make s negative to prevent the generation of anti-bumping restraints that would break the bond. Refinement with anti-bumping restraints provides a solvent model with acceptable hydrogen bonding distances that is consistent with the diffraction data. The anti-bumping restraints are regenerated before each refinement cycle. Anti-bumping restraints can also be added by hand using DFIX instructions with negative distances d.

SAME s1[0.02] s2[0.02] atomnames

The list of atoms (which may include the symbol '>' meaning all intervening non-hydrogen atoms in a forward direction, or '<' meaning all intervening non-hydrogen atoms in a backward direction) is compared with the same number of atoms which follow the SAME instruction. All bonds in the connectivity list for which both atoms are present in the SAME list are restrained to be the same length as those between the corresponding following atoms (with an effective standard deviation s1). The same applies to 1,3 distances (defined by two bonds in the connectivity list which share a common atom), with standard deviation s2. The default value of s1 is taken from the first DEFS parameter; the default value of s2 is twice this. s1 or s2 may be set to zero to switch off the corresponding restraints. The program automatically sets up the n*(n-1)/2 restraint equations required when n interatomic distances should be equal. This ensures optimum efficiency and avoids arbitrary unequal weights. Only the minimum set of restraints needs to be specified in the *.ins* file; redundant restraints are ignored by the program, provided that they have the same sigma values as the unique set of restraints. See also SADI and NCSY for closely related restraints.

The position of a SAME instruction in the input file is critical. This creates problems for programs such as SHELXPRO that provide a user interface to SHELXL, and for protein refinements SADI is to be preferred (e.g. to apply 4m local symmetry to a heme group); normally for proteins most of the 1,2- and 1,3-distances will be restrained to target values using DFIX and DANG respectivelly anyway. However SAME provides an elegant way of specifying that chemically identical but crystallographically independent molecules have the same 1,2 and 1,3 distances, e.g.

```
C1A
:
C19A
SAME C1A > C19A
C1B
:
C19B
SAME C1A > C19A
C1C
:
C19C
```

etc. This requires just n-1 SAME instructions for n equivalent molecules. In a more complicated example, assume that a structure contains several toluene solvent molecules that have been assigned the same atom names (in the same order!) and the same residue name (Tol) but different residue numbers, then one SAME instruction suffices:

```
SAME_Tol C1 > C7
```

This instruction may be inserted anywhere except after the last Tol residue; the program applies it as if it were inserted before the next atom that matches C1_Tol . This is convenient for proteins with repeated non-standard residues, since one command suffices to apply

suitable restraints, and no target values are needed, for compatibility with SHELXPRO tis SAME instruction has to be placed before the FVAR instruction. This is an exception to the usual rule that the action of a SAME instruction is position dependent; but it might be best to put it before a toluene residue with good geometry, since the connectivity table for this residue will be used to define the 1,2- and 1,3-distances. In this case it would also be reasonable to impose local two-fold symmetry for each phenyl ring, so a further SAME instruction could be added immediately before one toluene residue (the ring is assumed to be labeled cyclicly C1 .. C6 followed by the methyl group C7 which is attached to C1):

```
SAME C1 C6 < C2 C7
```

which is equivalent to:

```
SAME C1 C6 C5 C4 C3 C2 C7
```

Note that these two SAME restraints are all that is required, however many PHE residues are present; the program will generate all indirectly implied 1,2 and 1,3 equal-distance restraints! In this case it would also be sensible to restrain the atoms of each tolune molecule to be coplanar by a FLAT restraint:

```
FLAT_Tol C1 > C7
```

SADI s[0.02] atom pairs

The distances between the first and second named atoms, the third and fourth, fifth and sixth etc. (if present) are restrained to be equal with an effective standard deviation s. The SAME and SADI restraints are analyzed together by the program to find redundant and implied restraints. The same effect as is obtained using SADI can also be produced by using DFIX with d tied to a free variable, but the latter costs one more least-squares parameter (but in turn produces a value and esd for this parameter). The default effective standard deviations for SADI may be changed by means of a DEFS instruction before the instruction in question.

CHIV V[0] s[0.1] atomnames

The chiral volumes of the named atoms are restrained to the value V (in $Å^3$) with standard deviation s. The chiral volume is defined as the volume of the tetrahedron formed by the three bonds to each named atom, which must be bonded to three and only three non-hydrogen atoms in the connectivity list; the (ASCII) alphabetical order of the atoms making these three bonds defines the sign of the chiral volume. Note that RTAB may be used to list chiral volumes defined in the same way but without restraining them. The chiral volume is positive

has good convergence properties because it does not fix the orientation of the plane in its current position. s should be given in $Å^3$ as for CHIV, but for comparison with other methods the r.m.s. deviation from the plane is also printed. The default values of s is set by the second DEFS parameter.

DELU s1[0.01] s2[0.01] atomnames

All bonds in the connectivity list connecting atoms on the same DELU instruction are subject to a 'rigid bond' restraint, i.e. the components of the (anisotropic) displacement parameters in the direction of the bond are restrained to be equal within an effective standard deviation s1. The same type of restraint is applied to 1,3-distances as defined by the connectivity list (atoms 1, 2 and 3 must all be defined on the same DELU

both translation and libration of a large fragment will result in relatively similar U_{ij} components on adjacent atoms. SIMU may be combined with ISOR, which applies a further soft but quite different restraint on the U_{ij} components. SIMU may also be used when one or both of the atoms concerned is isotropic, in which case experience indicates that a larger esd (say 0.1 $Å^2$) is appropriate. The default value of s may be changed by a preceding DEFS instruction (st is then set to twice s).

DEFS sd[0.02] sf[0.1] su[0.01] ss[0.04] maxsof[1]

DEFS may be used to change the default effective standard deviations for the following DFIX, SAME, SADI, CHIV, FLAT, DELU and SIMU restraints, and is useful when these are to be varied systematically to establish the optimum values for a large structure (e.g. using R_{free}). sd is the default for s in the SADI and DFIX instructions, and also for s1 and s2 in the SAME instruction. sf is the default effective standard deviation for CHIV and FLAT, su is the default for both s1 and s2 in DELU, and ss is the default s for SIMU. The default st for SIMU is set to twice the default s.

maxsof is the maximum allowed value that an occupation factor can refine to; occupation factors that are fixed or tied to free variables are not restricted. It is possible to change this parameter (to say 1.1 to allow for hydrogen atoms) when refining both occupation factors and U's for solvent water in proteins (a popular but suspect way of improving the R factor).

ISOR s[0.1] st[0.2] atomnames

The named atoms are *restrained* with effective standard deviation s so that their U_{ij} components approximate to isotropic behavior; however the corresponding isotropic U is free to vary. ISOR is often applied, perhaps together with SIMU, to allow anisotropic refinement of large organic molecules when the data are not adequate for unrestrained refinement of all the U_{ij}; in particular ISOR can be applied to solvent water for which DELU and SIMU are inappropriate. ISOR should in general be applied as a weak restraint, i.e. with relatively large sigmas, for the reasons discussed above (see SIMU); however it is also useful for preventing individual atoms from becoming 'non-positive-definite'. However it should not be used indiscriminately for this purpose without investigating whether there are reasons (e.g. disorder, wrong scattering factor type etc.) for the atom going n.p.d. If (according to the connectivity table, i.e. ignoring attached hydrogens) his 0 TDı0.005 Tcı6 Twı[(axsof i)-kthen0(tho0(

the definition (and refinement) of a matrix transformation and mask. They are also very flexible, and can accommodate rotation of the molecule about hinges etc. Since for macromolecules at modest resolution the 1,2- and 1,3-distances are normally restrained to fixed target values by DFIX and DANG restraints, the NCS restraints are generated for equivalent 1,4-distances (if sd is non-zero or absent) and equivalent isotropic U-values (if su

```
Al1 ... ... ... ... 40.25 ...    ! 0.25 * fv(4)  four-fold axis, i.e. site
K1  ... ... ... ... 50.25 ...    ! 0.25 * fv(5)  symmetry 4]
```

This particular refinement would probably still be rather unstable, but the situation could be improved considerably by adding weak SUMP restraints for the elemental analysis. Such SUMP restraints may be used when elements are distributed over several sites in minerals so that the elemental composition corresponds (within suitable standard deviations) to an experimental chemical analysis.

SUMP may also be applied to BASF, EXTI and BASF parameters, including parameters used to describe twinning (TWIN) and anisotropic scaling (HOPE). The parameters are counted in the order overall scale and free variables, EXTI, then BASF.

7.6 Least-squares organization

`L.S. nls[0] nrf[0] nextra[0] maxvec[511]`

nls cycles of full-matrix least-squares refinement are performed, followed by a structure factor calculation. When L.S. (or CGLS) is combined with BLOC, each cycle involves refinement of a block of parameters which may be set up differently in different cycles. If no L.S. or CGLS instruction is given, 'L.S. 0' is assumed.

If nrf is positive, it is the number of these cycles that should be performed before applying ANIS. This two-stage refinement is particularly suitable for the early stages of least-squares refinement; experience indicates that it is not advisable to let everything go at once!

Negative nrf indicates which reflections should be ignored during the refinement but used instead for the calculation of free *R*-factors in the final structure factor summation; for example L.S. 4 –10 would ignore every 10th reflection for refinement purposes. It is desirable to use the same negative value of nrf throughout, so that the values of '*R*1(free)' and 'w*R*2(free)' are not biased by the 'memory' of the contribution of these reflections to earlier refinements. These independent *R*-factors (Brünger, 1992) may be used to calibrate the sigmas for the various classes of restraint, and provide a check as to whether the data are being 'over-refined' (primarily a problem for macromolecules with a poor data to parameter ratio). In SHELXL, these ignored reflections are not used for Fourier calculations.

nrf=–1 selects the R_{free} reference set that is flagged (with negative batch numbers) in the *.hkl* file (SHELXPRO may be used to do this). The division of the data into reference and working set is then independent of the space group and the MERG, OMIT and SHEL settings. However on merging reflections, to play safe a reflection is retained in the reference set only if all equivalents have the R_{free} flag set. Thus if equivalents are present, it is a good idea to use the SHELXPRO option to set the R_{free} flag in thin shells, so that all equivalents of a particular unique reflection are either all in the reference set or all in the working set. nrf=–1 is the recommended way of applying the R_{free} test in SHELXL.

nextra is the number of additional parameters which were derived from the data when performing empirical absorption corrections etc. It should be set to 44 for DIFABS [or 34 without the theta correction; Walker & D. Stuart (1983)]. It ensures that the standard deviations and GooF are estimated correctly; they would be underestimated if the number of

BLOC n1 n2 atomnames

If n1 or n2 are positive, the x, y and z parameters of the named atoms are refined in cycle |n1| or |n2| respectively.. If n1 or n2 are negative, the occupation and displacement parameters are refined in the cycle. Not more than two such cycle numbers may be

a final least-squares cycle be performed with little or no damping in order to improve these estimated standard deviations. Theoretically, damping only serves to improve the convergence properties of the refinement, and can be gradually reduced as the refinement converges; it should not influence the final parameter values. However in practice damping also deals effectively with rounding error problems in the (single-precision) least-squares matrix algebra, which can present problems when the number of parameters is large and/or restraints are used (especially when the latter have small esd's), and so it may not prove possible to lift the damping entirely even for a well converged refinement.

Note the use of 'DAMP 0 0' to estimate esds but not apply shifts, e.g. when a final L.S. 1 job is performed after CGLS refinement.

For CGLS refinements, damp is the multiplicative shift factor applied in the first cycle. In subsequent CGLS cycles it is modified based on the experience in the previous cycles. If a refinement proves unstable in the first cycle, damp should be reduced from its default value of 0.7.

If the maximum shift/esd for a L.S. refinement (excluding the overall scale factor) is greater than limse, all the shifts are scaled down by the same numerical factor so that the maximum is equal to limse. If the maximum shift/esd is smaller than limse no action is taken. This helps to prevent excessive shifts in the early stages of refinement. limse is ignored in CGLS refinements.

STIR sres step[0.01]

The STIR instruction allows a stepwise improvement in the resolution. In the first refinement cycle, the high-resolution limit (i.e. lowest d) is set at sres, in the next cycle to (sres–step), in the next (sres–2•step) etc. This continues until the limit of the data or the SHEL limit is reached, after which any remaining cycles to complete the number specified by CGLS or L.S. are completed with a constant resolution range. By starting at lower resolution and then gradually improving it, the radius of convergence for models with significant coordinate errors should be increased. This may be regarded as a primitive form of 'simulated annealing'; it could be useful in the early stages of refinement of molecular replacement solutions, or for getting rid of bias for R_{free} tests (in cases where the solution of the struture was - possibly of ncessity - based on all the data).

WGHT a[0.1] b[0] c[0] d[0] e[0] f[.33333]

The weighting scheme is defined as follows:

$$w = q / [\sigma^2(F_o^2) + (a*P)^2 + b*P + d + e*\sin(\theta)]$$

where P = [f * Maximum of (0 or F_o^2) + (1-f) * F_c^2]. It is possible for the experimental F_o^2 value to be negative because the background is higher than the peak; such negative values are replaced by 0 to avoid possibly dividing by a very small or even negative number in the expression for w. For twinned and powder data, the F_c^2 value used in the expression for P is the total calculated intensity obtained as a sum over all components. q is 1 when c is zero, $\exp[c*(\sin(\theta)\lambda)^2]$ when c is positive, and $1 - \exp[c*(\sin(\theta)/\lambda)^2]$ when c is negative.

The use of P rather than (say) F_o^2 reduces statistical bias (Wilson 1976). The weighting scheme is NOT refined if a is negative (contrast SHELX-76). The parameters can be set by

trial and error so that the variance shows no marked systematic trends with the magnitude of F_c^2 or of resolution; the program suggests a suitable WGHT instruction after the analysis of variance. This scheme is chosen to give a flat analysis of variance in terms of F_c^2, but does not take the resolution dependence into account. It is usually advisable to retain default weights (WGHT 0.1) until all atoms have been found and the refinement is essentially complete, when the scheme suggested by the program can be used for the next refinement job by replacing the WGHT instruction (if any) by the one output by the program towards the end of the *.res* file. This procedure is adequate for most routine refinements.

It may be desirable to use a scheme which does not give a flat analysis of variance to emphasize particular features in the refinement; for example c = +10 or -10 would weight up data at higher 2θ, e.g. to perform a 'high-angle' refinement (uncontaminated by hydrogen atoms which contribute little at higher diffraction angle) prior to a difference electron density synthesis (FMAP 2) to locate the hydrogens. The exponential weights which are obtained when c is positive were advocated by Dunitz & Seiler (1973). Weighting up the high angle reflections will in general give X-ray atomic coordinates which are closer to those from neutron diffraction.

Refinement against F^2 requires different weights to refinement against F; in particular, making all the weights equal ('unit weights'), although useful in the initial stages of refinement against F, is NEVER a sensible option for F^2. If the program suspects that an unsuitable WGHT instruction has been accidentally retained for a structure which had been refined previously with SHELX-76 or the XLS program in version 4 of the SHELXTL system, it will output a warning message.

`FVAR osf[1] free variables`

The overall scale factor is followed by the values of the 'free variables' fv(2) ... The overall scale factor is given throughout as the square root of the scale factor which multiplies F_c^2 in the least-squares refinement [to make it similar to the scale factor in SHELX-76 which multiplied F_c], i.e. $\mathrm{osf}^2 F_c^2$ is fitted to F_o^2.

SHELXL goes to some trouble to ensure that the initial value of the scale factor has very little influence. Firstly, if the initial scale is exactly 1.0, a quick structure factor summation with a small fraction of the total number of reflections is performed to estimate a new scale factor. If the values differ substantially then the new value is used. Secondly the scale factor is factored out of the least-squares algebra so that, although it is still refined, the only influence the previous value has is an indirect one via the weighting scheme and extinction correction.

Before calculating electron density maps and the analysis of variance, and writing the structure factor file (*name.fcf*), the observed F^2 values and esds are brought onto an absolute scale by dividing by the scale factor.

The free variables allow extra constraints to be applied to the atoms, e.g. for common site occupation factors or isotropic displacement parameters, and may be used in conjunction with the SUMP, DFIX and CHIV restraints. If there is more than one FVAR instruction, they are concatenated; they may appear anywhere between UNIT and HKLF (or END).

7.7 Lists and tables

The esds in bond lengths, angles and torsion angles, chiral volumes, Ueq, and coefficients of least-squares planes and deviation of atoms from them, are estimated rigorously from the full correlation matrix (an approximate treatment is used for the angles between least-squares

standard deviations (from the full covariance matrix). The angle to the previous least-squares plane (if any) is also calculated, but some approximations are involved in estimating its esd. na must be at least 3. If na is omitted the plane is fitted to all the atoms specified.

RTAB codename atomnames

Chiral volumes (one atomname), bonds (two), angles (three) and torsion angles (four atomnames) are tabulated compactly against residue name and number. codename is used to identify the quantity being printed; it must begin with a letter and not be longer than 4

m = 1: Write *h,k,l,* F_o, F_c and phase (in degrees) to .fcf in X-PLOR format. Only unique reflections after removing systematic absences, scaling [to an absolute scale of F(calc)], applying dispersion and extinction or SWAT corrections (if any), and merging equivalents including Friedel opposites are included. If F_o^2 was negative, F_o is set to zero. Reflections suppressed by OMIT or SHEL [or reserved for R(free)] are not included.

m = 2: List *h,k,l,* F_o,

2thetafull is used to specify the value of 2θ for which the program calculates the completeness of the data for the CIF output file as required by Acta Crystallographica. If no value is given, the program uses the maximum value of 2θ for the reflection data. If the data were collected to a specific limiting 2θ, or if a limit was imposed using SHEL, this would be a good choice. Otherwise the choice of 2thetafull is a difficult compromise; if it is too low, the paper will be rejected because the resolution of the data is not good enough; if it is higher, the lower completeness might lead to rejection by the automatic Acta rejection software!

Friedel opposites are merged after the least-squares refinement and analysis of variance but before calculating the Fourier synthesis. This will improve the map (and bring the maximum and minimum residual density closer to zero) compared with SHELX-76. In addition, since usually all the data are employed, reflections with $\sigma(F)$ relatively large compared with F_c are weighted down. This should be better than the use of an arbitrary cutoff on $F_o/\sigma(F)$. The rms fluctuation of the map relative to the mean density is also calculated; in the case of a difference map this gives an estimate of the 'noise level' and so may be used to decide whether individual peaks are significant. Usually FMAP 2 is employed to find missing atoms, but if a significant part of the structure is missing, FMAP 5 or 6 may be better. ACTA requires FMAP 2 so that the difference density is on an absolute scale.

If code is made negative, both positive and negative peaks are included in the list, sorted on the absolute value of the peak height. This is intended to be useful for neutron diffraction data.

code = 2: Difference electron density synthesis with coefficients (F_o–F_c) and phases ϕ(calc).

code = 3: Electron density synthesis with coefficients F_o and phases ϕ(calc).

code = 4: Electron density synthesis with coefficients ($2F_o$–F_c) and phases ϕ(calc). F(000) is included in the Fourier summations for code = 3 and 4.

code = 5: Sim-weighted ($2mF_o$-F_c) Fourier (Giacovazzo, 1992).

code = 6: Sim-weighted ($2mF_o$-F_c) Fourier with coefficients sharpened by multplying with $\sqrt{E/F}$.

`GRID sl[#] sa[#] sd[#] dl[#] da[#] dd[#]`
Fourier grid, when not set automatically. Starting points and increments multiplied by 100. s means starting value, d increment, l is the direction perpendicular to the layers, a is across the

is one with a radius of less than 0.4 Å. Peaks are assigned the radius of SFAC type 1, which is usually set to carbon. Peaks appear on the printout as numbers, but in the .res file they are given names beginning with 'Q' and followed by the same numbers. Peak heights are also written to the .res file (after the sof and dummy U values) in electrons $Å^{-3}$. See also MOLE for forcing molecules (and their environments) to be printed separately.

A default npeaks of +20 is set by FMAP; to obtain line printer plots, an explicit PLAN instruction with negative npeaks is required. If npeaks is positive the nearest unique atoms to each peak are tabulated, together with the corresponding distances. A table of shortest distances between peaks is also produced. For macromolecules and for users of the Siemens' SHELXTL system npeaks will almost always be positive! If npeaks is positive d1 and d2 have a different meaning. The default of d1 is then -1 and causes the full peaklist to appear in the .res file. If it is positive (say 2.3) then the full peaklist is still printed in the *.lst* file, but only suitable candidates for (full occupancy) water molecules appear in the *.res* file (with SFAC 4 and U set to 0.75). The water molecules must be less than 4 Å from an atom which begins with 'O', 'N' or 'W', and may not be nearer than d2 (default 3.0) from any atom which does not begin with 'O', 'N', 'W' or 'H', and may not be nearer than d1 to any 'O', 'N' or 'W' atom or to other potential waters which have larger peak heights. This facility is intended for extending the water structure of proteins in connection with BUMP and SWAT. To include the waters in the next refinement job, their names need to be changed and they need to be moved to before the HKLF instruction at the end of the atom list in the new *.ins* file. This can be performed automatically using SHELXPRO. It is recommended that the last water be called 'LAST' on the ISOR and CONN instructions so that its name does not need to be updated each job.

The heights and positions of the highest (difference) electron density maximum and the deepest minimum are output irrespective of the PLAN parameters.

MOLE n

Forces the following atoms, and atoms or peaks that are bonded to them, into molecule n of the PLAN output. n may not be greater than 99. n = 99 has a special meaning: the 'lineprinter plot' is suppressed for the following atoms, but the table of distances is still printed. This is sometimes useful for saving paper.

8. Strategies for Macromolecular Refinement

SHELXL is designed to be easy to use and general for all space groups and uses a conventional structure-factor calculation rather than a FFT summation; the latter would be faster, but in practice involves some small approximations and is not very suitable for the treatment of dispersion or anisotropic thermal motion. The price to pay for the extra generality and precision is that SHELXL is much slower than programs written specifically for macromolecules, but this is to some extent compensated for by the better convergence properties, reducing the amount of manual intervention required (and also the *R*-factor).

Recent advances in cryogenic techniques, area detectors, and the use of synchrotron radiation enable macromolecular data to be collected to higher resolution than was previously possible. In practice this tends to complicate the refinement because it is possible to resolve finer details of the structure; it is often necessary to model alternative conformations, and in a few cases even anisotropic refinement is justified. Although SHELXL provides a number of other features not found in many macromolecular refinement programs, it is probably the flexible treatment of disorder and the facilities for restrained anisotropic refinement that are most likely to be of immediate interest to macromolecular crystallographers.

An auxiliary program SHELXPRO (Chapter 9) is provided as an interface to other macromolecular programs. SHELXPRO is able to generate an *.ins* file from a PDB format file, including the appropriate restraints etc. SHELXPRO can also generate map files for the program O and can display the refinement results in the form of Postscript plots, as well as including the updated coordinates in the *.ins* file for the next refinement.. SHELXL produces PDB and CIF format files that can be read by SHELXPRO and used for archiving.

calculation of derivatives are larger that those in the structure factors (for the same grid intervals); this would also impede convergence.

8.2 Residues

Macromolecular structures are conventionally divided up into *residues*, for example individual amino-acids. In SHELXL residues may be referenced either individually, by '_' followed by the appropriate residue number, or as all residues of a particular class, by '_' followed by the class. For example 'DFIX 2.031 SG_9 SG_31' could be used to restrain a disulfide distance

The three bonds to a carbonyl carbon atom can be restrained to lie in the same plane by means of a *chiral volume restraint* (Hendrickson & Konnert, 1980) with a target volume of zero (e.g. 'CHIV_GLU 0 C CD' to restrain the carbonyl and carboxyl carbons in all glutamate residues to have planar environments). The planarity restraint (FLAT) restrains the chiral volumes of a sufficient number of atomic tetrahedra to zero; in addition the r.m.s. deviation of the atoms from the best planes is calculated. Chiral volume restraints with non-zero targets are useful to prevent the inversion of α-carbon atoms and the β-carbons of Ile and Thr, e.g. 'CHIV_ILE 2.5 CA CB'. It is also useful to apply chiral volume restraints to non-chiral atoms such as CB of valine and CG of leucine in order to ensure conformity with conventional atom-labeling schemes (from the point of view of the atom names, these atoms could be considered to be chiral !).

Anti-bumping restraints are distance restraints that are only applied if the two atoms are closer to each other than the target distance. They can be generated automatically by SHELXL, taking all symmetry

from 6 to less than 4 for typical organic structures, further restraints are often required for the successful anisotropic refinement of macromolecules.

The *similar ADP restraint* (SIMU) restrains the corresponding U_{ij}-components to be approximately equal for atoms which are spatially close (but not necessarily bonded because they may be in different components of a disordered group). The isotropic version of this restraint has been employed frequently in protein refinements. This restraint is consistent with the characteristic patterns of thermal ellipsoids in many organic molecules; on moving out along side-chains, the ellipsoids become more extended and also change direction gradually.

Neither of these restraints are suitable for isolated solvent (water) molecules. A linear restraint (ISOR) restrains the ADP's to be *approximately isotropic*, but without specifying the magnitude of the corresponding equivalent isotropic displacement parameter. Both SIMU and ISOR

R_{free} is invaluable in deciding whether a restrained anisotropic refinement is significantly better than an isotropic refinement. Experience indicates that both the resolution and the quality of the data are important factors, but that restrained anisotropic refinement is unlikely to be justified for crystals that do not diffract to better than 1.5 Å. An ensemble distribution created by molecular dynamics is an alternative to the harmonic description of anisotropic motion

same atom names, the atoms belonging to different components being distinguished only by their different PART numbers. This procedure enables the standard restraints etc. to be used unchanged, because the same atom and residue names are used. No special action is needed to add the disordered hydrogen atoms, provided that the disorder is traced back one atom further than it is visible (so that the hydrogen atoms on the PART 0 atoms bonded to the disordered components are also correct). Note that this very simple and effective treatment of disorder was not available in the original 1993 release of SHELXL.

8.7 Automatic water divining

It is relatively common practice in the refinement of macromolecular structures to insert water molecules with partial occupancies at the positions of difference electron density map peaks in order to reduce the *R*-factor (another example of ' *R*-factor cosmetics'). Usually when two different determinations of the same protein structure are compared, only the most tightly bound waters, which usually have full occupancies and smaller displacement parameters, are the same in each structure. The refinement of partial occupancy factors for the solvent atoms (in addition to their displacement parameters) is rarely justified by R_{free}, but sometimes the best R_{free} value is obtained for a model involving some water occupancies fixed at 1.0 and some at 0.5.

Regions of diffuse solvent may be modeled using *Babinet's principle* (Moews & Kretsinger, 1975); the same formula is employed in the program TNT, but the implementation is somewhat different. In SHELXL it is implemented as the SWAT instruction and usually produces a significant but not dramatic improvement in the agreement of the very low angle data. Anti-bumping restraints may be input by hand or generated automatically by the program, taking symmetry equivalents into account. After each refinement job, the displacement parameters of the water molecules should be examined, and waters with very high values (say *U* greater than 0.8 Å^2, corresponding to a *B* of 63) eliminated. The F_o-F_c map is then analyzed automatically to find the highest peaks which involve no bad contacts and make at least one geometrically plausible hydrogen bond to an electronegative atom. These peaks are then included with full occupancies and oxygen scattering factors in the next refinement job. This procedure is repeated several times; in general R_{free} rapidly reaches its minimum value, although the conventional *R*-index continues to fall as further waters are added. It should be noted that the automatic generation of anti-bumping restraints is less effective when the water occupancies are allowed to have values other than 1.0 or 0.5. This approach provides an efficient way of building up a chemically reasonable (but not necessarily unique) network of waters that are prevented from diffusing into the protein, thus facilitating remodeling of disordered side-chains etc. The occupancies of specific waters may also be tied (using free variables) to the occupancies of particular components of disordered side-chains where this makes chemical sense. This procedure may be facilitated by using SHELXPRO to convert the *.res* output file from one refinement job to the *.ins* file for the next, or fully automated using the program SHELXWAT that calls SHELXL repeatedly. A similar but much more sophisticated approach (ARP) described by Lamzin & Wilson (1993) may also be used in conjunction with SHELXL.

difference between R_1 and R_{free} was about 3% for jobs 4 to 7 and about 2% for the remaining jobs. Particularly noteworthy is the drop in the *R*-factors on introducing hydrogens (no extra parameters); a parallel job using exactly the same model but excluding hydrogens showed that 1.25% of the drop in R_{free} was contributed by the hydrogens. On the other hand the drop in job 12 is caused almost entirely by the improvements to the model; the same job with the original weights gave an R_{free} of 10.90%. After using R_{free} to monitor the refinement as discussed here, a final refinement was performed against all 80102 unique reflections without any further changes to the model; this converged to R_1 = 8.77%, essentially identical to the final R_1 for the working set.

SHELXL refinement of a serine protease (188 residues)

Job	Action taken	NP	NH	NW/NW½/NX	N_{par}	R_1	R_{free}
1	Final X-PLOR, 1.1-8Å	1337	0	176 / 0 / 19	6129	19.47	21.14
2	Same atoms, SHELXL	1337	0	176 / 0 / 19	6129	17.15	18.96
3	SWAT added	1337	0	176 / 0 / 19	6130	17.07	18.95
4	All atoms anisotropic	1337	0	176 / 0 / 19	13790	12.96	16.10
5	Disorder, added solvent	1376	0	207 / 0 / 34	14565	11.46	14.20
6	More disorder and solvent	1422	0	214 / 2 / 39	14831	11.35	14.22
7	Disorder, half occ. waters	1447	0	213 / 20 / 37	15478	11.13	14.10
8	Resolution: 0.96Å-Inf.	1447	0	218 / 28 / 37	15595	10.75	12.95
9	Riding Hydrogens added	1451	1088	220 / 38 / 40	15769	9.58	11.56
10	Minor adjustments	1477	1052	222 / 48 / 40	16114	9.15	11.19
11	Minor adjustments	1491	1042	211 / 64 / 48	16173	9.29	11.31
12	Weighting changed	1491	1029	222 / 84 / 38	16357	8.74	10.85
13	Further refinement	1499	1025	212 / 96 / 38	16353	8.76	10.79

NP = Number of protein atoms (including partially occupied atoms), NH = Number of hydrogens (all fully occupied), NW = Number of fully occupied waters, $NW_{½}$ = Number of half occupied waters, NX = Number of other atoms (inhibitor, formate, glycerol, some of them partially occupied), and N_{par} = Number of least-squares parameters.

8.10 Summary of useful SHELXL keywords for macromolecular refinement

The more important keywords for macromolecular refinement are summarized in the following table (* indicates significant changes from SHELXL-93):

DEFS Set global restraint esd defaults.

DFIX Restrain 1,2-distance to target (which may be a free variable).

DANG* Restrain 1,3-distance to target (which may be a free variable).

SADI Restrain distances to be equal without specifying target.

SAME Generate SADI automatically for 1,2- and 1,3-distances using connectivity.

CHIV Restrain chiral volume to target (default zero; may be a free variable).

FLAT* Planarity restraint.

DELU Generate rigid bond U_{ij} restraints automatically using connectivity.

SIMU Generate similar U (or U_{ij}) restraints automatically using distances.

ISOR 'Approximately isotropic' restraints.

BUMP* Generate anti-bumping restraints automatically (incl. symm. equivalents).

NCSY* Generate non-crystallographic symmetry restraints.

FVAR Starting values for overall scale factor and free varaibles.

SUMP Restrain linear combination of free variables.

PART Atoms with different non-zero PART numbers not connected by program.

9.1 Outline of the available features

The options provided by SHELXPRO can be divided into three general groups.

(a) Files and communication with other protein programs

[H] *.hkl* file from other data formats. This provides general interactive reformatting of reflection data files, avoiding the need to write a FORTRAN program or UNIX shell-script each time it is necessary to reformat reflection data.

[D] Convert DENZO/SCALEPACK *.sca* to *.hkl*. This is often the safest and quickest way of generating the *.hkl* reflection data file for SHELXL, SHELXS etc.

[V] R(free) files. This adds an R_{free} flag to selected reflections in an *.hkl* file; they may be chosen at random or in thin shells. This is the preferred method of calculating a *free R-factor* using SHELXL, and requires the SHELXL instructions CGLS n –1 or L.S. n –1.

[I] *.ins* from PDB file. This will normally be used when a structure is transferred from another program to SHELXL for the first time. It generates most of the restraints and other extra instructions automatically as well as converting the atoms to fractional coordinates in SHELX format. For editing and updating between SHELXL refinement cycles the following [U] option should be used instead.

[U] Update *.res* (and *.pdb*) to *.ins* file. This should be used to read the *.res* output file from a SHELXL refinement job and update it to create the *.ins* input file for the next job. Alterations such as extra residues or disorder components may be added from a PDB format file written by a graphics program such as O or XtalView.

[G] Generate PDB file from *.res* or *.pdb*. Although SHELXL can write a PDB format file directly, this option provides for more user interaction, e.g. for setting up a PDB format file containing symmetry equivalents or modified temperature factors for use with molecular replacement programs such as AMoRe.

[B] PDB deposition. Collects the information needed for PDB deposition from the *.lst* and *.pdb* files written by SHELXL and creates a file according to the current specifications for deposition with the Brookhaven PDB. The resulting file contains all the compulsory records, but still requires some hand editing e.g. to include information about the data collection..

[F] New output filename. New *.ps* and *.pro* files are started and the previous *.ps* and *.pro* files closed. This enables the Postscript plots to be viewed in another window without leaving SHELXPRO etc.

[C] Color plots (now on). This option toggles color on or off in the Postscript output files. For some journals it may be necessary to produce black and white diagrams rather than color.

[Q] Quit. Terminates SHELXPRO and returns to the command line prompt.

(b) Creation of map (and pdb) files for various graphics packages

[M] Map file for O from *.fcf*. This creates a map file that can be read by O and some versions of FRODO. A variety of maps may be created, including Sigma-A maps. SHELXPRO reads the *.fcf* file written by SHELXL (it contains calculated structure factors and phases) and the *.pdb* file (in order to work out the extent of the map).

[W] Write Turbo-Frodo map. Very similar to the corresponding option for O.

[O] PDB file for O. The otherwise exemplary program O is unfortunately not able to read standard PDB format files (as written by e.g. SHELXL) when they contain disordered groups. This option provides a (not very elegant) work-around.

[X] Write XtalView map coefficients. Writes a *.phs* file with coefficients for various types of map including Sigma-A maps for input to XtalView. XtalView should be instructed to calculate an F_o-map whatever type of map is actually required! This produces MUCH better maps than inputting the atoms from SHELXL as a *.pdb* file into XtalView and repeating the structure factor calculation in XtalView (because of various incompatibilities such as the solvent model, anisotropic temperature factors, complex scattering factors as well as approximations made by XtalView in the structure factor calculation).

(c) Analysis of a structure after refinement with SHELXL

[P] Progress of LS refinement diagram. Produces a diagram of the *R*-factor as a function of the refinement cycle, with special action for automated water divining (SHELXWAT). The *R*-factors are extracted from the REM instructions in the current *.res* file, which are accumulated there when the U option in SHELXPRO is used to update the *.res* file written by one refinement job to create the *.ins* file for the next.

[T] Thermal displacement analysis. Creates bar-plots to show the variation of B-value (and anisotropy) with residue number for main-chain and side-chain atoms.

[R] Ramachandran Phi-Psi plot. A Ramachandran plot is created and the outliers listed. Reads the *.lst* file that must contain the necessary torsion angles calculated in SHELXL using RTAB instructions.

[K] Kleywegt NCS plot. A Kleywegt plot is a Ramachandran plot with NCS-related residues joined by straight lines. The lines cross the edges of the plot and reappear at the other side if necessary. If the plot is too hairy you may be in trouble..

[N] NCS analysis. Creates bar-plots of differences in B-values and various torsion angles between NCS related monomers. These are read from the *.lst* file so the torsion angles should have been calculated using RTAB instructions in SHELXL.

[S] Reflection statistics from *.fcf*. *R*-factors, data completeness, mean(I/sigma) etc. may be calculated for user-specified resolution ranges.

[L] Luzzati plot. Similar to [S] but the resolution ranges are fixed by the program and a Luzzati plot of *R*-factor against resolution is created as well as the statistics.

[E] Esd analysis. Graphical analysis of the esds estimated by a (blocked) full-matrix refinement using SHELXL.

[Z] Least-squares fit. Allows parts of one or more structures to be fitted to each other and r.m.s. deviations calculated. The deviations may be plotted against residue number as bar plots and superimposed structures may be output in suitable format for preparing diagrams

Tis

9.2 Communication with other programs

The various options will now be described in more detail. Much of this information is provided by the program when an option is chosen. This section contains useful information on the best ways of using SHELXL for protein refinements.

[H] .hkl file from other data formats

The program can read a variety of reflection data file formats and write a *.hkl* file in SHELX *.hkl* format. If the original file contained *F*-values, the *.hkl* file should be read into SHELXL with HKLF 3; if the original file contained intensities, HKLF 4 is appropriate. The input file should contain one reflection per line, but lines may be stripped from the beginning and end, e.g. to process data transferred by email. On reading the file, the first line is displayed. To skip this line and move to the next, hit the <Enter> key. To read *h,k,l, F* (or F^2) and $\sigma(F)$ [or $\sigma(F^2)$] from this and subsequent lines in free format, enter the character $*$ followed by <Enter>; to read in fixed format, fill the positions under these quantities with H,K,L,F or S. Thus to read a correctly formatted *.hkl* file, enter the line:

HHHHKKKKLLLLFFFFFFFFSSSSSSSS

For technical reasons, the following option [D] should always be used instead of [H] to read files produced by SCALEPACK.

[D] Convert DENZO/SCALEPACK .sca to .hkl

The SCALEPACK *.sca* and SHELXL *.hkl* formats look very similar, but there are some subtle differences. The .sca file has three lines of header information but *.hkl* has no header. The *.hkl* file may be terminated by a line with all items zero that is not present in the .sca file; however both are also terminated by the end of the file. Unlike *.hkl*, the *.sca* file may contain floating-point numbers in 'I8' format. If the 'anomalous' flag was applied, the .sca file may contain reflections $\mathbf{h}_+$ and $\mathbf{h}_-$ on the same line, with dummy values if not measured. The [D] option handles these differences and may also be used to extract anomalous ΔF values (with esds) for heavy-atom location using Patterson or direct methods in SHELXS.

[V] R(free) files

This command is used to flag say 5 or 10% of the reflections in the *.hkl* file for use as a reference set in calculating free *R*-values (Brünger, 1992). As a rule of thumb, at least 500 reflections or 5% of the total number should be flagged, whichever is larger. It is difficult to obtain statistically meaningful free *R*-values for datasets containing a total of less than 5000 reflections before division into reference and working sets. The flag is applied by making the 'batch number' at the end of each line in the *.hkl* file negative. The unflagged reflections constitute the working set. The *.hkl* file is read into SHELXL in the normal way using HKLF 4 (or 3), and the flags are ignored (i.e. all reflections are used for refinement and no free *R* is calculated) unless the second number on the CGLS (or L.S.) instruction is –1, in which case

only the working set is used for the refinement, and only the reference set is used to calculate the free *R*-values. It is customary to perform the final refinement using all the data, but not increasing the number of independent parameters or reducing the weights of the restraints. This may be done by simply deleting the second number on the CGLS or L.S. instruction.

The reference set may either be chosen at random or in thin shells. The latter option is strongly recommended if a twinned structure is being refined or if NCS restraints are applied, because otherwise the reference and working sets will not be independent. When the

[U] Update .res (and .pdb) to .ins file

This option converts a SHELXL .res file to a new *.ins* file by including new or changed atoms from PDB format files such as those written by the graphics programs O, Turbo-Frodo and XtalView. All other SHELXL commands are retained unchanged. This option also provides for setting up disorder refinement and updating the list of solvent molecules. The .res file should not contain instructions other than RESI, AFIX, PART and atoms between FVAR and HKLF, and both FVAR and HKLF must be present. Note that although it is possible to set up

For example, residue 1001 in this example would become chain A residue 1. Similarly, residue 2189 becomes chain B residue 189. The solvent water that used to start at residue 1 now starts at residue 201.

For the deposition of reflection data, the CIF format *.fcf* file written by SHELXL may be used directly.

9.3 Creation of map (and pdb) files for various graphics packages

In a computer utopia, interactive graphics packages would all read the CIF format *.fcf* file written by SHELXL directly; this contains all the information necessary for generating maps. For the couple of years before this comes to pass, SHELXPRO provides the necessary generation of maps or (in the case of XtalView) coefficients. For the programs O and Turbo-Frodo, it is also necessary to define the region of space for which the map is calculated; SHELXPRO does this by scanning a PDB file to find the maximum and minimum atomic coordinates in each direction. Furthermore, O is liable to be confused by disordered residues even if these are specified exactly according to the PDB rules (as SHELXL does), so it is also necessary for SHELXPRO (option [O]) to be able to modify the PDB file so that all disorder components are given separate residue numbers. Note that the option [U] provides the reverse procedure, i.e. separate residues obtained using O may be recombined as different disorder components of the same residue for refinement using SHELXL. SHELXPRO does not make the changes that may be required to the *all.dat* connectivity file read by O.

The [M], [O], [W] and [X] options should be self-explanatory. The following questions are asked by the program; usually the answers suggested by the program are suitable, so most of the questions are answered by <Enter>.

Name of .fcf file created using SHELXL and LIST 6 [name.fcf]:

Enter name of PDB file [name.pdb]:

Include all waters in the volume covered by map? [Y]:

Number of grid points per cell in x, y and z (the first two MUST be powers of 2, and the last MUST be a multiple of 8) [64 64 88]:

Origin of map along x, y and z (grid points) [-32 -24 24] (must all be multiples of 8):

Extent of map along x, y and z (grid points) [128 136 88] (must all be multiples of 8):

Fourier type (-3=mFo-DFc (Sigma-A difference map), -2=2mFo-DFc (Sigma-A map), -1=Fo-Fc, 0=Fc, 1=Fo, 2=2Fo-Fc, n=nFo-(n-1)Fc [-2]:

Enter reference/working set Sigma-A ratio from SHELXL [0.97]:

Apply sharpening (Y or N) ? [N]:

Enter name of map file [sigmaa.map]:

For XtalView, the questions about the grid are skipped. Note that there is a choice of maps. Thus the input '3' for the Fourier type generates a $3F_o$-$2F_c$ map; '4' gives a $4F_o$-$3F_c$ map, etc. The sigma-A ratio is calculated in each SHELXL job that uses the free *R*-factor; it is designed

to correct the sigma-A weight for overfitting. For refinement at low resolution this might be about 0.8, for medium resolution 0.9; the default is appropriate for structures with a high ratio of data to parameters. If the free *R*-factor was not used in the refinement, a estimated value should be input. 'Sharpening' multiplies the coefficients by $\langle F^2 \rangle^{¼}$, where $\langle F^2 \rangle$ is the mean reflection intensity in the appropriate resolution shell (this factor is used in preference to the almost identical factor $\sqrt{(E/F)}$ because the latter involves a statistical factor for certain reflections that is inappropriate for this application). Finally, the program outputs the maximum and minimum electron density (in

information) and the side-chain plots by residue type. The color schemes are defined in the .pro output file.

Alpha-helices and beta-strands are entered one per line with 'A n1 n2' or 'B n1 n2' respectively, where n1 and n2 are the first and last residues of the helix or strand. The letters may be upper or lower case. The list is terminated with a blank line. Thus:

a 21 45
b 48 55
a 67 108

would define two alpha-helices (residues 21 to 45 and 67 to 108 resp.) and one beta-strand (48 to 55). The alpha-helix regions are colored blue, the beta-strands green, and the rest red. There may be up to four diagrams on one page, starting at the top. Each should be defined by entering three characters: a symbol to label the diagram, then either B (B-values) or A (anisotropy), followed by M (main-chain) or S (side-chain) and then the numbers of the first and last residues. END terminates the list. The program will suggest suitable parameters. A typical sequence, selecting these defaults by <Enter> each time, would be:

Next diagram [aBM 1 204]:

Maximum value and step for vertical scale [50 10]:

Next diagram [bAM 1 204]:

Next diagram [cBS 1 204]:

Maximum value and step for vertical scale [60 10]:

Next diagram [dAS 1 204]:

Note that no scale needs to be specified for the anisotropy, because the range is always from 0 to 1.

[R] Ramachandran Phi-Psi plot

The [R] option reads the SHELXL .lst output file and extracts the psi and phi torsion angles to make Ramachandran plots. If the main-chain is disordered, only the PART 1 (and of course PART 0) atoms are used. Glycines are included optionally as open squares; prolines are treated as normal residues. A list of outliers appears on the screen and in the *.pro* file. Residues are color-coded according to residue type unless black and white Postscript has

[K] Kleywegt NCS plot

This is the same as the normal Ramachandran plot (option [R] above) except that the phi/psi dots for each residue are smaller and residues related by non-crystallographic symmetry (NCS) are joined by lines (Kleywegt, 1996). The lines may cross the edges of the plot and reappear at the other side if this makes the differences between the angles smaller. Ramachandran outliers (as defined by Kleywegt and Jones) are also reported. This plot gives an immediate indication of how well NCS is obeyed for the main-chain atoms, and is also a good indicator of the overall quality of the structure. If the main-chain is disordered, only PART 0 and PART 1 atoms are considered. Glycines are optionally included as open squares; prolines are treated as normal residues. Unless color has been switched off (option [C]) the dots and lines are color-coded according to residue type. The refinement should have been performed with FMAP 2 and the RTAB instructions needed to calculate the

The default diagrams are:

aMH (diagram a; maximum absolute deviation of phi angles)
bMY (diagram b; maximum absolute deviation of psi angles)
cMO (diagram c; maximum absolute deviation of omega angles)
dMT (diagram d; maximum absolute deviation of all chi angles)
eMM (diagram e; maximum absolute deviation of main-chain B)
fMS (diagram f; maximum absolute deviation of side-chain B)
gAM (diagram g; average main-chain B)
hAS (diagram h; average side-chain B)

[S] Reflection statistics from .fcf

This option creates reflection statistics from a *.fcf* file written by SHELXL in response to a LIST 6 instruction.. The user must specify the resolution ranges, e.g. to be the same as those used for data reduction. A table of data completeness, *R*-factors etc. is written to the console and to the .pro output file.

[L] Luzzati plot

This plots the resolution vs. *R*1. The *.fcf* file must have been created using LIST 6 in SHELXL. SHELXPRO outputs a Postscript Luzzati (1952) plot, which gives estimates of the average errors in atomic coordinates for an incompletely refined structure assuming perfect data, NOT (as widely assumed by people who have not read this paper which happens to be in French) estimates of the esds in the atomic positions. For small proteins and high resolution data, esds in individual bond lengths and atomic positions may be estimated rigorously using SHELXL (see the [E] option in SHELXPRO described below). Nevertheless, a plot of *R*-factor against resolution is always entertaining.

[E] Esd analysis

This option reads SHELXL *.lst* file and prepares Postscript scatter-plots of esds in atom positions and n the at e9 ues. The refinemtht should normally
have been performed with the SHELXL instructions L.S. 1, DAMP 0 0, BLOC 1 and BOND. If geometrical restrL lts were used in the refinemtht the n the a esds will be very low, but
high resolution data are re8thred to perform such a refinemtht without restrL lts. Similarly the damping has to be switched off becautheeirthisnamesinatbe(tad to)-20(underestimated)-20(es

[Z] Least-squares fit

The [Z] option may be used to perform a least-squares fit of two molecules, taken from the same or different structures. The iterative quaternion method is employed. This option is of necessity rather complex, and it is important to read each request for information by the program carefully because the default action (<Enter>) may well not be suitable and an incorrect answer can lead to complications.

It is necessary first to define the first molecule (called 'current structure'), which is extracted from a PDB format file. The a second molecule ('model') is obtained from another (or possibly the same) PDB file. Both PDB files may be as output by SHELXL or may be taken directly from the PDB databank, so 'chains' may be present. Since the residues may be numbered differently in the two molecules, it is necessary to convert the residue numbers in both molecules to a matching set of residue numbers referred to as SHELXPRO residue numbers. These numbers are also used to annotate the plots etc. The set of residues used for fitting is in general a subset of those used for the plots and calculation of r.m.s. esds.

After performing the fit for specified atoms in each of the specified residues, the program prints the r.m.s. deviation of the atoms fitted and the largest individual deviations (greater than 2σ). Then appears the question:

New current structure (C), new model (M), Repeat fit (R), write PDB file (P), XP file (X), Postscript bar plot of differences (D) or exit (E) [E]:

'R' repeats the fit (possibly using different residues and atoms) of the 'model' (second molecule) to the 'current structure' (first molecule). 'M' replaces the 'model' but keeps the 'current structure'.. 'C' starts again with a new 'current structure'. 'P' writes a new PDB format file that contains the two molecules as two separate chains with the SHELXPRO residue numbers; this can be used as input to the program MOLSCRIPT. 'X' writes an orthogonal coordinate file that can be read by the Siemens' SHELXTL program XP and used to make a (stereo) Cα-trace of the superposition. 'D' prepares a Postscript bar plot of the differences between the two molecules, using all stored residues, not just those that were fitted.

[A] Anisotropic scaling (Hope & Parkin)

This option reads an *.fcf* file created using the LIST 6 instruction in SHELXL, and writes a NEW *.hkl* file after application of anisotropic scaling by the method of Parkin, Moezzi & Hope (1995). The modification of the observed structure factors in this way is scientifically suspect and is intended for testing purposes only. It is much better to use the HOPE instruction in SHELXL so that parameter correlations are taken into account and the observed data are not modified. The SHELXPRO correction provides a quick test as to whether HOPE in SHELXL will result in a significant improvement; in this case the question about the filename for corrected data should be answered with <Enter>. A 'local' R_{free} test is applied to establish how many parameters [none(!), 12, 18 or 24] may justifiably be fitted. A significant improvement is not to be expected if anisotropic refinement has been performed or if a large number of symmetry equivalents were merged in the data reduction.

10. SHELXWAT: Automated Water Divining

A simple program **SHELXWAT** has been added that iteratively recycles SHELXL to provide automatic water divining. This may be regarded as a cheap and inadequate imitation of ARP (V. Lamzin & K.S. Wilson, *Acta Cryst.* **D49** (1993) 129-147), but is relatively easy to use and useful if you intend to take a holiday. SHELXWAT is started by means of a command line with OPTIONAL UNIX-type switches (the filename must come last):

```
shelxwat name
```

or e.g.

```
shelxwat -n10 -s4 -u0.6 -r0.8 -m50 -f name
```

These are the default settings for the switches -n (number of overall cycles), -s (scattering factor number for oxygen), -u (starting isotropic U for new waters), -r (water rejected if U refines to greater than this value), -m (maximum number of waters to be added in one cycle) and -h (half/full occupancies) or -f (full occupancies only). All switches present must come before 'name'.

Standard SHELXL files *name.ins* and name.hkl are required; the *.ins* file should contain 'CGLS 3 -20', 'FMAP 2', 'PLAN 200 2.4' or 'PLAN 200 -2.4' (half occupancies allowed), 'CONN 0 O_501 > LAST', 'BUMP' or similar instructions (the free R test is not obligatory) and MUST include at least one water at the end of the atom list. The waters will then be assigned dynamical residue numbers starting with the residue number of this water (501 in the above example) and should all have residue class 'HOH' and atom name 'O' with one atom per residue and no PART numbers. On starting, SHELXWAT makes a backup copy (*name.bak*) of the *.ins* file, since the *.ins* file is repeatedly overwritten during the recycling. The recycling may be terminated tidily before the preset number of iterations has been performed by creating a file *name.end* in the same directory; this operates like the *name.fin* file for SHELXL, but is 'deleted' by SHELXWAT once per iteration.

SHELXWAT calls SHELXL once each cycle, then edits the resulting *.res* file to prepare the *.ins* file for the next cycle. The *R*1 (and $R1_{free}$, if present) indices are extracted from the *.lst* file and included in the *.res* files as remarks; these and other remarks (REM) provide a protocol of the refinement, and may be converted to a Postscript plot using the "P" option in SHELXPRO. Note that the SHELXPRO option "U" provides the facilities necessary to update that solvent etc. interactively, in much the same way that SHELXWAT does automatically.

By changing the PLAN instruction to (say) 'PLAN 200 1 1' and leaving out the BUMP instruction it might be possible to emulate ARP in its structure extension mode; this has yet to be tested, but might be useful for completing high resolution (better than 2Å) structures.

11. Examples of Macromolecular Refinement

The following extracts from the file *6rxn.ins* (provided together with *6rxn.hkl* as an example) illustrate a number of points. The structure was determined by Stenkamp, Sieker & Jensen, (1990) who have kindly given permission for it to be used in this way. As usual in *.ins* files, comments may be included as REM instructions or after exclamation marks. The resolution of 1.5Å does not quite justify refinement of all non-hydrogen atoms anisotropically ('ANIS' before the first atom would specify this), but the iron and sulfur atoms should be made anisotropic as shown below. Note that it would be better to flag the R_{free} reflections randomly using

```
REM Special restraints etc. specific to this structure follow:

REM HFIX 43 C1_1       !
DFIX C1_1 N_1 1.329    ! O=C(H)- (formyl) on N-terminus
DFIX C1_1 O1_1 1.231   ! incorporated into residue 1
DANG N_1 O1_1 2.250    !
DANG C1_1 CA_1 2.435   !

DFIX_52 C OT1 C OT2 1.249     !
DANG_52 CA OT1 CA OT2 2.379   ! Ionized carboxyl at C-terminus
DANG_52 OT1 OT2 2.194         !

SADI_54 0.04 FE SG_6 FE SG_9 FE SG_39 FE SG_42 ! Equal but unknown Fe-
S
SADI_54 0.08 FE CB_6 FE CB_9 FE CB_39 FE CB_42 ! distances around Fe

REM HFIX 83 SG_38 SG_138  ! -SH for remaining cysteine (disordered)

DFIX C_18 N_26 1.329            ! Patch break in numbering - residues
DANG O_18 N_26 2.250            ! 18 and 26 are bonded but there is a
DANG CA_18 N_26 2.425           ! gap in numbering for compatibility
DANG C_18 CA_26 2.435           ! with other rubredoxins that have an
FLAT 0.3 O_18 CA_18 N_26 C_18 CA_26  ! extra loop
RTAB Omeg CA_18 C_18 N_26 CA_26      !
RTAB Phi C_18 N_26 CA_26 C_26        !
RTAB Psi N_18 CA_18 C_18 N_26        !

REM DFIX from CSD and R.A.Engh & R.Huber, Acta Cryst. A47 (1991) 392.
REM Remove 'REM ' before HFIX to activate H-atom generation

REM HFIX_ALA 43 N
REM HFIX_ALA 13 CA
REM HFIX_ALA 33 CB

REM HFIX_ASN 43 N
REM HFIX_ASN 13 CA
REM HFIX_ASN 23 CB
REM HFIX_ASN 93 ND2

REM HFIX_ASP 43 N
REM HFIX_ASP 13 CA
REM HFIX_ASP 23 CB

   ... etc ...

REM HFIX_VAL 43 N
REM HFIX_VAL 13 CA CB
REM HFIX_VAL 33 CG1 CG2

REM Peptide standard torsion angles and restraints
```

```
RTAB_* Omeg CA C N_+ CA_+
RTAB_* Phi C_- N CA C
RTAB_* Psi N CA C N_+
RTAB_* Cvol CA

DFIX_* 1.329 C_- N
DANG_* 2.425 CA_- N
DANG_* 2.250 O_- N
DANG_* 2.435 C_- CA

FLAT_* 0.3 O_- CA_- N C_- CA

REM Standard amino-acid restraints etc.

CHIV_ALA C
CHIV_ALA 2.477 CA

DFIX_ALA 1.231 C O
DFIX_ALA 1.525 C CA
DFIX_ALA 1.521 CA CB
DFIX_ALA 1.458 N CA
DANG_ALA 2.462 C N
DANG_ALA 2.401 O CA
DANG_ALA 2.503 C CB
DANG_ALA 2.446 CB N

RTAB_ASN Chi N CA CB CG

CHIV_ASN C CG
CHIV_ASN 2.503 CA

DFIX_ASN 1.231 C O  CG OD1
DFIX_ASN 1.525 C CA
DFIX_ASN 1.458 N CA
DFIX_ASN 1.530 CA CB
DFIX_ASN 1.516 CB CG
DFIX_ASN 1.328 CG ND2
DANG_ASN 2.401 O CA
DANG_ASN 2.462 C N
DANG_ASN 2.455 CB N
DANG_ASN 2.504 C CB
DANG_ASN 2.534 CA CG
DANG_ASN 2.393 CB OD1
DANG_ASN 2.419 CB ND2
DANG_ASN 2.245 OD1 ND2

RTAB_ASP Chi N CA CB CG

CHIV_ASP C CG
```

```
CHIV_ASP 2.503 CA

DFIX_ASP 1.231 C O
DFIX_ASP 1.525 C CA
DFIX_ASP 1.530 CA CB
DFIX_ASP 1.516 CB CG
DFIX_ASP 1.458 CA N
DFIX_ASP 1.249 CG OD1 CG OD2
DANG_ASP 2.401 O CA
DANG_ASP 2.462 C N
DANG_ASP 2.455 CB N
DANG_ASP 2.504 C CB
DANG_ASP 2.534 CA CG
DANG_ASP 2.379 CB OD1 CB OD2
DANG_ASP 2.194 OD1 OD2

RTAB_CYS Chi N CA CB SG

CHIV_CYS C
CHIV_CYS 2.503 CA

DFIX_CYS 1.231 C O
DFIX_CYS 1.525 C CA
DFIX_CYS 1.458 N CA
DFIX_CYS 1.530 CA CB
DFIX_CYS 1.808 CB SG
DANG_CYS 2.401 O CA
DANG_CYS 2.504 C CB
DANG_CYS 2.455 CB N
DANG_CYS 2.462 C N
DANG_CYS 2.810 CA SG

  ... etc ...

RTAB_VAL Chi N CA CB CG1
RTAB_VAL Chi N CA CB CG2

CHIV_VAL C
CHIV_VAL 2.516 CA

DFIX_VAL 1.231 C O
DFIX_VAL 1.458 N CA
DFIX_VAL 1.525 C CA
DFIX_VAL 1.540 CA CB
DFIX_VAL 1.521 CB CG2 CB CG1
DANG_VAL 2.401 O CA
DANG_VAL 2.462 C N
DANG_VAL 2.497 C CB
DANG_VAL 2.515 CA CG1 CA CG2
DANG_VAL 2.479 N CB
```

```
DANG_VAL 2.504 CG1 CG2

WGHT    0.100000
FVAR       1.00000   0.5   0.5   0.5   0.5

RESI   1  MET
C1      1  -0.01633   0.35547   0.44703  11.00000   0.11817
O1      4   0.01012   0.32681   0.48491  11.00000   0.17896

N       3   0.00712   0.35446   0.37983  11.00000   0.11863
CA      1   0.05947   0.33273   0.35391  11.00000   0.06229
CB      1   0.07411   0.33732   0.27909  11.00000   0.15678
CG      1   0.03196   0.28864   0.22872  11.00000   0.14569
SD      5   0.04907   0.31846   0.14359  11.00000   0.23570
CE      1   0.11380   0.29170   0.12261  11.00000   0.21476
C       1   0.10634   0.38738   0.39766  11.00000   0.09178
O       4   0.10329   0.45513   0.41972  11.00000   0.16480
RESI   2  GLN
N       3   0.14741   0.35678   0.40741  11.00000   0.08599
CA      1   0.18940   0.39931   0.45565  11.00000   0.09291
CB      1   0.22933   0.34643   0.45886  11.00000   0.13253
CG      1   0.27354   0.38674   0.51173  11.00000   0.09866
CD      1   0.24547   0.38838   0.58387  11.00000   0.05748
OE1     4   0.22482   0.32772   0.60689  11.00000   0.16301
NE2     3   0.24704   0.46053   0.62045  11.00000   0.10164
C       1   0.22198   0.47895   0.43826  11.00000   0.08193
O       4   0.25019   0.48377   0.38408  11.00000   0.10402

RESI   3  LYS
N       3   0.21781   0.54034   0.48673  11.00000   0.07413
CA      1   0.25088   0.62006   0.47934  11.00000   0.05181
CB      1   0.21991   0.68311   0.51795  11.00000   0.09646
CG      1   0.16130   0.66288   0.49255  11.00000   0.10455
CD      1   0.12843   0.72146   0.52924  11.00000   0.22324
CE      1   0.10532   0.70085   0.60053  11.00000   0.26354
NZ      3   0.05943   0.74195   0.62796  11.00000   0.40338
C       1   0.30678   0.63497   0.50917  11.00000   0.05714
O       4   0.31462   0.59598   0.55179  11.00000   0.07986

  ... etc ...

RESI  12  GLU
N       3   0.41413   1.09215   0.48246  11.00000   0.06790
CA      1   0.37955   1.01183   0.48195  11.00000   0.05761
PART   1
CB      1   0.32666   1.01321   0.52971  21.00000   0.12219
CG      1   0.29679   0.93111   0.54638  21.00000   0.15333
CD      1   0.25357   0.93709   0.60700  21.00000   0.20272
OE1     4   0.24346   1.00278   0.63210  21.00000   0.26315
OE2     4   0.23012   0.87537   0.63031  21.00000   0.21375
```

```
RESI 201  HOH
O       4   0.13450   0.53192   0.60802  11.00000   0.13132
RESI 202  HOH
O       4   0.84795   0.53873   0.69488  11.00000   0.15273
RESI 203  HOH
O       4   0.27771   0.95750   0.25086  11.00000   0.11315
RESI 204  HOH
O       4   0.37066   0.71872   0.90376  11.00000   0.10854

   ... etc ...

RESI 233  HOH
O       4   0.27813   1.38725   0.25914  11.00000   0.10698

HKLF 3
END
```

12. SHELXS - Structure Solution

SHELXS is primarily designed for the solution of 'small moiety' (1-200 unique atoms) structures from single crystal at atomic resolution, but is also useful for the location of heavy atoms from macromolecular isomorphous or anomalous ΔF data. The use of the program with SIR, OAS or MAD F_A data is described in Chapter 15. SHELXS is general and efficient for all space groups in all settings, and there are no arbitrary limits to the size of problems which can be handled, except for the total memory available to the program. Instructions and data are taken from two standard (ASCII) text files, compatible to those used for SHELXL, so that input files can easily be transferred between different computers.

12.1 Program and file organization

The way of running SHELXS and the conventions for filenames will of course vary for different computers and operating systems, but the following general concept should be adhered to as much as possible. SHELXS may be run on-line by means of the command:

shelxs name

where *name* defines the first component of the filename for all files which correspond to a particular crystal structure. On some systems, *name* may not be longer than 8 characters. On UNIX systems, all filenames (including SHELXS) MUST be given in ***lower case***. Batch operation will normally require the use of a short batch file containing the above command etc.

Before starting SHELXS, at least one file - *name.ins* - MUST have been prepared; it contains instructions, crystal and atom data etc. It will usually be necessary to prepare a *name.hkl* file as well which contains the reflection data; the format of this file (3I4,2F8.2) is the same as for all versions of SHELX. This file should be terminated by a record with all items zero. The reflection order is unimportant. This .*hkl* file is read each time the program is run; unlike SHELX-76, there is no facility for intermediate storage of binary data. This enhances computer independence and eliminates several possible sources of confusion. SHELXS requires a single set of input data, and ignores batch numbers, direction cosines or wavelengths if they are present at the end of each record in the *name.hkl* file.

A brief summary of the progress of the structure solution appears on the console (i.e. the standard FORTRAN output), and a full listing is written to a file *name.lst*, which can be printed or examined with a text editor. After structure solution a file *name.res* is written; this contains crystal data etc. as in the *name.ins* file, followed by potential atoms. It may be copied or edited to *name.ins* for structure refinement using SHELXL or partial structure expansion with SHELXS (Chapter 14).

Two mechanisms are provided for interaction with a SHELXS job which is already running. The first, which it is not possible to implement for all computer systems, applies to 'on-line' runs. If the <ctrl-I> key combination is hit, the job terminates almost immediately (but without the loss of output buffers etc. which can happen with <ctrl-C> etc.). If the <Esc> key is hit during direct methods, the program does not generate any further phase permutations but completes the current batch of phase refinement and then procedes to *E*-Fourier recycling etc. If the <Esc> key is hit during Patterson interpretation, the program stops after completing the

calculations for the current superposition vector. Otherwise <Esc> has no effect. On computer consoles with no <Esc> key, <F11> or <Ctrl-[> usually have the same effect.

The second mechanism requires the user to create the file *name.fin*; the program tries at regular intervals to delete this file, and if it succeeds it takes the same

All instructions commence with a four (or less) letter word (which may be an atom name); numbers and other information follow in free format, separated by one or more spaces. Upper and lower case input may be freely mixed; with the exception of the text strings input using TITL it is all converted to upper case for internal use in SHELXS. The TITL, CELL, ZERR, LATT, SYMM, SFAC and UNIT instructions must be given in that order; all remaining instructions, atoms, etc. should come between UNIT and the last instruction, which is almost always HKLF (to read in reflection data).

Defaults are given in square brackets in this documentation; '#' indicates that the program will generate a suitable default value based on the rest of the available information. Continuation lines are flagged by '=' at the end of a line, the instruction being continued on the next line which must start with at least one space. Other lines beginning with one or more spaces are treated as comments, so blank lines may be added to improve readability. All characters following '!' or '=' in an instruction line are ignored, except after TITL or SYMM (for which continuation lines are not allowed). AFIX, RESI and PART instructions may be present in the .*ins* file for compatibility with SHELXL but are ignored.

12.3 Instructions common to all modes of structure solution

TITL [ ]
Title of up to 76 characters, to appear at suitable places in the output. The characters '!' and '=' may form part of the title. The title could include a chemical formula and/or space group, but one must be careful to update these if the UNIT or SYMM instructions are later changed !

CELL λ **a b c** α β γ
Wavelength and unit-cell dimensions in Angstroms and degrees.

ZERR Z esd(a) esd(b) esd(c) esd(α**) esd(**β**) esd(**γ**)**
Z value (number of formula units per cell) followed by the estimated errors in the unit-cell dimensions. This information is not actually required by SHELXS but is allowed for compatibility with SHELXL.

LATT N [1]
Lattice type: 1=P, 2=I, 3=rhombohedral obverse on hexagonal axes, 4=F, 5=A, 6=B, 7=C. N must be made negative if the structure is non-centrosymmetric.

SYMM symmetry operation
Symmetry operators, i.e. coordinates of the general positions as given in International Tables.

preceded by '$' but this is not obligatory (the '$' character is allowed for logical consistency with certain SHELXL instructions but is ignored). The program uses absorption coefficients from International Tables for Crystallography (1991), Volume C. For organic structures the first two SFAC types should be C and H, in that order; the E-Fourier recycling generally assigns the first SFAC type (i.e. C) to peaks.

SFAC a1 b1 a2 b2 a3 b3 a4 b4 c df' df" mu r wt

Scattering factor in the form of an exponential series, followed by real and imaginary corrections, linear absorption coefficient, covalent radius and atomic weight. Except for the atomic weight the format is the same as that used in SHELX-76. In addition, a 'label' consisting of up to 4 characters beginning with a letter (e.g. Ca2+) may be included before a1 (the first character may be a '$', but this is not obligatory). The two SFAC formats may be used in the same *.ins* file; the order of the SFAC instructions (and the order of element names in the first type of SFAC instruction) define the scattering factor numbers which are referenced by atom instructions. Not all numbers on this instruction are actually used by SHELXS, but the full data must be given for compatibility with SHELXL. For neutron data, c should be the scattering length (which may be negative) and a1..b4 will usually all be zero.

UNIT n1 n2 ...

Number of atoms of each type in the cell, in SFAC order.

REM

Followed by a comment on the same line. This comment is ignored by the program but is copied to the results file (*.res*). Note that comments beginning with one or more blanks are only copied to the *.res* file if the line is completely blank; REM comments are always copied.

MORE verbosity [1]

More sets the amount of (printer) output; verbosity takes a value in the range 0 (least) to 3 (most verbose).

TIME t [#]

If the time t (measured in seconds from the start of the job) is exceeded, SHELXS performs no further blocks of phase permutations (direct methods), but goes on to the final *E*-map recycling etc. In the case of Patterson interpretation, no further vector superpositions are performed after this time has expired. The default value of t is installation dependent, and is usually set to a little less than the maximum time allocation for a particular job class. Usually t is 'CPU time', but on some simpler computer systems (eg. MSDOS) the elapsed time has to be used instead.

OMIT s [4] 2θ(lim) [180]

Thresholds for flagging reflections as 'unobserved'. Note that if no OMIT instruction is given, ALL reflections are treated as 'observed'. Internally in the program s is halved and applied to F_o^2, so the test is roughly equivalent to suppressing all reflections with $F_o < s\sigma(F_o)$, as required for consistency with SHELX-76. Note that s may be set to 0 (to suppress reflections with negative F_o^2) or even to a negative threshold (to suppress very negative F_o^2) which has no equivalent in SHELX-76. If 2θ(lim) is POSITIVE, it specifies a 2θ value above which the data are treated as 'unobserved'; if it is negative, the absolute value is used as a lower 2θ cutoff.

OMIT h k l

The reflection h k l is flagged as 'unobserved' in the list of merged reflections after data reduction. It will not be used directly in phase refinement or Fourier calculations, but is

FMAP code [#] axis [#] nl [#]
The unique unit of the cell for performing the Fourier calculation is set up automatically unless specified by the user using FMAP and GRID. The program chooses a 53 x 53 x nl or 103 x 103 x nl grid depending the the resolution of the data, provided sufficient memory is available in the latter case.

code = 1 (F^2-Patterson), 3 (Patterson with coefficients input using HKLF 7; negative coefficients are allowed. 4 (*E*-map without peak-list optimisation, e.g. because the peaks correspond to unequal atoms), 5 (Fourier with *A* and *B* coefficients input using HKLF 3), 6 (*EF* Patterson), code > 6 (E-map followed by [code–6] cycles peak-list optimization). Note that the peak-list optimization assigns very strong peaks to heavy atoms (if specified by SFAC) and all remaining peaks to scattering factor type 1, so for many structures this should be specified as carbon on a SFAC instruction. FMAP 4 may be used with atoms but without TEXP etc. for an *E*-map based on calculated phases.

GRID sl [#] sa [#] sd [#] dl [#] da [#] dd [#]
Fourier grid, when not set automatically. Starting points and increments are multiplied by 100. s means starting value, d increment, l is the direction perpendicular to the layers, a is across the paper from left to right, and d is down the paper from top to bottom. Note that the grid is 53 x 53 x nl points, i.e. twice as large as in SHELX-76, and that sl and dl need not be integral. The 103 x 103 x nl grid is only available when it is set automatically by the program (see above).

PLAN npeaks [#] d1 [0.5] d2 [1.5]
If npeaks is positive it is the number of highest unique Fourier peaks which are written to the *.res* and *.lst* files; the remaining parameters are ignored. If npeaks is given as negative, the program attempts to arrange the peaks into unique molecules taking the space group symmetry into account, and to 'plot' a projection of each such molecule on the printer (i.e. the .lst file). Distances involving peaks which are less than r1+r2+d1 (the covalent radii r are defined via SFAC; 1 and 2 refer to the two atoms concerned) are considered to be 'bonds' for purposes of the molecule assembly and tables. Distances involving atoms and/or peaks which are less than r1+r2+d2 are considered to be 'non-bonded interactions'. Such interactions are ignored when defining molecules, but the corresponding atoms and distances are included in the line-printer output. Thus an atom may appear in more than one map, or more than once on the same map. Negative d2 includes hydrogen atoms in these non-bonds, otherwise they are ignored (the absolute value of d2 is used in the test). Peaks are always always assigned the radius of SFAC type 1, which is usually set to carbon. Peaks appear on the printout as numbers, but in the *.res* file they are given names beginning with 'Q' and followed by the same numbers.

To simplify interpretation of the lineprinter plots, extra symmetry-generated atoms are added, so that atoms or peaks may appear more than once. A table of the appropriate coordinates and symmetry transformations appears at the end of the output. See also MOLE for forcing molecules (and their environments) to be printed separately.

MOLE n [#]
Forces the following atoms, and atoms or peaks that are bonded to them, into molecule n of the PLAN output. n may not be greater than 99.

HKLF n [0] s [1] r11...r33 [1 0 0 0 1 0 0 0 1] wt [1] m [0]

Before running SHELXS, a reflection data file *name.hkl* must usually be prepared. The HKLF command tells the program which format has been chosen for this file, and allows the indices

PSEE m [200] 2θ(max) [#]

The largest |m| *E*-values and the complete Patterson map are dumped into the *name.res* file in fixed format for use by Patterson search programs (in particular PATSEE) etc. 2θ(max) should be used to limit the resolution of the *E*-values generated; the default value uses sinθ= λ/2. The 2θ(max) value is also written to the *.res* file, so it is possible to restrict the resolution of the *E*-values actually used by PATSEE to a lower 2θ(max) by editing this file without rerunning SHELXS; of course the *E*se by Pattn maittbmit ad(grams 128fı0.52 0 TDı-0.00from(m

13. Structure Solution by Direct Methods

13.1 Routine structure solution

Usually direct methods will be initiated with the single SHELXS command TREF; for large structures brute force (e.g. TREF 5000) may prove necessary. In fact there are a large number of parameters which can be varied, though the program is based on experience of many thousands of structures and can usually be relied upon to choose sensible default values. A summary of these parameters appears after the data reduction output, and should be consulted before attempting any direct methods options other than 'TREF n'.

'temperature' is higher; this corresponds to a larger value of Boltz), and the sign is chosen to give the best agreement with the negative quartets (if there are no negative quartets involving the reflection in question, a random sign is used instead). After each cycle through all ns phases, a new value for Boltz is obtained by multiplying the old value by cool; this corresponds to a reduction in the 'temperature'. To save time, only ns reflections are refined using the strongest mtpr triplets and mnqr quartets for each reflection (or less, if not so

reflections are unreliable (i.e. have high standard deviations), e.g. because data were collected using the default options on a CAD-4 diffractometer, then the NQUAL figure of merit is less reliable. If the space group does not possess translation symmetry, it is essential to obtain good negative quartets, i.e. to measure ALL reflections for an adequate length of time.

Only the TREF instruction is essential to specify direct methods; appropriate INIT, PHAN, FMAP, GRID and PLAN instructions are then generated automatically if not given.

13.3 What to do when direct methods fail

If direct methods fail to give a clearly correct

Å range with the number theoretically possible (assuming that OMIT is at its default setting of 4) as printed out by the program. If this ratio is less than one half, it is unlikely that the structure will be ever be solved by direct methods. This criterion may be relaxed somewhat for centrosymmetric structures and those containing heavy atoms. It also does not apply to the location of heavy atoms from macromolecular ΔF data because the distances between the 'atoms' are much larger. If the required resolution has not been reached, there is little point in persuing direct methods further; the only hope is to recollect the data with a larger crystal, stronger radiation source, longer measurement times, area detector, real-time profile fitting and lower temperature, or at least as many of these as are simultaneously practicable.

If the data reduction diagnostics give no grounds for suspicion and no direct methods solution gives good figures of merit, a brute force approach should be applied. This takes the form of TREF followed by a large number (e.g. TREF 5000); it may also be necessary to set a larger value for TIME. If either of the methods for interrupting a running job are available (see above), an effectively infinite value may be used (TREF 999999). Any change in this number of phase permutaions will also change the random number sequence employed for the starting phases. It may also be worth increasing the second TREF parameter (WE) in steps of say 10%.

If more than one solution has good R_α and Nqual values, it is possible that the structure has been solved but the program has chosen the wrong solution. The list of one-phase seminvariant signs printed by the program can be used to decide whether two solutions are equivalent or not. In such a case other solutions can be regenerated without repeating the complete job by means of 'TREF -n', where n is a solution code number (in fact the random number seed). Because of the effect of small rounding errors the 'TREF -n' job must be performed on the same computer as the original run. No other parameters should be changed when this option is used.

In cases of pseudosymmetry is may be necessary to modify the *E*-value normalization (i.e. by increasing the renorm parameter on the ESEL instruction to 0.9, or by setting a non-zero value of axis on the same instruction). E(min) should be set to 1.0 or a little lower in such cases.

When direct methods only reveal a fragment of the structure, it may well be correctly oriented but incorrectly translated relative to the origin. In such cases a non-centrosymmetric triclinic expansion with 'ESEL -1' may enable the symmetry elements and hence the correct translation (and perhaps the correct space group) to be identified.

Finally, if any heavier (than say sodium) elements are present, automatic Patterson interpretation should be tried.

14. Patterson Interpretation and Partial Structure Expansion

The Patterson superposition procedure in SHELXS was originally designed for the location of heavier atoms in small moiety structures, but it turns out that it can also be used to locate heavy atom sites for macromolecular ΔF data (see Chapter 15). For further details and examples see Sheldrick (1996) and Sheldrick, Dauter, Wilson, Hope & Sieker (1993).

14.1 Patterson interpretation algorithm

The algorithm used to interpret the Patterson to find the heavier atoms in the new version of SHELXS is totally different to that used in SHELXS-86; it may be summarized as follows:

1. One peak is selected from the sharpened Patterson (or input by means of a VECT instruction) and used as a superposition vector. This peak must correspond to a correct heavy-atom to heavy-atom vector otherwise the method will fail. The entire procedure may be repeated any number of times with different superposition vectors by specifying 'PATT nv', with |nv| > 1, or by including more than one VECT instruction in the same job.

2. The Patterson function is calculated twice, displaced from the origin by +U and -U, where U is the superposition vector. At each grid point the lower of the two values is taken, and the resulting 'superposition minimum function' is interpolated to find the peak positions. This is a much cleaner map than the original Patterson and contains only 2N (or 4N etc. if the superposition vector was multiple) peaks rather than N^2. The superposition map should ideally consist of one image of the structure and its inverse; it has an effective 'space group' of $P\overline{1}$ (or $C\overline{1}$ for a centered lattice etc.).

3. Possible origin shifts are found which place one of the images correctly with respect to the cell origin, i.e. most of the symmetry equivalents can be found in the peak-list. The SYMFOM figure of merit (normalized so that the largest value for a given superposition vector is 99.9) indicates how well the space group symmetry is satisfied for this image.

4. For each acceptable origin shift, atomic numbers are assigned to the potential atoms based on average peak heights, and a 'crossword table' is generated. This gives the minimum distance and Patterson minimum function for each possible pair of unique atoms, taking symmetry into account. This table should be interpreted by hand to find a subset of the atoms making chemically sensible minimum interatomic distances linked by consistently large Patterson minimum function values. The PATFOM figure of merit measures the internal consistency of these minimum function values and is also normalised to a maximum of 99.9 for a given superposition vector. The Patterson values are recalculated from the original F_o data, not from the peak-list. For high symmetry space groups the minimum function is calculated as an average of the two (or more) smallest Patterson densities.

5. For each set of potential atoms a

the atom list are retained during partial structure expansion, the rest are thrown away after calculating phases. At least one atom MUST be given! TEXP automatically generates appropriate FMAP, GRID and PLAN instructions.

TEXP (and/or PHAS) may be used in conjunction with TREF to generate fixed phases for use in direct methods; the special TEXP option na = 0 provides point atom phases for ALL reflections, which are then refined during the phase annealing and tangent expansion stages of direct methods (as specified on the PHAN and TREF instructions). It is not necessary to use different starting phases for the different phase sets, because

position. It may also be found by dividing the multiplicity of the special position (as as given in International Tables) by the multiplicity of the general position. Thus an atom on a fourfold axis will usually have s.o.f. = 10.25 (i.e. 0.25, fixed by adding 10).

positions; thus $P2_12_12_1$ (the only space group to fulfill all three criteria) is good but P1, C2, R3 and I4 are unsuitable.

If the standard direct methods run fails to find convincing heavy-atom sites, it should first be checked that the program has put out a comment that it has set the defaults for macromolecular data. The number of phase permutations may have to be increased (the first TREF parameter) or the number of large *E*-values for phase refinement may have to be changed (one should aim for at least 20 triplets per refined phase), but if too many phases are refined the performance is degraded because the ΔF-values only identify the strongest *E*-values reliably. The probability estimates may be changed by modifying the UNIT instruction, or more simply by changing the third TREF parameter, which multiplies the products of the three *E*-values in the triplet probability formula; for small molecules a value in the range 0.75 to 0.95 gives the best probability estimates, but it may be necessary to go outside this range for ΔF-data.

15.4 Patterson interpretation

For location of the heavy-atom site by Patterson interpretation of ΔF-data it may well be necessary to increase the number of superposition vectors to be tried (the first parameter on the PATT instruction), since the heavy-atom to heavy-atom vectors may be well down the Patterson peak-list. This number can be made negative to increase the 'depth of search' at the cost of a significant increase in computer time. The second number (the minimum vector length for the superposition vector) should be set to at least 8 Å (and to a larger value if the cell is large), and it can usually be made negative to indicate that special positions are not to be considered as possible heavy atom sites. An advantage of Patterson as opposed to direct methods is that such false solutions can be eliminated at a much earlier stage.

The third PATT parameter is also fairly critical for macromolecular ΔF-data; it is the apparent resolution, and is used to set the tolerances for deconvoluting the superposition map. If - as can easily happen with area detector data - a few ΔF-values are at appreciably higher resolution than the rest of the data, this may fool the program into setting too high an effective resolution. In such cases it is worth experimenting with several different values, e.g. 3.5 Å instead of 3.0 etc. The only other parameter which may need to be altered is maxat, if more than 8 sites are expected.

A typical ΔF PATT run (e.g. PATT 10 -12 2.5) will produce a relatively large number of possible solutions, some of which may be equivalent. The 'correlation coefficient' (which is defined in the same way as in most molecular replacement programs) is the only useful figure of merit for comparison purposes. Hand interpretation of the 'crossword table' is not as easy as for small molecules, because the minimum interatomic distances are not so useful; it is however still necessary to find a set of atoms for which the Patterson minimum function values are consistently high for at least most of the pairs of sites involved. This information tends to be more decisive for the higher symmetry space groups, because when there are more

16. CIF, CIFTAB and Electronic Publication

16.1 CIF archive format

The *CIF* format represents a major step forward in the archiving, publication and communication of crystallographic data. At last it is possible to publish crystal structures and incorporate structural data into the crystallographic databases without the expensive and error-prone retyping of tables by hand. CIF format also provides a convenient method of transferring data from one program system to another. The ACTA instruction instructs SHELXL to write two CIF-format files: *name.fcf* contains the reflection data and 'name.cif' all other data. These files contain all the items needed for archiving the structure; those answers not known to SHELXL (e.g. the color of the crystal) are left as a question mark. In general the final 'name.cif' file should be edited using CIFTAB or any text editor to replace most of these question marks. The file is then suitable for deposition in the CSD (organic) and ICSD (inorganic crystal structure) databases.

For publication of a routine structure determination via electronic mail it will normally be necessary to add the authors' names, title, text etc., which may also be done in CIF-format; this is followed by the edited contents of one or more *.cif* files each describing one structure (or possibly the same structure at different temperatures etc.). In general SHELXL provides all the CIF identifiers required by Acta Cryst. except those that begin with '_publ'. Further details are given below, and an example of a paper submitted to Acta Cryst. in this way may be found in the file *example.cif* (it has been brought up to date for the 1997 requirements for authors; whether it would pass the new stricter quality controls is another matter!). SHELXL users are strongly recommended to familiarize themselves with the definitive paper by the I.U.Cr. Commission on Crystallographic Data by Hall, Allen & Brown (1991), and with the current Acta Crystallographica Instructions for Authors.

where name is the first component of the filenames for the structure in question. CIFTAB enables tables to be produced from the *.cif* or *.fcf* files written by SHELXL and provides the following facilities, which may be selected from a simple menu.

Tables of crystal data, atom parameters, bond lengths and angles, anisotropic displacement parameters and hydrogen atom coordinates may be produced in a format specified in a file *ciftab.???* (where

The above directive, if present, should be the first line of the format file.

The directive $symops:n, where n is an integer, prints the symmetry operations used to

which are unknown to SHELXL, can be incorporated from separate files. This is more reliable

PostScript file of the manuscript; this can be printed or viewed by appropriate software. A useful feature is the highlighting (in bold) of any items which may subsequently be queried by editorial staff, and it may be possible to deal with these potential problems now, before final submission.

When everything is ready and checked, the CIF is e-mailed to med@iucr.ac.uk; after automatic checking is complete, a reply will list any problems requiring attention, will give a Co-Editor reference, and will ask for further material to be sent. This includes structure factor data, figures (diagrams), a copyright transfer form, and a formal signed letter of submission. The last two must still be sent by normal mail, but the others can be transferred electronically (ftp), using the method specified in the Instructions for Authors and the submission

17. SHELXA: Empirical Absorption Corrections

The program **SHELXA** has been kindly donated to the system by an **anonymous user**. This applies "absorption corrections" by fitting the observed to the calculated intensities as in the program DIFABS. SHELXA is intended for **EMERGENCY USE ONLY**, eg. when the world's only crystal falls off the diffractometer before there is time to make proper absorption corrections by indexing crystal faces or by determining an absorption surface experimentally by measuring equivalent reflections at different azimuthal angles etc.

SHELXA reads an *.fcf* file written by SHELXL (using LIST 4 or LIST 6 and any combination of MERG, OMIT etc.) and a .raw file in SHELX HKLF 4 format containing "direction cosines", and writes a new SHELX *.hkl* file in HKLF 4 format**. THIS WILL OVERWRITE AN EXISTING .hkl FILE !** A SHELXL-93 .fcf file is not suitable because some information is missing. The following restrictions apply to the use of SHELXA:

(a) The structure should not be twinned (racemic twinning is allowed), the data should have been collected from one crystal (inter-batch scale factors should not have been refined), and there may not be a re-orientation matrix on the HKLF instruction. Otherwise there are no

The data may be re-processed when, for example, extra atoms are added; however, as with DIFABS, best results are obtained if the procedure is last run with the final ISOTROPIC model; re-running it after anisotropic refinement will result in a deterioration of the structure and (most important) the *R*-factors. The ΔU fudge should not be used repetitively, because the effects will be cumulative !

Note that all esd's estimated by SHELXL using data "corrected" in this way will be invalid unless the number of parameters used in the absorption model is input as the third L.S.

18. Frequently Asked Questions

Q1: Please send me a copy of SHELX-76. I am afraid that I cannot use the new version because **my diffractometer measures *F*-values, not intensities**.

A: Buy a CCD detector. They measure intensities! [In fact, diffractometers

Q5: The structure could only be solved in **P1**, not $P\bar{1}$, but on refinement some of the bond lengths and U-values are wildly different in the two molecules. If I use SAME the geometries of the two molecules become very similar but how do I restrain the U_{ij} components of equivalent atoms to be the same?

A: You could use EADP, but it might be better to look for the inversion center instead, otherwise you will probably be **'marshed'**.

Q6: I included batch numbers in the *.hkl* file and BASF parameters in the *.ins* file, but the stupid program still **didn't refine the batch scale factors**!?

A: You need MERG 0 (the default MERG 2 will average the batch numbers).

Q7: How do I obtain the molecular replacement program **PATSEE**?

A: PATSEE has been maintained by its author, Ernst Egert, since he moved from Göttingen to the University of Frankfurt. He can be contacted by fax (+49-69-7982-9128) or email (bolte@chemie.uni-frankfurt.d400.de).

Q8: What should I do about **'may be split'** warnings?

A: Probably nothing. The program prints out this warning whenever it might be possible to interpret the anisotropic displacement of an atom in terms of two discrete sites. Such atoms should be checked (e.g. with the help of an ORTEP plot) but in many cases the single-site anisotropic description is still eminently suitable.

Q9: I get the message ' ** **UNSET FREE VARIABLE**'marshed'1r4 1aoyEo0(two 5ETOM ... **' i)-8I7(l

Q11: The program prints out a **Flack *x* parameter** of 0.3 with an esd of 0.05. Is the crystal racemically twinned?

A: Not necessarily! The Flack parameter estimated by the program in the final structure factor calculation ignores correlations with all other parameters (except the overall scale factor). Since these parameters may have refined so as best to fit a wrong absolute structure, it is quite possible to get an estimate of about 0.3 for the Flack parameter when the true value is 1, i.e. the structure needs to be inverted and is not racemically twinned. On the other hand a value close to zero with a small esd is a strong indication that the absolute structure is correct.

19. SHELX-97 Installation

Before trying to install the programs, it is worth checking with the SHELX homepage at http://shelx.uni-ac.gwdg.de/SHELX/ to see if there are any last-minute changes and whether other users have encountered problems on particular machines. The ftp site and CDROM contain the following files and subdirectories:

readme - Installation instructions, last-minute changes, changes since SHELXL-93 etc..

shelx.htm and ***shelxman.htm*** - On-line help in HTML format: requires a browser such as Netscape. *shelx.htm* contains the same general information as in README, and calls *shelxman.htm* that includes summaries of commands etc. ***applfrm.htm*** is the application form in html format. The file extensions will need to be changed to .html to active these files. three-letter extensions are used in the release for compatibility with MSDOS.

Subdirectory ***'unix'*** contains the sources of all programs for relatively standard UNIX systems. These should also compile successfully on many other operating sytems too (except VMS).

Subdirectory ***'vms'*** contains the VMS sources for Digital computers.

Subdirectory ***'doc'*** cotains the full manual in WINWORD 6 format, one file per chapter. It is designed to print on letter sized paper.

Subdirectory ***'ps'*** cotains the full manual in Postscript format, one file per chapter. It is designed to print on letter sized paper.

Subdirectory ***'egs'*** contains the test jobs and other examples files.

Subdirectory ***'ibm'*** contains the IBM RS6000 executables (these also execute on the IBM Power-PC series).

Subdirectory ***'sgi'*** contains the SGI IRIX executables; they should run under IRIX 5.3 or later with the R4000 series processors. For other systems it is desirable to recompile to obtain programs that execute faster even if the precompiled versions run correctly.

Subdirectory ***'linux'*** contains the LINUX executables for Intel processors.

Subdirectory ***'dos'*** contains the pure MSDOS executables. These may or may not run in the DOS windows under WINDOWS or OS2.

In addition, the ftp login directory contains gzipped tar files of the above subdirectories (e.g. ***unix.tgz***). These are convenient for down-loading with ftp as shown in the next section. The current sizes of these files in bytes are given on the SHELX homepage and should be checked to ensure that transmission is complete.

19.1 Installing the precompiled versions

In many cases it will be possible to use the precompiled versions provided. The executable programs (and the file ciftab.def) should simply be copied from the appropriate directory on the CDROM or ftp site to a directory on your machine. This directory should be specified in the 'PATH' so that the executables can be found. On UNIX systems the lazy way is to copy the programs into /usr/bin; on MSDOS systems they are usually copied to C:\EXE and this directory name is then added to the PATH specified in AUTOEXEC.BAT. You may also wish to copy the documentation and examples files.

As an example we shall take a PC running Linux; the following files should be fetched to your working directory by ftp (binary transfer !); for most other UNIX systems the installation procedure is similar:

linux.tgz, ps.tgz, egs.tgz, *shelx.htm* and *shelxman.htm*

The three gzipped tar files can then be expanded:

```
gunzip *.tgz
tar -xvf ps.tar
tar -xvf egs.tar
tar -xvf linux.tar
```

which will create the subdirectories ***ps, egs*** and ***linux***. The executables can be copied to */usr/bin* (needs system manager priviledges !):

```
cp linux/* /usr/bin
```

Under LINUX it is particularly easy to print the documentation, because lpr can recognize and print Postscript even on a non-Postscript printer:

```
lpr ps/*.ps
```

The on-line help files *shelxl.htm* and *shelxman.htm* should be renamed (mv) to *shelxl.html* and *shelxman.html* (the three-letter extension was needed for MSDOS systems !) and copied to a generally accessible directory; they may then be viewed with Netscape or any other HTML browser. These files are NOT copyrighted and you are welcome to improve and extend them as you wish for non-commercial purposes. *shelx.htm* calls *shelxman.htm* and applfrm.htm (the application form) It contains all the information from 'README' (which is a plain ASCII text file) plus a summary of the documentation (the full documention is available in WINWORD 6 format in subdirectory ***'doc'*** and in Postscript form in subdirectory ***'ps'***).

19.2 Program compilation under UNIX (and other operating systems)

The UNIX version has been designed to be easy to compile on a wide rangeof UNIX (and other) systems. The resulting compiled programs do not needany environment variables or hidden files to run; it is simply necessarythat the executable program is accessible via the PATH or an alias. The simplest way is to copy the executables into */usr/bin*.

The SGI executable of SHELXL was compiled under IRIX 5.3 as follows:

```
f77 shelxl.f -O3 -c
f77 shelxlv.f -O3 -c
f77 shelxl.o shelxlv.o -o shelxl
```

The compilation for other UNIX systems should be similar. **IT IS NECESSARY TO BE VERY CAREFUL ABOUT OPTIMIZATION**. The safest is to compile without any optimization first (-O0 rather than -O3 in this case), run the ags4 and 6rxn tests, and rename the resulting output files **.res, *.lst, *.fcf* and **.pdb*. Then recompile with highest optimization (-O3), rerun the tests, and use the UNIX diff instruction to compare the results with those from the unoptimized version. Small differences in the last decimal place do not matter, and of course the CPU times will differ, but if there are significant differences then the optimization level should be lowered and the tests repeated. For some systems (including certain SG Challenge and Digital Alpha systems), only the shelxlv.f file (containing the rate-determining routines) can be compiled with the highest optimization level; shelxl.f must be compiled at a lower level.

The shelxl.f source contains the following routines that

$fort/opt/ass=(noac,nodu)/align=all shelxs+shelxsv
$link shelxs
$fort/opt/ass=(noac,nodu)/align=all shelxl+shelxlv
$link shelxl
$fort/noopt shelxpro
$link shelxpro
$fort/noopt shelxwat
$link shelxwat
$fort/noopt shelxa
$link shelxa
$fort/noopt ciftab
$link ciftab

It may be necessary to split up the programs into subroutines to prevent the compiler running out of virtual memory. The files produced by the test jobs for SHELXL and SHELXS MUST be compared with those obtained using unoptimized versions of these programs (compiled with /noopt instead of /opt; note that /opt is usually the default) since optimizing errors are common for Digital compilers; there is a DIFF instruction in VMS that can be used for this. The remaining programs are not very CPU-intensive and so should not be optimized. If optimization causes errors, it is worth trying just to optimize *shelxsv.f* and *shelxlv.f* (which contain the rate determining routines) but not the rest. The executables need to be defined as follows:

shelxs :== $ disk:[directory]shelxs etc.

where 'disk' and 'directory' should be replaced by the appropraite local names and the programs are run (after preparing the files name.ins and name.hkl) by e.g.

shelxl name

SHELXWAT and SHELXA accept UNIX-type switches (even under VMS); they MUST come before the filename, e.g.

shelxwat -h name

No other files or parameter settings are required to run the programs, except that the file ciftab.def or a user-produced format definition file should be in the current directory when CIFTAB is run; if this file cannot be found in the current diectory, CIFTAB searches for it in a directory specified in the source.

19.4 Parallel and vector machines

SHELXL and SHELXS are designed to run very efficiently on vector computers (such as older Cray and Convex machines); no changes should be needed to the code. Unfortunately the crystallographic algorithms involved are less suitable for parallel computers (or multiprocessor systems); in such cases the avaliable computer resources are more efficiently used by running several jobs simultaneously, one per processor.

19.5 SHELXH - version of SHELXL for very large structures

SHELXH is a special version of SHELXL for the refinement of very large structures (with more than about 10000 unique atoms). The only difference between shelxh.f and shelxl.f is the first FORTRAN statement in which the array dimensions are specified by means of a PARAMETER statement; shelxh wascompiled (using shelxlv.f etc.) exactly as described above for shelxl. Large versions of shelxs, shelxpro and shelxa may be created in the same way, but it is rather unlikely that they will ever be required. Further details are provided by comments

20. SHELX-97 Application Form

SHELXL-97 LICENSE REGISTRATION FORM

Title/name:

Postal address:

Fax:

Email (legible!):

[ ] I wish to license SHELX-97 for use at the following for-profit firm or institution. I agree that within two months I will either destroy all copies of the programs in my possession or pay the license fee of US$2499. This license fee covers the use of the complete SHELX-97 for an unlimited time on an unlimited number of computers of any type at one geographical location:

[ ] I wish to license SHELX-97 for exclusively non-commercial purposes at the following not-for-profit institution only:

Please tell me how to obtain SHELX-97 by ftp [ ]; I already possess a copy of SHELX-97 [ ]

Please supply it on CDROM* [ ] / 100MB ZIP diskette* [ ] (*$99 for academic users)

[ ] I agree to cite SHELX-97 in all publications for which it was useful.

[ ] I agree that the author has no liability for any damage or loss caused by the programs.

[ ] Please send me a receipt for enclosed cheque [ ] Please send me an invoice

[ ] Please send direct bank transfer information [ ] No payment required (academic/ftp)

Signed: Date:

This form should be returned to George Sheldrick, Institut Anorg. Chemie, Tammannstr. 4, D37077 Göttingen, Germany by post or fax (+49-551-392582). Unsigned, emailed, incomplete (are the right boxes ticked?) or illegible forms will be returned by normal post for completion!

Hendrickson, W. A. & Konnert, J. H. (1980). *Computing in Crystallography,* edited by R. Diamond, S. Ramaseshan & K. Venkatesan. pp. 13.01 - 13.25. I.U.Cr. and Indian Acad. Sci.: Bangalore, India.

Hirshfeld, F. L. (1976). *Acta Cryst.* **A32**, 239 - 244.

Hirshfeld, F.L. & Rabinovich, D. (1973). *Acta Cryst.* **A29,** 510 - 513.

Hoenle, W. & von Schnering, H. G. (1988). *Z. Krist.* **184**, 301 - 305.

Irmer, E. (1990). Ph.D. Thesis, University of Göttingen. Germany.

Jameson, G. B., Schneider, R., Dubler, E. & Oswald, H. R. (1982). *Acta Cryst.* **B38**, 3016 - 3020.

Jones, P. G. (1988). *J. Organomet. Chem*. **345**, 405.

Kilimann, U., Noltemeyer, M., Schäfer, M., Herbst-Irmer, R., Schmidt, H. G. & Edelmann F. T. (1994). *J. Organomet. Chem.* **469**, C27 - C30.

Kld 12 220.77 53991y708 TDı0.004 ee(.Mg3ganomet. Chem.)Tjıst-Ir.

Pratt, C. S., Coyle, B. A. & Ibers J. A. (1971). *J. Chem. Soc.* 2146 - 2151.

Richardson, J.W. & Jacobson, R.A. (1987). *Patterson and Pattersons*, edited by J. P. Glusker, B. K. Patterson & M. Rossi, 310 - 317. I.U.Cr. and O.U.P.: Oxford.

Roesky, H. W., Gries, T., Schimkowiak, J. & Jones, P. G. (1986). *Angew. Chem. Int. Edn*. **25**, 84 - 85.

Rollett, J. S. (1970). *Crystallographic Computing,* edited by F. R. Ahmed, S. R. Hall & C. P. Huber, pp. 167 - 181. Copenhagen, Munksgaard.

Sheldrick, G.M. (1990). *Acta Cryst.* **A46**, 467 - 473.

Sheldrick, G.M. (1992). *Crystallographic Computing*, edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 145 - 157. I.U.Cr. and O.U.P.: Oxford, UK.

Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. C. (1993). *Acta Cryst.* **D49**, 18 - 23.

Sheldrick, G. M. & and R. O. Gould, R. O. (1995). *Acta Cryst.* **B51**, 423-431.

Steorxamp R. OE, Sceker, L. C. Joensn, M. CH (1990).

Index