Chapter 6.1.2

SHELXL-97

Structure Refinement

George W. Sheldrick
Dept. of Structural Chemistry
Tammannstraße 4
37077 Göttingen, Germany
email: gsheldr@shelx.uni-ac.gwdg.de

2. SHELXL - Structure Refinement

SHELXL is a program for the refinement of crystal structures from diffraction data, and is primarily intended for single crystal X-ray data of small moiety structures, though it can also be used for refinement of macromolecules against data to about 2.5 Å or better. It uses a conventional structure factor summation, so it is much slower (but a little more accurate) than standard FFT-based macromolecular programs. SHELXL is intended to be easy to install and use. It is very general, and is valid for all space groups and types of structure. Polar axis restraints and special position constraints are generated automatically. The program can handle twinning, complex disorder, absolute structure determination, CIF and PDB output, and provides a large variety of restraints and constraints for the control of difficult refinements. An interface program SHELXPRO enables macromolecular refinement results to be displayed in the form of Postscript plots, and generates map and other files for communication with widely used macromolecular programs. An auxiliary program CIFTAB is useful for tabulating the refinement results via the CIF output file for small molecules.

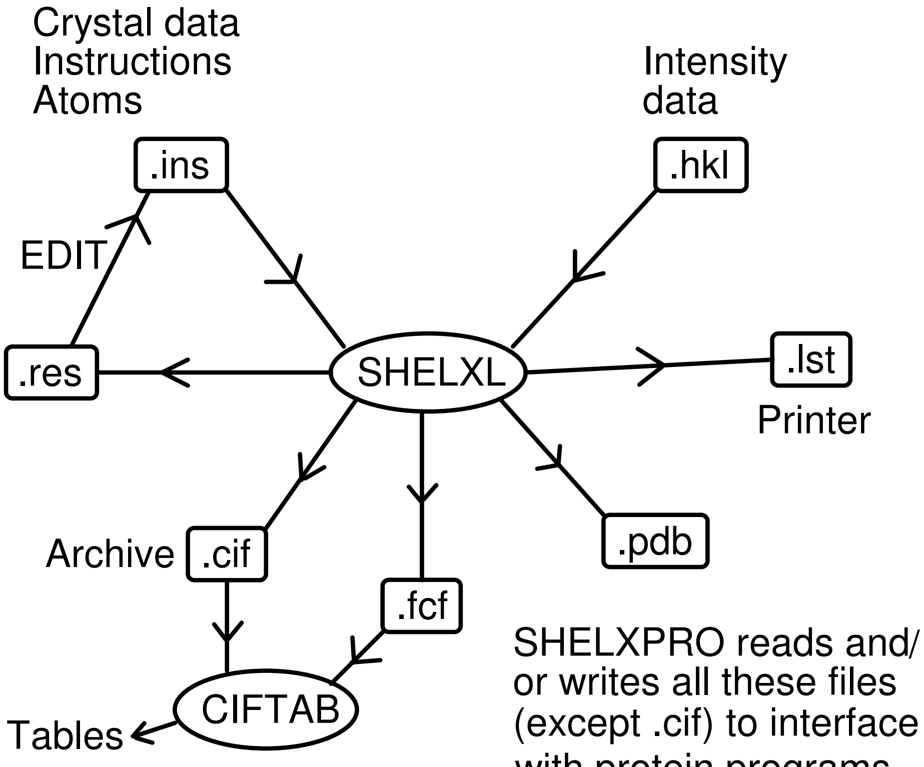
2.1 Program organization

To run SHELXL only two input files are required (atoms/instructions and reflection data); since both these files and the output files are pure ASCII text files, it is easy to use the program on a heterogeneous network. The reflection data file (*name.hkl)* contains *h, k, l,* F^2 and $\sigma(F^2)$ in standard SHELX format (section 2.3); the program merges equivalents and eliminates systematic absences; the order of the reflections in this file is unimportant. Crystal data, refinement instructions and atom coordinates are all input as the file *name.ins*; further files may be specified as 'include files' in the *.ins* file, e.g. for standard restraints, but this is not essential. Instructions appear in the *.ins* file as four-letter keywords followed by atom names, numbers, etc. in free format; examples are given in the following chapters. There are sensible default values for almost all numerical parameters. SHELXL is normally run on any computer system by means of the command:

shelxl *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. Batch operation will normally require the use of a short batch file containing the above command etc. The executable program must be accessible via the 'PATH' (or equivalent mechanism). No environment variables or extra files are required.

A brief summary of the progress of the structure refinement appears on the console, and a full listing is written to a file *name.lst*, which can be printed or examined with a text editor. After each refinement cycle a file *name.res* is (re)written; it is similar to *name.ins*, but has updated values for all refined parameters. It may be copied or edited to *name.ins* for the next refinement run. The MORE instruction controls the amount of information sent to the *.lst* file; normally the default MORE 1 is suitable, but MORE 3 should be used if extensive diagnostic information is required. The ACTA instruction produces CIF format files for archiving or electronic publication, and the LIST 4 instruction (generated automatically by ACTA) produces a CIF format reflection data file (*name.fcf*). For PDB deposition of macromolecular results, WPDB and LIST 6 should be used. The program SHELXPRO should then be used to complete the PDB file.



Two mechanisms are provided for interaction with a SHELXL job which is already running. The first is used by the MSDOS and some other 'on-line' versions: 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. Usually the <Tab> key may be used as an alternative to <ctrl-I>. If the <Esc> key is hit during least-squares refinement, the program completes the current cycle and then, instead of further refinement cycles, continues with the final structure-factor calculation, tables and Fourier etc. 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 contents of this file are irrelevant); the program tries at regular intervals to delete it, and if it succeeds it takes the same action as after <Esc>. The *name.fin* file is also deleted (if found) at the start of a job in case it has been accidentally left over from a previous job. This approach may be used with batch jobs under most operating systems.

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^2 value is too large to fit format F8.2, then format F8.0 may be used instead. - the decimal point overrides the FORTRAN format specification.

The use of a threshold for ignoring weak reflections may introduce bias which primarily affects the atomic displacement parameters; it is only justified to speed up the early stages of refinement. In the final refinement ALL DATA should be used except for reflections known to suffer from systematic error (i.e. in the final refinement the OMIT instruction may be used to omit specific reflections - although not without good reason - but not ALL reflections below a given threshold). Anyone planning to ignore this advice should read Hirshfeld & Rabinovich (1973) and Arnberg, Hovmöller & Westman (1979) first. Refinement against F^2 also facilitates the treatment of twinned and powder data, and the determination of *absolute structure*.

2.5 Initial processing of reflection data

SHELXL automatically rejects systematically absent reflections. The sorting and merging of the reflection data is controlled by the MERG instruction. Usually MERG 2 (the default) will be suitable for small molecules; equivalent reflections are merged and their indices converted to standard symmetry equivalents, but Friedel opposites are not merged in non-centrosymmetric space groups. MERG 4, which merges Friedel opposites and sets $\delta f''$ for all elements to zero, saves time for macromolecules with no significant dispersion effects. Throughout this documentation, F_o^2 means the EXPERIMENTAL measurement, which despite the square may possibly be slightly negative if the background is higher than the peak as a result of statistical fluctuations etc. R_{int} and R_{sigma} are defined as follows:

$$R_{\mathrm{int}} = \Sigma \mid F_o^2 - F_o^2(\mathrm{mean}) \mid \; / \; \Sigma [F_o^2]$$

where both summations involve all input reflections for which more than one symmetry equivalent is averaged, and:

$$R_{\mathrm{sigma}} = \Sigma [\sigma(F_o^2)] \; / \; \Sigma [F_o^2]$$

over all reflections in the merged list. Since these R-indices are based on F^2, they will tend to be about twice as large as the corresponding indices based on F. The 'esd of the mean' (in the table of inconsistent equivalents) is the rms deviation from the mean divided by the square root of (n-1), where n equivalents are combined for a given reflection. In estimating the $\sigma(F^2)$ of a merged reflection, the program uses the value obtained by combining the $\sigma(F^2)$ values of the individual contributors, unless the esd of the mean is larger, in which case it is used instead.

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* and an OMIT threshold, a conventional index *R*1 based on observed *F* values larger than $4\sigma(F_o)$ is also printed.

$$wR2 = \{ \Sigma [w(F_o^2-F_c^2)^2] \; / \; \Sigma [w(F_o^2)^2] \}^{1/2}$$

$$R1 = \Sigma \; | \; |F_o| - |F_c| \; | \; / \; \Sigma \; |F_o|$$

The *Goodness of Fit* is always based on F^2:

$$\text{GooF} = S = \{ \Sigma [w(F_o^2-F_c^2)^2] \; / \; (n-p) \}^{1/2}$$

where *n* is the number of reflections and *p* is the total number of parameters refined.

The WGHT instruction allows considerable flexibility, but in practice it is a good idea to leave the weights at the default setting (WGHT 0.1) until the refinement is essentially complete, and then to use the scheme recommended by the program. These parameters should give a flat 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 etc. It is recommended that a difference electron density synthesis is performed at the end of each refinement job; it is quick and of considerable diagnostic value. In contrast to SHELX-76, SHELXL finds the asymmetric unit for the Fourier synthesis automatically; the algorithm is valid for all space groups, in conventional settings or otherwise. Before calculating a Fourier synthesis, the Friedel opposites are always merged and a dispersion correction applied; a value of $R1$ is calculated for the merged data (without a threshold). Reflections with F_c small compared to $\sigma(F_o)$ are down-weighted in the Fourier synthesis. The rms density is calculated to give an estimate of the 'noise level' of the map.

2.9 The connectivity array

The key to the automatic generation of hydrogen atoms, molecular geometry tables, restraints etc. is the connectivity array. For a non-disordered organic molecule, the connectivity array can be derived automatically using standard atomic radii. A simple notation for disordered groups enables most cases of disorder to be processed with a minimum of user intervention. Each atom is assigned a 'PART' number *n*. The usual value of *n* is 0, but other values are used to label components of a disordered group. Bonds are then generated for atoms that are close enough only when either (a) at least one of them has n=0, or (b) both values of *n* are the same. A single shell of symmetry equivalents is automatically included in the connectivity array; the generation of equivalents (e.g. in a toluene molecule on an inversion center) may be prevented by assigning a negative 'PART' number. If necessary bonds may be added to or deleted from the connectivity array using the BIND or FREE instructions. To generate additional bonds to symmetry equivalent atoms, EQIV is also needed.

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 damping the refinement (see above) leads to underestimates of the esds; in difficult cases a final cycle may be performed with DAMP 0 0 (no damping, 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 find possible hydrogen bonds. This is a convenient way of finding the 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 placing of hydrogen atoms has assigned the hydrogens of 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 finds the most suitable hydrogen atom to form the hydrogen bond D-H•••A, and outputs the geometric data for this hydrogen bond to the *.lst* file and the *.cif* file (if ACTA is present).