Changes in SHELXL since SHELXL-97

Most of the changes in SHELXL are intended for small molecule applications. A number of improvements for macromolecules are also planned, but some of them will require major restructuring of the program, for example the introduction of chain IDs. With a few small unavoidable exceptions, the program should be upwards compatible with SHELXL-97 and even SHELX76.

1. General comments

The major changes from SHELXL-97 (in addition to fixing bugs) so far are as follows:

New instructions: ABIN, ANSC, ANSR, NEUT, PRIG, RIGU, TWST, WIGL, XNPD
Instructions deleted: HOPE, TIME (and MOLE is simply ignored).

GUI writers can achieve a preliminary quick fix by simply ignoring the above new instructions. Except for ANSC and WIGL they should be copied to a new *.ins* file.

Multiple CPUs are supported (thanks to Kay Diederichs). The number of CPUs actually employed and the array sizes may be set with runtime flags, e.g.

```
shelxl name -t4 -b5000
```

but this will rarely be necessary. For further details run the program without any command-line parameters. Note that to obtain optimum performance it may be necessary to switch off hyperthreading or to use the -t command line switch to reduce the number of threads started, e.g. -t4 for an Intel i7 system that has 4 CPUs plus 4 hyperthreading channels.

SHELXL is available as 32- and 64-bit executables. If there is a need (in order to be able to do extremely large full-matrix refinements) a version with 64-bit addressing could be generated. This will require at least 12GB RAM, otherwise it will pulverize the hard disk and never finish. It was originally planned to perform all calculations with double-precision arithmetic. This had some minor advantages but on average was about a factor of two slower, probably as a result of cache misses, so the idea has been abandoned.

Except that TIME is no longer available (not compatible with multi-CPU operation) and that HOPE is now replaced by more conventional anisotropic scaling (ANSC), the new version should be able to read the *.ins* and *.hkl* files used by any SHELX version of the last 40 years.

Hydrogen (H) and deuterium (D) atoms are now both treated as special cases for all purposes, in particular for AFIX. The positions of H and D on the SFAC instruction(s) no longer matter. This special treatment is modified by the NEUT instruction. If there are errors, e.g. connectivity errors caused by missing atoms, in HFIX/AFIX etc., all such errors are reported before aborting. There is no longer any effective limit on the number of atoms that can be implied by a single instruction such as SIMU $C_* $N_* $O_* or DELU N_1 > LAST.

This was previously a problem with some very large structures, and was caused by the packing of more than one data item into one 32-bit word. It required major code restructuring to change this and so allow an unlimited number of atoms to be (implicitly) addressed with a single instruction.

2. Changes to the CIF output files

There are extensive changes to the output CIF files. In general the program tries to output all it knows but nothing else. For example the temperature will now appear as '?' unless set by a TEMP instruction. The *.cif* file now includes embedded copies of the final *.res* file and the input *.hkl* file (HKLF 3, 4 or 5). The space group number, name and Hall symbol are generated from LATT and SYMM, but only for the 249 most common settings. For example, these CIF items are generated automatically for the space group $P2_1/c$ in the non-standard setting $P2_1/n$ but not for the less common non-standard setting $B2_1/d$, even though there was never a problem in refining in $B2_1/d$ (using LATT 6 and SYMM 3/4+X, 1/2-Y, 1/4+Z) with SHELXL.

Here is what the CIF output now looks like when SIZE is specified:

```
_exptl_transmission_factor_min    ?
_exptl_transmission_factor_max    ?
_exptl_crystal_size_max           0.298
_exptl_crystal_size_mid           0.193
_exptl_crystal_size_min           0.126
_exptl_absorpt_coefficient_mu     0.098
_shelx_estimated_absorpt_T_min    0.971
_shelx_estimated_absorpt_T_max    0.988
_exptl_absorpt_correction_type    ?
_exptl_absorpt_correction_T_min   ?
_exptl_absorpt_correction_T_max   ?
_exptl_absorpt_process_details    ?
```

The transmission factors are intended for the numbers from a scaling program such as SADABS, and are already in the core CIF directory. The numbers estimated from SIZE and mu that previously appeared as _exptl_absorpt_correction.. are now in _shelx_estimated_...

The old but undocumented **LIST 7** may still be used to produce a list of h, k, l, F_o^2 and $\sigma(F_o^2)$, followed on the same line by F_c^2 for each twin component (-1 if absent). The new **LIST 8** writes h, k, l, F_o^2, $\sigma(F_o^2)$, F_c^2, phase angle in degrees, d-spacing in Ångströms, and $\sqrt{(1/w)}$ to the *.fcf* file, where w is the weight used in the refinement. $\sqrt{(1/w)}$ should tend to be only slightly greater than $\sigma(F_o^2)$ for very weak reflections when a normal weighting scheme is used. The list is detwinned and sort/merged, but without eliminating Friedel opposites and anomalous contributions (except when calculating the phase angle, which would otherwise correspond to complex electron density) unless the structure is centrosymmetric. F_o^2, F_c^2, $\sigma(F_o^2)$ and $\sqrt{(1/w)}$ are on an absolute scale.

The **CONF** instructions optionally takes one or two numerical parameters in addition to the string of atom names (if any). “CONF min_d min_a” followed by atoms removes torsion angles generated by CONF from the *.cif* and *.lst* files if the 2-3 distance is greater than min_d or at least one of the 1-2-3 and 2-3-4 angles is more than min_a. The first parameter may be used to eliminate torsion angles when atom 2 or 3 is a metal atom. The torsion angles become meaningless (with a large s.u.) when either of the bond angles approaches 180 degrees. Such cases may be removed with the second parameter. These two tests together can produce a significant simplification of the torsion angle table for organometallic and coordination compounds. Often the default parameters will suffice, improving downwards compatibility.

Bond angles no longer appear in the *.cif* file when the atoms 1 and 3 have different non-zero PART numbers. It is still necessary to include these angles in the triangular tables in the *.lst* file. Symmetry equivalents of atoms with negative PART numbers now no longer appear in the connectivity table and bond length and angle tables. These were real (if relatively harmless) bugs in SHELXL-97!

As suggested by Brian McMahon (IUCr), _atom_site_site_symmetry_multiplicity has been corrected to _atom_site_site_symmetry_order. The SHELX programs do not need or use the concept of multiplicity; when an atom lies on a special position the occupancy is divided by the 'order' of that position. This is not a problem when an atom is exactly on the special position but the conversion between SHELX and CIF formats can be problematic when a disordered molecule of lower symmetry than the special position occupies the region of space around the position. A common example is a toluene molecule on an inversion center. If an atom in such a molecule accidentally lies close to the special position, it might be assigned an order and hence occupancy different from those of the rest of the disordered molecule, and this assignment could even change after the next refinement! To resolve this problem, SHELXL now always sets the order to 1.0 for PART -N atoms, the occupancy in the CIF file is then the same as in the SHELXL *.ins* file. For other atoms, the SPEC tolerance (used to detect special positions for generating U_{ij} constraints) is now also used to detect whether the atom is 'special' or not. For the large majority of CIF files the only difference to the files generated by SHELXL-97 will be the replacement of '_multiplicity' by '_order', but other differences could arise in some esoteric cases. For example a toluene molecule lying on a position of 2/m symmetry will be assigned an order of 1.0 (because of the PART -1) and will thus be given a CIF occupancy of 0.25, although it could be argued that the order should be 2 and the CIF occupancy 0.5 for all seven carbon atoms, or even that the order should be 4 and the CIF occupancy 1.0! However even in this case, the SHELX occupancy will always be equal to the CIF occupancy divided by the order.

3. Changes to the .pdb output file

The space group name (in a Coot-compatible format) is included in the CRYST1 record in the *.pdb* output file for 249 common space group settings. The residue name is now right-justified in the *.pdb* file.

4. Determination of absolute structure

A byproduct of the CIF changes is that the **ACTA** instruction now requires **MERG 2** (or less) for a non-centrosymmetric space group to avoid losing required information. This enables the Flack parameter to be estimated by the Parsons' quotient method based on $[I_+-I_-]$ / $[I_++I_-]$. Refining the Flack parameter with **TWIN** / **BASF** by least-squares is no longer considered to be the optimal way of determining the Flack parameter, see Flack *et al., Acta Cryst.* **A67** (2011) 21-34. The quotients are particularly robust because they involve the cancellation of errors that affect both I_+ and I_-. This approach requires the estimated intensity errors to be on an approximately absolute scale, so the the Flack parameter and its standard uncertainty will only be reliable when the goodness of fit at the end of the refinement is close to unity, as will normally be the case if the refinement has converged and the weighting scheme optimized. The Parsons' method may also be used for twinned structures because the calculation is performed after the data have been 'detwinned'. The Parsons' method cannot be used with **TWIN** / **BASF**, but where it is possible to interpret unambiguously the **TWIN** and **BASF** values in terms of a Flack parameter and its s.u., it is automatically included in the *.cif* file. For the ags4 test structure provided with SHELXL, one extra line is now output to the console:

```
Flack x = 0.012 ( 14) from 271 selected quotients (Parsons' method)
```

and the following information about the Friedel completeness appears in the .cif file:

```
_diffrn_reflns_theta_max          27.499
_diffrn_reflns_theta_full         25.242
_diffrn_reflns_Laue_measured_fraction_max    0.995
_diffrn_reflns_Laue_measured_fraction_full   0.994
_diffrn_reflns_point_group_measured_fraction_max   0.832
_diffrn_reflns_point_group_measured_fraction_full  0.906
_reflns_Friedel_coverage          0.531
_reflns_Friedel_fraction_max      0.636
_reflns_Friedel_fraction_full     0.798
```

In the case of a twinned structure (**TWIN/HKLF 4** or **HKLF 5**), **TWST** N[0] sets the twin component to which these statistics refer. Only single and composite reflections involving component N are used for the statistics. If N is zero or **TWST** is omitted, all components are used (in version 2013/2 or later). 'max' refers to the maximum resolution of the data used in the refinement, and 'full' to the CheckCIF standard (sin(theta)/lambda < 0.6). '_Friedel_fraction_' refers to the number of Friedel pairs actually present divided by the number theoretically possible; unlike the '_Friedel_coverage_' currently implemented in coreCIF, it should be 1.000 in the ideal case where no reflections are missing, and so provides an immediate check on the suitability of the data for the determination of absolute structure. Later in the *ags4.cif* file the following appears:

```
_refine_ls_abs_structure_details
;
 Flack x determined using 271 [(I+)-I(-)]/[(I+)+(I-)]
 quotients (Parsons and Flack (2004), Acta Cryst. A60, s61)
;
_refine_ls_abs_structure_Flack    0.012(14)
```

5. Estimation of standard uncertainties

A change with far-reaching consequences (suggested by Ton Spek) involves the estimation of standard uncertainties in all refined parameters and derived quantities and also the calculation of the goodness of fit. The standard least-squares approach involves division by the square root of (N-P), where N is the number of observations and P is the number of parameters refined. In previous versions, N was the number of reflections or in the case of a twin, the total number of composite and single reflections. This was the subject of many often heated discussions, because least-squares theory assumes that these obvervations are independent. It was argued (in particular by referees and editors) that this assumption is not valid in the case of **HKLF 5** refinement of twinned structures, with the unfortunate result that perfectly good experimental data were often thrown away so that only unique reflections could be used for refinement. A similar problem arose for refinement of non-centrosymmetric structures against reflections that had been merged according to the point group (**MERG 2**) rather than the Laue group (**MERG 3**). On the other hand Friedel opposites are not exact equivalents, and improvements in hardware and software have made their small differences more significant, so it was difficult to decide when to use **MERG 2** or **MERG 3**. Ton's elegant solution was to ***replace N by the number of unique reflections (according to the Laue group, not the point group) instead of the number of observations in all calculations,*** so it no longer necessary to ensure that the data used in the refinement are unique or to set the third **L.S.** parameter to correct for non-unique data. In fact **MERG 2** is now recommended for all non-centrosymmetric structures (and is obligatory if **ACTA** is specified) so that more complete statistics can be calculated and included in the *.cif* file. For small molecule structures the increase in CPU time is no longer important, and for macromolecules the **ACTA** instruction to create the *.cif* file is not necessary, so **MERG 3** may still be used. The **MERG 2** esds may be slightly larger than with **MERG 3** for data with insignificant anomalous differences, but **MERG 3** esds tended to be underestimated especially for non-centrosymmetric structures anyway. The third number on the **L.S.** instruction should now only be needed for 'squeezed' structures.

6. Including precalculated partial structure factors

The new **ABIN** instruction reads h, k, l, A and B from the file *name.fab*, where A and B are the real and imaginary components of a partial structure factor. This file is read in free format (numbers separated by one or more spaces) with one reflection per line, and is terminated by the end of the file (a 0 0 0 reflection is not required and is ignored if present). The reflections may be in any order, duplicates and systematic absences are ignored. Symmetry equivalents are generated automatically. At least one equivalent of each reflection used in the refinement, including all reflections in all twin components, should be present in this list, superfluous reflections in the *.fab* file (e.g. outside the resolution limits) are ignored. In the case of twinning, the A and B values should refer to the untwinned structure, but they are used to calculate the structure factors for all twin components. Thus the *.fcf* file created using the new **LIST 8**, which has already been 'detwinned' and merged (using Friedel's law only for centrosymmetric structures) but still contains the anomalous contributions, may be used as an aid to generating them. **ABIN** takes two free variable numbers n_1 and n_2 as parameters.

The input A and B values are multiplied by k.exp($-8\pi^2 U\sin^2\theta/\lambda^2$), where k is the value of free variable n_1 and U is the value of free variable n_2. If n_2 is omitted, U is set to zero, and if n_1 is also omitted, k is fixed at 1.0 (in which case the A and B values should be on an absolute scale of electrons per unit-cell). SUMP restraints may be applied to these free variables. The partial structure factor contributions might come from a solvent mask (for a macromolecule) or a blob of unresolved solvent density for a small molecule, e.g. in a channel along a cell axis, as modeled by the **squeeze** method in **PLATON**. In the latter case it might be appropriate to set the third L.S. parameter to the number of parameters that would have been required to model such a solvent region by fitting disordered solvent molecules to it, so that the standard uncertainties are estimated correctly. n_1 may be made negative to force the program not to assume Friedel's law when generating equivalents of the input A and B values; this is only required when the partial structure factors have significant anomalous contributions.

7. New ADP restraints and constraints

The enhanced rigid bond restraint **RIGU** described in *Acta Cryst.* A**68** (2012) 448-451 has been implemented. The syntax is the same as for DELU but three times as many restraints are generated. The default esds are usually adequate. The parameter p described in the paper may be changed using the new **PRIG** instruction, but this will rarely be necessary..

The new instruction **XNPD** sets a lower bound for the eigenvalues of the U_{ij} tensor of all anisotropic atoms or the U of an isotropic atom. The default (i.e. assumed if XNPD is not given) is 'XNPD -0.001'. This has the effect that non-positive-definite (NPD) atoms are still detected and reported, but they are prevented from causing the refinement to explode. The number of 'may be split' and NPD atoms is output to the console at the end of the refinement.

8. Refinement against neutron diffraction data

To facilitate refinement against neutron diffraction data, the **NEUT** instruction has been implemented. This has three effects:

1. The special treatment of H and D atoms is switched off except for the generation and refinement of hydrogen atoms with HFIX and AFIX. AFIX 87 and 137 take the negative scattering length of H into account when interpreting the circular difference density map.

2. If **NEUT** comes ***before*** a **SFAC** instruction that contains atom names but not numbers, neutron scattering factors and absorption coefficients are used for those elements. Except for D that is assumed to be pure isotopes, the scattering lengths are the weighted mean values for natural isotopic abundances. **DISP** can be added if required, e.g. for cadmium. The full form of the **SFAC** instruction may still be used to input scattering factors that are not changed by **NEUT**.

3. If the isotropic U of an atom (usually H or D) is given a value –k where -0.5 > k > 5.0, it is set to –k times the U-value of the last normal atom, but in contrast to the similar action when **NEUT** is not set, both atoms contribute to the calculation of the derivatives used in the least-squares calculations, so both atoms must be isotropic. This significantly reduces the number of degrees of freedom of the refinement and so would be expected to reduce the gap between $R1_{free}$ and R1 for a macromolecular refinement, especially when the number of neutron data is limited. A relatively large k value (say 2.5) appears to be appropriate for D_2O and H_2O molecules when refining against neutron data.

9. Other new facilities

The default value of the **ACTA** parameter is now the 2θ value at which sin(θ)/λ is exactly 0.6 (to comply with the little-known CheckCIF standard) and the default value of the fifth **DEFS** parameter has been increased to 4.0 (a higher value might be misunderstood as a free variable refence). The default value of the third **SIMU** parameter has been increased from 1.7 to 2.0 Angstroms. The previous value was often too low for e.g. C-S bonds in methionine side-chains, with the result that the SIMU restraint was not being applied in such cases. Note that his change increases the number of restraints applied and so can affect the refinement results when default parameters are used.

HTAB without atom names now writes the appropriate **EQIV** and full **HTAB** instructions (needed for CIF output of hydrogen bonds) to the *.res* file after the **HKLF** instructions but before the Fourier peaks. Programs that use these Q-peaks should have no problems with this because they can ignore the **EQIV** and **HTAB** instructions. This **HTAB** instruction also generates non-classical C-H•••O hydrogen bonds for O•••H-C-X and O•••H-C-C-X where X is an electronegative atom, but does not generate hydrogen bonds that violate the **PART** rules. These instructions should be checked carefully before moving them in front of the **HKLF** instruction for the next refinement run, because there can be surprises!

The **TWIN** instruction may now be used for more general cases of twinning, including rhombohedral obverse/reverse twins (indexed on a hexagonal cell) and many pseudo-merohedral twins. However there is still the restriction that only integer reflections indices may be read from the *.hkl* file

ANSC (anisotropic scaling) is only usually of practical use for isotropic refinements of macromolecules, because it would be 100% correlated with the individual anisotropic ADPs, though it might be applicable in cases where only the heavy atoms are refined anisotropically. In the first job, **ANSC** is entered without any parameters in the first run and is written out to the .res file with six parameters and can be reinput for the next refinement. Although six parameters are specified, the program automatically applies the constraints appropriate for the crystal system, so usually less than six parameters are actually refined. **ANSR** sets the esd of a restraint that is applied with target values of zero to these parameters to prevent instabilities in full-matrix refinement, especially when all atoms are also refined anisotropically, introducing correlations of 100%. **ANSR** should usually be left at the default value of 0.001.

The third **SHEL** parameter sets the high-resolution limit for the free R reflections so that the same free-R set can be used when varying the resolution cutoff (the second **SHEL** parameter) for the rest of the data. If not specified, the third parameter is made equal to the second.

The new **WIGL** instruction displaces all atoms not on special positions by the specified average distance in a random direction. The default is not to apply these shifts. The shifts are applied after generating the connectivity table (which would otherwise be compromised) and before generating hydrogen atoms. Fixed coordinates (e.g. for special positions) are not changed. **WIGL** is useful for removing free-R memory effects and checking convergence properties. Shifts greater than about 0.5A can result in some atoms moving out of density and not finding their way back home.

SADI without any atom names produces a list of **SADI** instructions in the *.res* file (after HKLF). They are derived from all the input **SAME** and **SADI** instructions with duplicates removed. This list can be edited if necessary and used instead of the original restraints in order to, for example, model disorder within a **SAME** group.

TWST N[0] specifies the twin component number to be used for the completeness and Friedel completeness statistics. Only single or composite reflections containing this twin component are used for these statistics. It applies to both **TWIN+HKLF 4** and **HKLF 5** data, but is most useful for the latter. If there is no twinning, this parameter has no effect. The default N = 0 (version 2013/2 and later) causes reflections from all components to be used in calculating the completeness and Friedel completeness.