PAUL SCHERRER INSTITUT

Dr. Tim Grüne :: Paul Scherrer Institut :: tim.gruene@psi.ch

SHELX for experimental phasing and refinement

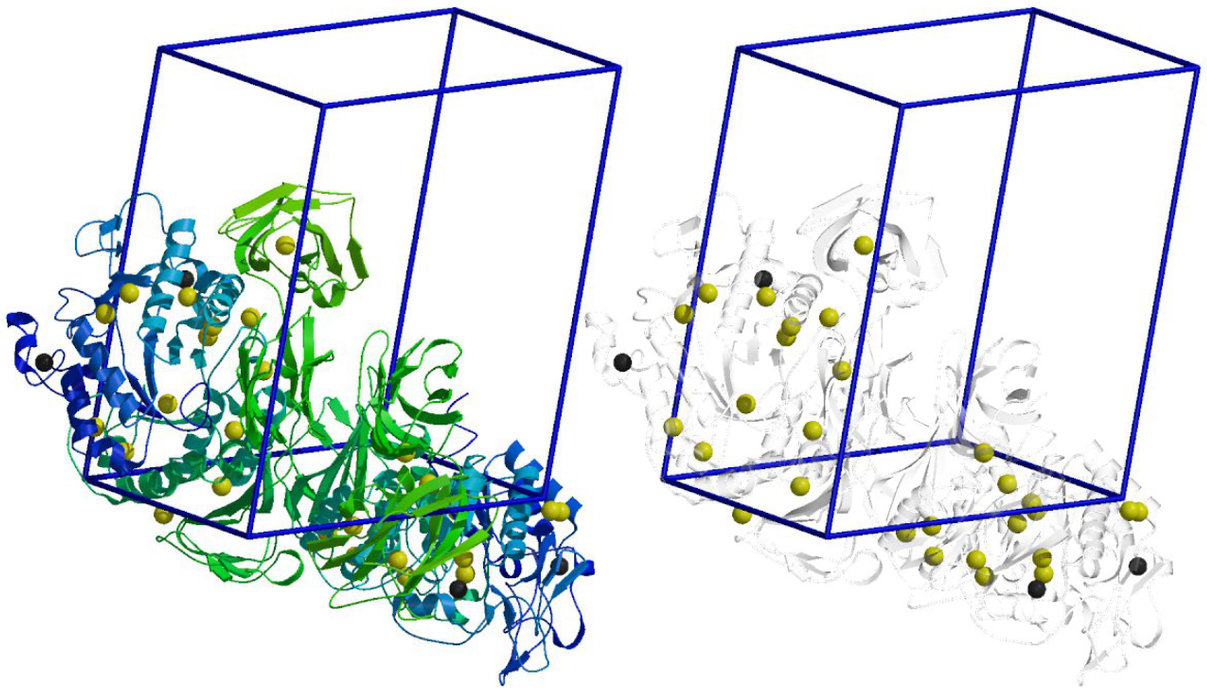
iNEXT Course Oulu, 2017

17th May 2017

1 - Overview

- Phasing and the Substructure
- Using SHELX C/D/E
- Structure Refinement with SHELXL

The Substructure



- Coordinates of anomalous scatterers
- Anomalous difference

$$\big||F^{+}(hkl)| - |F^{-}(hkl)|\big| \approx |F_{\text{sub}}(hkl)|$$

  corresponds to small molecule data set
- Shelxd: solve substructure with *direct methods*
- Harker Construction: Expand phases to full data set

The Crystallographic Phase Problem

1. Crystal diffraction yields intensities $I(hkl)$, and thus structure factor amplitude $|F(hkl)| \propto \sqrt{I(hkl)}$

2. Model building requires a map, i.e. $\rho(x,y,z) = \sum_{h,k,l} |F(hkl)| e^{i\phi(hkl)} e^{-2\pi i(hx+ky+lz)}$

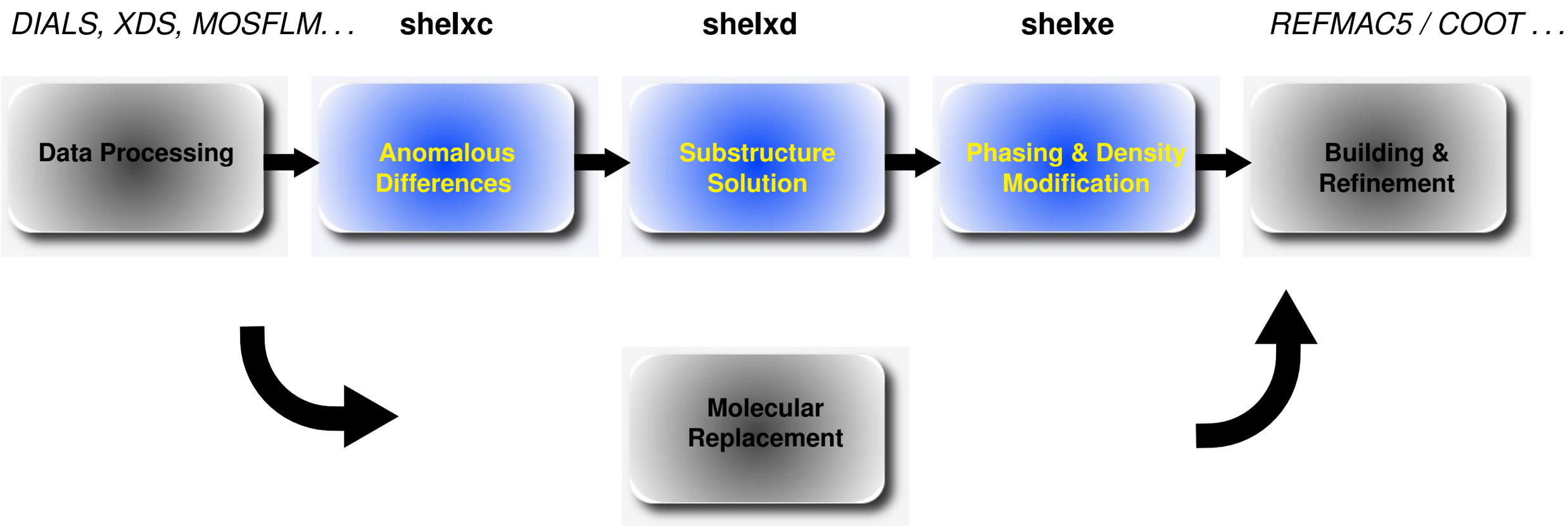
3. **Solving a structure** = determination of the phase angles $\phi(hkl)$ good enough to create an **interpretable map**

4. Small molecule crystallography: Problem mostly solved with **direct method**

5. SHELXD solves **small molecule** data at $d < 1.2$Å

6. SHELXD solves the **substructure** even at $d = 5$Å (and worse) because substructure atoms are **still resolved**

Context within Structure Solution



shelx c/d/e [1]

shelxc ➡ **shelxd** ➡ **shelxe**

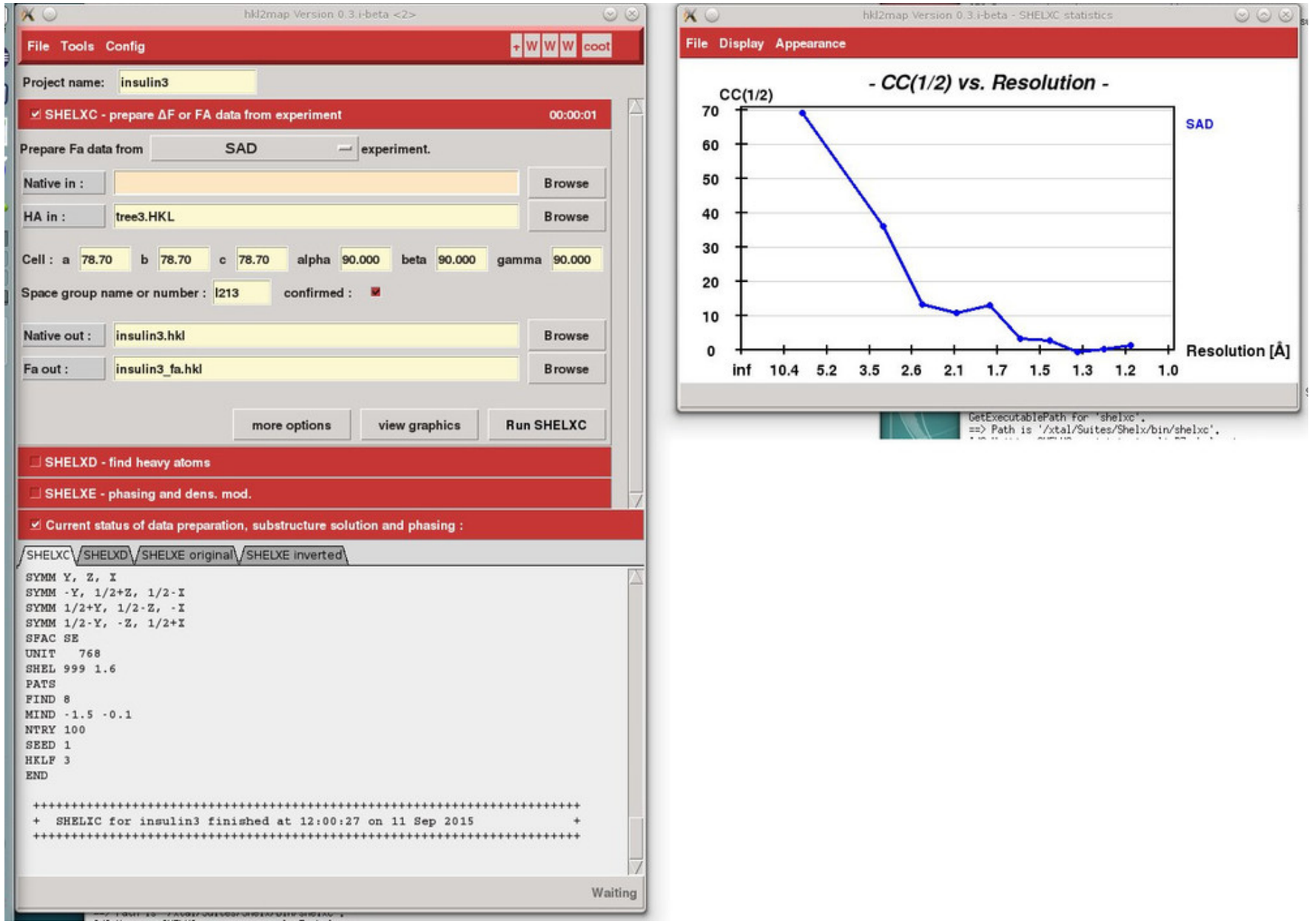
- Extracts anomalous signal
- Prepares native data
- Analyses data sets

- Determines substructure

- Density modification
- Poly-ALA autobuilding

2 - Using SHELX C/D/E

Graphical User Interface HKL2MAP



http://webapps.embl-hamburg.de/hkl2map/

Documentation

Most shelx programs issue “short” usage instruction when called without an argument.

```
tg@slartibartfast:~$ shelxc

  ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
  +  SHELXC - Create input files for SHELXD and SHELXE - Version 2013/2  +
  +  Copyright (c) George M. Sheldrick 2003-13                           +
  +                                  Started at 13:59:57 on 10 Jun 2014  +
  ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

 SHELXC reads a filename stem (denoted here by 'xx') on the command line
 plus some instructions from 'standard input'. It writes some statistics to
 'standard output' and prepares the three files needed to run SHELXD and
 SHELXE. SHELXC can be called from a GUI by a command line such as:

 shelxc xx <t

 which would read the instructions from the file t, or (under most UNIX
 systems) by a simple shell script that includes the instructions, e.g.

 shelxc xx <<EOF
 CELL 49.70 57.90 74.17 90 90 90
 SPAG P212121
 SAD elastase.sca
 FIND 12
 <<EOF
 shelxd xx_fa
 shelxe xx xx_fa -s0.37 -m20 -h -b
 shelxe xx xx_fa -s0.37 -m20 -h -b -i
```

More information including tutorials available at http://shelx.uni-goettingen.de/SHELX/.

Shelxc Data Preparation: Keywords

shelxc can be used for six different phasing scenarios:

SAD
- `SAD`
- (`NAT`)

SIR
- `SIR`
- `NAT`

SIRAS
- `SIRA`
- `NAT`

MAD
- (`NAT`)
- `PEAK`
- `INFL`
- `LREM`
- `HREM`

RIP
- `BEFORE`
- `AFTER`
- (`NAT`)

Each keyword takes the filename of the corresponding integrated dataset.

Running shelxc

1. Create input command file `shelxc.inp` with text editor

```
 CELL 49.70   57.90   74.17  90.000  90.000  90.000
SPAG P212121
FIND 12
NTRY 100
SFAC S
SAD elastase.sca
```

2. `shelxc mysad < shelxc.inp`

Shelxc Output Files

The command "`shelxc mysad < shelxc.inp`" creates three files:

mysad_fa.ins Text file with instructions for shelxd

mysad_fa.hkl Artificial substructure data set from which shelxd determines substructure coordinates. Each line contains

$$h, k, l, \left| |F^{+}(hkl)| - |F^{-}(hkl)| \right|, \alpha$$

α is not used by shelxd, but by shelxe to calculate an initial phase estimate for the protein structure as

$$\phi_T(hkl) = \phi_A(hkl) + \alpha(hkl)$$

MAD/SIRAS: exact α; SIR or SAD: rough estimate of α

ϕ_A is the phase angle calculated from the substructure coordinates determined by shelxd.

mysad.hkl native data used by shelxe for phasing and density modification

Shelxc: Resolution Cut-off for Anomalous Signal

```
 85349 Reflections read from SAD file XDS_pk1pk2.HKL

12186 Unique reflections, highest resolution  7.199 Angstroms
141.7 Friedel pairs used on average for local scaling

Resl.    Inf. 16.01 12.71 11.10 10.09  9.36  8.81  8.37  8.01  7.70  7.43  7.20
N(data)    1149  1124  1115  1099  1108  1102  1126  1078  1091  1122  1072
Chi-sq     1.30  1.22  1.24  1.27  1.37  1.55  1.58  1.62  1.49  1.35  1.43
<I/sig>    41.7  28.7  24.4  22.1  16.5  13.3   9.2   7.4   5.3   3.8   2.2
%Complete  98.9  99.9 100.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0  98.2
<d"/sig>   4.82  2.79  1.90  1.53  1.26  1.04  1.00  0.89  0.85  0.80  0.71
CC(1/2)    94.8  81.8  65.2  53.8  39.6  25.0  23.6   9.3   5.3  -6.6 -14.4
```

- anomalous signal where $CC(1/2) > 30\%$
- anomalous signal where `<d"/sig> > 1.3`
- $CC > 30\%$ usually more reliable than `<d"/sig> > 1.3`

3 - Shelxd

Shelxd — Finding the Substructure

```
shelxd mysad_fa
```

reads the "substructure data" `mysad_fa.hkl` and its instructions from `mysad_fa.ins`. The most important entries in `mysad_fa.ins`:

SFAC SE atom type to look for

FIND 12 expected number of substructure atoms, should be within 20 % of the actual number (try several for *e.g.* a soak where the number is not known)

SHEL 999 3.3 resolution limits of the **anomalous signal**. High resolution limit can be critical, but the default of $d_{\text{min}} + 0.5$ Å works well in many cases.

NTRY 10000 number of trials.

Shelxd Output

While `shelxd` runs, the best solution is written to `mysad_fa.res` which contains the substructure coordinates in fractional coordinates and which is later read by shelxe.

```
REM Best SHELXD solution:   CC 60.74   CC(weak) 49.22   CFOM 109.96
TITL mysad_fa.ins MAD in C2
CELL  0.98000  109.02   61.75   71.74   90.00   97.08   90.00
LATT  -7
SYMM -X, Y, -Z
SFAC SE
UNIT   192
SE01   1  0.758774  0.508636  0.246391   1.0000  0.2
SE02   1  0.792908  0.398262  0.138903   0.8845  0.2
                [...]
SE10   1  0.925819  0.231575  0.191291   0.5569  0.2
SE11   1  0.495239  0.183609  0.416278   0.5352  0.2
SE12   1  0.643097  0.029221  0.210653   0.4897  0.2 <---
SE13   1  0.811539  0.048553  0.227752   0.1453  0.2 <---
SE14   1  0.600281  0.156860  0.149628   0.0764  0.2
HKLF 3
END
```

The sixth column contains the occupancy of the corresponding atom. A sharp drop (here between SE12 and SE13) is a promising sign of a correct solution. The correlation coefficient (CC and CCweak) in the first line measures the reliability of the solution

For SAD, a CC of more than 30 % is a safe sign of a correct solution, for MAD the limit is about 40 %.

Shelxd Output

SHELXD is very robust. Attention should be paid to

1. The **resolution** at which the data are truncated, *e.g.* where the internal CC (CC1/2) between the signed anomalous differences of two randomly chosen reflection subsets falls below 30%.

2. The **number of sites** requested should be within about 20% of the true value.

3. In the case of a soak, the rejection of sites on **special positions** should be switched off.

4. For S-SAD, DSUL (**search for disulfides**) can be very useful.

5. In difficult cases it may be necessary to run more trials (say 50000).

Fine tuning SHELXD substructure solution

SHELXD is very fast and robust, but achieves this with the help of drastic assumptions.

In borderline cases it may be worth using the LLG (log likelihood gain) to distinguish substructure solutions, e.g. using the programs SHARP, CRANK2 or PHASER. For details see:

SHARP Methods Enzymol. 276 (1997) 472-494; Acta Cryst. D59 (2003) 2023-2030.

CRANK2 Acta Cryst. D67 (2011) 331-337; Nat. Commun. 4:2777 (2013).

PHASER (for experimental phasing) Acta Cryst. D60 (2004) 1220-1228; Acta Cryst. D67 (2011) 338-344.

These programs could also be used to refine and extend the heavy atom substructure before density modification and poly-Ala tracing with SHELXE. In general LLG-based methods require more detailed information (e.g. which elements are present) than SHELXC/D/E, and they tend to be slower.

4 - Shelxe

Shelxe: Phasing, Density Modification, Model Building

No `.ins`-tructions file. All parameters provided as command line options **after** data file names.

A typical and one of the most simple command line could be

```
shelxe mysad mysad_fa -s0.65 -h -a
```

`mysad` read native data `mysad.hkl`

`mysad_fa` read angle estimate for α from `mysad_fa.hkl`, substructure coordinates from `mysad_fa.res` (the shelxd output)

`-s0.65` Assume a solvent content of 65%. It should be reasonably well estimated.

`-h` substructure atoms present in native data `mysad.hkl`

`-a` run 5 (default) cycles of poly-ALA autotracing.

Shelxe `-i`: Inverted Substructure

It is impossible to distinguish the substructure from its enantiomorph with the anomalous data and there is a 50 % chance that the coordinates in `mysad_fa.res` are inverted w.r.t. the correct substructure.

Therefore shelxe must always be run **twice**

- with the **direct substructure**
- with the **inverted substructure**, *i.e.* with the same options as the direct hand *plus* the switch `-i`. This inverts the hand and takes care of everything necessary
  - inversion of screw axes, $P4_1$ to $P4_3$
  - off-axis inversion for $I4_1$ (1-x, 1/2-y, 1-z); $I4_122$ (1-x, 1/2-y, 1/4-z); $F4_132$ (1/4-x, 1/4-y, 1/4-z)
- output files are automatically amended by `_i` to distinguish the two runs.

N.B. if the inverted hand turns out to be the correct hand, your **space group may change** - *e.g.* in the **presence of screw axes**. Keep this in mind when you convert your native data to *e.g.* mtz-format!

Caveat: Substructure Resolution

- "Normal" macromolecular structure: Determine atom positions even at *e.g.* 5 Å resolution **because of restraints**.

- Substructure unrestrained

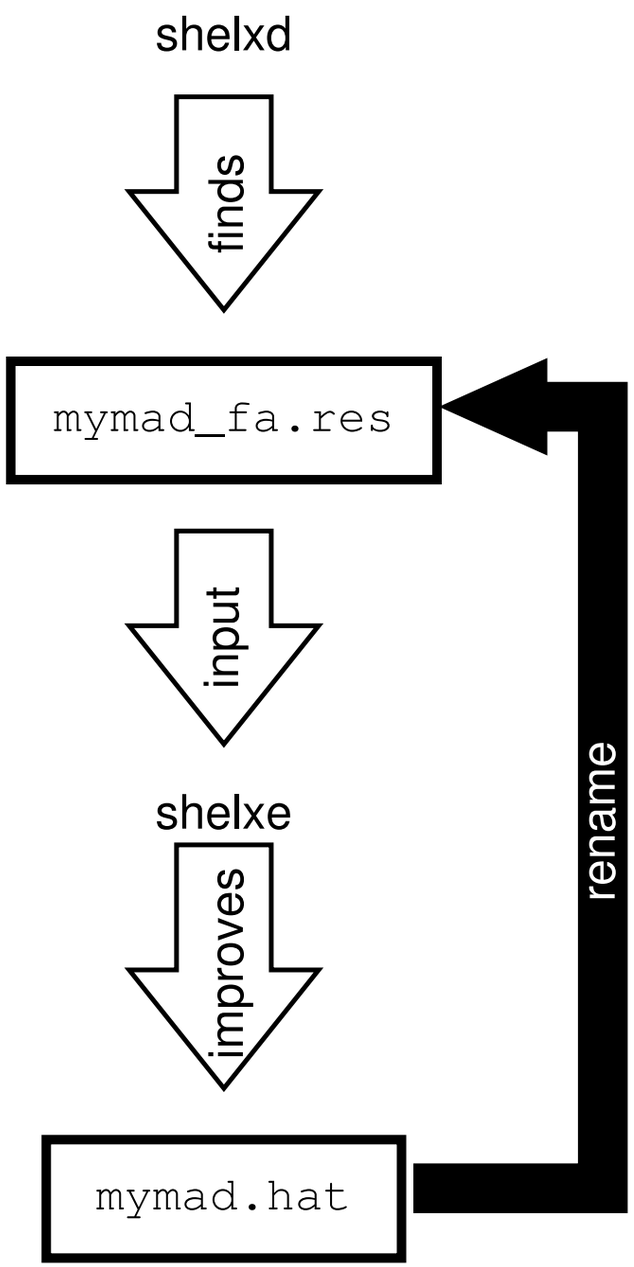
$\Rightarrow$ coordinates only know within resolution of anomalous signal, often much worse than 3 Å

Way out:

1. *e.g.* Sharp improves substructure coordinates **before** density modification

2. "**substructure recycling**" with shelxe

Shelxe: Substructure Recycling



- Better substructure = better starting phases = better map
- **Caveat:** If the inverted structure turns out to be the correct hand (*i.e.* `mysad_i.hat` from the `-i`-run of shelxe), the second run of shelxe must be run **without** the `-i` switch:

Shelxe: Did it work?

Criteria to tell if phasing worked:

1. Correct hand shows better **Contrast**, especially at early cycles of density modification.

2. Correct hand has higher **map correlation coefficient** throughout resolution range:

```
d     inf - 4.66 - 3.70 - 3.23 - 2.93 - 2.72 - 2.56 - 2.43 - 2.33 - 2.24 - 2.15
<mapCC> 0.626  0.795  0.775  0.754  0.819  0.804  0.756  0.694  0.620  0.582    direct
<mapCC> 0.810  0.877  0.845  0.844  0.874  0.856  0.840  0.830  0.839  0.809    inverse
```

3. A reasonable poly-ALA trace (average 10 residues per chain) and a $CC > 25\%$

When using the auto-tracing option (`-a`) in shelxe, the first two figures (contrast/ mapCC) become meaningless, but in this case the poly-ALA trace is much more conclusive.

Shelxe: Structure Solved?

Indicators from shelxe:

```
TITLE  elastase.pdb   Cycle   3   CC = 41.91% 226 residues in  4 chains
TITLE  elastase_i.pdb Cycle   3   CC =  7.26%  61 residues in  7  chains
```

1. CC>25%
2. average chain length > 10 (here: 56.5 *vs.* 8.7)
3. jump in CC over many cycles (*e.g.* with `-a50`)

Shelxe: Structure Solved?

Coot reads `mysad.pdb` (poly-ALA trace) and `mysad.phs` (map).

Elastase SAD tutorial

Direct hand Inverse hand

SHELXE: Current and Future Developments

- SHELXE started with **-x** and a reference PDB file **name.ent** is present: the mean phase error is output at various stages. The necessary origin shift is determined on the fly.

- If **-h** is also set, the program finds the atom in the reference file nearest to each heavy atom site. This is particularly useful for checking the substructure.

- The density modification has been improved for **SAD phasing**. For 20 test structures the mean phase improvement after the first round of density modification was 4.6°.

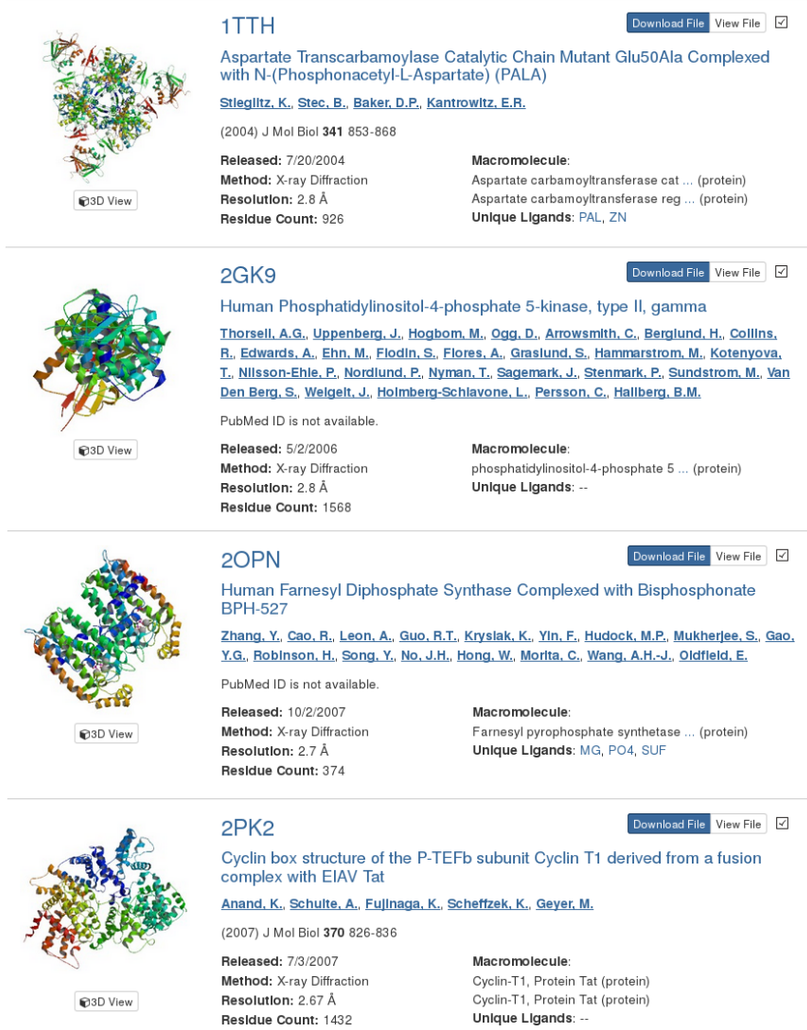
- These improvements are already in the **current distributed** version. In addition, SHELXC is being adapted to handle multiple SAD datasets and a major rewrite of SHELXE is in progress.

5 - Structure Refinement with SHELXL

Applications for SHELXL

- High resolution structures < 1.4Å (inorganic, organic, and macromolecular structures, *cf.* Cambridge Crystallographic Database, $> 875,000$ structures, Inorganic Crystallographic Database, $> 76,000$ structures)
- Occupancy refinement [3]
- Standard Uncertainties for parameters [4]
- Highly reliable across all 230 spacegroups
- Extremely versatile, complicated chemical situations
- Applicable to X–ray, neutron [5], and electron diffraction

Against common belief, SHELXL is suitable for mid–resolution, large complex structures. This is in particular true at the end stage of refinement for special studies of occupancy refinement or .



Some Features of SHELXL

- All parameters can be fixed, or refined, or tightened together
- *Free variables* enable complicated networks of disorder
- Full control over parameters, restraints, and constraints
- Twin refinement

Further reading: The SHELXL book, Müller, P., Herbst-Irmer, R., Spek, A., Schneider, T.R. & Sawaya, M.R. (2006). Crystal Structure Refinement: A crystallographer's guide to SHELXL. IUCr/Oxford University Press.

Running SHELXL

Input:

`myfilename.ins`	`myfilename.hkl`
Instructions: Coordinates + Re–/Constraints	Data: Reflections

Corresponds to:

PDB–file	mtz–file

```
#> shelxl myfilename
```

Output:

`myfilename.res`	`myfilename.fcf`	`myfilename.lst`	(`myfilename.cif`)
Updated `ins`: Coordinates + Parameters	Map file Coot: Model Building	Log–file	Deposition Validation

Getting Started

`ins`–file Use the program `pdb2ins` (A. Luebben, distributed *via* SHELX website). Converts PDB–file to ins–file, including

- Engh–Huber restraints
- and instructions for hydrogen atoms (`AFIX`–command)

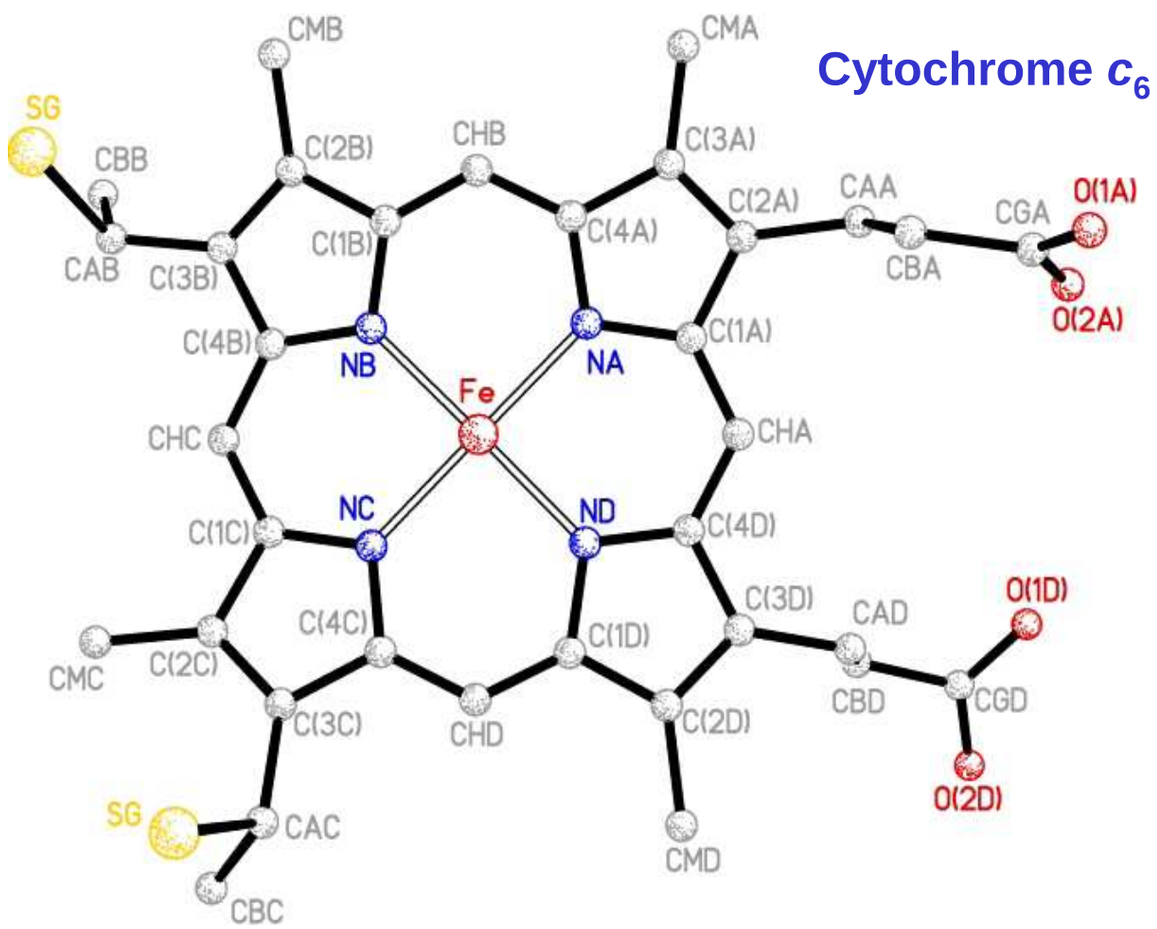
`hkl`–file : `xprep` or `shelxc` to converts `XDS_ASCII.HKL` to `hkl`–format;

`mtz2hkl` converts mtz–file to `hkl`–format (DIALS / MOSFLM)

Free Variables (slide courtesy George Sheldrick)

Use of free variables to obtain mean distances with esds

The following input refines fv 2, 3 and 4 to be the mean Fe-N, N-C and N…CH distances. Because of the 4- and 8-fold redundancy, accurate values are obtained that can be used as restraints.



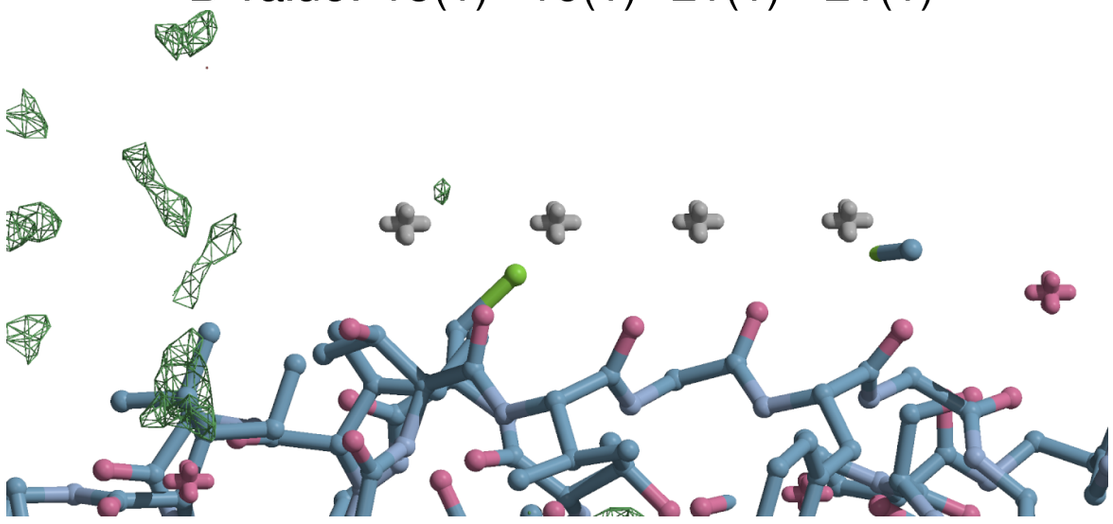

```
FVAR 1.0 1.8 1.4 2.4
DFIX_HEM 21 Fe NA Fe NB Fe NC Fe ND
DFIX_HEM 31 NA C1A NA C4A NB C1B NB C4B NC C1C NC C4C ND C1D ND C4D
DFIX_HEM 41 NA CHA NA CHB NB CHB NB CHC NC CHC NC CHD ND CHD ND CHA
etc...
```

Example: Ion occupancy in K^+ Channels

3LDC: MthK pore with 100mM K^+, $d_{\text{min}} = 1.45$Å

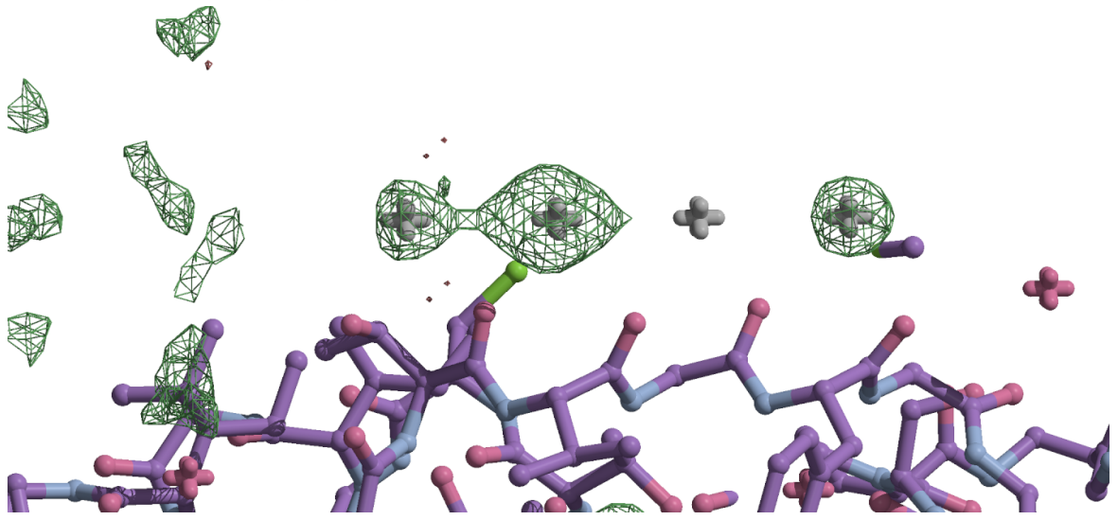
Knock–on, Refined occupancies

	S4	S3	S2	S1
occupancy:			1.01(3)	
B-value:	18(1)	16(1)	21(1)	21(1)



***KWKW*–Fixed occupancies: 50% K^+, 50% H_2O**

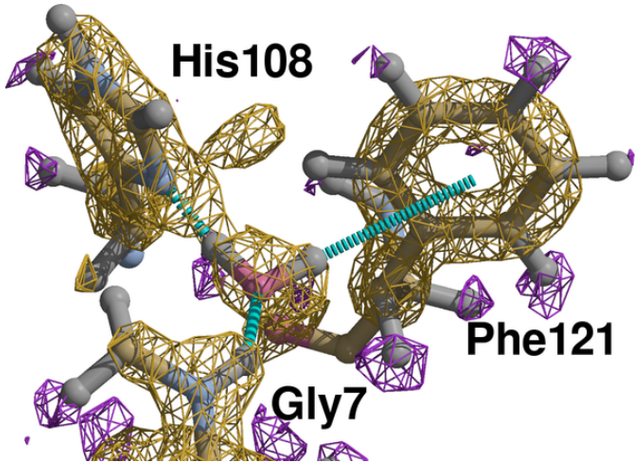
	S4	S3	S2	S1
occupancy:	0.50	0.50	0.50	0.50
B-value:	12	12	11	12



Crystallography supports the knock–on model of K^+ ion exchange

Example: Refinement against Neutron Data [5]

1) Hydrogen bond analysis with the `HTAB` command (PDB–ID 2ZOI):



D-H$\cdots$A	$d(D-H)$	$d(H\cdots A)$	$d(D\cdots A)$	$\angle(DHA)$
(Gly7) N-H$\cdots$O	1.01(19)Å	1.8(2)Å	2.7(3)Å	141(20)$^\circ$
O-D1$\cdots$ND1 (His108)	1.0(2)Å	1.5(2)Å	2.5(3)Å	175(25)$^\circ$

Example: Refinement against Neutron Data [5]

2) Deuterium Saturation in Perdeuterated Proteins (PDB–ID 3RZT) (Deuterium is very hygroscopic)

- Group three classes of chemical bonds for $H \leftrightarrow D$ exchange for $N-D$ and $O-D$ (including all water molecules)
- Refine group occupancy (*via* `free variable`)
- Calculate fraction of D and H

$$f_2 * b_c(D) = p * b_c(D) + (1-p) * b_c(H)$$
$$p = \frac{6.674 * f_2 - (-3.741)}{6.674 - (-3.741)}$$

- Result : $p = 93\%$

References and Further Reading

1. SHELX C/D/E: G. M. Sheldrick, Acta Cryst. (2010), D**66**, 479–485
2. Visit the shelx web page for documentation, tutorials, *etc.*: `http://shelx.uni-goettingen.de`
3. *Ion Permeation in K^+ Channels Occurs by Direct Coulomb Knock-On*, D. A. Köpfer, C. Song, T. Gruene, G. M. Sheldrick, U. Zachariae, B. L. de Groot, Science (2014), Vol. 346, 352–355
4. *Unexpected tautomeric equilibria of the carbanion-enamine intermediate in pyruvate oxidase highlight unrecognized chemical versatility of the thiamin cofactor*, Meyer, D., Neumann, P., Koers, E., Sjuts, H., Ludtke, S., Sheldrick, G. M., Ficner, R., Tittmann, K., PNAS (2012), Vol. 109, 10867–10872
5. *Refinement of Macromolecular Structures against Neutron Data with SHELXL–2013*, T Gruene, HW Hahn, AV Luebben, F Meilleur, GM Sheldrick, J. Appl. Cryst (2014), Vol. 47, 462–466

Availability

SHELX is available free for academic use via the SHELX homepage `http://shelx.uni-goettingen.de/`. Extensive documentation and many links to useful programs may also be found there. SHELX C/D/E are also distributed along with CCP4.