SHELXD - Direct Methods for Larger Structures

SHELX Workshop, Göttingen September 2011

George M. Sheldrick

http://shelx.uni-ac.gwdg.de/SHELX/

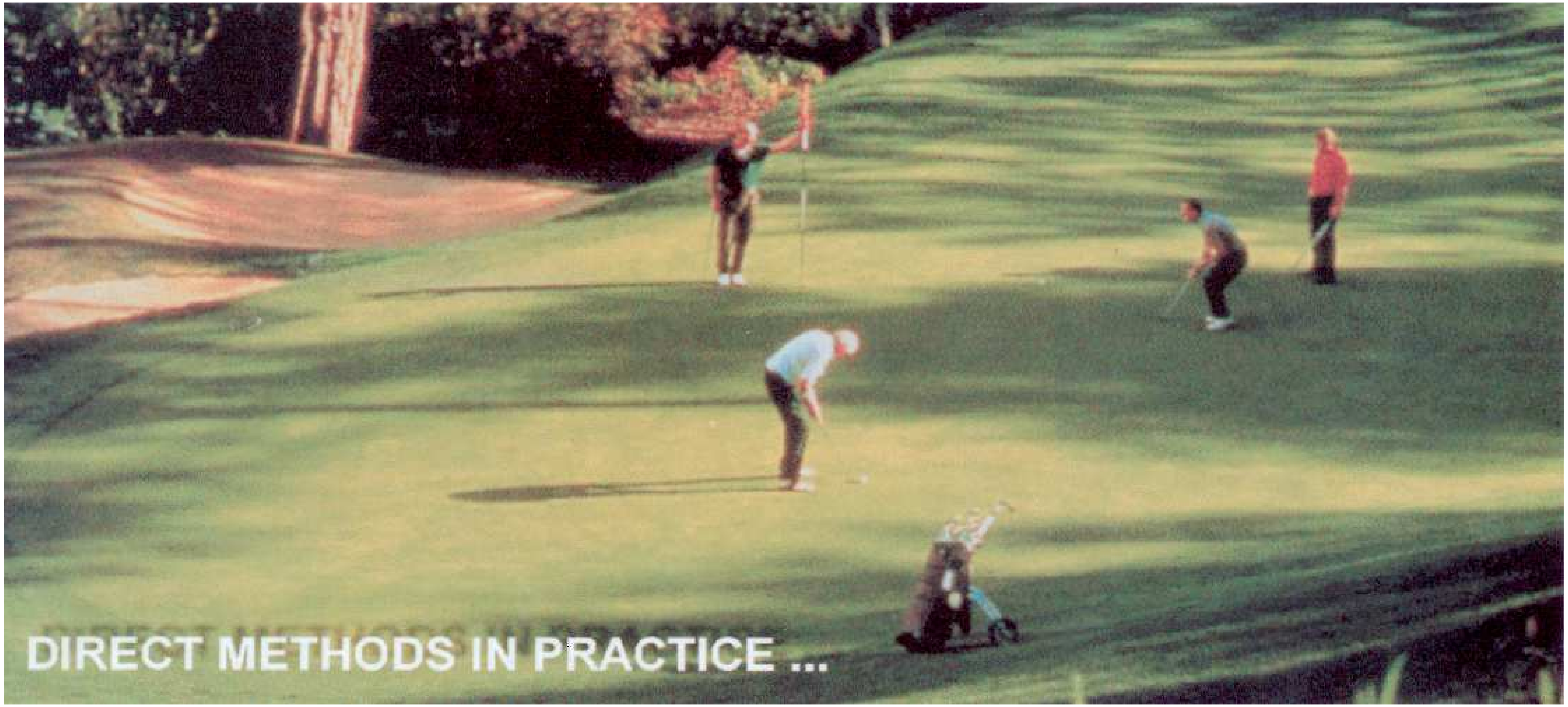
The crystallographic phase problem

- **In order to calculate an electron density map, we require both the intensities $I = |F|^2$ and the phases ϕ of the reflections *hkl*.**
- **The information content of the phases is appreciably greater than that of the intensities.**
- **Unfortunately, it is almost impossible to measure the phases experimentally!**

This is known as the *crystallographic phase problem* and would appear to be difficult to solve!

Despite this, for the vast majority of small-molecule structures the phase problem is solved routinely in a few seconds by black box *direct methods*.

Finding the minimum



Normalized structure factors

Direct methods turn out to be more effective if we modify the observed structure factors to take out the effects of atomic thermal motion and the electron density distribution in an atom. The normalized structure factors E_h correspond to structure factors calculated for a point atom structure.

$$E_h^2 = (F_h^2/\varepsilon) / \langle F^2/\varepsilon \rangle_{\text{resl. shell}}$$

where ε is a statistical factor, usually unity except for special reflections (e.g. 00*l* in a tetragonal space group). $\langle F^2/\varepsilon \rangle$ may be used directly or may be fitted to an exponential function (Wilson plot).

The tangent formula (Karle & Hauptman, 1956)

The tangent formula, usually in heavily disguised form, is a key formula in small-molecule direct methods:

$$\tan(\phi_h) = \frac{\sum_{h'} | E_{h'} E_{h-h'} | \sin(\phi_{h'} + \phi_{h-h'})}{\sum_{h'} | E_{h'} E_{h-h'} | \cos(\phi_{h'} + \phi_{h-h'})}$$

The sign of the sine summation gives the sign of sin(ϕ_h) and the sign of the cosine summation gives the sign of cos(ϕ_h), so the resulting phase angle is in the range 0-360°.

In the program MULTAN (Main & Woolfson) that was widely used between 1969 and 1990, this formula was used to extend and refine phases, starting with fixed or permuted (e.g. 45/135/225/315°) phases for a small number of large E-values. Later Yao Jia-Xing started from random phases (RANTAN).

The limits of the tangent formula

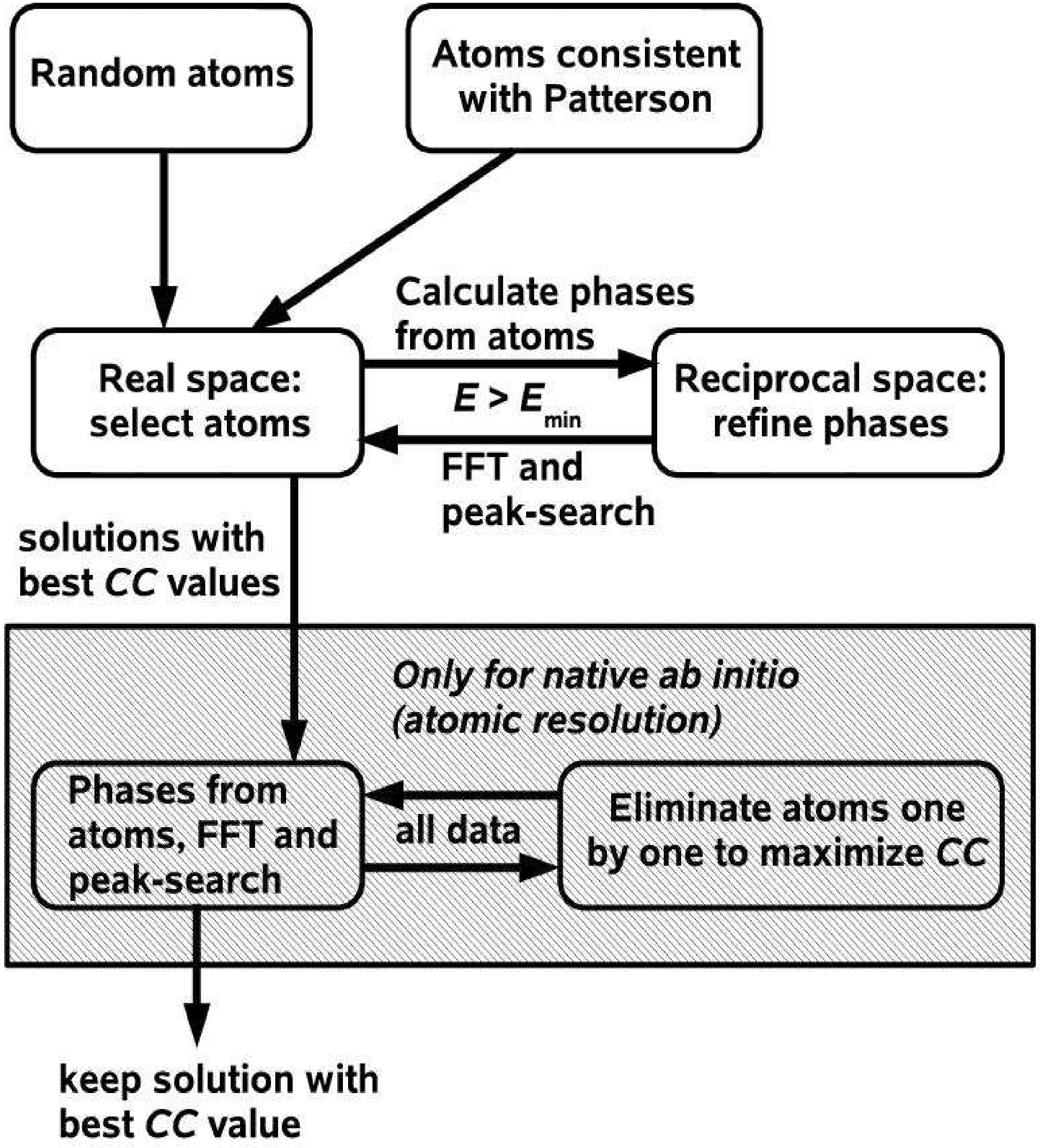
Conventional direct methods based on ‘improved’ versions of the tangent formula and implemented in programs such as MULTAN, RANTAN, SAYTAN, DIRDIF, SIR, SHELXS etc. were extremely efficient at solving small molecule structures up to 100 unique atoms but only succeeded in solving a handful of structures larger than about 200 unique atoms.

The efficiency of the tangent formula as a phase space search engine lies in its ability to correlate phases widely distributed in reciprocal space. Despite its computational efficiency, the tangent formula tends to lose enantiomorph discrimination and drifts towards *uranium atom* types of false solution, especially for larger structures. It was clearly necessary to constrain the phase set more tightly to be chemically reasonable.

Dual space recycling in SHELXD

Dual space recycling was introduced with the SnB program by Weeks, Miller, Hauptman *et al.* in 1993. The real space part of the cycle imposes a strong *atomicity* constraint on the phases that are refined in the reciprocal space part.

CC is the correlation coeff. between E_{obs} and E_{calc}.


Random atoms

Atoms consistent with Patterson

Calculate phases from atoms

Real space: select atoms

$E > E_{min}$

Reciprocal space: refine phases

FFT and peak-search

solutions with best *CC* values

Only for native ab initio (atomic resolution)

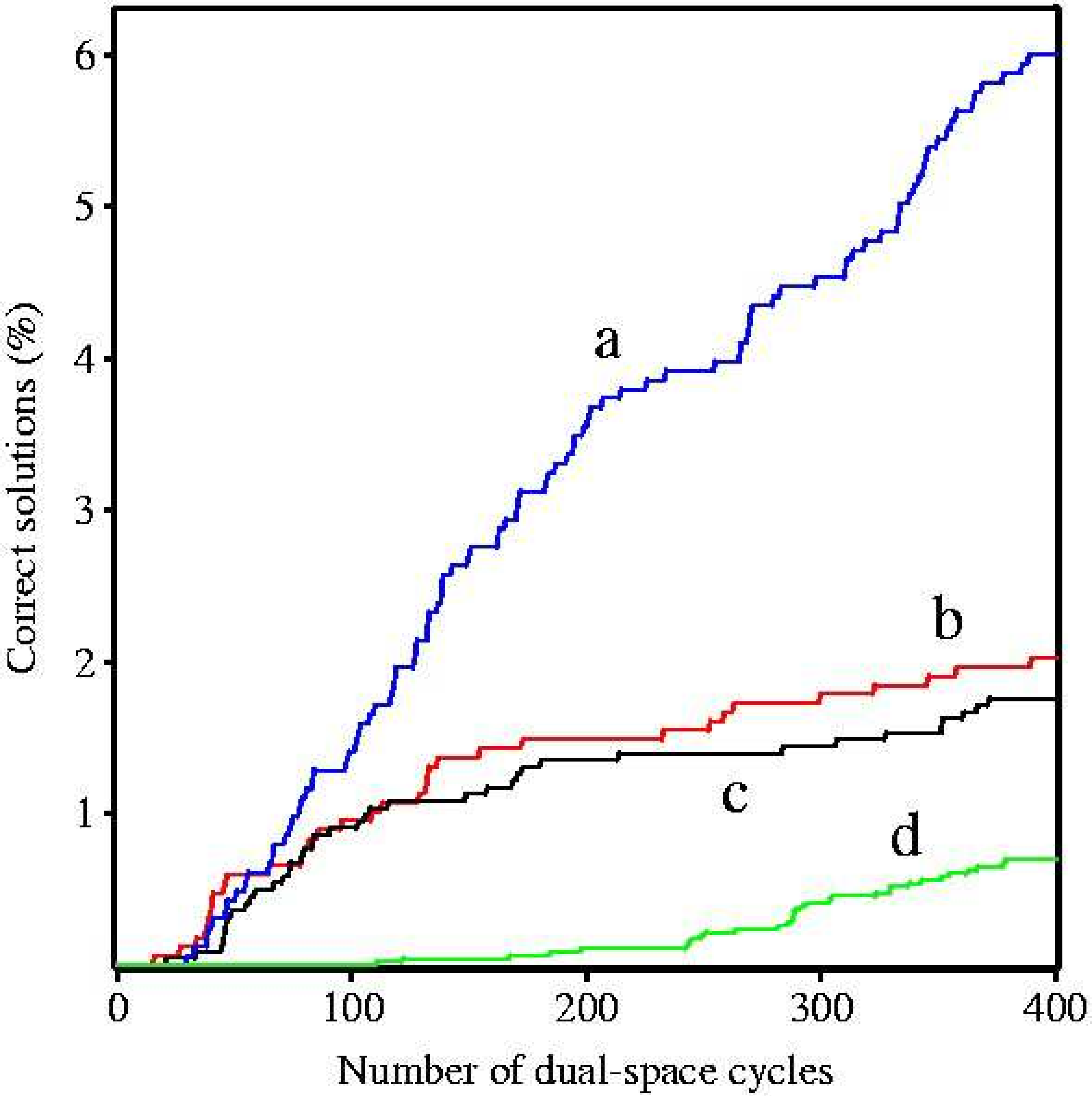
Phases from atoms, FFT and peak-search

all data

Eliminate atoms one by one to maximize CC

keep solution with best *CC* value

Gramicidin A (N=317) - different strategies



a: random omit + tangent expansion

b: random omit + minimal function

c: top N peaks + minimal function

d: random omit + no phase refinement

Random OMIT maps

Omit maps **were frequently used by protein crystallographers to reduce** ***model bias*** **when interpreting unclear regions of a structure. A small part (<10%) of the model is deleted, then the rest of the structure refined (often with simulated annealing to reduce memory effects) and finally a new difference electron density map is calculated.**

A key feature of SHELXD is the use of ***random omit maps*** **in the search stage. About 30% of the peaks are omitted at random and the phases calculated from the rest are refined. The resulting phases and observed** ***E*****-values are used to calculate the next map, followed by a peaksearch. This procedure is repeated 20 to 500 times.**

Unknown structures solved by SHELXD

Compound	Sp. Grp.	N(mol)	N(+solv)	HA	d(Å)
Hirustasin	$P4_32_12$	402	467	10S	1.20
Cyclodextrin	$P2_1$	448	467		0.88
Decaplanin	$P2_1$	448	635	4Cl	1.00
Cyclodextrin	P1	483	562		1.00
Bucandin	C2	516	634	10S	1.05
Amylose-CA26	P1	624	771		1.10
Viscotoxin B2	$P2_12_12_1$	722	818	12S	1.05
Mersacidin	$P3_2$*	750	826	24S	1.04
Feglimycin	$P6_5$*	828	1026		1.10
Tsuchimycin	P1	1069	1283	24Ca	1.00
rc-WT Cv HiPIP	$P2_12_12_1$	1264	1599	8Fe	1.20
Cytochrome c3	$P3_1$	2024	2208	8Fe	1.20

*twinned

The largest substructure solved so far was probably 197 correct Se out of a possible 205 by Qingping Xu of the JCSG (PDB 2PNK).

Probabilistic Patterson Sampling

When large structures contain heavier atoms (e.g. metal atoms or even sulfur), a very effective approach is to first to find them and then to expand to the full structure.

Each unique general Patterson vector of suitable length is a potential HA-HA vector, and may be employed as a 2-atom search fragment in a translational search based on the *Patterson minimum function*. For each position of the two atoms in the cell, the Patterson height P_j is found for all vectors between them and their symmetry equivalents, and the sum (PSUM) of the lowest (say) 35% of P_j calculated. An effective approach is to generate many different starting positions by simply taking the best (in terms of PSUM) of a finite number of random positions of the two-atom vector each time.

The *full-symmetry Patterson superposition minimum function* can be used to expand from the two atoms to a much larger number before entering the dual-space recycling.

Structure solution in P1

It is well established [e.g. Sheldrick & Gould, *Acta Cryst.* B51 (1995) 423-431; Xu et al., *Acta Cryst.* D56 (2000) 238-240; Burla et al., *J. Appl. Cryst.* 33 (2000) 307-311] that it may be more efficient to solve structures in P1 and then search for the symmetry elements later. This works particularly well for solving $P\bar{1}$ structures in P1.

A threefold Patterson minimum function or a rotation (but not translation) search for a fragment of known geometry are efficient ways of initiating the P1 phase determination.

In general the success rate is much higher in P1, but the quality of the resulting map is better in the true space group (as a result of averaging over symmetry equivalents). For this reason, expansion to P1 is not recommended for very high symmetry space groups.

Fine tuning SHELXD for large structures

1. **Use the inner loop (FIND) just to find heavy atoms (with the help of a *super-sharp* Patterson, PSMF –4) and the outer loop (PLOP) to expand to the full structure.**

2. **If the distribution of peak heights indicates a tendency to produce uranium atom solutions, increase the number of phases held fixed in tangent expansion by increasing the second TANG parameter (e.g. TANG 0.95 0.6).**

3. **Expand the data to P1 (TRIC), then find the true symmetry later (e.g. with ADDSYM SHX in PLATON).**

4. **If the data are borderline for atomic resolution, rename the resulting *name.res* file to *name.ins* and use SHELXE to produce a map (e.g. shelxe name −m20 −s0.4 −e1). This is also a good way of inverting the structure if required (direct methods cannot distinguish enantiomorphs), and also now enables polypeptides to be *autotraced*.**

The 1.2 Å rule

“Experience with a large number of structures has led us to formulate the empirical rule that if fewer than half the number of theoretically measurable reflections in the range 1.1-1.2 Å are “observed”, it is very unlikely that the structure can be solved by direct methods” (Sheldrick, 1990).

When heavier atoms such as Fe or even S are present, this rule can be relaxed, and most of the larger structures solved *ab initio* contain such atoms.

Charge flipping

Charge flipping (Oszlányl & Sütő, *Acta Cryst.* A60 (2004) 134-141; A64 (2008) 123-134) is a dual-space algorithm in which the density is modified by flipping low and negative density. It has one highly critical parameter: the electron density level δ below which the density should be 'flipped':

$$\text{If } \rho \geq \delta \text{ then } \rho' = \rho; \quad \text{if } \rho < \delta \text{ then } \rho' = -\rho$$

A suitable value for δ is about 1.3 times the square root of the variance of the density. Charge flipping is simple and easy to program, but requires data expanded to space group P1. As with other direct methods, it appears to be better to use *E*-values rather than *F*.

It is essential to tidy up the solution by e.g. several Fourier cycles in which low density is truncated but no flipping is applied.

A very similar idea – 'solvent flipping' – has been widely used for improving the electron density maps of macromolecules [Abrahams, *Acta Cryst.* (1997) D53, 371-376].

Implementations of charge flipping

PLATON can expand the data to P1, apply charge flipping and then search for the correct space group and labelling of the axes. The graphical display of the progress of the charge flipping is instructive.

OLEX2 can use the smtbx toolbox from Ralf Grosse-Kunstleve to do the same.

SUPERFLIP and EMDA (Palatinus & Chapuis, *J. Appl. Cryst.* 40 (2007) 786-790; Palatinus & van der Lee, *J. Appl. Cryst.* 41 (2008) 975-984) also do the same, but are valid for any number of dimensions and so are suitable for solving modulated structures.

Advantages and disadvantages of charge flipping

Charge flipping does not need to know the chemical formula but tends to lose light atoms in the noise in the presence of heavy atoms. Because it does not explicitly assume atoms and is easily extended to more than three dimensions, it can also solve modulated structures.

Charge flipping has the advantage that it uses all the data but the disadvantage that it also requires rather complete data! The resolution requirements are at least as critical as for dual-space methods that assume atoms.

Charge flipping uses all the data, which may be an advantage for pseudo-symmetric structures in which some classes of reflections are much stronger than others, SHELXD just uses the strongest *E*-values.

Both charge flipping and SHELXD can be quite slow for large equal-atom structures. The multi-CPU version of SHELXD is about 29 times faster on a 32-CPU workstation.

The role of the weak reflections

Expansion to P1 and subsequent search for the space group can work well even when the systematic absences are misleading. However, the fact that in most cases such missing reflections help to define the translational symmetry suggests that it is important to retain them when expanding the data to P1.

Whether or not a very weak reflection is included in the Fourier summation has little effect on the resulting electron density. On the other hand, a reflection that is calculated to be strong but was measured as weak tells us that something is wrong with the model (i.e. the current electron density).

One way to use this information would be to use a 2$|F_o|–|F_c|$ Fourier rather than a $|F_o|$ Fourier to calculate the electron density. Another is to add 90° to the calculated phase and use $|F_c|$ rather than $|F_o|$ as the amplitude (Oszlányl & Sütő, *Acta Cryst.* A61 (2005) 147-152); this is not easy to explain but works well in practice.

Conclusions

Expansion to P1, provided that the true symmetry is not too high, results in a major improvement in the success rate of most direct methods. It is particularly useful when the space group is uncertain.

Methods that are based on atoms are less sensitive to poor resolution or missing data. The presence of a few heavier atoms considerably improves the success rate of all methods and also allows ‘Patterson seeding’, but light atoms can get lost in the noise in the presence of very heavy atoms.

The use of only the largest *E*-values (required for the tangent formula) is likely to fail if strong pseudosymmetry causes some classes of reflections to be much stronger than others, though renormalization according to parity group can help.

It is always useful to have a choice of different algorithms, it is often not clear why some work better than others for a particular structure!

Acknowledgements

I am particularly grateful to Isabel Usón and many SHELX users for suggestions and discussions.

Standard SHELX reference:
Sheldrick, G.M. (2008). *Acta Cryst*. A64, 112-122.

SHELXD: Usón & Sheldrick (1999), *Curr. Opin. Struct. Biol.* 9, 643-648; Sheldrick, Hauptman, Weeks, Miller & Usón (2001), *International Tables for Crystallography Vol. F*, eds. Arnold & Rossmann, pp. 333-351.