Data scaling with the Bruker programs SADABS and TWINABS

ACA Philadelphia, July 28th 2015

George M. Sheldrick

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

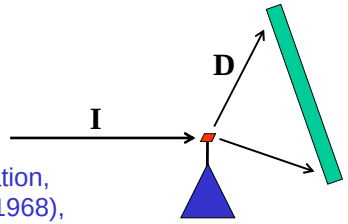
SADABS strategy

1. Determine scaling and absorption parameters by fitting individual intensities to the mean corrected intensities (averaged over equivalents). Outliers are flexibly downweighted but not rejected in this stage. For parameter determination Friedel opposites should be treated as equivalent (i.e. Laue group symmetry imposed) but they should not be merged in the resulting *.hkl* file.

2. Delete a small number of reflections that are completely incompatible with their equivalents, e.g. reflections blocked by the beam stop etc. Then determine an error model for the remaining reflections by fitting χ^2 to unity to put $\sigma(I)$ onto an absolute scale.

3. Output diagnostic statistics (graphically) and corrected data.

Scaling using equivalent reflections in SADABS



Scaling is based on the following approximation, similar to that used by Kopfmann & Huber (1968), North, Phillips & Mathews (1968), Blessing (1997) and many other papers and programs :

$$I_c = I_o\, S(n)\, P(u,v,w)\, Q(\mu r, 2\theta)$$

S = Incident beam correction (one scale factor per frame n), *P* = Diffracted beam correction using direction cosines *u,v,w, Q* = Spherical crystal factor.

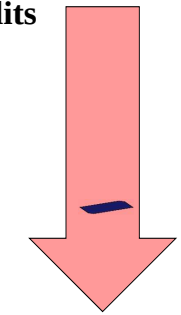
S and *P* are refined alternately to minimize $\Sigma w(\langle I_c\rangle - I_c)^2$, where $\langle I_c\rangle$ is the mean of a group of equivalents. These refinements are linear so no starting values are required. Outliers are dynamically down-weighted.

Sealed tube plus slits vs. focused beam

Sealed tube + slits: If the beam is wider than the crystal and absorption is small, the diffracted intensity should not vary much with crystal orientation. An empirical correction using spherical harmonics can easily cover this. If the crystal absorbs strongly, a face-indexed correction assuming that the crystal is uniformly bathed in the beam works well.

Microsource (or synchrotron) focused beam: The beam will often have an approximately Gaussian profile and the crystal is not uniformly bathed in the beam; often parts of the crystal are outside the beam. **Empirical corrections for this are essential** and require a ***high redundancy*** and (if the absorption is strong) also ***small crystals***.

Monochromator + slits



then and now

Focussing optics

Crystal bathed in beam of uniform intensity. Effective diffracting volume does not change on rotating.

Crystal smaller than beam of approximate Gaussian profile. Effective diffracting volume varies with crystal orientation.

Shutterless data collecton

Although CMOS detectors are noisier than CCDs, this is compensated for by the much larger active area and very fast readout, leading to a higher redundancy in the same overall time. ***Shutterless*** data collection, in which the crystal is rotated at constant speed, also eliminates systematic errors associated with opening and closing the shutter and stop/start motion of the goniometer.

A disadvantage of shutterless data collection is that it is too late to stop and repeat a frame if it is found to contain overloads. In SADABS 2014/4, one or more low angle **fast scans** are scaled to the other scans, but only used to replace overloaded reflections.

The error model

After deleting outliers, the parameters *k* and *g* in the expression:

$$\sigma^2_{corrected} = [k\sigma_{raw}]^2 + [g\,\langle I\rangle]^2$$

are adjusted so that χ^2 becomes close to unity. Since only two parameters are refined and the resolution does not appear in this equation, values of χ^2 close to unity over the full range of intensity and **especially resolution** are a particularly good sign that the error model is good!

SADABS 2014/4 and later provides a wide range of options for refining *k* and *g*, e.g. *k*=1, *g*=0 (favored by some users of the charge density program XD) or one *k* per scan and one overall *g* (the default).

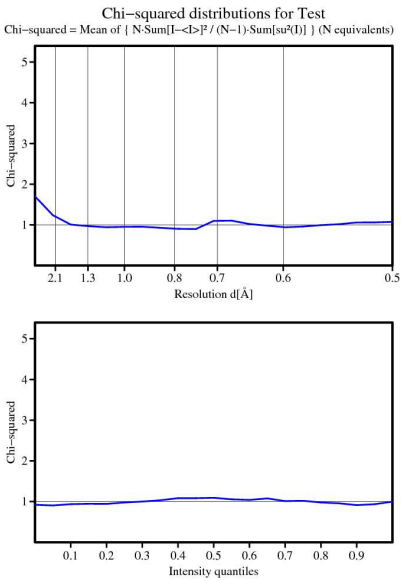
The equation separates ***random*** $[k\sigma_{raw}]^2$ and ***systematic*** $[g\langle I\rangle]^2$ errors.

We might expect random noise to be higher for CMOS, but compensated for by the higher redundancy. **To compare CCD and CMOS we should use the *g*-value and R_{pim} and the final *R*1 for the merged data, not R_{int}!**

Error model for a shutterless dataset

The rise at very low resolution is seen for almost every crystal and detector. It indicates that these reflections are still affected by residual systematic errors.

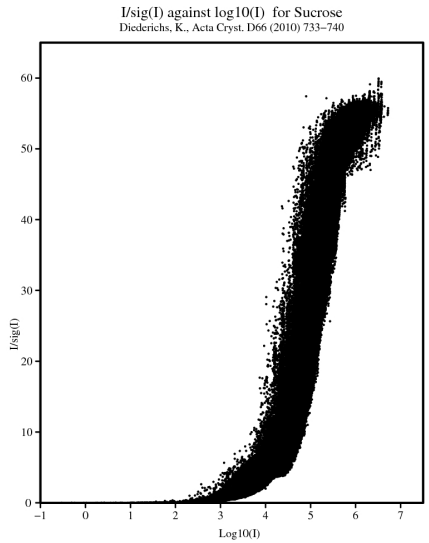
Note that for refinement of the structure, it is still necessary to allow for the systematic errors that affect all equivalents of a given reflection equally (e.g. with WGHT in SHELXL).



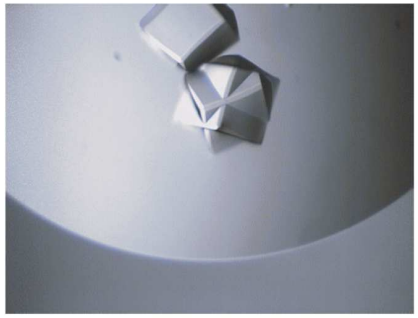
The Diederichs plot

Diederichs (2010) proposed using a plot of $I/\sigma(I)$ against $\log(I)$ for the *unmerged* data as a measure of how much the data are affected by systematic errors. This plot should have a sinusoidal shape, and the limiting maximum $I/\sigma(I)$ indicates how much the data are affected by systematic errors. A value less than 30 is indicative of problems.

The SADABS plot for a Photon-100 IμS shutterless dataset is shown. The limiting maximum value of $I/\sigma(I)$ (jn this case 56) is given by $1/g$.



Non-merohedral twins

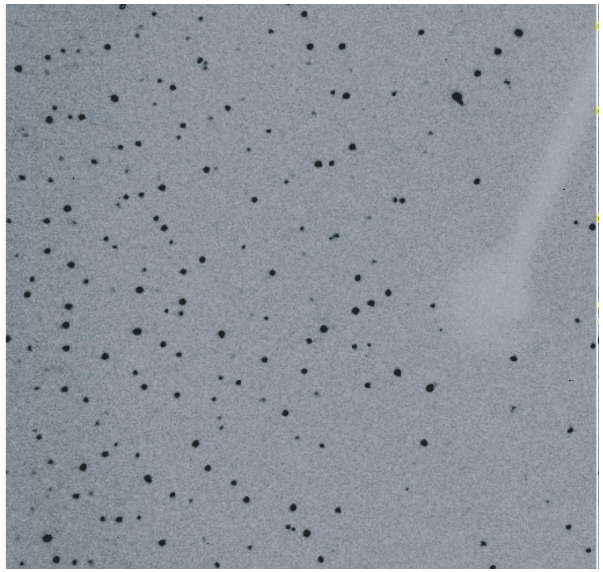


Cubic insulin twin

Glucose isomerase triple crystal

Note that whereas the two components of the cubic insulin twin are interpenetrant and so have approximately the same center, the three components of the GI 'drilling' have well separated centers.

Diffraction pattern of glucose isomerase *drilling*



This diffraction pattern contains single reflections from the 3 different twin components, plus spots with different degrees of overlap of 2 or 3 components.

Indexing non-merohedral twins

The diffraction pattern of a non-merohedral twin will consist of:

A: Separate reflections from the individual twin components

B: Partially overlapping reflections from different components

C: Almost perfectly overlapping reflections from different components

The trick is to find a cell that fits a reasonable fraction of the spots well (groups A and C involving the same component) whilst ignoring misfits.

The reciprocal lattice of the first component is then rotated to fit most of the not yet indexed spots to a second component. If necessary this is repeated for further twin components. Partially resolved reflections (group B) are best left out whilst indexing; once the orientation matrices are known, the complete diffraction pattern can be used.

Twins are indexed in this way using the Bruker program cell_now.

The cell_now algorithm

This is a brute-force algorithm and is intended only for use when all other methods fail. Multiple random real-space starting vectors d with lengths between user-input limits d_{min} and d_{max} are refined by iterative linear least-squares, using all reflections, minimizing the weighted sum of squares of the differences between $t=d_x x+d_y y+d_z z$ and m, the nearest integer to t for each reflection.

The trick is to use weights $w = \sqrt{I}\, p^2 / (p^2+(m-t)^2)$, where p^2 is the *precision factor*, usually 0.01, and I is intensity. This weights down reflections that do not fit well, i.e. [$(m–t)^2$ >> p^2], and so concentrates the refinement on the better-fitting reflections, which after recycling will tend to belong to the same twin domain.

Amazingly, this algorithm can index one component of a multicomponent twin, given enough CPU time.

Finding further twin components

The reflections that do not fit the first twin domain can be used to search for a second domain. Starting from a large number of random starting values of three rotation angles, the cell for the first domain is rotated to give a good fit to as many of these reflections as possible, using the same weighted least-squares criterion as for the first domain.

The rotation required (often 180° about a real or reciprocal axis) provides a description of the type of twinning.

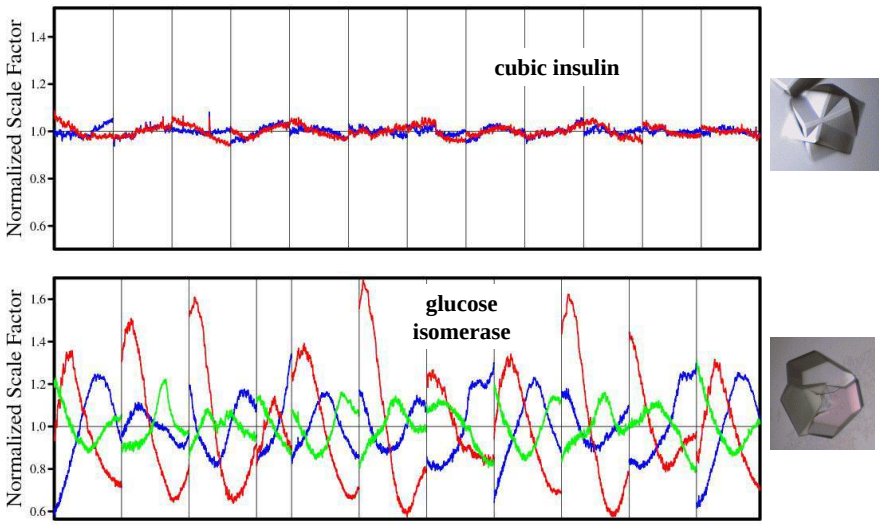
After assigning reflections to the second domain, the reflections that do not fit either domain may be used to search for a third domain etc. in the same way. Cell_now writes the multidomain *.p4p* file needed for integration of such twins with the Bruker program SAINT.

Scaling non-merohedrally twinned data

Scaling and absorption parameters may be determined in the usual way by fitting individual intensities to the mean corrected intensities (averaged over equivalents). However for non-merohedral twins the *equivalents* may be single reflections or groups of overlapping reflections with the same pattern of contributors.

We can either refine against the total intensities only (SHELX HKLF 5 format) or try to extract the best list of unique reflections using all the available information (SHELX HKLF 4 format). The latter has the advantage that structure solution and refinement can be performed using any suitable programs in exactly the same way as for an untwinned crystal.

Scale factors for twin components



(The smallest crystal (shown here in red) was furthest from the center)

Equivalent reflections and groups (HKLF 5 format)

h	*k*	*l*		component	(assuming point group mmm)
1	-2	3	.....	1	} equivalent singles
-1	-2	-3	.....	1	
-1	-2	-3	.....	2	— not equivalent to the above singles
-1	-2	-3	.....	-2	
2	0	-4	.....	1	} equivalent groups
1	2	-3	.....	-2	
-2	0	-4	.....	1	
4	1	1	.....	-2	
1	-2	-3	.....	-3	} not equivalent to the other groups shown here
-1	1	2	.....	1	

In SHELX HKLF 5 format, a group of overlapping reflections is defined by negative component numbers for all but the last reflection in the group. For scaling purposes the patterns of component numbers MUST match.

The HKLF 4 format output file

The measured total intensities I_m of the single or composite reflections are given by a set of equations that can be solved to obtain the relative twin masses k_n and the intensities of the unique reflections I_h:

$$I_m = k_1 I_h + k_2 I_{h'} + \ldots + k_n I_{h''}$$

This algorithm is robust, converges fast and (by using sparse matrix techniques) can process several million reflections in a few seconds. The twin ratios obtained are close to those from the HKLF 5 refinement with SHELXL.

Inconsistent component indexing

The algorithm to generate HKLF 4 format data revealed a subtle and unexpected elephant trap. In the cases where the Laue symmetry is lower than the metric symmetry of the lattice, the components may be indexed inconsistently, even when the second and subsequent orientation matrices were obtained by rotating the first!

For the insulin twin this indeed happened. In the lower cubic Laue group there are two ways of indexing a crystal, related to one another by a rotation about a twofold axis of the metric symmetry that is not a symmetry axis of the Laue group. This is similar to the generation of a *merohedral* twin. The warning sign was a high R_{int} for the least-squares deconvolution.

The scaling is not affected by this inconsistent indexing!!

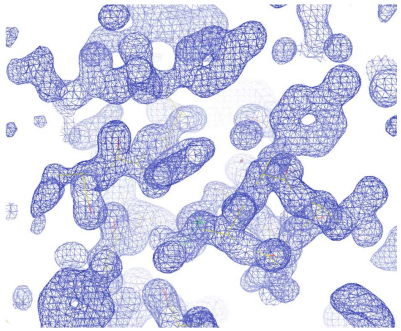
Solving the phase problem

After correcting the inconsistent indexing in the case of insulin and creating (HKLF 4 format) unique reflection files, SAD structure solution with SHELXC/D/E was rather straightforward:

Cubic insulin

SHELXD: CC 50.5, CCweak 29.1 for 6S

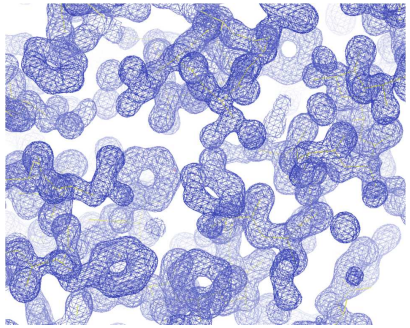
SHELXE: 50 out of 51 residues traced;
MPE 21°, 94% of CA atoms within 0.5Å



Glucose isomerase

SHELXD: CC 31.4, CCweak 20.4 for 2Mn

SHELXE: 383 out of 388 residues traced
MPE 21°, 95% of CA atoms within 0.5Å



Refinement tests

Refinement tests were performed using the cubic insulin and glucose isomerase data. The free *R*-values are similar for both HKLF 4 and HKLF 5 formats, and are in the range that would be expected for isotropic refinements without H-atoms and TLS. Care was taken in setting up the HKLF 5 free *R* set (only single spots, all other single and composite reflections involving the same indices were deleted from the working set).

	Insulin		Glucose isomerase	
HKLF 4:	$R1_{work}$	18.9	$R1_{work}$	18.3
	$R1_{free}$	23.3	$R1_{free}$	22.1
HKLF 5:	$R1_{work}$	14.6	$R1_{work}$	13.0
	$R1_{free}$	22.3	$R1_{free}$	22.8

The R_{work} values for HKLF 5 format are suspiciously low, suggesting a statistical artefact! This might be another reason for preferring HKLF 4 data.

Recommendations and Acknowledgements

Collecting data from non-merohedrally twinned crystals leads to *an increase* in the number of reflections that can be collected in a given time *and improves the completeness of the data* – minor components do not suffer from overloads – so *should become normal practice*! Concentrated solutions and cooling are recommended for growing the twinned crystals.

Particularly for macromolecules, the deconvolution to produce a unique reflection set (HKLF 4 format) has many advantages.

I should like to thank the many Bruker users who provided test data and suggestions for improving SADABS and TWINABS, especially Madhumati Sevvana, Regine Herbst-Irmer, Ina Dix, Michael Ruf, Jürgen Graf, Holger Ott, Jörg Kärcher and Lennart Krause.

The following paper contains extensive details and test results using SADABS, and could be cited when SADABS is used: Krause, Herbst-Irmer, Sheldrick & Stalke, *J. Appl. Cryst.* **48** (2015) 3-10.