SADABS (Version 2.03)
Bruker/Siemens area detector absorption and other corrections

The program SADABS is designed to exploit data redundancy to correct 3D-integrated (thin slice) data from Bruker CCD and MWPC area detectors. SADABS provides useful diagnostics and can correct for errors such as variation in the volume of the crystal irradiated, incident beam inhomogeneity, absorption by the crystal support (e.g. when the goniometer head passes under the collimator during an omega scan on a Bruker Platform goniometer), and crystal decay, as well as improving the esds of the intensities, so it is strongly recommended that it is used to process ALL data, whether or not absorption is significant. These corrections also enable larger crystals to be used for weakly diffracting crystals without introducing systematic errors; for an impressive example, see C.H. Görbitz, *Acta Cryst.* **B55** (1999) 1090-1098.

SADABS reads the *.raw* files generated by the Bruker integration program SAINT; no other files, environment variables etc. are required. SADABS is currently available as a stand-alone executable for the following operating systems: Windows 95, Windows 98, Windows NT, Windows 2000, IRIX 5.3, IRIX 6.5 and Linux/Intel. The program is started from a command line (in an MSDOS window under Windows) by the command:

sadabs

or (in the case of Windows) by double-clicking on a program icon, in which case it will open a temporary dialog window. The executable must be in a directory that is in the current PATH (under Linux this could be /usr/local/bin). Note that under Windows the PATH may be set in a batch file that is called when a MSDOS window is opened.

User interaction with SADABS is by means of question and answer. Almost all questions should be answered by <*Enter*> to accept the defaults suggested by the program unless there is a very good reason not to. The action of the program is divided into three parts:

1. Input of data and modeling of absorption and other systematic errors.

2. Error analysis and derivation of 'correct' standard uncertainties for the corrected intensities.

3. Output of Postscript diagnostic plots and corrected data.

This documentation follows the order of a typical SADABS session, so it is a good idea to open it in a browser window whilst running SADABS until you are familiar with the program.

Before running SADABS, you will need to know the Laue group (either from SMART, or in tricky cases by preliminary examination of a *.raw* file with XPREP without correction using SADABS), and you should have prepared either a single merged .data file **m.raw* or one *.raw* file for each scan using SAINT. All the *.raw* files must be from the same crystal indexed consistently (i.e. the orientation matrices should be similar but not necessarily identical); data from more that one crystal should be processed separately with SADABS and merged using XPREP. If data were inadvertently indexed inconsistently for different scans or in a way that does not correspond to the conventional setting of the Laue group, the 'T' option in the 'D' submenu in XPREP may be used to transform the indices and direction cosines.

Note that SADABS can now read the merged *.raw* file (not possible in previous versions), and that it is strongly recommended to set the 'instrument error factor' (well hidden in a SAINT submenu) to 0 when processing the data with SAINT. A non-zero value can make it impossible for SADABS to find a good error model. SADABS now detects the use of a wrong value for the machine error and offers to try to repair the damage done (useful for reading old *.raw* files for which the frames have been lost).

1. Input of data and modeling of absorption and other systematic errors

On starting SADABS, the first question asks for the name of a file used for a protocol of the SADABS run. It is a good idea to give a name that identifies the crystal; the extension '*.abs*' will be added if there is no dot in the name.

The program then gives a list of Laue groups, the default number '2' is not always correct!

Then the program asks about the treatment of Friedel opposites for the purpose of determining the model used for correcting systematic errors. This does NOT affect the final reflection list, where Friedel and other equivalents are never merged. The answer to this question should usually be 'Y', unless you have a high redundancy and know what you are doing! The answer 'N' halves the data to parameter ratio for the determination of the absorption and other parameters. If you answer 'N', the program allows you to define the point group so that the Friedel opposites can be correctly identified. The default answer to this question is always the chiral point group, i.e. the one that is appropriate for proteins, oligonucleotides etc. Previous versions of SADABS assumed this point group if the Friedel question was answered with 'N'; in the case where the real point group was non-centrosymmetric but non-chiral, this could lead to a (slight) increase in the apparent amount of 'racemic twinning' and to (more important) difficulties in getting an optimal error model, i.e. reflection esds were a little overestimated which could lead to low GooF values in the refinement.

Now SADABS asks for the names of the *.raw* files and reads in the data. The extension *.raw* is assumed if there is no dot in the name typed in. It is a good idea to name these

files so that the character immediately preceding the dot is 1,2,3 ... for the different files so that SADABS can generate correct default names for the second and subsequent files. If a merged reflection file **m.raw* from SAINT is input to SADABS, no further *.raw* files will be requested; the two types of files should not be mixed because it could lead to confusion with the scan numbers. Although SADABS can read a merged and scaled **t.raw* file from SAINT, this is NOT recommended; the scaling and filtering algorithms in SADABS are much more sophisticated than in SAINT, and deleting reflections prematurely screws up the statistics. These files can be large and may take a little time to read in over a network. The current version of SADABS is dimensioned to hold 2 million reflections (1.5 million for the Windows version so that it still runs under Windows 95 in 32 MB). If this is not enough please let me know!

If a scan has been processed with the "machine error factor" in SAINT not set to the recommended value of zero, SADABS will detect this and offer to repair the damage. This offer should always be accepted, otherwise there may be problems with the error model.

After reading in the data, the program checks the direction cosines for consistency. The mean error should not exceed about 0.005. Small non-zero values may be caused by the crystal wobbling during data collection etc., but large values indicate that something is seriously wrong and that the data integration with SAINT should be investigated and possibly repeated. The program also estimates the maximum two-theta and wavelength from the direction cosines and other information in the *.raw* file. These estimates are output only as a rough check on the consistency of the direction cosines etc.; they suffer from rounding errors and so should not be treated as definitive.

Having input the data successfully, the program commences with 'Part 1', the determination of a model for the systematic errors. It is possible to return to this point later to repeat the remaining calculations without having to read the data in again.

The program prints out the total number of reflections and the number that are unique (this depends on the Laue group and the treatment of Friedel opposites), followed by an analysis of redundancy and mean ***I/sigma(I)***. The program then asks for the mean[***I/sigma(I)***] threshold for including a group of equivalents in the subset of reflections used for parameter refinement. If the data were processed with a modern version of SAINT the default value of 3 will usually be good. If the data are exceptionally weak (this will be clear from the statistics that immediately precede the question) a value of 2.5 or even 2.0 could be tried. If the data were processed with an old version of SAINT that tended to underestimate the ***sigma(I)***-values, it might be better to enter '5' here.

The next question requests a high resolution threshold for the data used for parameter refinement. <*Enter*> causes no resolution threshold to be applied. If the resolution threshold was specified too optimistically when running SAINT, it is advisable to input a realistic limit here (e.g. 0.9 when the data were processed to 0.7A but the outer 0.2 A was

mainly noise). Again, this only affects the parameter determination, not the final data processing.

Now the program asks for a value for the parameter ***g*** for the weighting scheme:

$$w = [\ sigma^2(I) + (g<I_c>)^2\ ]^{-1}$$

that is used for parameter refinement. Note that ***g*** is determined later in the error analysis (Part 2), so the default value of 0.02 can be used in a first pass but if the the program subsequently determines a very different value, this stage could be repeated using the new value for ***g***. SAINT uses a similar expression with ***g*** as the "machine error", but with the fundamental flaw that ***g*** is multiplied by the intensity of an individual reflection ***I***, not the corrected mean value $<I_c>$. This will tend to weight up the equivalents with the lowest intensities, although they are the one most likely to suffer from absorption or other errors! A detailed discussion of this subtle statistical pitfall may be found in the HKL2000 manual; it suffices to say here that if the machine error is set to zero in SAINT, SADABS can get the statistics right. The "machine error factor" depends on the crystal quality as well as on the characteristics of the individual detector employed, so it is advisable to refine it rather than use a fixed value.

The next question asks for the restraint esd for consecutive scale factors, expressed as a fraction of their values. This should almost always be in the range 0.001 to 0.005, and the default of 0.002 is a good first try. The best value to use depends on the number of reflections per frame and the redundancy and quality of the data, so it is difficult to guess in advance. The best guide is the appearance of the Postscript plot of the incident beam correction; this should be smooth but still show a slight amount of high-frequency noise. If the value is too small, the correction may be over-restrained and the merging R-values (R_{int}) will be appreciably higher; if the value is too large, the plot will be noisy and the data may be over-fitted, leading to artificially low merging ***R***-values. In general the ***R1*** value at the end of the structure refinement will show a shallow minimum as a function of the value of this restraint; in critical cases this can be used to obtain the optimum value. The original version of SADABS used Savitsky-Golay smoothing instead of restraints, but restraints are more flexible and can handle the case of a very small number of reflections per frame better.

The program then asks for the highest orders to be used for the spherical harmonic absorption correction of the diffracted beam. If absorption is small the default values of 4 and 1 are recommended; for moderate absorption 6 and 3 are suitable and for strong absorption 8 and 5. Theoretically the odd order can be lower if the crystal shape is centrosymmetric. For higher orders the program will be significantly slower.

Now the program asks if special treatment of a thin plate crystal is required. If the answer is 'Y', the indices of the prominent face are requested. The Bruker CCD-microscope attachment and software are well able to determine these indices, but if one has accidentally forgotten to measure them, it is possible to use SADABS to find them by trial and error. The indices are usually small numbers such as 0 0 1, and if unsuitable

values are used the value of ***mu*t*** (***mu*** is the linear absorption coefficient, ***t*** the thickness) simply refines (asymptotically) to zero. The 'minimum glancing angle' enables reflections for which either the incident or the diffracted beam glances the plate to be ignored for the purposes of parameter refinement (a different value may be specified later for correction of all the data). These reflections suffer from large and uncertain absorption corrections; the absorption is also affected by slight bending of the plate and by the beam divergence, so it is better to leave such reflections out - with luck, equivalents that do not suffer from this problem will be present in the dataset. It is always worth trying the correction with and without the special thin plate treatment; in many cases the general spherical harmonic treatment (with orders 8 and 5) is just as good, and converges significantly faster.

The program then enters the parameter refinement. If a plate-like crystal was specified, the default number of cycles is 20, otherwise it is 10. I am still trying to speed up the convergence of the thin plate correction. Each refinement cycle consists of two subcycles. In the first, the scale factors ***S(n)*** (one for each frame, restrained as discussed above) are refined, together with three extra parameters for thin plates. In the second, the diffracted beam absorption ***P(u,v,w)*** is modeled using spherical harmonic functions of the orthogonalized diffracted beam direction cosines ***u***, ***v*** and ***w*** [Blessing, *Acta Cryst*. **A51** (1995) 33-38]:

$$I_c = I_o \, S(n) \, P(u,v,w)$$

The frame number ***n*** is non-integral (the centroid of a reflection will fall between two frames) so a linear interpolation is required:

$$x = n - N \; (N \text{ integer}; \; 0 < x < 1); \; S(n) = (1-x) \, S_N + x \, S_{N+1}$$

Including the restraints (with esds ***e*** after conversion from fractional to absolute values) for adjacent scale factors, the quantity minimized is:

$$Sum \, [\, w \, (\langle I_c \rangle - I_c)^2 \,] + Sum \, [\, e^{-2} \, (S_N - S_{N+1})^{\,2}]$$

where $\langle I_c \rangle$ is the mean corrected intensity of a group of equivalent reflections. The values of R_{int} are printed after each half-cycle (for the reflections used for parameter determination only).

If the thin-plate correction is included, the right hand side of the expression for I_c is divided by $(T + f)$, where ***T*** is the transmission factor calculated from ***mu*t*** for an infinite lamina [Sheldrick & Sheldrick, *Acta Cryst*. **B26** (1970) 1334-1338] and ***f*** is either f_{opp} (incident and diffracted beam on opposite sides of the plate) or f_{same} (both beams on the same side). The refined values of ***mu*t***, f_{opp} and f_{same} are printed each cycle; they correct for edge effects and warping of the plate etc. This formulation should be regarded as tentative, it tends to give a value of ***mu*t*** on the low side because some of the thin plate correction is mopped up by ***S*** and ***P*** (this also gives rise to high correlation factors that slow down the convergence when the thin plate correction is used).

The refinement is performed using the robust/resistant least-squares technique pioneered by Prince (1982) in his book *Mathematical Techniques in Crystallography and Materials Science*. This involves weighting down outliers (determined using the weighting scheme) and is much more stable than arbitrarily eliminating some reflections but not others. This makes it unnecessary to use the FILTER option in SAINT before running SADABS, in fact excessive use of FILTER will screw up the statistics and lead to wrong esds for the intensities. It is far better not to eliminate any reflections in SAINT, i.e. do NOT use FILTER!

After the required number of cycles have been performed, the R_{int} value for the reflections used for parameter refinement is printed. This is usually lower than the R_{int} value for all the data determined by XPREP, especially if a large number of weak data were not used for parameter determination. The program then asks if the parameter refinement should be repeated with new settings or not. This is useful if one is investigating possible thin plate indices, but usually it is better to complete the error model analysis and look at the Postscript plots before repeating the optimization of the absorption (etc.) model.

2. Error analysis and derivation of 'correct' standard uncertainties for the corrected intensities

The next stage (Part 2) is the determination of an error model. First it is necessary to specify which reflections should be considered to be erroneous and left out for the purposes of defining the correct ***sigma(I)***-values and preparing the diagnostic plots, and from the final output file of corrected intensities. A resolution limit (possibly different from that used for parameter determination) should be specified. In the case of a thin-plate correction, a glancing angle test may also be applied, which may also be different from the limit used for parameter determination. Then a ***error***/***sigma(I)*** limit may be applied (default 4.0***sigma***).

The idea is to eliminate reflection measurements suffering from serious systematic errors (e.g. a reflection cut off by the beam-stop or close to a strong reflection from an ice crystal or other impurity), not to throw out a large number of reflections in order to reduce the merging ***R***-values. If the data conform to a normal distribution and the weights are correct, 0.27% will deviate by more than 3***sigma*** and 0.05% by more than 3.5***sigma***; less than 0.01% should deviate more than 4.0***sigma***. Since this assumes purely random errors and some systematic errors inevitably remain, these represent lower limits on the percentages of outliers that should be retained in the data to avoid violating the statistical treatment. In practice a cutoff of about 4.0***sigma*** catches the real errors without upsetting the statistics too much. The program prints out the total number of data and the number of unique data before and after applying these rejection tests, and the user is given the opportunity to experiment with different cutoff values. Note the logic of applying this

rejection threshold after modeling absorption and other errors, rather than before (as would be the case using the FILTER option in SAINT).

After the user has decided which reflections to ignore, the program tries to find the best ***g*** value for the error model:

$$su^2(I_c) = k\,[\,sigma^2(I_c) + (g\langle I_c\rangle)^2\,]$$

where ***k*** is a scaling factor and ***su(I)*** is the corrected standard uncertainty of the corrected intensity I_c, and $sigma(I_c) = sigma(I_o)\ S(n)\ P(u,v,w)$, where $sigma(I_o)$ is the esd of the intensity output by SAINT. The best test of success in establishing a good error model is that the Postscript plots of Chi^2 against intensity and against resolution should be horizontal lines with Chi^2 equal to one.

After establishing the weighting scheme, the program prints a table giving for each scan the following information: R_{int}, minimum and maximum values for the incident and diffracted beam transmission factors, ***k*** (see above), the number of reflections, and the number with intensity greater than ***2su***. At this stage it is again possible to repeat the parameter refinement or the determination of the error model.

It must be emphasized that the R_{int} value, although traditional, is a very poor guide to the quality of the data; it is very easy to reduce it artificially by overfitting the data (e.g. by making the scale factor restraint esd larger, or by using high order spherical harmonics when there is no absorption) or by rejecting too many reflections [Diederichs & Karplus, *Nature Struct. Biol*. **4** (1997) 269-275]. The final ***R1*** value, bond length and angle standard uncertainties, and largest peak and hole in the difference electron density map after refinement are a much better guide (provided the same number of reflections are compared), and Chi^2 values of unity over the full resolution and intensity range are also a good indication that the processed data are free from serious systematic errors.

3. Output of Postscript diagnostic plots and corrected data

The Postscript plots provide essential diagnostics, so it is strongly recommended that they are created with a suitable file name for preservation (the extension *.eps* will be added if there is no dot in the name). They should be given a short title that will appear on each plot. The Postscript file is closed before the next question appears, so it is a good idea to examine the diagrams with GhostView (or GSView) before proceeding.

The first plot gives the variation of scale factors and smoothed R_{int} as a function of scan and frame number. The R_{int} plot may be subject to relatively wild fluctuations if there are very few reflections per frame. The scale factor plot should be almost smooth, just showing a little noise (see the discussion of the scale factor restraint above).

The next plot page shows the variation of R_{int} and R_{sigma} [$= Sum(su) / Sum\langle I_c\rangle$] as a function of resolution. It provides an indication of the resolution cutoff to be applied to

the data, and often shows very clearly the improvement of the data as a result of high redundancy. In the plot of $|E^2-1|$ as a function of resolution, the curve should stay close to the 0.968 line for a centrosymmetric structure or the 0.736 line for a non-centrosymmetric structure, especially for large organic or macromolecular structures. Values that are uniformly lower than expected may indicate twinning, and values that are uniformly higher than expected ay be caused by pseudo-translational symmetry. A systematic drop or rise at high resolution may indicate problems with the SAINT integration (e.g. integrating data that weren't there) Macromolecules may show solvent artifacts at low resolution but should otherwise fall fairly closely to the non-centrosymmetric line. For inorganic structures with heavy atoms on special positions this plot is less reliable.

The third page shows the distribution of Chi^2 as a function of resolution and intensity. The closer these two plots are to a horizontal line at Chi^2 equal to one, the better the fit to the error model and the more reliable the standard uncertainties of the corrected intensities. Small excursions to higher Chi^2 at low resolution are not unusual. Chi^2 is defined as follows:

$$Chi^2 = Mean\ of\ \{\ N\ Sum[I_c-\langle I_c\rangle]^2\ /\ (N-1)\ Sum[su^2(I_c)]\ \}$$

where N equivalents contribute to a given unique reflection (reflections with no equivalents are not included in either summation).

The only optional plots are those that display the distribution of outliers relative to the detector area for each different detector ***2theta*** angle. These can show up bad pixels or e.g. a shadow of the beam-stop (if the active pixel mask was not set correctly in SAINT!), ice rings and other spatial artifacts, but may result in the Postscript file becoming large for large datasets.

The user then has the option of writing the corrected intensities and their standard uncertainties to file in HKLF 4 format for further processing by XPREP. It is possible to apply an additional absorption correction assuming a spherical crystal with given **mu*r**, where mu is the linear absorption coefficient and **r** is the radius of the equivalent sphere; if **mu** is in mm-1 then **r** should be in mm; the dimensions cancel. **r** should be chosen so that it is biased towards the smallest crystal dimension; e.g. if the crystal is a blick with dimensions 0.1x0.2x0.3 mm, then 0.07mm would be a good value for **r**. This correction is included because the theta-dependend part of the absorption cannot be modeled well by comparing equivalent reflections, because these invariably have the same **2-theta** values. However it should not be applied if absorption is absent. The main affect of applying it will be to increase the equivalent isotropic displacement parameters in the resulting refinement.

The data can be corrected for ***lambda/2*** contamination [Kirschbaum, Martin & Pinkerton, *J. Appl. Cryst.* **30** (1997) 514-516]. The default correction factor of 0.0015 is typical for a sealed-tube *MoKa* SMART system, but should be set to zero for a MWPC with *CuKa* (because this detector can successfully discriminate against ***lambda/2*** radiation, in

contrast to a CCD or image plate). Since the correction factor is essentially constant for a given system provided that the mA and kV settings of the generator are not changed, it is recommended that each system is calibrated. The best plan is to find a strongly scattering crystal in a space group with many systematic absences (e.g. ***Pbca*** - a centered lattice can also be used, provided that SAINT is told that it is primitive, but this puts a strain on the integration) and try various values of the ***lambda/2*** factor so that the mean ***I/sigma(I)*** for the systematic absences output by XPREP is about unity (or slightly less). This correction has virtually no affect on the reflections that should not have zero intensity, but it is desirable to apply it so that the systematic absences can be recognized and the space group assigned correctly.

The program allows the user to leave out (or reinstate) particular scans without having to read the data in again, and start again with the determination of the parameters to model systematic errors. The table at the end of Part 2 may indicate that a particular scan is much worse than the others, in which case one can try again leaving it out.

Finally the SADABS run may be terminated with 'Q'.

George Sheldrick gsheldr@shelx.uni-ac.gwdg.de