

---

<b>PREFACE</b>	<b>5</b>
<b>STARTING THE PROGRAM AND SETTING THE DATA SETS</b>	<b>7</b>
<b>Step 1: How to start the HKL-2000 program?</b>	<b>7</b>
How do I start?	7
Can I start from any directory?	7
What are the files that make up the <i>HKL-2000</i> suite, and what do they do?	8
<b>Step 2: Detector and site info</b>	<b>8</b>
What is a site definition?	9
I don't see my detector listed!	9
How do I make a new site if mine isn't listed?	9
What is the Goniostat Type button?	10
How do I verify that the site information is correct?	10
What if I choose the wrong site because I wasn't sure?	11
What is the meaning of the goniostat alignment windows?	11
<b>Step 3: Raw data source</b>	<b>11</b>
How do I use the Directory Tree?	12
How do I set the path?	12
Can't I just type the path into the little boxes?	12
<b>Step 4: Output data files</b>	<b>12</b>
I don't see the "proc" directory. How do I make one?	12
Do I have to have a "proc" directory, or can I put the files anywhere?	13
How much disk space am I going to need for this?	13
<b>Step 5: Setting up Data Files</b>	<b>13</b>
What am I looking at in the dialog box that comes up?	13
Should I click Show All Sets?	14
How to recover from ambiguous name schemes?	14
What if I don't see any frames listed?	15
Why would I want to click Scale Sets Only?	15
Can I click more than one set of data? Why would I want to do this?	15
<b>Step 6: Experimental details</b>	<b>16</b>
What if I selected a set by mistake and want to remove it?	16
What do the Collect, Denzo, and Scale buttons in each window do?	16
What if I don't want to process all of the frames in this set or I want to process some of them separately?	17
The Distance that came up in the Experiment Geometry window is wrong. How do I fix it?	17
Frame Geometry: It looks like I can put in more frames than I actually have. What is this for?	18
<b>AUTO-INDEXING AND CELL REFINEMENT</b>	<b>21</b>
<b>Step 7: Display Your Image</b>	<b>21</b>
Wow, my crystal diffracts great! Is that the correct resolution in the green box?	22
How many peaks do I want to use?	22
What if the Peak Search picks lots of peaks in the ice ring, or picks peaks from a second lattice?	22
Can I get rid of some peaks I know shouldn't be included?	22

Can I add some peaks that I think the auto-index should use?	22
How can I flip through my frames at this stage to see if they all look ok?	22
Can I index on another frame? Can I index on peaks collected from several contiguous frames?	23
Adjusting the display: How can I make the diffraction image look “better”? You know, darker peaks, lower background.	23
<b>Step 8: Auto-indexing</b>	<b>24</b>
How do I define the beamstop shadow and why should I do this?	24
What does the blue highlight in the Sets Pending box mean?	25
Why would I want to put in any other resolution limits?	26
At what point should I be concerned with the Refinement Options?	26
Indexing keeps flashing. How long is this supposed to take?	26
What if there aren't any green circles? What if they don't line up on real spots?	26
<b>Step 9: Lattice parameter refinement</b>	<b>27</b>
Oops! The preds don't line up with the spots anymore!	28
I think the direct Beam Position is wrong. How do I change it?	31
What should I set at the 3D window?	32
What's the procedure for choosing the correct lattice?	32
What if no lattice other than primitive triclinic comes up green?	34
What are the Positional and Partiality $\chi^2$ values and what should they be? What are the numbers in parentheses?	34
What are all those the Refinement Information values for, and what do they mean?	36
Someone told me to gradually increase the resolution and alternate cycles of fitting a few parameters and fixing things. I'm patient, and I want to be careful. Why shouldn't I do this?	36
<b>Step 10: New lattice refinement</b>	<b>37</b>
Problems – predicted reflections don't line up with spots, things get worse, etc.	37
How do I start over? How do I get back to P1?	37
<b>Step 11: Resolution Limits, Fitting Radius and Spot Size</b>	<b>37</b>
What is the Profile Fitting Radius and how do I determine what it should be?	37
How do I set the spot size? Is it all that important? Why do larger spots make my $\chi^2$ 's worse?	38
What is the Elongation Limit?	39
Resolution Limits: should I set this here or not worry about it?	39
What are all these parameters Rotx, Roty, Rotz, X-beam, Y-beam, crossfire, mosaicity?	39
What is the Reference Zone? How is this useful?	41
<b>Step 12: 3D Window and Mosaicity</b>	<b>41</b>
What is the 3D Window?	42
Why do I have to keep the 3D Window at 1 for image plate data?	42
What are the limitations for 3D Window usage?	42
Why wouldn't I just want to set the 3D Window to encompass all of my frames?	42
What is Mosaicity? What does the histogram under Check Mosaicity mean?	42
How do I set the Mosaicity? Should I try to increase the mosaicity until all of the spots on the image are covered with preds?	43
How do I increase the Mosaicity?	43
<b>STRATEGY AND SIMULATION</b>	<b>47</b>
<b>Step 13: Strategy</b>	<b>47</b>
Why would I want to use a strategy and a simulation?	47
How do I start a strategy session?	47

How can I use I/Sigma and Crystal Dose buttons?	49
<b>Step 14: Simulation</b>	<b>51</b>
What is the simulation for, and why would I want to use it?	51
How do I use the simulation?	51
Can I change anything other than a frame width?	52
What do I do once I have an acceptable simulation?	52
What is a Combine button for?	52
<b>INTEGRATION AND SCALING THE DATA</b>	<b>55</b>
<b>Step 15: Integration</b>	<b>55</b>
The Integrate button is red. Is this a problem?	55
The Crystal Slippage and Special integration buttons.	55
What are these plots supposed to look like and how do I know if everything is ok?	56
What is an acceptable range for $\chi^2$ ?	56
“Houston, we’ve got problems”. The $\chi^2$ values are getting larger!	57
Why do the bars in the mosaicity histogram get thicker or thinner?	57
How long should this take? How do I know if the program is stuck?	57
Waiting for image frames.	57
Once the integration is complete, should I update the site?	58
How do I stop and start over?	58
<b>Step 16: Scaling</b>	<b>58</b>
Why do I want to do postrefinement? Which one should I choose?	58
Scaling Options: What are Scale Restrain, B Restrain, Fix B, Ignore Overloads, etc.?	59
Should I change the resolution now?	62
What are the three <i>Scalepack</i> output Files? Can I rename them? Where do they go?	62
Someone told me my statistics will be better if I decrease the number of zones. Should I do this?	62
What is the error model and should I adjust it?	63
How many cycles should I refine for and how do I know how to set this?	63
What’s the rejection file and why am I writing it?	63
<b>Step 17: Output plots and Error Scale Factor</b>	<b>63</b>
What is the Error Scale Factor and which way should I adjust it?	64
What are reasonable values for the error scale factor and what if I need really high ones?	64
What should I be looking for in the charts that show up after scaling?	64
How many reflections should I be rejecting?	66
What’s the difference between a reflection being marked for rejection and actually rejected?	66
How do I treat data that contains an anomalous signal?	67
Space Group Identification: How do I do this?	68
How to start scaling from .x files without re-integrating diffraction images?	70
Reindexing.	71
Statistics and <i>Scalepack</i> .	71
The Resolution of Diffraction Data.	73
Understanding the <i>Scalepack</i> log file <code>scale.log</code> .	74
<b>APPENDIX - INSTALLATION NOTES</b>	<b>76</b>
<b>TABLE OF FIGURES</b>	<b>81</b>



# Preface

The *HKL-2000* suite is a program package based on extended versions of the programs in the HKL suite: *Denzo*, *XdisplayF* and *Scalepack*. In addition to the *HKL* suite, *HKL-2000* consists of several more subprograms. The most important new elements are strategy and simulation, 3-D processing, mosaicity refinement during processing, variable spot size, easy change of the space group, and report generation, etc. The graphical command center coordinates all steps of experiment and enables performing an interactive experiment with on-line data analysis, where results are automatically updated when new data are collected. The experimenter can adjust the data collection strategy to assign the desired data completeness. A proper data collection strategy can help avoid later problems in data reduction and structure solution.

The software is driven by the graphical user interface (GUI), which greatly improves convenience of working with various data sets. The emphasis of the *HKL-2000* manual is on how to run *Denzo*, *XdisplayF* and *Scalepack* from the graphical user interface. The figures in this manual are for the 0.98.703x version of the *HKL-2000* program. Since the program continues to undergo constant revision and development, the version of the *HKL-2000* package you are using may not be the same one described here.

To find more about the features of and algorithms used in *Denzo* and *Scalepack* please read the article:

Z. Otwinowski and. W. Minor "**Denzo & Scalepack**"  
*International Tables for Crystallography Volume F: Macromolecular Crystallography*, p. 226-235, 2001,  
M. G. Rossmann, Kluwer Academic Publishers (Dordrecht/Boston/London).

Question about the program described here should be directed to Zbyszek Otwinowski or Wladek Minor at the addresses below, with CC: to HKL Research:

HKL WWW homepage:	<a href="http://www.hkl-xray.com">http://www.hkl-xray.com</a>
HKL e-mail :	<a href="mailto:hkl@hkl-xray.com">hkl@hkl-xray.com</a>
Zbyszek Otwinowski:	<a href="mailto:zbyszek@work.swmed.edu">zbyszek@work.swmed.edu</a>
Wladek Minor:	<a href="mailto:wladek@iwonka.med.virginia.edu">wladek@iwonka.med.virginia.edu</a>

To reference the *HKL-2000* package for publication, please use the following citation:

Z. Otwinowski and W. Minor, "**Processing of X-ray Diffraction Data Collected in Oscillation Mode**",  
*Methods in Enzymology*, Volume **276**: Macromolecular Crystallography, part A, p.307-326, 1997, C. W. Carter, Jr. &  
R. M. Sweet, Eds., Academic Press (New York).

**Denzo®**, **HKL®** and **HKL-2000®** are registered trademarks of HKL Research, Inc.



# Starting the Program and Setting the Data Sets

## Step 1: How to start the HKL-2000 program?

### How do I start?

The *HKL-2000* suite must be installed and configured with a license file before it can be run. Installation and configuration of the suite is described in the **Appendix**. If the *HKL-2000* suite has been installed in the default location (`/usr/local/bin`) you can typically just type `HKL2000`. (or `/usr/local/bin/HKL2000` if the directory is not in your executable path).

Alternatively, the suite can be installed in any directory. However, if the suite is not installed in its default location, an additional environmental variable `$HKLPATH` has to be set to point to that directory. For instance if all programs are installed in the directory `/home/user/prog2000`, it is necessary to execute the commands below for C-type (`csh` or `tcsh`) shells:

```
$ setenv HKLPATH /home/user/prog2000
$ alias HKL2000 $HKLPATH/HKL2000
```

or for Bourne-type (`sh`, `bash`, `dash`, etc.) shells:

```
$ export HKLPATH=/home/user/prog2000
$ alias HKL2000=$HKLPATH/HKL2000
```

Then type

```
$ HKL2000
```

### Can I start from any directory?

No. The user running the program must have permission to write files in your working directory, because the program writes the files `reject`, `peaks.file`, `hklpredictions`, `den.in`, `den.out`, `scl.in`, `xdisp.in`, and `xdisp.out`, to the working directory. What are these files?

`reject` is the list of rejected reflections created in scaling

`peaks.file` is the file that contains peaks selected by PeakSearch routine

`hklpredictions` is a binary file that contains a listing of predicted reflection positions

`den.in` is a listing of the commands used to run the indexing

`den.out` is a modified log file of the indexing

`scl.in` is the *Scalepack* input file

`xdisp.in` and `xdisp.out` are input and output files for the display program.

**What are the files that make up the *HKL-2000* suite, and what do they do?**

HKL2000 - The main program with the graphical user interface  
access\_prod - The program that prepares file necessary for license key generation. It has to be executed on each computer that will run the *HKL-2000* suite.  
anomalous - Anomalous signal analysis  
cdir\_t - Directory manipulation tools  
denzo\_3d - Auto-indexing, refinement and integration; 3-D processing  
hklfilename - File name manipulation tools  
hklsysinfo - Program that gathers information about computer resources  
readheader - File header reader  
scalepack - Scaling data comprising up to 2000 frames and 2 million reflections  
scalepack16m - Scaling data comprising up to 2000 frames and 16 million reflections  
scalepack8m - Scaling data comprising up to 2000 frames and 8 million reflections  
scalepacklvirus - Scaling data comprising up to 3000 frames and 50 million reflections  
scalepackmanyframes - Scaling data comprising up to 4000 frames and 2 million reflections  
scalepackmf6500 - Scaling data comprising up to 6500 frames and 3 million reflections  
scalepackribo - Scaling data comprising up to 2000 frames and 32 million reflections  
scalepackvirus - Scaling data comprising up to 2000 frames and 32 million reflections  
strat - Strategy and simulation  
suser - Internal communication program  
xdark - Dark current calculation  
xdisp - Image display and measurement

**Step 2: Detector and site info**

The first dialog to appear after invoking the program asks you to choose a site definition (detector and other site information) from the menu (Figure 1). You must select a site definition before continuing to the program.

### What is a site definition?

A site definition contains information about the X-ray detector/camera, goniostat and X-ray source. The definition includes the detector name and type, the axes of rotation and misalignment parameters of the goniostat, the X-ray beam properties, and properties of the diffraction image files collected, especially the expected suffix for these files (e.g. \*.img, \*.osc, etc.). It also has the default or updated (current) X and Y coordinates of the direct beam position on the detector, which are crucial pieces of information. This information is stored in /usr/local/hkclint or in \$HKLDIR as a series of subdirectories, one for each site (Figure 1). Each site definition directory contains an individual “site file” called `def.site` where most of the settings are stored. These subdirectories may also contain other files such as detector calibration data (in the form of files called `def.cal` and `denzo_nonunf`) which are necessary for processing uncalibrated data. For synchrotron data processing, these files should be obtained from the synchrotron beam-line.



Figure 1. The list of site definition directories

### I don't see my detector listed!

Then you have to create a new site definition. You cannot process data if the program does not know what sort of experimental setup you used.

### How do I make a new site if mine isn't listed?

To make a new site, click on the **NEW SITE** button. It then opens a new window with a field to give the new site name. This will create a new subdirectory in /usr/local/hkclint or \$HKLINT with the name you give (note that the user running the program will have to have permission to write to this directory). It's best to use a descriptive name, like `our_beatup_raxis` or `ALS-BL8.2.2-Apr2006-ADSC-Q315` that does not contain spaces (Figure 2). Now click on the **DETECTOR TYPE** button, and find your detector. Some detector manufacturers have different models and these are listed under each general detector heading. If your detector is not listed at all (unlikely, but theoretically possible), you will have to contact HKL Research (or your beamline coordinator) for the appropriate site file. Once you've selected the correct detector, hit **OK**. The new site definition directory will be created with the appropriate files, including a `def.site` file. You will now get the main **HKL-2000** window.

- ➔ The newly created `def.site` file will contain default values for the **X Beam** and **Y Beam Positions** and are almost always incorrect. You can change these by clicking on **Site Configuration**, and make the changes here. While not recommended, it is also possible to edit the appropriate line in the `def.site` file in the newly created subdirectory and then restart the program.

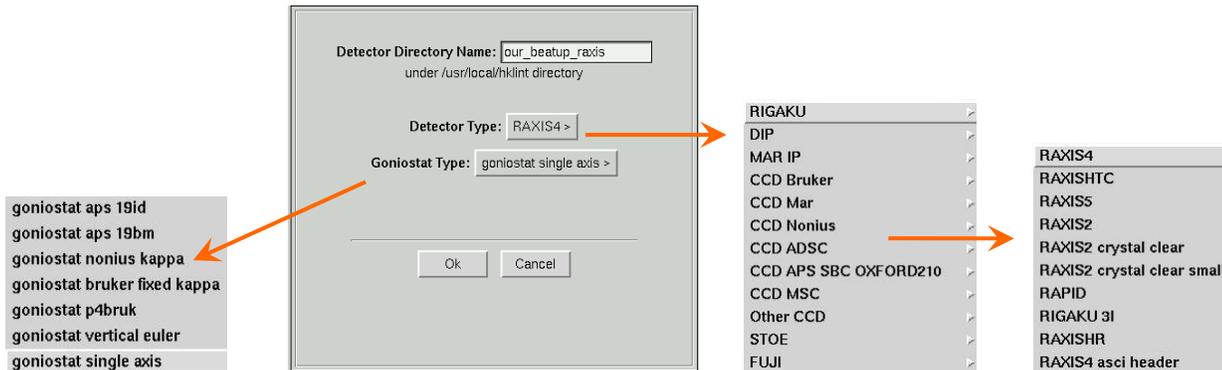


Figure 2. The New Site window

### What is the Goniostat Type button?

Choose the appropriate goniostat type from the predefined list (Figure 2). Presently the following types of goniostats are available:

- single axis – most common
- vertical euler – goniostat used with Rigaku Rapid, Saturn and Raxis HTC detectors
- nonius kappa – goniostat used by Nonius KappaCCD and Kappa2000 systems
- Bruker fixed kappa – used with Smart 1000 and Smart 6000 Bruker systems
- p4bruk – P4 goniostat used with Smart 6000
- aps 19id – minikappa goniostat used on id19 (undulator) beamline at the APS
- aps 19bm – minikappa goniostat used on bm19 (bending magnet) beamline at the APS

Note that not all goniostat types are predefined.

### How do I verify that the site information is correct?

Click on *Site Configuration* on the top menu bar and this will bring up a window with the site parameters (Figure 3). You can verify that the detector is the right one, and you can specify things like the **X Beam** and **Y Beam Positions** and the **Wavelength** and so on. There are a couple of things to note about this window:

First, you can load in data from an existing site file by clicking on the **LOAD SITE INFO** button. This brings up a file navigation window where you can choose a site file (`def.site`).

Second, the anode type (and thus the wavelength) for rotating anode X-ray sources is specified in this window, while the wavelength for synchrotron X-rays is specified by data set in the main *HKL-2000* window (and is hopefully, but not guaranteed, to be correctly specified in the image file header – check to make sure!).

Third, clicking on **SAVE SITE INFO** means that you are permanently editing or updating the file called `def.site` (in `/usr/local/hkclint/sitewhatever`). For this you will need a password (which is provided to the *Contact Person*). So this is not a typical user option. Also, the user running the program must have permission to write to the directory (`/usr/local/hkclint/sitewhatever`, but see the next comment, below).

Fourth, however, you can use your updated site parameters by clicking **CLOSE** button and these will remain in force until you quit the program, after which time they will revert to the original values.

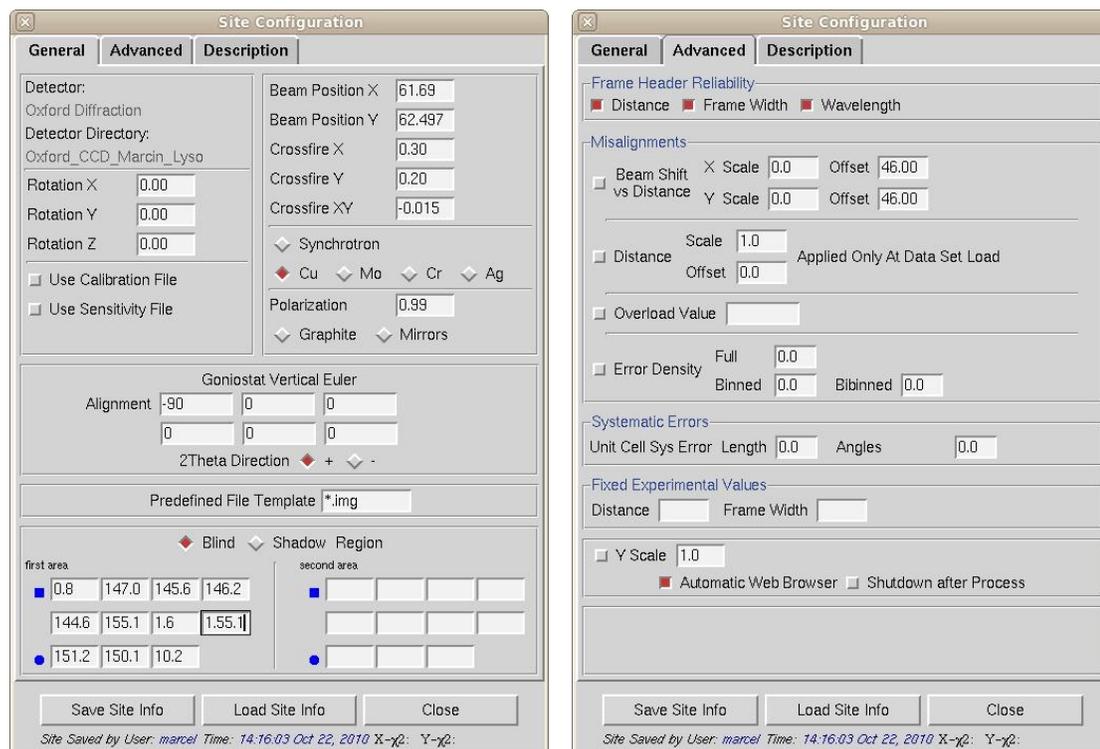


Figure 3. The Site Configuration window

### What if I choose the wrong site because I wasn't sure?

Well, most likely the indexing and integration will fail because file formats will be unreadable, etc. If you are lucky and the detector is correct and the goniostat description is compatible you may be able to get somewhere.

### What is the meaning of the goniostat alignment windows?

The windows describe the six mis-setting angles of a particular goniostat. The choice of the single axis goniostat allows for only first two mis-setting angles to be entered:

- 1) Angle around the beam
- 2) Angle describing the non-perpendicularity of the angle between beam and omega/spindle axis

The other goniostat types require all six mis-setting angles to be defined. The four remaining angles are:

- 3) Omega zero
- 4) Deviation between nominal and actual kappa (chi) – omega angle
- 5) Kappa zero
- 6) Deviation between nominal and actual kappa (chi) – phi angle

A change of the spindle axis direction requires a corresponding change of the first mis-setting angle by 180 degrees. This is equivalent to the Denzo command **film rotation 180**.

## Step 3: Raw data source

Set up your input data files using the [Directory Tree](#) and the **New Raw Data Dir** field.

### How do I use the Directory Tree?

Double clicking on a file folder displays the subdirectory contents of that directory. It will not display the files in that directory, so don't be puzzled when you can't see your data files. To go up one or more directories, double click on a parent directory. To make a new subdirectory, single click on the folder under which you want to create a new directory, click on **CREATE DIRECTORY** and a dialog box will open with the path displayed. Append your subdirectory name onto the end of this path and click **CREATE**. If your data is in a directory configured to auto mount but is not currently mounted, click **AMD ACCESS** (auto mount directory), type the name of the directory and click **ACCESS** to mount (Figure 4).

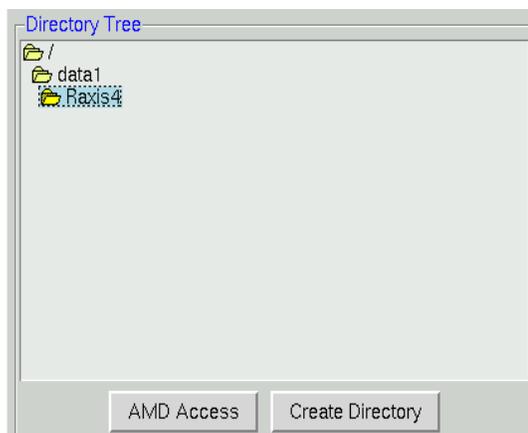


Figure 4. The directory tree

### How do I set the path?

Easy! Single click the subdirectory that contains your data (which will highlight it), then click the “»” button to the right under **New Raw Data Dir**. The full path will be written in the field next to “»” button (Figure 5).



Figure 5. The Raw Data directory

### Can't I just type the path into the little boxes?

You sure can, but the clicking method above is a lot faster, much less prone to typing errors, and is also, for some strange reason, kind of fun.

## Step 4: Output data files

Set up your output files using the **Directory Tree** and **New Output Data Dir**. The suggested convention is to have a subdirectory called `proc` in the directory that holds your frames, for the files created during data processing.

### I don't see the "proc" directory. How do I make one?

To make a `proc` subdirectory, single click on the folder under which you want to create this new directory (Figure 4). Usually this will be within the directory that contains your diffraction image files (“frames”). For example, if your frames are in `/data/gewirth/X25/complex`, then the `proc` would be a subdirectory within `complex`, i.e. `/data/gewirth/X25/complex/proc`. Click on **CREATE DIRECTORY** and a dialog box will open with the path

displayed. Append your subdirectory name onto the end of this path. When the data set is already created, you can change data output directory by right mouse button click only (Figure 6).



Figure 6. The Output Data directory

### Do I have to have a "proc" directory, or can I put the files anywhere?

You can put them in your parent directory, or you can put them in the same directory as the frames. The nice thing about having `proc` subdirectories is that if you reprocess the data multiple times (say, in different space groups) you can have a different `proc` (e.g. `proc1`, `proc2`, `proc_p222`, `proc_p21`, etc.) subdirectory for each attempt and not worry about overwriting data or having to make new file name patterns.

### How much disk space am I going to need for this?

Each indexed raw data file yields a new file, called an `*.x` file, with the extracted spot information from the frame in it. The size of the `*.x` files is therefore linearly proportional to the number of reflections on the image. As you might imagine, the `*.x` files for a lysozyme data set would be quite a bit smaller than those for a ribosomal subunit data set when the crystal diffracts well. Generally, the `*.x` file size depends on the unit cell dimensions and resolution. You always should have at least a few gigabytes of free disk space.

## Step 5: Setting up Data Files

After selecting the raw data directory, select the frames in that directory to be processed using the **LOAD DATA SETS** button (Figure 7) and single clicking on one or more sets of frames.

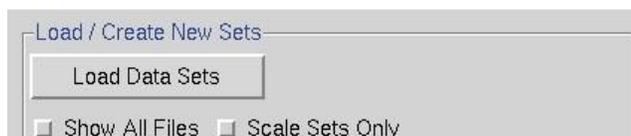


Figure 7. The Load/Create New Sets window

- ➔ When collecting diffraction data, avoid creating frames with names ending in string of 5 or more numbers because the set parser cannot distinguish a contiguous data set among the frames. For example, if the root of your data file names is `nc401`, and then you collected frames 001 through 140, you might expect files called `nc401001.img` through `nc401140.img`. Unfortunately, it is very difficult for a computer to parse this type of file name and distinguish between: 1-140, 1001-1140, or 401,001-401,140 frames. You may keep the root `nc401`, but add other non-numeric characters in between (for example: `nc401_set1_0001.img`, `nc401_set 2_0001.img`...)

### What am I looking at in the dialog box that comes up?

If you have not clicked on one of the checkboxes below the **LOAD DATA SETS** button, you should just see what the program considers to be the sets of data frames in the **New Raw Data Dir**, with # signs standing for digits of the frame numbers (Figure 8). Single click on the set of frames you want to process - it will turn blue - then hit **OK**.



Figure 8. The Load Data Sets window

When the set has been added, a blue box will appear in the main window summarizing the properties of the set of frames you have added (for example, as seen in Figure 13). To edit the properties of the data set you just created, push the **EDIT SET(S)** button (see Figure 12) you should see the **Raw Data Dir**, **Output Data Dir** and **File Template** boxes filled in, along with the starting number and the number of frames (Figure 9). If you are processing data concurrently with data collection, and the data collection process is not controlled by *HKL-2000*, you can set manually the expected **Number of Frames** box and the integration process will wait for incoming frames until all expected frames are collected. Assuming that the program properly parsed the frame headers, you should not need to change any of the other parameters of the data set.

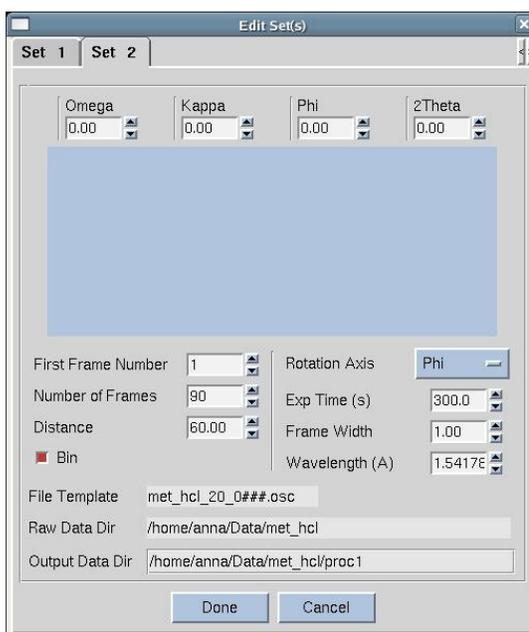


Figure 9. The Edit Set(s) window

**Should I select *Show All Sets*?**

Usually it is not necessary because the normal file selection procedure works fine if the site is set up correctly and the proper frame file name suffix has been specified. However, this option is still useful because you can navigate through the directory tree after activating the **SHOW ALL SETS** button (see Figure 7).

**How can I recover from ambiguous name schemes?**

If you've already collected data with ambiguous file names, you don't need to manually change all the file names! Consider the the `nc401001.img` to `nc401140.img` example described above. Load in the data set as described above, even if the parser does not extract the range of frame numbers properly. Go to the *Edit Set(s)* dialog, change the name in **File Template** from `nc#####.img` to `nc401###.img` and move the mouse. The program will parse the range of frames and a dialog box will ask for approval. After selecting **OK**, the proper values of **First Frame**

**Number** and **Number of Frames** should appear and **File Template** will be automatically adjusted according to the new scheme.

### What if I don't see any frames listed?

This means that either you have no data frames in the **New Raw Data Dir**, or else they are not named and numbered in such a way that the program can identify them as data frames. For example, your files may be of the form `*.mccd`, but the **Site Configuration** is expecting files with the suffix `*.img`. The best way is to go to **Site Configuration** and edit **Predefined File Template**. Alternatively, you can try clicking on the **SHOW ALL SETS** button before clicking on **LOAD DATA SETS**, and this will open a new window: **Select Data Sets – Show All Files** with all of the files in the **New Raw Data Directory** (Figure 10). In the window you can manually define a file template as well as a set start and end.

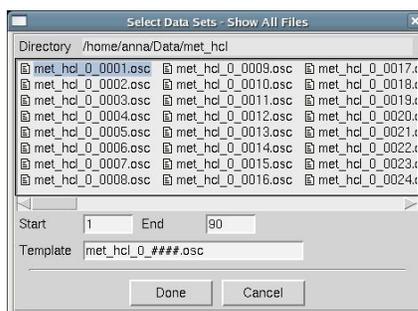


Figure 10. All files are listed after activating the Show All Sets button

### Why would I want to select Scale Sets Only?

This will display only the `*.x` files that in the **New Raw Data** or **New Output Data** directory paths (Figure 11). You might want to do this if you were just interested in scaling the integrated data, rather than reindexing and reintegrating the raw frames (say, if the raw frames were no longer on your disk drive).



Figure 11. Files listed after activating the Scale Sets Only button

### Can I select more than one set of data at once? Why would I want to do this?

Yes, you can select multiple sets of data within the **SELECT** data window. Clicking once selects, clicking again deselects. You do not have to hold down any additional keys to select more than one – just click on each set you want to add.

As to why you might want to do this; this is extremely useful when you have collected multiple data sets from a single crystal and want to process them all together. Examples of this include:

1. inverse-beam experiments,
2. data sets which differ by the  $2\theta$  offset of the detector,
3. low-pass and high-pass data sets that differ in the crystal-to-detector distance, or
4. multi-wavelength experiments.

In all cases, if the goniostat is aligned (and this alignment must be in the site configuration [def.site file]), the crystal has not been moved relative to the goniostat, and the header information of the image files is reasonably accurate (so

that detector distance and position information is present), then you should be able to process all of the frames together from a single auto-indexing.

## Step 6: Experimental details

As you select each data set, the program will print its parameters at the bottom of the window under **Experiment Geometry Frame Geometry**, Exposure Time and Options; make sure they are correct. In addition, having the correct *x beam* and *y beam* values in the Site Configuration is critical. Other parameters of the experimental geometry are not as important and mistakes can be found and corrected during the initial refinement. For instance, if the **Oscillation Range** is wrong, you will see that increasing the size of the **3D Window** does not increase the number of refined spots and instead results in an increase in the  $\chi^2$ . (The **3D Window** is the number of frames that the auto-indexing, refinement and integration process simultaneously. See page 42 below for more details.) In that case one should move to a **3D Window** size of 1, process the data and observe the **Crystal Rotation** diagnostic window. The most common cause is the opposite rotation of the spindle axis that can be corrected by correction of the goniostat definition in the Site Configuration.

### What if I selected a set by mistake and want to remove it?

That's easy! In the main window, click the **SELECT** button of the set you want to remove. Then click the **REMOVE SET** button on the right middle side of the main window (Figure 12). The program will ask if you really want to do this, and if so, the set will be removed.



Figure 12. The Set Control window

### What do the Collect, Integrate, and Scale checkboxes in each window do?

This tells the program what you want to do with the data in this set. If you want to collect the data (**COLLECT**), index and integrate the frames (**INTEGRATE**), and scale the data (**SCALE**), you would click all of these checkboxes (Figure 13). The default is for all the available options to be selected when the set panel is displayed. (*HKL-2000* control of data collection is still under development and currently supports only data collected on SBC beamline at APS as well as with a few selected detectors. If your configuration of *HKL-2000* does not support direct data collection, the **COLLECT** checkbox may be missing.) This is most useful if you want to jump to the end and scale frames that have already been indexed and integrated. In that case, just unselect the **INTEGRATE** checkbox, but leave the **SCALE** checkbox selected, and the program will know that you are passing that data set directly to *Scalepack*.



Figure 13. The Select window options

### What if I don't want to process all of the frames in this set or I want to process some of them separately?

Notice that under the panel called **Set Controls** there is an option called **SPLIT SET**. This splits your data set into two or more groups, which can then be integrated consecutively, independently, or selectively. Why would you want to do this? One reason could be if you had a bad or blank frame somewhere in the series of frames. Another possibility is that you changed the wavelength in the middle of the data collection. In general, sets are split whenever there is an interruption in a contiguous series of data collection. To split a set, select the desired set and then click on the **SPLIT SET** button in the **Set Controls** panel. A new box will appear which has two sliders (Figure 14). Move the top slider to define the end of the first set, and then move the bottom slider to define the start of the second set. If you want to skip one or more frames, just make sure that they are in the "gap" between the two sets. When done, click on the **SPLIT** button and a new set will appear (Figure 15). Make sure that the **Distance** and **Oscillation Start** values are correct for each set (they should be read from the header of the data files, but you never know).

- ➔ In principle you can split a data set as many times as you need to, however, keep in mind that 3D integration cannot work across the boundaries between split sets.

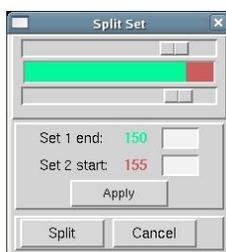


Figure 14. The Split Set window

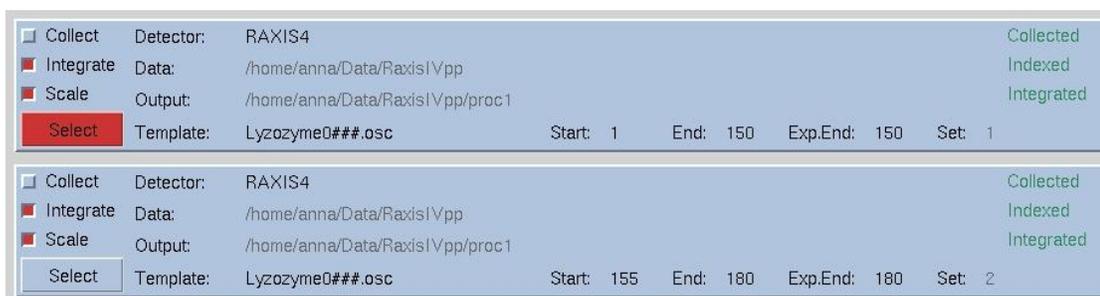


Figure 15. Data set divided into two sets with removal of four images in the middle

### The Distance that came up in the Experiment Geometry window is wrong. How do I fix it?

At the bottom of the page you see the boxes labeled **Experiment Geometry**, **Frame Geometry**, **Exposure Time and Options** (see Figure 16). For processing, the most important parameters are the **Distance** and the **Frame Width**. In most cases, these values, as well as the values mentioned below, should be automatically extracted from the frame headers. However, they can be changed in the table under the **Summary** tab (see Figure 25) or/and the **Edit Set(s)** window (see Figure 9). The **Distance** may be incorrect because the header of many data files does not contain this information, or the information in the header itself is incorrect. **Frame Width** is more likely to be right. You should also make sure that the **Wavelength** is correct, too. For home data collection setups the wavelength is usually defined in the Site Configuration, but obviously for synchrotron data the wavelength will vary for each experiment. You can fill in the correct values for the **Exposure Time** and **Oscillation Start** and so on, although these won't affect the results from processing a single contiguous data set. However, if you are processing more than one data set with a shared set of parameters (like crystal orientation, etc.) **Oscillation Start** will be important. When in doubt, put assumed values and try manual refinement. If you see that **Distance**, **Width**, or **Wavelength** are consistently wrong contact the beamline or detector vendor and ask for them to correct their image file headers.

### Frame Geometry: It looks like I can put in more frames than I actually have. What is this for?

This is particularly useful if you are processing data as it is being collected and you want the program to integrate the frames in the same session, but you don't want to wait to get started. In the **Edit Set(s)** window (see [Figure 9](#)), enter the correct **Number of Frames** you are planning to process in this batch. You will notice that as you enter this number the **Number of Frames** value will increase in the **Frame Geometry** window, as well as the **Exp. End** in **SELECT** panel at the top of the page. Notice as well that the **End** value in these two fields does not change; this is the number of frames that the program knows is on hand at the moment. For example, suppose you are collecting 100 frames of data, and you have 20 collected so far. You would enter 100 in the **Number of Frames** box before you started the integration of the 20 frames collected so far. You could then start the integration job and the first 20 frames will be integrated. When the last frame on hand is processed, the program will wait – giving the message “Waiting for Frames” – and when the new frames are written to disk the program will continue to process them. This is especially helpful at places like synchrotrons, where the frame integration rate is a substantial fraction of the data collection rate and you want to maximize your efficiency.

- Another way you can use the **Number of Frames** value is to stop the integration before all the frames are processed. This is helpful if you have a lot of frames that have lots of spots and so take a long time to integrate. Perhaps you just want to process a wedge of this data and get some idea of its quality. For example, say you have 300 frames but just want to look at the first 30. Enter 30 in the **Number of Frames** box and start processing. The first 30 frames will be integrated and then the program will stop. This is easier than splitting the data set and more civilized and nicer than hitting the **ABORT INTEGRATION** button.

Experiment Geometry		Frame Geometry		Exposure Time			Options	
Distance	127.38	Oscillation Range	90.0	Time / Frame	181.0	Sec	Wavelength	0.71074 A
Oscillation Start	0.00	Number of Frames	90		3.017	Min		17444.4 eV
Omega	0.00	Frame Width	1.00	Time / Degree	181.0	Sec	<input type="checkbox"/>	Narrow Frame
Chi	20.00	Rotation Axis	Omega		3.017	Min		
Phi	0.00			Time / Set	04:31:30			
2 Theta	0.00							

Figure 16. The experimental details panel





# Auto-indexing and Cell Refinement

## Step 7: Display Your Image

Select the data set to display, then click on the **DISPLAY** button to display the first frame (Figure 17).



Figure 17. The Display image button

Click **PEAK SEAR[CH]** and **MORE PEAKS** (if necessary) to select peaks for auto-indexing. Then click **OK** (Figure 18). Occasionally, you may notice that number of peaks that are picked by the program is low. In that case adjust the expected size of the spot by clicking the **SIZE PEAK UP** or **SIZE PEAK DOWN** buttons.

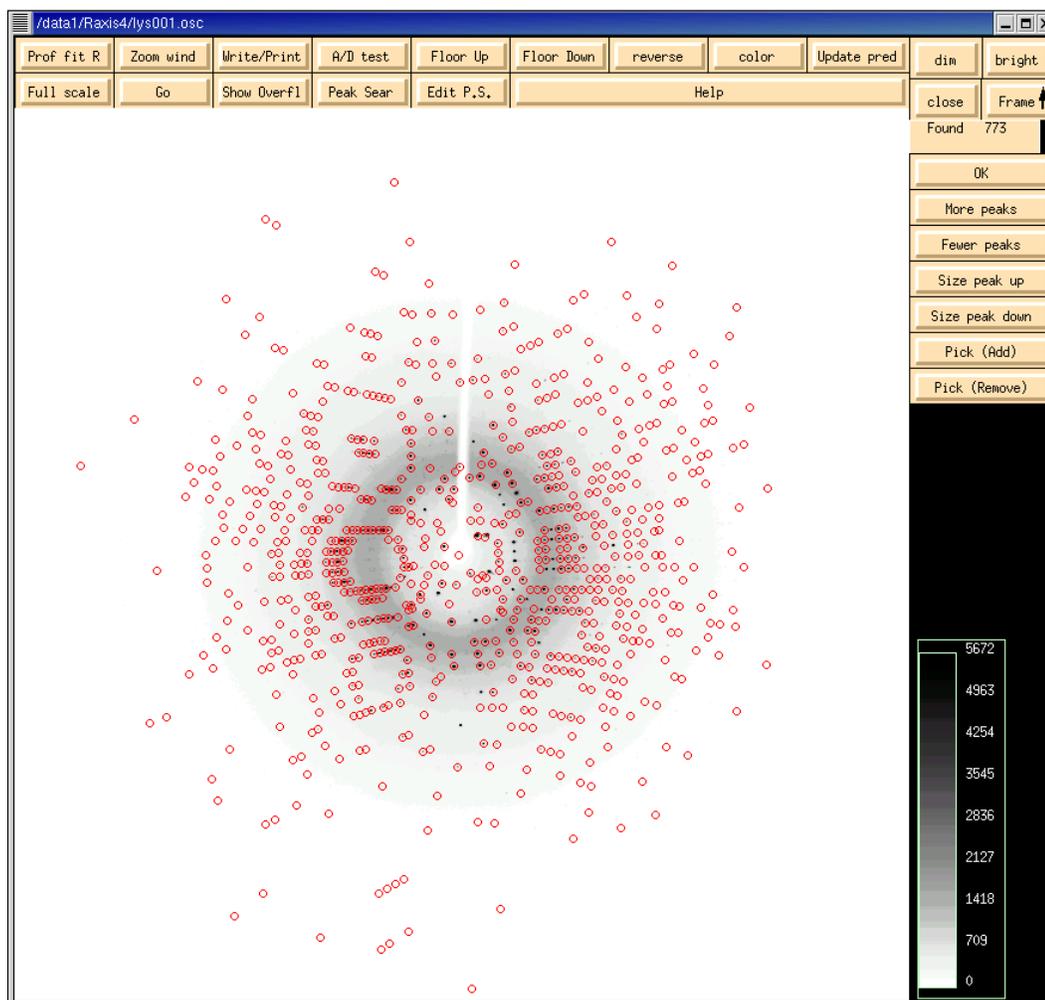


Figure 18. The Display window with activated "Peak Search" option

### Wow, my crystal diffracts great! Is that the correct resolution in the green box?

Sorry, I wish it was, but most likely it is not. The display program does not know the crystal-to-film distance yet, and so the **Distance** and **Wavelength** are set to defaults, which are likely to be different from your actual conditions. Once you have indexed the first frame, the display program knows the correct distance and the numbers in the green box of the image display are accurate.

### How many peaks do I want to use?

The main object is to pick peaks for auto-indexing. If the program picked around 100 good peaks (clearly belonging to the major lattice, not coming from ice rings, etc.) then you are OK. If lots of the peaks are very weak, you can increase the chances of getting some good ones by clicking on **MORE PEAKS**. Alternatively, you can click on the **FRAME** button to display the next frame, and the peak search will be repeated on this frame, using the same criteria used for the first frame (e.g. **MORE PEAKS**, **FEWER PEAKS**, etc.). These will be added to the peak search from the first frame. This is a way to index poorly diffracting crystals. If you have a double lattice, the best way is to increase weak level to, let's say, 20, to select only the stronger peaks – this way you may index only one lattice.

### What if the Peak Search picks lots of peaks in the ice ring, or picks peaks from a second lattice?

Since *Denzo* assesses the quality of each peak before using it in auto-indexing, most spurious peaks, such as those coming from ice rings, can be ignored by the user. However, if you have a second distinct lattice contaminating the diffraction pattern, you will have to make some effort to assure that the selected peaks come mainly from the primary lattice. The first way is to hope that *Denzo* can automatically distinguish between the primary lattice and any secondary lattices. If the secondary lattice is weak or at a significantly different orientation from the primary lattice, this is not unreasonable. Alternatively, you can try indexing on a different frame, where the secondary lattice may not be so pronounced. However, if this does not work, you can assist the program by editing the peak search list in the *Image Display* window.

### Can I get rid of some peaks I know shouldn't be included?

Yes, you can assist the program by editing the peak search peaks in the *Image Display* window. If you click on the **PICK (REMOVE)** button, the cursor will change to a cross, and this can be used to select peaks from the peak search for removal (Figure 19). The peaks to be removed will turn green. Click **OK** to remove these peaks from the peak list. Also, you may change the resolution to 4 Å (for example) to remove weaker, higher resolution peaks from the auto-indexing.

### Can I add some peaks that I think the auto-index should use?

Yes. Simply click on the **PICK (ADD)** button (Figure 19); the cursor should turn into a cross. Use this cross to select the peaks you want to add by clicking on them. A little green circle will be left over your selections. Note that this does not work in the *Zoom* window, even though clicking on peaks leaves little green circles in here as well – but they are not added to the peak list. Click **OK** to add these peaks to the peak list.

- ➡ The **PICK (ADD)** button should be used in extreme cases when you have very few (less than 10) very weak spots per frame.



Figure 19. The Pick (Add) and Pick (Remove) buttons from the Display window

### How can I flip through my frames at this stage to see if they all look ok?

In the *Image Display* window, there is a button labeled **FRAME**, with an up or down arrow in it. Clicking on the middle mouse button causes the displayed frame to increment in the direction of the arrow (up means a higher numbered frame, and *vice versa*). To change the direction of the arrow, click on the **FRAME** button with the right or left mouse button, respectively (Figure 20). The **FRAME** button will only increment one frame at a time. If you want to quickly go

to a frame that is far away from the current one, you'll have to close the *Image Display* window and enter the frame number in the box by **Frame Number** in the **Main** page (Figure 17).



Figure 20. The Frame button from the Display window

### Can I index on another frame? Can I index on peaks collected from several contiguous frames?

You can select which frame to display by changing the **Frame Number** box. Most of the time it's best to start with the first frame of the series, because when *Denzo* starts auto-indexing it will assume you are working with the first frame of a series unless you split the data set. If you're not working with the first frame and you haven't split the data set, then the program will not be able to find the peaks you selected for auto-indexing. This is because while you may have picked peaks from frame 10, *Denzo* is looking for these peaks on frame 1. You'll get an error message that says, "Error – Peaks none." So the rule is that the peak search has to be done on the first frame(s) of a set, and if you want to index starting from the middle you have to split the set so that one series starts there in the middle.

- ➔ One nice feature of *HKL-2000* is that you can index on peaks collected from several contiguous frames. Why is this helpful? Well, suppose you have a very poor crystal that doesn't diffract well and only gives a few low resolution spots. Alternatively, you might have a very sparse reciprocal lattice that only results in a few spots per image. Ordinarily this would be hard or impossible to index. However, if you could combine peaks from several consecutive frames, this might yield enough good reflections so that auto-indexing could work.

To accumulate peaks from consecutive images for auto-indexing, first display the first image of the series and run your normal peak search. If you need to, you can click on **MORE PEAKS** or **FEWER PEAKS**, but don't hit the **OK** button (Figure 21). Instead, use the middle mouse button to click on the **FRAME** button. The next frame will be displayed, and the peak search will be run on the next image. Notice that if you clicked **MORE PEAKS** or **FEWER PEAKS** the search on the second frame will behave as if you have already clicked this. You can select the number of frames and the new peaks are appended to the `peaks.file` file in the directory from which you launched *HKL-2000*. Note that for auto-indexing to make use of the peaks coming from the second and subsequent frames that the **3D Window** (found under the **Index** tab) must be set equal to or larger than the number of frames on which the peak search was performed. In other words, if you did the peak search on 4 consecutive frames but your **3D Window** is set to 3 frames, then the peak search peaks from only the first 3 frames will be used in the auto-indexing. Similarly, if your **3D Window** is set to 5 but you only did a peak search on the first 2 frames, then only the first two frames will be used for the initial indexing.



Figure 21. The More Peaks and Fewer Peaks buttons

### Adjusting the display: How can I make the diffraction image look "better"? You know, darker peaks, lower background.

The displayed image can be made darker by clicking on the **BRIGHT** button in the *Image Display* window. Conversely, the image can be made lighter by clicking on the **DIM** button. Small increments in either are made by clicking with the left mouse button. Larger jumps are made by clicking with the middle or right mouse buttons. To change the displayed dynamic range, use the **FLOOR UP** or **FLOOR DOWN** buttons. **FLOOR UP** narrows the dynamic range, while **FLOOR DOWN** increases it. If you have fairly weak diffraction you can enhance the contrast (i.e. make it look "better" than it actually is) by alternately clicking the **BRIGHT** and **FLOOR UP** buttons (Figure 22).



Figure 22. The buttons for adjusting the image

## Step 8: Auto-indexing

To start auto-indexing go back to the *HKL-2000* window, click the **Index** tab. Set the **Resolution Limits** (the default is to the edge; changing it later is also OK), then auto-index by clicking the **INDEX** button. The program will try to fit the diffraction pattern in each of the 14 Bravais lattices (described in more detail below) and open the *Bravais Lattice* window to allow you to select a different lattice if necessary. Keep the primitive triclinic (space group P1) lattice for now and close the *Bravais Lattice* window. Check to make sure the green circles identified by the auto-indexing line up with real spots by examining the *Image Display* window (click the **DISPLAY** button if it's not already open) and toggling on and off the **UPDATE PRED** button. Click on **SET BLIND REGION** and define the beamstop shadow.

### How do I define the beamstop shadow and why should I do this?

It is important to define the area of the beamstop shadow. This is because unless the beamstop shadow region is defined explicitly, *Denzo* may fail under certain circumstances of high background to reject potential reflections, which are in fact obscured by the beamstop. These few inaccurately measured “reflections” will severely degrade the quality of your anomalous signal. To define the beamstop shadow, keep the *Image Display* window open and return to the **Index** page. Click on **Set Blind Region** and then click the top large green box in the new window that appears (Figure 23). Now go to the *Image Display* window and click with the left mouse button four times to define the four corners of the rectangle that corresponds to the beamstop arm (obviously, this does not apply to mylar film beamstops that leave no arm shadow). After the fourth click, the blocked out region will appear red. If you don't like the results you can repeat the process and after selecting your four new vertices the original rectangle will be replaced by your new choice. Similarly, to define the circular area of the beamstop shadow, click on the big green circle, then move to your *Image Display* window and click once on the center of the circular area. After this you can move the mouse to enlarge or reduce the blind region. You don't have to hold down a mouse button to do this. When you are satisfied, click a second time with the left mouse button and a filled red circle will appear. Going back to the **Blind Region** dialog box, click **APPLY BOTH**. If you don't like your selection, click **CLEAR CURRENT**. If you want to begin integrating, click **APPLY CURRENT**. If you want to store these blind regions permanently, click **UPDATE SITE** and later this information will be forwarded to your **Site Configuration**, so you will be able to use that for all subsequent processing. (Please note that this may or may not be appropriate given the geometry of your experimental setup—for example, the size of the beamstop shadow may change as the detector distance changes.) Click **CLOSE** when you are done.

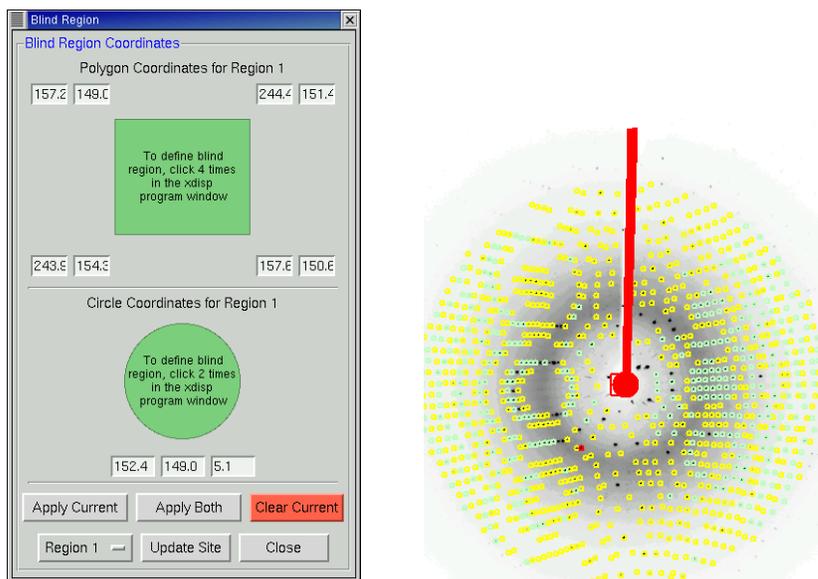


Figure 23. The Blind Region definition window and image with assigned blind region

What does the blue highlight in the Sets Pending box mean?

All of the data sets that you are integrating will appear in the Pending Sets box in the upper left corner of the Index page (which pending sets are present are controlled by selecting the Integrate checkbox for each data set on the Data tab). You can click on a set and it will be highlighted with blue background (Figure 24). This is the set that is currently being refined. Additionally, when one set is highlighted, you can change the parameters of the integration for that set, while leaving the parameters for the other set(s) unchanged. For example, set 1 may be set for processing to 3 Å resolution, while you may want to process set 2 to 2.5 Å. Similarly, perhaps the Spot Size is 0.3 for the first set, but 0.25 for the second set. Or perhaps the 3D Window is 5 frames for the first set but only 3 for the second. The Pending Sets box allows you to toggle between the pending data sets in preparation for integrating them in the same run. You can also accomplish the same thing by going to the Summary tab page and editing the parameters for each set (Figure 25).

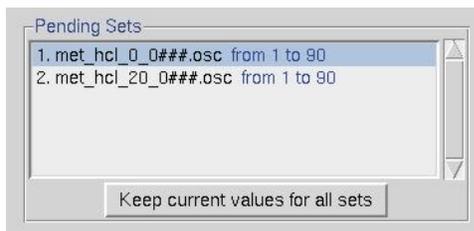


Figure 24. The Pending Sets box

Set	Raw Data Template	Start	End	Osc Start	Osc Width	Osc Range	Phi	Distance	Res Min	Res Max	3D Window	Profile Fitting Radius	Wavelength
	Default parameters			0.00	1.00		0.00	100.00	50.00	edge	1	30.0	
1	lys###.osc	1	20	198.00	1.00	20.0	198.00	75.14	50.00	edge	1	30.0	1.5418
2	lys###.osc	25	60	222.0	1.00	36.0	198.00	75.14	50.00	edge	1	30.0	1.5418

Figure 25. The summary information about processed sets

### Why would I want to put in any other resolution limits?

Before auto-indexing, the only parameter you will want to consider changing from the default values is the **Resolution**. If your data is good to the edge of the plate, then leave the default settings, as they are (Figure 26). In other cases, however, for example if you have ice rings in your data, or the data are weak beyond a certain resolution, it may be advantageous to change the **Max** resolution value that doesn't extend all the way to the edge or corner of the detector. For example, if you have an ice ring, try setting the **Max** value to 4 Å. If you have a secondary lattice which is more pronounced at higher resolutions, try indexing at low resolution.

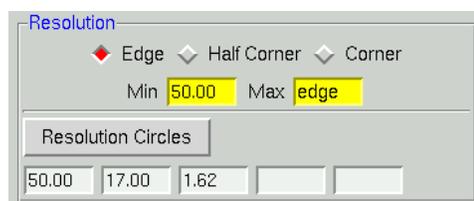


Figure 26. The Resolution limits box

### At what point should I be concerned with the Refinement Options?

Not yet. The initial parameters are those of **FIT BASIC** and these are appropriate for auto-indexing and the first refinement step. In general, thanks to eigenvalue filtering, you can integrate your data with **FIT ALL** and any parameters that are too highly correlated will be restrained by the program (Figure 27). The days of selective manual fitting and fixing are over. The program almost always does a better job.

- Manual intervention may still be appropriate in cases of multiple lattices or other crystal pathologies.

### Indexing keeps flashing. How long is this supposed to take?

For auto-indexing, the computer has to read in all the frames of the **3D Window** for which peak searches have been performed and then do the calculations. Generally auto-indexing takes less than a minute, but is roughly proportional to the number of reflections to be considered for auto-indexing.

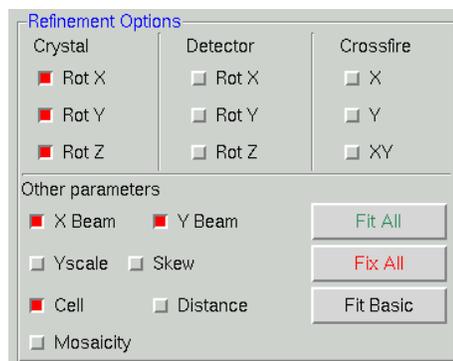


Figure 27. The Refinement Options box

### What if there aren't any green circles? What if they don't line up on real spots?

The green circles are a subset of the initial peaks selected in the peak search (Figure 28). This is the subset that has been used in the auto-indexing. How are these green circles selected? First, the image frame is read into memory by the auto-indexing program and the peak search coordinates are mapped onto this image. The auto-indexing program then decides whether in fact these coordinates represent usable peaks or not. The criteria for usability include falling within the resolution limits specified, high enough signal to noise, sufficiently low background (which helps to

eliminate spots in ice rings and overlaps), and identification of a lune. If there aren't any green circles after auto-indexing, one or more of the following problems may be the cause:

1. The **Peak Search** was done on a different frame from that being used to auto-index. In this case you should have already gotten the error message "Error – Peaks none." The remedy is to make sure these are the same. The rightmost column of the file `peaks.file`, generated during peak searching, tells which frame the peak search peaks were derived from.
2. None of the spots are strong enough to qualify for use in the auto-indexing routine. The remedy is to take a longer exposure, pick more peaks on the frame or on several frames, or lower the **Refinement Sigma Cutoff**, which is normally set at  $5\sigma$ . Lowering the sigma cutoff is a dangerous option, however, and is not recommended because weak reflections are not high quality reflections, and therefore may be noise.

Rejection criteria for background. *Denzo* assesses the quality of the background of each spot against criteria for slope and uniformity, and if certain fraction of the pixels in the background fails the test, then the whole spot is rejected. This is controlled by the keyword **reject fraction**. This default value for this rejection is 0.75 i.e. the 75% of the background pixels must be acceptable. If you have an ice ring, you can increase **reject fraction** until most of the spots in the ice ring are not longer picked up. Useful increments are 1-2% at a time. Do this while you are looking at the display, just as you would do for spot shape and mosaicity.

3. There weren't enough spots to describe a lune (sparse lattices, low mosaicity, and low resolution). The remedy is to pick more peaks or include more frames in the peak search.

A further discussion of auto-indexing failures can be found further down.

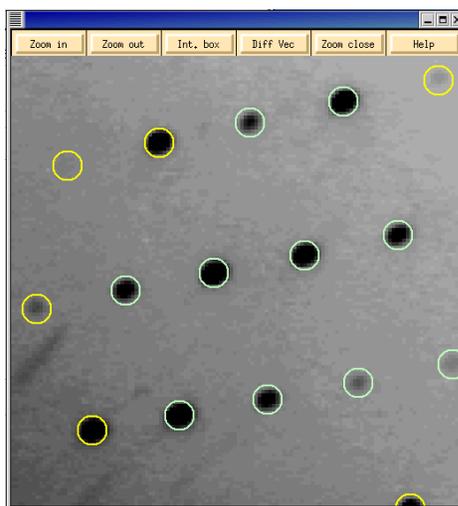


Figure 28. The Zoom window with image after auto-indexing

## Step 9: Lattice parameter refinement

Click the **REFINE** button to refine the lattice parameters. Check the image to make sure the predicted reflections sit on real spots. Make sure the direct beam position is correct and looks reasonable. Click the **FIT ALL** button and **REFINE** again. Click the **BRAVAIS LATTICE** button, select the appropriate lattice for your crystal, and then click the **APPLY & CLOSE** bar.

Oops! The preds don't line up with the spots anymore!

Occasionally the auto-indexing routine does not give clear-cut results. There are four classes of data-reduction problems:

1. **Utter failure**
2. **Misprediction**
3. **Kind-of-failed**
4. **The predicted positions agree perfectly with the peaks, but data scale poorly**

**1. Utter failure** means that there is no reasonable predicted lattice, the crystal unit cell parameters are unreasonable, and the predicted diffraction pattern does not correspond to the observed image. In most cases failure can be ascribed to problems in three areas:

- experimental conditions,
- the parameters of auto-indexing, and/or
- the peak search (rarely).

■ **Experimental conditions:** Problems with defining experimental conditions account for the vast majority of the problems with auto-indexing single crystal data. If any of the following parameters is widely off, your auto-indexing is also off (listed in order of importance):

1. The **X BEAM** and **Y BEAM** positions. These can be measured from the *Image Display* using the cursor and reading off the position from the little green box. A direct beam image or a good educated guess (say, based on the intersection of Laue streaks, the symmetry of the diffraction pattern, etc.) is often necessary. At synchrotrons, it is very helpful to have a **beam stop** which has been thinned sufficiently at its center so that a tiny amount of the direct beam passes through and makes a spot on each frame whose position can be measured. **Bottom line:** beam position error has to be smaller than half of the distance between diffraction spots. It is relatively easy to find a beam position for a test crystal with short unit cells and apply it to the **Site Configuration**.

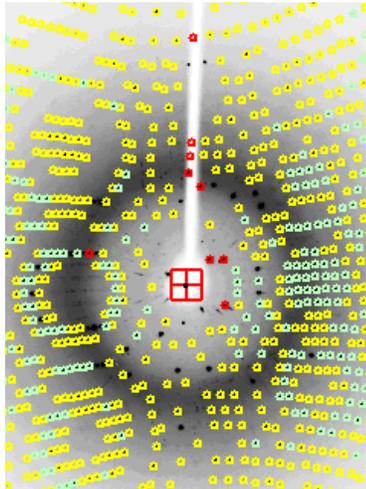


Figure 29. The beam stop thinned in the center marks the direct beam position

2. Specification of the **detector format** (believe it, it's happened).
3. The crystal-to-detector **DISTANCE**. The distance should be known accurately (within about 1%) and can be measured with a tape measure as a last resort.

4. The detector `rotx` and `roty` values. Misset detector `rotx` and `roty` values may be due to  $2\theta$  angles other than zero.
5. **Oscillation range.** Very rarely auto-indexing will fail because too large or too small an oscillation range was collected. In the case of **large** oscillations, the lunes may severely overlap and the auto-indexing program will have a problem with the non-uniqueness of the mapping from detector to reciprocal space. In the case of **small** oscillations or stills collected on non-mosaic crystals with a very parallel beam too few reflections may be present to describe a lune. As a rule of thumb the following types of oscillation ranges should be used:
  - **viruses:** 0.25° oscillations,
  - **proteins:** 1.0° oscillations,
  - **small molecules:** 2.5° - 5.0° oscillations,

assuming well diffracting samples. There is a fairly wide margin on either side, by at least a factor of 2.

Knowing the detector parameters only makes sense. If you want to deduce some unknowns (crystal lattice, cell and orientation parameters), you need some 'knowns' too.

■ **Auto-indexing parameters:** Before auto-indexing *Denzo* edits the peak list produced by the Peak Search algorithm of the *XdisplayF* program. Problems with the auto-indexing parameters account for a few of the remaining problems with auto-indexing. These may have to be adjusted in the case of non-single crystals, very large or small unit cells, lots of noise peaks, or ice rings. These parameters (with the *Denzo* keywords in parentheses) are the longest vector (**LONGEST VECTOR**), resolution limits (**RESOLUTION LIMITS**), refinement sigma cutoff (**WEAK LEVEL**), integration box size (**BOX**) and integration box spot size (**SPOT**). All of these parameters except for the longest vector can be changed through the *HKL-2000* window (the longest vector may be changed via a macro on the *Macros* tab).

1. The **LONGEST VECTOR** may need to be adjusted if the peaks are sparse due to a small unit cell or low mosaicity. In this case, it should be set to equal the longest vector of the reduced primitive real space unit cell. If you did not understand that last sentence, simply use the longest unit cell axis (which is the same in primitive space groups, anyway). Better to err on the side of larger rather than smaller values.
2. **Resolution limits** can be adjusted, if you have:
  - An **ice ring**, which interferes with the indexing. You can set the resolution limits to, say, 100 - 4.1 Å to eliminate these reflections.
  - A **small satellite crystal**, which contributes some reflections at low resolution but not at high resolution. You can set the resolution limits to, say, 8 - 2 Å, thereby excluding the strongest contributors from the contaminating lattice.
  - A **cracked crystal**, where the two lattices are the same space group only 2-3 degrees apart. You can often adjust the resolution limits, as in the case of an ice ring, so that you index at low resolution, where the two lattices are approximately identical, and then use manual refinement to zero in on both lattices. The idea here is to index each lattice separately and to combine the data later in *Scalepack*.
  - A **twinned crystal**, where the two lattices have a non-random orientation relative to one another. The program may fail spectacularly by finding a "**superlattice**," which indexes both lattices simultaneously, but cannot be used to integrate the data because the superlattice does not usually fit both lattices exactly. Then it's time to call an expert or grow new crystals.
3. **Resolution sigma cutoff.** If you have a number of spurious peaks the image distributed randomly over the image, you can adjust the value of the resolution sigma cutoff. This will affect the number of peaks accepted by the auto-indexing routine. By increasing the resolution sigma cutoff, you can exclude weak, spurious peaks, then run auto-indexing, and examine the superimposition of the auto-indexing predictions (green circles) on the image. If most of them cover real reflections and not noise, then you are succeeding.

4. **Box** and **spot** sizes may have to be adjusted if too many spots are being removed by the overlap or background rejection criteria. Auto-indexing does not use spots which *Denzo* determines are overlaps.

■ **Peak search.** Finally, you can try changing the number of peaks you select in the peak picking routine. However, since *Denzo* edits the list of peaks anyway, this is unlikely to yield much improvement if the suggestions above do not work. In that case, *do you have a better crystal?*

**2. Misprediction** results in location of reflection masks not corresponding to the positions of the Bragg peaks. It is visibly obvious and is disastrous. It may be due to:

- misindexing,
- forcing the wrong space group symmetry, or
- a serious detector misconfiguration or malfunction.

**Misindexing.** Auto-indexing is based on the assumption that the diffraction spots are correctly mapped from detector coordinates to diffraction (reciprocal) space. The origin of the diffraction space is defined by the position of the direct beam on the detector. A substantial error in the beam position can shift the indexing of the diffraction pattern by an integer vector. Such misindexing can be totally self-consistent until the stage when symmetry related reflections are compared. For any assumed (starting) value of the beam position, the origin of the diffraction space during indexing will be shifted to the nearest grid point of the best primitive lattice. An initial error in the direct beam position by 0.48 times the distance between reflections will lead to correct indexing, while an error of 0.52 times the same distance will cause a misindexing of the diffraction pattern by one index. **Misindexing by one is never corrected by subsequent refinement** of the crystal and detector parameters. Misindexing often produces poor agreement between the predicted and the observed positions of the reflections, but for some crystal orientations, the agreement between the predicted and the observed positions can be equally good for both correctly indexed and misindexed cases. This property of the diffraction geometry creates a potential trap for the unwary crystallographer.

Sometimes the **crystal asymmetric unit** may have molecules related by an approximate translation by a fraction (typically one half) of a unit cell edge or diagonal. The resulting diffraction pattern will have odd index reflections much weaker than even index reflections. Auto-indexing may find one of the two possible solutions, the choice depending on whether odd reflections are weak enough to be assumed (within experimental error) systematically absent or not. This depends on what is the fraction of odd reflections in the peak search. If there are only a few odd reflections, then most of the peak search result may be explained with a smaller real-space unit cell. To prevent auto-indexing from finding such a smaller cell, one should enhance the fraction of odd reflections used in auto-indexing by changing the peak-search criteria or by using only high-resolution reflections in auto-indexing. If one still cannot index odd reflections, then one should consider ignoring them all together. In such a case, one can solve the structure in a smaller unit cell and the resulting structural error will not be very significant if odd reflections are much weaker than even ones.

**3. Kind-of-failed** means that a lattice other than primitive triclinic was predicted, but it had a relatively high distortion index, say 1-1.5% (the lattice distortion index is described in more detail under “What’s the procedure for choosing the correct Bravais lattice?” below). This may not be a “failure” but simply an indication that further refinement of parameters is needed, because the initial detector parameters in the Site Configuration were very far off. You should just run a few cycles of refinement in P1 and change **Bravais Lattice**. Unfortunately, misindexing may give similar results. A quick check of the superposition of the predicted spot positions on the reflections should tell you if the refinement is likely to succeed. Note that refinement may converge to low  $\chi^2$  values even with misindexed crystals, but eventually the problem will surface and you will have to index correctly. This is now less of a problem with the **3D Window** option because the requirement to integrate with consistent values over a range of frames is much more stringent a test. You are much less likely to refine misindexed data to low  $\chi^2$  values using a multiple-frame **3D Window**.

**4. Predicted positions agree perfectly with the peaks** and no detector or diffraction artifacts are visible, **but data scale poorly** and produce many outliers. This problem may be either a simple mistake in data processing (such as using the wrong file format or wrong goniostat definitions), or a non-uniform exposure during crystal oscillation.

The non-uniform exposure may be caused by spindle motor backlash, shutter malfunction (opening too early or too late), ionization chamber electronics failure (if used), decay or variation of the X-ray beam intensity (if an ionization chamber is not used), variable speed of the spindle motor, etc. Non-uniform exposure is best diagnosed by exclusion of other problems that may affect data quality. Graphical feedback provides confidence that the problem cannot be at the indexing/integration stage.

Large variations in absorption of X-rays by the crystal will make data scale poorly and will produce visible variation of the background; however it will not affect positional agreement. The variation in the absorption can be avoided easily by properly mounting the crystal. The correction for absorption is a whole field in itself.

Errors in detector orientation will produce distorted mapping from detector to reciprocal space. A wrong crystal-to-detector distance will increase (or decrease) the apparent reciprocal space separation between Bragg peaks. This error will not be uniform in three directions; in the first approximation, along the beam direction the error will be proportional to the square root of the error in the distance, and in the other two directions the error will have linear dependence. In most cases a significantly wrong crystal-to-detector distance (say an error of 10%) will not make the auto-indexing step fail immediately; however, the calculated unit cell will be quite wrong. The length of unit cell along the beam direction will be 5% shorter, in the perpendicular directions 10% shorter. If the crystal is diagonally oriented (no principal axis along the beam direction) then the apparent angles between axes will violate the lattice symmetry. Incorrectly defined angles between detector and the beam will result in incorrect angles between crystal axes. Therefore, the interpretation of the lattice symmetry is dependent, to some extent, on how precisely the detector position parameters are known *a priori*.

### I think the direct Beam Position is wrong. How do I change it?

There are two ways to change the direct **Beam Position**: analytically and empirically. Analytically means that you just input the direct **Beam Position** as coordinates into the **Site Configuration** menu. Unless you have the password you will not be able to save this information to the site definition file, but it will be in use until you quit the program. If you do change the direct **Beam Position** this way, you may want to click **Abort Indexing** and re-run the auto-indexing. *HKL-2000* also allows you to change the position of the direct beam empirically by moving a cursor around on the image and seeing how this changes the positions of the predicted reflections. To do this, click on the button that says **SET BEAM POSITION**. A dialog box will appear (Figure 30).

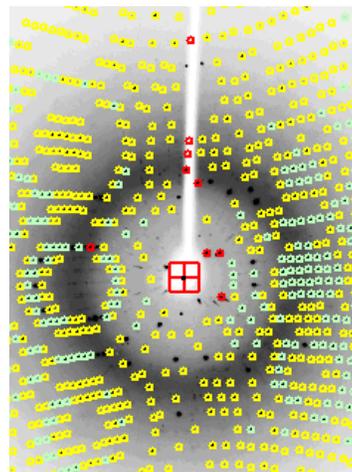
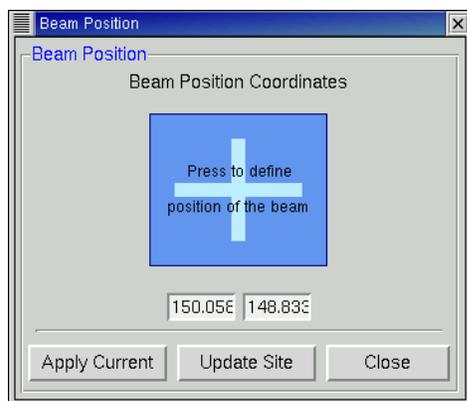


Figure 30. The Beam Position coordinates box and its position defined on the image by the red cross

Clicking on the big blue square will activate the cursor such that the coordinates of wherever you click in the main **Image Display** window will be entered into the fields for the new **X** and **Y Beam Positions**. At this time only the main **Image Display** window is active for this operation, not the **Zoom** window. These new positions do not take effect, however, until you click on the button that says **APPLY CURRENT**. Clicking here does two things. First, it enters new **Beam Position** coordinates into the refinement, and second, it gives a **GO** command to *Denzo*, which starts the

refinement (equivalent to clicking the **REFINE** button). This allows you to see the effect of the new **Beam Position** on the predicted reflections right away. You can also enter these new values into the **Site Configuration** file by clicking the **UPDATE SITE** button.

### What should I set at the 3D window?

A good rule of thumb is to set the number of oscillation degrees covered by the **3D Window** (after the resolution limits and lattice have been determined) to be at least twice the refined mosaicity. For example, if the mosaicity is about  $1.0^\circ$ , and the frame width is  $0.5^\circ$ , a value of 5 is a good choice. The default value of 5 is reasonable for macromolecular data, with a frame width of  $1.0^\circ$ . Since all of the frames in the **3D Window** are held in memory during the refinement, this may also restrict the number of frames used in the **3D Window** on older computer systems. For small molecule structure determinations, where the reciprocal lattice is sparsely populated, a **3D Window** of 30 frames is a good choice (Figure 31). The definition of the 3D window is on page 42.

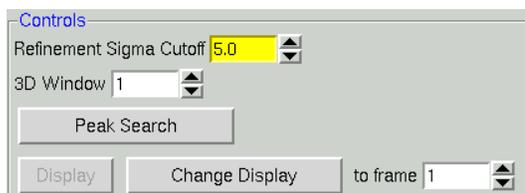


Figure 31. The 3D window in Controls box

### What's the procedure for choosing the correct lattice?

The lattice is best chosen after the first frames of the **3D Window** have been refined in space group P1, but before the whole set of frames is integrated. Since primitive triclinic is the lowest possible symmetry for your crystal, if you can't succeed in refining in P1, choosing a higher symmetry lattice with fewer refineable parameters will be even worse.

Clicking on the **REFINE** button starts the refinement of the auto-index lattice. Make sure that you are under the **FIT ALL** regime – you do not want to be fixing refinement parameters. You should notice the number of spots used in the refinement go way up, and the  $\chi^2$  errors in X and Y positions approach 1.0. Before proceeding further, go back to your diffraction **Image Display** window and examine the positions of the yellow or green circles with respect to the diffraction spots. If indexing and refinement worked, the predicted reflections (circles) should match up with the diffraction pattern.

If the predictions do not line up with the diffraction spots, two key values to check are the **Distance** and the **X and Y Beam Positions**. THESE MUST BE CORRECT. If these are correct but the predictions still do not line up with the spots you have probably misindexed and you should examine the parameters of the experimental setup. Note that some instruments do not produce proper information in the image header and you may have to manually input some parameters.

If the  $\chi^2$  values are high (orange or red colored), things to examine are:

1. The **Spot Size**. Too large a **Spot Size** will result in larger  $\chi^2$  values. There will be more discussion of **Spot Size** later in this tutorial.
2. The second thing to suspect is an error with the **Site Configuration** - for example, the spindle rotation direction, or even orientation, may be specified incorrectly – or the **Frame Width**. An easy way to test for this is to reduce the size of the **3D Window** from 5 frames (the default for CCD detectors) to 1 frame, hit **REFINE**, and note the  $\chi^2$  values again. If they improve (move closer to 1.0 and change to green in color) this tells you that there is a problem with specifying the relationship between successive oscillation frames.

- a. The most likely candidate is an incorrect **Frame Width** (perhaps the header was written incorrectly). This can be changed in the **Frame Geometry** panel of the **Main** page.
- b. A second candidate, especially with data collected from customized goniostats, is that the spindle rotation direction is incorrectly specified. To change the goniostat misalignment parameters, go to the **Site Configuration** box and change the sign of the motor axis value. For example, if the axis is specified as 0 1 0, you could try changing it to 0 -1 0. Changing the order of the numbers, e.g. to 1 0 0 means that you are specifying a different rotation axis orientation, a *severe* error.

As in all things, the proof is in the pudding. If the  $\chi^2$  values improve (after aborting the refinement and re-indexing to get more accurate values of the **Distance** and the crystal rotations), then continue on.

Assuming that the predicted reflections still match up with the diffraction pattern and the  $\chi^2$  values are not red colored (i.e. above 5), now is the time to select the proper Bravais lattice. So click on the **BRAVAIS LATTICE** button and examine the table ([Figure 32](#)). The **Bravais Lattice** table shows the changes that need to be made to the current unit cell parameters ( $a$ ,  $b$ ,  $c$ ,  $\alpha$ ,  $\beta$  and  $\gamma$ ) in order to force that unit cell into each lattice, along with a distortion index, which is a percentage of how distorted the resulting geometry is with respect to the primitive triclinic geometry. The lower the distortion index, the more likely the given lattice is correct; by definition, the distortion index for the primitive triclinic lattice is 0%. If you know (really know, not “kind of” know) the space group ahead of time, you can select the correct lattice. If not, select the lattice that has the highest symmetry and the lowest distortion index.

If a lower symmetry lattice (e.g. primitive monoclinic at 0.13%) has a lower distortion index than a higher symmetry lattice (e.g. primitive tetragonal at 0.57%) then sometimes this may be the better choice, even if the higher symmetry lattice (e.g. tetragonal) is colored green. Why is this? Thanks to the use of the **3D Window** in the refinement process, the indexing now uses many more reflections and encompasses a thicker slice of diffraction space than the previous incarnation of auto-indexing and refinement in *Denzo*. As a result, the distortion indices are much more accurate, and may reflect real deviations from crystallographic symmetry, rather than just an incomplete refinement of crystal parameters.

What’s the best way to proceed? A systematic way to choose the correct lattice is to march your way up the list of candidate lattices (candidate lattices being the green ones) and look for a significant **change** in the distortion index. For example let’s say you think you have a tetragonal crystal, and the candidate lattices are primitive triclinic, C-centered monoclinic, primitive monoclinic, C-centered orthorhombic, primitive orthorhombic, and primitive tetragonal. After primitive triclinic, you would select primitive monoclinic, refine, and examine the **Bravais Lattice** table again. The distortion index for monoclinic (and all lattices of lower symmetry – take note!) should be zero. Are the distortion indices for the higher symmetry lattices still close? If there is a significant jump (say > 0.2%) to go to a higher symmetry lattice, then perhaps those lattices are not correct and you should stick with the lower symmetry lattice.

Keep in mind that a crystal misindexed to a lower symmetry lattice will merely be characterized by a high degree of “non-crystallographic” symmetry and can reliably be solved and refined as long as complete data was collected, whereas a crystal misindexed to a higher symmetry lattice will cause you no end of grief as you try to force idealized symmetry onto a system that lacks it. That’s why it is important to accurately determine the lattice of your crystal early on in a data collection session, so that you will not be surprised and shocked when your best so-called “tetragonal” crystal turns out to be monoclinic and you’ve only collected half a data set and the crystal is lying in a puddle on the floor of the synchrotron hutch.

Once you have selected the lattice, by clicking on the little radio-button by the one you chose, click the **APPLY** bar to input the new lattice.

Bravais Lattice Table								
<i>Autoindexing performed for unit cell between 1.6 to 49 Angstroms</i>								
◇ primitive cubic	44.26%	24.26	5.29	7.04	89.90	90.02	90.01	
		12.19	12.19	12.19	90.00	90.00	90.00	
◇ I centred cubic	62.60%	25.26	8.80	24.83	97.33	20.24	102.86	
		19.63	19.63	19.63	90.00	90.00	90.00	
◇ F centred cubic	61.33%	8.80	49.31	8.81	100.27	106.17	87.16	
		22.30	22.30	22.30	90.00	90.00	90.00	
◇ primitive rhombohedral	15.95%	25.26	24.83	24.26	12.30	16.18	20.20	
		24.78	24.78	24.78	16.23	16.23	16.23	
		7.92	7.92	73.31	90.00	90.00	120.00	
◇ primitive hexagonal	15.29%	7.04	5.29	24.26	90.01	89.98	90.10	
		6.16	6.16	24.26	90.00	90.00	120.00	
◇ primitive tetragonal	6.78%	7.04	5.29	24.26	89.99	89.98	89.90	
		6.16	6.16	24.26	90.00	90.00	90.00	
◇ I centred tetragonal	8.00%	7.04	5.29	49.31	96.16	98.20	89.90	
		6.16	6.16	49.31	90.00	90.00	90.00	
◇ primitive orthorhombic	0.04%	5.29	7.04	24.26	89.98	90.01	90.10	
		5.29	7.04	24.26	90.00	90.00	90.00	
◇ C centred orthorhombic	2.57%	5.29	48.80	7.04	89.99	89.90	96.21	
		5.29	48.80	7.04	90.00	90.00	90.00	
◇ I centred orthorhombic	4.27%	5.29	7.04	49.31	98.20	83.84	90.10	
		5.29	7.04	49.31	90.00	90.00	90.00	
◇ F centred orthorhombic	8.06%	8.80	8.81	49.31	92.83	100.27	73.83	
		8.80	8.81	49.31	90.00	90.00	90.00	
◇ primitive monoclinic	0.01%	5.29	24.26	7.04	90.02	90.10	89.99	
		5.29	24.26	7.04	90.00	90.10	90.00	
◇ C centred monoclinic	2.56%	5.29	48.80	7.04	90.01	90.10	96.21	
		5.29	48.80	7.04	90.00	90.10	90.00	
◇ primitive triclinic	0.00%	5.29	7.04	24.26	89.98	89.99	89.90	

*If you would like to change the crystal lattice: select desired bravais lattice, press Apply button and close window, otherwise just close window.*

Apply      Apply & Close

Figure 32. The Bravais Lattice table

### What if no lattice other than primitive triclinic comes up green?

This could mean that indeed you have a triclinic lattice and that you will have to collect 360° (go “around the world”) to get a complete (anomalous) data set. Alternatively, your data may be misindexed (see the discussion about misindexing on pp. 30). If you are not able to see a particular pathology that would contribute to misindexing (e.g. incorrect direct beam position or incorrectly specified goniostat or detector parameters) then the best thing to do is to continue along the tutorial and integrate the frames. If the frames process well in P1 and the **Bravais Lattice** table continues to show high distortion indices for lattices other than primitive triclinic, then that is your lattice. If the frames do not process well in P1, then there is indeed something wrong with your indexing or with your data. In other words, if P1 indexes and integrates well, the lattice could be triclinic. If P1 does not index and integrate well: the data or their processing is bad and you cannot rule out other lattices yet.

### What are the Positional and Partiality $\chi^2$ values and what should they be? What are the numbers in parentheses?

The  $\chi^2$  values represent the average ratio, squared, of the error in the fitting divided by the expected error. In other words, it represents how close your observed errors are to those you predicted you would observe.  $\chi^2$  analysis is the foundation of Bayesian statistics. The numbers in parentheses are the number of overlaps for each class of reflections.

Refinement Information				
Space Group: P4				
Resolution: 50.00 - 1.46				
Positional:	1098	X- $\chi^2$ :	0.78	Y- $\chi^2$ : 0.71 (0)
Partiality:	1270	$\chi^2$ :	1.10	(84)
X Beam:		150.055	Y Beam: 148.844	
a:	79.06	b:	79.06	c: 38.06
$\alpha$ :	90.00	$\beta$ :	90.00	$\gamma$ : 90.00
Crystal Rotation X:		140.382	0.011	0.004
Crystal Rotation Y:		1.722	-0.002	0.003
Crystal Rotation Z:		2.408	0.002	0.008
Detector Rotation X:		0.010	-0.008	0.008
Detector Rotation Y:		0.013	-0.003	0.009
Detector Rotation Z:				
Crossfire X:		0.223	0.022	0.034
Crossfire Y:		0.228	-0.022	0.033
Crossfire XY:		-0.009	-0.001	0.049
Distance:		75.138	0.026	0.016
Mosaicity:				
Y Scale:		0.99898	0.00026	0.00024
Skew:		0.00033	-0.00003	0.00017

Figure 33. The refinement information e.g. positional and partiality  $\chi^2$

A good refinement will have  $\chi^2$  values near 1.0. In the early stages of refinement these values can be high, but towards the end they should be below 2 or so. All other things being equal,  $\chi^2$  values for x and y positions are sensitive to the **Spot Size** - a larger **Spot Size** will result in larger  $\chi^2$  values. It makes sense when you think about it: choosing a very large spot allows you considerable latitude when choosing the center of the reflection, since small changes in position won't appear to have much effect on the integrated spot, and thus you will not have "pinned down" the center of the spot very well. A very small spot size, on the other hand, will be quite sensitive to the center of the reflection, since small changes here would have you integrating the sides of the reflection and not the center, and so you will detect that one place is good to be and another is worse.

The magnitude of the  $\chi^2$  values is not a critical test for the success of *Denzo*, since they represent only the comparison of the spatial differences between the observed and predicted reflections to an error model that is typically quite strict in the prediction of small errors. Assuming default values of the error estimates,  $\chi^2$  values of 2 or even 3 are acceptable, because the position of the predicted reflection, and hence the intensity, is still very accurate. Thus, the  $\chi^2$  values may not reflect errors in the integration of the reflections.

- ➡ Note that you can "improve" your  $\chi^2$  values (get them closer to 1) by changing the error model. The most likely candidate for change is the value for "error positional." This value represents the error, in mm, in the positional measurement of the spot center. The default values of **Error Positional** are set in the program for "typical" crystal for particular detector.

This can be changed by going to the **Macros** tab and entering the keyword "**ERROR POSITIONAL** [value]" under the field **Immediate Execution**, where [value] is replaced by your new error, e.g. 0.03 (Figure 34).



Figure 34. The Macros box

You should be cautious about increasing this value, especially for a well-tested and characterized detector, because the high  $\chi^2$  values may reflect underlying problems with your crystal, rather than an actual error with the detector. For example, your crystal may be slightly cracked, with incomplete spot splitting, and the spot shape will be enlarged and distorted, reflecting the contribution from the multiple lattices. This will lead to ambiguities in determining the centers of the diffraction spots, which will be reflected in larger  $\chi^2$  values.

It is a good idea to keep your error model roughly the same from crystal to crystal so that you can use the  $\chi^2$ 's to pick up any other variations in your data collection. For example if you are getting  $\chi^2$ 's in the range of 2 to 5 consistently

and your crystal is good you probably have problems with the detector calibration. Note that most people think that their crystals, like their children, are outstanding. At least for crystals, this is not always the case.

If the  $\chi^2$  values are very high, ( $> 10$  or so for default values of the error parameters) something is seriously wrong with the indexing, refinement, or the detector. This will be apparent from visual inspection of the displayed image in any case.

Since the goal of *Denzo* is to produce a list of *hkl*'s and unscaled intensities, it does not matter that the errors in the positions differ a bit from the standard expected error. On the other hand,  $\chi^2$  values substantially different from 1 in scaling should be investigated, because this directly affects a critical result of the scaling procedure, namely the  $\sigma$  value assigned to each scaled intensity. So that's where you should pay attention to finer deviations of  $\chi^2$  values from 1.

### What are all those Refinement Information values for, and what do they mean?

In **FIT ALL** regime you can successfully refine all the parameters – crystal rotations, detector rotations, detector distance, unit cell dimensions, etc. – simultaneously, if their initial values are close enough. The **FIT BASIC** options select the most critical values. These are usually the crystal rotation and unit cell parameters, and the beam position. The resolution limits must be set first (discussion on page 26).

If you are under the **FIT BASIC** regime, then you have refined the parameters listed in the information panel (see [Figure 33](#)):

- $\chi^2$  values for good refinements will be close to 1
- X beam and Y beam position of the direct beam at the detector (in mm)
- refined crystal unit cell length (in Ångstroms) and angles within symmetry of selected Bravais lattice
- the rotations of the crystal (in degree) applied to the predicted reflections (preds) from initial (absolute) conditions; values are entered as absolute rotations from zero

If you are under **FIT ALL** regimes there are listed values for additionally refined parameters:

- detector rotation X, Y, Z
- crossfire X, Y, XY
- refined detector to crystal distance (in mm) can be different than the value provided in [Experiment Geometry](#) box on the main page
- mosaicity (when **3D Window** is  $> 1$ )
- Y scale
- skew

The detector and crystal parameters are refined by a least squares method that minimizes the deviation of the reflection centroids from their predicted positions.

Someone told me to gradually increase the resolution and alternate cycles of fitting a few parameters and fixing things. I'm patient, and I want to be careful. Why shouldn't I do this?

This is the **wrong** way to go about indexing and integrating your images. For example, back in the old days, people would get an initial index at low resolution, fixing all parameters except for, say, the crystal rotations. Then they would fix the rotations and refine the distance. Then they would fix the distance and refine the unit cell parameters. Then maybe they would increase the resolution a bit and repeat the process. The problem with this approach is that single parameter refinement may not allow correlated parameters to be sufficiently refined. In other words, you could get trapped into overfitting one variable and underfitting another, resulting in both variables being incorrectly determined. So unless you know with near-absolute certainty that a variable is correct, cycles of fixing and fitting is the wrong way to go.

People used to take these kinds of approaches because refinements were unstable and simultaneous fitting of multiple correlated variables led to instability in the refinement. High correlation makes it possible for the errors in one

parameter to compensate partially for the errors in other parameters. If the compensation is 100%, the parameter would be undefined, but the error compensation by other parameters would make the predicted pattern correct.

*HKL-2000* overcomes this problem in two main ways. First, **eigenvalue filtering** (described in *Numerical Recipes* as Singular Value Decomposition) is employed to remove the most correlated components from the refinement and make it numerically stable. Eigenvalue filtering works reliably when the starting parameters are close to the correct values but it may fail to correct large errors in the input parameters. Therefore, once the whole data set is integrated, the global refinement (sometimes called **postrefinement**) can refine crystal parameters more precisely. The unit cell used in further calculations should come from the global refinement (in *Scalepack*) and not from the *Denzo* refinement.

Second, refinements are generally carried out at the maximum resolution of the diffraction data, and, with the **3D Window**, over multiple successive frames. More data allows more parameters to be fit simultaneously, and increasing the dimensionality (by adding a third dimension) further constrains variables that might be poorly fit in only 2 dimensions (like the distance, or the third unit cell dimension). These advances are made possible by improvements in the program as well as improvements in computer speed, data storage, and memory.

## Step 10: New lattice refinement

Click the **REFINE** button again and refine in the new lattice. Check the image to make sure the predicted reflections line up on real spots.

**Problems – predicted reflections don't line up with spots, things get worse, etc.**

Your crystal may still be misindexed (see discussion see on page 30). Alternatively, you may have chosen the wrong lattice and will have to choose another lattice, usually with lower symmetry. However, these sorts of problems don't usually show up until the set integration process has begun.

**How do I start over? How do I get back to P1?**

To start over completely, you can go to the **File** menu and quit the program. However, if you just want to try auto-indexing again you can click on the **ABORT REFINEMENT** button, go back to the image and re-do the peak search, and proceed as before. To re-set the space group without aborting the refinement, you can re-open the **Bravais Lattice** table and choose another lattice, like primitive triclinic. Click **APPLY & CLOSE**.

## Step 11: Resolution Limits, Fitting Radius and Spot Size

While referring to the indexed image, carefully set the correct **Resolution** limits, **Profile Fitting Radius**, and **Spot Size**. Click the **REFINE** button to update the display after changing any of these parameters. If the crystal *rot<sub>y</sub>* value displayed in the **Refinement Information** panel is colored red (i.e. for  $45^\circ < rot_y < 135^\circ$ ) click on the **REFERENCE ZONE** button in the **Controls** panel and choose a new zone with a better value for *rot<sub>y</sub>*.

**What is the Profile Fitting Radius and how do I determine what it should be?**

The **Profile Fitting Radius** is the radius of the area around each particular spot, in mm, containing neighboring spots, used to calculate the average spot profile (Figure 35). The spot in question is fitted to the average profile of all the spots within the specified radius. Generally, the radius is set so those spots on roughly 3-5% of the area of the detector are included in the averaging.

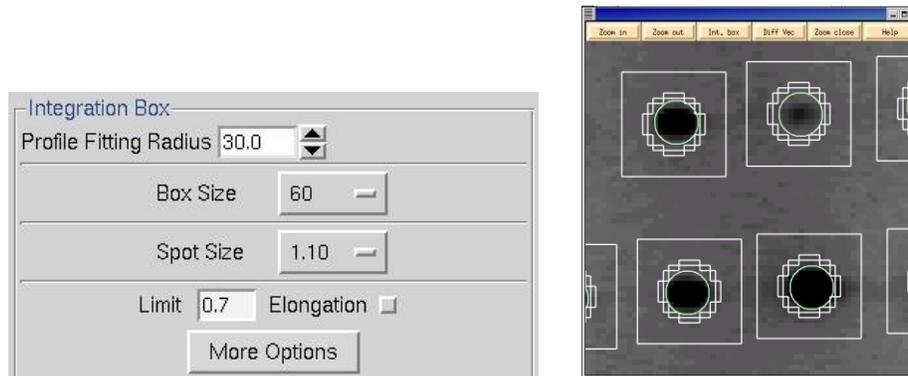


Figure 35. The Integration Box and Zoom windows with Profile Fitting Radius and Spot Size boxes

The particular size of the area chosen depends on how much the profiles vary across the film and the density of the spots on the film. For example, if the spot profiles vary a lot across the film, then you would choose a smaller radius. If you have a small lattice and the spots are widely separated, you might want to choose a larger radius.

The calculation of the average profile is a time consuming task and it is proportional to the number of spots in the profile fitting radius circle, but too small a radius will not capture enough spots and lead to a noisy average profile. Too large a radius will average out significant profile variations.

To see the **Profile Fitting Radius** in the *Image Display* window click **PROF FIT R** button (Figure 36). A thin circle corresponding to the profile fitting radius will appear and will move around with the cursor. The pred display changes now too. Instead of seeing all of the predicted reflections, you see only those preds that correspond to the reflections that are stronger than the refinement sigma cutoff (usually set at  $5\sigma$ ). Only this subset is shown because these are the only reflections used in the calculation of the average spot profile within the radius. (Note: to return to the normal mode, click **PROF FIT R** button again).

Prof fit R

Figure 36. Profile fitting radius button

A good rule of thumb is that the **Profile Fitting Radius** should enclose about 10 to 50 spots whose intensities are above the weak level. Useful increments of the profile fitting radius are 2.5% of the detector size (e.g. 5 mm for a 200 mm wide IP).

The radius can be set in the **Index** window under the **Integration Box** panel in the lower left quadrant of the window (Figure 35). The value is given in millimeters, and you can adjust it iteratively, hitting the **REFINE** button in between your guesses. In general it is preferable to enlarge the radius so as to make sure that there are no orphan reflections (reflections that, when placed in the center of the circle, have no other non-weak reflections within the circle). However, keep in mind that if you are using a **3D Window** greater than 1, that the **Profile Fitting Radius** is actually the “Profile Fitting Sausage,” since it encompasses the stack of frames in the **3D Window**, and more spots may be used than first seems from examining a single frame.

**How do I set the spot size? Is it all that important? Why do larger spots make my  $\chi^2$  's worse?**

The final parameter to set is the **Spot Size**. This is generally a function of the size of your crystal and the quality of the X-ray beam (collimation, crossfire, parallelism, etc.) but it must be set for proper integration.

If you choose too small a **Spot Size** significant data is lost because it is cut off. Choosing too large a **Spot Size** runs the risk of 1) rejecting reflections due to overlaps, 2) including portions of the neighboring reflection in the measured intensity, and 3) including too much of the background in your intensity measurement.

Setting the **Spot Size** is somewhat of a tradeoff, since not all the spots on an image are of the same size, and unfortunately, one spot size and shape must suffice for the entire set of images (with the exception of correcting for the elongation caused by the intersection of reflections at high angle to the detector face, which is corrected by the parameter **Elongation Limit**). However, it is unlikely that your data will suffer too much because profile fitting limits the damage caused by including the background in the spot. As a general rule of thumb, then, it is best to try to adjust the spot shape to fit some of the medium to strong spots on the image.

The spot can be either circular or elliptical. Clicking on the **MORE OPTIONS** button allows you to specify elliptical spots if your spot shapes are sufficiently elongated (Figure 37).

- ➡ However that there is only so much that can be done to compensate for problems with crystal order, split crystals,  $\text{CuK}\alpha_1$ ,  $\text{CuK}\alpha_2$  splitting, etc. Bad crystals are still going to give bad data.

After making the changes to your spot shape, or turning on the **Elongation Limit** option, you can verify that the spot shape has changed by examining the image and the predicted spots in the **Zoom** window of the **Image Display** window (you may have to click **UPDATE PRED** to show the predictions).

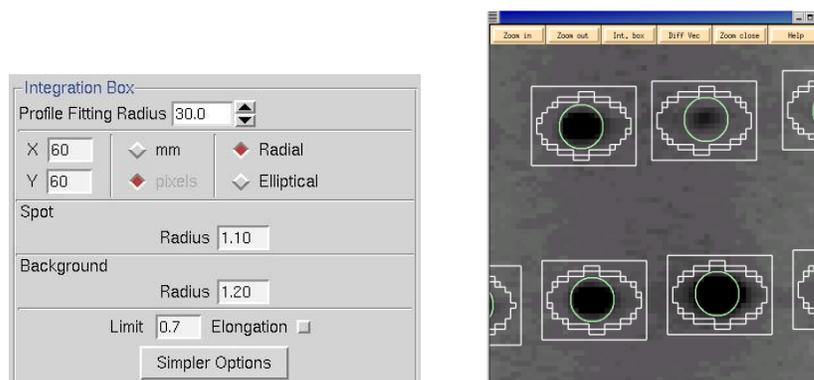


Figure 37. The Integration Box and Zoom windows with more option for spot size and shape

### What is the Elongation Limit?

Note that *Denzo* now has an option to take into account the radial elongation to the spots that arises from the intersection of the diffracted X-rays at high angle with the detector. This is activated by checking the **Elongation Limit** button in the **Integration Box** panel. The default value for **Elongation Limit** (expressed as the fraction of the box size the spot is allowed to elongate) is 0.7. The range is from 0 to 1.

### Resolution Limits: should I set this here or not worry about it?

Examine the indexed diffraction pattern carefully. The cursor can be used to determine the resolution of a particular spot. In general, you want to be on the generous side (say 0.2 Å) with setting the resolution limits, since they can always be lowered in **Scaling** if you were too optimistic. Conversely, if you are too stingy with the resolution limits in *Denzo* you will have to go back and re-integrate all your frames again, since *Scalepack* cannot increase the resolution limits beyond what you have set in *Denzo*. You may have to darken the image—by clicking on the **BRIGHT** button several times—in order to see the high resolution reflections clearly. Using the **Zoom** window also helps. Once you have determined the proper resolution limits, go back to the **Index** window and input the new resolution limits into the proper box. Then hit the **REFINE** button again and the new predicted reflections will appear in the diffraction pattern window. You may have to do this a couple of times until you get it right. Look in all quadrants of the diffraction image, (and if possible, multiple frames) as crystals often diffract anisotropically.

### What are all these parameters Rotx, Roty, Rotz, X-beam, Y-beam, crossfire, mosaicity?

These are parameters that define the crystal and the geometry of the diffraction experiment.

The crystal **Rotx**, **Roty**, and **Rotz** angles define the orientation of the crystal lattice in space for the calculation of the predicted reflections.

It is inconvenient and non-intuitive to work with images in the data conventions, because we are used to thinking about diffraction geometry in terms of the X-ray beam, the spindle axis, and the laboratory frame of reference (horizontal, vertical). There are two possible conventions within this "intuitive" set:

### Gravity/beam system (not used in *Denzo*)

In this system, the X-ray beam is defined as being perpendicular to the direction of the force vector due to gravity ("down"). The beam is the *z* axis, gravity is parallel to the *y* axis, and the *x* axis is perpendicular to both of these. In terms of crystal orientation rotations

- `rotz` would denote rotations of the crystal around the **beam** axis,
- `roty` would denote rotations of the crystal about the **vertical** axis and
- `rotx` would denote rotations of the crystal about the **horizontal** axis.

This convention has the advantage that it does not change with different camera spindle geometry. In addition, it corresponds directly to the common orthogonal detector directions, in that the detector directions *x* and *y* on the film or IP are perpendicular to the X-ray beam. A disadvantage of this system is that the *x* axis may or may not correspond to the spindle axis, depending upon the  $\chi$  (chi) and  $\Omega$  (omega) setting angles of the 3-axis goniostat. For the purposes of aligning your crystal, this makes life more complicated, as we are used to thinking about aligning crystals by moving the arcs on a goniometer head. For  $\chi$  equal 90 or 270 degrees and  $\Omega$  equal to zero, however, the spindle axis is parallel to the *x* axis and thus `rotx` corresponds to rotations about the spindle axis.

Although this system will not (formally) work in outer space (where the gravity vector is essentially zero), nor in the event that the X-ray beam is vertical, the fact remains that (so far) all macromolecular crystallography is done with horizontal X-ray beams on the surface of good old planet Earth, and thus it is a viable option. While *Denzo* does not use this convention, *XdisplayF* does.

### Spindle/beam system (used in *Denzo*)

The spindle/beam system is the *Denzo* crystal and cassette orientation convention. In this system,

- the *z* axis is again parallel to the **beam**,
- the *x* axis is parallel to the **spindle** axis and
- the *y* axis, also termed the "**vertical**" axis (for want of a better word), is perpendicular to the spindle and beam axes.

In terms of crystal orientation rotations:

- `rotz` would again denote rotations of the crystal around the **beam** axis,
- `rotx` would denote rotations of the crystal about the **spindle** axis and
- `roty` would denote rotations about the axis perpendicular to the beam and the spindle.

The advantage of this system is that it is very intuitive to the crystallographer. In addition, with Eulerian  $\chi$  values of 0, 90, 180 or 270 degrees the spindle and vertical axes are parallel to the *x* or *y* axes of the detector. The main disadvantage of this system is that it is dependent on knowing the geometry of the camera.

While the crystal and cassette orientations follow the spindle/beam convention for *Denzo*, the beam, boxprintout, spot, margin, film width and length in the *Denzo* log file follow the data convention

The **X-beam** and **Y-beam** values are the distance from the edge of the detector data collection area to the beam spot, in mm.

**Crossfire** is a measure of the X-ray beam divergence and focusing as it leaves the collimator and illuminates the crystal. Crossfire, being a symmetric tensor, has *x*, *y*, and *xy* components. It affects the prediction of partial

reflections and their positions, not their angular width. It is expressed as angular divergence of the beam. The default value is zero crossfire, i.e. a perfectly parallel beam or a beam focused on the detector.

**Mosaicity** is defined in *Denzo* as the rocking angle, in degrees, in both the vertical and the horizontal directions, which would generate all the spots seen on a still diffraction photograph. It includes contributions due to X-ray bandwidth, beam crossfire, etc. Mosaicity is refineable when integrating your frames if the **3D Window** is set to be at least twice as large as the refined mosaicity. For example, if your crystal has a mosaicity of about 1 degree, then a **3D Window** encompassing 2 degrees or more of data should allow for proper refinement. If mosaicity grows beyond a reasonable limit during refinement it indicates misindexing or a wrong goniostat definition. Most common is the movement of the crystal in a direction opposite to the one assumed.

All of these parameters are subject to refinement by the integration program.

### What is the Reference Zone? How is this useful?

The **Reference Zone** shows the current zone and other zones which are crystallographically equivalent to the current zone. Keep in mind that upon indexing the crystal for the first time (which is just another way of stating that the crystal orientation parameters are being determined) *Denzo* makes an arbitrary choice of reference zone (Figure 38). Being able to choose an equivalent zone is useful in a number of ways.

1. If the crystal `roty` happens to fall between 45 and 135 degrees, this has a higher correlation than other zones and makes the refinement of the crystal orientation more difficult (although it will still work). Choosing an equivalent zone from the list is an easy way to overcome this.
2. If, for some reason, two or more data sets are collected from the same crystal, without removing the crystal or changing its orientation relative to the goniostat, and the data sets are indexed independently of one another, then it is possible that *Denzo* will arbitrarily assign a different reference zone to the two data sets. Unfortunately, this will result in a different original index for reflections which otherwise should be equivalent, and so merging reflections between data sets will not work. However, by choosing the same reference zone for both data sets, scaling and merging reflections can be done correctly and accurately.

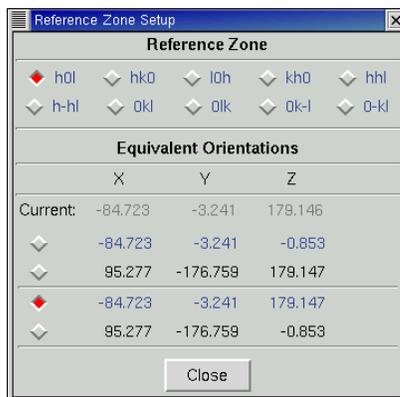


Figure 38. The Reference Zone window

## Step 12: 3D Window and Mosaicity

Set the **3D Window** so that the **Oscillation Range** encompassed by this sector of reciprocal space is at least twice as large as the refined mosaicity. Typically a value of 5 frames is used for CCD detector data. For IP data you should keep the **3D Window** at 1 initially, due to a high probability that consecutive frames are not scanned in the same way.

### What is the 3D Window?

The **3D Window** specifies the number of consecutive frames from which the reflections will be integrated simultaneously. Typically for CCD data a value of 5 frames is used. One principal advantage of this feature is that binning frames together can include more reflections in the refinement, thereby increasing the accuracy of indexing on frames that have very few reflections (e.g. small molecule data sets, low mosaicity / small unit cell macromolecular data sets, low resolution macromolecular data sets). The disadvantage of grouping frames together is that local variations in frames are averaged out (e.g. "bad" frames, frames where the crystal slipped, etc.)

- For data collected on IP scanners (Raxis, Mar, DIP) errors in the reproducibly positioning the IP with respect to the scanning machinery preclude the use of a **3D Window** greater than 1. That's why the default value for the **3D Window** is 1 for these sites.

### Why do I have to keep the 3D Window at 1 for image plate data?

Unlike CCD detectors, which are solid state, image plate detectors move the IP relative to the scanning head in the course of data acquisition, data read-out, and plate erasure. While these devices are very well made, there is, nonetheless, a small non-reproducibility in their operation that results in slight variations in the actual position of the direct beam and the **Distance**. Until it has been established that this variation is small relative to the pixel size, the **3D Window** should be kept at 1 to avoid inaccurate mapping of the pixels of sequential frames onto each other.

### What are the limitations for 3D Window usage?

The limitation for 3D Window usage is primarily computer memory. All the frames of the **3D Window** must be held in memory simultaneously; however, most modern computers have sufficient memory to handle reasonable values of the **3D Window** parameter.

- For small molecule data collection, the reciprocal lattice is very sparsely populated and it is reasonable to set the **3D Window** to 30 or more frames.

### Why wouldn't I just want to set the 3D Window to encompass all of my frames?

One principal advantage of the **3D Window** is that binning frames together can include more reflections in the refinement, thereby increasing the accuracy of indexing on frames that have very few reflections (e.g. small molecule data sets, low mosaicity / small unit cell macromolecular data sets, low resolution macromolecular data sets). The disadvantage of grouping frames together is that local variations in frames are averaged out (e.g. "bad" frames, frames where the crystal slipped, etc.) A good rule of thumb is to set the **3D Window** (after the resolution limits and lattice have been determined) to be at least twice the refined mosaicity. The default value of 5 is a reasonable compromise for most cases.

### What is Mosaicity? What does the histogram under Check Mosaicity mean?

**Mosaicity** is defined in *Denzo* as the rocking angle, in degrees, in both the vertical and the horizontal directions, which would generate all the spots seen on a still diffraction photograph. It includes contributions due to X-ray bandwidth, beam crossfire, etc. Mosaicity is refineable when integrating your frames if the **3D Window** is set to be at least twice as large as the refined mosaicity. For example, if your crystal has a mosaicity of about 1°, then a **3D Window** encompassing 2° or more of data should allow for proper refinement.

The histogram is the mosaicity histogram for the reflections of the frame or of the frames of the **3D Window**. The reflections are sorted into 20 zones, which range from minus the input or refined mosaicity divided by 2 to plus the input mosaicity divided by 2 (Figure 39).

Each zone represents the shortest angular distance of the center of the reflection from the surface of the Ewald sphere at the end of the oscillation range, and is formally defined as the average observed partiality of the reflections in each zone. In other words, if, at the end of the oscillation range a particular reflection was -0.3° away from (i.e. past) the surface of the Ewald sphere, then it would contribute to the "bar" in the -0.3° zone.

Only reflections that are single partials are included in the analysis. A reflection in negative zones means that the center of the reflection has already passed through the Bragg condition. In all cases, no matter what mosaicity you have input to *Denzo*, the histogram should pass through 50% in the zero zone. If the mosaicity you guessed and input was too small, then the histogram will not descend to zero at  $+\text{mos}/2$ . If the mosaicity you guessed and input was too large, then you will see a histogram which looks like a step function, falling to zero just past the zero zone.

If the mosaicity was chosen correctly or slightly overestimated (which is the preferred side to err on) the histogram will descend smoothly to zero at  $+\text{mos}/2$ . Frequently this is not strictly the case and the histogram tails off due to the effects of diffuse scattering. If there are only a few reflections, the histogram may be quite choppy. The number of reflections can be increased by increasing the size of the **3D Window**. If the histogram does not resemble any of the shapes mentioned here, it may be indicative of motor, spindle, or shutter problems.

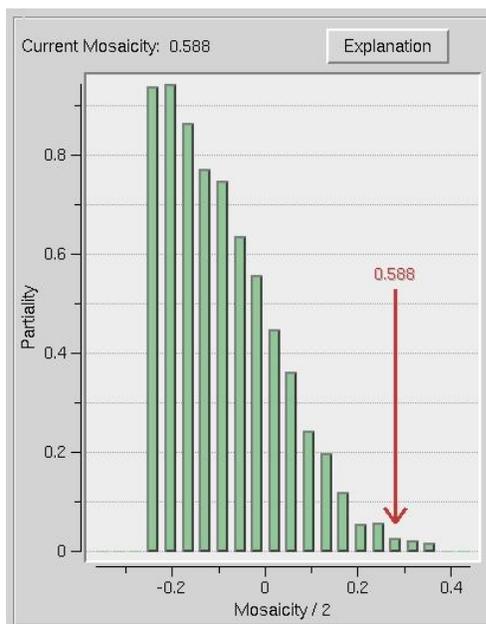


Figure 39. The mosaicity histogram

**How do I set the Mosaicity? Should I try to increase the mosaicity until all of the spots on the image are covered with preds?**

The best way to set the mosaicity is to use the **3D Window** and let *Denzo* refine it in the course of integrating the reflections. If you are unable to use a **3D Window** greater than 1, and the mosaicity is greater than half your oscillation range (which is usually the case), then you will have to set the mosaicity manually. You do this by going to the **Crystal Information** button and entering the **Mosaicity** in the field provided.

A correctly set mosaicity will not always result in predictions covering all of the reflections on an image. In particular, very strong low order reflections may require unreasonably large values of the mosaicity for coverage. These reflections may not represent the intrinsic mosaic spread of the crystal, but rather may result from particularly intense diffraction from that lattice point.

A mosaicity that is set too high is better than one that is set too low, but setting the mosaicity too high could result in an unnecessarily high number of overlaps.

**How do I increase the Mosaicity?**

You do this by going to the **Crystal Information** tab (Figure 40) and entering the **Mosaicity** in the field provided.

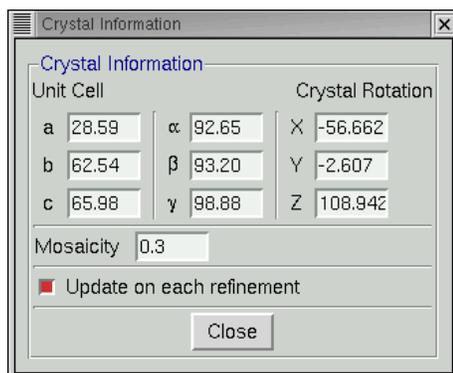


Figure 40. The Crystal Information window





# Strategy and Simulation

## Step 13: Strategy

### Why would I want to use a strategy and a simulation?

Data collection is not an automatic process. In a perfect world, crystals would be perfect and have negligible mosaicity, crystal lifetime would be unlimited, crystal exposures would be short, data collection time would be unlimited, and disk storage capacity would be infinite. In this case, one would just choose a tiny oscillation range and collect 360 degrees of data without worrying about the consequences. However, as we all know, crystals are imperfect and decay fast, data collection time is not unlimited, and storage space is not infinite. Thus, we have to make strategic decisions about how we will go about collecting a diffraction data set to avoid having an incomplete data set or one that is ruined by overlapped spots. This involves making decisions about the oscillation range to collect, the width of each frame, and the exposure time per frame. The strategy and simulation utilities in *HKL-2000* can assist with these decisions.

### How do I start a strategy session?

Index a frame or set of frames as usual, making sure that the correct Bravais lattice is selected, the resolution is properly chosen, and refine with the **FIT ALL** regime (including mosaicity with the use of **3D window**). Do not proceed to the integration step yet!

Once the first frame(s) are indexed, click on the **Strategy** tab. You will see four blank charts and three control boxes: the first for strategy, the second for simulation and the third for combining results of different simulations ([Figure 41](#)). The strategy is performed first.

Click on the button that says **STRATEGY**. In a few seconds you will see a series of colored lines in the **Strategy** plot, which shows **Delta Omega** vs. **Omega Start**—or depending on the goniostat, the rotation angle (Kappa, Chi, Phi) used for the oscillation ([Figure 42](#)). At a given **Omega** (Kappa, Chi, Phi) **Start**, the value of **Delta Omega** for each line is the number of degrees of Omega you have to collect starting at **Omega Start** to get the percent completeness represented by the line. In other words, each line corresponds to the percentage of a complete data set that can be collected for a chosen starting oscillation angle (**Omega Start**) and total number of degrees collected (**Delta Omega**).

Thus, considering the example in [Figure 42](#), if only a 20% complete (dark blue line) data set is desired (although why you'd want that is somewhat mystifying), only a 6° wide oscillation needs to be collected, regardless of starting angle. On the other hand, most people want 95+% complete data sets, so those plots are more relevant. A triangular gray region present on [Figure 42](#) marks the region that is not available for data collection due to hardware limitations.

Without clicking the mouse, move it around in the window and a cross hair will appear. Let say, you want to collect a 100% complete data set. Line the crosshairs up on the line corresponding to 100% complete. As you move it around, notice that the **Omega Start** and **Delta Omega** values in the control box are updated. In order to use and collect the most complete data set with the smallest possible oscillation range, choose a point on the 100% line at an **Omega Start** value where the **Delta Omega** is at a minimum.

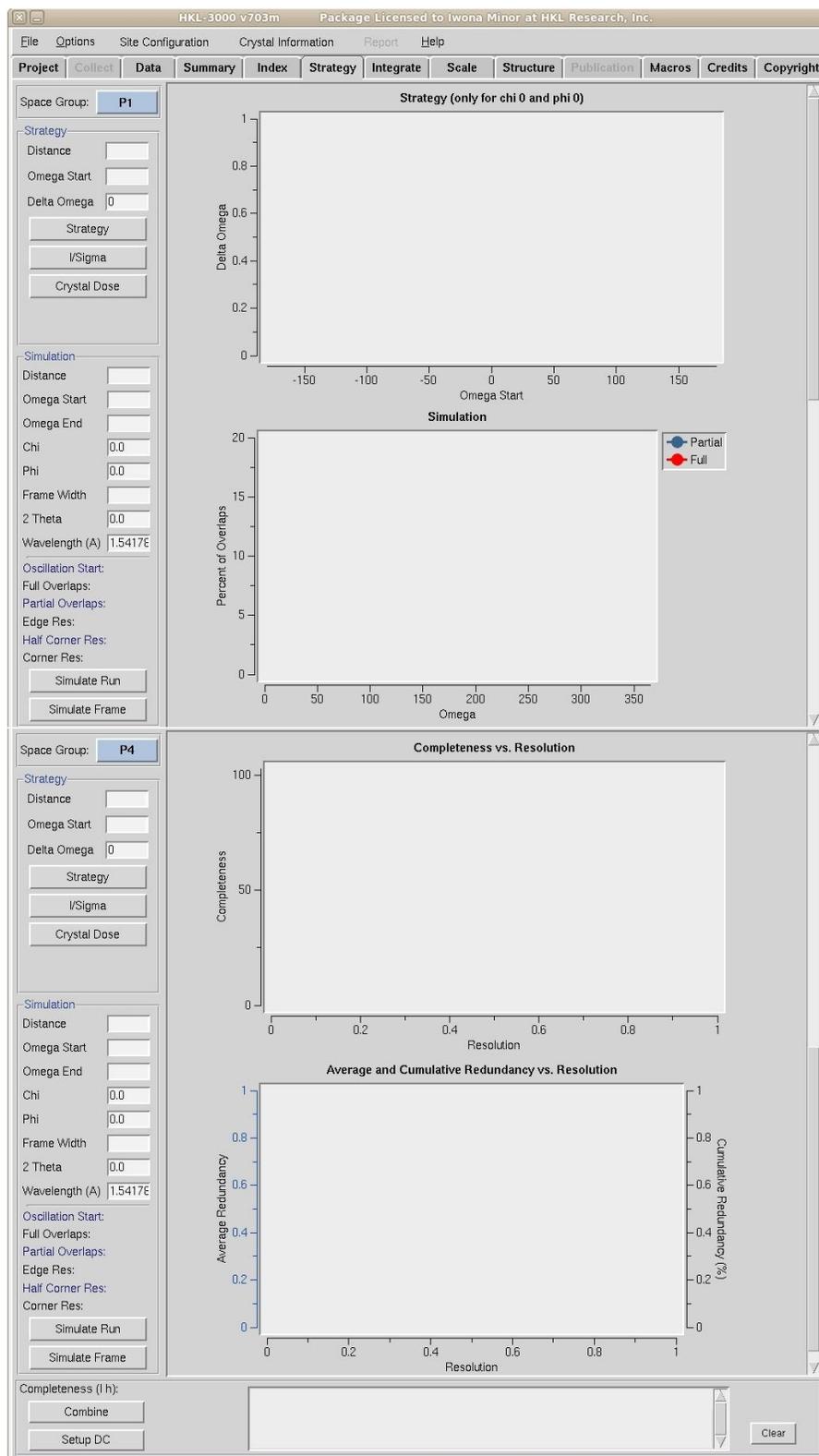


Figure 41. The Strategy/Simulation window

Note that for most high symmetry space groups the plots will resemble a sawtooth with a rapidly rising leading (left hand side) edge. In this case, it is better to err slightly on the side of caution, and select an **Omega Start** value a few degrees smaller than the point with the minimum **Delta Omega**, in order to avoid the high side of the sawtooth. If you do not have a high symmetry space group, the plot will look more like a horizontal line or a gentle wave. In those cases, there is not much strategy to ponder, since the lattice does not allow for much economy in data collection.

Once you have settled on a **Delta Omega** and **Omega Start** point on the plot, click on the mouse and the values will be entered into the appropriate boxes in the **Simulation** control panel.

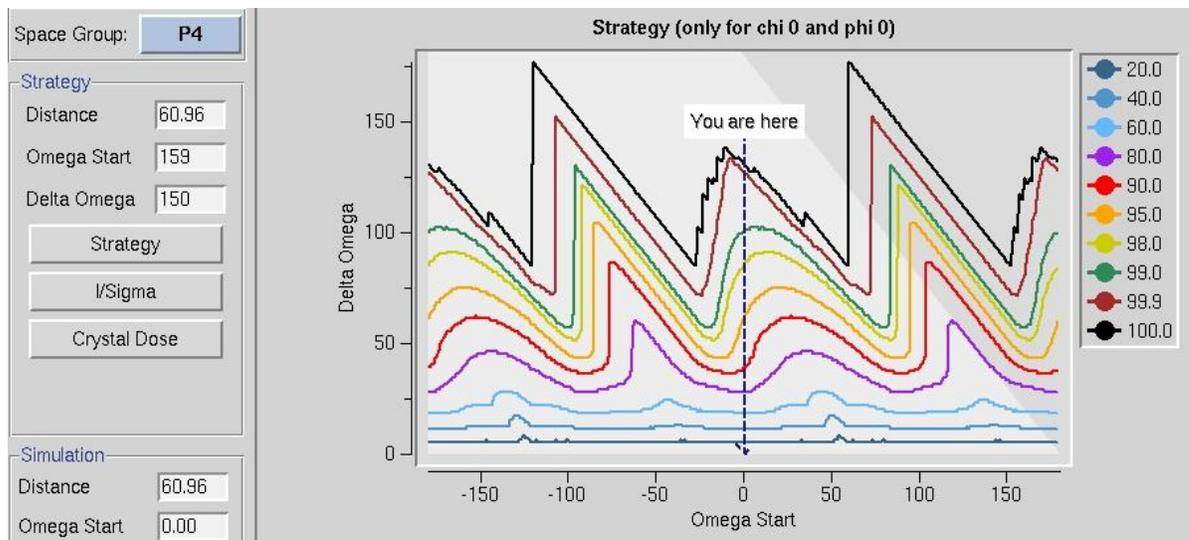


Figure 42. The Delta Omega vs. Omega Start plot

### How can I use I/Sigma and Crystal Dose buttons?

You need to collect at least 3 frames to use the **I/SIGMA** button, and the program will estimate (after a quick integration and scaling)  $I/\sigma$  up to the resolution value at the edge of the detector for the specified detector distance. A new window **Current I/Sigma** opens with an **I/Sigma vs. Resolution** plot (Figure 43). Two gray lines are drawn on the plot, at  $I/\sigma = 10$  and  $I/\sigma = 2$  respectively.

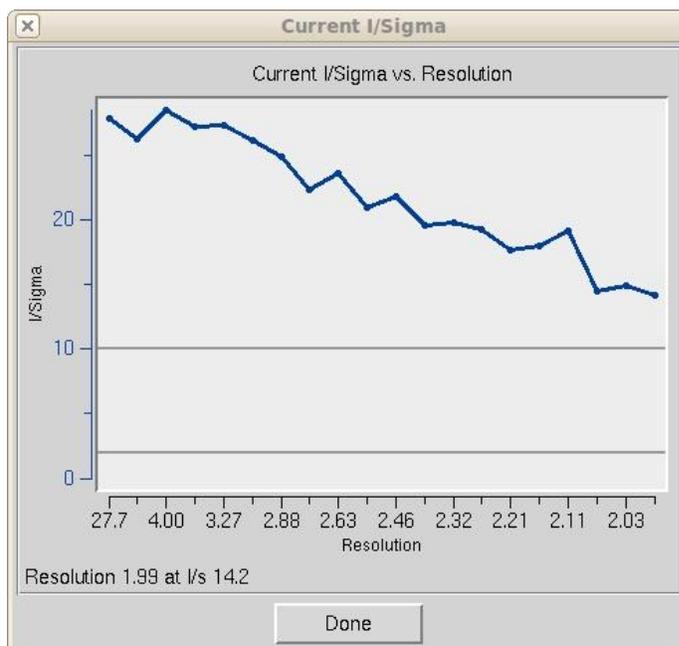


Figure 43. The I/Sigma vs. Resolution plot

The **CRYSTAL DOSE** button is to be used for X-ray sources with very high flux (typically synchrotron beam lines). It opens a new window **Estimated Crystal Dose** where you insert a correct **Beam Flux** value. The program runs a simulation estimating how radiation damage will cause the  $I/\sigma$  values to decrease at the end of data collection and shows the result as on the  $I/\sigma$  vs. resolution for **Current and Last Frame** plot (Figure 44). Note that this analysis requires the sequence of the protein in the crystal, which can be added by clicking the **EDIT PROJECT** button on the **Project** tab.

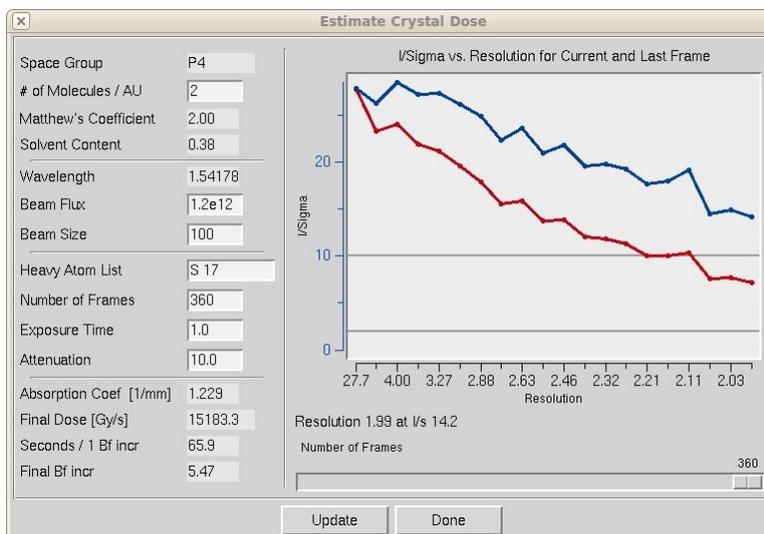


Figure 44. The Estimated Crystal Dose window

## Step 14: Simulation

### What is the simulation for, and why would I want to use it?

The simulation utility is useful for predicting the number of overlaps in an entire data collection run, as a function of frame width, distance, and oscillation range. This is particularly important if your crystal has high mosaicity or long unit cell axes. In those cases, the spots from some frames may be separated just fine, but then subsequent frames may be rendered useless by overlapping spots as (for example) a long axis that is perpendicular to the spindle axis rotates into a position parallel to the beam. The simulation utility allows you to simulate an entire data collection run, to view the potential overlaps, and then to adjust your data collection parameters before the crystal is exposed and time is wasted. It is really quite useful.

### How do I use the simulation?

Once the **Omega Start** and **Omega End** are input (either by clicking on the **Strategy** plot or by entering them directly), click on the **SIMULATE RUN** button. You will see two things. First, a plot will be drawn with red and blue lines (which may overlap), corresponding to the percentage of fully and partially recorded reflections that are predicted to be overlapped for each frame you plan on collecting (Figure 45). Remember, if a reflection is overlapped on another, it will not be useful for your data set and be excluded during integration and scaling. Second, on your **Image Display** image window a new series of predicted reflections will be displayed, corresponding to each frame in your simulated data collection run. Obviously, since you haven't yet collected the data the underlying diffraction image will not change or correspond to the predicted reflections, but that is OK (this is just a simulation). This will show you visually how your diffraction image will index over the data set collection.

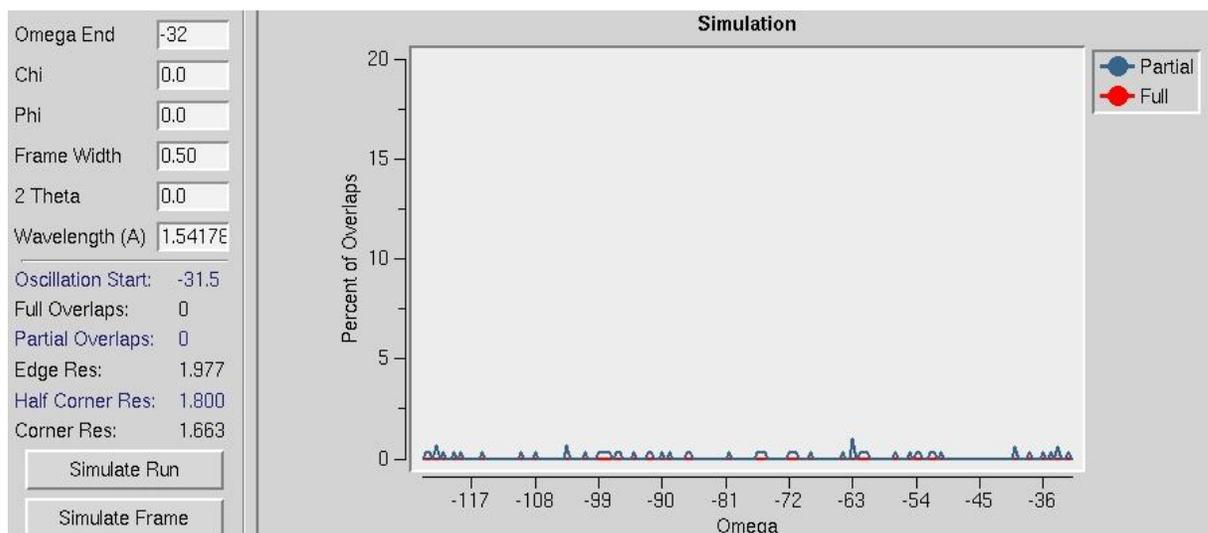


Figure 45. The Simulation window

The thing to pay attention to is the percentage of overlaps. Typically you'd like to lose no more than 5% of the reflections to overlap. This is where the simulation is useful. In the control box, you can replace the original frame width with a new value. Got a lot of overlaps? Decrease the frame width and simulate the run again by clicking on the **SIMULATE RUN** button. Examine the results; the overlaps should decrease. Only got a few overlaps and want to increase the percentage of fully-recorded reflections? Increase the frame width and simulate the run again. (Note that this is not the best strategy because exposure time per frame increases with increasing frame width, and background increases linearly with exposure time. All other things being equal, it's best to collect thin frames rather than thick in order to minimize background noise.)

### Can I change anything other than a frame width?

Typically at some point you will find a reasonable frame width (0.2-1.0° for macromolecules and 1.0-5.0° for small molecules) that yields an acceptably few number of overlapped reflections. This is not guaranteed, of course, since large unit cell axes and high mosaicity can conspire to ruin a run despite your best intentions. However, in this case, you may also consider the effect of changing (increasing) the crystal-to-detector distance. This can be input into the simulation control panel and the **SIMULATE RUN** hit again. As the detector is moved back, the spot separation should increase and the percentage of overlaps will decrease. Naturally this will limit how high the resolution of the data that can be collected, but that is life. Perhaps you are one of the unlucky people that have the triple whammy of high resolution, large unit cell, and high mosaicity. In that case you will be perpetually frustrated, but at least you will be able to develop the most rational data collection plan before exposing your crystals to an excessive amount of radiation. At some point an acceptable compromise should be reached. You can also consider collecting multiple oscillations of data with varied data collection parameters and then scaling the sets together (see “What is the Combine button for?” below).

### What do I do once I have an acceptable simulation?

Once you have settled on an **Omega Start**, **Delta Omega**, **Distance**, and **Frame Width**, you can input these values into the control software for the data collection run and begin your data collection. Typically the whole strategy and simulation exercise should take only a couple of minutes or less. It’s actually easier to do it than to read about it!

While collecting your data with your newly input parameters, it’s always a good idea to run the strategy and simulation again, just to confirm your previous analysis. If anything is wrong, you can stop the data collection before you’ve wasted a lot of time. Errors or inaccuracies in the strategy or simulation, when they do occur, usually arise from an inaccurate estimation of the mosaicity or the unit cell axes, especially when these are derived from just a few frames. Of course, nowadays with high brilliance synchrotrons, large CCD detectors, frozen crystals (that hopefully are not too radiation sensitive), and terabytes of disk storage, you can also forget about the entire discussion above and just collect thin frames and go “around the world.” But isn’t it more fun to think about what you are doing?

### What is the Combine button for?

The plots **Completeness vs. Resolution** and **Average and Cumulative Redundancy vs. Resolution** ([Figure 46](#)) show statistics for a single simulation (colors of the lines are same as explained for the *Scaling* tab of *HKL-2000* package as explained below).

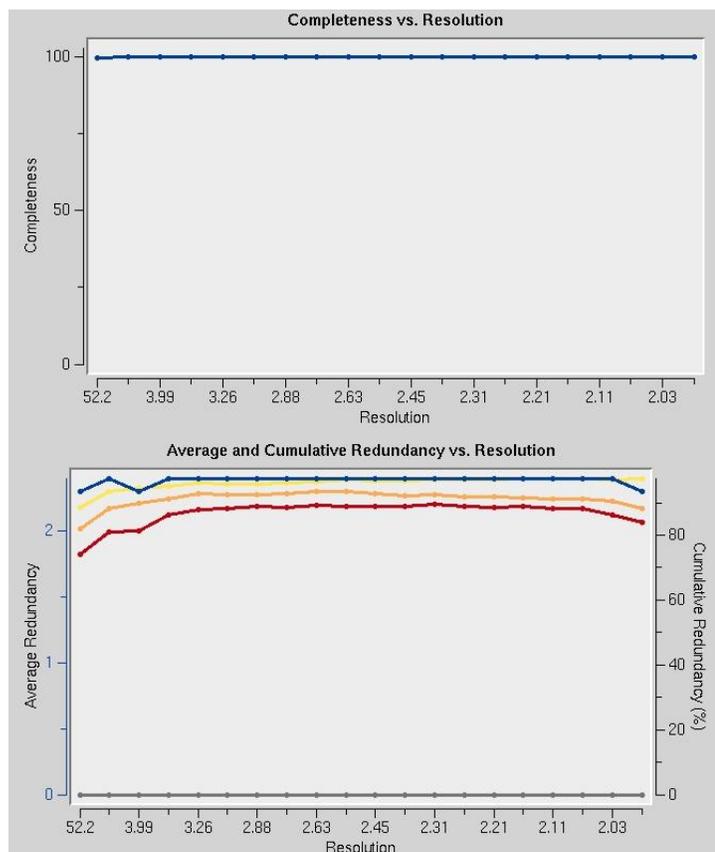


Figure 46. Completeness and Redundancy vs. Resolution plots

Sometimes due to hardware limitations it is not possible to collect a complete data set while using a simple strategy. Then you may simulate several different runs, and their short descriptions will appear in the panel next to the **COMBINE** button (Figure 47). The completeness in low (l) and high (h) resolution shells is shown above the **COMBINE** button. You may select any combination of simulated runs by checking or unchecking the appropriate boxes. By clicking the **COMBINE** button you will obtain statistics and plots for the selected runs. You may remove all simulated runs from the last panel by clicking the **CLEAR** button. A **SETUP DC** button allows exporting a chosen strategy to the **Collect** tab of the main *HKL-2000* window (if your version of *HKL-2000* is configured with the data collection module).



Figure 47. The Combine panel



# Integrating and Scaling the Data

## Step 15: Integration

Integrate the set of frames by clicking on the **INTEGRATE** button in the *Index* tab. Monitor the progress of the integration by examining the  $\chi^2$ , **Cell Constants**, **Crystal Rotations**, **Mosaicity**, and **Distance vs. Frame** plots, as well as the agreement between the predicted reflections and the spots on each image. Ideally, all of these plots will yield horizontal lines, indicating little or no change of the refinement parameters for different 3D Windows in the data set.

**The Integrate button is red. Is this a problem?**

The lettering in the button should be colored green. If it is orange, that means that either the  $\chi^2$  values are suspiciously high and should be checked, or else you have failed to select a lattice other than primitive triclinic. If the lettering is red, this warns you that the  $\chi^2$  values are unacceptably high and should be investigated (Figure 48). You can still continue, but there is probably a problem and you will have to address it eventually.

Integrate	Positional: 1634	X- $\chi^2$ : 0.98	Y- $\chi^2$ : 0.77	(0)
	Partiality: 1420	$\chi^2$ : 1.10		(2)
Integrate	Positional: 754	X- $\chi^2$ : 1.53	Y- $\chi^2$ : 2.47	(0)
	Partiality: 754	$\chi^2$ : 0.66		(0)
Integrate	Positional: 268	X- $\chi^2$ : 13.71	Y- $\chi^2$ : 4.95	(0)
	Partiality: 825	$\chi^2$ : 7.35		(16)

Figure 48. The Integrate buttons colored green, orange, red

**The Crystal Slippage and Special integration buttons.**

The **Special** radio-button opens up the command file used in the integration of the diffraction images. If you want to customize the integration settings passed to *Denzo* (for example, putting in many more “go” statements or some other modification to the standard procedure), this is the place to make those changes (Figure 49).

➡ You can also make these changes under the **Macros** tab in the *HKL-2000* interface.

The **Crystal Slippage** button is a shorthand way of performing the integration in two resolution stages. The first stage is done at low resolution (4 Å), and then is expanded to the final resolution. Normally, integration is done to the highest resolution immediately. However, if the crystal is slipping in position during data collection, the crystal orientation parameters may not be consistent enough from frame to frame to process frames automatically. Using only the low order reflections at first allows a larger radius of convergence, which helps account for slipping crystals.

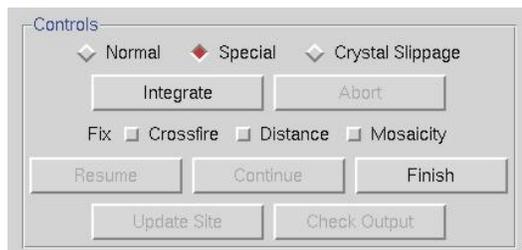


Figure 49. The special integration option

### What are these plots supposed to look like and how do I know if everything is ok?

After clicking, the display will change to a graphical plot of the Integration progress, and show the currently refined mosaicity as well. One of the most useful plots (diagnostically speaking) is  $\chi^2$  vs. **Frame** (Figure 50). Ideally both lines (the positional  $\chi^2$  for X and Y) will be in the range of 1 to 2 over the set of processed frames. All of the plots should be examined, with particular attention paid to detecting bad frames, which will show up as strong outliers in the **Integration Information** plots. Ideally all of the plots –  $\chi^2$ , **Unit Cell Constants**, **Distance**, and **Mosaicity** - should be flat as a function of frame number. You'll also notice that the points for each plot appear in bunches that correspond to the number of frames in the 3D window. If in the  $\chi^2$  vs. **Frame** plot some frames have unreasonably high  $\chi^2$  values, it makes sense to exclude those bad frames from scaling. To do this, split the data set in two or more sets and re-integrate.

Figure 50. The diagnostic plots:  $\chi^2$  vs. Frame and Distance Change vs. Frame

### What is an acceptable range for $\chi^2$ ?

Ideally, the  $\chi^2$  x and y values should be around 1 (i.e. colored green in the Refinement Information panel; see [Figure 33](#)). Sometimes they rise up into the 3 to 4 range. As the integration proceeds, the Image Display window will continually update. If the predicted reflections continue to superimpose on the observed reflections as the integration proceeds, then there is no immediate cause for alarm. However, if you begin refining to very high mosaicity values

(greater than  $2^\circ$  for frozen crystals), or the crystal orientation, cell, or distance parameters are changing dramatically, you should consider re-indexing or choosing another lattice. In some particularly problematic integrations, especially when the calculated unit cell parameters are changing substantially, you may try fixing the mosaicity by clicking the **Fix Mosaicity** button on the **Controls** panel (Figure 49). This is particularly relevant for cases where there are several lattices in the beam at the same time. In the best-case scenario only one of the lattices would dominate the diffraction pattern and be indexed, but sometimes as the crystal rotates other lattices become more prominent and the program may start to confuse them. By restricting the mosaicity you may be able to effectively limit the number of nearby spots that can be included in the integration.

“Houston, we’ve got problems”. The  $\chi^2$  values are getting larger!

Ideally the  $\chi^2$  values will be close to 1.  $\chi^2$  values that go up to 3 and even 4 are sometimes normal and may not indicate a serious problem. However,  $\chi^2$  values that are off the plot that is displayed during integration (i.e.  $>5$ ) probably mean that something is wrong. The most likely culprit, for otherwise “normal” data sets (i.e. where the diffraction pattern on the poorly indexed frames still looks ok, and there weren’t any shutter or goniostat problems) is misindexing due to an incorrect **X Beam** or **Y Beam** value. Experimenting with different beam position values may solve the problem. This is relatively easy to do using the **SET BEAM POSITION** button found in the **Index** panel (described on page 31 above). Crystal idiosyncrasies (such as slight splitting, which may not be apparent from an inspection of the diffraction images) or shutter, spindle or goniostat problems can also lead to high  $\chi^2$  values. Problems with the experimental setup will often manifest themselves suddenly, and a sharp change in the  $\chi^2$  values will pinpoint the frames that have been corrupted. It is always a good idea to verify the experimental setup by collecting a data set from a known, high quality crystal in order to benchmark your setup.

Why do the bars in the mosaicity histogram get thicker or thinner?

The thickness of the bars is proportional to the number of single partial reflections being integrated in the **3D Window**. More reflections will lead to thicker bars.

How long should this take? How do I know if the program is stuck?

When the integration is complete the display will show the message “Integration Complete.” Obviously the speed of the integration depends on both hardware and data-specific factors. Among the hardware ones are: CPU speed, amount of physical memory, and the speed at which the computer can access the raw data from disk (local disks are faster than remote ones). Data-specific factors include the number of observed and predicted reflections on each image, the size of each image file (larger frames take longer to read from disk), and the number of frames in the **3D Window** (integrating 5 frames at a time will take about 5 times as long as integrating a single frame). For most modern computers and typical macromolecular data sets, this process is usually takes 1-3 minutes at most. In my experience, if several minutes go by with no additions to the plots, something is probably amiss. The most likely culprit is slowness in reading the image files from disk, particularly if your frames are on a remote file server.

Waiting for image frames.

A very nice feature of *HKL-2000* is the program’s ability to wait for image frames during processing. This is indicated by the message “Waiting for Frames” which appears above the **Integration Information** window during integration (Figure 51). The program knows to wait for frames based on the value you entered in the **Number of Frames** entry of the **Edit Set(s)** dialog on the **Data** panel. For example, let’s say you are collecting 60 images for a data set, but only 10 have been collected so far and the others are on the way (e.g. are still being collected or read from the detector, being copied from a file server or CD/DVD, coming over FTP, etc.). When you first set up the data files, only 10 frames will be seen by the program. However, if you enter 60 for **Number of Frames**, the program will know that eventually 60 frames will constitute the complete set of diffraction images. You can start the indexing and integration, and these will proceed through the first 10 (or however many you have) frames currently on disk. When those have been integrated, the message “Waiting for Frames” will appear. At this point, you don’t have to do anything. As more frames appear, they will be automatically read and integrated. This allows you to monitor the progress of your data collection as it occurs. It also allows you to use your data collection time (e.g. at the synchrotron) more efficiently, by working on another task while you wait for integration to be completed.

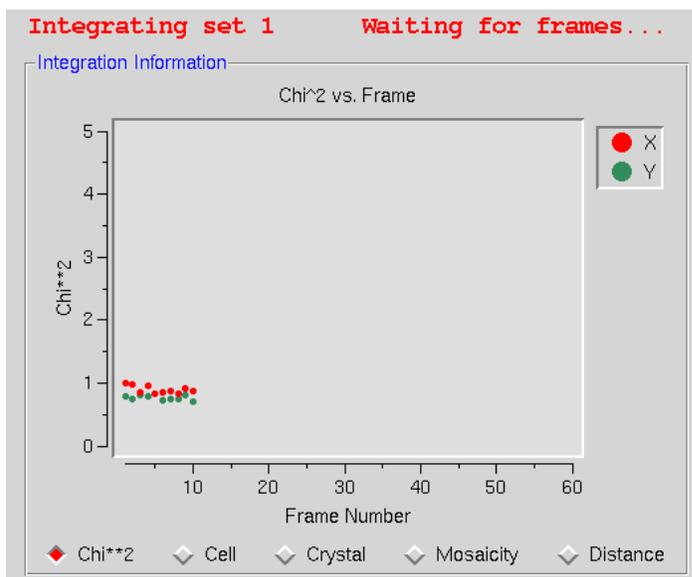


Figure 51. The program is waiting for new frames during processing the data

### Once the integration is complete, should I update the site?

Ordinarily you would not update the site unless:

- 1) you have the password to do so,
- 2) the refined values of the site parameters (**X and Y Beam Positions, Crossfire, Y Scale**) are substantially different from the original site values, and
- 3) the values you've refined are reliable.

Site parameters determined from weak, poorly diffracting crystals may not be much of an improvement (if at all!) from the default values. Generally you want to use a superior-diffracting crystal to determine updated site parameter values.

### How do I stop and start over?

To stop the integration hit **ABORT** and you will get a message box with the question: "Which crystal orientation and cell parameters would you like to keep?" You can choose between three options: "original values" (values as read from the header of data files), "indexed values" (values after the last auto-indexing round), and "current values." Now you can reindex if you'd like.

## Step 16: Scaling

Click on the tab to open the *Scaling* panel, which provides the interface to the *Scalepack* program. Select the appropriate **Global Refinement** choice (previously called **Postrefinement** in the old *Scalepack*): **Non-Slipping Crystal and Perfect Goniostat** or **Non-Slipping Crystal and Imperfect Goniostat** or **Small Slippage and Imperfect Goniostat** or **Custom Postrefinement** or **No Postrefinement** (Figure 52). Click on **Scale Sets** in the **Controls** panel to start the first round of scaling (see Figure 55).

### Why do I want to do postrefinement? Which one should I choose?

Due to the correlations between crystal and detector parameters the values of the unit cell parameters refined from a single image may be quite imprecise. This lack of precision is of little significance to the process of integration, as long as the predicted positions are on target. There is no contradiction here, because at some crystal and detector orientations the positions of the reflections may only weakly depend on the value of a particular crystal parameter. At the end of the data reduction process one would wish to get precise unit cell values. This is done in a procedure referred to as **Global Refinement** or **Postrefinement**. The implementation of this method in *Scalepack* allows for

separate refinement of the orientation of each image, but with the same unit cell value for the whole data set. In each batch of data (a batch is typically one image) different unit cell parameters may be poorly determined. However, in a typical data set there are enough orientations to determine precisely all unit cell lengths and angles.

**Global Refinement** is also more precise than the processing of a single image in the determination of crystal mosaicity and the orientation of each image. This is particularly important when deciding which reflections are fully recorded and which are partially recorded, which is why you should routinely do the **Global Refinement**.

Within the global refinement, the following parameters can be refined: the unit cell lengths and angles, the crystal orientation parameters (`rotx`, `roty`, and `rotz`), and the mosaicity. These can either be refined over the “Crystal” (generally this means over the whole data set), or over the “Batch” (generally this means each image). You can see these listed explicitly if you click on the 4<sup>th</sup> choice under the **Global Refinement** menu: **Custom Postrefinement**.



Figure 52. The global refinement options

The defaults are set so that the unit cell parameters and the mosaicity are refined to be the same over the entire data set. This makes sense for most data collections these days where the crystal is frozen and the entire data set is collected from a single crystal. However, if something happened to the crystal in the middle of the data collection (e.g. a temperature change that caused the unit cell constants to change, or the crystal began to decay severely), then one would want to refine these for each batch rather than over the entire data set.

Similarly, the defaults for the crystal orientation parameters in this case are refined for each batch, or frame, rather than over the entire data set. This allows you to spot any problems with the goniostat. If you do refine this for each batch and it doesn't change much, then on subsequent rounds of refinement you can refine the orientation once for the entire data set (crystal).

In general it is best to assume that the crystal is slipping and that the goniostat is imperfect (**Small Slippage Imperfect Goniostat**) when choosing **Global Refinement** options, and only restrict the refinement with the other options if the data warrants it.

#### Scaling Options: What are Scale Restrain, B Restrain, Fix B, Ignore Overloads, etc.?

These options affect how scaling of the integrated diffraction intensities is performed. The default options are: **Scale Restrain** 0.01, **B restrain** 0.1, **Write Rejection File**, **Ignore Overloads**, **Number of Zones** 10, **Error Scale Factor** 1.3, **Default Scale** 10, and **Resolution Limits** as determined from the \*.x file headers (Figure 53). What follows is a discussion of each of the options listed in this panel.

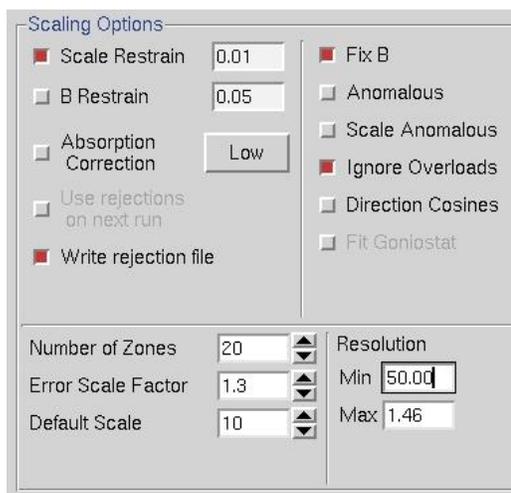


Figure 53. The Scaling Options panel

**SCALE RESTRAIN** can be used to restrain scale factor differences from consecutive films or batches. The value entered in the box represents the amount you will allow the scale factors to differ from consecutive frames or batches. It adds a factor of  $(\text{scale1} - \text{scale2})^2 / (\text{scale restrain})^2$  to the target function minimized in scaling. This only applies to batches between which you add partials. The value should roughly represent the expected relative change in scale factors between adjacent frames.

For very thin frames, or frames where most of the reflections are partially recorded (yellow, as opposed to green, spots on the image during integration, see [Figure 28](#)), or in sectors whose width is less than the mosaicity, or from data which consists of only a few frames, **SCALE RESTRAIN** is almost obligatory. In these cases, there may not be enough intersections between frames to get accurate scale and B factors. Indeed, what you may see is both scale and B factors ranging all over the place. If things get really bad, the program may crash due to floating point arithmetic exceptions when taking the exponent of unreasonable B factors. This does not mean that the data is unusable; it simply means that the scale and B factors must be restrained. Note that the restraints only apply to the frames over which you are adding up partially recorded reflections.

- It is not correct to try to “bin” the individual frames into larger batches to try to overcome the problem of few intersections between frames. This is because you then lose the ability to **ADD PARTIALS** between the new “bins.” You can, however, overcome the problem in another way, by adding the macro **number of iterations 0** to the **During Scaling** box on the **Macros** panel. In this case, the data will not be scaled, but simply merged. Obviously, this has its drawbacks.

**B RESTRAIN** can be used to restrain B factor differences from consecutive frames or batches. The value which follows the flag represents the amount in  $\text{Å}^2$  you will allow the B factors to differ from consecutive frames or batches.

**ABSORPTION CORRECTION** applies additional parameters that correct for variations in the degree of X-ray absorbance by the crystal in different orientations. For example, more X-rays are absorbed when the longer dimensions of the crystal are parallel to the incident beam. In essence, this procedure adjusts for the fact that most crystals are not perfect spheres! The absorption correction algorithm parameterizes this effect by an expansion of real spherical harmonics terms (the **Low** or **High** option controls how many terms are included in the correction). This procedure, along with a number of other scaling corrections applied by *Scalepack*, is described in more detail in Otwinowski et al. Acta Cryst A59: 228-234 (2003).

In the majority of cases, there is likely to be improvement of the quality of the scaling results (and no harm!) in including at least **Low** absorption corrections. The only serious drawback is in the speed of the scaling process, as *Scalepack* requires additional processing time to calculate the spherical harmonic expansions.

**USE AUTO CORRECTIONS** applies an algorithm to automatically adjust the error model.

**USE REJECTIONS ON NEXT RUN** means that the reflections that were written to the file `reject` are not used in subsequent scaling runs. Unless this is explicitly checked reflections will only be flagged for rejection, not actually rejected. Generally one uses this on the second, and all subsequent runs of scaling, unless an anomalous signal is expected.

**WRITE REJECTION FILE** tells the program to create a list of reflections, stored in the file called `reject` that meet the criteria for rejection. They are applied, i.e. rejected, when the **USE REJECTIONS ON NEXT RUN** button is checked.

**FIX B.** This flag tells the program not to fit B factors at all. Usually it is combined with the input of the B factors you want to apply but do not wish to refine anymore, or it is used for frozen crystals where you do not expect significant decay. This is in contrast to the default procedure, where the B factors are fit only after the convergence of the scaling. In the default procedure, if scaling does not converge in 20 (default) cycles of refinement, B factors will not be fitted.

**ANOMALOUS.** Flag for keeping Bijvoets (I+ and I-) separate in the output file. If the **ANOMALOUS** flag is on, anomalous pairs are considered equivalent when calculating scale and B factors and when computing statistics, but are merged separately and output as I+ and I- for each reflection. See the next section for a discussion of how to treat data that contains an anomalous signal.

**SCALE ANOMALOUS.** This is the flag for keeping Bijvoets (I+ and I-) separate both in scaling and in the output file. If the **SCALE ANOMALOUS** flag is on, anomalous pairs are considered non-equivalent when calculating scale and B factors and when computing statistics, and are merged separately and output as I+ and I- for each reflection.

This is a dangerous option because scaling may be unstable due to the reduced number of intersections between images. The danger is much greater in low symmetry space groups. **SCALE ANOMALOUS** will always reduce  $R_{\text{-merge}}$ , even in the absence of an anomalous signal, because of the reduced redundancy. However,  $\chi^2$  values will not be affected in the absence of an anomalous signal.

**IGNORE OVERLOADS.** Useful if you collect data at low and high exposures and is useful to ignore the saturated reflections at the high exposures. Applies to data read in after this command. If it is not checked, fitted profiles with some pixels missing (typically due to overload) are included in the scaling. Note that for summed profiles this does not apply, because only profile fitting can estimate the value of the overloaded pixels.

**DIRECTION COSINES** produce information that can be read by an outside absorption correction program, such as *Shelx*. Generally, this option is not needed as *HKL-2000* is capable of applying absorption corrections intrinsically (see above).

**FIT GONIOSTAT** gives the goniostat misalignment angles. It should really only be used by beamline staff or the maintainers of X-ray diffractometers.

**NUMBER OF ZONES.** Number of **resolution shells** the data is divided into for the basis of calculating the statistics summarized on the plots on the *Scale* panel. This input is required and must match the number of zones specified under the **ESTIMATED ERROR** keyword if you apply that on the *Macros* panel (though the same corrections can be applied using the **ADJUST ERROR MODEL** button). Handy tip: it's nice to set up the **NUMBER OF ZONES** to equal the number of zones used by CNS for the output of refinement statistics by shell, for when you get around to publishing your data and refinement statistics together in a paper. The default value is 20, but for higher resolution data a larger value allows to you detect small crystals of ice or salt not always easily visible on the diffraction image.

**ERROR SCALE FACTOR.** This is a single multiplicative factor which is applied to the input  $\sigma_1$ . This should be adjusted so the normal  $\chi^2$  (goodness of fit) value that is printed in the final table of the output comes close to 1. By default the input errors are used (**ERROR SCALE FACTOR** = 1.3). Reasonable values of the error scale factor are between 1 and 2. If you have to resort to higher values (i.e. >2) in order to get your overall  $\chi^2$  values to 1 then you should be suspicious

of some problem with your crystal, indexing, space group/lattice assignment, or detector or goniostat. That doesn't mean that your data is useless, but it does mean that there are problems with it.

**DEFAULT SCALE.** This is the overall scale factor used in the absence of an initial scale factor. This is useful if the data are too intense, which is sometimes the case with small molecules. It will reduce the output intensities by the factor entered.

**RESOLUTION.** This is the maximum and minimum resolution (or **d-spacing**) of the reflections used in this run. The defaults is the full resolution range found in the input integration data. Note that if you integrated your raw data correctly, the maximum resolution should slightly exceed the resolution at which your signal drops below  $2\sigma$ . Therefore, in subsequent rounds of scaling it is usually important to enter a new value for the maximum resolution over which scaling is performed. Here's an example: You collected data and integrated it over the resolution range of 50 to 2.2 Å. The first round of scaling was performed over this resolution range and you noted that the average  $I/\sigma$  dropped below 2 at 2.3 Å resolution. Therefore, you should do all subsequent scaling with a minimum resolution of 50 Å and a maximum resolution of 2.3 Å.

### Should I change the resolution now?

Don't change the resolution until after the first round of scaling, where you have already properly adjusted the error scale factor and rejected outliers. It is only at this point that you can determine where the signal drops below  $2\sigma$ . This is your high resolution limit. Once this has been determined, you should delete the `reject` file and start your scaling over again, this time with the proper resolution limits.

### What are the three *Scalepack* output Files? Can I rename them? Where do they go?

The three files and their default names are `output.sca`, `scale.out`, and `scale.log` (Figure 54).

`output.sca` is the file that contains your scaled reflections. The header lists the unit cell and space group, and the five columns correspond to  $h$ ,  $k$  and  $l$ ,  $I$ , and  $\sigma(I)$ . You may rename this file as you wish, although the convention is to retain the `*.sca` suffix to show that it was created by *Scalepack*. Many other programs that read *Scalepack* files expect this suffix.

`scale.log` is a log of the scaling operations. It is often quite long once the list of rejected reflections is included. You may rename this file as you wish also.

`scale.out` (not visible in an **Output File** window) is a file that contains the data for the plots seen in the user interface. As this file is only used internally by *HKL-2000*, you shouldn't rename it.

The files are all written to the output data directory you specified in the **Main** page. However, if you fill in the path explicitly in the box provided you can write the files anywhere you'd like.

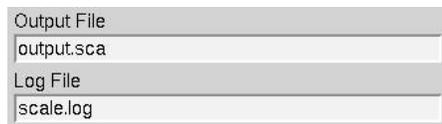


Figure 54. The *Scalepack* output files

### Someone told me my statistics will be better if I decrease the number of zones. Should I do this?

No. Combining those resolution shells where  $R_{merge}$  is rapidly changing into a single zone yields misleading statistics. In this case, the shell will be dominated by the strong data at the low resolution end of the zone and give the impression that the high resolution limit of the zone has better statistics than it really does. For example, if you combined all your data into a single zone, the  $R_{merge}$  in the *final* shell would be pretty good (i.e. equal to the overall  $R_{merge}$ ), when in fact it was substantially worse. It is more sensible to divide your zones into equal volumes and have enough of them so that you can accurately monitor the wholly expected decrease in the quality of the data as resolution increases.

### What is the error model and should I adjust it?

The error model is the estimate of the systematic error for each of the resolution shells. There will be exactly the same number of error estimates here as there are **NUMBER OF ZONES**.

The error estimates do not all have to be the same. The estimated error applies to the data which are read **after** this keyword, so you can apply different error scale factors to subsequent batches by repeating this input with different values. This is an important point if you enter data from a previous *Scalepack* output that does not need its  $\sigma$  to be increased.

The error estimates should be approximately equal to the R-factors for the resolution shells **where statistical errors are small**, namely the lower resolution shells where the data is strong. This is a crude estimate of the systematic error (to be multiplied by  $I$ ), and is usually invariant with resolution. The default is 0.03 (i.e. 3%) for all zones. Examine the difference between the total error and the statistical error in the final table of statistics (this can be viewed by clicking the **SHOW LOG FILE** button and scrolling to the bottom). The difference between these numbers tells you what contribution statistical error makes to the total error ( $\sigma$ ). If the difference is small, then reasonable changes in the estimated error values following the will not help your  $\chi^2$  much. This is because the estimated errors represent your guess and/or knowledge of the contribution of systematic error to the total error, and a small difference indicates that systematic error is not contributing much. If the difference between total and statistical error is significant, and the  $\chi^2$  are far from 1, then consider adjusting the estimated error values in the affected resolution shells.

### How many cycles should I refine for and how do I know how to set this?

The default for the **Number of Global Refinement Cycles** is 10. You can change this by changing the number in the box under **Global Refinement** (see [Figure 52](#)).

### What's the rejection file and why am I writing it?

The rejection file is called `reject` and contains the indices of the reflections to be rejected on subsequent rounds of scaling if the **Use Rejections on Next Run** button is set.

## Step 17: Output plots and Error Scale Factor

When the initial round of scaling is finished, examine the output plots and adjust the **ERROR SCALE FACTOR** to bring the overall  $\chi^2$  down to around 1 (the overall  $\chi^2$  is also explicitly specified at the end of the log file). Delete the `reject` file by clicking on the button in the lower **Controls** panel, and then click **SCALE SETS** again (Figure 55). Repeat these steps until the overall  $\chi^2$  is acceptable then proceed to subsequent rounds of scaling by activating **Use Rejections on Next Run** and scale again. Repeat the cycle of scaling and rejecting until the number of new rejections has declined to just a few. At this point the scaling is completed.

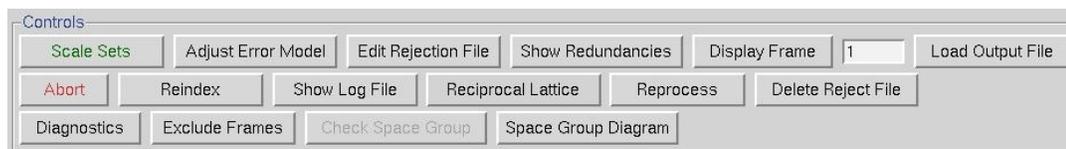


Figure 55. The scaling Controls panel

After integration some frames may appear to have strong outliers in the integration information plots like unacceptably high  $\chi^2$  values – i.e., far above the mean value. You can exclude those frames from scaling by clicking on the **EXCLUDE FRAMES** button. A table with the list of frames will appear (Figure 56). Select a single frame by clicking on its box or on a point on the plot. Select a range of frames by checking **Multiple Select** and clicking twice, once on the first frame and once on the last. You may have to move to a different page with the Previous or Next buttons to see all frames. Frames marked in red will be excluded when close the table and rescale.

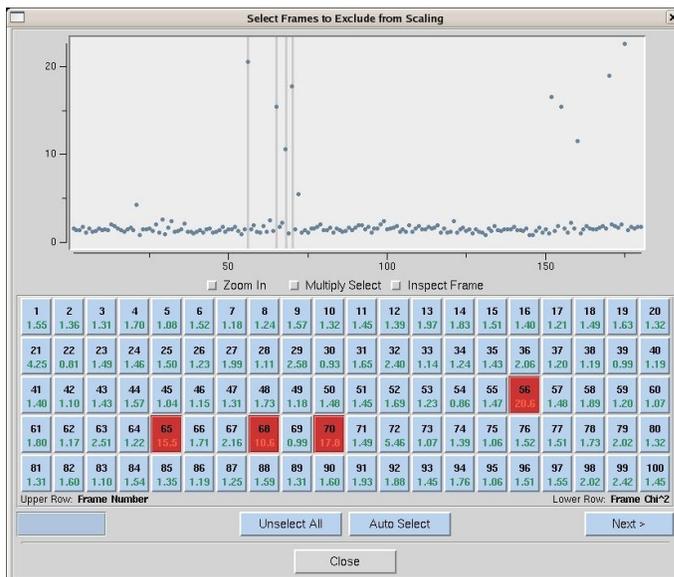


Figure 56. Frames selected to be excluded from scaling are marked red

### What is the Error Scale Factor and which way should I adjust it?

The **ERROR SCALE FACTOR** is a single multiplicative factor which is applied to the input  $\sigma_1$ . This should be adjusted so the normal  $\chi^2$  (goodness of fit) value that is printed in the final table of the output comes close to 1. By default the input errors are used (**ERROR SCALE FACTOR** = 1.3). The default value can be adjusted upwards if the overall  $\chi^2$  is greater than 1.

### What are reasonable values for the error scale factor and what if I need really high ones?

Reasonable values are between 1 and 2. The default is 1.3. If you need to use values much greater than 2, there is likely a problem.

### What should I be looking for in the charts that show up after scaling?

Ten charts are generated by *HKL-2000* after scaling is completed. These are **Scale and B vs. Frame**, **Completeness vs. Resolution**, **Chi<sup>2</sup> and R-Factor vs. Frame**, **Chi<sup>2</sup> and R-Factor vs. Resolution**, **I/Sigma vs. Resolution**, **I vs. Resolution**, **Chi<sup>2</sup> and R-Factor vs. Average I**, **Average and Cumulative Redundancy vs. Resolution**, **Mosaicity vs. Frame**, and **Low Resolution Completeness vs. Resolution**. Each of the plots has an explanation button underneath, which provides more information about the data plotted and the meanings of each color. Note that some plots have two vertical columns; in most cases the data displayed are color coded to match its axis. For example, blue data are plotted with respect to the blue axis and red data to the red axis. Some plots also have a button to toggle between linear and log representations.

The first chart to examine is **I/Sigma vs. Resolution**, which shows the ratio of the intensity to the error of the intensity, i.e.  $I/\sigma$  as a function of resolution (Figure 57). From this you can find the  $2\sigma$  cutoff and adjust your resolution limits accordingly.

The second chart to examine is **Completeness vs. Resolution**, which describes the completeness of the data for different resolution shells. The blue line represents all reflections, the purple reflections with  $I/\sigma$  greater than 3, the red reflections with  $I/\sigma$  greater than 5, the orange reflections with  $I/\sigma$  greater than 10, and the yellow reflections with  $I/\sigma$  greater than 20 (Figure 58). At this point you can determine if the **Strategy** you planned above was effective or if you need to collect additional data.

You should also look closely at the **Scale and B vs. Frame** (Figure 59) to see if there are significant variations in intensity and B-factor of frames across the beamset, which could indicate the crystal rotating out of the beam, radiation damage, etc.

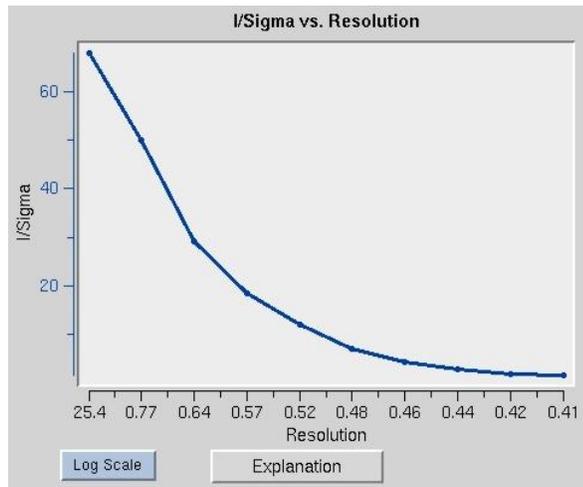


Figure 57. The output plot: *I/Sigma vs. Resolution*

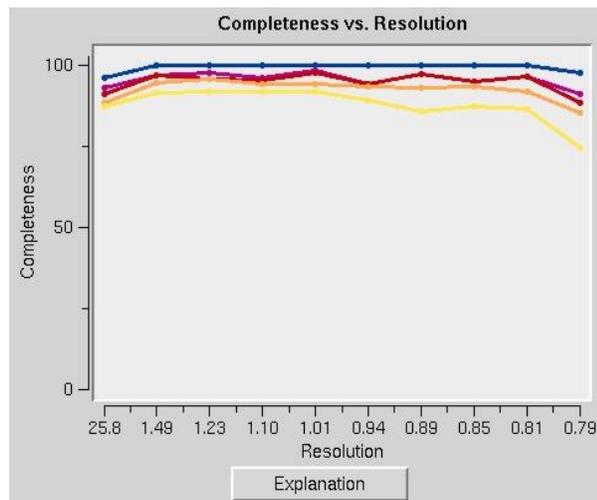


Figure 58. The output plot: *Completeness vs. Resolution*



Figure 59. The output plot: Scale and B vs. Frame

### How many reflections should I be rejecting?

Ideally you would not reject any reflections. However, in practice even excellent data sets have a few outliers that need to be rejected. Generally if fewer than 0.5% of the reflections are rejected this is considered normal. Rejections significantly greater than 1% usually indicate some problem with the indexing, integration, space group assignment, or the data itself (e.g. spindle or shutter problems).

### What's the difference between a reflection being marked for rejection and actually rejected?

The table that appears at the top of the page after each round of scaling has two categories: **Reflections Marked for Rejection**, and **Reflections Rejected** (Figure 60). The **Reflections Marked for Rejection** are the percentage of all reflections that have been written to the `reject` file. The **Reflections Rejected** represent the percentage of reflections, which have actually been omitted from the data in the latest round of scaling. Rejection of reflections is an iterative process that should converge after only a few rounds of rejection and refinement. For example, on the first round of scaling, 100 reflections may be flagged for rejection and written to the `reject` file. If **Use Rejections on Next Run** is checked, then the reflections in the `reject` file will be removed from the set of reflection data on the next round of scaling. Having done this, other reflections may now be flagged as outliers, and written to the end of the `reject` file. Perhaps there are 20 more this time. Repeating this procedure should converge to a point when no more reflections are being rejected (and thus the two percentages are identical).

Crystal - Global Refinement					
Space Group	P212121				
Resolution	50.00 - 0.70				
a	5.273 ± 0.001	α	90.000 ±		
b	7.021 ± 0.001	β	90.000 ±		
c	24.231 ± 0.001	γ	90.000 ±		
Cell Volume	897.2				
Mosaicity Range	0.45 - 0.49				
Global Statistics					
Total Number of Reflections	9249				
Number of Reflections Marked for Rejection	25				
Percentage of Reflections Marked for Rejection	0.27				
Percentage of Reflections Rejected	0.27				
Total Linear R-Factor	0.022				

Figure 60. Info table after first round of scaling

### How do I treat data that contains an anomalous signal?

Detecting an anomalous signal is very easy. In the first step the data are scaled and postrefined normally with the **ANOMALOUS** option selected (*not* the **SCALE ANOMALOUS** option!) This tells the program to output a \*.sca file where the I+ and I- reflections are separated. In this case *Scalepack* will treat the I+ data and the I- data as two separate measurements within a data set and the statistics that result from merging the two will reflect the differences between the I+ reflections and the I- reflections. Notice that in *HKL-2000* there is no need to go through a lot of jiffies, which separate the I+ and I- data and reformat them, etc. Obviously, for a centric reflection there is no I-, so the merging statistics will only reflect the non-centric reflections. You can tell what percentage of your data is being used to calculate the merging statistics by examining the redundancy table near the end of the log file. Under the column of redundancy > 2 you will find out what percentage of the data is being compared. Since you only have I+ and I-, you will never have a redundancy of more than 2.

The presence of an anomalous signal is detected by examining the  $\chi^2$  values on the **Chi<sup>2</sup> and R-Factor vs. Resolution** plot (Figure 61). The two  $\chi^2$  curves show the  $\chi^2$  values for each resolution shell, calculated for merged and unmerged Friedel pairs, in orange and blue respectively. Assuming that the errors scale factors were reasonable, and there is no useful anomalous signal in your data, both curves showing the  $\chi^2$  resolution dependence should be flat and average around 1 when scaling with either merged or unmerged Friedel pairs. On the other hand, if  $\chi^2 > 1$  and you see a clear resolution dependence of the  $\chi^2$  for scaling with merged Friedel pairs (i.e. the  $\chi^2$  increases as the resolution decreases), there is a strong indication of the presence of an anomalous signal. The resolution dependence allows you to determine where to cut off your resolution for the purposes of calculating an anomalous difference Patterson map with an optimal signal to noise. Please note, however, that this analysis assumes that the error model is reasonable and gives you a  $\chi^2$  close to 1 when the anomalous signal is not present.

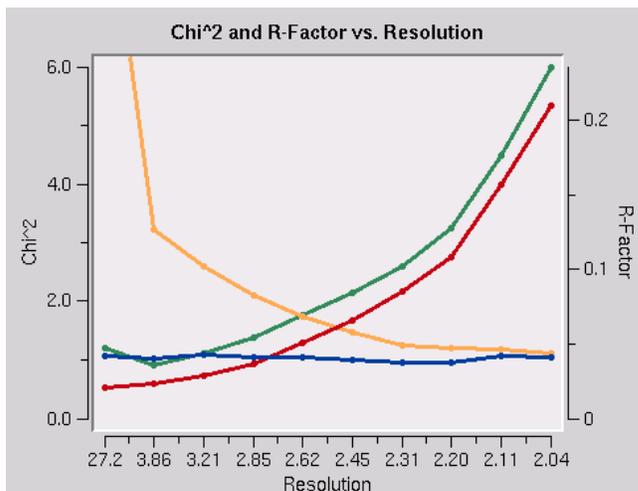


Figure 61. Anomalous signal detection

- You can only use **SCALE ANOMALOUS** when you have enough redundancy to treat the F+ and F- completely independently.

### Space Group Identification: How do I do this?

*Scalepack* can be used to determine the space group of your crystal. What follows is a description of how you would continue from the lattice type given by *Denzo* to determine your space group. This whole analysis of course only applies to enantiomorphic compounds such as proteins. It does not apply to small molecules, necessarily, which may crystallize in centrosymmetric space groups. If you expect a centrosymmetric space group, you should use any space group which is a subgroup of the Laue class to which your crystal belongs. You also need enough data for this analysis to work so that you can see systematic absences.

To determine your space group, follow these steps:

1. Determine the lattice type in *Denzo*.
2. Scale by the primary space group in *Scalepack*. The primary space groups are the first space groups in each Bravais lattice type in the table that follows this discussion. In the absence of lattice pseudosymmetries (e.g. monoclinic with  $\beta \approx 90^\circ$ ) the primary space group will not incorrectly relate symmetry related reflections. Note the  $\chi^2$  statistics. Now try a higher symmetry space group (next down the list) and repeat the scaling, keeping everything else the same. If the  $\chi^2$  is about the same, then you know that this is OK, and you can continue. If the  $\chi^2$  are much worse, then you know that this is the wrong space group, and the previous choice was your space group. The exception is primitive hexagonal, where you should try  $P6_1$  after failing  $P3_121$  and  $P3_112$ .
3. Examine the bottom of the log file (the **SHOW LOG FILE** button) or the simulated reciprocal lattice picture (the **RECIPROCAL LATTICE** button) for systematic absences. If this was the correct space group, all of the systemic absence reflections should be absent (or with very small values). Compare this list with the listing of reflection conditions by each of the candidate space groups. The set of absences seen in your data which corresponds to the absences characteristic of the listed space groups identifies your space group or pair of space groups. Note that you cannot do any better than this (i.e. get the handedness of screw axes) without phase information.
4. If it turns out that your space group is orthorhombic and contains one or two screw axes, you may need to reindex to align the screw axes with the standard definition. If you have one screw axis, your space

group is  $P222_1$ , with the screw axis along  $c$ . If you have two screw axes, then your space group is  $P2_12_12$ , with the screw axes along  $a$  and  $b$ . If the *Denzo* indexing is not the same as these, then you should reindex using the **REINDEX** button.

- So far, this is the way to index according to the conventions of the International Tables. If you prefer to use a private convention, you may have to work out own transformations. One such transformation has been provided in the case of space groups  $P2$  and  $P2_1$ .

Bravais Lattice	Primary assigned Space Groups	Candidates	Reflection Conditions along screw axes
Primitive Cubic	P213	195 P23	
	P4132	198 P213	(2n,0,0)
		207 P432	
		208 P4232	(2n,0,0)
		212 P4332	(4n,0,0)*
	213 P4132	(4n,0,0)*	
I Centered Cubic	I213	197 I23	*
		199 I213	*
	I4132	211 I432	
		214 I4132	(4n,0,0)
F Centered Cubic	F23	196 F23	
	F4132	209 F432	
		210 F4132	(2n,0,0)
Primitive Rhombohedral	R3	146 R3	
	R32	155 R32	
Primitive Hexagonal	P31	143 P3	
		144 P31	(0,0,3n)*
		145 P32	(0,0,3n)*
	P3112	149 P312	
		151 P3112	(0,0,3n)*
		153 P3212	(0,0,3n)*
		150 P321	
	P3121	152 P3121	(0,0,3n)*
		154 P3221	(0,0,3n)*
		168 P6	
	P61	169 P61	(0,0,6n)*
		170 P65	(0,0,6n)*
		171 P62	(0,0,3n)**
		172 P64	(0,0,3n)**
		173 P63	(0,0,2n)
	P6122	177 P622	
		178 P6122	(0,0,6n)*
179 P6522		(0,0,6n)*	
180 P6222		(0,0,3n)**	
181 P6422		(0,0,3n)**	
	182 P6322	(0,0,2n)	

<b>Primitive Tetragonal</b>	<b>P41</b>	<b>75</b>	<b>P4</b>	
		<b>76</b>	<b>P41</b>	<b>(0,0,4n)*</b>
		<b>77</b>	<b>P42</b>	<b>(0,0,2n)</b>
		<b>78</b>	<b>P43</b>	<b>(0,0,4n)*</b>
	<b>P41212</b>	<b>89</b>	<b>P422</b>	
		<b>90</b>	<b>P4212</b>	<b>(0,2n,0)</b>
		<b>91</b>	<b>P4122</b>	<b>(0,0,4n)*</b>
		<b>95</b>	<b>P4322</b>	<b>(0,0,4n)*</b>
		<b>93</b>	<b>P4222</b>	<b>(0,0,2n)</b>
		<b>94</b>	<b>P42212</b>	<b>(0,0,2n),(0,2n,0)</b>
		<b>92</b>	<b>P41212</b>	<b>(0,0,4n),(0,2n,0)**</b>
		<b>96</b>	<b>P43212</b>	<b>(0,0,4n),(0,2n,0)**</b>
<b>I Centered Tetragonal</b>	<b>I41</b>	<b>79</b>	<b>I4</b>	
		<b>80</b>	<b>I41</b>	<b>(0,0,4n)</b>
	<b>I4122</b>	<b>97</b>	<b>I422</b>	
		<b>98</b>	<b>I4122</b>	<b>(0,0,4n)</b>
<b>Primitive Orthorhombic</b>	<b>P212121</b>	<b>16</b>	<b>P222</b>	
		<b>17</b>	<b>P2221</b>	<b>(0,0,2n)</b>
		<b>18</b>	<b>P21212</b>	<b>(2n,0,0),(0,2n,0)</b>
		<b>19</b>	<b>P212121</b>	<b>(2n,0,0),(0,2n,0), (0,0,2n)</b>
<b>C Centered Orthorhombic</b>	<b>C2221</b>	<b>20</b>	<b>C2221</b>	<b>(0,0,2n)</b>
		<b>21</b>	<b>C222</b>	
<b>I Centered Orthorhombic</b>	<b>I212121</b>	<b>23</b>	<b>I222</b>	*
		<b>24</b>	<b>I212121</b>	*
<b>F Centered Orthorhombic</b>	<b>F222</b>	<b>22</b>	<b>F222</b>	
<b>Primitive Monoclinic</b>	<b>P21</b>	<b>3</b>	<b>P2</b>	
		<b>4</b>	<b>P21</b>	<b>(0,2n,0)</b>
<b>C Centered Monoclinic</b>	<b>C2</b>	<b>5</b>	<b>C2</b>	
<b>Primitive Triclinic</b>	<b>P1</b>	<b>1</b>	<b>P1</b>	

- ➡ Note that for the pairs of similar candidate space groups followed by the \* (or \*\*) symbol, scaling and merging of diffraction intensities cannot resolve which member of the possible pair of space groups your crystal form belongs to.

### How do I start scaling from \* .x files without reintegrating diffraction images?

This is easy. In the **Data** panel, set up the **Output Data Dir** to be the directory where your \* .x files are located. It is not necessary to set the **Raw Data Dir**. Next, click the **Scale Sets Only** button followed by **LOAD DATA SETS**. You will see a list of \* .x files in the little dialog box. Select the set you want to scale, followed by **OK**. If you want to scale additional sets of \* .x files, go through and find these too and add them to the list. Make sure the **SCALE** button is selected for each set you want to scale together. For example, if you want to scale two sets of \* .x files together (say, from two different crystals, or from high and low resolution passes of data collection, or from a native and a derivative, etc.) make sure that **SCALE** is clicked for each one. If you don't want them to be scaled together, then make sure that **SCALE** is not clicked. You don't have to worry about the **Image Display** or the **Experiment Geometry**, since this was done when you first generated the \* .x files.

Now go to the **Scaling** page. You should see the set or sets of \*.x files listed in the **Pending Sets** list. These are the frames that will all be scaled together. Delete the reject file by clicking on the **DELETE REJECT FILE** button. Then scale away!

### What is reindexing, and how do I do it?

Reindexing involves reassigning indices from one unit cell axis to another. This becomes an important issue when comparing two or more data sets that were collected and processed independently. This is because *Denzo*, when confronted with a choice of more than one possible indexing convention, makes a random choice. This is no problem, except that if it makes a different choice for a second data set, the two will not be compatible for scaling together without first reindexing. One cannot distinguish non-equivalent alternatives without scaling the data, which is why this is not done in *Denzo* (i.e. at the indexing or integration step). You can tell if you need to reindex one of two data set if the  $\chi^2$  values upon merging the two are very high (e.g. 50). This makes sense when you consider that scaling two or more data sets together involves comparing reflections with the same *hkl* or index. If the two indexing schemes are equivalent but not identical, chaos will result.

No reindexing, no new auto-indexing, and **nothing** except changing the sign of **Y SCALE** in *Denzo* can change the sign of the anomalous signal. At the moment reindexing does not work in *HKL-2000*. To reindex you'll have to follow the scenarios listed in the *HKL* manual.

### Statistics and Scalepack.

The quality of X-ray data is initially assessed by statistics. In **small molecule** crystallography there is almost always a large excess of strong data, so this allows the crystallographer to discard a substantial amount of suspect data and still accurately determine a structure. Compared to small molecules, however, **proteins** diffract poorly. Moreover, important phase information comes from weak differences and we must be sure these differences do not arise from the noise caused by the limitations of the X-ray exposure and detection apparatus. As a result, we cannot simply throw away or statistically downweight marginal data without first making a sophisticated judgment about which data are good and which are bad.

To accurately describe the structure of a protein molecule, we often need higher resolution data than the crystal provides. That is life. One of the main judgments the crystallographer makes in assessing the quality of his data is thus the **resolution** to which his crystal diffracts. In making this judgment, we wish to use the statistical criteria which are most discriminatory and which are the least subjective. In practice, there are two ways of assessing the high resolution limit of diffraction. The first is **the ratio of the intensity to the error of the intensity**, i.e.  $I/\sigma$ . The second way, which is traditional, but inferior, is the agreement between symmetry related reflections, i.e.  $R_{merge}$ .

From a statistical point of view,  $I/\sigma$  is a *superior criterion*, for two reasons. First, it defines a resolution "limit" since by definition  $I/\sigma$  is the *signal to noise* of your measurements. In contrast, the  $R_{merge}$  is an unweighted statistic that does not take into account the measurement error.

Second, the  $\sigma$  assigned to each intensity derives its validity from the  $\chi^2$ 's, which represent the *weighted* ratio of the difference between the observed  $I$  and the average value  $\langle I \rangle$  squared, divided by the square of the error model and multiplied by a factor correcting for the correlation between  $I$  and  $\langle I \rangle$ . Since it depends on an explicit declaration of the expected error in the measurement, the user of the program is part of the Bayesian reasoning process behind the error estimation.

The essence of Bayesian reasoning in *Scalepack* is that you bring  $\chi^2$  (or technically speaking, the goodness-of-fit, which is related to the total  $\chi^2$  by a constant) close to 1.0 by manipulating the parameters of the error model.  $R_{merge}$ , on the other hand, is an *unweighted* statistic which is independent of the error model. It is sensitive to both intentional and unintentional manipulation of the data used to calculate it, and may not correlate with the quality of the data. An example of this is seen when collecting more and more data from the same crystal. As the redundancy goes up, the final averaged data quality definitely improves, yet the  $R_{merge}$  also goes up. As a result,  $R_{merge}$  is only really useful when comparing data which has been accumulated and treated the same. This will be discussed again later.

In short,  $I/\sigma$  is the *preferred way* of assessing the quality of diffraction data because it derives its validity from the  $\chi^2$  (likelihood) analysis. Unless all of the explicit and implicit assumptions which have been made in the calculation of

an  $R_{\text{merge}}$  are known, this criterion is less meaningful. This is particularly true when searching for a “number” which can be used by others to critically evaluate your work.

There are two modes of analysis of data using  $\chi^2$ 's. The first mode keeps the  $\chi^2$  (or more precisely, the goodness-of-fit) constant and compares the error models. Basically, this means that you are adjusting your estimates of the errors associated with the measurement until the deviations within observations agree with the expectation based on the error model.

The second mode keeps the error model constant and compares  $\chi^2$ 's. This mode is computationally much faster and is used in refinement procedures. Of the two modes, the first is more informative, because it forces you to consider changes in the error model. Which mode you use generally depends on what you are comparing. When assessing the general quality of your detector, the first mode is used. When comparing a *derivative* to a *native*, the second mode is used due to an incomplete error model which does not take into account important factors like *non-isomorphism*. Thus, the  $\chi^2$  of scaling between a native and a derivative provides a measure of non-isomorphism, assuming the detector error is accurately modeled for both samples.

$R_{\text{merge}}$  was historically used as an estimate of the non-isomorphism of data collected on film using several different crystals, and for this purpose it still has some validity because we do not account for non-isomorphism in our error model. It is not so important now that complete X-ray data sets are collected from single, frozen crystals.

One of the drawbacks of using  $R_{\text{merge}}$  as a measure of the quality of a data set is that it *can be intentionally and unintentionally manipulated*. Unintentional factors which can artificially lower  $R_{\text{merge}}$  generally have the effect of reducing the redundancy of the data or eliminating weaker observations. In crystallography, the greater the redundancy of the data is the worse is the  $R_{\text{merge}}$ , because of the correlation between  $I$  and  $\langle I \rangle$  which reduces the  $R_{\text{merge}}$ . The greater the redundancy, the lower is the correlation. For two measurements with the same  $\sigma$ , the correlation is 50%, so  $R_{\text{merge}}$  is underestimated by  $\sqrt{2}$  compared to the case of no correlation. Known unintentional factors which lower  $R_{\text{merge}}$  include the following:

1. Data collected so that lower resolution shells, where the data is strong, have a higher redundancy than the higher resolution shells, where the data is generally weaker. This can be accomplished by collecting data on detectors where  $2\theta \neq 0$ , or including data from the corners of rectangular or square image plates. There is nothing wrong with using this data; it will just artificially lower the  $R_{\text{merge}}$ .
2. Inclusion of single measurements in the calculation of  $R_{\text{merge}}$  in one widely used program, which is why a table using this erroneous calculation used to be presented in the *Scalepack* output. Although the bug in the program was unintentional, it nonetheless reduced the  $R_{\text{merge}}$  and this may have accounted for the parameter's longevity. A second, more subtle bug that reduced  $R_{\text{merge}}$  prompted the introduction of the keyword **BACKGROUND CORRECTION**. Fortunately, both bugs have now been fixed, but the point is that errors of this type can persist.
3. Omission of negative or weak reflections from the calculation of  $R_{\text{merge}}$ . This is often undocumented behavior of crystallographic data scaling/merging software. Examples include:
  - a) elimination from the  $R_{\text{merge}}$  calculation the reflections that have negative intensities.
  - b) conversion of  $I < 0$  to  $I = 0$  before the calculation of  $R_{\text{merge}}$  and inclusion of this reflection in the data set (the statistics of such a type are included in the *Scalepack* output for reasons of comparison. This is the first  $R_{\text{merge}}$  table in the log file, not the final one).
  - c) omitting reflections with  $\langle I \rangle < 0$  from the calculation of  $R_{\text{merge}}$  but inclusion of these reflections in the output data set. Default  $\sigma$  cutoffs set to unreasonable values, like 1. This is in fact the default of the software used commonly to process image plate data.
4. Use of unreasonable default or recommended rejection criteria in other programs. These eliminate individual  $I$ 's which should contribute to  $R_{\text{merge}}$  and yet are still statistically sound measurements.
5. Use of the eigenvalue filter to determine the overall B factor of a data set collected on a non-frozen, decaying crystal. In this case, the eigenvalue filter will calculate an overall B factor which is appropriate

for the middle of the data set, yet apply this to all data. As a result, the high resolution data will be downweighted compared to data processed with the first, least decayed frame as the reference. The high resolution data is generally weaker than the low resolution data, and as a result is more likely to result in higher  $R_{merge}$ . By down-weighting the high resolution data, the  $R_{merge}$  is artificially lowered. Any program which does not allow the option of setting the reference frame will have this problem. Of course, there is no problem with non-decaying crystals.

There are also intentional ways of lowering your  $R_{merge}$ . Like those ways listed above, they generally result from the statistically invalid elimination of weak reflections, reduction of the redundancy of the data, or de-emphasis of weak data. The difference between these methods and those listed above is that they are generally under the control of the user.

1. Use of an unreasonable **SIGMA CUTOFF** (e.g.  $\geq 0$ ). The rejection of weak data will always improve  $R_{merge}$ . There is a further discussion of **SIGMA CUTOFF** in the *Scalepack Keywords Descriptions* section.
  2. Use of a resolution limit cutoff. Again, the omission of weak data will improve  $R_{merge}$ . A reasonable resolution cutoff is the zone where  $1/\sigma < 2$ .
  3. Combining into a single zone for the purposes of calculations those resolution shells where  $R_{merge}$  is rapidly changing. In this case, the shell will be dominated by the strong data at the low resolution end of the zone and give the impression that the high resolution limit of the zone has better statistics than it really does. For example, if you combined all your data into a single zone, the  $R_{merge}$  in the *final* shell would be pretty good (=  $R_{merge}$  overall), when in fact it was substantially worse. As described above, it is more sensible to divide your zones into at least 10 (or preferably more) equal volumes.
  4. Omitting partially recorded reflections. This has the effect of a) reducing the redundancy, and b) eliminating poorer reflections. Partially recorded reflections will always have a higher  $\sigma$  associated with them because they have a higher total background, due to the inclusion of background from more than one frame in the reflection.
  5. Scaling I+ and I- reflections separately in the absence of a legitimate anomalous signal (**SCALE ANOMALOUS**). This has the effect of reducing the redundancy.
  6. Ignoring overloaded reflections using the **IGNORE OVERLOADS** in *Scalepack*. The intensity of overloaded or saturated, reflections cannot be directly measured, because obviously some of the pixels are saturated. Profile fitting only measures these reflections indirectly, by fitting a calculated profile to the spot using the information contained in the *wings* or *tail* of the spot. Ignoring the inaccuracies inherent in this method by ignoring overloads may have a dramatic effect on  $R_{merge}$ . Note that in case of molecular replacement the option **INCLUDE OVERLOADS** should be used.
- ➔ **IGNORE OVERLOADS** is often a useful tool, however. For example, when calculating anomalous differences you do not want to use overloaded reflections because you are looking for very, very small differences and want to use only the most accurate data. Another time you might ignore overloads is when you collect multipass data. In this case, a crystal is exposed twice, once for a short time, the other for a longer time. The longer exposure is to sufficiently darken the high resolution reflections, but will result in saturated low resolution reflections. Since the low resolution reflections can be obtained from the short exposures, the overloaded ones can be ignored in the long exposures.

### The Resolution of Diffraction Data.

A statistically sensible cutoff for the resolution of a diffraction data set is that shell where the average  $1/\sigma$  is 2 after correctly integrating and scaling the data. The resolution of the data is a distinct criterion from its completeness. The best data will be 100% complete. Completeness may suffer due to anisotropic diffraction, overlapped reflections, or geometrical constraints (data from the corners of the detector was used, beam stop or cooling nozzle shadows, etc.). In order to properly estimate the resolution of the data you have to take into account the  $1/\sigma$ ,  $R_{sym}$  and the completeness of the data.

### Understanding the *Scalepack* log file `scale.log`.

The *Scalepack* log file, `scale.log`, can be examined to see if the data scaling went well. It can be very long, since it contains the list of all `*.x` files read in and a record of every cycle of scaling. The log file is divided into several sections. In order these are:

1. the list of `*.x` files read in, and the list of reflections from each of these files that will be rejected;
2. the output file name;
3. the total number of raw reflections used before combining partial reflections;
4. the initial scale and B factors for each frame, goniostat parameters, and the space group; and
5. the initial scaling procedure.

The log file should be examined after each iteration. In particular, the errors,  $\chi^2$  and both *R* factors should be checked.



## Appendix - Installation Notes

### Downloading and unpacking HKL-2000 executable files

Executables of all programs can be obtained from our FTP server `hkl.hkl-xray.com`. Please check our web page [www.hkl-xray.com](http://www.hkl-xray.com) for the most recent information. You may use a graphical FTP application, or use the `ftp` command from a terminal window (to open a terminal window in MAC OS X, open the Finder's Go menu, choose Applications, then choose Utilities and select Terminal). To download the programs, type the following instructions:

```
ftp hkl.hkl-xray.com
```

```
login: hkl2000ftp
```

password: *The password can be obtained through the automatic password system (see the web page) or by contacting HKL Research, Inc. via e-mail. If you have successfully connected to the server you may type:*

```
ftp> ls
```

In `ver_098.###` directory there is a file called `anomal.dat` and the following subdirectories:

```
Linux:  
> HKL2000_Linux_ver_0.98.###.tar.Z  
> individual_programs
```

```
MAC-INTEL:  
> HKL2000_Macintel.###.tar.Z  
> individ_programs
```

```
SGI:  
> HKL2000_SGI_ver_0.98.###.tar.Z  
> individ_programs
```

Each `xxx.tar.Z` file contains compressed and “tarred” executables for the corresponding operating system, while the `individual_programs` directory contains individual executable files, which are not tarred or compressed. Remember that you have to use **binary mode** to transfer these files.

To download the appropriate file, e.g. `HKL2000_Macintel.###.tar.Z` use the `get` command. For example:

```
> get HKL2000_Macintel.###.tar.Z
```

After you've transferred the files, you have to uncompress them using the command:

```
> uncompress file-name.tar.Z
```

for example:

```
> uncompress HKL2000_Linux_ver_0.98.###.tar.Z
```

and “untar” the file using the command:

```
> tar -xvf file-name.tar
```

for example:

```
> tar -xvf HKL2000_Linux_ver_0.98.###.tar
```

You should have all the executable programs now.

```
access_prod
anomalous
cdirt
denzo
denzo_3d
HKL2000
hklfilename
hklsysinfo
readheader
scalepack
scalepack16m
scalepack8m
scalepack1virus
scalepackmanyframes
scalepackmf6500
scalepackribo
scalepacktest
scalepackvirus
strat
suser
xdark
xdisp
```

### Installing HKL-2000

Typically the executables are placed in the directory (requires system administrator privileges):

```
/usr/local/bin
```

You can also place executables in an arbitrary directory, for instance:

```
/home/joed/programs/hkl
```

In that case, you have to set the environmental variable `$HKLPATH` to that directory. The exact command will depend on the shell you are using to run programs (this can be determined by typing `echo $SHELL`). For C-type shells (`csh`, `tcsh`, etc.) and the example above, the command is:

```
setenv HKLPATH /home/joed/programs/hkl
```

For Bourne-type shells (e.g. `sh`, `bash`, `dash`, etc.), the command is:

```
export HKLPATH=/home/joed/programs/hkl
```

HKL-2000 requires also a directory `hklint` for site files. Its default location is:

```
/usr/local/hklint
```

(typically, accessing this location requires administrator privileges), but it can be changed to an arbitrary location, by setting the environmental variable `$HKLDIR`. For example:

```
setenv HKLDIR /home/joed/programs/hkl
```

(for C-type shells, and `export HKLDIR=<path>` for Bourne-type shells as above). The `anom.dat` file should be placed in this directory together with subdirectories containing the site files. The site file (`def.site`) of each detector (home or synchrotron) should be stored in its own directory. Each such directory should be appropriately named to easily identify the specific facility and the detector.

Example:

```
/usr/local/hkclint/Raxis4  
/usr/local/hkclint/Spring8-BL23-Quantum4  
/usr/local/hkclint/BNLX25-Quantum4
```

You can check how the environmental variables were set by typing the command:

```
env | grep HKL
```

To keep the values of the HKL environmental variables for subsequent logins, you should modify the configuration files for the shell you are using (`~/.tcshrc`, `~/.cshrc`, `~/.bashrc`, etc.).

### License key

Additionally the license key file, `cr_info`, needs to be present in the directory:

```
/usr/local/lib
```

in order for the *HKL-2000* programs to run. One `cr_info` file may contain information about many computers and detectors. The program looks through the `cr_info` file for the right combination of fields that matches the given detector family and computer information. Other fields are ignored.

To obtain the license key, you need to run the `access_prod` program on each machine that you plan to use for running the *HKL-2000* programs. Executing the `access_prod` results in the following text-based dialog (example answers to the prompts are shown in red):

- enter the name of your institution: **State University**
- enter the name of PI: **John Doe**
- enter the name of person responsible for HKL maintenance: **Jane Doe**
- enter names of all frame formats (detectors) you want to use: **mar raxis4**
- enter the number of detectors in your lab which are currently supported by HKL package: **2**
- enter the name of company from which you received your license: **HKL Research, Inc.**
- enter HKL version number: **0.98.698**
- if this is temporary license, enter the number of days you request, for permanent license enter 'noexpir': **noexpir**
- do you want to send automatically e-mail to [hkl@hkl-xray.com](mailto:hkl@hkl-xray.com) in order to activate your license (y) ? **n**

If the answer is `y` (yes), the `access_prod` program may automatically send an e-mail with a request for an access file: `cr_info`. It does not always

work, though (depending upon how the email system is configured on your computer. If the answer is n (no) the program will produce the following info file. You may save the info file and afterwards “manually” send it to [hkl@hkl-xray.com](mailto:hkl@hkl-xray.com).

Please:

- ✓ Do not edit or alter the **info** file - it is processed automatically after verification.
- ✓ Do not send an **info** file as an attachment.
- ✓ Note that the **access\_prod** program does not work properly if the setup of the computer operating system is not consistent.

The example of an info file created by the access\_prod program is given below (comments in italic):

HOST-NAME: p4.med.state.edu	<i>(extracted by the program - do not edit!)</i>
	<b>NOTE: the field is limited in length to 30 characters</b>
HOST-ID: -2035379618	<i>(extracted by the program - do not edit!)</i>
HOSTNAME: state.edu	<i>(extracted by the program - do not edit!)</i>
HW_PROV: sgi	<i>(extracted by the program - do not edit!)</i>
CPU-SERIAL: sgi,CPU,1762854231	<i>(extracted by the program - do not edit!)</i>
INSTITUTION: State University	<i>Licensee's Institution name</i>
PI: John Doe	<i>Principal Investigator or Group Leader</i>
CONTACT: Jane Doe	<i>Contact Person</i>
DETECTORS: mar raxis4	<i>Detector Format Families</i>
NO-OF-DET: 2	<i>Number of Detectors at Home Laboratory</i>
PROVIDER: HKL	<i>License Provider: Rigaku/MSC, HKL Research, etc.</i>
VERSION: 0.98.698	<i>Requested Version of the Program</i>
EXPIR: noexpir	

The following are the detector families to be used in the above questionnaire:

**dip** - DIP family of scanners  
**raxis2** - RAXIS-II  
**raxis4** - RAXIS-IV  
**raxis5** - RAXIS-V  
**mar** - MAR imaging plate detectors (except MAR345)  
**mar345** - MAR345 detector  
**brandeis** - Brandeis 1k CCD detector on line X12c at BNL  
**un-brand** - CCD Brandeis 2x2  
**unsup** - ADSC Quantum-4 CCD detector  
**unsup210** - ADSC Quantum-210 detector  
**unsup315** - ADSC Quantum-315 detector  
**un-mar** - MAR CCD detector 130mm  
**un-mar165** - MAR CCD detector 165mm

### Starting HKL-2000

After you have installed *HKL-2000* and placed the license file `cr_info` in the `/usr/local/lib` directory, the program can be started from a terminal window by typing:

```
HKL2000
```

If you receive an error message "no display name and no \$DISPLAY environment variable", it means that the X windows system is not working or not configured properly. On Mac OS X, you need to start the X11 application (from the Finder's **Go** menu, choose **Applications**, then choose **Utilities** and select **X11**). You should see the **X11** icon in the dock, double click on it to open an **xterm** window and start *HKL-2000* from this window.

If you receive an error message "Command not found", it means that HKLPATH is not in the executable command search path. Modify the environmental variable PATH or type a full path e.g. `/usr/local/bin/HKL2000`.

# Table of Figures

Figure 1. The list of directories with the def.site files	9
Figure 2. The New Site window	10
Figure 3. The Site Configuration window	11
Figure 4. The directory tree	12
Figure 5. The Raw Data directory	12
Figure 6. The Output Data directory	13
Figure 7. The Load/Create New Sets window	13
Figure 8. The Load Data Sets window	14
Figure 9. The Edit Set(s) window	14
Figure 10. All files are listed after activating the Show All Sets button	15
Figure 11. Files listed after activating the Scale Sets Only button	15
Figure 12. The Set Control window	16
Figure 13. The Select window options	16
Figure 14. The Split Set window	17
Figure 15. Data set divided into two sets with removal of four images in the middle	17
Figure 16. The experimental details panel	18
Figure 17. The Display image button	21
Figure 18. The Display window with activated "Peak Search" option	21
Figure 19. The Pick (Add) and Pick (Remove) buttons from Display window	22
Figure 20. The Frame button from the Display window	23
Figure 21. The More Peaks and Fewer Peaks buttons	23
Figure 22. The buttons for adjusting the image	24
Figure 23. The Blind Region definition window and image with assigned blind region	25
Figure 24. The Pending Sets box	25
Figure 25. The summary information about processed sets	25
Figure 26. The Resolution limits box	26
Figure 27. The Refinement Options box	26
Figure 28. The Zoom window with image after auto-indexing	27
Figure 29. The beam stop thinned in the center marks the direct beam position	28
Figure 30. The Beam Position coordinates box and its position defined on the image by the red cross	31
Figure 31. The 3D window in Controls box	32
Figure 32. The Bravais Lattice table	34
Figure 33. The refinement information e.g. positional and partiality $\chi^2$	35
Figure 34. The Macros box	35
Figure 35. The Integration Box and Zoom windows with Profile Fitting Radius and Spot Size boxes	38
Figure 36. Profile fitting radius button	38
Figure 37. The Integration Box and Zoom windows with more option for spot size and shape	39
Figure 38. The Reference Zone window	41
Figure 39. The mosaicity histogram	43
Figure 40. The Crystal Information window	44
Figure 41. The Strategy/Simulation window	48
Figure 42. The Delta Omega vs. Omega Start plot	49
Figure 43. The I/Sigma vs. Resolution plot	50
Figure 44. The Estimated Crystal Dose window	50

<i>Figure 45. The Simulation window</i>	51
<i>Figure 46. Completeness and Redundancy vs. Resolution plots</i>	53
<i>Figure 47. The Combine panel</i>	53
<i>Figure 48. The Integrate buttons colored green, orange, red</i>	55
<i>Figure 49. The special integration option</i>	56
<i>Figure 50. The diagnostic plots: <math>\chi^2</math> vs. Frame and Distance Change vs. Frame</i>	56
<i>Figure 51. The program is waiting for new frames during processing the data</i>	58
<i>Figure 52. The global refinement options</i>	59
<i>Figure 53. The Scaling Options panel</i>	60
<i>Figure 54. The Scalepack output files</i>	62
<i>Figure 55. The scaling Controls panel</i>	63
<i>Figure 56. Frames selected to be excluded from scaling are marked red</i>	64
<i>Figure 57. The output plot: <math>1/\Sigma</math> vs. Resolution</i>	65
<i>Figure 58. The output plot: Completeness vs. Resolution</i>	65
<i>Figure 59. The output plot: Scale and B vs. Frame</i>	66
<i>Figure 60. Info table after first round of scaling</i>	67
<i>Figure 61. Anomalous signal detection</i>	68