Notes on MicroED data processing

Introduction

These notes focus on MicroED data processing with the CCP4 programs, XDS, and Phenix. Over the past few years, these suites have been adapted for MicroED, and they are the ones most commonly used in our laboratory. While not covered here, it is possible to use other programs as well.

autoPROC — Work is underway to enable processing of MicroED data with the Global Phasing tools.

DIALS — MicroED processing with DIALS has progressed rapidly, but will not covered here. The course environment does provide DIALS, and interested users are encouraged to try it on the example data sets on the stick.

HKL3000 — Recent versions of the HKL suite contain some modifications to allow processing of MicroED data. Familiarity with the command-line interface of denzo and scalepack may still be required.

Even though these notes may appear to do so, it is not possible to give a failsafe step-by-step protocol for MicroED data processing. In particular, MicroED data processing is usually not a linear process. Previous steps often have to be repeated given information learned during later steps. These notes attempt to point out what to look for when deciding whether to backtrack and how far. Lastly, note that all programs covered here have their own documentation and often their own tutorials. The material covered here barely scratches the surface.

User instructions

Procedure 1. Booting the USB stick

MicroED data processing is interactive and makes extensive use of graphical user interfaces; laptop users may be well served by an external mouse if working with the built-in trackpad is too exhausting.

1. The computer must have at least one free USB type A port to boot the stick. Computers that only have USB type C ports may be able to boot the stick with a suitable adapter. To boot into the USB stick, the host system's boot order may need to be adjusted. This step depends on the particular computer involved.

    - On a PC, look for something like Boot Menu, Boot Options, or some such during the boot sequence. On our test machines the function is predominantly bound to the **F12** key. Other common options are the **DEL**, **F2**, or any of the **F9** through **F11** keys.

    - On Apple hardware, hold an **alt/option** key during the boot tune. Then select the EFI Boot icon, and click the upwards pointing arrow underneath it.

    - If the computer is running Linux or macOS and already has a sufficiently complete SBGrid installation, there may be no need to boot into the stick. It should be possible to complete the computational part of the course by accessing its data partition from the native operating system. The source code for the tvips-tools package can be downloaded from http://cryoem.ucla.edu/downloads; precompiled, stable binaries are available from http://www.ccpem.ac.uk/download.php, and the development snapshots used during this course can be obtained from http://cryoem.ucla.edu/downloads/snapshots.

2. Go through gnome-initial-setup and answer the questions to taste. Browse the Help items in Getting Started if desired.

Note

Users with a HiDPI display, or Apple users with Retina hardware, may want to adjust the scaling factor. In a Terminal window, give the command

```
$ gsettings set org.gnome.desktop.interface scaling-factor 1
```

Note

Wireless network adapters *should* work within the Linux environment, and *should* be able to connect to the `UCLA_WEB` wireless network.

Note

The system installed on the stick does not touch any data stored on the host system's hard drives. While it is possible to explicitly mount the host's file systems when booted into the stick, it may be safer to mount the stick's data partition from the host's native operating system instead to transfer data. However, the home directory, where all the work is done, is generally not visible when the stick is mounted from the computer's native operating system. This can be overcome by creating a zip-file containing the relevant files on the stick's `data` partition, which *is* visible from *e.g.* macOS or Windows.

```
$ zip -r /mnt/data/my_data.zip dir1 dir2
```

This will compress *`dir1`* and *`dir2`* in the current directory into a `my_data.zip` file on the stick's `data` partition.

Warning

Do not perform any system updates when booted into the stick! This can break the stick because not all required drivers may be available for more recent versions of the operating system.

Warning

Do not remove the stick unless the operating system has been shut down properly! Do not remove the power while booted into the USB stick! The file system does not journal regular writes, which increases performance but risks loosing data if the system is not shut down normally.

Warning

The stick has several partitions, not all if which are recognized by *e.g.* Windows operating systems, but they are all required for the stick to function properly. Windows may suggest formatting these partitions when the stick is inserted, but doing so will prevent the stick from booting into Linux. Do not format any unrecognized partitions in Windows–the data partition will be accessible, even though the other partitions are not.

Provided datasets

Table 1. Data collection parameters

Dataset, camera, filename(s)	Distance (mm)	Rotation rate (°/s)	Exposure time (s)	Binning	Gain	σ	ADC offset
Lysozyme, TVIPS TemCam-F416							

Dataset, camera, filename(s)	Distance (mm)	Rotation rate (°/s)	Exposure time (s)	Binning	Gain	σ	ADC offset
movie4s_01_00[01].tvips	2640	-0.09	4.1	2×2	2	5	18
movie4s_23_000.tvips	2640	-0.09	4.1	2×2	2	6	17
Proteinase K, Thermo Fisher CetaD							
lam11/*.mrc	2055	+0.2	3.0	2×2	14	N/A	16
lamv/*.mrc	2055	+0.2	3.0	2×2	14	N/A	16
Ketone, Thermo Fisher Falcon III							
2_20190913_162354.mrc	1053	+0.9	1.0	1×1	5	N/A	16
3_20190913_163300.mrc	1053	+0.9	1.0	1×1	4	N/A	16

These parameters are needed to process the given datasets. The data are located at `/mnt/data/`*`sample`*`/raw` where *`sample`* is either `lysozyme`, `proteinasek`, or `ketone`. The lysozyme data can also be accessed from its DOI, 10.15785/SBGRID/222 [http://dx.doi.org/10.15785/SBGRID/222]. A negative rotation rate implies the data set should be processed with the Reverse direction of spindle rotation box checked in iMosflm, or `ROTATION_AXIS` set to `-1 0 0` in XDS.

Table 2. Symmetries, unit cells, and molecular replacement search models

Sample	Unit cell						Spacegroup	MR model
	a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)		
Lysozyme	75.96	75.96	37.22	90	90	90	$P4_32_12$	1iee
Proteinase K	67.30	67.30	106.43	90	90	90	$P4_32_12$	6cl7
Ketone	6.61	9.08	19.92	90	90	90	$P2_12_12_1$	N/A

Ideally, the unit cell is determined during indexing. The spacegroup will generally not be known until the structure is solved. Knowing both in advance tremendously helps MicroED data processing. The actual values obtained will differ slightly, but should be within a few percent of those tabulated here.

Note

This text assumes all steps are completed in order, as the output of one step becomes the input of the next. It *is* possible to skip steps because sample output of the individual steps is provided (see Appendix A, *The `data` partition*).

Image conversion

Data reduction software developed for X-ray crystallography does generally not understand the MRC or TVIPS set file format. Instead, the data are first converted to SMV, a simple and widely supported image file format. While the conversion tools can sometimes extract some of the necessary metadata from the files (*e.g.* binning, pixel size, timestamp, or wavelength in the case of TVIPS set files), diffraction data processing requires some information that is *not* recorded by the camera. This has to be supplied on the command line while converting the images.

Note

All commands are executed from Terminal, a terminal emulator. The prompt, `$`, is not part of the command, nor is a backslash (\) if it occurs at the end of a line. The latter merely indicates that

the line is too long to fit onto the page; characters on lines separated by backslashes should be entered as a single line into the Terminal application.

Data processing will create a lot of files. Processing each sample in a separate directory helps keeping things organized. Before starting work on the lysozyme sample for instance, it makes sense to create a new directory and change into it.

Example 1.

```
$ mkdir lysozyme
$ cd lysozyme
```

Procedure 2. Convert TVIPS set files

1. To convert the first lysozyme dataset:

   ### Example 2.

   ```
   $ tvips2smv -d 2640 -r 0.09 -v -o lyso_01/lyso_###.img \
       /mnt/data/lysozyme/raw/movie4s_01_00?.tvips
   Estimated exposure time: 4.10 s \
       (from 120 timestamp differences in the range [4.09, 4.11] s)
   ```

   For this dataset the sample-detector distance is 2640 mm and the rotation rate of the stage is 0.09°/s (see Table 1, "Data collection parameters"; note that the conversion tools expect the *modulus* of the rotation rate—the rotation *direction*, given by its sign, will be entered in Procedure 4, "Set initial parameters for iMosflm processing" below). The output images are written to `lyso_01/lyso_001.img`, `lyso_01/lyso_002.img`, *etc*. In this case the exposure time was calculated to 4.10 s; this value should lie in the indicated range, which in turn should be small. **tvips2smv** uses the exposure time together with the supplied rotation rate to calculate the oscillation range of each image.

2. (Optional) The TVIPS TemCam-F416 camera in rolling-shutter mode truncates negative pixel values. These pixel values cannot be recovered, but they can be modeled. To attempt modeling negative pixel values, and applying a pedestal to ensure all pixel values are positive, replace the **tvips2smv** command above with:

   ### Example 3.

   ```
   $ img2px -d 2640 -r 0.09 -s 5 -v -o lyso_01/lyso_###.img \
       /mnt/data/lysozyme/raw/movie4s_01_00?.tvips
   ```

   This will apply the modeling algorithm to all pixels with standard deviations greater than 5 ADU. The purpose of the threshold is to exclude pixels in the beam stop shadow, since their values do not follow the distribution of pixel values that record diffracted electrons. A suitable threshold may be determined by inspecting `stddev.img` (see the use of **adxv** below), generated by running **img2px** without the `-s` option and finding a pixel value that is greater than those in the beam stop shadow, but smaller than those along the borders of the image. The pixel values in `stddev.img` are 1000-fold multiples of the actual standard deviation; to determine a suitable value for the `-s` option, the pixel value determined by inspecting `stddev.img` must be divided by 1000. The pixels actually used during modeling are indicated in `mask.img`, and the pedestal, or ADC offset, required to ensure the modeled pixel values are greater than zero is output just before **img2px** terminates; any other output is of little interest here.

3. (Optional) The TVIPS tools come with online documentation. To view the manual page, run

Example 4.

```
$ man tvips2smv
```

4. Inspect the images visually. It is generally a good idea to know what the images look like. In real cases, this often determines whether it is worth spending any time on attempting to process the data at all.

Example 5.

```
$ adxv lyso_01/*.img
```

Note that while different methods will have to be used to convert data from different cameras, the same images are suitable for processing in both MOSFLM and XDS.

Procedure 3. Convert MRC files

- Standard MRC files contain almost no metadata at all. Therefore, most of it has to be supplied on the command line.

Example 6.

```
$ mrc2smv -d 2055 -k 0 -r 0.2 -w 0.0251 -B 2 -E 3 -M 16 -P 0.014 \
-v -o prot_11/prot_###.img /mnt/data/proteinasek/raw/lam11/*.mrc
```

The `-d`, `-r`, `-v`, and `-o` options are identical to those used with the **tvips2smv** or **img2px** commands and denote the distance, rotation rate, verbosity, and output, respectively. Where an option requires a value, they are tabulated in Table 1, “Data collection parameters”.

-k *rotation*	Rotate the images counterclockwise by *rotation* × 90°. This can be used to align the tilt axis to the convention of the downstream processing suite. The default value is **0**, so this option is not strictly necessary in the example above, but is included for illustration purposes.
-w *wavelength*	Set wavelength to *wavelength* Å in the output files. All datasets here were collected at 200 kV, corresponding to a de Broglie wavelength of 0.0251 Å. The de Broglie wavelength corresponding to a given acceleration voltage *ht* kV may be found using **ht2wavelength *ht***.
-B *binning*	This option defines the binning factor. Binning is assumed to be equal in the horizontal and vertical directions.
-E *exposure_time*	This option specifies the exposure time, in seconds per frame.
-M	Additive offset for linear pixel value transformation. Unlike TVIPS set files, MRC files *do* contain negative pixel values. However, these cannot be represented in the output SMV files, and integration software does not always gracefully deal with them. This option should be used to ensure the number of underflows is small (less than one or, at most, a few percent), while not being so large as to cause overflows. **mrc2smv** will output the number of over- and underflows, one line at the time, as the dataset is traversed frame by frame.
-P	The physical side length of a square pixel, in mm. The pixels on the CetaD as well as the Falcon III measure 0.014 mm along its sides.

Indexing and integration

iMosflm

In this section, the lysozyme dataset is used as an example; the procedure for processing proteinase K is analogous, but small-molecule structures, such as ketone, are best processed with XDS.

Procedure 4. Set initial parameters for iMosflm processing

1. Start iMosflm, the graphical user interface for MOSFLM.

   ```
   $ imosflm &
   ```

   Before loading the images, select the Processing options item from the Settings menu and choose the Indexing tab. Uncheck the Automatically index after spot finding option. In MicroED, indexing often requires a specific subset of images to work, and time can be saved by turning this option off in favor of explicitly invoking autoindexing. This change only has to be made the first time iMosflm is run.

2. To load images, select the Add images... item from the Session menu. Navigate to the images converted previously and double-click any one of them.

3. Access the Experiment settings from the Settings menu in the main iMosflm window.

   a. If the rotation rate is negative (see Table 1, “Data collection parameters”) check the Reverse direction of spindle rotation box in the Experiment tab.

      #### Note

      Owing to the small wavelength in electron diffraction (~0.0251 Å), it will not always be immediately obvious when the rotation direction is wrong. Datasets will frequently index in both directions, but for certain orientations the reflection predictions from the model will disagree wildly with the actual spots. Otherwise, these problems may not become apparent until multiple datasets are merged.

   b. Set Gain according to Table 1, “Data collection parameters” and ADC offset to `0` in the Detector tab. This is necessary because MOSFLM treats the converted data as images from an ADSC Q210 detector, which is different from the TVIPS TemCam F-416 or Thermo Fisher CetaD cameras used here.

      If **img2px** was used to convert the images, the ADC offset should instead be set to whatever the program output before terminating. The wavelength is automatically transported from the `.tvips` files during image conversion.

      #### Note

      The gain is detector-dependent. During integration, iMosflm will warn if the gain appears to be wrong, and it is advisable to heed that warning. XDS will determine the gain itself, without any user intervention.

4. Access the Processing options in the Settings menu. In the Integration section of the Advanced integration tab, change Null pixel threshold to `-1`. This is necessary because unless **img2px** was used for format conversion, the images generally contain pixels with zero pixel values. For this particular camera, many reflections will be erroneously discarded during integration unless the Null pixel threshold is adjusted.

5. Find the beam center.

   - In the image display window, zoom into the beam center and drag out the red mask. Center the mask on the halo around the circular part of the beamstop shadow by dragging the beam center indicated by the magenta cross.

   - Alternatively, find the beam center using Friedel mates. If two Friedel mates can be identified on a single image, the beam center will be the midpoint on the line between them.

     ## Note

     For datasets where the beam center moves considerably over the course of the dataset (see Procedure 6, “Integration”), indexing can be very sensitive to the choice of beam center. In these cases, small changes to the beam center can make or break indexing.

6. Adjust the three masks.

   green — The green mask is for general exclusion. Here, it should be used to cover the shadow of the beam stop. A polygonal exclusion mask can be created using the Masking tool in the image viewer’s toolbar.

   red — The red mask controls the spot finding area but does not directly affect integration. It usually does not need to be changed. The red rectangle indicates the area used for estimation of the image background. It should not cover any areas that deviate significantly from the overall image (*e.g.* the beam stop shadow).

   blue — The blue mask indicates the resolution limits. This affects integration, but not spot finding or indexing. For initial assessment of a dataset it may be advisable to set the high-resolution limit to the edge of the image. For final integration, it may be preferable to drag the high-resolution limit off the image entirely; this will integrate the high-resolution spots in the corner of the detector and allows the resolution limit to be determined during subsequent scaling. This strategy is however not advisable for weakly diffracting data: significant instability may ensue if there is little or no signal in the area between the blue circles. The low-resolution cutoff can be moved to behind the beamstop shadow.

Procedure 5. Indexing

Note

Indexing and integration are presented as two separate steps. In reality they are often intertwined and performed iteratively for MicroED data.

iMosflm will try to use two images 90° apart in ϕ (the rotation, or oscillation, angle) for indexing. MicroED data rarely span 90°, and two images rarely contain sufficiently many spots for indexing. Instead:

1. Choose the Indexing task in iMosflm.

2. Chose a set of images spanning a wedge of approximately 10° to 20°. If the wedge is too small, the underlying three-dimensional lattice will not be apparent in the volume of reciprocal space spanned, and autoindexing will fail. If the wedge is too large, experimental errors (*e.g.* moving beam center, rotation rate jitter) may accumulate and prevent successful indexing.

   The number of spots in the wedge should be in the hundreds. Thousands of spots are too many, and tens of spots are too few, but exceptions do occur. The provided datasets can generally be indexed using images 10, 20, 30, 40, and 50.

3. Indexing solutions will be sorted by their penalty and the root mean square difference between predicted and observed spot positions (labeled Pen. and sigma(x, y), respectively, in the table of solutions after the Index button has been pressed). Ideally, the penalty increases dramatically for solutions immediately succeeding the correct alternative, and the correct crystal system will be automatically highlighted. Since the spacegroup is known (see Table 2, "Symmetries, unit cells, and molecular replacement search models"), it can be chosen from the drop-down menu.

4. Choose the mosaicity. For MicroED images the value suggested by the procedure invoked by pressing the Estimate button it is rarely reliable. Instead it often has to be determined iteratively during integration (see Procedure 6, "Integration"). It is usually safer to initially overestimate the mosaicity than to underestimate it.

5. (Optional) If indexing fails:

   - Check the Max cell edge. For lysozyme an appropriate value is 120 Å; for proteinase K 200 Å may be better. The indexing algorithm can not find a solution with a cell edge larger than this value; choosing an appropriate value therefore requires some prior knowledge (see Table 2, "Symmetries, unit cells, and molecular replacement search models").

   - Choose a different set of images for indexing.

   - Tweak the beam center (see Step 5 in Procedure 4, "Set initial parameters for iMosflm processing").

   - Use the general exclusion mask to also cover regions near the rotation axis. It appears that the wide spots (*i.e.* spots apparent on many abutting frames and marked by green boxes during integration) in this area of the image can confuse autoindexing. The added mask should be removed prior to integration.

Note

Cell refinement (or postrefinement) is usually skipped for MicroED data. When indexing using many frames, iMosflm already performs postrefinement of the chosen solution, and the procedure invoked by the Cell refinement task tends to be fragile with MicroED data.

Procedure 6. Integration

1. Choose the Integration task in iMosflm. Change the MTZ filename to something that reflects the crystal to be integrated. The chosen path will be the name of the output MTZ file from iMosflm. Take care to choose a unique name for each integrated dataset, otherwise previous results may be overwritten.

2. Fix Tilt and Twist by checking the corresponding boxes in the Fix column. These should be fixed at zero and never change.

3. (Optional) Fix Mosaicity and Distance. These may be unstable during refinement and often behave as general error sinks; if they diverge too much, they may need to be fixed.

4. Turn on Show predictions on images during processing. This allows the predictions to be verified against the actual spots during integration and is essential during the first iterations of integration.

5. Process the data set. If the model does not cover the actual spots, integration will yield suboptimal results. If so, reindex using a different set of images. This will yield a slightly different crystal orientation.

6. Adjust the mosaicity, and optionally the mosaic block size in the Images task. Ideally the mosaicity should be set such that predictions cover spots while they are visible in the image, but not much more.

Note that the weak tails of the spots may not be visible, but should nevertheless be integrated. If the crystal orientation is correct it is often possible to adjust the mosaicity to the refined value of the first frames of the dataset.

7. (Optional) Observe the beam center movement over the course of integration; it is particularly pronounced for the lysozyme datasets. Smooth variations indicate that MOSFLM is able to correctly track the motion.

8. If iMosflm warns about parameters that have refined to questionable values when integration finishes or is aborted, it is generally a good idea to Reset them. Also pay attention to the warnings in the lower, right corner: these may indicate aspects that can be adjusted to improve integration results.

9. Once the entire dataset is processed, perform a QuickScale. Reference statistics for different datasets are given in Table 2, “Symmetries, unit cells, and molecular replacement search models”.

   Diverging mean(k) and 0k scale factors are classically indicative of radiation damage; the later frames of the dataset may have to be rejected during subsequent merging owing to dose considerations. Ideally, POINTLESS will suggest the correct spacegroup from Table 2, “Symmetries, unit cells, and molecular replacement search models”.

   The output paths of the QuickScale operation are controlled by the <program_name>_ field in the Pointless, Aimless/Scala & Truncate MTZ output identifier section in the Sort Scale and Merge tab of the Processing options window. Note that this field appears not to be updated when the MTZ filename is changed in the main window; take care if the output of QuickScale is to be preserved.

Note

MOSFLM will throw `XML format errors` when integrating data from the CetaD. The error message *can* be dismissed, but will reappear as soon as the next image is processed. This is annoying, but does not appear to adversely affect integration.

XDS

Data processing with XDS follows the same general principles as with MOSFLM, but the lack of a real-time graphical user interface makes the procedure more opaque. In this course, XDS can be used to process the proteinase K and the ketone data sets, but there is nothing preventing XDS from being applied to the lysozyme data as well.

Procedure 7. Indexing and integration with XDS

1. Start xdsgui, a graphical user interface for XDS.

   ```
   $ xdsgui &
   ```

   A directory for processing a single dataset, or a project, may be created by choosing the Choose or create new folder button, and clicking the Create new folder on the right side of the top row. Then, select the desired directory and click Choose. It is recommended to process each dataset in a separate directory, and one can easily switch between different projects by clicking their corresponding paths in the Projects tab; the currently active project is displayed in the title bar.

2. Load an image from the dataset, by clicking the Frame tab followed by Load. Then navigate to the folder with the converted images, select an image and click Open. Clicking generate XDS.INP will cause xdsgui to generate an initial input file for XDS, `XDS.INP`. The program does this by inspecting the images in the directory containing the chosen image and deriving default values based on their metadata, much of which were provided during image conversion.

3. Switch to the XDS.INP tab and adjust the generated input file. The input file is a plain text file with keywords and values separated by an equals (=) sign. Any text following an exclamation mark (!) is a comment and will be ignored by XDS.

   - Find the beam center, analogously to how it is done in iMosflm. This can be done by mousing over the tentative beam center in the Frame tab, and assigning the values displayed for x and y to **ORGX** and **ORGY** in `XDS.INP`. XDS appears to be more tolerant to errors in the beam center than MOSFLM.

   - Set **REFINE(CORRECT)=CELL BEAM ORIENTATION AXIS**. By default, XDS will attempt to refine all parameters during its `CORRECT` step. This is often unstable with MicroED data, hence the number of refined parameters is reduced.

   - For small molecules, set **STRONG_PIXEL=6** and **MINIMUM_NUMBER_OF_PIXELS_IN_A_SPOT=6**. The spots in diffraction patterns from a small molecules are often large; these settings are intended to prevent spurious peaks in the images from derailing indexing.

   - Add **OFFSET=*adc_offset***. This tells XDS that *adc_offset* ADU were added to each pixel, and should be subtracted for the purpose of estimating the signal of the integrated reflection. The value of *adc_offset* should equal the `ADC offset` value output by **img2px**, or whatever value was given with the `-M` option if **mrc2smv** was used for image conversion.

   - Set **ROTATION_AXIS** according to the rotation direction (*i.e.* the sign of the rotation rate) in Table 1, "Data collection parameters". The proteinase K and ketone datasets were all collected in the forward direction, for which the default value of **ROTATION_AXIS=1 0 0** is suitable.

   Save the modified input file by clicking Save and run start processing by clicking Run XDS. Pay attention to the output in the Terminal window. If XDS fails, parameters may need to be adjusted, or the failing steps may be restarted. The output of the individual steps is also available in the XYCORR, INIT, COLSPOT, IDXREF, DEFPIX, INTEGRATE, and CORRECT tabs.

4. (Optional) If integration fails:

   - The most common problem for MicroED data is an insufficient number of indexed spots to continue past the `IDXREF` stage. The simplest workaround, even though not a proper solution, is to modify the **JOB** line to only include the stages past `IDXREF`, *i.e.* `DEFPIX`, `INTEGRATE`, and `CORRECT`.

   - XDS's **INCLUDE_RESOLUTION_RANGE** fills a role similar to the space between the blue circles in iMosflm. Spots that fall in this range are classified as trusted, and used for *e.g.* parameter refinement. The included range may need to be changed if refinement fails.

   - It is not uncommon in MicroED that the frames at the either end of a dataset differ in some critical characteristics from their expectations. For instance, the first and/or last frames may have been exposed longer, or they may have been recorded at different rotation rates (because the stage was ramping up or down). In XDS, the frames to include for integration is controlled using **DATA_RANGE** and **EXCLUDE_DATA_RANGE**.

   - Adjust **BACKGROUND_RANGE**, the first and last image used for background estimation. By default, XDS uses the first 5° of data, but if the background is variable, it may make sense to increase this.

   - Adjust **SPOT_RANGE**, the first and last images used for locating strong spots. This is similar to the images chosen for spotfinding in iMosflm and defaults to the same value as `BACKGROUND_RANGE`.

- Problems with symmetry determination may be addressed by adjusting **TEST_RESOLUTION_RANGE**. This parameter controls the lower and upper resolution limits of strong reflections used for deciding on spacegroup.

- The **DELPHI** parameter controls the number of spot profiles for integration. This parameter can be increased from its default value of 5° if there are too few spots to establish adequate profiles.

- Lastly, xdsgui provides a means to update the geometry from a previous (suboptimal) integration iteration. To apply this, navigate to the tools tab, choose Optimizing data quality and click copy latest geometry description over previous one and copy BEAM_DIVERGENCE, REFLECTING_RANGE from INTEGRATE.LP to XDS.INP. The latter step will update the current `XDS.INP` file, and xdsgui asks the user whether the updated file is to be Reloaded or discarded.

Note

XDS will not attempt to find screw axes. For instance, XDS will report P222 where the real symmetry is $P2_12_12_1$, and P422 where $P4_32_12$ would have been expected. This is not a big problem here, because symmetry can be determined during phasing, either by molecular replacement or by direct methods.

Table 3. Reference values for integration

Dataset	**Lysozyme**				**Proteinase K**				**Ketone**	
	01		**23**		**lamv**		**lam11**		**2**	**3**
	tvips2smv	**img2px**	**tvips2smv**	**img2px**	**MOSFLM**	**XDS**	**MOSFLM**	**XDS**	**XDS**	**XDS**
Mosaicity (°)	0.65	0.56	0.92	0.90	0.99	0.393	1.54	0.242	N/A	N/A
Results										
Symmetry	P422	C222	P422	C121	P422	P1	P422	P1	P1	P1
Highest resolution shell (Å)	2.16–2.10	2.16–2.10	2.16–2.10	2.16–2.10	1.83–1.80	1.57–1.48	1.83–1.80	1.56–1.47	0.96–0.90	0.96–0.90
R_{merge}	0.220 (0.637)	0.345 (1.756)	0.192 (0.510)	0.272 (1.084)	0.197 (0.828)	0.257 (N/A)	0.331 (1.541)	0.152 (2.337)	0.078 (0.232)	0.172 (0.335)
$CC_{1/2}$	0.960 (0.184)	0.913 (0.017)	0.981 (0.492)	0.745 (0.194)	0.969 (0.418)	0.970 (0.0)	0.946 (0.102)	0.976 (0.068)	98.1 (79.0)	94.7 (71.1)
Completeness (%)	72.1 (67.2)	58.2 (58.1)	50.3 (50.9)	18.2 (18.9)	64.9 (64.8)	17.3 (7.5)	59.6 (56.3)	53.8 (23.1)	73.1 (70.2)	73.7 (70.0)
Multiplicity	2.9 (2.2)	2.8 (2.8)	4.4 (4.6)	3.1 (3.1)	3.0 (2.8)	1.4 (1.0)	3.2 (2.5)	1.8 (1.2)	1.7 (1.6)	1.8 (1.6)

Reflections were generally integrated to the edge of the detector; for lysozyme, they were integrated into the corners. Symmetry denotes the symmetry as determined by POINTLESS when processing with MOSFLM; XDS integrates in P1.

Scaling and merging

Procedure 8. Merging and phasing preparation

1. Start the CCP4Interface

   ```
   $ ccp4i &
   ```

   The first time the CCP4Interface is started, the Directories & Project Directory window opens automatically; whenever the project directory needs to be changed on subsequent runs, the window can be accessed by pressing the Directories&ProjectDir button. Fill in the Project (*e.g.* setting it to `lysozyme`) and uses directory fields (click Browse... and navigate to the directory where the lysozyme data were processed) and Add project. Click Apply&Exit.

   #### Note

   These notes use the old version of the CCP4 interface. It is entirely possible to use the new interface, ccp4i2, but that is not covered here.

2. If applicable, download the coordinates and structure factors for the molecular replacement search model.

   - Open Firefox, navigate to http://www.rcsb.org, and search for the appropriate PDB id (see Table 2, “Symmetries, unit cells, and molecular replacement search models”). From Download Files choose PDB Format to obtain the PDB-formatted coordinates and then Structure Factors (CIF) to retrieve the structure factors in CIF format. Move the downloaded files to the project directory, *e.g.*

     ```
     $ mv ~/Downloads/1iee* .
     ```

     Convert the CIF structure factors to MTZ file format. Choose Convert to/modify/extend MTZ from the Reflection Data utilities task in the CCP4Interface. Choose Import reflection file in mmCIF format and create MTZ file. Set the In path to that of the downloaded CIF file, and note (or change) the Out path. Choose Run Now from the Run drop-down menu.

   - Alternatively, use **phenix.fetch_pdb** from the Phenix suite.

     ```
     $ phenix.fetch_pdb --mtz 1iee
     ```

     When invoked with the `--mtz` option, **phenix.fetch_pdb** will download the structure factors and PDB-formatted coordinates and generate an MTZ file.

Procedure 9. Merging with AIMLESS

1. Select Symmetry, Scale, Merge (Aimless) from the Data Reduction and Analysis task in the CCP4Interface.

2. Set the path of HKLIN #1 to the output MTZ file from the integration step using the Browse button. Once additional integrated datasets are available, they can be added by clicking Add File. To visualize the coverage of reciprocal space of an individual dataset, click the View button and select the HKL Zones icon.

Note

It is possible to use POINTLESS and AIMLESS to scale and merge intensities integrated in XDS as well by assigning XDS's `XDS_ASCII.HKL` to the HKLIN input for AIMLESS.

3. Set the Project name, crystal name, and dataset name to taste. Set HKLOUT to *e.g.* `lyso_merged.mtz`.

4. Check Ensure unique data & add FreeR column for 0.05 fraction of the data and Copy FreeR from another MTZ. Set the MTZ with FreeR to the MTZ file created from the RCSB CIF file previously (see Procedure 8, "Merging and phasing preparation") using the Browse button.

5. Click Run Now from the Run drop-down menu.

6. Once AIMLESS has finished, look at the log file by double-clicking on the job in the central pane in CCP4Interface.

7. (Optional) Even though individual datasets may have merged well through iMosflm's QuickScale option, it is possible that multicrystal merging is not well-behaved.

   - Sometimes, this can be overcome by excluding known poor batches (check Exclude a range of batches under the Resolution and batch exclusions tab, then Add batches). The Mn(k) & 0k(theta=0) v. batch, Rmerge v Batch, and Maximum resolution limit, I/sigma > 1.0 graphs in the AIMLESS log can aid in deciding on what, if any, batches to exclude.

   - Another option is to cut the resolution. The CC(1/2) v resolution, Completeness v Resolution and Wilson plot graphs in the AIMLESS log are useful tools to determine a resolution cutoff.

     R_{merge} is widely recognized as an unsuitable metric for determining the resolution cutoff. For MicroED data the signal-to-noise ratio at high resolution often appears inflated, which makes Mn(I/sd) a similarly poor determinant for this purpose.

   - It is possible to override the spacegroup suggested by POINTLESS. Check Customize symmetry determination and Choose a previous solution, then check and set Choose solution from search by Space group name in the Options for Pointless section.

   - MOSFLM integrates every spot by straight intensity summation as well as using a profile derived from strong nearby spots. By default (combine in the Observations Used & Handling of Partials section), AIMLESS attempts to automatically decide which integrated intensities to use during merging. For some datasets merging works better when only using summation intensities: choose summation intensities or profile fitted intensities (IPR) in the Observations Used & Handling of Partials section.

   - Particularly for high-mosaicity datasets, the default acceptance criteria for partials may be too strict. It is sometimes beneficial to widen the gap between the upper and lower limits of the Only accept partials with total fraction between item in the Observations Used & Handling of Partials section. This will generally increase the number of accepted reflections at the cost of decreasing internal consistency.

   - The rejection criteria for outliers in the Reject Outliers section may need to be adjusted. This will generally increase the internal consistency at the cost of decreasing the number of accepted reflections.

   - Datasets processed in the opposite rotation direction may go undetected until the merging stage. Often, such data appear to merge fine when considered on their own, but will cause a multicrystal

merge to exhibit very poor statistics. In these cases, it is advisable to carefully review the integration results to isolate the wrongly processed data.

Procedure 10. Merging with XSCALE

1. To scale and merge data from several crystals processed in XDS into a single dataset, navigate to the XSCALE tab. Add an `INPUT_FILE` line with the path to `XDS_ASCII.HKL` for each additional dataset to be included. Optionally, append a `INCLUDE_RESOLUTION_RANGE`, after each `INPUT_FILE` line to limit the data to be included from the particular dataset.

   The result of the merge are tabulated in `XSCALE.LP`. Pay particular attention to the `STATISTICS OF SCALED OUTPUT DATA SET` table. Completeness and/or multiplicity should increase when merging multiple crystals, but other merging statistics will generally deteriorate slightly due to non-isomorphism. If the quality indicators differ significantly, merging with XSCALE can be troubleshot analogously to merging with AIMLESS.

2. If the data are to be phased by molecular replacement, convert the merged intensities to MTZ file format. This can be done from the XDSCONV tab in xdsgui, by setting `OUTPUT_FILE` to *e.g.* `temp.hkl CCP_I+F`. When clicking Run XDSCONV this will produce an MTZ file called `temp.mtz` in the project directory.

   To add the free flags from the molecular replacement search model, choose Merge MTZ Files (Cad) from the Reflection Data Utilities in the CCP4Interface, and use the Browse buttons to give the merged MTZ as the first file and the MTZ from the search model as the second. Use all columns from the merged data, but only selected columns, namely the R-free-flags from the search model's data. It may also make sense to limit the resolution of the second file to that of the first file, or to extend the free set, in case the second file spans a smaller resolution range than the first.

   ### Note

   REFMAC has trouble refining the test data sets unless the scaled and merged MTZ has the correct spacegroup. While this, strictly speaking, will remain undetermined until after the next step, Phasing, the merged file may for simplicity be amended with the correct spacegroup here. This can be done by checking the Override input space group in MTZ file and output data in space group box, and providing the appropriate spacegroup (see Table 2, "Symmetries, unit cells, and molecular replacement search models") in the adjacent text box.

Reference statistics for different samples are given in Table 4, "Reference values for scaling and merging".

Table 4. Reference values for scaling and merging

Dataset	**Lysozyme**		**Proteinase K**		**Ketone**
		img2px	**MOSFLM**	**XDS**	**XDS**
Symmetry	P422	P422	P422	P422	P222
Highest resolution shell (Å)	2.16–2.10	2.16–2.10	1.84–1.80	1.85–1.80	0.96–0.90
R_{merge}	0.220 (0.584)	0.331 (1.968)	0.292 (1.248)	0.235 (1.549)	0.270 (0.522)
$CC_{1/2}$	0.970 (0.365)	0.956 (0.094)	0.966 (0.267)	0.970 (0.160)	90.8 (26.5)
Completeness (%)	97.8 (97.3)	99.1 (99.6)	90.2 (88.8)	91.1 (91.8)	90.9 (90.0)
Multiplicity	4.4 (3.9)	5.1 (5.4)	4.3 (3.6)	2.8 (2.7)	2.8 (2.5)

Symmetry denotes the symmetry used for merging. Values in parentheses refer to the highest resolution shell.

Phasing

Procedure 11. Molecular replacement

- Phasing by molecular replacement can be done either using MOLREP from CCP4 or Phaser from the Phenix suite. MOLREP is often faster than Phaser, but does not support electron scattering factors. For the simple molecular replacement problems posed by the data at hand, electron scattering factors are not crucial to finding a solution.

  - **Molecular replacement using MOLREP**

    a. Select Run Molrep - auto MR from the Molecular Replacement task in the CCP4Interface.

    b. Set Data to the MTZ file with the merged intensities from the section called “Scaling and merging”. Set Model to the coordinate file downloaded from the RCSB in Procedure 8, “Merging and phasing preparation”. Set Solution to something reflective of the sample, *e.g.* `lyso_molrep.pdb`.

    c. (Optional) If the spacegroup was not defined during any of the previous steps, it will be necessary to set SG to use to Laue class instead of the default As is. The correct spacegroup should then be found during molecular replacement; it should have a significantly higher contrast than any of the alternatives. It will be necessary to rerun MOLREP and setting SG to use to the correct spacegroup.

    d. Click Run Now from the Run drop-down menu.

    e. Once MOLREP has finished, inspect the log file by double-clicking the job in the central panel of the main CCP4i window. At the bottom of the file, MOLREP prints the contrast. According to the MOLREP manual page [http://www.ccp4.ac.uk/html/molrep.html#score], a contrast greater than 3 means that MOLREP has definitely found a solution.

  - **Molecular replacement using Phaser**

    a. To run Phaser from Phenix, first start the Phenix graphical user interface.

    ```
    $ phenix &
    ```

    The first time the Phenix user interface is started, it asks for a project to be created. Set the Project ID to reflect the nature of the sample (*e.g.* `lysozyme`), and select the appropriate Project directory using the Browse... button. On subsequent invocations, new projects can be created by clicking the New project button.

    b. Choose Molecular replacement in the right part of the frame, and select Phaser-MR (simple one-component interface).

    c. Select the Input files tab. Add the search model (Add file and choose the coordinate file downloaded from the RCSB) and the data to search (Add file and choose the MTZ file with the merged intensities from the section called “Scaling and merging”).

    d. Set the Sequence identities to 100; in all cases here, the search model is expected to exactly represent the contents of the crystal.

    e. In the Search options tab click Other settings. Set Scattering Form Factor type to Use electron scattering. Click OK.

f. (Optional) If the spacegroup was not defined during any of the previous steps, it will be necessary to select All possible in same pointgroup from the Also try alternative space group(s) drop-down menu. The correct spacegroup should then be found during molecular replacement.

g. Click the Run icon and select Run now.

h. Check the translation function Z-score (TFZ) in the Run status tab when Phaser is finished. According to the PhaserWiki [http://www.phaser.cimr.cam.ac.uk/index.php/Molecular_Replacement] a value greater than 8 means Phaser has definitely solved the molecular replacement problem.

- **Direct methods with SHELX**

a. To prepare data scaled and merged by XSCALE for SHELX, go to xdsgui's XDSCONV tab, and convert the intensities to SHELX format by setting **OUTPUT_FILE** to *e.g.* **ketone.hkl SHELX**, and clicking run XDSCONV.

b. Using *e.g.* nano, prepare a text file called `ketone.ins` in the project directory with the following contents:

```
TITL Ketone
CELL 1.0 a b c alpha beta gamma
LATT latt_record
SYMM symm_record
SFAC C H N O
UNIT 20 0 0 0
NTRY 1000
HKLF 4
EOF
```

where *a*, *b*, *c*, *alpha*, *beta*, *gamma* denote the unit cell after merging, and the *latt_records* and *symm_records* can be generated from the relevant spacegroup using *e.g.* https://cci.lbl.gov/cctbx/shelx.html. The **SFAC** and **UNIT** cards inform SHELX about the expected composition (here: 20 carbons and no hydrogens, nitrogens, or oxygens). **NTRY** limits the number of trials, and **HKLF 4** indicates that intensities are used instead of structure factor amplitudes.

c. Run SHELXT

```
$ shelxt ketone -a -m1000 -y
```

to try all compatible spacegroups (`-a`) in 1000 dual-space iterations (`-m1000`) and to look for chemical spacing (`-y`). Different solutions will be named `ketone_a.res`, `ketone_b.res`, *etc.*, as characterized in the **shelxt** output. The solution can be checked by loading the output `ketone_a.res` into *e.g.* coot. The shelxt command will have to be repeated until a sensible solution is obtained.

d. A solution may be further refined by copying *e.g.* `ketone_a.res` to `sol1.ins` and `ketone.hkl` to `sol1.hkl`. Then

```
$ shelxl sol1
```

Reference statistics for different samples are given in Table 5, "Reference values for molecular replacement".

Table 5. Reference values for molecular replacement

Dataset	Lysozyme		Proteinase K	
	tvips2smv	img2px	MOSFLM	XDS
Symmetry	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$
Contrast (MOLREP)	26.44	19.27	35.87	32.07
Top LLG (Phaser)	1432	1702	6160	6350
Top TFZ (Phaser)	33.3	36.4	69.0	70.8

Refinement

Procedure 12. Refinement and validation

1. Macromolecular refinement can be done either with REFMAC from CCP4 or phenix.refine from the Phenix suite. phenix.refine is often slower than REFMAC, but offers a richer graphical user interface with ready access to intuitive graphs.

   - **Refinement with REFMAC**

     a. To use REFMAC for refinement, choose Run Refmac5 from the Refinement task in the CCP4Interface. Use the Browse buttons to assign paths to the MTZ in and PDB in input files, and choose where the MTZ out and PDB out files are to be written. Output lib will need to be assigned, too, even though it is not used here.

     b. Select Run&View Com File from the Run drop-down menu. This will open a window where the command line arguments and the input script can be edited. The REFMAC input scripts consists of keywords, subsidiary keywords, and values; to aid readability, continued lines are indicated by a minus (-) character. To enable electron scattering factors, open a new line before any keyword, and enter **`source EC MB`**. Then click Continue without display.

   - **Refinement with phenix.refine**

     a. To refine from Phenix, first start the Phenix graphical user interface.

     ```
     $ phenix &
     ```

     b. Choose Refinement in the right part of the frame, and select phenix.refine.

     c. Select the Input data tab. Add the molecular replacement solution (Add file and choose the PDB file produced by MOLREP or Phaser; the Phaser output will be found in a directory called `phaser_N`, where *N* is the number of the Phaser job) and the data to refine against (Add file and choose the MTZ file with the merged intensities from the section called “Scaling and merging”).

     d. Set Scattering table to electron in the Other options section.

     e. Click the Run button and select Run locally.

2. (Optional) For the first refinement after molecular replacement it often makes sense to include a rigid body refinement step early in the refinement process. In phenix.refine, this is accessed from the Rigid body check box in the Refinement settings tab; for REFMAC, there is a rigid body refinement option in the top, left drop-down menu. It may also be a good idea to run a few more cycles than normal, given that this is the point where the residual between model and data is the largest.

3. When refinement is done, inspect the R-factors. After the first round of refinement, they may be high. R-factor distributions derived from the structures deposited to the PDB can be viewed using

   ```
   $ phenix.r_factor_statistics 2.1
   ```

   on the command line. The command above displays the statistics for resolutions around 2.1 Å.

4. At this stage, it makes sense to visually verify the molecular replacement and the initial refinement.

   ```
   $ coot &
   ```

   This will start Coot. Chose Open Coordinates... from the File menu and select the refined model. For refinement in phenix.refine, this will generally be the only file ending in `.pdb` in a directory called `Refine_`*N*, where *N* is the number of the refinement job; for SHELX, it is the file that ends in `.res`. Similarly, open the phased MTZ file (Auto Open MTZ... in the File menu) found in the same directory as the refined model, or the file that ends in `.fcf` for SHELX.

   Step through the model, look for discrepancies between the density and the model. Further validation options are available under Coot's Validate menu. Alternatively molprobity can be accessed via http://molprobity.biochem.duke.edu.

5. Validation typically involves cycling the refinement program with visual inspection in Coot. It may also turn out that the data needs cutting, because even though the data processed fine, it did not refine so well. Often the model can be significantly improved by:

   - Automatically modeling waters. Both phenix.refine and REFMAC have the option to automatically add and remove water molecules.
   - Turning on weight optimization for temperature factors and/or stereochemistry terms. This is automated in phenix.refine, but not in REFMAC.

Reference statistics for different samples are given in Table 6, "Reference values for initial refinement.".

Table 6. Reference values for initial refinement.

Dataset	Lysozyme		Proteinase K	
	tvips2smv	img2px	MOSFLM	XDS
REFMAC				
R_{work}/R_{free}	27.52/32.66	24.81/29.97	24.51/27.09	23.88/26.65
Bond/angle r.m.s.d. (Å/°)	0.0097/1.5666	0.0068/1.3817	0.0076/1.4196	0.0085/1.4551
phenix.refine				
R_{work}/R_{free}	21.73/30.30	20.94/28.43	21.63/25.00	21.76/25.06
Bond/angle r.m.s.d. (Å/°)	0.008/0.991	0.008/0.992	0.007/0.914	0.008/0.946

Output from MOLREP was refined with REFMAC, output from Phaser was refined with phenix.refine.

A. The `data` partition

The sample directories on the data partition of the USB stick have several subdirectories, which contain intermediate results of data processing.

`raw`	The raw data as written by the camera system. Note that large datasets may be chunked such that a dataset is spread over several files.
`integration`	Integrated but unmerged intensities produced by iMosflm or XDS. Each dataset yields one MTZ file for iMosflm, or one `XDS_ASCII.HKL` file for XDS.
`merging`	Scaled and merged intensities produced by AIMLESS or XSCALE. There is one MTZ file, or a pair of `.ahkl` and `.hkl` files, for each sample. This directory also contains files output by SFTOOLS for refinement in REFMAC.
`rcsb`	Where applicable, this directory contains the search model for molecular replacement, downloaded from http://www.rcsb.org. The model coordinates are in PDB format, and the structure factors in CIF as well as MTZ format. The latter can be used to duplicate the set of free observations during merging.
`mr`	Molecular replacement solutions in PDB format from MOLREP and Phaser.
`refinement`	Model coordinates and phased reflections in PDB and MTZ format, respectively. After further refinement and validation, this is what would be deposited into the PDB.

Additionally, the partition has a directory with publications on MicroED.