Analysis tab user guide

Contents

PDF version

This guide provides a brief overview of the GMCA data processing capability recently added to the JBluIce tab Analysis. The new tab is designed to provide users with a more interactive and uniform interface to access a variety of crystallographic data analysis tools. This guide is mostly presented in a Q and A form so that you may jump to the relevant sections.

Overview of the gmcaproc pipeline and directory structure

The files for gmcaproc are located at the gmcaproc sub-directory. Briefly, the gmcaproc pipeline without specifying space group works as follows:

- Data sweeps are processed independently of each other in space group P1 using XDS in the **XDS-p1** directory (in parallel).
- From the pointless-determined space group for each sweep, the corresponding Laue group is chosen.
- The sweep that gives rise to the best overall I/sigmaI value will be chosen as the "correct" solution.
- CORRECT will be rerun with the "correct" solution as reference.
- All sweeps are scaled together in directory **scale-1** using XSCALE with no resolution cutoff applied, and statistics recalculated in aimless, with and without a resolution cutoff. Data for each wavelength (up to 4 significant digits) are written as separate files with w-x.xxxx as part of file-name.

If the optimization option is chosen for the gmcaproc pipeline without giving a space group, the above steps will run first followed by additional steps:

- The orientation matrix of the "correct" solution will be copied to **XDS** directory for each sweep.
- Data for each sweep are processed (integrate and correct) with the same orientation matrix.
- The process is iterated two more times with improved parameters for each sweep.
- The data are then scaled together in directory **scale-2**. The results are shown as a new row/entry (gmcaproc-ref).

If a space group is given by the user, then gmcaproc (and gmcaproc w optimization) XDS processing happens in directory **XDS-spgname**. It will process data in the specified space group, followed by 2 cycles of refinement, and then scaling in the scale-1 directory. The optimization flag has no additional effect.

All other supported third-party pipelines will run in the their own separate sub-directories. By default, the new pipelines (both processing and strategy calculation) assuming I+ $\neq$ I- (i.e. data may contain an anomalous signal). If you want to calculate results only for native data, please toggle on "native data" button in the processing dialog. For data processed automatically, the user can tell the program how to treat I+/I- by setting it in the "Collection" tab (under "XDS Proc").

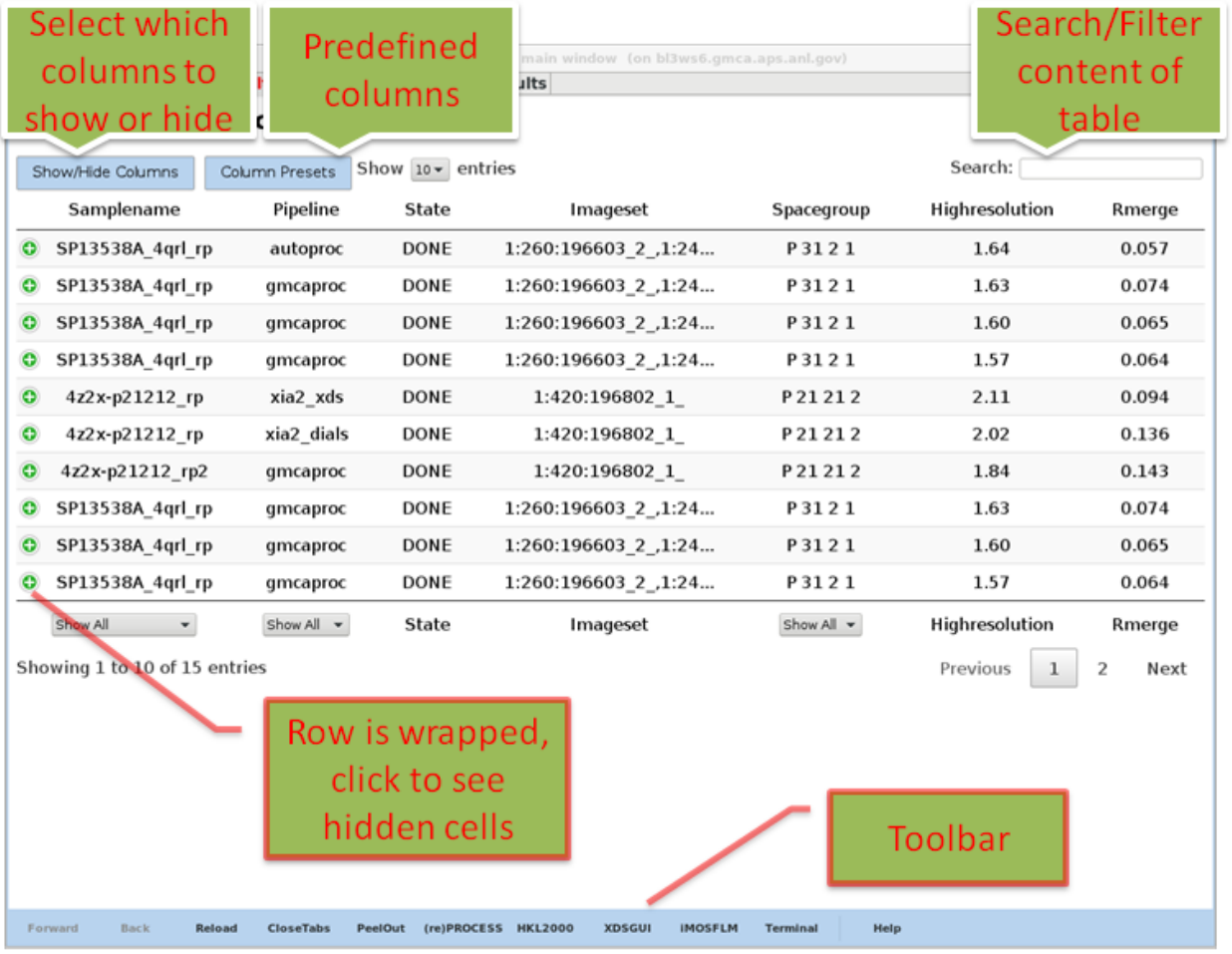

Figure 1: **Fig 1. GUI overview**

GUI

Main GUI

The new GUI is powered by an embedded browser, so it behaves similarly to that of a web browser. Operations revolve around two interactive tables (Fig 1 below), "Summary of Screening Results" for initial characterization (spot finding, indexing and strategy calculation, Matthews' coefficient, etc.) and "Summary of Processing Results" for data processing. Each table (initially empty) is populated by auto-processing jobs started by JBluIce or other processing jobs started by the user via the processing GUI dialog. Additional task-dependent tabs will be created on the fly, and can be closed. The number of columns shown will be automatically adjusted based on available screen real estate (window size). Thus, if there exists columns that do not fit the window, rows will be wrapped (as indicated by a green + sign at the beginning of each row). A user will need to click on the "green +" (or increase window size) in order to see the rest of row content.

A "Show/Hide Columns" button at the top of the table controls which of many columns are displayed (Fig. 2).

A few predefined presets (each defining a combination of columns to show) were also provided (Fig. 3). By default, the "Simple" preset is selected. A user may also start from one of the presets, and customs the views to his preference.

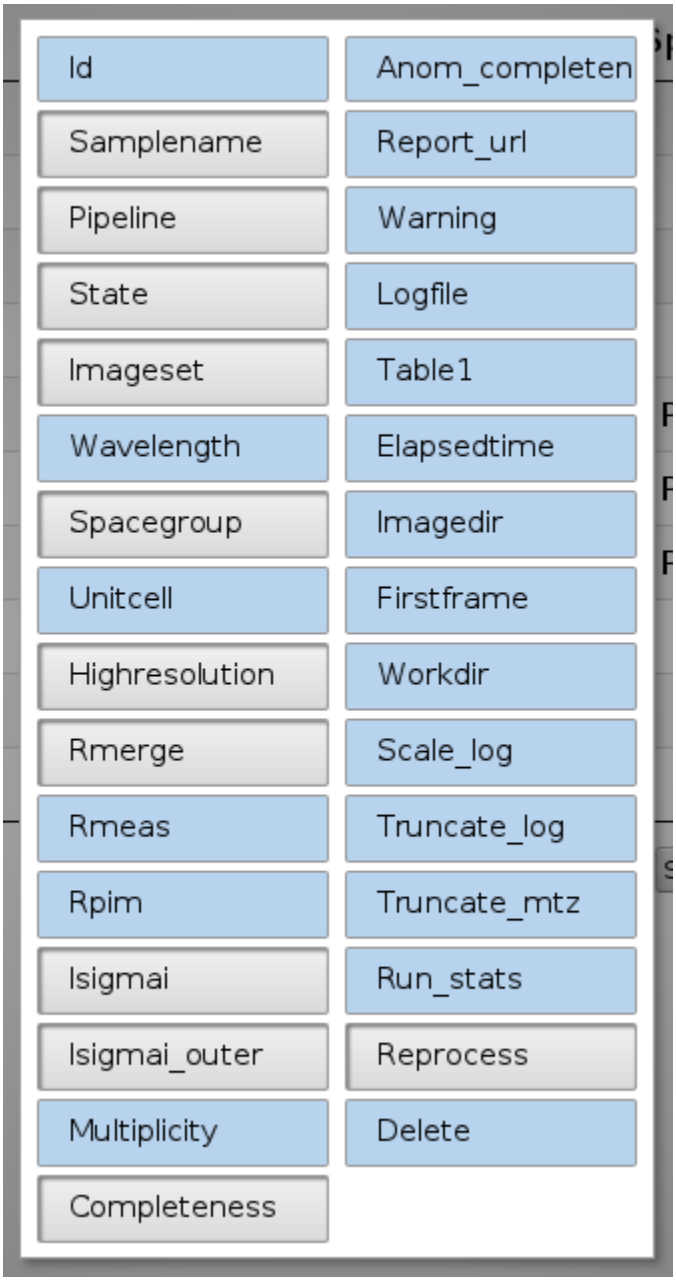

Figure 2: **Fig 2. All available columns for data processing. A user can select which column(s) to show via the Show/Hide Columns button. User choice will be saved until changed.**

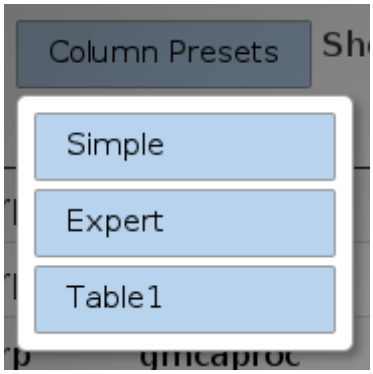

Figure 3: **Fig 3. Predefined column sets for easy access. By default, "Simple" is selected.**

Figure 4: **Fig 4a. Toolbar located at the bottom of the Analysis tab**

ToolBar and Context menu

The two Summary pages are augmented by a toolbar:

- Forward: history forward
- Back: history back
- Reload: regenerate the current page (refresh)
- CloseTabs: close all tabs except for the two Summary pages
- PeelOut: start a new Analysis window outside JBluIce window
- Process or Strategy: launches a dialog for processing data (also activated by keyboard shortcut Ctrl-X)
- XDSGUI: starts xdsgui in the current collect dir
- iMOSFLM: starts iMOSFLM in the current collect dir
- HKL2000: starts HKL2000 in the current collect dir
- DIALS: start DIALS gui in the current collect dir
- KAMO: quick launch for small wedge data processing using KAMO/XDS, for other cases, please use the Process button
- Terminal: Open terminal in directory based on the current location
- Setting: disable or change processing directory structure
- Help: brings up this document in a new tab (also activated by keyboard shortcut Ctrl-H)

Each page also has a context menu, which can be activated by a right mouse click

The function of each menu item is as follows:

- Reload page: reload the page manually (also activated by keyboard shortcut Ctrl-R), same as Reload

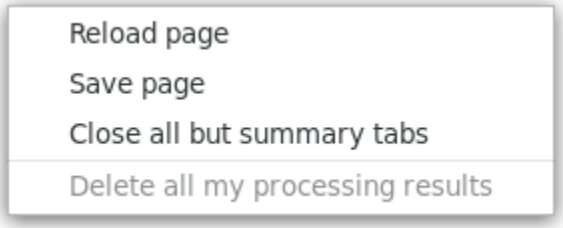

Figure 5: **Fig 4b. Context menu, activated by mouse right-click**

in Toolbar
- Save page: save the page into a file on disk. Results are saved periodically in the root collection directory. These html files can be opened using a browser.
- Close all child tabs: close all tabs except for the two Summary pages, same as CloseTabs in Toolbar
- Delete all my processing results: the processing statistics in both Summary pages are deleted, **NB this feature is disabled currently, user can delete processing row individually via the "Delete row" column (hidden by default)**.

User-driven processing dialog

A user may start a processing job through the processing dialog, as show in Fig. 5 below. It is accessible through the context menu (Fig. 4), or keyboard shortcut Ctrl-X, or the button "Start" of each row (Fig 1). Only the image directory field is mandatory. For each field, a hint may be provided if the mouse hovers over the field.

Directory explorer and integration of third-party applications

The new GUI provides an integrated file explorer for inspecting the processing results without leaving JBluIce. This function is accessible for each processing job by clicking the hyperlinks "Open" (Fig 1). A new tab (located to the right of the summary pages) will be created for a file browser. For crystallographic files, appropriate applications may be launched if the browser cannot handle the format. For example, if an image file is clicked, JBluIce will launch adxv for that image.

Detailed data processing report

For pipelines that provide detailed reports about the data processing (such as autoPROC and xia2-dials), the html report can be accessed by clicking the "Full report" hyperlink for each successfully completed job.

Questions and answers

General

Can I run the data processing GUI by itself, for example, as a second day user?

The analysis tab in JBluIce does run by itself, on any gmca bl computer, open a terminal, type

```
> jbluice-process
```

How does JBluIce start autoprocessing for me?

JBluIce runs the new gmcaproc pipeline at (25%), 50% and 100% checkpoints within each data collection run. At the end of the data collection, fast_dp, xia2-dials and autoPROC (only for academic users) pipelines will process and scale all data of the run. If a user pause data collection, fast_dp will run at the pause point (gmcaproc will run if the run contains multiple sweeps/segements), and autoPROC and xia_dials if pause happens further into the data collection.

All four indexing programs (mosflm, dials, labelit and xds) will be run automatically for each crystal screened, or for snapshots in Collection tab 0. All auto-processing files are located in the **process** sub-directory of the collection directory. Results will be displayed in either the Processing page or the Screening page with the newest results on the top. Background jobs are checked periodically by JBluIce, and each page will update itself when it detects state changes. From the 2019-1 run, the frequency of the periodic update of processing

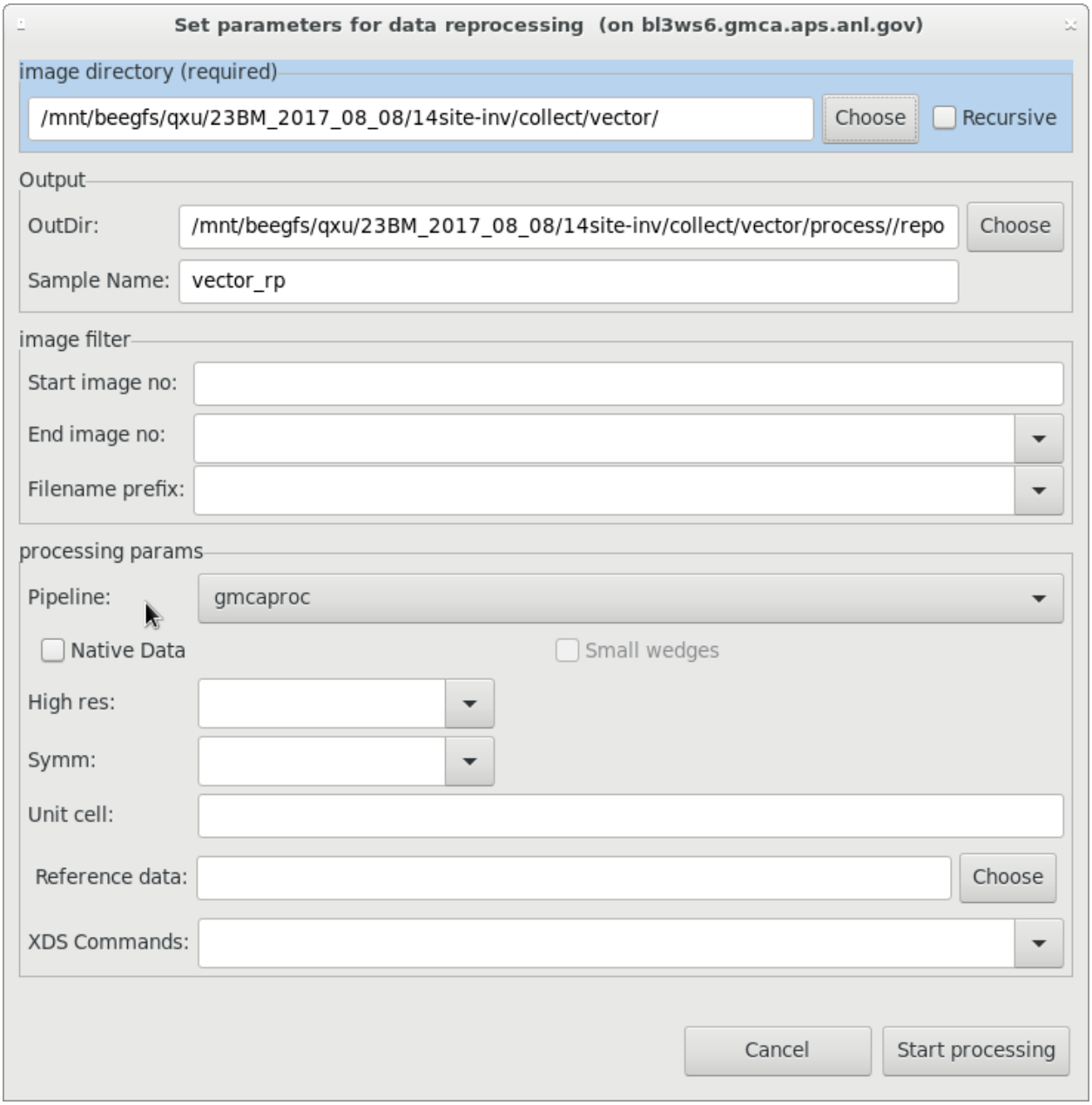

Figure 6: **Fig 5. Data processing dialog**

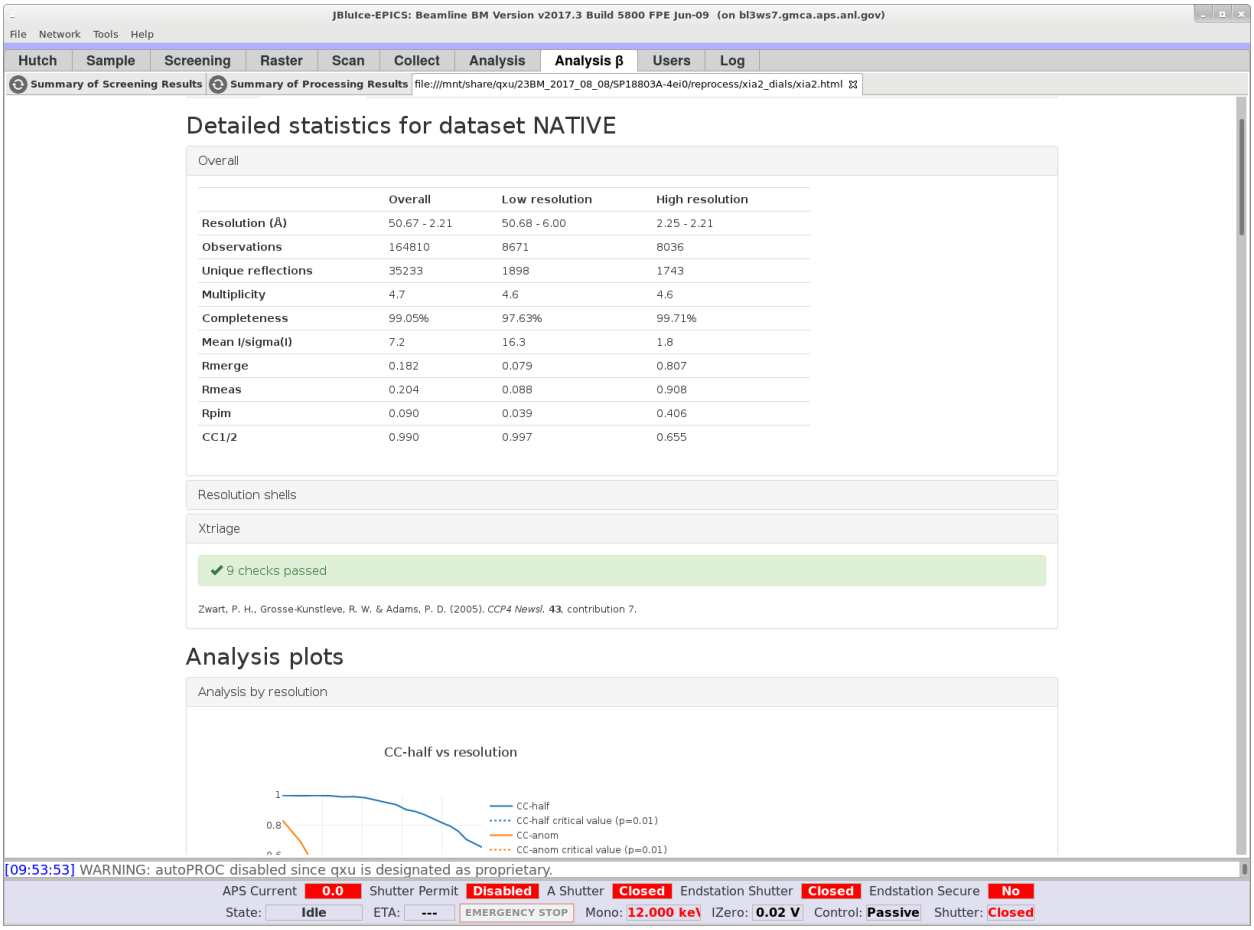

Figure 7: **Fig 6. Xia2 DIALS report in jBluIce**

results in jBluIce will be reduced if the Analysis Tab is NOT visible. It will switch to fast update if a user changes the focus to the Analysis tab.

Users can (re)process any of their data on GMCA disks using the same or additional pipelines with the provided GUI. The results are displayed together with those of the auto-processing.

Can I disable auto processing, or relocate them to somewhere else?

By default, jBluIce put processing files close to the data/images of each crystal. Alternatively, jBluIce offers the options to put all processing files for all crystals together, under the collection root (23IDx...) under a directory "process". autoprocessing can generate a lot of files. If you prefer not to use auto processing at all or want to choose the alternate of directory structure for processing files, you can do that using Tools/Options/Options or using Settings located in the bottom toolbar of the Analysis tab.

How are the result pages are updated?

All computing jobs are submitted to a queue waiting to be executed. The job status will not show until the submitted job starts running (usually within 10s or so). A background process checks for changes in result pages at fixed intervals. If the analysis tab is visible to the user, the interval will be faster (2sec), otherwise, the interval is set to 1hr (to minimize the impact on other jBluIce operations). If a change in job state is detected, the result pages will be regenerated/updated.

Browsing Results

How do I check processing results when I get home?

The processing results in JBluIce will be updated and saved automatically whenever a job status is changed. They correspond to three files present in the data collection root directory 23ID..._cbf (these files are automatically backed up by the GMCA rsync script).

- Summary-of-Processing-Results.html, all data processing results. This file contains all the information user need to look the processing results when they get home, including file locations. User can use a web browser to open this file (the relative links in it will only works if the original directory structure is kept). Please note that additional columns (e.g. truncate, working directory) may be hidden, they can be unhidden using the "Show/Hide Columns" button at the top left corner.
- Summary-of-Screening-Results.html, all indexing and strategy results (from the Screening Tab or Collect Tab 0).
- datasets.txt, a text summary of datasets collected. This file is used by KAMO.

Additionally, user can also save the result page explicitly by right-clicking on the corresponding page and then selecting "Save page".

I clicked on the "table1" button (or some other button) on the top, nothing showed up.

Each row will wrap automatically dependent on the size of the window. If a row is wrapped – indicated by a green + sign at the beginning of the row, you will need to click on the green + sign at the left end of the row to unwrap and see the hidden content. If you have a big screen, maximizing the window may also help.

Due to limitation in screen real estate, additional fields are hidden by default, which are controlled under the "Show/Hide Columns" button (a field will show if the corresponding button is pushed down). A few pre-defined views are provided on its right side. **Please notice that your selections of views or Show/Hide columns will be remembered by jBluIce during auto-reloads until you change them next time.**

There is too much information, how can I locate or filter results?

Use the search box on the top right corner. You may search for words as long as they are present in one of the fields. For example, type "Final" in the search box will show only results with "Final" in the directory name, ie. filtering out immediate processing results. Type "autoproc" will only show autoPROC results. You may also try multiple words separated by spaces.

Indexing and Strategy

How do I run or rerun indexing images ?

The indexing/strategy request must be launched from the Summary of Screening Results page (by clicking the Strategy button at the bottom toolbar, or the "Reproc" button of each row). It has the same interface as data processing, except for the pipeline selection field. Please note that strategy calculations in mosflm_strategy, dial_strategy and labelit_strategy options are carried out by MOSFLM. If the automatically picked solution is not what you want, you will need to rerun it with more specific parameters, such as the space group.

My space group was not chosen by indexing program...

Indexing pipelines usually choose the highest metric symmetry within some penalty threshold. Your true space group is likely a sub group of the default chosen solution. If you know the space group, it is necessary to rerun the indexing pipeline by specifying your space group explicitly so that the indexing program uses it to provide better strategy results. MOSFLM calculations usually take $<$ 10-30 sec.

How was unit cell content was calculated?

If the indexing is deemed successful, the indexing pipeline will attempt to analyze the asu content (+/- 50 aa) by performing matthews_coeff calculation assuming that (1) you have protein, (2) the solvent content is ~50%. Please note that this is an estimated value, more accurate estimation requires more specific knowledge about macromolecules (e.g. size and type).

The resolution estimation in Screening Results is not good.

During initial characterization stage (indexing), only strong spots, which tend to be dominated by low resolution reflections, are picked. Thus, the resolution estimation during screening should be taken with a grain of salt. It is not uncommon to see the final resolution being more than 1 Å better. Additionally, ice spots could result in over-estimation of resolution.

Why is the oscillation angle (osc) always 0.2 deg during strategy calculation?

We have set the upper bound of the osc angle as 0.2 deg. We recommend that our users collect thinner-sliced data to take advantage of the newer detectors (e.g. better signal-to-noise ratio). Please note that, for thinner-sliced data, you may need to reduce the exposure time accordingly to avoid over-exposure.

Can I index raster images?

At the GMCA beamlines, the raster images are STILL images with oscillation angle being 0 deg (much weaker as well). This causes problems for MOSFLM, XDS, and LABELIT. However, DIALS does not seem to have such restriction, so you could try to run indexing using the dials_strategy option in the Screening page.

Can I run indexing using one, three or more images?

Conventional wisdom suggests that two images 90 deg apart often produce better indexing results. However, it is possible to index using other than 2 images. The trick is to filter out unwanted images using start & end numbers, or file-name prefix.

Processing

How do I process or reprocess a dataset in the GUI?

On the Summary of Processing Results page, click "Process" in the bottom toolbar (keyboard shortcut Ctrl-X), select the directory that contains the image data, and optionally edit other fields to fine tune your input and click the "Start processing" button. You may get some hints about each input field by hovering the mouse over it. For some fields, such as space group symbol, you can check whether the typed input is valid by hitting the return key after inputting it. If you want to process a dataset using the same pipeline more than once (e.g. using different parameters), please make sure to define a unique output directory name for each run (otherwise, the processing files will overwrite each other).

The job will be submitted to a queue, and may not start immediately. The processing status will be checked periodically, and pages will be reloaded automatically upon state changes (You may also refresh the page manually by selecting Refresh on mouse right click menu, or typing Ctrl-R). Results are sorted such that results from the latest jobs will be at the top.

Another entry point for processing is through the reprocess button for each processing entry/row (click on "Reproc"; if the row is wrapped, first click the green + icon). This brings up the same interface but populated with parameters from the auto-processing.

How is the high resolution cutoff chosen?

The gmcaproc and fast_dp pipelines choose the high resolution cutoff based on the CC1/2 value. It will attempt to cut the resolution when CC1/2 drops below 30%. **If your diffraction is anisotropic, the resolution of the best diffracting axis will be chosen, even the overall CC may drop below 30%**. If you want to cut your data anisotropically, please refer to the UCLA anisotropy server and the STARANISO server (also part of the autoPROC). The gmcaproc pipeline also gives user the data before a resolution cutoff was applied, i.e. to the corner of the detector. These files are located in sub-directory scale-1 with files containing ALLDATA as part of file-names. For the autoPROC pipeline, a similar cutoff scheme as that of gmcaproc is chosen by default, but is more conservative. For xia2 pipelines, the default was used.

I have magic XDS keywords I would like to use.

For difficult cases, additional XDS paramters may be necessary. You can add additional XDS paramters in the XDS command field with the following format: KEYWORD1=xx;KEYWORD2=yy (i.e. key=value pairs separated by semicolon). Alternatively, you can modify the generated XDS.INP file to add your keywords and rerun XDS. When a user is exploring the working directory, XDSGUI will be started automatically when a user clicks a XDS.INP file. Results for manual XDS processing will not be tracked.

Multiple-crystal Strategy and Data Processing

How do I perform a multi-crystal strategy calculation using the new interface?

Multi-crystal strategy calculation using XDS is implemented. Assuming that you have a processed (partially complete) dataset in XDS format and have collected one or more snapshot images of the current crystal. The

detailed procedure is as follows:

- Mount and center a new crystal
- Collect one or more images using the Collect tab or the Screening tab
- Open Analysis/Summary of Screening Results and check whether the crystal was indexed correctly
- If yes, click on "reproc" button of the corresponding crystal, in the "reference dataset" field, choose the partial data with the directory browser, in the pipeline field, choose xds_strategy, and submit the job. The partial data is ideally in the XDS XDS_ASCII format, but Scalepack and mtz formats are also accomondated. The strategy type (anomalous or native) must match the partial data (i.e. if you provided a partial data with anomalous data, anomalous strategy will be calculated).
- Once the job finish, you can export the default strategy using the Export button or select a different set of start/end parameters from a drop-down menu. By default, the result for 90% overall completeness (if not achievable, then 10 degree total rotation)
- If the crystal does not index as expected, you may still us the procedure above to see whether including symmetry and cell parameters helps. Alternatively, you may specify which image to use by giving a more specific File-name prefix, otherwise, the program will use the last collected image(s). For example, if you have three images, A1_0_000001.cbf, A1_0_000045.cbf, A1_0_000090.cbf, and you define File-name prefix as A1_0_000045, the second image will be used in the strategy calculation.

I want to process data from multiple crystals.

The current pipelines will process the data in a single directory (as long as it is the same type of crystal, otherwise the program will fail at the scaling stage). gmcaproc can also find images to process recursively starting from a top level directory. To select images, multiple prefixes (format: prefix1;prefix2 etc) may be defined in the prefix field, furthermore, subdirectories can be also selected via either inclusion or exclusion. Alternatively, you may create a temporary directory and create symbolic links in it to all the images you want to process, and select the new directory containing links as the input image directory for processing.

For processing many crystals (e.g. 10s-100s), please try the KAMO pipeline. The provided GUI in jBluIce, activated by Process button and select kamo_dials or kamo_xds as the pipeline, sets things up so that KAMO can run on GMCA computers and data collection/processing directory structure.

For data processing of multiple crystals during data collection (aka live mode), user can just click the KAMO button in the Toolbar located at the bottom of the Analysis tab. KAMO interface will launch (which may take some time), and monitor appearance of new datasets defined by a text file named datasets.txt at the collection root directory (generated by jBluIce) and process them. User can choose to merge datasets periodically. This can be done by selecting datasets (to be merged), and click "Multi-merge strategy" button to launch "Pre multi merge" dialog window, then choose group/symmetry/workdir and click Proceed button. A background process will be started to run the merging process on the behalf of the user. Each merging result is saved as a new entry in the Analysis table. It may be better to view these results using the Column Presets/Multi-Xtals view. **It is not advisable to close KAMO and restart it if there are processing jobs running since KAMO will lose track of those running jobs.**

Structural Determination (Experimental)

How are structure determination jobs run?

If there is strong anomalous signal present in the data, fast_ep will run automatically for gmcaproc and fast_dp pipeline. Currently, it only works with signal-wavelength data (SAD).

For ligand screening, dimple pipeline in CCP4 will be run if user provides the model information, which can be configured in the screening spreadsheet file. A new column with the name "ModelPath" must be

added, the case-sensitive filename of the apo model must be provided for each crystal. The filename can be in absolute path (e.g. /mnt/beegfs/user/model/model1.pdb), or just model filename only (e.g. model1.pdb). In later case, the model file must be placed in the user's Desktop or Downloads directory. Please note that the model must conform to dimple requirements, e.g. ligand(s) should be remove from the file beforehand.

A quick way for preparing the model and the spreadsheet files is to send the files to one's webmail account, and then save them to GMCA computers from one's webmail account. The ModelPath column can be editted before data collection of the corresponding crystal, there is no need to reupload the spreadsheet to jBluIce, as long as the filename and location of the spreadsheet is not changed.

Other Topics

Difference between gmcaproc and fast_dp?

gmcaproc and fast_dp are very similar for processing single-sweep data. The data are initially processed in space group P1, followed by space group determination by pointless (P1 corrected data for gmcaproc vs. P1 integrated data for fast_dp). One difference at the sweep level is images for spot finding and indexing: fast_dp uses only three small wedges, while gmcaproc uses the first half of the sweep. gmcaproc runs two CORRECT steps, one to generate P1 data for pointless, second CORRECT in the space group identified by pointless, while fast_dp runs only one correct step at the end (with only one CORRECT/scaling step it is more difficult to apply a high resolution cutoff). gmcaproc can handle multiple-sweep data from single or multiple crystal(s), multiple wavelength, or multiple sites, while fastdp can process only single-sweep data.

In short, fast_dp processes single-sweep data and may be slightly faster, while gmcaproc is more versatile to handle more complicated data collection schemes and formats (e.g. hdf5) at GM/CA, and is capable of producing "refined" datasets, i.e. two-pass integration and scaling.

old GMCAproc vs the new gmcaproc?

The new gmcaproc pipeline is completely rewritten in python to handle more complicated data collection schemes (e.g. multiple-wavelength in a single run), to accept multiple image formats (hdf5), and to incorporate more third-party pipelines. The previous GMCAproc pipeline uses XDS for space group determination and processes each sweep of data in serial manner assuming a common crystal orientation. The newer gmcaproc pipeline processes all sweeps in parallel independent of each other, and the sweeps are reconciled (re-indexed) together at the scaling stage. As a result, the new gmcaproc pipeline is a factor of # sweeps faster (if there is no limitation on computing resources).

If the optimization step for the new pipeline is chosen, additional steps of refinement (rerun of integrate-correct step with refined parameters) will be run by assuming the same orientation matrix of the reference sweep (the sweep with the best overall I/sigmaI value). As a result, if you must process your data using the same orientation matrix, you should reprocess the data by choosing the gmcaproc with optimization option without specifying a space group (it is not run by default for sake of speed).

In summary, the new pipeline is faster and capable of processing more complicated data (e.g. multiple-wavelength MAD).

JBluIce runs the new gmcaproc pipeline at fixed checkpoints during each Collect run, while the old gmcaproc pipeline starts processing using a more complicated logic that may appear random to users.

I found a bug or need a new feature.

Users are encouraged to submit bug reports (and feature requests) since they help us improve the software that benefits all our users. If you find a bug (or bugs) or need certain features in the software, please contact

Qingping Xu at qxu@anl.gov or your host, or provide your feedback on the user feedback form. In some cases, we may need to ask you to share some data, which of course will be used strictly for debugging purposes.