P211

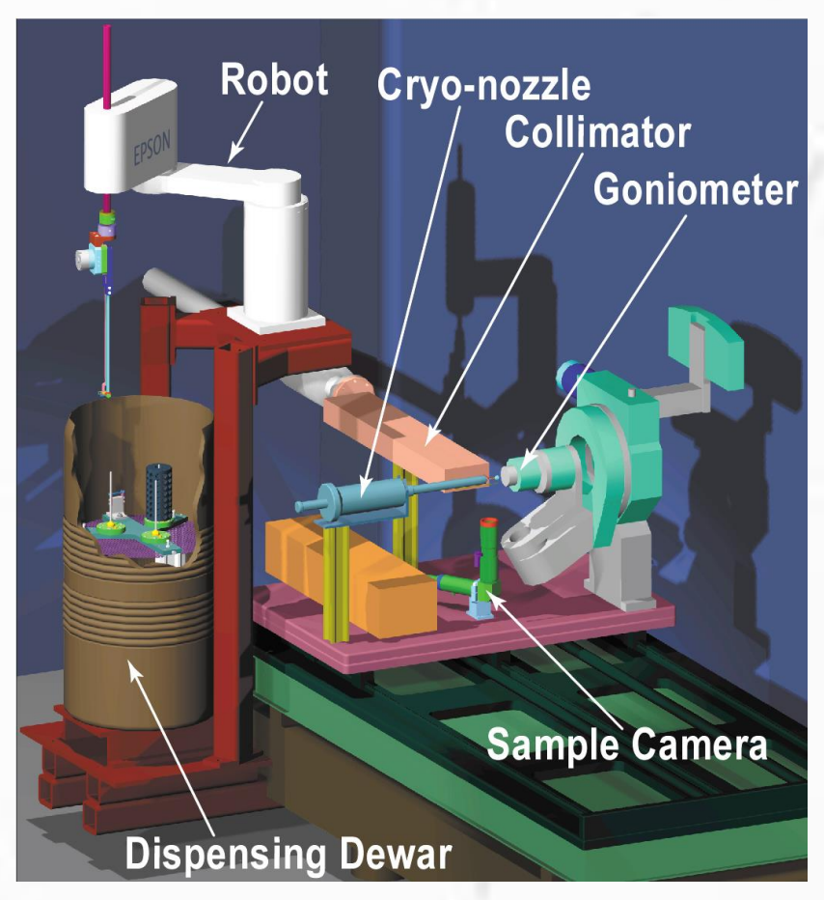
Crystal Ranking: Automatic Analysis of Screening Images

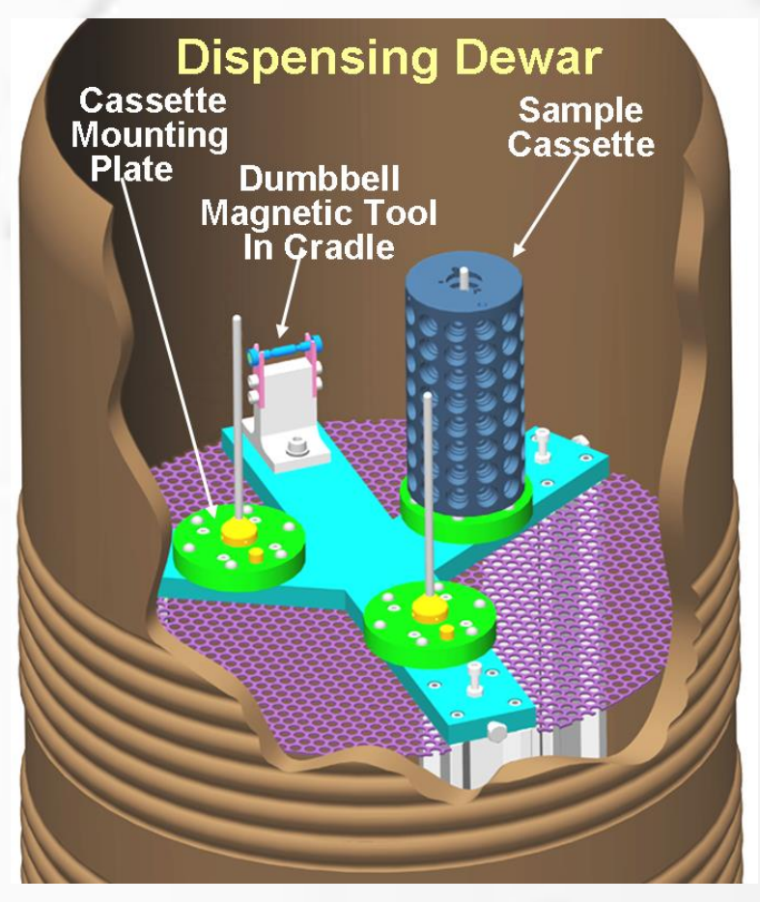
Qingping Xu[1,2], Zepu Zhang[1,2], Nick Sauter[3], Henry van den Bedem[1,2], Ana Gonzalez[2], Clyde Smith[2], Ashley Deacon[1,2]
[1]Joint Center for Structural Genomics, [2]SSRL, Stanford University, Menlo Park, CA, [3]Dept of Physical Biosciences, LBL, Berkeley, CA

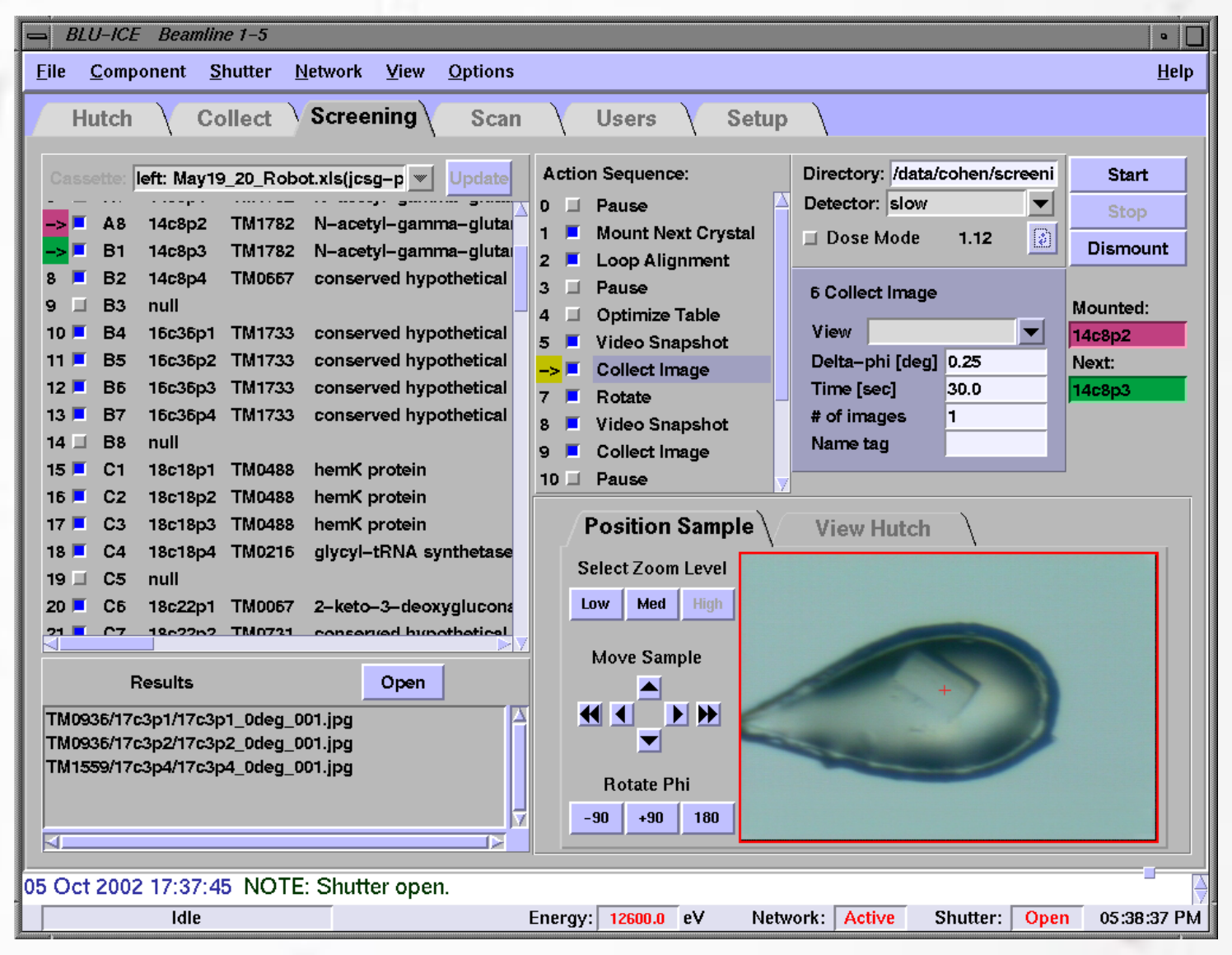
Abstract

The ability to screen crystals and find ones suitable for crystallographic studies is essential to any crystallographic project. The rapid adoption of crystal mounting robots at synchrotron beamlines allows automatic screening of hundred of crystals in a couple of hours. Quick and automatic analysis of the diffraction images from such screening is essential in order to reduce human labor and make efficient usage of beam time. The goals are to assemble and record the diffraction properties of screened crystals and select ones suitable for further structural studies. We attempt to address this problem by a two-step process. The first step is to conduct a statistical analysis of the diffraction images, through the identification of diffraction peaks and the detection of ice rings. The statistics include a resolution estimation, spot shape analysis, diffraction strength, spot split percentage, and ice ring strength/location. A score is assigned to each image based on these statistics. This step is implemented in a C++ library DISTL (Diffraction Image Scoring Tools Library). Autoindexing is then used to provide a more accurate picture of the crystal quality for promising crystals.

Automatic Crystal Screening at SSRL







Cassette: left: May19_20_Robot.xls(jcsg-p

->	A8	14c8p2	TM1782	N-acetyl-gamma-glutai
->	B1	14c8p3	TM1782	N-acetyl-gamma-glutai
8	B2	14c8p4	TM0667	conserved hypothetical
9	B3	null		
10	B4	16c36p1	TM1733	conserved hypothetical
11	B5	16c36p2	TM1733	conserved hypothetical
12	B6	16c36p3	TM1733	conserved hypothetical
13	B7	16c36p4	TM1733	conserved hypothetical
14	B8	null		
15	C1	18c18p1	TM0488	hemK protein
16	C2	18c18p2	TM0488	hemK protein
17	C3	18c18p3	TM0488	hemK protein
18	C4	18c18p4	TM0216	glycyl-tRNA synthetase
19	C5	null		
20	C6	18c22p1	TM0067	2-keto-3-deoxyglucona

Action Sequence:
0 Pause
1 Mount Next Crystal
2 Loop Alignment
3 Pause
4 Optimize Table
5 Video Snapshot
-> Collect Image
7 Rotate
8 Video Snapshot
9 Collect Image
10 Pause

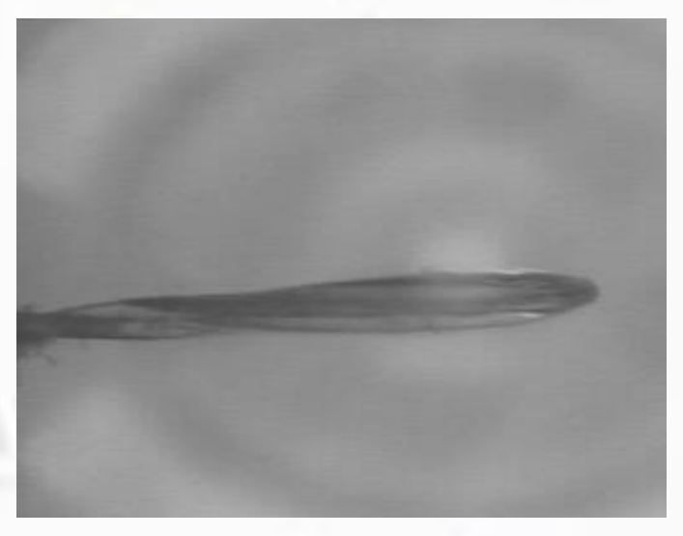
Directory: /data/cohen/screeni
Detector: slow
Dose Mode 1.12

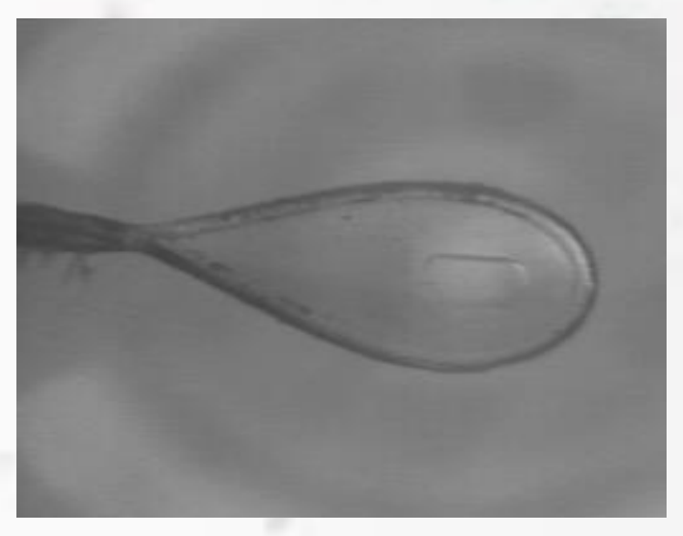
6 Collect Image
View
Delta-phi [deg] 0.25
Time [sec] 30.0
# of images 1
Name tag

Mounted: 14c8p2
Next: 14c8p3

Results
TM0936/17c3p1/17c3p1_0deg_001.jpg
TM0936/17c3p2/17c3p2_0deg_001.jpg
TM1559/17c3p4/17c3p4_0deg_001.jpg

05 Oct 2002 17:37:45 NOTE: Shutter open.
Idle Energy: 12600.0 eV Network: Active Shutter: Open 05:38:37 PM





Reference: Cohen, Ellis, Miller, Deacon, and Phizackerley. J. Appl. Crystallogr. 35, 720–726 (2002).

Stanford Auto-Mounter

- Easy shipping, storage and screening of samples
- Large-scale capacity (3 cassettes of 96 samples in hutch)
- Allows remote and unattended operation
  - Automated sample mounting
  - Automated sample alignment
  - Automated diffraction images
- Integration with BLU-ICE data collection environment
- Available to general users
- Efficient and robust, 3.5 min per screen, screens 285 samples in under 17 hr, 1% failure loop alignment

Standard Crystal Screening Protocol

After load crystal information into system database:

- Mount a sample
- Automatic loop centering
- Take a snapshot of crystal
- Rotate crystal by 90 degree
- Take a snapshot of crystal and collect another image
- Dismount sample and proceed to next crystal
- Computationally analyze the screened crystal when the next crystal is screened to improve efficiency (BLU-ICE and WEB-ICE at SSRL)

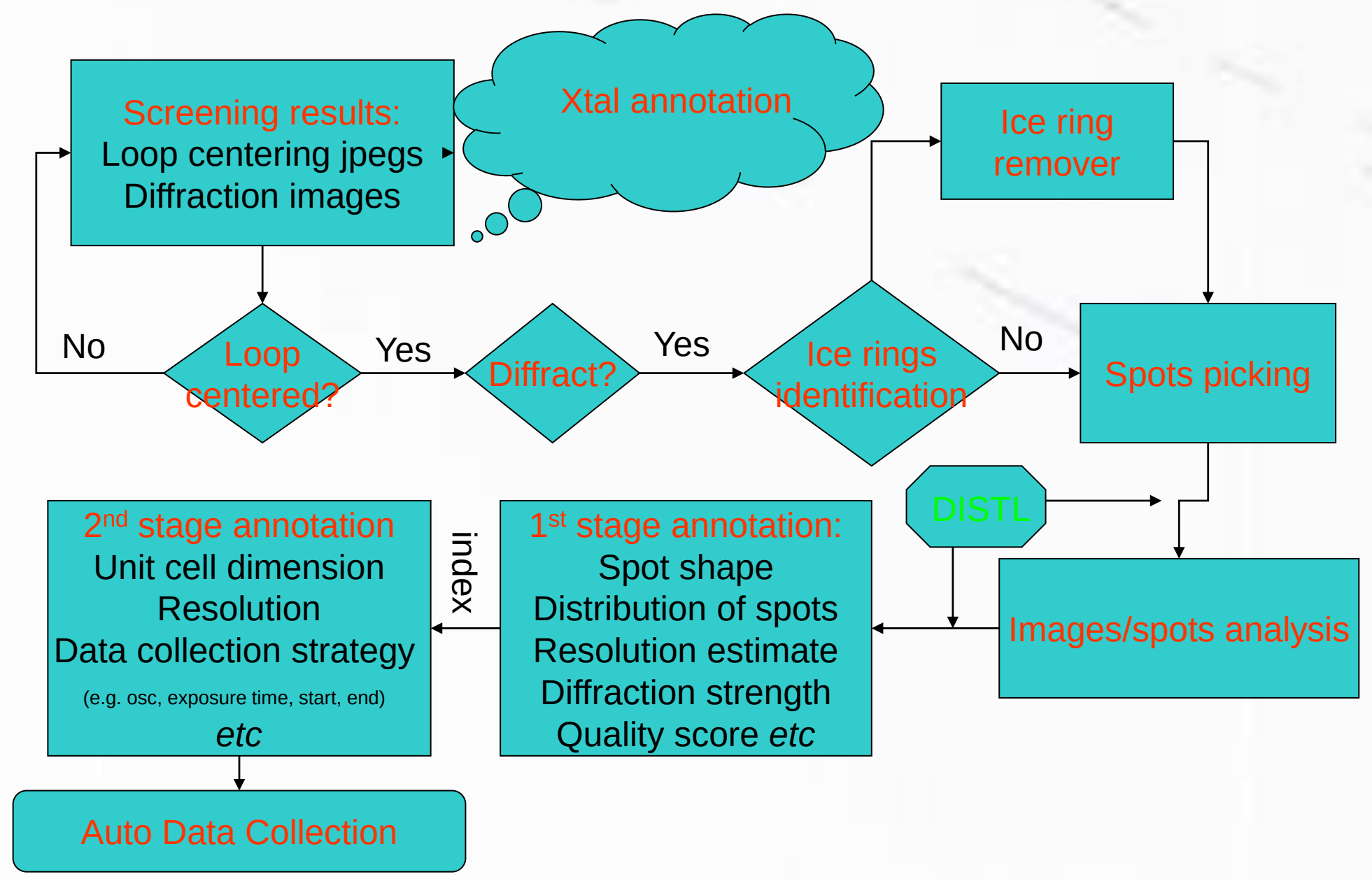
Manually process the failed samples, analyze all screening results and proceed with data collection.

DISTL: Diffraction Image Scoring Tools Library

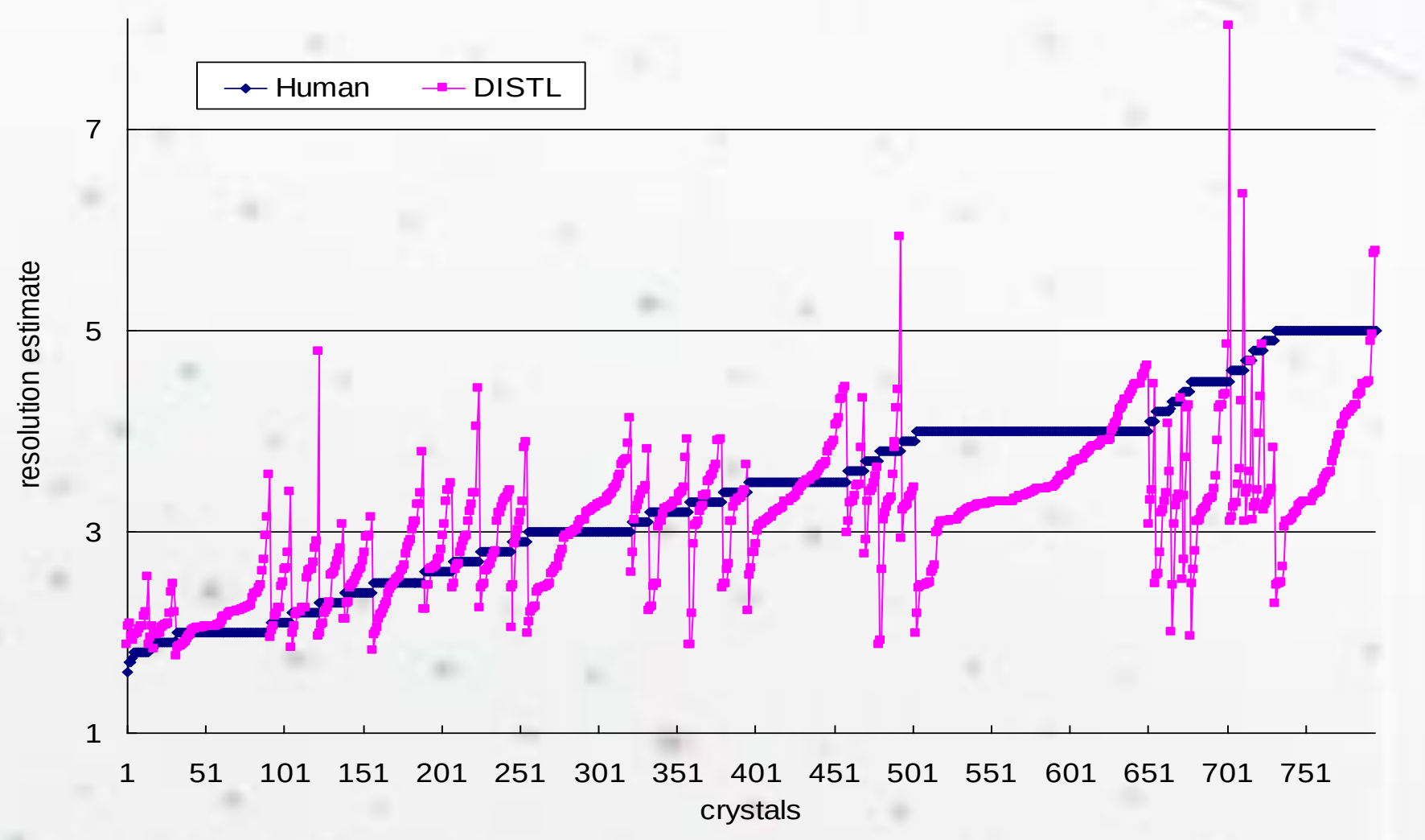
It is tedious and time consuming for human to look the large amount of screening images to find the best diffracting crystals. The development of automation technologies at the synchrotrons, such as installation of robotic sample mounting systems, has improved the efficiency of the beamlines tremendously, This aids in the selection of the best diffracting crystals by increasing sampling the number of crystals screened at the synchrotron sites.

- A object-oriented C++ library for diffraction image analysis
- Current capabilities
  - Read/write diffraction images
  - Robust spots picking and display
  - Identify ice rings
  - Output list of spots for external programs such as LABELIT, MOSFLM
  - Simple analysis of spots, shape, resolution, strength, distribution, etc
  - Score each image
- Results are presented in graphical and textual forms
- Analysis of large number of images are handle by scripts
- Developed by Zepu Zhang, Ashley Deacon *et al.* at SSRL.

Automatic Crystal Screening Annotation System



Results and Discussion



Comparison of resolution estimates between DISTL and human for crystals diffracting beyond 5 Å

Analysis of ~2000 crystals indicated that DISTL results are very promising. The software agrees with human decisions on whether a crystal diffracts in ~75% of the cases. For crystals diffracting beyond 5 Å, 98.5% of the diffracting crystals were correctly identified. For these crystals, the average difference in resolution estimation (defined as <|resolution$_{Human}$-resolution$_{DISTL}$|>) is less than 0.5 Å. The ice rings or ice spots can be identified correctly around 70% of the time, while strong ice rings can always be correctly identified. The current DISTL/SPOTFINDER program is capable of providing a list of good candidates for further investigation, although human intervention is still necessary to double check the results to prevent false positives/negatives.

DISTL/SPOTFINDER:

- C++ library, component or platform for further development
- Generate a more robust set of diffraction peaks that can be used for auto-indexing. It has been incorporated into Lawrence Berkeley Lab Index Toolbox (LABELIT, Nick Sauter *et al.*, ALS)
- Reliable resolution estimations, agrees well with human
- Can identify and locate ice rings automatically
- Analysis and scoring crystals without indexing, useful for crystals which cannot be indexed automatically
- Give an initial score for each image based on the following parameters: resolution, diffraction strength, ice rings, spot shape

XRANK:

- External script to evaluate screening results after screening is finished,
- Parse the screening data directory tree, run DISTL/SPOTFINDER and collect statistics
- Output statistics in format suitable for external programs such as EXCEL when can then be sorted and analyzed by user
- Run autoindex for crystals with good initial scores, calculate 2nd generation score for crystals indexable (ongoing)
- Determine data collection strategy (ongoing)
- Fast, can be run in parallel

BLU-ICE and WEB-ICE (Ana Gonzalez *et al.* SSRL):

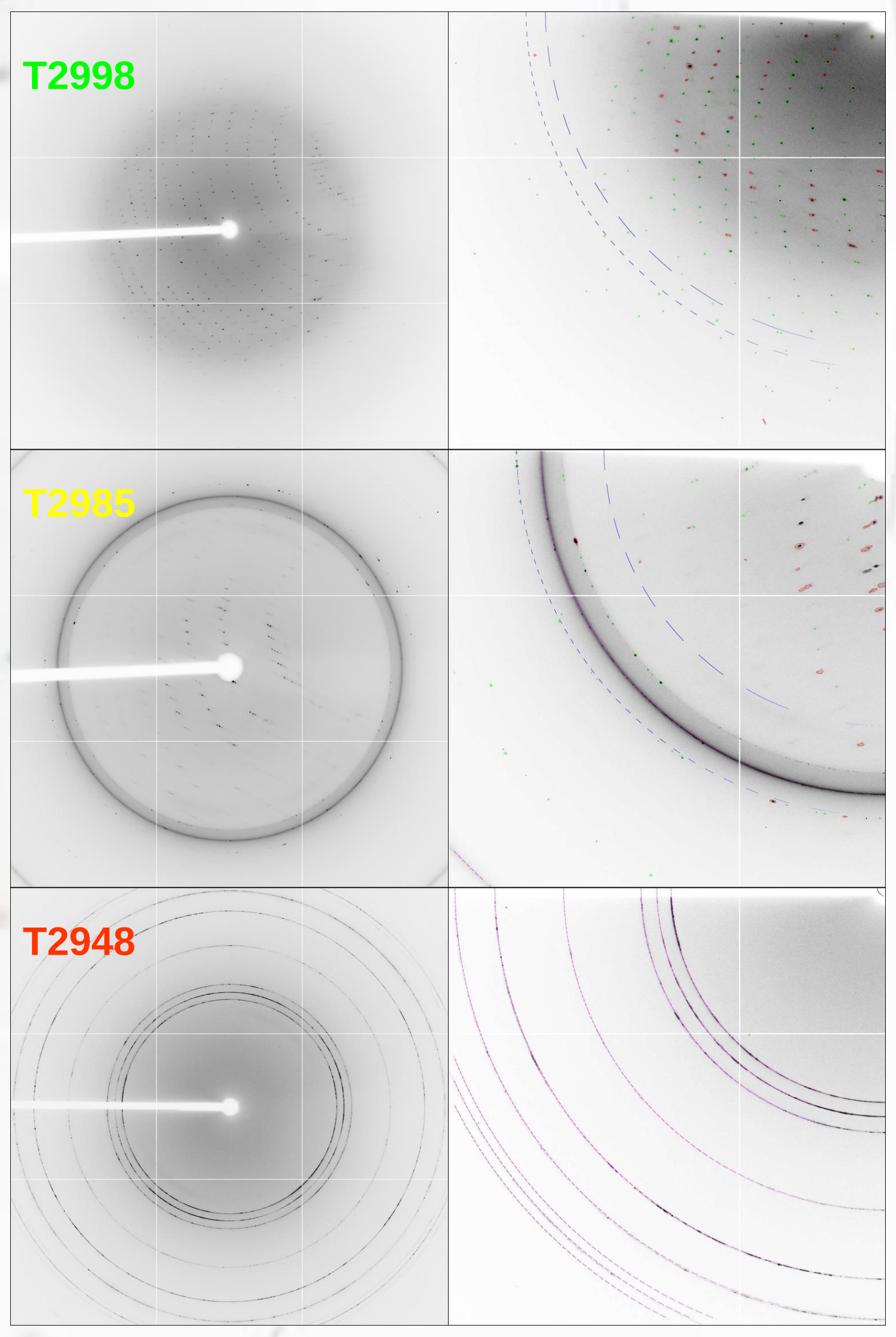
- Screening tab, analyze each crystal right after it was screened, results are available online immediately
- Results are presented on the web and allows remote access
- All above functionalities
- For general users

PROBLEMS:

- Resolution estimation can be difficult for crystals barely diffracting
  A few spots generated by crystals diffracting vs “spots” from noise

An Example: MB3402A

	Human			DISTL		
UniqueID	resolution	icerings	spot quality	resolution	icerings	spot shape
T2998	2.5	0	8	2.55	0	0.7566
T2993	2.7	5	6	3.101	7	0.5029
T2938	2.8	0	6	3.378	0	0.5862
T2997	2.9	0	7	3.837	0	0.6922
T2996	3	0	7	3.277	0	0.733
T2988	3	1	8	3.381	0	0.5531
T2940	3.1	0	8	3.37	0	0.6357
T2937	3.2	3	7	99	3	0.6972
T2991	3.2	7	8	3.101	7	0.7798
T2989	3.5	2	8	4.124	3	0.6775
T2959	3.7	0	8	3.393	0	0.7779
T2985	4	5	5	3.804	1	0.5365
T2986	4	6	8	4.655	5	0.3126
T2987	5	3	7	5.76	4	0.6709
T2956	6	0	6	3.212	0	0.6581
T2953	6	0	4	4.706	0	0.4933
T2943	7.5	6	5	99	5	0.7814
T2941	8	5	8	4.246	3	0.7308
T2942	8	5	6	4.278	3	0.7879
T2983	9	6	2	99	3	0.7714
T2995	10	0	6	99	4	0.7705
T2984	12	6	2	99	2	0.5932
T2944	20	4	7	99	2	0.3792
T2939	20	5	7	99	3	0.3123
T2945	20	5	6	99	3	0.6004
…	99	0	0	99	0	…
T2990	99	0	0	99	0	0.7778
T2950	99	2	0	99	1	0.6249
T2946	99	5	0	99	4	0.7659
T2947	99	6	0	99	9	0.3652
T2949	99	6	0	99	9	0.2273
T2948	99	7	0	99	9	0.6899



Comparison of DISTL results to human evaluations indicated that DISTL is able to identify good candidates for data collection. In this example, a total of 38 crystals were screened for target MB3402A (2636322). The structure was solved with three-wavelength MAD using the crystal T2998 (1VLI)

From images to Structure: Development of fully automated structure determination procedures (autoXDS)

Although solving a structure by an experienced crystallographer with SAD/MAD data is relatively straightforward once crystallographic data were collected. However, there are few fully automated procedures that can process the data and solve a structure without any human intervention in a consistent fashion. This is because most current crystallographic packages are not yet powerful enough to use in a “black box” fashion and the parameter space need to be explored in a structure determination process can be pretty large. As a result, human intelligence is often needed to steer the process towards the correct direction. In order to further expand the capability of Xsolve, we have implemented a prototype procedure (autoXDS) that automatically processes diffraction images, phases the structure and produces a good initial trace with minimum of user input. autoXDS is a Perl script (~3000 lines) drives popular crystallographic packages XDS (data processing), SHELX (heavy atom position and phasing), autoSHARP (heavy atom refinement and phasing), RESOLVE and ARP/WARP (tracing) and CCP4 utilities.

For Se-MAD/SAD structure determination, the only user inputs required are the directory name where the images are located and a protein sequence (or a target name for a JCSG target). The script scans the image directory and decides on how to set up data processing with XDS. First, a batch of images (usually the ones collected first) are indexed and integrated in a reduced cell in space group P1, the Laue space group used for integration is determined automatically by scaling the resulting P1 integrated data in different point groups according to the indexing result. The resulting data are used as a reference dataset for all following processing. The Laue group and refined unit cell are then used to integrate all the batches. Reindexing is checked systematically by finding maximum correlation between an integrated dataset and the reference dataset. The scaling is done individually for each wavelength first and then across the wavelengths. The heavy atom sites and enantiomorph are searched systematically by SHELXC/D/E in different resolution ranges in all relevant space groups. The correct space group is selected based on results of SHELXD and SHELXE. The heavy atom sites are then refined by autoSHARP and initial trace is obtained from ARP/WARP and RESOLVE. The results are collected into a single location for uploading to a central database. Useful statistics are extracted at each stage and collected into a single log file.

autoXDS is designed to be fully automated and robust, while allowing some flexibility when manual intervention is needed. The tests have indicated that autoXDS can 1. process data (indexing, point group determination, integration and scaling) fully automatically in a very consistent fashion (>95% success rate); 2. solve >90% of the MAD structures (heavy atom site location and refinement, space group and heavy atom enantiomorph determination, density modification and tracing) with no human intervention within a few hours. The data quality and the initial tracing model produced by autoXDS rival or are often better than what humans or other programs can generate.

Above structure can be solve with one command line:

```
unix>autoXDS –data=/data/jcsg/ssrl/9_1/20040520/collection/MB3402A/T2998a -seq=MB3402A.seq
```

The Joint Center for Structural Genomics is funded by the National Institutes of Health (NIH) Protein Structure Initiative. SSRL is operated by Stanford University on behalf of the U.S. Department of Energy (DOE), Office of Basic Energy Sciences. The SSRL Structural Molecular Biology Program is supported by the Department of Energy, Office of Biological and Environmental Research, and by the National Institutes of Health, National Center for Research Resources, Biomedical Technology Program, and the National Institute of General Medical Sciences.

Monday, May 30, 5:30-7:30pm