



# HKL approach to handling Eiger images

Wladek Minor

HDMRX, BNL, May 2016

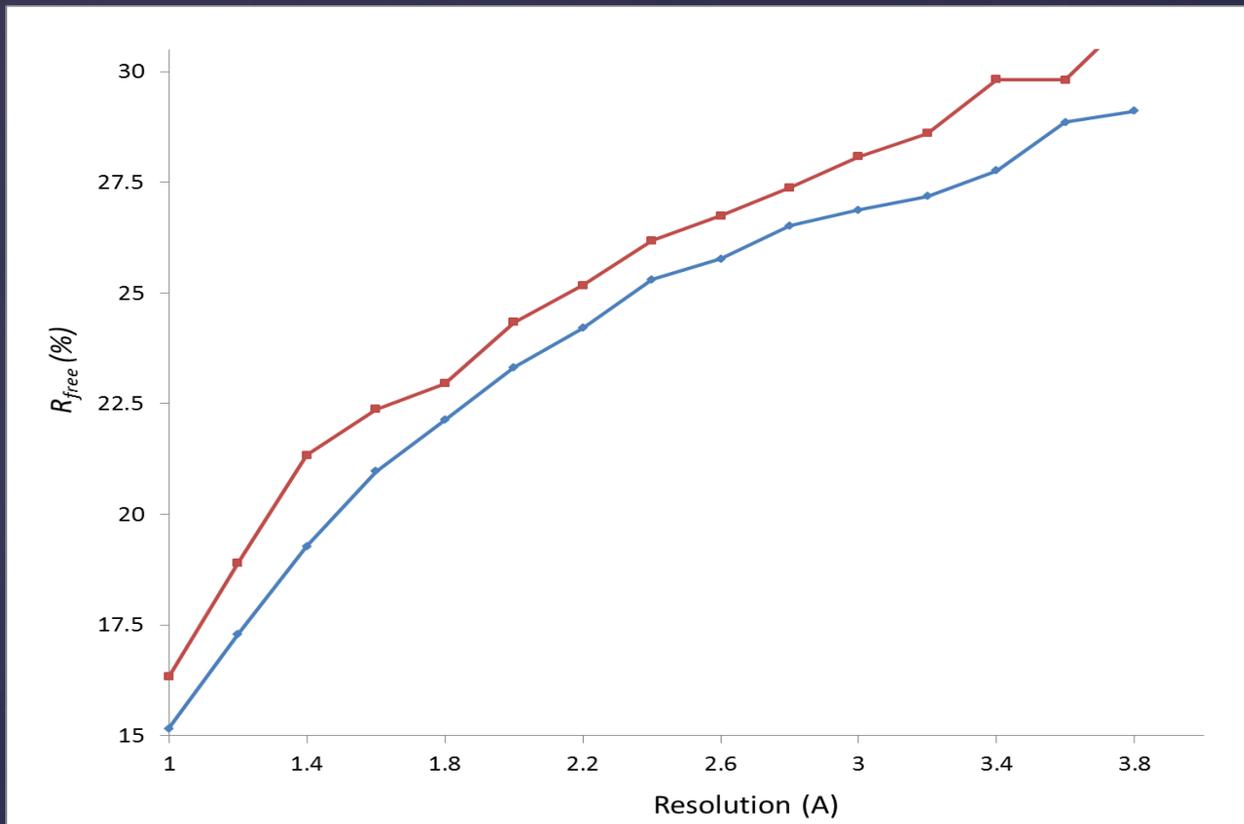
# What experimenters know about data collection ?

```
REMARK 3 ESTIMATED OVERALL COORDINATE ERROR.
REMARK 3 ESU BASED ON R VALUE (A) : NULL
REMARK 3 ESU BASED ON FREE R VALUE (A) : NULL
REMARK 3 ESU BASED ON MAXIMUM LIKELIHOOD (A) : NULL
REMARK 3 ESU FOR B VALUES BASED ON MAXIMUM LIKELIHOOD (A**2) : NULL
REMARK 3
REMARK 3 RMS DEVIATIONS FROM IDEAL VALUES.
REMARK 3 DISTANCE RESTRAINTS. RMS SIGMA
REMARK 3 BOND LENGTH (A) : NULL ; NULL
REMARK 3 ANGLE DISTANCE (A) : NULL ; NULL
REMARK 3 INTRAPLANAR 1-4 DISTANCE (A) : NULL ; NULL
REMARK 3 H-BOND OR METAL COORDINATION (A) : NULL ; NULL
REMARK 3
REMARK 3 PLANE RESTRAINT (A) : NULL ; NULL
REMARK 3 CHIRAL-CENTER RESTRAINT (A**3) : NULL ; NULL
REMARK 3
REMARK 3 NON-BONDED CONTACT RESTRAINTS.
REMARK 3 SINGLE TORSION (A) : NULL ; NULL
REMARK 3 MULTIPLE TORSION (A) : NULL ; NULL
REMARK 3 H-BOND (X...Y) (A) : NULL ; NULL
REMARK 3 H-BOND (X-H...Y) (A) : NULL ; NULL
REMARK 3
REMARK 3 CONFORMATIONAL TORSION ANGLE RESTRAINTS.
REMARK 3 SPECIFIED (DEGREES) : NULL ; NULL
REMARK 3 PLANAR (DEGREES) : NULL ; NULL
REMARK 3 STAGGERED (DEGREES) : NULL ; NULL
REMARK 3 TRANSVERSE (DEGREES) : NULL ; NULL
REMARK 3
REMARK 3 ISOTROPIC THERMAL FACTOR RESTRAINTS. RMS SIGMA
REMARK 3 MAIN-CHAIN BOND (A**2) : NULL ; NULL
REMARK 3 MAIN-CHAIN ANGLE (A**2) : NULL ; NULL
REMARK 3 SIDE-CHAIN BOND (A**2) : NULL ; NULL
```

# What experimenters know about data collection ?

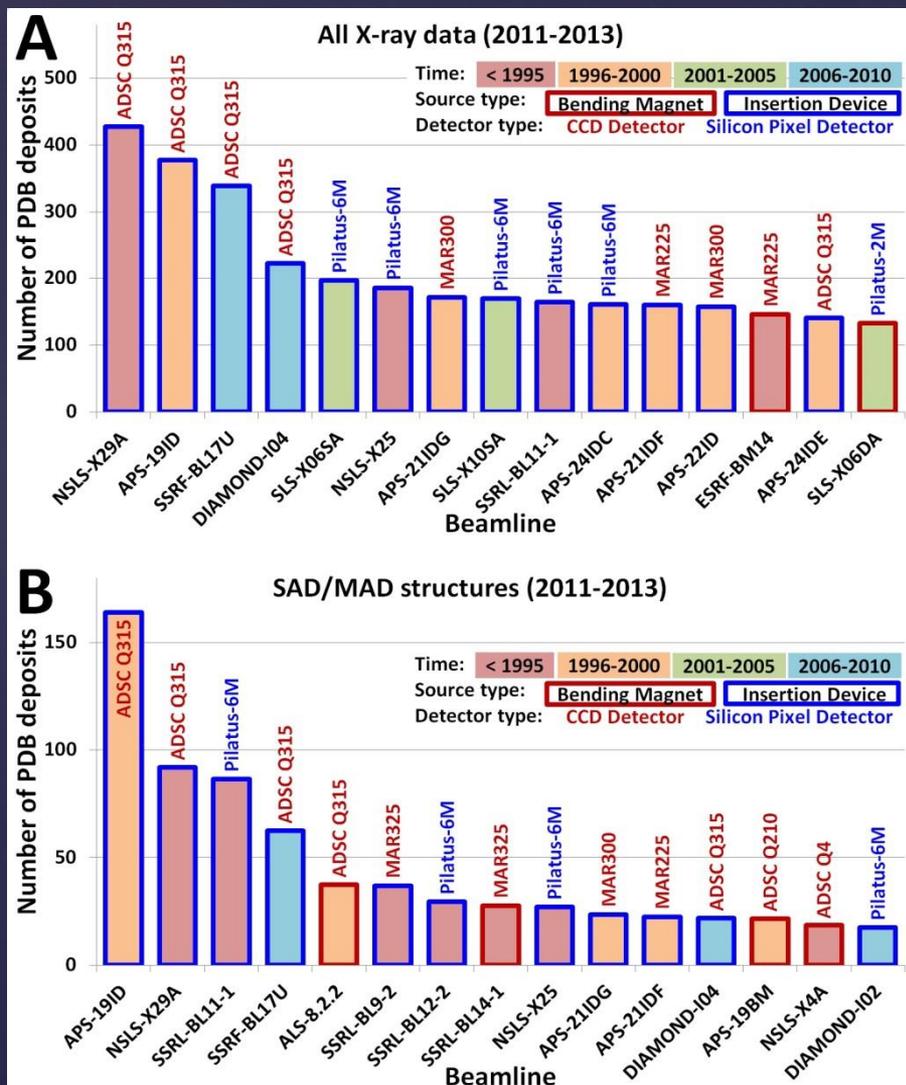
```
REMARK 200 DETECTOR TYPE : CCD
REMARK 200 DETECTOR MANUFACTURER : ADSC QUANTUM 4
REMARK 200 INTENSITY-INTEGRATION SOFTWARE : BLU-ICE
REMARK 200 DATA SCALING SOFTWARE : MOSFLM, CCP4, SCALEPACK
REMARK 200
REMARK 200 NUMBER OF UNIQUE REFLECTIONS : 17575
REMARK 200 RESOLUTION RANGE HIGH (Å) : 2.950
REMARK 200 RESOLUTION RANGE LOW (Å) : 47.870
REMARK 200 REJECTION CRITERIA (SIGMA(I)) : NULL
REMARK 200
REMARK 200 OVERALL.
REMARK 200 COMPLETENESS FOR RANGE (%) : 94.3
REMARK 200 DATA REDUNDANCY : 3.200
REMARK 200 R MERGE (I) : 0.07800
REMARK 200 R SYM (I) : 0.08600
REMARK 200 <I/SIGMA(I)> FOR THE DATA SET : NULL
REMARK 200
REMARK 200 IN THE HIGHEST RESOLUTION SHELL.
REMARK 200 HIGHEST RESOLUTION SHELL, RANGE HIGH (Å) : 2.95
REMARK 200 HIGHEST RESOLUTION SHELL, RANGE LOW (Å) : 3.03
REMARK 200 COMPLETENESS FOR SHELL (%) : 92.4
REMARK 200 DATA REDUNDANCY IN SHELL : 2.80
REMARK 200 R MERGE FOR SHELL (I) : NULL
REMARK 200 R SYM FOR SHELL (I) : NULL
REMARK 200 <I/SIGMA(I)> FOR SHELL : NULL
REMARK 200
REMARK 200 DIFFRACTION PROTOCOL: SINGLE WAVELENGTH
REMARK 200 METHOD USED TO DETERMINE THE STRUCTURE: MAD SE-MET
REMARK 200 SOFTWARE USED: SNB, MLPHARE, CCP4, SOLVE, HKL, RESOLE
REMARK 200 STARTING MODEL: NULL
REMARK 200
REMARK 200 REMARK: NULL
REMARK 280
REMARK 280 CRYSTAL
REMARK 280 SOLVENT CONTENT, VS (%) : NULL
REMARK 280 MATTHEWS COEFFICIENT, VM (ANGSTROMS**3/DÅ) : NULL
REMARK 280
```

# Unexpected correlation?

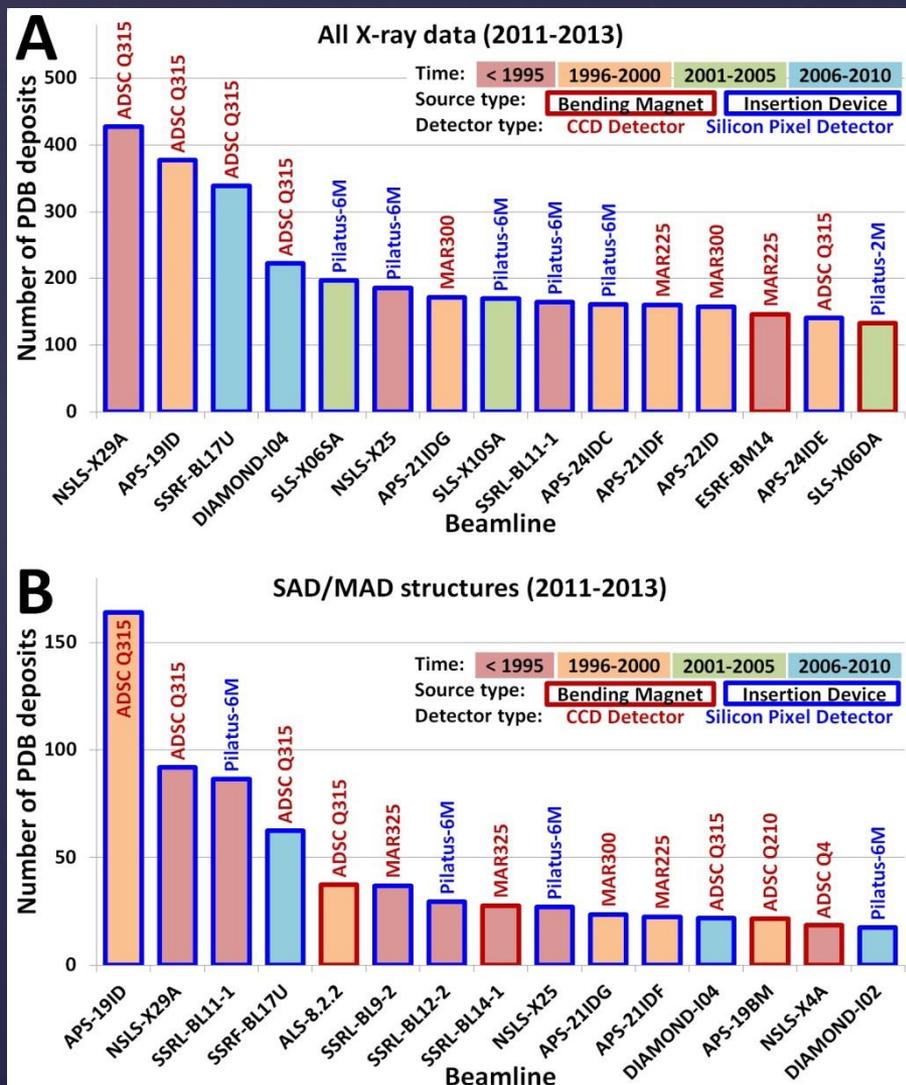


Average  $R_{free}$  by resolution bin (with a width of 0.2 Å for X-ray crystallography PDB structures deposited after January 1, 2001, divided into two groups by the number of missing data items (“NULLs”) in the PDB file. The means for “high-completion” deposits (20 NULLs or less) are shown in blue, and the means for “low-completion” deposits (50 or more NULLs) are shown in red.

# Where we should collect data ?



# High Data-Rate?



# Diffraction experiment - the last experiment before deposition to PDB

Dataset – 2minutes, sample change 2minutes -> 10minutes

6 datasets/hour -> 144 datasets/day

180 days -> 25920 datasets/day -> **2.5 PDB**

125 synchrotron stations -> 324 PDB

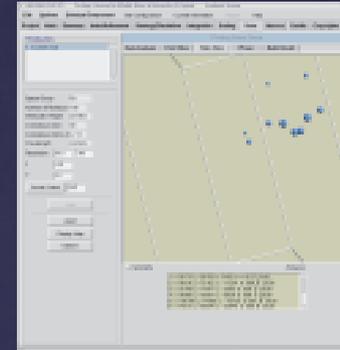
Efficiency -> **0.3%**

# HKL-3000 at SBC



# Database-controlled pipeline

lab e-book

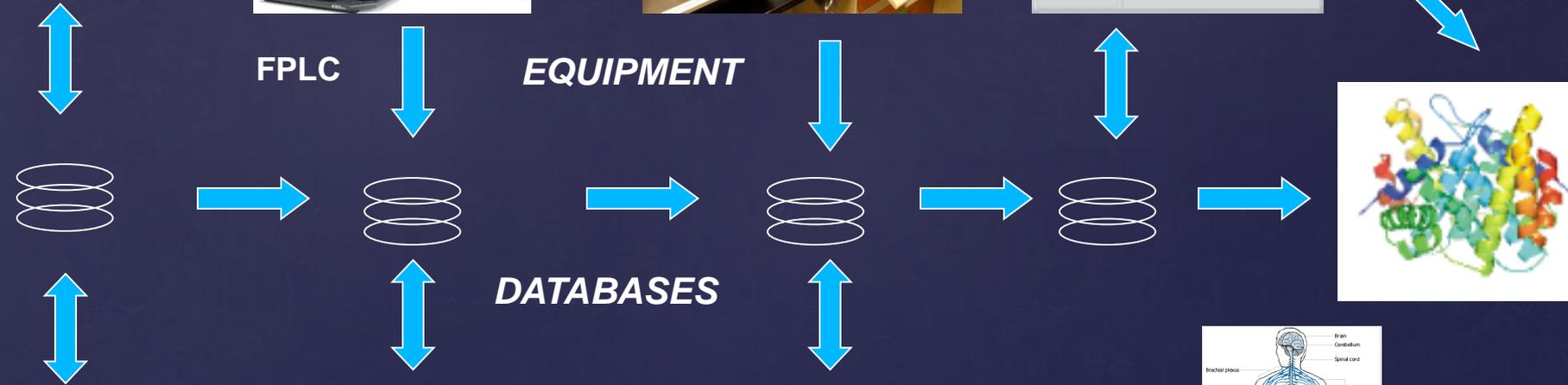


HKL-3000

FPLC

EQUIPMENT

DATABASES



WET LAB DATABASE

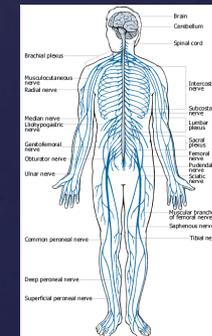
Accession ID	Accession Name	Accession Type	Accession Date	Accession Status	Accession Location	Accession Description	Accession Author	Accession Contact
WET001	Sample 1	Protein	2001-01-01	Active	Lab 1	Sample 1 description	Author 1	Contact 1
WET002	Sample 2	Protein	2001-01-01	Active	Lab 1	Sample 2 description	Author 1	Contact 1
WET003	Sample 3	Protein	2001-01-01	Active	Lab 1	Sample 3 description	Author 1	Contact 1

PeptideDB @ UNIVERSITY OF VIRGINIA

Accession ID	Accession Name	Accession Type	Accession Date	Accession Status	Accession Location	Accession Description	Accession Author	Accession Contact
PeptideDB001	Sample 1	Protein	2001-01-01	Active	Lab 1	Sample 1 description	Author 1	Contact 1
PeptideDB002	Sample 2	Protein	2001-01-01	Active	Lab 1	Sample 2 description	Author 1	Contact 1
PeptideDB003	Sample 3	Protein	2001-01-01	Active	Lab 1	Sample 3 description	Author 1	Contact 1

peptide list

Accession ID	Accession Name	Accession Type	Accession Date	Accession Status	Accession Location	Accession Description	Accession Author	Accession Contact
peptide list 001	Sample 1	Protein	2001-01-01	Active	Lab 1	Sample 1 description	Author 1	Contact 1
peptide list 002	Sample 2	Protein	2001-01-01	Active	Lab 1	Sample 2 description	Author 1	Contact 1
peptide list 003	Sample 3	Protein	2001-01-01	Active	Lab 1	Sample 3 description	Author 1	Contact 1



# Big brother?

## Statistics / Progress in Minor Lab LIMS by researcher

Last week (17 Apr 2015 - 24 Apr 2015)

Person	Clones	Exprs	Purifs	Macro preps	Plates	Drops	Crystals	Datasets processed	Structure refs	Kinetic assays	Thermal shift assays
<a href="#">Cooper, David</a>	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>	0	<u>23</u>	<u>18</u>	<u>0</u>	0	0
<a href="#">Handing, Katarzyna</a>	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>	0	<u>51</u>	<u>53</u>	<u>13</u>	0	0
<a href="#">Hou, Jing</a>	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>	<u>1</u>	30	<u>0</u>	<u>1</u>	<u>1</u>	0	0
<a href="#">Kowiel, Marcin</a>	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>	0	<u>1</u>	<u>8</u>	<u>3</u>	0	0
<a href="#">Shabalin, Ivan</a>	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>	0	<u>125</u>	<u>14</u>	<u>9</u>	0	0
<a href="#">Shumilin, Igor</a>	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>	0	<u>0</u>	<u>3</u>	<u>2</u>	0	0
<a href="#">Szlachta, Karol</a>	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>	0	<u>34</u>	<u>20</u>	<u>3</u>	0	0

Last month (25 Mar 2015 - 24 Apr 2015)

# HDF5 vs CBF

## Structural Biologist Perspective

	Size	Processing time	Rmerge/Rmeas/Rpim
HDF5	3GB	21 min	0.054/0.057/0.018
CBF	3GB	21 min	0.054/0.057/0.017

Changing compiler suite **GCC** -> **Intel**  
processing time **21 min** -> **4 min**

# HDF5 vs CBF

## Structural Biologist Perspective

The screenshot displays a software interface with two main panels. On the left is a hierarchical tree view of an HDF5 dataset. The tree is expanded to show the 'detector' group, which contains various parameters. The 'detector\_distance' parameter is highlighted in blue. On the right is a 'TableView' window showing a single row of data with the value '73.3' circled in red. Below the table view, there is a text box with the following text:

Reported detector distance **73.3 m**

Actual detector distance **65.6 mm**

At the bottom of the interface, a metadata window for 'detector\_distance' is visible, showing the following information:

```
detector_distance (91160624, 2)  
32-bit floating-point, 1  
Number of attributes = 1  
units = m
```

The 'units = m' line is circled in red.

# HDF5 vs CBF

## Structural Biologist Perspective

The screenshot displays a software interface with a hierarchical tree on the left and four data tables on the right. The tree structure is as follows:

- entry
  - data
  - instrument
  - sample
    - goniometer
      - chi
      - chi\_end
      - chi\_range\_average
      - chi\_range\_total
      - kappa
      - kappa\_end
      - kappa\_range\_average
      - kappa\_range\_total
      - omega
      - omega\_end
      - omega\_range\_average
      - omega\_range\_total
      - phi
      - phi\_end
      - phi\_range\_average
      - phi\_range\_total

The four data tables are labeled **chi**, **kappa**, **omega**, and **phi**. Each table contains 52 rows of data, with the first column representing an index from 0 to 52 and the second column representing a numerical value (all values are 0.0). Each table also has a small bar chart icon at the top.

**Each frame: 0**

phi\_range\_total (91237952, 2)  
32-bit floating-point, 1  
Number of attributes = 1  
units = degree

Log Info Metadata

# HDF5 *vs* CBF *vs* 321 other formats

## Structural Biologist Perspective

### New format new trivial inconveniences:

- File renaming
  - need to modify links in HDF5 master file
- New libraries and plugins
  - LZ4 compression plugin compatible only with newer versions of HDF5 library

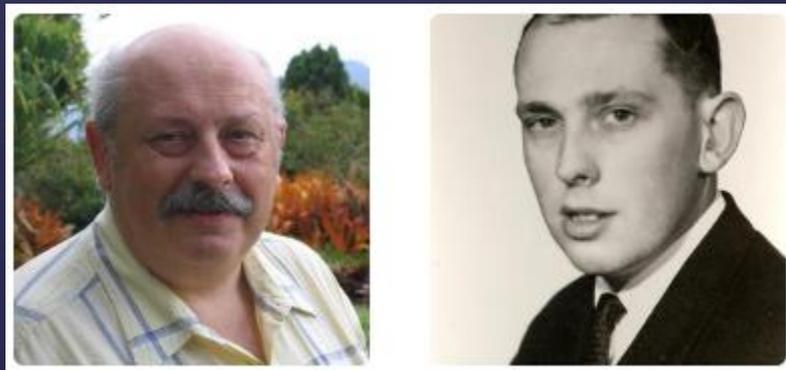
What is special in Eiger and HD5 ?

CBF – MAF

HDF5 – MAF

# What is special in Eiger and HD5 ?

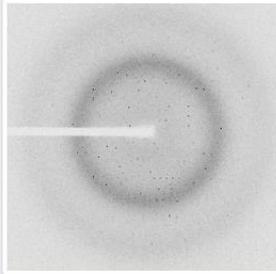
..... - MAH  
CBF – MAF  
HDF5 – MAF



# http://www.proteindiffraction.org



## Integrated Resource for Reproducibility in Macromolecular Crystallography



This project is being funded by the [Targeted Software Development](#) award 1 U01 HG008424-01 as part of the [BD2K \(Big Data to Knowledge\)](#) program of the National Institute of Health. The project is developing tools for "wrangling" data from protein diffraction experiments. We are also creating a growing repository of diffraction experiments used to determine protein structures in the [PDB](#), contributed by the [CSGID](#), [SSGCID](#), [JCSG](#), [MCSG](#), [SGC](#) and other large-scale projects, as well as individual research laboratories.

Currently indexed datasets: **2899**

[Read more...](#)



### Search examples

Find a specific PDB ID: [4K6A](#)

Free format search: ['potential drug target'](#)

Combining searches: [drug AND cholera](#)

Specific beamline: [beamline=21-ID-G](#)

Fuzzy search: [authors ~ Shabalin](#)

Resolution limit (Angstroms): [resolution<1.25](#)

Search by tag: [workshop](#)



Browse & search

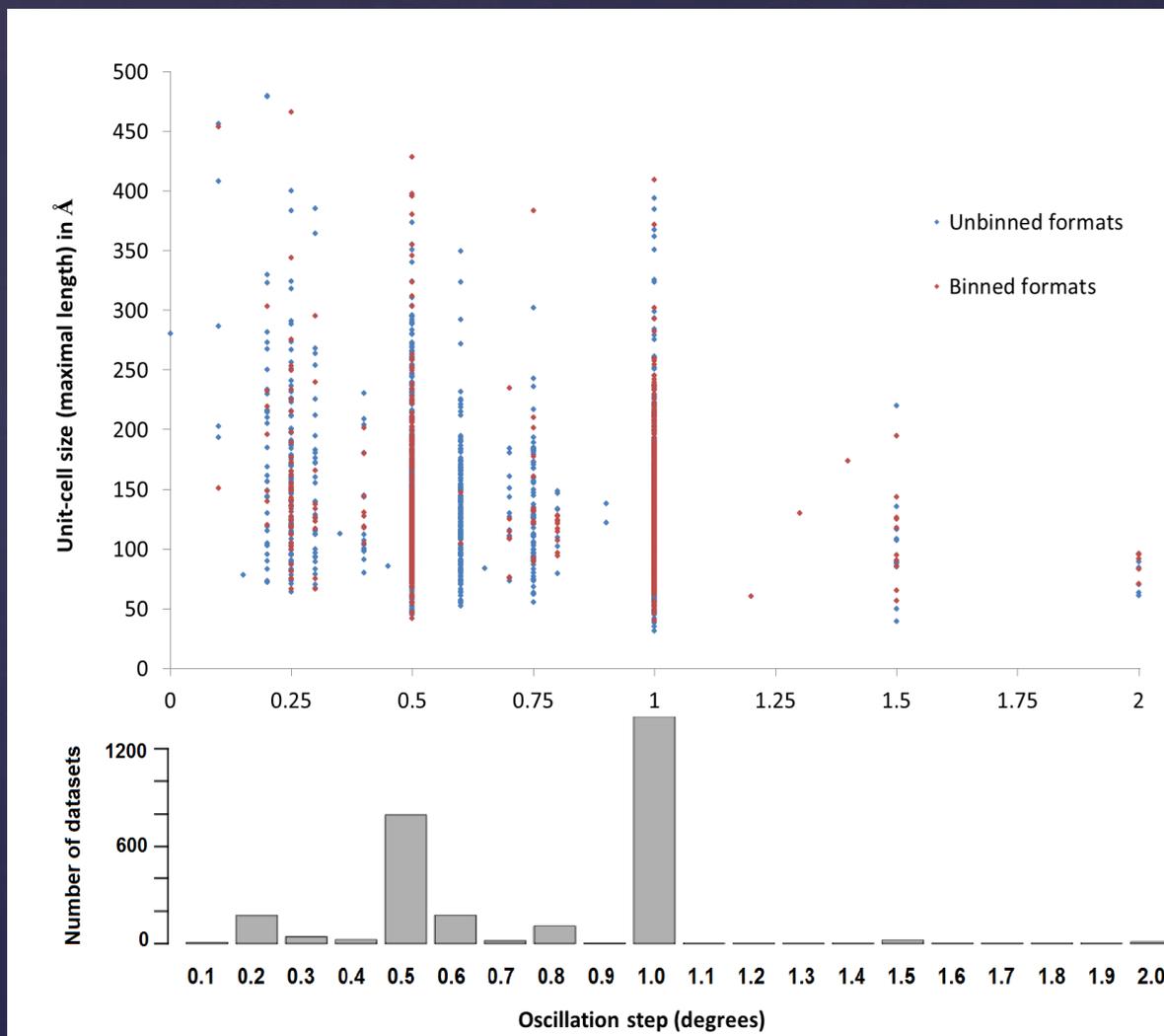


Statistics



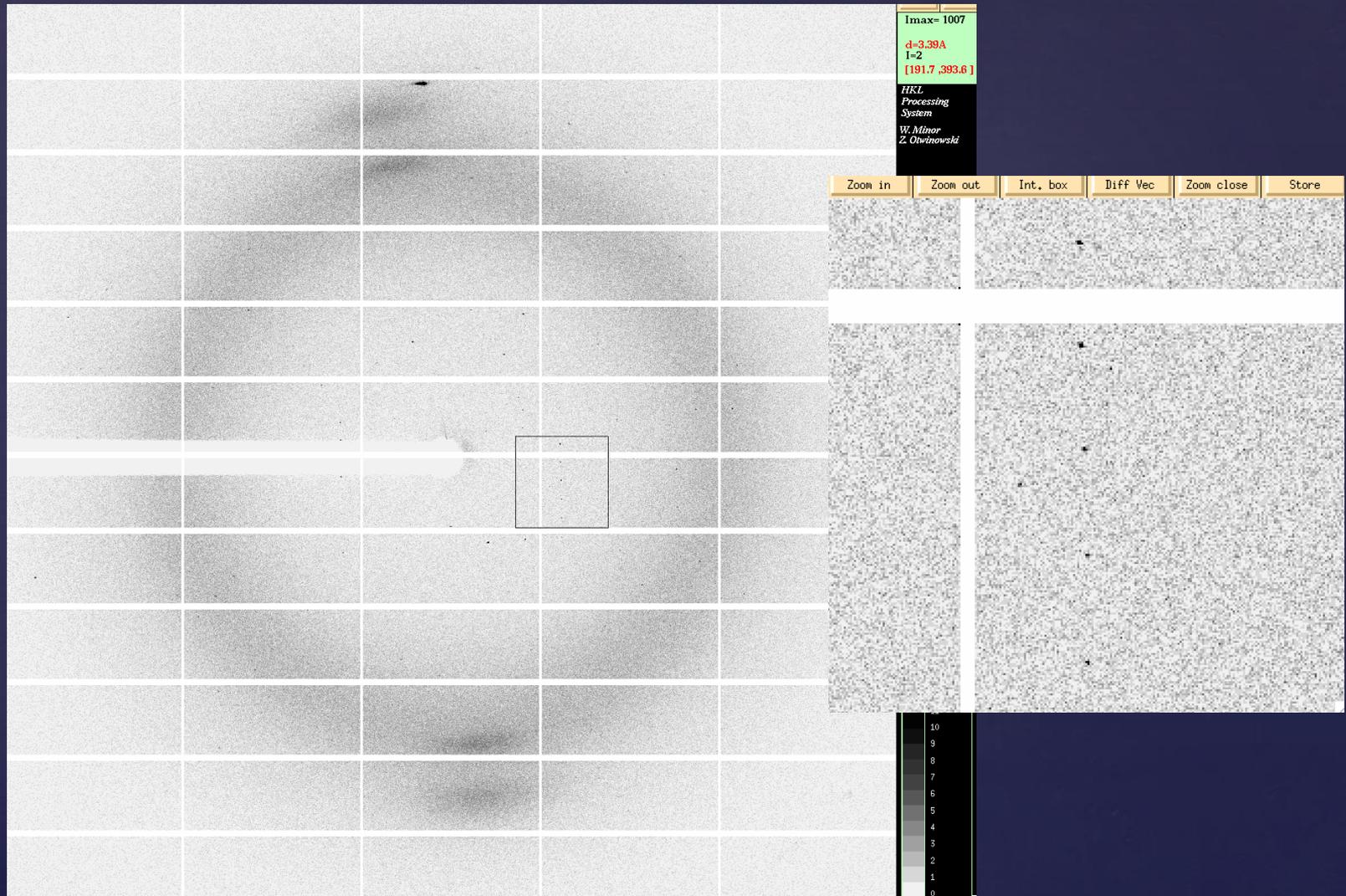
Submit data

<http://www.proteindiffraction.org>

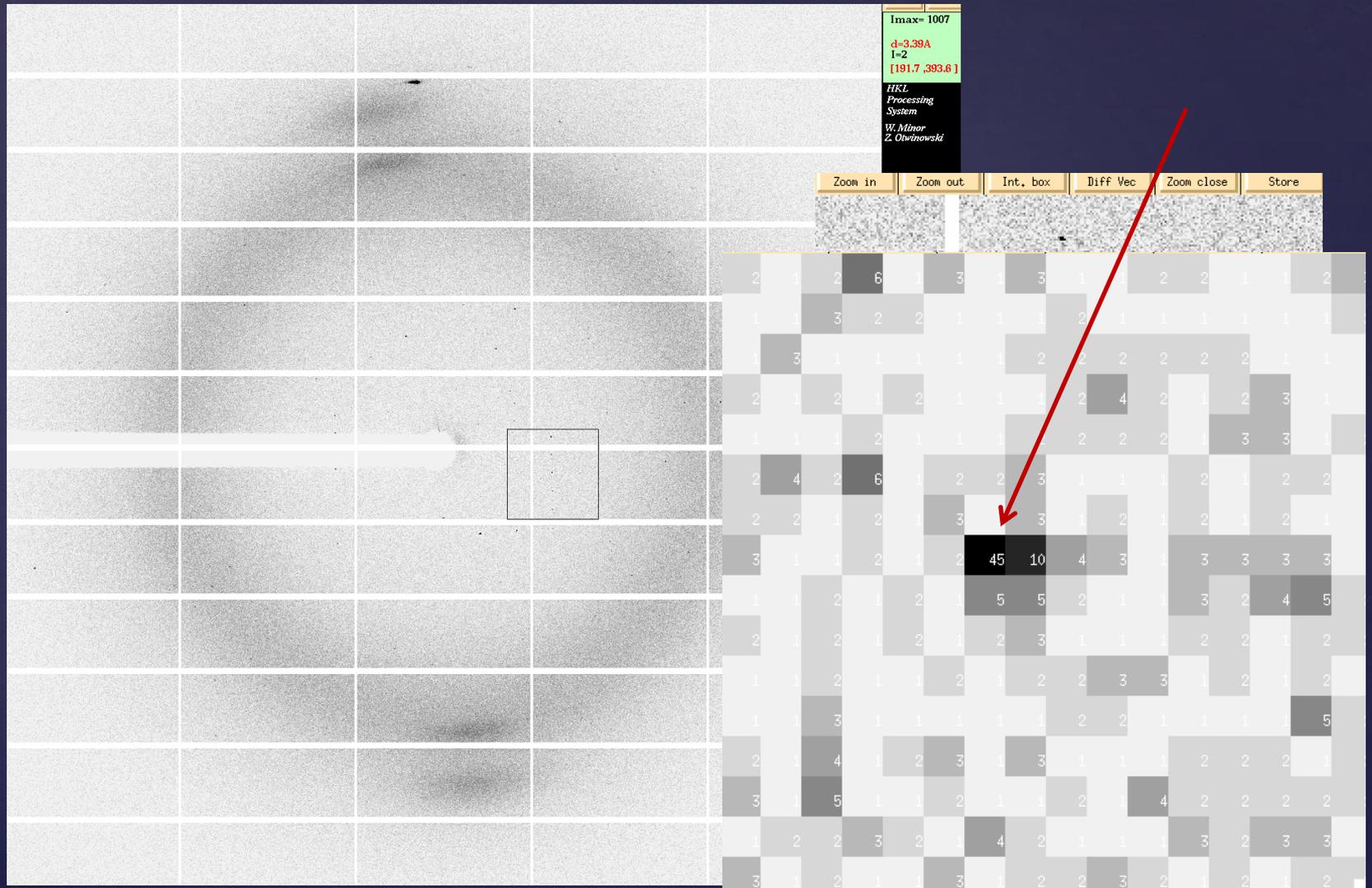




# Optimal data collection ?

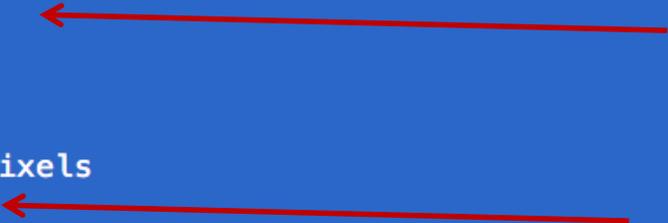


# Optimal data collection ?



# Header – is CBF header a MAH ?

```
# 2015/May/06 10:30:40
# Pixel_size 172e-6 m x 172e-6 m
# Silicon sensor, thickness 0.001 m
# Oscillation_axis omega
# Excluded_pixels: badpix_mask.tif
# Chi 0.0000 deg.
# Angle_increment 0.1000 deg.
# Polarization 0.99
# file_comments
# N_oscillations 2500
# Beam_xy (1223.03, 1256.56) pixels
# Exposure_time 0.020000 s
# Phi 0.0020 deg.
# Energy_range (0, 0) eV
# Start_angle 160.6000 deg.
# Detector_distance 0.617619 m
# Detector_Voffset 0.0000 m
# Alpha 0.0000 deg.
# Flat_field: (nil)
# Threshold_setting 7619 eV
# Exposure_period 0.020950 s
# N_excluded_pixels: = 321
# Kappa 0.0020 deg.
# Tau = 0 s
```



Do you like this image?



Do you like this image ?



# How expensive is bright lens ?



See more choices

Canon EF 85mm f1.2L II USM Lens  
for Canon DSLR Cameras - Fixed  
by Canon

**\$1,999.00** ✓Prime

Get it by **Monday, Aug 24**

More Buying Choices

**\$1,999.00** new (22 offers)

**\$1,499.99** used (24 offers)

Trade-in eligible for an Amazon gift card

★★★★★ ▾ 159



See Style Options

Canon EF 85mm f/1.8 USM Medium  
Telephoto Lens for Canon SLR  
Cameras - Fixed  
by Canon

**\$369.00** ✓Prime

Get it by **Monday, Aug 24**

More Buying Choices

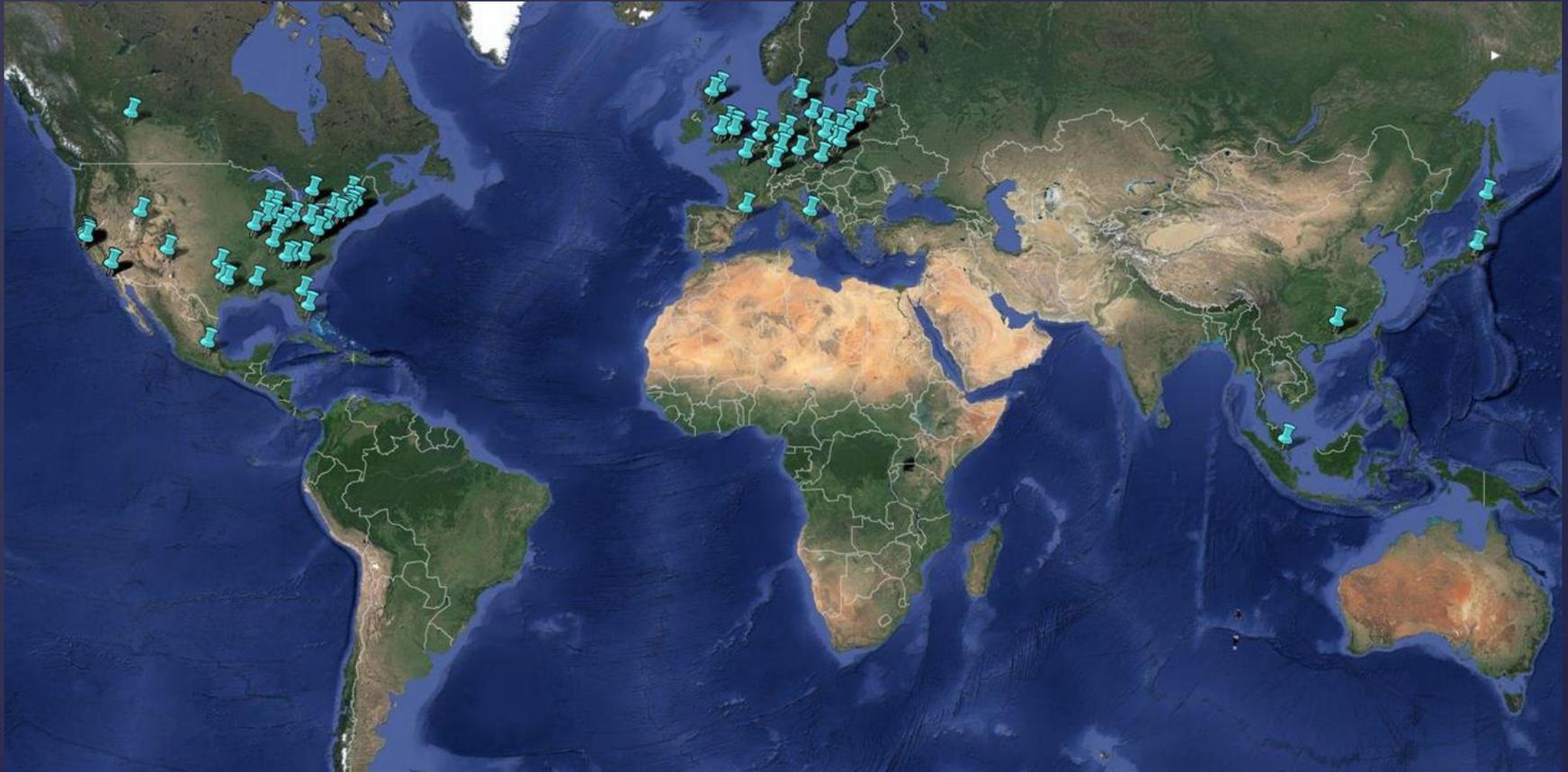
**\$369.00** new (27 offers)

**\$298.00** used (26 offers)

Trade-in eligible for an Amazon gift card

★★★★★ ▾ 770

# Collaborations documented by at least one paper



# Acknowledgment



# Acknowledgments

Wladek Minor

- Przemek Porebski
- Marcin Cymborowski
- Matt Zimmerman (CSIRC)
- Marek Grabowski
- Heping Zheng
- Karol Langner (Google)
- Piotr Sroka
- Ivan Shabalin
- Katherine Handing

Zbyszek Otwinowski

- Dominika Borek

Andrzej Joachimiak

MCSG and SBC staff

Wayne Anderson and CSGID staff

Steve Almo and NYSGRC Staff

Ian Wilson, Marc Elslinger and JCSG staff

Steven Burley, John Westbrook and PDB staff

Tom Terwilliger

Zbyszek Dauter

**Grants:**

**U01-HG008424**

NIH GM53163, GM62414, GM74942

GM093342, GM094585, GM094662

DOE, NCI

NIAID HHSN272200700058C

NIAID HHSN272201200026C

HKL Research. Inc.