# Enhancing Peak Precision in Crystallography


Adam Kurth

CXFEL Labs, Biodesign Institute, Arizona State University


BioXFEL — A National Science Foundation Science and Technology Center

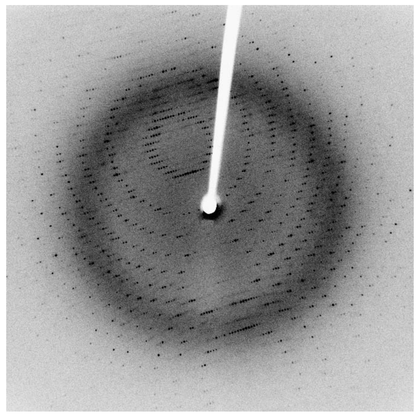
ASU Biodesign Institute — Arizona State University

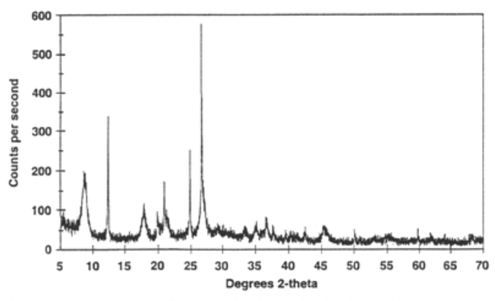
CXFEL Labs

## Problem: Three-Ring Integration

Water background subtraction is crucial for accurate structure determination, especially in diffraction data with interference, emphasizing the importance of separating signal from noise, including water background.

The goal of water background subtraction in X-ray crystallography is to accurately measure peak intensity values ($I$) while effectively separating them from background noise ($\sigma I$) in the image. This involves estimating peak values and coordinates while considering nearby background noise to achieve reliable differentiation. Three-ring integration is used for this purpose, ensuring precise measurement of peak intensities amid varying background noise levels.






https://commons.wikimedia.org/wiki/File:X-ray_diffraction_pattern_3clpro.jpg#/media/File:X-ray_diffraction_pattern_3clpro.jpg
https://serc.carleton.edu/research_education/geochemsheets/techniques/XRD.html

## Implementation

### 1. File and Data Preparation

Initiates with the *load_file_h5* function to verify and load the specified HDF5 file, ensuring data availability. It introduces two key classes:

- *PeakThresholdProcessor* for threshold-based peak identification.
- *ArrayRegion* for managing and extracting specified data regions.

This stage sets the foundation for accurate data analysis by preparing the data structures and loading the necessary data from files.

- User inputs an HDF5 image and specifies intensity threshold

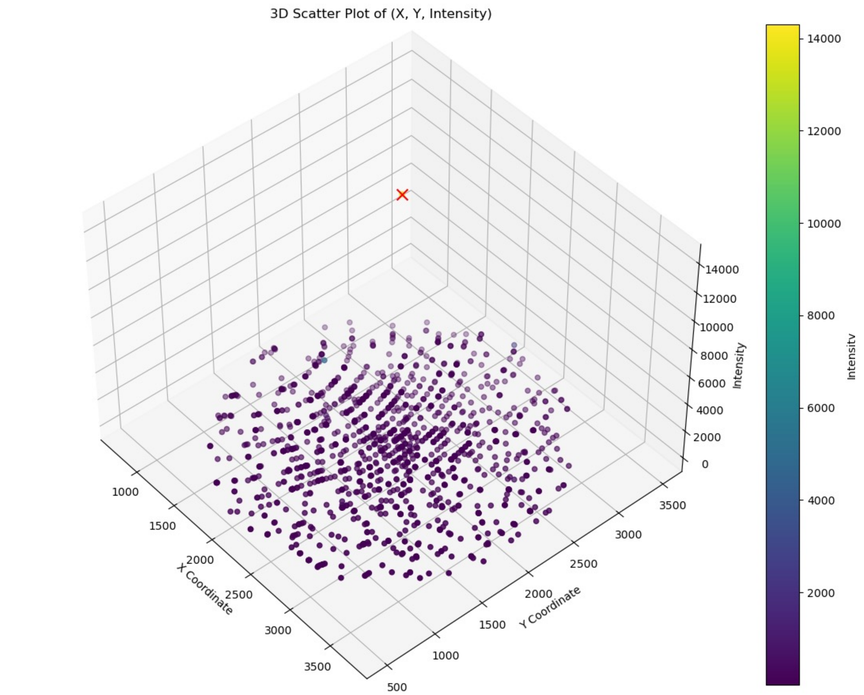
```
DISPLAY REGION
[[    0.0000000367    0.0000000464    0.0000104391    0.0000000963    0.0000285485    0.0000022639    0.1650659740    0.0068235416    0.0014431686]
 [    0.0000000150    0.0000001350    0.0000146119    0.0000026415    0.0009533833    0.0004932618    0.1020206884    0.3589628339    0.0003718563]
 [    0.0000000029    0.0000576651    0.0000061747    0.0000836480    0.0010862813    0.0081905210   11.6344747543    0.0011066927    0.0001987342]
 [    0.0000028895    0.0000019056    0.0000319810    0.0000735537    0.0554109924    0.4253226817    0.0084624672    0.0000047701    0.0000219900]
 [    0.0000236321    0.0000169144    0.0000338372    0.2086020261 14298.4199218750    0.0000311176    0.0002345003    0.0000009032    0.0000043385]
 [    0.0000085413    0.0000353902    0.0001237222    0.0000154377    0.0000287614    0.0000942235    0.0001058824    0.0000002992    0.0000386222]
 [    0.0000000721    0.0000000143    0.0000001399    0.0000414735    0.0000001738    0.0000001966    0.0000477113    0.0000000665    0.0000106145]
 [    0.0000000023    0.0000000088    0.0000001241    0.0000014439    0.0000004072    0.0000000037    0.0000000042    0.0000000502    0.0000007026]
 [    0.0000000010    0.0000051898    0.0000014510    0.0000000161    0.0000002331    0.0000000006    0.0000000004    0.0000000006    0.0000006530]]
SEGMENT
 [[    0.0000061747    0.0000836480    0.0010862813    0.0081905210   11.6344747543]
 [    0.0000319810    0.0000735537    0.0554109924    0.4253226817    0.0084624672]
 [    0.0000338372    0.2086020261 14298.4199218750    0.0000311176    0.0002345003]
 [    0.0001237222    0.0000154377    0.0000287614    0.0000942235    0.0001058824]
 [    0.0000001399    0.0000414735    0.0000001738    0.0000001966    0.0000477113]]
BOOLEAN
 [[False False False False False]
 [False False False False False]
 [False False False False False]
 [False False False False False]
 [False False False False False]]
```

### 2. Data Processing

Focuses on analyzing the data through a series of steps designed to isolate and evaluate peak intensities within the dataset. Utilizes the *extract_region* function for precise region extraction around potential peaks, and employs the *PeakThresholdProcessor* for identifying significant coordinates based on threshold values. The *coordinate_menu* function allows for interactive exploration and processing of selected peaks, facilitating detailed analysis of the data's critical points.

```
(row,col) (0, 0) with a value of  6.1747137e-06
(row,col) (1, 0) with a value of  3.1981028e-05
(row,col) (2, 0) with a value of  3.3837216e-05
(row,col) (3, 0) with a value of  0.00012372217
(row,col) (4, 0) with a value of  1.3989028e-07
(row,col) (0, 1) with a value of  8.364805e-05
Passed (row, col) (1, 1) 7.355371e-05
Passed (row, col) (2, 1) 0.20860203
Passed (row, col) (3, 1) 1.5437696e-05
(row,col) (4, 1) with a value of  4.147347e-05
(row,col) (0, 2) with a value of  0.0010862813
Passed (row, col) (1, 2) 0.055410992
Peak point to be skipped: (2, 2) 14298.42
Passed (row, col) (3, 2) 2.8761428e-05
(row,col) (4, 2) with a value of  1.73839e-07
(row,col) (0, 3) with a value of  0.008190521
Passed (row, col) (1, 3) 0.42532268
Passed (row, col) (2, 3) 3.1117648e-05
Passed (row, col) (3, 3) 9.422347e-05
(row,col) (4, 3) with a value of  1.9659073e-07
(row,col) (0, 4) with a value of  11.634475
(row,col) (1, 4) with a value of  0.008462467
(row,col) (2, 4) with a value of  0.00023450026
(row,col) (3, 4) with a value of  0.00010588242
(row,col) (4, 4) with a value of  4.7711346e-05
```

- Traverses and integrates pixel values for values in the “radius”.
- Skips the actual peak point.
- Gathering information about background around that peak.




### 3. Visualization and Synthesis

Concludes with the generation of a 3D scatter plot via the *create_scatter* function, visualizing the spatial distribution and intensity of identified peaks. This visualization aids in the interpretation of data, highlighting the relationship between peak locations and their intensities. The main process orchestrates the workflow from data preparation to visualization, encapsulating the analysis in a comprehensive manner that supports the identification and evaluation of peak intensities against background noise.

- Confirm all points have been traversed, then compute and view summary peak estimate.
- Scatter plot will be shown after all rings of “radius” $r = \{1, 2, 3, 4\}$.

```
[[ True  True  True  True  True]
 [ True  True  True  True  True]
 [ True  True  True  True  True]
 [ True  True  True  True  True]
 [ True  True  True  True  True]]
Number of traversed cells 25
Peak point to be skipped: (2, 2)
Total sum: 11.652923464859015
Average surrounding peak: 0.4661169385943606
Intensity value 14298.42

Peak Estimate for ring 2 : 14297.953804936406 with radius of 2
```

## Enhanced Prediction

In this process, simulations are conducted to generate images with noise using the *pattern_sim* tool. Key parameters, such as *--nphotons* (in kilo electron volts, keV), are specified to simulate both "high" (8 keV) and "low" (7 keV) intensity images. These images are then indexed using *indexamajig* to produce the stream files needed for analysis. The program extracts the indexed peak information from the "low" keV stream file and replaces the background in the "high" keV stream file with more precise peak values, resulting in an overwritten stream file, with more accurate peak estimates.

- Overwritten enhanced peak estimate values.

```
Reflections measured after indexing
   h    k    l          I   sigma(I)       peak background  fs/px  ss/px panel
 -45   -8   -1       0.00       1.00      0.00       0.00   73.6 2210.5 panel0
 -45   -2   -8       0.00       1.00      0.00       0.00   20.4 2964.0 panel0
 -44  -18    5       0.00       1.00      0.00       0.00    9.9 1434.8 panel0
 -44  -10    1       0.00       1.00      0.00       0.00  134.1 2080.3 panel0
 -44   -6   -2       0.00       1.00      0.00       0.00  158.1 2335.7 panel0
 -44   -5   -3       0.00       1.00      0.00       0.00  155.9 2440.0 panel0
 -44   -4    4       0.00       1.00      0.00       0.00  153.7 2545.0 panel0
 -44   -1   -9       0.00       1.00      0.00       0.00   76.6 3071.9 panel0
 -43  -17    5       0.00       1.00      0.00       0.00  103.7 1479.9 panel0
 -43  -11    2       0.00       1.00      0.00       0.00  191.0 1897.7 panel0
 -43    0  -10       0.00       1.00      0.00       0.00  132.6 3179.3 panel0
 -42  -16    5       0.00       1.00      0.00       0.00  192.2 1522.5 panel0
 -42  -14    4       0.00       1.00      0.00       0.00  220.0 1662.2 panel0
 -42  -12    3       0.00       1.00      0.00       0.00  246.0 1797.0 panel0
 -42  -10    2       0.00       1.00      0.00       0.00  272.7 1927.1 panel0
 -42   -7    0       0.00       1.00      0.00       0.00  296.1 2151.9 panel0
 -42   -2   -5       0.00       1.00      0.00       0.00  287.7 2657.0 panel0
 -42    1  -14       0.00       1.00      0.00       0.00   78.0 3658.5 panel0
 -42    1  -13       0.00       1.00      0.00       0.00  116.8 3527.2 panel0
 -42    1  -12       0.00       1.00      0.00       0.00  153.5 3403.4 panel0
 -41  -24    8       0.00       1.00      0.00       0.00   93.1  971.9 panel0
```


## Acknowledgments

This work was funded by the National Science Foundation STC award number 1231306.

**This work supported in part by the National Science Foundation under midscale RI-2 award No. 2153503**