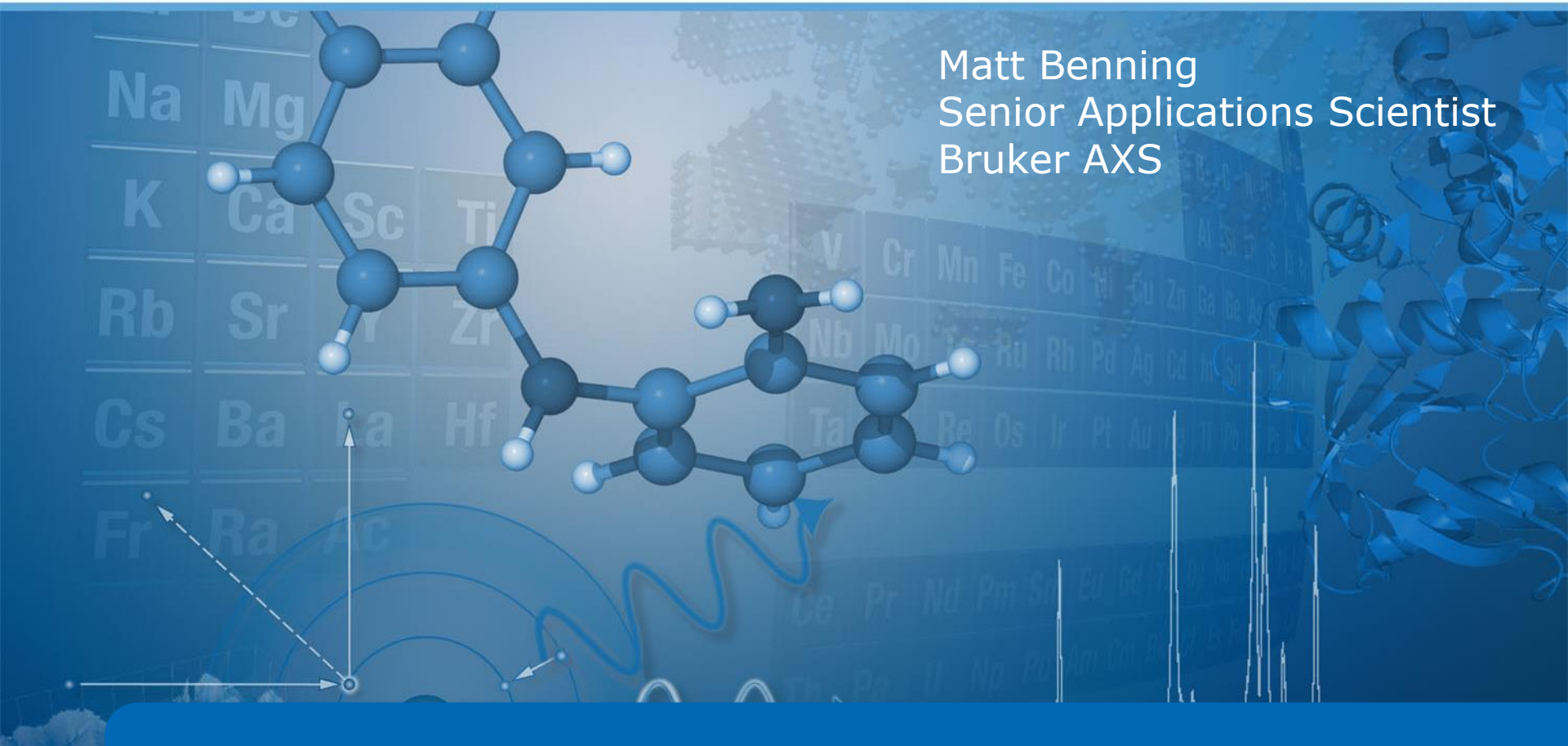
Current Trends in Native SAD Phasing



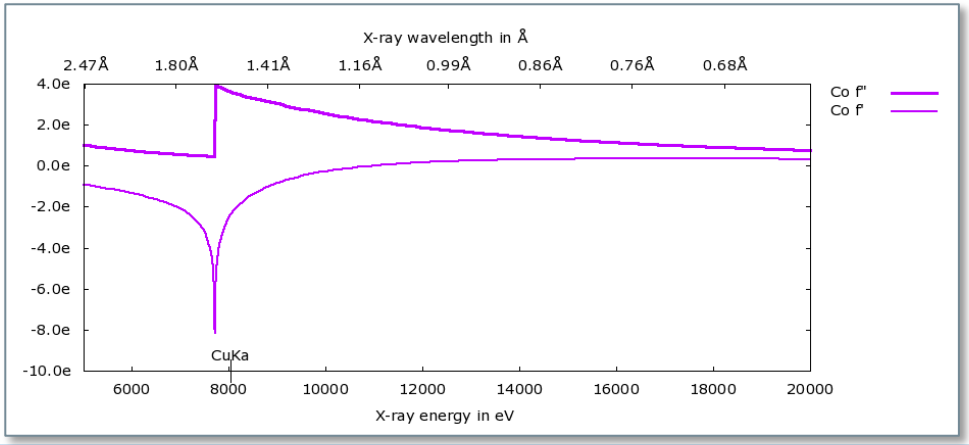
Matt Benning
Senior Applications Scientist
Bruker AXS

Different ways of getting phases

- Molecular replacement
  - Uses a similar model to determine the unknown structure
  - Rotation search based on Patterson map, relative orientation of the molecule
  - Translation search in the unit cell

- Anomalous dispersion
  - X-ray data is inherently centrosymmetric, Friedel's law HKL = -H-K-L
  - When you are near the absorption edge of an atom, you get a phase shift
  - Utilize the differences in Friedel's mates to find marker atoms
  - Cobalt anomalous scattering plot

  - $f = f_0 + f' + if''$



Different ways of getting phases
Heavy atom methods

Anomalous dispersion methods

- Multi-wavelength (MAD)
  - Synchrotron data at two or more wavelengths near the absorption edge of an element
- Single-wavelength (SAD)
  - Single wavelength at an element's absorption edge
  - Native SAD
    - Sulfur atoms
    - Naturally occurring metals

Native SAD
Marker atoms

- F” values for common marker atoms in proteins
- Metals are the best since they typically produce a strong anomalous signal and are tightly bound
- CYS and MET residues provide sulfurs as potential marker atoms but MET have a higher degree of conformation freedom so can be hard to find

	CuKα	GaKα
P	0.43	0.33
S	0.56	0.43
Se	1.15	0.88
Mg	0.18	0.15
Ni	0.53	3.29
Zn	0.68	0.53
Ca	1.3	1
Mn	2.76	2.24
Fe	3.2	2.56
Co	3.65	2.92
Cd	4.66	3.67

Experimental phasing

Except in relatively rare cases where atomic resolution data permits the phase problem to be solved by *ab initio* direct methods, experimental phasing usually implies the presence of heavy atoms to provide reference phases. We then calculate the phases ϕ_T of the full structure by:

$$\phi_T = \phi_A + \alpha$$

Where ϕ_A is the calculated phase of the heavy atom substructure.

α can be estimated from the experimental data.

The phase determination requires the following stages:

- Location of the heavy atoms
- Determine reference phases and phase shift
- Improvement of these phases by density modification

Analysis of MAD data

Karle (1980) and Hendrickson, Smith & Sheriff (1985) showed by algebra that the measured intensities in a MAD experiment should be given by:

$$|F \pm|^2 = |F_T|^2 + a|F_A|^2 + b|F_T||F_A|\cos\alpha \pm c|F_T||F_A|\sin\alpha$$

$|F_T|$ (native *F* including heavy atoms but ignoring *f′* and *f″* contributions)
$|F_A|$ (heavy atom structure factor)
α phase shift

where:

$a = (f''^2 + f'^2)/f_0^2$,
$b = 2f'/f_0$,
$c = 2f''/f_0$ and
$\alpha = \phi_T - \phi_A$

a, *b* and *c* are different for each wavelength.

Analysis of MAD data

- Provided that $|F \pm|^2$ has been measured at two or more wavelengths we can solve the equations for all the variables
- So for MAD (or SIRAS) phasing, all we need to do is to use find the heavy atoms from F_A, use them to get ϕ_A and then calculate a map with amplitudes $|F_T|$ and phases
  $\phi_T = \phi_A + \alpha$.

SAD phasing

The $|F_T|$ and $\phi_T = \phi_A + \alpha$ approach used in MAD phasing is so convenient that it is desirable to extend it to SAD

For SAD, we can approximate:

$|F_T| = (|F+| + |F-|) / 2$ and

$\alpha = 90^\circ$ for $|F+| > |F-|$ or

$\alpha = 270^\circ$ for $|F+| < |F-|$

$|F+| - |F-| \approx c\,|F_A|\sin\alpha$, can use ΔF values to find HA

The weight (fom) to be assigned to ϕ_T should be large when both the calculated heavy atom structure factor $|F_H|$ and $||F+| - |F-||$ are large, and small when either or both are small

Native SAD Phasing
Experiment

- Number of HA sites
  - Sulfur requires 2-3 sites per 100 residues
  - Heavier atoms require fewer sites
- Data Collection strategies
  - Higher random multiplicity, different crystal orientations
  - Inverse beam using wedges of data
  - Friedel mates on the same image
  - Multi-crystal averaging
- Resolution limit
  - For the substructure solution, 3.5 – 2.5 Å
  - Interpretable electron density map, > 3.5 Å
- Solvent content
  - SAD phasing is highly dependent on density modification so the higher the solvent content the better

Native SAD Phasing
Will it work?

- Data metrics, correlation coefficients
  - ΔF vs σ, for all the Bijvoet pairs
  - CC (1/2), internal anomalous consistency within a data set
  - Anomalous scattering ratio (R_{as})
    - Comparison of symmetry related centric and acentric reflections
- Data collection
  - Optimize sample prep
  - Select the proper crystal mount
  - Minimize radiation damage
  - Collection strategy
    - Increase multiplicity evenly between I^+ and I^-
    - Sulfur requires 13 – 25 fold
    - Heavier atoms 5 – 15 fold

SHELX Suite

SHELX is a set of programs written by George Sheldrick for structure determination and model refinement from single crystal diffraction data

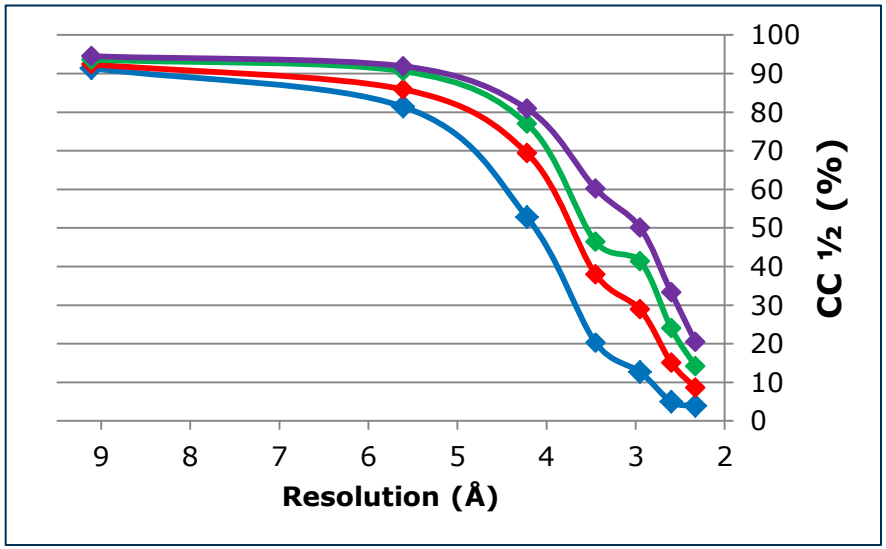
- SHELXC - prepare files for phasing with SHELXD/E
- SHELXD - determine the heavy atom substructure
- SHELXE - density modification and main-chain trace
- XC, XM, XE are the Bruker versions of SHELX

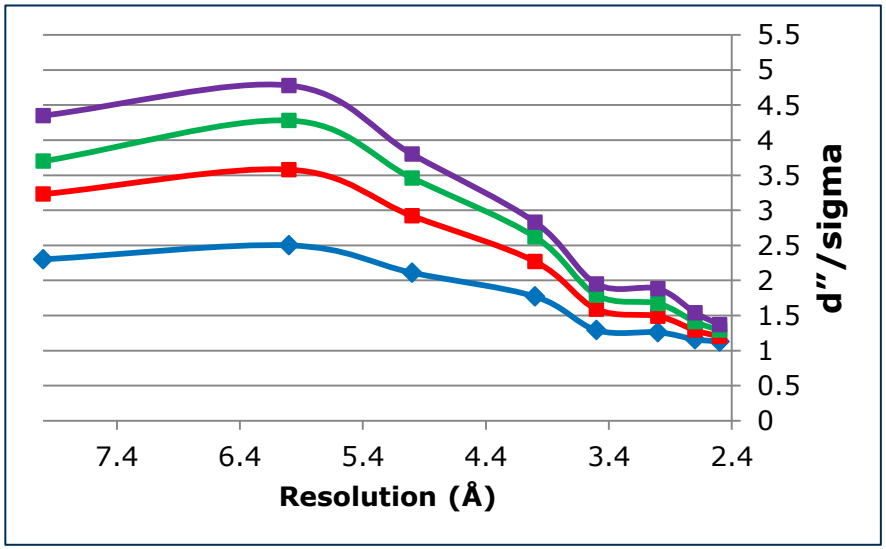
XC, SHELXC
Data analysis

- Creates a file with anomalous coefficients and phase shift
  - Preforms local scaling to help remove systematic errors
  - F_H, α
- Outputs data metrics
  - Data completeness and I/σI in resolution bins
  - Signal to noise for anomalous differences, $\Delta F/\sigma$ (d”/sig)
  - CC for anomalous differences between wavelengths, MAD
  - $CC_{1/2}$ to measure internal consistency within a data set, SAD

SAD Phasing
Higher multiplicity

- Metrics improve as the multiplicity increases
- Example of an cubic Insulin data set



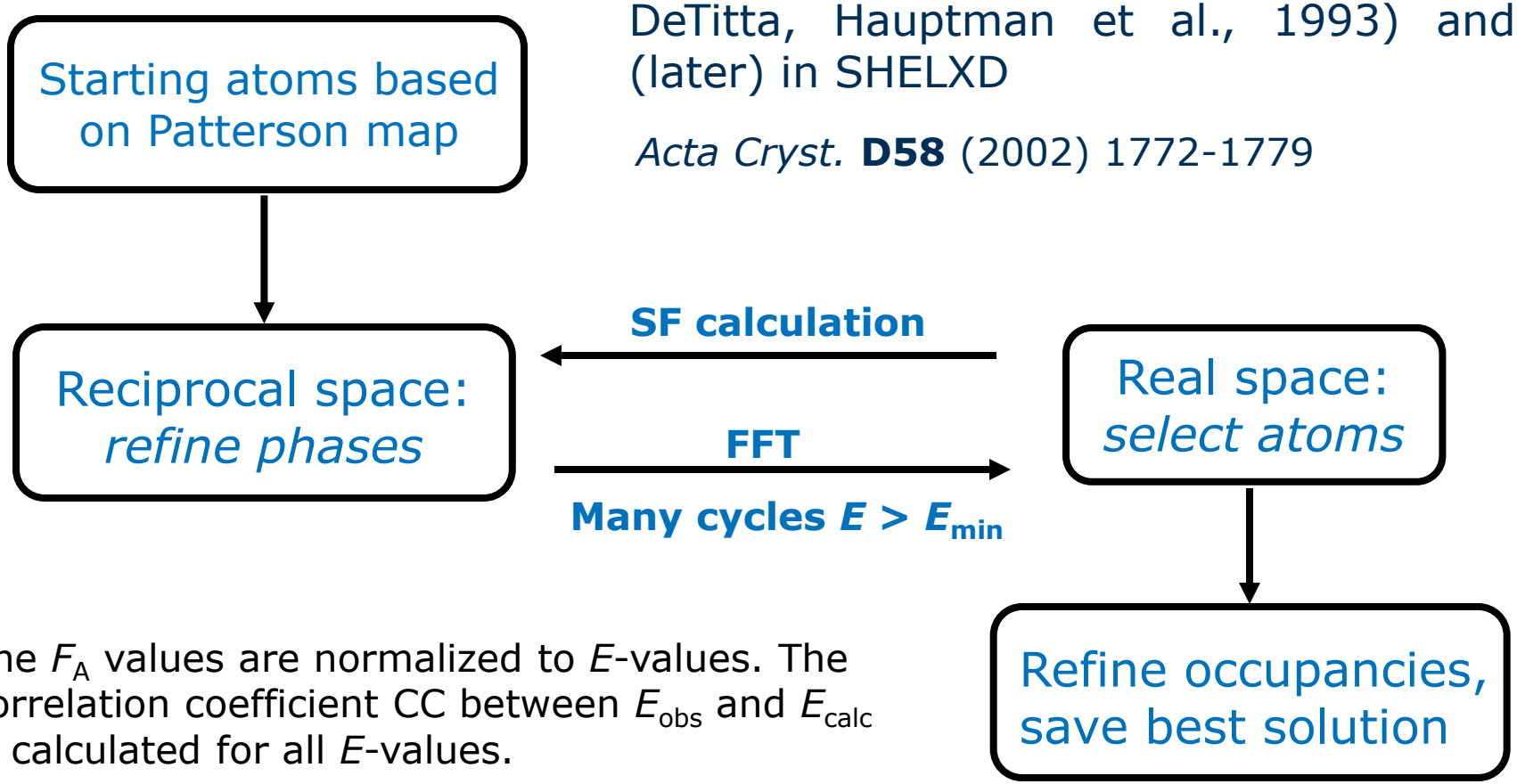


- 45 degrees
- 90 degrees
- 135 degrees
- 180 degrees

XM, SHELXD
Dual-space methods

Implemented in SnB (Weeks, Miller, DeTitta, Hauptman et al., 1993) and (later) in SHELXD

Acta Cryst. **D58** (2002) 1772-1779



The F_A values are normalized to *E*-values. The correlation coefficient CC between E_{obs} and E_{calc} is calculated for all *E*-values.

XM, SHELXD
Critical parameters

- High resolution data is not so critical; 2.5 - 3.5 Å is fine because the distance between sites is still normally greater. Vary the resolution range to find the best solution.
- The *resolution* at which the ΔF-data are truncated, e.g. where the internal CC ($CC_{1/2}$) between the signed anomalous differences of two randomly chosen reflection subsets falls below 30% or $\Delta F/\sigma$ falls below 1.3.
- The *number of sites* requested should be within about 20% of the true value so that the occupancy refinement works well (and reveals the true number).
- In the case of a soak, you can allow sites on *special positions* but in general this should be disabled.
- DSUL command allows a two-atom search treating disulfide bridges as super sulfurs.

In difficult cases it may be necessary to fine-tune these settings and run more trials (say >10000 rather than 1000).

XM, SHELXD
Correlation coefficients

- Compares the Normalized structure factor calculated from the substructure to that found from the difference data
- The CC is calculated with all data (CC-all) and also with only the weaker *E*-values (CC-weak) that were not used in the dual-space recycling
  - CCall >30%
  - CCweak >15%
  - CFOM >45%

XE, SHELXE
Density modification and Phasing

Sphere of Influence

- Assign the protein and solvent regions by calculating the variance of each voxel (volume element) on spherical surface in the density map
- Voxel with large variances belong to the protein region, lowest variances solvent
- A crossover region or fuzzy solvent boundary is used to prevent the density modification from being locked into a poor solvent boundary

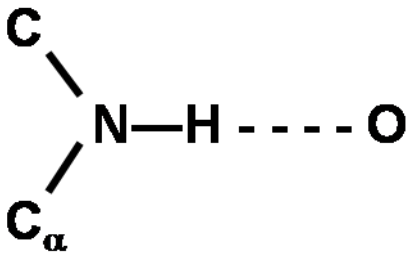
XE, SHELXE
Autotracing algorithm

A fast autotracing algorithm has been incorporated into the density modification in SHELXE

- Find potential α-helices in the density and try to extend them at both ends. Then find other potential tripeptides and try to extend them at both ends in the same way.
- Tidy up and splice the traces as required, applying any necessary symmetry operations.
- Use the traced residues to estimate phases and combine these with the initial phase information using sigma-A weights, then restart density modification. The refinement of one B-value per residue provides a further opportunity to suppress wrongly traced residues.

XE, SHELXE
Criteria for accepting chains

- The overall fit to the density should be good.
- The chains must be long enough (in general at least 7 amino-acids); longer chains are given a higher weight.
- There should not be too many Ramachandran outliers.
- There should be a well defined secondary structure (ϕ / φ pairs should tend to be similar for consecutive residues).
- On average, there should be significant positive density 2.9 Å from N in the N→H direction (to a hydrogen bond acceptor).

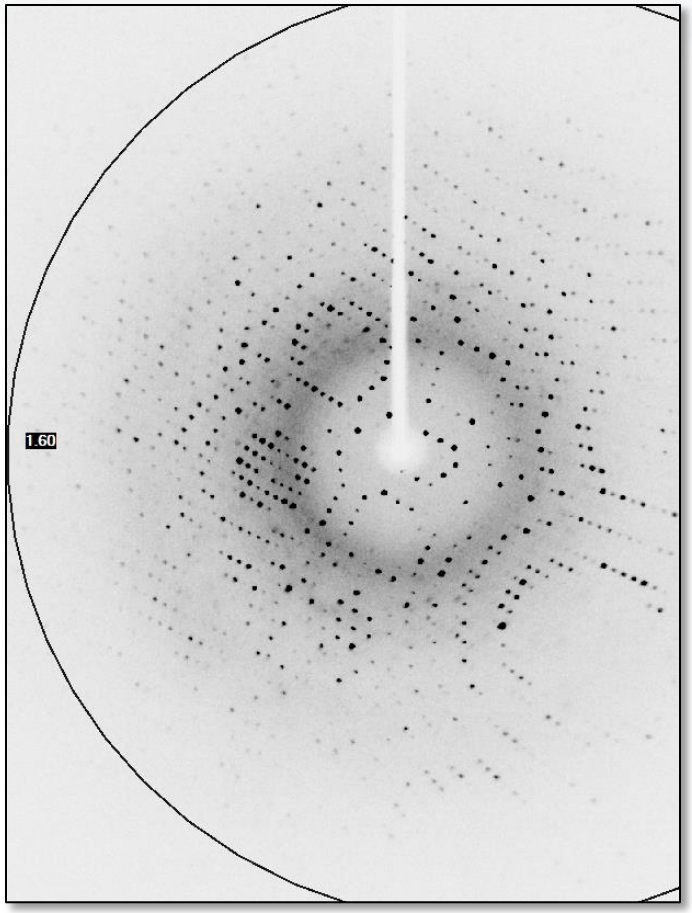


XE, SHELXE
Figures of merit

- Contrast
  - Comparison of the variance for the solvent and protein regions
  - Should be higher for the correct HA enantiomorph
  - Clear separation usually indicates a correct substructure
- Connectivity
  - Fraction of adjacent pixels that have similar assignments
  - Should be higher for the correct HA enantiomorph
- Model correlation coefficient
  - CC for the trace and native data
  - Above 25% usually means that the model is correct

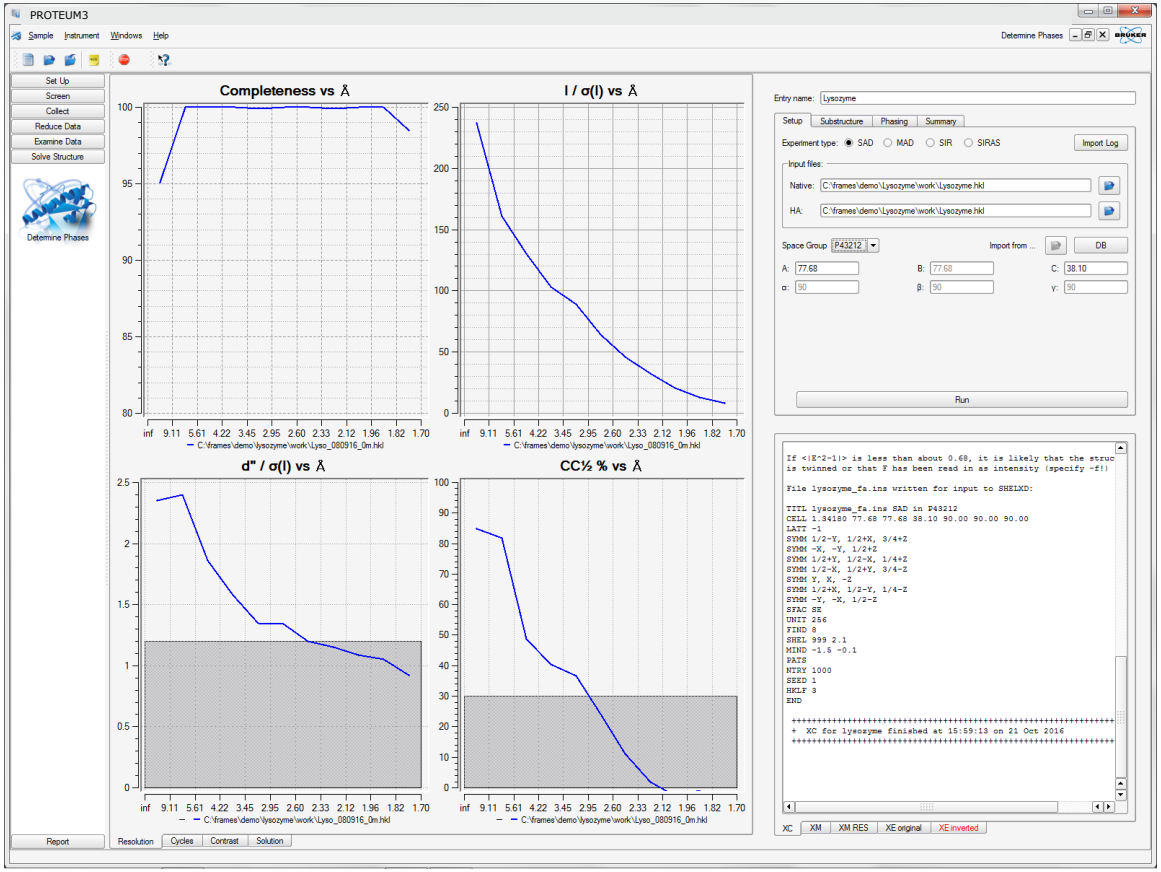
Lysozyme
Native SAD phasing

- Space group $P4_32_12$
- Cell constants $a=b=77.6$, $c=38.1$ Å
- Data collected 165°
- Exposure time 2 sec/image
- Image width 0.5°
- Wall time 11 min
- Max resolution 1.6 Å
- Rmerge 3.12 (17.6)
- $I/\sigma I$ 33.3 (7.5)
- Multiplicity 8.1 (11)
- $CC_{½}$ >97%
- D8 VENTURE
  - METALJET, GaKα
  - PHOTON II



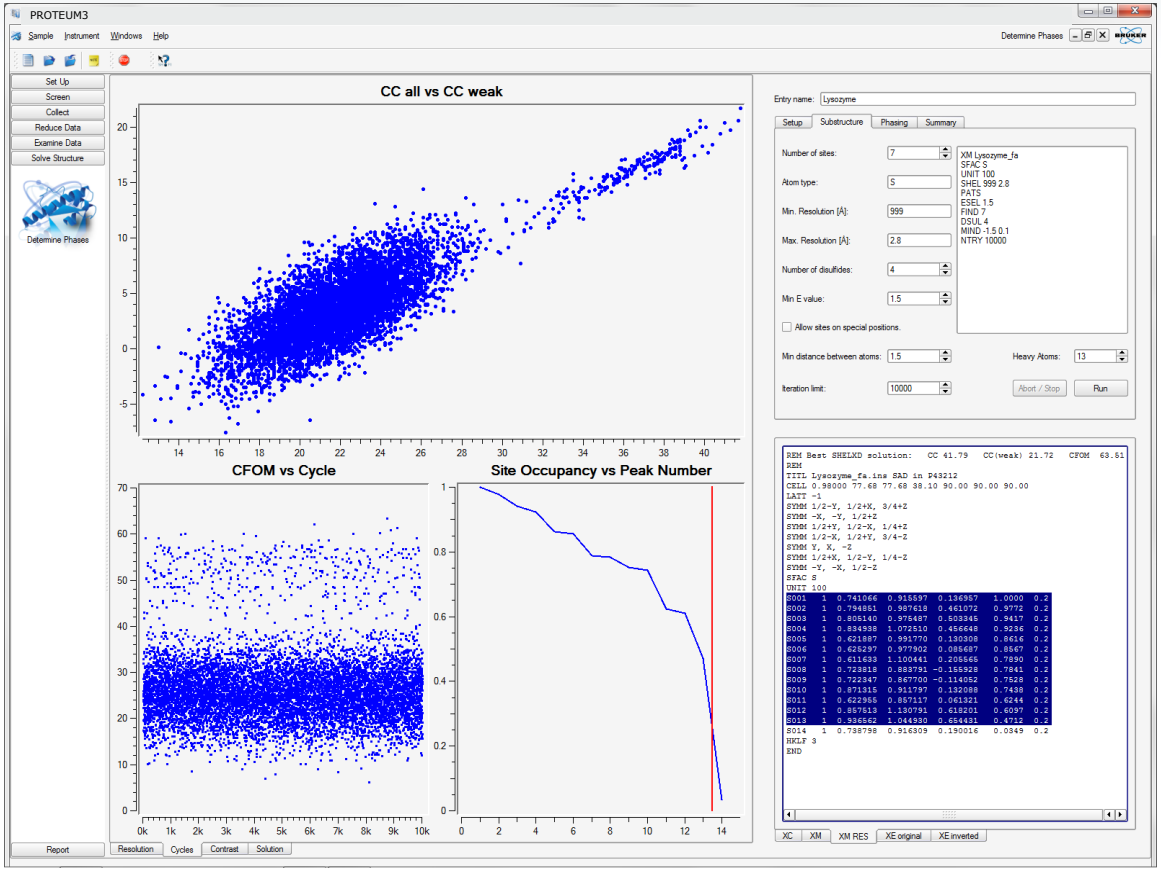
Lysozyme
Setup

- SHELXC
- Based on the data metrics, truncate the data at 2.6-2.8 Å to conduct substructure search



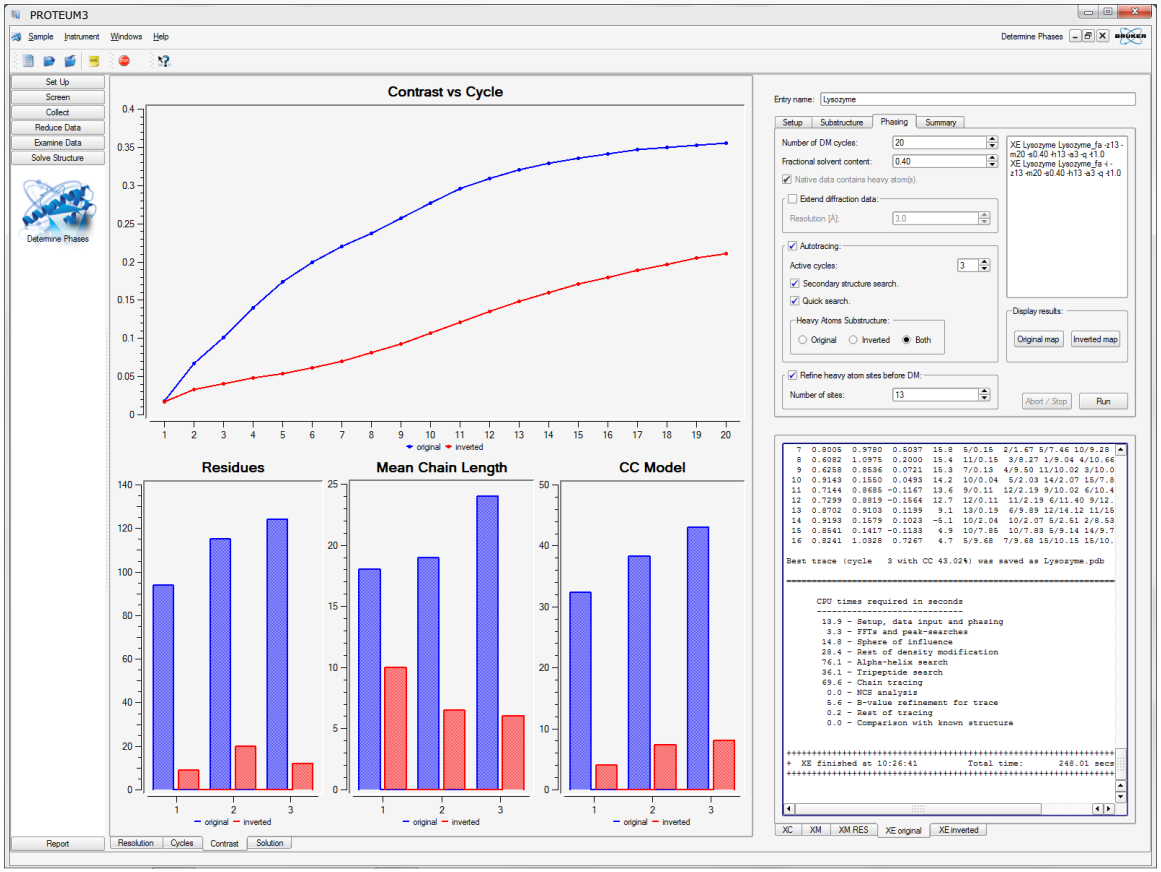
Lysozyme
Substructure

- Input the number of sites based on the sequence or know metal sites
- Select the resolution limit, try lowering in steps of 0.5 Å to find a solution
- Select disulfides if know to be present in structure, these should be input as a single site
- 1000 trials is a good start
- If there are two distinct groups of solutions, that's a good sign that you have a correct substructure
- Look for a separation between the real peaks and noise for phasing



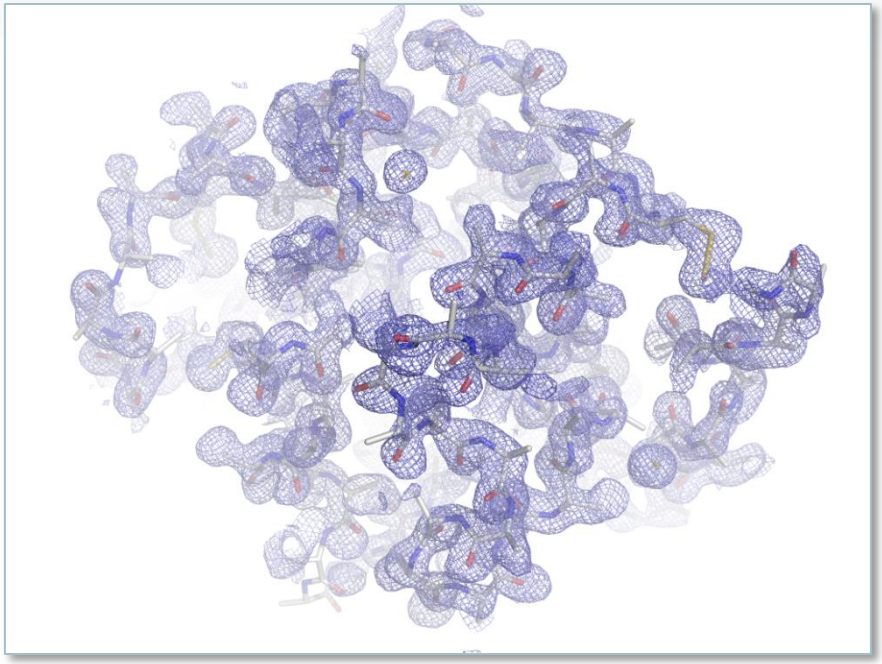
Lysozyme
Phasing

- Input the number of DM cycles, 20 works well
- Select the solvent content
- For autotracing, 3-6 cycles is a good start
- Select the method of model tracing and which enantiomorph to use
- Can refine the HA sites
- Usually a good sign if there is a clear difference in the contrast between enantiomorphs
- Residues fit, Mean chain length and the Model CC are displayed



Lysozyme
Native SAD phasing

- SHELXD
  - 10 Sulfur sites
    - 8 Cys
    - 2 Met
    - 3 other sites
  - Resolution 2.8 Å
  - Cycles 1000
- SHELXE
  - Residues found 124
  - Final CC 43.02

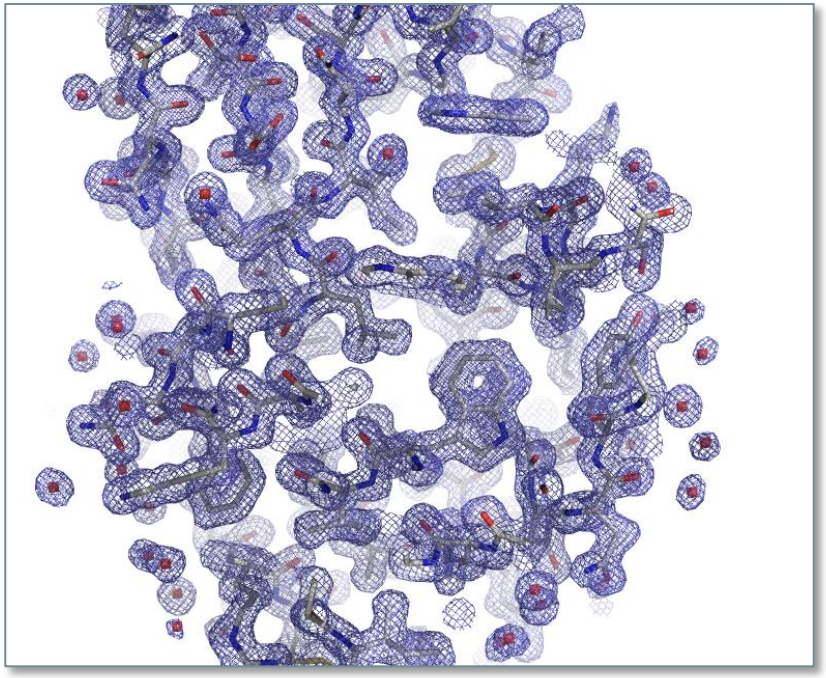


Electron density map calculated using Native SAD phases from SHELXE

Lysozyme
Native SAD phasing

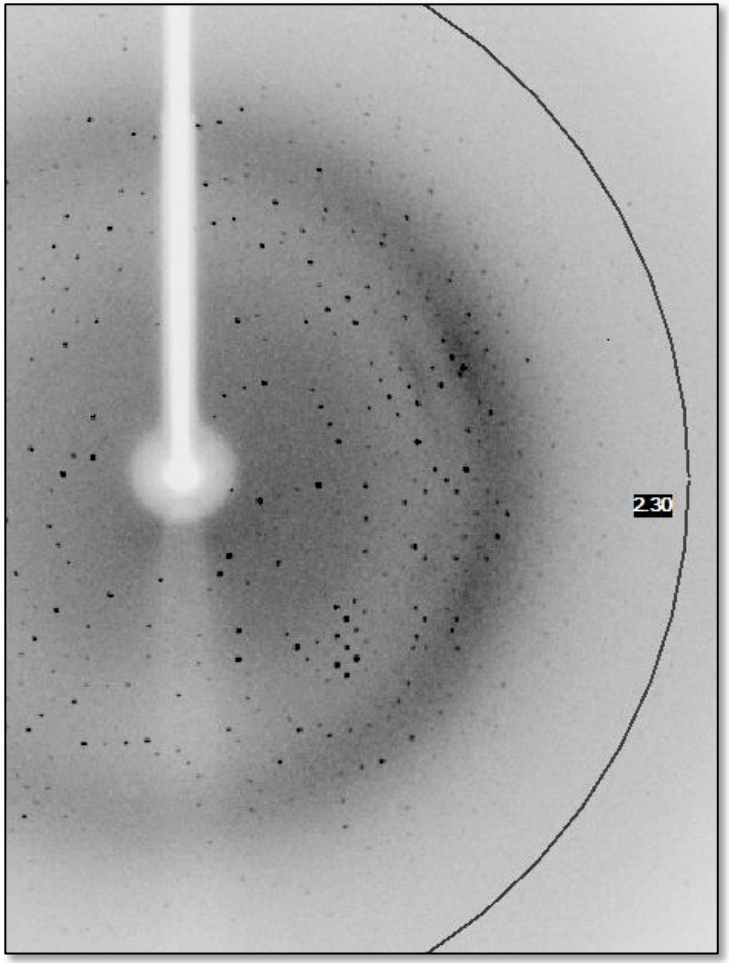
- Phases from SHELXE input into Buccaneer (CCP4)
  - 129 residues fit
- Refmac results

Final statistics	
R factor	0.1840
R free	0.2096
Rms Bond Length	0.010
Rms Bond Angle	1.39
Rms Chiral Volume	0.089



Glycoside Hydrolase PSLG
Native SAD phasing with Cd

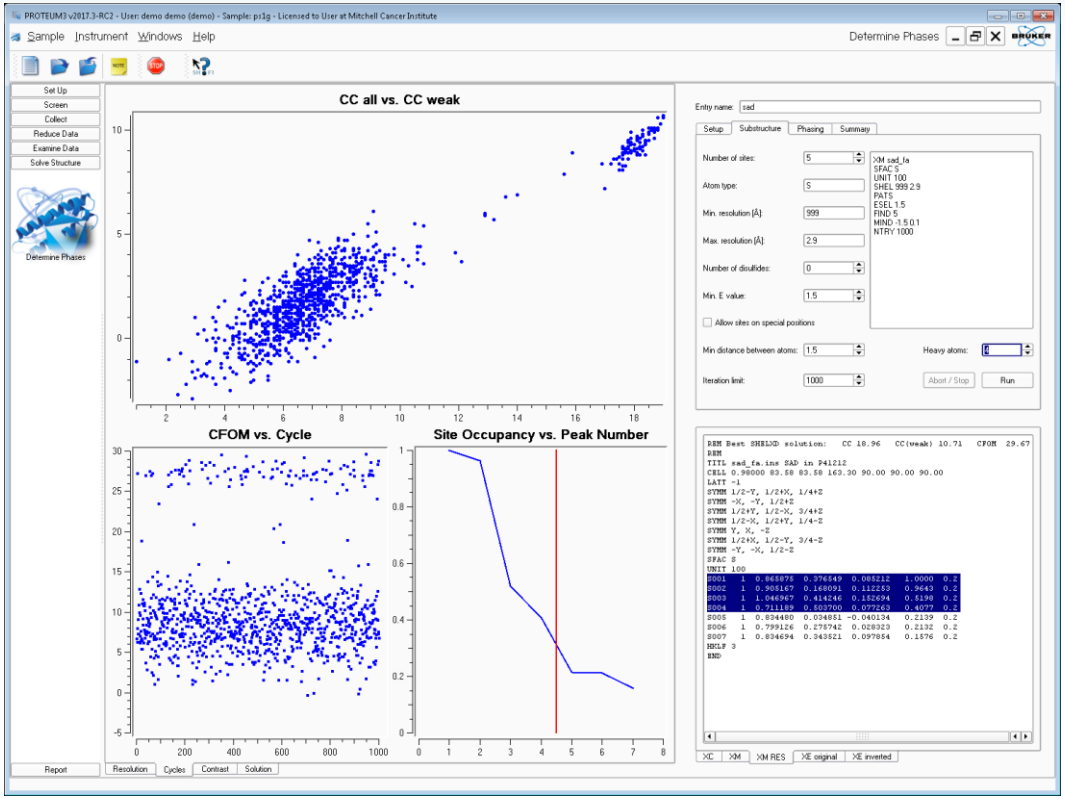
- *Pseudomonas aeruginosa*
- 416 aa
  - 4 Cd bound
  - 9 MET
- Space group $P4_12_12$
- Cell constants $a=b=86.3$, $c=163.4$ Å
- Data collected 225°
- Exposure time 180 sec/deg
- Wall time 11.3 hrs
- Max resolution 2.25 Å
- Rmerge 11.7 (44.9)
- I/σI 14.6 (3.37)
- Multiplicity 6.3 (7.6)
- D8 VENTURE
  - METALJET, GaKα
  - PHOTON II



Glycoside Hydrolase PSLG
Substructure

- SHELXD finds the four Cd sites

Sites	X	Y	Z	OCC
S1	0.094476	0.828214	0.113359	1.000
S2	0.133167	0.623676	0.084805	0.818
S3	0.293849	0.494024	0.074887	0.412
S4	-0.05802	0.581758	0.153117	0.392
S5	0.064199	0.799539	0.124275	0.150
S6	0.120617	0.862052	0.126271	0.032



Data metrics

Res	9.11	5.61	4.22	3.45	2.95	2.6	2.33	2.12	1.96	1.82	1.7
<ΔF/σ>	1.69	1.36	1.08	0.97	0.86	0.82	0.76	0.74	0.72	0.68	0.69
CC(1/2)	47.4	51.5	30.2	24.5	5.3	-10.1	-1.7	1.6	-3	-10	-11.3

Glycoside Hydrolase PSLG
Phasing

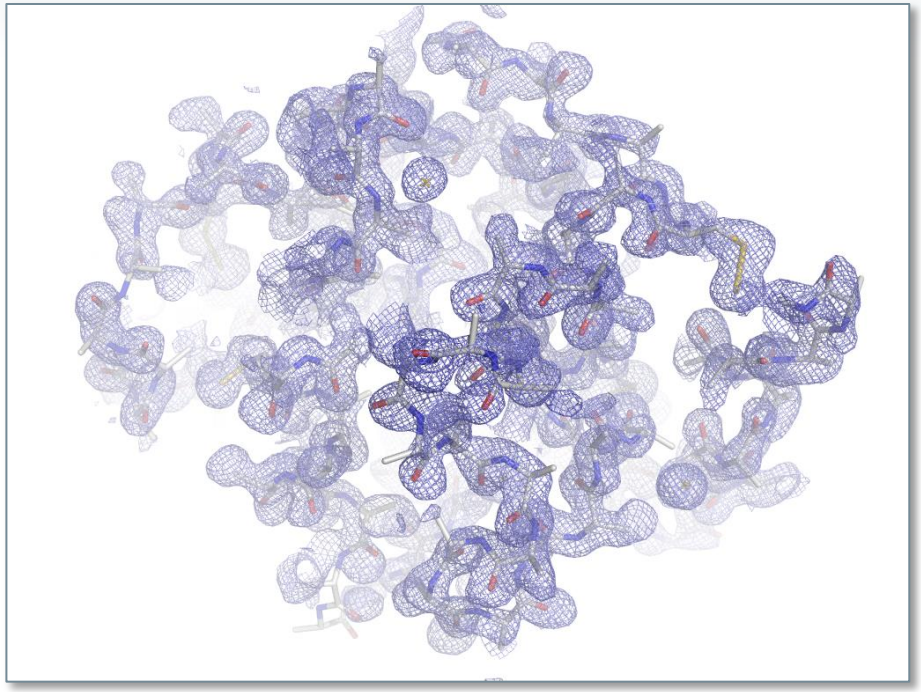
- Clear difference in the contrast between the different HA hands
- Takes a while but eventually the phase error drops enough where most of the backbone can be fit

Glycoside Hydrolase PSLG
Native SAD phasing

- SHELXD
  - 4 Cd sites
  - Resolution 2.9 Å
  - Cycles 1000
- SHELXE
  - Residues found 312
  - Final CC 34.9

Anomalous differences (σ)*	
19.31	CD_CD
2.07	SD_MET
1.03	CA_HIS
0.97	CE_MET
0.8	CG_MET
0.75	C_HIS
0.63	CB_HIS
0.59	O_HIS
0.59	OE1_GLU

*ANODE, Thorn (2011)

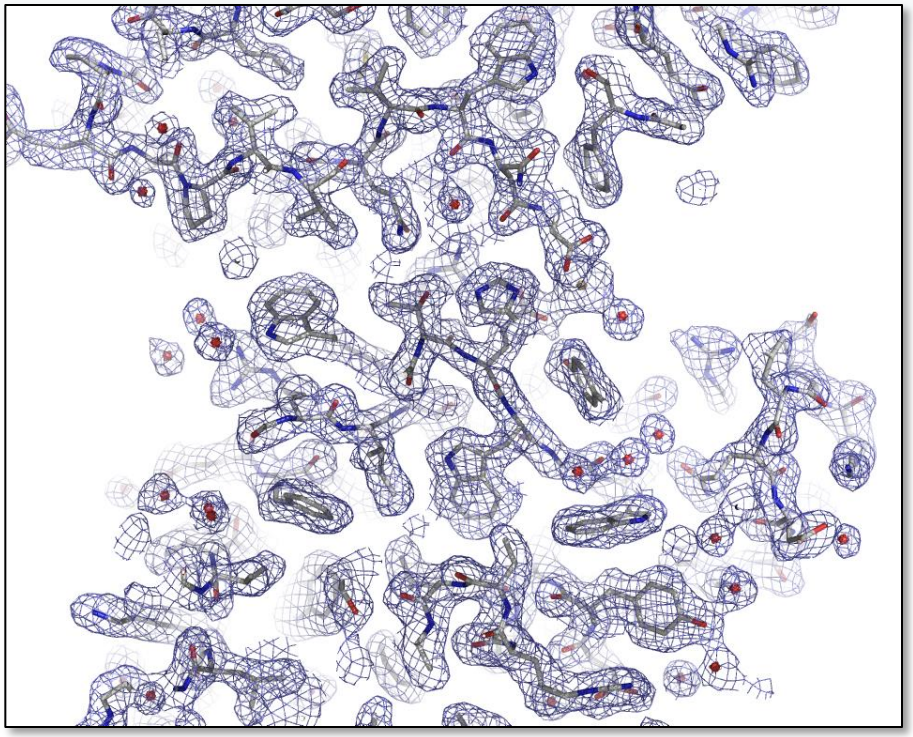


Electron density calculated from SHELXE phases. The final model was superimposed on the map using Molrep

Glycoside Hydrolase PSLG
Native SAD phasing

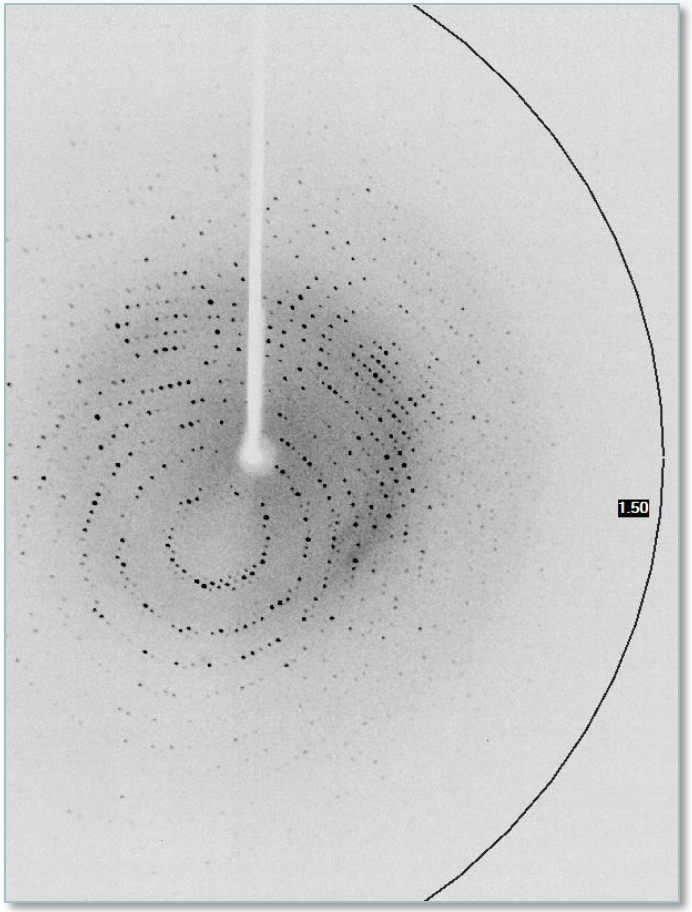
- Buccaneer (CCP4)
  - Fit 416 residues
  - Manual building
  - Solvent search

Final statistics	
R factor	0.179
R free	0.225
Rms Bond Length	0.014
Rms Bond Angle	1.62
Rms Chiral Volume	0.094



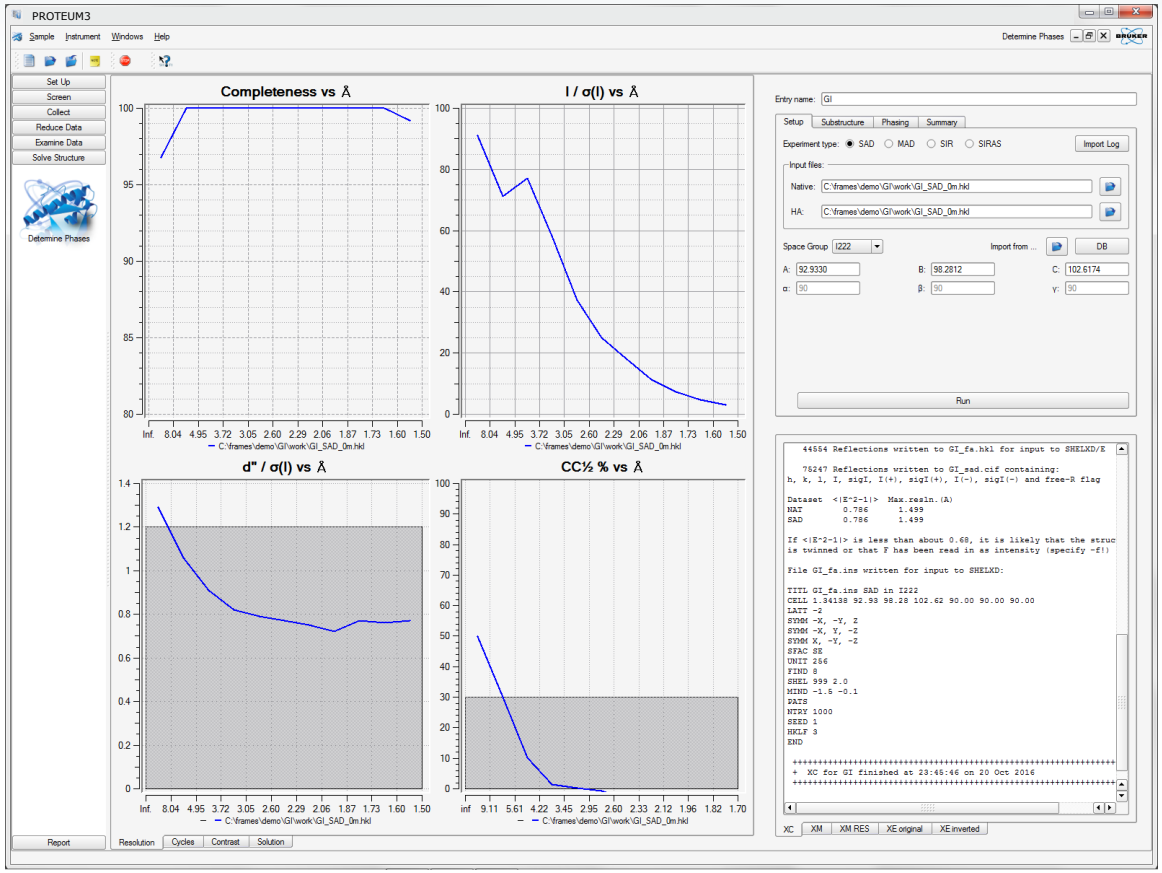
Glucose Isomerase
Native SAD phasing

- *Streptomyces rubiginosus*
- Space group I222
- Cell constants *a*=92.9, *b*=98.3, *c*=102.6 Å
- Data collected 370°
- Exposure time 2 sec/image
- Image width 0.2°
- Wall time 1 hour
- Max resolution 1.5 Å
- Rmerge 6.89 (54.8)
- I/σI 19.2 (2.73)
- Multiplicity 8.7 (13.5)
- $CC_{½}$ >80%
- D8 VENTURE
  - METALJET, GaKα
  - PHOTON II



Glucose Isomerase
Setup

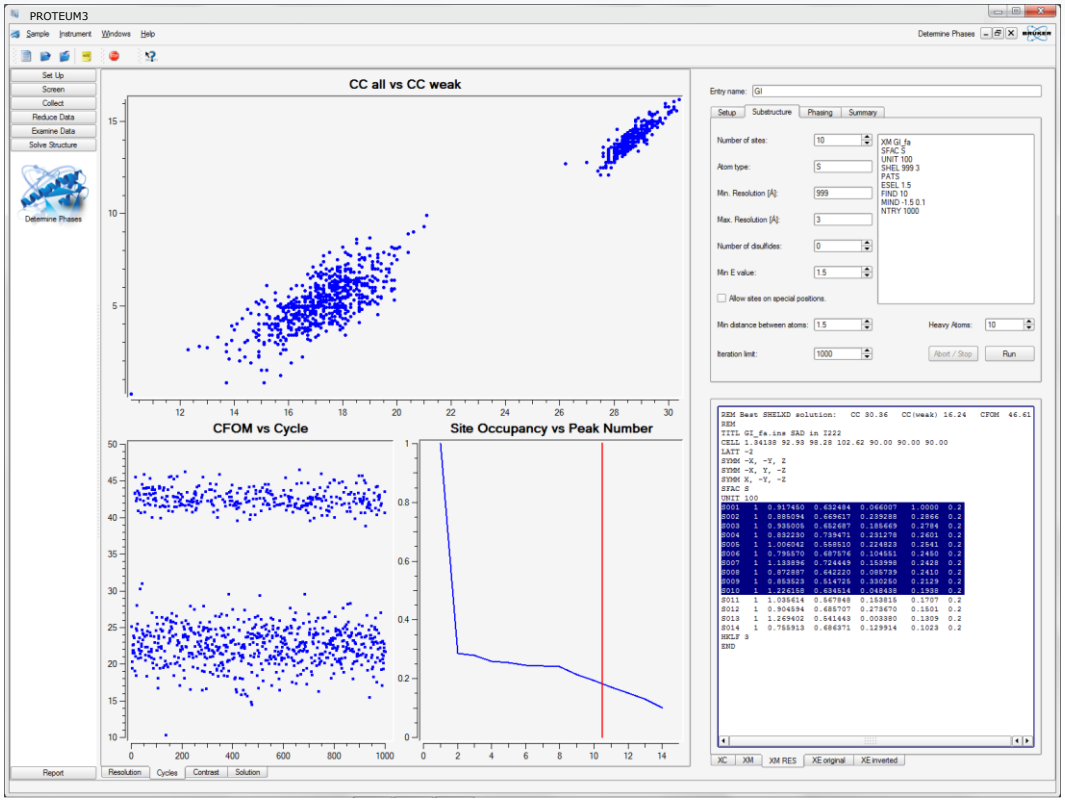
- From the statistics, it appears that finding a substructure might be difficult



Glucose Isomerase
Substructure

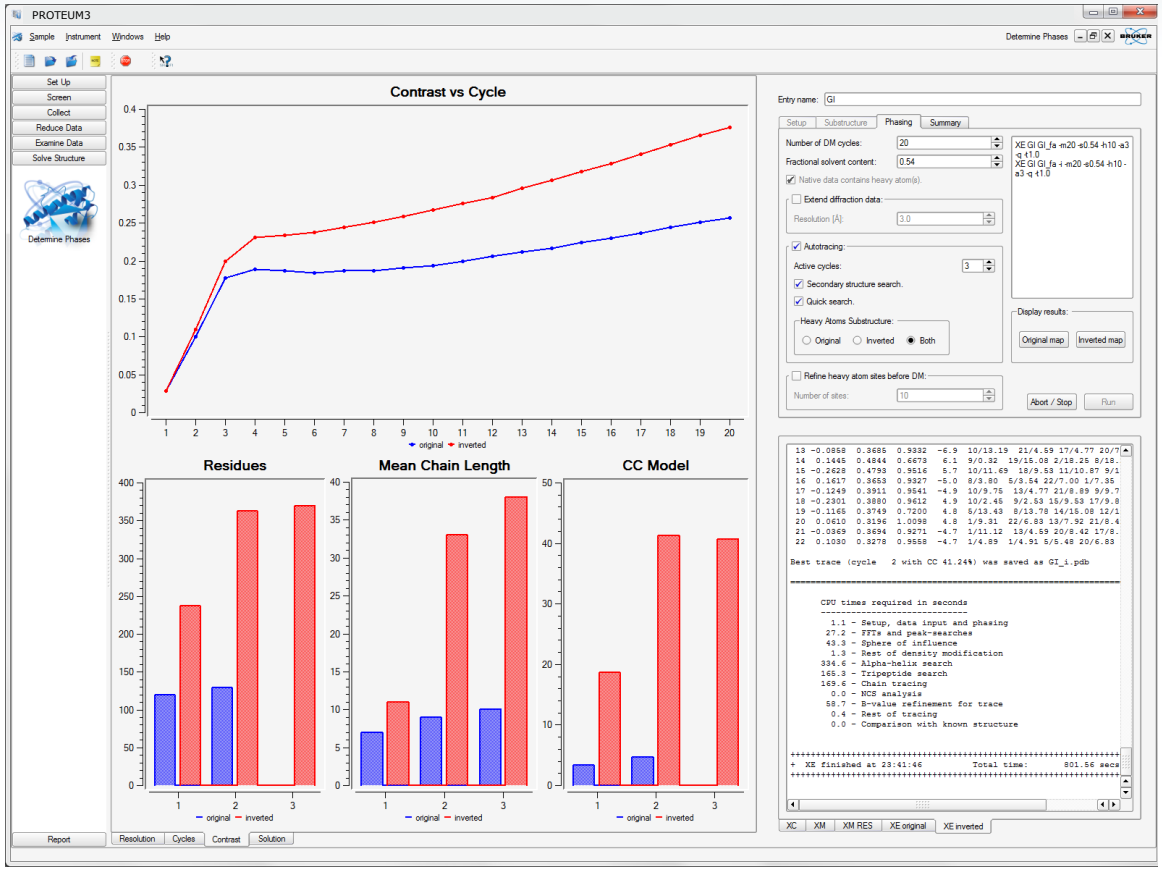
- 388 residues
- 7 Met, 1 Cys
- 2 potential metal binding sites, Mg
- From the CC values, there are clearly two groups of possible solutions
- With a mix of metals and sulfur sites, the metals will have a much higher occupancy

Sites	X	Y	Z	OCC
S1	0.91745	0.632484	0.066007	1.000
S2	0.885094	0.669617	0.239288	0.2866
S3	0.935005	0.652687	0.185669	0.2784
S4	0.83223	0.739471	0.231278	0.2601
S5	1.006042	0.55851	0.224823	0.2541
S6	0.79557	0.687576	0.104551	0.245
S7	1.133896	0.724449	0.153998	0.2428
S8	0.872887	0.64222	0.085739	0.241
S9	0.853523	0.514725	0.33025	0.2129
S10	1.226158	0.634514	0.048438	0.1938



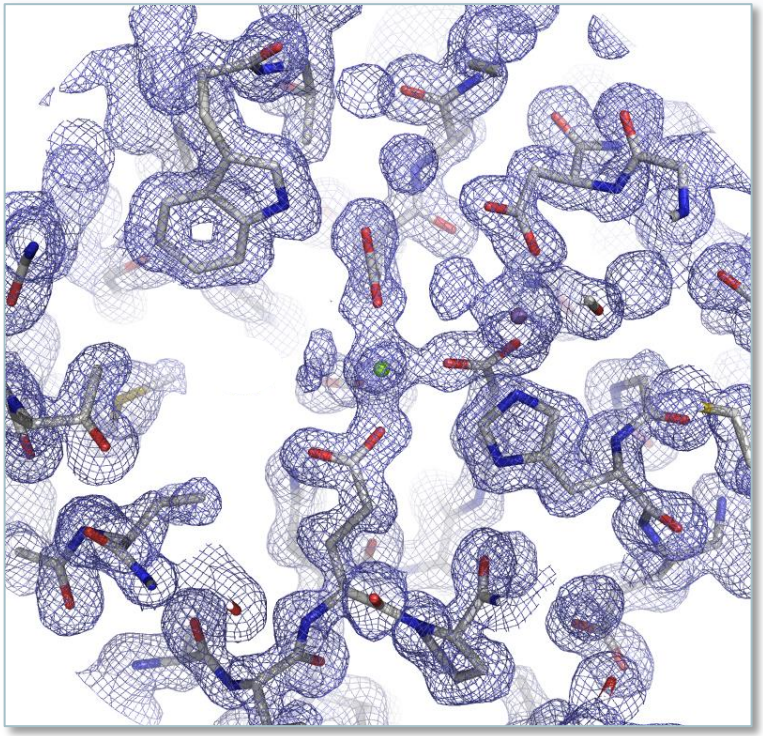
Glucose Isomerase
Phasing

- Clear difference in the contrast between enantiomorphs indicates which heavy atom hand is correct
- Traced 363 residues
- Final CC for the model was 41 %

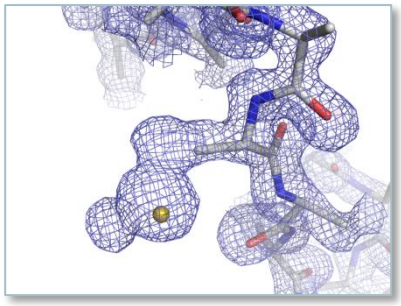


Glucose Isomerase
Native SAD phasing

- SHELXD
  - 10 sites
    - 2 metal (Mn, Mg)
    - 7 Met
    - 1 Cys
  - Resolution 3.0 Å
  - Cycles 1000
- SHELXE
  - Resides found 363
  - Final CC 41.24 %



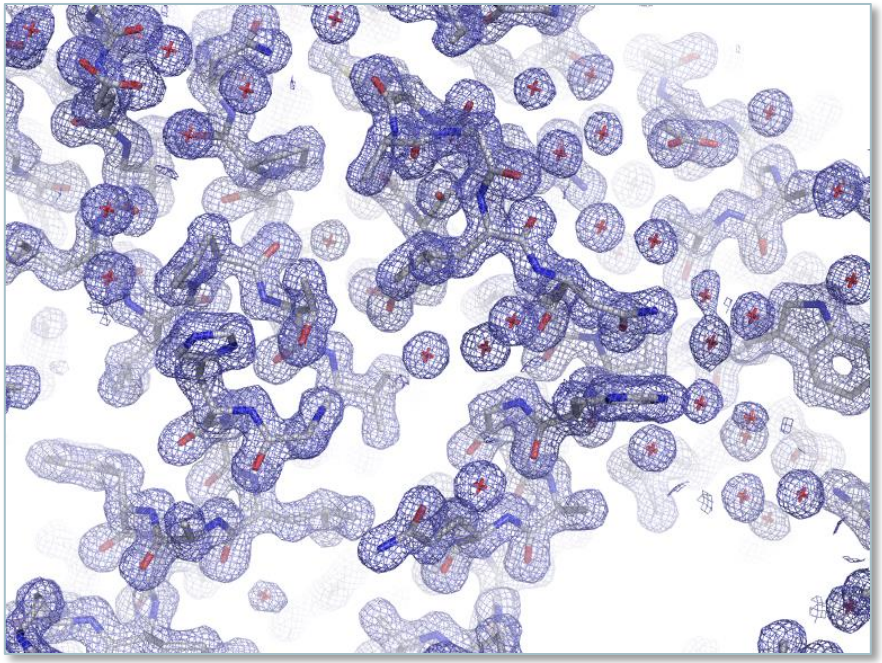
Electron density map calculated using Native SAD phases from SHELXE



Glucose Isomerase
Native SAD phasing

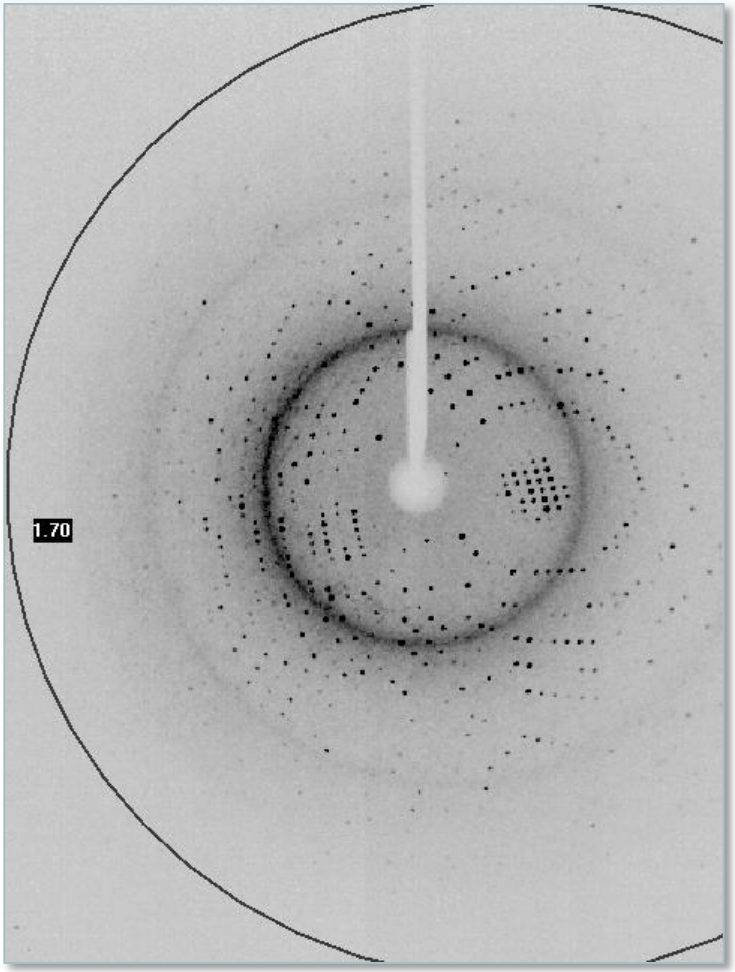
- Phases from XE input to PHENIX
  - 382 residues fit
- Refinement results

Final statistics	
R factor	0.216
R free	0.230
Rms Bond Length	0.010
Rms Bond Angle	1.22
Avg B-factor	13.9



CYCLODIPHOSPHATE SYNTHASE
Native SAD phasing

- Space group C2
- Cell constants *a*=117.2, *b*=68.2, *c*=60.0 Å
  β=96.6°
- Zn-dependent enzyme
- Trimer, 186 residues
  - 3 Met
  - 1 Cys
- Wall time 6.5 hrs
- D8 VENTURE
  - IμS, CuKα
  - PHOTON II



CYCLODIPHOSPHATE SYNTHASE
Data collection

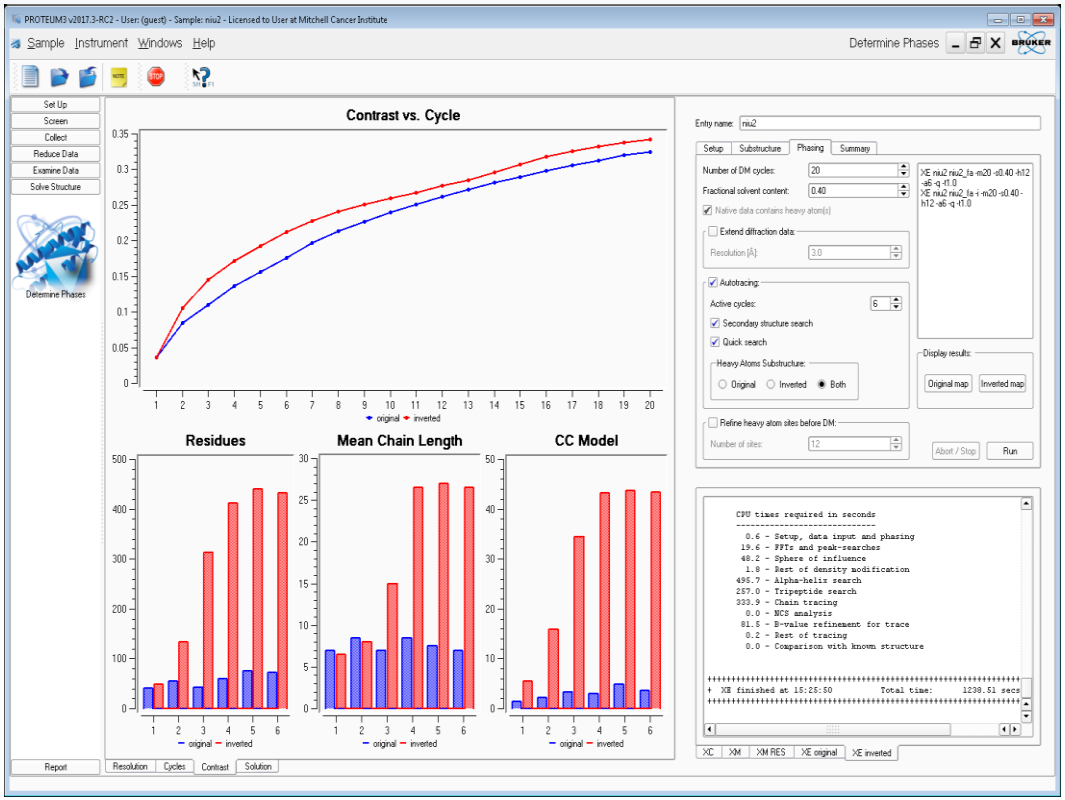
- Collect high and low angle data separately
  - High angle 40 sec/deg
  - Low angle 13 sec/deg
- Only integrate the low angle runs at the substructure cutoff (2.5 Å)

	High Angle	Combined
Resolution (Å)	18.0 –1.70	18.0 –1.70
Rmerge	0.052 (0.58)	0.051 (0.58)
<I/σI>	24.5 (2.58)	28.2 (2.58)
CC (1/2)	(65.0)	(65.0)
Completeness	99.0	99.9
Multiplicity	7.0 (9.1)	9.3 (16.5)
Time	5.3 hrs	1.3 hrs

CYCLODIPHOSPHATE SYNTHASE

Native SAD phasing

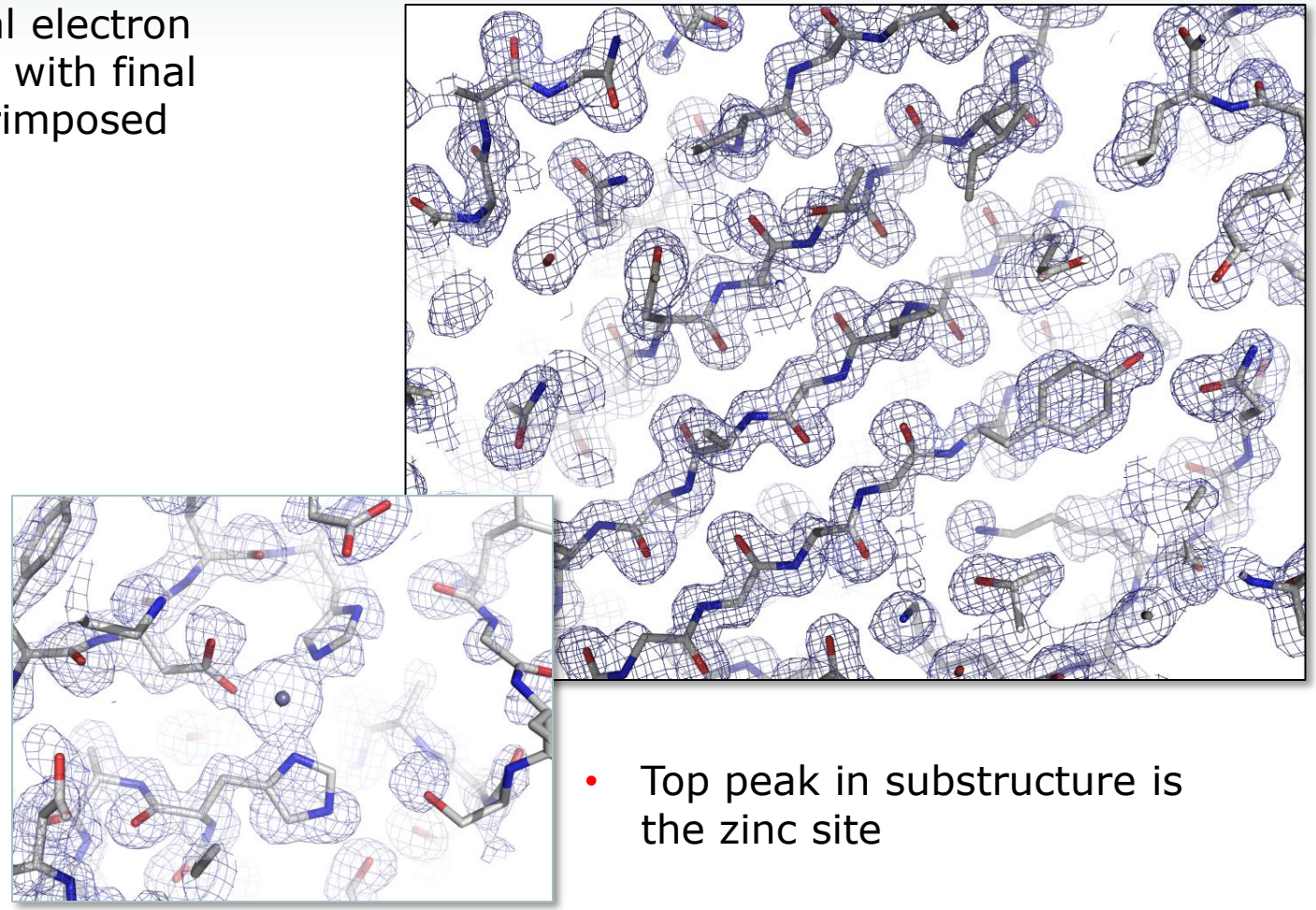
- SHELXD
  - 12 sites
    - Zn
    - 11 Met
    - 3 Cys
  - Resolution 2.5 Å
  - Cycles 1000
- SHELXE
  - Resides found 435
  - Final CC 45.3 %



Res	9.11	5.61	4.22	3.45	2.95	2.6	2.33	2.12	1.96	1.82	1.7
<ΔF/σ>	1.75	1.28	1.1	0.9	0.77	0.76	0.7	0.66	0.68	0.72	0.8
CC(1/2)	44.4	43.6	26.3	7.3	5.3	2.4	-2.6	-2.8	-6.5	-13	-20.3

CYCLODIPHOSPHATE SYNTHASE
Native SAD phasing

- Experimental electron density map with final model superimposed

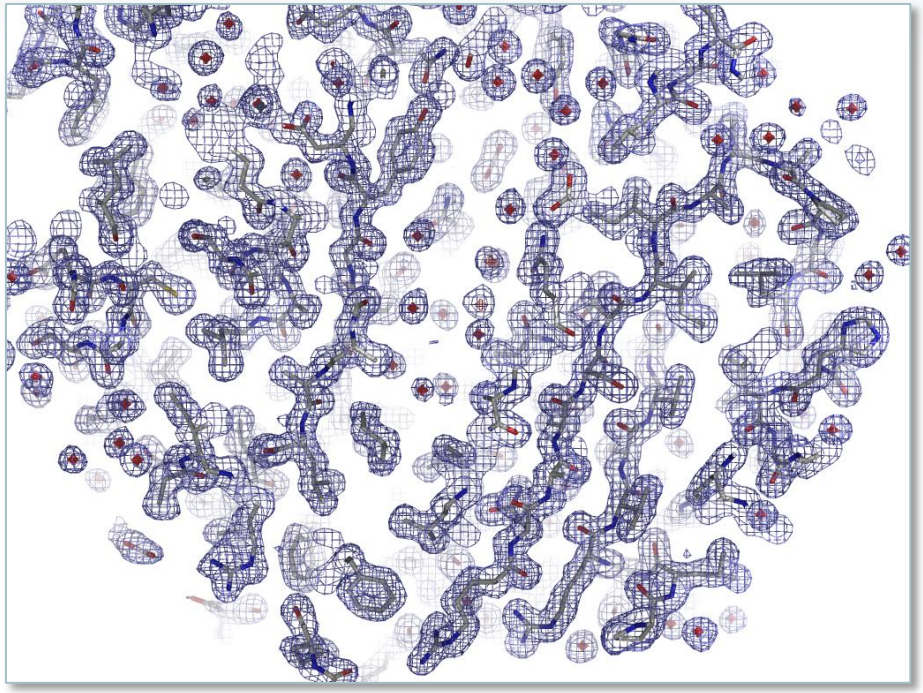


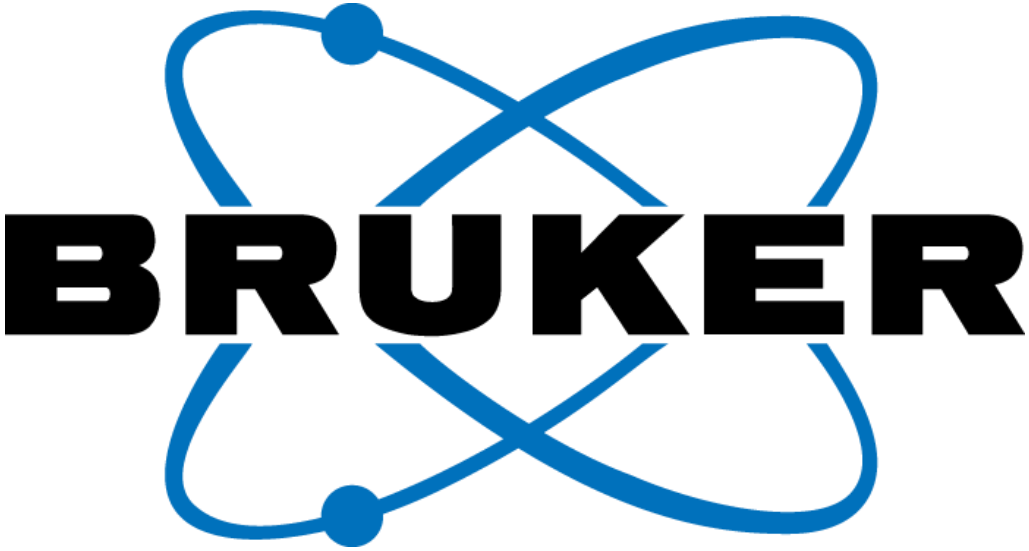
- Top peak in substructure is the zinc site

CYCLODIPHOSPHATE SYNTHASE
Native SAD phasing

- Phases from XE input to Buccaneer
  - 490 residues fit
- Model tweaked and solvent search
- Refinement results

Final statistics	
R factor	0.186
R free	0.224
Rms Bond Length	0.010
Rms Bond Angle	1.36
Rms ChirVolume	0.077





Innovation with Integrity