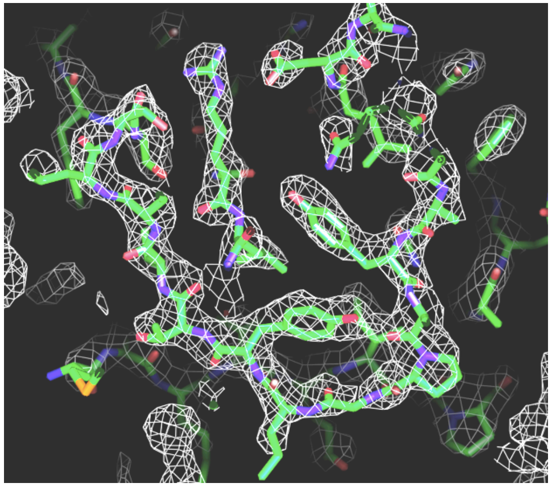
Phenix

Planning and carrying out automated structure determination using SAD phasing

SBGrid/NECat Phenix Workshop
Boston, November 10, 2016

Tom Terwilliger, Li-Wei Hung, Los Alamos National Laboratory
Randy Read, Airlie McCoy, University of Cambridge
Pavel Afonine, Paul Adams, Lawrence Berkeley National Laboratory

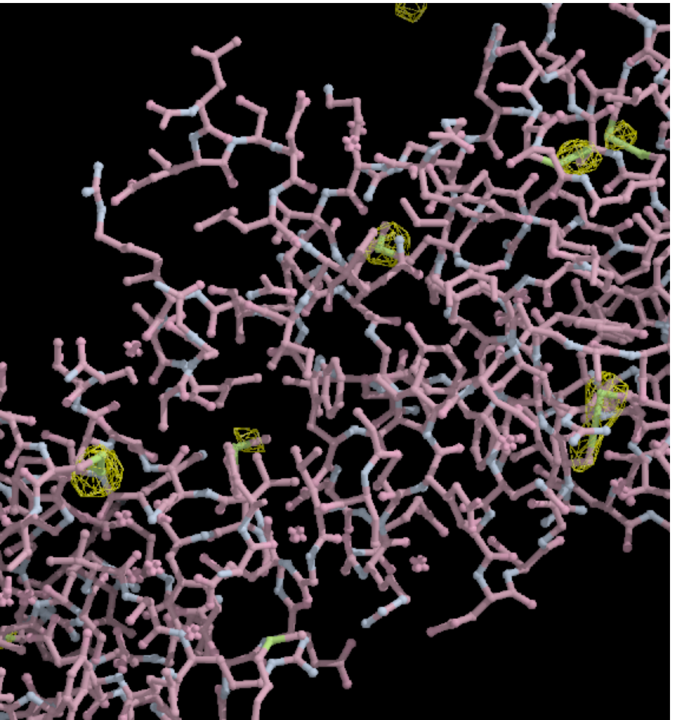
Steps in Single Wavelength Anomalous Diffraction (SAD) Structure Determination



- **Plan the experiment**
- **Measure the data**
- **Scale the data**
- **Evaluate the accuracy of the anomalous differences**
- **Find the anomalous sub-structure**
- **Identify hand of sub-structure**
- **Calculate experimental phases and a map**
- **Improve the map with density modification**
- **Build and refine a model**

Planning a SAD experiment

Will I find the sites of anomalously-scattering atoms?

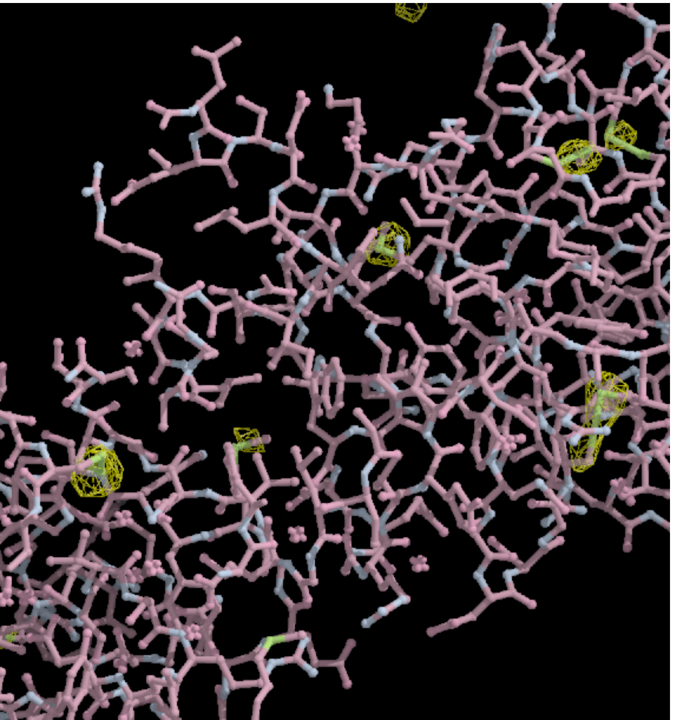


Planning a SAD experiment

How many sites?
How many reflections?

What is the anomalously-scattering atom?
What is the wavelength?

Are the sites (on average) well ordered?
Are the data well-measured?

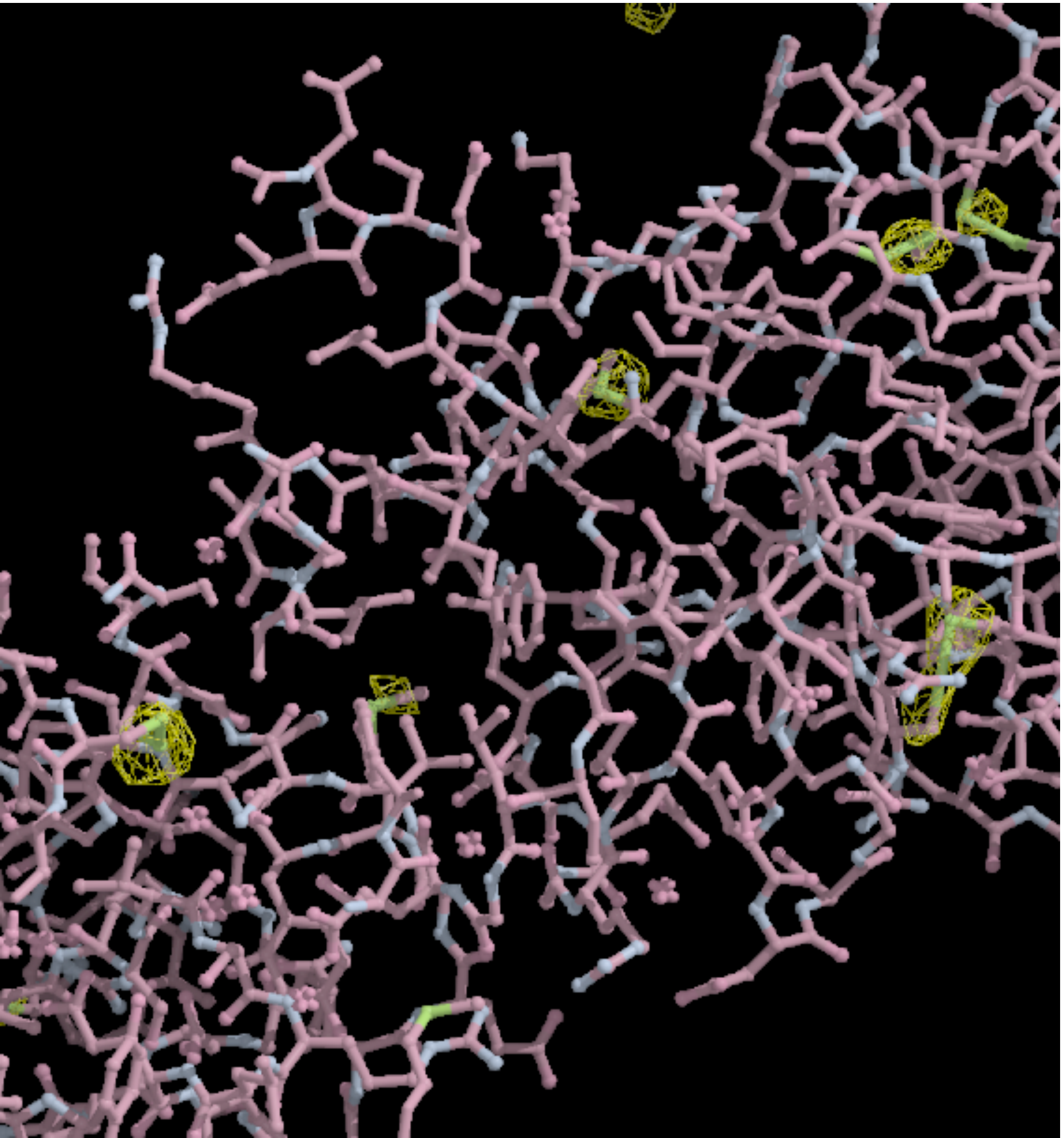


Maximizing the anomalous signal and the anomalous correlation

The ***anomalous correlation*** CC_{ano} *is a measure of the accuracy of each anomalous difference*
(correlation to ideal anomalous data from your structure)

The ***anomalous signal*** *is a measure of how much total information per site is present in the anomalous differences*
(peak height in anomalous difference Fourier)

Anomalous signal

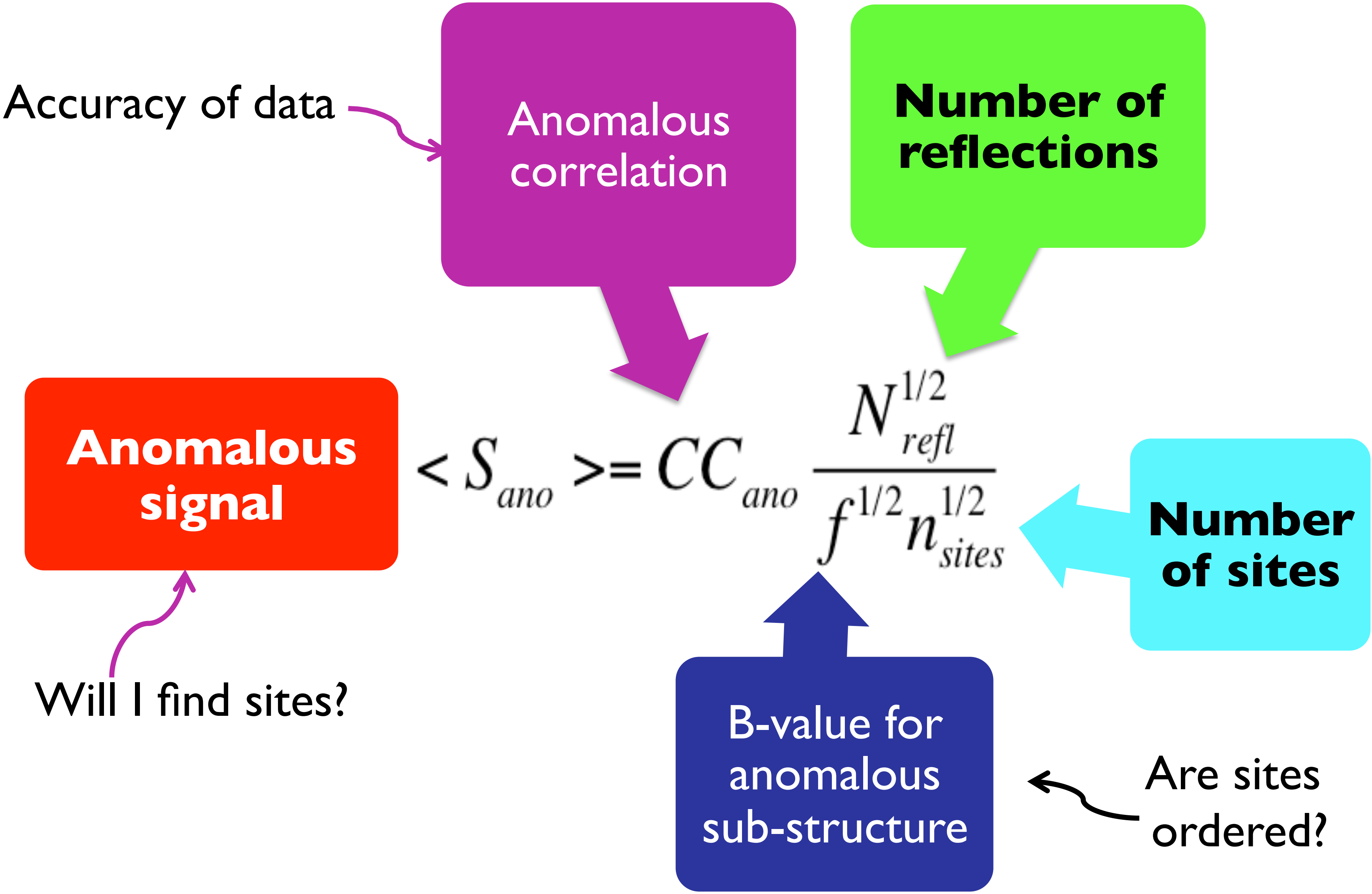


Peak height in anomalous difference Fourier at coordinates of anomalously-scattering atoms

Typical values of S_{ano} for solved datasets: 10-20

Anomalous difference Fourier with observed data and model phases

What determines if I will find the sites?



Accuracy of data

Anomalous correlation

Number of reflections

Anomalous signal

$$< S_{ano} > = CC_{ano} \frac{N_{refl}^{1/2}}{f^{1/2} n_{sites}^{1/2}}$$

Number of sites

Will I find sites?

B-value for anomalous sub-structure

Are sites ordered?

How big will my anomalous signal be?

Expected value of anomalous signal S_{ano}

$$< S_{ano} >= CC_{ano} \frac{N_{refl}^{1/2}}{f^{1/2} n_{sites}^{1/2}}$$

f is 2nd moment of the anomalous scattering factor

(f is large if B-value for anomalously-scattering atoms is high)

$$f = \frac{< (f^{h})^{2} >}{< f^{h} >^{2}}$$

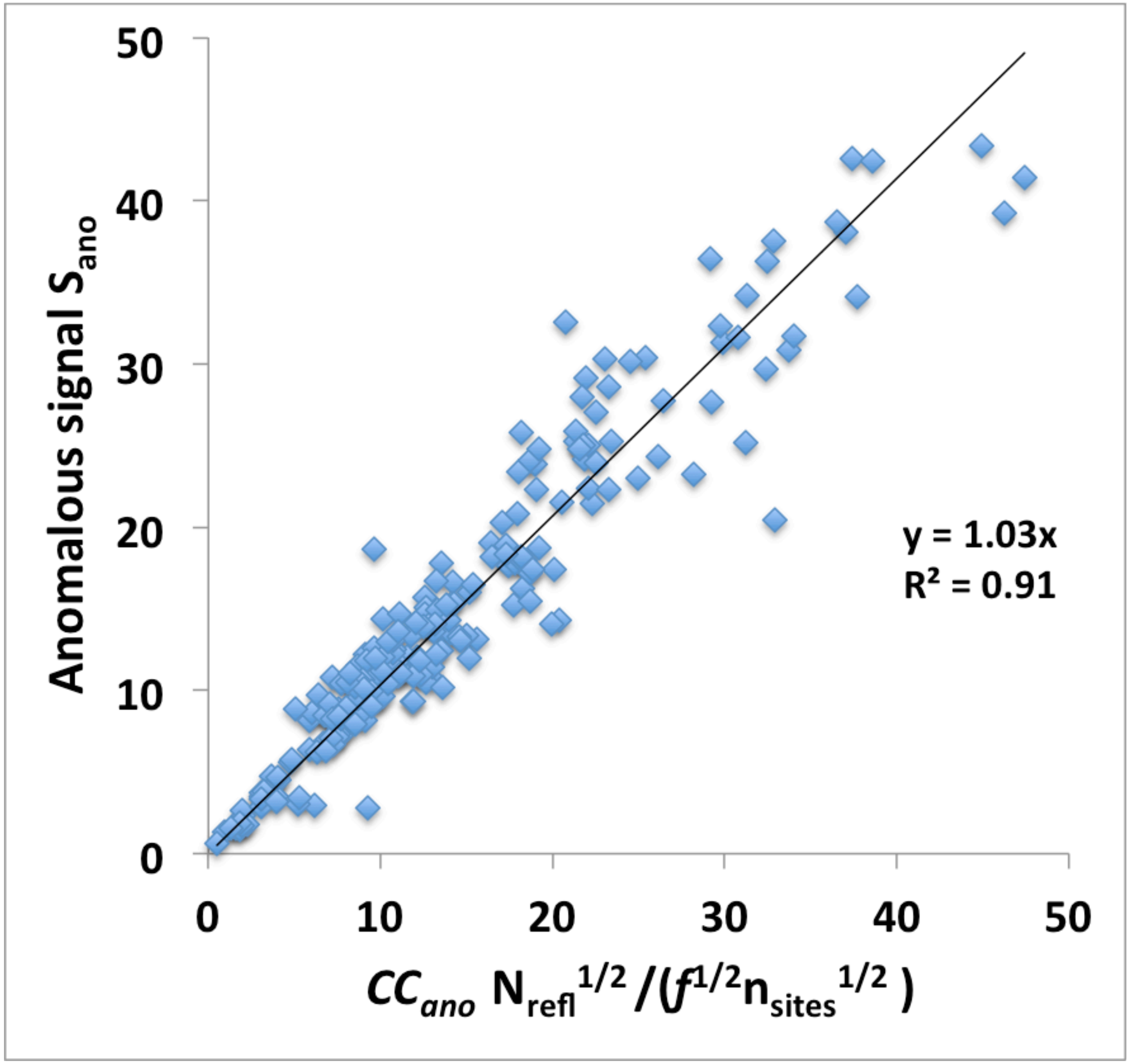
(f^{h} in this equation is the anomalous scattering factor)

$$f^{h} \equiv f'' e^{-B \, (\sin^{2} \theta_{h} / \lambda^{2})}$$

Perfect data (20,000 reflections, 8 sites): S_{ano}= $(20000/8)^{1/2}$ = 50

Good data (overall CC_{ano}=0.36 f=2.0): S_{ano}= 12.6

Checking our simple model for anomalous signal



$$< S_{ano} > = CC_{ano} \frac{N_{refl}^{1/2}}{f^{1/2} n_{sites}^{1/2}}$$

CC_{ano}: Correlation of anomalous differences with model differences

S_{ano}: Peak height in model-phased difference Fourier

218 SAD datasets 1.2 – 4.5 Å

CysZ multi-crystal sulfur-SAD data

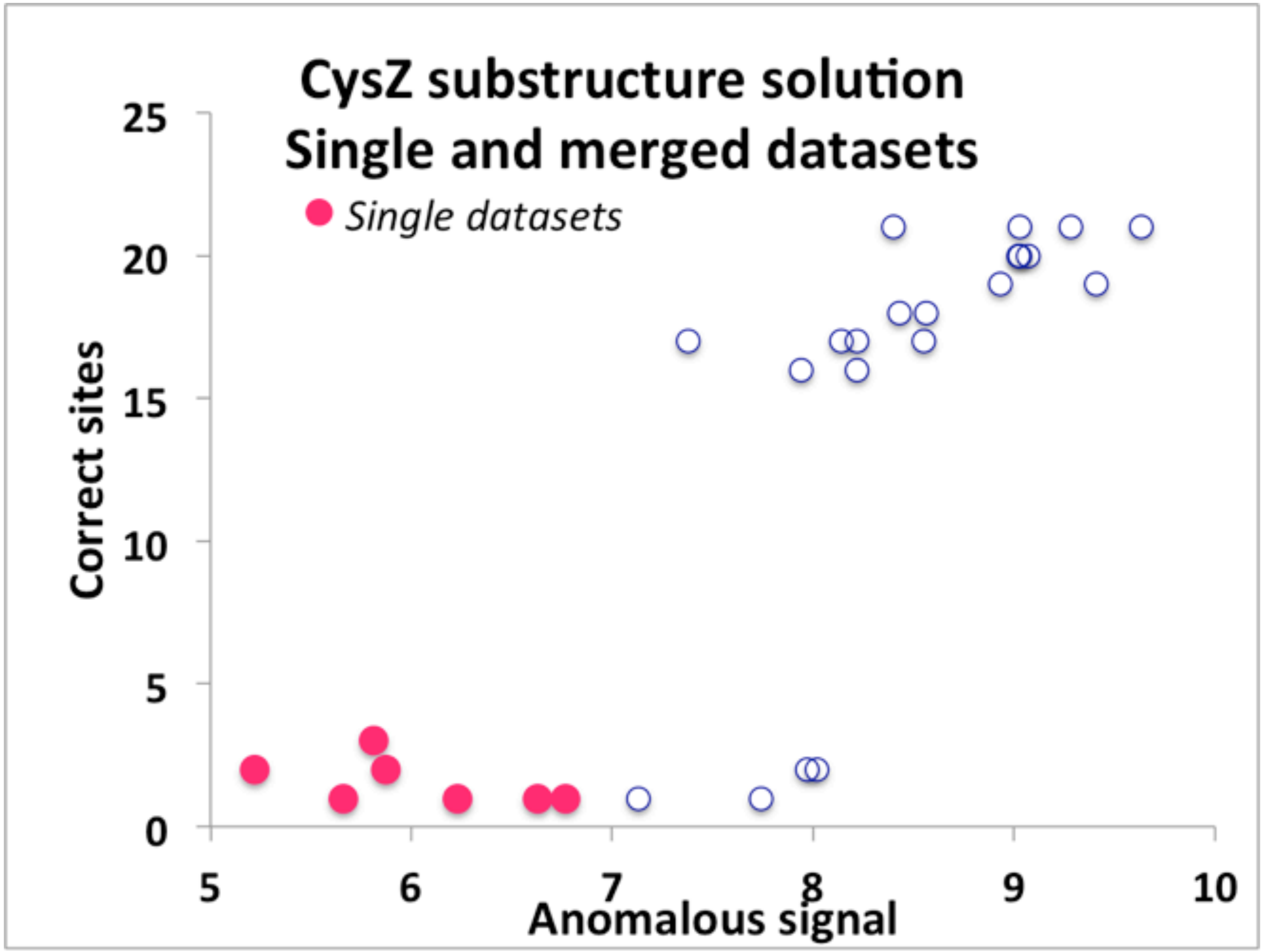
Qun Liu, Tassadite Dahmane, Zhen Zhang, Zahra Assur, Julia Brasch, Lawrence Shapiro, Filippo Mancia, Wayne Hendrickson (2012). Science 336, 1033-1037

Data from 7 crystals collected at wavelength of 1.74 Å to resolution of 2.3 Å

Can anomalous signal tell us which merged datasets will be solved?

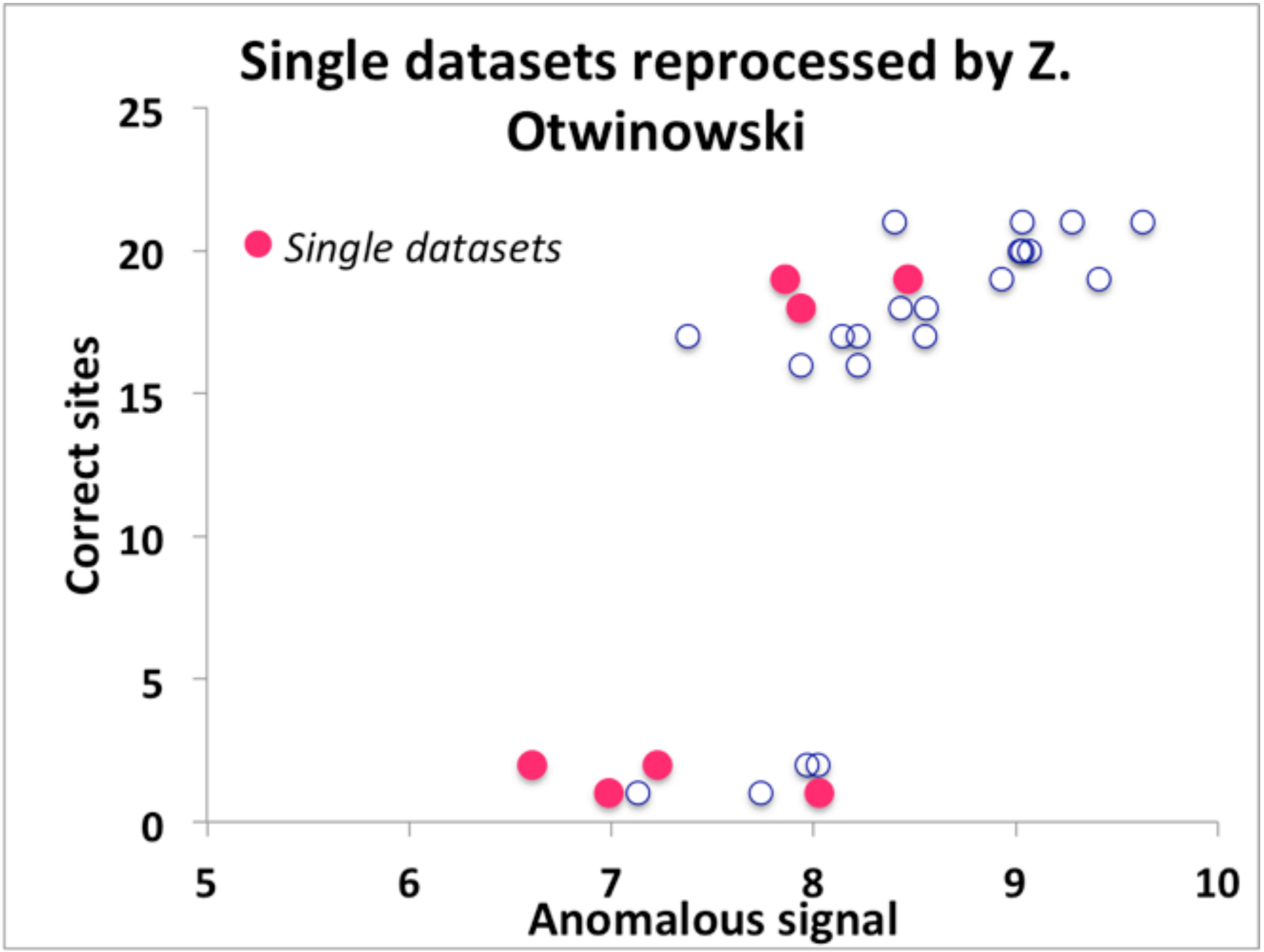
CysZ multi-crystal sulfur-SAD data

(Hyss LLG brute-force substructure determination)



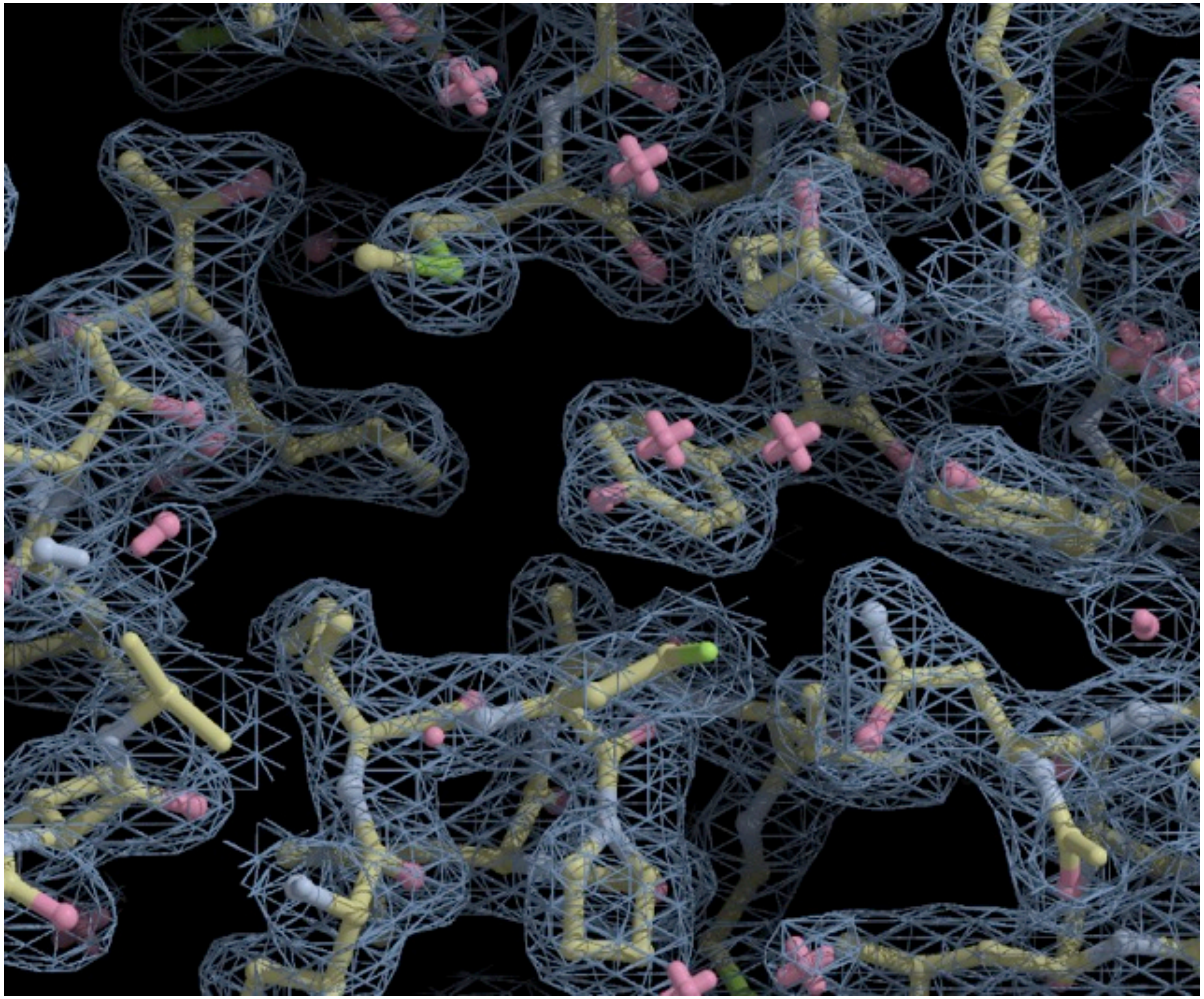
CysZ multi-crystal sulfur-SAD data

(Hyss LLG brute-force substructure determination)



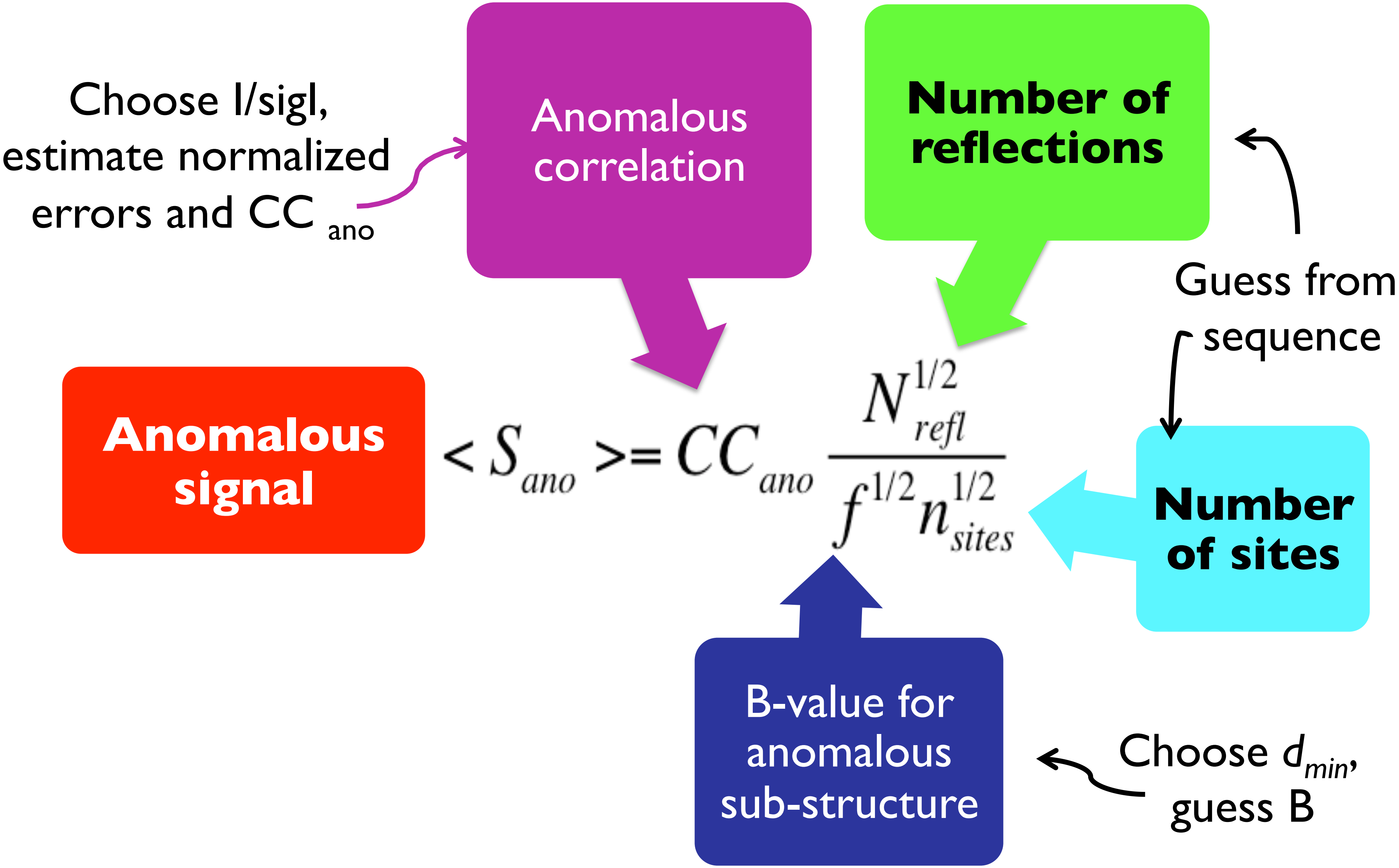
CysZ single-crystal sulfur-SAD data

Crystal 6 *AutoSol R/Rfree=0.24/0.27*



phenix.plan_sad_experiment

Design an experiment that will give you enough anomalous signal



Choose I/sigI, estimate normalized errors and CC $_{ano}$

Anomalous correlation

Number of reflections

Guess from sequence

Anomalous signal

$$< S_{ano} > = CC_{ano} \frac{N_{refl}^{1/2}}{f^{1/2} n_{sites}^{1/2}}$$

Number of sites

B-value for anomalous sub-structure

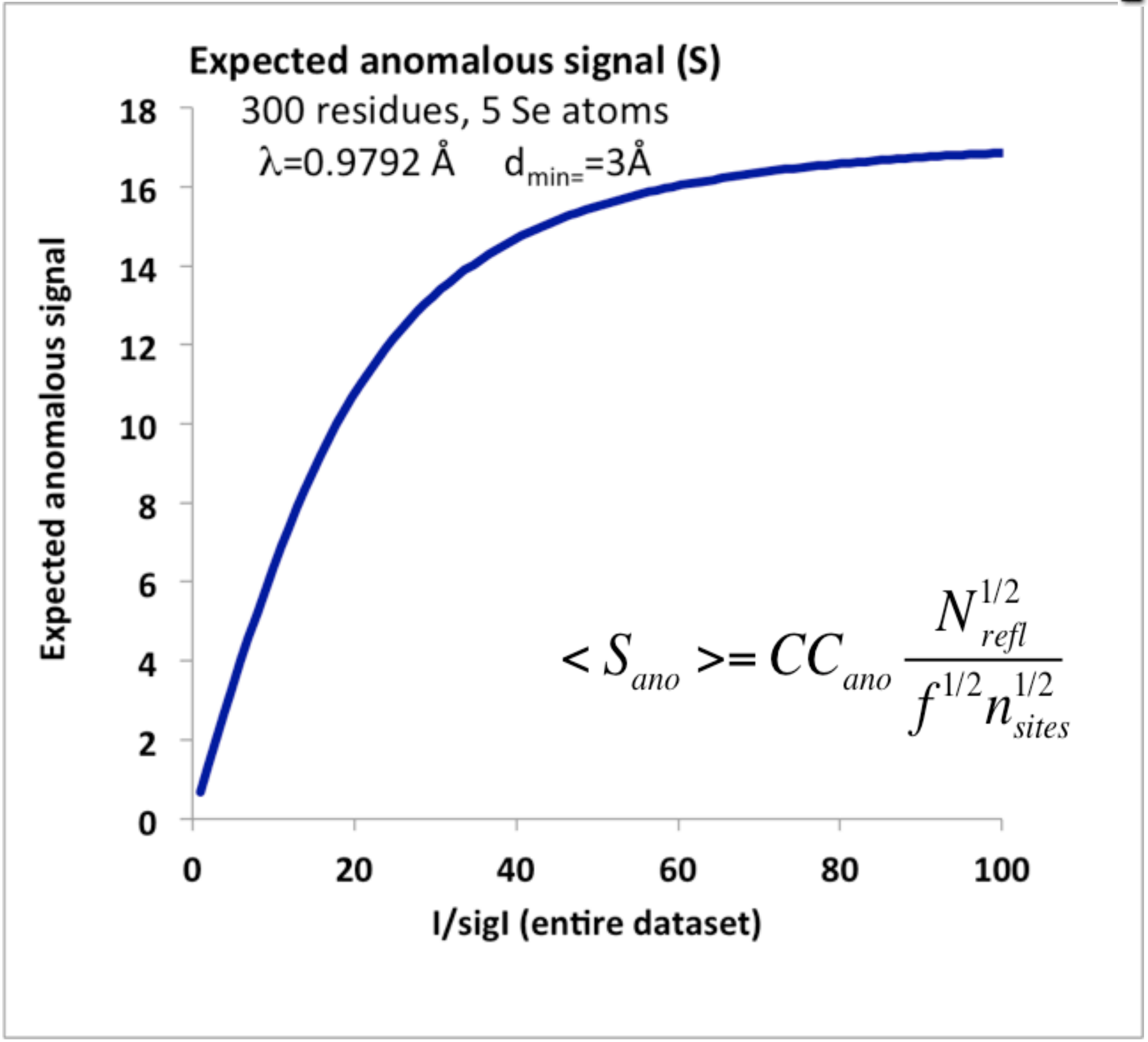
Choose d_{min}, guess B

Will I solve my structure?

Simulate experiment with *phenix.plan_sad_experiment* based on:

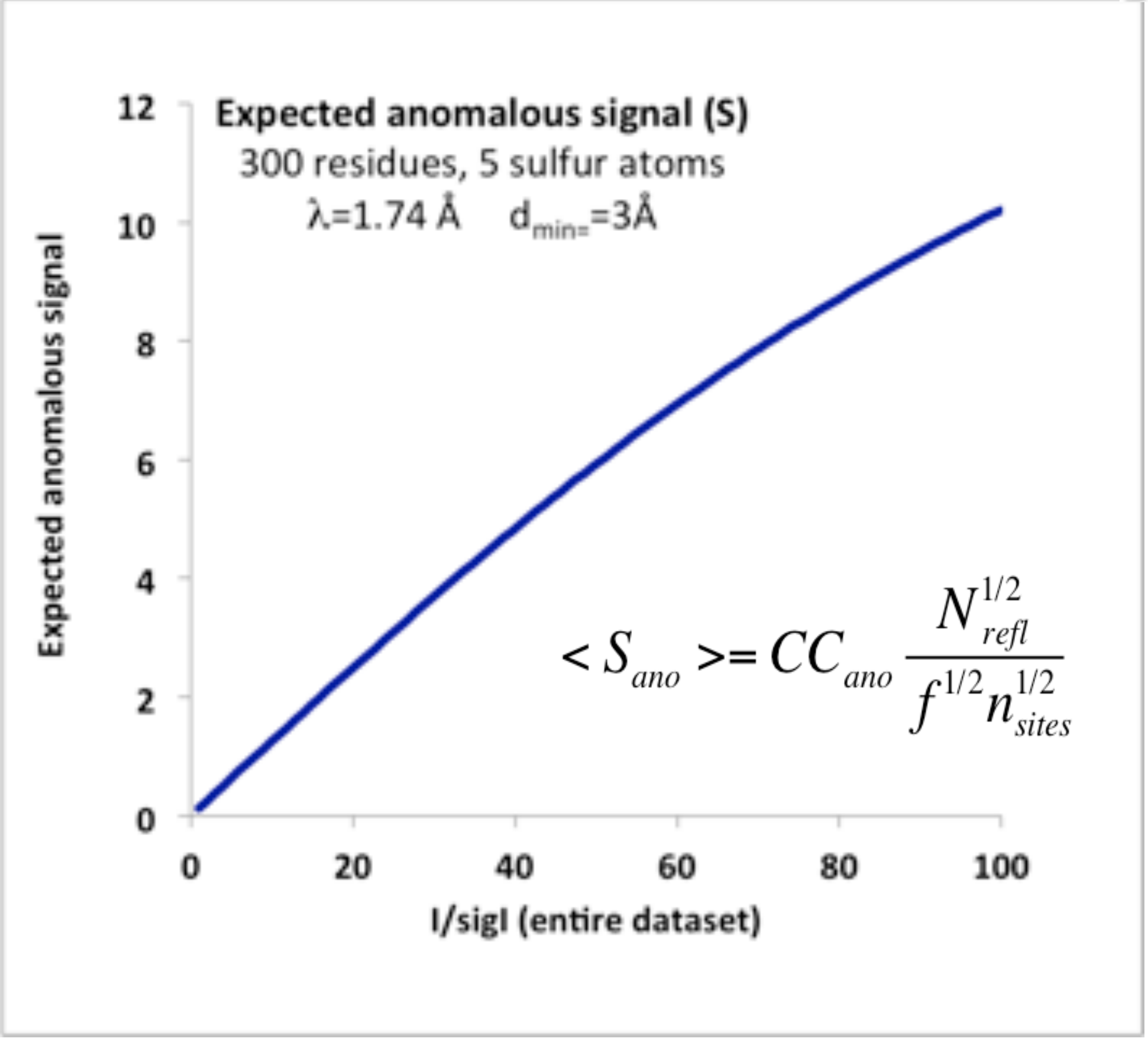
- I/sigma (errors in measurement)
- Anomalously-scattering atom, wavelength (f")
- Sequence (other atoms)
- Resolution of data
- Number of sites

Anomalous signal depends on I/sigI



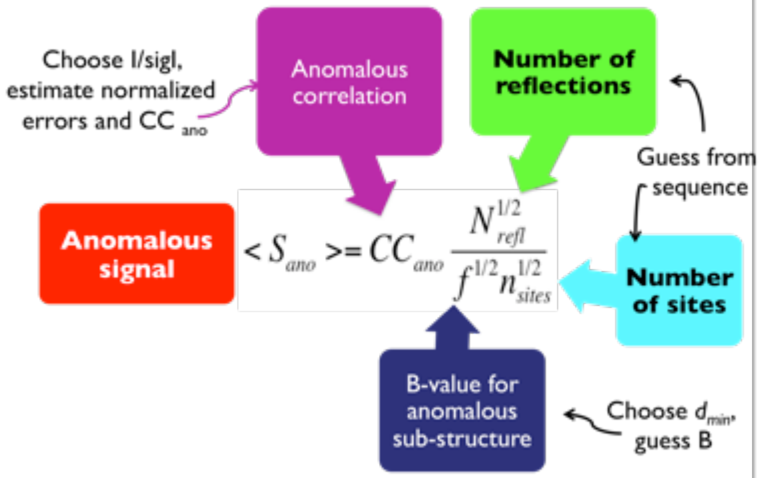
$$<S_{ano}> = CC_{ano}\frac{N_{refl}^{1/2}}{f^{1/2}n_{sites}^{1/2}}$$

Anomalous signal depends on f” (S vs Se)



$$<S_{ano}> = CC_{ano}\frac{N_{refl}^{1/2}}{f^{1/2}n_{sites}^{1/2}}$$

Estimating the anomalous signal before collecting the data



Choose I/sigI, estimate normalized errors and CC $_{ano}$

Anomalous correlation

Number of reflections

Guess from sequence

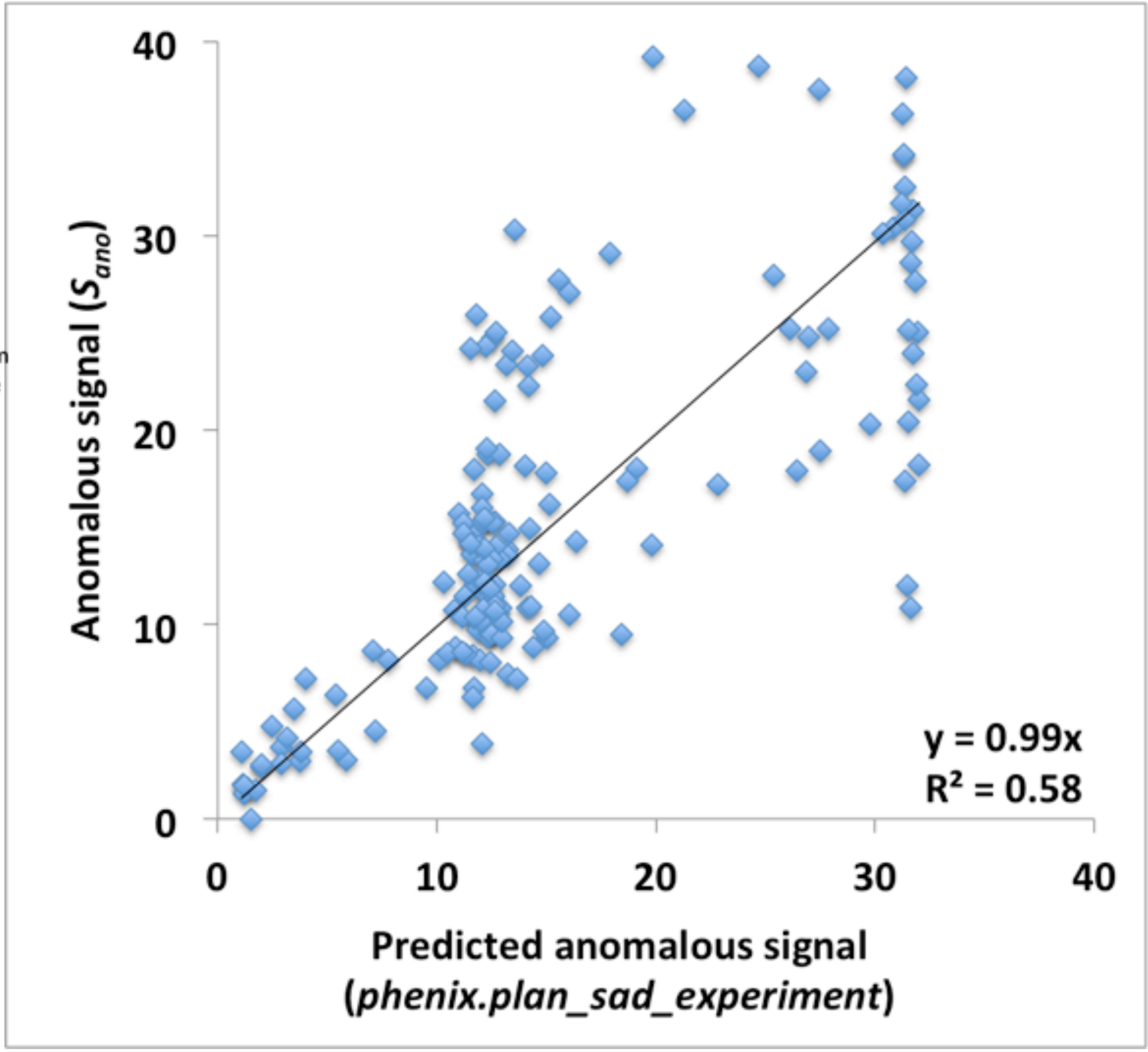
Anomalous signal

$$< S_{ano} > = CC_{ano} \frac{N_{refl}^{1/2}}{f^{1/2} n_{sites}^{1/2}}$$

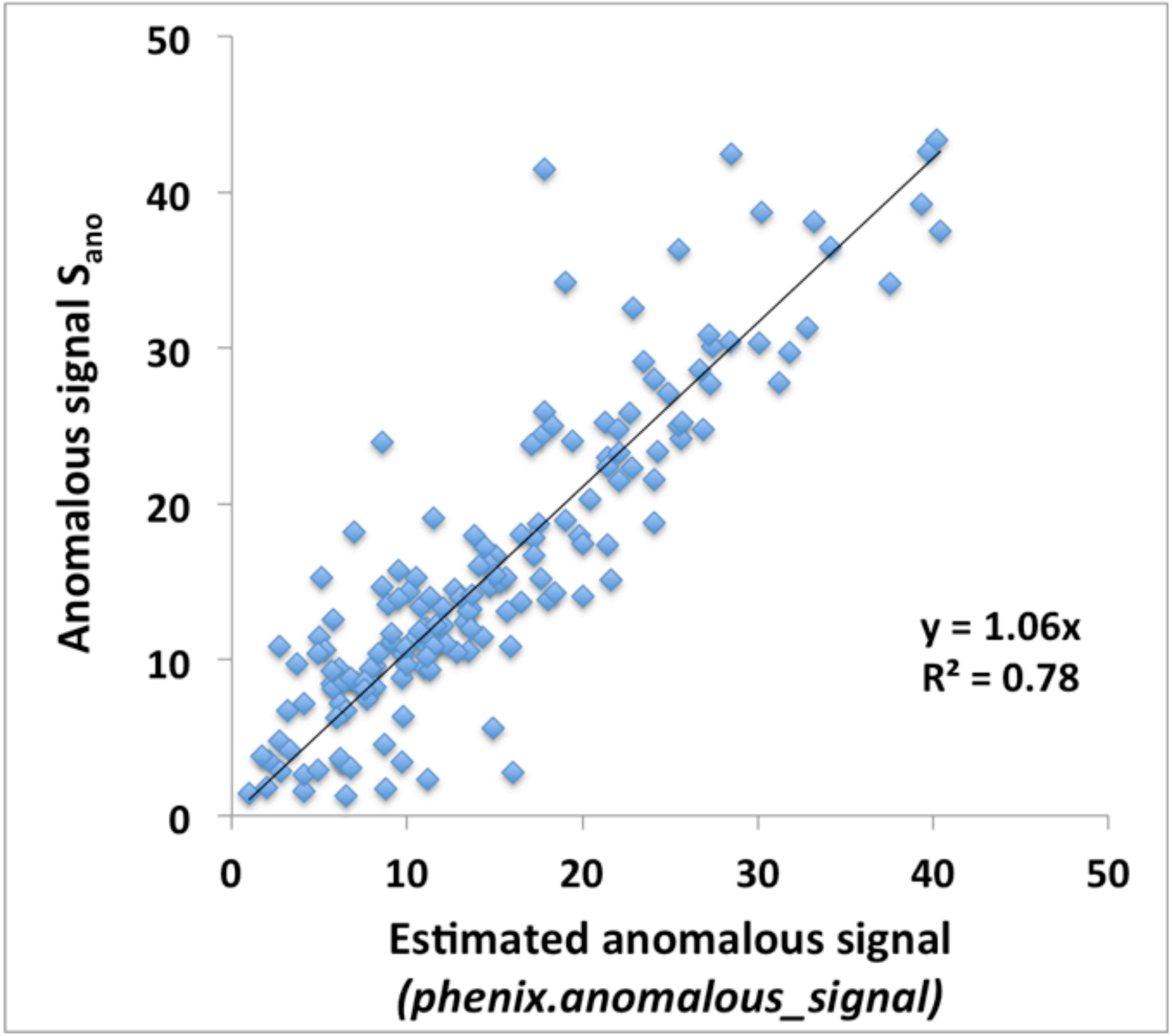
Number of sites

B-value for anomalous sub-structure

Choose d_{min}, guess B



Estimating the anomalous signal after collecting the data



Finding the anomalous sub-structure

Using the SAD likelihood function to find sites

"The likelihood of measuring the observed anomalous data

given

a potential sub-structure"

Using the SAD likelihood function to find the anomalous sub-structure

Start with guess about the anomalous sub-structure

From anomalous difference Patterson

Random

Any other source

Find additional sites that increase the likelihood

LLG completion based on log-likelihood gradient maps*

Iterative addition of sites

Related to using an anomalous difference Fourier—but better

*La Fortelle, E. de & Bricogne, G. (1997). Methods Enzymol. 276, 472-494
McCoy, A. J. & Read, R. J. (2010). Acta Cryst. D66, 458-469.

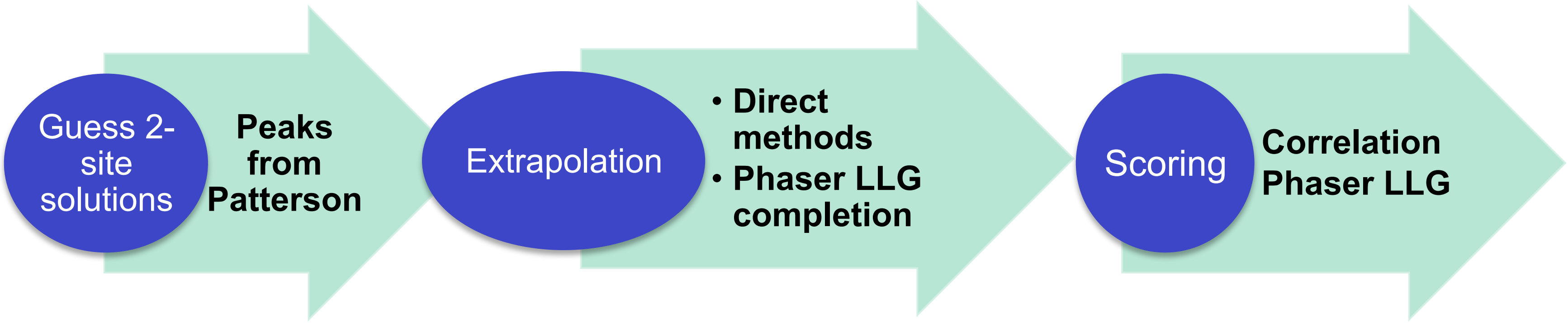
LLG sub-structure searches in HySS

Test cases

164 SAD datasets from PDB (largely JCSG MAD data)

Using peak, remotes, inflection as available to include data with low anomalous signal

Finding anomalous substructure with LLG completion



Guess 2-site solutions

Peaks from Patterson

Extrapolation

- Direct methods
- Phaser LLG completion

Scoring

Correlation
Phaser LLG

Use LLG score to compare solutions

- **Range of resolution**
  Variable number of Patterson solutions

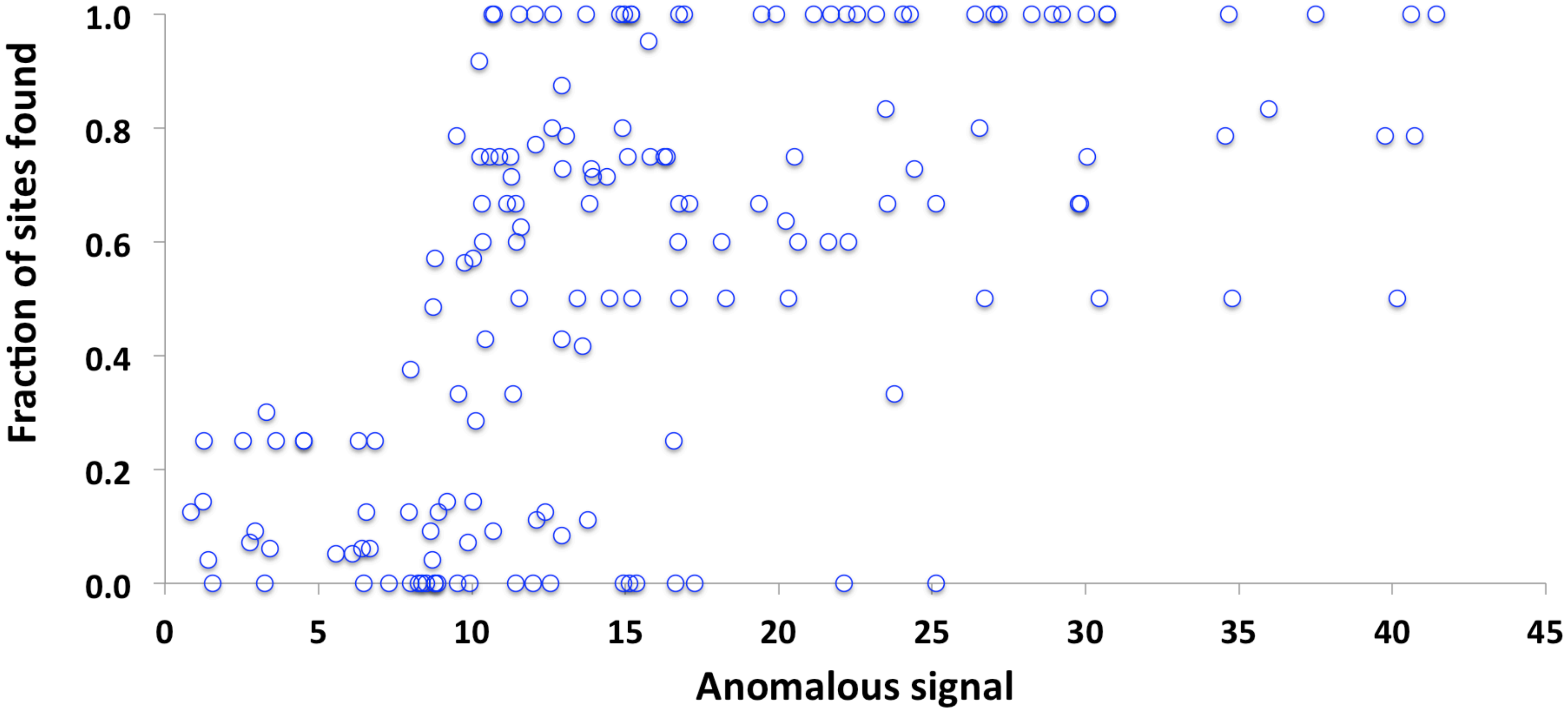
Adjustable LLGC_SIGMA

(cut-off for peak height)

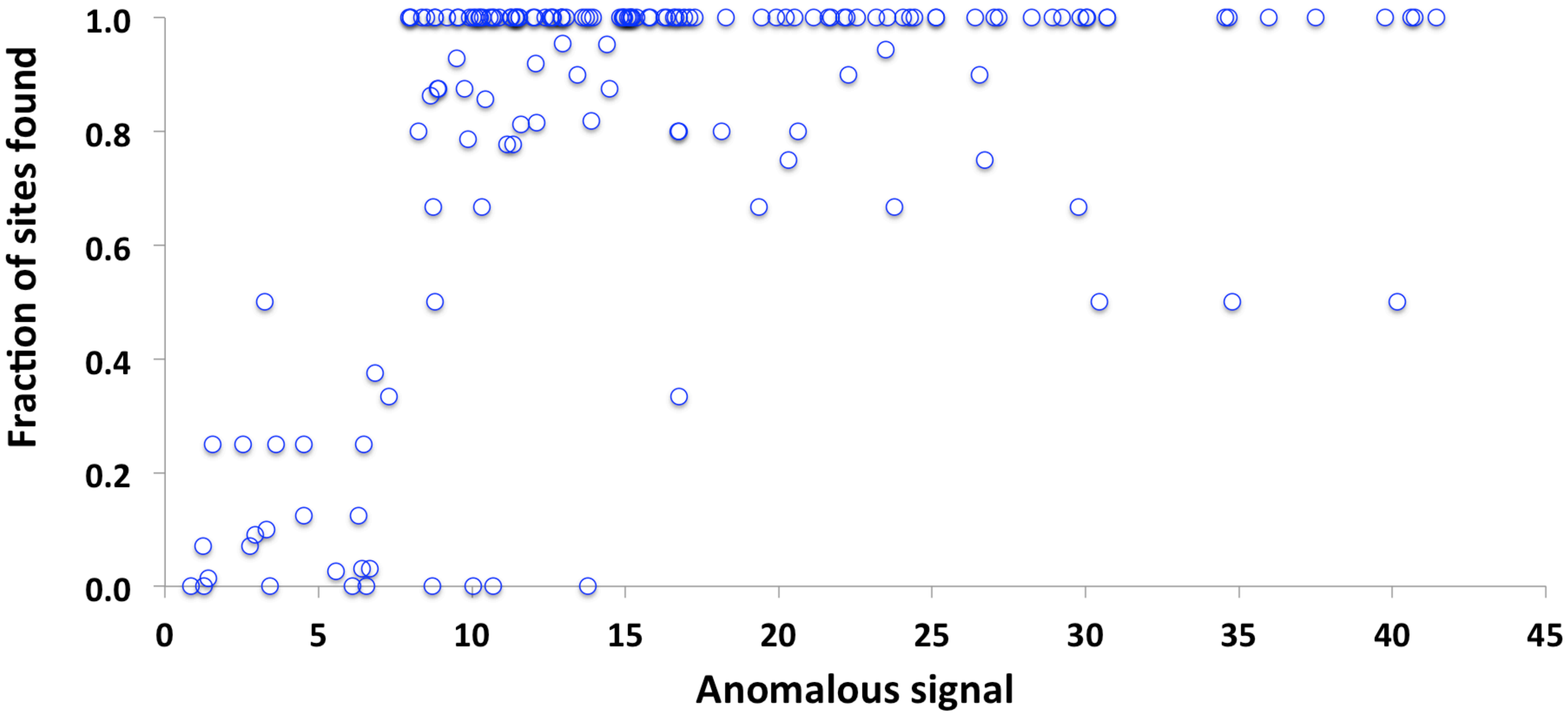
Terminate early if same solution found several times

Run quick direct methods first

Dual Space Sub-structure Completion

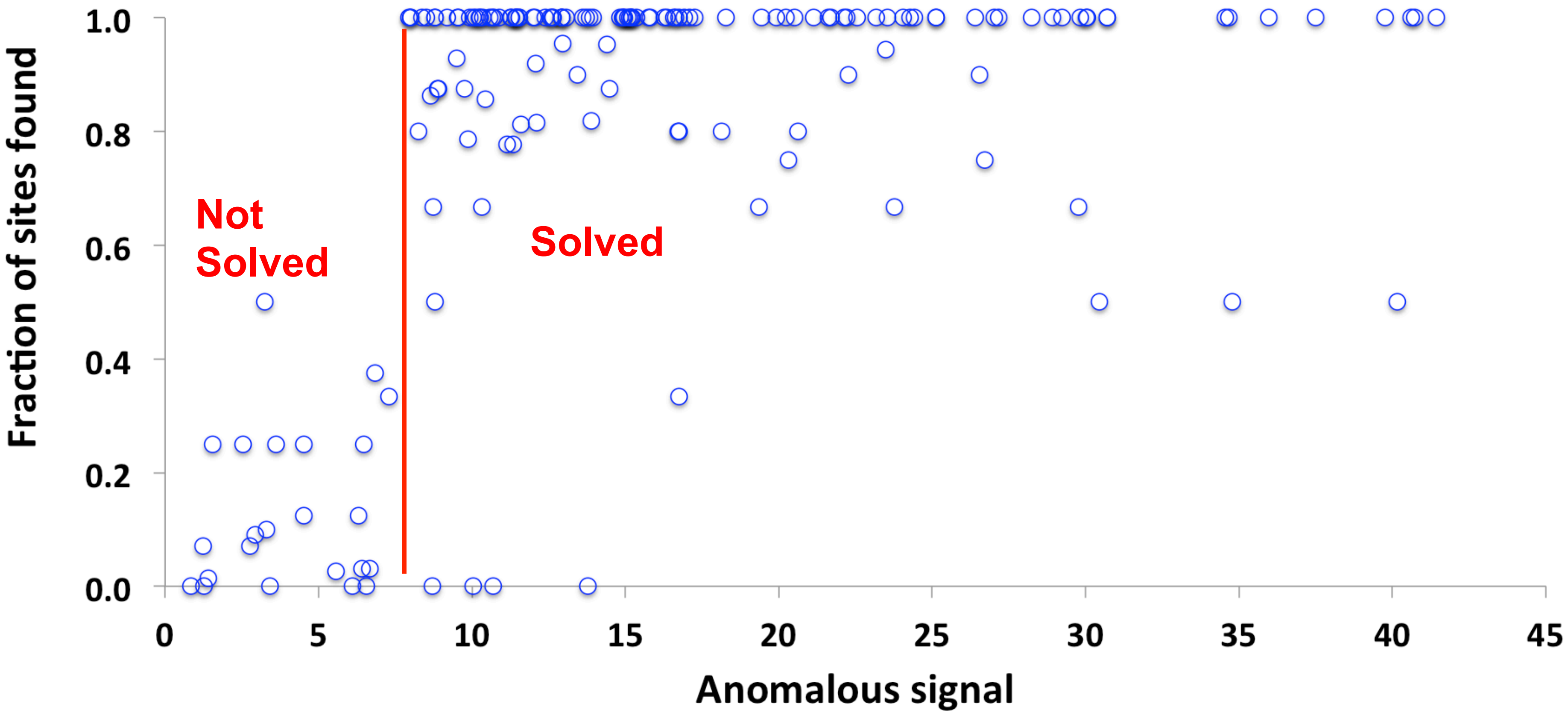


LLG Sub-structure Search



Bunkóczi et al., Nature Methods 12, 127–130 (2015).

Anomalous signal indicates if a dataset can be solved



Optimizing scaling and merging of SAD data

(phenix.scale_and_merge)

Why F+,F- differ from one crystal to another

Crystal 1

F+,F-

Errors in measurement (σ_{obs})

Crystals really are different ($\sigma_{crystal}$)

Crystal 2

F+,F-

Optimizing estimates of F+,F-

Crystal 1

F+,F-

Local scaling to reduce systematic errors

Use of $\sigma_{crystal}$ in weighting

Crystal 2

F+,F-

Applying inter-dataset variances in weighting

Crystal 1

Δ_{ano}

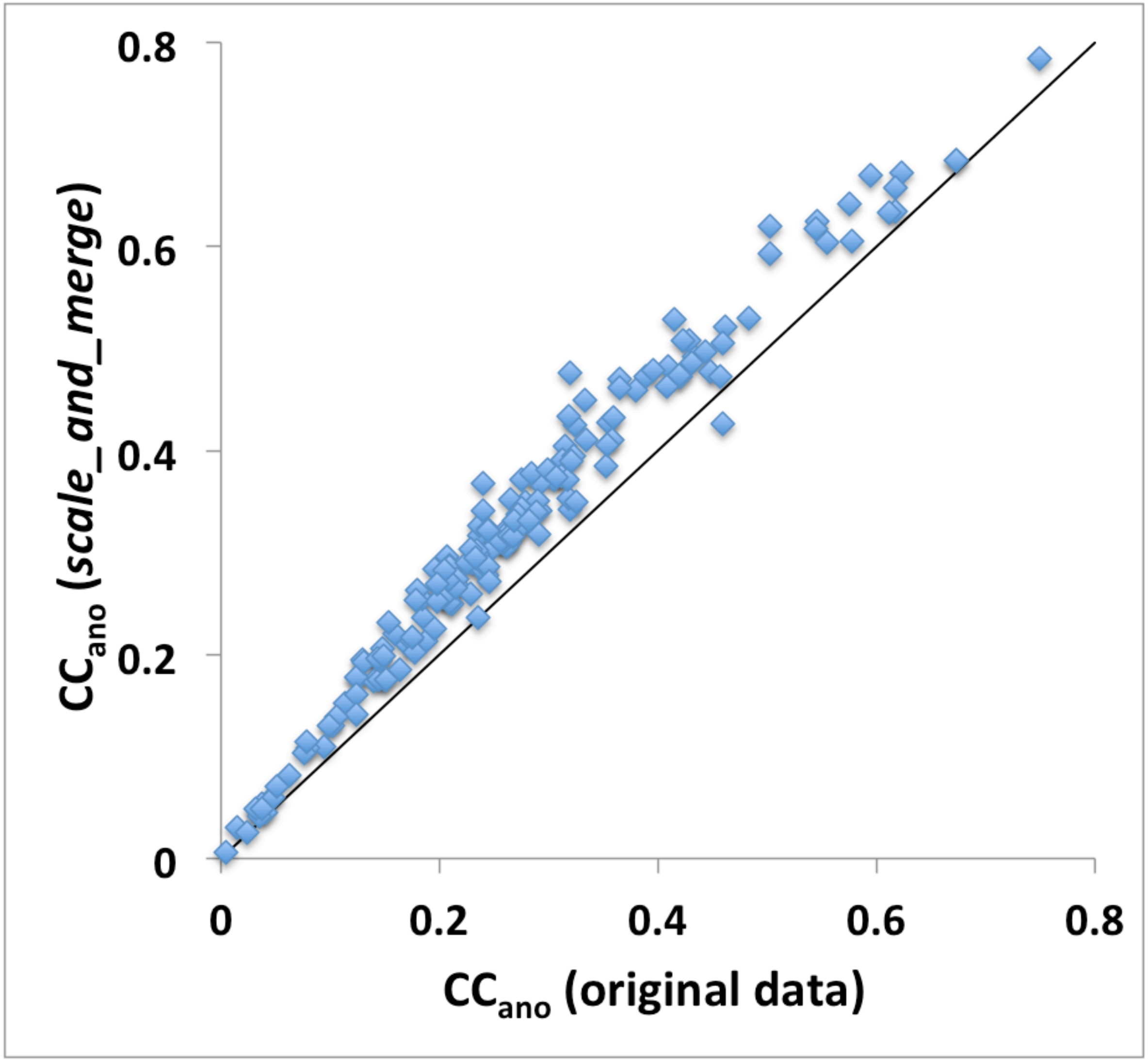
Weighting for data from an individual crystal:

$$\sigma^2_{total} \approx \sigma^2_{obs} + \sigma^2_{crystal}$$

Average of all crystals

Δ^{AVG}

Improvement in anomalous correlation using local scaling in *phenix.scale_and_merge*



SBGrid SAD/MAD datasets

reprocessed by Kay Diederichs

Set	PDB	Atom	Expt	Resolution (Å)	Sites	Molecule	Reference
#3	3TRZ	ZN	SAD	2.8	12	Lin28/let-7d microRNA complex	(Nam et al., 2011)
#97	1XBN	FE	MAD	2.6	1	bacterial nitric oxide sensor	(Nioche et al., 2004)
#111	4TSO	BA	SAD	2.6	1	Fluorescent RNA aptamer	(Warner et al., 2015)
#123	3M1C	SE	SAD	2.7	9	herpesvirus fusion regulator complex gH-gL	(Chowdary et al., 2010)

3TRZ (12 ZN SAD)

Scaling of data with phenix.scale_and_merge

Resolution	CC1/2_ano
6.0	0.64
5.5	0.44
5.0	0.25
4.5	0.17
4.0	0.09
3.5	-0.01
3.0	-0.01
2.8	-0.02

phenix.scale_and_merge *XDS_ASCII.HKL*

3TRZ (12 ZN SAD)

Analysis of anomalous data with phenix.anomalous_signal

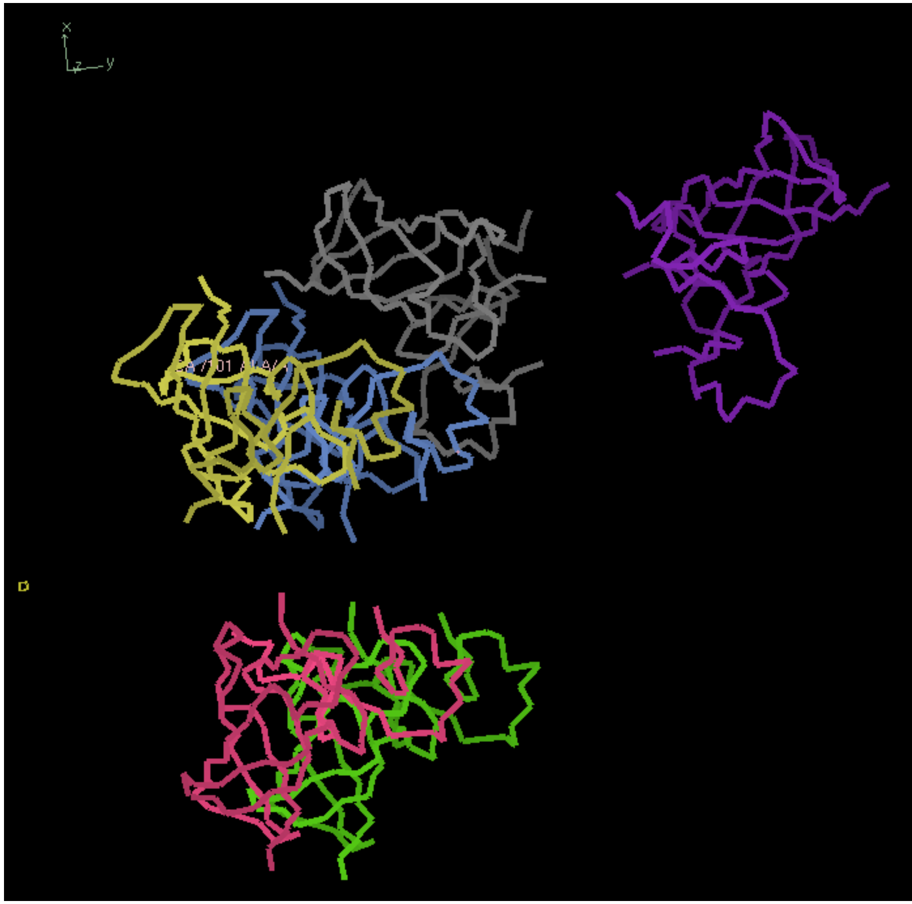
Resolution	CC1/2_ano	CCano*	Anomalous Signal	P(substr)	FOM*
6.0	0.63	0.66	12	76	0.3
5.5	0.59	0.66	13	81	0.3
5.0	0.51	0.63	15	91	0.3
4.5	0.42	0.61	16	97	0.3
4.0	0.31	0.55	17	99	0.3
3.5	0.15	0.39	14	86	0.3
3.0	0.06	0.26	11	73	0.2
2.8	0.05	0.23	10	65	0.2

```
phenix.anomalous_signal    data=scaled_data.mtz scaled_data.mtz
half_dataset_a=half_dataset_a.mtz    half_dataset_b=half_dataset_b.mtz
seq_file=seq.dat    atom_type=ZN      sites=12
```

3TRZ (12 ZN SAD)

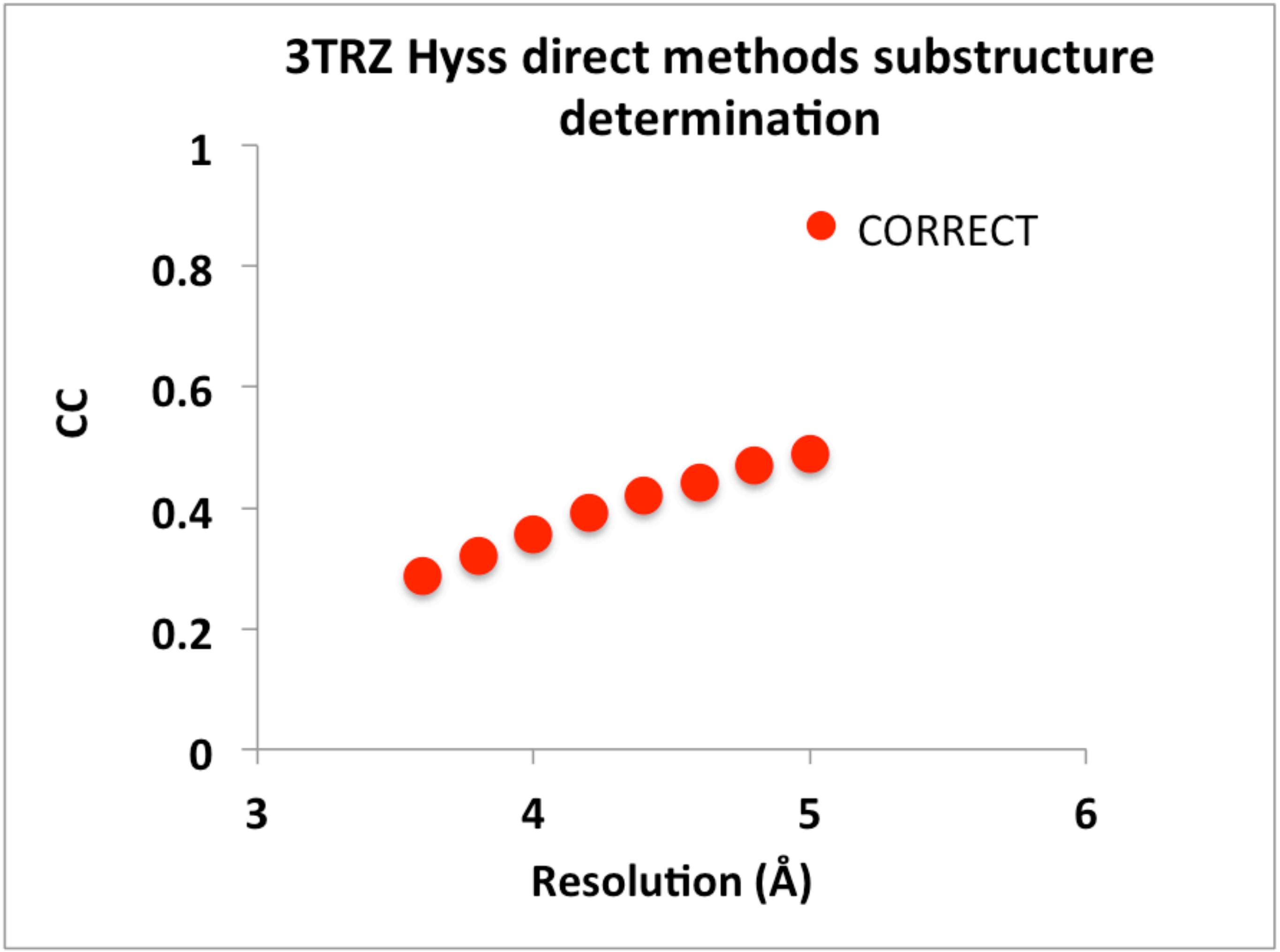
Structure contains translational non-crystallographic symmetry

phenix.xtriage scaled_data.mtz

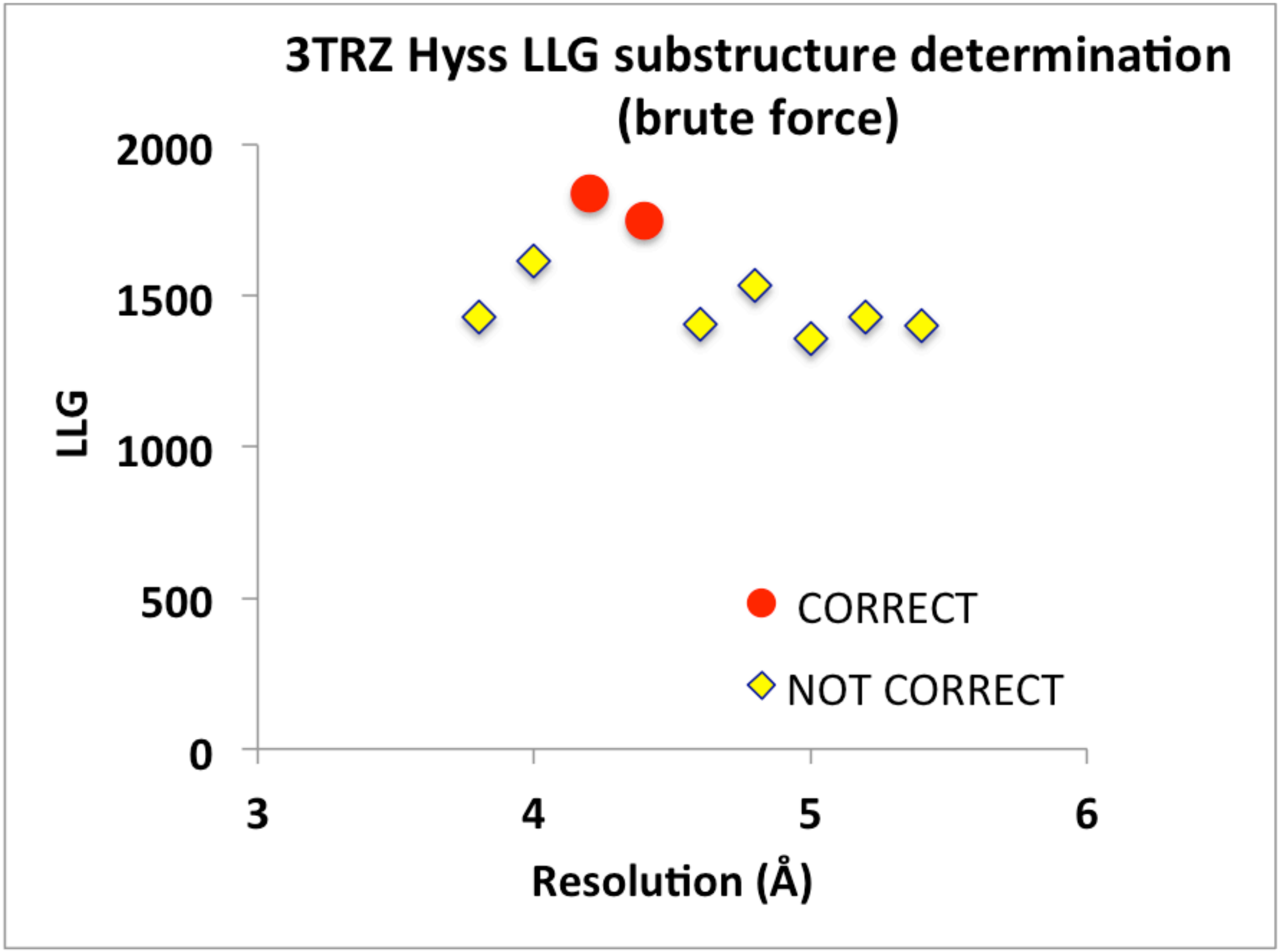


Peak in native Patterson function:
(0,0,1/3)   36% of origin
→ Strong translational NCS present

(Note: Phaser SAD LLG scoring does not yet account for tNCS)



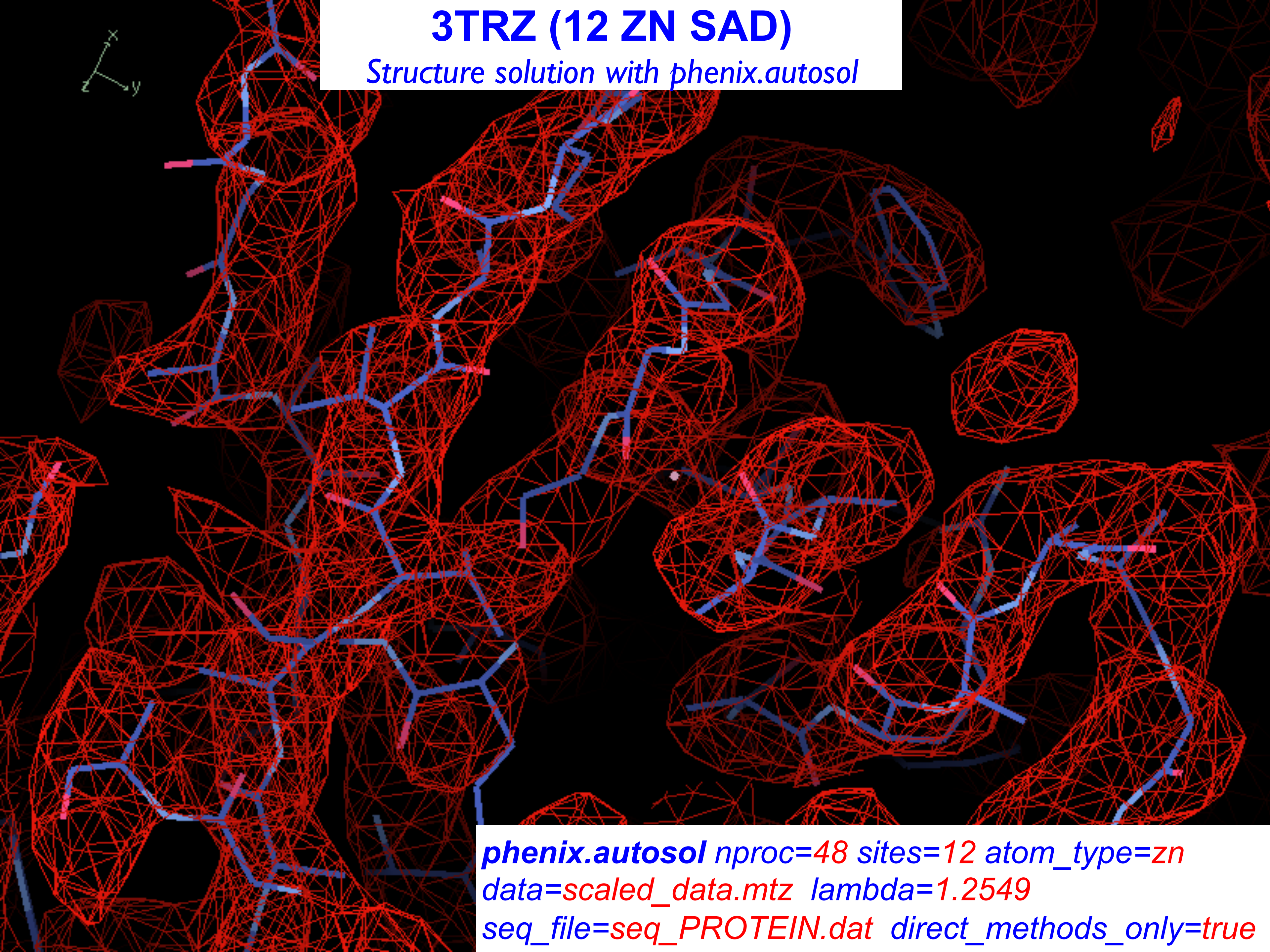
```
phenix.hyss  nproc=48  scaled_data.mtz  12 ZN   wavelength=1.2549  resolution=4.2
```



```
phenix.hyss nproc=48 scaled_data.mtz  12 ZN  wavelength=1.2549 \
  resolution=4.2    rescore=phaser-complete    strategy=brute_force
```

3TRZ (12 ZN SAD)

Structure solution with phenix.autosol



phenix.autosol *nproc=48 sites=12 atom_type=zn*
data=scaled_data.mtz lambda=1.2549
seq_file=seq_PROTEIN.dat direct_methods_only=true

3TRZ (12 ZN SAD)

Protein model building with phenix.autobuild

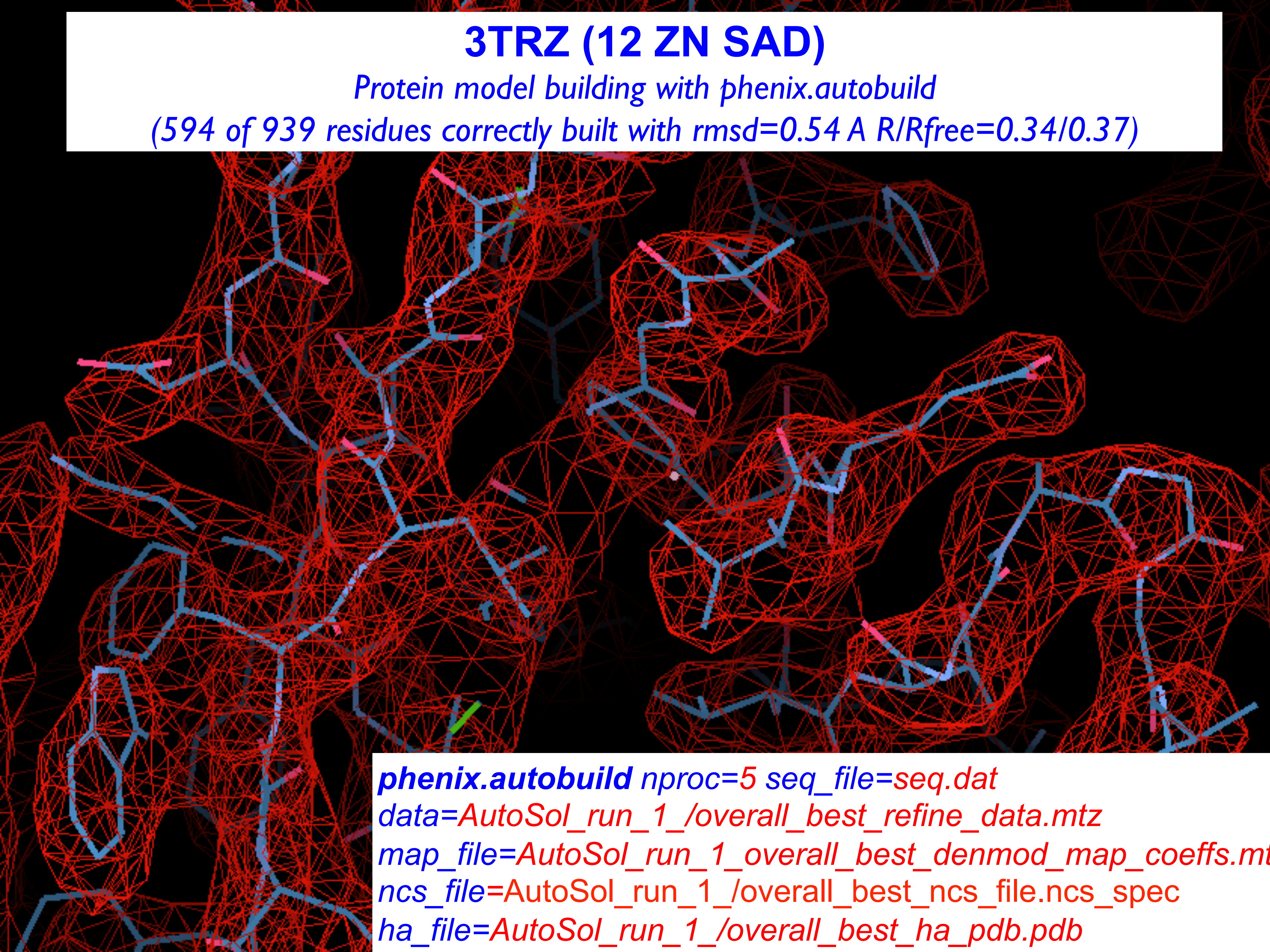
(594 of 939 residues correctly built with rmsd=0.54 A R/Rfree=0.34/0.37)

```
phenix.autobuild nproc=5 seq_file=seq.dat
data=AutoSol_run_1_/overall_best_refine_data.mtz
map_file=AutoSol_run_1_overall_best_denmod_map_coeffs.mt
ncs_file=AutoSol_run_1_/overall_best_ncs_file.ncs_spec
ha_file=AutoSol_run_1_/overall_best_ha_pdb.pdb
```

3TRZ (12 ZN SAD)

Protein model building with phenix.autobuild

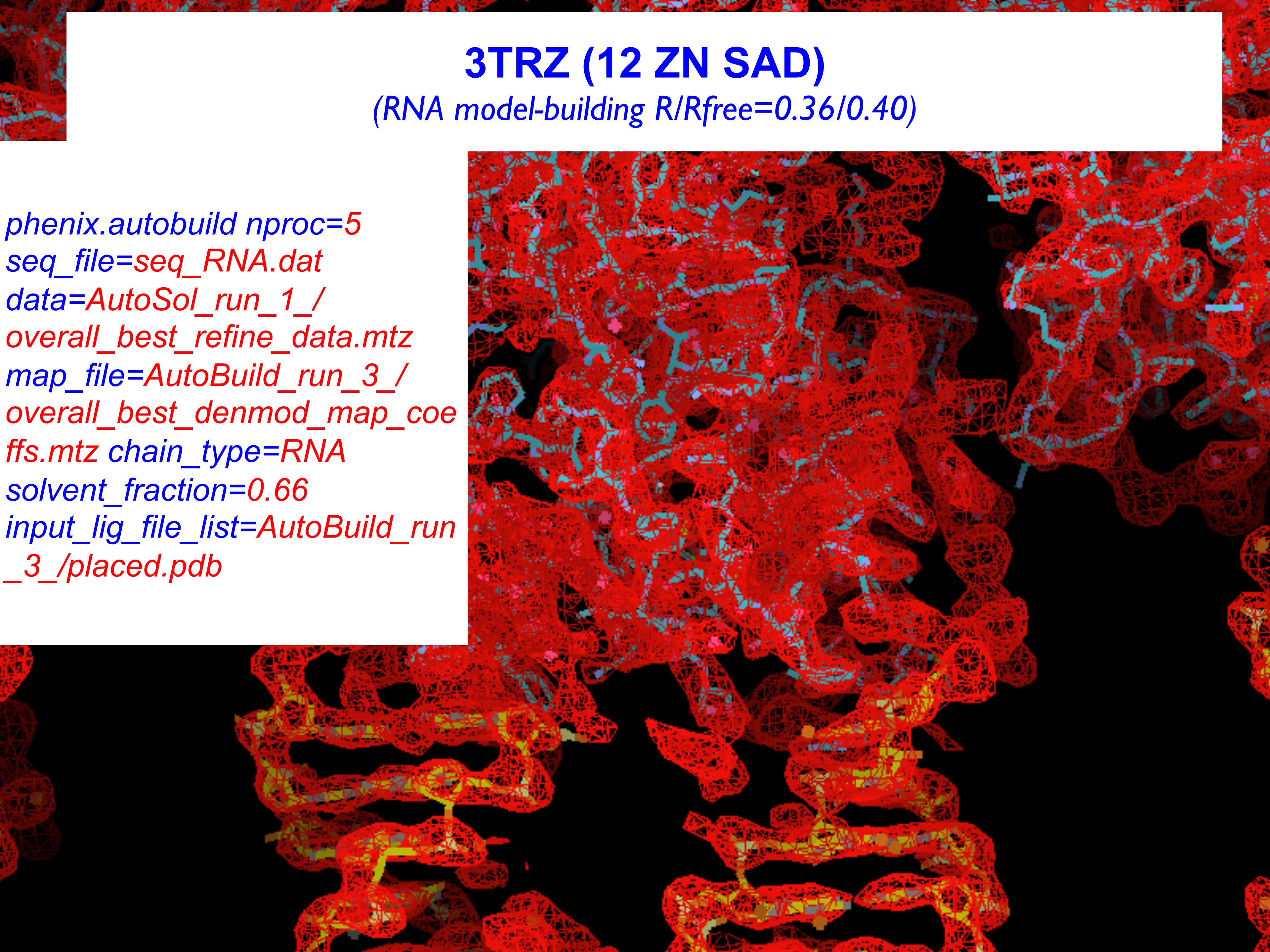
(594 of 939 residues correctly built with rmsd=0.54 A R/Rfree=0.34/0.37)



```
phenix.autobuild nproc=5 seq_file=seq.dat
data=AutoSol_run_1_/overall_best_refine_data.mtz
map_file=AutoSol_run_1_overall_best_denmod_map_coeffs.mt
ncs_file=AutoSol_run_1_/overall_best_ncs_file.ncs_spec
ha_file=AutoSol_run_1_/overall_best_ha_pdb.pdb
```

3TRZ (12 ZN SAD)

(RNA model-building R/Rfree=0.36/0.40)



```
phenix.autobuild nproc=5 seq_file=seq_RNA.dat data=AutoSol_run_1_/overall_best_refine_data.mtz map_file=AutoBuild_run_3_/overall_best_denmod_map_coeffs.mtz chain_type=RNA solvent_fraction=0.66 input_lig_file_list=AutoBuild_run_3_/placed.pdb
```

1XBN (Fe MAD)

Analysis of anomalous data with phenix.autosol

```
phenix.autosol mad.eff
```

Parameters file:

```
autosol {
  seq_file = seq.dat
  sites = 1
  atom_type = Fe
  wavelength {
    data = e1.HKL
    lambda = 1.738729
  }
  wavelength {
    data = e2.HKL
    lambda = 1.624747
  }
  wavelength {
    data = e3.HKL
    lambda = 1.740630
  }
}
```

1XBN (Fe MAD)

Analysis of anomalous data with SOLVE in phenix.autosol (AutoSol_run_1_/solve_2.prt)

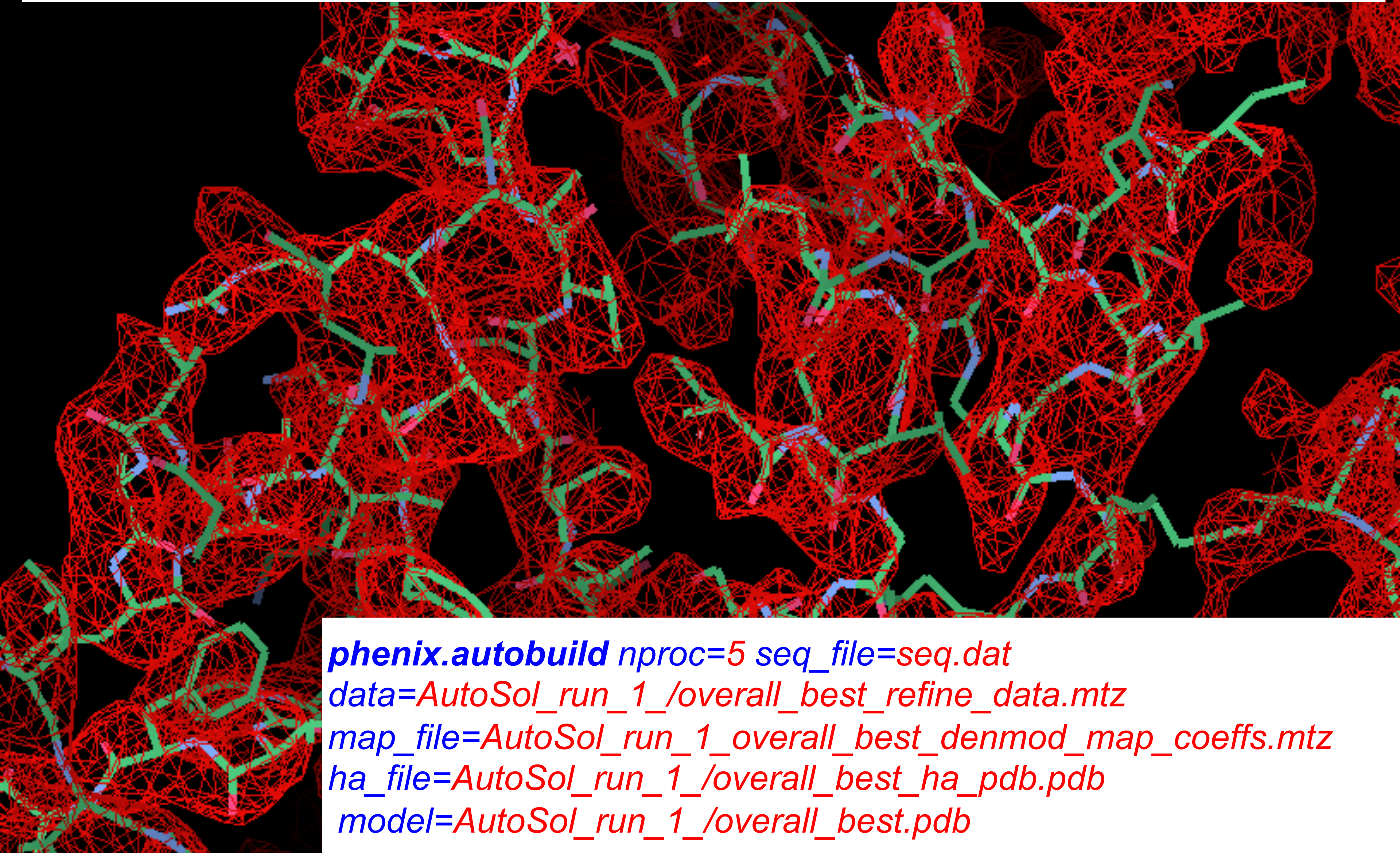
Correlation of anomalous differences between wavelengths

Resolution (Å)	CC 1 vs 2	CC 1 vs 3	CC 2 vs 3
5.3	0.84	0.81	0.75
4.0	0.56	0.56	0.38
3.7	0.26	0.37	0.2
3.5	0.21	0.33	0.03
3.3	0.07	0.26	-0.01
3.2	0.06	0.35	-0.02
3.0	-0.04	0.23	-0.03
2.9	-0.01	0.3	0.12
2.8	0.13	0.16	-0.02
2.6	-0.07	0.24	-0.07

1XBN (Fe MAD)

Analysis of anomalous data with phenix.autobuild

R/Rfree=0.35/0.41

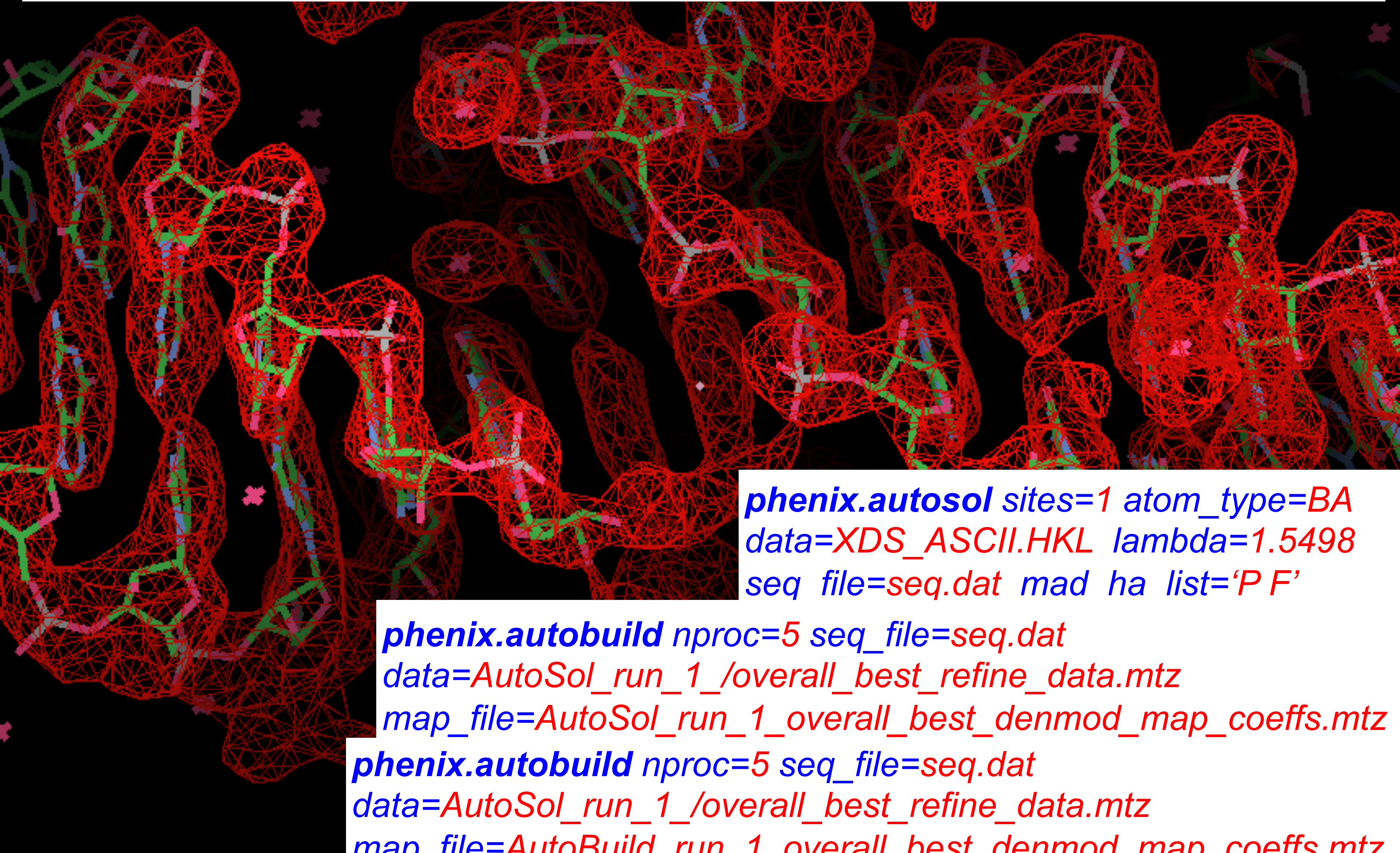


```
phenix.autobuild nproc=5 seq_file=seq.dat
data=AutoSol_run_1_/overall_best_refine_data.mtz
map_file=AutoSol_run_1_overall_best_denmod_map_coeffs.mtz
ha_file=AutoSol_run_1_/overall_best_ha_pdb.pdb
 model=AutoSol_run_1_/overall_best.pdb
```

4TS0 (Ba SAD)

Analysis of anomalous data with phenix.autosol; phenix.autobuild

47 of 87 residues built with rmsd 0.54 A.   R/Rfree=0.43/0.48



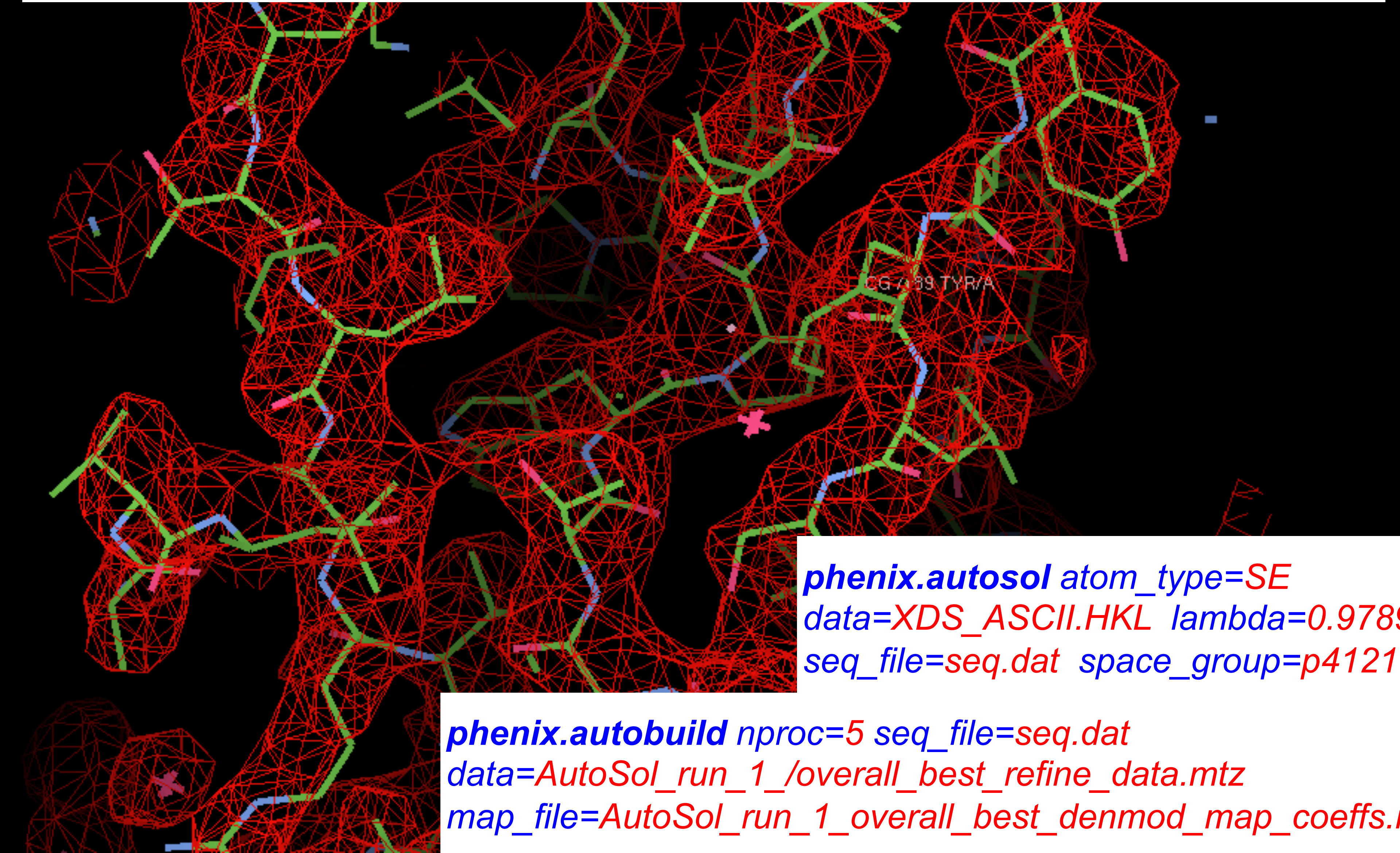
```
phenix.autosol sites=1 atom_type=BA
data=XDS_ASCII.HKL  lambda=1.5498
seq  file=seq.dat  mad  ha  list='P F'
```

```
phenix.autobuild nproc=5 seq_file=seq.dat
data=AutoSol_run_1_/overall_best_refine_data.mtz
map_file=AutoSol_run_1_overall_best_denmod_map_coeffs.mtz
```

```
phenix.autobuild nproc=5 seq_file=seq.dat
data=AutoSol_run_1_/overall_best_refine_data.mtz
map_file=AutoBuild_run_1_overall_best_denmod_map_coeffs.mtz
```

3M1C (SE-SAD)

Analysis of anomalous data with phenix.autosol; phenix.autobuild
698 of 865 residues built with rmsd 0.51 A.   R/Rfree=0.29/0.34



```
phenix.autosol atom_type=SE
data=XDS_ASCII.HKL  lambda=0.978
seq_file=seq.dat  space_group=p4121
```

```
phenix.autobuild nproc=5 seq_file=seq.dat
data=AutoSol_run_1_/overall_best_refine_data.mtz
map_file=AutoSol_run_1_overall_best_denmod_map_coeffs.
```

The Phenix Team

Lawrence Berkeley Laboratory

Paul Adams, Pavel Afonine, Nigel Moriarty, Nicholas Sauter, Oleg Sobolev, Billy Poon

Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung

Duke University

Jane & David Richardson, Christopher Williams, Bradley Hintze

Randy Read, Airlie McCoy, Gabor Bunkóczi, Robert Oeffner

Cambridge University

An NIH/NIGMS funded Program Project