Phenix

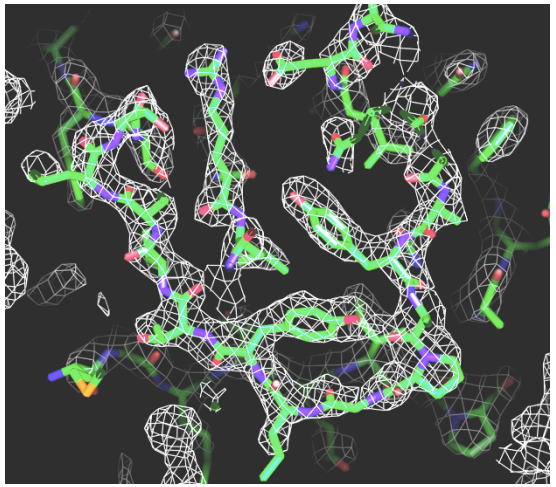
Planning and carrying out automated structure determination using SAD phasing

Methods and techniques in integrated structural biology:
Beyond black boxes

Strasbourg, October 5, 2016

Tom Terwilliger
Los Alamos National Laboratory

Steps in Single Wavelength Anomalous Diffraction (SAD) Structure Determination

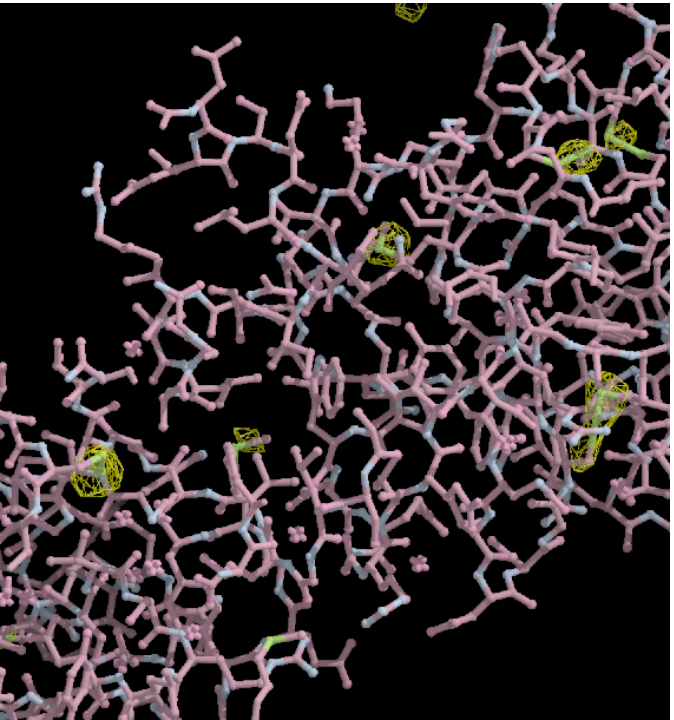


- Plan the experiment
- Measure the data
- Scale the data
- Evaluate the accuracy of the anomalous differences
- Find the anomalous sub-structure
- Identify hand of sub-structure
- Calculate experimental phases and a map
- Improve the map with density modification
- Build and refine a model

Phenix

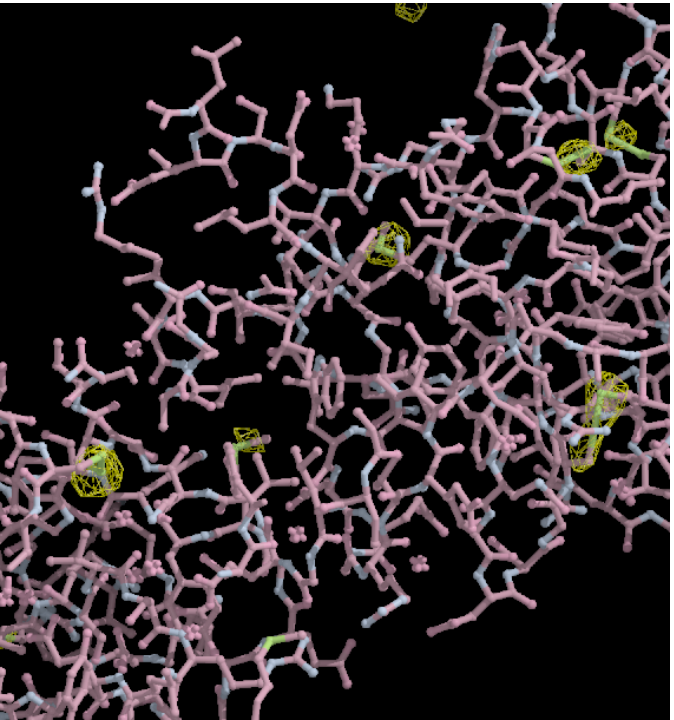
Planning a SAD experiment

Will I find the sites of anomalously-scattering atoms?



Phenix

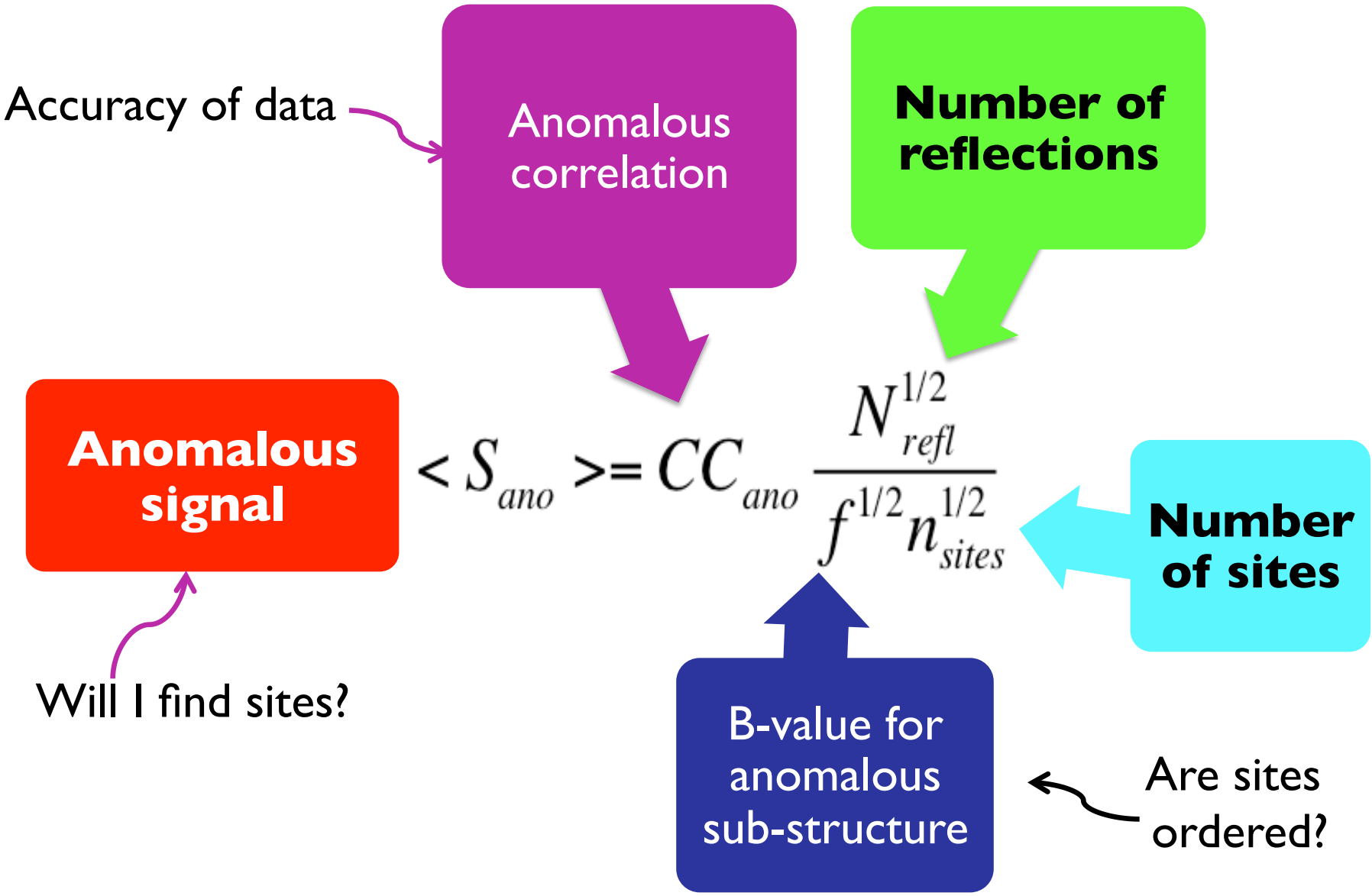
Planning a SAD experiment



How many sites?
How many reflections?

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

What determines if I will find the sites?



Phenix

Maximizing the anomalous signal and the anomalous correlation

*The **anomalous correlation** CC_{ano} is a measure of the accuracy of each anomalous difference*

*The **anomalous signal** is a measure of how much total information per site is present in the anomalous differences*

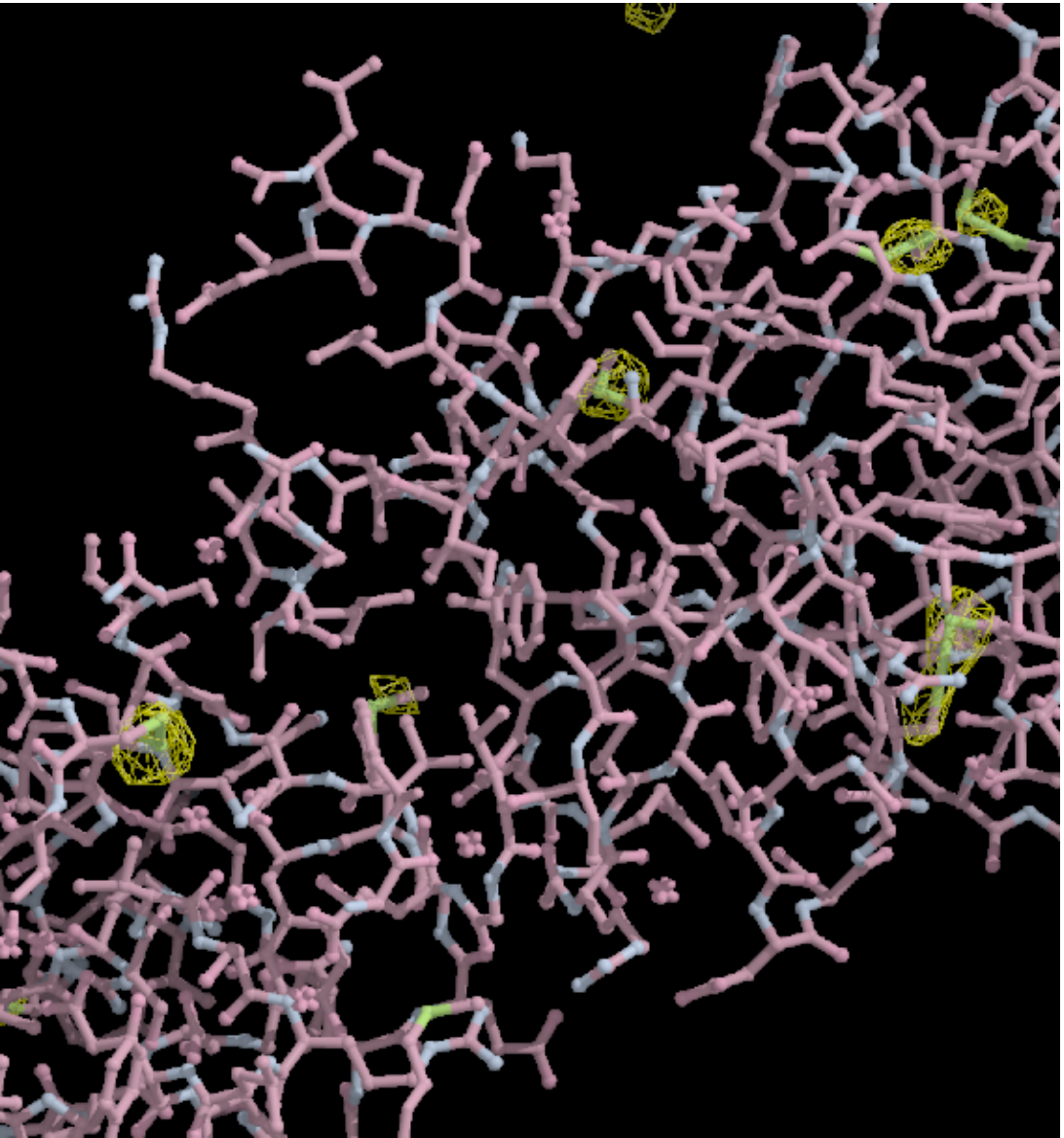
Anomalous correlation: accuracy of anomalous differences

Correlation of observed and sub-structure anomalous differences

$$CC_{ano} \equiv \frac{<\Delta_{ano,j}\Delta^{obs}_{ano,j}>}{<\Delta_{ano}>^{1/2}<\Delta^{2,obs}_{ano}>^{1/2}}$$

CC_{ano} indicates how much of each anomalous difference is useful (on average)

Anomalous signal: peak height at coordinates of anomalously-scattering atoms



$$S_{ano} = \frac{< \rho_{ano}(x_j) >}{< \rho_{ano}^2 >^{1/2}}$$

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

Anomalous difference Fourier with observed data and model phases

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)

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

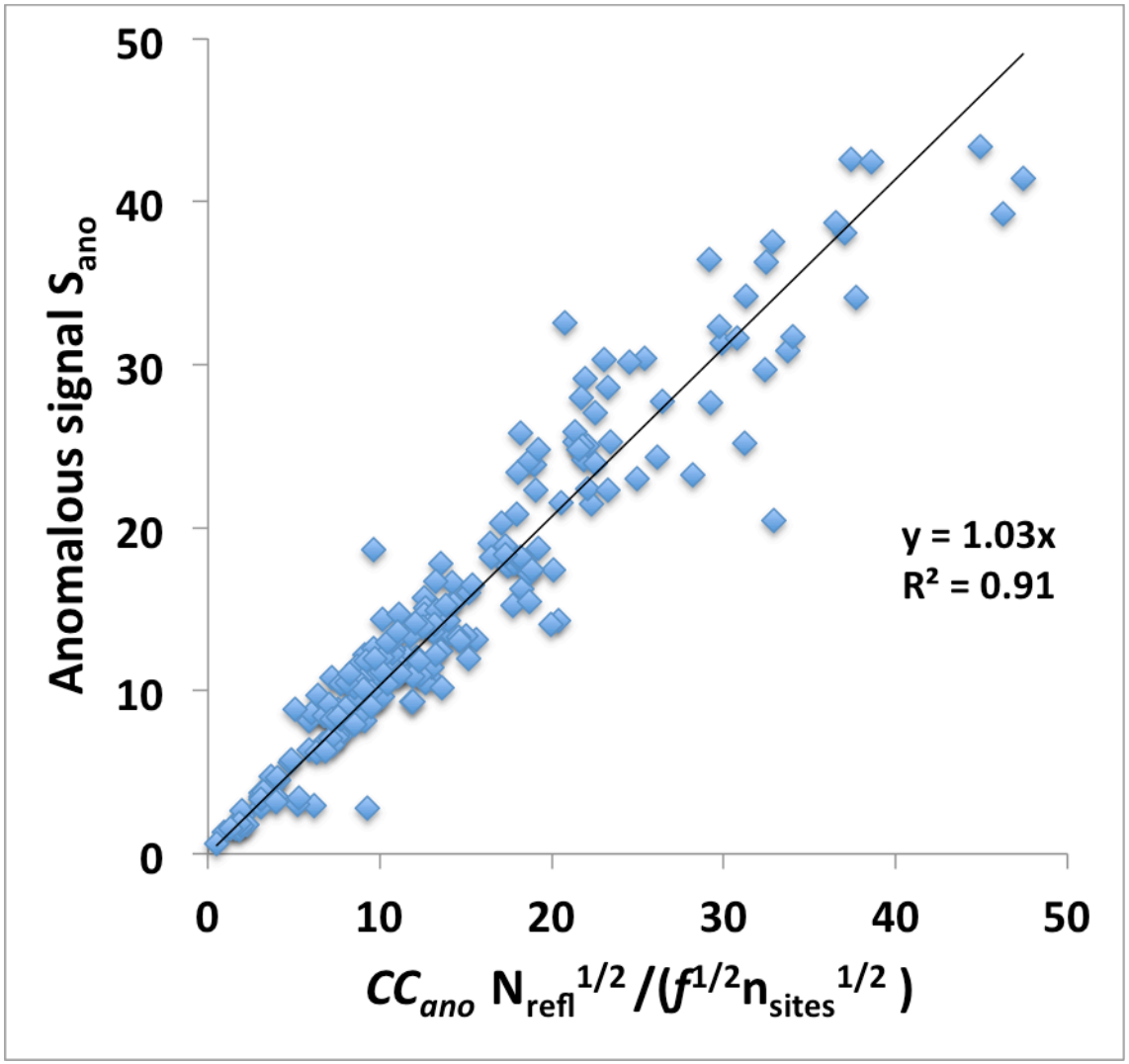
(f^h in this equation is the anomalous scattering factor)

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

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

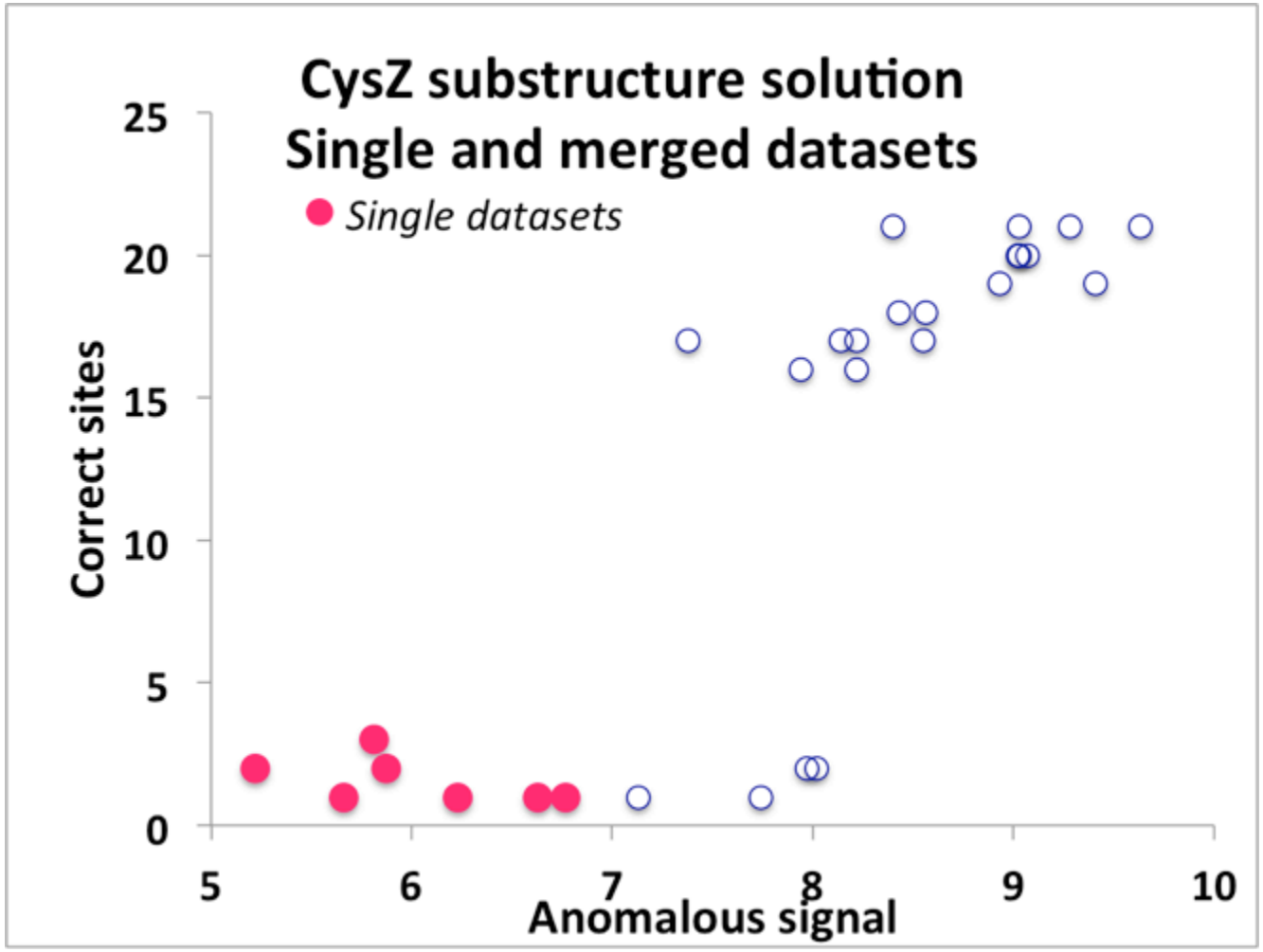
CysZ multi-crystal sulfur-SAD data

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

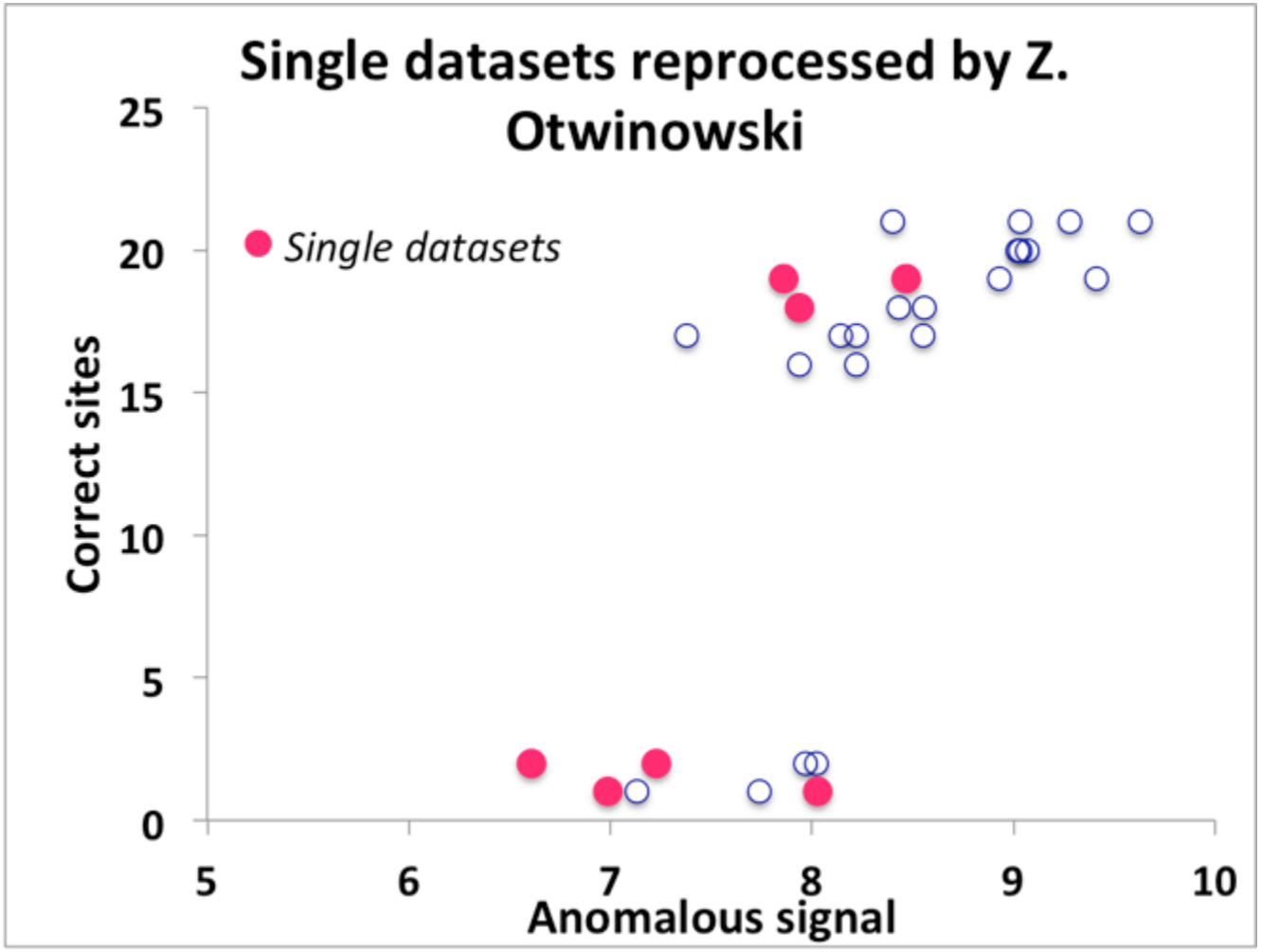
Data from 7 crystals collected at wavelength of 1.74 Å to resolution of 2.3 Å

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

CysZ multi-crystal sulfur-SAD data

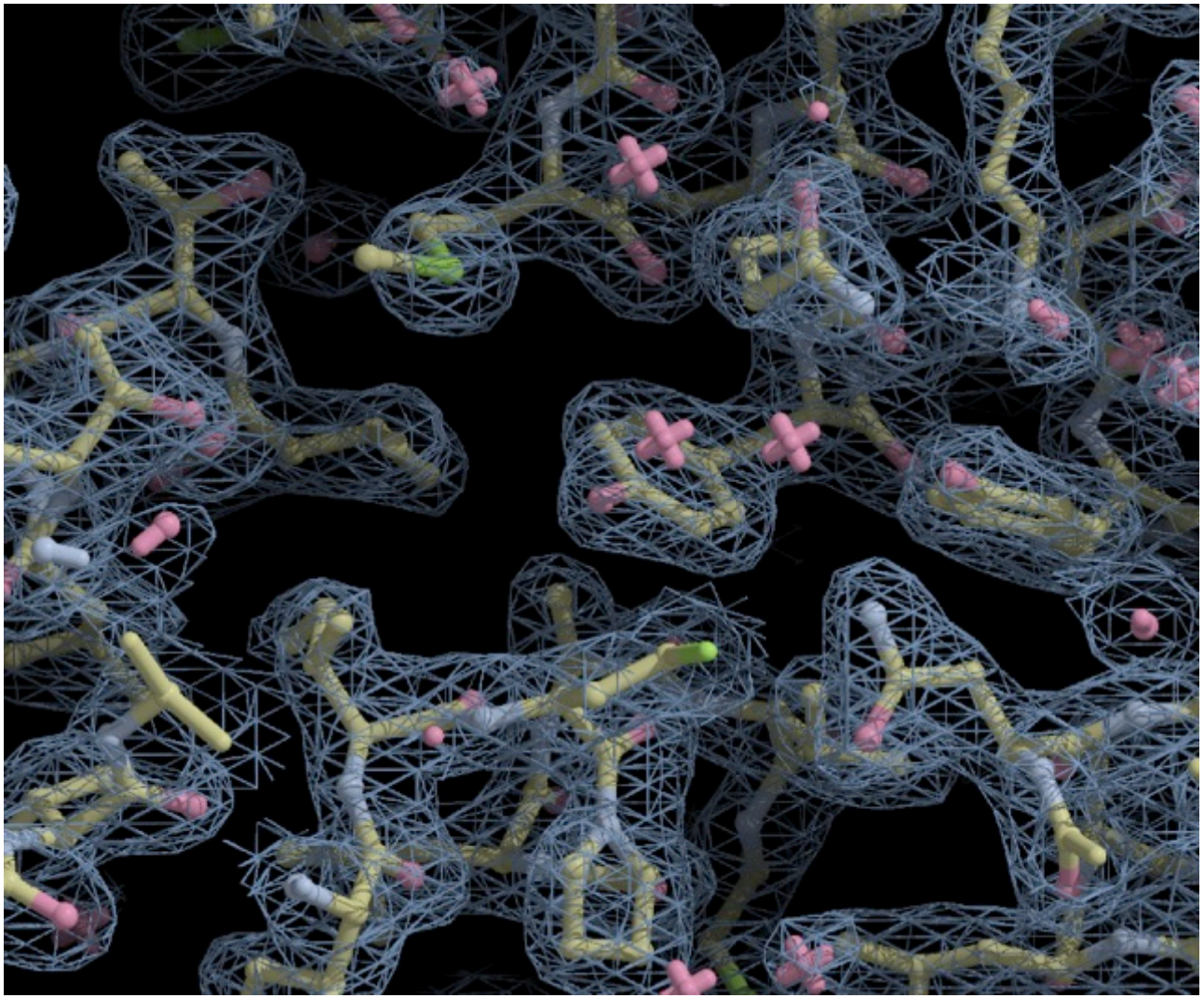


CysZ multi-crystal sulfur-SAD data



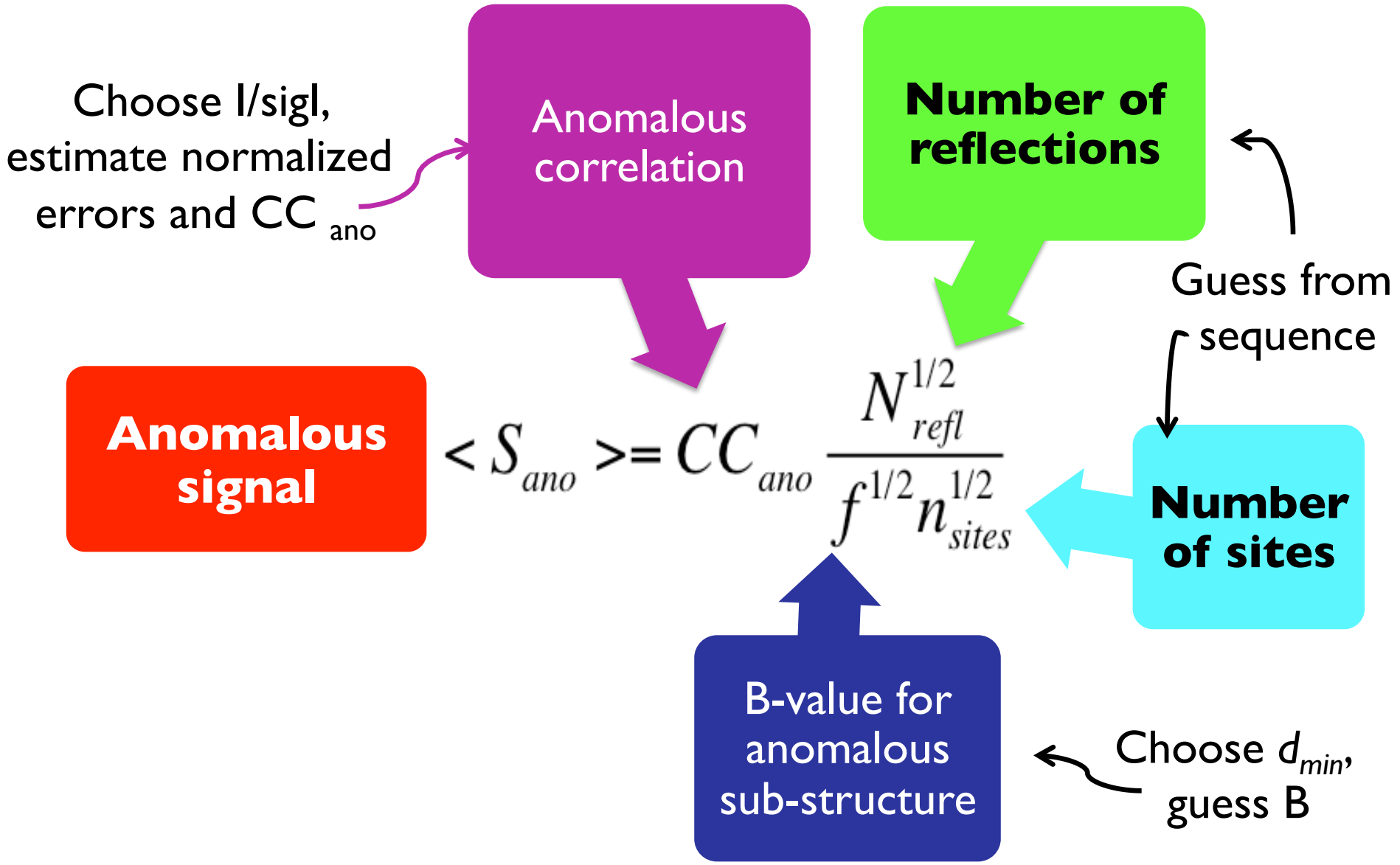
CysZ single-crystal sulfur-SAD data

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



phenix.plan_sad_experiment

Design an experiment that will give you enough anomalous signal



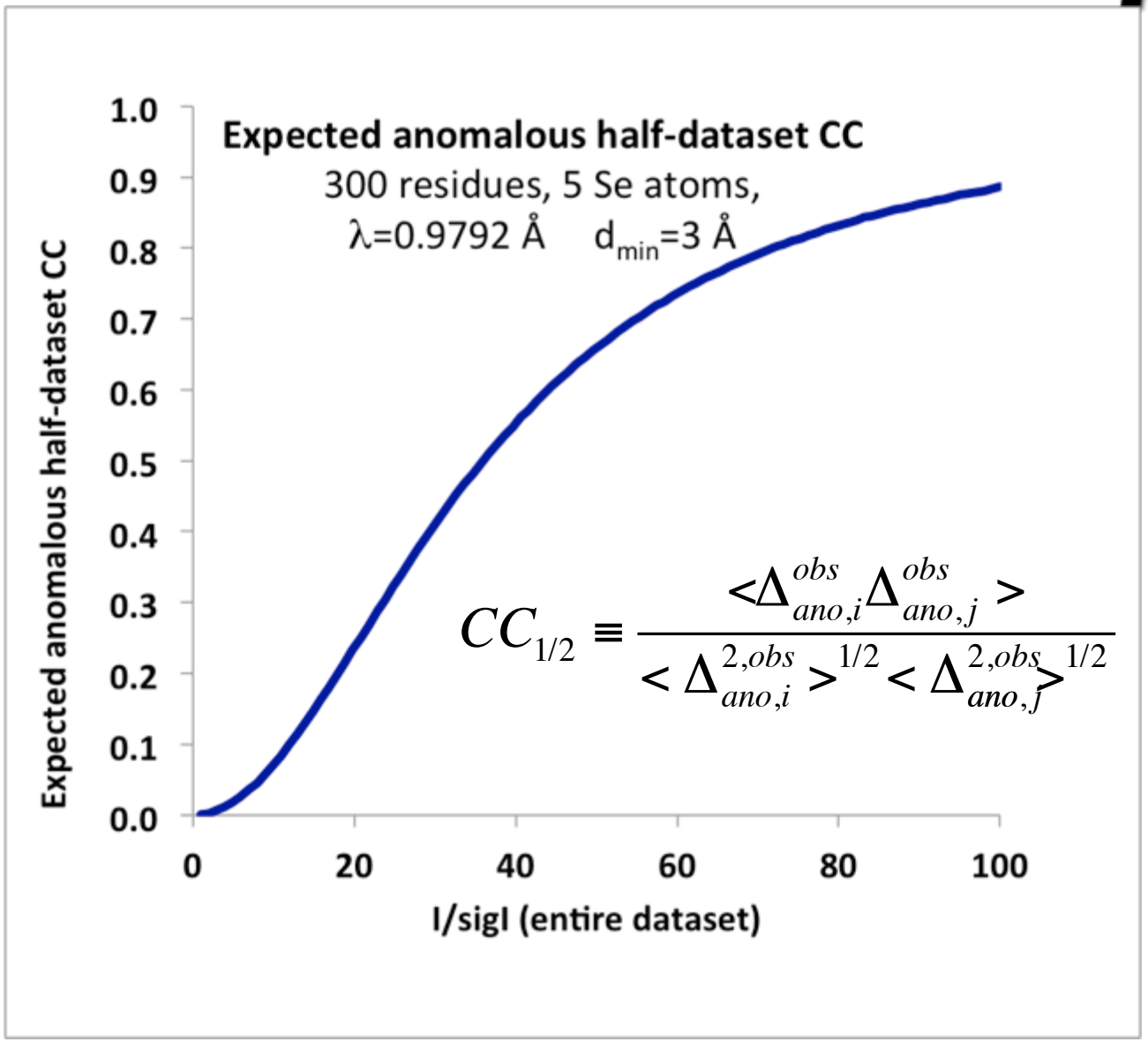
Phenix

Will I solve my structure?

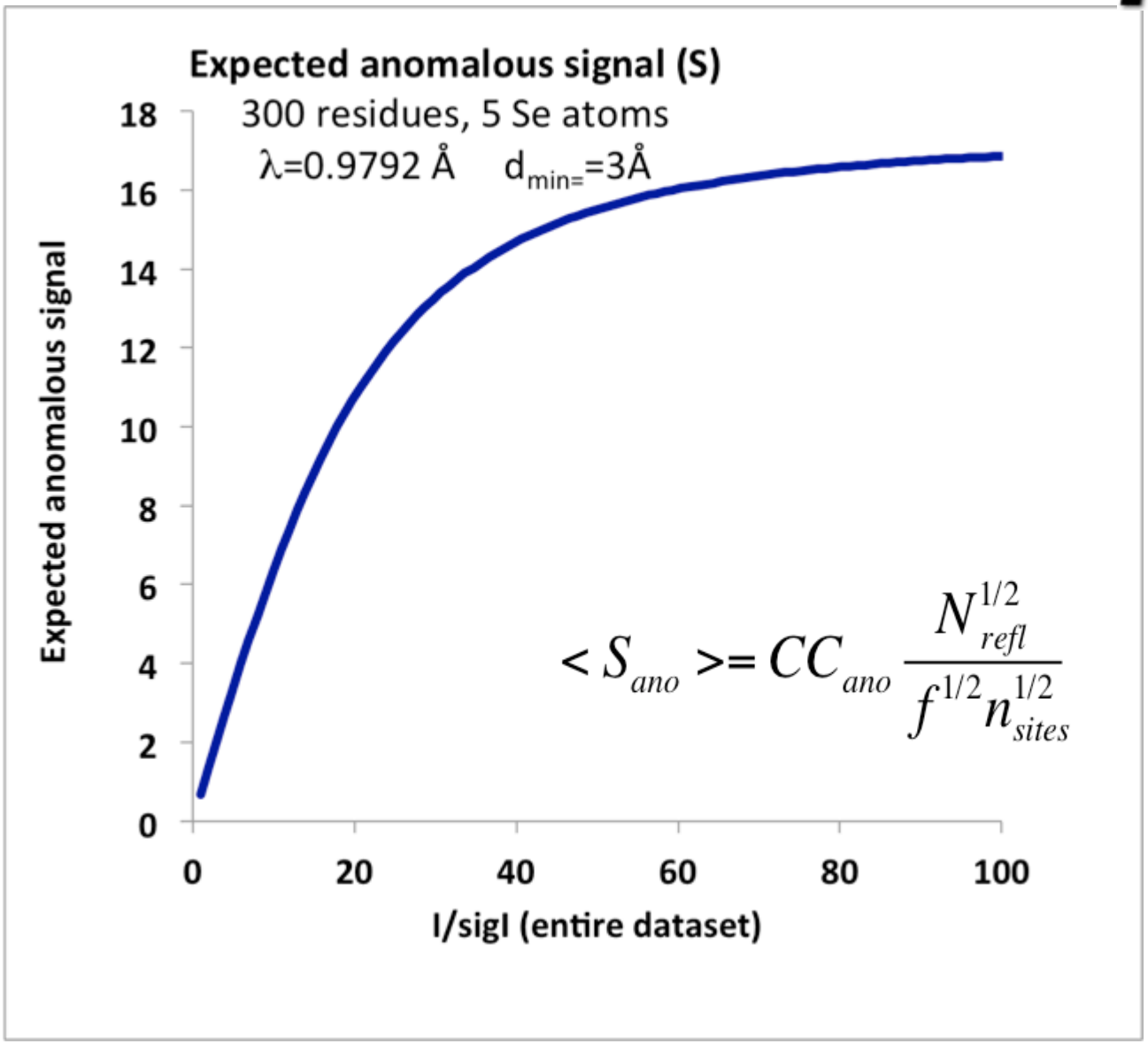
Simulate experiment with
phenix.plan_sad_experiment based on:

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

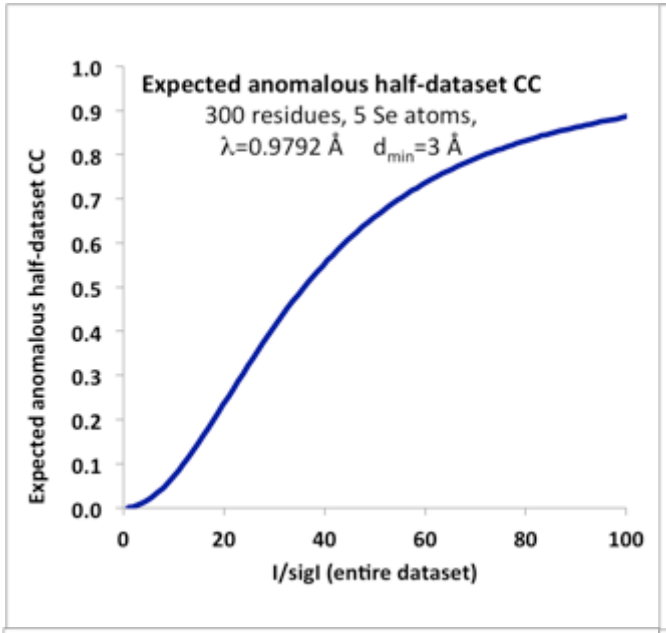
Anomalous half-dataset correlation depends on I/sigI

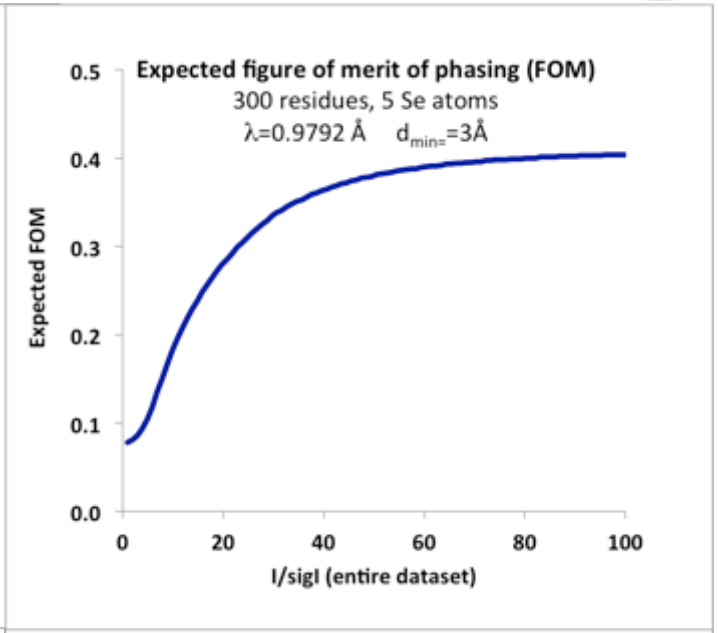


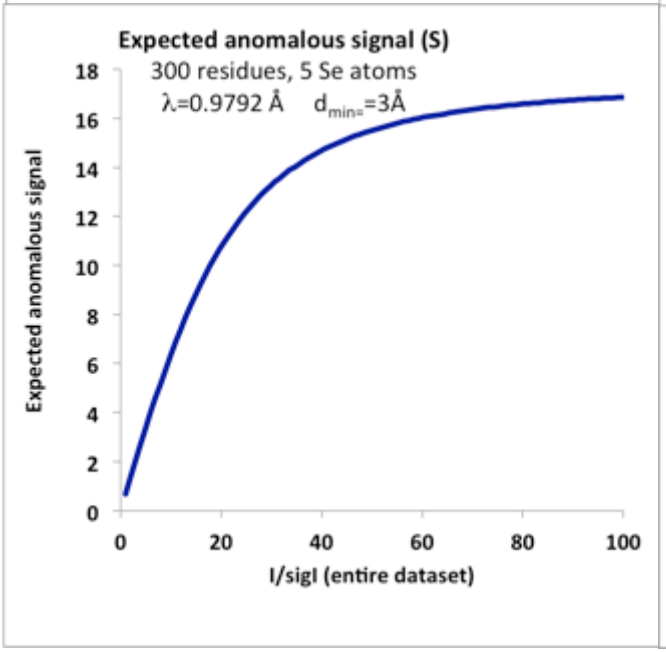
Anomalous signal depends on I/sigI

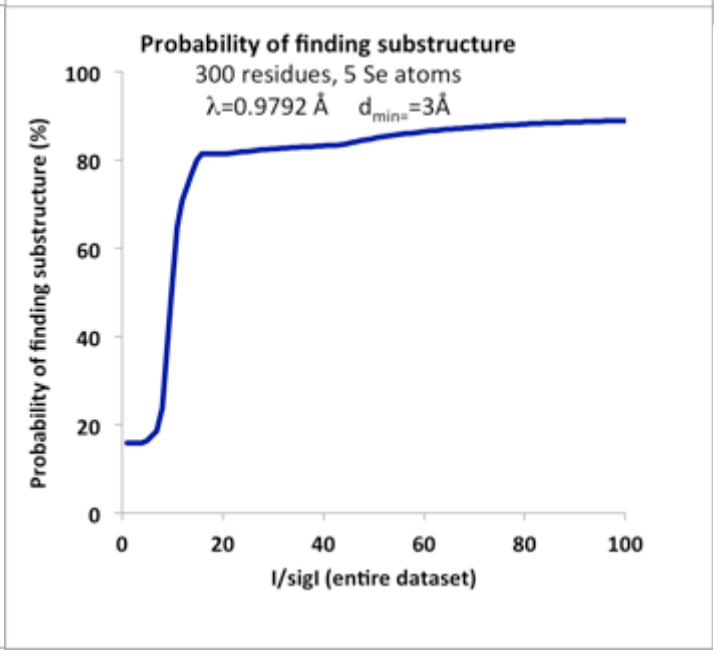


Probability of finding substructure depends on I/sigI



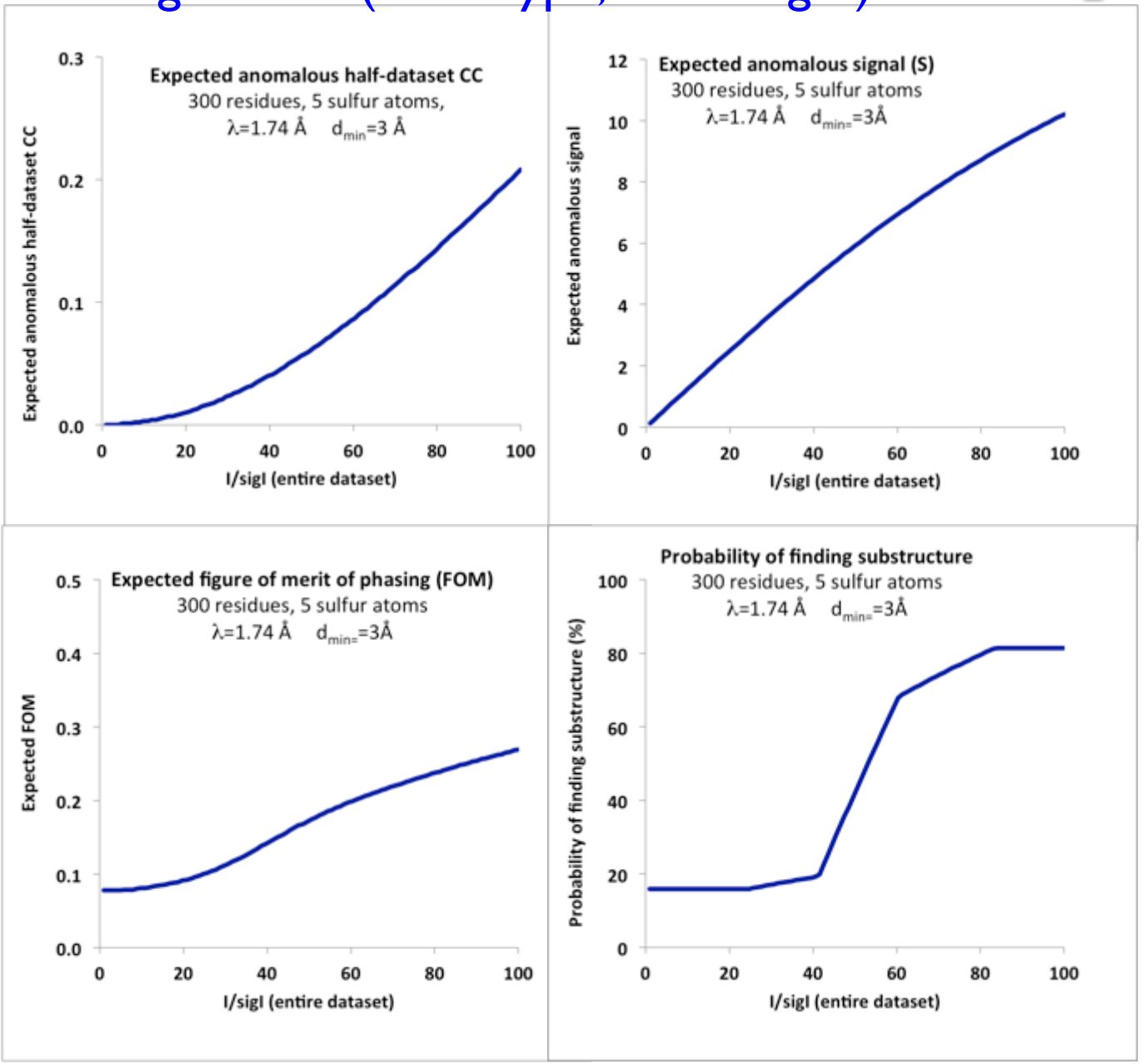




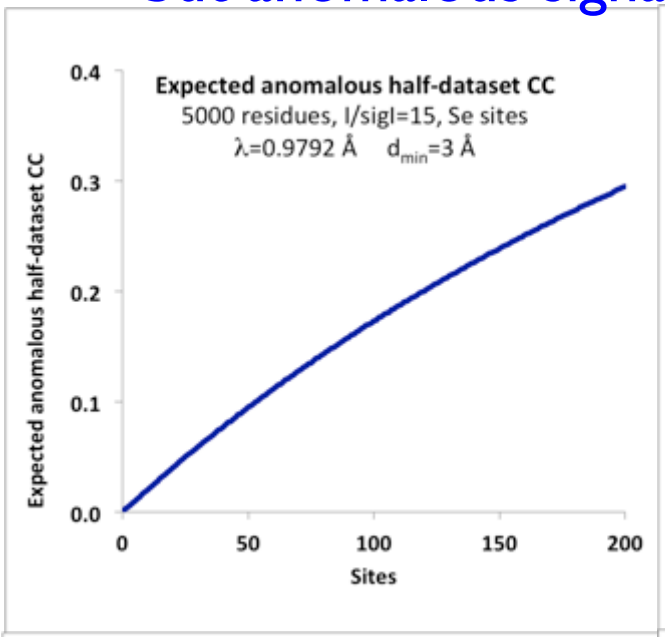


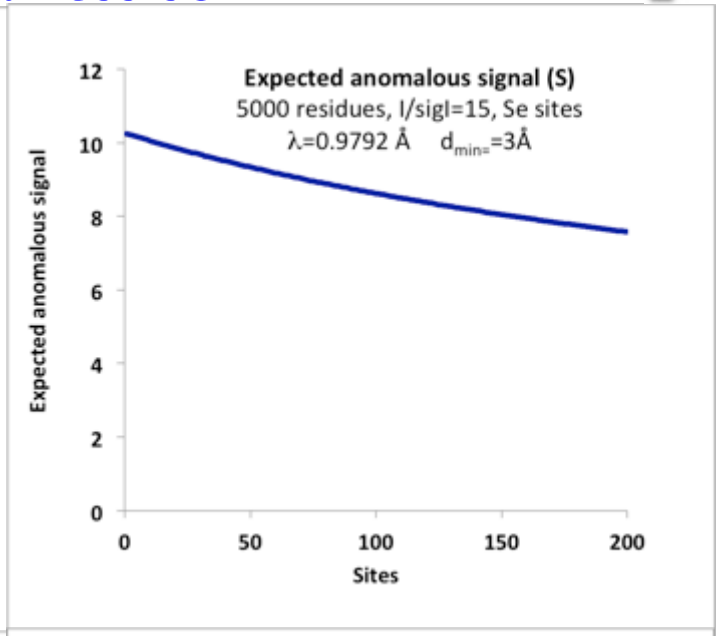
Anomalous data quality depends on I/sigI … and scattering factors (atom type, wavelength)

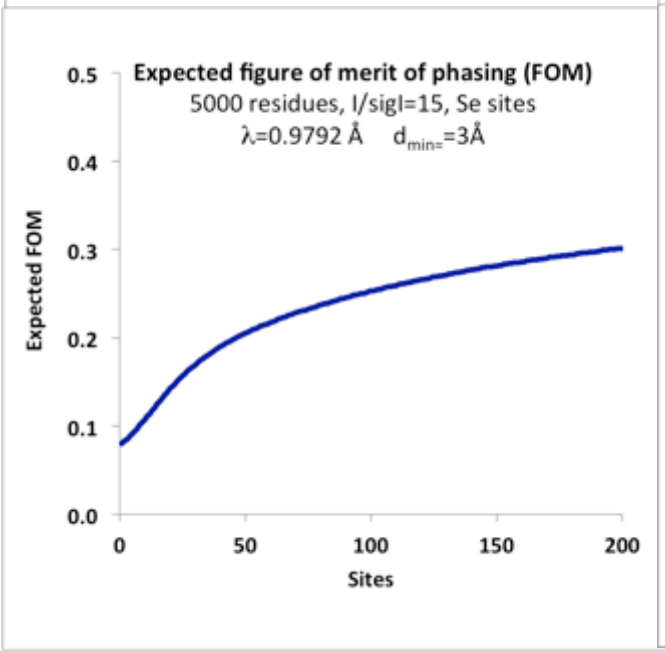
Phenix

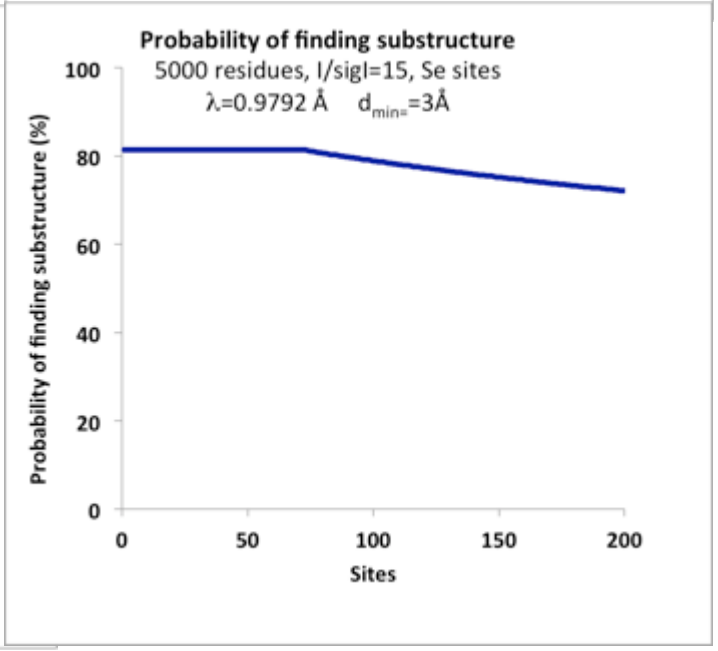


Phasing quality depends a lot on number of sites… but anomalous signal less so



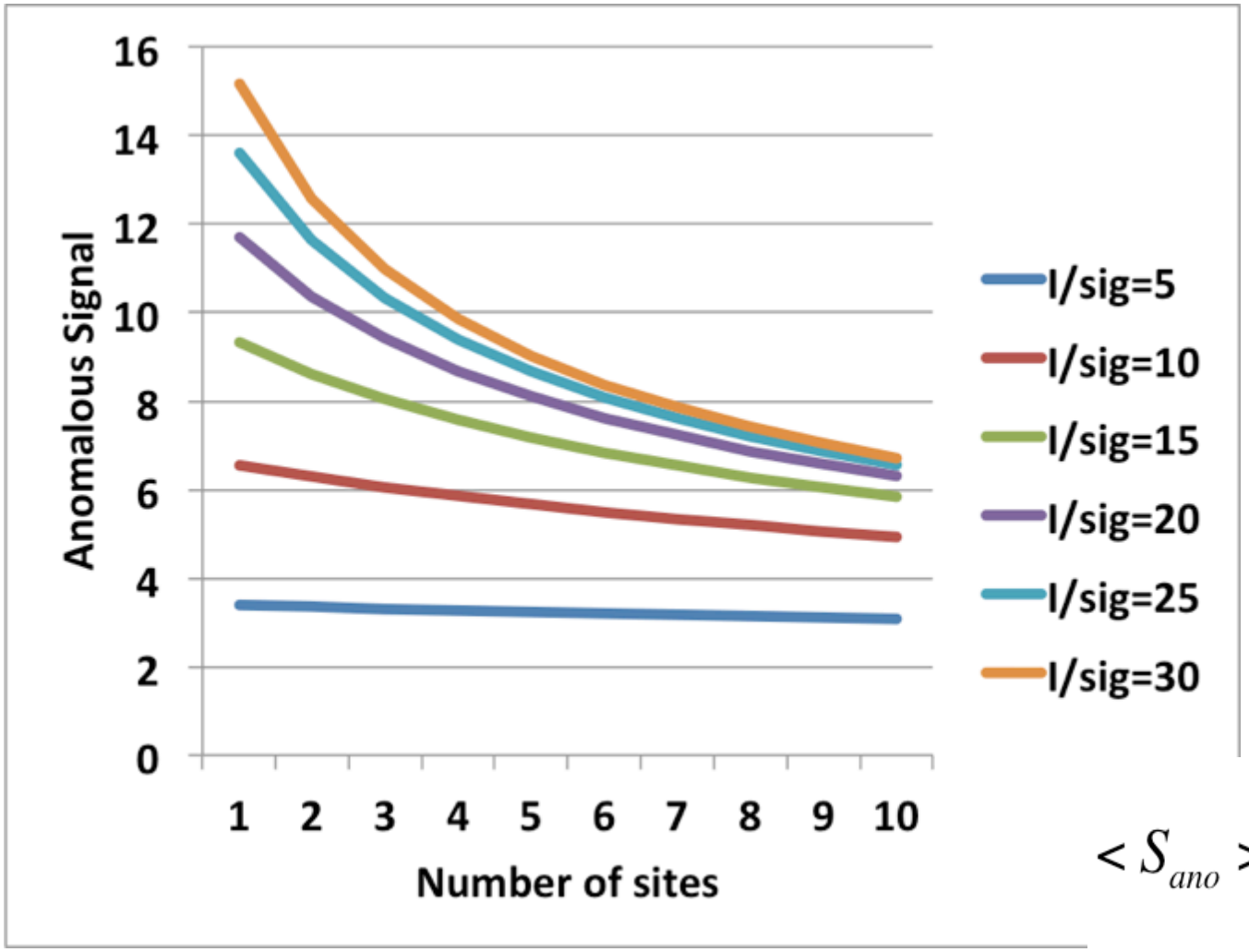






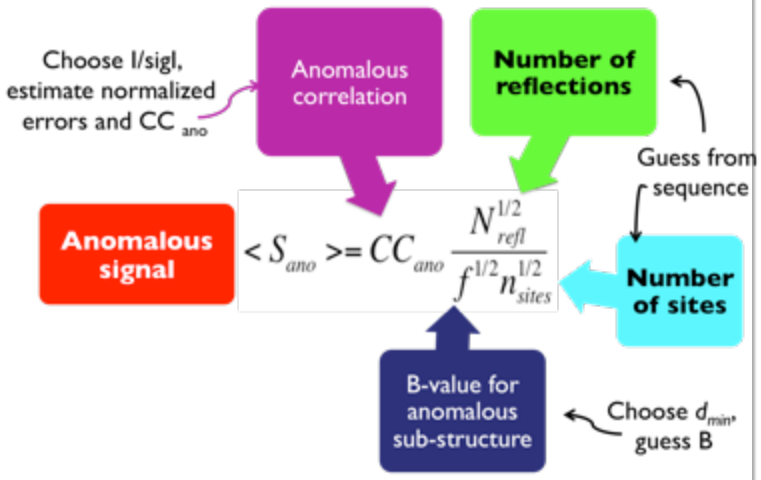
Anomalous signal vs I/sigI and sites

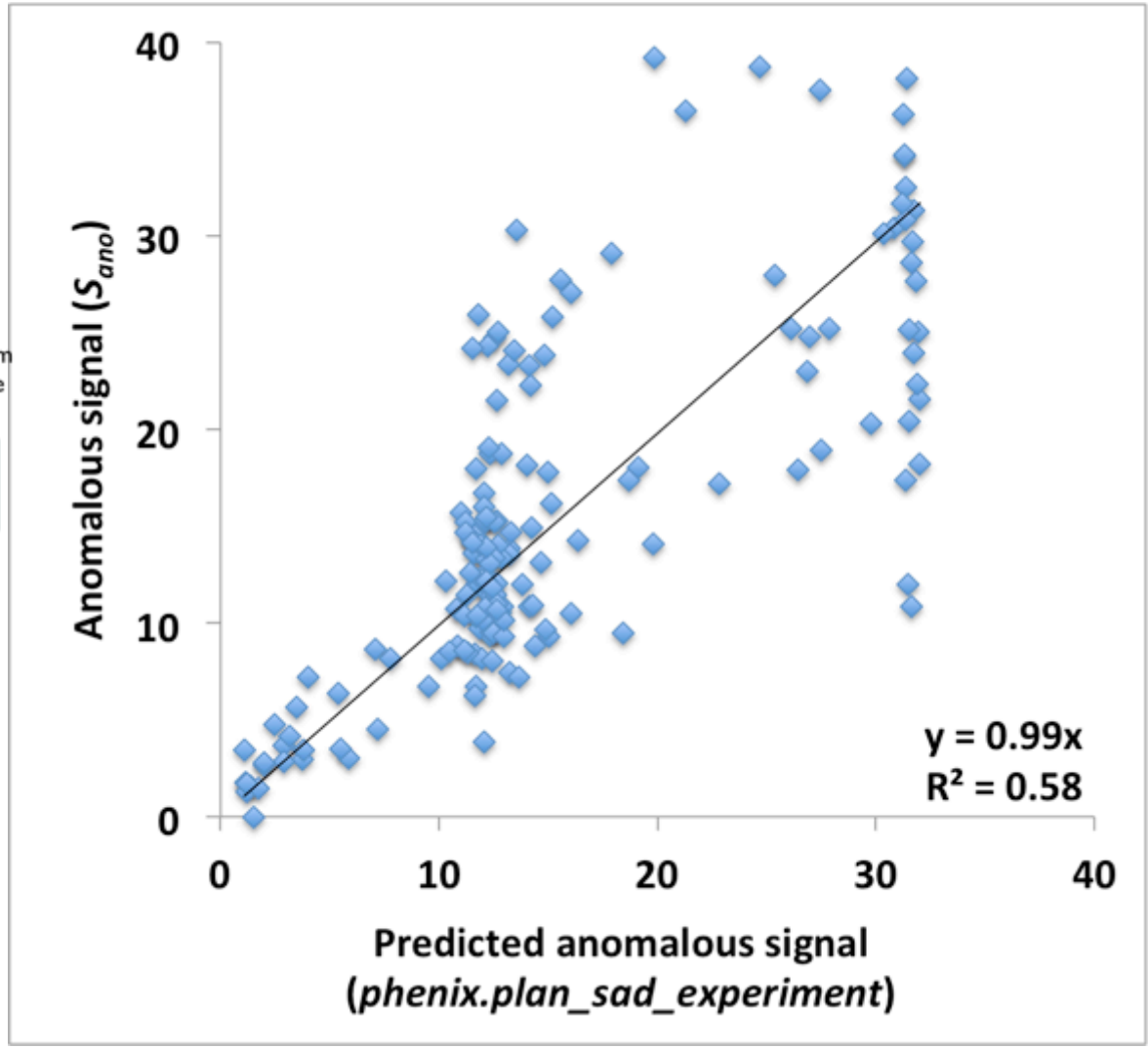
100 residues, varying Se, varying I/sigma



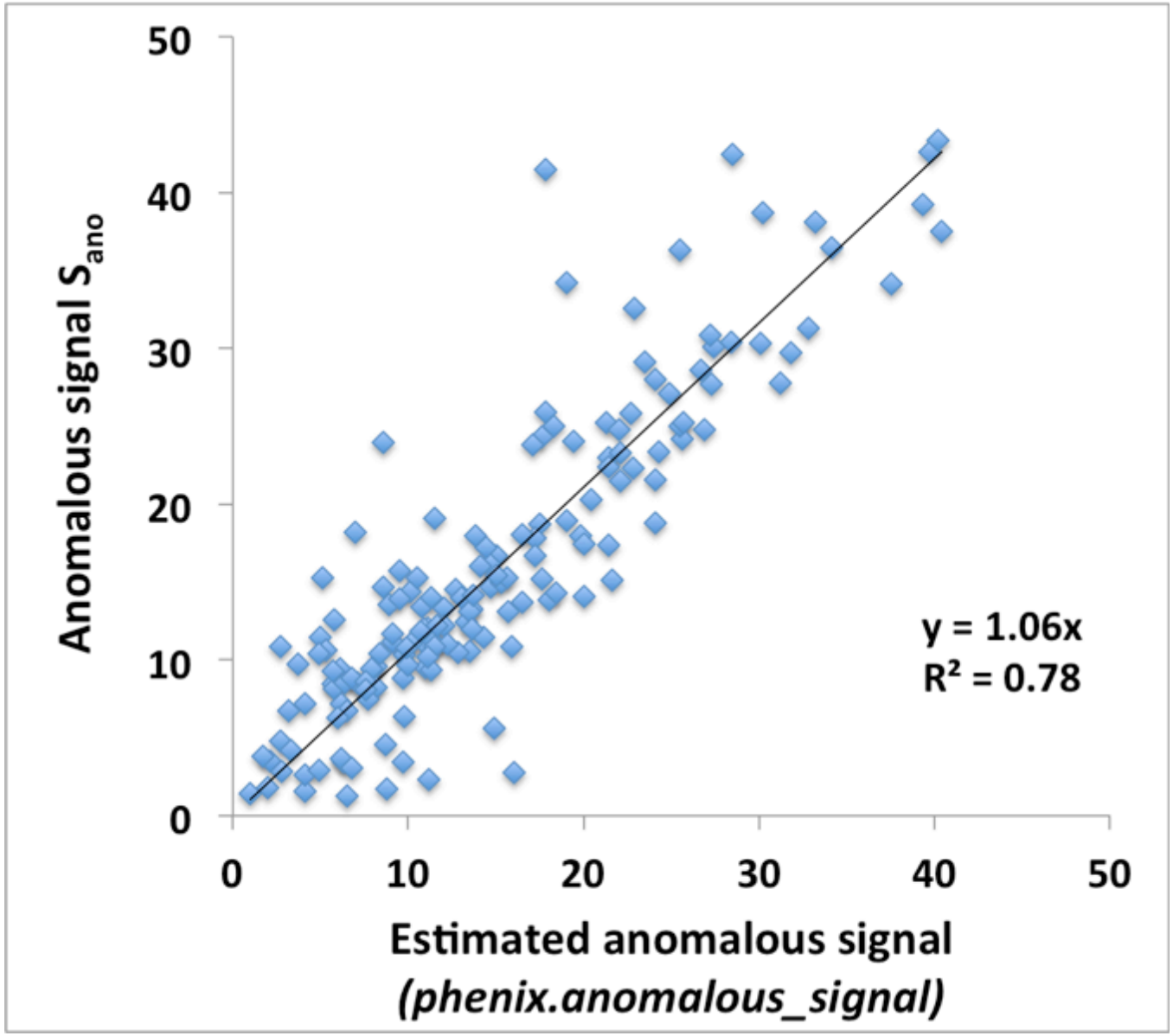
$$< S_{ano} >= CC_{ano} \frac{N_{refl}^{1/2}}{f^{1/2} n_{sites}^{1/2}}$$

Estimating the anomalous signal before collecting the data





Estimating the anomalous signal after collecting the data



Phenix

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.

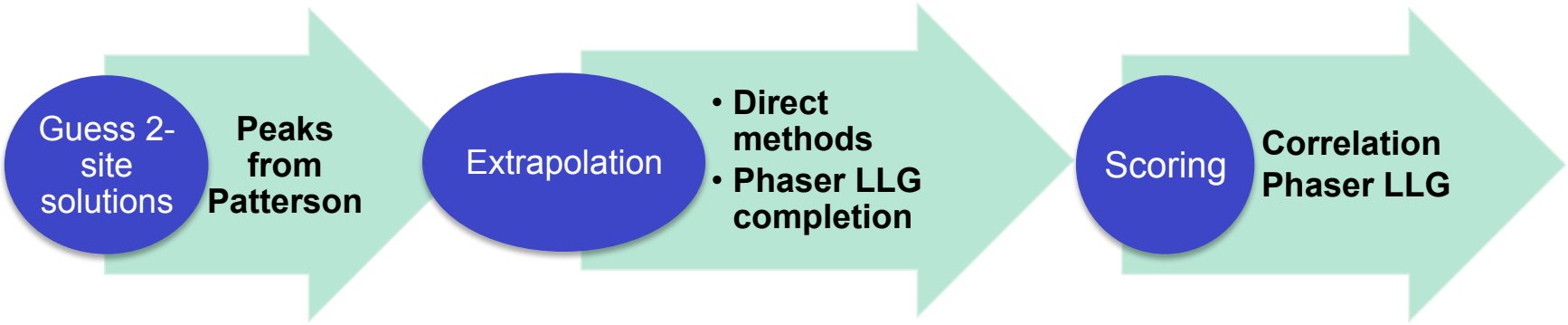
LLG sub-structure searches in HySS

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



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

Adjustable LLGC_SIGMA

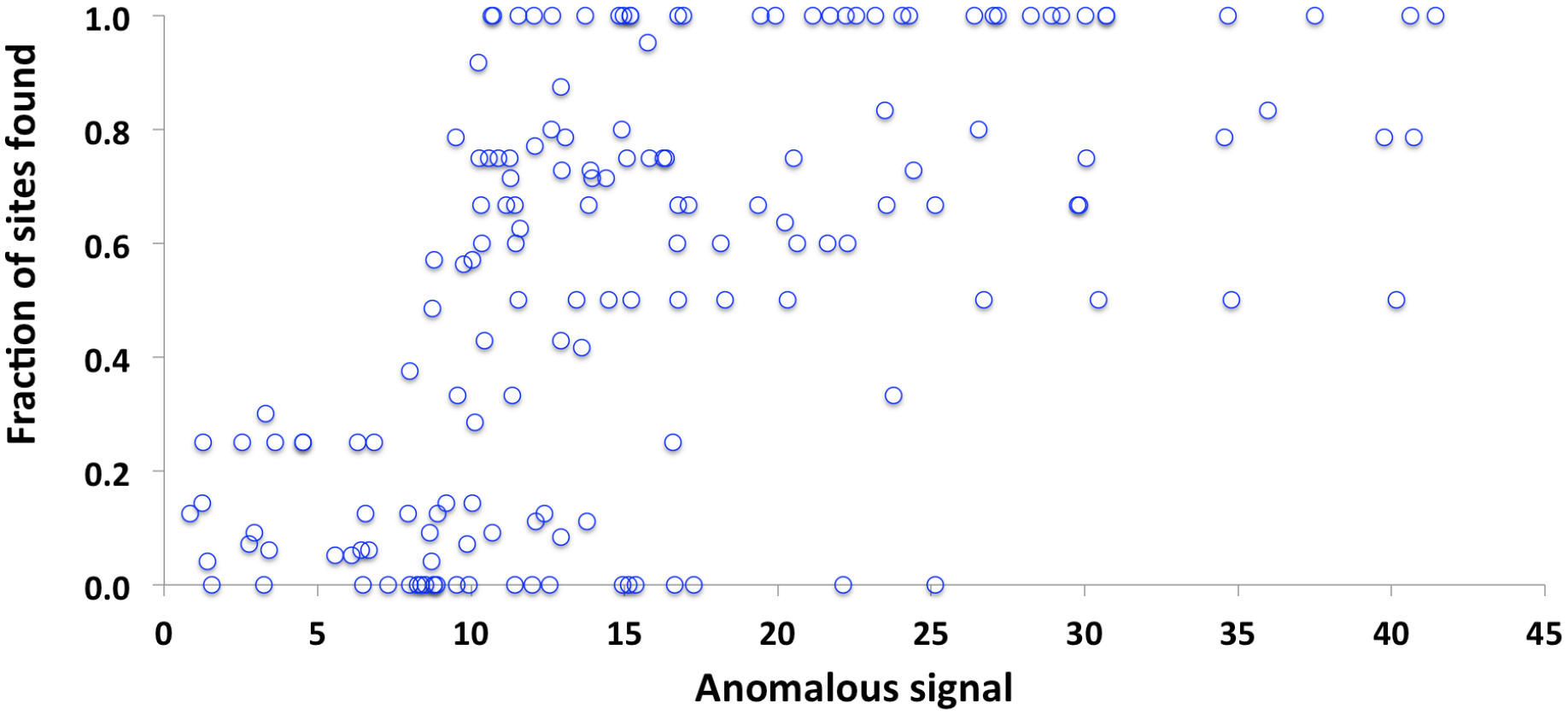
(cut-off for peak height)

Use LLG score to compare solutions

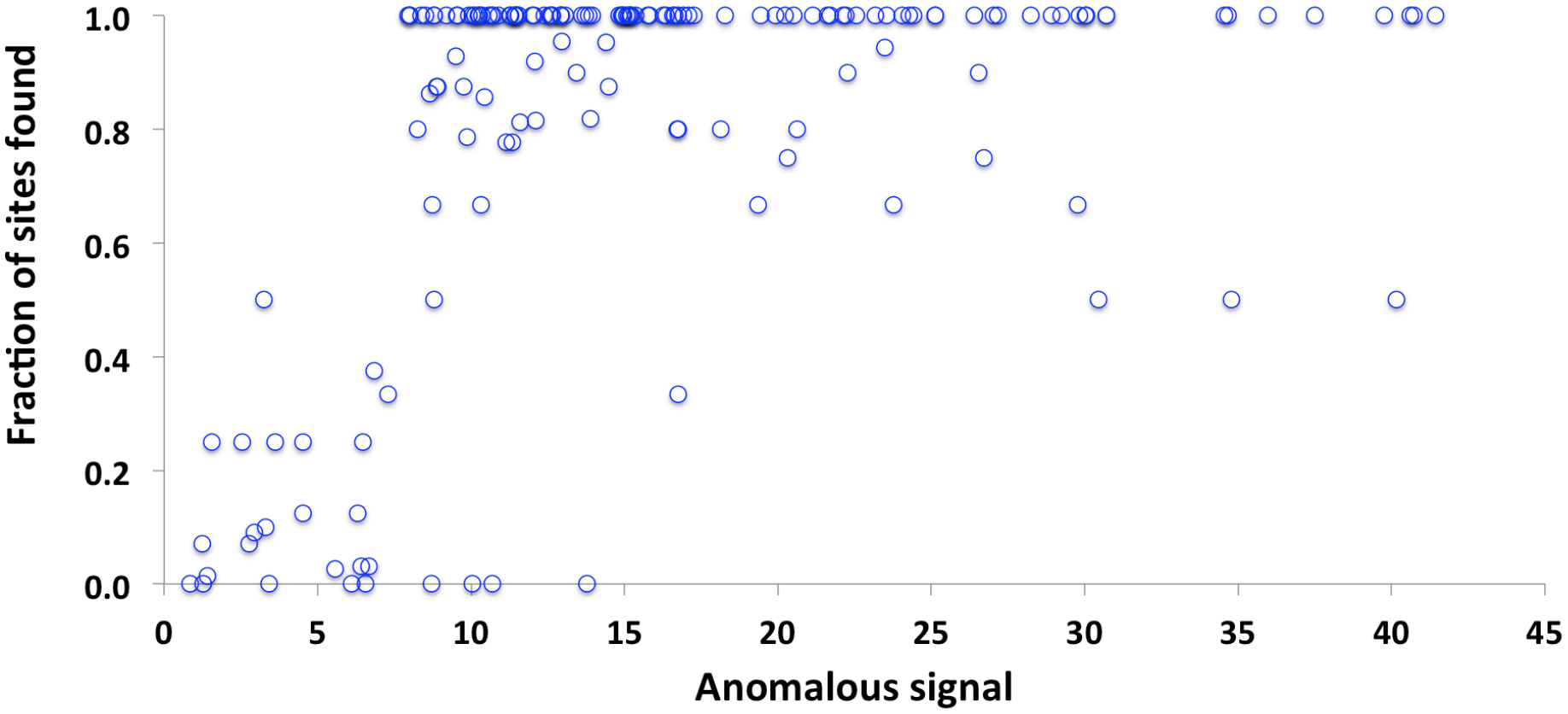
Terminate early if same solution found several times

Run quick direct methods first

Dual Space Sub-structure Completion

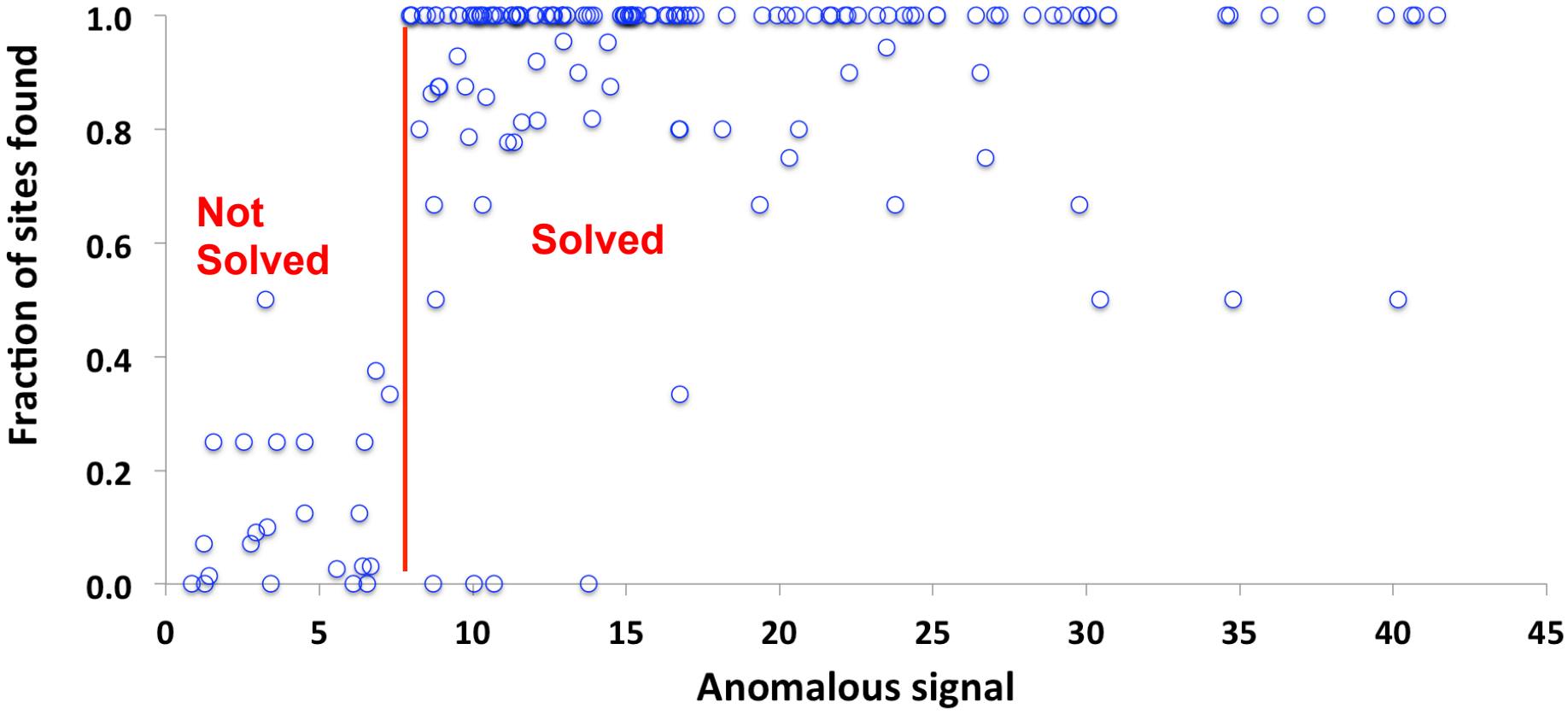


LLG Sub-structure Search



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

Anomalous signal indicates if a dataset can be solved



The Phenix Team

Phenix

Lawrence Berkeley Laboratory

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

Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung

Cambridge University

Randy Read, Airlie McCoy, Gabor Bunkóczi, Robert Oeffner

Duke University

Jane & David Richardson, Christopher Williams, Bradley Hintze

An NIH/NIGMS funded Program Project