Phenix

Structure solution from weak anomalous data

Diffraction Methods in Structural Biology

Gordon Research Conference

Bates College, Maine

July 30, 2014

Gábor Bunkóczi, Airlie McCoy, Randy Read (Cambridge University)
Nat Echols, Ralf Grosse-Kunstleve, Paul Adams, James Holton (Lawrence Berkeley National Laboratory)
Tom Terwilliger (Los Alamos National Laboratory)

Phenix

Structure solution from weak anomalous data

Problems with weak signal

Quantifying the anomalous signal

Solving the anomalous sub-structure with weak signal

Solving structures with weak signal

Estimating the anomalous signal from the data

Weak anomalous signal

Reasons:

Few anomalous scatterers, sulfur SAD, weak diffraction, wavelength far from peak

Consequences:

Substructure identification is difficult

Phasing is poor

Iterative density modification, model-building and refinement works poorly

Quantifying the anomalous signal I

Phenix

CC_{ano}: How accurate are the anomalous differences?

Anomalous differences measured with errors ε_j

$$\Delta_{ano,j}^{obs} = \Delta_{ano,j} + \varepsilon_j$$

Correlation of observed and true anomalous differences

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

Fraction of observed anomalous differences that is noise

$$E_{ano}^{2} = \frac{<\sigma_{ano}^{2}>}{<\Delta_{ano}^{2,obs}>}$$

Expected value of CC_{ano}

$$CC_{ano} \sim [1 - E_{ano}^{2}]^{1/2}$$

Quantifying the anomalous signal II

Phenix

Anomalous signal S_{ano}: How accurate are maps based on the anomalous differences?

Anomalous difference Fourier with model phases

$$\rho(x) = \frac{1}{V}\sum_h \Delta^{obs}_{ano,h} e^{i(\varphi^c_h - \frac{\pi}{2})} e^{-2\pi i(h.x)}$$

Peak height at coordinates of anomalously-scattering atoms

$$S_{ano} \equiv \frac{<\rho(x_k)>}{<\rho^2>^{1/2}}$$

Expected value of signal S_{ano}

$$S_{ano} \sim CC_{ano}\frac{N^{1/2}_{refl}}{N^{1/2}_{sites}(\frac{5}{4})^{1/2}}$$

Phenix

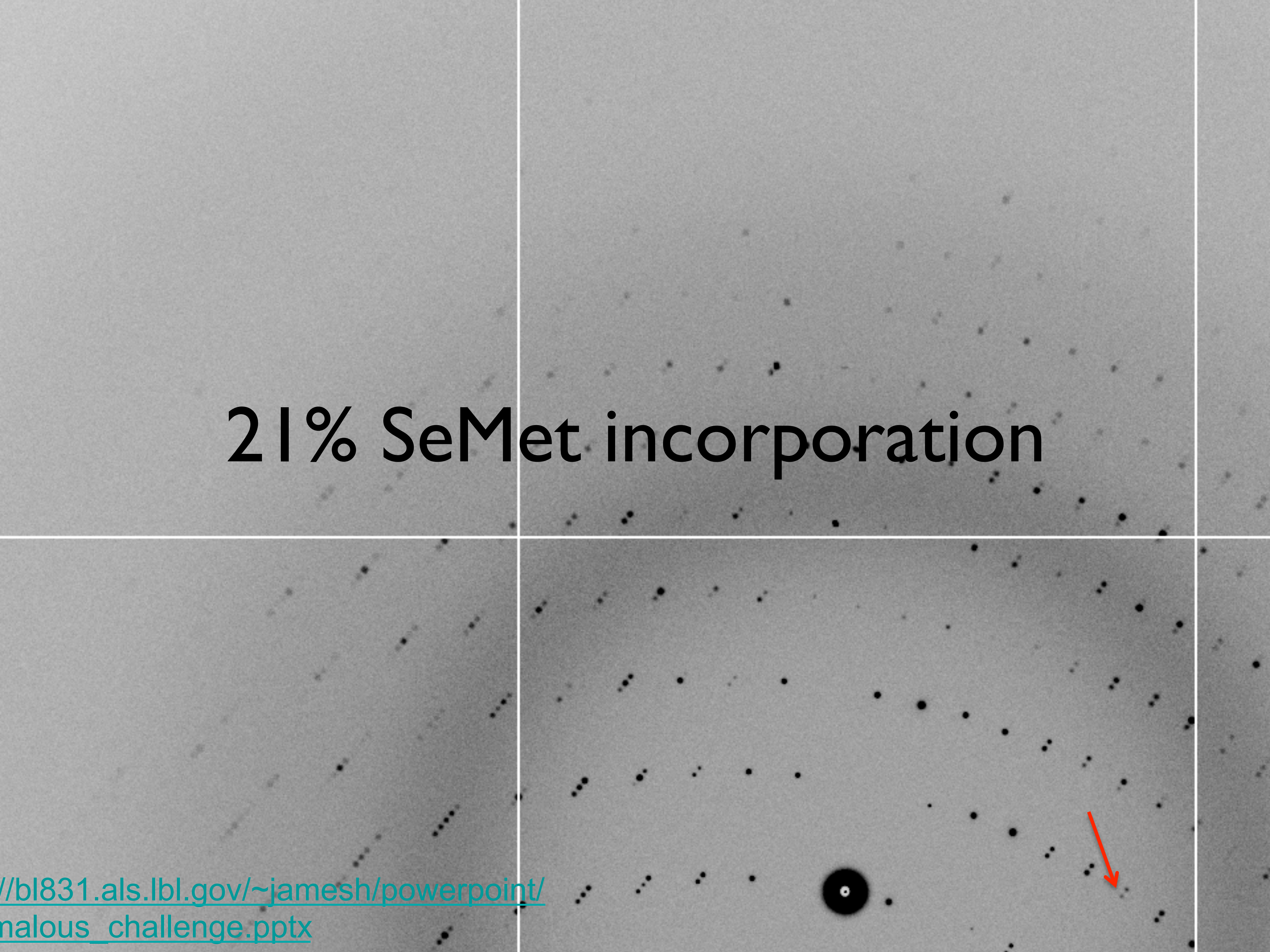
Example of anomalous signal S_{ano}
Holton Challenge data

bl831.als.lbl.gov/~jamesh/challenge/anom/

Simulated diffraction data from 3dk0 to 1.8 Å (useful to 2.3 Å)

0% to 100% occupancy of Se in selenomethionine

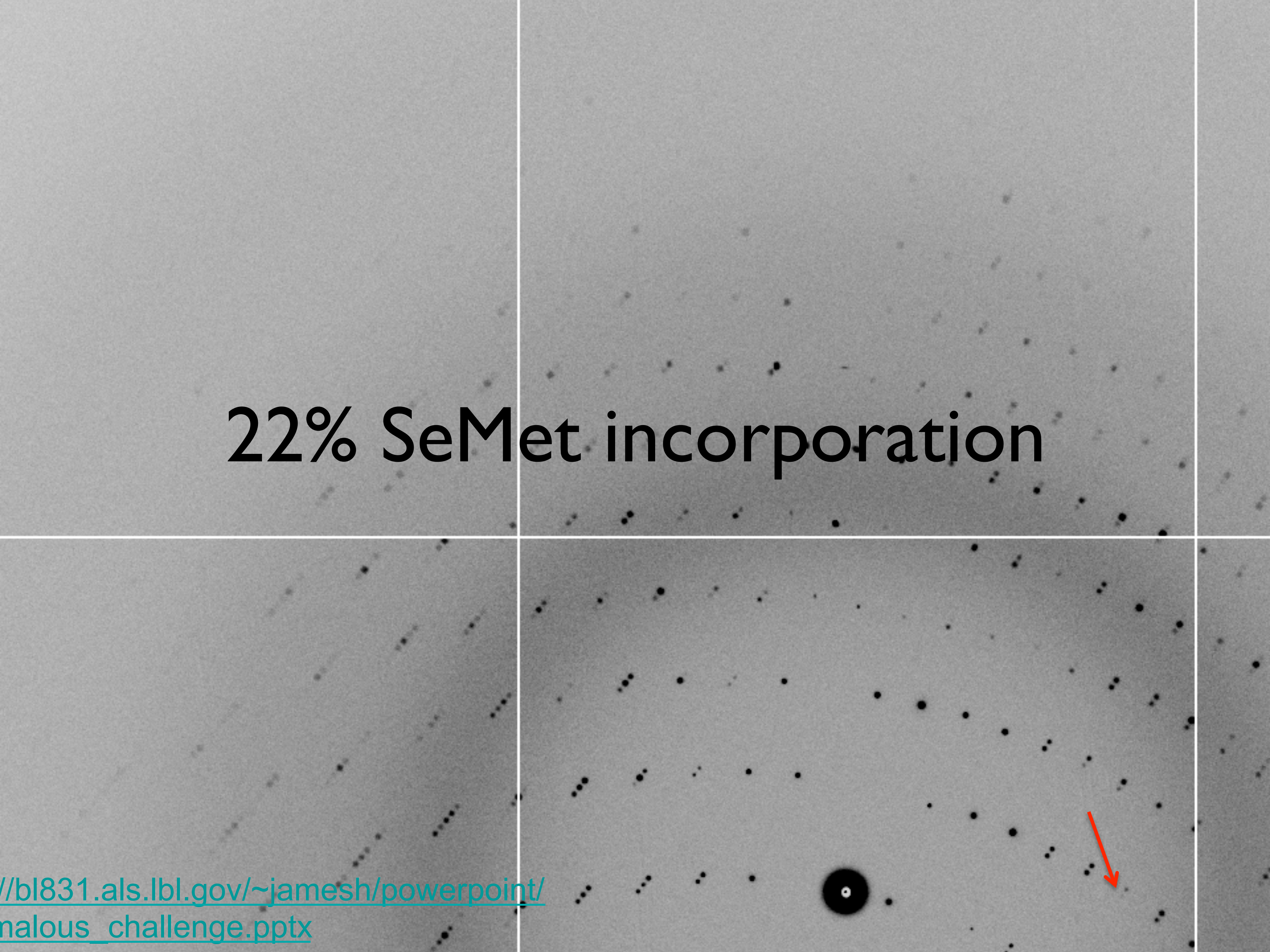
"*Impossible.mtz*" has fraction Se of 0.21



21% SeMet incorporation

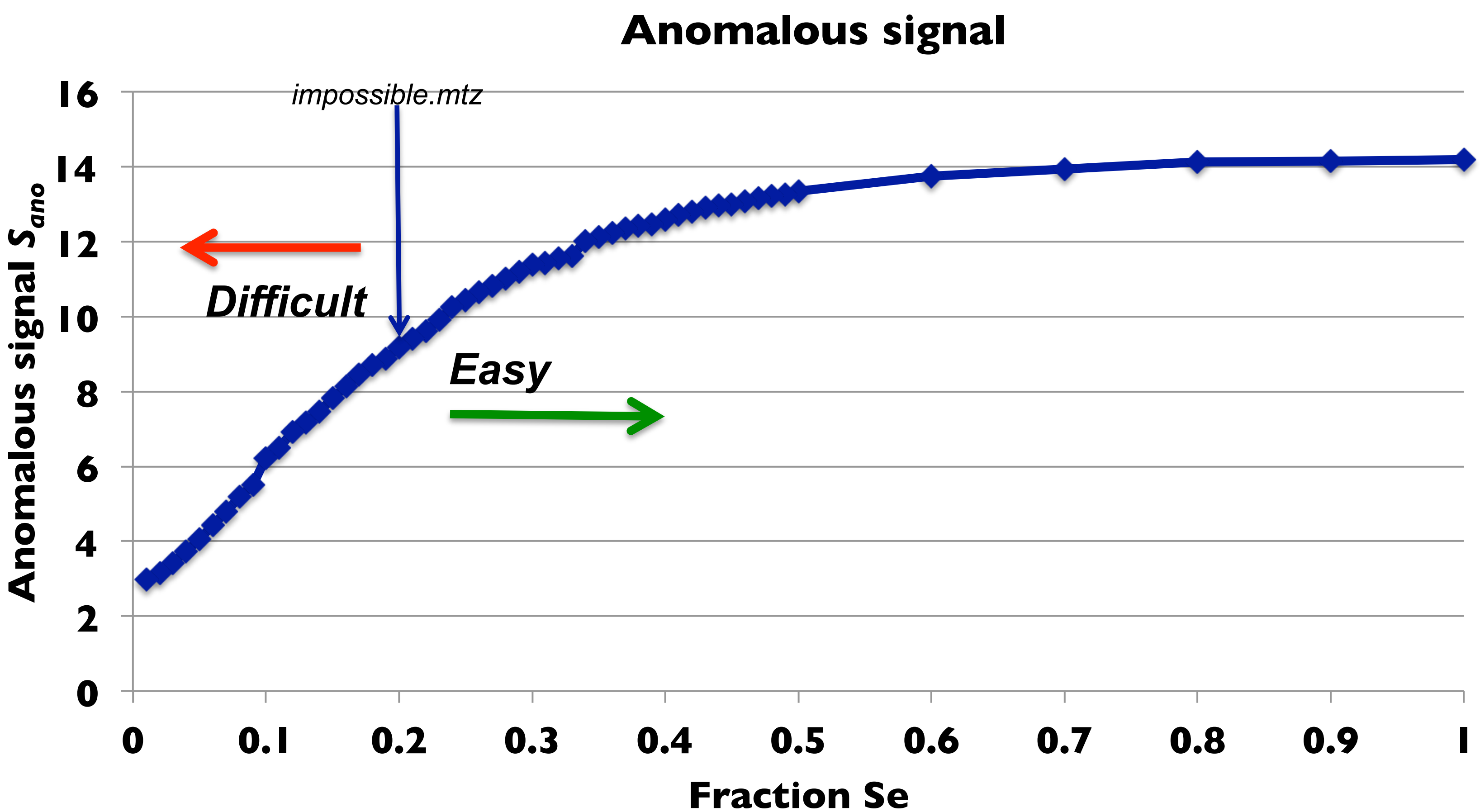
//bl831.als.lbl.gov/~jamesh/powerpoint/
nalous_challenge.pptx

22% SeMet incorporation



//bl831.als.lbl.gov/~jamesh/powerpoint/
malous_challenge.pptx

Example of anomalous signal S_{ano}
Holton Challenge data



Phenix

Finding the anomalous sub-structure with weak anomalous signal

Current approaches

Dual-space methods (Shelxd, HySS, Crunch2)

Difference Fourier (Solve)

Limitation of these approaches

Anomalous differences are only approximately proportional to the structure factors for anomalously-scattering atoms

Finding the anomalous sub-structure with weak anomalous signal

Most powerful source of information about substructure before phases are known is the SAD likelihood function:

The likelihood of measuring the observed anomalous data given a partial model

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 a difference Fourier—but much 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.

Using LLG completion and dual-space completion in HySS

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

Using LLG completion in HySS

Test cases

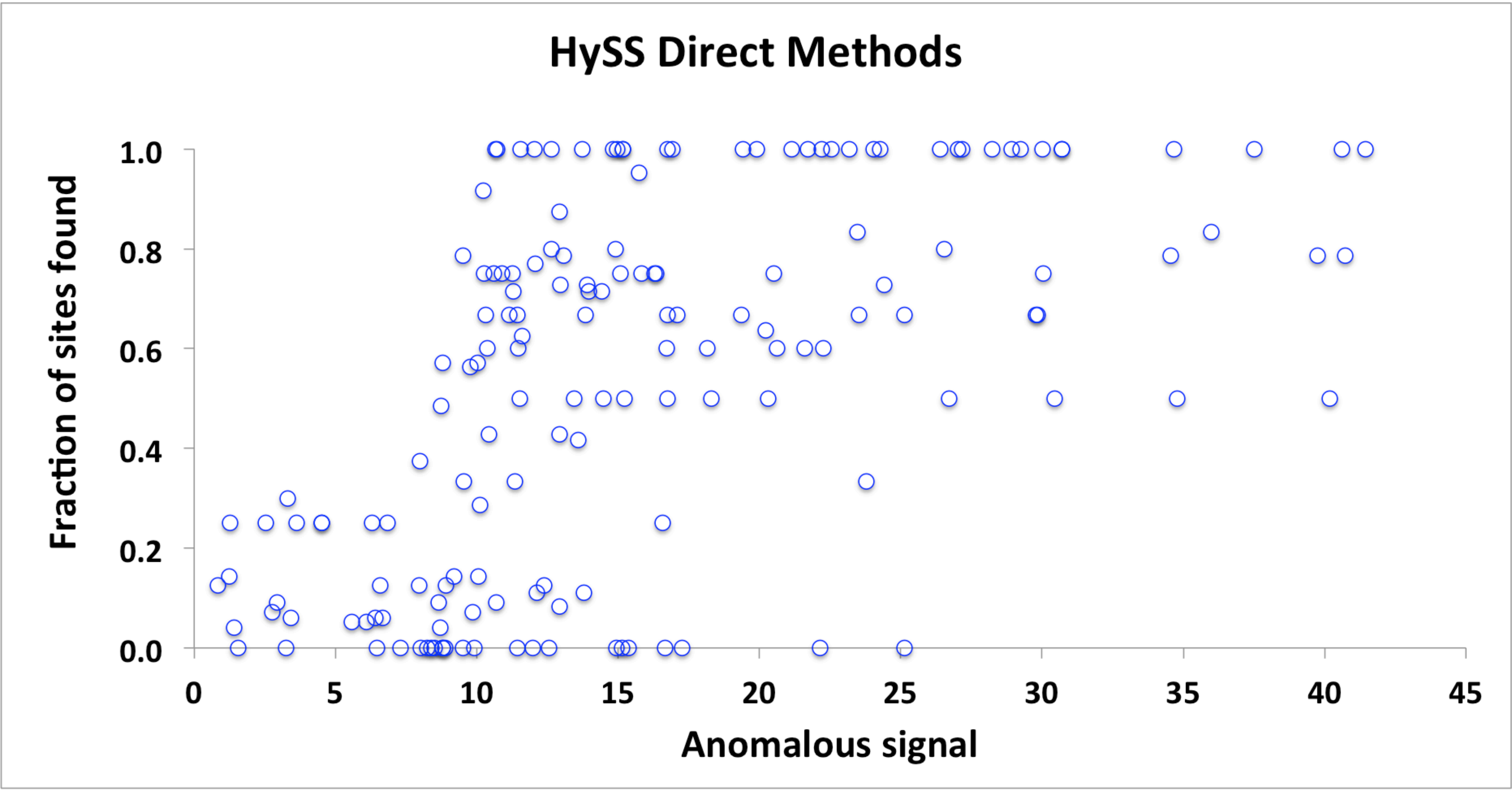
164 SAD datasets from PDB (largely JCSG MAD data)

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

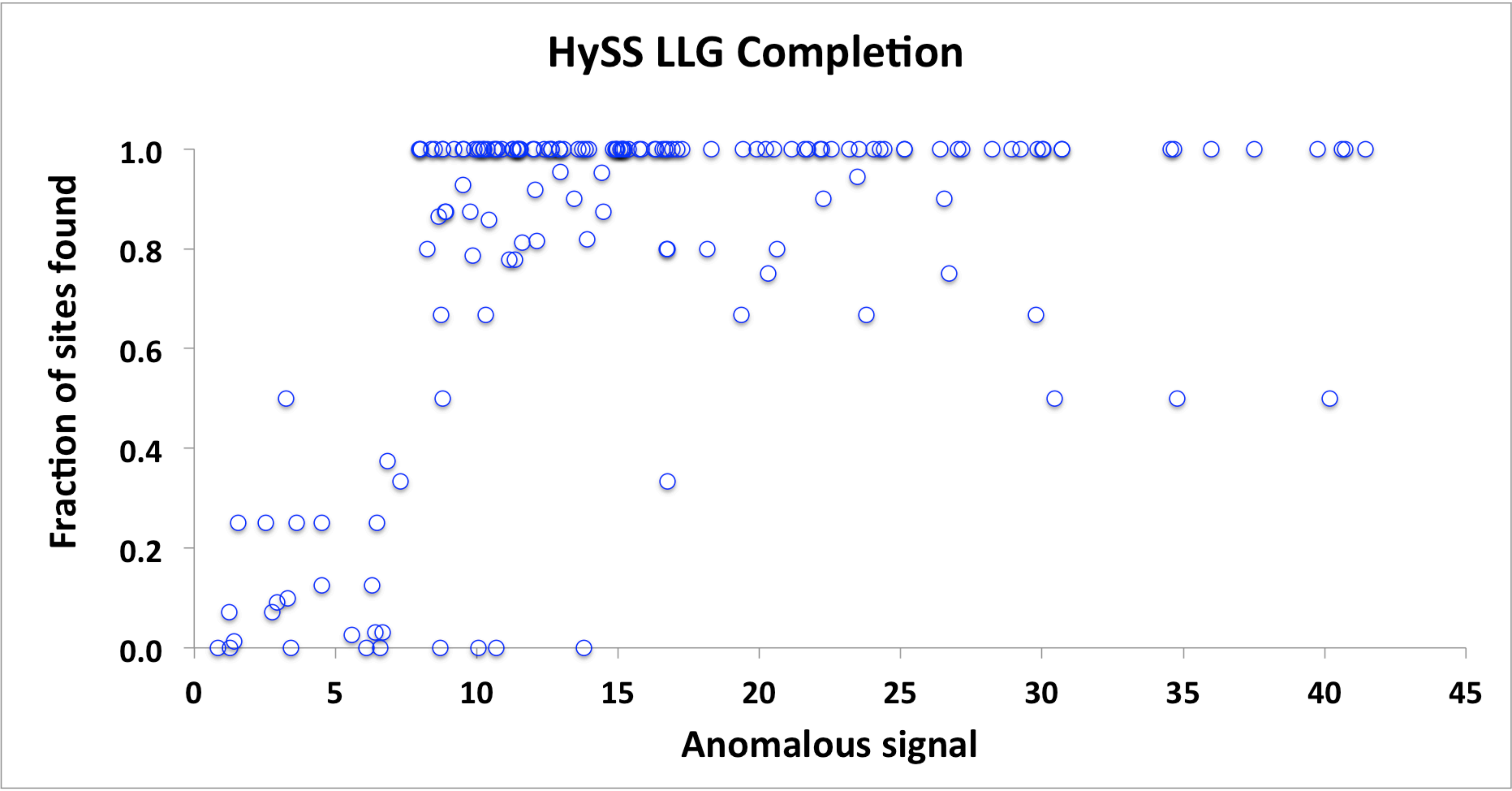
Setting up test data on 165 datasets

- *phenix.fetch_pdb 2o7t*
- *phenix.python $PHENIX/phenix/phenix/autosol/sad_data_from_pdb.py 2o7t*
- Splits out each wavelength (peak, edge, remote etc) for MAD and run separately
- Run HySS with dual-space methods or LLG completion

Direct methods vs LLG completion
164 SAD datasets from PDB



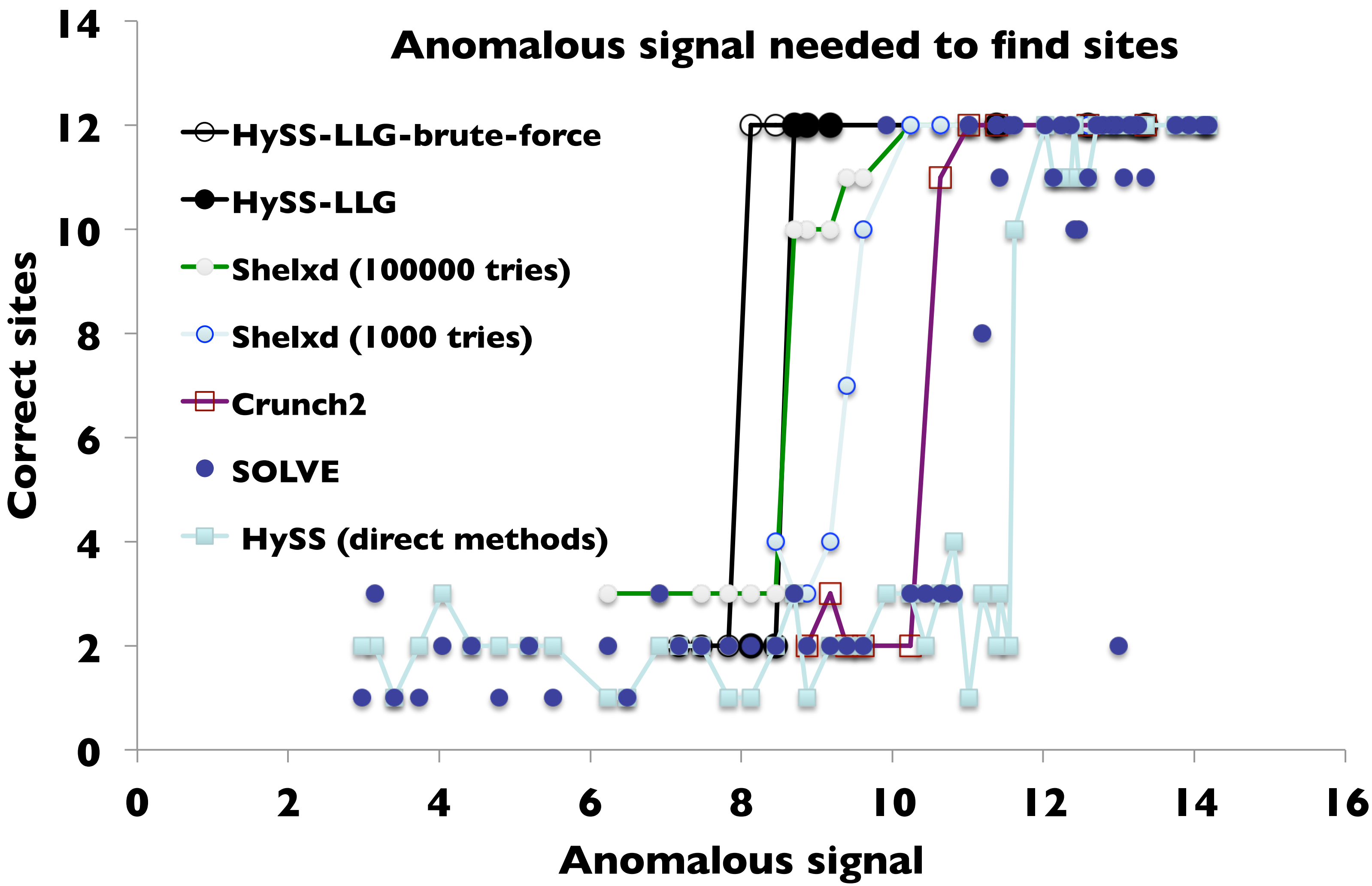
Direct methods vs LLG completion
164 SAD datasets from PDB



Phenix

Holton Challenge data

Correct sites found vs anomalous signal S_{ano}



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 1.74 Å

Only merged data could be solved

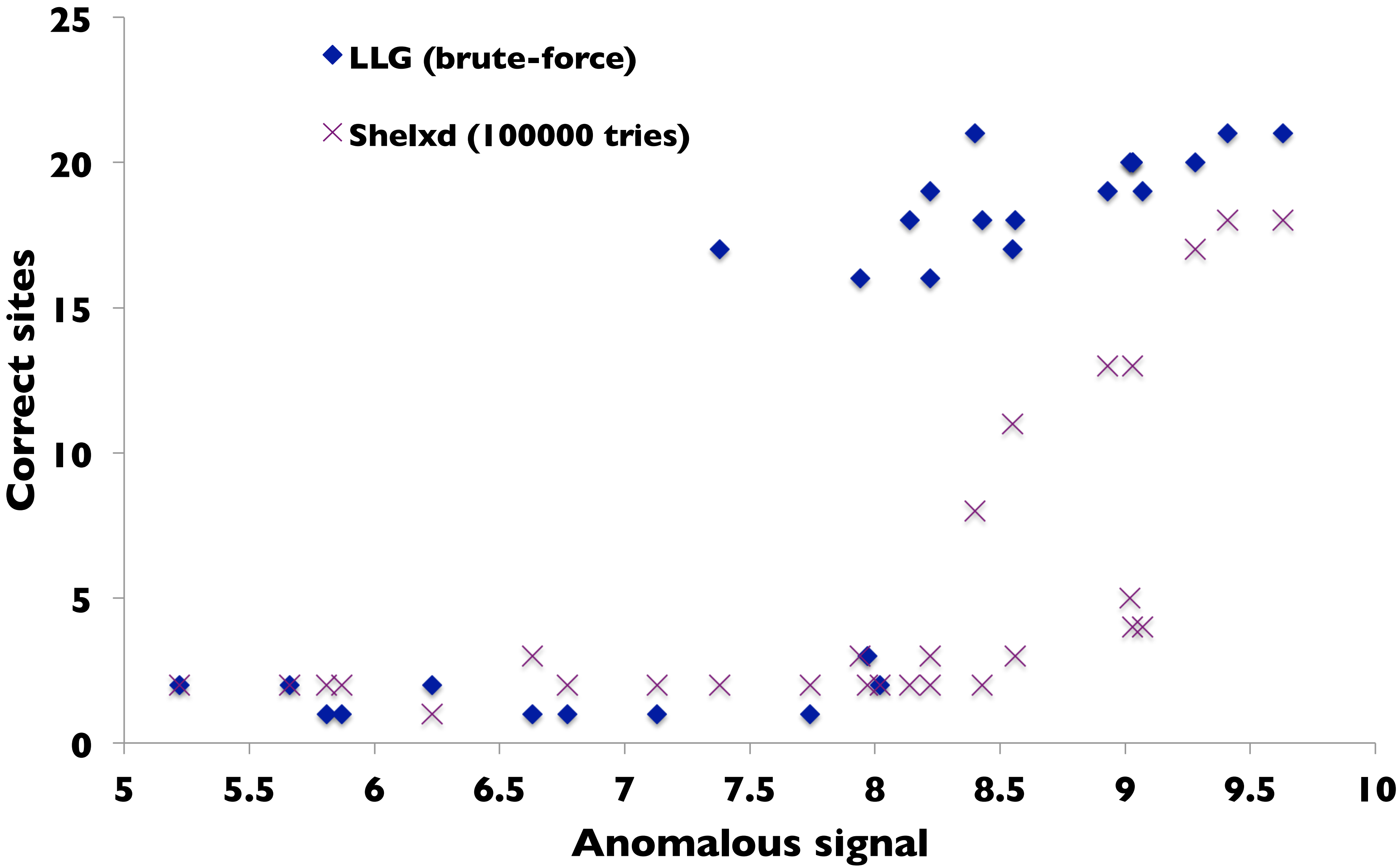
What is the minimum number of crystals that could have been used?

CysZ multi-crystal sulfur-SAD data

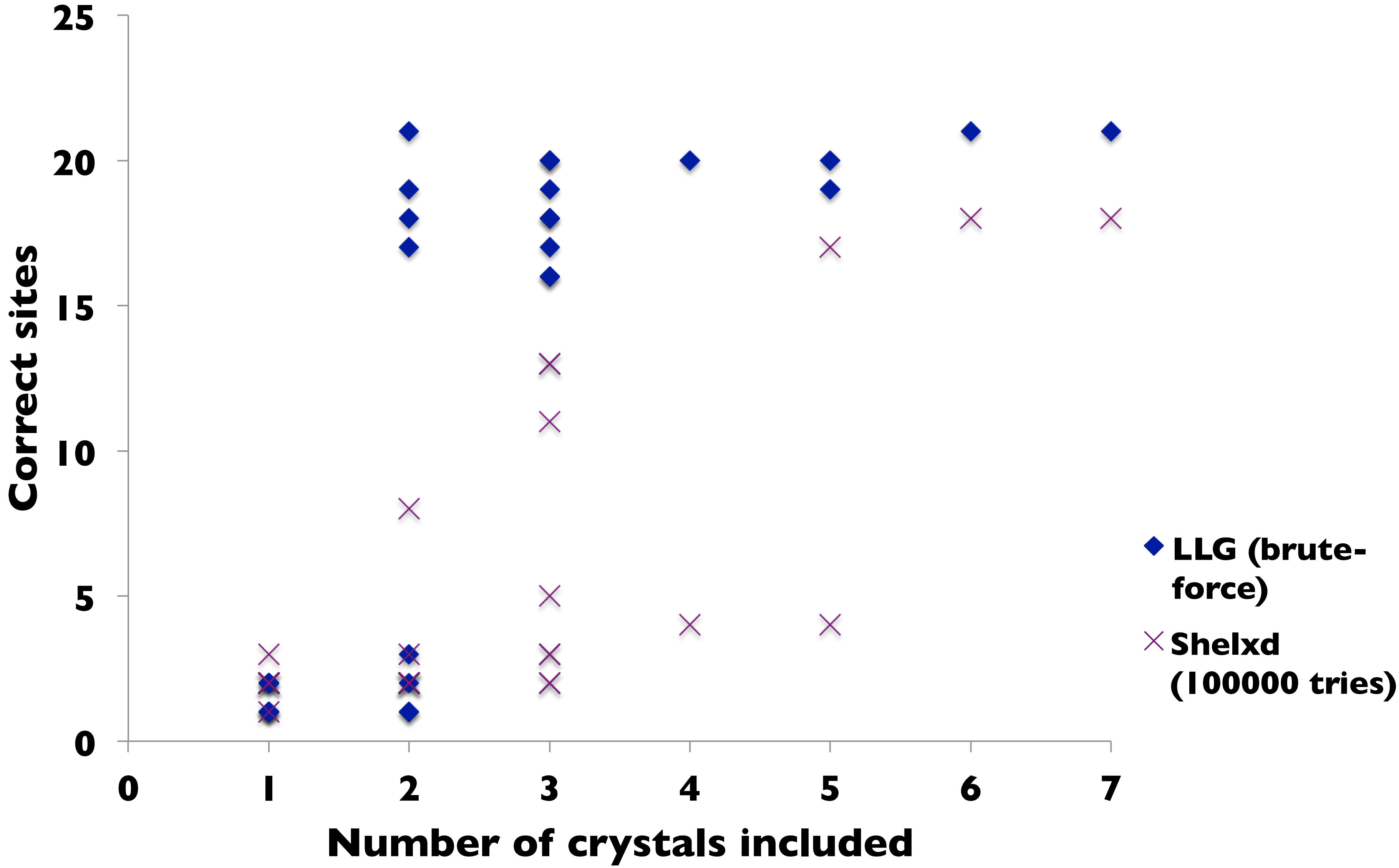
Phenix

Datasets	Anomalous signal
5	5.22
1	5.66
4	5.81
2	5.87
6	6.23
7	6.63
3	6.77
56	7.13
561	7.94
67	8.22
273	9.02
2734	9.03
27345	9.07
27346	9.28
273456	9.41
2734561	9.63

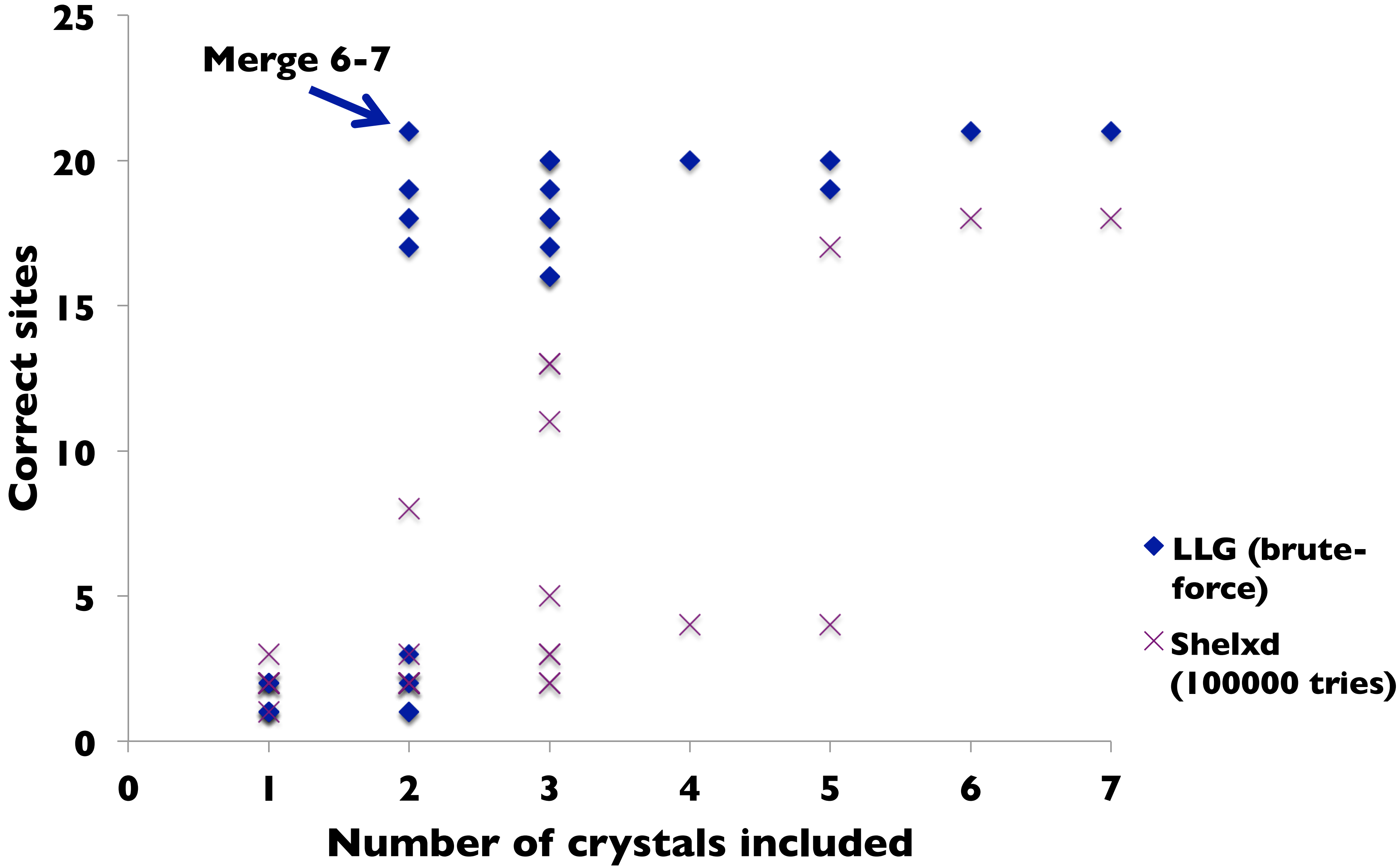
CysZ multi-crystal sulfur-SAD data



CysZ multi-crystal sulfur-SAD data



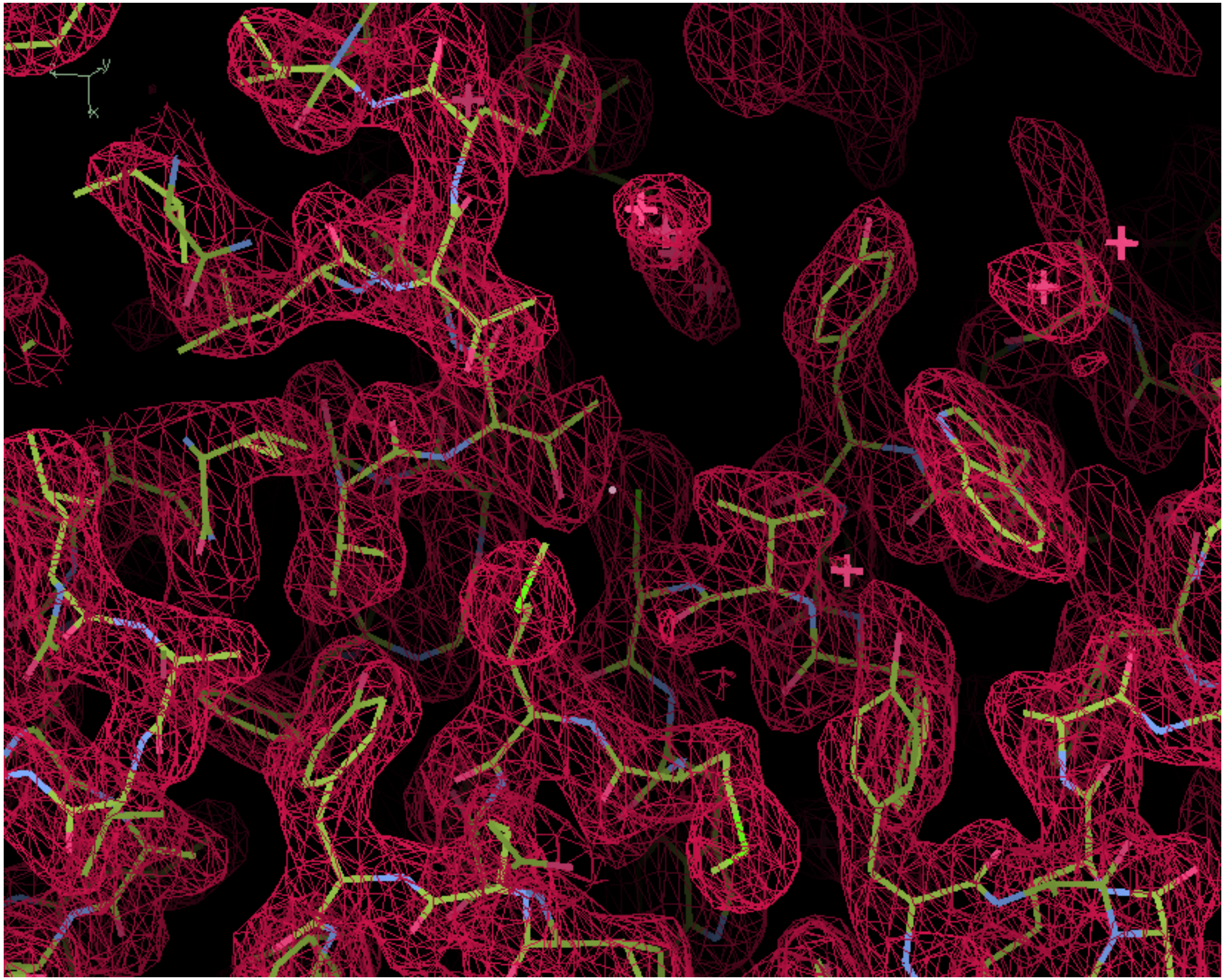
CysZ multi-crystal sulfur-SAD data



CysZ multi-crystal sulfur-SAD data

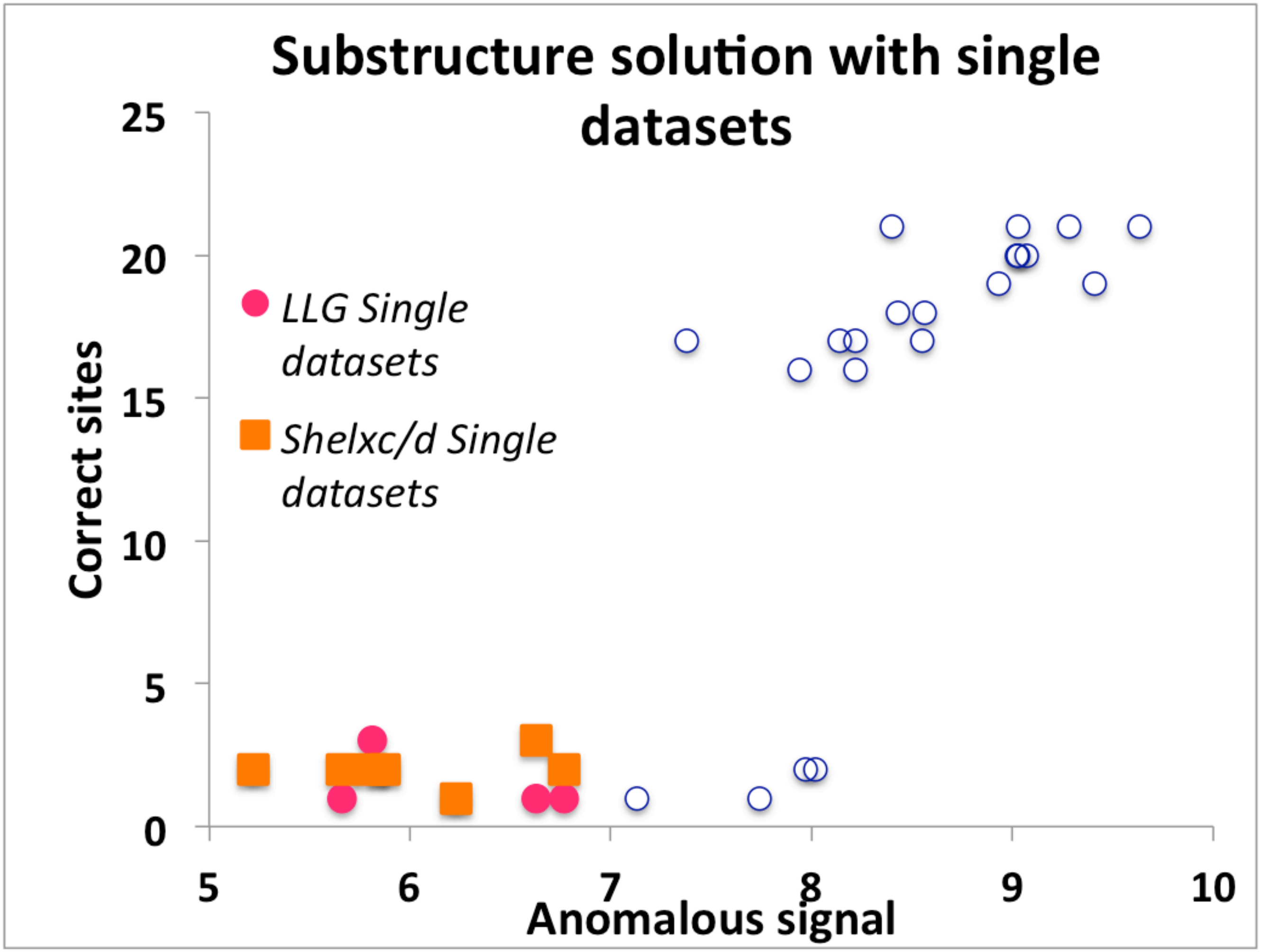
Merge of crystals 6, 7

AutoSol/Autobuild R/Rfree=0.22/0.26



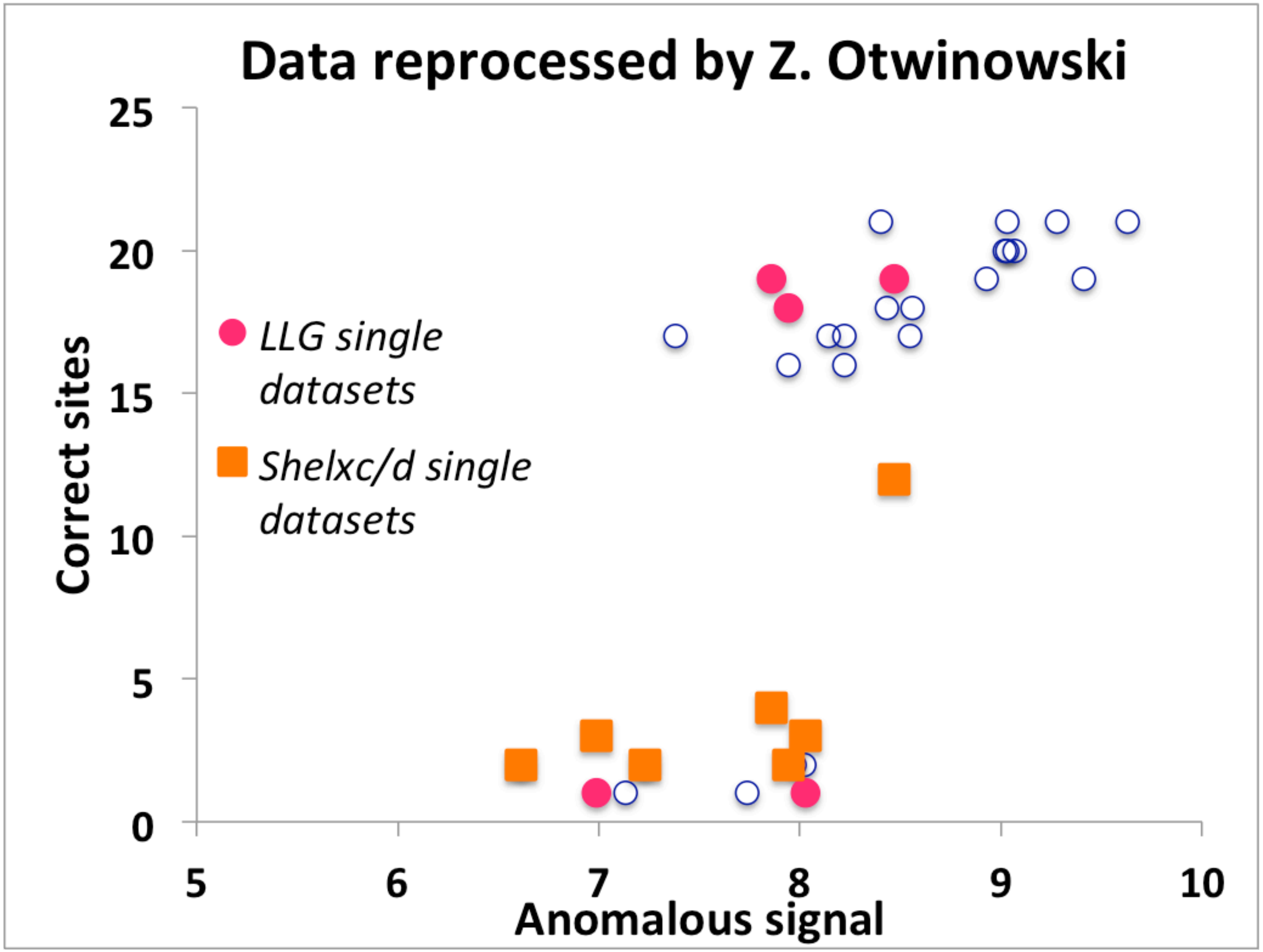
CysZ multi-crystal sulfur-SAD data

Phenix



CysZ multi-crystal sulfur-SAD data

(The minimum number of datasets for this structure is 1)



Structure determination with weak anomalous signal

AutoSol:

Substructure solution, phasing, density modification, preliminary model-building

AutoBuild

Iterative model-building, refinement, density modification

Parallel AutoBuild

Parallel runs of AutoBuild with map averaging and picking best models

Structure solution with phenix.autosol

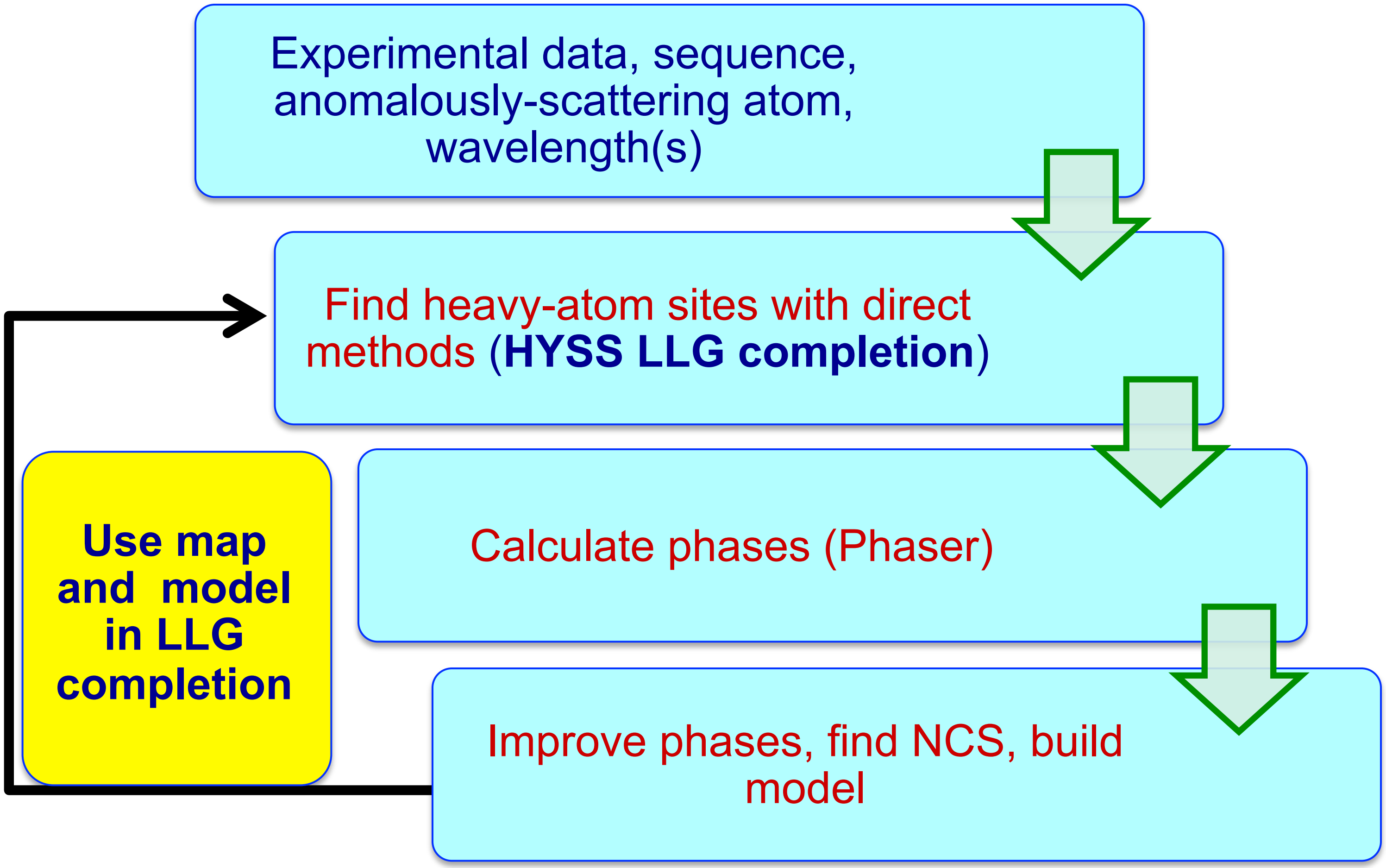
Experimental data, sequence, anomalously-scattering atom, wavelength(s)

Find heavy-atom sites with direct methods (HYSS dual-space)

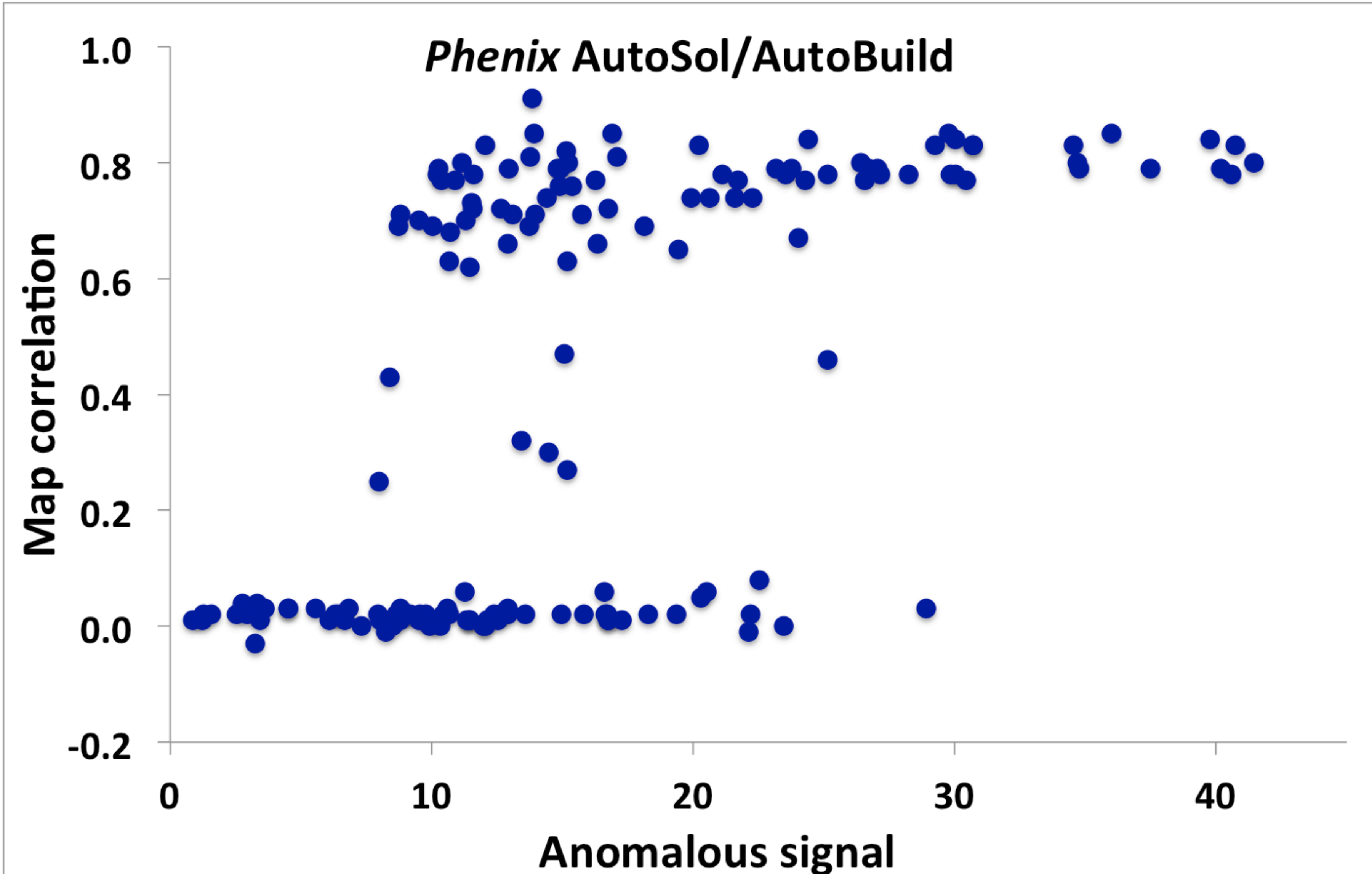
Calculate phases (Phaser)

Improve phases, find NCS, build model

Structure solution with phenix.autosol: enhancements for weak SAD data

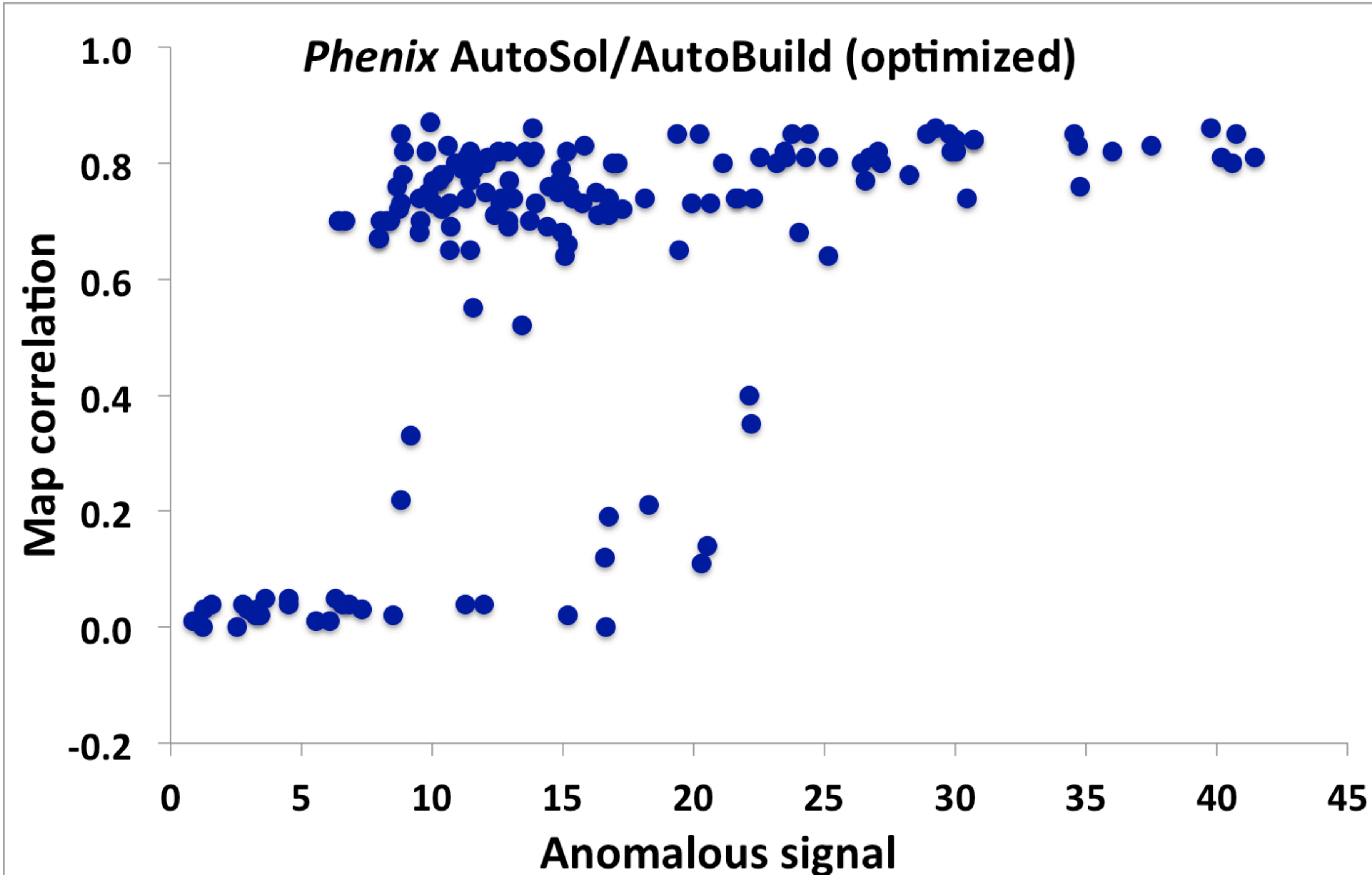


AutoSol structure solution
164 SAD datasets from PDB
(including inflection/remote datasets not previously used as SAD data)

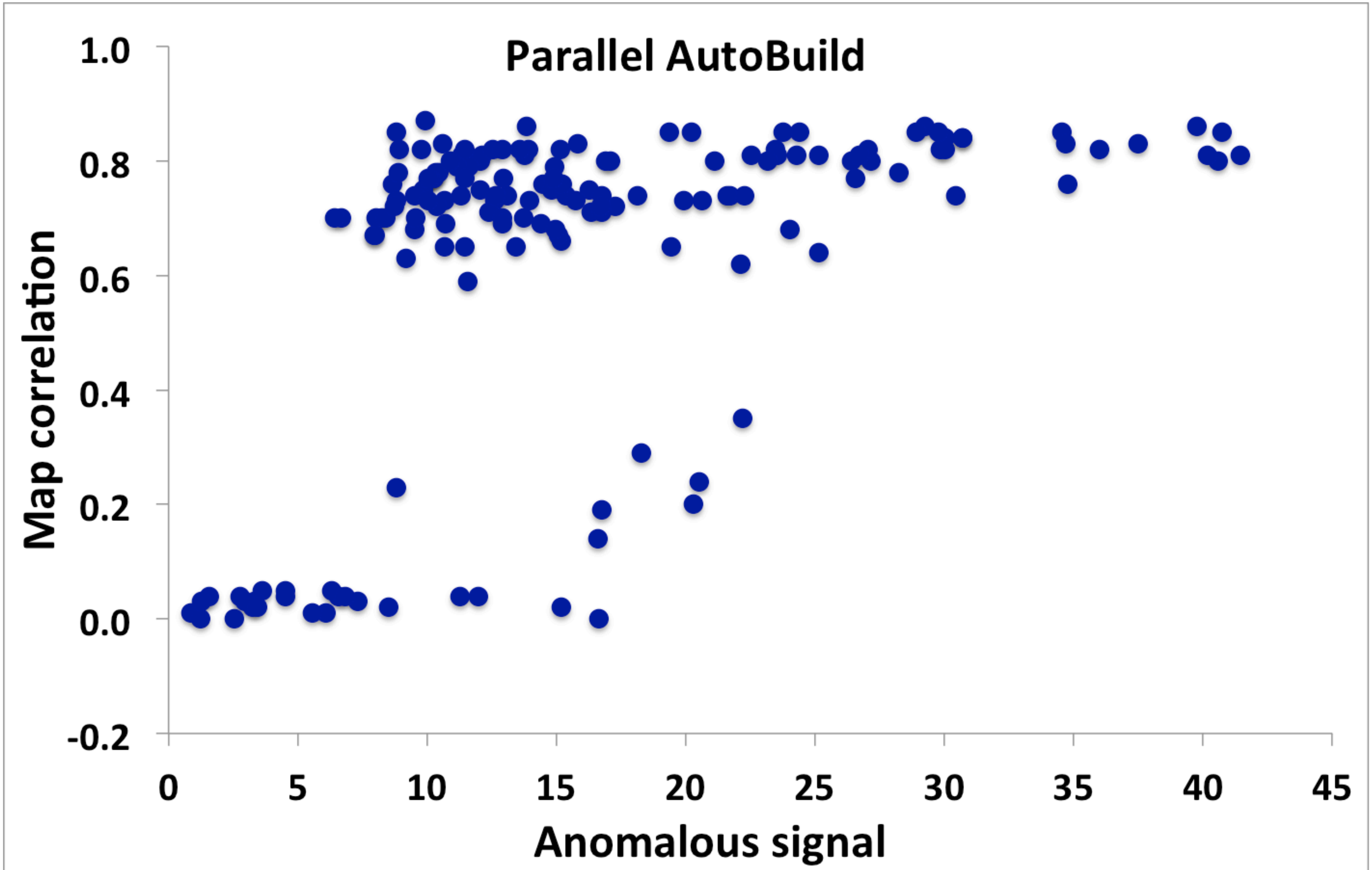


AutoSol structure solution
164 SAD datasets from PDB

Phenix



AutoBuild model-building
164 SAD datasets from PDB

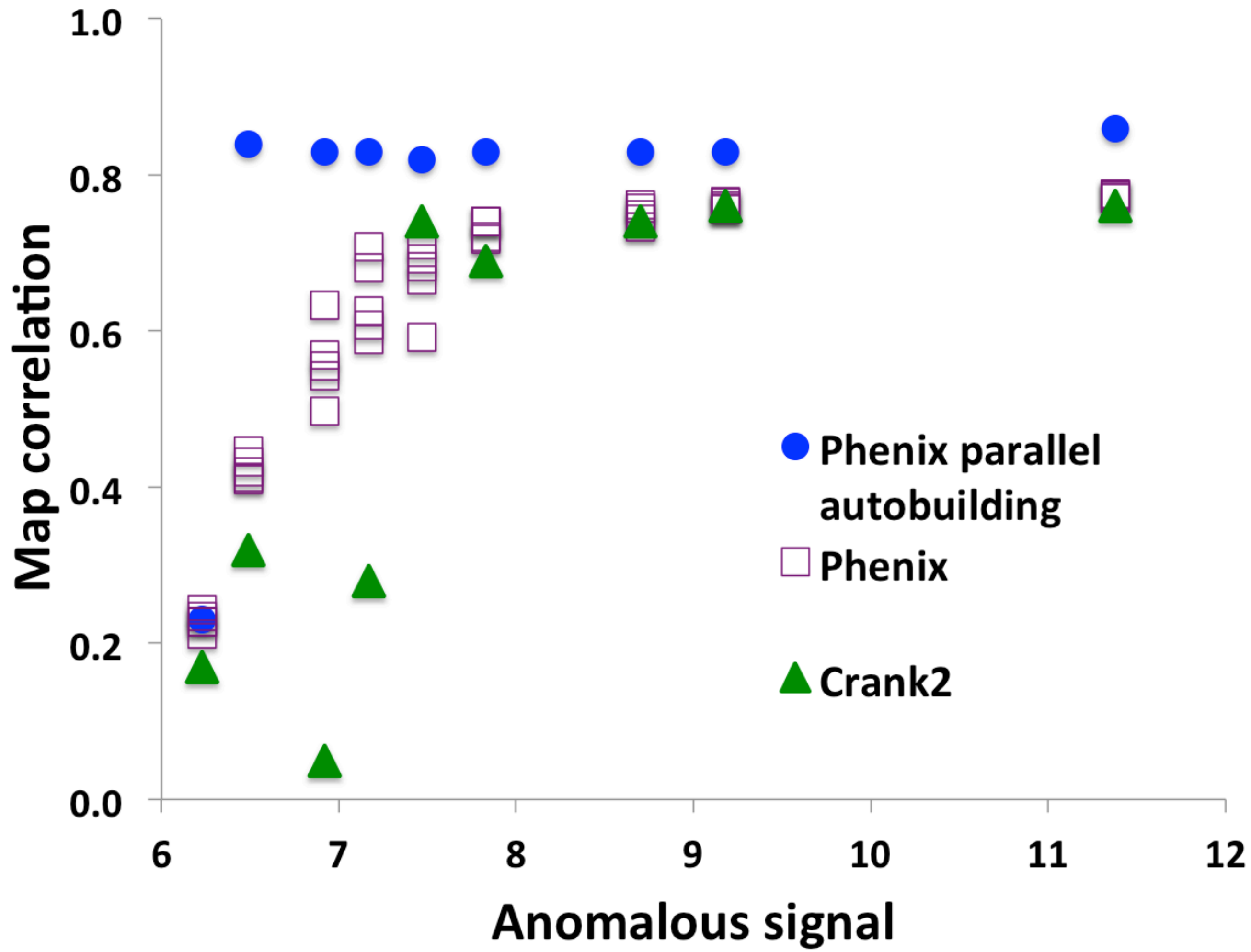


Holton Challenge data

Starting point: *known sites.*

Calculate phases, carry out iterative density modification, model-building and refinement.

Final map correlation vs anomalous signal-to-noise.



Phenix

Estimating the anomalous signal from the data

Gold standards for the anomalous information:

Correlation of true and observed differences:

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

Peak height in model-phased Difference Fourier:

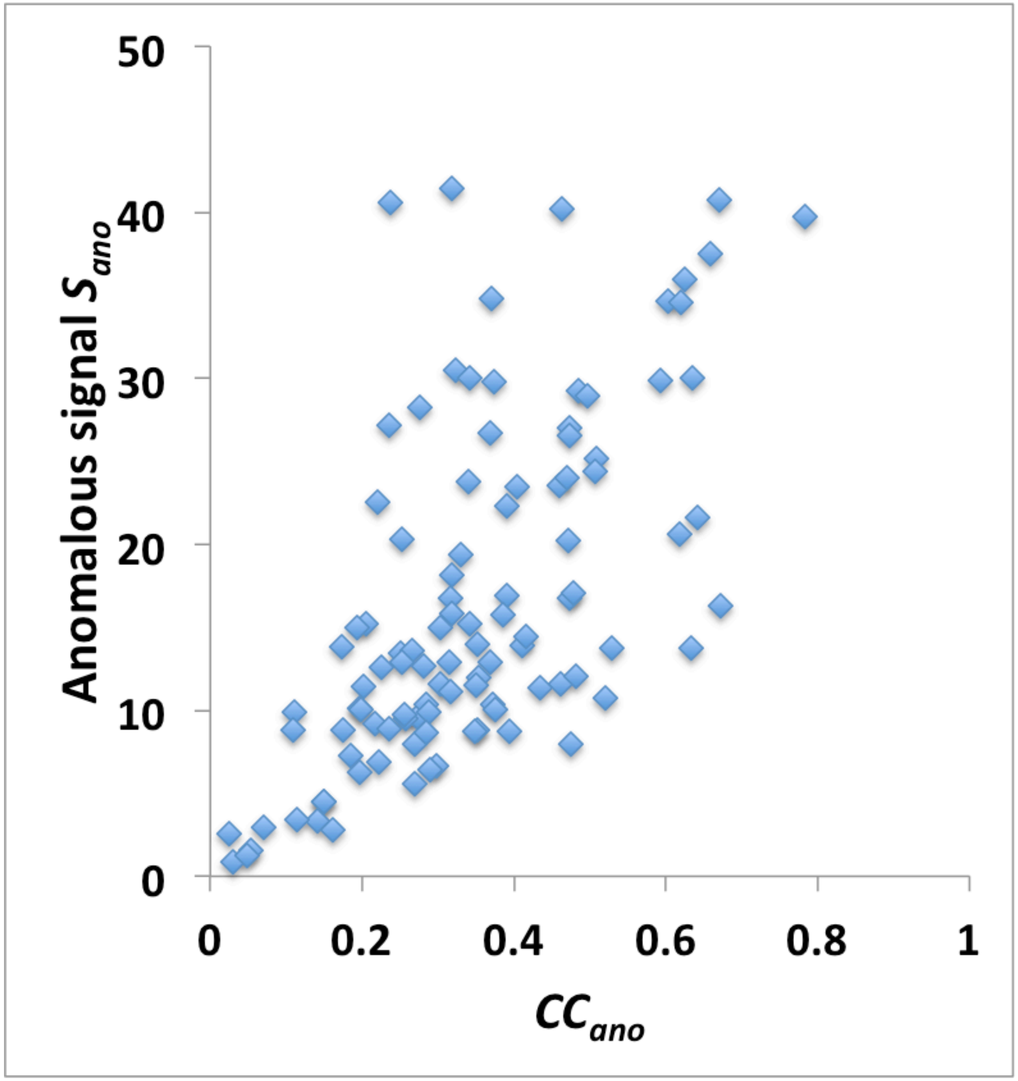
$$S_{ano} \equiv \frac{<\rho(x_k)>}{<\rho^2>^{1/2}}$$

Relationship between CC_{ano} *and* S_{ano}

$$S_{ano} \sim CC_{ano} \frac{N_{refl}^{1/2}}{N_{sites}^{1/2}(\frac{5}{4})^{1/2}}$$

Phenix

Checking the relationship between CC_{ano} and S_{ano}



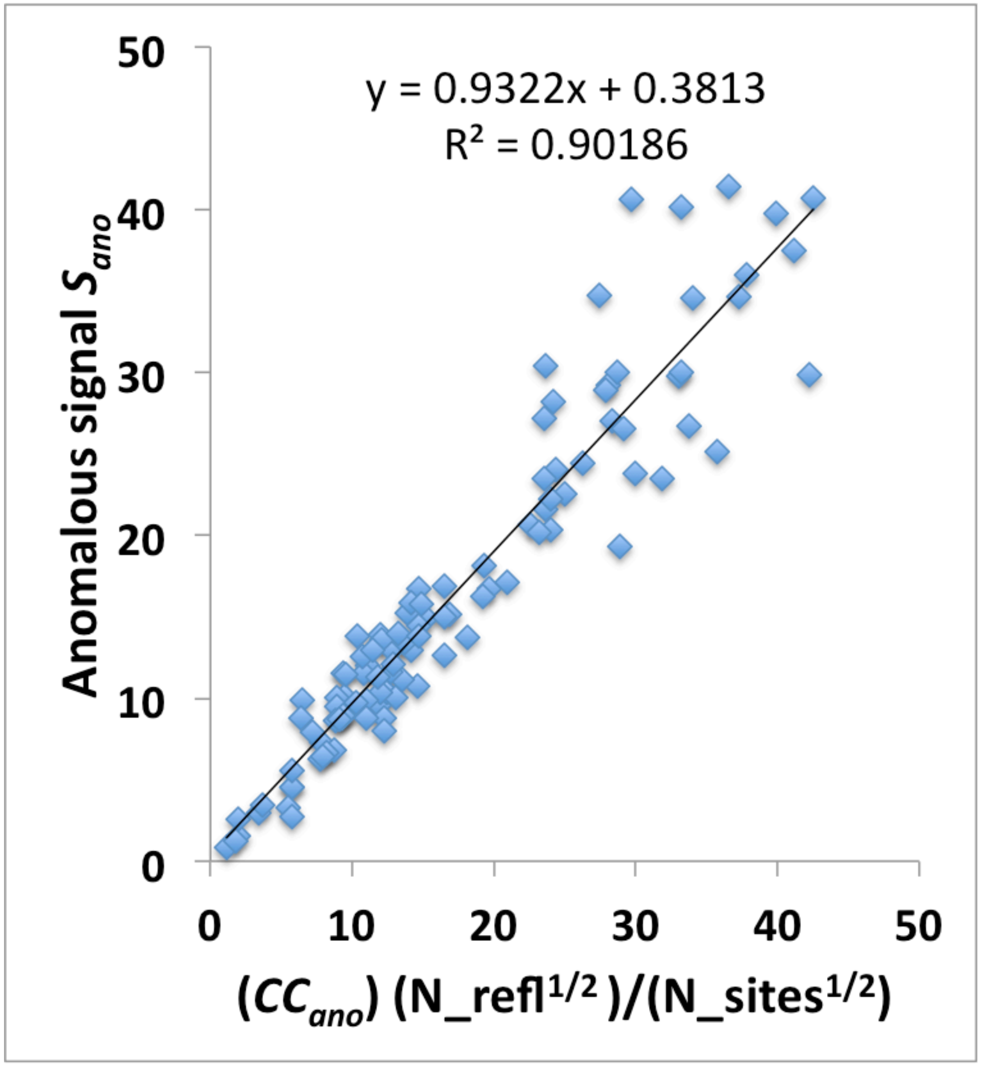
$$S_{ano} \sim CC_{ano} \frac{N_{refl}^{1/2}}{N_{sites}^{1/2}\left(\frac{5}{4}\right)^{1/2}}$$

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

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

Phenix

Checking the relationship between CC_{ano} and S_{ano}



$$S_{ano} \sim CC_{ano} \frac{N_{refl}^{1/2}}{N_{sites}^{1/2} \left(\frac{5}{4}\right)^{1/2}}$$

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

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

Estimating the anomalous correlation CC_{ano} from the data

Phenix

CC_{ano} estimates based on simple theory:

Estimated from experimental uncertainties and anomalous differences

$$E_{ano}^2 = \frac{<\sigma_{ano}^2>}{<\Delta_{ano}^{2,obs}>}$$

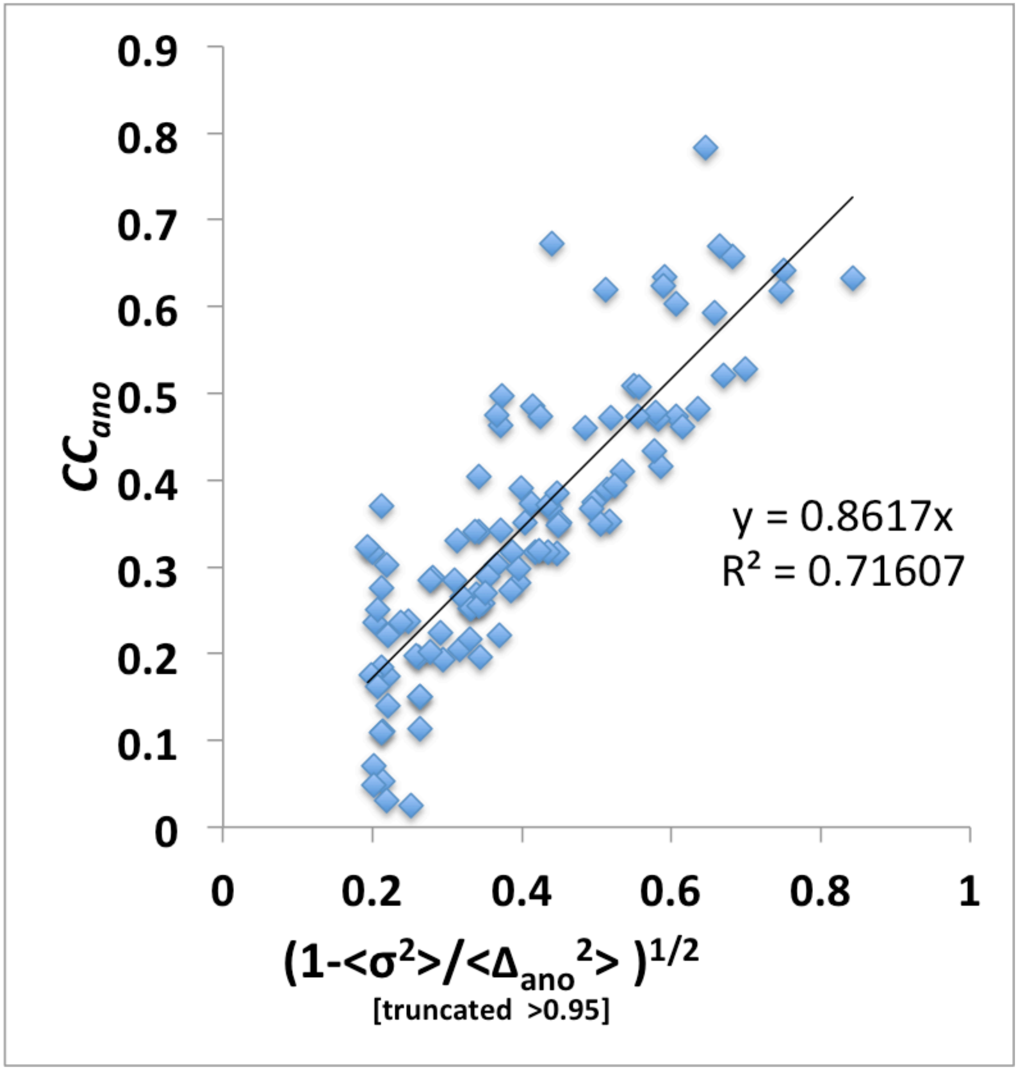
$$CC_{ano} \sim [1 - E_{ano}^2]^{1/2}$$

Estimated from half-dataset correlation of experimental anomalous differences

$$CC_{ano}^* = [\frac{2CC_{ano}^{half\_dataset}}{1 + CC_{ano}^{half\_dataset}}]^{1/2}$$

$$CC_{ano} \sim CC_{ano}^*$$

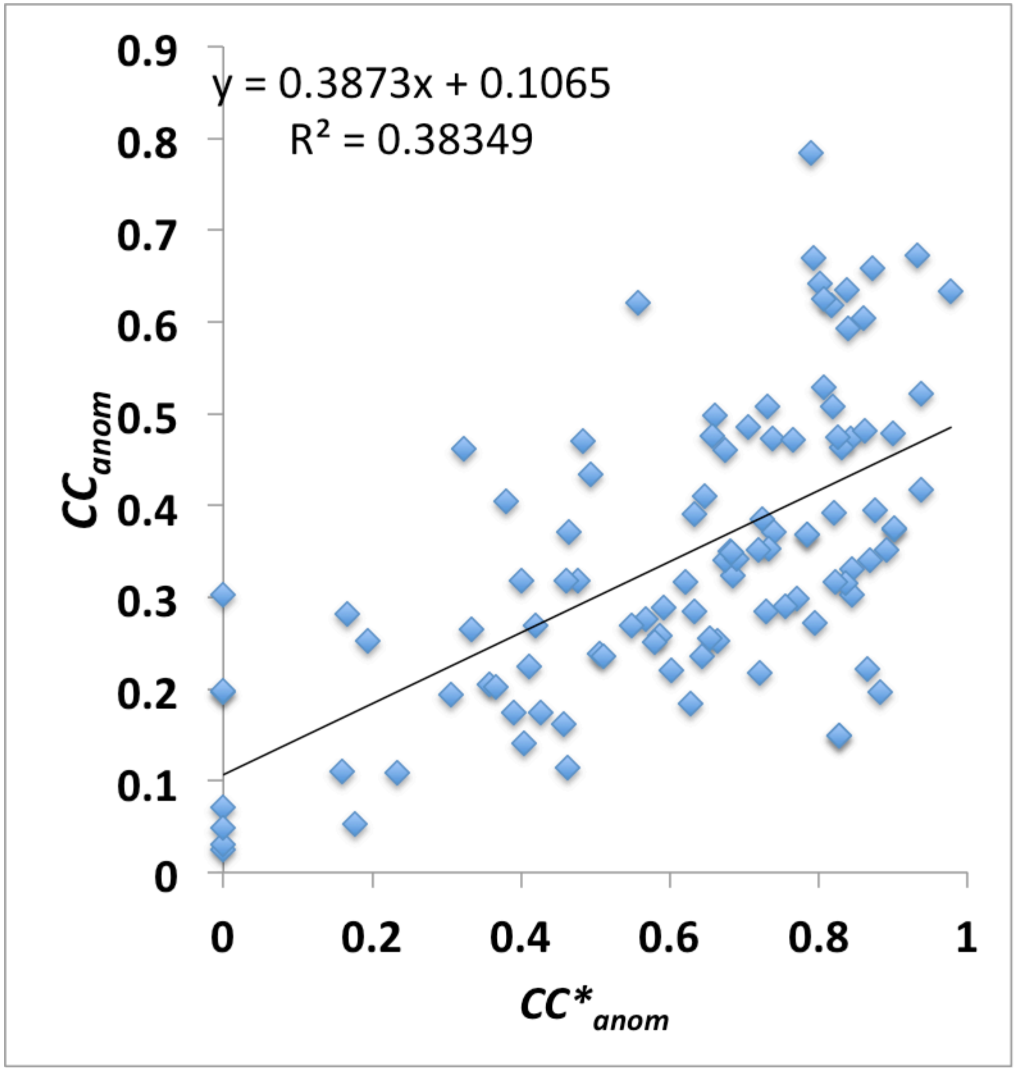
Estimating CC_{ano} from experimental uncertainties and anomalous differences



$$E_{ano}^2 = \frac{<\sigma_{ano}^2>}{<\Delta_{ano}^{2,obs}>}$$

$$CC_{ano} \sim [1 - E_{ano}^2]^{1/2}$$

Estimating CC_{ano} from the half-dataset anomalous correlation.



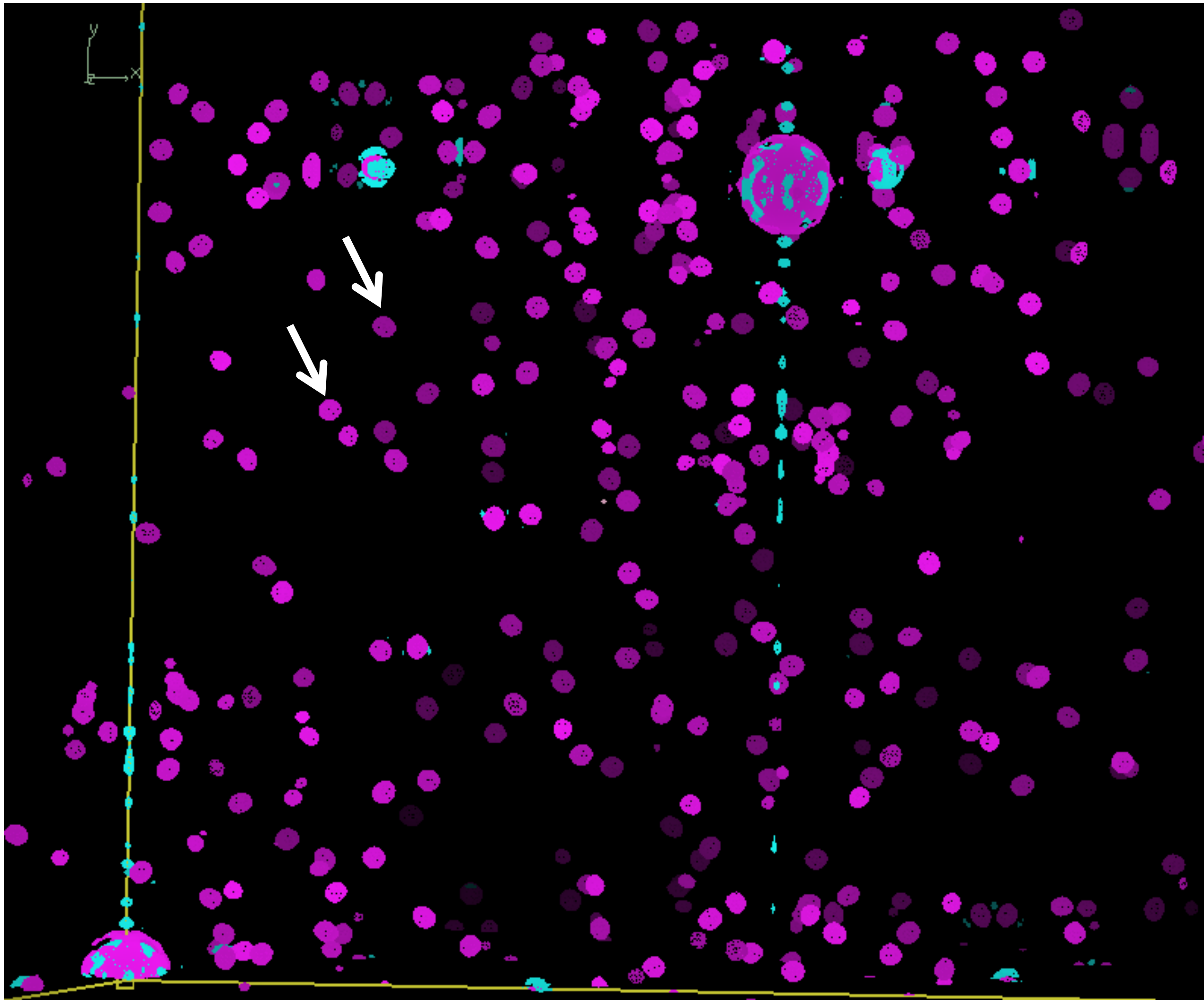
$$CC^{*}_{ano} = \left[\frac{2CC^{half\_dataset}_{ano}}{1+CC^{half\_dataset}_{ano}}\right]^{1/2}$$

$$CC_{ano} \sim CC^{*}_{ano}$$

Skew of anomalous difference Patterson

Anomalous difference Patterson for 2a3n (14 Se sites, 1.3 Å)
Contours at +/-4σ.
Positive pink, negative blue

Model anomalous differences

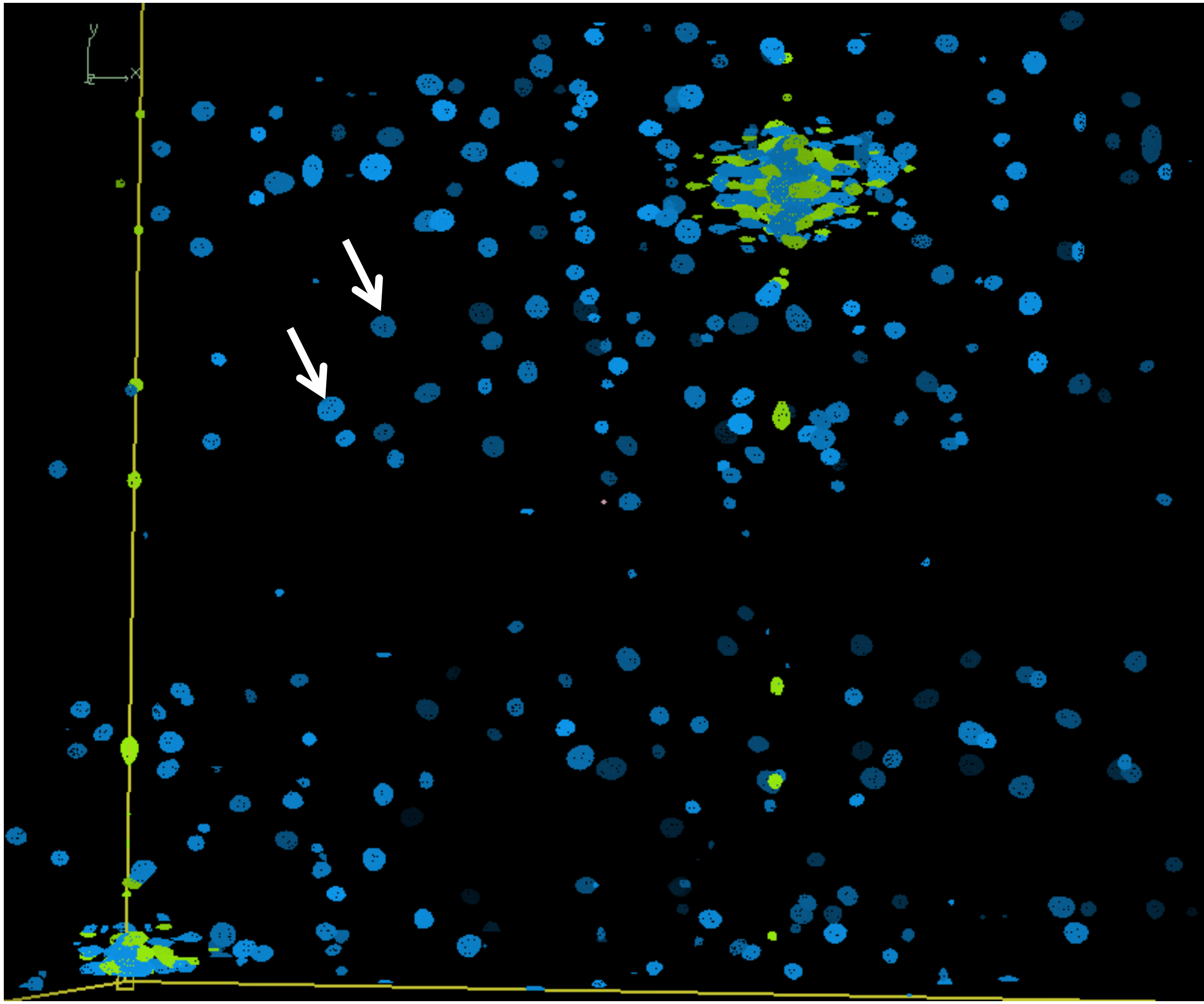


Skew of anomalous difference Patterson

Phenix

Anomalous difference Patterson for 2a3n (14 Se sites, 1.3 Å)
Contours at 4σ.
Positive blue, negative green

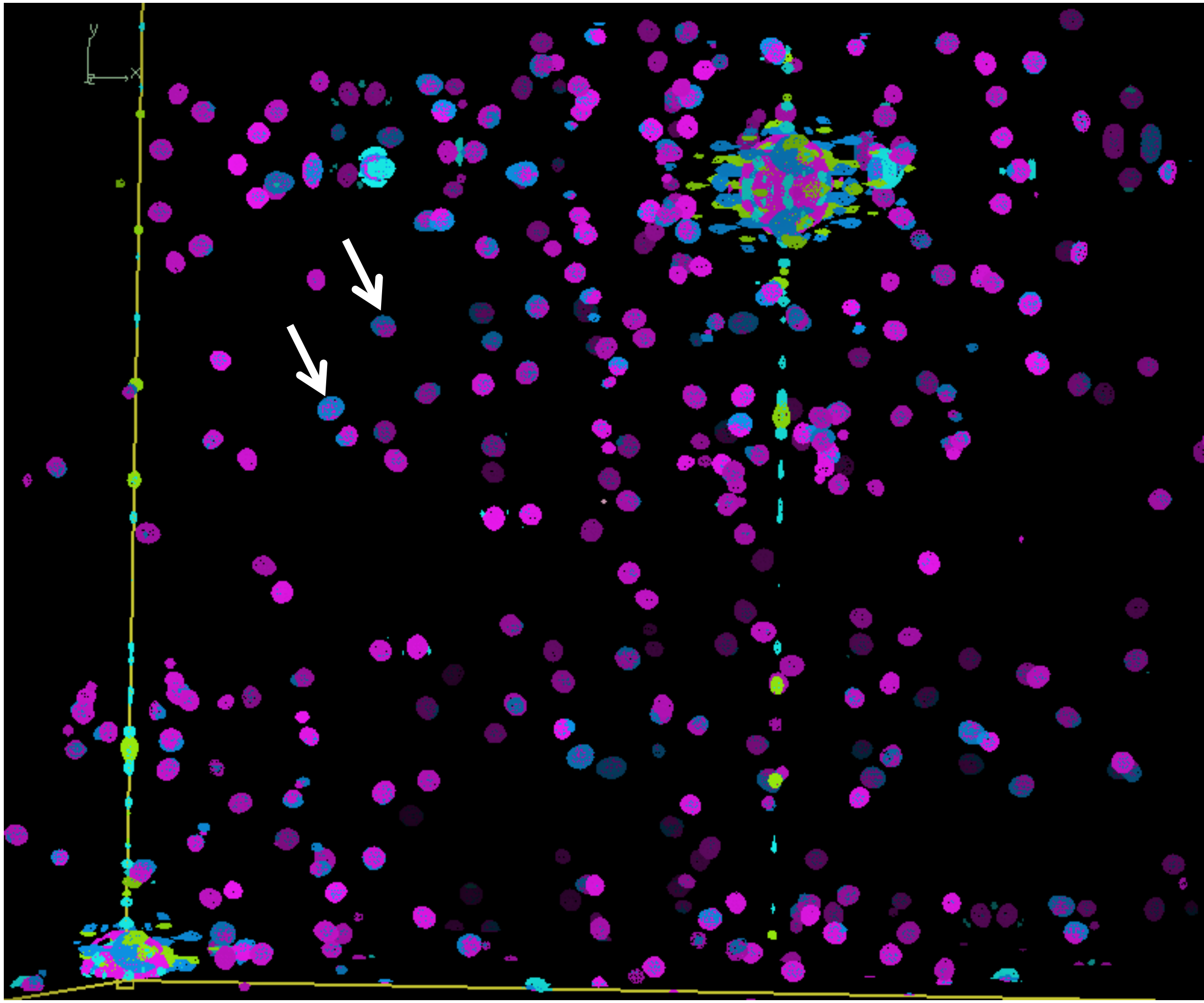
Measured anomalous differences



Skew of anomalous difference Patterson

Anomalous difference Patterson for 2a3n (14 Se sites, 1.3 Å)
Contours at 4σ.

Model (pink) and experimental (blue) anomalous differences

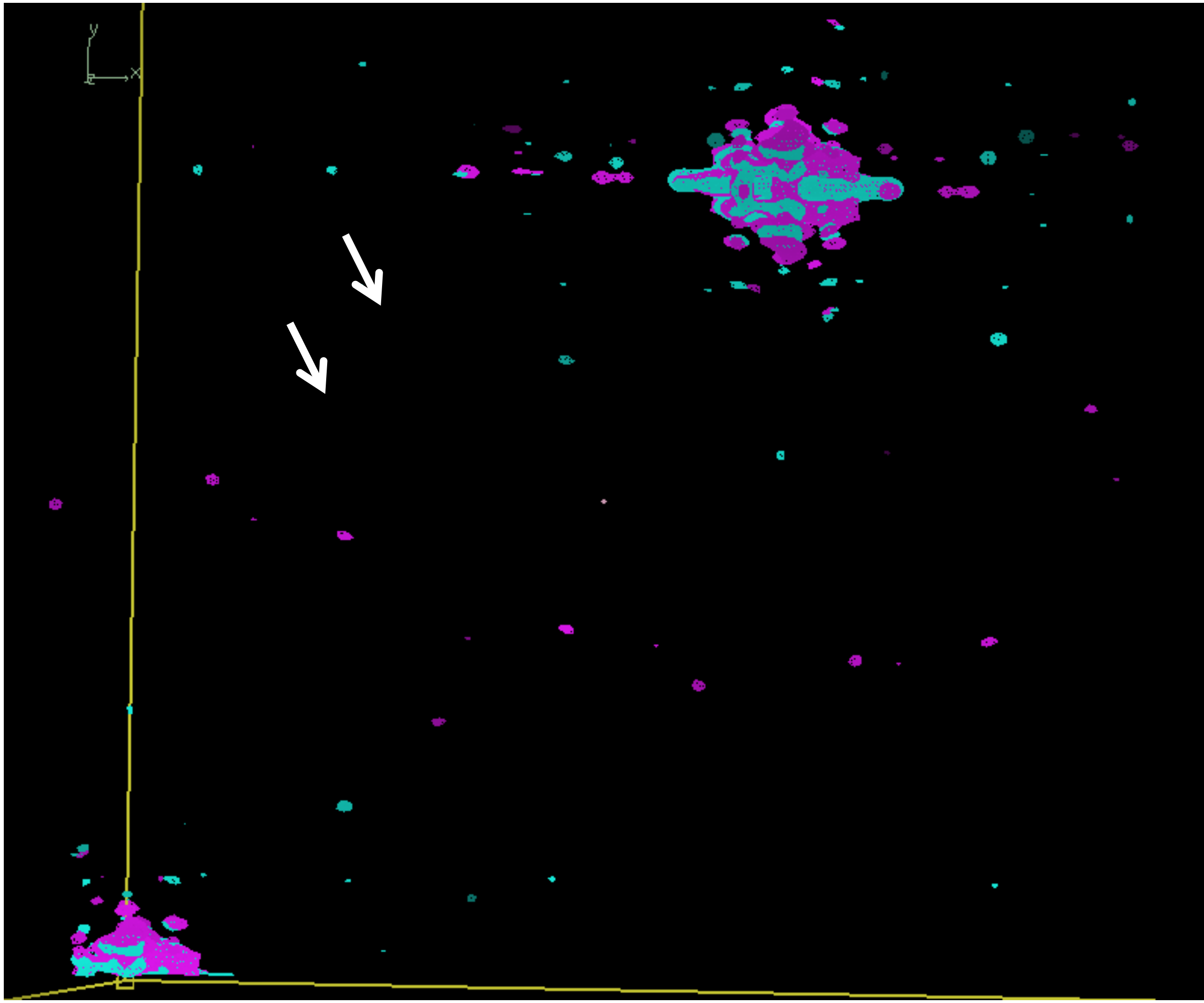


Phenix

Skew of anomalous difference Patterson

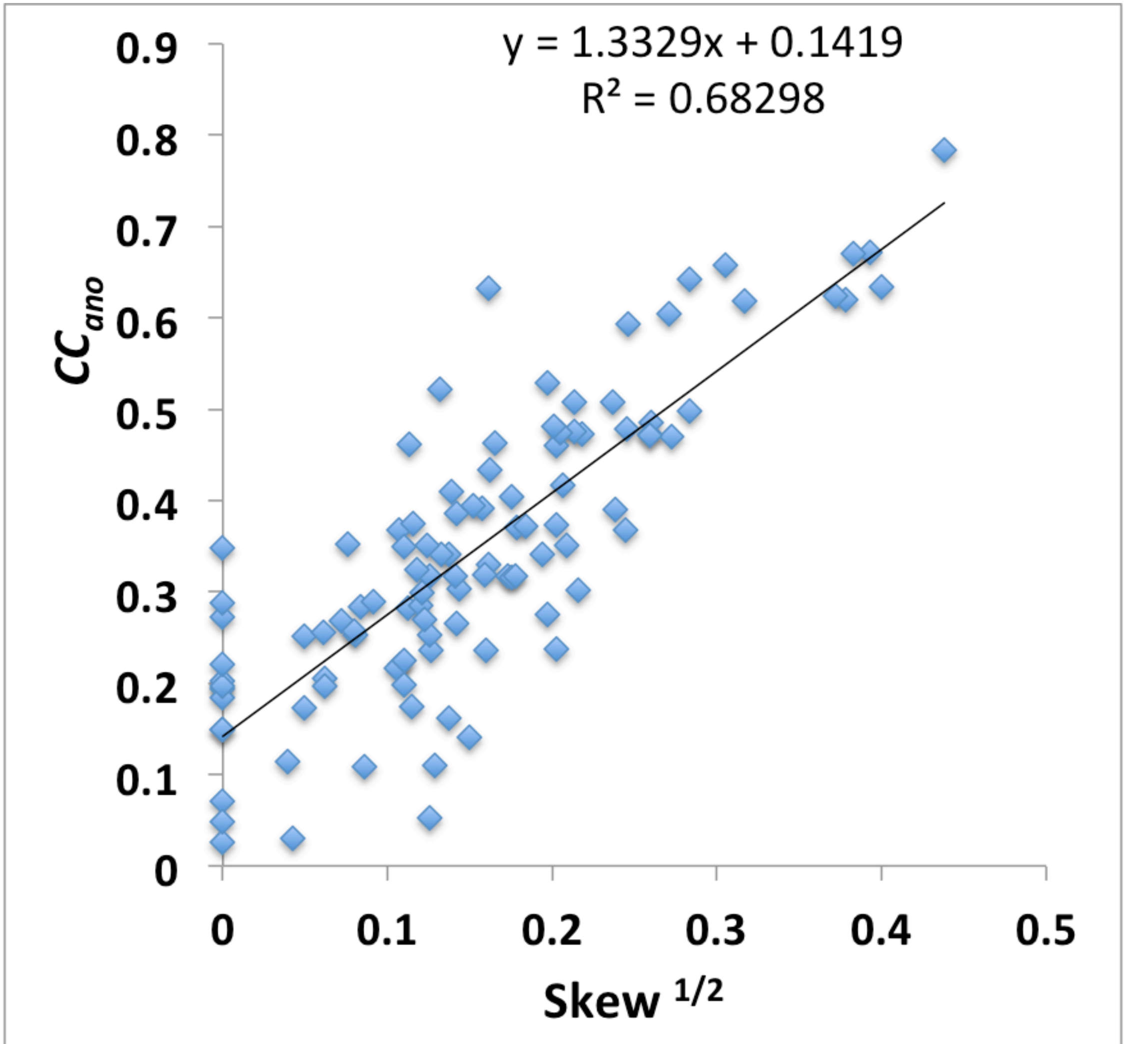
Anomalous difference Patterson for 2a3n (14 Se sites, 1.3 Å)
Contours at 4σ.
Positive blue, negative pink.

Randomized anomalous differences



Estimating CC_{ano} from skew of the anomalous difference Patterson

Phenix



$$CC_{ano} \sim skew_{Patterson}^{1/2}$$

Estimating the anomalous correlation CC_{ano}

Phenix

Estimated fraction of observed anomalous differences that is noise

$$E_{ano}^2 = \frac{<\sigma_{ano}^2>}{<\Delta_{ano}^{2,obs}>}$$

$$CC_{ano} \sim [1 - E_{ano}^2]^{1/2}$$

Half-dataset CC of anomalous differences

$$CC_{ano}^* = [\frac{2CC_{ano}^{half\_dataset}}{1 + CC_{ano}^{half\_dataset}}]^{1/2}$$

$$CC_{ano} \sim CC_{ano}^*$$

Skew of anomalous difference Patterson

$$CC_{ano} \sim skew_{Patterson}^{1/2}$$

Estimating the anomalous signal S_{ano}

Estimation of S_{ano} requires the value of CC_{ano} and the number of sites

$$S_{ano} \sim CC_{ano} \frac{N_{refl}^{1/2}}{N_{sites}^{1/2}(\frac{5}{4})^{1/2}}$$

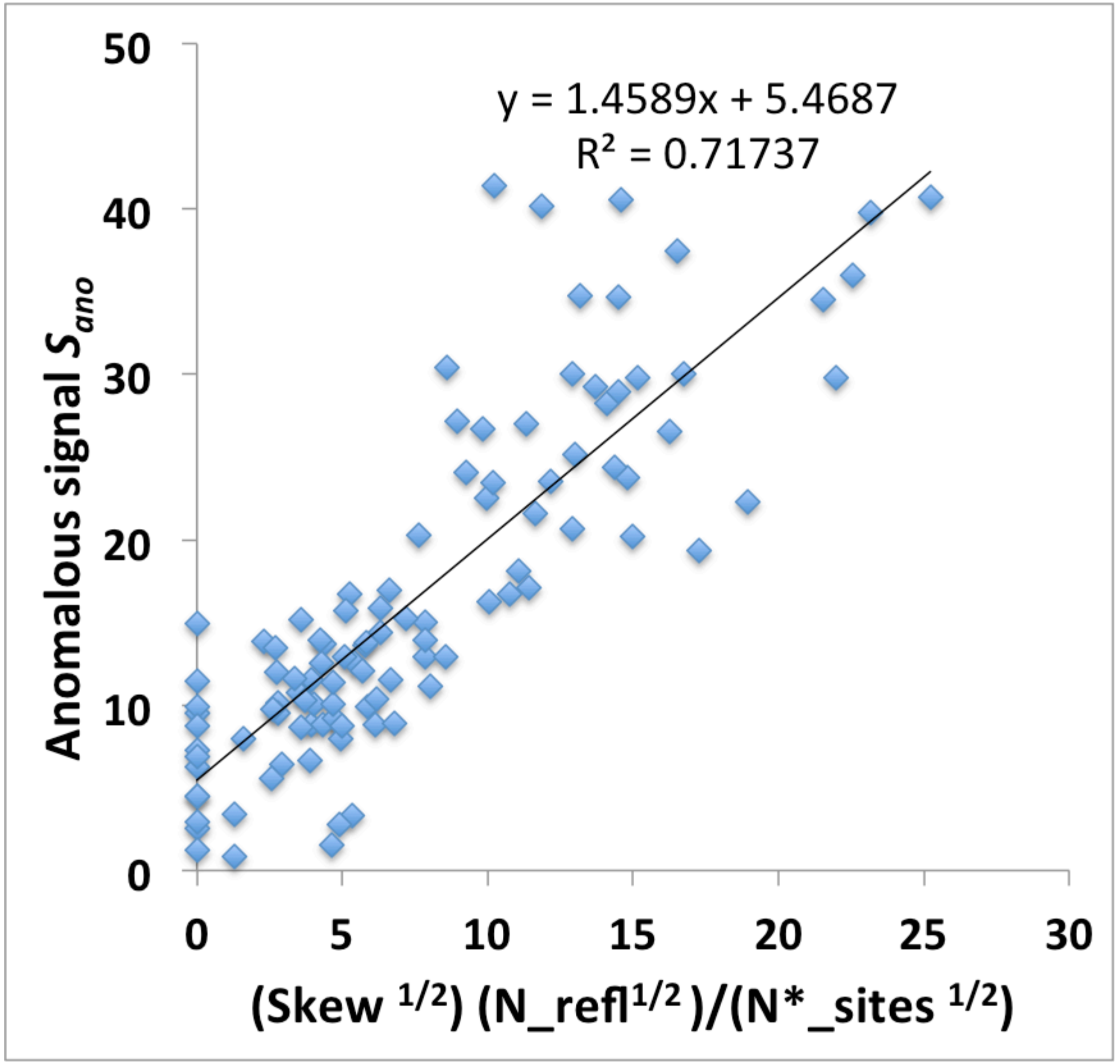
Use *phenix.autosol* estimate of number of sites

Based on sequence, asymmetric unit volume

Guess of number of NCS copies

Guess of number of sites for atoms other than S, Se (typically 1-2 per 100 residues)

Estimating S_{ano} from skew of the anomalous difference Patterson

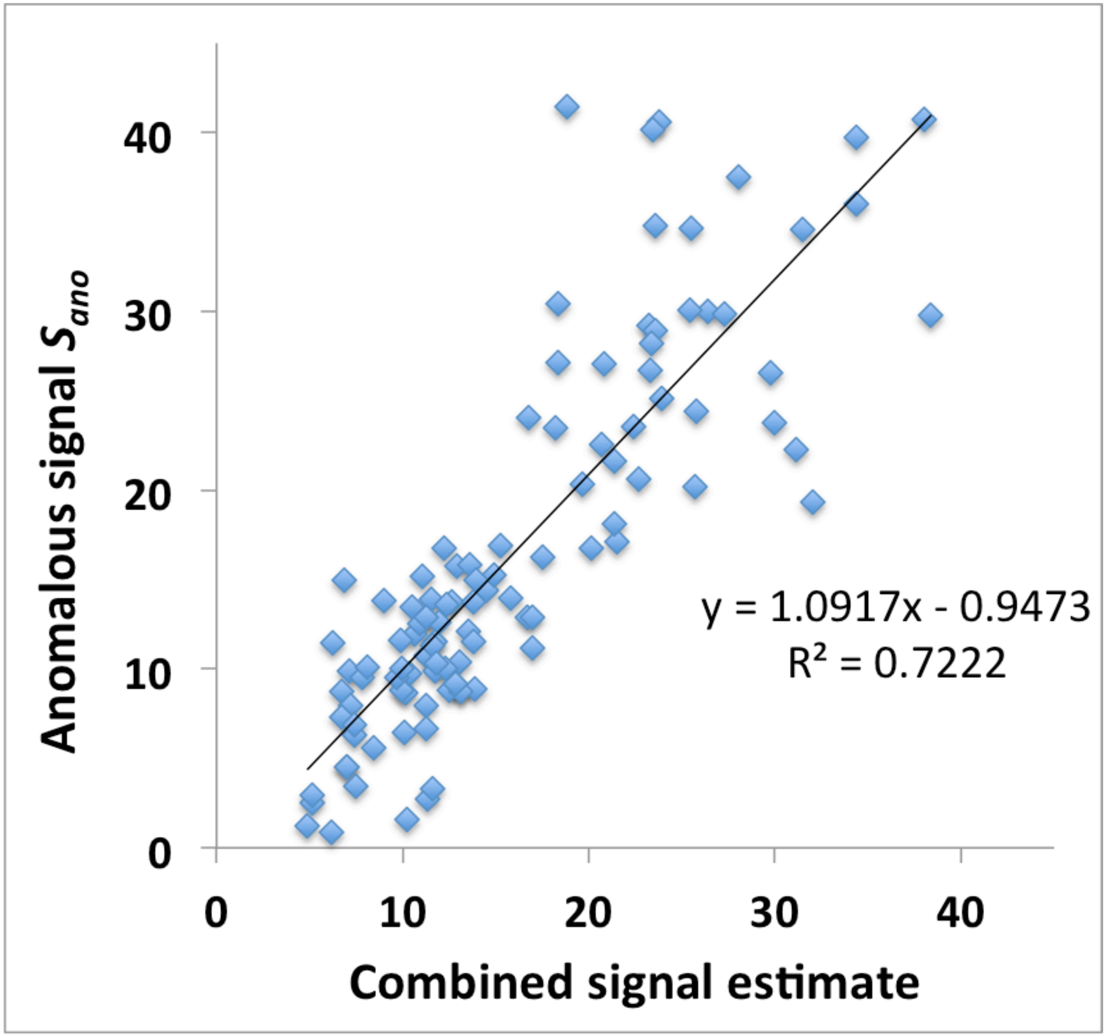


$$CC_{ano} \sim skew^{1/2}_{Patterson}$$

$$S_{ano} \sim CC_{ano} \frac{N^{1/2}_{refl}}{N^{1/2}_{sites}(\frac{5}{4})^{1/2}}$$

Estimating S_{ano} from all 3 measures of anomalous correlation

Phenix



$$CC_{ano} \sim [1 - E_{ano}^2]^{1/2}$$

$$CC_{ano} \sim [\frac{2CC_{ano}^{half\_dataset}}{1 + CC_{ano}^{half\_dataset}}]^{1/2}$$

$$CC_{ano} \sim skew_{Patterson}^{1/2}$$

$$S_{ano} \sim CC_{ano} \frac{N_{refl}^{1/2}}{N_{sites}^{1/2} (\frac{5}{4})^{1/2}}$$

Using the anomalous signal S_{ano} and correlation CC_{ano}

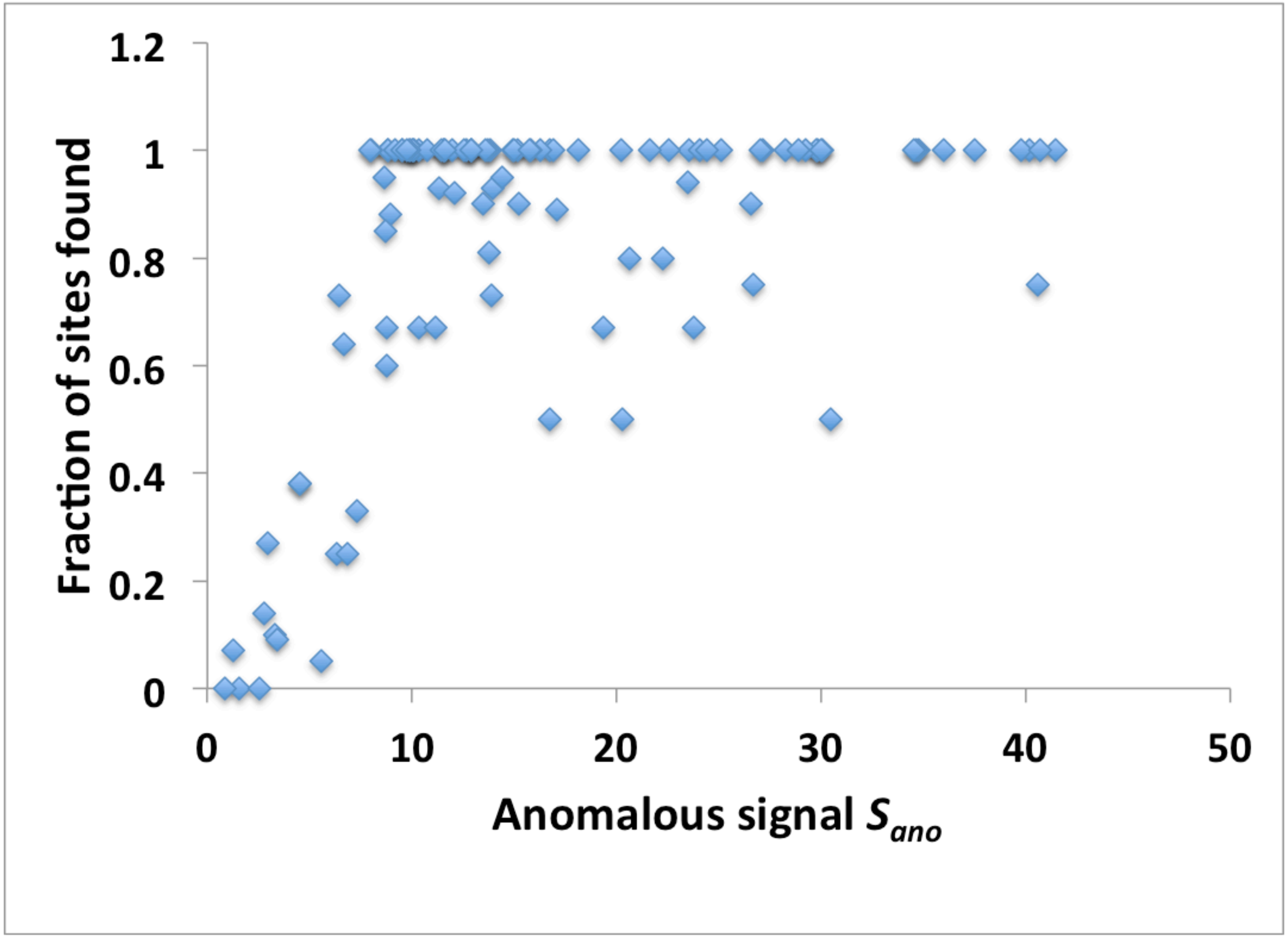
What do we expect:

Finding sites may be most closely related to map quality (S_{ano})

Experimental phase quality may be most closely related to the accuracy of the anomalous differences (CC_{ano})

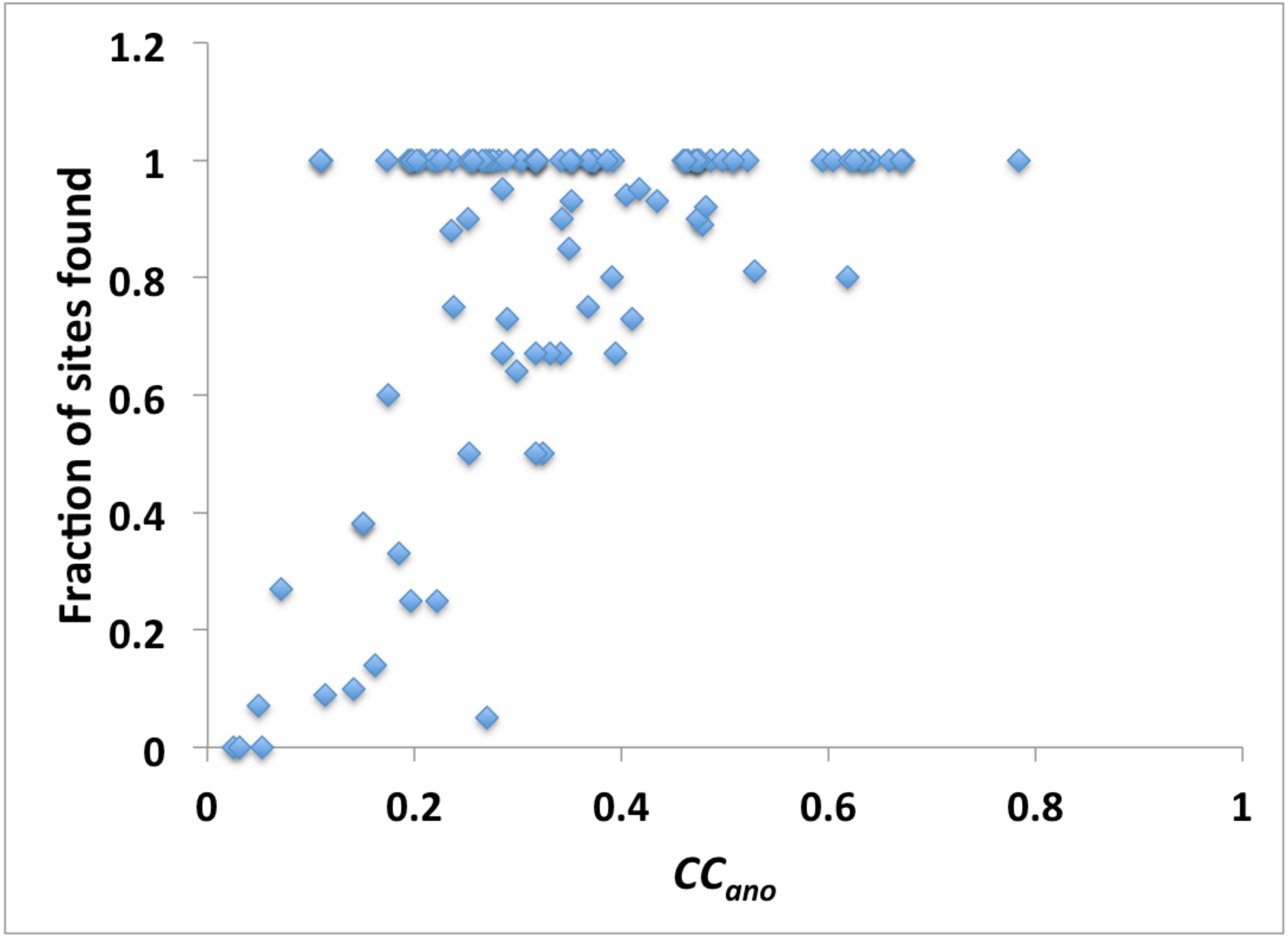
Can I find the substructure: Using the anomalous signal S_{ano} to guess

Best possible case: using known signal S_{ano}



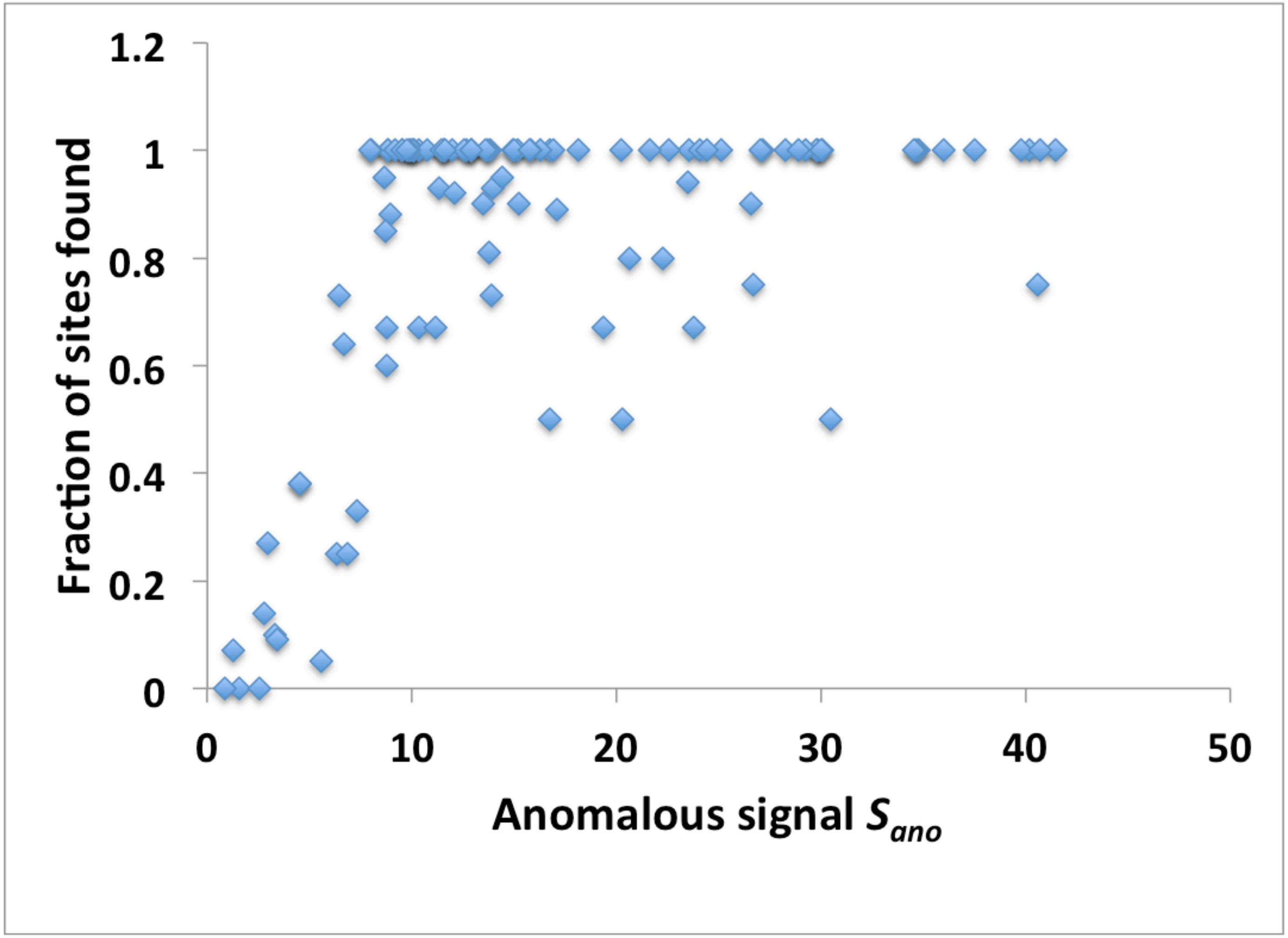
Can I find the substructure: Using the anomalous correlation CC_{ano}

Best possible case: using true CC_{ano}



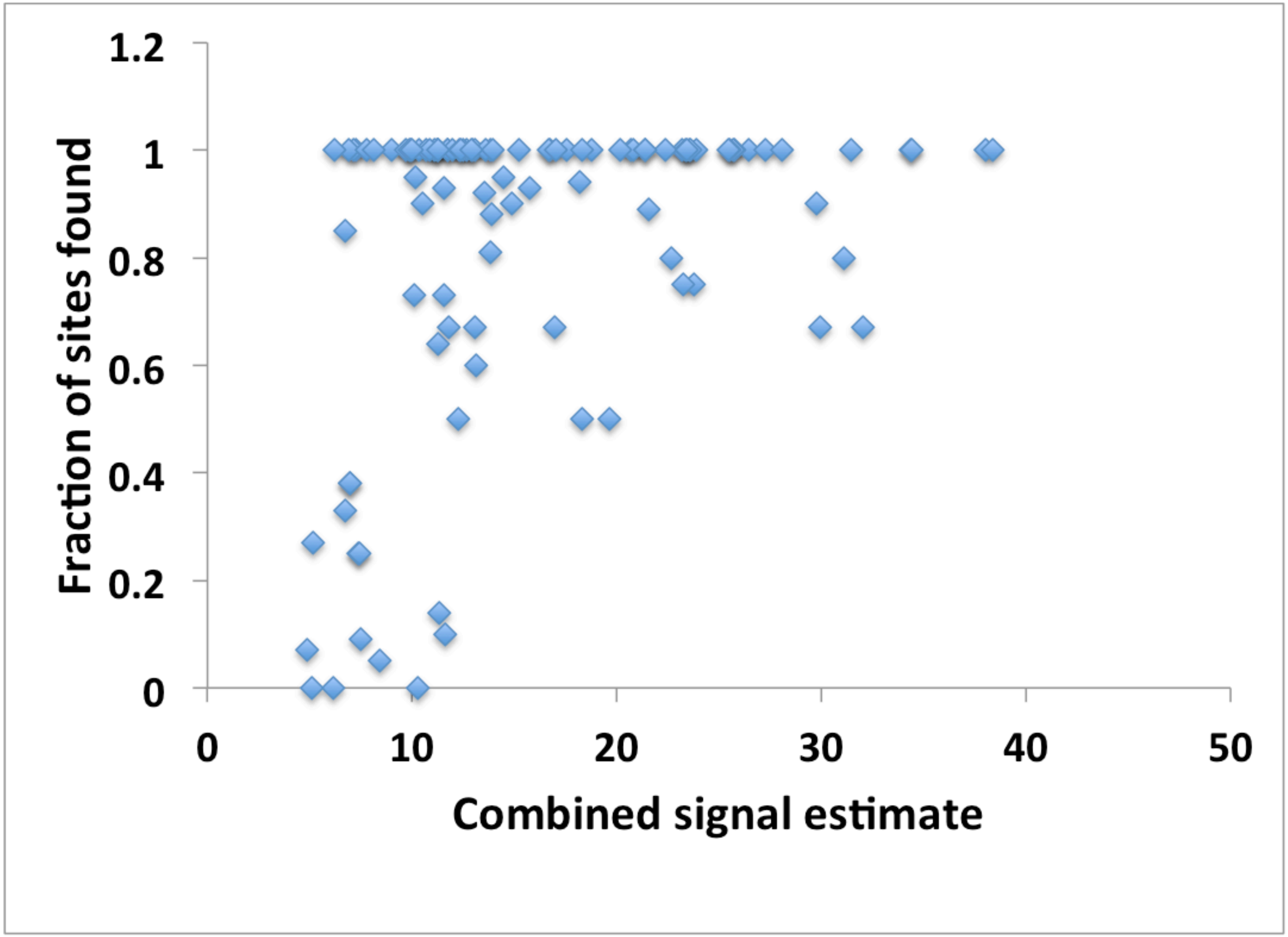
Can I find the substructure: Using the anomalous signal S_{ano} to guess

Best possible case: using known signal S_{ano}

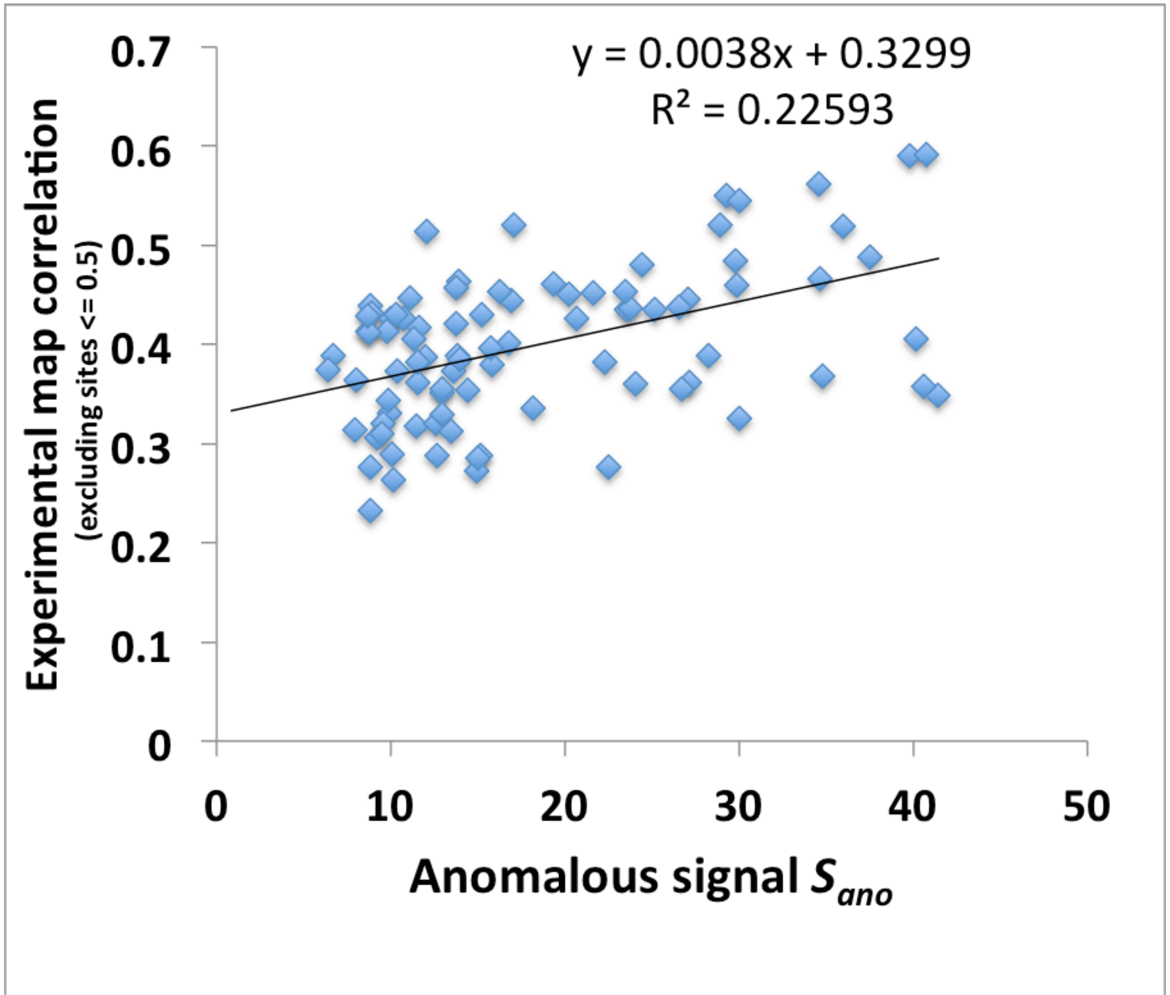


Can I find the substructure: Using the anomalous signal S_{ano} to guess

Real-world case: S_{ano} estimated from the data



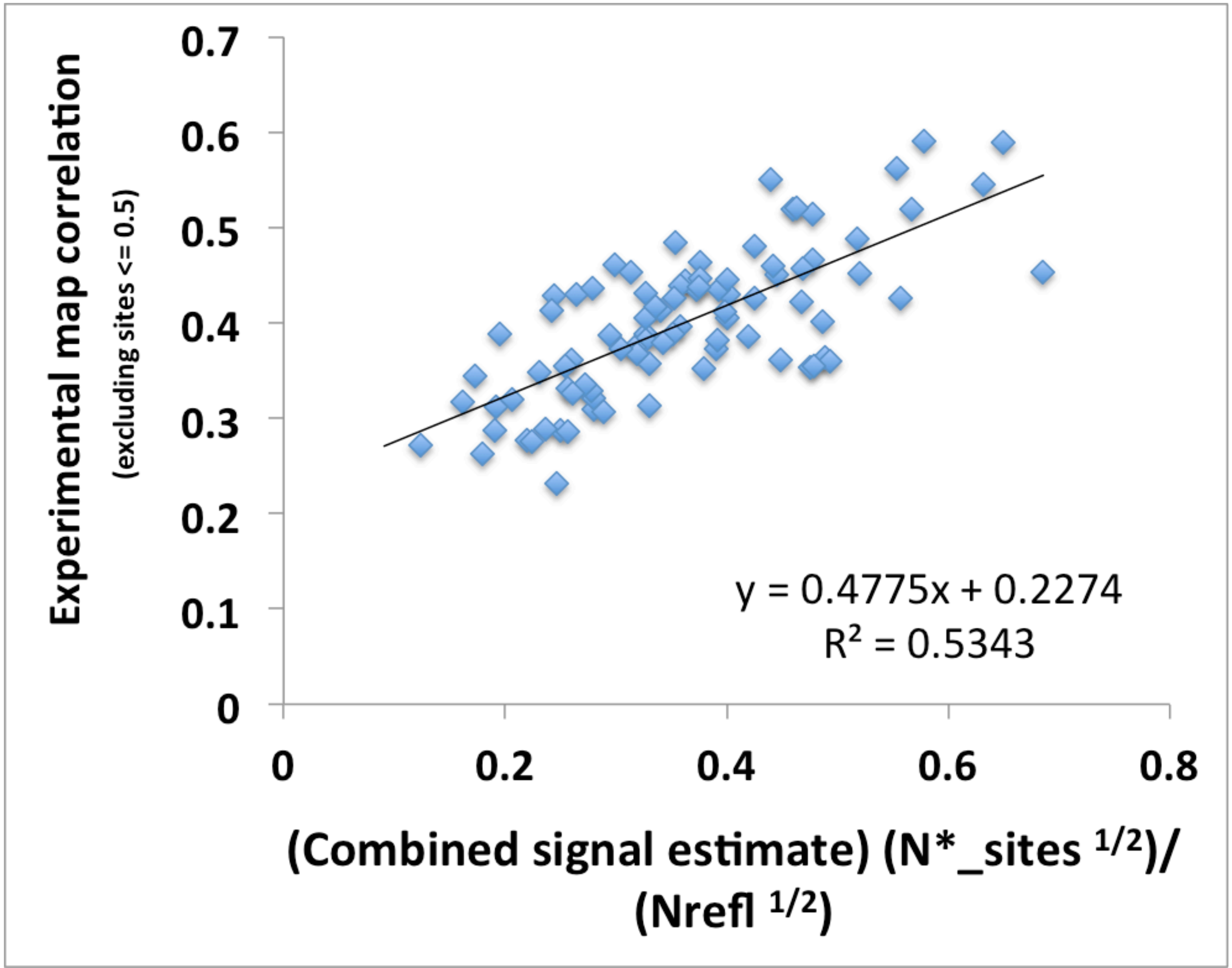
How good will the phasing be: Could we use the anomalous signal S_{ano}?



How good will the phasing be: Using the anomalous correlation CC_{ano} to guess

Phenix

Real-world case: CC_{ano} estimated from the data



Structure solution from weak anomalous data: Perspectives

Anomalous signal and anomalous correlation are useful measures of quality and can be estimated from the data

Likelihood-based methods for finding the anomalous substructure are powerful even with weak signal

Structures can be solved with weak signal

The PHENIX Project

Phenix

Lawrence Berkeley Laboratory

Paul Adams, Pavel Afonine, Nat Echols, Ralf Grosse-Kunstleve, Nigel Moriarty, Nicholas Sauter, Peter Zwart

Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung

Berkeley Lab

Los Alamos National Laboratory

University of Cambridge

Randy Read, Airlie McCoy, Gabor Bunkoczi, Rob Oeffner, Richard Mifsud

Cambridge University

Duke University

Jane & David Richardson, Jeff Headd, Vincent Chen, Chris Williams, Bryan Arendall, Swati Jain, Bradley Hintze, Lizbeth Videau

An NIH/NIGMS funded Program Project

Berkeley Lab

Phenix