Phenix

Experimental phasing, including recent developments in SAD phasing

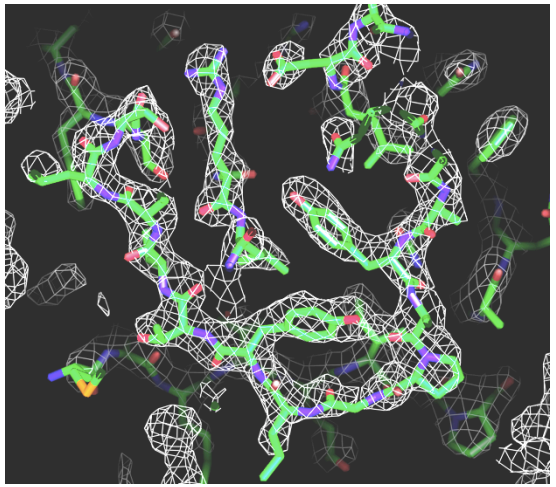
Utilizing The Phenix Software Suite For Protein Crystallography

August 6-7, 2015

University of Kansas

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

Planning a SAD experiment

Maximizing the anomalous signal and the anomalous correlation

*The **anomalous correlation** is a measure of the accuracy of each anomalous difference*

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

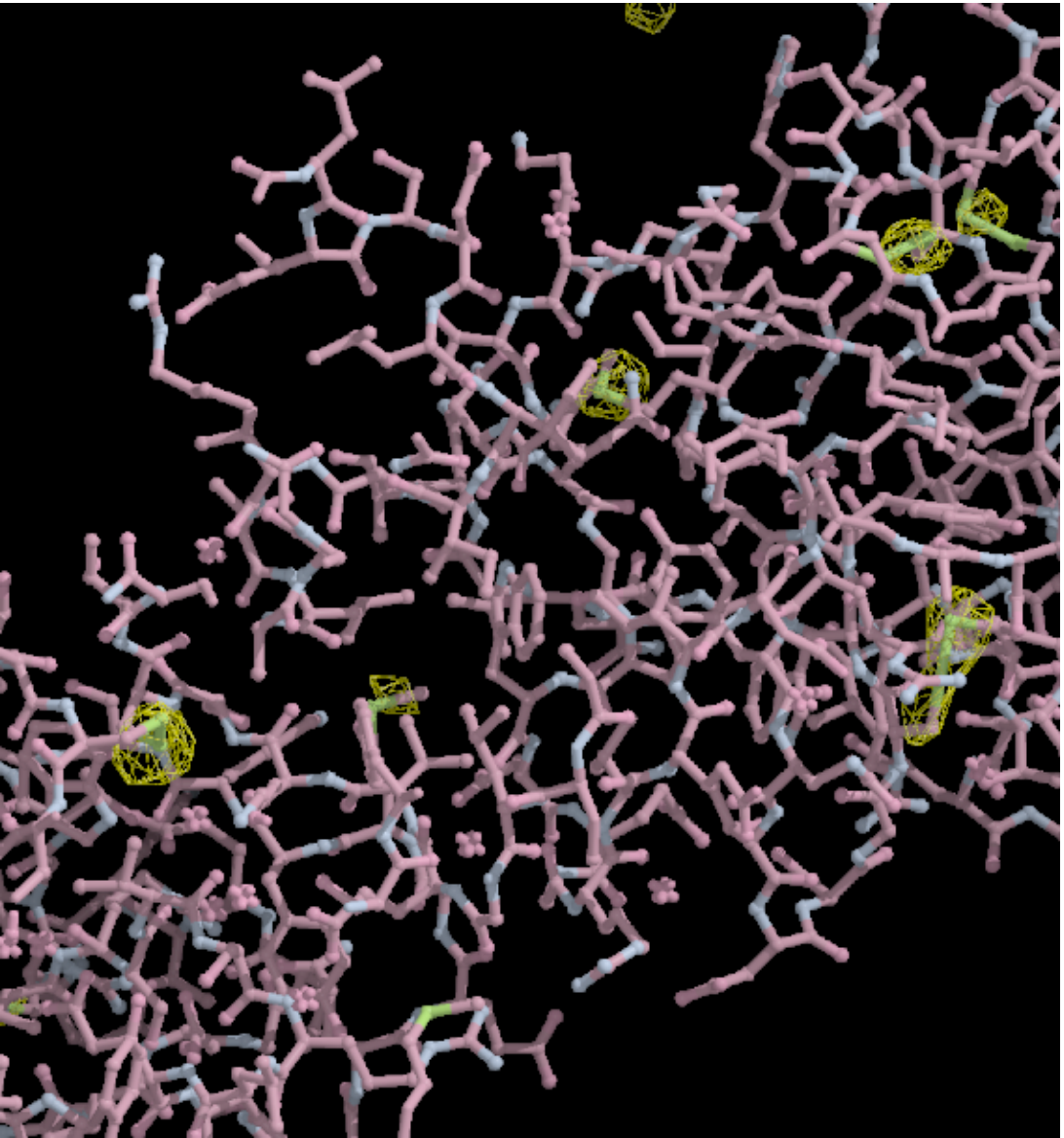
Anomalous correlation: accuracy of anomalous differences

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

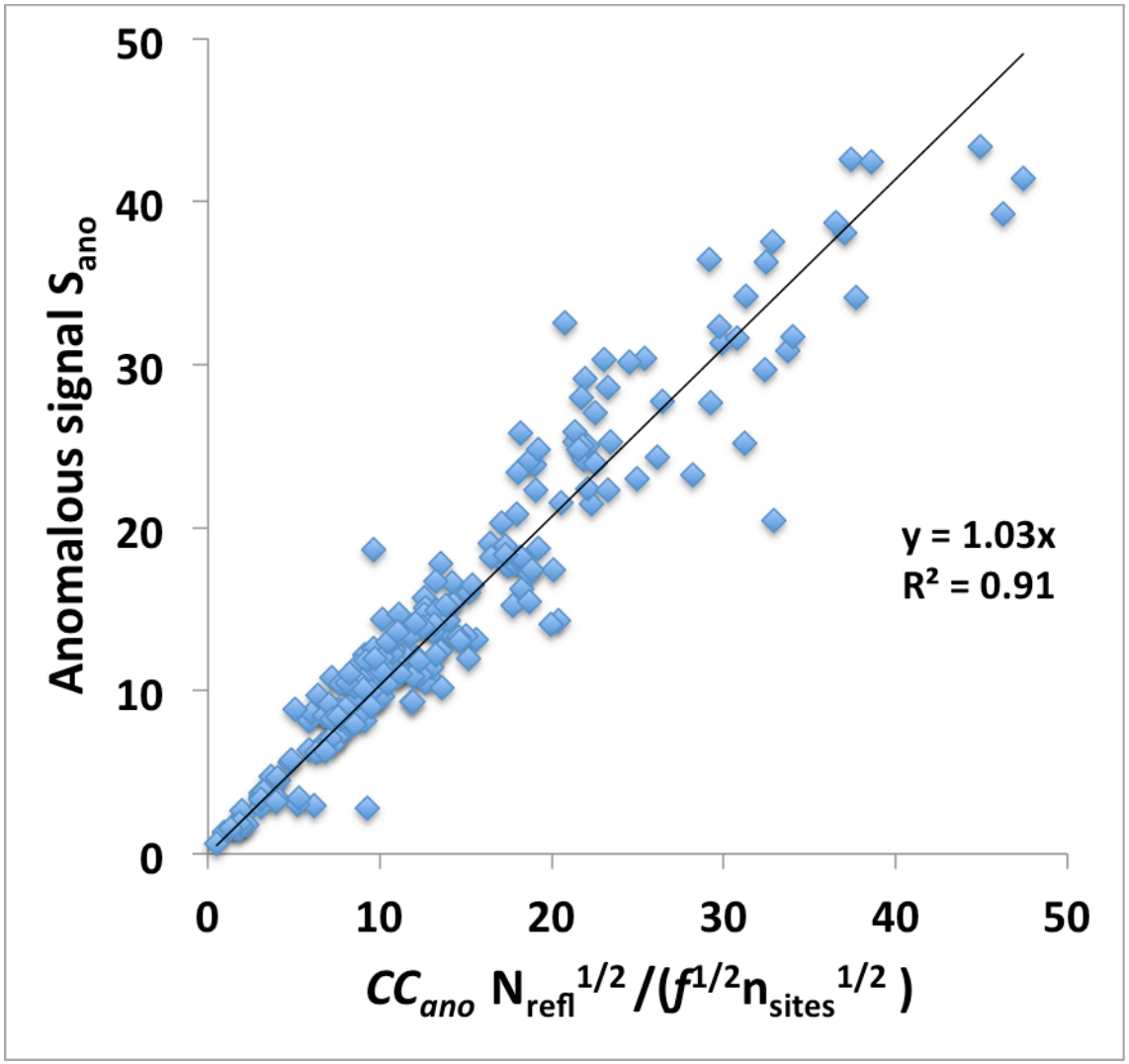
$$f = \frac{< (f^h)^2 >}{< f^h >^2}$$

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



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

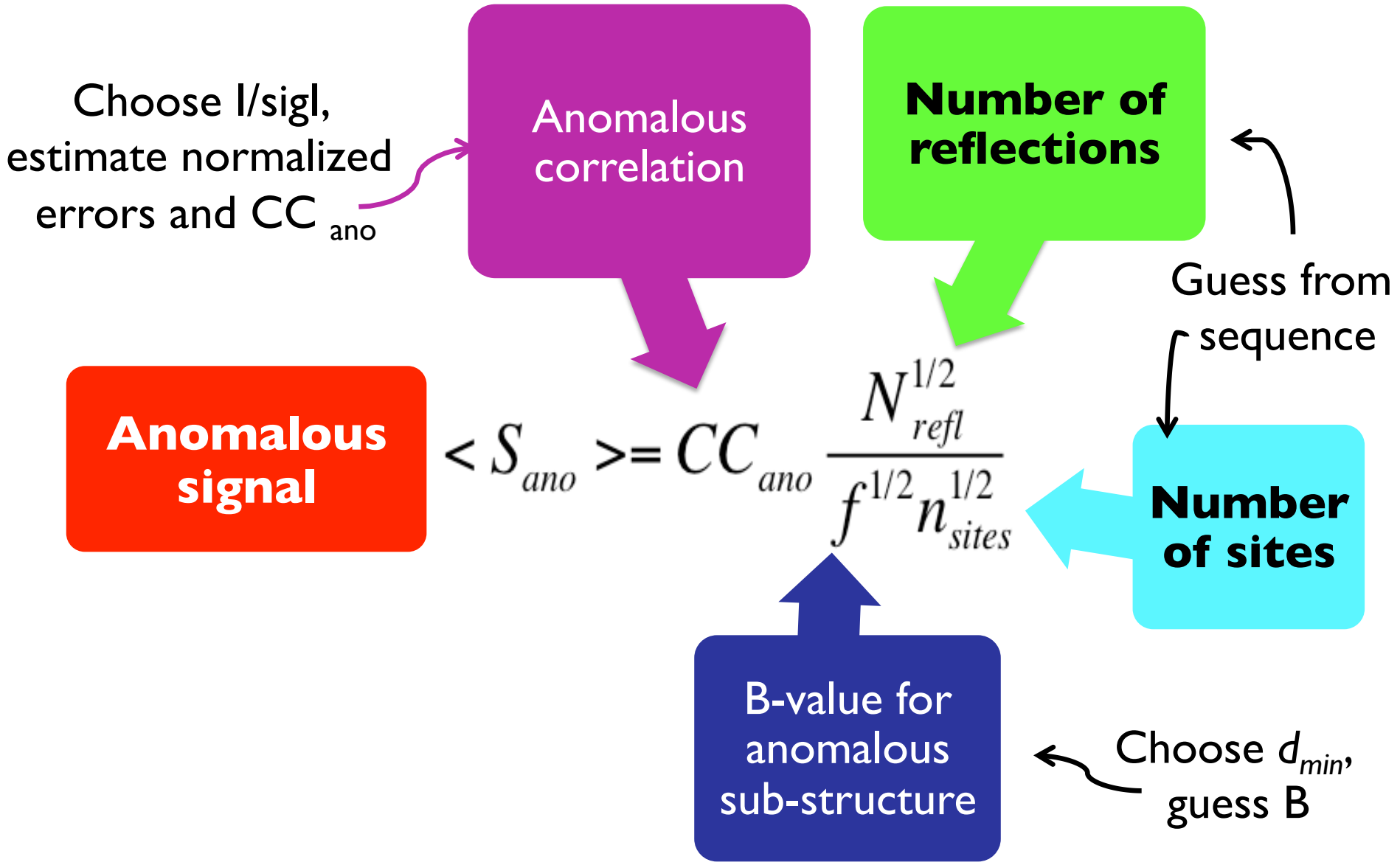
CC_{ano}: Correlation of anomalous differences with model differences

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

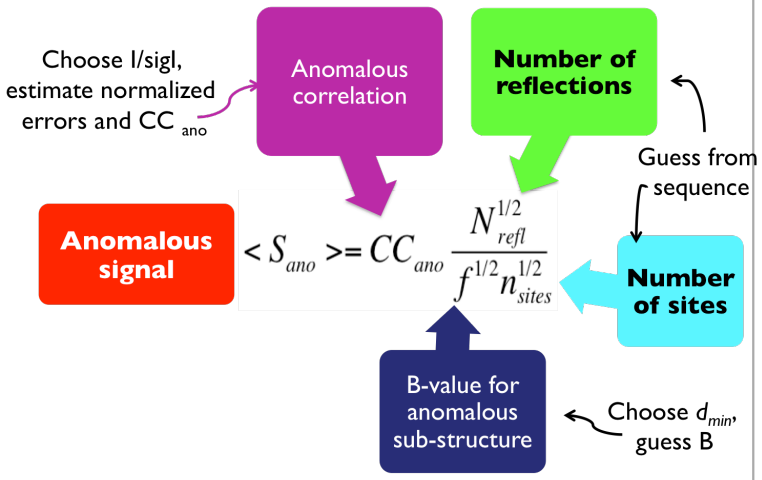
218 SAD datasets 1.2 – 4.5 Å

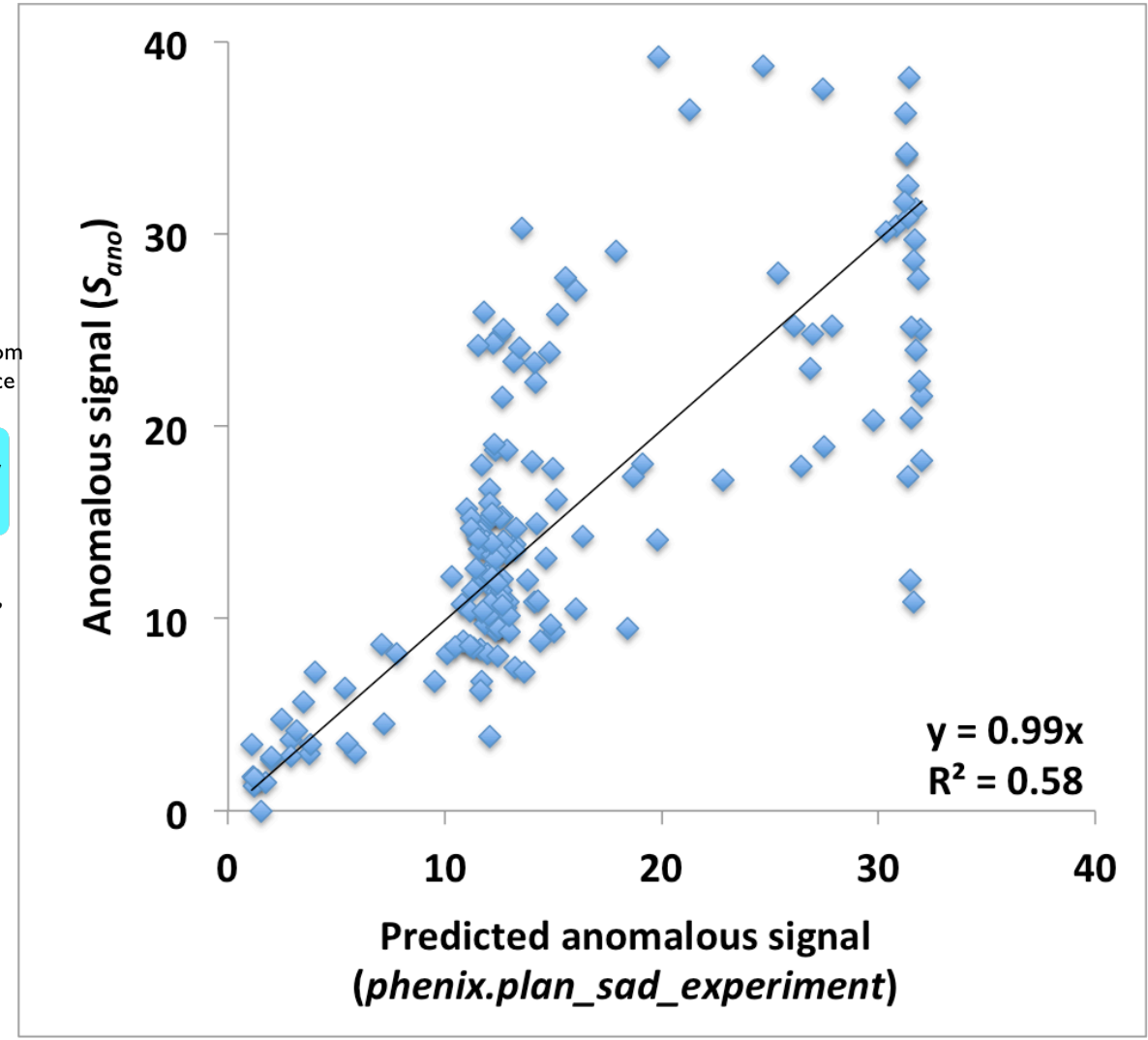
phenix.plan_sad_experiment

Design an experiment that will give you enough anomalous signal



Estimating the anomalous signal before collecting the data





Optimizing scaling and merging of SAD data

(phenix.scale_and_merge)

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

Crystal1
F+,F-

Errors in measurement (σ_{obs})

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

Crystal 2
F+,F-

Optimizing estimates of F+,F-

Crystal1
F+,F-

Local scaling to reduce systematic errors

Use of $\sigma_{crystal}$ in weighting

Crystal 2
F+,F-

Applying inter-dataset variances in weighting

Crystal1

Δ_{ano}

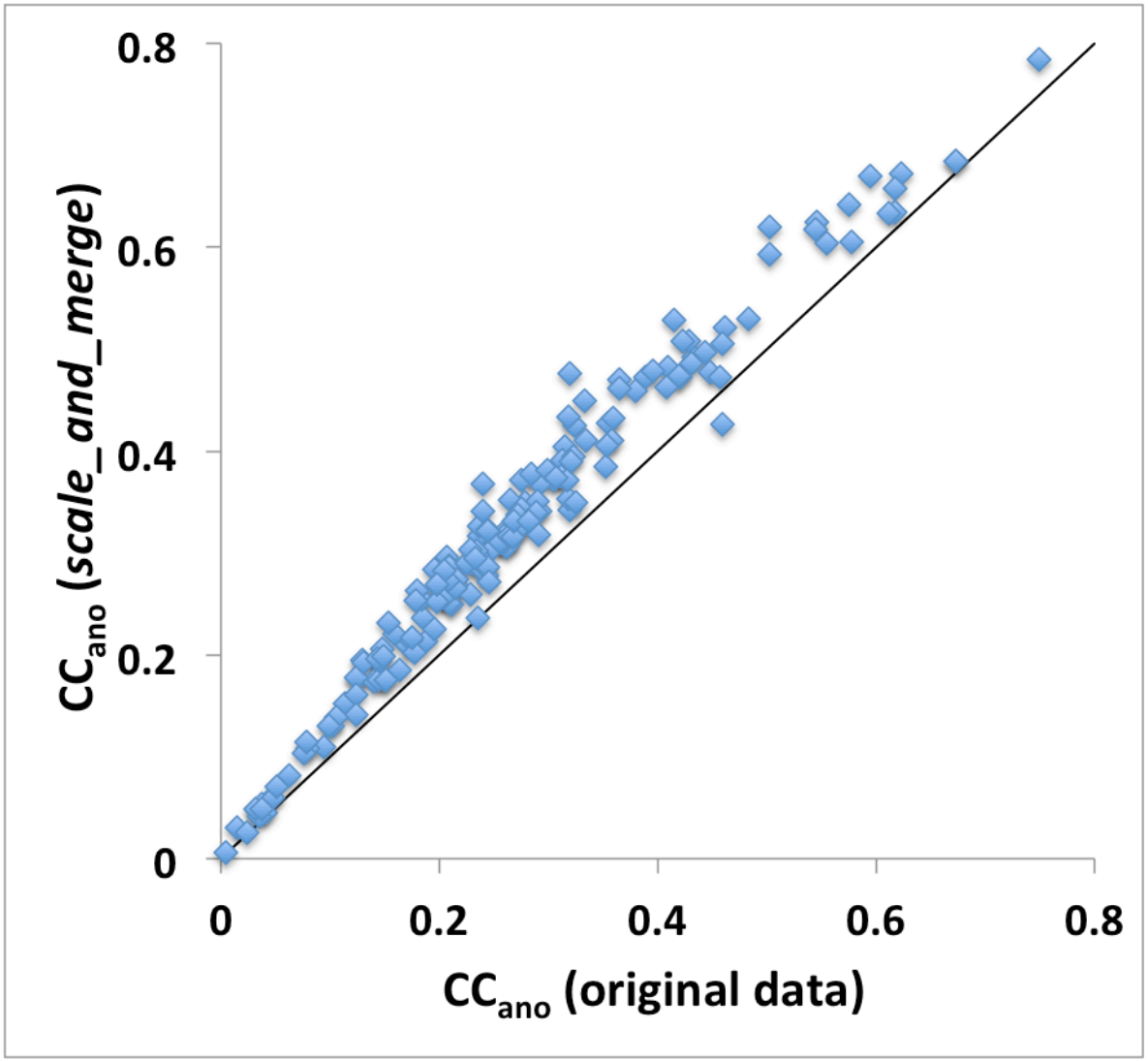
Weighting for data from an individual crystal:

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

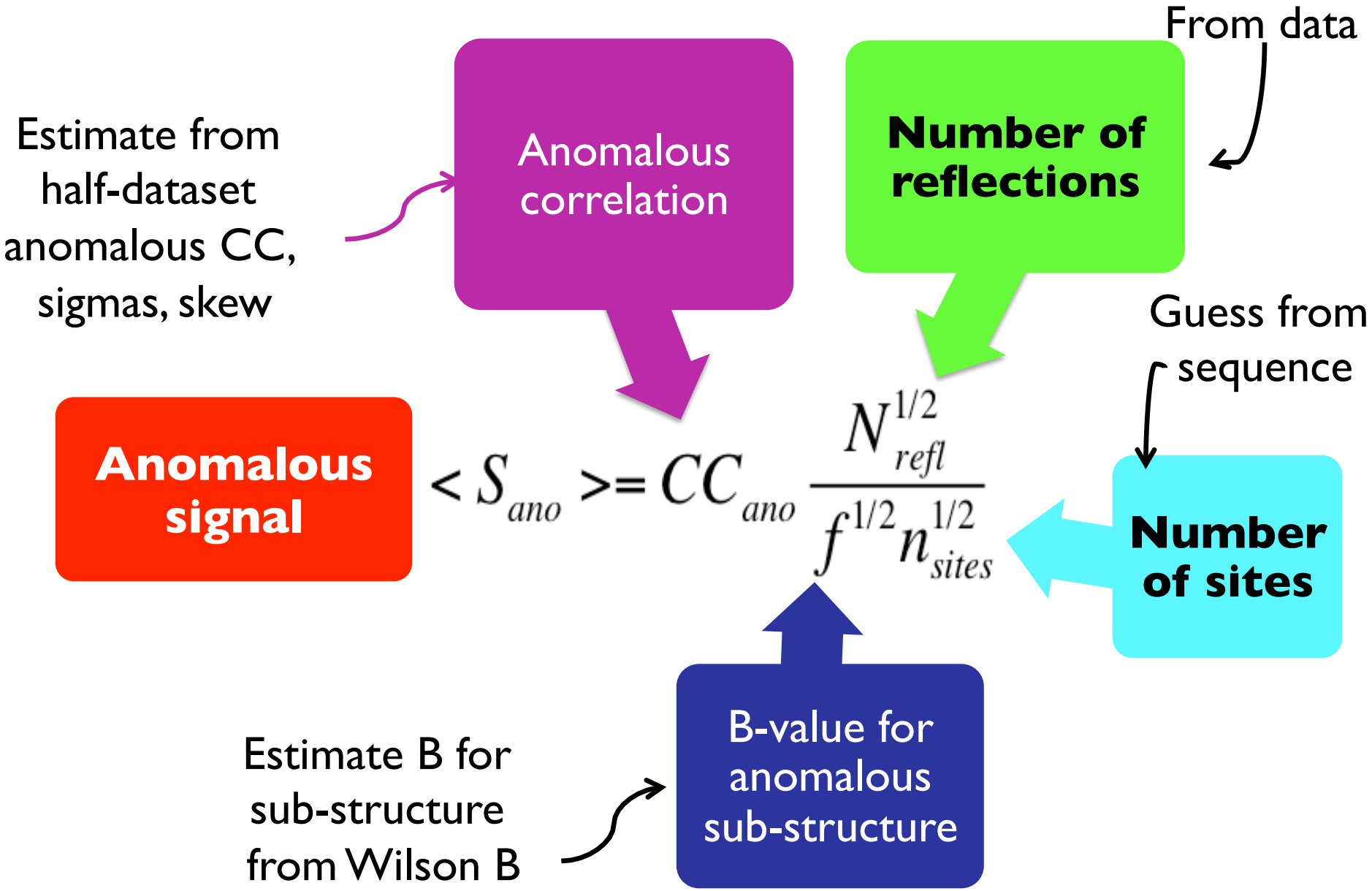
Average of all crystals

Δ^{AVG}

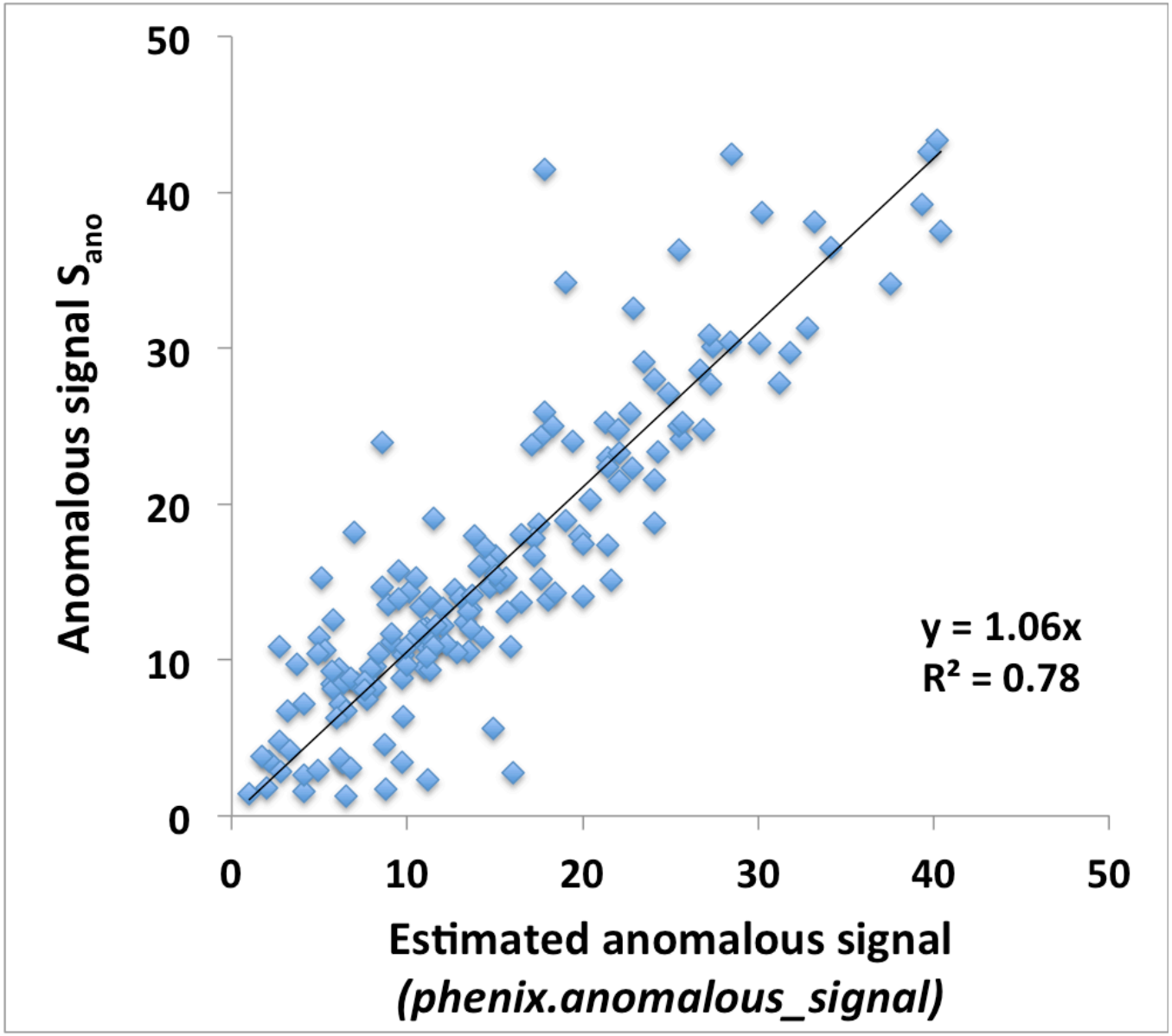
Improvement in anomalous correlation using local scaling in *phenix.scale_and_merge*



Estimating the anomalous signal after collecting the data



Estimating the anomalous signal after collecting the data



Finding the anomalous sub-structure with the SAD likelihood function

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

Most powerful source of information about the sub-structure before phases are known

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.

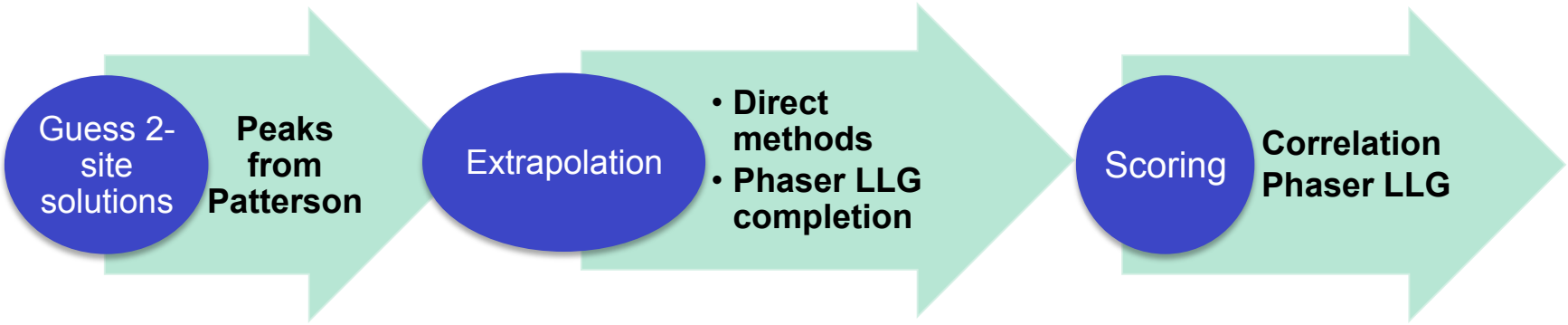
LLG sub-structure searches in HySS

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

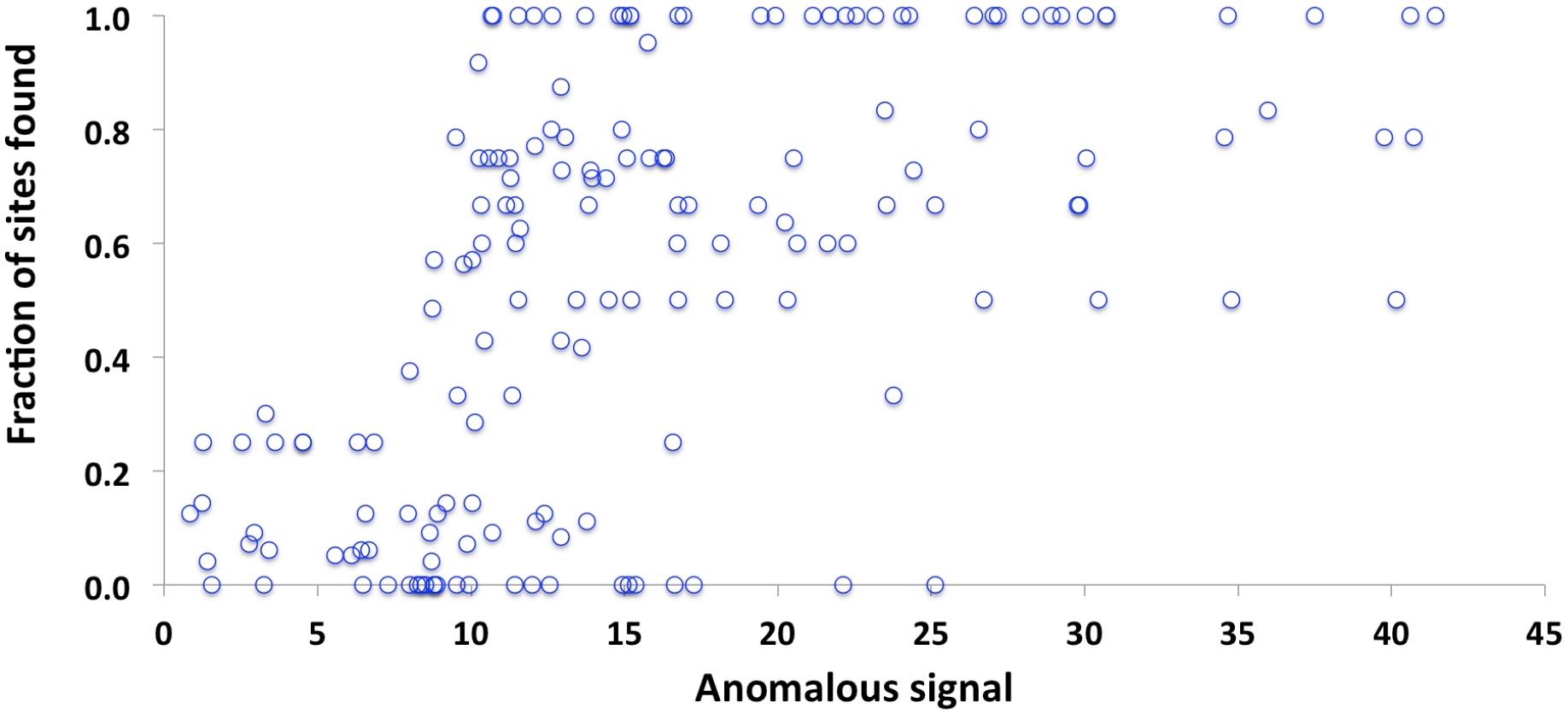
(cut-off for peak height)

Use LLG score to compare solutions

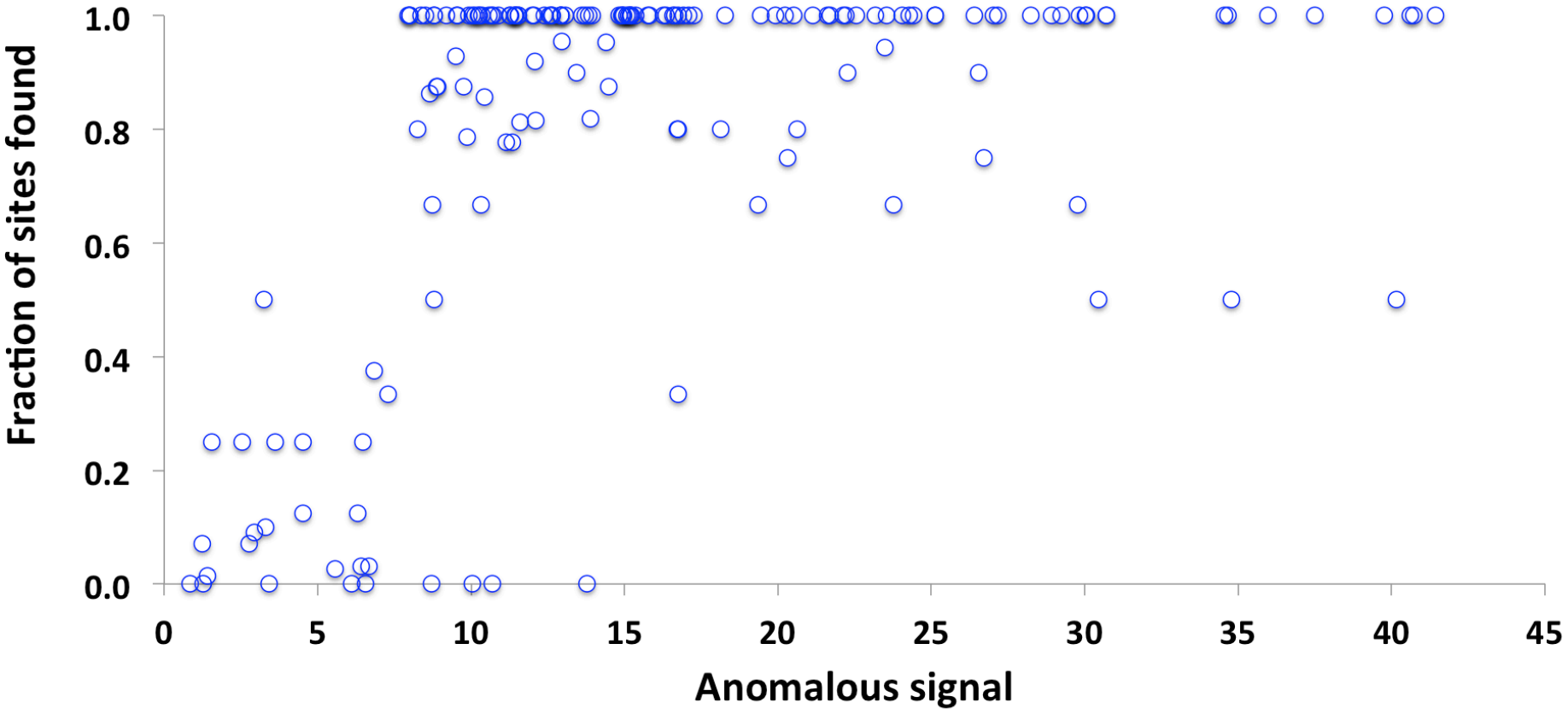
Terminate early if same solution found several times

Run quick direct methods first

Dual Space Sub-structure Completion

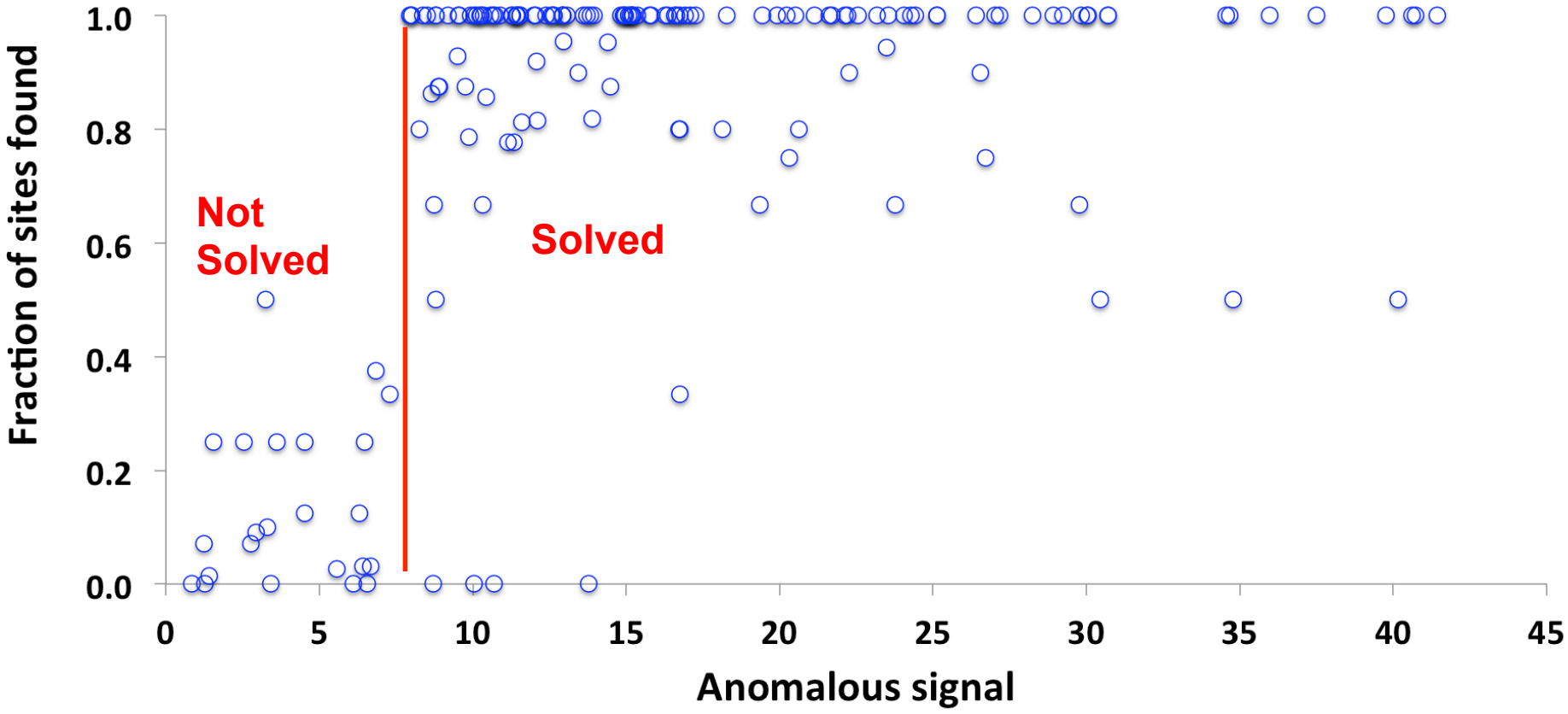


LLG Sub-structure Search



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

Anomalous signal indicates if a dataset can be solved



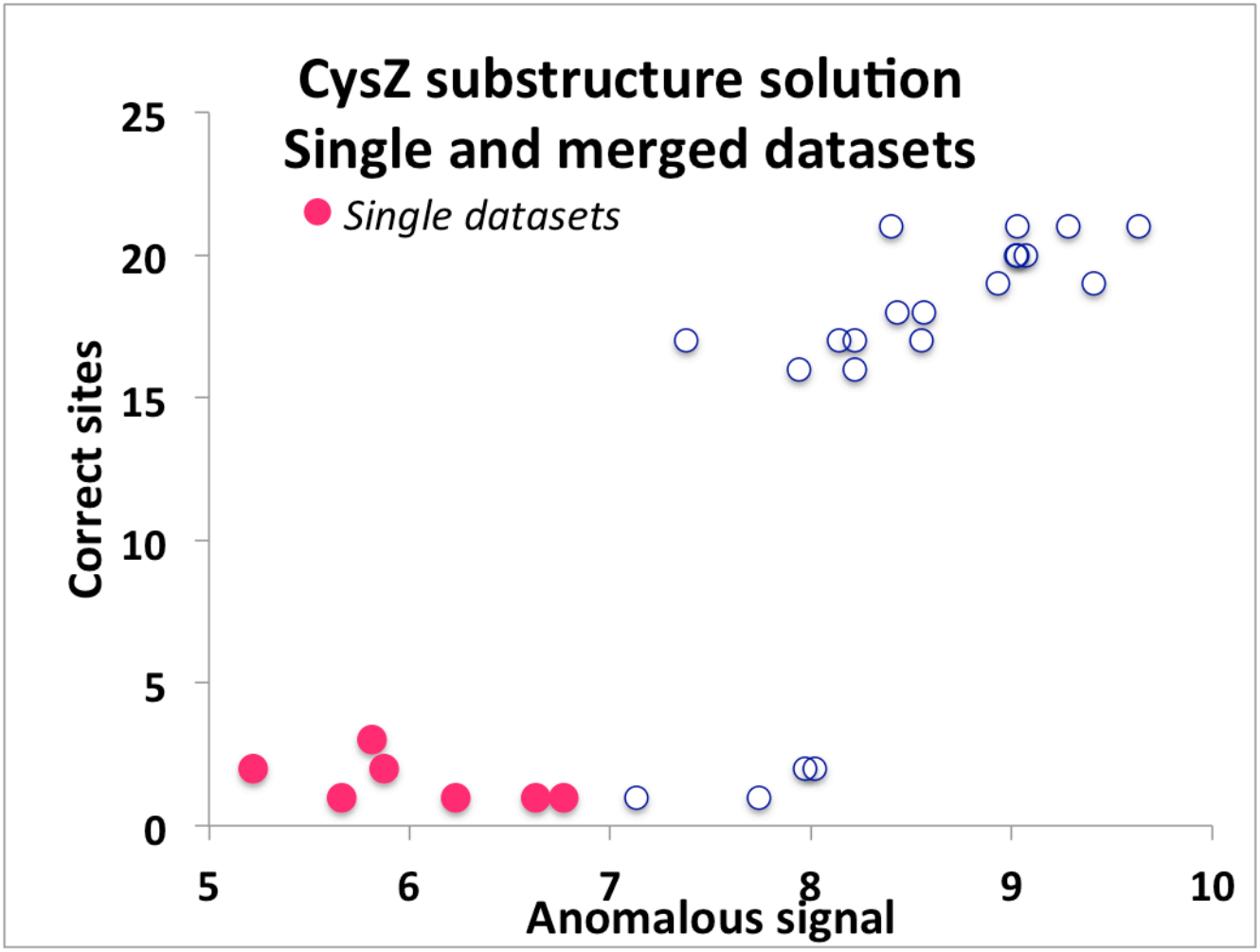
CysZ multi-crystal sulfur-SAD data

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

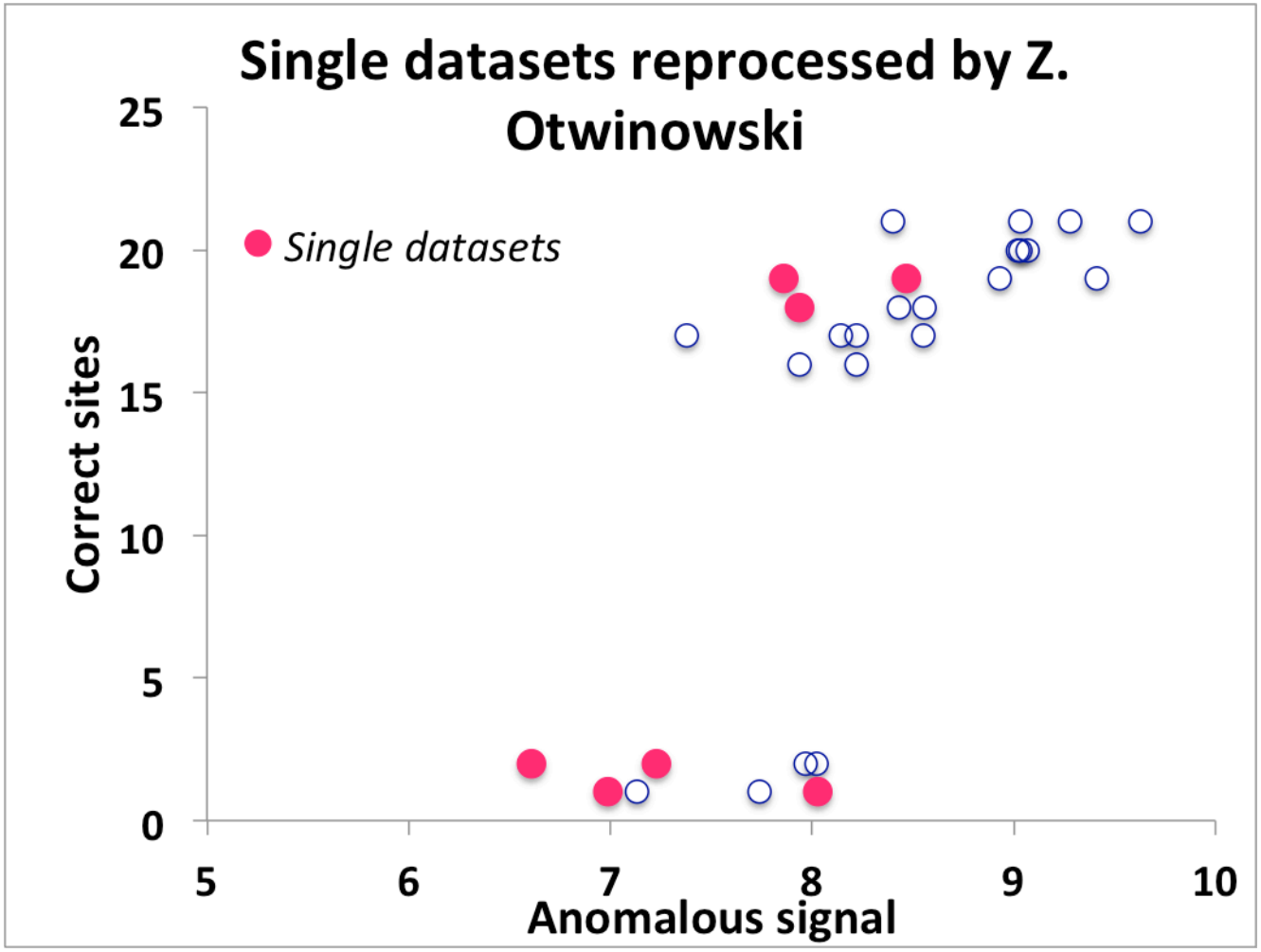
Data from 7 crystals collected at wavelength of 1.74 Å to resolution of 2.3 Å

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

CysZ multi-crystal sulfur-SAD data

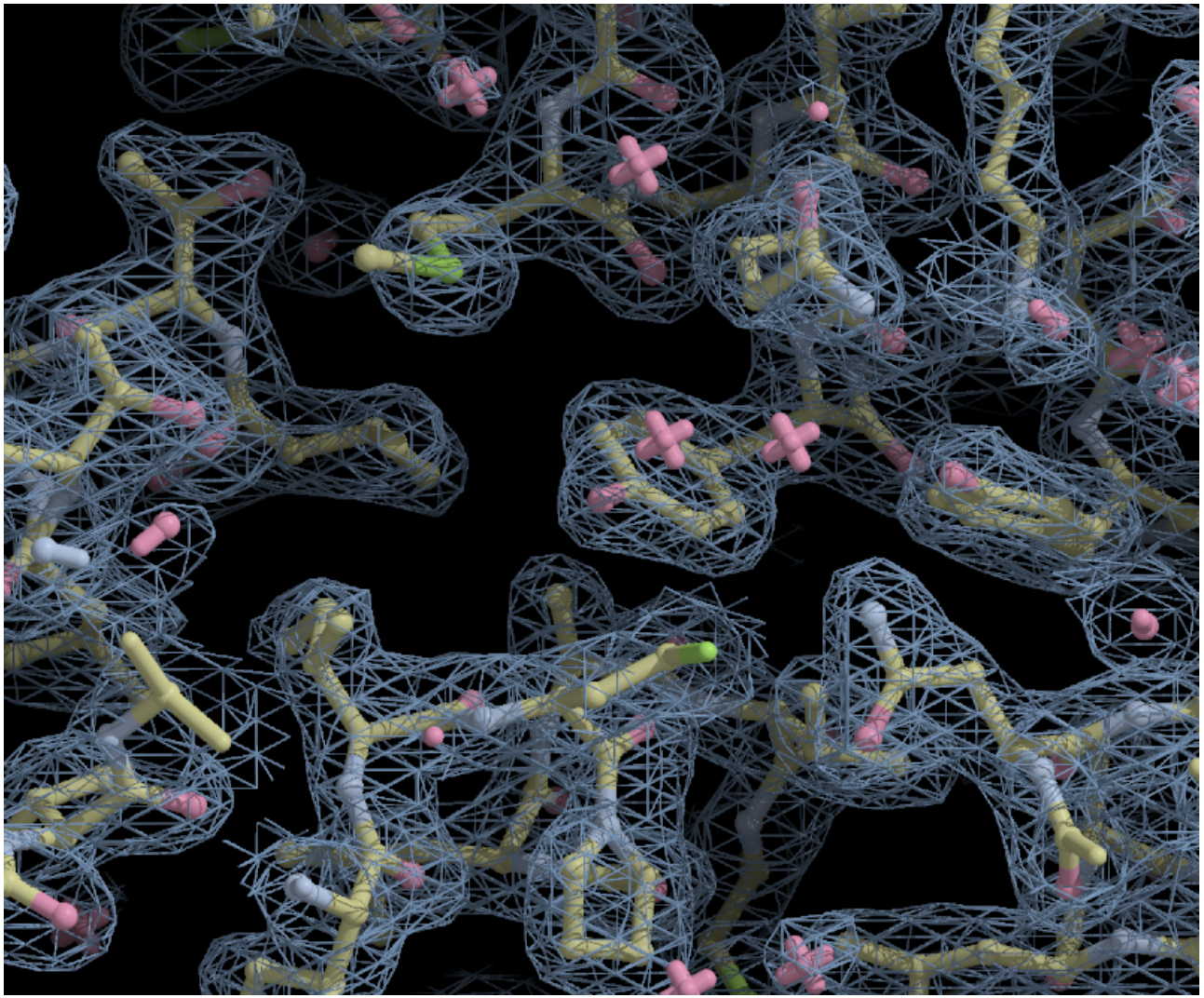


CysZ multi-crystal sulfur-SAD data



CysZ single-crystal sulfur-SAD data

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



Choosing the hand of the sub-structure

Decision-making based on map quality

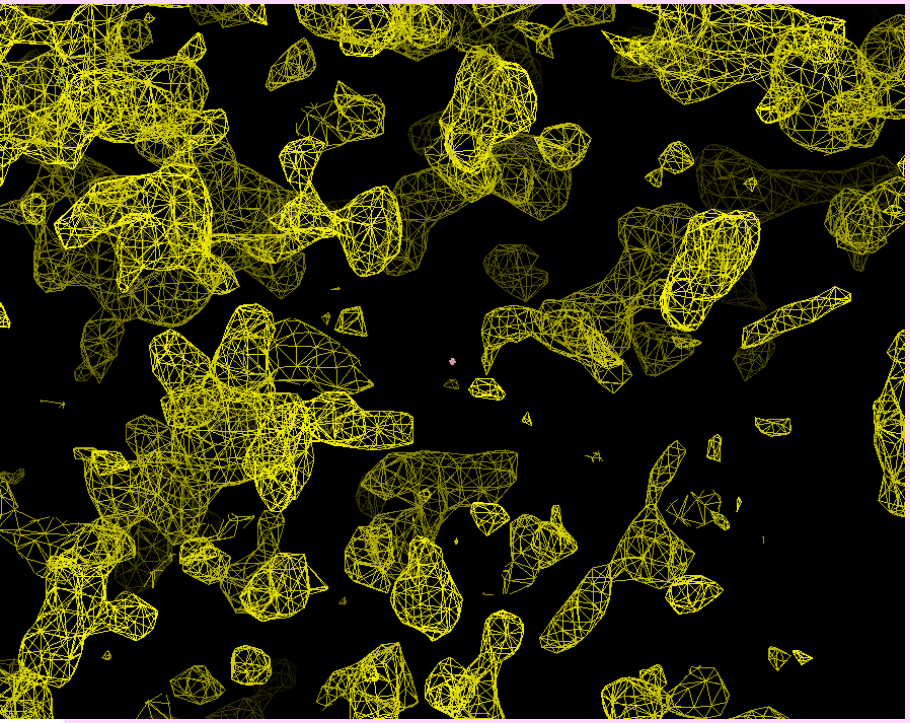
(phenix.autosol)

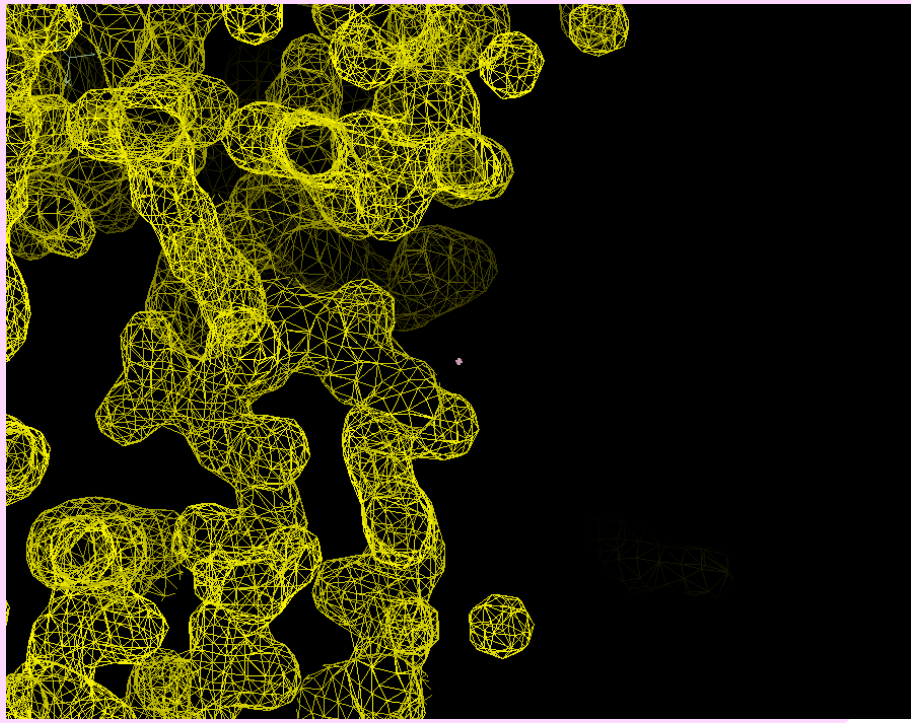
Deciding what is good:

Measures of the quality of an electron-density map:

Which solution is best?

Are we on the right track?



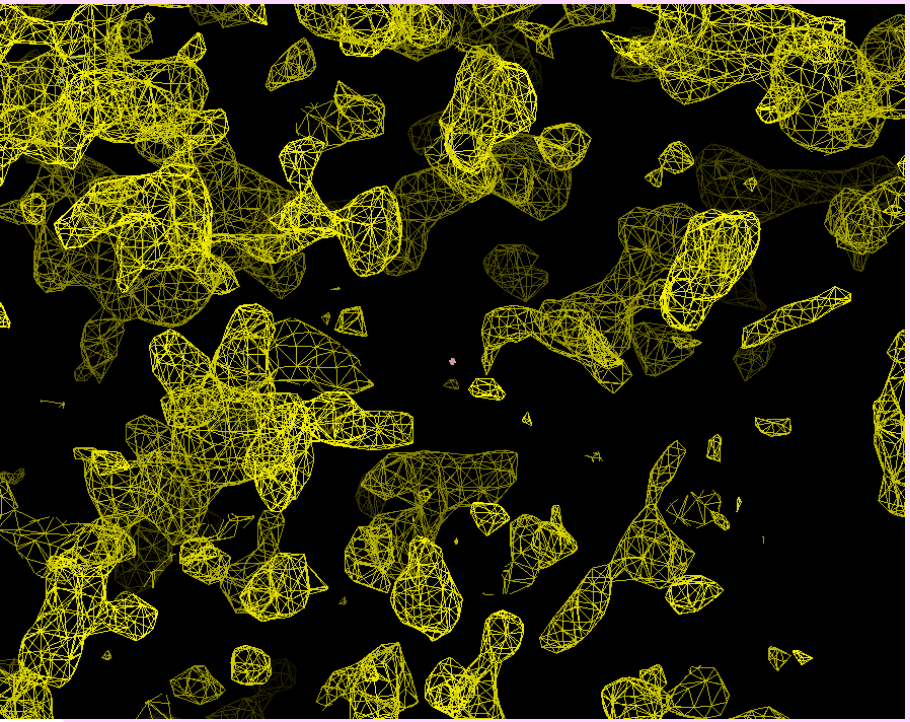


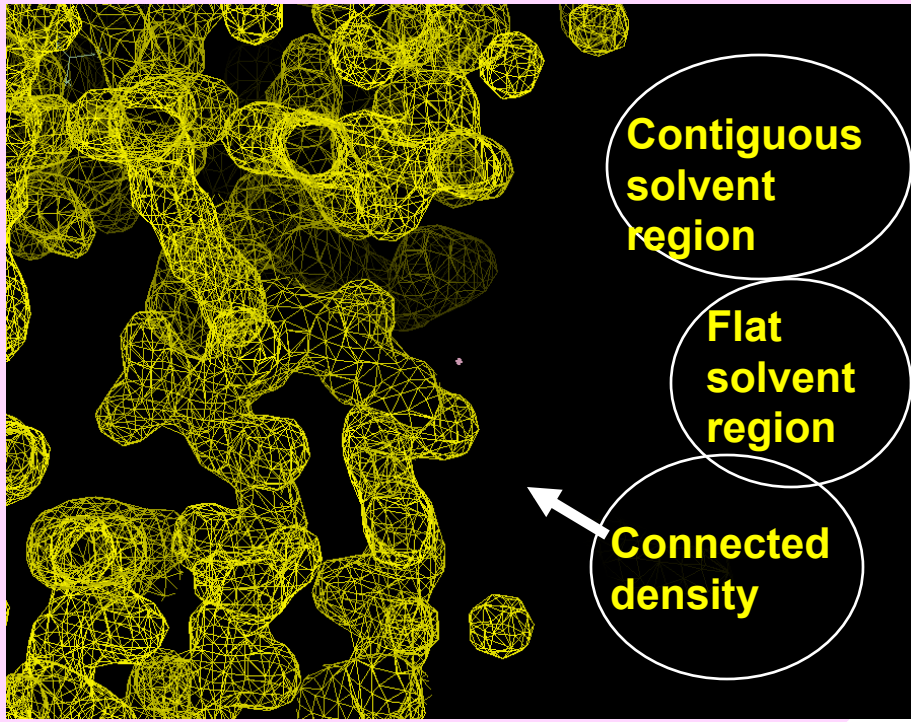
Deciding what is good:

Measures of the quality of an electron-density map:

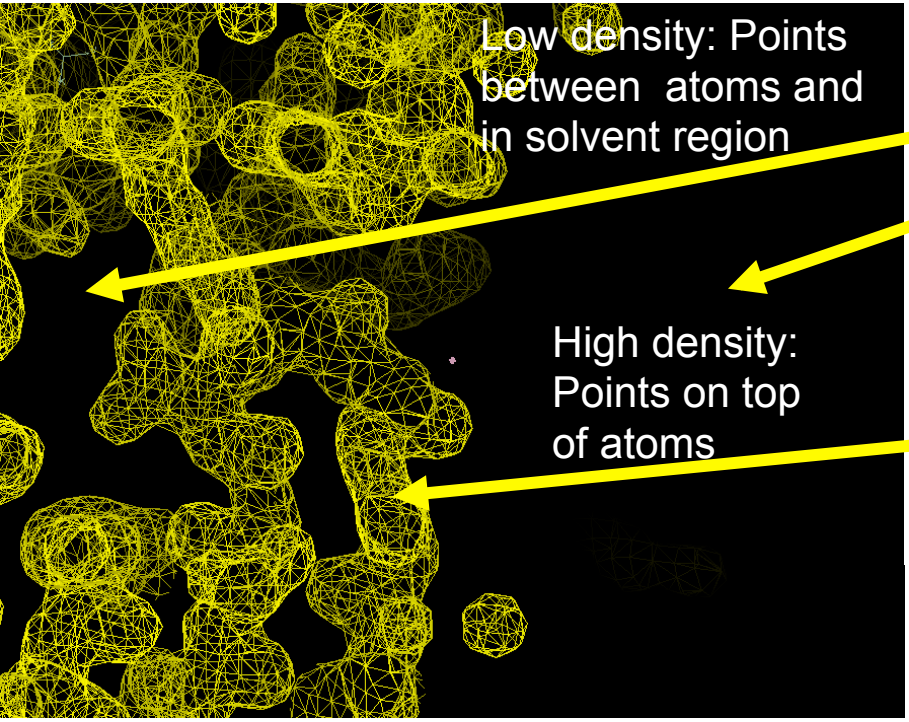
Which solution is best?

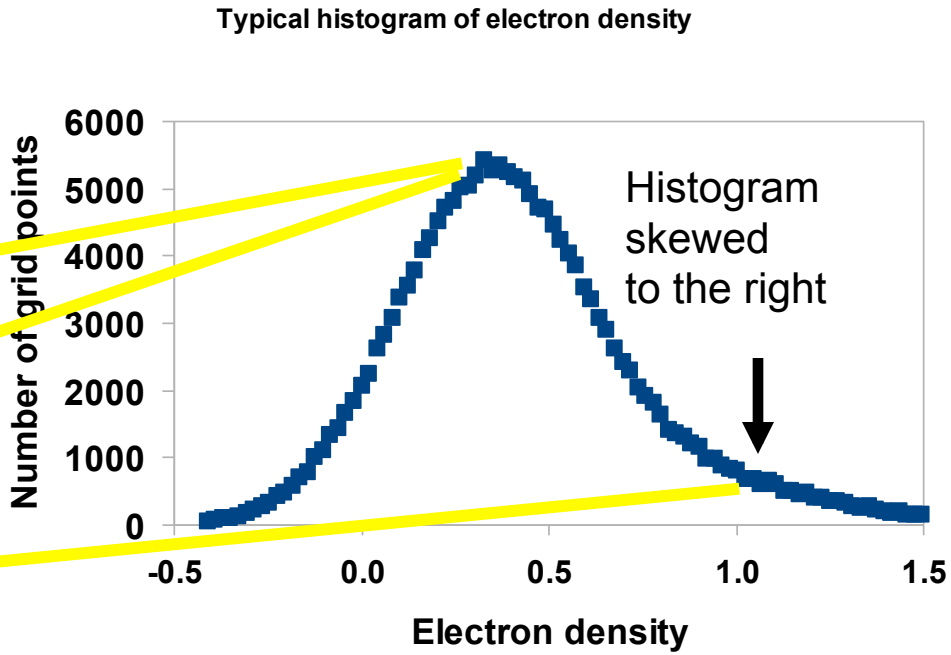
Are we on the right track?



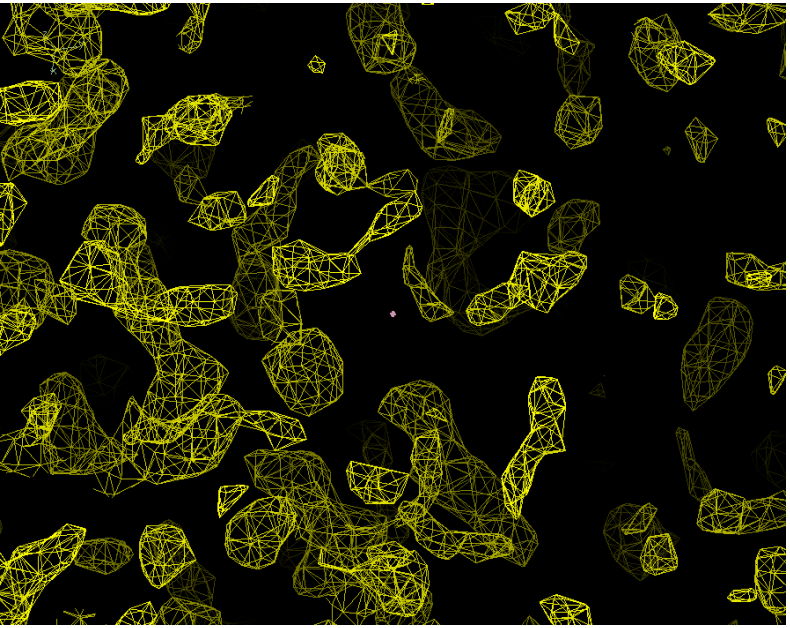


Histogram of electron density values has a positive “skew”

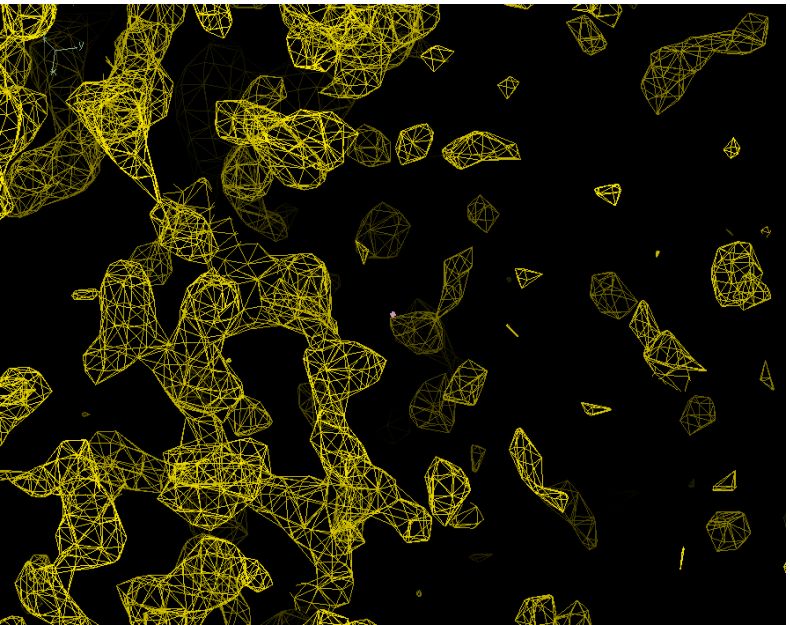




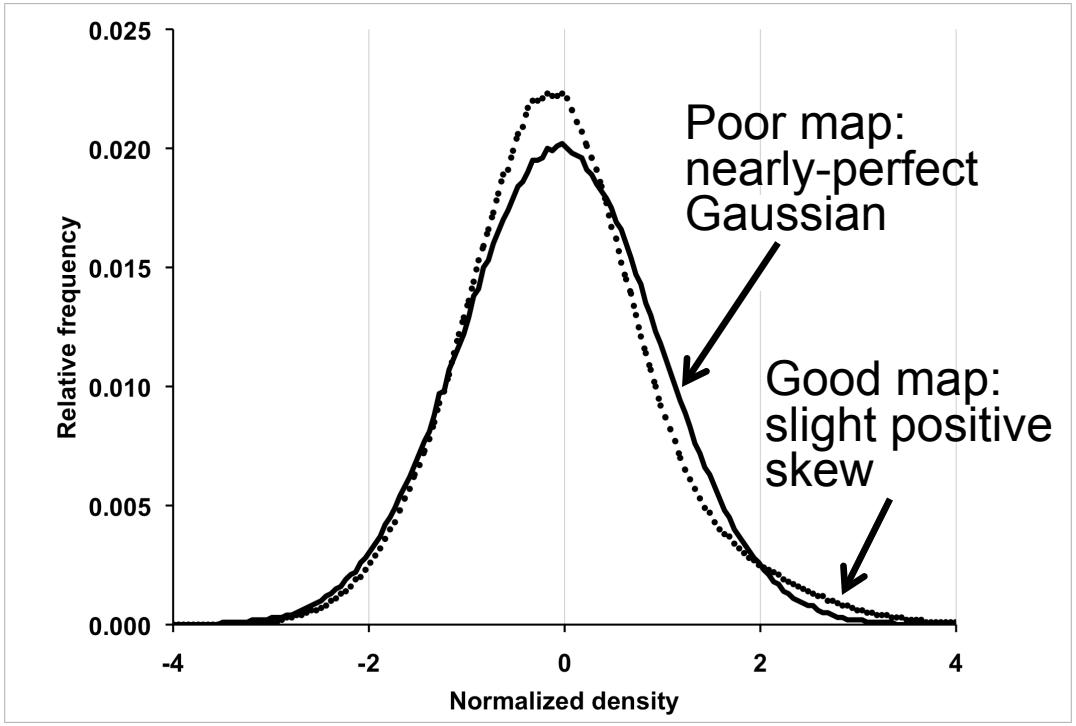
Skew of electron density differentiates poor and good maps – even when the difference is barely visible



Poor map
(inverse hand)



Good map



Evaluating electron density maps

Basis	*Good map*	*Random map*
Skew of density (Podjarny, 1977)	Highly skewed (very positive at positions of atoms, zero elsewhere)	Gaussian histogram
Connectivity of regions of high density (Baker, Krukowski, & Agard, 1993)	A few connected regions can trace entire molecule	Many very short connected regions
Correlation of local rms densities (Terwilliger, 1999)	Neighboring regions in map have similar rms densities	Map has uniform rms density
R-factor in 1st cycle of density modification (Cowtan, 1996)	Low R-factor	High R-factor

How well does the skew reflect map quality?

Create real maps

Score the maps based on skew

Compare the scores with the actual quality of the maps

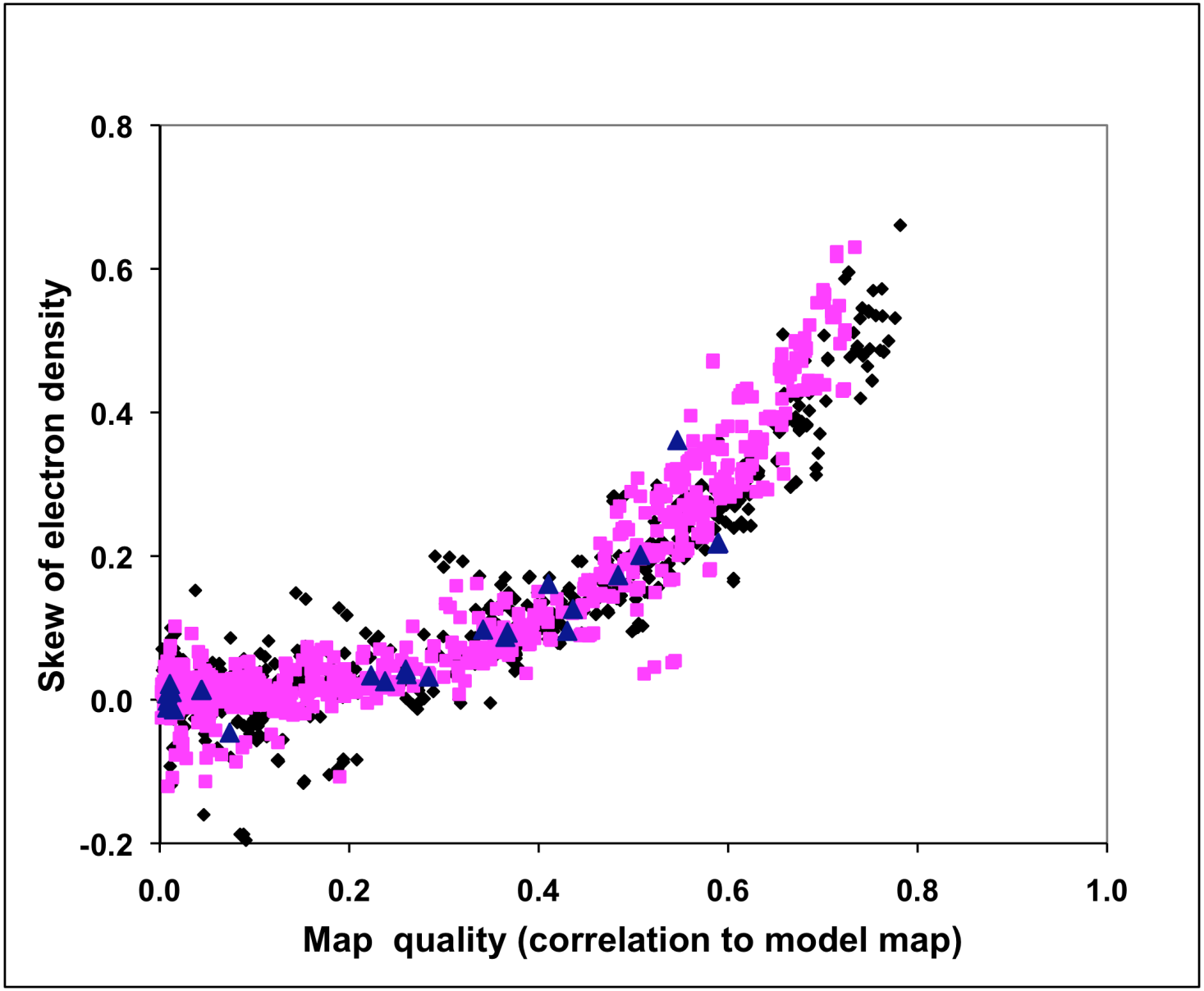
Creating real maps

247 MAD, SAD, MIR datasets with final model available
(*Phenix* library and JCSG publicly-available data)

Run *phenix.autosol* on each dataset.

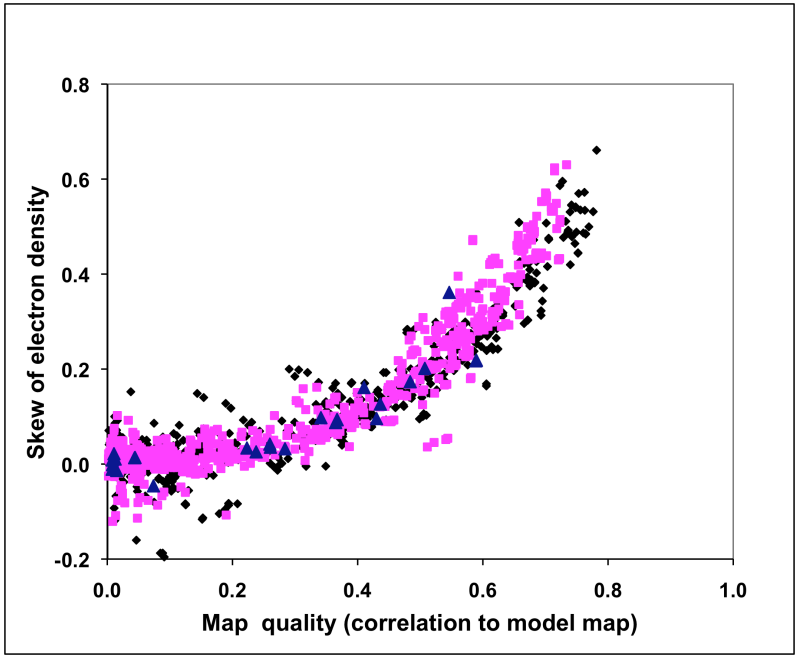
Calculate maps for each solution considered
(opposing hands, additional sites, including various derivatives for MIR)

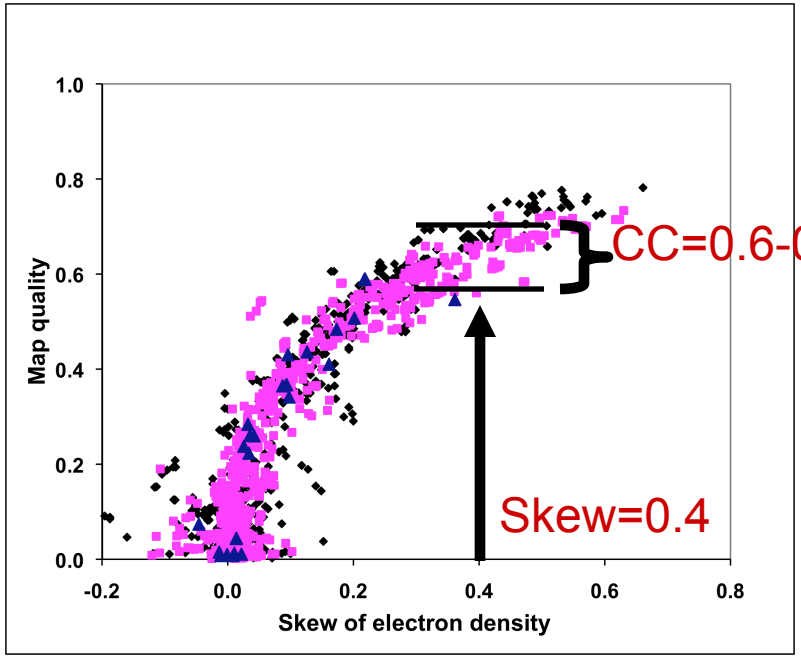
Skew of electron density – positive skew of density values



Using scoring criteria to estimate the quality of a map

Skew depends on map quality → Estimate map quality from skew





Estimated map quality in practice

Evaluating solutions to a 2-wavelength MAD experiment (JCSG Tm3681, 1VPM, SeMet 1.6 Å data)

Data for HYSS	Sites	Estimated CC ± 2SD	Actual CC	
Peak	12	0.73 ± 0.04	0.72	←
Peak (inverse hand)	12	0.11 ± 0.43	0.04	
F_A	12	0.73 ± 0.03	0.72	
F_A (inverse)	12	0.11 ± 0.42	0.04	
Sites from diff Fourier	9	0.70 ± 0.17	0.69	

Phenix

Density modification

Improving phase quality by including expectations about the map

(phenix.autosol; phenix.autobuild)

Statistical density modification

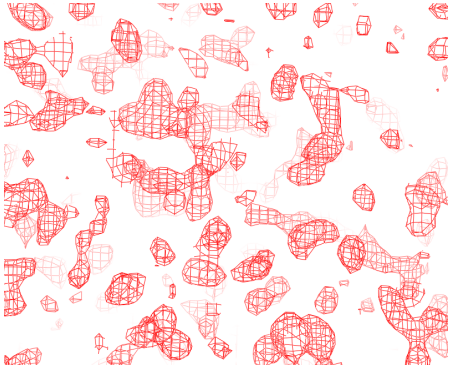
•Principle: phase probability information from probability of the map and from experiment:

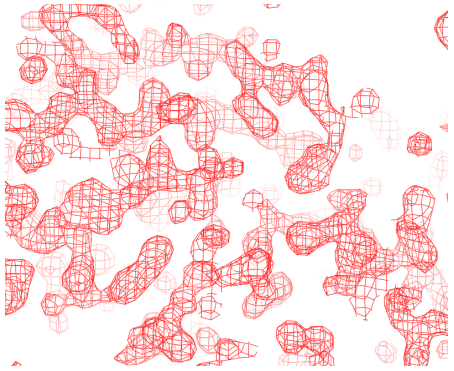
•$P(\phi) = P_{\text{map probability}}(\phi)\, P_{\text{experiment}}(\phi)$

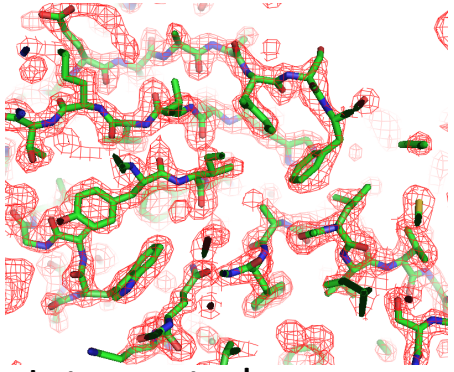
•"Phases that lead to a believable map are more probable than those that do not"

•A believable map is a map that has…
•a relatively flat solvent region
•NCS (if appropriate)
•A distribution of densities like those of model proteins

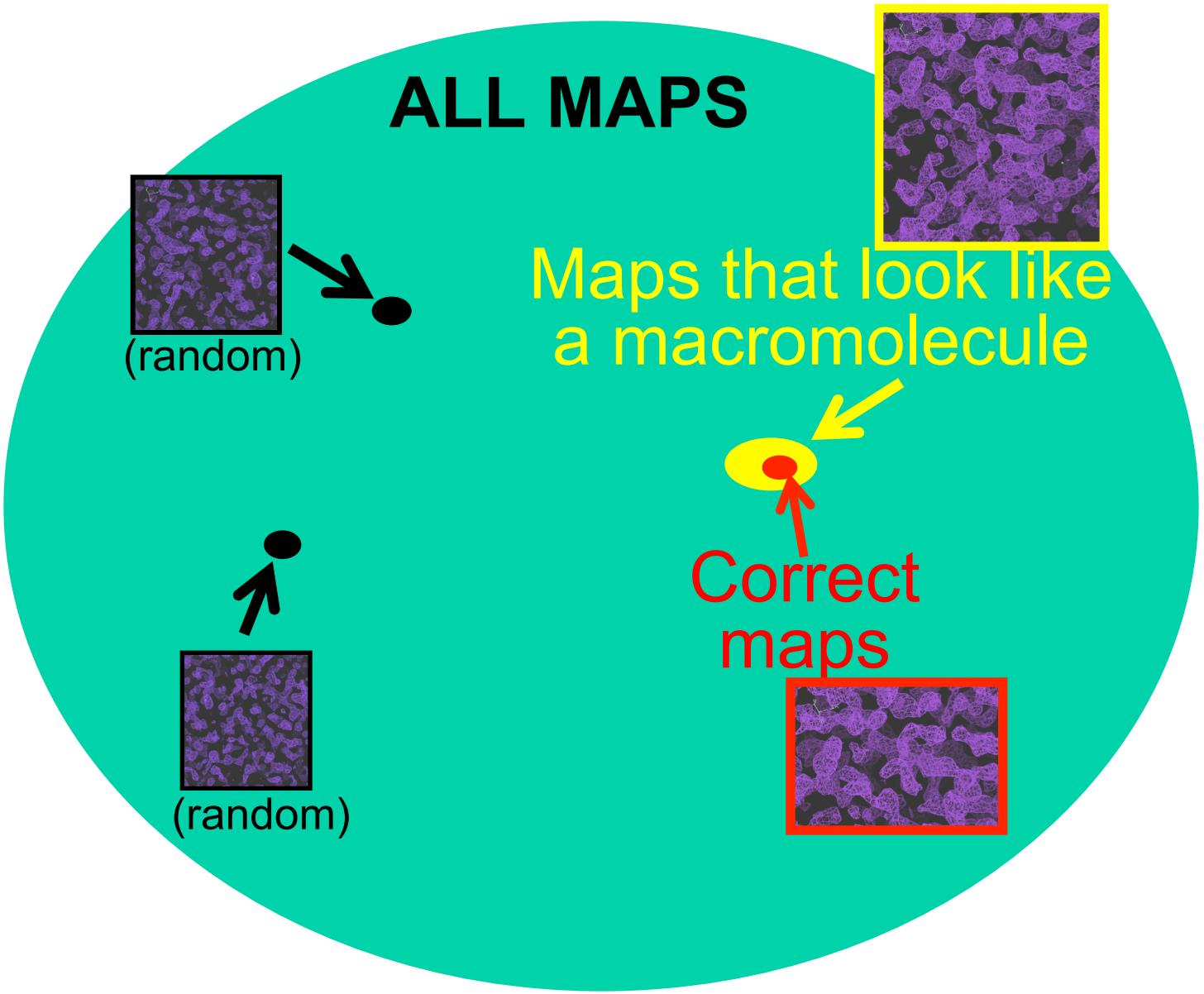
•Method:
•calculate how map probability varies with electron density ρ
•deduce how map probability varies with phase ϕ
•combine with experimental phase information


Experimental map


Density-modified

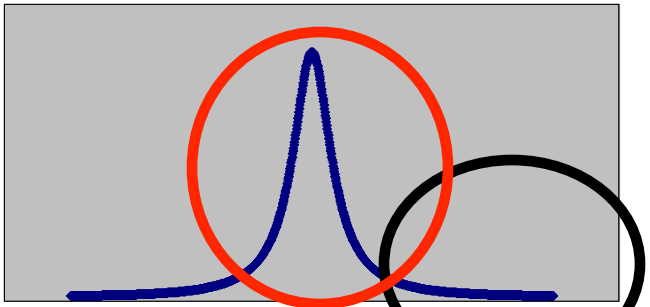

Interpreted

Maps that look like proteins are MUCH more likely to be correct than ones that do not

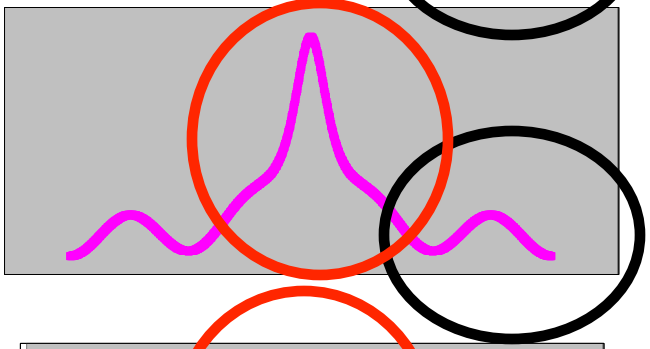


Map probability phasing: Getting a new probability distribution for each phase given estimates of all others

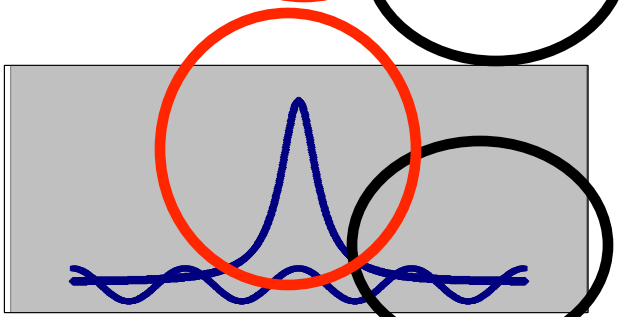
1. Identify expected features of map (flat far from center)
2. Calculate map with current estimates of all structure factors except one (k)
3. Test all possible phases ϕ for structure factor k (for each phase, calculate new map including k)
4. Probability of phase ϕ estimated from agreement of map with expectations
5. Phase probability of reflection k from map is independent of starting phase probability because reflection k is omitted from the map



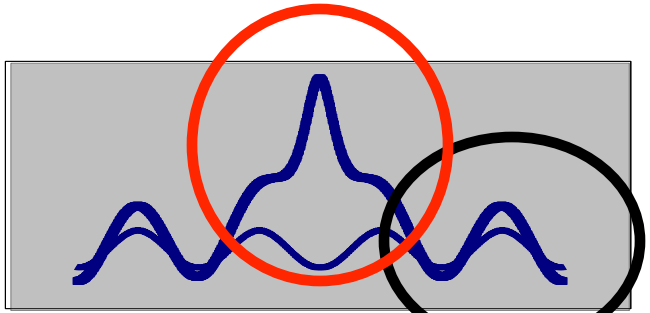
A function that is (relatively) flat far from the origin



Function calculated from estimates of all structure factors but one (k)



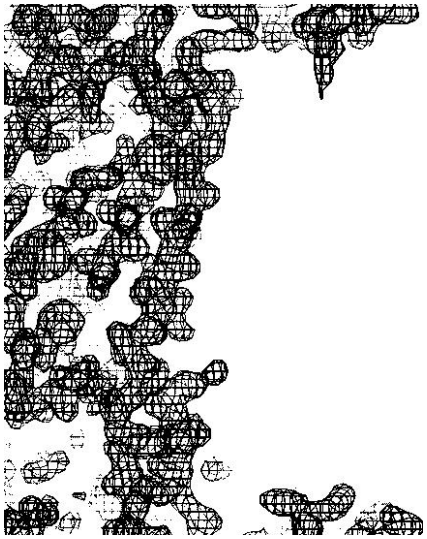
Test each possible phase of structure factor k. P(ϕ) is high for phase that leads to flat region



A map-probability function – allowing different weighting of information from different parts of the map

Log-probability of the map is sum over all points in map of local log-probability

$$LL^{MAP}(\{\mathbf{F_h}\}) \approx \frac{\mathrm{N_{REF}}}{\mathrm{V}} \int_{\mathbf{V}} LL(\rho(\mathbf{x}, \{\mathbf{F_h}\}))\mathrm{d}^3\mathbf{x}$$



A map with a flat (blank) solvent region is a likely map

Local log-probability is believability of the value of electron density ($\rho(x)$) found at this point

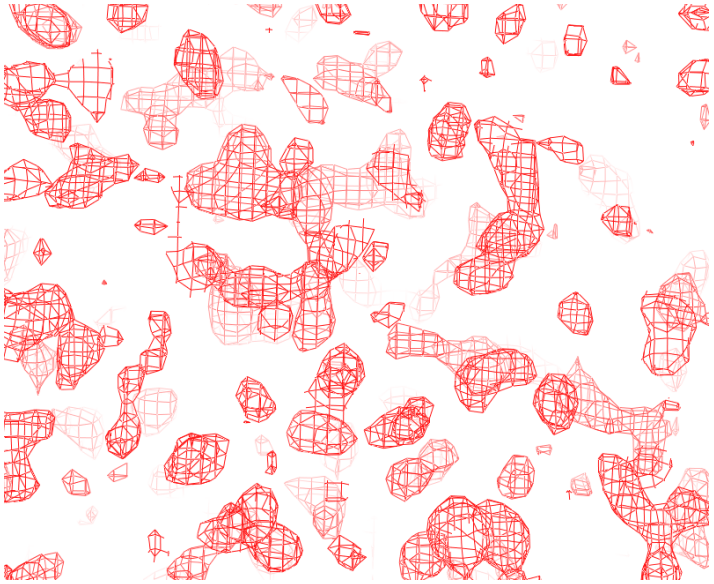
$$LL(\rho(\mathbf{x}, \{\mathbf{F_h}\})) = \ln[p(\rho(\mathbf{x})|PROT)p_{PROT}(\mathbf{x}) + p(\rho(\mathbf{x})|SOLV)p_{SOLV}(\mathbf{x})]$$

If the point is in the PROTEIN region, most values of electron density ($\rho(x)$) are believable

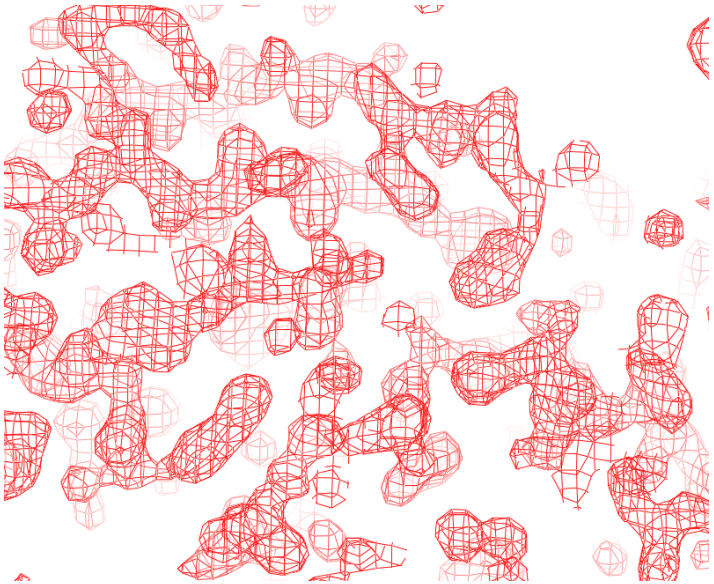
If the point is in the SOLVENT region, only values of electron density near zero are believable

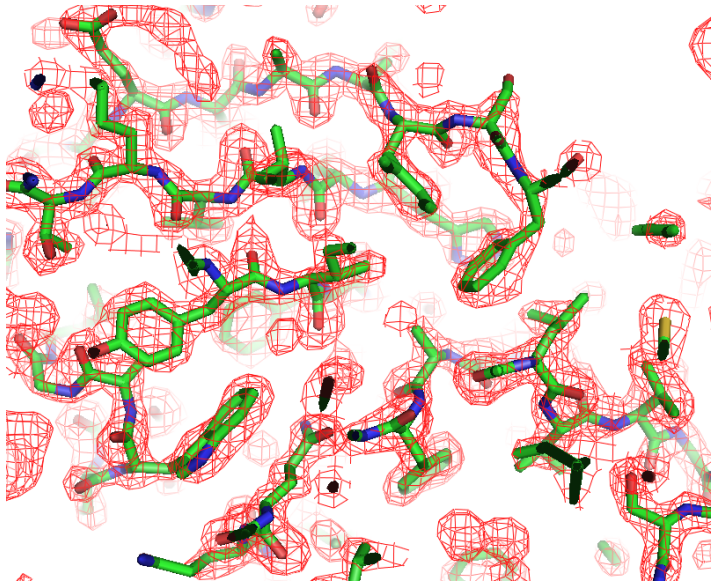
Statistical density modification
(nsf-N SAD map , 2Å, no NCS, 50% solvent)

Phaser SAD map
(CC=0.43)

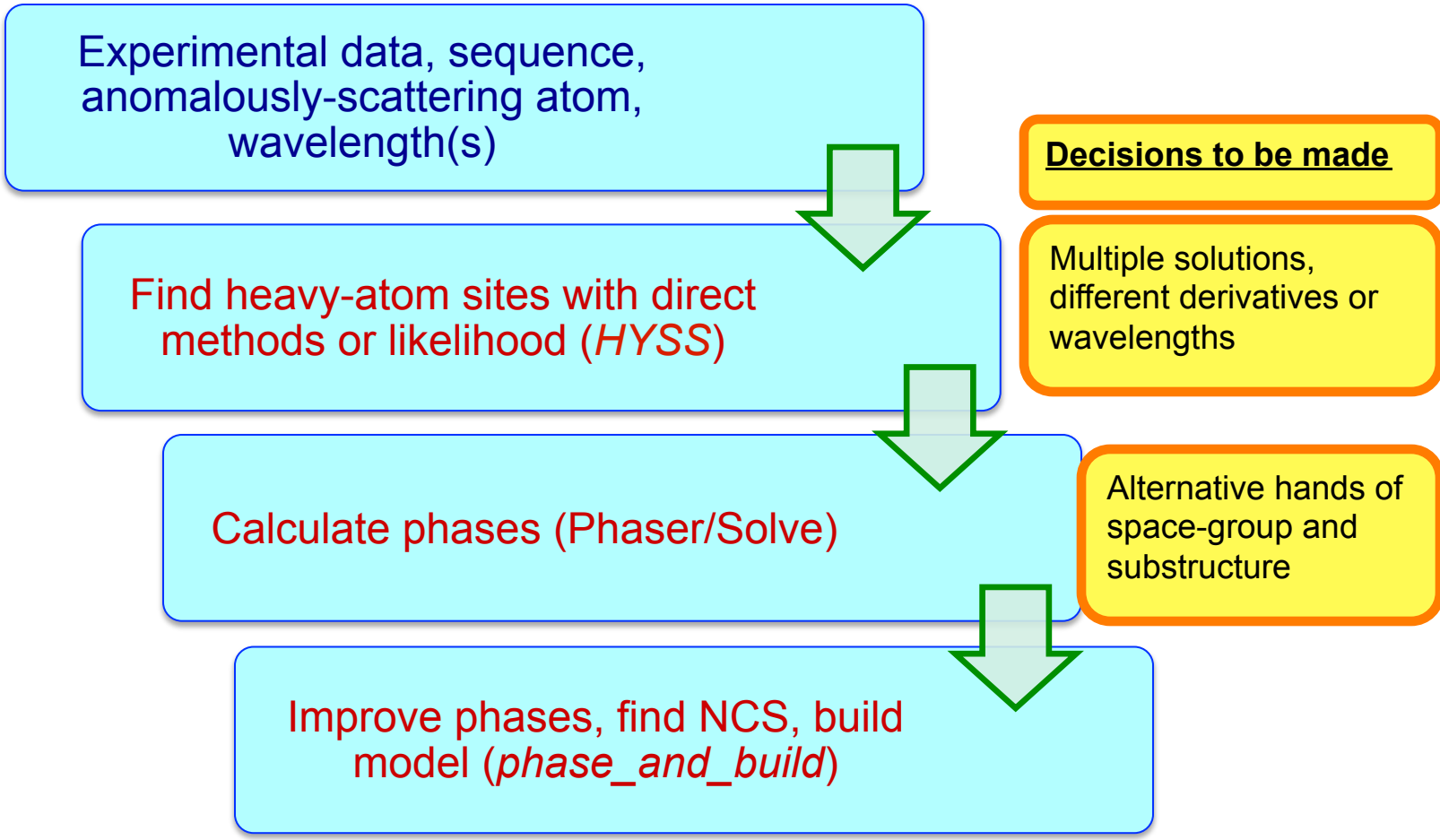


Phaser +RESOLVE
(CC=0.79)

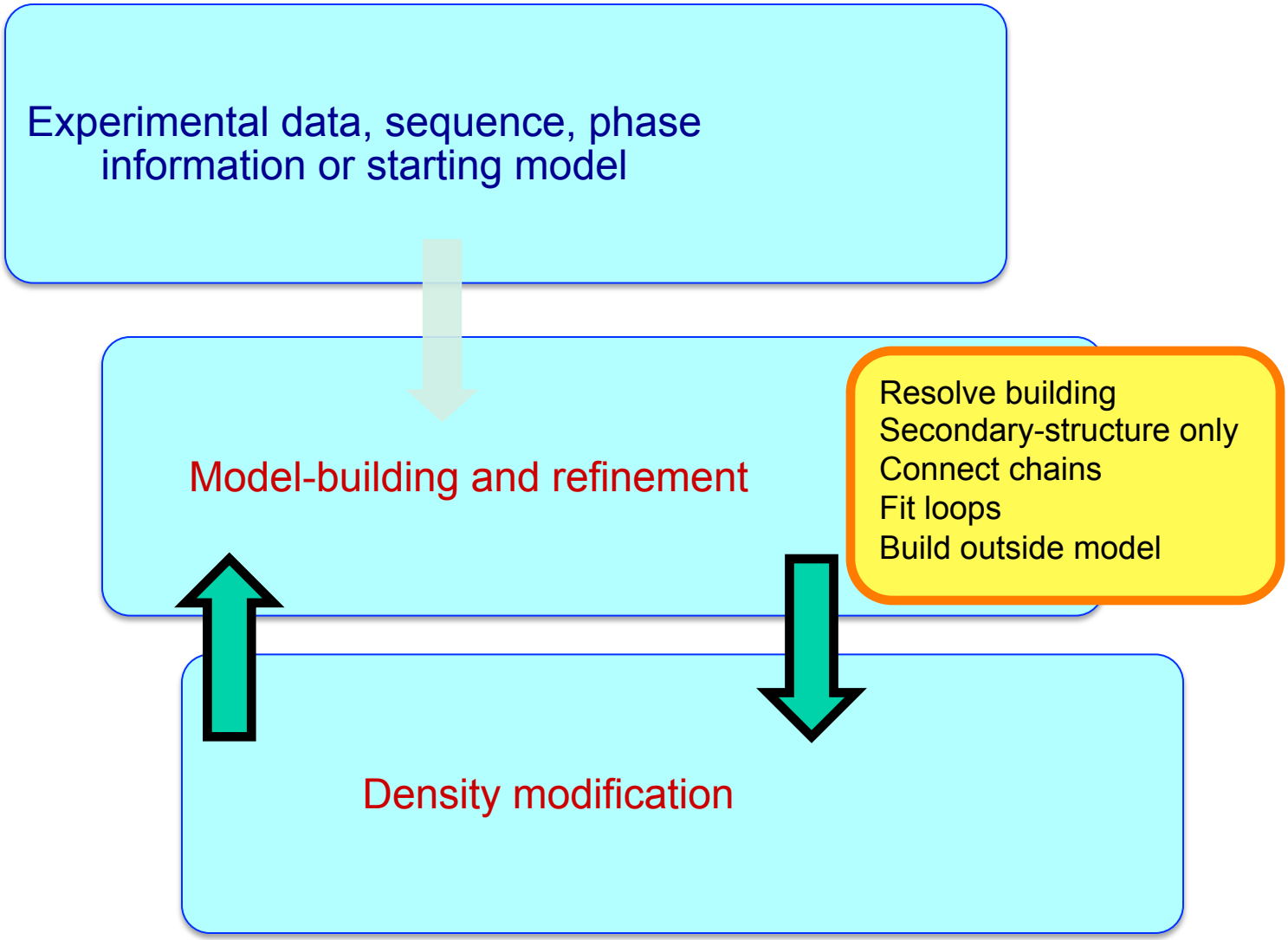




Structure solution with *phenix.autosol*



Iterative density modification, model-building and refinement with *phenix.autobuild*



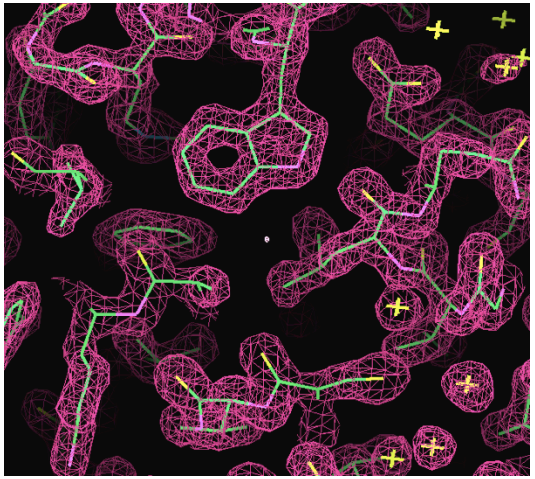
Model-building at moderate or high resolution

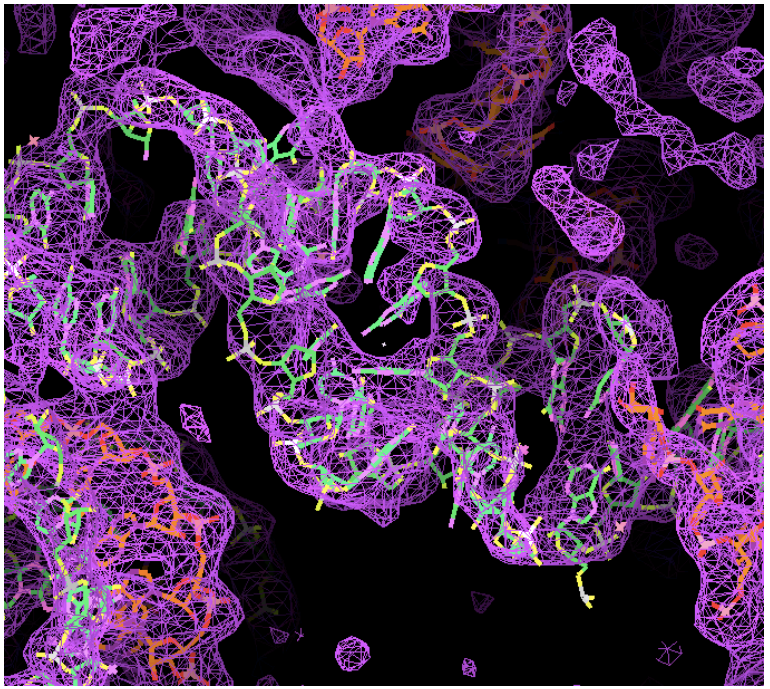
Phenix

•FFT-based identification of regular secondary structure

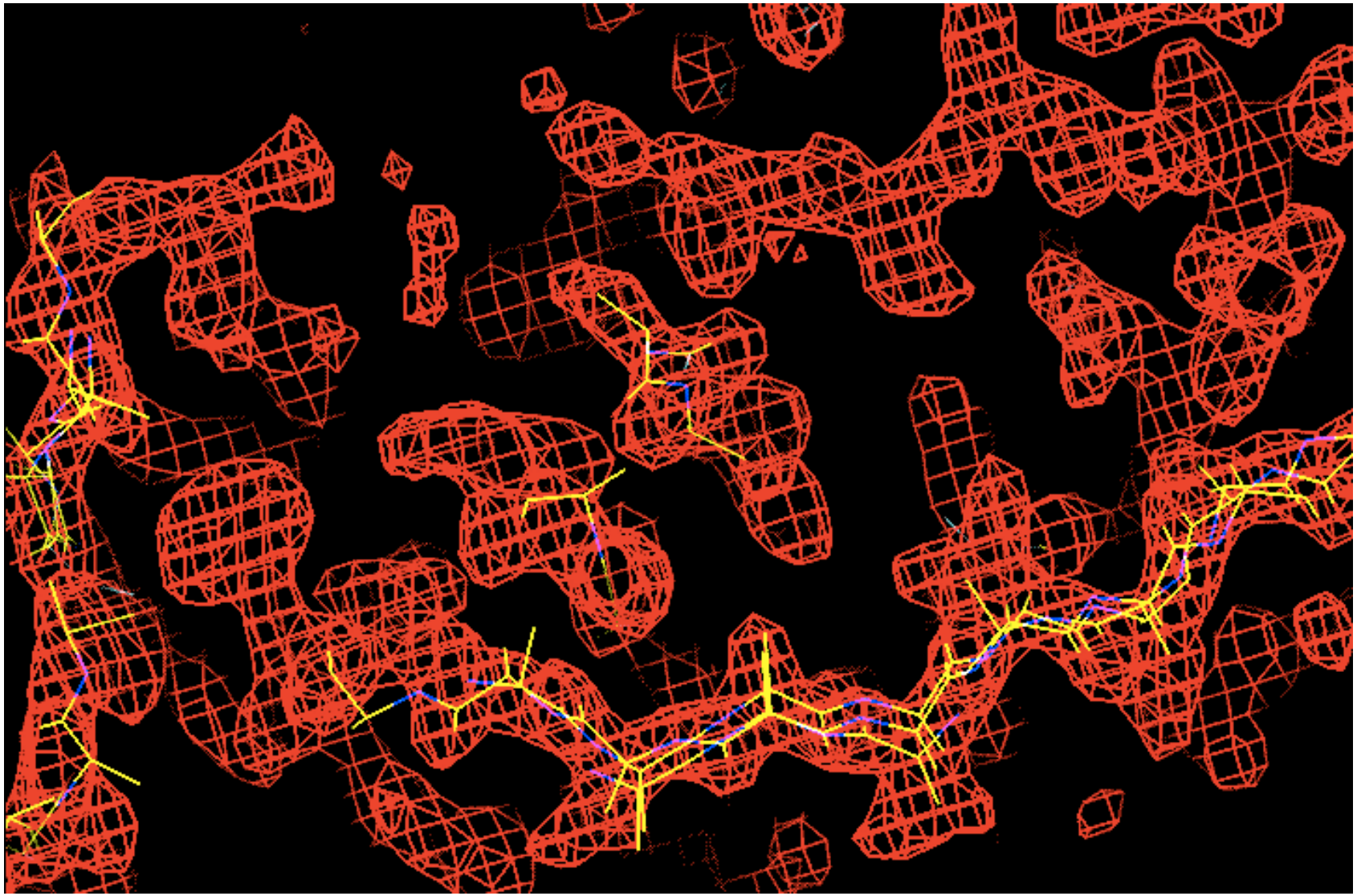
•Extension with short fragments from high-resolution structures

•Probabilistic sequence alignment

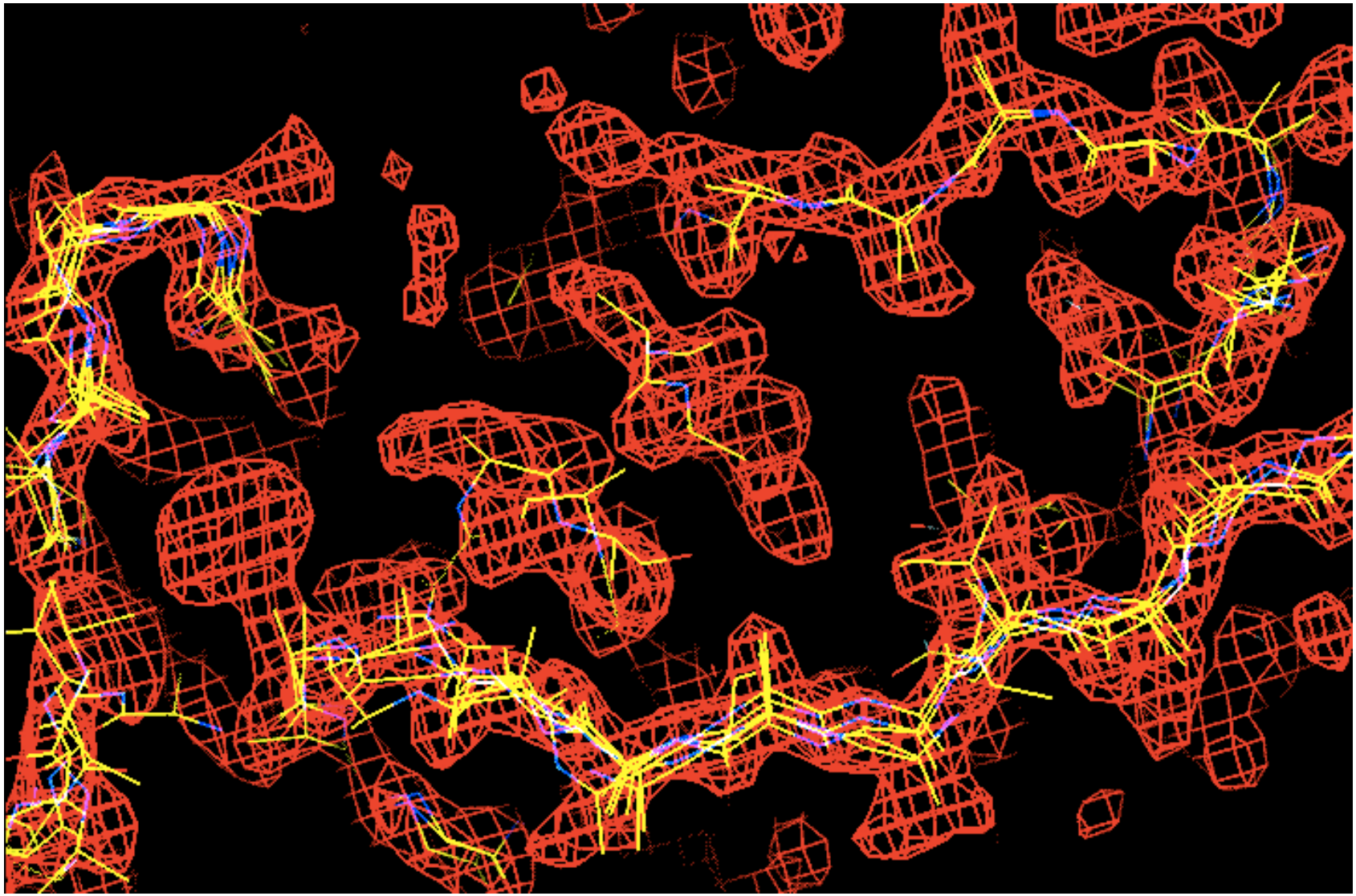




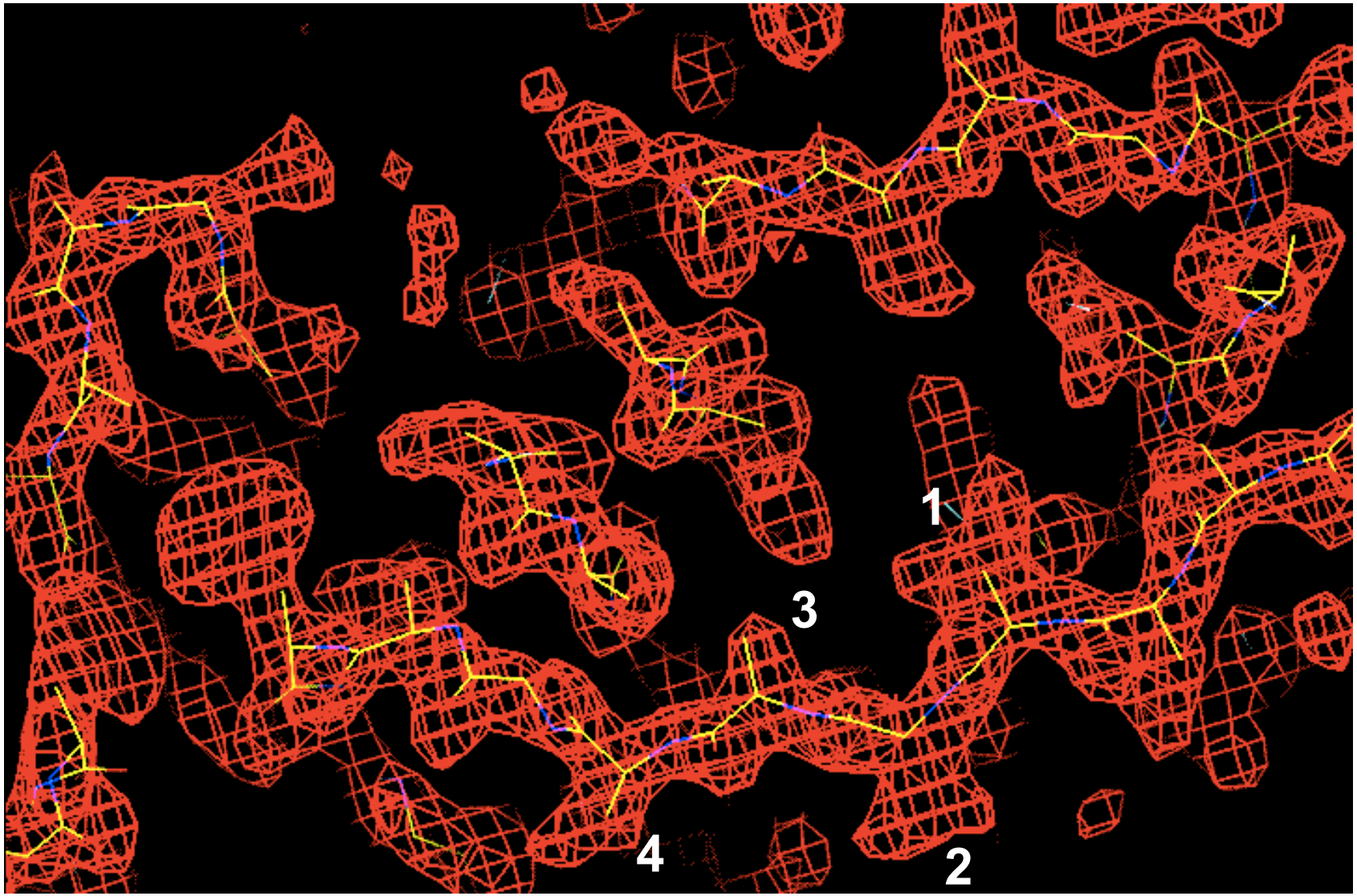
Initial model-building – strand fragments



Chain extension
(result: many overlapping fragments)



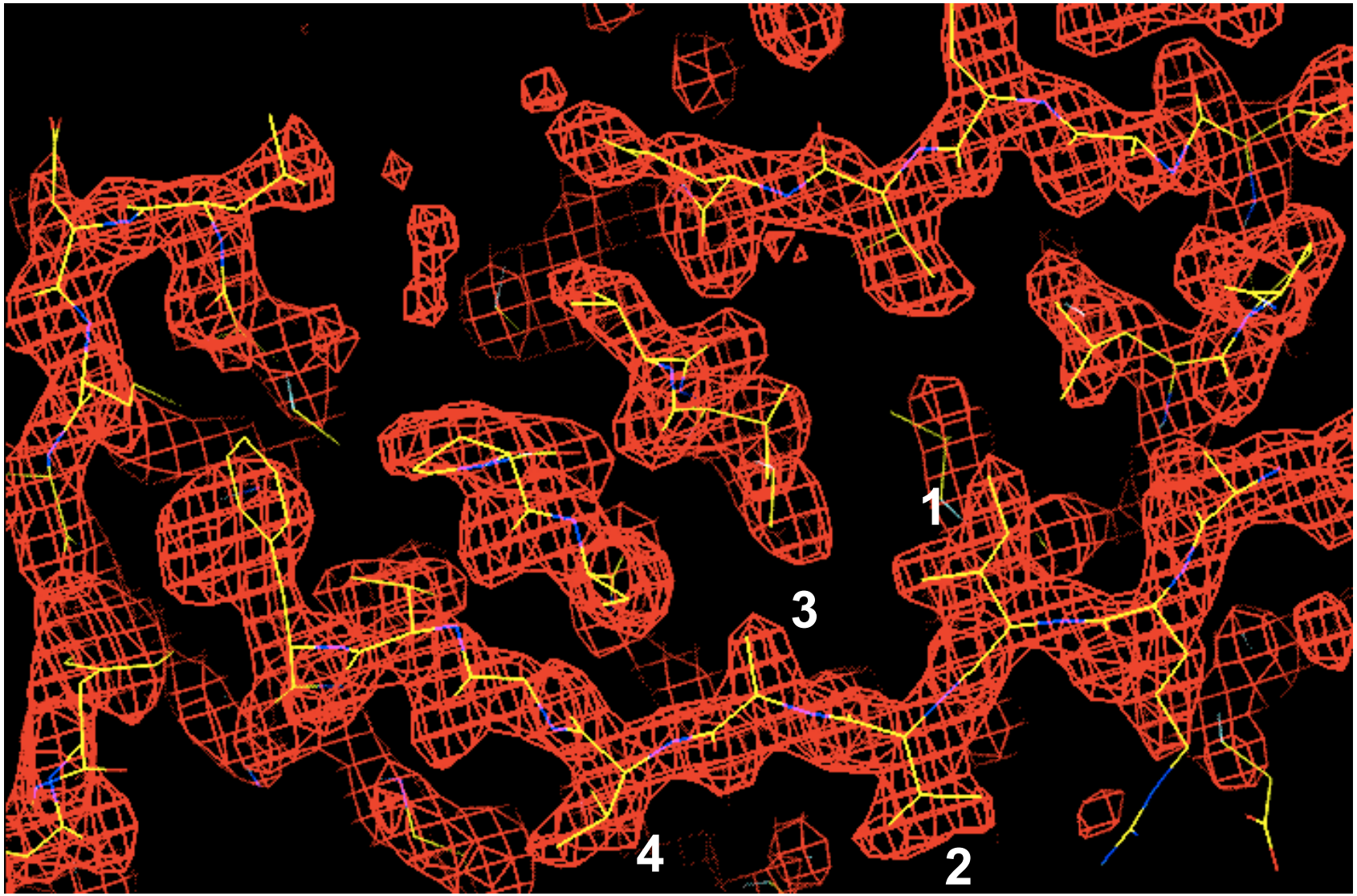
Main-chain as a series of fragments (choosing the best fragment at each location)



Side-chain template matching to identify sequence alignment to map (IF5A data)
Relative probability for each amino acid at each position
(Correct amino acids in bold)

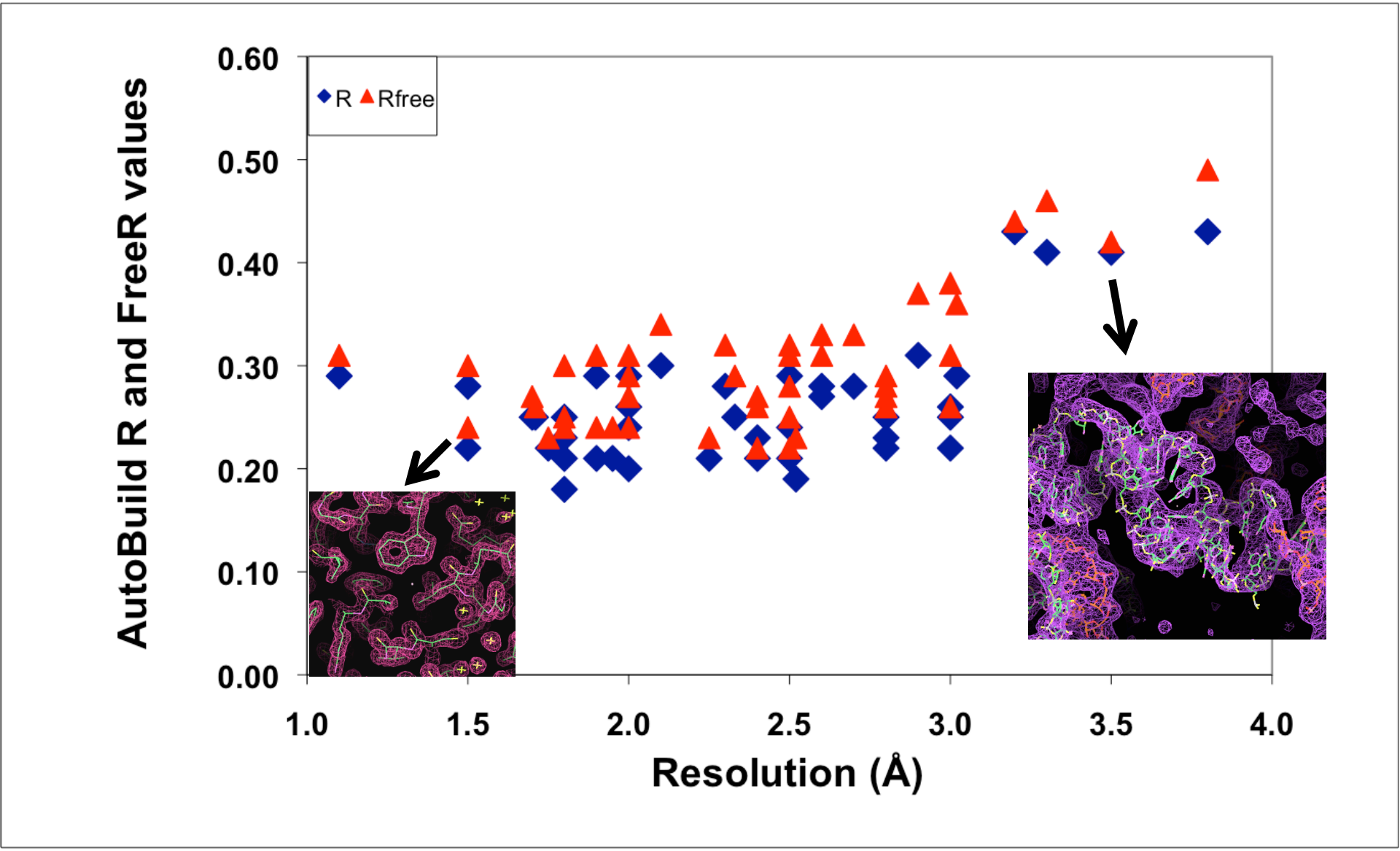
#	G	A	S	V	I	L	M	C	F	Y	K	R	W	H	E	D	Q	N	P	T
1	6	5	4	18	**18**	6	1	1	1	2	6	2	2	1	9	6	1	0	1	4
2	4	11	14	**37**	5	2	0	2	0	0	2	3	0	0	1	2	0	0	0	6
3	11	**23**	5	12	5	3	2	0	1	3	7	3	1	0	5	3	2	0	2	2
4	7	9	6	**16**	8	5	2	0	1	3	8	4	1	0	7	6	2	0	3	4
5	**31**	7	3	7	4	2	1	0	1	3	5	4	1	0	6	2	2	0	11	1
6	1	3	3	**41**	14	8	0	0	0	0	2	1	0	0	2	4	0	0	1	9
7	0	0	0	0	0	0	0	0	**15**	**63**	1	0	**17**	1	0	0	0	0	0	0
8	2	3	6	23	10	6	2	1	0	1	4	3	0	0	5	**16**	1	0	1	6
9	**96**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Addition of side-chains to fixed main-chain positions



AutoBuild – tests with structure library

Fully automated iterative model-building, final R/Rfree



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