

SAD Phasing and automated structure determination

Phenix workshop

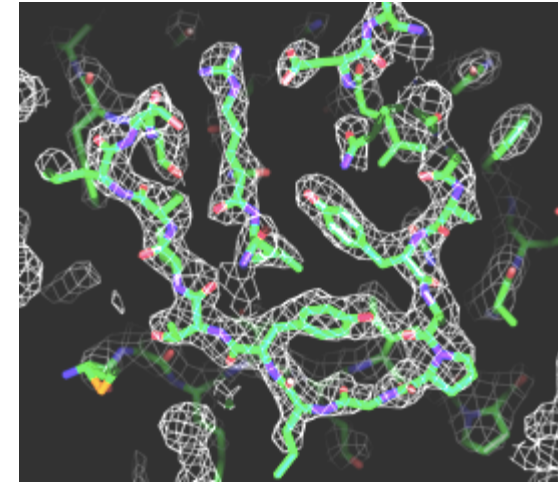
Shanghai, China
Jan. 14, 2016

Tom Terwilliger
Los Alamos National Laboratory



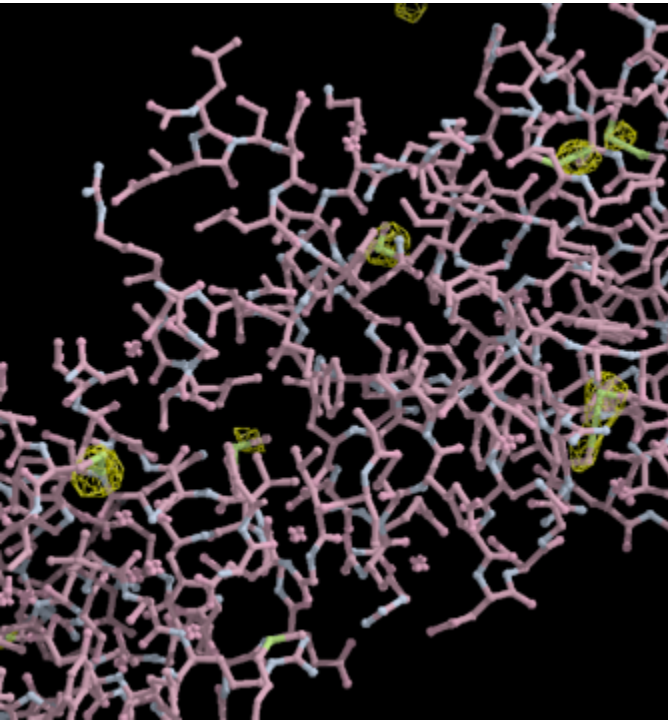
Steps in Single Wavelength Anomalous Diffraction (SAD) Structure Determination

- **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**



Planning a SAD experiment

Will I find the sites of anomalously-scattering atoms?



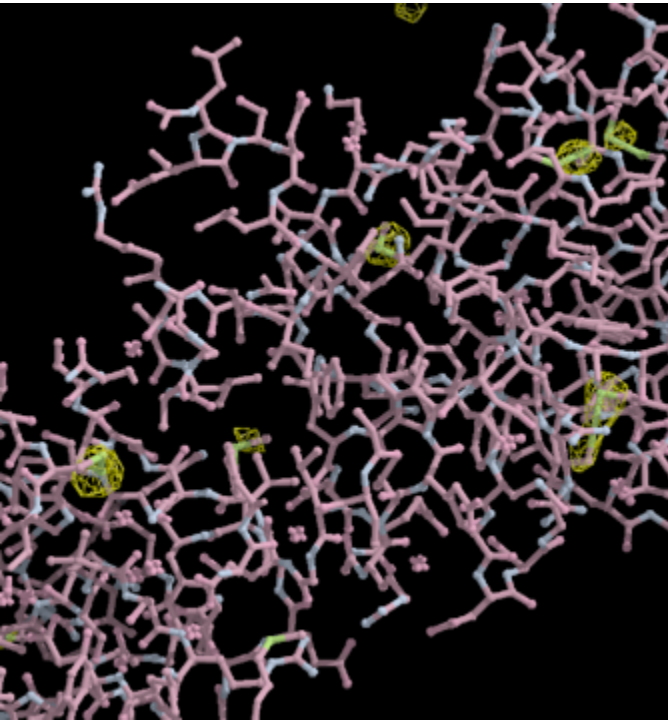
Planning a SAD experiment

How many sites?

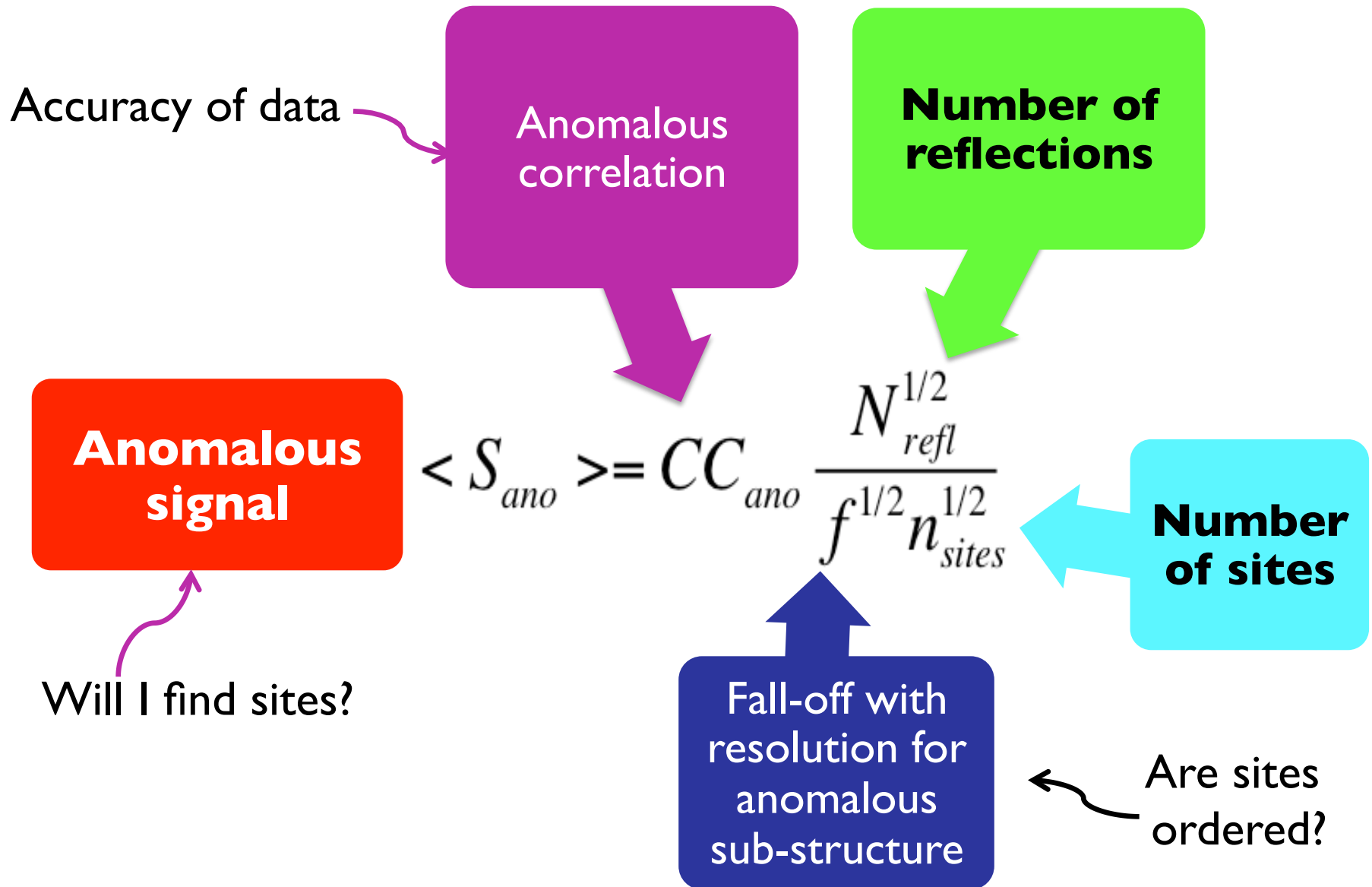
How many reflections?

Are the sites (on average) well ordered?

Are the data well-measured?



What determines if I will find the sites?



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

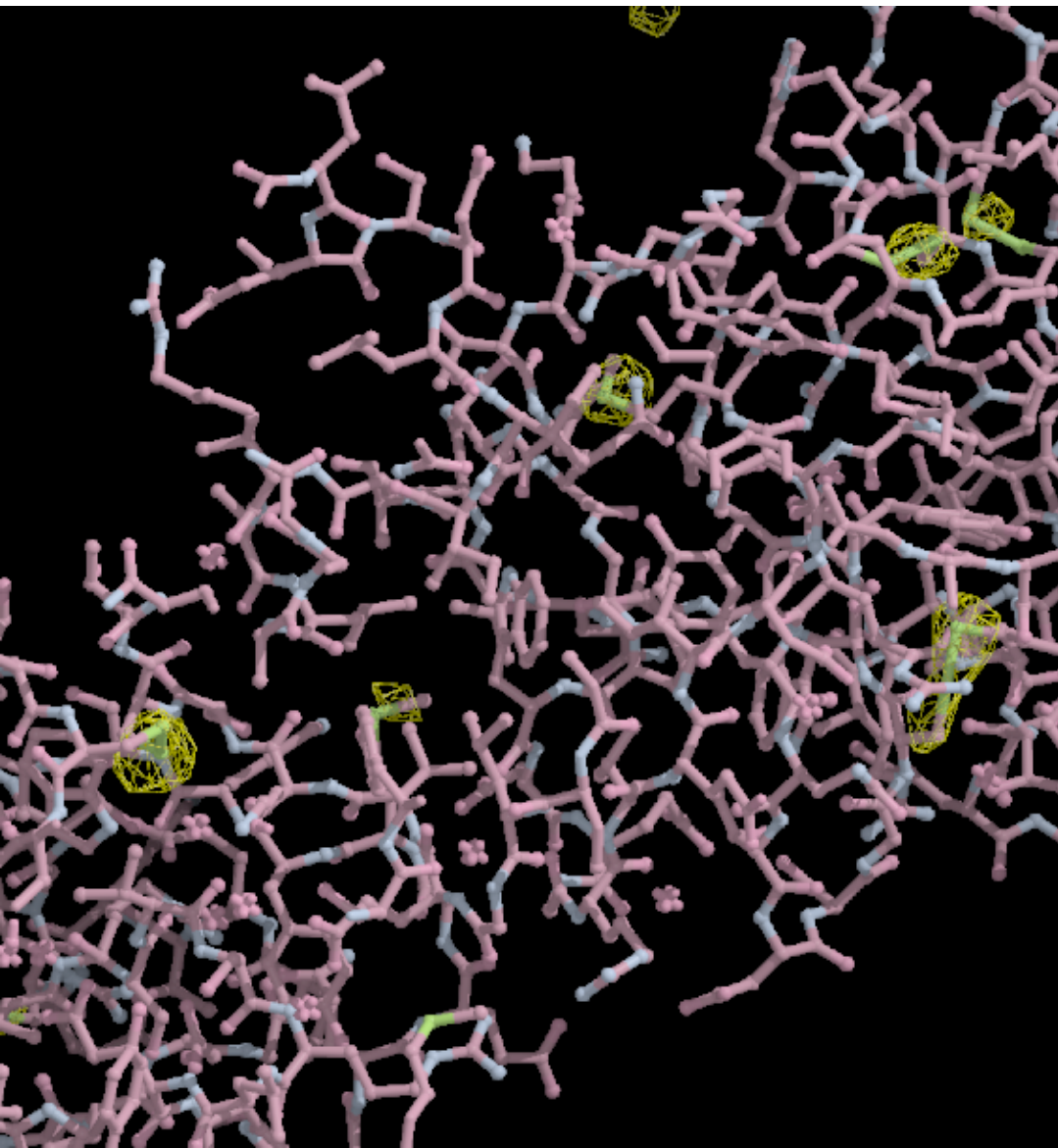
Anomalous correlation: accuracy of anomalous differences

Correlation of observed and
sub-structure anomalous
differences

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

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

Anomalous signal: peak height in anomalous difference Fourier at coordinates of anomalously-scattering atoms



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

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}

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

f is 2nd moment of the
anomalous scattering factor
(accounts for weak high-resolution data)

$$f = \frac{\langle (f^h)^2 \rangle}{\langle f^h \rangle^2}$$

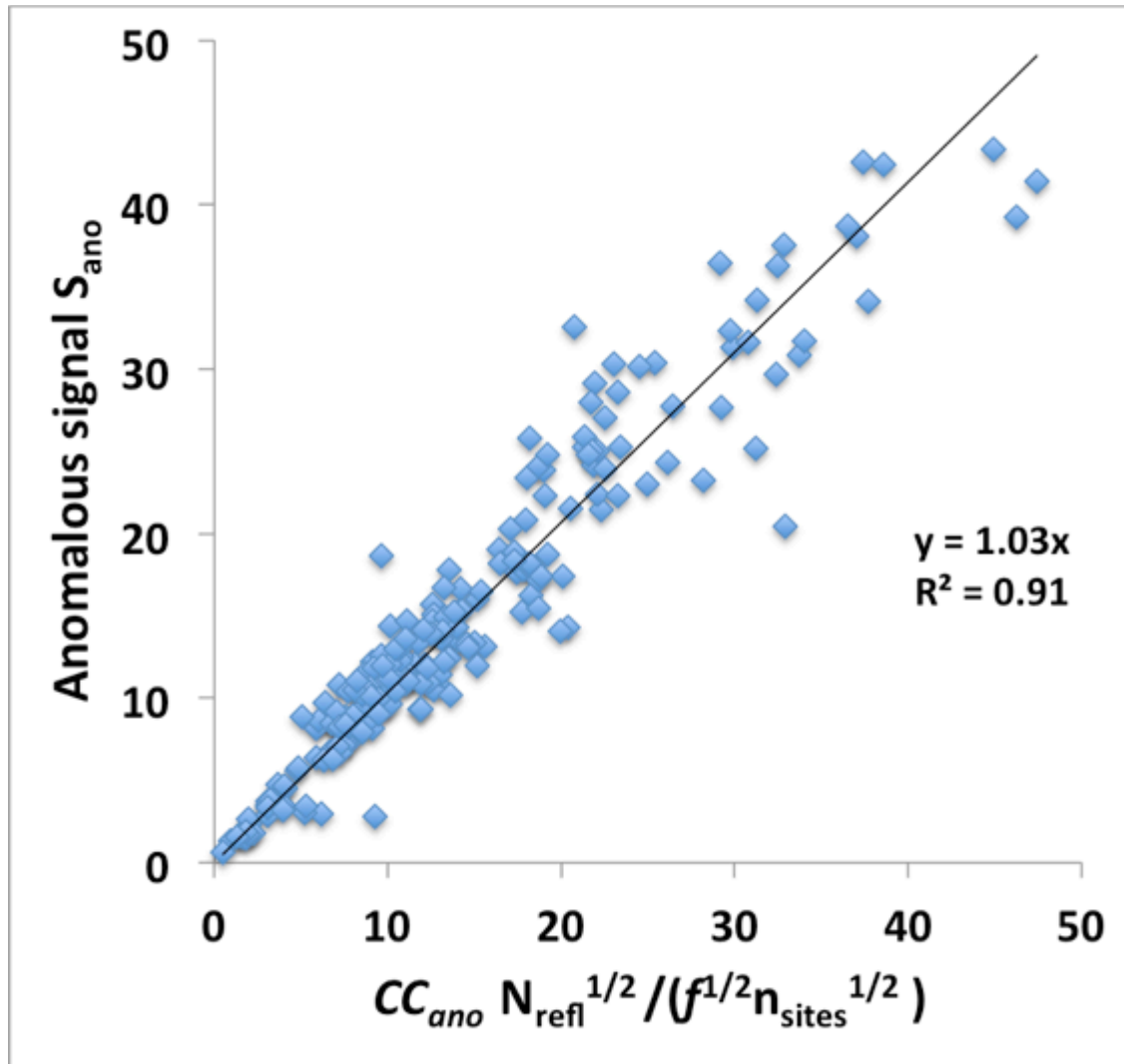
Anomalous scattering factor

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

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



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

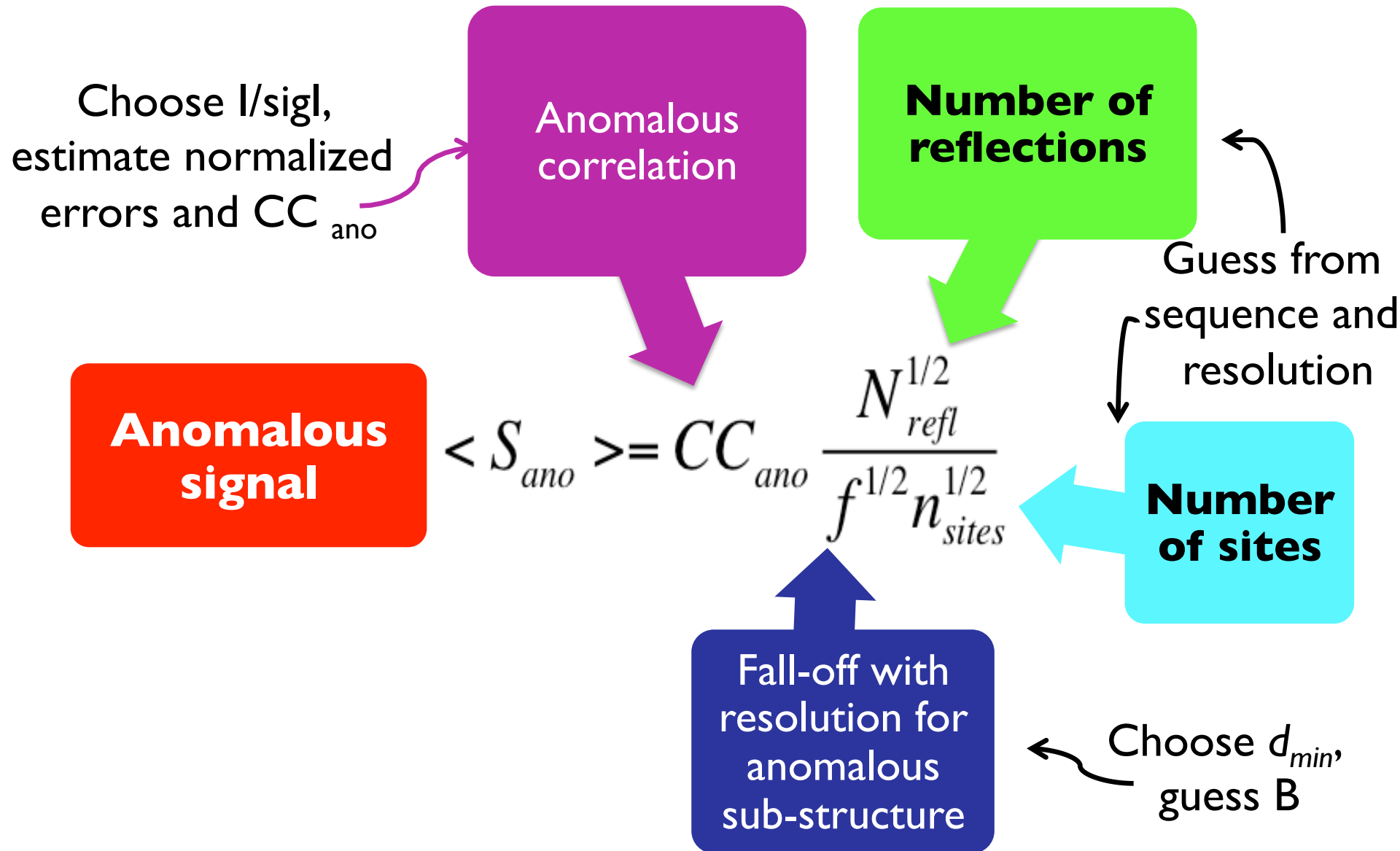
CC_{ano} : Correlation of anomalous differences with model differences

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



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.

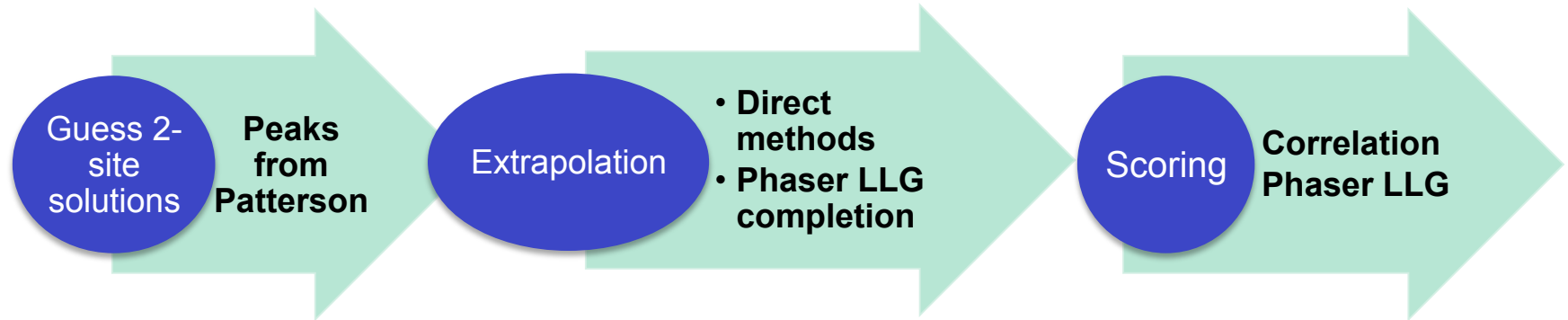
LLG sub-structure searches in Phenix

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

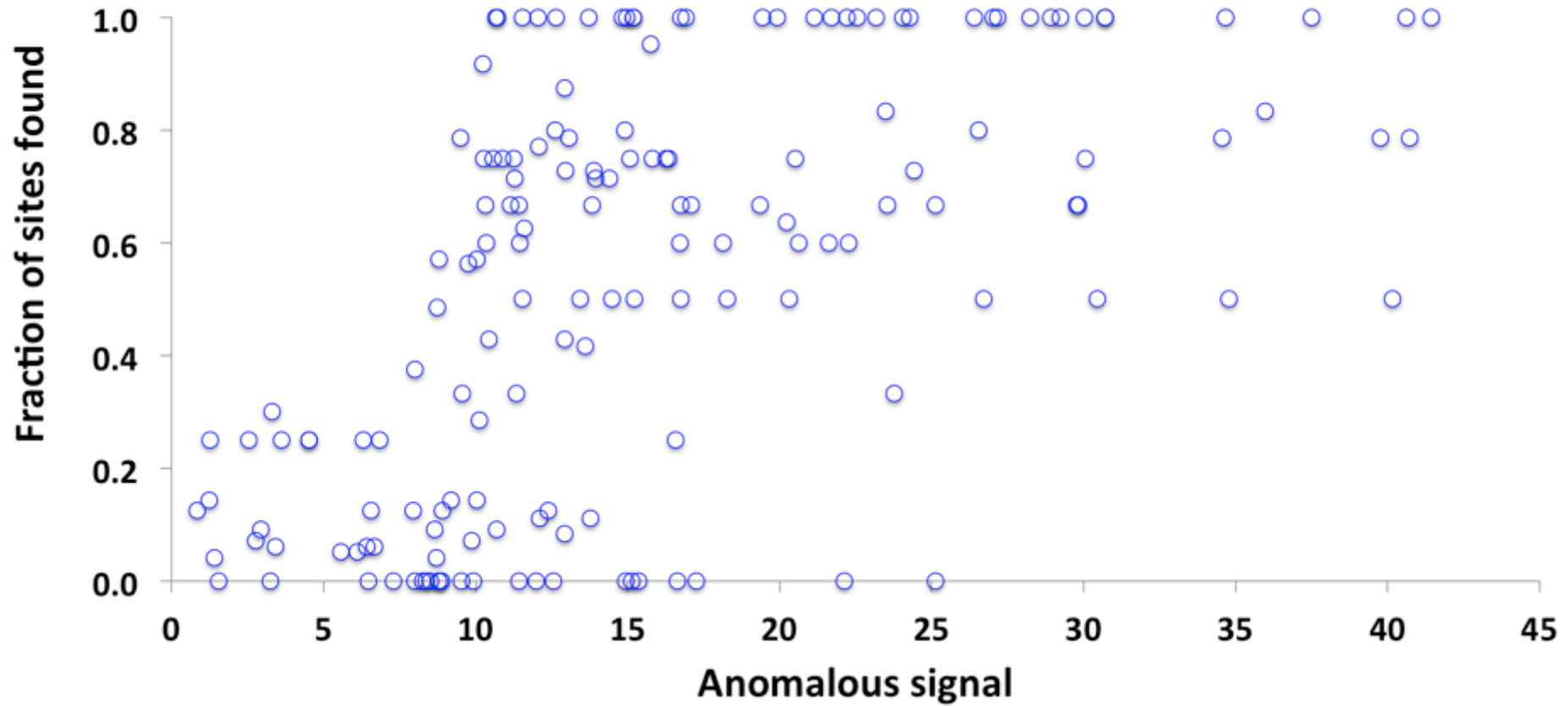
Adjustable
LLGC_SIGMA
(cut-off for peak height)

Use LLG score to
compare solutions

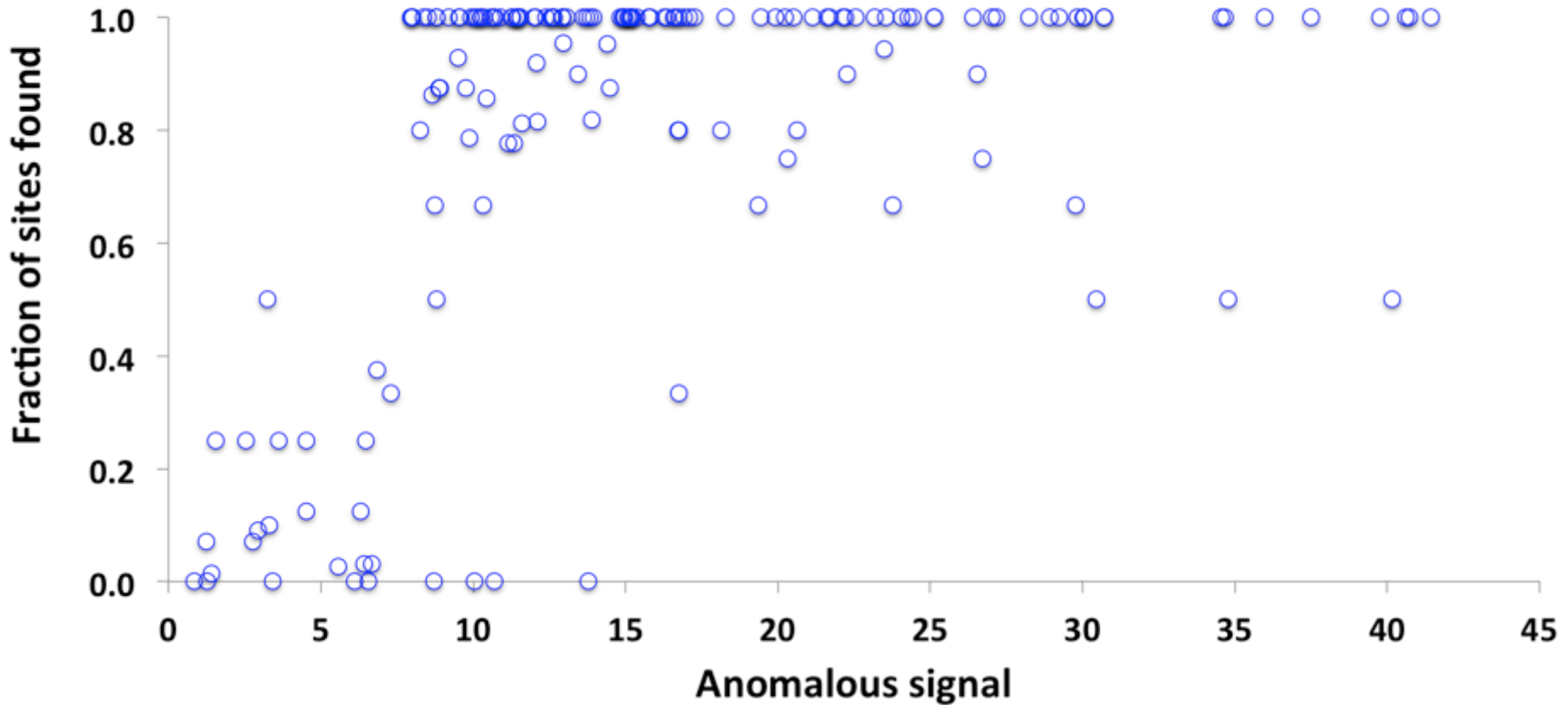
Terminate early if same
solution found several
times

Run quick direct
methods first

Dual Space Sub-structure Completion

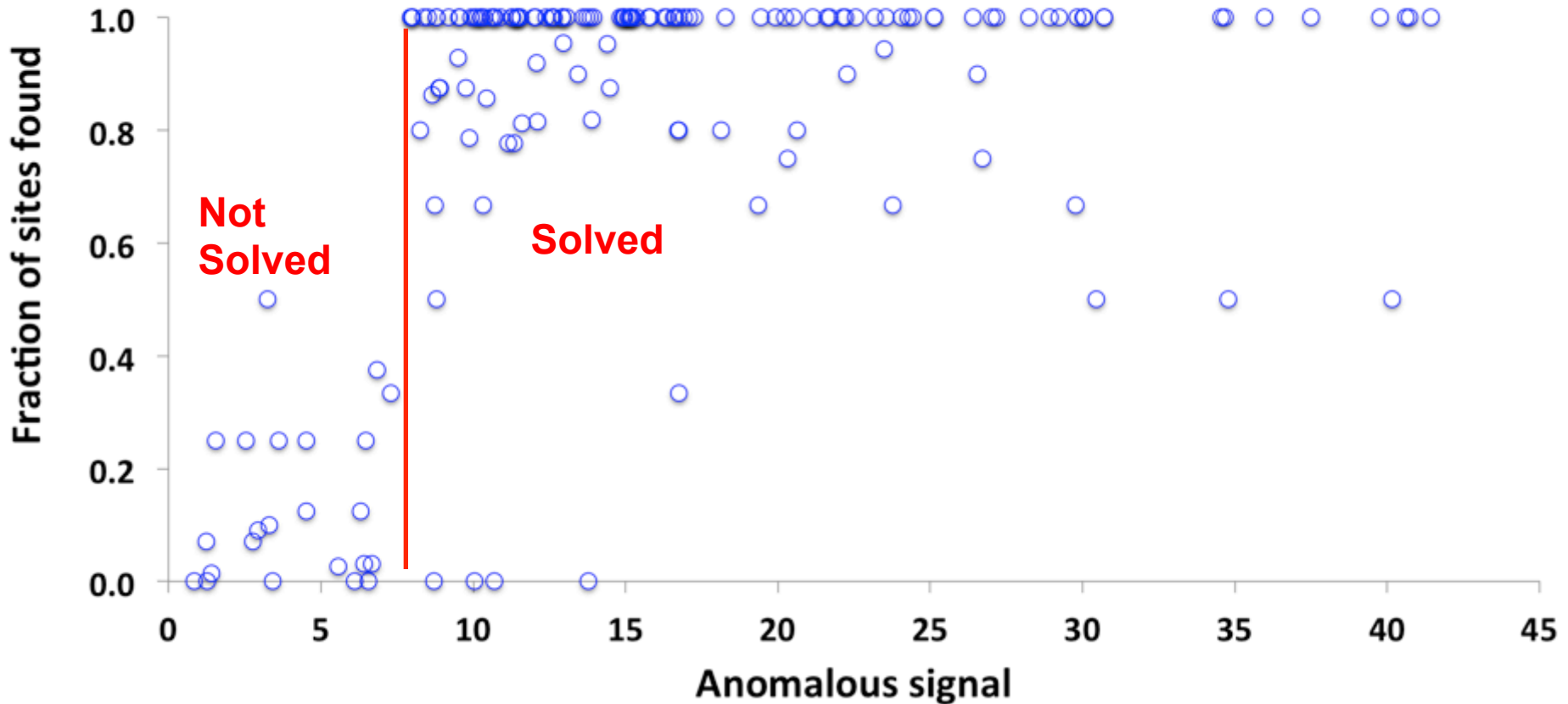


LLG Sub-structure Search



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

Anomalous signal indicates if a dataset can be solved



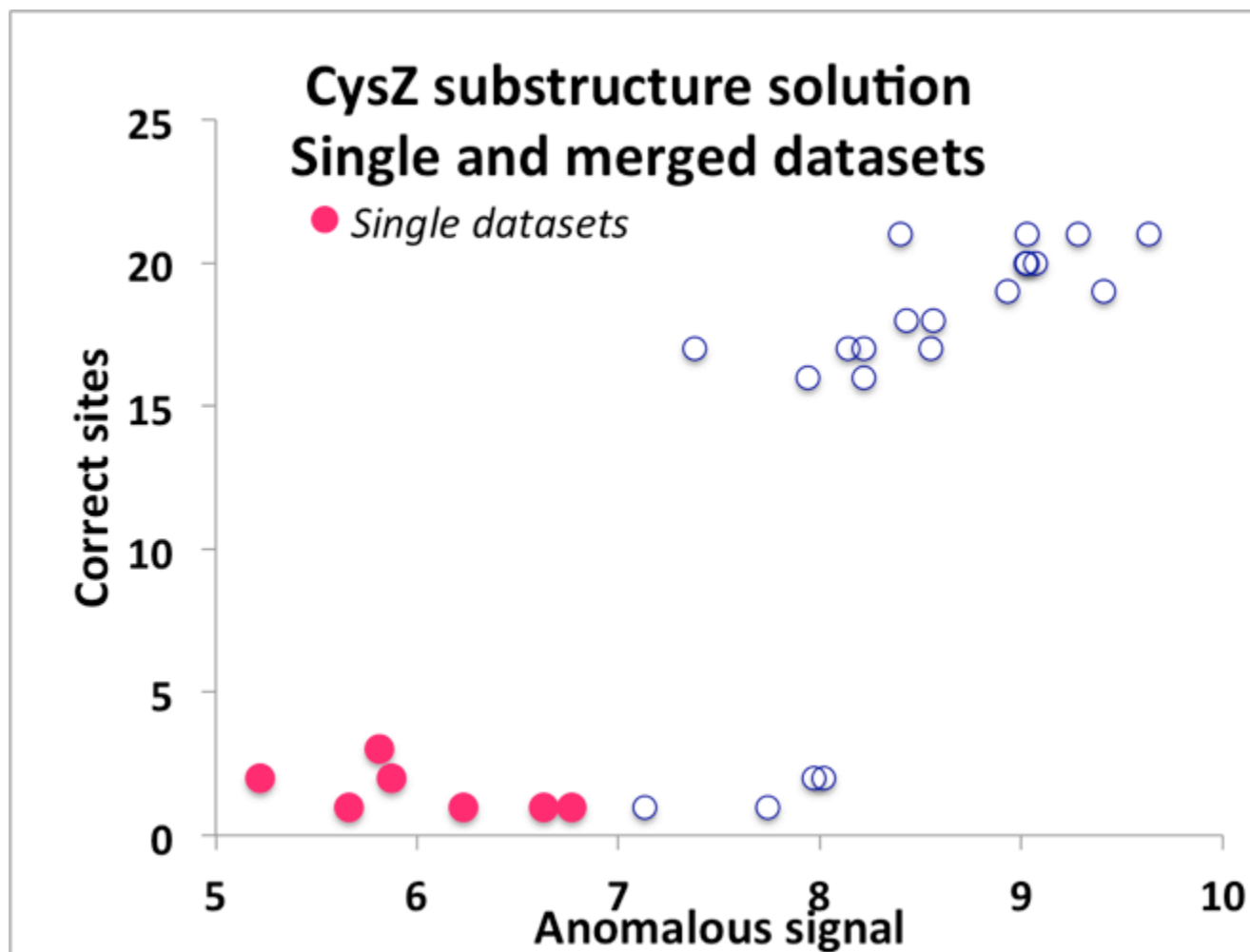
CysZ multi-crystal sulfur-SAD data

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

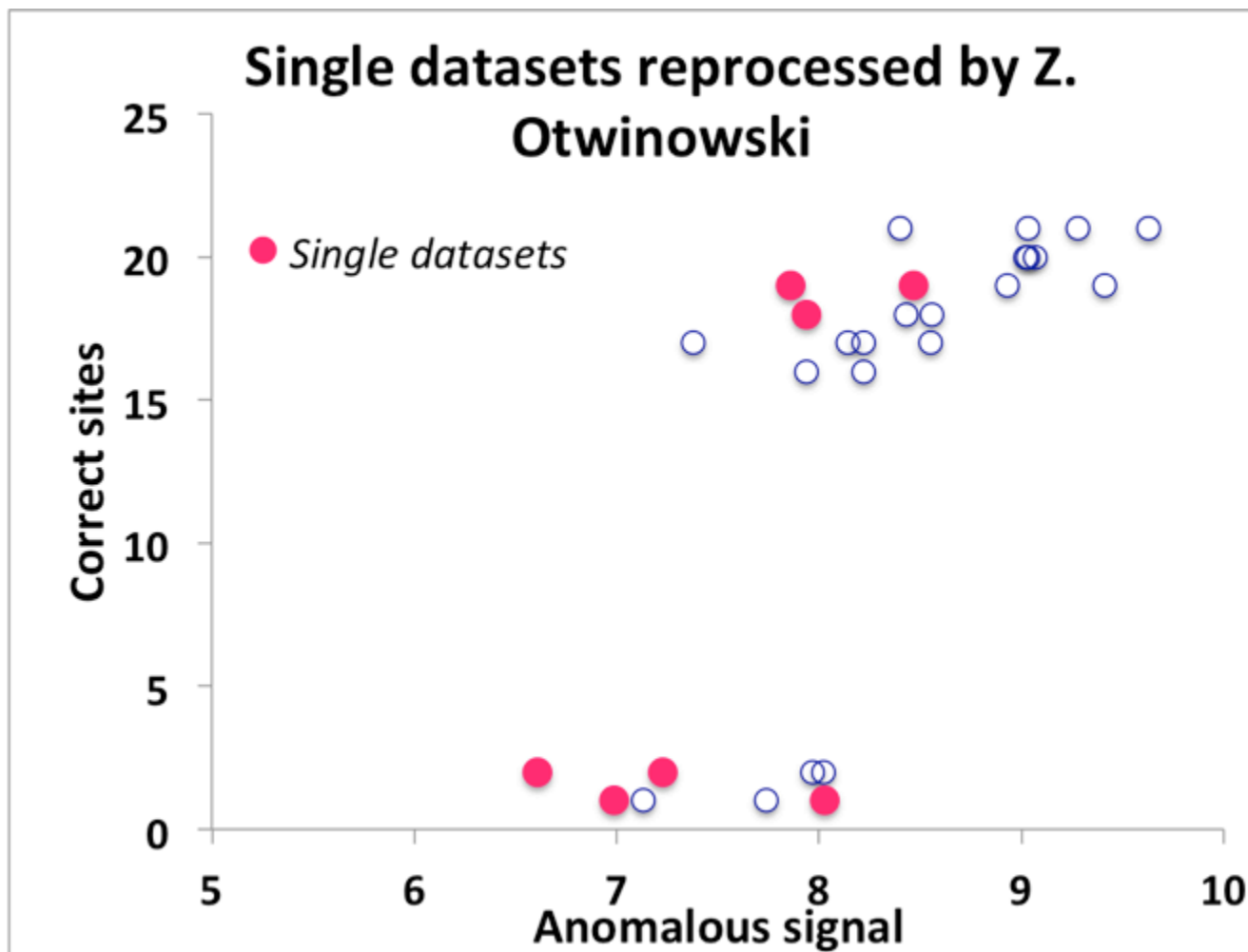
**Data from 7 crystals collected at wavelength of 1.74 Å to
resolution of 2.3 Å**

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

CysZ multi-crystal sulfur-SAD data

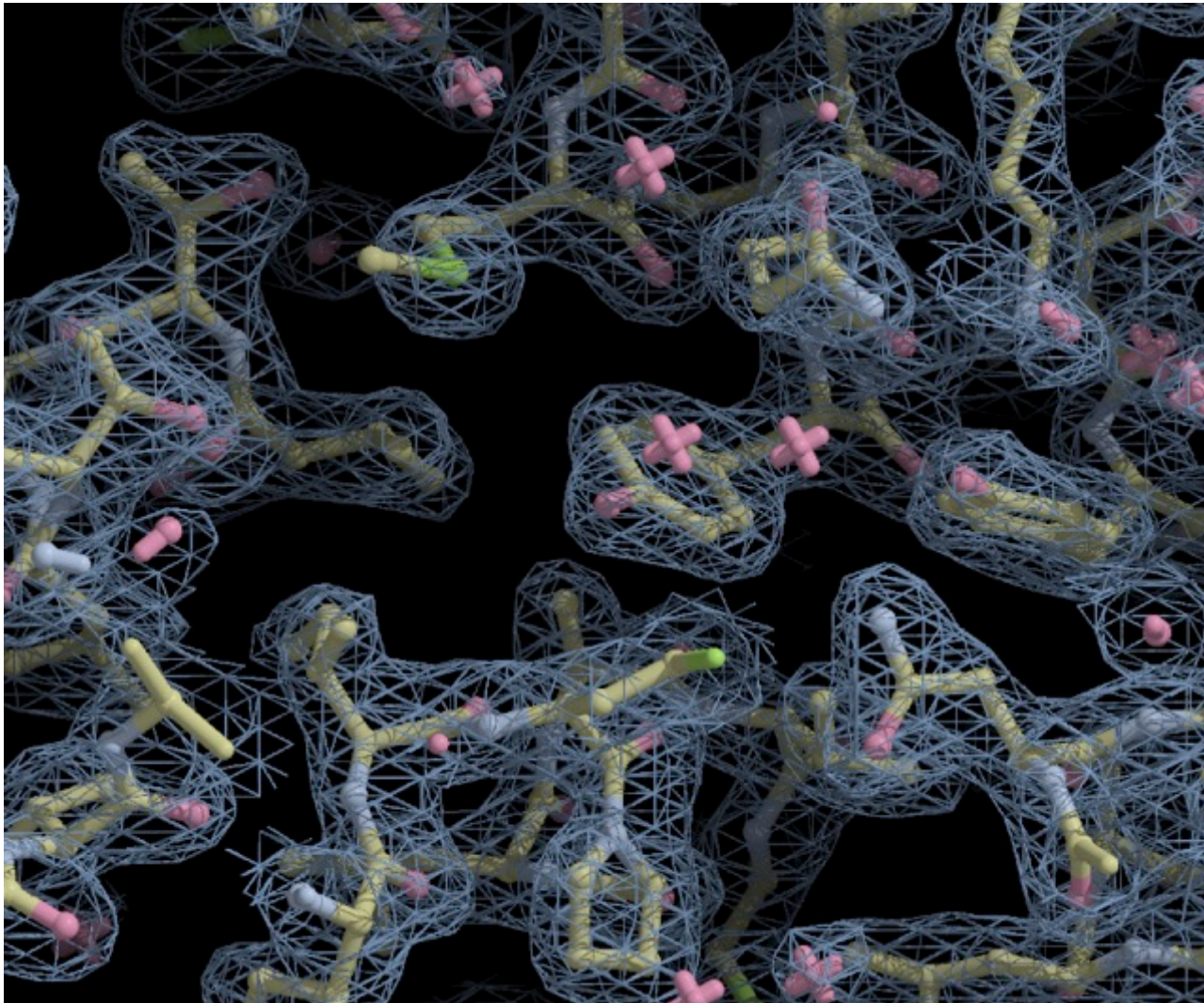


CysZ multi-crystal sulfur-SAD data



CysZ single-crystal sulfur-SAD data

Crystal 6 *AutoSol* $R/R_{\text{free}}=0.24/0.27$

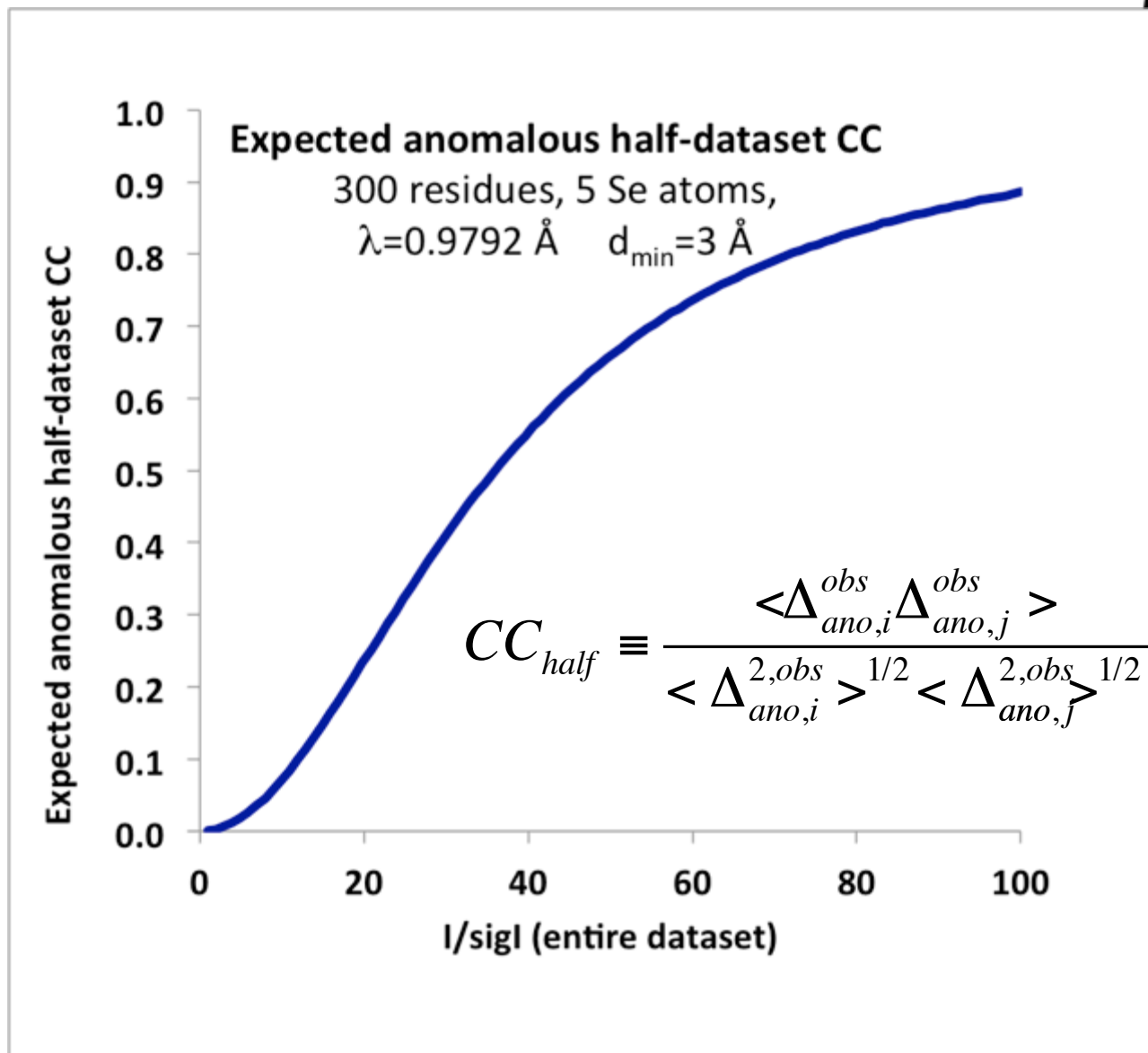


Will I solve my structure?

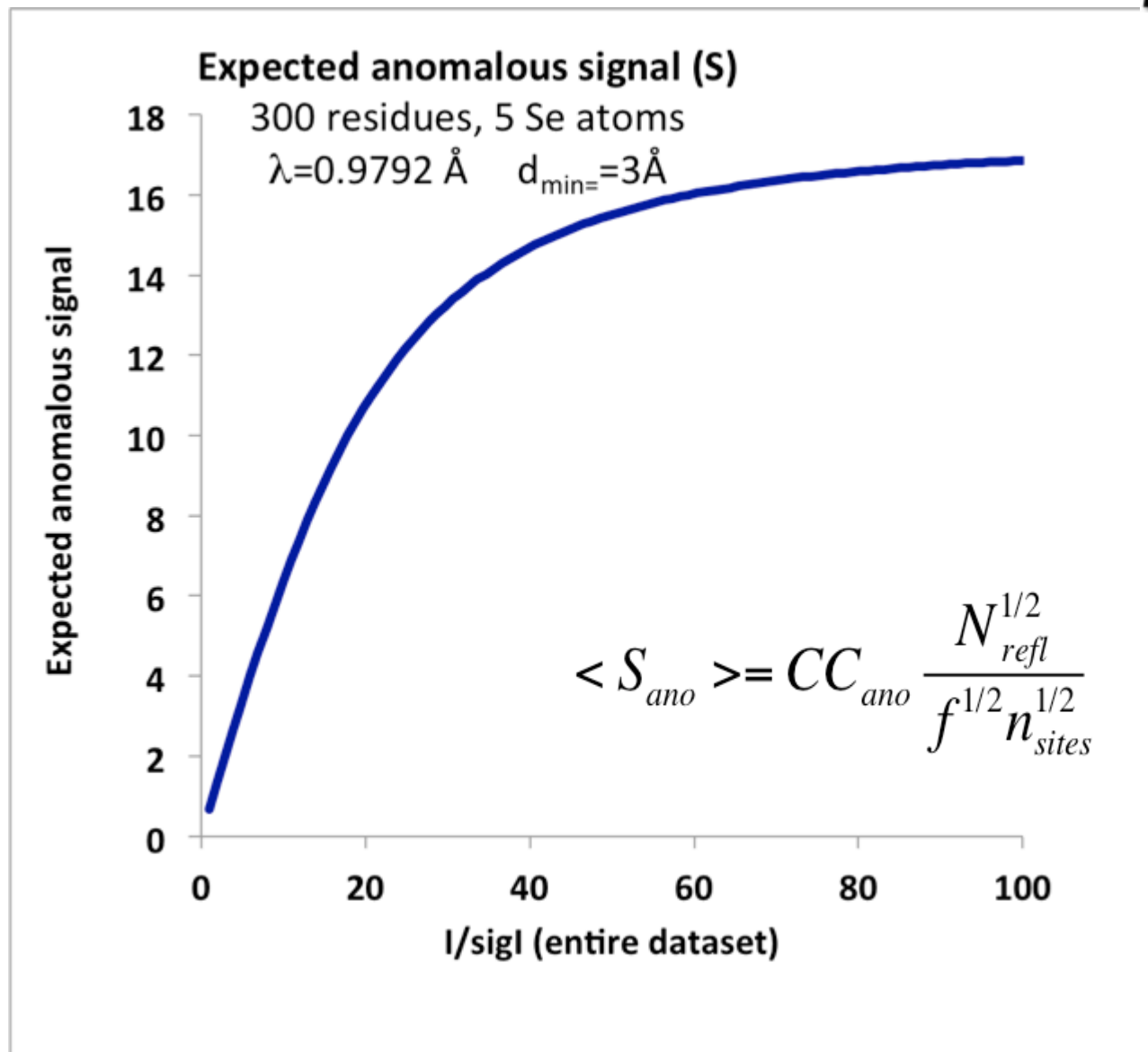
Simulate experiment with
phenix.plan_sad_experiment based on:

- I/σ (errors in measurement)
- Anomalous-scattering atom (f'')
- Sequence (other atoms)
- Resolution of data
- Number of sites

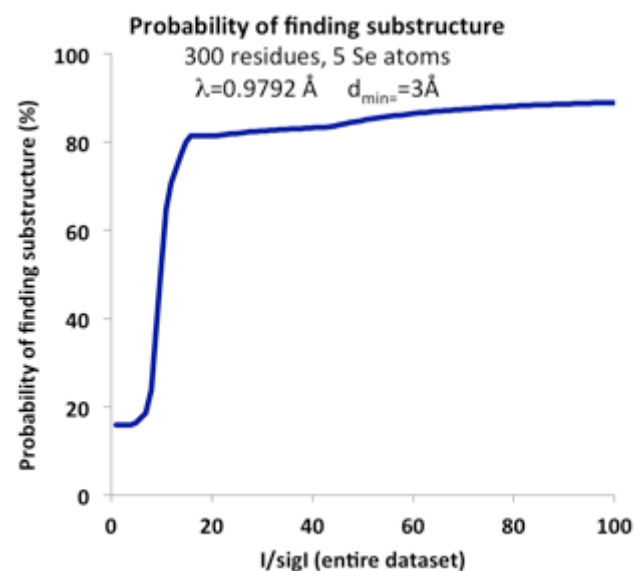
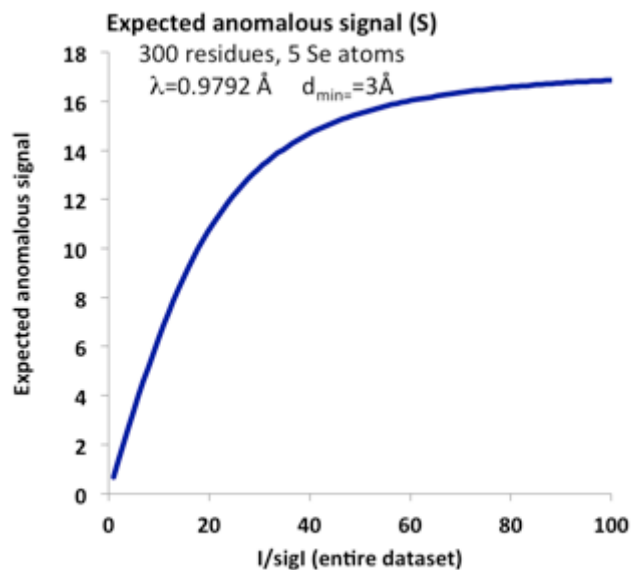
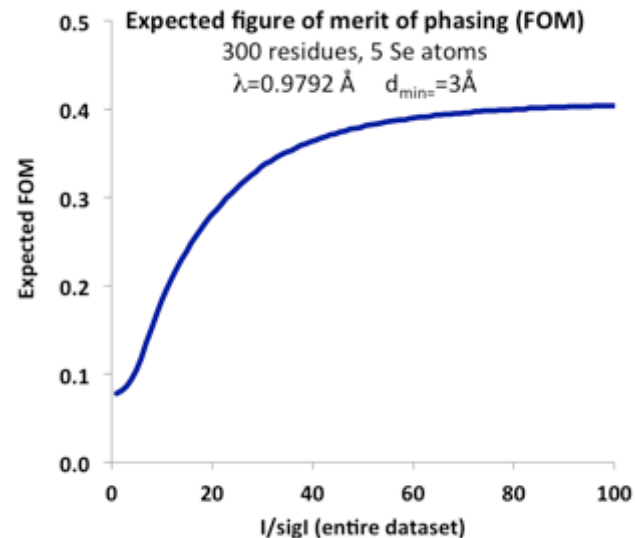
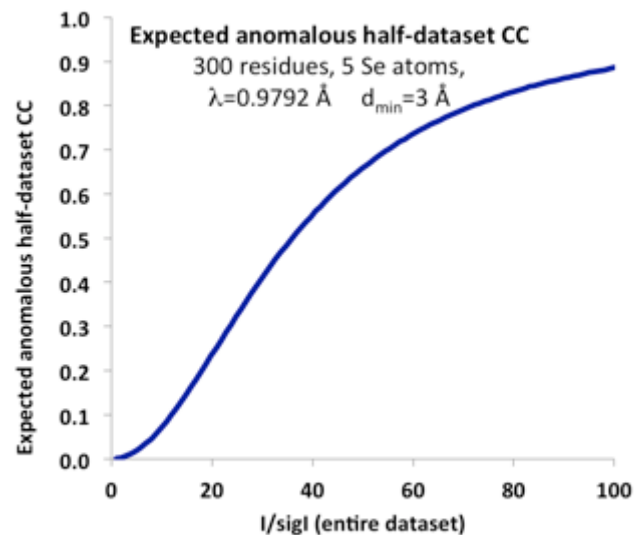
Anomalous data quality depends on I/σ_I



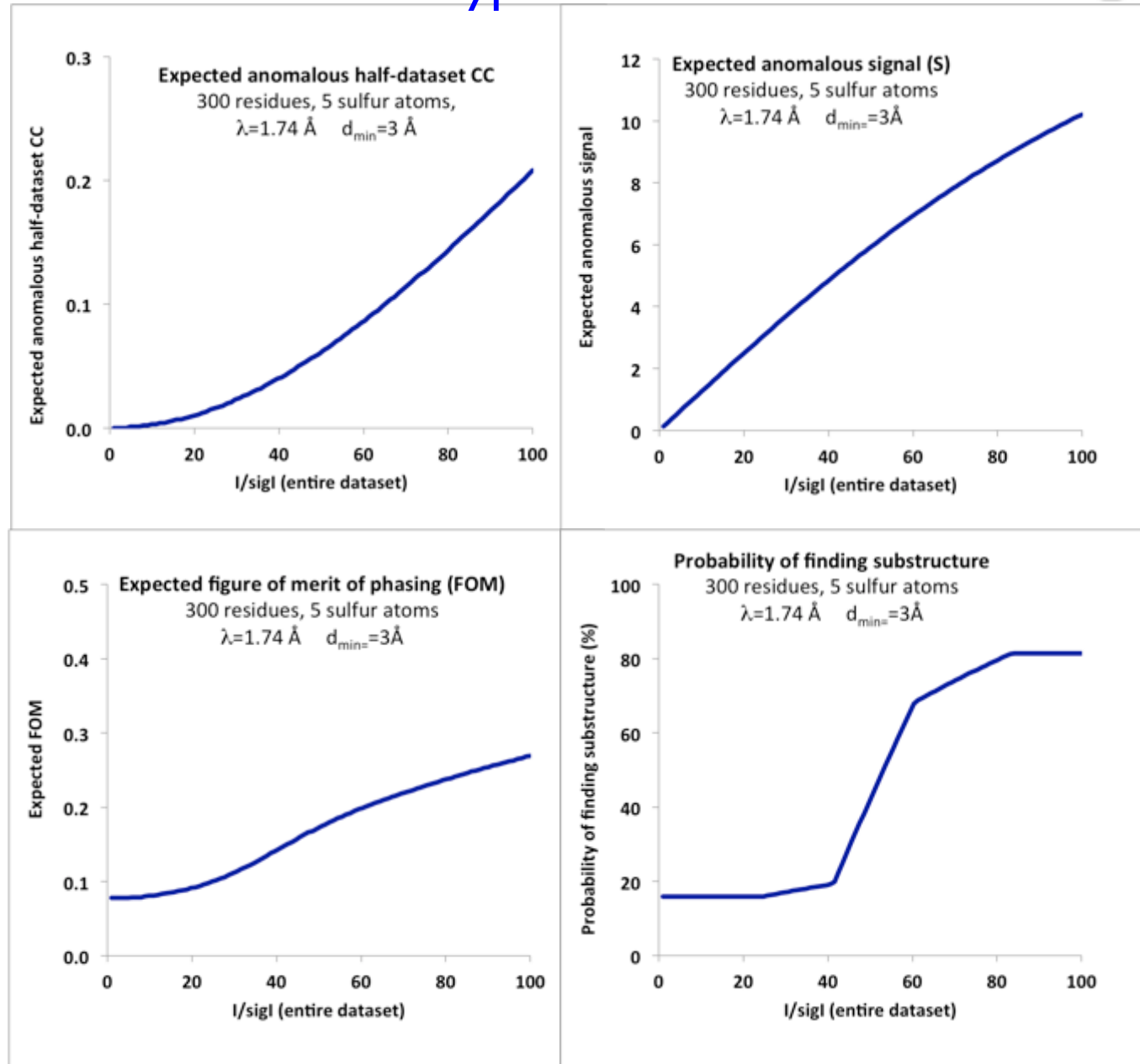
Anomalous data quality depends on I/sigI



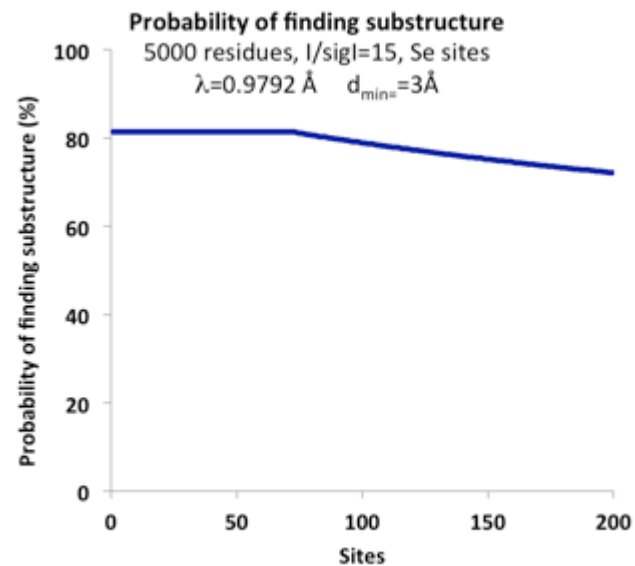
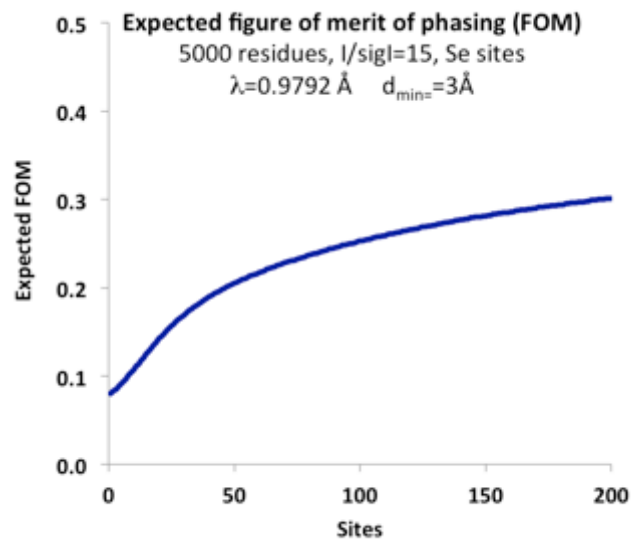
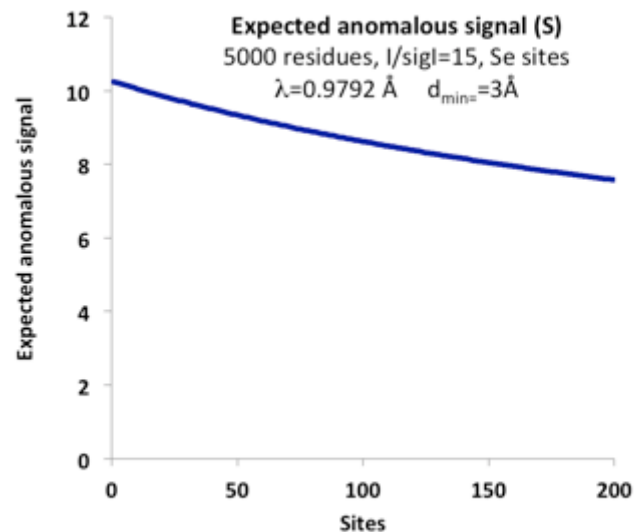
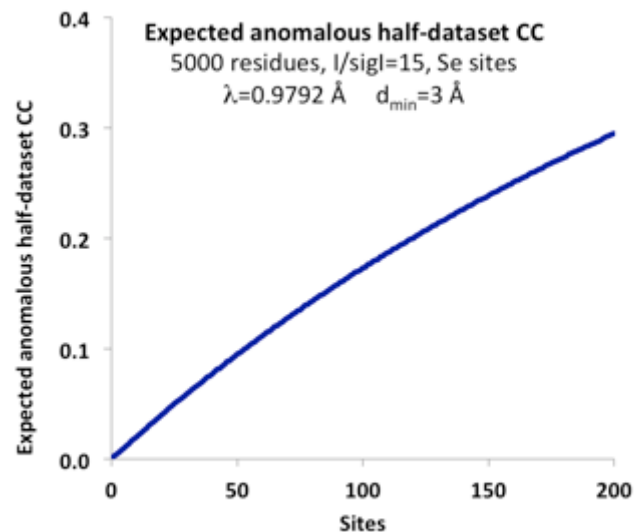
Anomalous data quality depends on $I/\sigma I$



Anomalous data quality depends on I/σ_I ... and atom type

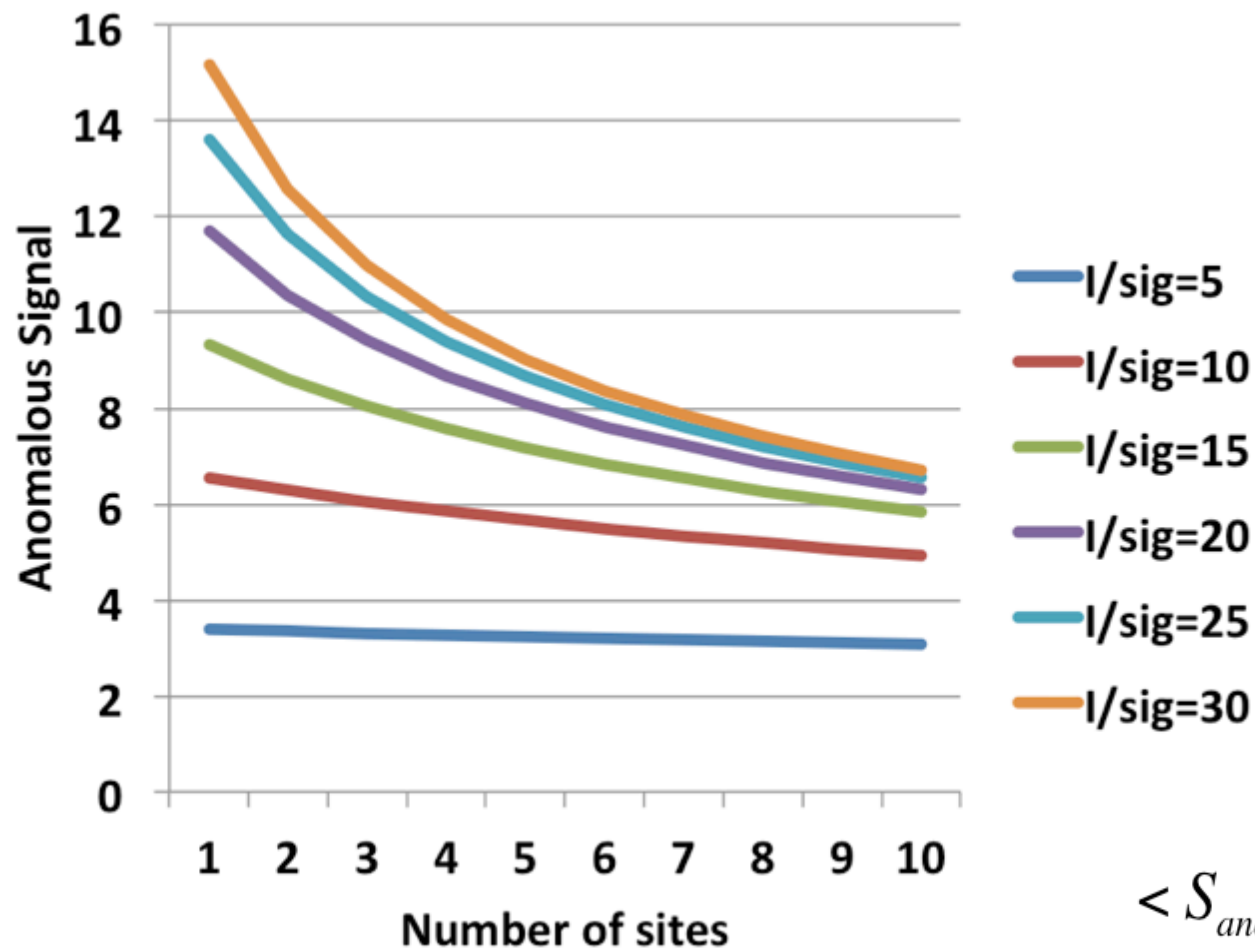


Phasing quality depends a lot on number of sites... but anomalous signal less so



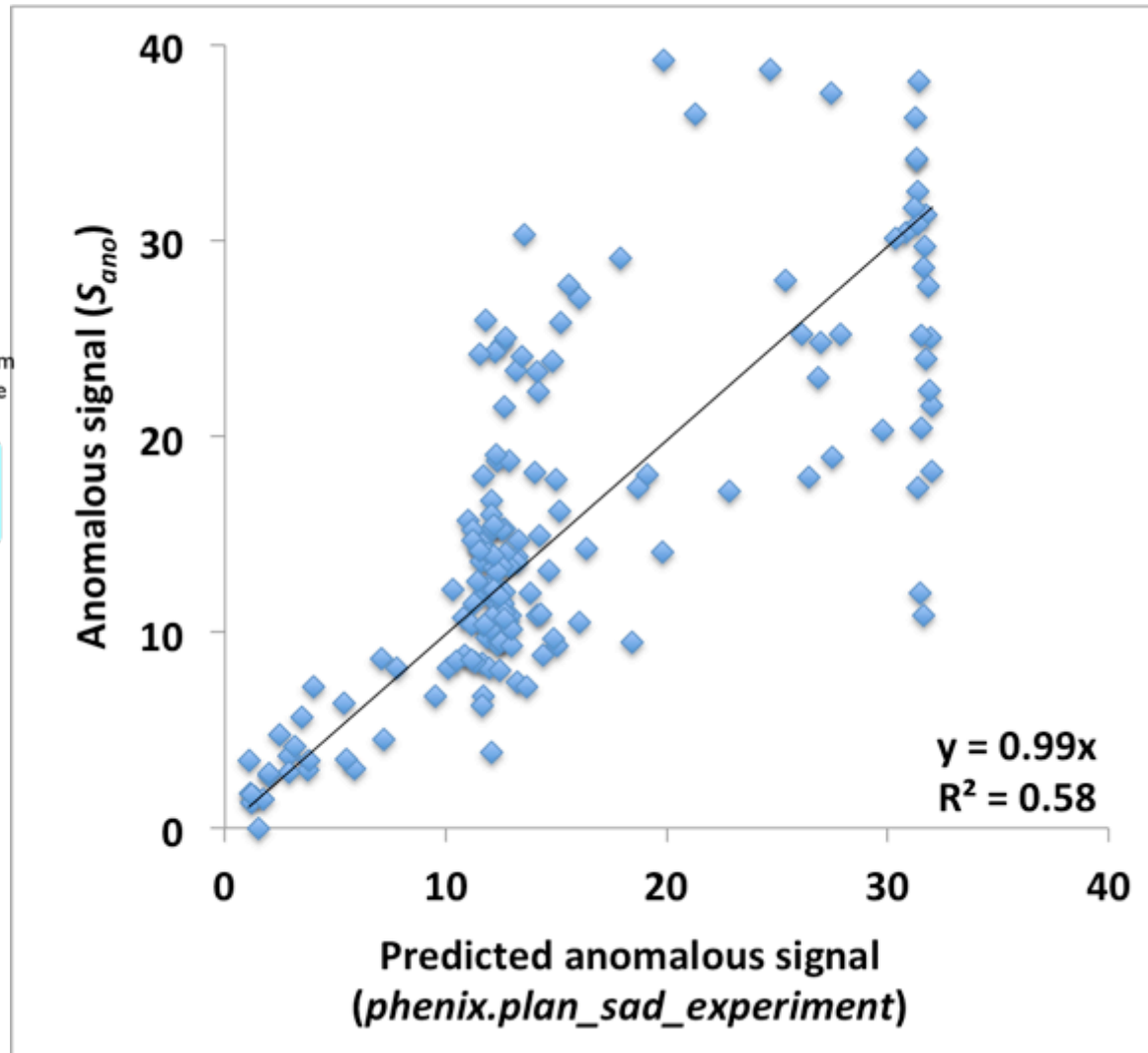
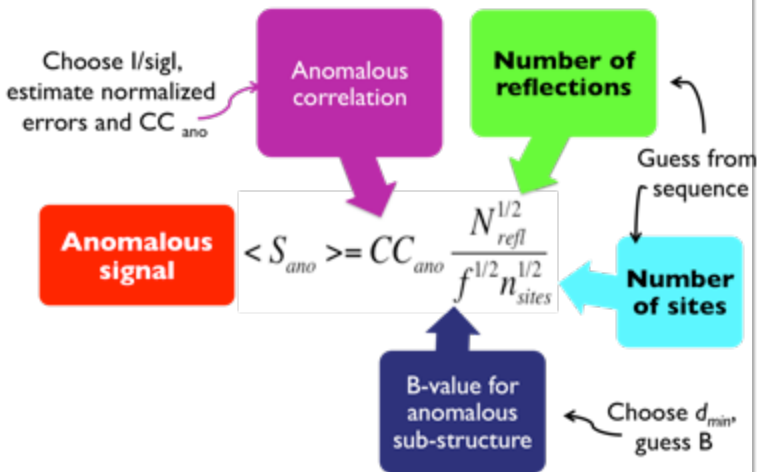
Anomalous signal vs I/σ and sites

100 residues, varying S_e , varying I/σ

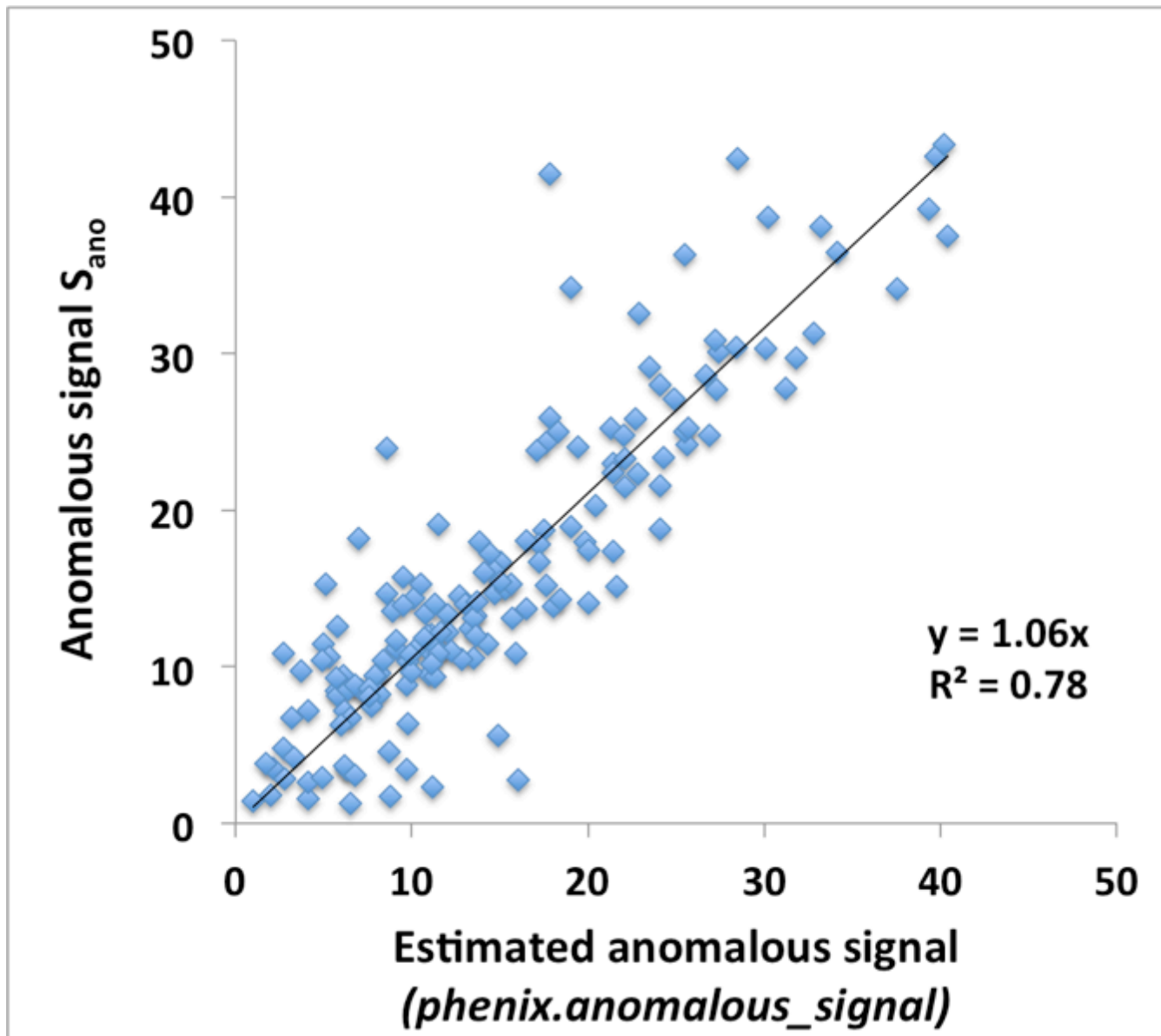


$$\langle S_{ano} \rangle = CC_{ano} \frac{N_{refl}^{1/2}}{f^{1/2} n_{sites}^{1/2}}$$

Estimating the anomalous signal before collecting the data

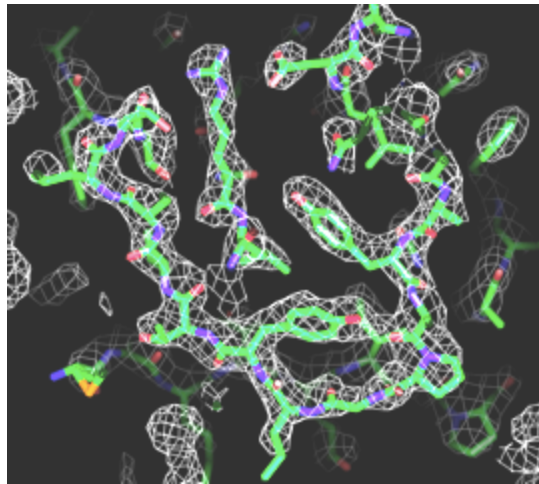


Estimating the anomalous signal after collecting the data



Planning an experiment: Summary

- **Plan the experiment:** *What overall I/sigI do I need to solve this structure? (phenix.plan_sad_experiment)*
- **Measure the data:** *Make sure I/sigI is high enough*
- **Scale the data:** *(phenix.scale_and_merge)*
- **Evaluate the accuracy of the anomalous differences** *(phenix.anomalous_signal)*
- **Find the anomalous sub-structure** *(phenix.hyss, phenix.autosol)*



Automation of structure determination

Automation...

makes straightforward cases accessible to a wider group of structural biologists

makes difficult cases more feasible for experts

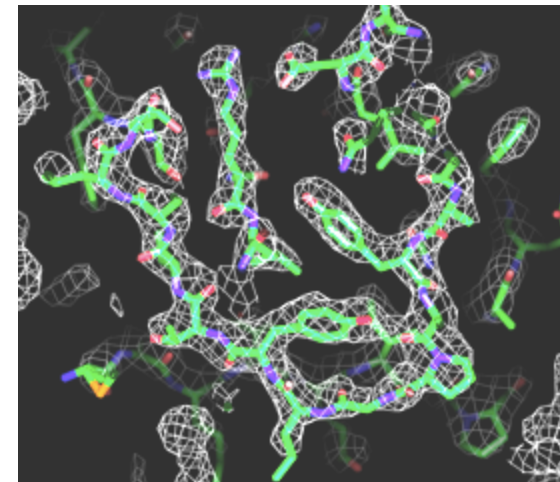
can speed up the process

can help reduce errors

Automation also allows you to...

try more possibilities

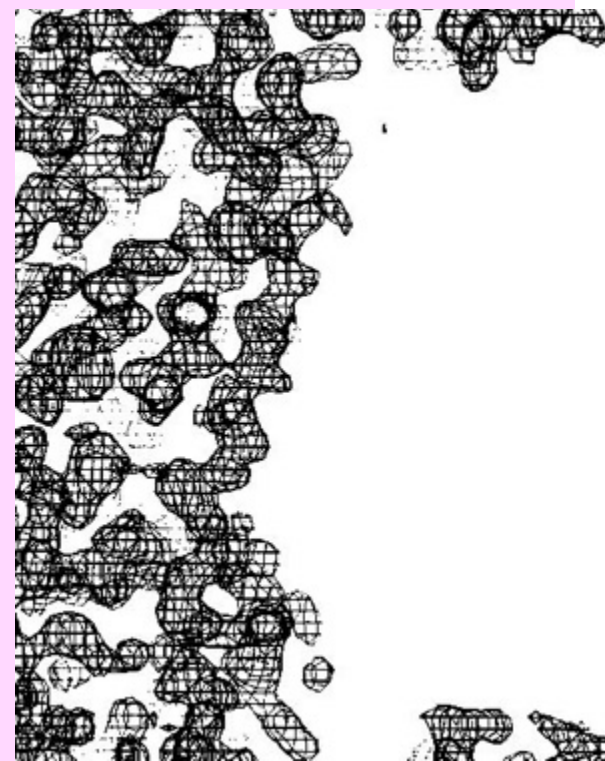
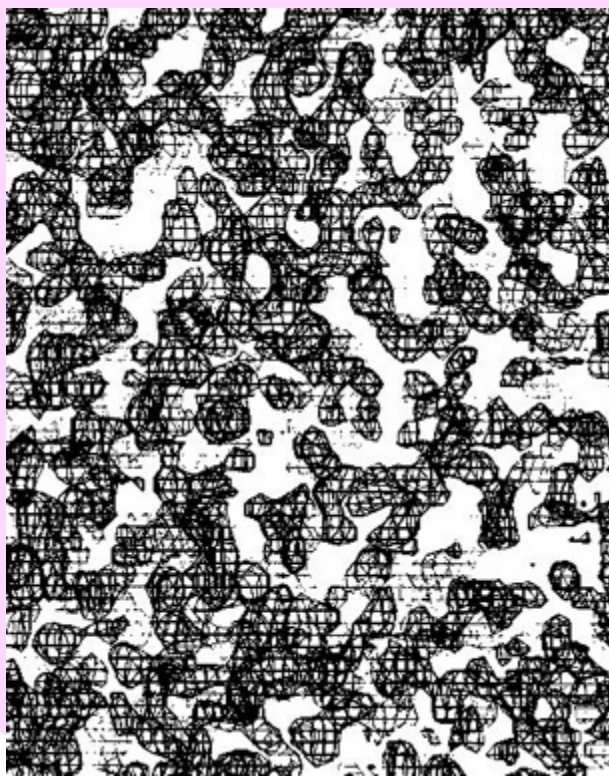
estimate uncertainties



Deciding what is good:
Measures of the quality of an electron-density map:

Which solution is best?

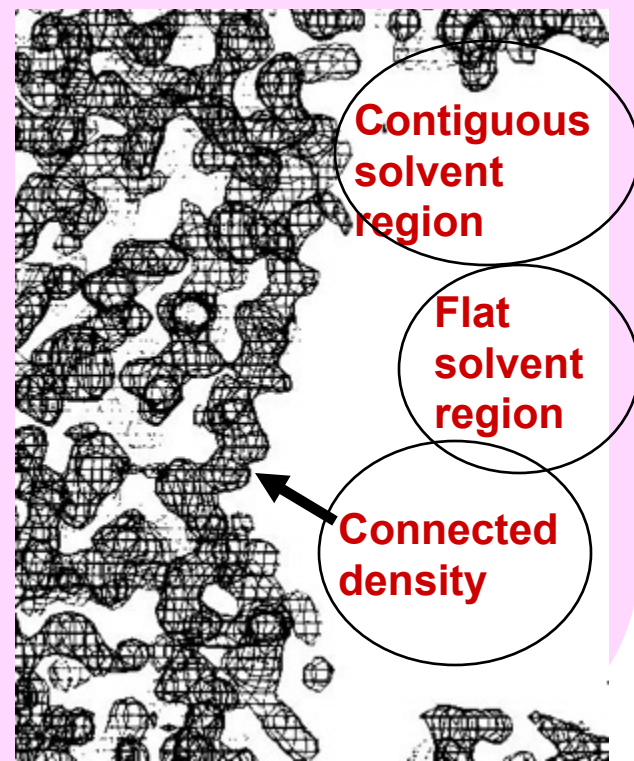
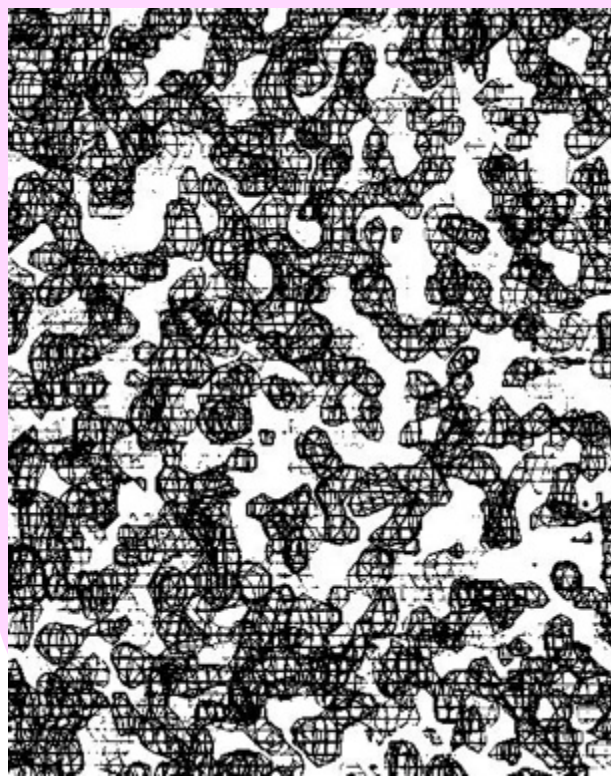
Are we on the right track?



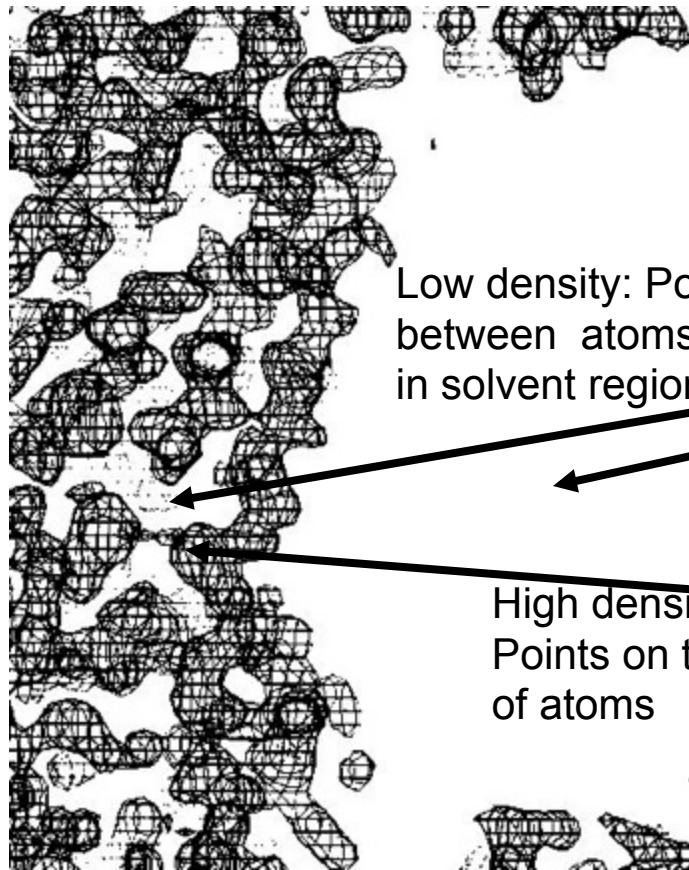
Why we need good measures of the quality of an electron-density map:

Which solution is best?

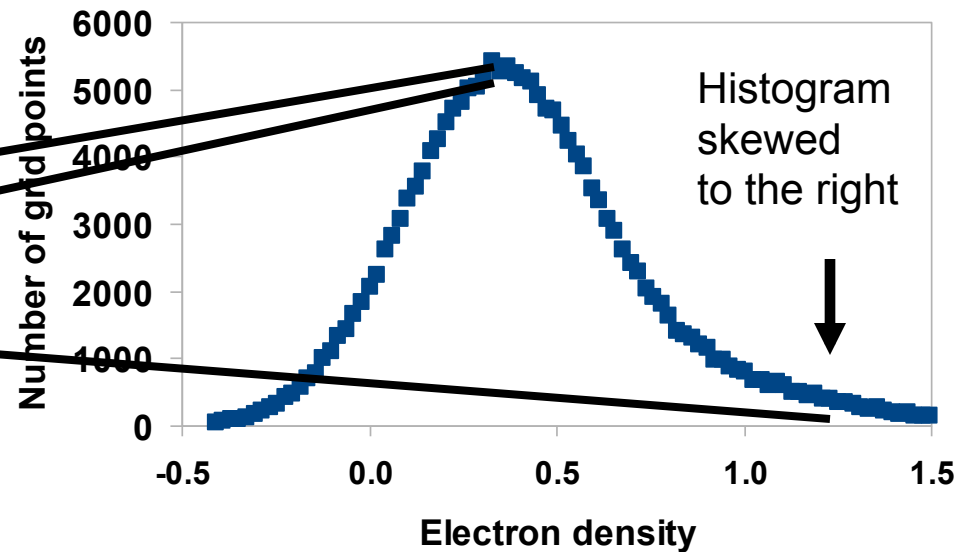
Are we on the right track?



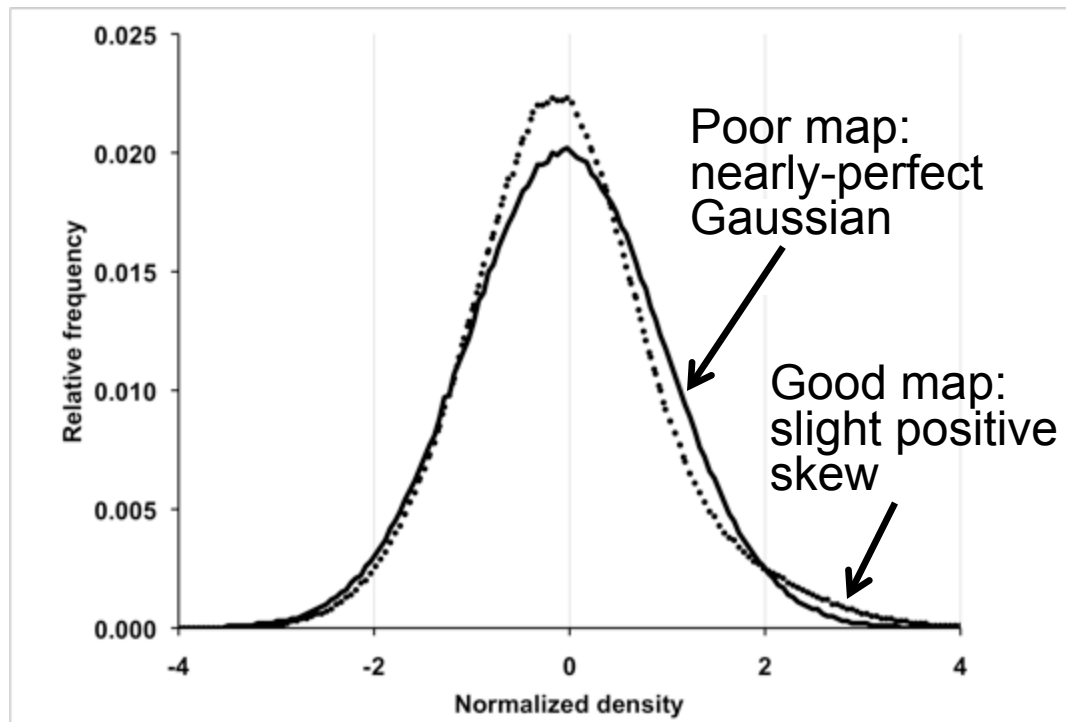
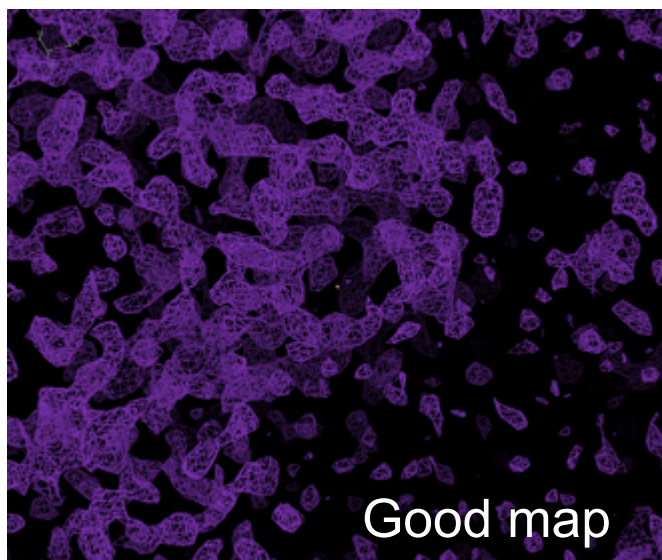
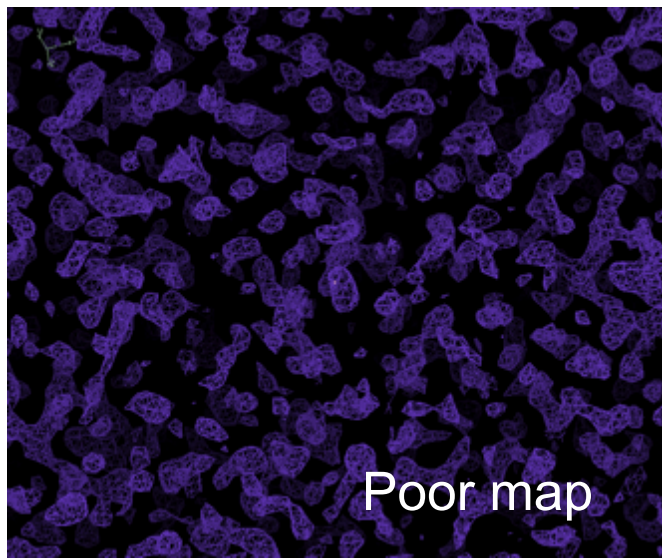
Histogram of electron density values has a positive “skew”



Typical histogram of electron density



Skew of electron density for poor and good maps



Evaluating electron density maps

<i>Basis</i>	<i>Good map</i>	<i>Random map</i>
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 1 st 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

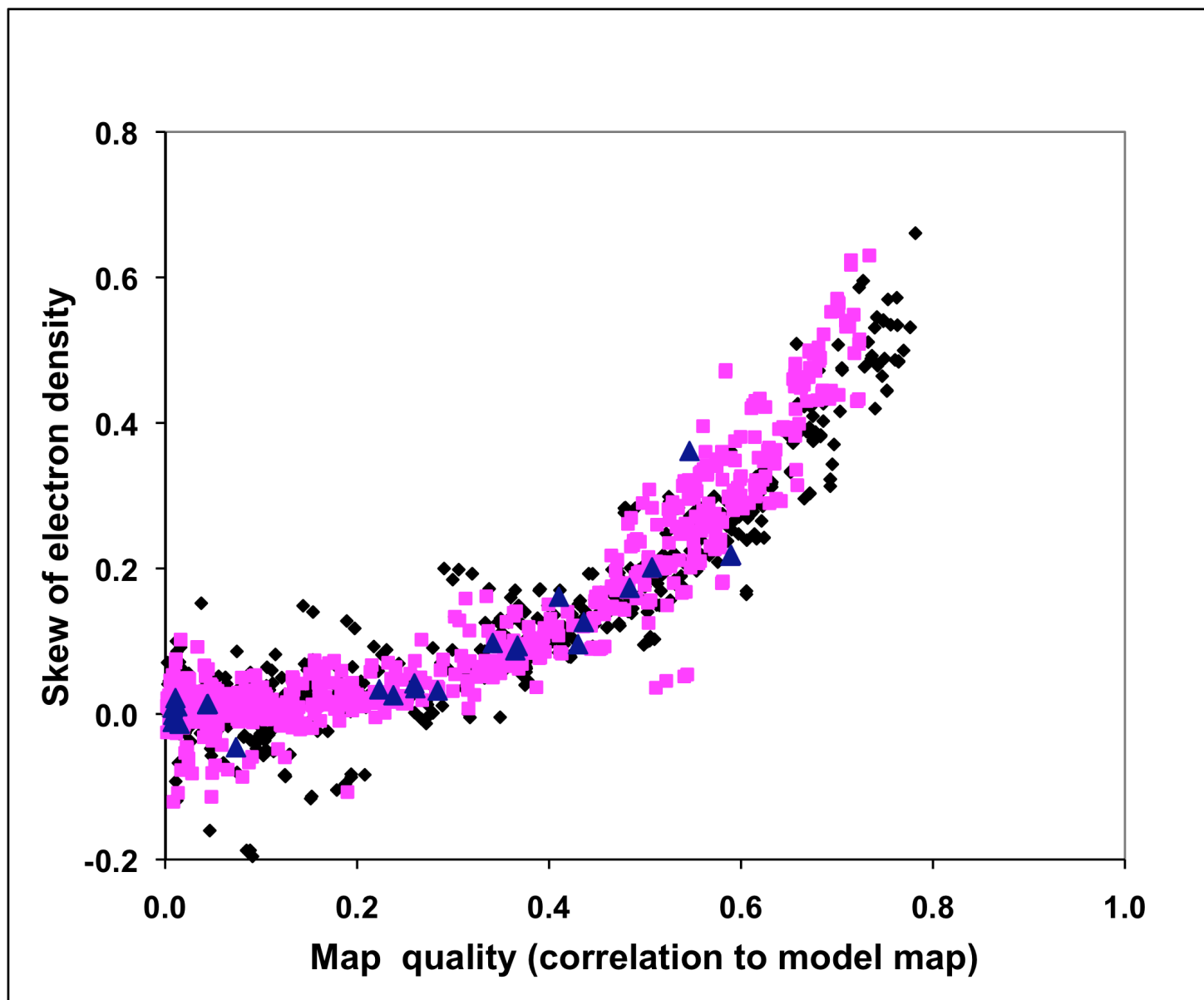
Creating real maps

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

Run AutoSol Wizard 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

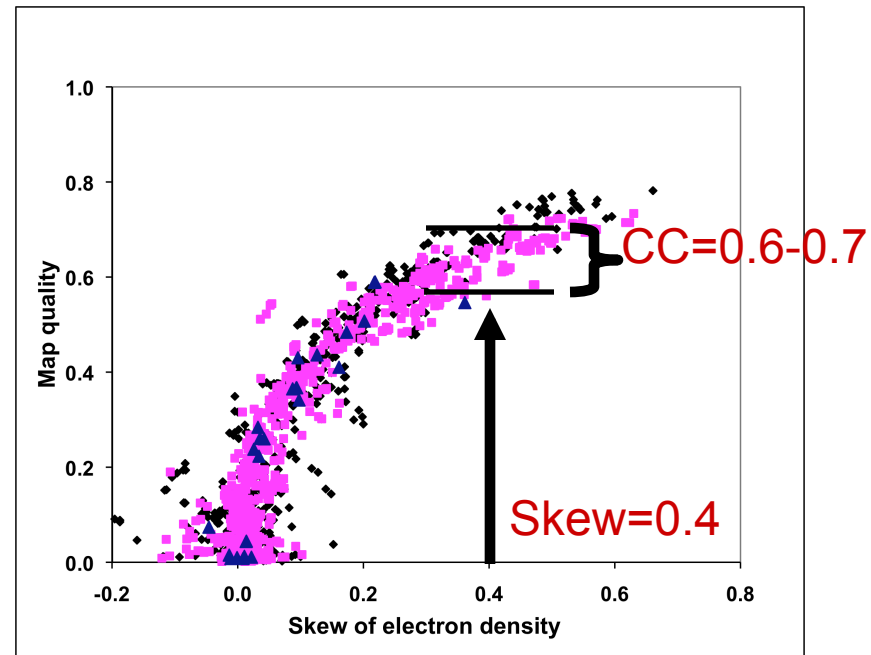
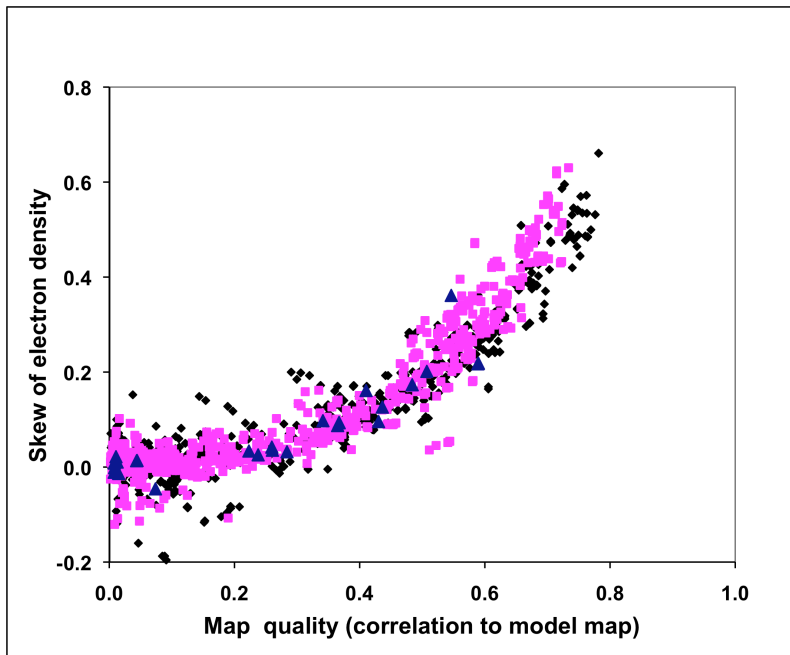


Using scoring criteria to estimate the quality of a map

Skew depends on map quality



Estimate map quality from skew



Estimated map quality in practice

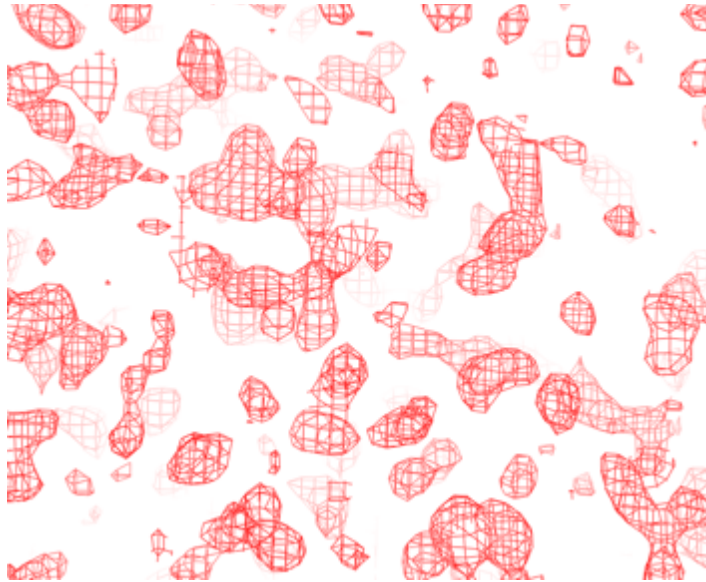
*Evaluating solutions to a 2-wavelength MAD experiment
(JCSG Tm3681, 1VPM, SeMet 1.6 Å data)*

Data for HYSS	Sites	Estimated CC $\pm 2\text{SD}$	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

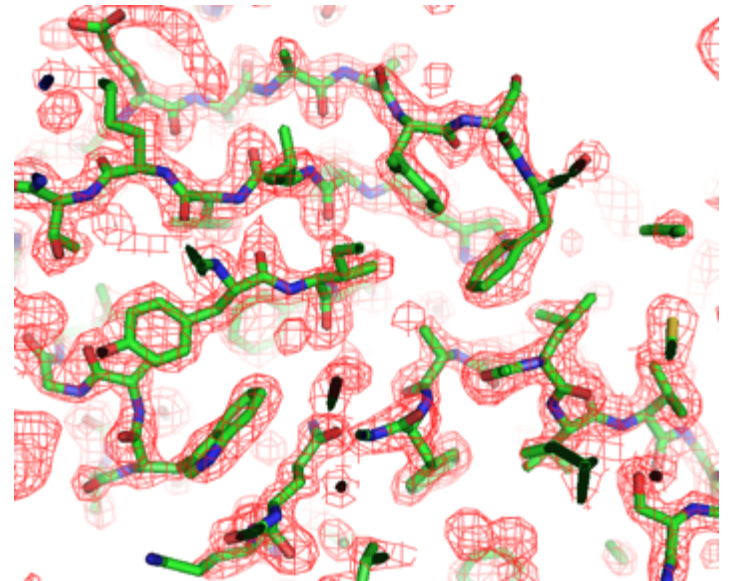
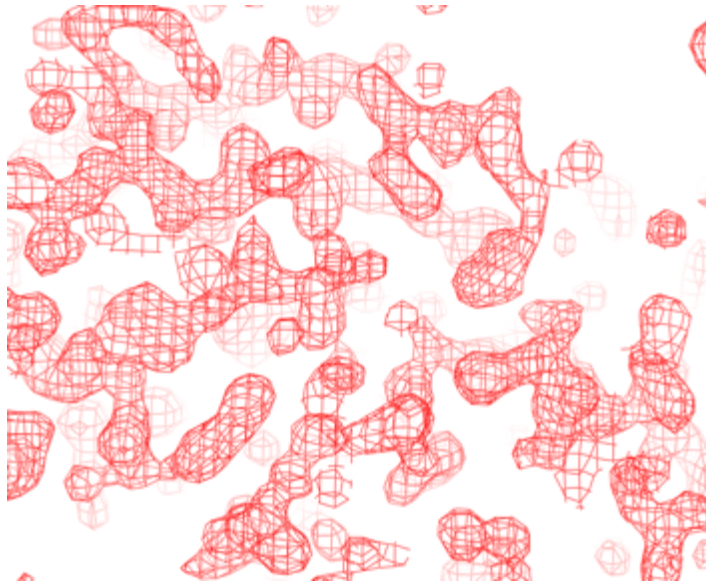
Improving map quality with density modification

SAD map , 2Å, no NCS, 50% solvent)

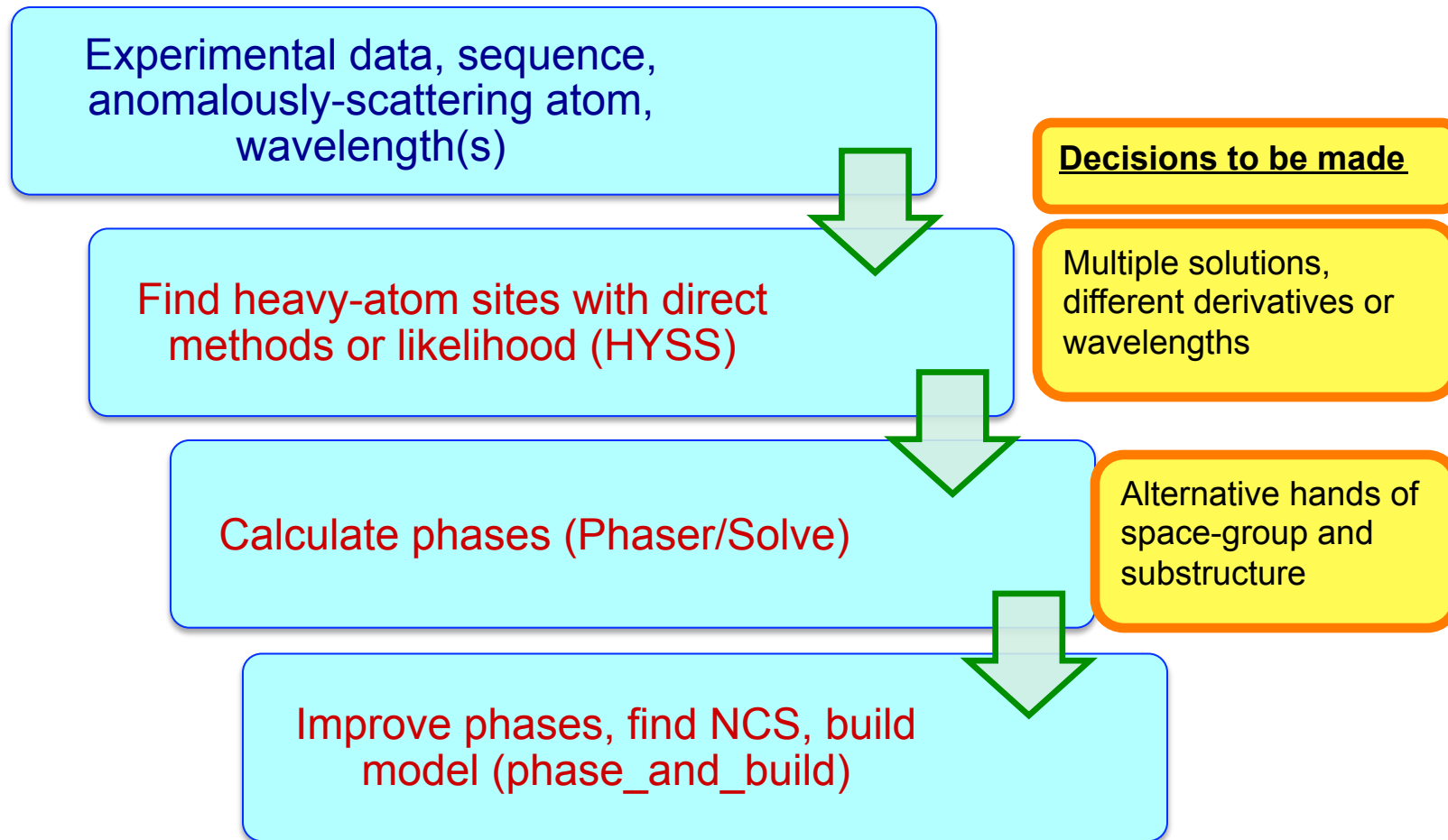
Phaser SAD map
(CC=0.43)



Phaser +RESOLVE
(CC=0.79)

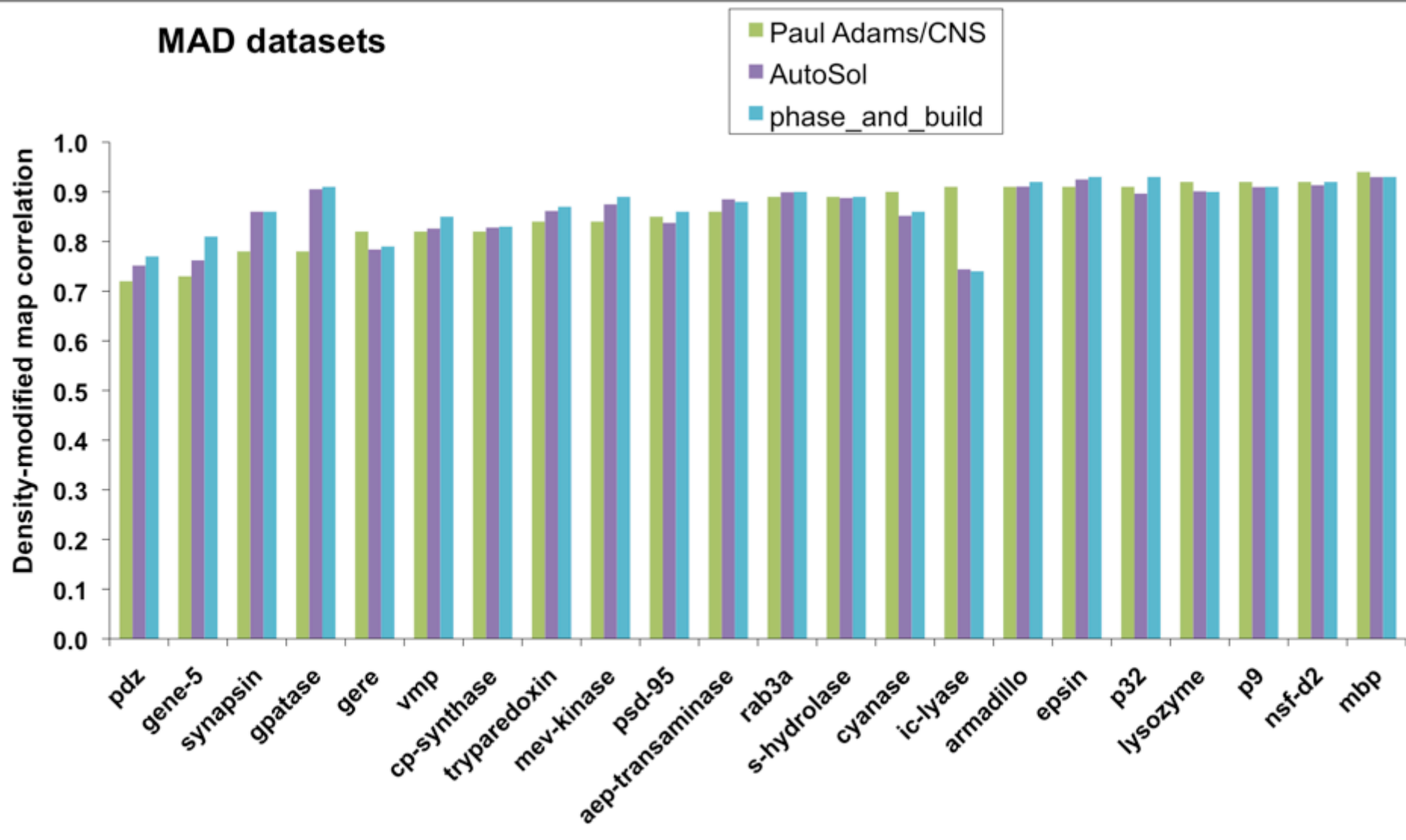


Structure solution with phenix.autosol

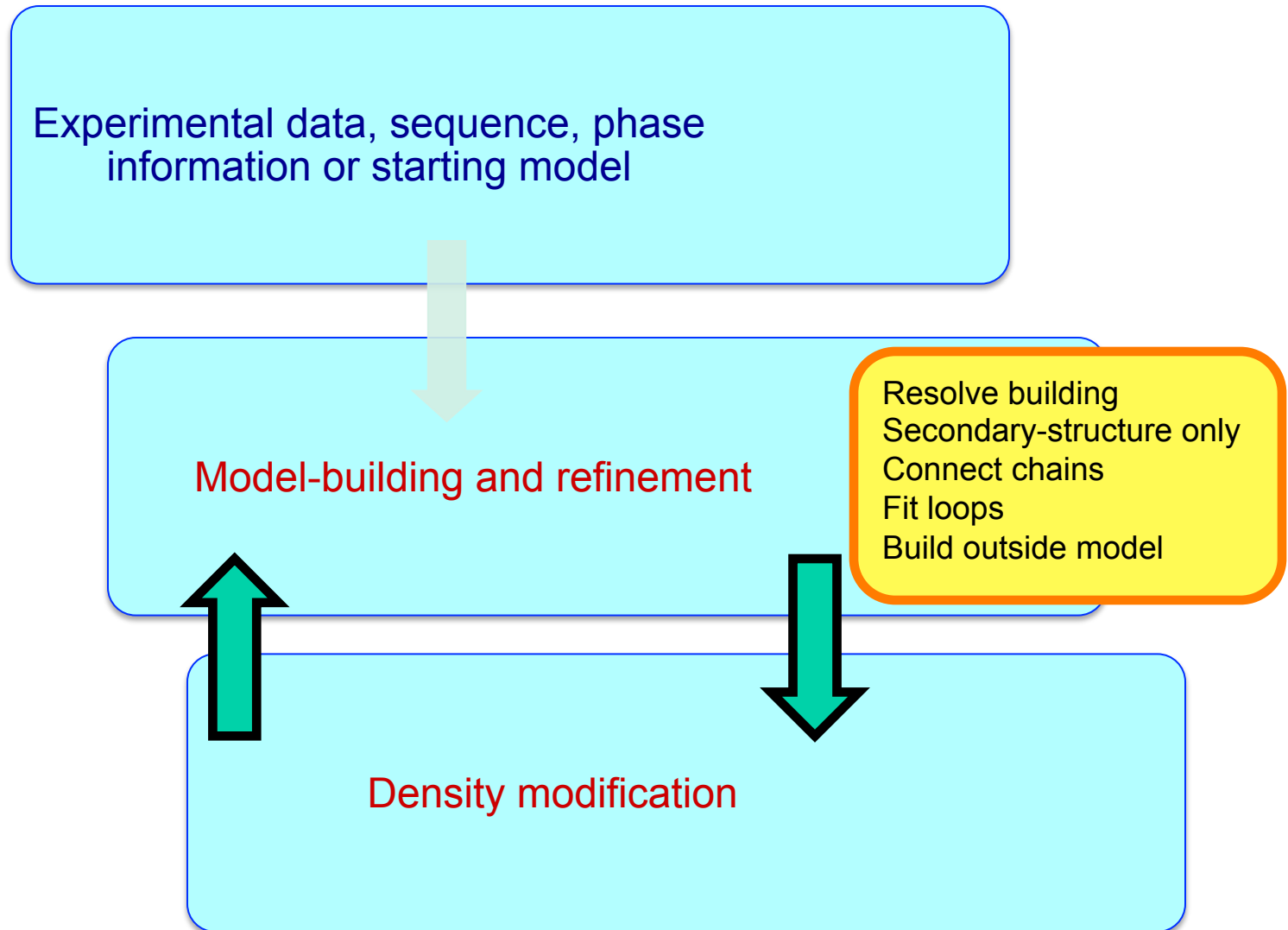


AutoSol – fully automatic tests with structure library
(MAD datasets, HYSS search, SOLVE)
RESOLVE/ phase_and_build maps

MAD datasets

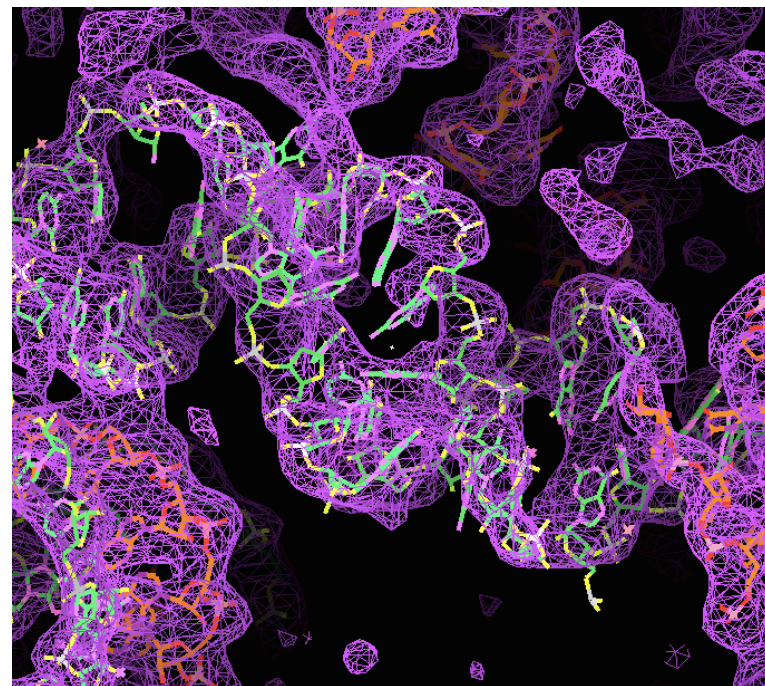
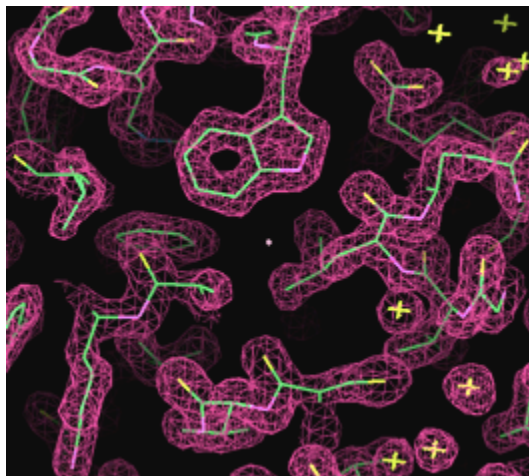


Iterative density modification, model-building and refinement with phenix.autobuild



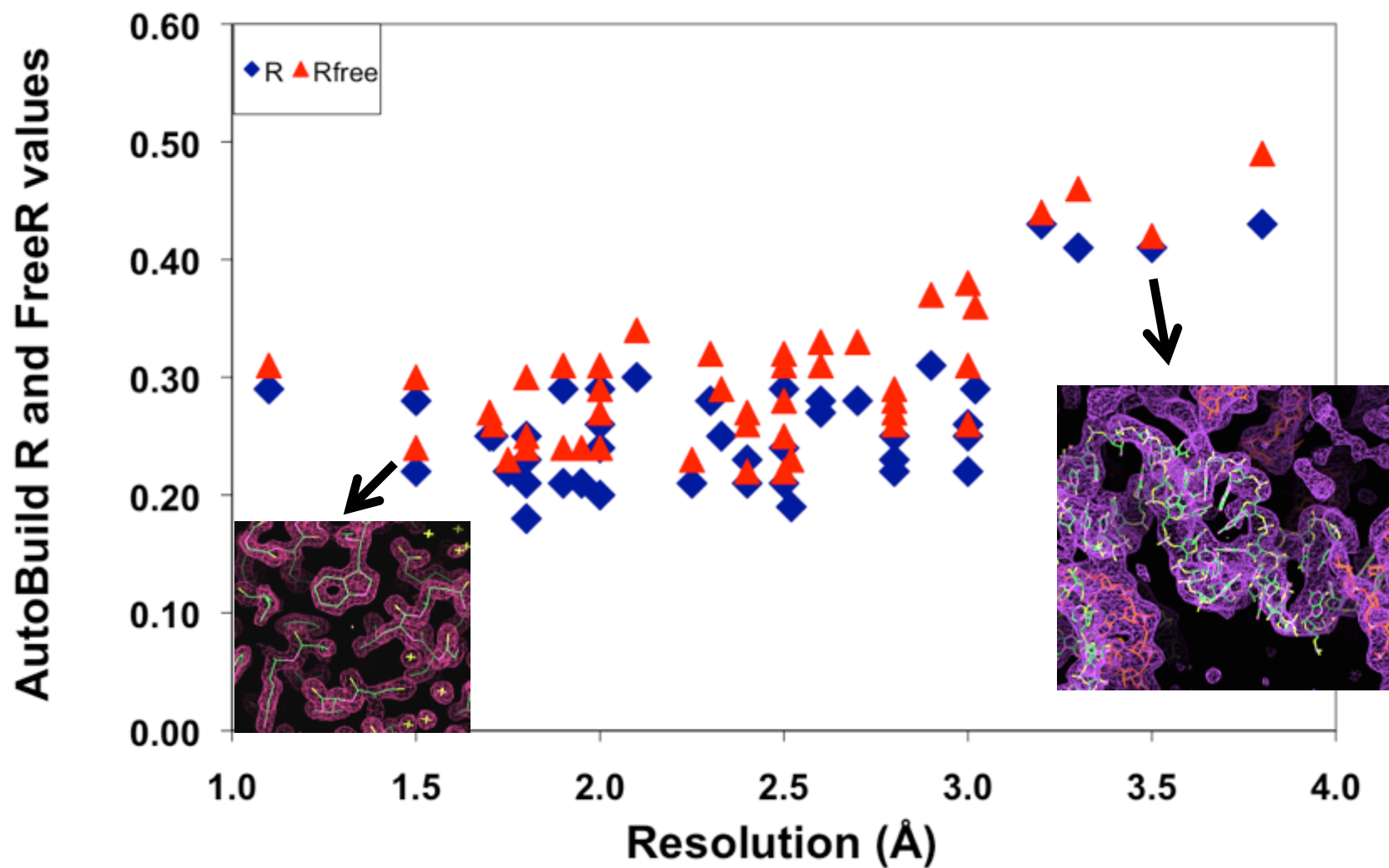
Model-building at moderate or high resolution

- FFT-based identification of regular secondary structure
- Extension with short fragments from high-resolution structures
- Probabilistic sequence alignment



AutoBuild – tests with structure library

Fully automated iterative model-building, final R/Rfree



The Phenix Project



Lawrence Berkeley Laboratory

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



Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung



Randy Read, Airlie McCoy, Gabor Bunkoczi, Rob Oeffner

Cambridge University



Duke University

Jane & David Richardson, Chris Williams, Bryan Arendall, Bradley Hintze



*An NIH/NIGMS funded
Program Project*

Model-building and Density Modification

Phenix workshop

Shanghai, China
Jan. 14, 2016

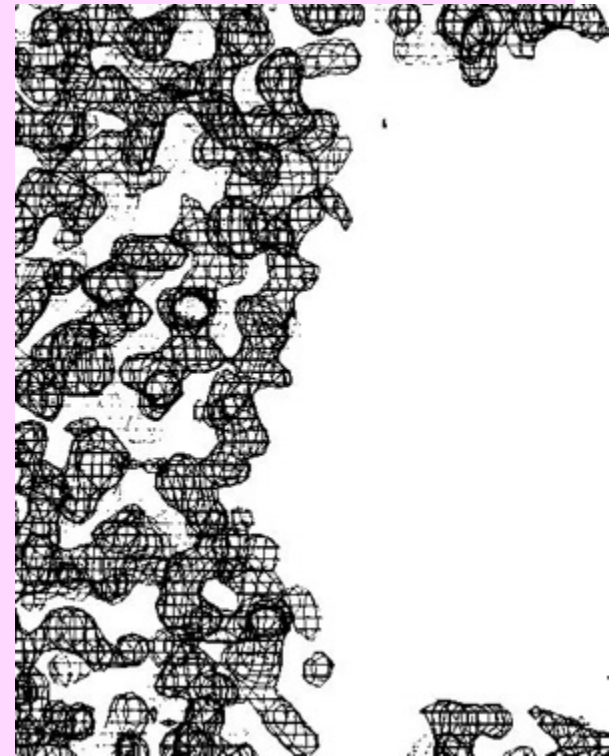
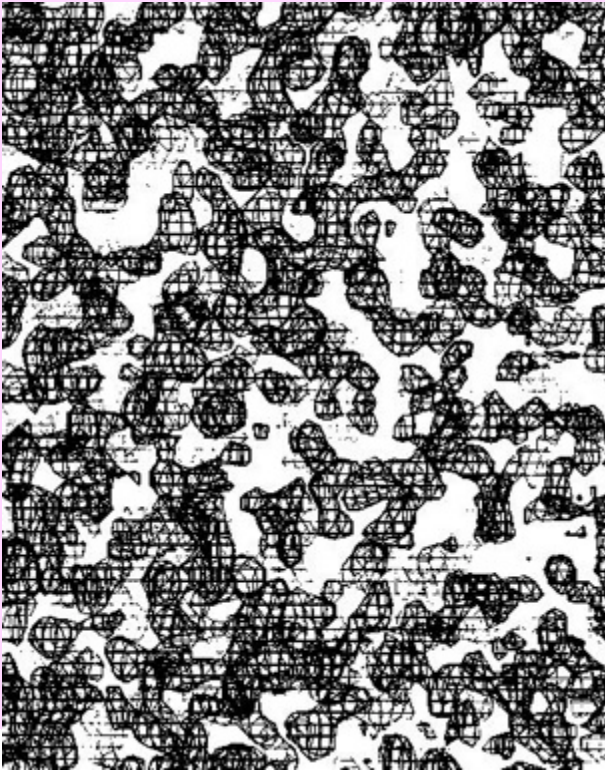
Tom Terwilliger
Los Alamos National Laboratory



Deciding what is good:
Measures of the quality of an electron-density map:

Which solution is best?

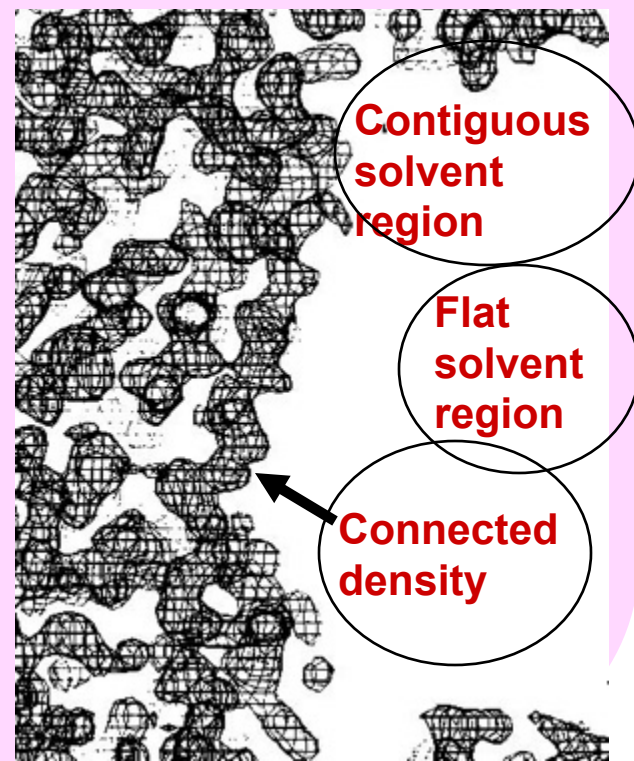
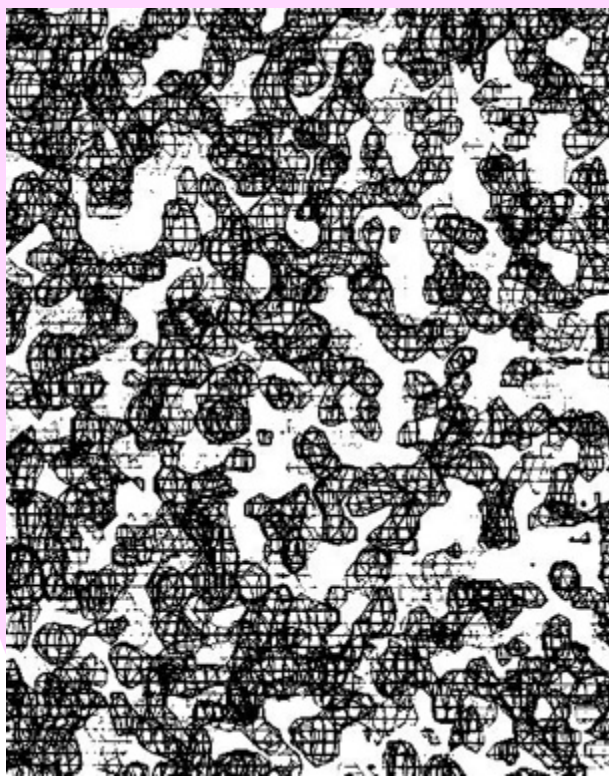
Are we on the right track?



Why we need good measures of the quality of an electron-density map:

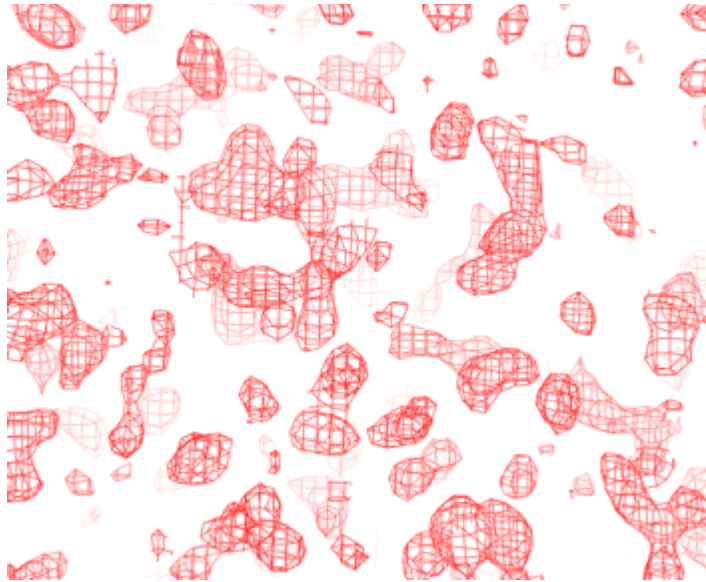
Which solution is best?

Are we on the right track?

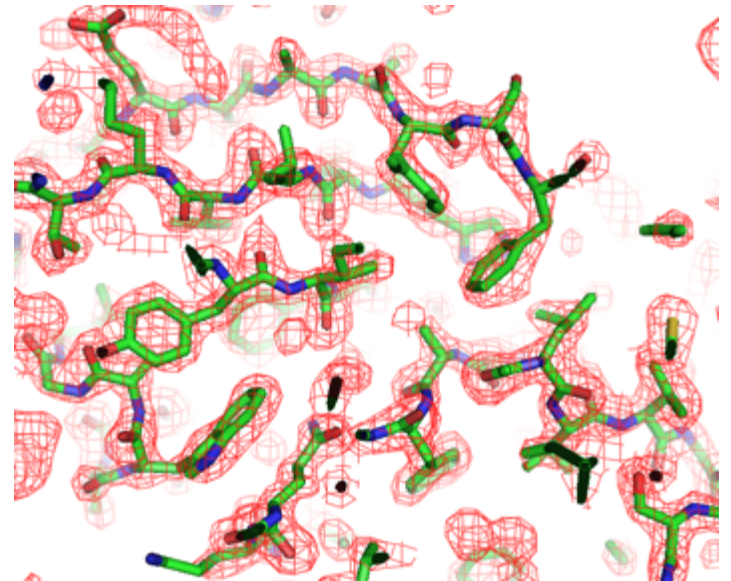
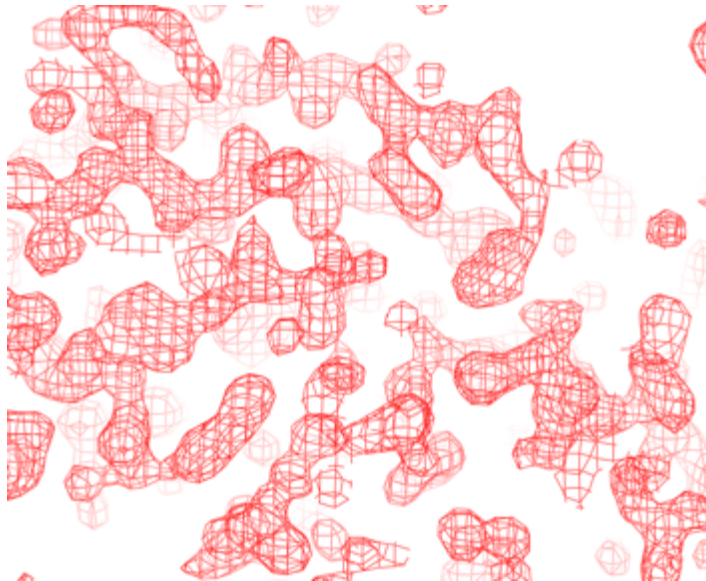


Improving crystallographic map quality with density modification (SAD map , 2Å, no NCS, 50% solvent)

Phaser SAD map
(CC=0.43)

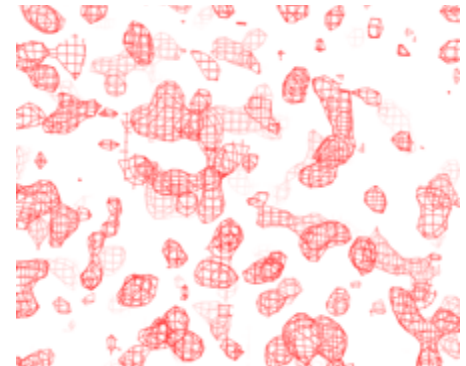


Phaser +RESOLVE
(CC=0.79)

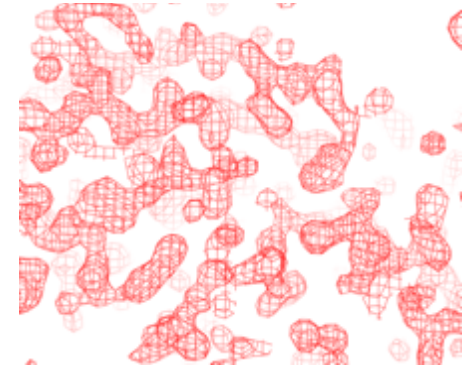


Statistical density modification

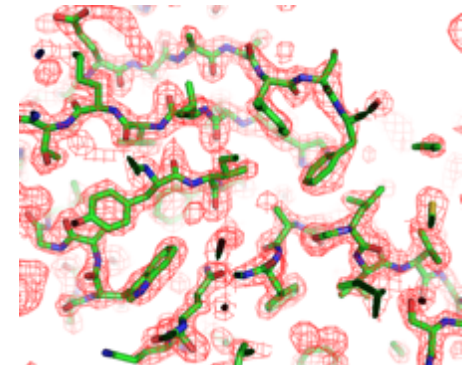
- Principle: phase probability information from probability of the map and from experiment:
- $P(\phi) = P_{\text{map probability}}(\phi) P_{\text{experiment}}(\phi)$
- “Phases that lead to a believable map are more probable than those that do not”
- A believable map is a map that has...
 - a relatively 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

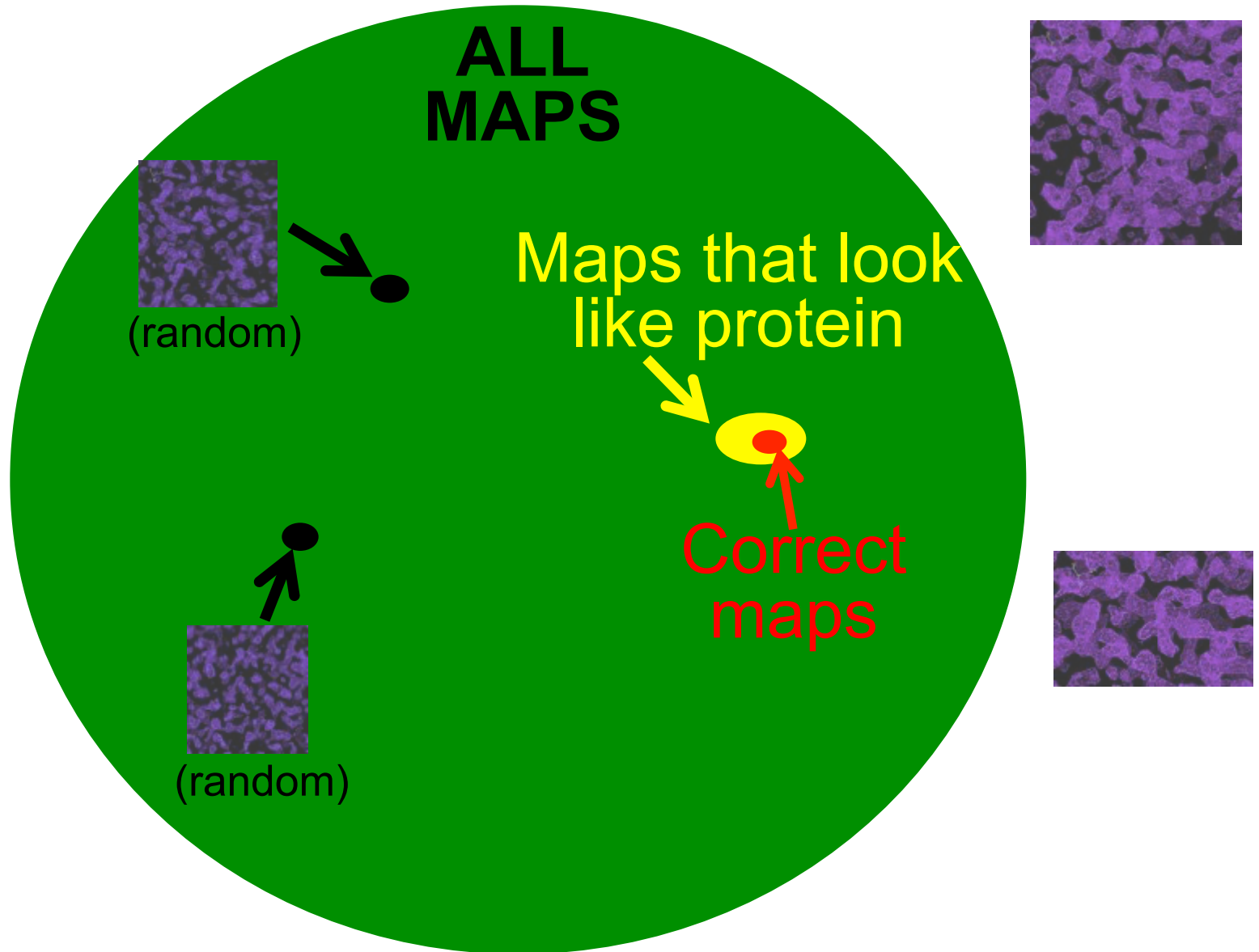


Density-modified



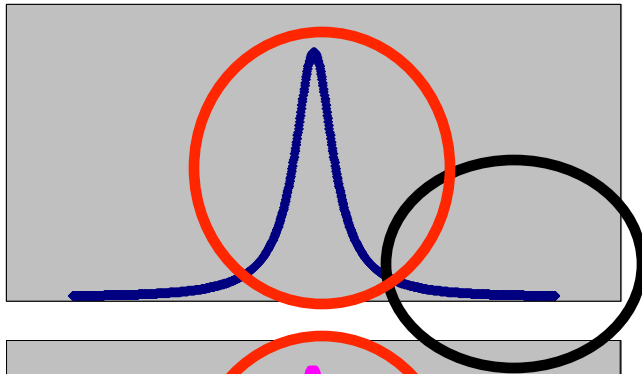
Interpreted

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

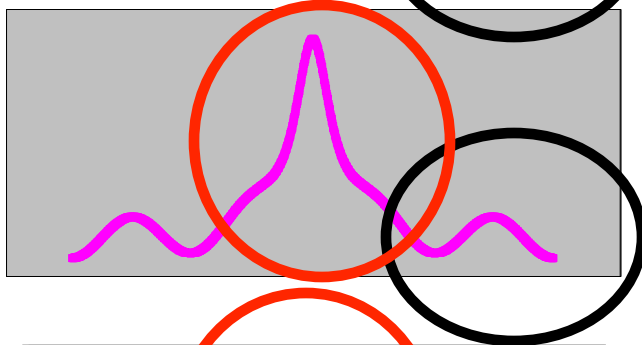


Map probability phasing: Getting a new probability distribution for each phase given estimates of all others

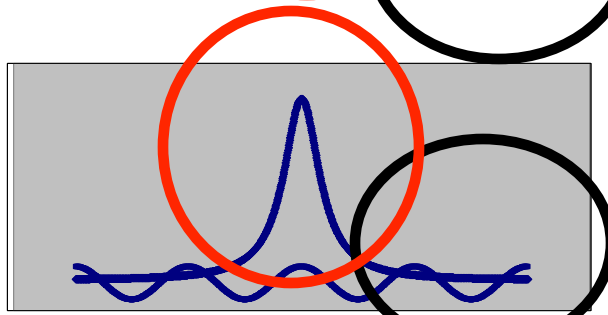
1. Identify expected features of map (flat far from center)
2. Calculate map with current estimates of all structure factors except one (k)
3. Test all possible phases ϕ for structure factor k (for each phase, calculate new map including k)
4. Probability of phase ϕ estimated from agreement of map with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



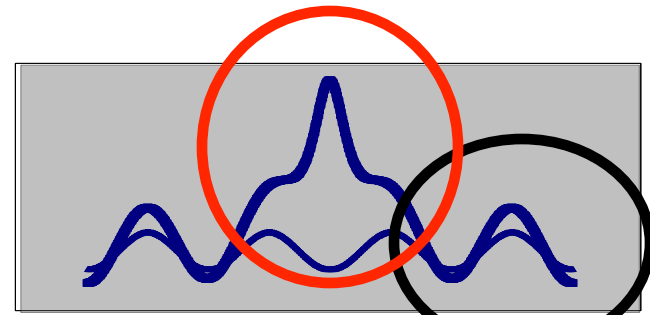
A function that is (relatively) flat far from the origin



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

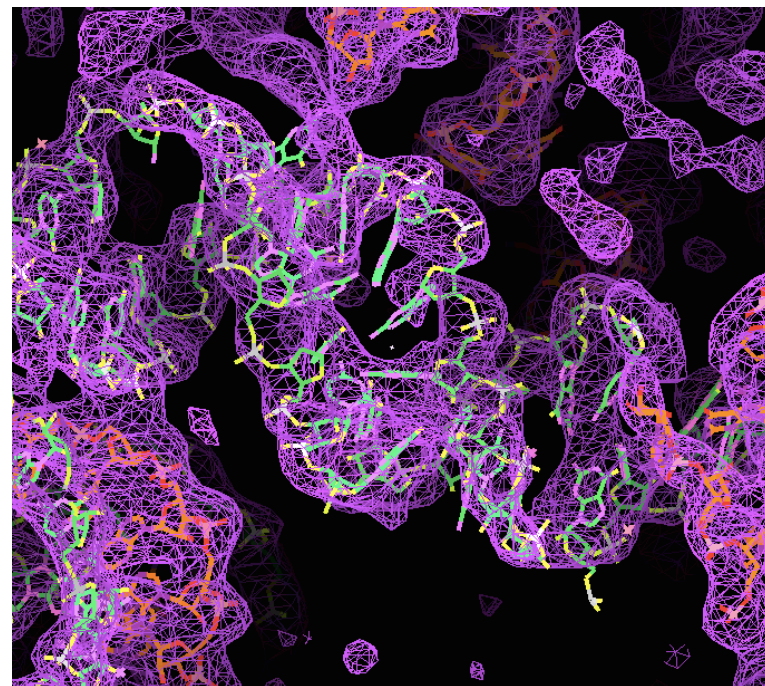
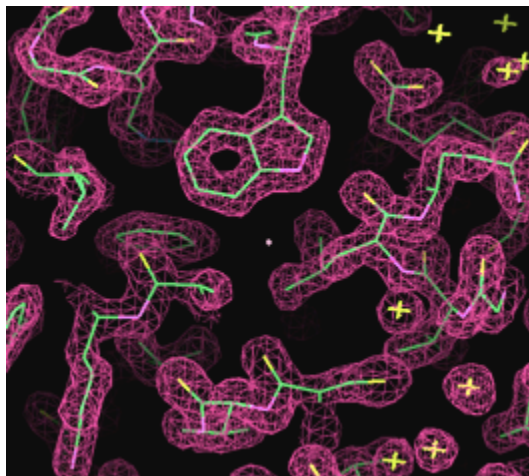


Test each possible phase of structure factor k . $P(\phi)$ is high for phase that leads to flat region

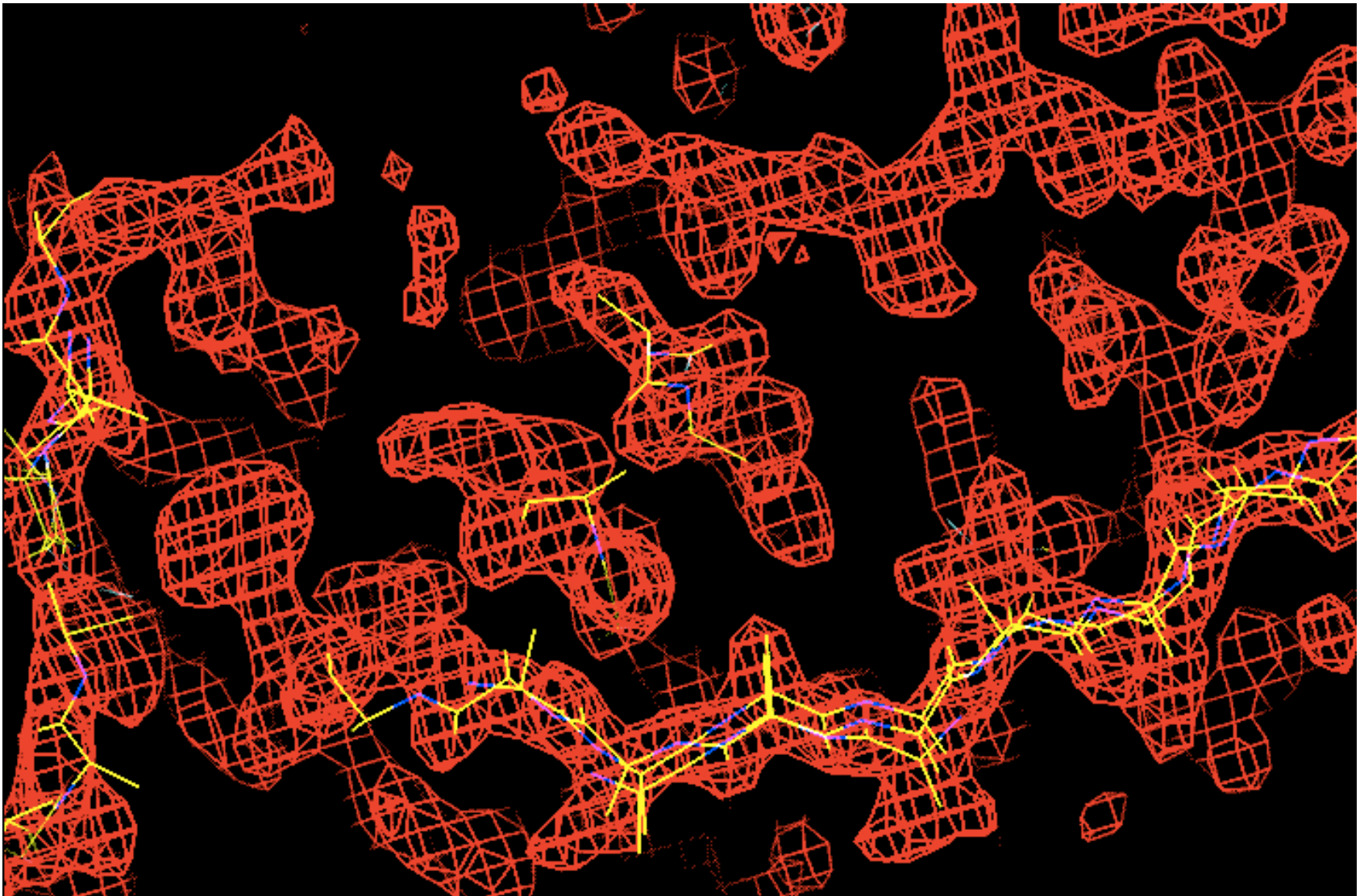


Model-building at moderate or high resolution

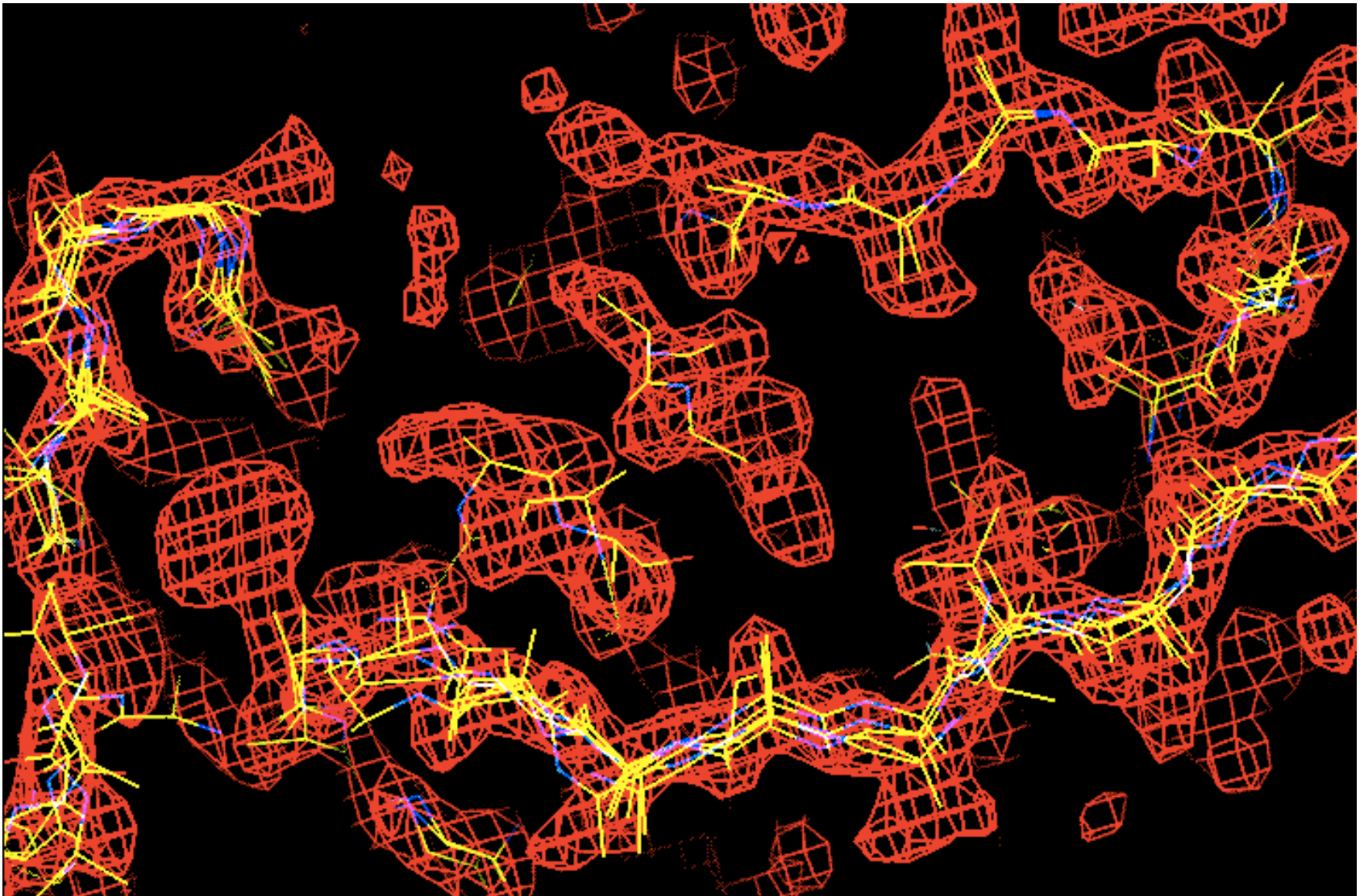
- FFT-based identification of regular secondary structure
- Extension with short fragments from high-resolution structures
- Probabilistic sequence alignment



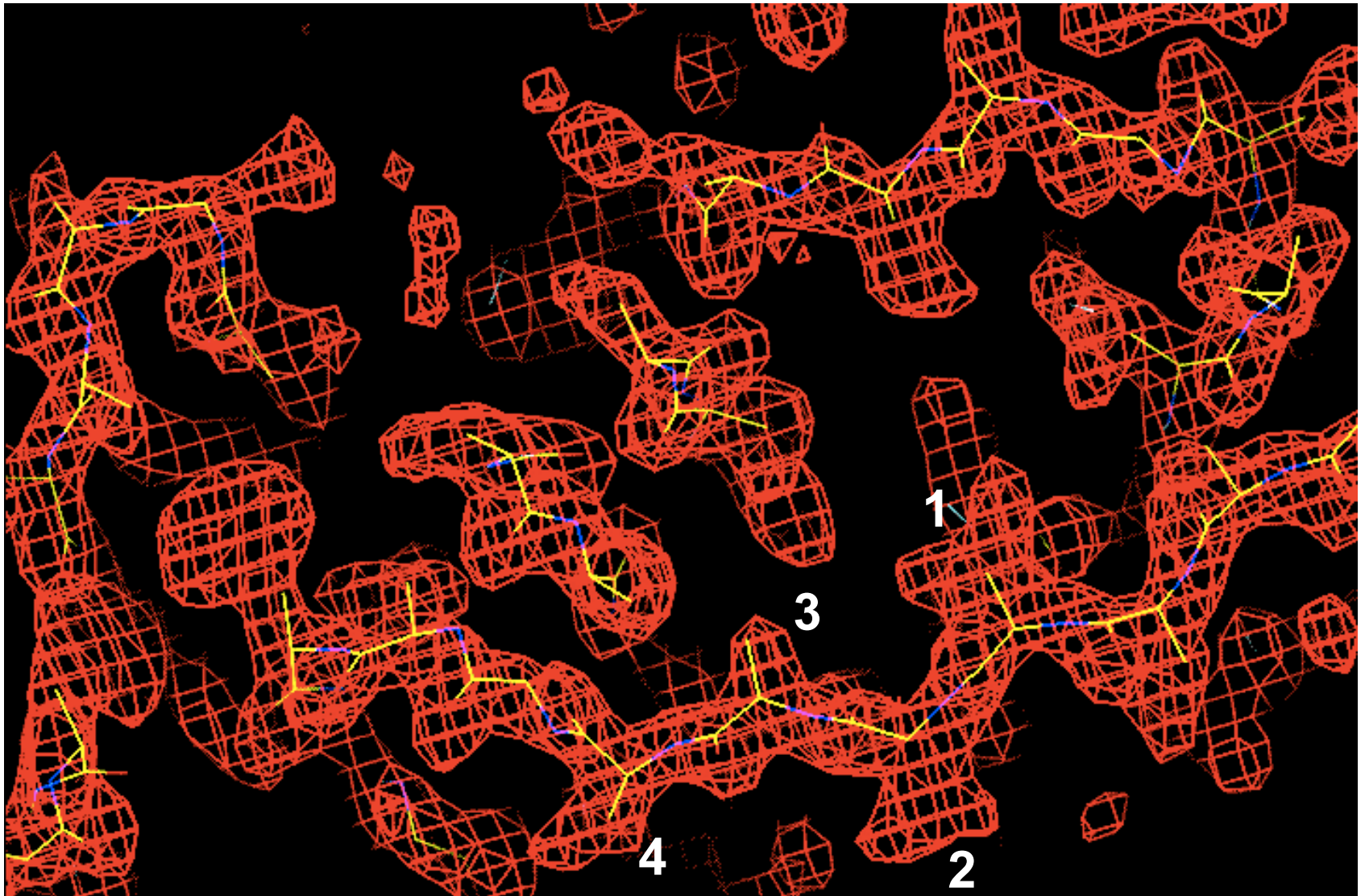
Initial model-building – strand fragments



*Chain extension
(result: many overlapping fragments)*



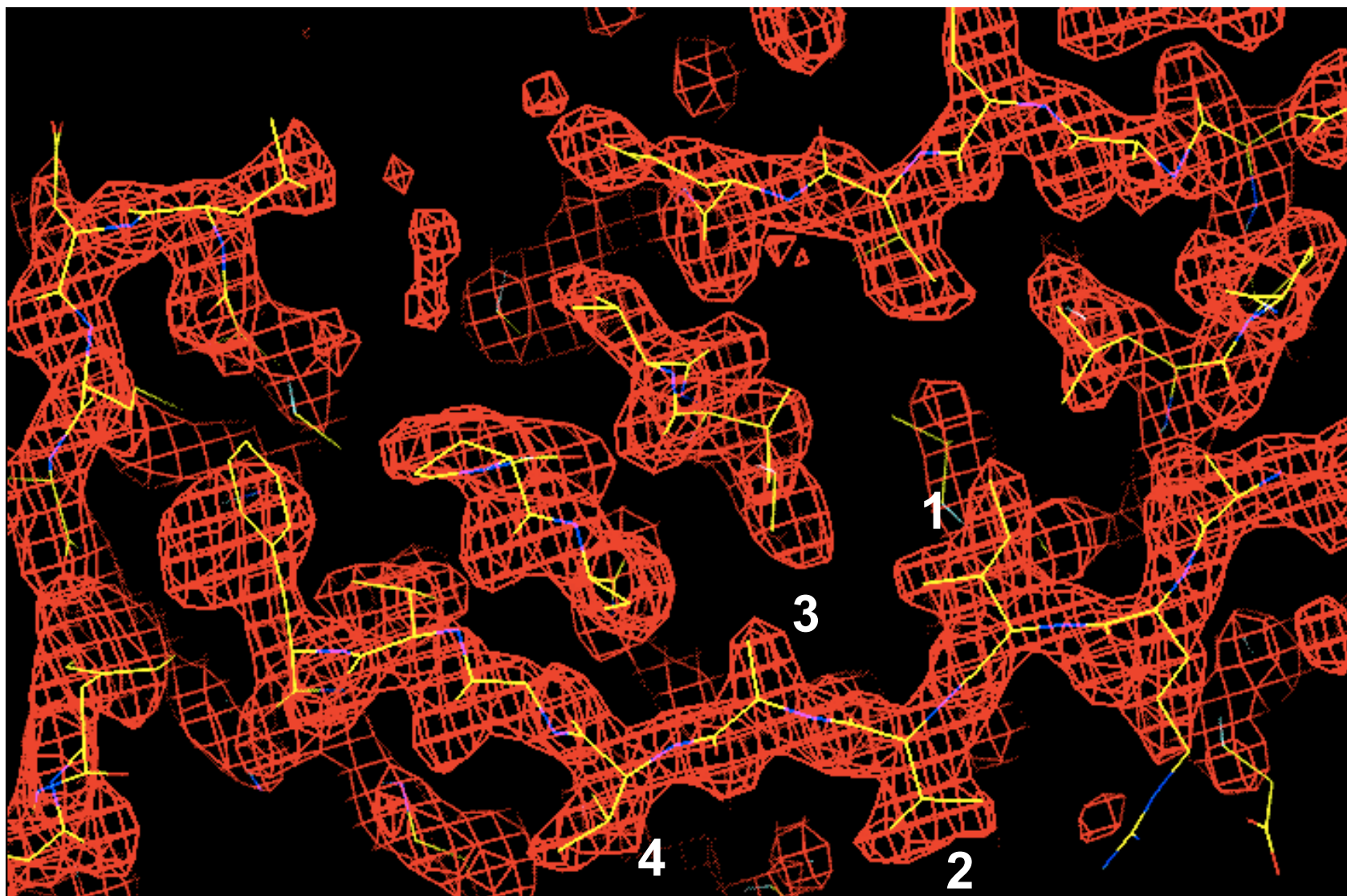
*Main-chain as a series of fragments
(choosing the best fragment at each location)*



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)

[illegible]

Addition of side-chains to fixed main-chain positions



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

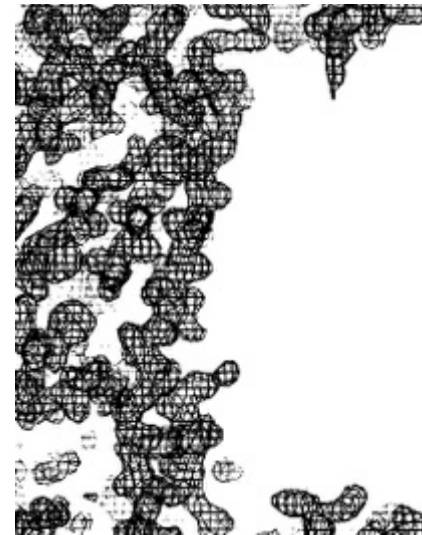
$$LL^{MAP}(\{\mathbf{F}_h\}) \approx \frac{N_{REF}}{V} \int_V LL(\rho(\mathbf{x}, \{\mathbf{F}_h\})) d^3\mathbf{x}$$

Local log-probability is believability of the value of electron density ($\rho(\mathbf{x})$) 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 ($\rho(\mathbf{x})$) are believable

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

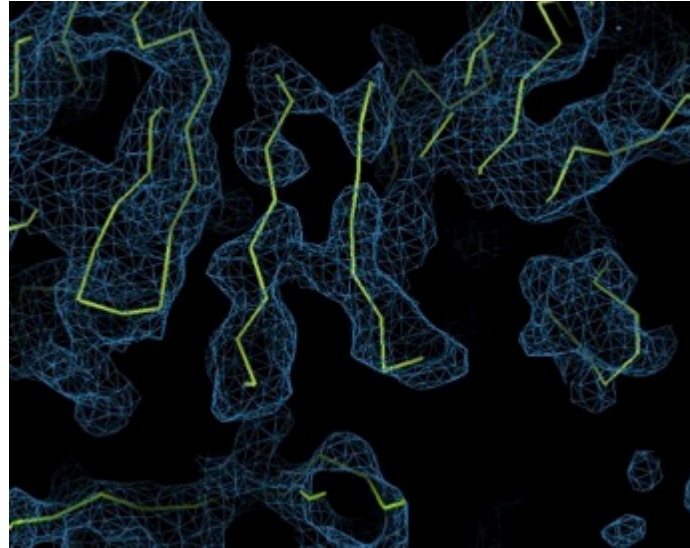


A map with a flat (blank) solvent region is a likely map

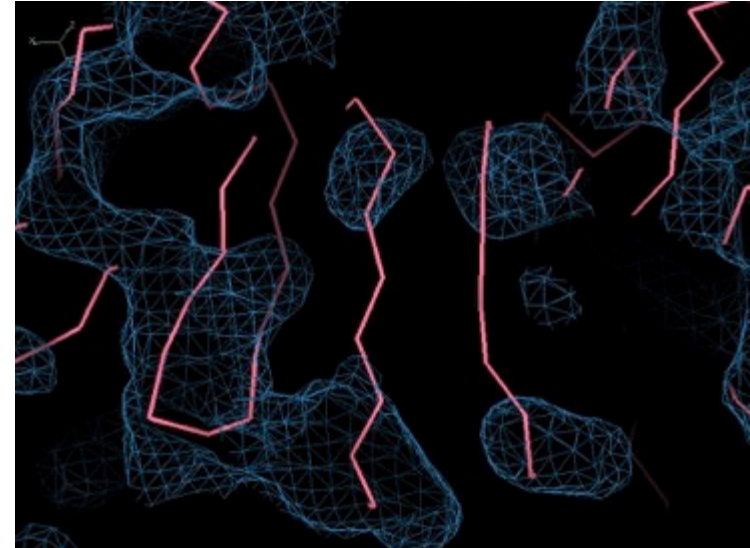
Statistical density modification with cross-crystal averaging

Cell receptor at 3.5/3.7 Å. Data courtesy of J. Zhu

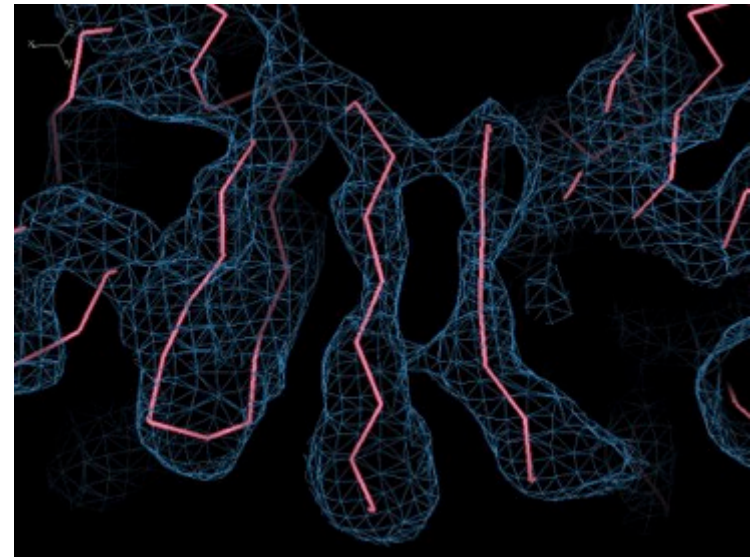
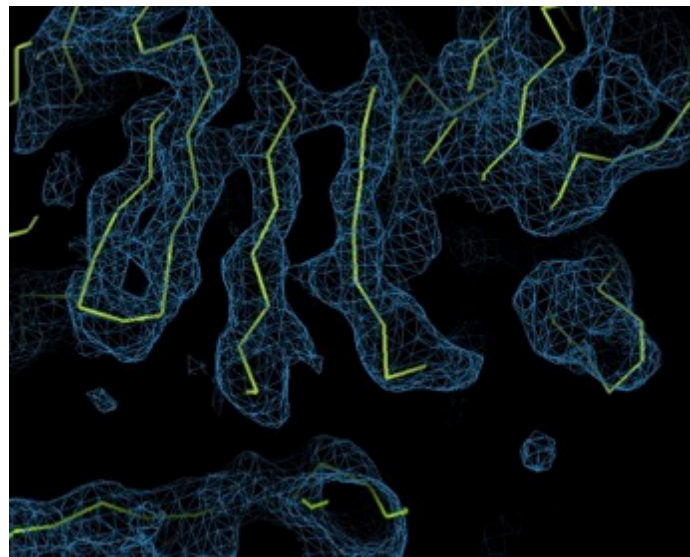
Crystal 1 (4 copies)



Crystal 2 (2 copies)



RESOLVE
density
modification



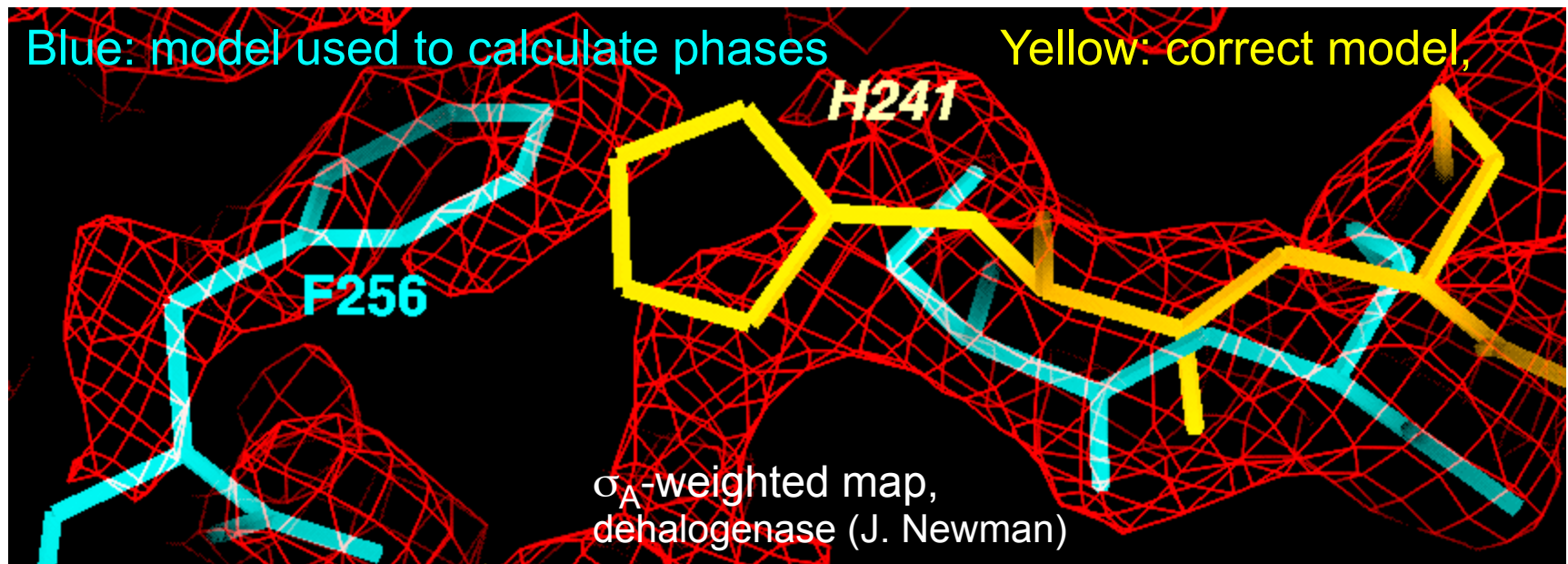
PHENIX
Multi-crystal
averaging

Removing model bias with prime-and-switch phasing

The problem:

Atomic model used to calculate phases $\rightarrow$ map looks like the model

Best current solution: σ_A -weighted phases



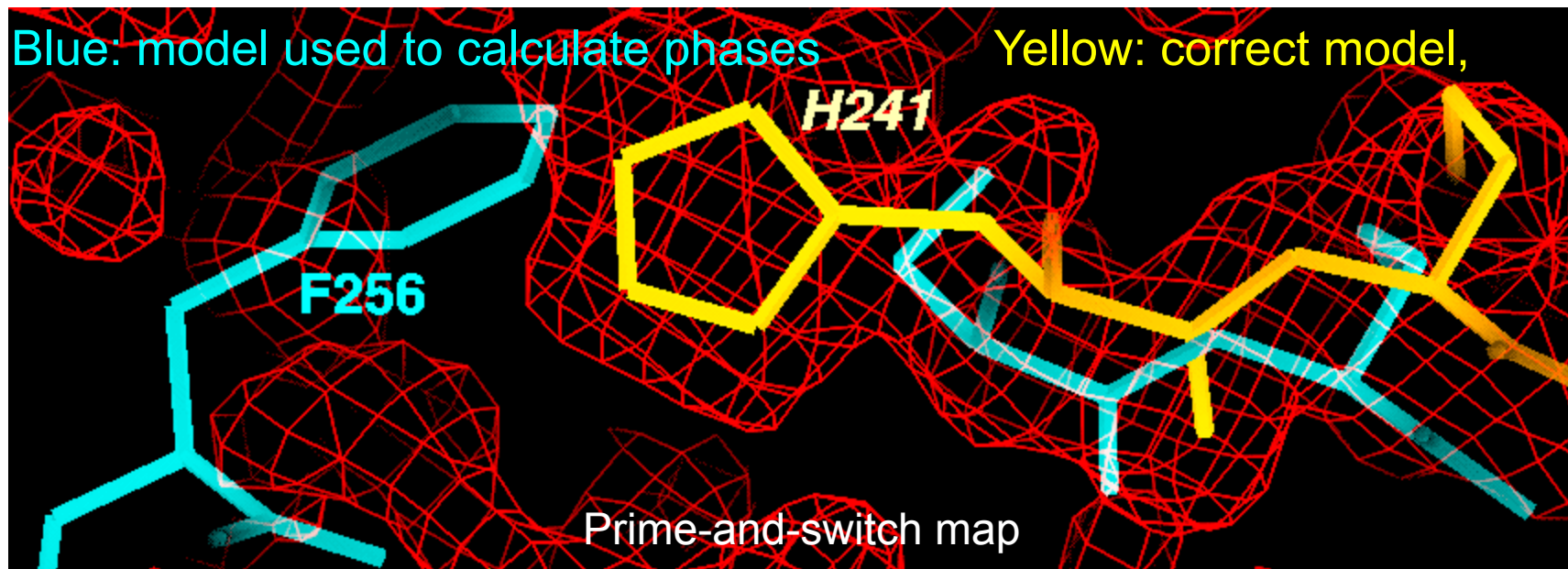
Prime-and-switch phasing

A solution:

Start with σ_A -weighted map

Identify solvent region (or other features of map)

Adjust the phases to maximize the probability of the map – **without biasing towards the model phases**

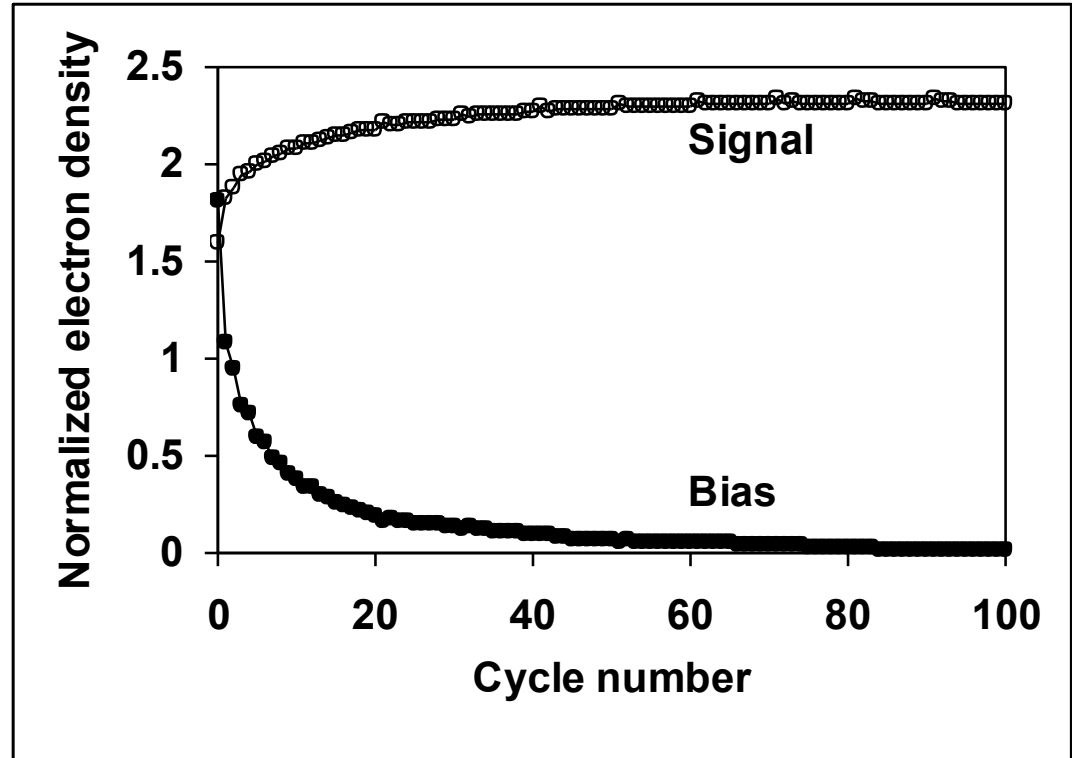


Prime-and-switch phasing

Why it should work...

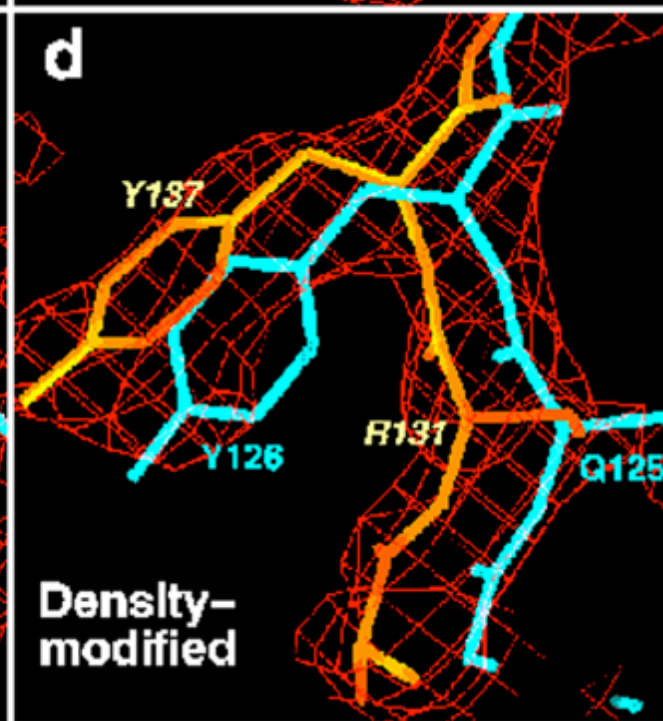
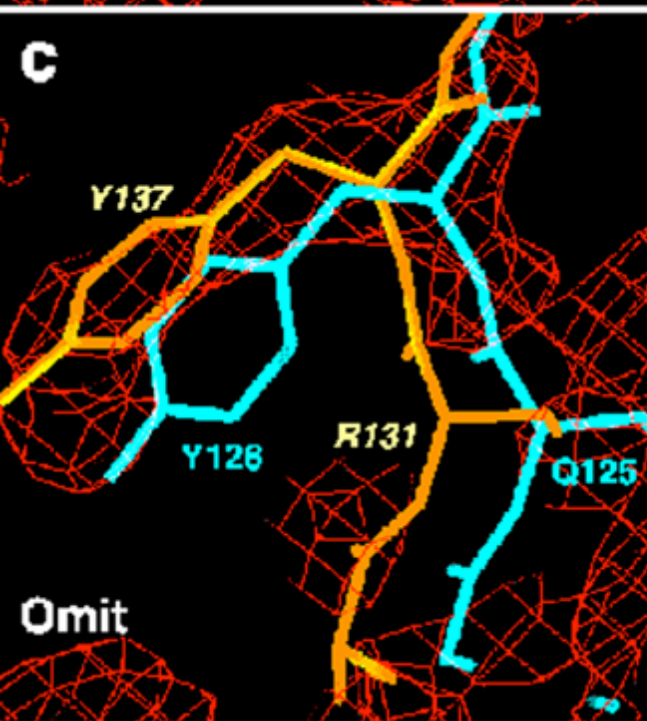
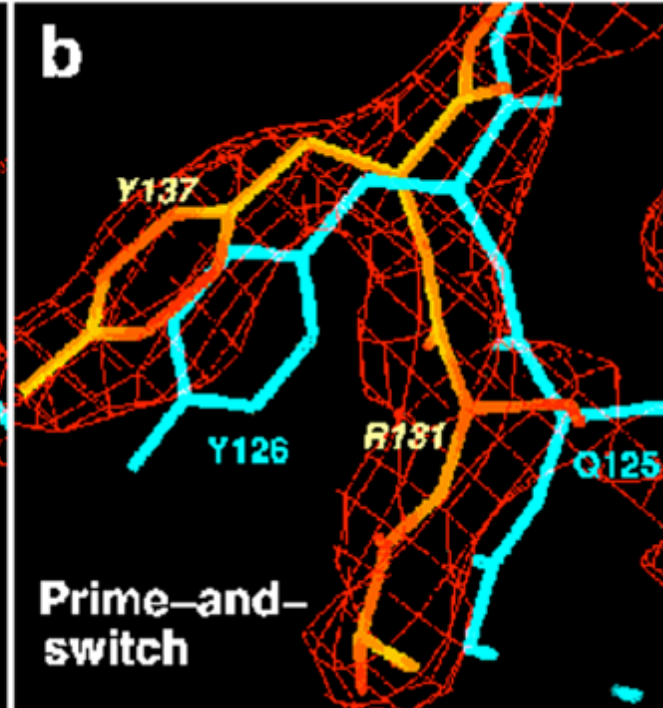
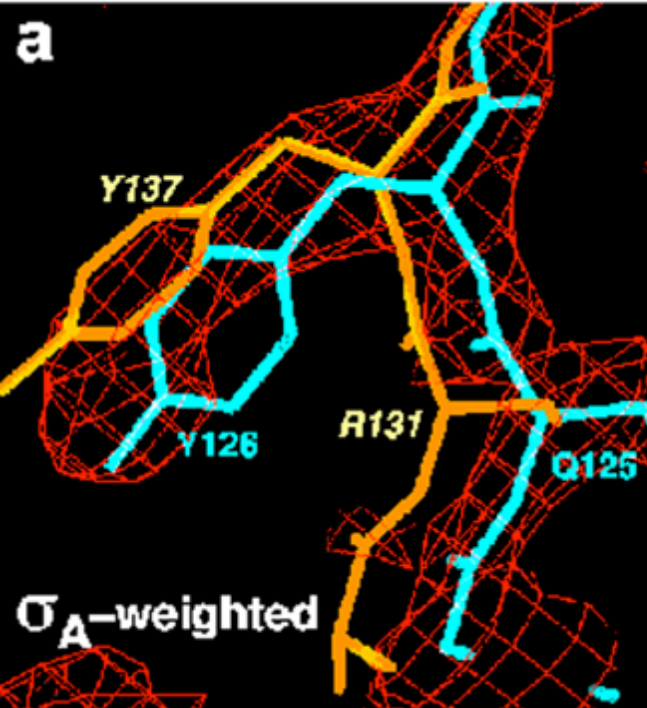
Priming: Starting phases are close to correct ones...but have bias towards misplaced atoms

Switching: Map-probability phase information comes from a different source...which reinforces just the correct phase information



Signal: peak height at correct atomic positions

Bias: peak height at incorrect atoms in starting model



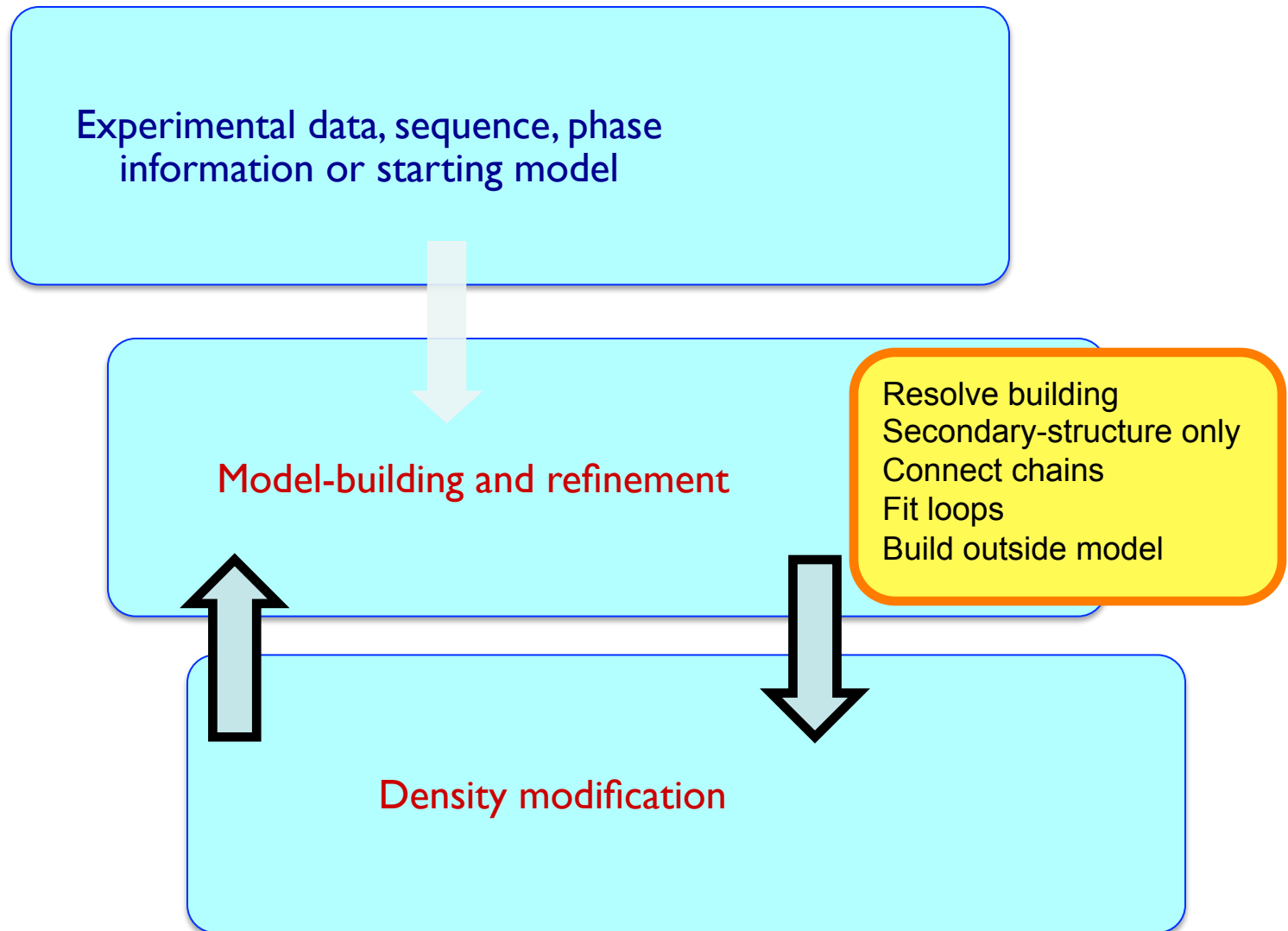
Prime-and-switch example

(IF5A, T. Peat)

Orange:
correct model

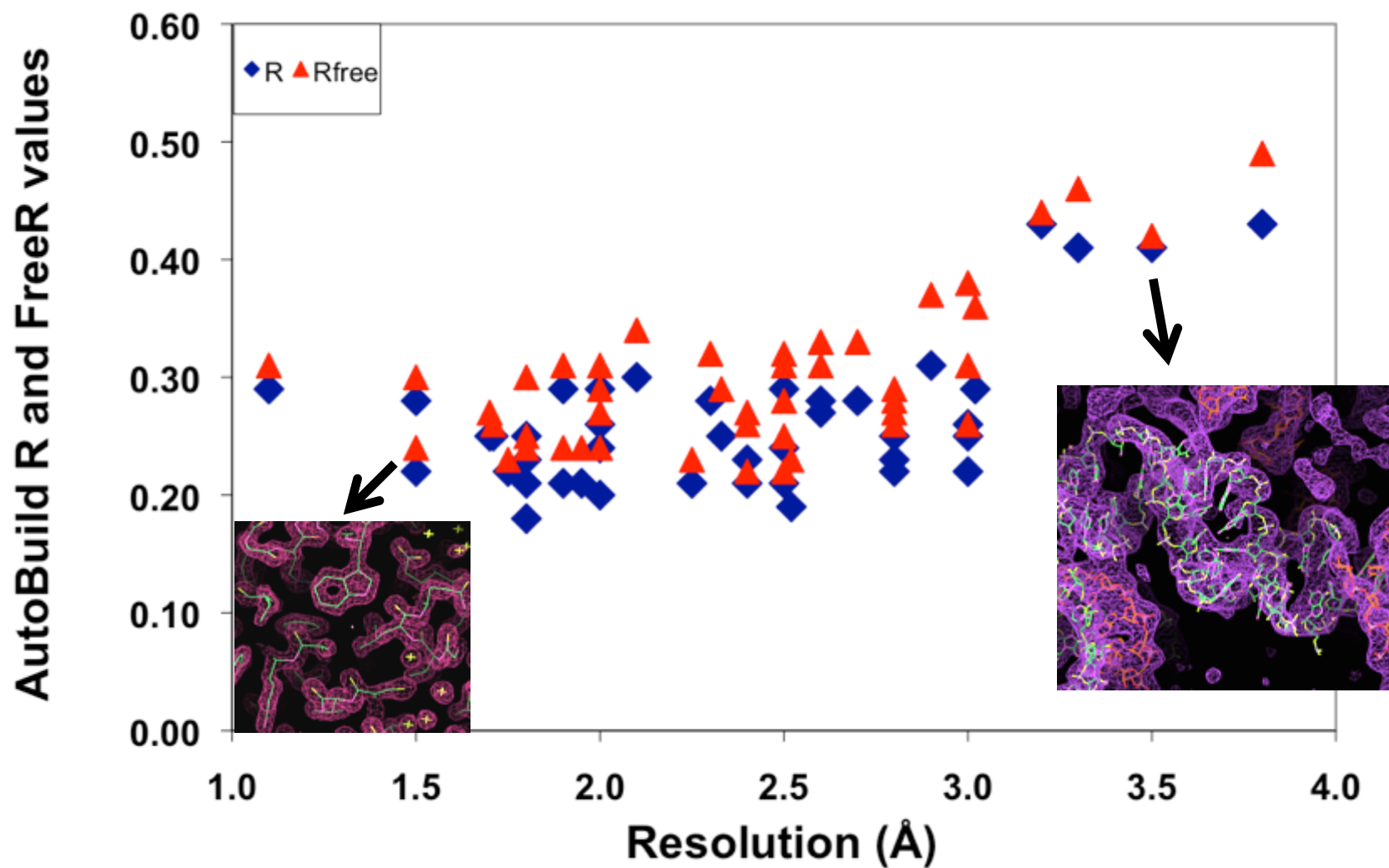
Blue: model used
to calculate
phases

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

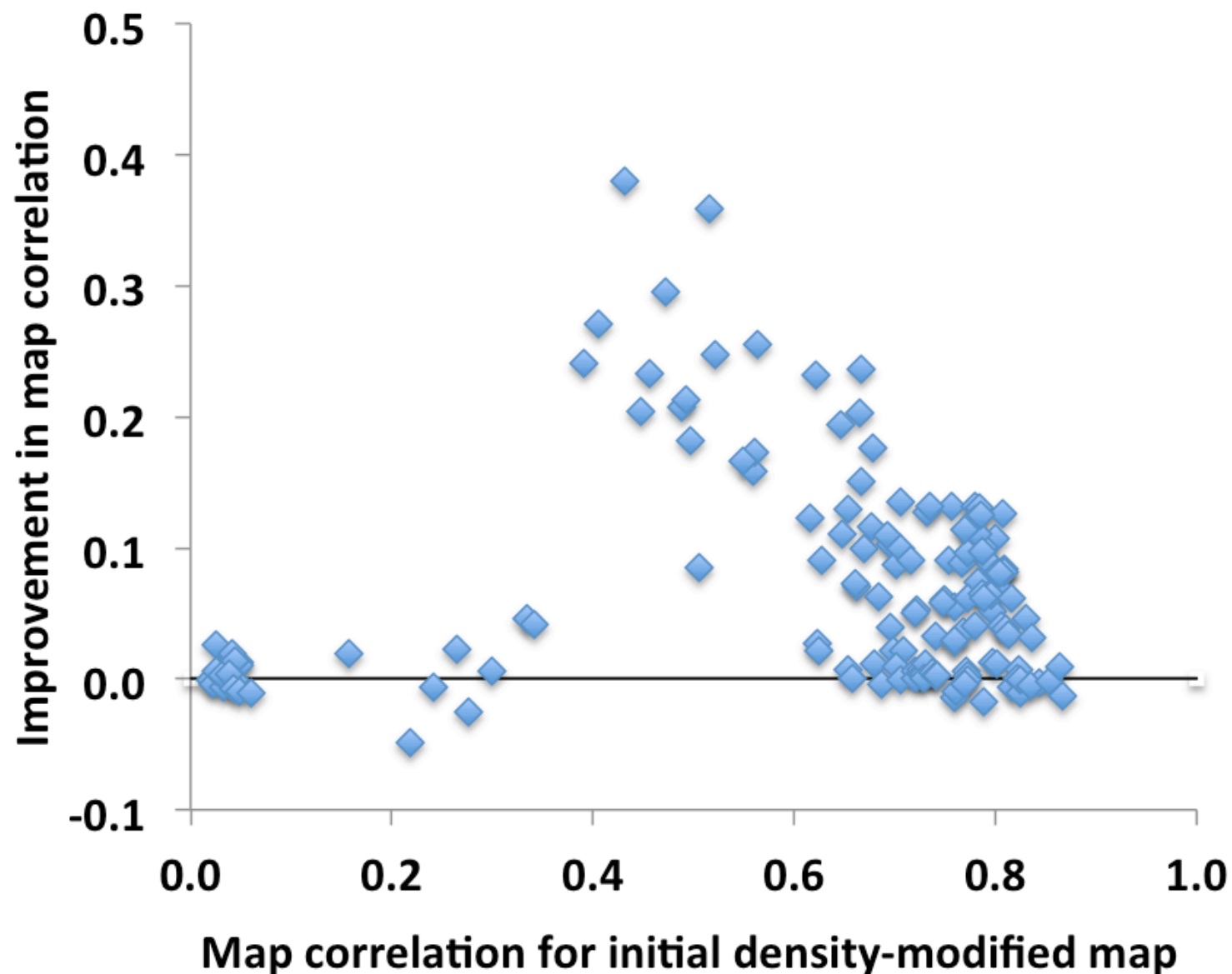


AutoBuild – tests with structure library

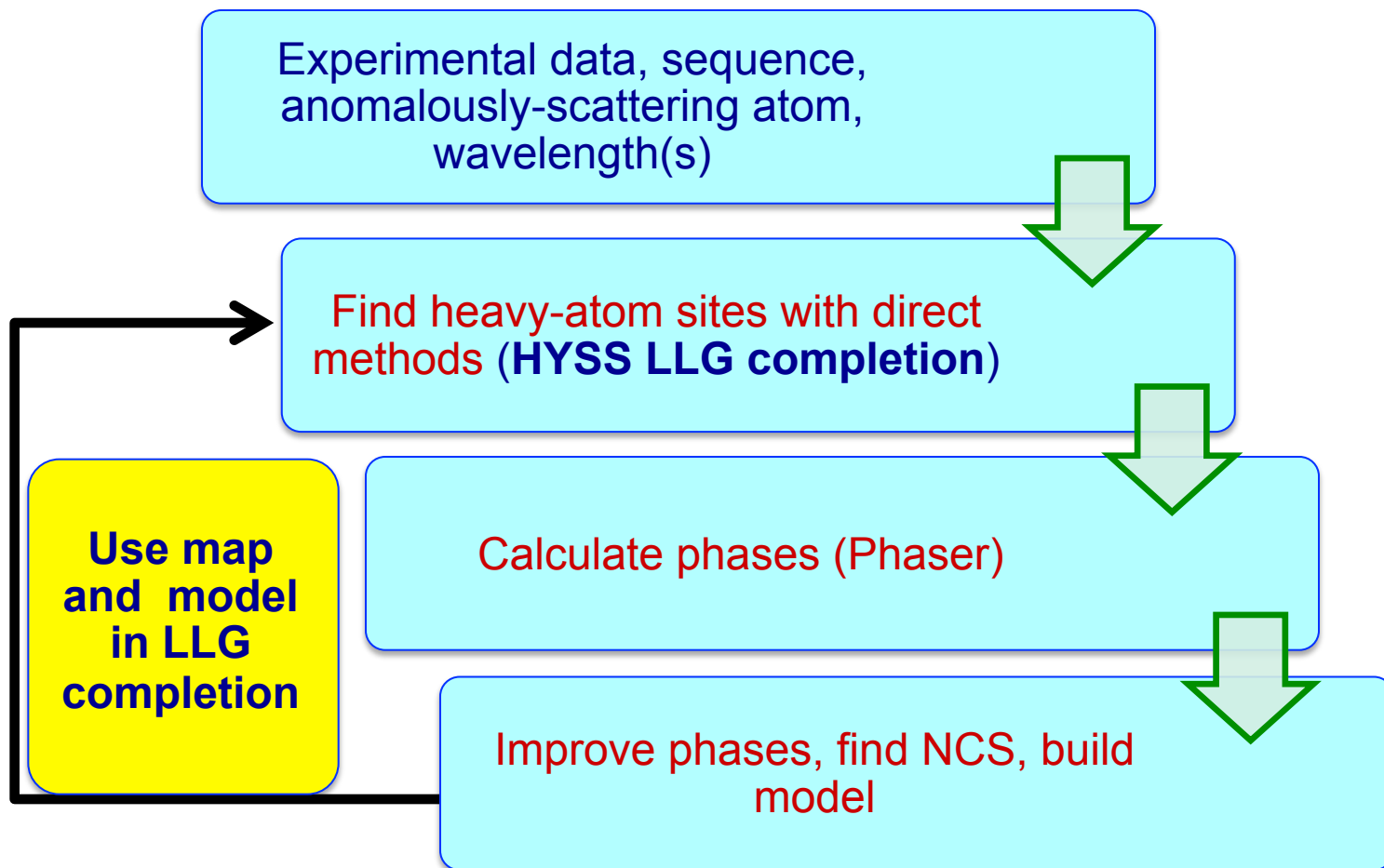
Fully automated iterative model-building, final R/Rfree



Map improvement from iteration of model-building, density modification and refinement

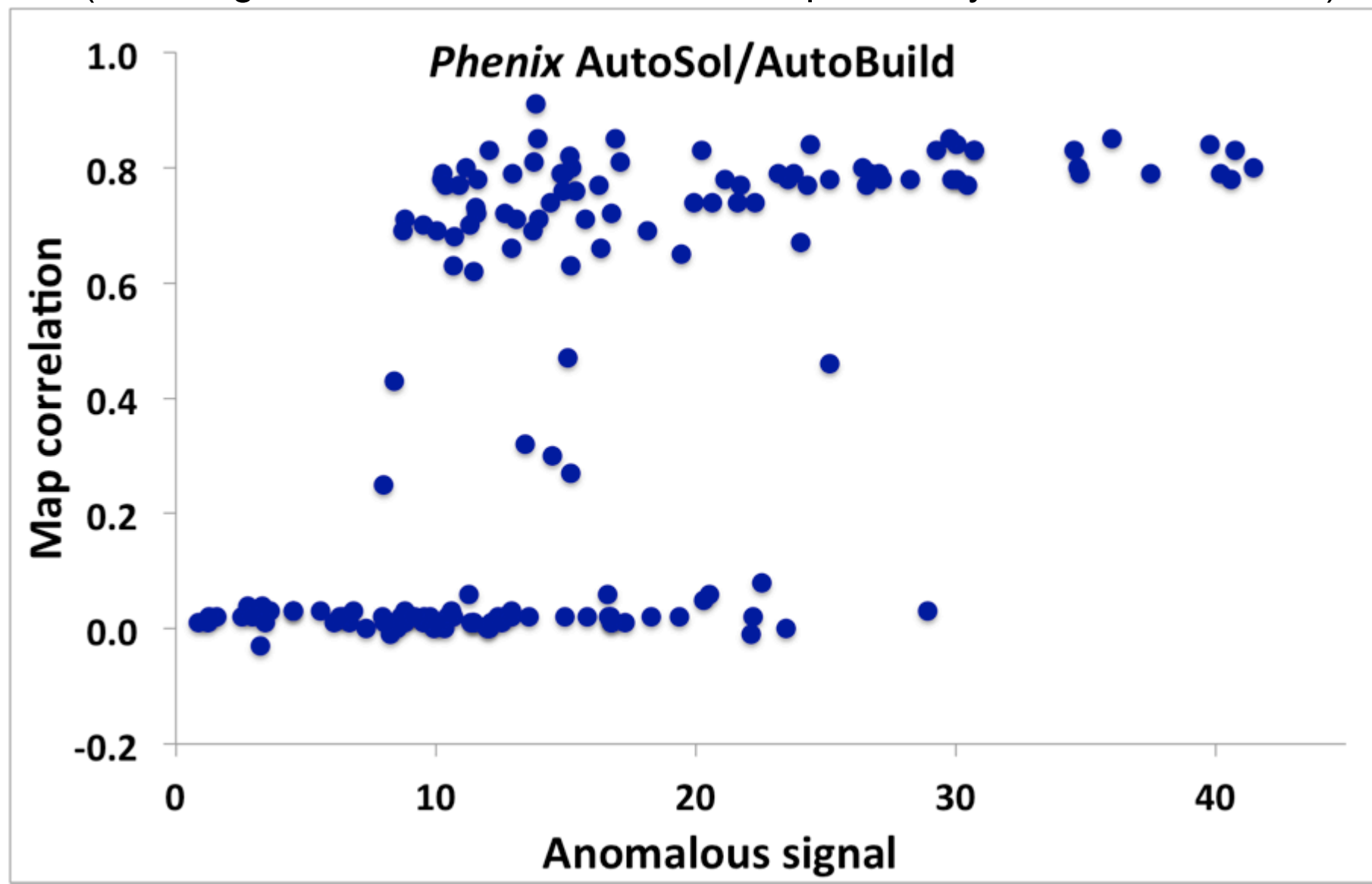


Structure solution with *Phenix*: enhancements for weak SAD data

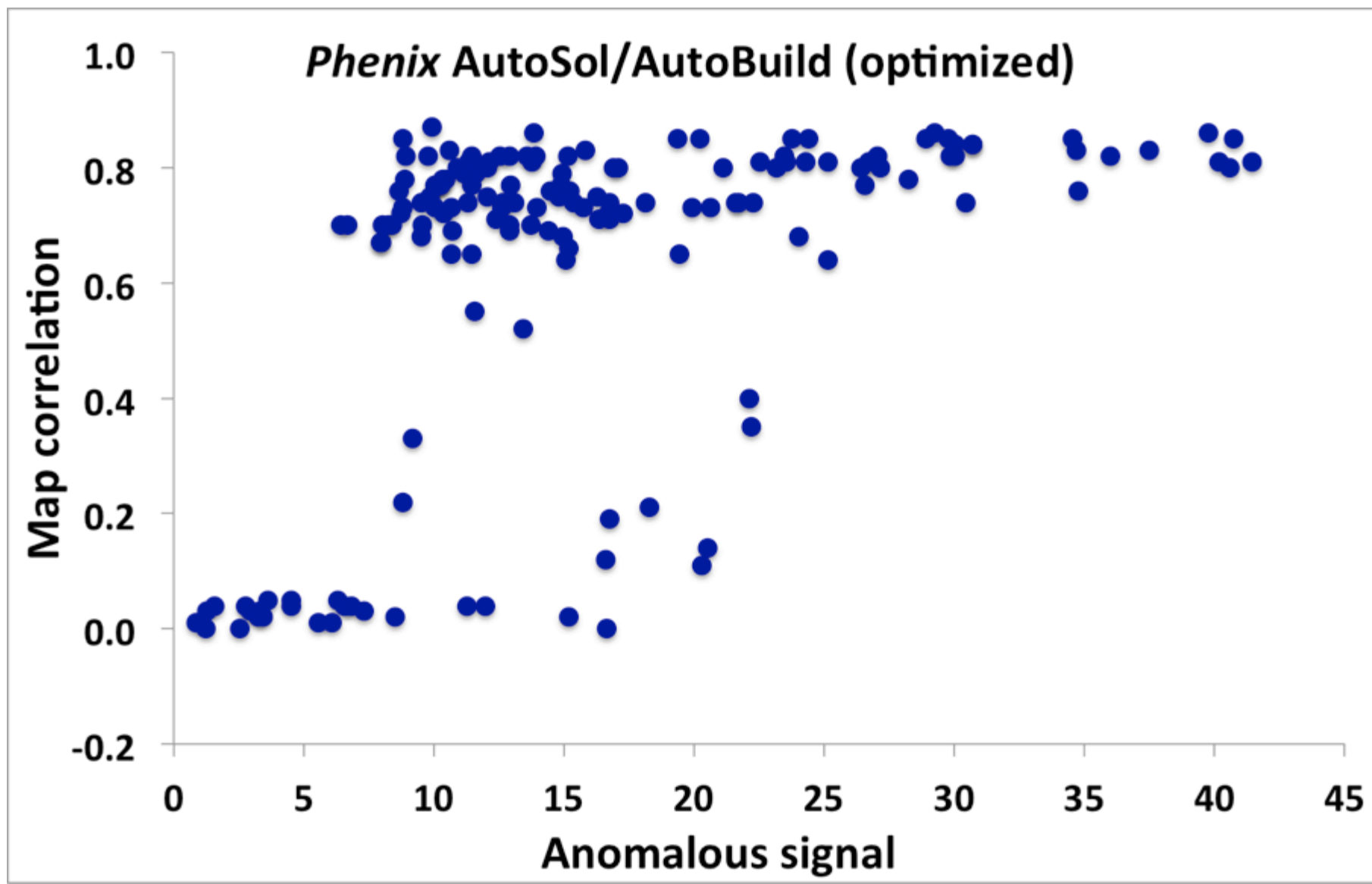


AutoSol structure solution 164 SAD datasets from PDB

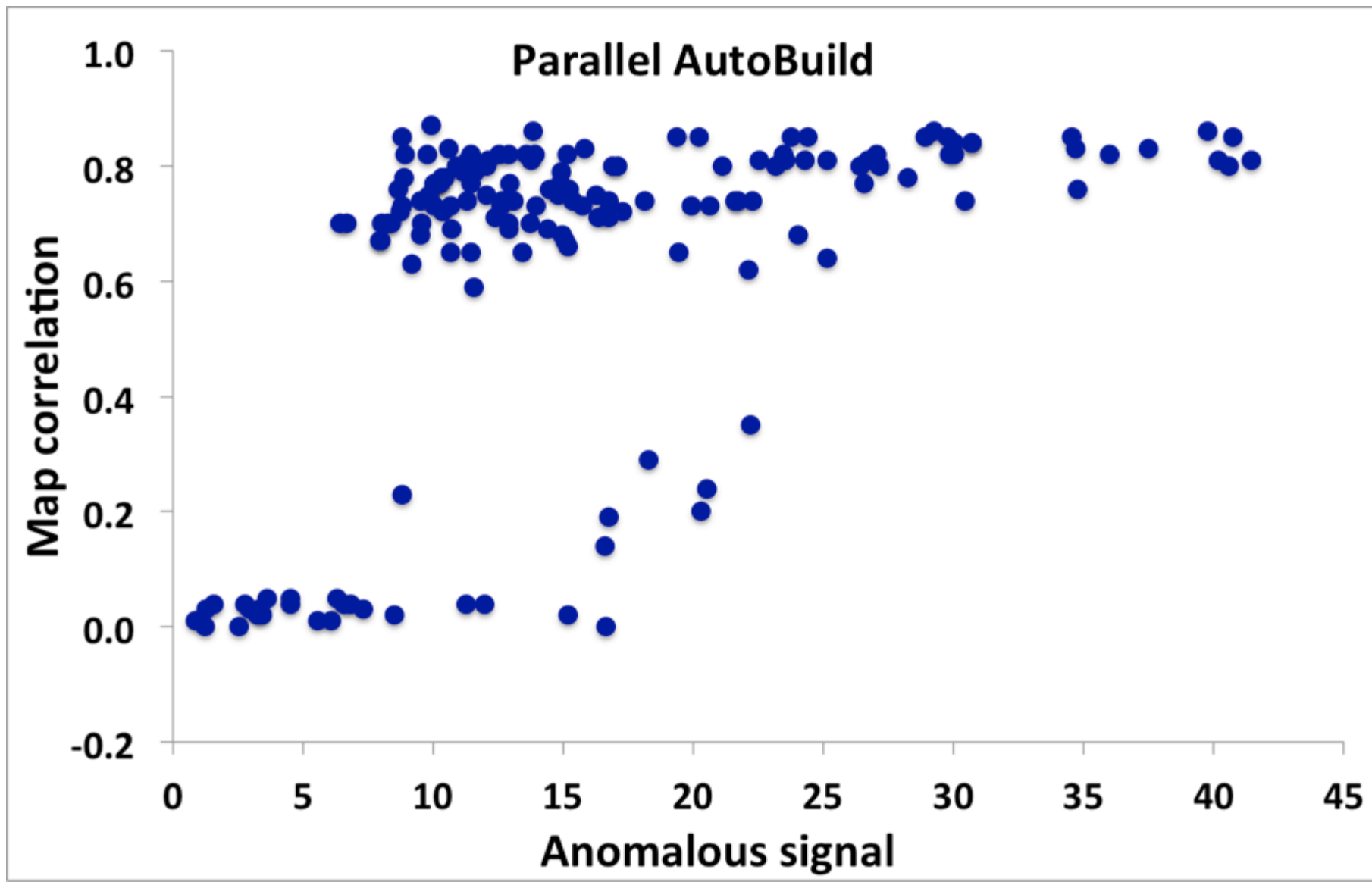
(including inflection/remote datasets not previously used as SAD data)



AutoSol structure solution
164 SAD datasets from PDB



AutoBuild model-building
164 SAD datasets from PDB



What can you do with automated procedures for structure solution and model-building?

If a task is modular and automated...

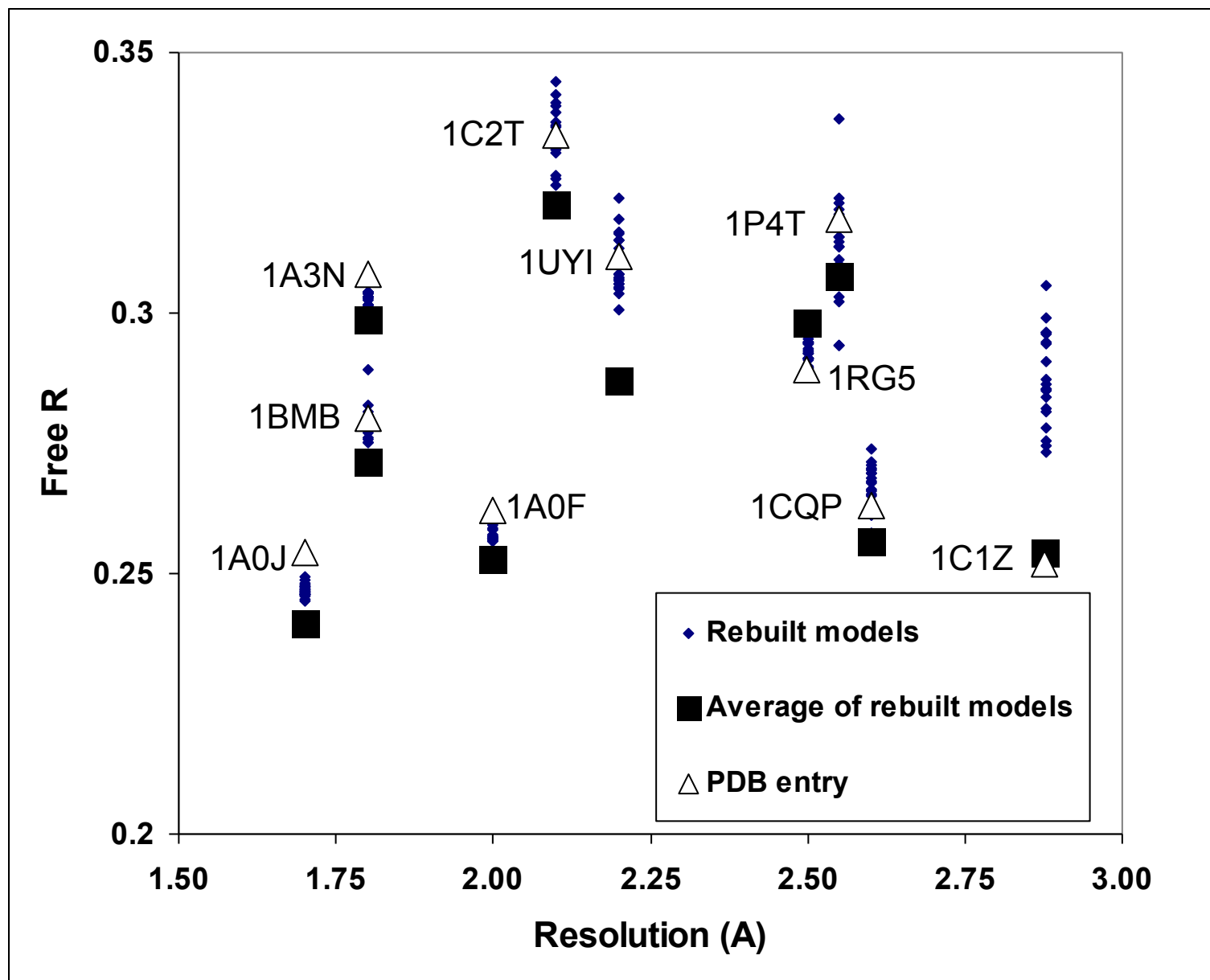
you can run it many times

...checking different space groups, datasets to use

...checking if your model is biasing your map

...checking if you always get the same model

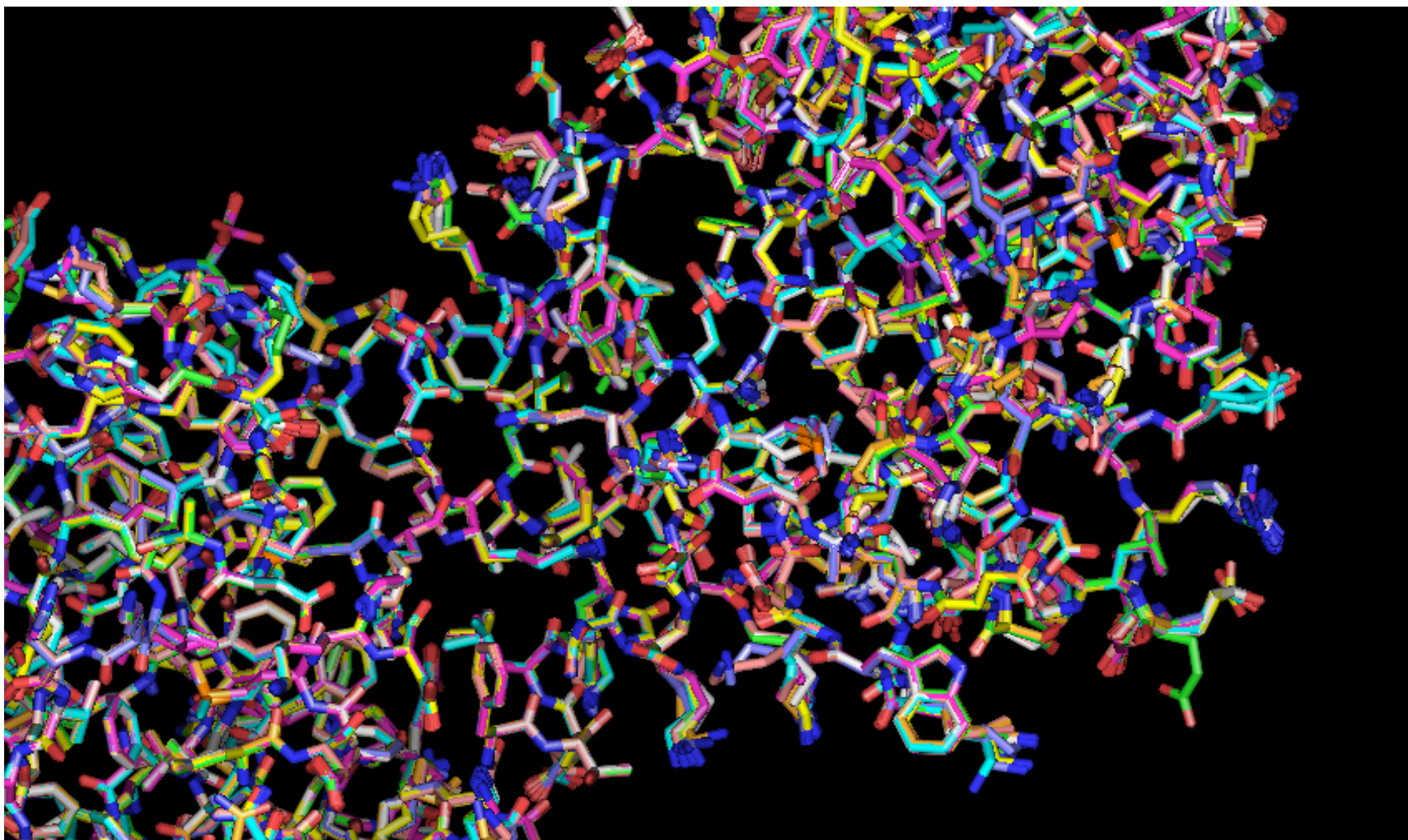
Building 20 models for each of 10 structures



Multiple-model representation of uncertainties

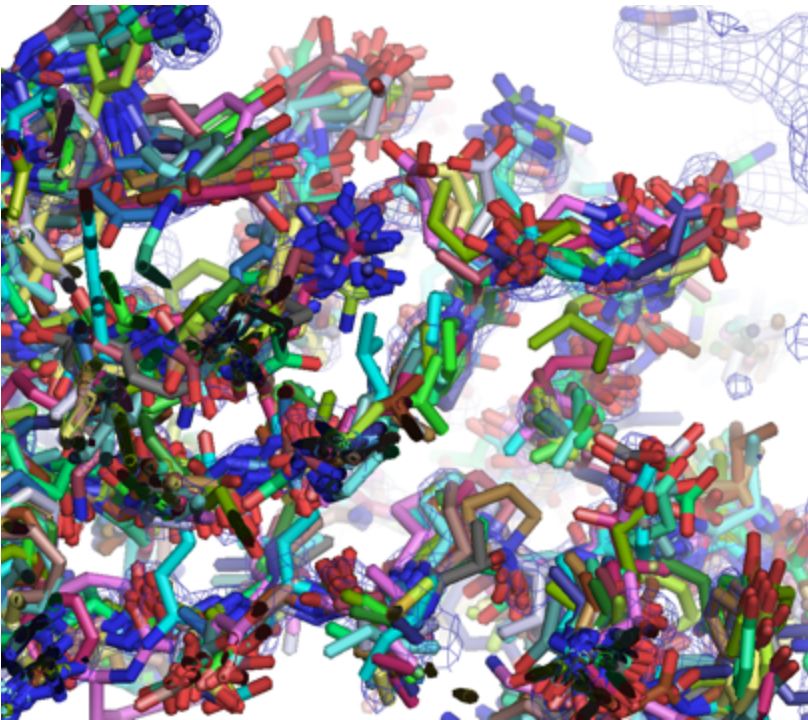
20 models built for 1CQP, no waters, $D_{\min}=2.6 \text{ \AA}$ $R=0.19-0.20$; $R_{\text{free}}=0.26-0.27$

The variation among models is a lower bound on their uncertainty

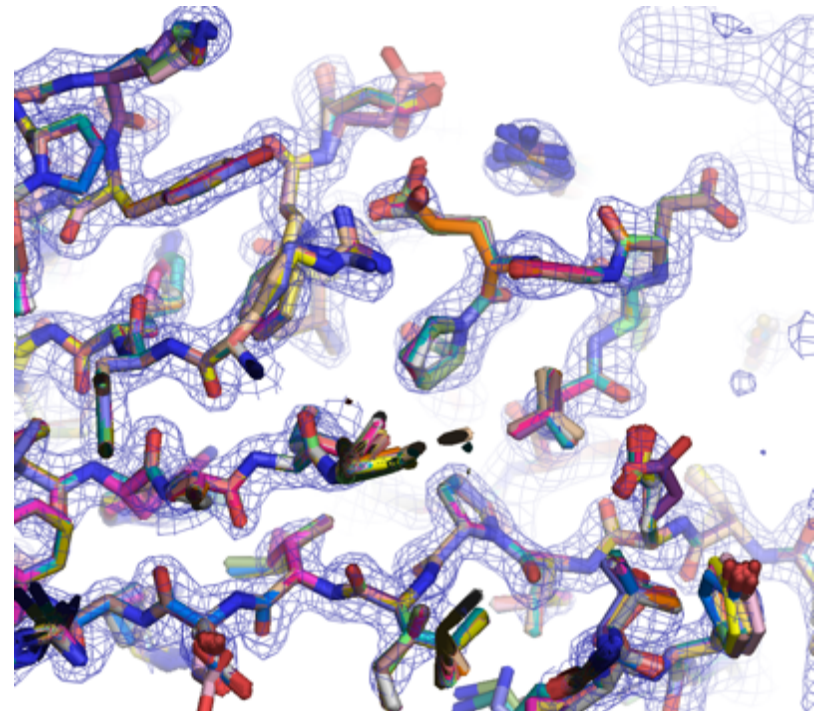


->The RMSD among models tells us (a lower bound on) the uncertainty in our models

(It is not the RMSD of true structures in the crystal)



Rebuild with 4.5 Å data



Rebuild with 1.75 Å data

The Phenix Project

Lawrence Berkeley Laboratory

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



Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung



Randy Read, Airlie McCoy, Gabor Bunkoczi, Rob Oeffner

Cambridge University



Duke University

Jane & David Richardson, Chris Williams, Bryan Arendall, Bradley Hintze



*An NIH/NIGMS funded
Program Project*

Phenix



Low-Resolution Model-building

Phenix workshop

Shanghai, China
Jan. 14, 2016

Tom Terwilliger
Los Alamos National Laboratory



Rapid building of models for regions containing regular secondary-structure

Helices:

Identification: rods of density at low resolution

Strands:

Identification: β structure as nearly-parallel pairs of tubes

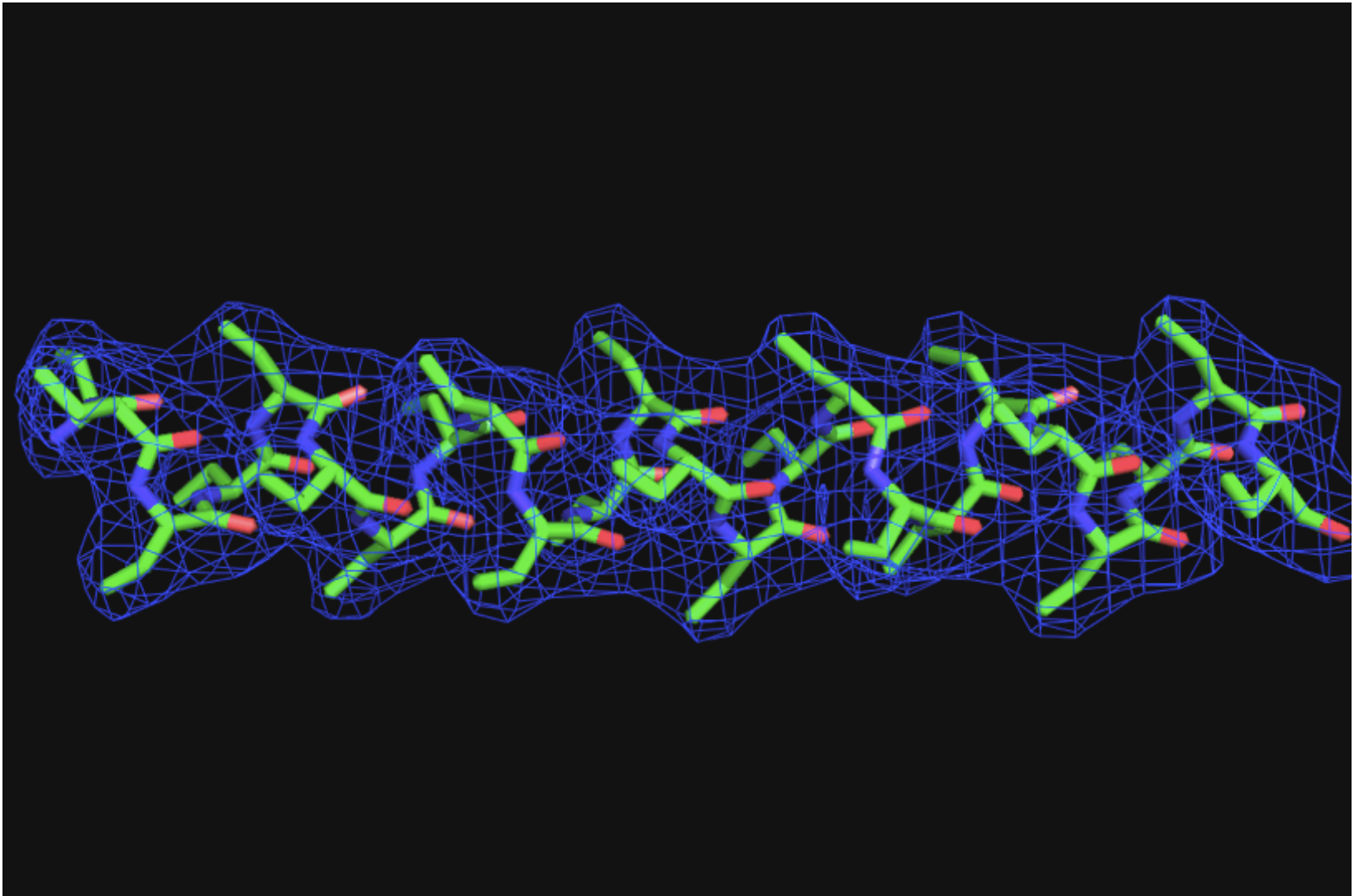
Any protein chains (trace_chain):

Identification: $C\alpha$ positions consistent with density and geometry of protein chains

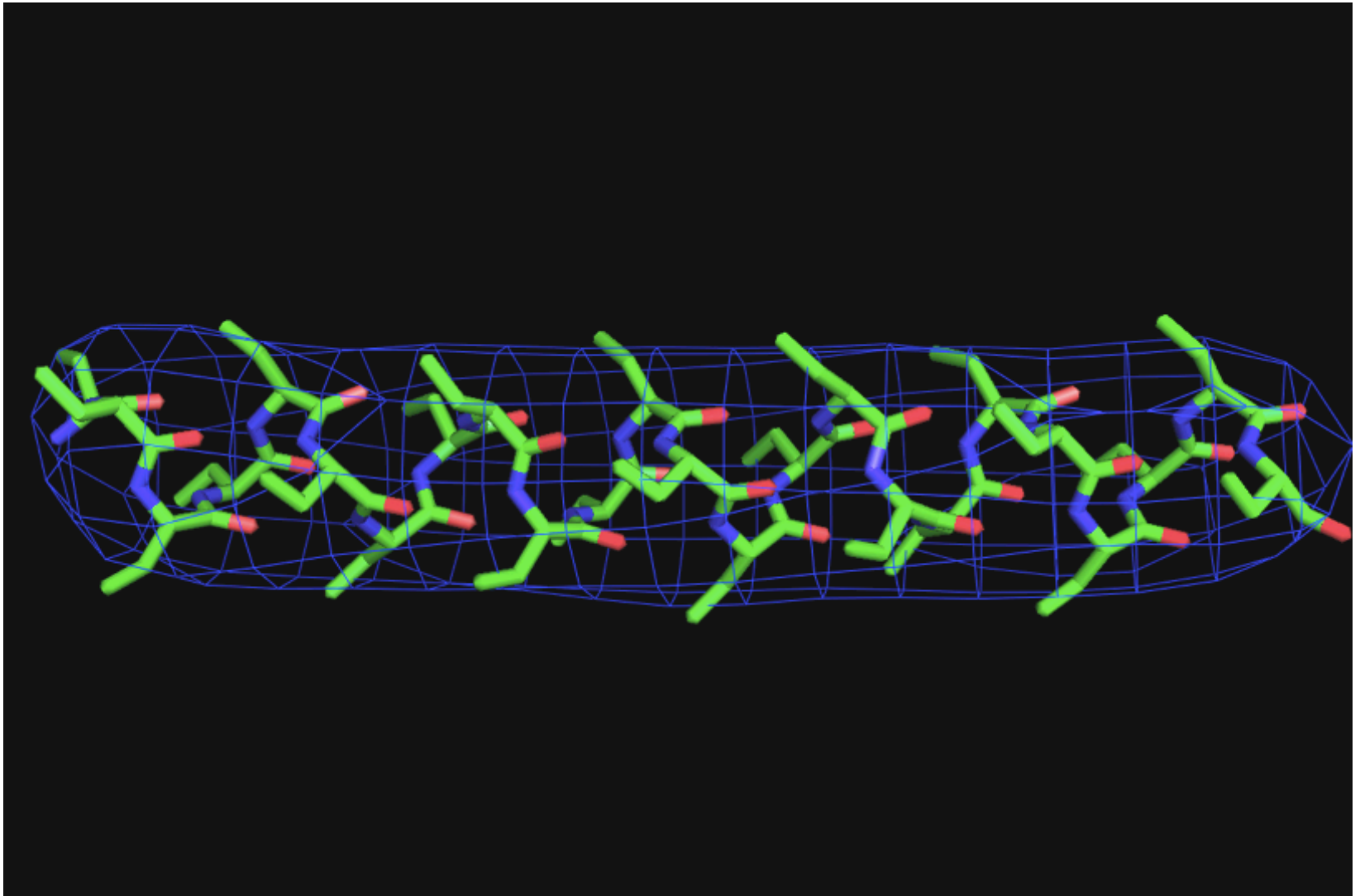
RNA/DNA:

Identification: match of density to averaged A or B-form template

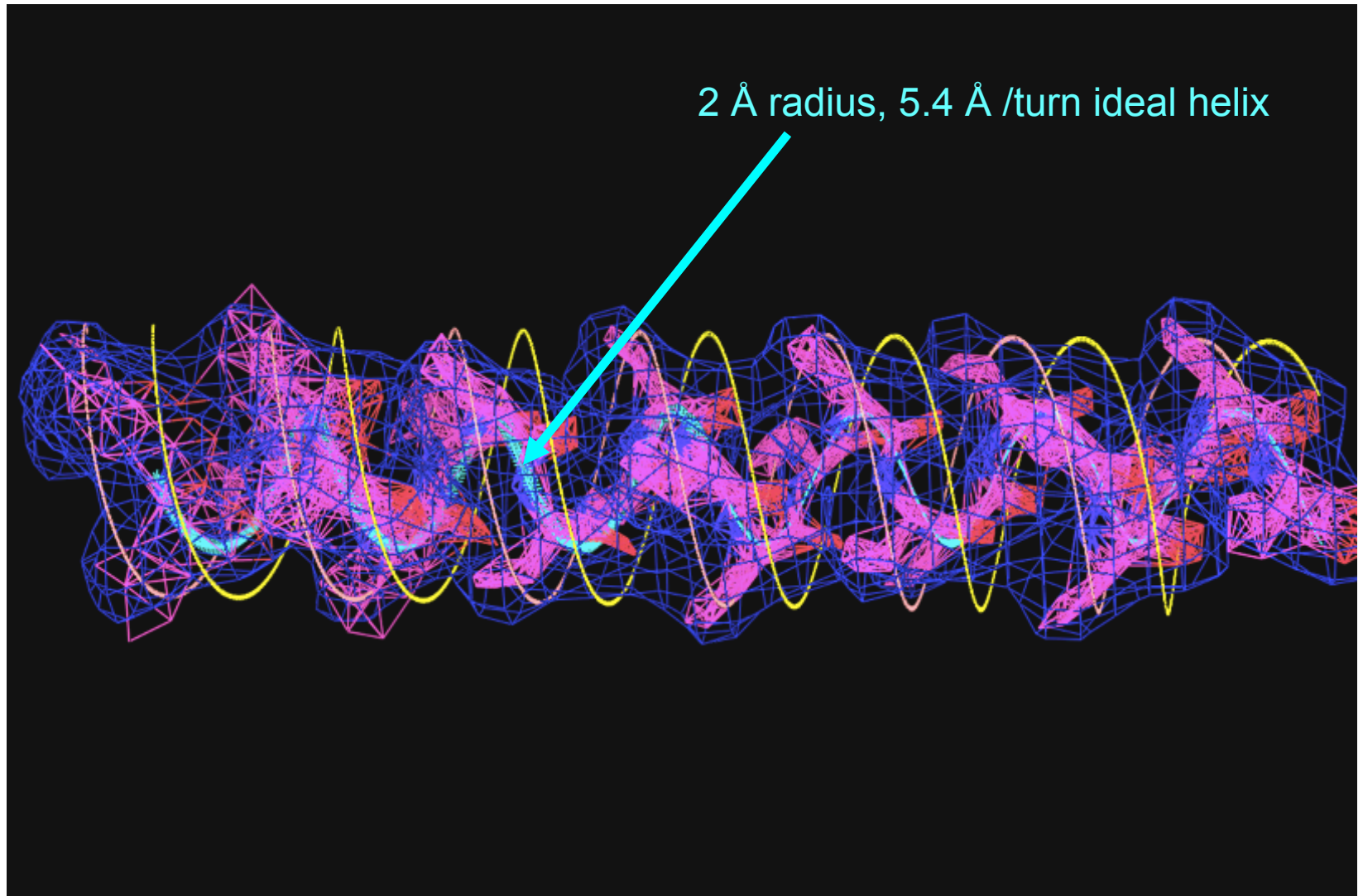
Model α -helix; 3 Å map



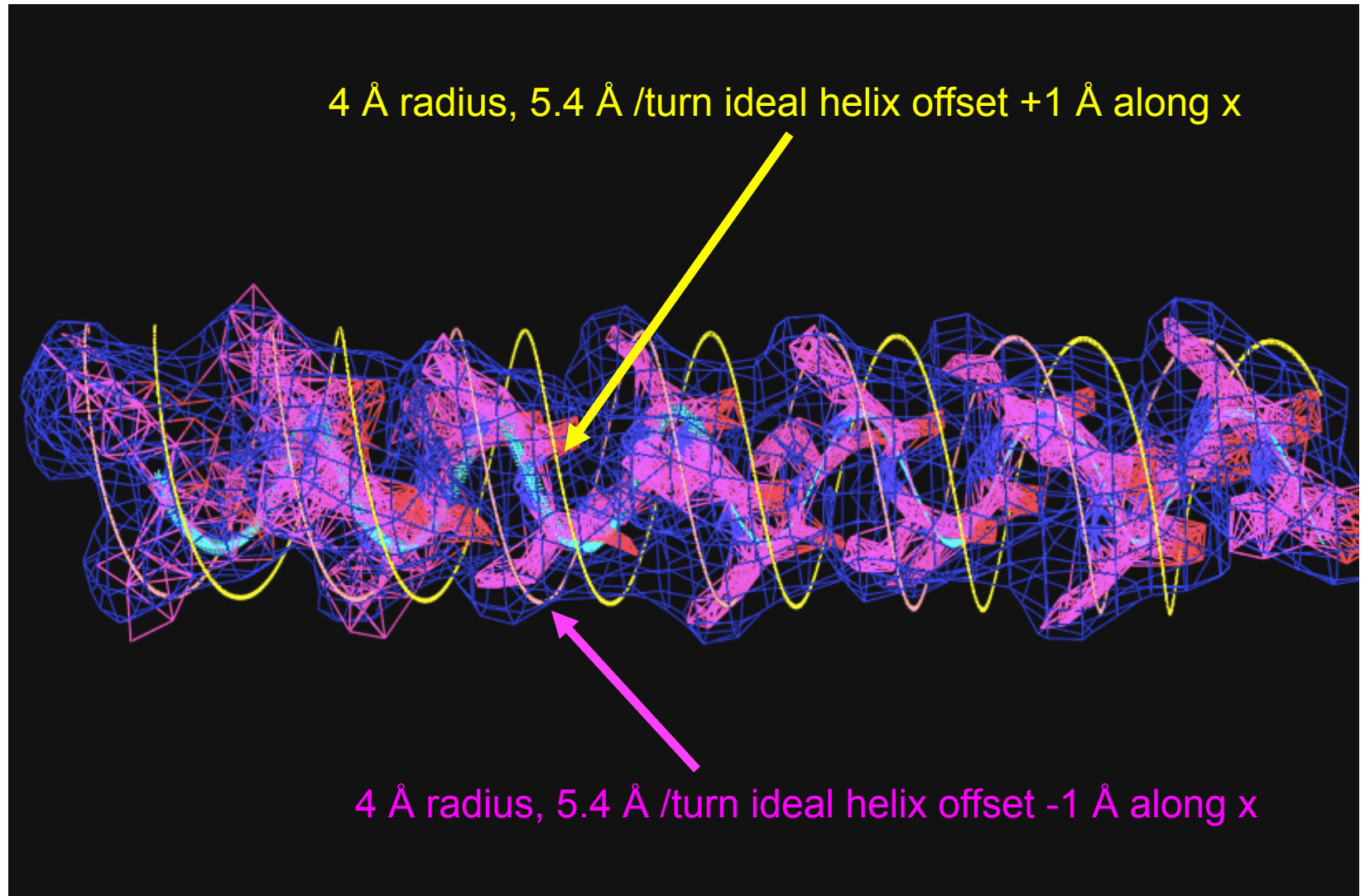
Model α -helix; 7 Å map



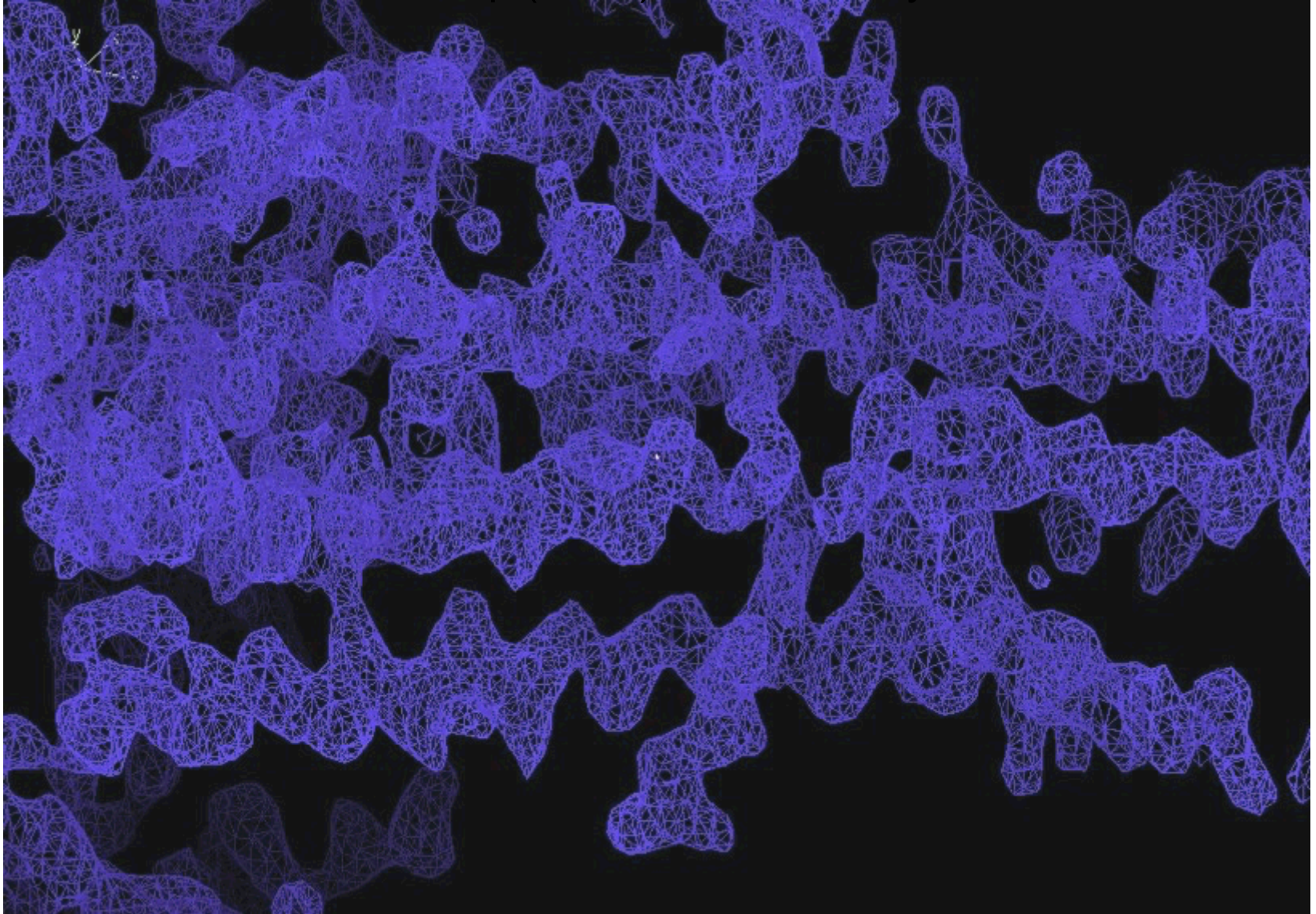
Trace main-chain with ideal helix, allowing curvature



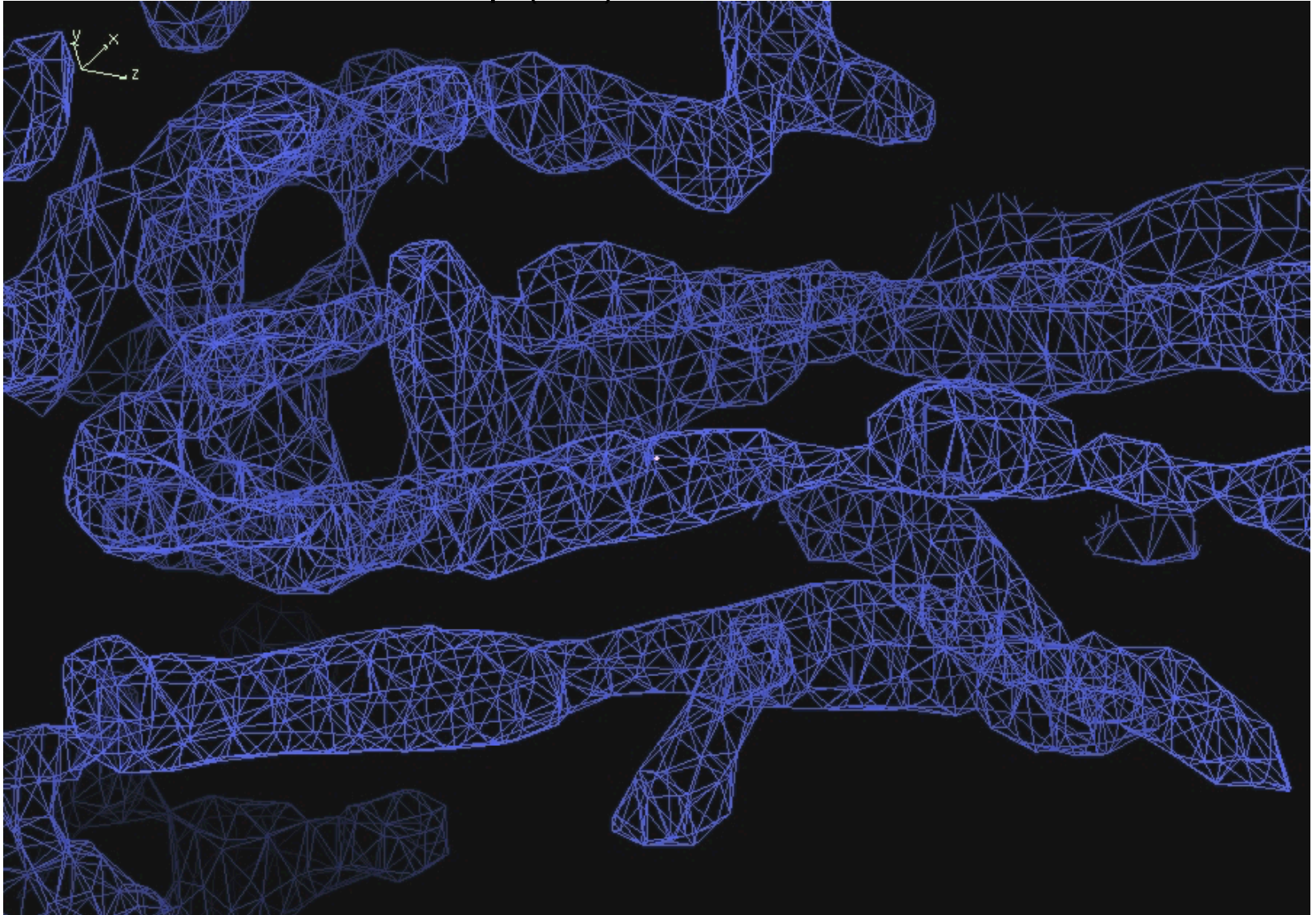
Identify direction and $C\alpha$ position from overlap with 4 Å radius helices offset ± 1 Å from main-chain



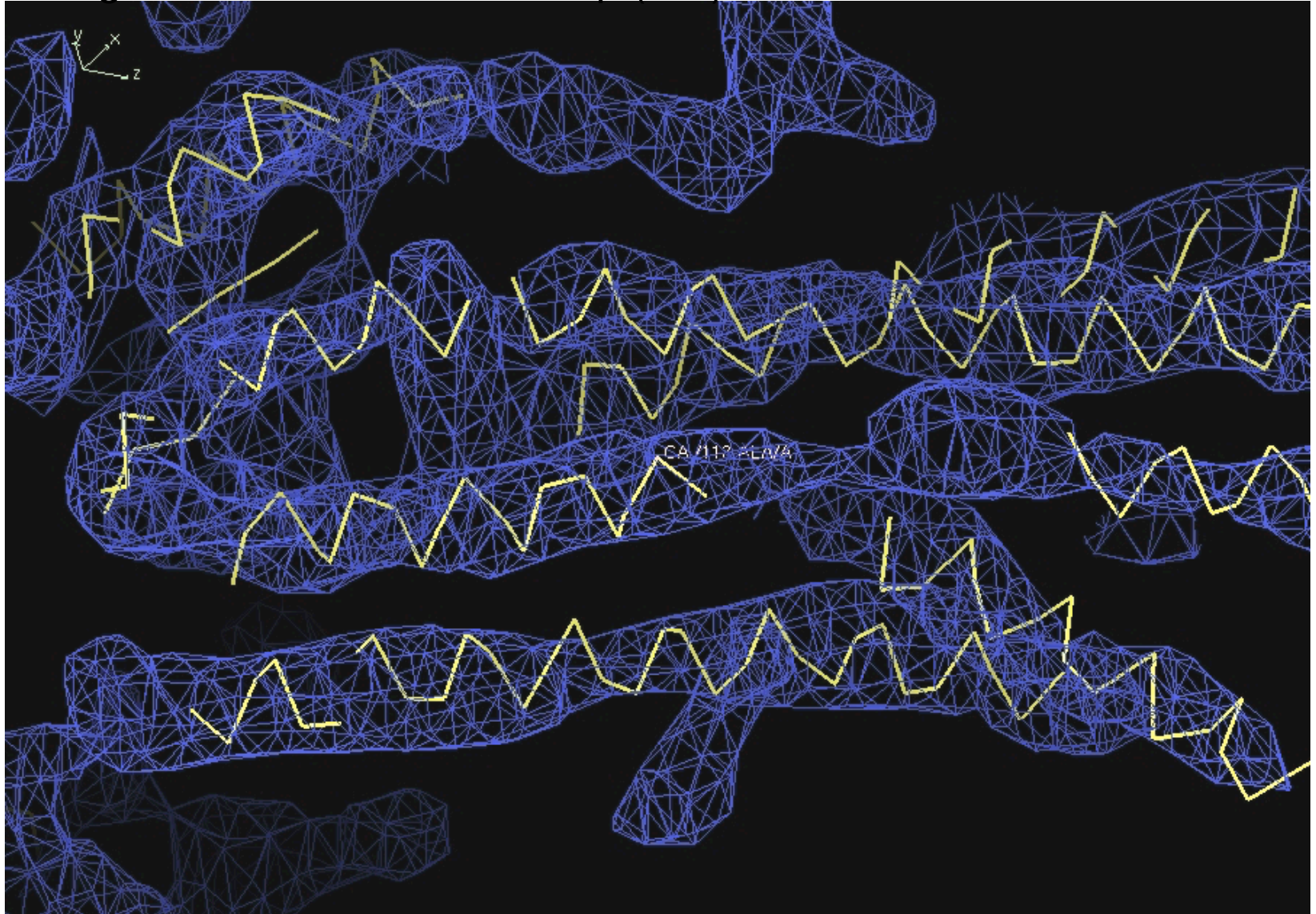
A real case: 1T5S SAD map (3.1 Å) Data courtesy of P. Nissen



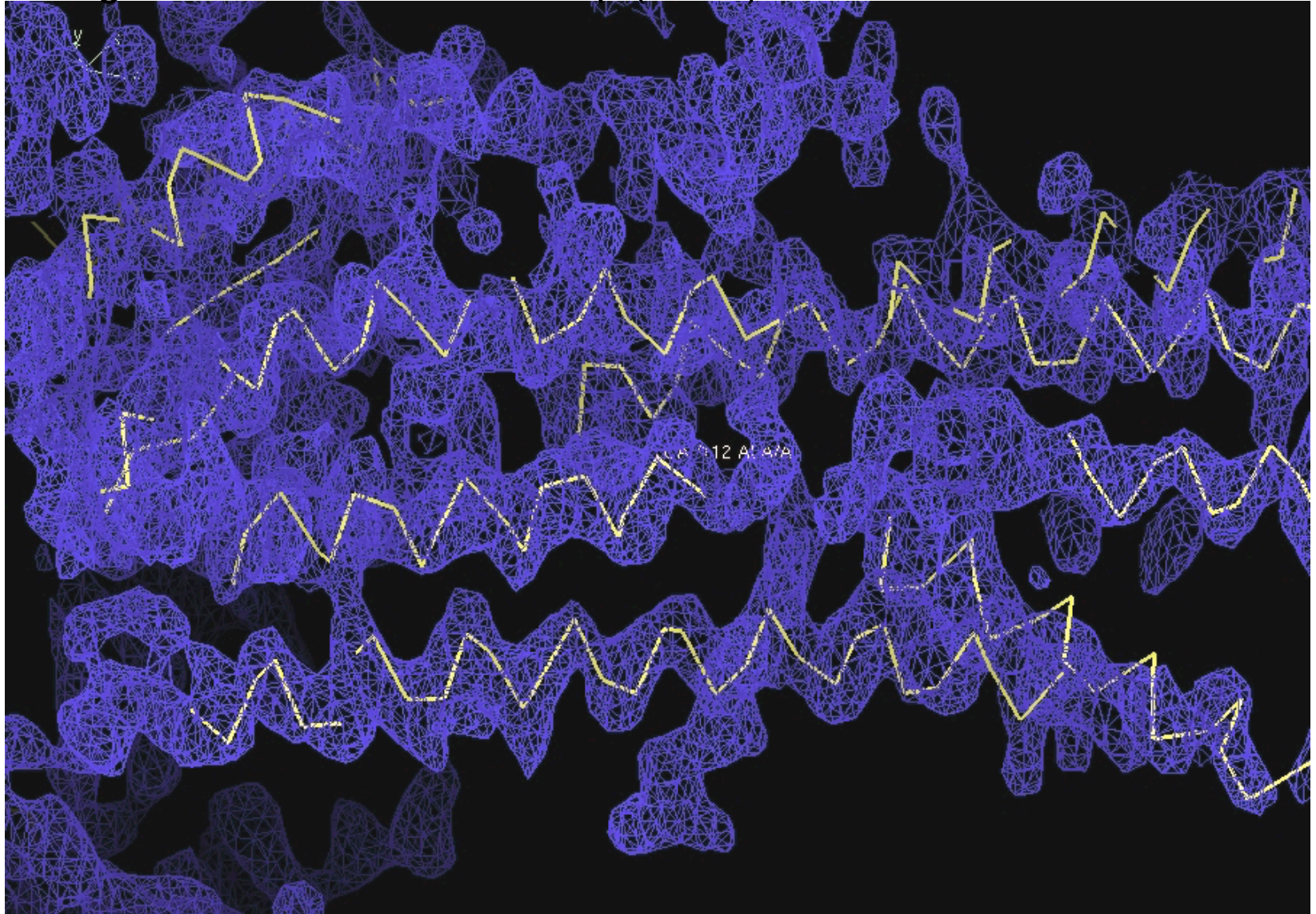
A real case: 1T5S SAD map (7 Å)



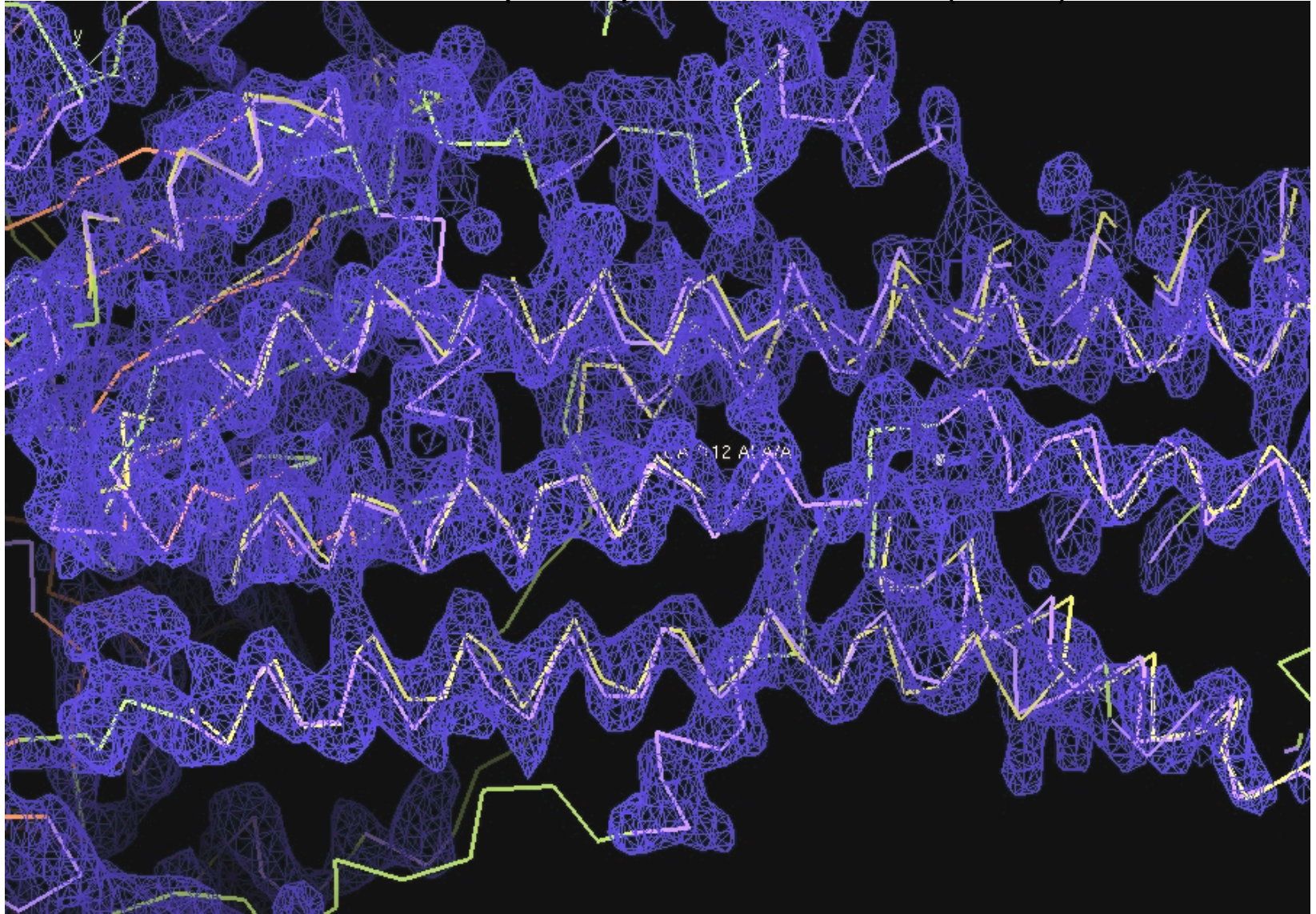
Finding helices in 1T5S SAD map (7 Å)



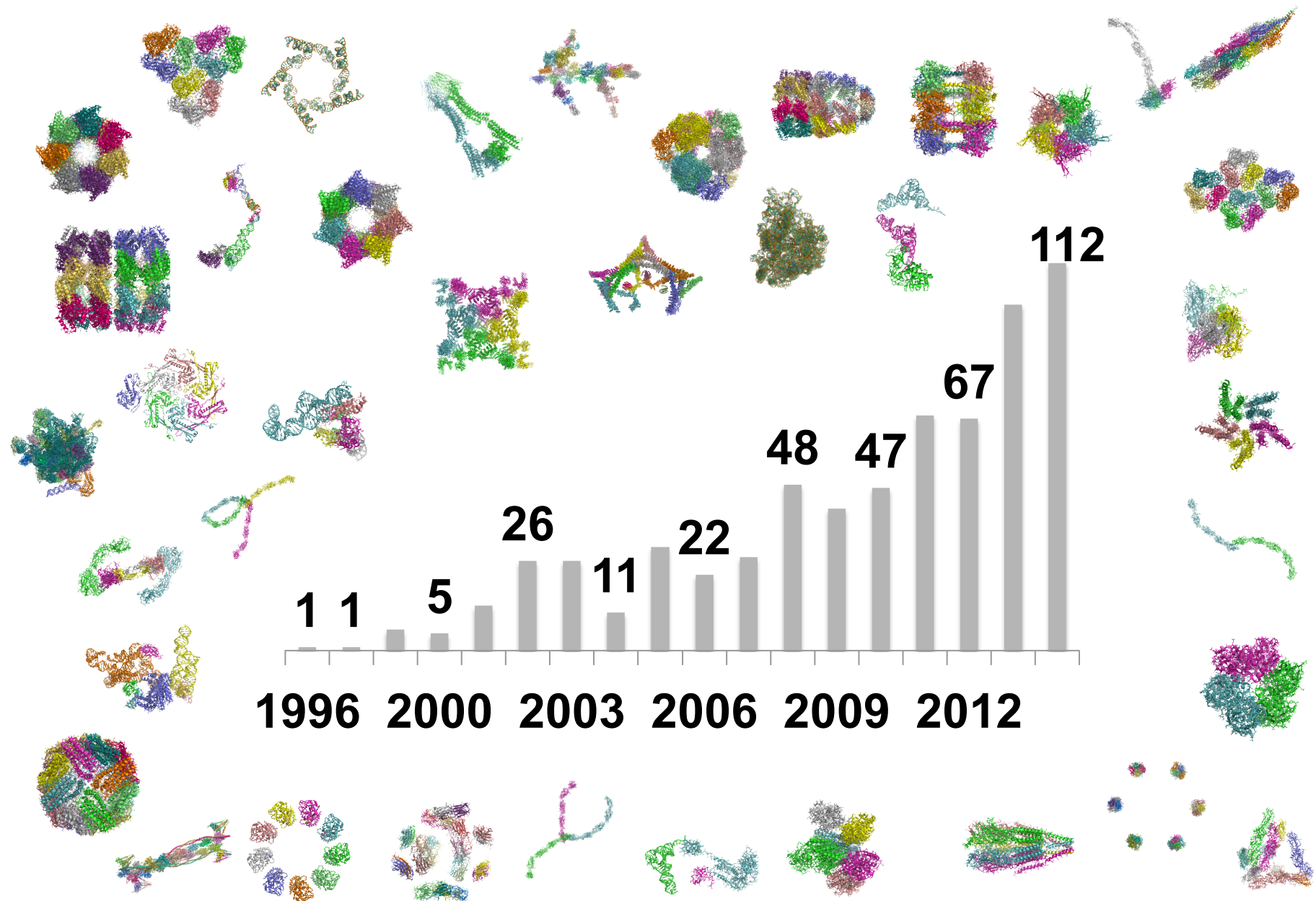
Finding helices in 1T5S SAD map (3.1 Å)



Helices from 1T5S SAD map compared with 1T5S (3.1 Å)

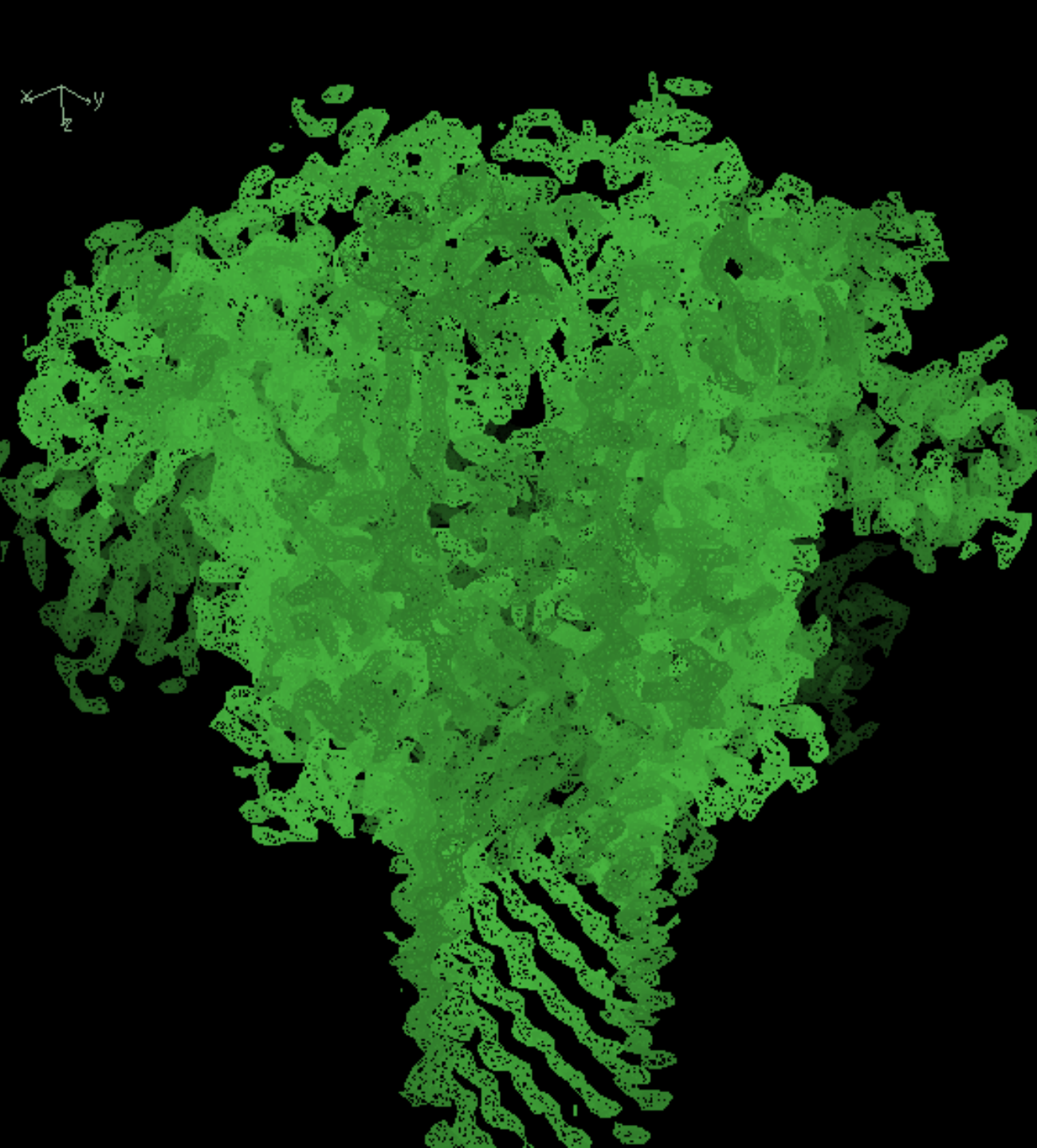


About 750 Cryo-EM structures in PDB

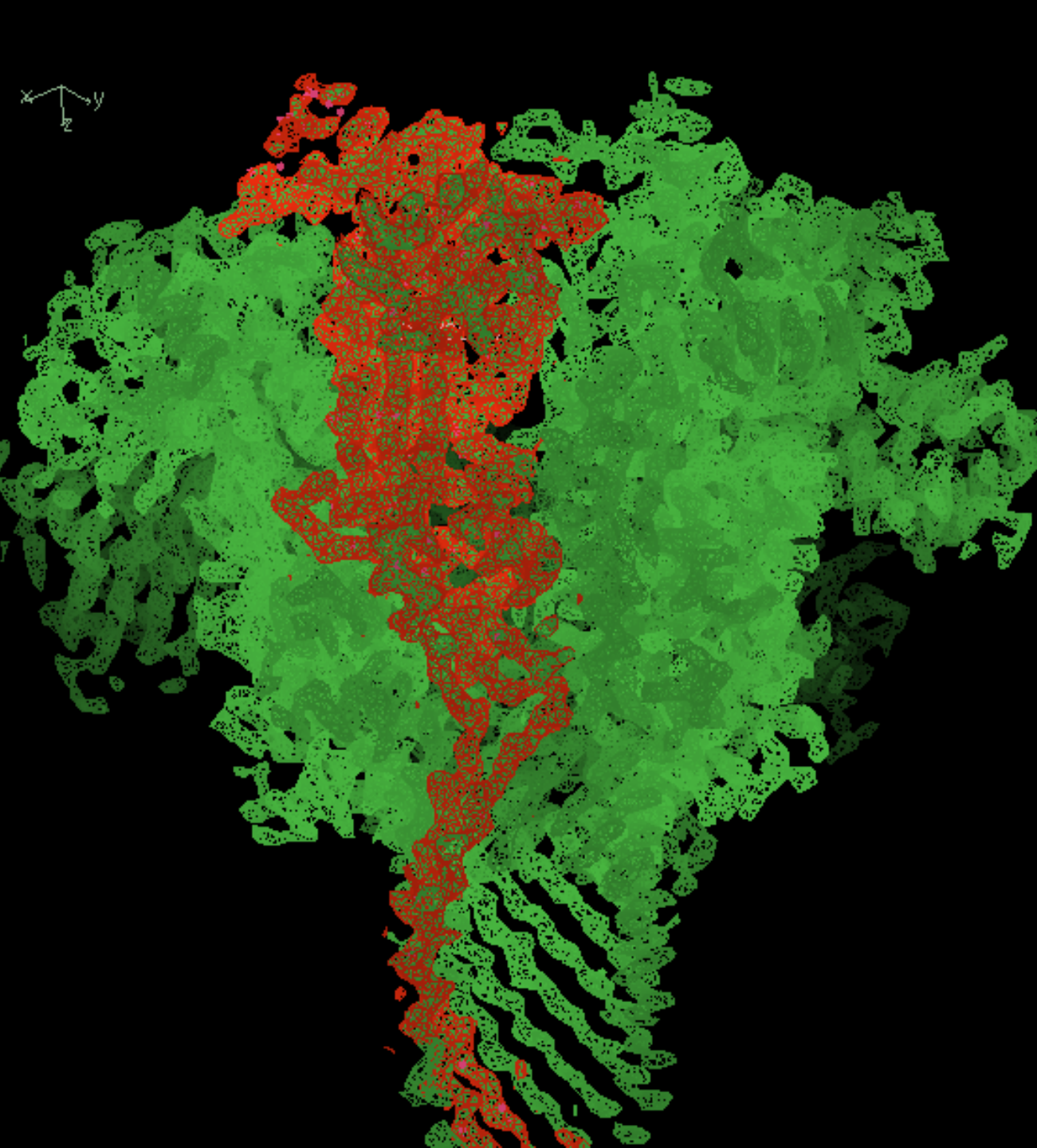


Building into cryo-EM maps

- Automatically segment maps and extract asymmetric unit of reconstruction
- Create maps emphasizing information at various resolutions by variable map sharpening
- Trace protein main chain using nearly-constant C α -C α -C α distances and angles
- Identify direction of main-chain in models by fit to density

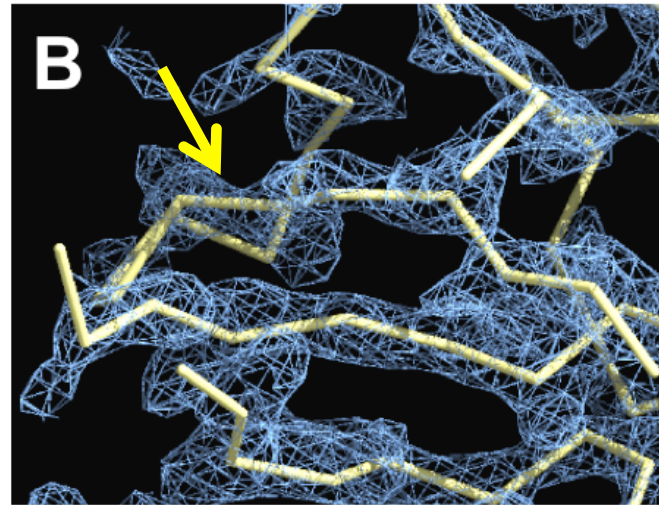
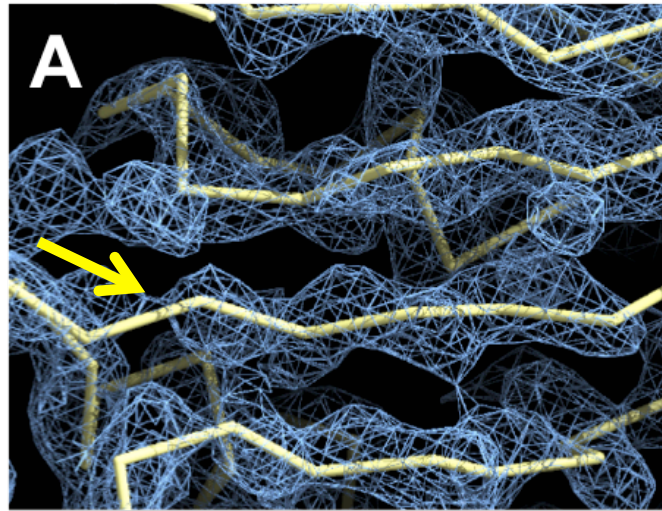


Automated
segmentation of
emd_6224 (anthrax
toxin protective
antigen pore at 2.9
Å; Jiang et al 2015)

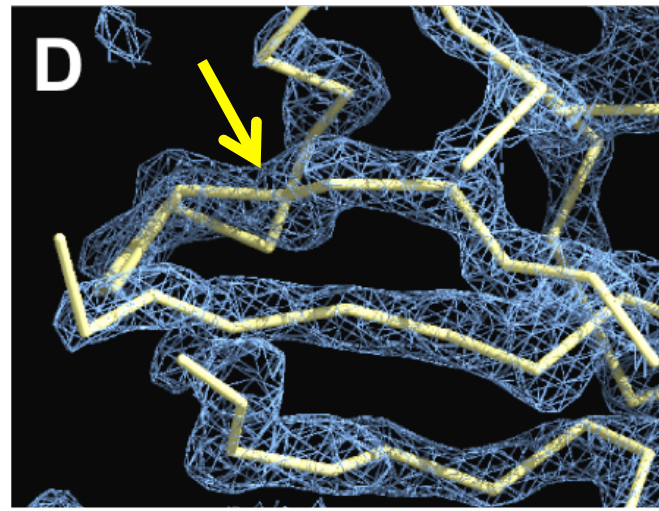
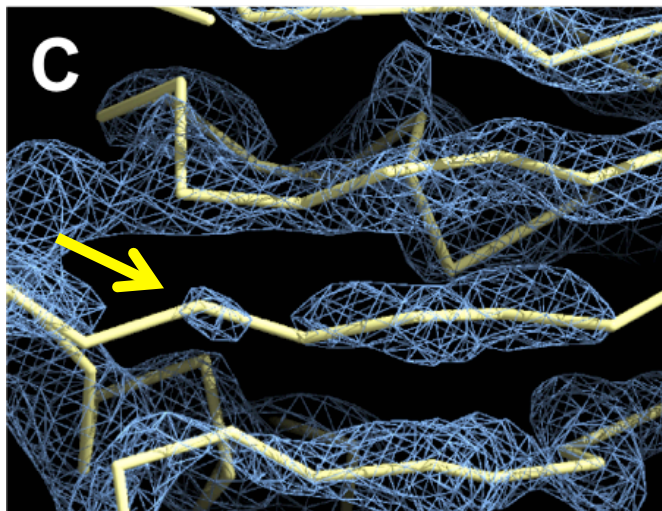


Automated
segmentation of
emd_6224 (anthrax
toxin protective
antigen pore at 2.9
Å; Jiang et al 2015)

Accurate low-resolution information in cryo-EM maps



Original



Blurred

X-ray

Cryo-EM

Tracing polypeptide backbone in a map

- Alternative to finding helices/strands
- Can be rapid
- Suited for lower-resolution maps where the backbone is clear but not side chains

***Tracing backbone step 1:
Points in high density***

(s-hydrolase, PDB entry 1A7A)

Move points to ridgelines

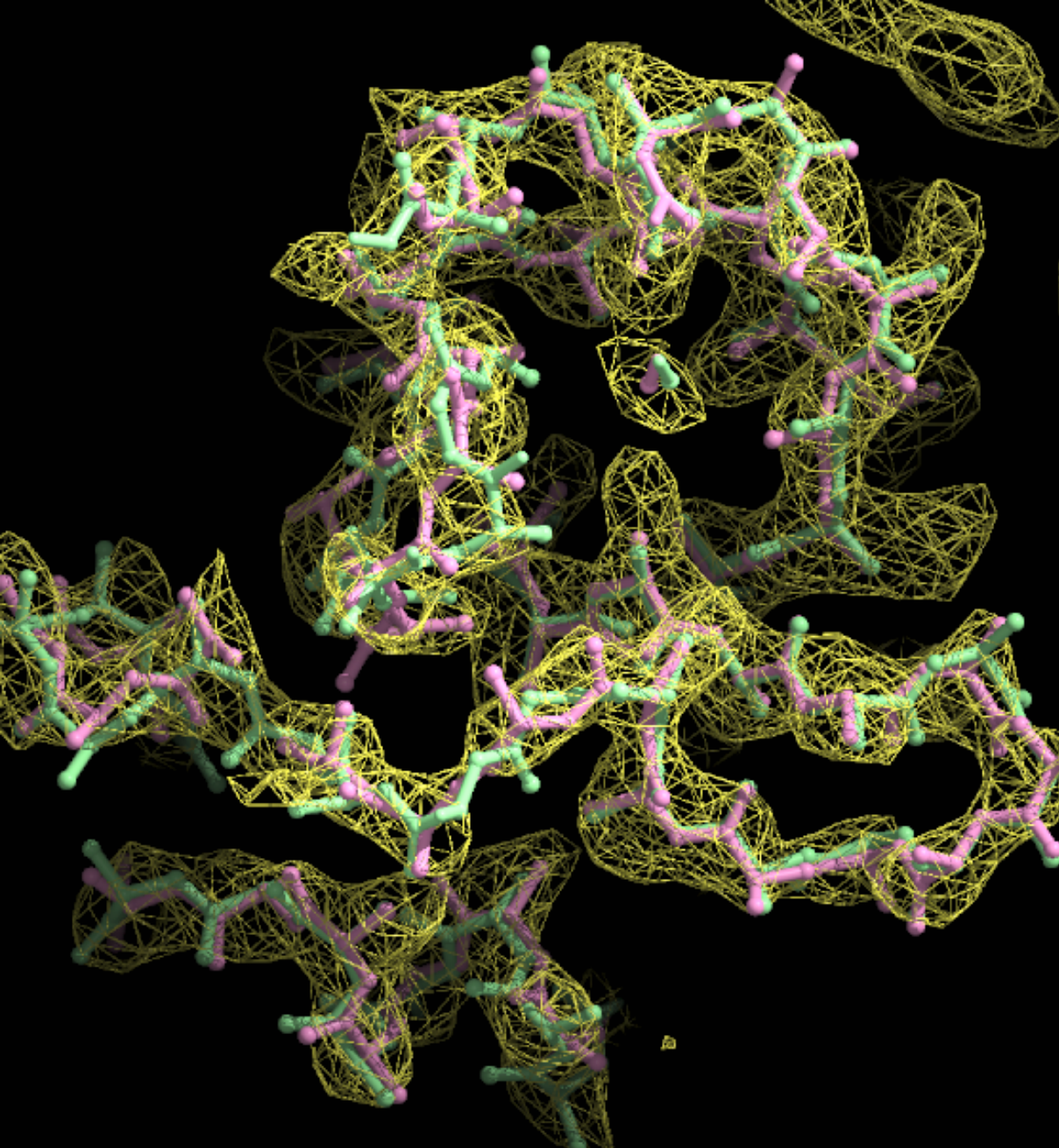
(s-hydrolase, PDB entry 1A7A)

Find possible C_{α} pairs
(3.8 Å apart, high density between, points near line)

Find possible C_α trimers
(Pairs sharing C_α ; 110° angle; points near line extending from vertex)

C_{α} tracing
(s-hydrolase, PDB entry 1A7A)

C_{α} tracing
(mevalonate kinase, PDB entry 1KKH, 9 sec)



Cryo-EM map from yeast
mitochondrial ribosome
(chain I of large subunit,
3.2 Å, Amunts et al.,
2014)

Autobuilt model (pink)
Deposited model (green)
(main-chain and C β atoms)

The Phenix Project

Lawrence Berkeley Laboratory

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



Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung



Randy Read, Airlie McCoy, Gabor Bunkoczi, Rob Oeffner

Cambridge University



Duke University

Jane & David Richardson, Chris Williams, Bryan Arendall, Bradley Hintze



*An NIH/NIGMS funded
Program Project*

