

Automation of Structure Solution with PHENIX

*Cold Spring Harbor Course
X-ray methods in structural biology
Oct. 21, 2010*

Tom Terwilliger
Los Alamos National Laboratory



Why automate structure determination?

Automation...

makes straightforward cases accessible to a wider group of structural biologists

makes difficult cases more feasible for experts

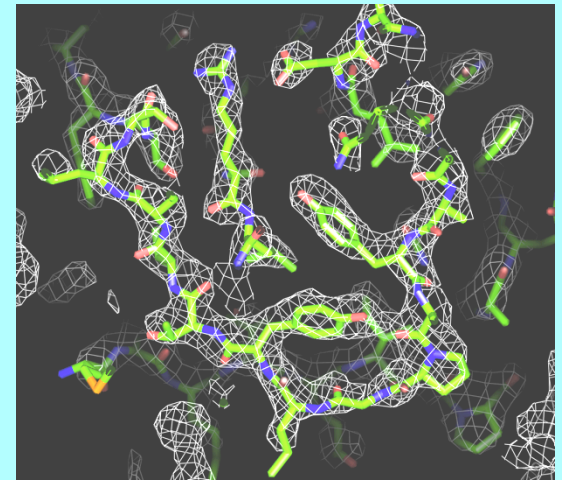
can speed up the process

can help reduce errors

Automation also allows you to...

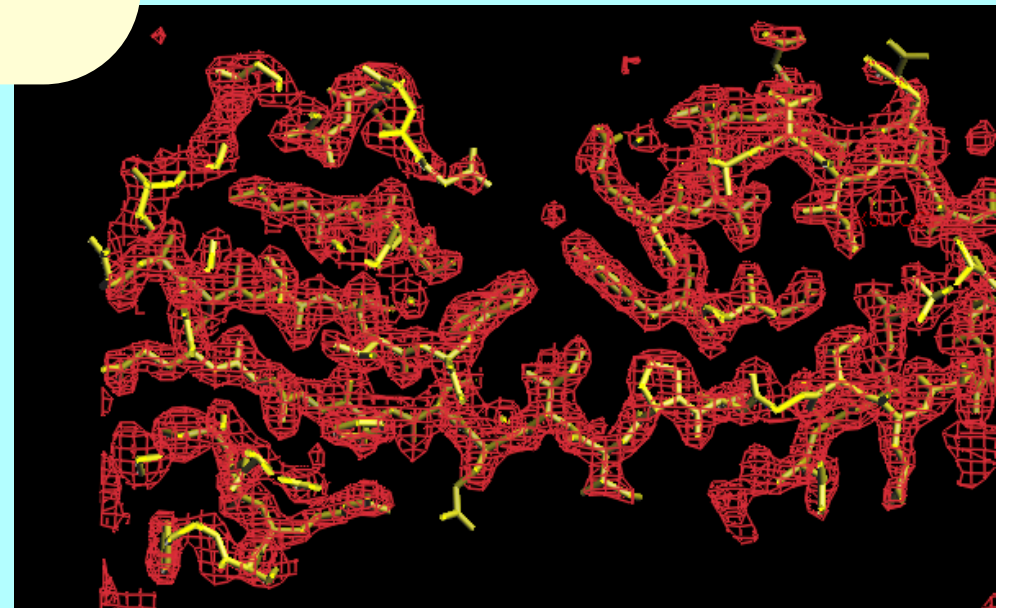
try more possibilities

estimate uncertainties



Requirements for automation of structure determination of macromolecules by X-ray crystallography

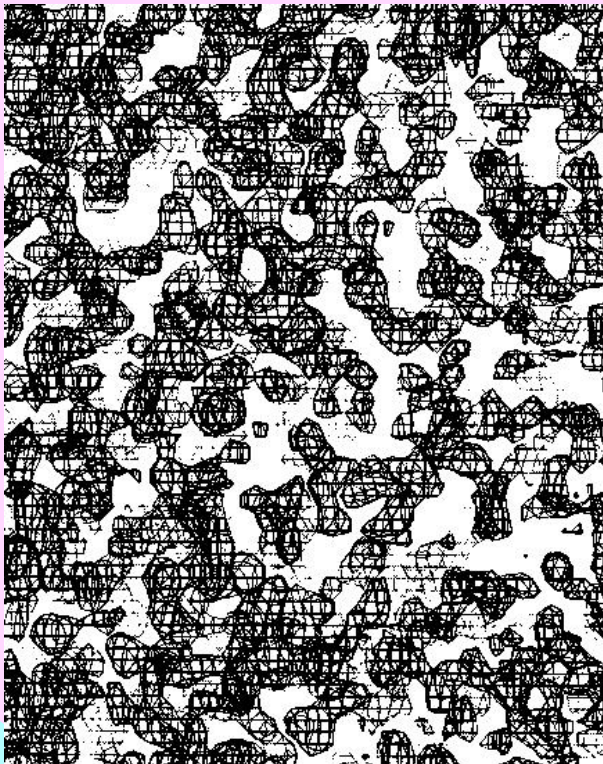
- (1) **Software carrying out individual steps**
- (2) **Seamless connections between steps**
- (3) **A way to decide what is good**
- (4) **Strategies for structure determination and decision-making**



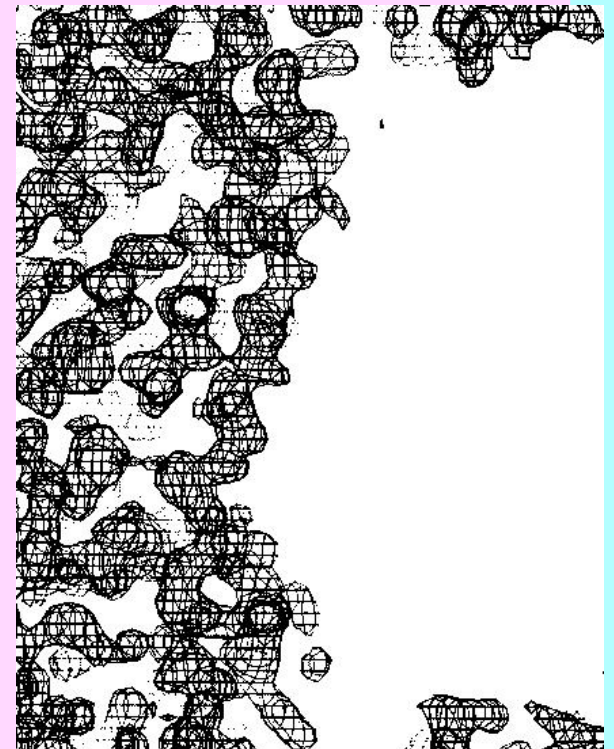
Why we need good measures of the quality of an electron-density map:

Which solution is best?

Are we on the right track?



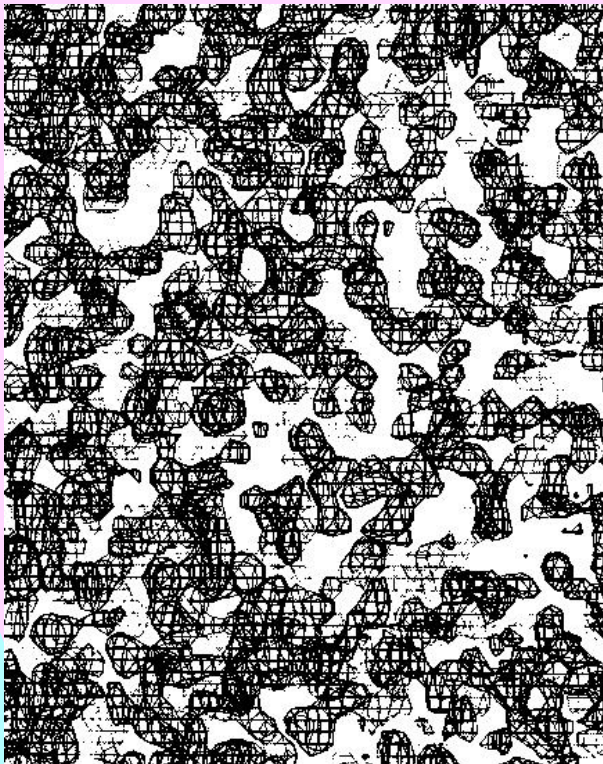
**If map is good:
It is easy**



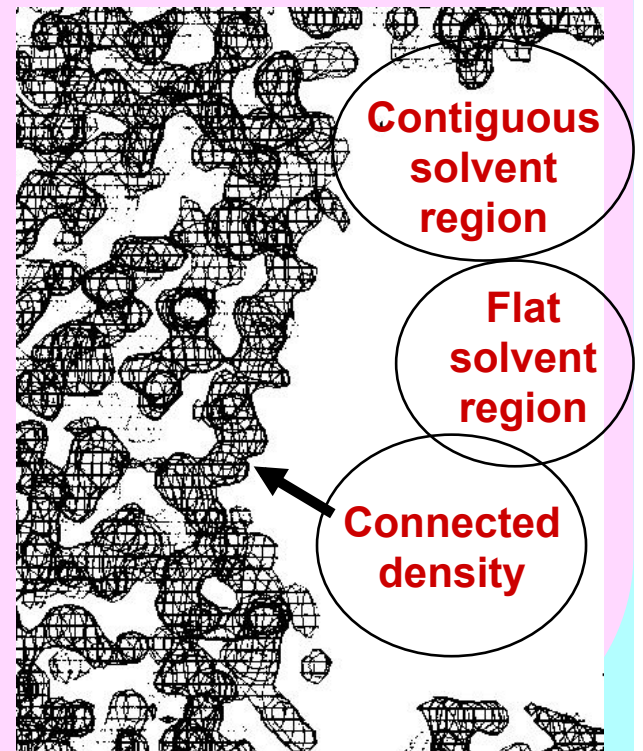
Why we need good measures of the quality of an electron-density map:

Which solution is best?

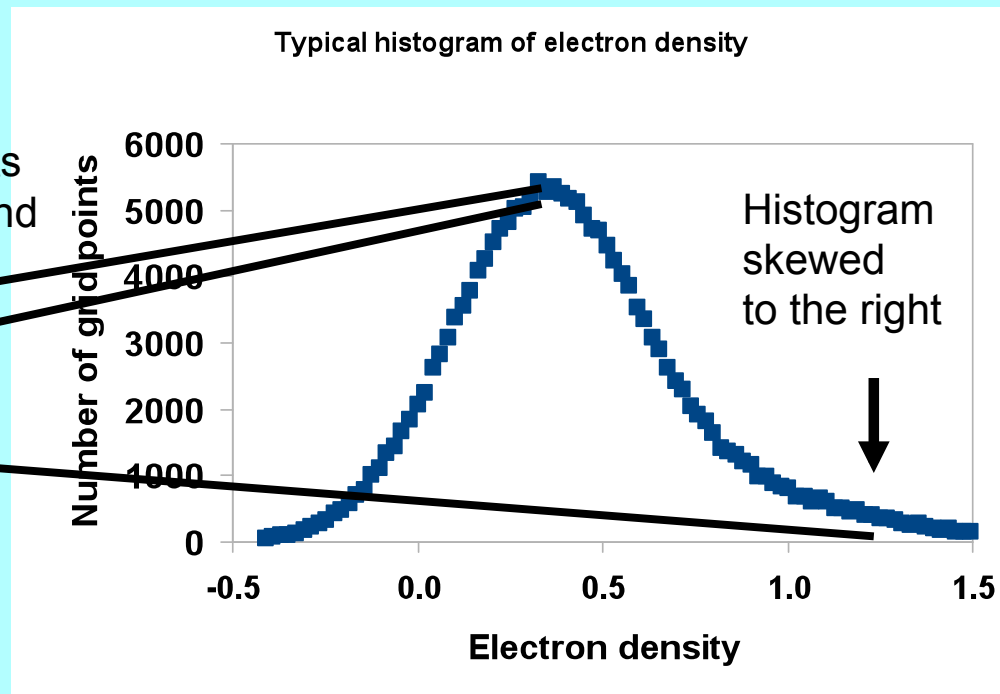
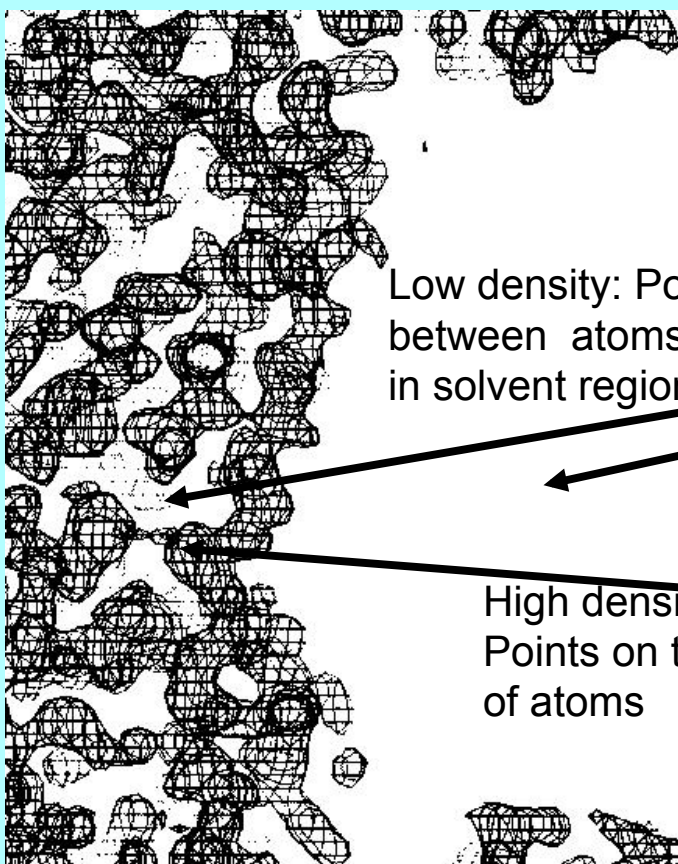
Are we on the right track?



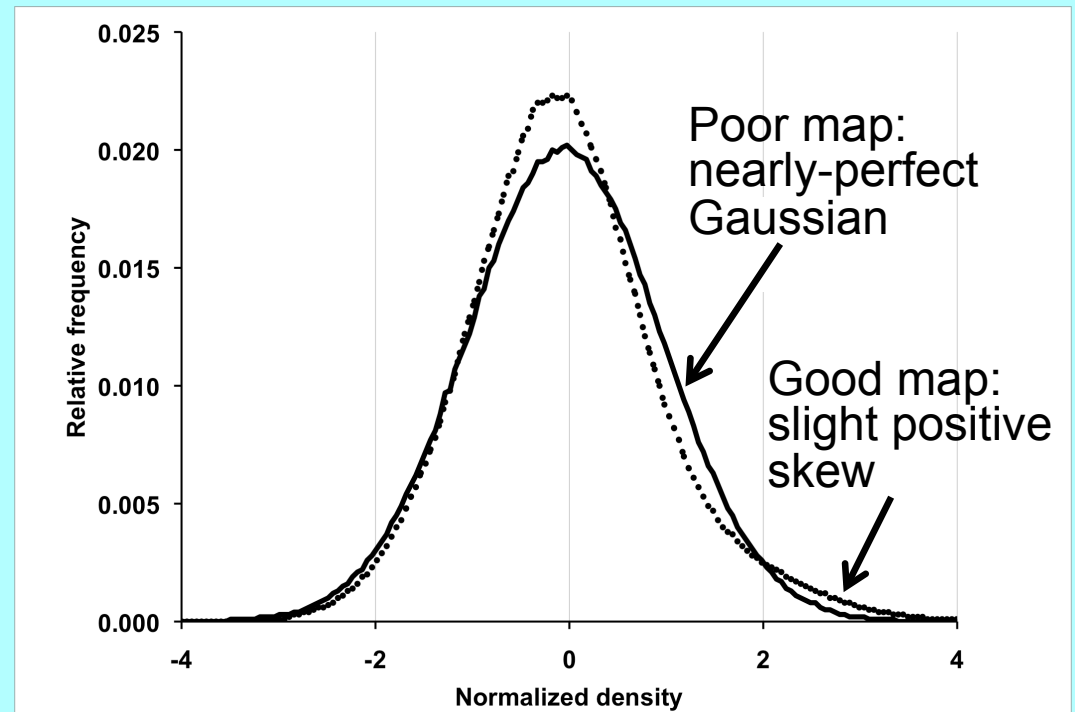
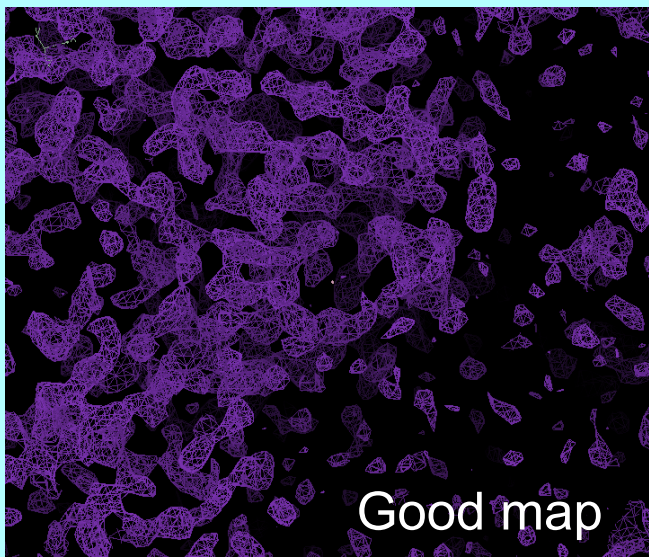
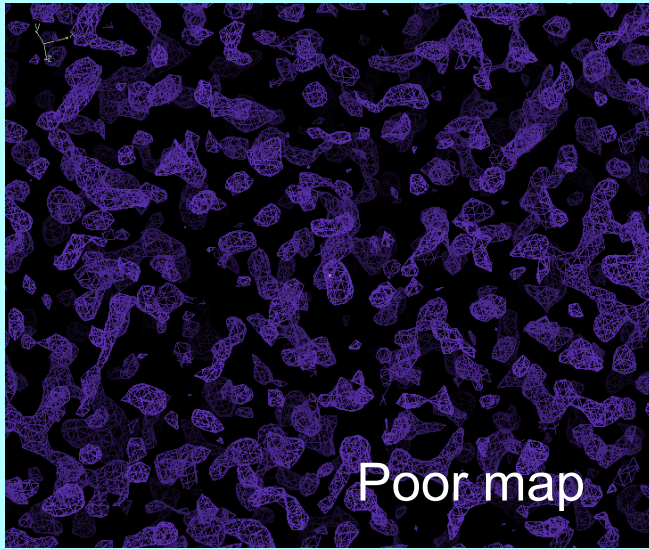
If map is good:
It is easy



Histogram of electron density values has a positive “skew”



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

Which scoring criteria best reflect the quality of a map?

Create real maps

Score the maps with each criteria

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)

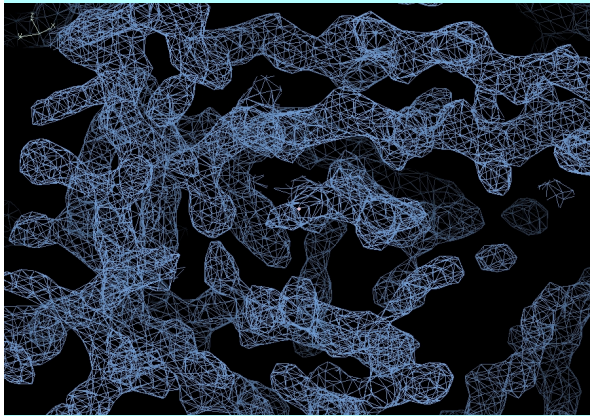
Run AutoSol Wizard on each dataset.

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

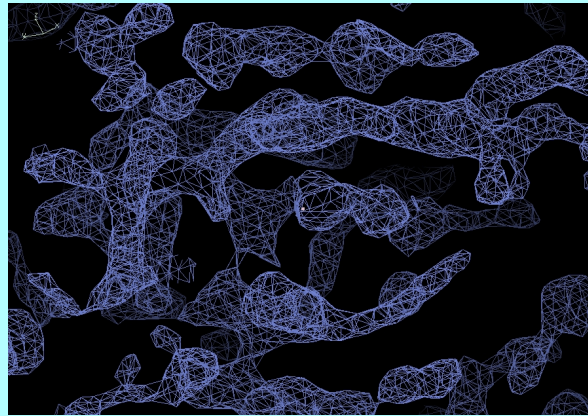
Score maps based on each criteria

Calculate map correlation coefficient (CC) to model map
(no density modification, shift origin if necessary)

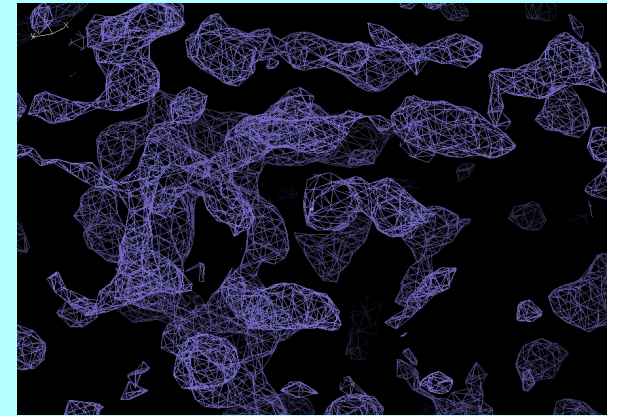
Model map
1VQB, 2.6 Å, SG C2



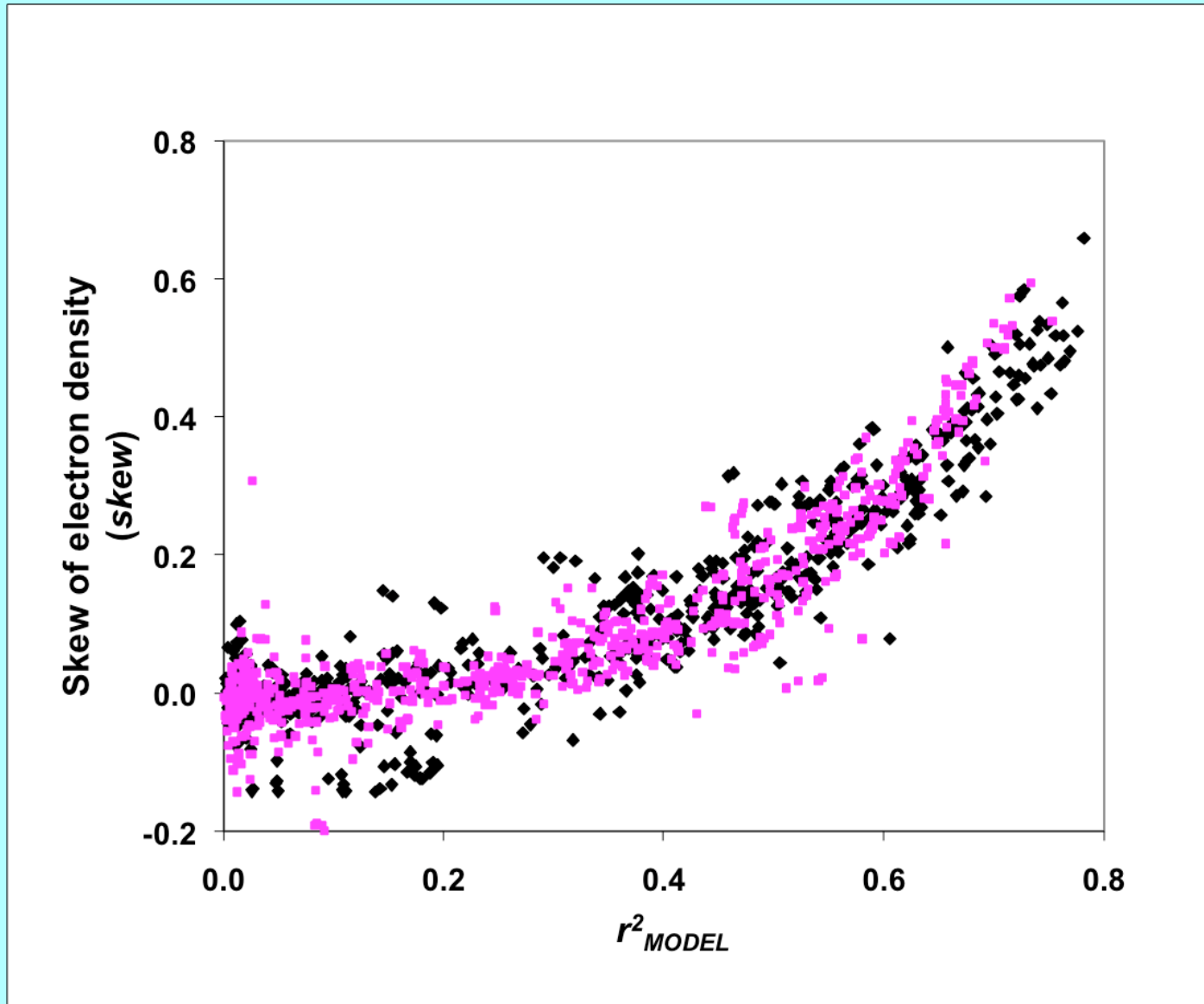
SOLVE MAD map
CC=0.62



Inverse-hand map
CC=0.55



Skew of electron density – positive skew of density values

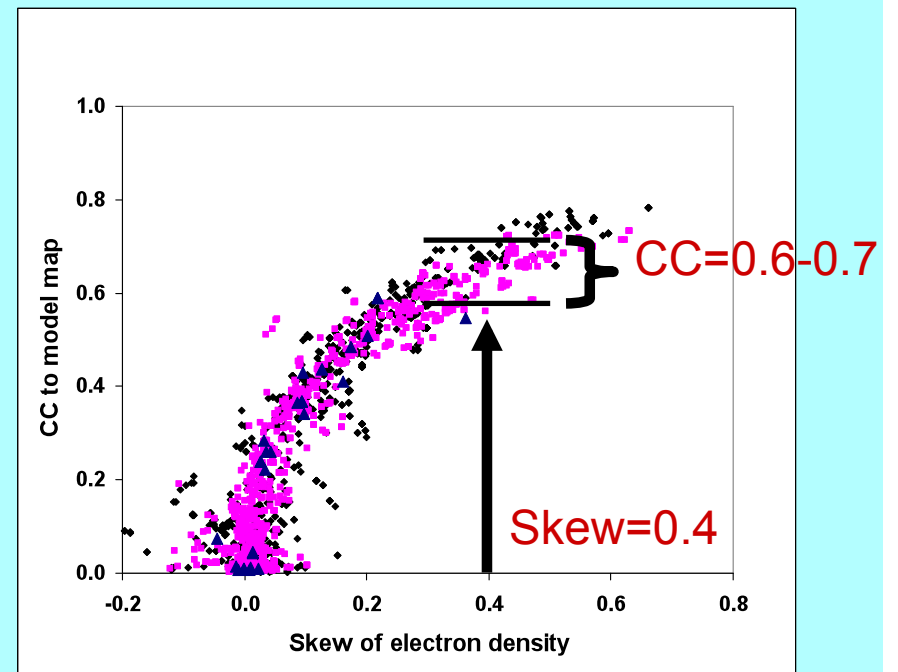
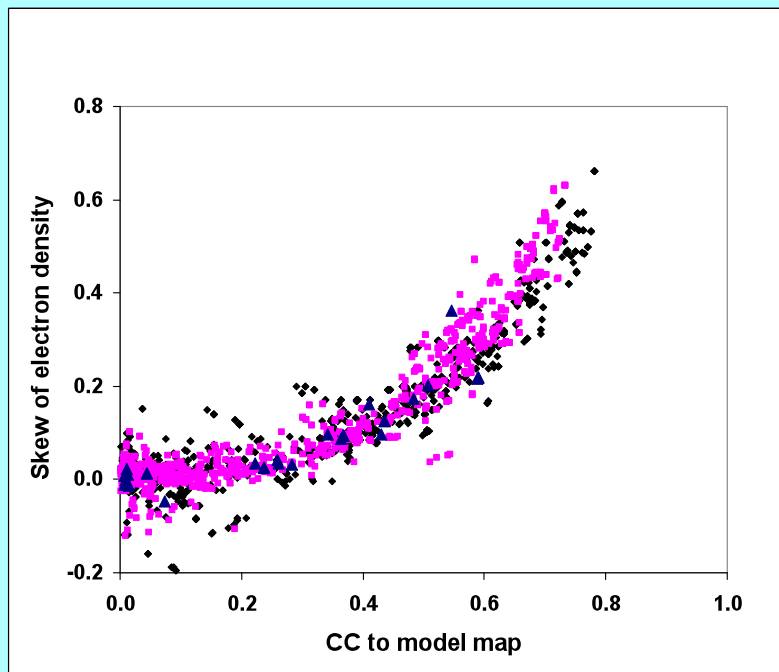


Using scoring criteria to estimate the quality of a map

Skew depends on CC

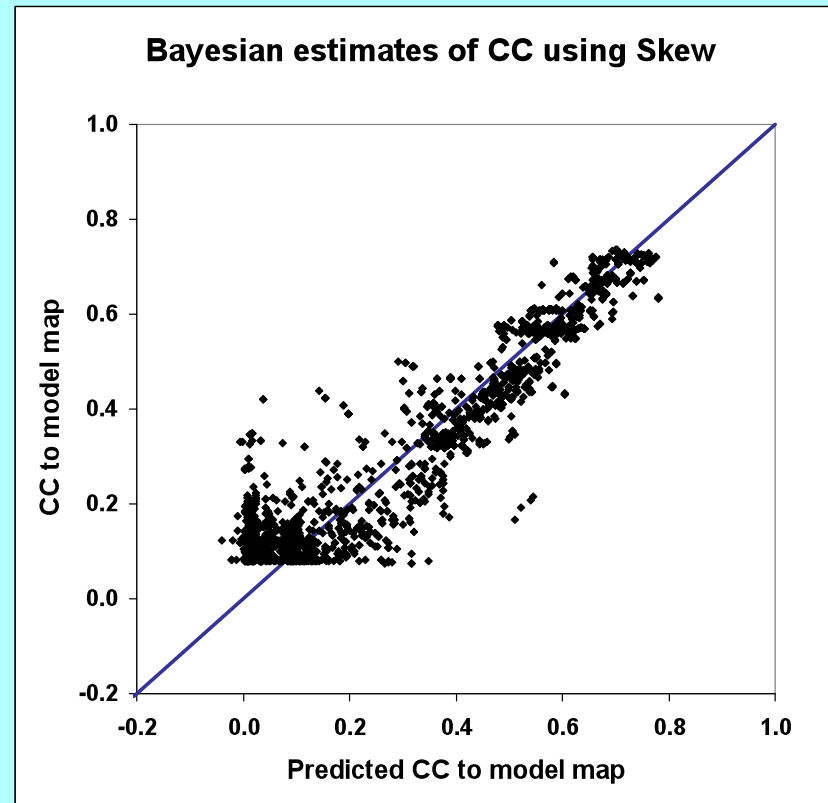


Estimate CC from skew



How accurate are estimates of map quality?

Actual
quality



Estimated quality

Cross-validated estimates of quality

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 $\pm 2SD$	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

Statistical density modification (RESOLVE)

•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”

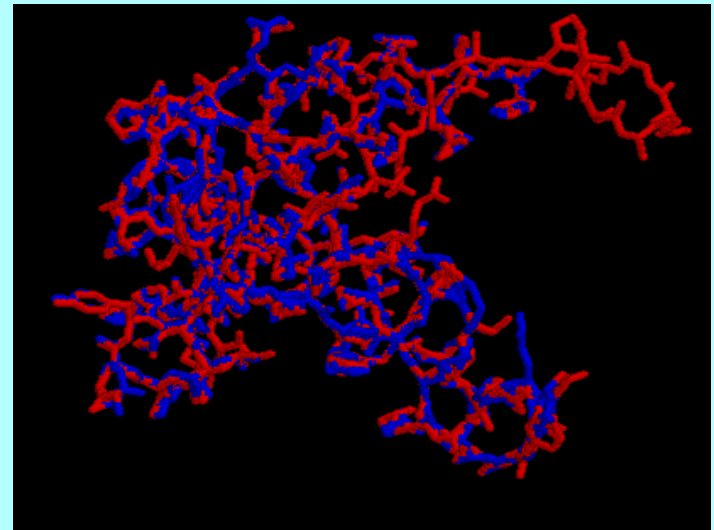
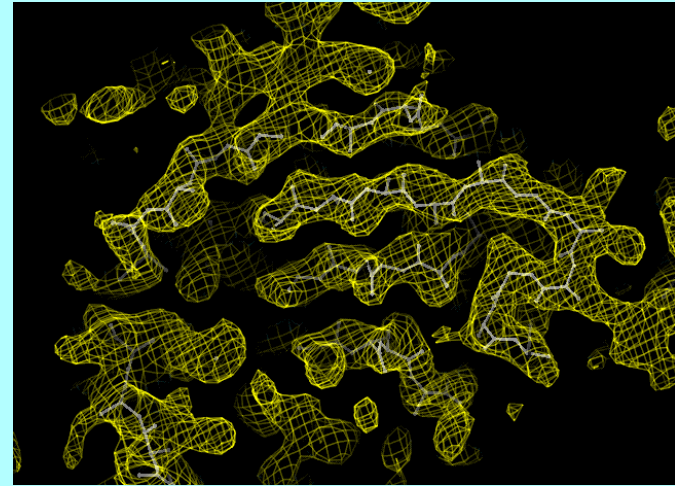
•A believable map is a map that has...

- a relatively flat solvent region
- NCS (if appropriate)

•A distribution of densities like those of model proteins

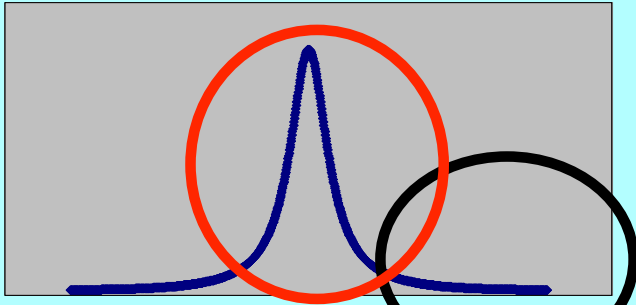
•Method:

- calculate how map probability varies with electron density ρ
- deduce how map probability varies with phase ϕ
 - combine with experimental phase information

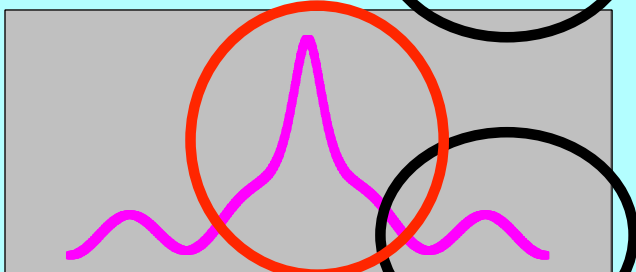


Map probability phasing: Getting a new probability distribution for each phase given estimates of all others

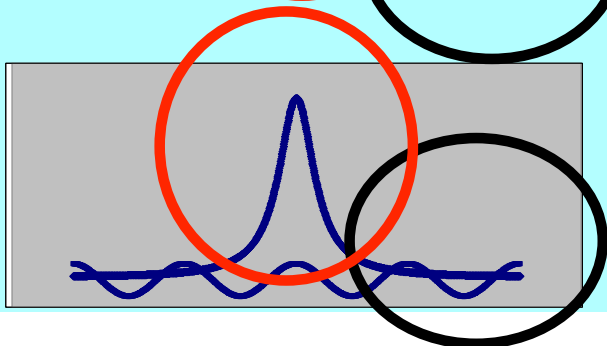
1. Identify expected features of map (flat far from center)
2. Calculate map with current estimates of all structure factors except one (k)
3. Test all possible phases ϕ for structure factor k (for each phase, calculate new map including k)
4. Probability of phase ϕ estimated from agreement of map with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



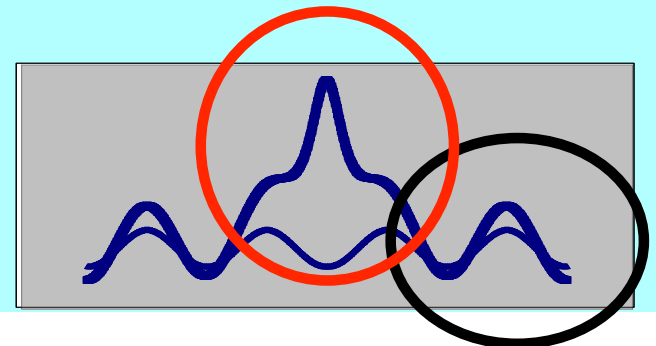
A function that is (relatively) flat far from the origin



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



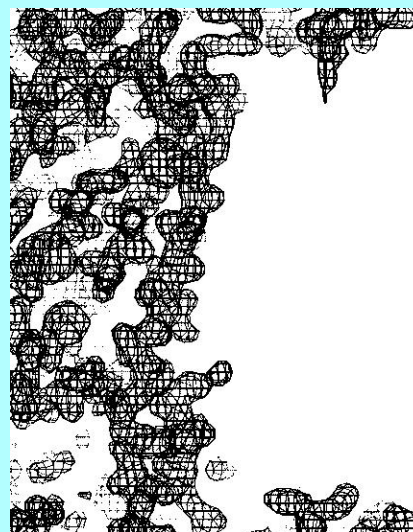
Test each possible phase of structure factor k . $P(\phi)$ is high for phase that leads to flat region



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

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



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

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

Structure solution with *phenix.autosol*

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

Decisions to be made

Find heavy-atom sites with direct
methods (HYSS)

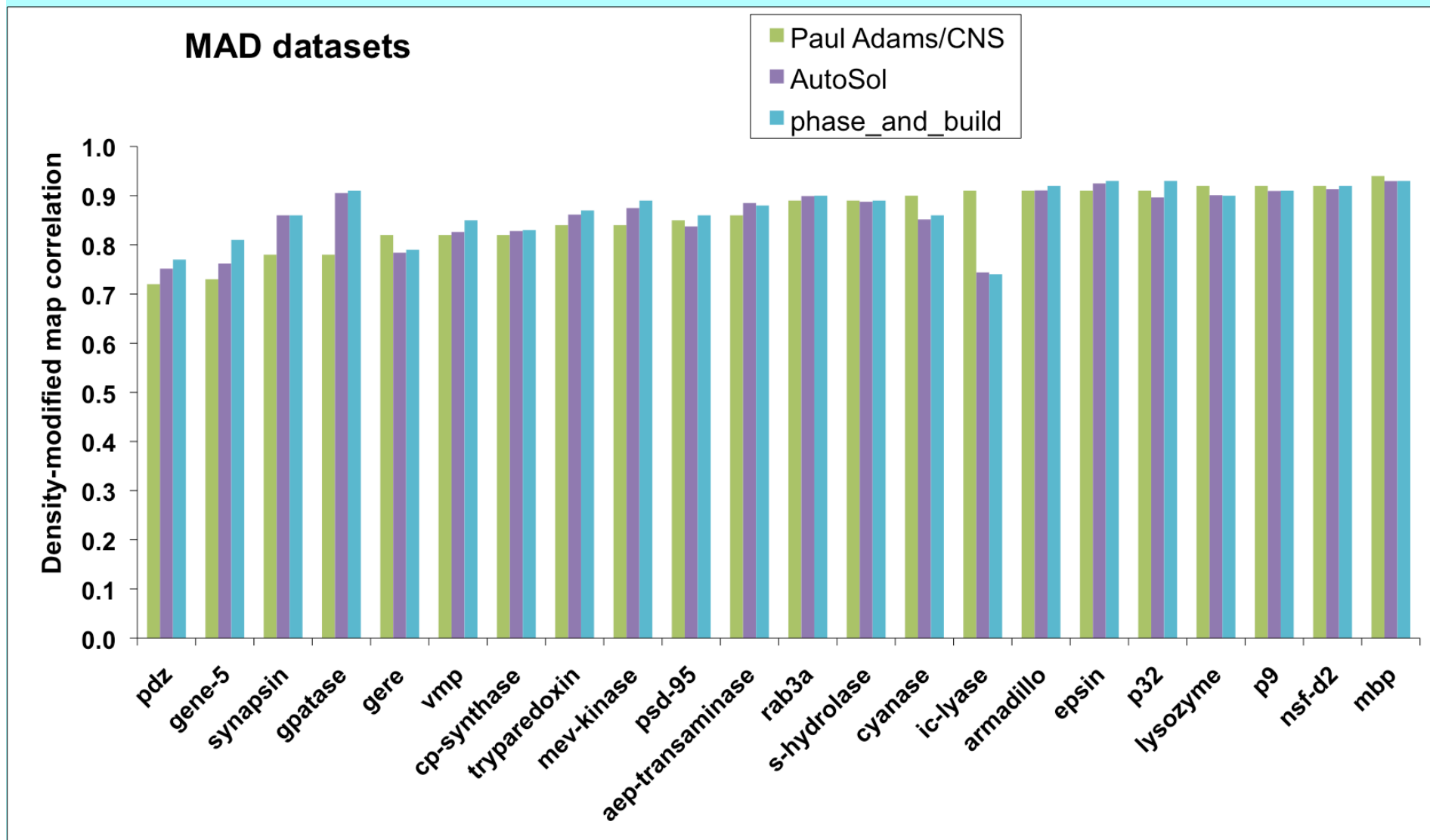
*Multiple solutions,
different derivatives or
wavelengths*

Calculate phases (Phaser/Solve)

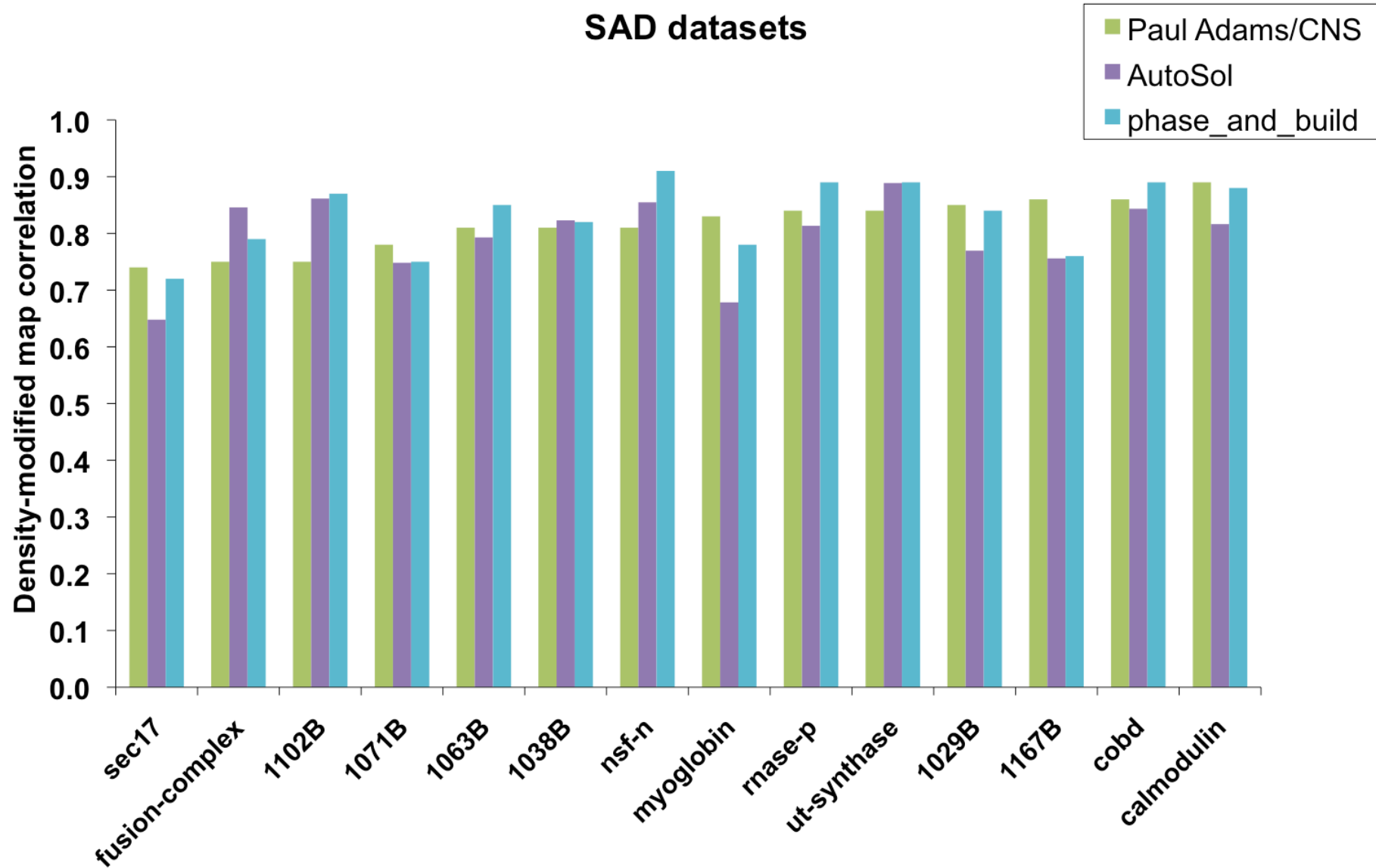
*Alternative hands of
space-group and
substructure*

Improve phases, find NCS, build
model (phase_and_build)

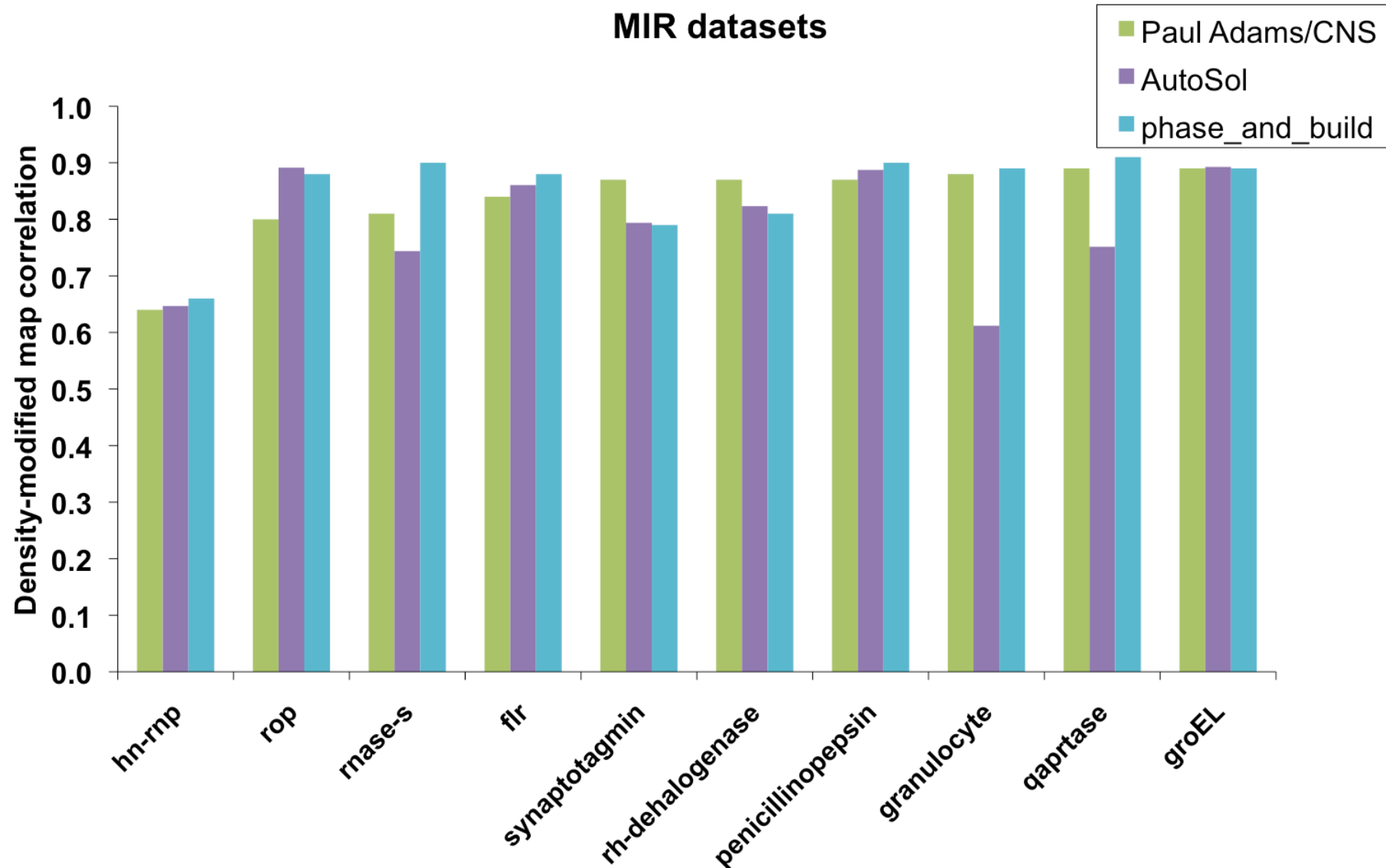
AutoSol – fully automatic tests with structure library
(MAD datasets, HYSS search, SOLVE)
RESOLVE/ phase_and_build maps



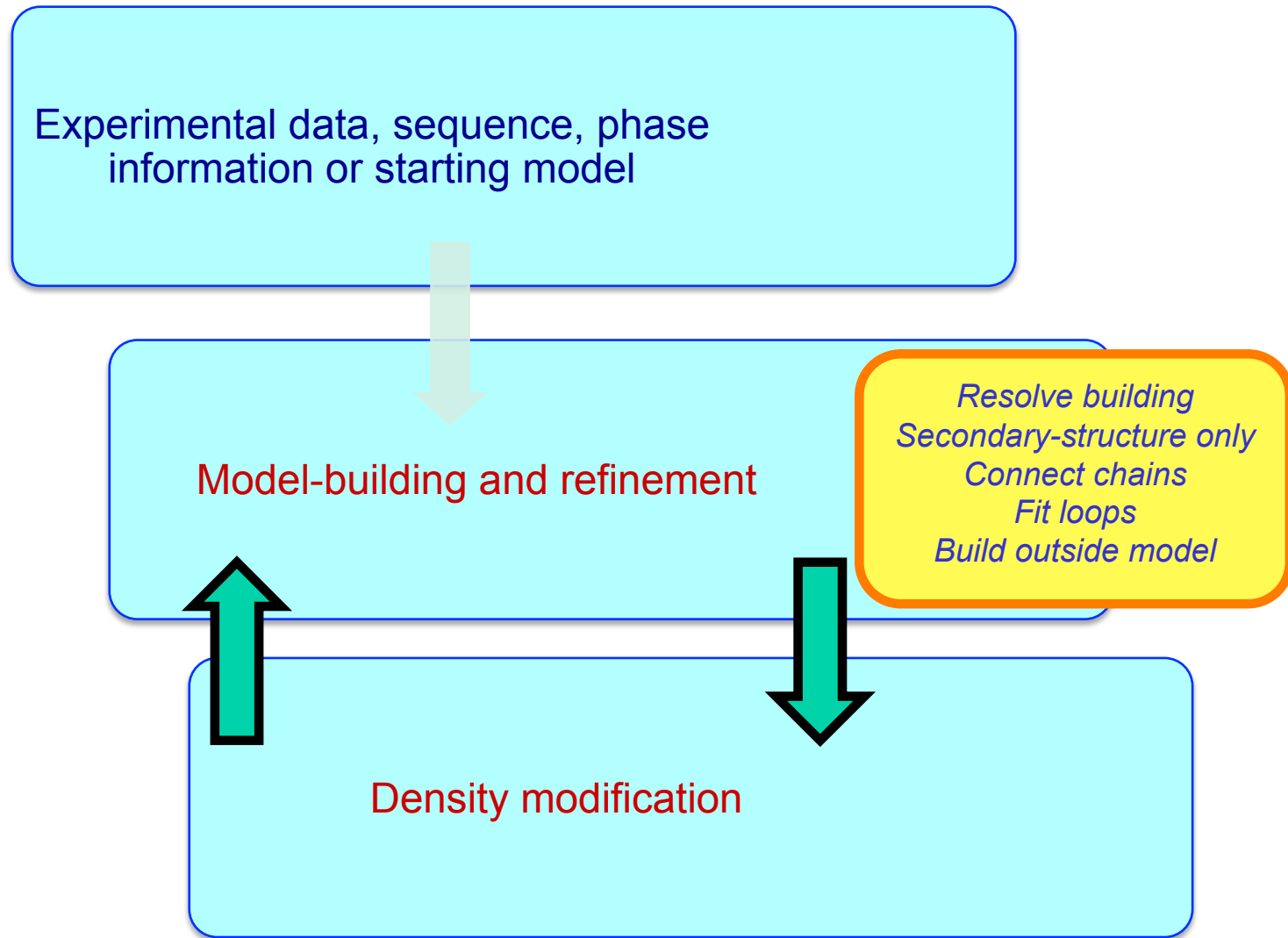
AutoSol – fully automatic tests with structure library
(SAD datasets, HYSS, Phaser)
RESOLVE/ phase_and_build maps



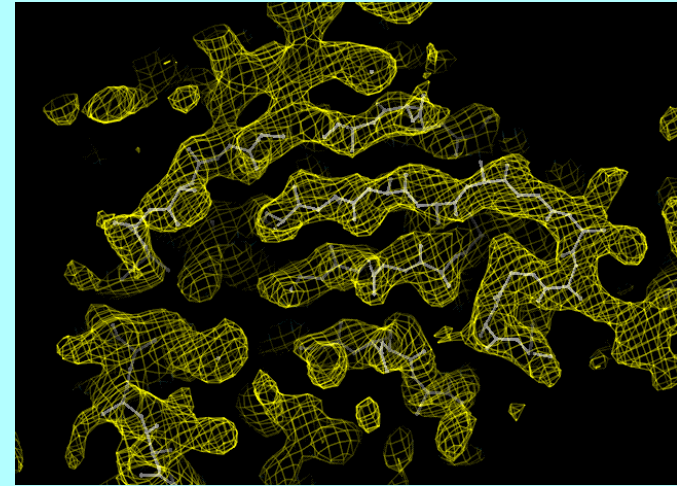
AutoSol – fully automatic tests with structure library
(SAD datasets, HYSS, Phaser)
RESOLVE/ phase_and_build maps



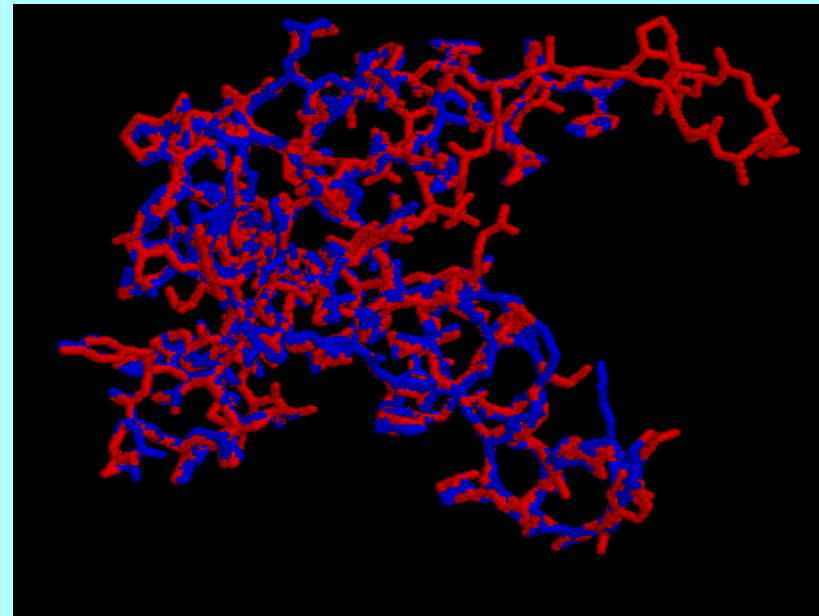
Iterative density modification, model-building and refinement with *phenix.autobuild*



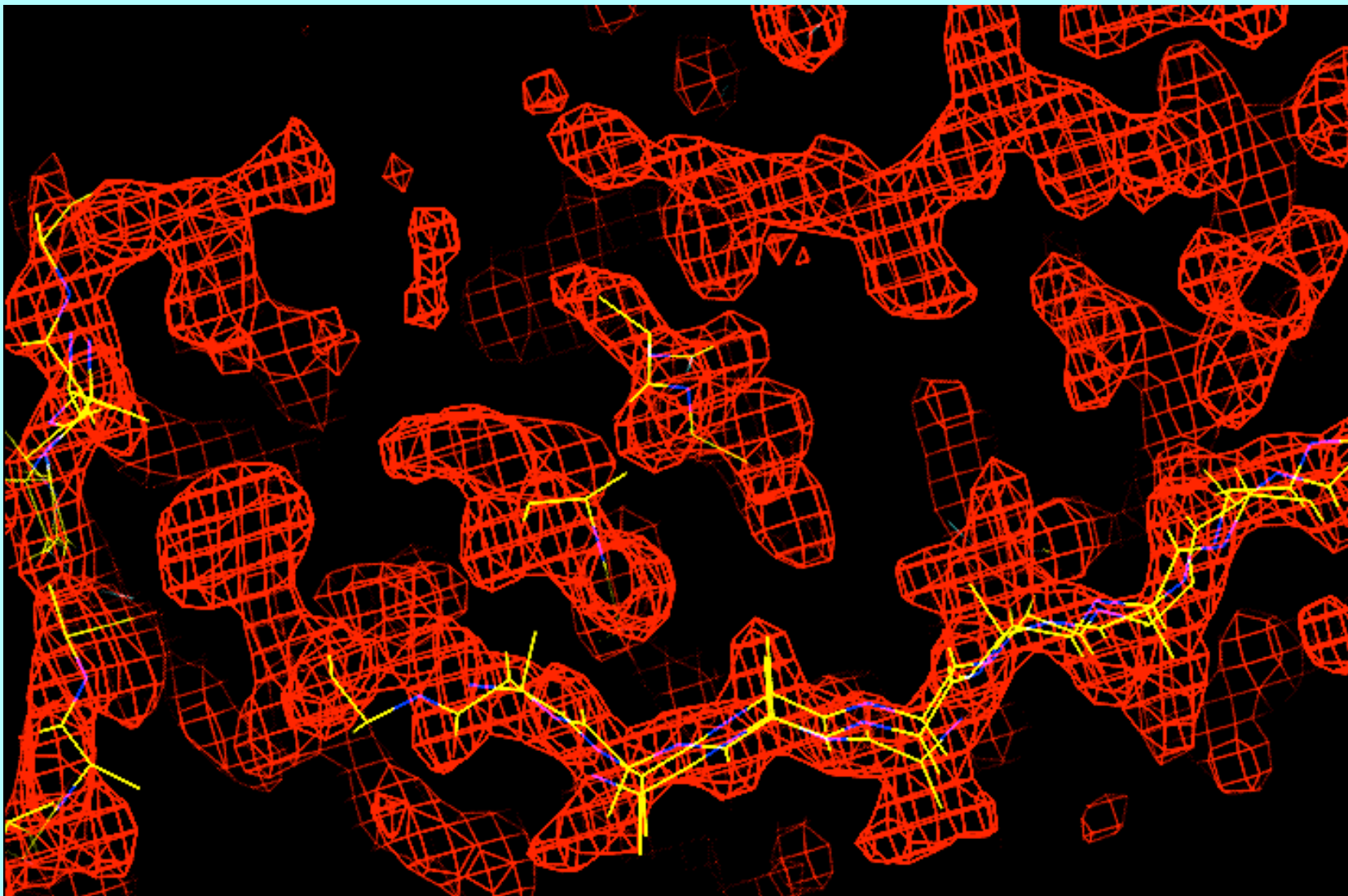
RESOLVE model-building at moderate resolution



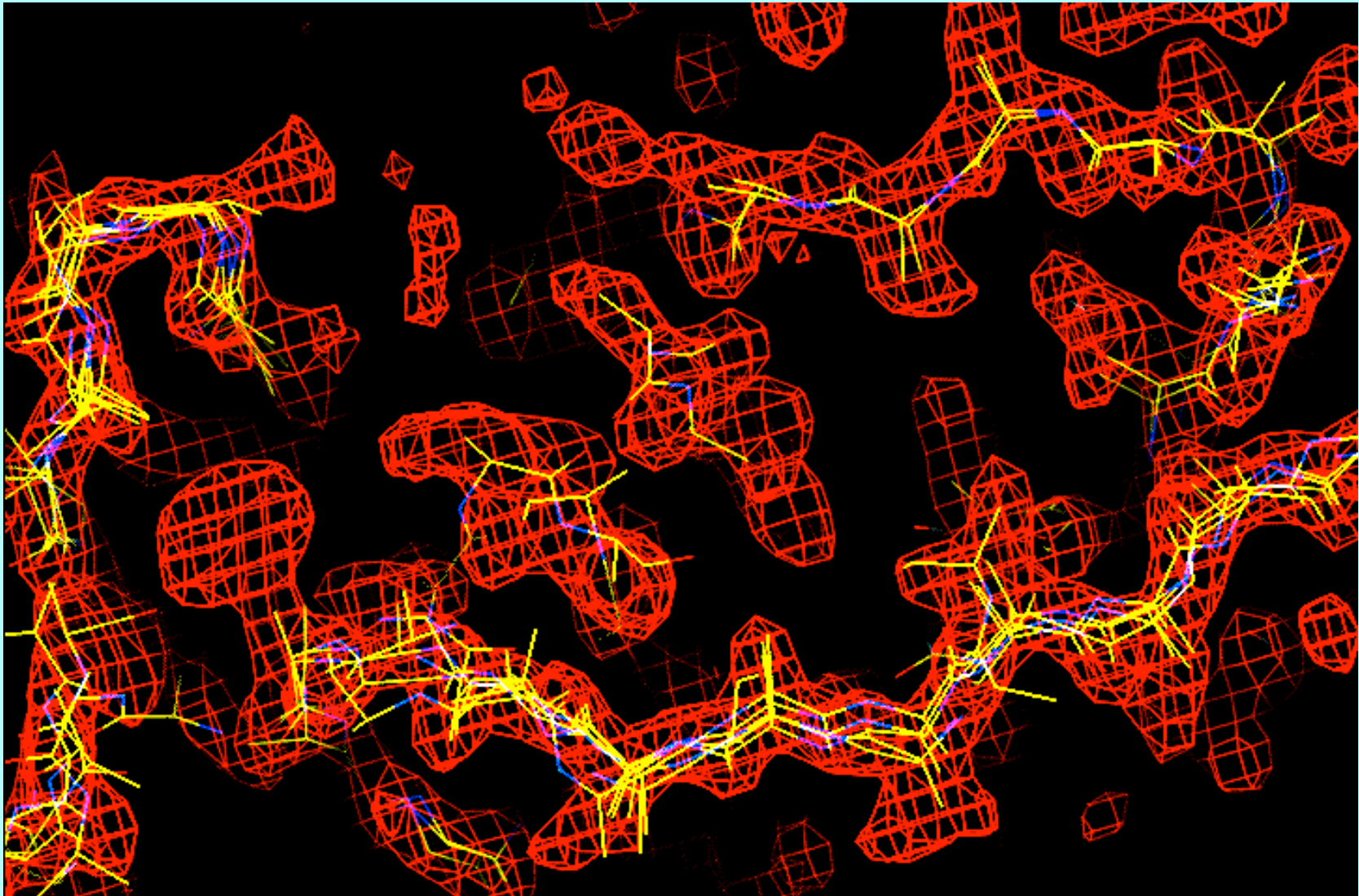
- *FFT-based identification of helices and strands*
- *Extension with tripeptide libraries*
- *Probabilistic sequence alignment*
- *Automatic molecular assembly*



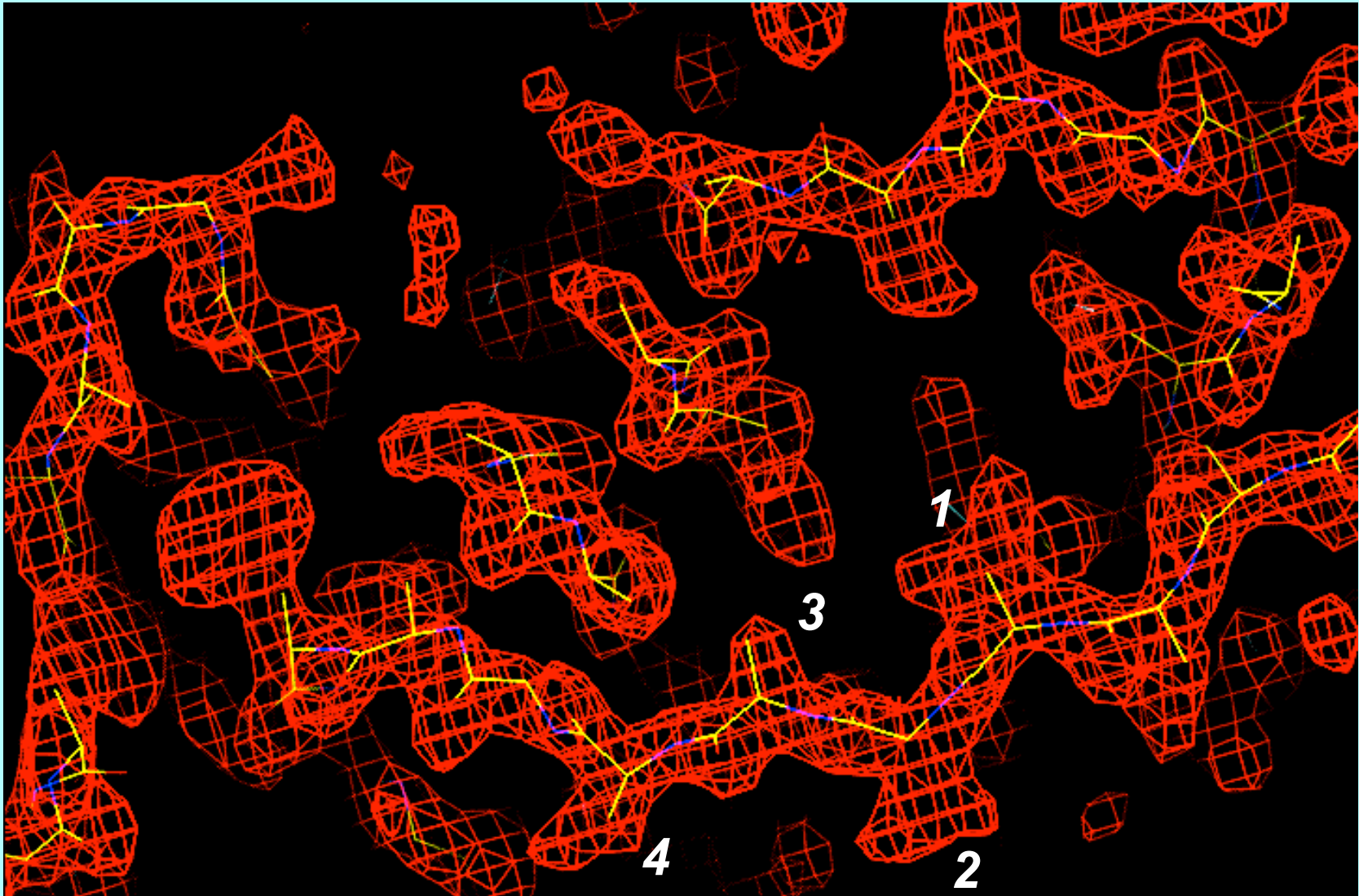
Initial model-building – strand fragments



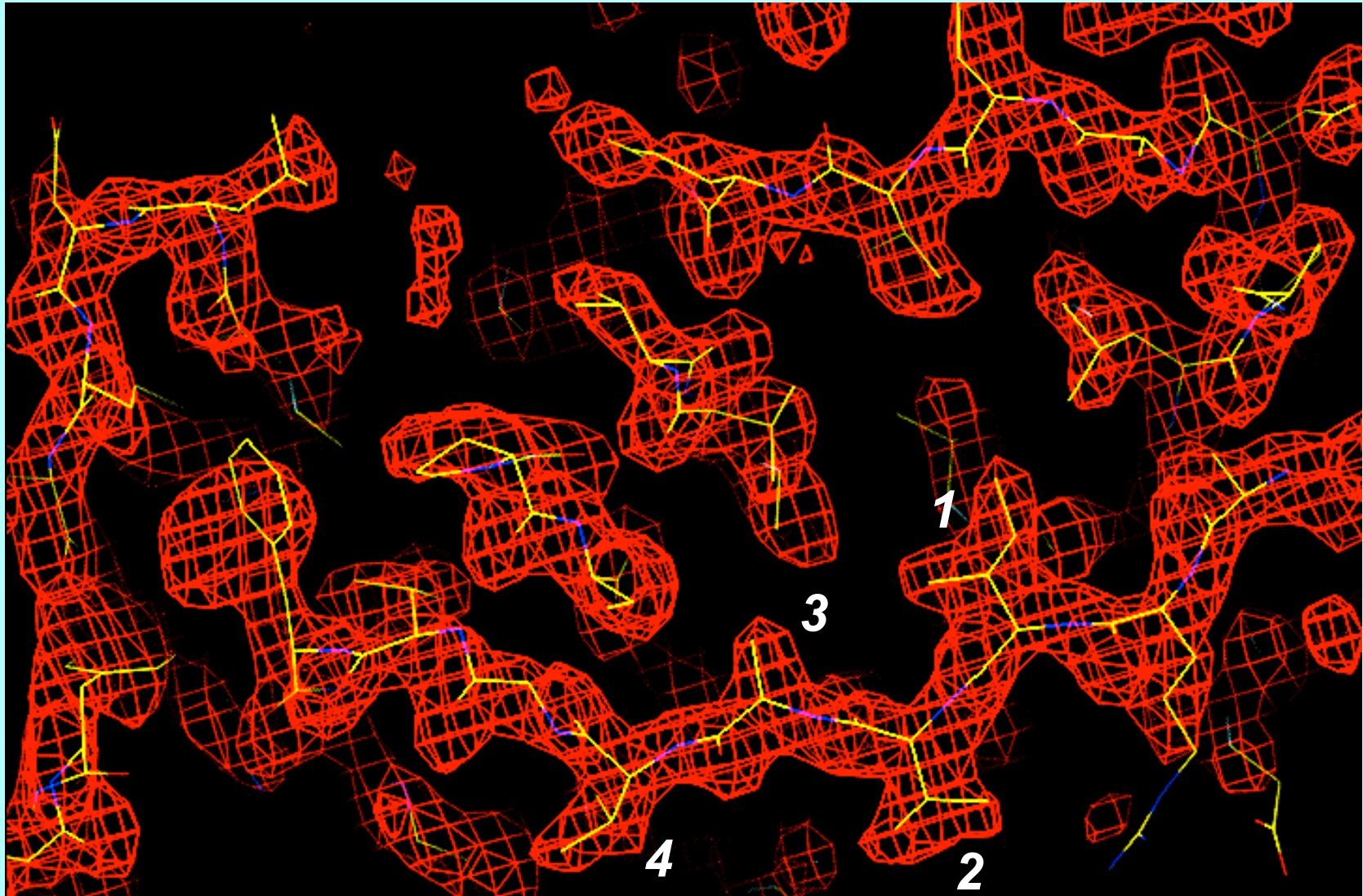
*Chain extension
(result: many overlapping fragments)*



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

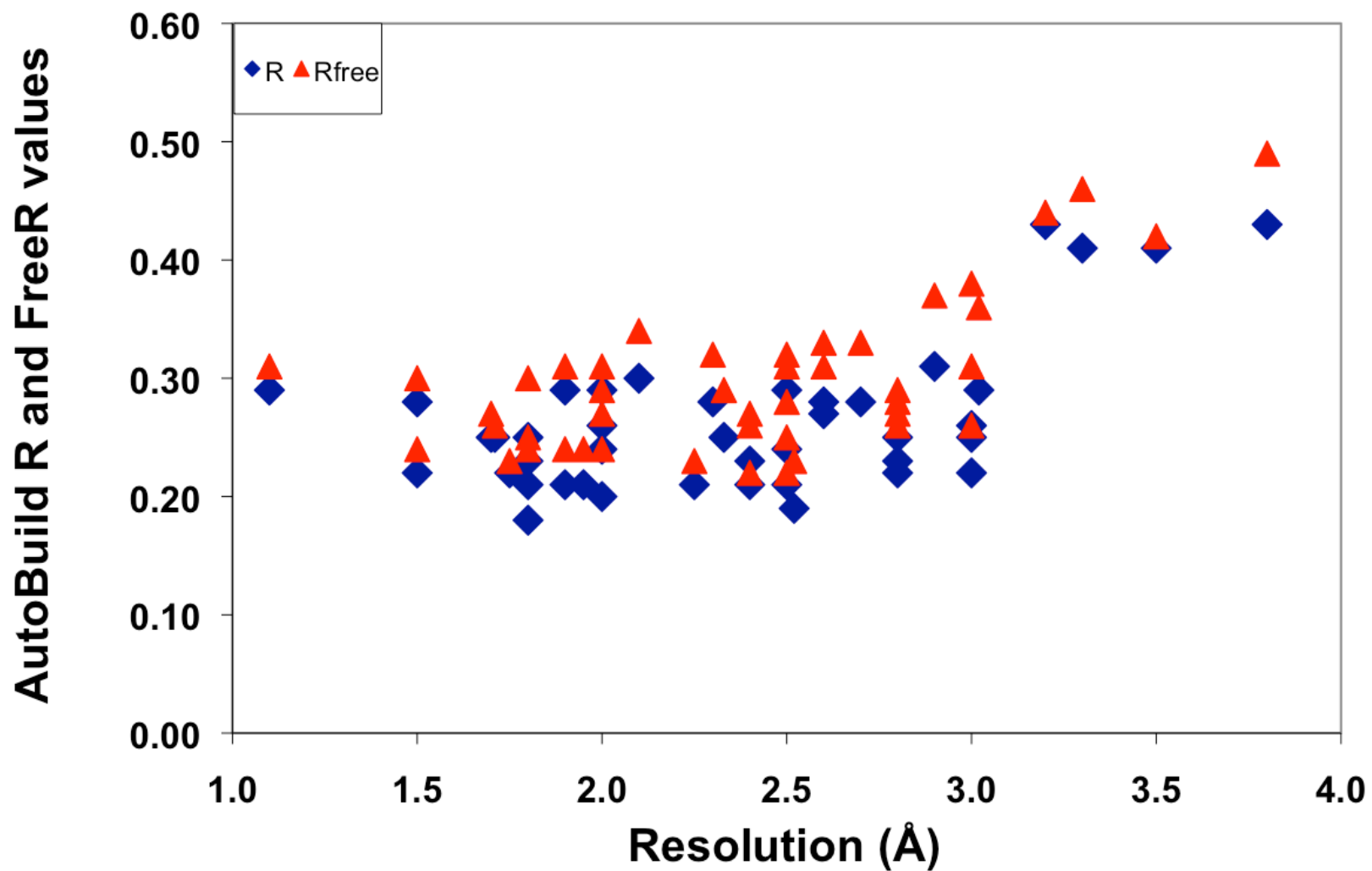


Addition of side-chains to fixed main-chain positions



AutoBuild – tests with structure library

Fully automated iterative model-building, final R/Rfree



Rapid building of models for regions containing regular secondary-structure

Helices:

Identification: rods of density at low resolution

Strands:

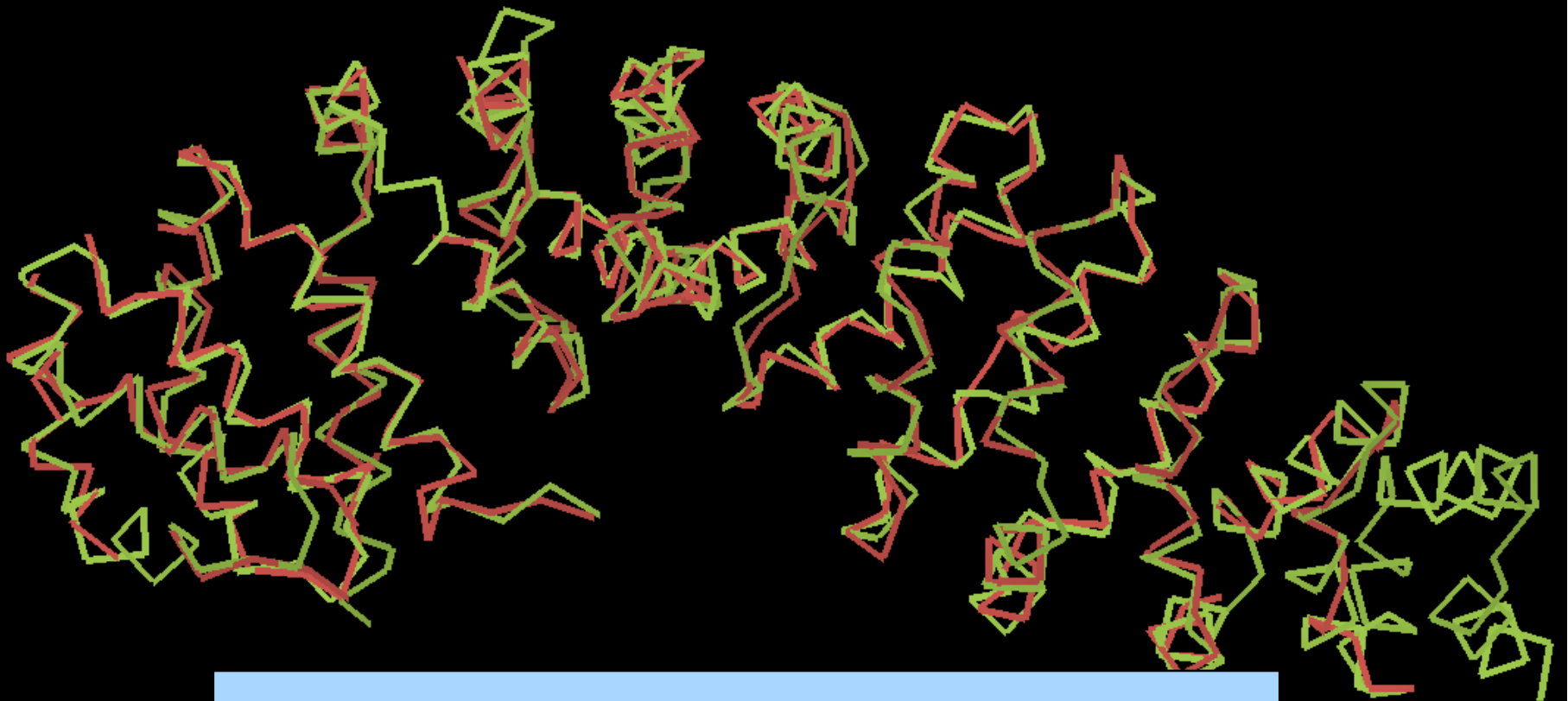
Identification: β structure as nearly-parallel pairs of tubes

Any protein chains (trace_chain):

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

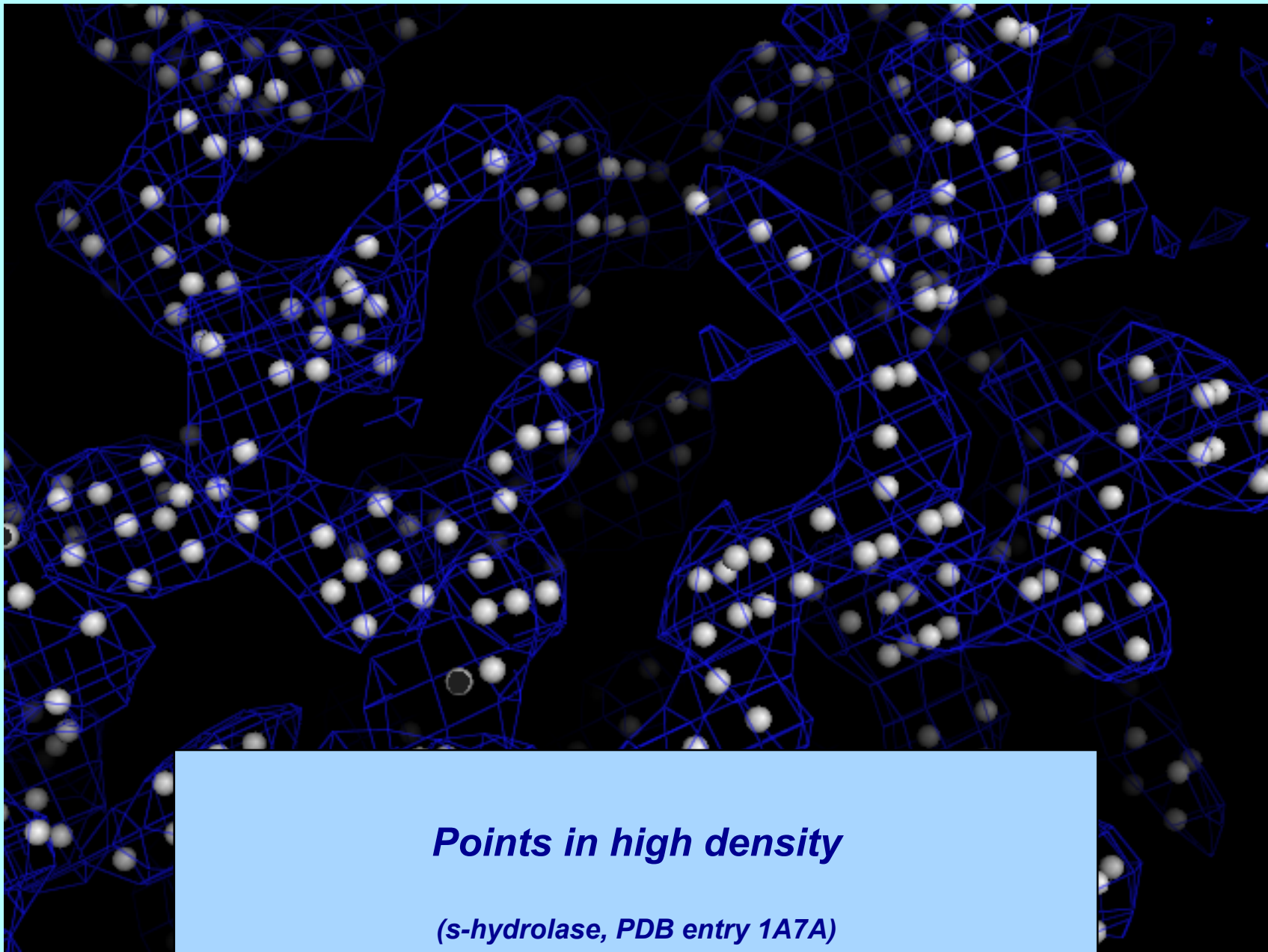
RNA/DNA:

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



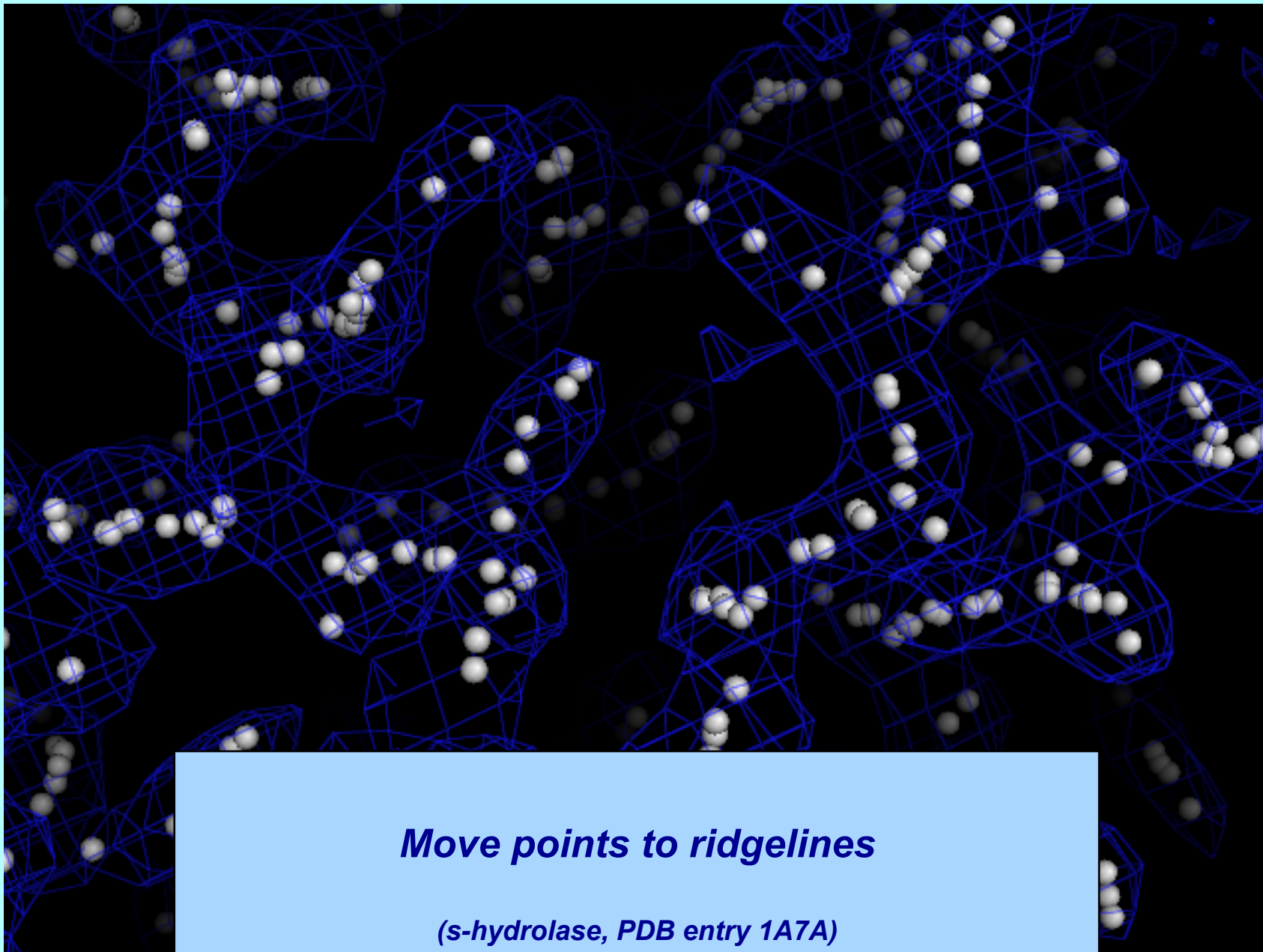
***Rapid chain-tracing
for evaluation of map quality***

(armadillo repeat of β -catenin, 369 residues, 23 sec)



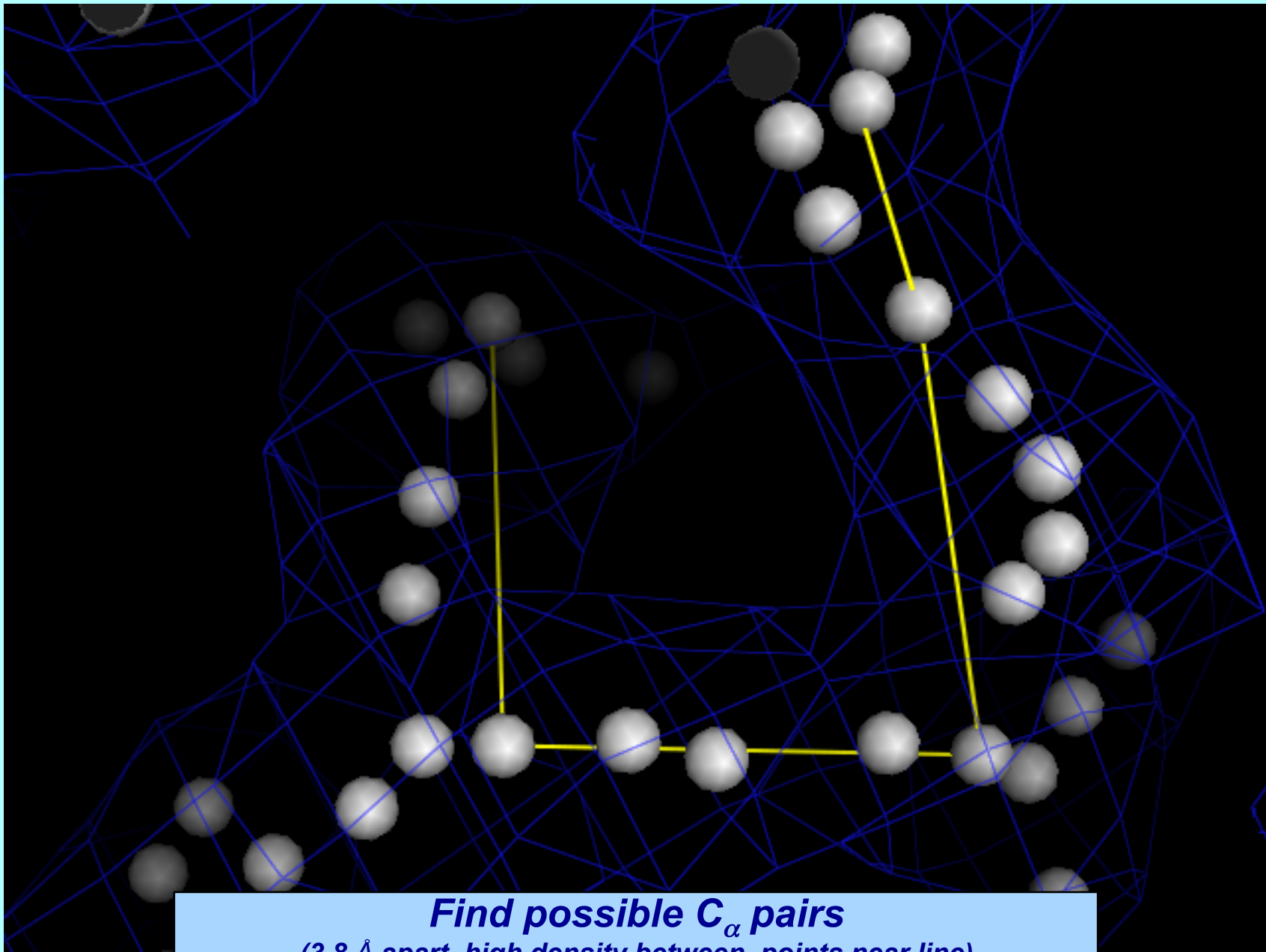
Points in high density

(s-hydrolase, PDB entry 1A7A)

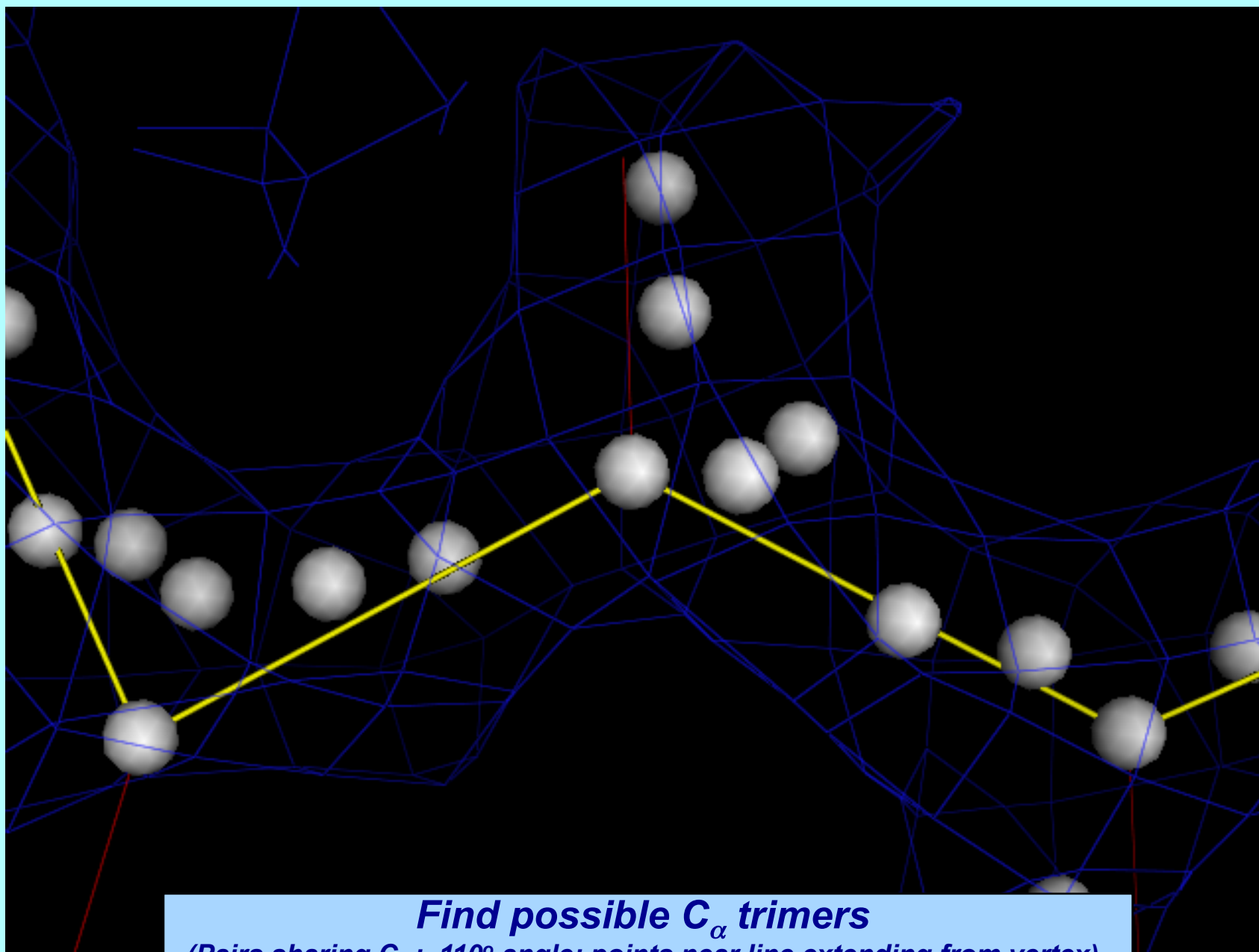


Move points to ridgelines

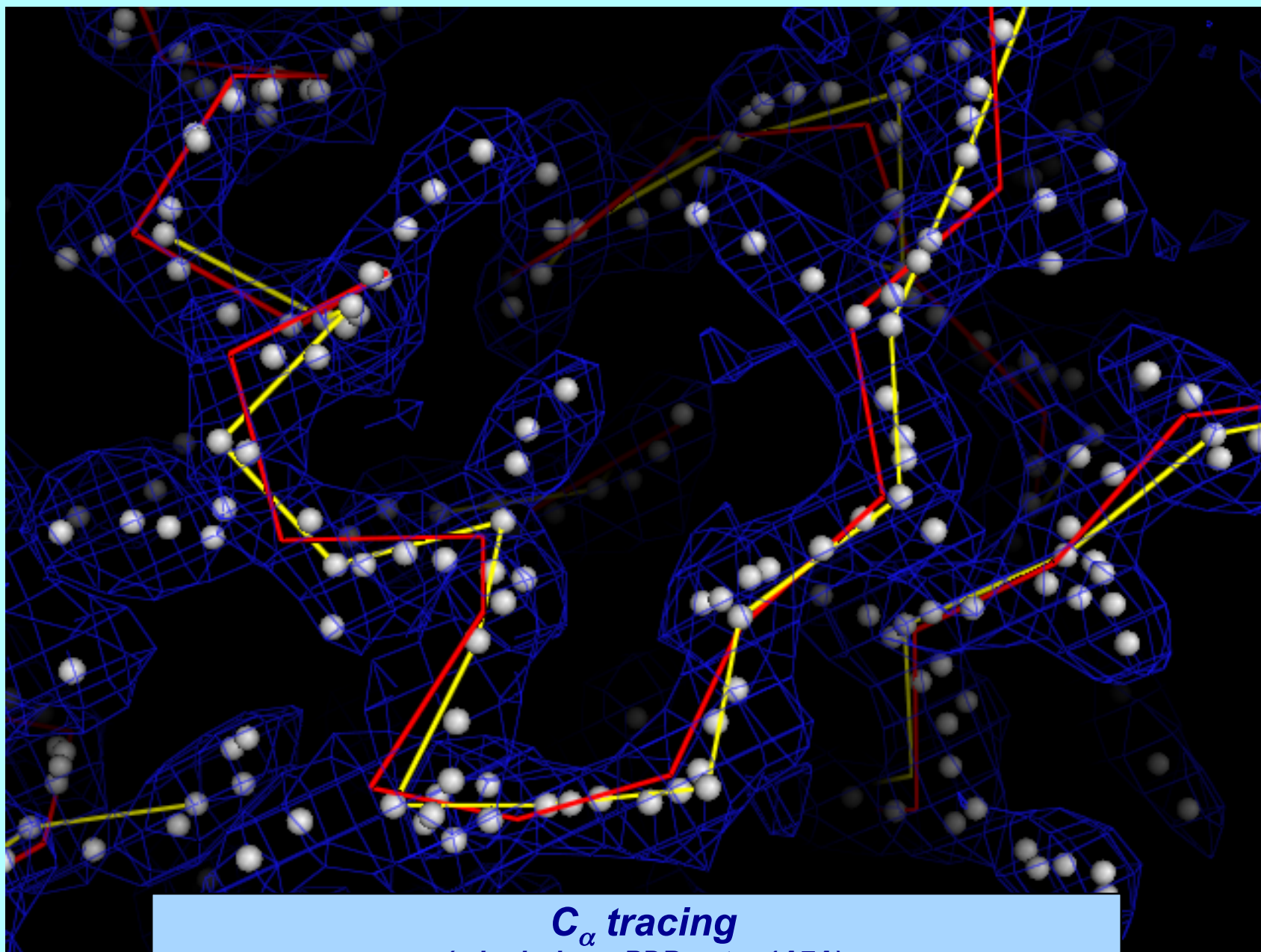
(s-hydrolase, PDB entry 1A7A)



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



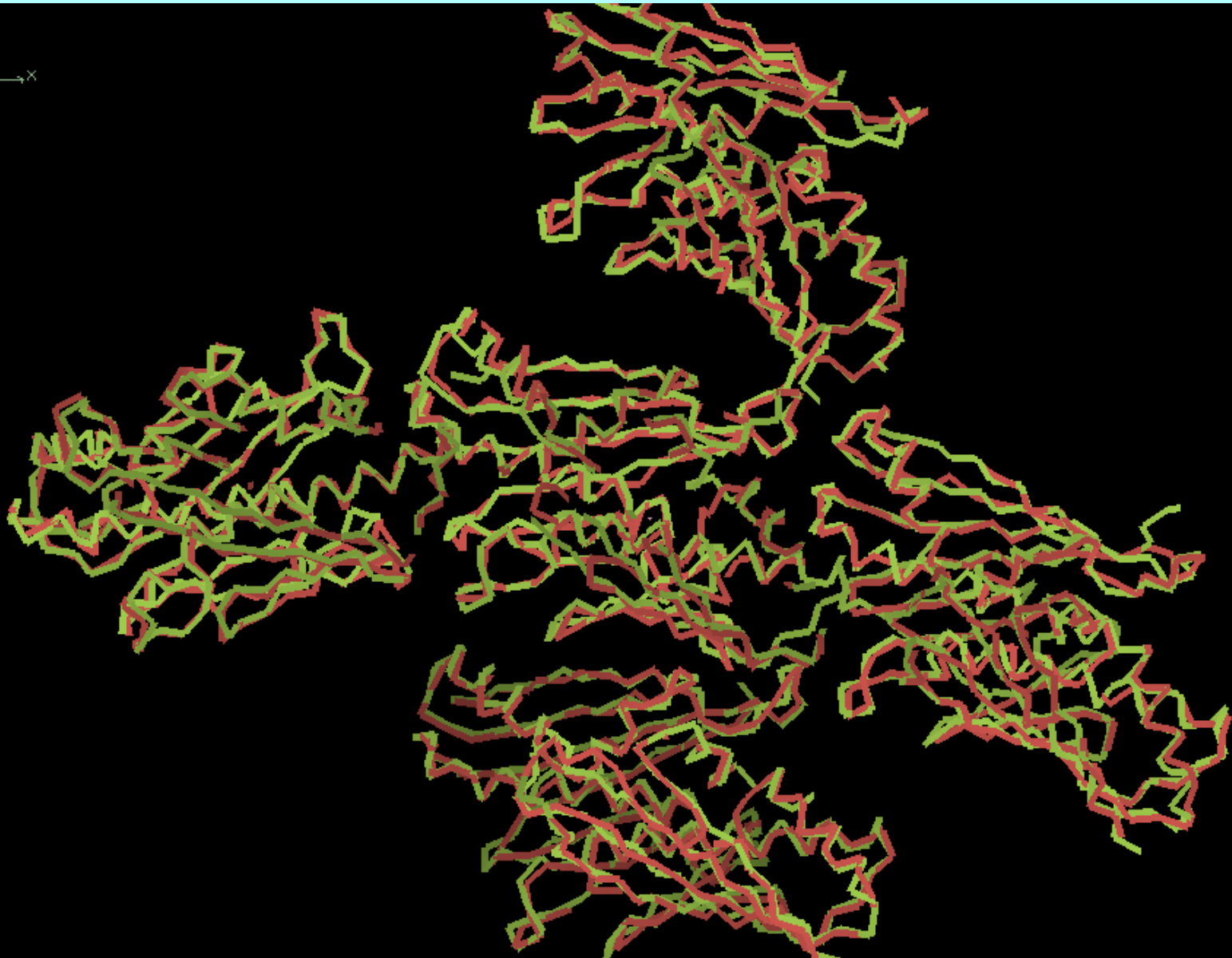
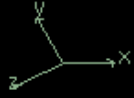
Find possible C_α trimers
(Pairs sharing C_α ; 110° angle; points near line extending from vertex)



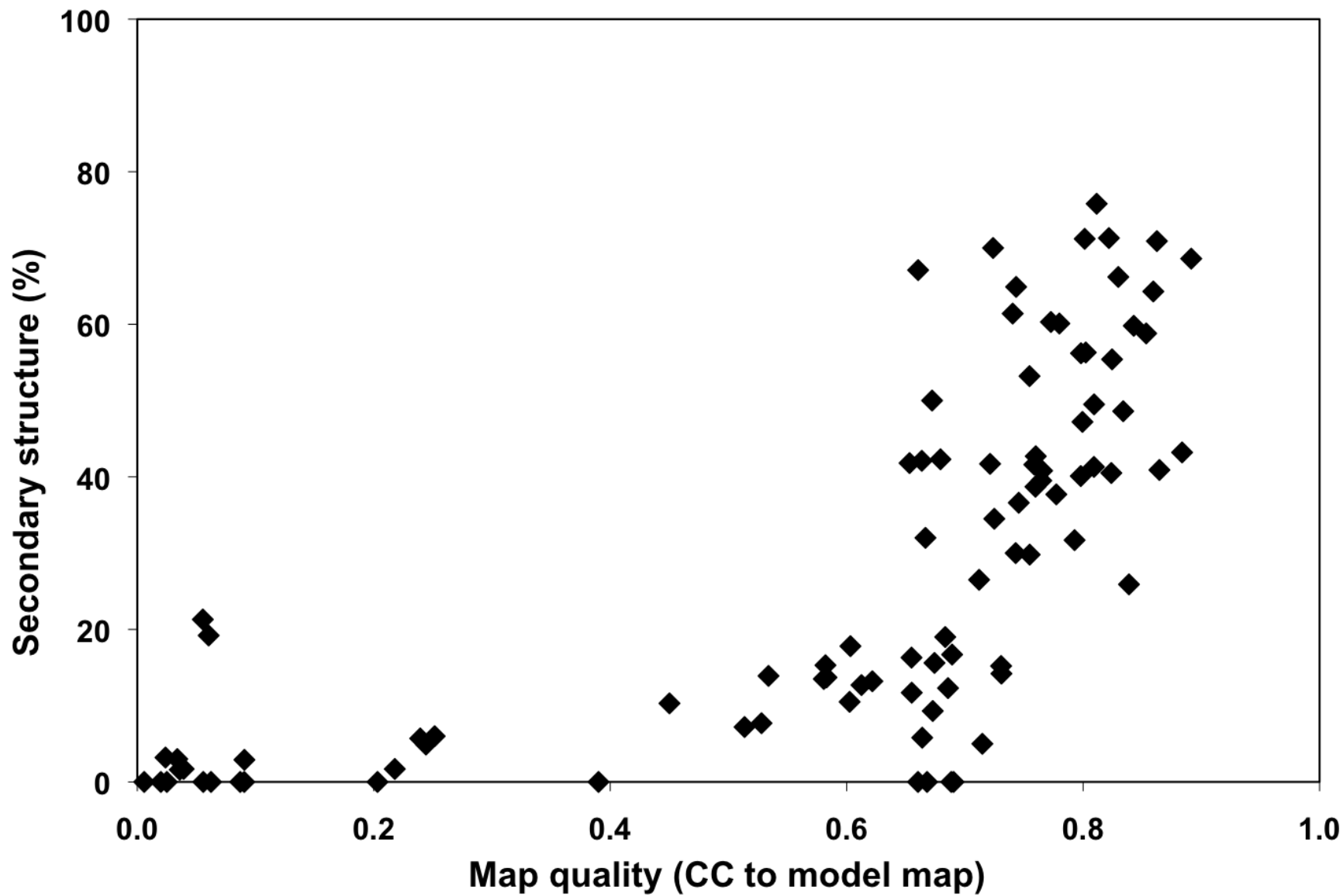
C_{α} tracing
(s-hydrolase, PDB entry 1A7A)



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



C_{α} tracing
(1038B, PDB entry 1LQL, 114 sec)



Using secondary structure content to evaluate map quality

Rapid phase improvement and model-building with *phenix.phase_and_build*

First improve the map

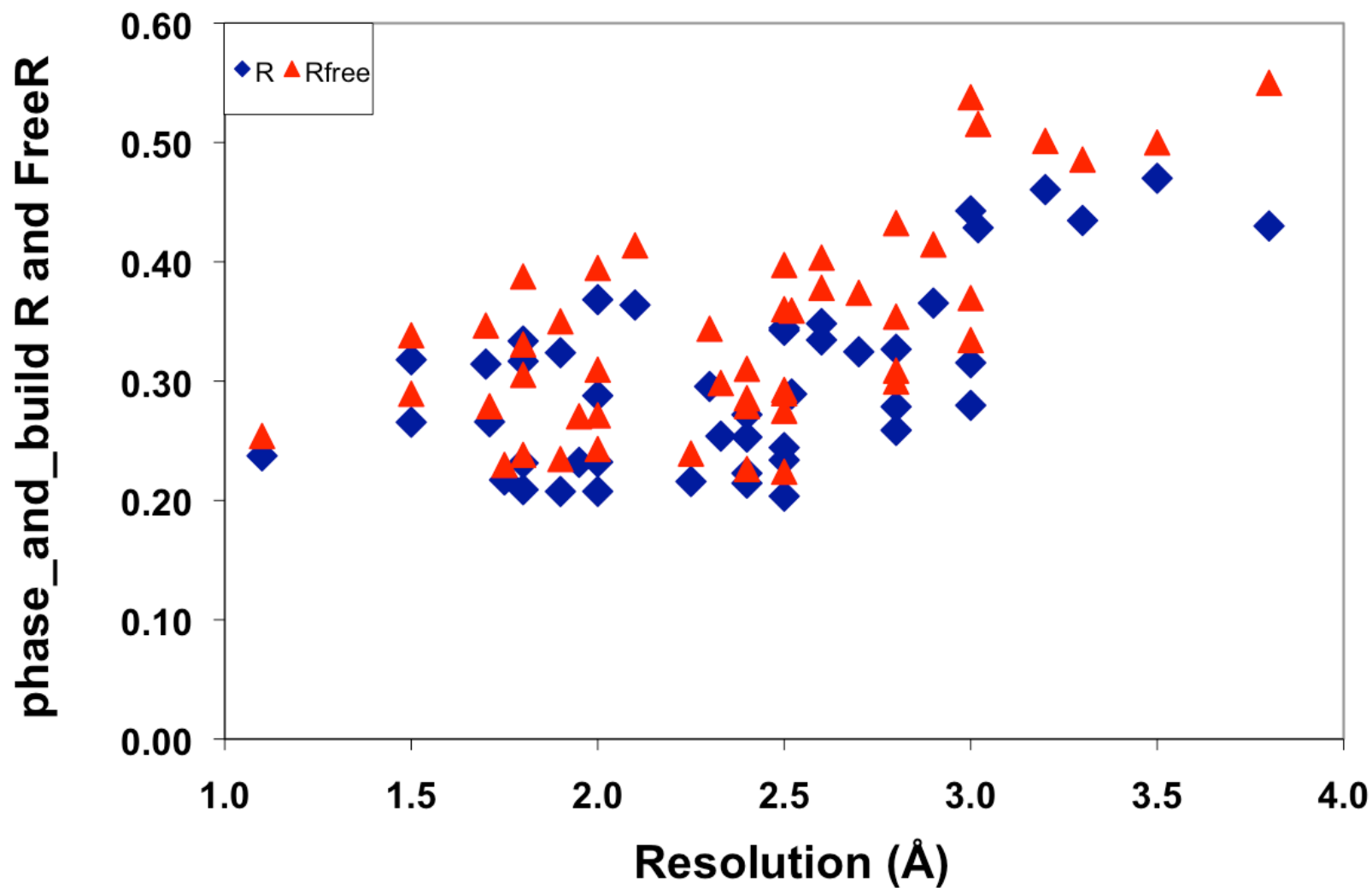
NCS identification from density
Iterative rapid model-building and density modification



Then build a full model

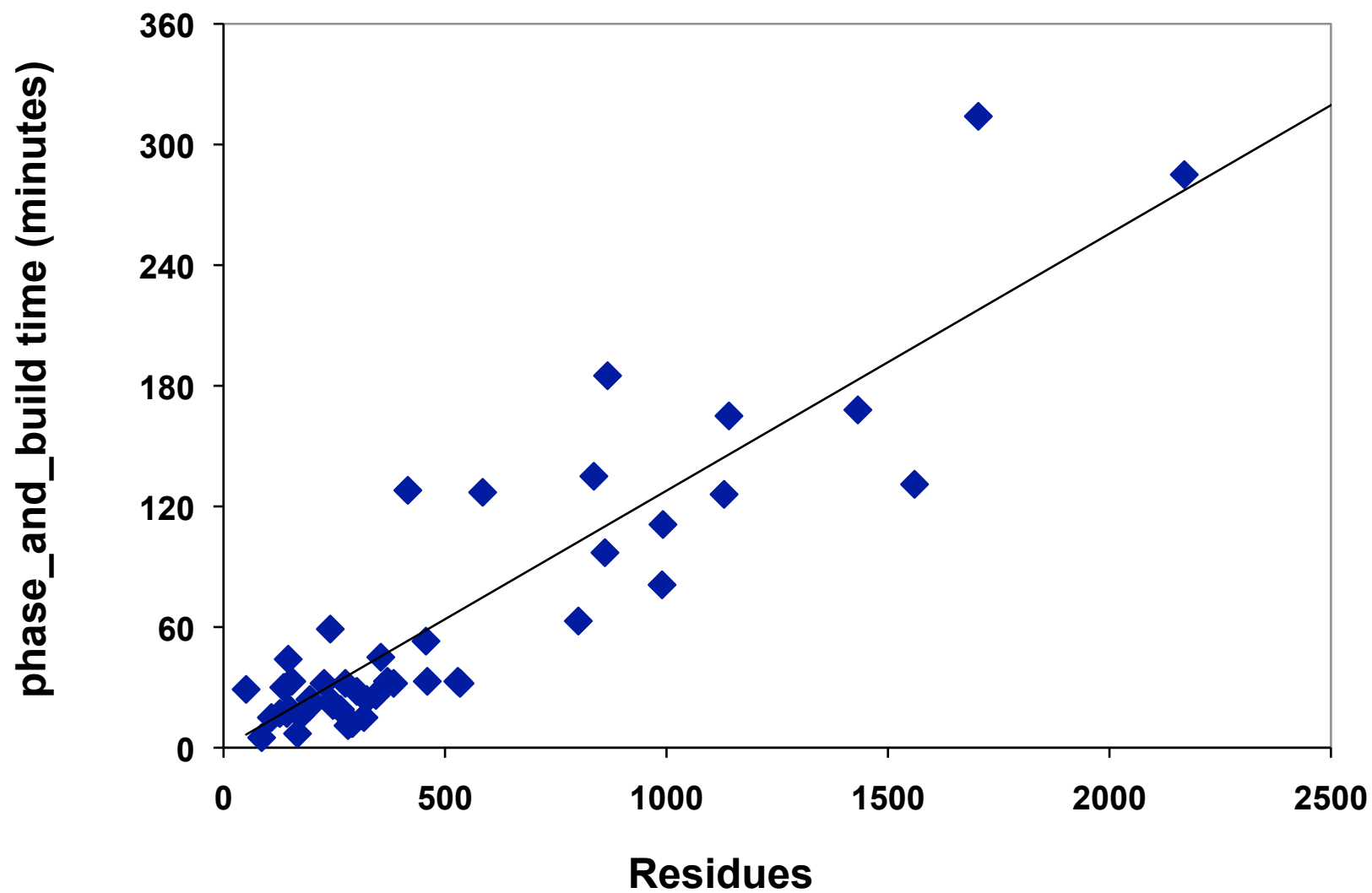
Model-building and refinement with NCS
Comprehensive sequence assignment
Loop fitting

phenix.phase_and_build – tests with structure library
Final R/Rfree



phase_and_build – tests with structure library

One cycle (approx 500 residues/hour)



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

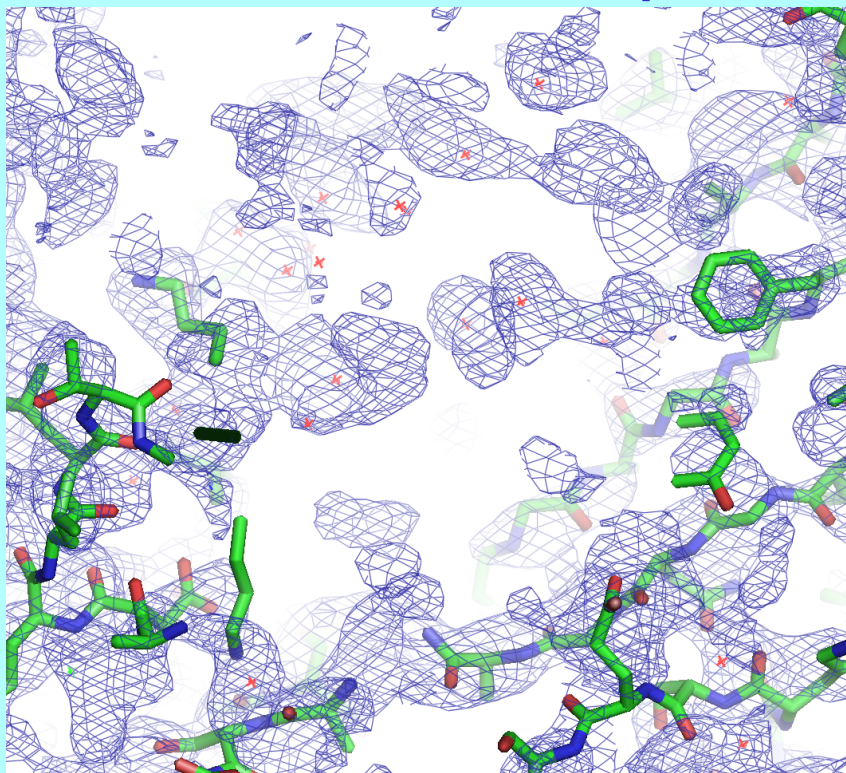
...checking if your model is biasing your map

...checking if you always get the same model

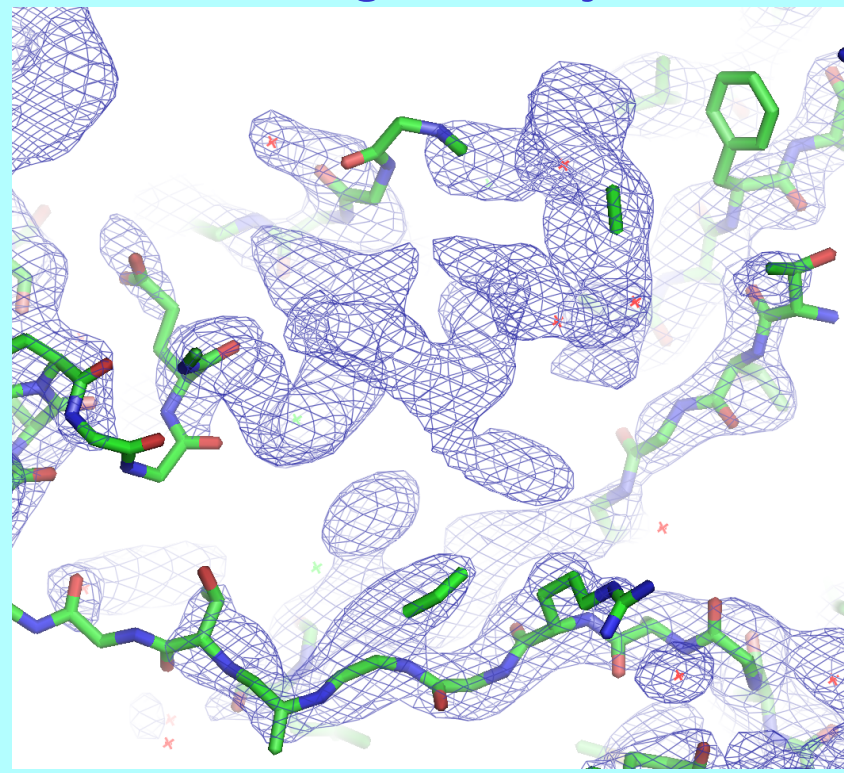
Iterative-Build OMIT procedure

“Is the density in my map biased by the model?”

2mFo-DFc omit map



*After building outside
OMIT region 10 cycles*

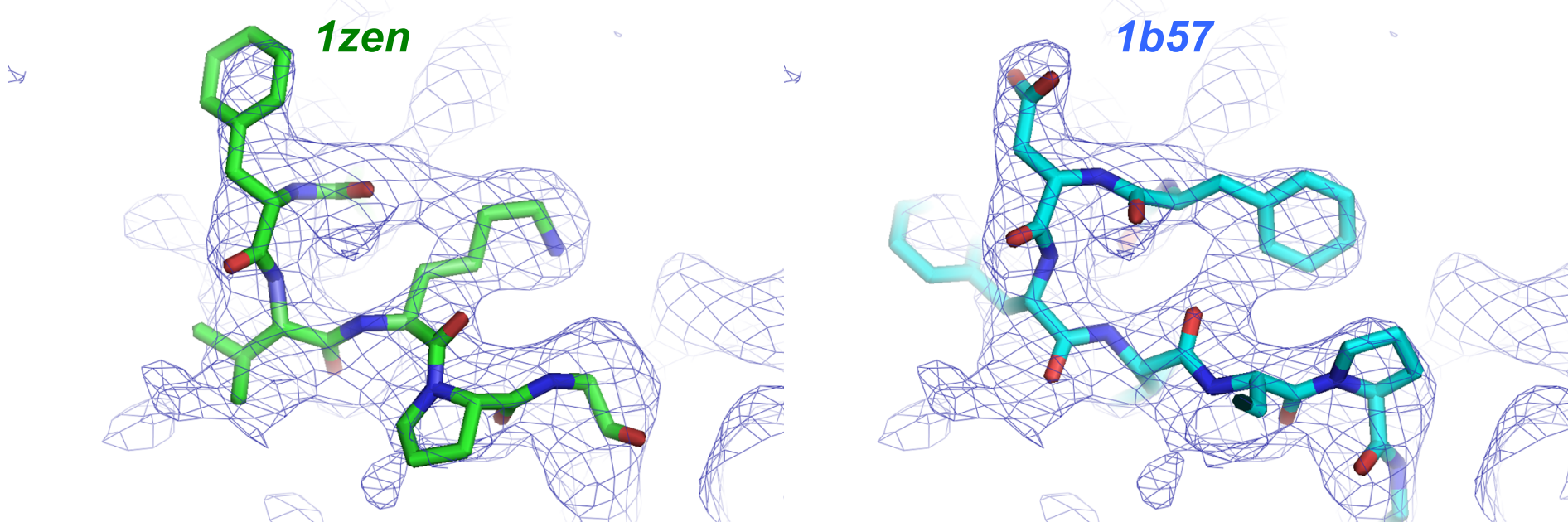


*1HP7 molecular replacement with 1AS4
R/Rfree after initial refinement: 0.41/0.48*

Iterative-Build OMIT procedure

“Removing model bias”

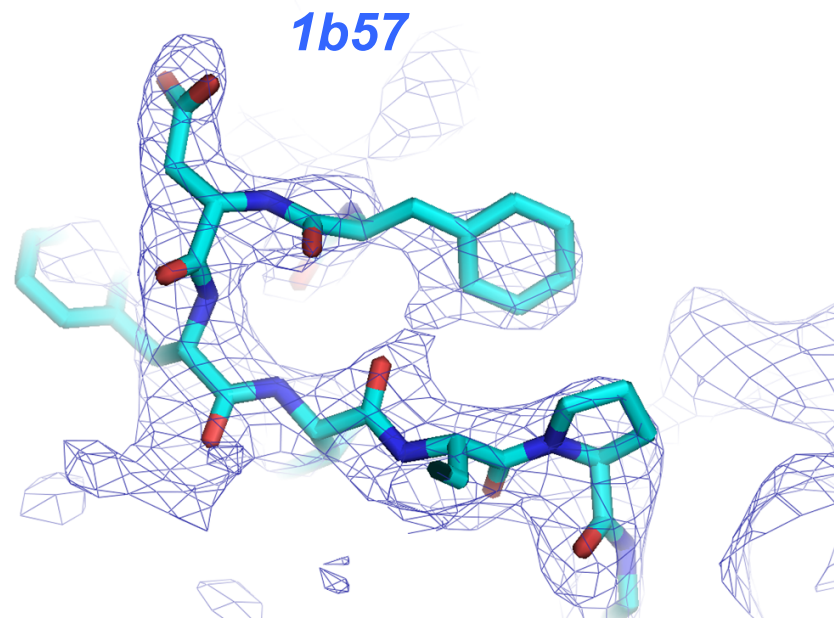
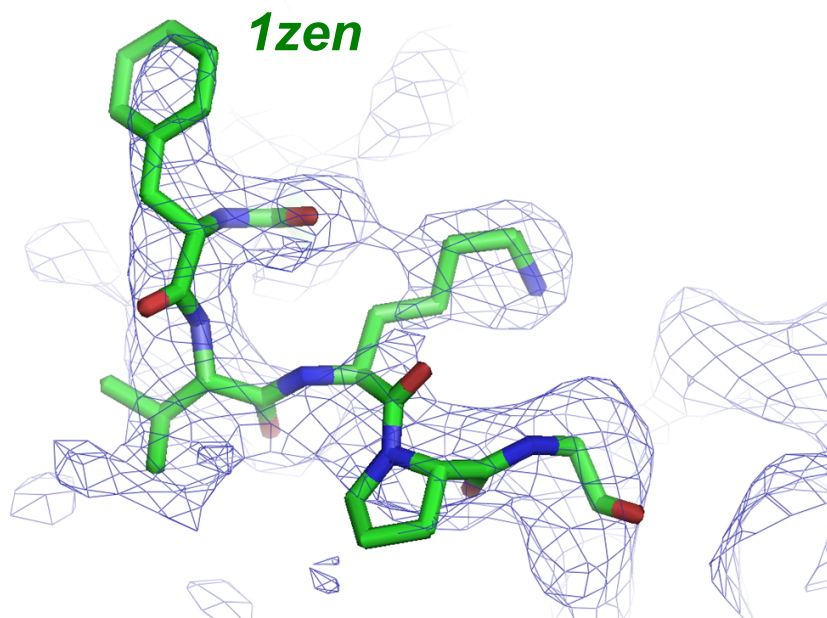
2mFo-DFc map
Phased with 1zen model



Iterative-Build OMIT procedure

“Removing model bias”

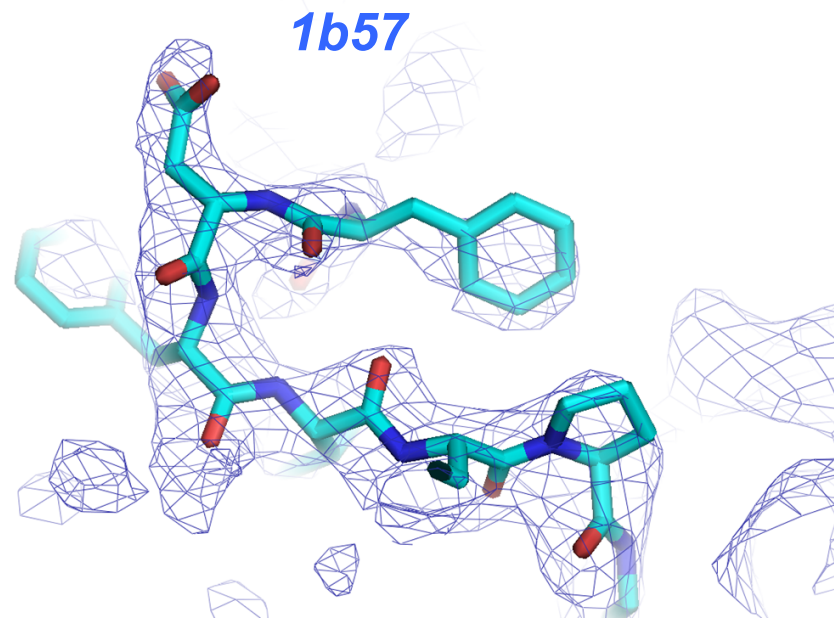
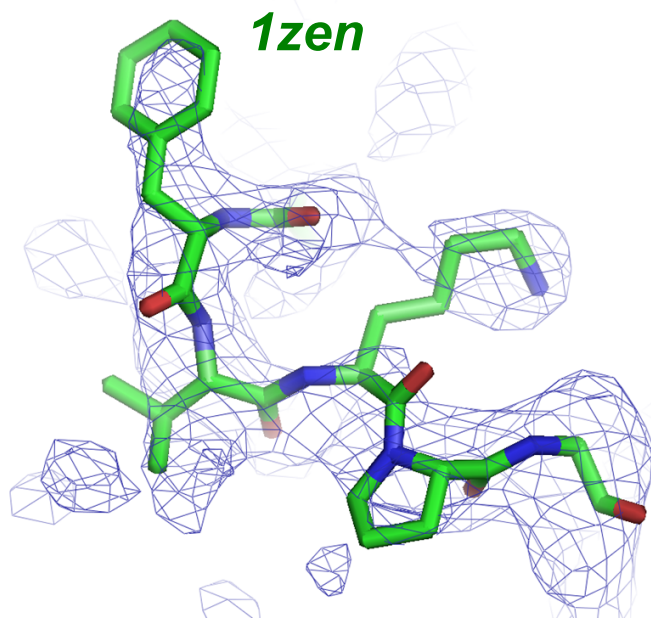
2mFo-DFc omit map
Phased with 1zen model



Iterative-Build OMIT procedure

“Removing model bias”

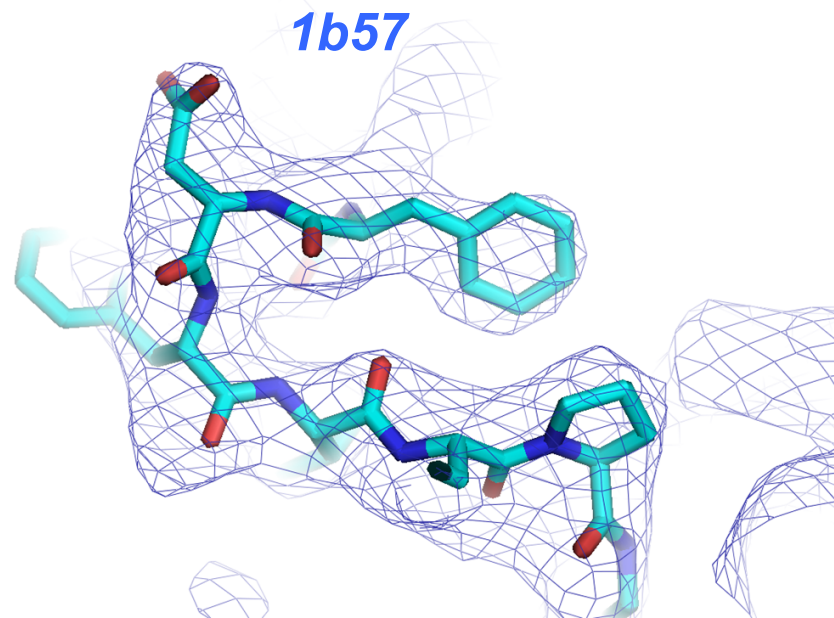
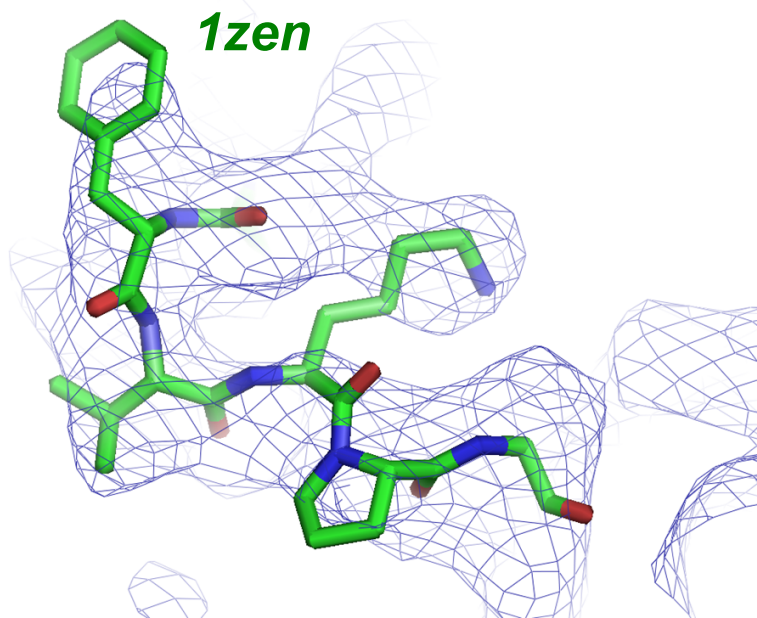
2mFo-DFc SA-omit map
Phased starting with 1zen model



Iterative-Build OMIT procedure

“Removing model bias”

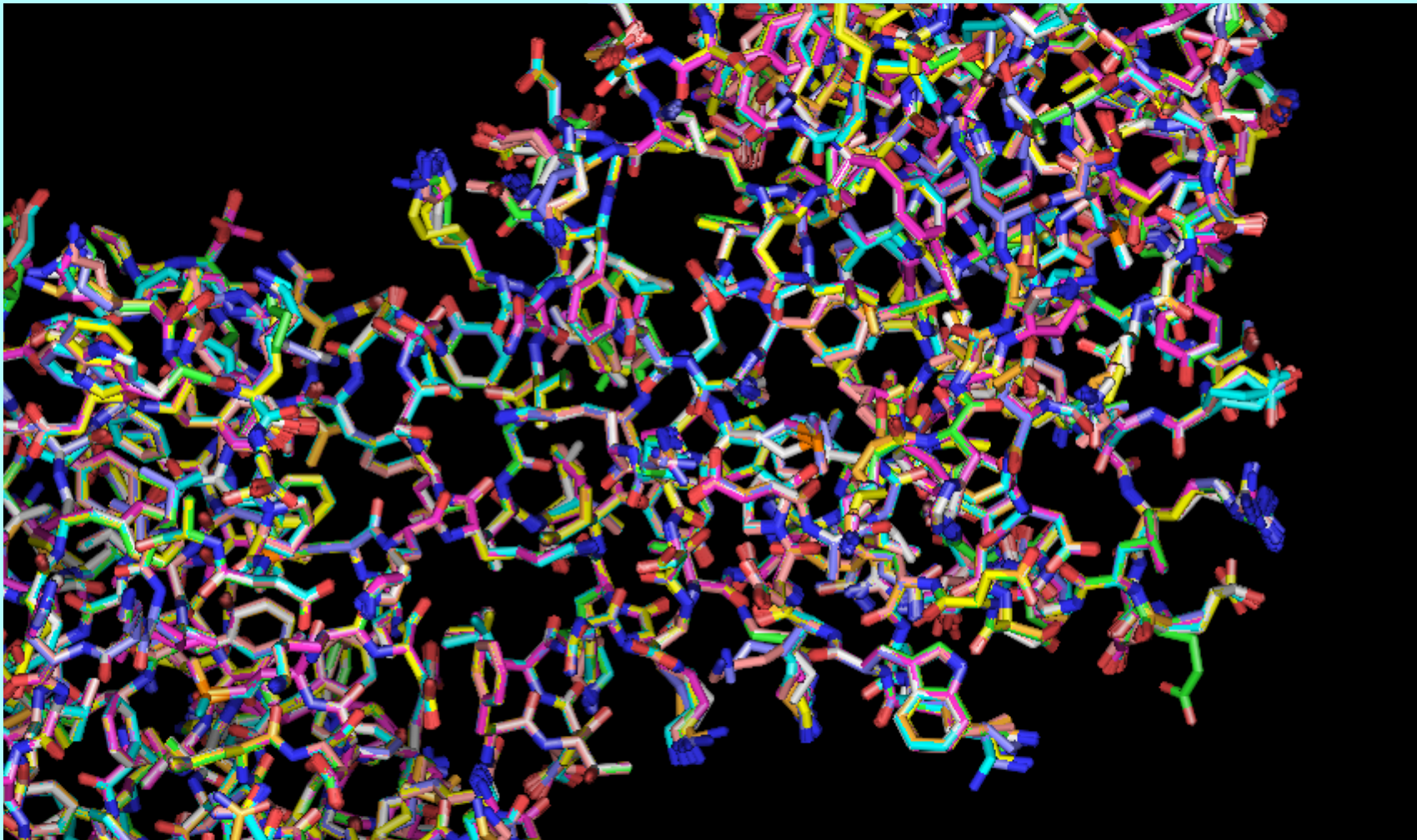
2mFo-DFc iterative-build omit map
Phased starting with 1zen model



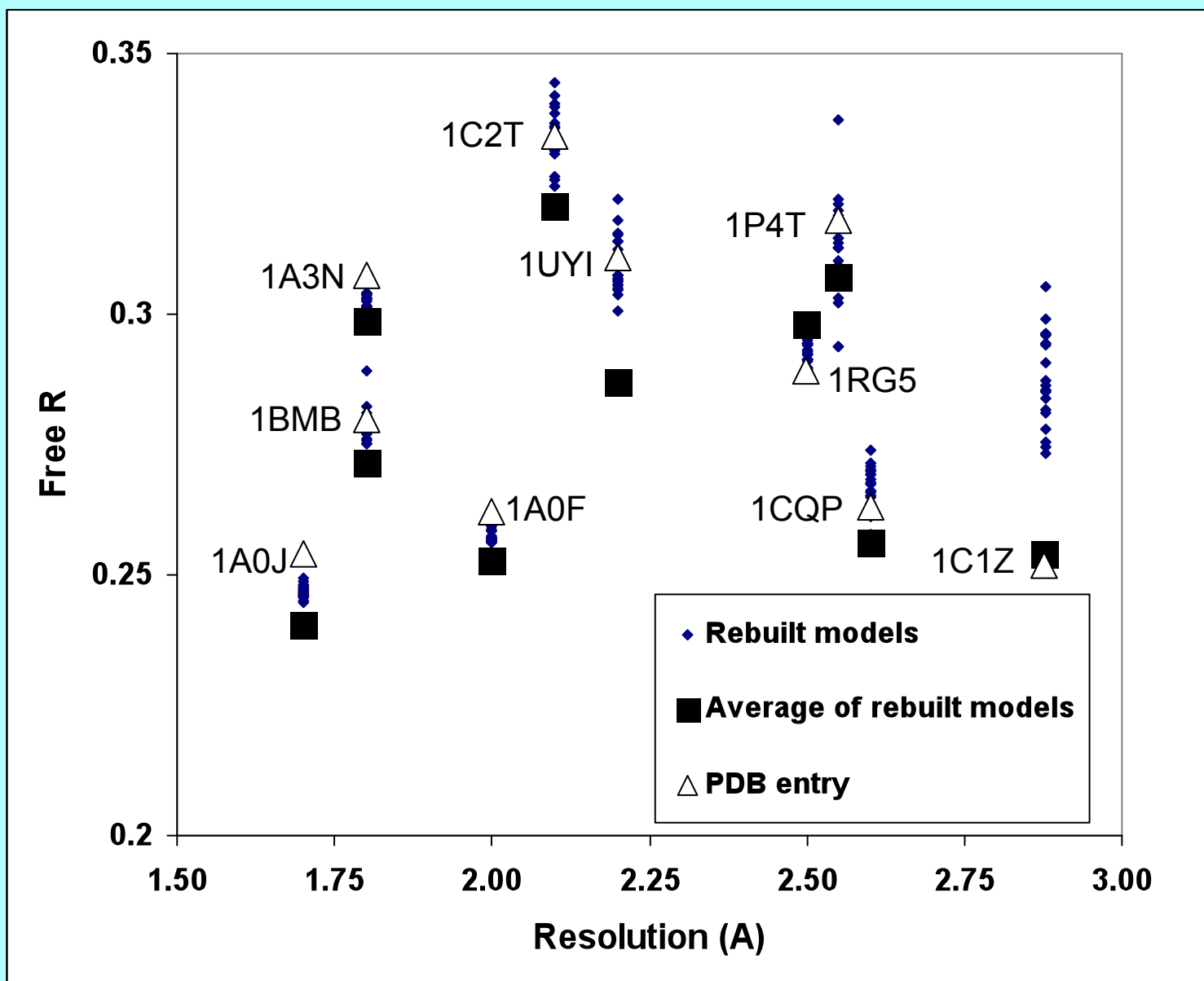
Multiple-model representation of uncertainties

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

The variation among models is a lower bound on their uncertainty

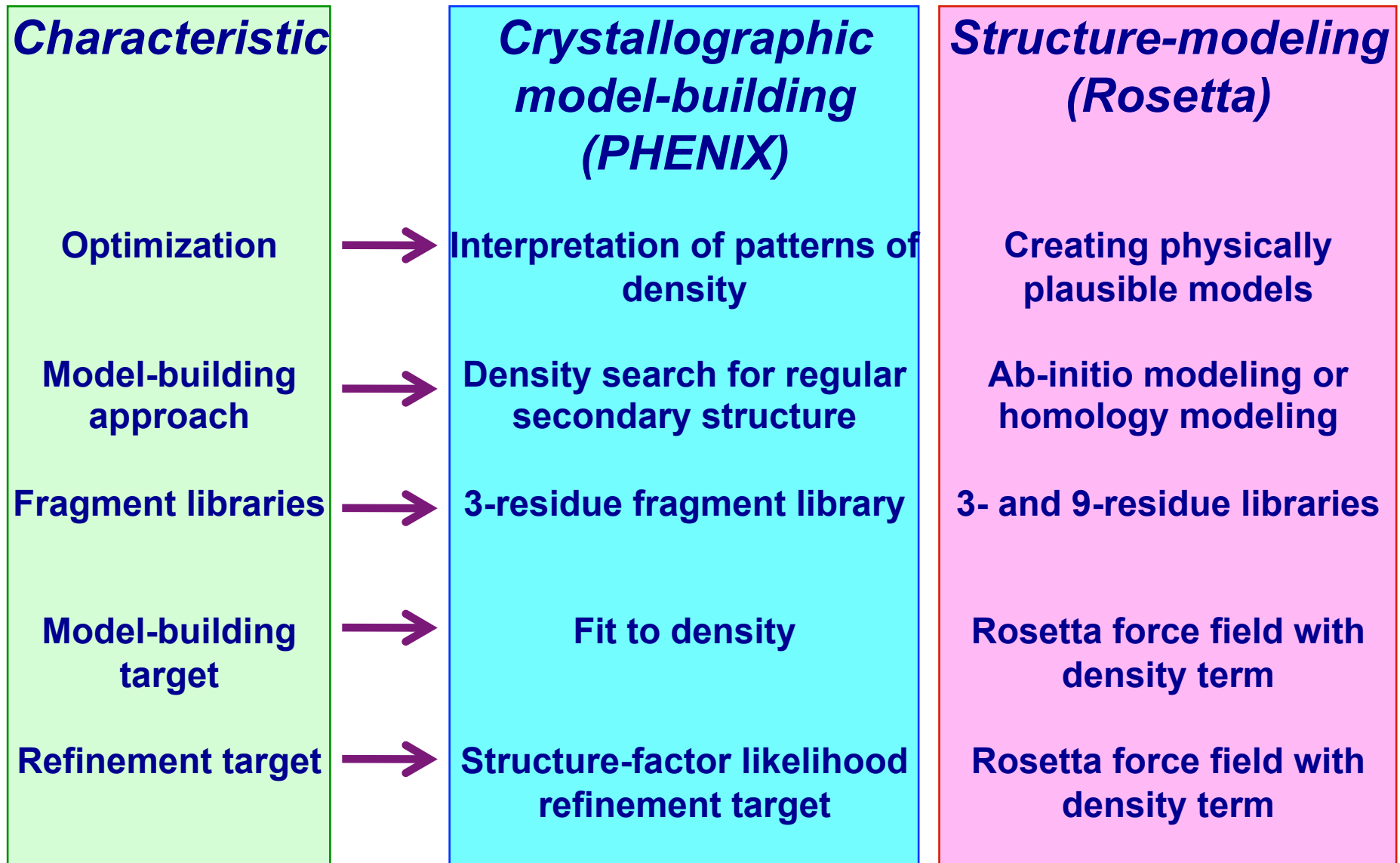


Building 20 models for each of 10 structures



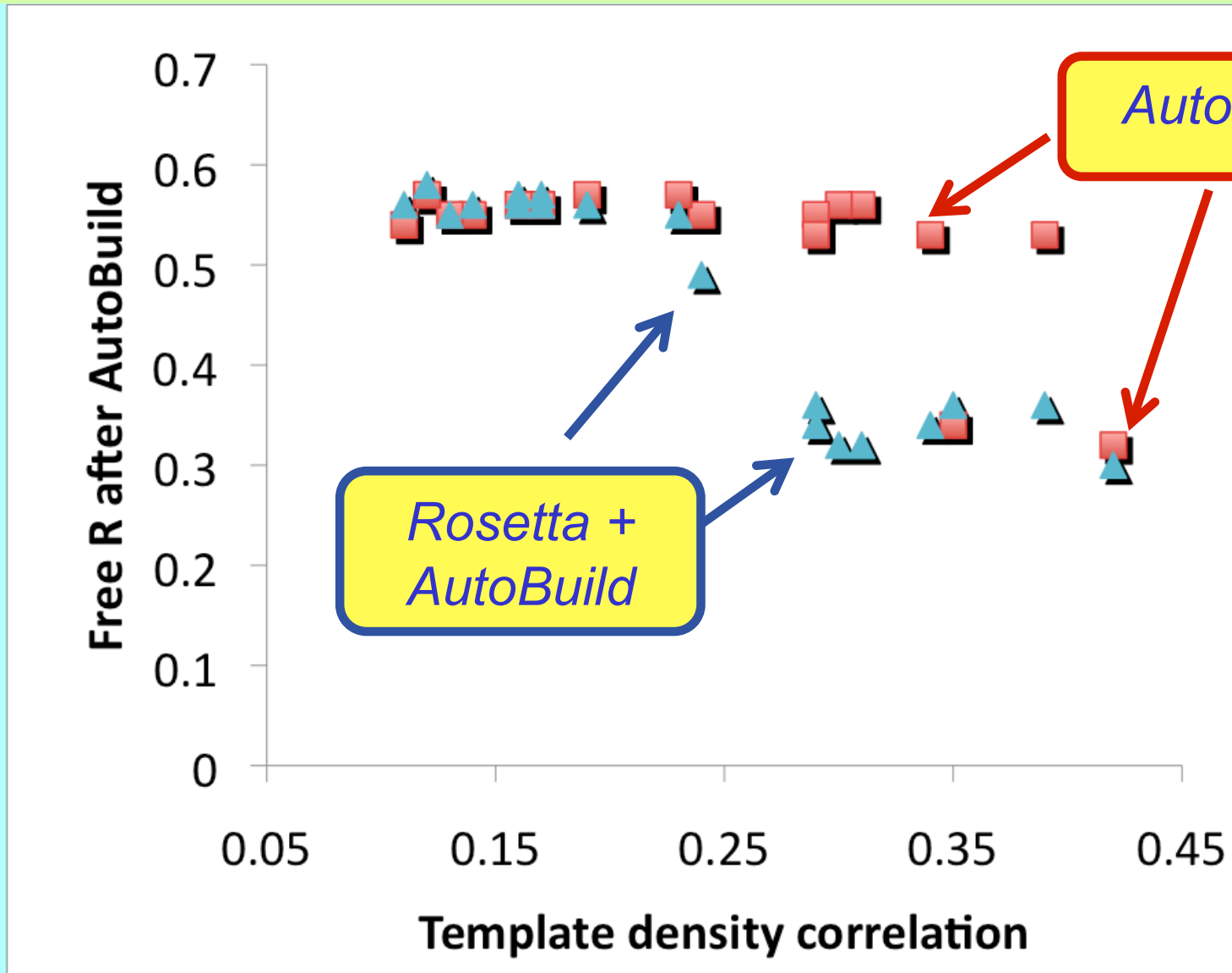
Complementarity of PHENIX and Rosetta model-building

(Randy Read, David Baker, Frank DiMaio)



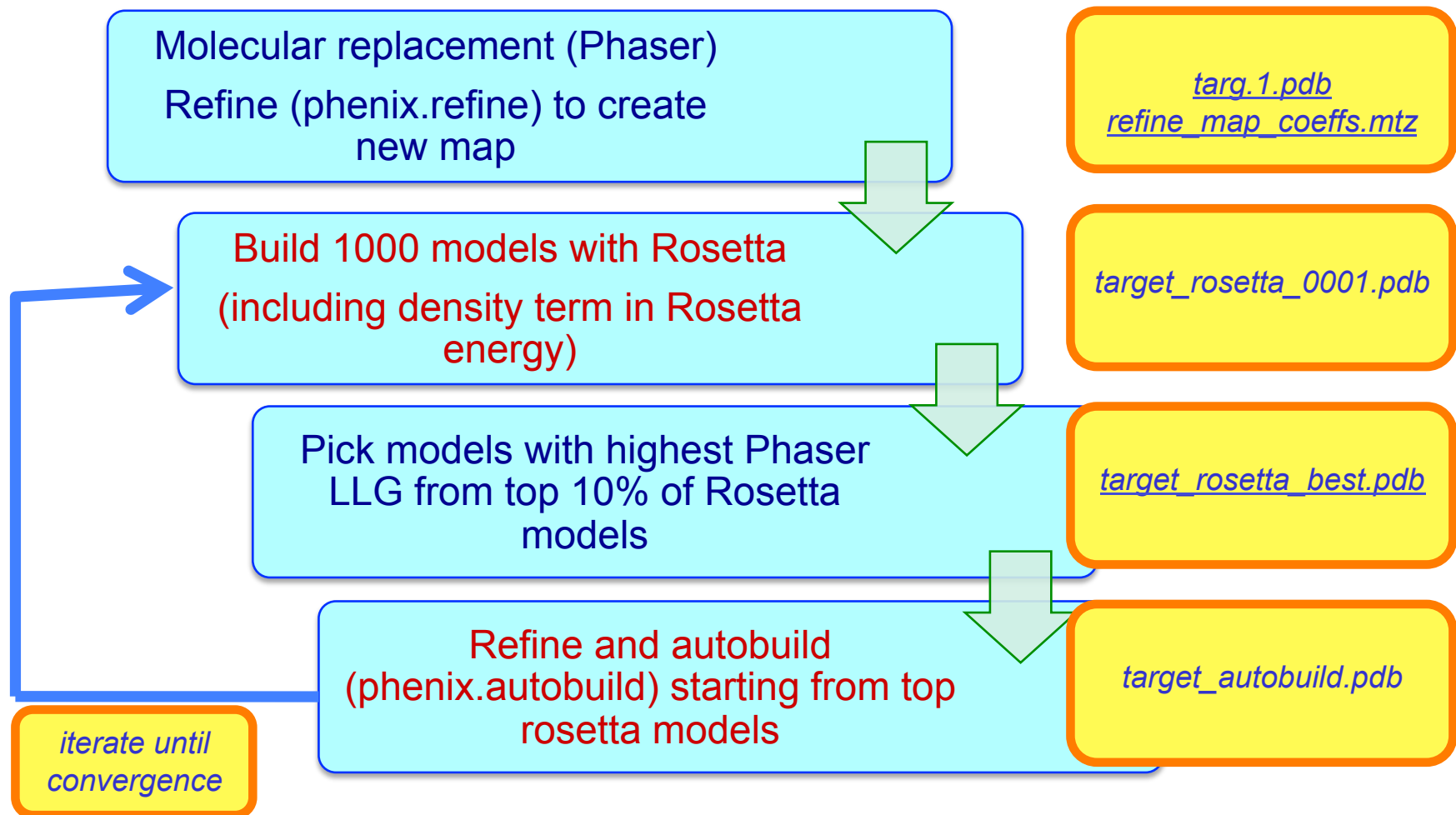
Combining structure-modeling with crystallographic model-building

20 templates for 1XVQ from PDB (optimally superimposed)



Molecular replacement using distant homology models with PHENIX and Rosetta (*phenix.mr_rosetta*)

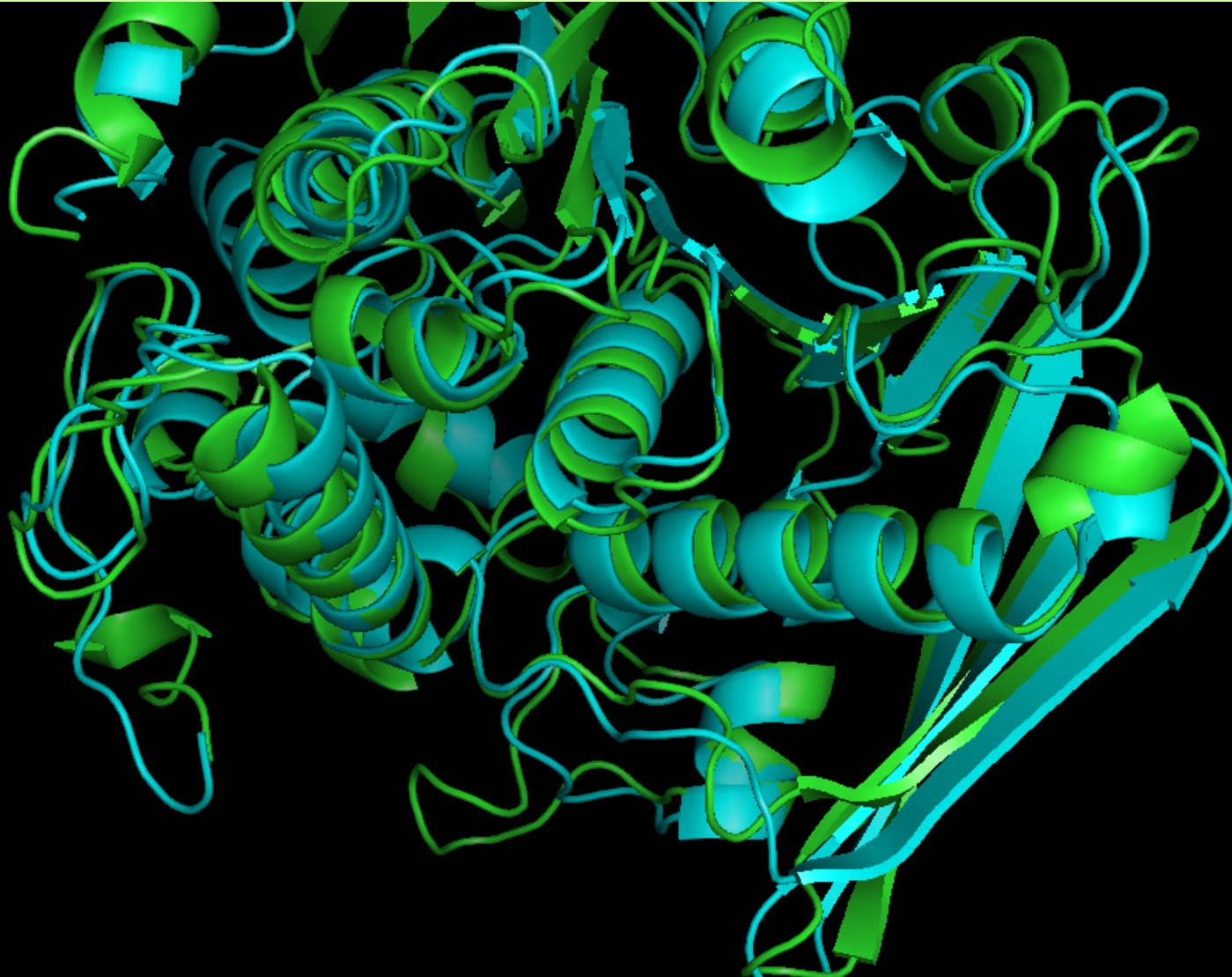
Molecular replacement and model-building



Structure determination of cab55348 (using template supplied by user)

1.9 Å, 28% sequence identity (*AutoMR alone fails with R/Rfree=0.47/0.53*)

MR model: blue, Final model: green

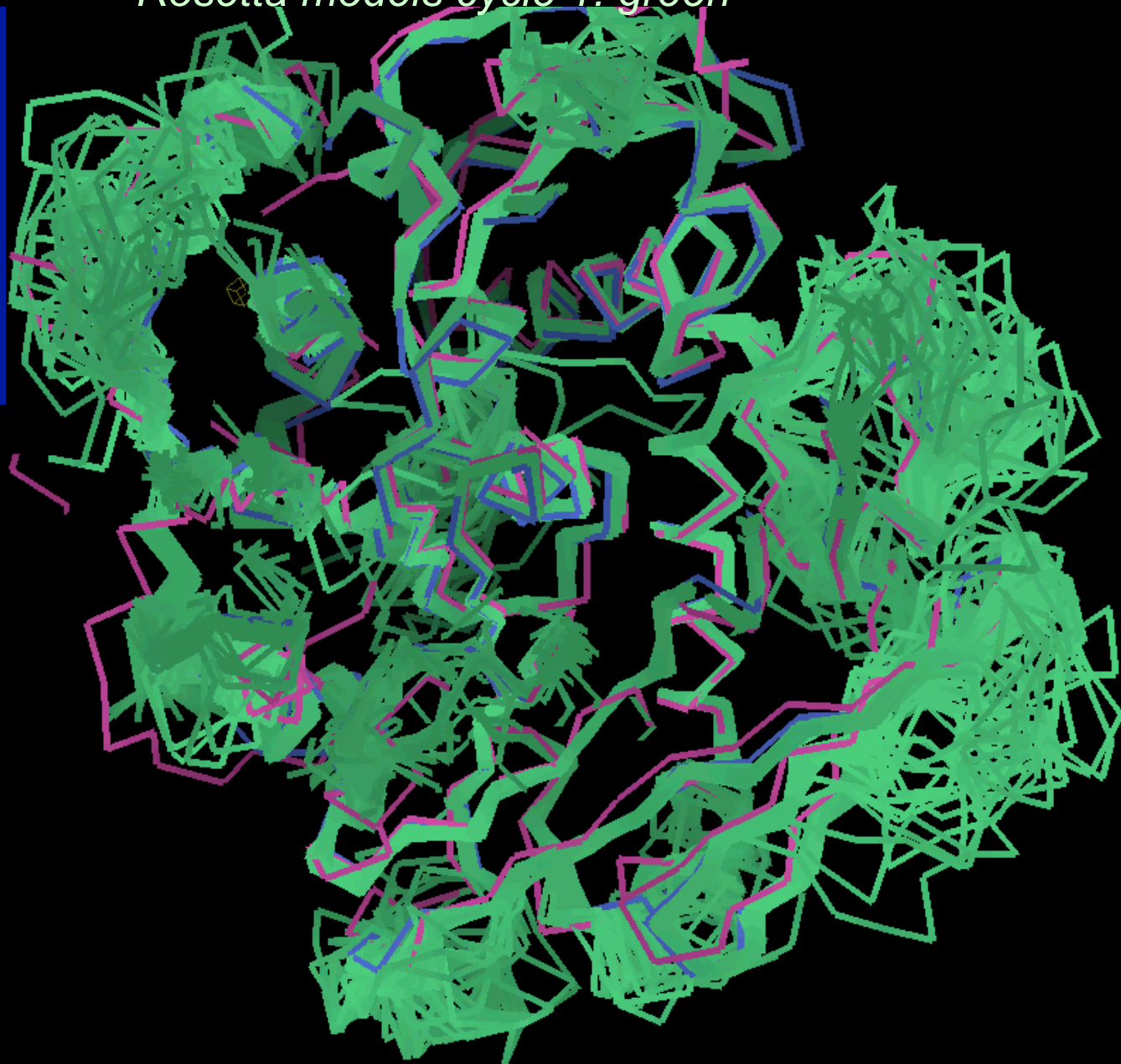


MR model : blue

Final model: pink

Rosetta models cycle 1: green

Sample
Rosetta
models in
cycles 1 and
2,

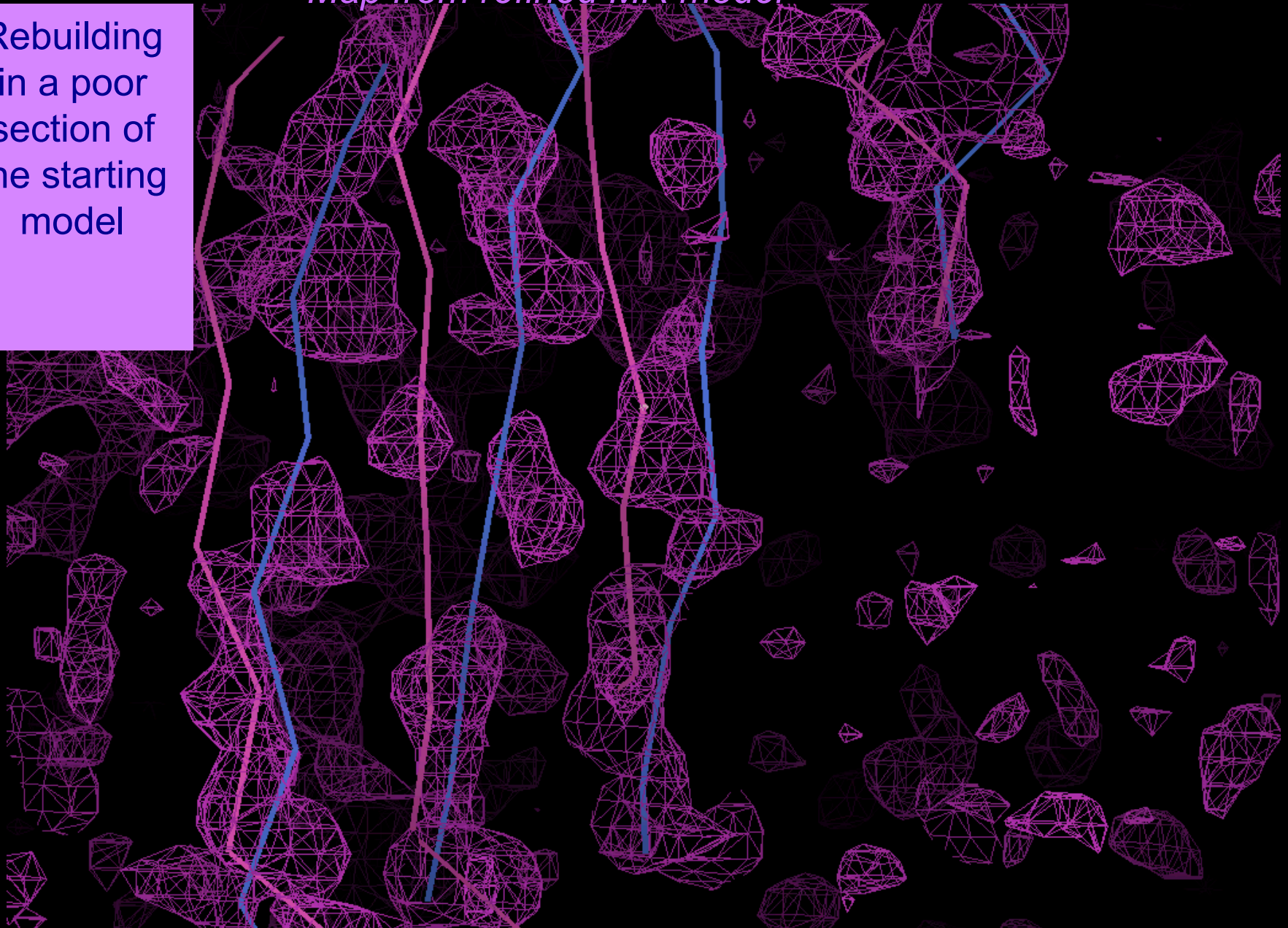


MR model : blue

Final model: pink

Map from refined MR model

Rebuilding
in a poor
section of
the starting
model



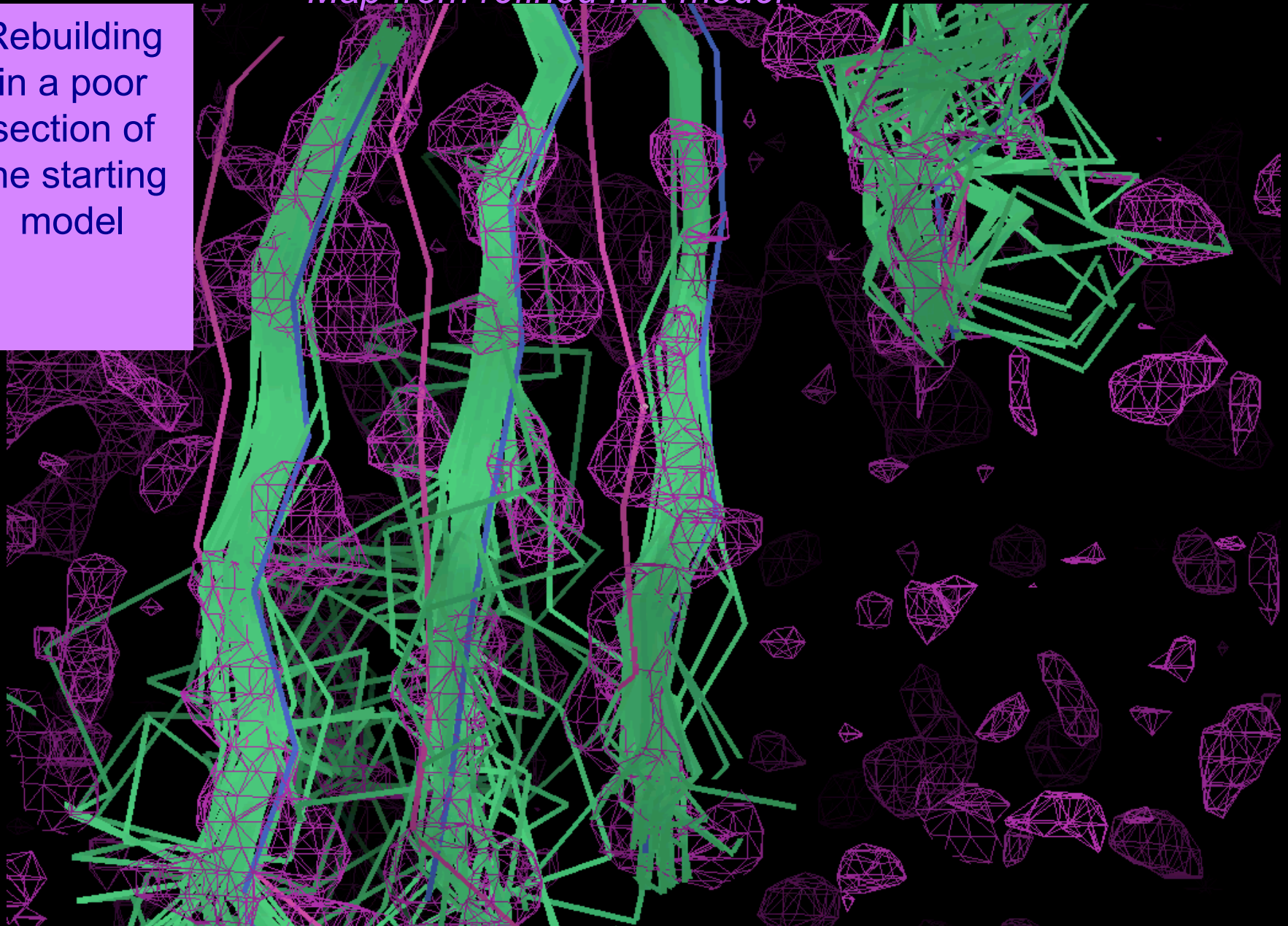
MR model : blue

Final model: pink

Rosetta models cycle 1: green

Map from refined MR model

Rebuilding
in a poor
section of
the starting
model



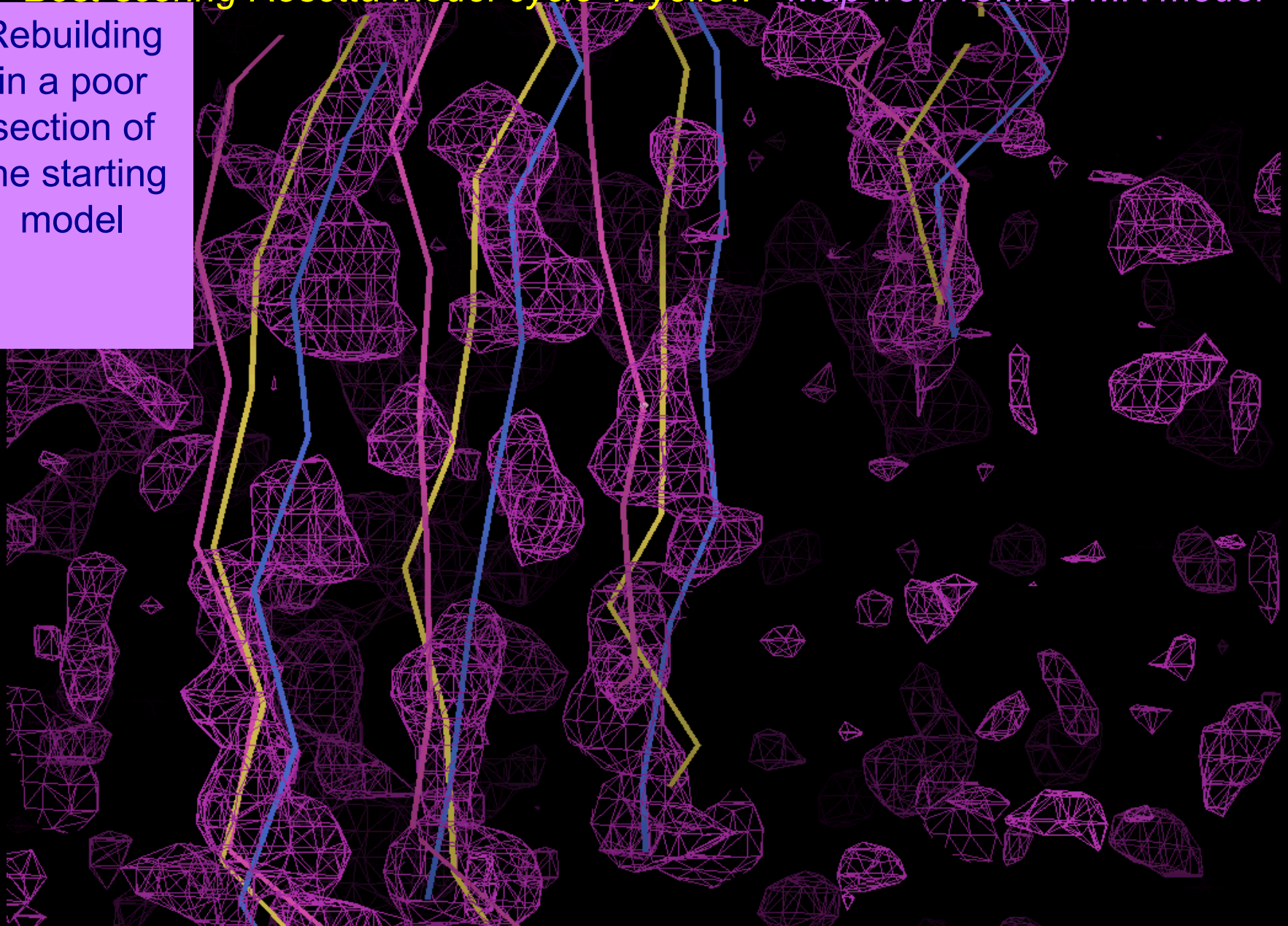
MR model : blue

Final model: pink

Best-scoring Rosetta model cycle 1: yellow

Map from refined MR model

Rebuilding
in a poor
section of
the starting
model



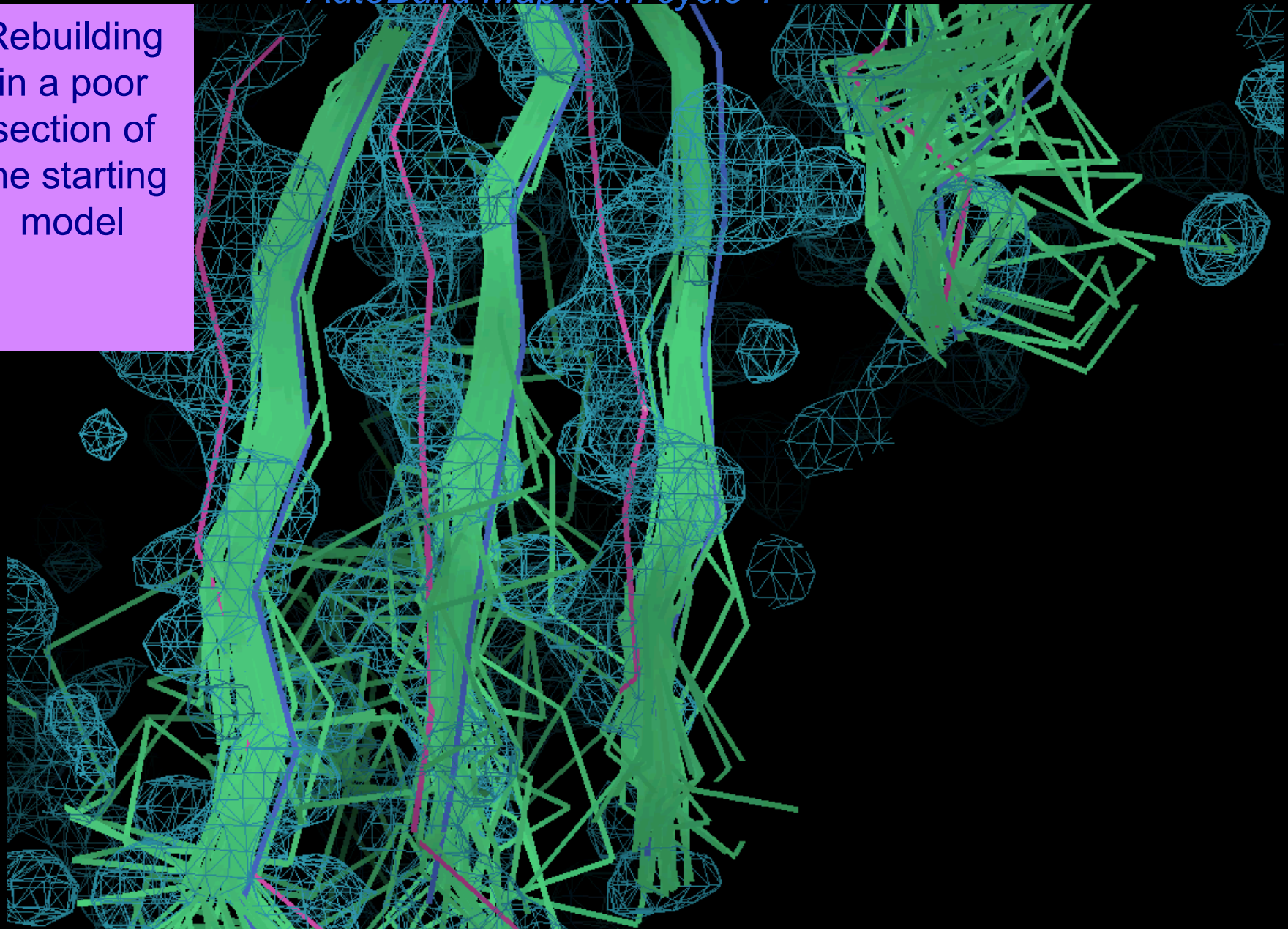
MR model : blue

Final model: pink

Rosetta models cycle 1: green

AutoBuild Map from cycle 1

Rebuilding
in a poor
section of
the starting
model



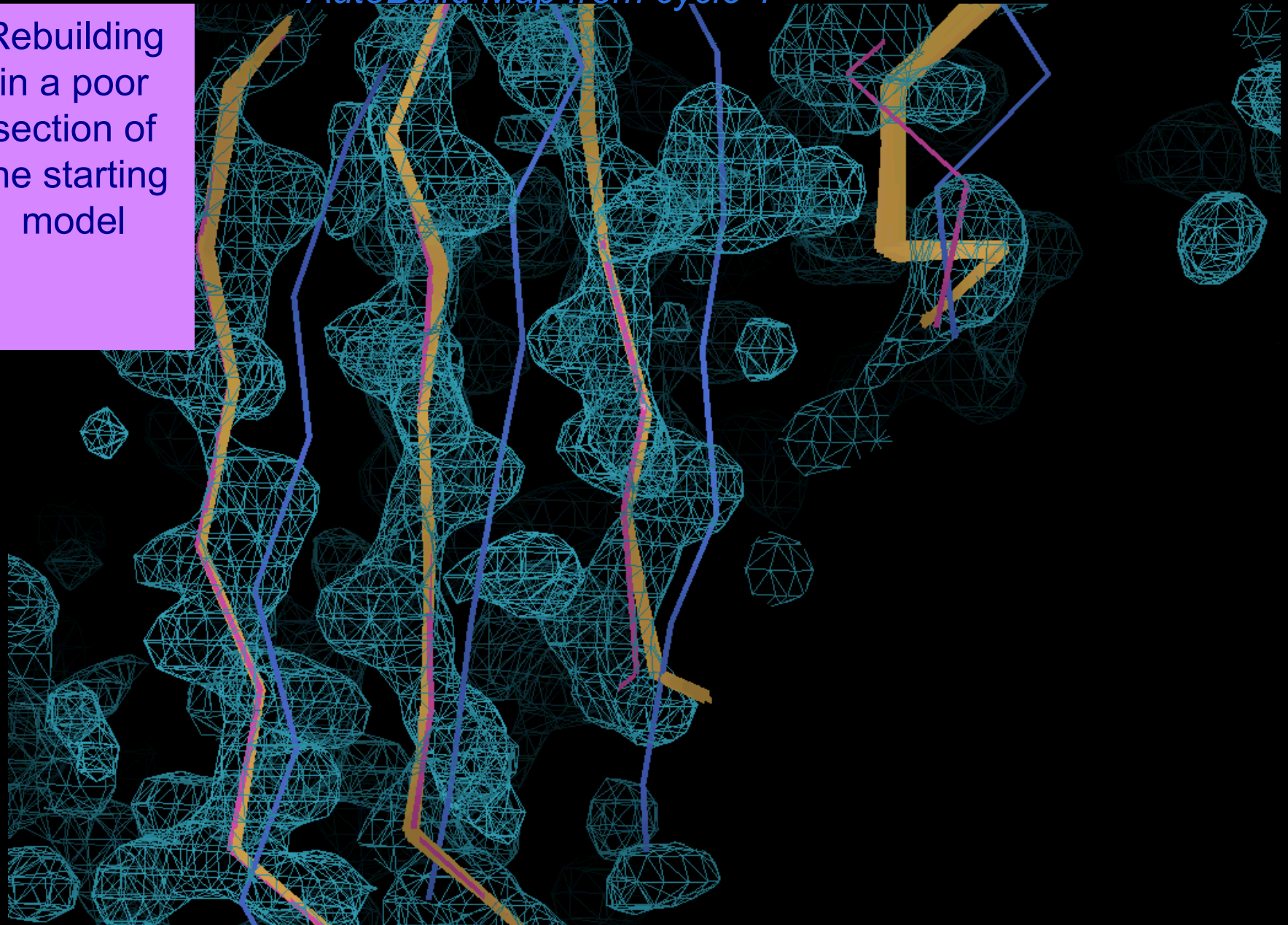
MR model : blue

Final model: pink

Rosetta models cycle 2: yellow

AutoBuild Map from cycle 1

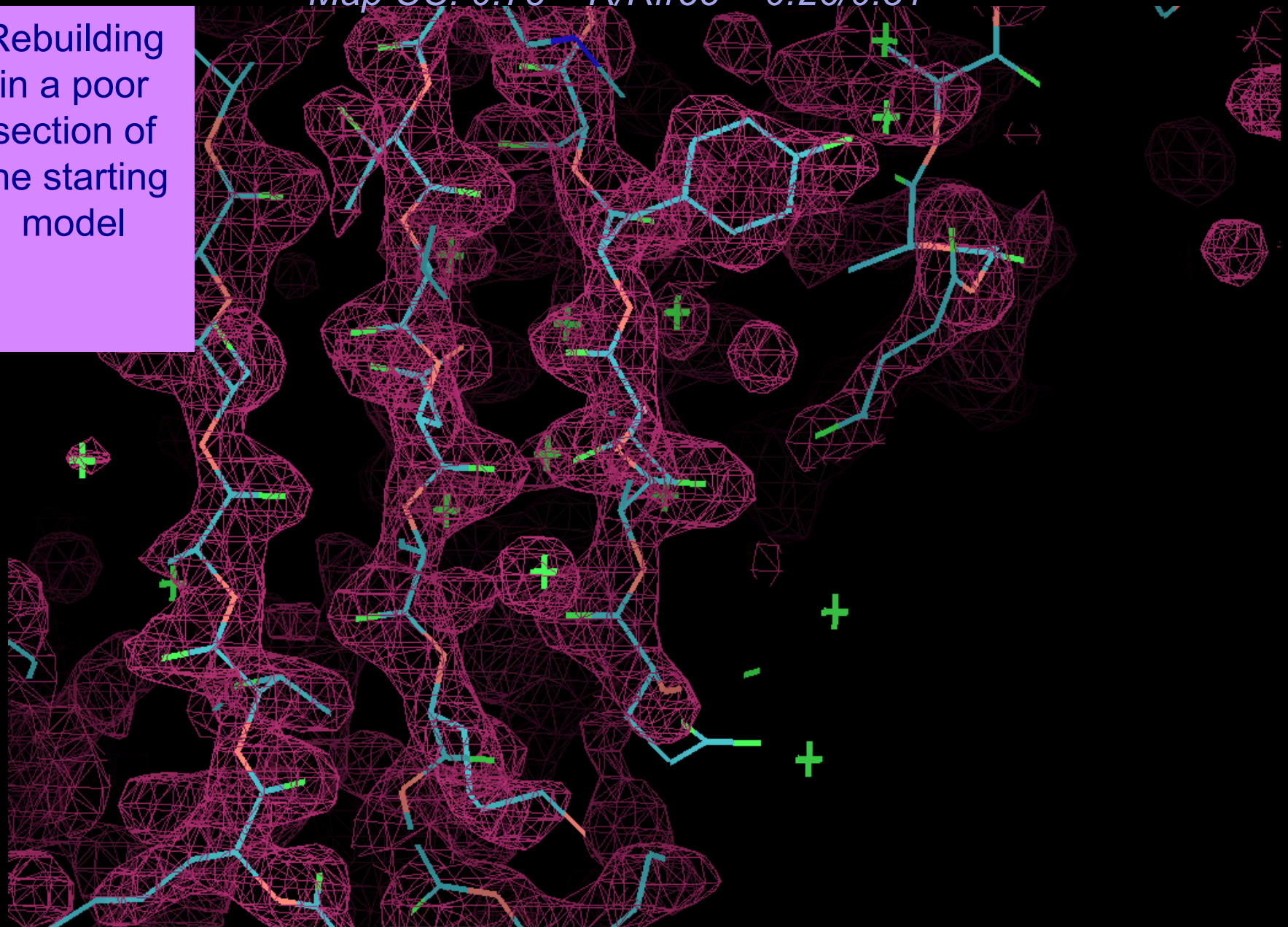
Rebuilding
in a poor
section of
the starting
model



AutoBuild model cycle 2

Map CC: 0.78 R/Rfree = 0.26/0.31

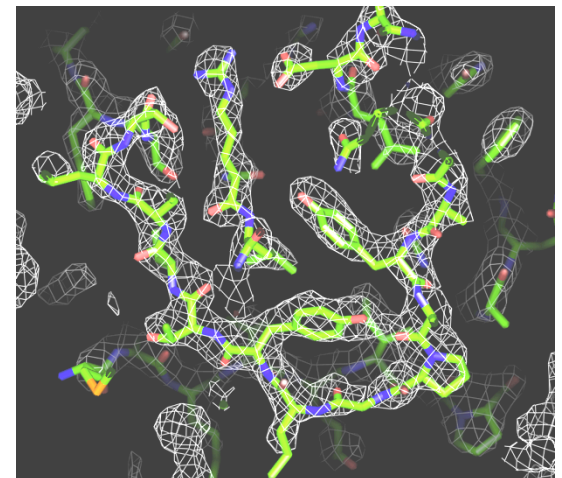
Rebuilding
in a poor
section of
the starting
model



Wizards



- AutoSol Wizard: Structure solution (MIR/MAD/SAD) with HYSS/Phaser/Solve/Resolve
- AutoBuild Wizard: Iterative density modification, model-building and refinement with Resolve/phenix.refine/Elbow; model rebuilding in place; touch-up of model; simple OMIT; SA-OMIT; Iterative-build OMIT; OMIT around atoms in a PDB file; protein, RNA, DNA model-building
- LigandFit Wizard: automated fitting of flexible ligands
- AutoMR Wizard: Phaser molecular replacement followed by automatic rebuilding



MODEL-BUILDING TOOLS



- phenix.find_ncs: Find and evaluate NCS from density, heavy-atom sites, or model
- phenix.apply_ncs: Apply NCS operators to a single chain
- phenix.build_one_model: Resolve rapid model-building with real-space refinement
- phenix.phase_and_build: Improve map by model-building and refinement, then build full model
- phenix.find_helices_strands: Trace chain or build secondary structure from a map

REFINEMENT AND TOOLS



- phenix.refine: fully automatic/fully flexible refinement, SA-refinement, NCS identification, TLS, torsion-angle refinement, twin refinement
- phenix.xtriage: twinning, twin laws, anisotropy, anomalous signal, outliers, space group
- phenix.builder: ligand structures and CIF definitions from SMILES, PDB....
- phenix.ligand_identification: identify ligand density with class-specific libraries
- phenix.validation, phenix.model_vs_data, phenix.real_space_correlation, phenix.get_cc_mtz_mtz: Molprobity and density analysis of structures and density maps
- phenix.pdbtools, phenix.reflection_file_editor: manipulate PDB and mtz files
- ...and many more: see [phenix.doc](#) and www.phenix-online.org

The PHENIX Project



Lawrence Berkeley Laboratory

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



Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung



Randy Read, Airlie McCoy, Gabor Bunkoczi, Rob Oeffner

Cambridge University



Duke University

Jane & David Richardson, Vincent Chen, Chris Williams, Bryan Arendall, Laura Murray



*An NIH/NIGMS funded
Program Project*