Automation of Structure Solution

Diffraction Methods in Molecular Biology
Gordon Research Conference
July 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

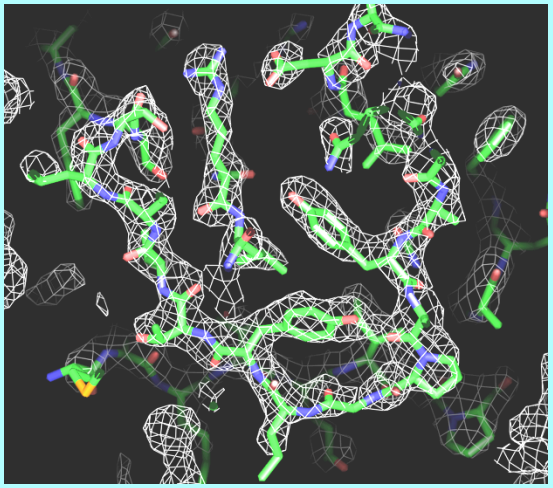
can speed up the process

can help reduce errors

Automation also allows you to…

try more possibilities

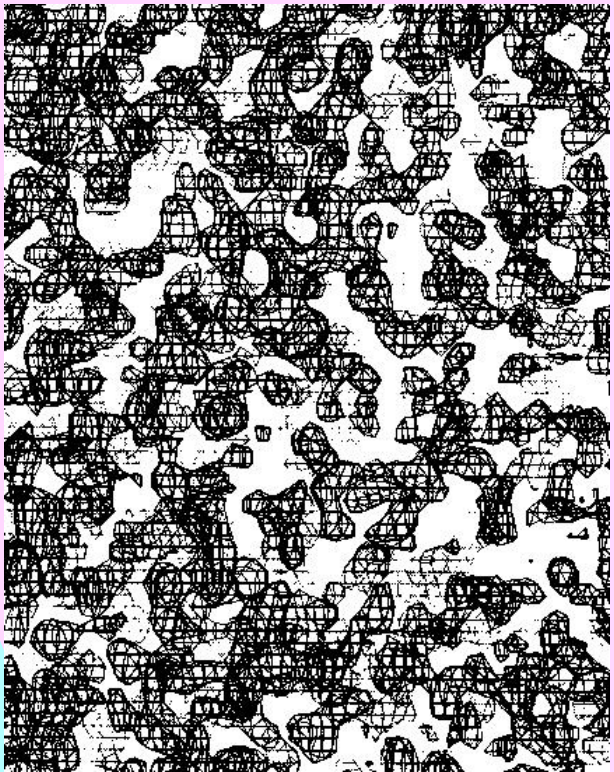
estimate uncertainties



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

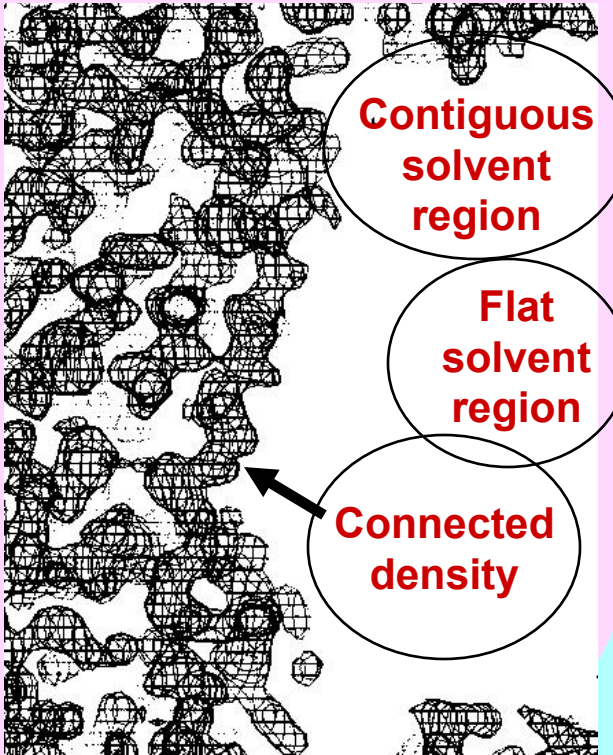
Which solution is best?

Are we on the right track?

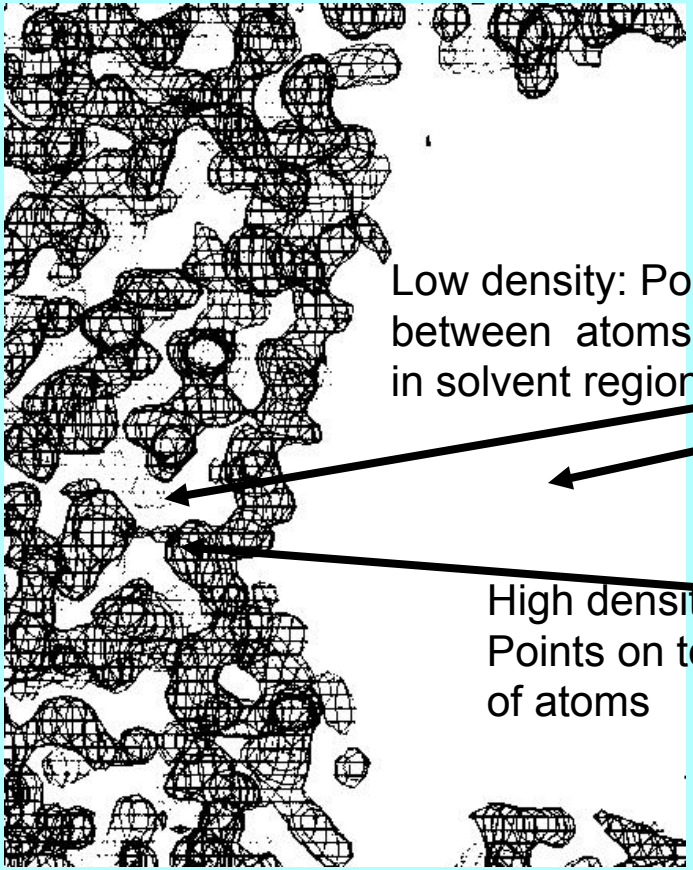


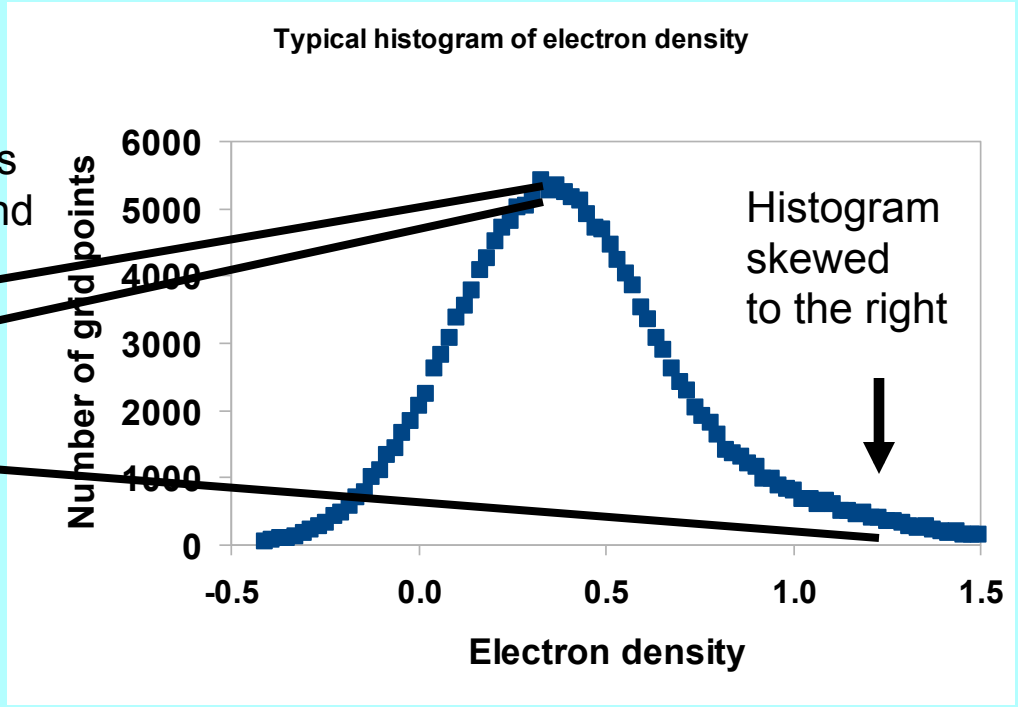
If map is good:

It is easy

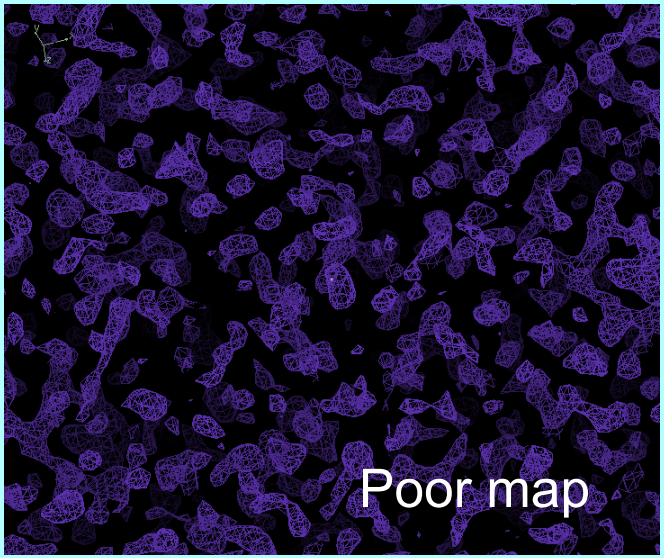


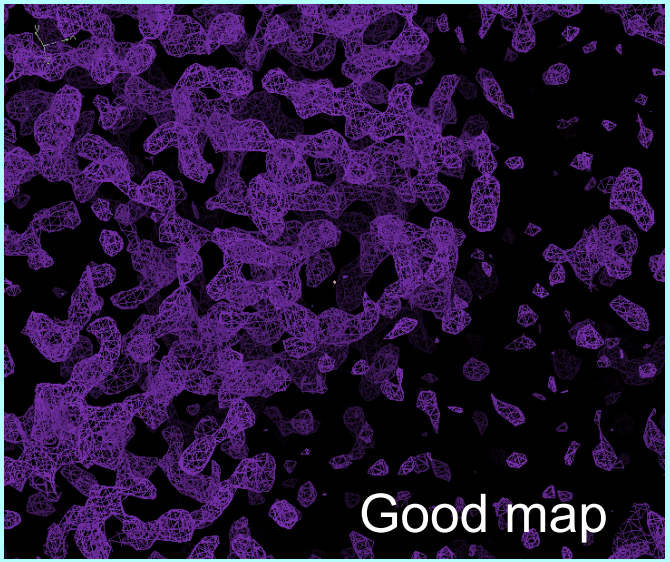
Histogram of electron density values has a positive “skew”

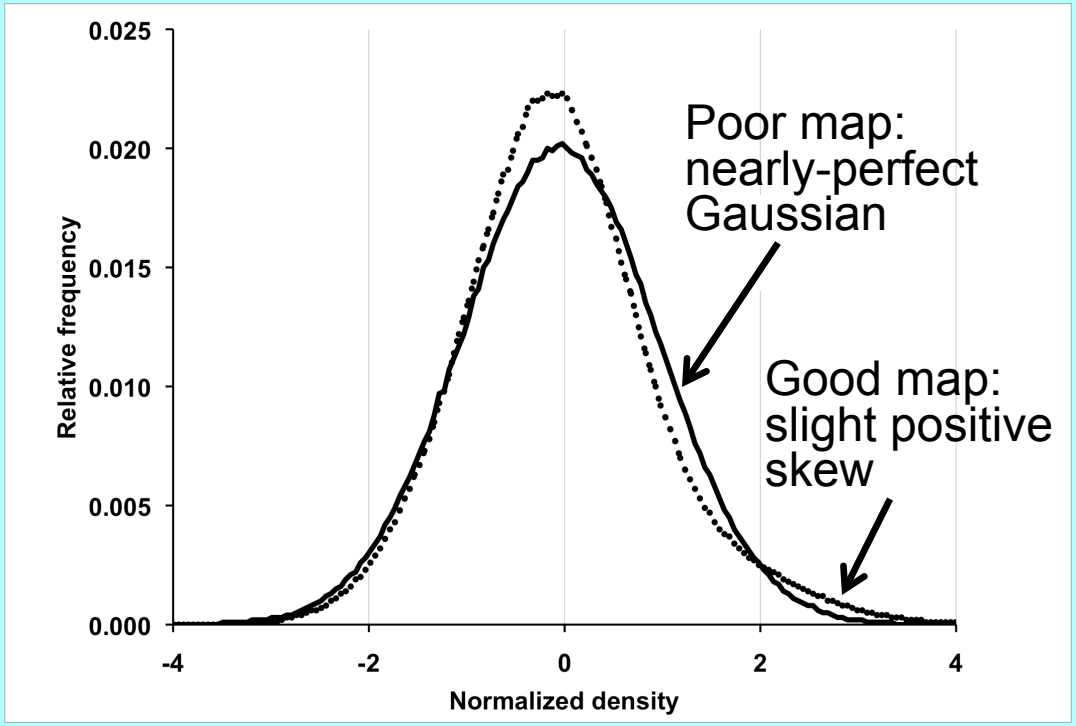




Skew of electron density for poor and good maps







Evaluating electron density maps

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

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)

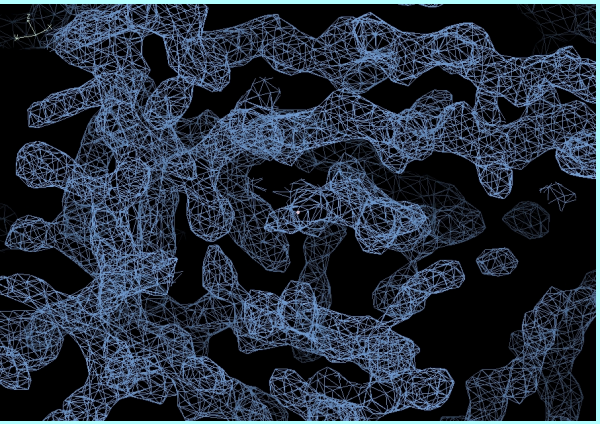
Run AutoSol Wizard on each dataset.

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

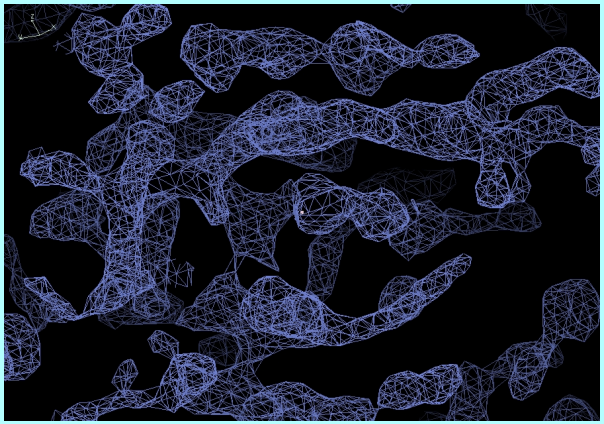
Score maps based on each criteria

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

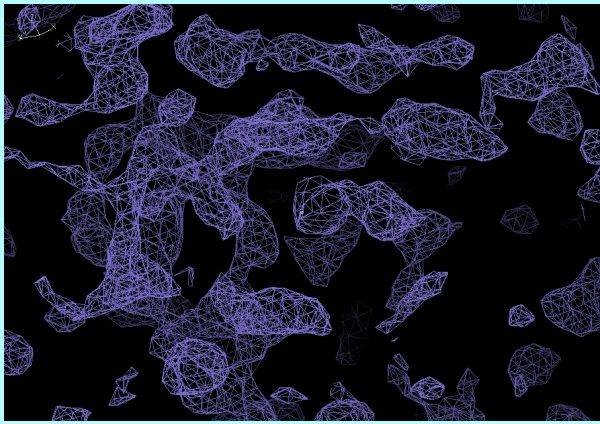
Model map
1VQB, 2.6 Å, SG *C2*



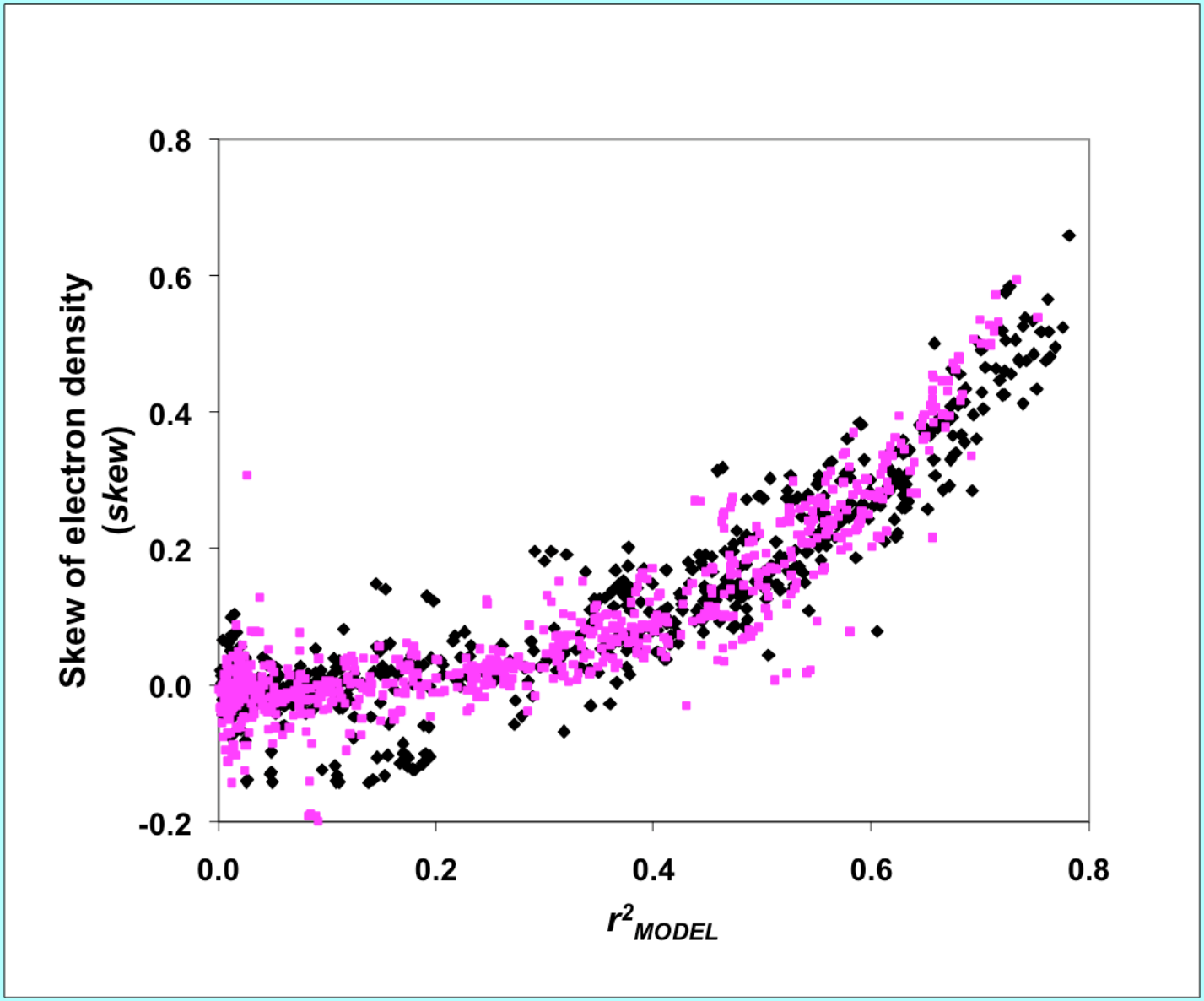
SOLVE MAD map
CC=0.62



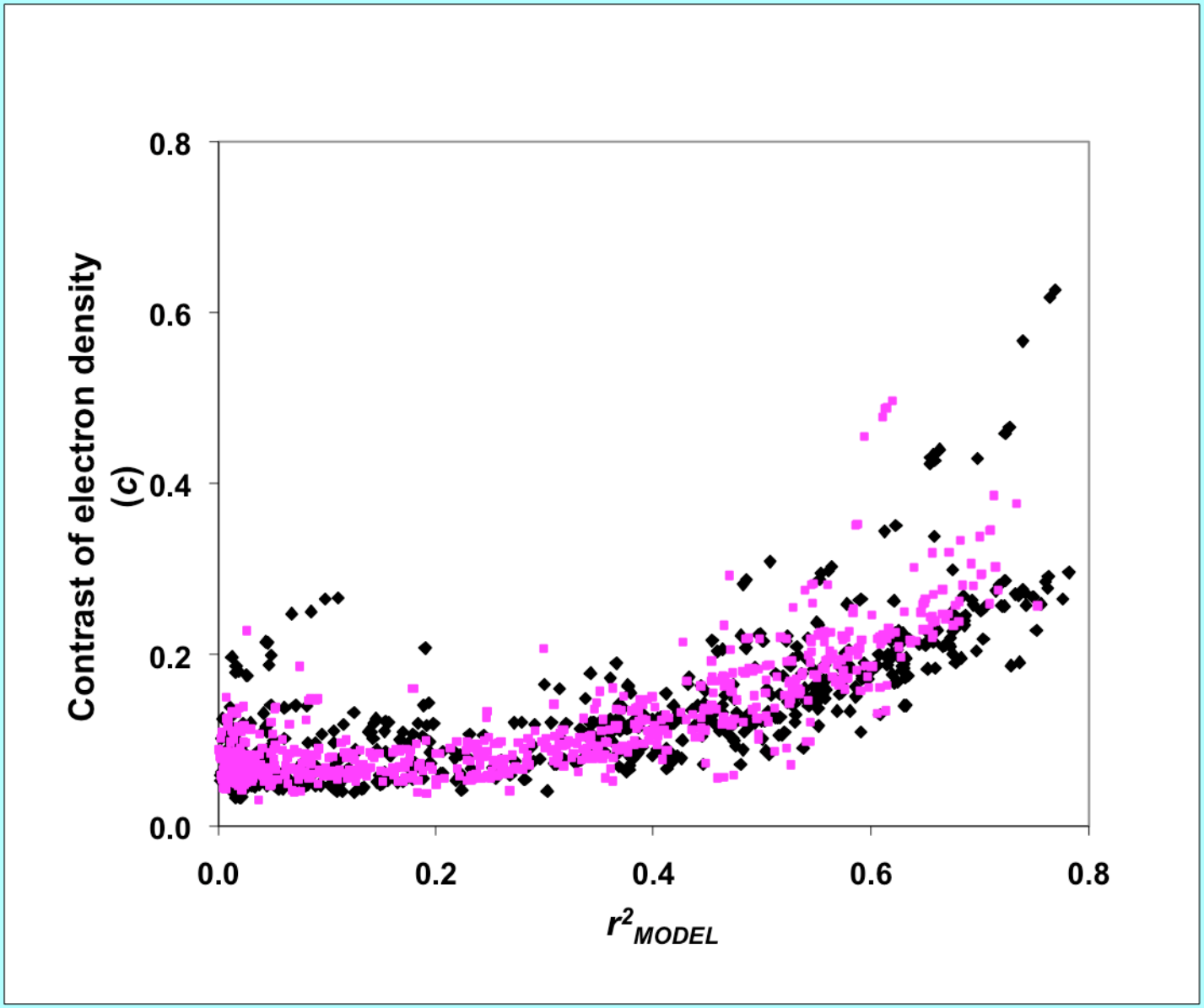
Inverse-hand map
CC=0.55



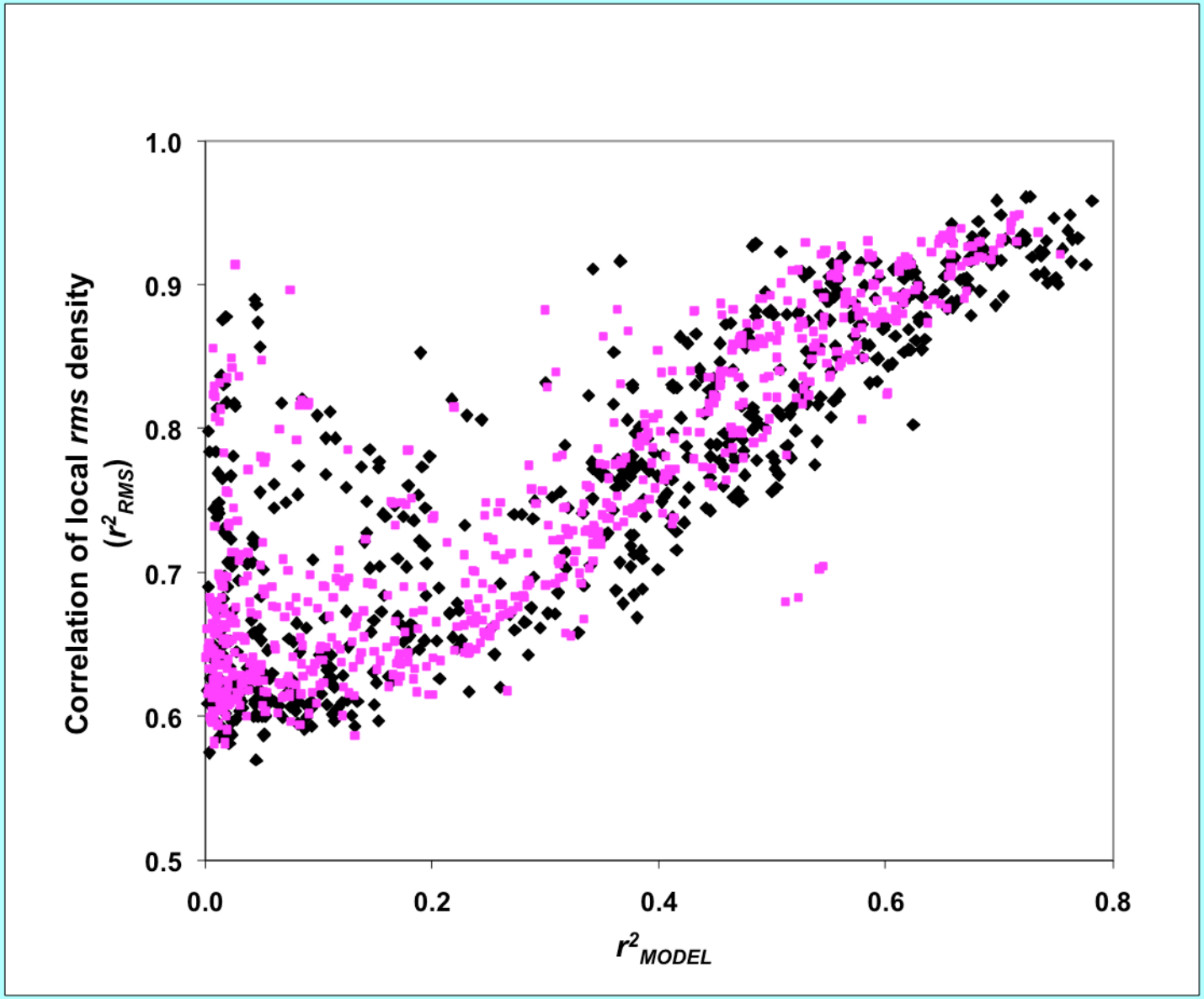
Skew of electron density – positive skew of density values



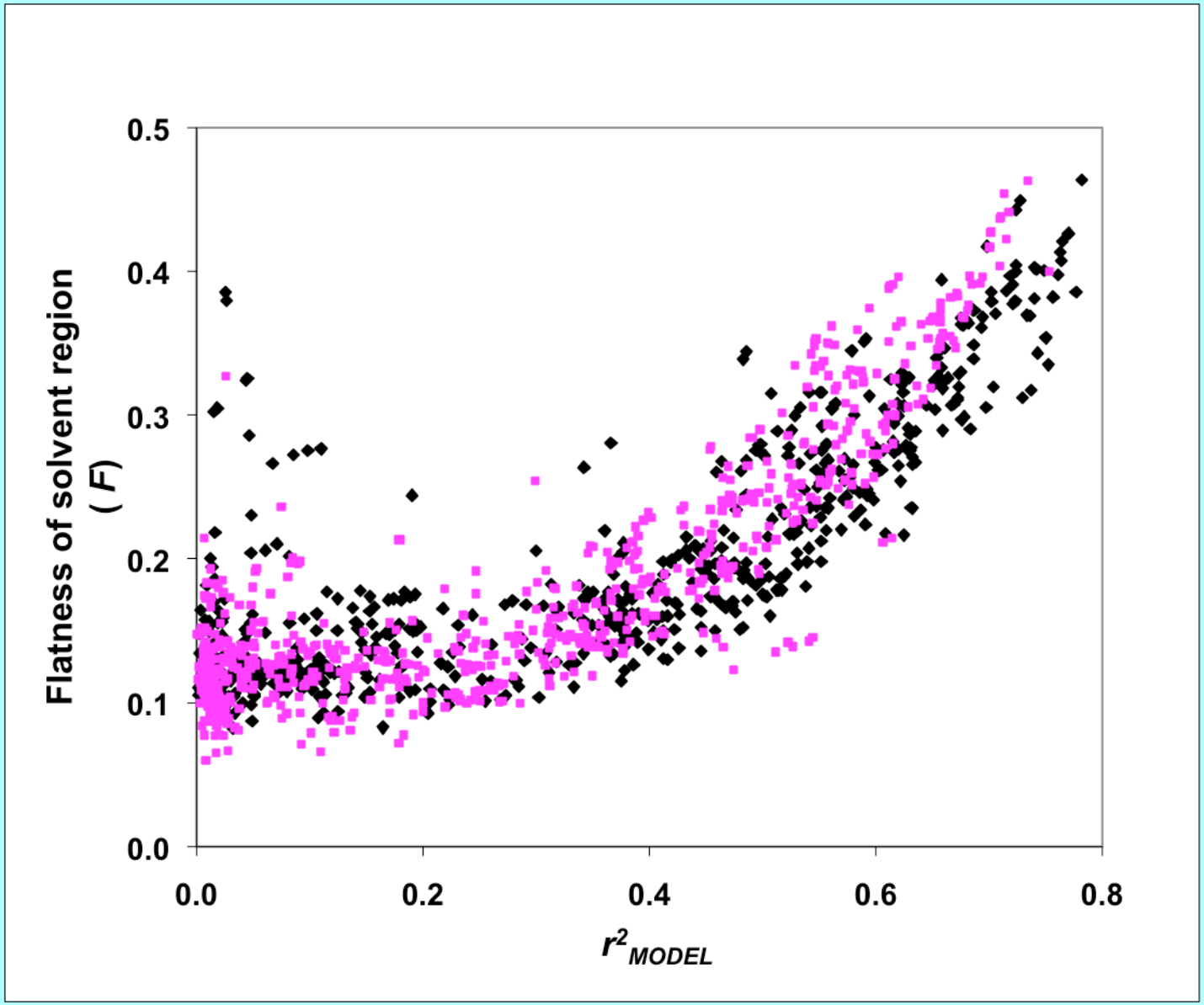
Contrast of density – variability of local rms density



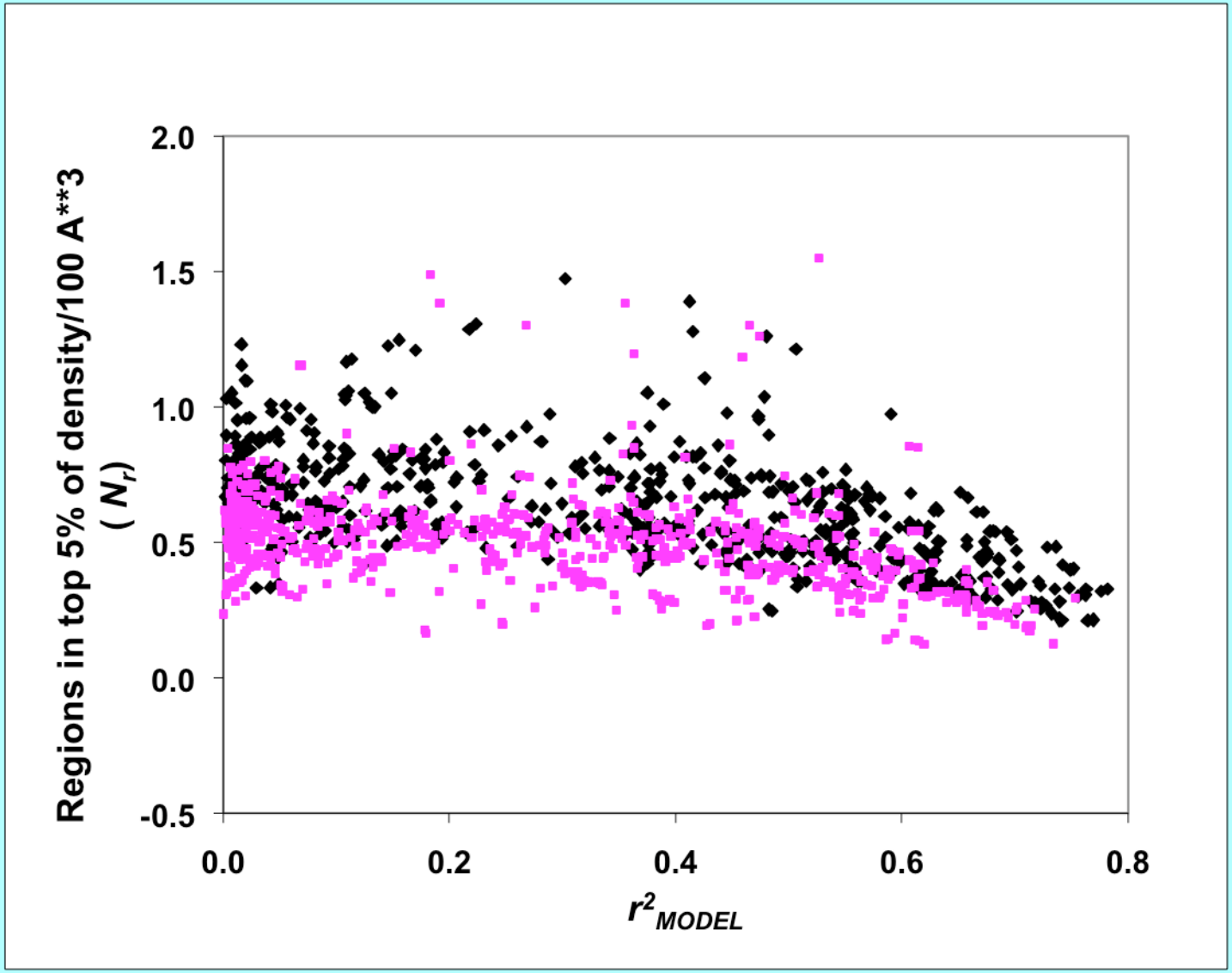
Correlation of local RMS density – large solvent and protein regions



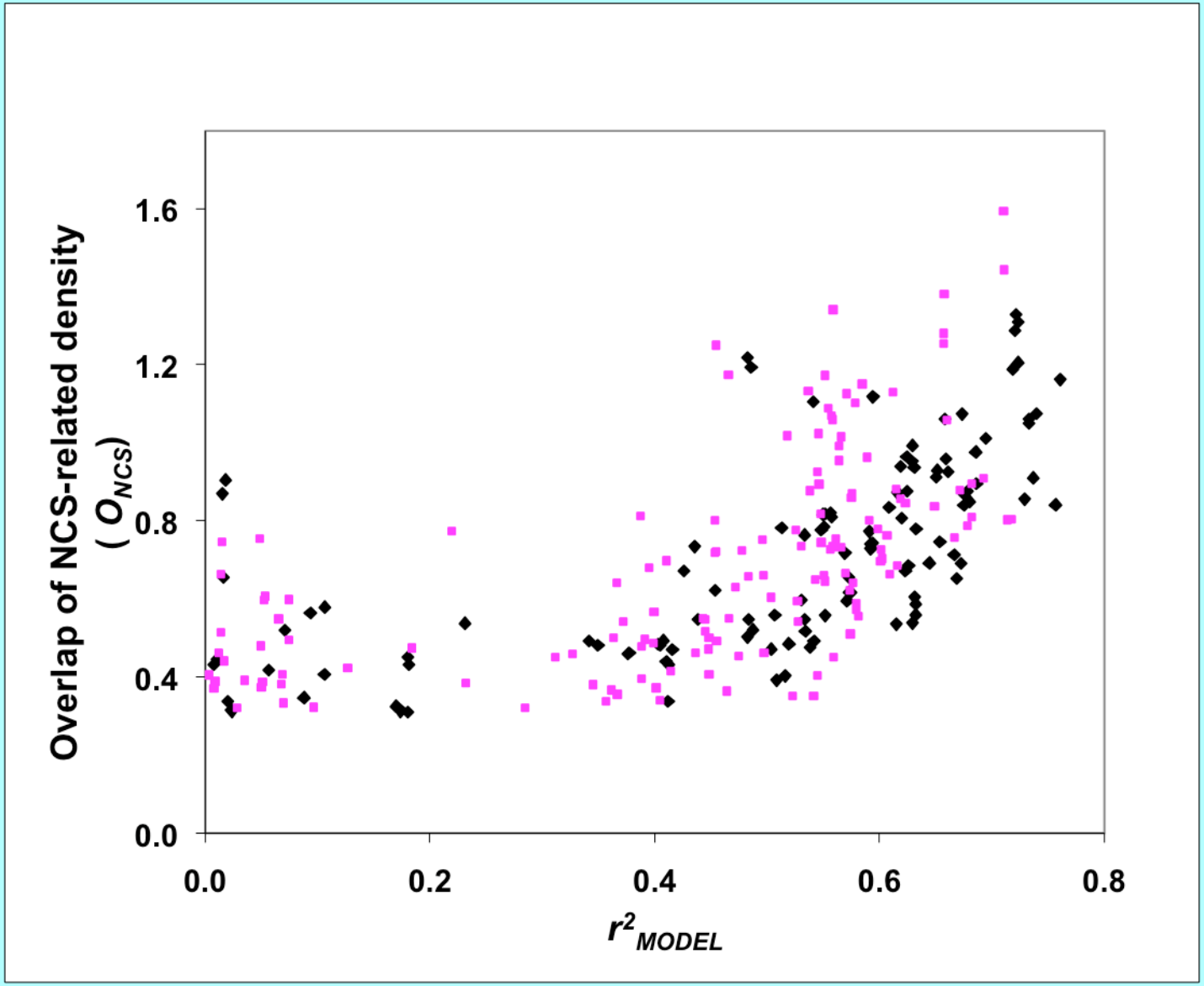
Flatness of solvent region (rms in solvent relative to protein)



Contiguous protein density



NCS in density



Correlation of prime-and-switch density modified phases with experimental phases

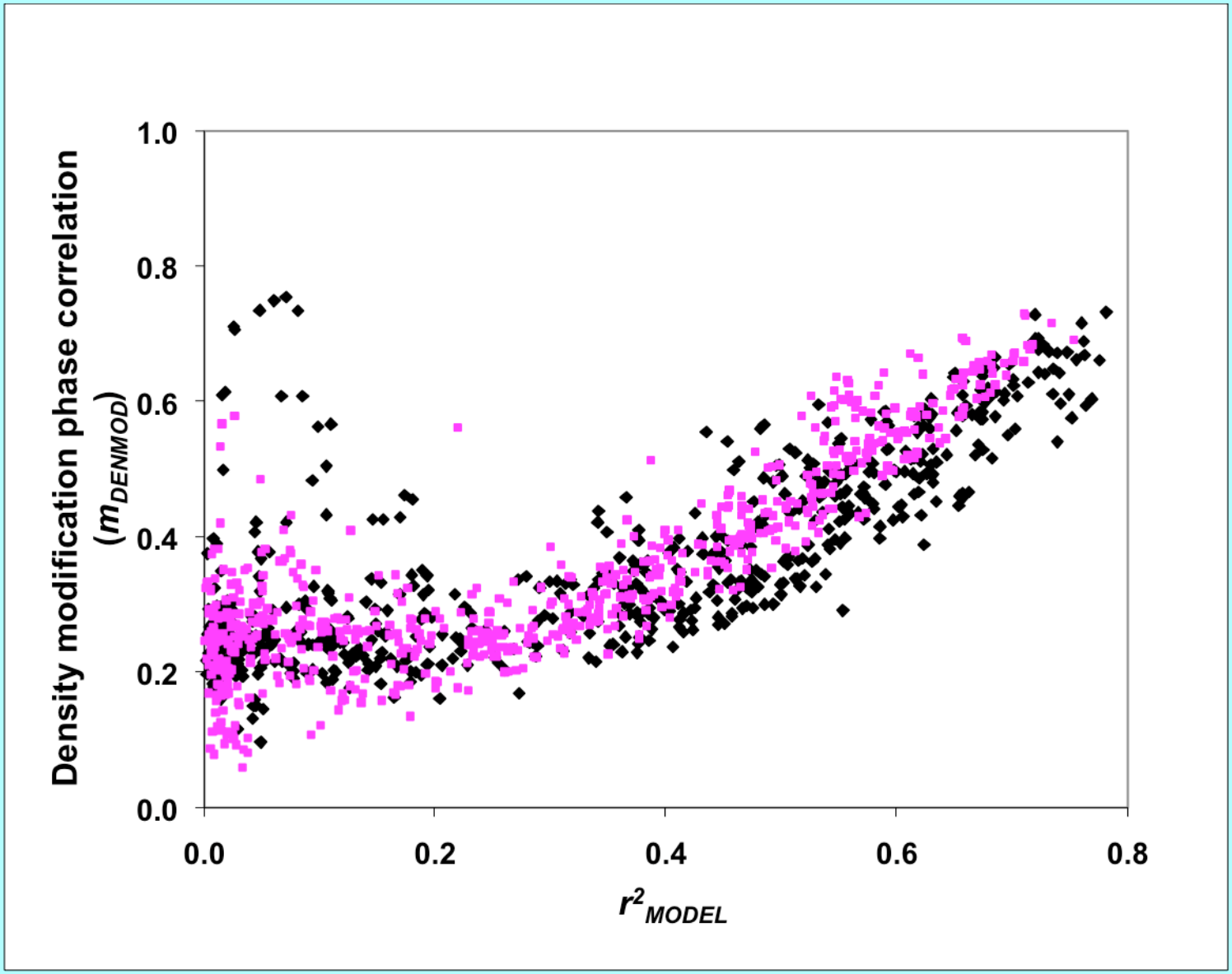
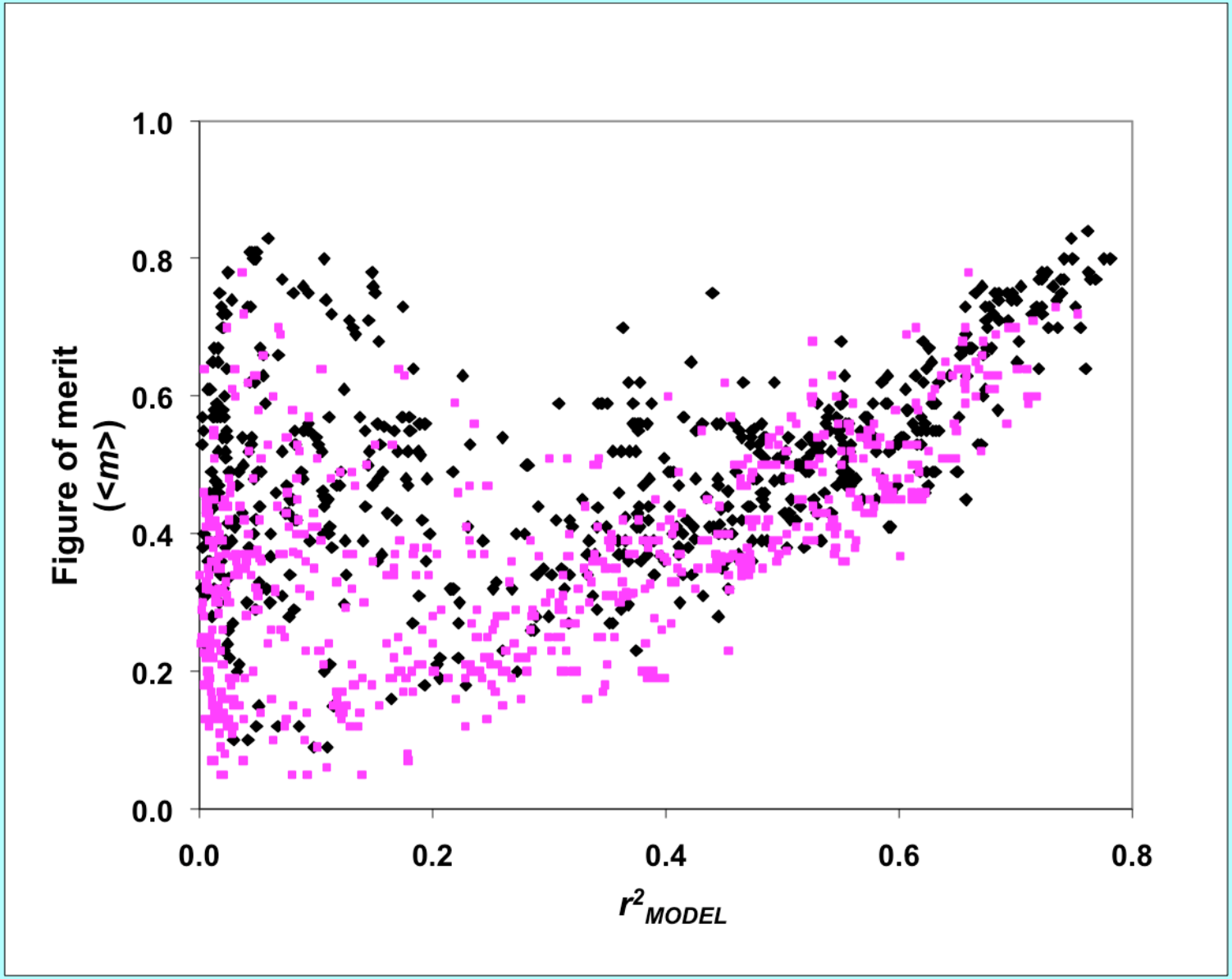
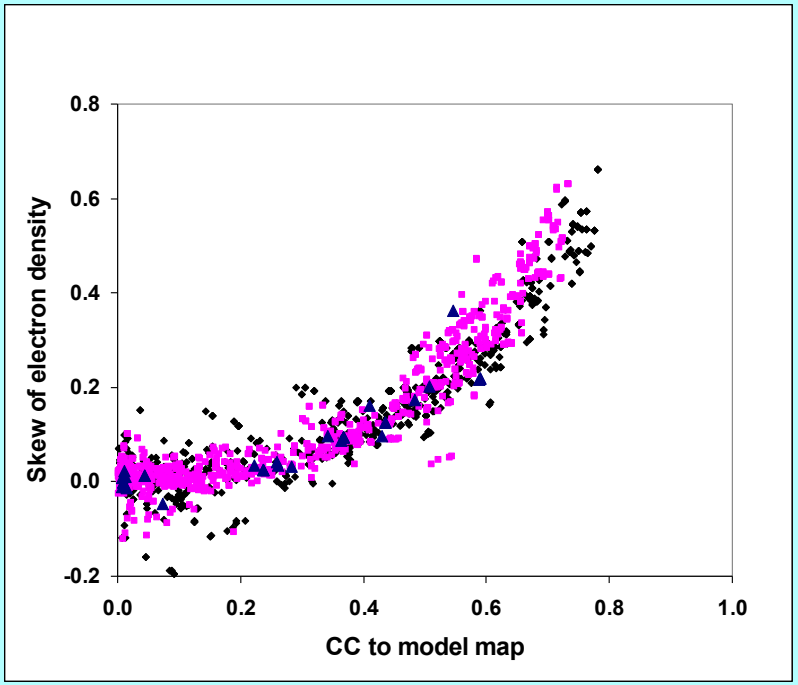


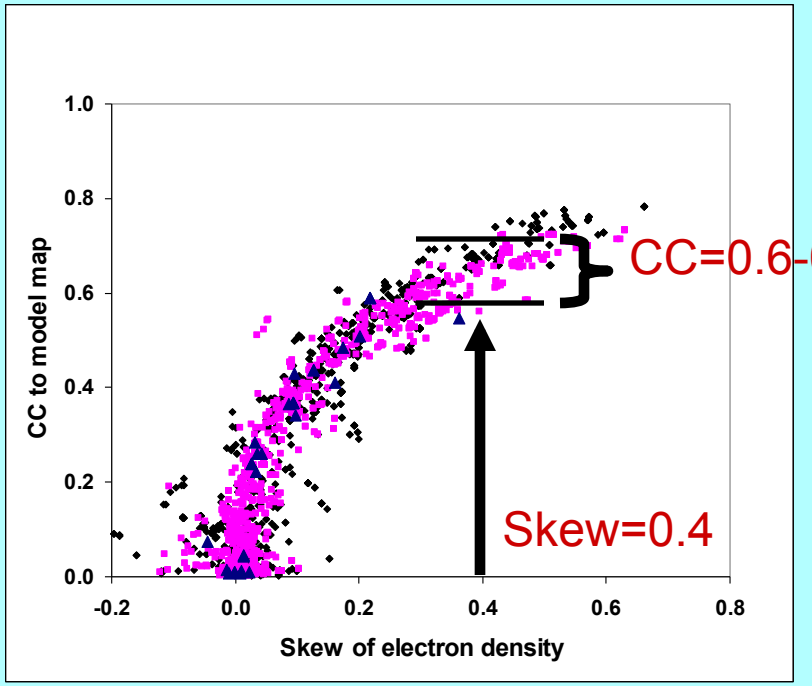
Figure of merit of phasing



Using scoring criteria to estimate the quality of a map

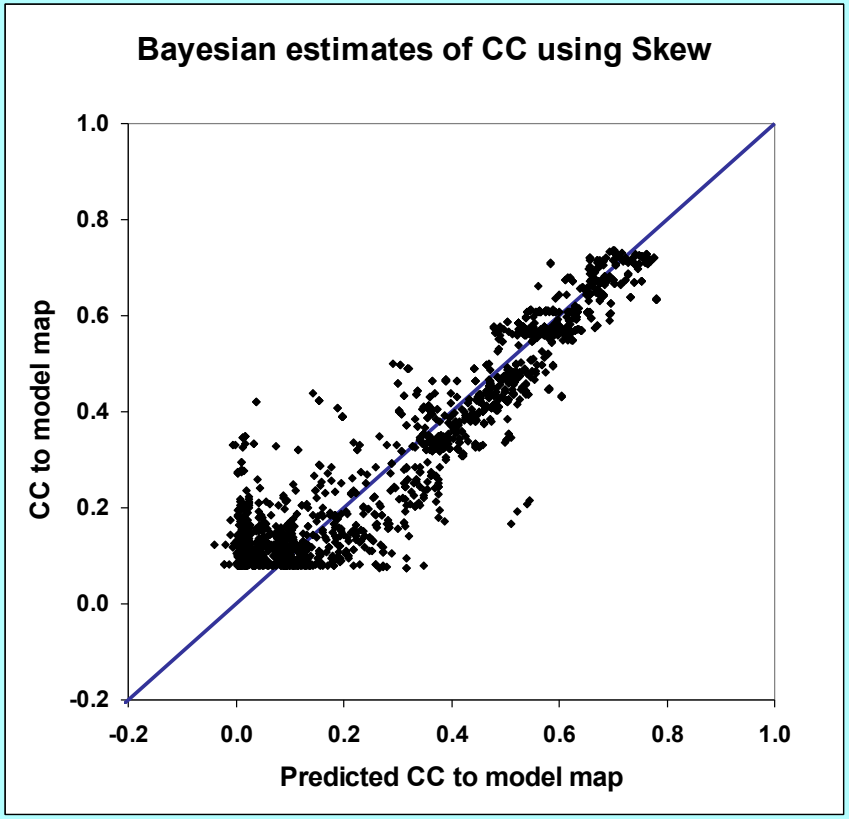
Skew depends on CC → Estimate CC from skew





How accurate are estimates of map quality?

Actual quality



Estimated quality

Cross-validated estimates of quality

Prediction accuracy of measures of experimental map quality

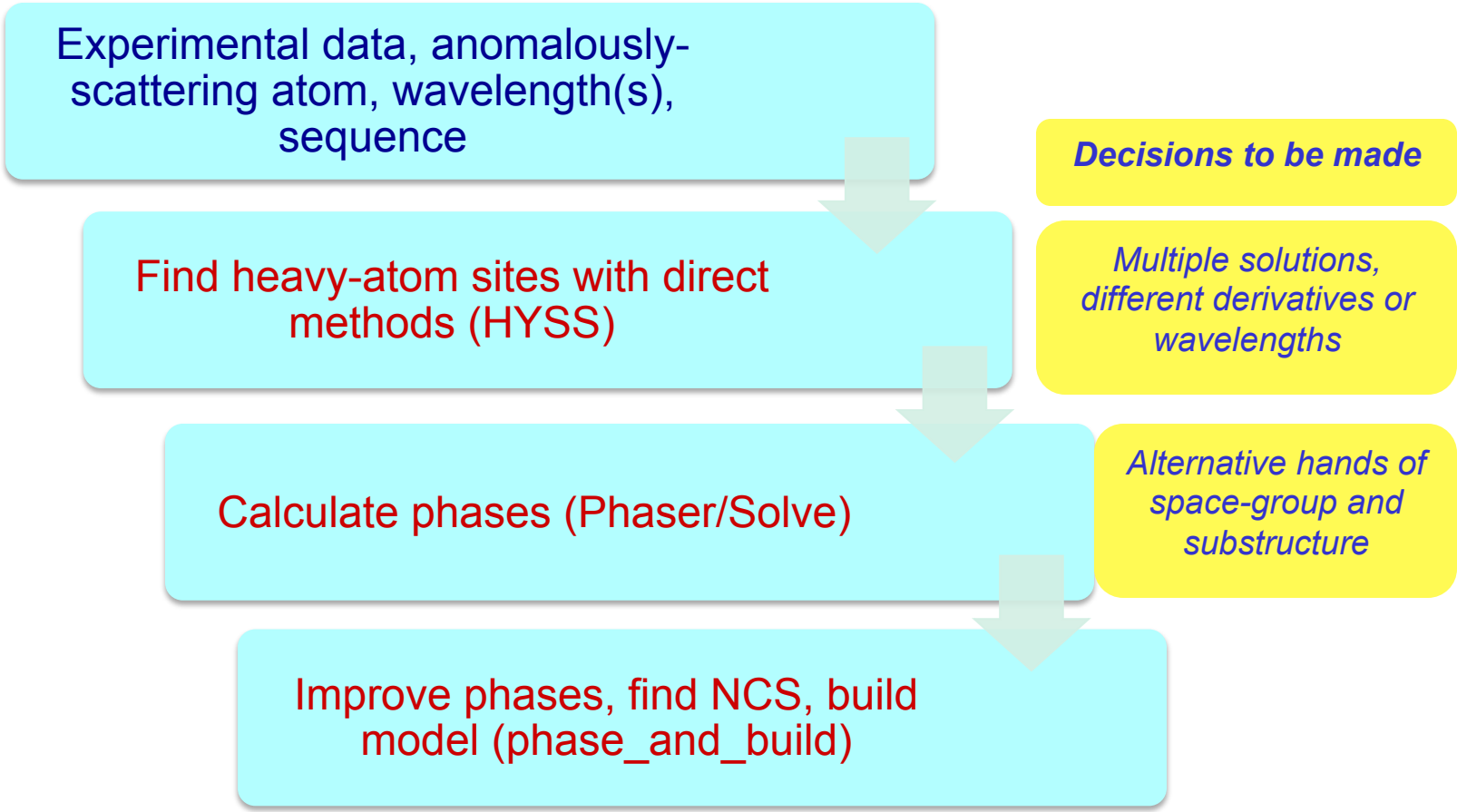
Measure	Quality prediction correlation	Quality prediction error
Skew	0.90	0.10
Correlation of local rms	0.85	0.12
Flatness of solvent	0.80	0.14
Prime-and-switch phase correlation	0.80	0.10
Contrast	0.78	0.15
Density modification R	0.77	0.14
Contiguous density	0.42	0.20
Figure of merit of phasing	0.42	0.21

Estimated map quality in practice

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

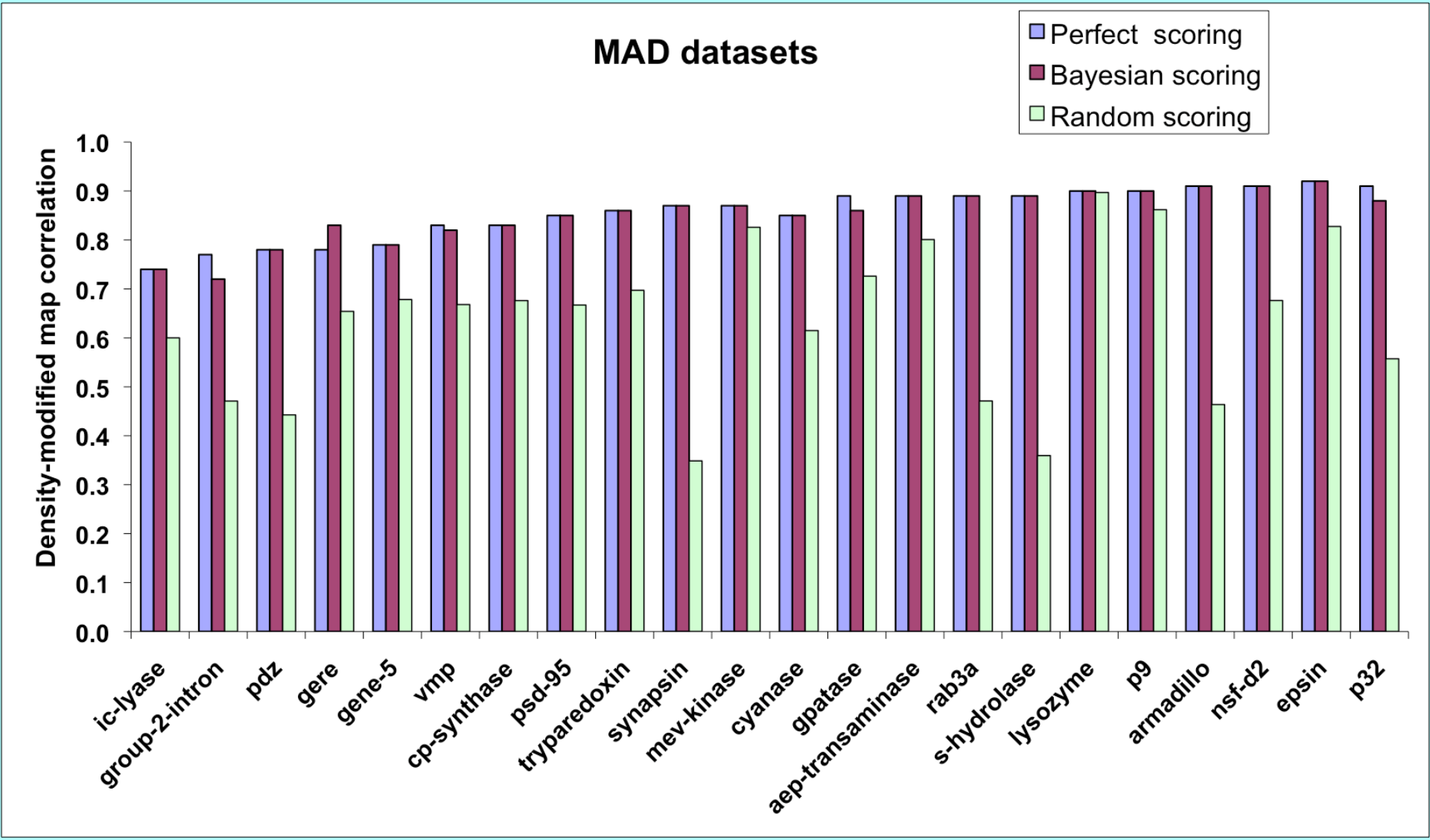
Data for HYSS	Sites	Estimated CC ± 2SD	Actual CC	
Peak	12	0.73 ± 0.04	0.72	←
Peak (inverse hand)	12	0.11 ± 0.43	0.04	
F_A	12	0.73 ± 0.03	0.72	
F_A (inverse)	12	0.11 ± 0.42	0.04	
Sites from diff Fourier	9	0.70 ± 0.17	0.69	

Structure solution with *phenix.autosol*



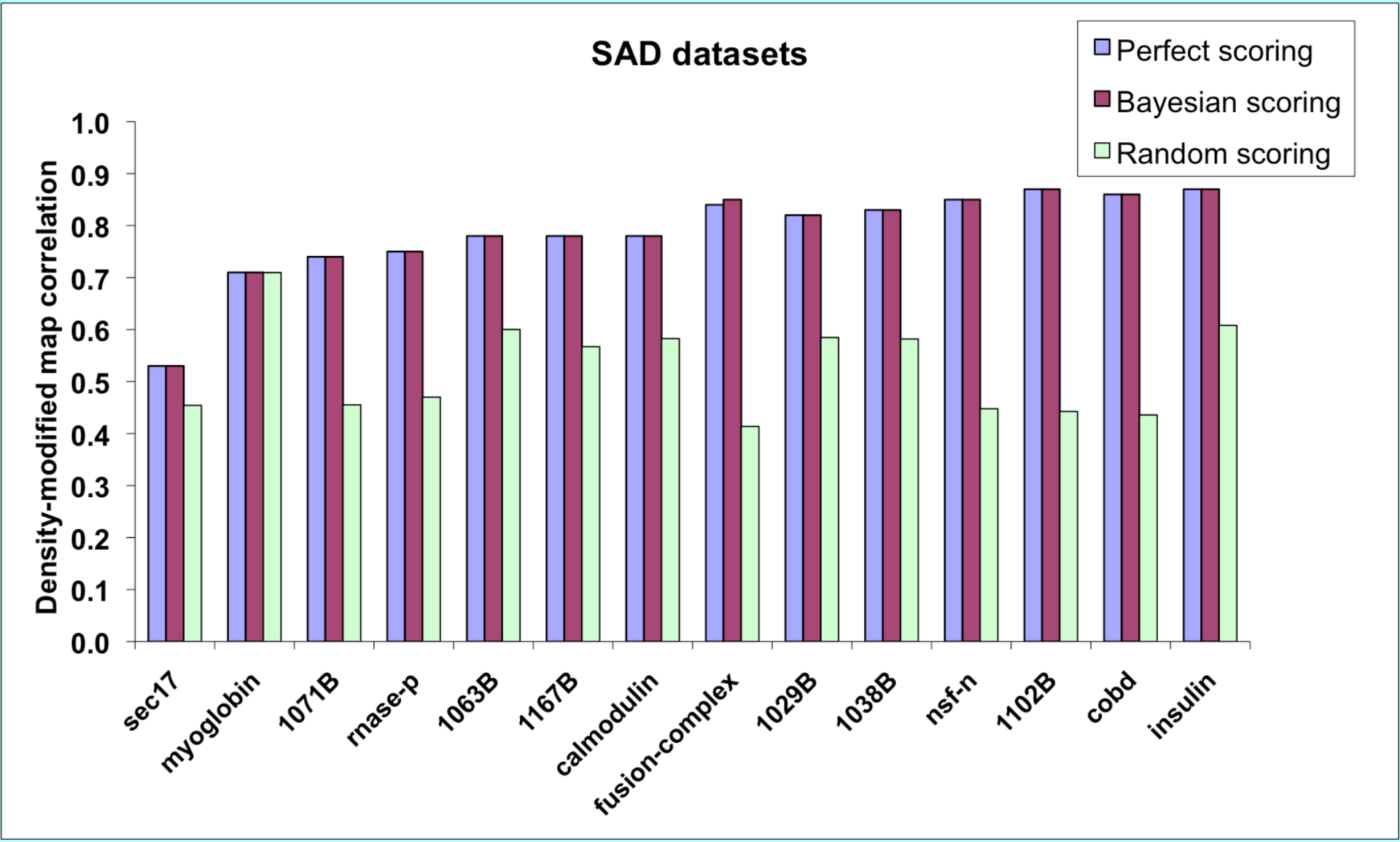
Scoring makes a difference in AutoSol automated structure solution

Perfect scoring *vs* Bayesian scoring *vs* random scoring (mean of 10 runs)



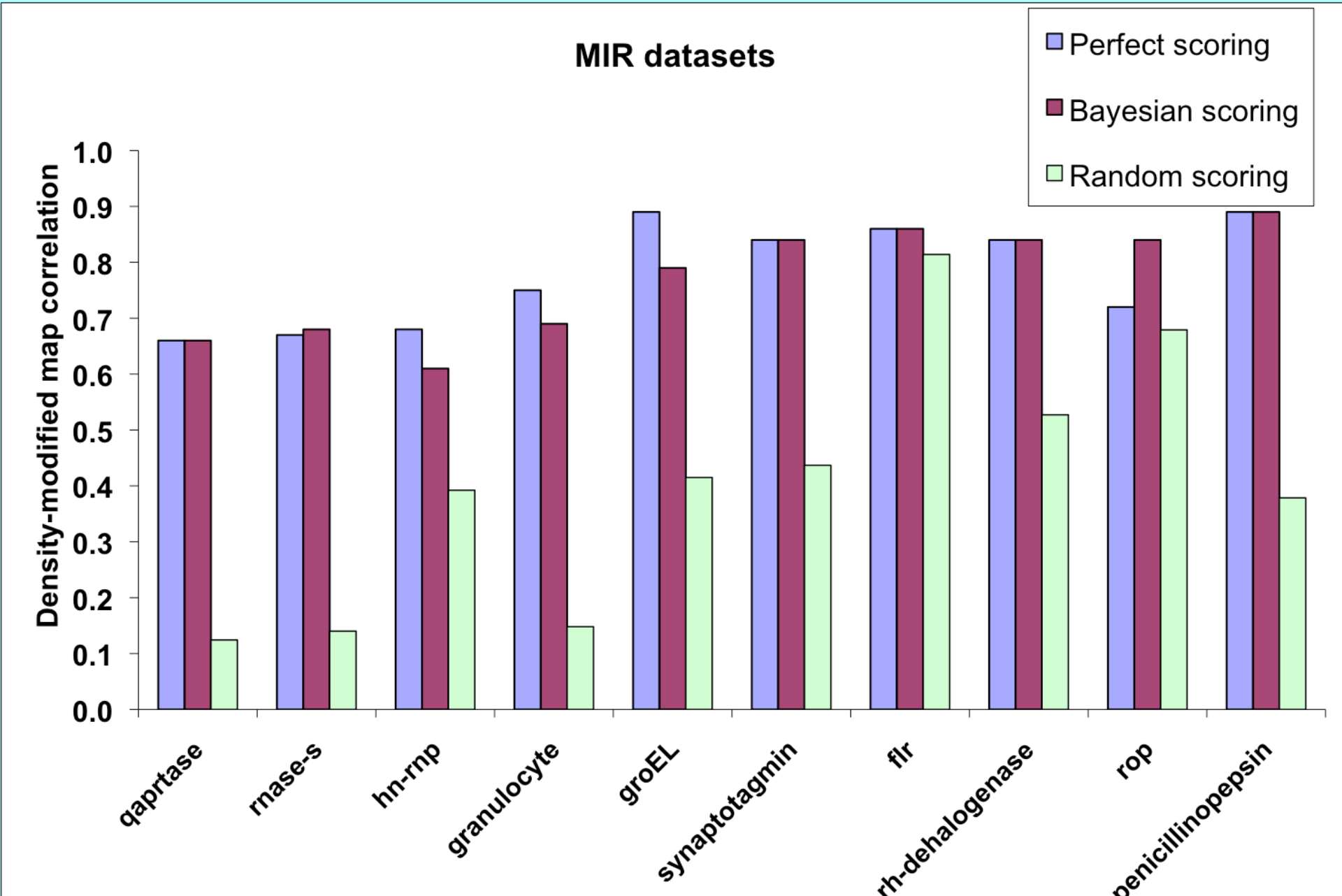
Scoring makes a difference in AutoSol automated structure solution

Perfect scoring *vs* Bayesian scoring *vs* random scoring (mean of 10 runs)

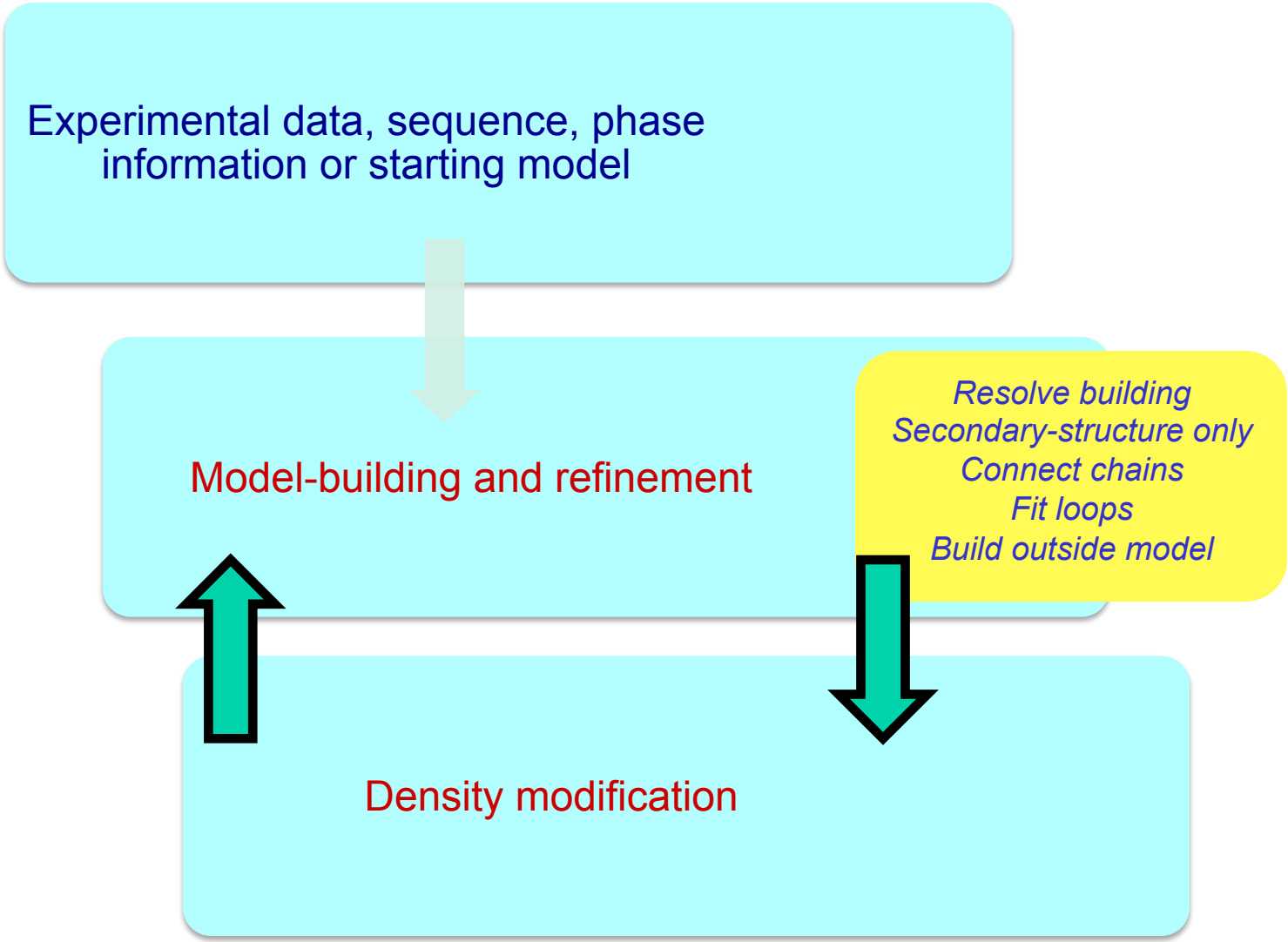


Scoring makes a difference in AutoSol automated structure solution

Perfect scoring *vs* Bayesian scoring *vs* random scoring (mean of 10 runs)

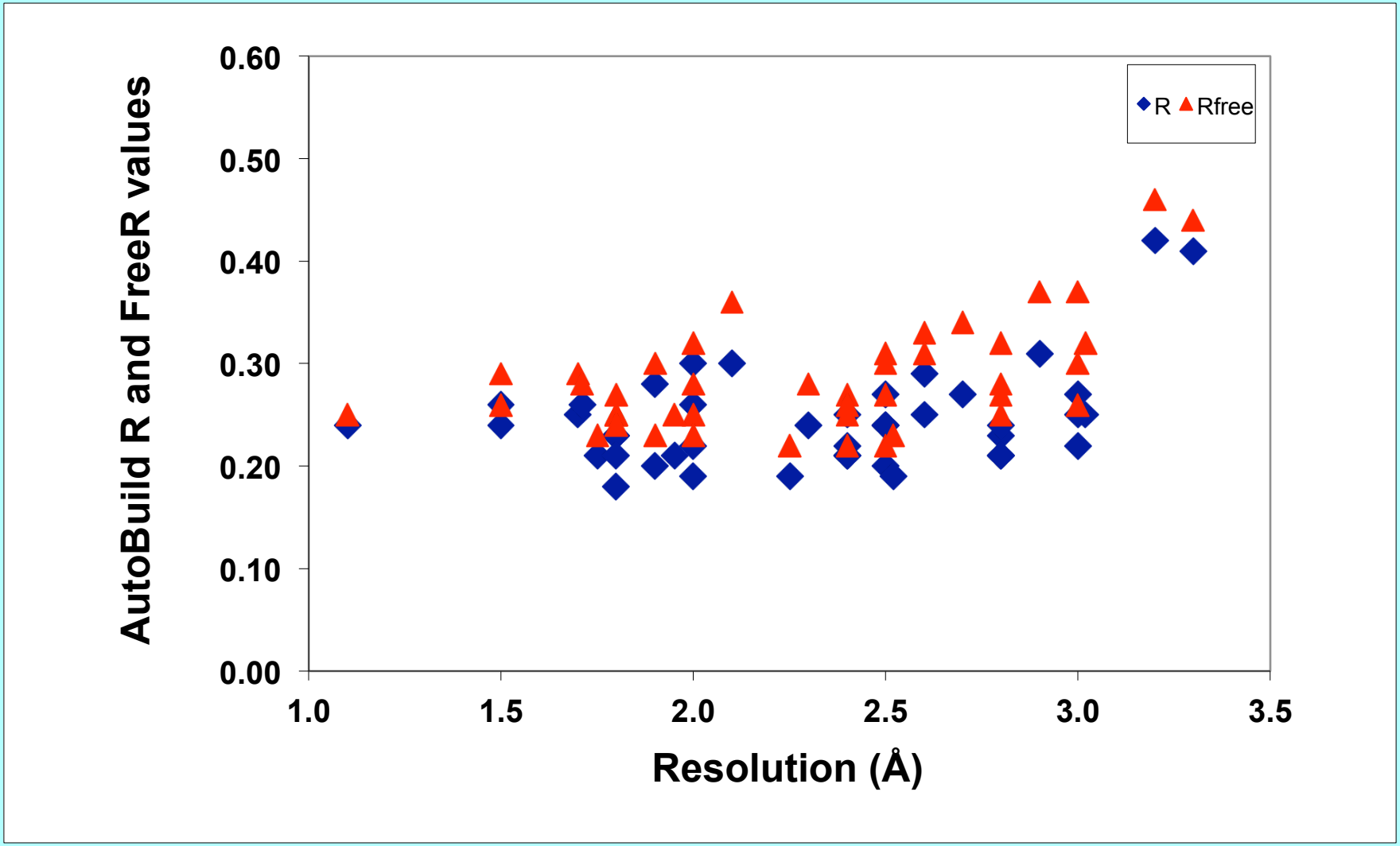


Iterative density modification, model-building and refinement with *phenix.autobuild*



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

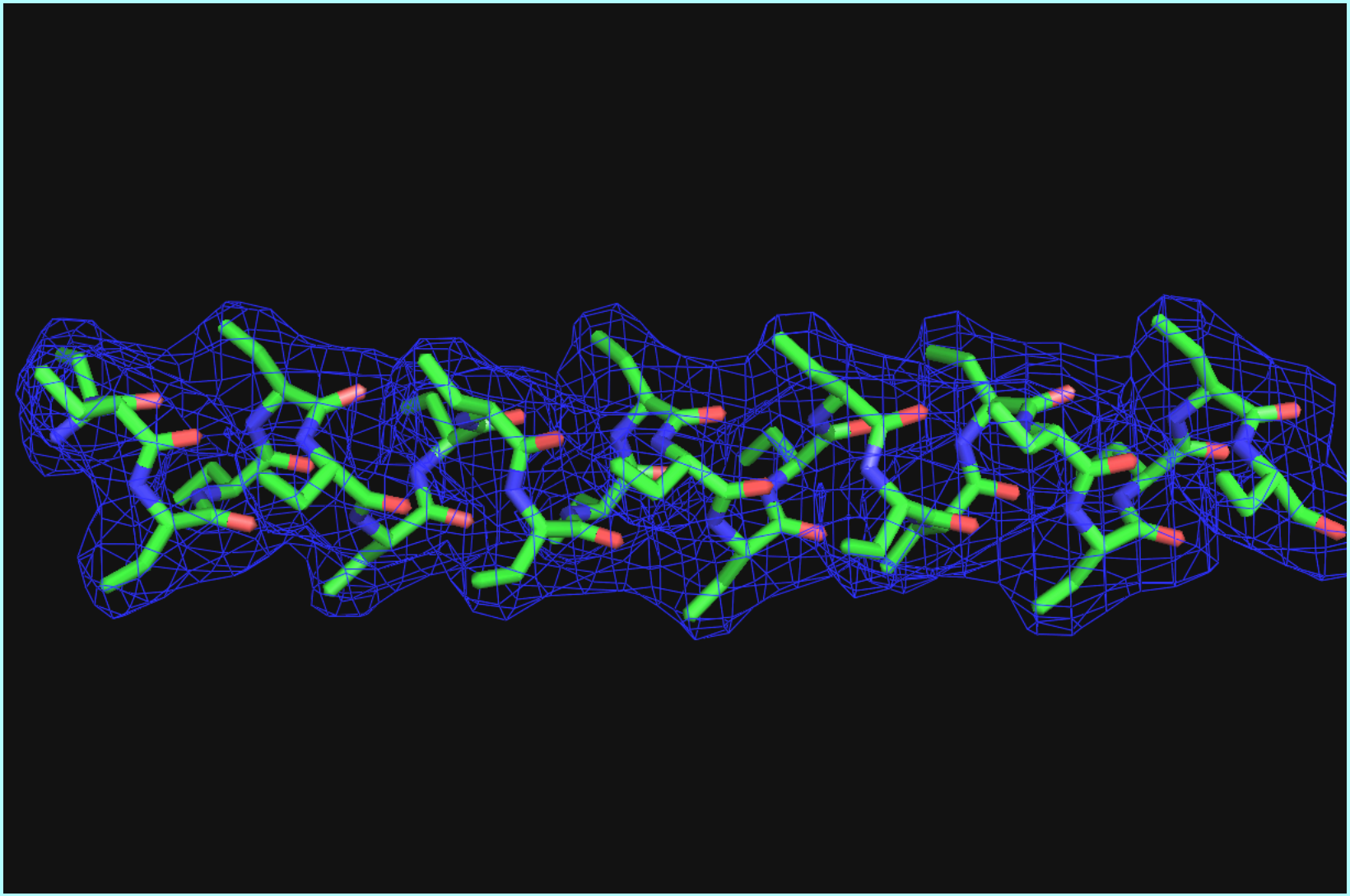
Any protein chains (trace_chain):

Identification: Cα positions consistent with density and geometry of protein chains

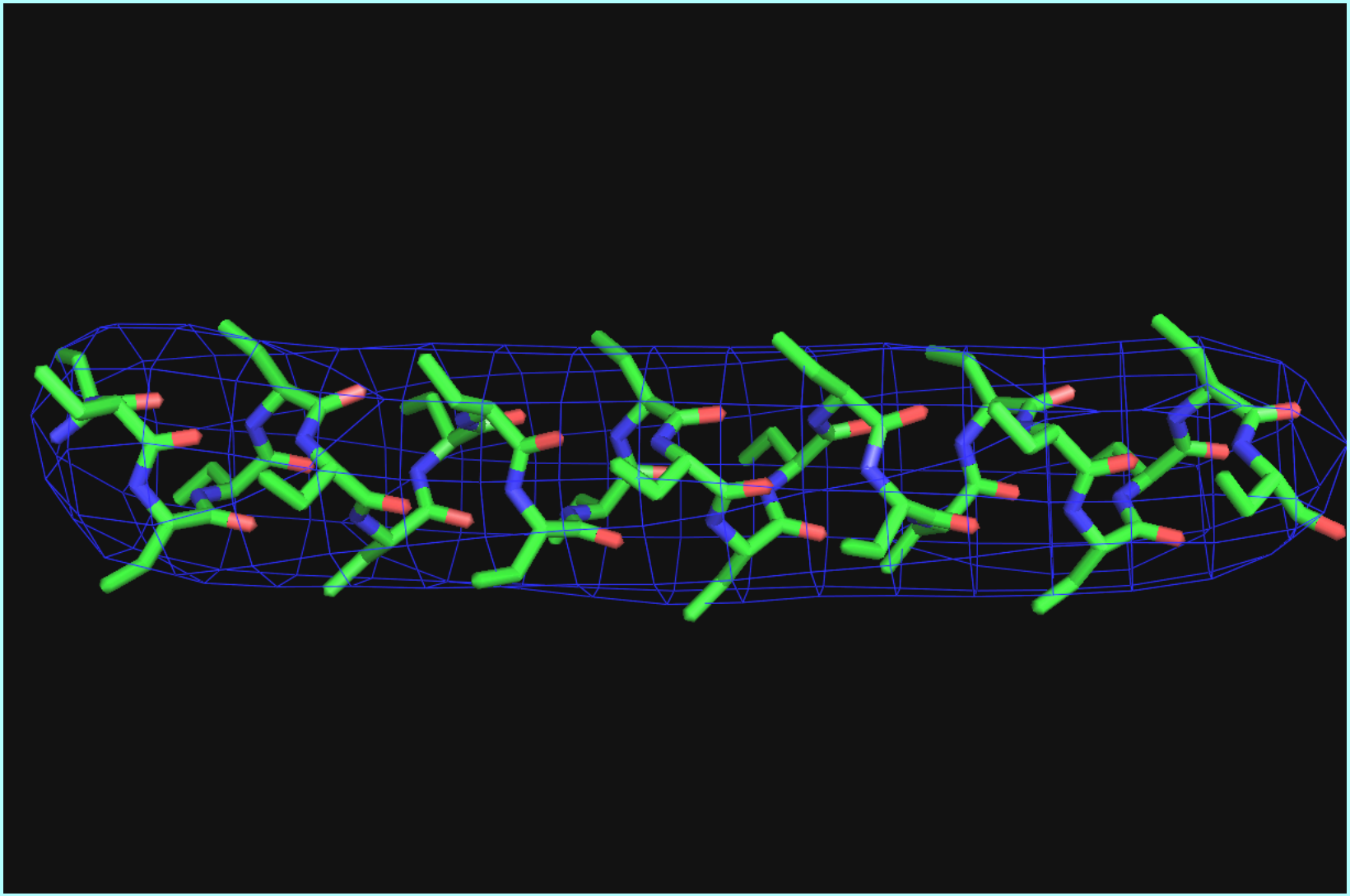
RNA/DNA:

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

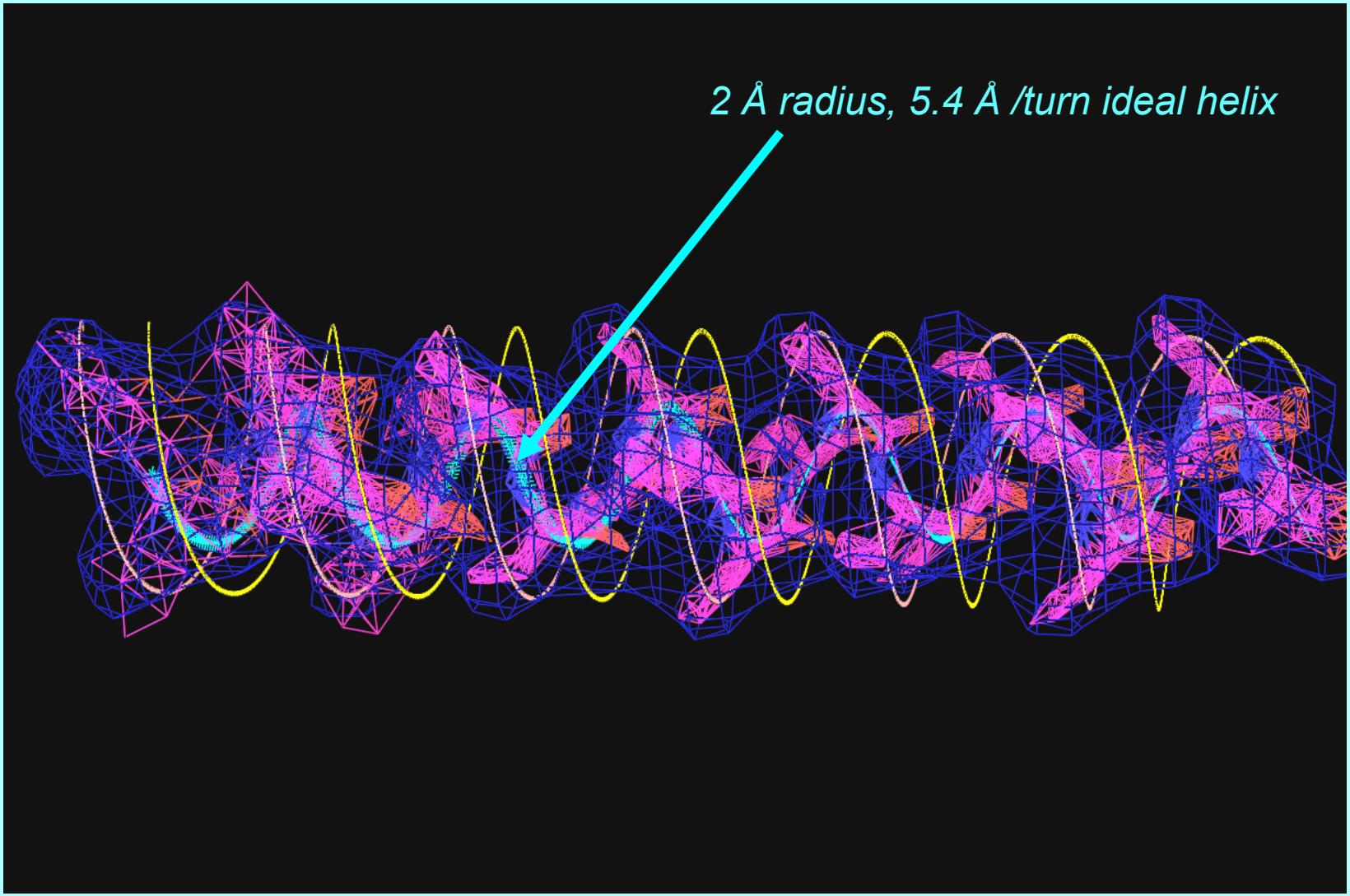
Model α-helix; 3 Å map



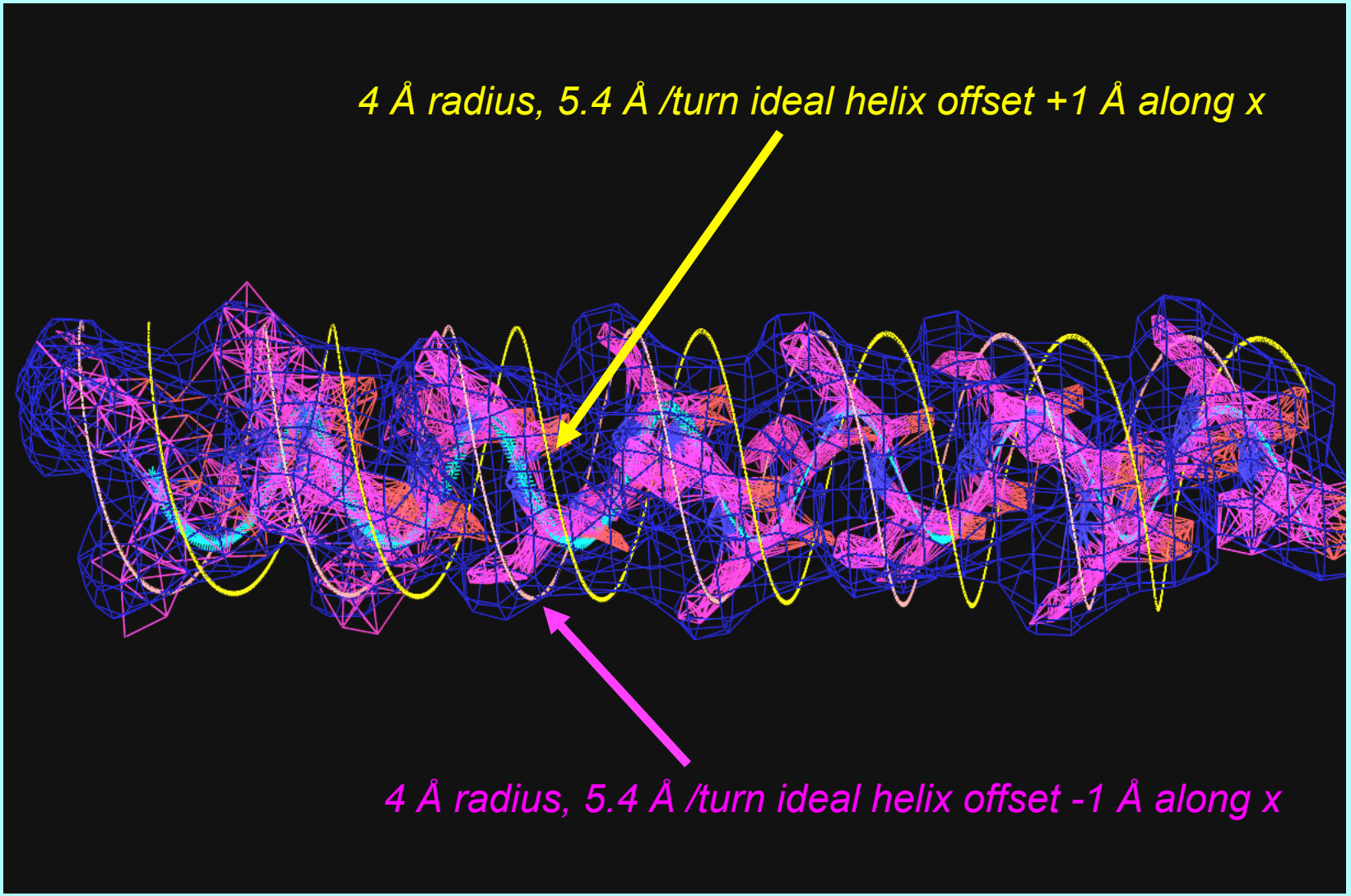
Model α-helix; 7 Å map



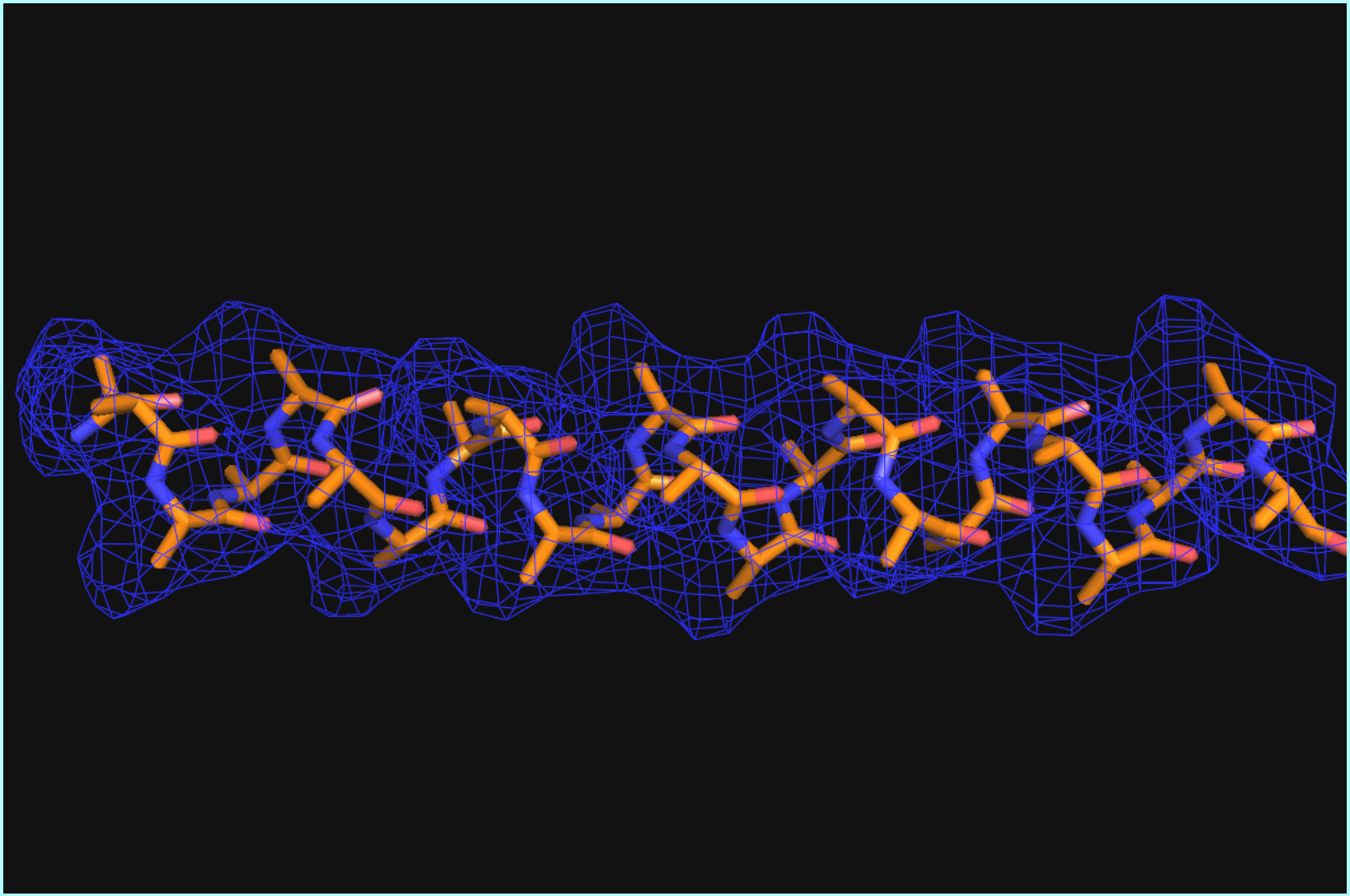
Trace main-chain with ideal helix, allowing curvature



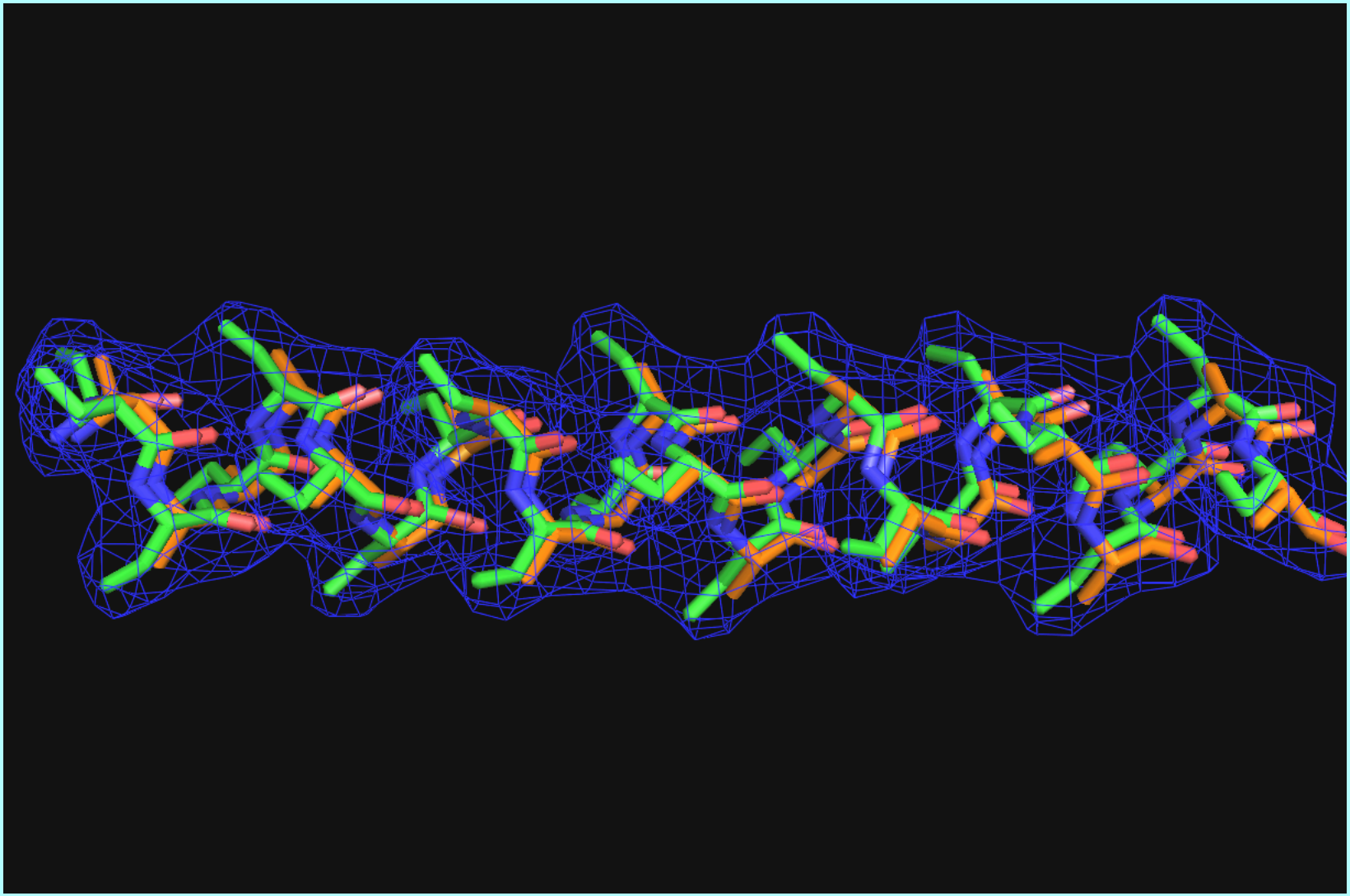
Identify direction and Cα position from overlap with 4 Å radius helices offset +/- 1 Å from main-chain



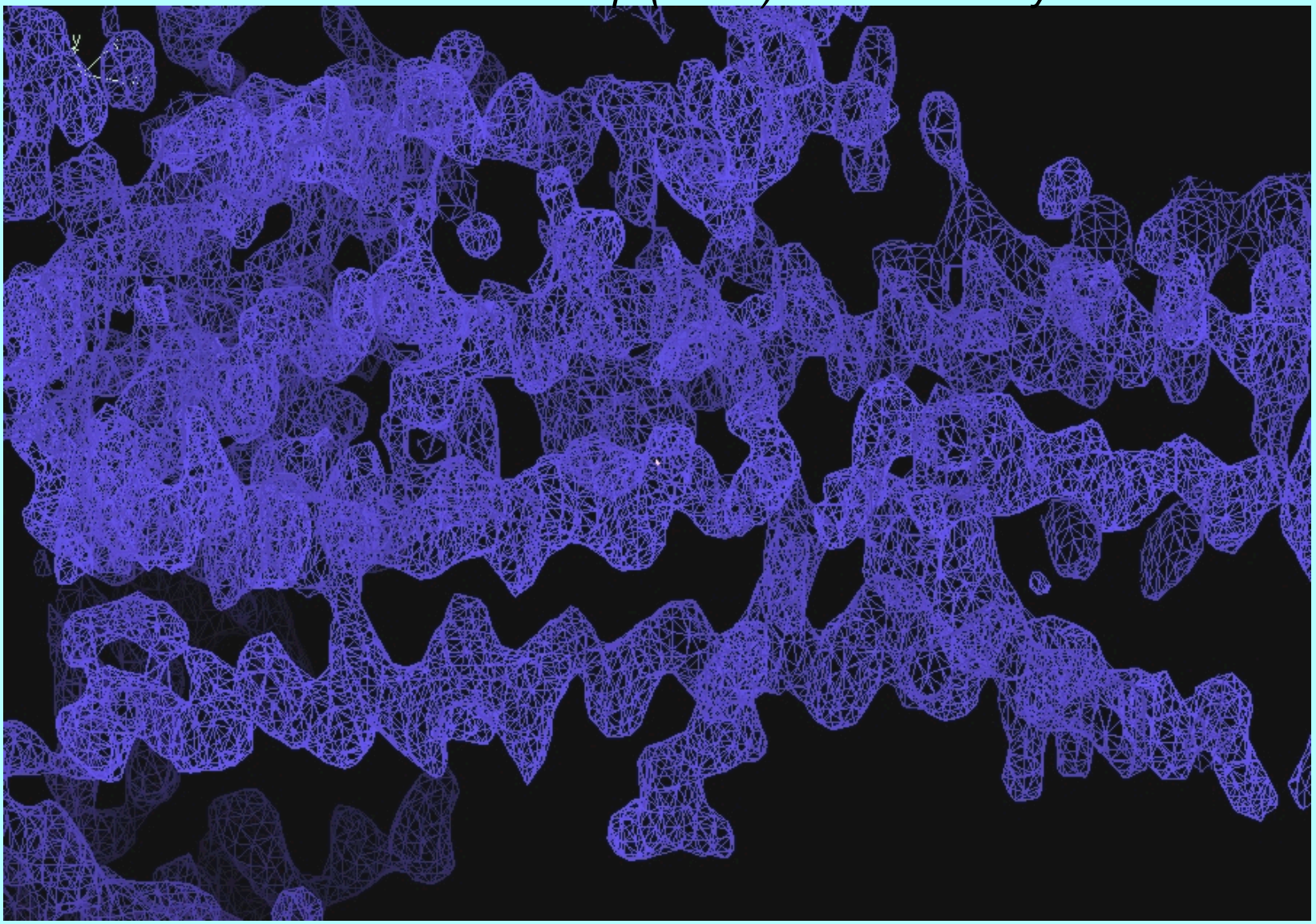
Choose best-fitting helices; link together if necessary



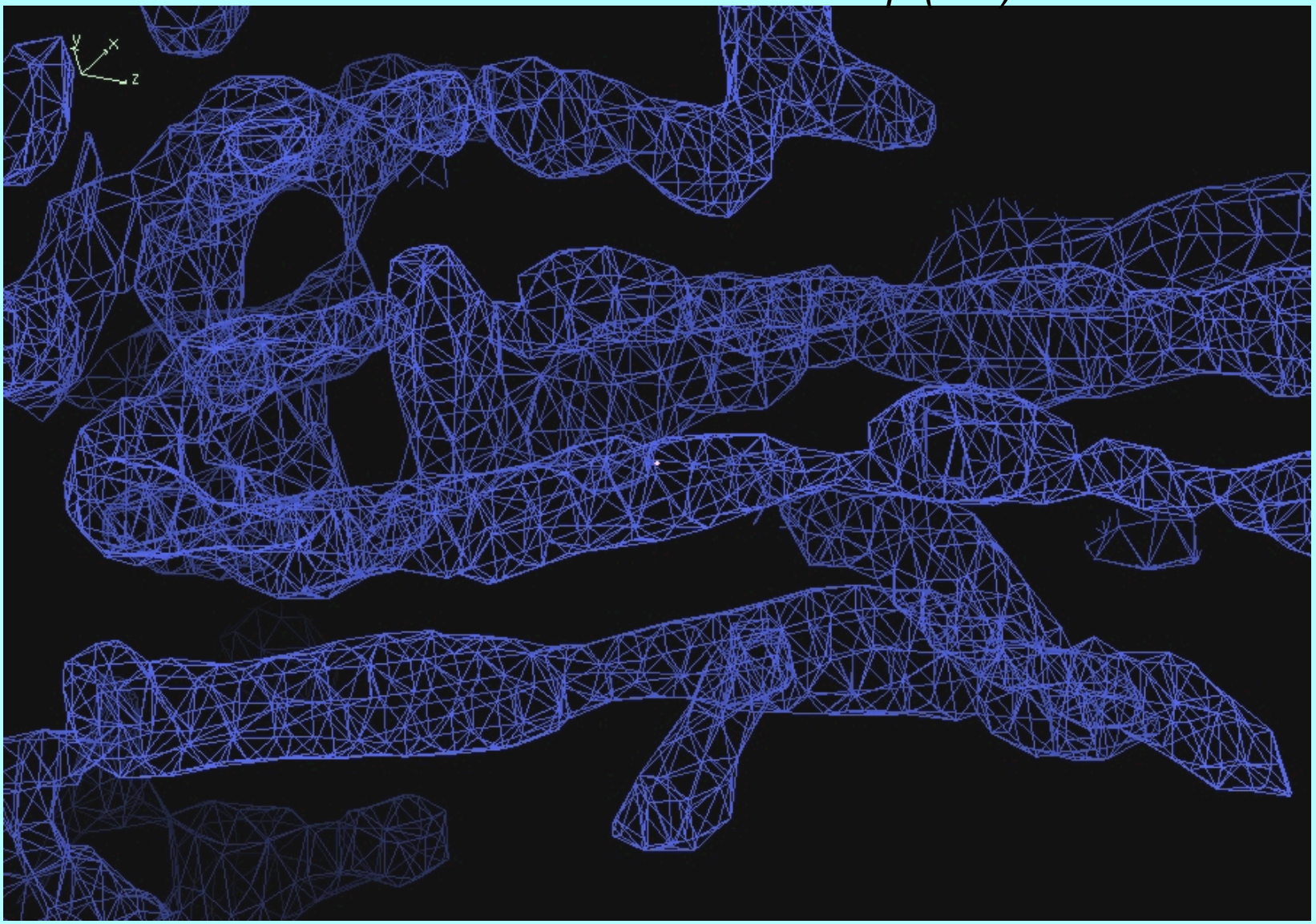
Comparison with model helix



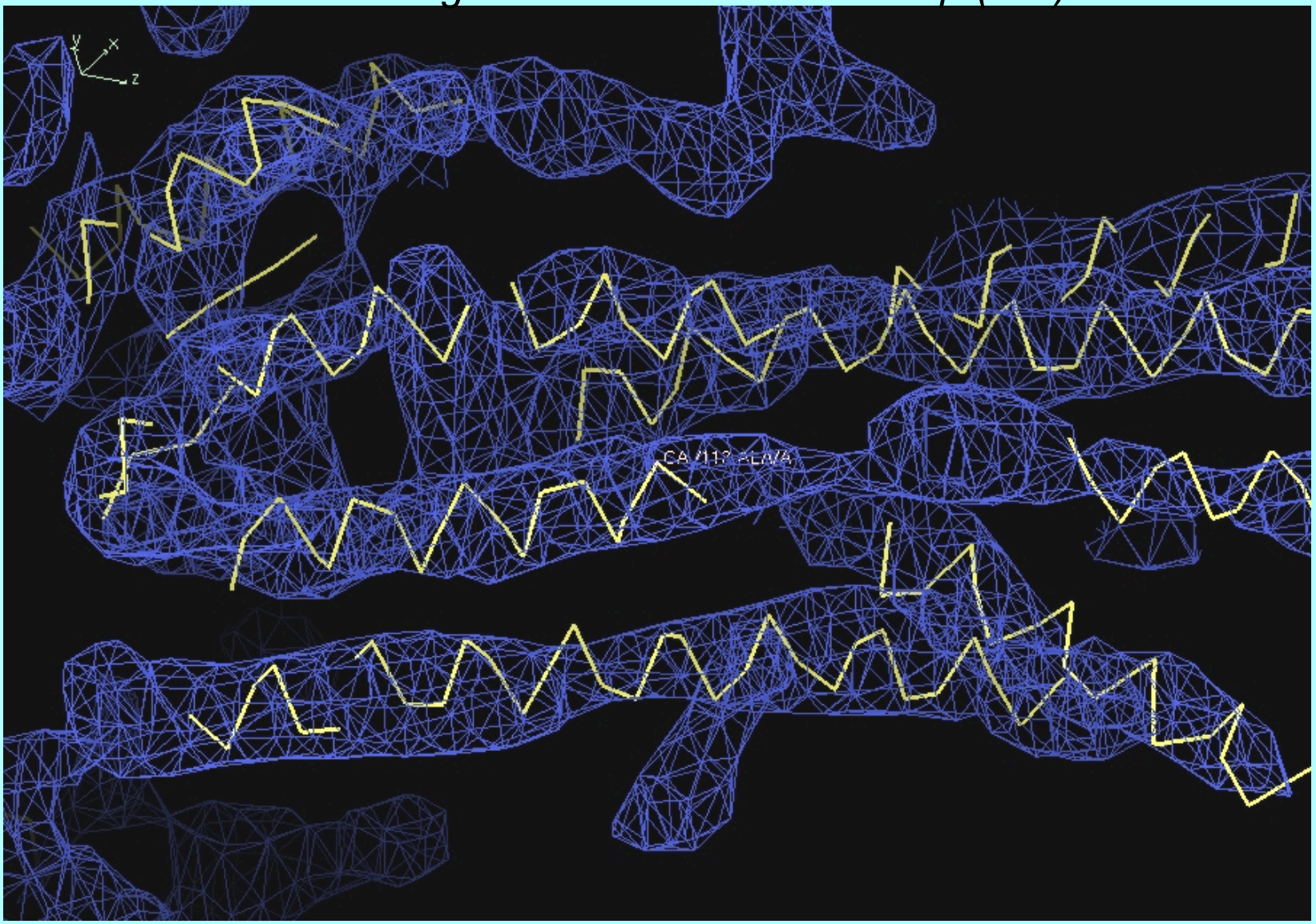
A real case: 1T5S SAD map (3.1 Å) Data courtesy of P. Nissen



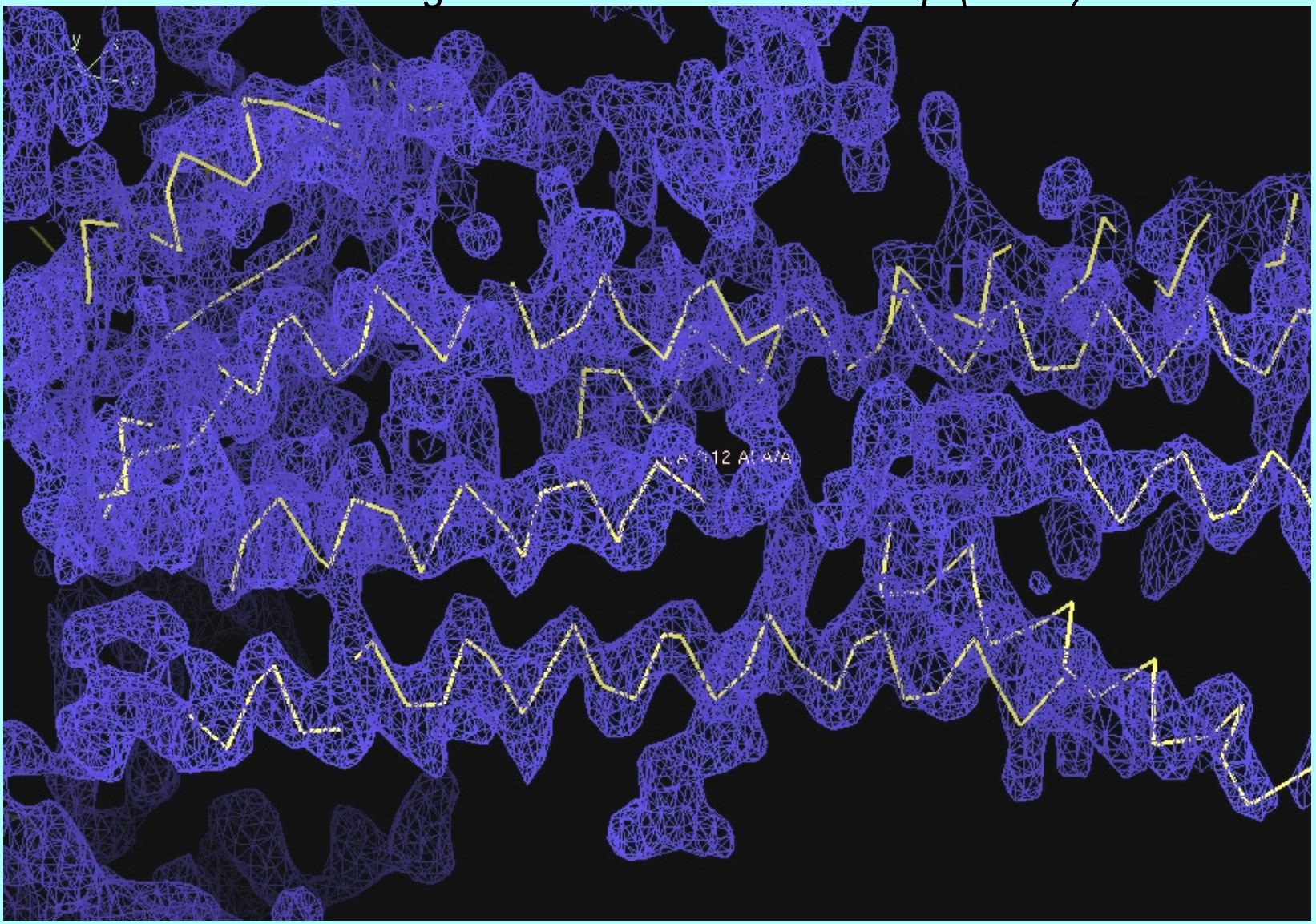
A real case: 1T5S SAD map (7 Å)



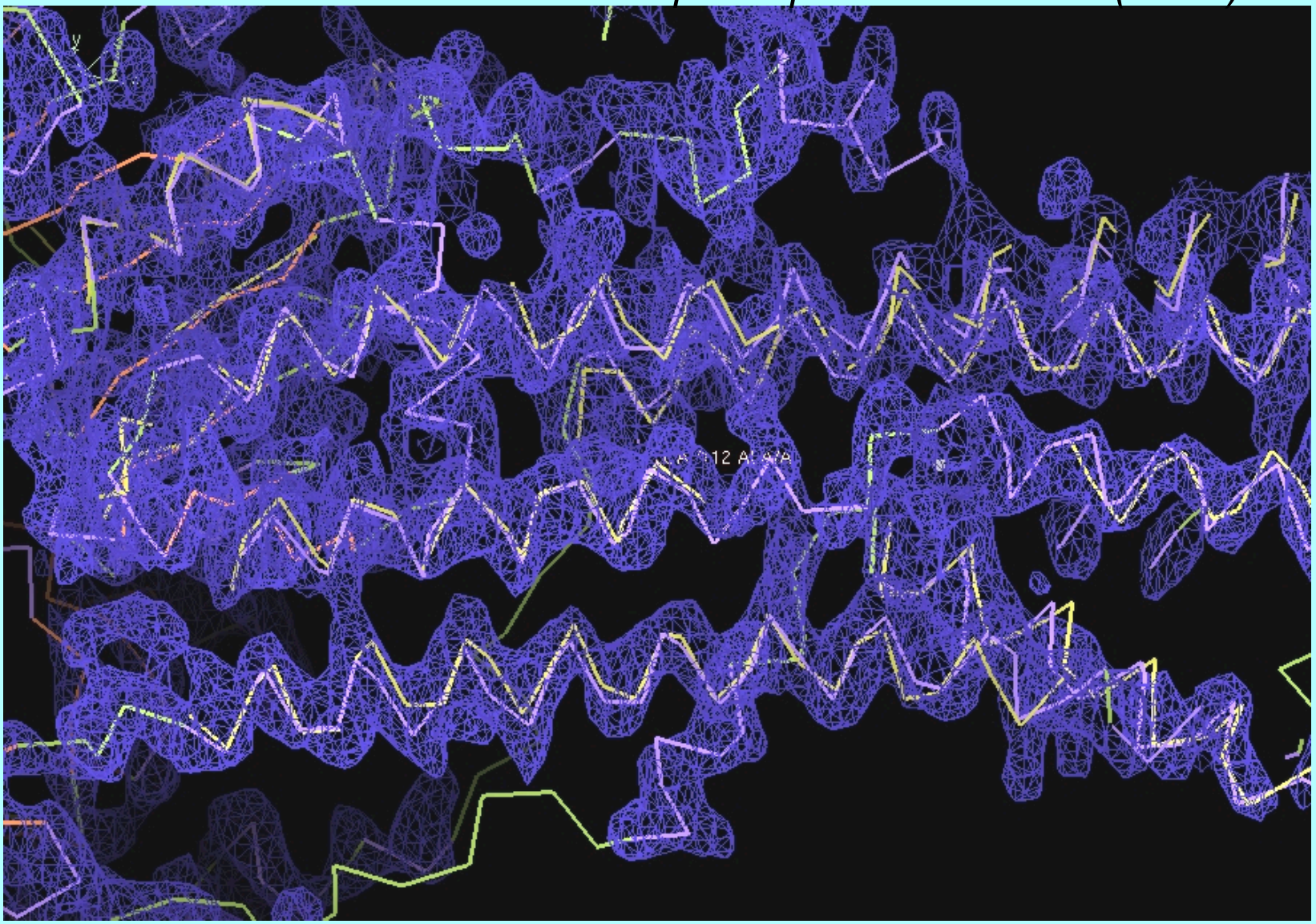
Finding helices in 1T5S SAD map (7 Å)

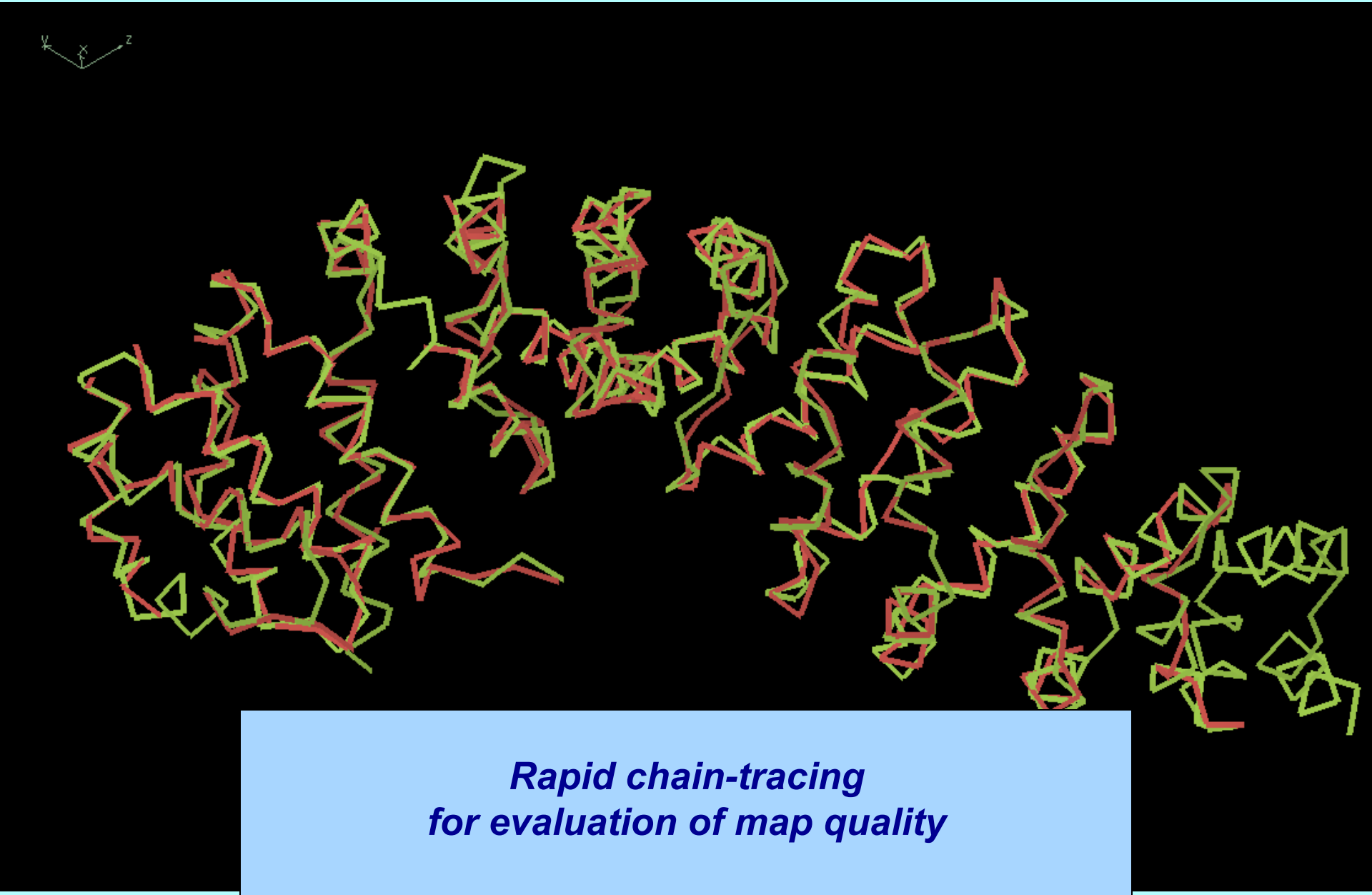


Finding helices in 1T5S SAD map (3.1 Å)



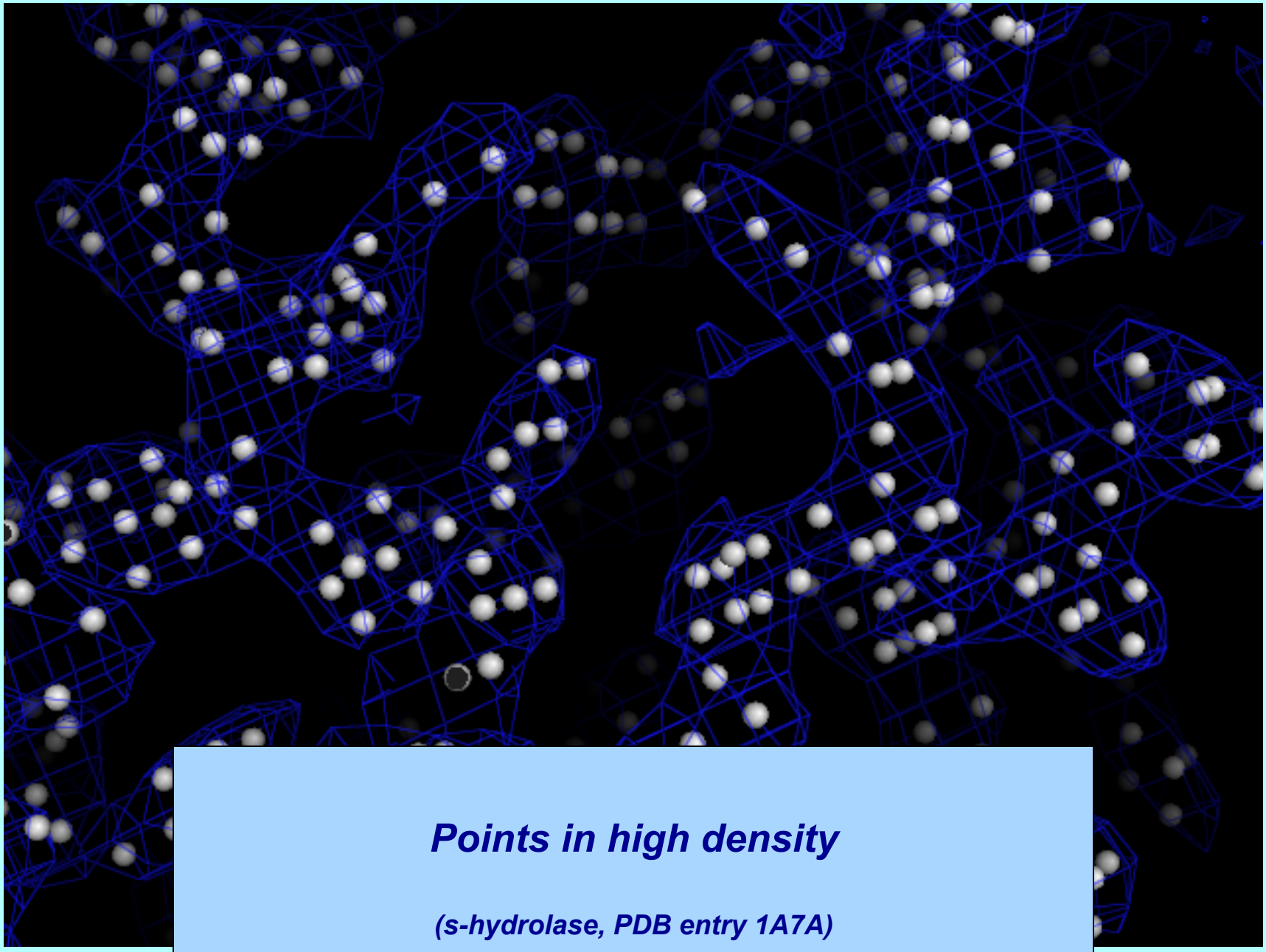
Helices from 1T5S SAD map compared with 1T5S (3.1 Å)





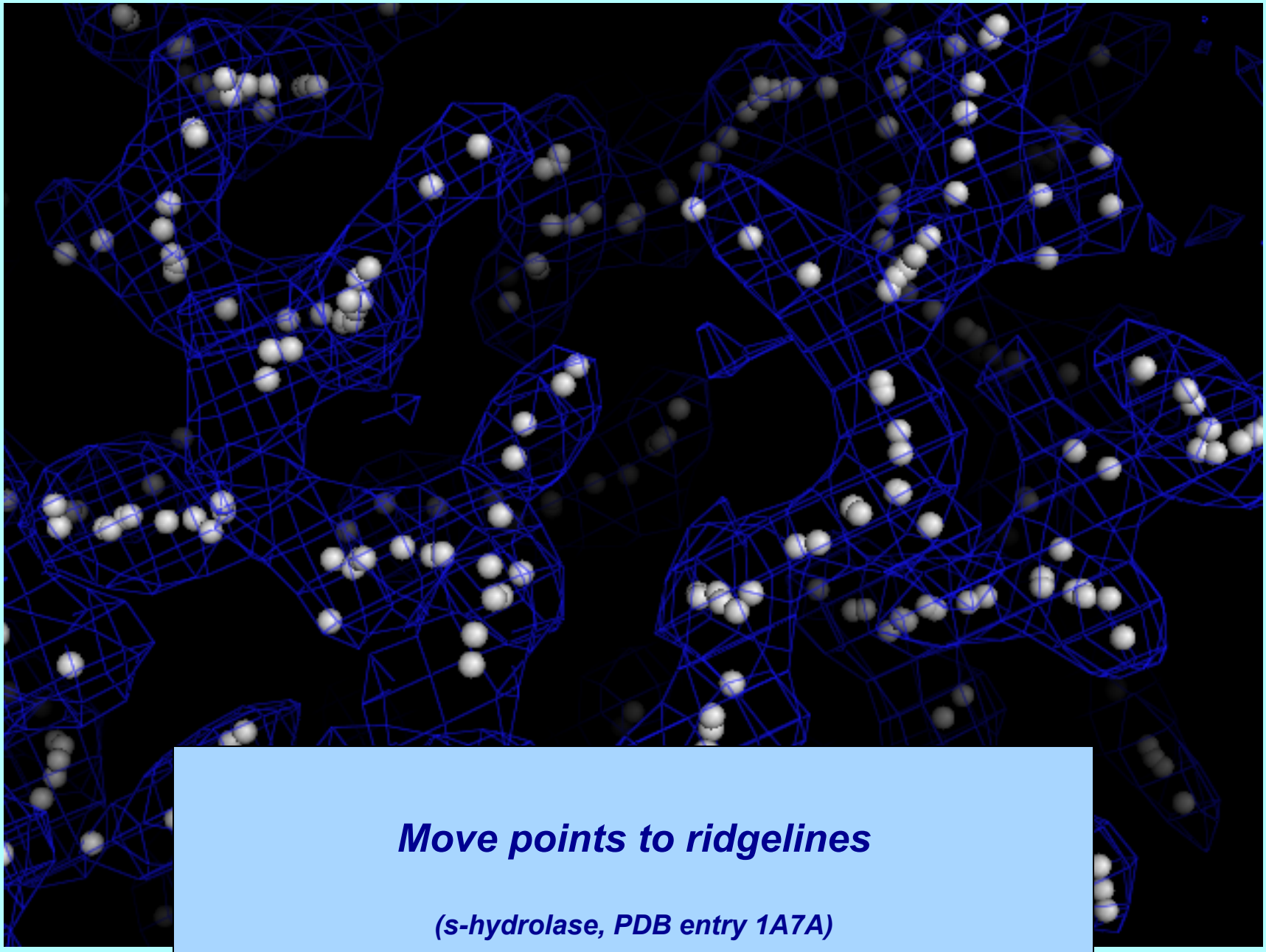
Rapid chain-tracing
for evaluation of map quality

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



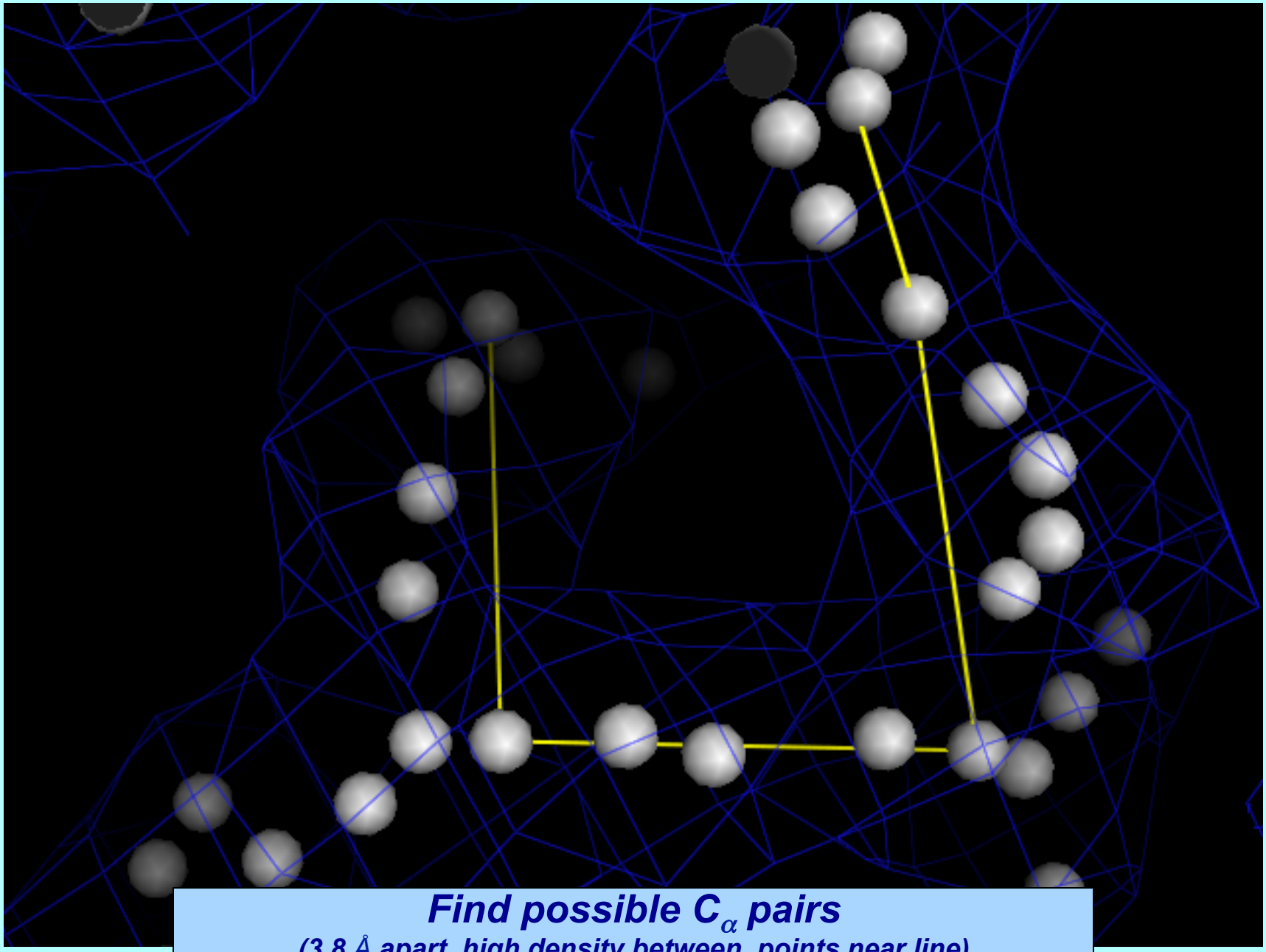
Points in high density

(s-hydrolase, PDB entry 1A7A)



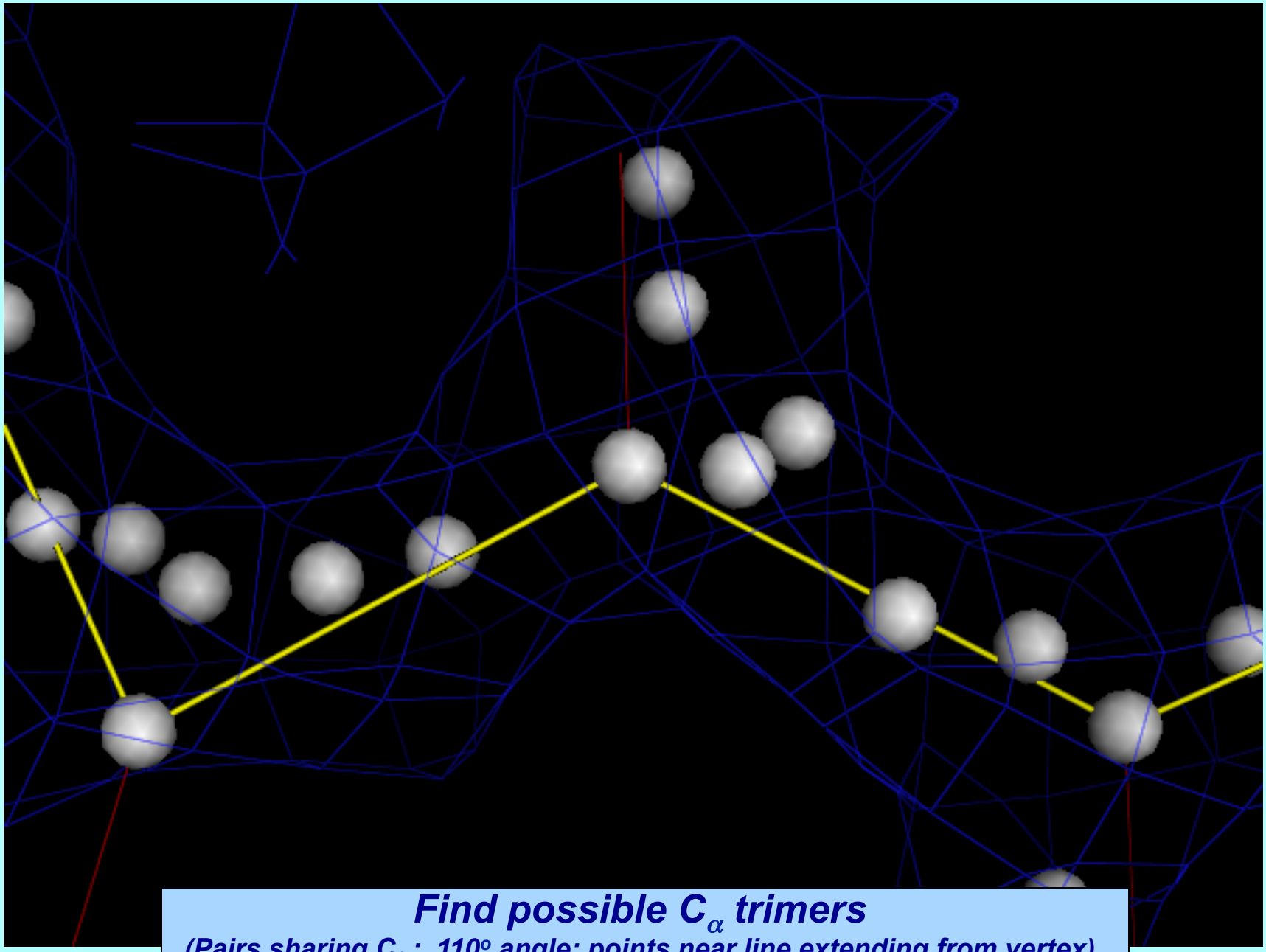
Move points to ridgelines

(s-hydrolase, PDB entry 1A7A)



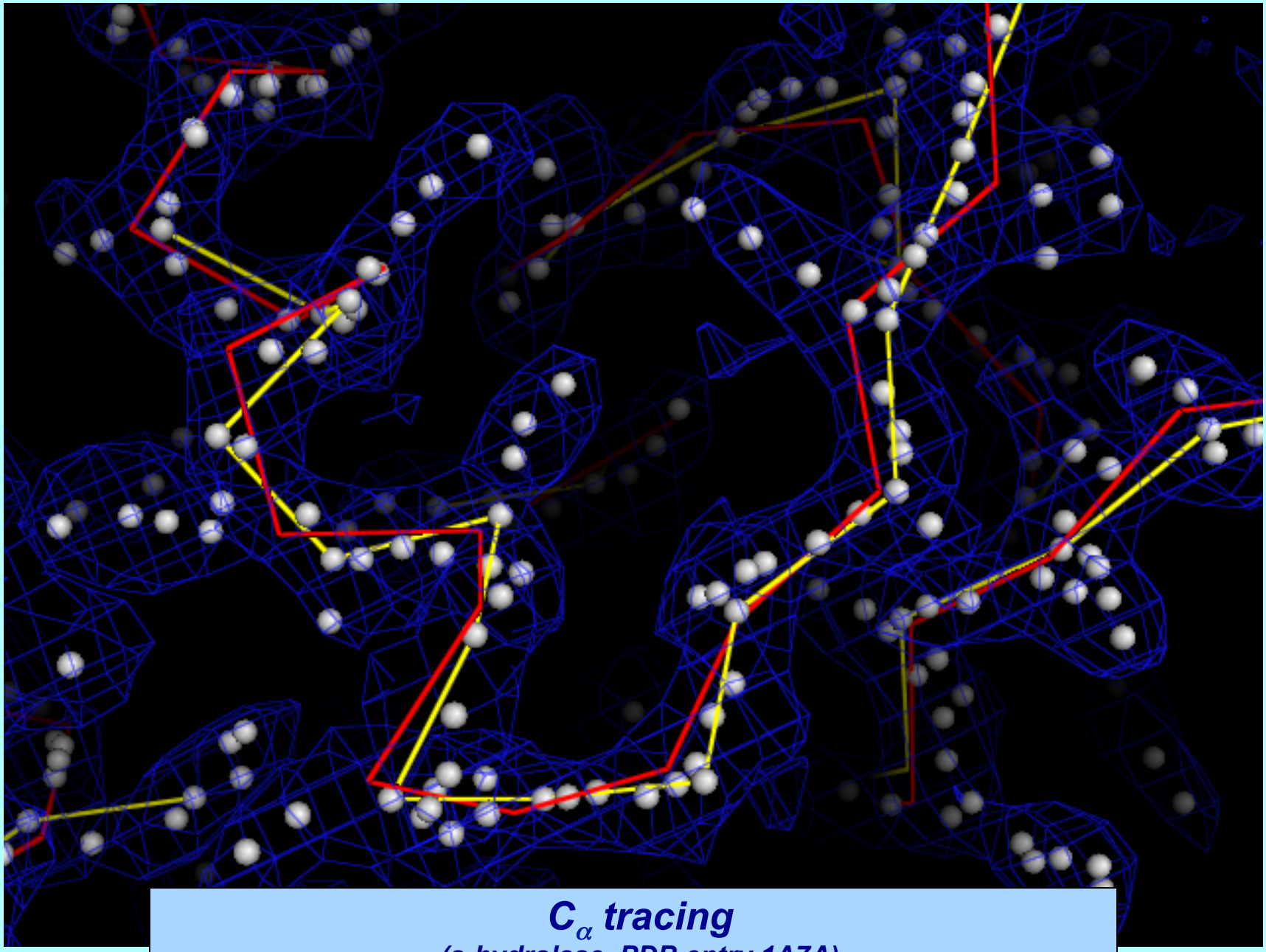
Find possible C_α pairs

(3.8 Å apart, high density between, points near line)



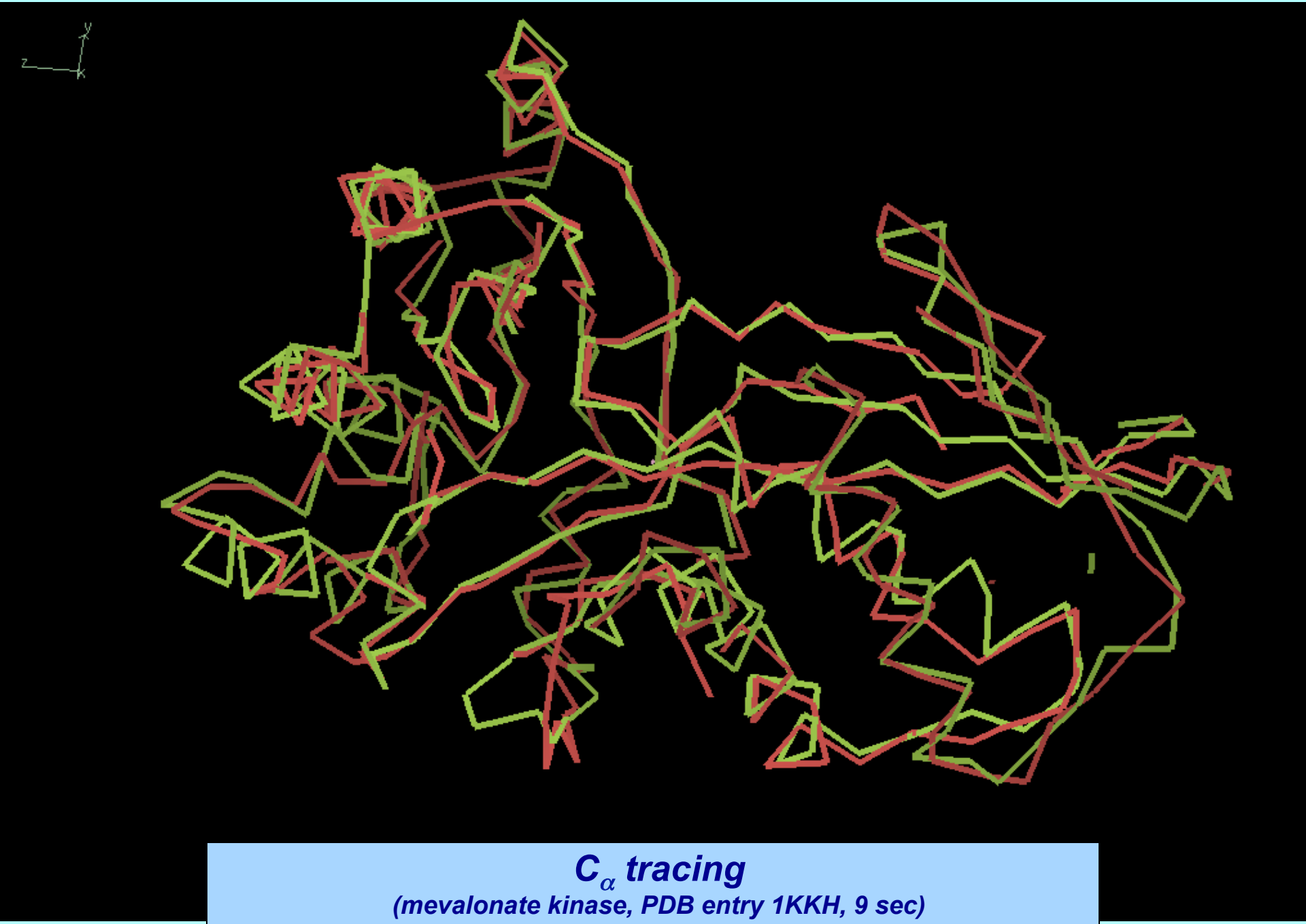
Find possible C_α trimers

(Pairs sharing C_α; 110^o angle; points near line extending from vertex)



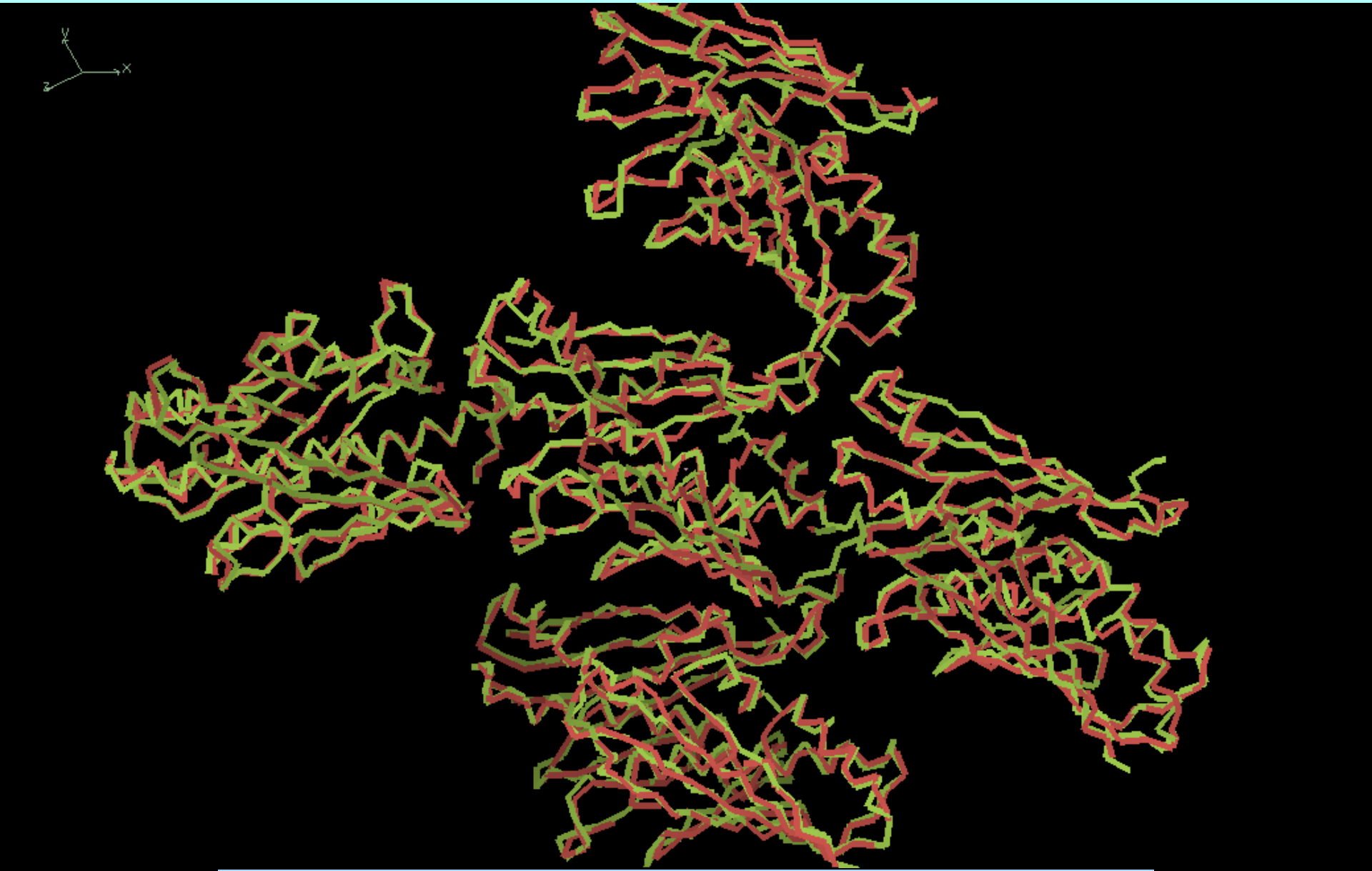
C_α tracing

(s-hydrolase, PDB entry 1A7A)



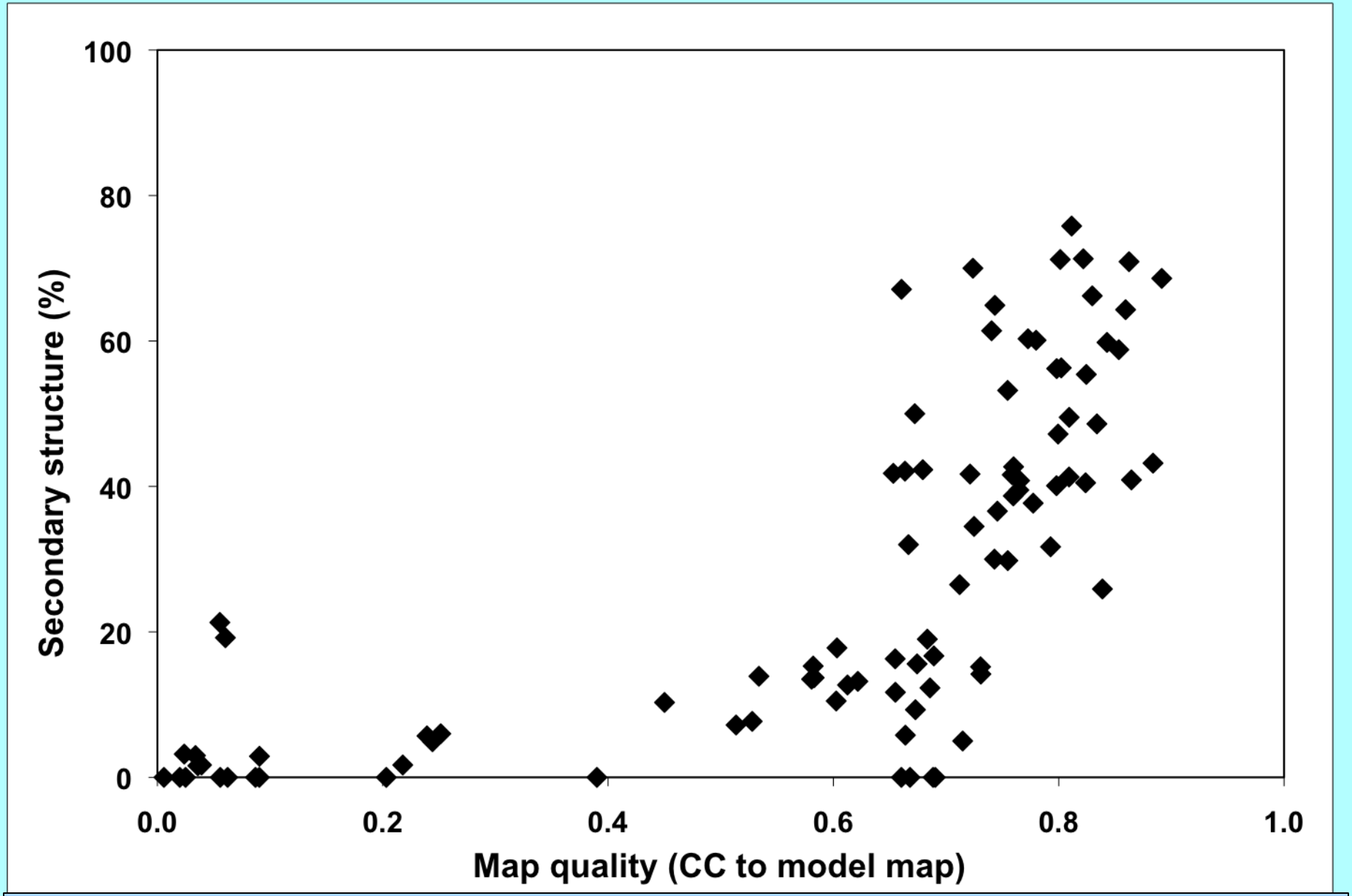
C_α tracing

(mevalonate kinase, PDB entry 1KKH, 9 sec)



C_α tracing

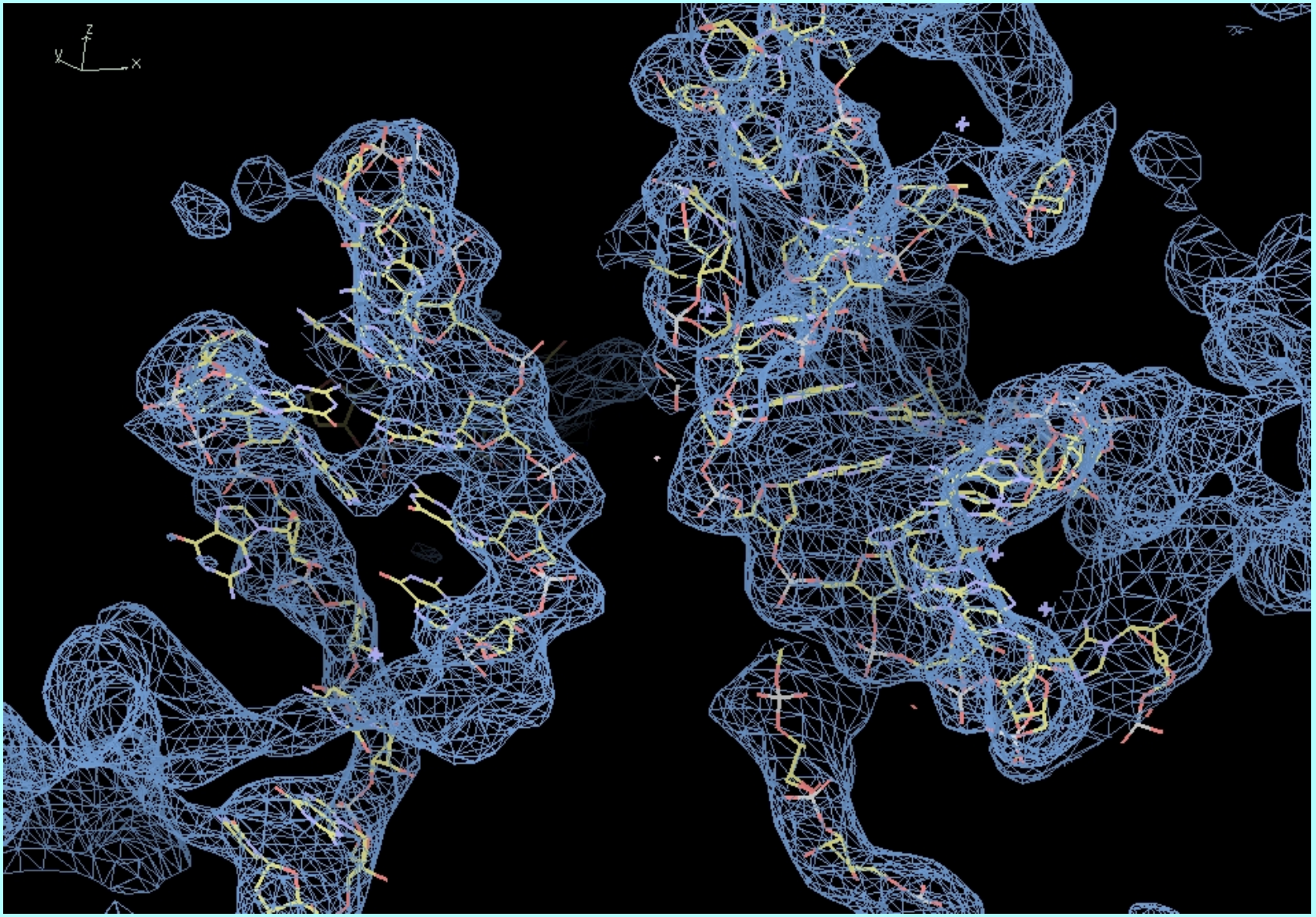
(1038B, PDB entry 1LQL, 114 sec)



Using secondary structure content to evaluate map quality

Building RNA

Group II intron at 3.5 Å. Data courtesy of J. Doudna

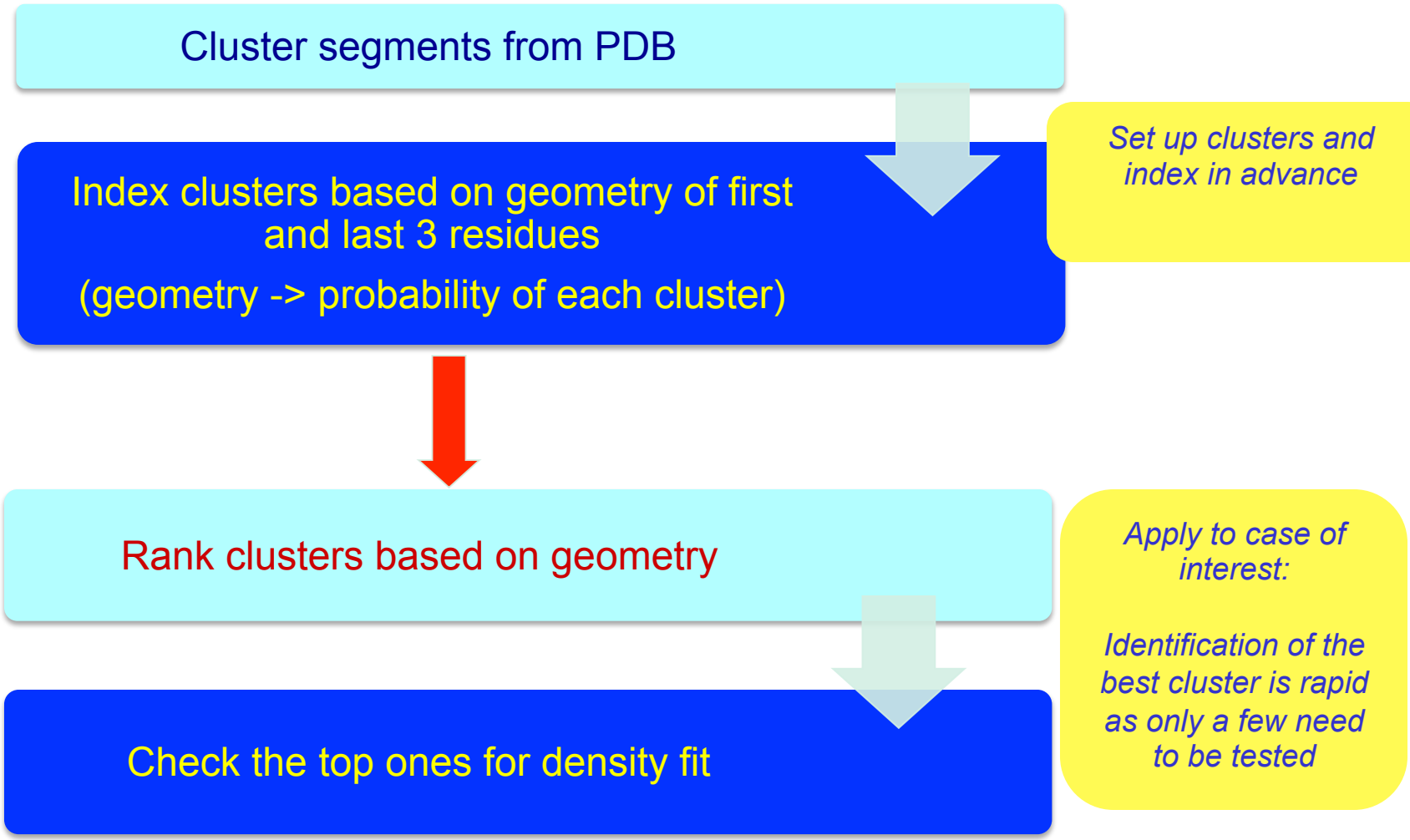


Rapid model-building options in PHENIX

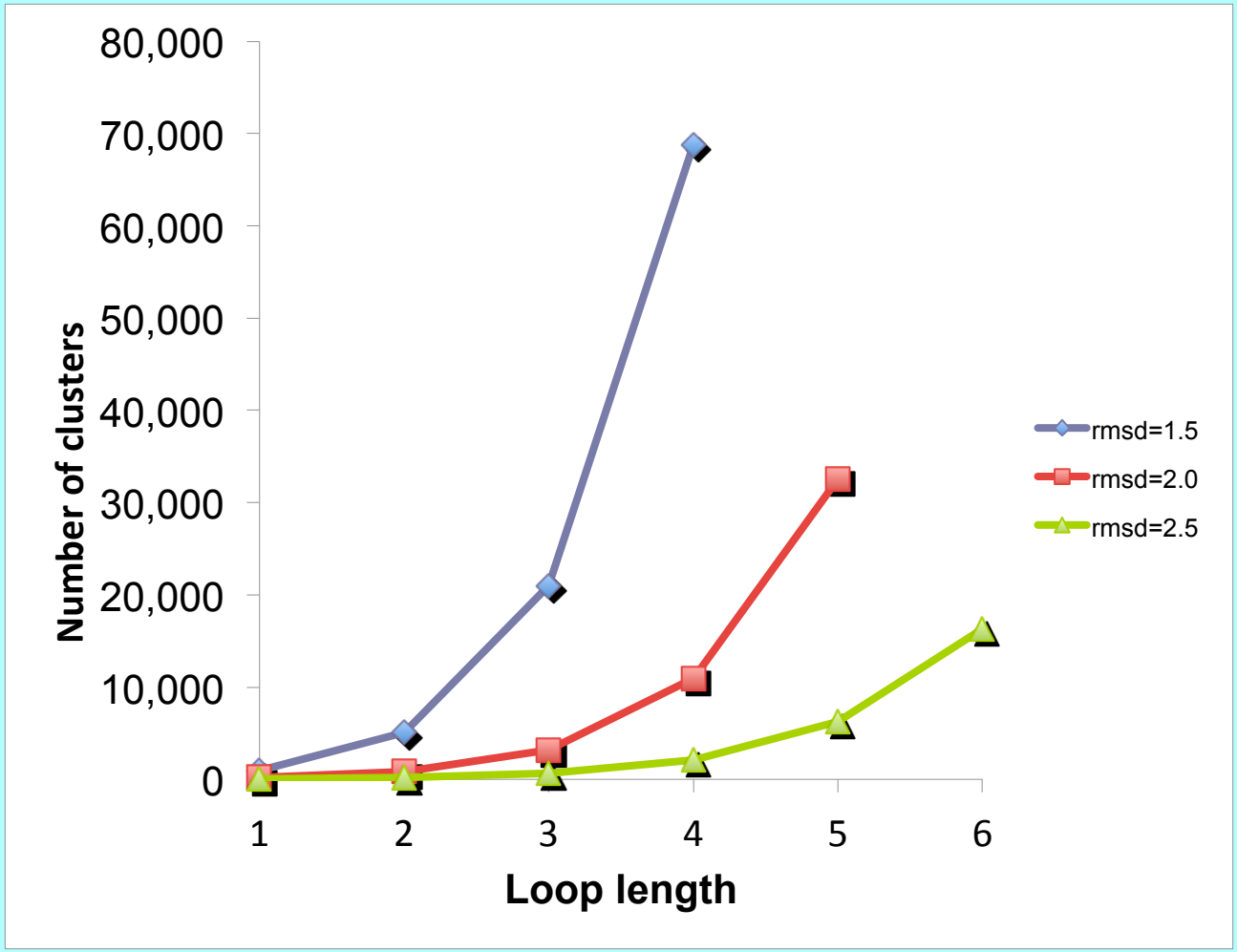
(42 structures; 26651 residues)

Method	Residues built	RMSD (Å)	Time (sec)	Residues/sec
trace_chain	21428	1.61	1441	14.9
helices_strands	12322	1.24	5331	2.3
RESOLVE	19037	1.16	16933	1.1

Fitting loops with an indexed loop library

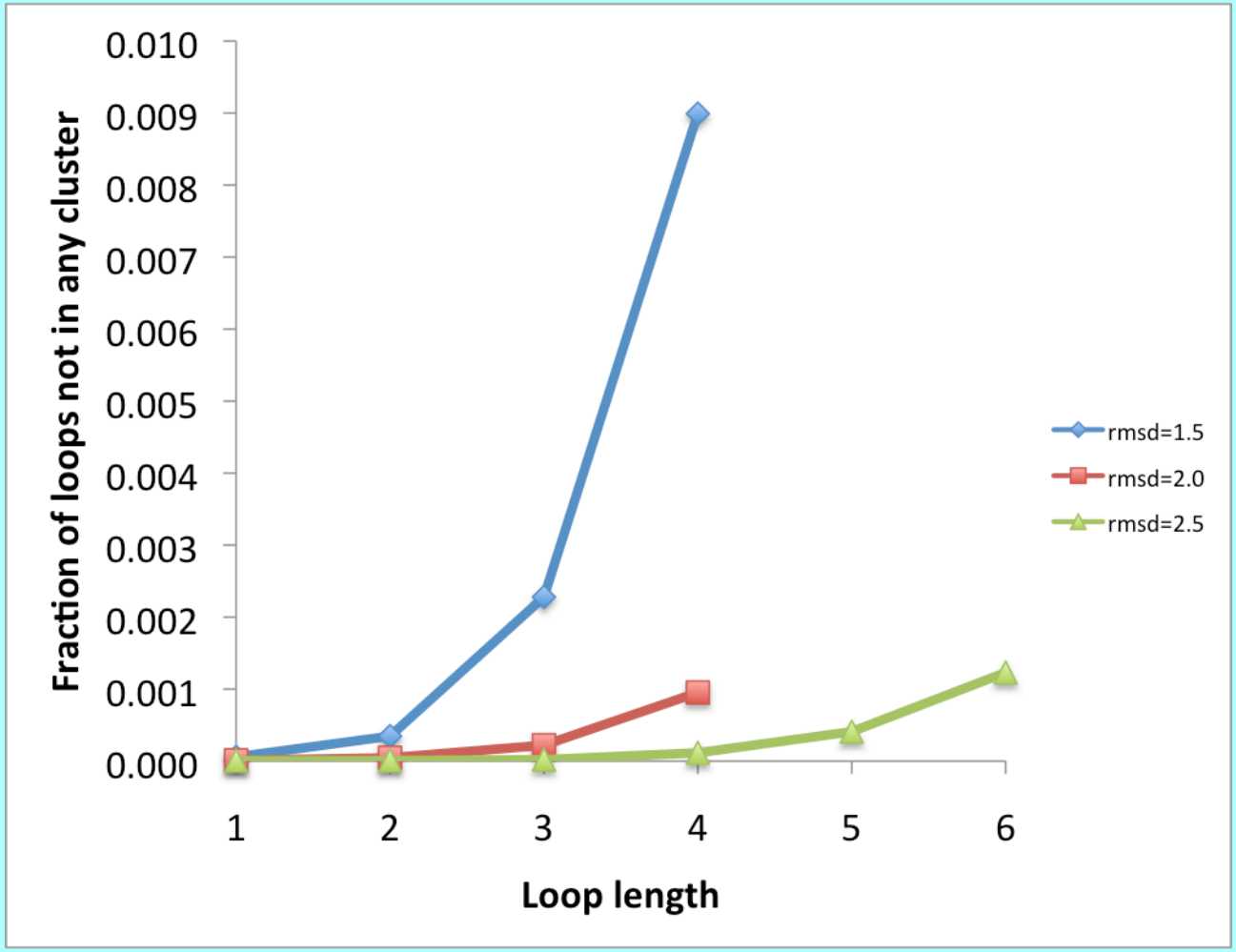


Loop libraries: how many clusters are there?



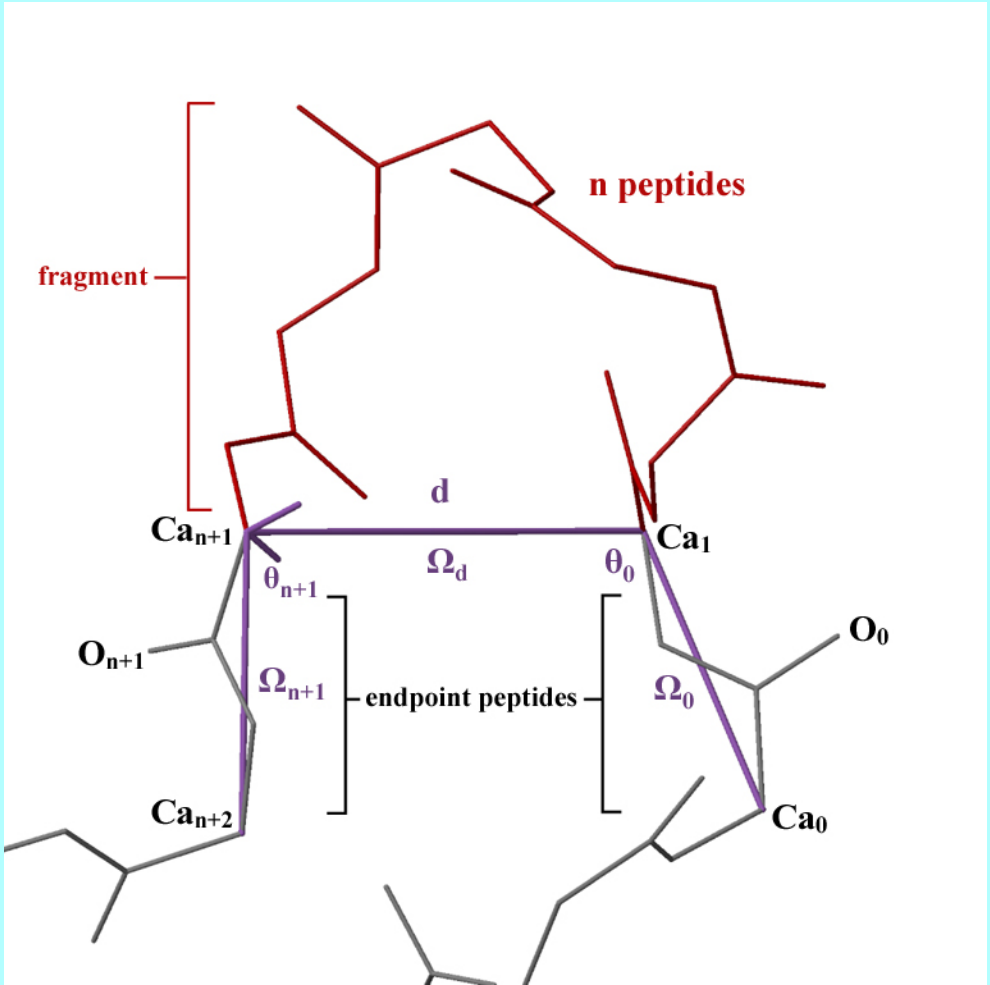
NOTE: clustering based on loop + 3 residues on each end

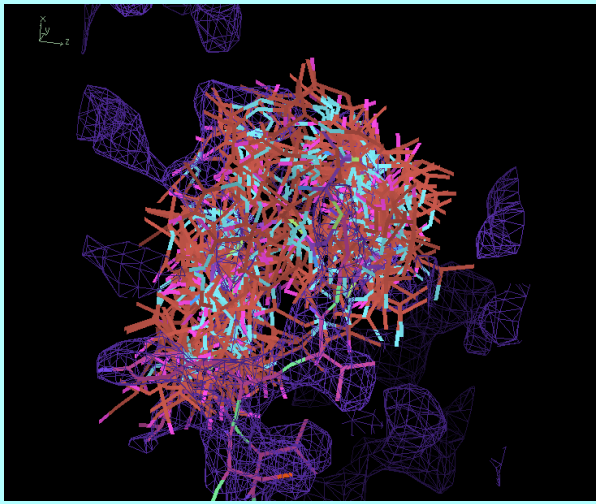
Loop libraries: how complete is coverage by clusters?



Indexing loops using 7 parameters describing loop length and existing loop ends

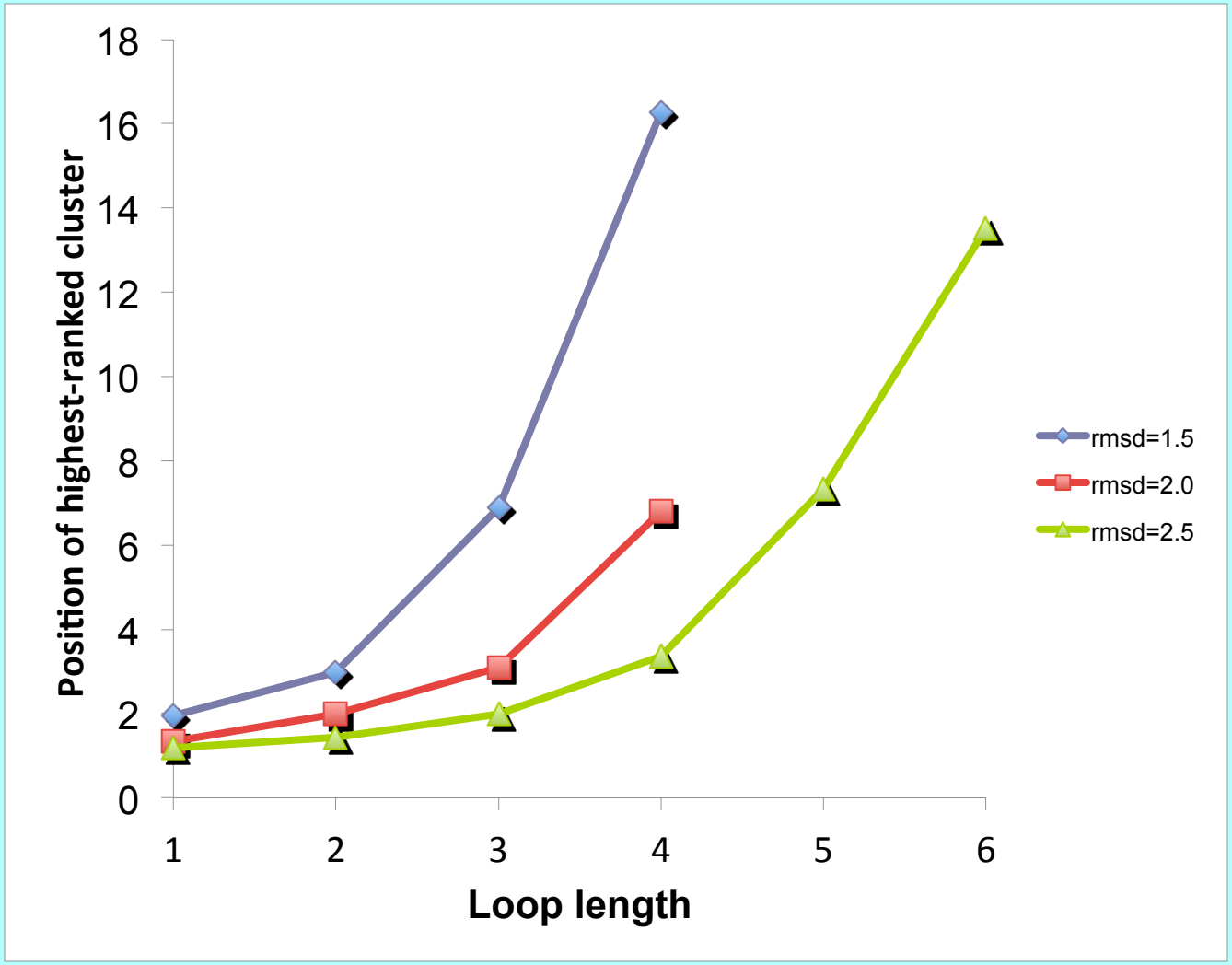
Number of peptides in loop
Distance between C-alpha positions of existing ends
2 angles describing chain direction at existing ends
3 dihedrals describing chain orientation at existing ends

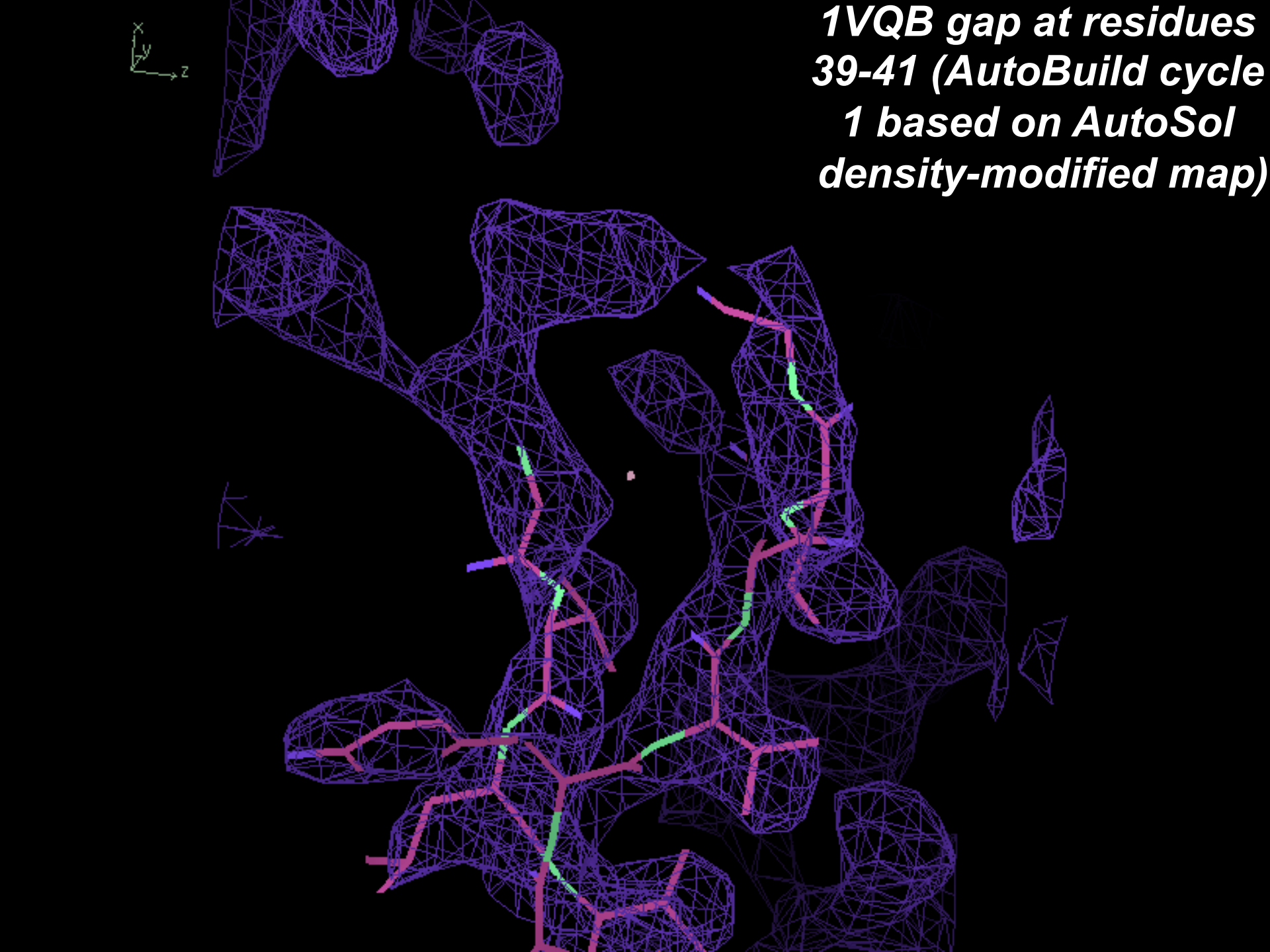


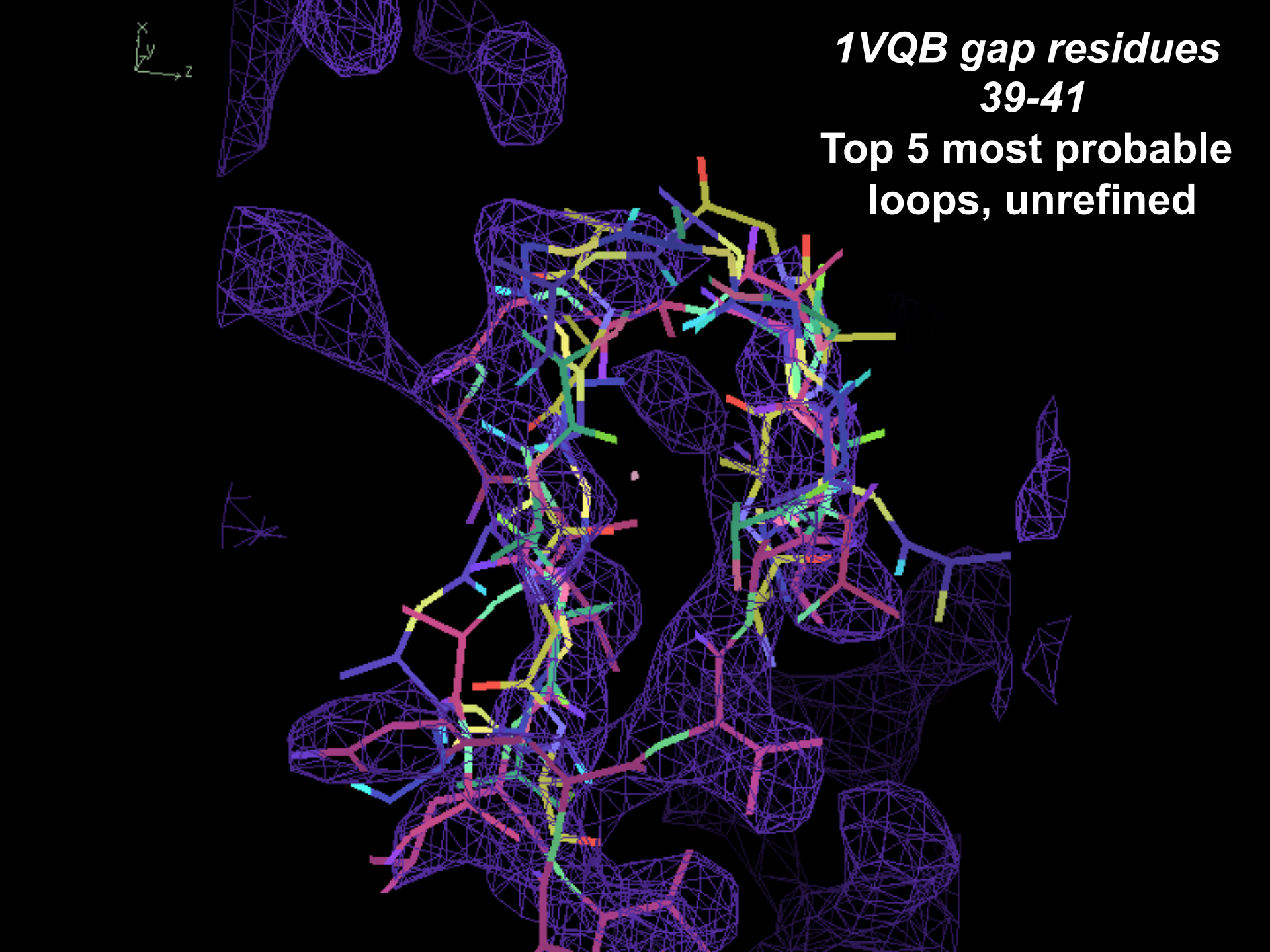


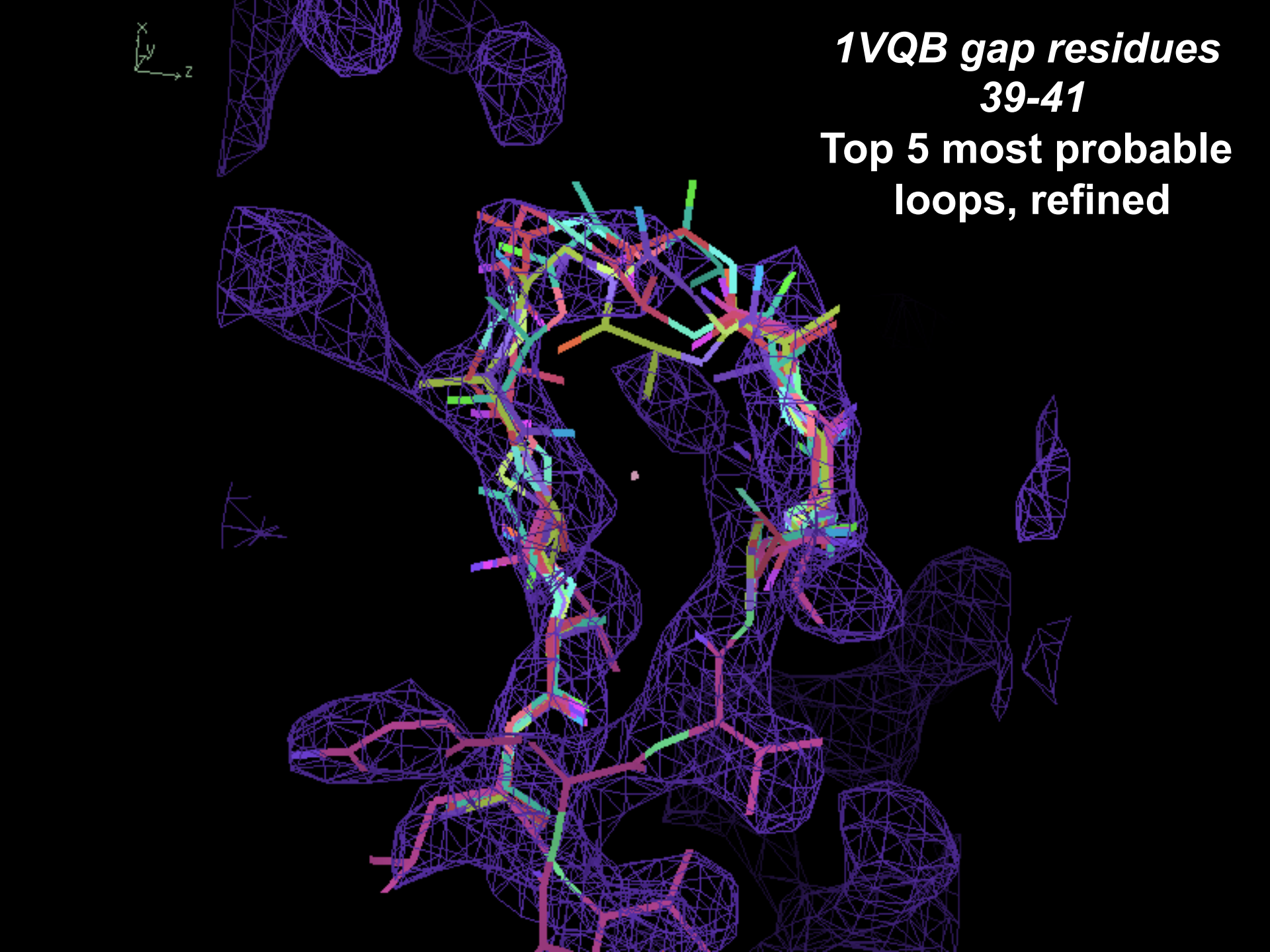
Identification of the cluster that has the lowest rmsd to a target fragment using index based on 3 residues at each end

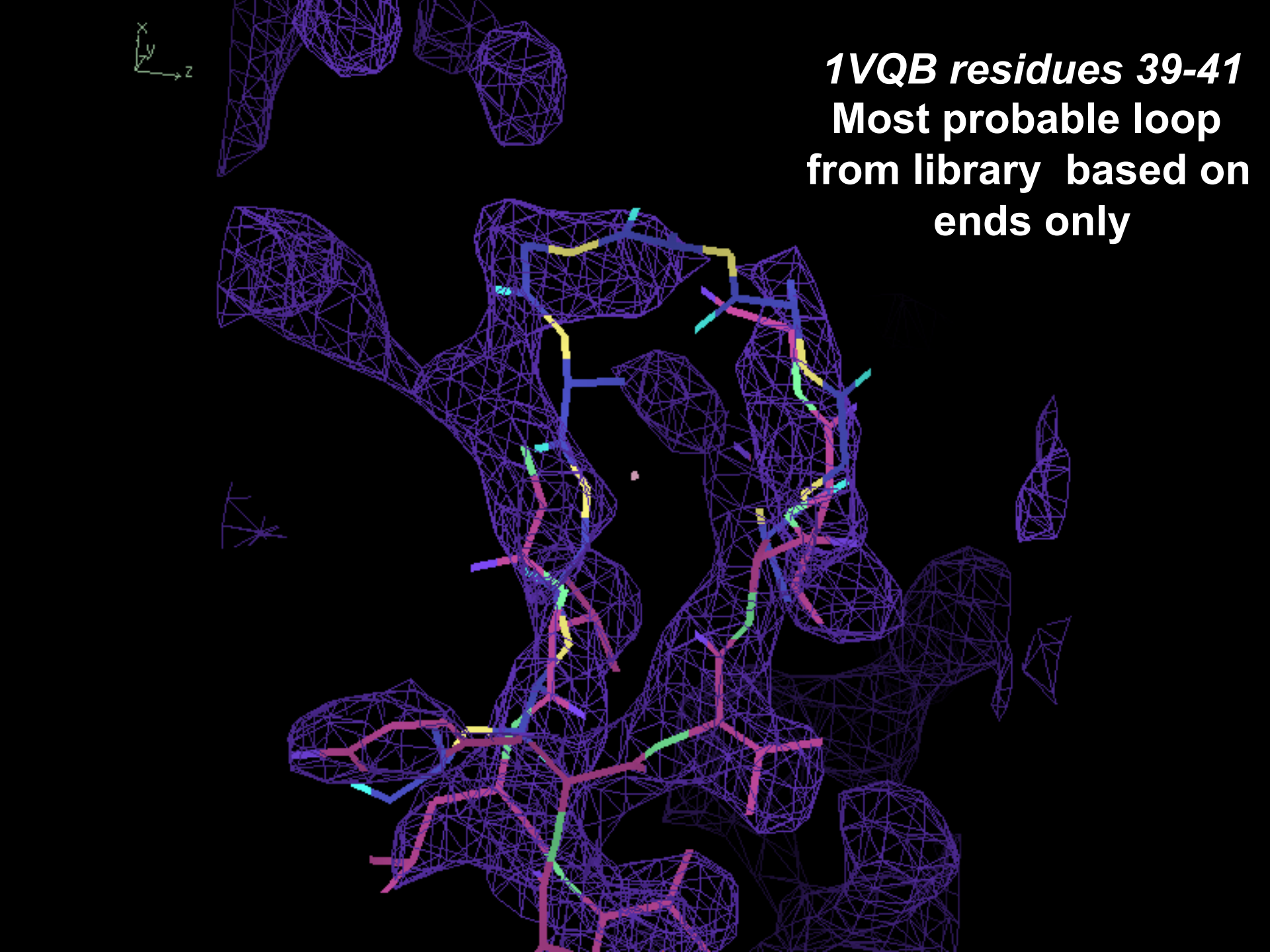
(test using non-overlapping database of structures)

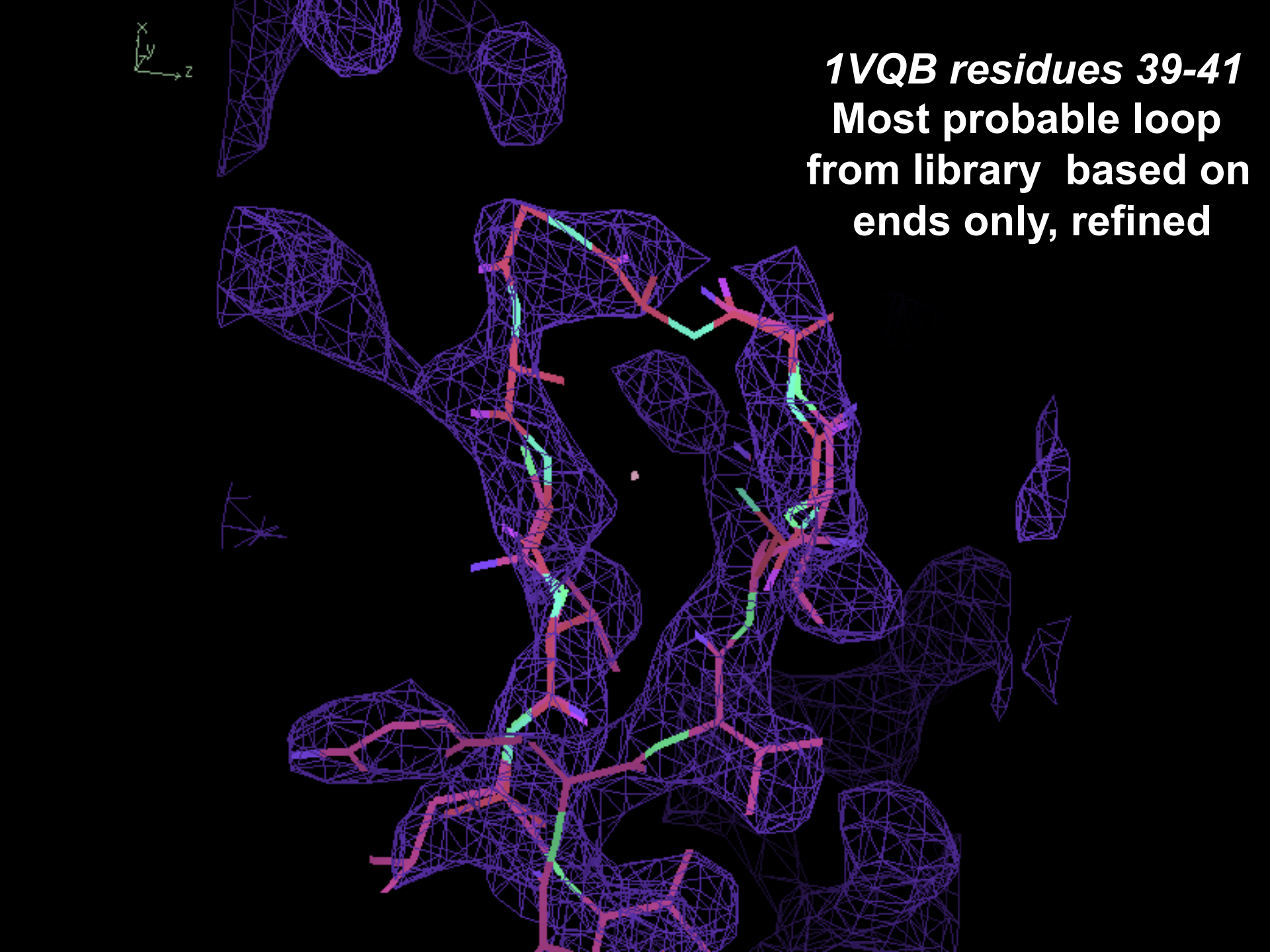


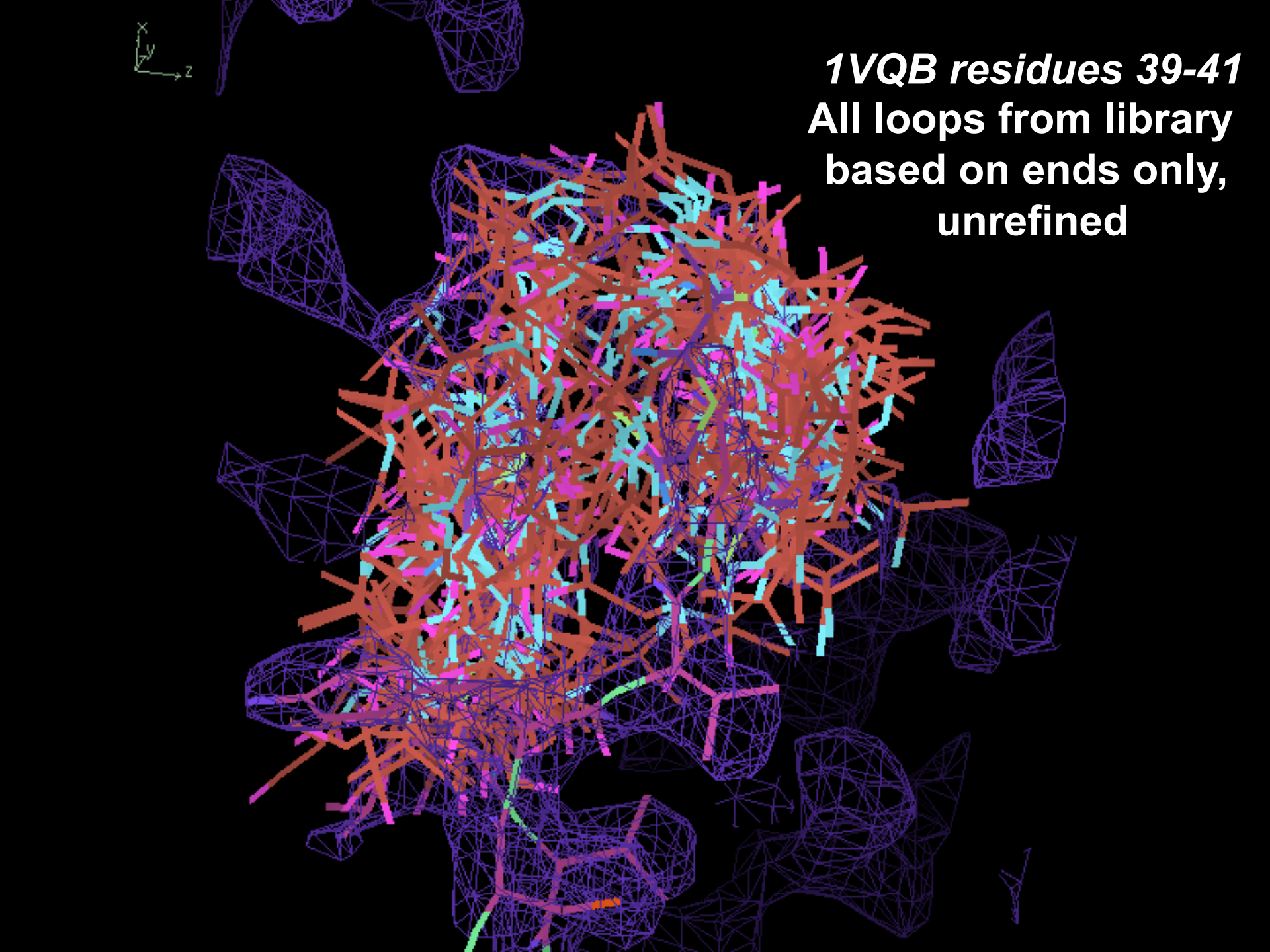

1VQB gap at residues 39-41 (AutoBuild cycle 1 based on AutoSol density-modified map)

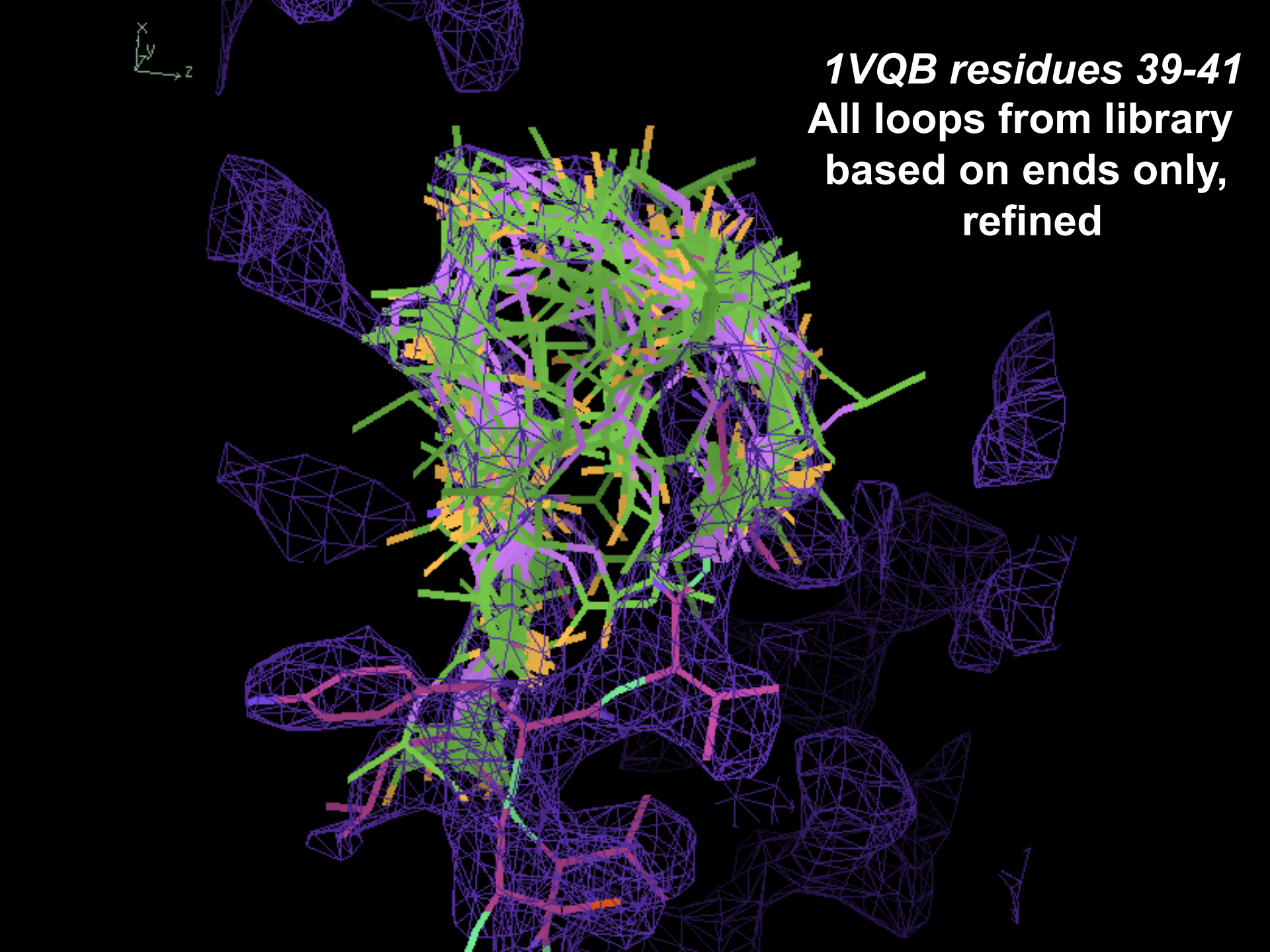

1VQB gap residues
39-41
Top 5 most probable
loops, unrefined


1VQB gap residues
39-41
Top 5 most probable
loops, refined

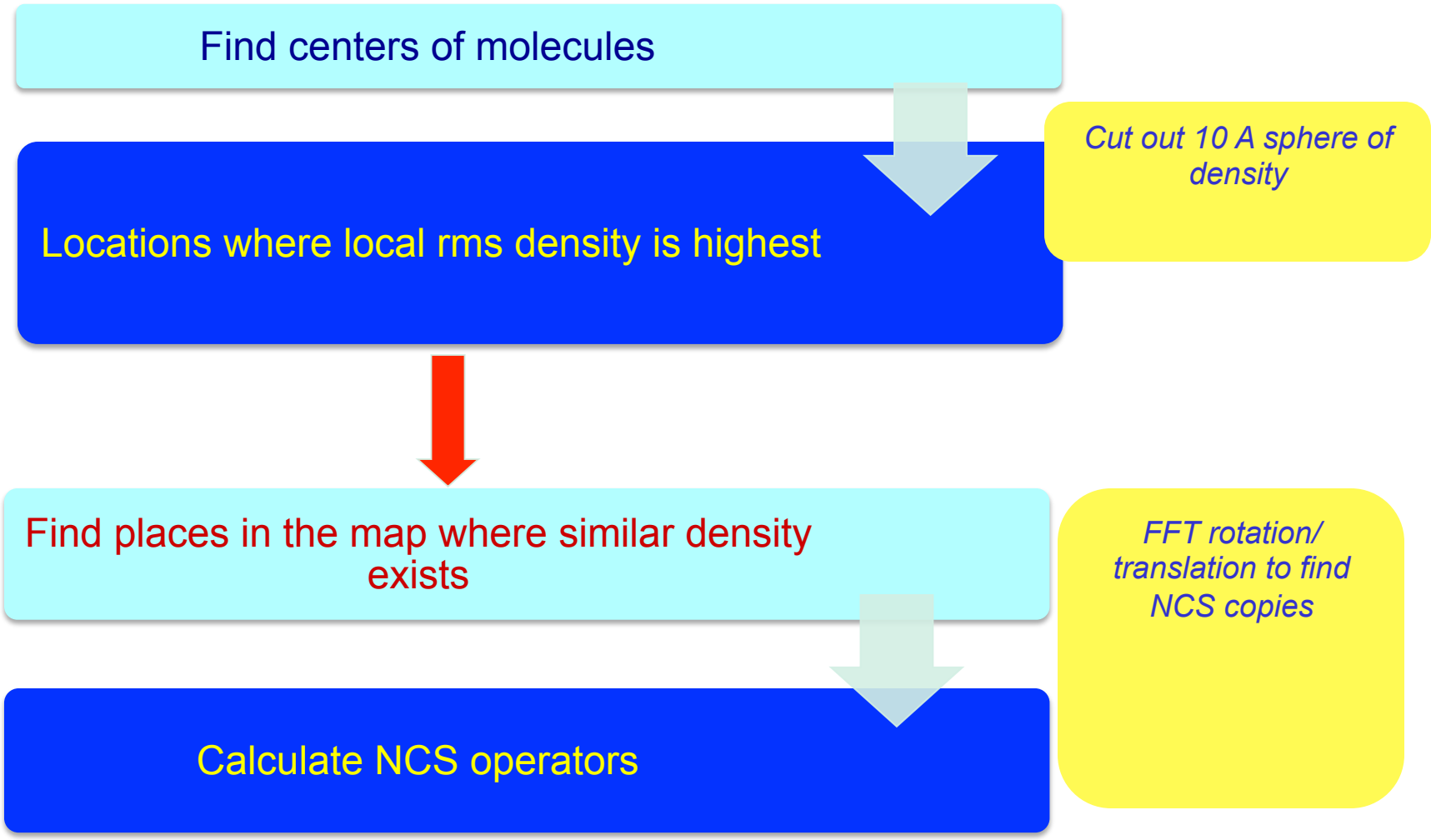

1VQB residues 39-41
Most probable loop
from library based on
ends only


1VQB residues 39-41
Most probable loop
from library based on
ends only, refined


1VQB residues 39-41
All loops from library
based on ends only,
unrefined

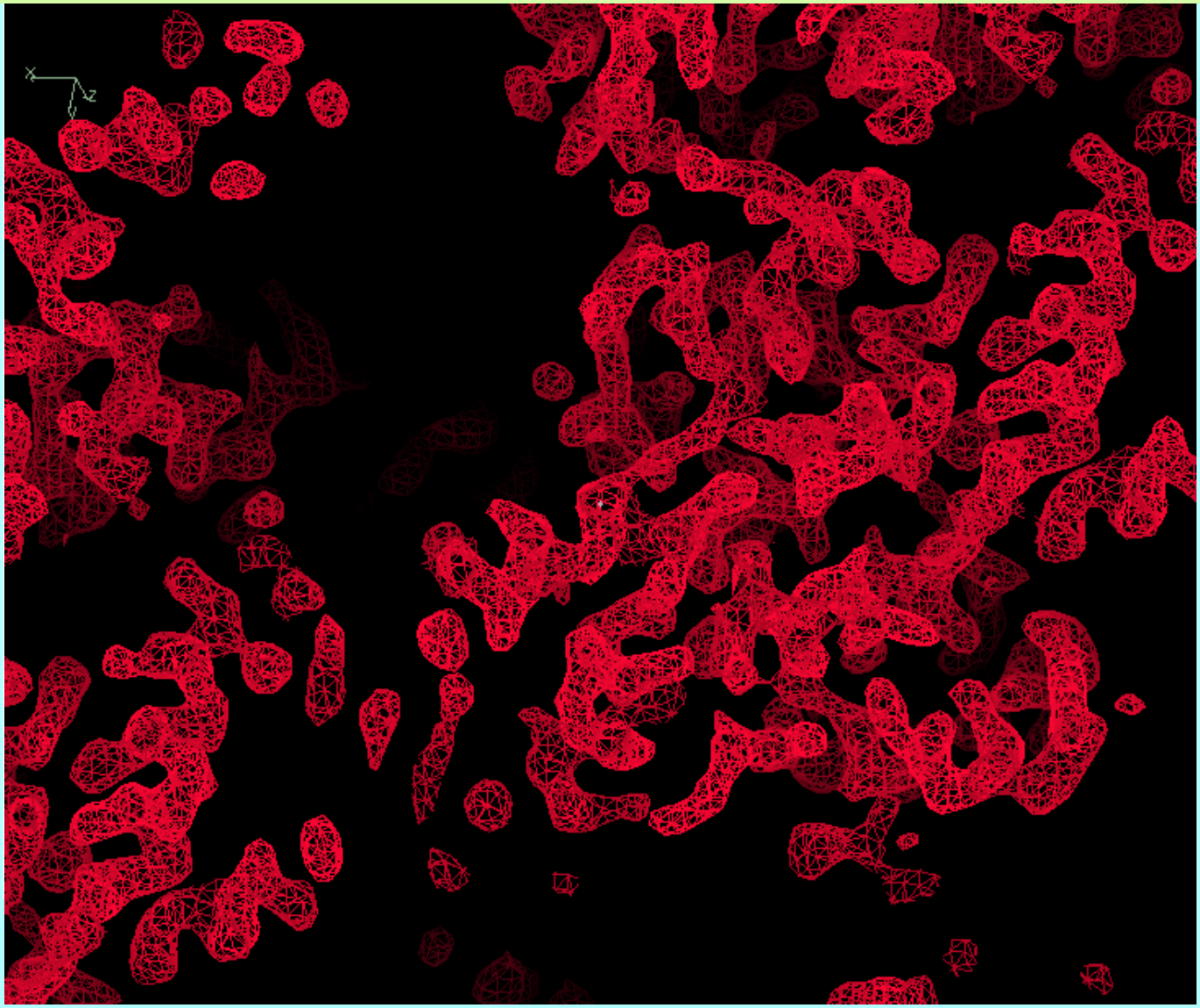

1VQB residues 39-41
All loops from library
based on ends only,
refined

Finding NCS from an electron density map

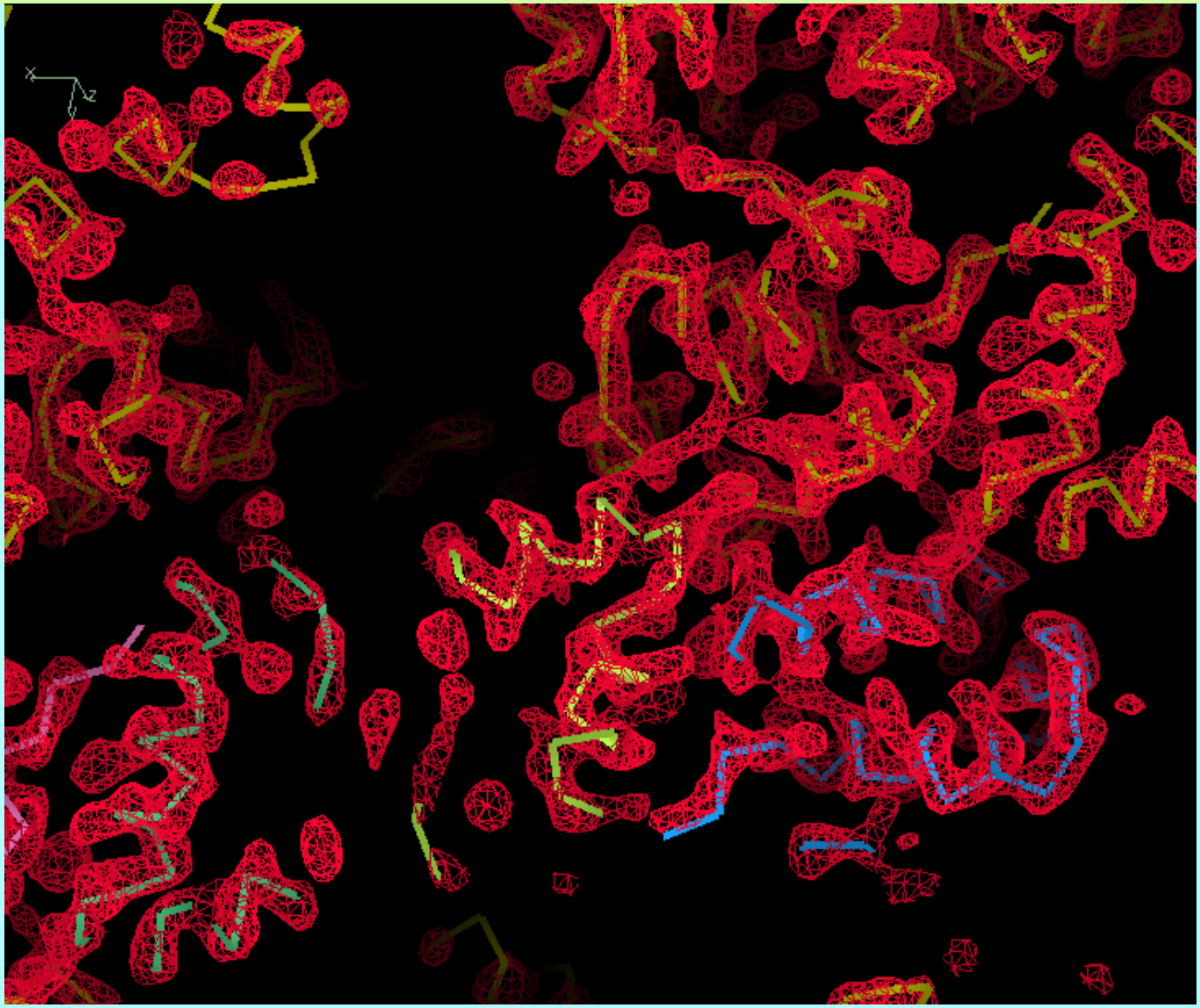


Finding NCS copies

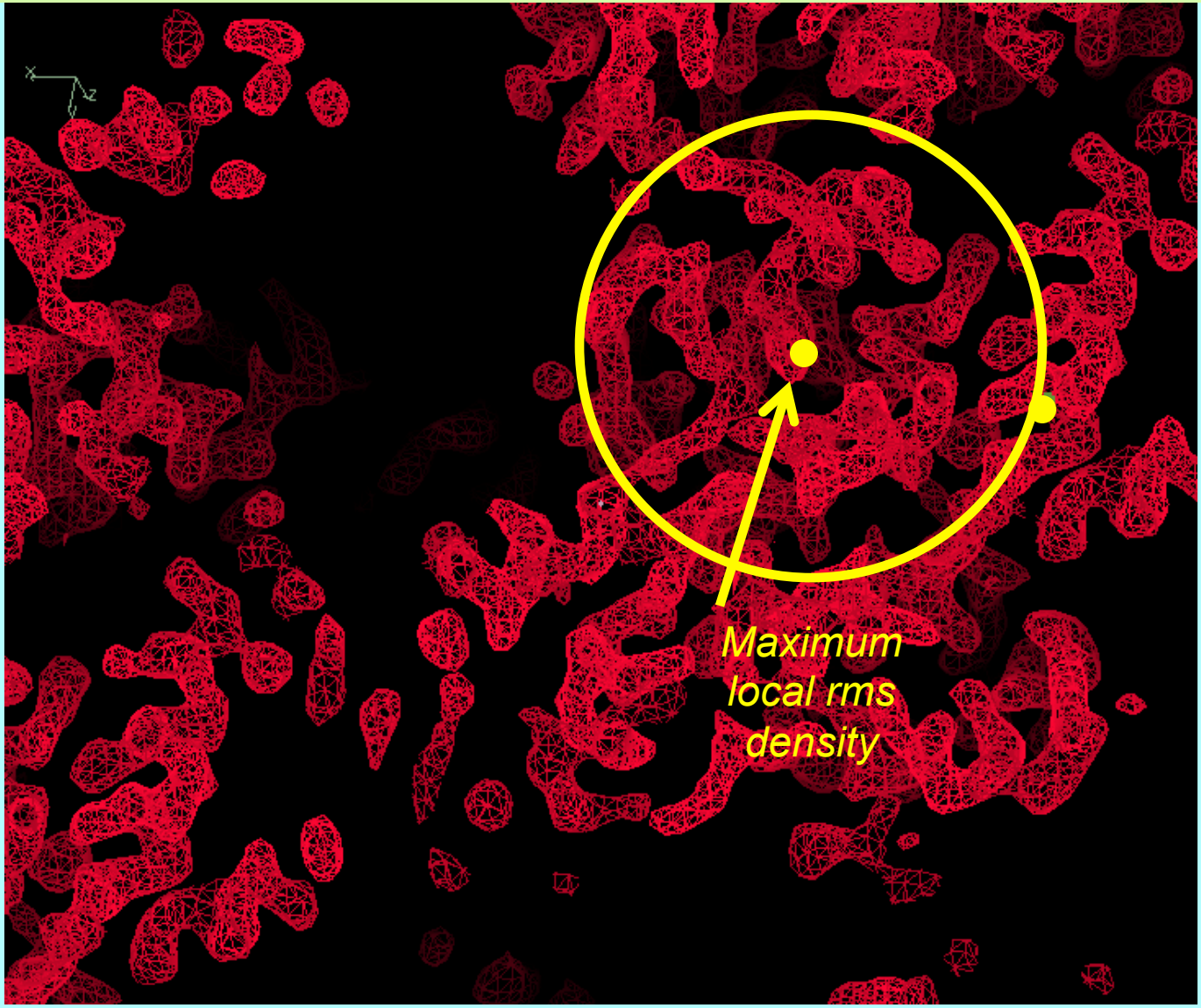
gerE 1FSE, Ducros et al., 2001



Finding NCS copies

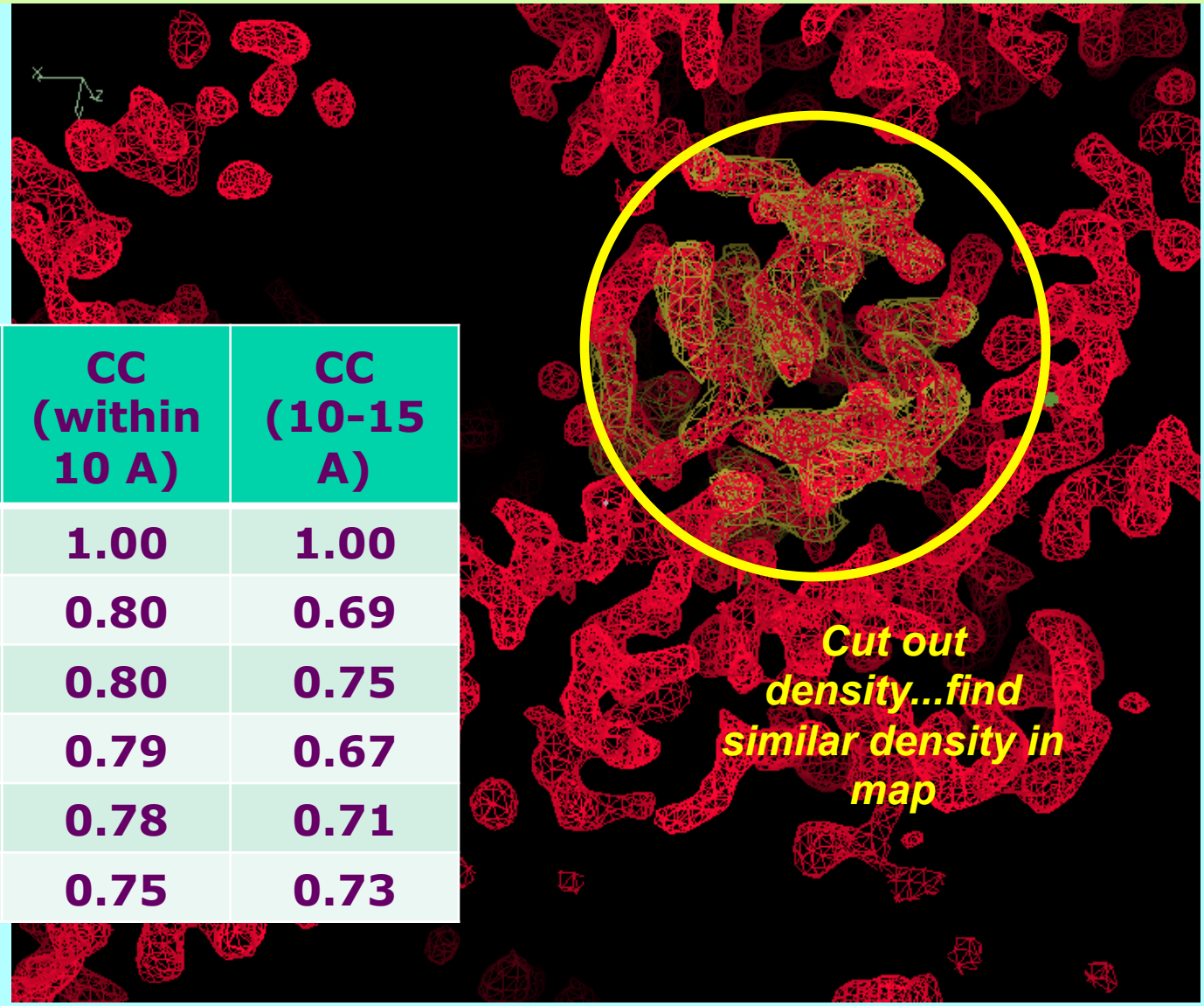


Finding NCS copies



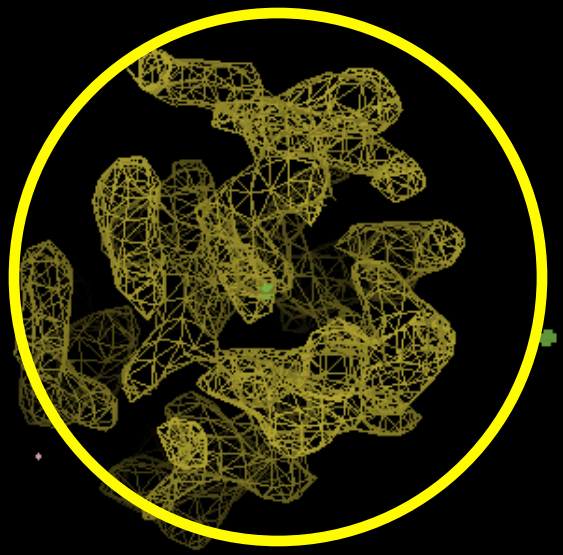
Finding NCS copies

peak	CC (within 10 A)	CC (10-15 A)
1	1.00	1.00
2	0.80	0.69
3	0.80	0.75
4	0.79	0.67
5	0.78	0.71
6	0.75	0.73



Finding NCS copies

peak	CC (within 10 A)	CC (10-15 A)
1	1.00	1.00
2	0.80	0.69
3	0.80	0.75
4	0.79	0.67
5	0.78	0.71
6	0.75	0.73



Local average of 6 NCS copies of density in map

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

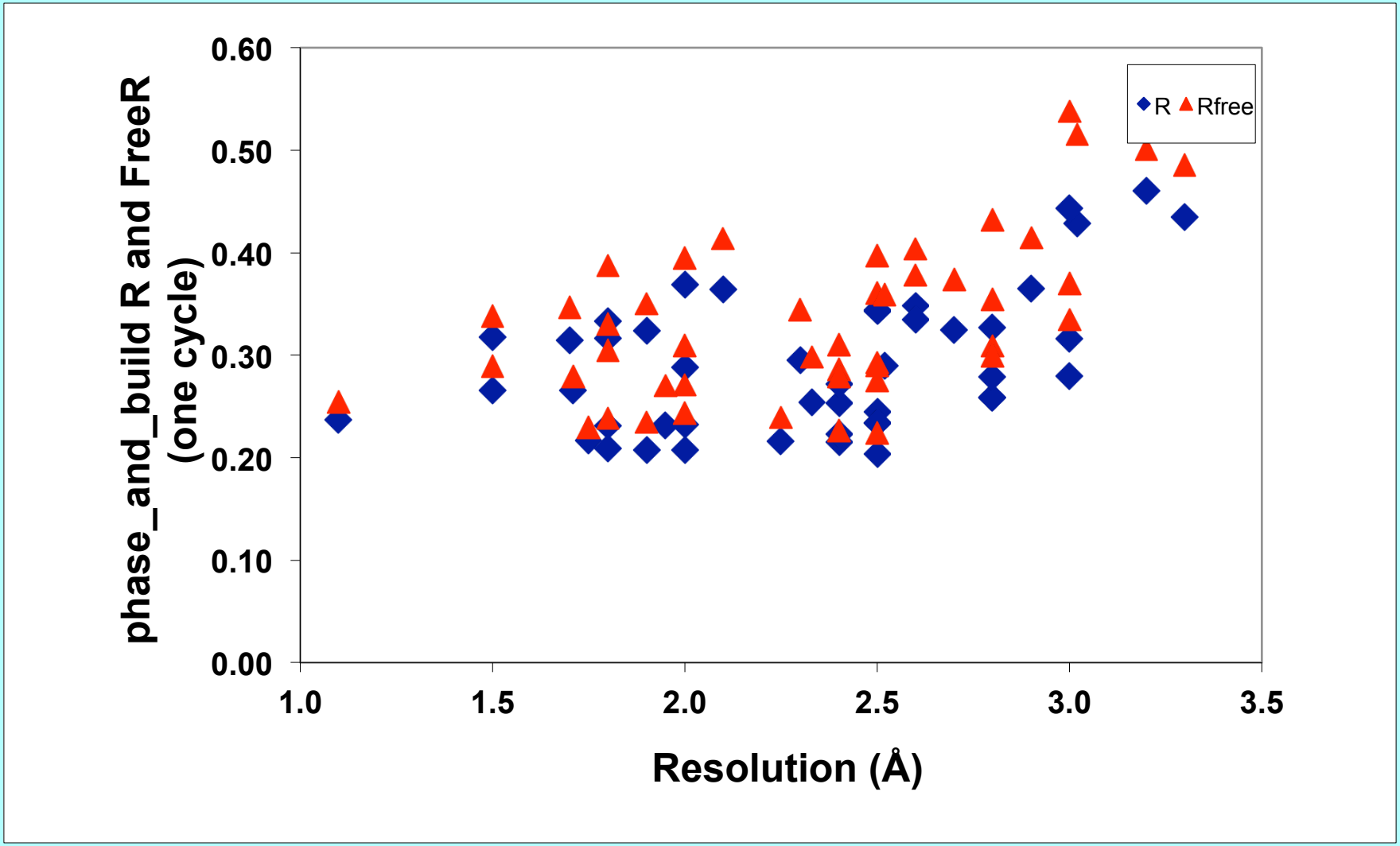
Model-building and refinement with NCS

Comprehensive sequence assignment

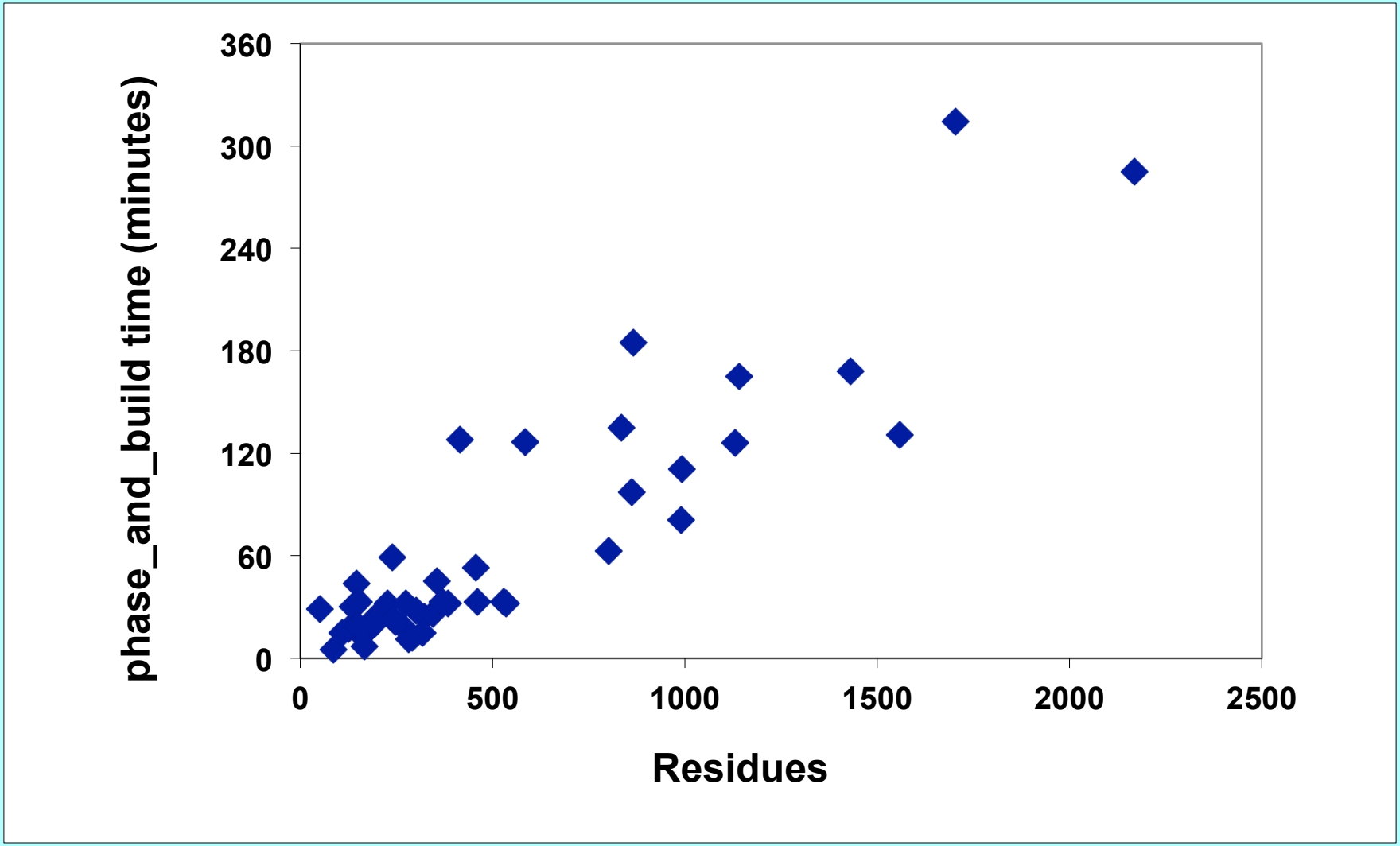
Loop fitting

phenix.phase_and_build – tests with structure library

One cycle, final R/Rfree



phase_and_build – tests with structure library
One cycle, time required



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

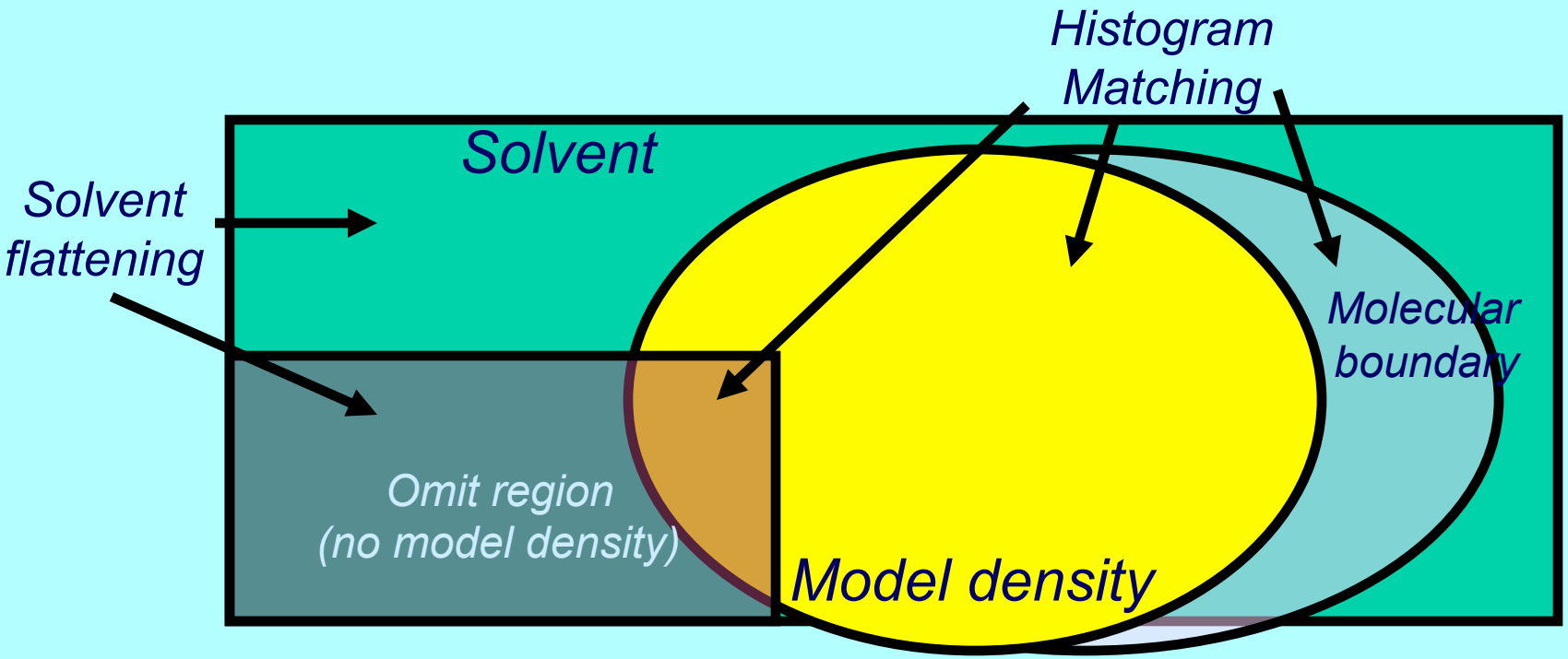
…checking different space groups, datasets to use

…checking if your model is biasing your map

…checking if you always get the same model

Composite omit map with statistical density modification

Statistical density modification allows a separate probability distribution for electron density at each point in the map: can specify that "missing" density is within molecular boundary

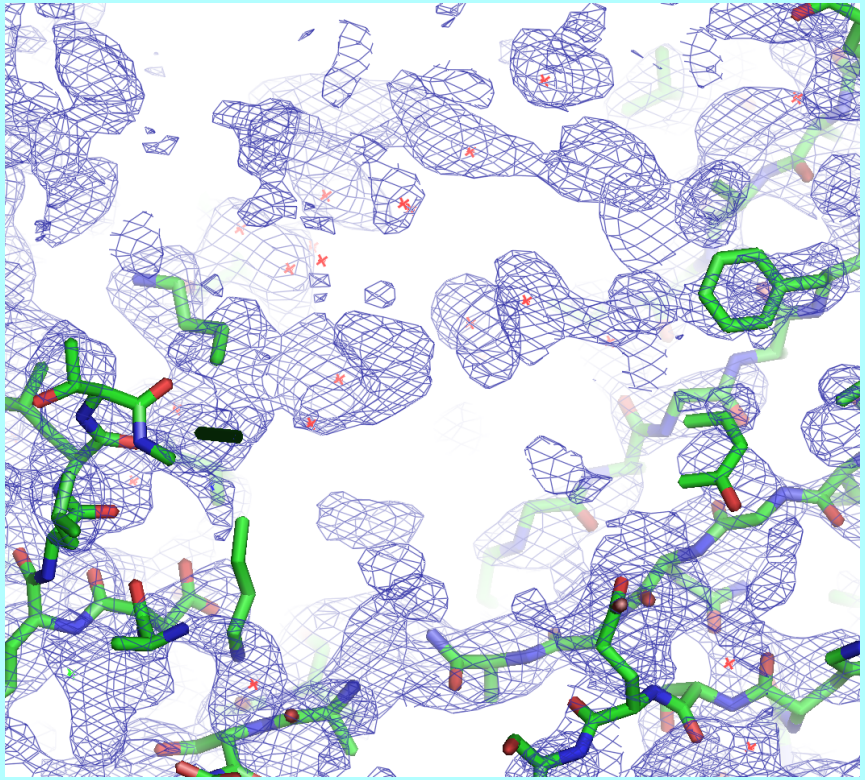


Can be used with or without experimental phases…(and with or without omit)

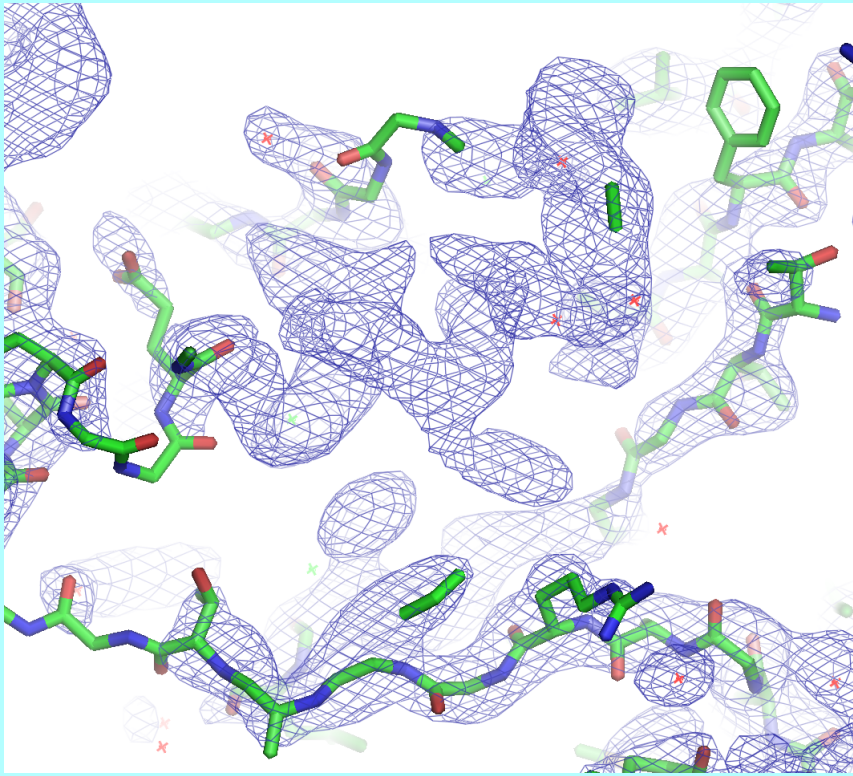
Iterative-Build OMIT procedure

"Is the density in my map biased by the model?"

2mFo-DFc omit map



After building outside OMIT region 10 cycles



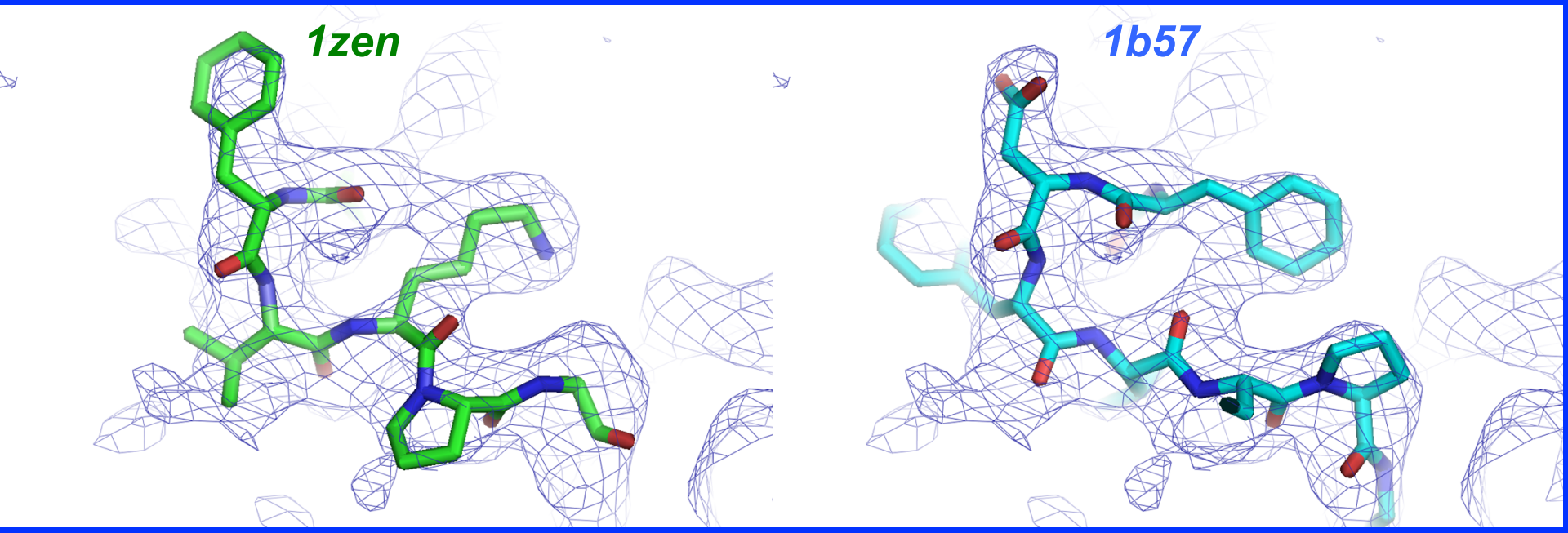
1HP7 molecular replacement with 1AS4

R/Rfree after initial refinement: 0.41/0.48

Iterative-Build OMIT procedure

"Removing model bias"

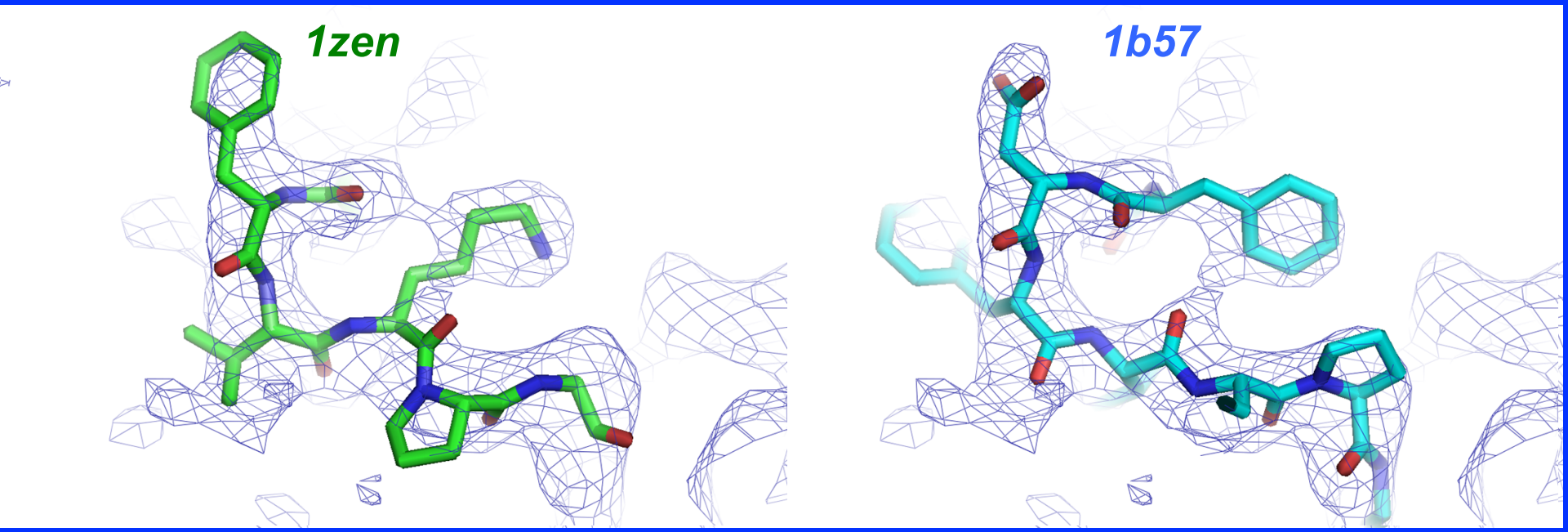
2mFo-DFc map
Phased with 1zen model



Iterative-Build OMIT procedure

"Removing model bias"

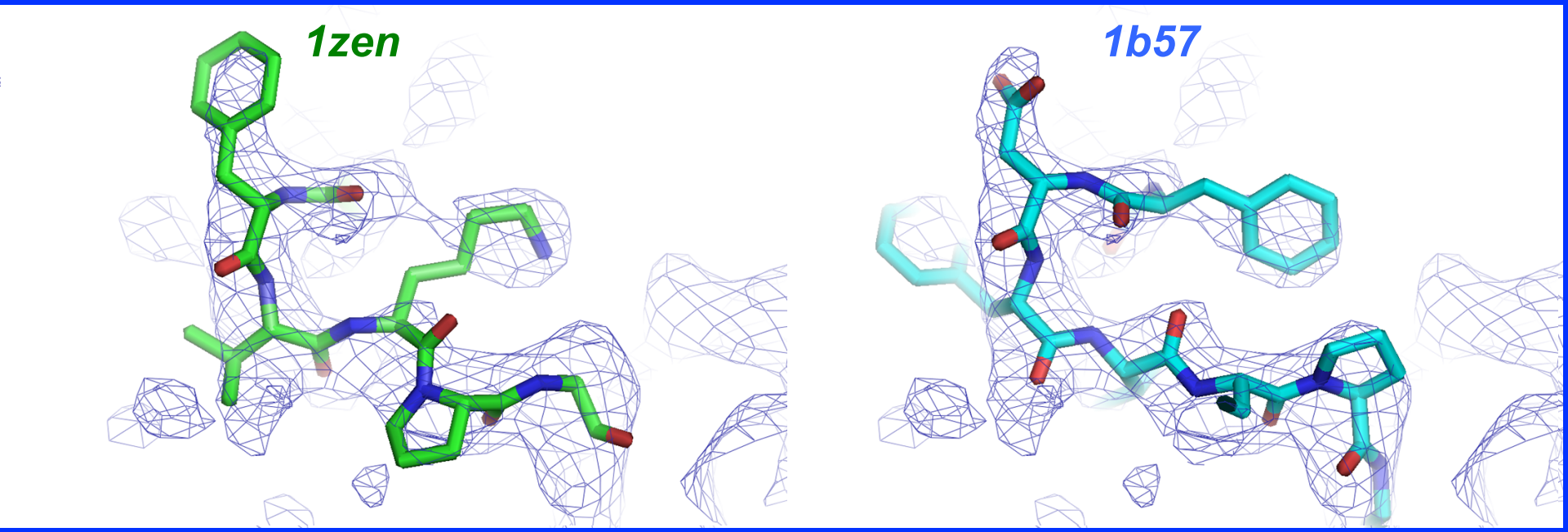
2mFo-DFc omit map
Phased with 1zen model



Iterative-Build OMIT procedure

"Removing model bias"

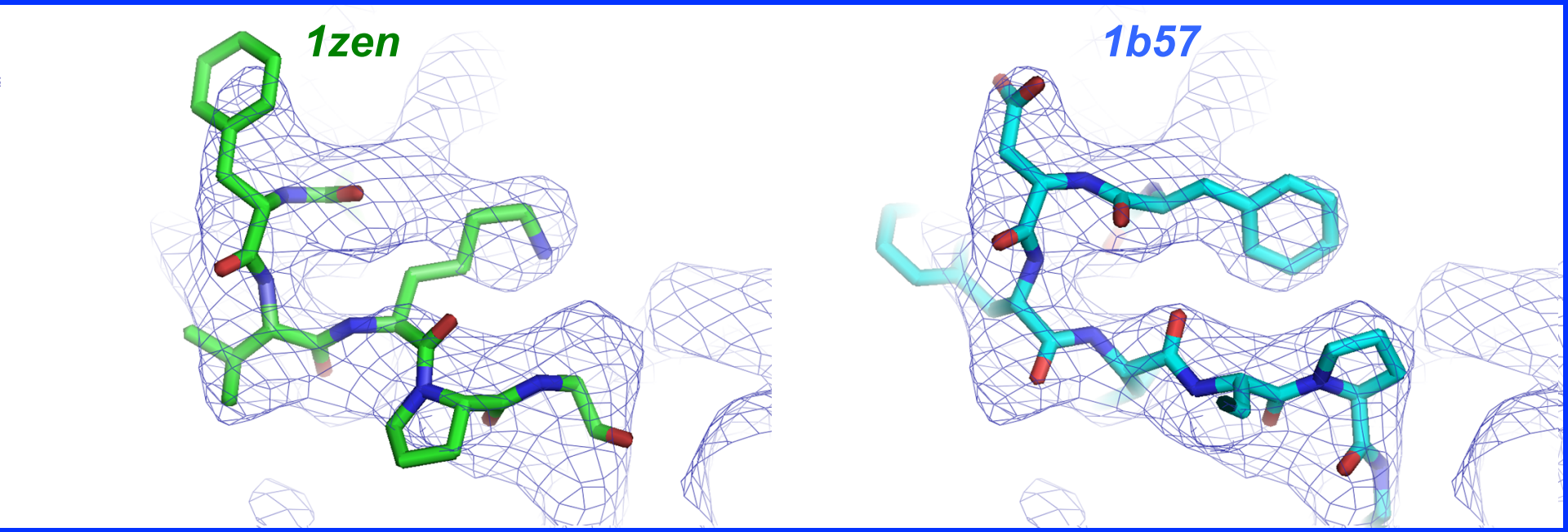
2mFo-DFc SA-omit map
Phased starting with 1zen model



Iterative-Build OMIT procedure

"Removing model bias"

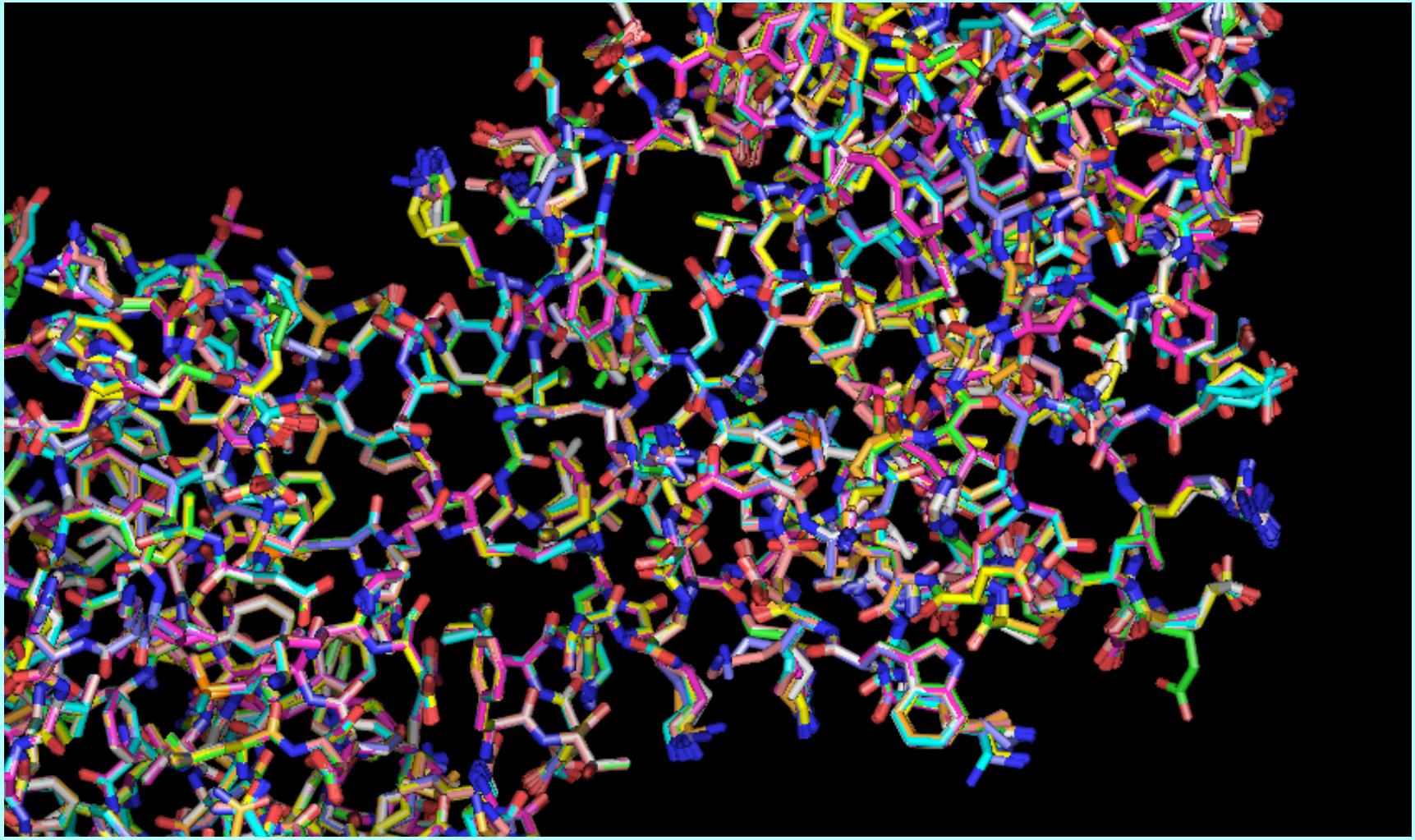
2mFo-DFc iterative-build omit map
Phased starting with 1zen model



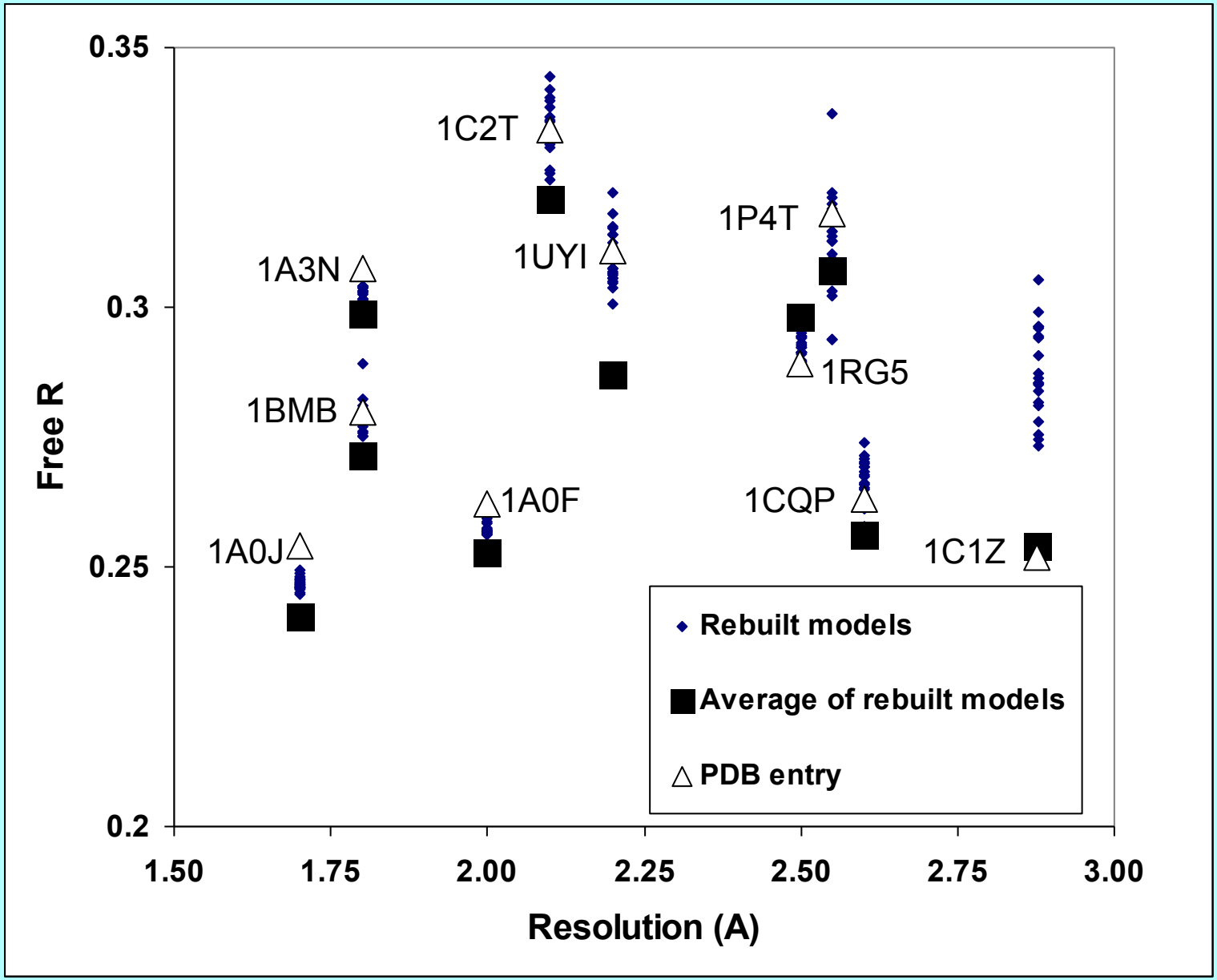
Multiple-model representation of uncertainties

20 models built for 1CQP, no waters, Dmin=2.6 A R=0.19-0.20; Rfree=0.26-0.27

The variation among models is a lower bound on their uncertainty

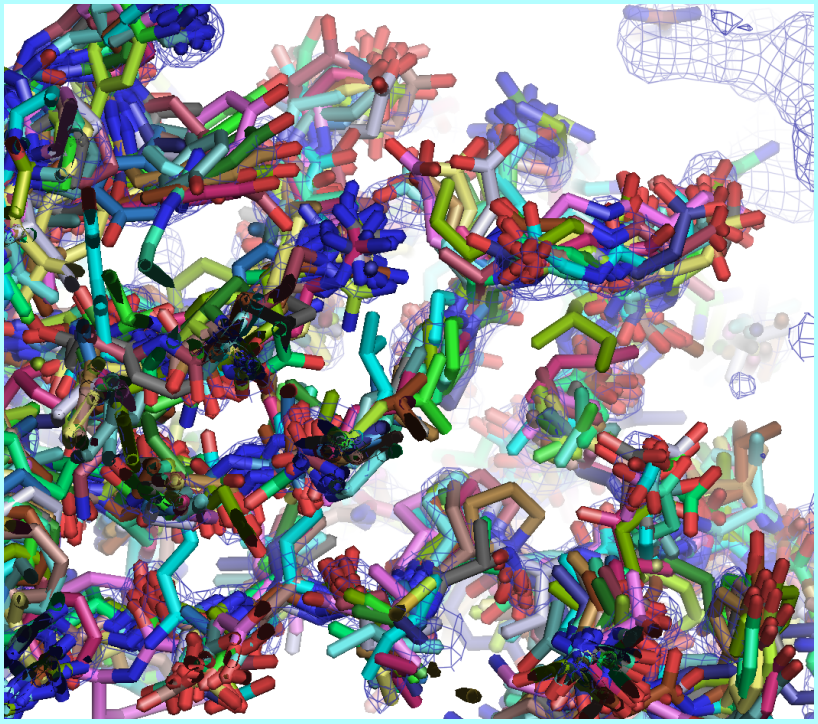


Building 20 models for each of 10 structures

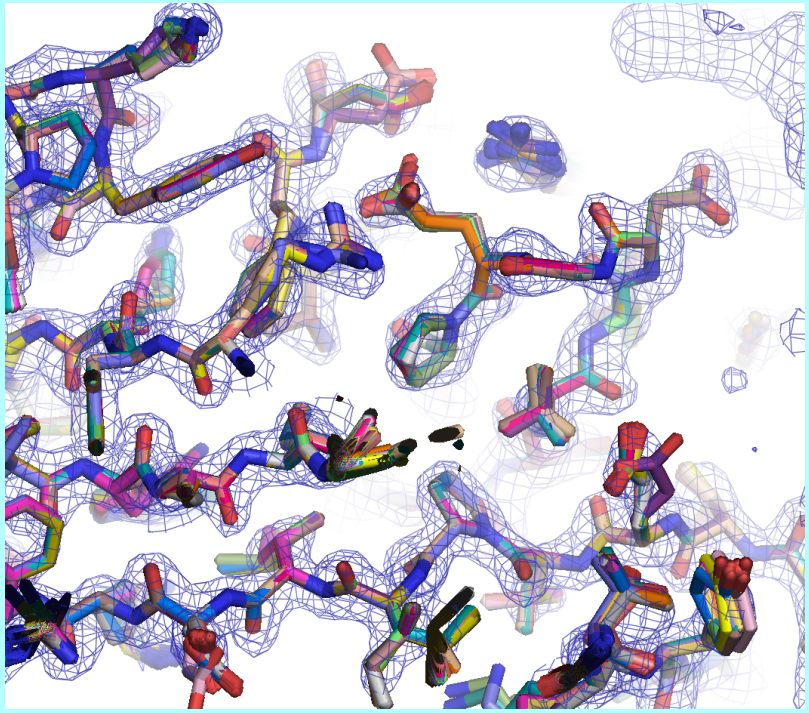


->The RMSD among models tells us (a lower bound on) the uncertainty in our models

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



Rebuild with 4.5 Å data



Rebuild with 1.75 Å data

Wizards

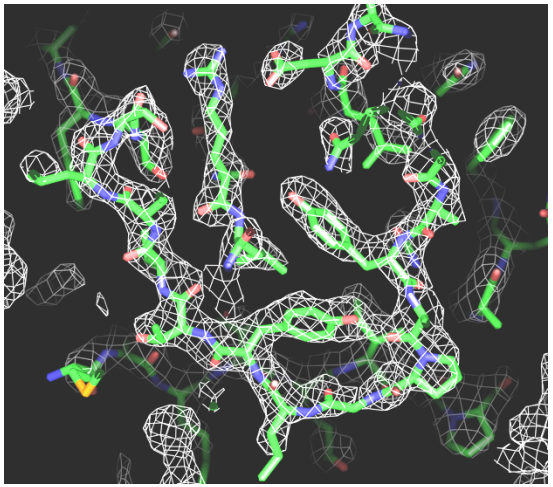
Phenix

•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

- 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

- 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, 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, Jeff Headd, Chris Williams, Bryan Arendall, Laura Murray

An NIH/NIGMS funded Program Project