

PHENIX Wizards and Tools



Tom Terwilliger

Los Alamos National Laboratory terwilliger@lanl.gov

The PHENIX project

Computational Crystallography Initiative (LBNL)

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



Los Alamos National Lab (LANL)

Tom Terwilliger, Li-Wei Hung



Cambridge University

*Randy Read, Airlie McCoy, Gabor Bunkoczi,
Rob Oeffner*



Duke University

Jane Richardson, David Richardson, Jeff Headd, Vincent Chen



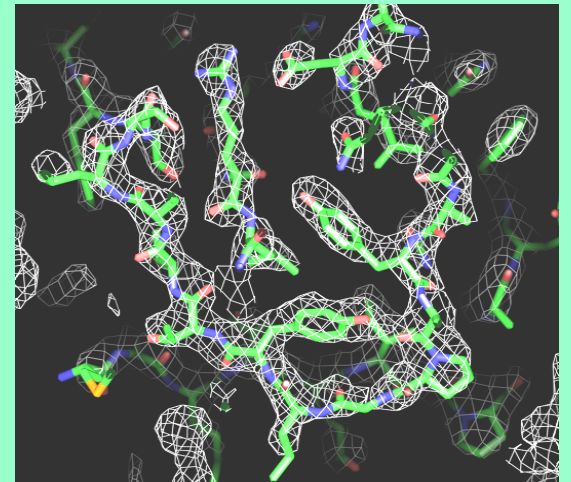
Texas A&M University

Tom Ioerger, Jim Sacchettini



PHENIX 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



Determining a SAD structure with *P*HENIX

- Solve the structure: `phenix.autosol sad.sca 12 se`
- AutoBuild a model and improve phases:
`phenix.autobuild after_autosol=true`
- Find ligands:
`phenix.ligandfit sad.sca model=partial.pdb ligand=ATP`
- Refine the model carefully:
`phenix.refine exptl_fobs_freeR_flags.mtz \`
`overall_best.pdb #and many more commands`

Why automate structure determination?

Automation...

makes straightforward cases accessible to a wider group of structural biologists

makes difficult cases more feasible for experts

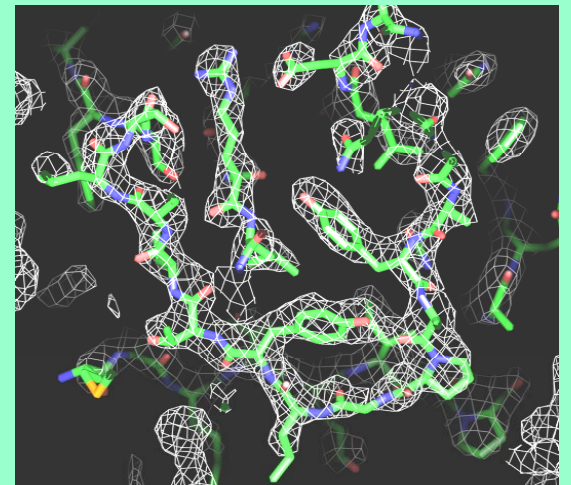
can speed up the process

can help reduce errors

Automation also allows you to...

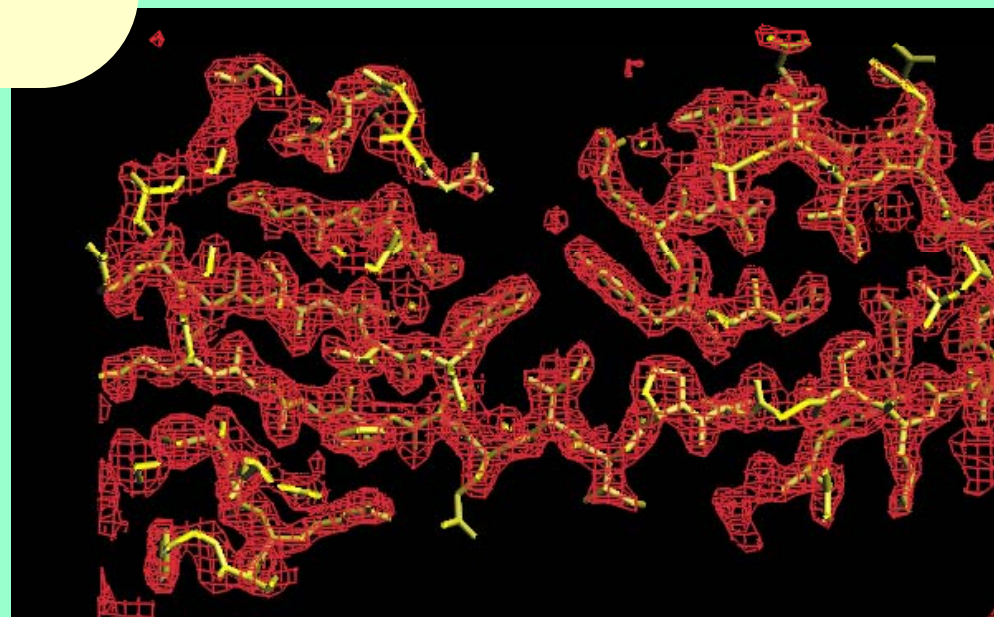
try more possibilities

estimate uncertainties



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

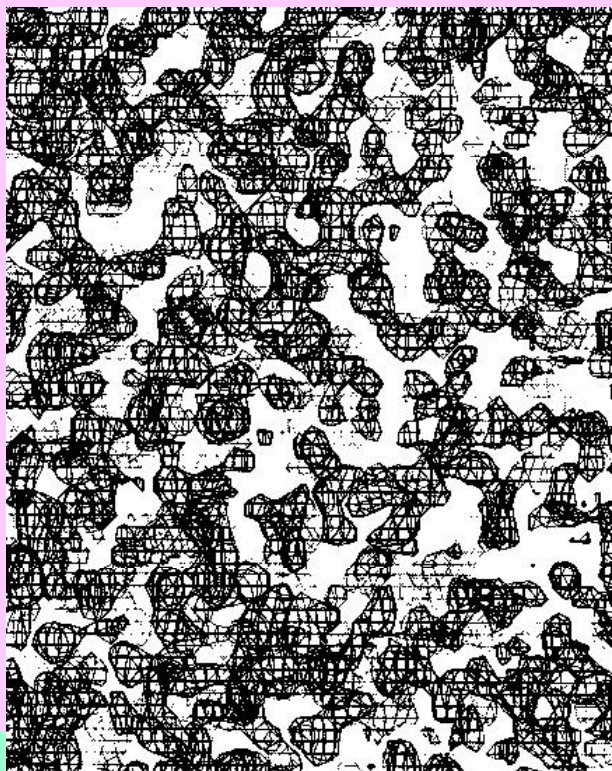
- (1) **Software carrying out individual steps**
- (2) **Seamless connections between steps**
- (3) **A way to decide what is good**
- (4) **Strategies for structure determination and decision-making**



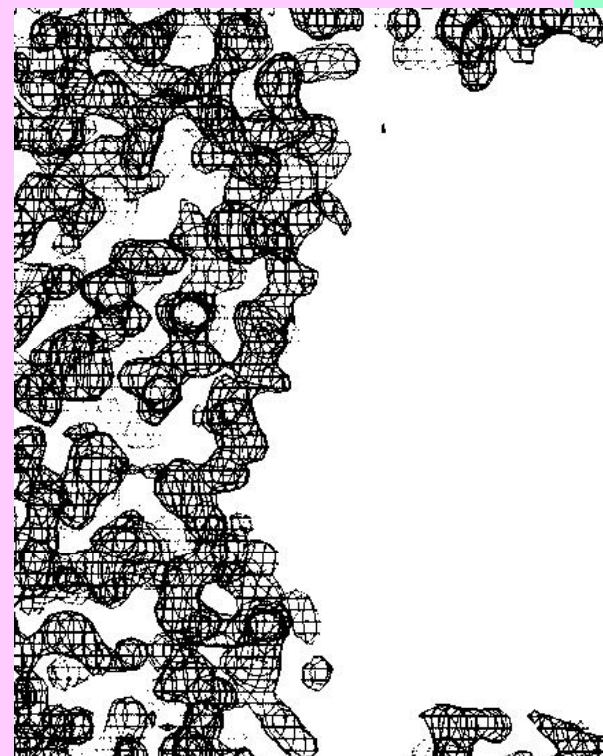
Why we need good measures of the quality of an electron-density map:

Which solution is best?

Are we on the right track?



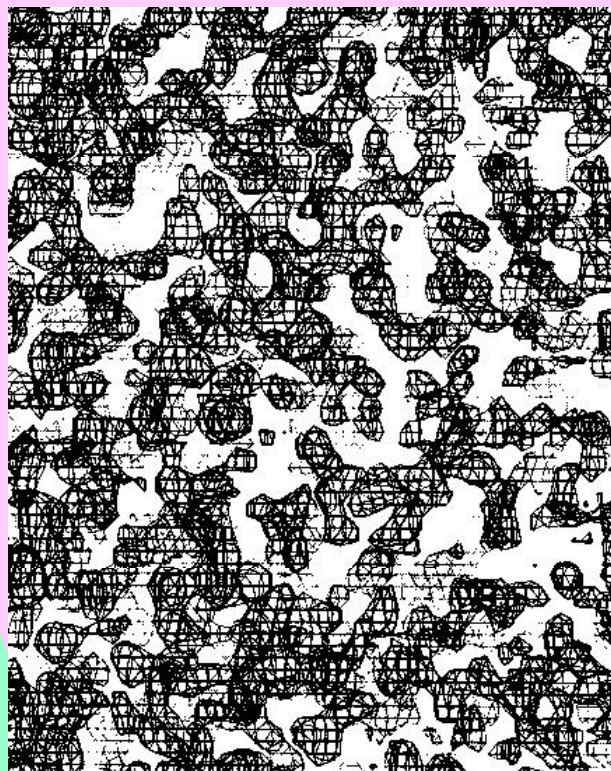
**If map is good:
It is easy**



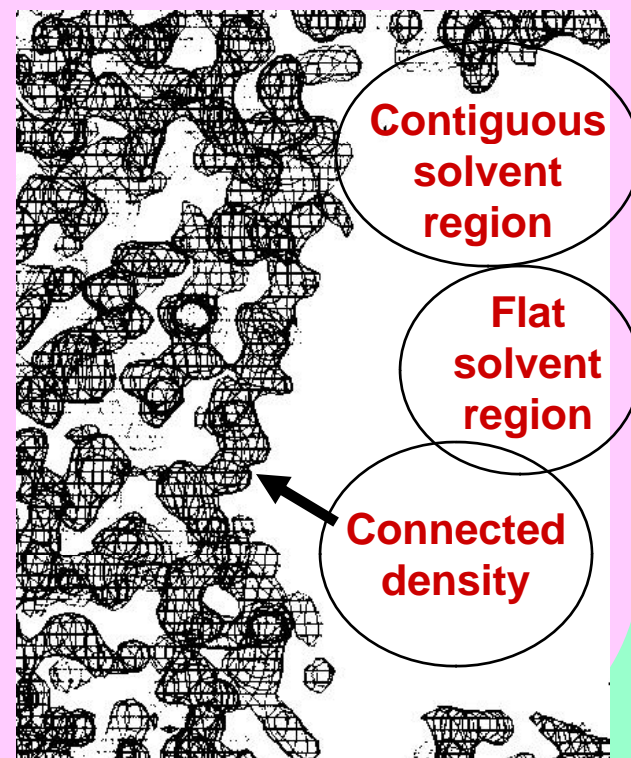
Why we need good measures of the quality of an electron-density map:

Which solution is best?

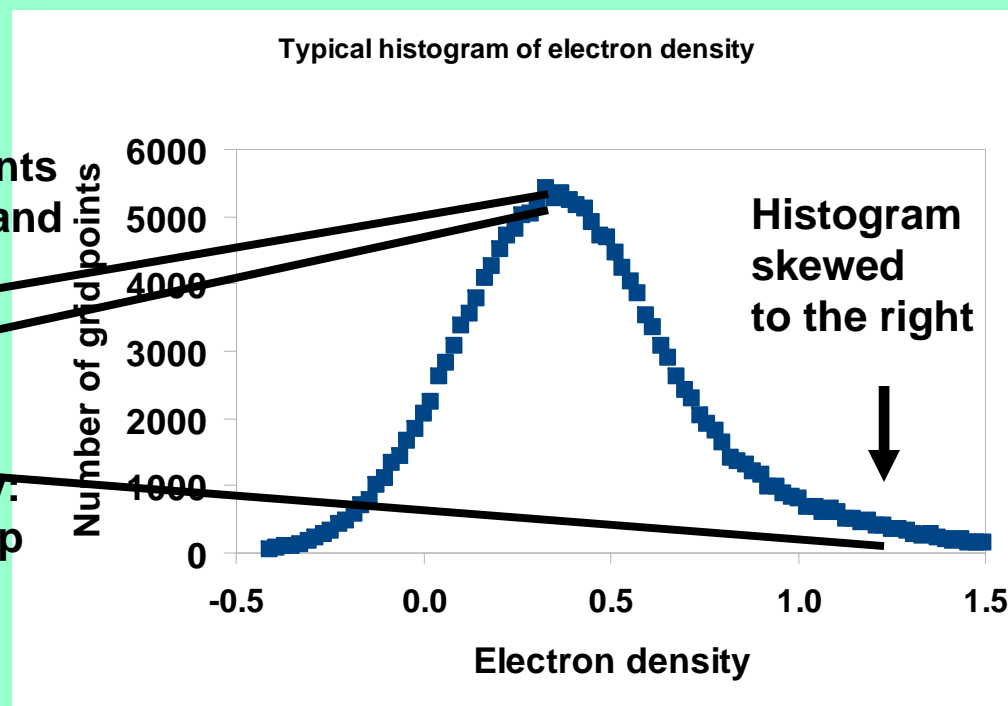
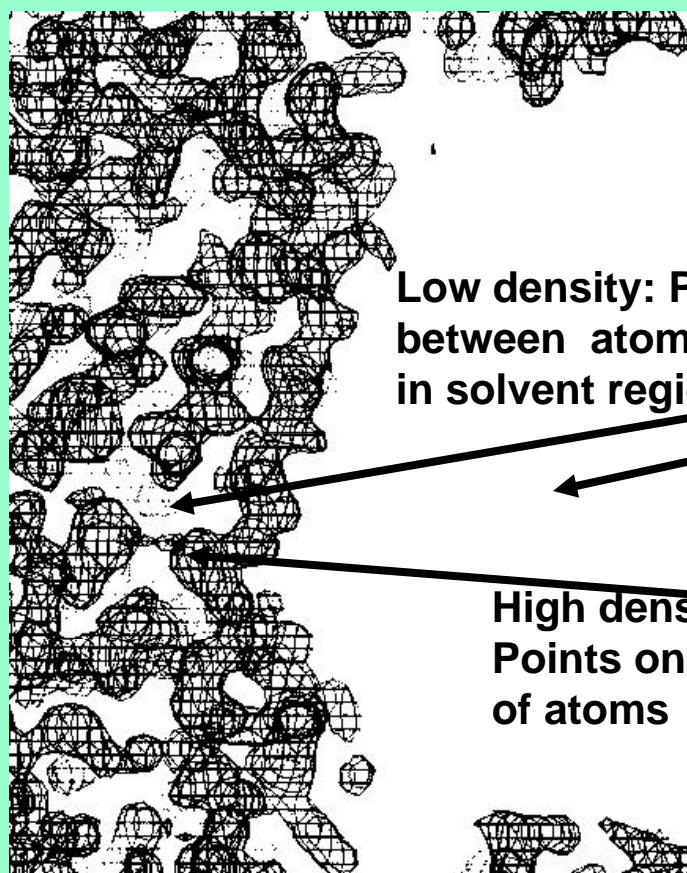
Are we on the right track?



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



Histogram of electron density values has a positive “skew”



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)**

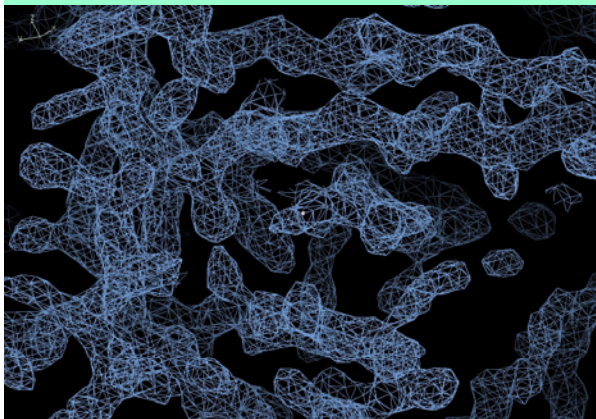
Run AutoSol Wizard on each dataset.

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

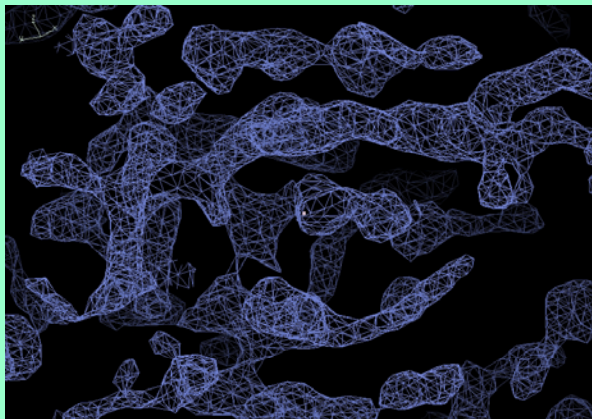
Score maps based on each criteria

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

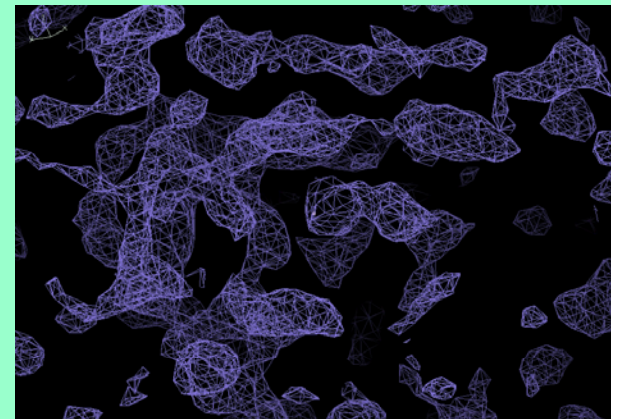
Model map
1VQB, 2.6 Å, SG C2

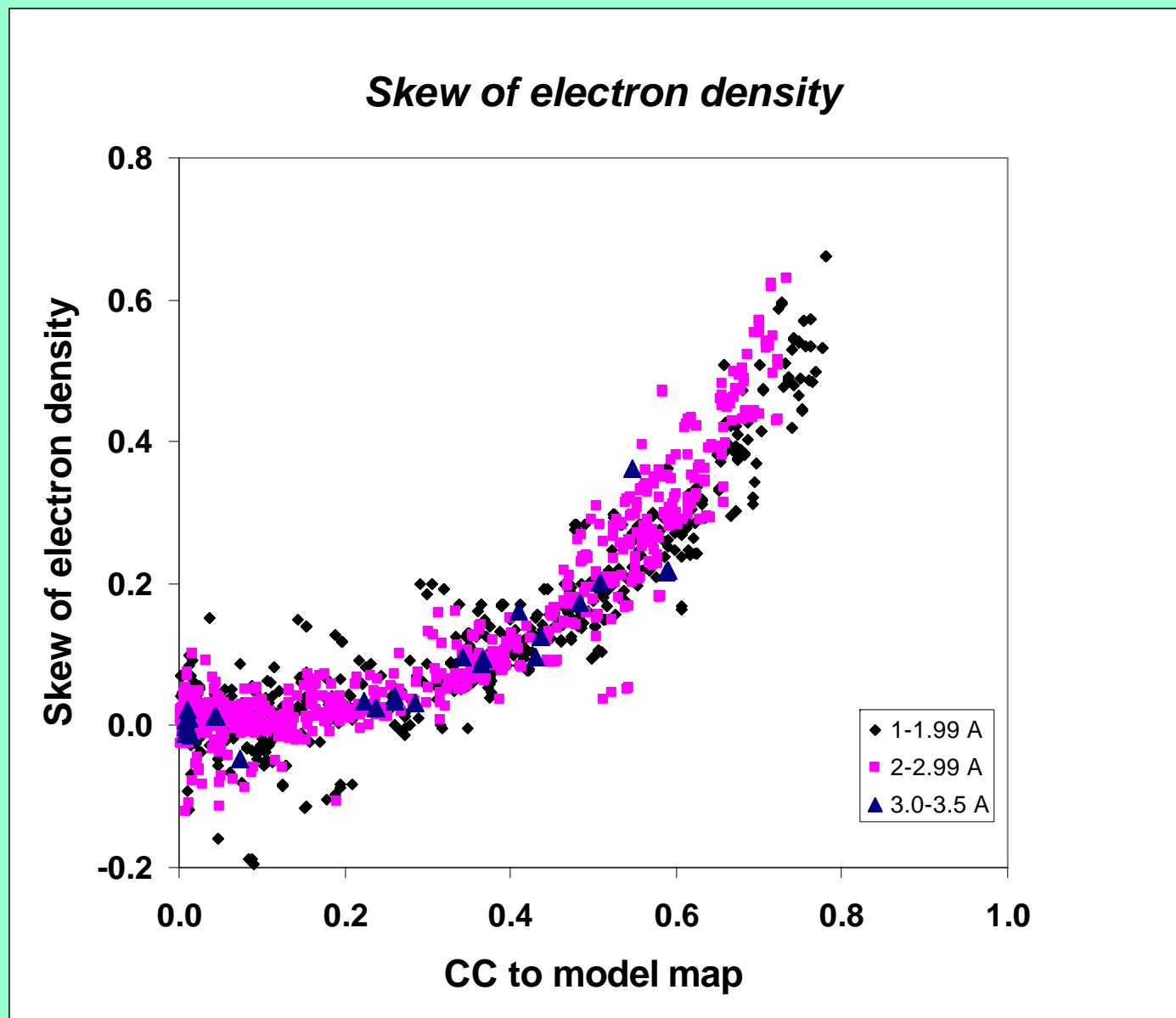


SOLVE MAD map
CC=0.62

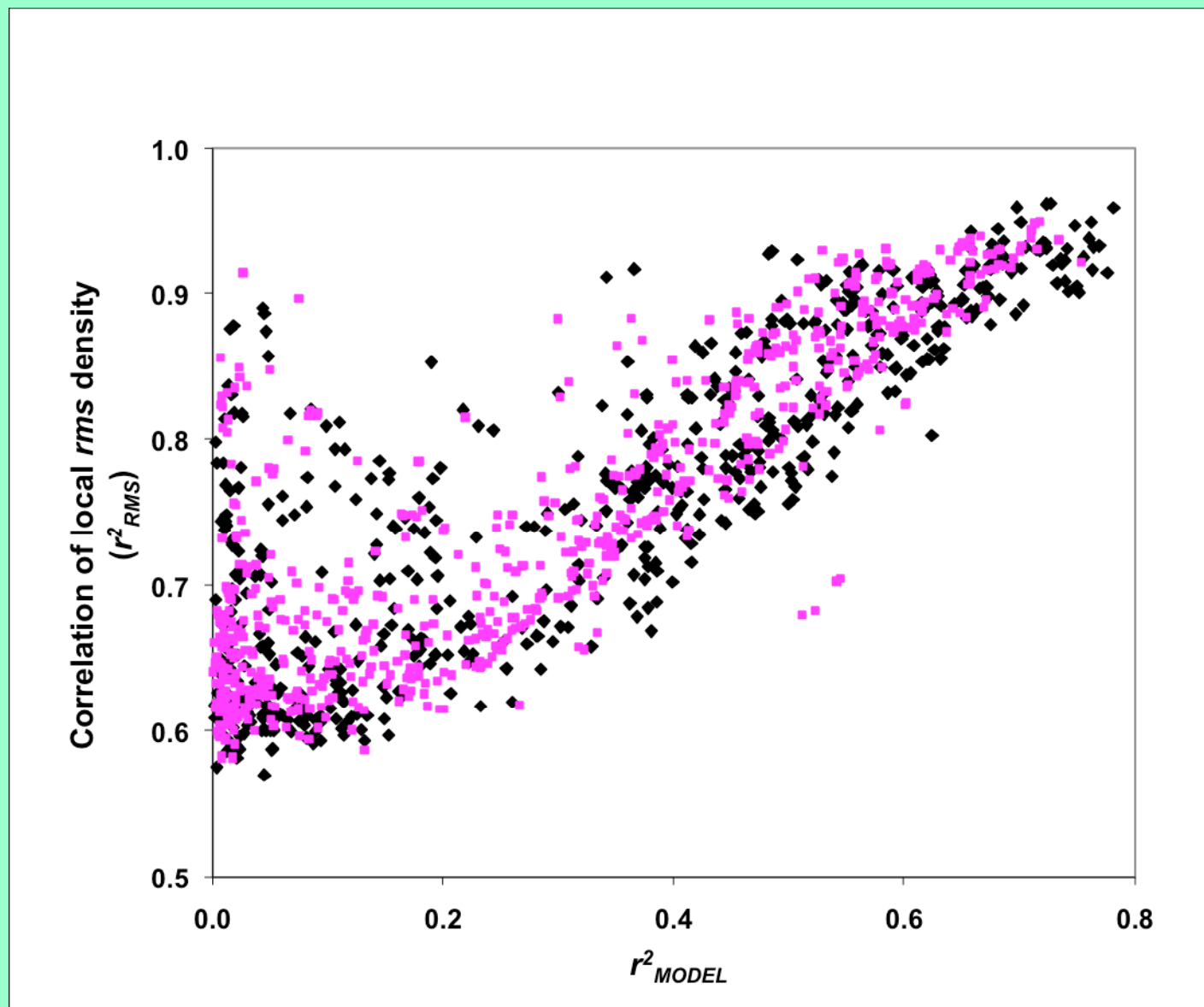


Inverse-hand map
CC=0.55



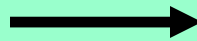


Correlation of local RMS density
(Solvent next to solvent, protein next to protein)

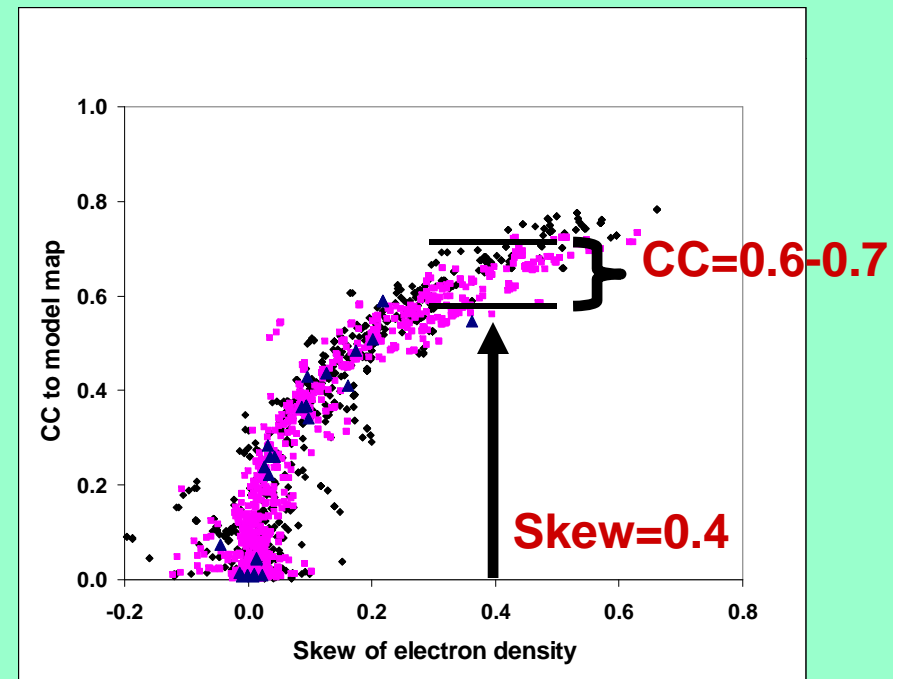
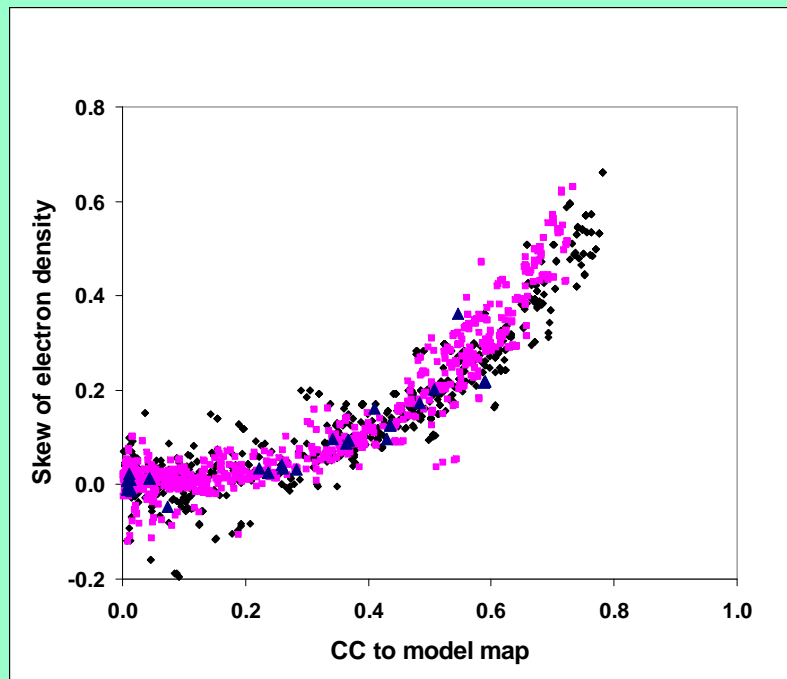


Using scoring criteria to estimate the quality of a map

Skew depends on CC

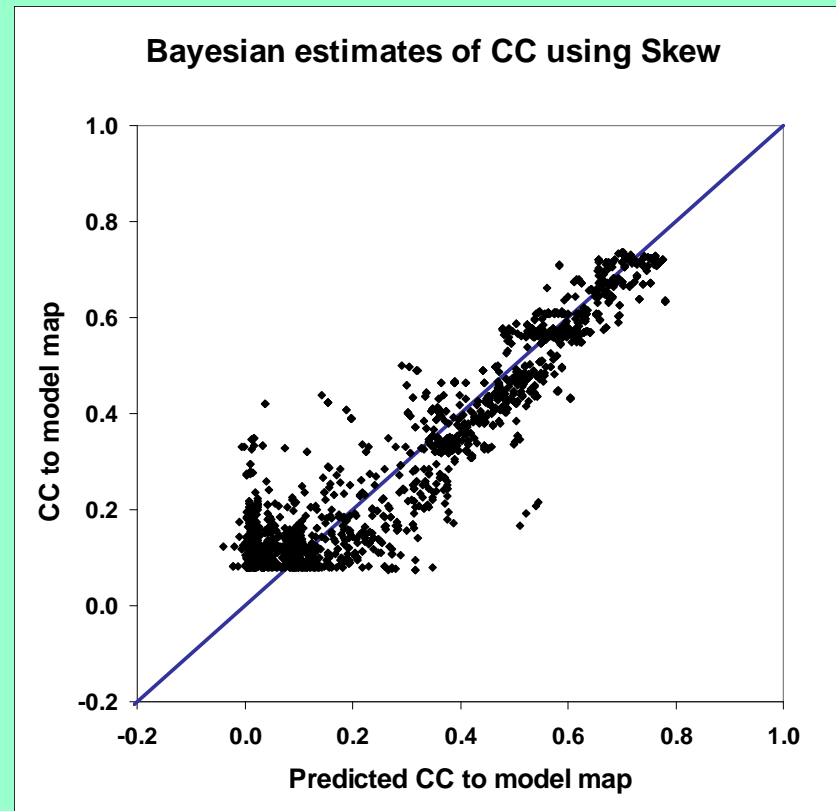


Estimate CC from skew



How accurate are estimates of map quality?

**Actual
quality**



Estimated quality

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 \pm 0.04	0.72 ←
Peak (inverse hand)	12	0.11 \pm 0.43	0.04
F_A	12	0.73 \pm 0.03	0.72
F_A (inverse)	12	0.11 \pm 0.42	0.04
Sites from diff Fourier	9	0.70 \pm 0.17	0.69

*What to do next: Follow up on all the solutions that MIGHT
be the best (within 2 SD of the top)*

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”

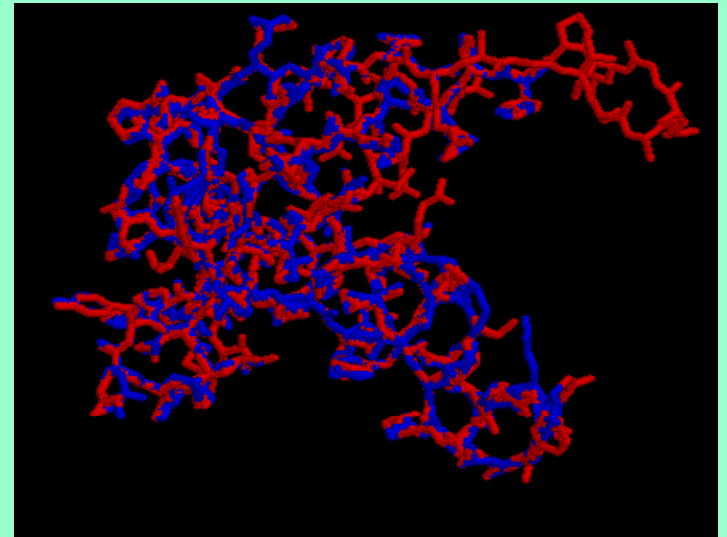
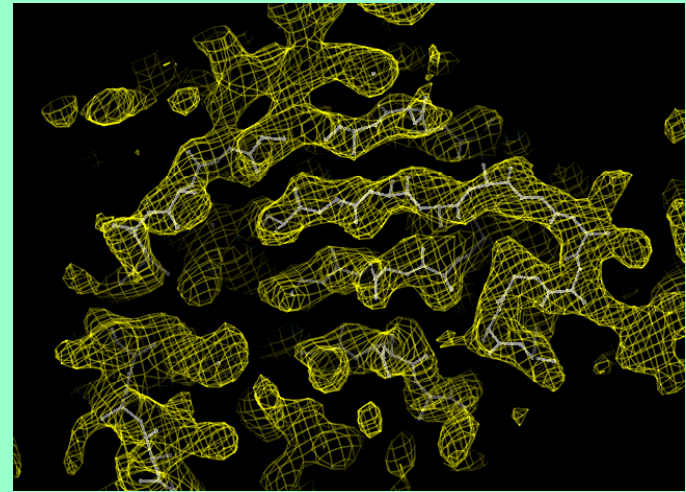
- A believable map is a map that has...

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

- A distribution of densities like those of model proteins

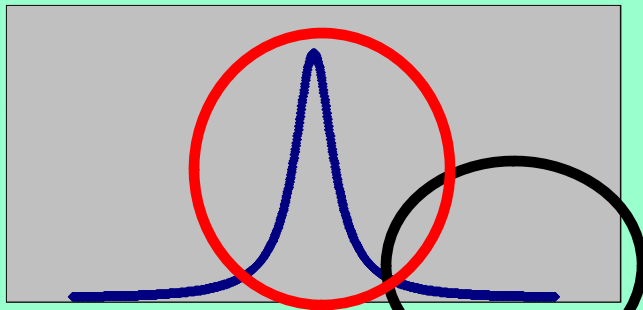
- Method:

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

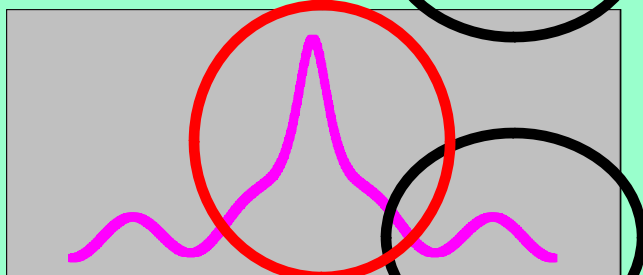


Map probability phasing: Getting a new probability distribution for each phase given estimates of all others

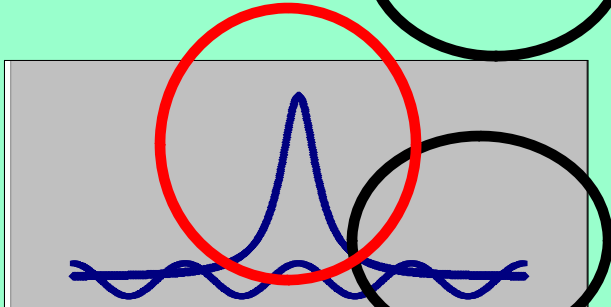
1. Identify expected features of map (flat far from center)
2. Calculate map with current estimates of all structure factors except one (k)
3. Test all possible phases ϕ for structure factor k (for each phase, calculate new map including k)
4. Probability of phase ϕ estimated from agreement of map with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



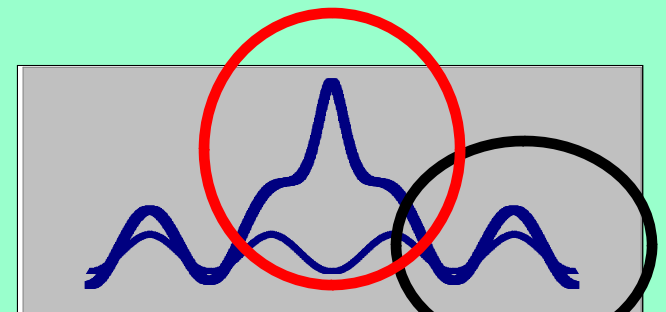
A function that is (relatively) flat far from the origin



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



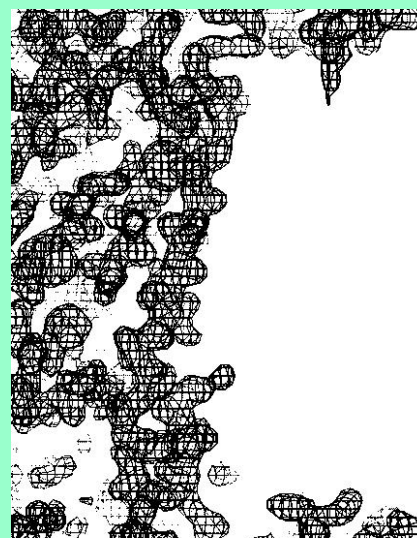
Test each possible phase of structure factor k . $P(\phi)$ is high for phase that leads to flat region



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

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

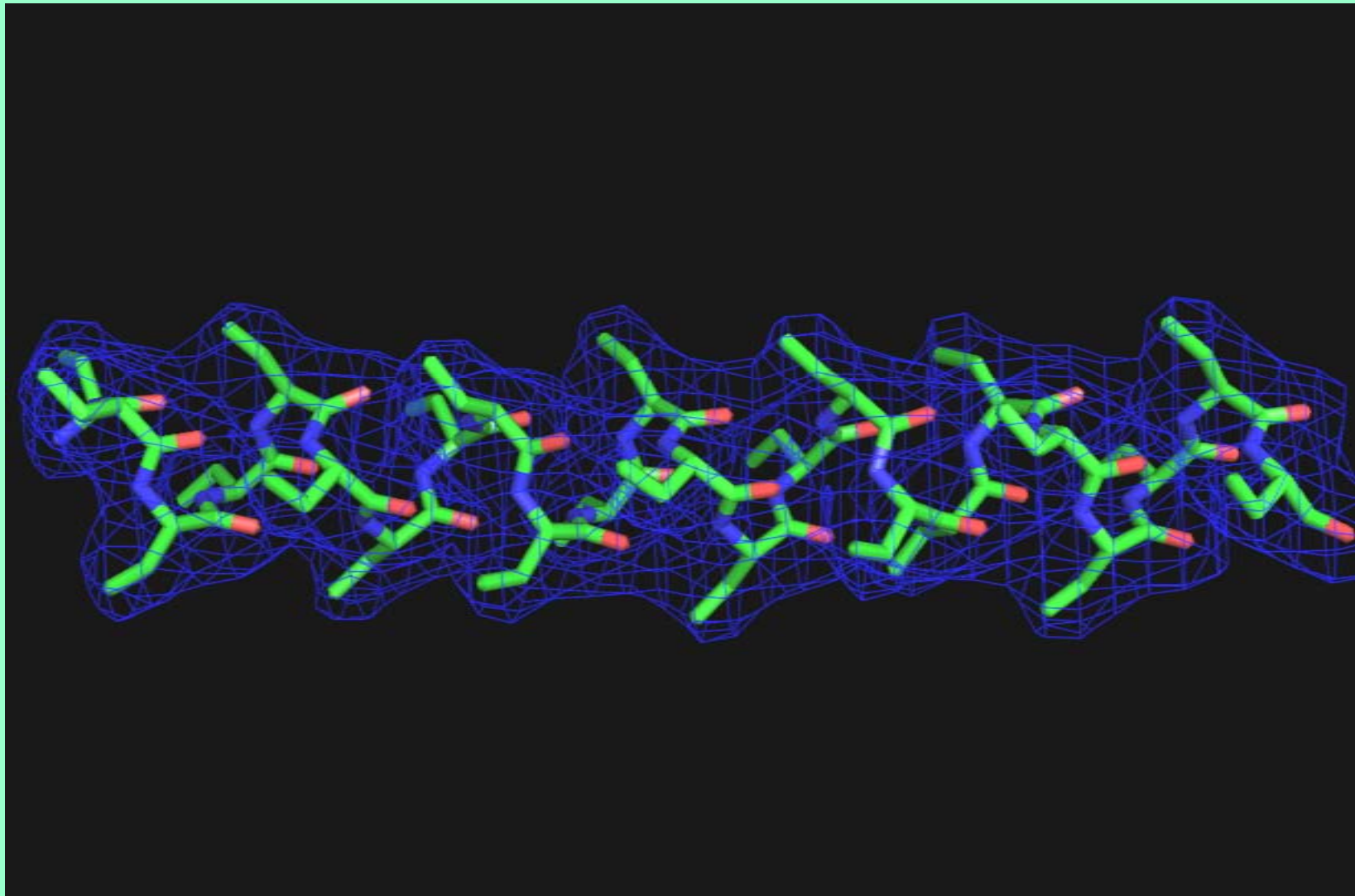
Any protein chains (trace_chain):

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

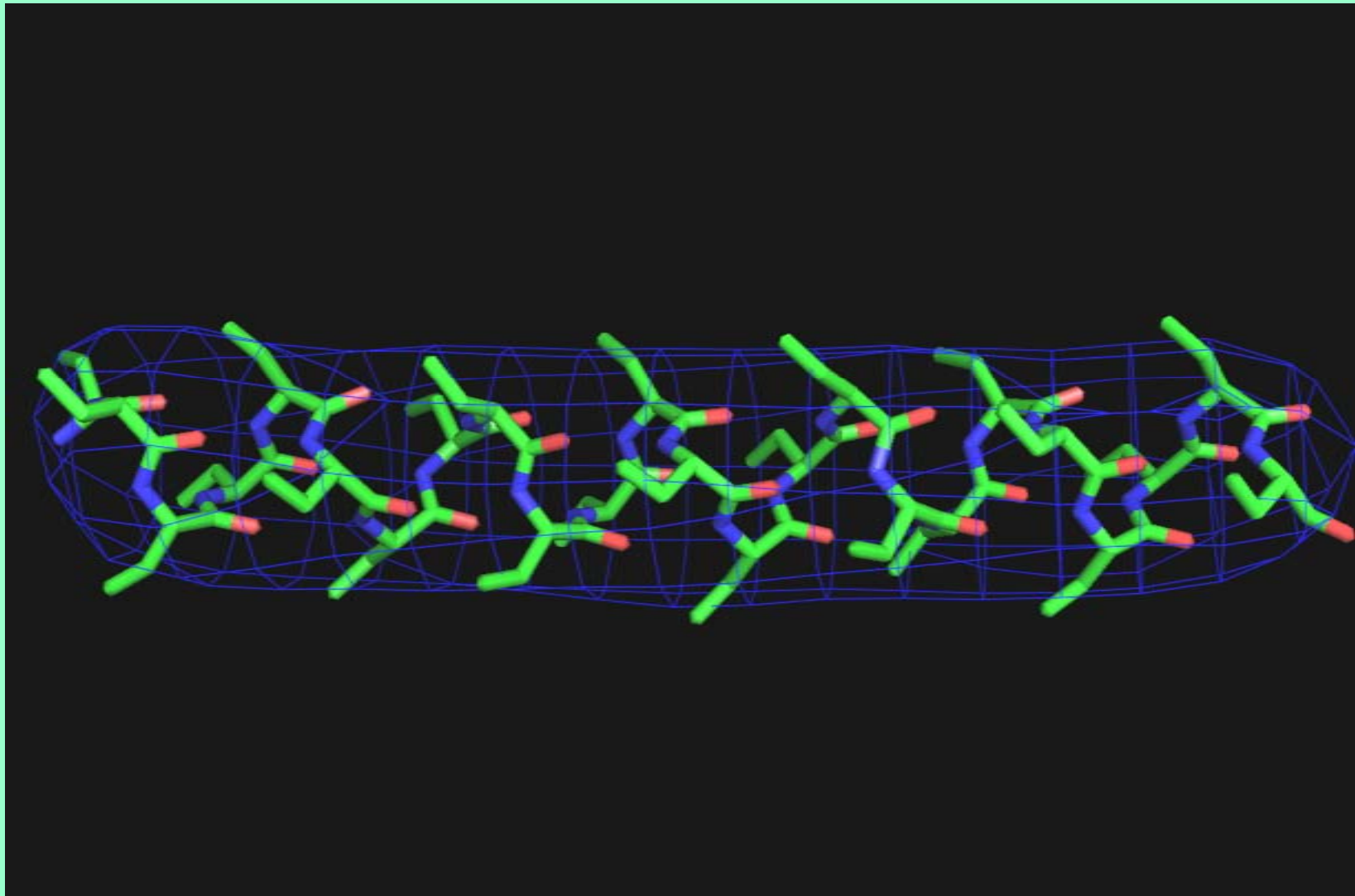
RNA/DNA:

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

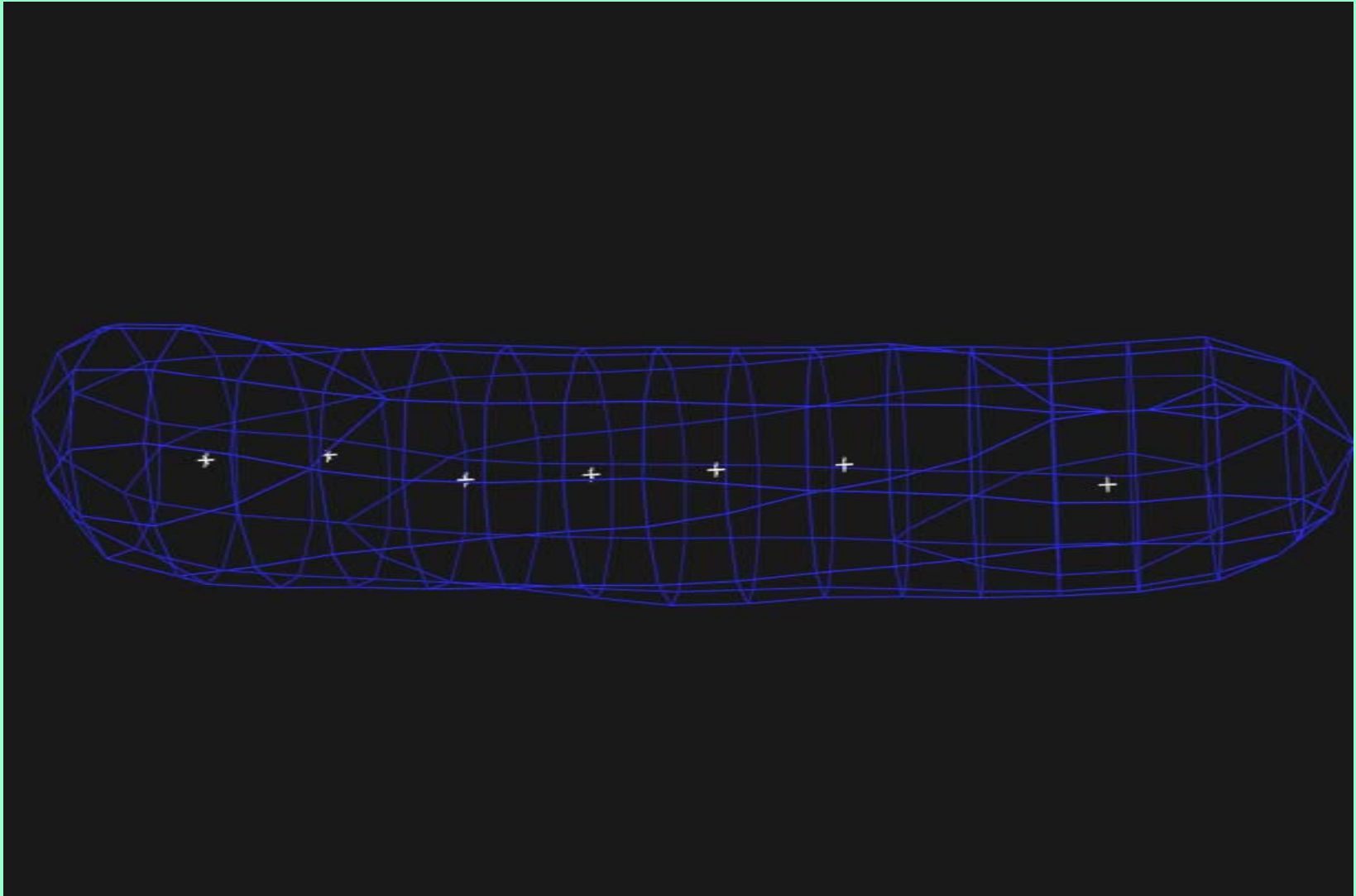
Model α -helix; 3 Å map



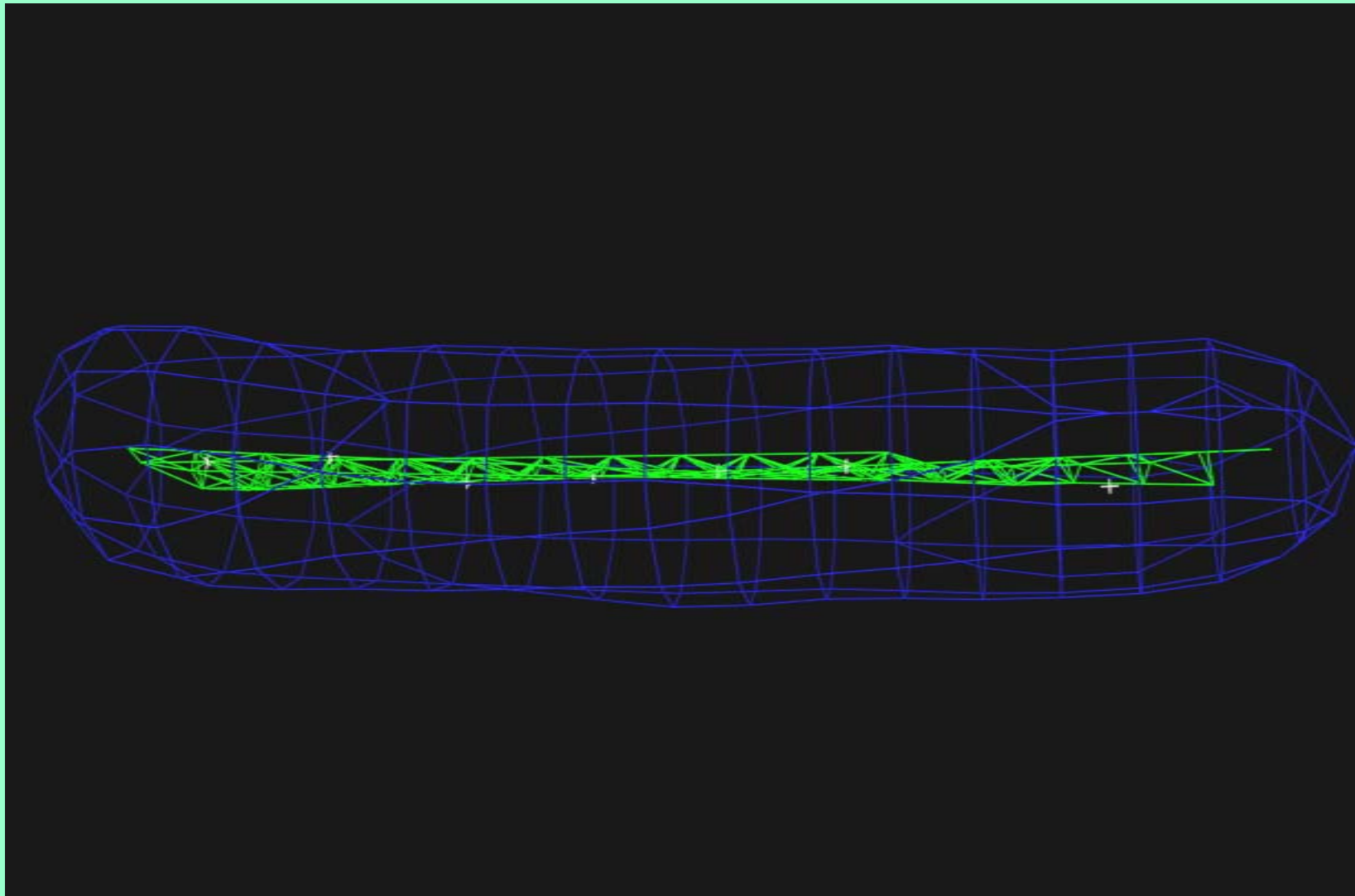
Model α -helix; 7 Å map



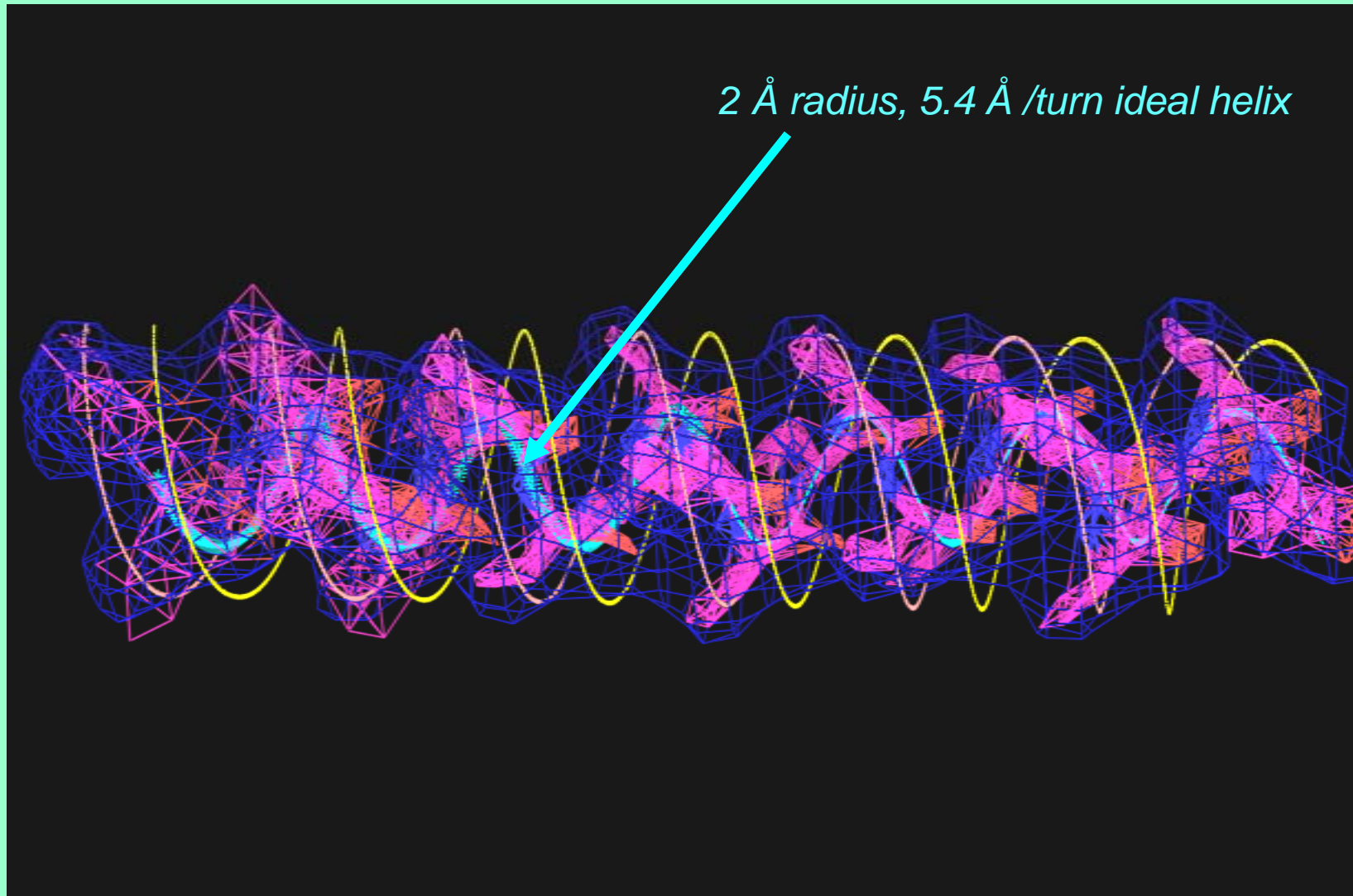
Find points along tubes of density in 7 Å map



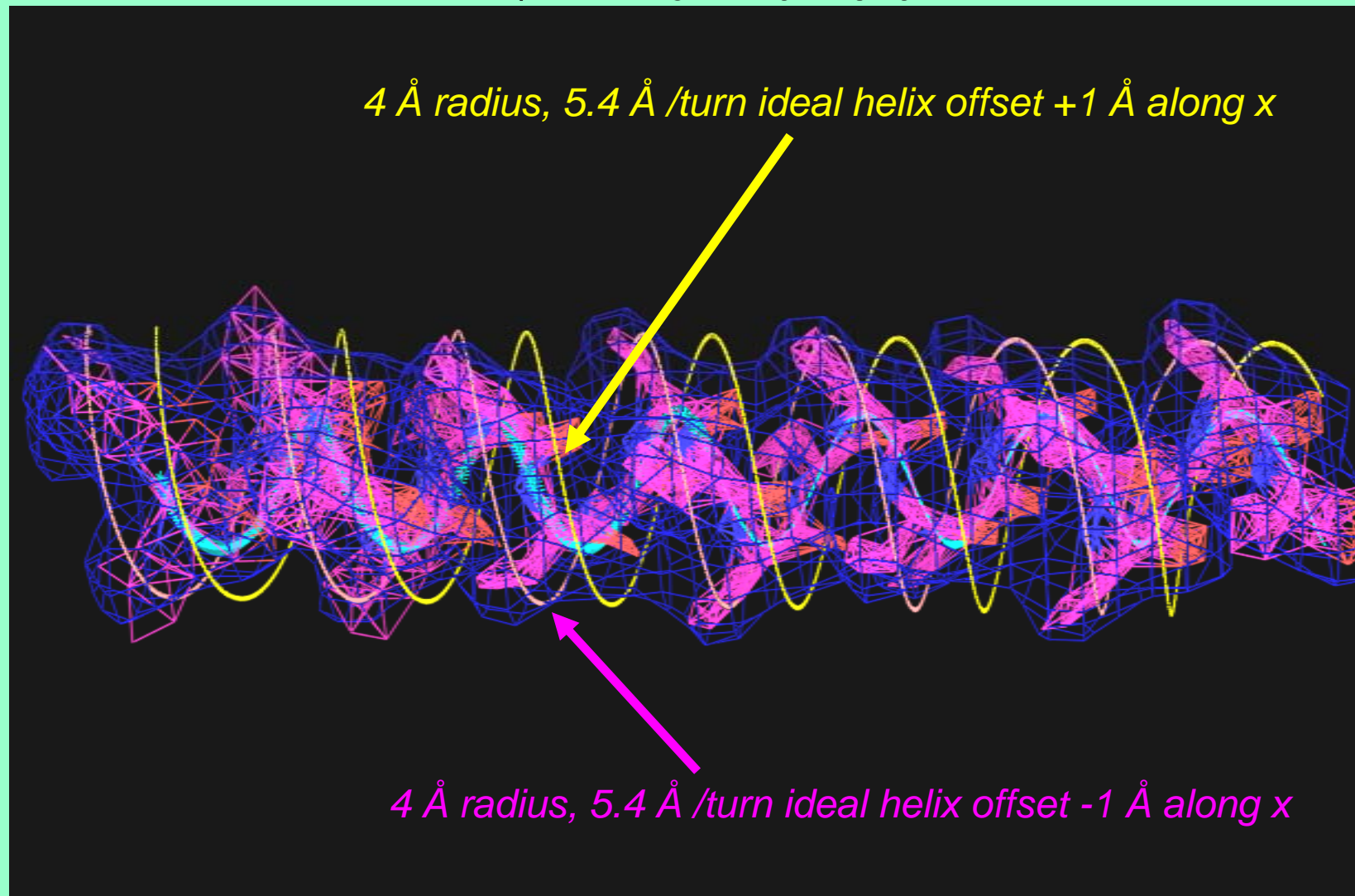
Trace along tubes of density in 7 Å map



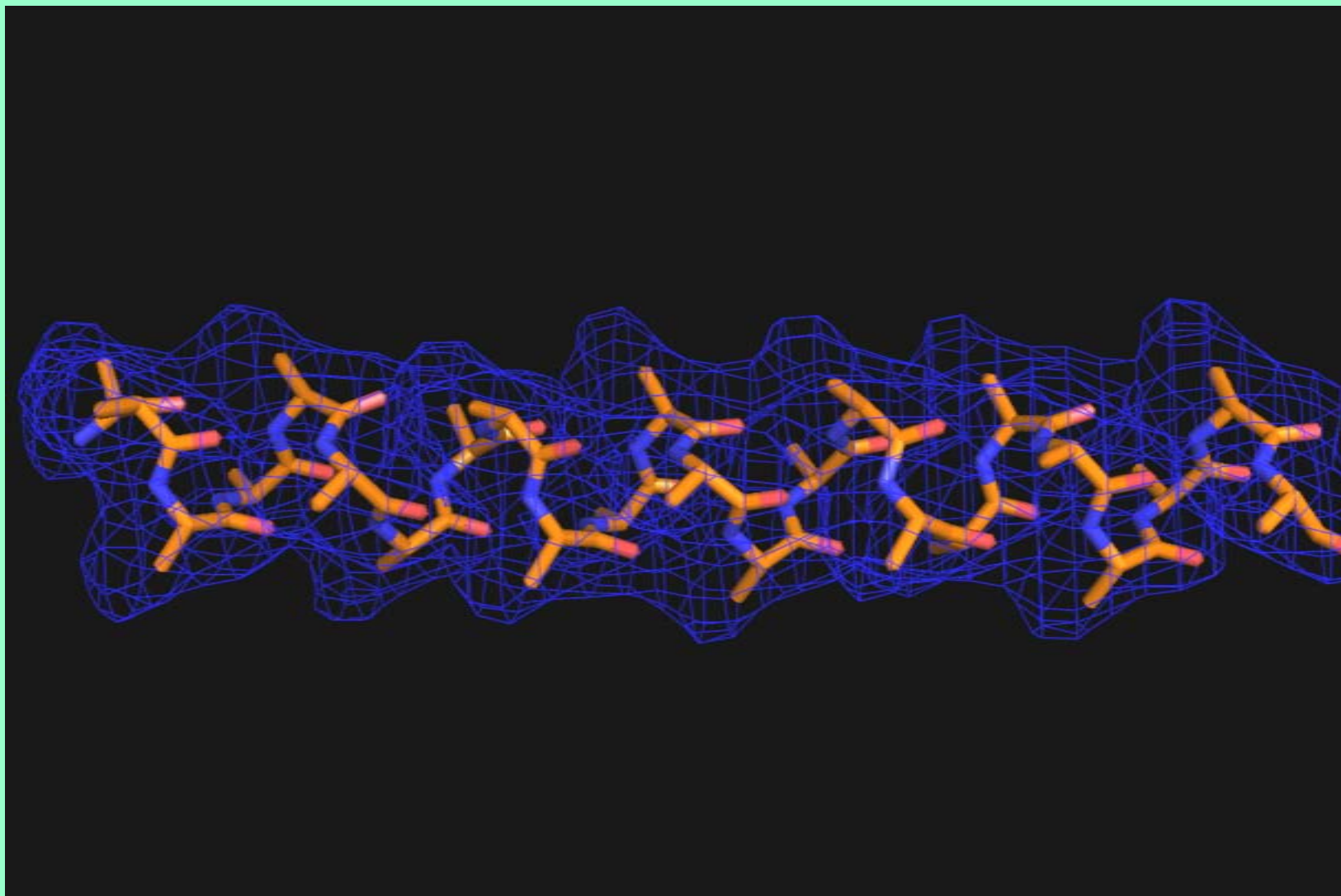
Trace main-chain with ideal helix, allowing curvature



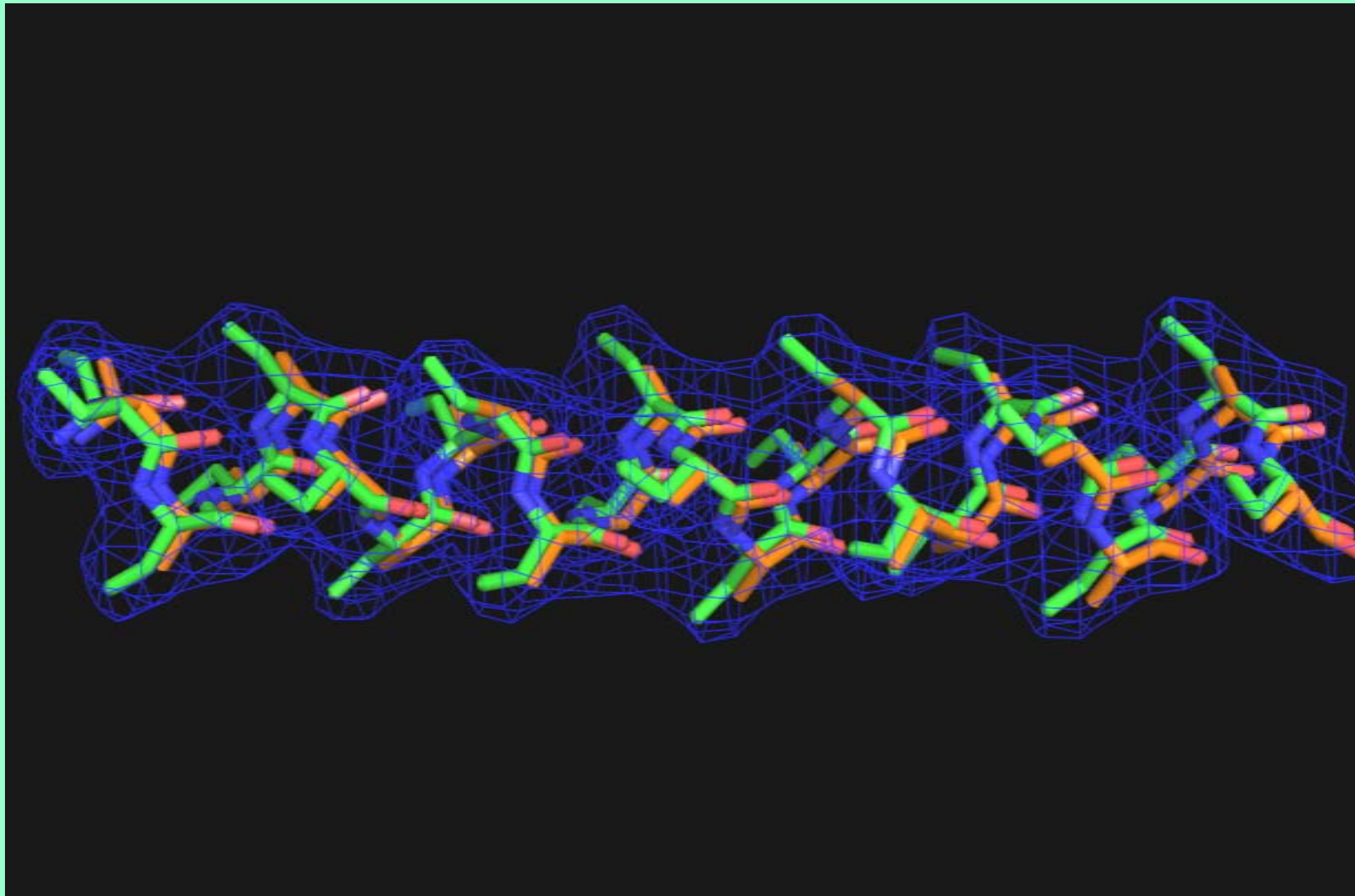
Identify direction and $C\alpha$ position from overlap with 4 Å radius helices offset ± 1 Å from main-chain



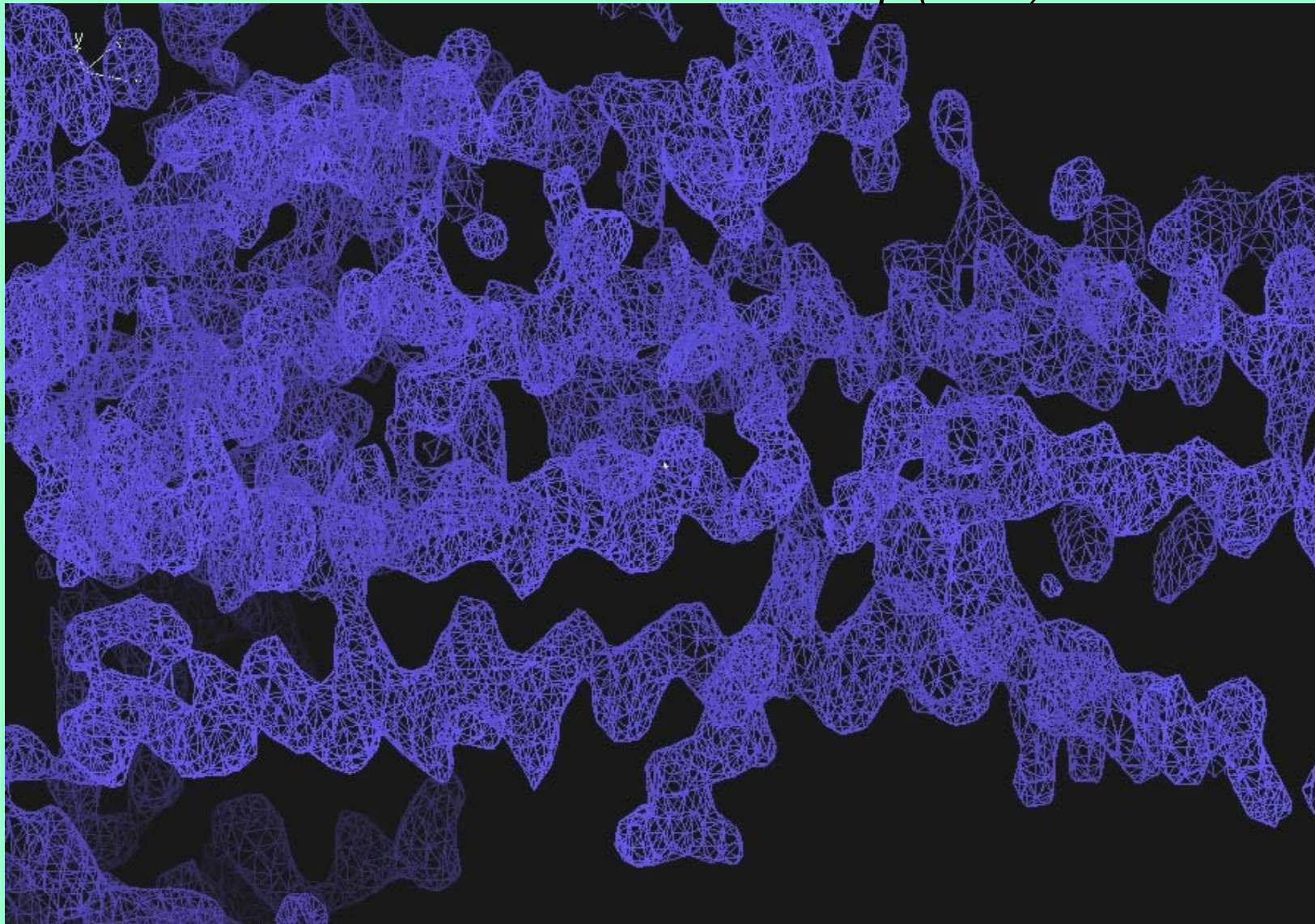
Choose best-fitting helices; link together if necessary



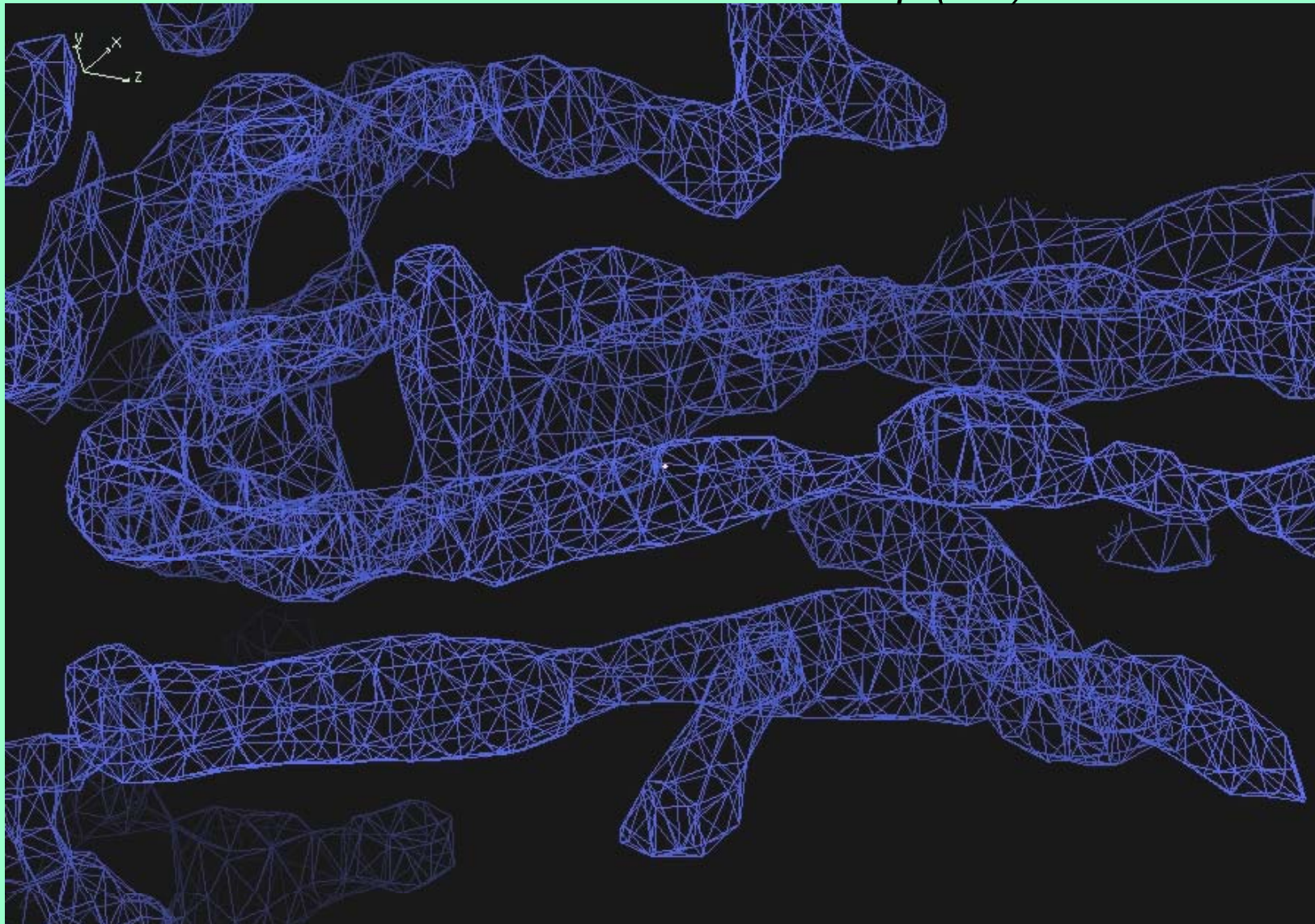
Comparison with model helix



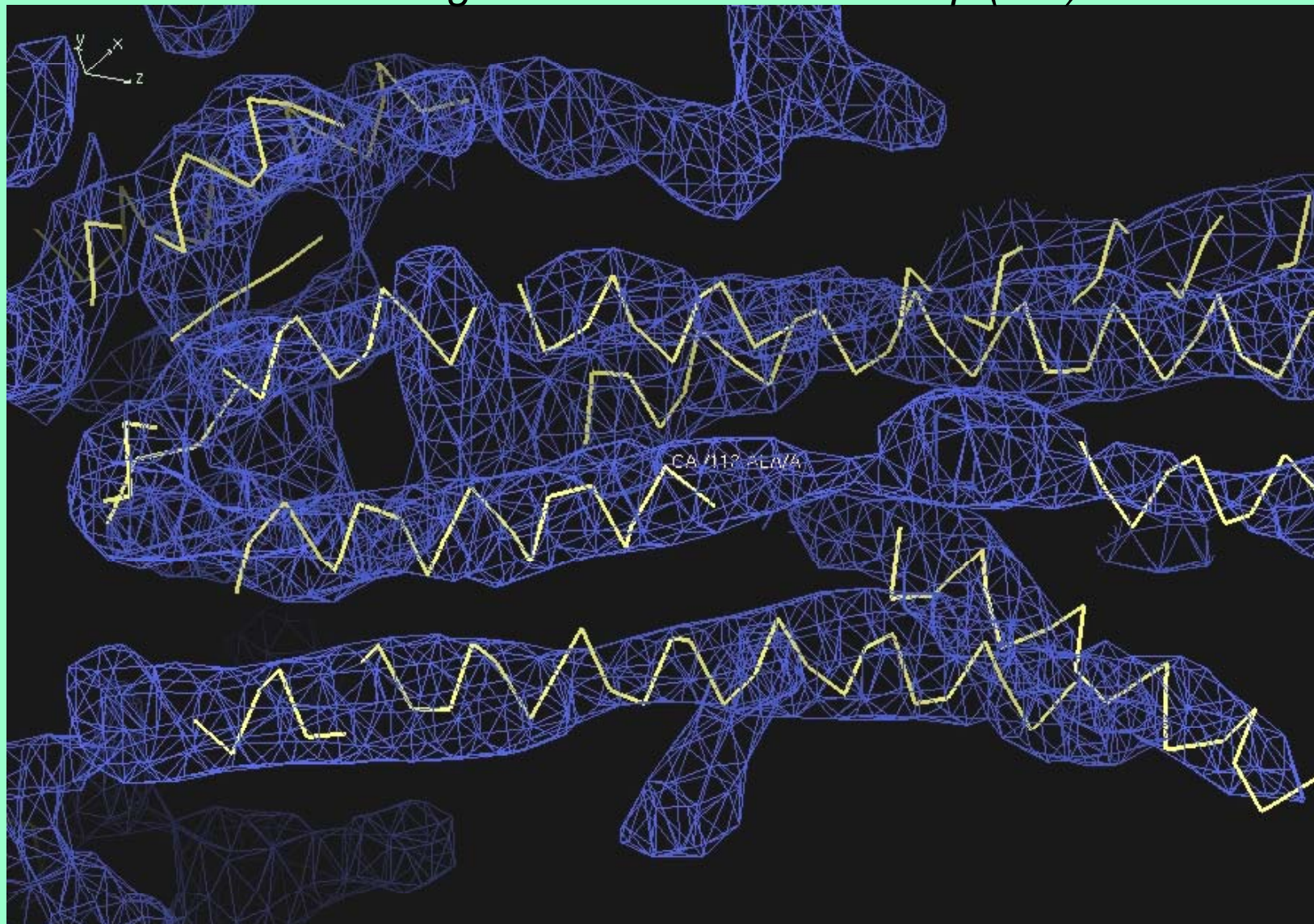
A real case: 1T5S SAD map (3.1 Å)



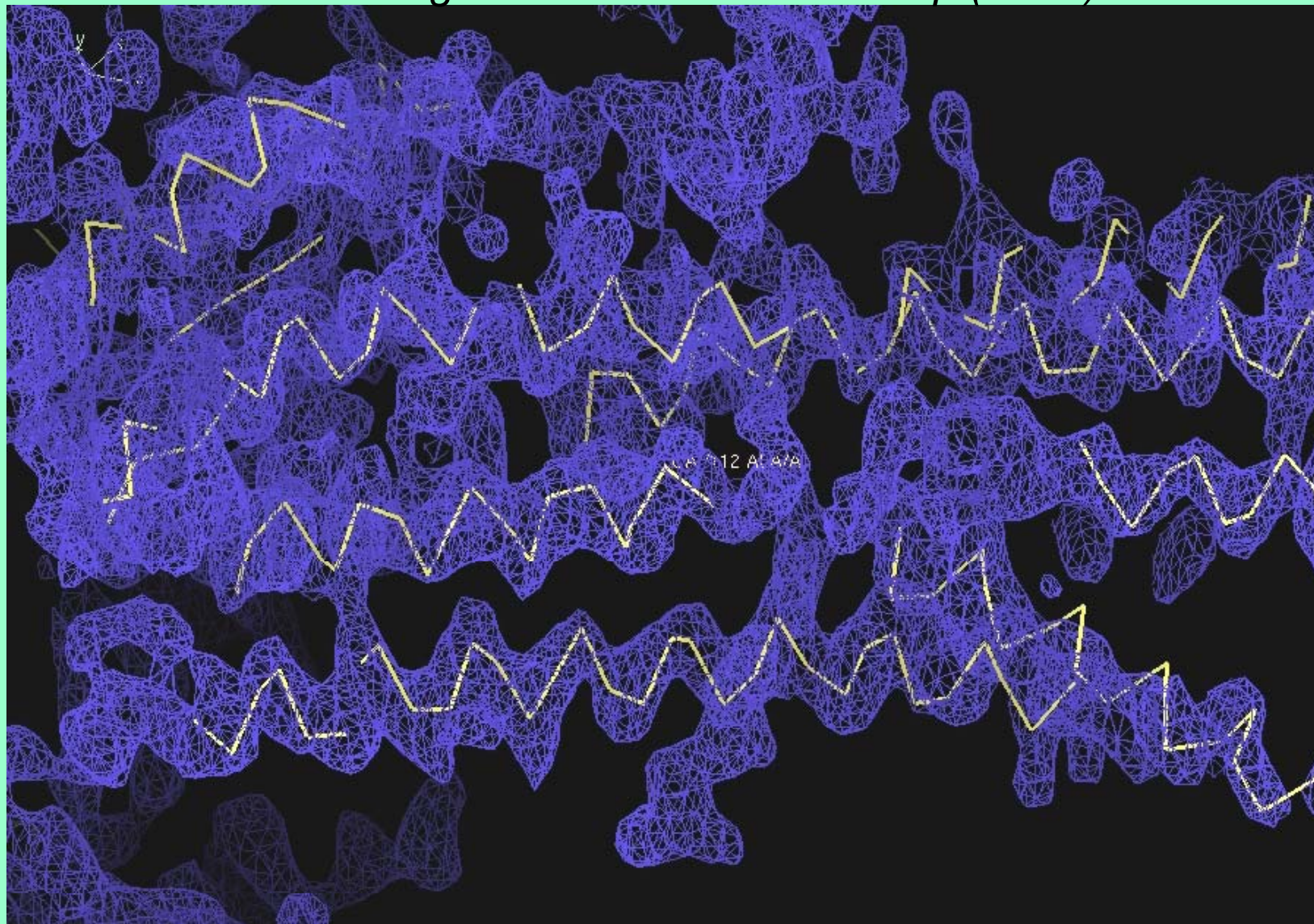
A real case: 1T5S SAD map (7 Å)



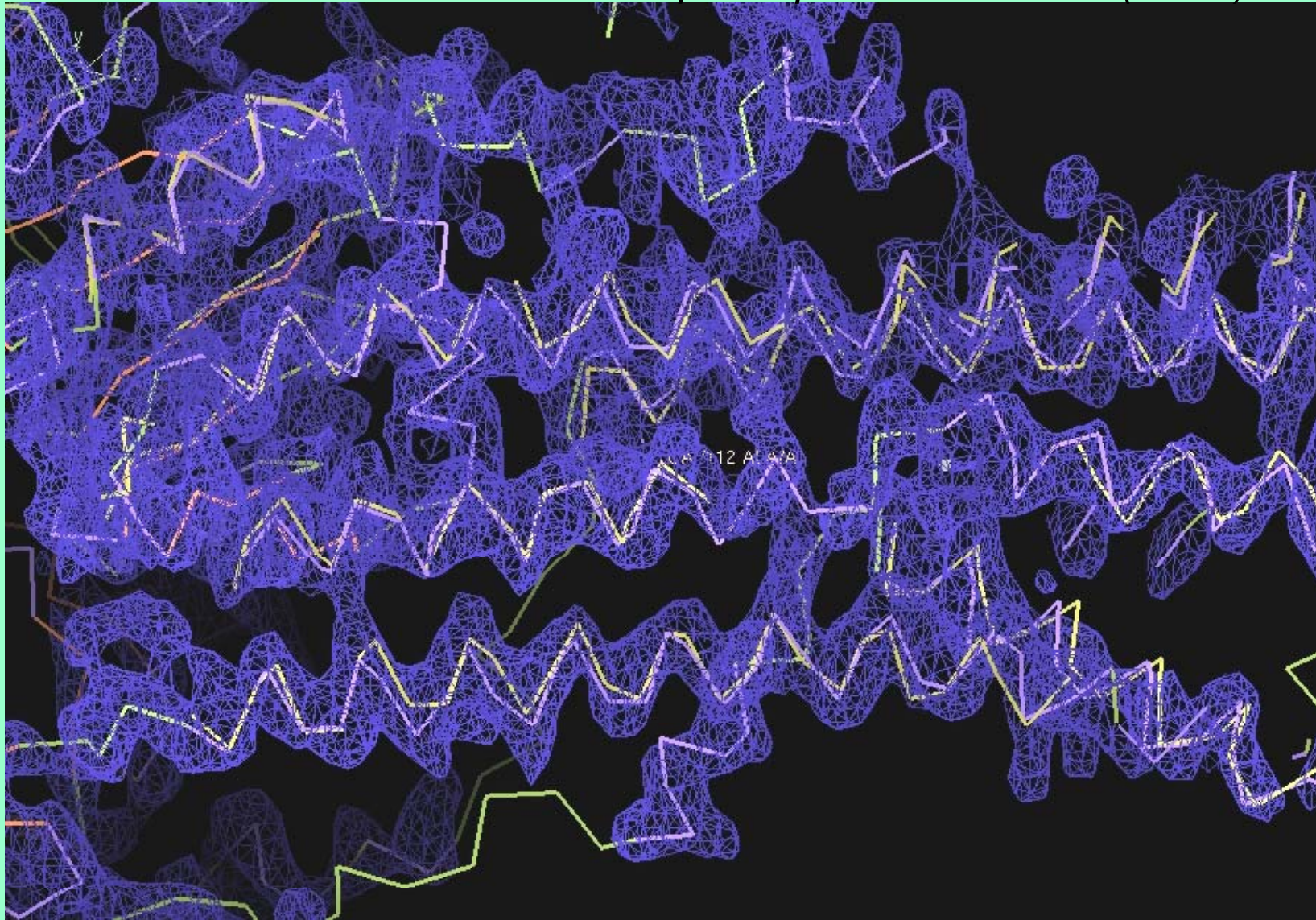
Finding helices in 1T5S SAD map (7 Å)



Finding helices in 1T5S SAD map (3.1 Å)



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



PHENIX AutoSol Wizard

*Scale and analyze X-ray data
(SAD/MAD/MIR/Multiple datasets)*



*HYSS heavy-atom search on each
dataset*



*Initial phasing (2.5 Å)
Score and rank solutions
(Which hand? Which dataset gave best solution?)*

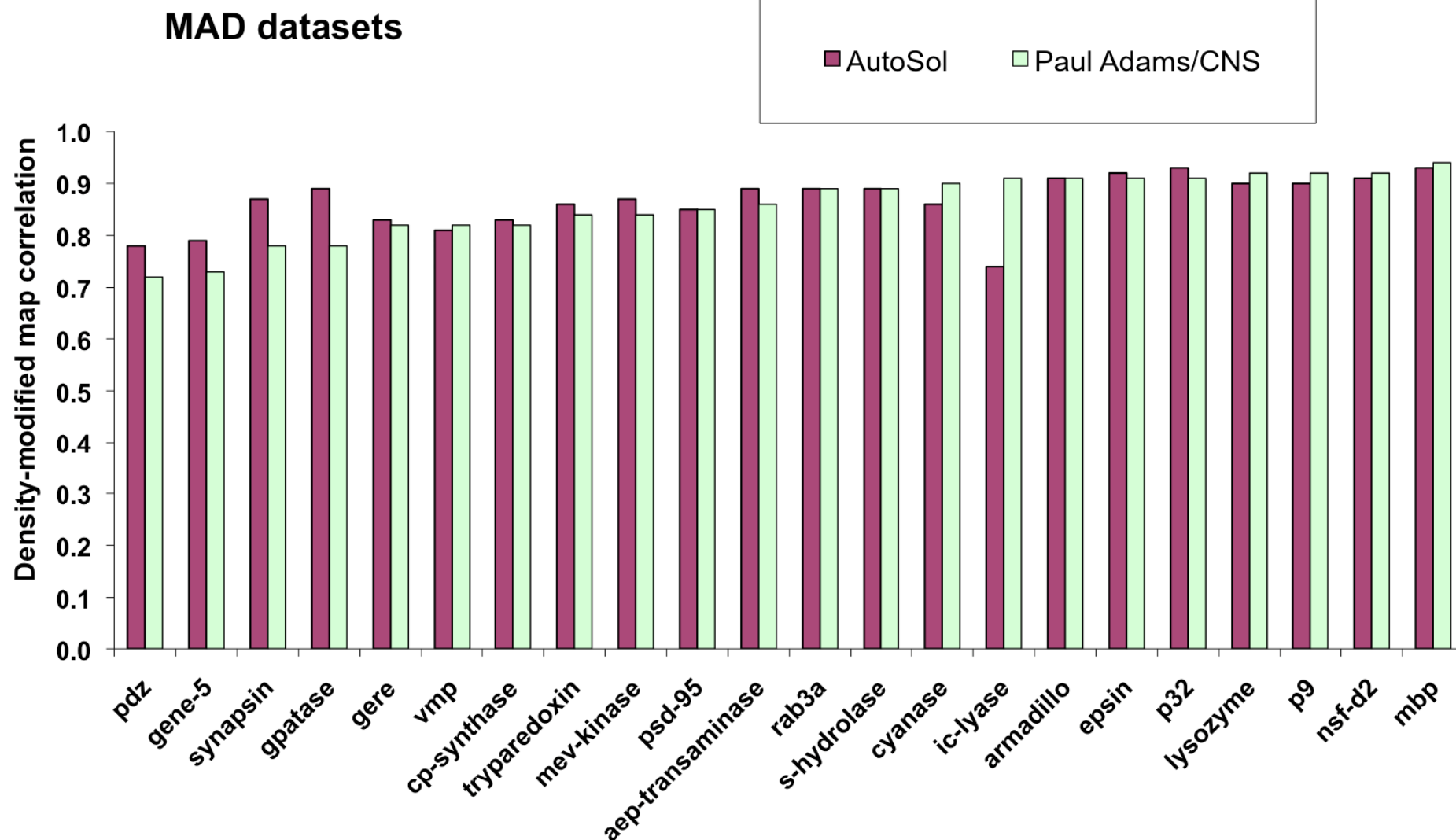


Final phasing and density modification

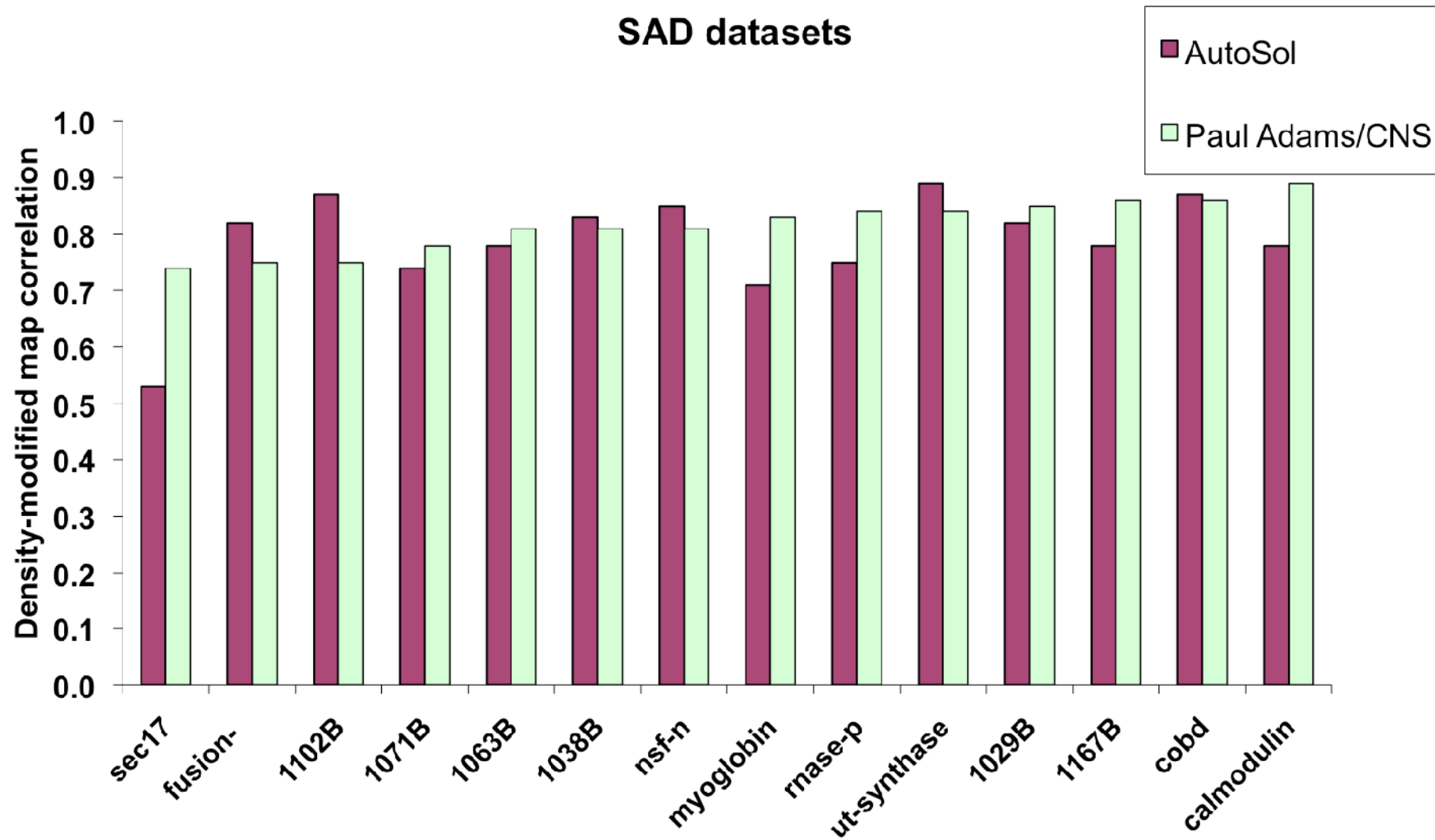


Build secondary-structure-only model

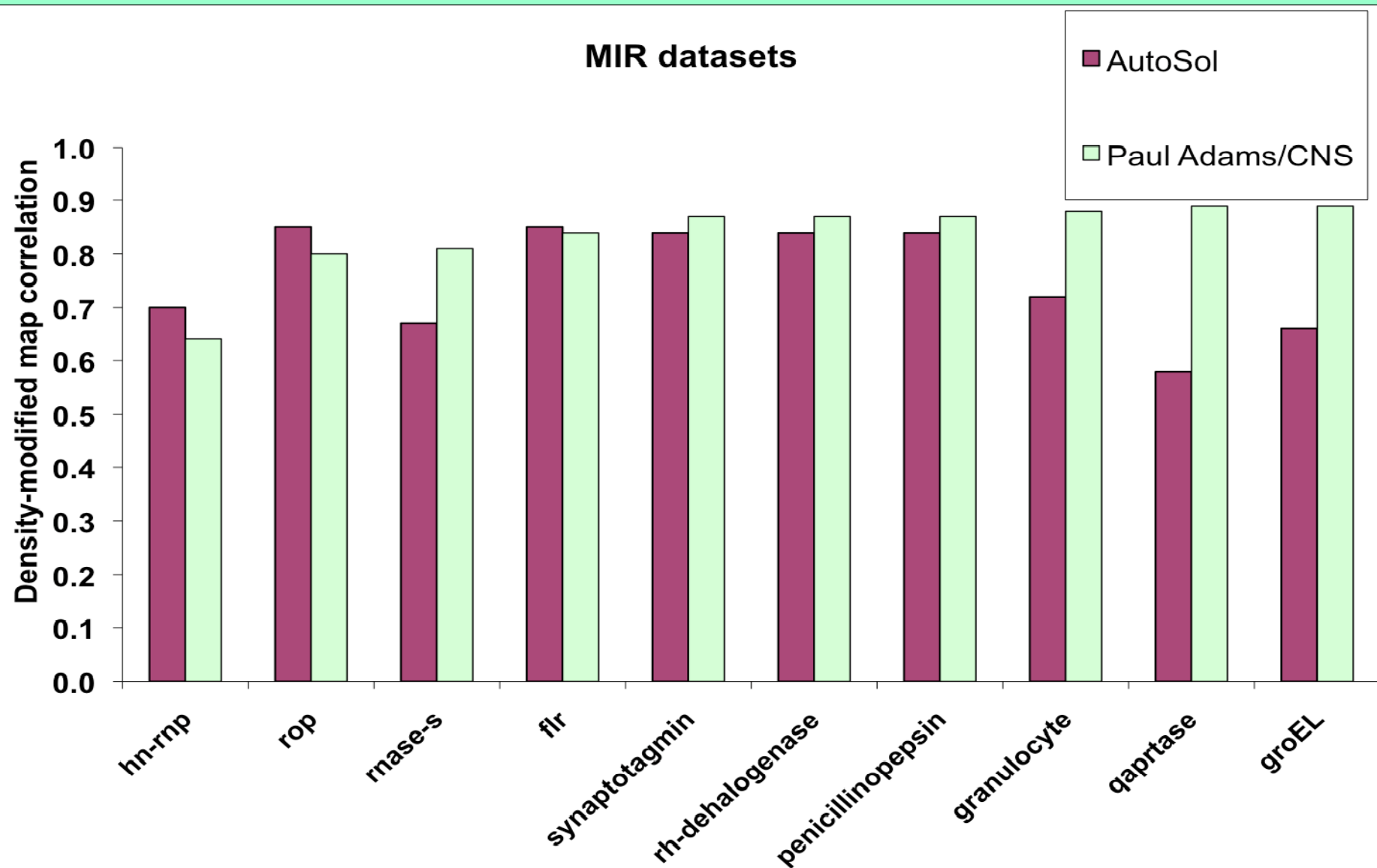
AutoSol – fully automatic tests with structure library (MAD datasets, HYSS search, SOLVE/RESOLVE phases)



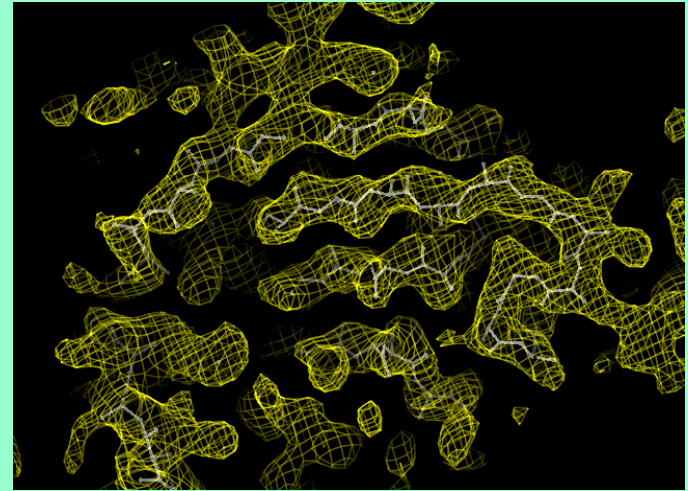
AutoSol – fully automatic tests with structure library (MAD datasets, HYSS search, Phaser phases)



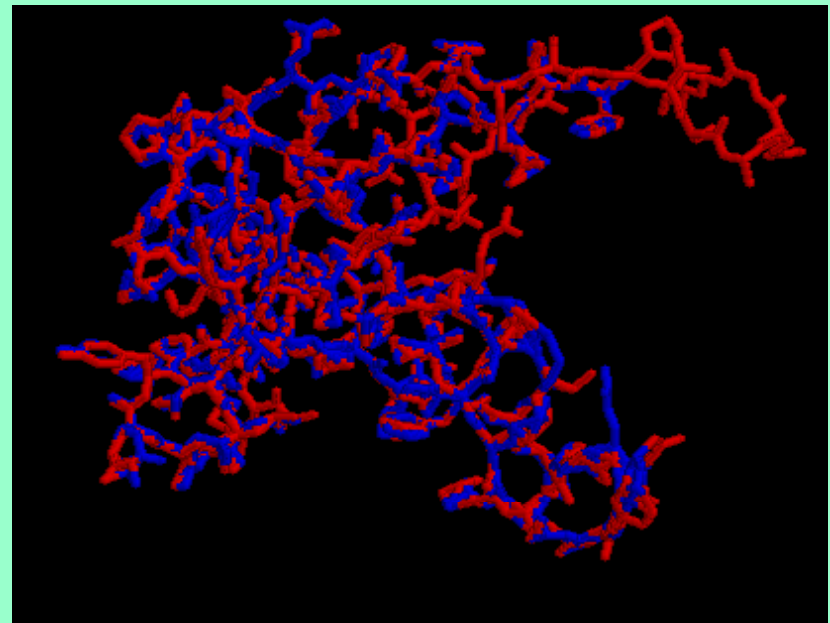
AutoSol – fully automatic tests with structure library (MIR datasets, HYSS search, SOLVE/RESOLVE phases)



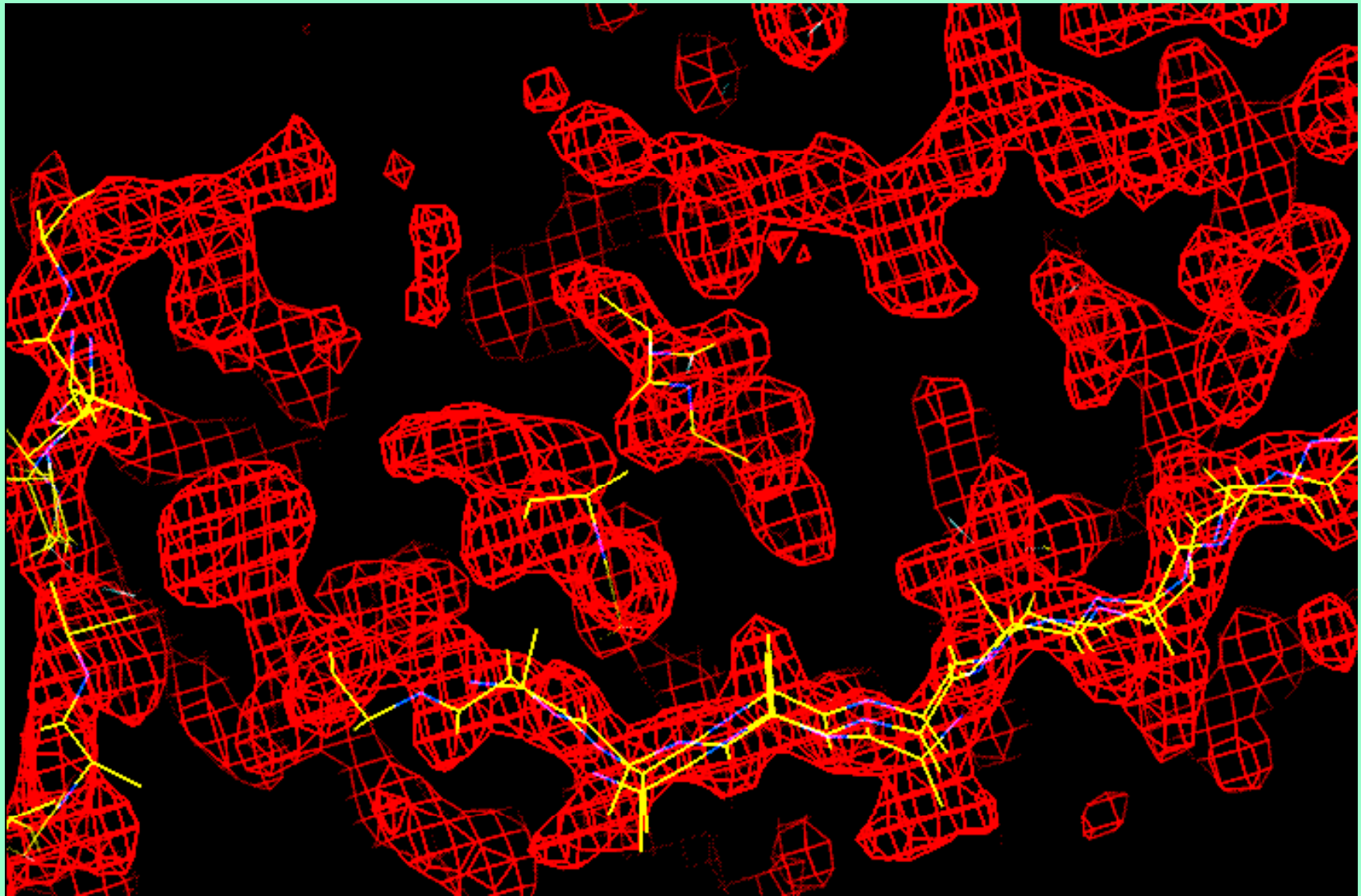
RESOLVE model-building at moderate resolution



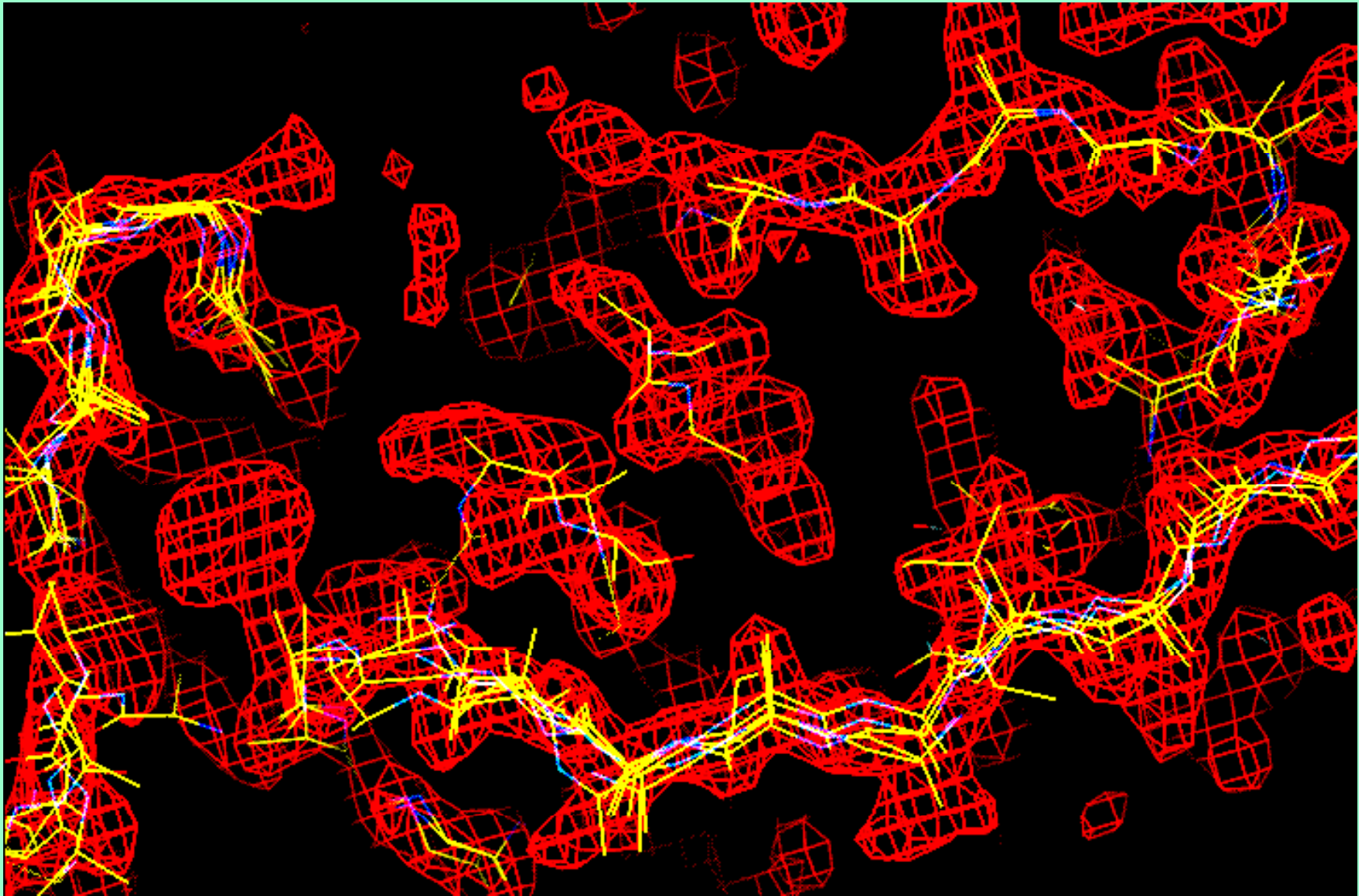
- *FFT-based identification of helices and strands*
- *Extension with tripeptide libraries*
- *Probabilistic sequence alignment*
- *Automatic molecular assembly*



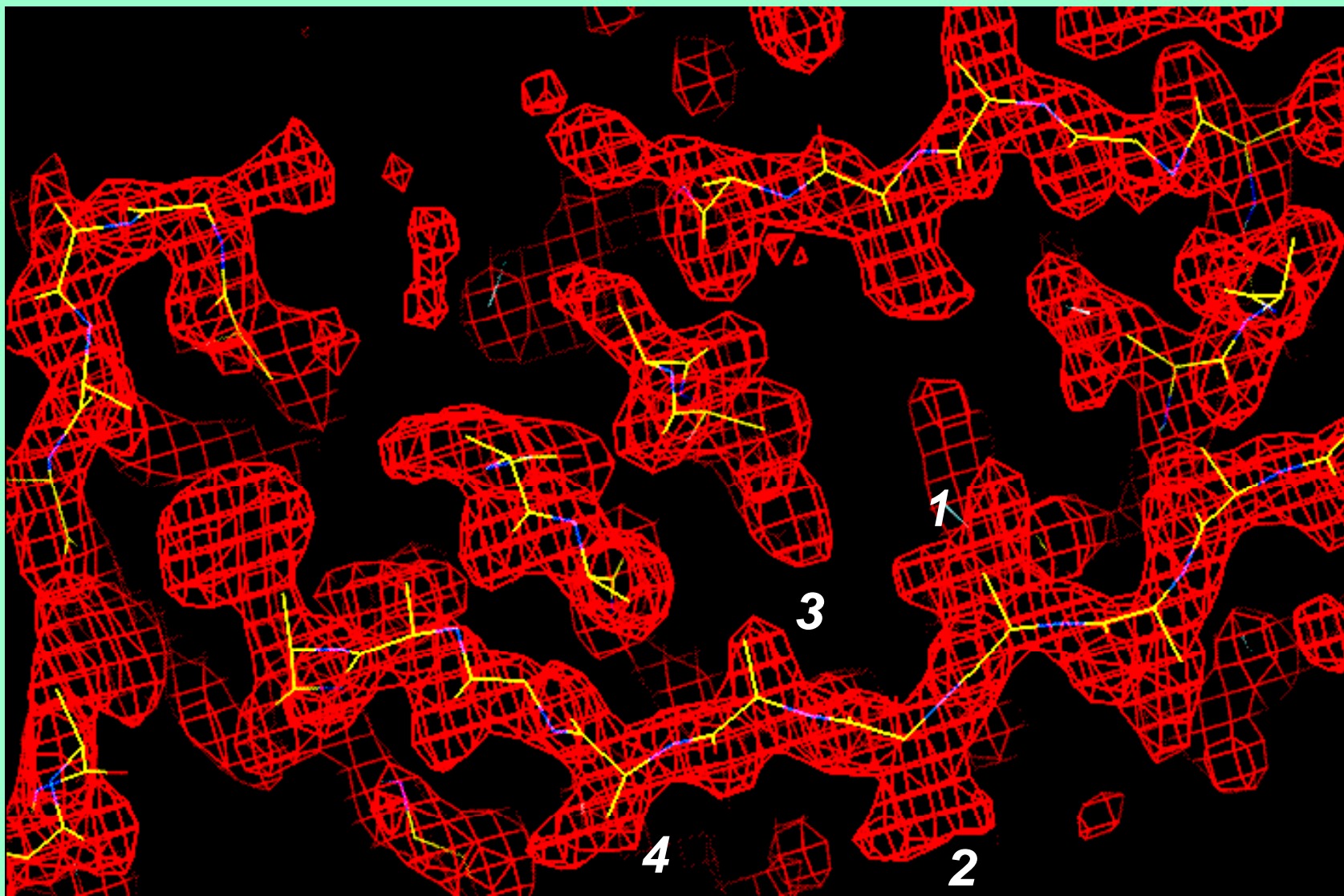
Initial model-building – strand fragments



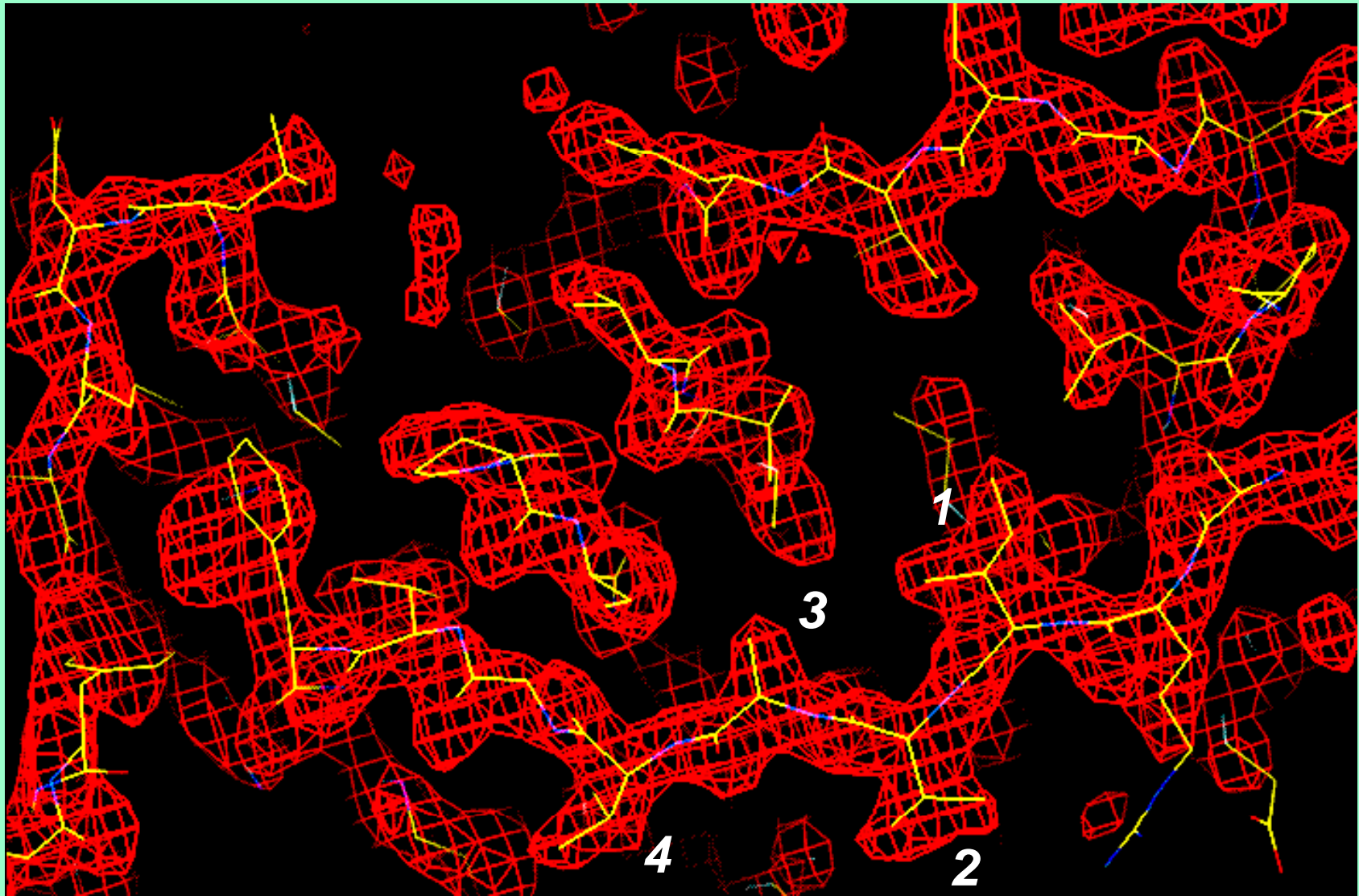
Chain extension
(result: many overlapping fragments)



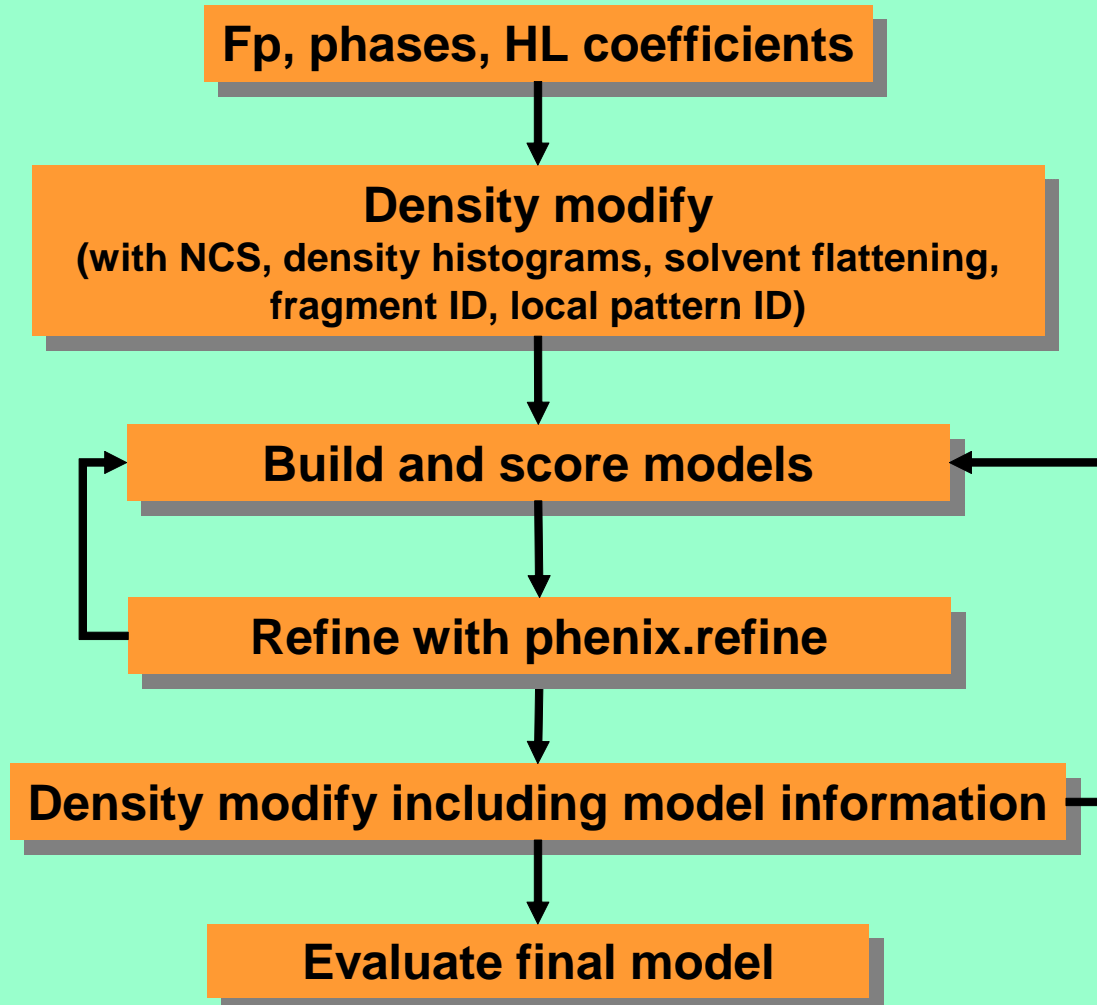
*Main-chain as a series of fragments
(choosing the best fragment at each location)*



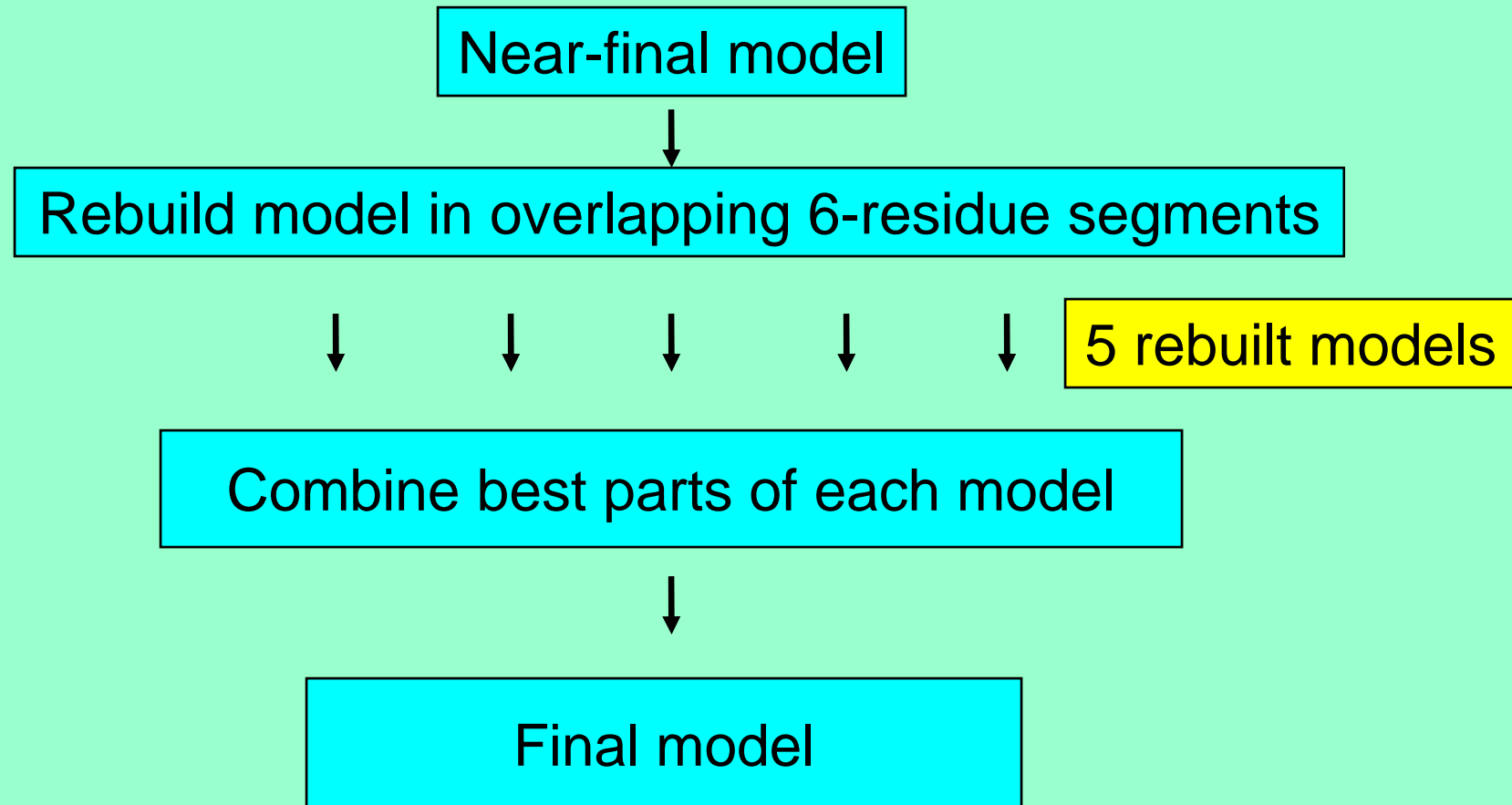
Addition of side-chains to fixed main-chain positions



Iterative density modification, model-building and refinement with the PHENIX AutoBuild Wizard

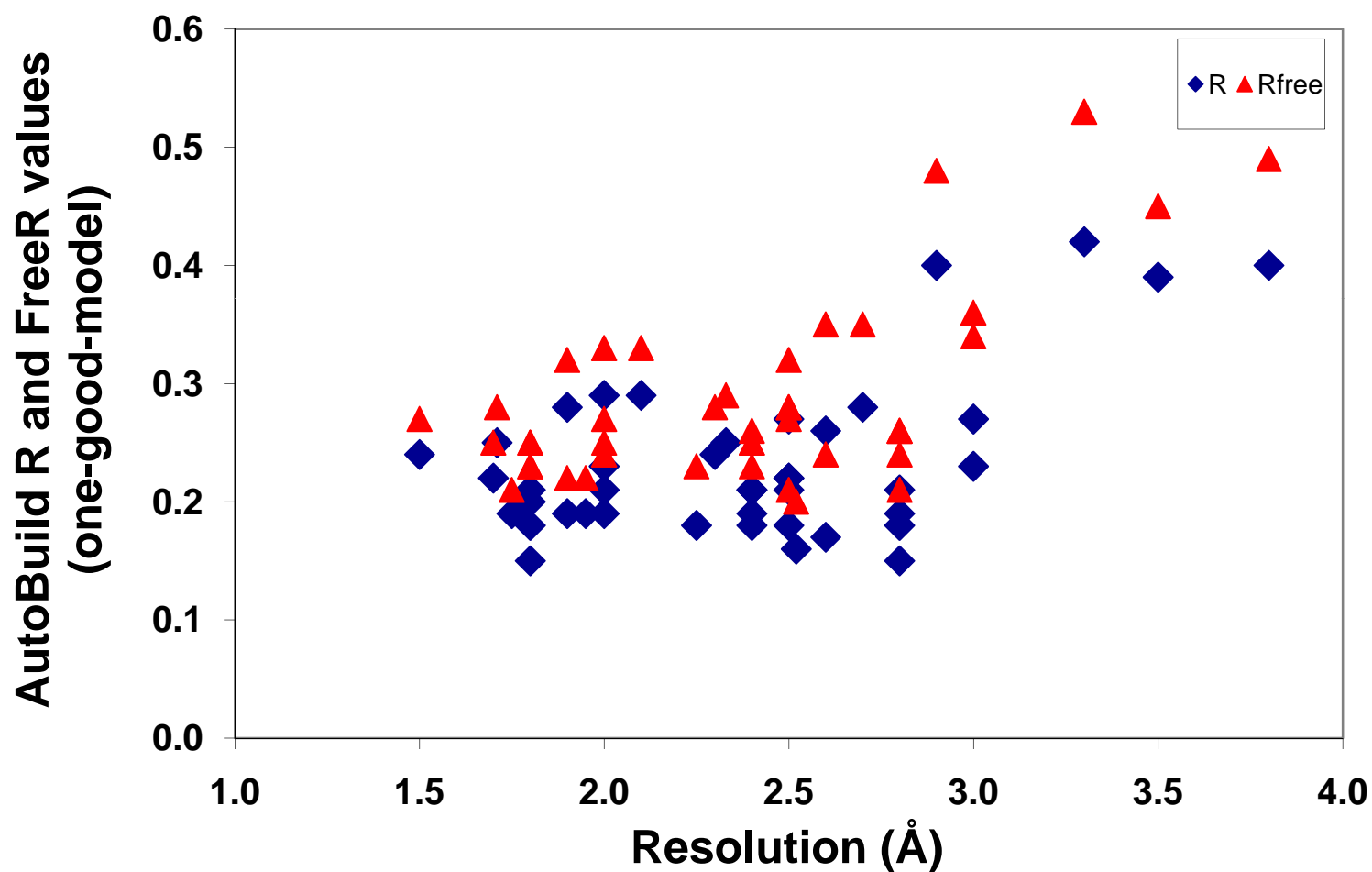


Generating one very good model with the PHENIX AutoBuild Wizard



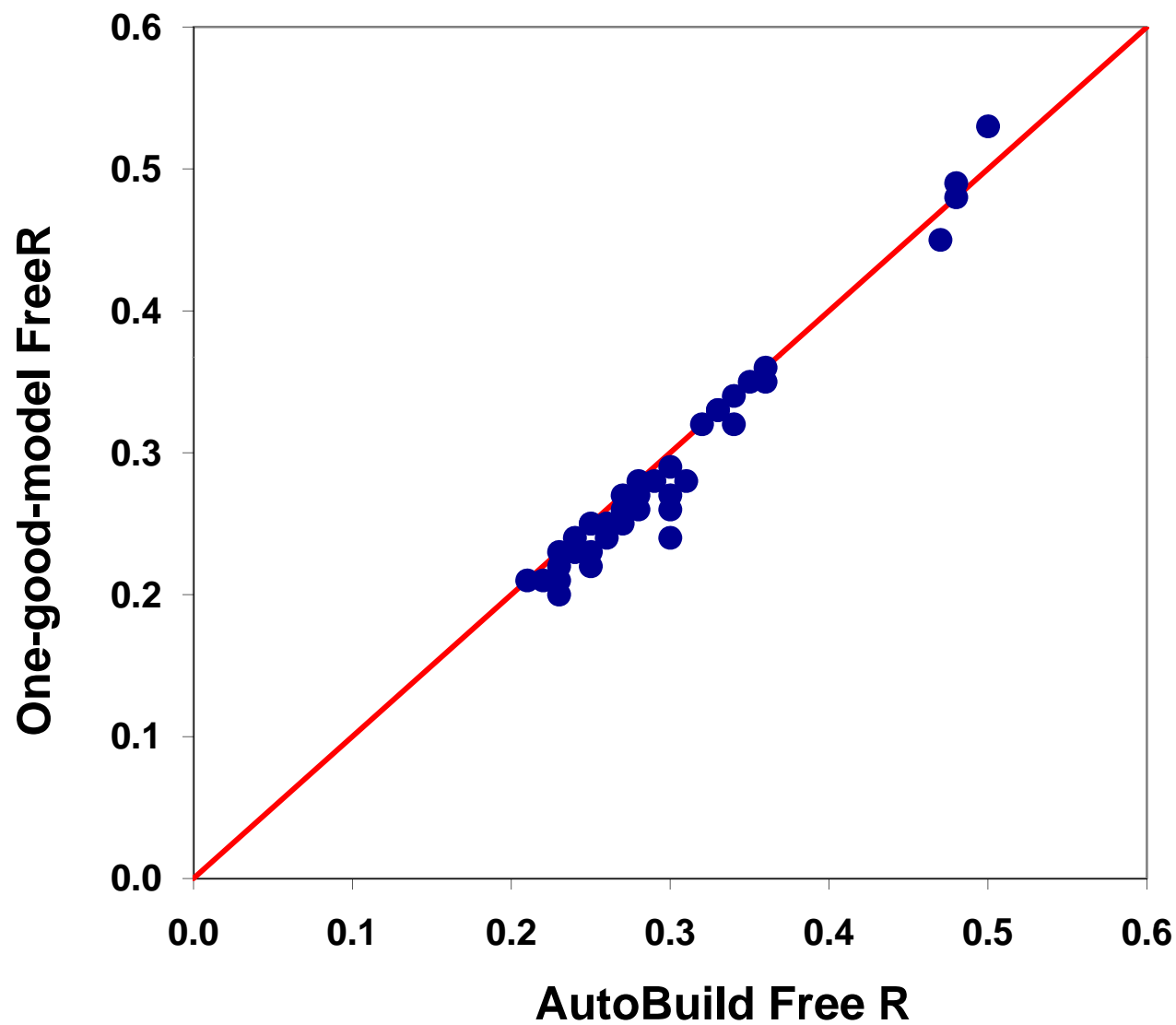
AutoBuild – tests with structure library

Fully automated iterative model-building, final R/Rfree



AutoBuild – tests with structure library

Final Rfree with one-good-model vs standard AutoBuild



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

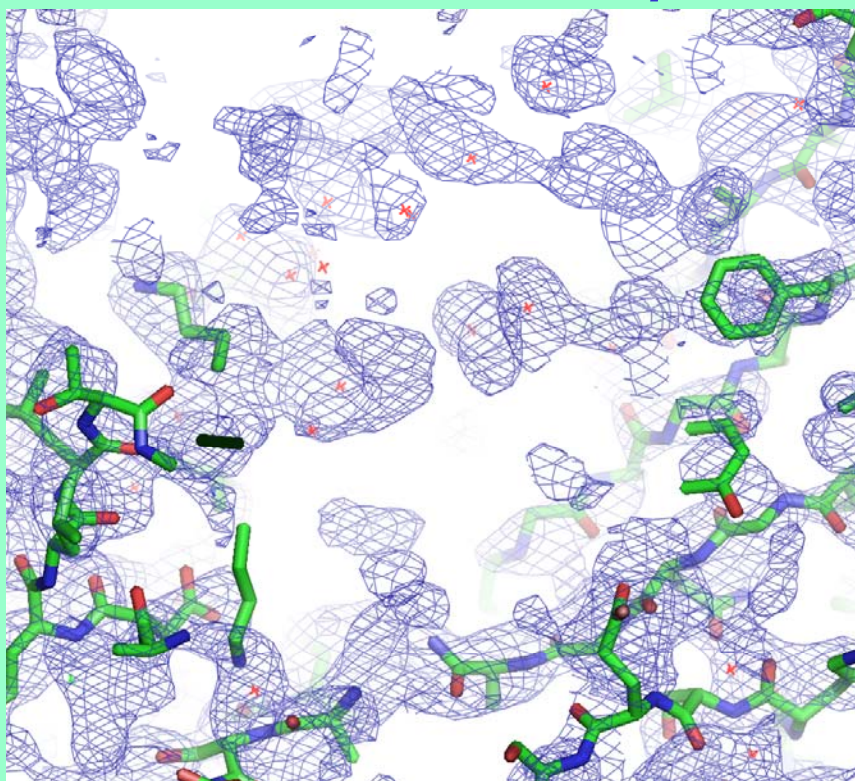
...checking different space groups, datasets to use

...checking if your model is biasing your map

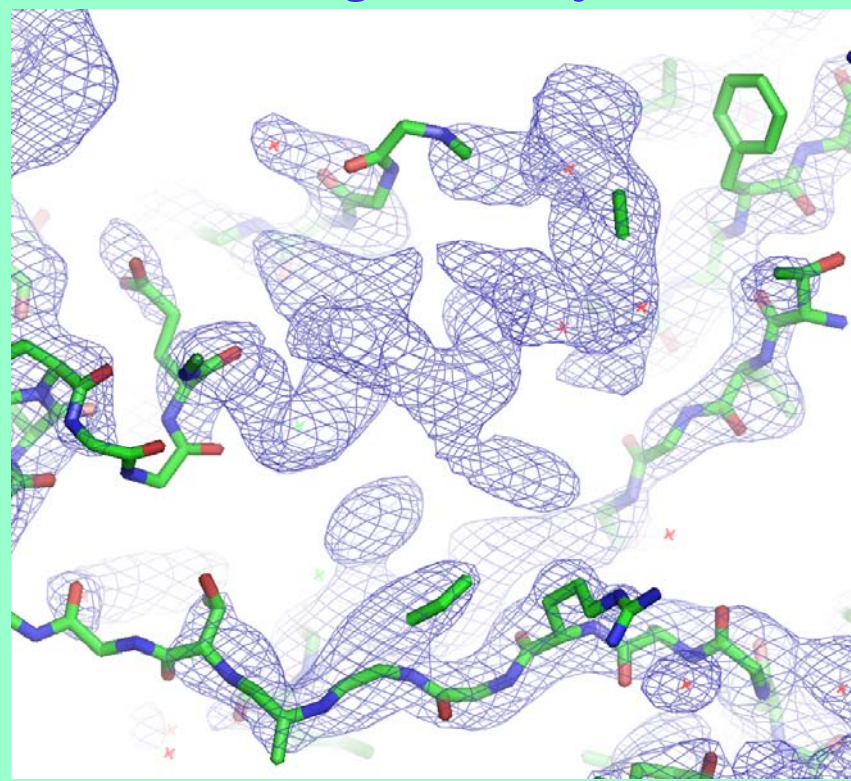
...checking if you always get the same model

Iterative-Build OMIT procedure

2mFo-DFc omit map



*After building outside
OMIT region 10 cycles*

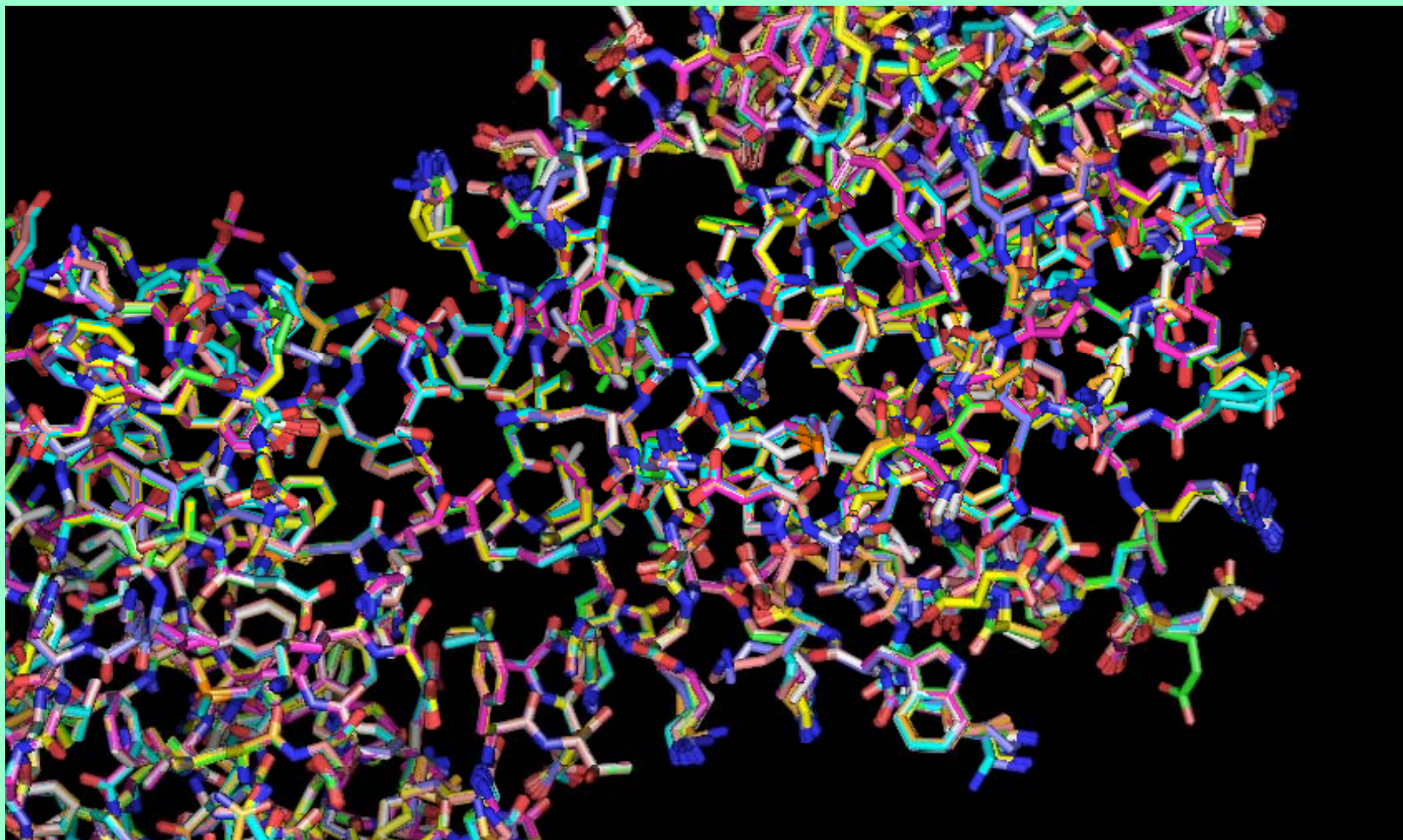


*1HP7 molecular replacement with 1AS4
R/Rfree after initial refinement: 0.41/0.48*

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



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

If a task is modular and automated...

you can run it focusing on different parts of the structure

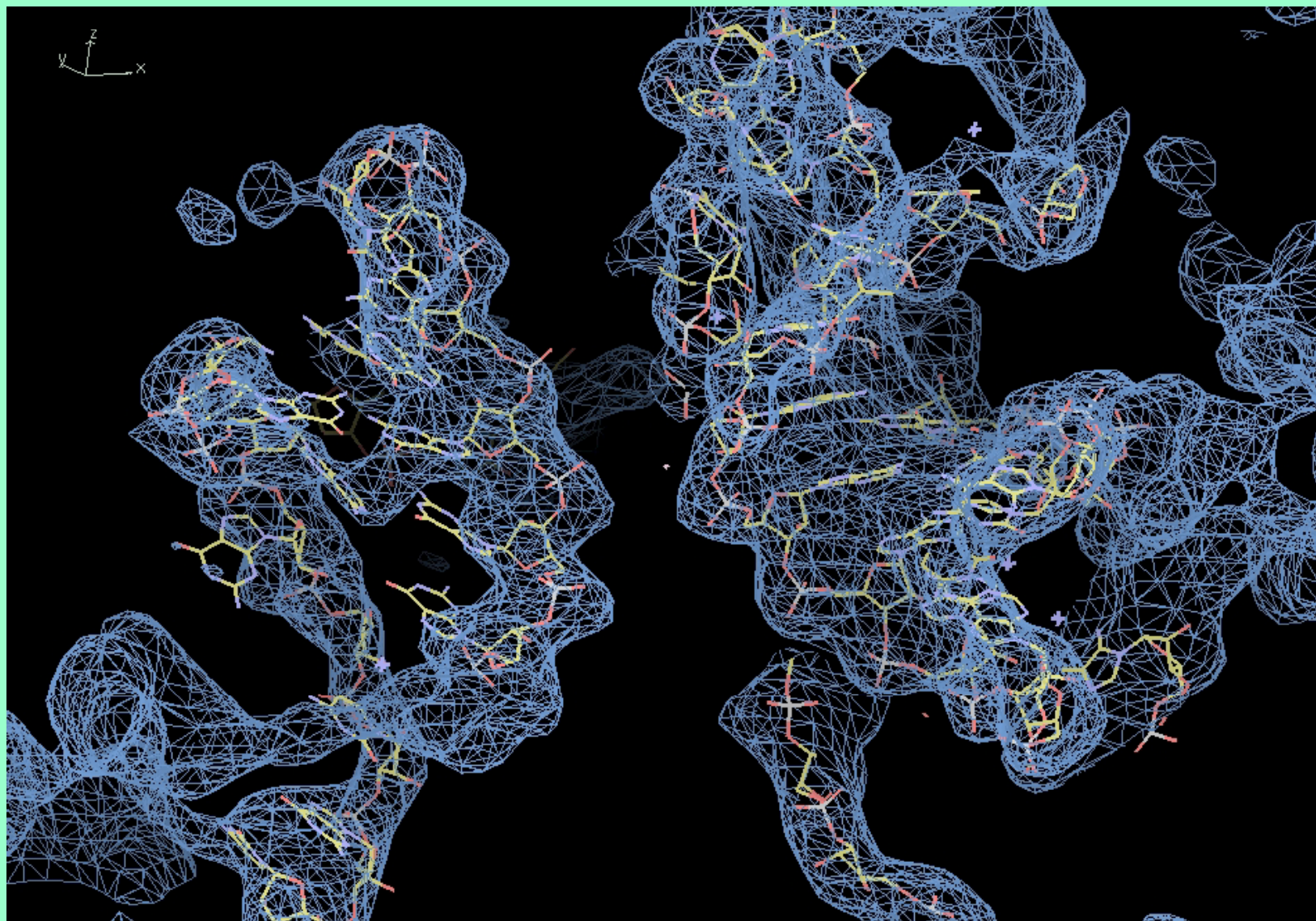
...build the RNA and then the protein

...build the helices in a low resolution map

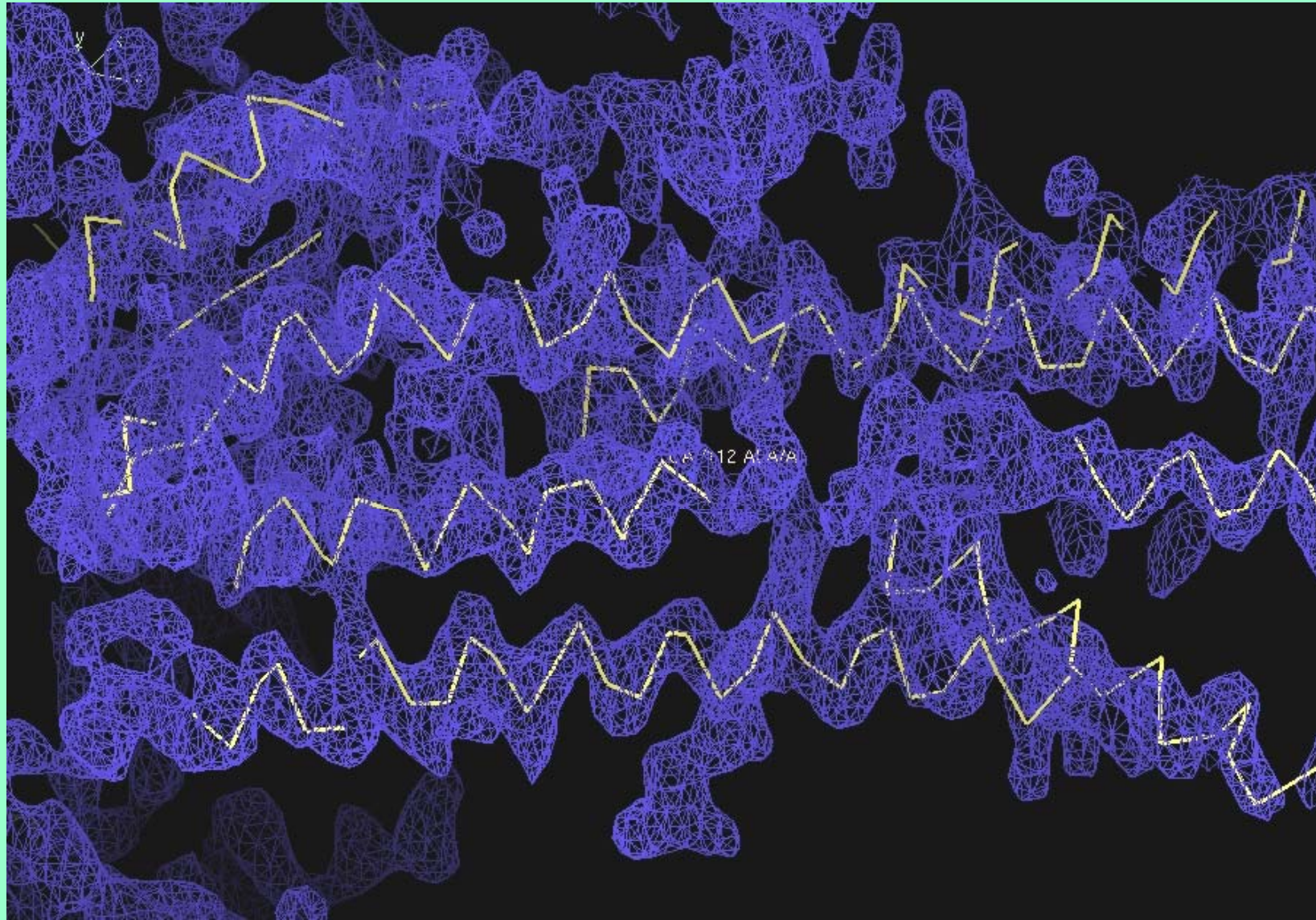
... use cross-crystal averaging in density modification

...build a protein model and then add ligands

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



Finding helices
Ca²⁺ ATPase SAD map at 3.1 Å. Data courtesy of P. Nissen



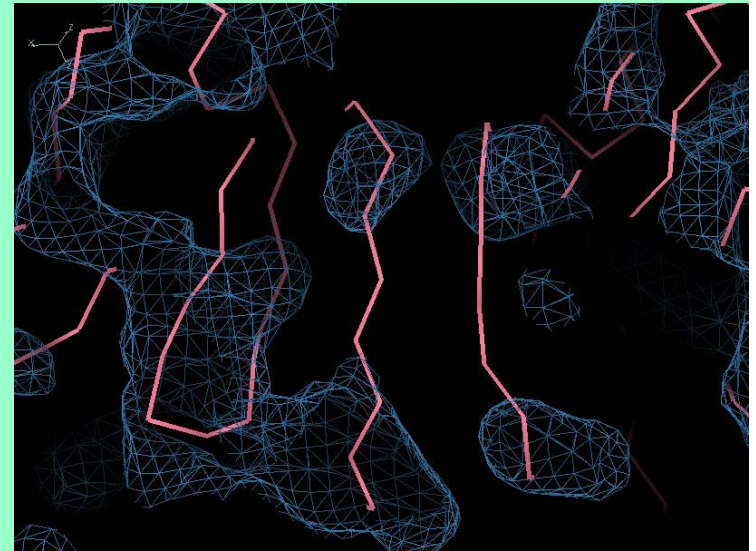
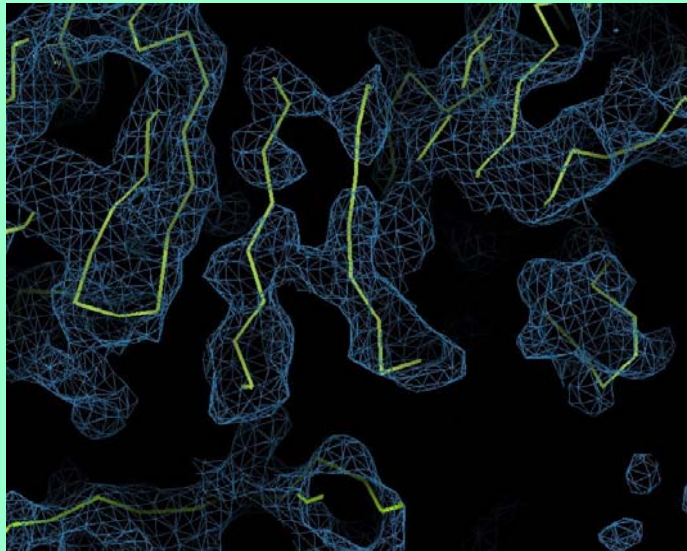
Statistical density modification with cross-crystal averaging

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

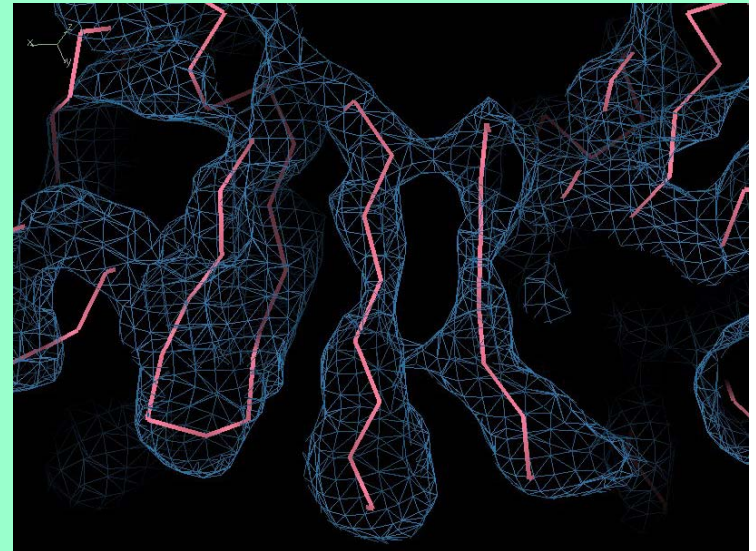
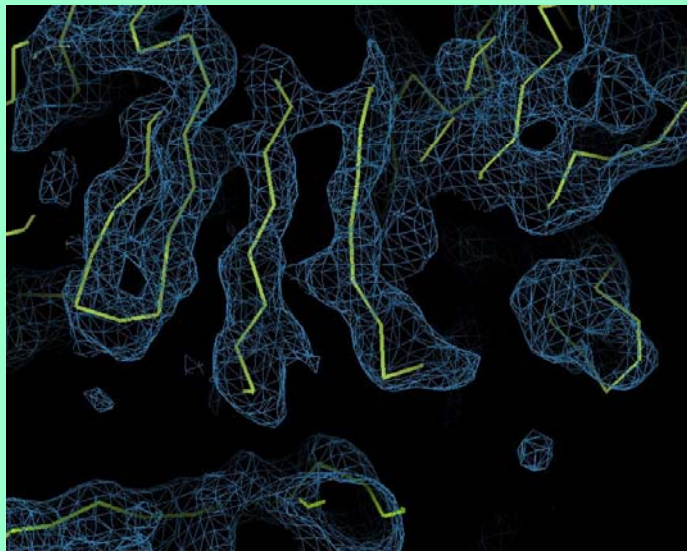
Crystal 1 (4 copies)

Crystal 2 (2 copies)

**RESOLVE
density
modification**

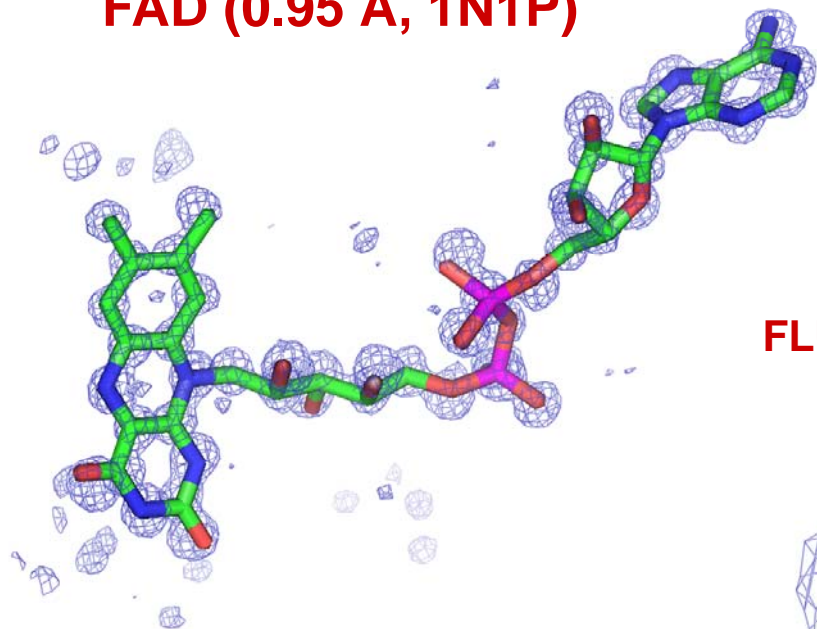


**PHENIX
Multi-crystal
averaging**

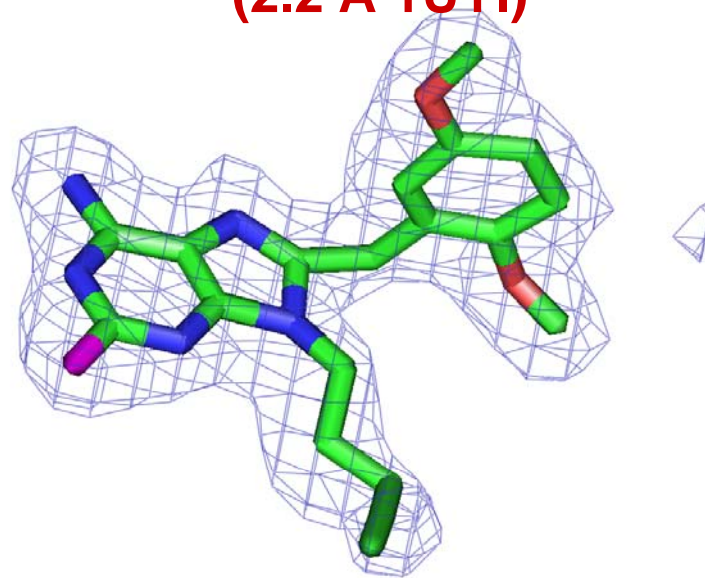


Automated fitting of flexible ligands

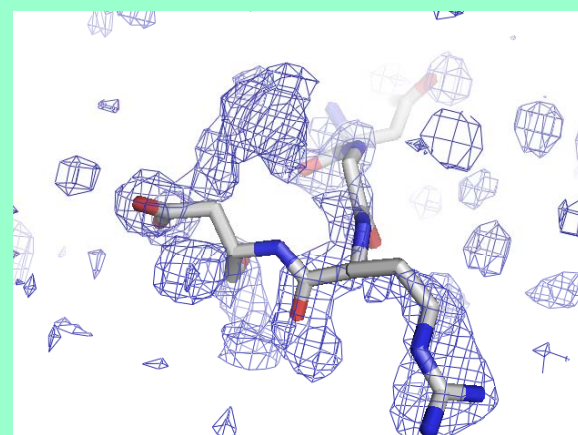
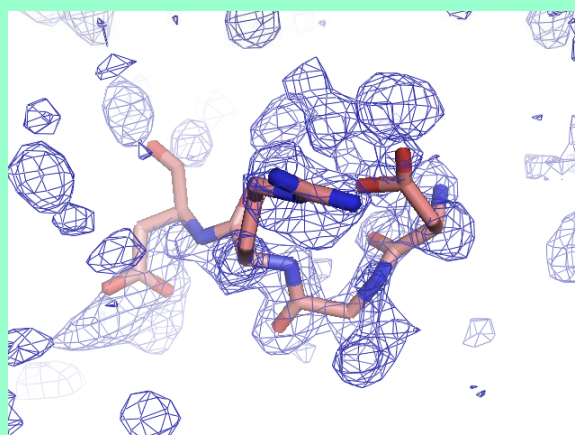
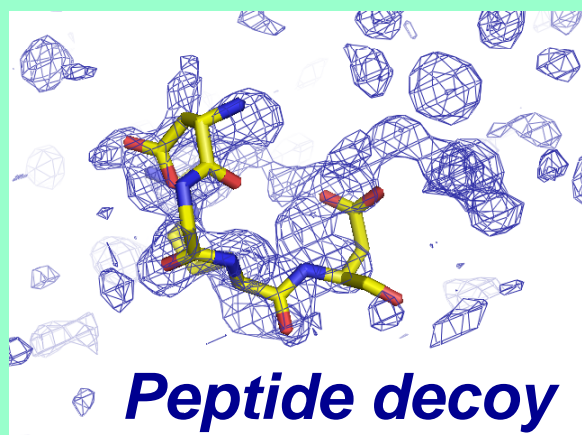
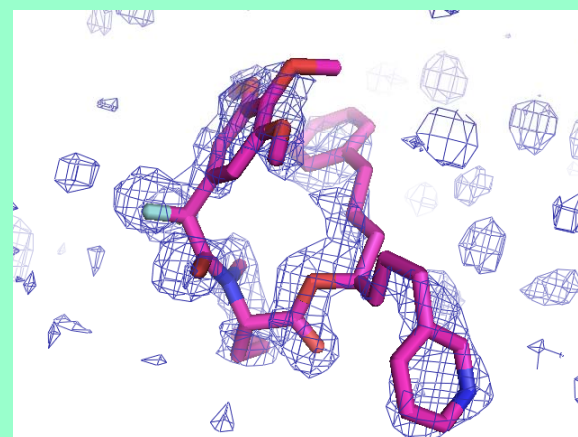
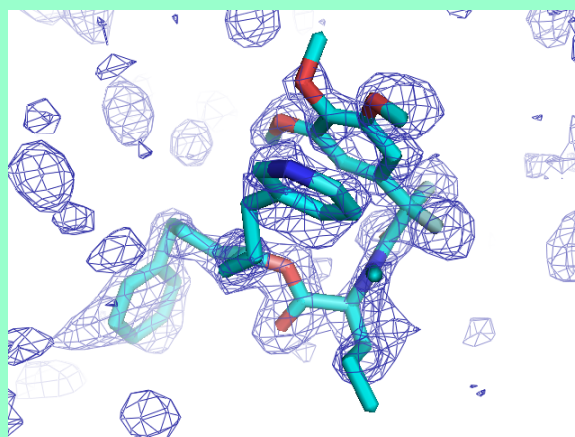
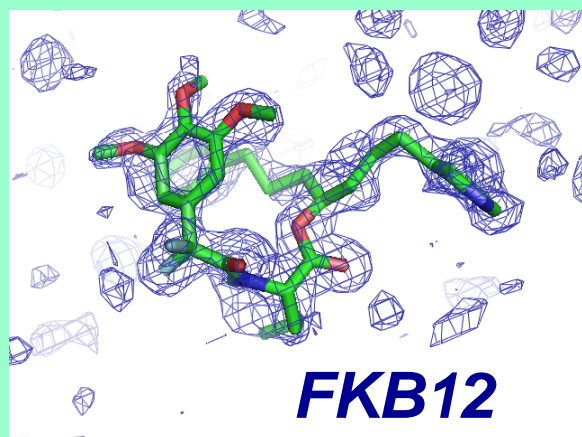
FAD (0.95 Å, 1N1P)



**8-(2,5-DIMETHOXY-BENZYL)-2-
FLUORO-9-PENT-9H-PURIN-6-YLAMINE
(2.2 Å 1UYI)**



phenix.find_all_ligands – 1J4R (3 molecules of FKB12)



Site 1

Site 2

Site 3

The future: many hard problems remain in macromolecular crystallography

Automatically identifying and building all ligands, metals, waters

Building multiple conformers

Building poorly-defined regions

Building complexes of protein and nucleic acid

Representation of uncertainties in models

Choosing optimal data (multiple crystals, multiple soaks) to use

Automatic analysis of radiation damage

Optimal structure solution in the presence of twinning

...and many more

PHENIX AT ARGONNE CCP4 WORKSHOP

- Paul Adams
- Tom Terwilliger
- www.phenix-online.org
- phenix.doc for help

phenix.autosol, phenix.autobuild
phenix.refine ...

The PHENIX project

Computational Crystallography Initiative (LBNL)

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



Los Alamos National Lab (LANL)

Tom Terwilliger, Li-Wei Hung



Cambridge University

*Randy Read, Airlie McCoy, Gabor Bunkoczi,
Rob Oeffner*



Duke University

Jane Richardson, David Richardson, Jeff Headd, Vincent Chen



Texas A&M University

Tom Ioerger, Jim Sacchettini



<http://www.phenix-online.org>