Molecular Replacement Structure Solution

Macromolecular Crystallography School
Madrid, May 2017

Paul Adams

Lawrence Berkeley Laboratory and
Department of Bioengineering UC Berkeley

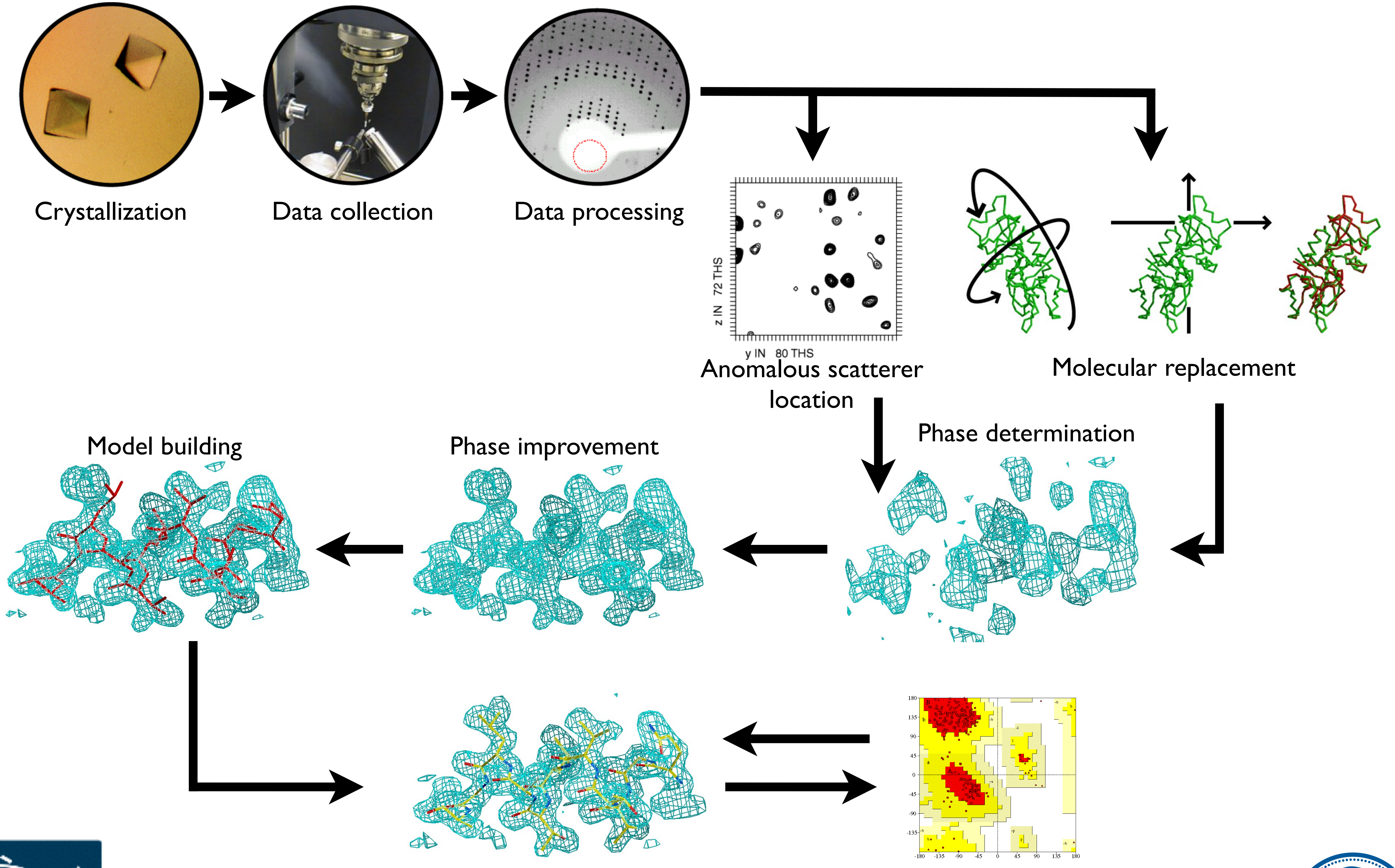
Eruditio et Religio

Los Alamos
NATIONAL LABORATORY
EST.1943

BERKELEY LAB

UNIVERSITY OF
CAMBRIDGE

The Crystallographic Process



Crystallization

Data collection

Data processing

Anomalous scatterer location

Molecular replacement

Phase determination

Phase improvement

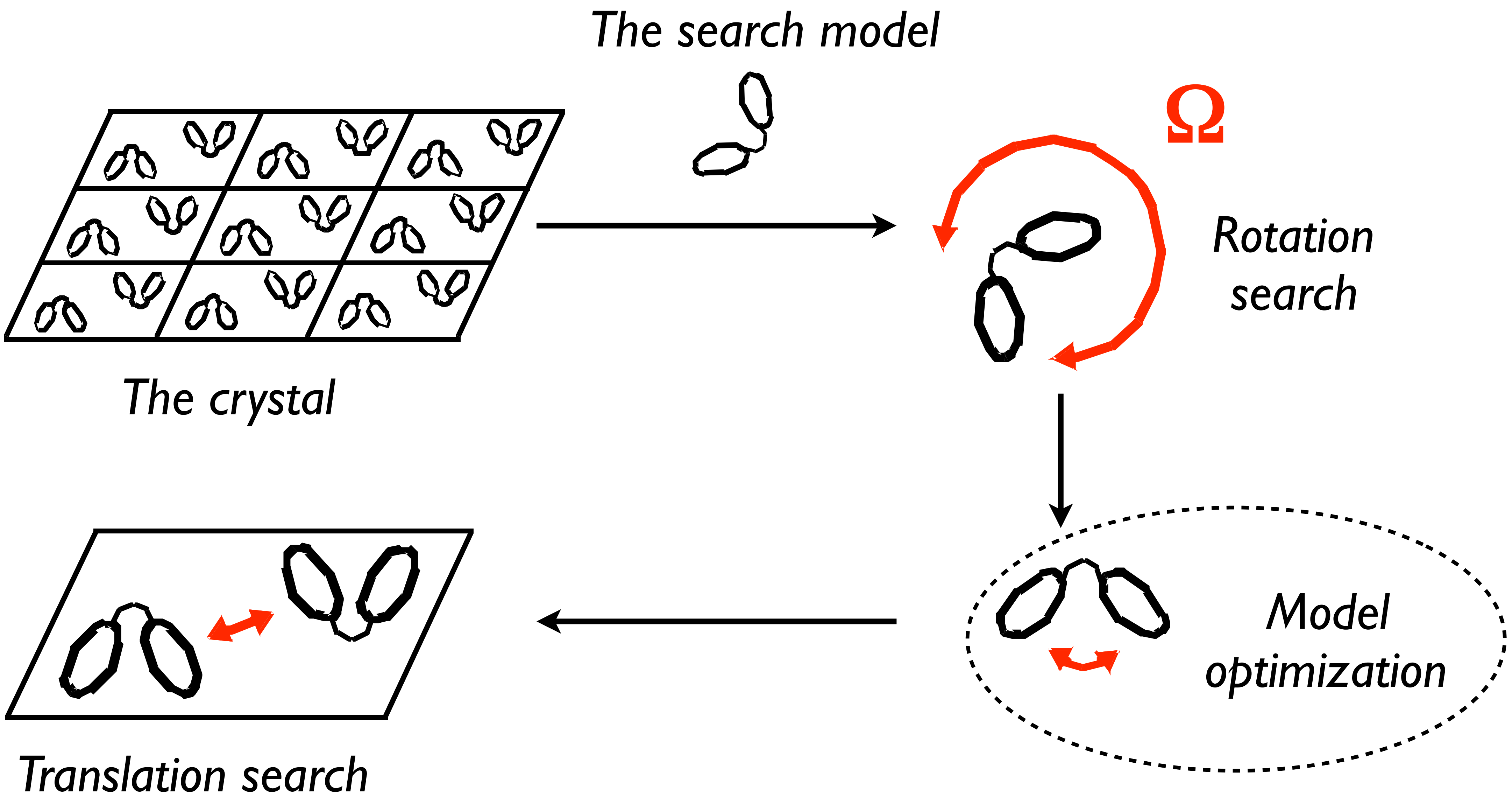
Model building

Model refinement

Validation

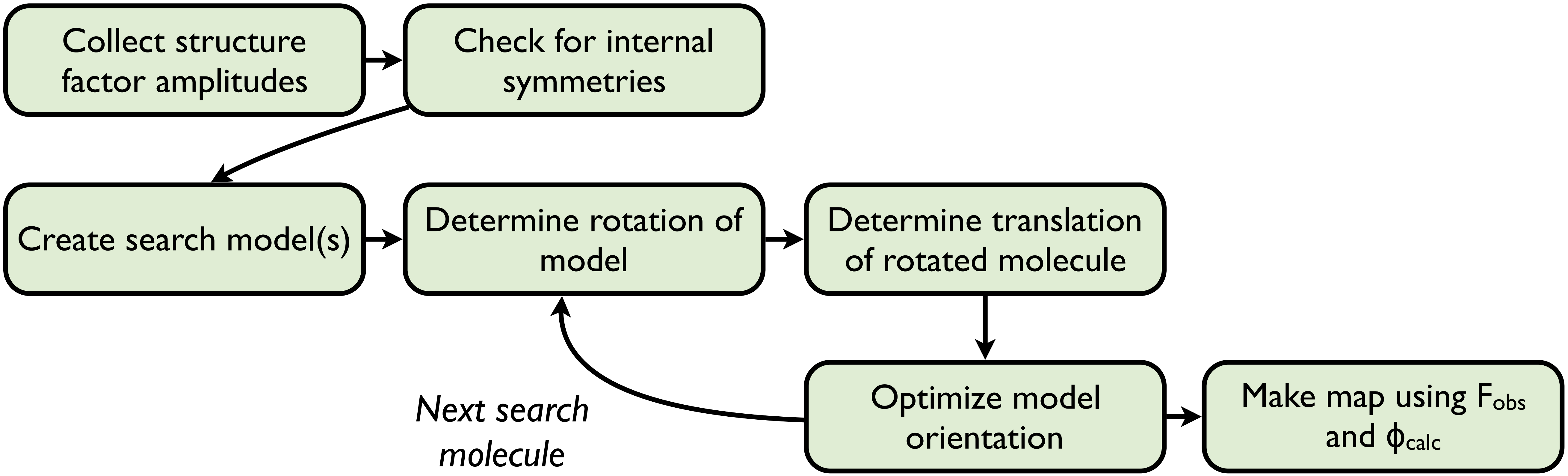
Phenix

The Divide and Conquer Approach



- The orientation and translation of models is searched on a grid
- The grid parameters depend on the resolution of the data and the symmetry of the crystal
- Rigid body refinement allows the model to move off the predefined search grid

Overview of Molecular Replacement



Collect structure factor amplitudes

Check for internal symmetries

Create search model(s)

Determine rotation of model

Determine translation of rotated molecule

Optimize model orientation

Make map using F_{obs} and ϕ_{calc}

Next search molecule

- Methods rely on the magnitude of measured amplitudes (not differences)
- Shares some methods with substructure location
- Sensitive to missing or poorly measured data (especially at low resolution)
- Can be automated for many cases
- ~75% or more of structures solved annually are by molecular replacement

Scoring Functions

- Traditional Rotation Function:
  - Patterson product function

$$\mathrm{Rot}(\Omega) = \int\limits_U P_{\mathrm{obs}}(u) P_{\mathrm{model}}(\Omega u)\,\mathrm{d}u$$

- Direct Rotation Function:
  - Correlation of squared normalized structure factors ($X=E^2$)

$$\mathrm{CC}(\Omega) = \frac{\sum\limits_H (X_{H,\mathrm{obs}} - \langle X_{\mathrm{obs}} \rangle)(X_{H,\Omega} - \langle X_\Omega \rangle)}{\left[\sum\limits_H (X_{H,\mathrm{obs}} - \langle X_{\mathrm{obs}} \rangle)^2\right]^{1/2} \left[\sum\limits_H (X_{H,\Omega} - \langle X_\Omega \rangle)^2\right]^{1/2}}$$

Translation Functions

- Amplitude-based or phased translation functions.
- Variety of target functions:
  - Standard linear correlation of observed and calculated quantities (E, $|E|^2$, F, $|F|^2$)
  - Residual
- Fast Translation Function for correlation of $|F|^2$

$$C(\mathbf{t}) = \frac{\sum\left(|\mathbf{F}_O|^2 - \overline{|\mathbf{F}_O|^2}\right)\left(|\mathbf{F}_C(\mathbf{t})|^2 - \overline{|\mathbf{F}_C(\mathbf{t})|^2}\right)}{\sqrt{\sum\left(|\mathbf{F}_O|^2 - \overline{|\mathbf{F}_O|^2}\right)^2 \sum\left(|\mathbf{F}_C(\mathbf{t})|^2 - \overline{|\mathbf{F}_C(\mathbf{t})|^2}\right)^2}}$$

Likelihood

- Best model is most consistent with the data
- Measure consistency by probabilities
- Likelihood target:
  - probability of observed amplitude given (set of) model structure factor contributions
  - account for effect of unknown relative phases
- Benefits of likelihood
  - account for expected size of errors in model
  - account for lack of completeness of model
  - exploit knowledge from partial solutions
  - allow ensemble of possible models
    - useful for MR with NMR

Likelihood in Practice

- The search methods are very similar, but different target functions are used.

$$P\text{-RF}_r = \frac{2F_\text{o}}{\Sigma_S + \sigma_F^2} \exp\left(- \frac{F_\text{o}^2 + D^2 F_\text{big}^2}{\Sigma_S + \sigma_F^2} \right) I_0 \left(\frac{2F_O D F_\text{big}}{\Sigma_S + \sigma_F^2} \right)$$

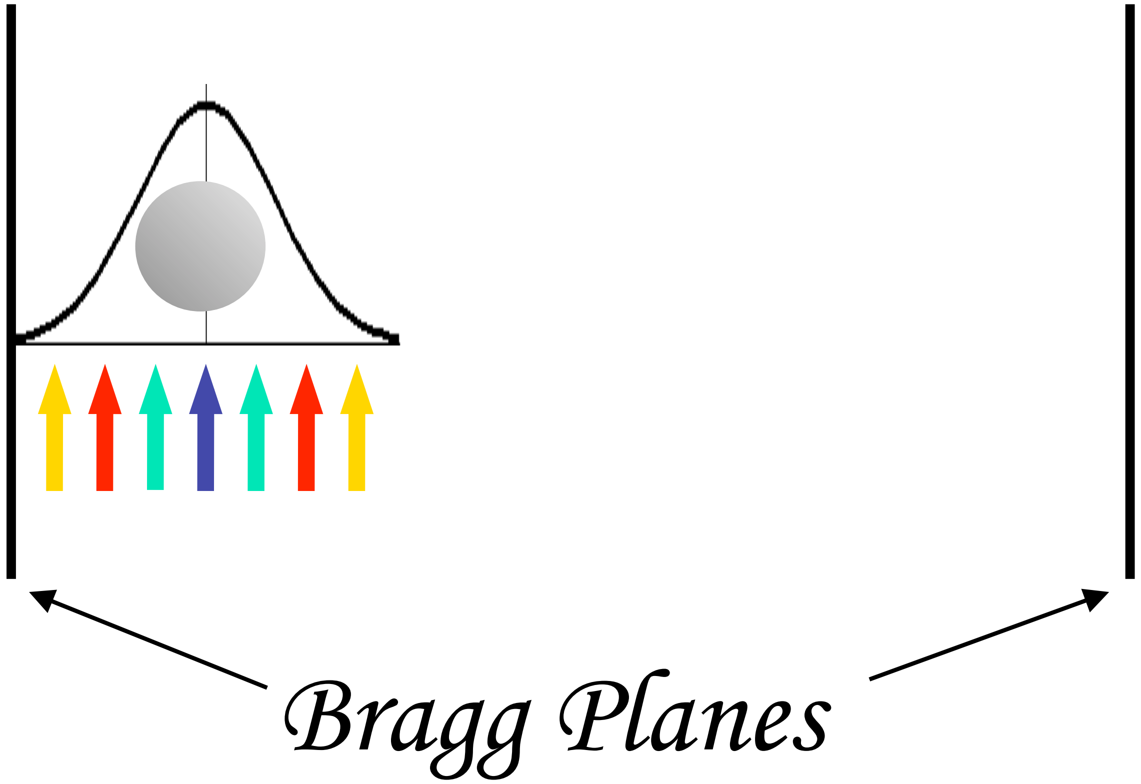
$$P\text{-TF}_r = P\text{-Xray}_r$$

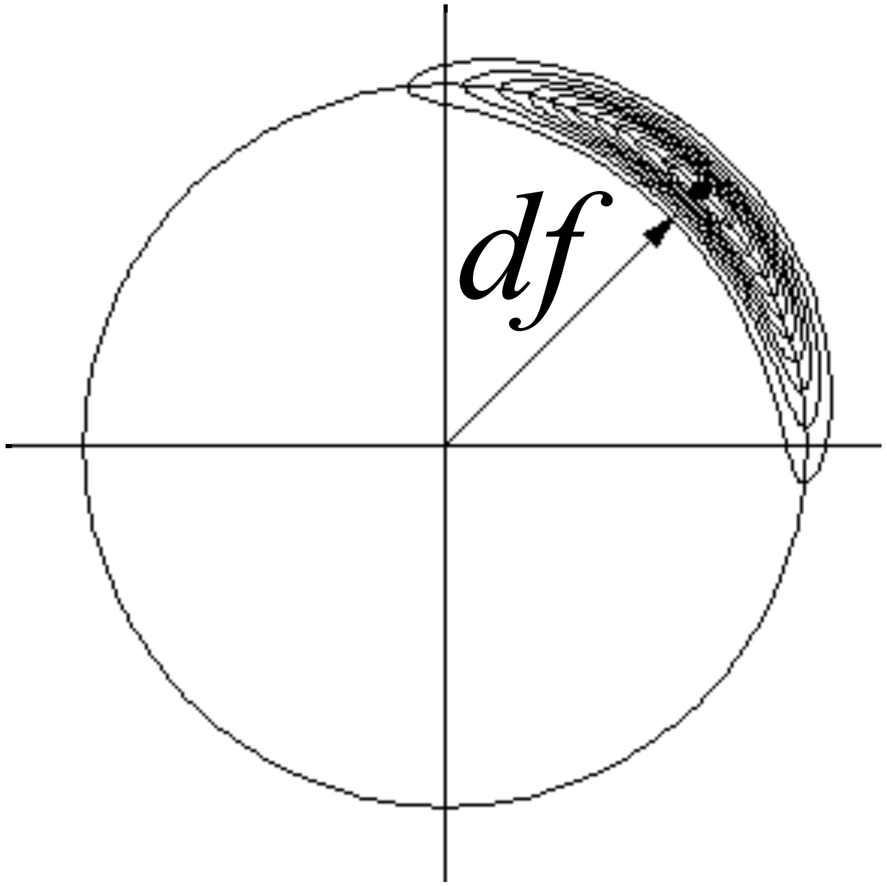
$$= \frac{2F_\text{o}}{\sigma_\Delta^2 + \sigma_F^2} \exp\left(- \frac{F_\text{o}^2 + D^2 F_\text{c}^2}{\sigma_\Delta^2 + \sigma_F^2} \right) I_0 \left(\frac{2F_\text{o} D F_\text{c}}{\sigma_\Delta^2 + \sigma_F^2} \right)$$

- Approximations can be used to calculate the rotation and translation functions rapidly using FFTs.
- Allows prior information to be used even in the rotation search.
- Requires a way to describe how similar/different the search model is to the expected structure (an error model)

Effect of Errors in Atomic Position

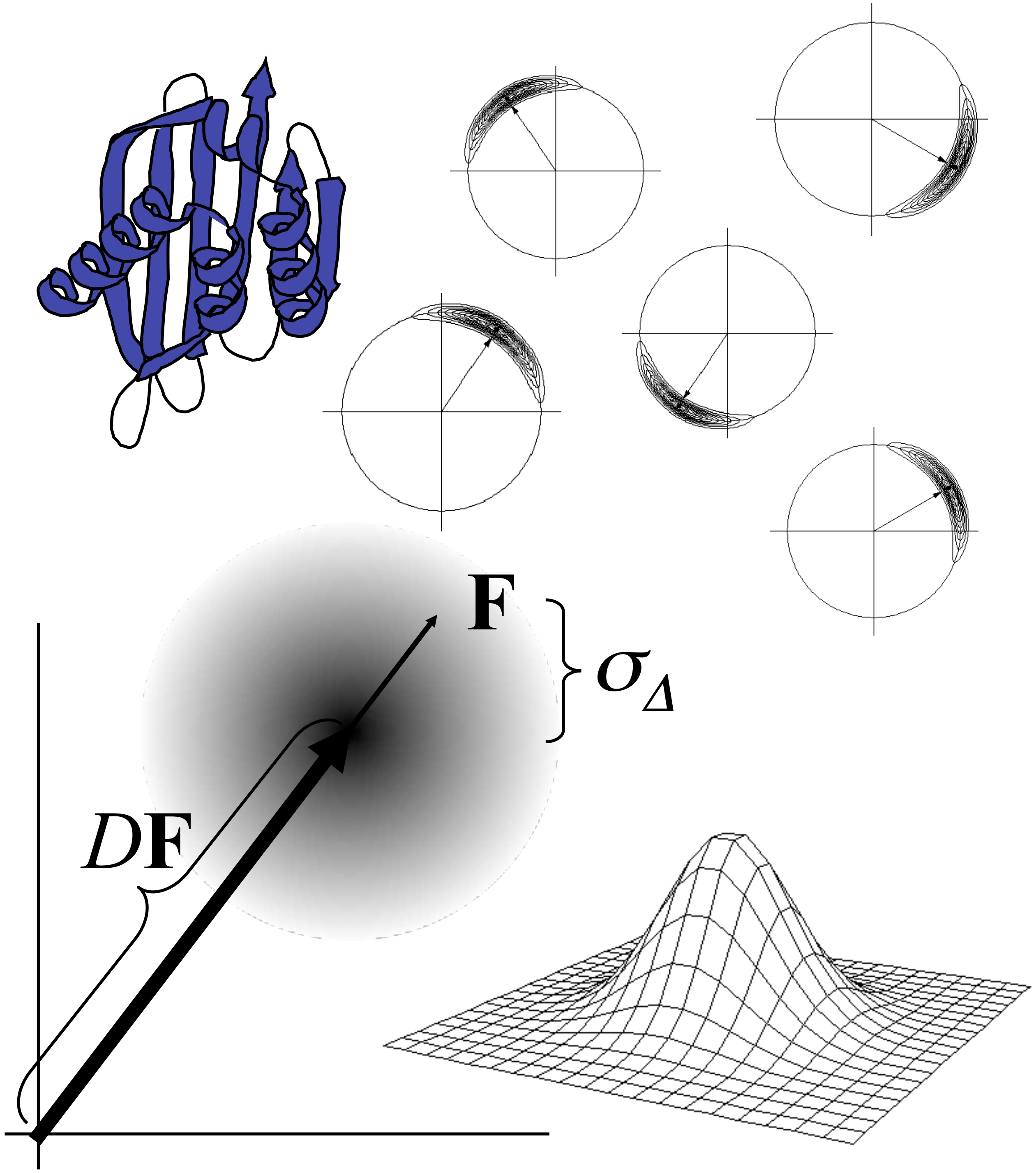
- Atomic errors give "boomerang" distribution of possible atomic contributions
- Portion of atomic contribution is correct





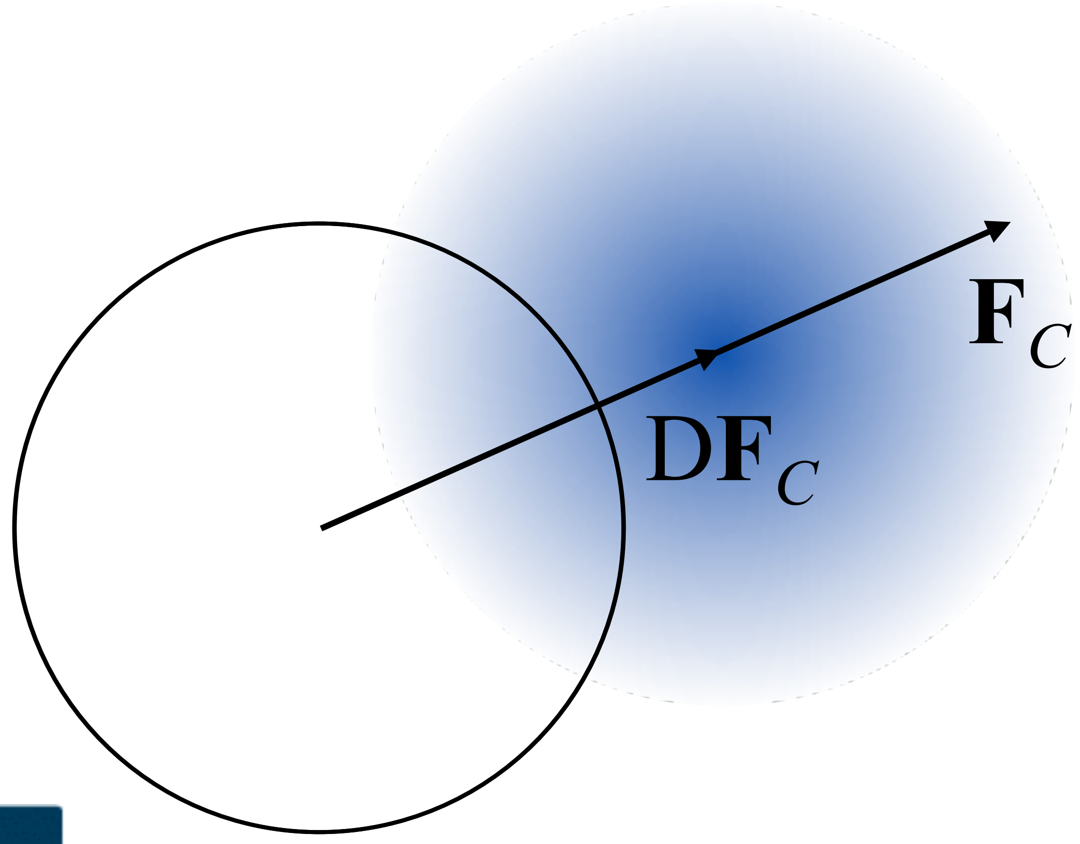
Structure factor with coordinate errors

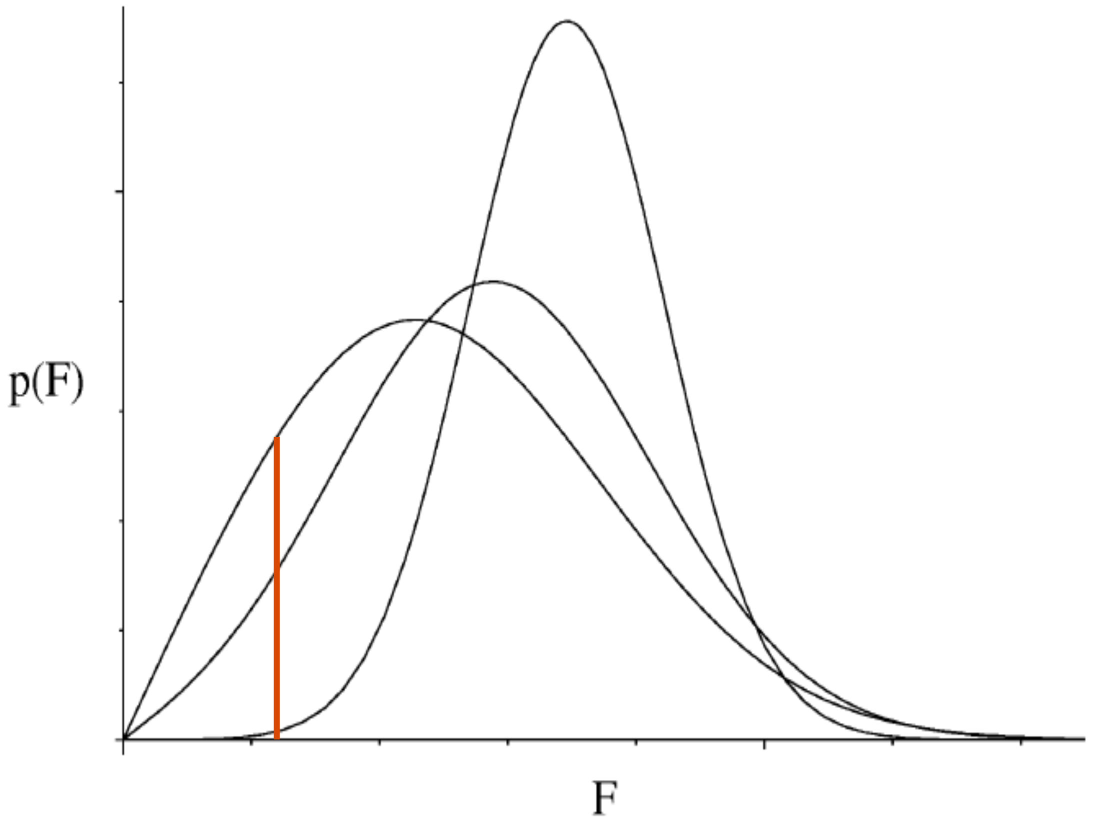
- Same direction as the sum of the atomic *f*
  - but shorter by 0< *D* <1
  - *D*=f(resolution)
- Central Limit Theorem
  - Many small atoms
  - Gaussian distribution for the total summed F
  - σ_Δ=f(resolution)



Calibrating the Likelihood Function

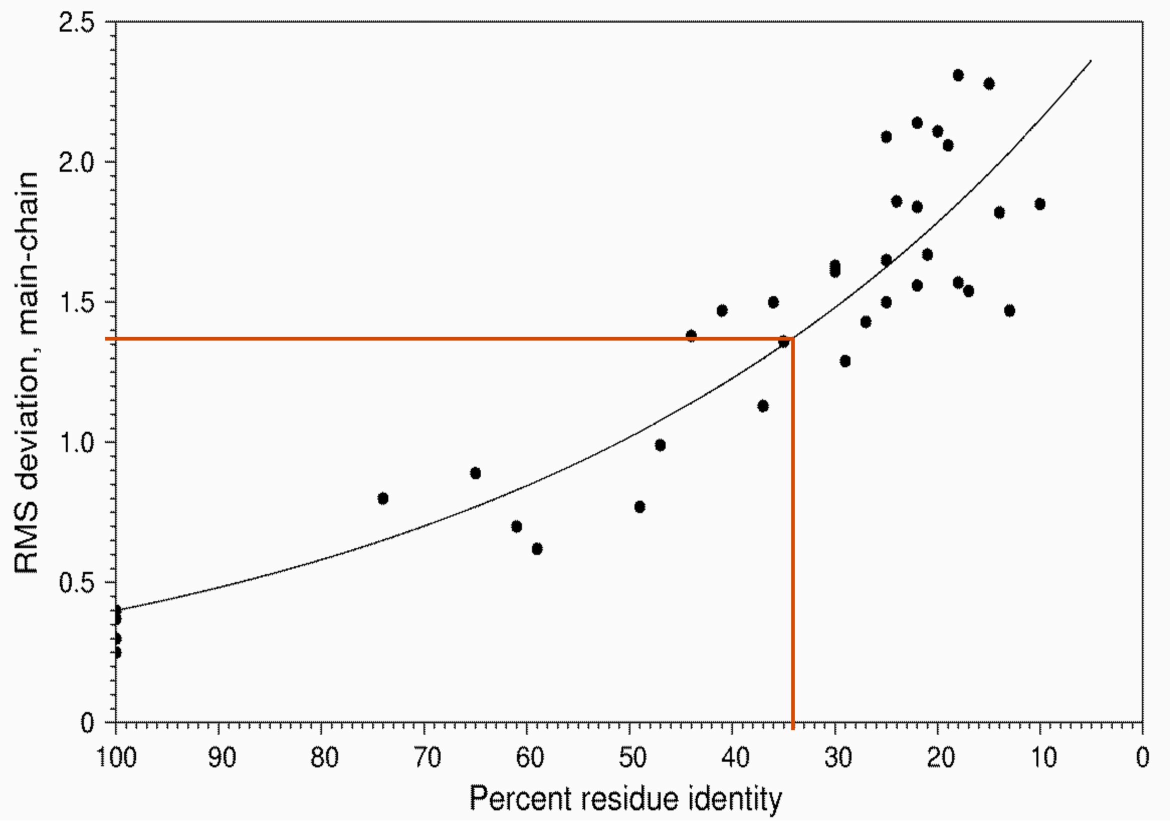
- Depends on the parameter σ_A
  - combined measure of model error and completeness
- For refinement, σ_A determined by comparing $|F_o|$ and $|F_c|$
  - $|F_c|$ unknown at the start of molecular replacement



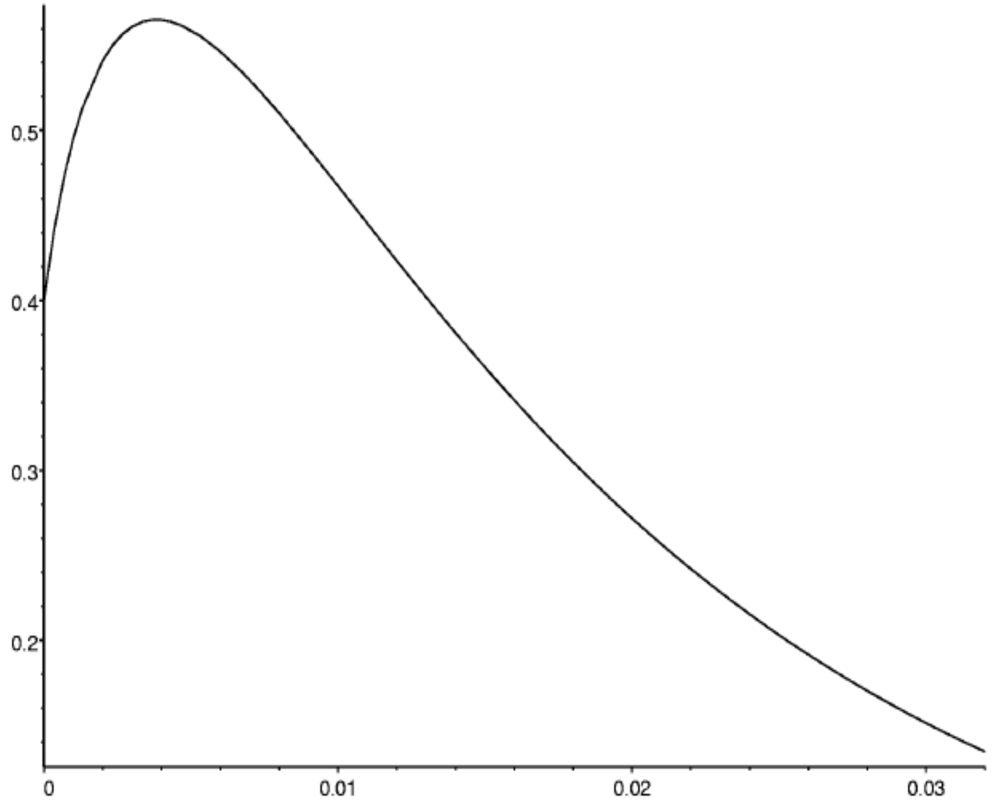


Defining an Error Model

- Depends on multiple factors: completeness, disordered solvent, model errors
- Chothia & Lesk (EMBO J., 1986) related sequence identity to rms deviation



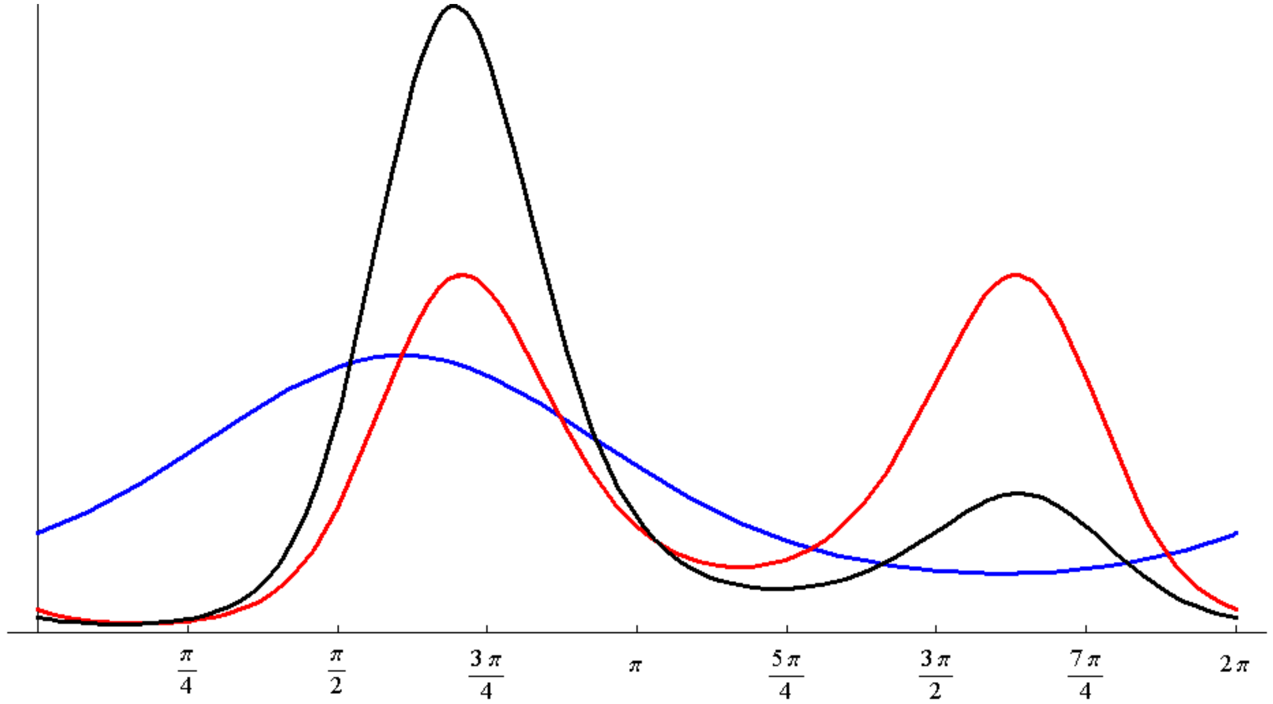
Relationship between identity and RMS deviation



SigmaA curve (error model) calculated from a given RMSD

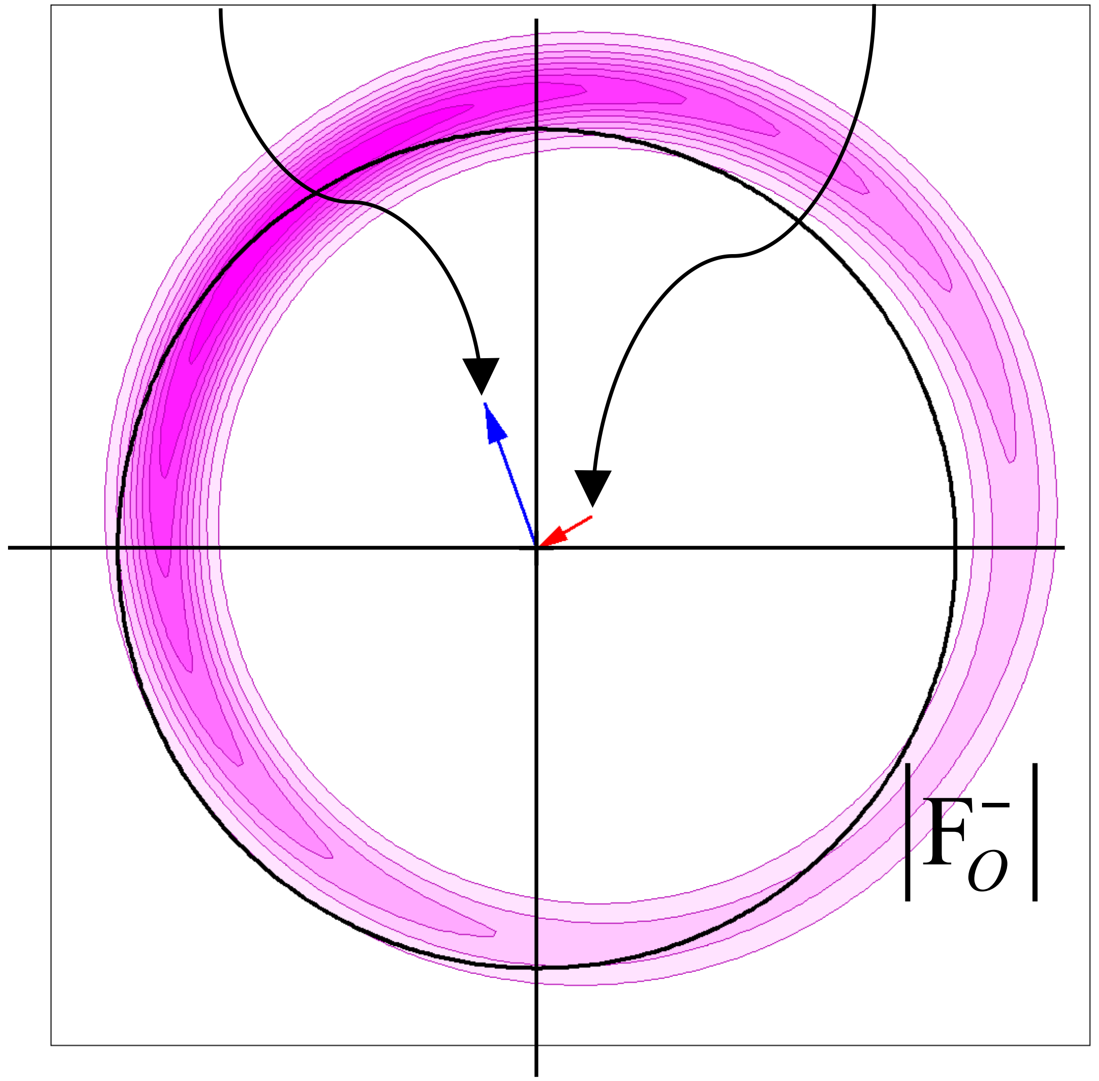
Combining MR and SAD

- Amplitudes from an MR solution can be treated as a heavy atom model in phasing



Expected value of $\mathbf{F}^{-*}$ $(\mathbf{H}^{-*})$

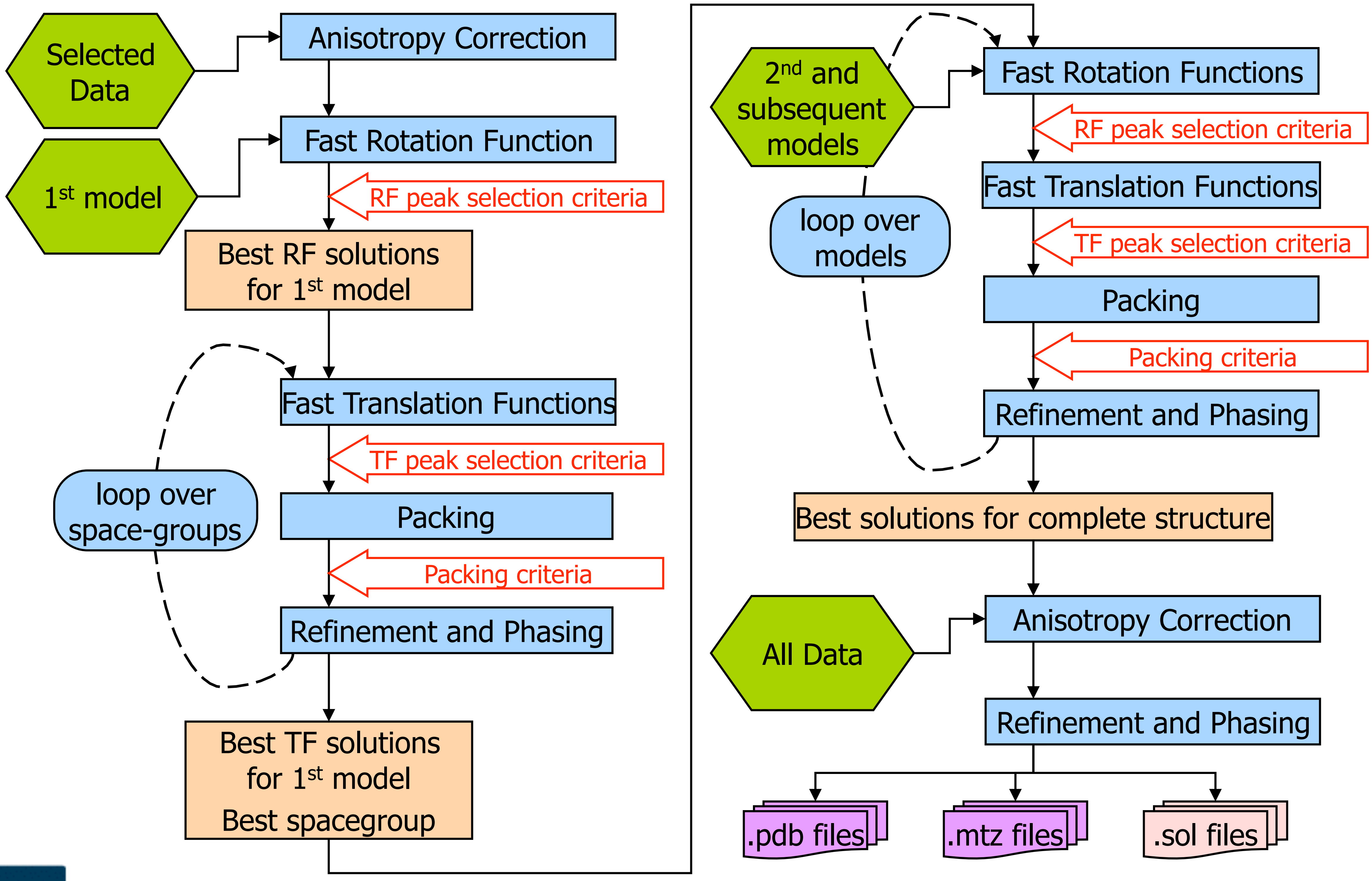
Expected difference between $\mathbf{F}^{+}$ and $\mathbf{F}^{-*}$



Automation in Phaser

- MR_AUTO mode
  - Searches over possible space-groups
  - Checks potential solutions for packing
  - Refines solutions away from search grid to optimal orientation and position
  - Uses parts of the structure already found to bootstrap the entire solution
- Protocol fine-tuned with difficult MR problems

Automated Molecular Replacement



BERKELEY LAB

Phenix

The Search Model

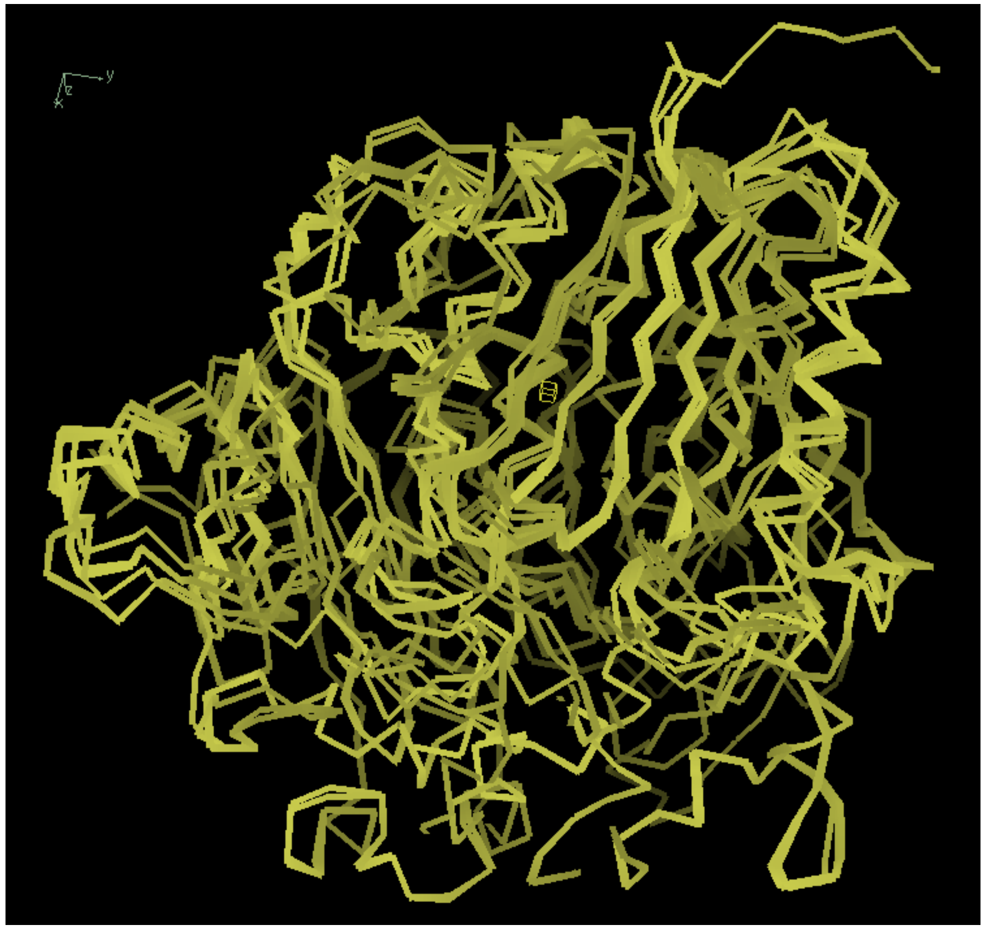
- There are many variables in constructing a search model:
  - Sequence alignment methods
  - Domain identification/juxtaposition
  - Sequence editing
    - Poly-ala, "mixed", "all-atom", C-alpha only
    - Combinatorial selection of models for ensembles
  - Perturbation along normal modes
- Must select those to use from potential models
  - Single "best" model
  - Ranking of models for MR trials
  - Use multiple models simultaneously

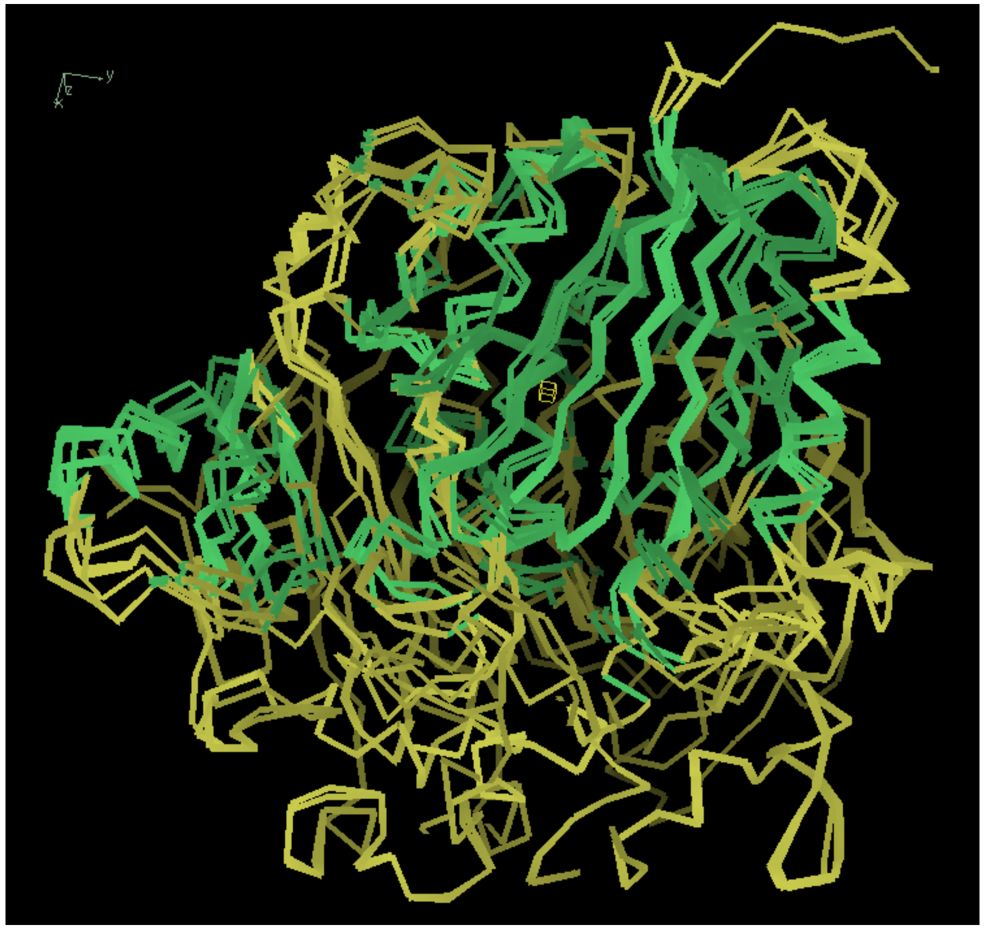
Model Manipulation in Phenix

- Sculptor
  - use sequence alignment to:
    - trim parts of template not in target
    - adjust B-factors of poorly-conserved regions
  - use surface accessibility to:
    - adjust B-factors of surface regions
- Ensembler
  - multiple structure superposition to make ensemble of possible models

Ensembler

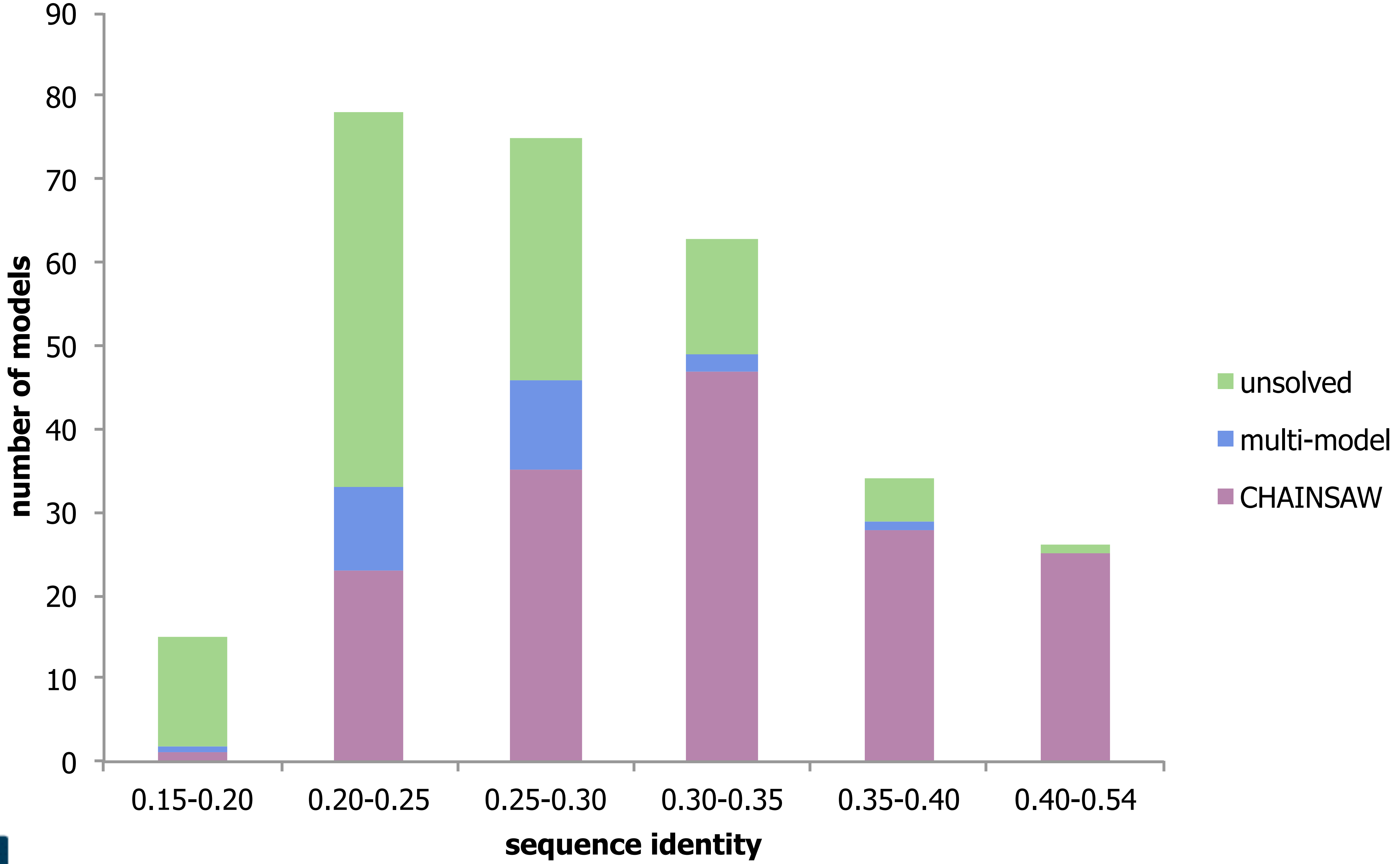
- Initial alignment with SSM or Muscle
- Iterative weighting of structural alignment
- Trim regions that are not conserved among models





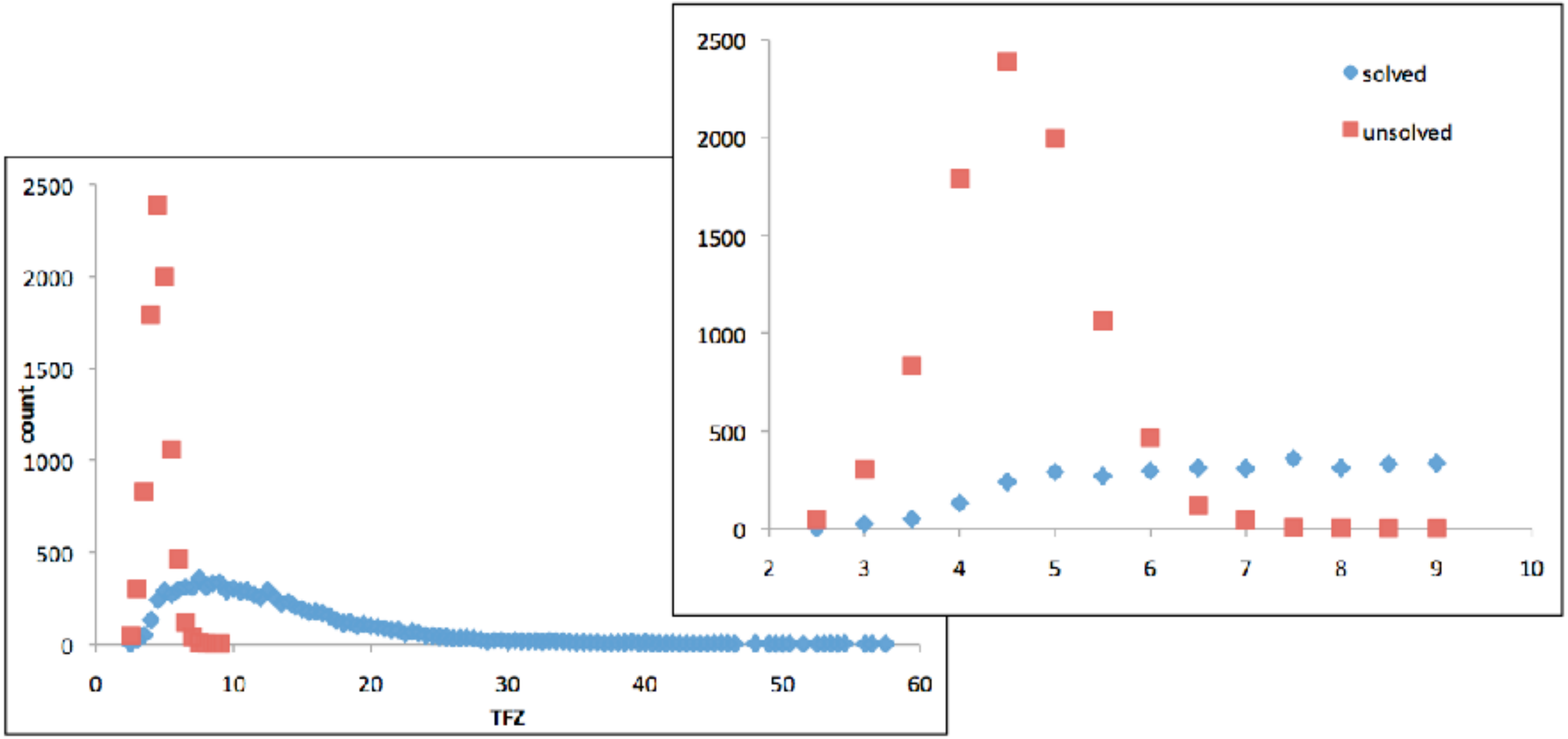
Multi-model Strategy with Sculptor/Ensembler

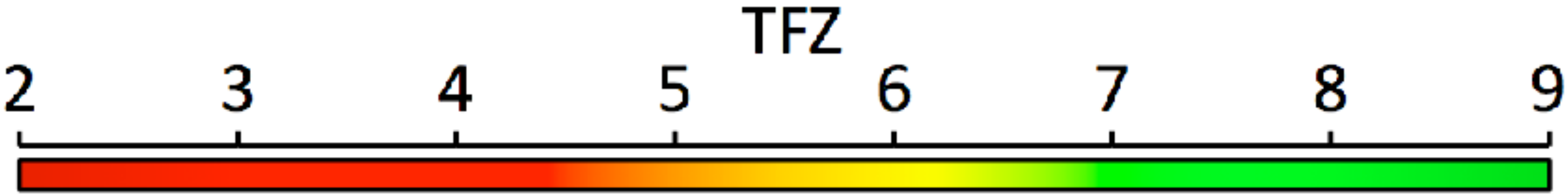
Calculations with CLUSTALW alignments



Phenix

Has the Molecular Replacement Worked?





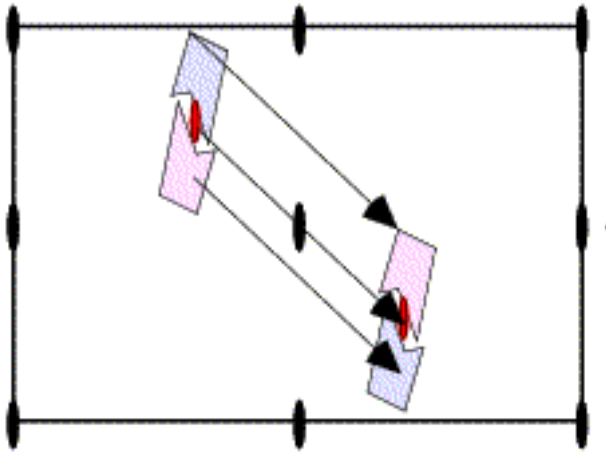
Rob Oeffner, Cambridge

Some Limitations of Molecular Replacement

- Unusual intensity distributions frustrate standard likelihood functions
  - Translational non-crystallographic symmetry
- What can be done if there is no search model in the protein databank?
- What to do when a solution is found but cannot be used to rebuild/refine the structure?

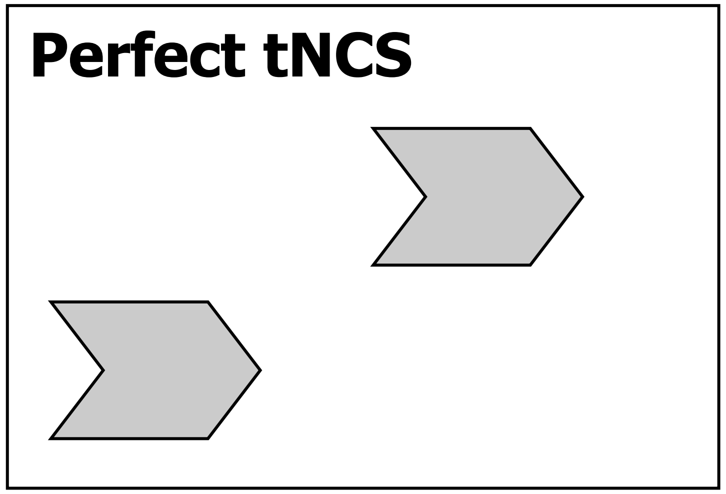
Translational NCS

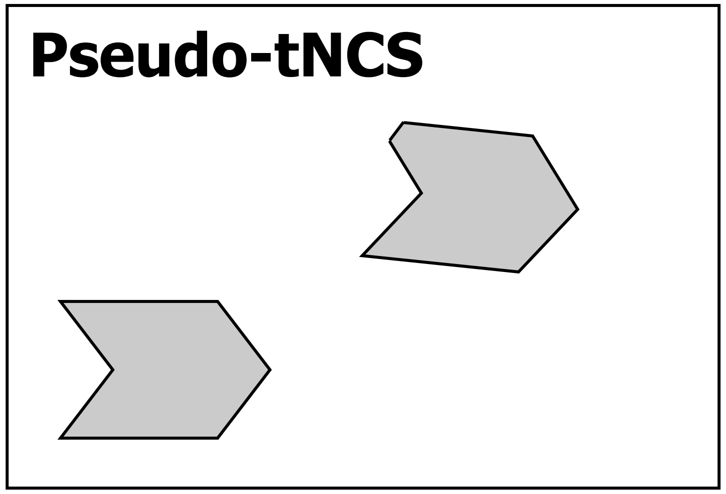
- Non-crystallographic symmetry is found in about 1/3 to 1/2 of crystal structures
- Often parallel to crystallographic symmetry axis
  - combination gives translational NCS (tNCS)
- Largest class of problems where default maximum likelihood functions fail
  - changes expected intensities, but not modelled

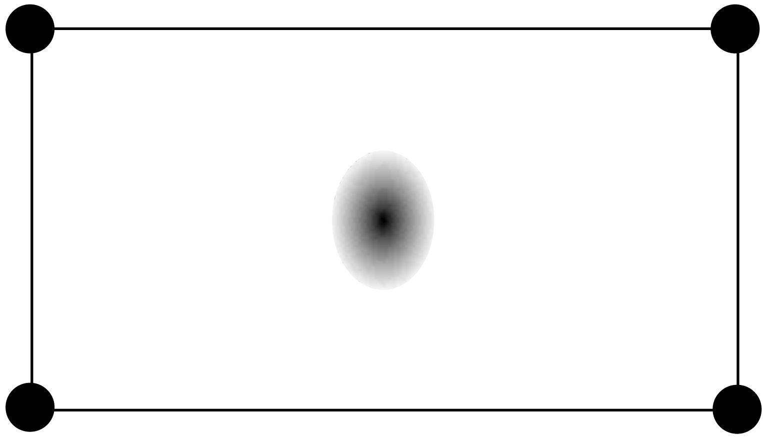


Pseudo-translational NCS

- tNCS is not perfect
  - There is usually a rotational component (ncsR)
  - There is non-isomorphism between structures
    - Differences in coordinates and scattering
    - Gives rise to D values (ncsD)
  - Vector (ncsT) often different slightly from cell or centering translation
    - have to refine the exact translation, perhaps test alternatives







Modelling pseudo-translational NCS

- Generalized ε-factor

$$\varepsilon_{hkl} = f\left(ncsD_s, G_{s,ncsR}, ncsT, \text{symmetry}\right)$$

- The ε-factors are no longer integers
- The ε-factors are found by maximizing the probability of the data
  - Probability described by the Wilson distribution
  - Similar to anisotropy correction

$$P_{tNCS}\left(F_{hkl}\right) = \frac{2F_{hkl}}{\varepsilon_{hkl}\Sigma_N^{hkl}} \exp\left(-\frac{F_{hkl}^2}{\varepsilon_{hkl}\Sigma_N^{hkl}}\right)$$

BERKELEY LAB

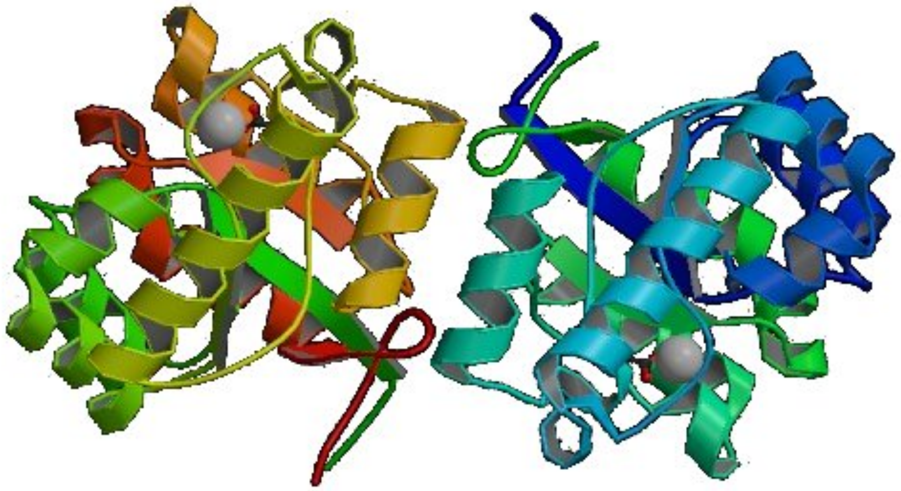
Phenix

Example Detection and Refinement

```
-------------------------------
PSEUDO-TRANSLATIONAL NCS VECTOR
-------------------------------

  Patterson Symmetry: P -1
  Resolution of All Data (Number):      28.93 - 1.90 (47848)
  Resolution of Patterson (Number):      10.00 - 5.00 (2319)
  There were 2 non-origin distinct peaks (i.e. more than 15 angstroms from the origin)

  46.6% origin:   FRAC 0.250 0.500 0.750   (ORTH   -7.5   16.3   42.7)

  31.3% origin:   FRAC 0.500 0.000 0.500   (ORTH   22.0   -6.0   28.5)

...

  Pseudo-translational NCS rotation angle 1.44607 -2.0814 -1.66689
  for pseudo-translational NCS translation vector 0.245175 0.493209 0.742281
  D corresponding to RMS deviation of NCS related structure:
     Range (low resolution - high resolution): 0.9009 - 0.3886
```

Example - Acetylxylan Esterase

- Problem case from Gideon Davies, York
  - P212121 crystal form
  - Two molecules in ASU
  - Related by tNCS (0.38, 0, 0.5)



Taylor et al JBC April 21 2006

- Attempt solution with Phaser MR_AUTO
  - First RF gives a weak signal
  - First TF fails to find correct translation
    - hence second RF and second TF fail

Results

	No tNCS correction	pure tNCS	pseudo tNCS
RF Correct	4.93	4.85	5.46
RF Top Incorrect	4.38	4.83	4.19
TF Correct	-	7.61	12.68
TF Top Incorrect	5.4	5.89	-

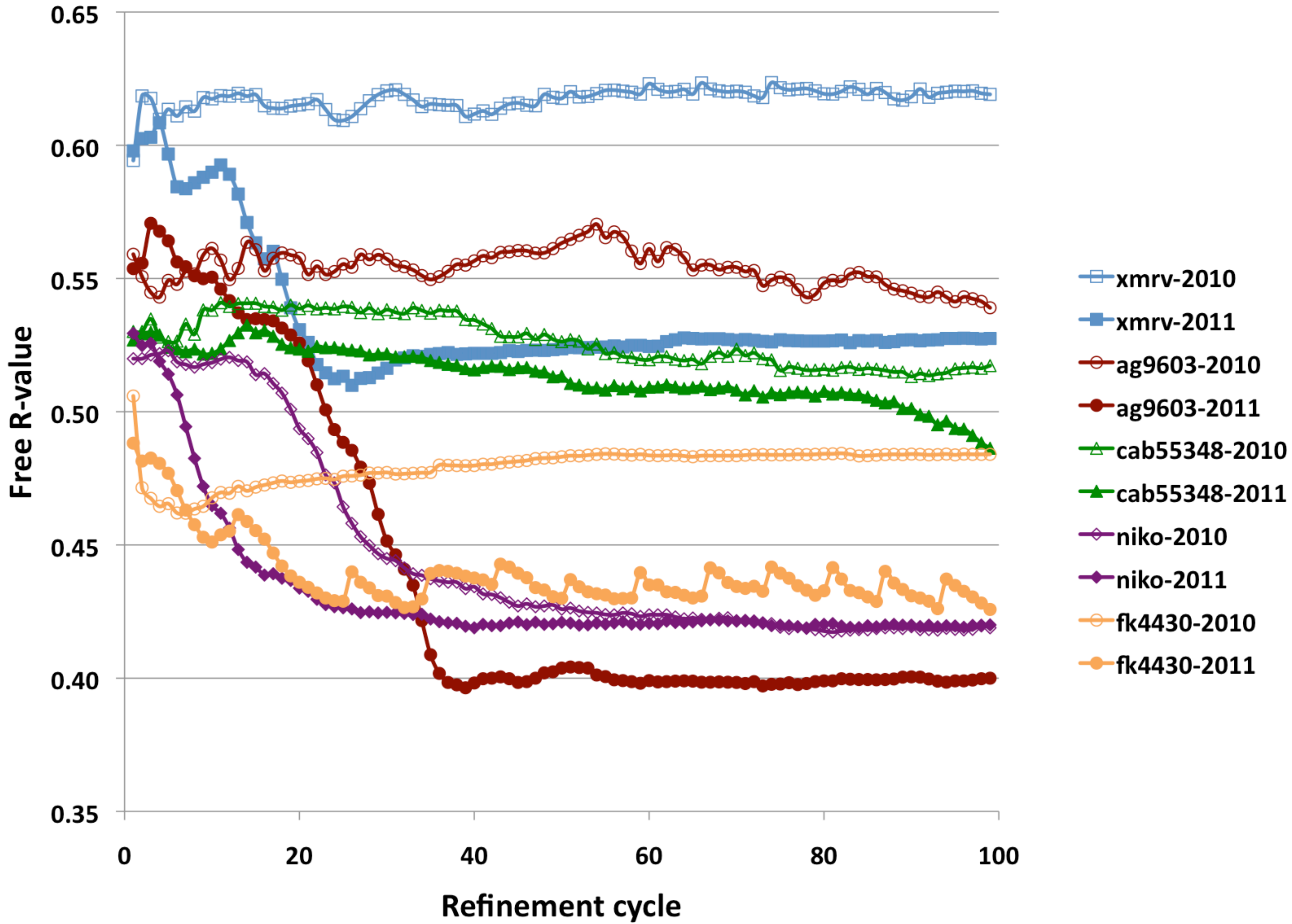
- Translation vector refines from 0.378, 0, 0.5 to 0.377, 0, 0.498
  - cancellation is less exact, especially for 0kl
- Rotation refines from 0 to small rotation, mostly 1.8° around x
  - agrees well with final orientation difference
- *ncsD* values refine close to 1 (0.98 – 0.89)

Extending Molecular Replacement

- For low sequence similarity models often a solution can be found, but the model cannot be used or refined to generate maps good enough to interpret
- How can we improve the model enough to generate phases for the true structure?
  - Modify the model using molecular modelling methods - “mr_rosetta”
  - Modify the model using the current electron density map - “morphing”

Extensive Refinement

- Refinement can improve some models

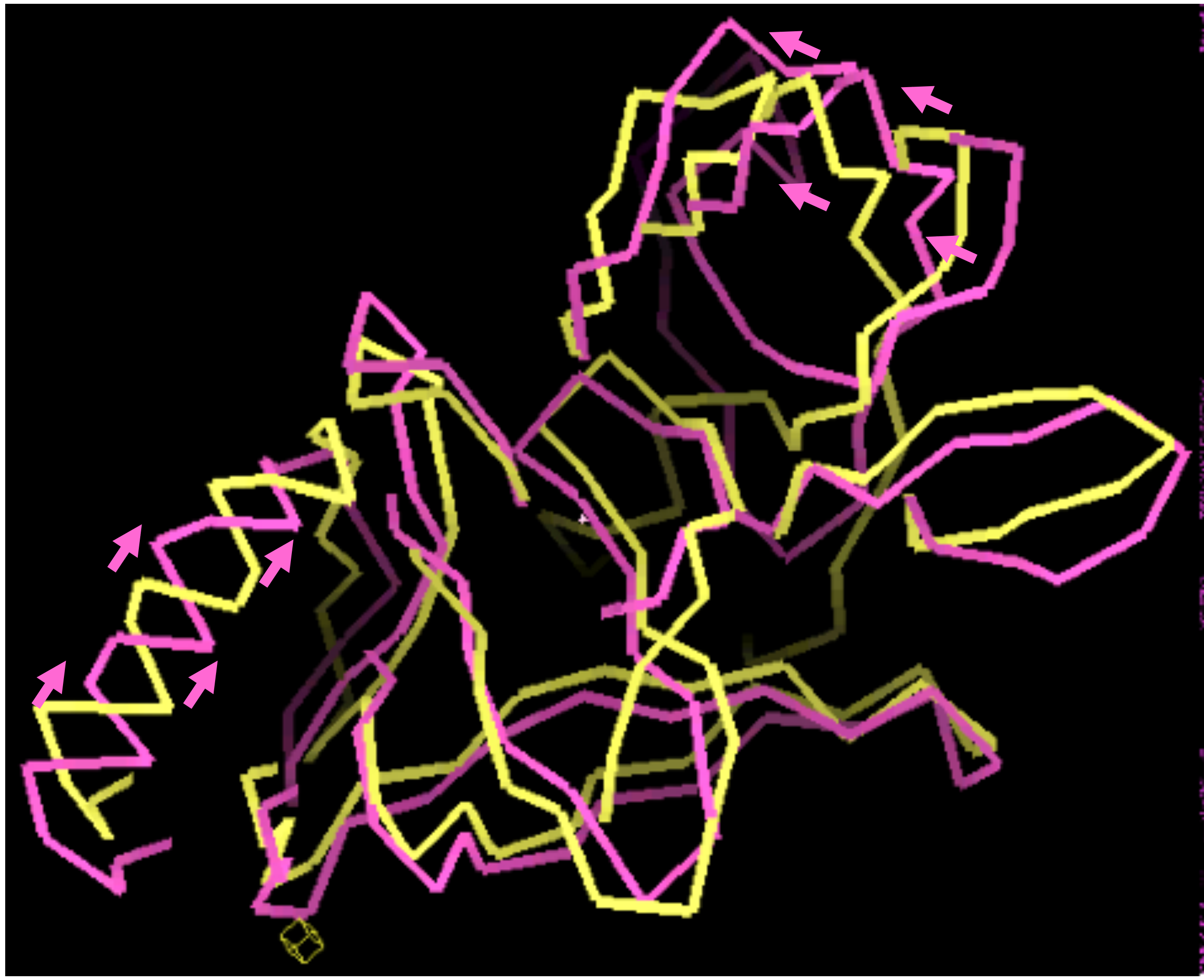


Phenix

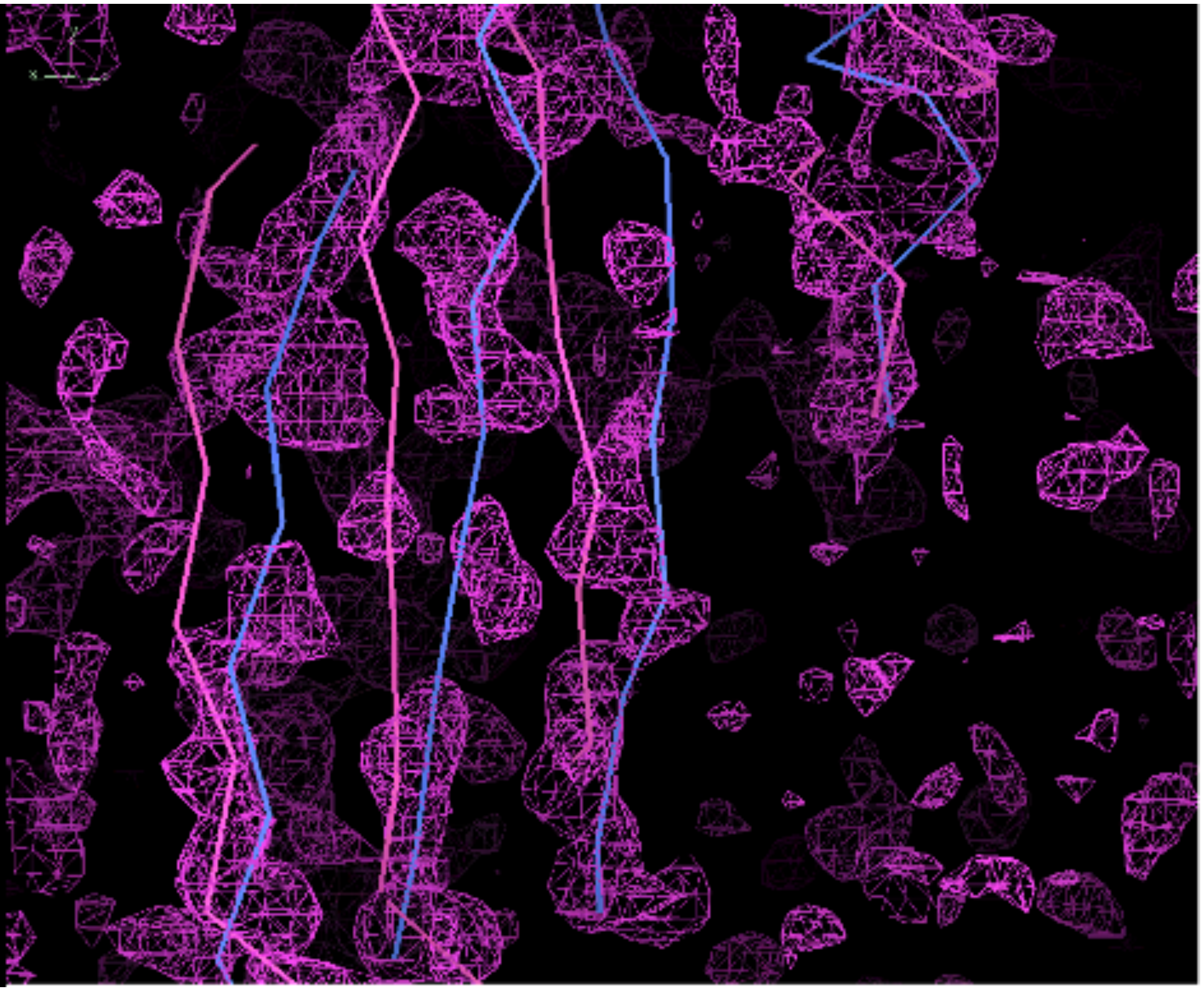
Tom Terwilliger, Los Alamos National Laboratory

Difficult MR

- Model is different enough locally to generate very poor electron density maps



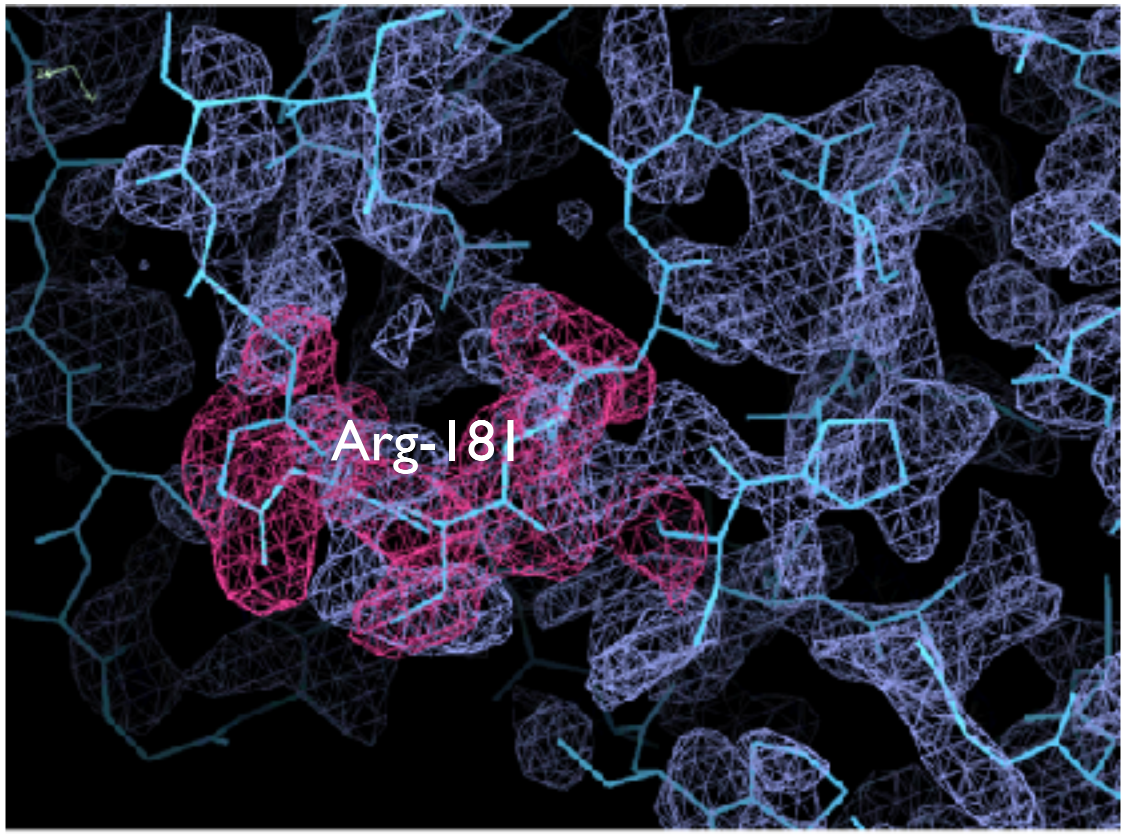
ag9603; NMR model (pink),
true structure (yellow)



cab55348; MR solution (blue),
true structure (pink)

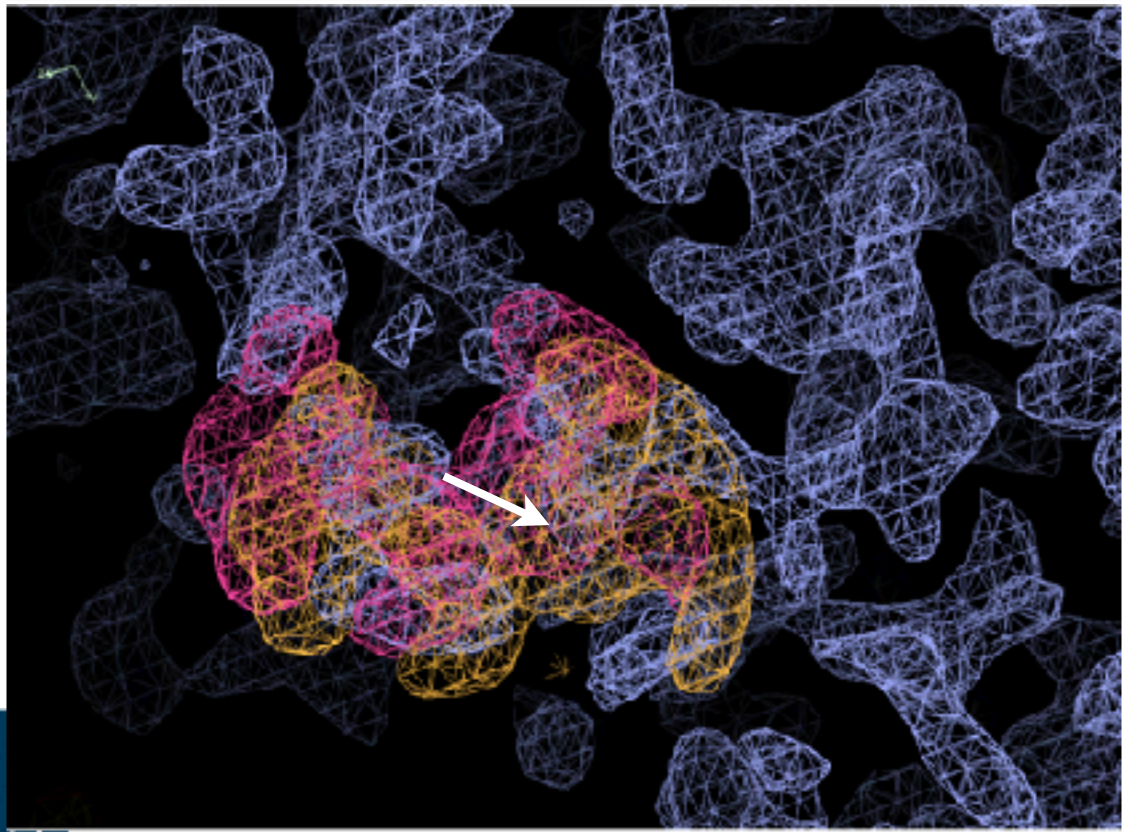
*Tom Terwilliger, Los Alamos
National Laboratory*

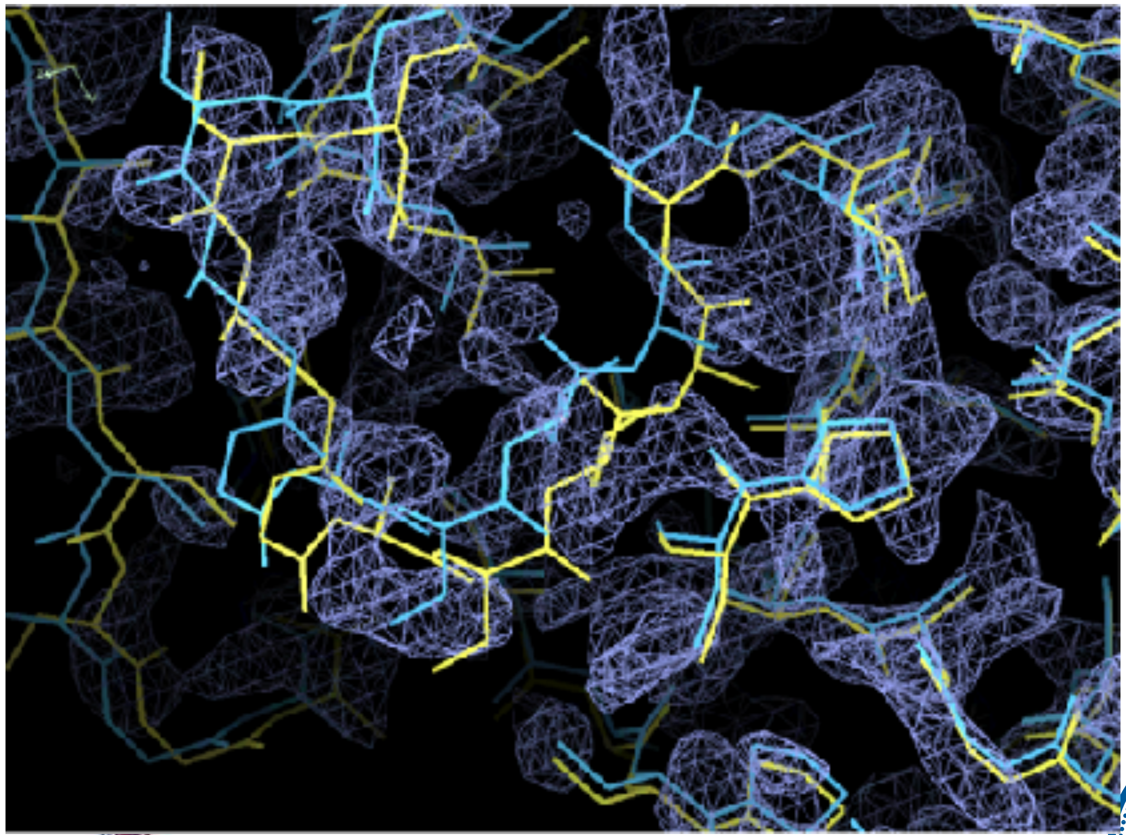
Morphing Procedure



- Identify local translation to apply to one C_α atom and nearby atoms
- Smooth the local translations in window of 10 residues
- Apply the smoothed translation to all atoms in the residue

Tom Terwilliger, Los Alamos National Laboratory

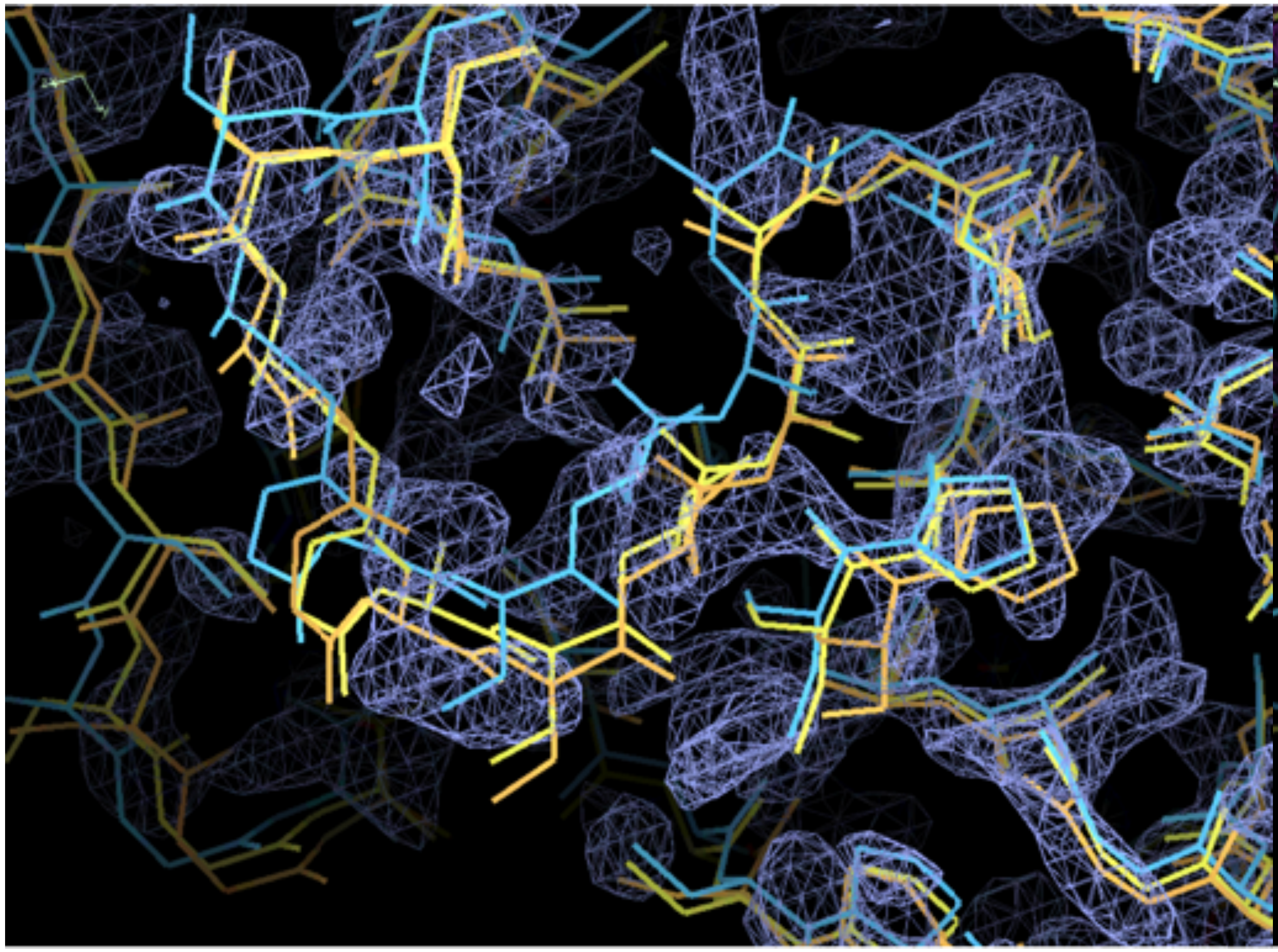




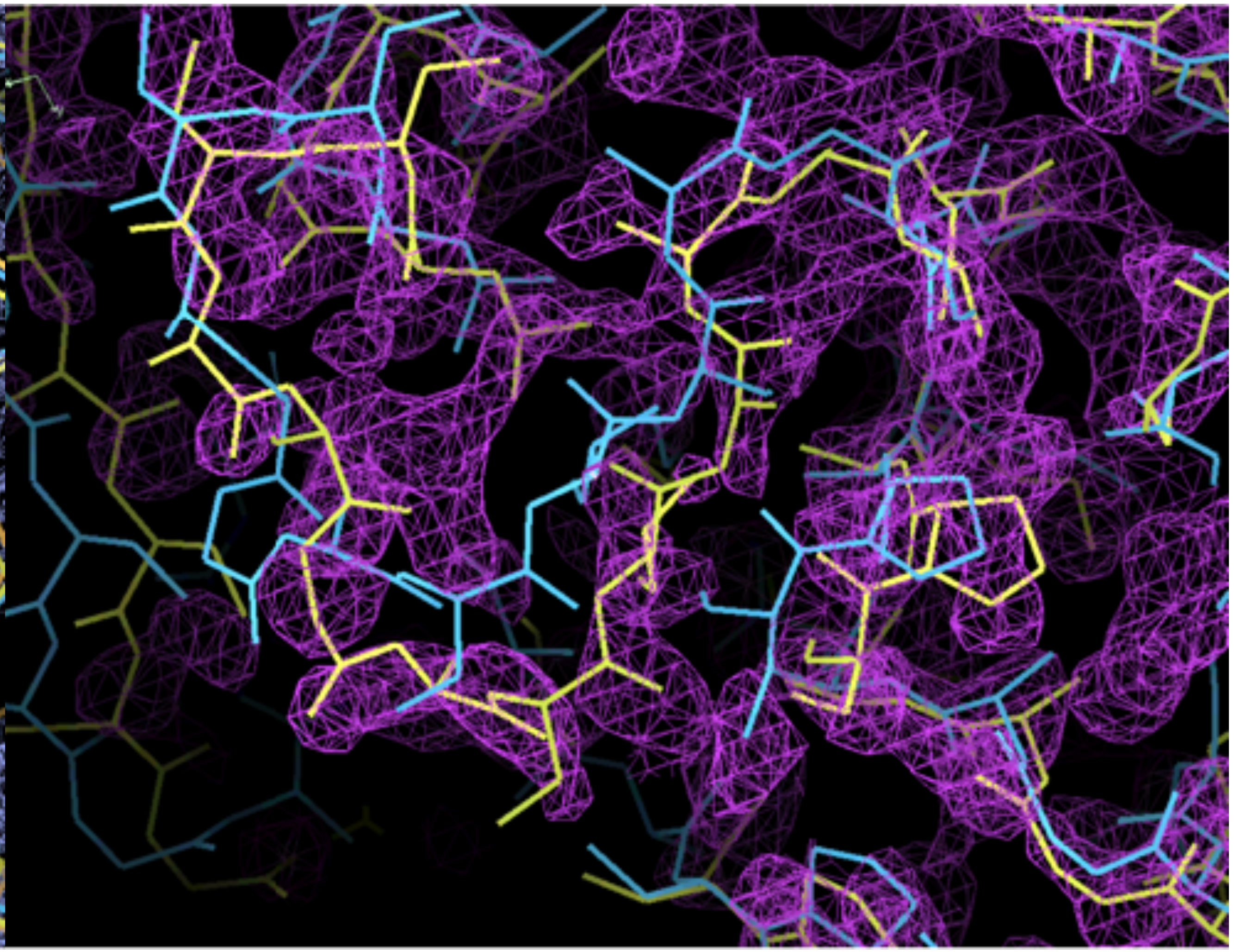
Morphing Procedure

- The geometry between the morphed fragments will be poor: standard refinement is applied to correct the model
- The process is iterated

Tom Terwilliger, Los Alamos National Laboratory



3PIC, 32% identity, (blue)
Morphed model (yellow)
Refined morphed model (orange)

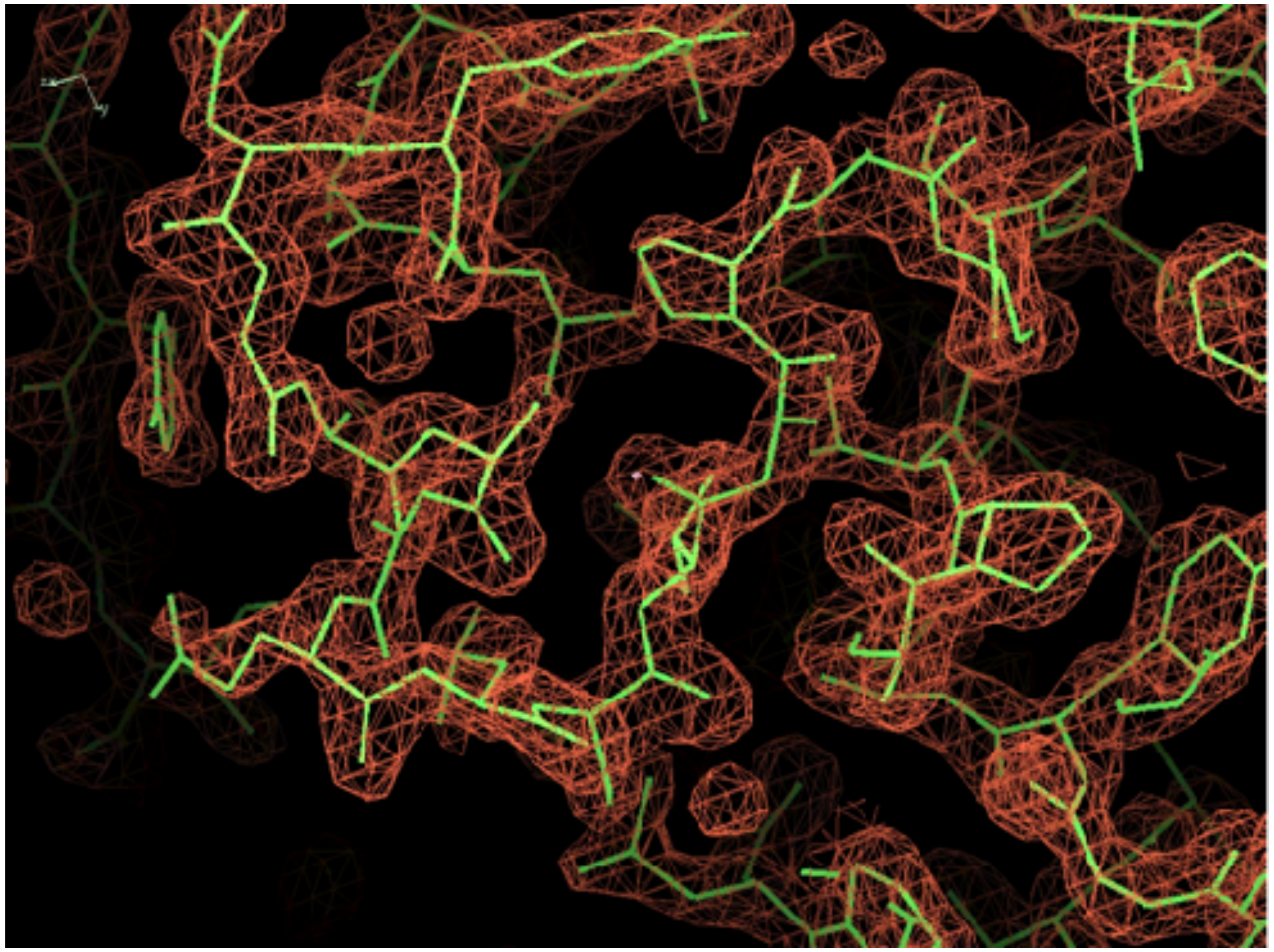


3PIC, 32% identity, (blue)
Refined morphed model (yellow)
Updated prime-and-switch map (purple)

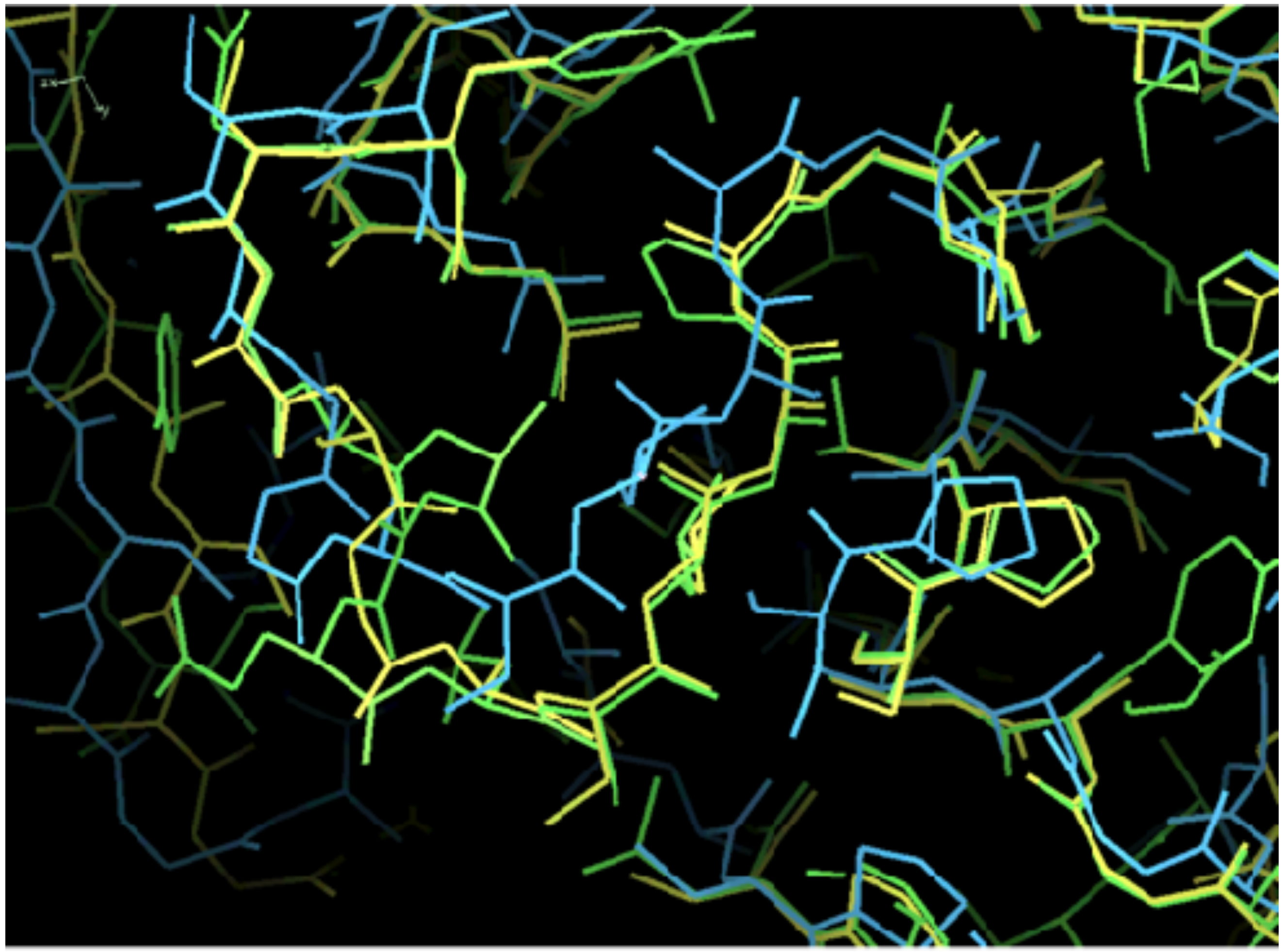
Improved Phases

- The map and morphed model can then be used as the input to automated building

Tom Terwilliger, Los Alamos National Laboratory



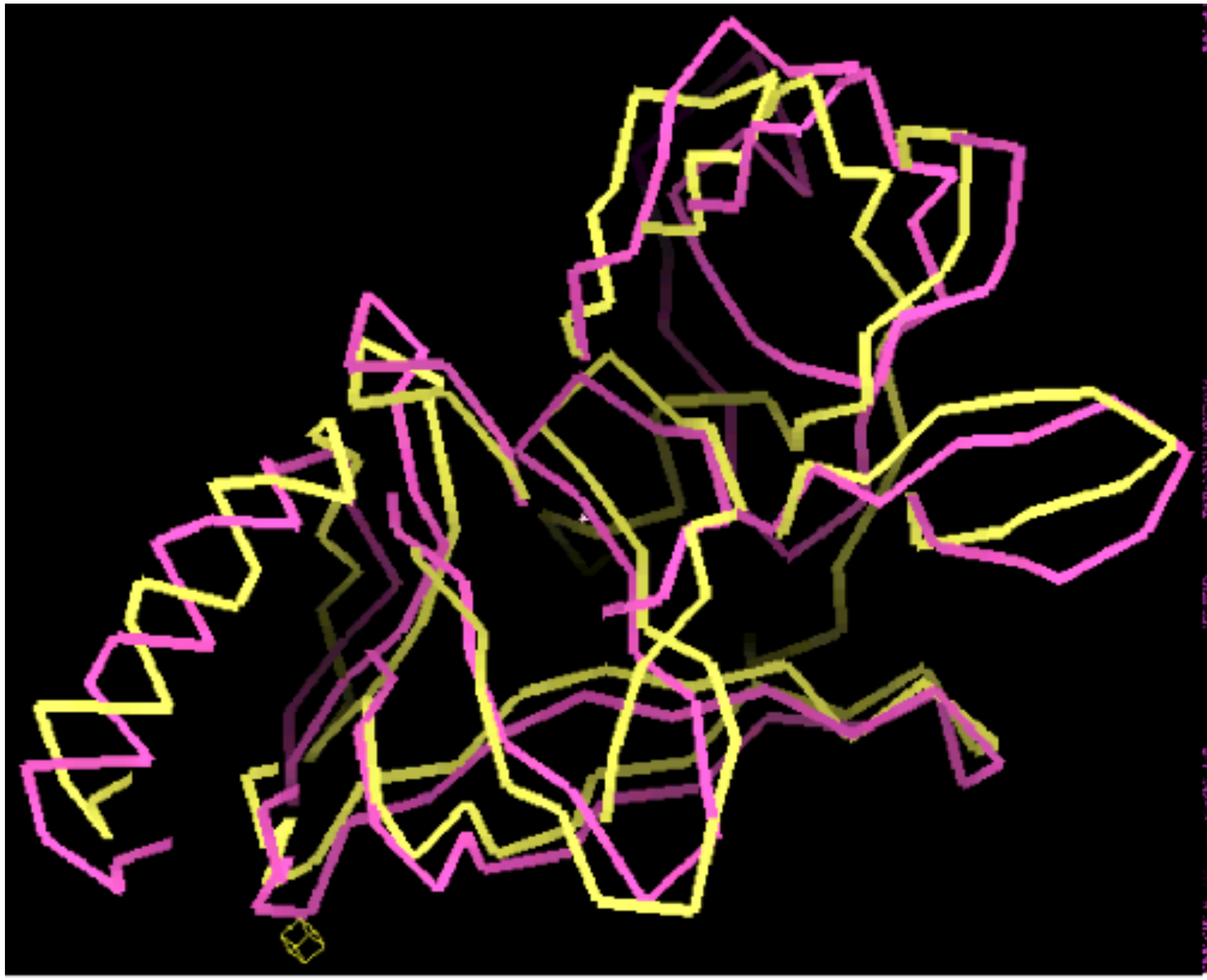
Autobuilt model (green)
Density modified map (red)



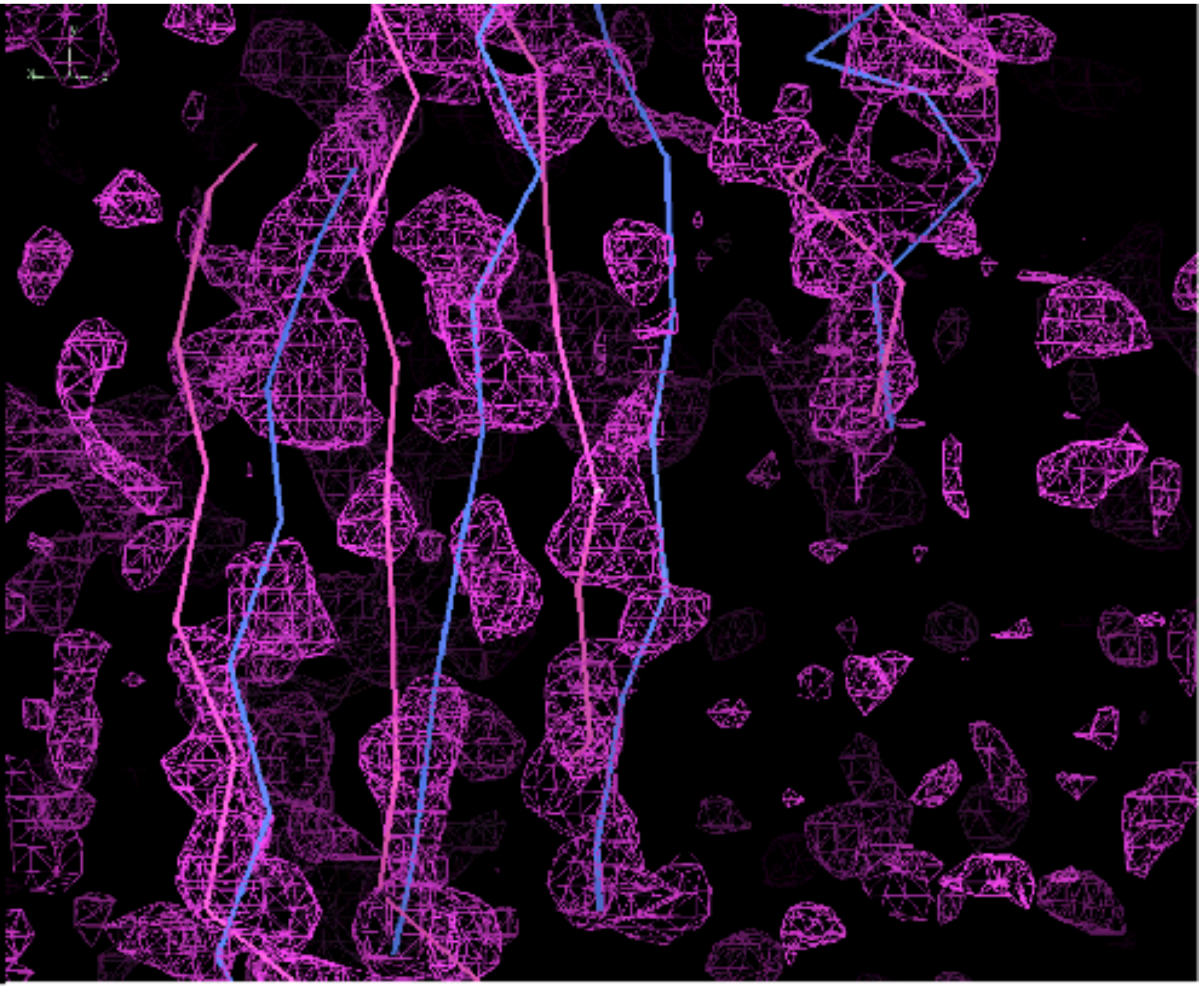
Starting model (blue)
Refined morphed model (yellow)
Autobuilt model (green)

Difficult MR

- Model is different enough locally to generate very poor electron density maps



ag9603; NMR model (pink), true structure (yellow)



cab55348; MR solution (blue), true structure (pink)

Tom Terwilliger, Los Alamos National Laboratory

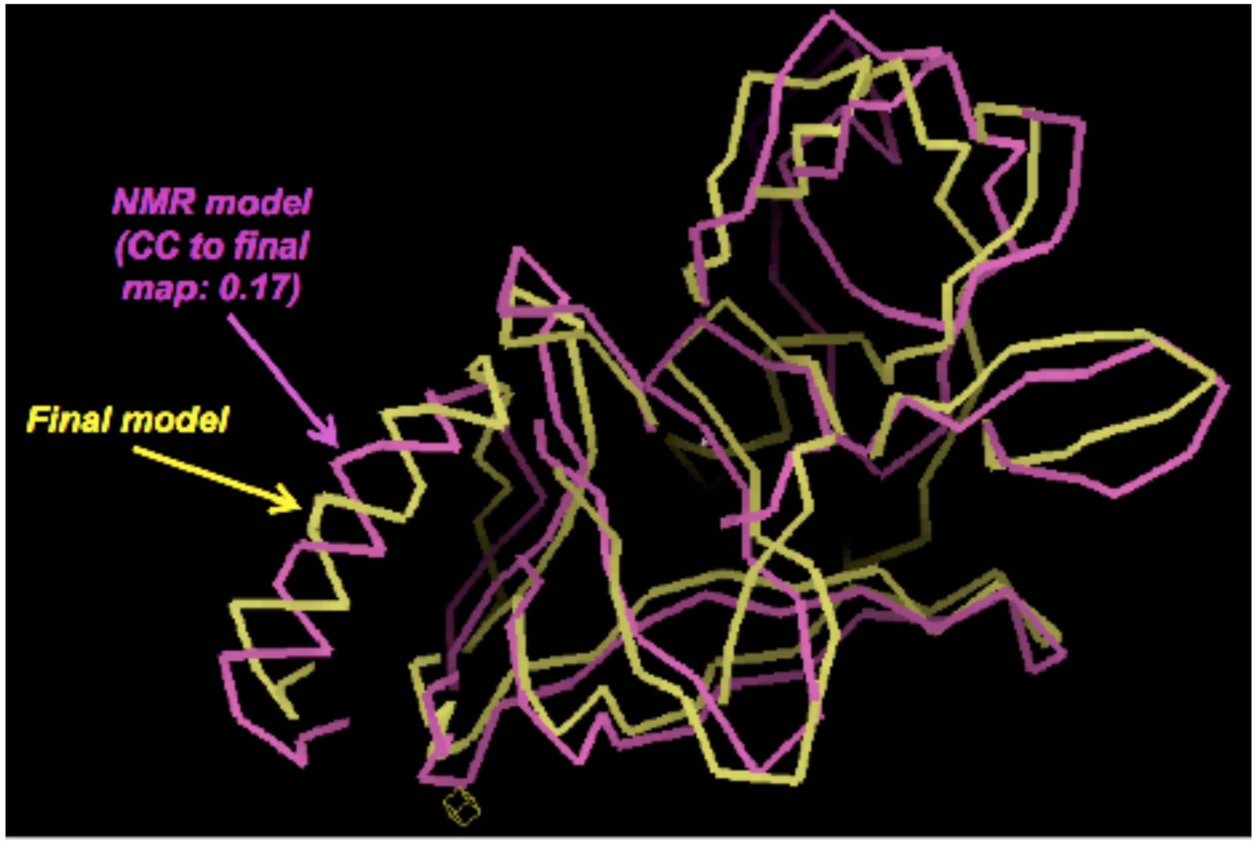
Making Use of Homology Modelling

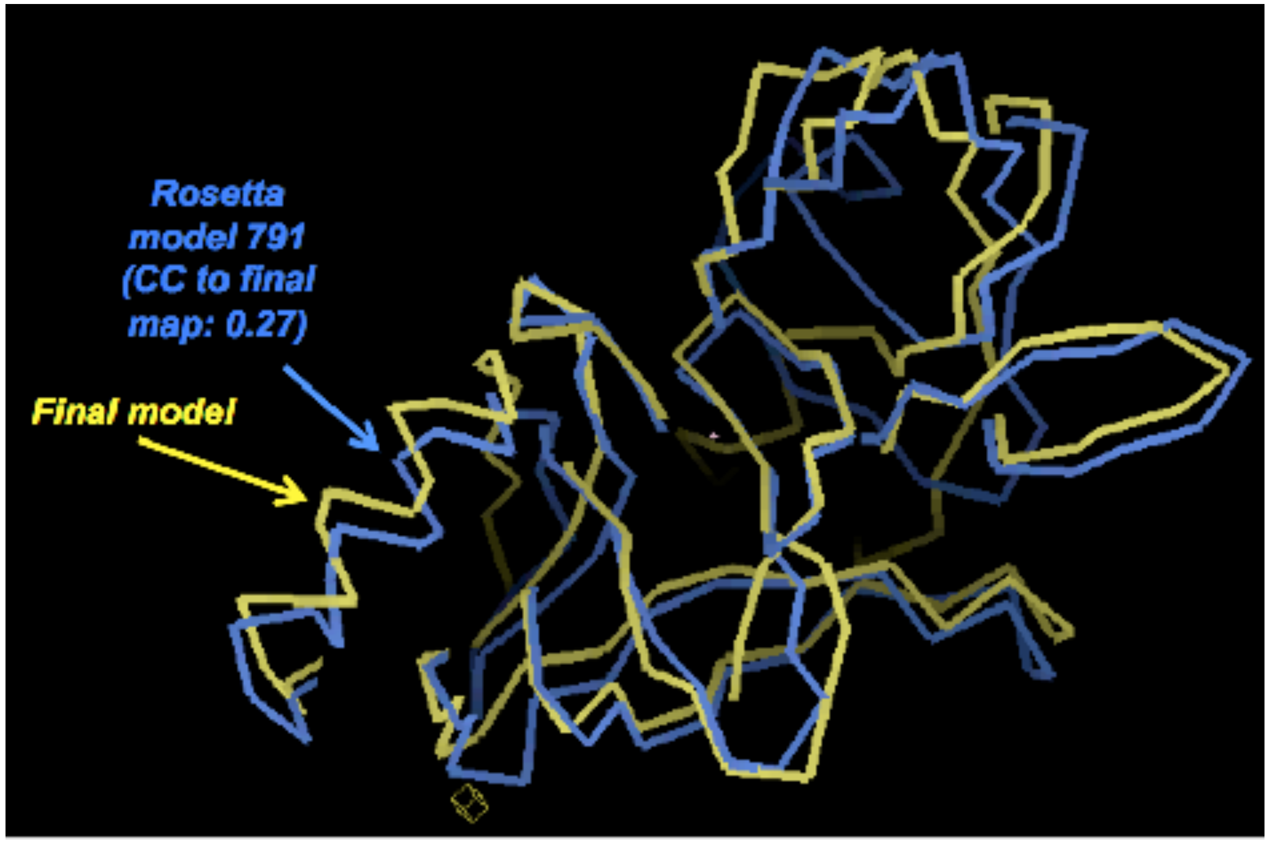
- Use homology modelling methods to improve the model

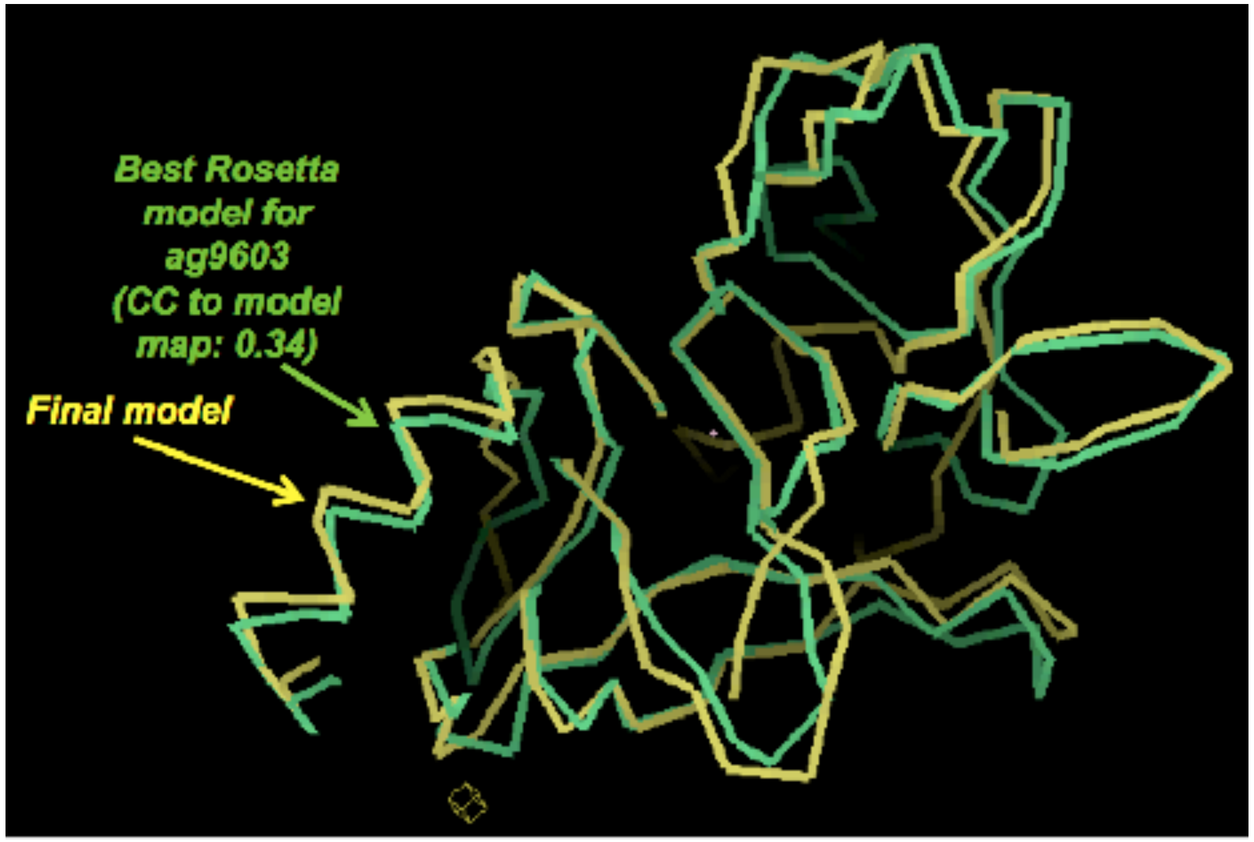
	Crystallographic model building (Phenix)	**Structure Modelling (Rosetta)**
Optimization	Interpretation of density patterns	Creating physically reasonable models
Model building approach	Search for fragments (e.g. helices) in density	*Ab initio* or homology modelling
Fragment libraries	3-residue library	3- and 9-residue libraries
Target	Fit to density	Rosetta force field (density optional)
Refinement target	Reciprocal space likelihood function plus geometry	Rosetta force field (density optional)

Tom Terwilliger, Los Alamos National Laboratory

Does Homology Modelling Help?

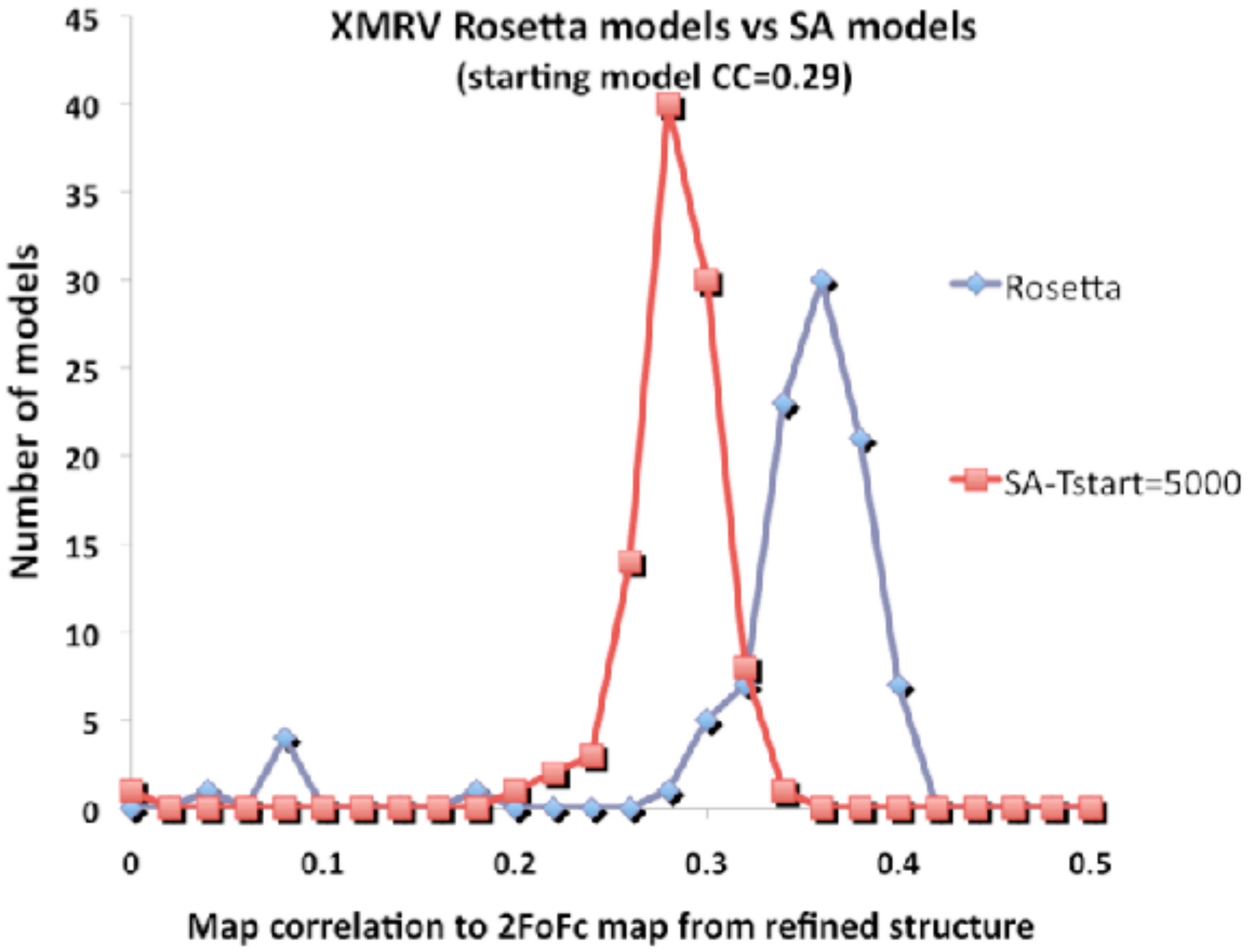






Tom Terwilliger, Los Alamos National Laboratory

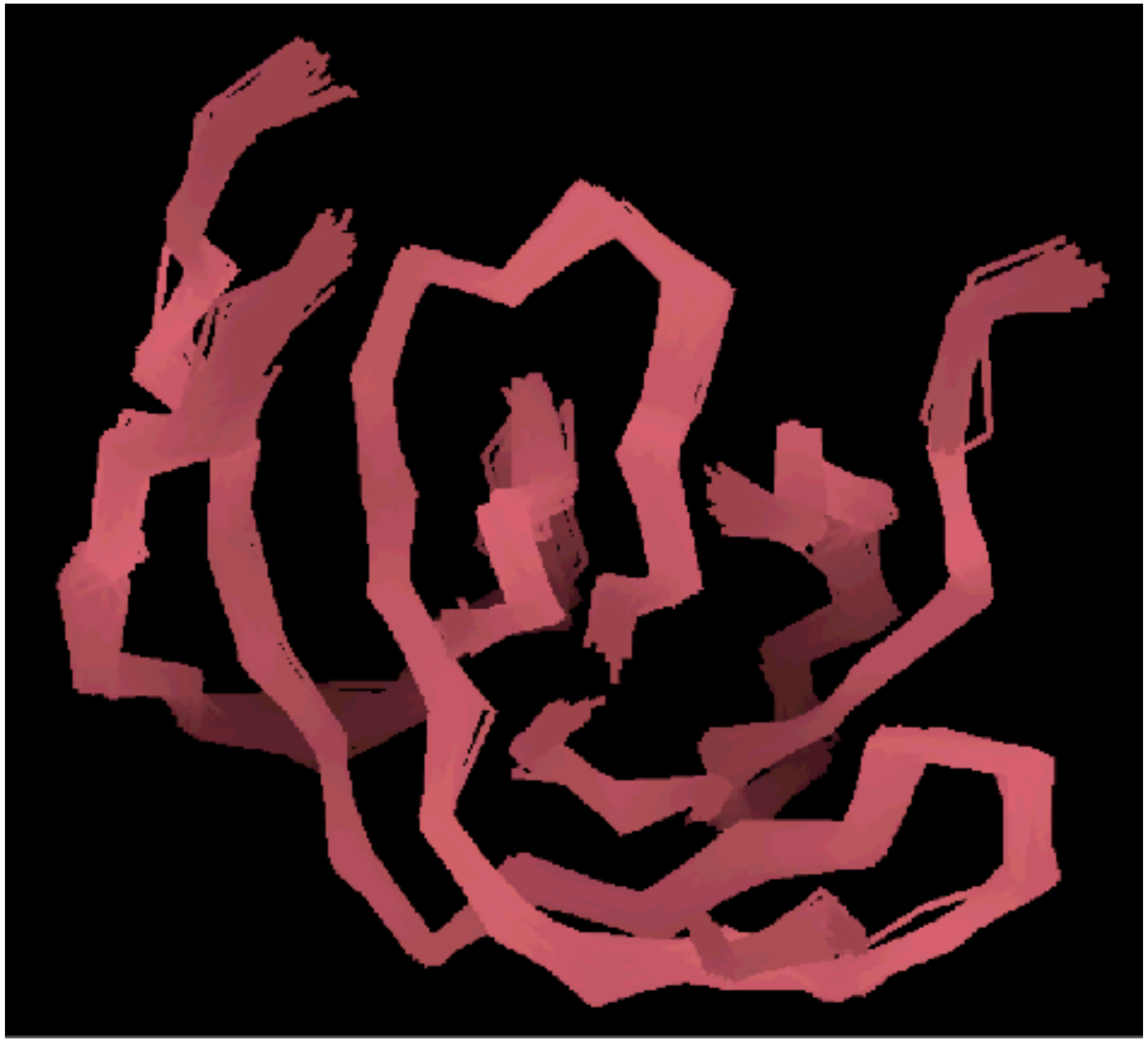
Comparison with Refinement (SA)



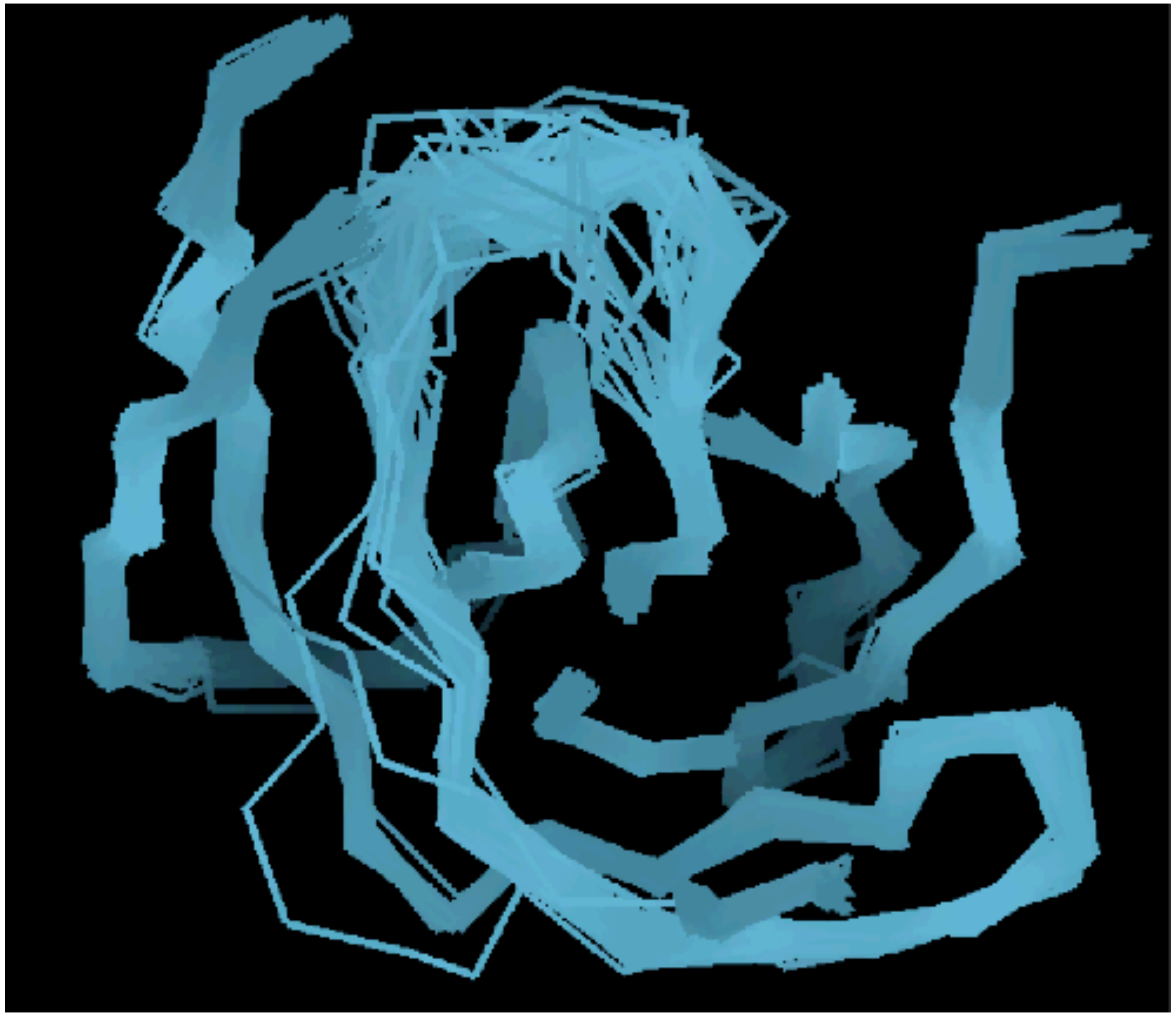
Tom Terwilliger, Los Alamos National Laboratory

Comparison with Refinement (SA)

- Rosetta can explore more of conformation space



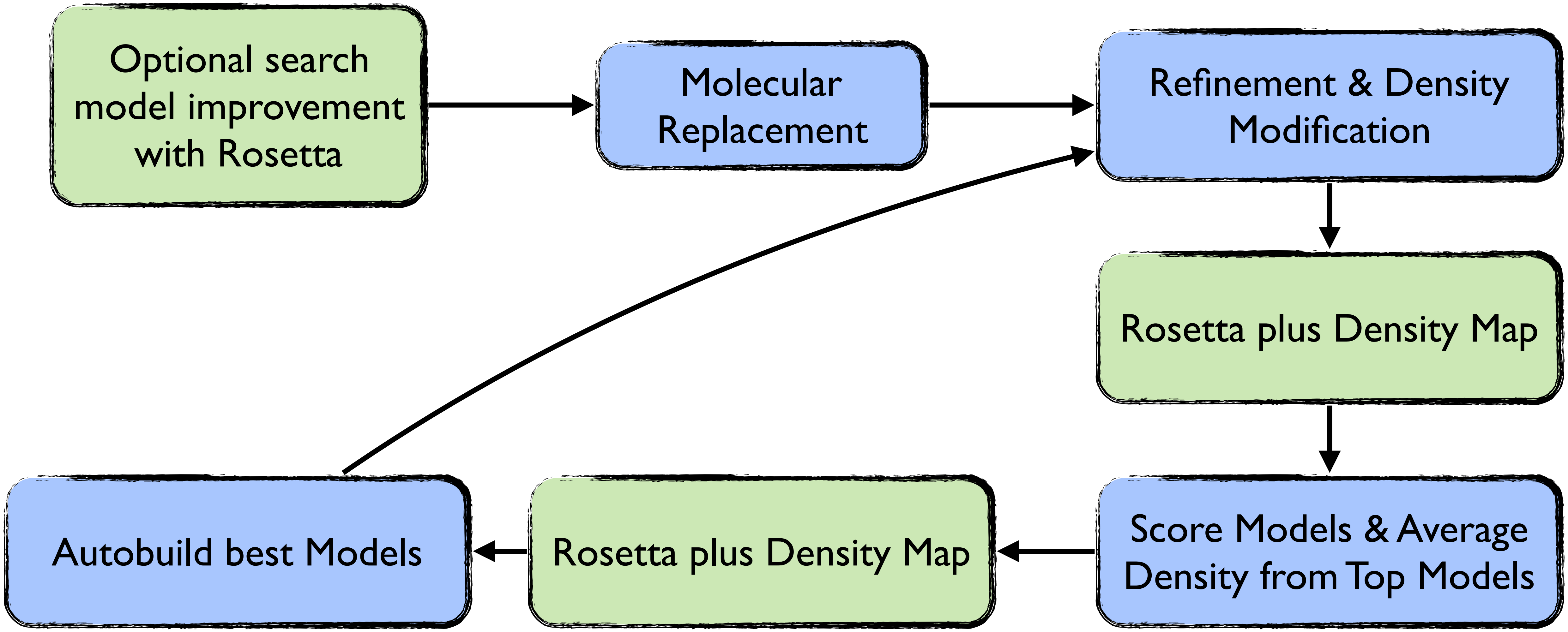
100 models from annealing



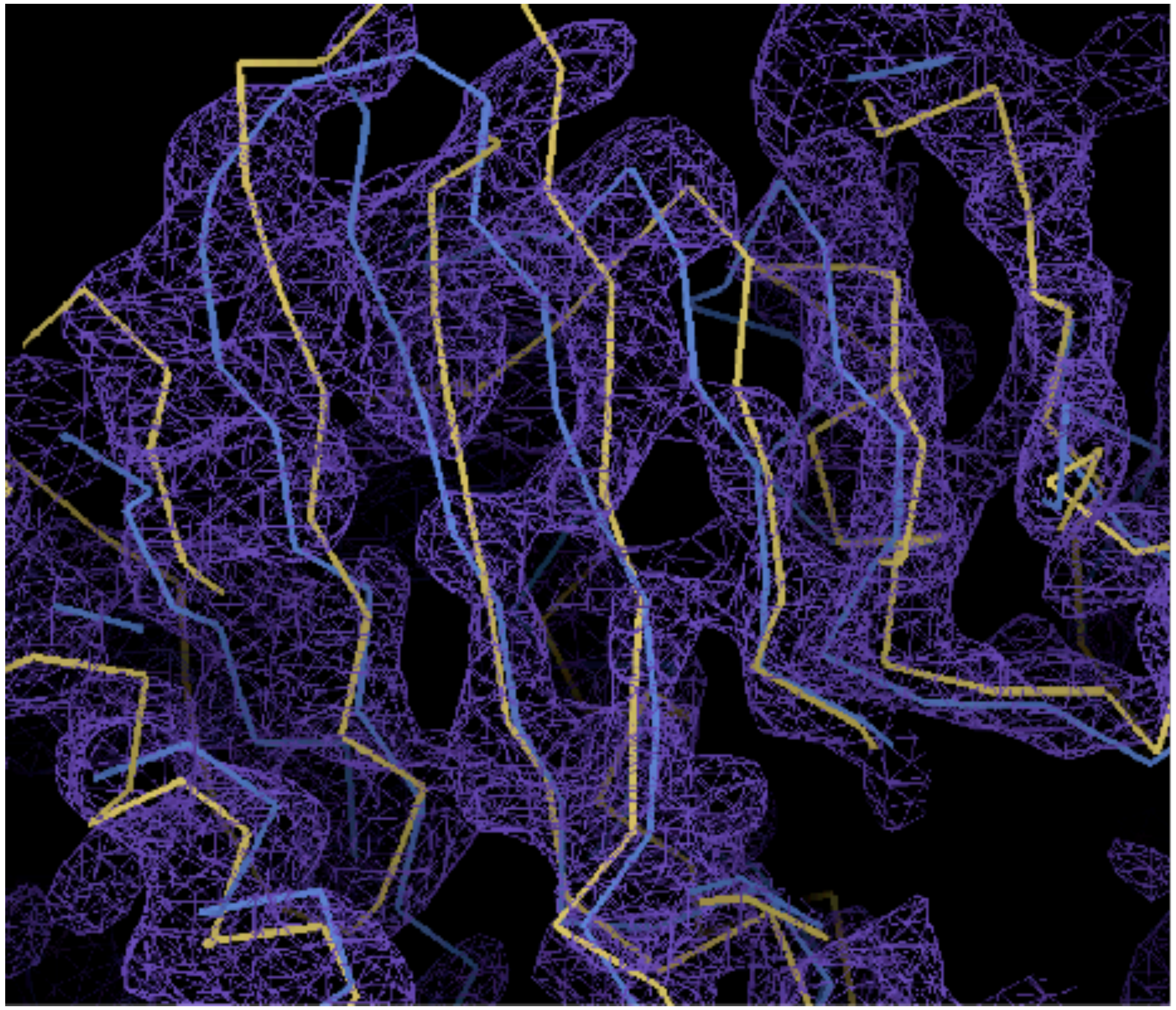
100 models from Rosetta

Tom Terwilliger, Los Alamos National Laboratory

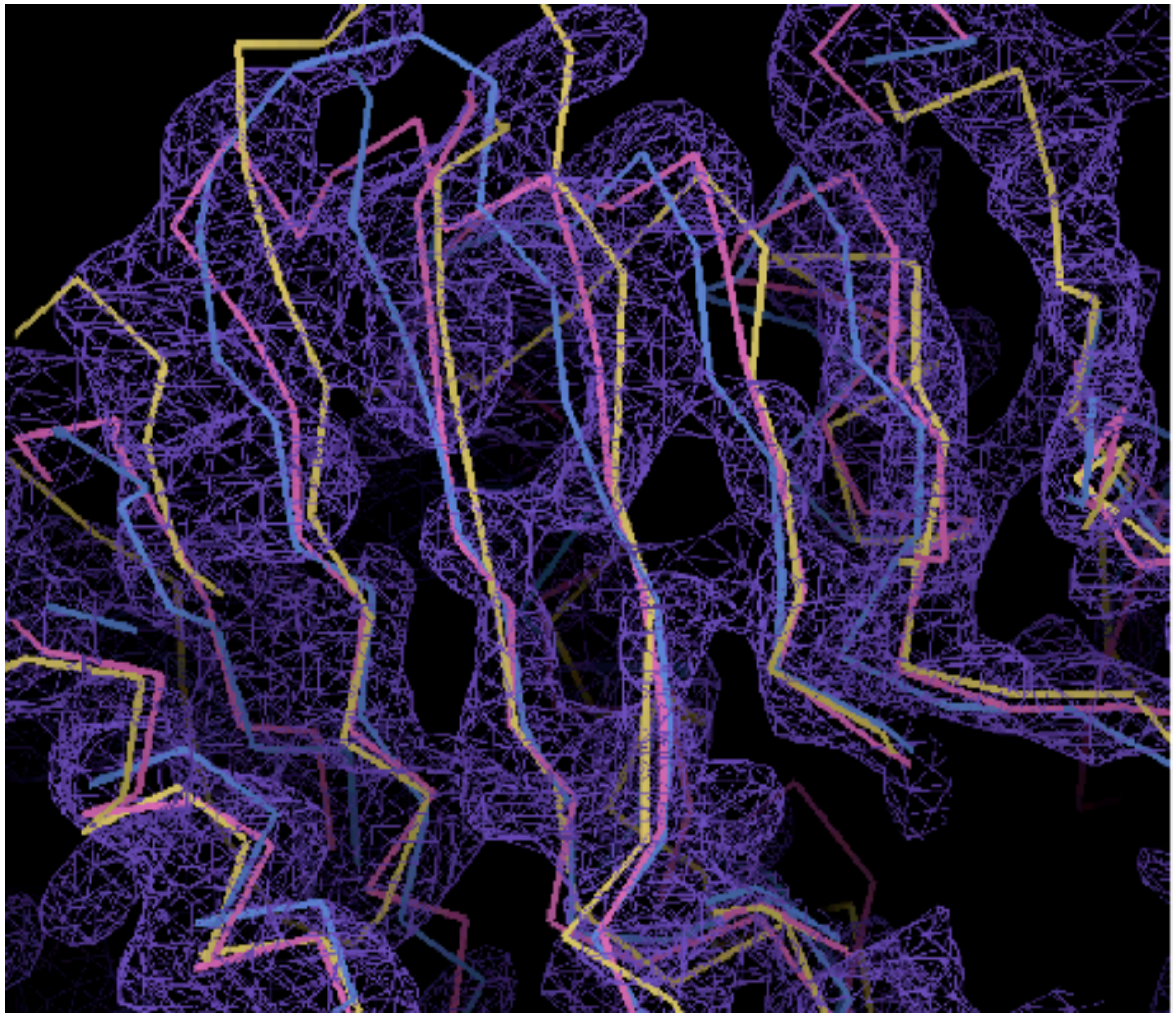
MR_Rosetta Procedure


Optional search model improvement with Rosetta
Molecular Replacement
Refinement & Density Modification
Rosetta plus Density Map
Score Models & Average Density from Top Models
Rosetta plus Density Map
Autobuild best Models

Rosetta Moves Models Closer to the True Structure



hp3342; MR solution, 22% identity (blue)
Final model (yellow)
Density modified map, 3.2Å (purple)



hp3342; MR solution, 22% identity (blue)
Final model (yellow)
Density modified map, 3.2Å (purple)
Best Rosetta model (magenta)

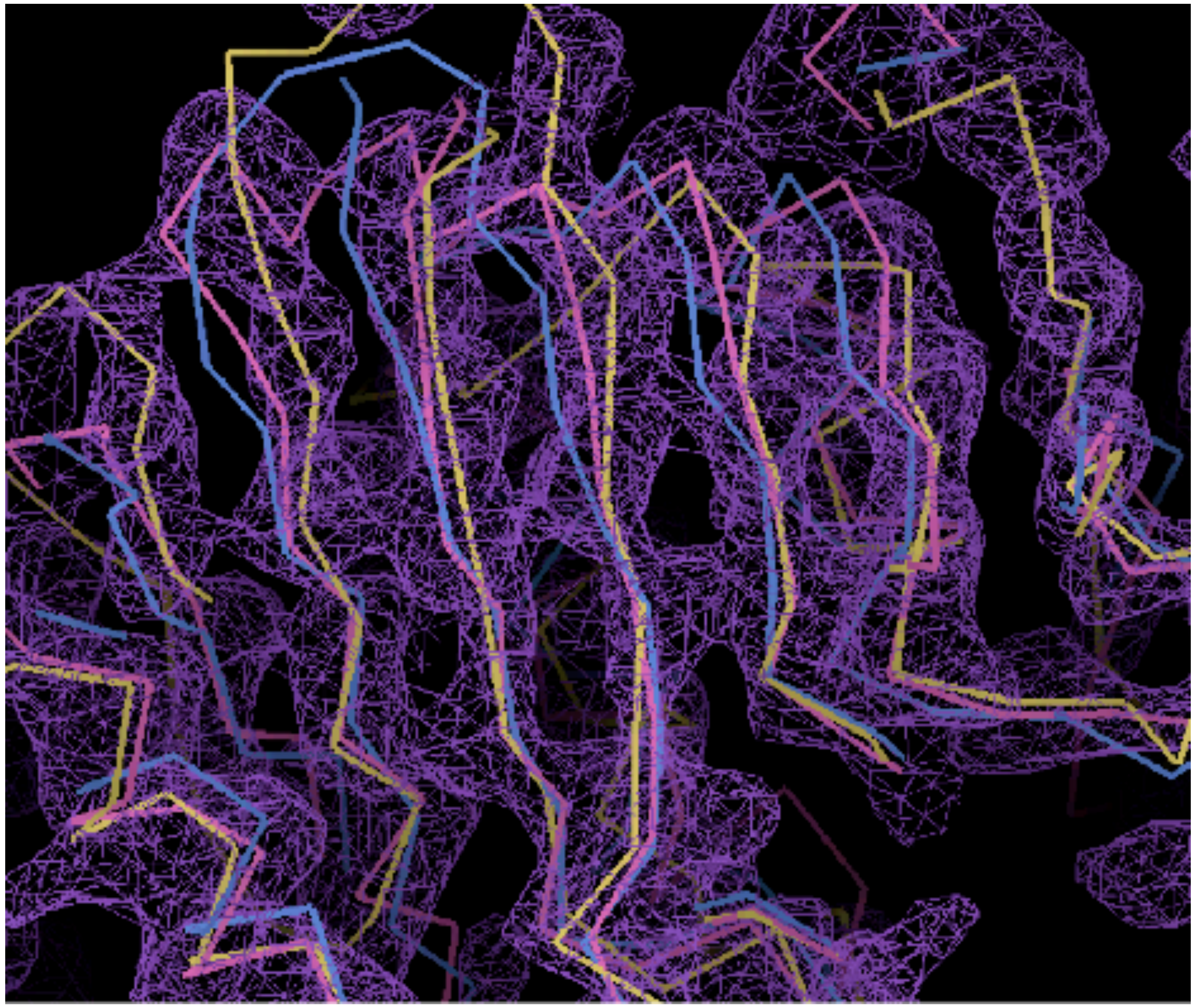
Tom Terwilliger, Los Alamos National Laboratory

BERKELEY LAB

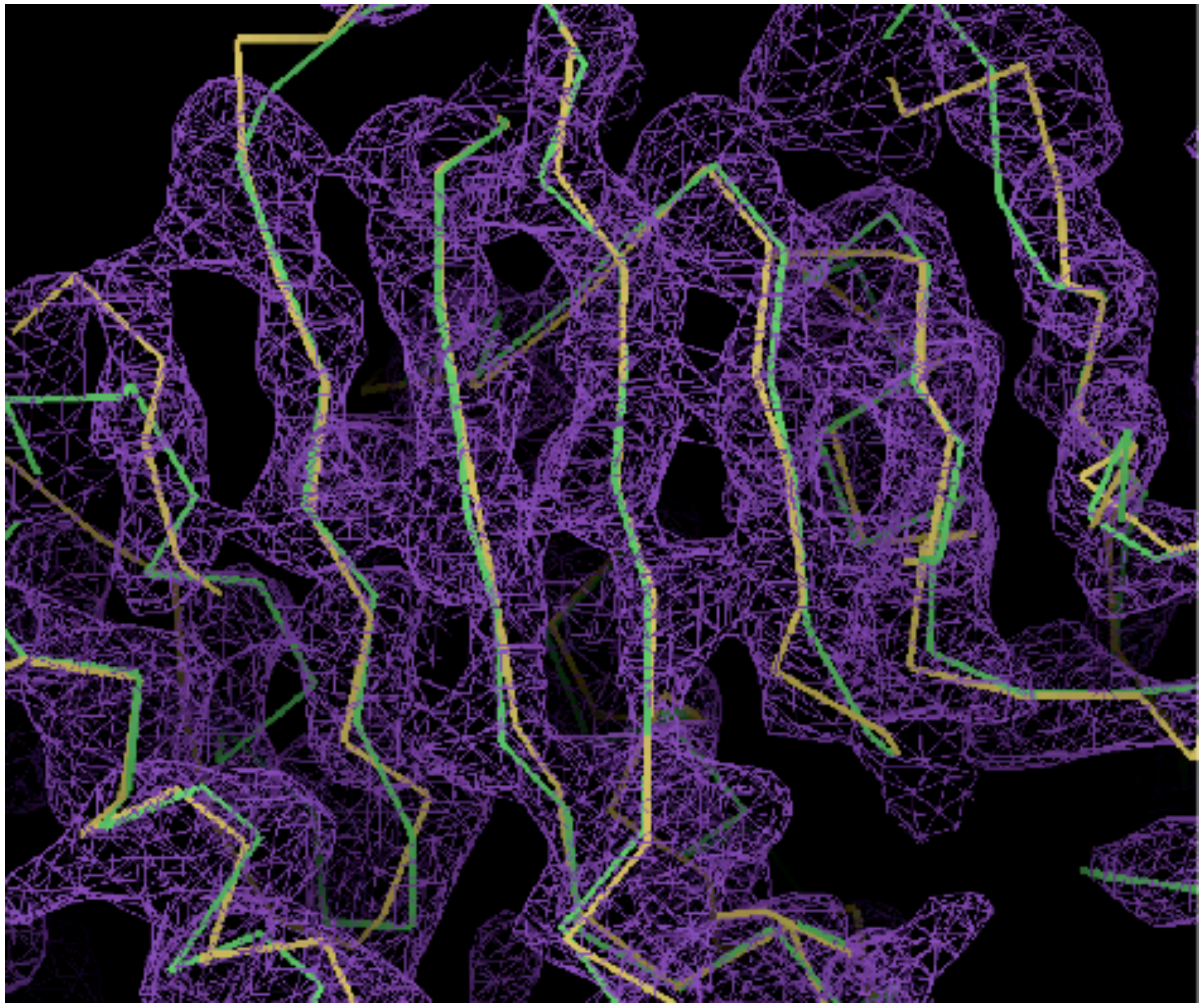
Phenix

Phases are Improved

Tom Terwilliger, Los Alamos National Laboratory

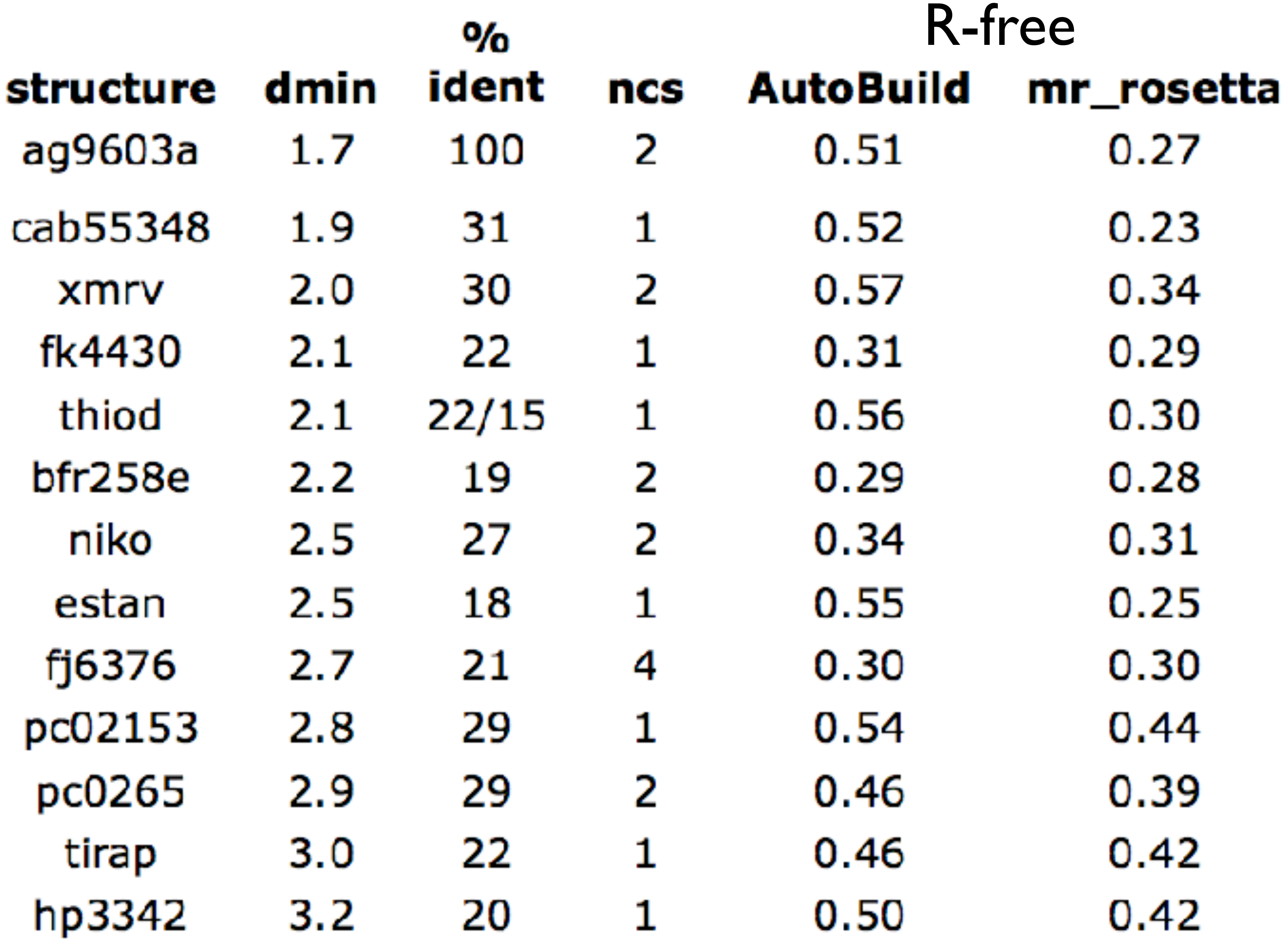


hp3342; MR solution, 22% identity (blue)
Final model (yellow)
Density modified map based on Rosetta model, 3.2Å (purple)
Best Rosetta model (magenta)



hp3342; MR solution, 22% identity (blue)
Final model (yellow)
Density modified map based on Rosetta model, 3.2Å (purple)
Autobuilt model (green)

Can be Applied at Low and High Resolution



				R-free	
structure	dmin	% ident	ncs	AutoBuild	mr_rosetta
ag9603a	1.7	100	2	0.51	0.27
cab55348	1.9	31	1	0.52	0.23
xmrv	2.0	30	2	0.57	0.34
fk4430	2.1	22	1	0.31	0.29
thiod	2.1	22/15	1	0.56	0.30
bfr258e	2.2	19	2	0.29	0.28
niko	2.5	27	2	0.34	0.31
estan	2.5	18	1	0.55	0.25
fj6376	2.7	21	4	0.30	0.30
pc02153	2.8	29	1	0.54	0.44
pc0265	2.9	29	2	0.46	0.39
tirap	3.0	22	1	0.46	0.42
hp3342	3.2	20	1	0.50	0.42

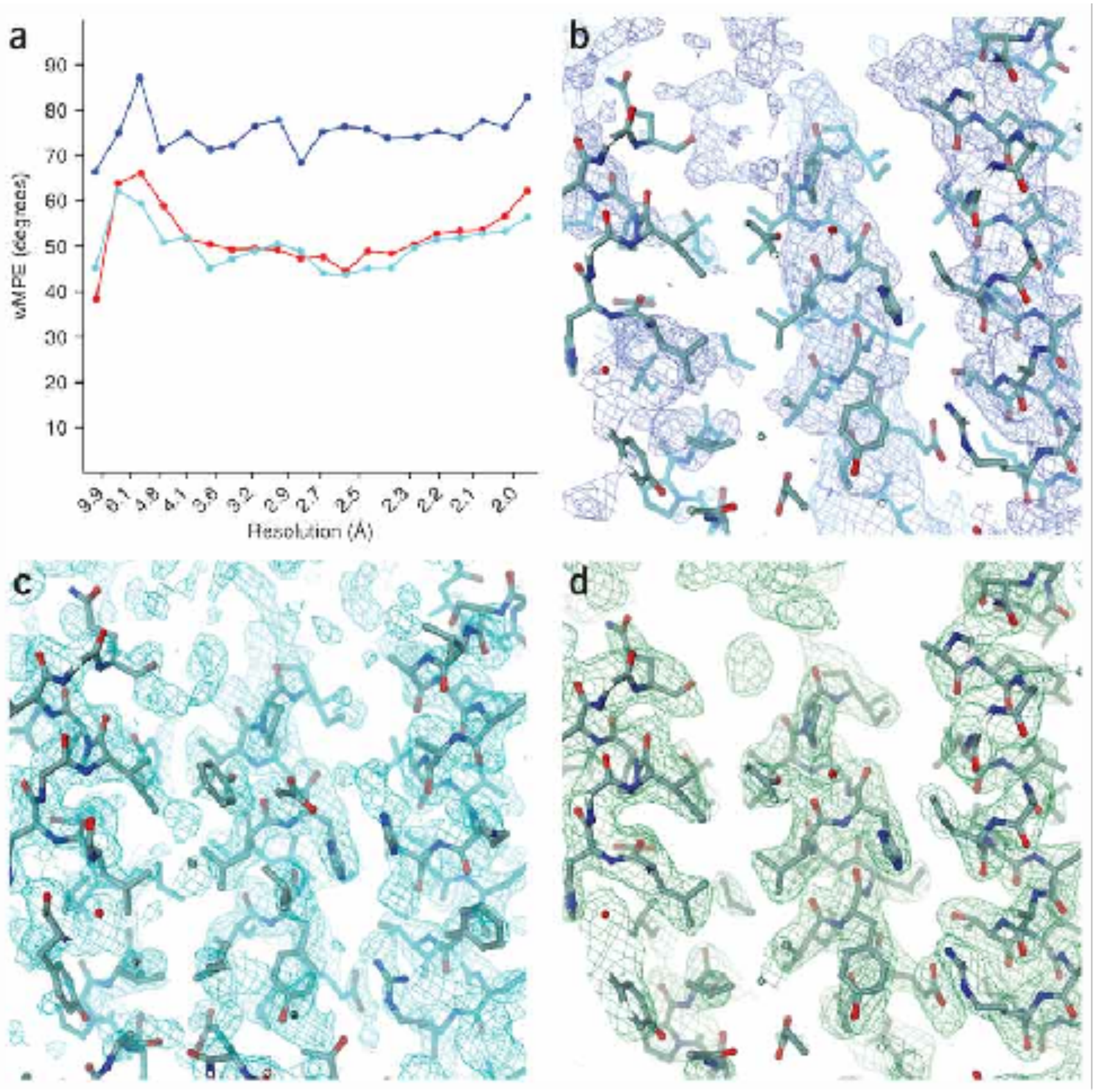
DiMaio F et al: Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature*. 2011 **473**:540-3

Extending Molecular Replacement

- In some cases the sequence identity can be so low as to suggest there is no structure of similar structure known
- What are the prospects for solving such molecular replacement cases?

Ab Initio Structure Solution

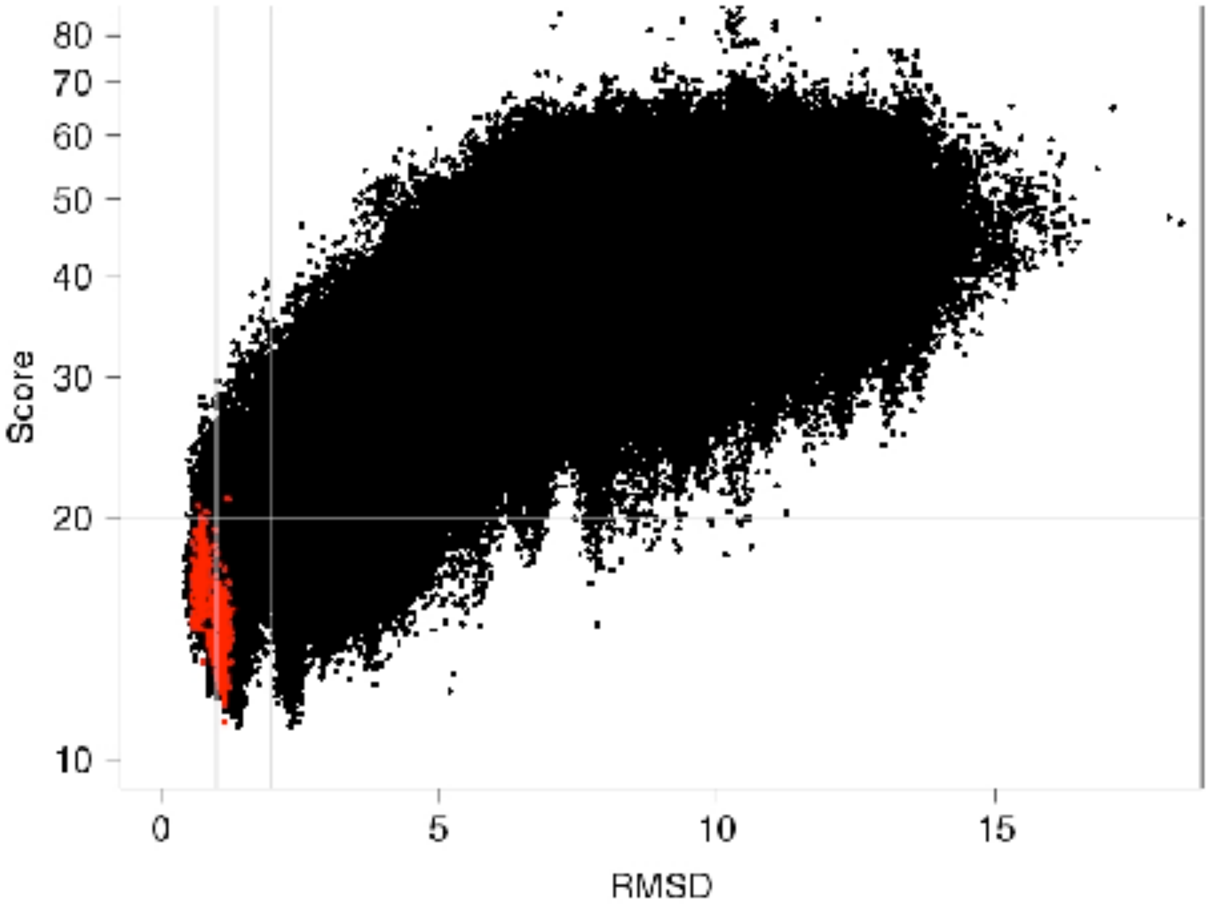
- *Arcimboldo:* Combining molecular replacement with small fragments, data extension, and automated rebuilding
- Dimer of 5-helix bundles (2x111 residues)
- Place 14-residue helices with Phaser
  - 1,473 potential 3-helix solutions (12% of atoms)
- Subject each solution to DM and autotracing with SHELXE
  - First at 1.95Å, then extend to 1.7Å with the "free lunch" algorithm
  - 3 of 1,473 gave an interpretable map



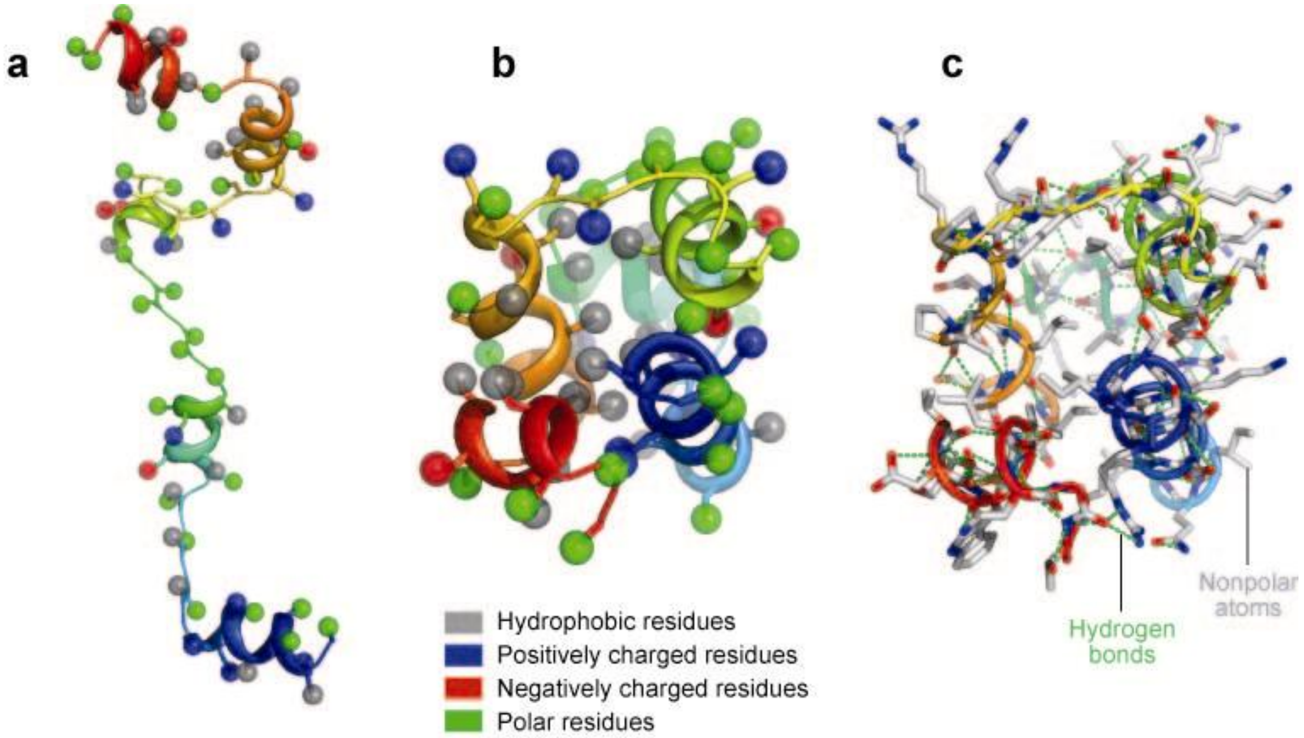
Rodríguez, Grosse, Himmel, González, de Ilarduya, Becker, Sheldrick & Usón, "Crystallographic ab initio protein structure solution below atomic resolution", Nature Methods 6: 651-653, 2009.

Rosetta

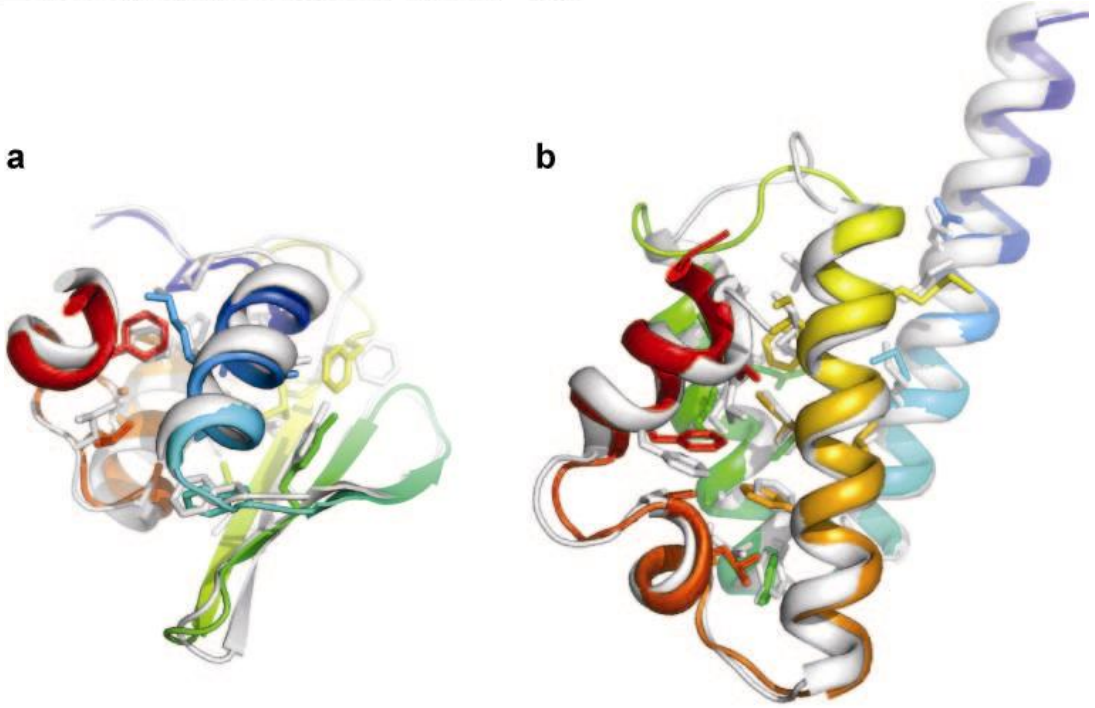
- *ab initio* model generation and model optimization
- Requires extensive computational sampling



Black - Rosetta *ab initio* models, Red - Crystal structure after Relax protocol



AR Das R, Baker D. 2008.
Annu. Rev. Biochem. 77:363–82.

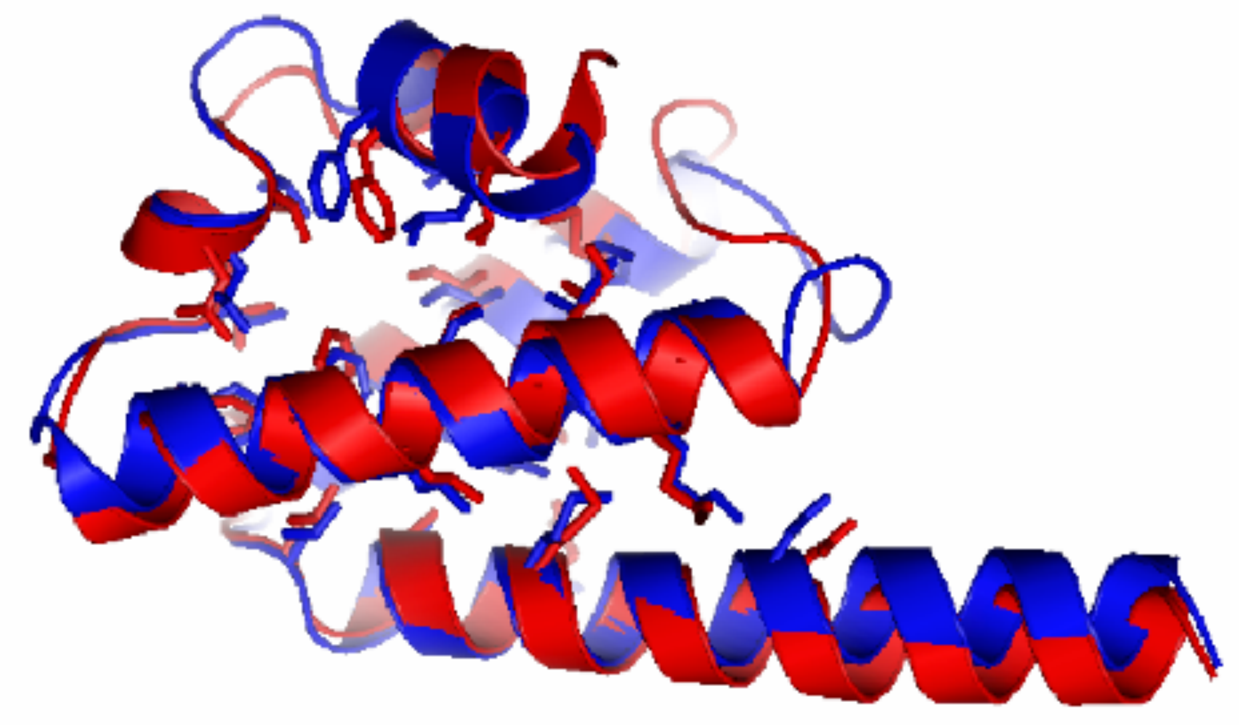


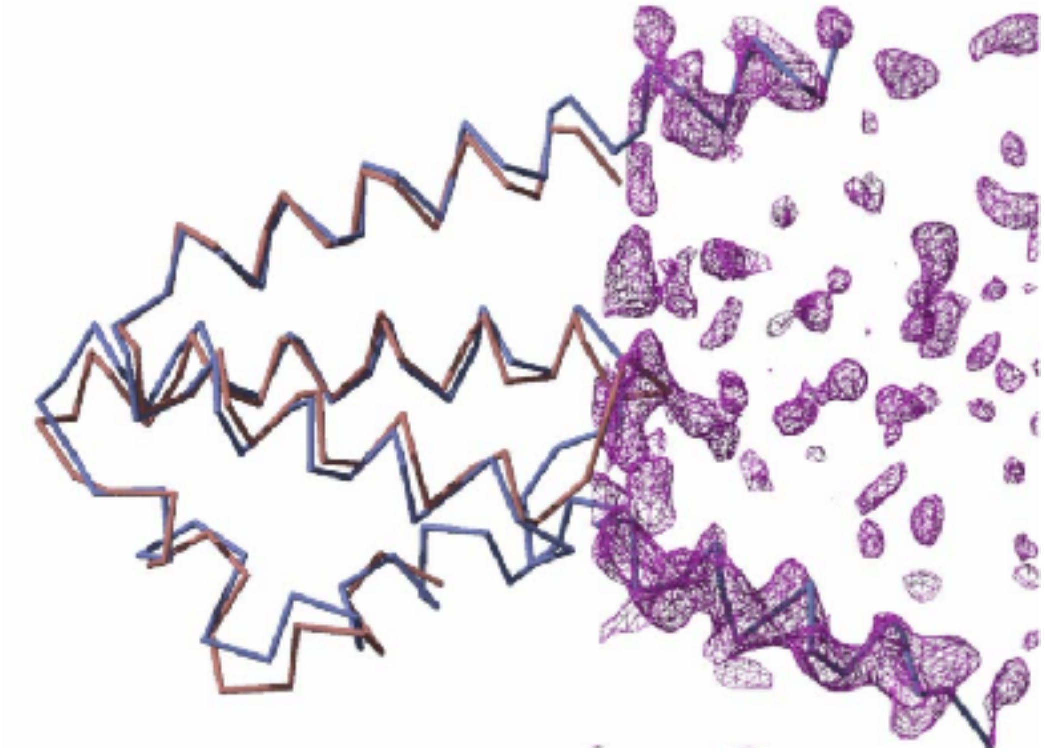
AR Das R, Baker D. 2008.
Annu. Rev. Biochem. 77:363–82.

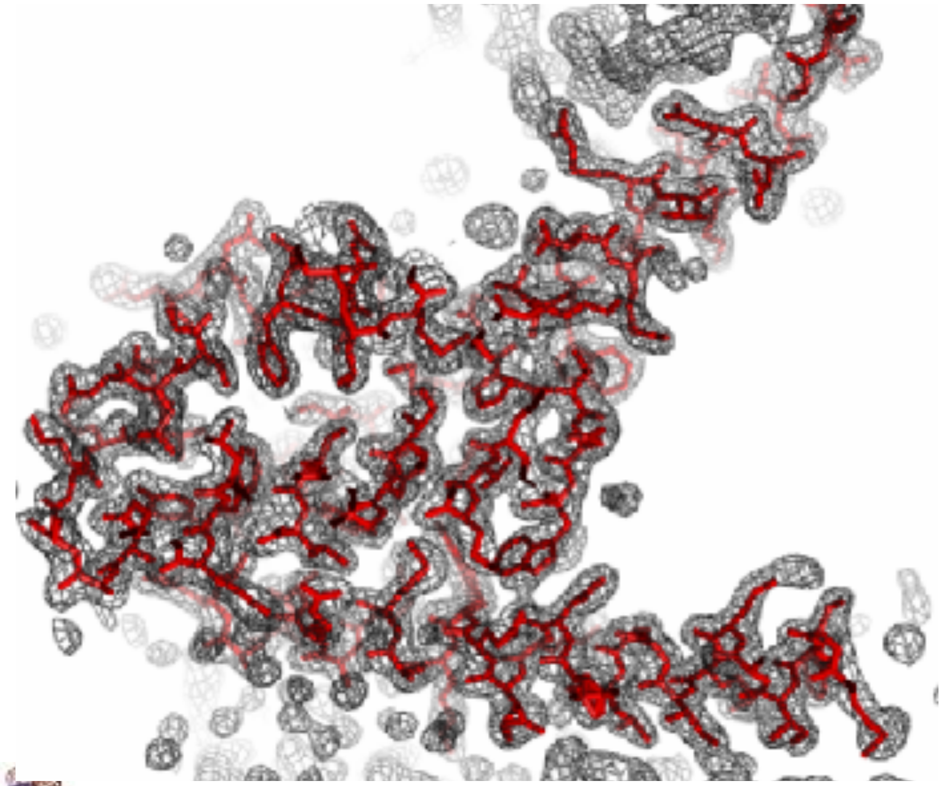
Phenix

Ab Initio Structure Solution

- Rosetta (Baker group) is a method for *ab initio* protein structure prediction
- Models were used in MR to solve a novel structure (no close enough models were available in the PBD)
- Automated model building methods complete the structure







Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D Nature. 2007 Nov 8;450(7167):259-64.

Phenix

Summary

- New algorithms increase the success rate of molecular replacement
- Suggested approach:
  - Apply standard methods
    - Anisotropy & tNCS addressed automatically in Phaser
  - Analyze indicators of success (Z-scores), packing, R-factors
    - Also check the PDB for cell dimensions and space group (did you use lysozyme to lyse your cells?)
  - If not obvious, try extensive refinement (100 cycles)
  - If still unclear, try morphing
  - If still not OK, try MR_Rosetta
  - Desperate? - try Rosetta or similar tools for *ab initio* model generation (limits on the size of molecule)
- Include experimental phase information if you have it

Acknowledgments

- **Lawrence Berkeley Laboratory**
  - Pavel Afonine, Youval Dar, Nat Echols, Jeff Headd, Richard Gildea, Ralf Grosse-Kunstleve, Dorothee Liebschner, Nigel Moriarty, Nader Morshed, Billy Poon, Ian Rees, Nicholas Sauter, Oleg Sobolev, Peter Zwart
- **Los Alamos National Laboratory**
  - Tom Terwilliger, Li-Wei Hung
- **Cambridge University**
  - Randy Read, Airlie McCoy, Laurent Storoni, Gabor Bunkoczi, Robert Oeffner
- **Duke University**
  - Jane Richardson & David Richardson, Ian Davis, Vincent Chen, Jeff Headd, Christopher Williams, Bryan Arendall, Laura Murray, Gary Kapral, Dan Keedy, Swati Jain, Bradley Hintze, Lindsay Deis, Lizbeth Videau
- **University of Washington**
  - Frank DiMaio, David Baker
- **Oak Ridge National Laboratory**
  - Marat Mustyakimov, Paul Langan
- **Others**
  - Alexandre Urzhumtsev & Vladimir Lunin
  - Garib Murshudov & Alexi Vagin
  - Kevin Cowtan, Paul Emsley, Bernhard Lohkamp
  - David Abrahams
  - PHENIX Testers & Users: James Fraser, Herb Klei, Warren Delano, William Scott, Joel Bard, Bob Nolte, Frank von Delft, Scott Classen, Ben Eisenbraun, Phil Evans, Felix Frolow, Christine Gee, Miguel Ortiz-Lombardia, Blaine Mooers, Daniil Prigozhin, Miles Pufall, Edward Snell, Eugene Valkov, Erik Vogan, Andre White, and many more
- **Funding:**
  - NIH/NIGMS:
    - *P01GM063210, P50GM062412, P01GM064692, R01GM071939*
  - Lawrence Berkeley Laboratory
  - PHENIX Industrial Consortium

BERKELEY LAB

Phenix