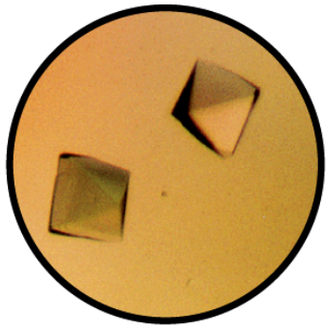
Model Building in Phenix

Macromolecular Crystallography School
Madrid, May 2017

Paul Adams

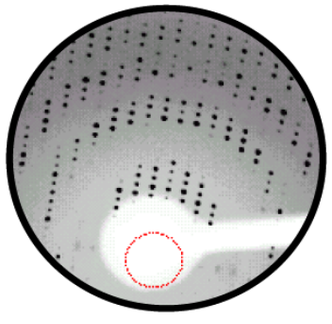
Lawrence Berkeley Laboratory and
Department of Bioengineering UC Berkeley

The Crystallographic Process

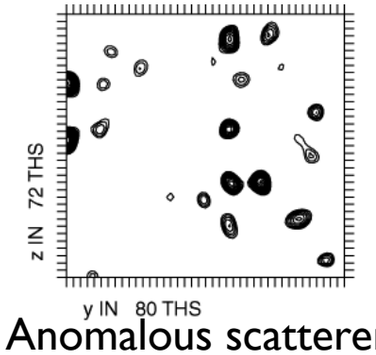


Crystallization

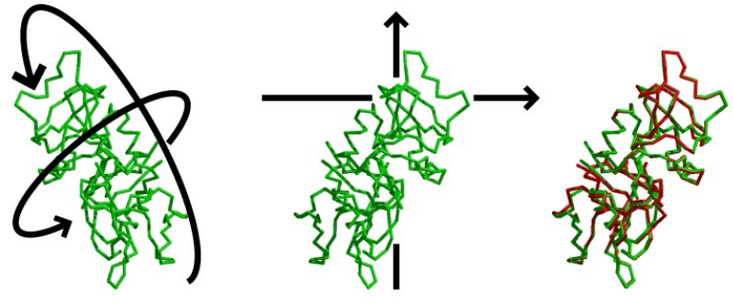
Data collection



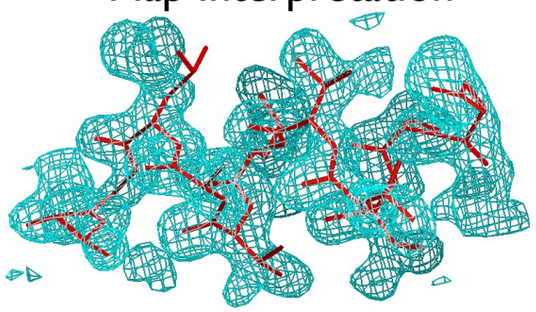
Data processing



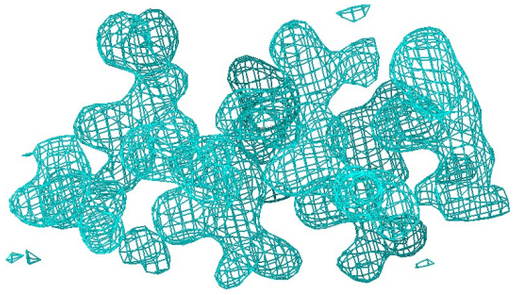
Anomalous scatterer location



Molecular replacement

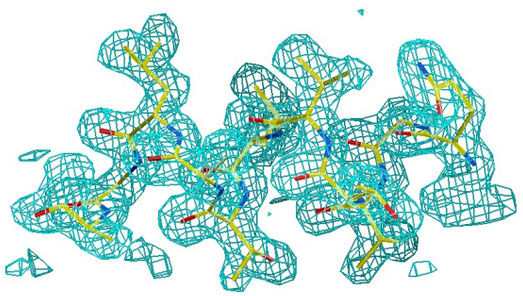


Map Interpretation

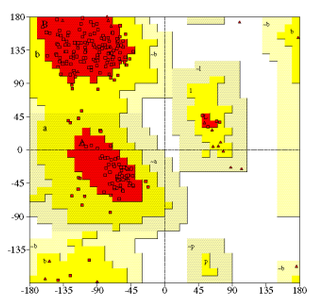


Phase improvement

Phase determination



Model refinement



Validation

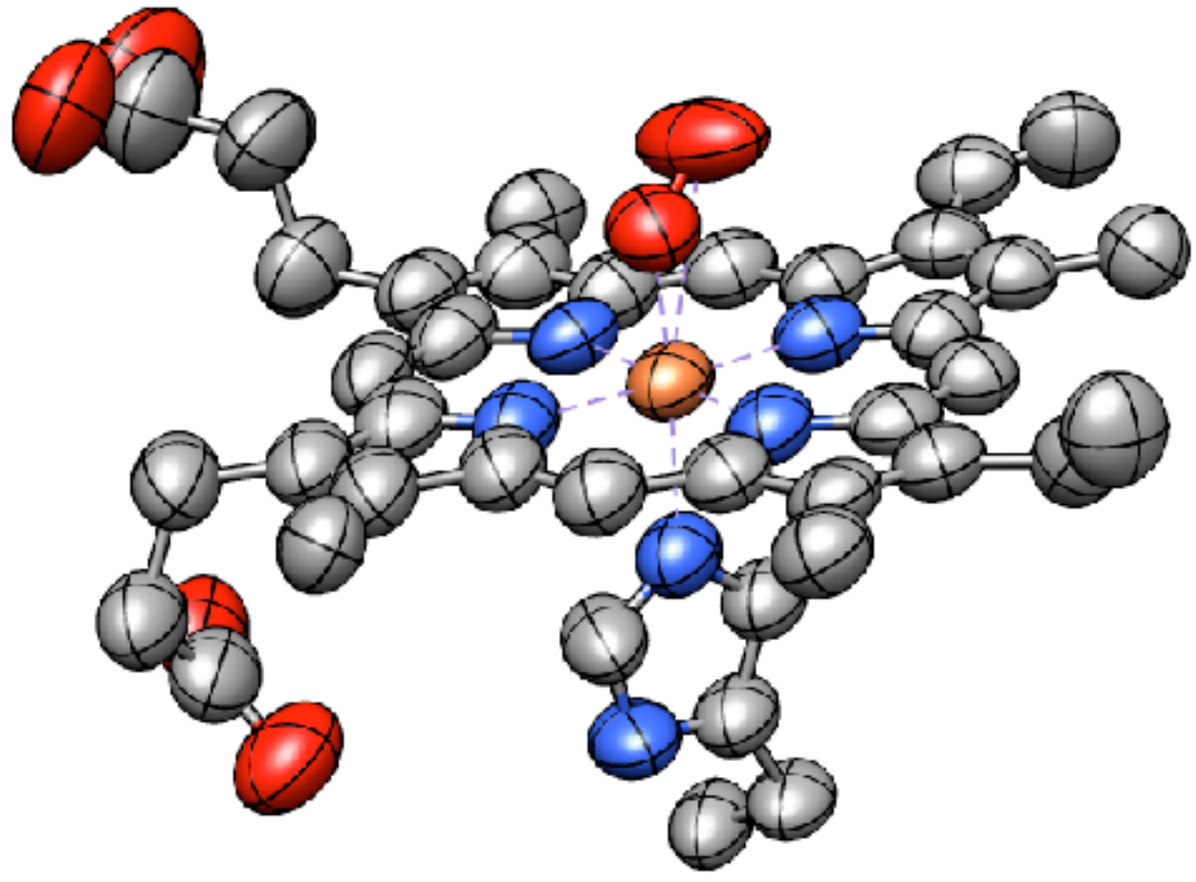
BERKELEY LAB

Phenix

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868

The Crystallographic Model

- Atoms (spherical or ellipsoid)
- Mean square displacements (B-factors)
- Occupancy
- Chemical restraints (e.g. bond lengths, angles, etc)



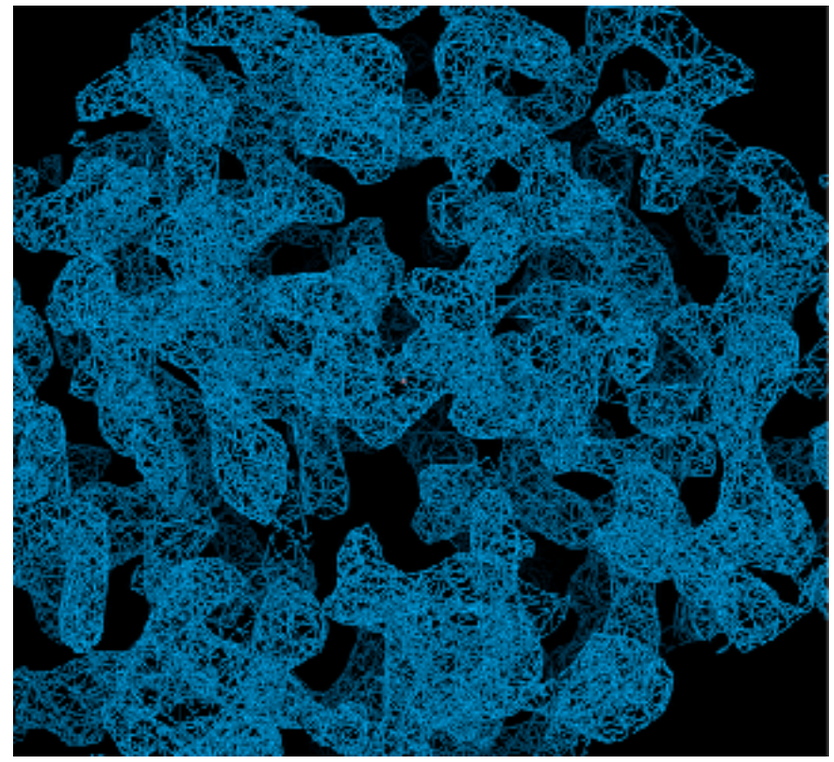
BERKELEY LAB

Phenix

Electron Density Maps

- Real and reciprocal spaces are related by Fourier Transformation

$|F|e^{i\Phi_{obs}}$ FT



- Experimental phasing

$$m_{obs}F_{obs}e^{i\phi_{obs}}$$

$$m_{dm}F_{obs_{dm}}e^{i\phi_{dm}}$$ [Any missing Fobs generated from density modification]

- Molecular Replacement

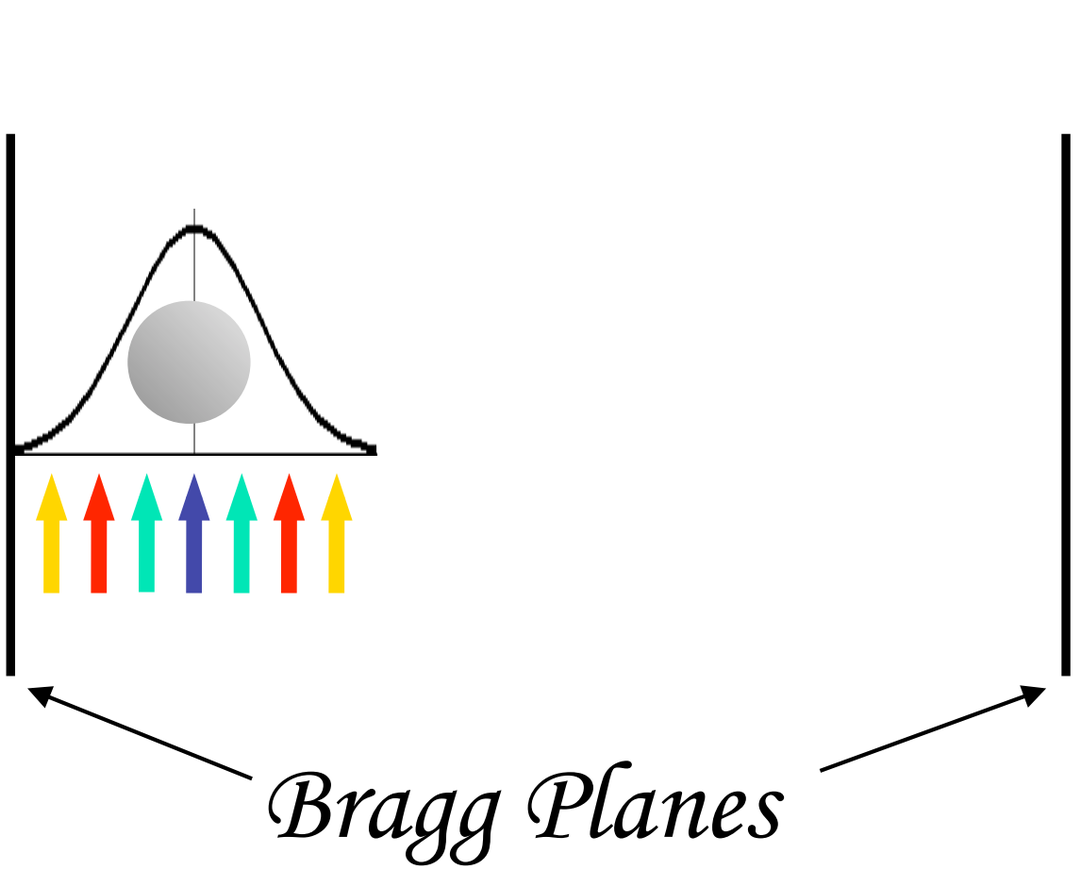
$$m_{dm}F_{obs_{dm}}e^{i\phi_{dm}}$$ [Any missing Fobs generated from density modification]

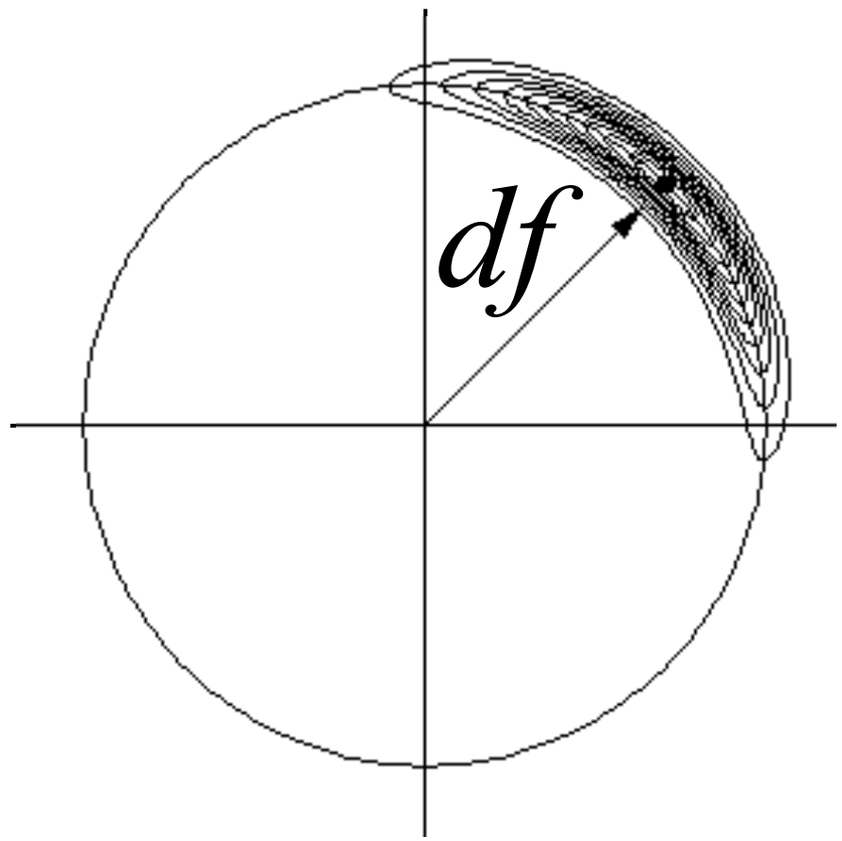
BERKELEY LAB

Phenix

Effect of Errors in Atomic Position

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



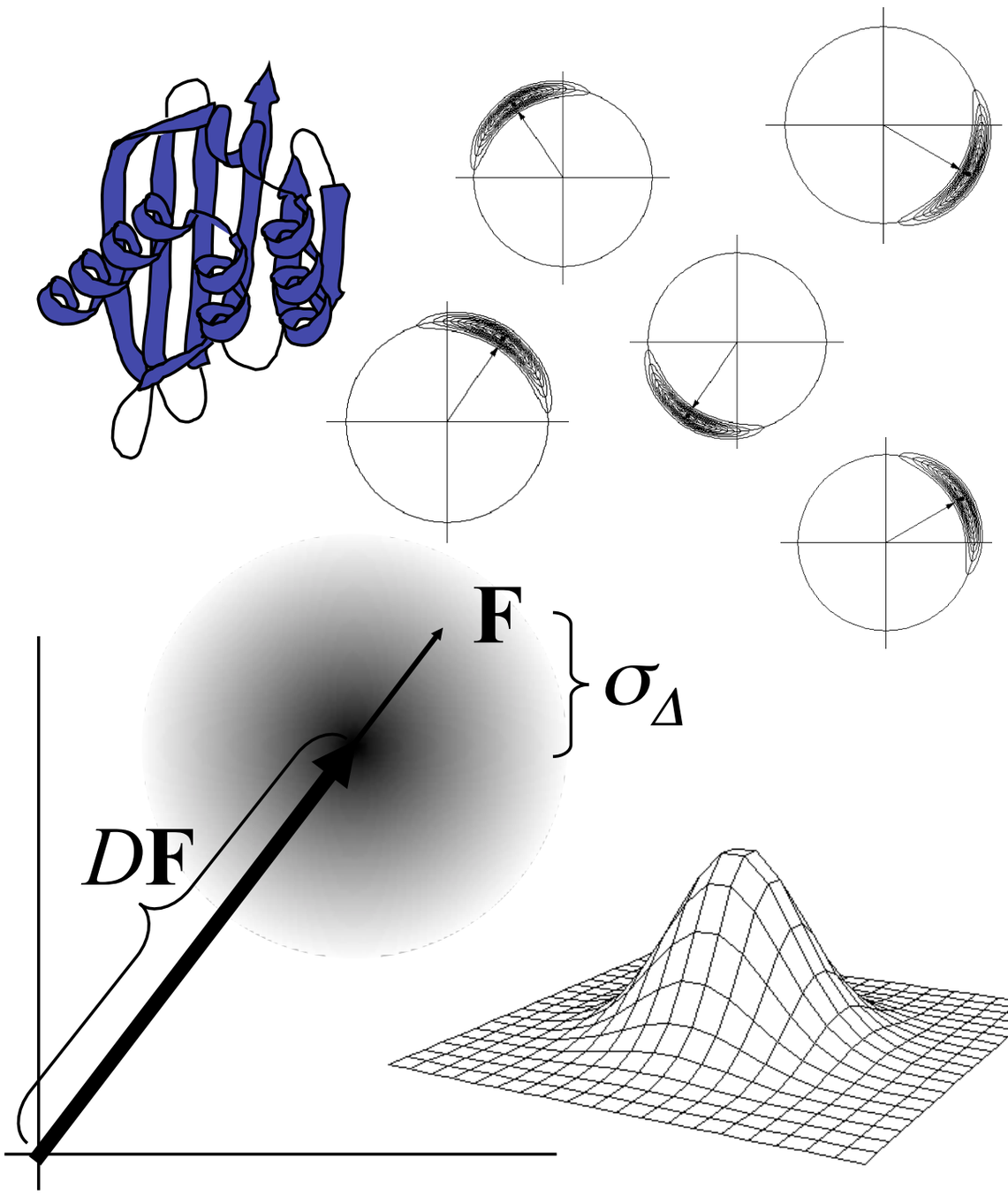


BERKELEY LAB

Phenix

Structure factor with coordinate errors

- Same direction as the sum of the atomic *f*
  - but shorter by 0< *D* <1
  - *D*=f(resolution)



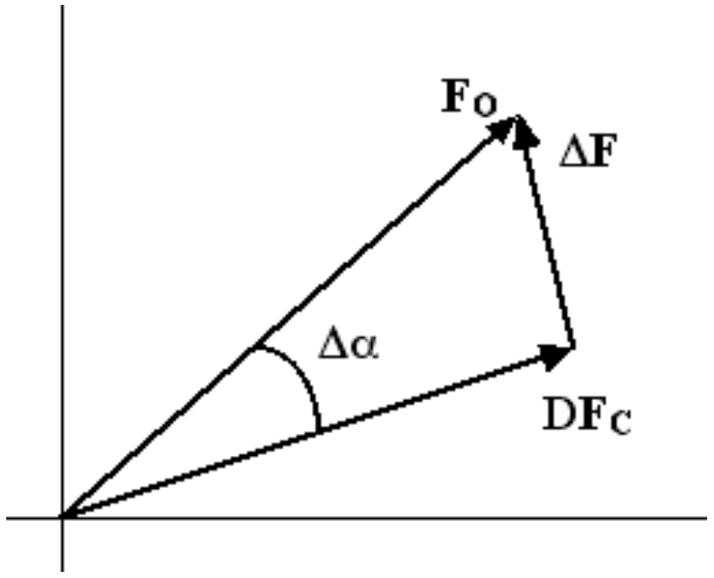
- Central Limit Theorem
  - Many small atoms
  - Gaussian distribution for the total summed F
  - σ_Δ=f(resolution)

BERKELEY LAB

Phenix

Estimated Error in Maps - σ_A

- When an atomic model is available, estimates of errors arising from the model can be made.



$$|\Delta \mathbf{F}|^2 = |\mathbf{F}_O|^2 + D^2|\mathbf{F}_C|^2 - 2D|\mathbf{F}_O||\mathbf{F}_C|\cos(\Delta\alpha)$$

$$\mathbf{F}_O \approx \left(2m|\mathbf{F}_O| - D|\mathbf{F}_C|\right)\exp(i\alpha_C)$$

$\left(2m|\mathbf{F}_O| - D|\mathbf{F}_C|\right)\exp(i\alpha_C)$ Model phased map + Difference map

$\left(m|\mathbf{F}_O| - D|\mathbf{F}_C|\right)\exp(i\alpha_C)$ Difference map

$$m_{comb} 2F_{obs} e^{i\phi_{comb}} - D_{\sigma_A} F_{calc} e^{i\phi_{calc}}$$

$$m_{comb} F_{obs} e^{i\phi_{comb}} - D_{\sigma_A} F_{calc} e^{i\phi_{calc}}$$ [Difference map]

$$m_{\sigma_A} 2F_{obs} e^{i\phi_{calc}} - D_{\sigma_A} F_{calc} e^{i\phi_{calc}}$$

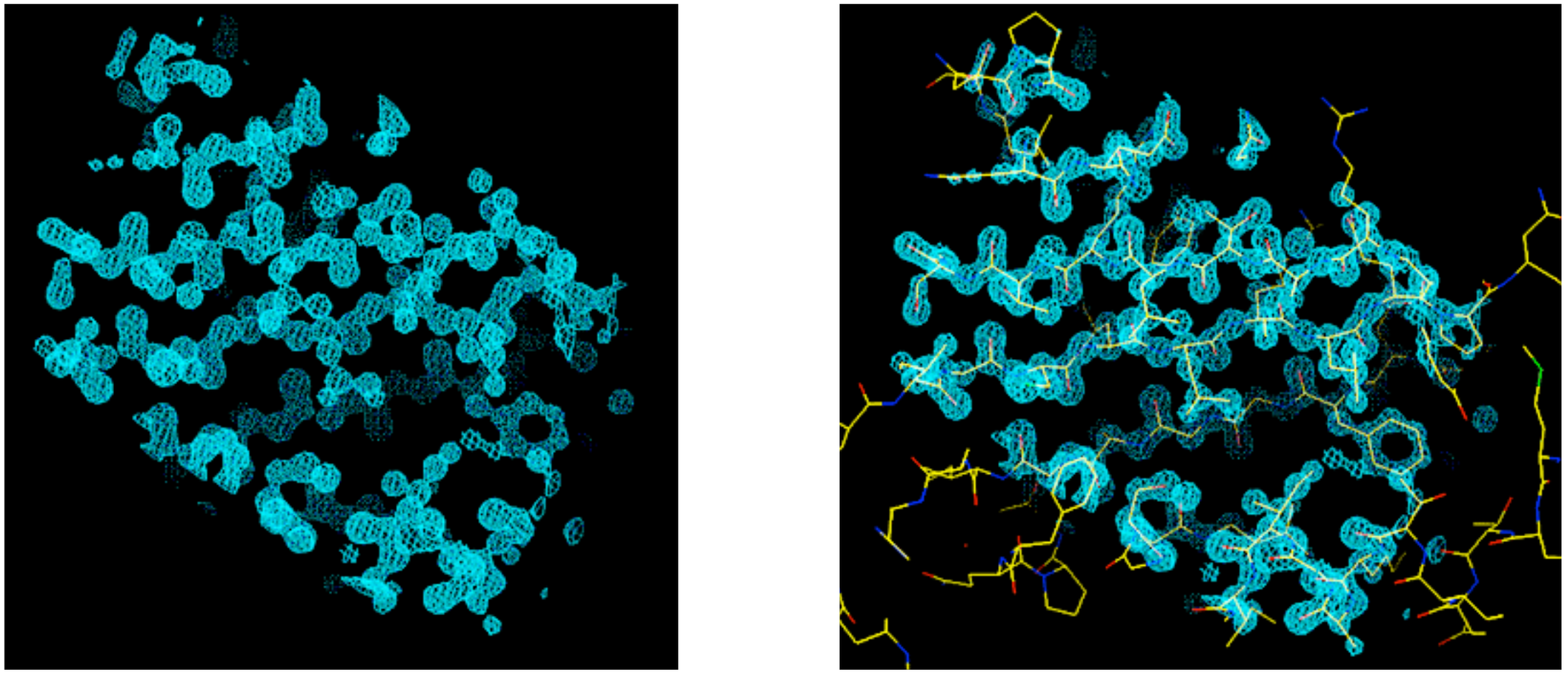
$$m_{\sigma_A} F_{obs} e^{i\phi_{calc}} - D_{\sigma_A} F_{calc} e^{i\phi_{calc}}$$ [Difference map]

BERKELEY LAB

Phenix

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868

High Resolution Maps



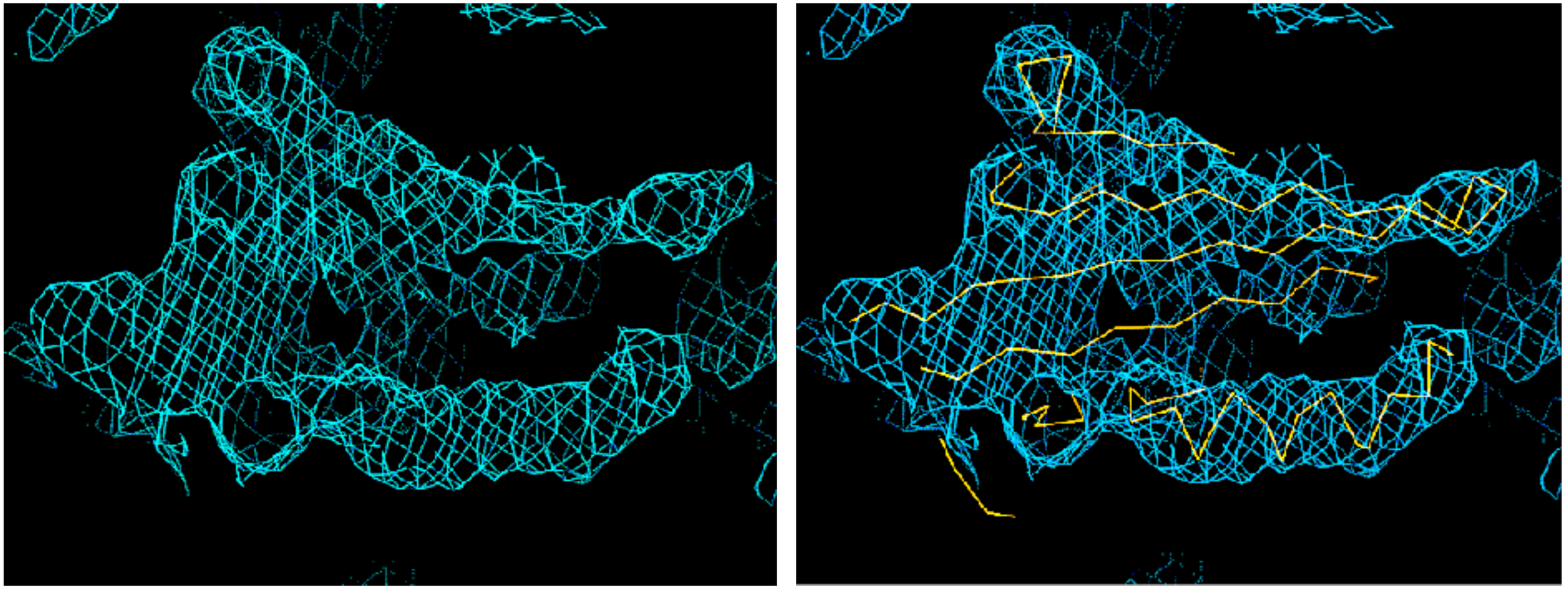
- High resolution maps (1.5Å or better) are typically easy to interpret, although time consuming)
- Biggest challenge is recognizing and modelling discrete disorder and atomic motion

Image from Phil Evans, LMB MRC Cambridge

BERKELEY LAB

Phenix

Low Resolution Maps



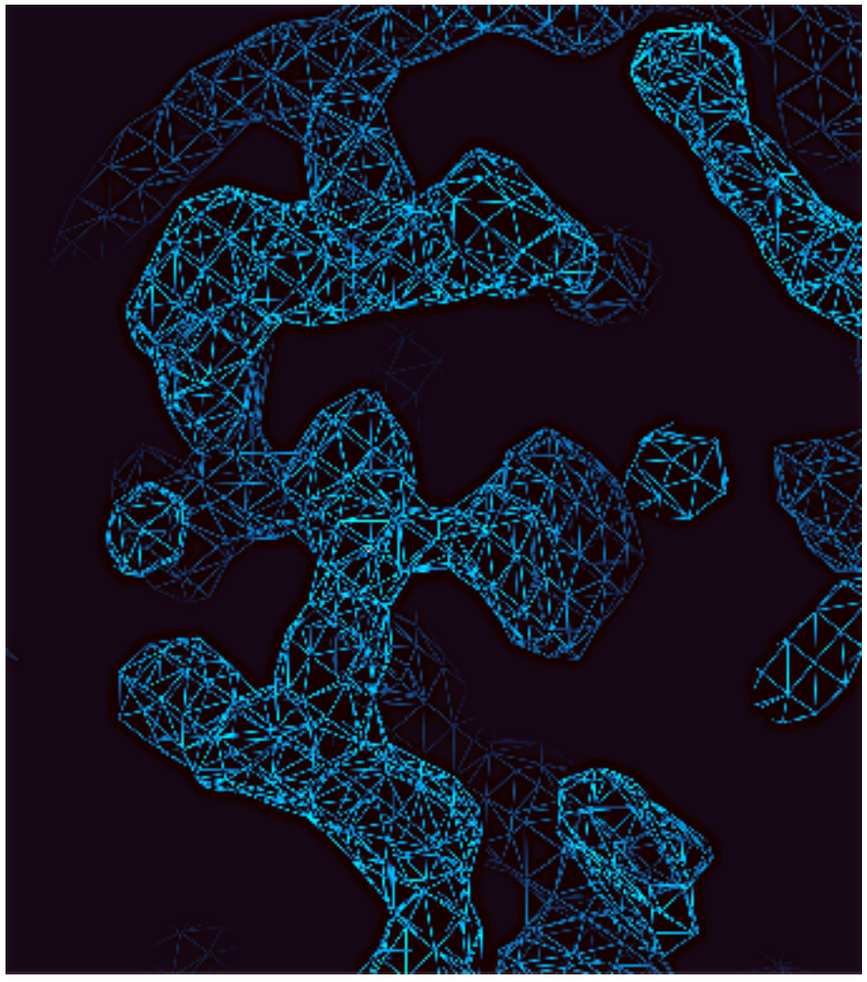
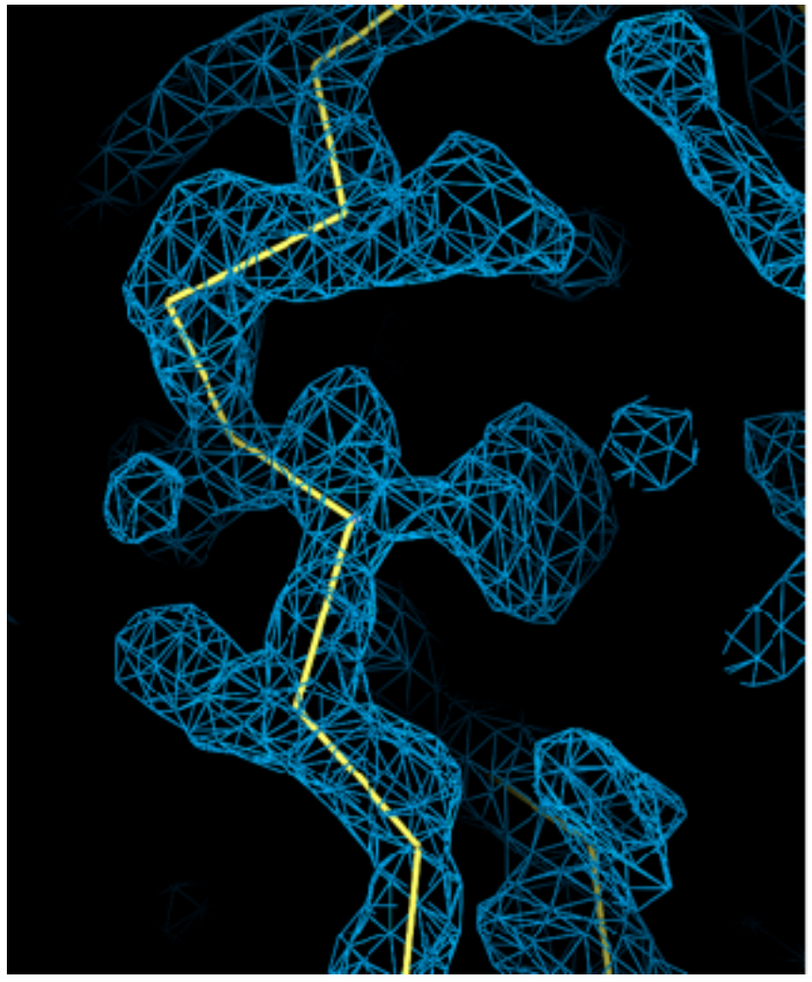
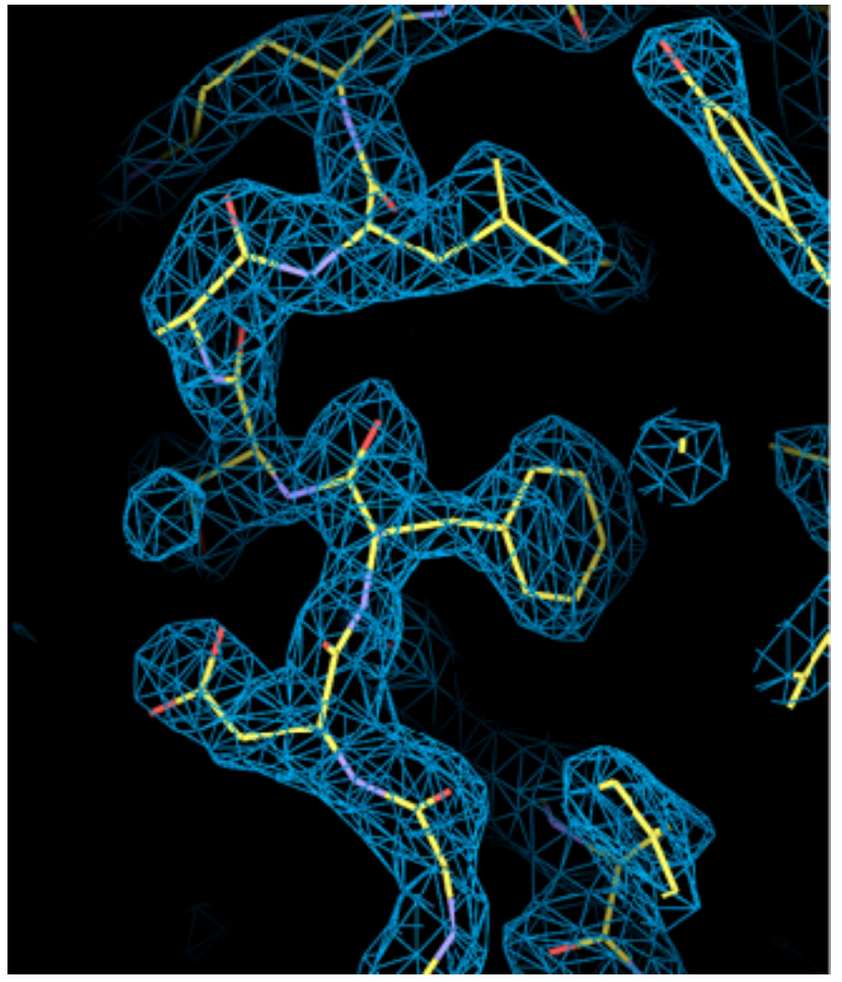
- Low resolution maps (3.5Å and worse) are typically very difficult to interpret.
- The lack of detail makes it difficult to determine the identity of residues.
- At very low resolution the use of similar structural motifs can greatly aide the process

Image from Phil Evans, LMB MRC Cambridge

BERKELEY LAB

Phenix

Divide-and-Conquer

- Manual model building typically requires that the map interpretation be divided up into different stages:
  - Tracing the polymer backbone then adding the chemical identities for the polymer units (e.g. amino acids)

BERKELEY LAB

Phenix

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868

Automated Model Building

- The process of map interpretation can also be performed computationally.
  - Is less time consuming for the user
  - Object decision-making can minimize errors
  - Automated methods typically rely on some kind of pattern matching algorithm to extract information from the map.

BERKELEY LAB

Phenix

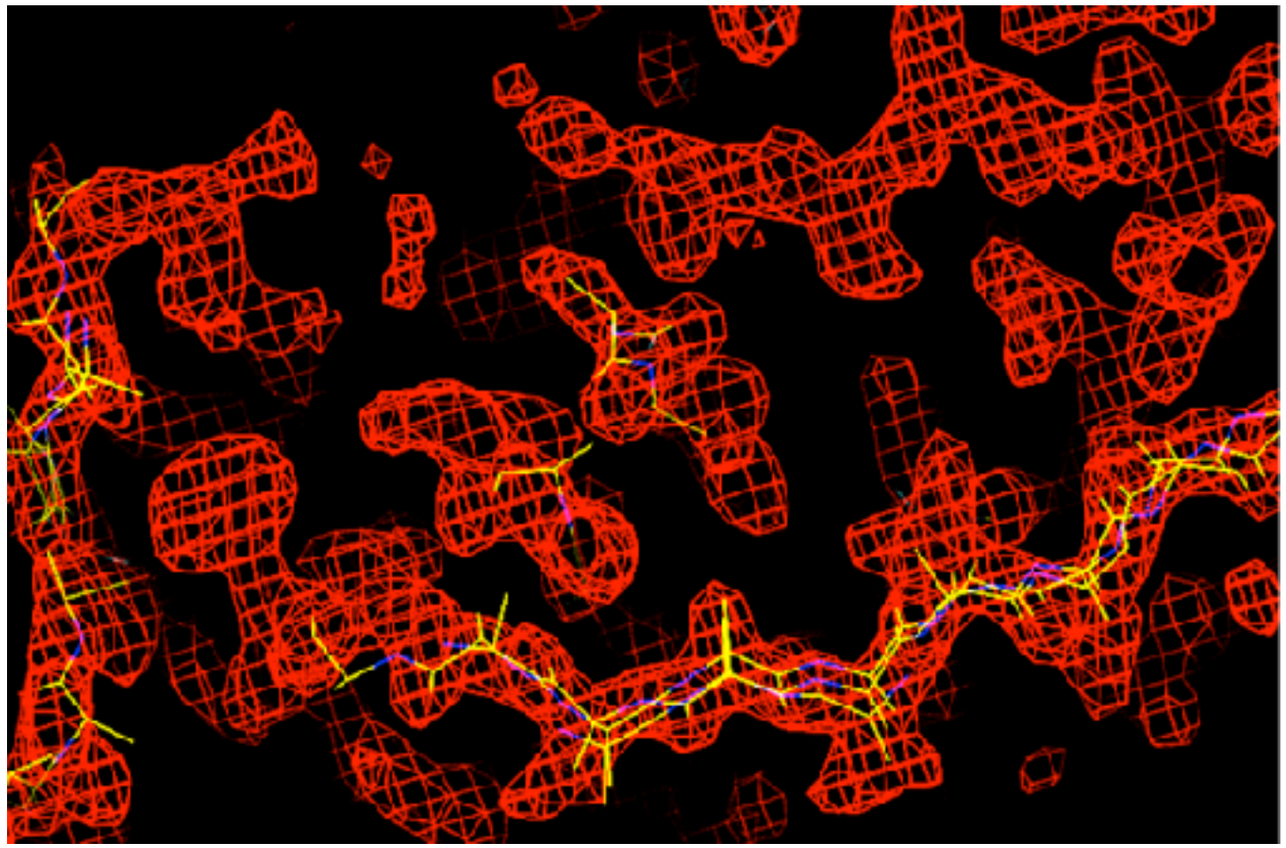
THE UNIVERSITY OF CALIFORNIA BERKELEY 1868

Map Interpretation with Larger Fragments

- RESOLVE uses pattern matching methods to automate the model building process:
  - FFT-based identification of helices and strands
  - Extension with tri-peptide libraries
  - Probabilistic sequence alignment
  - Automatic molecular assembly
- RESOLVE uses larger fragments than individual atoms so is able to perform well even at medium to low resolution

Phenix

Locating Fragments



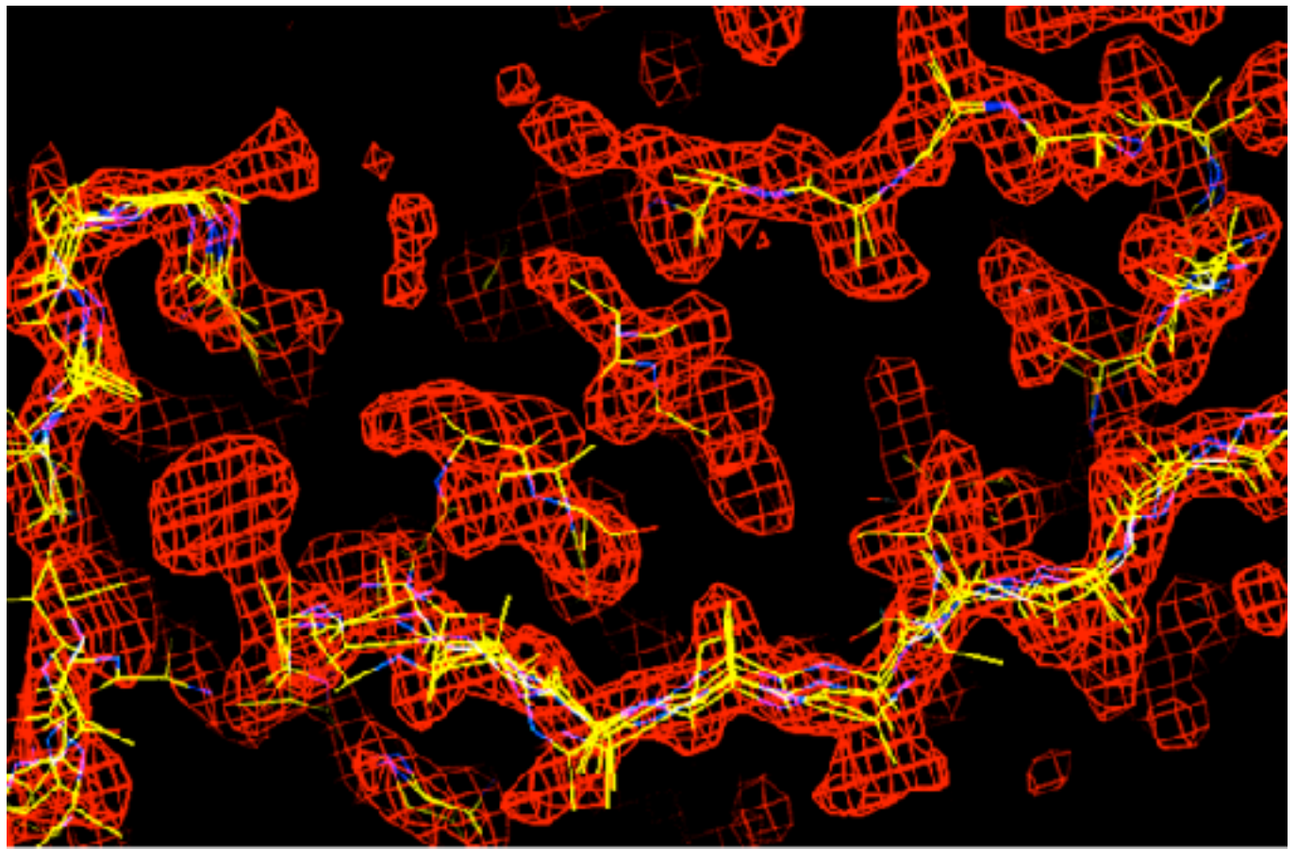
- Fragments:
  - Helical template: 6 amino acids, average density from ~200 6-amino acid helical segments
  - Helix fragment library: 53 helices 6-24 amino acid long
  - Beta-sheet template: 4 amino acid, average density
  - Beta-sheet fragment library: 24 strands 4-9 amino acid long
- Identify possible template locations with FFT-based convolution search
- Maximize correlation coefficient of template with map
- Superimpose each fragment in corresponding library (helix, sheet) on template
- Identify longest segment in good density, score = <density>*sqrt(Natoms)

BERKELEY LAB

Phenix

Image from T. Terwilliger, Los Alamos National Laboratory

Fragment Extension



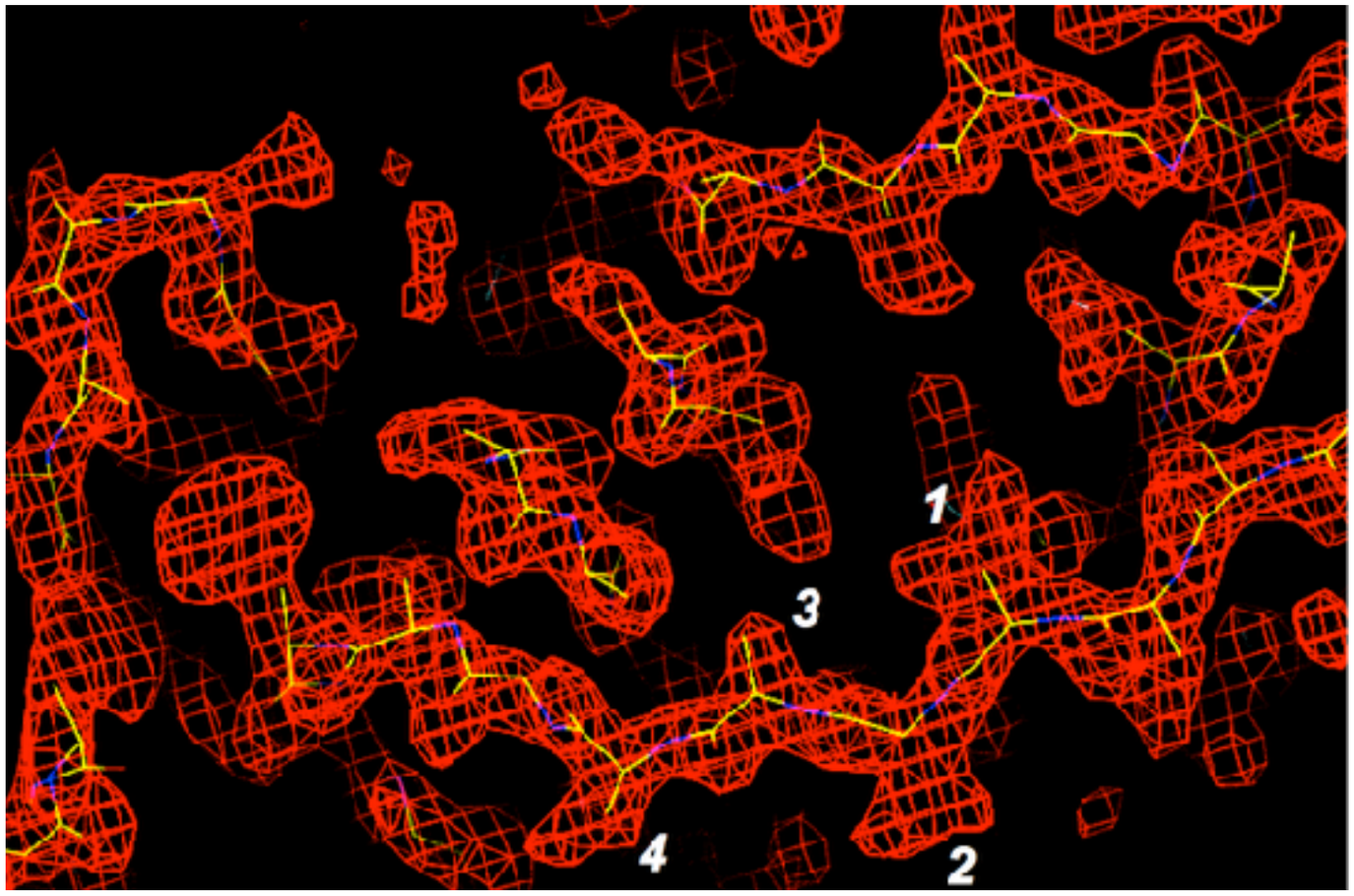
- Tri-peptide fragment library
  - N-terminal extension (3 full amino acids), 9232 members
  - C-terminal extension (CA C O + 2 full amino acids), 4869 members
- Look-ahead scoring: find fragment that can itself be optimally extended
- Each of 10000 fragments: superimpose CA C O on same atoms of last residue in chain (extending by 2 residues): pick best 10
- Each of best 10: extend again by 2 residues and pick best 1:
  - Score for 2-residue extension= best <density> for 4-residue extension based on this 2-residue extension

BERKELEY LAB

Phenix

Image from T. Terwilliger, Los Alamos National Laboratory

The Final Mainchain Trace



- Choose highest-scoring fragment
- Test all overlapping fragments as possible extensions
- Choose one that maximizes score when put together with current fragment
- When current fragment cannot be extended: remove all overlapping fragments, choose best remaining one, and repeat

BERKELEY LAB

Phenix

Image from T. Terwilliger, Los Alamos National Laboratory

Assigning the Sequence

- The sequence is assigned to the mainchain by a probabilistic alignment method, determining the relative probability of every amino acid at each position (based on density and sequence composition)

#	G	A	S	V	I	L	M	C	F	Y	K	R	W	H	E	D	Q	N	P	T
1	6	5	4	18	18	6	1	1	1	2	6	2	2	1	9	6	1	0	1	4
2	4	11	14	37	5	2	0	2	0	0	2	3	0	0	1	2	0	0	0	6
3	11	23	5	12	5	3	2	0	1	3	7	3	1	0	5	3	2	0	2	2
4	7	9	6	16	8	5	2	0	1	3	8	4	1	0	7	6	2	0	3	4
5	31	7	3	7	4	2	1	0	1	3	5	4	1	0	6	2	2	0	11	1
6	1	3	3	41	14	8	0	0	0	0	2	1	0	0	2	4	0	0	1	9
7	0	0	0	0	0	0	0	0	15	63	1	0	17	1	0	0	0	0	0	0
8	2	3	6	23	10	6	2	1	0	1	4	3	0	0	5	16	1	0	1	6
9	96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

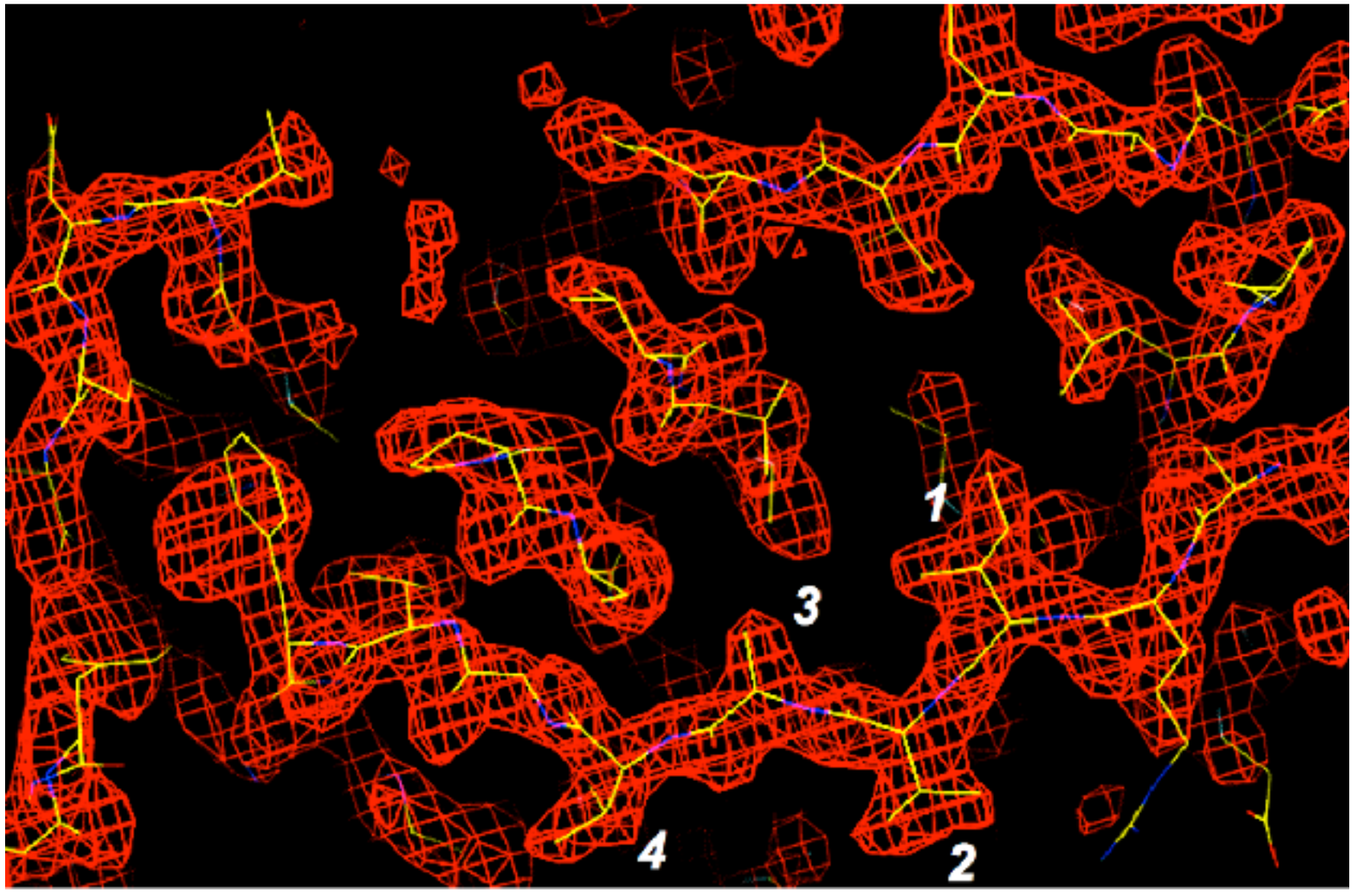
Image from T.Terwilliger, Los Alamos National Laboratory

BERKELEY LAB

Phenix

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868

Side Chain Addition



- Best rotamers, based on correlation coefficient are used
- Refinement is required

Image from T. Terwilliger, Los Alamos National Laboratory

Rapid Secondary Structure Fitting

- Secondary structure elements have recognizable features, even at low resolution (e.g. α-helices look like tubes)
- The essential features (e.g. the long axis of the α-helix) can often be identified
- Once identified, the density can be analyzed further to determine position and orientation

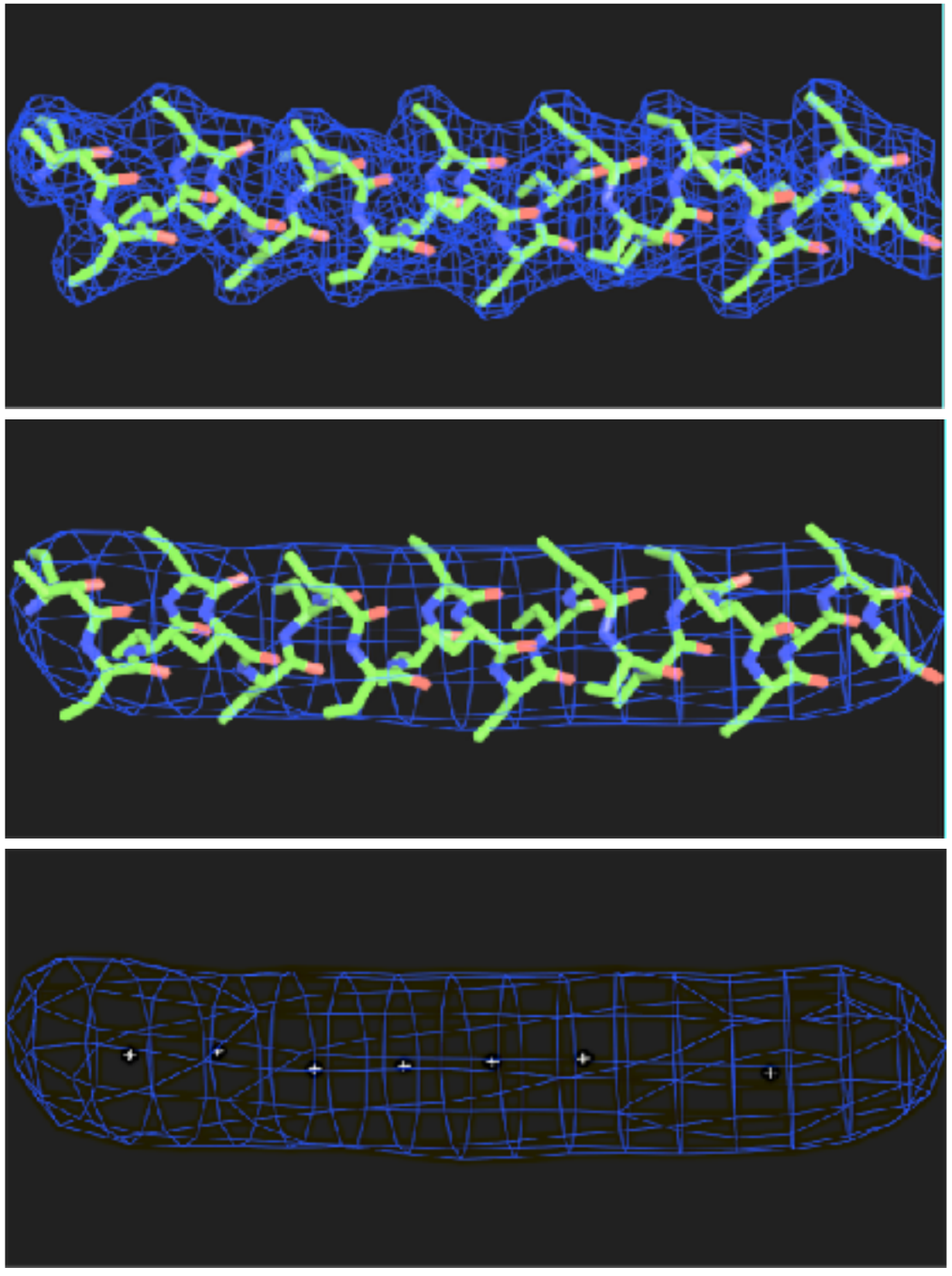


Image from T. Terwilliger, Los Alamos National Laboratory

BERKELEY LAB

Phenix

Rapid Secondary Structure Fitting

- The distribution of density at the main chain atomic positions and the sidechains can be used to determine direction and derive accurate C_α positions
- This is very quick (seconds to minutes)
- Can be followed by sidechain fitting to create a fairly complete model
- Similar methods can be applied to find β-sheets

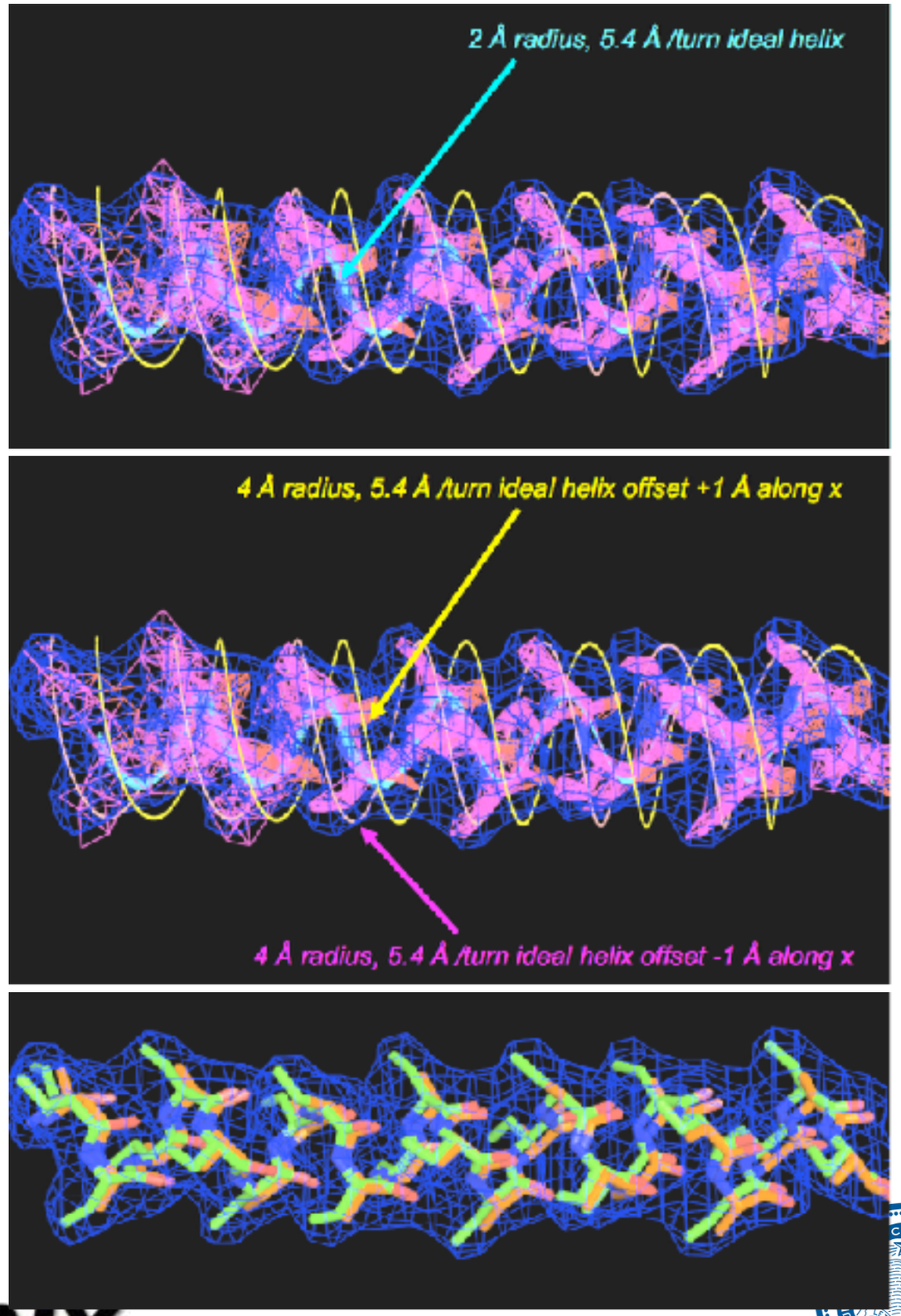
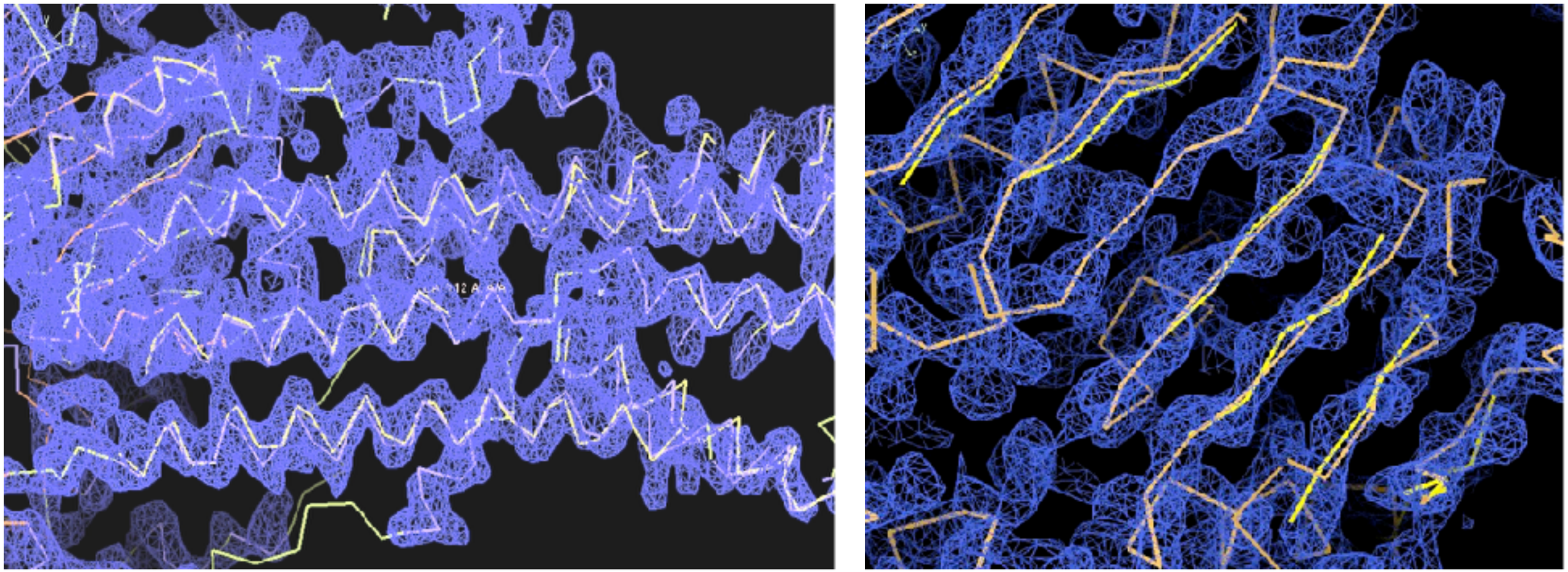


Image from T. Terwilliger, Los Alamos National Laboratory

High Resolution Data is Not Required



Calcium release channel 3.1 Å. Data courtesy of P. Nissen

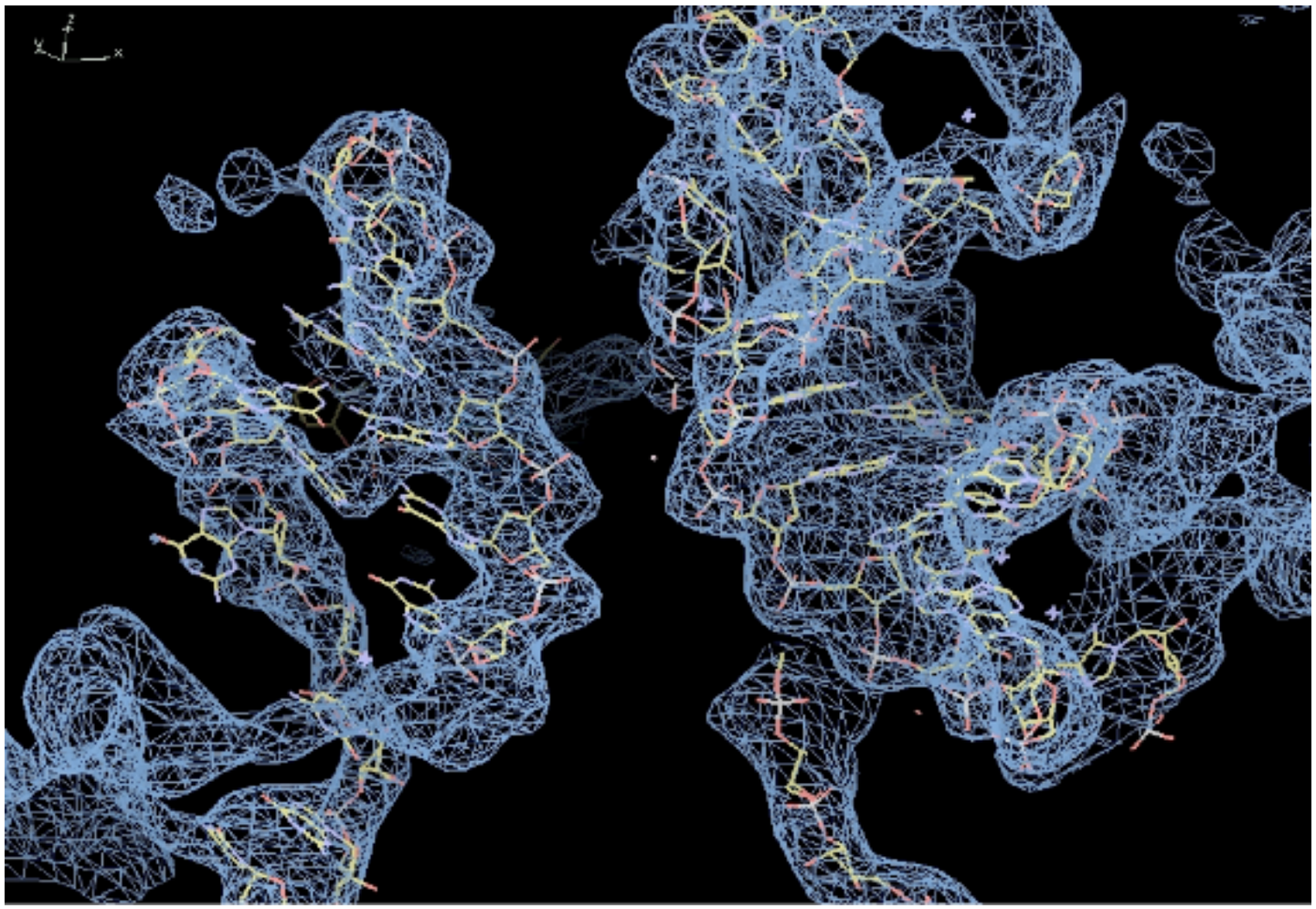
- This rapid method works even at modest resolution
- Can be used to determine if structure solution is likely given the current experimental phases
- Success will depend on the quality of the phases (more than the resolution)

BERKELEY LAB

Phenix

Image from T. Terwilliger, Los Alamos National Laboratory

Nucleic Acid Model Building



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

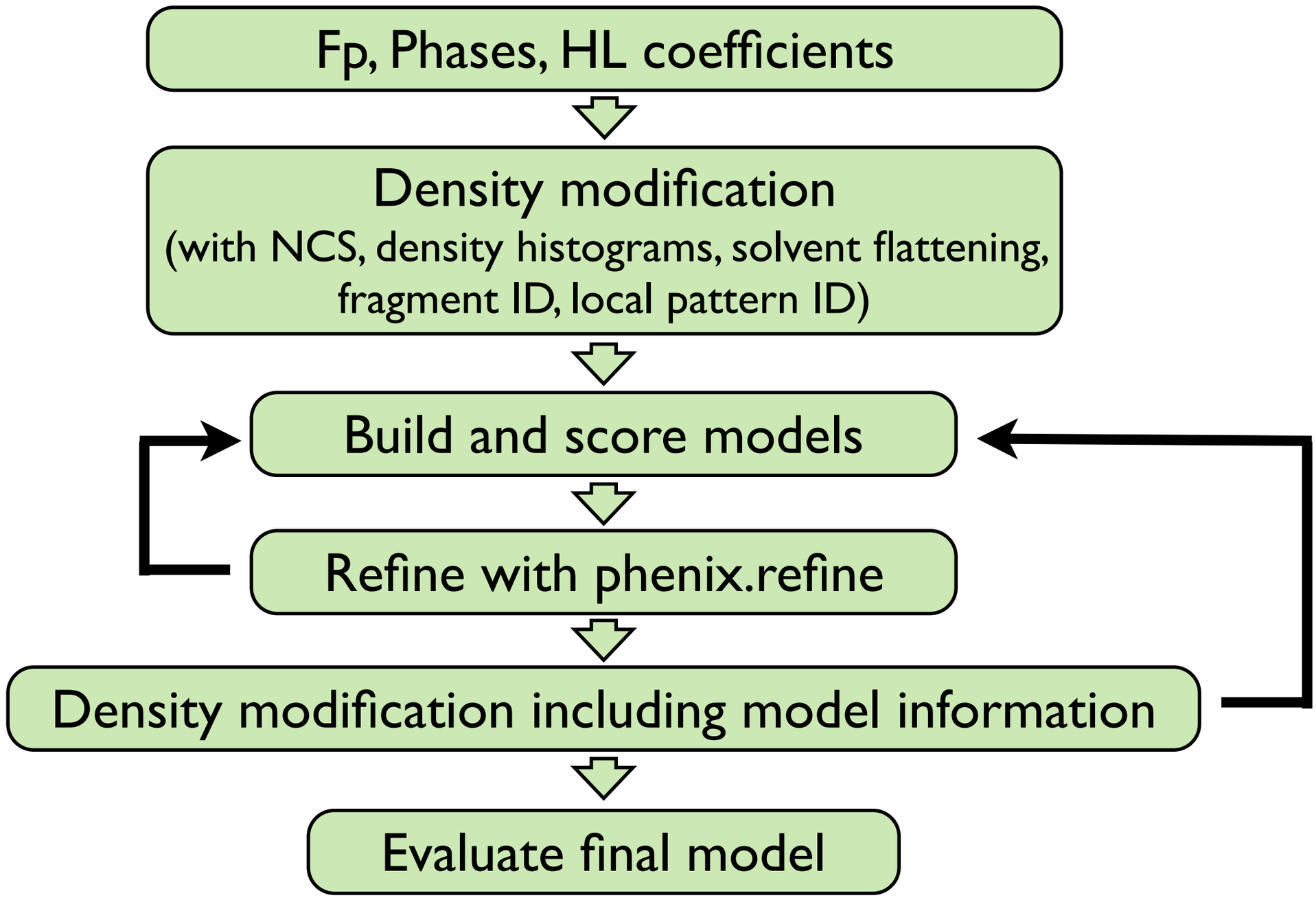
- Nucleic acid structures can be built using fragment location (short A-form or B-form helices), followed by extension
- Works well even with low resolution
- Current limitation is the simultaneous building of protein and nucleic acid

BERKELEY LAB

Phenix

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868

Automated Model Building/Rebuilding



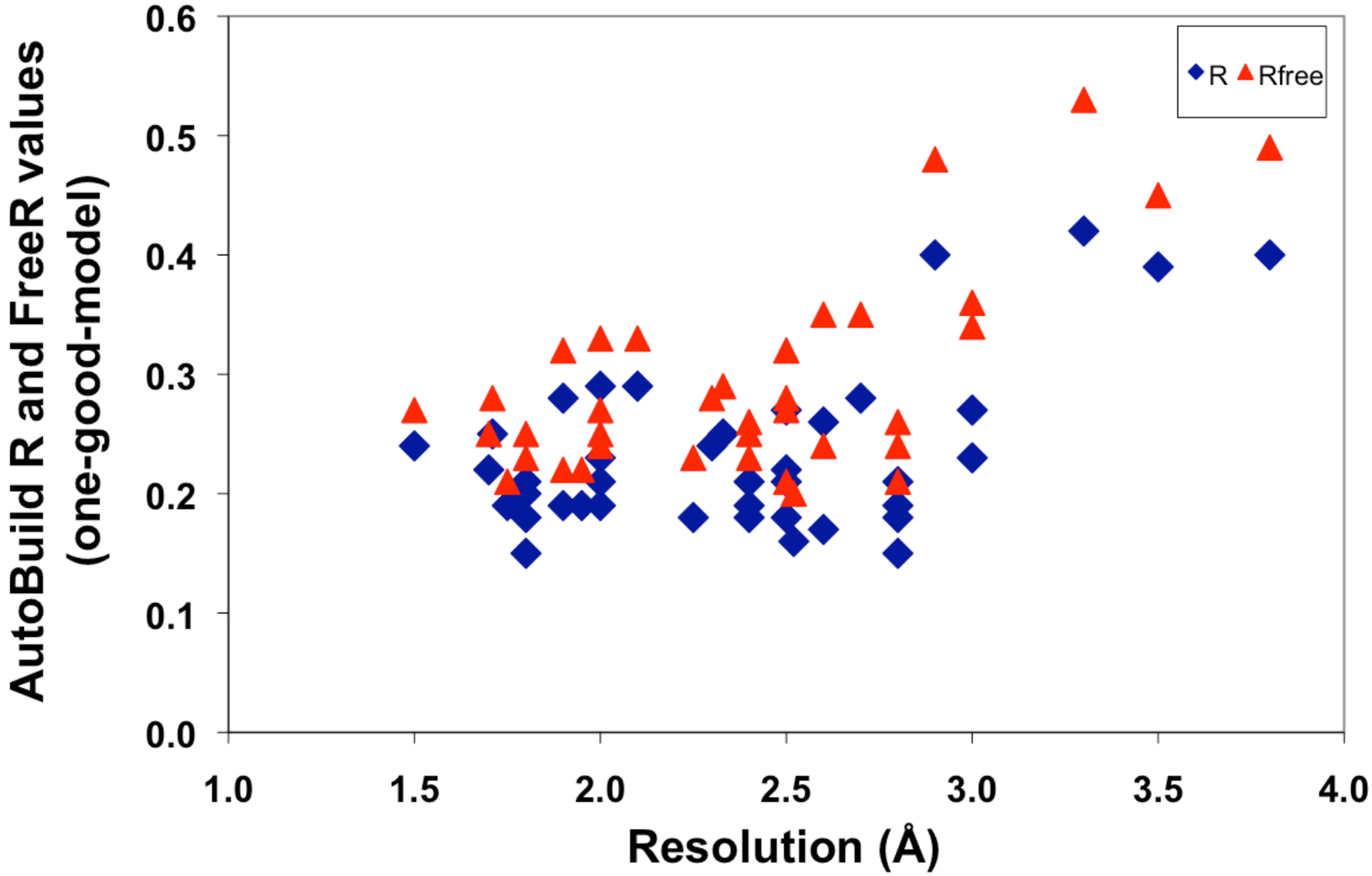
Acta Cryst. 2007, **D63**:597-610.

Acta Cryst. 2008, **D64**:61-69.

Acta Cryst. 2008, **D64**:515-524.

Phenix

Automated Building Depends on Data Quality



- Automated model building results are relatively independent of resolution
- Results are more dependent on data quality and intrinsic quality of the electron density map

BERKELEY LAB

Phenix

Tom Terwilliger, Los Alamos National Laboratory

Automated Model Building for Cryo-EM

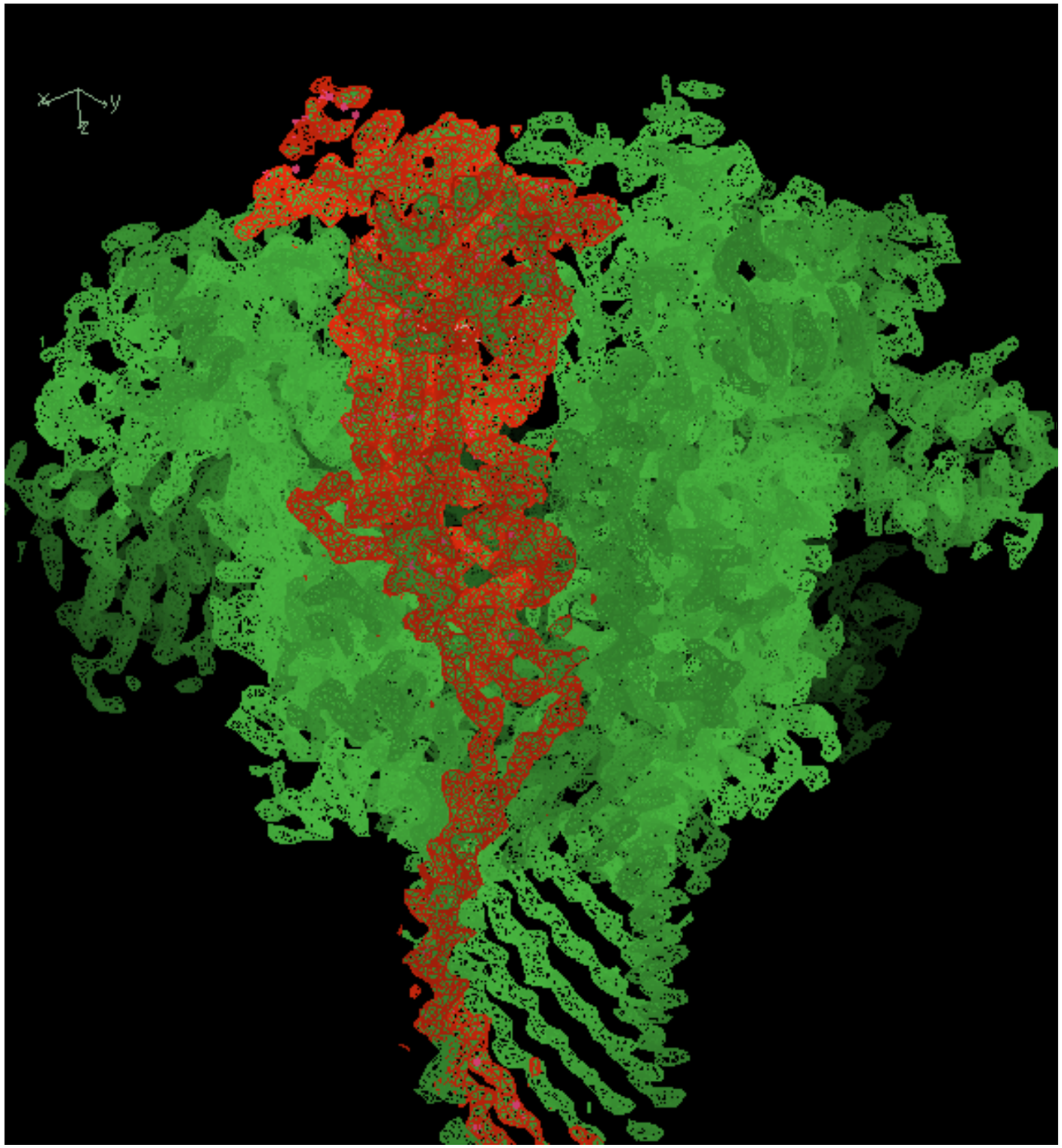
- Higher resolution (4.5Å and better) makes automated building possible
- Being developed in Phenix by Tom Terwilliger (Los Alamos National Lab):
  - Automatically segmenting maps and extracting the asymmetric unit of reconstruction
  - Create maps that emphasize information at various resolutions by variable map sharpening
  - Trace the protein main chain using nearly-constant $C\alpha$-$C\alpha$-$C\alpha$ distances and angles
  - Identify direction of the main-chain in models by fit to density

BERKELEY LAB
Lawrence Berkeley National Laboratory

Phenix

THE UNIVERSITY OF CALIFORNIA · BERKELEY · 1868 · LET THERE BE LIGHT

Automated Model Building



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

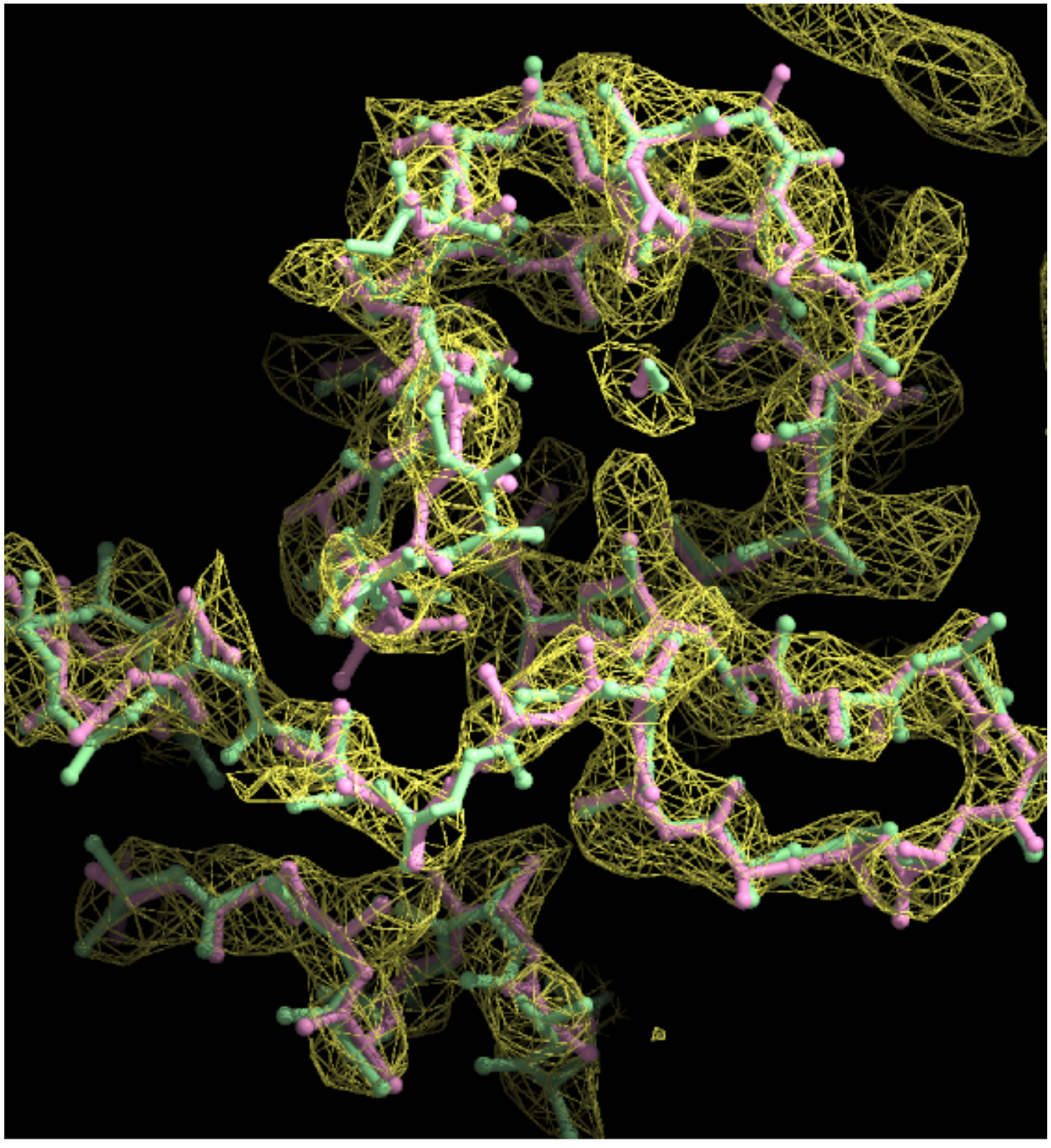
Tom Terwilliger (LANL), Oleg Sobolev (LBNL)

BERKELEY LAB
Lawrence Berkeley National Laboratory

Phenix

THE UNIVERSITY OF CALIFORNIA · BERKELEY · 1868

Automated Model Building



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

Autobuilt model (pink)

Deposited model (green)

(only main-chain and C_β atoms shown)

Tom Terwilliger (LANL), Oleg Sobolev and Pavel Afonine (LBNL)

BERKELEY LAB
Lawrence Berkeley National Laboratory

Phenix

THE UNIVERSITY OF CALIFORNIA · BERKELEY · 1868 · LET THERE BE LIGHT

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