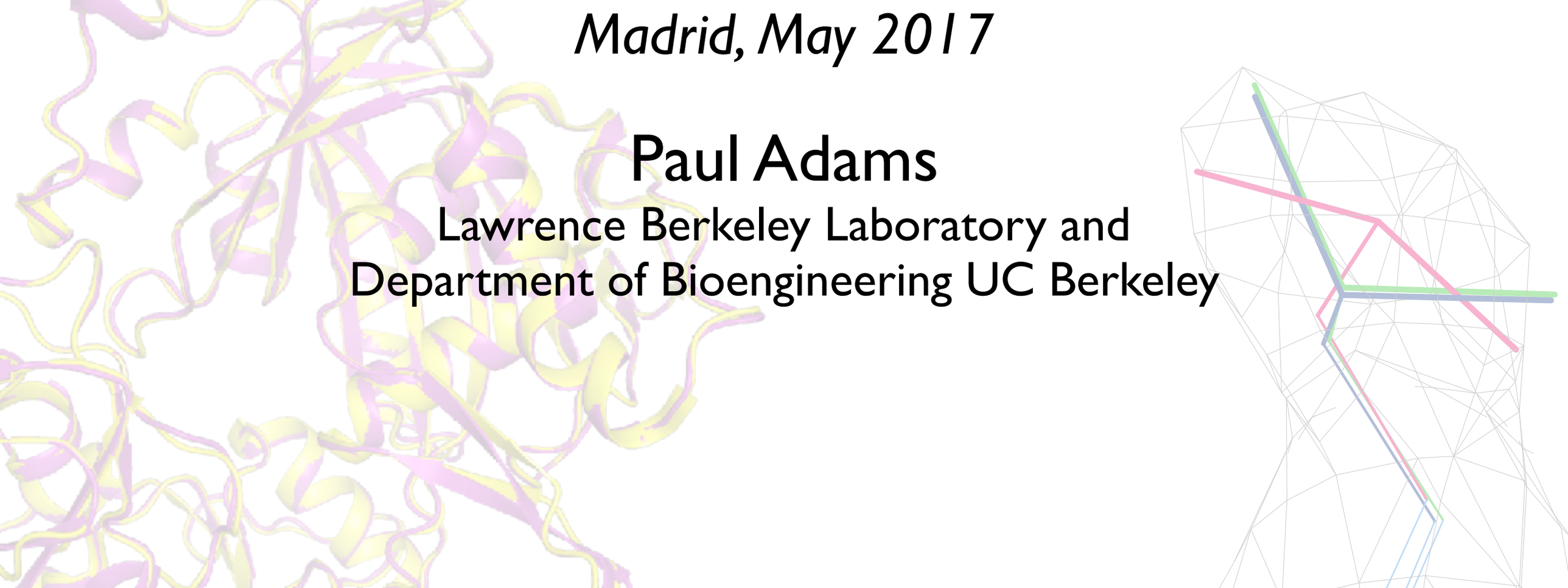
Low Resolution Refinement

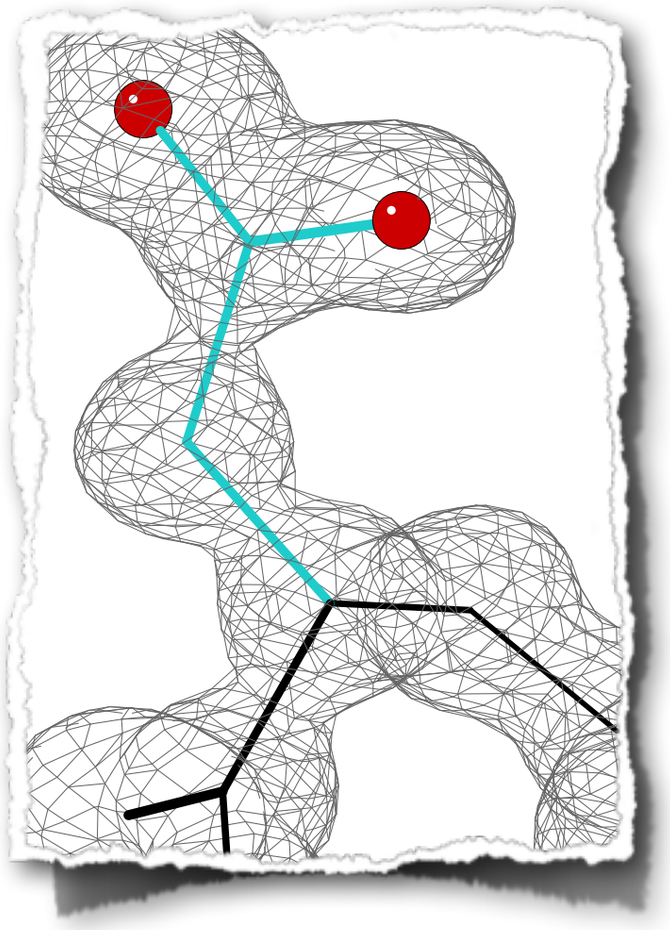
Macromolecular Crystallography School
Madrid, May 2017

Paul Adams
Lawrence Berkeley Laboratory and
Department of Bioengineering UC Berkeley

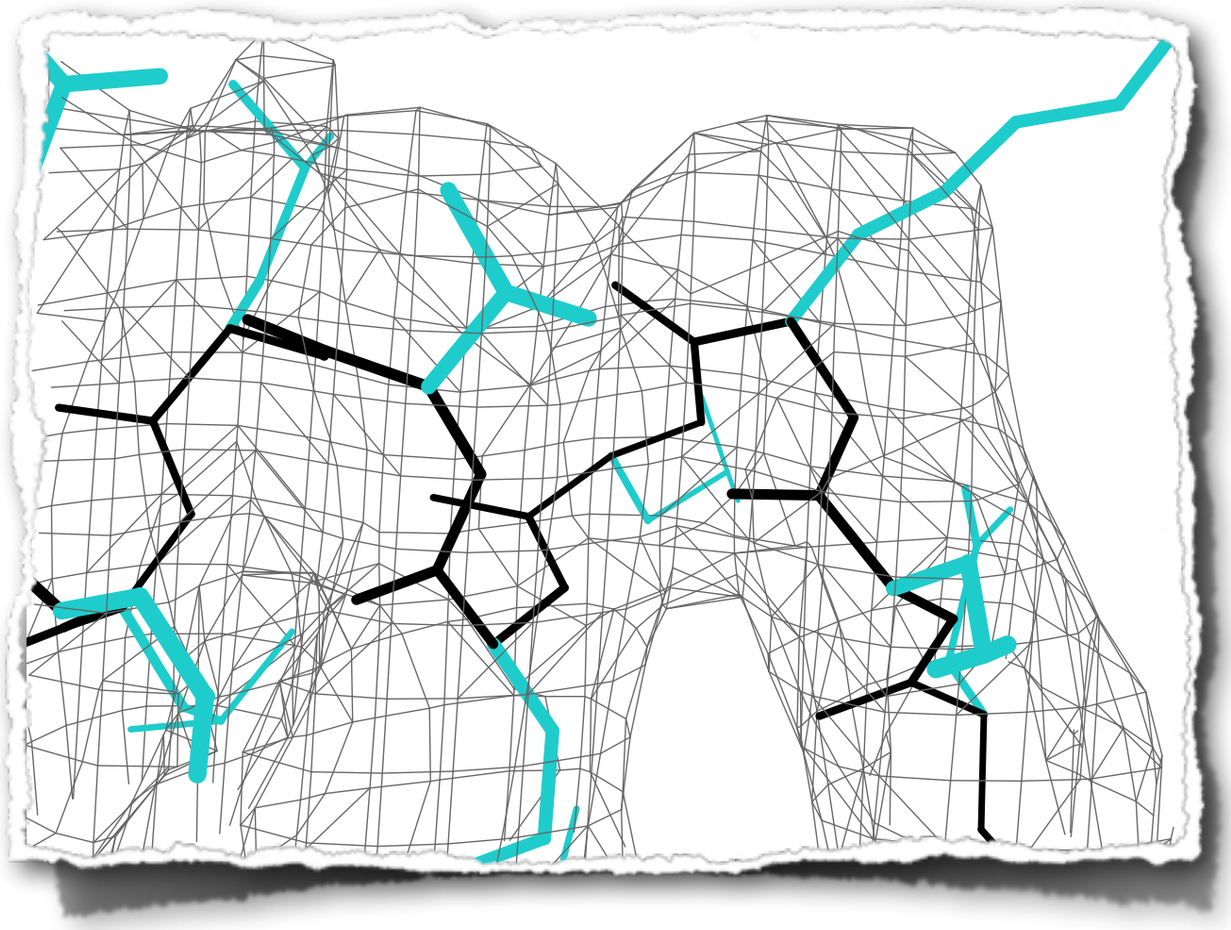


Macromolecular Crystallography

PDBID: 2gkg
Resolution: 1.00Å



PDB ID: 3k7a
Resolution: 3.80Å



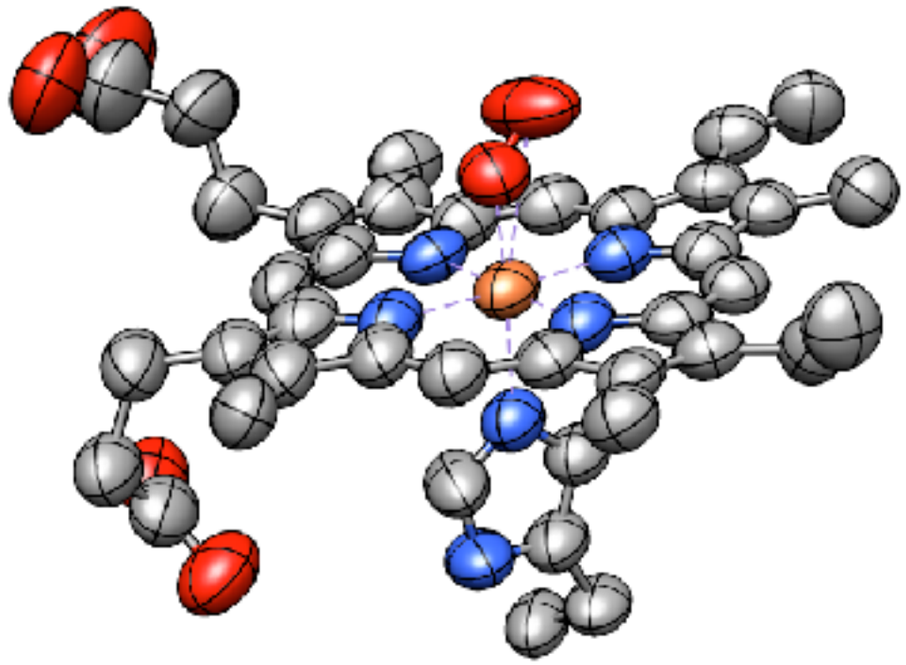
- Many challenges, but low resolution data is increasingly an issue:
  - How to interpret "featureless" maps (pattern matching, chemical constraints)
  - How to optimize models with sparse data (prior information)

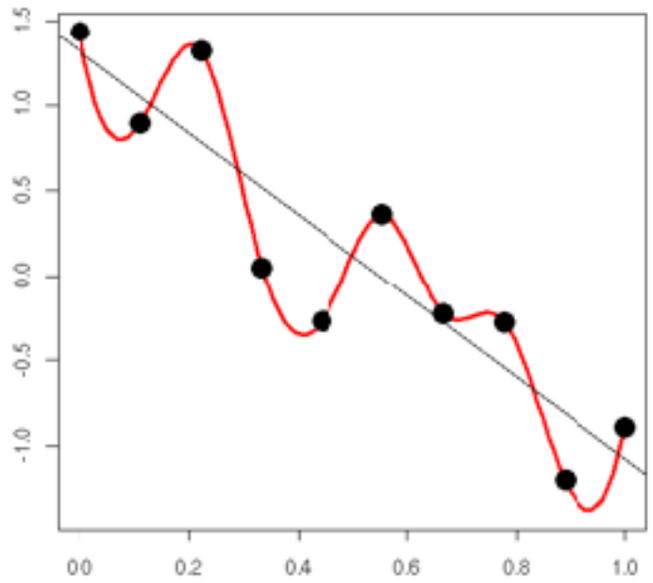
BERKELEY LAB
Lawrence Berkeley National Laboratory

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868

The Challenge of Too Few Data

- With only low resolution data we typically have too many parameters to optimize
  - Atomic coordinates, displacement parameters
- Underdetermined optimization problems lead to overfitting of the data
- To help address overfitting we can:
  - Add prior information to reduce the number of effective parameters
  - Remove parameters
- Current refinement methods do not define a reasonable chemical result in the absence of data





BERKELEY LAB
Lawrence Berkeley National Laboratory

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868

New Methods in Phenix for Improving Models

- Using prior structural knowledge as additional restraints:
  - Secondary structure
  - Protein mainchain conformations (Ramachandran)
  - Related high resolution structures as restraints
  - Multiple copies of the same molecule as restraints *(c.f. local NCS restraints in SHELX)*
- Automated correction of models during refinement using prior knowledge of stereochemistry:
  - Fixing of rotamers
  - Flipping of side chains

BERKELEY LAB
Lawrence Berkeley National Laboratory

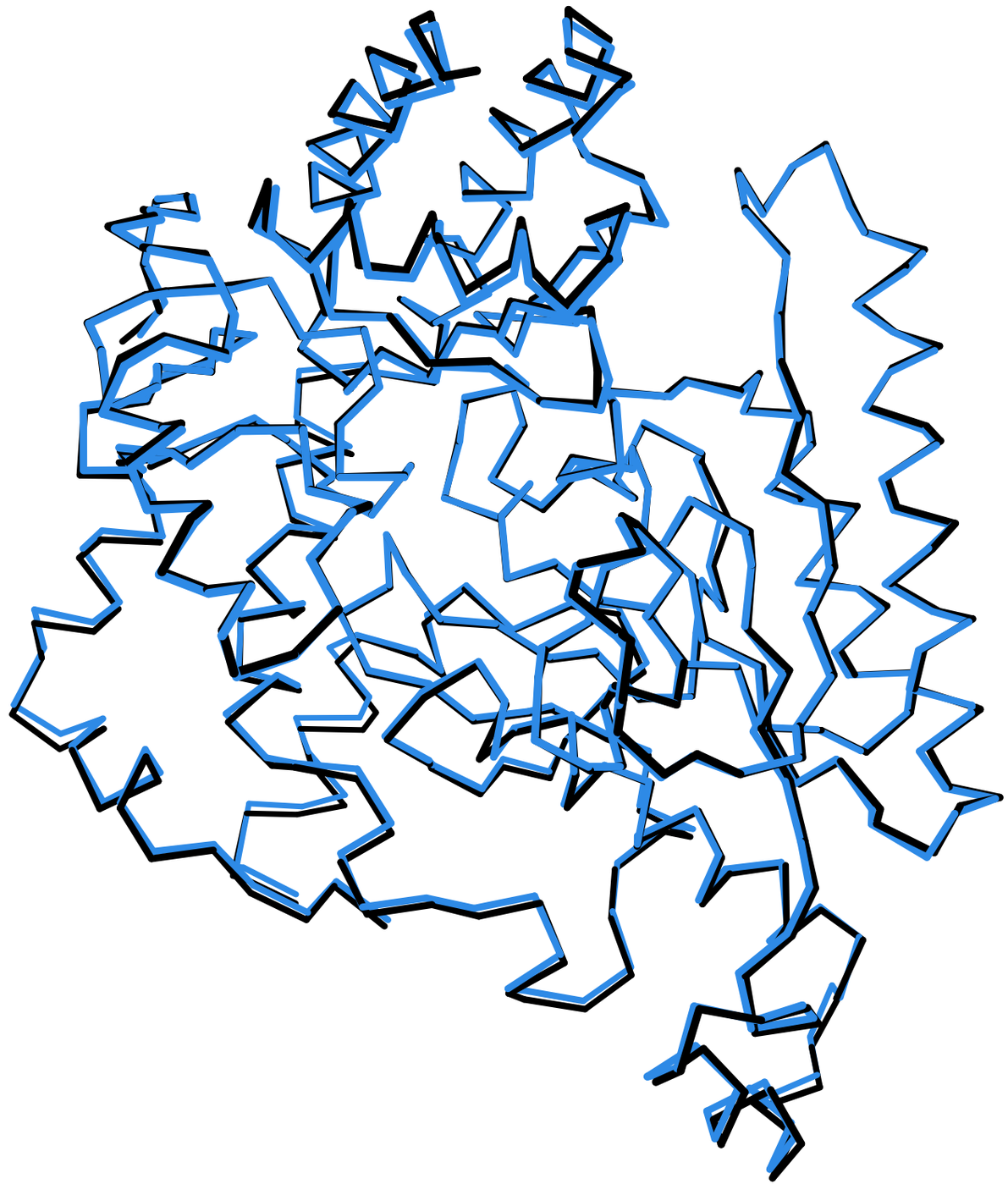
Reference model restraints

(Jeff Headd)

1GTX and 1OHV

1GTX: 3.0 Å

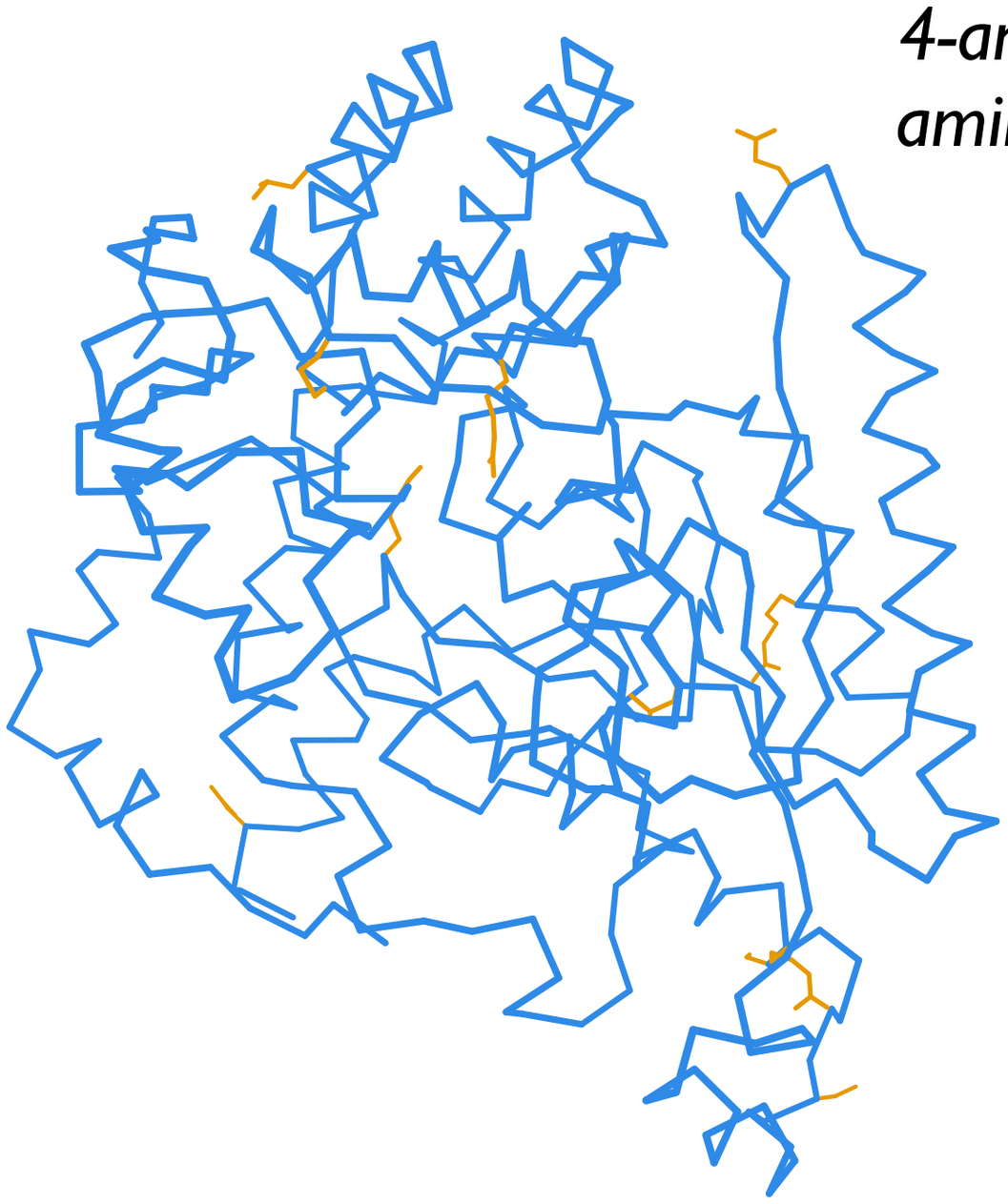
1OHV: 2.3 Å

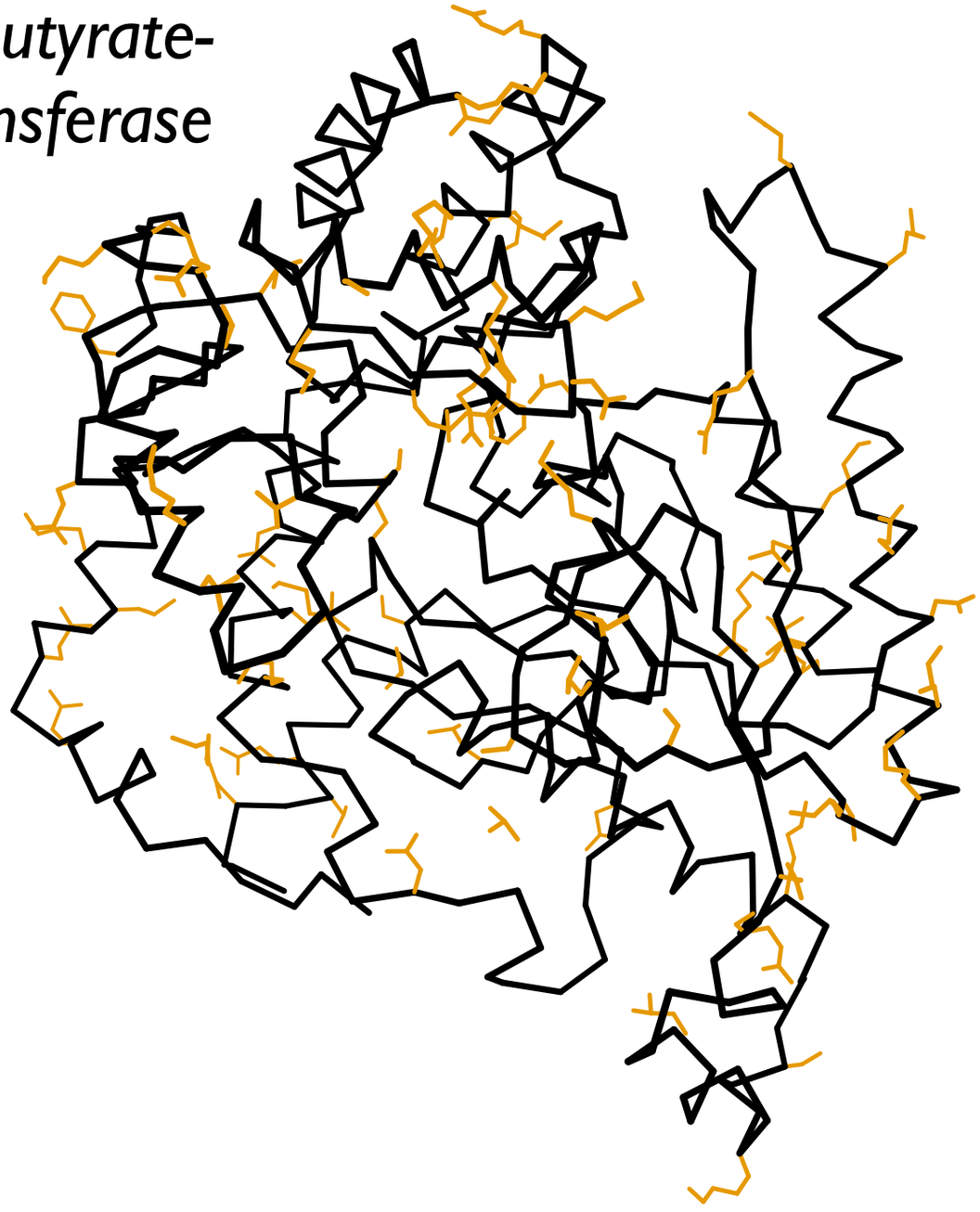


4-aminobutyrate-aminotransferase

1GTX and 1OHV

4-aminobutyrate-aminotransferase





1OHV: 2.3 Å

1GTX: 3.0 Å

BERKELEY LAB
Lawrence Berkeley National Laboratory

Reference Model Restraints

Combines two concepts:

- Pre-correct rotamer outliers
  - Set rotamer outliers in the model to match the torsion angles of the reference model if the reference model has an acceptable rotamer at that position and there is no significant clash or density mismatch
- Generate reference torsion restraints
  - Restrain each torsion angle in the working model to the corresponding torsion angle in the reference model
    - Chains are aligned using SSM alignment to allow for sequence differences
  - Restraints take the form of a modified harmonic 'top-out' potential that allows for structural differences

BERKELEY LAB
Lawrence Berkeley National Laboratory

Headd JJ et al., 2012, *Acta Cryst.* **D68**:381-390

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868

Reference model restraints

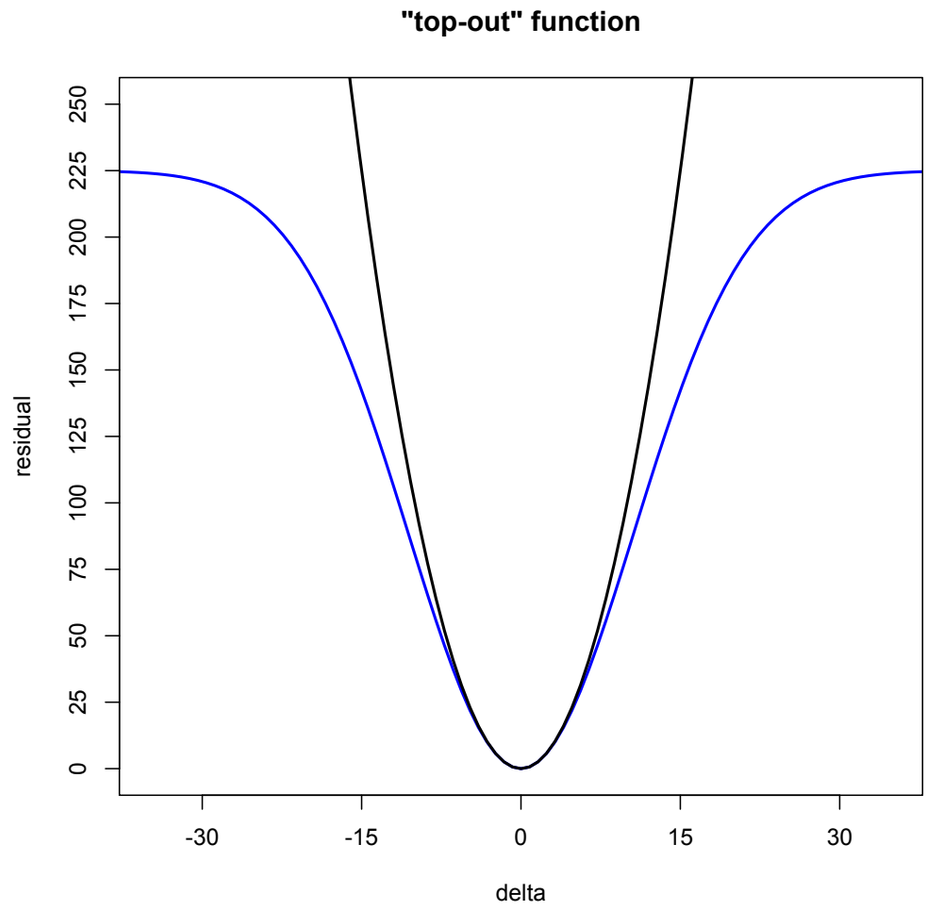
$$E_{total} = \sum_{i=1}^{n} E_i$$

Simple harmonic potential: $E_i = w\Delta_i^2$

'Top-out' potential: $E_i = \tau(1.0 - e^{\frac{-\Delta_i^2}{l^2}})$

$$\tau = wl^2$$

$$w = \frac{1}{\sigma^2}$$



Similar potentials are used in REFMAC5 and BUSTER - Geman-McClure robust estimator function

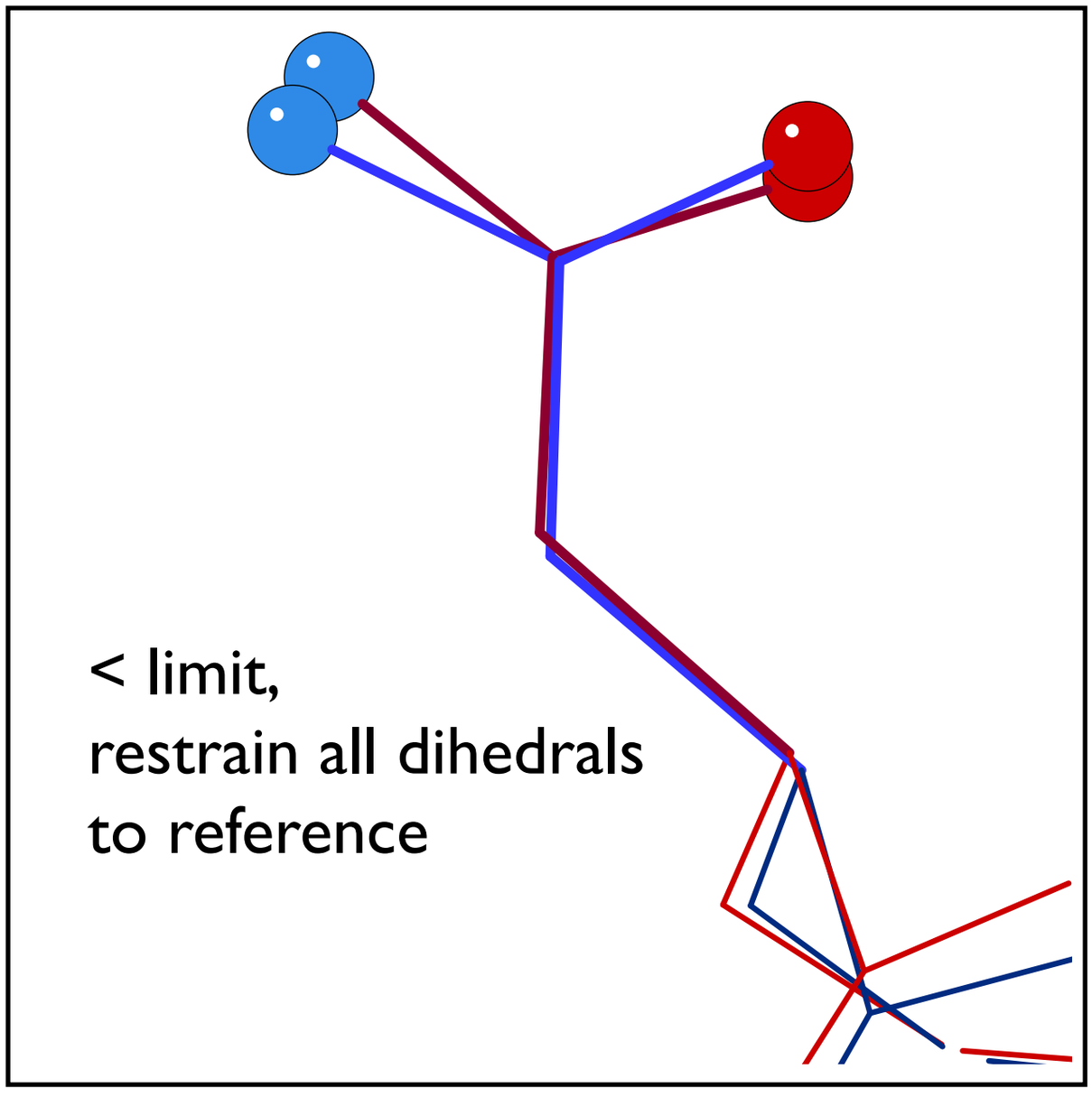
where σ is the ESD, Δ is the difference between the model dihedral and reference dihedral, and *l* is a 'limit' parameter that limits how far the model dihedral may vary from the reference dihedral before being shut off.

developed by Ralf Grosse-Kunstleve

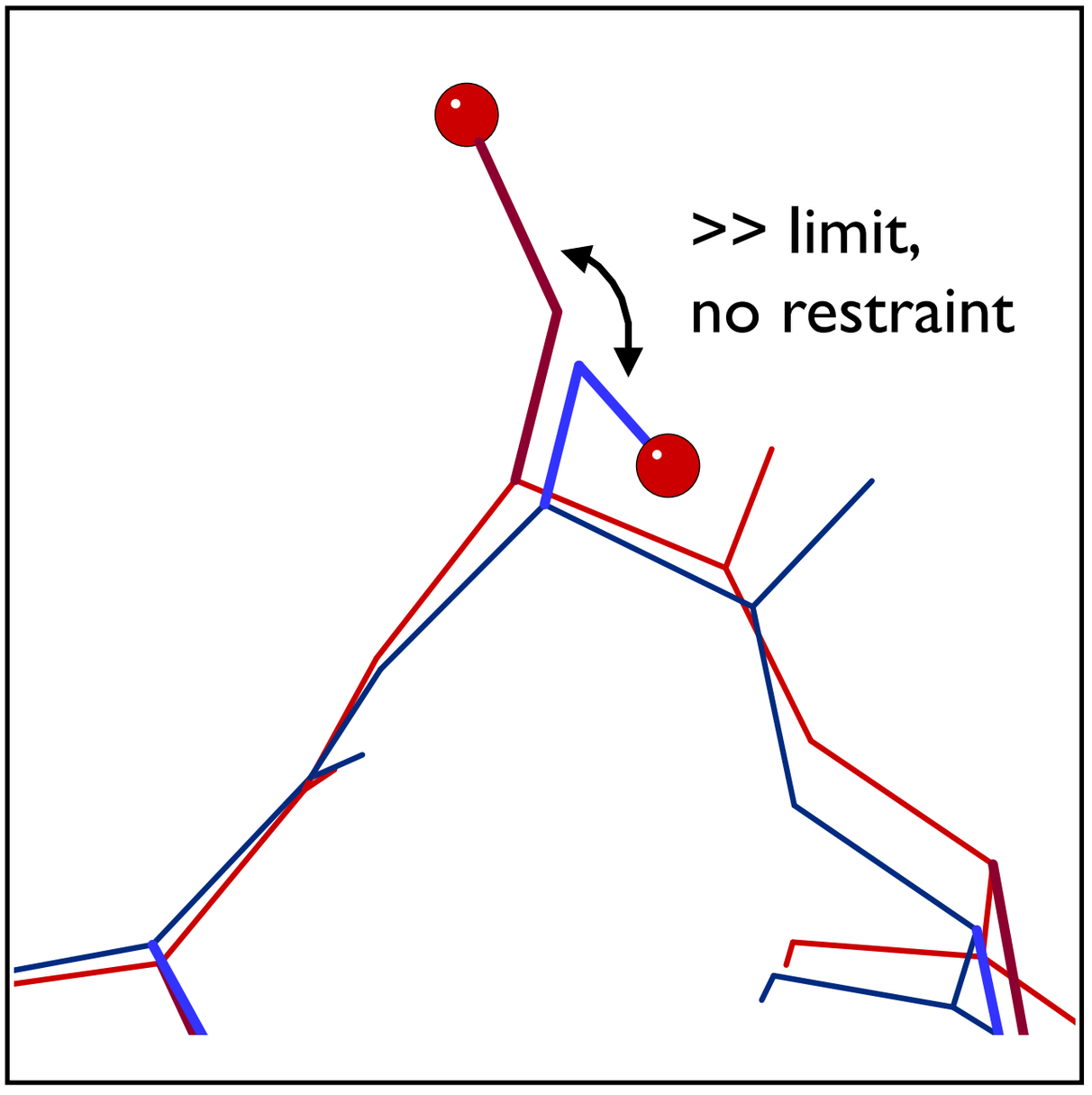
default: limit = 15.0°

The ‘limit’ parameter

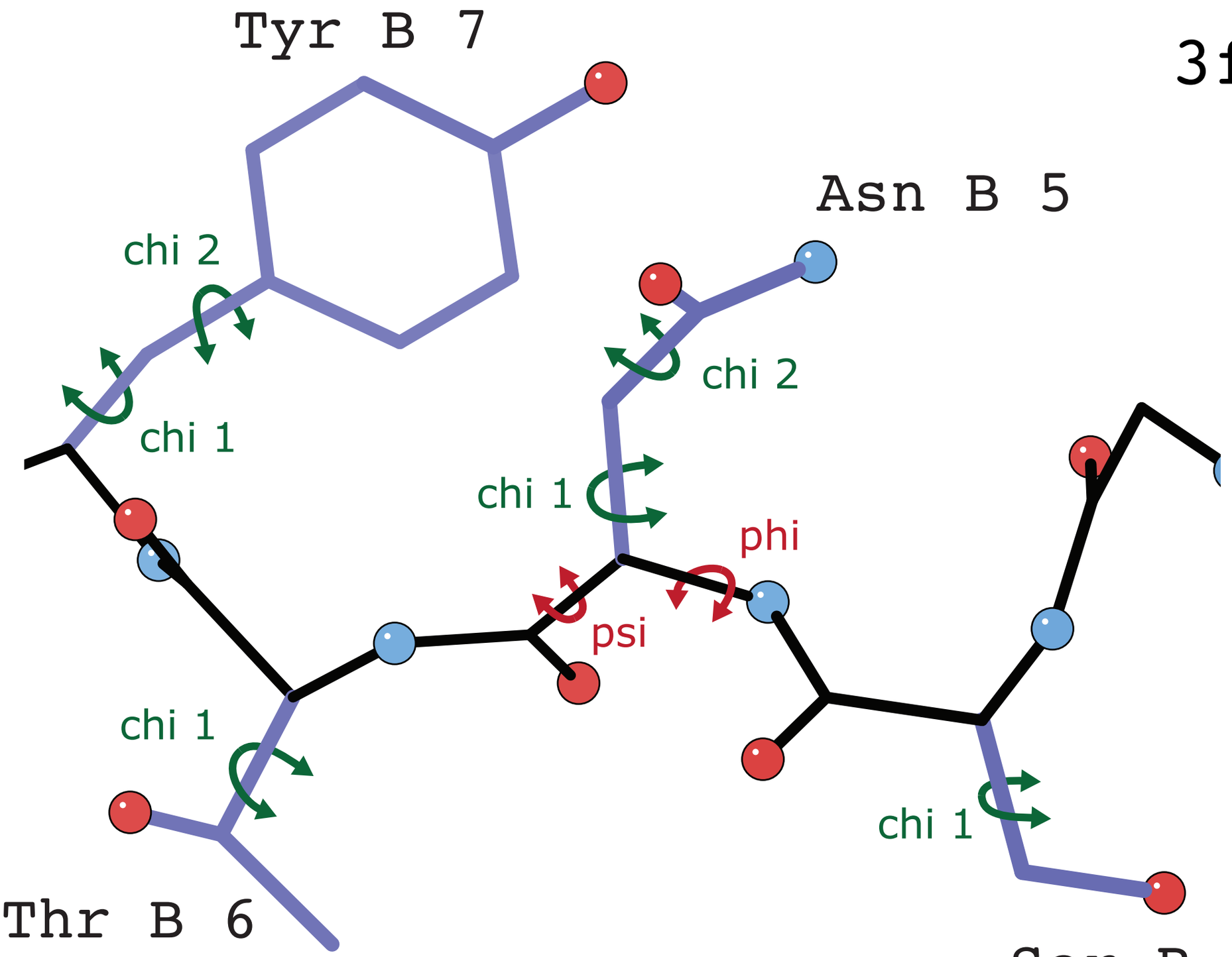
default: limit = 15.0°



< limit,
restrain all dihedrals
to reference

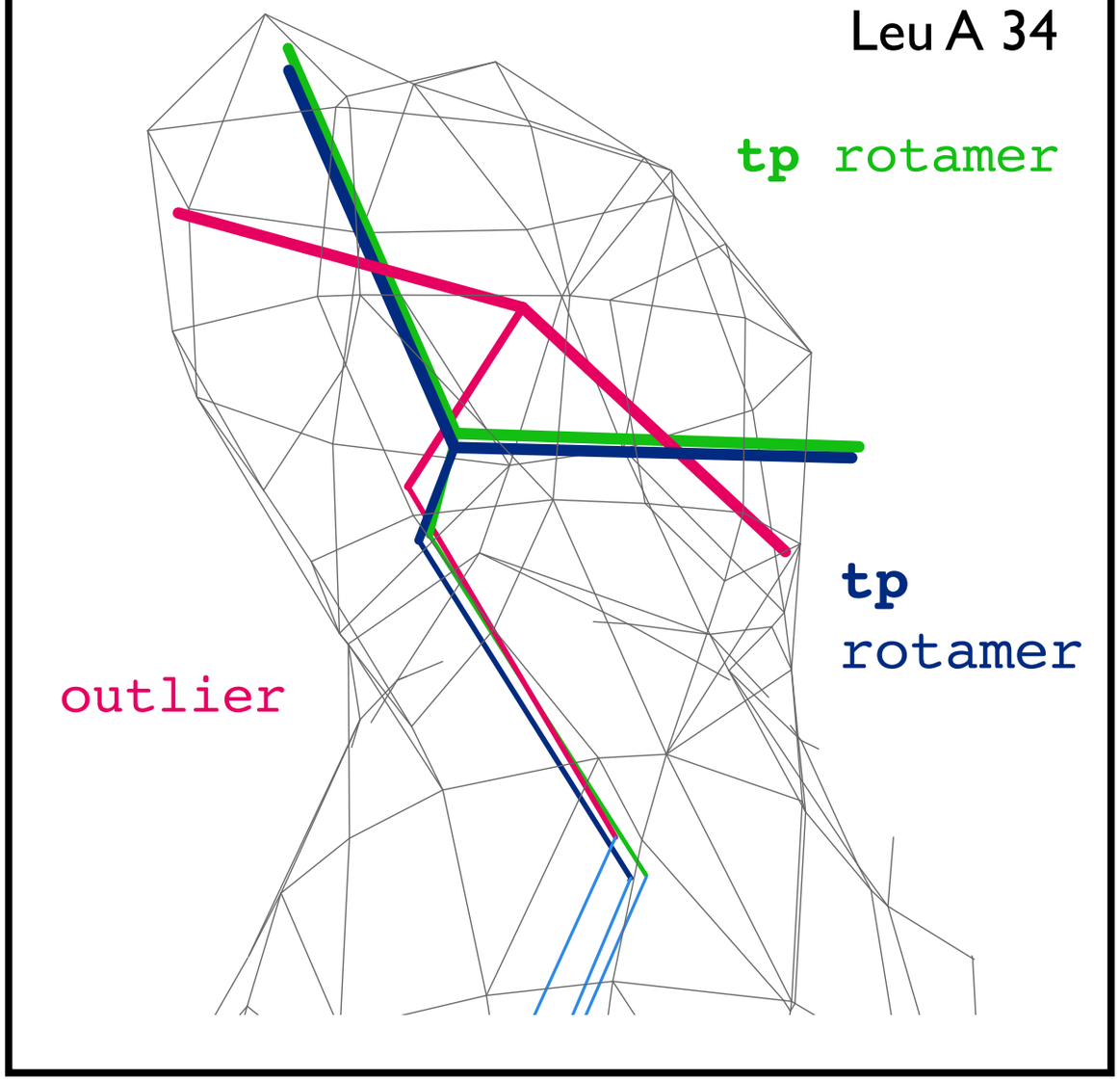


Why torsion angles?

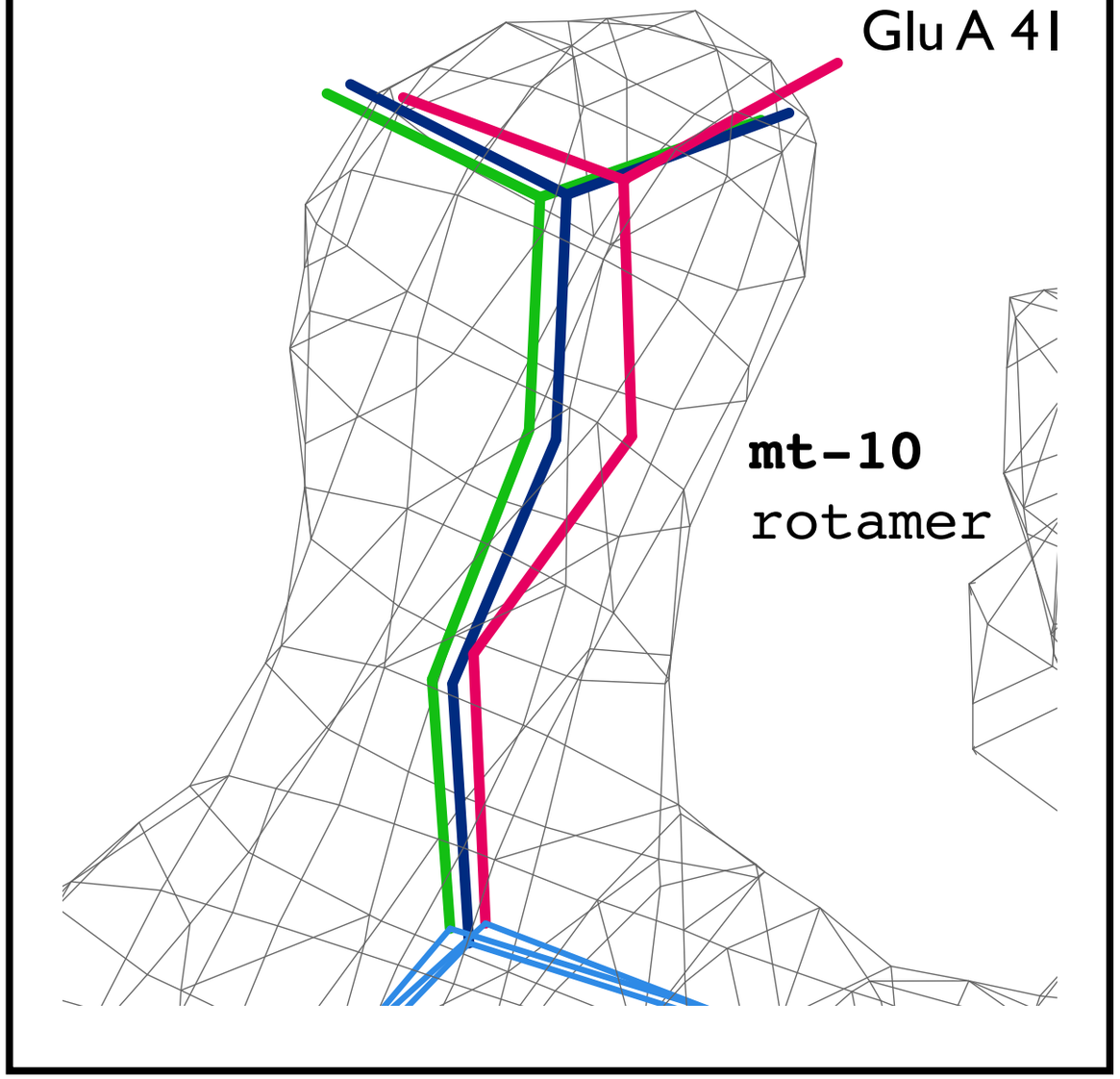


1GTX/1OHV reference example

outlier correction



restrained refinement



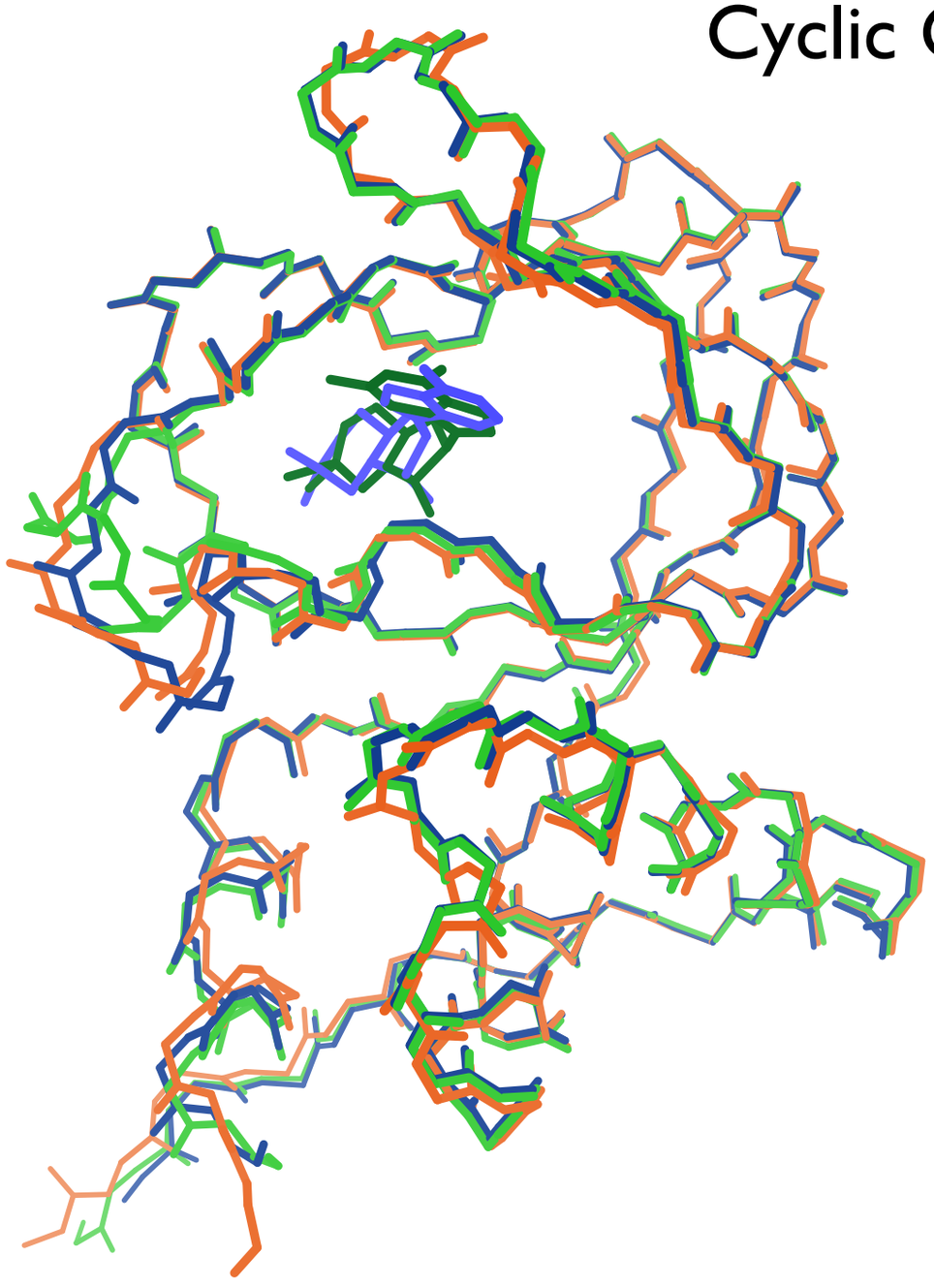
1GTX (3.0Å) 1OHV (2.3Å) 1GTX w/ 1OHV reference

5 macrocycles of *phenix.refine*
w/ reference restraints →

R_{free}: 0.2379 → 0.2186
ΔR: 0.833 → 0.60
MolProbity: 64th → 96th

Practical Example

Cyclic GMP-dependent protein kinases (PKG's)



cAMP bound: 2.49Å

cGMP bound: 3.20Å

APO form: 2.69Å

JJ Kim et al. (2011) Crystal structures of PKG Iβ (92-227) with cGMP and cAMP reveal the molecular details of cyclic nucleotide binding. *PLoS ONE*.

BERKELEY LAB
Lawrence Berkeley National Laboratory

Cyclic GMP-dependent protein kinase

cAMP bound: 2.49Å

cGMP bound: 3.20Å

APO form: 2.69Å

	Validation Criteria	cAMP bound
All-Atom Contacts	Clashscore, all atoms:	16.53
	Clashscore percentile	81st
Protein Geometry	Poor rotamers:	2.61%
	Rama outliers:	0.00%
	Rama favored:	98.80%
	Cβ dev. > 0.25Å:	0
	MolProbity score:	2.04
	MP score percentile	95th
	Res w/ bad bonds:	0.00%
	Res w/ bad angles:	0.00%
Residual	R-work	0.1960
	R-free	0.2264

cGMP bound
56.57
15th
18.58%
2.02%
85.48%
23
3.84
12th
2.38%
5.95%
0.2102
0.2582

→

cGMP bound
24.56
87th
4.00%
0.40%
96.00%
0
2.61
96th
0.00%
1.18%
0.1980
0.2397

APO
28.52
46th
10.53%
3.19%
89.02%
3
3.29
12th
0.79%
0.98%
0.2205
0.2612

→

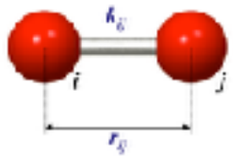
APO
19.5
15th
3.66%
0.60%
96.61%
0
2.43
89th
0.00%
0.20%
0.2166
0.2525

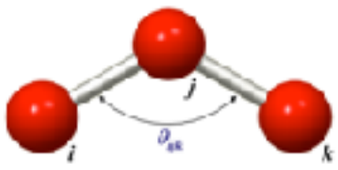
BERKELEY LAB
Lawrence Berkeley National Laboratory

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868

Sources of Prior Information

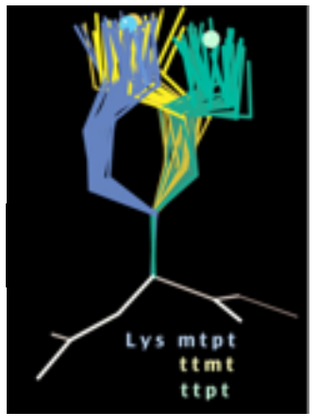
Covalent geometry

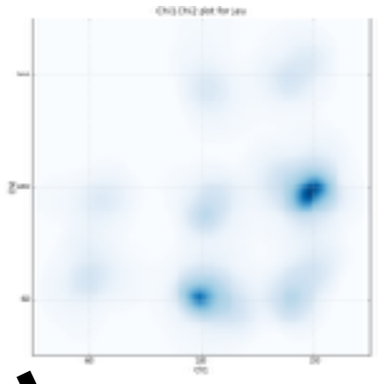


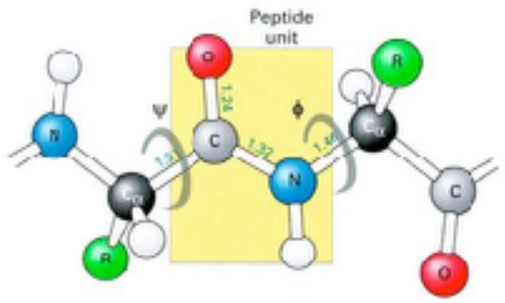


Images from PumMa web site (http://www.pumma.nl)

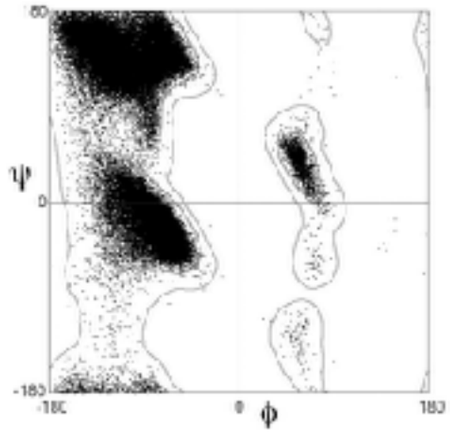
Sidechain distributions

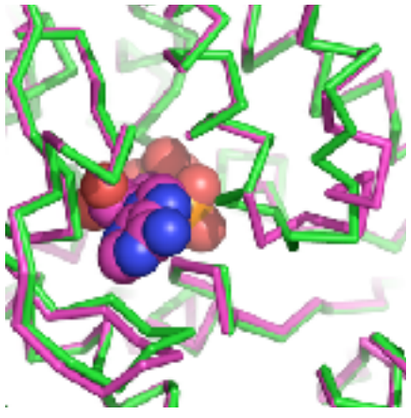




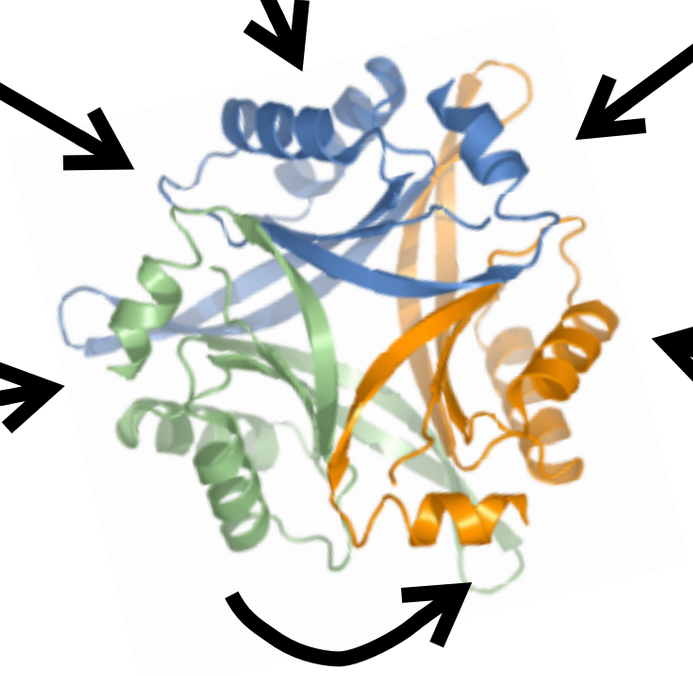


Mainchain distributions

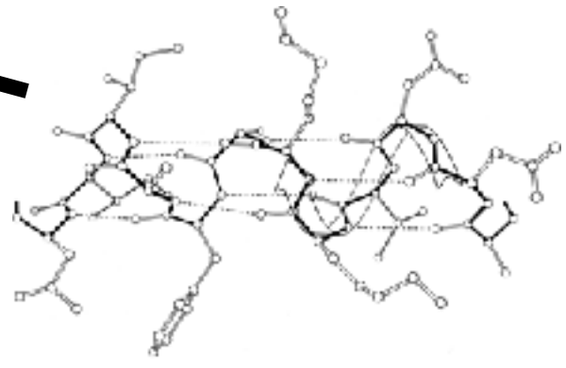




Related structures



Internal symmetry



Secondary structure

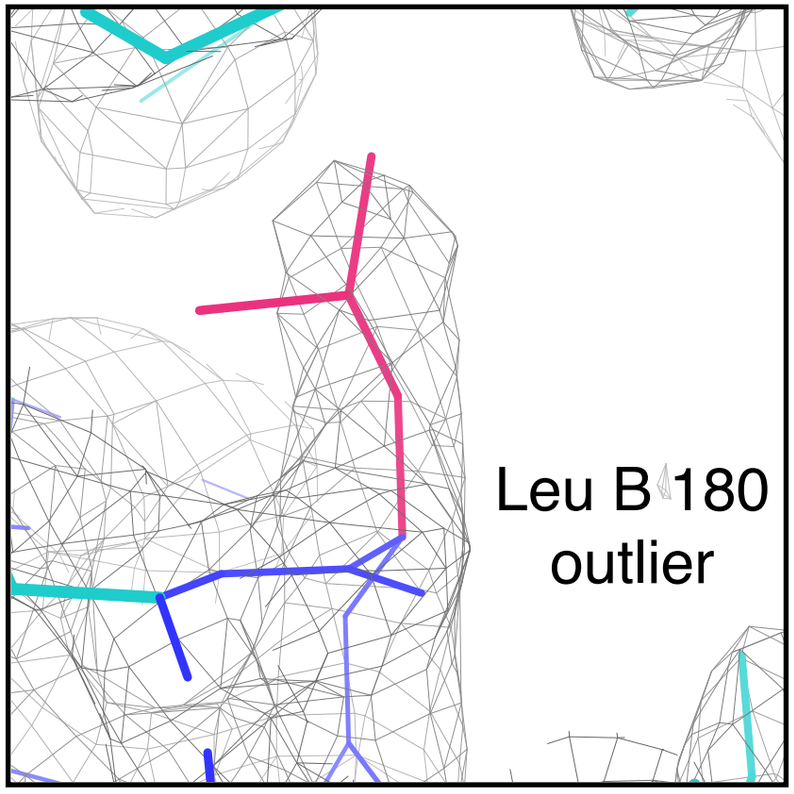
Torsion space NCS restraints

(Jeff Headd)

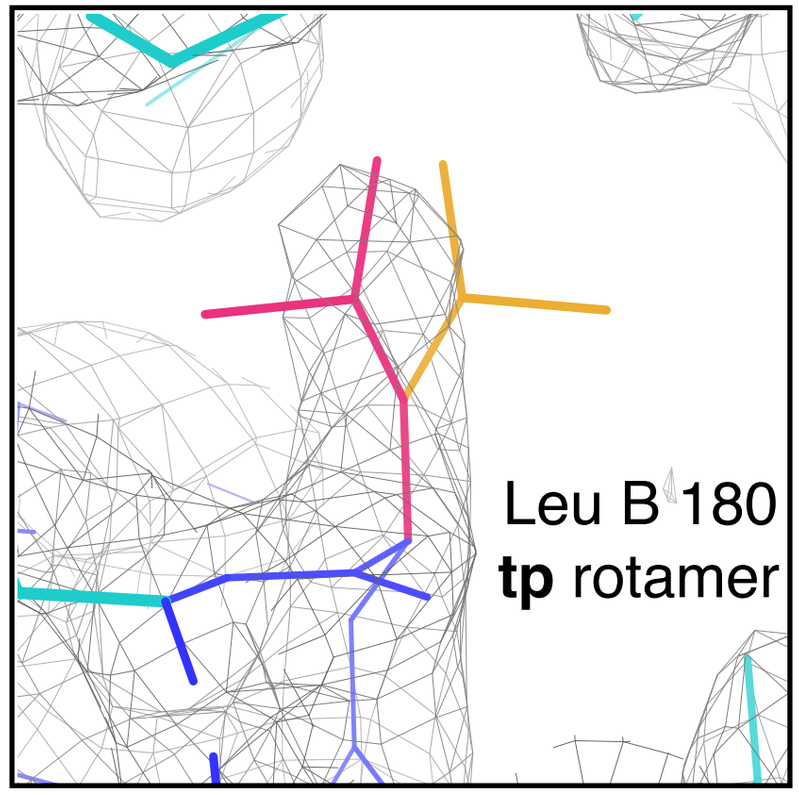
1b04: 2.8 Å DNA ligase

rotamer outlier correction

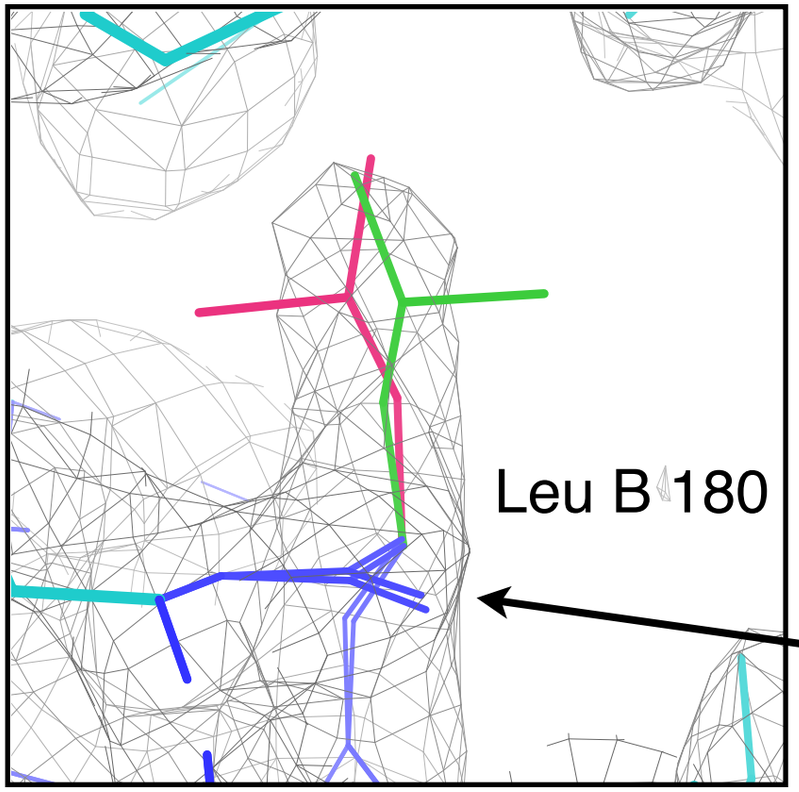
1. Identify rotamer outlier



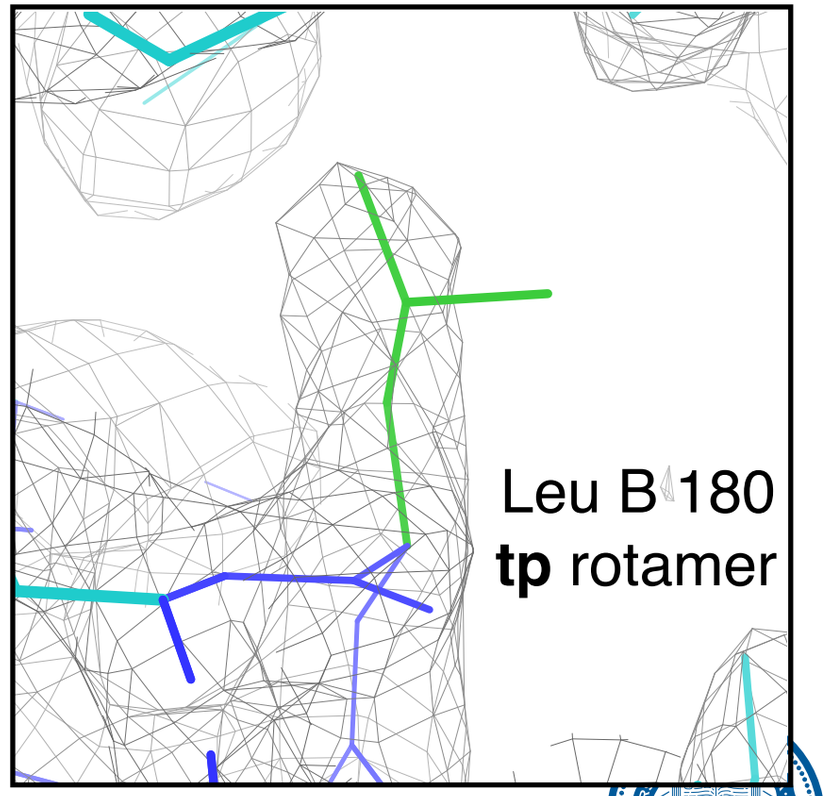
2. correct to corresponding rotamer in NCS-related chain by matching χ angles



3. ‘backrub’ search, then limited χ angle torsion search



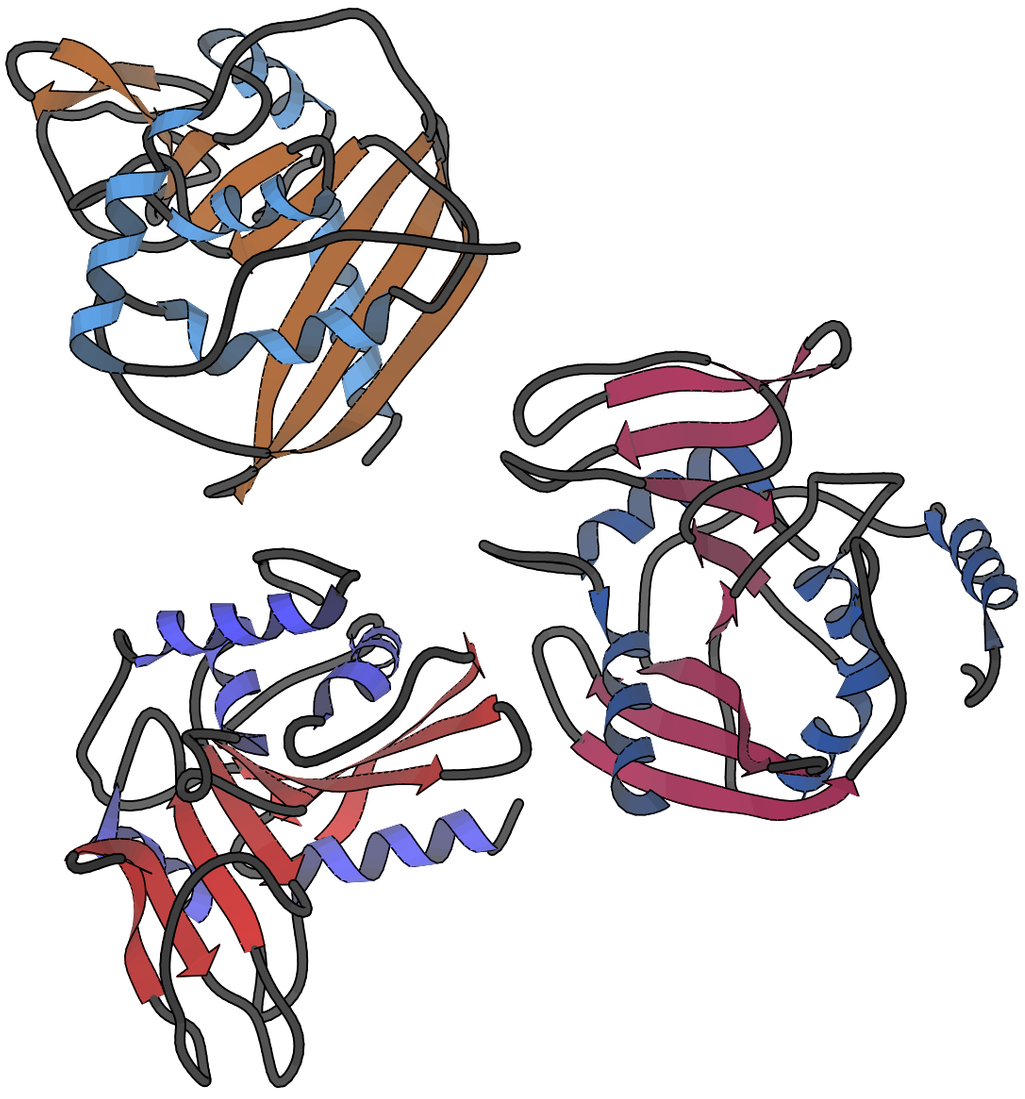
4. verify rotamer is still correct match



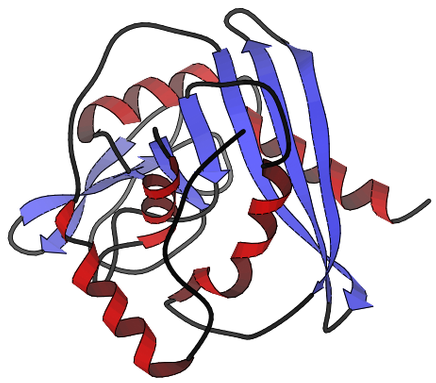
BERKELEY LAB
Lawrence Berkeley National Laboratory

molecular replacement ⟶ refinement

3hd0: 2.70 Å
endonuclease



MR w/ Phaser



2w35
2.15 Å

R_{work} = 0.3844

AutoBuild

- Rebuild in place
- NCS on for rebuilding
- NCS off for refinement
- No water picking

R_{work} = 0.1895
R_{free} = 0.2745

R_{gap} = 0.085

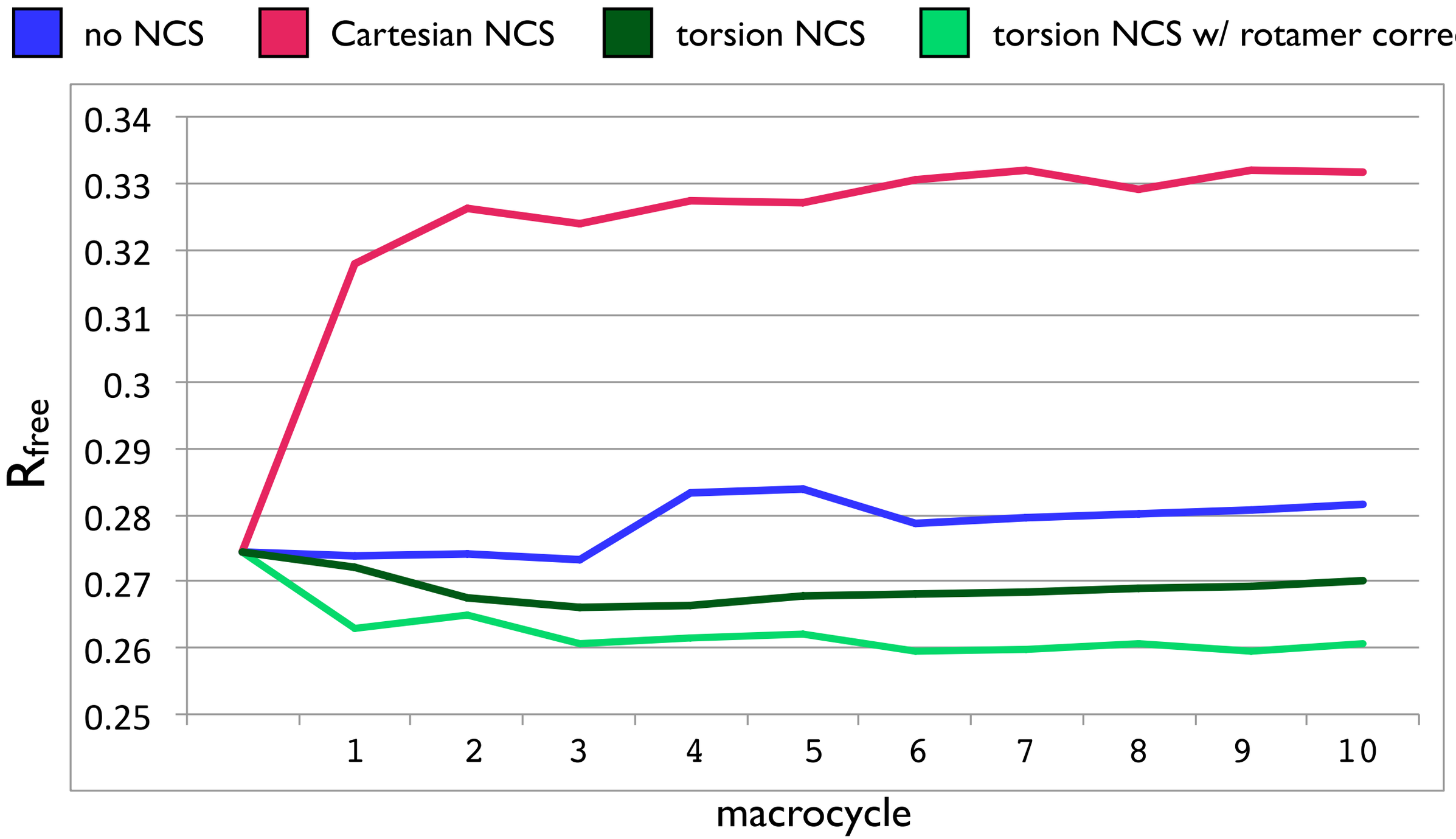
phenix.refine

- 10 macrocycles
- optimize weights
- No NCS, Cartesian NCS, torsion NCS w/ and w/o rotamer correction

BERKELEY LAB
Lawrence Berkeley National Laboratory

THE UNIVERSITY OF CALIFORNIA BERKELEY
1868

3hd0 refinement



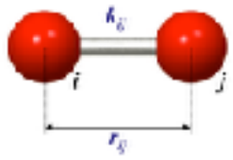
torsion NCS w/ rotamer correction ⟶ $R_{work} = 0.2040$ $R_{free} = 0.2606$ $R_{gap} = 0.056$

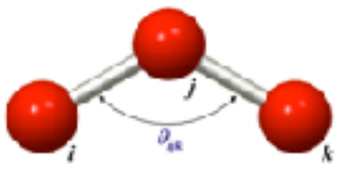
BERKELEY LAB
Lawrence Berkeley National Laboratory

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868

Sources of Prior Information

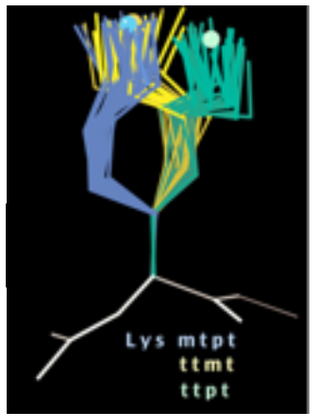
Covalent geometry

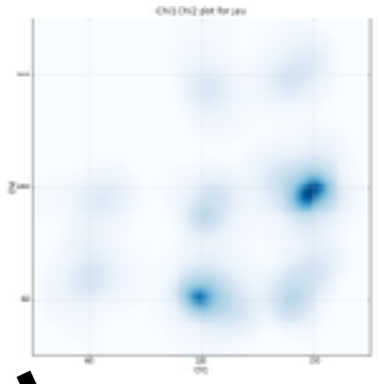


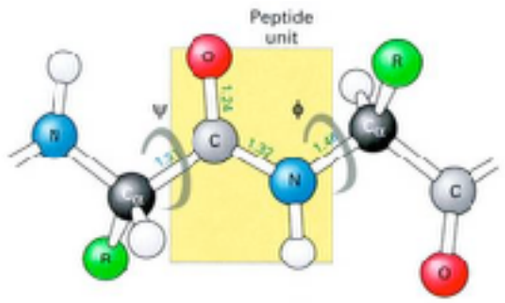


Images from PumMa web site (http://www.pumma.nl)

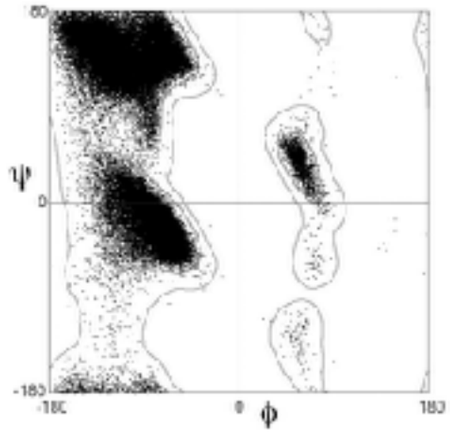
Sidechain distributions

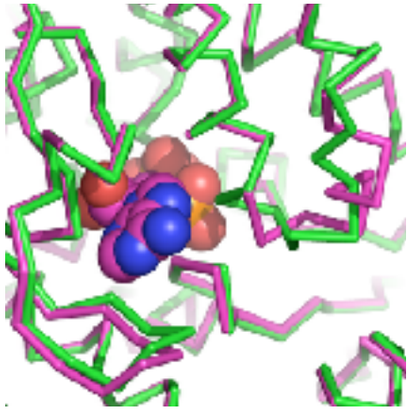




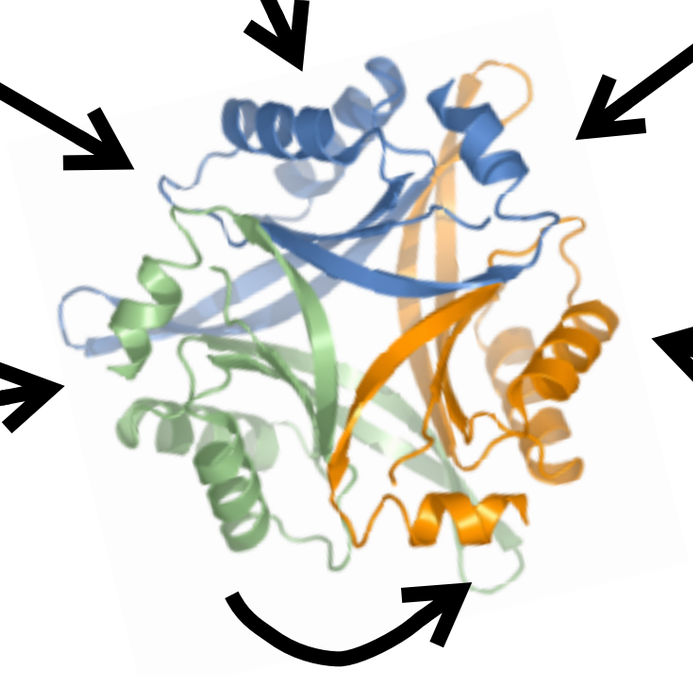


Mainchain distributions

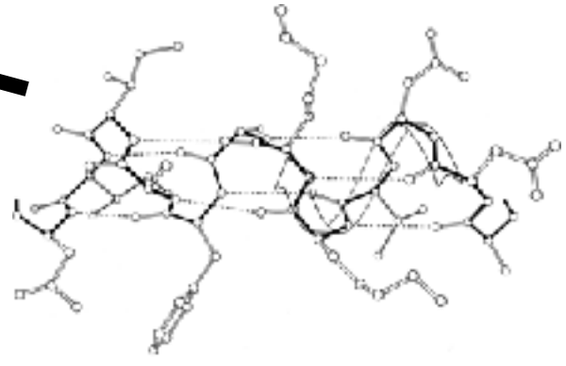




Related structures



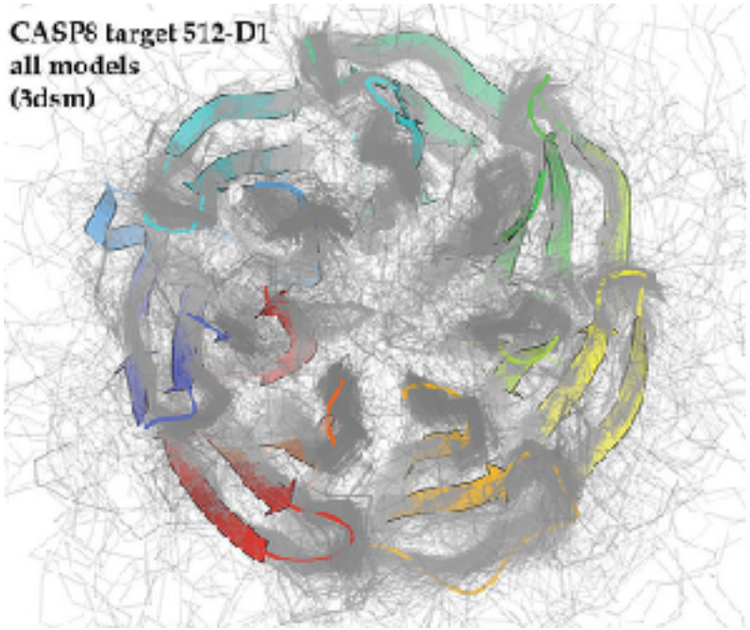
Internal symmetry



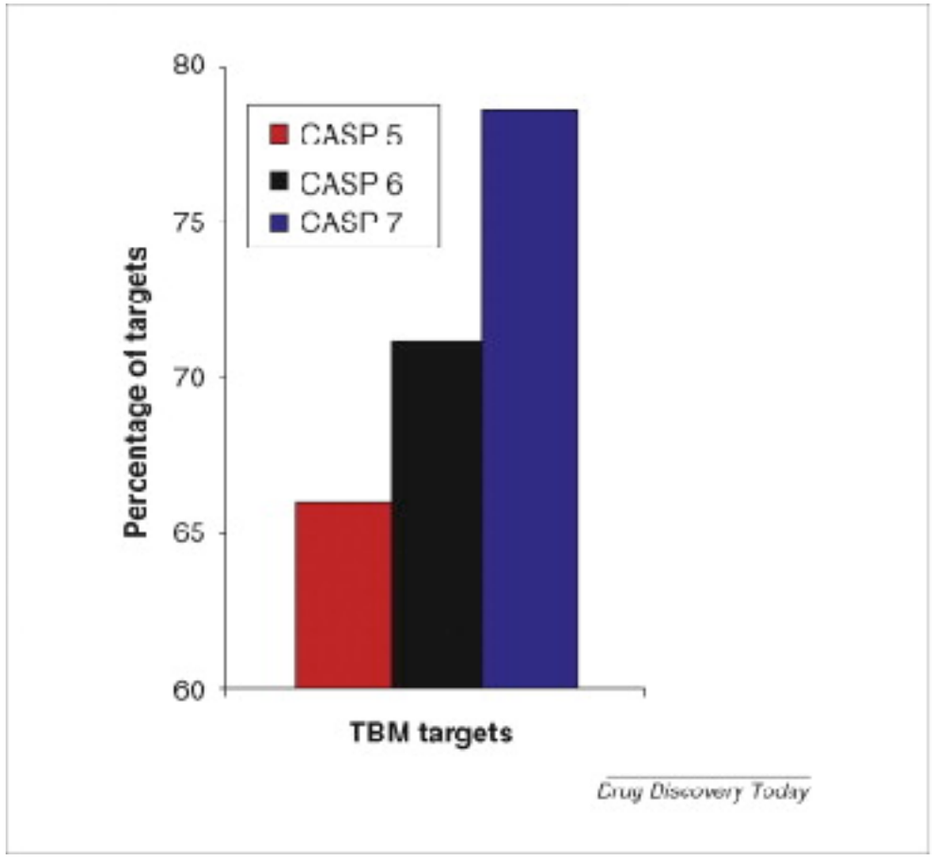
Secondary structure

More Prior Information

- As the number of observations decreases we need to increase the amount of prior information we include (or the number of constraints we apply)
- At the extreme - what if we had no data?
- Other fields have been trying to address this problem:
  - Structure prediction
  - Homology modelling
  - Protein folding



http://www.predictioncenter.org



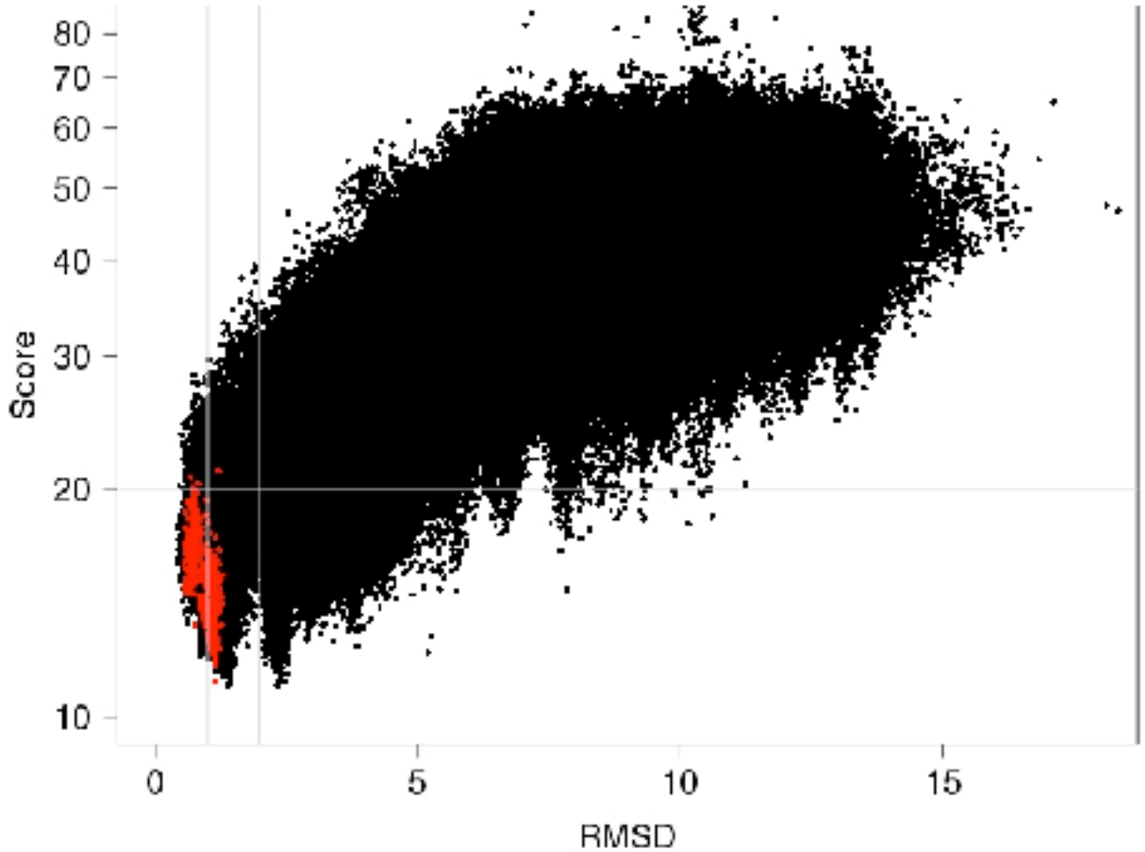
From: Kryshtafovych & Fidelis, Drug Discovery Today, 2009, 14:386–393

Physically Realistic Potentials (Rosetta)

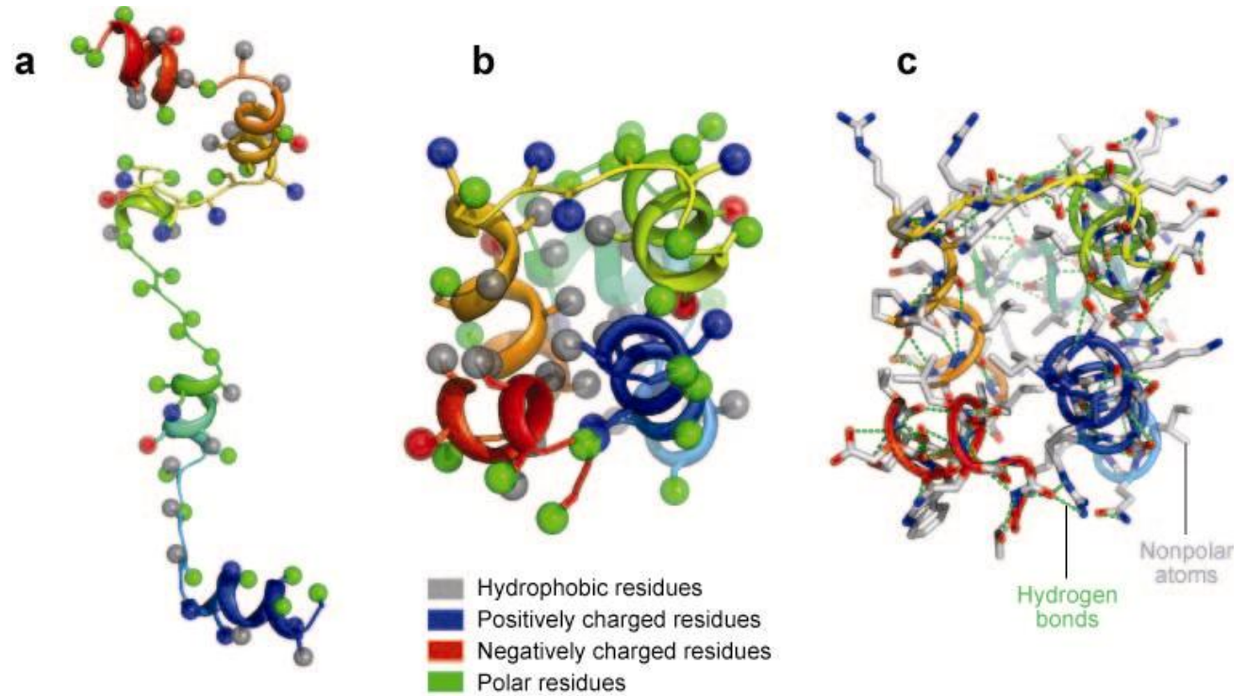
(Nat Echols & Frank DiMaio)

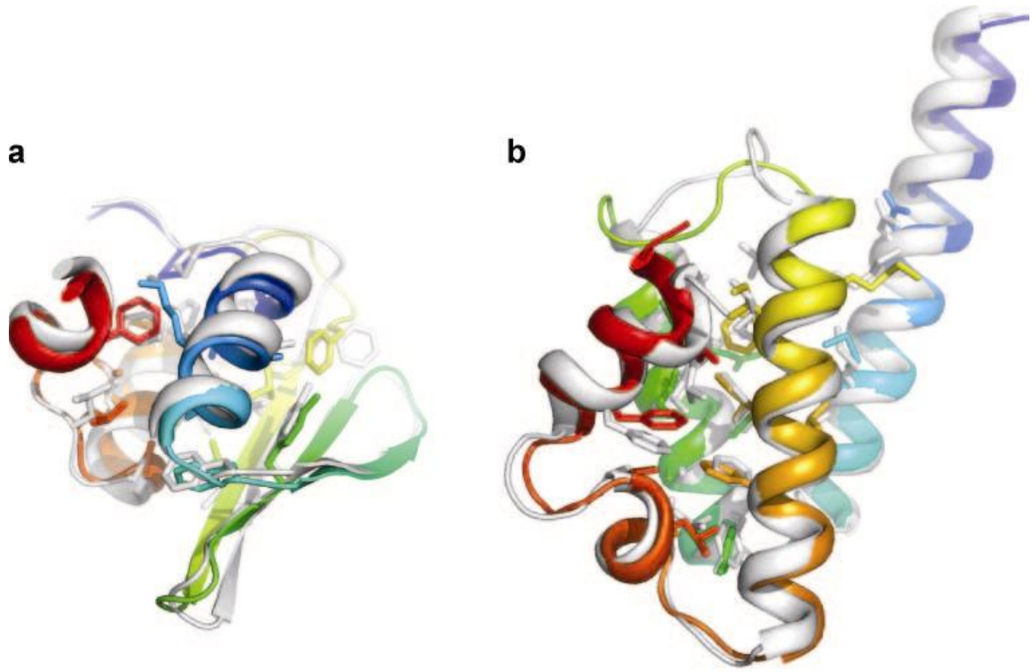
Rosetta

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



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





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

Why Rosetta

- Designed to recognize near-native structures among many possible models; combines empirical and physical potentials
  - All-atom force field, incorporates solvation effects, attractive forces, hydrogen bonds, knowledge-based dihedral restraints
- Can yield chemically realistic *ab initio* models without experimental data to guide assembly
  - Occasionally good enough for molecular replacement
- Shown to be useful for NMR structure determination with sparse data (CS-Rosetta), MR solution improvement (MR-Rosetta), RNA structure refinement (ERRASER)

Kuhlman et al. (2003) *Science* **302**:1364-8
Rohl et al. (2004) *Methods Enzymol.* **383**:66-93
Keedy et al. (2009) *Proteins* **77**:29-49

https://www.rosettacommons.org

Complementary Algorithms

Phenix

- Reciprocal space X-ray target functions (ML, MLHL, LS-twin)
- Bulk solvent correction
- B-factor refinement (including TLS)
- Map calculation
- Density modification (using RESOLVE)

Rosetta

- Physically realistic potentials
- Repacking to remove steric clashes and building rotameric sidechains
- Torsion-angle minimization
- Real-space target (refinement against electron density)
- Fragment-based rebuilding (optional, not currently used)

Python/C++ architecture facilitates combination

Low Resolution Protocol

- Sidechain repacking (using density)
- Coordinate refinement (reciprocal space torsion angle minimization and reduced nonbonded penalty)
- B-factor refinement

3 Cycles

- Sidechain repacking (using density)
- Coordinate refinement (real space and reciprocal space torsion angle minimization)
- B-factor refinement

5 Cycles

- Sidechain repacking (using density)
- Coordinate refinement (reciprocal space minimization with restrained bonds and angles)
- B-factor refinement

2 Cycles

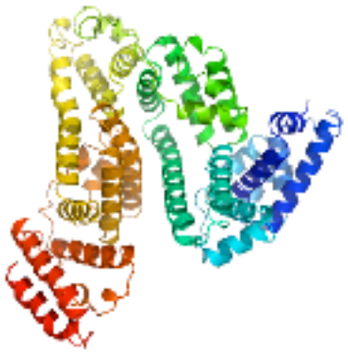
Protocol run 5 times in parallel and the best model selected based on R-free

BERKELEY LAB
Lawrence Berkeley National Laboratory

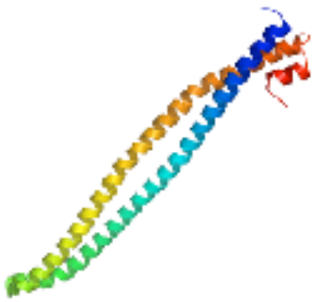
Test Cases

Solved by molecular replacement with same protein from another deposition at higher resolution

2j5f (3.0Å)

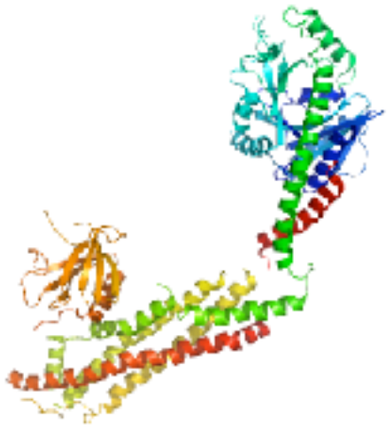


1bke (3.15Å)

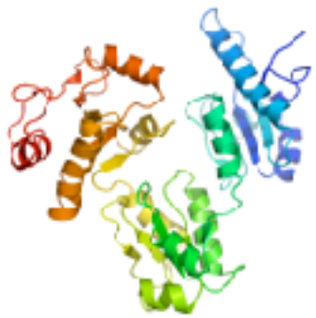


3mtt (3.3Å)

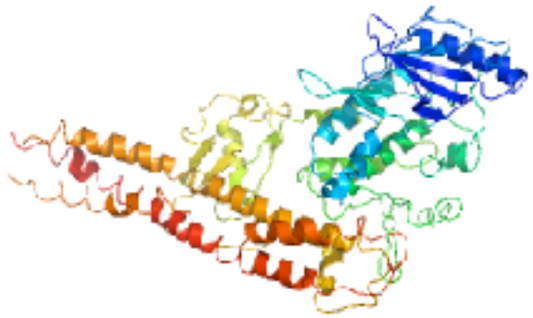
1kct (3.46Å)



3snh (3.7Å)

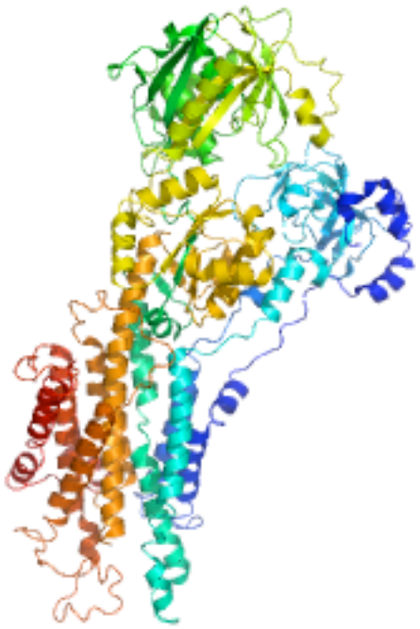


2vaf (3.8Å)

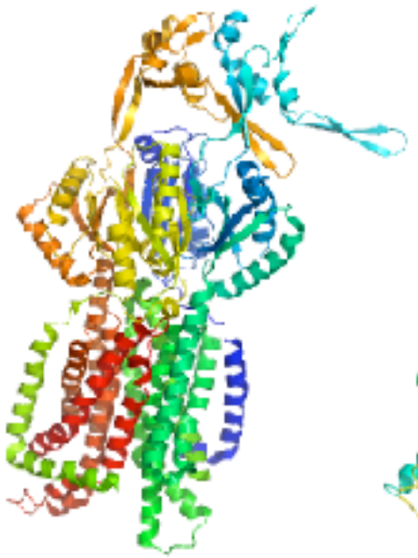


3rzf (4.0Å)

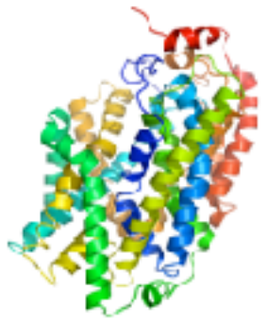
Membrane Proteins



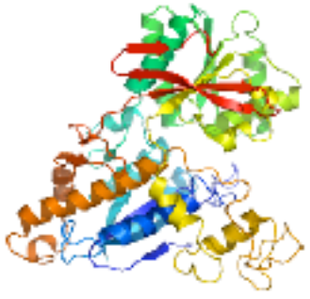
3fps (3.2Å)



3k07 (3.52Å)

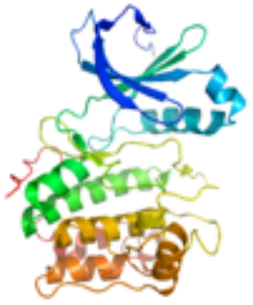


2x79 (3.8Å)

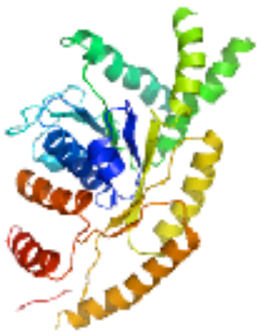


1isr (4.0Å)

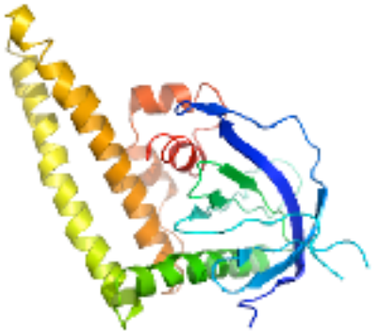
Solved using homologous proteins



3pwy (3.5Å)



3idq (3.7Å)

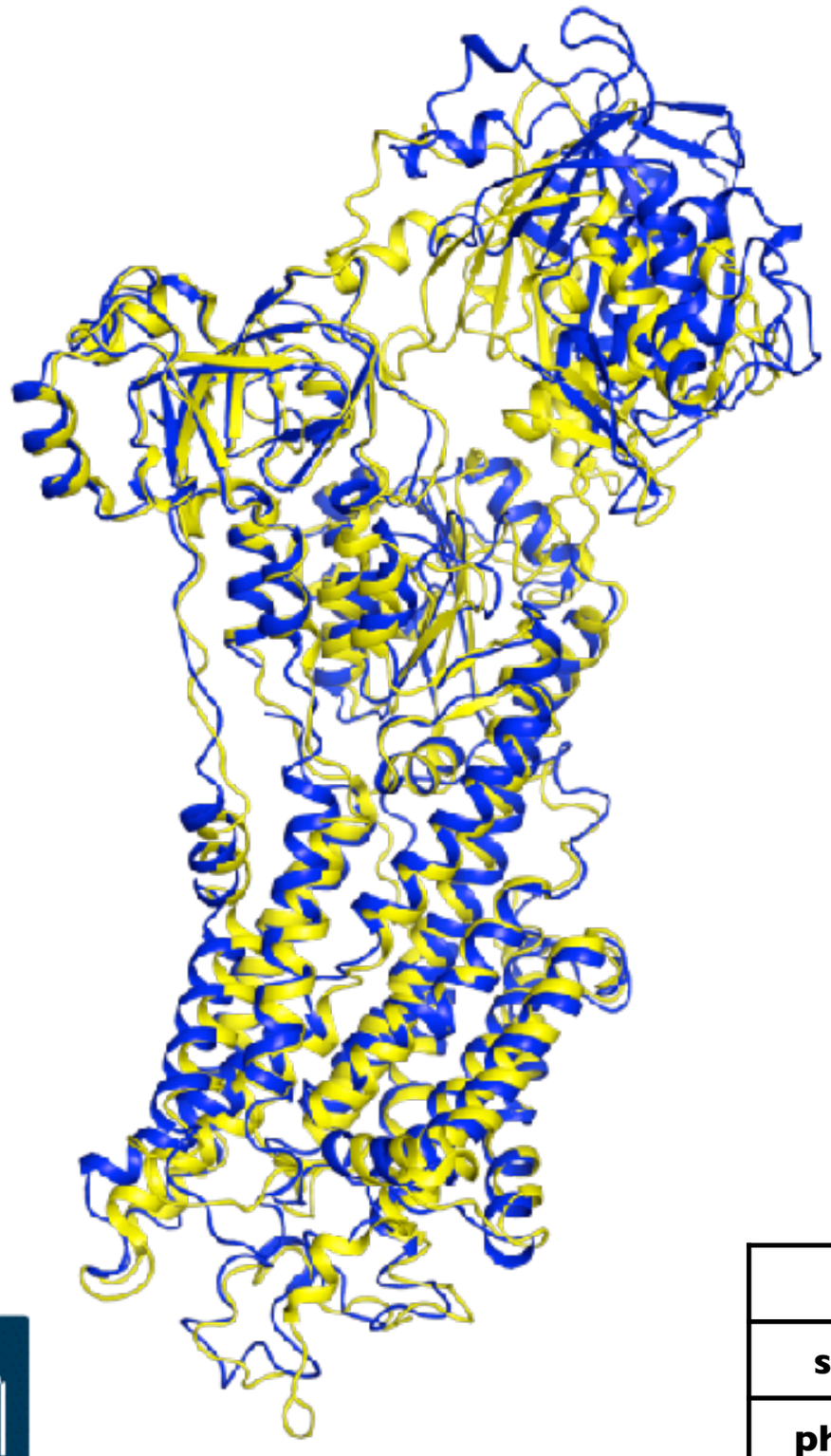


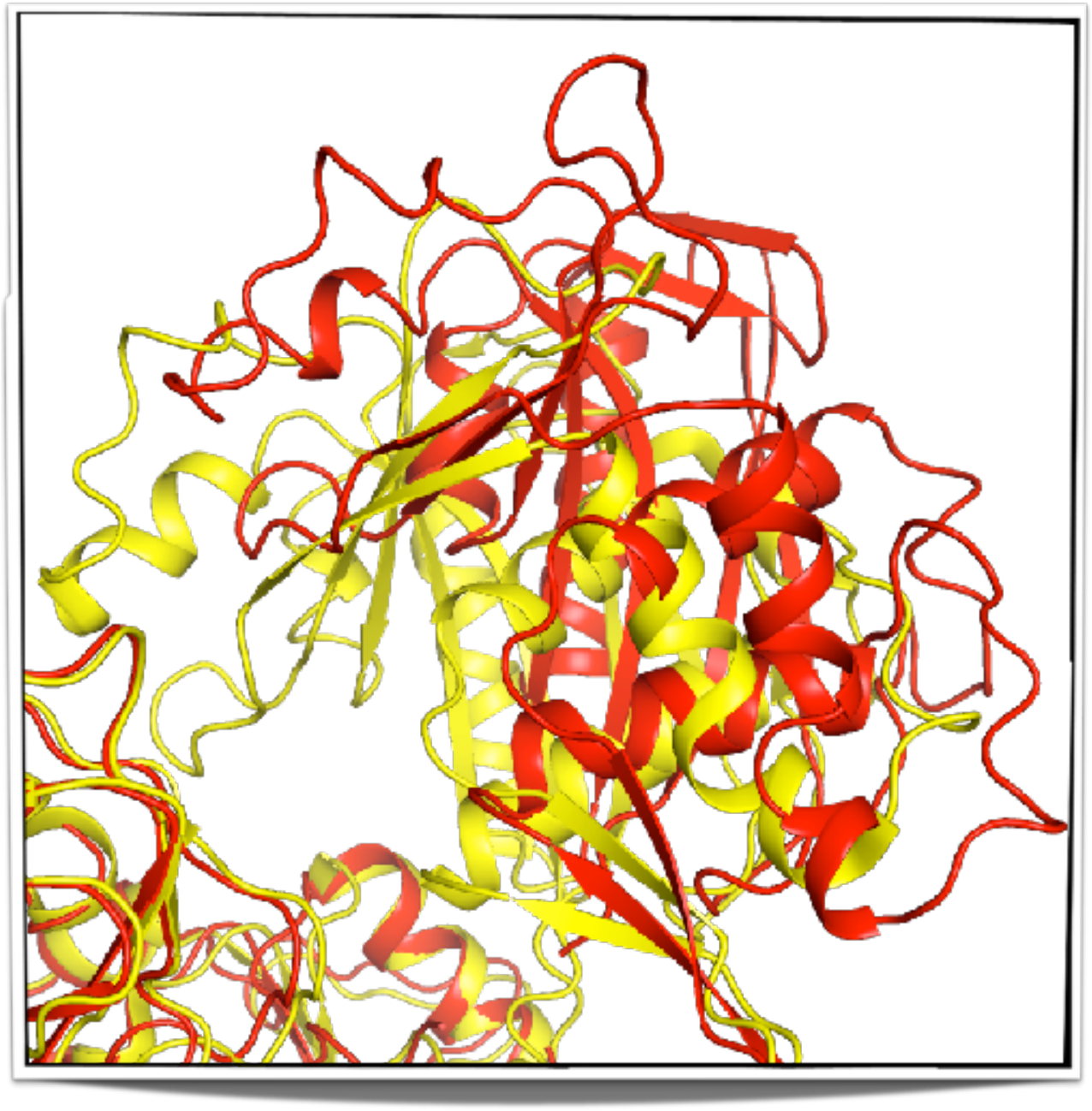
3a8n (4.5Å)

BERKELEY LAB
Lawrence Berkeley National Laboratory

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868

Calcium ATPase - phenix.refine

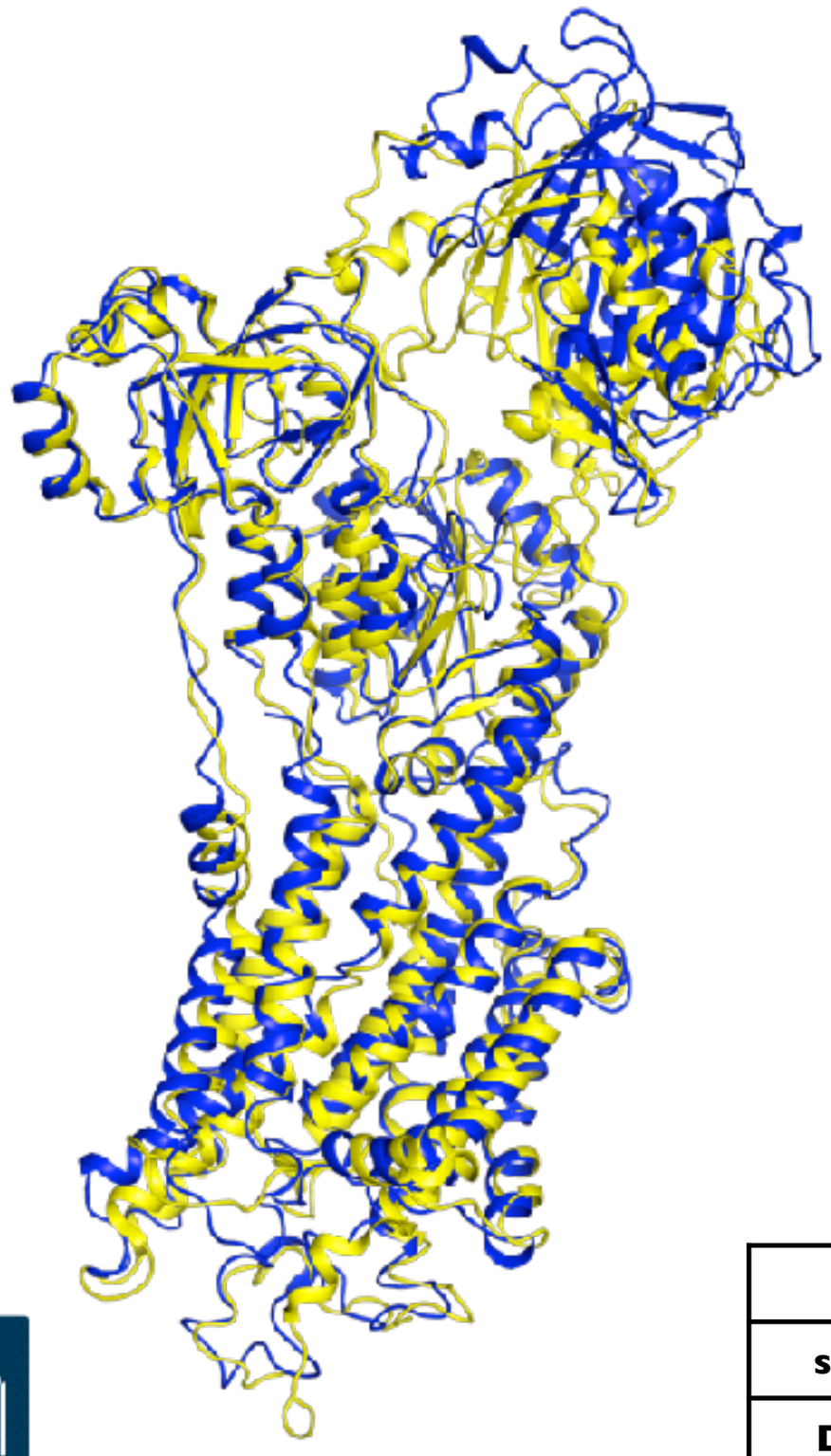


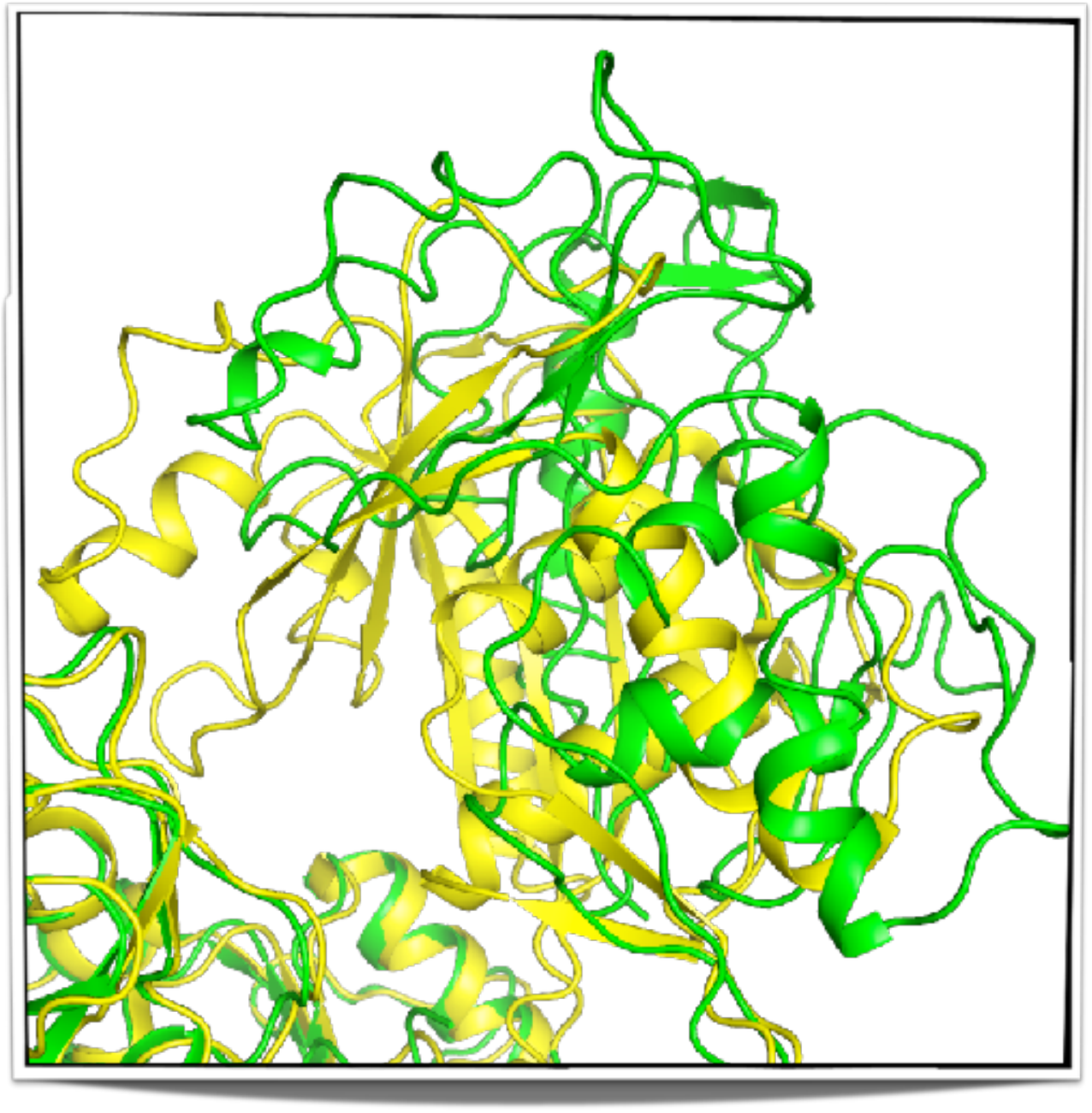


	R	R-free	mp score	RMSD
start	0.47	0.51	3.21	6.1
phenix	0.43	0.48	2.66	6.2

BERKELEY LAB
Lawrence Berkeley National Laboratory

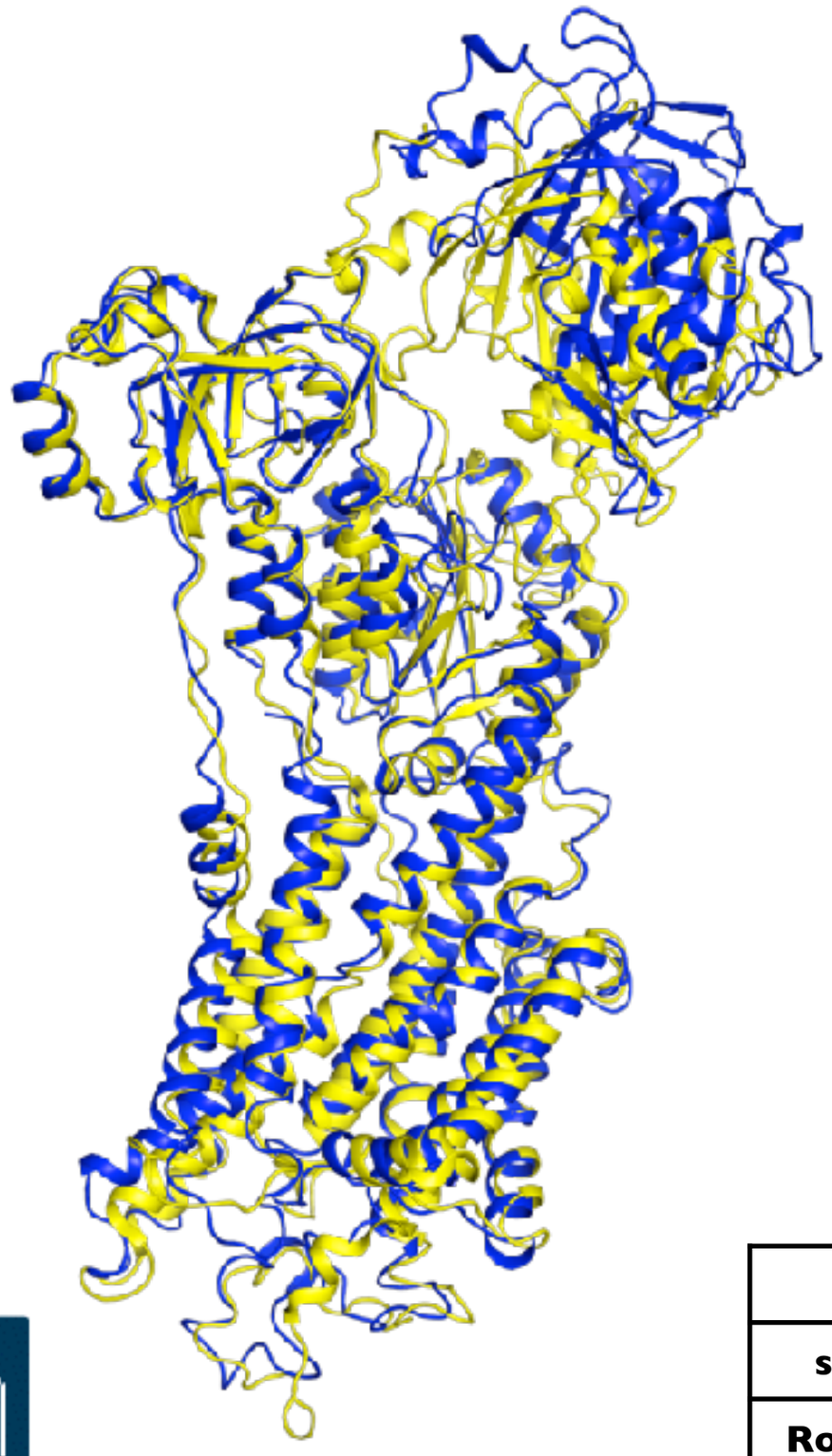
Calcium ATPase - DEN

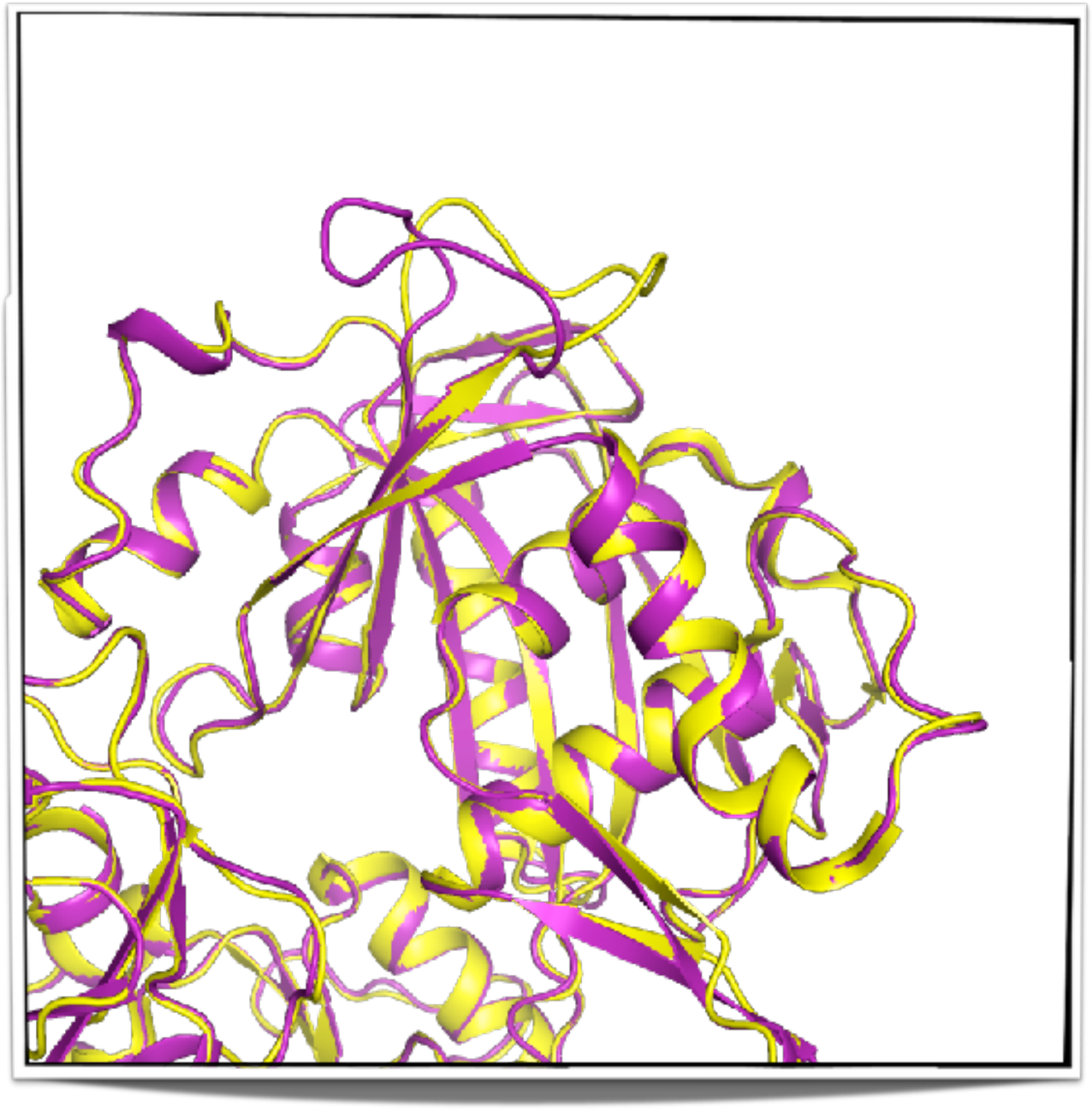




	R	R-free	mp score	RMSD
start	0.47	0.51	3.21	6.1
DEN	0.38	0.44	3.79	6.1

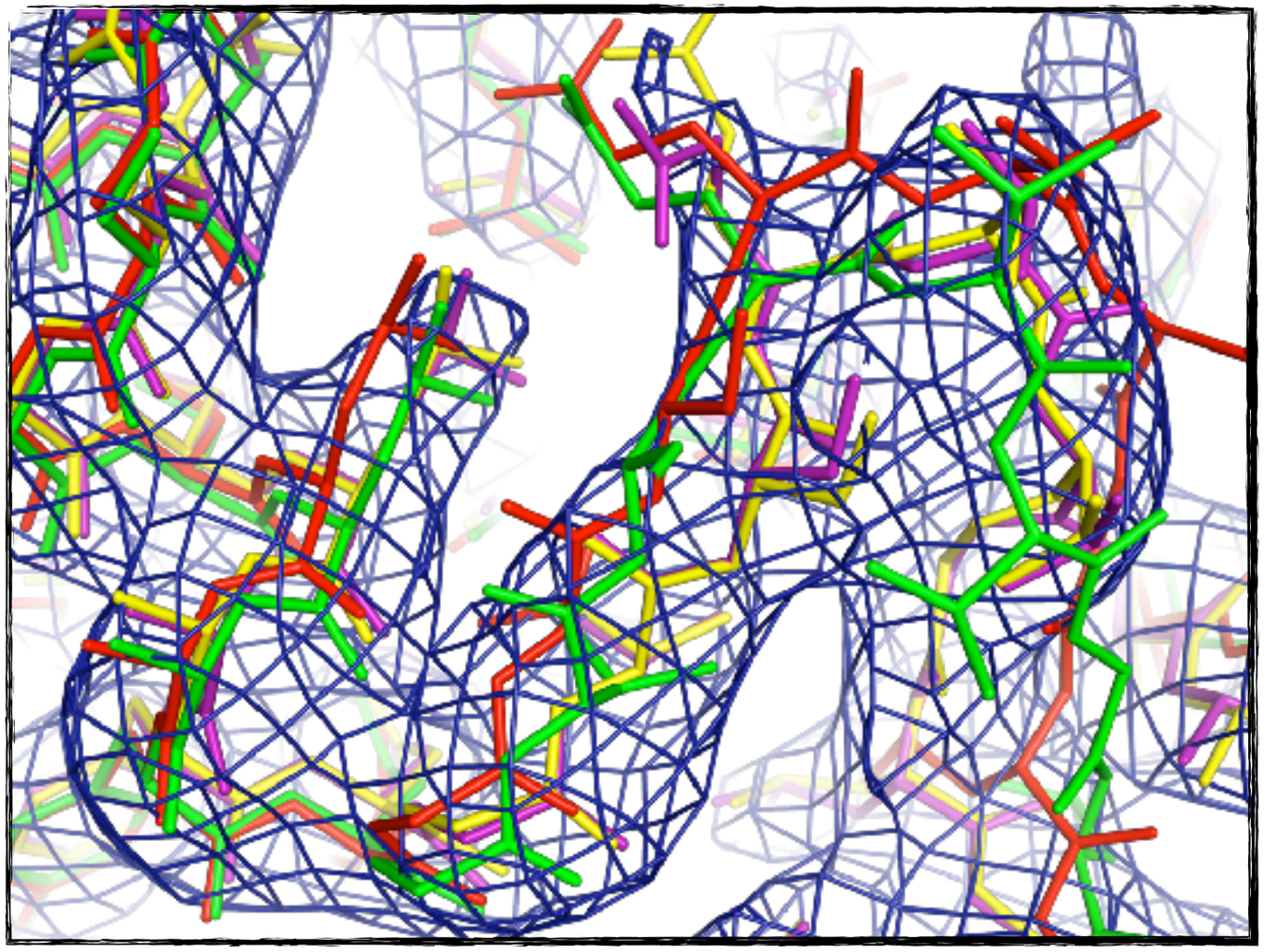
Calcium ATPase - Phenix-Rosetta

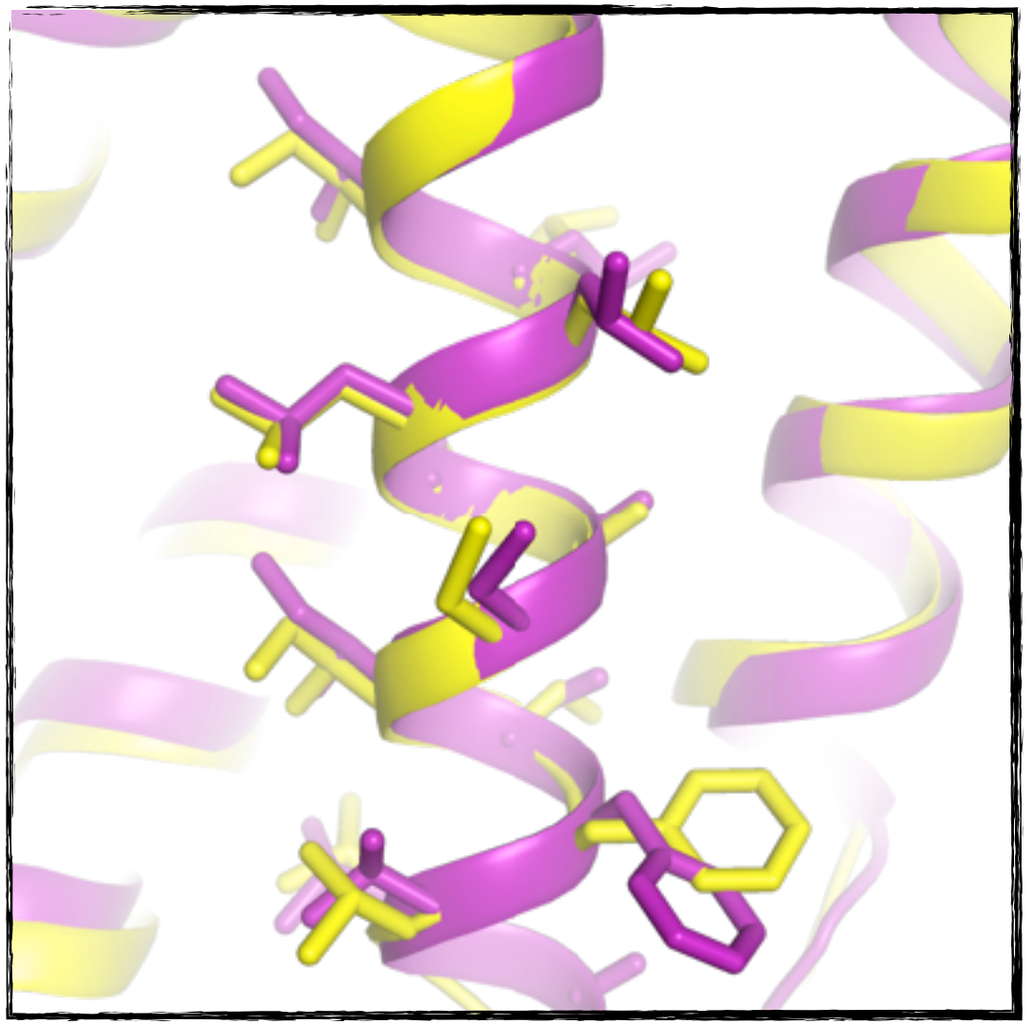




	R	R-free	mp score	RMSD
start	0.47	0.51	3.21	6.1
Rosetta	0.24	0.28	1.55	1.7

Calcium ATPase - Detail



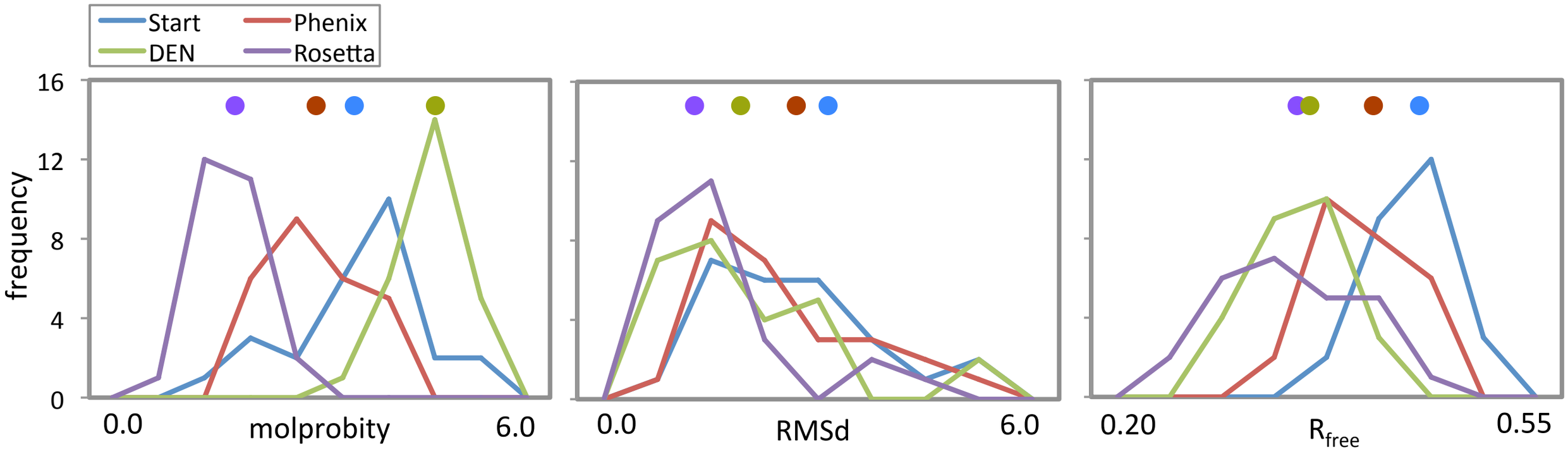


- Phenix-Rosetta model is very close to the deposited structure (even at the level of side chains) with better fit to density

BERKELEY LAB
Lawrence Berkeley National Laboratory

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

Improved Models



- Phenix-Rosetta typically has improved fit to the crystallographic data and models are closer to the known structure
- Phenix-Rosetta always has improved model quality, as judged by Molprobity
- Generally similar to DEN results but with much improved geometry, and generally faster

DiMiao et al., 2013, *Nature Methods* ***10***:*1102-1104*

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
Lawrence Berkeley National Laboratory

THE UNIVERSITY OF CALIFORNIA · BERKELEY · 1868