New tools for improving electron cryo-microscopy maps and validating models

Cryo-EM Validation in the Age of SARS-CoV-2: Methods, Tools and Applications

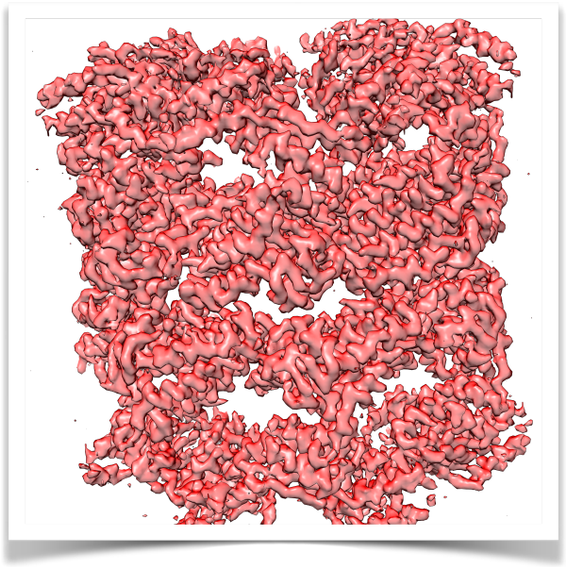
November 2020

Paul Adams

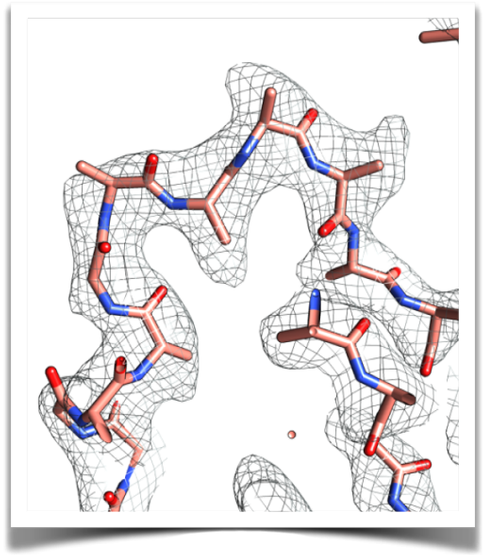
Lawrence Berkeley Laboratory and

Department of Bioengineering UC Berkeley

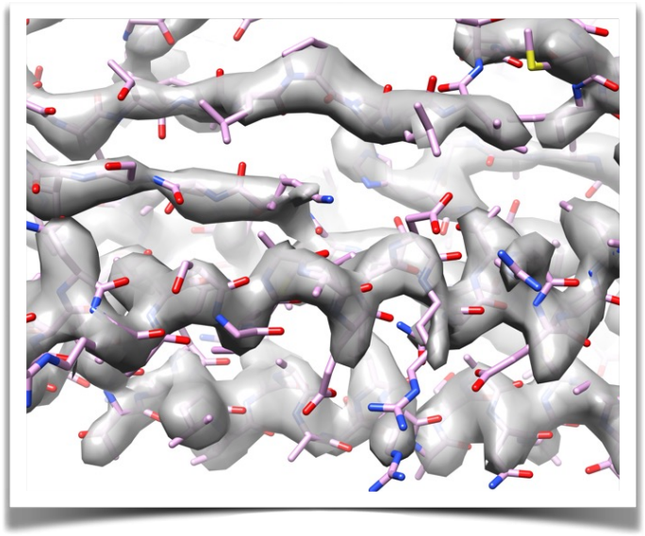
New Tools for Cryo-EM in Phenix



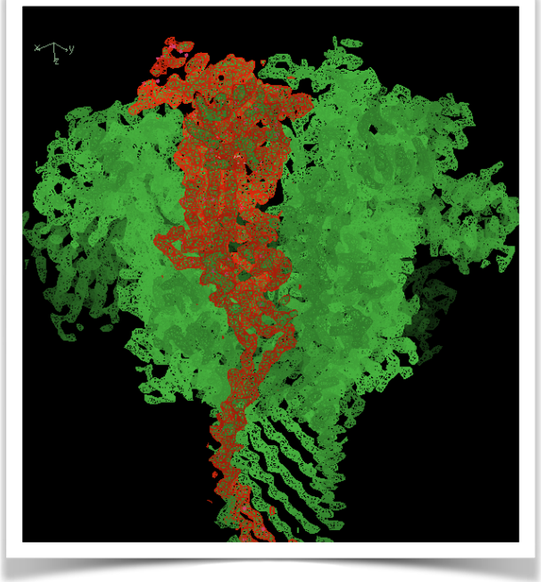
Symmetry from a map



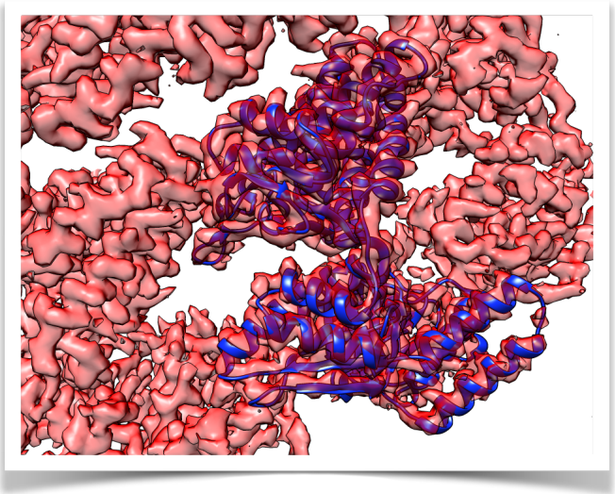
Automated map sharpening



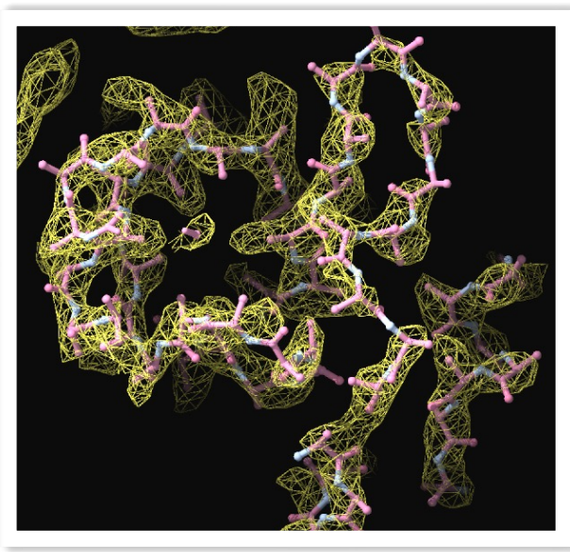
Density modification



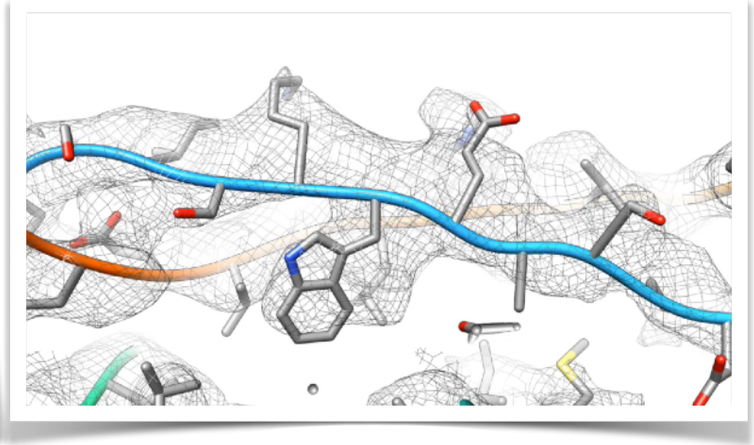
Map segmentation



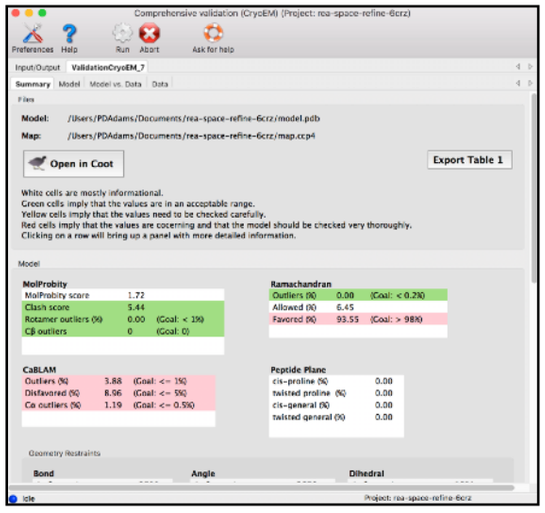
Rigid model docking



Automated model building



Real space refinement



Model and map validation

BERKELEY LAB
Lawrence Berkeley National Laboratory

THE UNIVERSITY OF CALIFORNIA · BERKELEY · 1868

The Phenix Project

Lawrence Berkeley Laboratory

Paul Adams, Pavel Afonine, Dorothee Liebschner, Nigel Moriarty, Billy Poon, Christopher Schlicksup, Oleg Sobolev

New Mexico Consortium
Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung

University of Cambridge

Randy Read, Airlie McCoy, Tristan Croll, Claudia Millán Nebot, Rob Oeffner, Massimo Sammito, Duncan Stockwell

Baylor College of Medicine

Matt Baker, Corey Hryc

Duke University

Jane & David Richardson, Chris Williams, Vincent Chen

An NIH/NIGMS funded Program Project

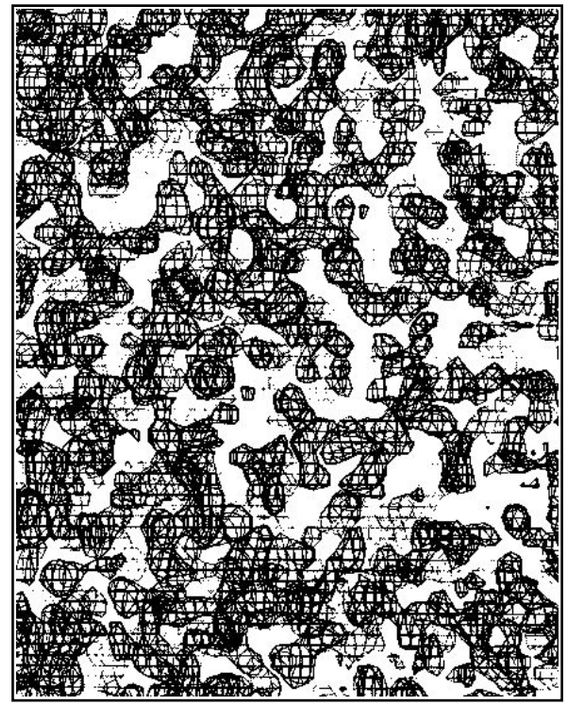
Liebschner et al., Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Cryst.* 2019 **D75**:861-877

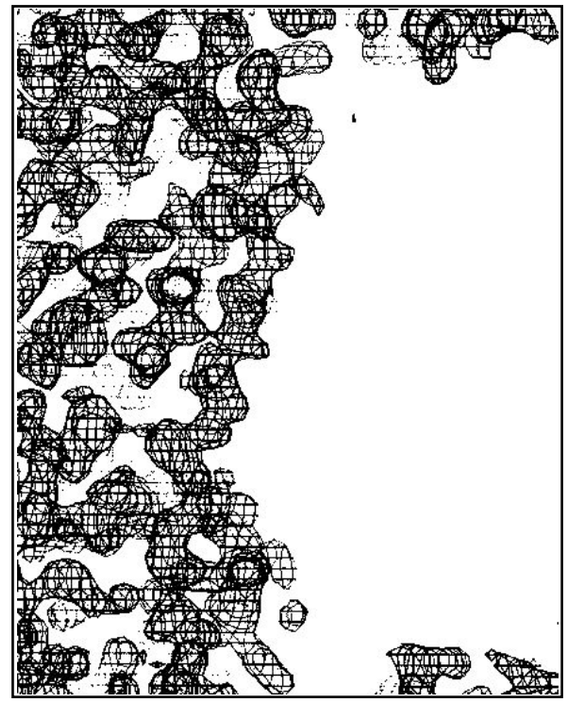
Map Improvement by Density Modification

Tom Terwilliger
Los Alamos National Laboratory
Steven Ludtke
Baylor College of Medicine
Randy Read
Cambridge University
Pavel Afonine
Lawrence Berkeley National Laboratory

Map Improvement

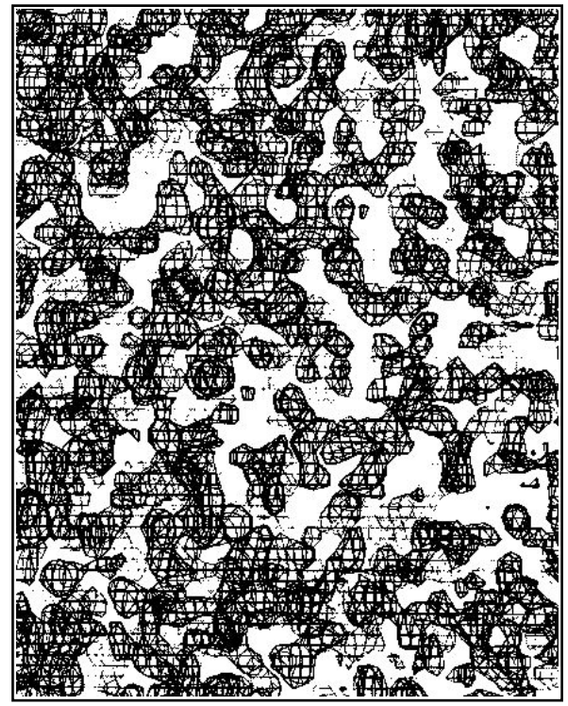
- Maps contain errors
- The maps can be improved by the application of real space constraints
- The Fourier coefficients are modified to produce a map most consistent with what we know about macromolecular structures:
  - Solvent density distribution (Solvent flattening)
  - Atomicity and positivity (Sayre's equation)
  - Macromolecular density distributions (histogram matching)
  - Similarity between molecules (symmetry averaging)

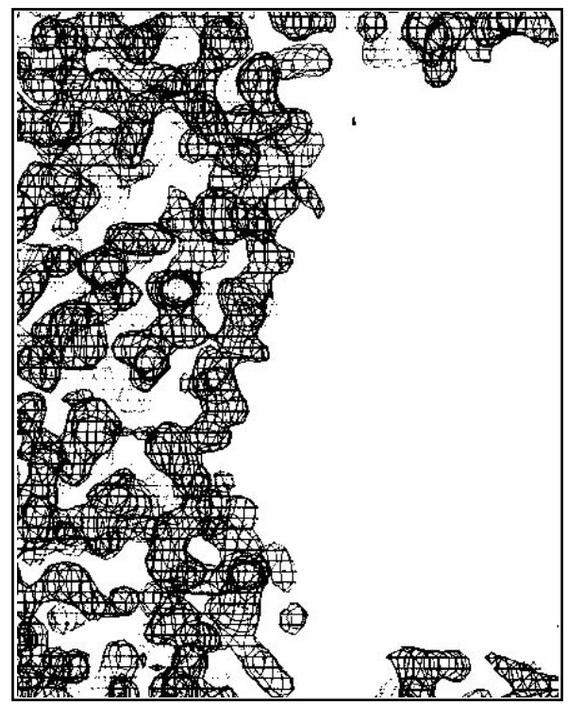




Statistical Phase Improvement

- Principle: phase probability information from probability of the map and from experiment:
  - $P(\varphi) = P_{\text{map probability}}(\varphi)\, P_{\text{experiment}}(\varphi)$
- 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
  - Symmetry (if appropriate)
  - A distribution of densities like those of model proteins
- Method:
  - calculate how map probability varies with the map ρ
  - deduce how map probability varies with phase φ
  - change map to maximize probability
  - combine with original map





BERKELEY LAB
Lawrence Berkeley National Laboratory

THE UNIVERSITY OF CALIFORNIA · BERKELEY · 1868

Overview of the Cryo-EM Procedure

Original half-maps → Averaged original half-maps

Maximum likelihood modification of Fourier coefficients

+

Map-phasing half-maps → Averaged map-phasing half-maps

Weighted combination using error model

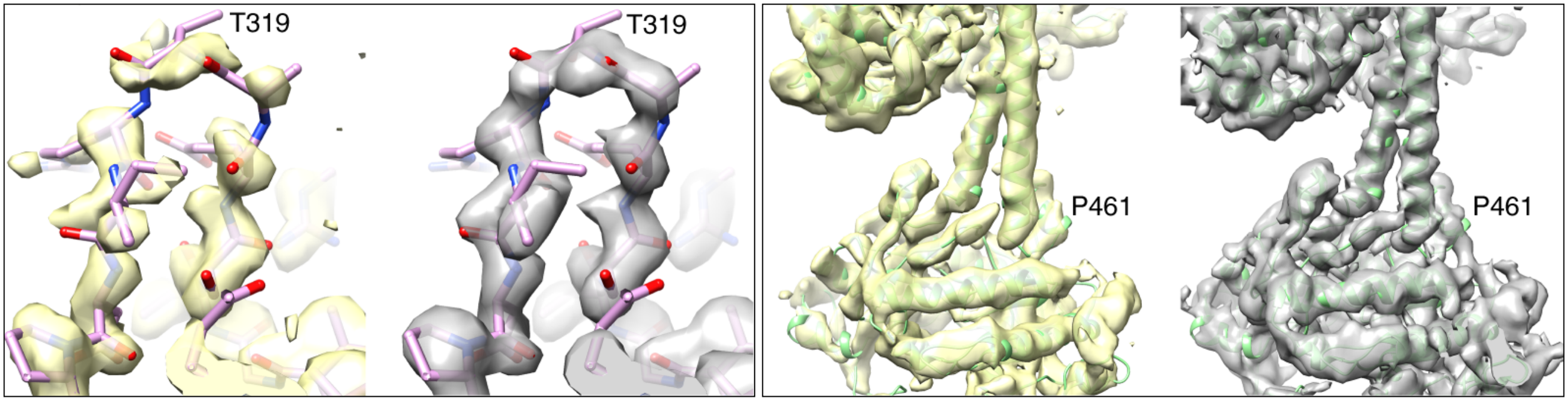
Final density modified full map

Terwilliger et al: Improvement of cryo-EM maps by density modification. *Nature Methods* 2020, **9**:923-927

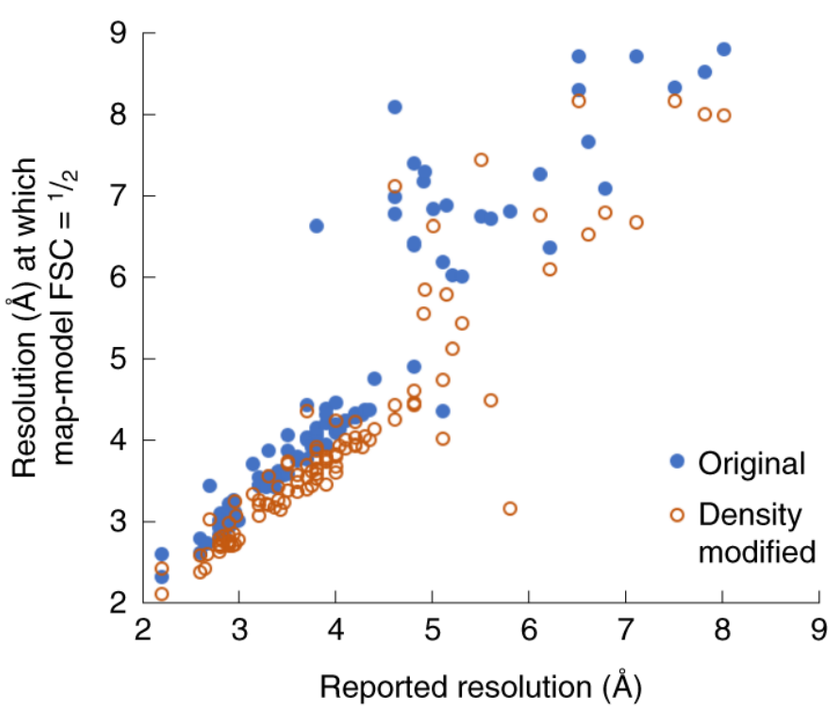
BERKELEY LAB
Lawrence Berkeley National Laboratory

Improved Maps

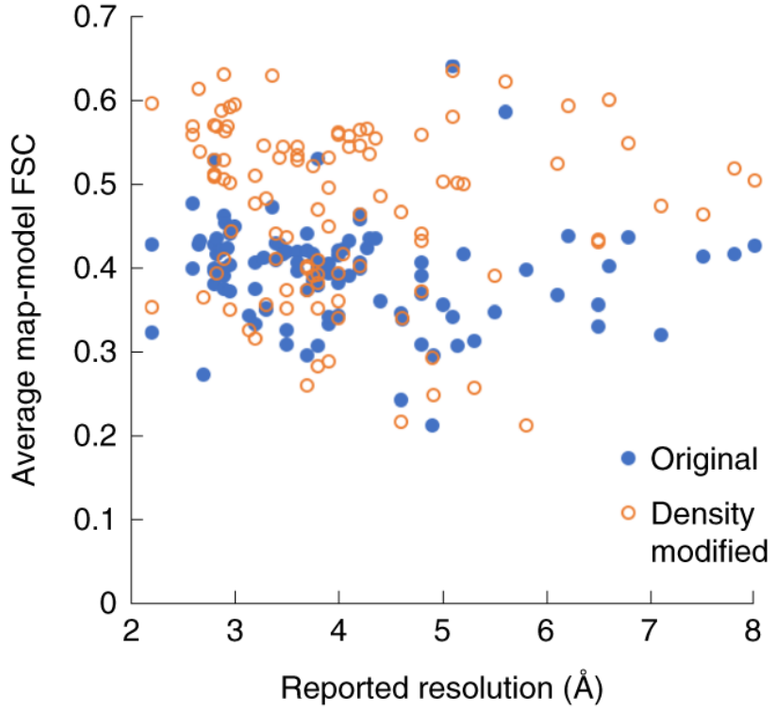
Original *Density Modified* *Original* *Density Modified*



β-galactosidase (2.2 Å, EMDB 2984)



Guanylate cyclase at 5.8 Å (EMDB 20282)



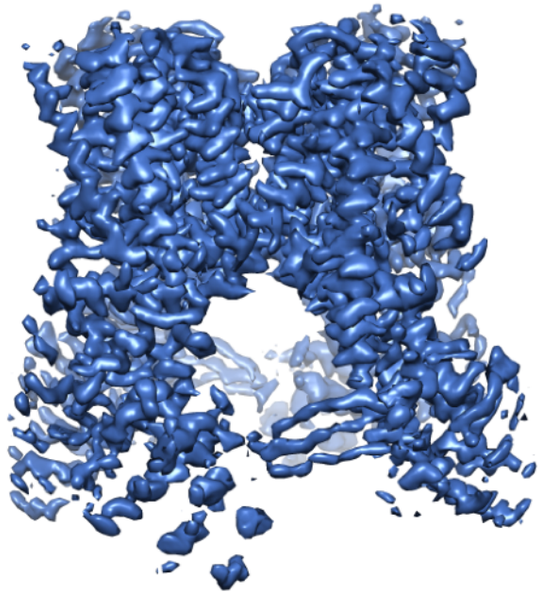
BERKELEY LAB
Lawrence Berkeley National Laboratory

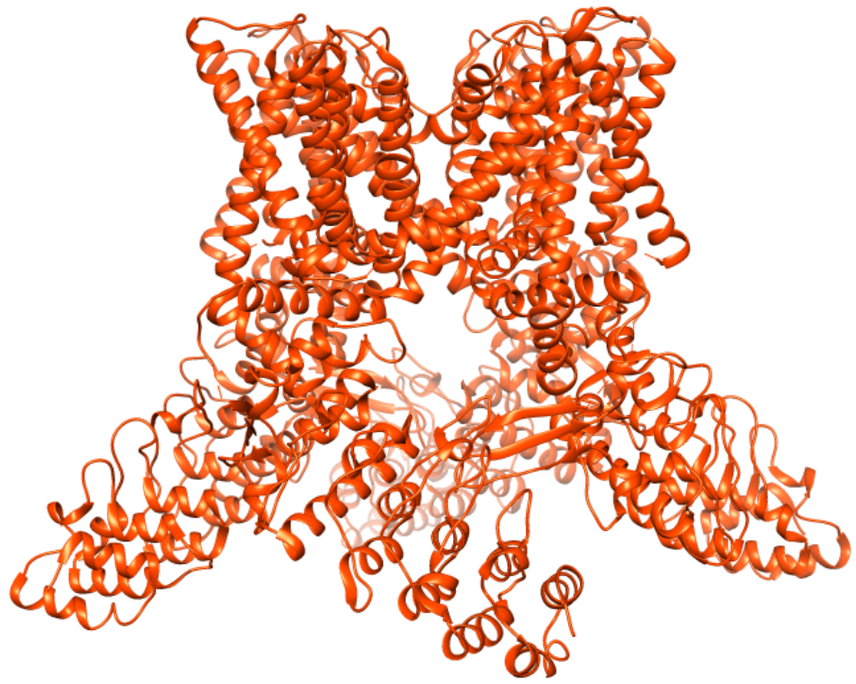
Validation

Christopher Williams, Jane Richardson, David Richardson
Duke University
Pavel Afonine, Oleg Sobolev, Nigel Moriarty
Lawrence Berkeley National Laboratory
Maarten Hekkelman, Robbie Joosten, Tassos Perrakis
Netherlands Cancer Institute

Validation and Cryo-EM

- Does the map make sense?
  - Gold Standard FSC of half maps
- Does the model make sense?
  - MolProbity
- Does the model fit the map?
  - Overall and local correlation





BERKELEY LAB
Lawrence Berkeley National Laboratory

Map Resolution and Map/Model Fit

Summary of map resolution estimates.

Metric	Objects used	Purpose	Values	Meaning, possible actions
d_{FSC}	Half-maps	Highest resolution at which the experimental data are confident	The higher the better	Resolution determined using half-maps method
d_{99}	Map	Resolution cutoff beyond which Fourier coefficients are negligibly small	$d_{99} \geq d_{FSC}$	Expected values
			$d_{99} < d_{FSC}$	Verify d_{FSC}; omit coefficients with $d_{99} \leq d < d_{FSC}$
			$d_{99} >> d_{FSC}$	Sharpen the map
d_{model}	Map and model	Resolution cutoff at which the model map is the most similar to the target map	$d_{model} \geq d_{FSC}$	Expected values
			$d_{model} < d_{FSC}$	Verify d_{FSC}; check ADP (too large?); validate map details
			$d_{model} >> d_{FSC}$	Sharpen the map
			$d_{model} << d_{99}$	Check ADP (too large?)
			$d_{model} >> d_{99}$	Check ADP (too small?); check the model
d_{FSC_model}	Map and model	Resolution cutoff up to which the model and map Fourier coefficients are similar	$d_{FSC\_model} \geq d_{FSC}$	Expected values
			$d_{FSC\_model} < d_{FSC}$	Verify d_{FSC}; omit coefficients with $d_{FSC\_model} \leq d < d_{FSC}$
			$d_{FSC\_model} \geq d_{FSC}$	Sharpen the map
			$d_{FSC\_model} >> d_{model}$	Omit coefficients with $d_{model} \leq d < d_{FSC\_model}$
			$d_{FSC\_model} << d_{model}$	Sharpen the map

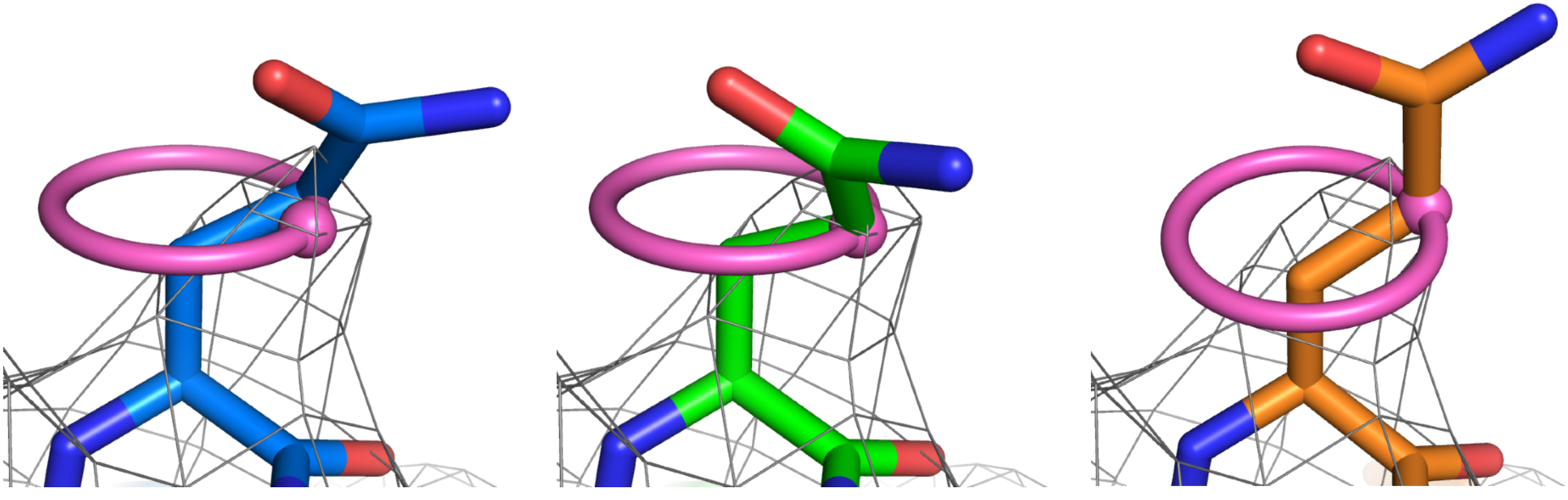
Summary of map correlation coefficients used in this work.

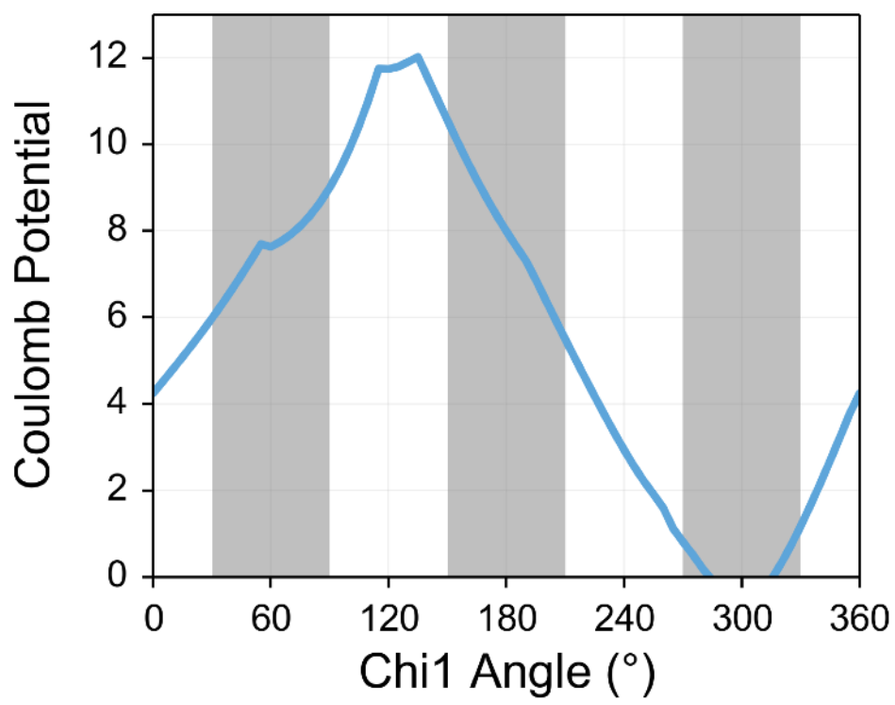
Metric	Region of the map used in calculation	Purpose
CC_{box}	Whole map	Similarity of maps
CC_{mask}	Jiang & Brünger (1994) mask with a fixed radius	Fit of the atomic centers
CC_{volume}	Mask of points with the highest values in the model map	Fit of the molecular envelope defined by the model map
CC_{peaks}	Mask of points with the highest values in the model and in the target maps	Fit of the strongest peaks in the model and target maps
CC_{vr_mask}	Same as CC_{mask} but atomic radii are variable and function of resolution, atom type and ADP	Fit of the atomic images in the given map

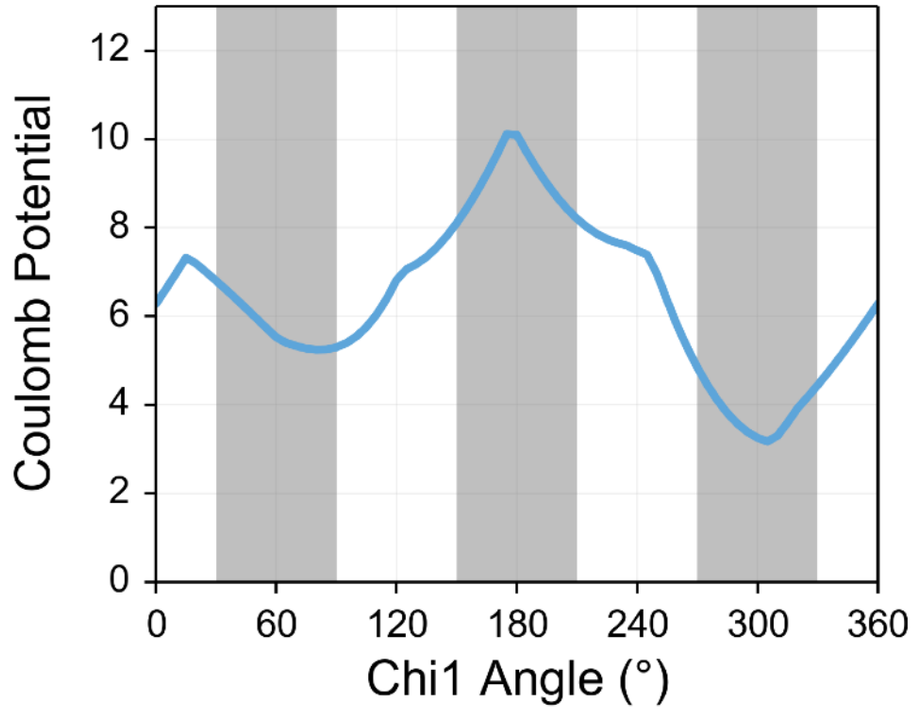
BERKELEY LAB
Lawrence Berkeley National Laboratory

Afonine et al: New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Cryst.* 2018, **D74**:814-840.

EMRinger reports on backbone placement







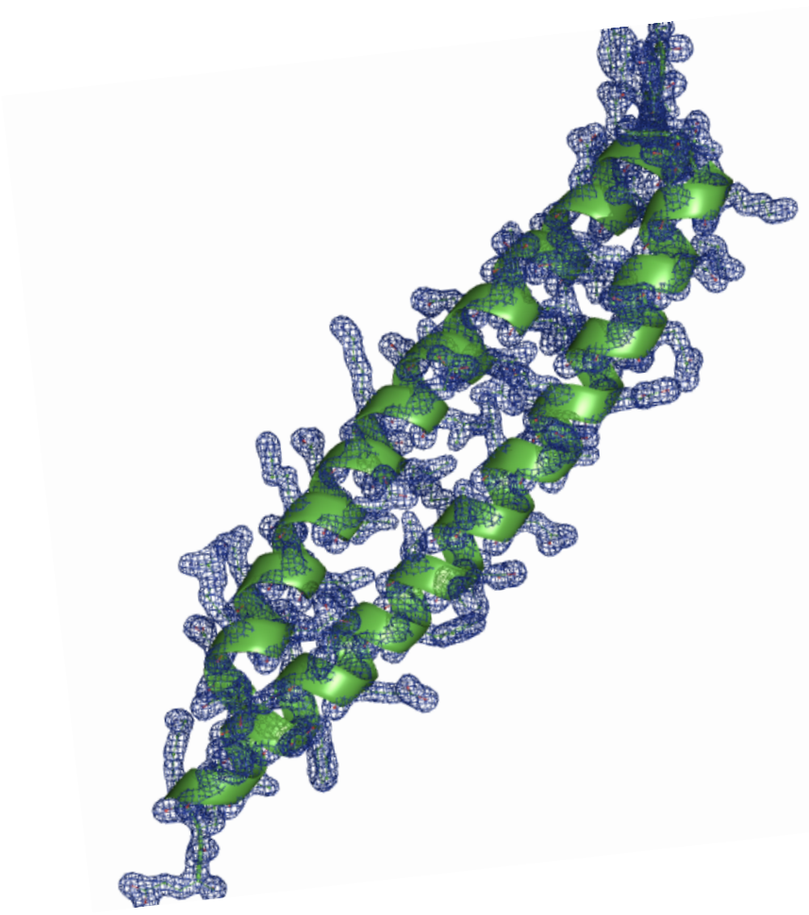
UCSF
University of California
San Francisco

Barad BA, et al. EMRinger: Side-chain-directed model and map validation for 3D Electron Cryomicroscopy. *Nature Methods* 2015, **12**:943-6

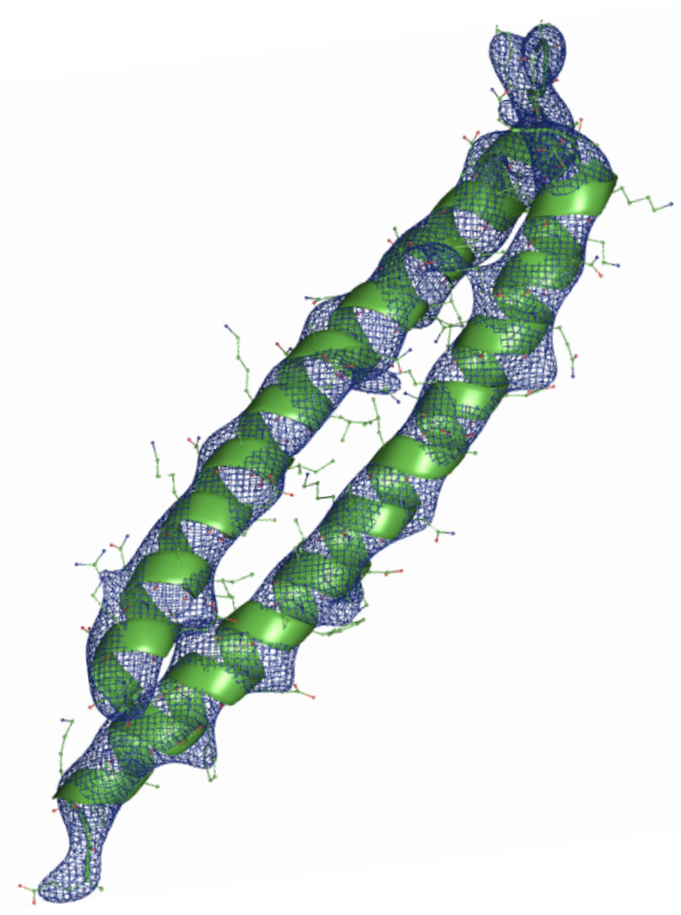
BERKELEY LAB
Lawrence Berkeley National Laboratory

Lower Resolution Requires Additional Information

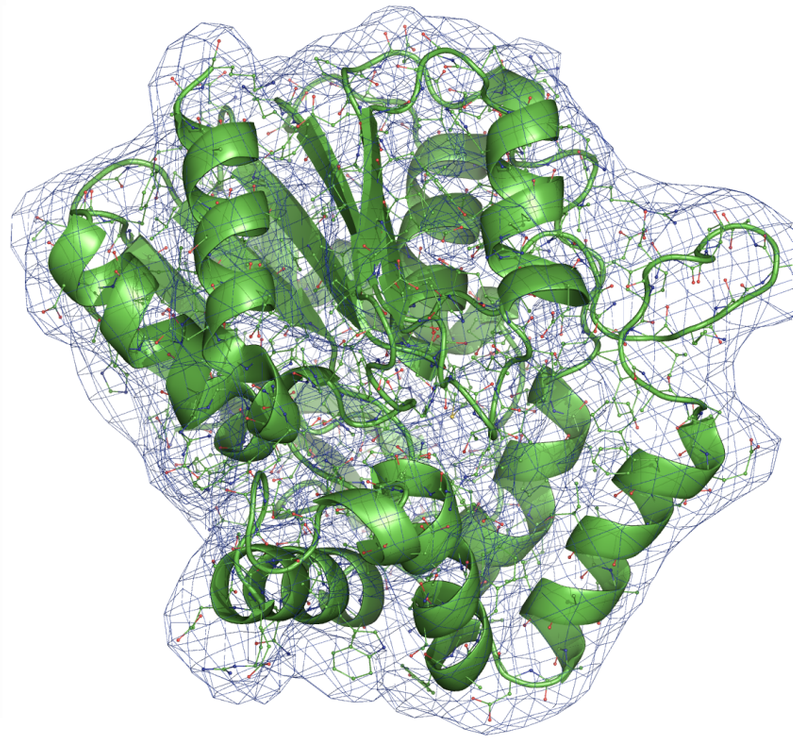
High Resolution **Low Resolution**



Side chains



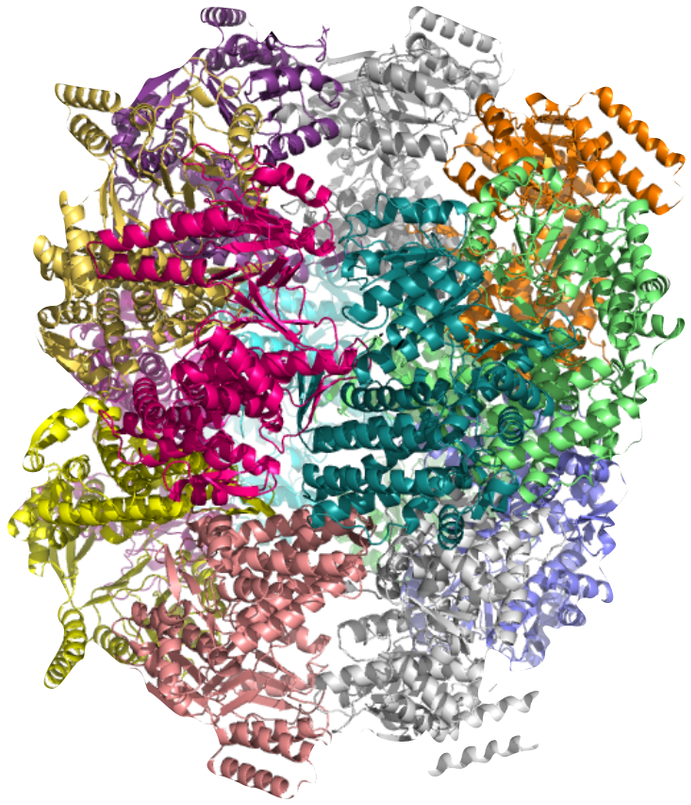
Secondary Structure



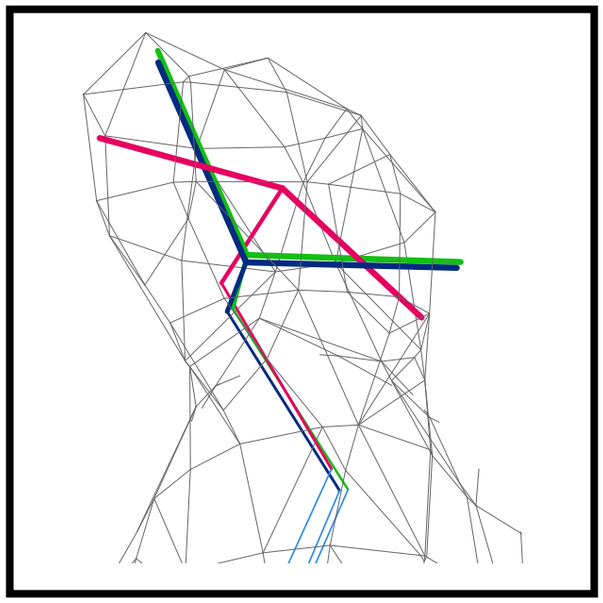
Molecule

BERKELEY LAB
Lawrence Berkeley National Laboratory

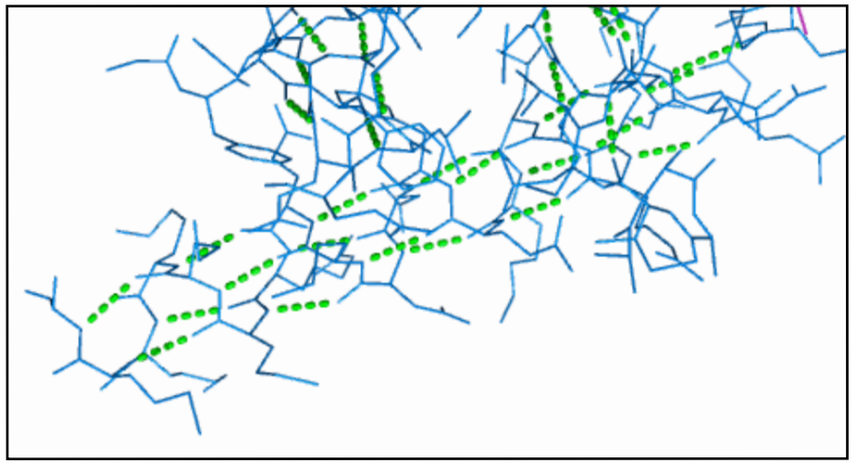
Additional Model Restraints



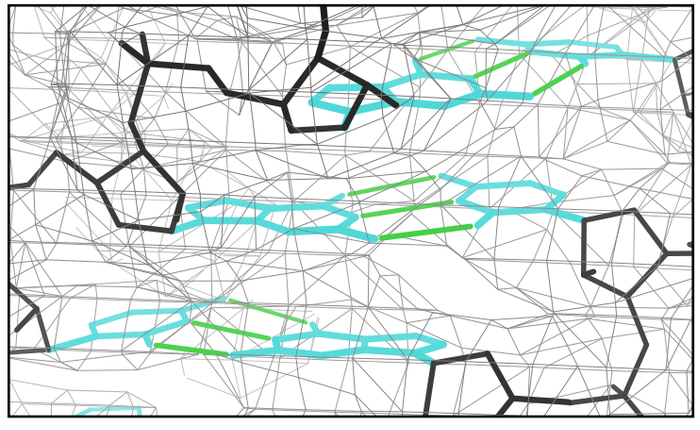
- Symmetry constraints
- Multiple symmetry groups
- Optimization of NCS operators (w.r.t density)
- Automatic expansion of monomer from symmetry records



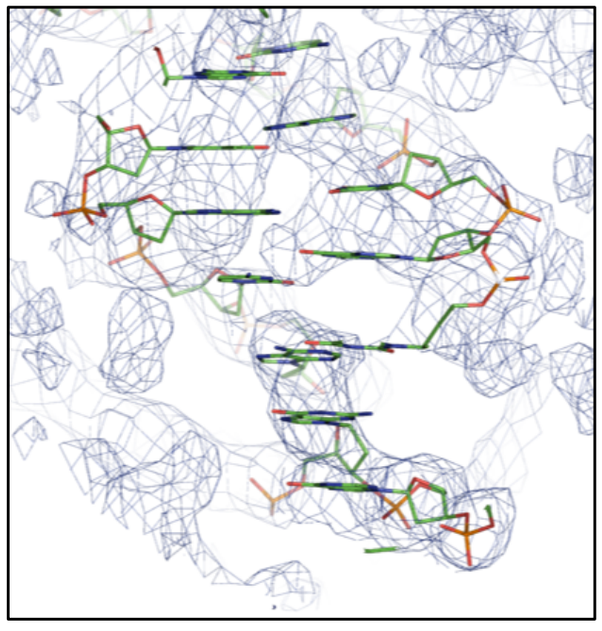
Reference model torsion angle restraints



Secondary structure restraints



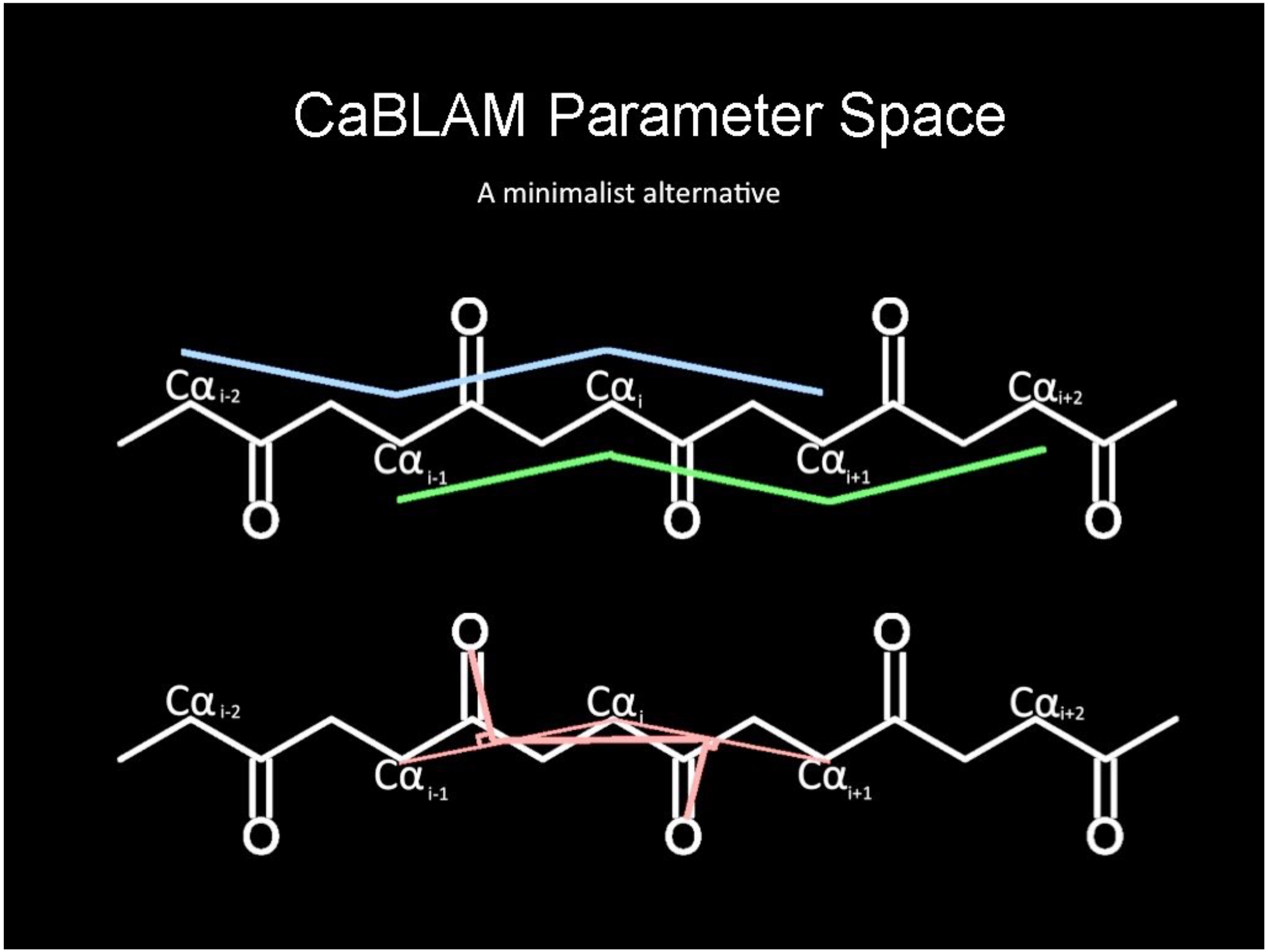
Base pairing restraints



Parallelity restraints

BERKELEY LAB
Lawrence Berkeley National Laboratory

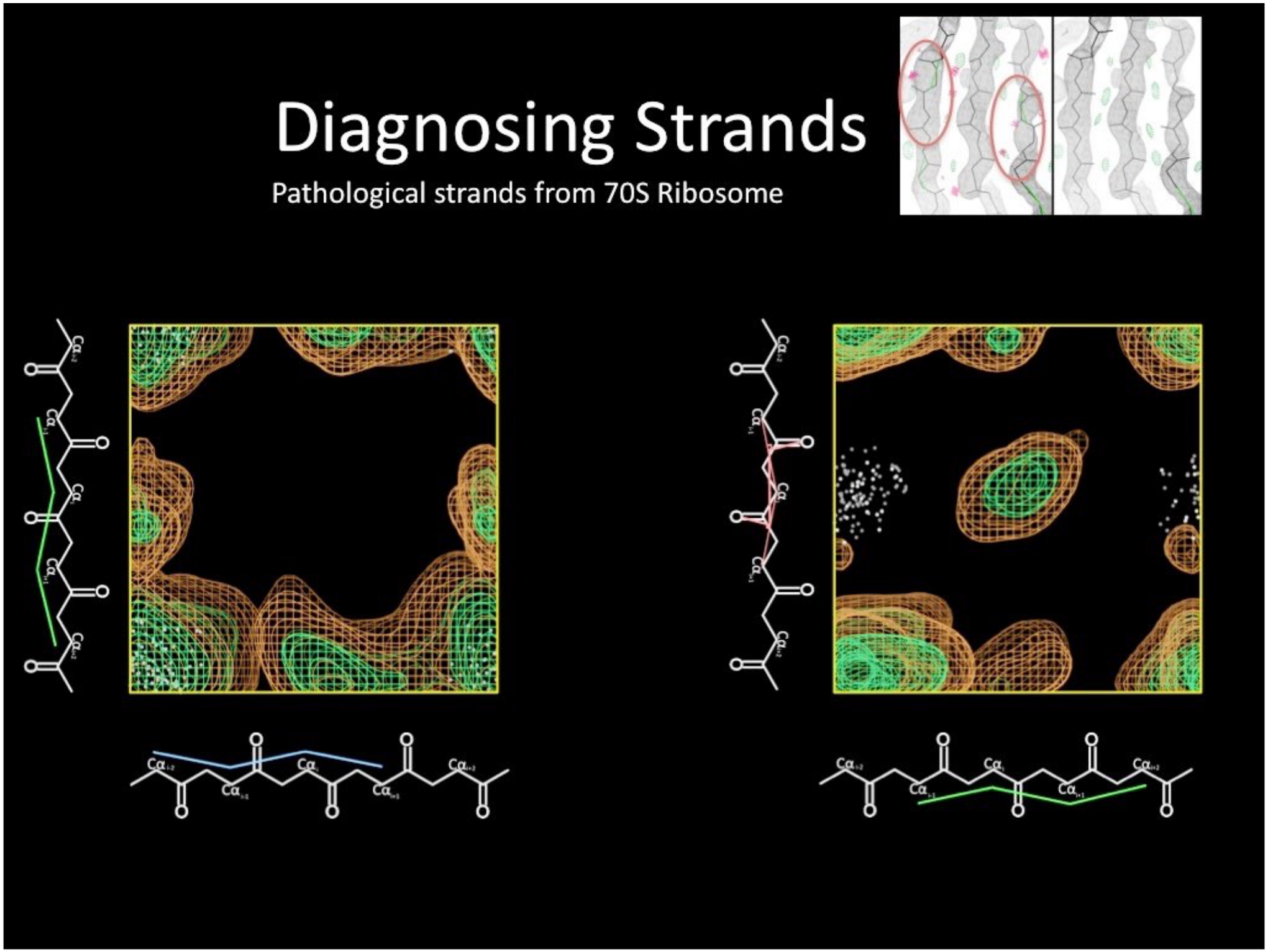
Validation Using Cα Atoms



Williams et al: MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* 2018, **27**:293-315

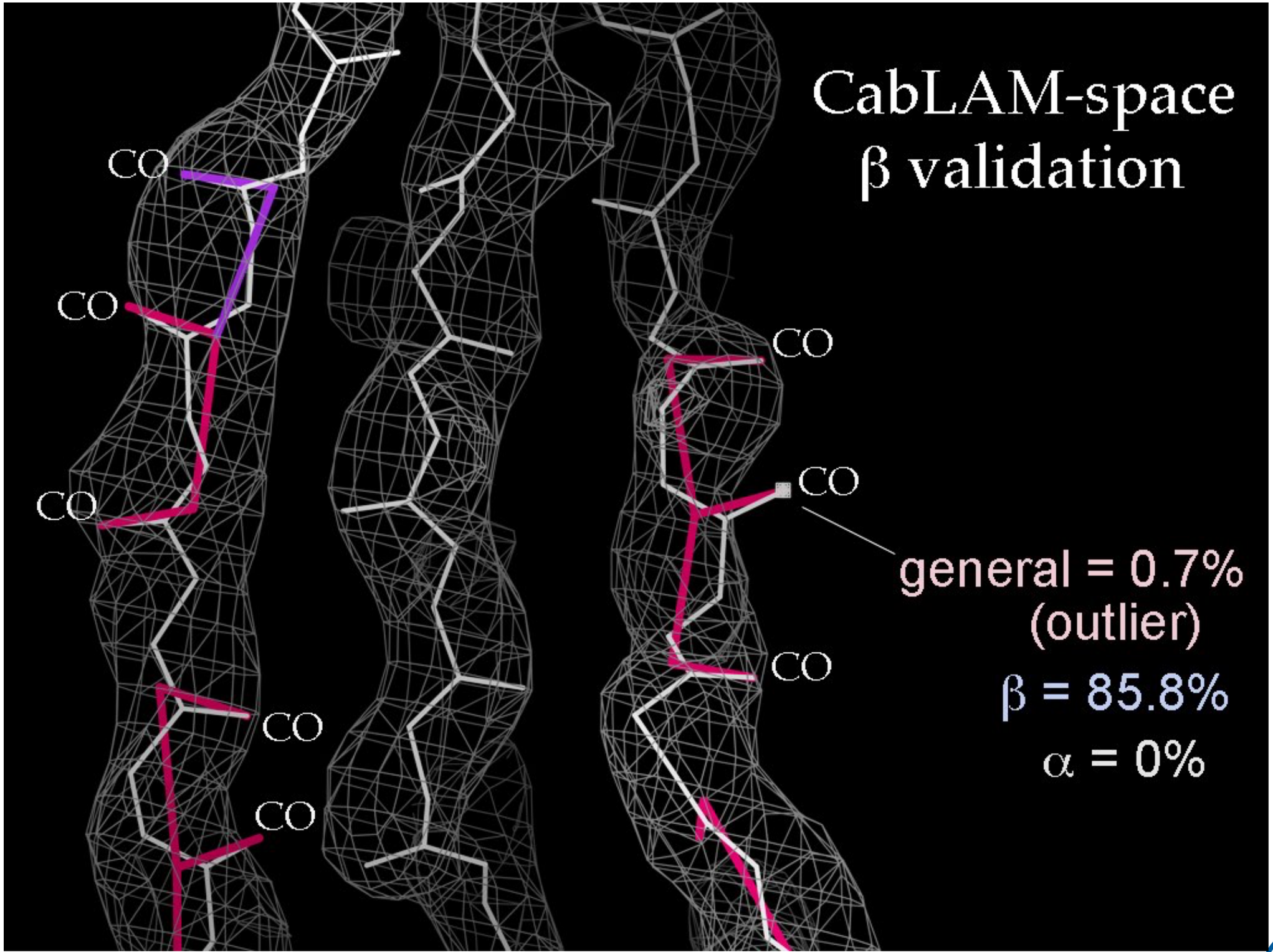
BERKELEY LAB
Lawrence Berkeley National Laboratory

Identifying Distorted Secondary Structure



Christopher Williams,
Duke University

Assessing Secondary Structure Probability



Christopher Williams,
Duke University

Comprehensive Validation

Comprehensive validation (CryoEM) (Project: rea-space-refine-6crz)

Preferences Help Run Abort Ask for help

Input/Output ValidationCryoEM_7

Summary Model Model vs. Data Data

Files

Model: /Users/PDAdams/Documents/rea-space-refine-6crz/mo

Map: /Users/PDAdams/Documents/rea-space-refine-6crz/ma

Open in Coot

White cells are mostly informational.
Green cells imply that the values are in an acceptable range.
Yellow cells imply that the values need to be checked carefully.
Red cells imply that the values are cocerning and that the model shou
Clicking on a row will bring up a panel with more detailed informatio

Model

MolProbity

MolProbity score	1.72	
Clash score	5.44	
Rotamer outliers (%)	0.00	(Goal: < 1%)
Cβ outliers	0	(Goal: 0)

CaBLAM

Outliers (%)	3.88	(Goal: <= 1%)
Disfavored (%)	8.96	(Goal: <= 5%)
Cα outliers (%)	1.19	(Goal: <= 0.5%)

Geometry Restraints

Bond **Angle**

Idle

Comprehensive validation (CryoEM) (Project: rea-space-refine-6crz)

Preferences Help Run Abort Ask for help

Input/Output ValidationCryoEM_7

Summary Model Model vs. Data Data

MolProbity Rotamers Ramachandran Clashes Geometry Restraints

These statistics are computed using the same underlying distributions as the MolProbity web server. The overall score represents the experimental resolution expected for a model of this quality; ideally the score should be lower than the actual resolution.

Overall scores

MolProbity score: 1.72 **Clash score:** 5.44

CaBLAM

Outliers (%): 3.88 **Disfavored (%):** 8.96 **Cα outliers (%):** 1.19

Chain	Residue	Evaluation	CaBLAM Score	CA Geometry Score	Secondary Struct
A	ILE 955	CaBLAM Disfav...	0.03762	0.01447	
A	PRO 969	CaBLAM Disfav...	0.02931	0.46424	try alpha helix
A	SER 1012	CaBLAM Outlier	0.00273	0.67504	try alpha helix
A	LEU 1016	CaBLAM Outlier	0.00086	0.07553	

Cβ deviation analysis

No Cβ position outliers detected.

Cis and twisted peptides

Cis conformations are observed in about 5% of Prolines.
Cis conformations are observed in about 0.03% of general residues.
Twisted peptides are almost certainly modeling errors.

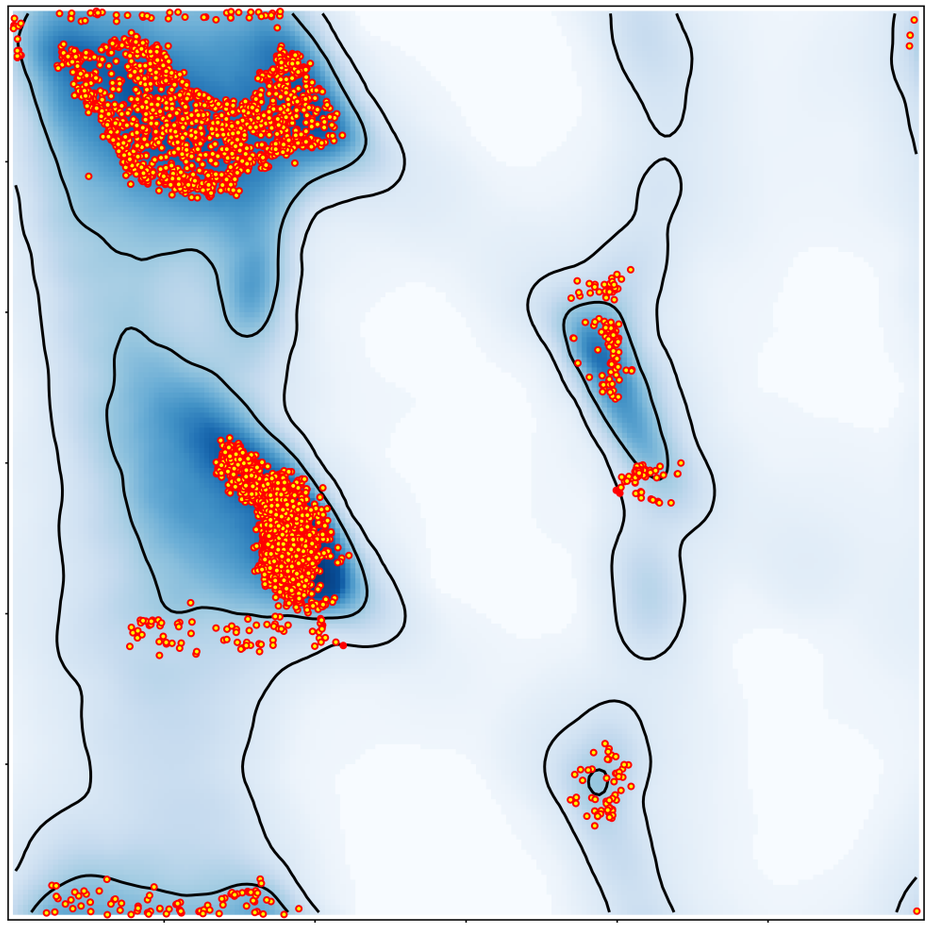
No non-trans peptides detected.

Idle Project: rea-space-refine-6crz

BERKELEY LAB
Lawrence Berkeley National Laboratory

Validating the Ramachandran Plot

- When restraints based on validation metrics are needed, care needs to be taken with interpretation of validation results

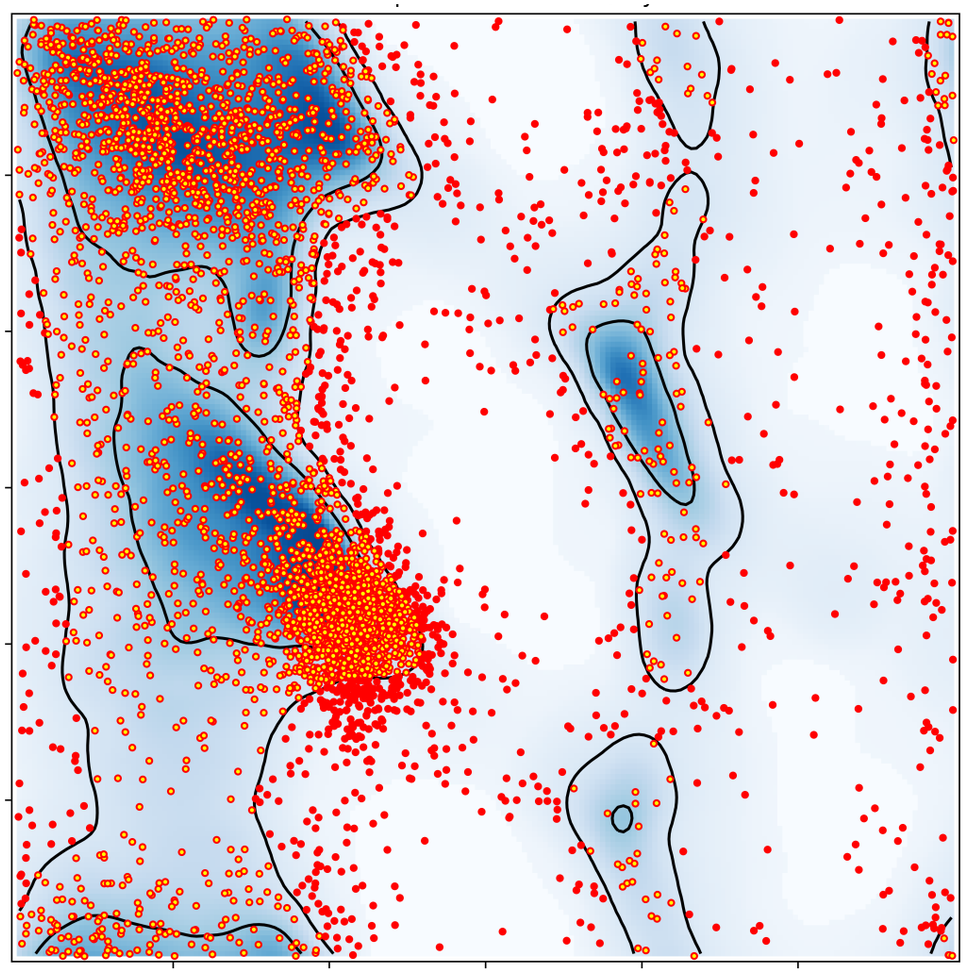


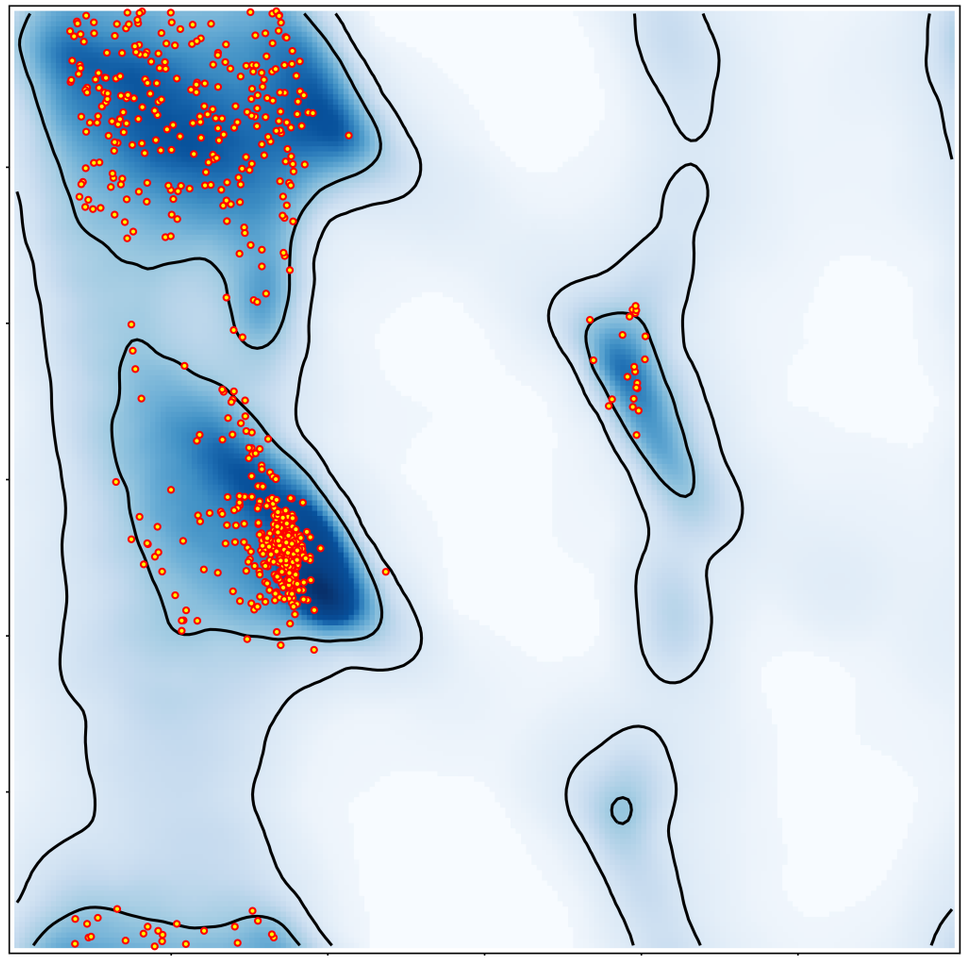
Favored	96.4
Outliers	0.2

BERKELEY LAB
Lawrence Berkeley National Laboratory

Detecting Unusual Distributions

- A poor model can have a clearly poor Ramachandran plot
- A poor model with inappropriately applied restraints may be less clear





BERKELEY LAB
Lawrence Berkeley National Laboratory

The Rama-Z (Ramachandran plot Z-score)

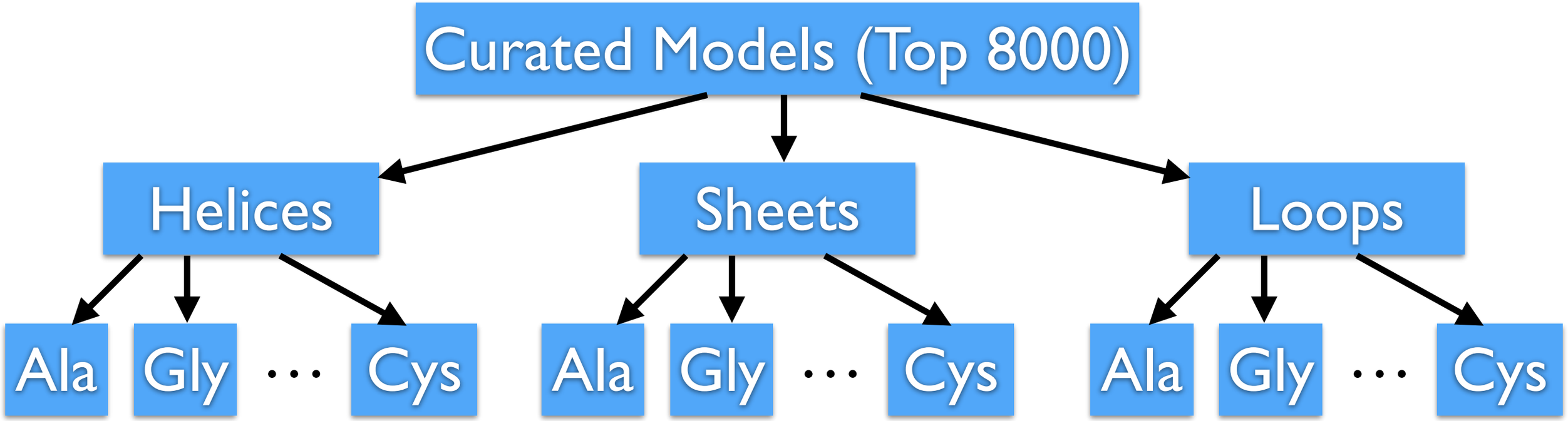
CABIOS

Vol. 13 no. 4 1997
Pages 425–430

Objectively judging the quality of a protein structure from a Ramachandran plot

Rob W.W.Hooft, Chris Sander and Gerrit Vriend

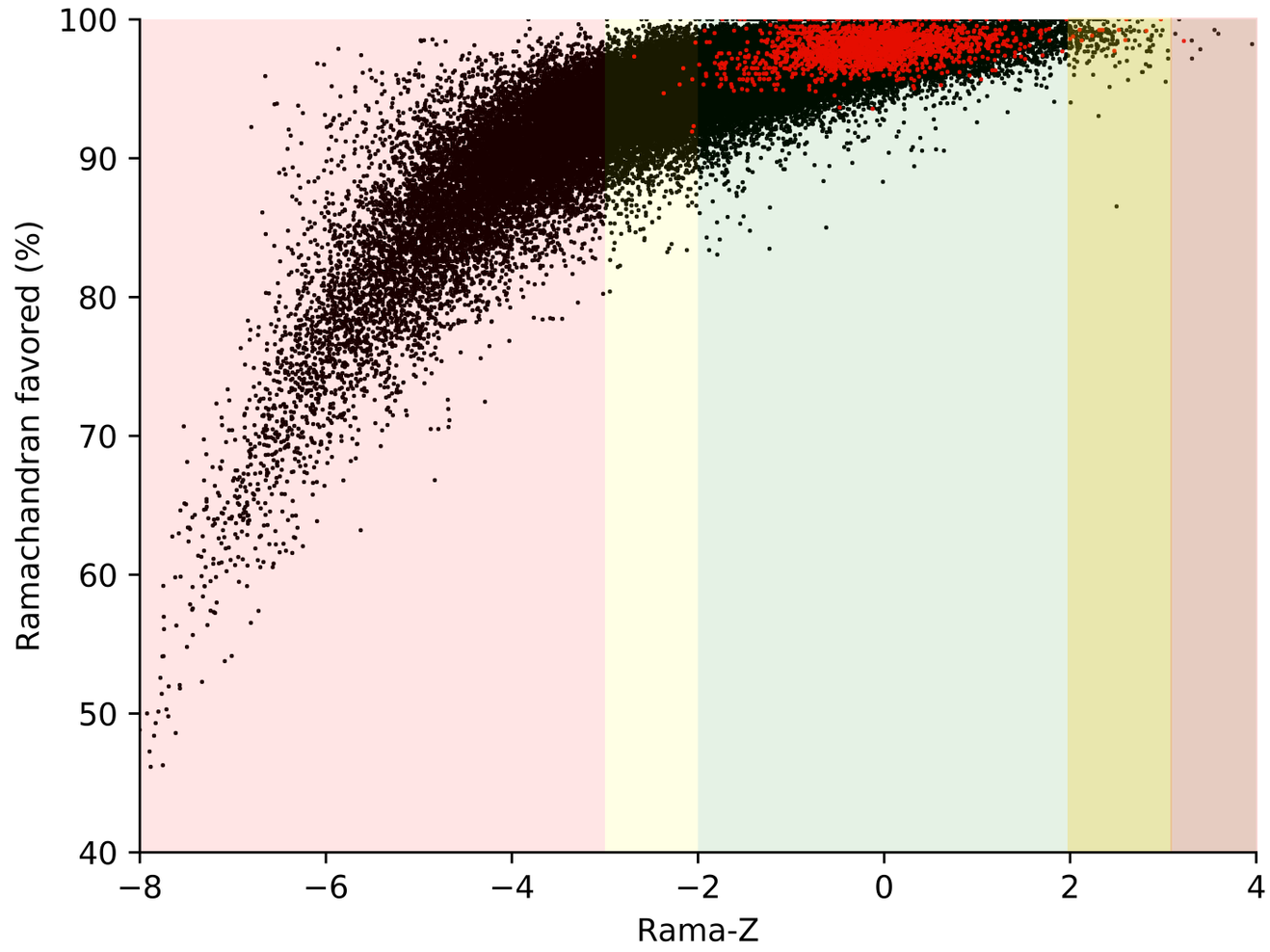
- Comparison of the distribution of (φ, ψ) of a particular model with reference distributions



BERKELEY LAB
Lawrence Berkeley National Laboratory

The Rama-Z (Ramachandran plot Z-score)

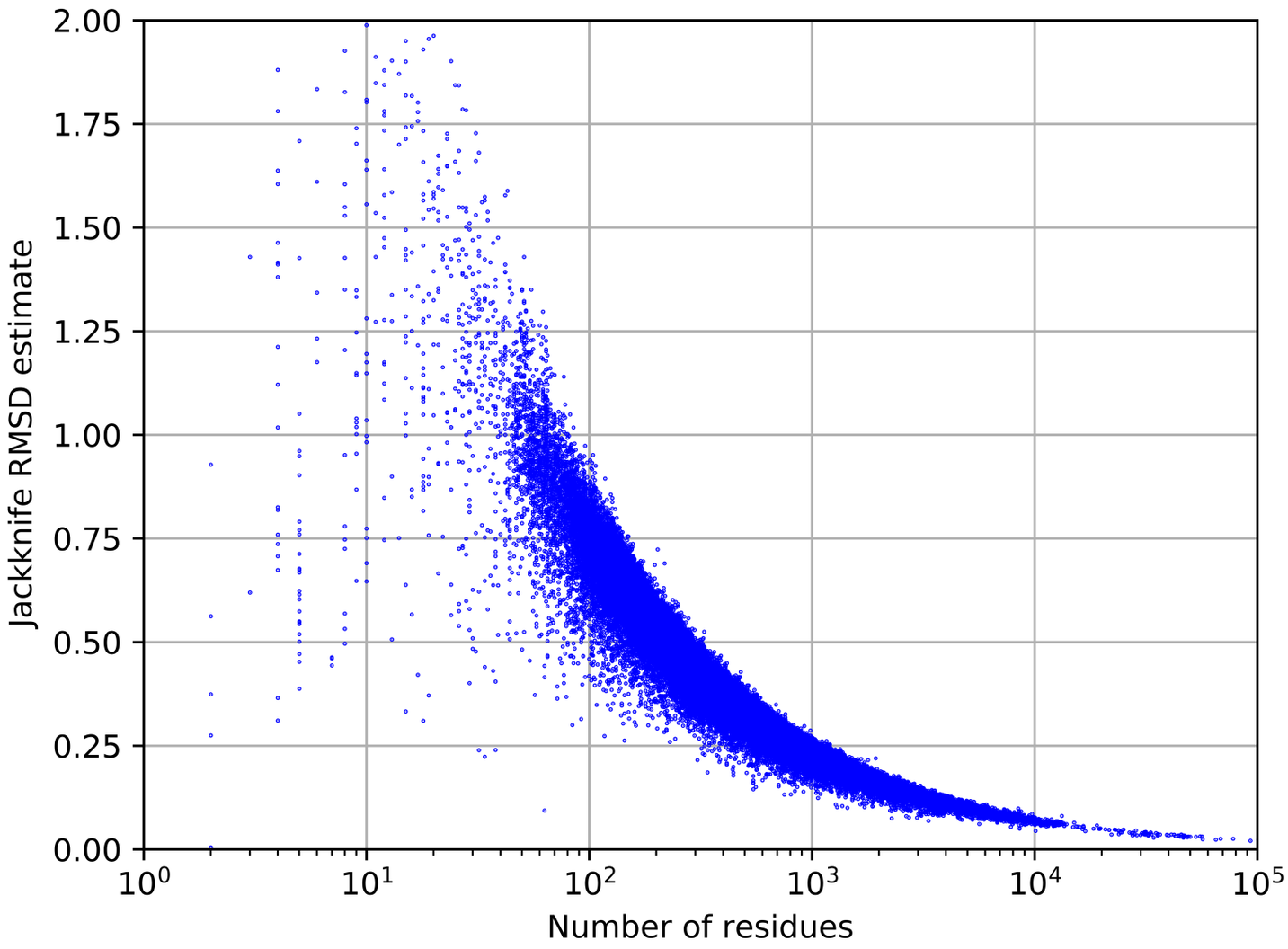
- Rama-Z score is good at identifying odd-looking Ramachandran plots
- Used in PDBREDO and WHAT_CHECK. Now implemented in Phenix.
- One number, simple criteria:
  - 0 < |Z| < 2: Good
  - 2 < |Z| < 3: Suspicious
  - |Z| > 3: Poor
- All models in PDB with resolution better than 1.2Å have Rama-Z > -3.



Sobolev et al: A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry. *Structure* 2020, 28:1249-1258

Rama-Z reliability (RMSD)

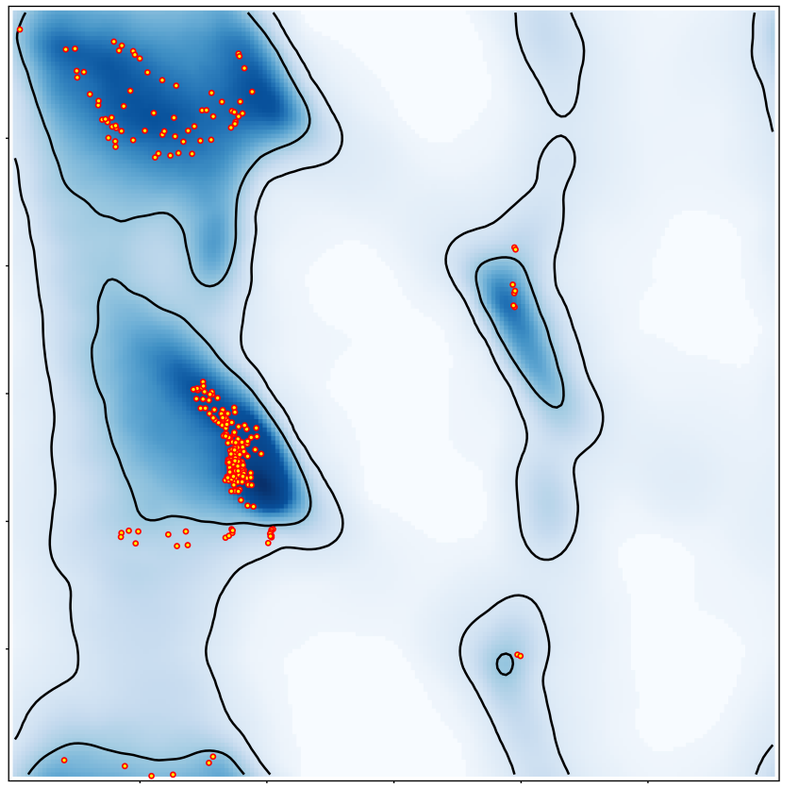
- Jackknife estimate of Rama-Z RMSD for a particular model
- More residues – more reliable Rama-Z score
- Rama-Z can be used to track progress during refinement runs



BERKELEY LAB
Lawrence Berkeley National Laboratory

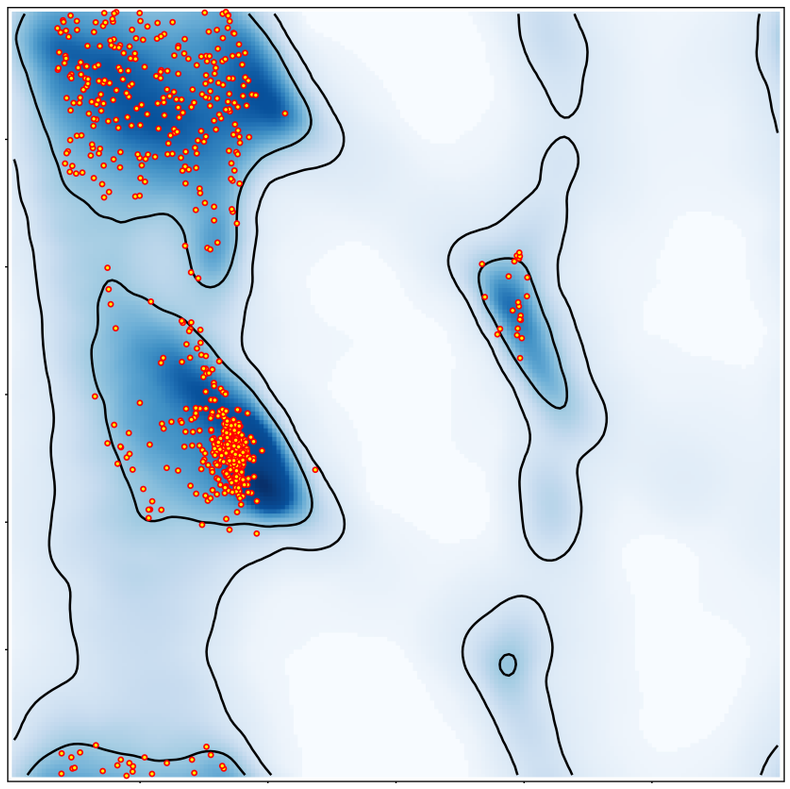
Unlikely Ramachandran Plots

6bu9 | 6.8 Å



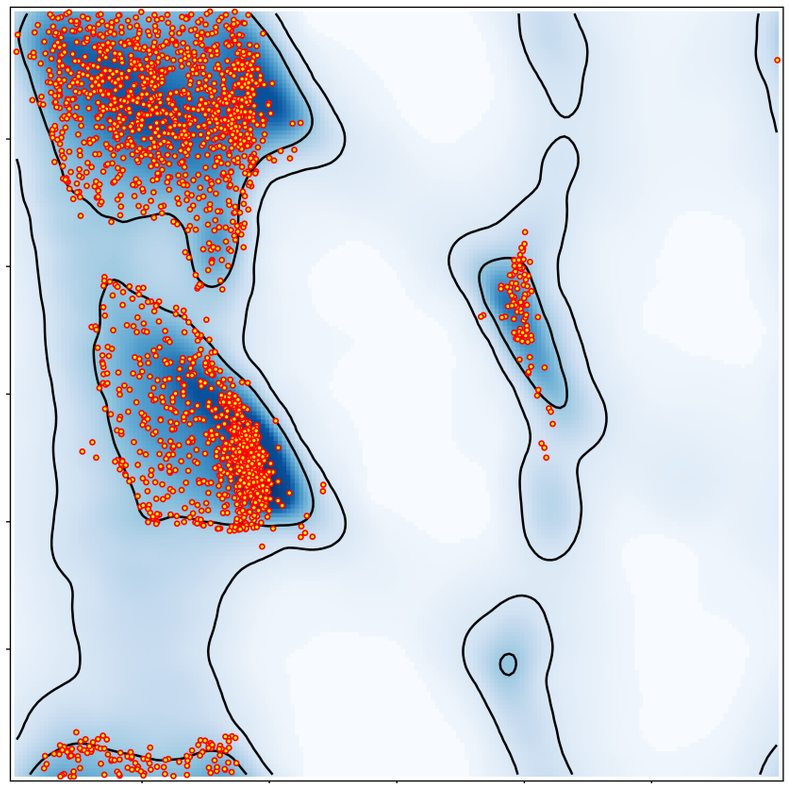
Rama-Z = -5.0

6dzv | 4.2 Å



Rama-Z = -4.1

6cs1 | 4.6 Å



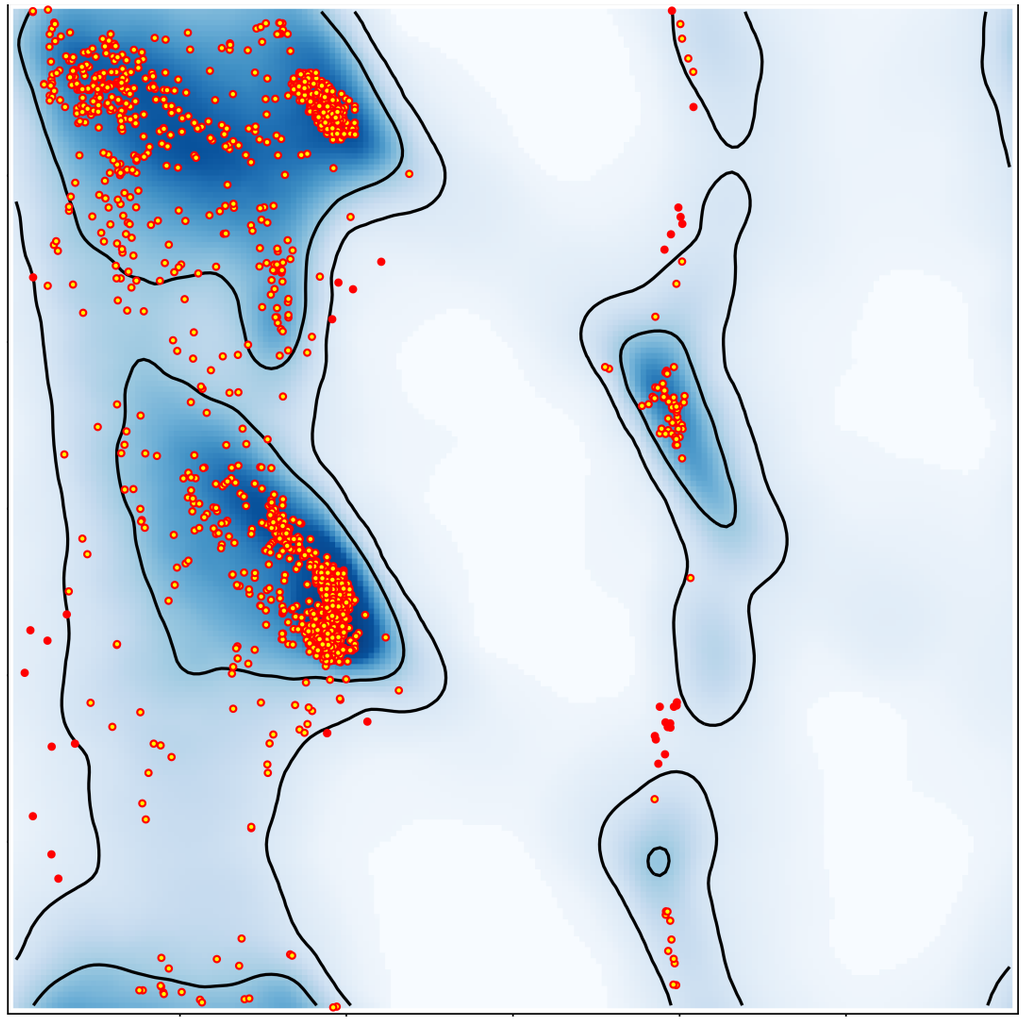
Rama-Z = -4.2

- |Z| > 3: Poor
- 2 < |Z| < 3: Suspicious
- |Z| < 2: Good

BERKELEY LAB
Lawrence Berkeley National Laboratory

Separate Rama-Z Scores

- Sometimes Rama-Z calculated for the whole model will not trigger a warning



- Separate Rama-Z scores:
  - Whole: -1.6
  - Helix: -1.9
  - Sheet: -2.5
  - Loop: -0.7
- Recommend checking separate Rama-Z scores when whole Rama-Z is OK.

BERKELEY LAB
Lawrence Berkeley National Laboratory

Availability in Phenix

- In phenix.refine and phenix.real_space_refine

Comprehensive Validation

Model: 5t4q_8359.pdb
Map: 5t4q_8359.map

MOL PROBITY

Open in Coot

Export Table 1

Model	
Composition (#)	
Chains	
Atoms	
Residues	
Water	
Ligands	
Bonds (RMSD)	
Length (Å)	
Angles (°)	
MolProbity score	
Clash score	
Ramachandran p	
Outliers	
Allowed	
Favored	
Rama-Z (Ramac	
whole (N =	
helix (N =	
sheet (N =	
loop (N = 2	
Rotamer outlier	
Cβ outliers (%)	
Peptide plane (%)	
Cis proline/general	0.0/0.3
Twisted proline/general	0.0/0.0
CaBLAM outliers (%)	9.18
ADP (B-factors)	
Iso/Aniso (#)	23568/0
min/max/mean	
Protein	0.52/500.00/328.57
Nucleotide	---
Ligand	23.19/44.57/31.76
Water	---
Occupancy	
Mean	1.00
occ = 1 (%)	99.97
0 < occ < 1 (%)	0.03
occ > 1 (%)	0.00

Allowed	14.48
Favored	81.93
Rama-Z (Ramachandran plot Z-score, RMSD)	
whole (N = 4732)	-6.15 (0.09)
helix (N = 1921)	-4.82 (0.03)
sheet (N = 272)	-4.38 (0.23)
loop (N = 2539)	-3.48 (0.11)
Rotamer outliers (%)	0.00

Idle

Project: test

Output PDB files

```
Geometry
---------------------------------------------------------------------------

GEOMETRY RESTRAINTS LIBRARY: GEOSTD + MONOMER LIBRARY
DEVIATIONS FROM IDEAL VALUES.
  BOND      :  0.004   0.049  23553
  ANGLE     :  1.051  10.386  32694
  CHIRALITY :  0.040   0.136   4352
                       0.002   4759
                      53.708   4805
                      CE : 2.499

                      : 7.12

                      0.00 %
                      0.00 %

                      0.00 %
                      0.28 %
                      0.00 %
                      0.00 %
```

```
LOT Z-SCORE):
INTERPRETATION: BAD < -3 | SUSPICIOUS < -2 | GOOD > -2
VALUES FOR HELIX/SHEET/LOOP ARE NOT ADDITIVE AND ARE INTERPRETED
INDEPENDENTLY.
  WHOLE: -6.15 (0.09), RESIDUES: 4732
  HELIX: -4.82 (0.03), RESIDUES: 1921
  SHEET: -4.38 (0.23), RESIDUES: 272
  LOOP : -3.48 (0.11), RESIDUES: 2539
```

BERKELEY LAB
Lawrence Berkeley National Laboratory

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868

Conclusions

- Density modification methods can be successfully applied to cryo-EM reconstructions
- Structure solution at low resolution presents some new challenges for validation, requiring new metrics
  - Higher-dimensional geometry distributions (CaBLAM) can identify problem regions and provide suggestions about secondary structure
  - Rama-Z is able to identify unusual Ramachandran plot distributions. It should be used together with standard outliers metrics
  - CaBLAM and Rama-Z should be included in standard validation reports provided by the wwPDB and "Table 1" reported in structural publications

BERKELEY LAB
Lawrence Berkeley National Laboratory

Acknowledgements

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, Christopher Schlicksup, Oleg Sobolev, Peter Zwart

Los Alamos Laboratory/New Mexico Consortium
Tom Terwilliger, Li-Wei Hung

Baylor College of Medicine
Matt Baker, Corey Hryc

Cambridge University
Randy Read, Airlie McCoy, Gabor Bunckozi, Tristan Croll, Kaushik Hatti, Claudia Millán Nebot, Rob Oeffner, Massimo Sammito, Duncan Stockwell, Laurent Storoni

Duke University
Jane Richardson & David Richardson, Ian Davis, Vincent Chen, Jeff Headd, Chris Williams, Bryan Arendall, Bradley Hintze, Laura Murray

UC San Francisco
Ben Barad, Yifan Cheng, Jaime Fraser

University of Washington
Frank DiMaio, Ray Wang, David Baker

Oak Ridge National Laboratory
Marat Mustyakimov, Paul Langan

Other Collaborators
Maarten Hekkelman, Robbie Joosten, Tassos Perrakis
Corey Hryc, Zhao Wang, Steve Ludtke, Wah Chiu
Pawel Janowski, David Case
Dale Tronrud, Donnie Berholz, Andy Karplus
Alexandre Urzhumtsev & Vladimir Lunin
Garib Murshudov & Alexi Vagin
Paul Emsley, Bernhard Lohkamp, Kevin Cowtan
PHENIX Testers & Users

Funding
- NIH/NIGMS: P01GM063210, P50GM062412, P01GM064692, R01GM071939
- PHENIX Industrial Consortium
- Lawrence Berkeley Laboratory

BERKELEY LAB
Lawrence Berkeley National Laboratory

Phenix

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868

CERES - Cryo-EM re-refinement system

cci.lbl.gov/ceres/goto_entry/7ctt_30469/10_2020

iCloud ATT Email Apple Apple System Status Google Maps Work News Journals Personal

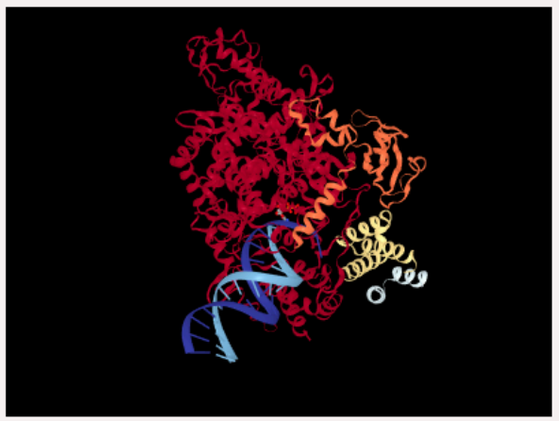
CERES Home About Glossary of terms Figures PDB/EMDB Contact Phenix

7CTT (map ID: 30469) - 10_2020

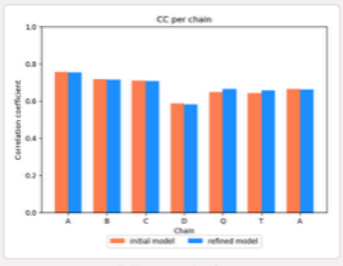
Summary | Model | Data | NGL Viewer | Downloads/Links

Cryo-EM Table 1

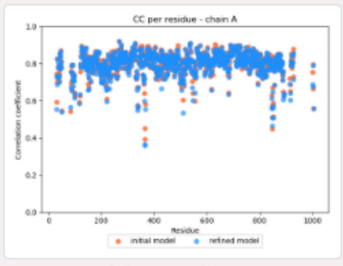
	PDB	Re-refined
Model		
Clashscore	10.17	12.21
RMSD bonds (Å)	0.009	0.003
RMSD angles (°)	0.91	0.57
Cβ deviations (count)	0.0	0.0
Rotamer outliers (%)	0.45	0.0
Min nonbonded distance (Å)	2.123	2.153
Cablam outliers (%)	2.89	2.89
Ramachandran (residues in %)		
outliers	0.0	0.0
allowed	4.7	4.4
favored	95.3	95.6
Rama Z-Score	-2.8	-0.85
RMSD initial vs. re-refined (Å)	0.43	
Data		
Resolution (Å)	3.2	
d_{model} (Å)	3.3	3.4
d_{99} (Å)	3.4	
Model vs data		
CC_{box}	0.71	0.71
CC_{mask}	0.76	0.76
EMRinger	2.65	2.92



Model from PDB



CC per chain



CC per residue

CERES © 2020 Privacy and Security Notice

BERKELEY LAB Lawrence Berkeley National Laboratory

https://cci.lbl.gov/ceres/

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868