Refinement in Phenix

Argonne, June, 2011

Paul Adams

Lawrence Berkeley Laboratory and Department of Bioengineering UC Berkeley

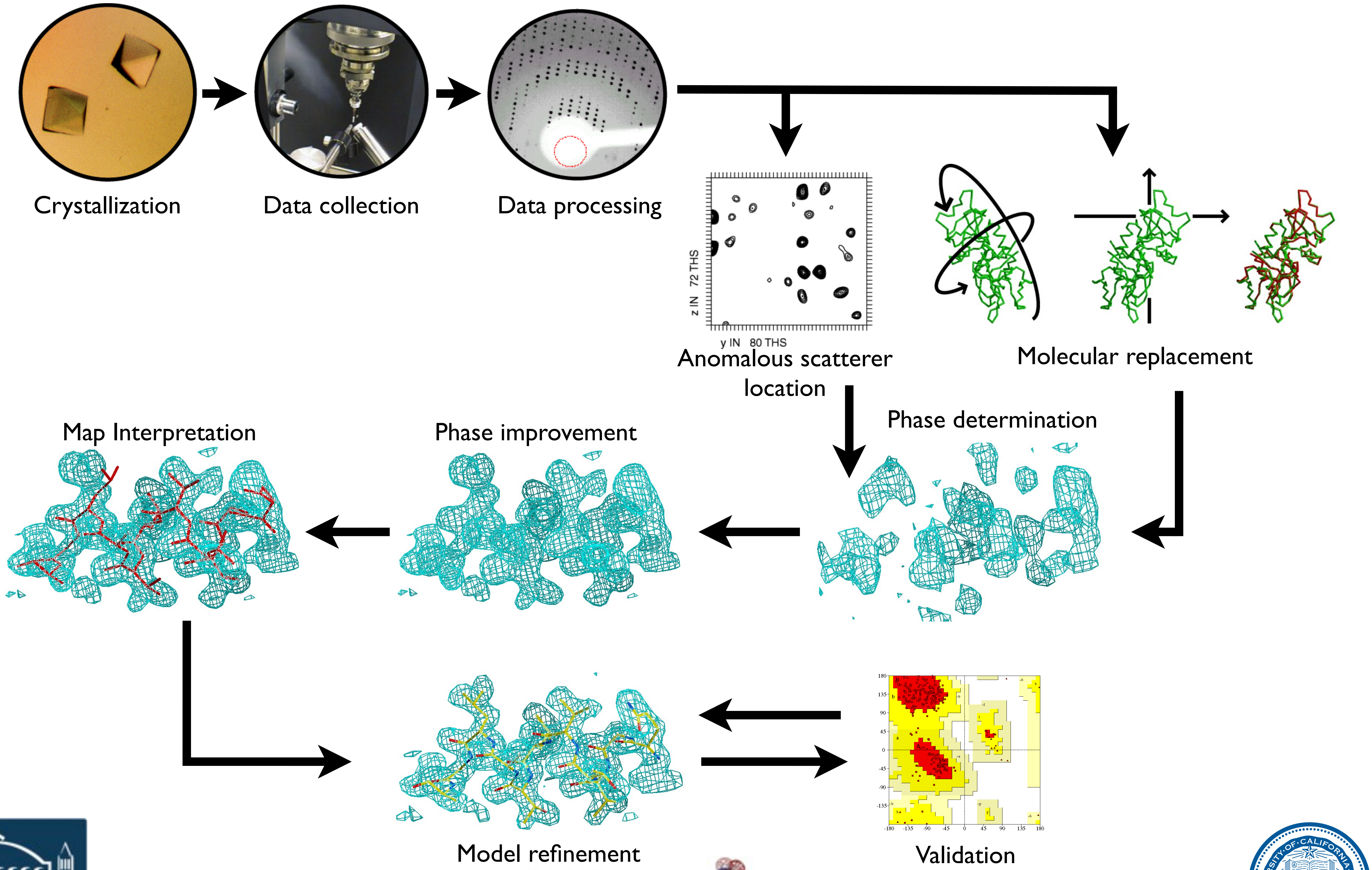
Eruditio et Religio

Los Alamos
NATIONAL LABORATORY

BERKELEY LAB

UNIVERSITY OF CAMBRIDGE

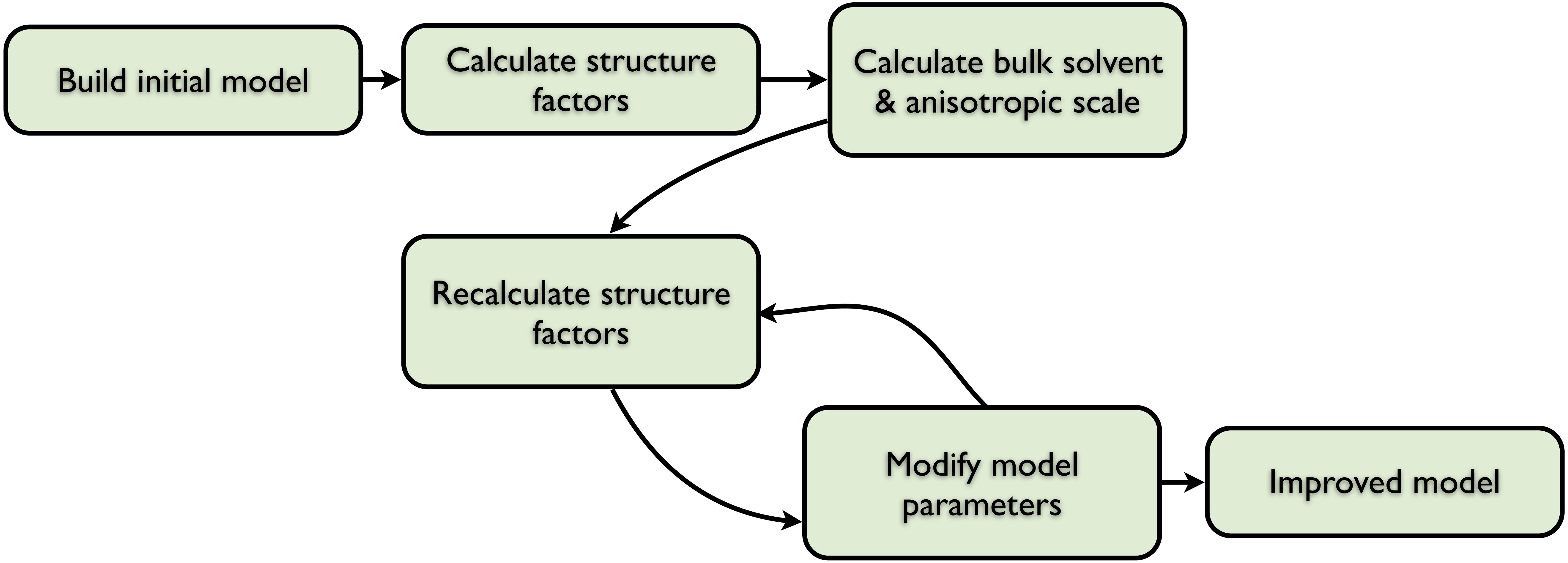
The Crystallographic Process



Phenix

BERKELEY LAB
Lawrence Berkeley National Laboratory

Overview of Structure Refinement


Build initial model
Calculate structure factors
Calculate bulk solvent & anisotropic scale
Recalculate structure factors
Modify model parameters
Improved model

- Structure refinement is an iterative process that changes the model parameters while improving the fit to the experimental data

Crystallographic Structure Refinement

- An *optimization* algorithm is used to minimize a *target function* by changing the *parameters* of the model
- Parameters:
  - coordinates, B-values, occupancies
- Optimization algorithm:
  - minimization, simulated annealing
- Target function (Objective function):
  - Function based on electron density (real-space refinement)
  - Function based on structure factors (reciprocal-space refinement)

$$E = E_{chem} + w_a \sum_{hkl} \frac{1}{\sigma^2}\left(|F_o| - |F_c|\right)^2$$

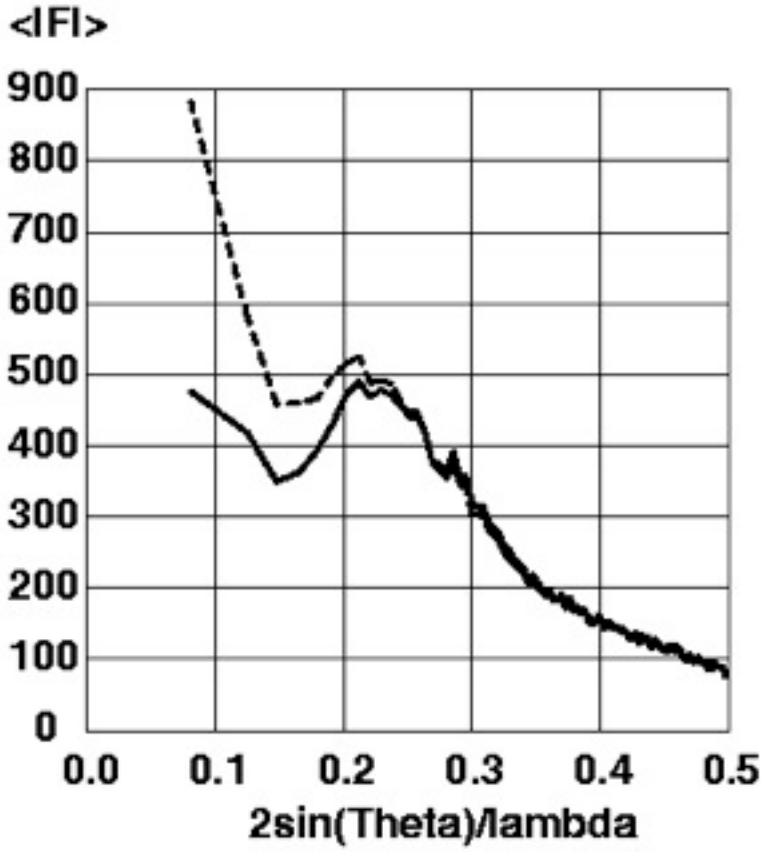
Why do we need Refinement?

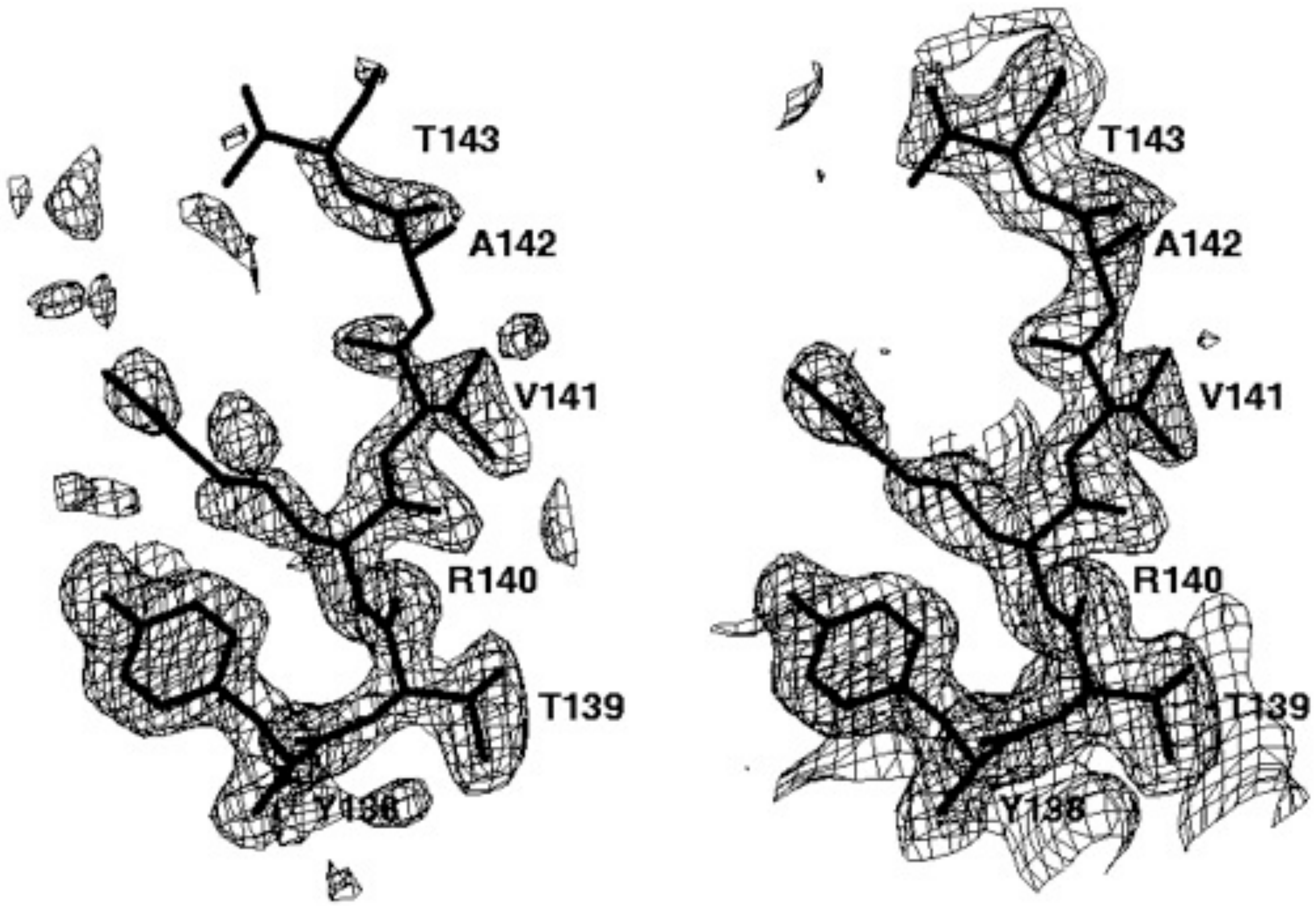
- The models generated by hand our automatically typically have errors and are incomplete:
  - Missing atoms that should be included (missing domains, loops, sidechains, ligands, water, …)
  - Atoms that that have been misplaced
- This is a result of:
  - Experimental phases are sometimes poor, especially at low resolution
  - Molecular Replacement phases can generate model bias
  - Every atom that has an error affects all calculated structure factors and thus changes the density at all other points in the map
- As the model is improved, the phases improve, revealing new aspects of the structure (loops, sidechains, ligands, water, …)

The Model

- Structure factors from the model are calculated using a FFT (by sampling the Gaussian form factors on a grid)
- The model has to include a contribution from the bulk solvent in the crystal (calculated using a mask around the protein)

$$\mathbf{F} = k\{\mathbf{F}_{\mathrm{calc}} \exp[-\Delta B(\sin\theta/\lambda)^2] + d_{\mathrm{solv}}\mathbf{F}_{\mathrm{solv}} \exp[-B_{\mathrm{solv}}(\sin\theta/\lambda)^2]\}$$

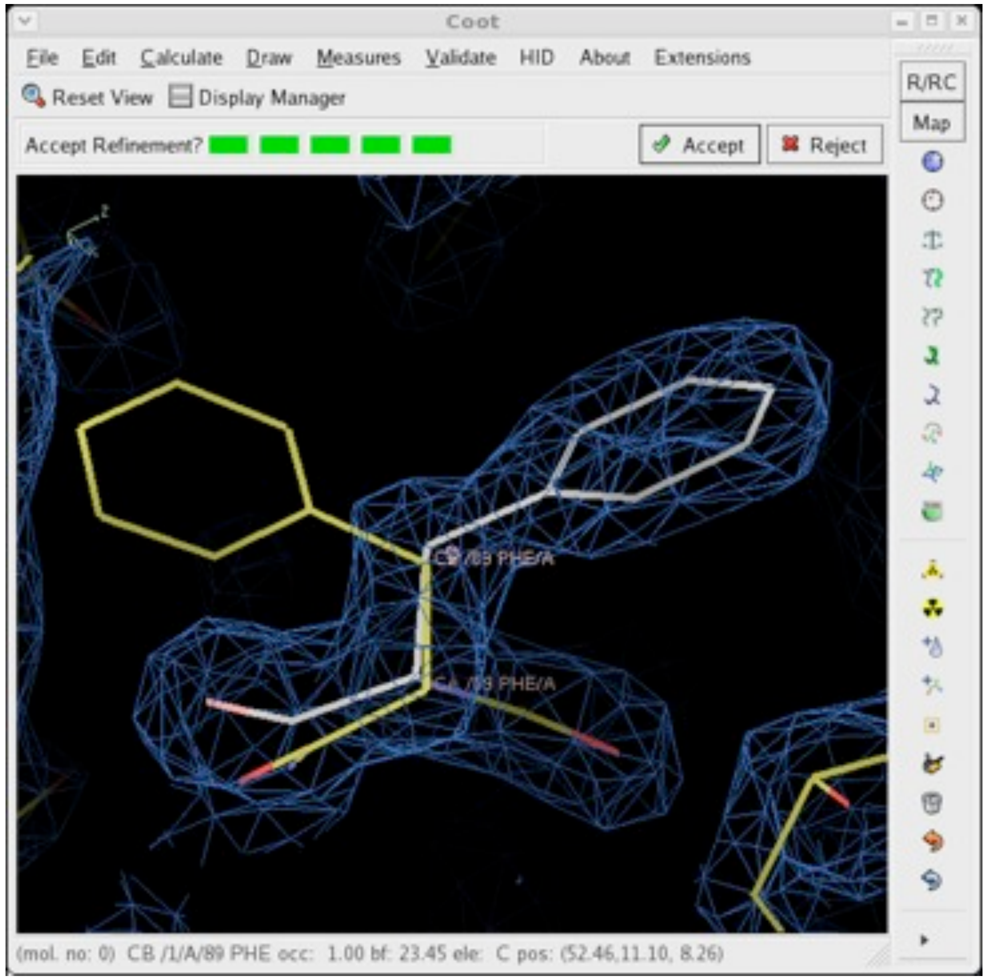




Phenix

The X-ray Term

- Real space:
  - Least-squares residual: $\Sigma\ (\rho_{obs} - \rho_{calc})^2$
  - Convolution product: $\Sigma\ \rho_{obs} \times \rho_{calc}$
  - Sum of differences: $\Sigma|\rho_{obs} - \rho_{calc}|$


Coot
File Edit Calculate Draw Measures Validate HID About Extensions
Reset View
Display Manager
Accept Refinement? 🟩 🟩 🟩 🟩 🟩
✓ Accept
✖ Reject
R/RC
Map
CB /89 PHE/A
CA /89 PHE/A

```
(mol. no: 0)  CB /1/A/89 PHE occ:  1.00 bf: 23.45 ele:  C pos: (52.46,11.10, 8.26)
```

Image from ccp4wiki

- Reciprocal space:
  - Least-squares residual: $\Sigma\ (|F_{obs}| - k\ |F_{calc}|)^2$
  - Correlation coefficient between $|F_{obs}|$ and $|F_{calc}|$
  - Functions including phases:
    - $\Sigma\ w\ [(A_{obs} - k\ A_{calc})^2 + (B_{obs} - k\ B_{calc})^2]$

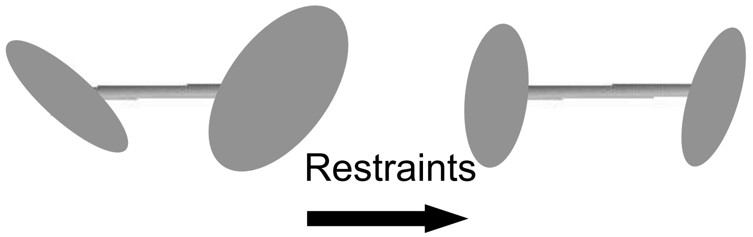
Observations and Parameters

- In contrast to small molecule crystallography we have:
  - Large unit cells, typically 50% disordered solvent, flexibility
  - Often limited resolution (2.5Å or worse)
  - Observation to parameter ratios close to 1 or worse

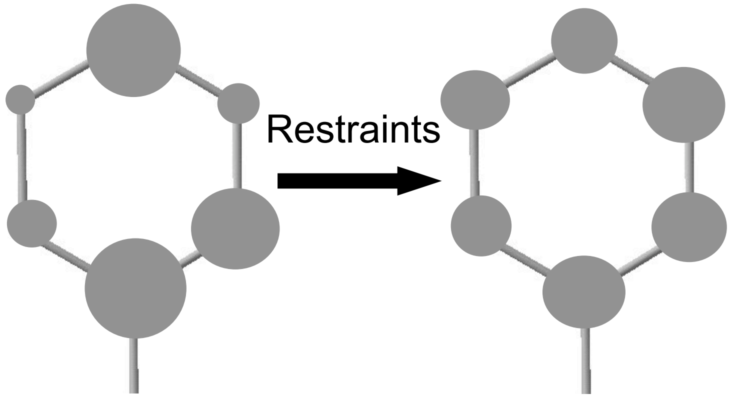
Resolution	Reflections	xyz	xyzB	xyzU
3.0	3,500	0.8	0.6	0.3
2.5	6,800	1.6	1.2	0.5
1.9	13,500	3.1	2.3	1.0
1.5	29,800	6.8	5.1	2.3
1.2	58,800	13.3	10.0	4.4
1.0	81,300	18.5	13.8	6.1

Improving the Observation to Parameter Ratio

- To make refinement practical the observation to parameter ratio is increased using restraints and constraints:
- Restraint
  - Model property ~ ideal value
  - Adds prior observed information (reduces the number of parameters refined)
  - Inclusion of chemical information in the objective function
- Constraint
  - Model property = ideal value
  - Removes one or more parameters from the model



Other Restraints



- Atomic displacement parameters
  - Bonded atoms should have similar displacement parameters
    - Restrain bonded atoms to have similar displacement values:
      - $E = \Sigma_{\text{bonds}}\, w\, (ADP_1 - ADP_2)^2$
    - Restrain displacement parameters for each atom to be similar to those of the atoms in their neighborhood:

$$E_{\mathrm{ADP}} = \sum_{i=1}^{N_{\mathrm{atoms}}} \left[\sum_{j=1}^{M_{\mathrm{atoms}}} \frac{1}{r_{ij}^{\mathrm{distance\_power}}} \frac{\left(U_i - U_j\right)^2}{\left(\dfrac{U_i + U_j}{2}\right)^{\mathrm{average\_power}}} \Bigg|_{\mathrm{sphereR}} \right]$$

Constraints

- Rigid-body refinement
  - For example, molecule consists of two domains, only refine position and orientation of each domain uses only 2 * (3 rotational + 3 translation) = 12 parameters
  - So few parameters it requires only low-resolution data
- Rigid groups
  - Torsion angle refinement
- Atomic Displacement Parameters
  - All atoms have the same B one parameter
  - All main-chain and all side-chain atoms in each residue have the same B one or two parameters per residue
  - TLS refinement 20 parameters per group
- Non-crystallographic symmetry
  - A number of N NCS-related molecules/domains are assumed to be identical
  - Reduces the number of parameters by a factor N

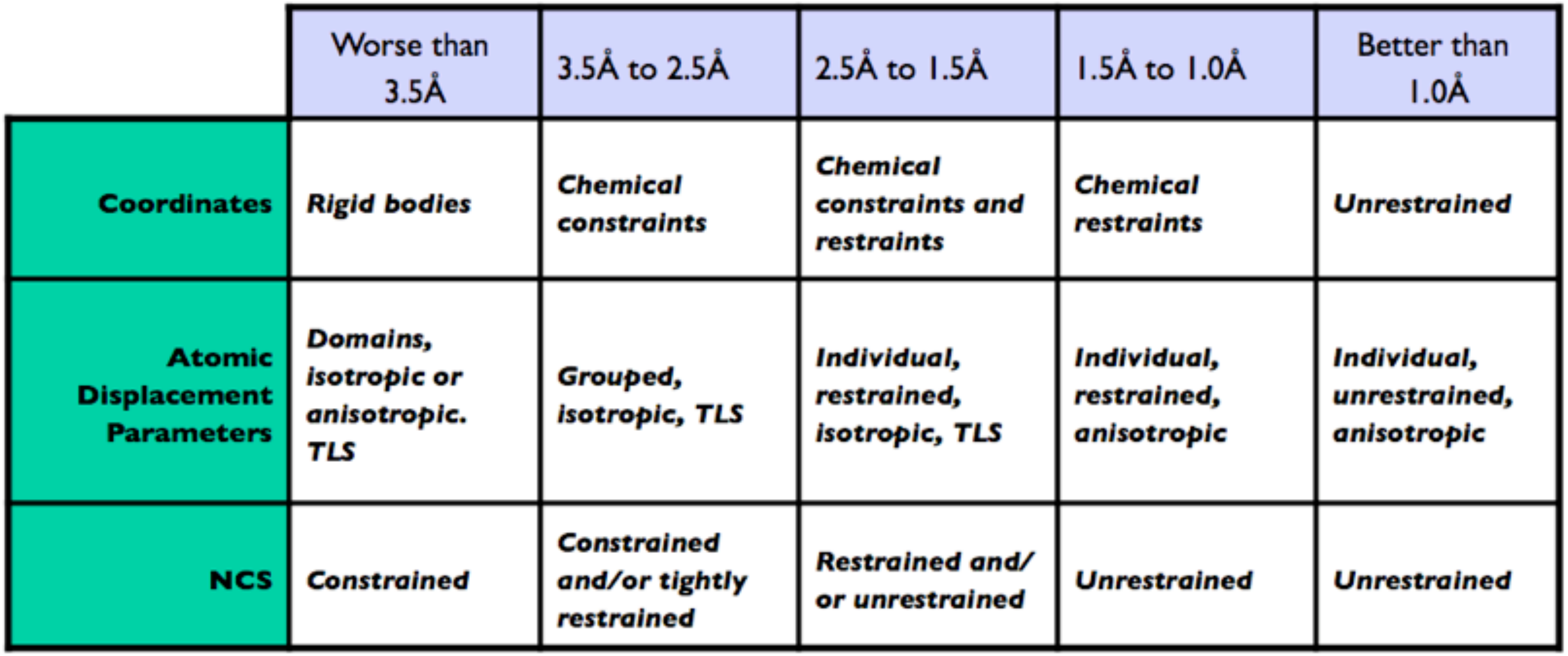
Restraint and Constraint Values

- Bond lengths and angles for proteins come from a study of Engh & Huber
  - They analysed the geometry of fragments of small molecule crystal structures similar to those found in amino acids
  - This yielded a list of distinct atom types, ideal bond lengths and angles, and estimates of their variance
  - Modifications of some values have been necessary over time (based on very high resolution structures)
- A similar analysis has been carried out for nucleic acids
- For other compounds values can be generated à la Engh & Huber, calculated by certain programs, or found in databases

Reducing Overfitting in Refinement

- Cross-validation
  - Brunger, Nature 355, 472, 1992
- Torsion angle dynamics refinement
  - Rice & Brunger, Proteins 19, 277, 1994
- Translation-Libration-Screw refinement
  - Winn et al., Acta Cryst. D 57, 122-133, 2001
- Maximum likelihood formulation of refinement
  - Bricogne, Meth. Enzymol. 276, 361, 1997
  - Murshudov, Dodson, Vagin, CCP4, 1996
  - Pannu & Read, Acta Cryst. A 52, 659-668, 1996
  - Adams, Pannu, Read, Brunger, PNAS 94, 5018, 1997

Number of Observations and Parameterizations

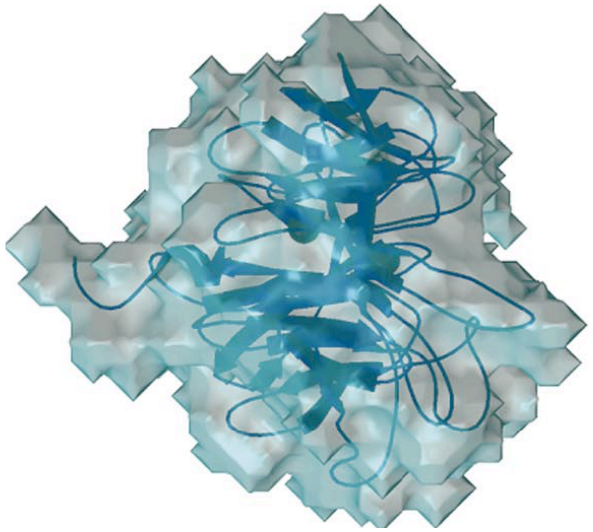


	Worse than 3.5Å	3.5Å to 2.5Å	2.5Å to 1.5Å	1.5Å to 1.0Å	Better than 1.0Å
Coordinates	***Rigid bodies***	***Chemical constraints***	***Chemical constraints and restraints***	***Chemical restraints***	***Unrestrained***
Atomic Displacement Parameters	***Domains, isotropic or anisotropic. TLS***	***Grouped, isotropic, TLS***	***Individual, restrained, isotropic, TLS***	***Individual, restrained, anisotropic***	***Individual, unrestrained, anisotropic***
NCS	***Constrained***	***Constrained and/or tightly restrained***	***Restrained and/or unrestrained***	***Unrestrained***	***Unrestrained***

- Start with the most conservative parameterization
- Only move to a less conservative parameterization after consulting minimally biased indicators (free R-value, Ramachandran plot, chemistry)
- Experimental phases usually permit a less conservative final parameterization

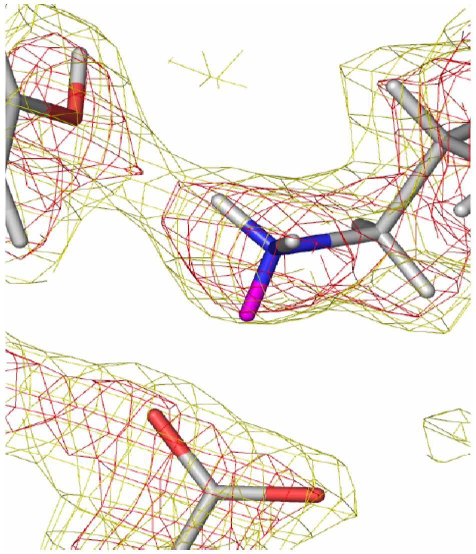
Comprehensive Structure Refinement

Low



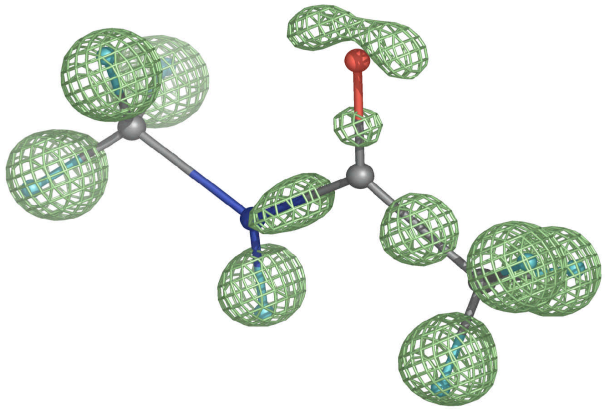
- Rigid body
- Group ADP
- Torsion angle constraints

Medium/High



- Restrained coordinates
- Restrained ADPs (iso/aniso)
- Automated water picking

Ultra-high



- Interatomic scatterers
- Unrestrained refinement
- Explicit hydrogens

- Simulated annealing
- NCS restraints (including automatic NCS determination and restraints generation)
- TLS refinement
- Occupancies (individual or group, automatically constrained for alternate side chains)
- Anomalous scattering factor refinement (individual or group)
- Twinned refinement target
- Joint refinement against X-ray and Neutron data

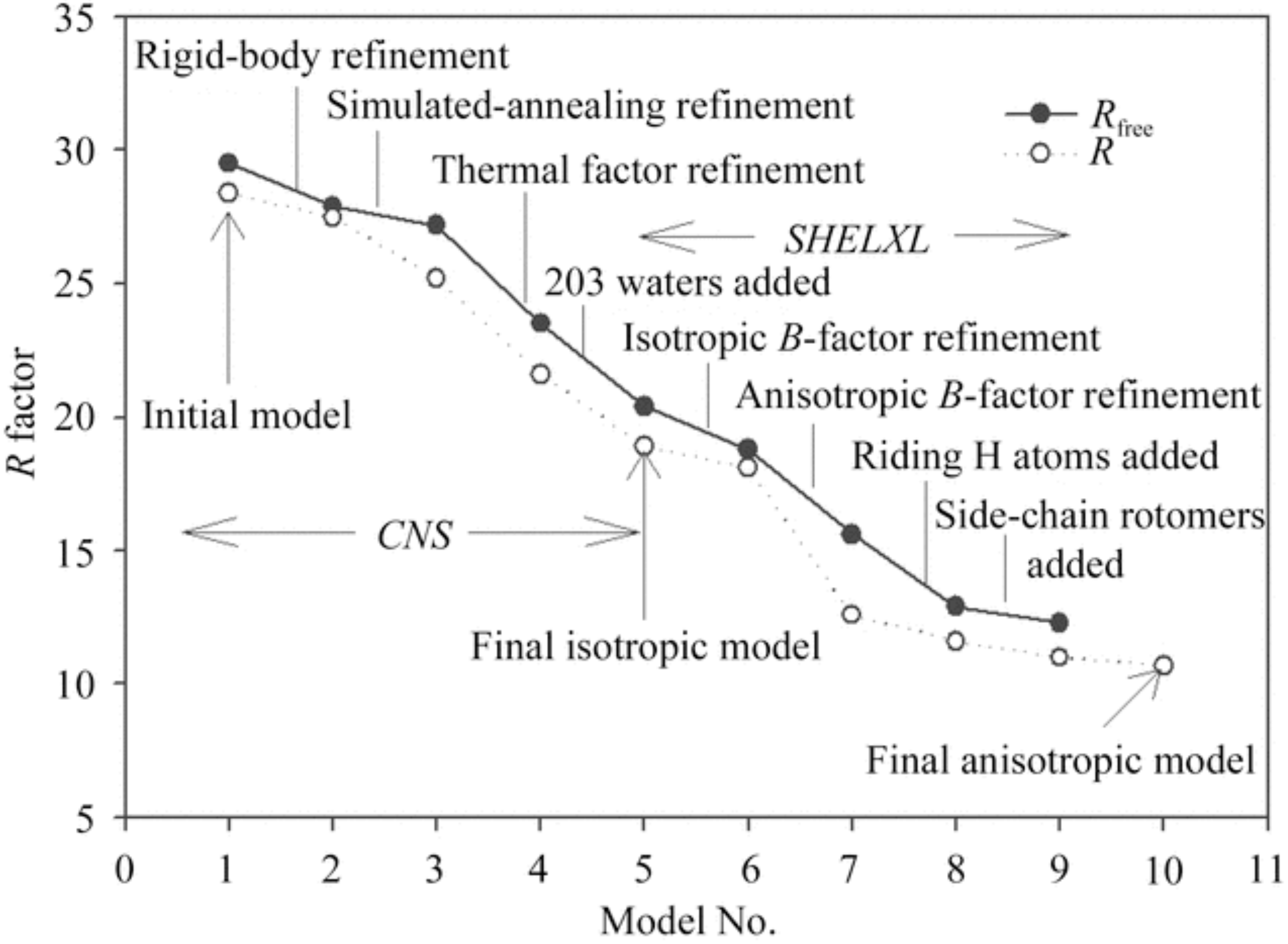
Pavel Afonine, Nat Echols, Ralf Grosse-Kunstleve & Peter Zwart, Lawrence Berkeley Laboratory

Phenix

Acta Cryst. 2005, **D61**:850-855.

Acta Cryst. 2007, **D63**:1194-1197.

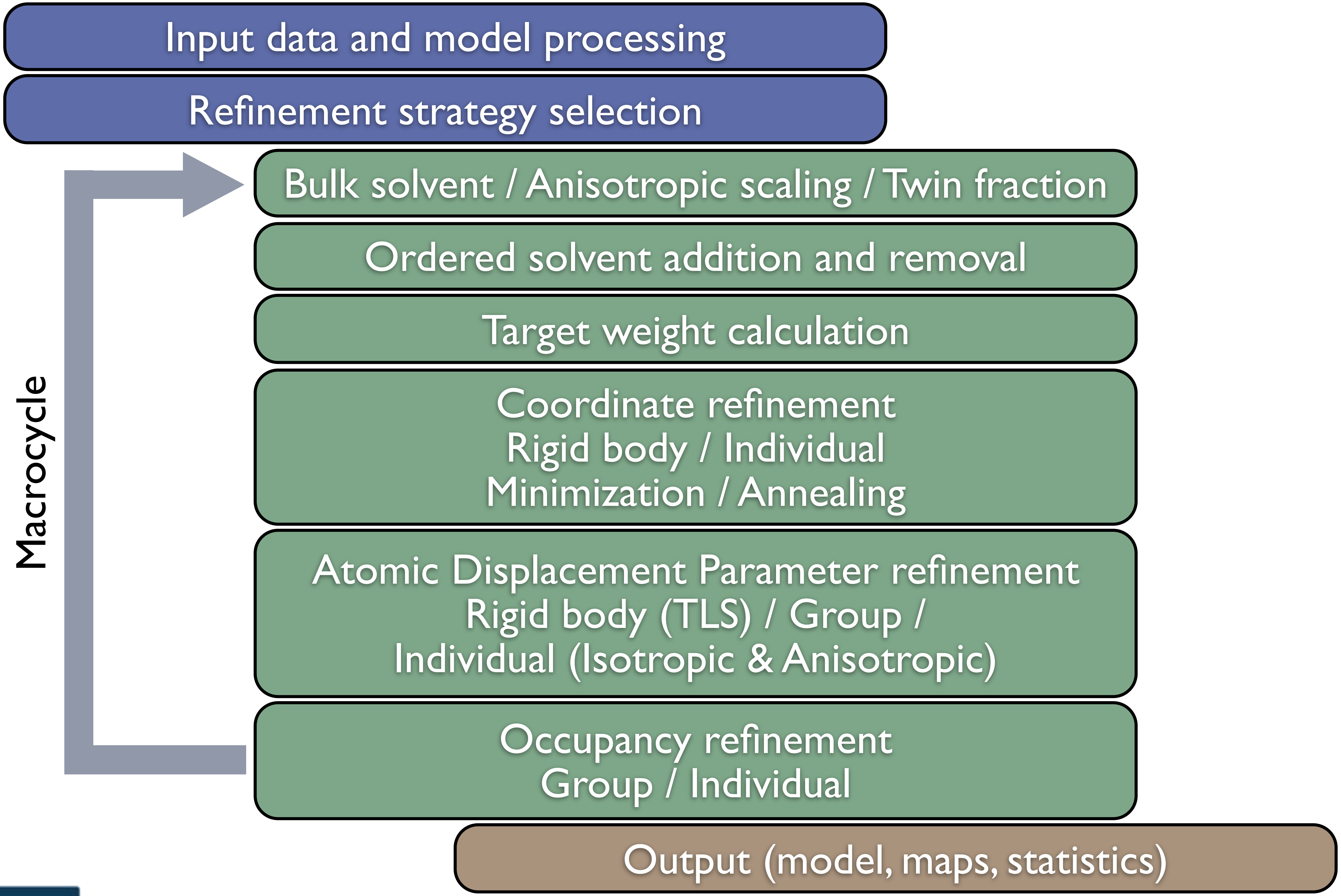
Why Automate Structure Refinement?



Acta Cryst. (2002). D58, 2009-2017, Yousef et al.

Phenix

Refinement Protocol



Input data and model processing

Refinement strategy selection

Macrocycle

Bulk solvent / Anisotropic scaling / Twin fraction

Ordered solvent addition and removal

Target weight calculation

Coordinate refinement
Rigid body / Individual
Minimization / Annealing

Atomic Displacement Parameter refinement
Rigid body (TLS) / Group /
Individual (Isotropic & Anisotropic)

Occupancy refinement
Group / Individual

Output (model, maps, statistics)

Pavel Afonine, Ralf Grosse-Kunstleve & Peter Zwart, Lawrence Berkeley Laboratory

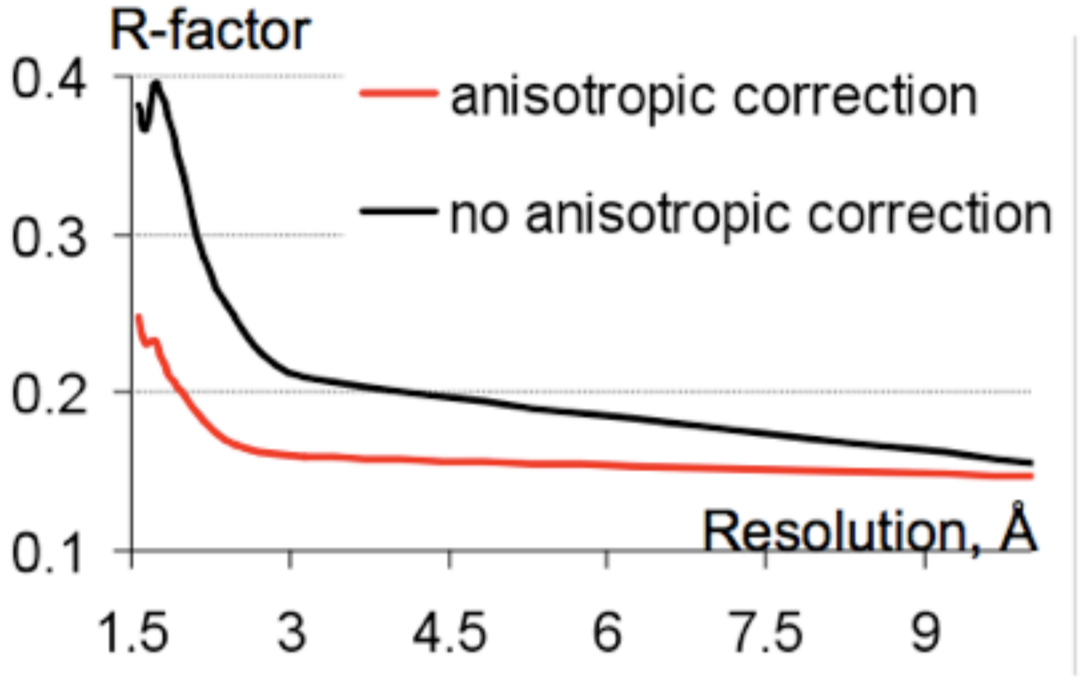
Phenix

Robust Scaling & Bulk Solvent Correction

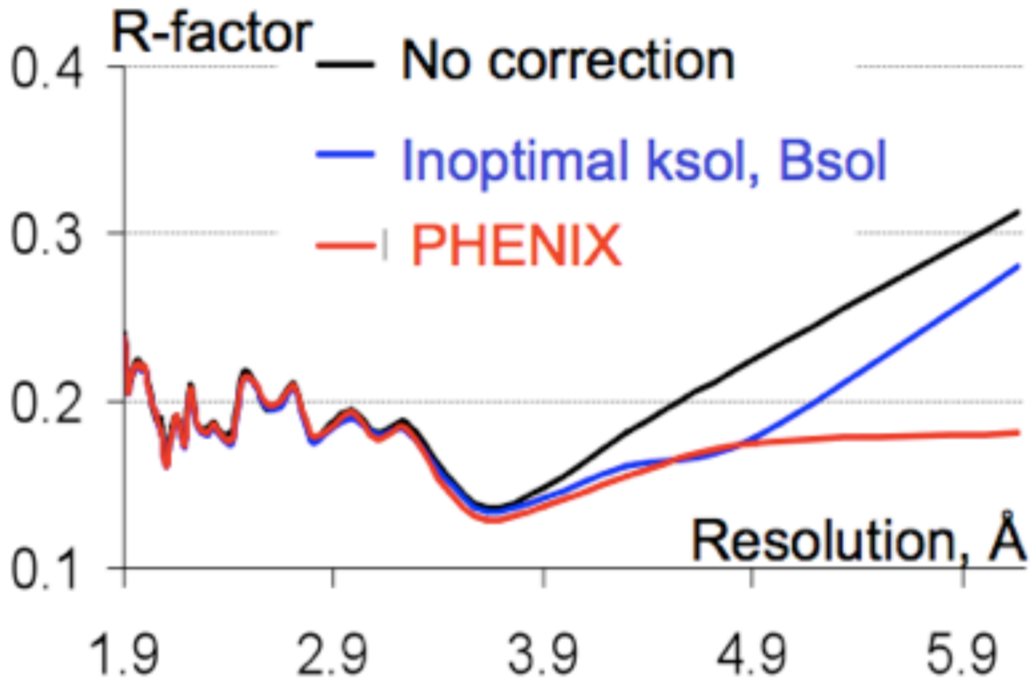
$$\mathbf{F}_{\mathrm{MODEL}} = k_{\mathrm{OVERALL}} e^{-\mathbf{s}\mathbf{U}_{\mathrm{CRYSTAL}}\mathbf{s}^{t}} \left(\mathbf{F}_{\mathrm{CALC\_ATOMS}} + k_{\mathrm{SOL}} e^{-\frac{B_{\mathrm{SOL}}\, s^{2}}{4}} \mathbf{F}_{\mathrm{MASK}} \right)$$

- Bulk solvent scaling uses a grid search with optimization
- Combines both bulks solvent and anisotropic scaling

Anisotropic scaling (PDB: 2mhr)



Effect of Bulk Solvent

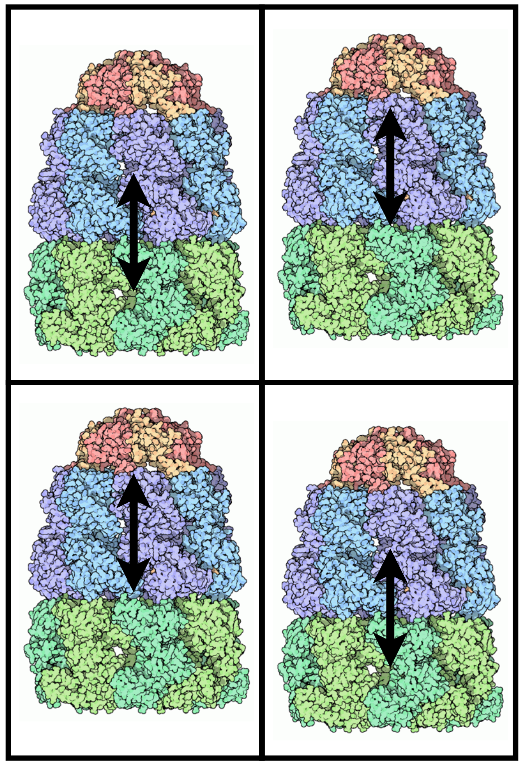


Pavel Afonine, Lawrence Berkeley Laboratory

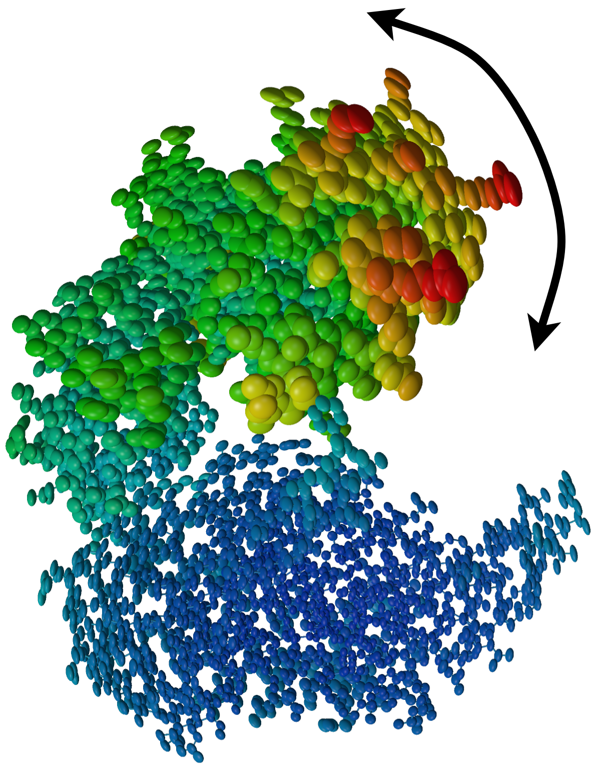
Acta Cryst. 2005, **D61**:850-855.

Modeling Atomic Displacements

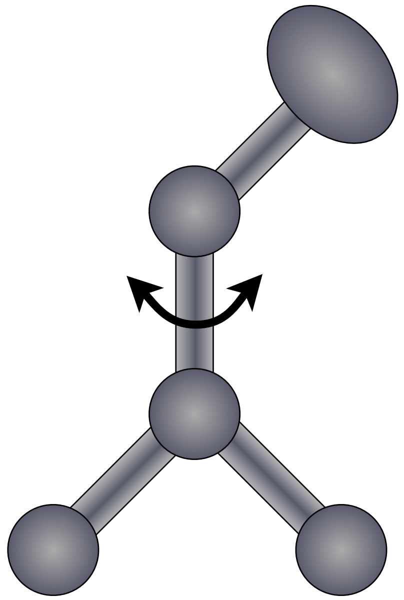
- Atom displacements are typically anisotropic
  - $U_{Total} = U_{Crystal} + U_{Rigid} + U_{Torsion} + U_{Atom}$



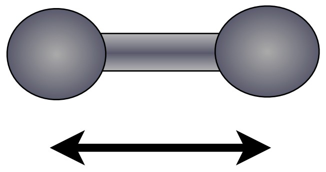
$U_{Crystal}$



U_{Rigid}



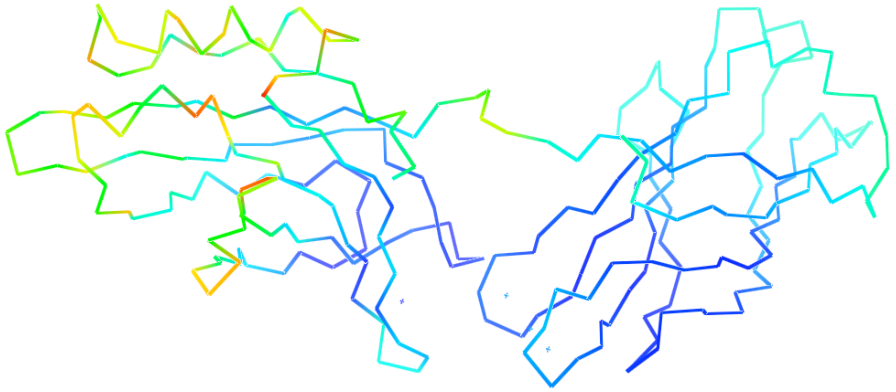
$U_{Torsion}$



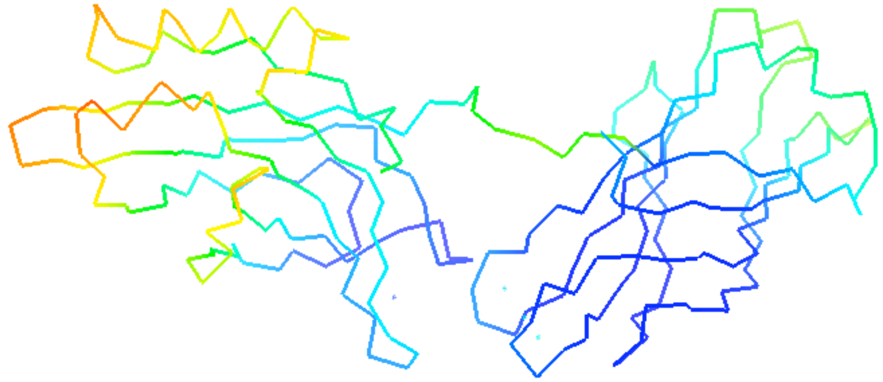
U_{Atom}

Improved ADP Refinement

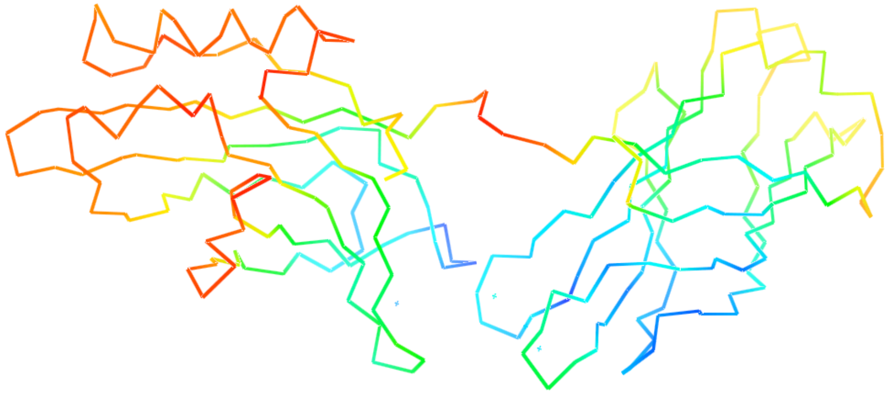
Synaptotagmin, 3.2Å



CNS
R-free=34%
R=29%

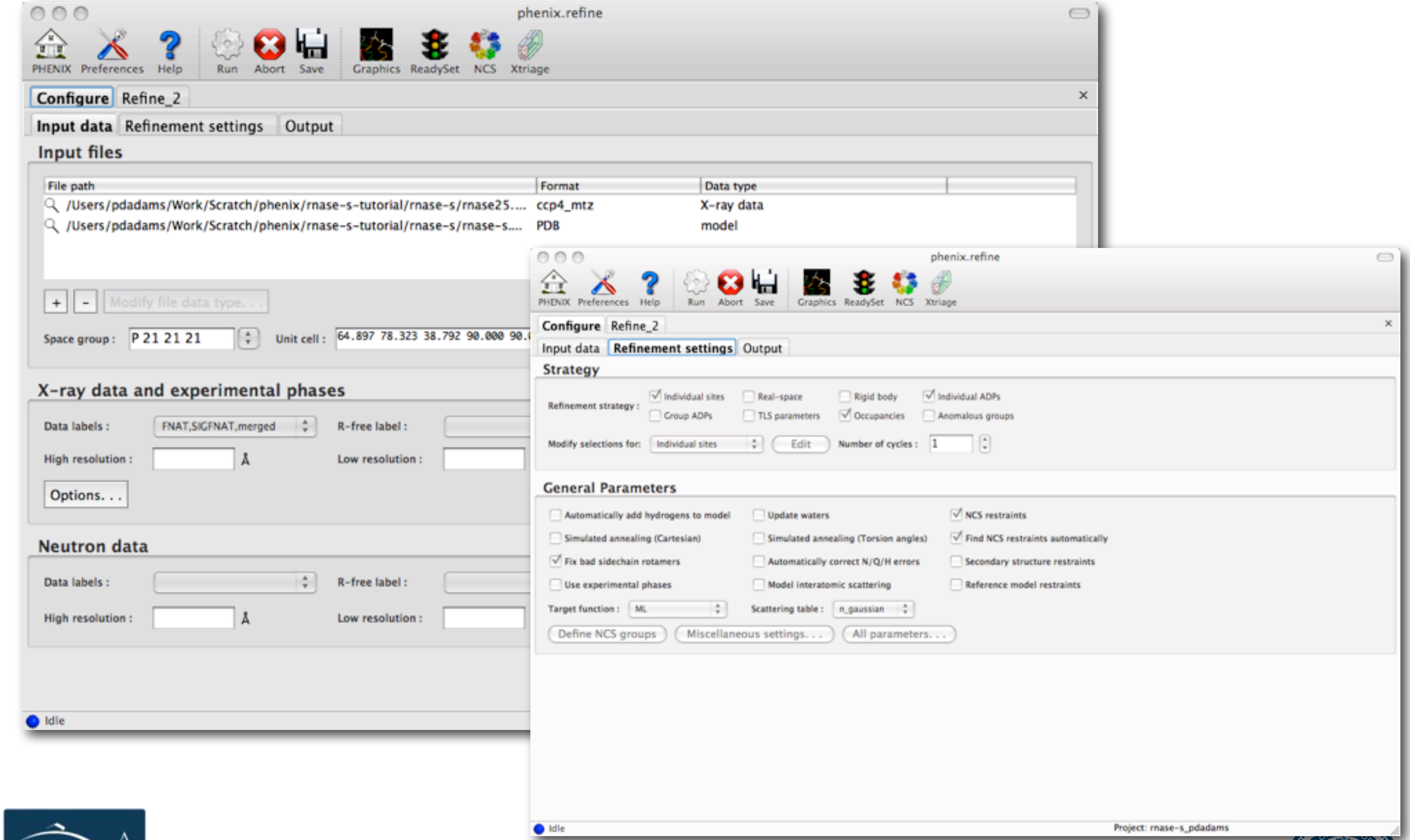


PHENIX – Isotropic restrained ADP
R-free=27.7%
R=24.6%



PHENIX – TLS + Isotropic ADP
R-free=24.4%
R=20.7%

Refinement GUI



Input files

File path	Format	Data type
/Users/pdadams/Work/Scratch/phenix/rnase-s-tutorial/rnase-s/rnase25....	ccp4_mtz	X-ray data
/Users/pdadams/Work/Scratch/phenix/rnase-s-tutorial/rnase-s/rnase-s....	PDB	model

+ - Modify file data type...

Space group : P 21 21 21

Unit cell : 64.897 78.323 38.792 90.000 90.(

X-ray data and experimental phases

Data labels : FNAT,SIGFNAT,merged

R-free label :

High resolution : Å

Low resolution :

Options...

Neutron data

Data labels :

R-free label :

High resolution : Å

Low resolution :

Idle

Strategy

Refinement strategy : ☑ Individual sites □ Real-space □ Rigid body ☑ Individual ADPs □ Group ADPs □ TLS parameters ☑ Occupancies □ Anomalous groups

Modify selections for: Individual sites Edit Number of cycles : 1

General Parameters

□ Automatically add hydrogens to model □ Update waters ☑ NCS restraints

□ Simulated annealing (Cartesian) □ Simulated annealing (Torsion angles) ☑ Find NCS restraints automatically

☑ Fix bad sidechain rotamers □ Automatically correct N/Q/H errors □ Secondary structure restraints

□ Use experimental phases □ Model interatomic scattering □ Reference model restraints

Target function : ML Scattering table : n_gaussian

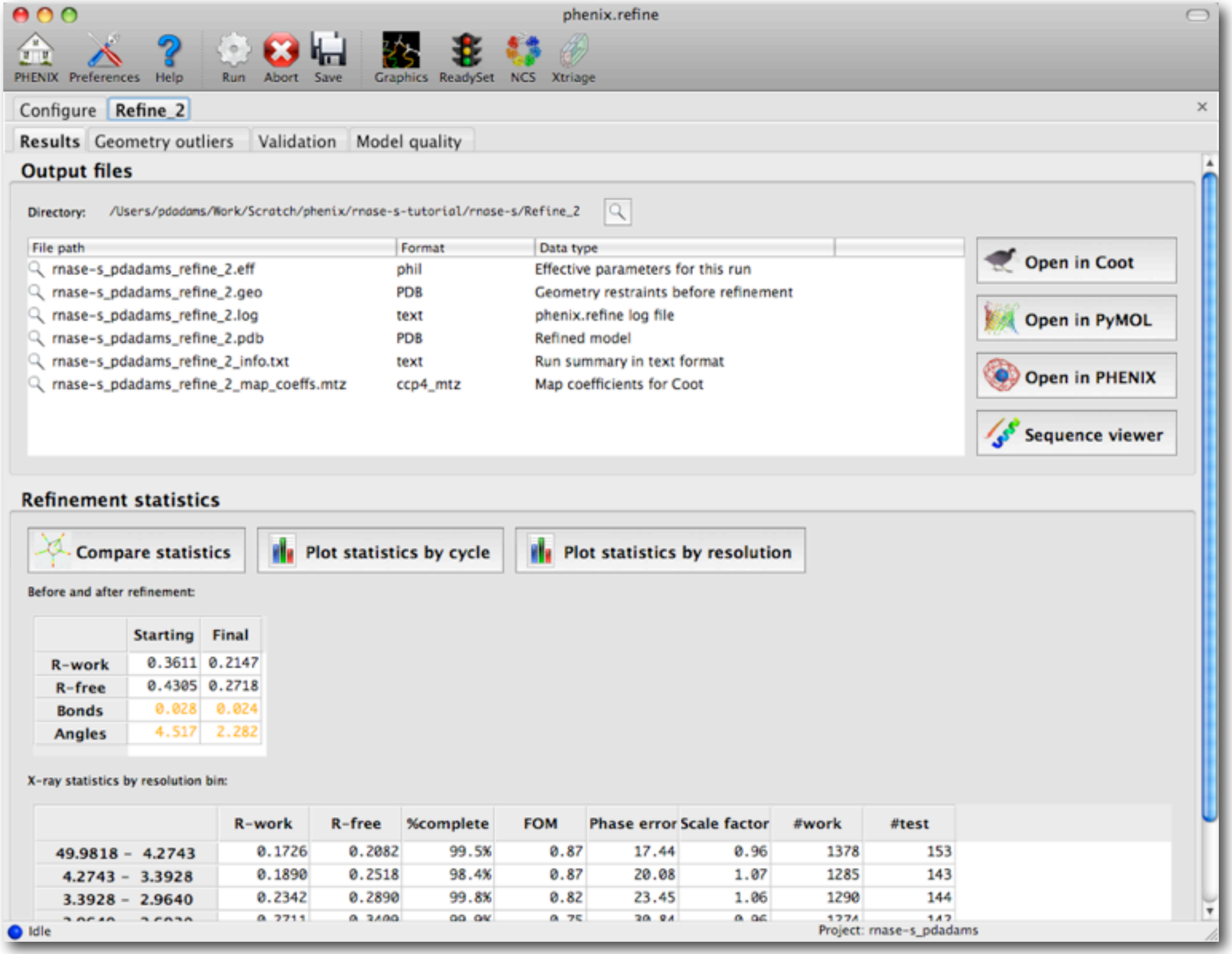
Define NCS groups Miscellaneous settings... All parameters...

Idle Project: rnase-s_pdadams

Phenix

GUI: Nat Echols (LBL)

Results - Summary



Output files

Directory: /Users/pdadams/Work/Scratch/phenix/rnase-s-tutorial/rnase-s/Refine_2

File path	Format	Data type
rnase-s_pdadams_refine_2.eff	phil	Effective parameters for this run
rnase-s_pdadams_refine_2.geo	PDB	Geometry restraints before refinement
rnase-s_pdadams_refine_2.log	text	phenix.refine log file
rnase-s_pdadams_refine_2.pdb	PDB	Refined model
rnase-s_pdadams_refine_2_info.txt	text	Run summary in text format
rnase-s_pdadams_refine_2_map_coeffs.mtz	ccp4_mtz	Map coefficients for Coot

Refinement statistics

Before and after refinement:

	Starting	Final
R-work	0.3611	0.2147
R-free	0.4305	0.2718
Bonds	0.028	0.024
Angles	4.517	2.282

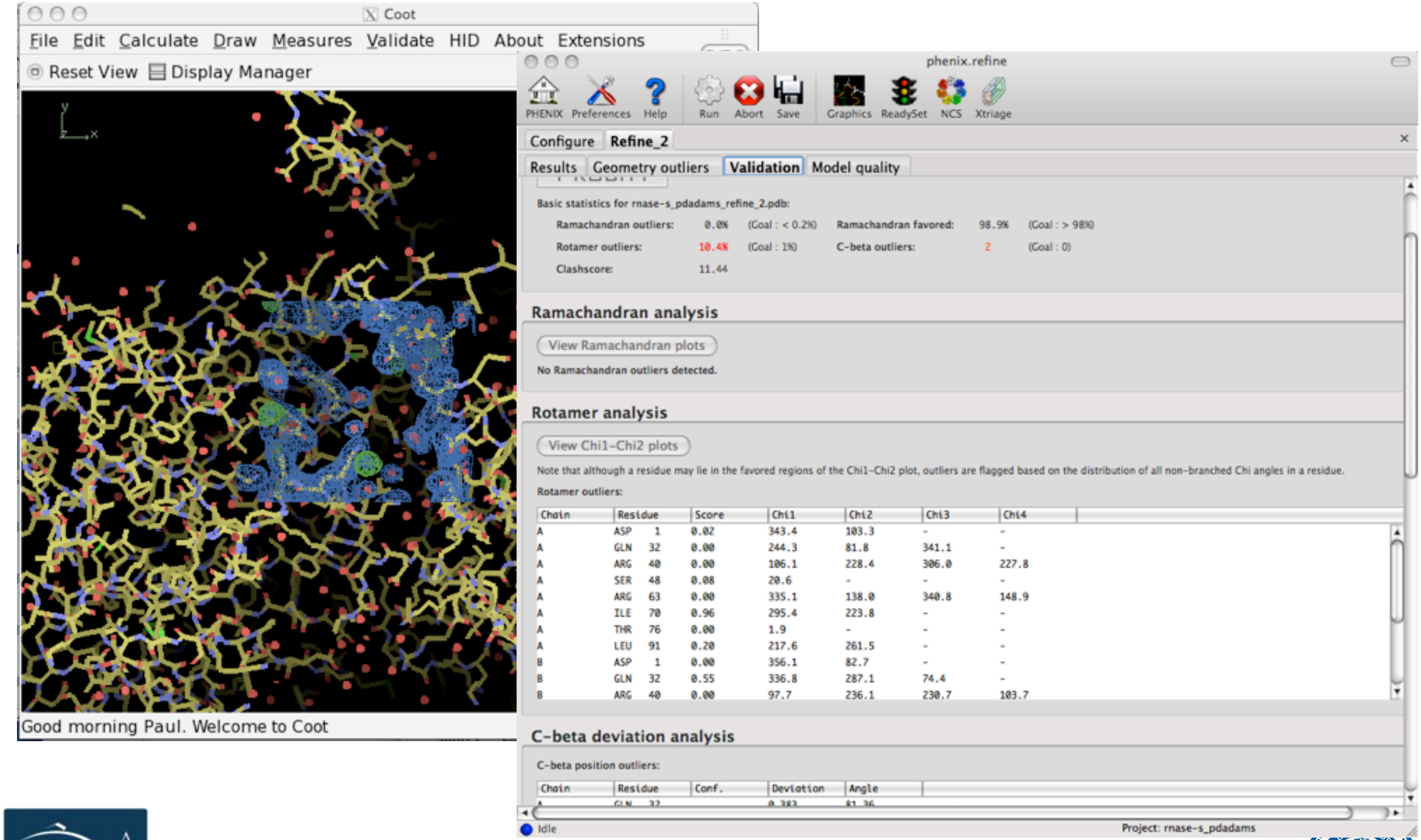
X-ray statistics by resolution bin:

	R-work	R-free	%complete	FOM	Phase error	Scale factor	#work	#test
49.9818 - 4.2743	0.1726	0.2082	99.5%	0.87	17.44	0.96	1378	153
4.2743 - 3.3928	0.1890	0.2518	98.4%	0.87	20.08	1.07	1285	143
3.3928 - 2.9640	0.2342	0.2890	99.8%	0.82	23.45	1.06	1290	144
2.9640 - 2.6020	0.2711	0.3409	99.9%	0.75	30.84	0.96	1274	142

Project: rnase-s_pdadams

Phenix

Results - Rebuilding and Validation



Basic statistics for rnase-s_pdadams_refine_2.pdb:

Ramachandran outliers:	0.0%	(Goal : < 0.2%)	Ramachandran favored:	98.9%	(Goal : > 98%)
Rotamer outliers:	10.4%	(Goal : 1%)	C-beta outliers:	2	(Goal : 0)
Clashscore:	11.44				

Ramachandran analysis

No Ramachandran outliers detected.

Rotamer analysis

Note that although a residue may lie in the favored regions of the Chi1-Chi2 plot, outliers are flagged based on the distribution of all non-branched Chi angles in a residue.

Rotamer outliers:

Chain	Residue	Score	Chi1	Chi2	Chi3	Chi4
A	ASP 1	0.02	343.4	103.3	-	-
A	GLN 32	0.00	244.3	81.8	341.1	-
A	ARG 40	0.00	106.1	228.4	306.0	227.8
A	SER 48	0.08	20.6	-	-	-
A	ARG 63	0.00	335.1	138.0	340.8	148.9
A	ILE 70	0.96	295.4	223.8	-	-
A	THR 76	0.00	1.9	-	-	-
A	LEU 91	0.20	217.6	261.5	-	-
B	ASP 1	0.00	356.1	82.7	-	-
B	GLN 32	0.55	336.8	287.1	74.4	-
B	ARG 40	0.00	97.7	236.1	230.7	103.7

C-beta deviation analysis

C-beta position outliers:

Chain	Residue	Conf.	Deviation	Angle
A	GLN 32		0.383	81.36

Phenix

GUI: Nat Echols (LBL)

Model Validation

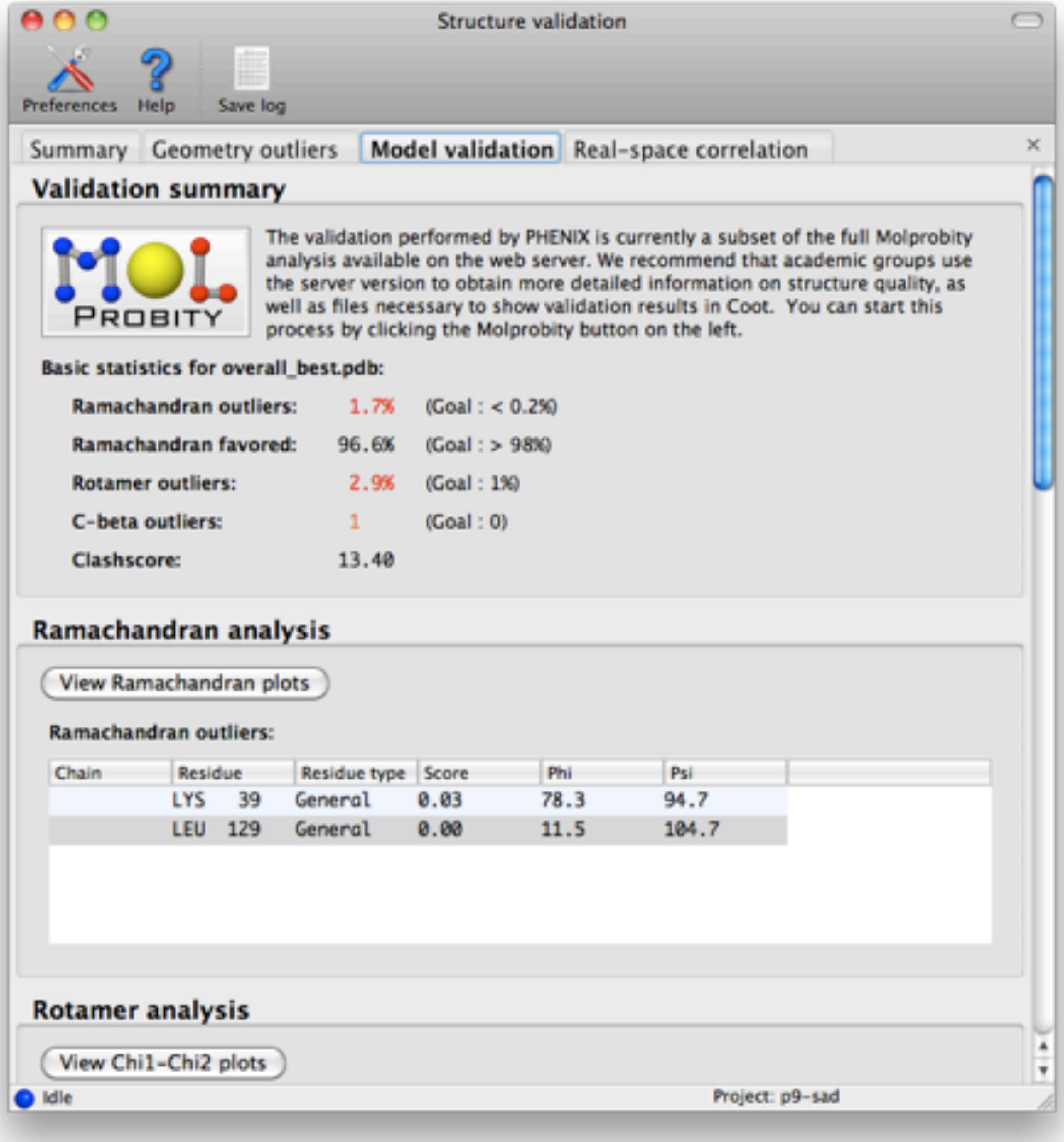
- In science we construct models to explain experimental observations
- We must always ask if the model is correct, or as correct as it can be given the experimental uncertainties
  - Does the model fit the experimental data?
  - Does the model confirm prior knowledge?
  - Does the model predict things that we can measure? (typically leads to other experiments)

Validation

- Global validators:
    - R-factors (e.g. Free-R-factor)
    - Overall deviations from ideal bond lengths and bond angles
- Local validators:
    - Deviations from ideal geometry
    - Deviations from known distributions of backbone torsion angles (protein)
    - Deviations from known distributions of side chain conformations (protein)
    - Local fit of model to electron density
    - Contacts between atoms (unlikely chemical interactions, too close atoms)

Validation

- Outlier lists recenter Coot view; Probe dots automatically loaded
  - optional real-space correlation (if reflections available), with B-factor analysis



Structure validation

Validation summary

The validation performed by PHENIX is currently a subset of the full Molprobity analysis available on the web server. We recommend that academic groups use the server version to obtain more detailed information on structure quality, as well as files necessary to show validation results in Coot. You can start this process by clicking the Molprobity button on the left.

Basic statistics for overall_best.pdb:

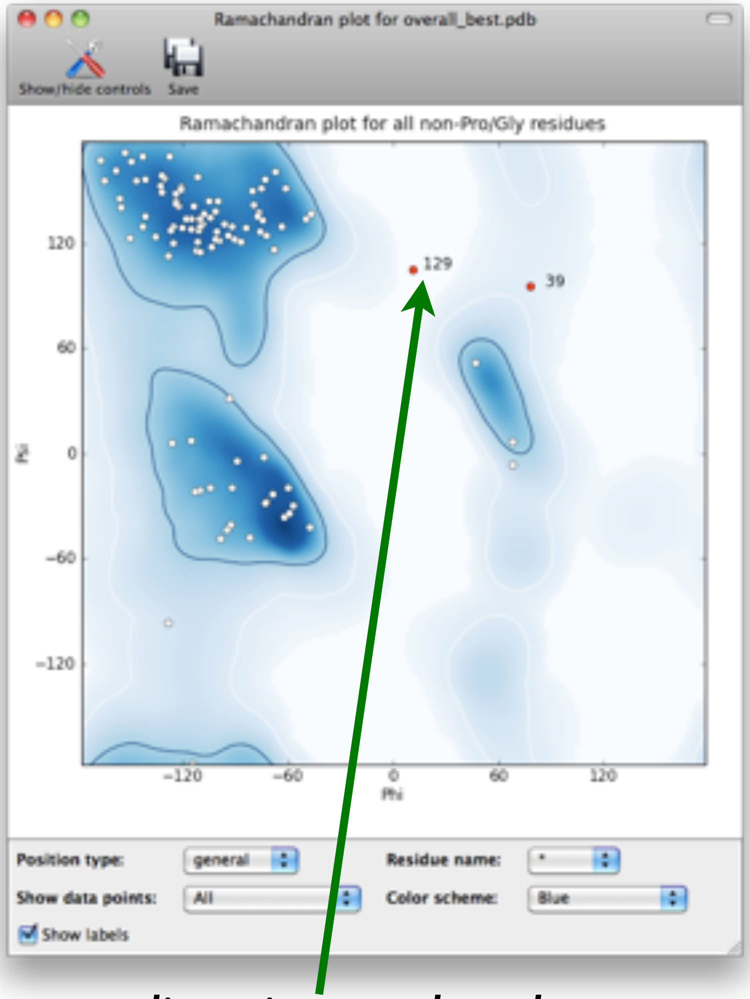
Ramachandran outliers:	1.7%	(Goal : < 0.2%)
Ramachandran favored:	96.6%	(Goal : > 98%)
Rotamer outliers:	2.9%	(Goal : 1%)
C-beta outliers:	1	(Goal : 0)
Clashscore:	13.40	

Ramachandran analysis

Ramachandran outliers:

Chain	Residue	Residue type	Score	Phi	Psi
	LYS 39	General	0.03	78.3	94.7
	LEU 129	General	0.00	11.5	104.7

Rotamer analysis



Ramachandran plot for overall_best.pdb

Position type: general

Residue name: *

Show data points: All

Color scheme: Blue

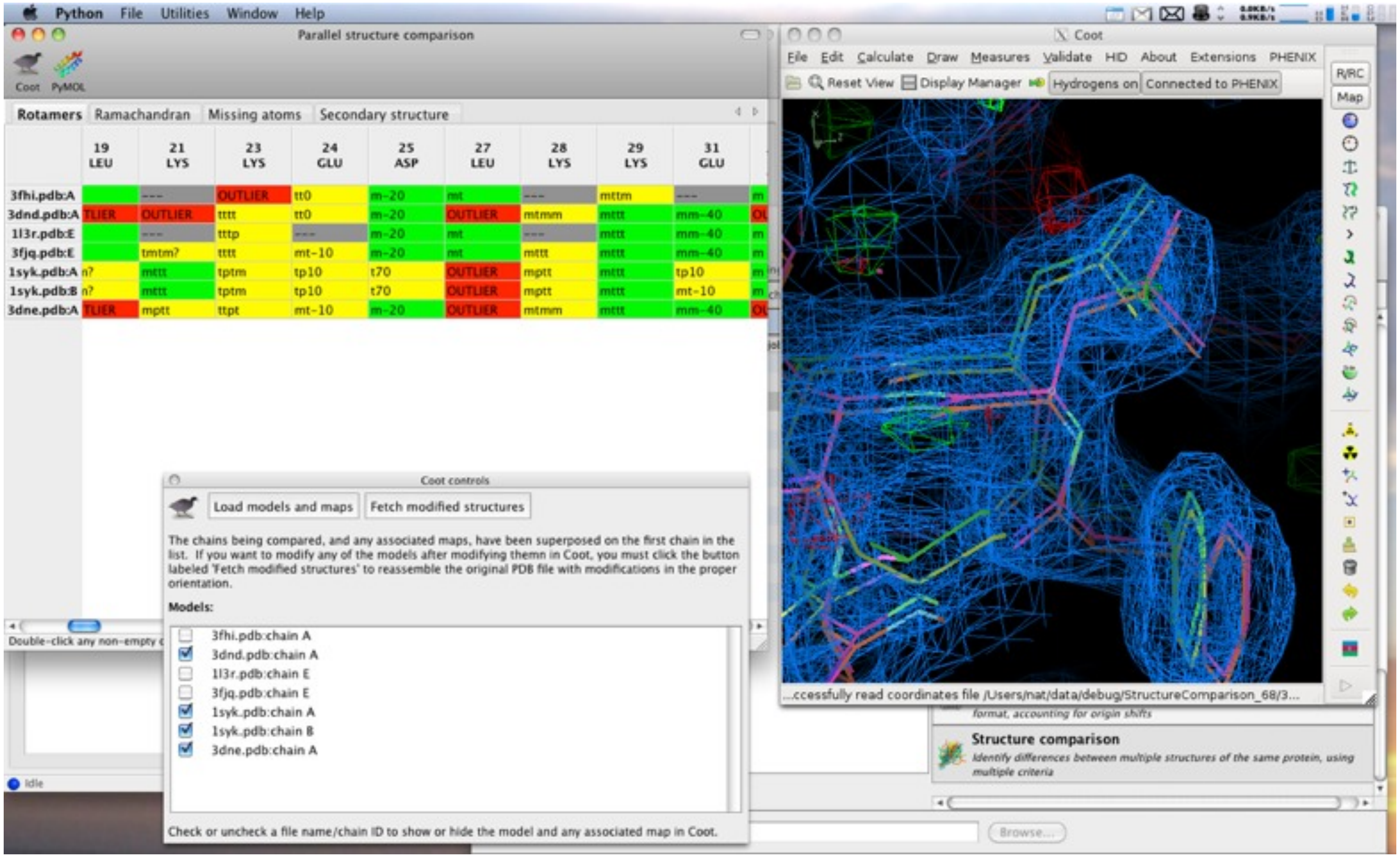
☑ Show labels

outliers in graphs also recenter Coot

MolProbity: Richardson Lab, Duke
GUI: Nat Echols (LBL)

Parallel validation of multiple structures

- Identifies points of difference between structures of the same protein, with optional map superpositioning



	19 LEU	21 LYS	23 LYS	24 GLU	25 ASP	27 LEU	28 LYS	29 LYS	31 GLU	
3fhi.pdb:A		---	OUTLIER	tt0	m-20	mt	---	mttm	---	m
3dnd.pdb:A	TLIER	OUTLIER	tttt	tt0	m-20	OUTLIER	mtmm	mttt	mm-40	OL
1l3r.pdb:E		---	tttp	---	m-20	mt	---	mttt	mm-40	m
3fjq.pdb:E		tmtm?	tttt	mt-10	m-20	mt	mttt	mttt	mm-40	m
1syk.pdb:A	n?	mttt	tptm	tp10	t70	OUTLIER	mptt	mttt	tp10	m
1syk.pdb:B	n?	mttt	tptm	tp10	t70	OUTLIER	mptt	mttt	mt-10	m
3dne.pdb:A	TLIER	mptt	ttpt	mt-10	m-20	OUTLIER	mtmm	mttt	mm-40	OL

Double-click any non-empty c

Idle

The chains being compared, and any associated maps, have been superposed on the first chain in the list. If you want to modify any of the models after modifying themn in Coot, you must click the button labeled 'Fetch modified structures' to reassemble the original PDB file with modifications in the proper orientation.

Models:

- ☐ 3fhi.pdb:chain A
- ☑ 3dnd.pdb:chain A
- ☐ 1l3r.pdb:chain E
- ☐ 3fjq.pdb:chain E
- ☑ 1syk.pdb:chain A
- ☑ 1syk.pdb:chain B
- ☑ 3dne.pdb:chain A

Check or uncheck a file name/chain ID to show or hide the model and any associated map in Coot.

...ccessfully read coordinates file /Users/nat/data/debug/StructureComparison_68/3...

format, accounting for origin shifts

Structure comparison

Identify differences between multiple structures of the same protein, using multiple criteria

Phenix

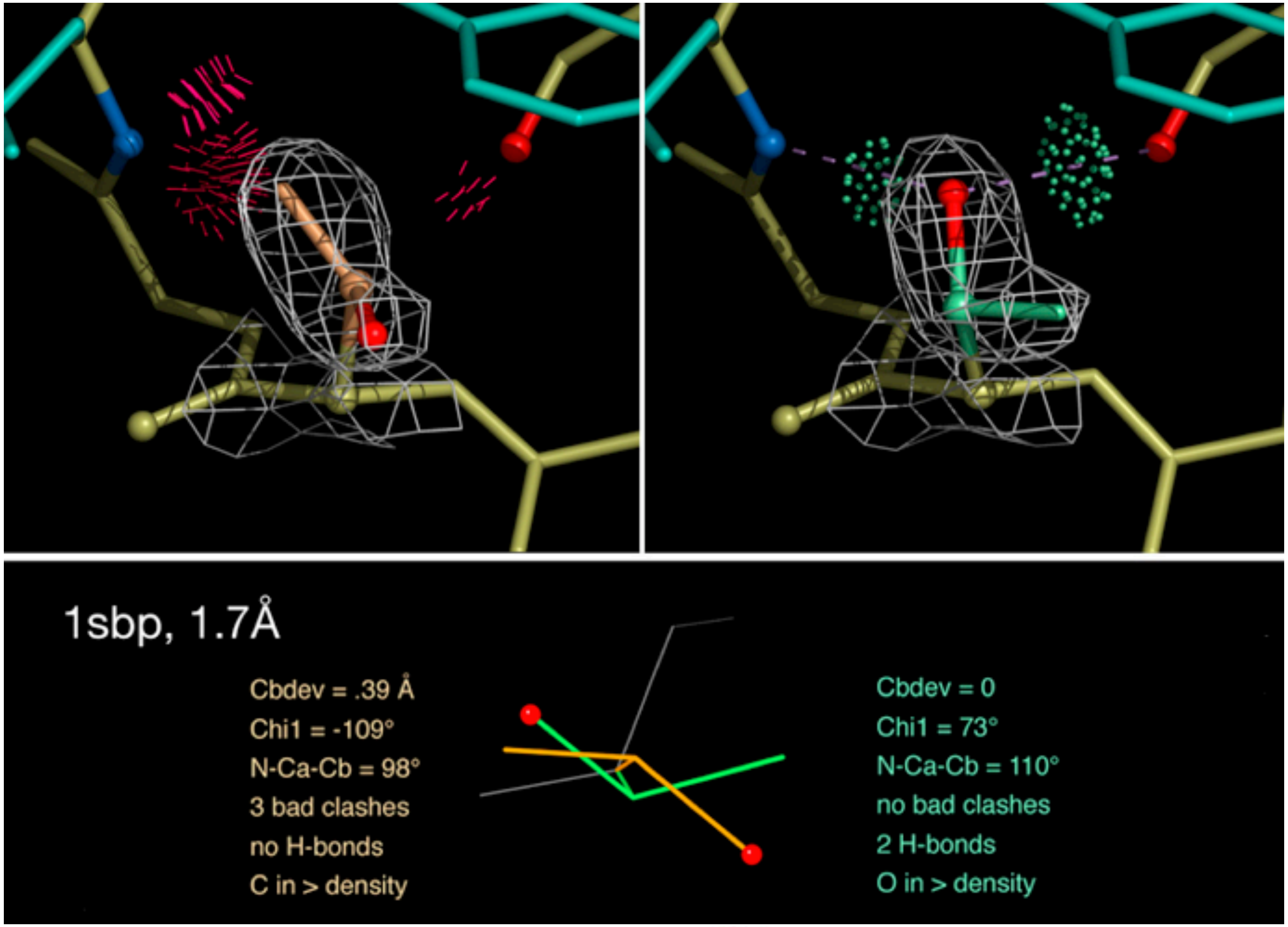
Nat Echols, Nigel Moriarty, Pavel Afonine, Ralf Grosse-Kunstleve (LBL) & Herb Klei (BMS)

Active use of Validation Measures

- Automated fixing of rotamers
- Automated flipping of side chains
- Accounting for local context
- Using prior knowledge about secondary structure as restraints
- Using similar high resolution structures as restraints

Automated Rotamer Fixing

- Electron density can often be ambiguous for some residues (e.g. Leu)
- Methods developed for validation (identifying incorrect rotamers) can be used to automatically fix problem residues



1sbp, 1.7Å

Cbdev = .39 Å
Chi1 = -109°
N-Ca-Cb = 98°
3 bad clashes
no H-bonds
C in > density

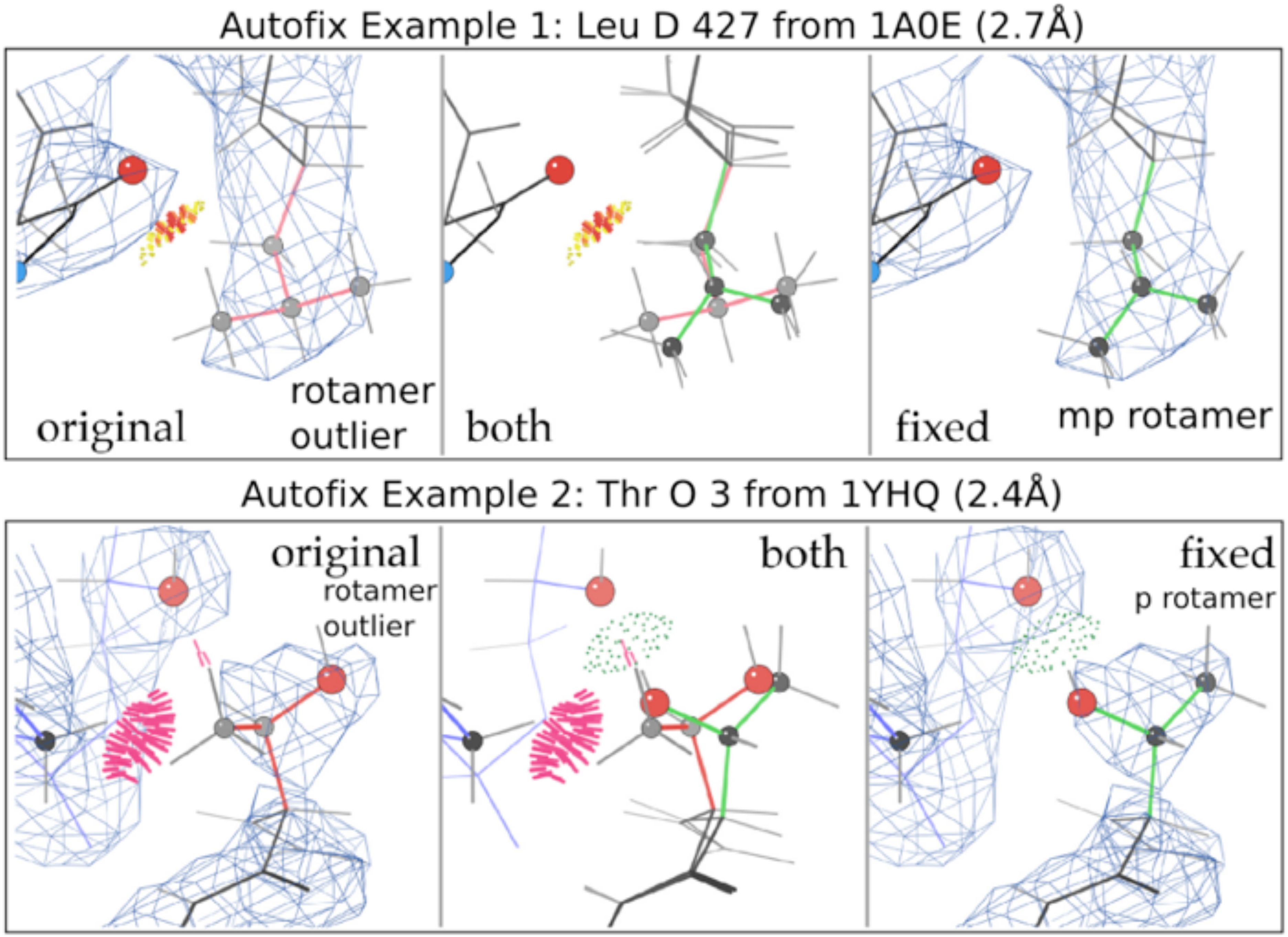
Cbdev = 0
Chi1 = 73°
N-Ca-Cb = 110°
no bad clashes
2 H-bonds
O in > density

BERKELEY LAB

Phenix

Jeff Headd, Duke University

Automated Rotamer Fixing


Autofix Example 1: Leu D 427 from 1A0E (2.7Å)
original
rotamer
outlier
both
fixed
mp rotamer
Autofix Example 2: Thr O 3 from 1YHQ (2.4Å)
original
rotamer
outlier
both
fixed
p rotamer

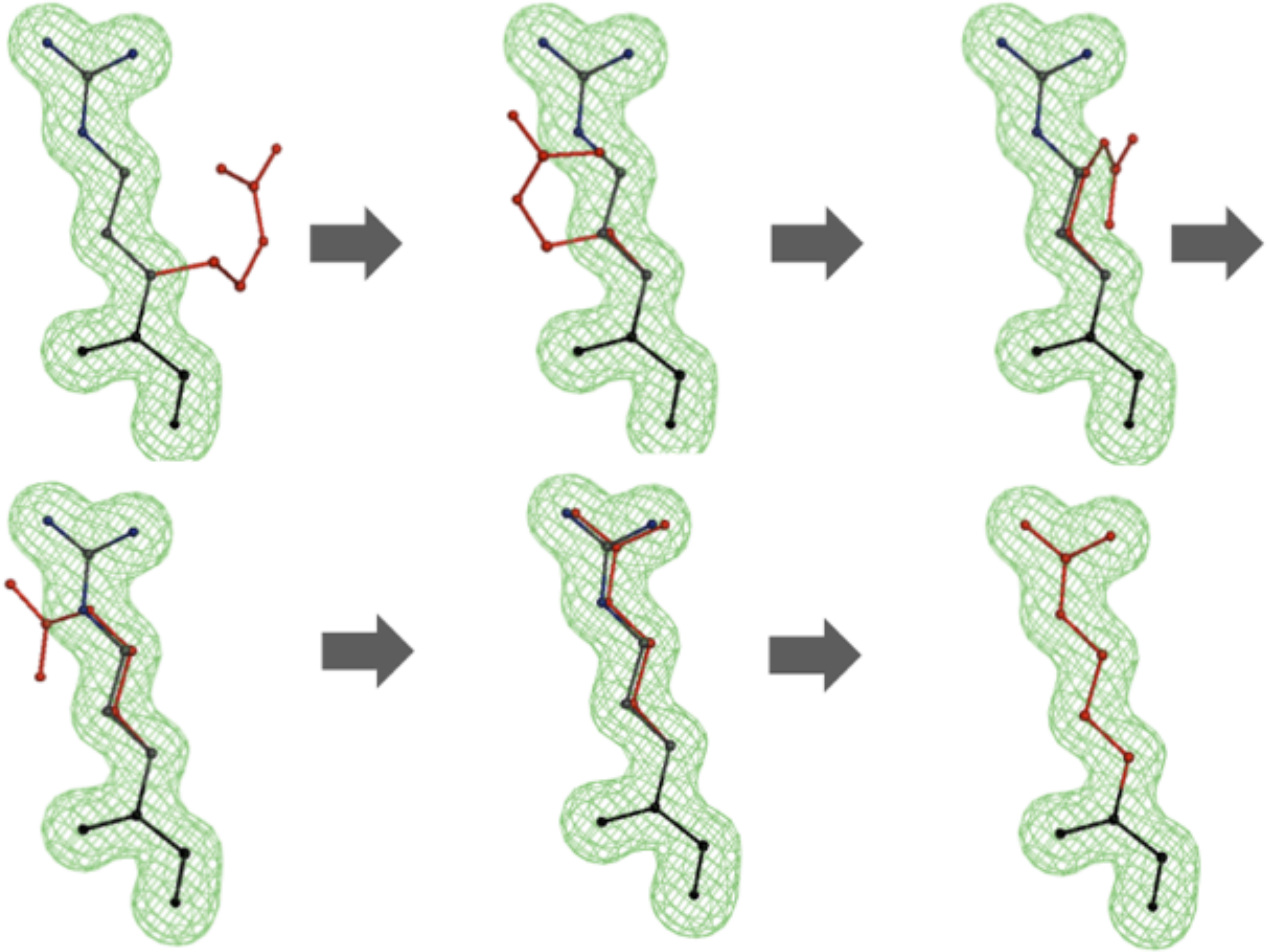
Headd JJ, Immormino RM, Keedy DA, Emsley P, Richardson DC, Richardson JS. Autofix for backward-fit sidechains: using MolProbity and real-space refinement to put misfits in their place. J Struct Funct Genomics. 2009 Mar;10(1):83-93.

Phenix

Jeff Headd, Duke University

Automated Rotamer Fixing in Refinement

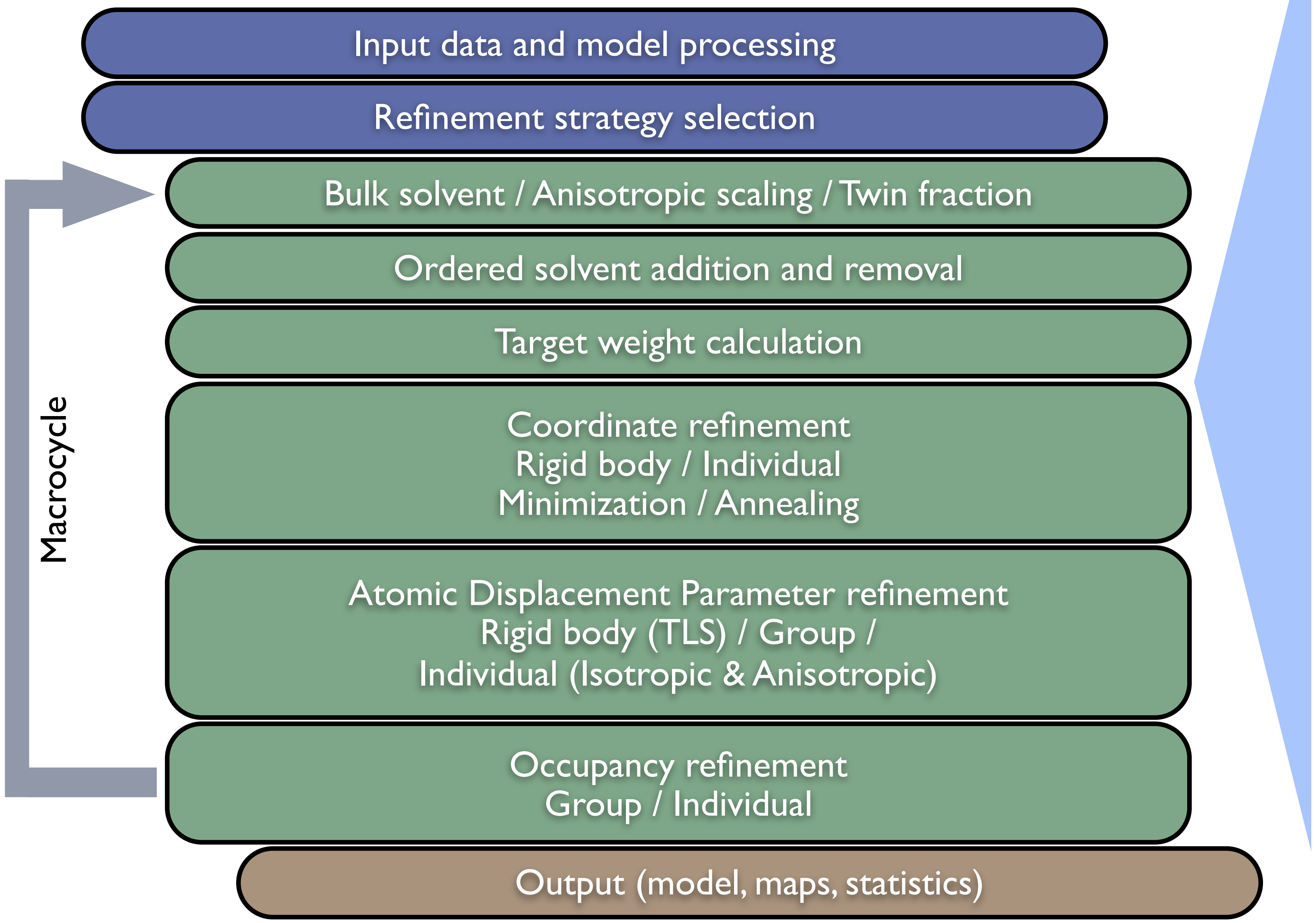
- Assessment of local quality of side chains by comparison to rotamer library
- Torsion angle search against density with real space refinement



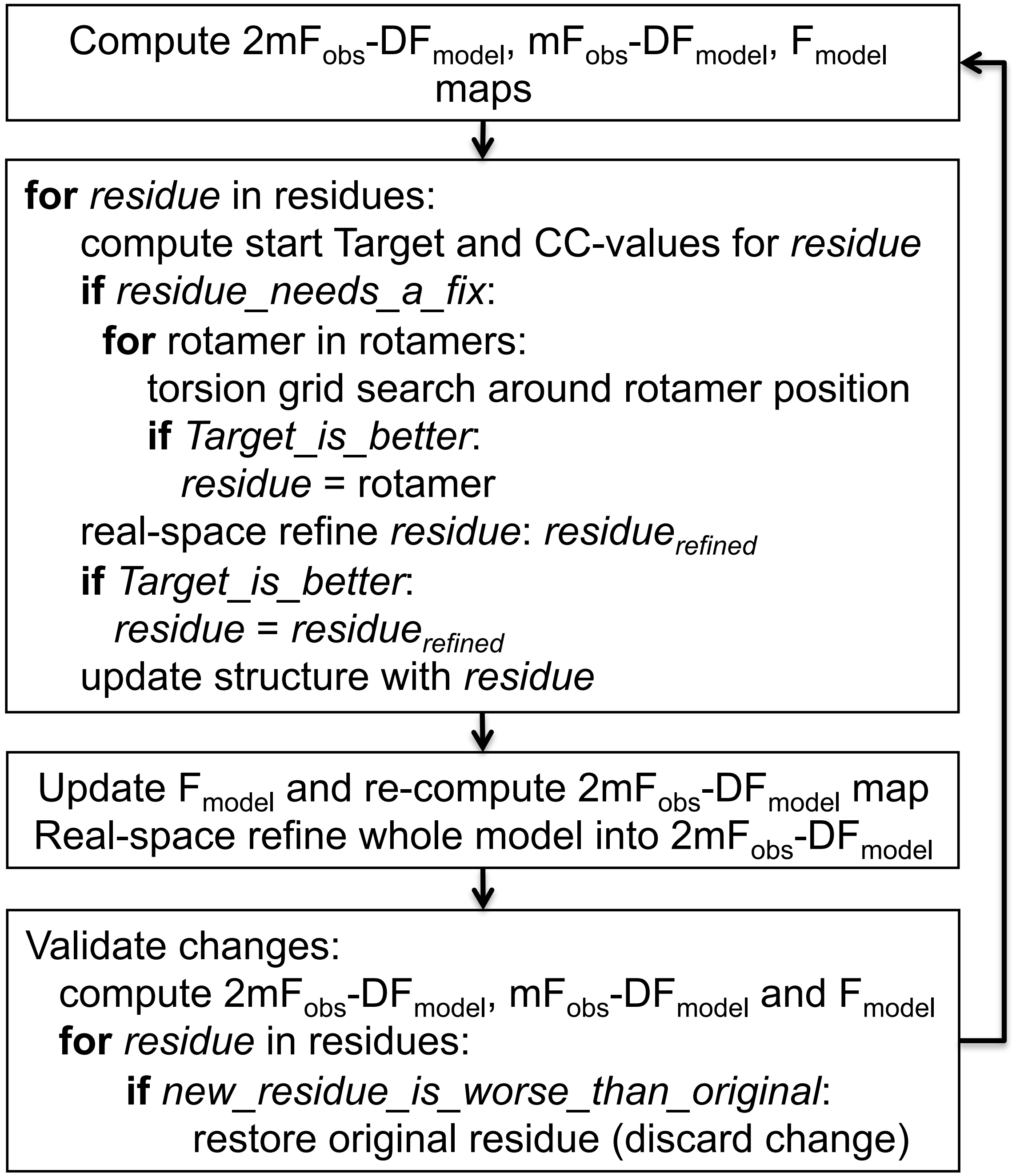
Phenix

Pavel Afonine, Ralf Grosse-Kunstleve, Jeff Headd

Protocol



Input data and model processing
Refinement strategy selection
Bulk solvent / Anisotropic scaling / Twin fraction
Ordered solvent addition and removal
Target weight calculation
Coordinate refinement
Rigid body / Individual
Minimization / Annealing
Atomic Displacement Parameter refinement
Rigid body (TLS) / Group /
Individual (Isotropic & Anisotropic)
Occupancy refinement
Group / Individual
Output (model, maps, statistics)
Macrocycle



```
Compute $2mF_{obs}$-$DF_{model}$, $mF_{obs}$-$DF_{model}$, $F_{model}$ maps

for residue in residues:
    compute start Target and CC-values for residue
    if residue_needs_a_fix:
      for rotamer in rotamers:
          torsion grid search around rotamer position
          if Target_is_better:
             residue = rotamer
    real-space refine residue: $residue_{refined}$
    if Target_is_better:
       residue = $residue_{refined}$
    update structure with residue

Update $F_{model}$ and re-compute $2mF_{obs}$-$DF_{model}$ map
Real-space refine whole model into $2mF_{obs}$-$DF_{model}$

Validate changes:
  compute $2mF_{obs}$-$DF_{model}$, $mF_{obs}$-$DF_{model}$ and $F_{model}$
  for residue in residues:
      if new_residue_is_worse_than_original:
          restore original residue (discard change)
```

```
% phenix.refine model.pdb data.hkl fix_rotamers=true
```

☑ Fix bad sidechain rotamers

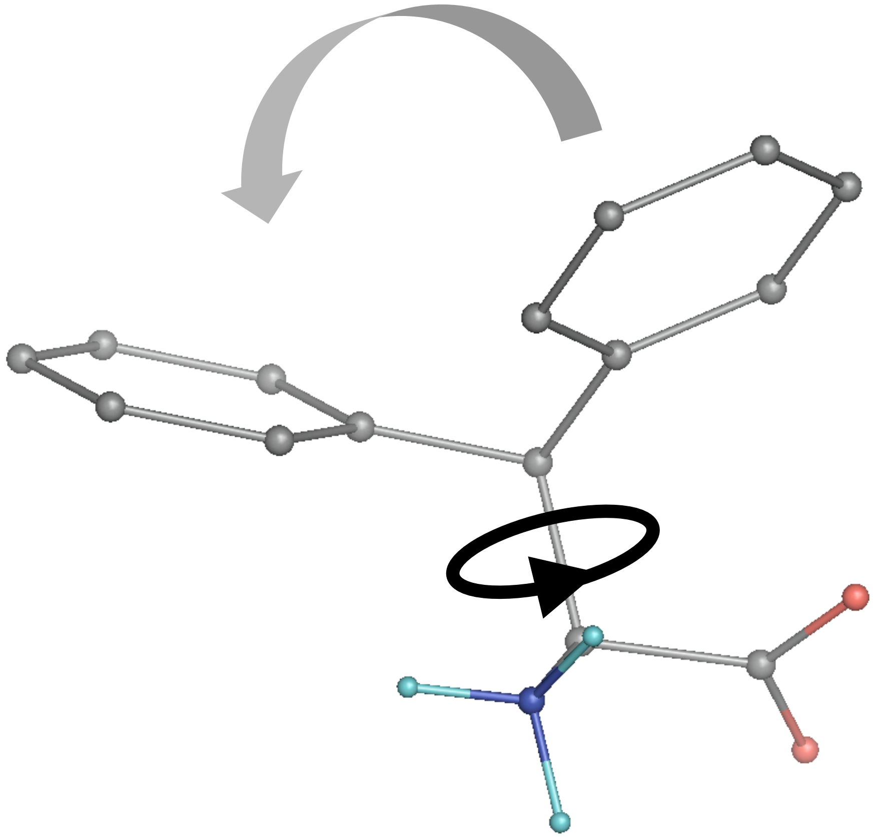
Pavel Afonine, LBL
Nat Echols, LBL

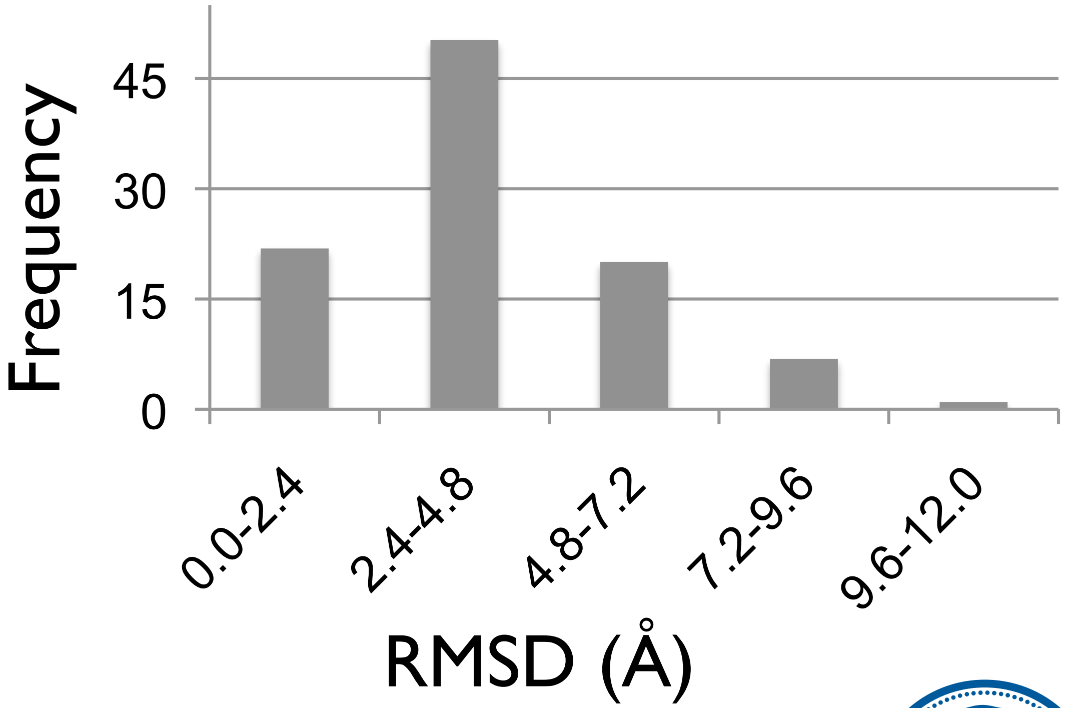
Testing Performance

Test refinement of 150 structures from PDB in resolution range 1.5-3.0Å:

- Refine original models
  - Basic refinement
  - Basic refinement + local real-space refinement
- *Generate distorted models:*
  - Remove water
  - For each residue select the most distant rotamer
  - Quick geometry regularization to remove bad clashes
- Refine distorted models
  - Basic refinement
  - Basic refinement + Simulated Annealing
  - Basic refinement + local real-space refinement

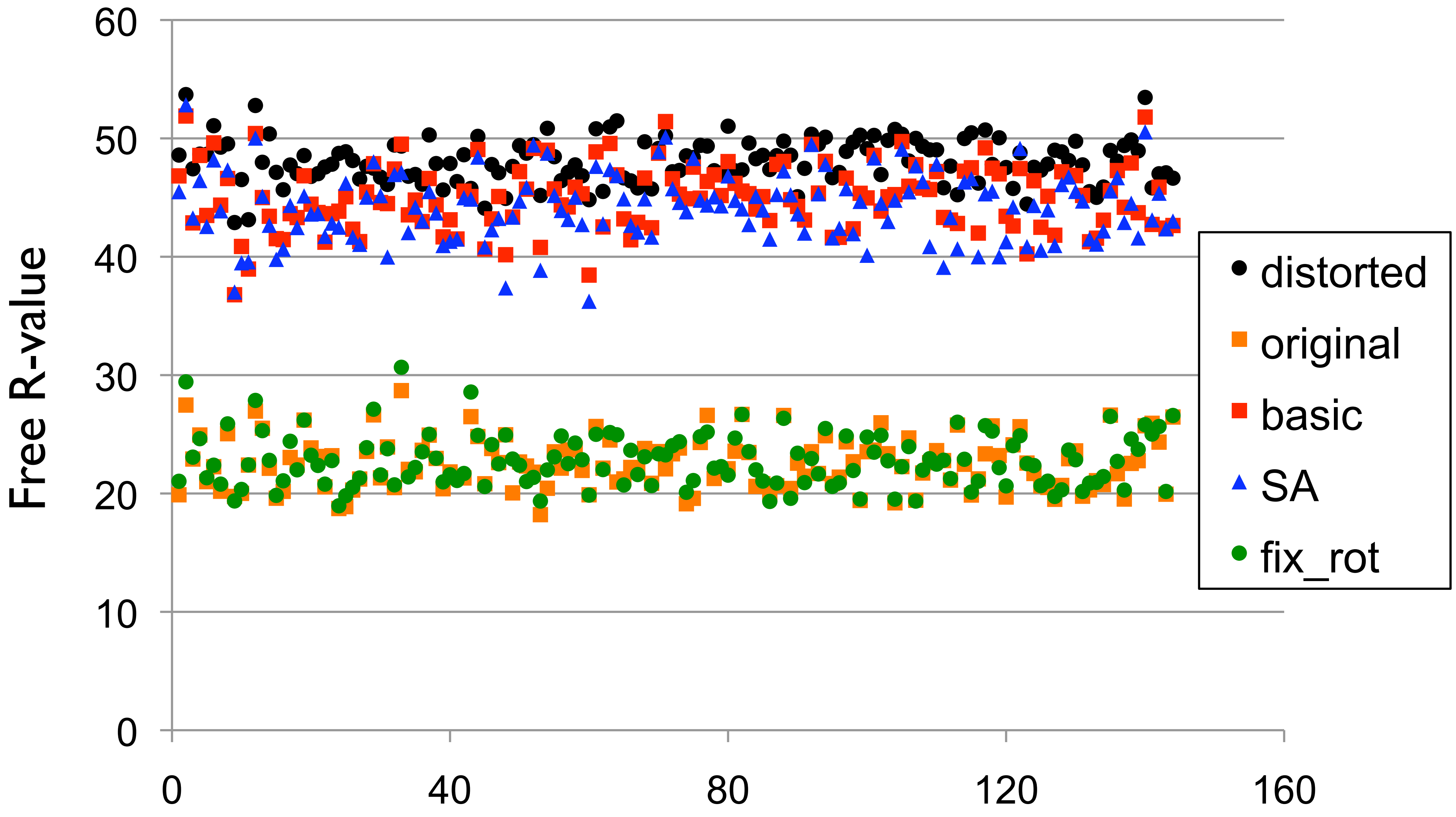
(Where basic refinement is individual coordinates, ADPs, occupancies, and solvent model update)





Pavel Afonine, LBL

Refinement of Distorted Models

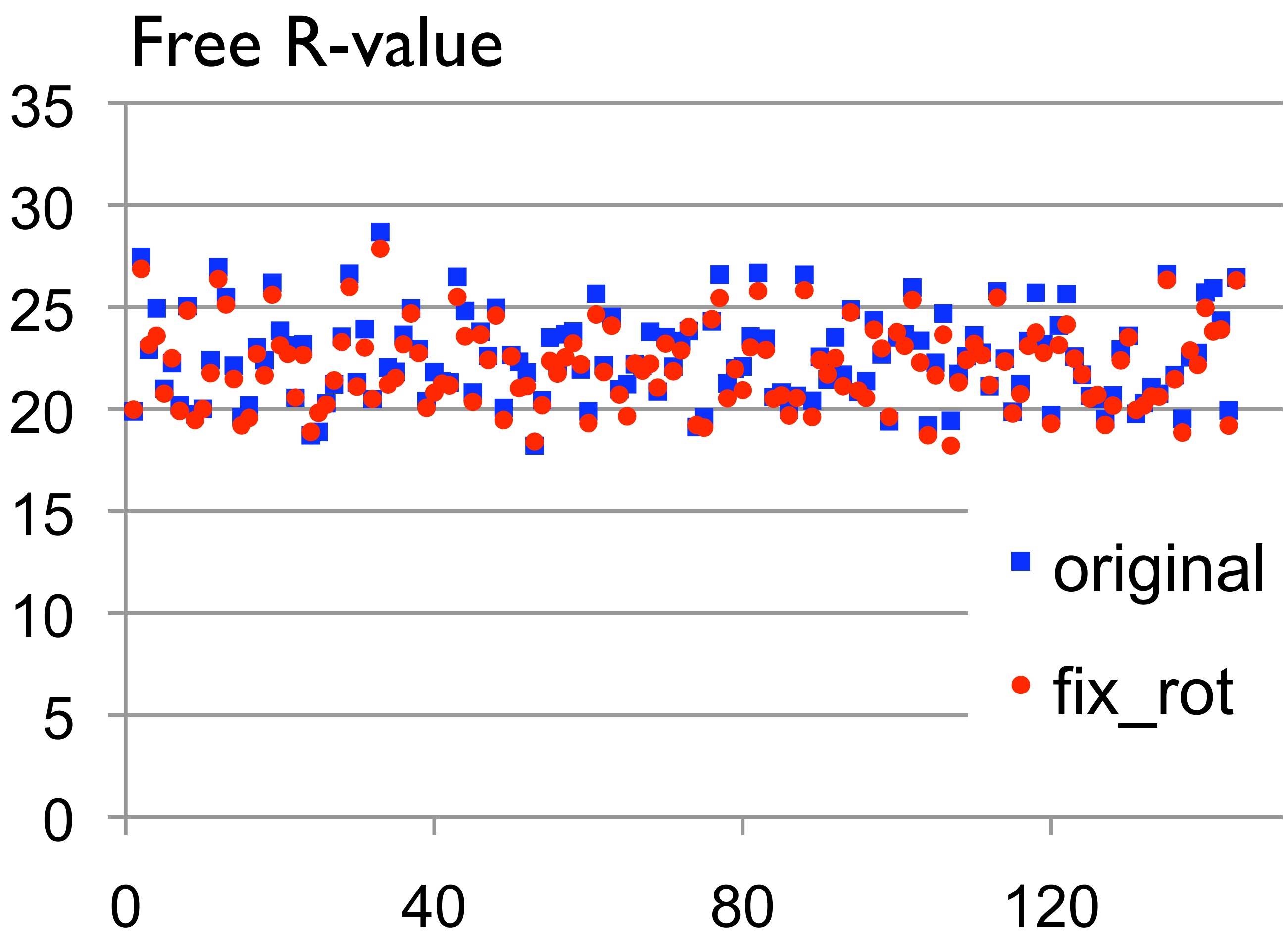


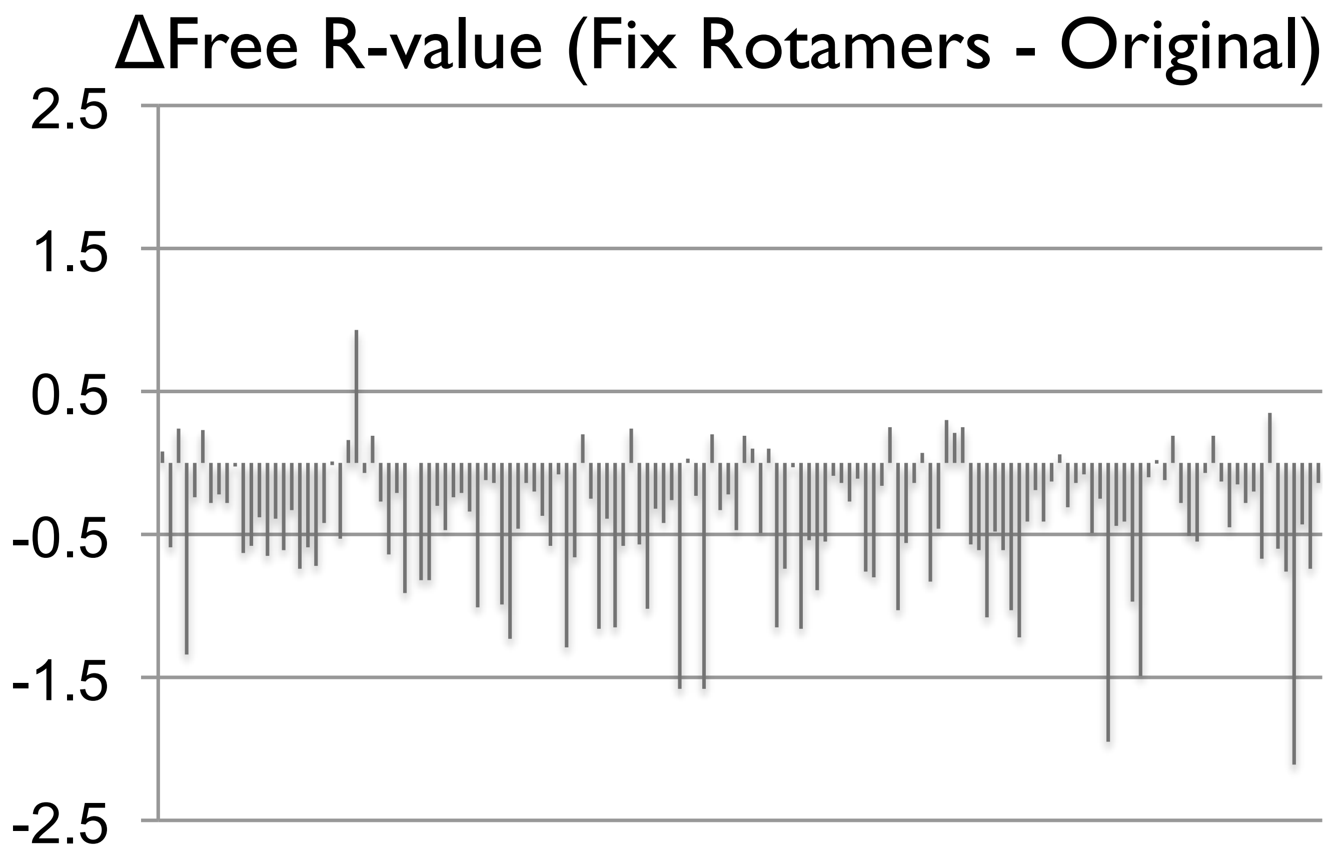
- Errors in rotamers are difficult to fix using gradient methods or simulated annealing
- Local searching and real space refinement can recover the correct rotamers in many cases

Phenix

Pavel Afonine, LBL

Refinement of Original Models

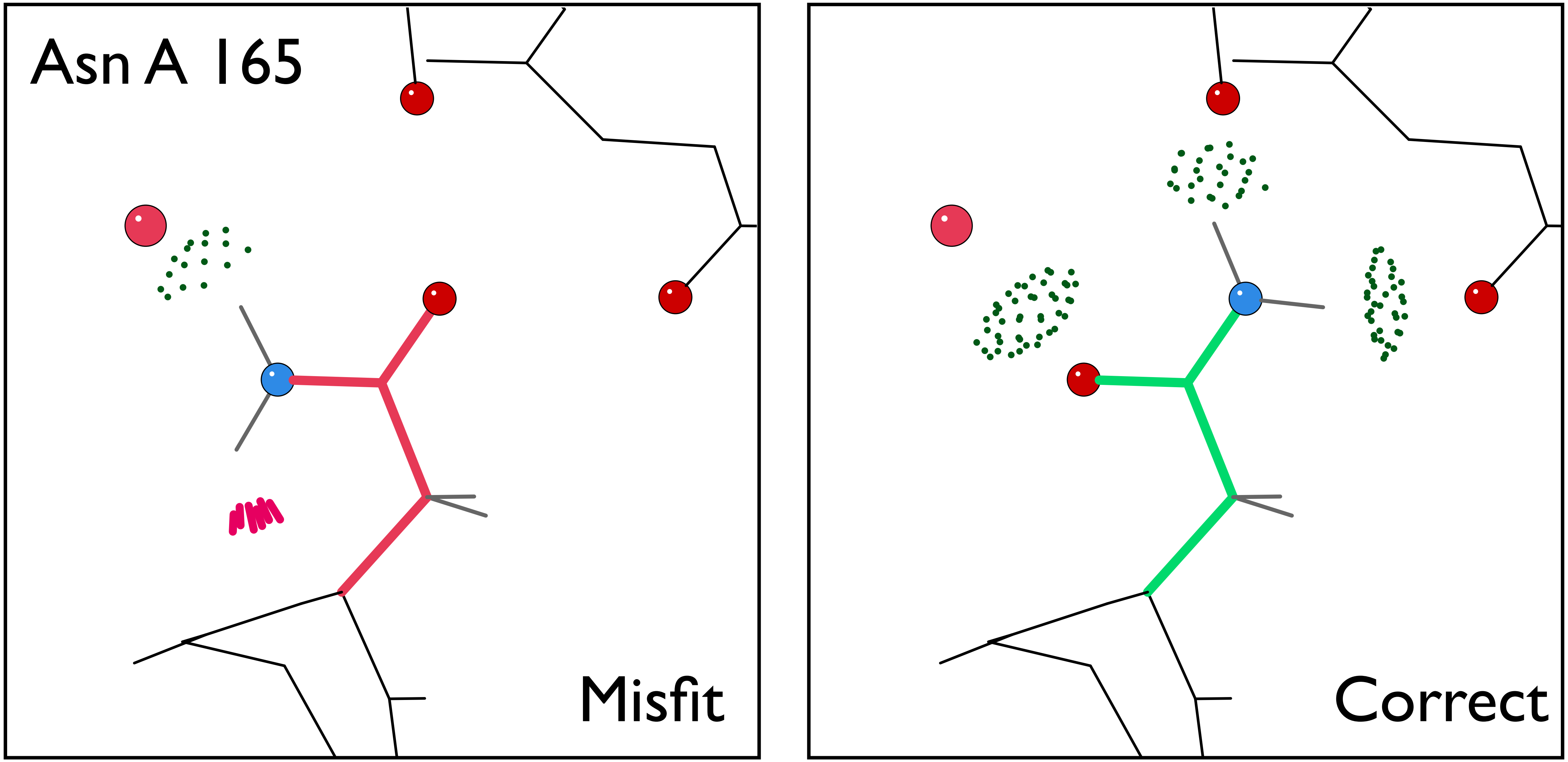

Free R-value


ΔFree R-value (Fix Rotamers - Original)

- Refinement with automated rotamer fixing typically improves free R-values
- Many structures in the PDB could have multiple rotamer errors that can be corrected
- More analysis is required (e.g. impact at low resolution)

Automated Asn/Gln/His Corrections

- Automatically detect and correct flipped N/Q/H residues at each macrocycle
- Uses MolProbity/Reduce methodology (H-bonds, clashes) to determine correct orientation



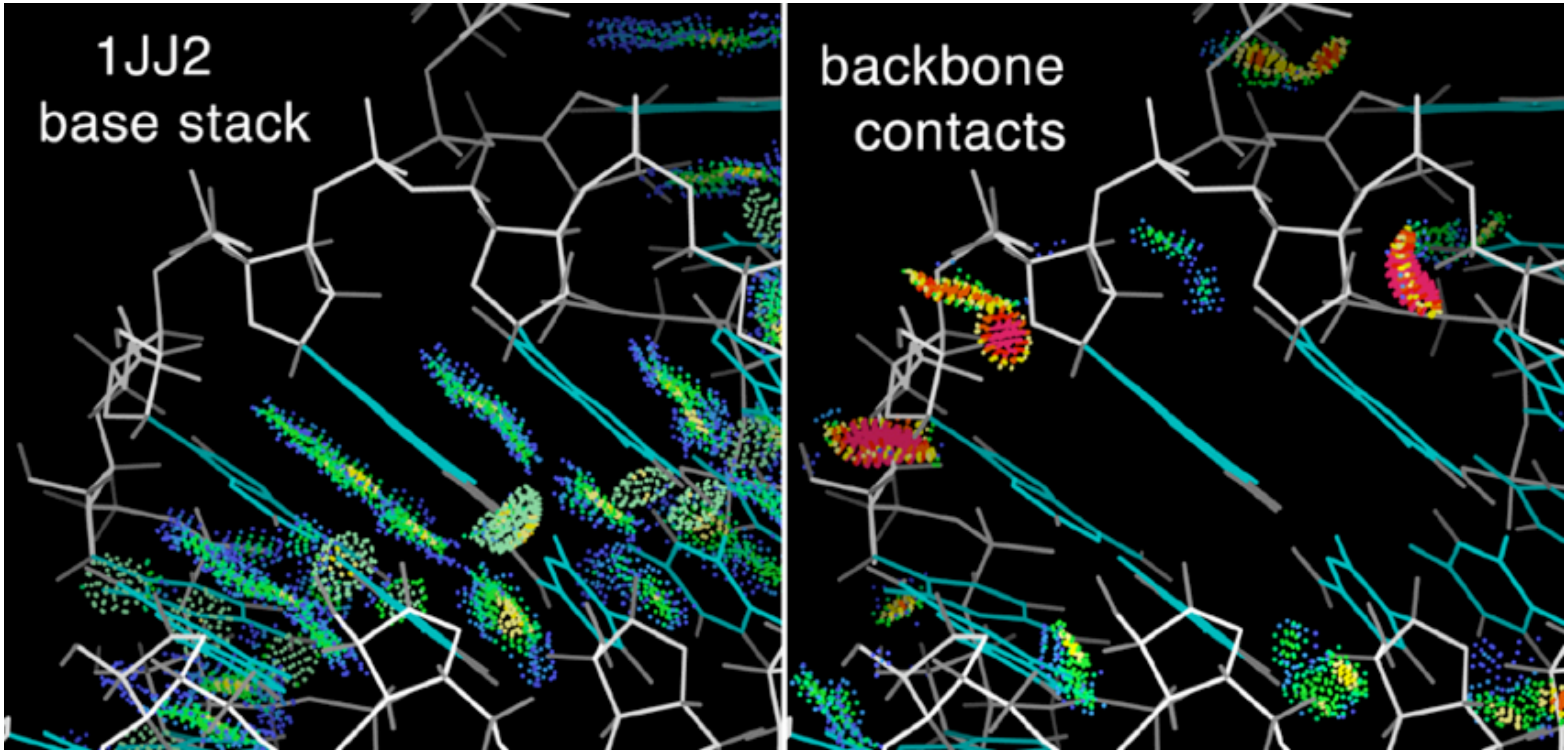
Sulfate Binding Protein (1SBP)

Phenix

Jeff Headd, LBL

Problems in Nucleic Acid Structures

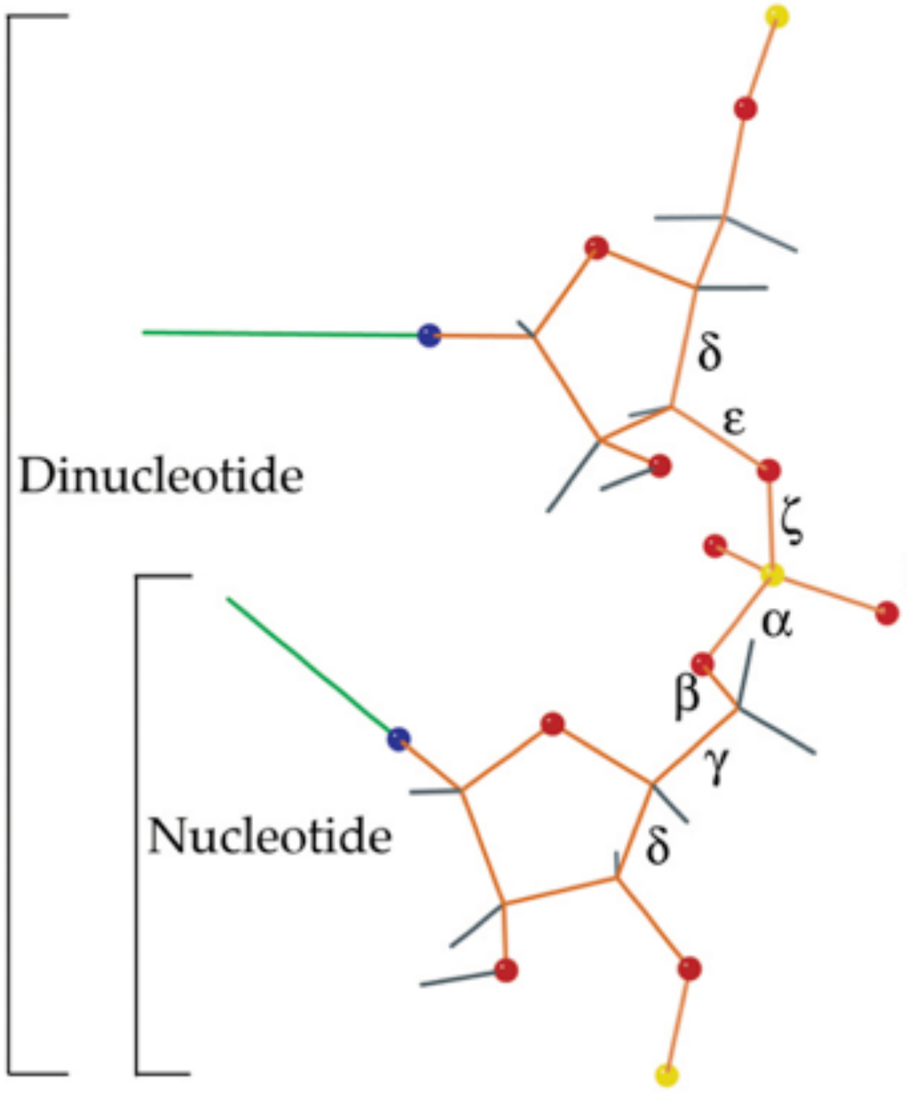
- Nucleic acid structures (esp. RNA) are often solved at low resolution
- The interactions between bases are often favorable
- It is common to see geometric problems with the backbone

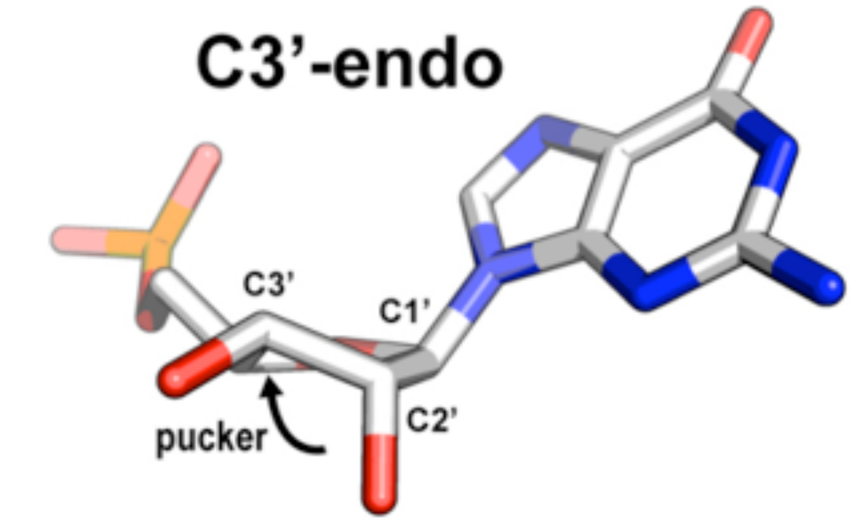


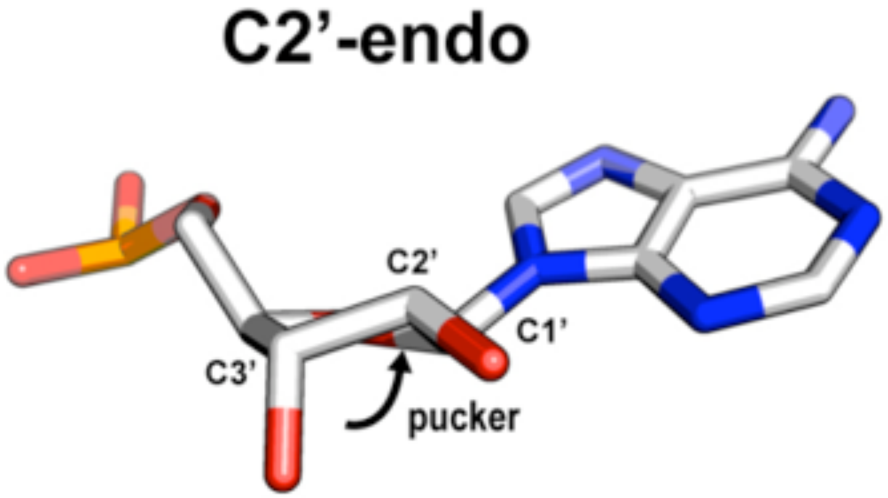
Jeff Headd & the Richardsons, Duke University

Conformation Dependent Geometry

- Nucleic acids have specific conformational variations in their backbone (arising from different sugar puckers)
- The different puckers lead to different local ideal geometries
- The best pucker is automatically recognized and the restraints dynamically modified







Richardson Lab, Duke University
Ralf Grosse-Kunstleve, LBL

Phenix

Secondary structure restraints

- For coordinate refinement, restrain hydrogen bond length (or N-O distance if hydrogens absent)
- Automatic annotation using KSDSSP* (`phenix.ksdssp`)
- Secondary structure groups for phenix.refine provided by `phenix.secondary_structure_restraints`

```
HELIX    1   1 ASP A   37  GLY A   48  1                                  12
SHEET    1   A 2 ARG A  13  ASP A  14  0
SHEET    2   A 2 LEU A  27  SER A  30 -1  O  ARG A  29   N  ARG A  13
```

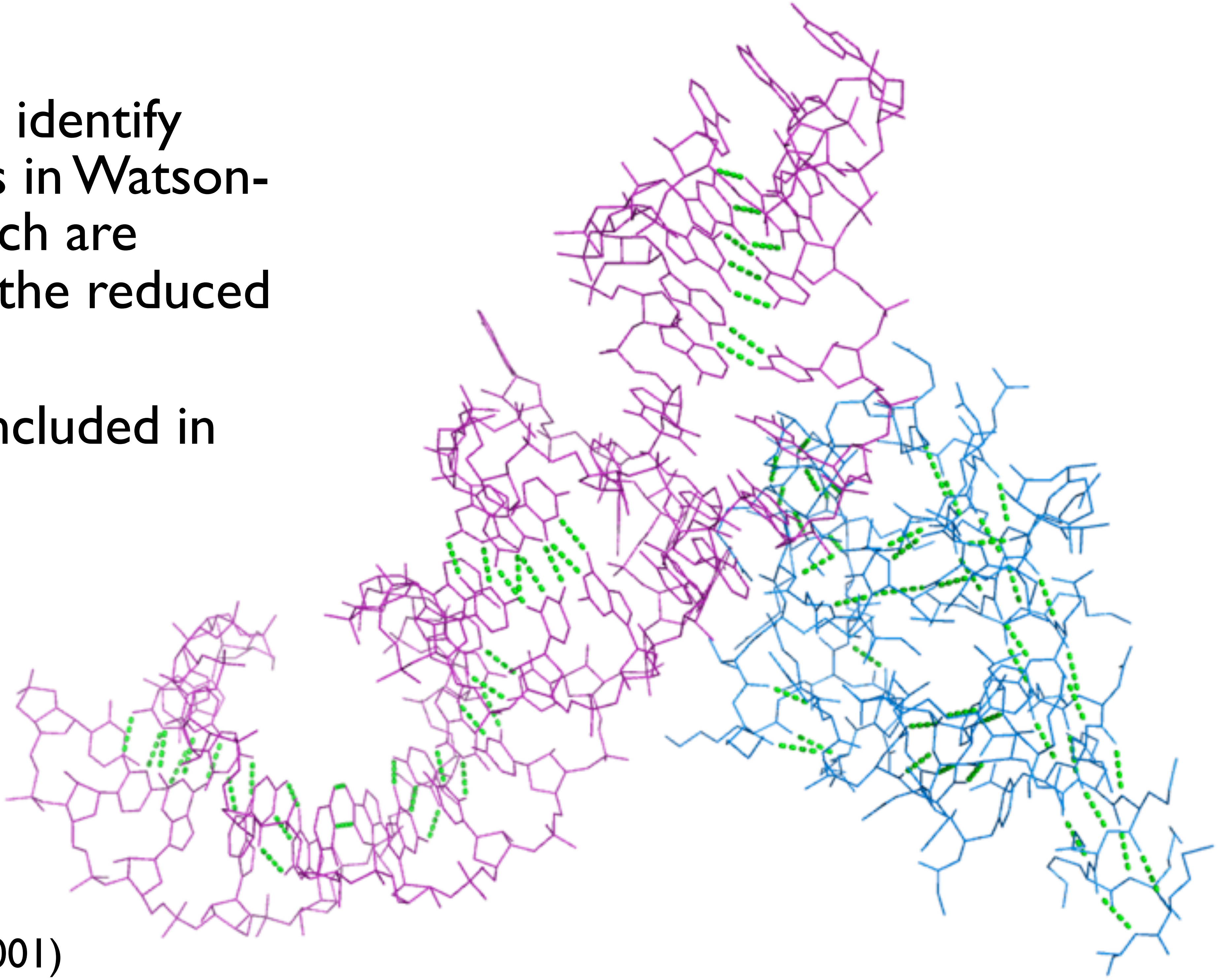
```
refinement.secondary_structure.helix {
  selection = "chain 'A' and resseq 263:275"
  helix_class = 1
}
refinement.secondary_structure.sheet {
  first_strand = "chain 'A' and resseq 13:14"
  strand {
    selection = "chain 'A' and resseq 27:30"
    sense = antiparallel
    bond_start_current = "chain 'A' and resseq 29"
    bond_start_previous = "chain 'A' and resseq 13"
  }
}
```

* Open-source (BSD-like) reimplementation of the DSSP algorithm, by authors of UCSF Chimera (http://www.cgl.ucsf.edu/Overview/software.html). The only free program of its type!

Nat Echols, LBL

Base pairing restraints

- Uses PROBE to identify hydrogen bonds in Watson-Crick pairs, which are converted into the reduced syntax
- Automatically included in refinement

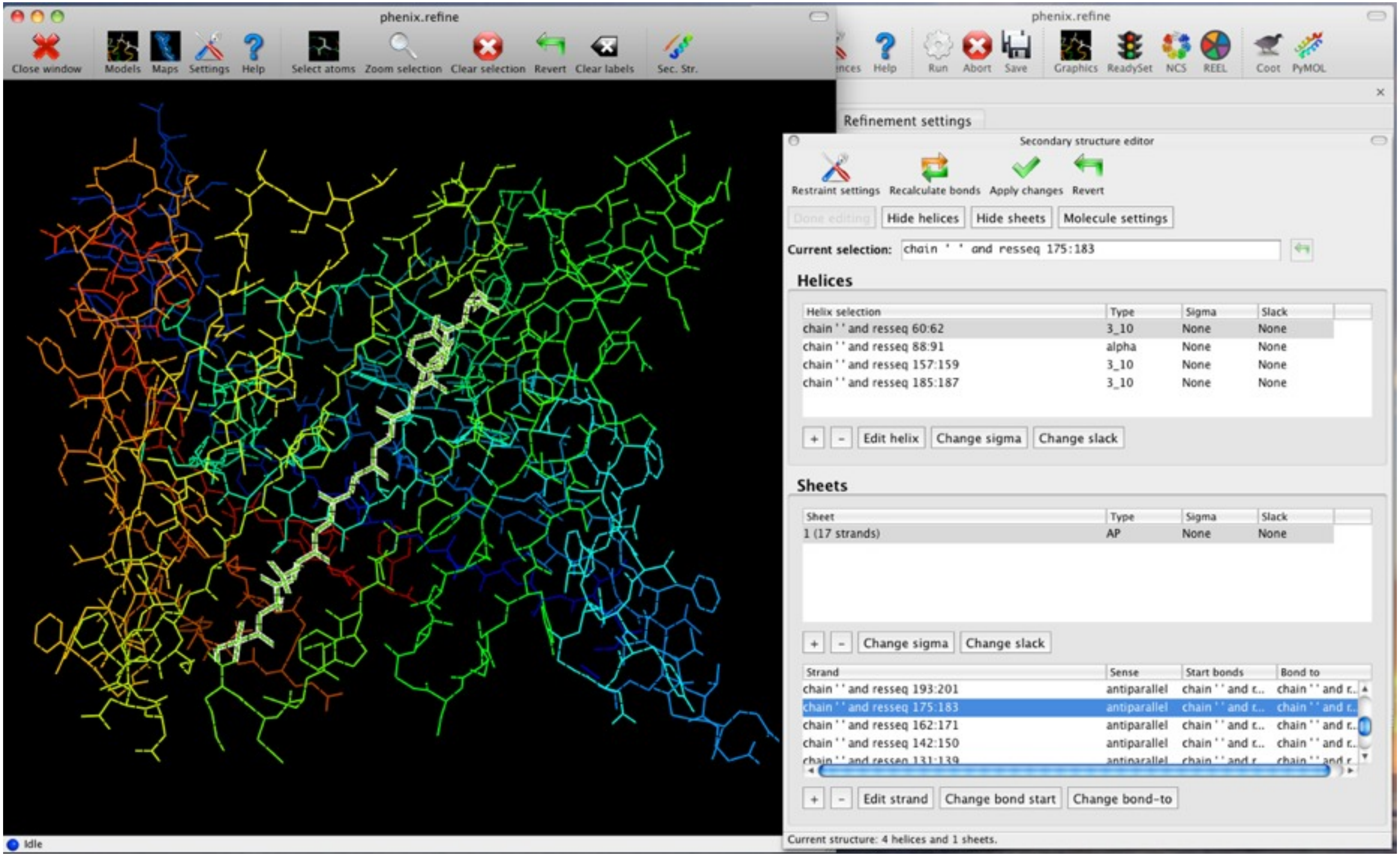


Example (protein+RNA):
Signal recognition particle
(Batey et al. JMB 307:229, 2001)
PDB ID: 1hq1

Phenix

Nat Echols & Jeff Headd, LBL

Editing secondary structure



Current selection: chain ' ' and resseq 175:183

Helices

Helix selection	Type	Sigma	Slack
chain ' ' and resseq 60:62	3_10	None	None
chain ' ' and resseq 88:91	alpha	None	None
chain ' ' and resseq 157:159	3_10	None	None
chain ' ' and resseq 185:187	3_10	None	None

Sheets

Sheet	Type	Sigma	Slack
1 (17 strands)	AP	None	None

Strand	Sense	Start bonds	Bond to
chain ' ' and resseq 193:201	antiparallel	chain ' ' and r...	chain ' ' and r..
chain ' ' and resseq 175:183	antiparallel	chain ' ' and r...	chain ' ' and r..
chain ' ' and resseq 162:171	antiparallel	chain ' ' and r...	chain ' ' and r..
chain ' ' and resseq 142:150	antiparallel	chain ' ' and r...	chain ' ' and r..
chain ' ' and resseq 131:139	antiparallel	chain ' ' and r	chain ' ' and r

Current structure: 4 helices and 1 sheets.

Phenix

Nat Echols, LBL

Secondary structure restraints: examples

- Automatic annotation with default settings, no H atoms
- DNA-binding protein, 3.1Å (early in refinement)*

SS	R-work	R-free	ΔR	Ramachandran outliers
-	0.2883	0.3689	0.0806	2.52%
+	0.2877	0.3652	0.0775	2.25%

* *data provided by A. Schoeffler, UC Berkeley*

- Bacterial protein, 2.25Å (AutoSol model)

SS	R-work	R-free	ΔR	Ramachandran favored**
-	0.2733	0.3246	0.0523	95.07%
+	0.2723	0.3221	0.0488	96.41%

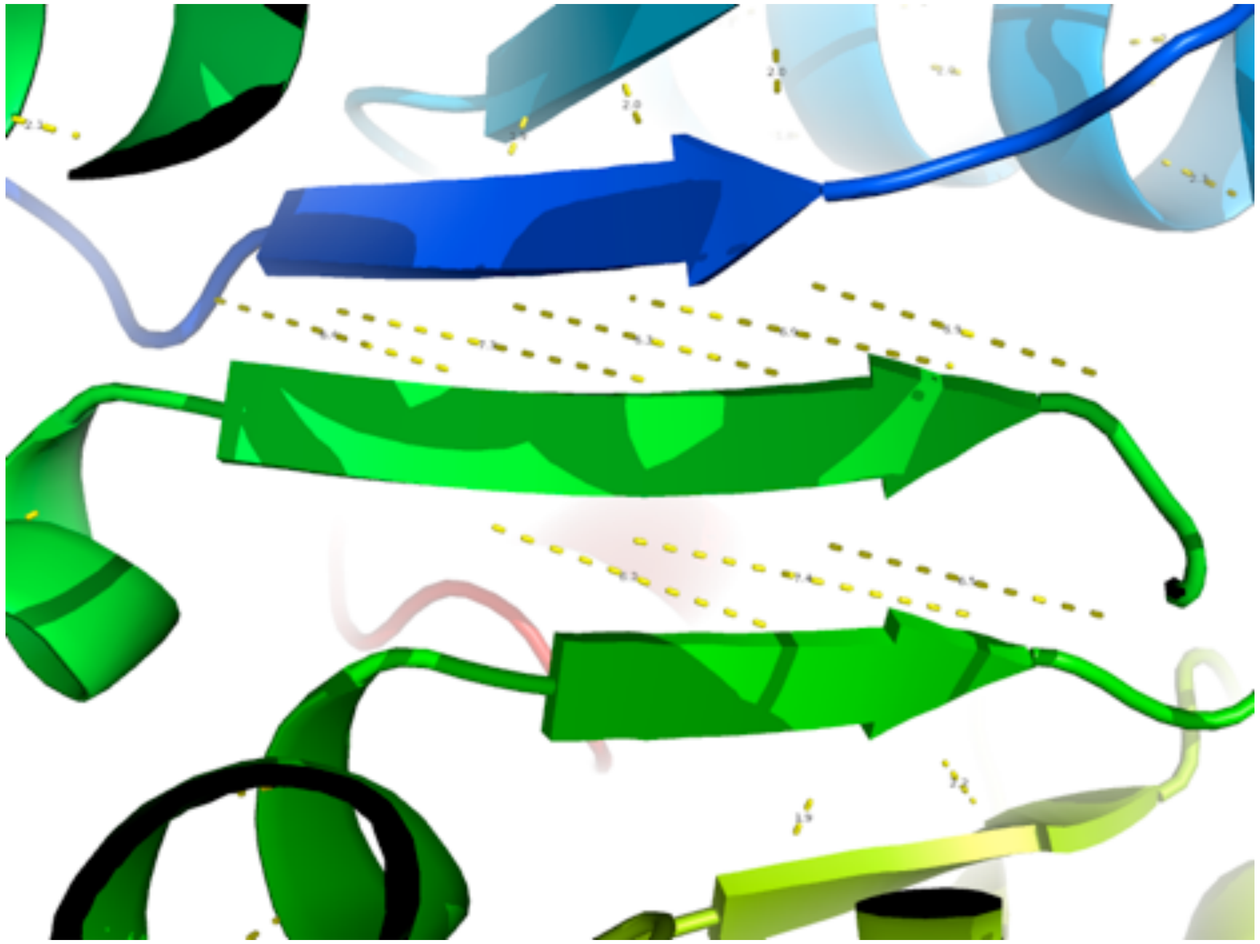
** *no outliers*

- Careful manual annotation may improve results

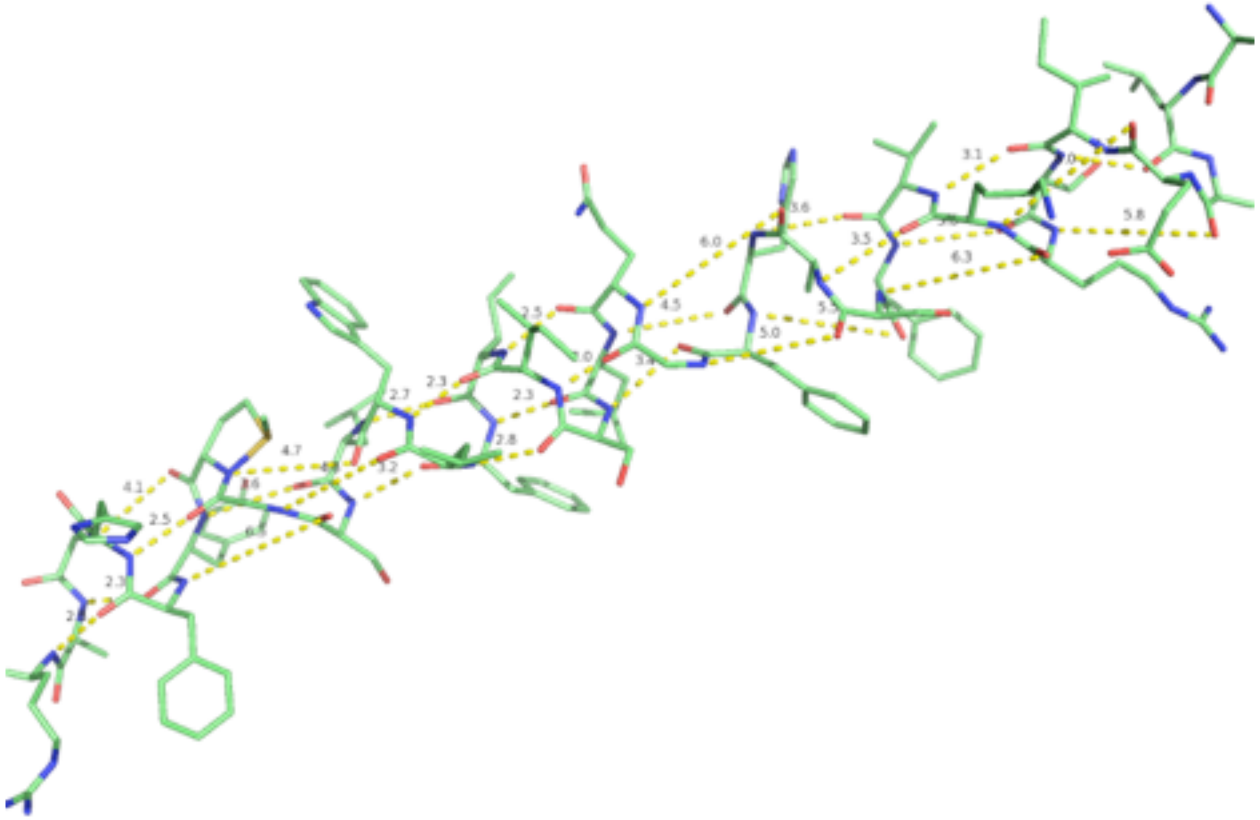
Nat Echols, LBL

Hydrogen bond quality control

- Automatic annotation is challenging - many false positives and negatives
- Outlier filtering throws out excessively long bonds, but not all of these are truly invalid
- Improved detection and/or prediction methods are needed



PDB ID 1a8i: SHEET records in PDB file are shifted



PDB ID 2o01: distorted geometry prevents automatic detection of helix

Phenix

Nat Echols, LBL

Reference Model Restraints for Low Resolution Refinement

- Improve low resolution refinement by using a related higher resolution structure as a reference.
- Generate reference dihedral restraints for all matching dihedral angles between the working model and the reference model.
- Restraints take the form of a simple harmonic:

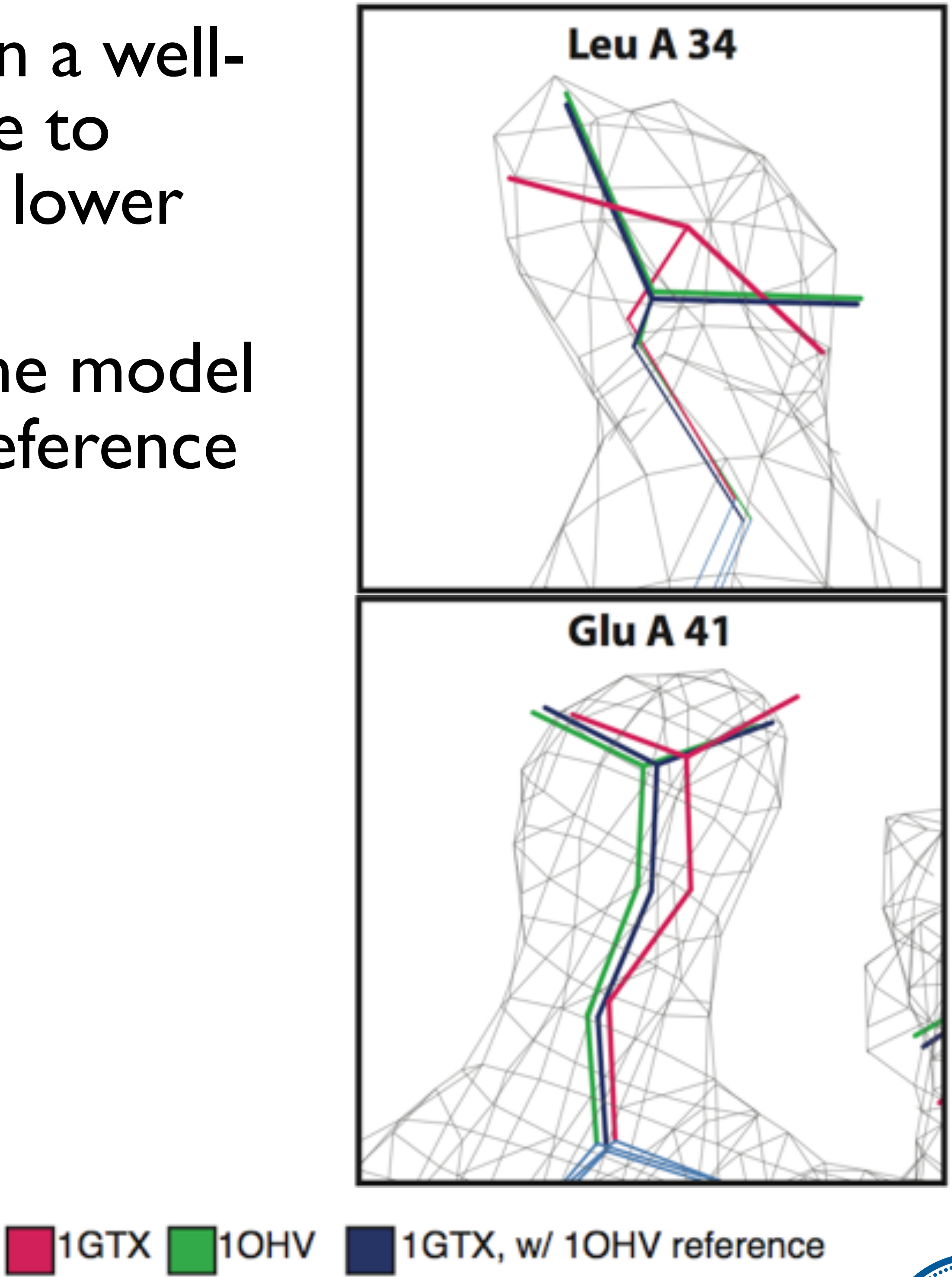
$$E_{total} = \sum_{i=1}^{n} E_i \qquad \left\{ \begin{array}{l} E_i = \omega_i \Delta_i^2, \quad \Delta_i \leq l \\ E_i = \omega_i l^2, \quad \Delta_i > l \end{array} \right\} \qquad \omega_i = \frac{1}{\sigma^2}$$

- 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.
- The 'limit' parameter allows differences between the working and reference models (e.g. hinges, conformational changes)
- Pre-correct rotamer outliers in the working model to match the χ angles of the reference model if the reference model has a proper rotamer at that position.

Reference Structures

- Use the information contained in a well-defined high resolution structure to improve models generated with lower resolution data
- Dihedral angle restraints pulls the model towards the higher resolution reference (until the deviation is too great)

		1GTX alone	1OHV	1GTX w/ ref.
Leu A 34	X_1	203.5°	186.4°	185.6°
	X_2	225.6°	45.6°	46.3°
	Rotamer	Outlier	**tp**	**tp**
Glu A 41	X_1	295.4°	287.7°	287.7°
	X_2	177.1°	172.6°	173.0°
	X_3	47.5°	73.2°	73.0°
	Rotamer	**mt-10**	**mt-10**	**mt-10**

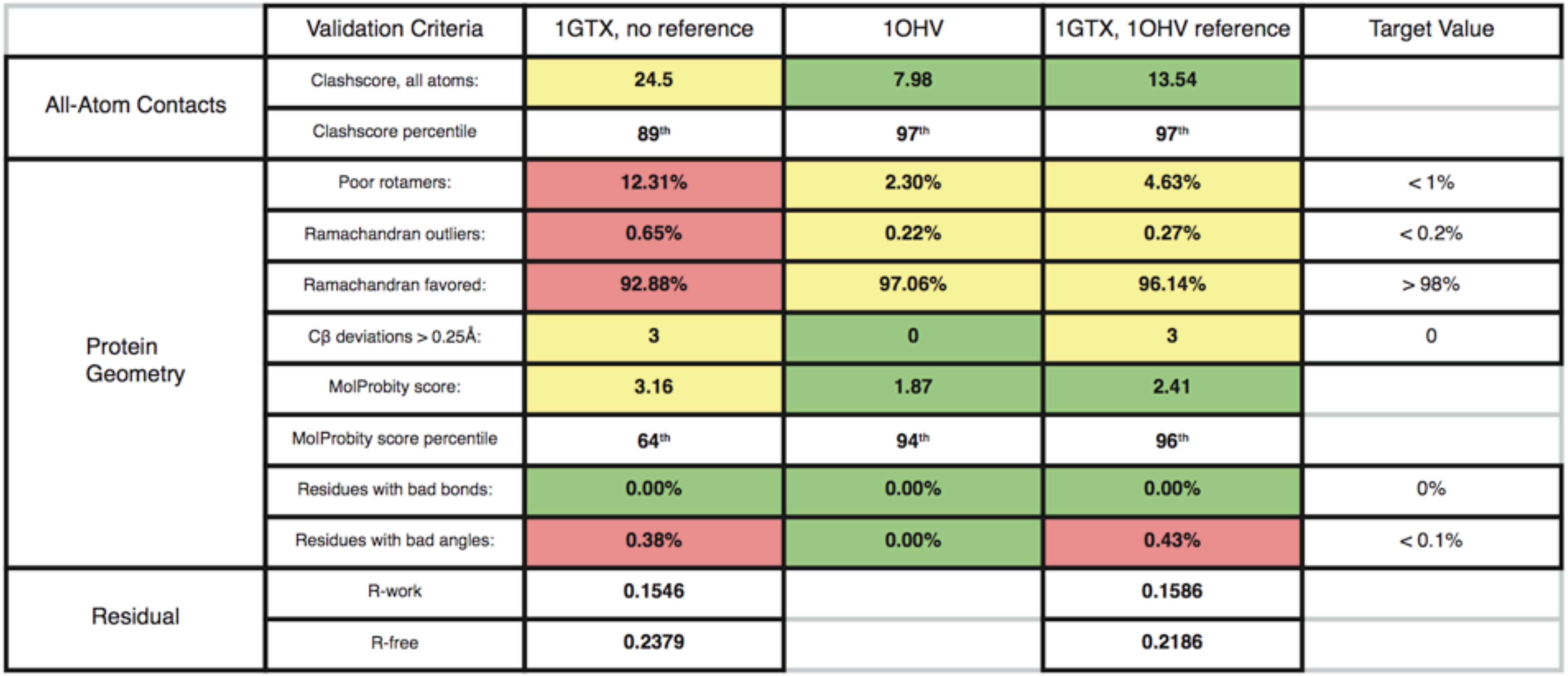


Phenix

Jeff Headd, LBL

Reference Structures

- Overall statistics are improved - better geometry and better fit to the experimental data

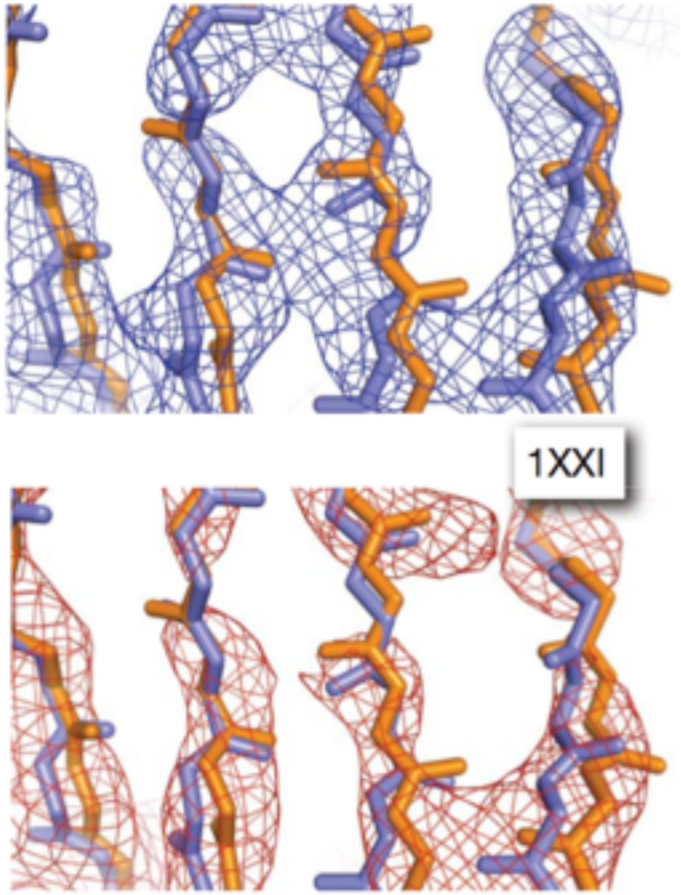


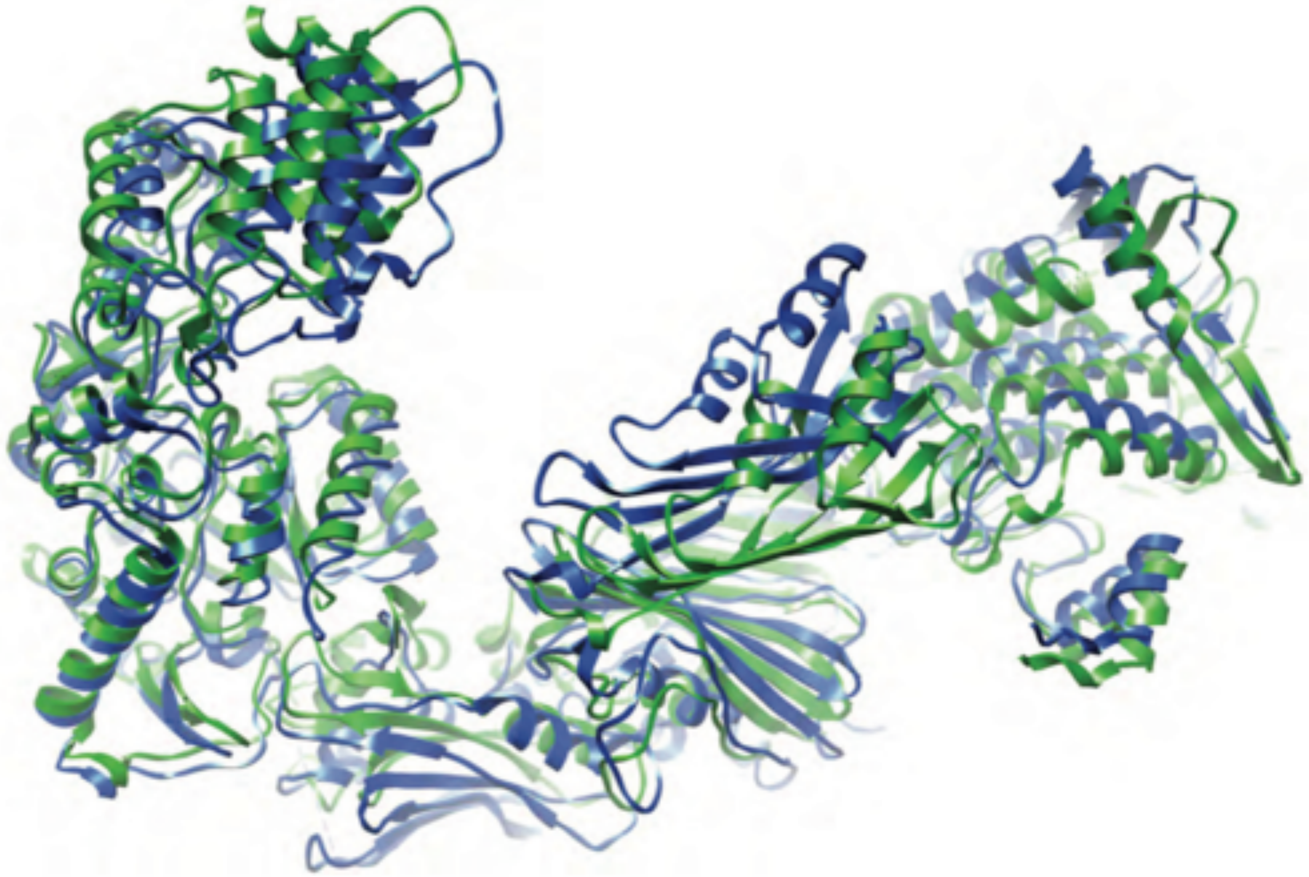
	Validation Criteria	1GTX, no reference	1OHV	1GTX, 1OHV reference	Target Value
All-Atom Contacts	Clashscore, all atoms:	24.5	7.98	13.54	
	Clashscore percentile	89th	97th	97th	
Protein Geometry	Poor rotamers:	12.31%	2.30%	4.63%	< 1%
	Ramachandran outliers:	0.65%	0.22%	0.27%	< 0.2%
	Ramachandran favored:	92.88%	97.06%	96.14%	> 98%
	Cβ deviations > 0.25Å:	3	0	3	0
	MolProbity score:	3.16	1.87	2.41	
	MolProbity score percentile	64th	94th	96th	
	Residues with bad bonds:	0.00%	0.00%	0.00%	0%
	Residues with bad angles:	0.38%	0.00%	0.43%	< 0.1%
Residual	R-work	0.1546		0.1586	
	R-free	0.2379		0.2186	

Jeff Headd, LBL

The DEN Method

- Researchers have developed other methods to add prior information into structure refinement and fitting (Schroeder et al., 2010)
- A deformable elastic network is used to restrain the model to an external structure
- Better models are produced (geometric and R-values)





Summary

- Algorithms previously used for validation can be used to automatically correct models during refinement
  - Automated rotamer refitting
  - Automated sidechain flips
- Low resolution structure solution and refinement is challenging, but can be improved
  - Inclusion of external information provides additional observations
    - Secondary structure restraints
    - High resolution reference models
- There is room for improvement of the geometric restraints used in refinement

Challenges Remain

- Low resolution structure solution and refinement
- Structure completion
  - Automated identification, fitting and refinement of ligands, metals, ions, and water
  - Identification, fitting and refinement of discrete disorder (multiple conformations)
  - Representing other forms of disorder
- Automated parameterization of models in refinement
  - ADPs, TLS groups, NCS, hydrogens
- Handling different kinds of twinning and integrating it into the whole structure solution process
- Automated understanding of chemistry

Acknowledgments

- **Lawrence Berkeley Laboratory**
  - Pavel Afonine, Nat Echols, Jeff Headd, Ralf Grosse-Kunstleve, Nigel Moriarty, Nicholas Sauter, 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, Chris Williams, Bryan Arendall, Laura Murray
- **Others**
  - Garib Murshudov & Alexi Vagin
  - Kevin Cowtan, Paul Emsley, Bernhard Lohkamp
  - Alexandre Urzhumtsev & Vladimir Lunin
  - 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