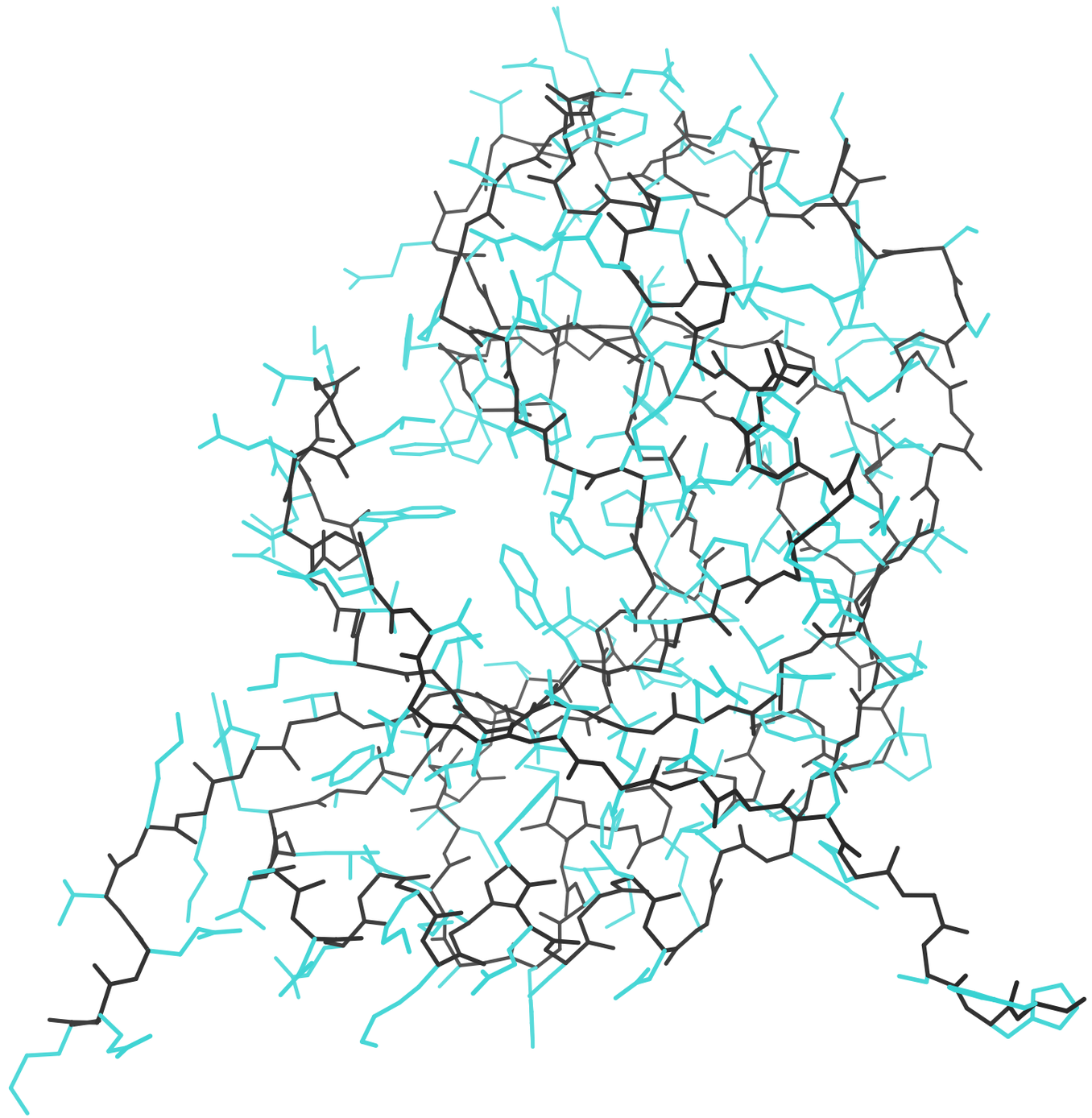
Validating your model with MolProbity and friends

Nat Echols, Jeffrey Headd, Pavel Afonine, Jane Richardson, et al.

These slides (and many more) are available online:
http://www.phenix-online.org/presentations

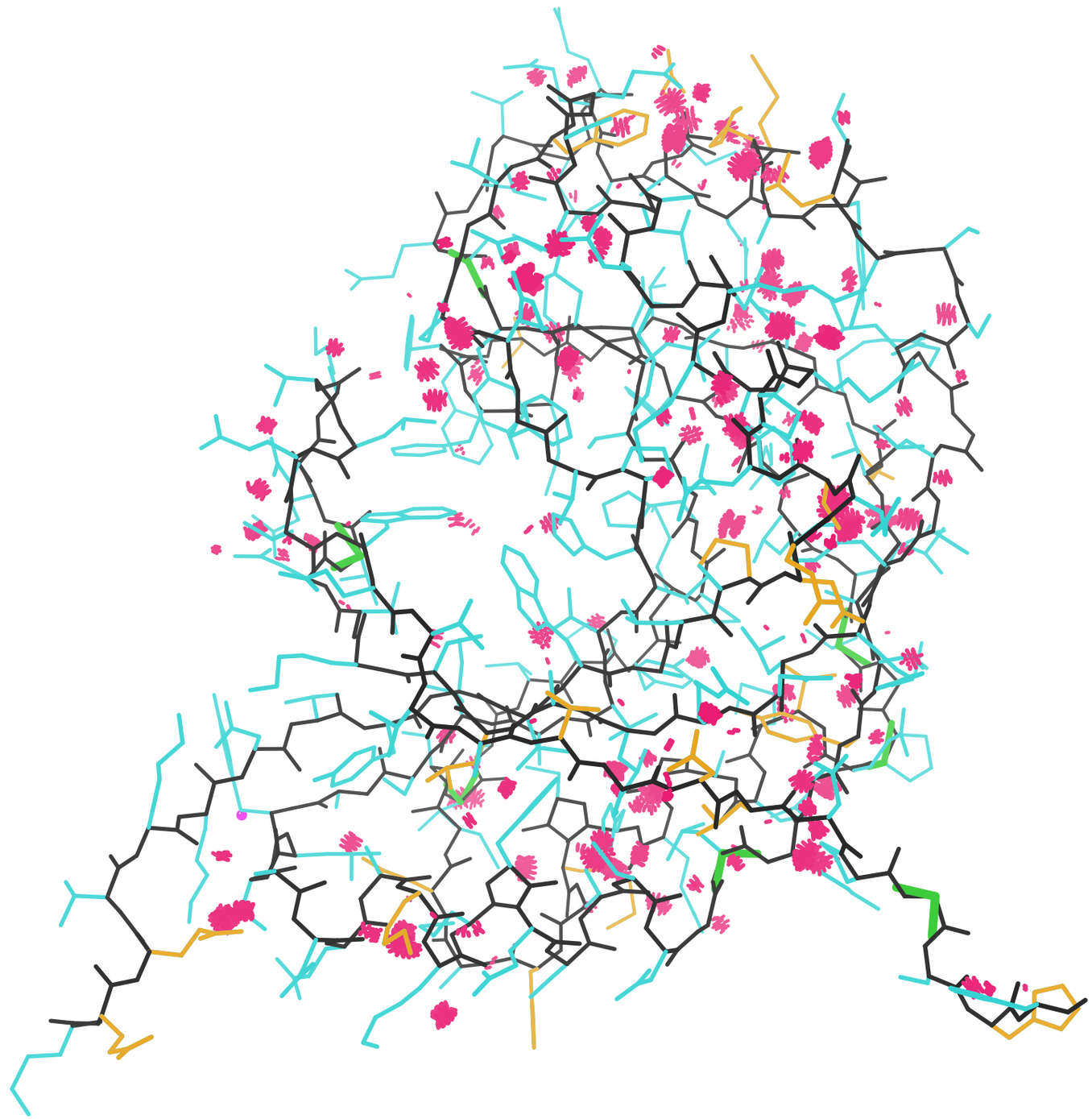
See also:
http://molprobity.biochem.duke.edu

What is model validation?

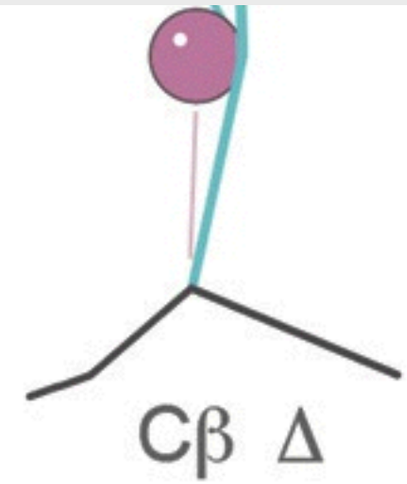


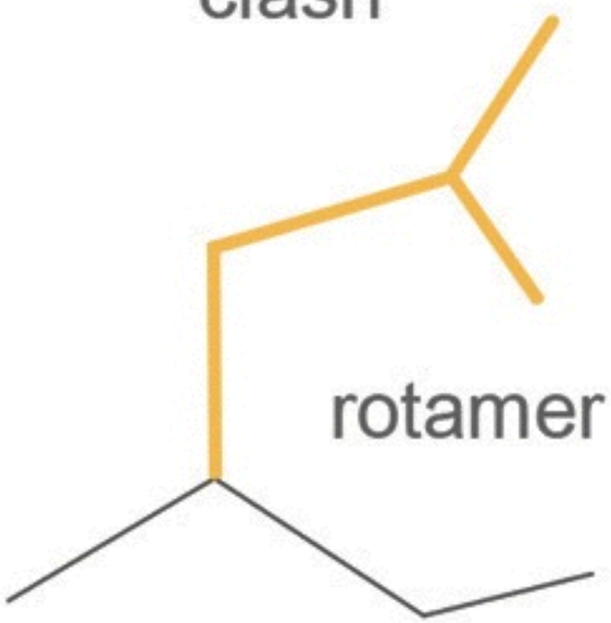
Cyclic Nucleotide Phosphodiesterase (2.4 Å)

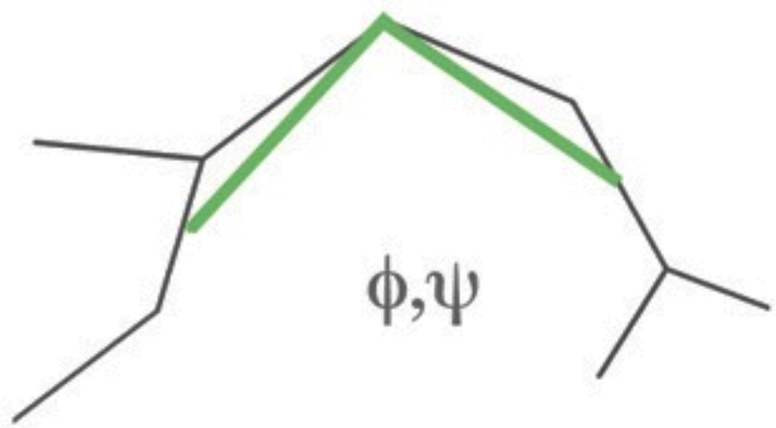
What is model validation?

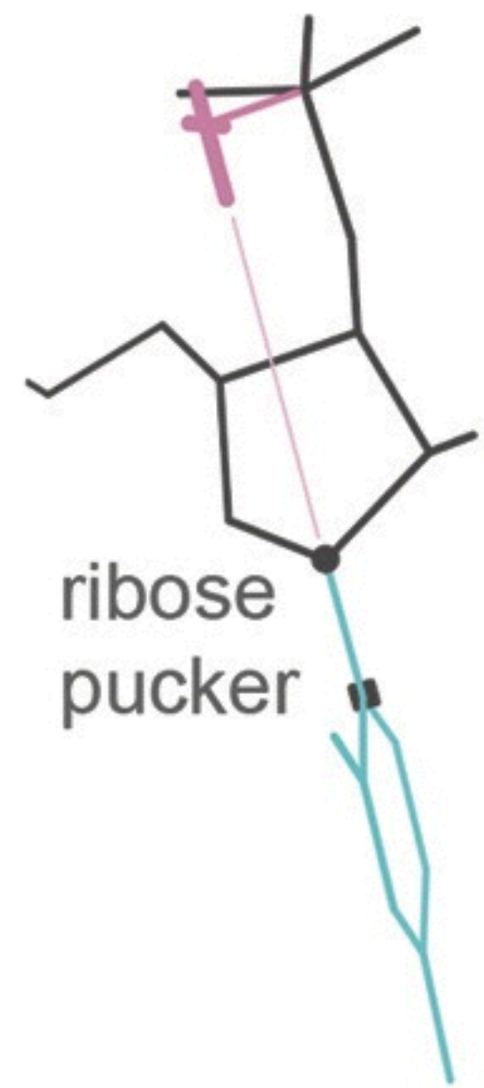


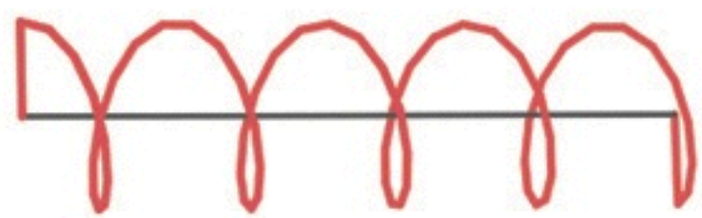
Cyclic Nucleotide Phosphodiesterase (2.4 Å)





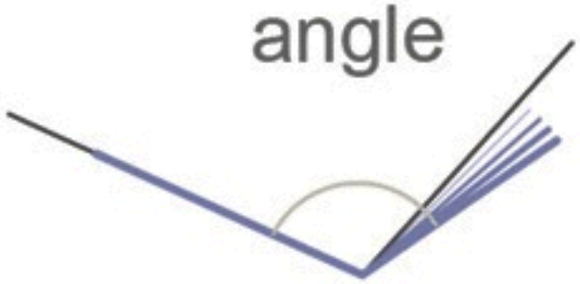


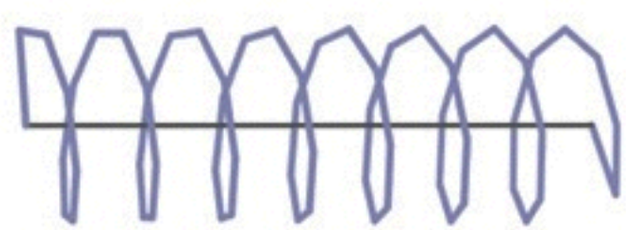




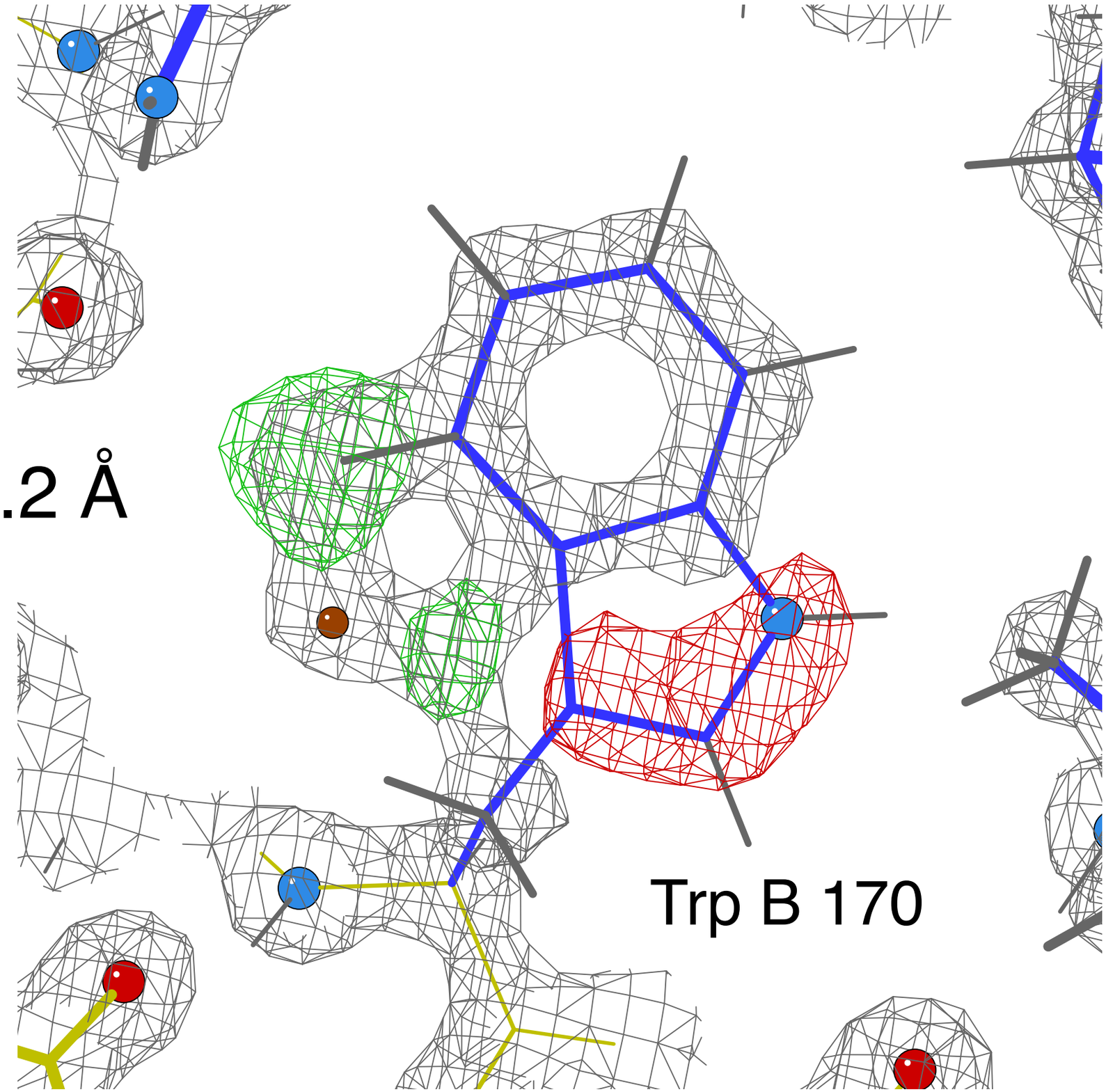
bond

angle



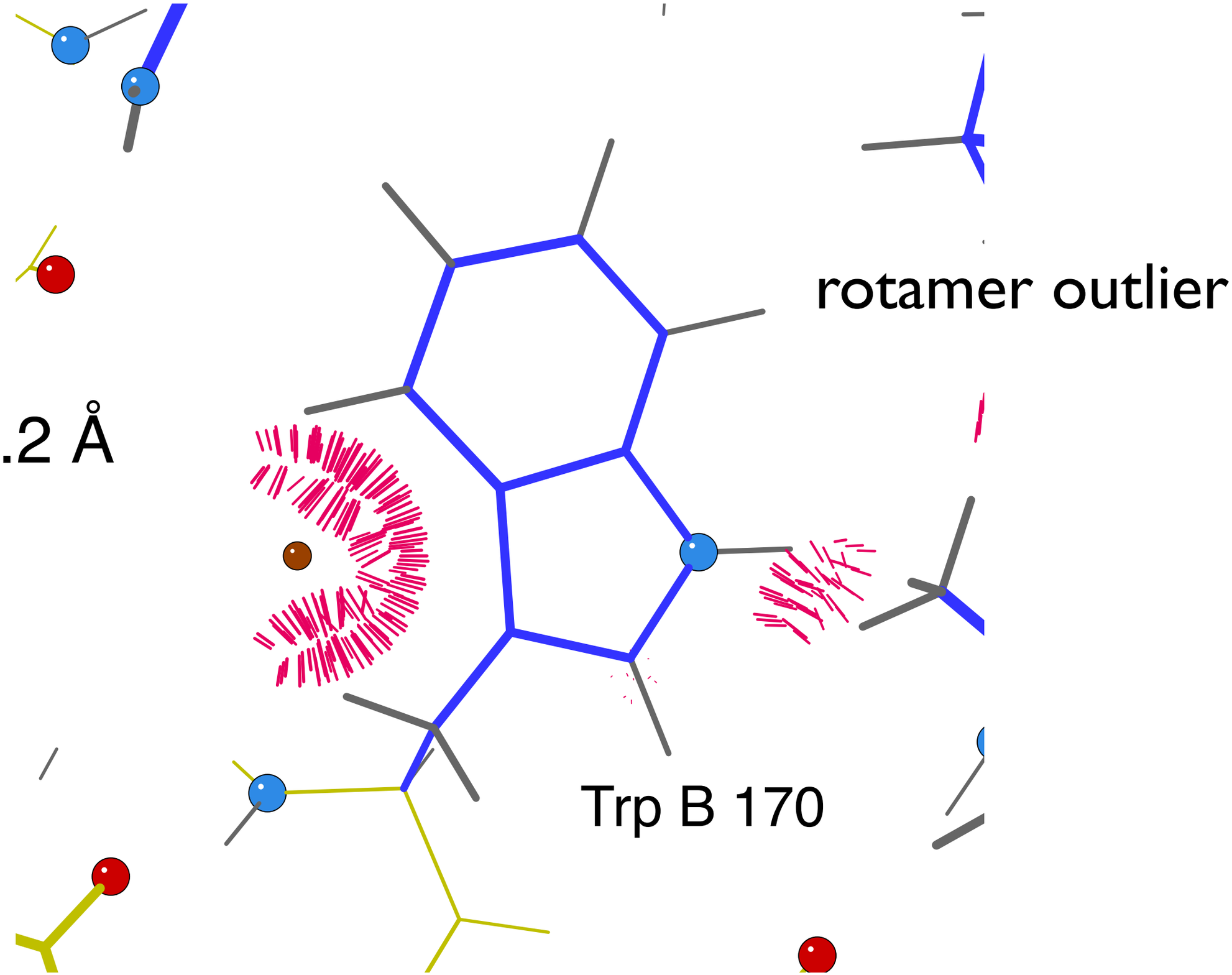


at least one person should look at the map...

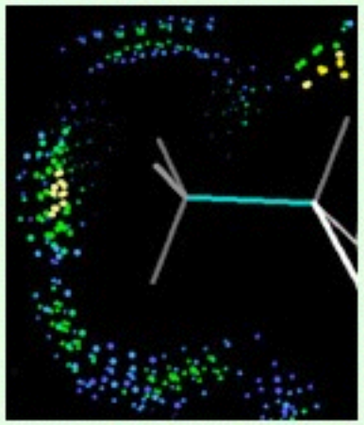


* courtesy of Dale Tronrud

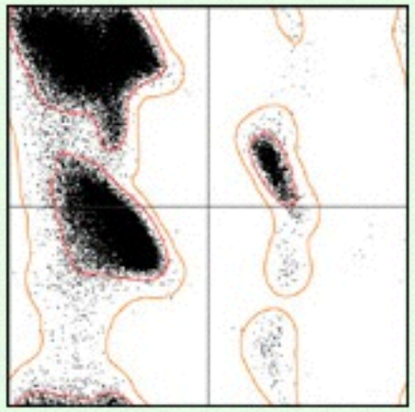
at least one person should look at the map...



All-atom contacts, clashscore

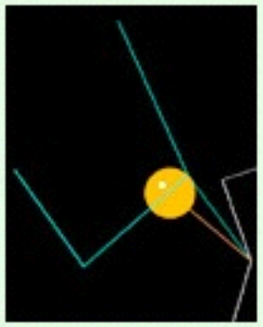


Ramachandran criteria

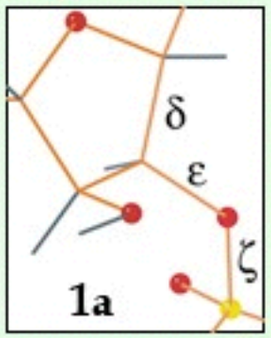


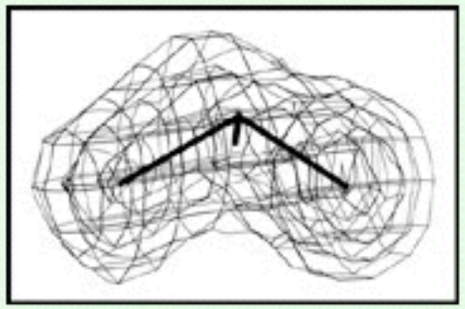
Sidechain rotamers

Geometry



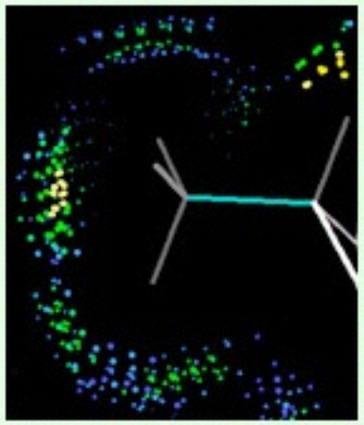
RNA bb



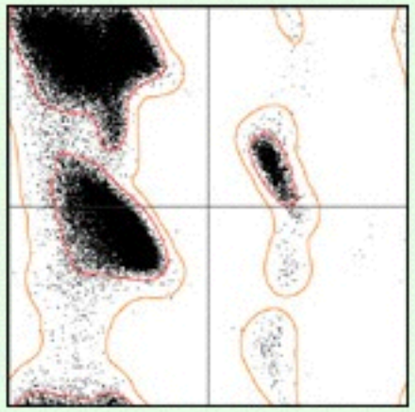


Crystallographic : R_{free}, electron density fit

All-atom contacts, clashscore

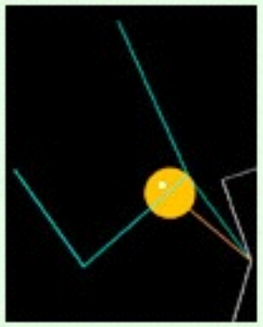


Ramachandran criteria

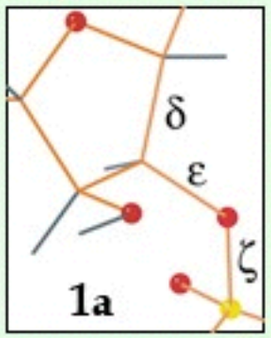


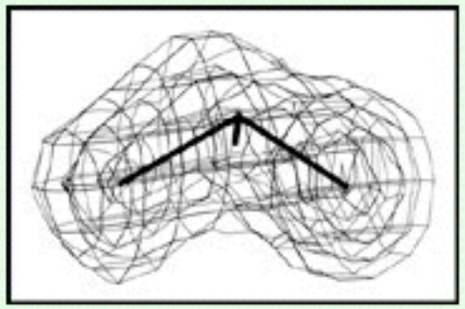
Sidechain rotamers

Geometry



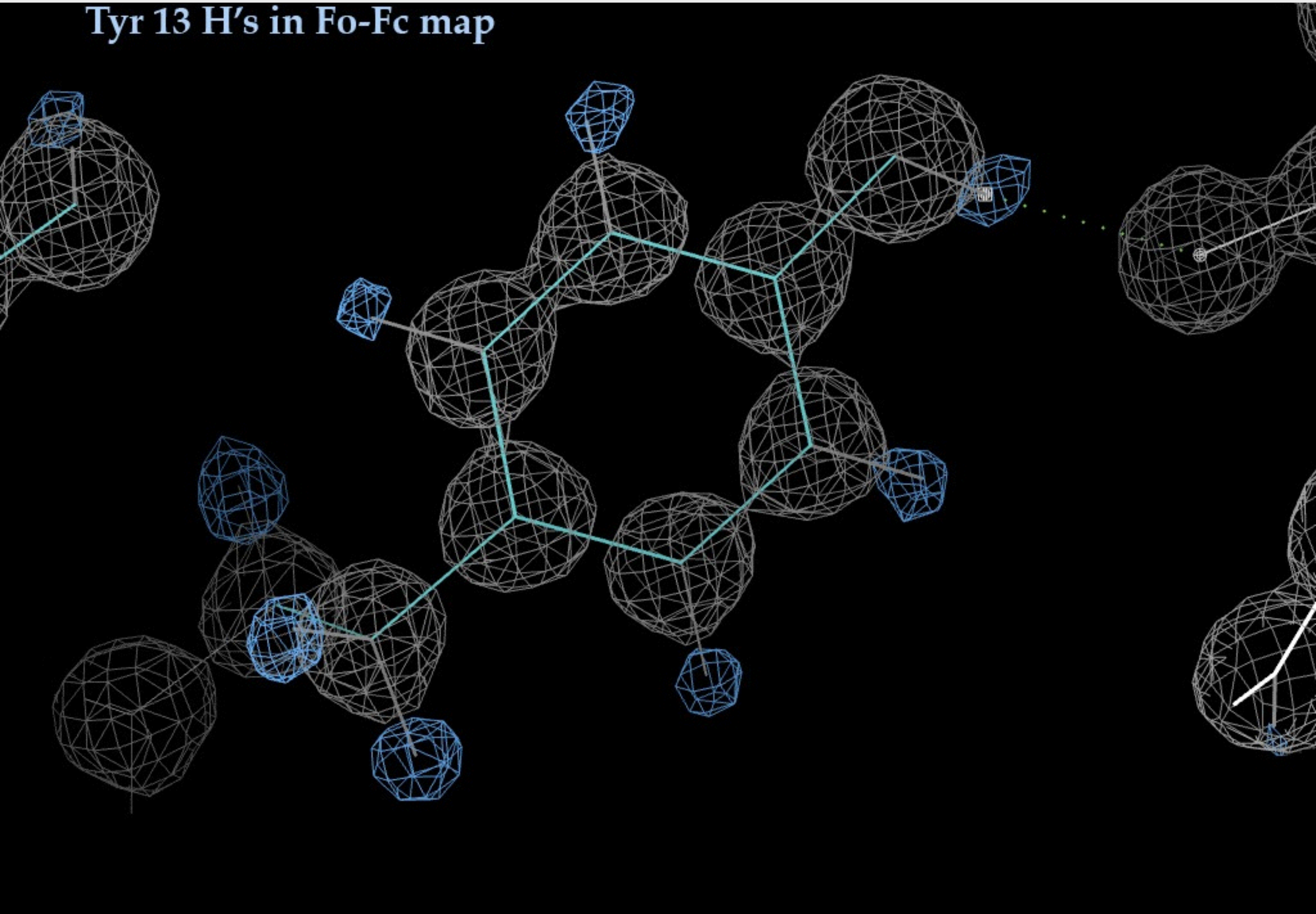
RNA bb

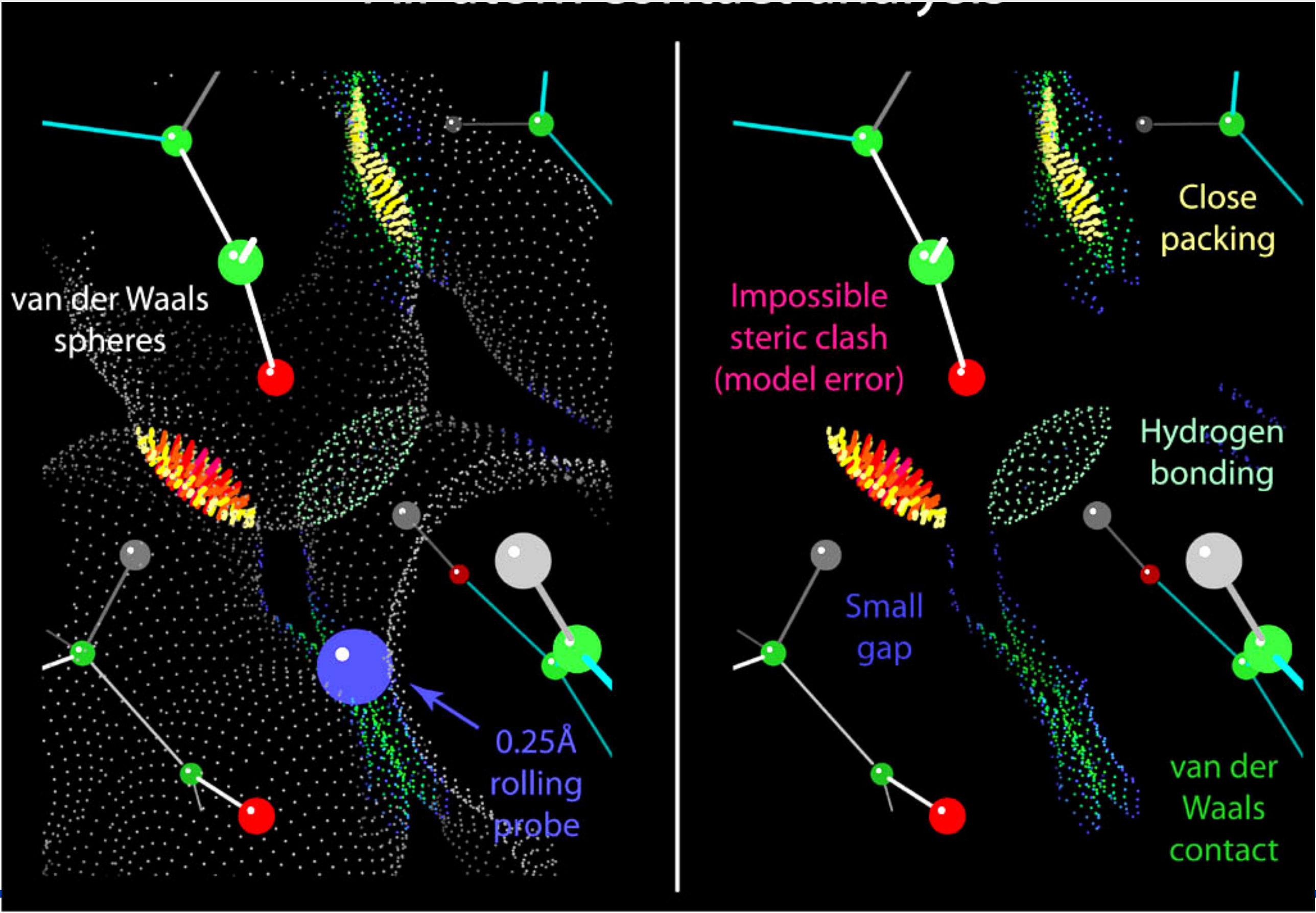




Crystallographic : R_{free}, electron density fit

Tyr 13 H's in Fo-Fc map

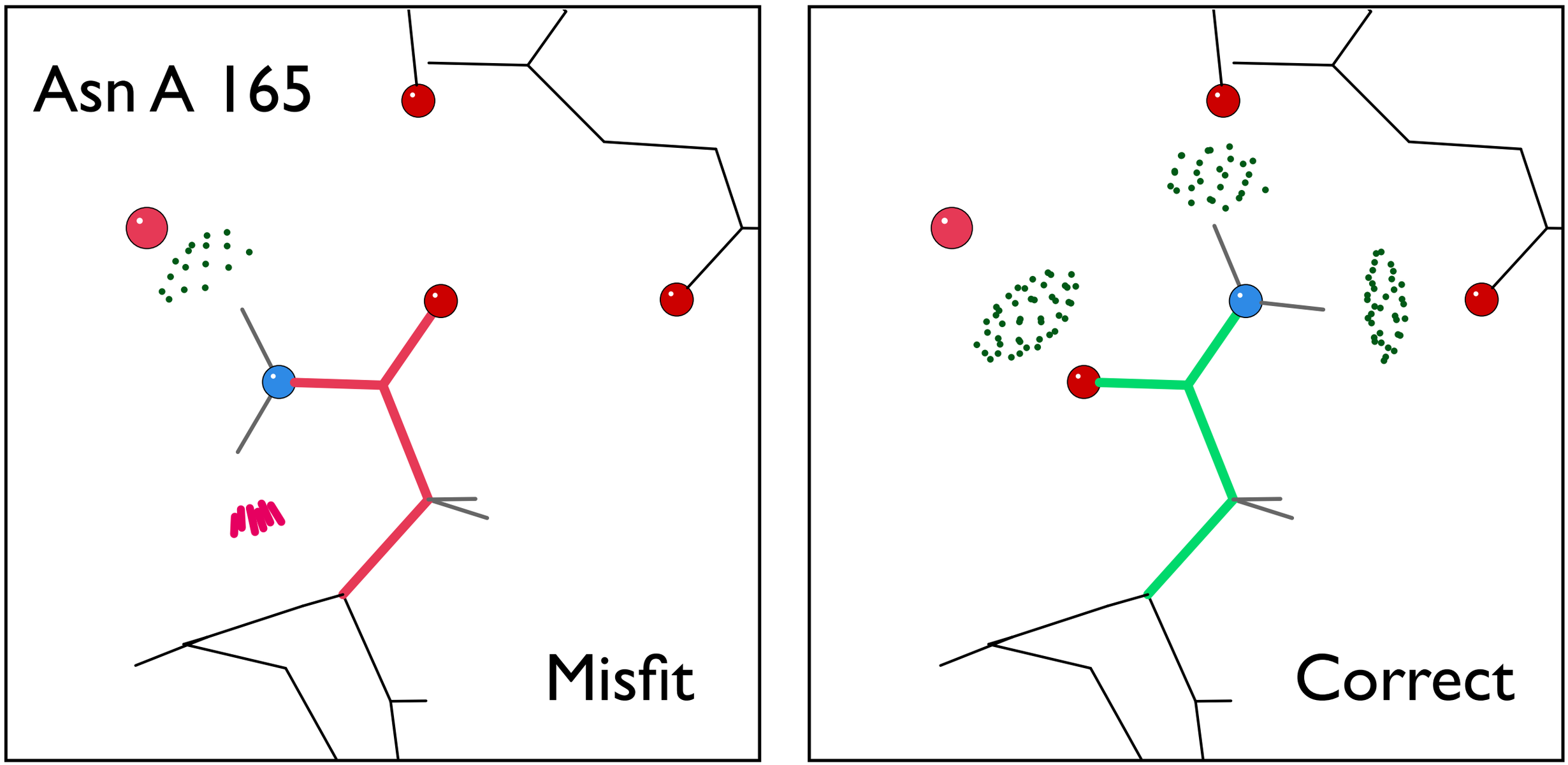



van der Waals
spheres
0.25Å
rolling
probe
Impossible
steric clash
(model error)
Close
packing
Hydrogen
bonding
Small
gap
van der
Waals
contact

Asn / Gln / His Correction

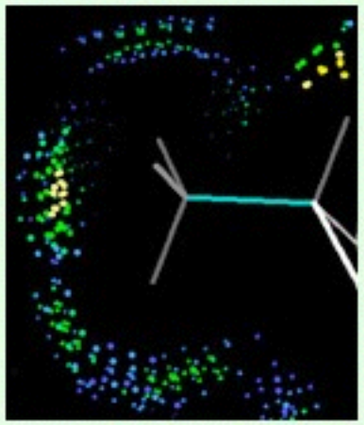
* Automatically detect and correct flipped N/Q/H residues at each macrocycle

* Uses MolProbity/Reduce methodology (H-bonds, clashes) to determine correct orientation

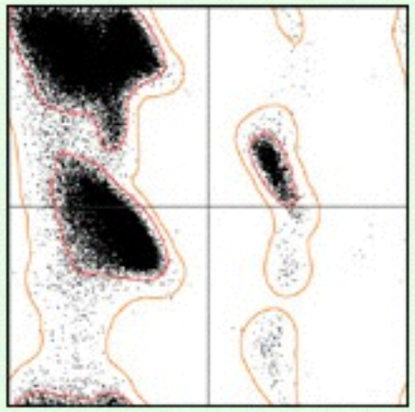


Sulfate Binding Protein (1SBP)

All-atom contacts, clashscore

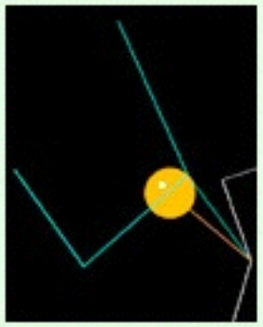


Ramachandran criteria

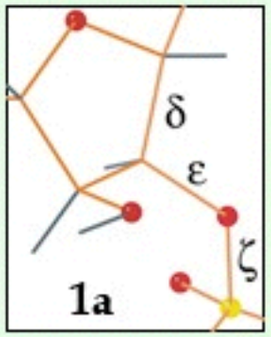


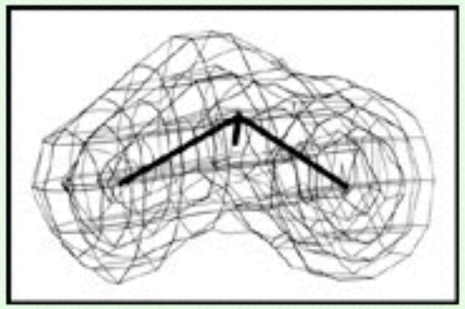
Sidechain rotamers

Geometry

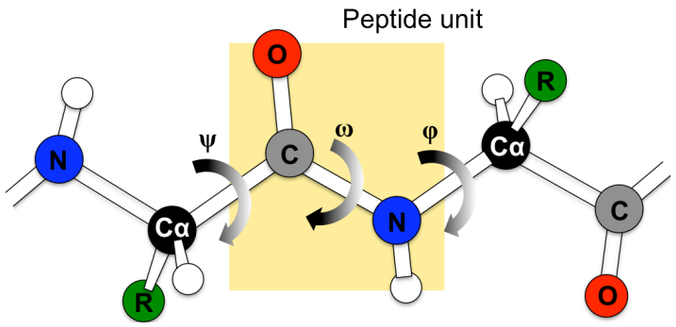


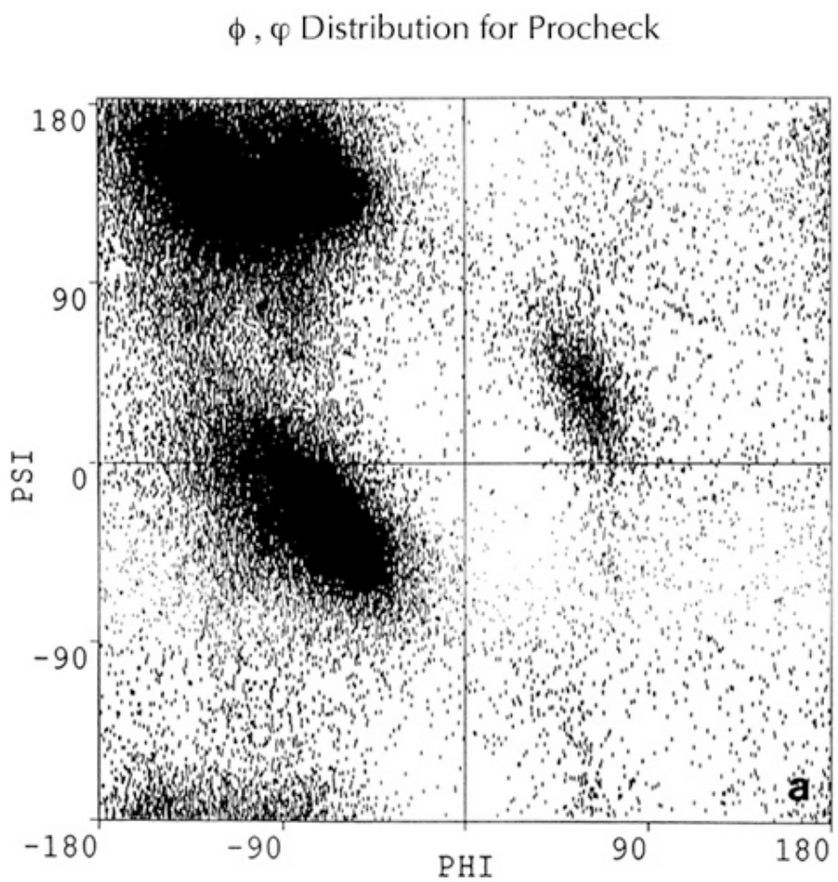
RNA bb





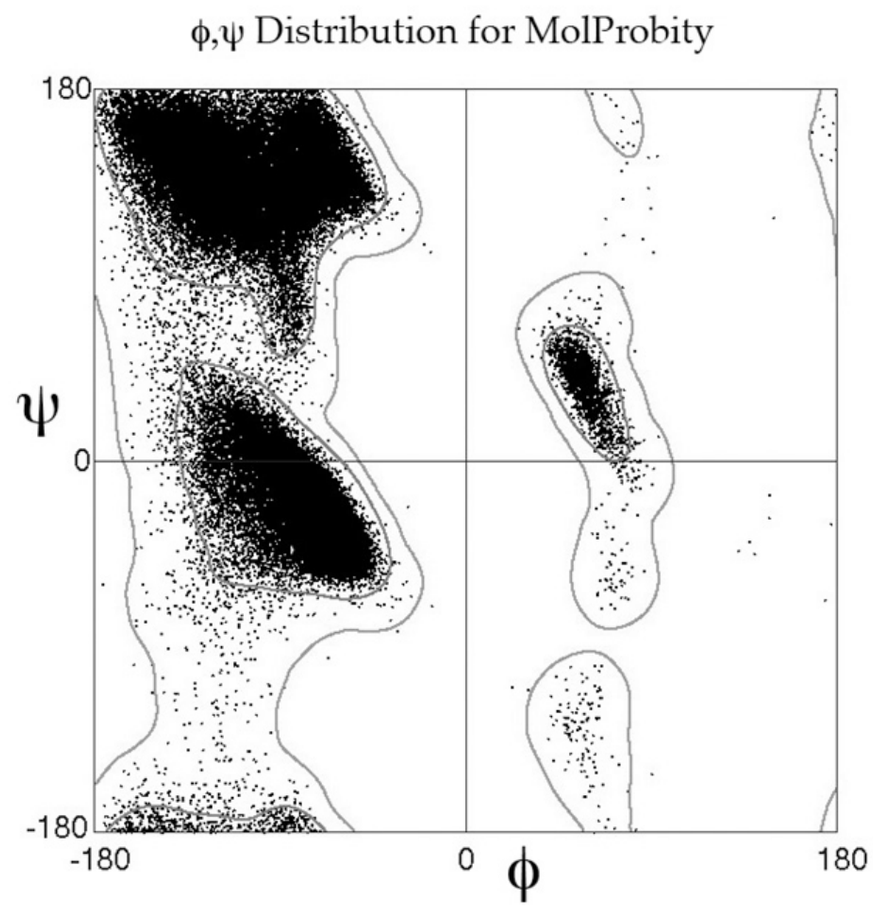
Crystallographic : R_{free}, electron density fit





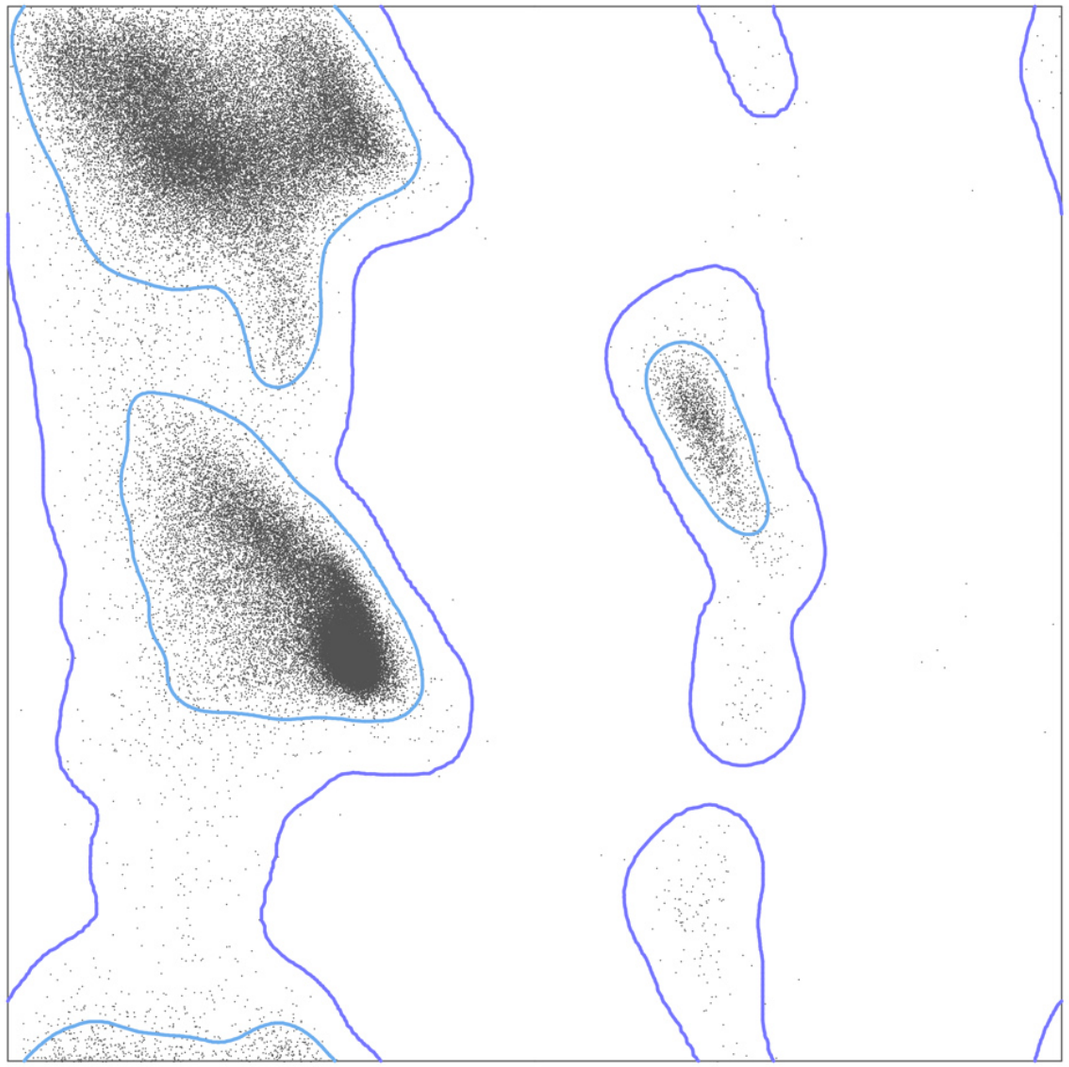
~ 100,000 residues, entire PDB as of 1992

reference: A.L. Morris, *et al.*, (1992) Proteins, **12**: 345-364.

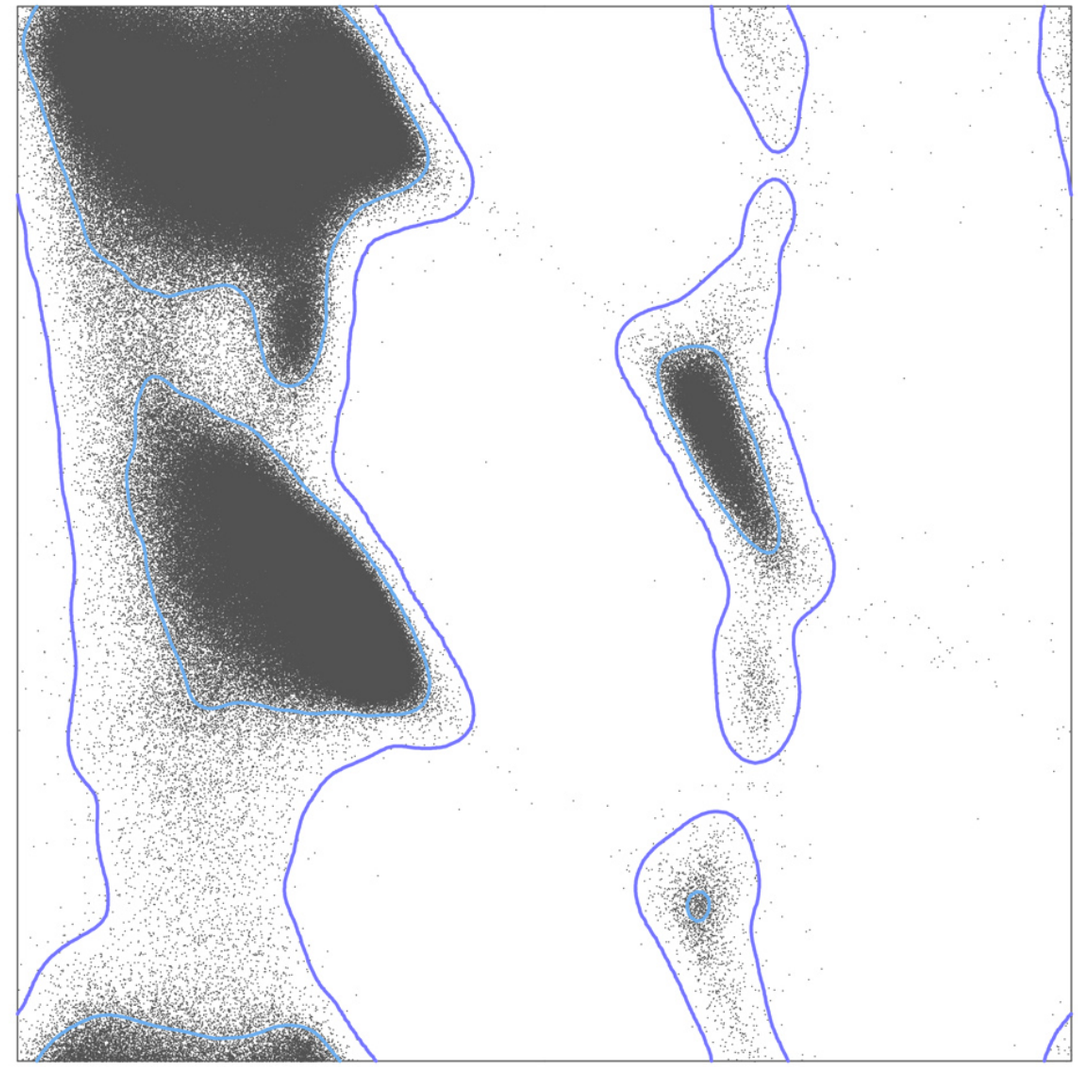


~ 100,000 residues, Top500 structures in 2003

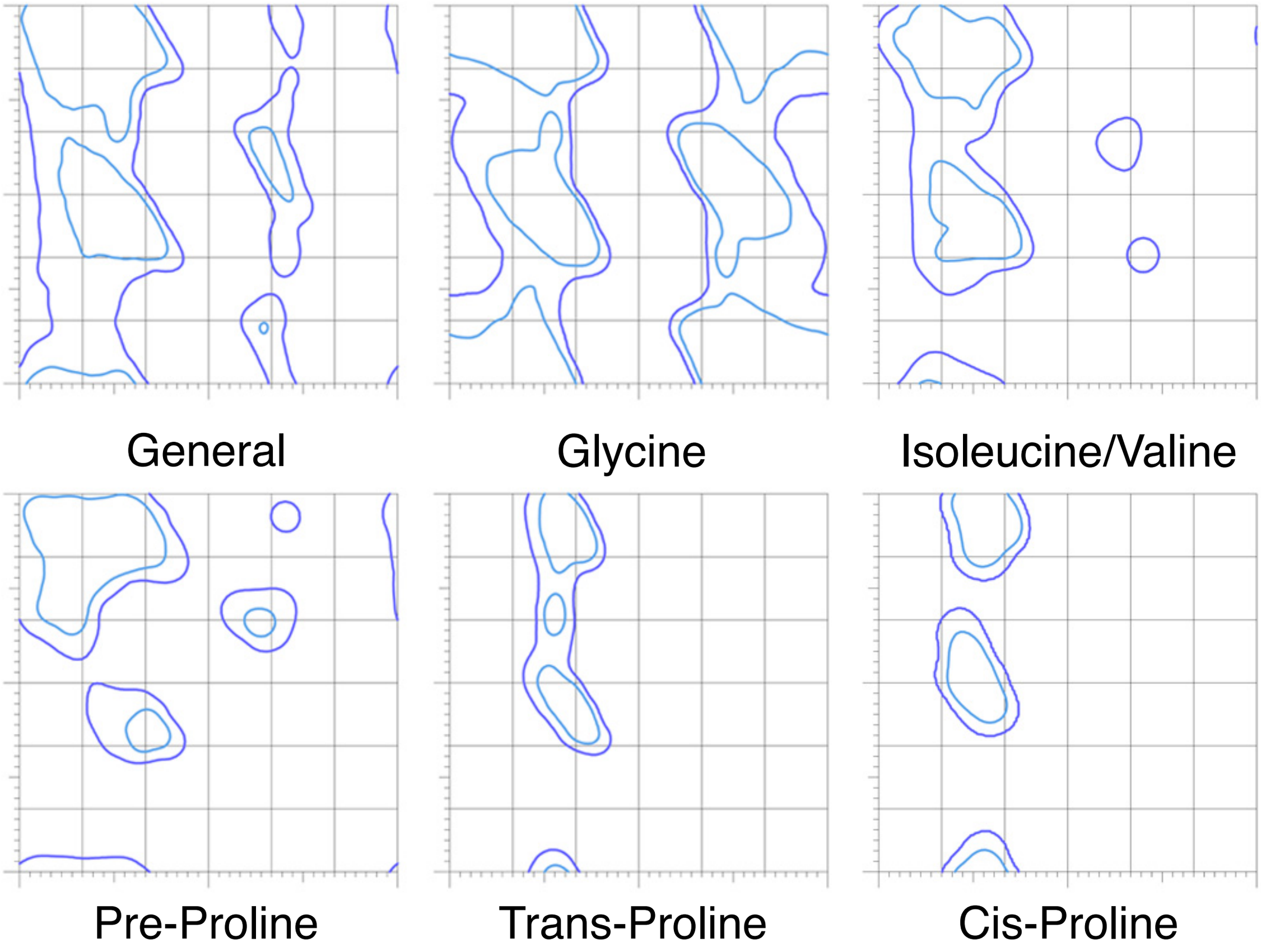
Top500



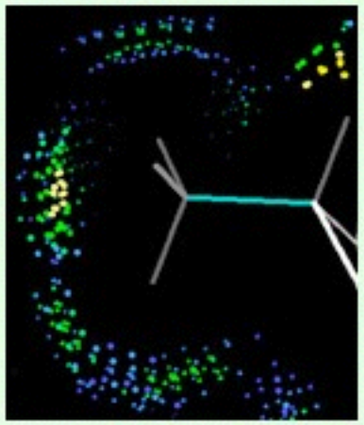
Top8000



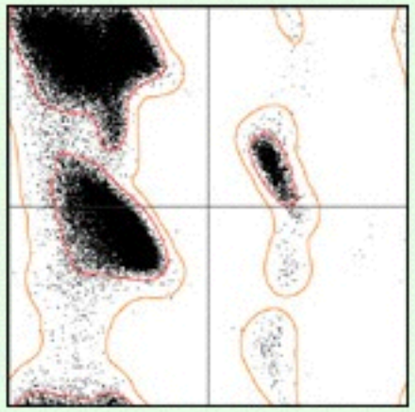
98% = Favored
99.95% = Allowed
Otherwise outlier


General
Glycine
Isoleucine/Valine
Pre-Proline
Trans-Proline
Cis-Proline

All-atom contacts, clashscore

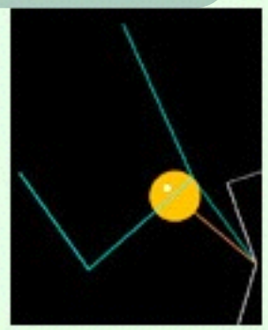


Ramachandran criteria

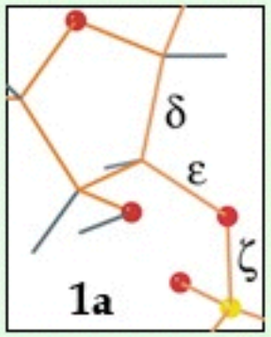


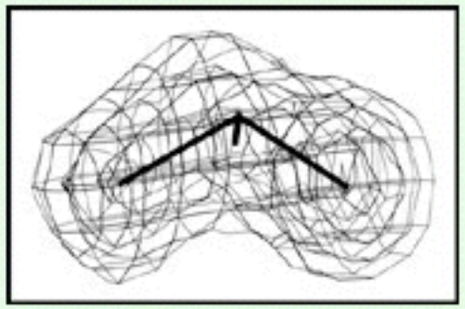
Sidechain rotamers

Geometry



RNA bb

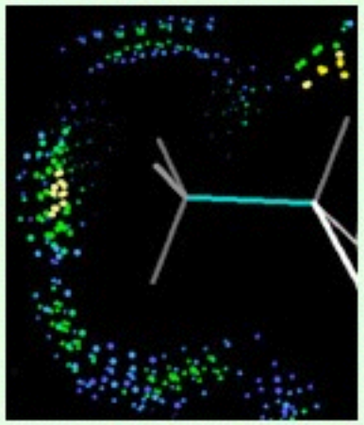




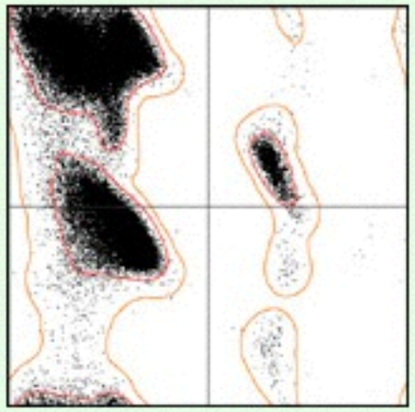
Crystallographic : R_{free}, electron density fit

Rotamers are tight and distinct

All-atom contacts, clashscore

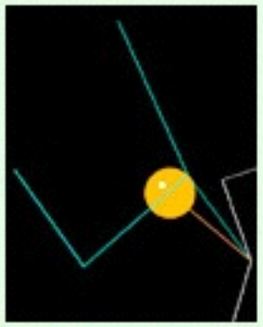


Ramachandran criteria

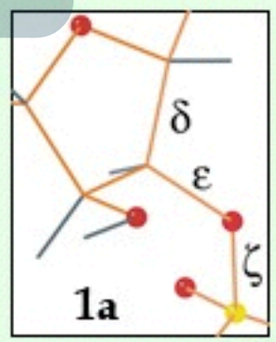


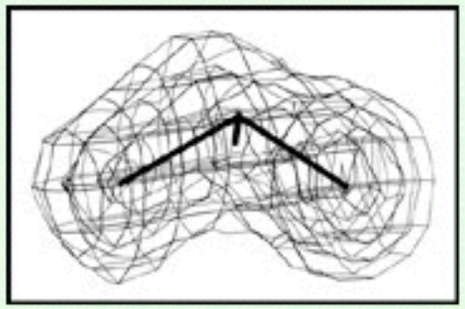
Sidechain rotamers

Geometry



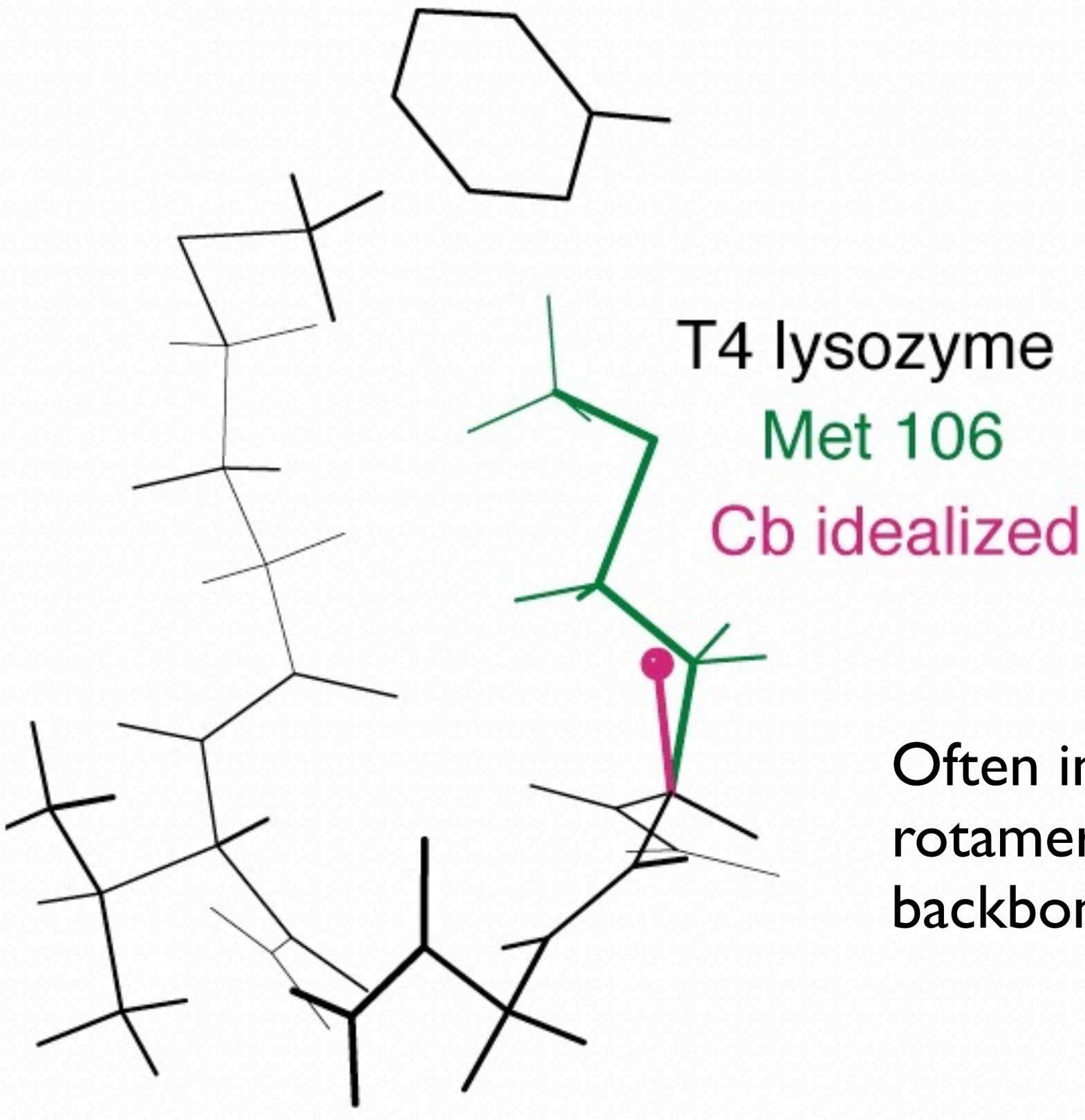
RNA bb





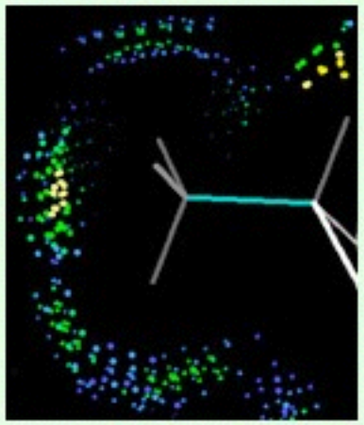
Crystallographic : R_{free}, electron density fit

Cβ Deviations

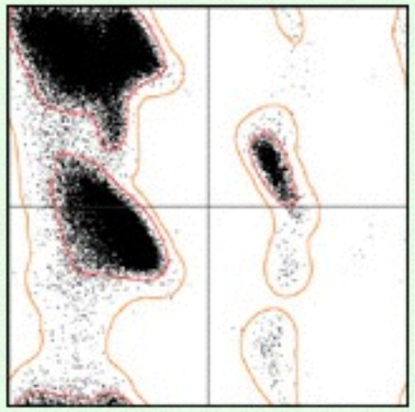


Often indicative of a wrong rotamer (or outlier) and backbone errors

All-atom contacts, clashscore

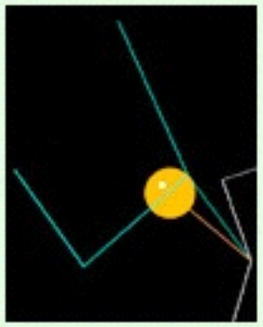


Ramachandran criteria

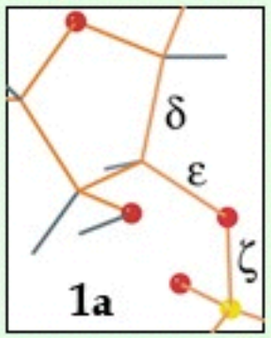


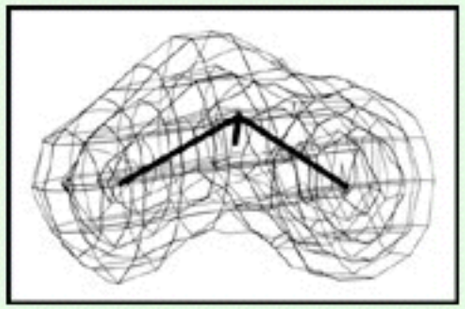
Sidechain rotamers

Geometry

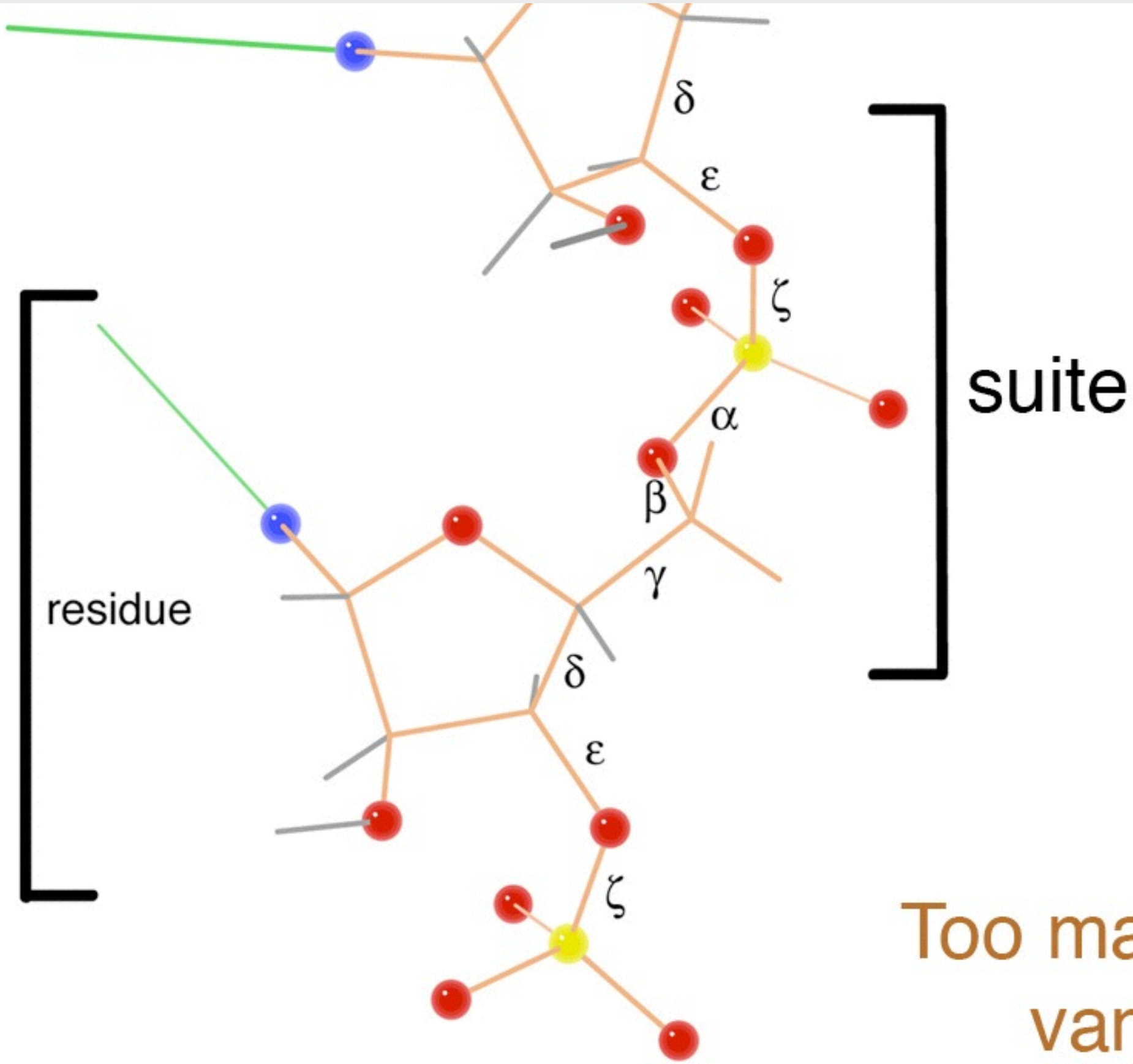


RNA bb

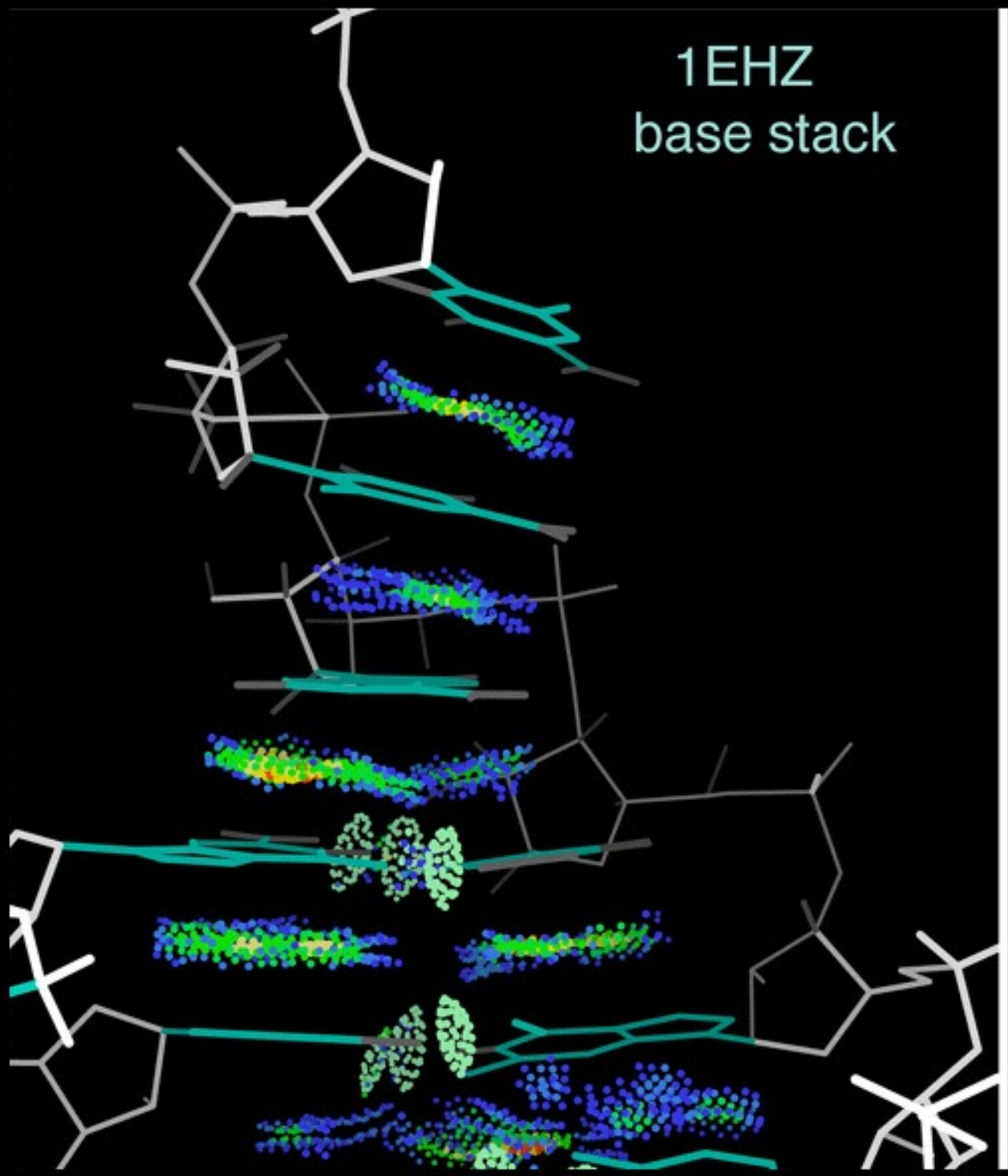


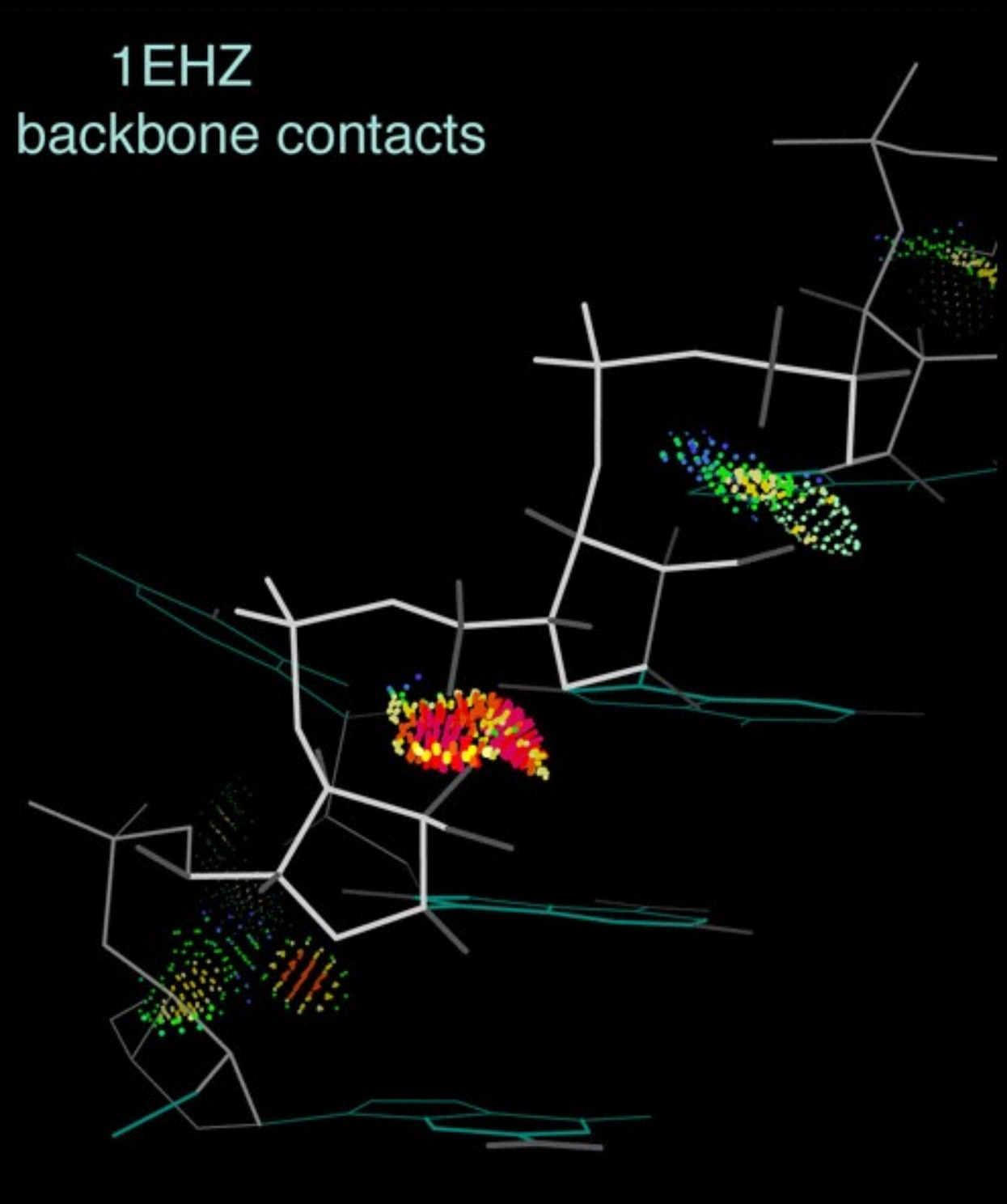


Crystallographic : R_{free}, electron density fit



Too many variables!


1EHZ
base stack


1EHZ
backbone contacts

Validation: basic recommendations

- The MolProbity server suggests these cutoffs:

  clashscore < **10**

  Ramachandran outliers **<= 0.2%**

  Ramachandran favored **>= 98%**

  Rotamer outliers **< 1%**

  C-beta deviations **= 0**

  Overall MolProbity score **<= d_min**

- There is no universal appropriate set of values for RMS(bonds) or RMS(angles); resolution dependent
  - but if these are above 0.02/2.0, there may be problems

General recommendations for better results

- <u>If you are running MR, make sure the starting model is as good as possible</u>
  - Re-refinement may be very helpful*
- Unless you have atomic-resolution data, make sure you optimize the X-ray versus geometry weight at the final stages to get the best possible geometry
- At low resolution, additional restraints are extremely helpful
- **Perform validation throughout refinement, not just before you deposit in the PDB or publish**

* See Joosten et al. (2009) for a general discussion. In our own internal tests with an automated wrapper for phenix.refine, we have found that at least 25% of PDB entries can be improved by a drop in R-free of 0.02 or greater, and another 25% by 0.01-0.02.

How to tell when your structure is "finished"

- <u>There is no objective, absolute set of criteria for this!</u>
- Better questions to be asking:
  - Have all obvious geometry errors been corrected?
  - Do all residues in the model have a reasonable fit to the 2mFo-DFc map?
  - Is the model complete? Have all interpretable difference map features been accounted for?
  - Are the various statistics consistent with (and ideally superior to) similar structures at the same resolution?
  - Does it make sense biologically?
  - If I were asked to review this structure from a competitor, would I recommend publication?

validation in PHENIX

phenix.refine results

protein kinase A
PDB ID: 3dnd

phenix.refine

Preferences Help Run Abort Save Graphics ReadySet NCS TLS Restraints Xtriage

Configure | Refine_2 | Refine_3

Log output | Run status | Results | MolProbity | Real-space correlation | Atomic properties

Output files

Directory: /Users/jheadd/labwork/LBNL/phenix_workshop/pka-compare/Refine_3

File name	Contents
pka-compare_refine_3.eff	Effective parameters for this run
pka-compare_refine_3.geo	Geometry restraints before refine...
pka-compare_refine_3.log	phenix.refine log file
pka-compare_refine_3.mtz	Map coefficients for Coot
pka-compare_refine_3.pdb	Refined model
pka-compare_refine_3_info.txt	Run summary in text format

Open in Coot

Open in PyMOL

Sequence viewer

Convert map coefficients MTZ to CCP4 maps

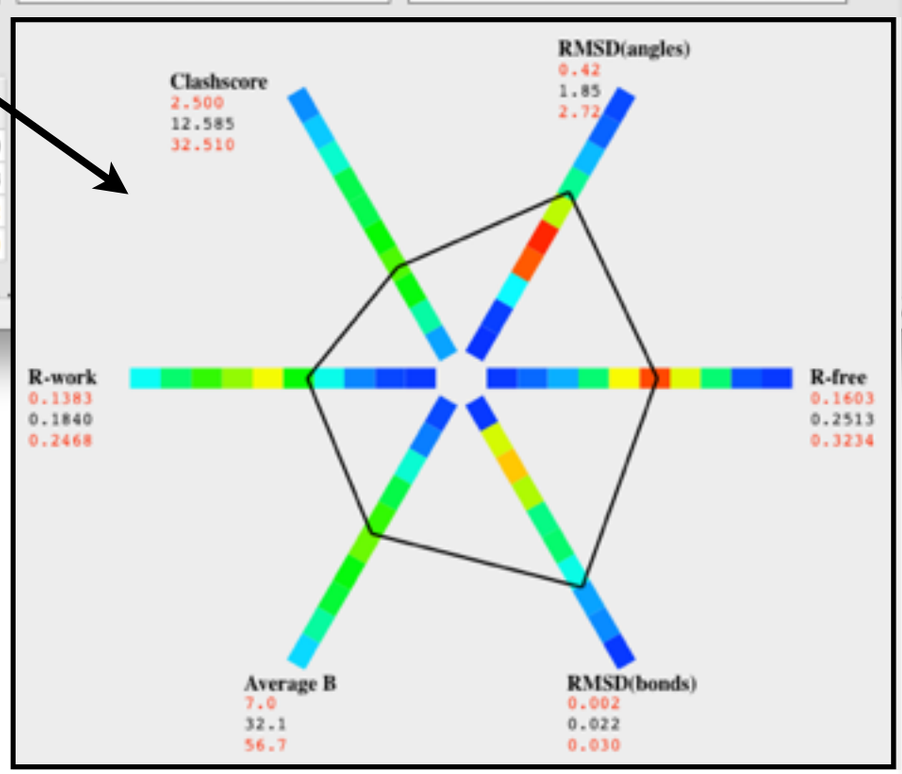
Refinement statistics

Compare statistics | Plot statistics by cycle | Plot statistics by resolution

Before and after refinement:

	Starting	Final
R-work	0.1908	0.1840
R-free	0.2533	0.2513
Bonds	0.022	0.022
Angles	1.850	1.850

Idle



Coot

File Edit Calculate Draw Measures Validate HID About Extensions PHENIX

Reset View | Display Manager | Hydrogens on | Connected to PHENIX | R/RC | Map

(mol. no: 33) CA /1/A/189 LYS occ: 1.00 bf: 19.88 ele: C pos: (11.92, 4.16,-14...

MolProbity summary

phenix.refine

Preferences | Help | Run | Abort | Save | Graphics | ReadySet | NCS | TLS | Restraints | Xtriage

Configure | Refine_2 | Refine_3

Log output | Run status | Results | MolProbity | Real-space correlation | Atomic properties

Summary | Basic geometry | Protein | Clashes

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. You can start this process by clicking the MolProbity button on the left.

Basic statistics for pka-compare_refine_3.pdb:

Ramachandran outliers:	0.3%	(Goal : < 0.2%)	**Ramachandran favored:**	97.5%	(Goal : > 98%)
Rotamer outliers:	7.7%	(Goal : 1%)	**C-beta outliers:**	6	(Goal : 0)
Clashscore:	12.59		**Overall score:**	2.40	

Show validation in KiNG

Missing atoms

No missing non-hydrogen atoms detected.

Idle

Project: pka

KING 2.20

File Edit Views Display Tools Help

Kinemage #1

pka-compare_refi

mainchain
Calphas
sidechain
water
chain A
hets
Rota outliers
Rama outliers
angle dev
length dev
Cbeta dev
chain I
vdw contact
small overlap
bad overlap
H-bonds
McMc contacts
ScSc contacts
McSc contacts
Hets contacts
dots

Animate

Zoom | Clipping | Pick center | Markers | Show text | Show hierarchy

basic geometry

phenix.refine

Preferences | Help | Run | Abort | Save | Graphics | ReadySet | NCS | TLS | Restraints | Xtriage

Configure | Refine_2 | Refine_3

Log output | Run status | Results | MolProbity | Real-space correlation | Atomic properties

Summary | Basic geometry | Protein | Clashes

Bond length restraints

Number of restraints: 3038

RMS(deviation): 0.022

Max. deviation: 0.163

Number of outliers > 4sigma: 4

List of outliers (sorted by deviation):

Atom 1	Atom 2	Ideal value	Model value	Deviation
P TPO A 197	O3P TPO A 197	1.610	1.482	6.4
P TPO A 197	O2P TPO A 197	1.610	1.509	5.1
C7 LL2 A 351	N10 LL2 A 351	1.430	1.335	4.8
CG1 ILE A 163	CD1 ILE A 163	1.513	1.676	4.2

Bond angle restraints

Number of restraints: 4101

RMS(deviation): 1.850

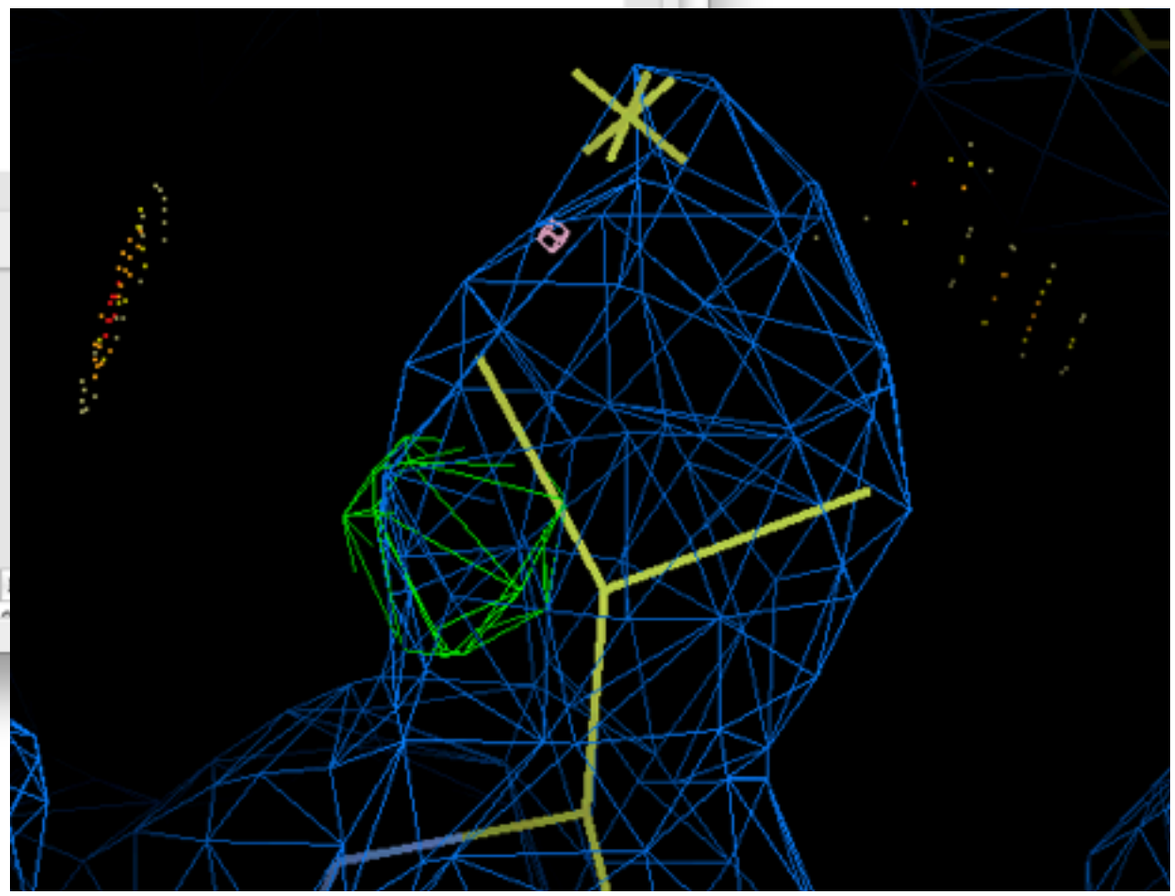
Max. deviation: 16.301

Number of outliers > 4sigma: 6

List of outliers (sorted by deviation):

Atoms	Ideal value

Idle



Ramachandran outliers

phenix.refine

Preferences | Help | Run | Abort | Save | Graphics | ReadySet | NCS | TLS | Restraints | Xtriage

Configure | Refine_2 | Refine_3

Log output | Run status | Results | MolProbity | Real-space correlation | Atomic properties

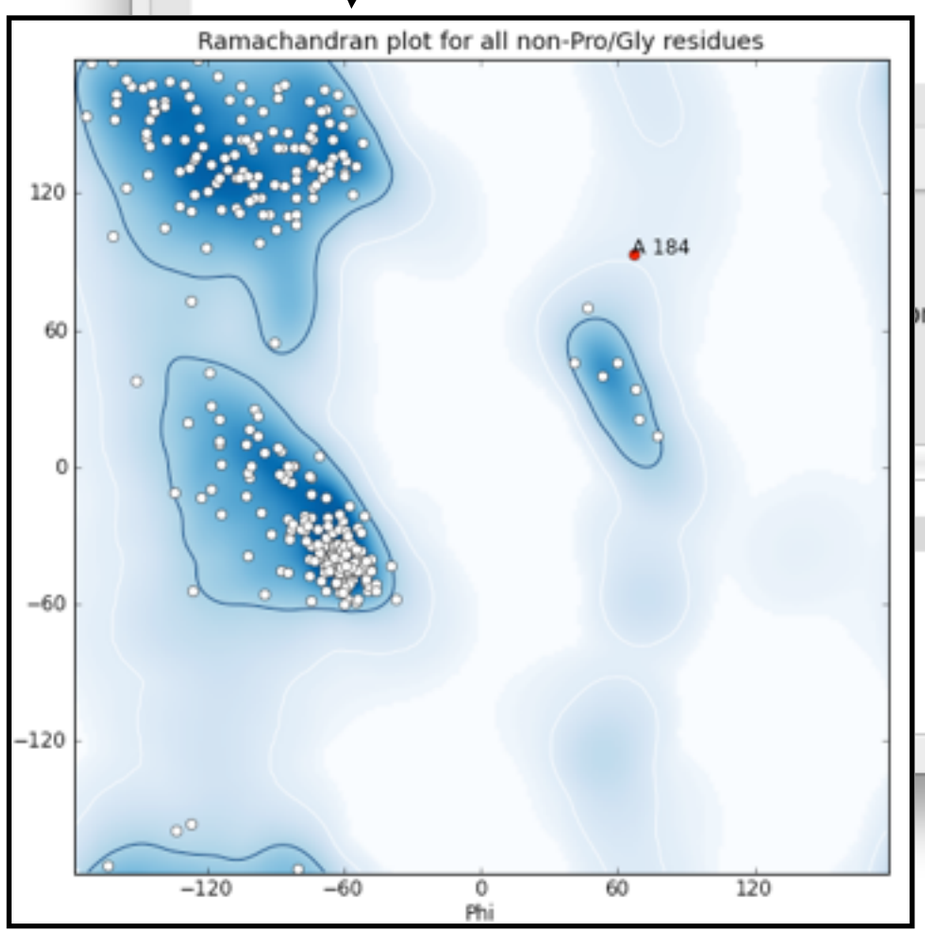
Summary | Basic geometry | Protein | Clashes

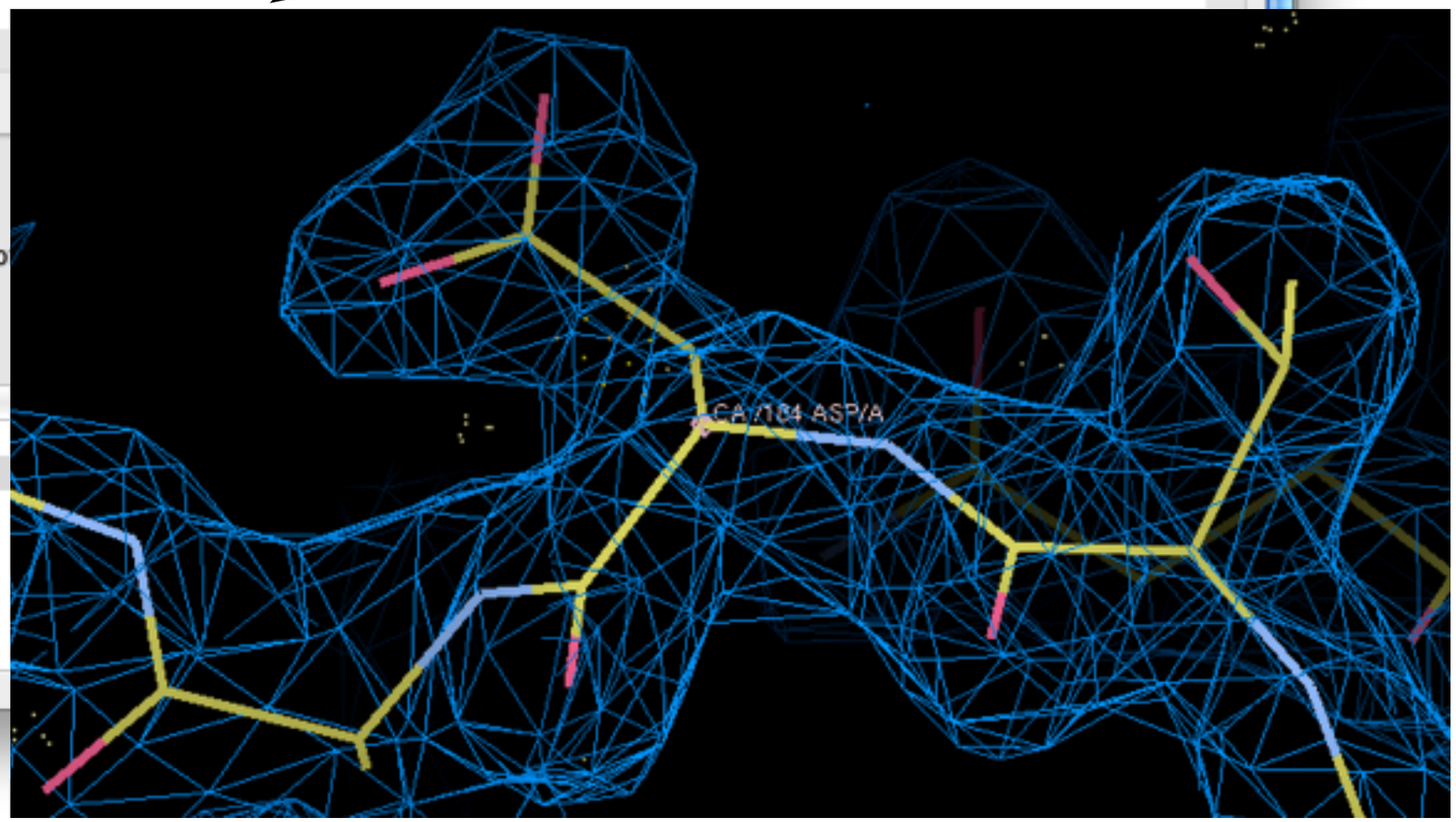
Ramachandran analysis

View Ramachandran plots

Ramachandran outliers:

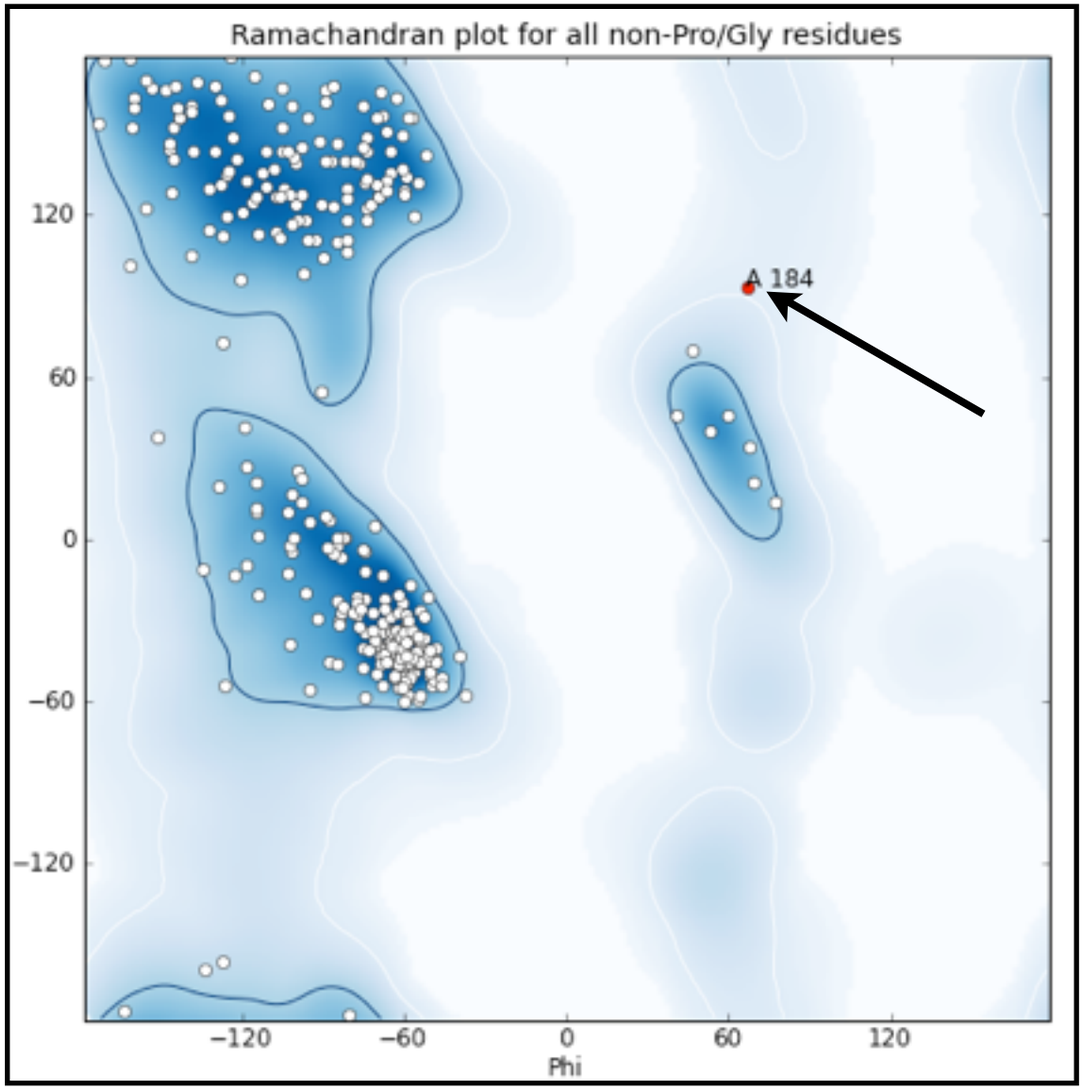
Chain	Residue	Residue type	Score	Phi	Psi
A	ASP 184	General	0.04	66.8	93.0



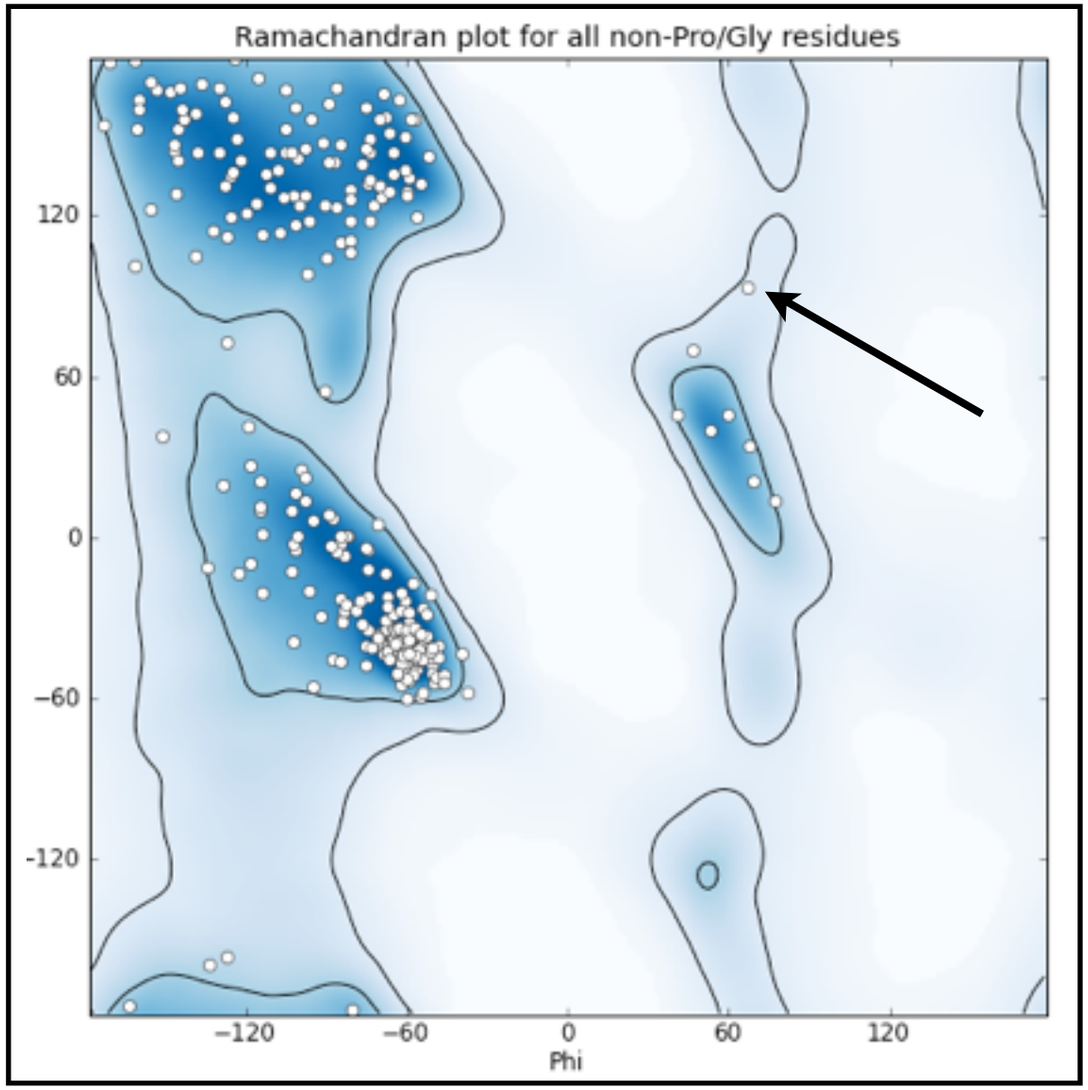


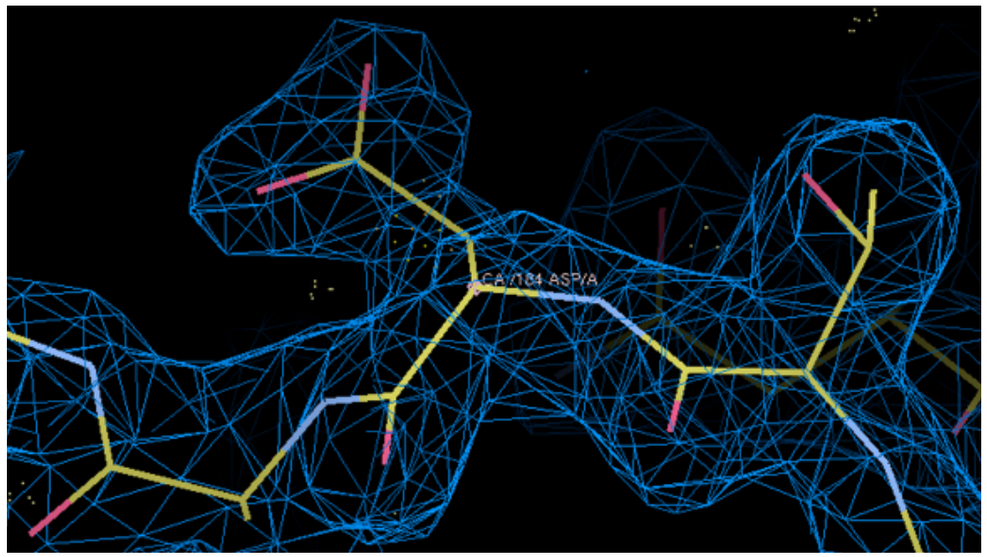
Ramachandran outliers

Top5000



Top8000





rotamer outliers

phenix.refine

Preferences Help Run Abort Save Graphics ReadySet NCS TLS Restraints Xtriage

Configure | Refine_2 | Refine_3

Log output | Run status | Results | MolProbity | Real-space correlation | Atomic properties

Summary | Basic geometry | Protein | Clashes

Rotamer analysis

View Chi1-Chi2 plots

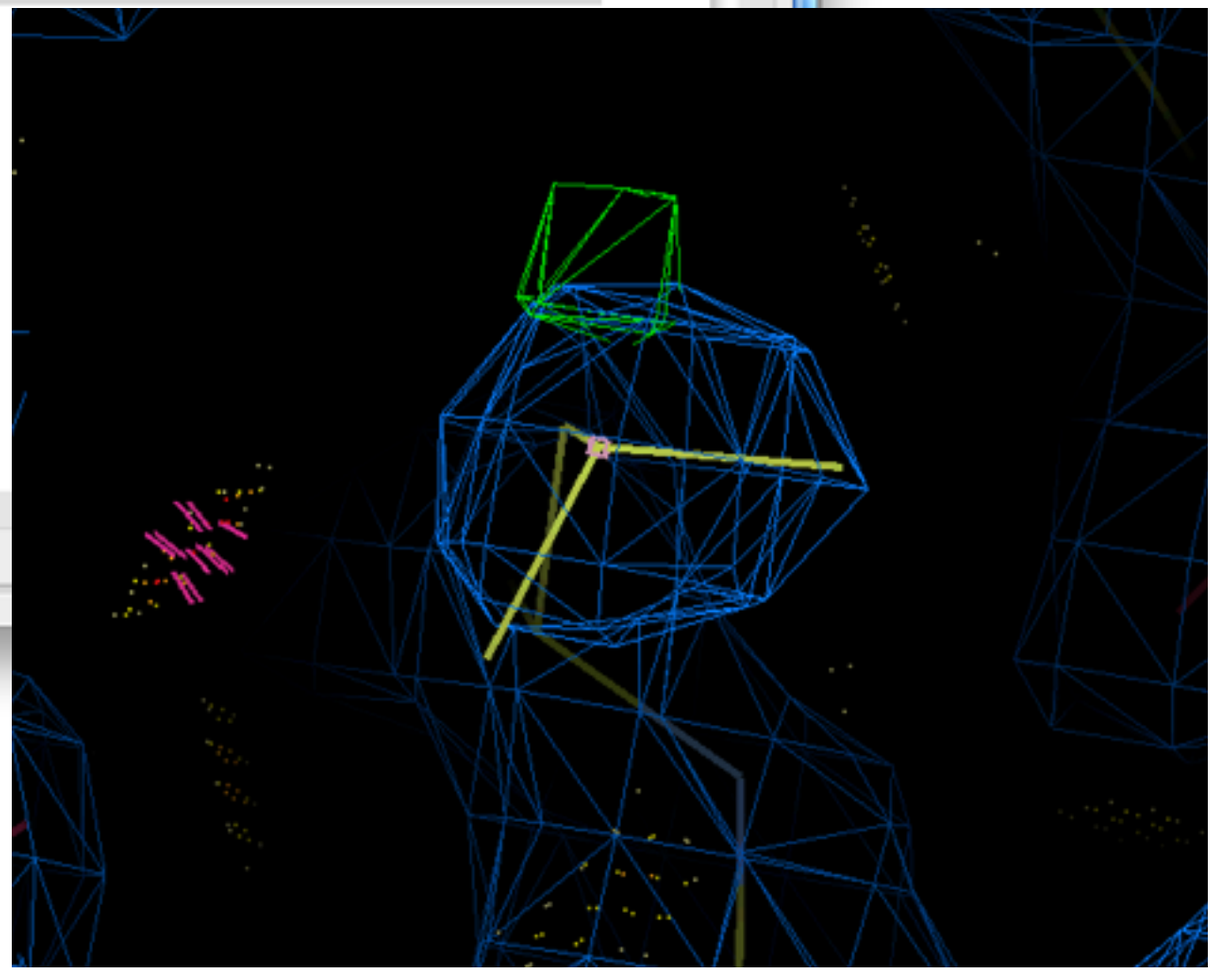
Note that although a residue may lie in the favored regions of the Chi1-Chi2 plot, outliers are flagged based on the distribution of all non-branched Chi angles in a residue.

Rotamer outliers:

Chain	Residue	Score	Chi1	Chi2	Chi3	Chi4
A	MET 63	0.44	274.5	116.6	58.9	-
A	LEU 74	0.84	297.8	330.6	-	-
A	LEU 160	0.00	273.9	12.6		
A	LEU 162	0.00	263.9	6.6		
A	VAL 191	0.16	221.0	-		
A	LEU 198	0.00	199.0	217.9		
A	LYS 217	0.99	[illegible]	158.8		
A	LEU 268	0.00	216.9	216.0		
A	LEU 269	0.00	258.0	11.4		
A	LEU 272	0.00	257.5	[illegible]		
A	LYS 285	0.00	66.7	[illegible]		
A	LYS 295	0.00	305.0	53.2		
A	LYS 309	0.01	39.0	191.1		
A	GLU 311	0.02	53.2	131.3		
A	LYS 317	0.00	343.7	64.5		
A	GLU 331	0.62	40.4	150.9		
A	LYS 345	0.00	30.2	263.2		

C-beta deviation analysis

Idle



Cβ deviations

phenix.refine

Preferences | Help | Run | Abort | Save | Graphics | ReadySet | NCS | TLS | Restraints | Xtriage

Configure | Refine_2 | Refine_3

Log output | Run status | Results | MolProbity | Real-space correlation | Atomic properties

Summary | Basic geometry | Protein | Clashes

A	LYS	317	0.00	343.7	64.5	166.7	27.6
A	GLU	331	0.62	40.4	150.9	100.7	-
A	LYS	345	0.00	30.2	263.2	118.2	60.2

C-beta deviation analysis

C-beta position outliers:

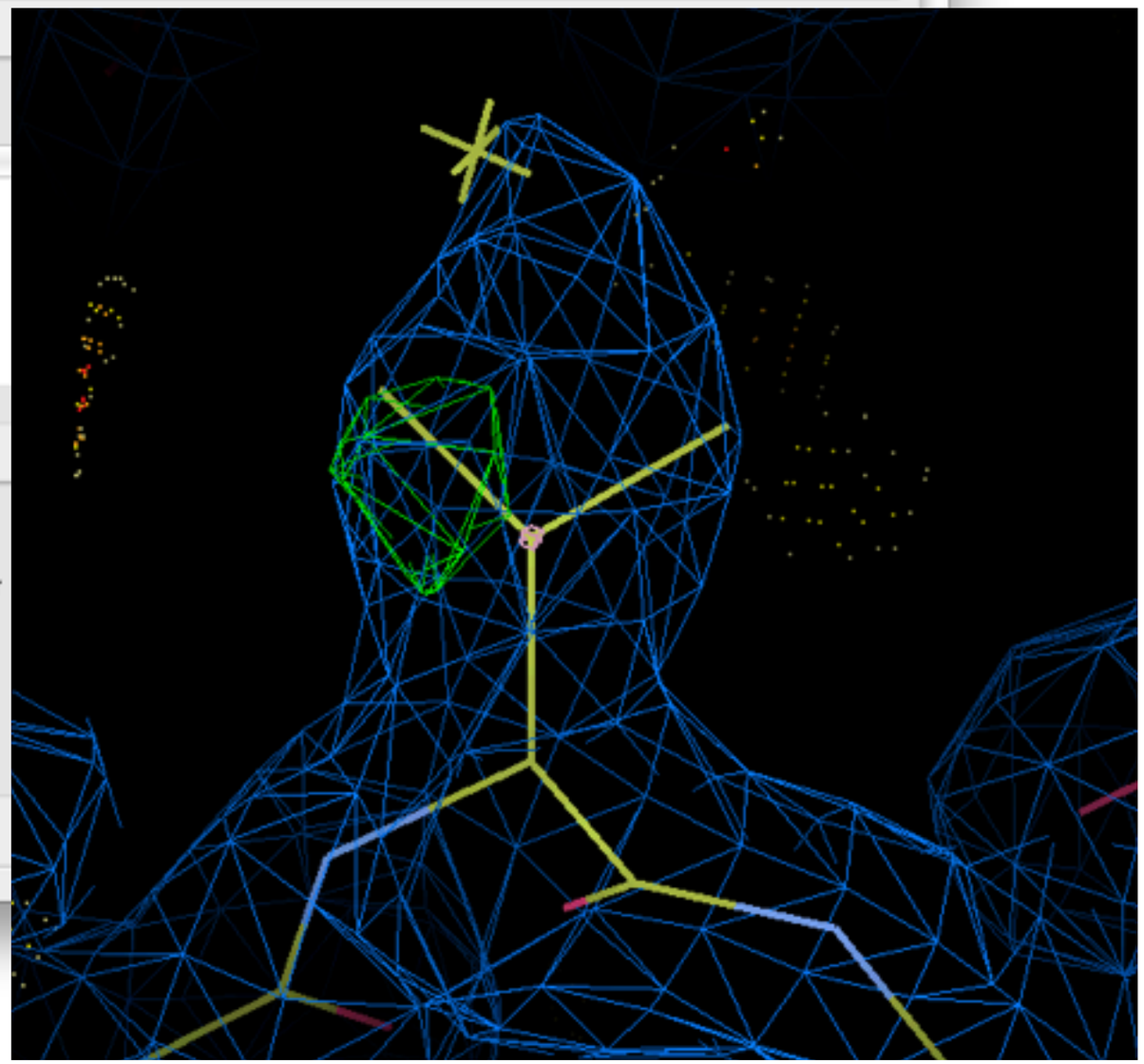
Chain	Residue	Altloc	Deviation	Angle
A	PHE 54		0.357	-124.13
A	GLU 140		0.306	-105.19
A	ILE 163		0.252	106.89
A	GLU 208		0.306	99.47
A	VAL 251		0.255	-121.68
A	LYS 309		0.288	-96.30

Backwards sidechains

Recommended sidechain flips:

REDUCE (phenix.reduce) has been run on your file to add hydrogens necessary for identifying clashes in the model. Asymmetric sidechains which required flipping have been identified; these have been changed in pka-compare_refine_3.reduce.pdb.

Idle



steric clashes

phenix.refine

Preferences | Help | Run | Abort | Save | Graphics | ReadySet | NCS | TLS | Restraints | Xtriage

Configure | Refine_2 | Refine_3

Log output | Run status | Results | MolProbity | Real-space correlation | Atomic properties

Summary | Basic geometry | Protein | Clashes

All-atom contact analysis

Coot display

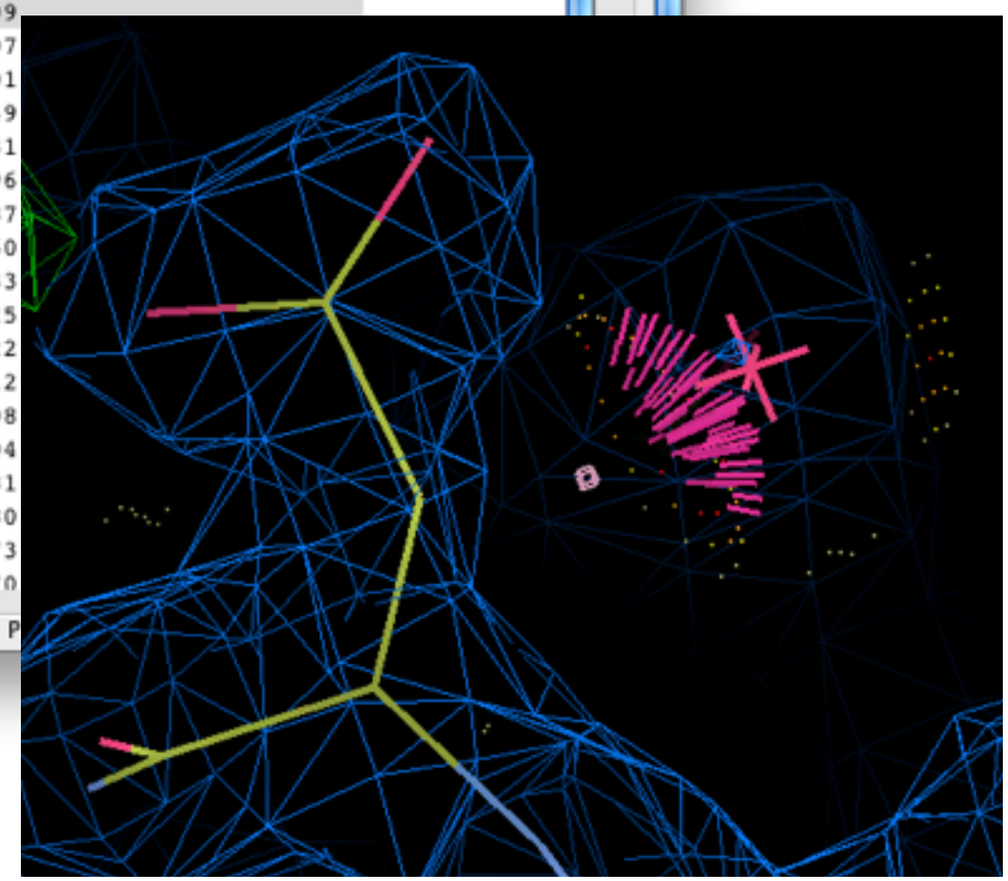
☑ Show Probe dots in Coot ☑ Only show bad overlaps Reload data Re-run PROBE

Bad contacts from PROBE: 74 overlapping atom pairs

This list summarizes all severe clashes (more than 0.4 Angstrom overlap) found by PROBE; you can view these graphically in Coot. If no hydrogens were present, REDUCE was used to add them prior to running PROBE.

Atom 1	Atom 2	Overlap
A 184 ASP HB2	A 546 HOH O	1.209
A 39 HIS CD2	A 41 ASP H	0.907
A 317 LYS H	A 317 LYS HD2	0.901
A 135 ILE HD11	A 585 HOH O	0.849
A 317 LYS CD	A 317 LYS H	0.831
A 295 LYS HD3	A 295 LYS N	0.796
A 295 LYS H	A 295 LYS HD3	0.787
A 39 HIS HD2	A 41 ASP H	0.760
A 295 LYS CD	A 295 LYS H	0.733
A 91 GLU OE2	A 353 HOH O	0.725
A 268 LEU HD22	A 272 LEU HD22	0.722
A 177 GLN HG3	A 554 HOH O	0.712
A 135 ILE CD1	A 585 HOH O	0.708
A 18 PHE HD2	A 19 LEU HD13	0.704
A 17 GLU O	A 21 LYS HD3	0.681
A 21 LYS HD2	A 21 LYS N	0.680
A 275 VAL HG21	A 577 HOH O	0.673
A 275 VAL CG2	A 577 HOH O	0.670

Idle



real-space correlation

phenix.refine

Preferences | Help | Run | Abort | Save | Graphics | ReadySet | NCS | TLS | Restraints | Xtriage

Configure | Refine_2 | Refine_3

Log output | Run status | Results | MolProbity | Real-space correlation | Atomic properties

Real-space correlation

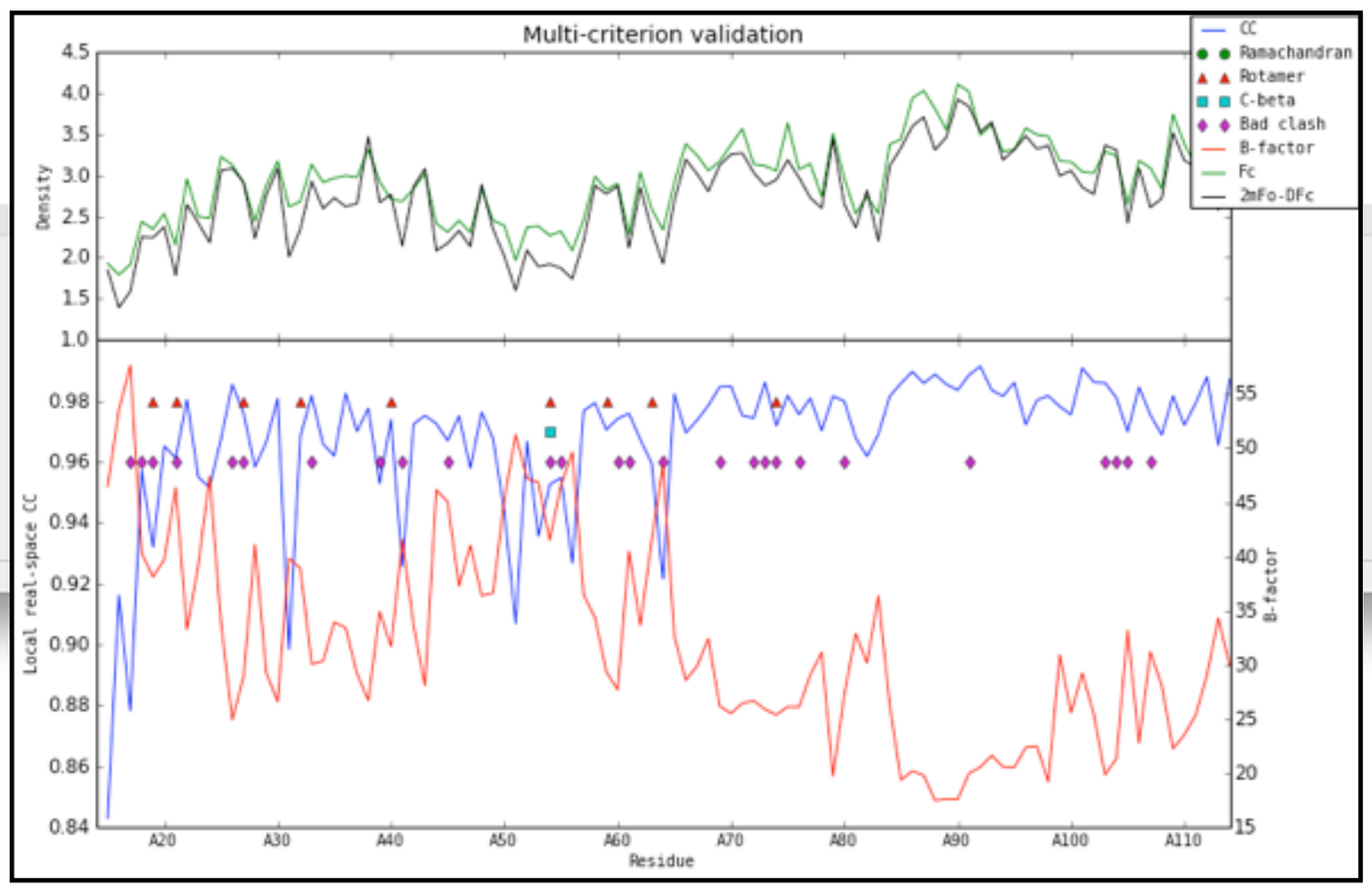
If you have a graphics window open, clicking an atom or residue in the list will recenter the display (you may disable this in the "Graphics" section of the preferences). Click on a column label to re-sort the list by values in that column.

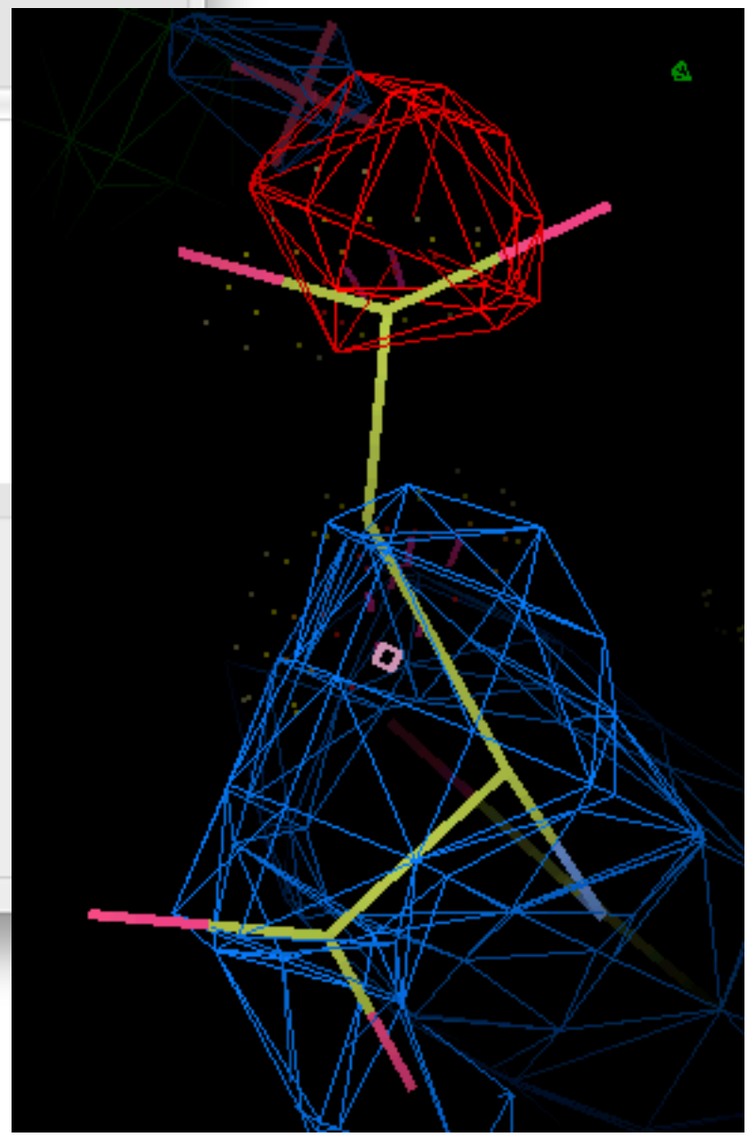
Multi-criterion plot

By residue: Show only residues with CC less than 0.80 Update list

Chain	Residue	CC	B-factor	Occ.	Fc	2mFo-DFc
I	ASP 24	0.750	65.57	1.00	1.81	1.17
A	GLU 334	0.795	81.62	1.00	1.29	0.84

Idle





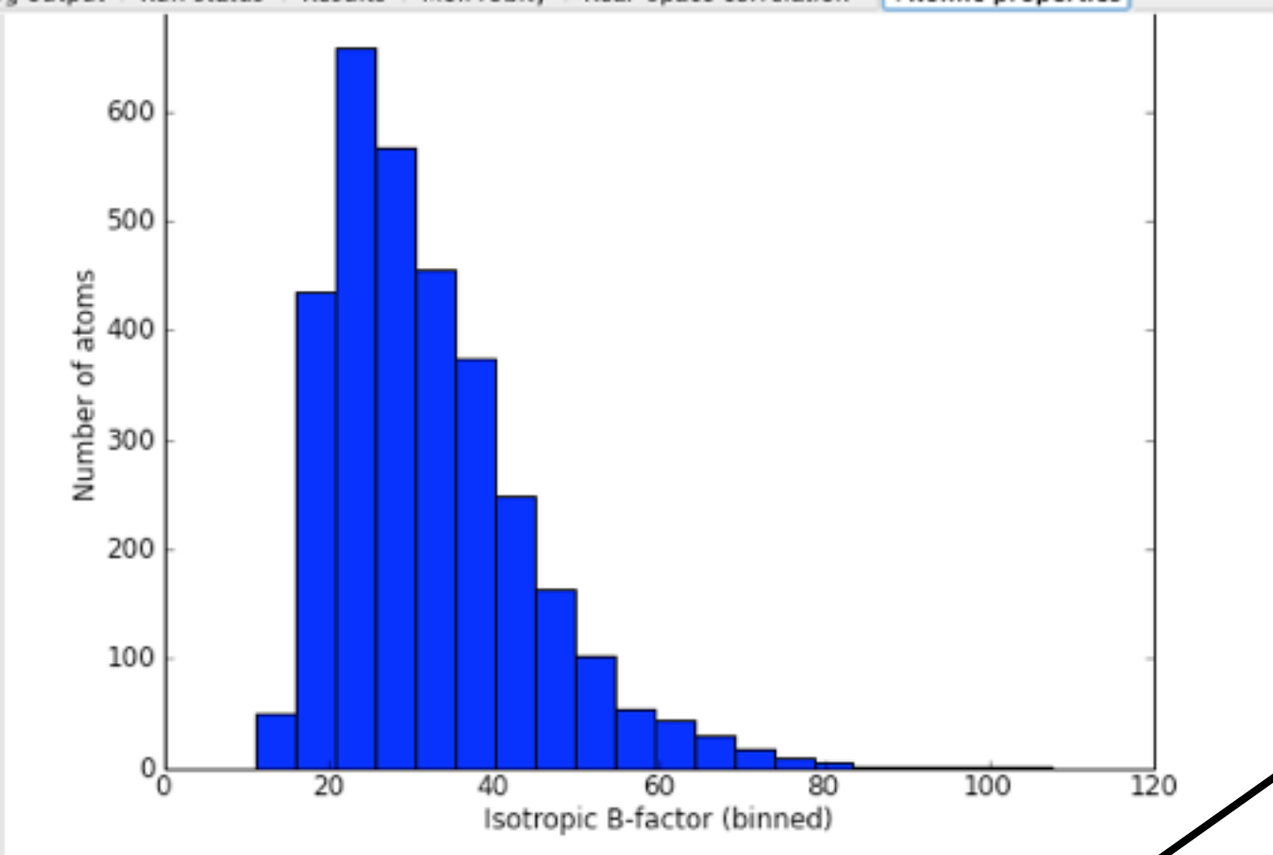
atomic properties

phenix.refine

Preferences Help Run Abort Save Graphics ReadySet NCS TLS Restraints Xtriage

Configure | Refine_2 | Refine_3

Log output | Run status | Results | MolProbity | Real-space correlation | Atomic properties



Suspiciously high B-factors

The table below lists all isotropic ADPs with values greater than four standard deviations above the mean value for this structure. Although a high B-factor is not necessarily wrong, it may be worth double-checking the atomic positions, occupancies, or (rarely) element types.

Atom ID	Isotropic B-factor	Occupancy
OE1 GLU A 334	107.58	1.00
CD GLU A 334	101.13	1.00
OE2 GLU A 334	93.83	1.00

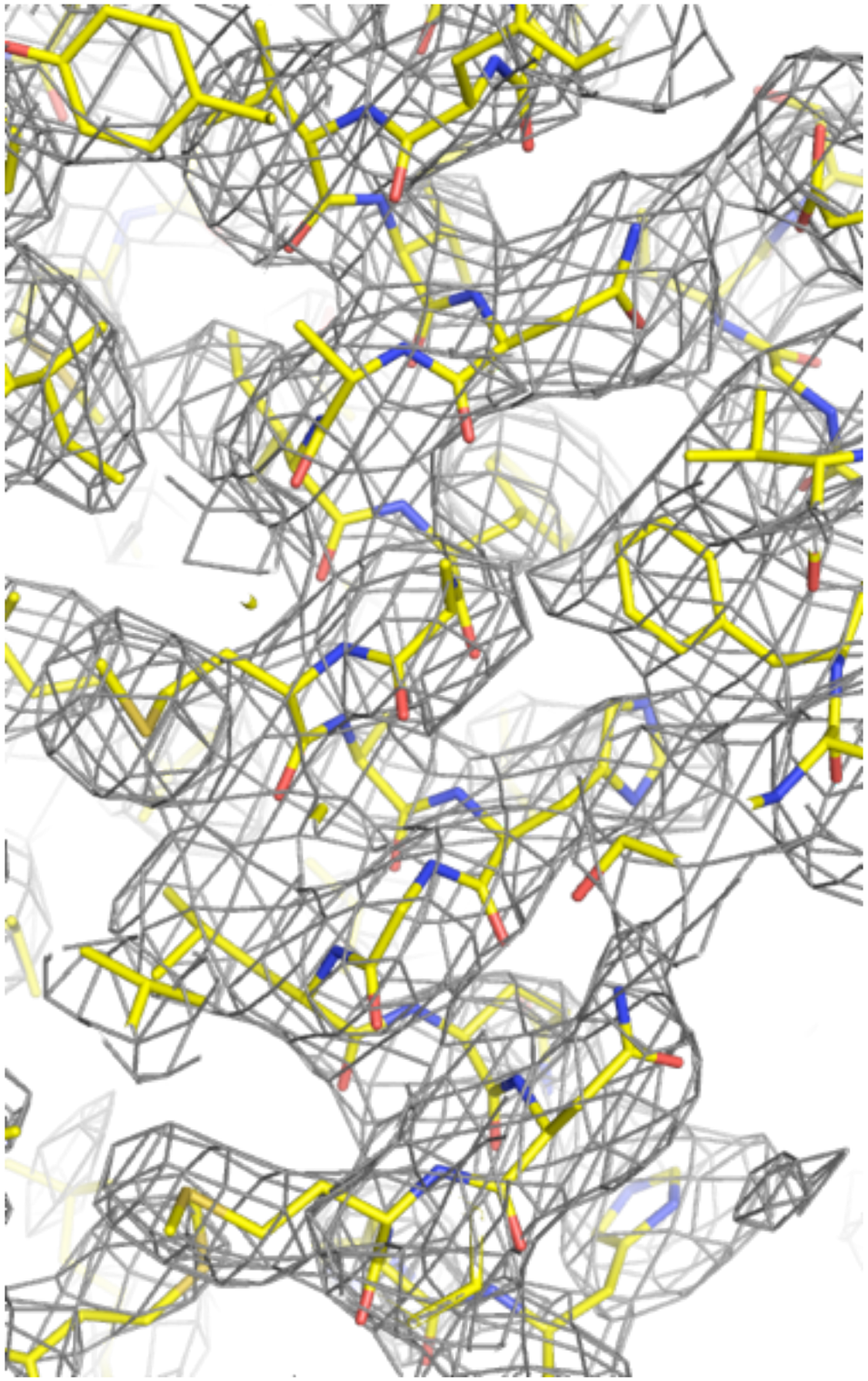
Idle

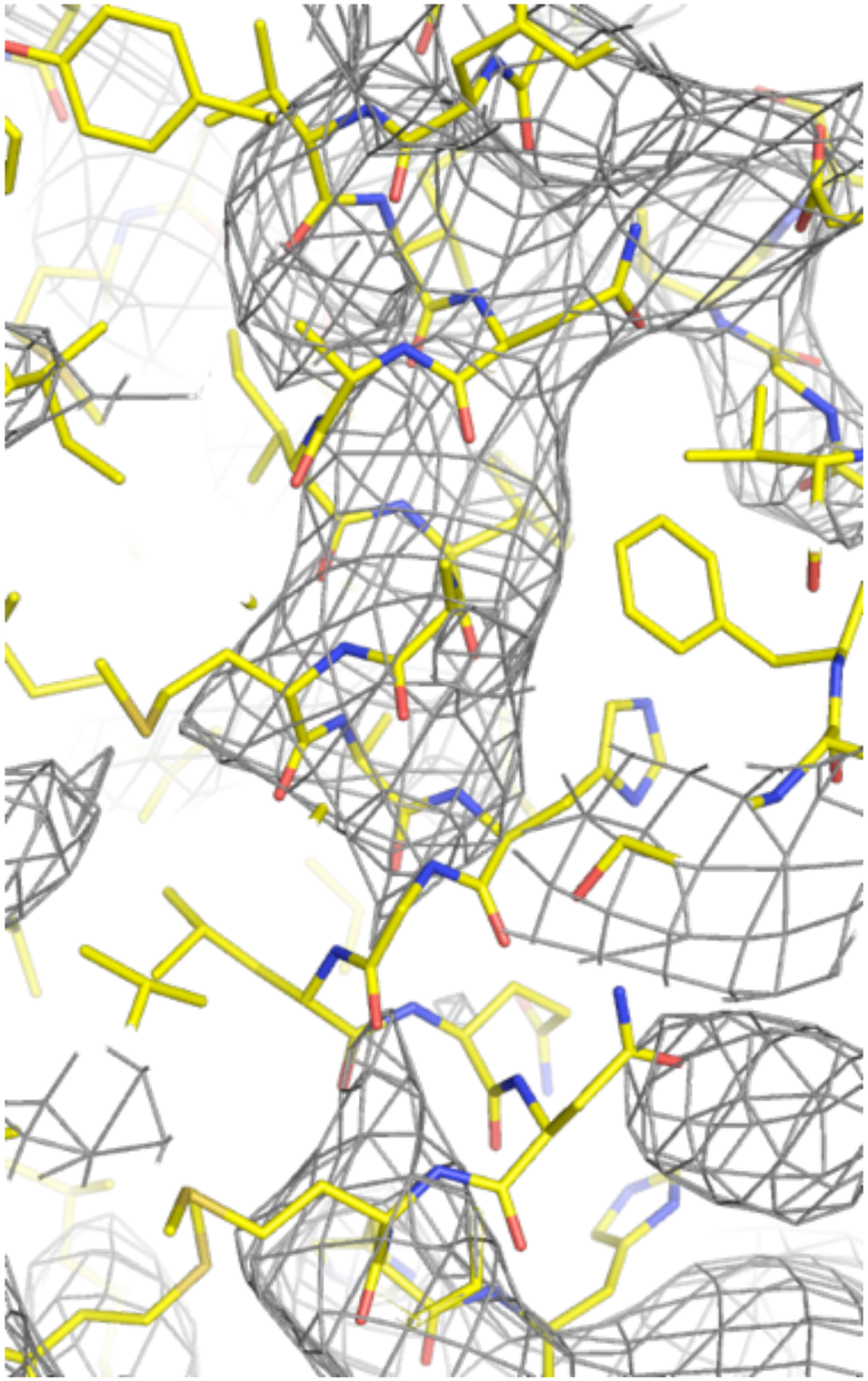
Project: pka-compare

Validating your electron density maps

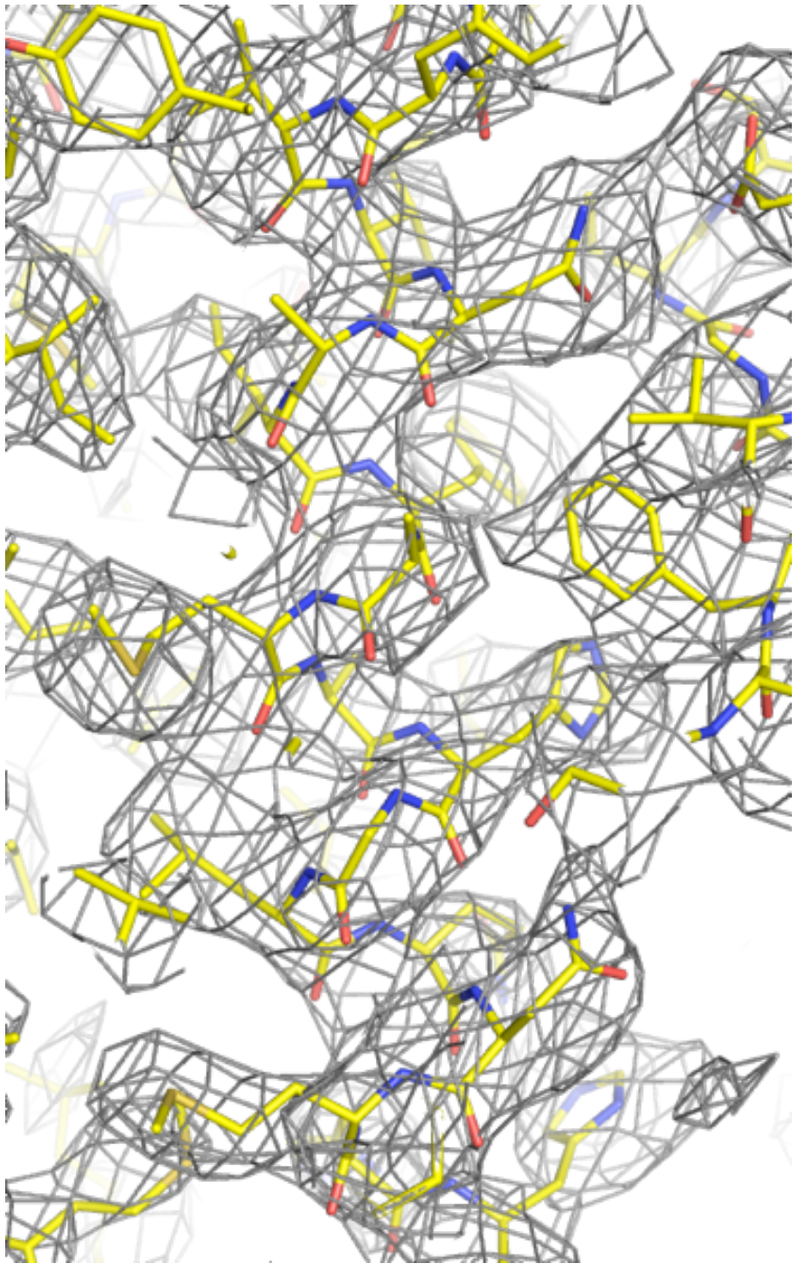
Model bias: a synthetic example

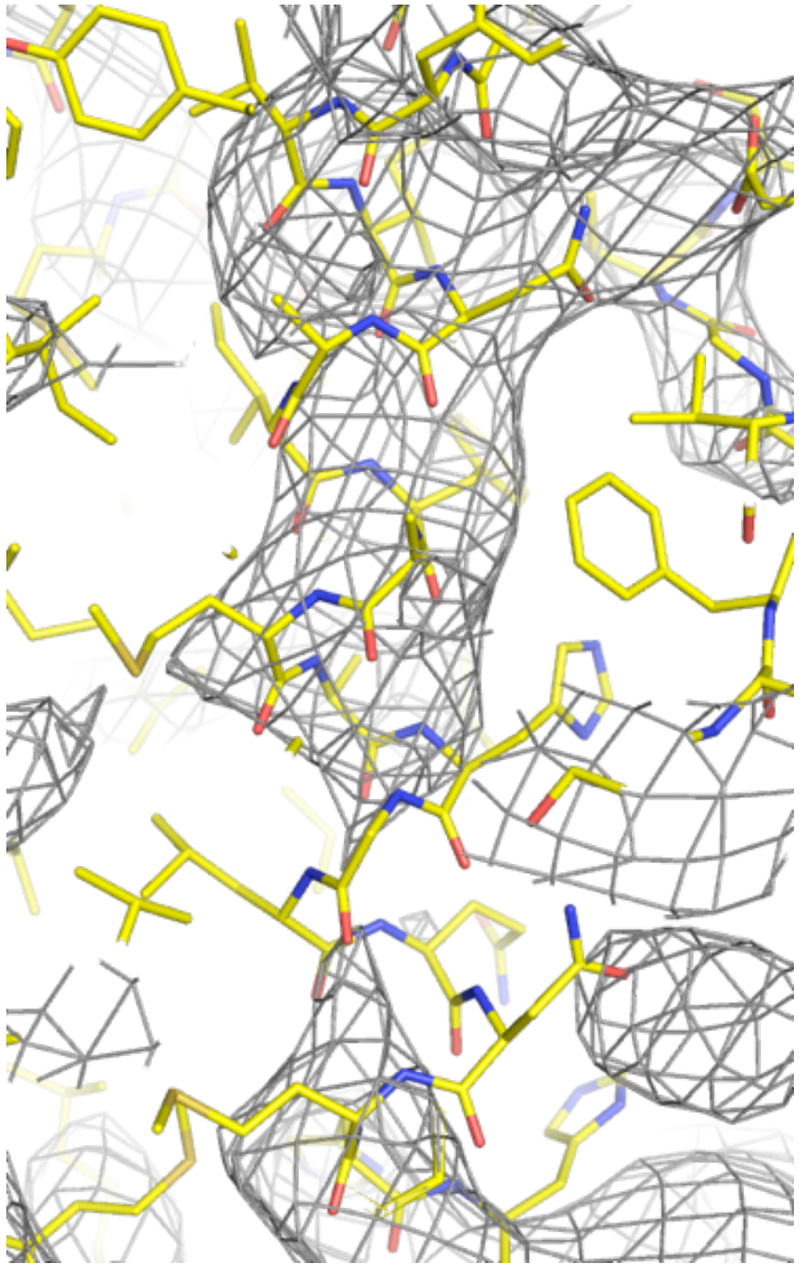
Which of these maps is real?

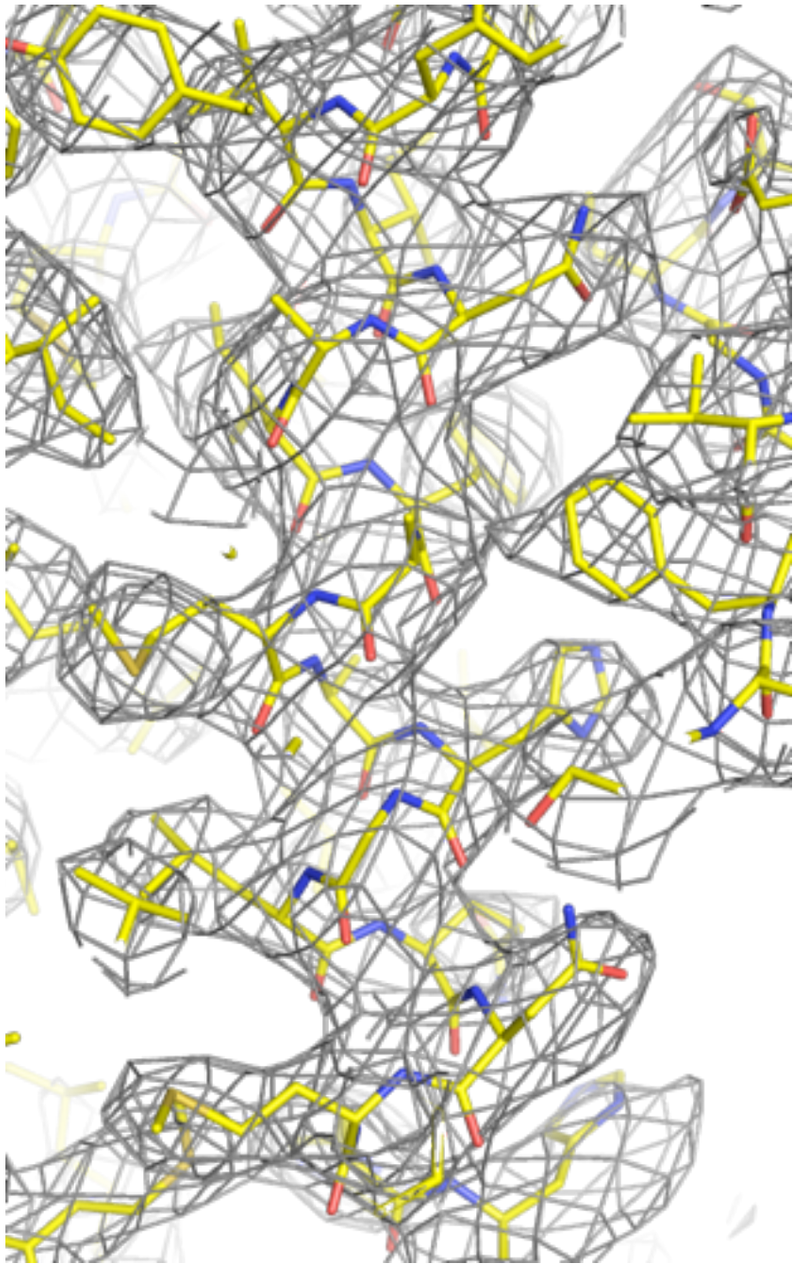




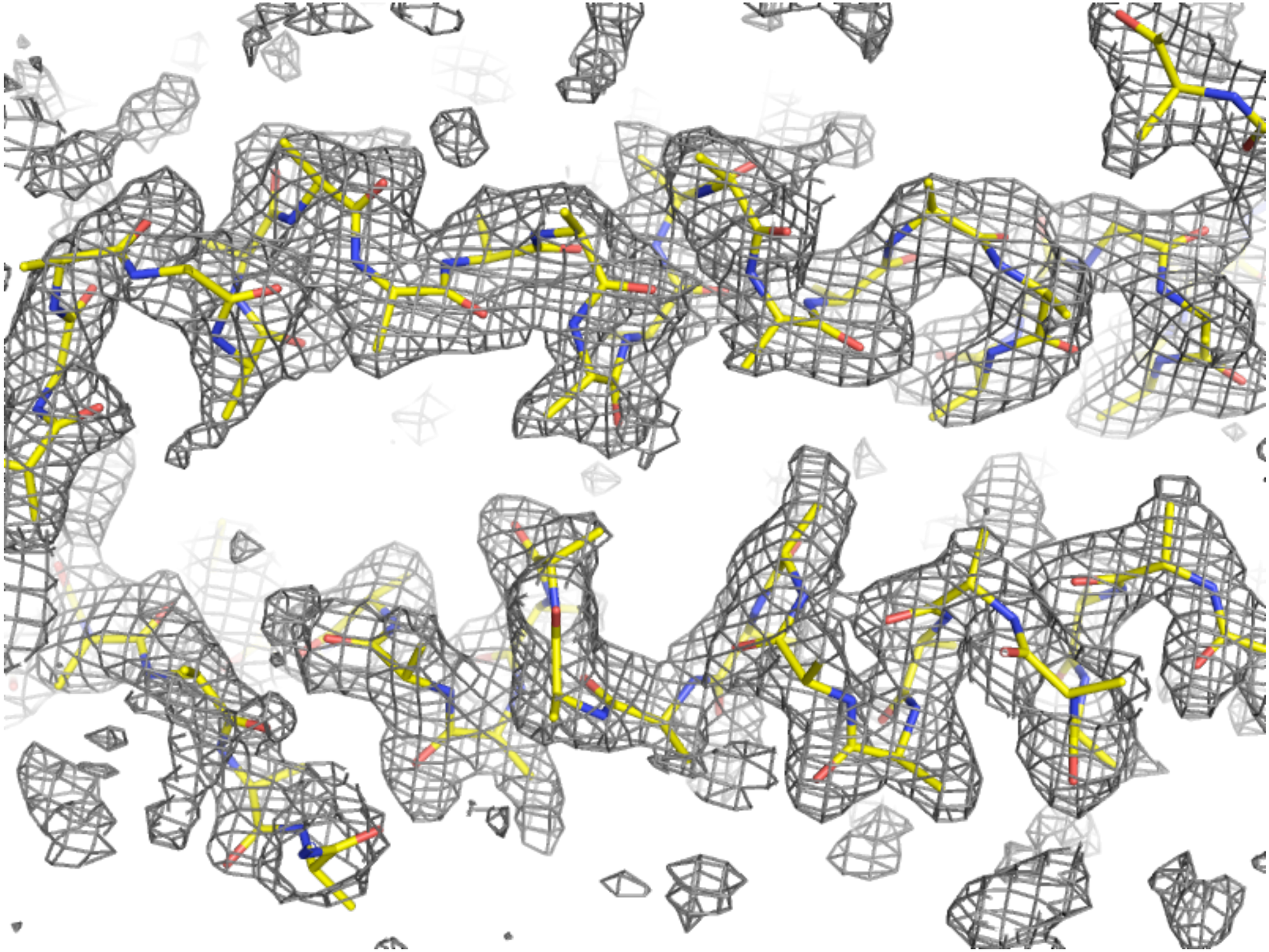
Model bias: a synthetic example at 4.0Å





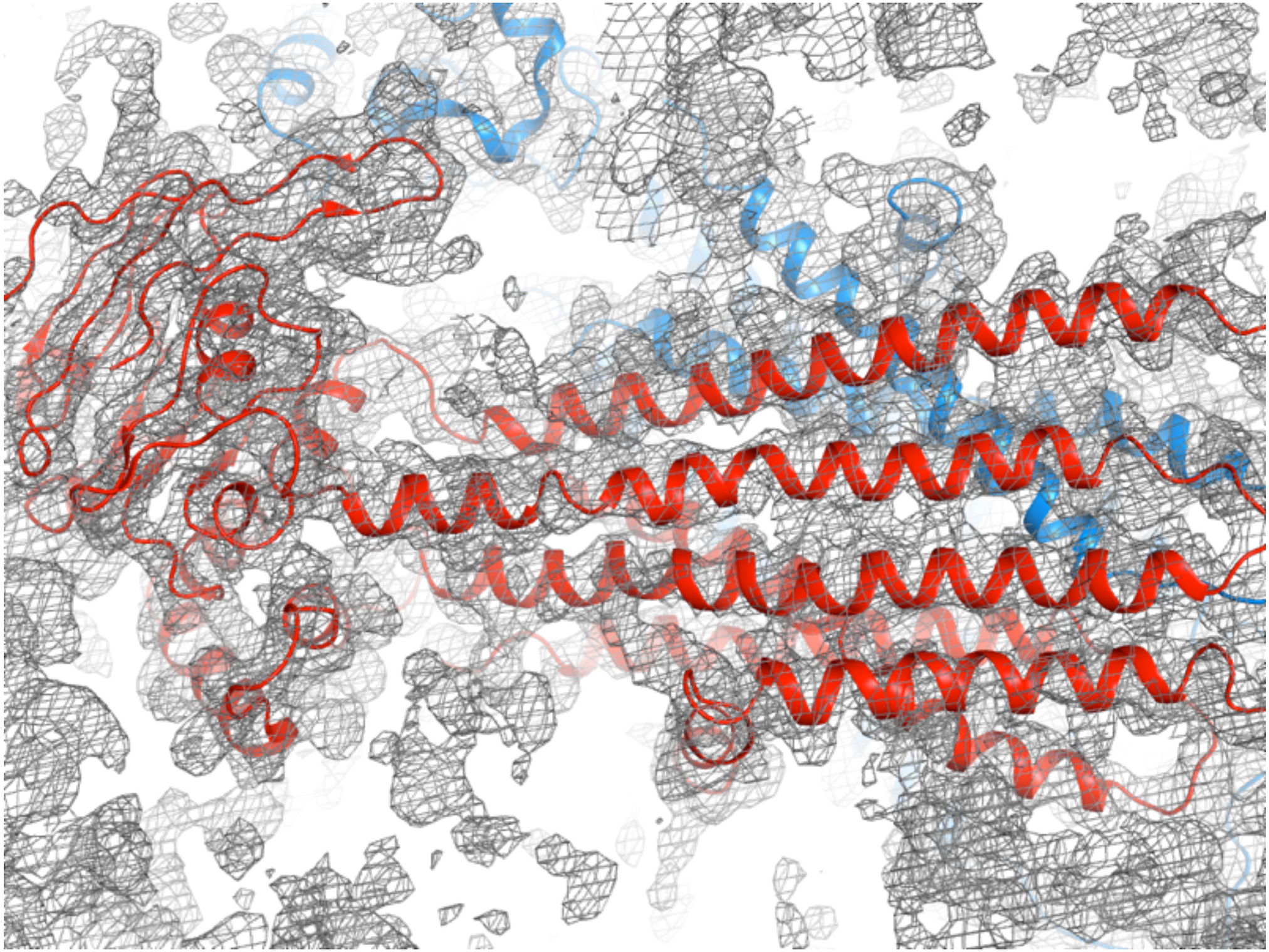


Model bias and you



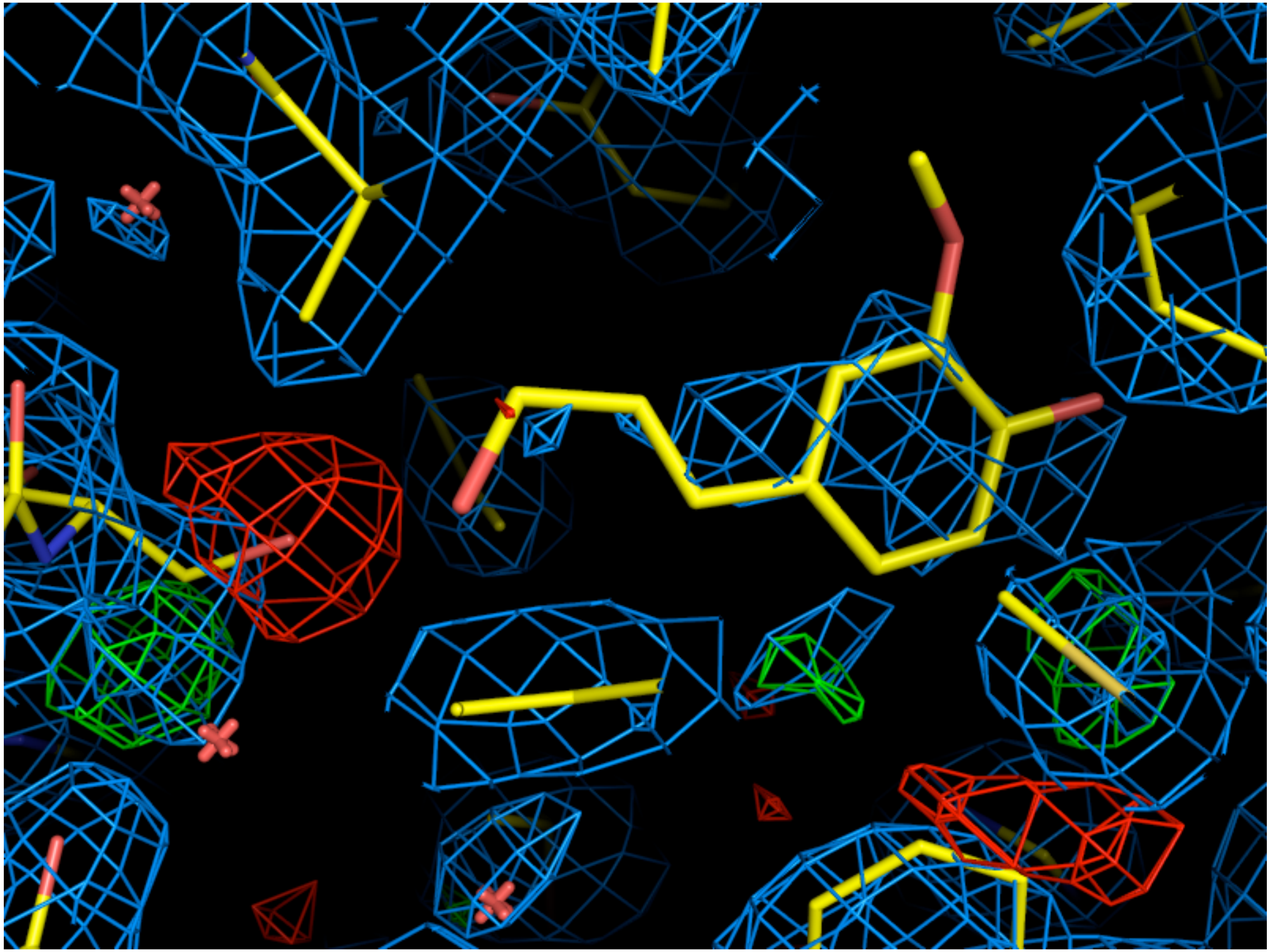
(output of twinned refinement of incorrect solution, from an anonymous Phenix user)

Model bias and you



PDB ID 1z2r - Reyes & Chang (2005) *Science* 308:1028-31 [retracted]

Confirmation bias: even worse than model bias



See also Pozharski et al. (2013) Acta Cryst D69:150-167.

Coping with model bias

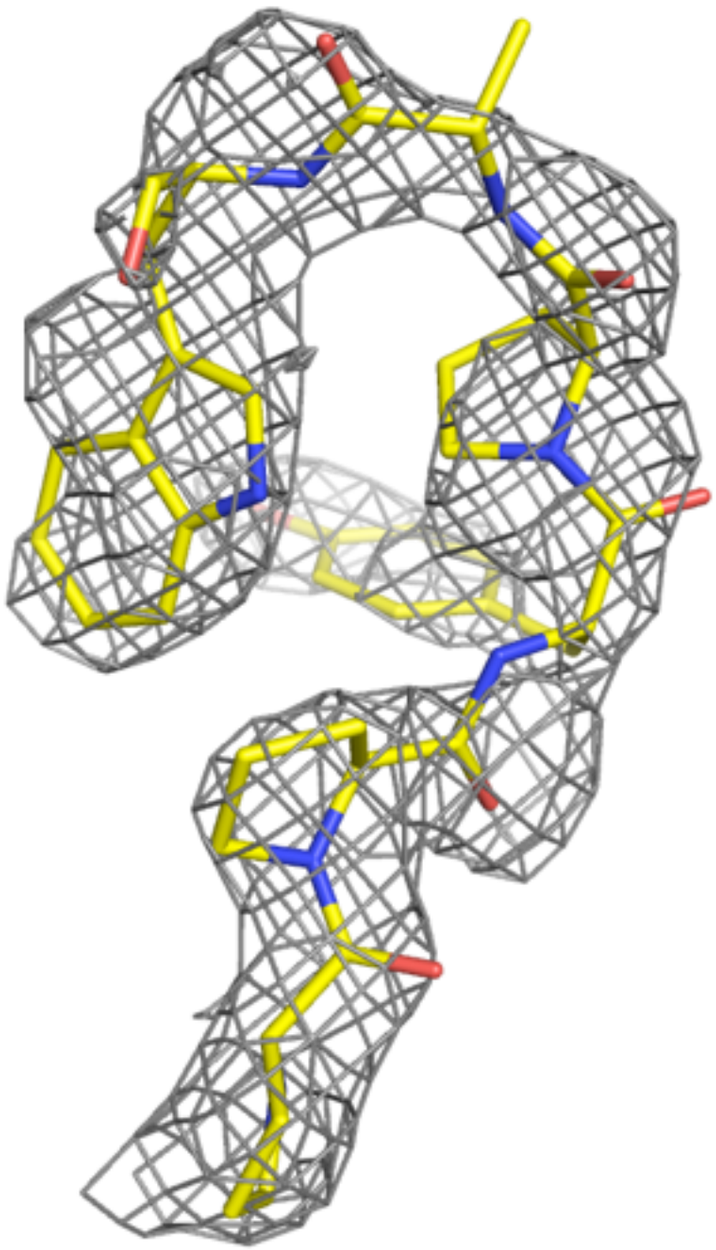
(Pavel Afonine)

- There are many methods to reduce model bias
  - likelihood-weighted σ_A-map: $2mF_{OBS}$-DF_{MODEL} (Read, 1986; Urzhumtsev et al., 1996)
    - this is what phenix.refine and REFMAC output by default
  - OMIT map (Bhat, 1988)
  - Simulated-annealing OMIT maps (Hodel et al., 1992; Brunger et al., 1998)
  - 'kicked' OMIT maps (Guncar et al., 2000)
  - Model rebuilding with randomization (Zeng et al., 1997; Reddy et al., 2003)
  - Prime-and-switch density modification (Terwilliger, 2004)
  - Carry out the usual model building and refinement avoiding a specific model part, such as ligand
  - 'ping-pong refinement' (Hunt & Deisenhofer, 2003)

- Most of the above methods may or may not remove the bias completely
- Many of these lead to reduced map quality - some may also take a long time to process

Contouring, sigma levels, and publication graphics

Many crystallographers are tempted to make figures like this to demonstrate the presence of a molecule:



Problems with this figure:

1. Calculated using model phases with peptide included
2. Contour level is both arbitrary and relatively low (0.8 sigma as shown here)
3. No context shown - what does the density for nearby atoms look like?
4. mFo-DFc difference map not shown

Validating your model with omit maps

Maps calculated without part of the model should still show clear density for the missing atoms:

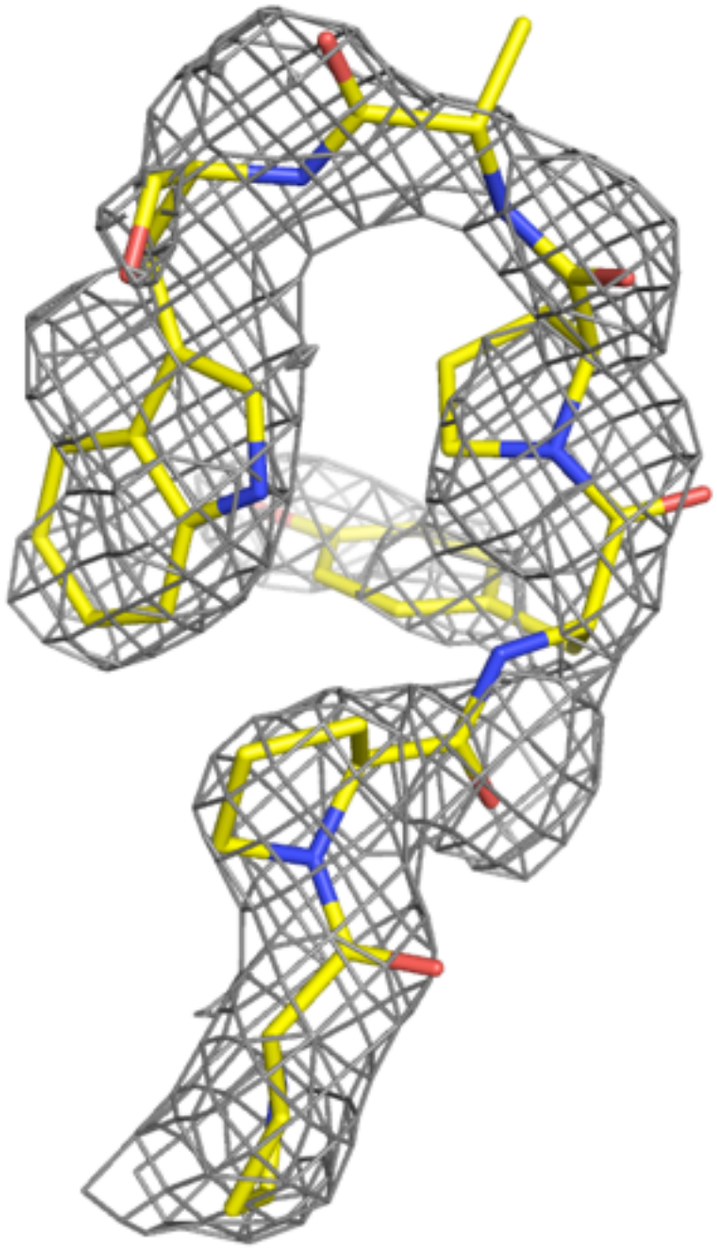
grey = 2mFo-DFc refined density @ 0.8 sigma

grey = 2mFo-DFc omit density @ 1.0 sigma
green = mFo-DFc omit density @ 3.0 sigma

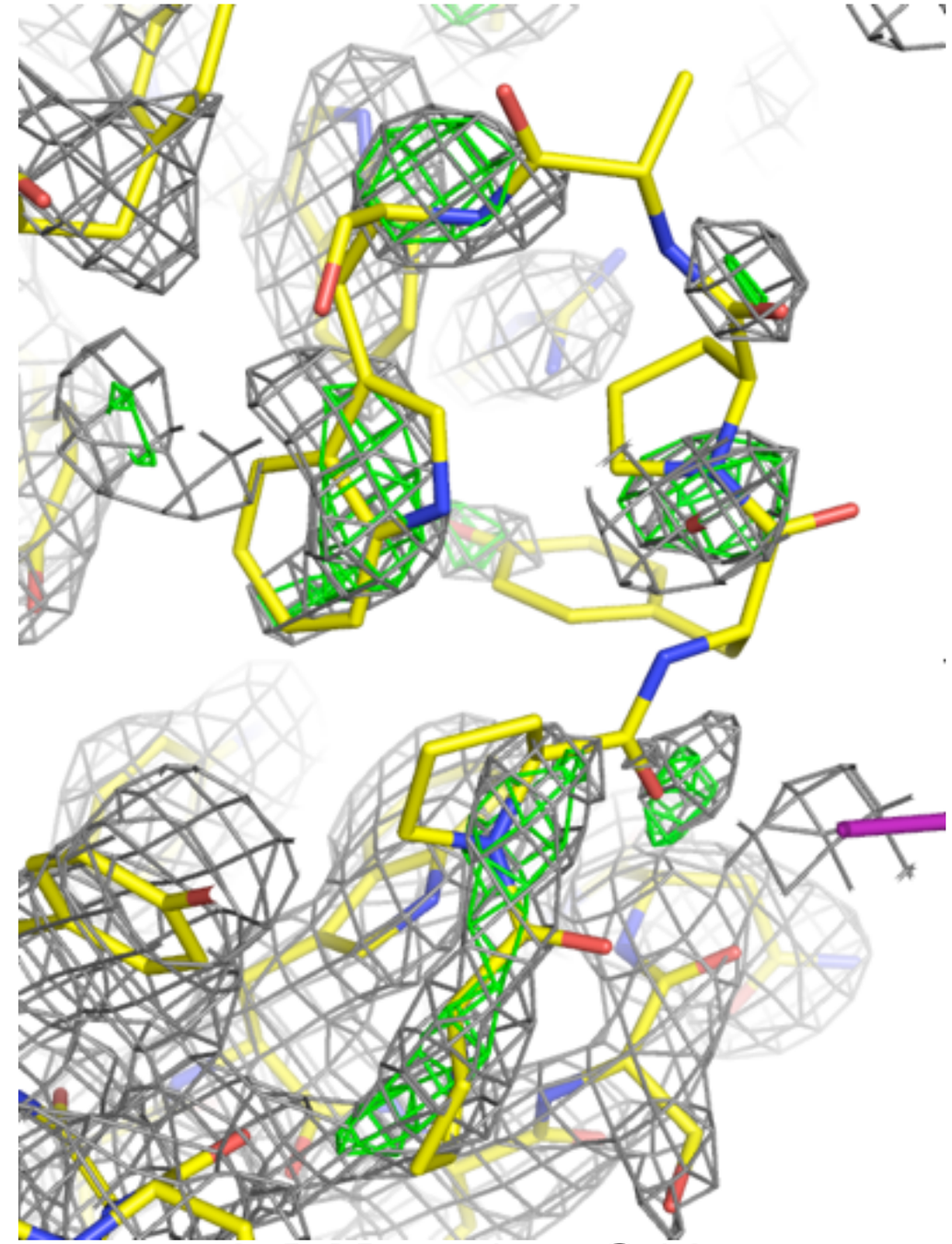
<u>To thoroughly avoid phase bias, simulated annealing or rebuilding is strongly recommended</u>

Validating your model with omit maps

The same peptide from two slides previous:



grey = 2mFo-DFc refined density @ 0.8 sigma



grey = 2mFo-DFc omit density @ 1.0 sigma
green = mFo-DFc omit density @ 3.0 sigma

The "peptide density" is obviously water molecules or buffer components!

Demonstrating ligand binding with electron density

If you want to show that a ligand is present in your crystal, follow these steps:

1. Solve and refine as far as possible **without the ligand**; save the final maps
2. Add your ligand, continue refinement
3. Use the maps from (1) with the model from (2) in your figures

This avoids the problem of model bias entirely, and is also easier!

If you already placed the ligand and don't want to re-do step (1), a **simulated annealing omit map** is the most rigorous (and reviewer-approved) method to remove bias