Validation Philosophy

- Visualizations > statistics
- Local conformations > structure-level averages

- “Outlier” thresholds are set statistically
  - Expect to see experimentally justified statistical outliers sometimes, especially at functional sites
  - Cherish these! You found something cool!

Outline

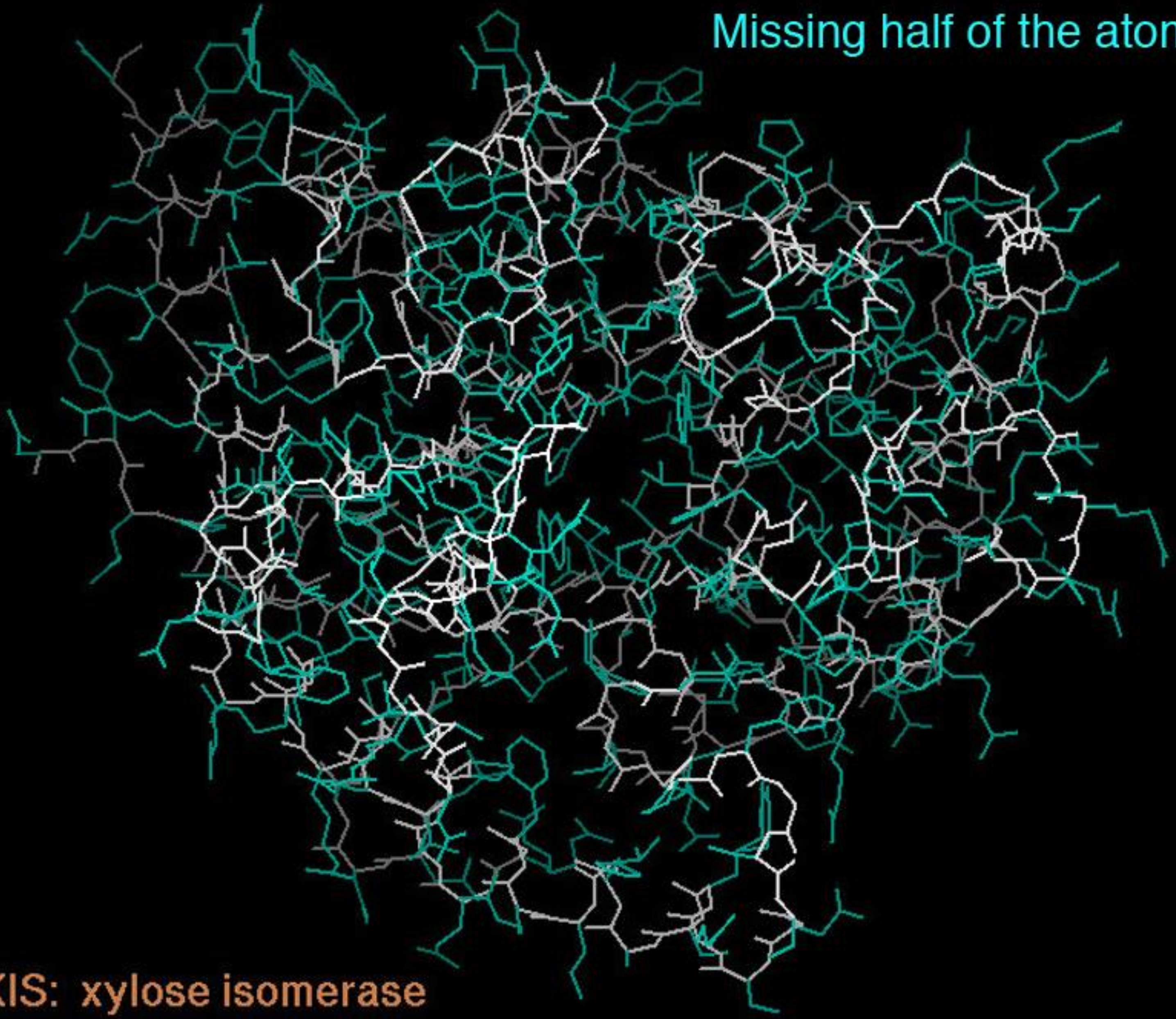
For each validation

- Method
  - Briefly, how the underlying idea or math works
- Visualization
  - How outliers are visually represented
- Probable causes
  - Example of a common or interesting type of error
  - Not comprehensive!

All-Atom Clashes and Contacts

Add hydrogens with phenix.reduce

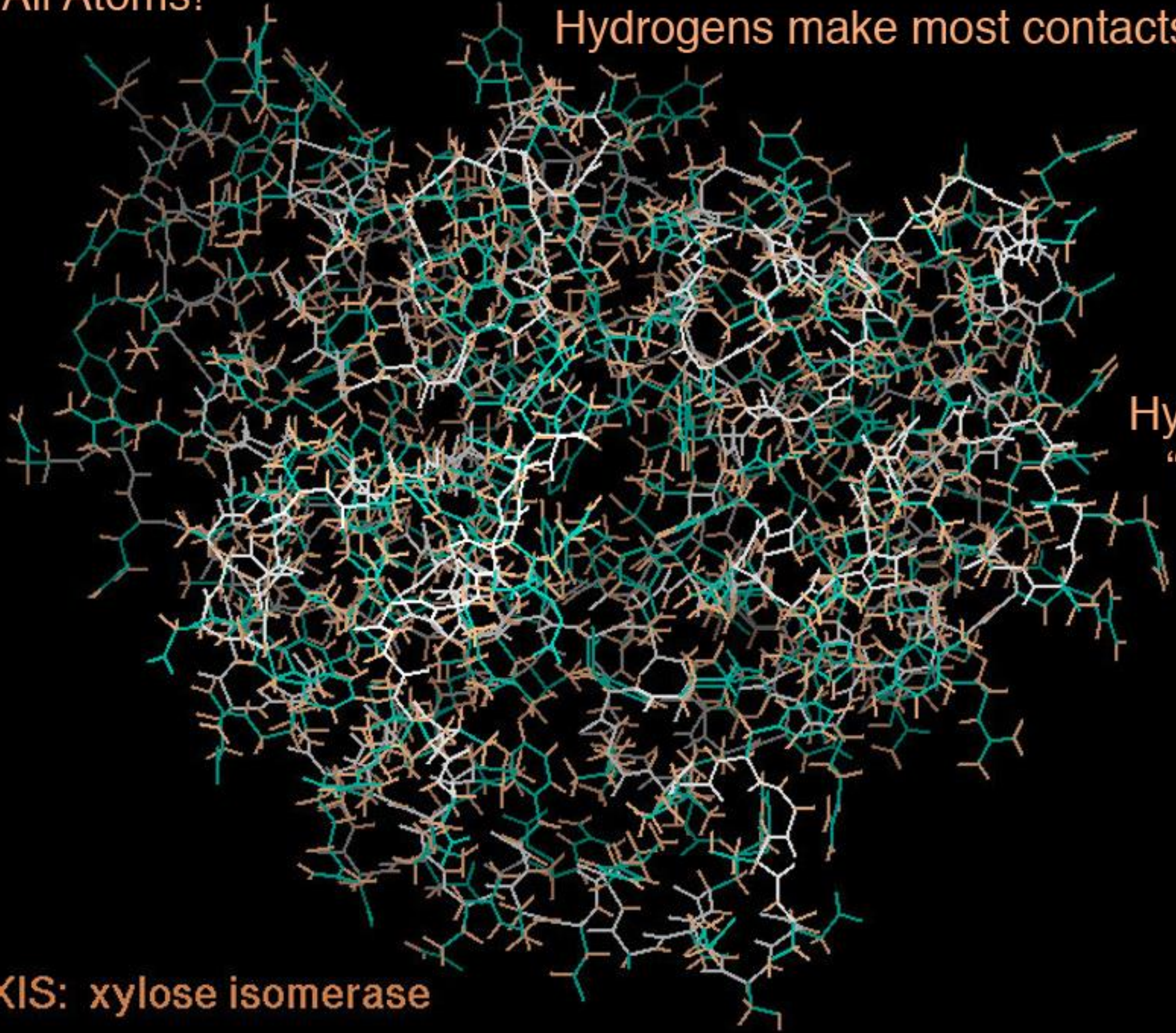
Missing half of the atoms!



4XIS: xylose isomerase

All Atoms!

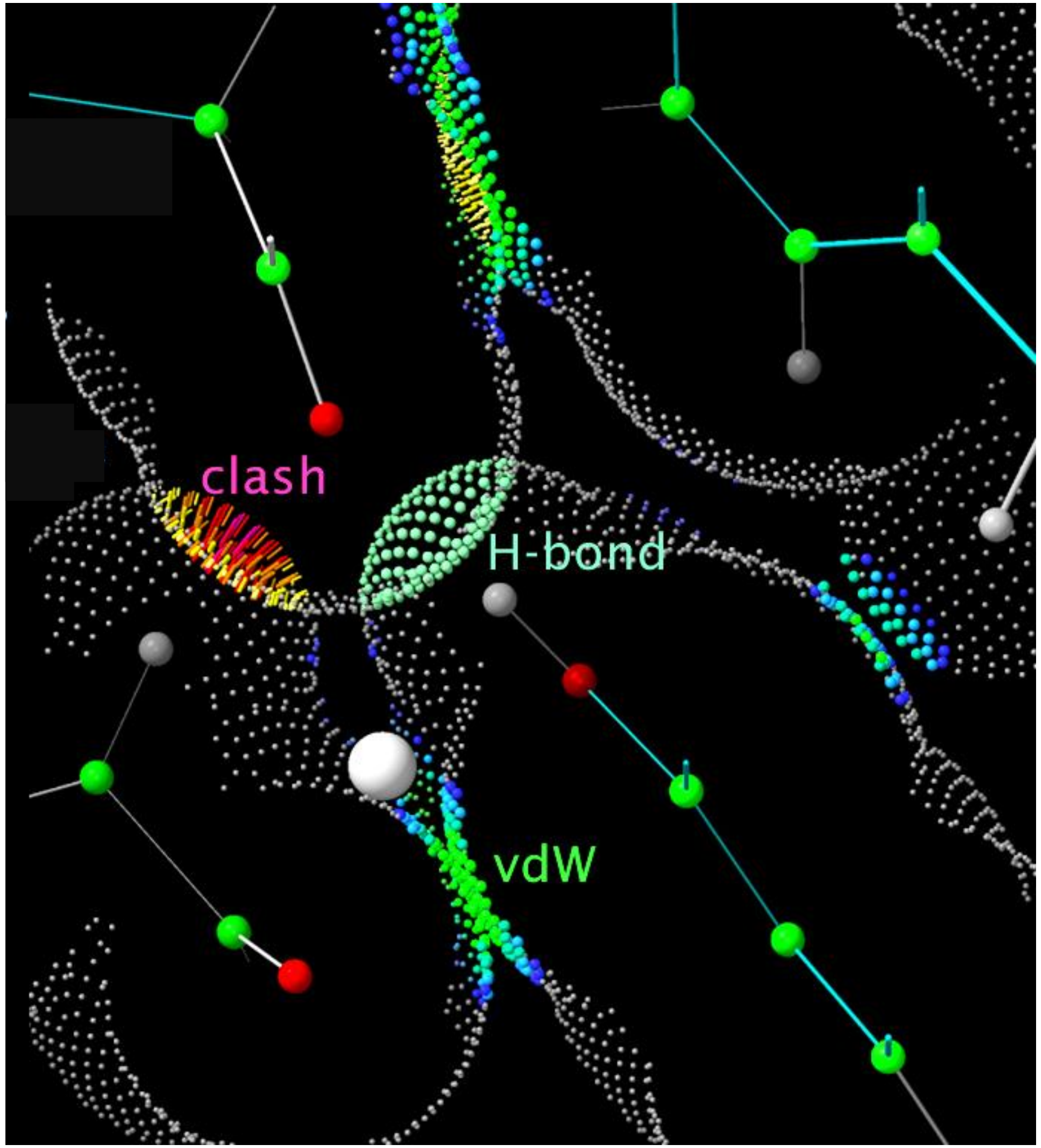
Hydrogens make most contacts



Hydrogens:
"twigs
on the
tree"

4XIS: xylose isomerase

All-Atom Contacts and Clashes: Method



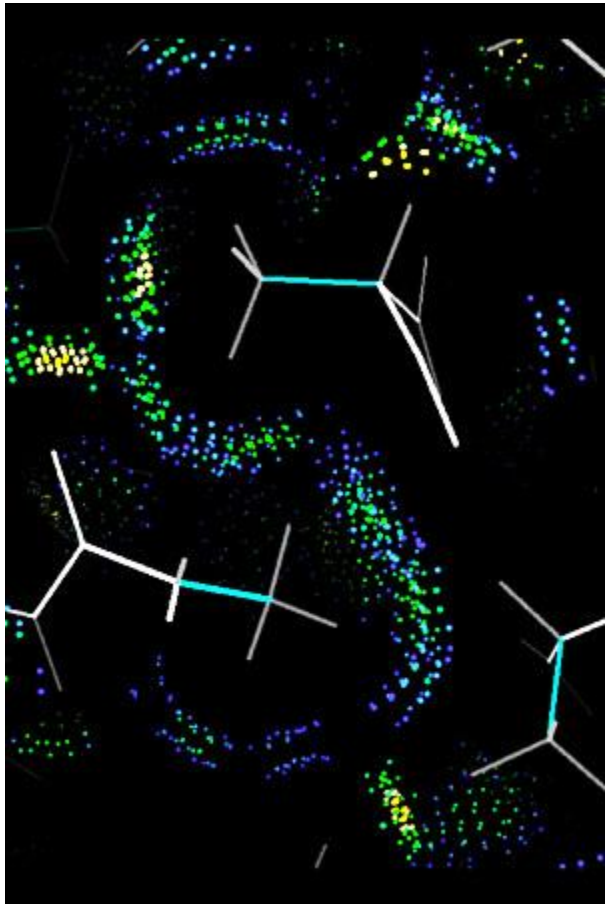
- Roll a 0.25Å radius "Probe" sphere over the van der Waals surface of each atom

- Mark where the probe touches or overlaps with another van der Waals surface

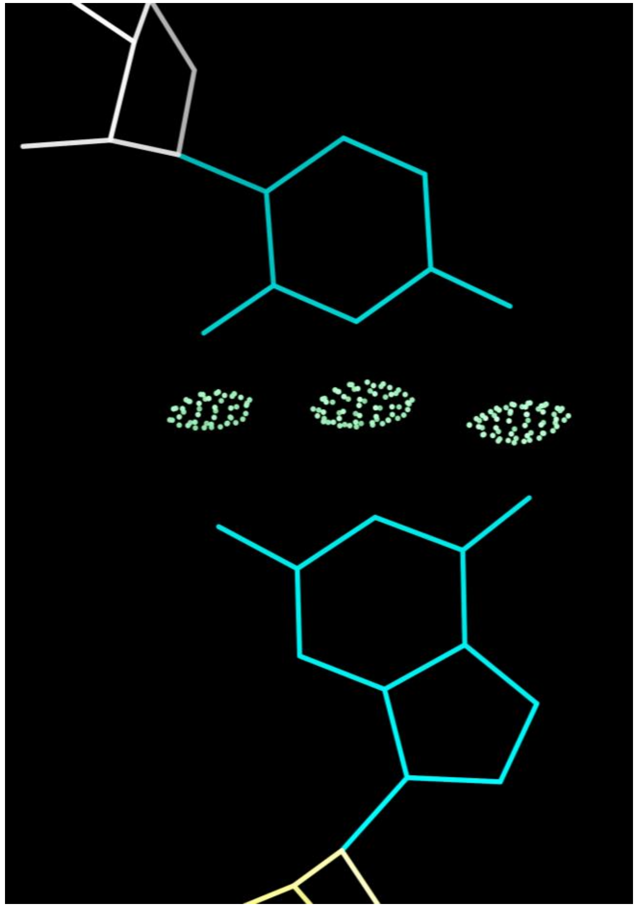
- Note that hydrogen atom surfaces can shield heavy atom surfaces

molprobity.clashscore

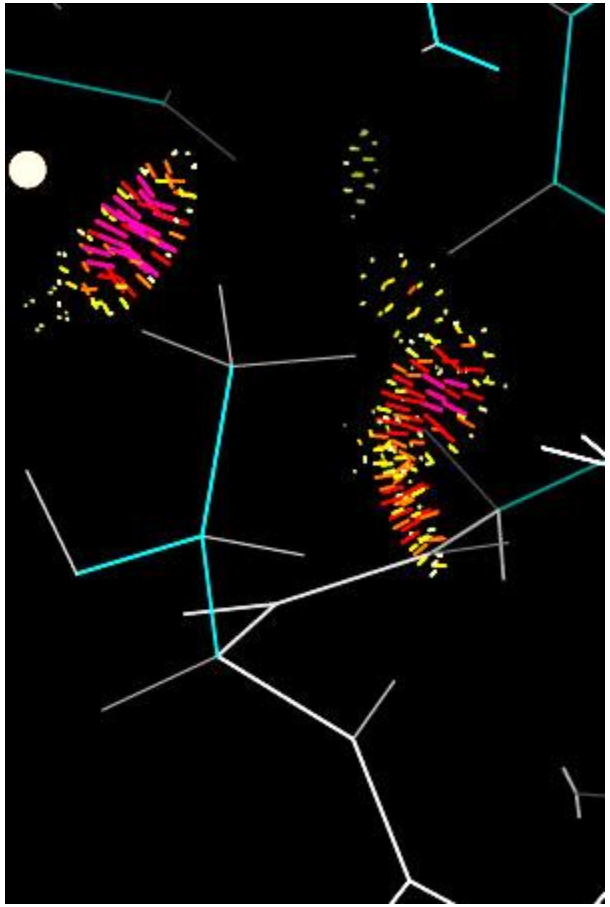
All-Atom Contacts and Clashes: Visualization



Favorable vdW packing in greens and blues

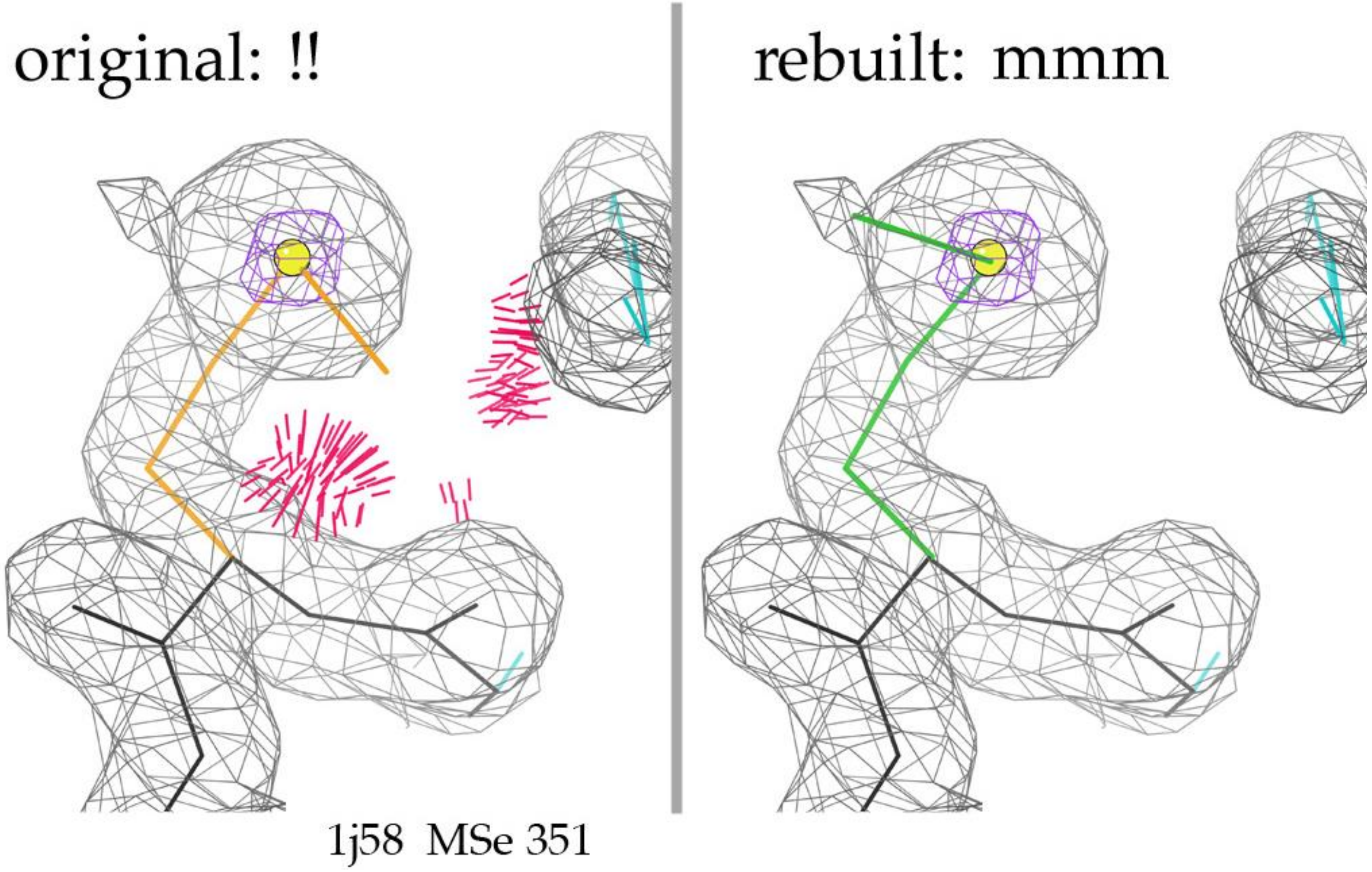


Favorable hydrogen bonding as light green pillows



Steric overlaps, aka “clashes”, as hot pink spikes

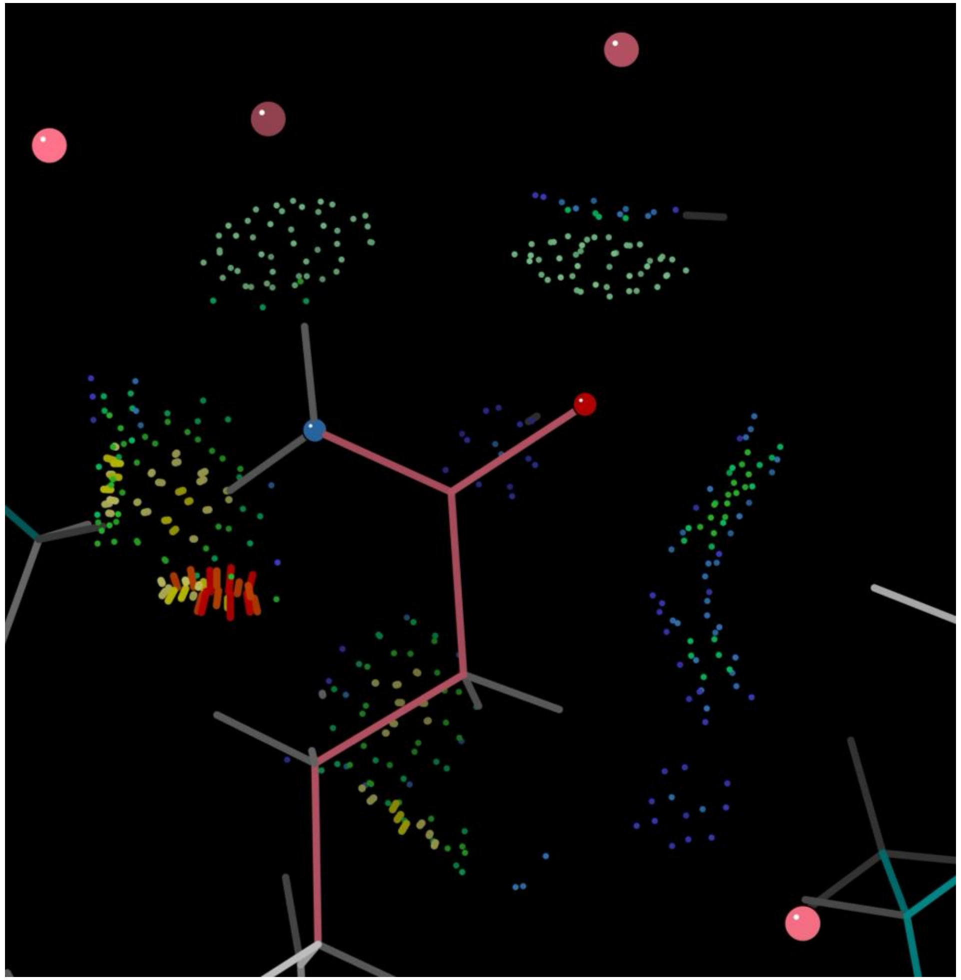
All-Atom Contacts and Clashes: Probable causes

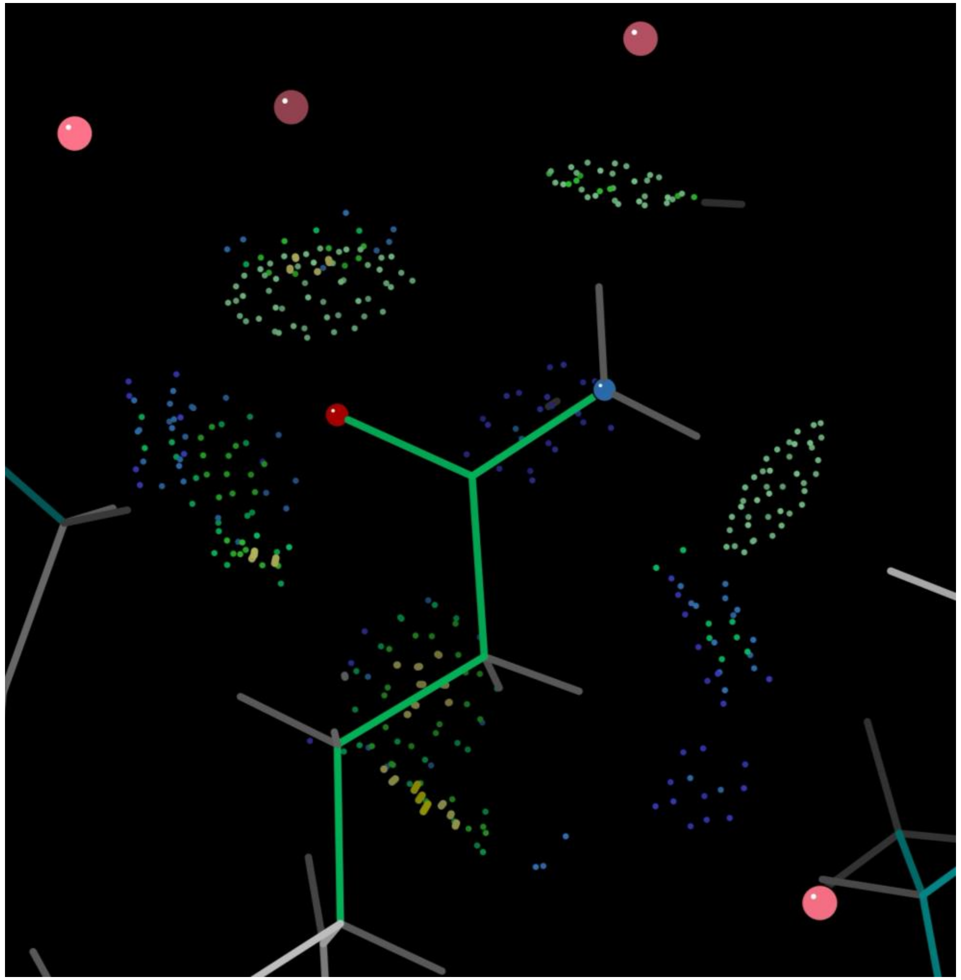


<u>Other outliers</u>

- Clashes usually occur alongside other outliers
- Emphasize modeling errors
  - *Real* rare features are less likely to have clashes
- Can imply direction for fixups

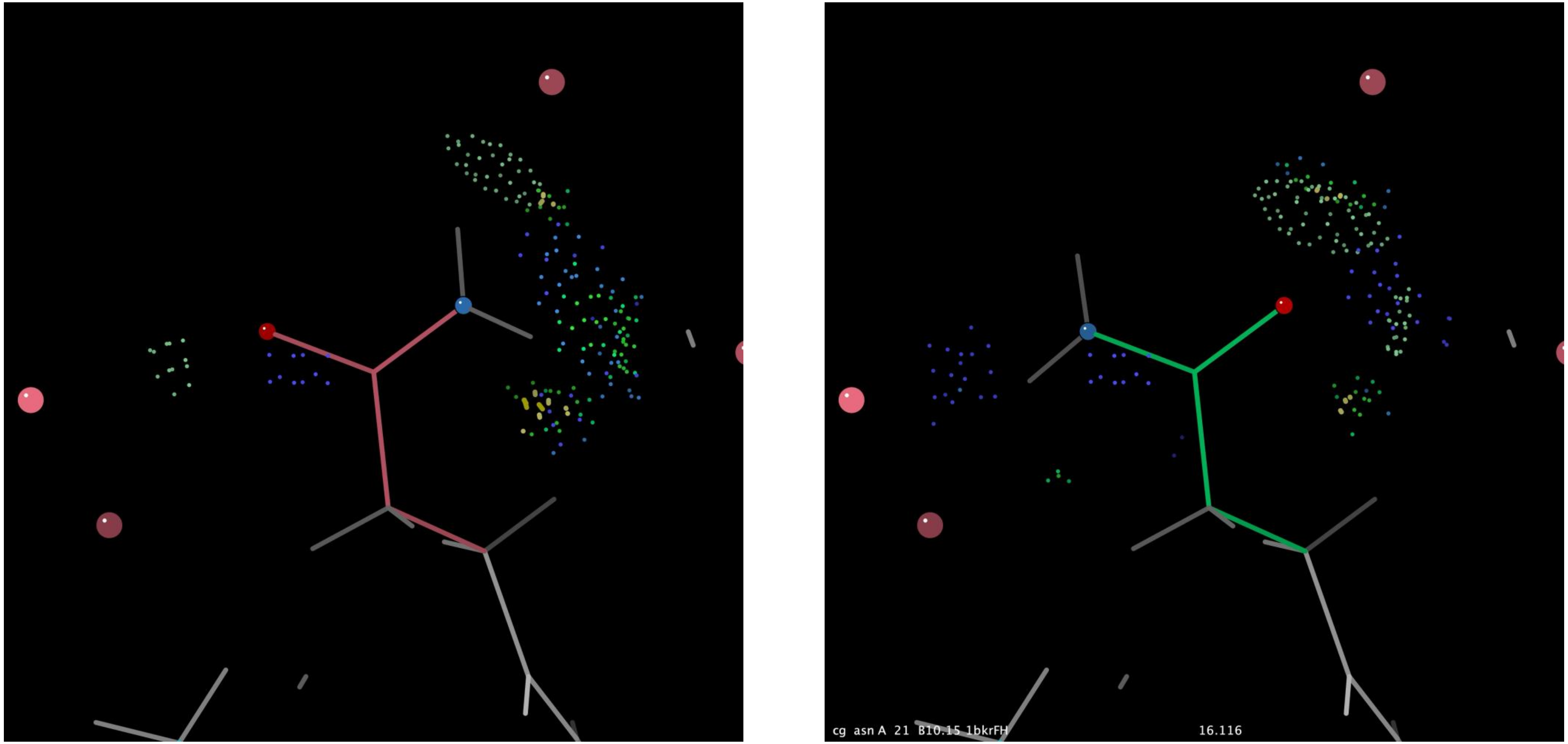
All-Atom Contacts and Clashes: Asn/Gln/His Flip corrections





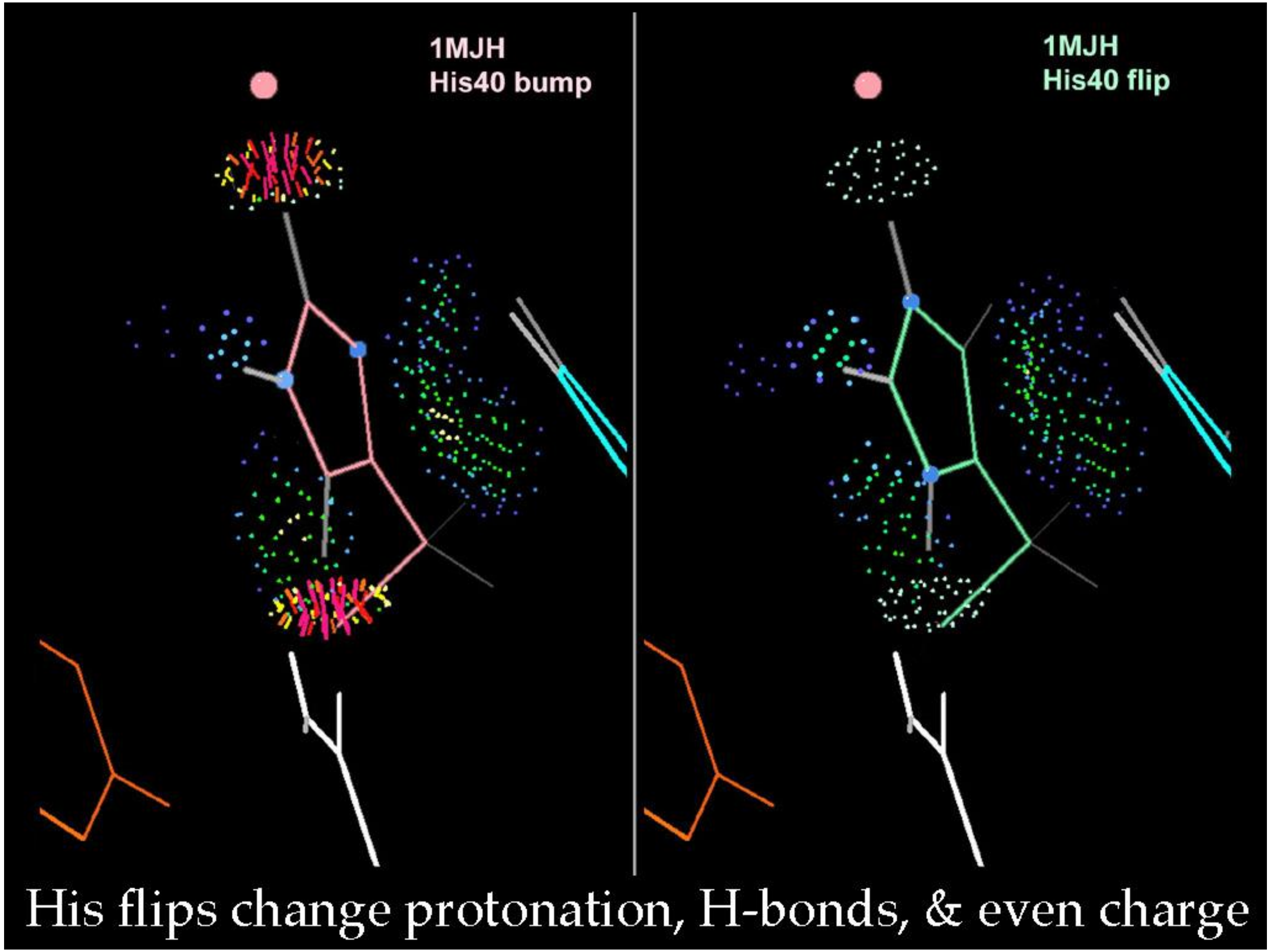
Which Gln is correct?

All-Atom Contacts and Clashes: Asn/Gln/His Flip corrections



Which Asn is correct?

All-Atom Contacts and Clashes: Probable causes

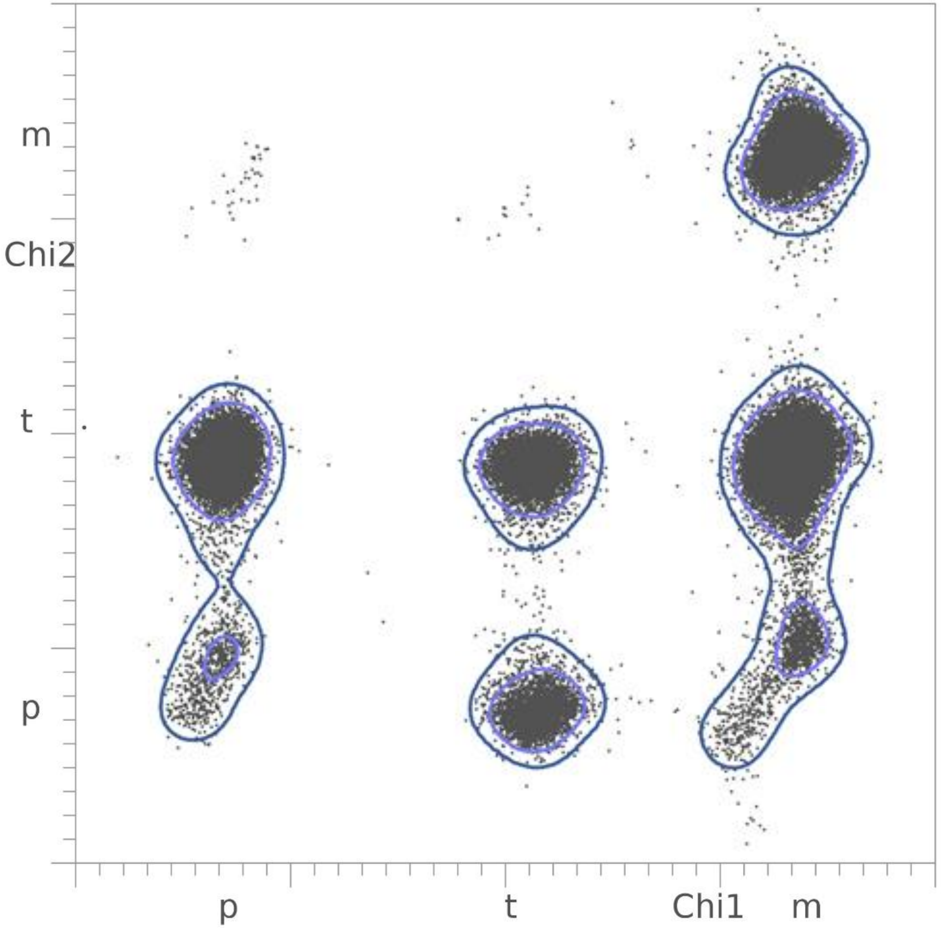


Sidechain flips

- Asparagine, Glutamine, and Histidine (N/Q/H) are pseudo-symmetric
- Wrong orientation can produce clashes without other error markup
- Fix with Reduce or Coot tools, then re-refine.

Sidechain Rotamers

Sidechain Rotamers: Method

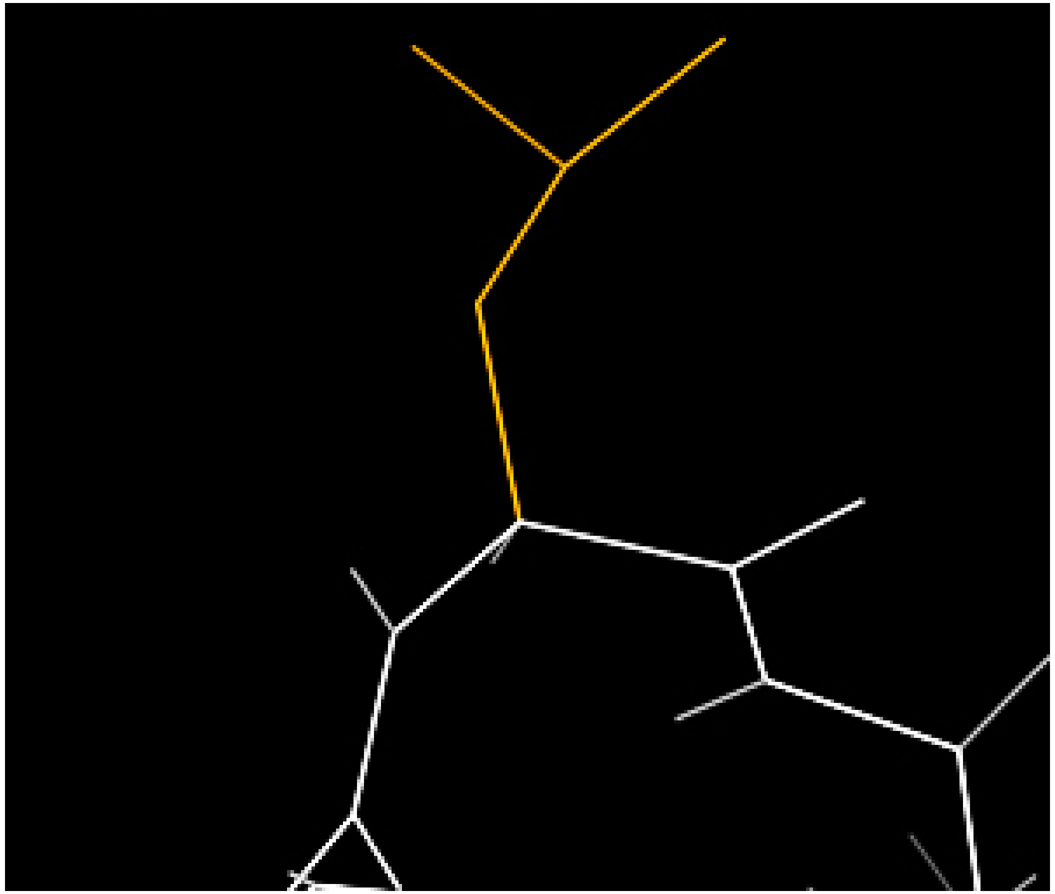


Rotamer distribution for Isoleucine in χ1/ χ2 space

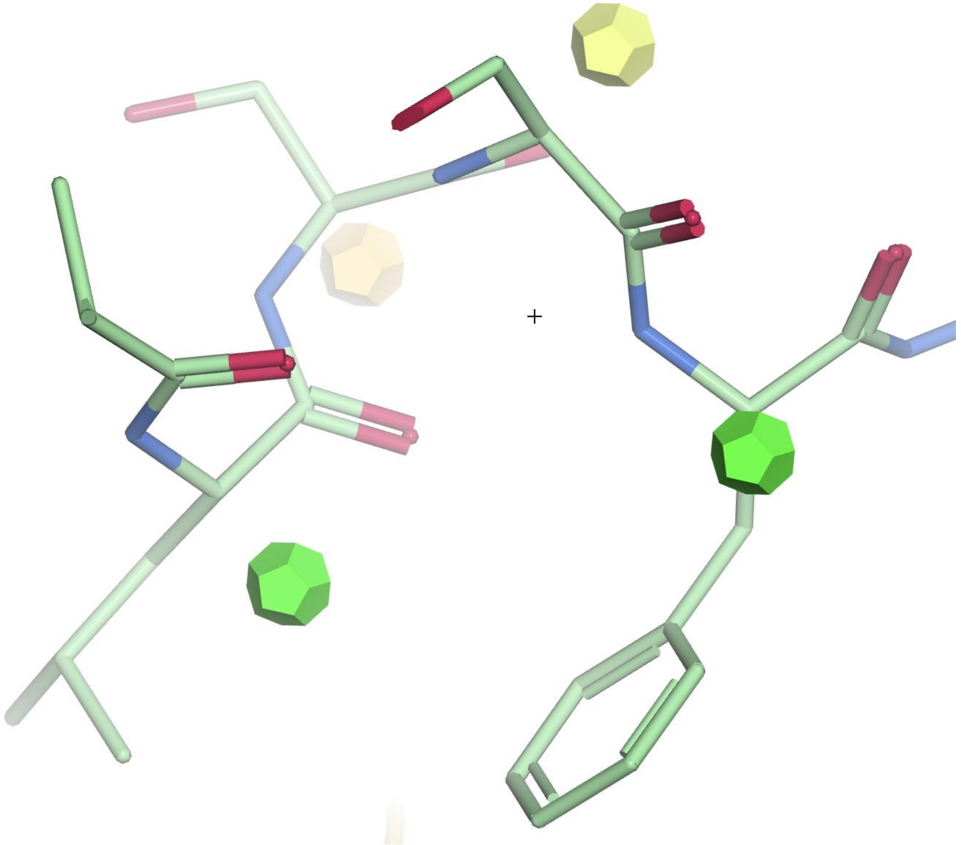
- Sidechain conformations are described by a series of χ (Chi) torsions
- Rotamers are statistically expected combinations of χ values
- For tetrahedral atom centers, this means staggered
  - p +60°
  - t 180°
  - m -60°
- For planar atom centers, rotamers are much more continuous
  - Rotamers are named with a central value
  - e.g m90 or p-80 for Histidine
- Updated in 2016:
  - Favored (98% of data) Allowed (99.7% of data)

molprobity.rotalyze

Sidechain Rotamers: Visualization

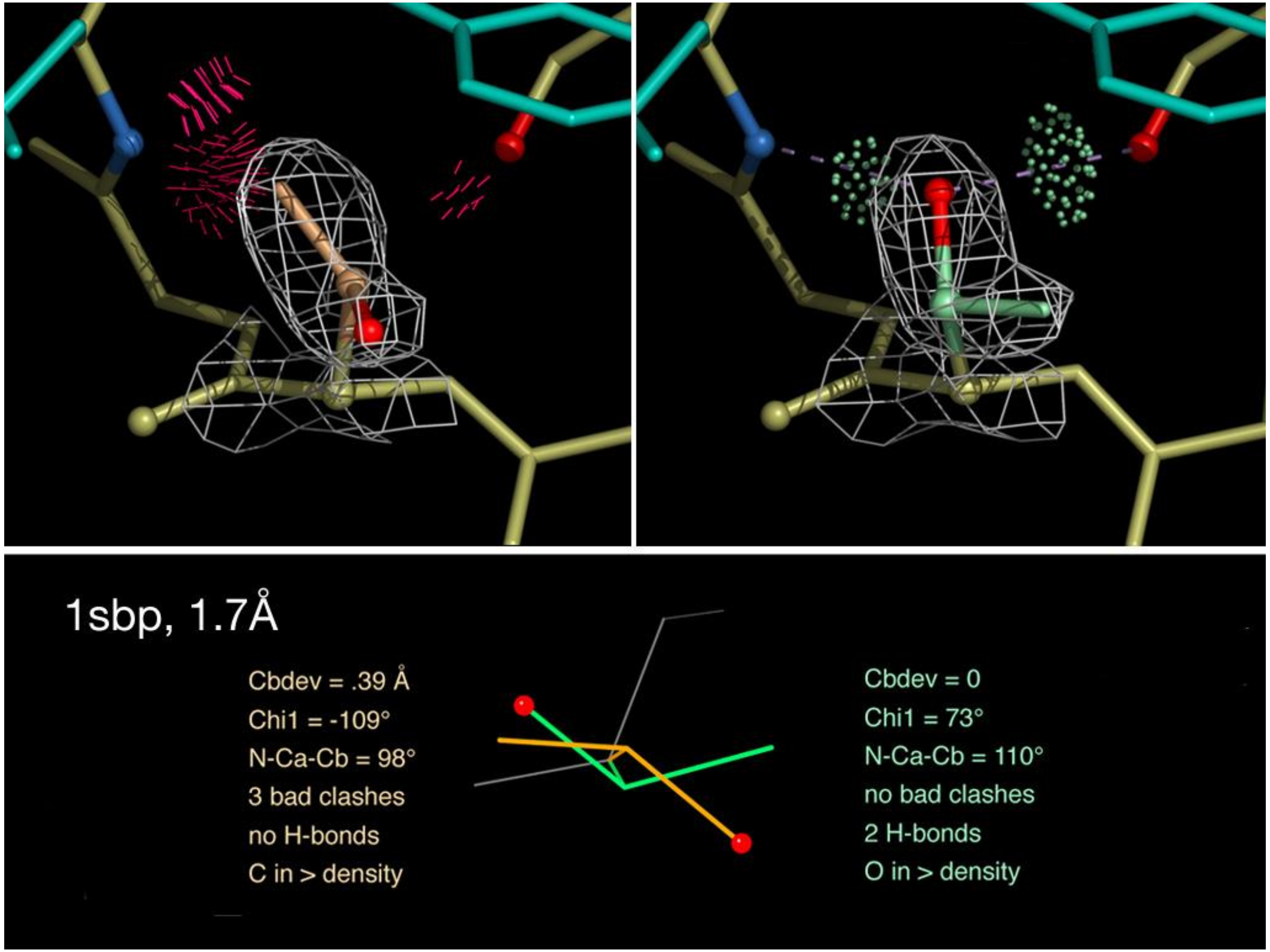


In KiNG, Rotamer outliers are traced in gold over the modeled sidechain



In Coot/Moorhen, Rotamers are marked with a colored dodecahedron

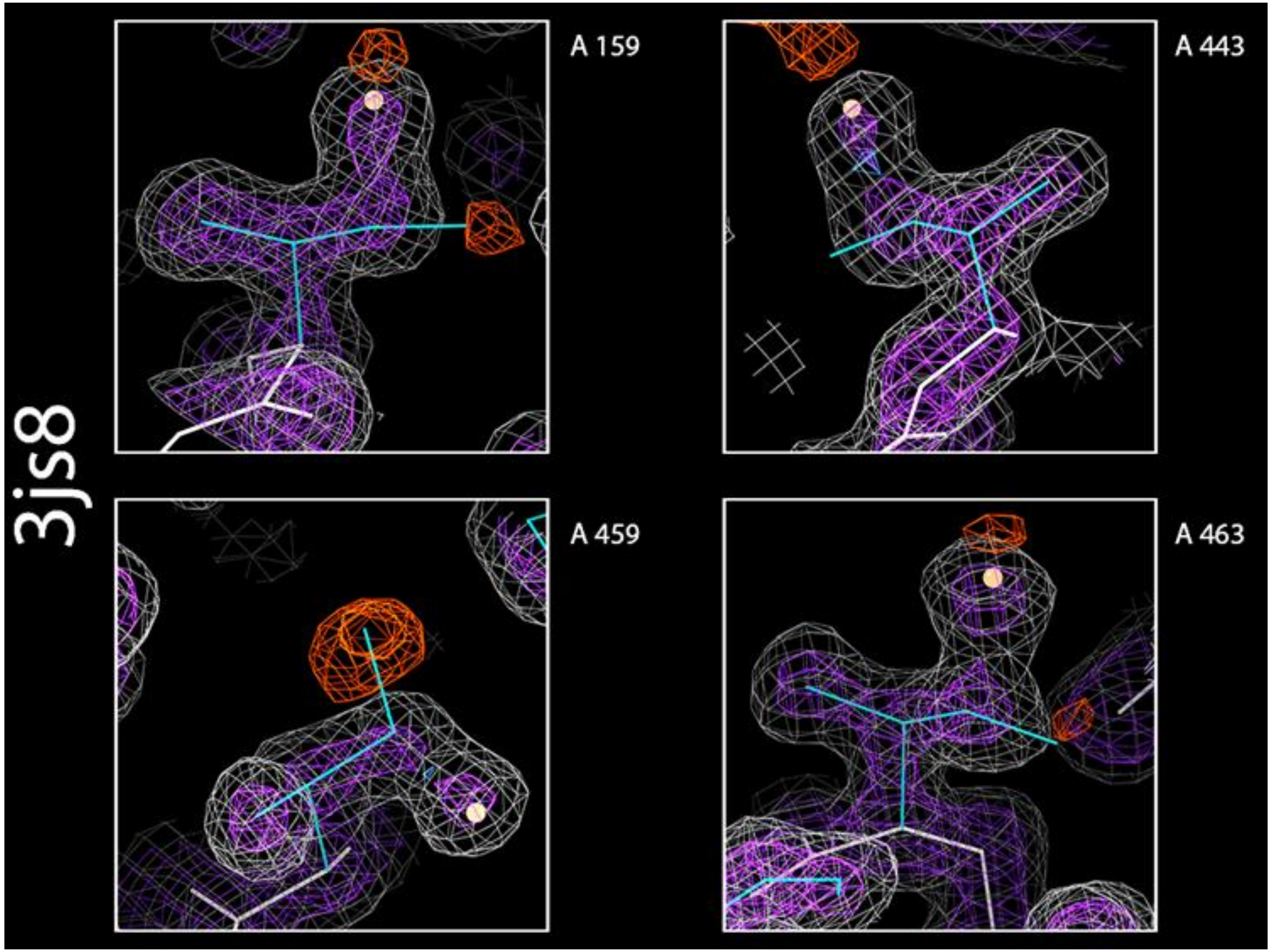
Sidechain Rotamers: Probable causes



Backwards Valine, Leucine, Threonine

- May find terminal atoms fit into density at the expense of the branch atom
- Simple to fix with a flip (then re-refinement)

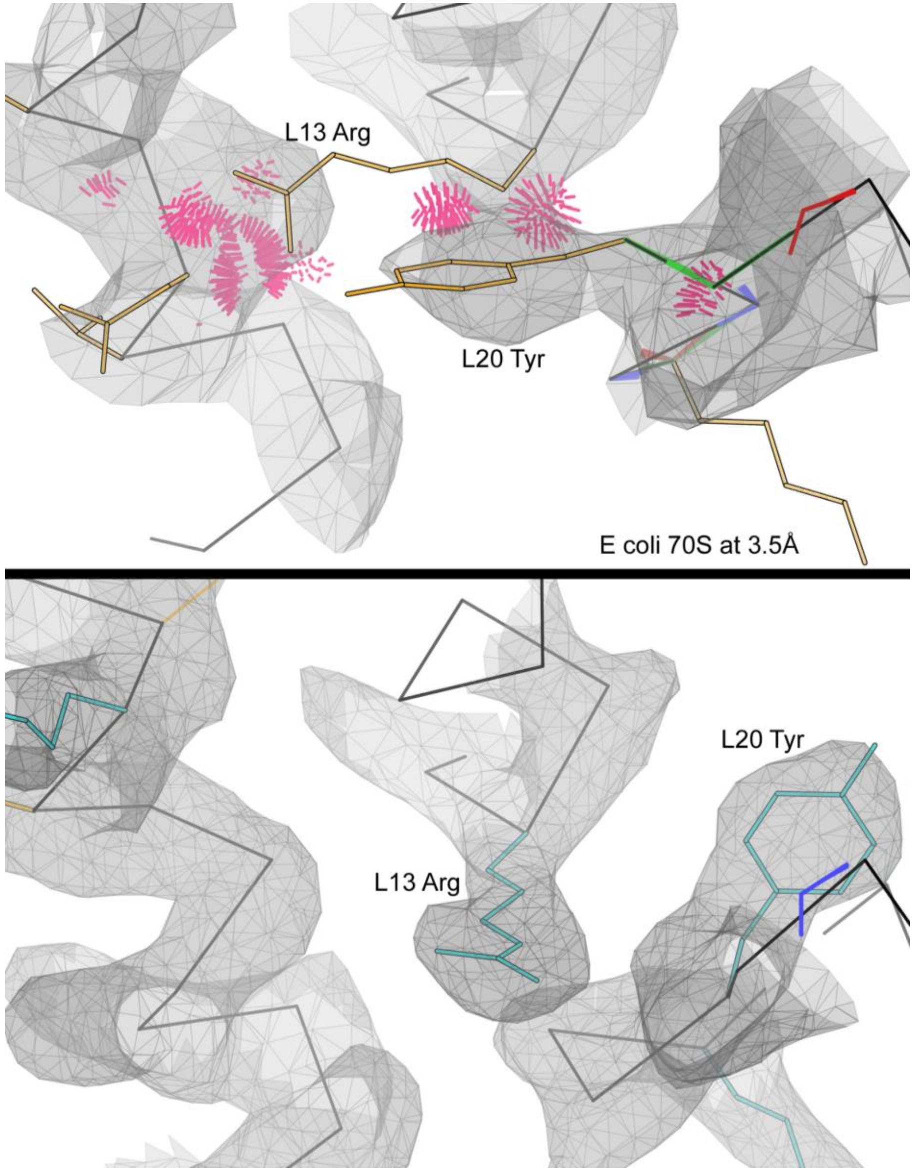
Sidechain Rotamers: Probable causes



Water problems

- Modeled water may co-opt sidechain density and create a rotamer outlier
- Isoleucine CD1 is especially vulnerable
- Delete water, rebuild sidechain

Sidechain Rotamers: Probable causes



Sidechains in wrong density

- Sidechains can get stuck in the density for other features
    - Other sidechains
    - Ligands
    - Backbone in ~3Å maps

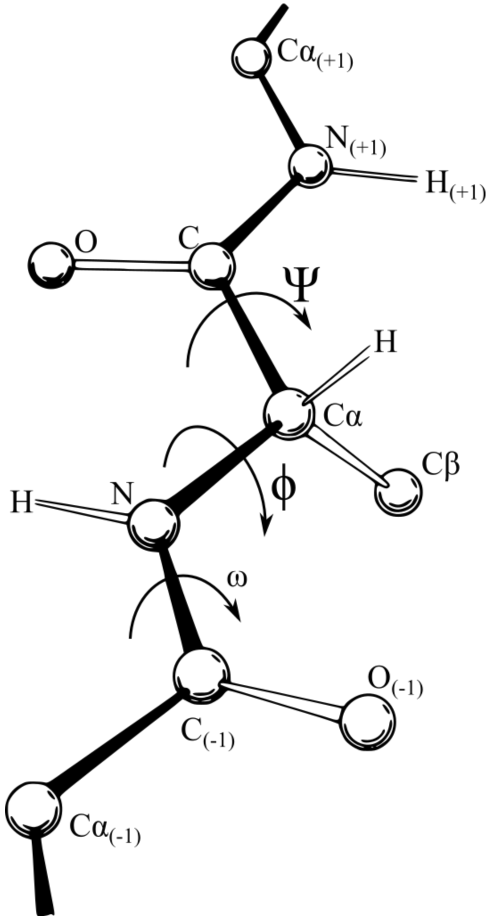
- Have to fix the whole network of misplacements

Protein Backbone Validation

Ramachandran
CaBLAM
Rama-Z

Ramachandran

Ramachandran: Method



- Phi and Psi torsions describe local protein backbone conformation

- Phi **ϕ** = C_{i-1}-N-CA-C
- Psi **ψ** = N-CA-C-N_{i+1}

- Each residue's ϕ/ψ pair is converted into cartesian coordinates and checked against contours of expected behavior

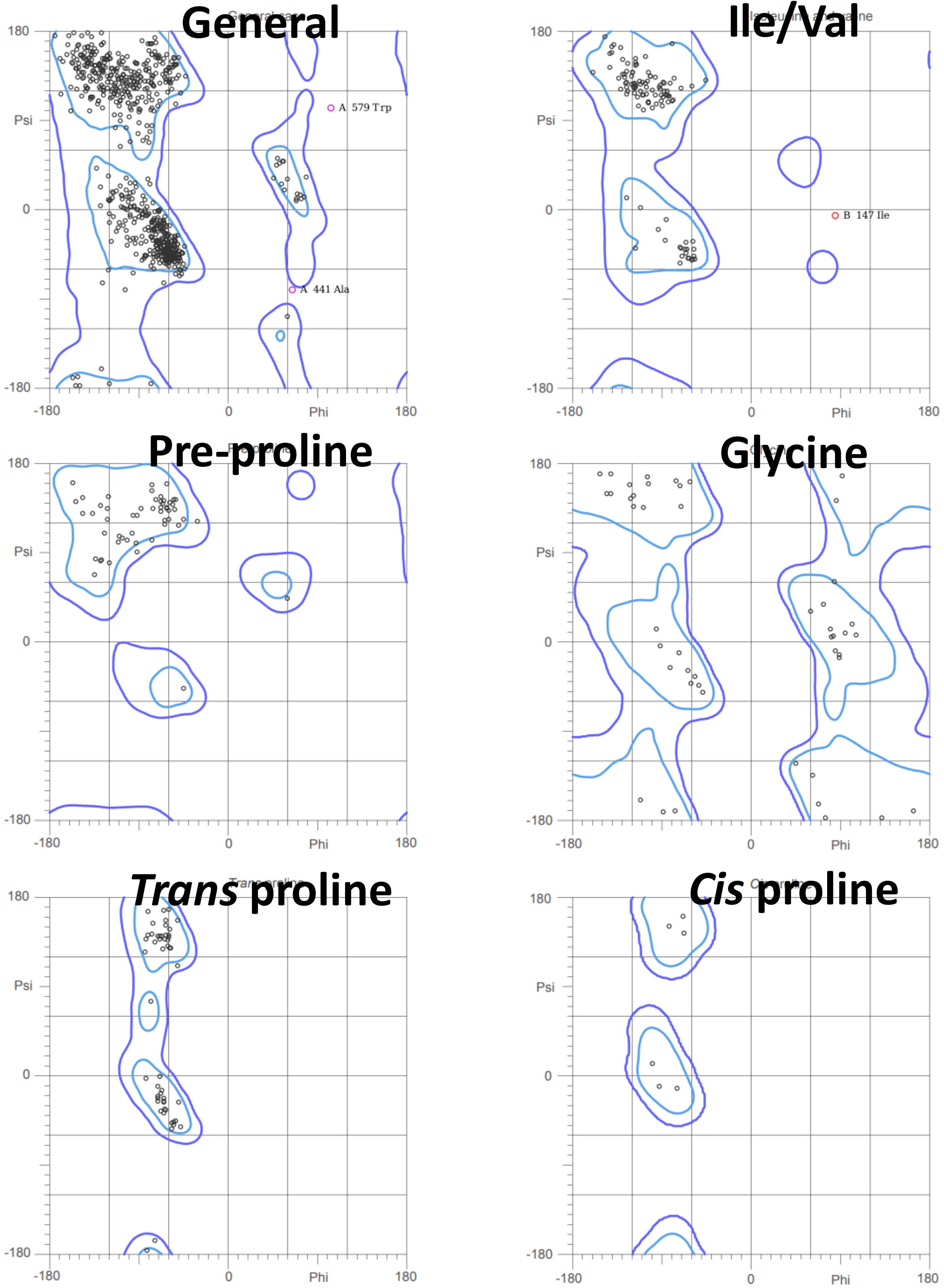
molprobity.ramalyze

Ramachandran: Visualization

Ramachandran plots shows location of each residue relative to contours of expected behavior

Different residue categories have very different expectations!

Glycine is permissive and symmetrical
Proline is restrictive
Branched C-Beta sidechain (Ile,Val) affect distribution



General

Ile/Val

Pre-proline

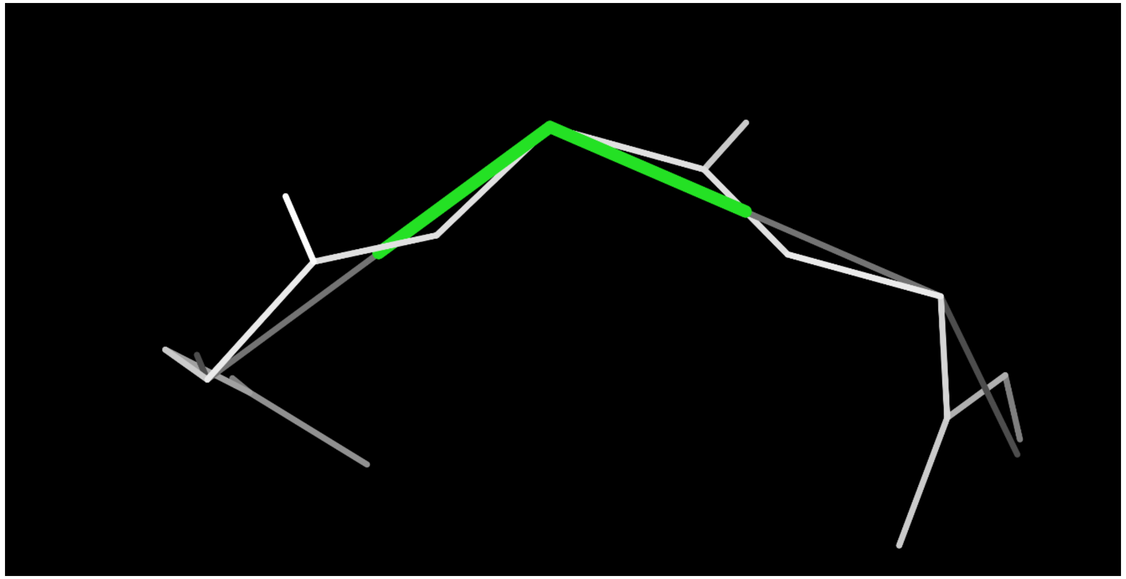
Glycine

Trans proline

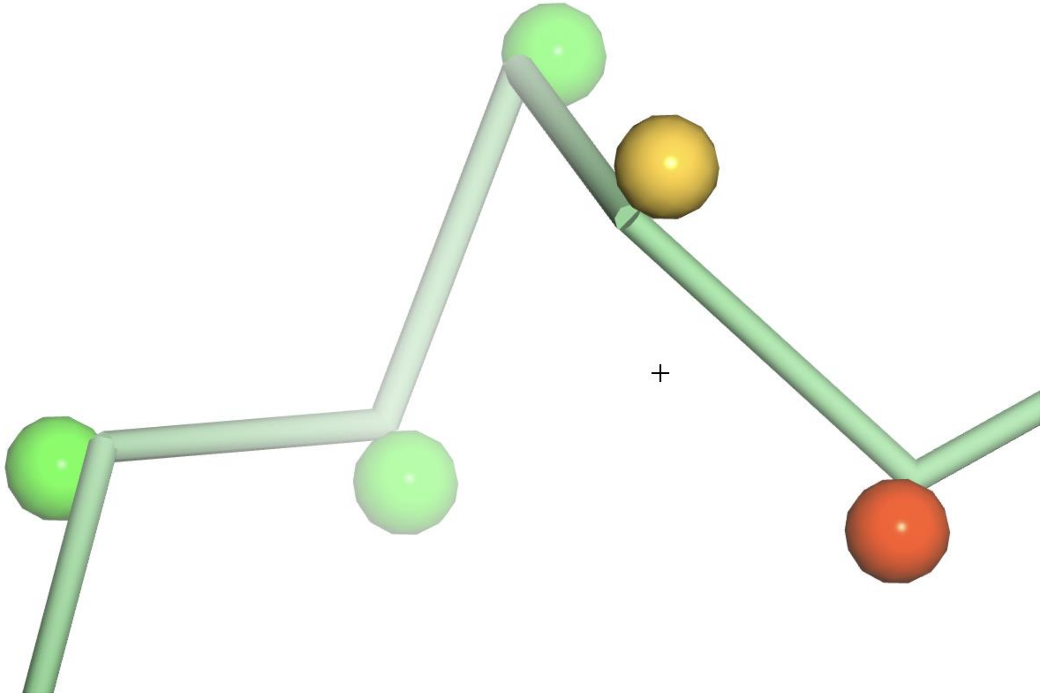
Cis proline

Favored (98% of data)   Allowed (99.5% of data)

Ramachandran: Visualization



KiNG markup highlights an outlier residue's CA in green, and extends to the peptide bonds on either side, along the CA-CA-trace

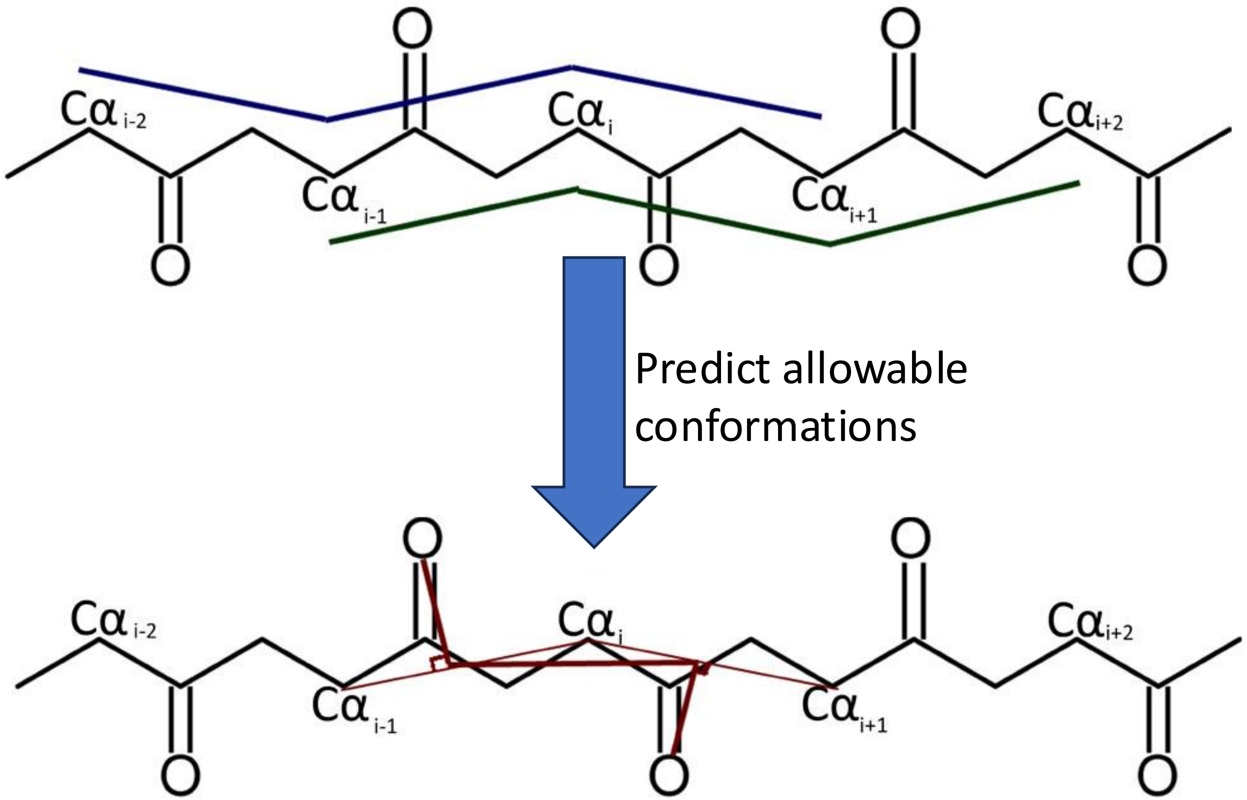


Coot/Moorhen markup places a ball at each CA, color-coded by Ramachandran favorability.

CaBLAM

CaBLAM: Method

CA-pseudodihedrals capture model “intent”



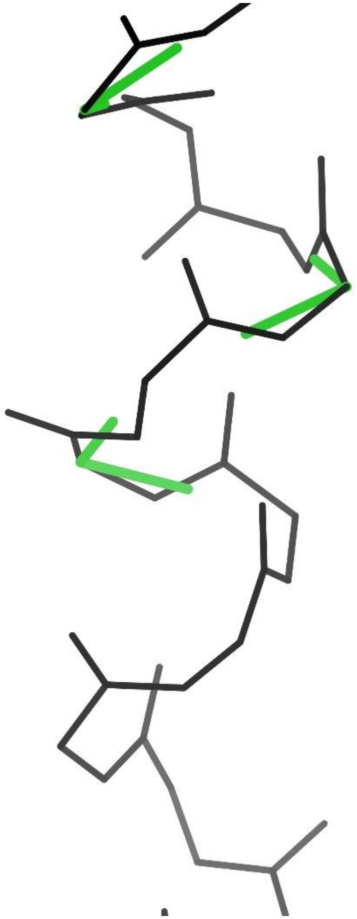
Predict allowable conformations

Peptide-peptide-pseudodihedral captures common model errors

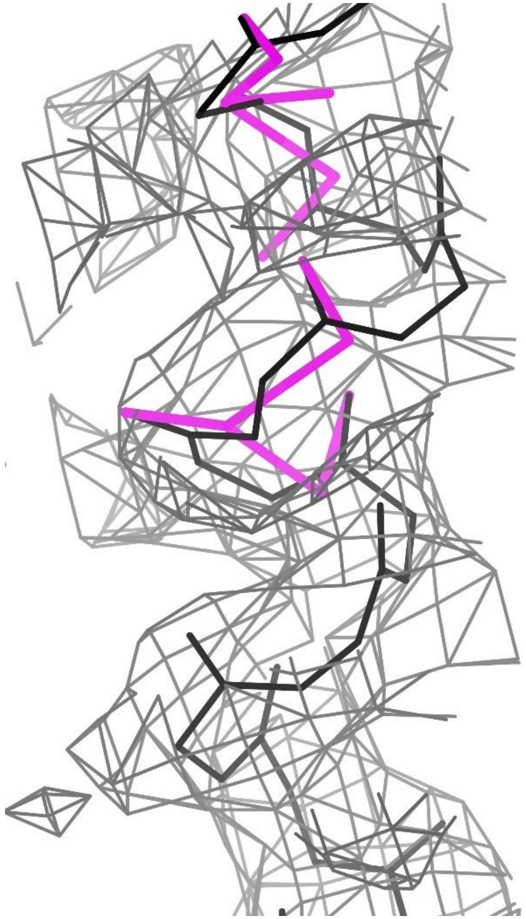
- At low resolution, the backbone CA trace is modeled better than the backbone details
- Common model errors involve wrong peptide plane orientation
- CaBLAM uses modeled CA trace geometry to predict likely peptide plane orientation, and marks the discrepancies

molprobity.cablam

Rama/CaBLAM: Probable causes



Rama markup



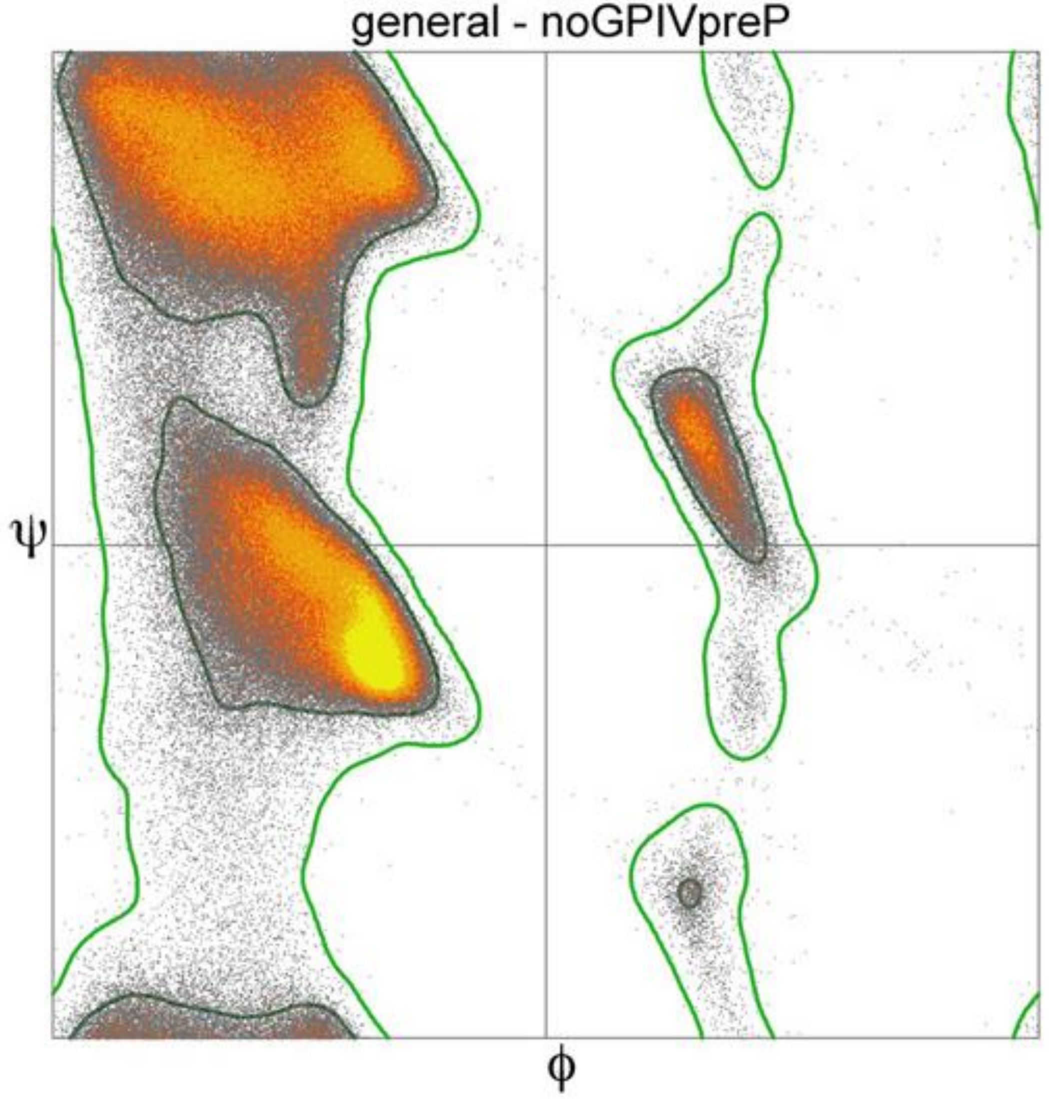
CaBLAM markup
(magenta bars)

Misplaced carbonyl oxygens

- At resolutions worse than ~2.5Å, carbonyl oxygen density disappears
  - O may be fit in arbitrary orientation
- Low-resolution density envelope allows multiple models
  - Not everything that fits is protein-like
  - Data doesn't have enough information to choose among models

Ramachandran Z-score

Ramachandran Z-score: Method



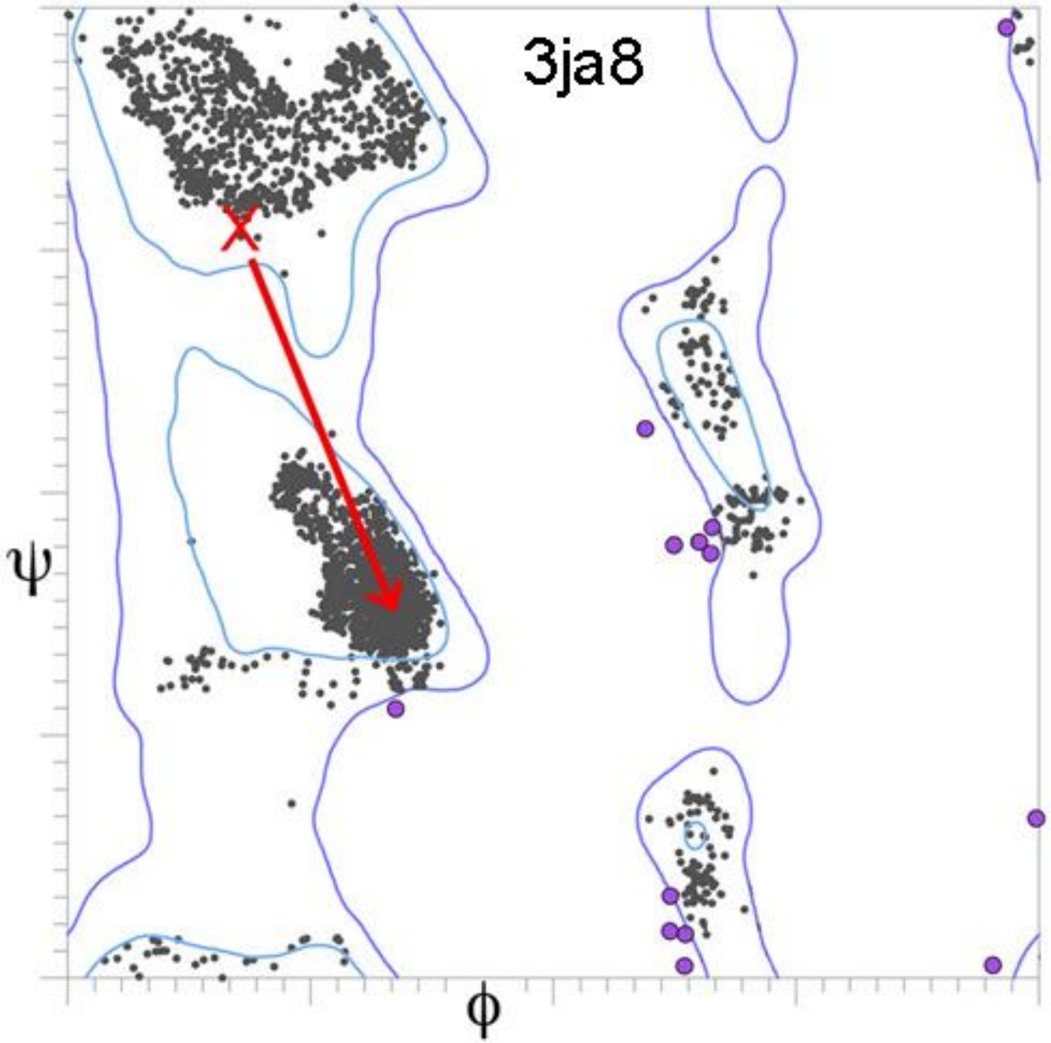
- Compare observed Ramachandran distribution against expected distribution (shown)

- Assign statistical Z-score based on distance from expectation

- |Z-score| <= 2 indicates a realistic distribution
- |Z-score| > 3 indicates a highly unrealistic distribution

phenix.rama_z

Ramachandran: Probable causes



Rama Z-score -4.26 ± 0.10

Overfitting to Rama criteria

- Some programs allow refinement of the Ramachandran plot
    - Hides/compounds rather than fixes errors, if used carelessly
    - Artificially improves Ramachandran and MolProbity scores
- Over-idealized distribution may be detectible by Rama Z-Score
- Use other methods to fix model errors
- Then (maybe) Rama restraints to hold good structure in place

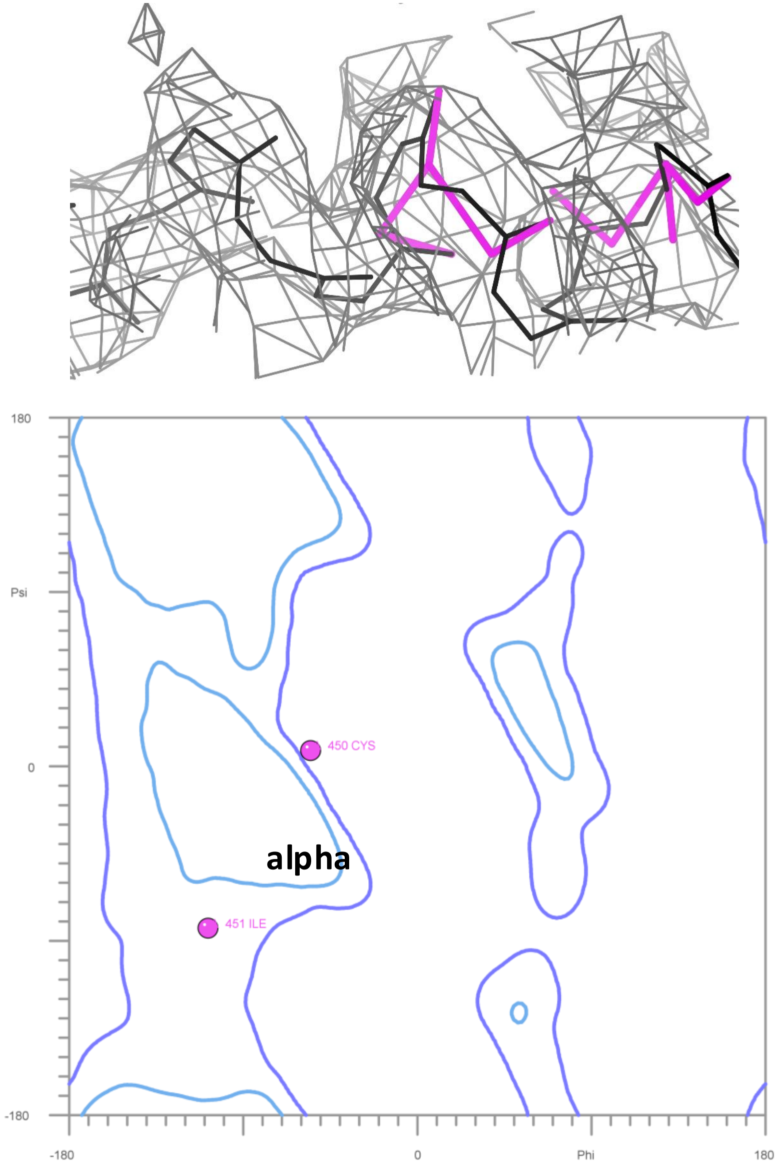
Rama/CaBLAM: Probable causes

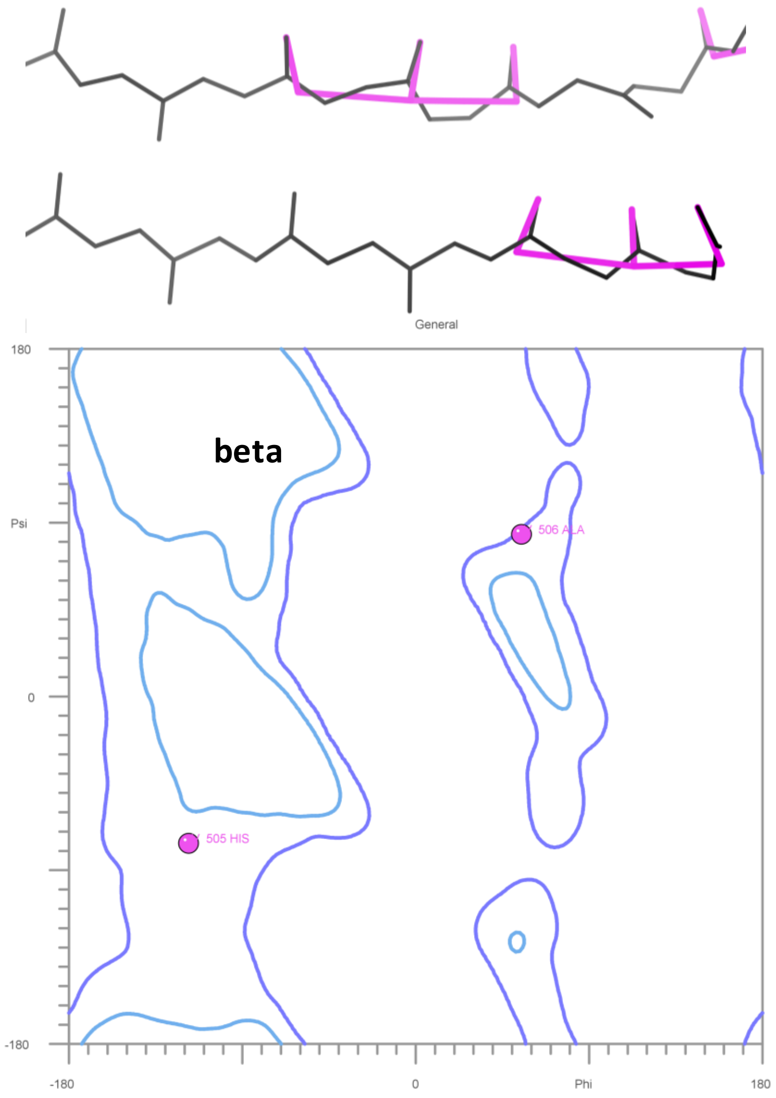
Current Rama position

does not predict

Correct Rama position

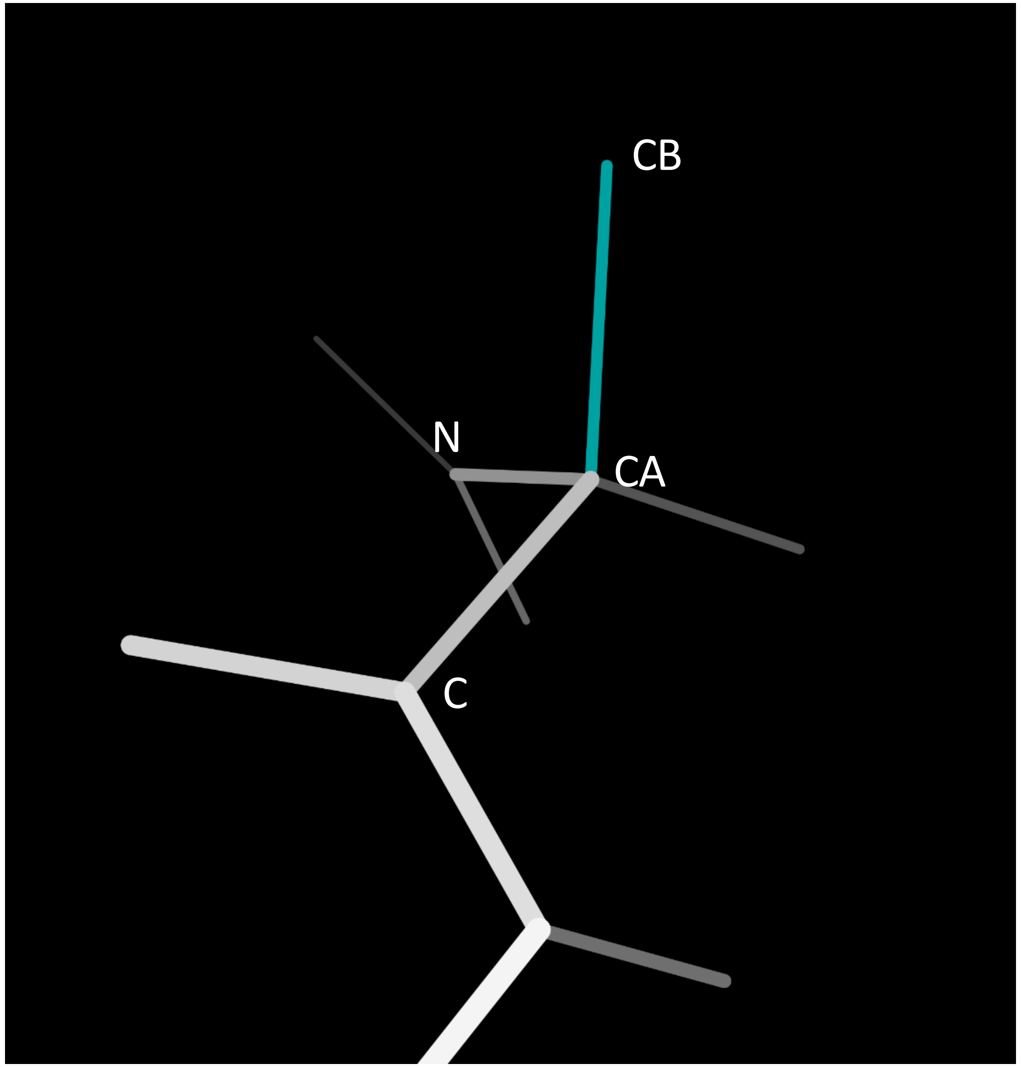
- If model errors are large, points in Rama space are displaced far from their intended regions
- 90° or even 180° peptide orientation errors are possible in low-resolution maps!





C-Beta Deviation

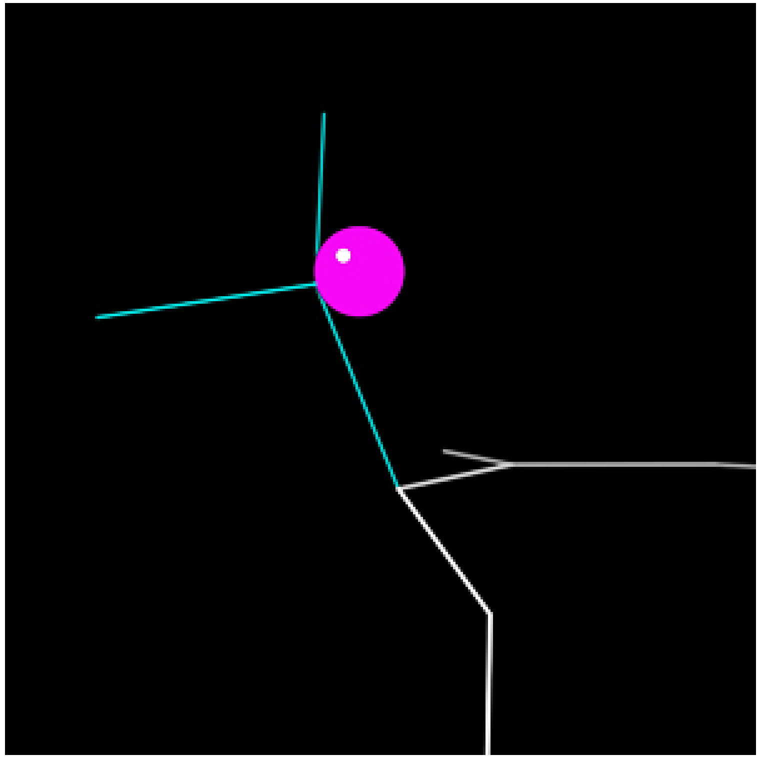
C-Beta Deviation: Method

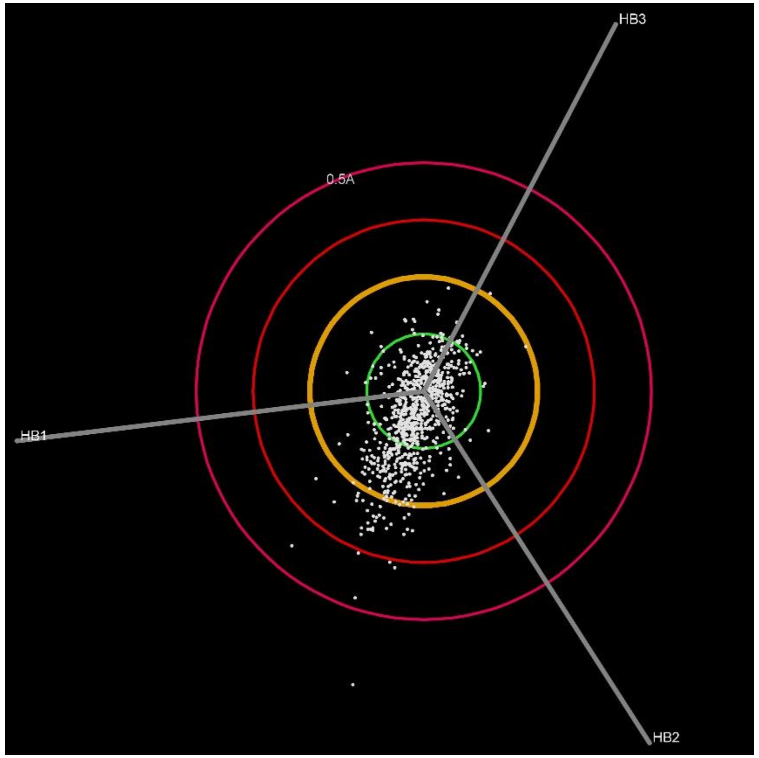


- Ideal CB position is defined by backbone geometry
- Calculate ideal position using average of two torsions
  - N-C-CA-CB
  - C-N-CA-CB
- CBs modeled >0.25Å from ideal position are outliers

molprobity.cbetadev

C-Beta Deviation: Visualization

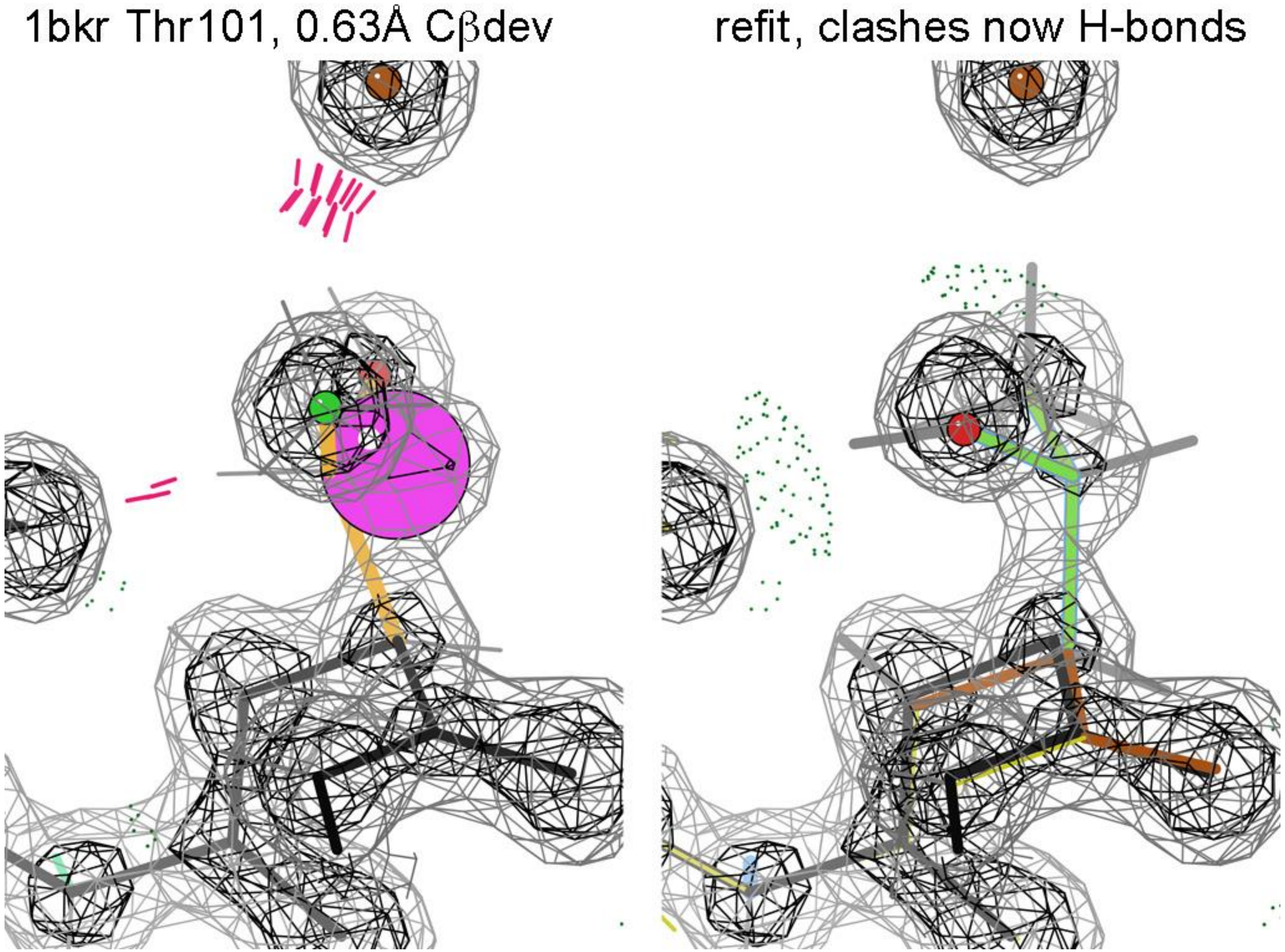




- In KiNG, a magenta sphere is drawn
  - Center at ideal CB position
  - Edge tangent to modeled position
  - Size of sphere proportional to severity of outlier

- Bullseye kinemage shows distribution and direction of all CB positions.
- Yellow circle is 0.25Å outlier cutoff

C-Beta Deviation: Probable causes



1bkr Thr101, 0.63Å Cβdev

refit, clashes now H-bonds

Misplaced sidechains

- CB deviation outliers are usually caused by misplaced sidechains
  - Especially branched sidechains fit backwards, like this Thr

Chirality errors

- If D amino acids are misnamed as L amino acids (e.g. ALA for DAL), or vice versa, very large Cbdevs result

Covalent Bond Geometry

Bond Geometry: Method

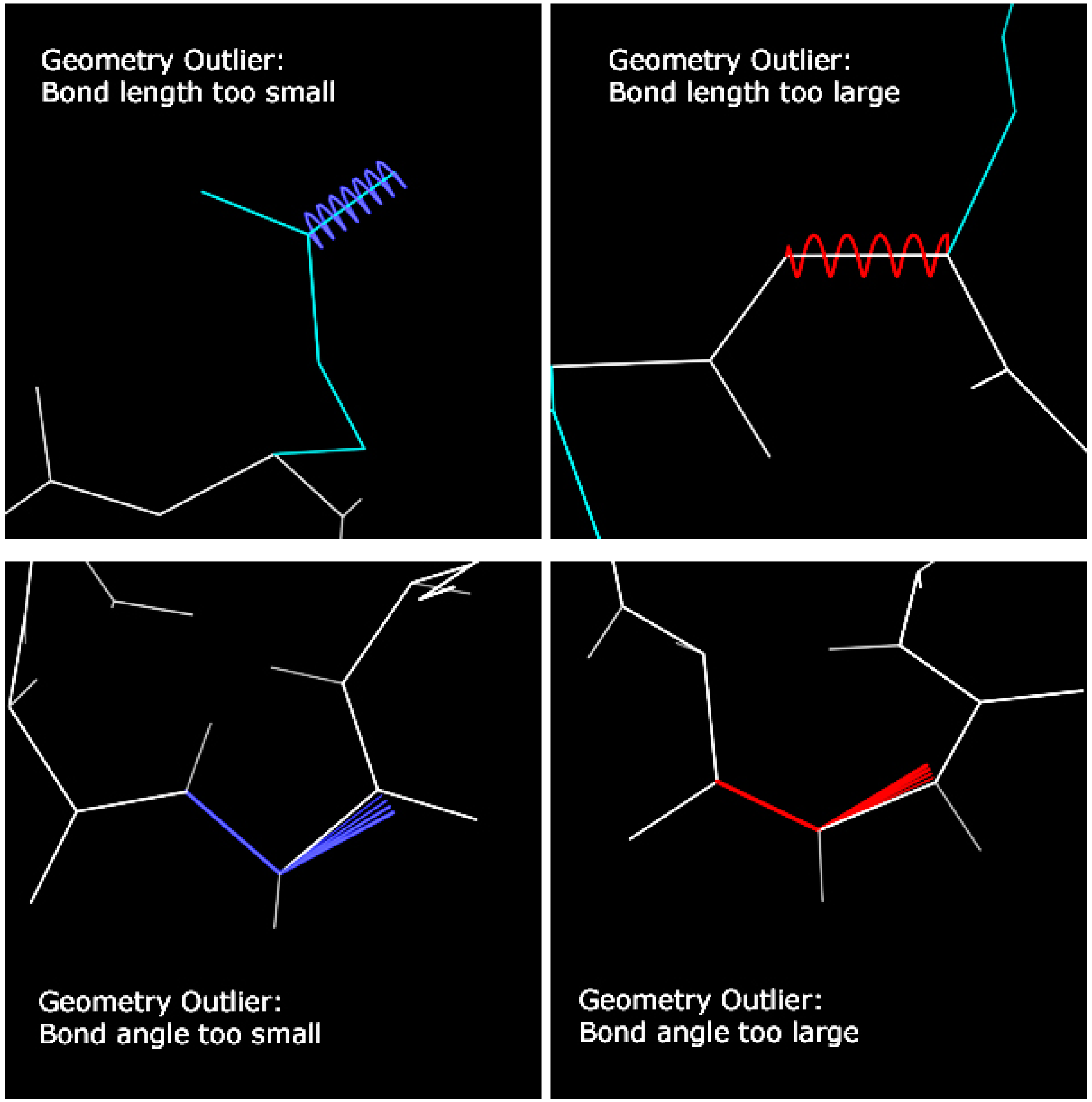
- Measure bond lengths and angles
- Check against a library of expected values
- >4σ deviation from expected = outlier

- Standard reference library has 1 value per bond or angle
- Derived from Engh and Huber
  - https://doi.org/10.1107/S0108767391001071

- Conformation-Dependent Library (CDL) has values that depend on local Ramachandran conformation
- Phenix default
- Derived from Karplus et al.
  - https://doi.org/10.1107/S2059798315022408

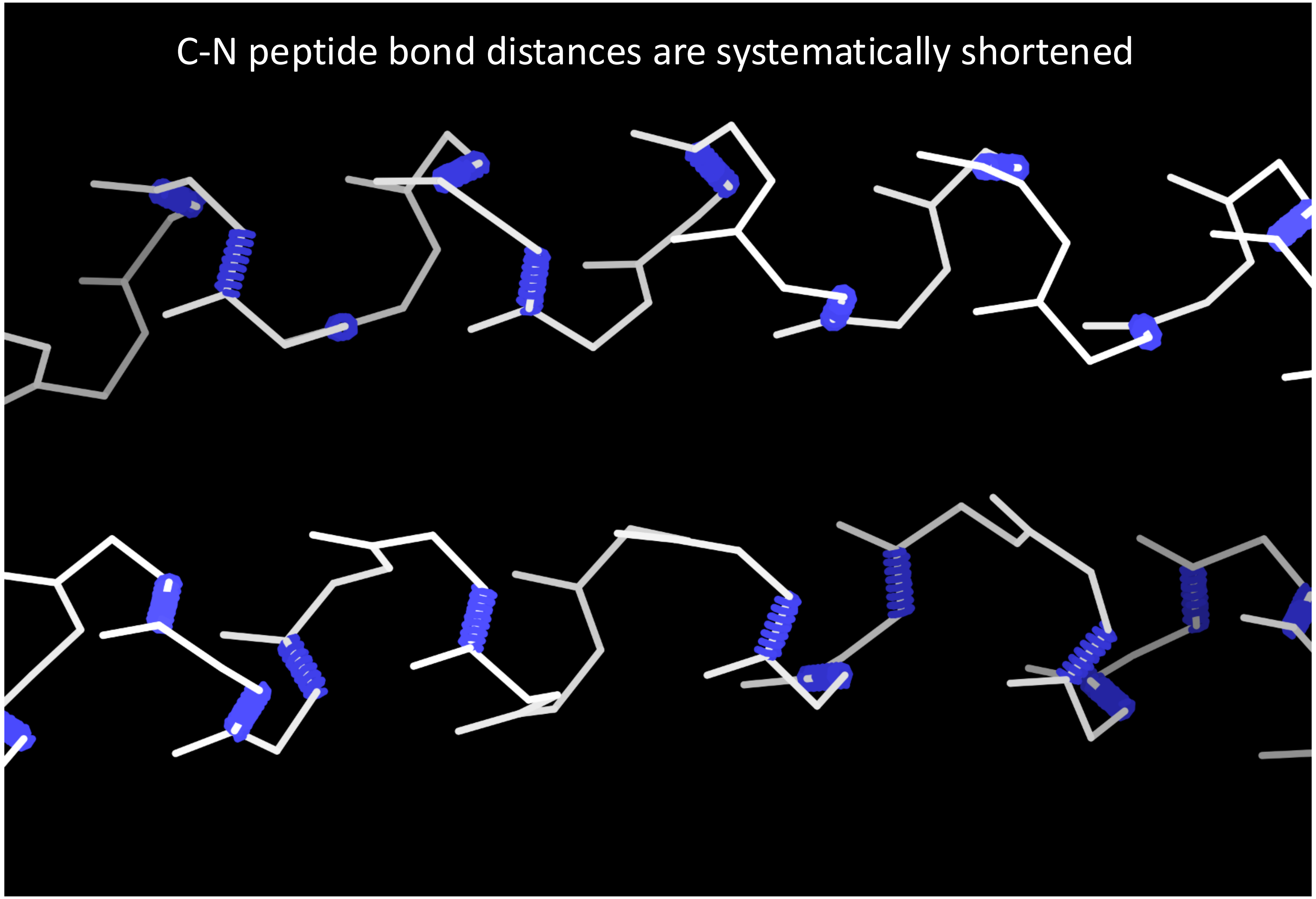
molprobity.mp_validate_bonds

Bond Geometry: Visualization



- Bond length outliers are drawn as springs
- Bond angle outliers are drawn as fans
- Color-coded
  - Red-shift = too far
  - Blue-shift = too close

Bond Geometry: Probable causes

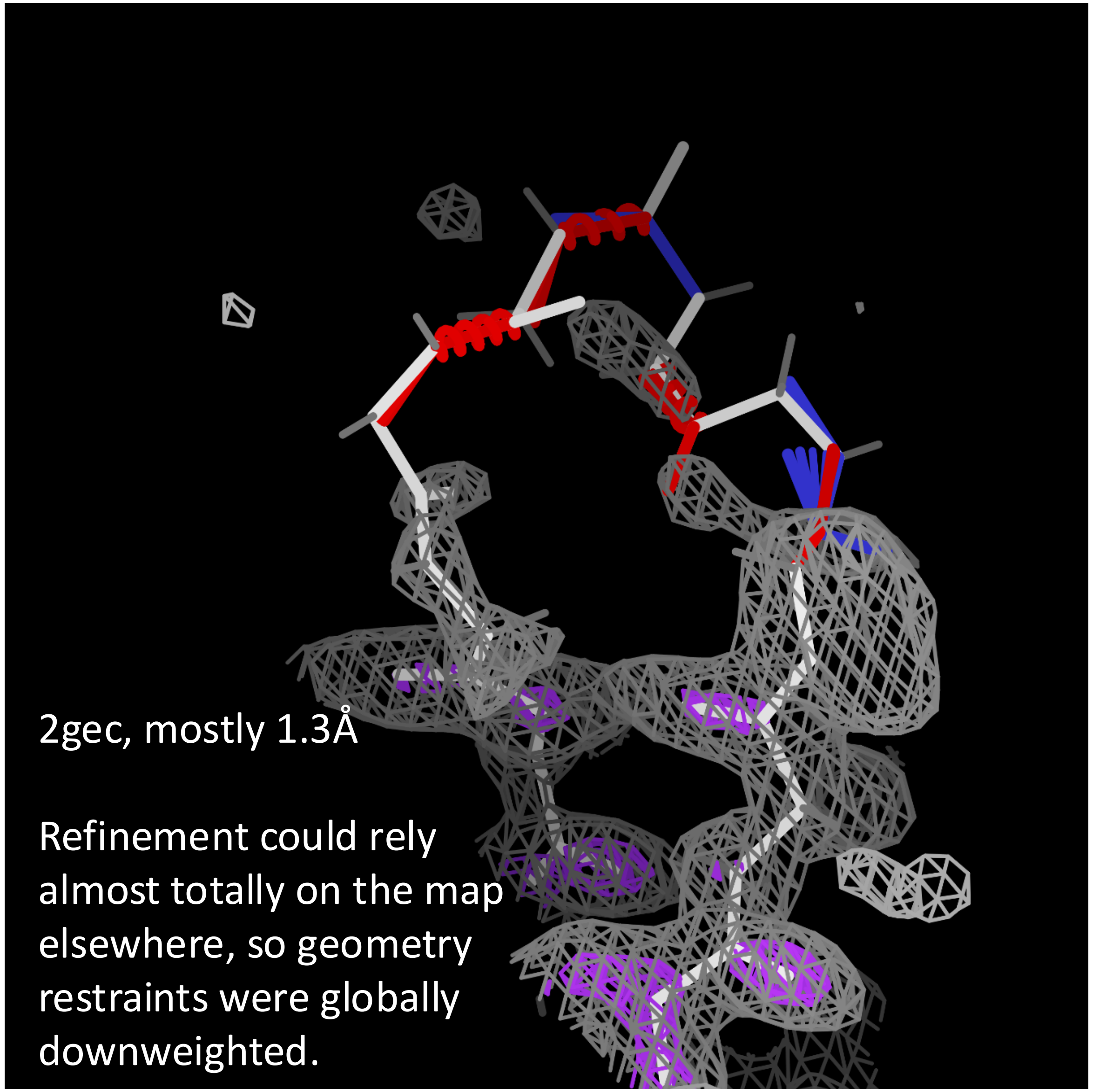


OmegaFold prediction for p81313, as of Sept 2022

Systematic

- Systematic geometry errors occur in programs with different libraries or expectations
- Be aware of what you import
- Do geometry minimization and/or re-refine.

Bond Geometry: Probable causes



2gec, mostly 1.3Å

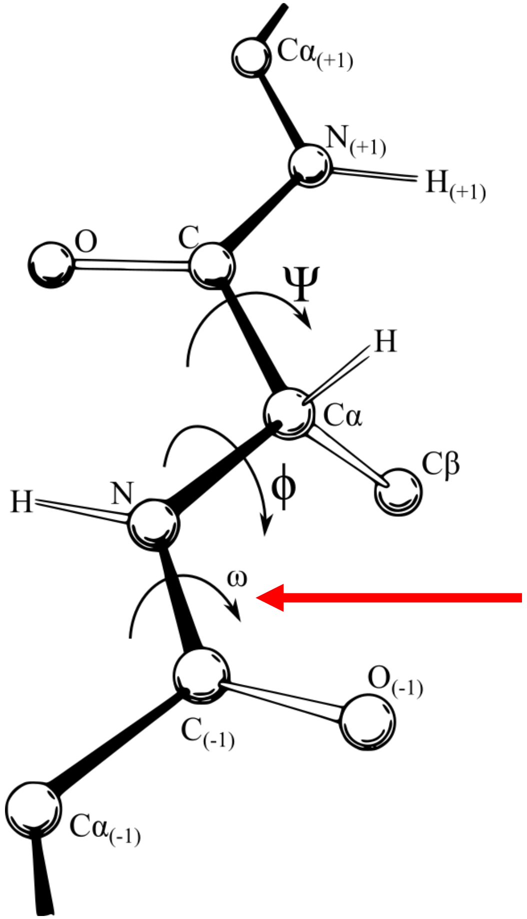
Refinement could rely almost totally on the map elsewhere, so geometry restraints were globally downweighted.

<u>Localized</u>

- Localized geometry outliers result from conformational strain and/or missing density
- Fix the source of strain
- Manually apply more restraints to low-data regions
- Leave it unmodeled if a good solution is impossible

Cis Peptides

Cis Peptides: Method



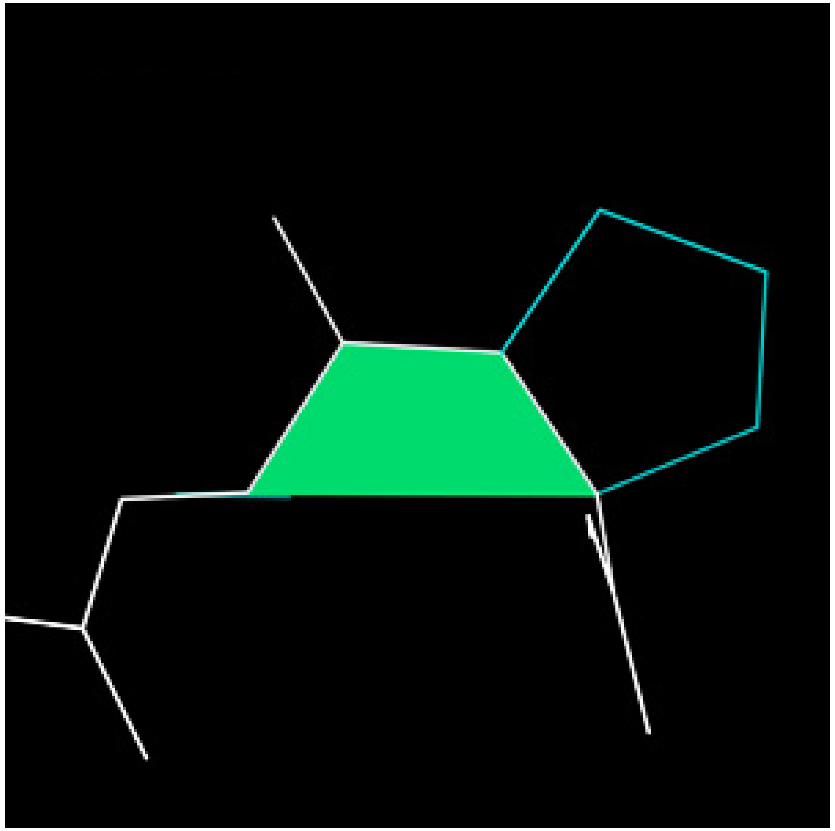
- The peptide bond that joins amino acids has partial double bond character and does not rotate freely

- CA-C-N-CA torsion
  - "Omega"

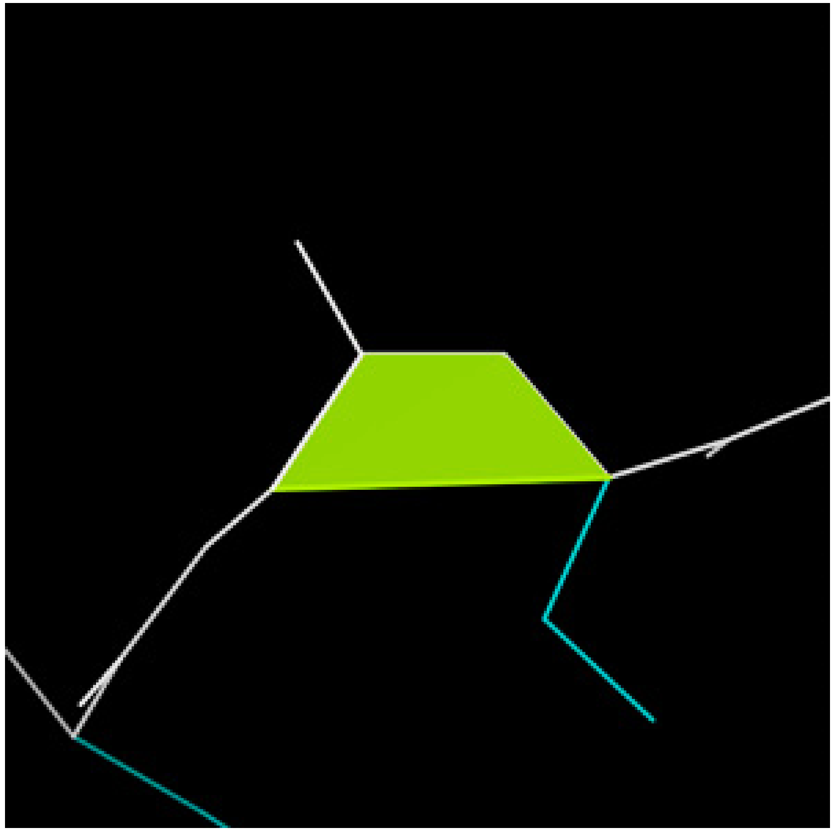
- Usually *trans* (CA on opposite sides)
- Rarely *cis* (both CA on same side)

molprobity.omegalyze

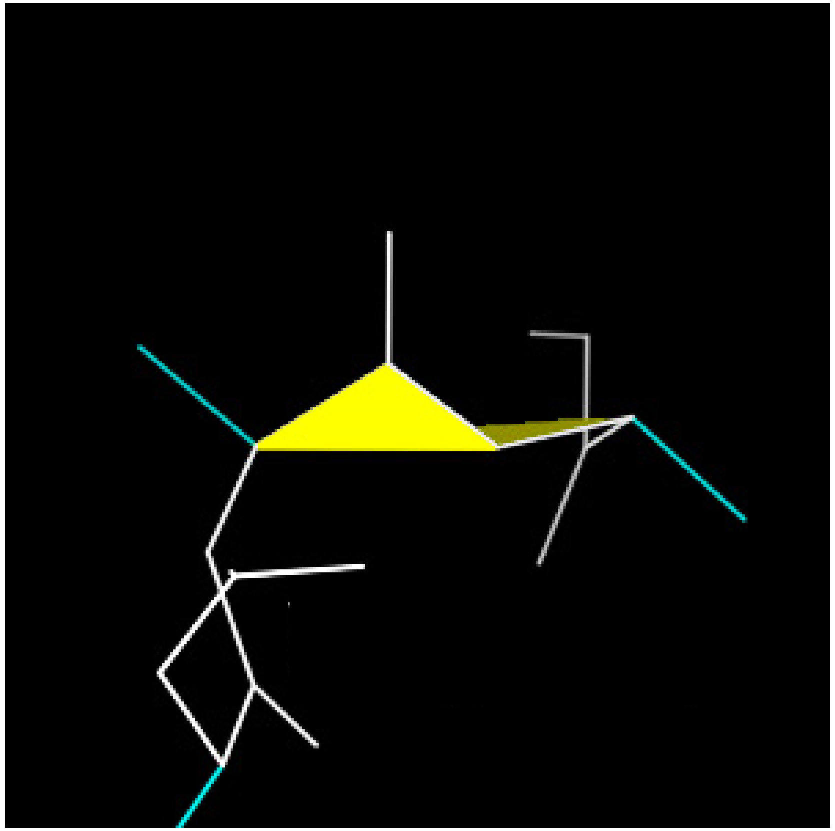
Cis Peptides: Visualization (KiNG)



- *Cis* peptide bond is much more common preceding Proline
  - ~5% of **Proline**
- Gentle green trapezoid fills the characteristic CA-CA space

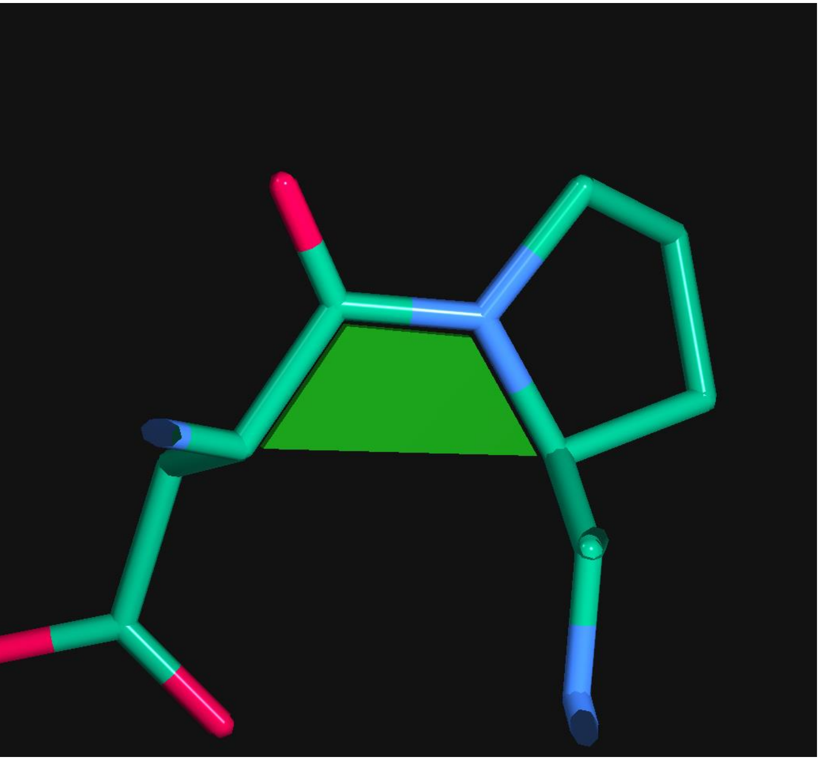


- *Cis* peptide bond is extremely rare preceding other residues
  - ~0.03% of **non-Proline**
- Unpleasantly lime trapezoid fills the characteristic CA-CA space

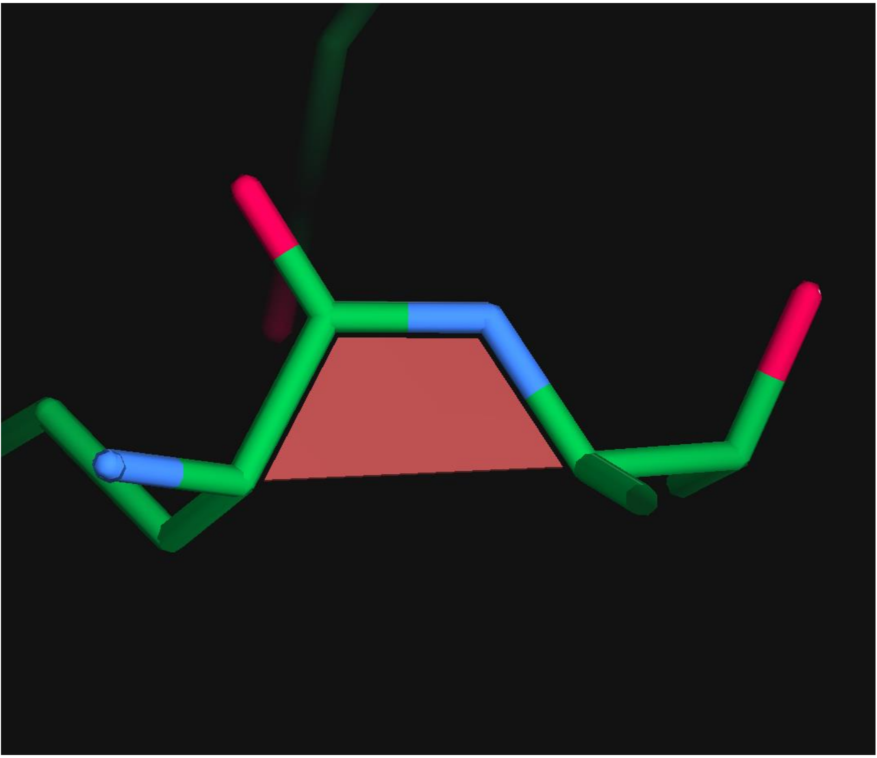


- Peptides **twisted** >30 from planar are severe geometry distortions
- Space is filled with yellow, angle between component planes approximates severity

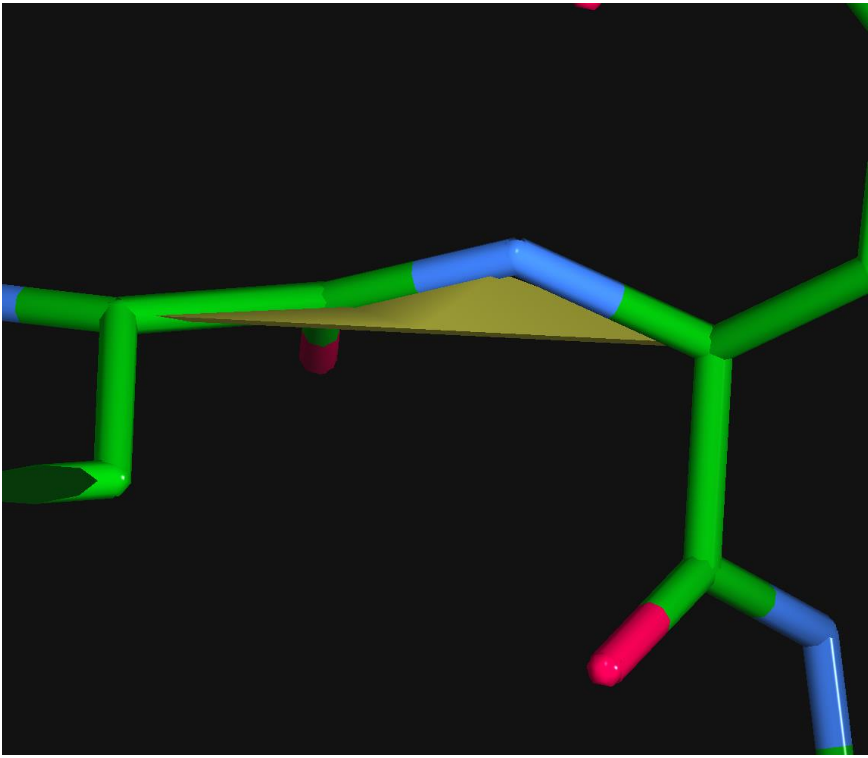
Cis Peptides: Visualization (Coot)



- *Cis* peptide bond is much more common preceding Proline
    - ~5% of **Proline**
- Gentle green trapezoid fills the characteristic CA-CA space



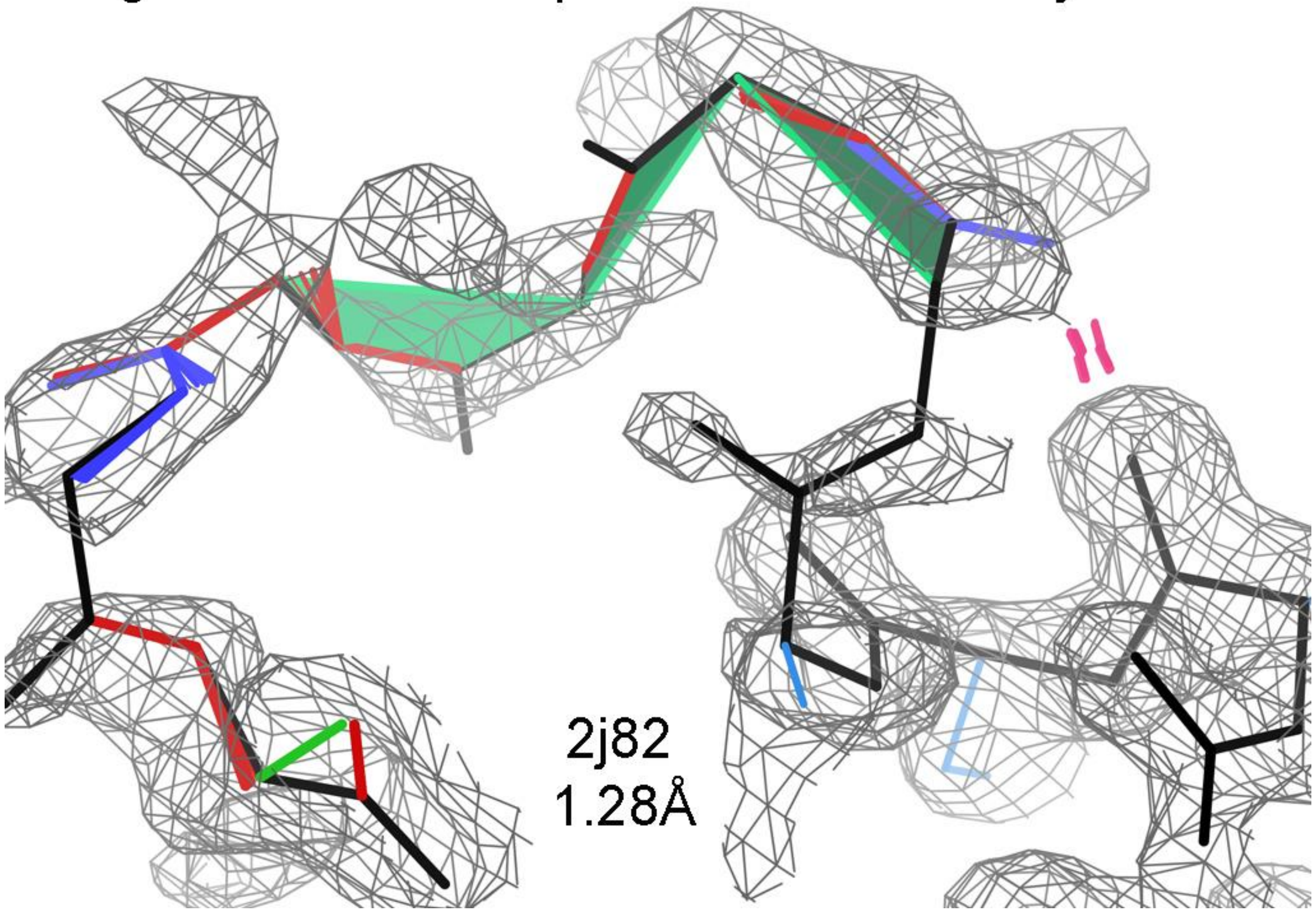
- *Cis* peptide bond is extremely rare preceding other residues
    - ~0.03% of **non-Proline**
- Warning red trapezoid fills the characteristic CA-CA space



- Peptides **twisted** >30 from planar are severe geometry distortions
- Space is filled with yellow, angle between component planes approximates severity

Cis Peptides: Probable causes

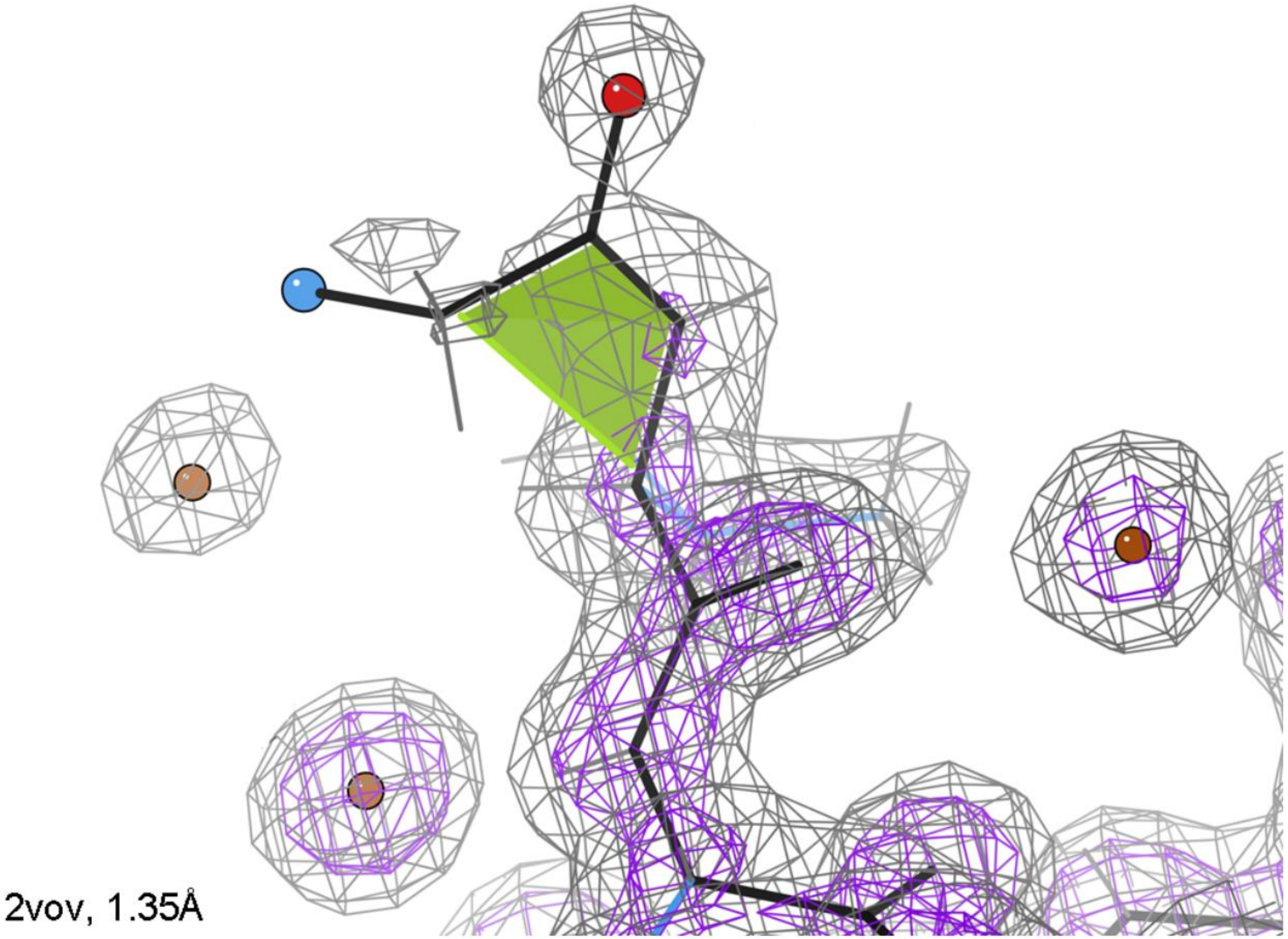
Arg-Gln-Asn-Ser triple *cis*-nonPro -- unjustified



Fit to small density

- The *cis* CA-CA distance is shorter and **seems** to fit better into fragmented density
- A conformation this rare requires more justification than a marginally better fit
- Flip it to *trans* unless density, chemistry, homology, or another source gives you clear support

Cis Peptides: Probable causes



Chain termini

- Non-Pro *cis* peptides at chain ends are always wrong
- Limited density and lack of other constraints *allows* them to be modeled
- But that same lack of constraints means there's nothing to hold an unusual conformation in place

RNA Validations

Rotameric backbone suites

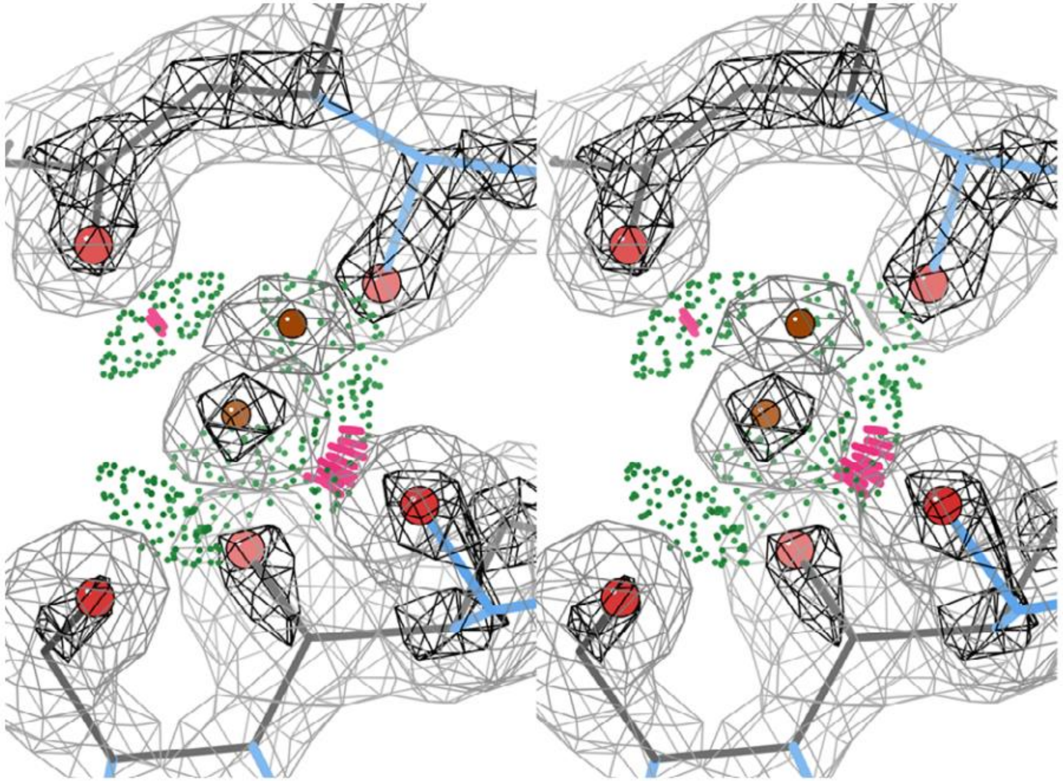
Ribose sugar puckers

(see extras for details)

molprobity.rna_validate
molprobity.suitename

Water Validation

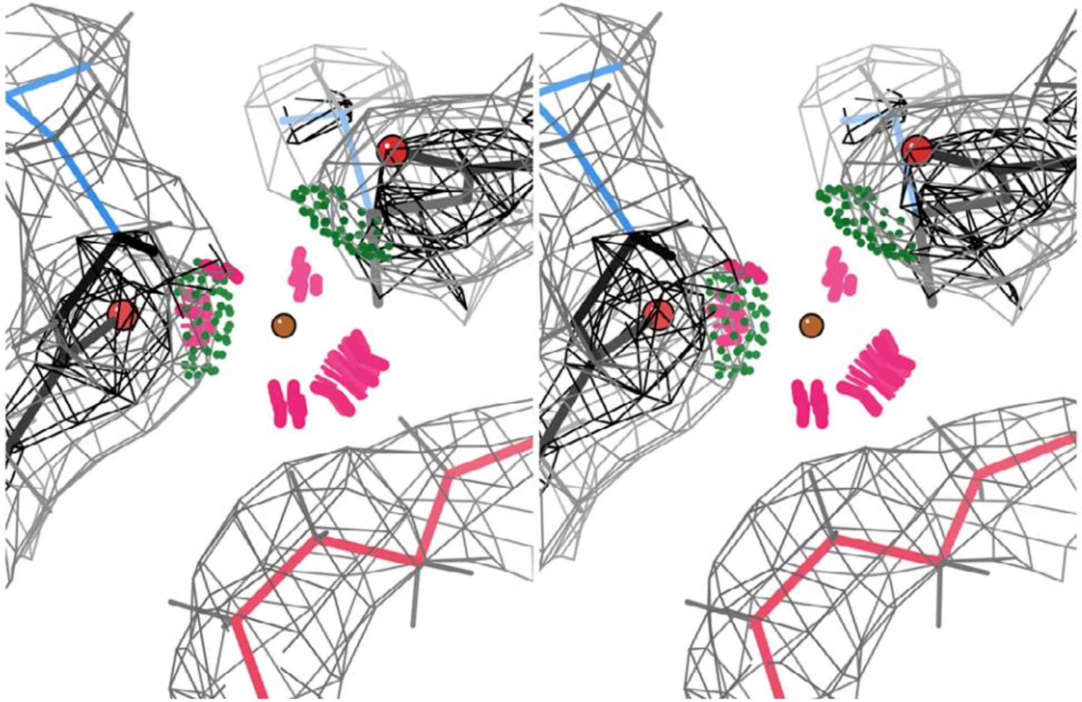
A water that should be an ion



(Stereo image)
HOH 606 from 6hhm, 1.23 Å

- Very strong density peak
- Octahedral contact geometry
  - (water is tetrahedral)
- Contacts are all polar groups (δ-)
- This is actually a + ion, probably Na+

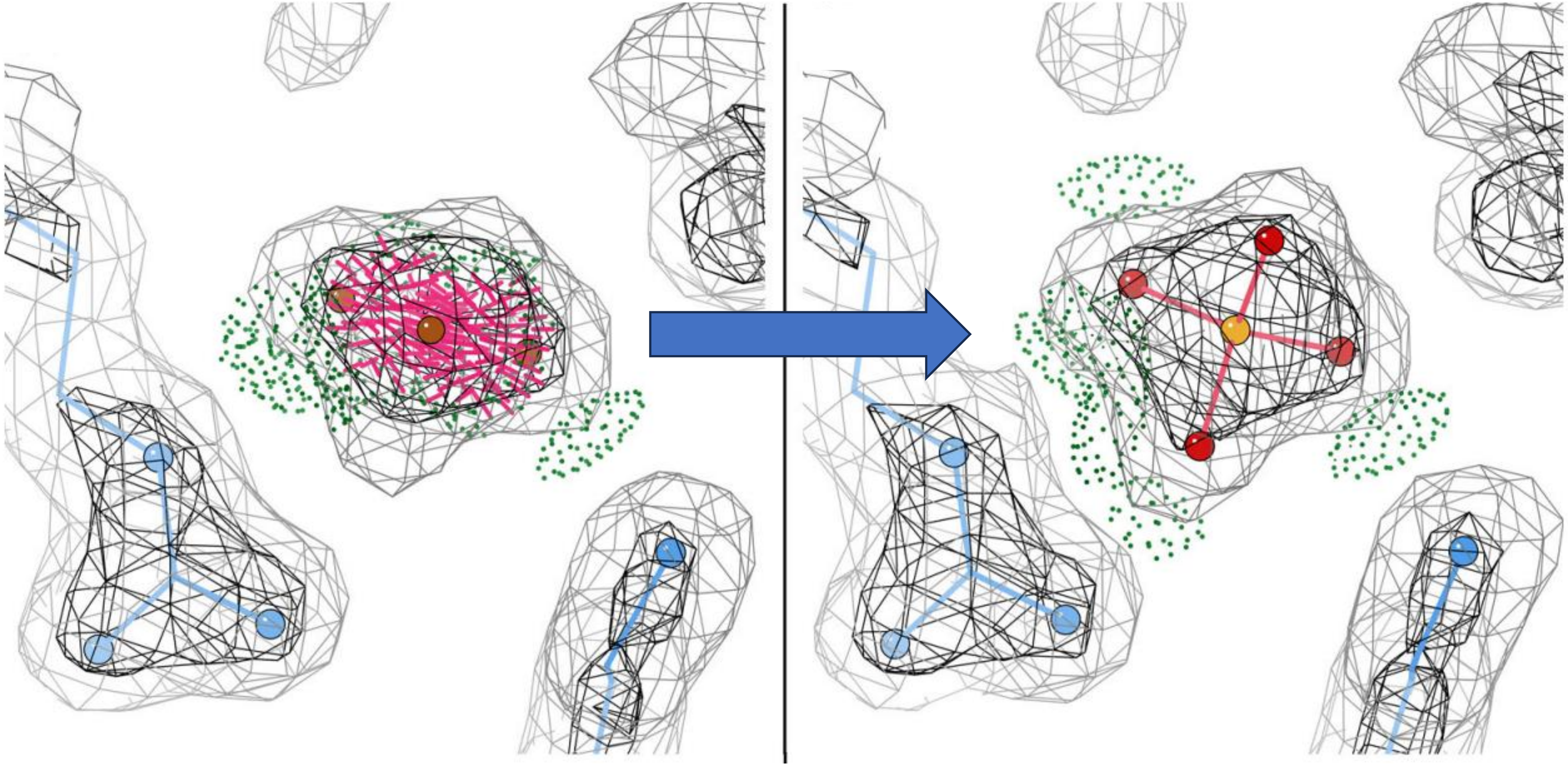
A water that should not be



(Stereo image)
HOH 504 from 5onu, 2.22 Å

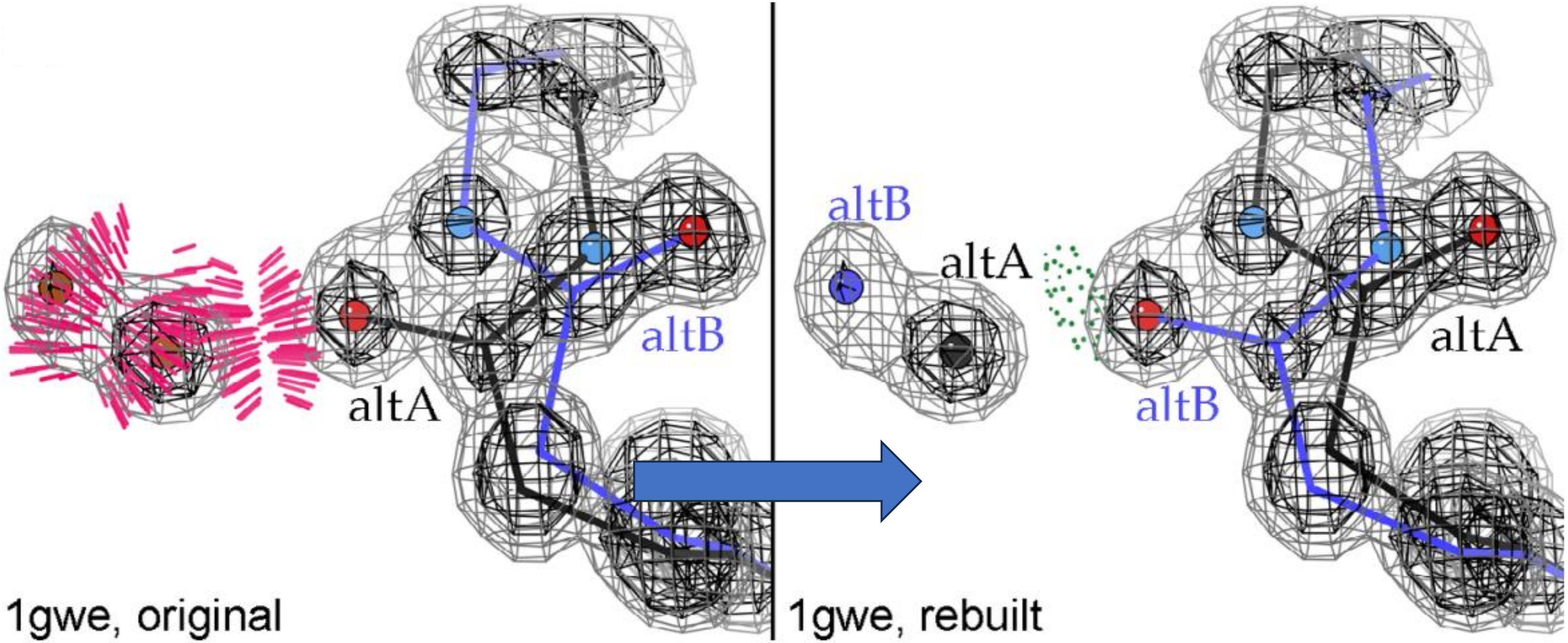
- No density peak
- Mix of polar and non-polar contacts
  - So unlikely to be a coordinated ion
- This water doesn’t really exist

Waters that should be ligands



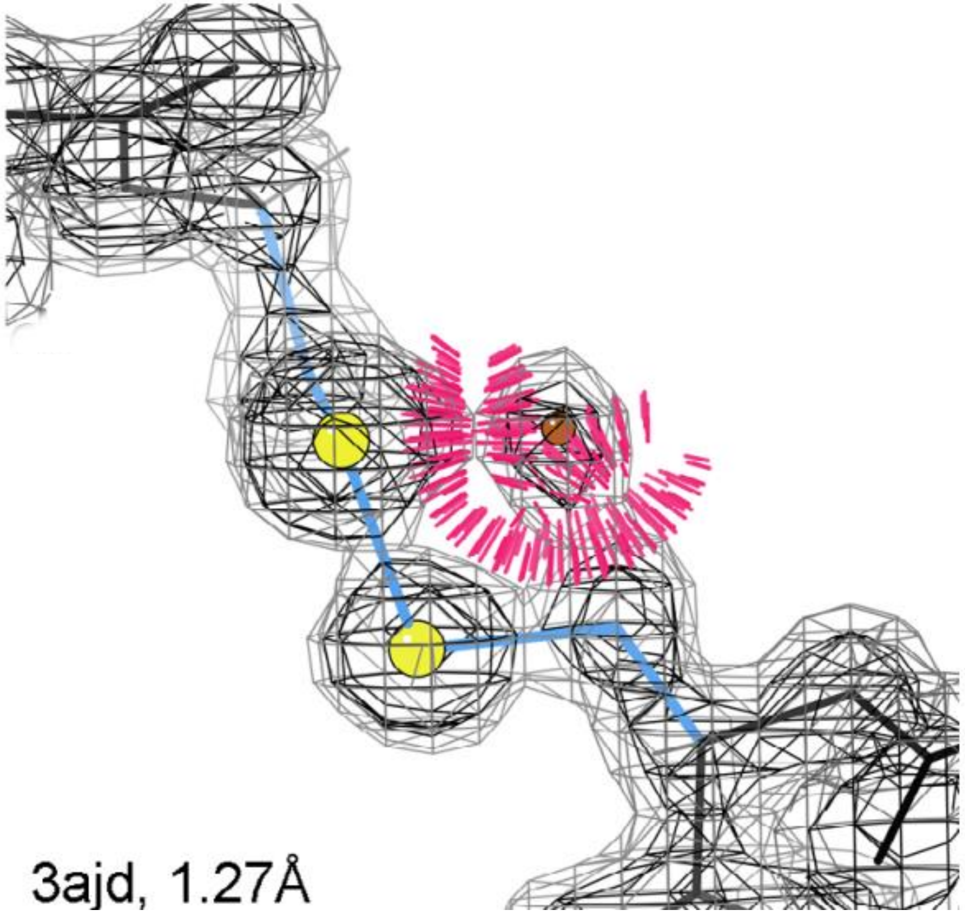
- Densely-clashing waters may actually be a ligand

Waters that should be partial occupancy



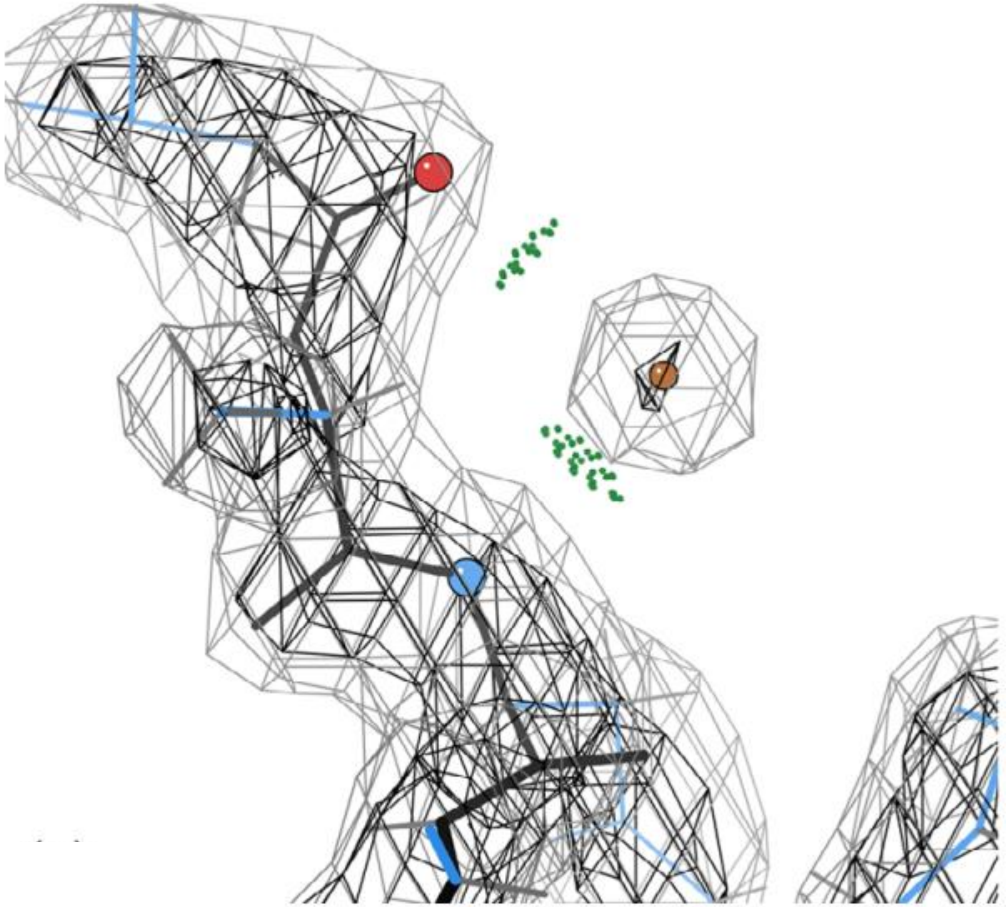
- Densely-clashing waters may actually be part of an alternate conformation network

Waters that replace alternates



- Very close contacts
  - (Covalent bond distance)
- Clash with non-terminal sidechain atoms
- Could be an unmodeled alternate conformation

Waters can be real, too!



- Clear density peak
  - Weaker than macromolecule density is fine
- Hydrogen bonds
- Contacts with both δ+ and δ- polar partners, so an ion is unlikely

MolProbity Score

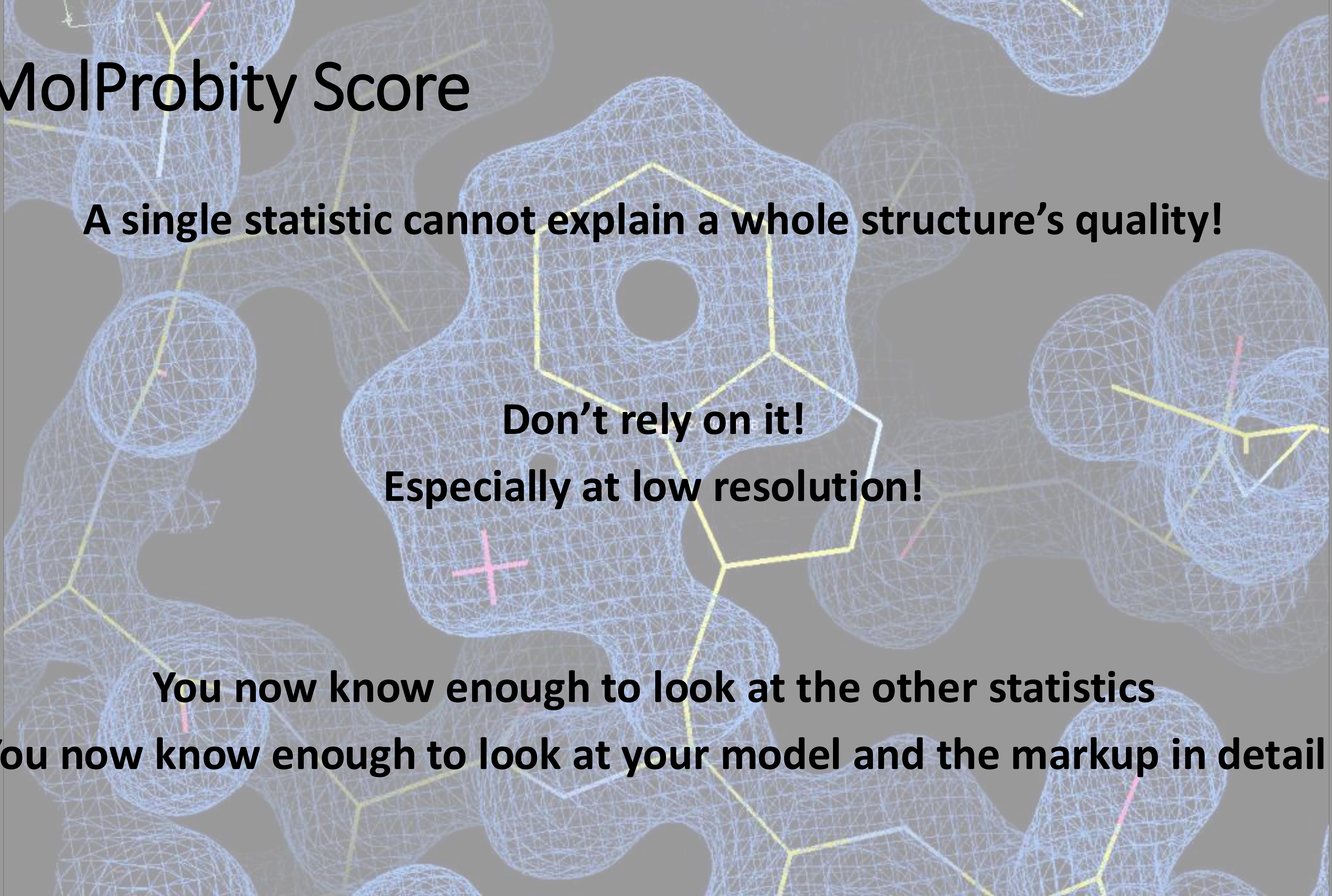
MolProbity Score

- The MolProbity Score combines validations and scales the result to look like a resolution
  - Clashscore
  - Ramachandran
  - Rotamers

- MolProbity better than model resolution is good
- MolProbity worse than model resolution is bad

molprobity.molprobity

MolProbity Score



A single statistic cannot explain a whole structure's quality!

Don't rely on it!

Especially at low resolution!

You now know enough to look at the other statistics

You now know enough to look at your model and the markup in detail

When do you stop?

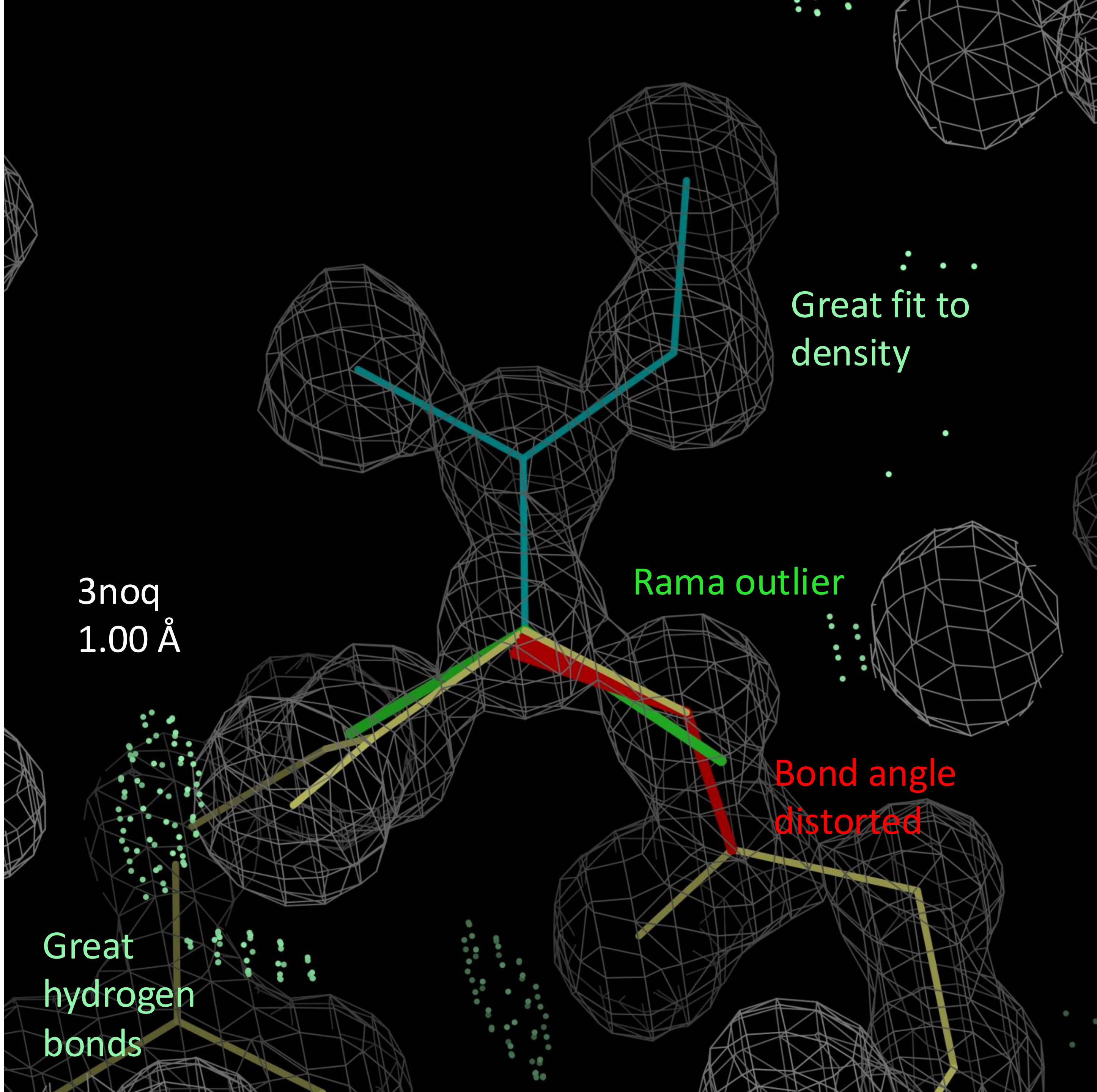
- Realistically? Do as much as you can.
  - Ideally stop when you – and refinement – can’t make the structure better

- Zero outliers is not the goal!
  - Some outliers are justified
  - Some outliers are not justified, but can’t be fixed

- If you can’t obtain a physically-reasonable solution, consider deleting the region.

Outliers can be real

- Zero outliers should **not** be the goal.
- Rama outlier, supported by data and environment.



AlphaFold validation

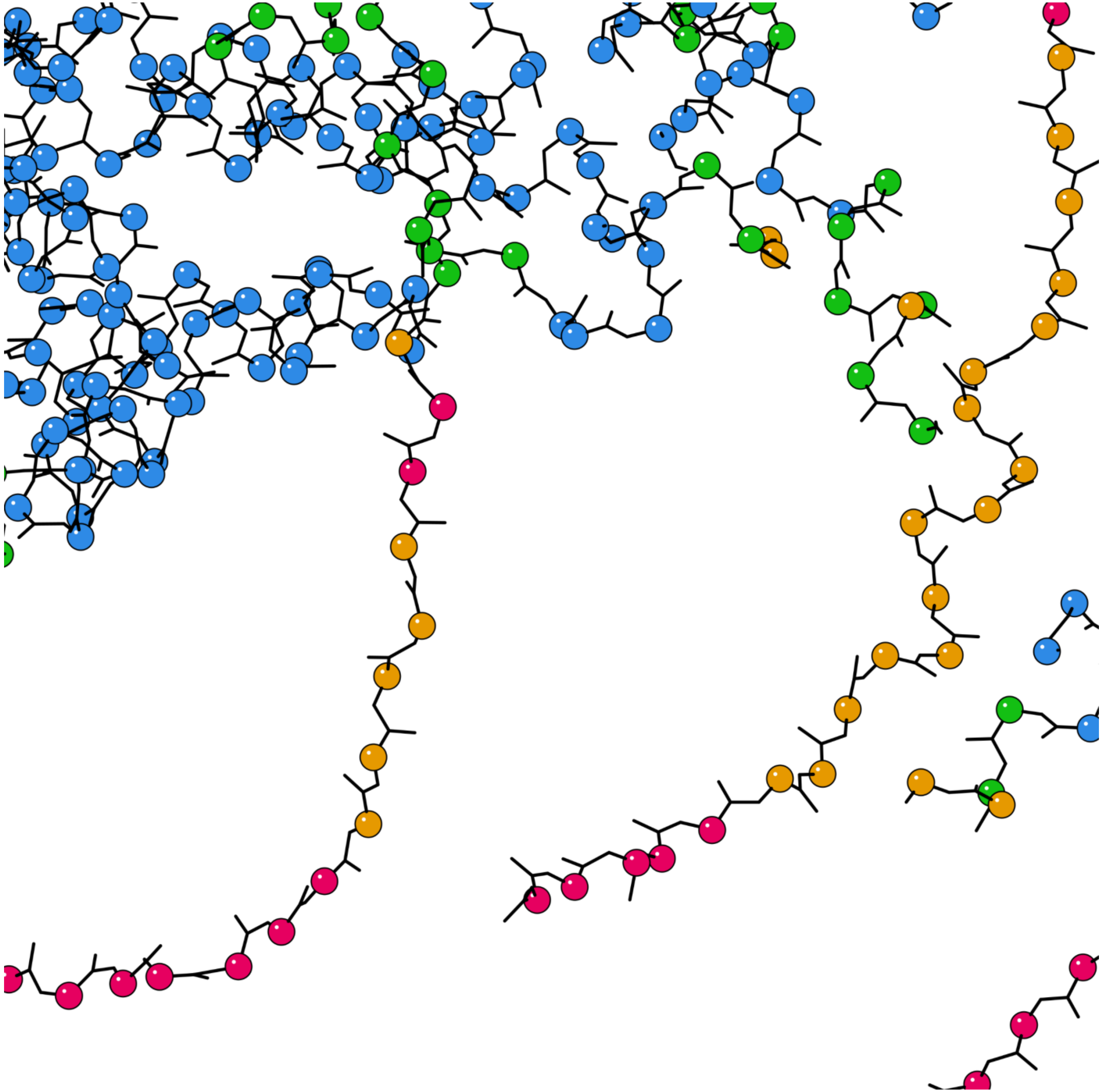
```
phenix.barbed_wire_analysis output.type=kin
```

(under development)

Validation tool

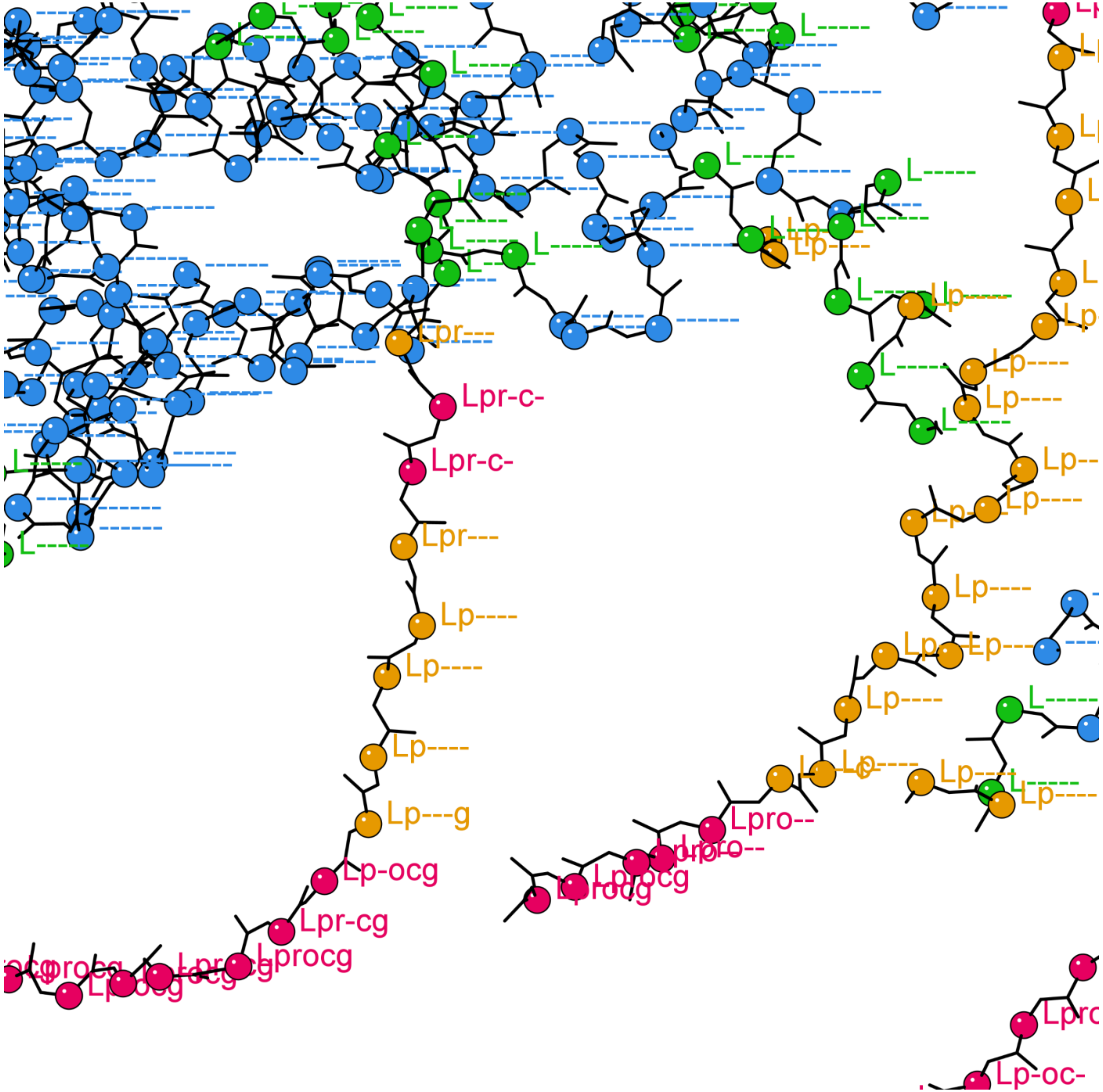
- Predictive (blue)
- Unpacked high pLDDT (gray)
- Near-predictive (green)
- Pseudostructure (gold)
- Barbed wire (hot pink)

- Note barbed wire/unpacked possible transitions



Validation tool

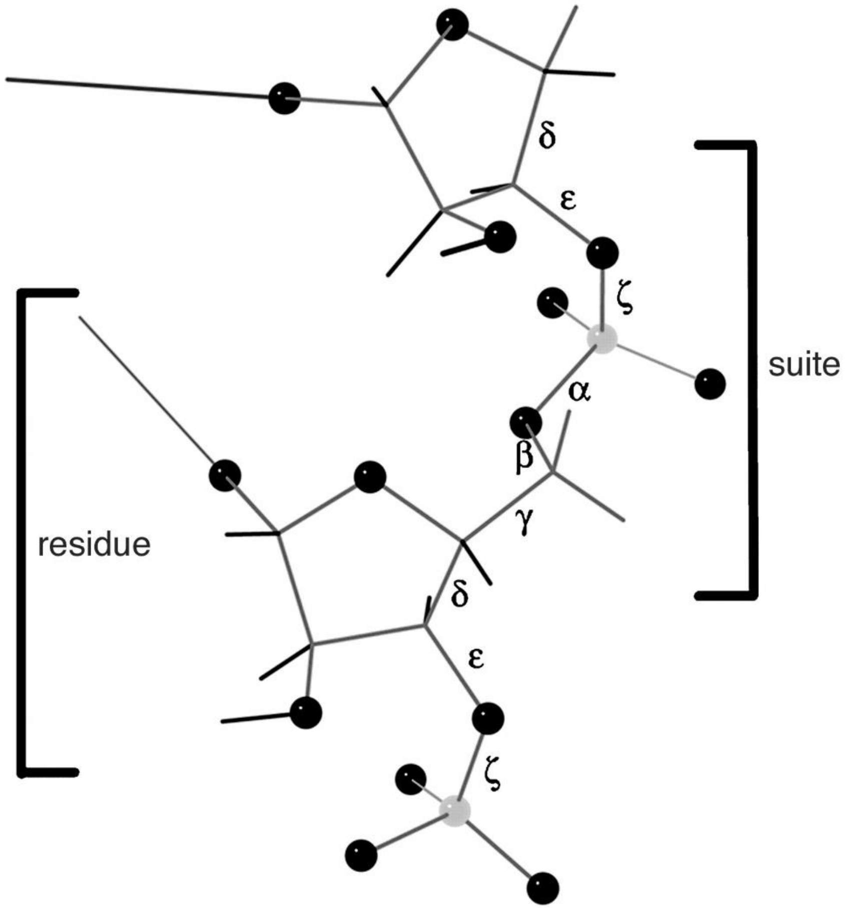
- Letter codes show assessment of each residue
- More letters = more barbed-wire-like
  - L = low pLDDT
  - p = low packing
  - r = bad Rama
  - o = bad omega (cis)
  - c = bad CaBLAM
  - g = bad bond geometry



(In KiNG, press “w” for larger font)

RNA Suites

RNA Suites: Method

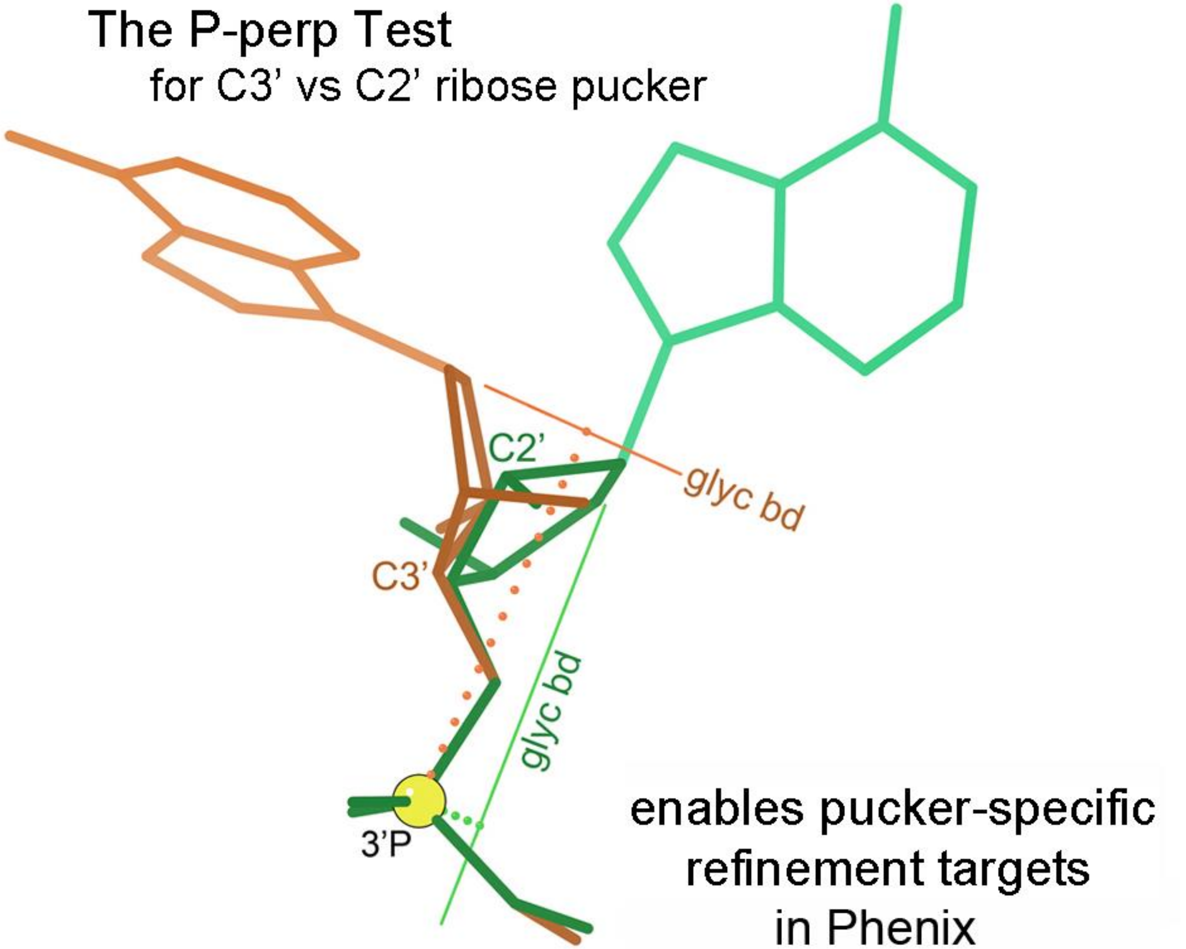


- Useful RNA backbone division is sugar-to-sugar suite, not P-to-P residue
- Suite conformation names are a combination of a number and a letter/character
  - e.g. 1A is the most common A-form helix conformation
- Outliers are named as !!
  - Pronounced “bang, bang”
  - Many !!’s are real, rare conformations

molprobity.suitename

RNA Ribose Puckers

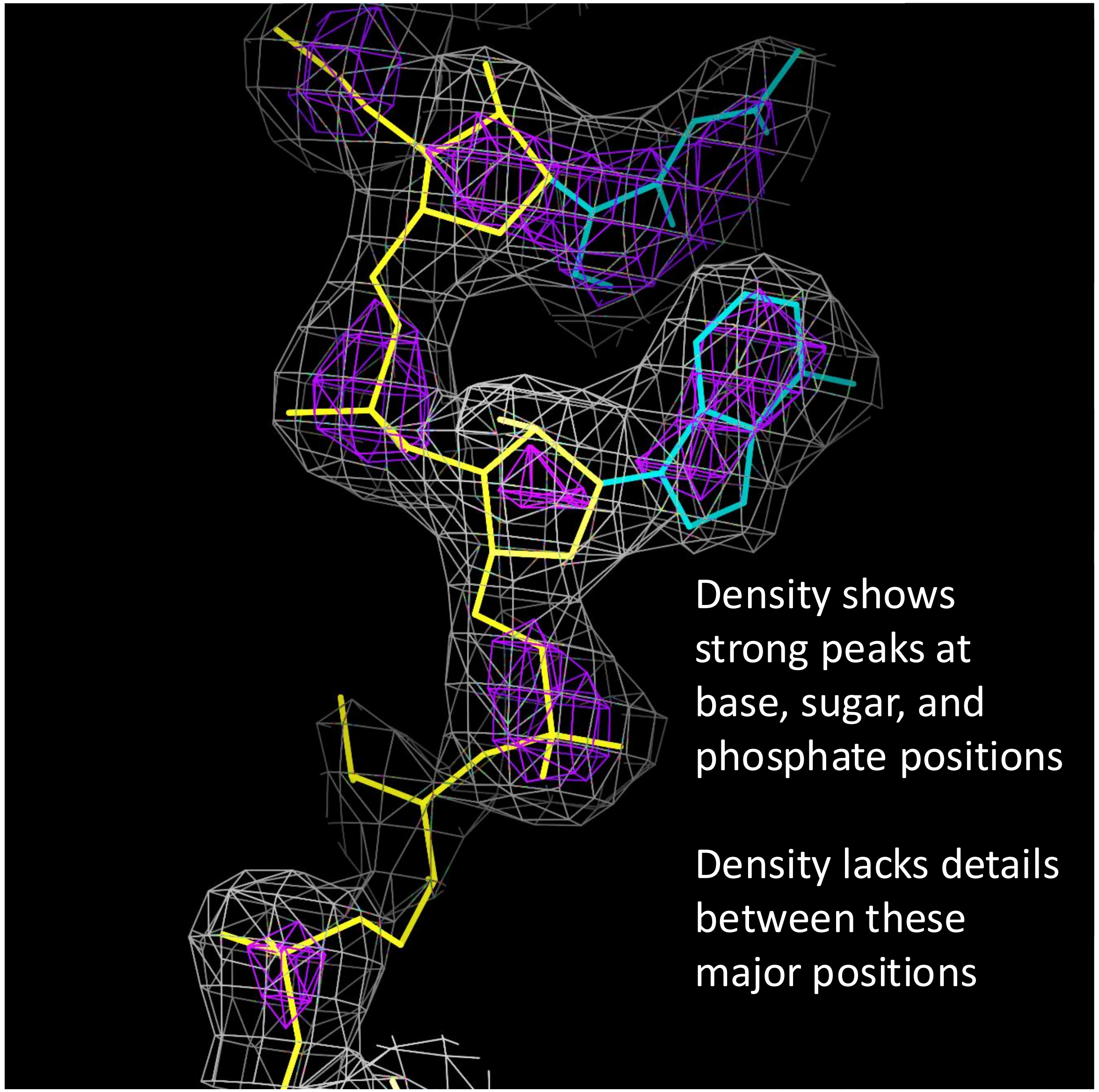
RNA Ribose Puckers: Method



- The backbone ribose in RNA can have one of two pucker states
  - C2' endo
  - C3' endo
- Ribose pucker correlates very strongly with perpendicular distance from the 3'phosphate to the glycosidic bond vector
  - Glycosidic bond joins ribose sugar to nucleobase
- At low resolution, perpendicular distance is easy to see, ribose pucker is hard to see
- If there's a mismatch, the pucker is probably wrong

molprobity.rna_validate

RNA Errors: Probable Causes



Density shows strong peaks at base, sugar, and phosphate positions

Density lacks details between these major positions

- RNA backbone has many degrees of freedom
- Electron density often leaves RNA backbone underdetermined
  - Even when bases are better resolved
- More tools to help with this are in development

Resolution and the Limits of Validation

At 1.5Å to 2.5Å

MolProbity is still very effective.

The density contains enough specific information
that where your model fits the density,
the simple validations (geometry, Rama, rotamers),
and the explicit-H all-atom contacts

then **it's pretty sure to be accurate !**

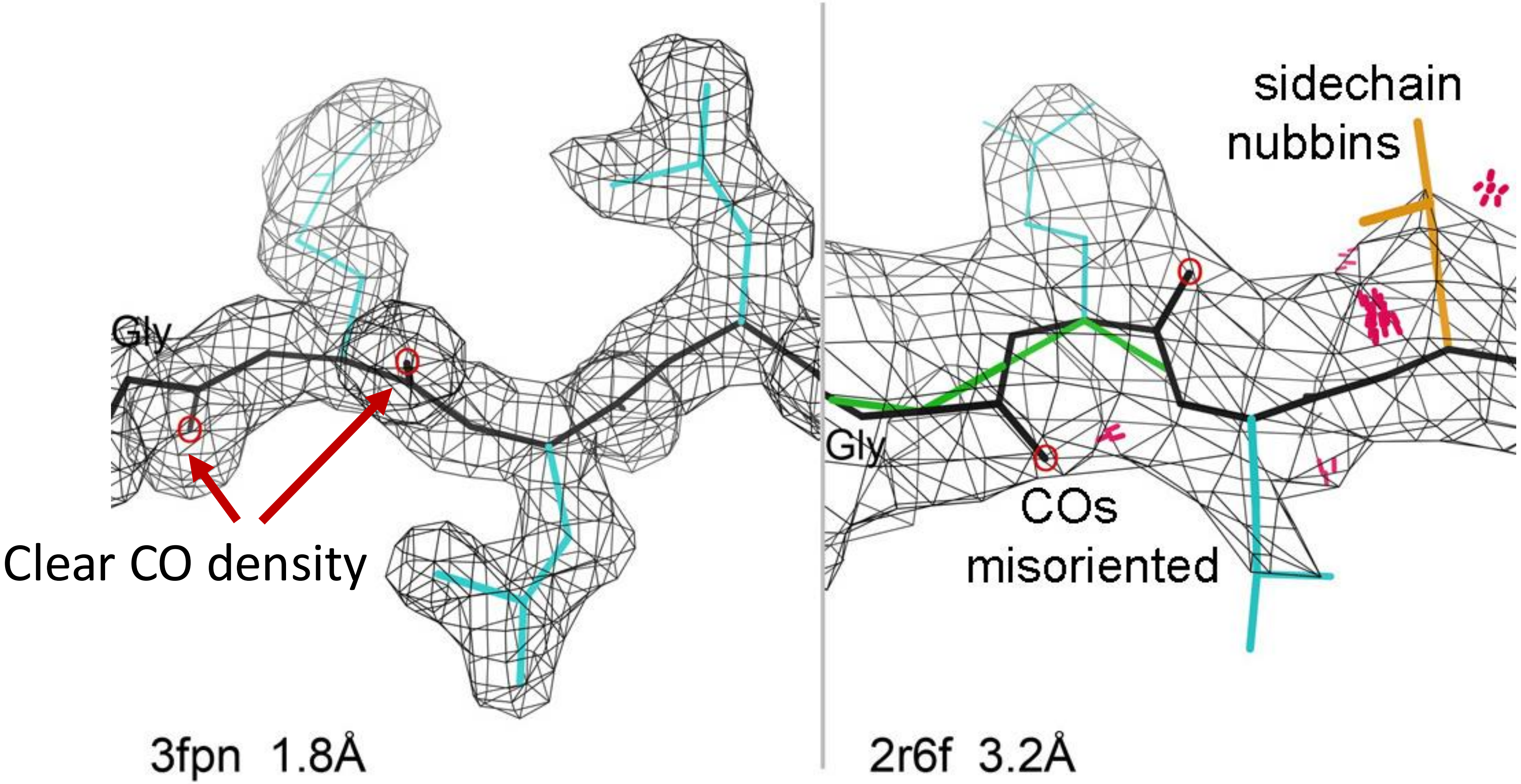
But that's not true at 3 to 4Å !!

Why does this happen ?

What are we doing about it ?

Tackling lower resolution (2.5 to 4Å)

Very challenging both for x-ray and for cryoEM



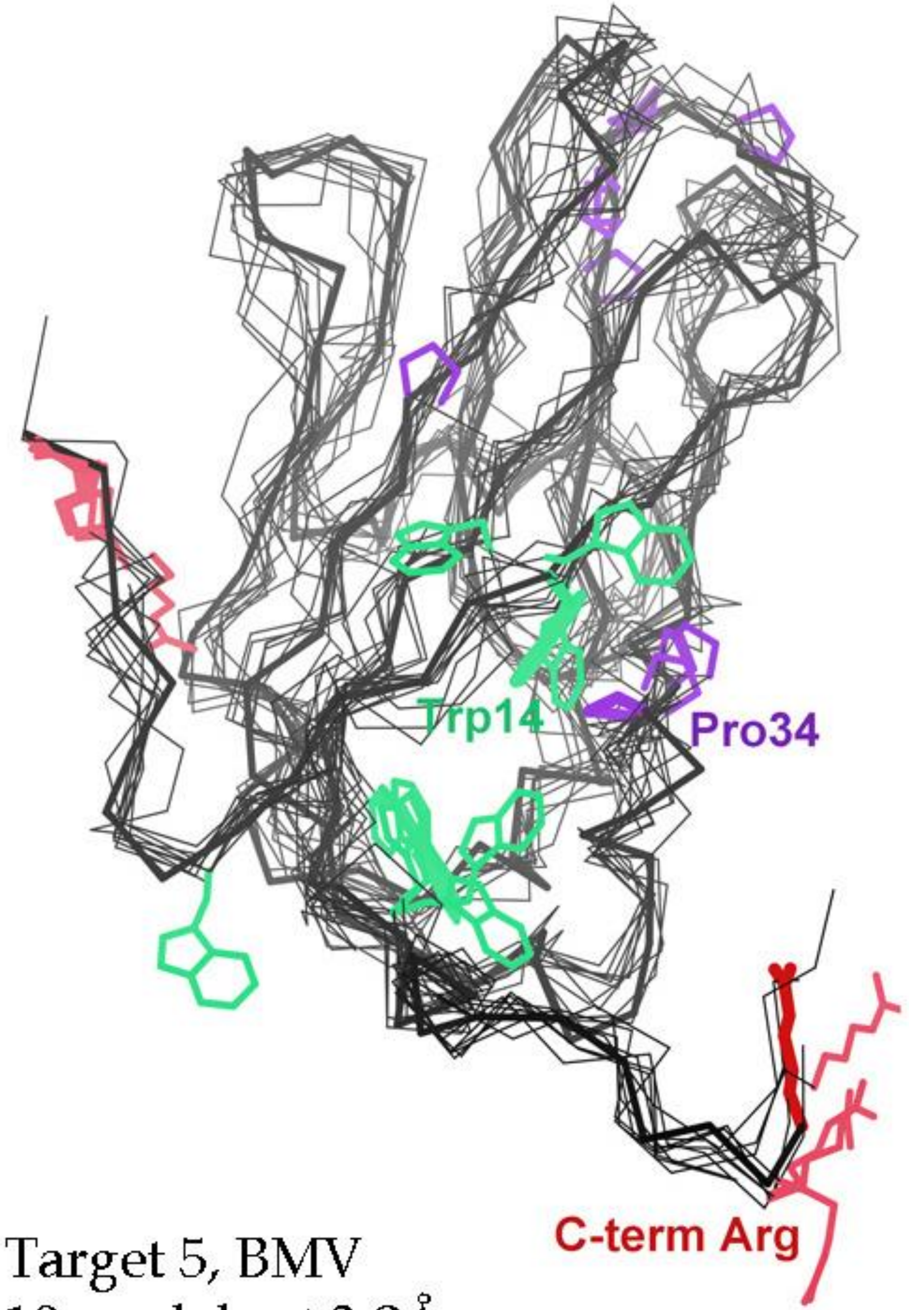
Clear CO density

3fpn 1.8Å

sidechain nubbins

COs misoriented

2r6f 3.2Å



Target 5, BMV
10 models at 3.8Å

At 3-4Å,
many distinct
models are equally
compatible with
the broad density

Much other information
is needed, which can
lead to overfitting
and systematic errors