Molecular modeling 2018 -- Lecture 12

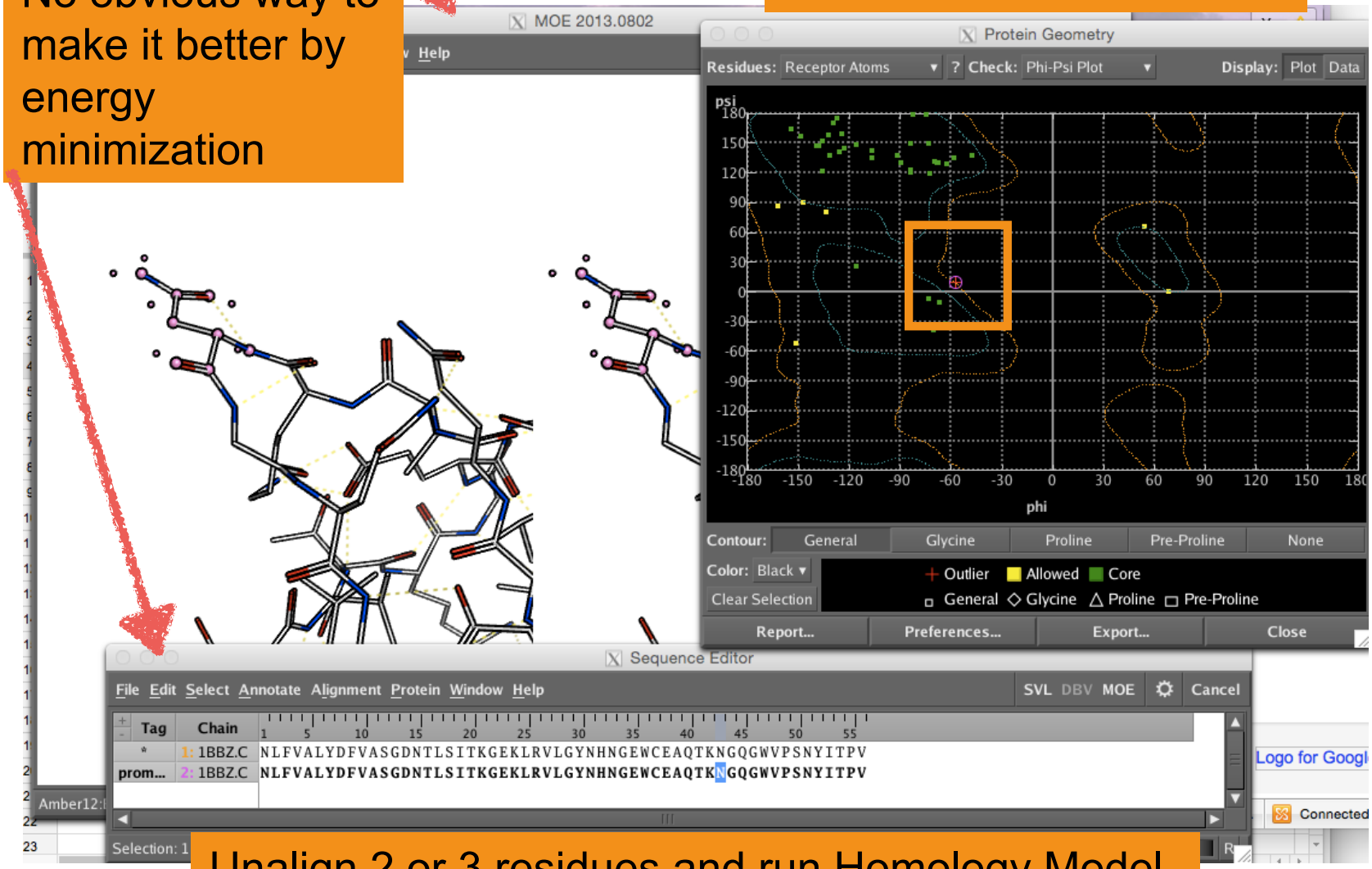
2nd pass modeling
Validation

2nd pass modeling

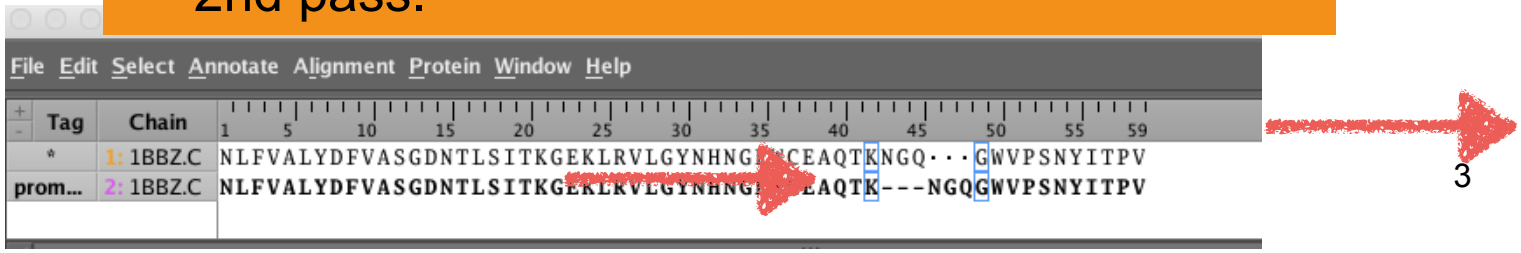
- After homology modeling using the automated script, you should inspect.
  - Search for (a) outliers in the Ramachandran plot, (b) buried charges, (c) hydrogen bonds in the core of the protein that are not made.
  - Fix the problems by ...
    - judicious energy minimization.
    - energy minimization with restraints
    - 2nd pass homology modeling:
      - Re-open the sequence file
      - Align
      - Unalign a few residues around where the problem is.
      - Run Homology Model again.

No obvious way to make it better by energy minimization

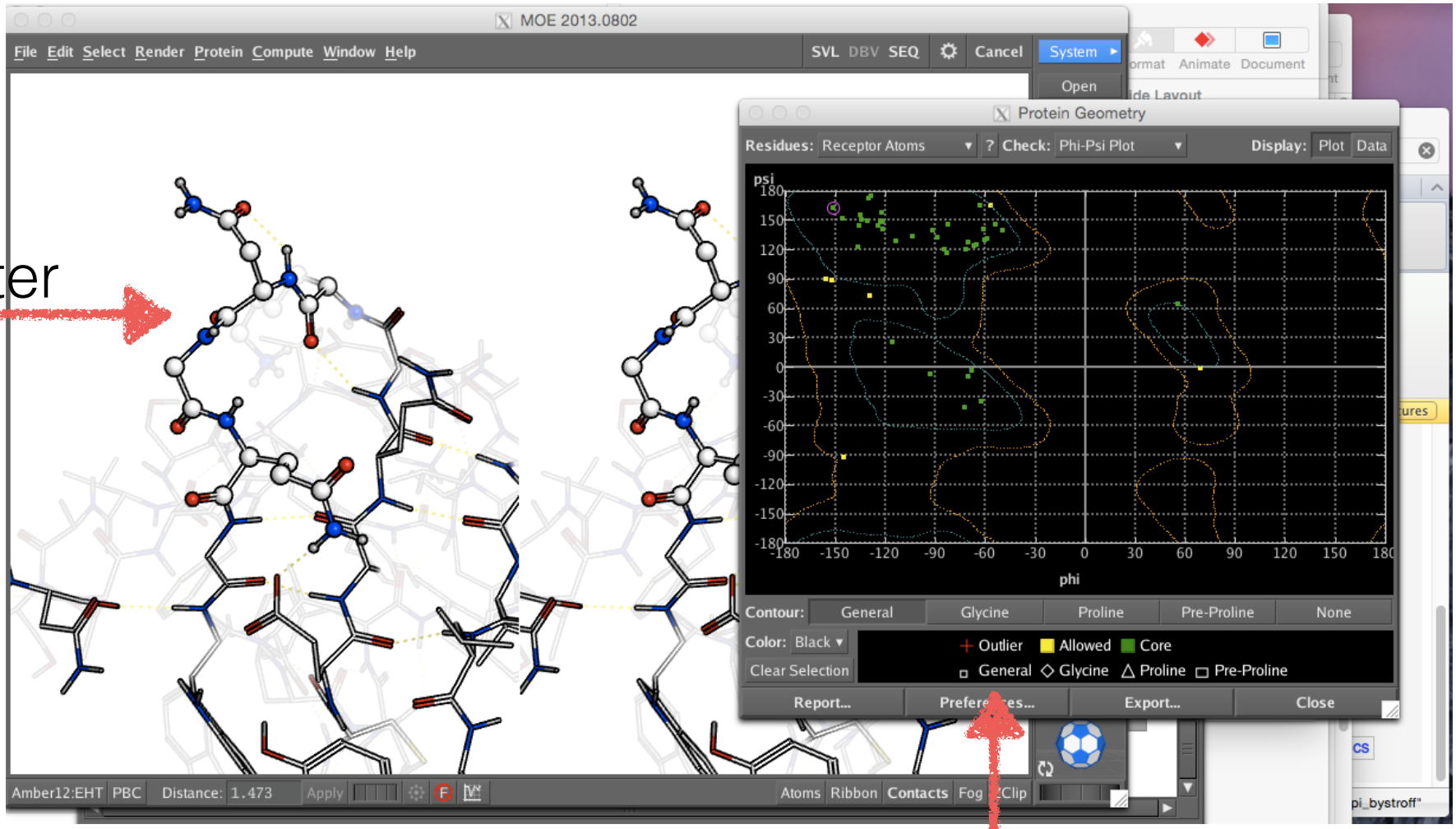
Phi-Psi outlier found



Unalign 2 or 3 residues and run Homology Model — 2nd pass.



After 2nd pass Homology Model…



New loop has no phi-psi outliers.

Validation of your model

- You can never know if the model is right.
- You can only know if the model is wrong.
- When you are "done" with a model, check:
  - H-bonding (view contacts. look for buried Ns and Os with no H-bonds)
  - Buried charges without counter-ions.
  - Excessive exposed hydrophobics (do a molecular surface and color by hydrophilicity)
  - Ramachandran outliers.
  - Buried cavities. (hydrate*, then do molecular surface and look inside)

*see future lecture on molecular dynamics!

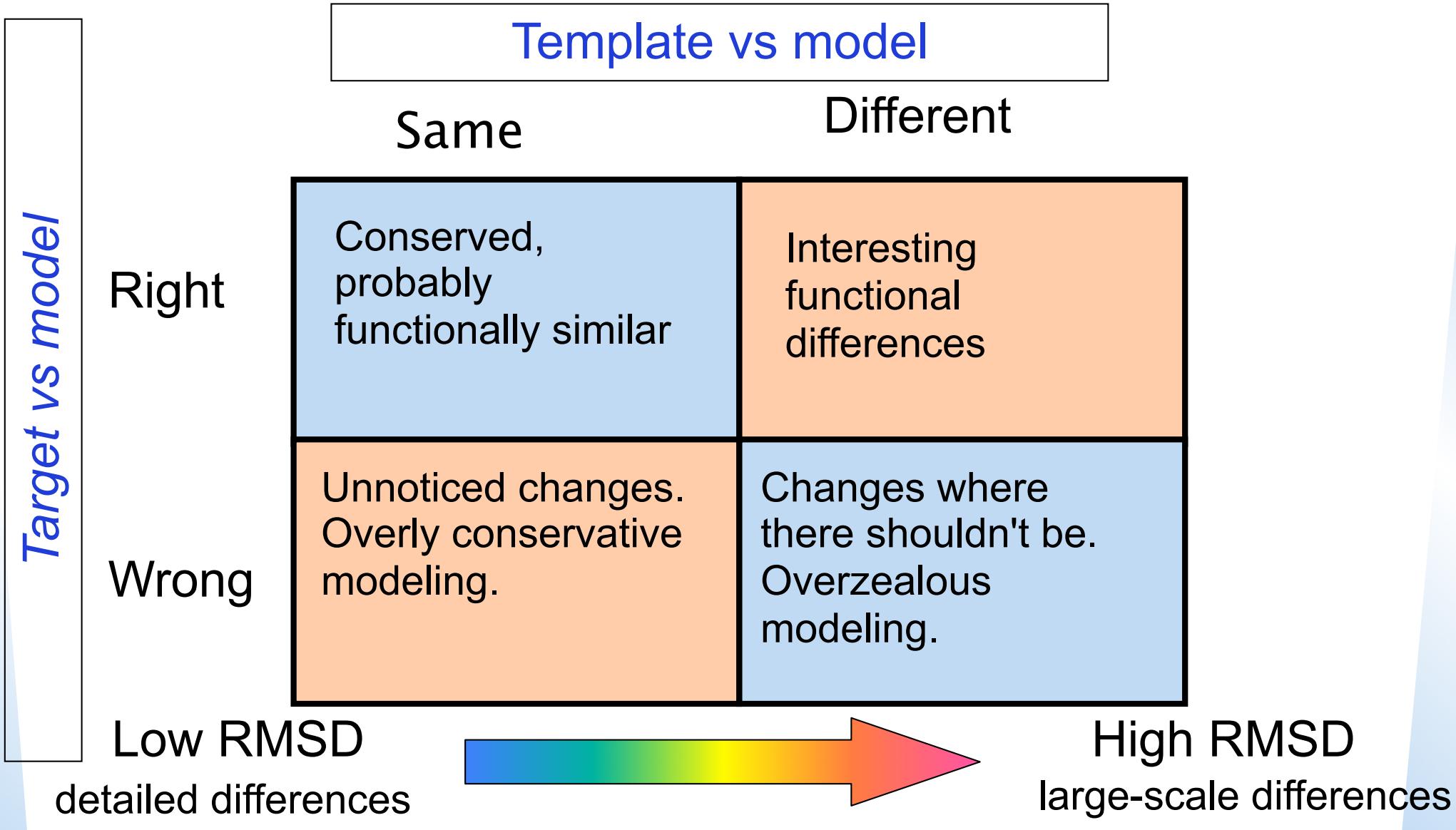
Modeling errors*

- Backbone errors
  - indel in helix — realign
  - indel in strand — realign
  - loose loops — realign, MD
  - missing H-bonds — restrain, minimize
- Sidechain errors
  - buried charges, polar sidechains — realign
  - too many exposed hydrophobics — realign
  - phi > 0 and not Gly, Asn — realign or minimize
  - (phi < -90 or phi > 0) and Pro — realign or minimize
- Voids — MD, rotamer search, minimize

*Here we are omitting obvious errors: collisions, stretched bonds, distorted planar groups, etc.

"Same/different" versus "right/wrong."

There are 2 dimensions to models: model vs template is something we **can** see. Model vs target is something we can't see, but can only **infer**.

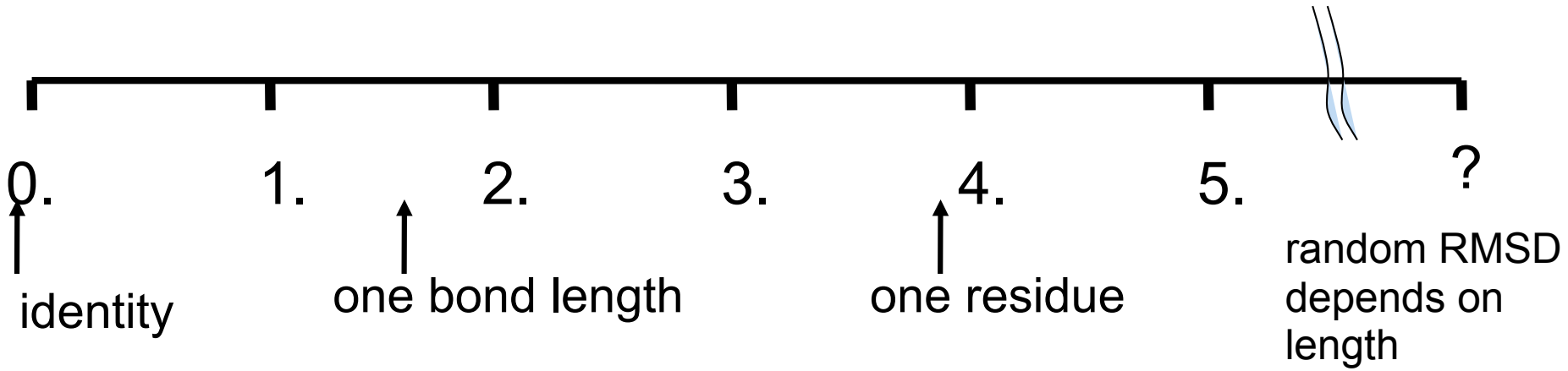


Cartesian coordinate differences: RMSD

- RMSD = root mean square deviation|

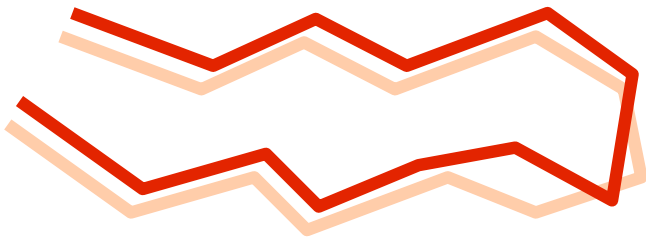
$$\sqrt{\frac{\sum_{i=1,N}\left(\vec{x}_i - \vec{y}_i\right)^2}{N}}$$

By far, the most widely used and accepted metric for structural difference.



Internal coordinate differences complement Cartesian ones

- Internal coordinates = bond distances, bond angles, torsion angles
- Deviations indicate **local** functional differences.
- MDA = maximum deviation in backbone angles
- Protein segments with mda < 120° almost always have superimposable structures.
- Superimposable structures do not always have mda < 120°.



Type-1 β-hairpin

Type-2 β-hairpin

low RMSD

high MDA

Internal coordinate differences: Distance Matrix Error

- DME = distance matrix error (average or RMS)
  Distance matrix D^x_{ij} = distance from i to j in structure x

$$\frac{\sum \left| D^x_{ij} - D^y_{ij} \right|}{\text{N choose 2}} \qquad \frac{\sum_{i<j=1,N} \left| D^x_{ij} - D^y_{ij} \right|}{N(N-1)/2}$$

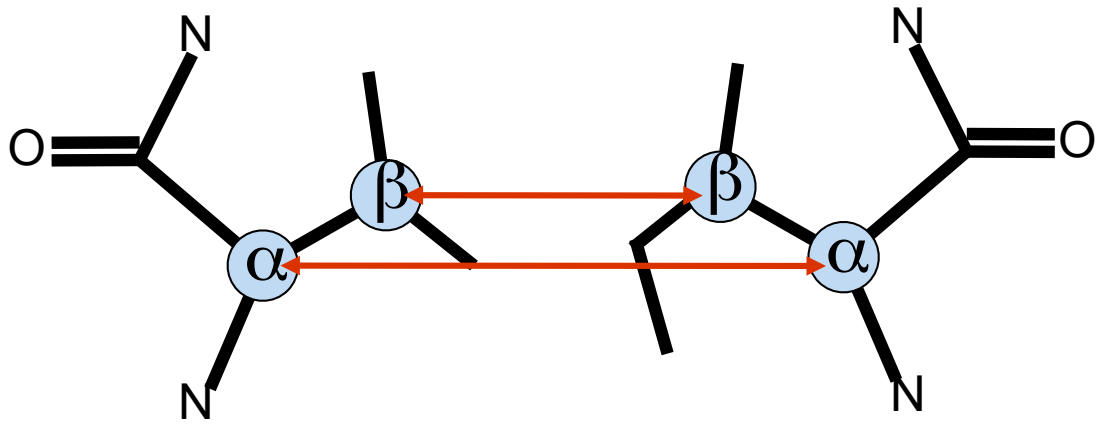
-------- or ---------

$$\sqrt{\frac{\sum \left(D^x_{ij} - D^y_{ij} \right)^2}{\text{N choose 2}}} \qquad \sqrt{\frac{\sum_{i<j=1,N} \left(D^x_{ij} - D^y_{ij} \right)^2}{N(N-1)/2}}$$

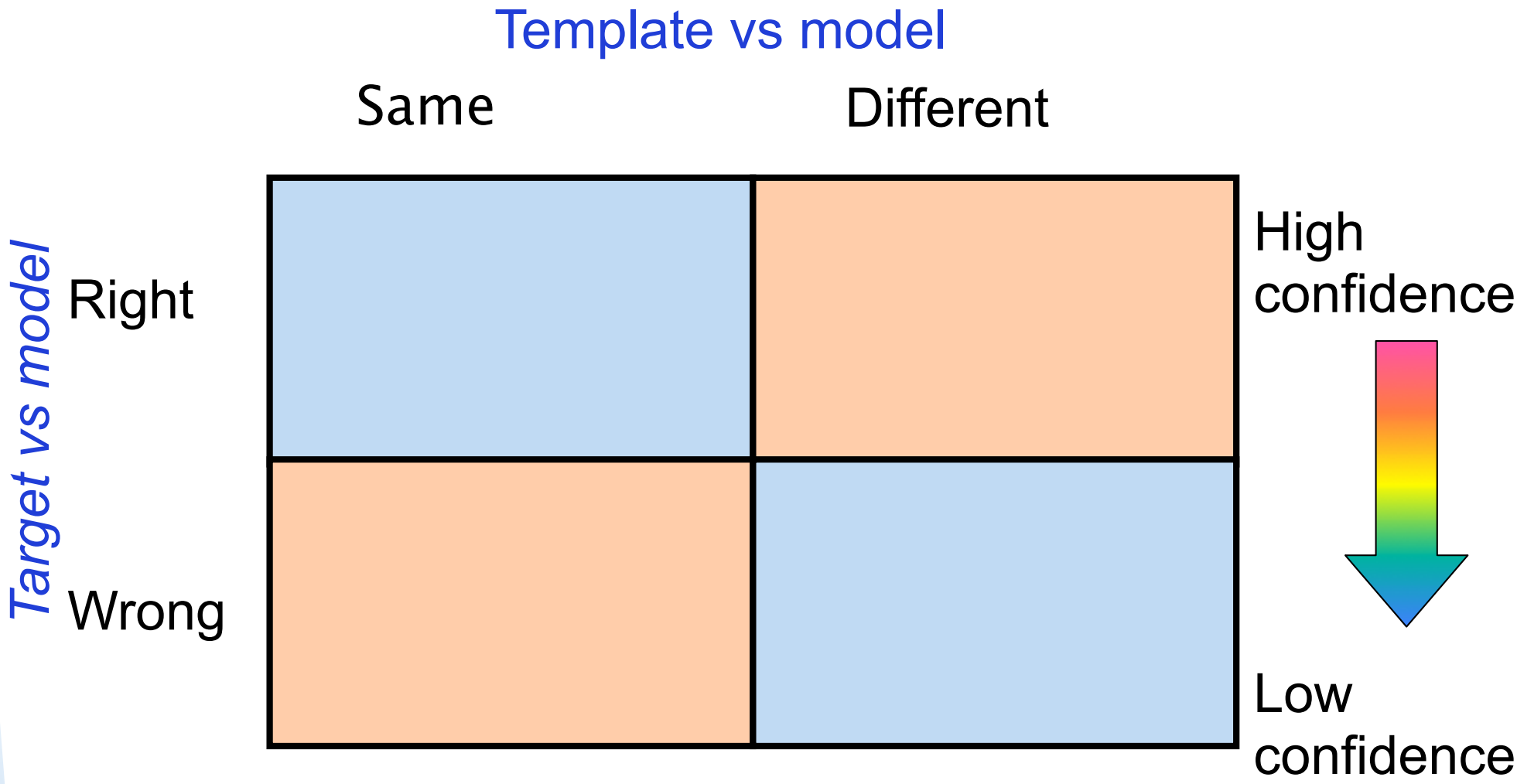
"N choose 2" = the number of pairs possible with N items = N(N-1)/2

DME, continued

- As for any difference metric, we must have an alignment first. The alignment associates D^{y}_{ij} with D^{x}_{ij}.
- D_{ij} may be measured from Cα to Cα, or from Cβ to Cβ. (In the latter case, if the residue is a Gly, then Cα is used instead.)



Confidence should measure correctness



Confidence

Confidence= the estimated probability of being right.

Physics-based confidence estimate:

Based on **modeling experience**, knowledge of **stereochemistry**, **function**, other factors, not statistics. Case specific.

Knowledge-based confidence estimate:

Based on **statistics** of known structures and repeated modeling experiments. **Empirical**, not theoretical. Not specific to one case.

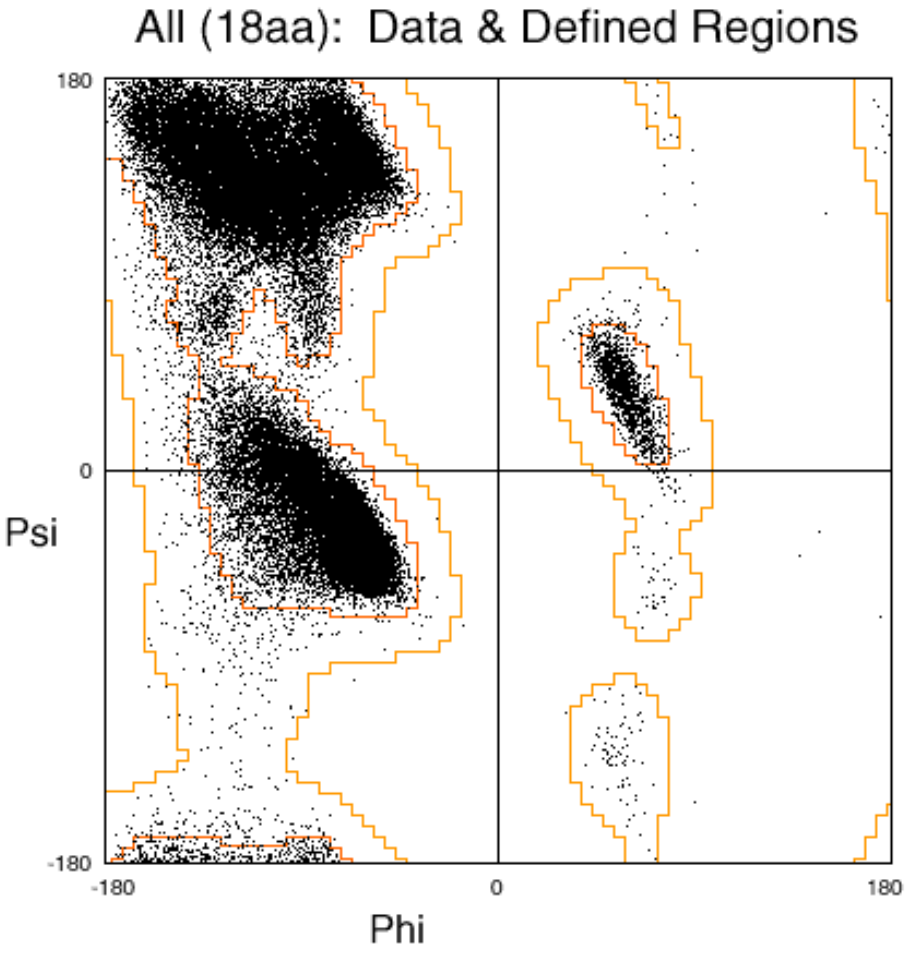
Knowledge-based statistics: Ramachandran allowed regions

- Check for other amino acids outside the allowed regions.
- If it is an outlier, is it conserved? Then it's real.

Remedies for <u>suspicious</u> outliers:

(1) energy minimize with restraint

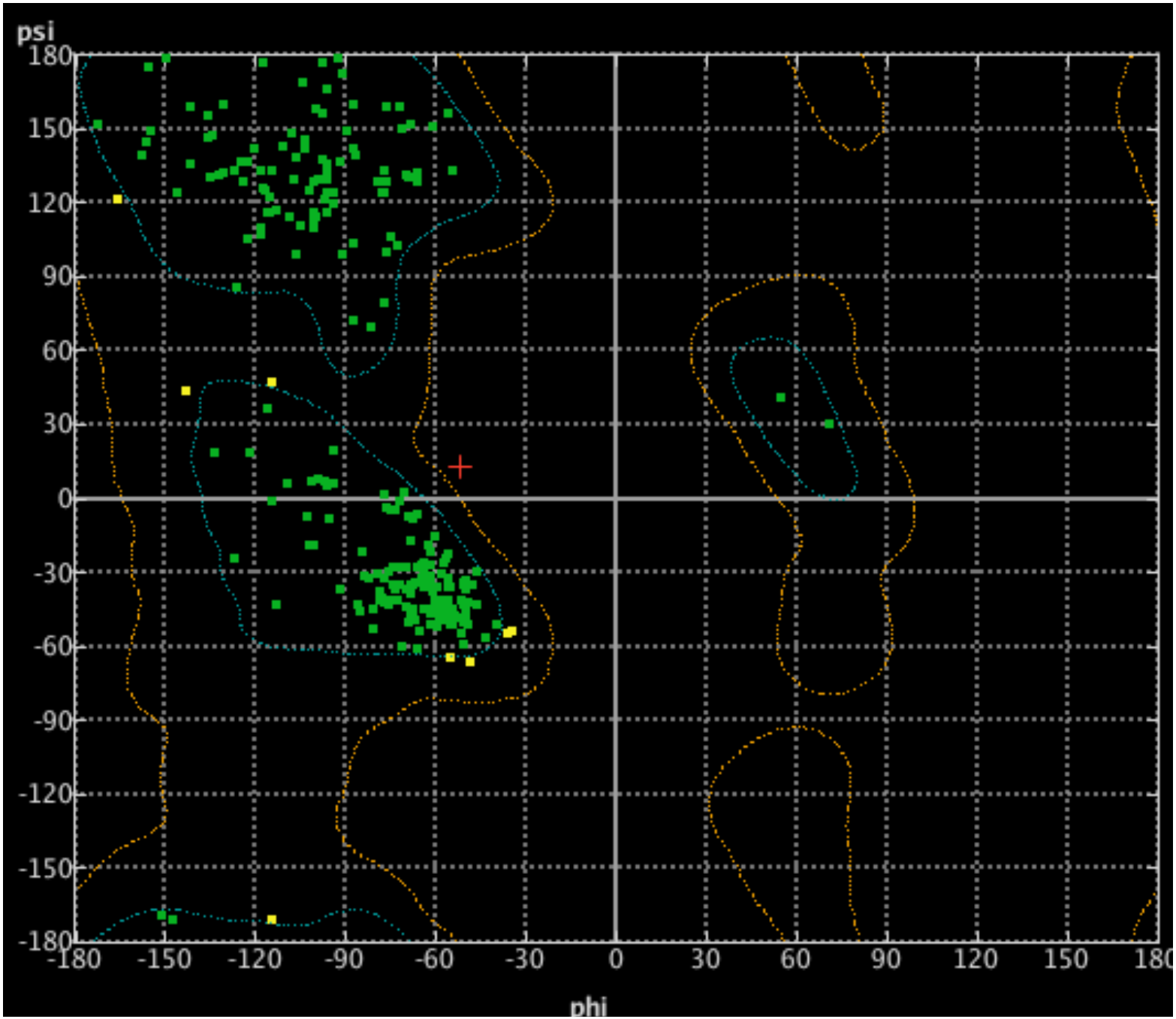
(2) Ignore it. Outliers happen.
But watch out. Too many outliers makes the whole model suspect...



Courtesy of Jane & David Richardson

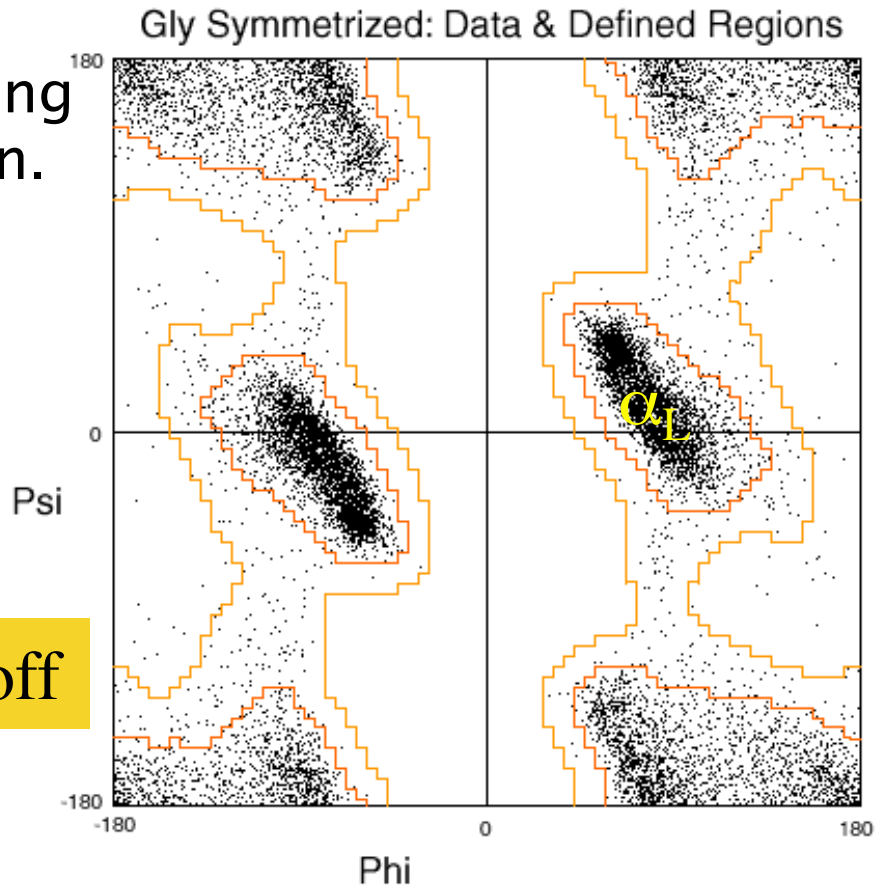
kinemage.biochem.duke.edu

Ramachandran plot: outliers should be rare



Knowledge-based confidence: positive phi angle at Glycine

- Glycines, lacking a C-beta, have a greater allowed Ramachandran region, including the "α_L", or positive phi, region.
- 2-fold symmetrized statistics for Glycine $\phi\psi$ angles show a more realistic picture of the energy landscape.



Courtesy of Jane & David Richardson

bet on α_L

bets are off

XXXXXGXXXXXG
XXXXXGXXXXXN
XXXXXGXXXXXN
XXXXXGXXXXXD
XXXXXGXXXXXG

Knowledge-based confidence: Proline phi angle

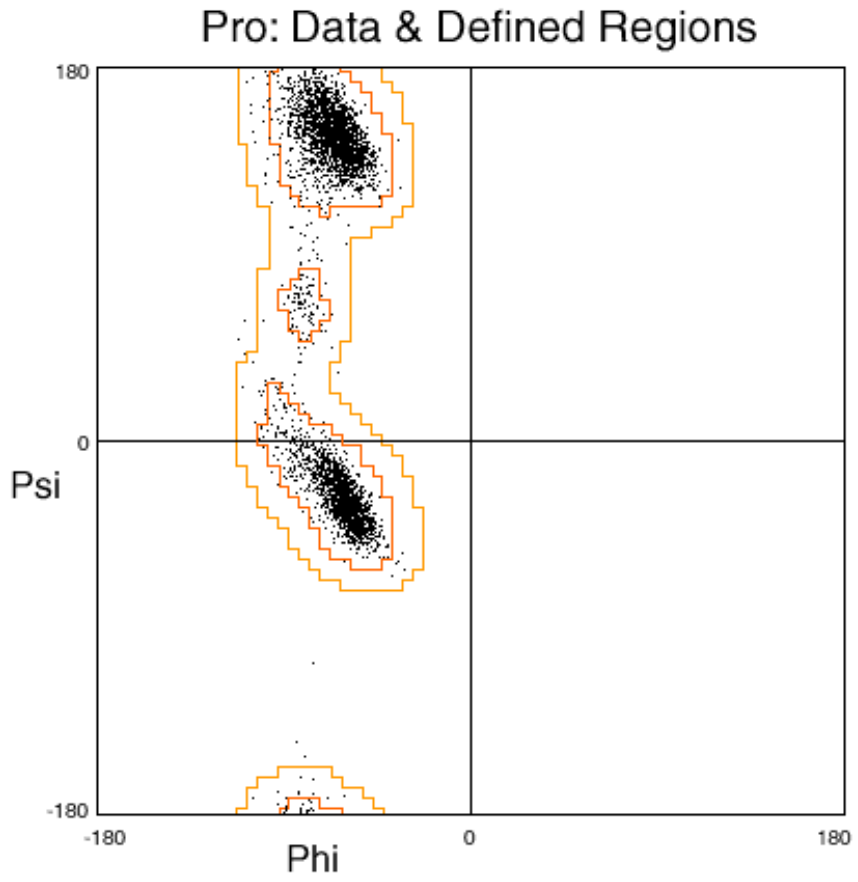
- Check for impossible phi angles at Proline positions.

If you find one, there are two possible remedies

(1) energy minimize it away

(2) re-align the Pro.

never leave it like that.

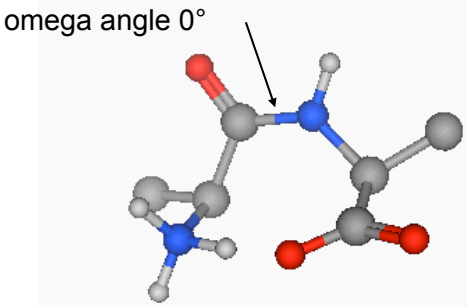


Courtesy of Jane & David Richardson

Knowledge-based confidence: cis peptide bond at X-Pro

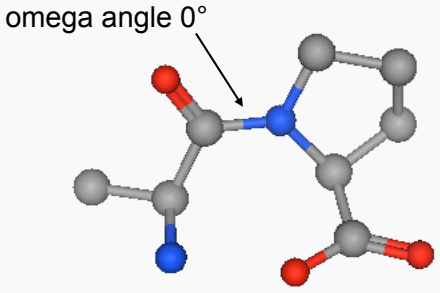
- "cis peptides" : ω (omega) torsion angle may only be 180° or 0° (because of double-bond character), but 0° is highly disfavored (and therefore rare!) unless the residue following the peptide bond is a Proline. Why is this true?
- X = the residue before Pro. X = big (F,Y,W) favors the *trans* state.

relatively problematic!

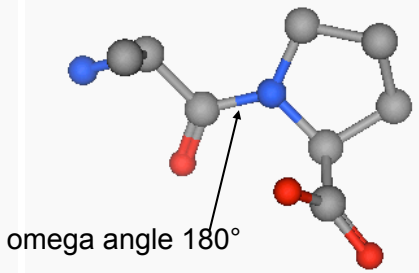


cis X-Ala

------no big energy difference-----



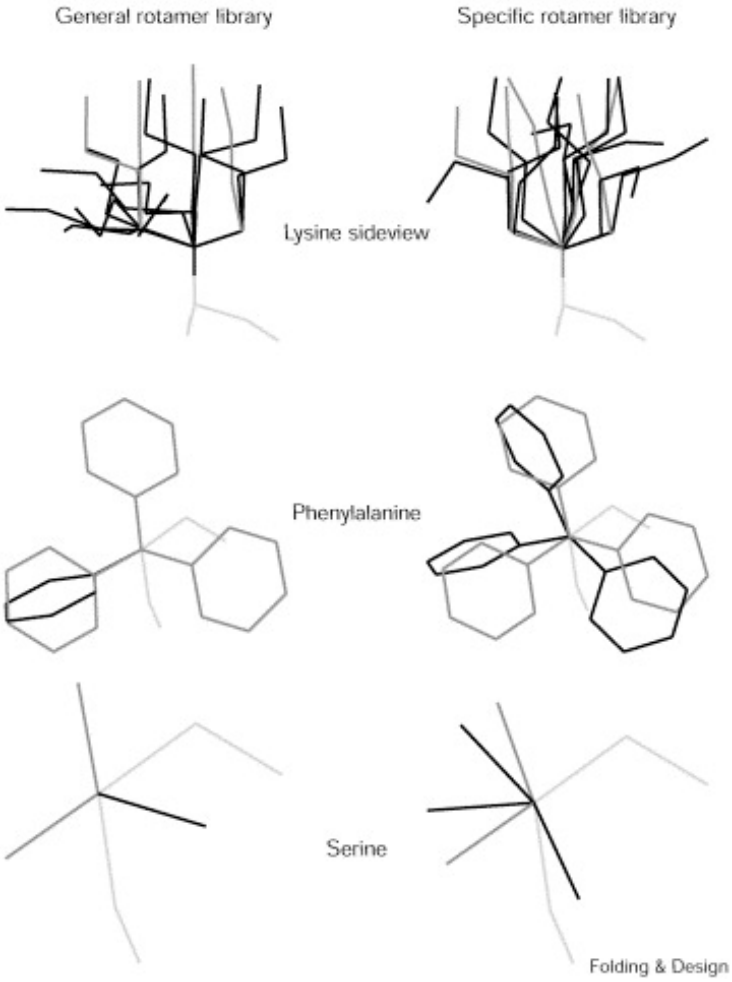
cis X-Pro



trans X-Pro

Knowledge-based statistics: Preferred rotamers

•**Rotamers** are preferred sidechain conformations, found by clustering database sidechains. •**Rotamer** sets (libraries) may be coarse grained or fine grained (pulldown menu in Rotamer explorer). •**Rotamers** have intrinsic energies, due to local interactions.



Compute | Biopolymer | Rotamer explorer

Allows modeler to test rotamer swaps.

Compute | Biopolymer | Protein geometry, rotamer

Finds side chains that need help.

Physics-based confidence: void regions

- Nature abhors a void.

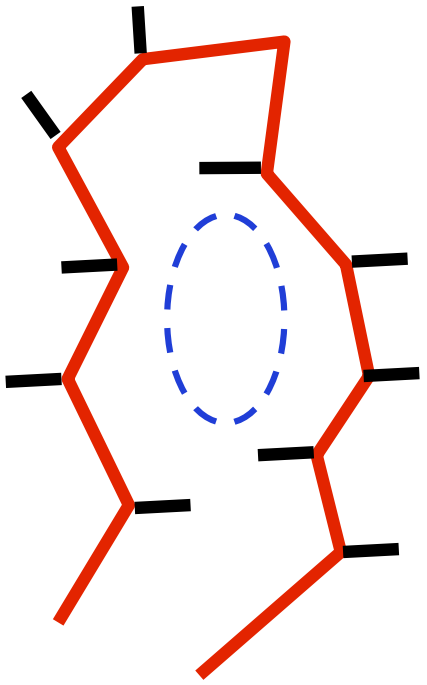
Remedies:

(1) re-pack sidechains with rotamer explorer.

(2) add waters.

(3) energy minimize with distance restraints

(4) Leave it alone. Voids may be functionally important. See (Paredes et al, BMC Bioinformatics 2011)



Physics-based confidence: buried charges

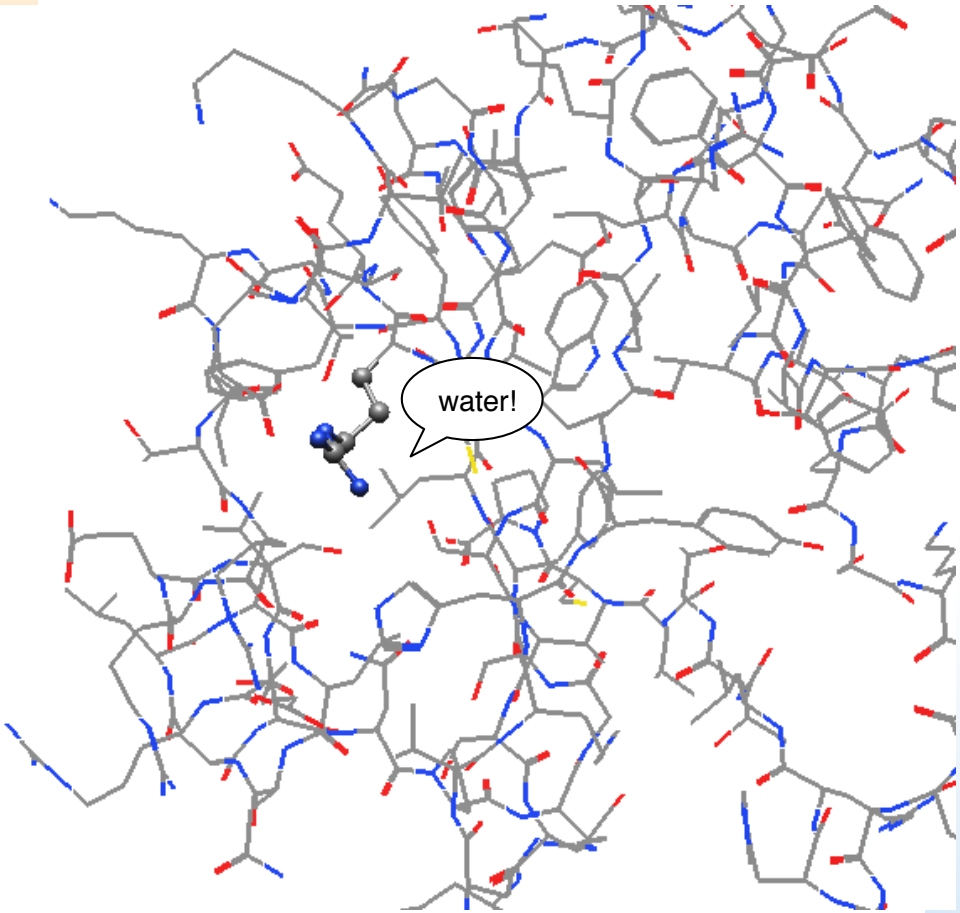
- Charges hate to be de-solvated.

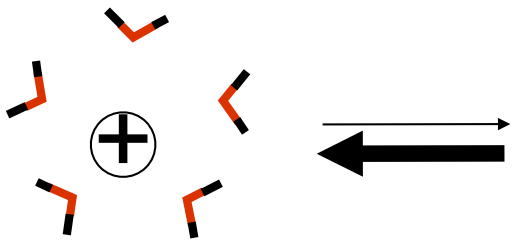
Remedies:

(1) re-pack sidechains. Find a sa
bridge.

(2) re-align. Put it on the outsid

(3) Leave it alone.





water dipoles
delocalize the charge

buried charge is like a
charge in a vacuum.

11.4 MOLProbity

guided tour

`molprobity.biochem.duke.edu`

- Automated checker for correctness of a model.

Midterm topics

- Practical portion: MOE.
  Time will be a factor. Complete as many of the assigned tasks as you can in the allotted 2 hours. If you have done the homeworks and exercises, then you can do these tasks.
- "Theory" portion will consist of multiple choice questions and problems. If a question/problem is asked during lectures, in the slides, or in homework assignments, then it could appear on the midterm.

- Pick slides for review session.
  Each student presents one slide (or more)