Molecular modeling 2019 -- Lecture 25

Model Validation

Validation of your model

- You can never ***know*** if the model is right.
- You can only ***know*** when the model is wrong.
- When you are "done" with a model, check:
  - –Bond distances, bond angles, D-amino acids, cis-peptides, clashes
  - –H-bonding, especially buried unsatisfied donors/acceptors.
  - –Buried charges without counter-ions.
  - –Excessive exposed hydrophobics
  - –Ramachandran outliers. Positive φ angle not in a glycine.
  - –Buried cavities or deep pockets.

ELIMINATE ALL REASONS TO DISBELIEVE THE MODEL.

Comparing model to template

Template vs model

Target vs model

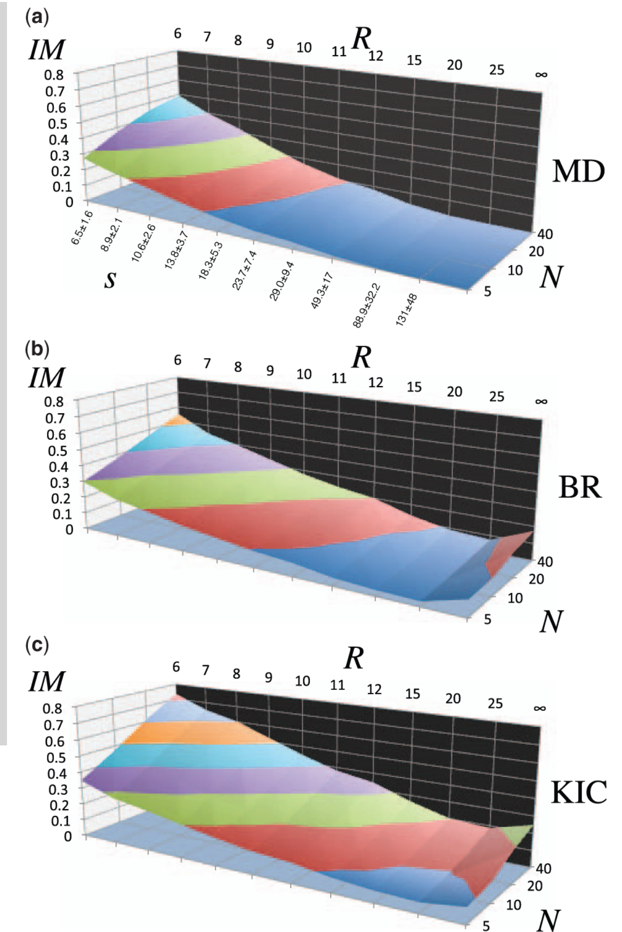
	Same	Different
Right	Conserved, functionally similar	Interesting differences
Wrong	Overly conservative modeling.	Overzealous modeling.

Low RMSD
detailed differences

High RMSD
large-scale differences

Stay close to the template!

Studies show loop searches (KIC method) , short molecular dynamics (MD) and monte carlo (backrub motions, BR) fail to sample the true backbone structure, more often make things *worse*.



Criterea for improvement (IM)
Template 1e40A (magenta tubes) and homolog target 1bglB (orange tubes) superimposed with a minimized, diversified *de novo* structure (thin gray string) based on the template. If the gray string more closely resembles target, then we say the method locally improved the model.

Fig. 5 Improvement (IM) of local substructures starting with template, using three methods. Small regions (R) can improve with many tries (N)

Bioinformatics, Volume 32, Issue 10, 15 May 2016, Pages 1454–1461, https://doi.org/10.1093/bioinformatics/btw001

The content of this slide may be subject to copyright: please see the slide notes for details.

OXFORD UNIVERSITY PRESS

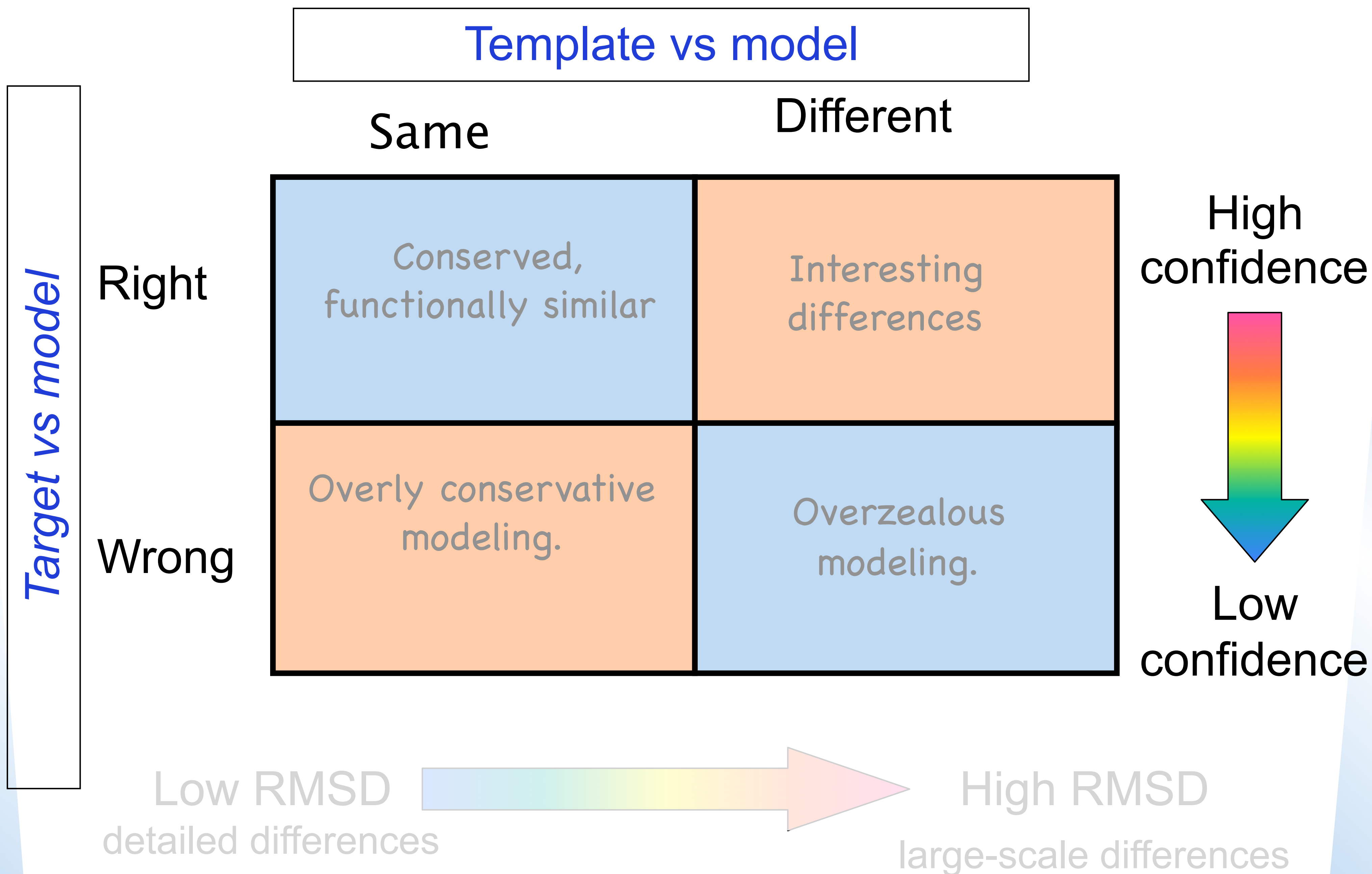
Confidence

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.

Confidence *should* measure correctness



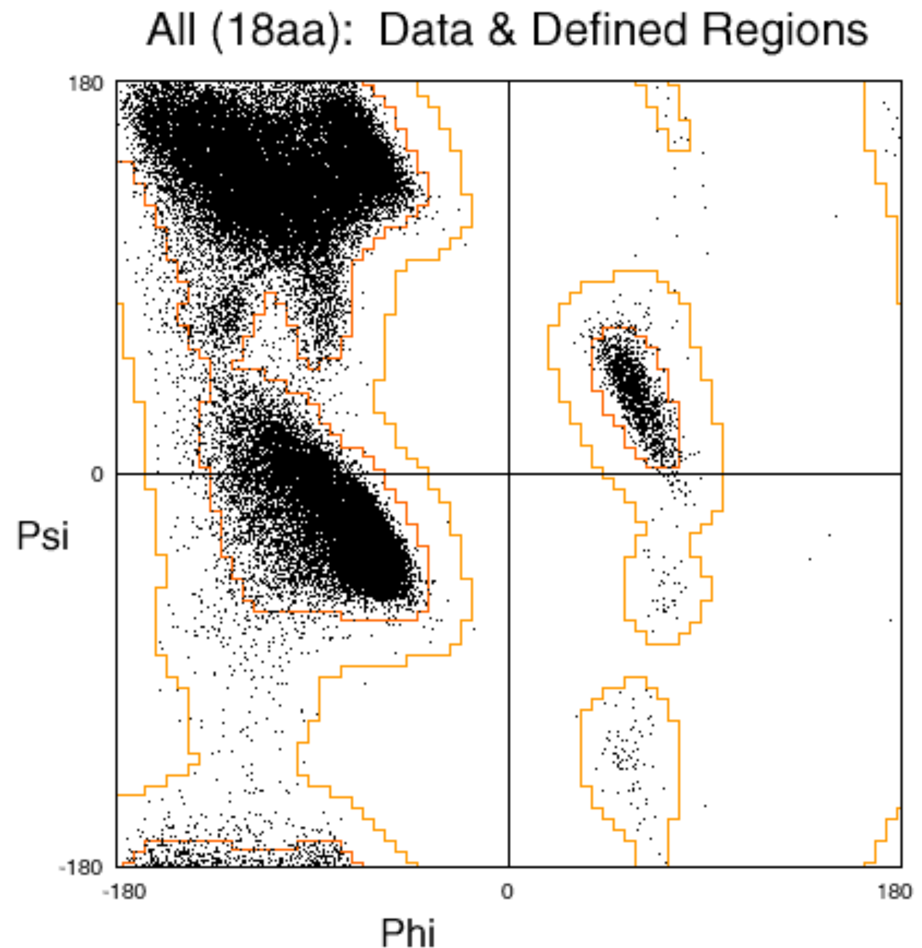
Knowledge-based statistics: Ramachandran allowed regions

- 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

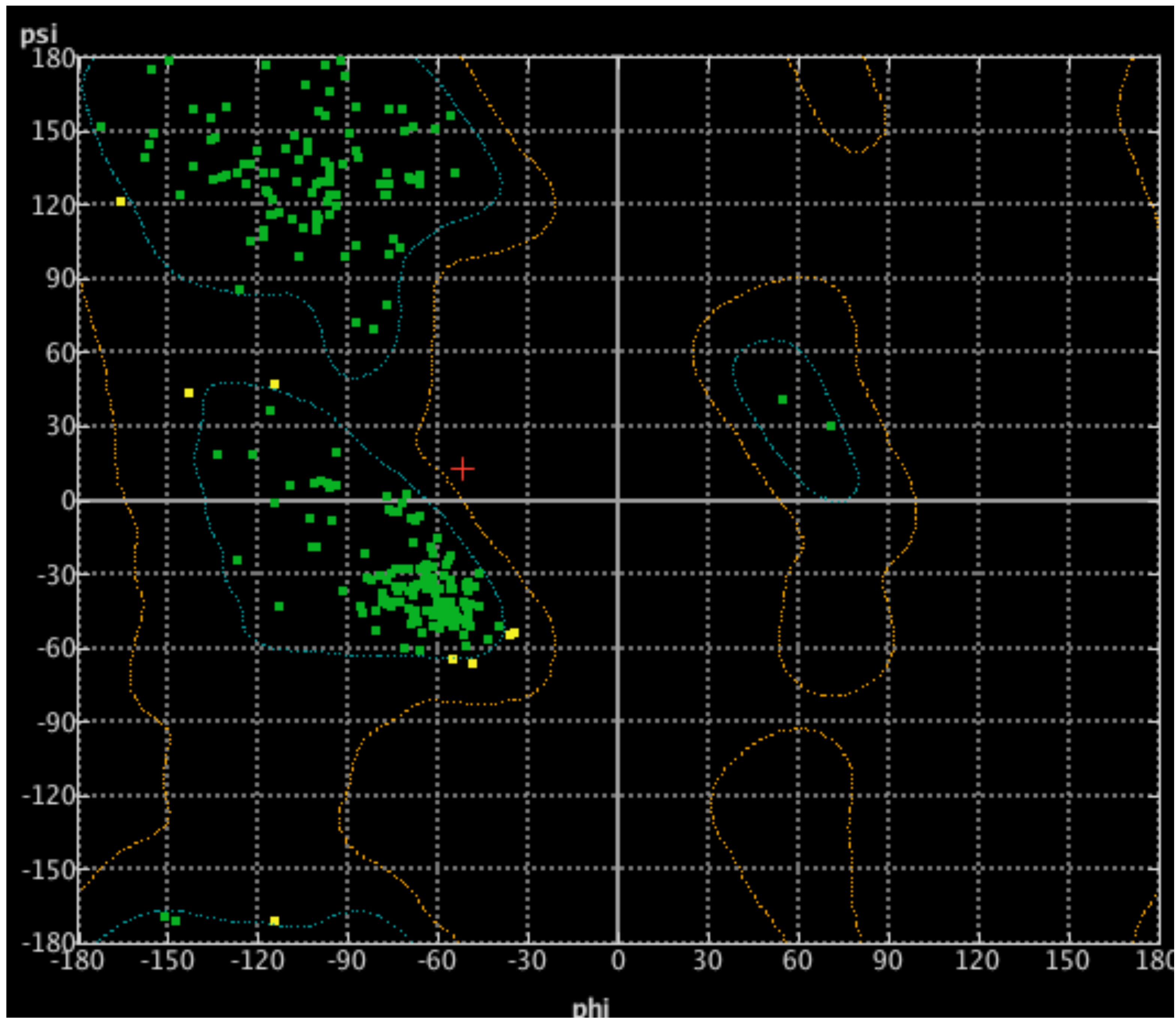
(2) Ignore it. Outliers happen.
But watch out. Too many outliers makes the whole model suspect...



Courtesy of Jane & David Richardson

kinemage.biochem.duke.edu

Ramachandran plot: outliers should be rare



Knowledge-based confidence: conserved glycines are probably

- Glycines are allowed in a wider Ramachandran region, including the "α_L" region.
- If glycine is <u>conserved</u>, you can bet it is is in one of the glycine-only zones. If not conserved, then it must remain in the standard Ramachandran zones.

bet on α_L

bets are off

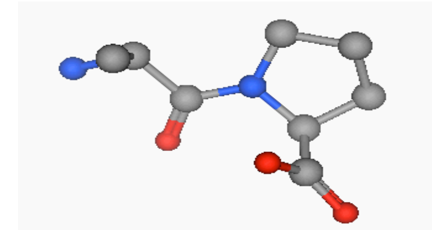
XXGXXXXXXG
XXXXXGXXXXXXN
XXXXXGXXXXXXN
XXXXXGXXXXXXD
XXXXXGXXXXXXG

Gly Symmetrized: Data & Defined Regions

180
0
-180
Psi
-180
0
180
Phi
α_L

Courtesy of Jane & David Richardson

Knowledge-based confidence: Proline phi angle always ≈-60°



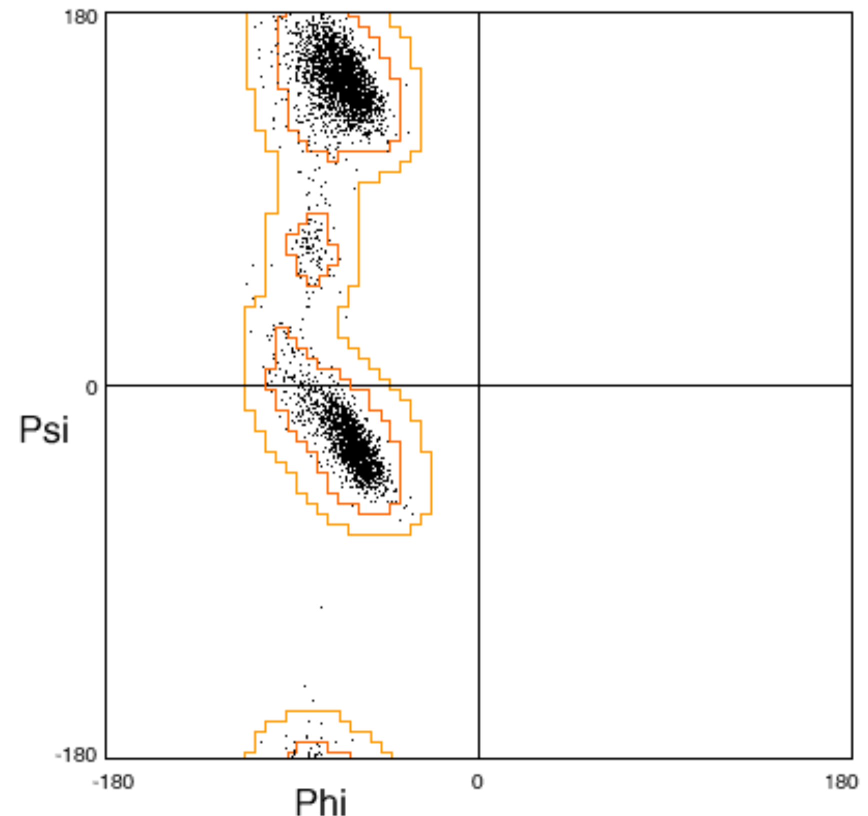
- Check for impossible phi angles at Proline positions.

If you find one, there are two possible remedies

(1) energy minimize it away

(2) re-align the Proline.

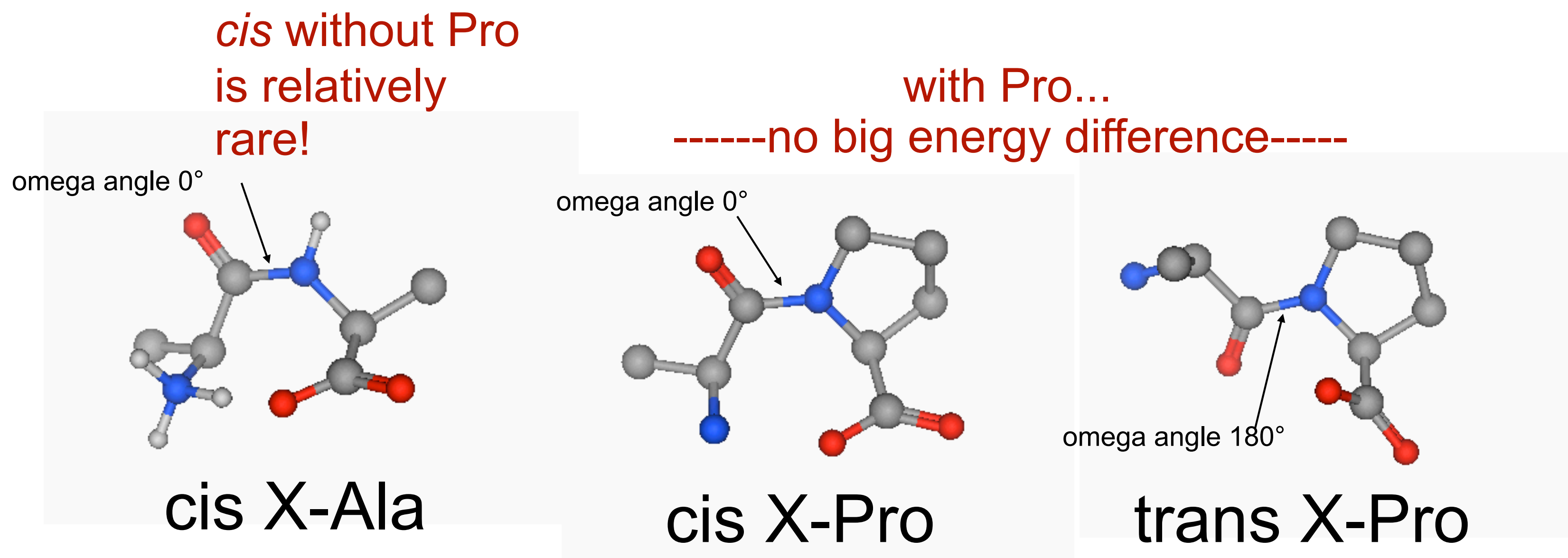
never leave it like that.



Courtesy of Jane & David Richardson

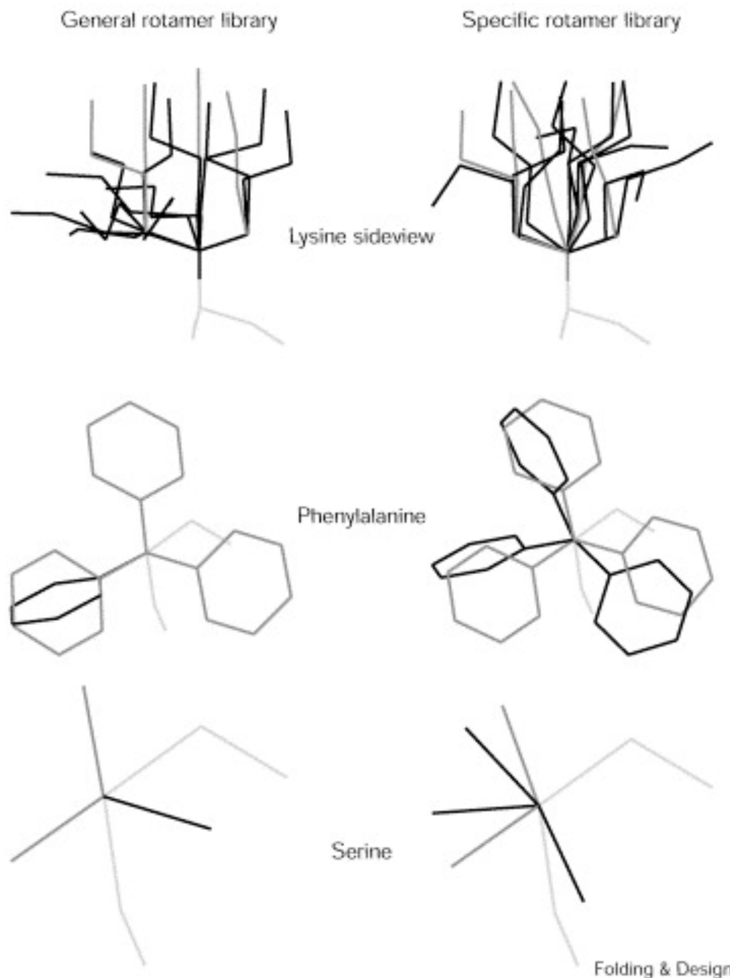
Knowledge-based confidence: cis peptide bond at X-Pro

- "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.



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.

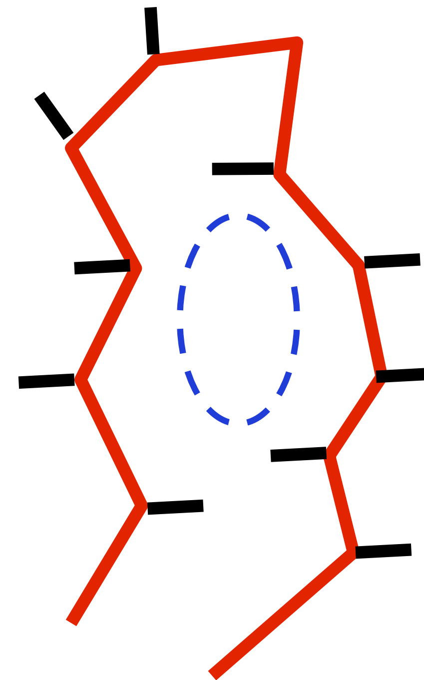
Remedies:

(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 are rare, always paired

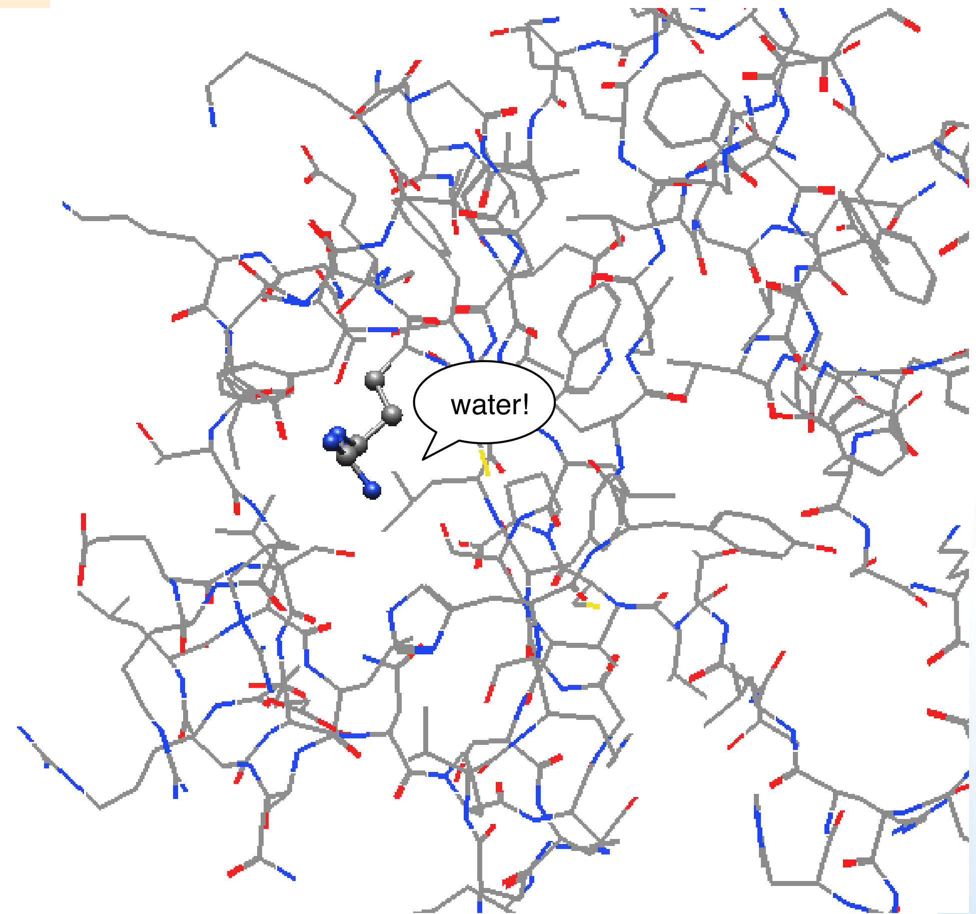
- Charges hate to be de-solvated.

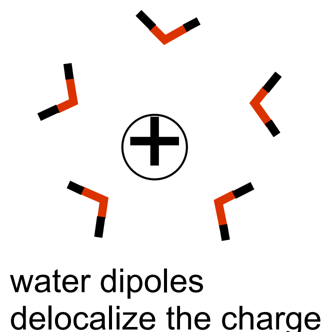
Remedies:

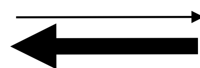
(1) re-pack sidechains. Find a sa
bridge.

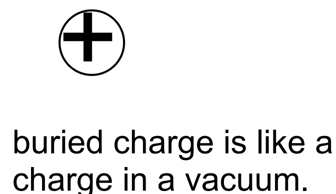
(2) re-align. Put it on the outsid

(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.

Summary

A model is as "correct" as it can be if....

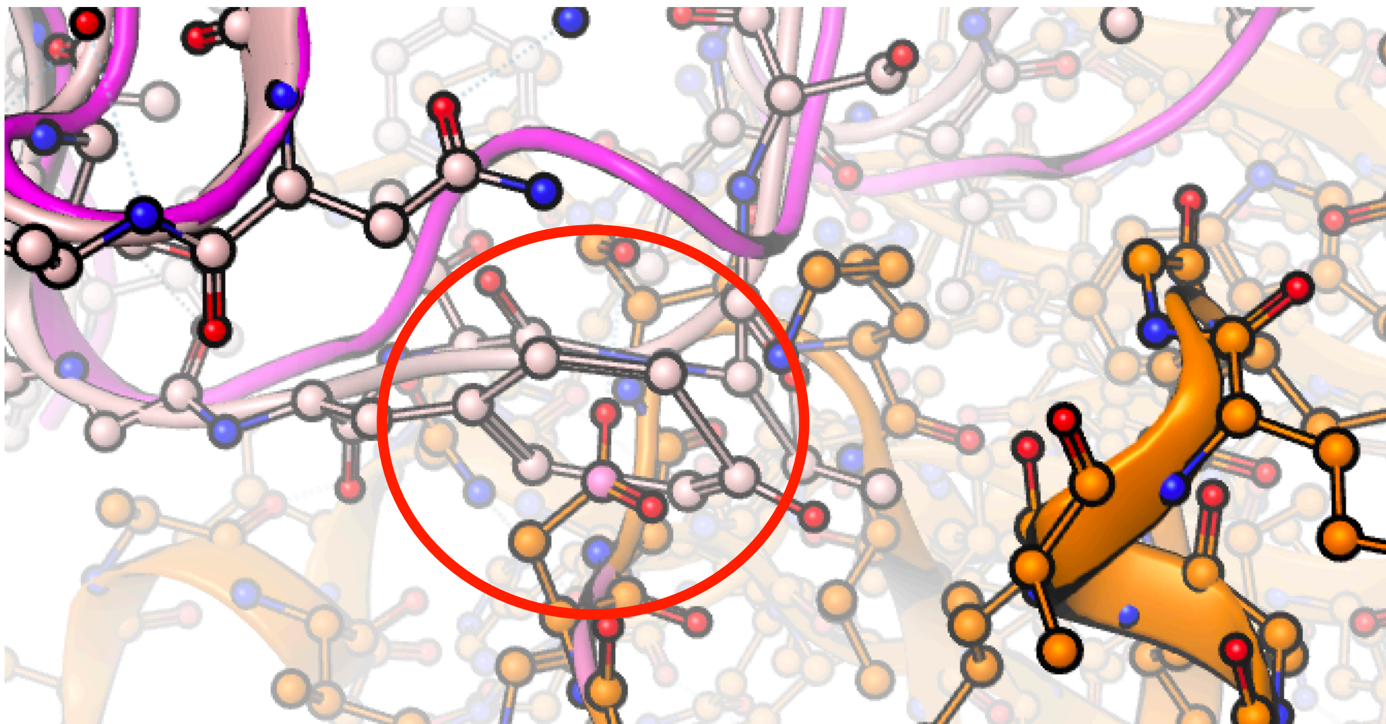
- It stays close to the template
- It breaks the fewest possible "rules." (buried H-bonds, voids, phi/psi outliers, etc.)
- Template/model differences are confidently predicted.

The philosophy of expert protein design

- Well-trained intuition is much faster than a random search.
- There are many, many, many right answers. We don't need the very best one.
- No force field is perfect anyway. We can't avoid the need for experimental confirmation.

Demo: docked too close!

- **Part 1**: Homology modeling when docking too close leads to distorted sidechains. Use Geometry panels to find distorted residues.



Demo: skewed docking

- **Part 2**: Homology modeling when docking to far leaves large space between ligand and receptor. Fix by redocking.

Demo: fine-tuning

- Part 3: Add residues to loop and terminus to fill space.

Exercise 25.1 -- start HW5

- Download **HW5_startdesign.moe** from course website. Open it.
- Select interface sidechains and ligand loop backbone atoms. **EPUSIEPF**. Turn on energy minimization (**SVL: run 'gizmin.svl'**)
- **Identify side chains to design (i.e. mutate)**
- For each designable side chain:
    - Inspect the site.
    - Decide what amino acid should be there. If ok as is, move on.
    - **Protein | Protein builder**. Change **1 Target...** to allow your selected amino acids (keep it to a few)
    - hit **Rotamer**.
    - Arrow through the options and pick the best.
    - Energy minimize (D*o not energy minimize while building rotamers! Do it after you pick one.*)
    - Select nearest neighbors and **Repack**.
    - Go on the next residue. Any order.
  - Keep going until you can't find any more improvements.
- **Unfix** (more) backbone atoms *in the vicinity*. But select the receptor backbone atoms (Continue designing until you can't find improvements. Be

New directions for protein design

- Docking + design. Doing both at the same time. (FlexPepDock)
- Receptor design -- biosensors, enzymes
- Designing for kinetic stability -- disulfides, permutation and core packing
- Vaccine design -- symmetrical clusters, loops.