Guide to MolProbity Model Validation

Presented by Christopher Williams
Validation can make a difference

Improvements driven by availability of validation tools
You can make a difference

% PDB deposits with $\geq 30$-fold too many cis-nonPro, by year

Improvements driven by community education and awareness
Validation Philosophy

- Hydrogens are half the atoms! Add them before validation, or your analysis is incomplete.

- 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!
Intervention Philosophy

• Refinement is great at details, bad at escaping local minima

• Human interventions should
  • Find the right local minimum
  • Preserve interesting features
  • Not sweat the details
MolProbity

http://molprobity.biochem.duke.edu/index.php

- Free, online structure validation server
- Also built into Phenix
- Confidential
  - Files are automatically deleted
- Open-source
  - https://github.com/rlabduke

Walkthroughs, tutorials, and usage FAQs:

- Evaluate X-ray structure: Typical steps for a published X-ray crystal structure or one still undergoing refinement.
- Evaluate NMR structure: Typical steps for a published NMR ensemble or one still undergoing refinement.
- Fix up structure: Rebuild the model to remove outliers as part of the refinement cycle.
- Work with kinemages: Create and view interactive 3-D graphics from your web browser.
- Guide to Reduce options: Learn about adding hydrogen to a structure for all-atom contact analysis.
- Guide to validation options: Choose validations appropriate to a structure.

More tutorials

Citations, science, and technical FAQs:


About hydrogen: Why have the hydrogen bondlengths changed?

Installing Java: how to make kinemage graphics work in your browser.

Download MolProbity: how can i run a private MolProbity server, or run from the command line?
Outline

For each validation
• Method
  • Briefly, how the underlying idea or math works

• Visualization
  • How outliers are visually represented in KiNG/NGL/Isolde

• Probable causes
  • Example of a common or interesting type of error
  • Not comprehensive!
All-Atom Clashes and Contacts
Add hydrogens
(phenix.reduce or MolProbity website)
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
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

Other outliers

- Clashes usually occur alongside other outliers

- Emphasize modeling errors
  - *Real* rare features are less likely to have clashes

- Can imply direction for fixups
Sidechain Rotamers
Sidechain Rotamers: Method

- Sidechain conformations are described by a series of $\chi$ (Chi) torsions

- Rotamers are statistically expected combinations of $\chi$ values

- For tetrahedral atom centers, this means staggered
  - $p +60^\circ$
  - $t 180^\circ$
  - $m -60^\circ$

- For planar atom centers, rotamers are much more continuous
  - Rotamers are named with a central value
  - e.g. $m90$ or $p-80$ for Histidine

Rotamer distribution for Isoleucine in $\chi1/\chi2$ space
Sidechain Rotamers: Visualization

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

In ISOLDE, Rotamer allowed/outliers are marked with a spiral, color coded by prior probability.
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

- 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 $\phi = C_{i-1}\text{-N-CA-C}$
- Psi $\psi = \text{N-CA-C-N}_{i+1}$
- Each residue’s $\phi/\psi$ pair is converted into cartesian coordinates and checked against contours of expected behavior
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
KiNG markup highlights an outlier residue’s CA in green, and extends to the peptide bonds on either side, along the CA-CA-trace.

ISOLDE markup places a ball at each CA, color-coded by Ramachandran favorability.
CaBLAM: Method

- 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
Rama/CaBLAM: Probable causes

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

• $|Z\text{-score}| \leq 2$ indicates a realistic distribution
• $|Z\text{-score}| > 3$ indicates a highly unrealistic distribution
Ramachandran: Probable causes

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 Z-score -4.26 ± 0.10
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
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

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
• Derived from Karplus et al.
  • https://doi.org/10.1107/S2059798315022408
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

C-N peptide bond distances are systematically shortened

**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.

OmegaFold prediction for p81313, as of Sept 2022
Localized

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

2gwe, mostly 1.3Å

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

These are red in Coot!
Cis Peptides: Visualization (ISOLDE)

- *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 \textbf{seems} to fit better into limited density
- A conformation this rare requires more justification than a marginally better fit
- Flip it to \textit{trans} unless density, chemistry, homology, or another source gives you clear support
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
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
RNA Errors: Probable Causes

- 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

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

Density lacks details between these major positions
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 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.
Useful links

• For the quick-and-dirty webpage version of this material:
  • http://molprobity.biochem.duke.edu/help/validation_options/validation_options.html
  • This also includes links to many of the relevant publications

• I deliberately skipped over structure-level statistics, but if you want to see the target values for Ramachandran Favored, CaBLAM Outliers, etc:
  • http://molprobity.biochem.duke.edu/help/validation_options/summary_table_guide.html
Bonus Content

Here are a few more examples of interesting model errors associated with certain validations.

These didn’t fit in the main presentation, but you should still get to see them.
AlphaFold validation

phenix.barbed_wire_analysis output.type=kin
(under development)
Validation tool

- Predictive (blue)
- Unpacked high pLDDT (gray)
- Near-predictive (green)
- Unpacked possible (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)
UnDowser
UnDowser: Method

- Undowser is a tool for finding incorrect waters
- Use all-atom contact analysis to find waters with steric clashes
- Identify probable substitutions for each problem water
  - Ions
  - Ligands
  - Sidechain alternates
  - Nothing!
UnDowser: Visualization

MolProbity has a dedicated chart for water analysis
- Each clashing water is listed
  - Colored by severity
  - Possible causes marked in table
- Recently added to Phenix commandline
- Coming soon to GUI

<table>
<thead>
<tr>
<th>Water ID</th>
<th>Clashes with</th>
<th>Water B</th>
<th>Contact B</th>
<th>Clash Severity</th>
<th>Clash with Polar May be ion</th>
<th>Clash with non-polar Unmodeled alt or acute</th>
<th>Clash with water Occ &lt;1 or ligand</th>
<th>Clash with atom Add or rename alt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HG3 of A: S1 :GLU</td>
<td>26.53</td>
<td>26.06</td>
<td>0.503</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: 187:HOH:</td>
<td>HAt2 of A: T10 :GLY</td>
<td>23.36</td>
<td>18.74</td>
<td>0.736</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HG3 of A: T10 :GLYS</td>
<td>24.10</td>
<td>20.04</td>
<td>0.501</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: 180:HOH:</td>
<td>HAt2 of A: 48 :LYS</td>
<td>24.10</td>
<td>20.04</td>
<td>0.425</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CE of A: 48 :LYS</td>
<td>24.10</td>
<td>20.04</td>
<td>0.425</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: 114:HOH:</td>
<td>O of A: 132 :HOH</td>
<td>27.11</td>
<td>26.13</td>
<td>0.651</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: 122:HOH:</td>
<td>O of A: 134 :HOH</td>
<td>26.13</td>
<td>27.11</td>
<td>0.651</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: 80:HOH:</td>
<td>HG3 of A: 59 :ASP</td>
<td>22.27</td>
<td>24.16</td>
<td>0.426</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SUMMARY: 6 waters out of 58 have clashes (10.34%)
UnDowser: Probable Causes

- Water problems are highly varied
  - Fit into other ligand/ion density
  - Incorrect occupancy/alternate
  - Shouldn’t be there at all

- For details, see [https://doi.org/10.1002/pro.3786](https://doi.org/10.1002/pro.3786) and [https://phenix-online.org/phenixwebsite_static/mainsite/files/newsletter/CCN_2019_07.pdf#page=2](https://phenix-online.org/phenixwebsite_static/mainsite/files/newsletter/CCN_2019_07.pdf#page=2)
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

2r6f 3.2Å

COs misoriented
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.

Target 5, BMV
10 models at 3.8Å
More Visualizations
CaBLAM: Visualization

- Colored bars are drawn along the dihedral relationship between peptide planes
  - Purple for disfavored
    - Matters in helix/sheet, not in loops
  - Pink for full outlier
    - Matters everywhere

- Colored wheels show CaBLAM evaluation if a peptide plane were rotated
RNA Ribose Puckers: Visualization

- A magenta cross is drawn for each incorrect ribose pucker
  - Long end of cross points along glycosidic bond vector
  - Cross is connected to 3’phosphate by the perpendicular distance line
Chiral Volume Outliers
(Very rare unless something is weird)
Chiral Volume Outliers: Method

- Tetrahedral atoms with 4 distinct substituents are chiral

- Do a little light vector math to find the volume enclosed by the chiral center and its three heaviest children
  - Magnitude of volume indicates how tetrahedral the bonding is
  - Sign of volume indicates handedness (L vs D)
Chiral Volume Outliers: Visualization and Causes

- True handedness swaps
  - D-amino acids with L names
  - L-amino acids with D names

- Squished or flattened geometry errors

- Atom naming errors
  - FeS clusters
  - Swapping CD1 and CD2 names in Leu
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.

His flips change protonation, H-bonds, & even charge
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
**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