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

1qw9 - 1.2 Å

Trp B 170

* courtesy of Dale Tronrud
at least one person should look at the map...

1qw9 - 1.2 Å

rotamer outlier

Trp B 170
All-atom contacts, clashscore

Ramachandran criteria
Sidechain rotamers
Geometry
RNA bb

Crystallographic: \( R_{\text{free}} \), electron density fit
All-atom contacts, clashscore

Ramachandran criteria
Sidechain rotamers
Geometry
RNA bb

Crystallographic: $R_{\text{free}}$, electron density fit
Tyr 13 H’s in Fo-Fc map
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

Asn A 165

Misfit

Correct

Sulfate Binding Protein (1SBP)
All-atom contacts, clashscore

Ramachandran criteria
Sidechain rotamers
Geometry
RNA bb

Crystallographic: $R_{\text{free}}$, electron density fit
$\phi, \psi$ Distribution for Procheck

$\phi, \psi$ Distribution for MolProbity

$\sim 100,000$ residues, entire PDB as of 1992

$\sim 100,000$ residues, Top500 structures in 2003

Top500

Top8000

98% = Favored
99.95% = Allowed
Otherwise outlier

*courtesy of Vincent Chen and Daniel Keedy*
General

Glycine

Isoleucine/Valine

Pre-Proline

Trans-Proline

Cis-Proline

*courtesy of Vincent Chen and Daniel Keedy*
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: $R_{\text{free}}$, 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: $R_{\text{free}}$, electron density fit
Validation: basic recommendations

• The MolProbity server suggests these cutoffs:

  \text{clashscore} < 10

  \text{Ramachandran outliers} \leq 0.2\%

  \text{Ramachandran favored} \geq 98\%

  \text{Rotamer outliers} < 1\%

  \text{C-beta deviations} = 0

  \text{Overall MolProbity score} \leq d_{\text{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

• If you are running MR, make sure the starting model is as good as possible

• 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”

• There is no objective, absolute set of criteria for this!

• 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
MolProbity 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 pkacompare_refine_3.pdb:

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

No missing non-hydrogen atoms detected.
basic geometry

Number of restraints: 3038
RMS(deivation): 0.022
Max. deviation: 0.163
Number of outliers > 4sigma: 4

List of outliers (sorted by deviation):

<table>
<thead>
<tr>
<th>Atom 1</th>
<th>Atom 2</th>
<th>Ideal value</th>
<th>Model value</th>
<th>Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>P TPO A 197</td>
<td>O3P TPO A 197</td>
<td>1.610</td>
<td>1.482</td>
<td>6.4</td>
</tr>
<tr>
<td>P TPO A 197</td>
<td>O2P TPO A 197</td>
<td>1.610</td>
<td>1.509</td>
<td>5.1</td>
</tr>
<tr>
<td>C7 LL2 A 351</td>
<td>N10 LL2 A 351</td>
<td>1.430</td>
<td>1.335</td>
<td>4.8</td>
</tr>
<tr>
<td>C9 ILE A 163</td>
<td>CD1 ILE A 163</td>
<td>1.513</td>
<td>1.676</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Number of restraints: 4101
RMS(deivation): 1.850
Max. deviation: 16.301
Number of outliers > 4sigma: 6

List of outliers (sorted by deviation):

<table>
<thead>
<tr>
<th>Atoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idle</td>
</tr>
</tbody>
</table>
Ramachandran outliers
Ramachandran outliers

Top5000

Top8000
rotamer outliers

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-branching Chi angles in a residue.

Rotamer outliers:

<table>
<thead>
<tr>
<th>Chain</th>
<th>Residue</th>
<th>Score</th>
<th>Chi1</th>
<th>Chi2</th>
<th>Chi3</th>
<th>Chi4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>MET 63</td>
<td>0.44</td>
<td>274.5</td>
<td>116.6</td>
<td>58.9</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>LEU 74</td>
<td>0.84</td>
<td>297.8</td>
<td>330.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>LEU 150</td>
<td>0.00</td>
<td>273.9</td>
<td>12.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>LEU 162</td>
<td>0.00</td>
<td>263.9</td>
<td>6.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>VAL 191</td>
<td>0.16</td>
<td>221.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>LEU 198</td>
<td>0.00</td>
<td>199.0</td>
<td>217.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>LYS 217</td>
<td>0.99</td>
<td>306.6</td>
<td>158.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>LEU 268</td>
<td>0.00</td>
<td>216.0</td>
<td>216.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>LEU 269</td>
<td>0.00</td>
<td>258.0</td>
<td>11.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>LEU 272</td>
<td>0.00</td>
<td>257.5</td>
<td>25.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>LYS 285</td>
<td>0.00</td>
<td>66.7</td>
<td>287.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>LYS 295</td>
<td>0.00</td>
<td>305.0</td>
<td>53.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>LYS 309</td>
<td>0.01</td>
<td>39.0</td>
<td>191.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>GLU 311</td>
<td>0.02</td>
<td>53.2</td>
<td>131.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>LYS 317</td>
<td>0.00</td>
<td>343.7</td>
<td>64.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>GLU 331</td>
<td>0.62</td>
<td>40.4</td>
<td>150.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>LYS 345</td>
<td>0.00</td>
<td>30.2</td>
<td>263.2</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
**Cβ deviations**

<table>
<thead>
<tr>
<th>Chain</th>
<th>Residue</th>
<th>Altloc</th>
<th>Deviation</th>
<th>Apple</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>PHE 54</td>
<td>0.357</td>
<td>-124.13</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>GLU 140</td>
<td>0.306</td>
<td>-105.19</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>ILE 163</td>
<td>0.252</td>
<td>106.89</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>GLU 208</td>
<td>0.306</td>
<td>99.47</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>VAL 251</td>
<td>0.255</td>
<td>-121.68</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>LYS 309</td>
<td>0.288</td>
<td>-96.30</td>
<td></td>
</tr>
</tbody>
</table>

**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.
steric clashes

<table>
<thead>
<tr>
<th>Atom 1</th>
<th>Atom 2</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 184</td>
<td>A 546</td>
<td>1.209</td>
</tr>
<tr>
<td>A 39</td>
<td>A 41</td>
<td>0.907</td>
</tr>
<tr>
<td>A 317</td>
<td>A 317</td>
<td>0.901</td>
</tr>
<tr>
<td>A 135</td>
<td>A 317</td>
<td>0.849</td>
</tr>
<tr>
<td>A 317</td>
<td>A 317</td>
<td>0.831</td>
</tr>
<tr>
<td>A 295</td>
<td>A 295</td>
<td>0.796</td>
</tr>
<tr>
<td>A 295</td>
<td>A 295</td>
<td>0.787</td>
</tr>
<tr>
<td>A 39</td>
<td>A 41</td>
<td>0.760</td>
</tr>
<tr>
<td>A 295</td>
<td>A 295</td>
<td>0.733</td>
</tr>
<tr>
<td>A 91</td>
<td>A 353</td>
<td>0.725</td>
</tr>
<tr>
<td>A 268</td>
<td>A 272</td>
<td>0.722</td>
</tr>
<tr>
<td>A 177</td>
<td>A 554</td>
<td>0.712</td>
</tr>
<tr>
<td>A 135</td>
<td>A 585</td>
<td>0.712</td>
</tr>
<tr>
<td>A 18</td>
<td>A 19</td>
<td>0.694</td>
</tr>
<tr>
<td>A 21</td>
<td>A 21</td>
<td>0.680</td>
</tr>
<tr>
<td>A 275</td>
<td>A 577</td>
<td>0.673</td>
</tr>
<tr>
<td>A 275</td>
<td>A 577</td>
<td>0.670</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Atom ID</th>
<th>Isotropic B-factor</th>
<th>Occupancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>OE1 GLU A 334</td>
<td>107.58</td>
<td>1.00</td>
</tr>
<tr>
<td>CD GLU A 334</td>
<td>101.13</td>
<td>1.00</td>
</tr>
<tr>
<td>OE2 GLU A 333</td>
<td>92.83</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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

Coping with model bias

- There are many methods to reduce model bias
  - likelihood-weighted $\sigma_A$-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
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

To thoroughly avoid phase bias, simulated annealing or rebuilding is strongly recommended.
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