Model Validation for Cryo-EM

SLAC Cryo-EM School
July 2019

Paul Adams
Lawrence Berkeley Laboratory and
Department of Bioengineering, UC Berkeley
Validation is Needed

- **Rotamer outliers:**
  - 30% of models have more than 10% outliers
  - 10% of models have more than 20% outliers

- **Ramachandran plot outliers:**
  - 90% of models have more than 5% outliers
  - 35% of models have more than 10% outliers

Cryo-EM structures 2015 and earlier
Accuracy & Precision

(a) Low accuracy
Low precision

(b) Low accuracy
High precision

(c) High accuracy
Low precision

(d) High accuracy
High precision

From http://extensionengine.com, by Furqan Nazeeri
Errors

• Random errors (noise)
  • Typically normally distributed
  • Can be reduced by increasing the number of observations
  • Affect the precision

• Systematic errors (bias)
  • Could arise from a poor experimental design or lack of understanding of the system being studied
  • Are reproducibly biased
  • Affect the accuracy

• Gross errors
  • Incorrect assumptions have been made or serious mistakes undetected
  • May be detectable as outliers compared to prior knowledge
Mistakes Still Happen

Retraction

WE WISH TO RETRACT OUR RESEARCH ARTICLE “STRUCTURE OF MsbA from E. coli: A homolog of the multidrug resistance ATP binding cassette (ABC) transporters” and both of our Reports “Structure of the ABC transporter MsbA in complex with ADP+vanadate and lipopolysaccharide” and “X-ray structure of the EmrE multidrug transporter in complex with a substrate” (1–3).

The recently reported structure of Ssv1866 (4) indicated that our MsbA structures (1, 2, 5) were incorrect in both the hand of the structure and the topology. Thus, our biological interpretations based on these inverted models for MsbA are invalid.

An in-house data reduction program introduced a change in sign for anomalous differences. This program, which was not part of a conventional data processing package, converted the anomalous pairs (I+ and I−) to (F− and F+), thereby introducing a sign change. As the diffraction data collected for each set of MsbA crystals and for the EmrE crystals were processed with the same program, the structures reported in (1–3, 5–6) had the wrong hand.

The error in the topology of the original MsbA structure was a consequence of the low resolution of the data as well as breaks in the electron density for the connecting loop regions. Unfortunately, the use of the multicyclic refinement procedure still allowed us to obtain reasonable refinement values for the wrong structures.

The Protein Data Bank (PDB) files 1JSQ, 1PF4, and 1Z2R for MsbA and 1S7B and 2F2M for EmrE have been moved to the archive of obsolete PDB entries. The MsbA and EmrE structures will be recalculated from the original data using the proper sign for the anomalous differences, and the new Coα coordinates and structure factors will be deposited.

We very sincerely regret the confusion that these papers have caused and, in particular, subsequent research efforts that were unproductive as a result of our original findings.

GEOFFREY CHANG, CHRISTOPHER B. ROTH, CHRISTOPHER L. REYES, OWEN PORNILLOS, YEN-JU CHEN, ANDY P. CHEN

Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037, USA.

References

Retraction: Cocryrstal structure of synaptobrevin-II bound to botulinum neurotoxin type B at 2.0 Å resolution

Michael A Hanson & Raymond C Stevens


In this paper, we described both the three-dimensional crystal structure of a botulinum toxin catalytic domain separated from the holotoxin (BoNT/B-LC, PDB 1F82) and a structure of the toxin catalytic domain in complex with a peptide (Sb2–BoNT/B-LC, PDB 1P83). The complex was later refined and deposited in the Protein Data Bank (PDB 3G94). The apo structure (PDB 1F82) remains valid. However, because of the lack of clear and continuous electron density for the peptide in the complex structure, the paper is being retracted. We apologize for any confusion this may have caused.
Register Errors

- Register errors typically start in loop regions (over or under building).
- Fixing these errors can be challenging (loop regions often have poor density) - estimated that 1% of structures in PDB have register errors.
- Real space analysis can help.
- Packing analysis (WHATCHECK, MolProbity).

1CHR, 3.0 (light) versus 2CHR (dark)

Image from Gerard Kleywegt, European Bioinformatics Institute

Terwilliger et al., Acta Cryst D64, 515-524, 2008
Other Kinds of Errors

- Systematic error in magnification
- Incorrect sequence (less common these days)
- Incorrectly placed waters or too many waters
- Waters fit instead of ions and side chains
- Small molecule geometry (where did you get the restraints from?)

\[ I\text{HBP, 1.9Å, 1993} \]

\[ \text{Semi-empirical calculated structure} \]

Phenix
Geometric Measures

- Some of the best measures for validation are from information not used in the model optimization (e.g. Free R-value).
- For geometry (of proteins) one of the best measures is the Ramachandran distribution - the main chain torsion angles.
- The handedness of amino acids, and the steric clashes that occur, given the side chain attachment to the mainchain, results in limits on the distribution of mainchain torsion angles.

G. N. Ramachandran
The Ramachandran Plot

All minus Pro & Gly

Proline

Glycine

Pre-proline

Images from Jane and David Richardson, Duke University

Phenix
The Ramachandran Plot

- A protein structure should in general conform to prior expectations (based on theory and prior observation)
- Most (98%+) residues should have a mainchain conformation consistent with the Ramachandran distribution
- A small percentage (0.2%) of residue may show Ramachandran outliers (note they are not necessarily errors)
  - Outliers can be seen in strained regions of the structure (e.g. in the active site)
- Any outliers need to be confirmed by detailed analysis
Rotamers

- There are steric clashes between atoms within amino acid side chains
- These clashes lead to preferred conformations, called rotamers
- Different rotamers are generated by rotation of side chain torsion angles ($\chi_1$, $\chi_2$ etc)

Image from Jane and David Richardson, Duke University
• As with the Ramachandran distribution, protein side chains are expected to conform to known rotamer distributions
• More variability because of interactions with other sidechains, mainchain or ligands
• Outliers may be meaningful, but need to be verified
• Sidechains on the protein surface will often have little density (disorder)
Hydrogens

- Macromolecules contain hydrogens
  - Approximately half of the atoms in a structure
- Hydrogens make the majority of contacts in a structure
- Typically ignored because they aren’t typically seen experimentally
- But, the hydrogens are there!
- The Richardson group (Duke University) have pioneered the use of hydrogens in calculating packing (and clashes) inside macromolecules
- The quality of packing and the nature of clashes can be used to validate and correct structures

Images from Jane and David Richardson, Duke University
All Atom Contacts:
Add H atoms
Roll 0.25 Å radius probe
... 3 terms -->

Image from Jane and David Richardson, Duke University
MolProbity

- MolProbity has been developed to validate structures (purely on coordinates)
- Performs all atom contacts, Ramachandran, rotamer and other geometry analyses

Analysis output: all-atom contacts and geometry for 3g5uH.pdb

Summary statistics

<table>
<thead>
<tr>
<th>All-Atom Contacts</th>
<th>Clashscore, all atoms: 159.56</th>
<th>2\textsuperscript{nd} percentile* (N=37, 3Å - 9999Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clashscore is the number of serious steric overlaps (&gt; 0.4 Å) per 1000 atoms.</td>
<td></td>
</tr>
<tr>
<td>Poor rotamers</td>
<td>20.10%</td>
<td>Goal: &lt;1%</td>
</tr>
<tr>
<td>Ramachandran outliers</td>
<td>11.33%</td>
<td>Goal: &lt;0.2%</td>
</tr>
<tr>
<td>Ramachandran favored</td>
<td>62.35%</td>
<td>Goal: &gt;98%</td>
</tr>
<tr>
<td>CB deviations &gt;0.25Å</td>
<td>6</td>
<td>Goal: 0</td>
</tr>
<tr>
<td>MolProbity score^</td>
<td>4.55</td>
<td>4\textsuperscript{th} percentile* (N=342, 3.25Å - 4.05Å)</td>
</tr>
<tr>
<td>Residues with bad bonds:</td>
<td>0.04%</td>
<td>Goal: 0%</td>
</tr>
<tr>
<td>Residues with bad angles:</td>
<td>3.85%</td>
<td>Goal: &lt;0.1%</td>
</tr>
</tbody>
</table>

* 100\textsuperscript{th} percentile is the best among structures of comparable resolution; 0\textsuperscript{th} percentile is the worst.

^ MolProbity score is defined as the following: \[0.42574 \times \log(1 + \text{clashscore}) + 0.32996 \times \log(1 + \max(0, \text{pctRotOut}-1)) + 0.24979 \times \log(1 + \max(0, 100 - \text{pctRamafavored}-2)) + 0.5\]

By adding H to this model and allowing Asn/Gln/His flips, we could automatically improve your clashscore by 2.05 points.

http://molprobity.biochem.duke.edu/
MolProbity

- Generates detailed problem list
- Problems can be fixed more easily by using validation lists viewed visually (e.g. Coot from Phenix)

<table>
<thead>
<tr>
<th>#</th>
<th>Res</th>
<th>High B</th>
<th>Clash &gt; 0.4Å</th>
<th>Ramachandran</th>
<th>Rotamer</th>
<th>Cβ deviation</th>
<th>Bond lengths.</th>
<th>Bond angles.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 33</td>
<td>VAL</td>
<td>207.38</td>
<td>0.761Å N with A 36 LEU HD11</td>
<td>-</td>
<td>5.4% (m) chi angles: 285.2</td>
<td>0.04Å</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A 34</td>
<td>SER</td>
<td>186.49</td>
<td>1.084Å HA with A 38 MET HB2</td>
<td>OUTLIER (0%) General case / 1.9, 66.3</td>
<td>82.4% (p) chi angles: 69.5</td>
<td>0.015Å</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A 35</td>
<td>VAL</td>
<td>204.94</td>
<td>1.221Å HG12 with A 359 TYR CE2</td>
<td>Allowed (0.14%) General case / -50.4, -77.0</td>
<td>19.1% (m) chi angles: 304.1</td>
<td>0.049Å</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A 36</td>
<td>LEU</td>
<td>142.23</td>
<td>1.095Å H with A 35 VAL HG23</td>
<td>Favored (42.64%) General case / -73.2, -49.2</td>
<td>1% chi angles: 55,4,111.8</td>
<td>0.07Å</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A 37</td>
<td>THR</td>
<td>170.59</td>
<td>0.723Å H with A 36 LEU HG</td>
<td>Favored (3.45%) General case / -69.8, -60.5</td>
<td>64.9% (m) chi angles: 296.4</td>
<td>0.052Å</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A 38</td>
<td>MET</td>
<td>155.79</td>
<td>1.084Å HB2 with A 34 SER HA</td>
<td>Favored (36.13%) General case / -49.2, -42.4</td>
<td>0% chi angles: 229,3,294,4,131.7</td>
<td>0.054Å</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A 39</td>
<td>PHE</td>
<td>122.57</td>
<td>1.127Å HB2 with A 35 VAL O</td>
<td>Favored (49.78%) General case / -51.0, -39.2</td>
<td>31.2% (θ) chi angles: 196,6,84</td>
<td>0.027Å</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A 40</td>
<td>ARG</td>
<td>102.85</td>
<td>0.683Å N with A 37 THR O</td>
<td>Allowed (0.94%) General case / -116.3,69.9</td>
<td>9.1% (θ) chi angles: 246,9,176,6,36.5,191.6</td>
<td>0.078Å</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Results - Rebuilding and Validation
Validation

- Outlier lists recenter Coot view; Probe dots automatically loaded
Using Validation Tools Improves Models

Images from Jane and David Richardson, Duke University
Cis-Peptides

cis-nonPro peptides are very rare (~0.03%), usually genuine and functionally important
Too Many Cis-Peptides

- Cis non-Prolines are chosen much more often than chance, because they are more compact than trans and fit better into the shrunken & rather featureless low-resolution density (esp. for loops).
- Automated building with a no-cis fragment library (Tom Terwilliger)
Flagging non-Pro cis & twisted peptides in Omegalyze

very rare: ~0.03% of non-Pro

normal: ~5% of Pro

almost never justified

Twisted Peptide = >30 degrees from planar
Validation Using Cα Atoms

CaBLAM Parameter Space

A minimalist alternative

Christopher Williams,
Duke University
Identifying Distorted Secondary Structure

Diagnosing Strands
Pathological strands from 70S Ribosome

Christopher Williams,
Duke University
Assessing Secondary Structure Probability

CabLAM-space

β validation

general = 0.7%
(outlier)

β = 85.8%

α = 0%

Phenix

Christopher Williams,
Duke University
Comprehensive Validation

These statistics are computed using the same underlying distributions as the MolProbity web server. The overall score represents the experimental resolution expected for a model of this quality; ideally the score should be lower than the actual resolution.

Overall scores

MolProbity score: 1.72  Clash score: 5.44

CaBLAM

Outliers (%): 3.88  Disfavored (%): 8.96  Cα outliers (%): 1.19

<table>
<thead>
<tr>
<th>Chain</th>
<th>Residue</th>
<th>Evaluation</th>
<th>CaBLAM Score</th>
<th>CA Geometry Score</th>
<th>Secondary Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>ILE 955</td>
<td>CaBLAM Disfav...</td>
<td>0.03762</td>
<td>0.01447</td>
<td>try alpha helix</td>
</tr>
<tr>
<td>A</td>
<td>PRO 969</td>
<td>CaBLAM Disfav...</td>
<td>0.02931</td>
<td>0.46424</td>
<td>try alpha helix</td>
</tr>
<tr>
<td>A</td>
<td>SER 1012</td>
<td>CaBLAM Outlier</td>
<td>0.00273</td>
<td>0.67504</td>
<td>try alpha helix</td>
</tr>
<tr>
<td>A</td>
<td>LEU 1016</td>
<td>CaBLAM Outlier</td>
<td>0.00086</td>
<td>0.07553</td>
<td></td>
</tr>
</tbody>
</table>

Cα deviation analysis

No Cα position outliers detected.

Cis and twisted peptides

Cis conformations are observed in about 5% of Prolines. Cis conformations are observed in about 0.03% of general residues. Twisted peptides are almost certainly modeling errors.

No non-trans peptides detected.
## Map Resolution and Map/Model Fit

### Summary of map resolution estimates.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Objects used</th>
<th>Purpose</th>
<th>Values</th>
<th>Meaning, possible actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{FSC}$</td>
<td>Half-maps</td>
<td>Highest resolution at which the experimental data are confident</td>
<td>The higher the better</td>
<td>Resolution determined using half-maps method</td>
</tr>
</tbody>
</table>
| $d_{99}$   | Map          | Resolution cutoff beyond which Fourier coefficients are negligibly small                    | $d_{99} \geq d_{FSC}$ $d_{99} < d_{FSC}$ $d_{99} \gg d_{FSC}$ | Expected values
Verify $d_{FSC}$; omit coefficients with $d_{99} \leq d < d_{FSC}$\nSharpen the map
Expected values
Verify $d_{FSC}$; check ADP (too large?); validate map details\nSharpen the map
Check ADP (too large?)\nCheck ADP (too small?); check the model\nExpected values\nVerify $d_{FSC}$; omit coefficients with $d_{FSC} \leq d < d_{FSC}$\nSharpen the map\nOmit coefficients with $d_{model} \leq d < d_{FSC}$\nSharpen the map |
| $d_{model}$| Map and model| Resolution cutoff at which the model map is the most similar to the target map               | $d_{model} \geq d_{FSC}$ $d_{model} < d_{FSC}$ $d_{model} \gg d_{model}$ $d_{model} \ll d_{99}$ | $d_{model} \gg d_{99}$\nCheck ADP (too large?)\nCheck ADP (too small?); check the model |
| $d_{FSC\_model}$ | Map and model | Resolution cutoff up to which the model and map Fourier coefficients are similar         | $d_{FSC\_model} \geq d_{FSC}$ $d_{FSC\_model} < d_{FSC}$ $d_{FSC\_model} \geq d_{FSC}$ $d_{FSC\_model} \gg d_{model}$ $d_{FSC\_model} \ll d_{model}$ | $d_{FSC\_model} \gg d_{model}$\nCheck ADP (too large?)\nCheck ADP (too small?); check the model |

### Summary of map correlation coefficients used in this work.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Region of the map used in calculation</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC&lt;sub&gt;box&lt;/sub&gt;</td>
<td>Whole map</td>
<td>Similarity of maps</td>
</tr>
<tr>
<td>CC&lt;sub&gt;mask&lt;/sub&gt;</td>
<td>Jiang &amp; Brünger (1994) mask with a fixed radius</td>
<td>Fit of the atomic centers</td>
</tr>
<tr>
<td>CC&lt;sub&gt;volume&lt;/sub&gt;</td>
<td>Mask of points with the highest values in the model map</td>
<td>Fit of the molecular envelope defined by the model map</td>
</tr>
<tr>
<td>CC&lt;sub&gt;peaks&lt;/sub&gt;</td>
<td>Mask of points with the highest values in the model and in the target maps</td>
<td>Fit of the strongest peaks in the model and target maps</td>
</tr>
<tr>
<td>CC&lt;sub&gt;vr_mask&lt;/sub&gt;</td>
<td>Same as CC&lt;sub&gt;mask&lt;/sub&gt; but atomic radii are variable and function of resolution, atom type and ADP</td>
<td>Fit of the atomic images in the given map</td>
</tr>
</tbody>
</table>

Resolution Determination

\[ FSC(r) = \frac{\sum_{r_i \in r} F_1(r_i) \cdot F_2(r_i)^*}{\sqrt{\sum_{r_i \in r} |F_1(r_i)|^2 \cdot \sum_{r_i \in r} |F_2(r_i)|^2}} \]
Cross Validation with Half Maps

- Perturb model (random shift of coordinates)
- Re-refine against 1 half map
- Calculate FSC of model against 2nd half map
- FSC curve shouldn’t show signal beyond the half map resolution
Model/Map Validation

Benjamin Barad, Yifan Cheng, Jaime Fraser
University of California San Francisco

Ray Yu-Ruei Wang, Frank DiMaio
University of Washington

Nat Echols
Lawrence Berkeley National Laboratory
Validation and Cryo-EM

- Do the map make sense?
  - Gold Standard FSC of half maps
- Does the model make sense?
  - MolProbity
- Does the model fit the map?
  - Overall and local density correlation
  - What about the detailed local fit?
Look at the Density Around Sidechains

Ringer


Phenix
Ben Barad, Jaime Fraser, UCSF
Look at the Density Around Sidechains

**EMRinger**

EMRinger reports on backbone placement

[Graphs and diagrams showing changes in Coulomb potential with Chi1 angle for different backbone placements.]

Phenix  
Ben Barad, Jaime Fraser, UCSF

University of California
San Francisco
EMRinger Score to Validate Model vs Data

- Quantify how well the model backbone puts side chains in places where there are density peaks consistent with rotameric conformations

$R^2 = 0.547$

Side chain information = random (on average)

Available in GUI and command line

phenix.emringer model.pdb map.ccp4

http://emringer.com
Ensembles

• At lower resolution ensemble models are probably more appropriate
• Can be used to help assess map variability (Herzik, Fraser, Lander. Structure. 2019)
Deposition Issues

- Successful re-analysis of cryo-EM data relies on accurate data/model deposition
- Current practice has led to significant issues:
  - Models misplaced wrt maps
  - Inconsistent map deposition (sharpened, masked, filtered, wrong map)
  - Absence of half-maps
  - Very variable assessments of resolution
  - Optimistic ligand placement (probably unintentional)

Conclusions

• Many of the validation metrics developed to assess models can be readily applied to cryo-EM structures
• Many of the pitfalls of low resolution from other fields apply to cryo-EM
• Care needs to be taken to ensure that validation metrics can be used when restraints are applied in refinement
• Additional validation metrics for the model w.r.t. the data are needed
• We do not have cross-validation metrics for the model/data
Acknowledgements

<table>
<thead>
<tr>
<th>Berkeley Laboratory</th>
<th>UC San Francisco</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pavel Afonine, Youval Dar, Nat Echols, Jeff Headd, Richard Gildea, Ralf Grosse-Kunstleve, Dorothee Liebschner, Nigel Moriarty, Nader Morshed, Billy Poon, Ian Rees, Nicholas Sauter, Oleg Sobolev, Peter Zwart</td>
<td>Ben Barad, Yifan Cheng, Jaime Fraser</td>
</tr>
<tr>
<td>Los Alamos Laboratory/New Mexico Consortium</td>
<td>University of Washington</td>
</tr>
<tr>
<td>Tom Terwilliger, Li-Wei Hung</td>
<td>Frank DiMaio, Ray Wang, David Baker</td>
</tr>
<tr>
<td>Baylor College of Medicine</td>
<td>Oak Ridge National Laboratory</td>
</tr>
<tr>
<td>Matt Baker</td>
<td>Marat Mustyakimov, Paul Langan</td>
</tr>
<tr>
<td>Cambridge University</td>
<td>Other Collaborators</td>
</tr>
<tr>
<td>Randy Read, Airlie McCoy, Gabor Bunckozi, Tristan Croll, Rob Oeffner, Kaushik Hatti, Massimo Sammito, Duncan Stockwell, Laurent Storoni</td>
<td>Corey Hryc, Zhao Wang, Wah Chiu, Pawel Janowski, David Case, Dale Tronrud, Donnie Berholz, Andy Karplus, Alexandre Urzhumtsev &amp; Vladimir Lunin, Garib Murshudov &amp; Alexi Vagin, Paul Emsley, Bernhard Lohkamp, Kevin Cowtan, David Abrahams, PHENIX Testers &amp; Users</td>
</tr>
<tr>
<td>Duke University</td>
<td>Funding</td>
</tr>
<tr>
<td>Jane Richardson &amp; David Richardson, Ian Davis, Vincent Chen, Jeff Headd, Chris Williams, Bryan Arendall, Bradley Hintze, Laura Murray</td>
<td>- NIH/NIGMS: P01GM063210, P50GM062412, P01GM064692, R01GM071939</td>
</tr>
<tr>
<td></td>
<td>- PHENIX Industrial Consortium</td>
</tr>
<tr>
<td></td>
<td>- Lawrence Berkeley Laboratory</td>
</tr>
</tbody>
</table>