Using AlphaFold predictions for structure determination

Phenix Workshop
March 18-19, 2024, University of Oklahoma

Slides by Tom Terwilliger
The New Mexico Consortium
Los Alamos National Laboratory

Presented by Christopher Williams,
with additional slides

Richardson Lab
Duke University, Biochemistry Department
AlphaFold predictions are great hypotheses

AlphaFold models can be…..
AlphaFold predictions and confidence estimates

Residue-specific confidence (pLDDT) identifies where errors are more likely

<table>
<thead>
<tr>
<th>AlphaFold confidence (pLDDT)</th>
<th>Median prediction error (Å)</th>
<th>Percentage with error over 2 Å</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;90</td>
<td>0.6</td>
<td>10</td>
</tr>
<tr>
<td>80 - 90</td>
<td>1.1</td>
<td>22</td>
</tr>
<tr>
<td>70 - 80</td>
<td>1.5</td>
<td>33</td>
</tr>
<tr>
<td>&lt;70</td>
<td>3.5</td>
<td>77</td>
</tr>
</tbody>
</table>

**AlphaFold confidence measure**  
(pLDDT, Predicted difference distance test)

**Confidence:**  
**Blue:** > 90  
**Green:** 80 - 90

**AlphaFold prediction for RNA helicase**  
(PDB entry 6i5i)

<table>
<thead>
<tr>
<th>AlphaFold confidence (pLDDT)</th>
<th>Median prediction error (Å)</th>
<th>Percentage with error over 2 Å</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;90</td>
<td>0.6</td>
<td>10</td>
</tr>
<tr>
<td>80 - 90</td>
<td>1.1</td>
<td>22</td>
</tr>
<tr>
<td>70 - 80</td>
<td>1.5</td>
<td>33</td>
</tr>
<tr>
<td>&lt;70</td>
<td>3.5</td>
<td>77</td>
</tr>
</tbody>
</table>

PAE matrix (Predicted aligned error)

PAE matrix identifies accurately-predicted domains

Confidence:

Blue: > 90
Green: 80 - 90

AlphaFold prediction for RNA helicase (PDB entry 6i5i)

Dark blue: uncertainty in relative positions < 5 Å
Strategy for structure determination in the AlphaFold era

1. Predict your structure
   - Design your experiment based on predicted models (choose experimental approach, consider trimming at domain boundaries)

2. Solve your structure
   - Cryo-EM or X-ray MR with trimmed predicted model, SAD

3. Update your prediction
   - Run AlphaFold with your best model as a template

4. Improve your structure
   - Use your new predictions as hypotheses

Iterate
Using your best model as a template in AlphaFold prediction

Why?

Because your new prediction might be better than your model ...and better than your original AlphaFold prediction.
Improving AlphaFold prediction using partial models as templates (Cryo-EM)

Data from 7mjs, Cater, R.J., et al. (2021). Nature 595, 315–319
Phenix AlphaFold prediction server

Available from the Phenix GUI

Predicts structures of protein chains
(one at a time)

Can use a template to guide the prediction

You do not need an MSA (multiple sequence alignment) if you supply a template

The template should not be an AlphaFold model

Many thanks for AlphaFold, ColabFold scripts, and the MMseqs2 server for MSAs
Process predicted model

Convert pLDDT to B-value

Trim low-confidence parts of model

Identify high-confidence domains

Compact high-confidence regions

Groupings of residues with low PAE values
Phenix tools for structure determination with AlphaFold

- **PredictModel** (Predict with AlphaFold)
- **ProcessPredictedModel** (Trim and identify domains)
- **ResolveCryoEM, LocalAnisoSharpen** (map improvement)
- **EMPlacement, DockInMap** (Docking of single, multiple chains)
- **DockAndRebuild** (Morphing and rebuilding)
- **RealSpaceRefine** (Refinement)
- **Phaser-MR** (Molecular replacement)
- **AutoBuild** (Density modification and rebuilding)
- **Phenix.refine** (Refinement)
- **PredictAndBuild** (Prediction and structure determination)

AlphaFold models

Cryo-EM

X-ray

Full automation
Low-pLDDT Alphafold predictions

• Most of the time, AlphaFold predictions are high-confidence and easy to interpret

• Most of the time, phenix.process_predicted_model is all you need

• So, let’s talk about the other times . . .
When automation struggles, Use visualization

Mol* viewer at https://alphafold.ebi.ac.uk

ChimeraX: “color bfactor palette alphafold”
Features to watch for

- High pLDDT
  - Unpacked helices

- Low pLDDT
  - Non-predictive “barbed wire”
  - Unpacked, physically possible regions
  - Near-predictive packed regions
Unpacked high pLDDT

- High-confidence, protein-like structure, touching nothing
  - Often helix
  - Often well-separated by PAE matrix
- Probably folded in biological multimer/complex
- May have to truncate the construct for solo crystallization

Homo sapiens
Uniprot P60228

M. Jannaschii
Uniprot Q58865
AlphaFold predictions and confidence estimates

Residue-specific confidence (pLDDT) identifies where errors are more likely

<table>
<thead>
<tr>
<th>AlphaFold confidence (pLDDT)</th>
<th>Median prediction error (Å)</th>
<th>Percentage with error over 2 Å</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;90</td>
<td>0.6</td>
<td>10</td>
</tr>
<tr>
<td>80 - 90</td>
<td>1.1</td>
<td>22</td>
</tr>
<tr>
<td>70 - 80</td>
<td>1.5</td>
<td>33</td>
</tr>
<tr>
<td>&lt;70</td>
<td>3.5</td>
<td>77</td>
</tr>
</tbody>
</table>

The low-pLDDT regime contains multiple behaviors

Awesome
Wrong
Distorted

Low pLDDT - Barbed wire

Low-confidence AlphaFold predictions often have wide coils like concertina wire
Barbed wire

• Extreme density of geometry outliers
  • (The protein is not actually drawn in this image, just the validation markup)

• This is a good thing!

• Along with pLDDT, this clearly and consistently marks regions where AlphaFold hasn’t made any prediction

• Different from “normal” modeling errors
Unpacked Possible

• Somewhat protein-like conformations

• Possibly folded in full biological context

• Unpacked and unidealized, but…

Zinc finger CCCH domain-containing protein 13
Residues 70-100
Homo sapiens
Uniprot Q5T200
Unpacked Possible

• Lacks validation outliers!

• Also lacks good hydrogen bonding

• More “real” than barbed wire, but no predictive value in most cases

Extended conformation is still Ramachandran-legal

Few or no H-bonds

Inconsistent helix identification

Zinc finger CCCH domain-containing protein 13
Residues 70-100
_Homo sapiens_
Uniprot _Q5T200_
Near-predictive

- Low pLDDT, but . . .
- Well-packed
- Protein-like fold
- Protein-like local geometry

*Homo sapiens*
Uniprot **P60228**
Near-predictive

*Homo sapiens*
Uniprot P60228

6zon.pdb, chain E  P60228 AlphaFold prediction
pLDDT comparison

Low pLDDT contains multiple behaviors
Protein-like regions with pLDDT ~45-70 *may* still be usable!
Whole-model statistics may be misleading

Barbed wire present, validation says “probably unusable”

Barbed wire removed, validation says “needs work”
Low-pLDDT tool in Phenix

• Barbed wire analysis combines:
  • pLDDT score
  • Packing quality
    • Ignores contacts within secondary structure
    • Ignores sequence-local contacts
  • Density of barbed wire-like validation problems

• *phenix.barbed_wire_analysis*
• *phenix.barbed_wire_analysis* output.type=kin
  • Colored balls kinemage markup
• *phenix.barbed_wire_analysis*
  output.type=selection_file
  • PDB-format file of just the Predictive and Near-predictive parts of the input
Low-pLDDT kinemage markup

• Predictive (blue)
• Unpacked high pLDDT (gray)
• Near-predictive (green)
• Unpacked possible (gold)
• Barbed wire (hot pink)

• This markup only available in KiNG/kinemage format for now.
• The low-pLDDT tool is still in development
An NIH/NIGMS funded Program Project

Sample workflows
X-ray structure determination with AlphaFold

Sequence file

Predict structure
Trim, identify domains

High-confidence domains

X-ray intensity data (mtz file)

Molecular replacement

Working structure and map

Density modification and autobuilding

Rebuilt model and optimized map

Predict using rebuilt chains as templates

Updated predictions

Iterate
Cryo-EM structure determination with AlphaFold

- Half-maps (optional processed map)
  - Density modification
  - or Anisotropic sharpening

- Optimized map
- Sequence file
  - Predict structure
  - Trim, identify domains

- High-confidence domains
  - Dock domains in map
- Docked domains
  - Morph full prediction onto domains and rebuild
- Rebuilt model
  - Predict using rebuilt chains as templates
- Updated predictions

Iterate
Input and output from structure determination with AlphaFold

Input:
- Experimental data (maps or X-ray data)
- Contents of asymmetric unit (sequence file)

Output:
- Rebuilt model
  - Optimized map
- Map and model ready for next steps
- Docked predicted models
- Useful as high-quality reference models
Improving AlphaFold prediction using partial models as templates
(X-ray crystallography)

AF predictions cycle 1 (green), cycle 2 (blue)