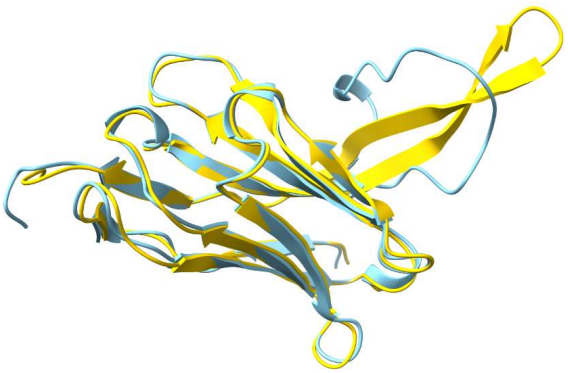
Using AlphaFold2 predictions for structure determination

Phenix Workshop
June 27-28, 2024, University of Montana



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

Phenix

Los Alamos NATIONAL LABORATORY

BERKELEY LAB

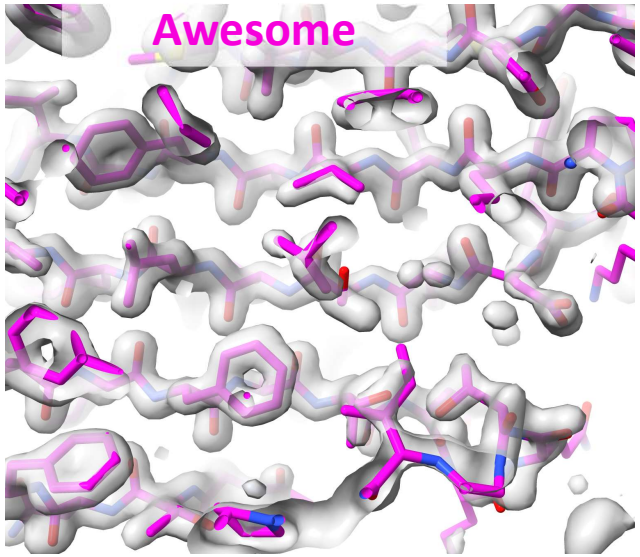
UNIVERSITY OF CAMBRIDGE

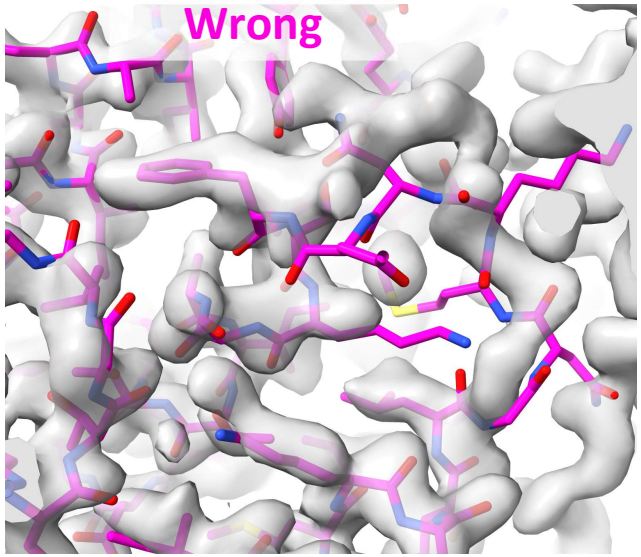
UTHealth The University of Texas Health Science Center at Houston

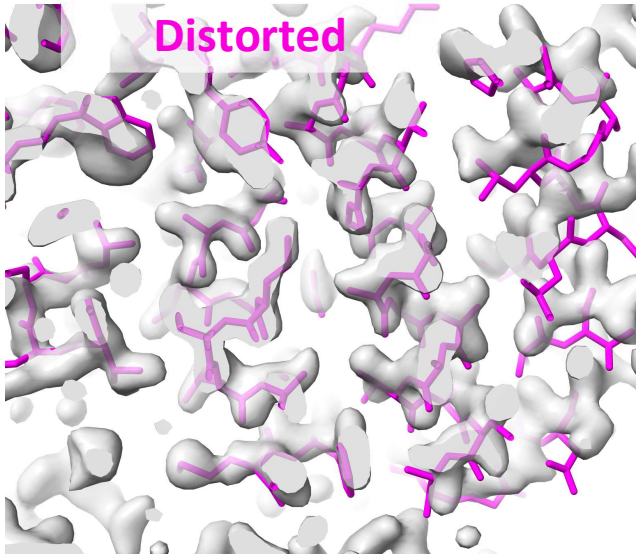
New Mexico CONSORTIUM

AlphaFold predictions are great hypotheses

AlphaFold models can be….



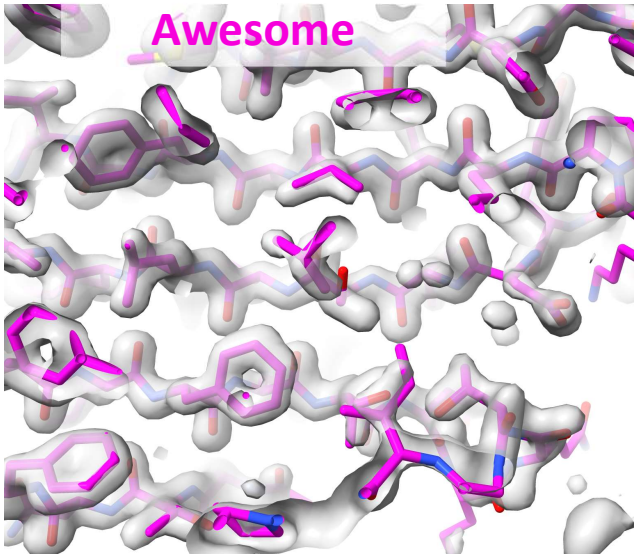


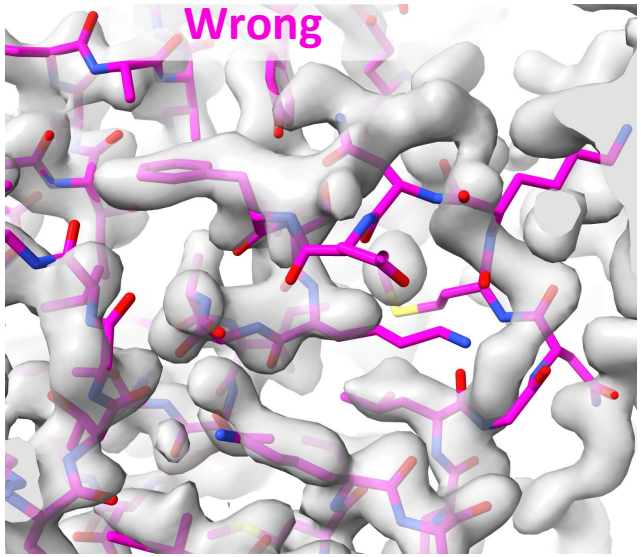


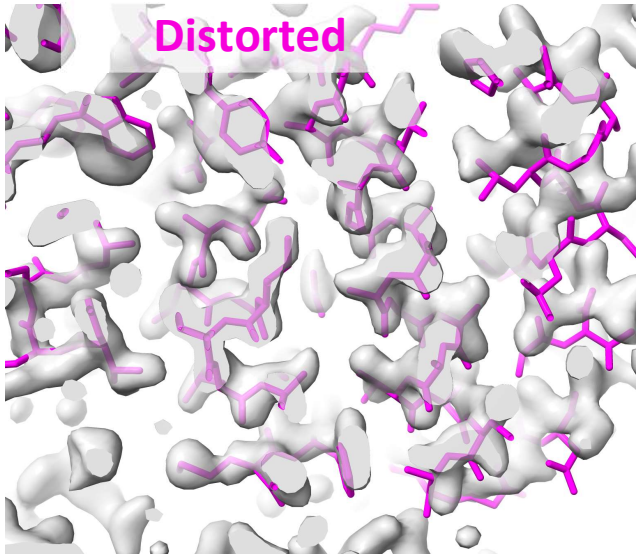
AlphaFold predictions and confidence estimates

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

AlphaFold confidence (pLDDT)	Median prediction error (Å)	Percentage with error over 2 Å
>90	0.6	10
80 - 90	1.1	22
70 - 80	1.5	33
<70	3.5	77



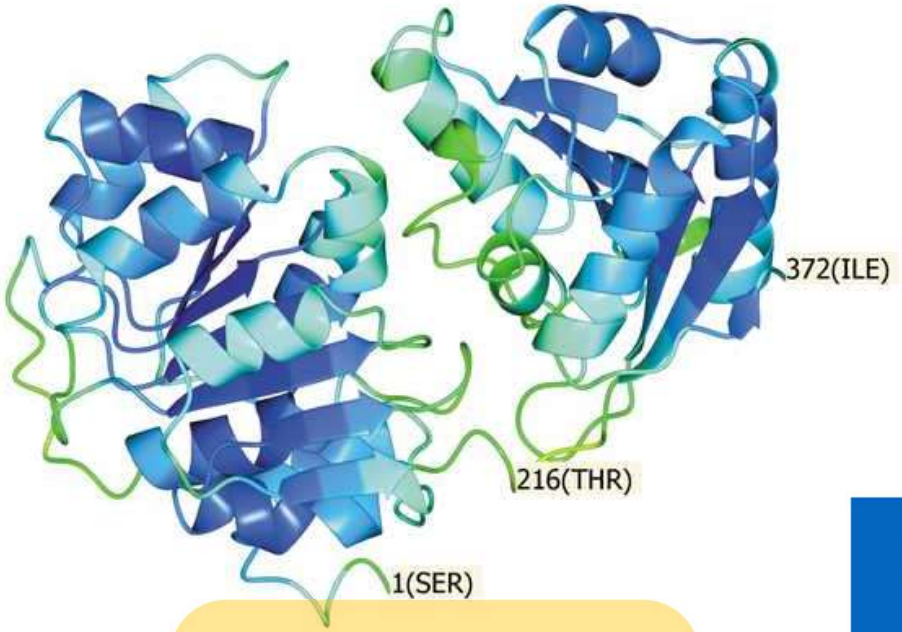




Terwilliger, Thomas C., et al. "AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination." Nature Methods 21.1 (2024): 110-116.

AlphaFold confidence measure

(pLDDT, Predicted difference distance test)



Confidence:

Blue: > 90

Green: 80 - 90

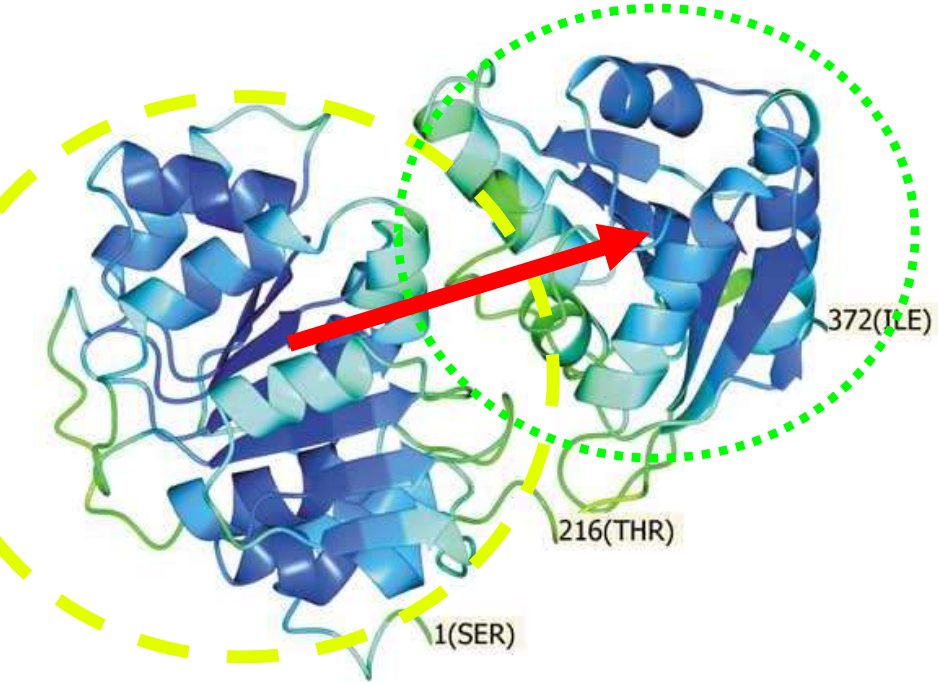
AlphaFold prediction for RNA helicase

(PDB entry 6i5i)

AlphaFold confidence (pLDDT)	Median prediction error (Å)	Percentage with error over 2 Å
>90	0.6	10
80 - 90	1.1	22
70 - 80	1.5	33
<70	3.5	77

Oeffner et al. (2022). Acta Cryst. D78, 1303-1314

PAE matrix *(Predicted aligned error)*



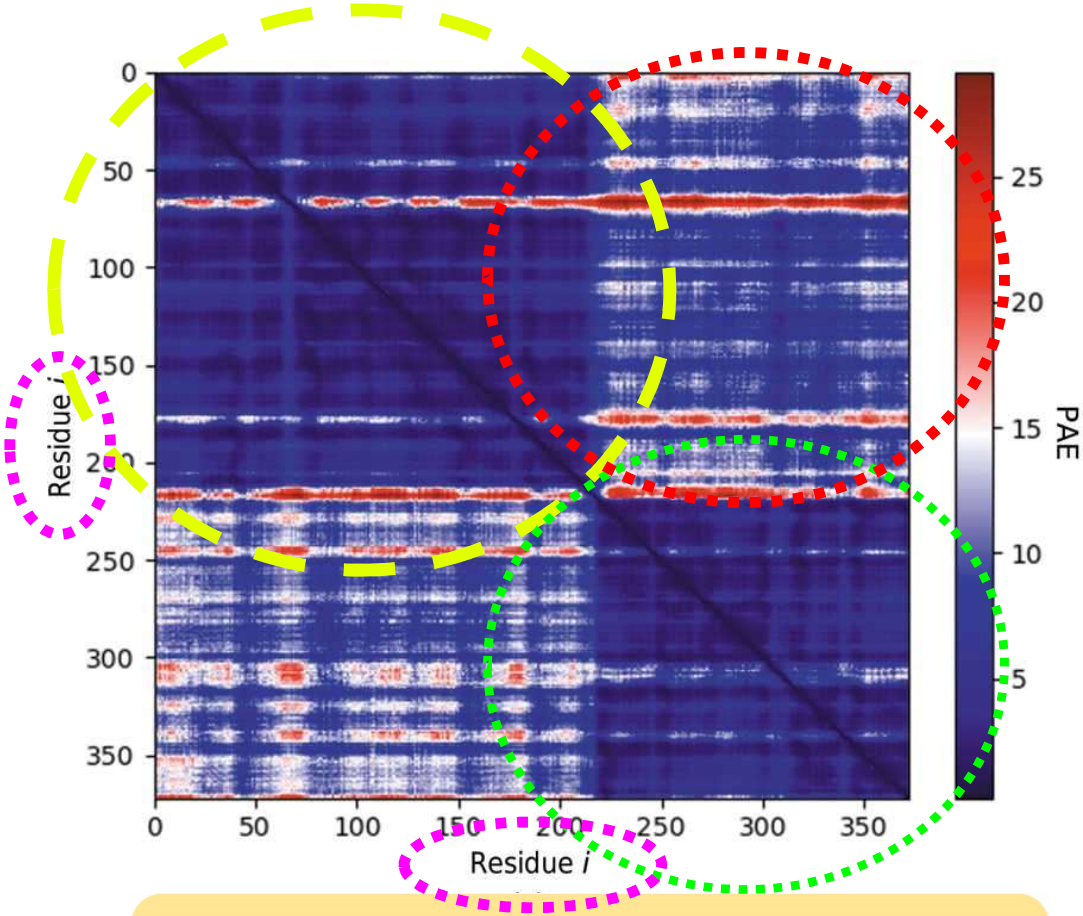
Confidence:

Blue: > 90

Green: 80 - 90

AlphaFold prediction for RNA helicase
(PDB entry 6i5i)

PAE matrix identifies accurately-predicted domains



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

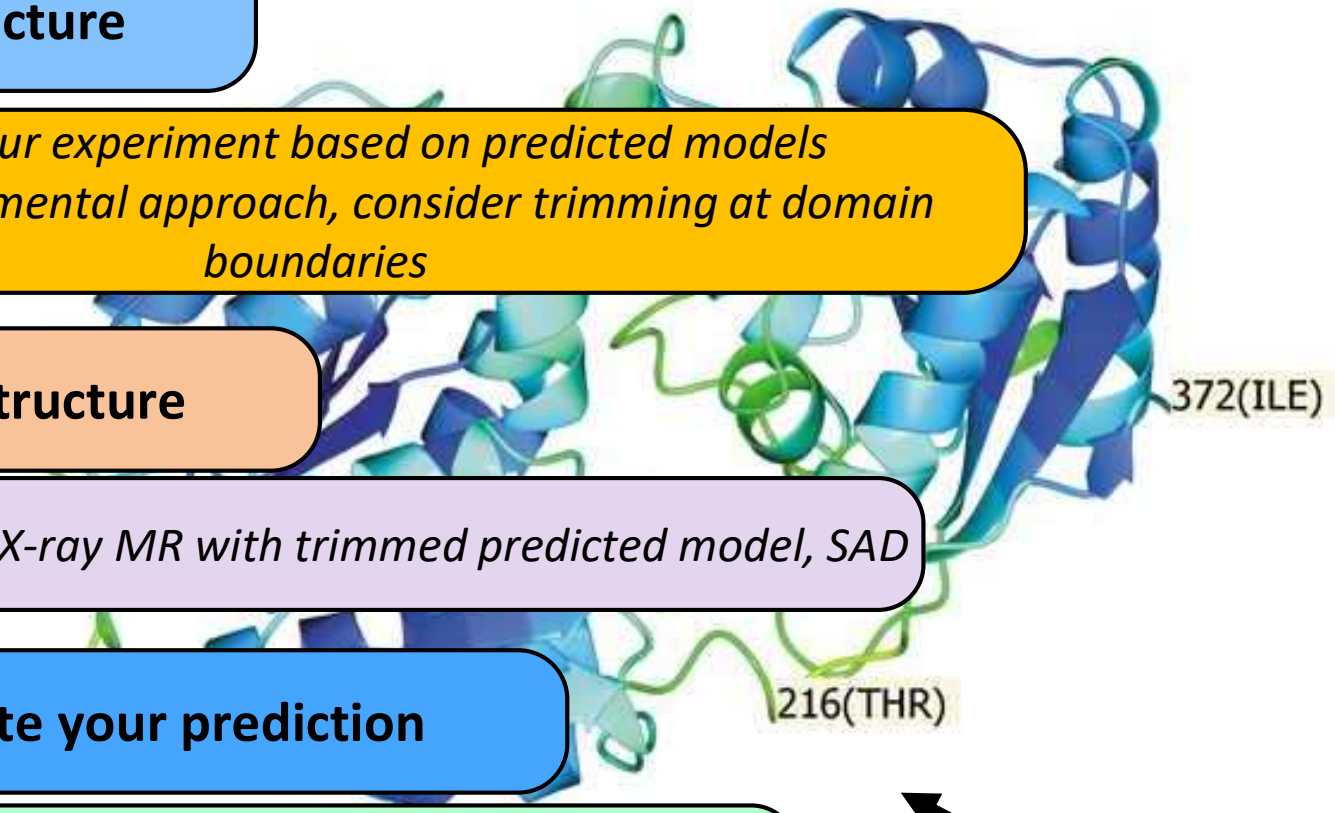
Cryo-EM or X-ray MR with trimmed predicted model, SAD

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

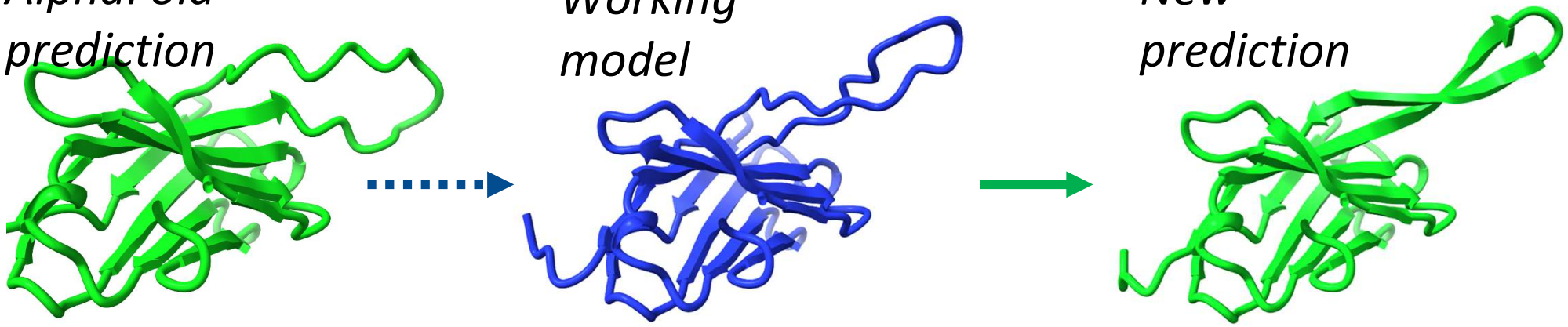
Why?

Because your new prediction might be better than your model ...and better than your original AlphaFold prediction

AlphaFold prediction

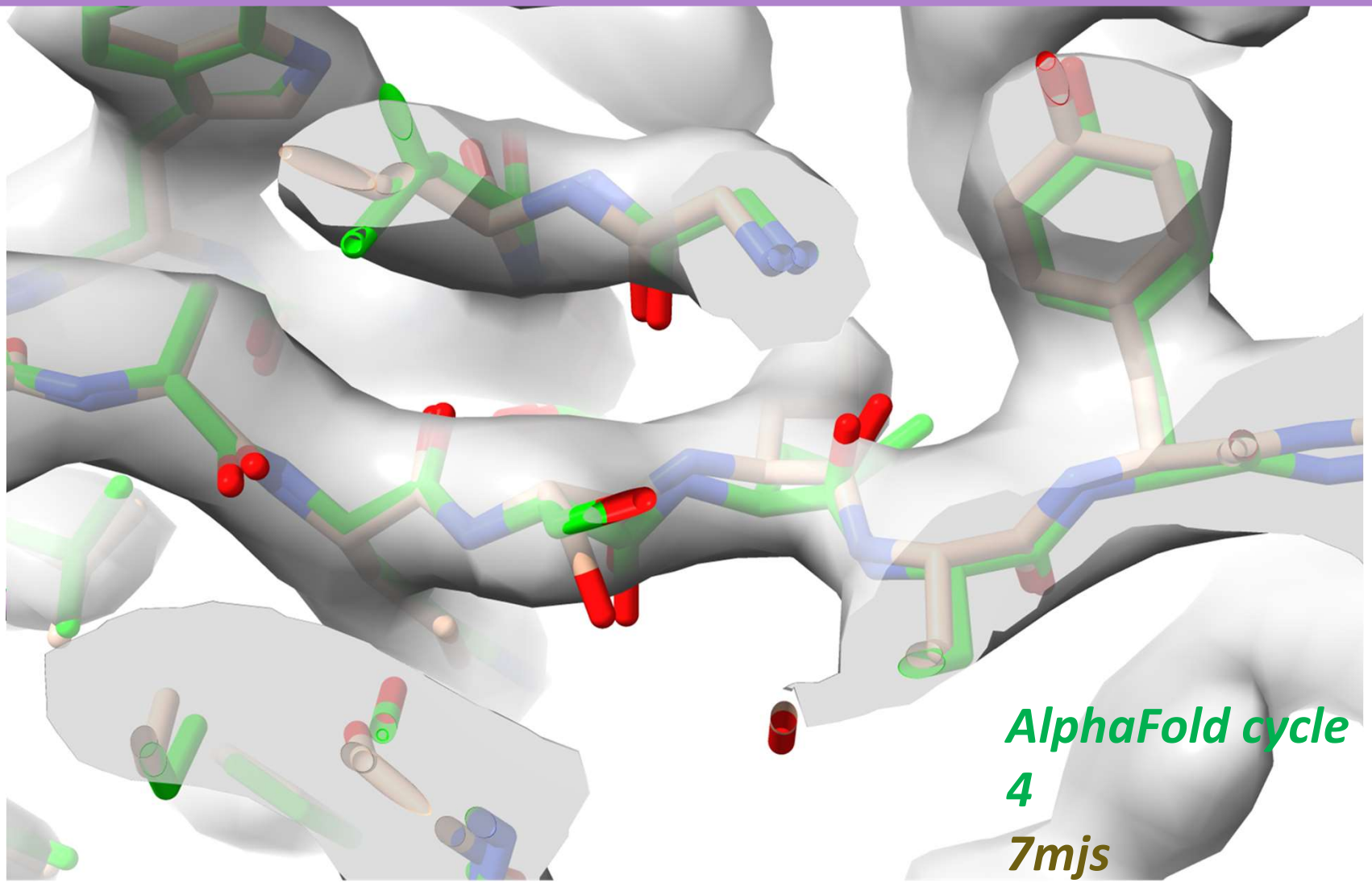
Working model

New 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

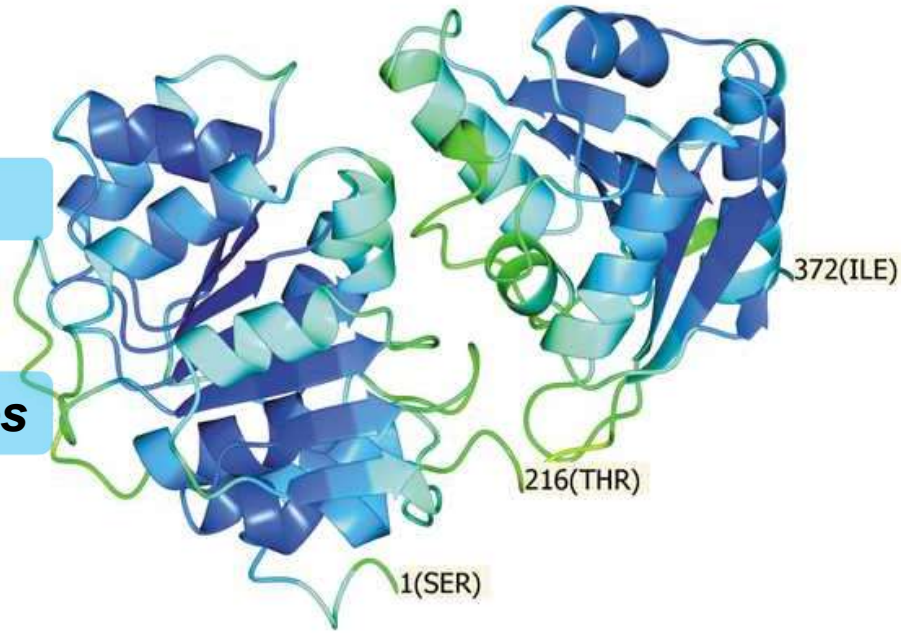
Convert pLDDT to B-value

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)

AlphaFold models

ResolveCryoEM, LocalAnisoSharpen (map improvement)

EMPlacement, DockInMap (Docking of single, multiple chains)

DockAndRebuild (Morphing and rebuilding)

RealSpaceRefine (Refinement)

Cryo-EM

Phaser-MR (Molecular replacement)

AutoBuild (Density modification and rebuilding)

Phenix.refine (Refinement)

X-ray

PredictAndBuild (Prediction and structure determination)

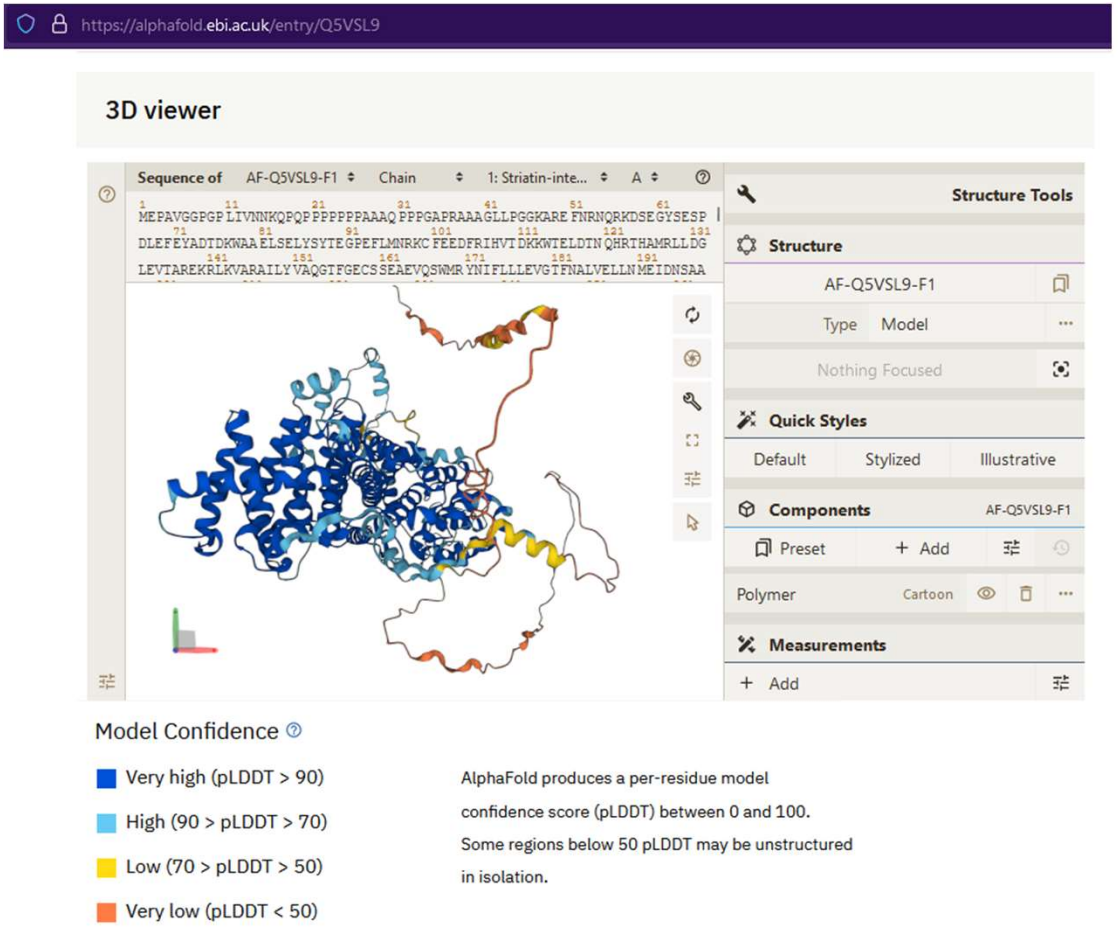
Full automation

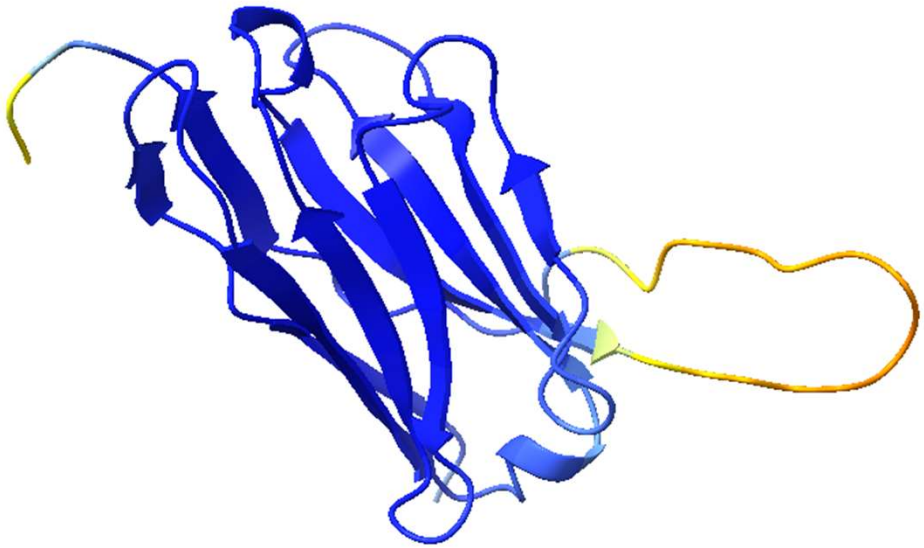
Low-pLDDT Alphafold predictions

- 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
or
PDB

ChimeraX: “color bfactor palette alphafold”

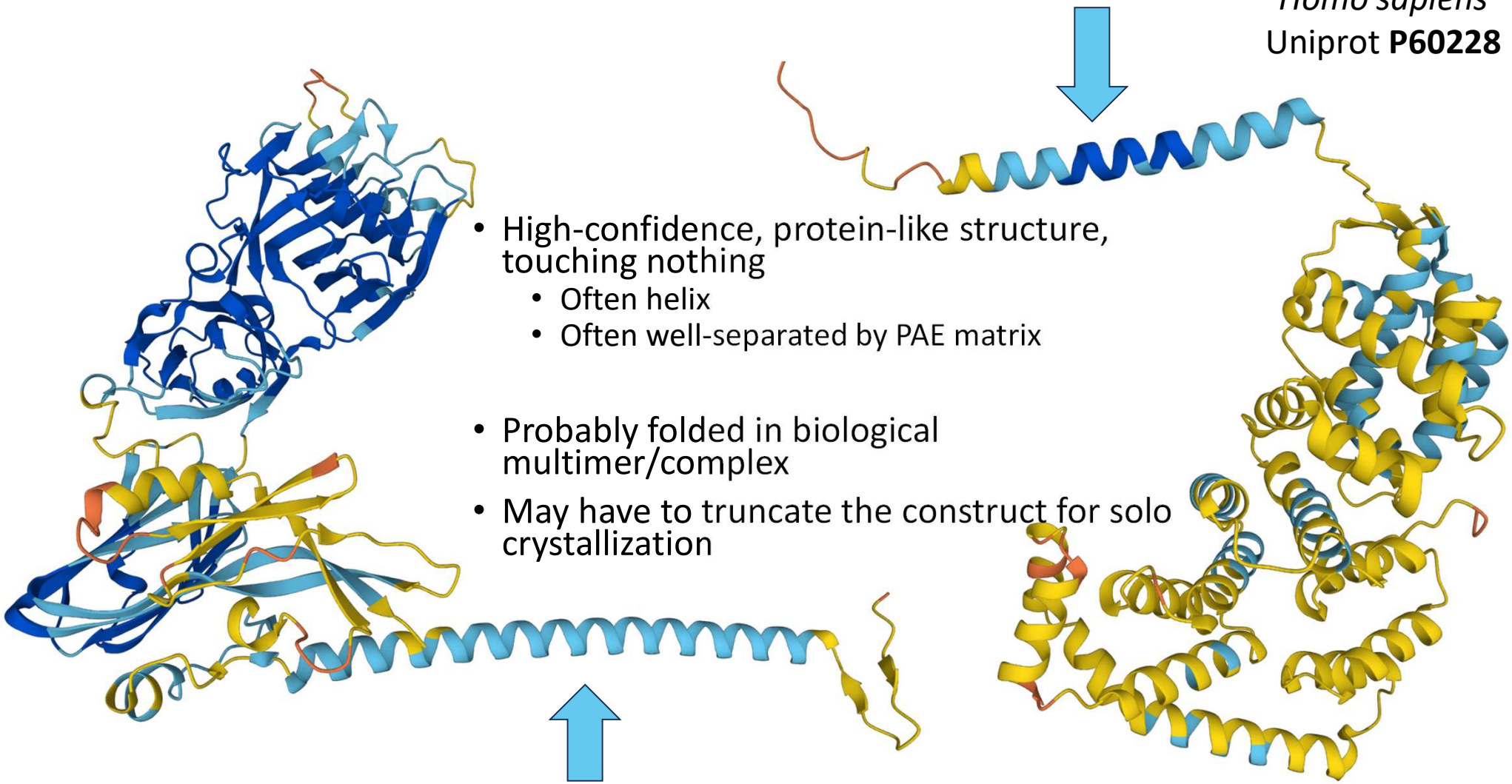
Features to watch for

- High pLDDT
  - Unpacked helices

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

Unpacked high pLDDT

Homo sapiens
Uniprot **P60228**



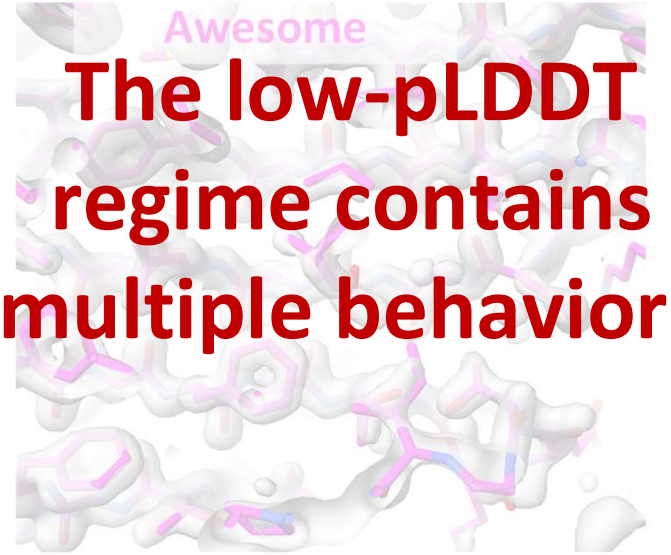
- High-confidence, protein-like structure, touching nothing
  - Often helix
  - Often well-separated by PAE matrix
- Probably folded in biological multimer/complex
- May have to truncate the construct for solo crystallization

M. Jannaschii
Uniprot **Q58865**

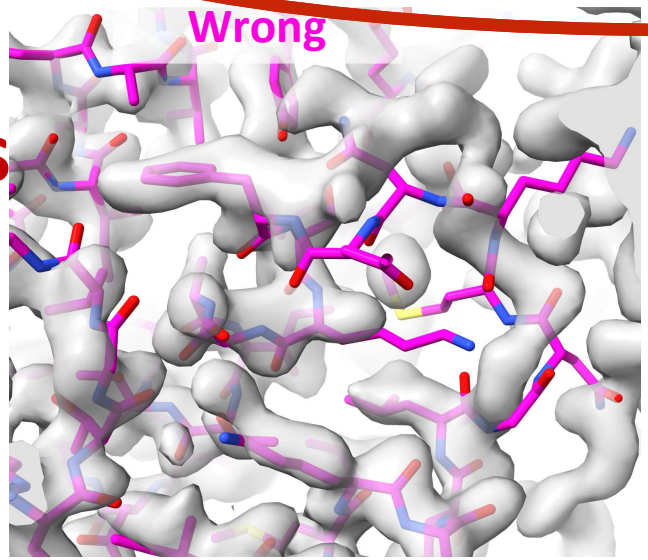
AlphaFold predictions and confidence estimates

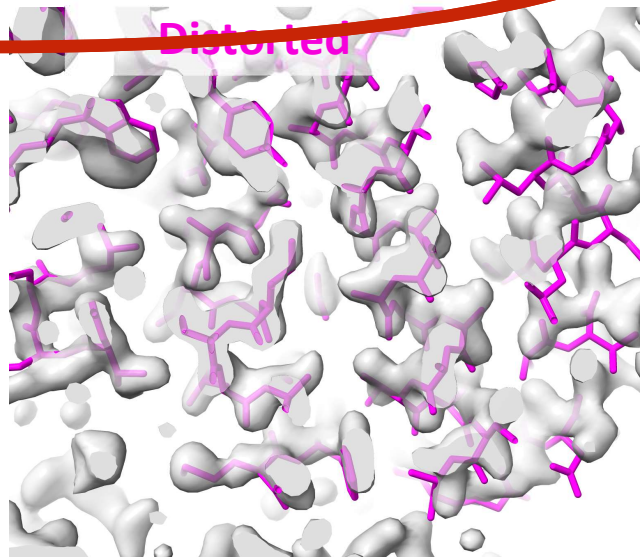
Residue-specific confidence (pLDDT) identifies where errors are more likely

AlphaFold confidence (pLDDT)	Median prediction error (Å)	Percentage with error over 2 Å
>90	0.6	10
80 - 90	1.1	22
70 - 80	1.5	33
<70	3.5	77



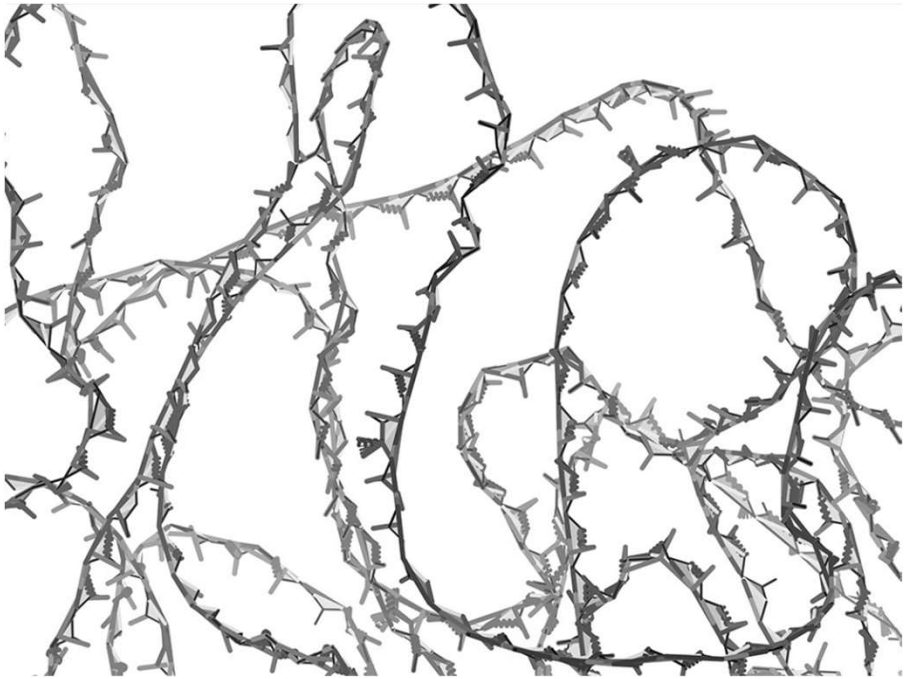
The low-pLDDT regime contains multiple behaviors





Terwilliger, Thomas C., et al. "AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination." Nature Methods 21.1 (2024): 110-116.

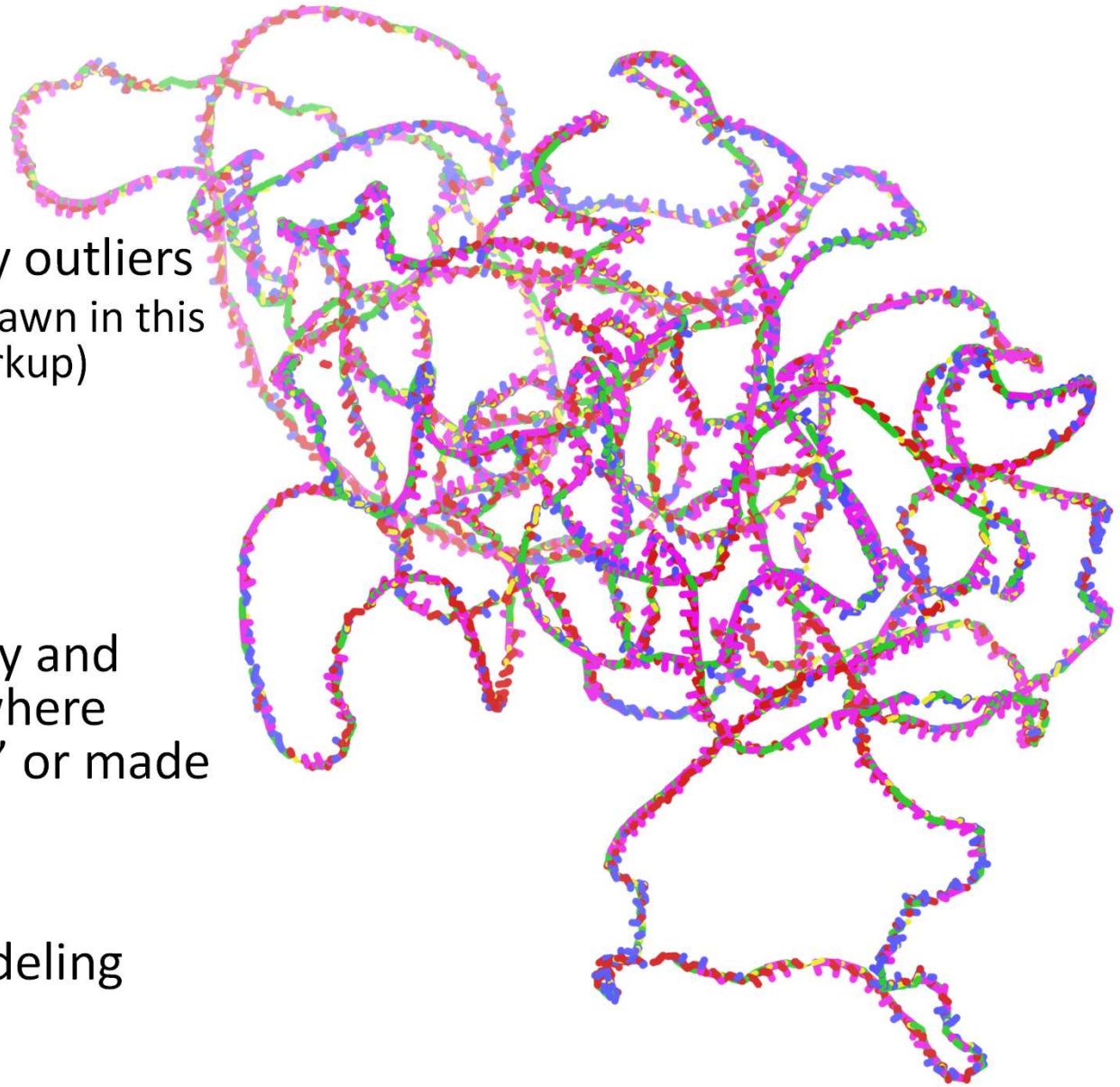
Low pLDDT - Barbed wire



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

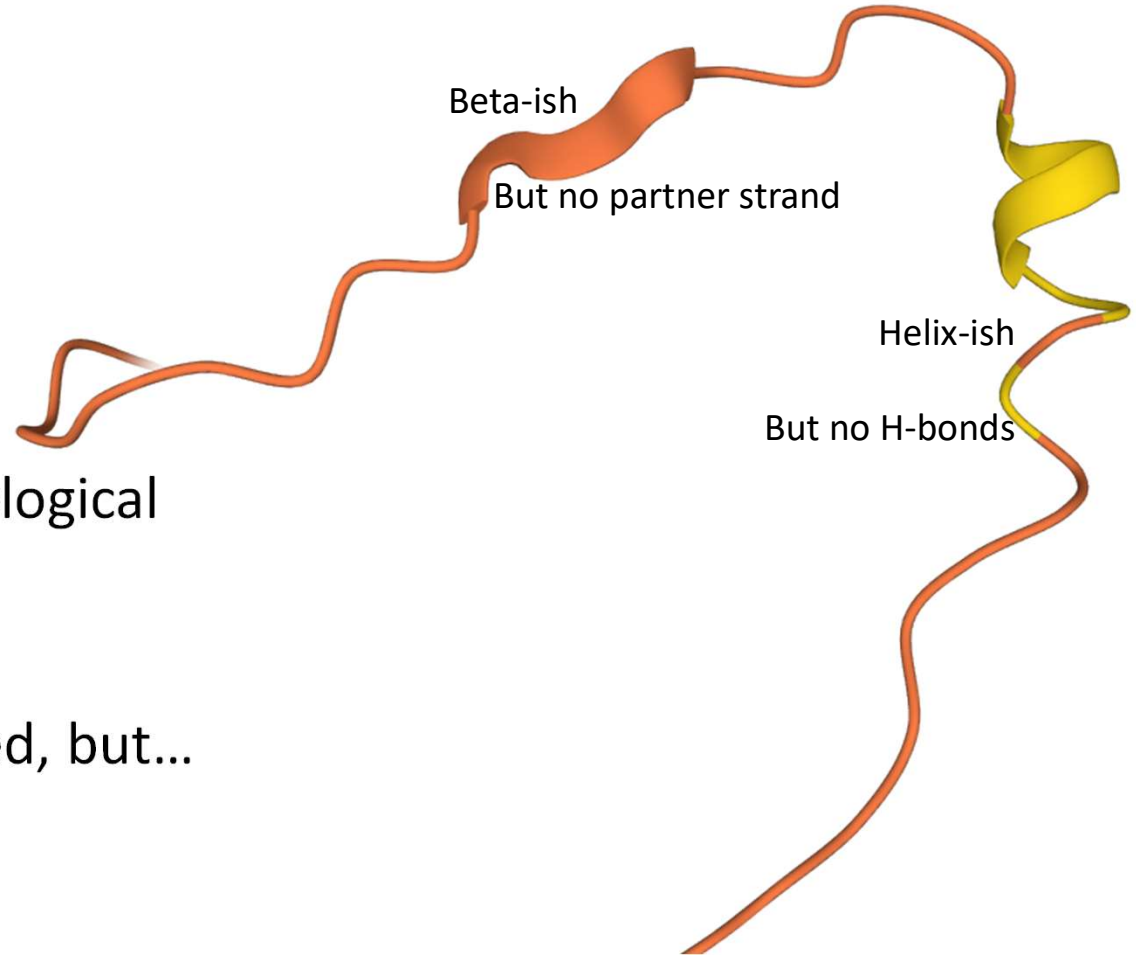
Barbed wire

- 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 has “hallucinated” or made no prediction
- Different from “normal” modeling errors



Unpacked Possible

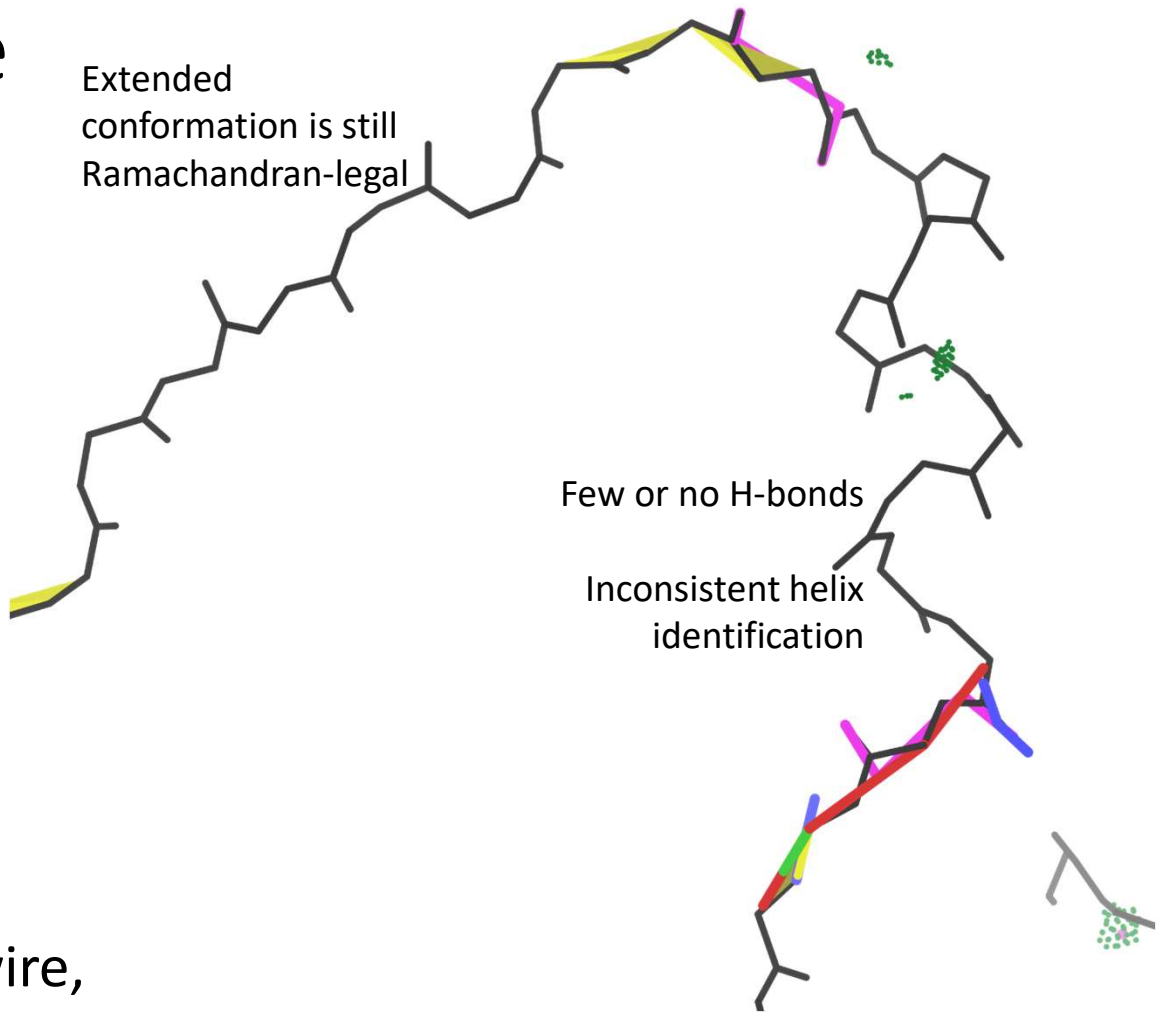
- Somewhat protein-like conformations
- Possibly folded in full biological context
- Unpacked and unidealized, but…



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

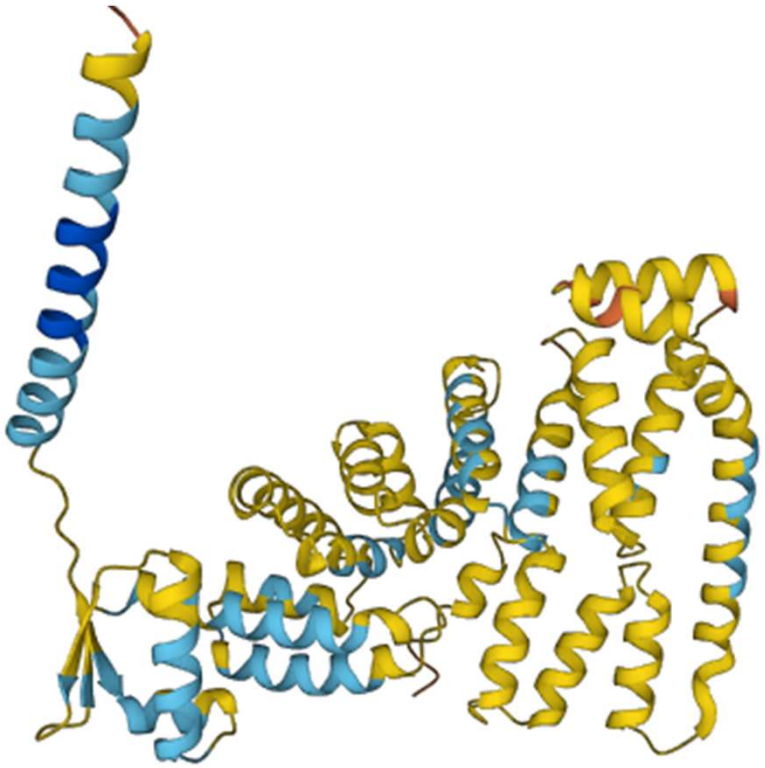
Unpacked Possible

- Lacks validation outliers!
- Also lacks good hydrogen bonding
- More “real” than barbed wire, but no predictive value in most cases. Probably a “hallucination”



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

Near-predictive

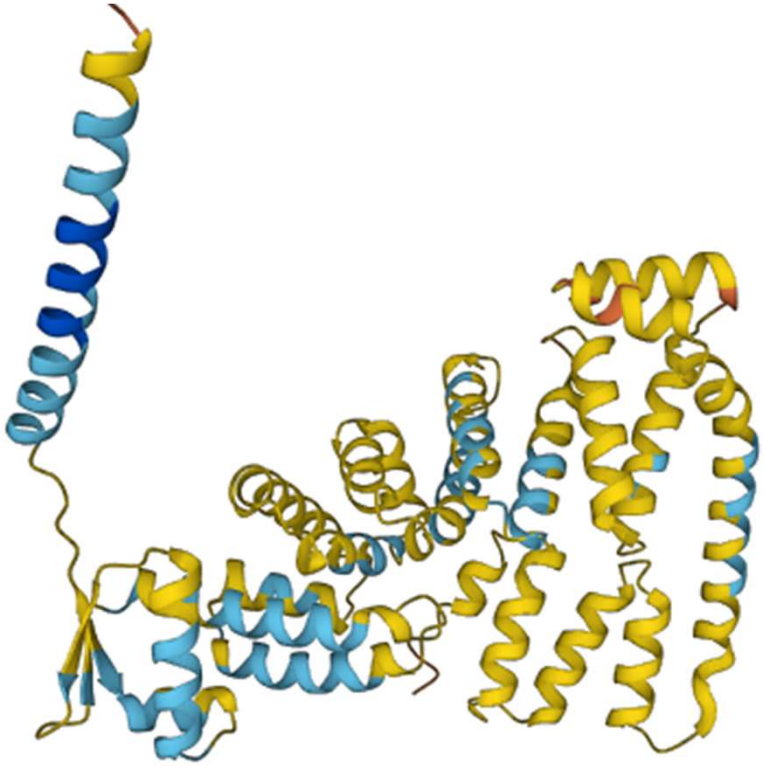


- Low pLDDT, but . . .

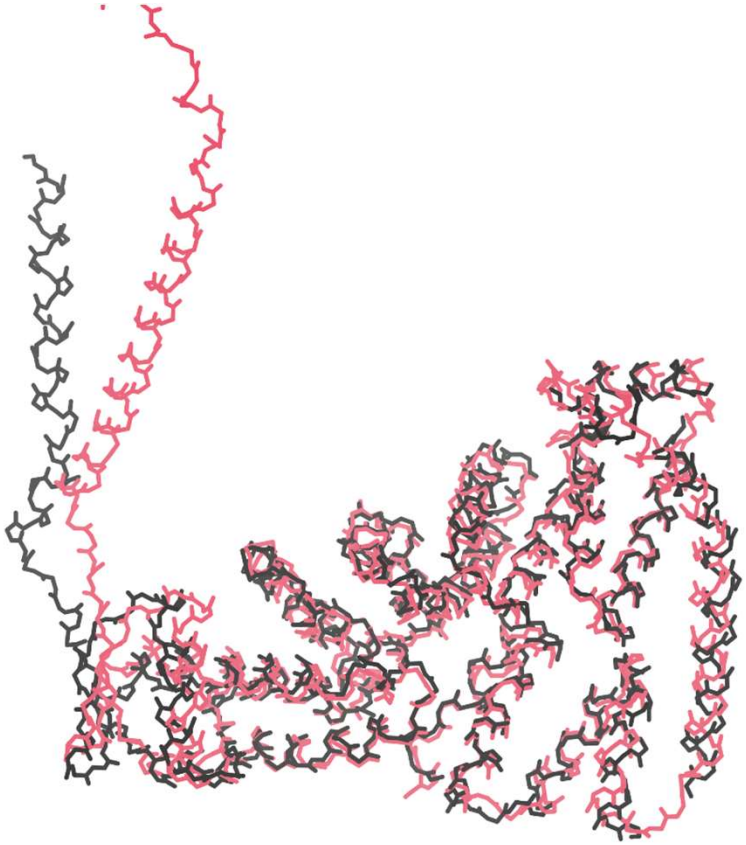
- Well-packed
- Protein-like fold
- Protein-like local geometry

Homo sapiens
Uniprot **P60228**

Near-predictive



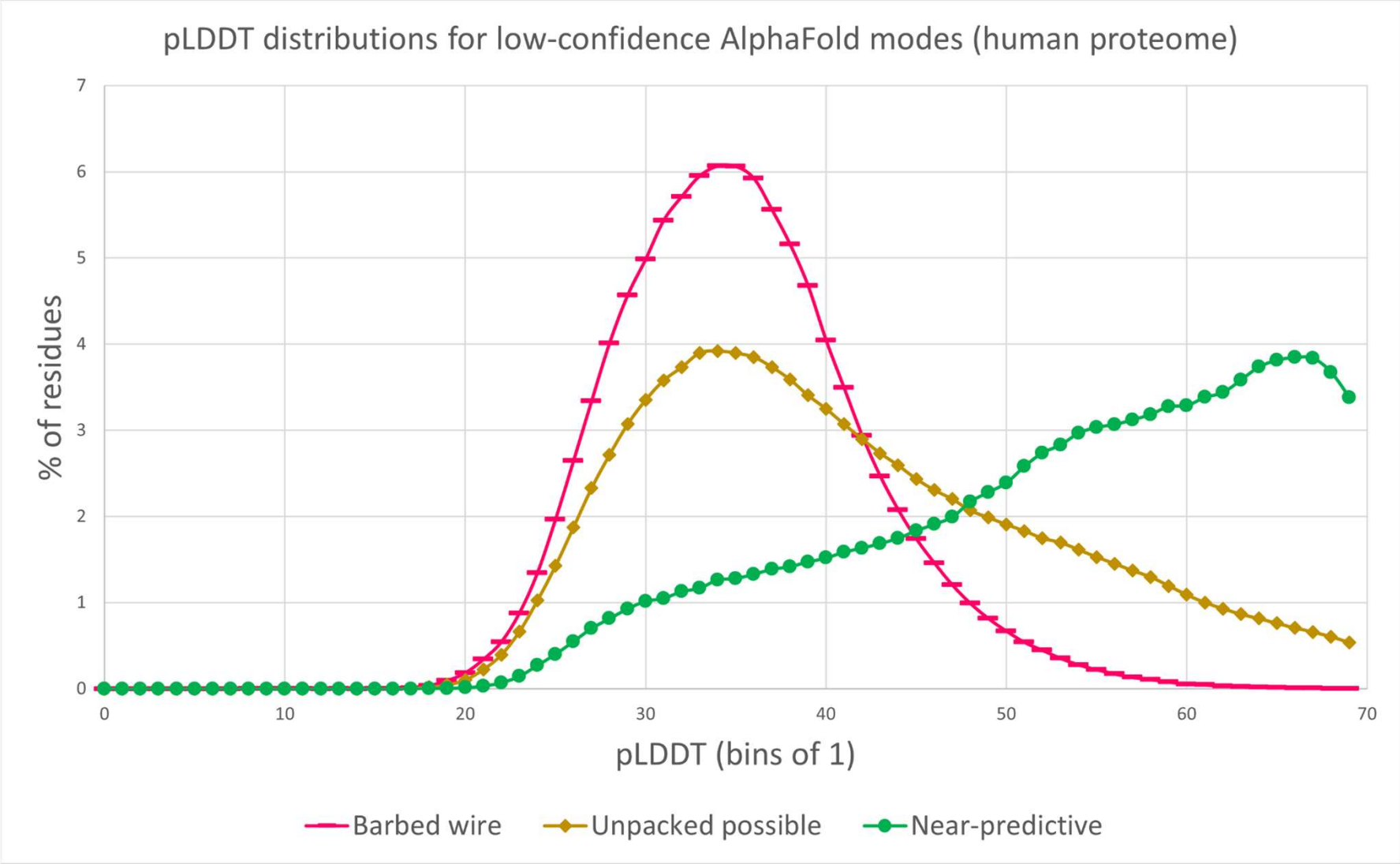
Homo sapiens
Uniprot **P60228**



6zon.pdb, chain E

P60228 AlphaFold prediction

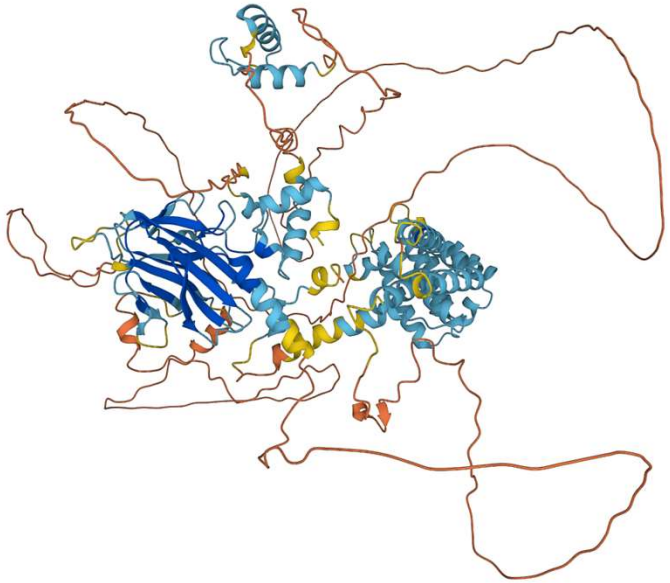
pLDDT comparison



Low pLDDT contains multiple behaviors

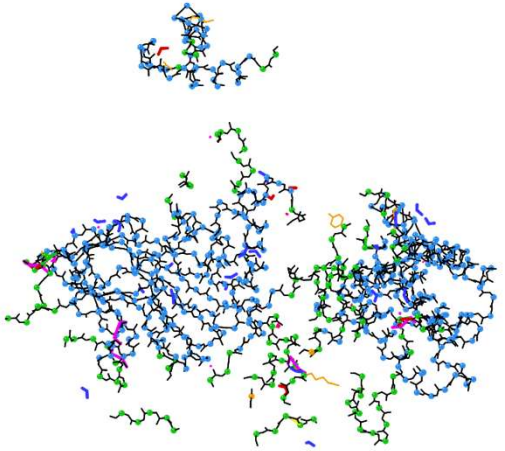
Protein-like regions with pLDDT ~45-70 *may* still be usable!

Whole-model statistics may be misleading



Clashscore, all atoms:	0.54	
Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.		
Poor rotamers	27	3.12%
Favored rotamers	791	91.55%
Ramachandran outliers	133	13.91%
Ramachandran favored	702	73.43%
Rama distribution Z-score	-3.50 ± 0.24	
MolProbity score^	1.87	
Cβ deviations >0.25Å	72	7.97%
Bad bonds:	0 / 7731	0.00%
Bad angles:	241 / 10452	2.31%
Cis Prolines:	3 / 28	10.71%
Cis nonProlines:	30 / 929	3.23%
Twisted Peptides:	152 / 957	15.88%
CaBLAM outliers	149	15.6%
CA Geometry outliers	144	15.09%
Tetrahedral geometry outliers	10	

Barbed wire present, validation says “probably unusable”



0.54	
per 1000 atoms.	
7	1.34%
509	97.32%
4	0.75%
505	94.22%
-0.75 ± 0.33	
1.17	
7	1.28%
0 / 4757	0.00%
30 / 6407	0.47%
0 / 18	0.00%
1 / 554	0.18%
6	1.2%
1	0.20%
0/707	

Barbed wire removed, validation says “needs work”

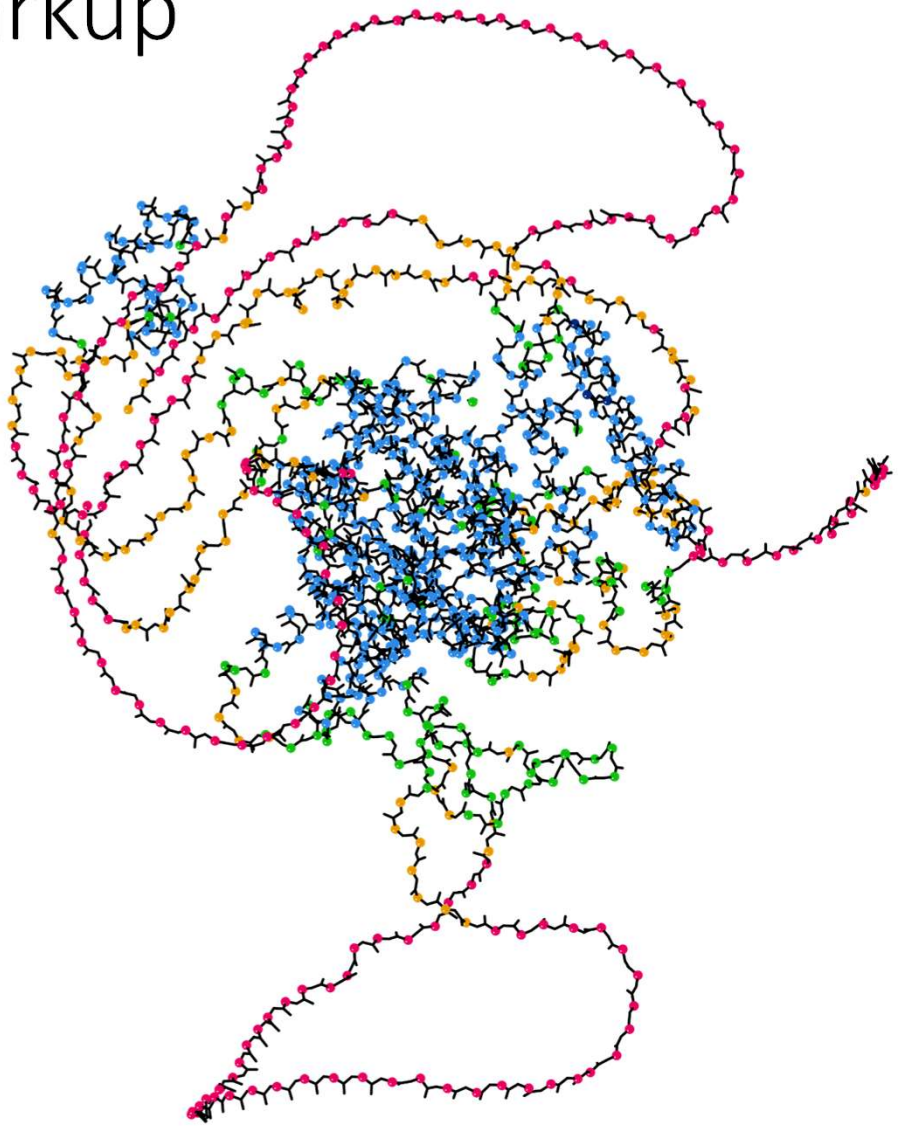
Low-pLDDT tool in Phenix

- 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

What about AlphaFold3?

- This presentation concerns AlphaFold2
- AlphaFold3 has now been released
  - Abramson, J., Adler, J., Dunger, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024). https://doi.org/10.1038/s41586-024-07487-w
- Offers centralized support for predicting ligands, multimers, modified residues, etc.
- Improves pLDDT accuracy for “near-folded” regions
- AF3 is not yet available in a form we can use for iterative prediction
- Stay tuned for developments

The Phenix Project

Lawrence Berkeley Laboratory

Paul Adams, Pavel Afonine, Dorothee Liebschner, Nigel Moriarty, Billy Poon, Oleg Sobolev, Christopher Schlicksup

University of Cambridge

Randy Read, Airlie McCoy, Alisia Fadini

Los Alamos National Laboratory
New Mexico Consortium

Tom Terwilliger, Li-Wei Hung

UTHealth

Matt Baker

Duke University

Jane Richardson, Vincent Chen, Michael Prisant, Christopher Williams

An NIH/NIGMS funded Program Project

Liebschner D, *et al.*, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*. Acta Cryst. 2019 **D75**:861–877

Sample workflows

X-ray structure determination with AlphaFold

Sequence file

Predict structure

Trim, identify domains

X-ray intensity data (mtz file)

High-confidence domains

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

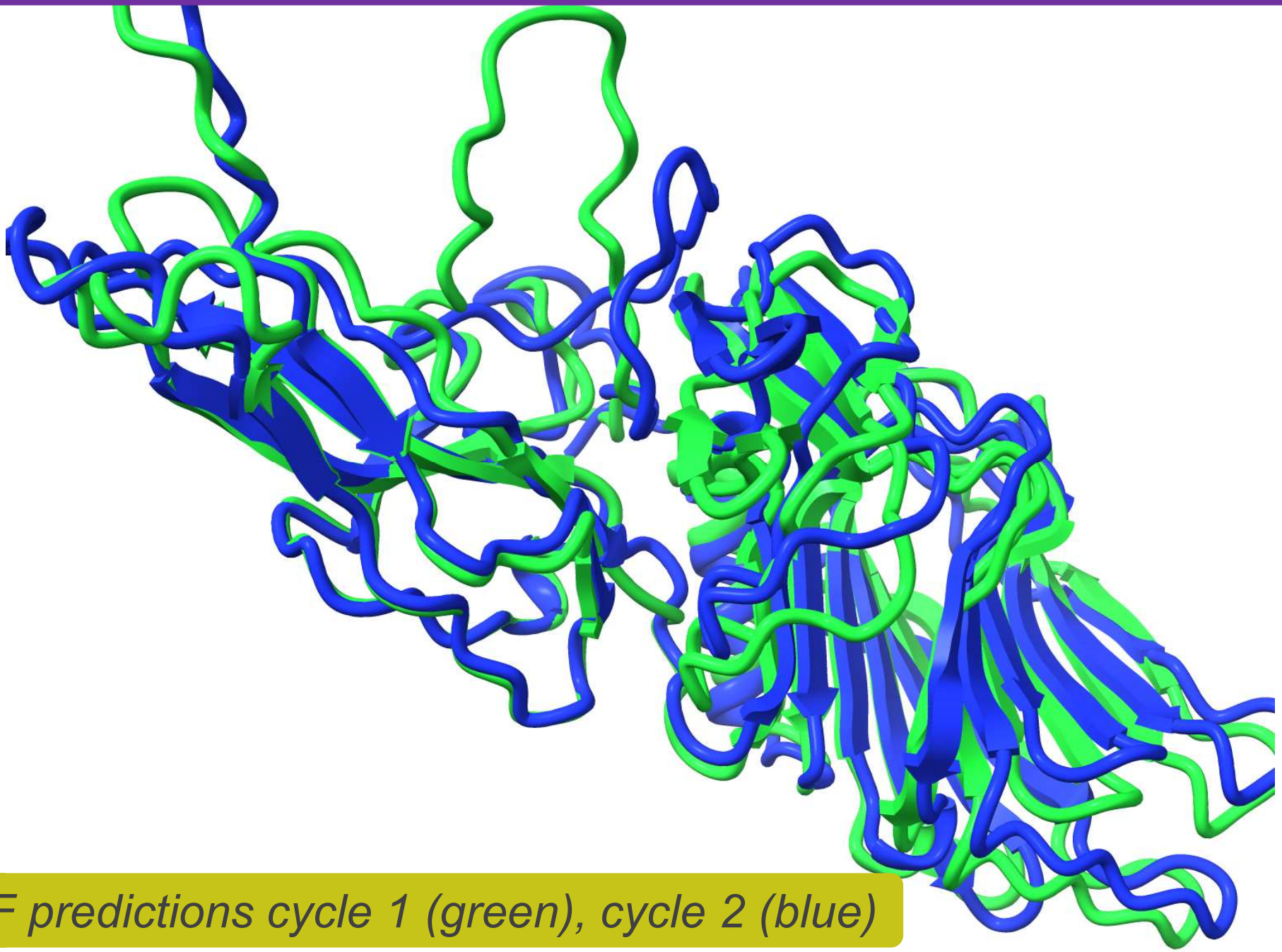
Docked predicted models

Map and model ready for next steps

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)