X-ray and cryo-EM structure solution strategies
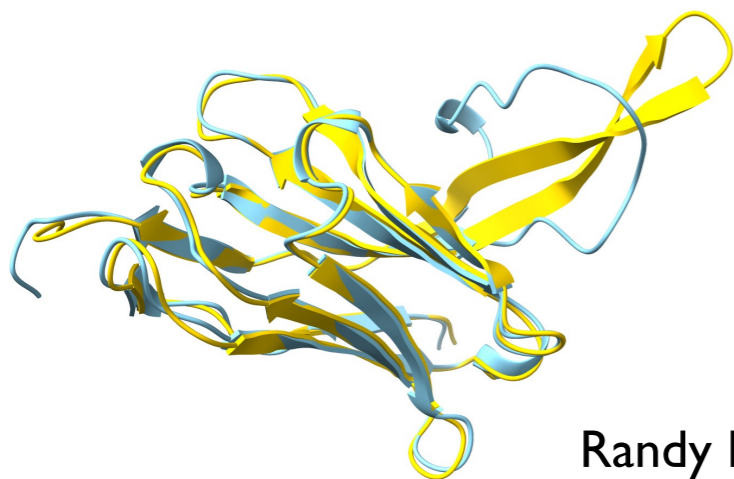*Taking advantage of AlphaFold models*

Tom Terwilliger

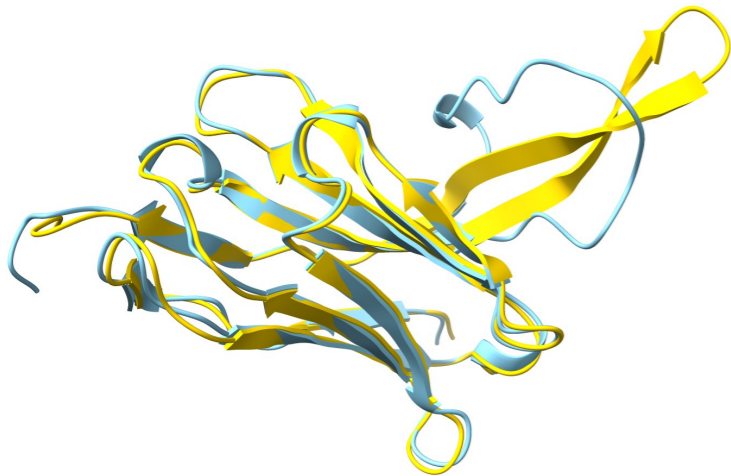The New Mexico Consortium
Los Alamos National Laboratory

Randy Read, Tristan Croll, Claudia Millán, (Cambridge, University), Paul Adams, Billy Poon, Pavel Afonine, Christopher J. Schlicksup (Lawrence Berkeley National Laboratory); Jane Richardson (Duke University)
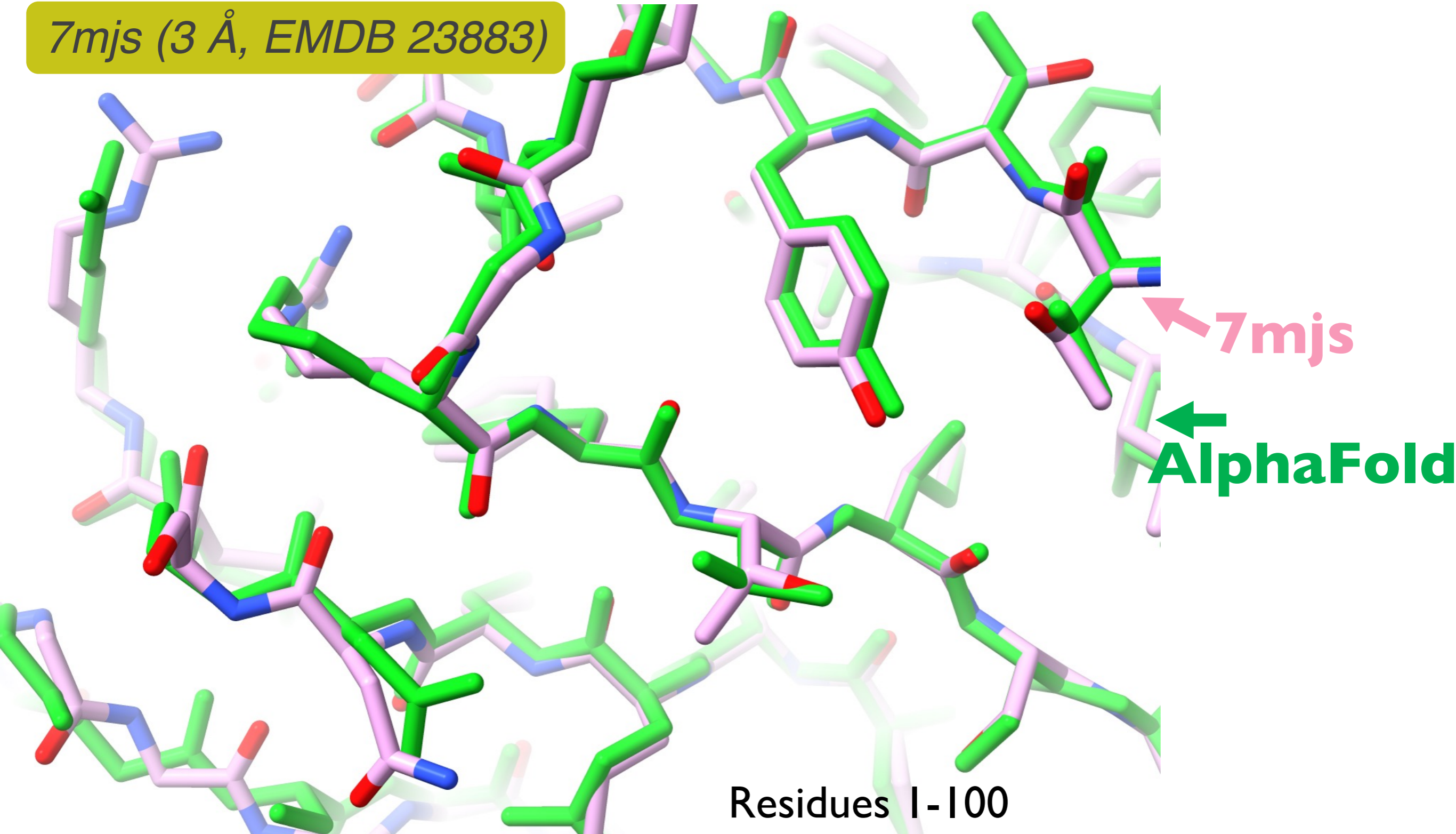
Why use AlphaFold models?

→ *They are great* **hypotheses** *for protein structures*

Models are accurate where sequence coverage is high

7mjs (3 Å, EMDB 23883)

← 7mjs

← AlphaFold

Residues 1-100
High sequence coverage and confidence

→High confidence residues can be accurate

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

...and less accurate where sequence coverage is low

7mjs (3 Å, EMDB 23883)

7mjs

AlphaFold

Residues 100-120
Low sequence coverage, low confidence, low accuracy

Residues 1-100
High sequence coverage and confidence
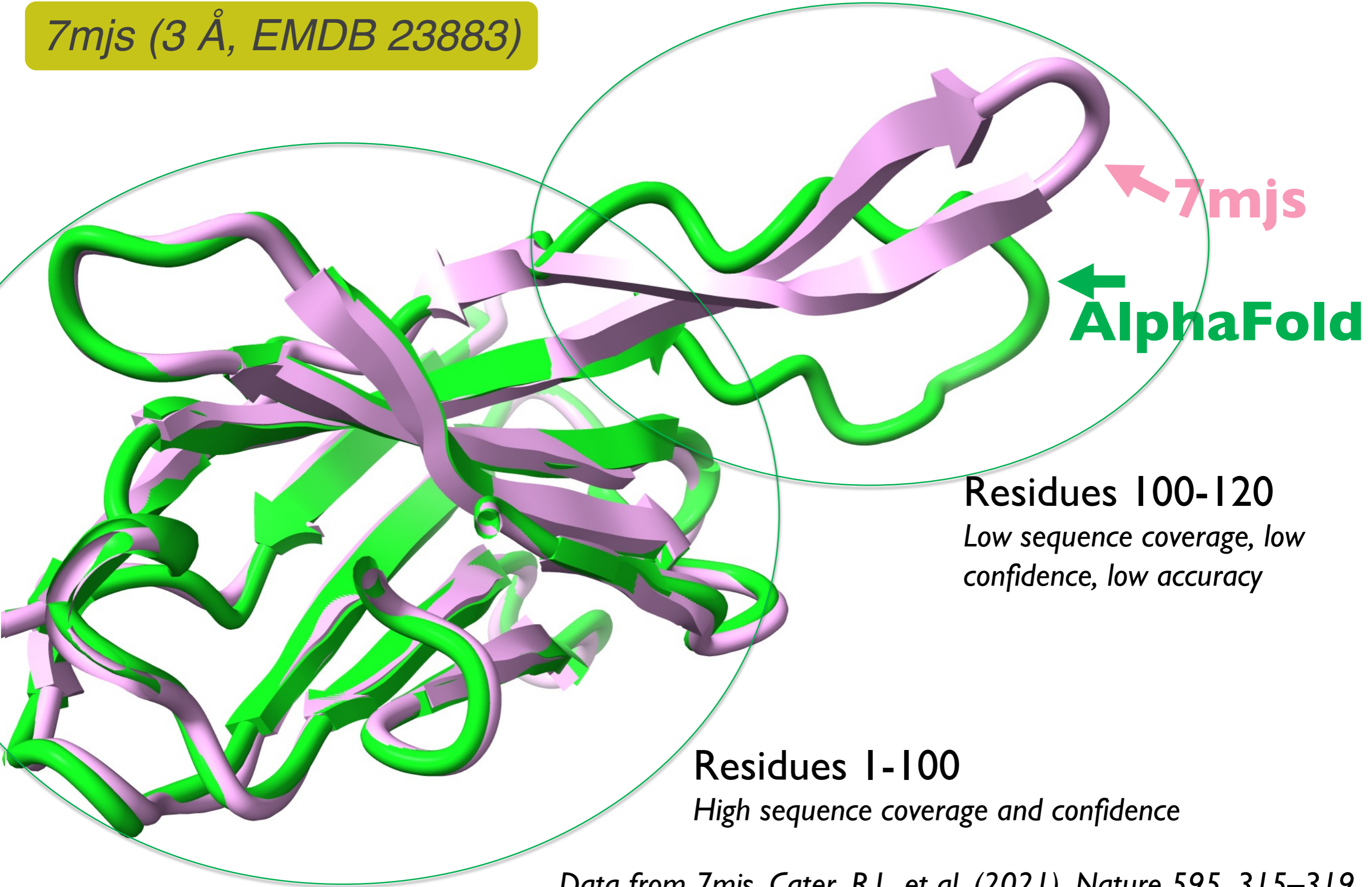
Data from 7mjs, Cater, R.J., et al. (2021). Nature 595, 315–319

Limitations

Only protein → **No water, ions, covalent modifications, carbohydrates, ligands, DNA, RNA**

Trained on good and poor structures → **Parameters may systematically include poor geometry**

Little information about residues that are far apart → **Models may have distortions and incorrect domain relationships**
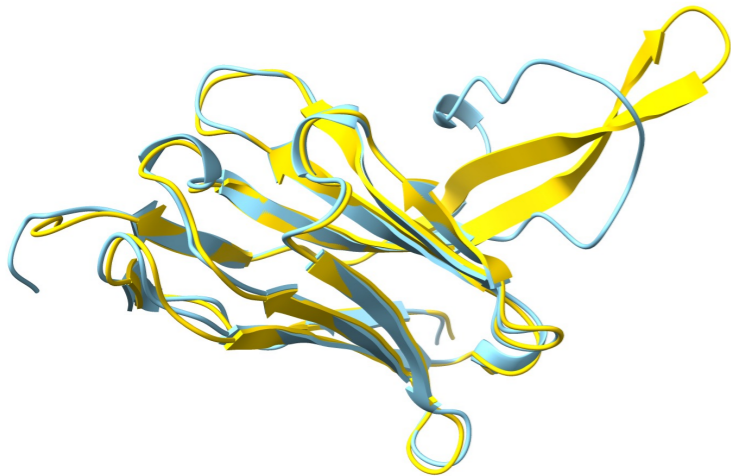
AlphaFold models are great for jump-starting structure determination

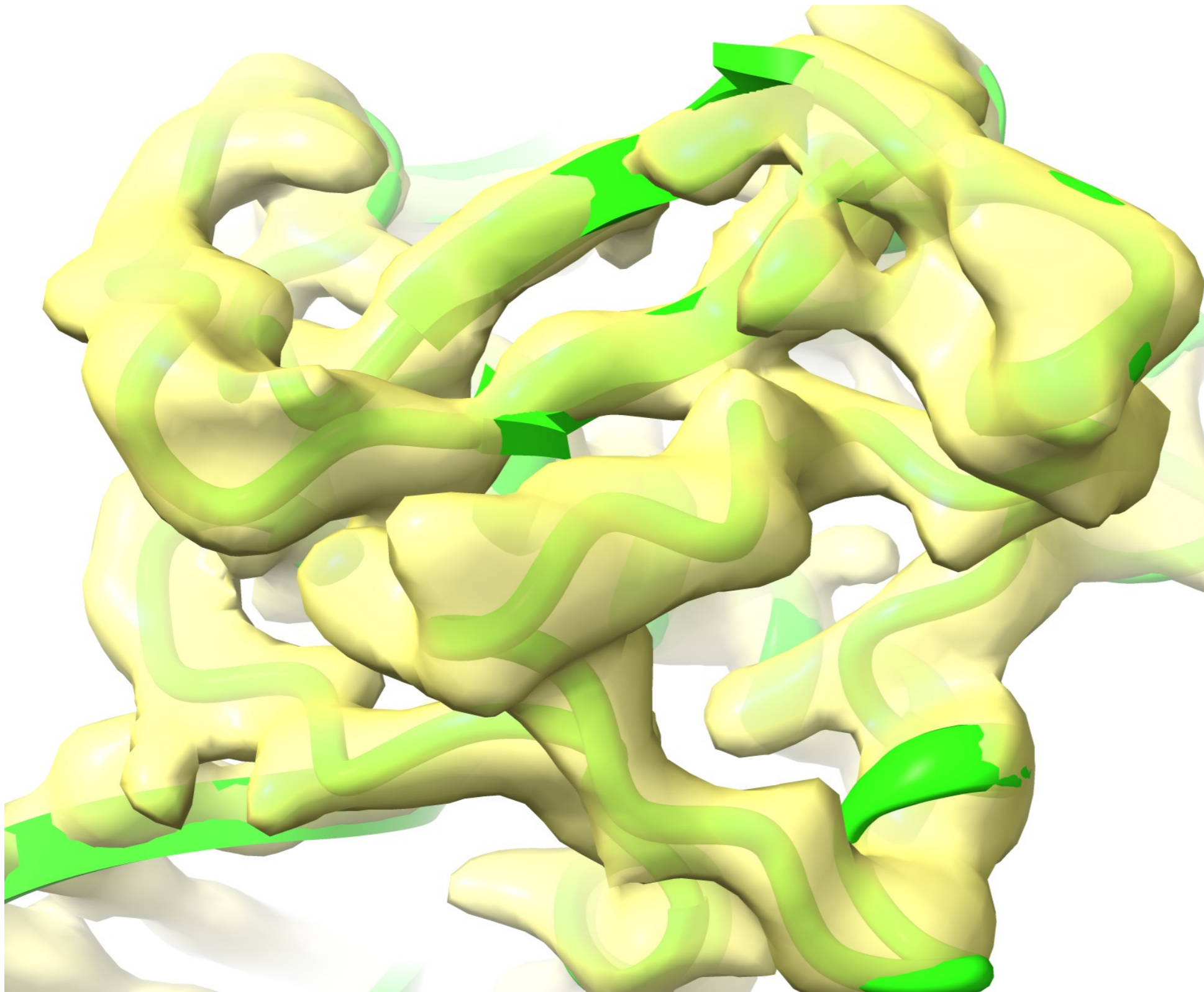*Example:* *Finishing a difficult crystal structure*

Repressor – DNA complex, solved with 2.6 Å SeMet SAD data and refined against 3.1 Å native data

Before AlphaFold, R/Rfree = 0.27/0.29

*AlphaFold model: A **hypothesis** about this structure*

After AlphaFold, R/Rfree = 0.21/0.24
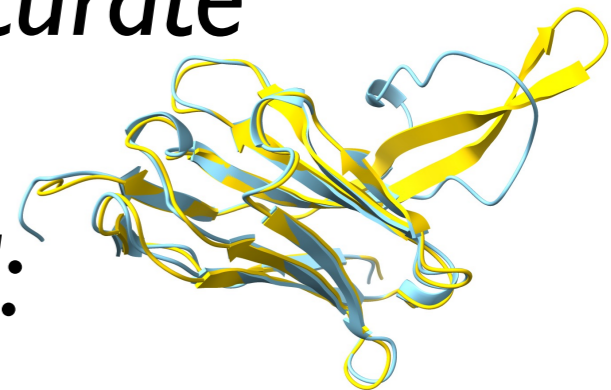*(it was a good hypothesis)*

Jamie Wallen, Western Carolina University

AlphaFold models are great for jump-starting structure determination

➡️ *Confidence measure (plDDT) allows pruning of worst parts of models*

➡️ *High-confidence parts are often accurate*

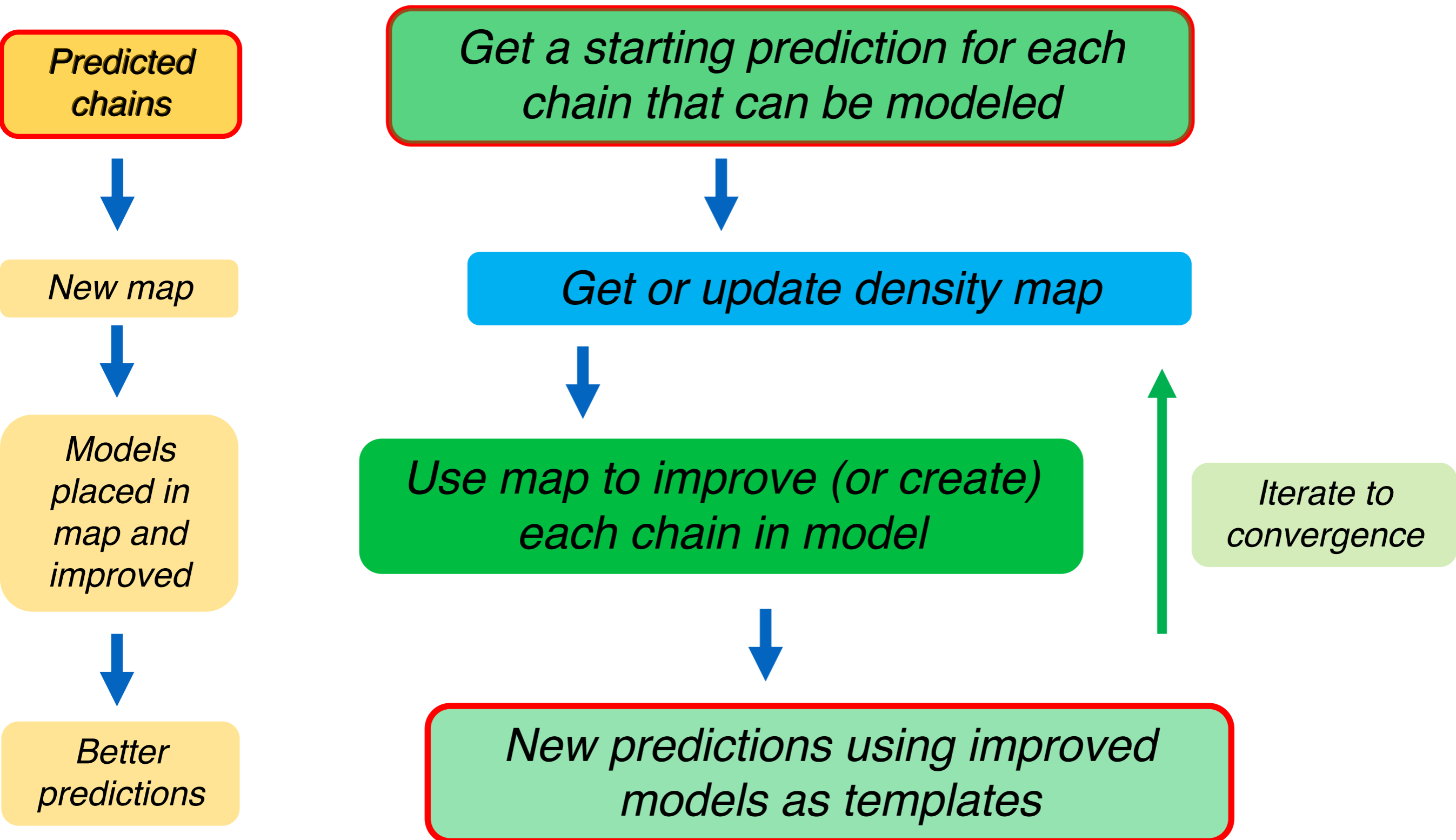*Better than a homology model:*

➡️ *No insertions and deletions in the sequence*

➡️ *High accuracy of entire domains helps MR*

Strategy for initial structure determination with AlphaFold
(Getting an accurate full model for each chain in structure)

Predicted chains

New map

Models placed in map and improved

Better predictions

Get a starting prediction for each chain that can be modeled

Get or update density map

Use map to improve (or create) each chain in model

New predictions using improved models as templates

Iterate to convergence
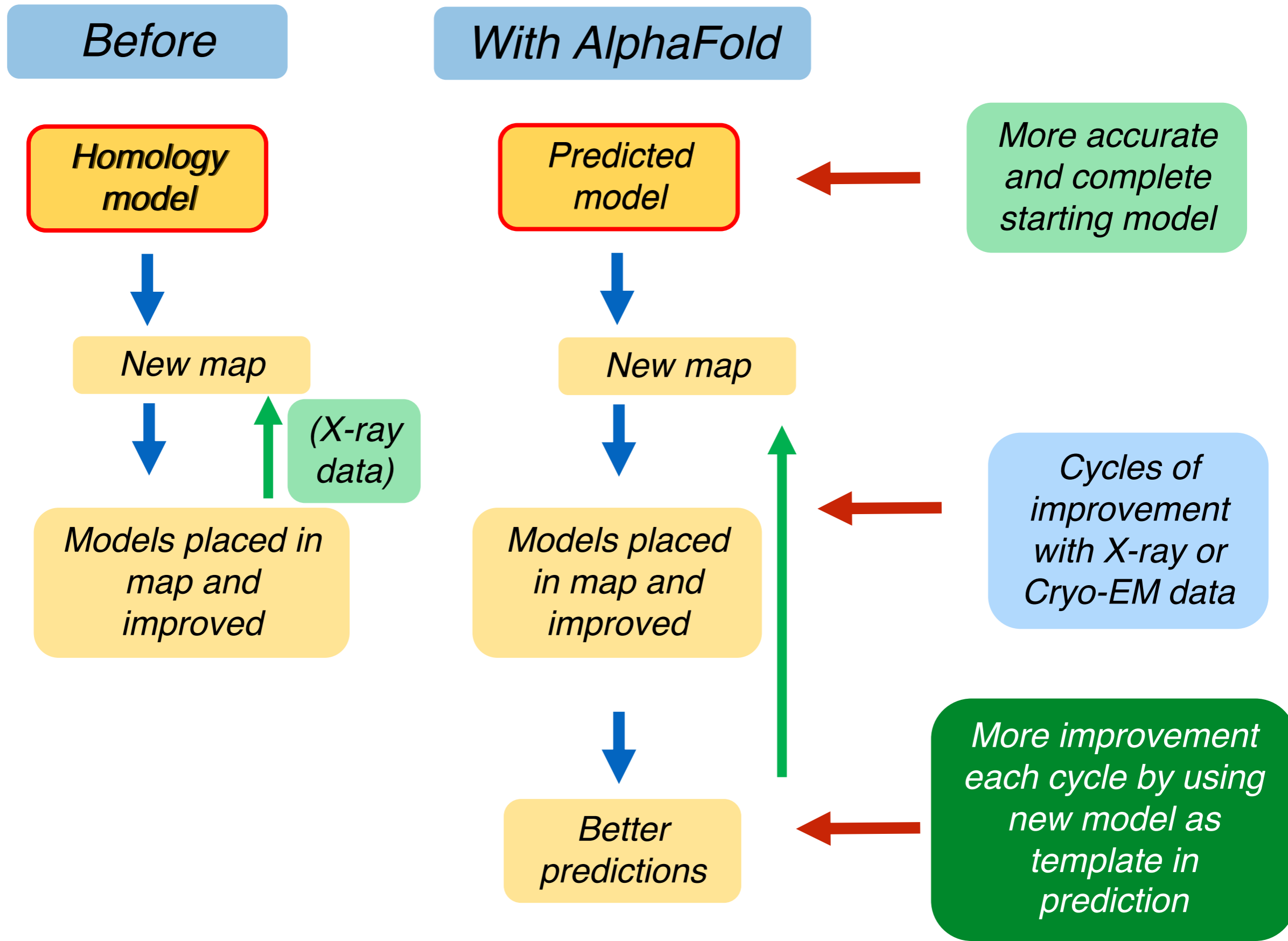
Fully automatic initial structure determination with AlphaFold
(Phenix PredictAndBuild, X-ray data)

X-ray intensity data (mtz file)

Contents of asymmetric unit (sequence file)

(wait 2-24 h)

Docked predicted models

Rebuilt model

Useful as high-quality reference models

Preliminary model ready for next steps

Phenix tools for generating and rebuilding AlphaFold models

GUI/ Colab/ Phenix Server tools

PredictModel
Predict structures of all chains in a sequence file

PredictAndBuild (X-ray or cryo-EM)
Predict, dock or run MR and rebuild all chains in structure with iteration
(sequence file and X-ray intensities or cryo-EM half-maps)

GUI tools

ProcessPredictedModel
Interpret plDDT values as B and trim

DockAndRebuild
Dock and rebuild one chain

Notes on AlphaFold prediction and PredictAndBuild

PredictModel

You can supply a template
AlphaFold will use the template in prediction
You do not need an MSA if you supply a template
The template should not be an AlphaFold model

PredictAndBuild (X-ray or cryo-EM)

Your sequence file should contain one sequence for every chain in the structure
For X-ray data, the space group in the data file or specified in the GUI needs to be at least in the same point group as the correct space group

Observations about AlphaFold predictions…

MSA's and templates tell AlphaFold what is close

With a good MSA, skip the templates

With a good template, skip the MSA

All inputs to AlphaFold go in a table with one column per residue

Incorrect alignments in MSA's will make prediction worse

Multiple conformations may yield complex MSA's

Templates that cannot be aligned to sequence are useless