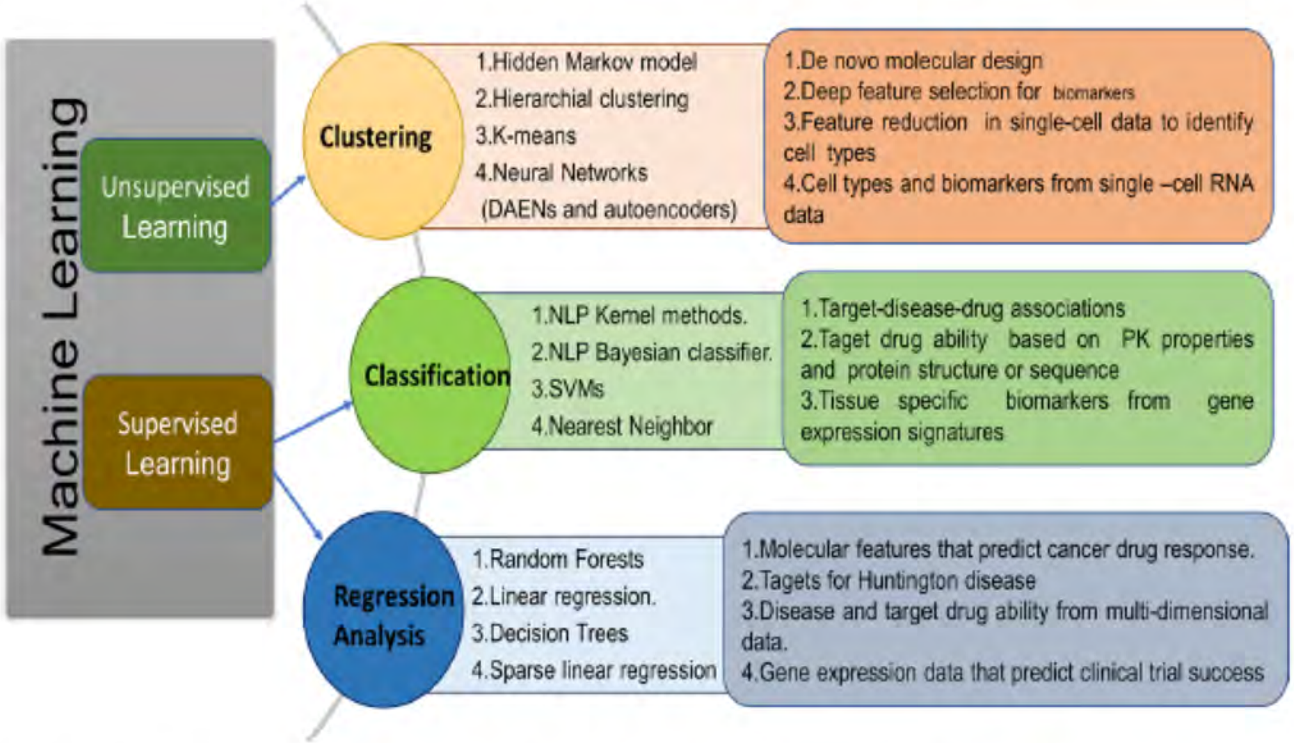
Machine Learning in Structural Biology: Some examples and Introduction to AlphaFold

George N. Phillips, Jr.

Rice University, December 7, 2022

Machine Learning in Drug Discovery

Dara, S et al., Artificial Intelligence Review 55:1947 (2022)



Machine Learning

Unsupervised Learning

Supervised Learning

Clustering

1.Hidden Markov model
2.Hierarchial clustering
3.K-means
4.Neural Networks (DAENs and autoencoders)

1.De novo molecular design
2.Deep feature selection for biomarkers
3.Feature reduction in single-cell data to identify cell types
4.Cell types and biomarkers from single –cell RNA data

Classification

1.NLP Kernel methods.
2.NLP Bayesian classifier.
3.SVMs
4.Nearest Neighbor

1.Target-disease-drug associations
2.Taget drug ability based on PK properties and protein structure or sequence
3.Tissue specific biomarkers from gene expression signatures

Regression Analysis

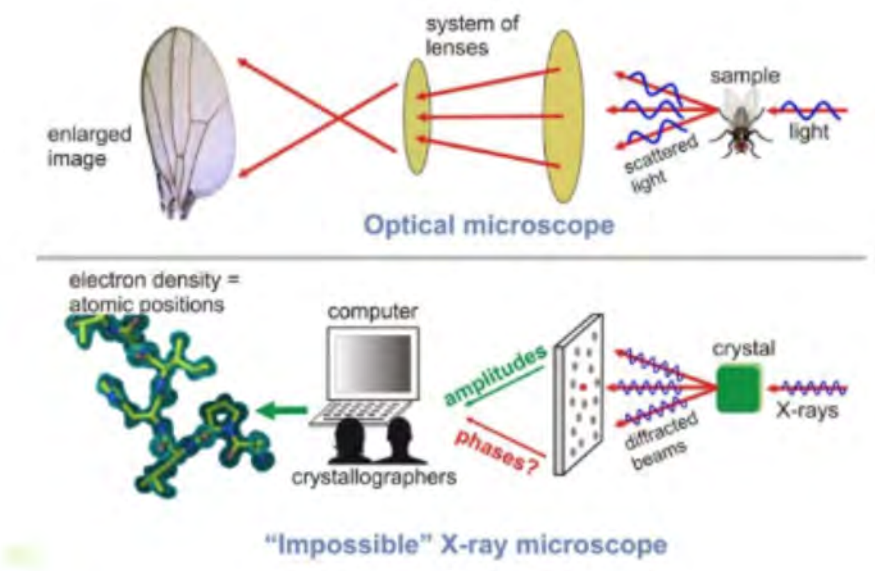
1.Random Forests
2.Linear regression.
3.Decision Trees
4.Sparse linear regression

1.Molecular features that predict cancer drug response.
2.Tagets for Huntington disease
3.Disease and target drug ability from multi-dimensional data.
4.Gene expression data that predict clinical trial success

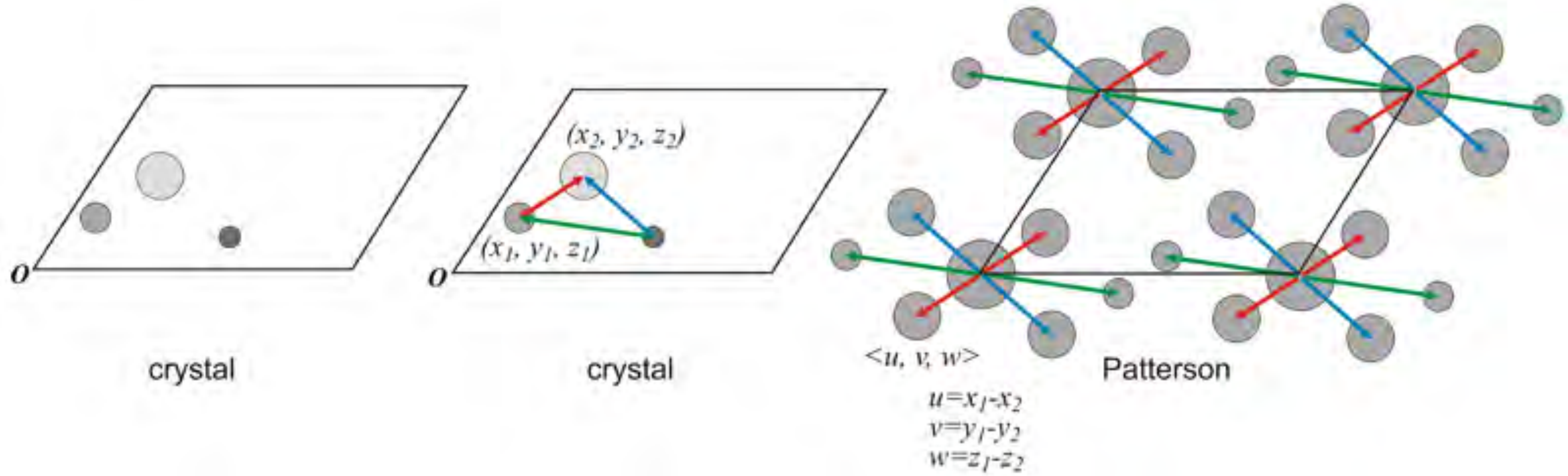
Fig. 2 Applications of AI in Drug discovery depicts the Machine learning mechanisms

Deep Learning for Solving the Phase Problem in X-ray Crystallography

Can a Neural Network be set up to solve Patterson maps to yield electron density maps?

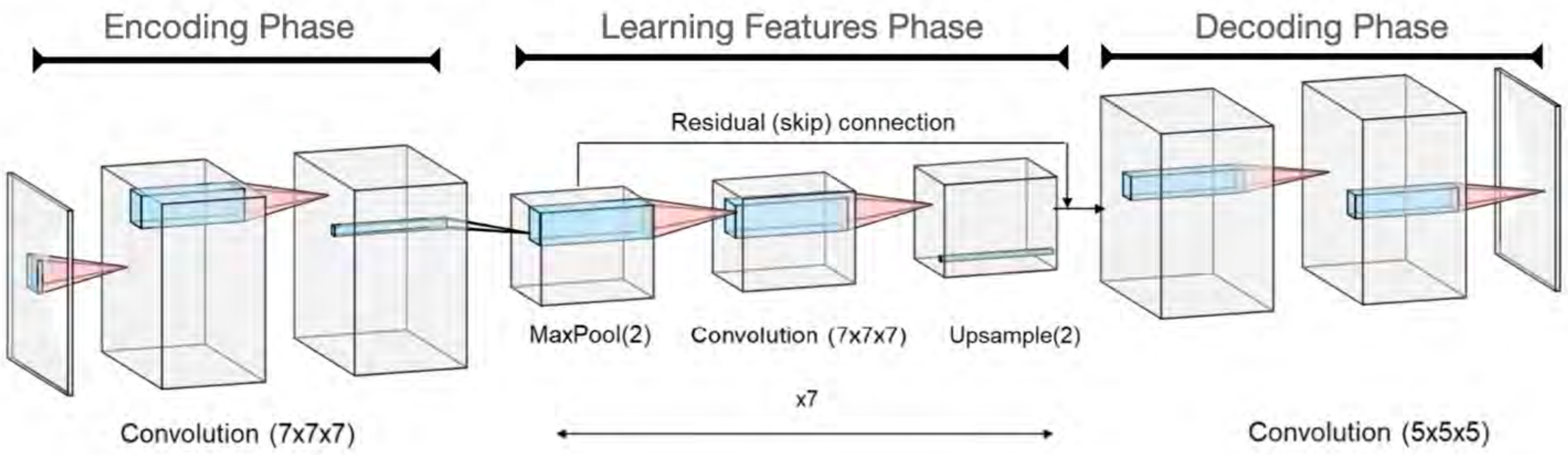


Patterson Maps can be made without phases



$$p(u, v, w) = \frac{1}{V} \cdot \sum_{h,k,l} |F(h, k, l)|^2 \cdot e^{-2\pi i(hu+kv+lw)}$$

Solving Simple Patterson Maps with ML



It works for simple cases

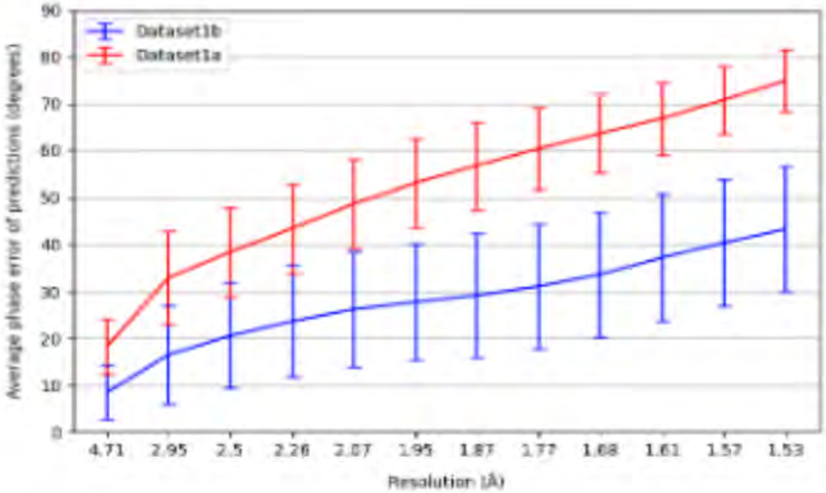
1C75_87

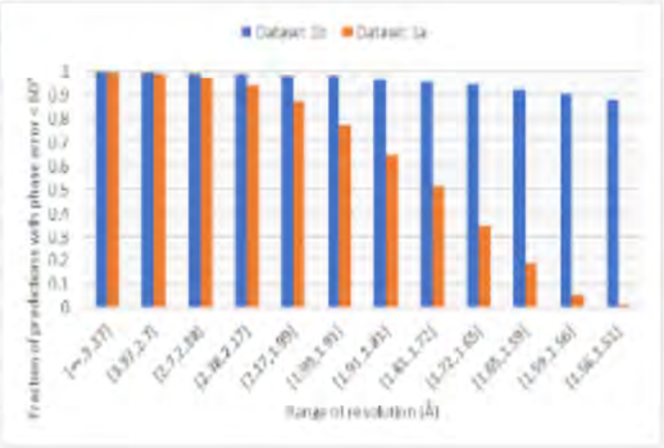
1ETY_17

1HB8_74

Summary of 7,390 validation examples (66,504 for training)

Dialanine phase error





Pre-AlphaFold History

- Protein data bank has 150000+ coordinate sets for protein (and other) structures
- It MIGHT be true that the set of all known single-domain structures are in the PDB
- Folks have recognized the power of co-evolution data for quite a while
- CASP competitions have undergone dramatics improvements of the past six years using co-evolution concepts
- GPU-based computing has taken off for Machine Learning applications
- The structural biology community provides ‘blind’ tests of new structures for CASP

Original AlphaFold (2018)

- Places first in 13th CASP competition (barely)
- Did best when no existing template was available but with good sequence data
- Used co-evolution ideas

AlphaFold2 (2020) for Structure Prediction of Proteins

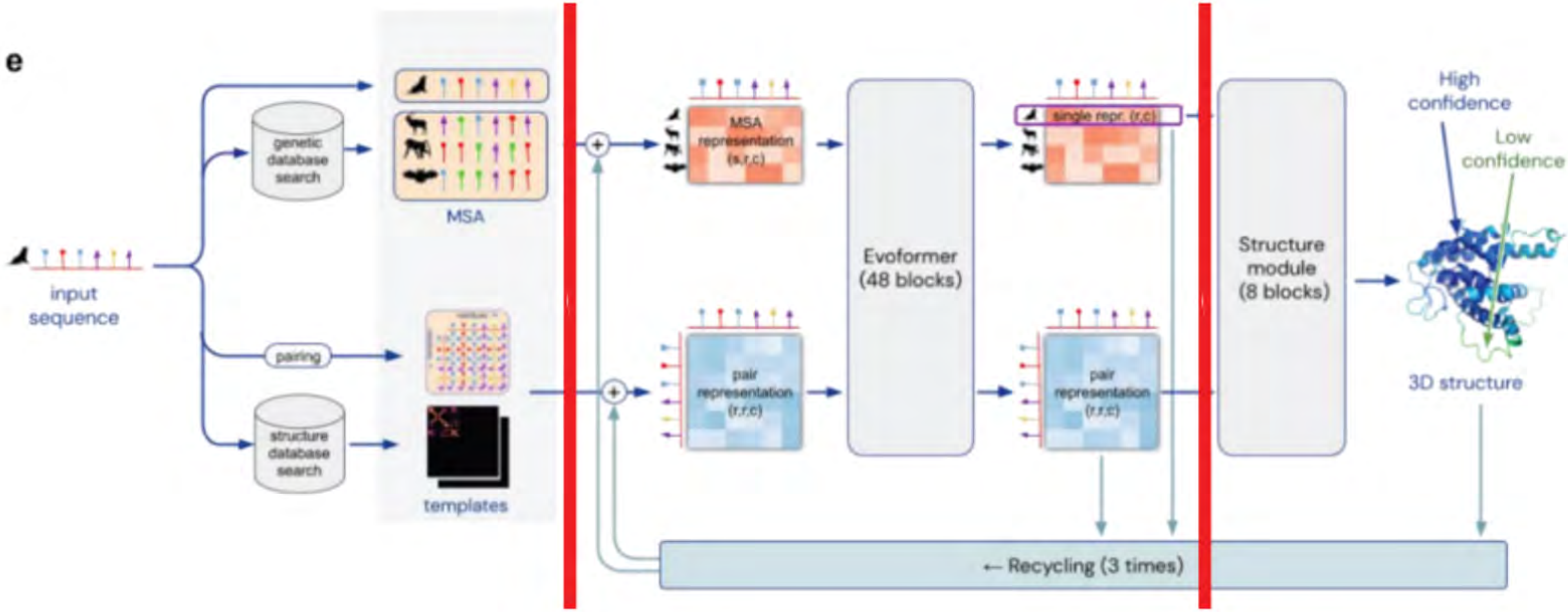
- Also from DeepMind folks at Google
- Performed essentially at the target accuracies set for the entire CASP goal
- Approaches experimental accuracies in many cases
- Method fully published, code available, widely deployed in Structural Biology
- Training involved millions of $ in computing time.

How does it work and what is it good and not good for?

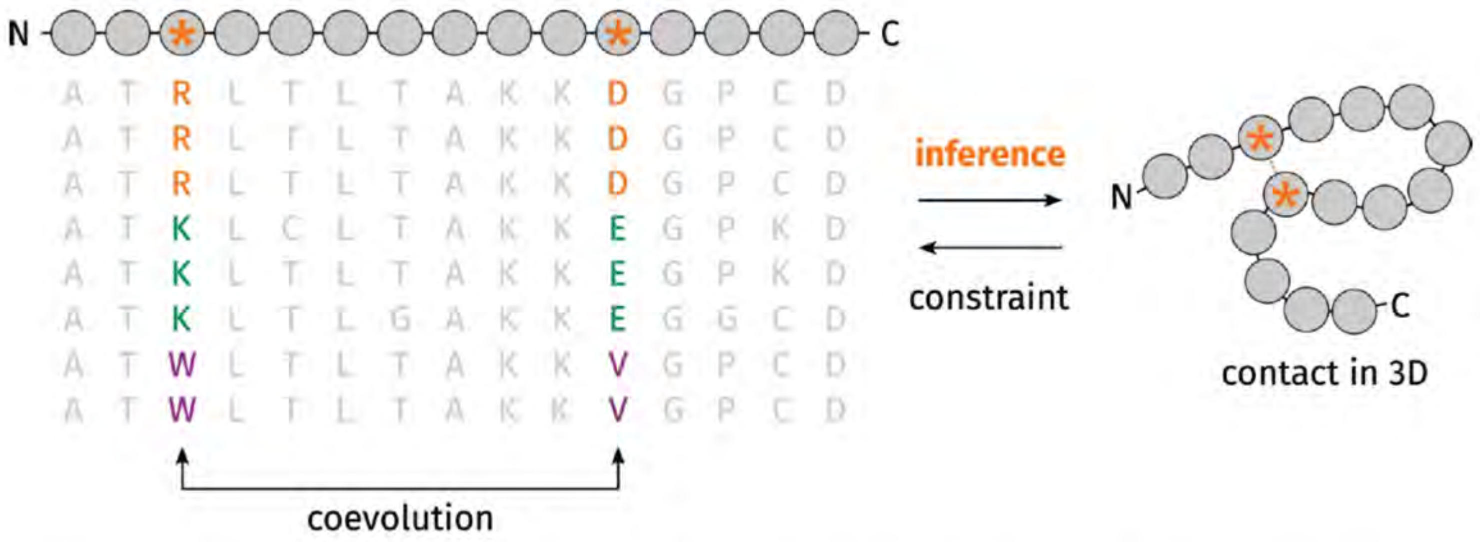
How does it work? (First in words)

- First prepare a multiple sequence alignment (MSA) to your unknown that you want to predict the structure of
- Look for structure templates already deposited in the PDB
- Correlate co-evolution data at sites in the sequence with physical distances in a template structure
- Evolve the MSA co-evolution data to focus in on physical sites
- Evolve the locations corresponding to atomic coordinates as if they were a ‘gas’ of unconnected amino acids
- Use geometrical constraints to connect the individual amino acids into a polypeptide, and ‘refine’ the structure
- Training with added simulated structures (self-distillation)

A peek under the hood of AlphaFold2

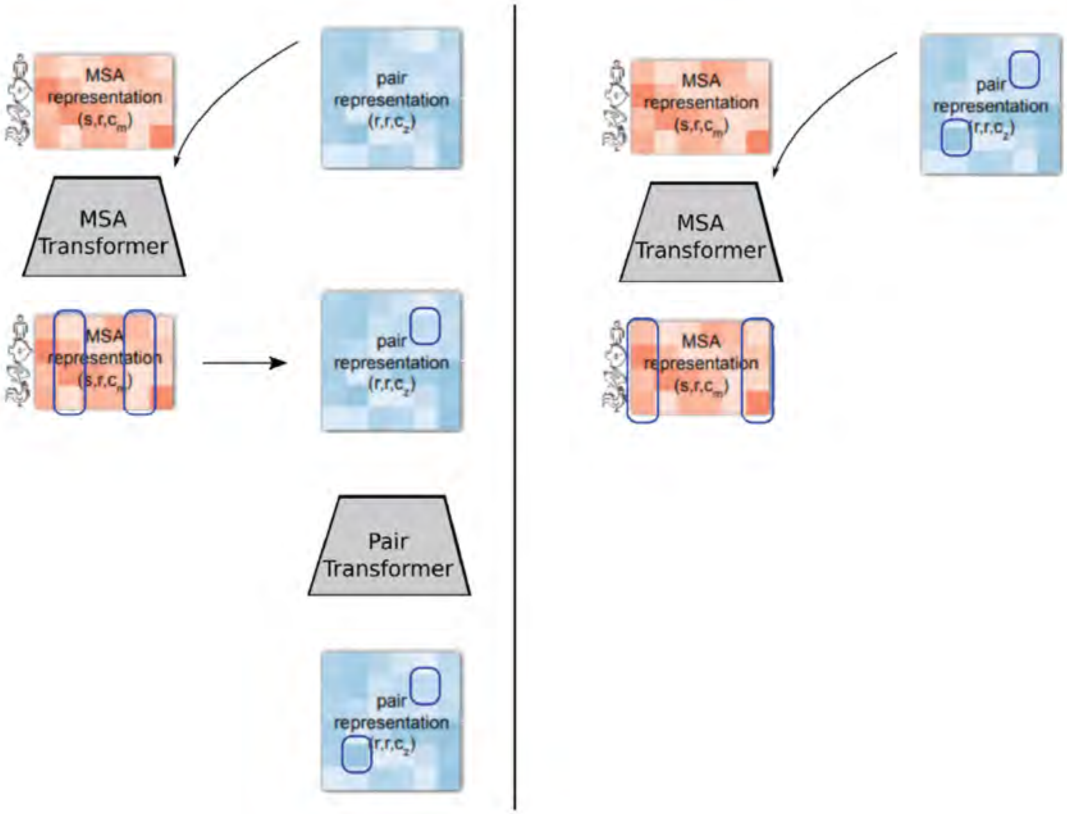


Co-evolution to distance inference

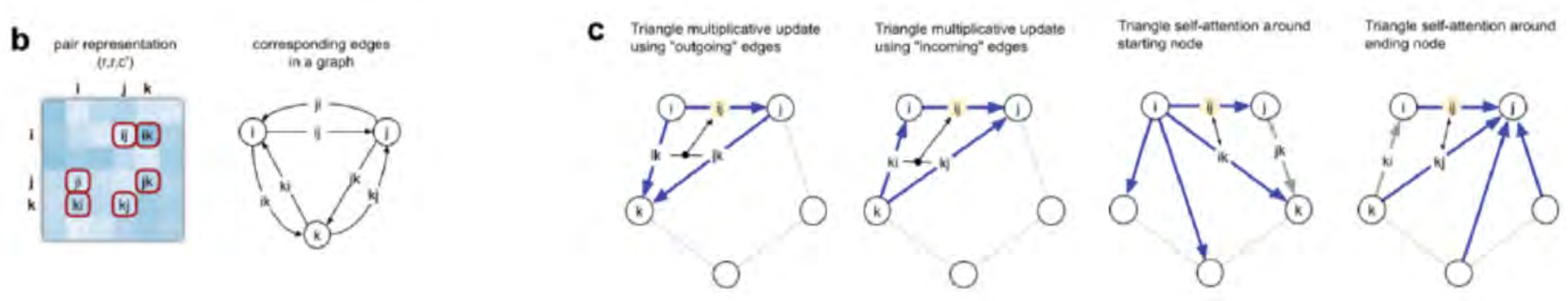


```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

Schematic of how co-evolution methods extract information about protein structure from a multiple sequence alignment (MSA). Image modified from doi: 10.5281/zenodo.1405369, which in turn was modified from doi: 10.1371/journal.pone.0028766

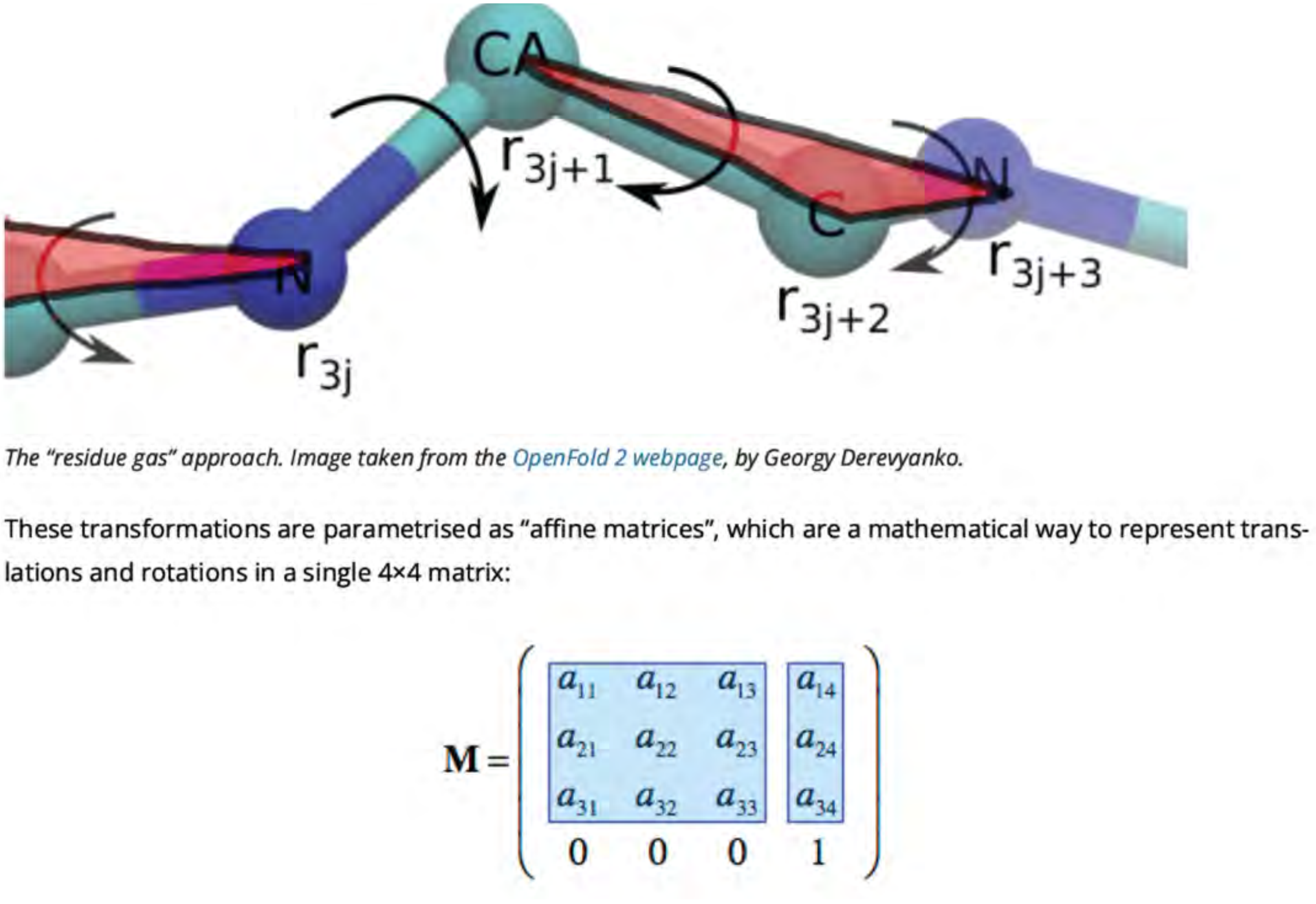

MSA representation (s,r,c_m)
pair representation (r,r,c_z)
MSA Transformer
MSA representation (s,r,c_m)
pair representation (r,r,c_z)
Pair Transformer
pair representation (r,r,c_z)
MSA representation (s,r,c_m)
pair representation (r,r,c_z)
MSA Transformer
MSA representation (s,r,c_m)

Triangle inequality to embed distance info between residue sites



Triangular attention, as published in the Nature paper.

Every residue in the “gas” is modeled as a triangle along the peptide backbone

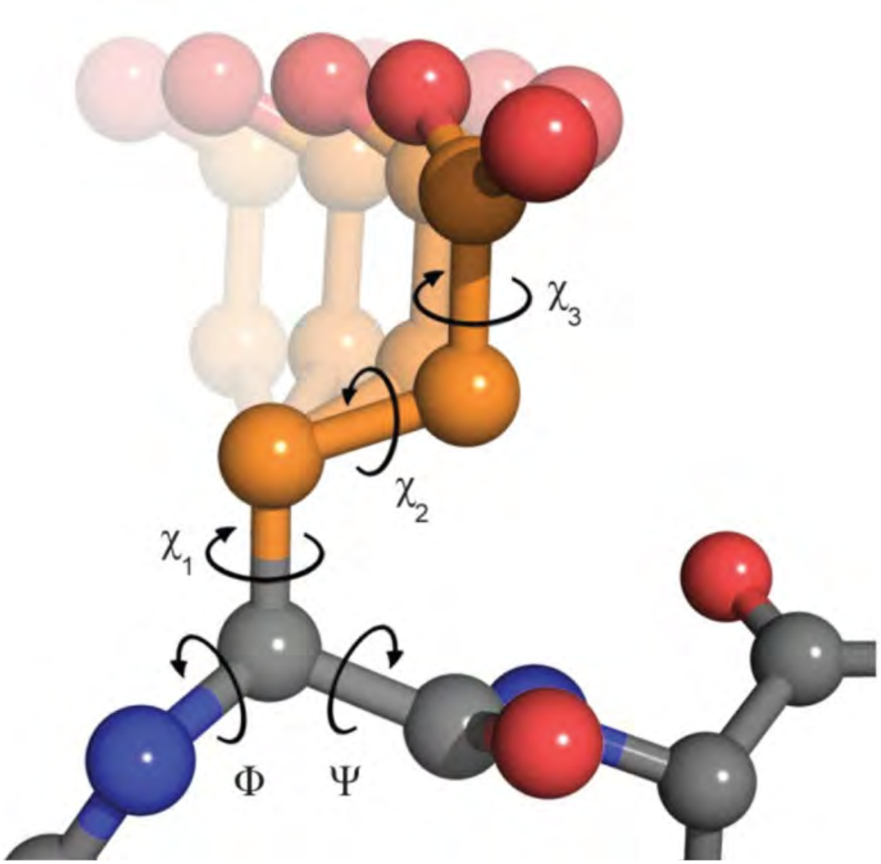


The “residue gas” approach. Image taken from the OpenFold 2 webpage, by Georgy Derevyanko.

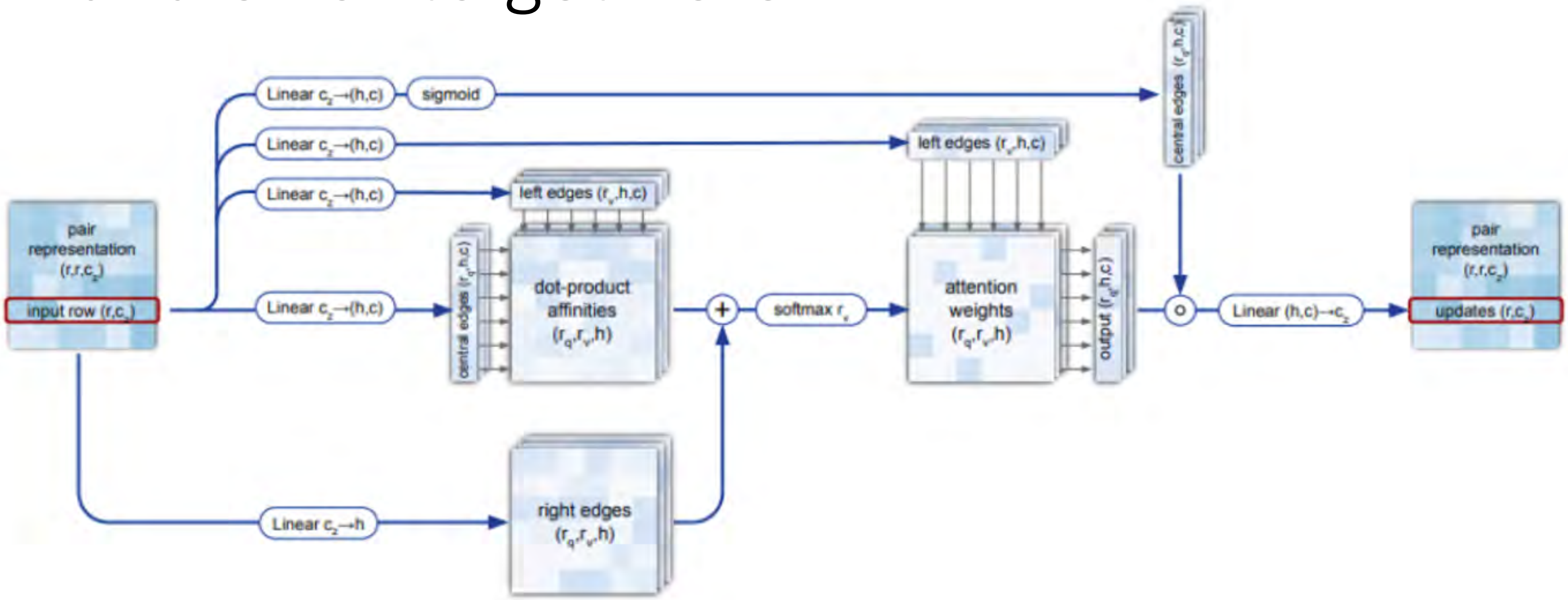
These transformations are parametrised as “affine matrices”, which are a mathematical way to represent translations and rotations in a single 4×4 matrix:

$$\mathbf{M} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Sdie chains are added next....

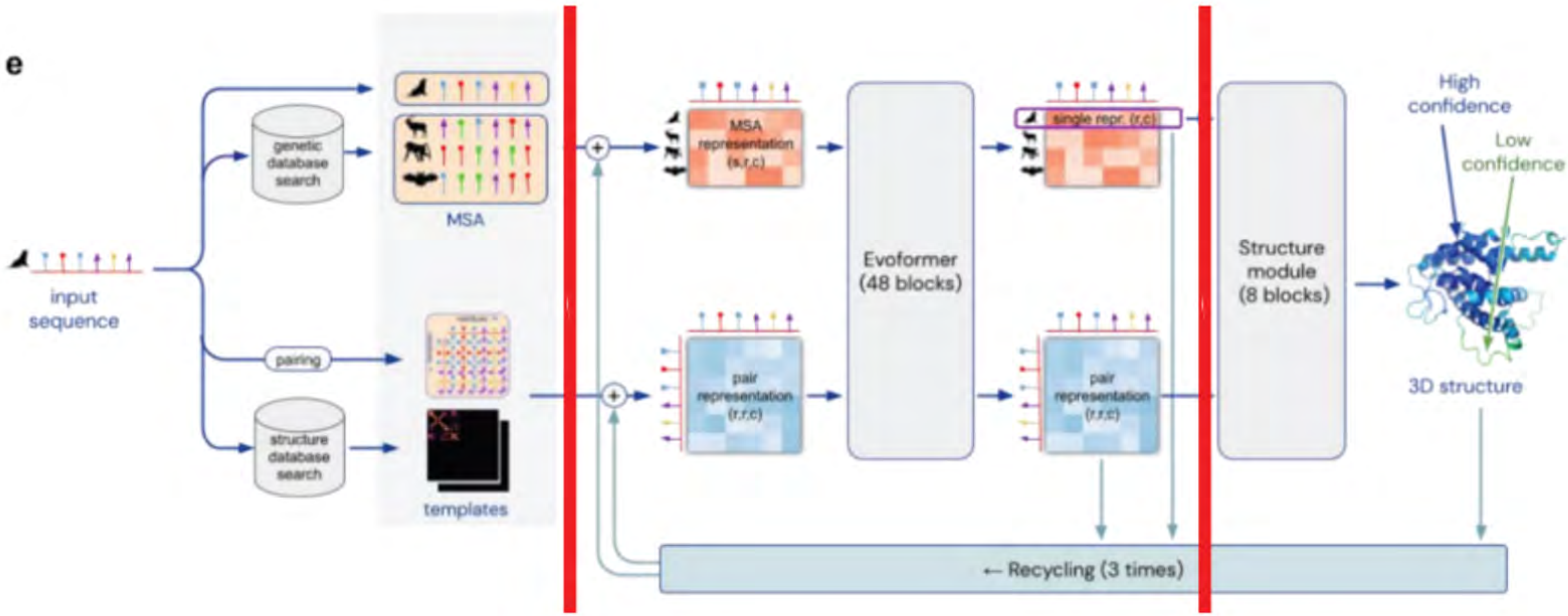


Very complicated data flow. Likely much trial and error to get here.



pair representation (r,r,c_z)
input row (r,c_z)
Linear c_z→(h,c)
sigmoid
Linear c_z→(h,c)
Linear c_z→(h,c)
left edges (r_v,h,c)
Linear c_z→(h,c)
central edges (r_q,h,c)
dot-product affinities (r_q,r_v,h)
Linear c_z→h
right edges (r_q,r_v,h)
+
softmax r_v
left edges (r_v,h,c)
attention weights (r_q,r_v,h)
output (r_q,h,c)
central edges (r_q,h,c)
o
Linear (h,c)→c_z
pair representation (r,r,c_z)
updates (r,c_z)

Summary

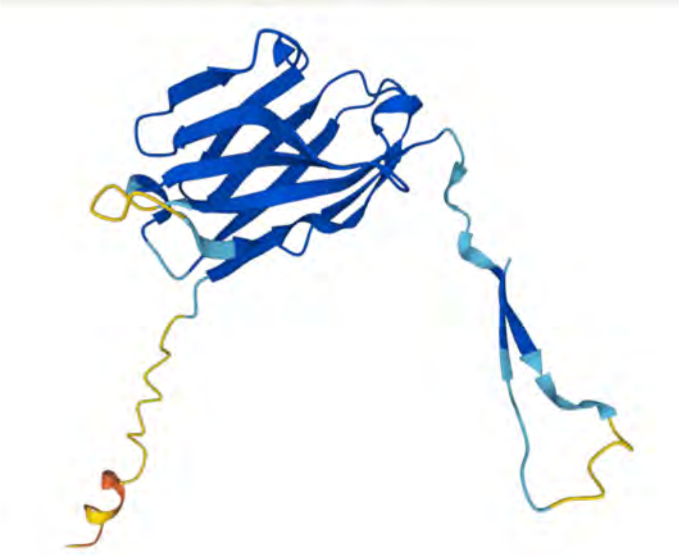

e
input sequence
genetic database search
pairing
structure database search
MSA
templates
MSA representation (s,r,c)
pair representation (r,r,c)
Evoformer (48 blocks)
single repr. (r,c)
pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (3 times)

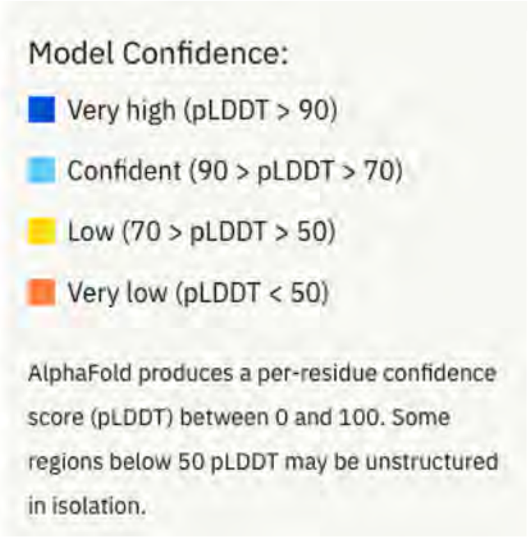
For grins, they trained a separate structure for each of the 48 blocks, to see the evolution of structure as a movie.

Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

Measures of reliability are included!

- Depends mainly on the number of MSA entries
- Reported in terms of plDDT
- Color coded in in presented structures





Model Confidence:

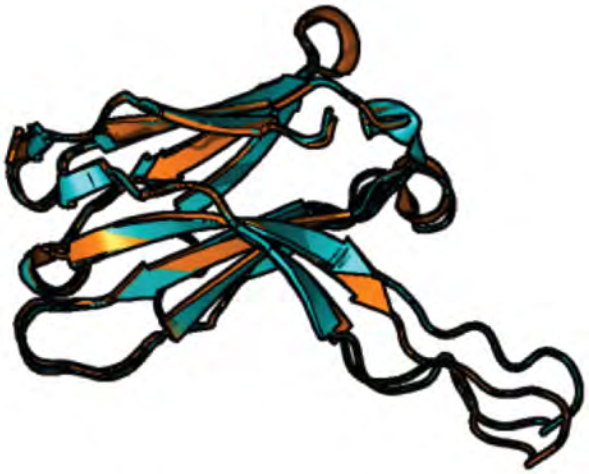
🟦 Very high (pLDDT > 90)

🟦 Confident (90 > pLDDT > 70)

🟨 Low (70 > pLDDT > 50)

🟧 Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.



Bibliography

Vamathevan, J., Clark, D., Czodrowski, P. *et al.* Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* **18**, 463–477 (2019). https://doi.org/10.1038/s41573-019-0024-5

Haberal, H. Oğul, Prediction of Protein Metal Binding Sites using Deep Neural Networks, *Mol. Inf.* **2019**, *38*, 1800169.

Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

BEST! https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/ CARLOS OUTEIRAL RUBIERA

AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function Jeffrey Skolnick, Mu Gao, Hongyi Zhou, and Suresh Singh
Journal of Chemical Information and Modeling **2021** *61* (10), 4827-4831
DOI: 10.1021/acs.jcim.1c01114

2022 conference on Neural Information Processing Systems
Workshop: Machine Learning in Structural Biology (very technical)