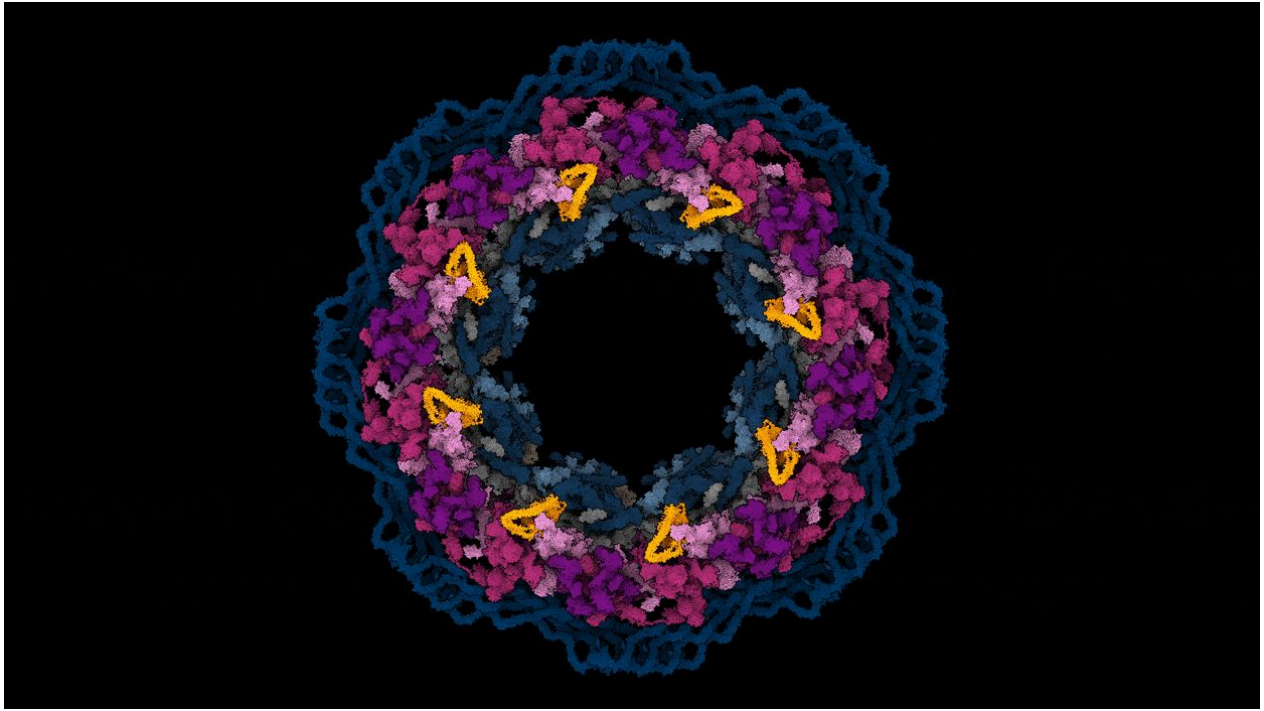
CSE7850/CX4803 Machine Learning in Computational Biology

Lecture 19: Protein Structure Prediction

Yunan Luo

Solving protein structure is EXTREMELY challenge



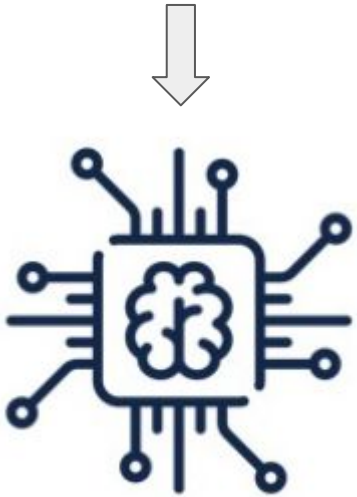
https://www.nature.com/articles/d41586-022-00997-5

Nuclear pore complex: controls the flow of molecules in and out of the nucleus of the cell, where the genome sits. Each is made up of more than 1,000 proteins that together form rings around a hole through the nuclear membrane.

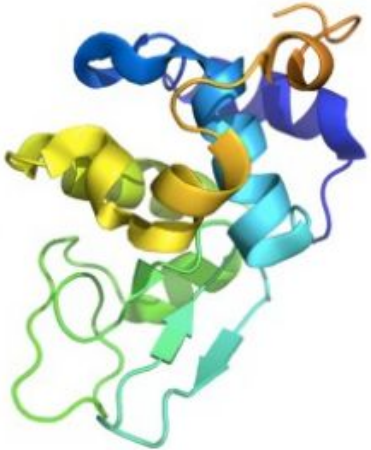
Protein structure prediction

Amino acid sequence

MEKVNFLKNGVLRLPPGFRFRPTDEELVVQYLKRKVFSFPLPASIIPEVEVYKSDPWDLPGDMEQEKYFFSTK
EVKYPNGNRSNRATNSGYWKATGIDKQIILRGRQQQQQLIGLKKTLVFYRGKSPHGCRTNWIMHEYRLAN
LESNYHPIQGNWVICRIFLKKRGNTKNKEENMTTHDEVRNREIDKNSPVVSVKMSSRDSEALASANSELKK



Algorithm / Model



Protein structure

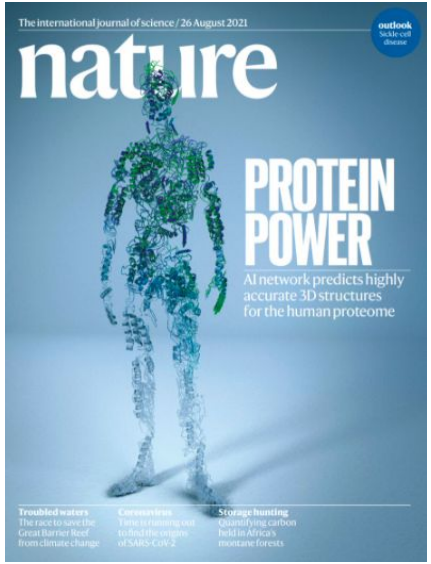
Slides credit: Jinbo Xu

Two papers today

- Paper #1: **AlphaFold**
  - A groundbreaking deep learning algorithm for protein structure prediction

- Paper #2: **OmegaFold**
  - Addressing a limitation of AlphaFold

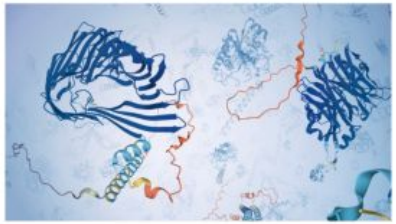
AlphaFold



FOCUS | 11 JANUARY 2022

Method of the Year 2021: Protein structure prediction

Protein structure prediction is our Method of the Year 2021, for the remarkable levels of accuracy achieved by deep learning-based methods in predicting the 3D structures of proteins and protein complexes, essentially solving this long-standing challenge.



One of biology's biggest mysteries 'largely solved' by AI

BBC

By Helen Briggs
BBC science correspondent

NEWS | BIOLOGY

'The game has changed.' AI triumphs at solving protein structures

Science

In milestone, software predictions finally match structures calculated from experimental data

NEWS | 30 November 2020

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

Nature

Google's deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.

A.I. Predicts the Shapes of Molecules to Come

NY Times

DeepMind has given 3-D structure to 350,000 proteins, including every one made by humans, promising a boon for medicine and drug design.

Weekendavisen

VM i molekylære puslespil

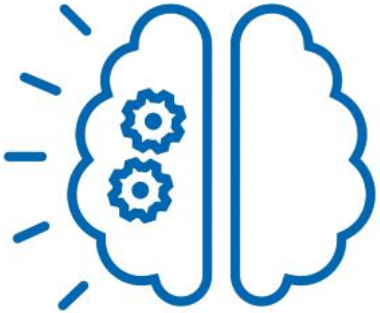
AMANDA BERING ABILDGAARD
JAKOB RAHR WINTHER
KRESTEN LINDORFF-LARSEN
RASMUS HARTMANN-PETERSEN

Slides credit: Kresten Lindorff-Larsen

What is AlphaFold?

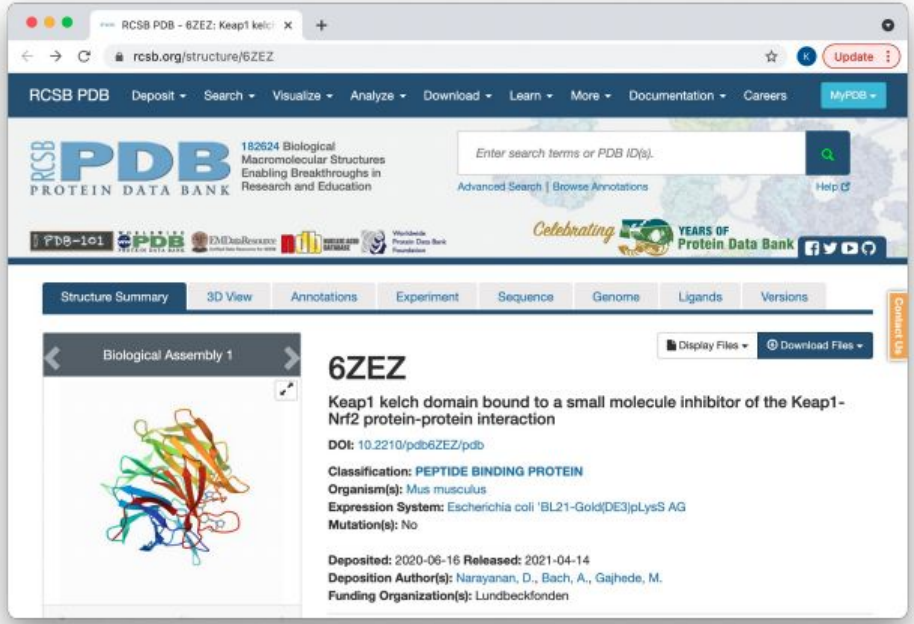
- A machine-learning-based model for predicting the 3D structure of proteins using only sequence as input

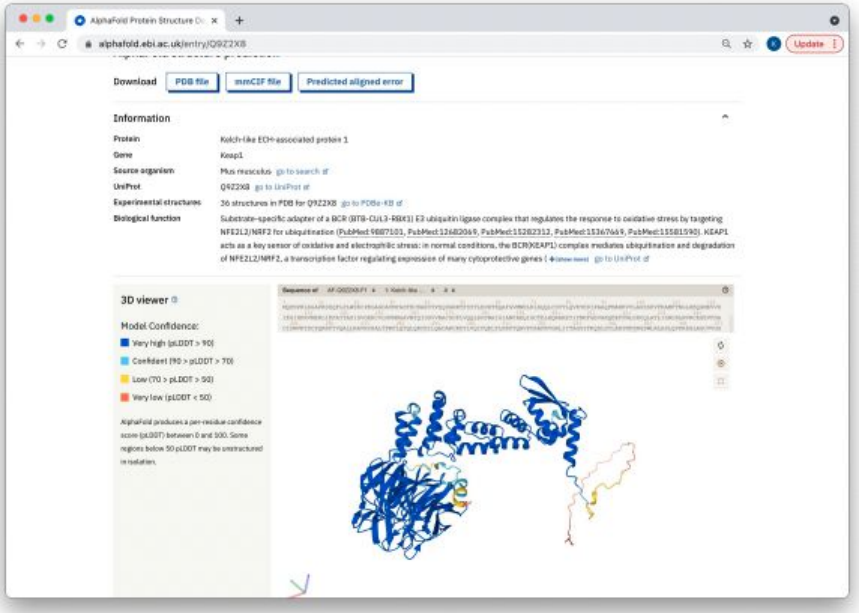
- Trained on known sequences and structures from the Protein Data Bank, as well as large databases of protein sequences



Slides credit: Kresten Lindorff-Larsen

What can AlphaFold do?





Pallesen et al, J Med Chem, 2021

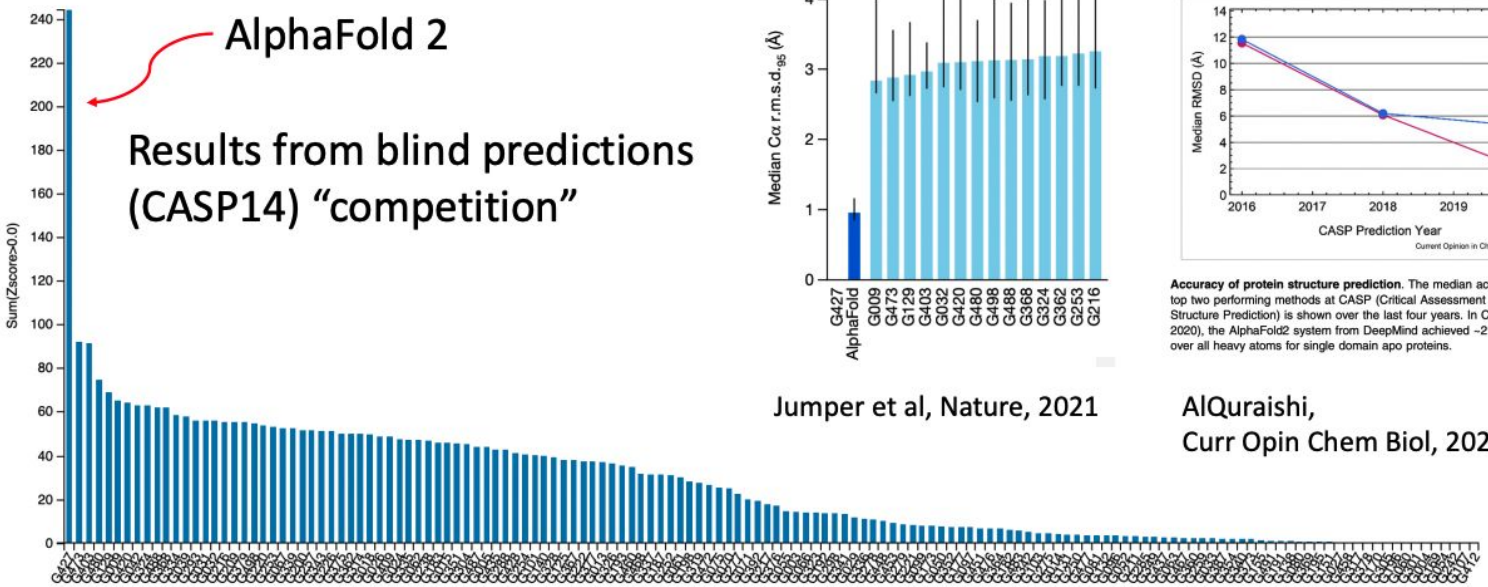
Tunyasuvunakool et al, Nature, 2021
Varadi et al, Nucl Acid Res, 2021

Slides credit: Kresten Lindorff-Larsen

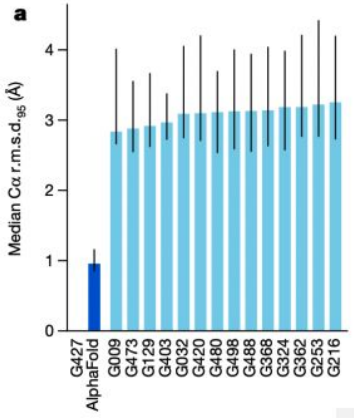
A startling success

Critical Assessment of protein Structure Prediction (CASP)

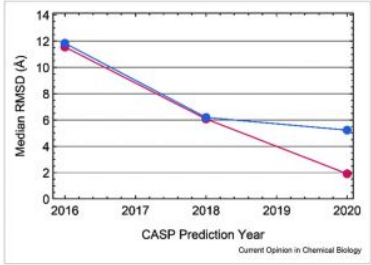
- Since 1994, every two years a contest is held to see who can best predict protein structures from sequences
- Targets structures are held from publication until results are in



Pereira et al, Proteins, 2021



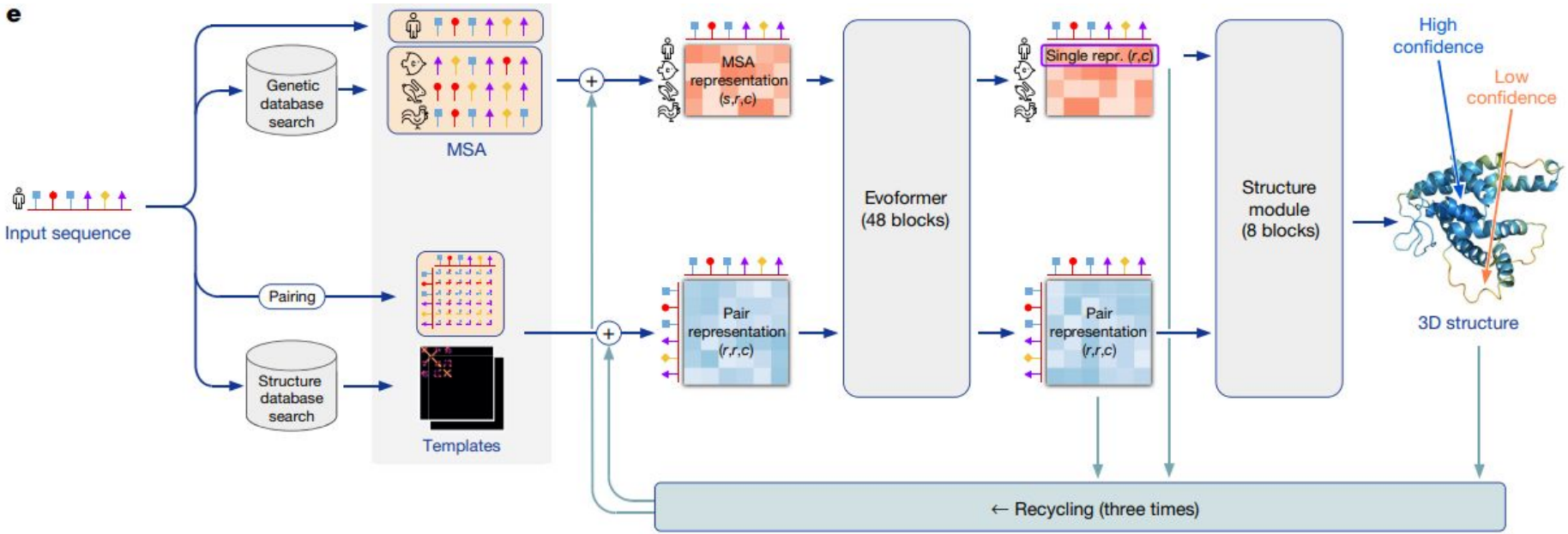
Jumper et al, Nature, 2021



Accuracy of protein structure prediction. The median accuracy of the top two performing methods at CASP (Critical Assessment of protein Structure Prediction) is shown over the last four years. In CASP14 (late 2020), the AlphaFold2 system from DeepMind achieved ~2 Å accuracy over all heavy atoms for single domain apo proteins.

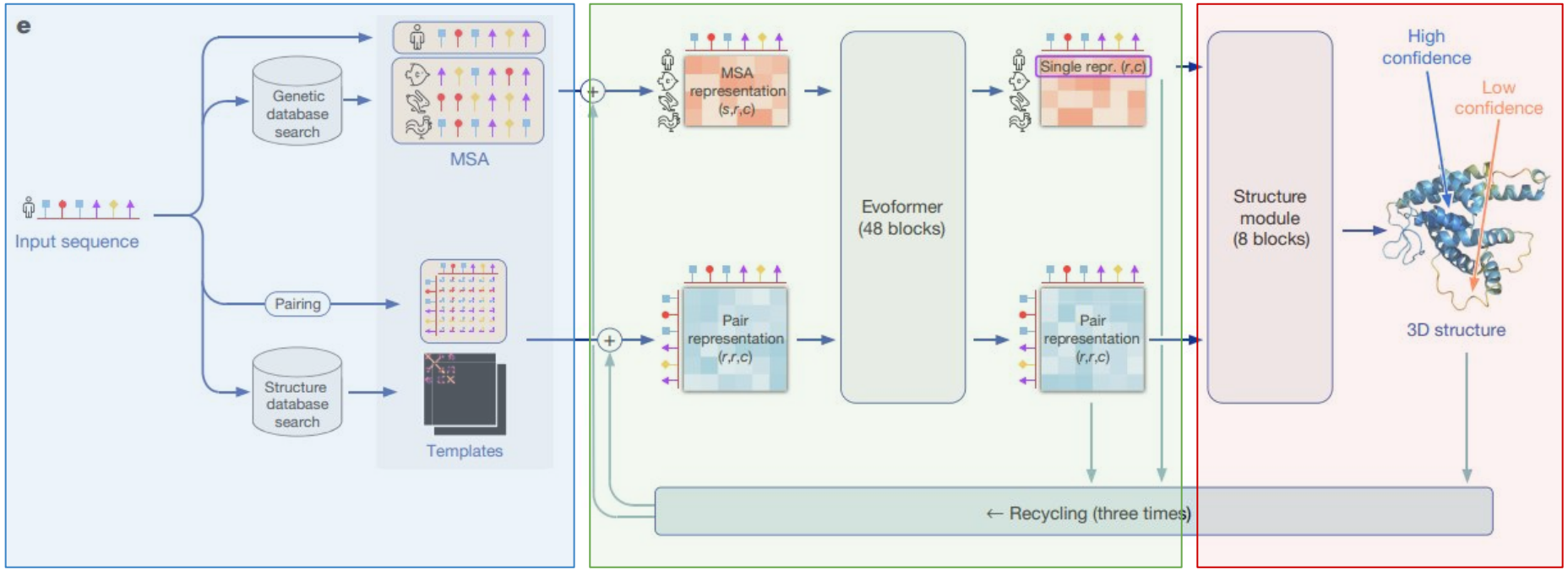
AlQuraishi,
Curr Opin Chem Biol, 2021

How does AlphaFold work?



Jumper et al., *Nature*, 2021

How does AlphaFold work?



Input module

Embedding module (Evoformer)

Structure module

Jumper et al., *Nature*, 2021

https://github.com/sokrypton/ColabFold

ColabFold: AlphaFold in your browser

AlphaFold2.ipynb

File Edit View Insert Runtime Tools Help

Share

+ Code + Text Copy to Drive

Connect

ColabFold v1.5.2: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using AlphaFold2 and Alphafold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and read our manuscript. Old versions: v1.4, v1.5.1

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: Making protein folding accessible to all. *Nature Methods*, 2022

Input protein sequence(s), then hit `Runtime` -> `Run all`

query_sequence: " PIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASK

- Use `:` to specify inter-protein chainbreaks for **modeling complexes** (supports homo- and hetro-oligomers). For example **PI...SK:PI...SK** for a homodimer

jobname: " test

num_relax: 0

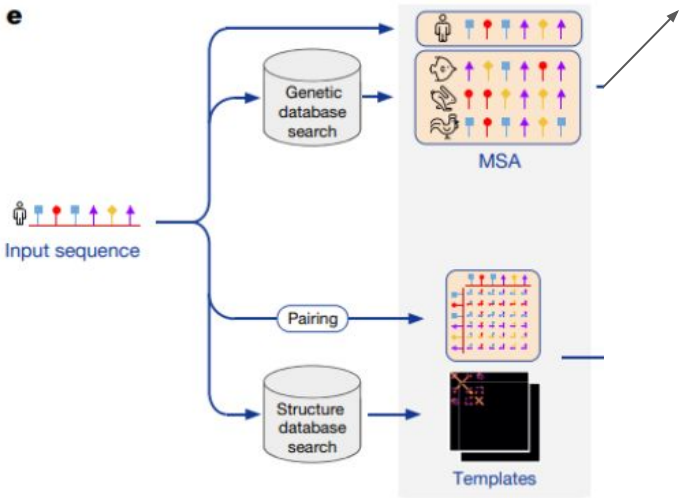
- specify how many of the top ranked structures to relax using amber

template_mode: none

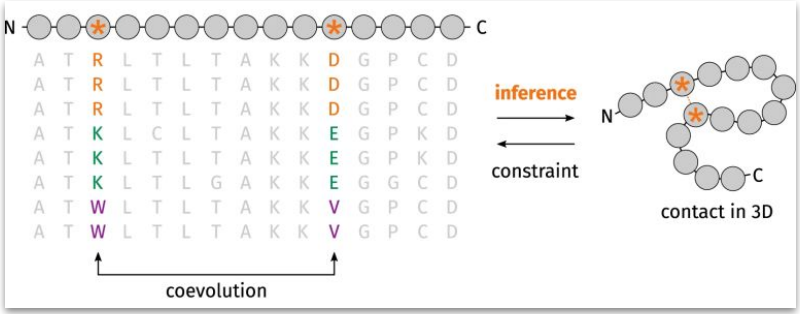
- `none` = no template information is used. `pdb70` = detect templates in pdb70. `custom` - upload and search own templates (PDB or mmCIF format, see notes below)

Show code

Key ingredient: Multiple sequence alignment



- Multiple sequence alignment (MSA)
  - Residues in contact tend to coevolve



(Image source)

Search homologous sequences to build MSA

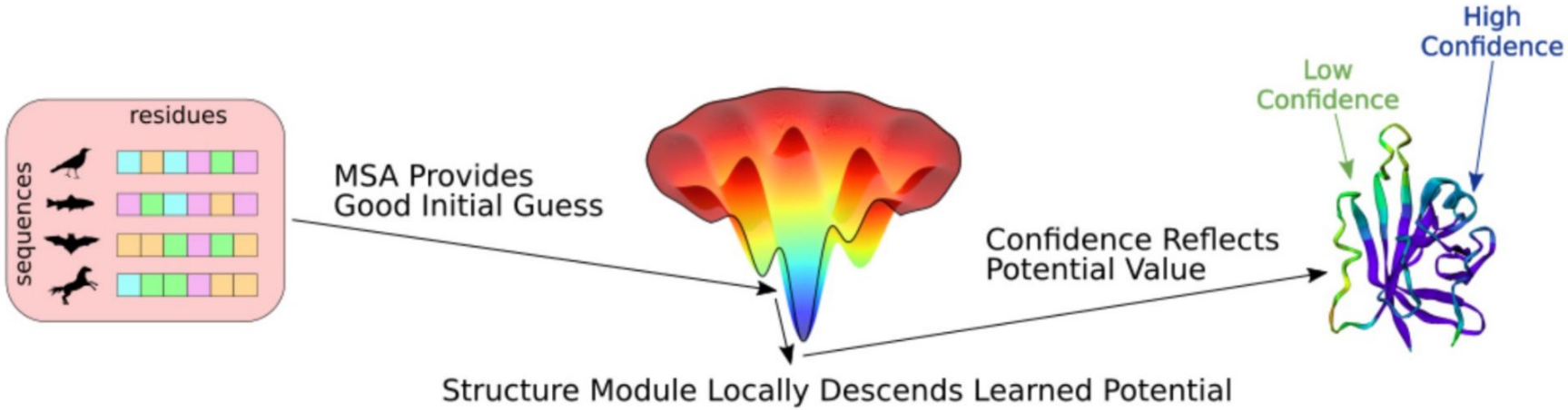
Protein sequence ADRLYLTKIHHEFEGD

Search homologous sequences

Multiple sequence alignment

```
AQKLYLTHIDAEVDGD
ADTLYMTKIHHQFQGD
ADRLFITEVKQVFEGD
ADRLYMTKIHHTFEGD
ADKLYCTLIHNSFDGD
ADRLYMTKIHHEFEGD
ADTLYLTMIHQKFQAD
TDTLYITHIDETFQGD
ADTLYLTQIRNKFQGD
TSRMYITKIGQEFEGD
ADRLYMTKIHHEFEGD
ADRLYITHIHHSFEGD
ADRLYMTKIHHEFEGD
```

A guided search in a good energy function?



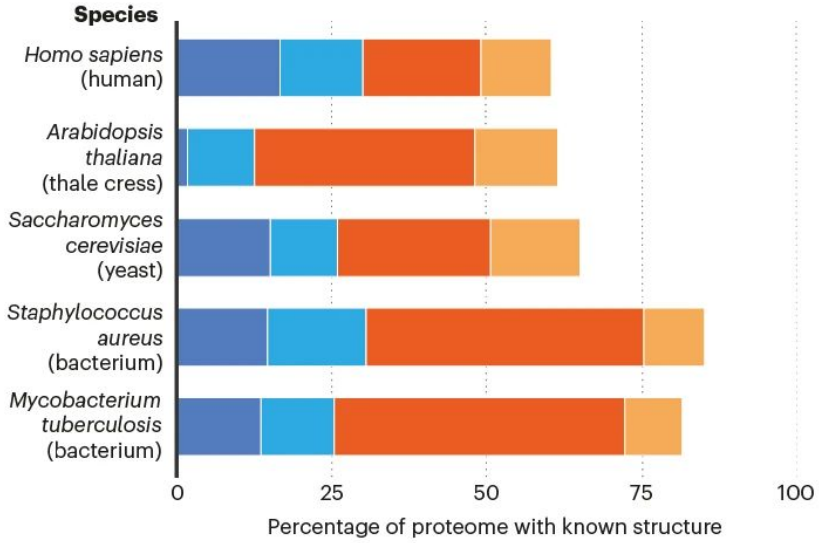
Roney and Ovchinnikov, bioRxiv, 2022

WHAT'S KNOWN ABOUT PROTEOMES

AlphaFold's predictions have greatly increased the proportion of confidently known structures in the human proteome — the collection of all human proteins. The software is even more useful for other species.

Source of knowledge about proteome

- High-quality experimental structures in the PDB*
- Structural knowledge derived from related proteins in the PDB*
- Knowledge from AlphaFold models only (high confidence)
- Knowledge from AlphaFold models only (intermediate confidence)



*PDB: Protein Data Bank. AlphaFold can also be used to calculate these structures — but doesn't add significantly to what's already known.

©nature

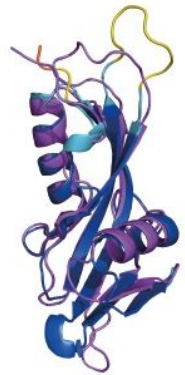
THE GOOD, THE BAD AND THE UGLY

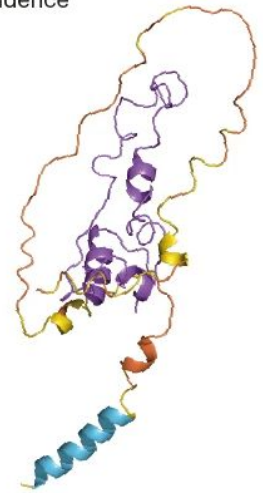
AlphaFold's predictions of a folded protein's structure come with confidence estimates. Superimposing each model on the experimentally determined structure (if available) shows the accuracy of the prediction.

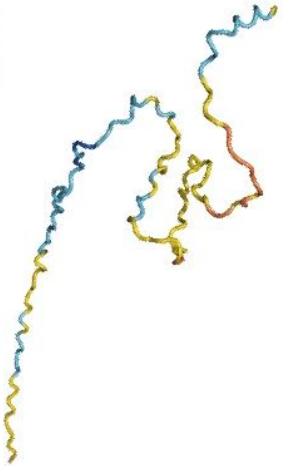
Protein Data Bank (PDB) structure

AlphaFold structure, with confidence estimates for each section.

Very high | High | Low | Very low







Good
AlphaFold model of phosphohistidine phosphatase overlaps closely with PDB structure.

Bad
AlphaFold model of human insulin bears no relation to the PDB structure.

Ugly
AlphaFold has little confidence across much of its prediction for this human ubiquitin-protein ligase. There is no PDB structure to compare it with.

©nature

https://www.nature.com/articles/d41586-022-00997-5

Limitations of AlphaFold?

- Multiple sequence alignments (MSAs) of homologous sequences are not always available
- Examples:
  - Orphan proteins
  - Fast-evolving proteins like antibodies
- But in the nature, a protein folds from its primary amino acid sequence into 3D structure, without consulting MSAs!

- This is the main motivation of our Paper #2 today: **OmegaFold**

OmegaFold: replacing MSAs with LMs