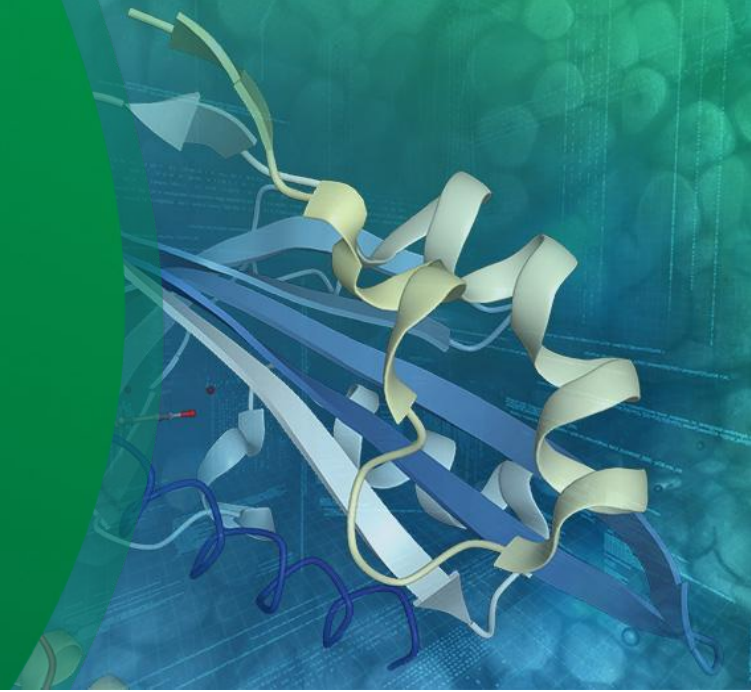
AlphaFold



PDBe
Protein Data Bank in Europe

Google DeepMind

EMBL-EBI

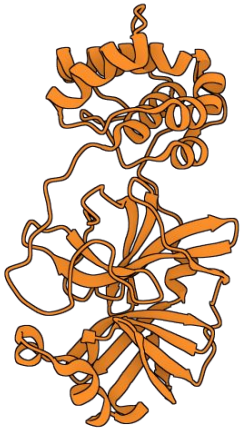
What are proteins?

- Each protein molecule consists of one or more chains, made by linking small molecules called amino acids. The amino acid residues are arranged in a specific sequence along each chain. These chains are folded into a distinctive 3D shape, determined by interactions between the amino acids.

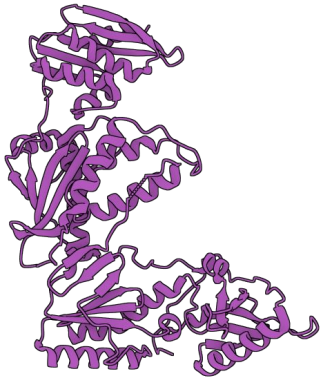
- There is a strong relationship between protein structure and protein function. The 3D shape of the protein determines the actions it can perform.



Salmonella alpha-2-macroglobulin (PDB ID: 4U48)
Provides protection against proteases secreted from hosts



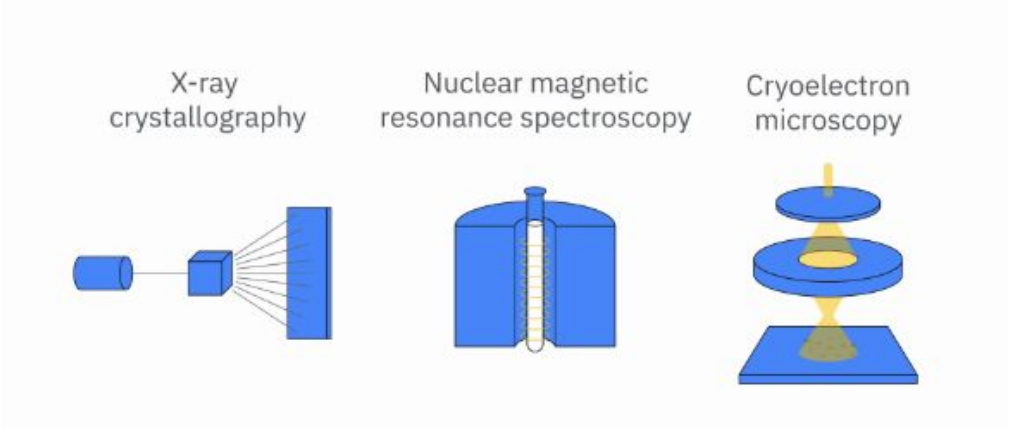
SARS-CoV main protease (PDB ID: 6Y7M)
This essential coronavirus protease processes the polyproteins translated from the viral RNA



HIV-1 Reverse Transcriptase (PDB ID: 8U6H)
One of the key proteins responsible for the replication cycle of HIV-1 in the host.

The struggle to determine protein structures

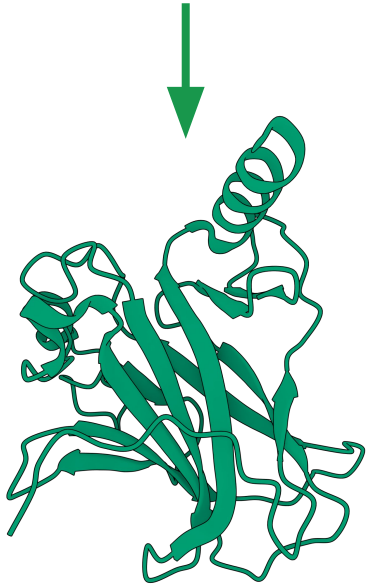
- For several decades, it has been possible to determine protein structures at atomic levels of detail, using experimental methods.
- However, despite the progress made in experimental techniques, determining protein structure experimentally is inherently time and labour-consuming, and involves many troubleshooting steps.



The protein folding problem

…TAKSVTCTYSPALNKLFCQLAKTCPVQLWV
DSTPPPGTRVRAMAIYKQSQHMTEVVRRCPH
HERCSDSDGLAPPQHLIRVEGNLRAEYLDDRN
TFRHSVVVPCEPPE…..

- **Anfinsen's dogma**[1], 3D structure of (most) natural proteins is determined only by its amino acid sequence.
- This hypothesis was important, because it suggested we should be able to reliably predict a protein's structure from its sequence. Decades of research into structural biology have since shown that Anfinsen was largely correct.

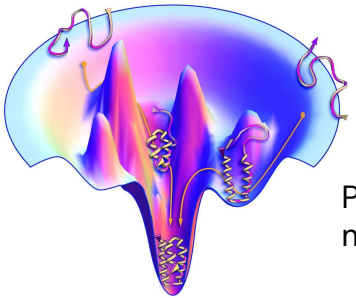


[1] Anfinsen, C B, Principles that Govern the Folding of Protein Chains, Science, 1973 Jul 20;181(4096):223-30.

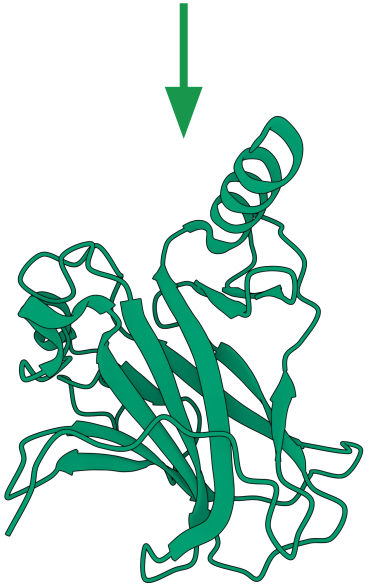
The protein folding problem

…TAKSVTCTYSPALNKLFCQLAKTCPVQLWV
DSTPPPGTRVRAMAIYKQSQHMTEVVRRCPH
HERCSDSDGLAPPQHLIRVEGNLRAEYLDDRN
TFRHSVVVPCEPPE…..

- **Levinthal's paradox**[1], states that the number of potential configurations a protein could adopt is astronomical, yet proteins fold rapidly, suggesting efficient folding pathways.
- It can be seen as a **computational problem** because it involves **predicting a highly complex**, 3D shape from a linear, one-dimensional string of amino acids.



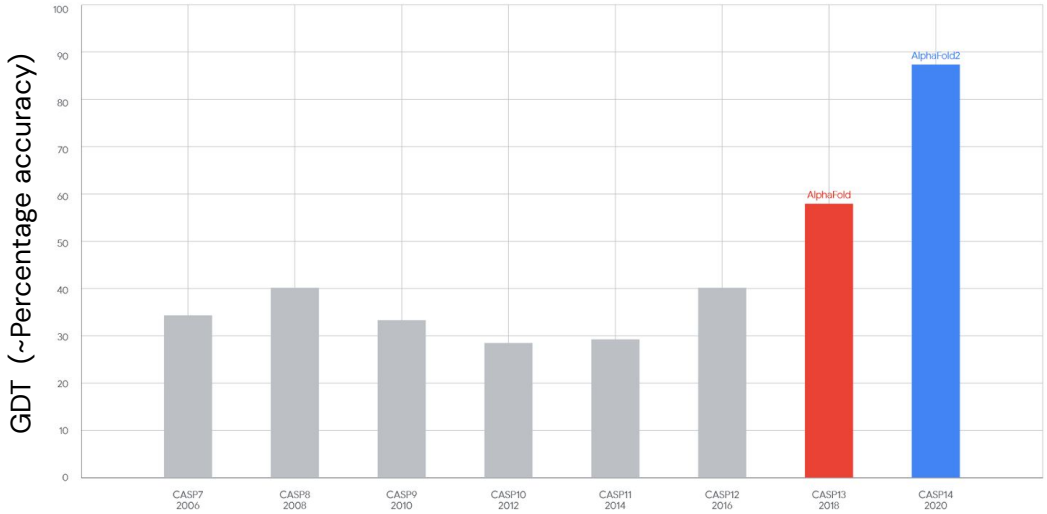
Proteins seek to “fold” into a structure that minimizes free energy.



[2] Levinthal C. Are there pathways for protein folding?. Journal de chimie physique (1968)

AlphaFold's success in CASP

Critical Assessment of Structure Prediction (CASP) is an experimental test of protein structure predictions. It has been carried out every two years since 1994. The assessment is open to anyone.



Overall success at protein structure prediction in CASPs over the years. AlphaFold drove rapid improvements in 2018 and 2020.

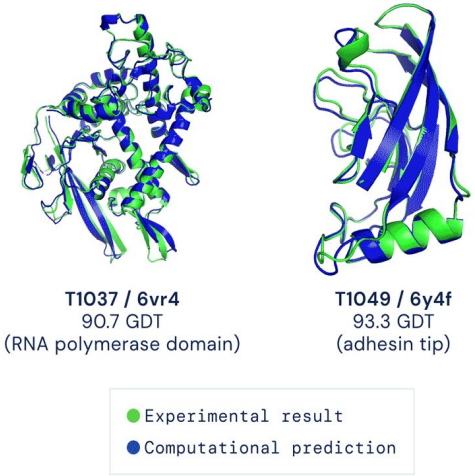
- Google DeepMind previously entered an earlier version of AlphaFold in 2018's CASP13. It took the first place but by a small margin.
- Those predictions were not accurate enough, so the protein structure prediction problem was not considered solved.

Google DeepMind

PDBe Protein Data Bank in Europe EMBL-EBI

What is AlphaFold 2?

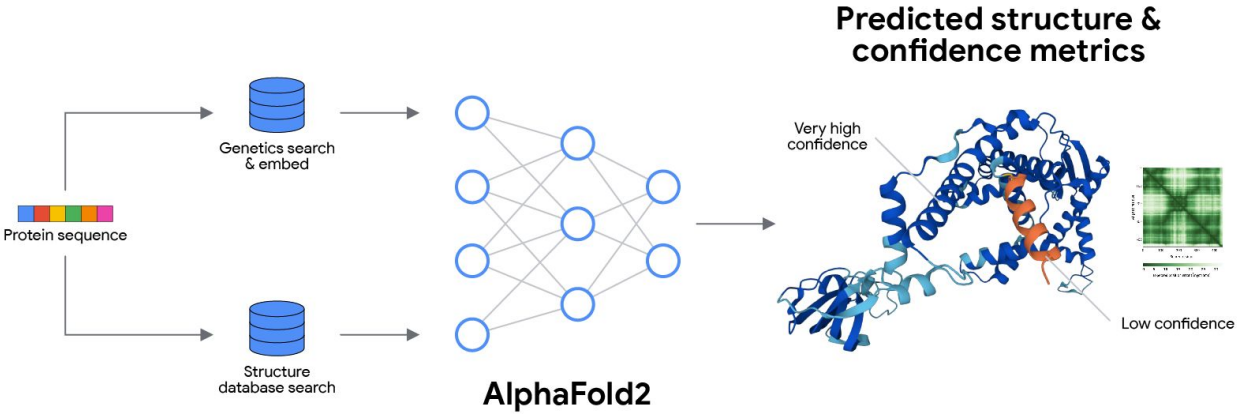
- AlphaFold 2[1] was developed by Google DeepMind.
  - It can create highly **accurate theoretical models** of protein structures **using amino acid sequence**.
- In 2020, AlphaFold 2 achieved an unprecedented accuracy score at CASP14 [2]



[1] Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021)
[2] Jumper, John et al. "Applying and improving AlphaFold at CASP14." Proteins (2021)

AlphaFold2 is an advanced deep learning algorithm

- The training data for AlphaFold2 came from the Protein Data Bank (PDB).
- Crucially, AlphaFold also provides confidence metrics, such as pLDDT, pTM and PAE. If it is not sure about some part of the structure, it will tell you this in the form of a confidence score, enabling critical interpretation



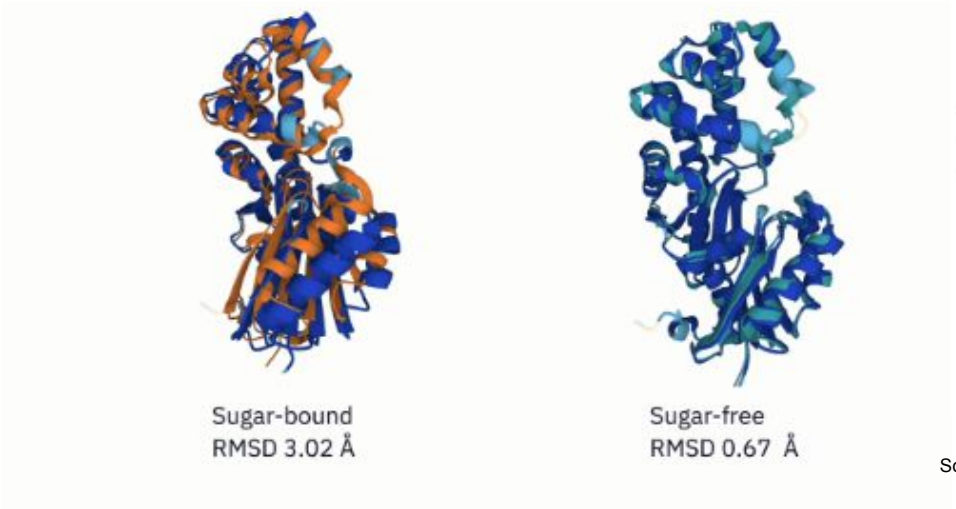
Google DeepMind

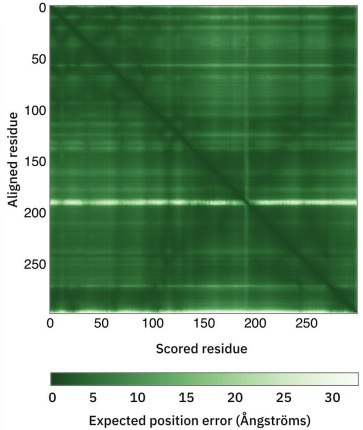
Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021)

PDBe Protein Data Bank in Europe EMBL-EBI

Strengths and limitations of AlphaFold 2- conformations

By default, AlphaFold2 does not capture such conformational changes, as it was designed to predict static structures, i.e. structural snapshots.





Source: https://pdbe-kb.org/proteins/Q96Y14
(Hexokinase)

Hexokinase (Q96Y14) adopts distinct conformations in the presence (orange, left) and absence (green, right) of sugar. Notably, AlphaFold's structure prediction aligns with the sugar-free state (as could be seen both visually and via RMSD value).

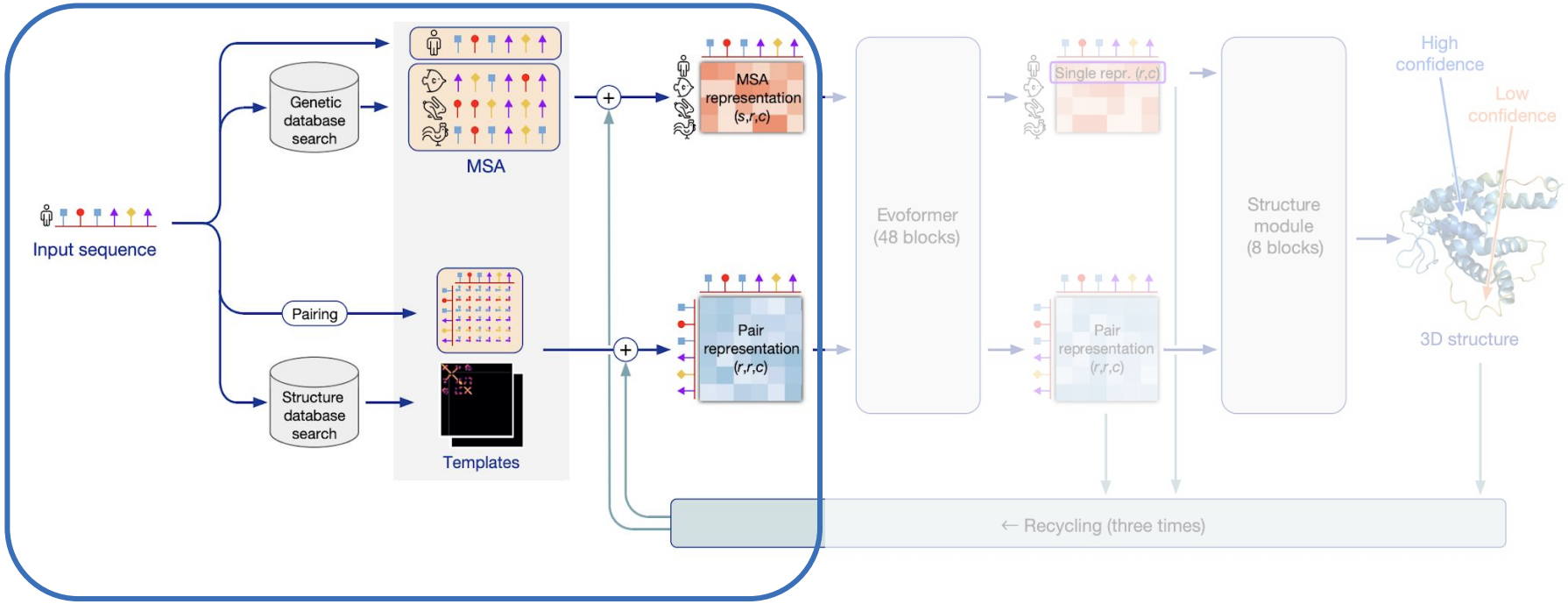
Strengths and limitations of AlphaFold 2

AlphaFold2 predicts	AlphaFold2 struggles to predict	AlphaFold2 doesn't predict
• Single protein chains • Protein multimers • Multisubunit protein-protein complexes	• Multiple conformations for the same sequence • Effects of point mutations • Antigen-antibody interactions	• Protein-DNA and protein-RNA complexes • Nucleic acid structure • Ligand and ion binding • Post-translational modifications • Membrane plane for transmembrane domains

AlphaFold: High-level overview



Inputs



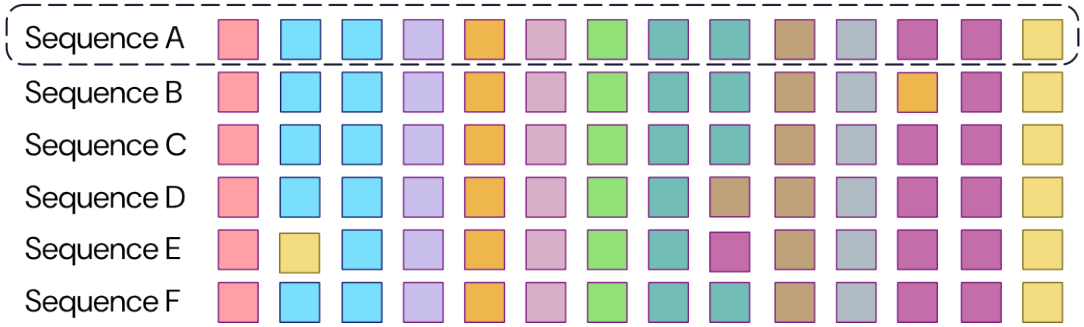
Google DeepMind

Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021).

PDBe Protein Data Bank in Europe EMBL-EBI

The role of multiple sequence alignment (MSA)

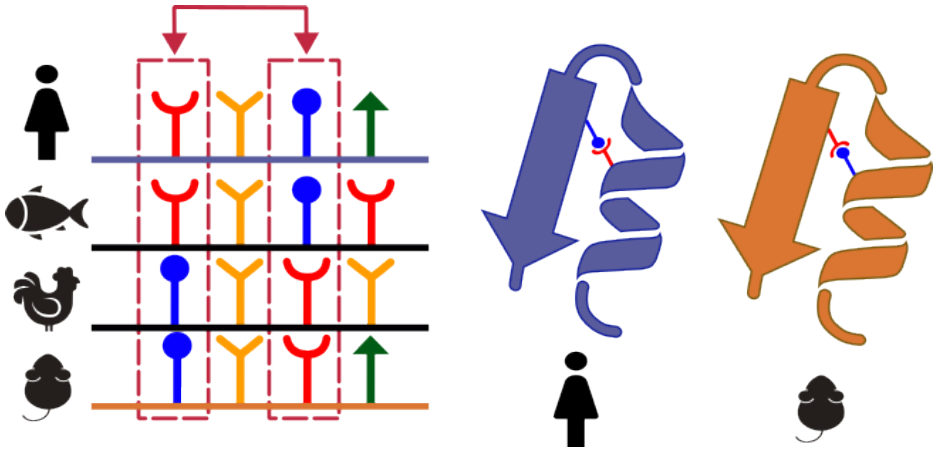
- The primary input for AlphaFold2's neural network is then the MSA.
- AlphaFold2 uses MSAs to compare and analyse the sequences of similar proteins from different organisms. It highlights similarities and differences, which helps understand the evolutionary relationships between the proteins.



Co-evolution

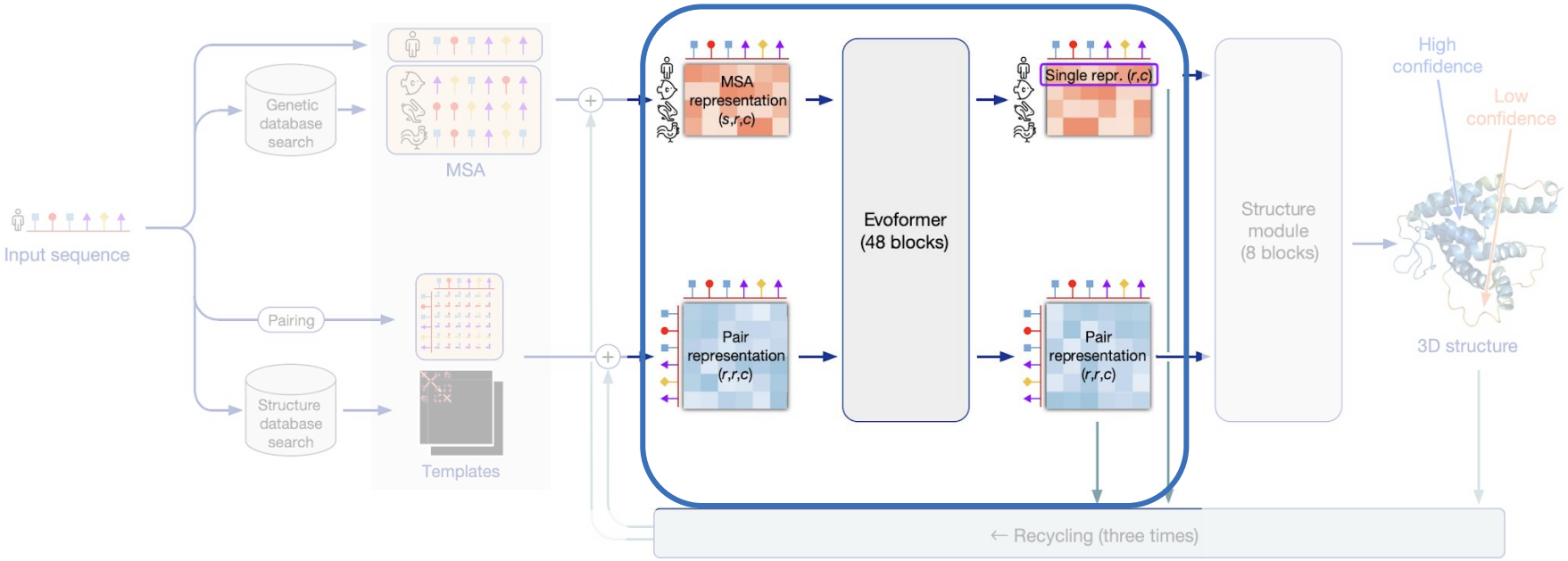
If two amino acids in a protein are in close contact, mutations in one of them will probably be followed by mutations of the other. This preserves the structure of the protein, and is known as co-evolution or covariation.

The opposite is also true: if two regions of a protein are changing and evolving independently from each other, it is likely that they are not in direct contact.



Göbel et al.,. Correlated mutations and residue contacts in proteins. Proteins 1994
William R. Taylor, Kerr Hatrick, Compensating changes in protein multiple sequence alignments, Protein Engineering, (1994)

Evoformer



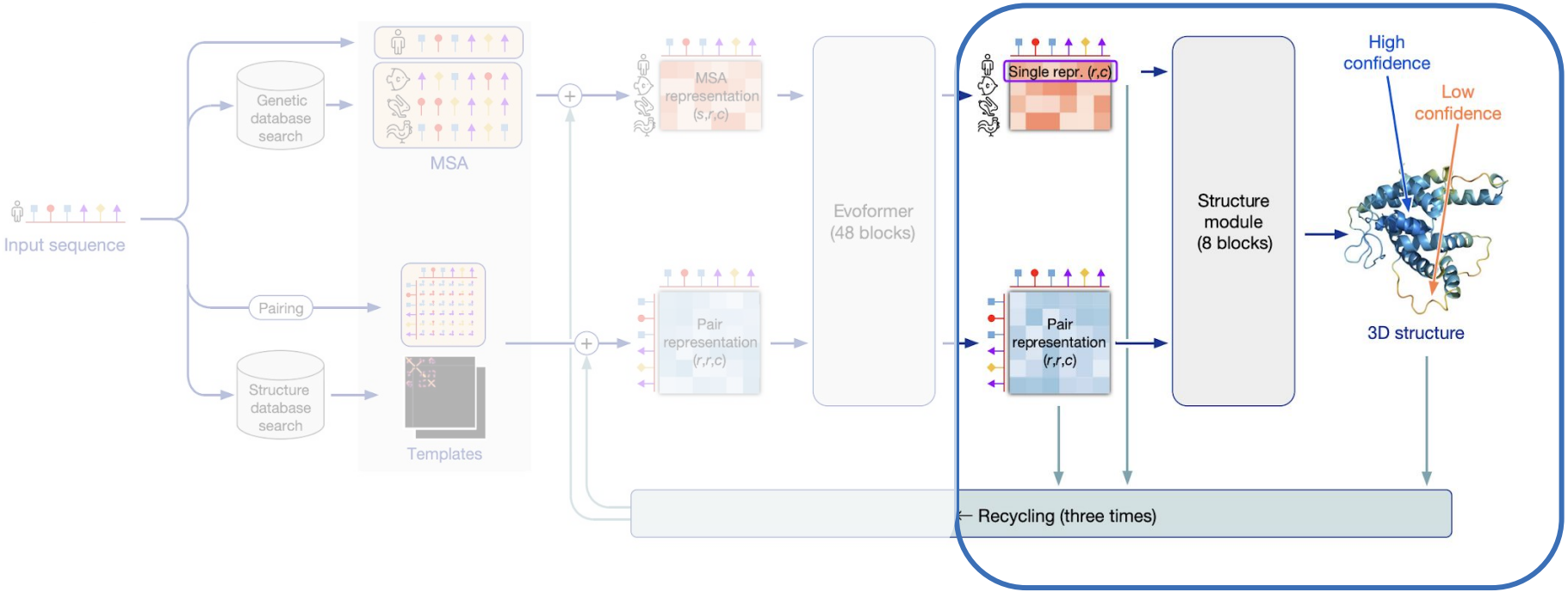
The Evoformer interprets and updates both the MSA and the pair representations. The important aspect of this network is the continuous flow of information between the MSA and the pair representations. This enables reasoning about spatial and evolutionary relationships, which refines the structural hypothesis.

Google DeepMind

Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021).

PDBe Protein Data Bank in Europe EMBL-EBI

Structure module

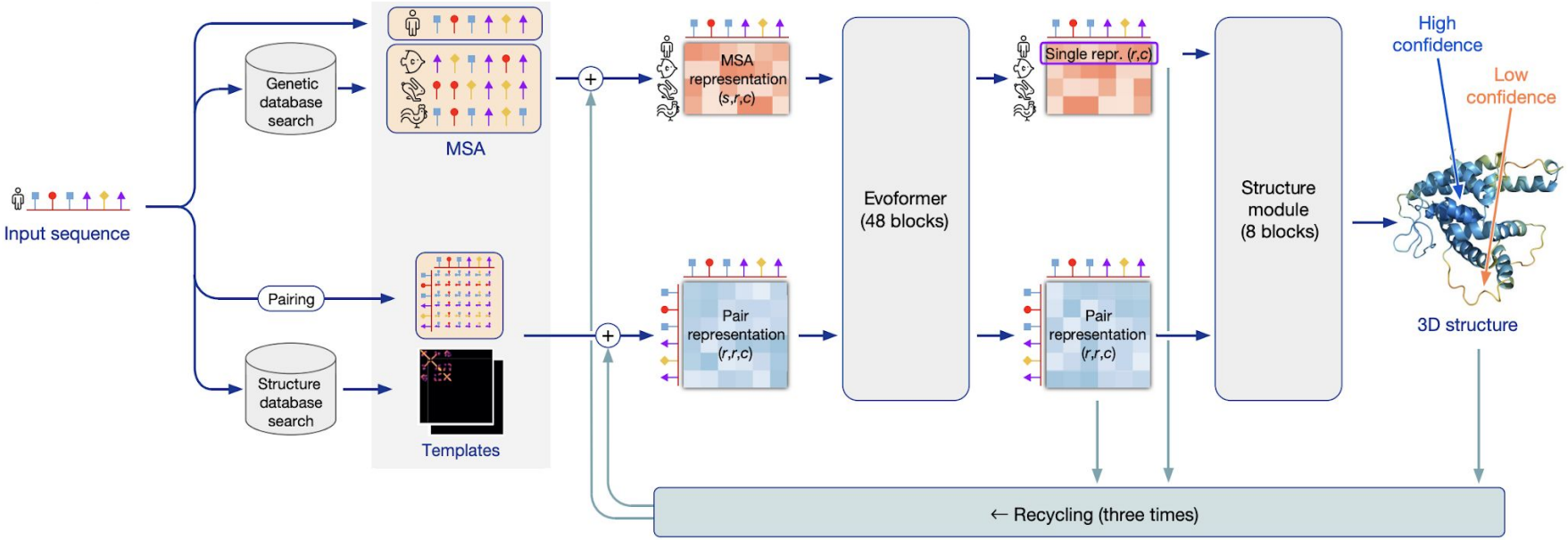


Google DeepMind

Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021).

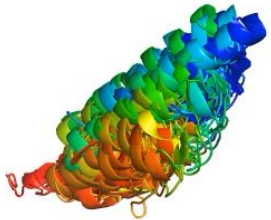
PDBe Protein Data Bank in Europe EMBL-EBI

Recycling



Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021).

AlphaFold, a high-level overview. Video



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

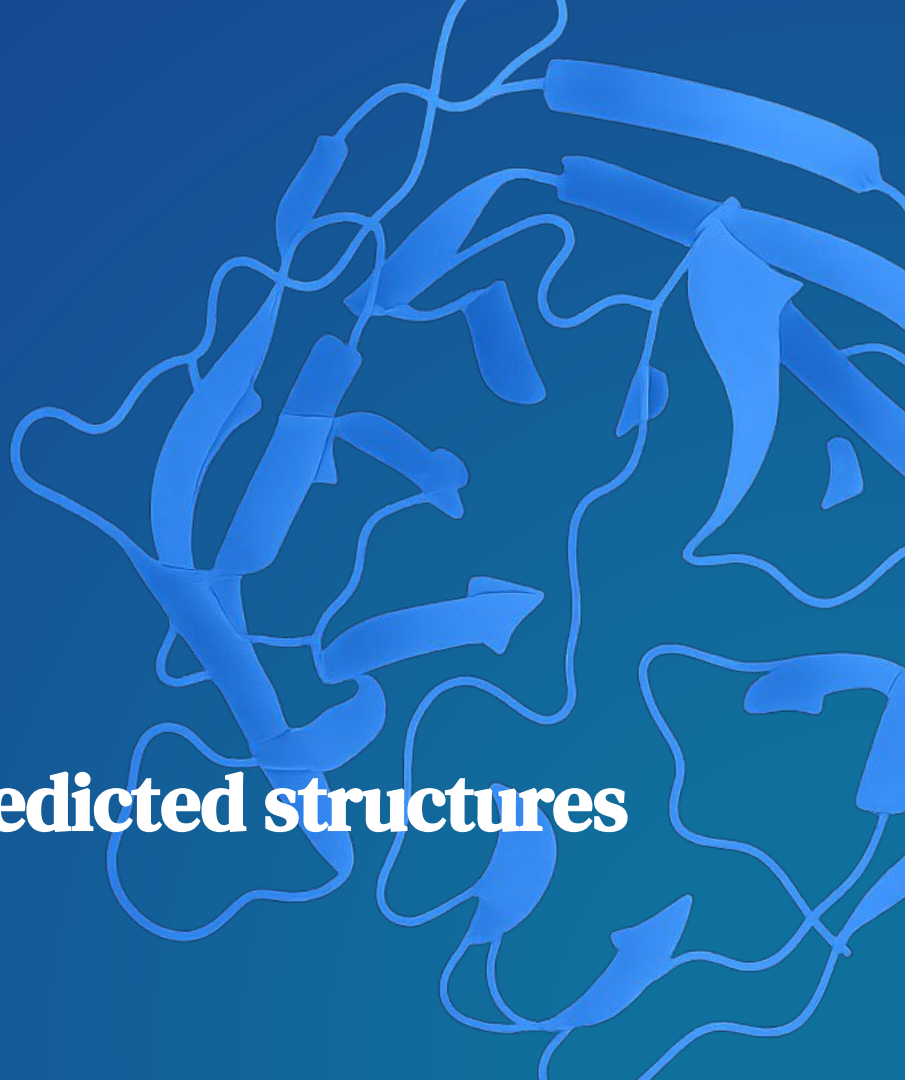
CASP14 target T1044 (RNA polymerase of crAss-like phage)

Google DeepMind

Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021).

PDBe Protein Data Bank in Europe EMBL-EBI

Evaluating AlphaFold2's predicted structures using confidence scores



“Nothing in **biology** makes sense except in the light of **evolution** ”

Theodosius Dobzhansky, 1973

“Nothing in **AlphaFold predictions** makes sense except in the light of **confidence metrics** ”

Google DeepMind

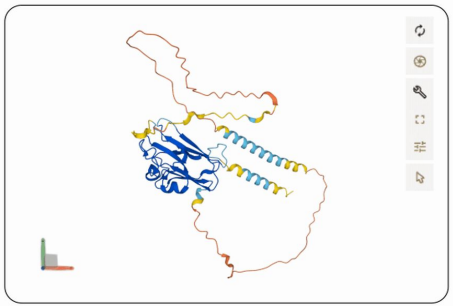
PDBe Protein Data Bank in Europe EMBL-EBI

pLDDT confidence metric

- pLDDT measures confidence in the local structure, estimating how well the prediction would agree with an experimental structure. It is based on the local distance difference test Cα (lDDT-Cα), which is a score that does not rely on superposition but assesses the correctness of the local distances.

- The pLDDT score can vary significantly along a protein chain. This means AlphaFold2 can be very confident in the structure of some parts of the protein, but less confident in other regions.

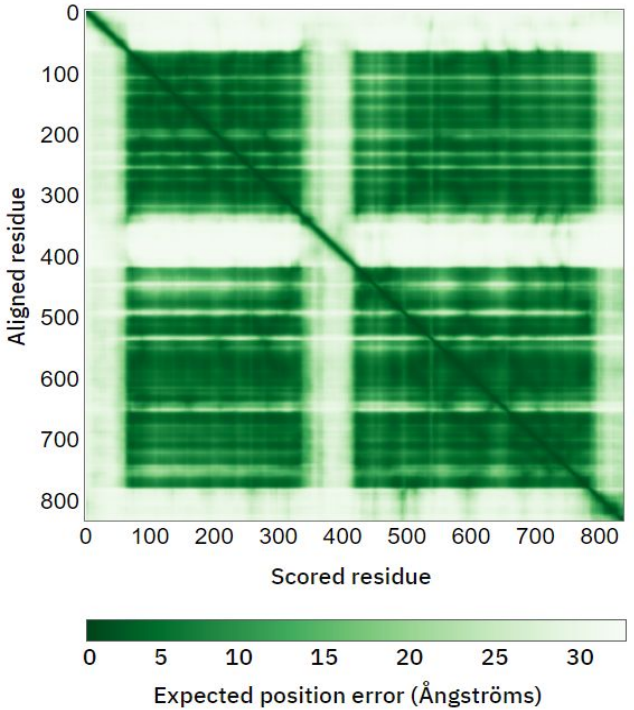
- A high pLDDT score for all domains of a protein does not necessarily mean that AlphaFold2 is confident in the relative positions or orientations of those domains. pLDDT does not measure confidence at such large scales, so a different metric is required.





Predicted Aligned Error (PAE)

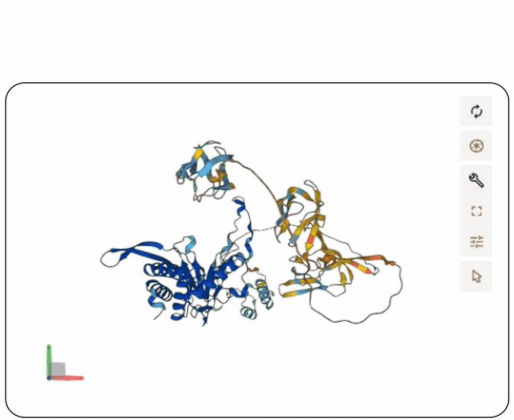
- PAE is a measure of how confident AlphaFold2 is in the relative position of two residues within the predicted structure.
- PAE is defined as the expected positional error at residue X, measured in Ångströms (Å), if the predicted and actual structures were aligned on residue Y.
- The PAE plot will always have a dark green diagonal line running from top-left to bottom-right. This represents residues being aligned against themselves, where confidence is always high by definition, so it is not informative and can be ignored. The biologically-relevant information, in terms of relative orientations, is contained in the regions away from the diagonal.

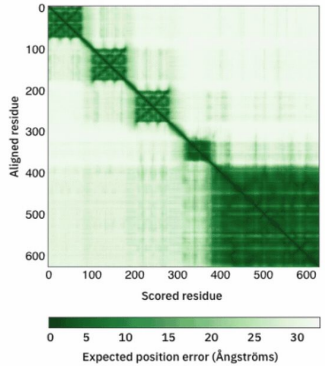


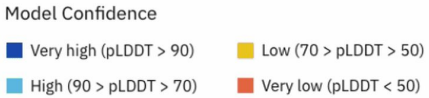
Google DeepMind

PDBe Protein Data Bank in Europe EMBL-EBI

Predicted Aligned Error (PAE)

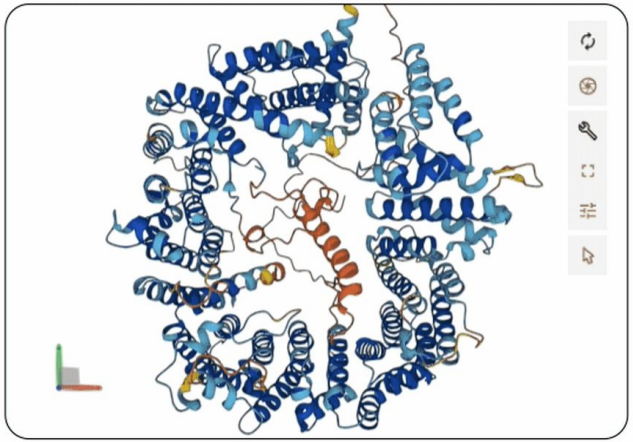


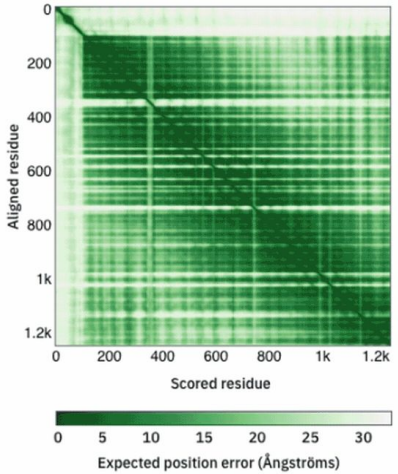




- PAE is effectively a measure of how confident AlphaFold2 is that the domains are well packed and that the relative placement of the domains in the predicted structure is correct.
- A low PAE score between two residues from two different domains means low predicted error. This in turn means AlphaFold2 is confident in the position of these residues. Conversely, a high PAE score means that AlphaFold2 is not confident in their relative position.

Important to use both outputs together





Model Confidence

Very high (pLDDT > 90)

Low (70 > pLDDT > 50)

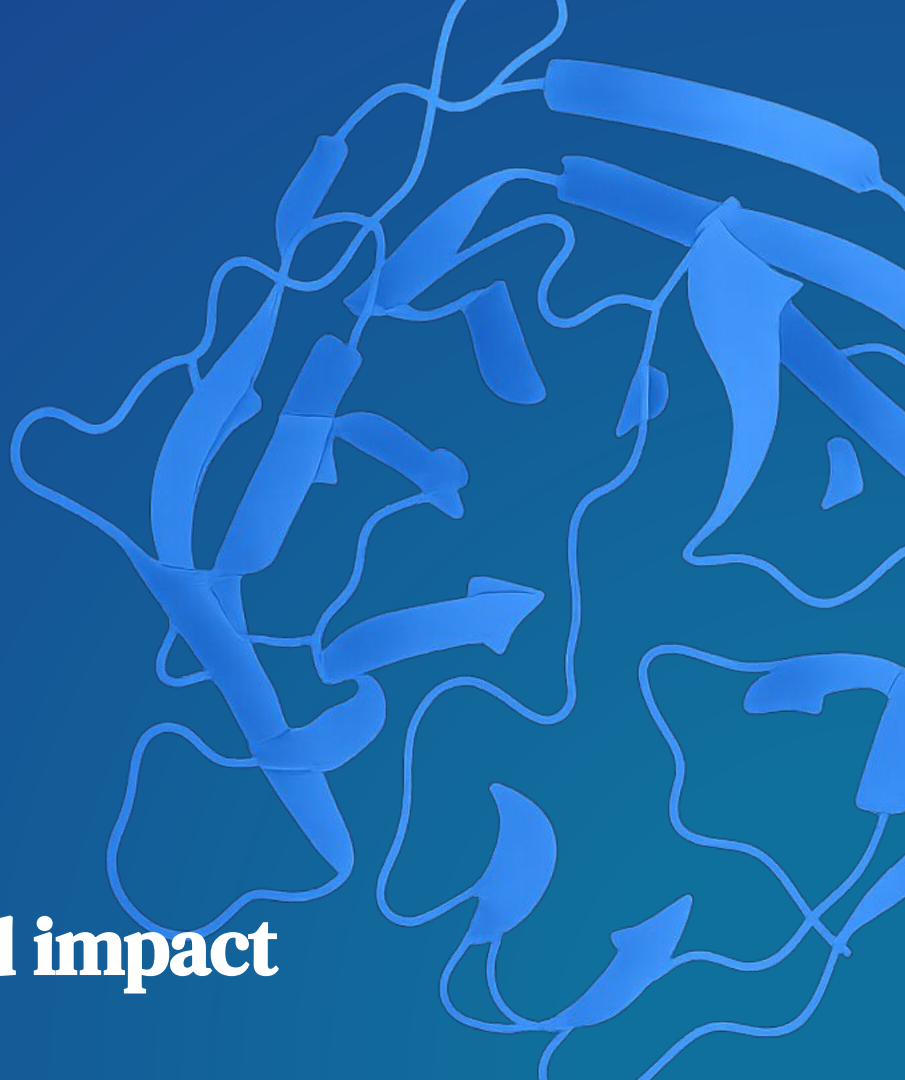
High (90 > pLDDT > 70)

Very low (pLDDT < 50)

Exportin-4 protein
https://alphafold.ebi.ac.uk/entry/A0A1I9LPW2

Google DeepMind

PDBe Protein Data Bank in Europe EMBL-EBI

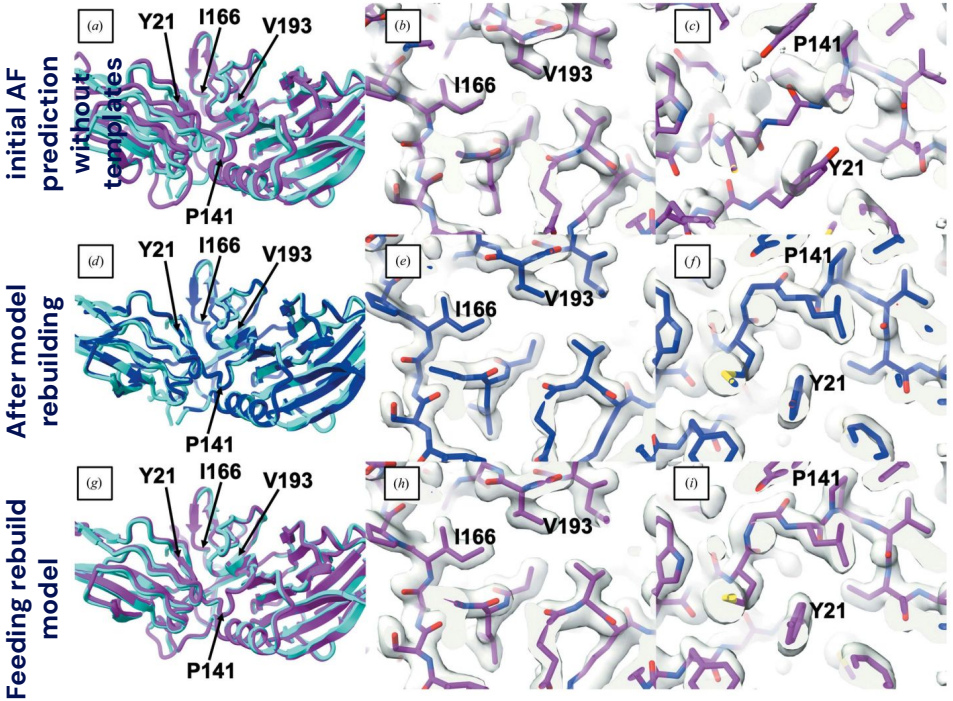


AlphaFold 2: Validation and impact

AlphaFold structures closely resemble crystal structures

- This procedure was applied to X-ray data for 215 structures recently released by the Protein Data Bank.
- In 87% of cases procedure yielded a model with at least 50% of Cα atoms matching those in the deposited models within 2 Å.
- In 7 cases the initial MR step did not yield any solution and the analysis was not continued.

"AlphaFold predictions alone are usually accurate enough to solve the crystallographic phase problem with molecular replacement."

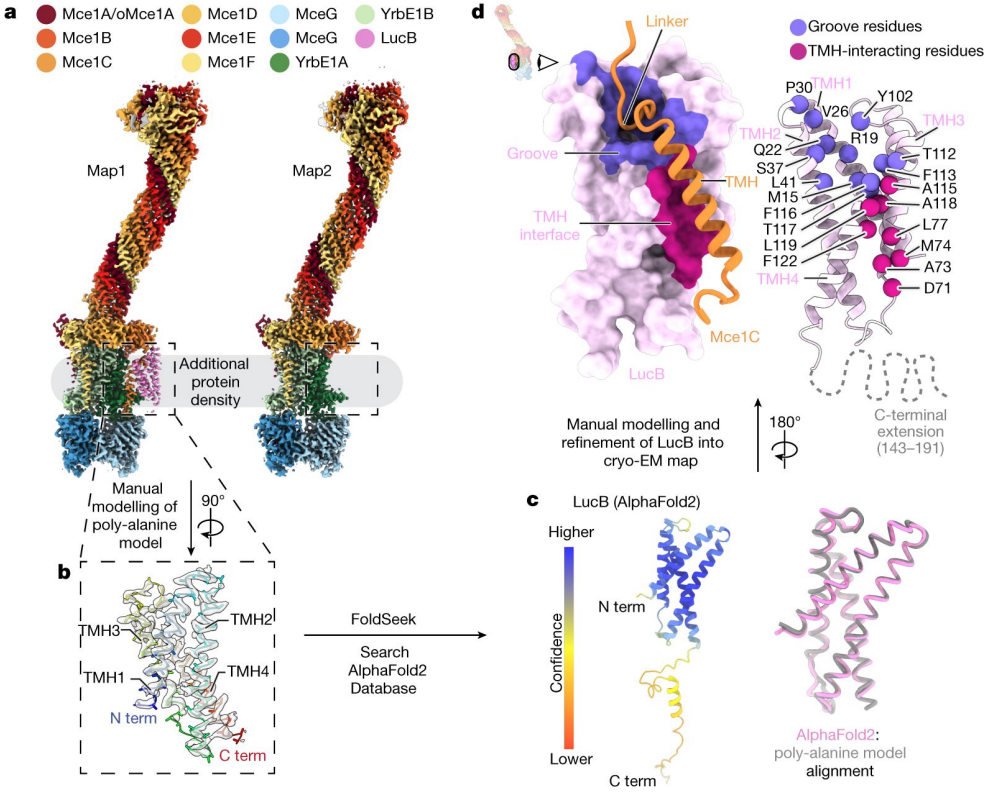


Iterative AlphaFold prediction and model rebuilding with X-ray data for PDB entry 7oa7 (cyan).

Terwilliger T. C. et al., ‚Accelerating crystal structure determination with iterative AlphaFold prediction. Acta Crystallographica Section D Structural Biology (2023)

Identifying unknown subunits

- AlphaFold2 was instrumental in resolving the structure of Mce1, a protein used by the tuberculosis bacterium to scavenge nutrients from host cells.
- Electron density revealed an additional unknown subunit associated with the ABC transporter.
- A structure-based search against the AlphaFold Database returned a hit that matched the unknown subunit, later named LucB.
- Fitting the AFDB model into the EM density required minimal adjustment apart from a few side-chain rotamer changes.

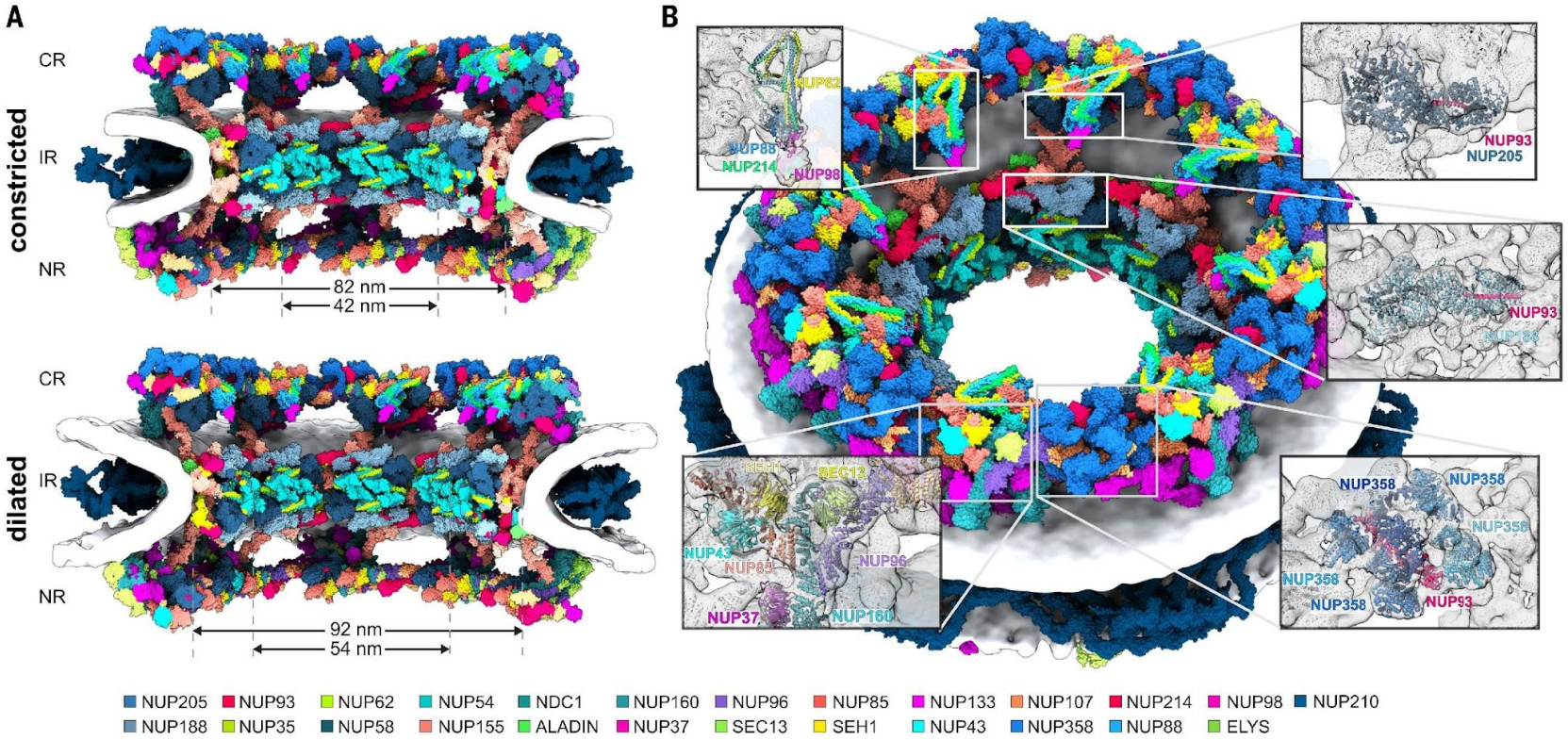


Chen, J., . et al. Structure of an endogenous mycobacterial MCE lipid transporter. Nature (2023).

Google DeepMind

PDBe Protein Data Bank in Europe EMBL-EBI

AlphaFold to solve huge complexes



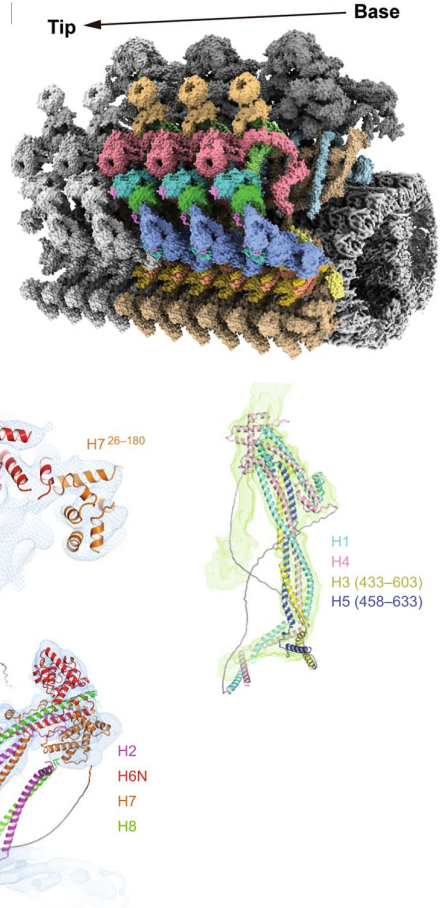
Google DeepMind

Shyamal Mosalaganti et al. ‚AI-based structure prediction empowers integrative structural analysis of human nuclear pores.Science (2022)

PDBe Protein Data Bank in Europe EMBL-EBI

And just large complexes

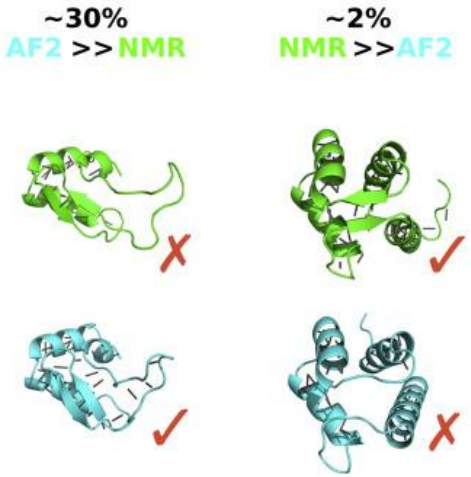
- Intraflagellar train IFT-A and IFT-B - **0.8 MDa** (Ma et al., 2023, McCafferty et al., 2022, Hesketh et al., 2022)
- Full-length kinesin-1 heavy chain and kinesin-1 heavy and light chains - **0.4 MDa** complex (Tan et al., 2023).
- Augmin complex (Gabel et al., 2022, Zupa et al., 2022).
- Components of the yeast small subunit processome (Zhao et al., 2022).
- Component of the eukaryotic lipid transport machinery (Cai et al., 2022).
- "**Better structure reconstruction can be obtained by combining AlphaFold predicted structure models and cryo-EM data**" (Giri et al., 2023).
- Human Bre1 complex bound to the nucleosome, using cryo-EM and AlphaFold (Onishi, S. et al., 2024).
- **2.6 MDa** complex of yeast fatty acid synthase (Tüting et al., 2023)



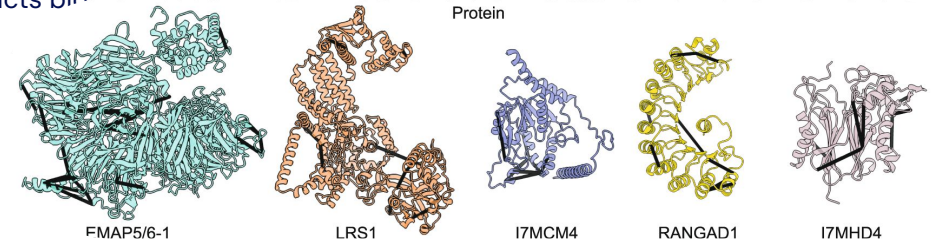
Google DeepMind

PDBe Protein Data Bank in Europe EMBL-EBI

What about NMR and *in vivo*?



- AlphaFold models fit NMR data as good as, or even better, than original NMR ensembles (Fowler, N.J. et al., 2022). On a set of 904 human proteins:
  - **2%** of NMR ensembles had ANSURR score at least 50 greater than the AF2 structure.
  - **30%** of AF2 structures had ANSURR score at least 50 greater than NMR ensembles.
  - AlphaFold models tend to contain more hydrogen bonds than NMR ensembles built based purely on NMR data, and these hydrogen bonds tend to persist in solution.
- Study of the 100 best-sampled proteins of the intact Tetrahymena thermophila cilia (compared intramolecular distance restraints obtained from in situ chemical cross-linking and cross-links) (McCafferty, C.L. et al., 2023).
  - **86.2%** of cross-linked residues map onto AlphaFold structures.
  - **43%** of the proteins showed no violations at all, and *inconsistencies observed in the other proteins occurred in low-confidence regions* or between protein domains.
  - AlphaFold "predicts biologically relevant protein conformations"



Google DeepMind

PDBe Protein Data Bank in Europe EMBL-EBI

AlphaFold structures closely resemble crystal structures

- Structural biology experiments demonstrate that AlphaFold structures (or well-defined parts of the predicted structures, like protein domains) **work well as search models for molecular replacement** (MR) in X-ray crystallography.
  - Molecular replacement works well if RMSD < 1–1.5 Å for at least 30% of the atoms.
  - Availability of high-accuracy models is then definitely critical for successful structure solution.

The X-ray crystallography phase problem solved thanks to *AlphaFold* and *RoseTTAFold* models: a case-study report

I. Barbarin-Bocahu and M. Graille

The breakthrough recently made in protein structure prediction by deep-learning programs such as *AlphaFold* and *RoseTTAFold* will certainly revolutionize biology over the coming decades. The scientific community is only starting to appreciate the various applications, benefits and limitations of these protein models. Yet, after the first thrills due to this revolution, it is important to evaluate the impact of the proposed models and their overall quality to avoid the misinterpretation or overinterpretation of these models by biologists. One of the first applications of these models is in solving the `phase problem' encountered in X-ray crystallography in calculating electron-density maps from diffraction data. Indeed, the most frequently used technique to derive electron-density maps is molecular replacement. As this technique relies on knowledge of the structure of a protein that shares strong structural similarity with the studied protein, the availability of high-accuracy models is then definitely critical for successful structure solution. After the collection of a 2.45 Å resolution data set, we struggled for two years in trying to solve the crystal structure of a protein involved in the nonsense-mediated mRNA decay pathway, an mRNA quality-control pathway dedicated to the elimination of eukaryotic mRNAs harboring premature stop codons. We used different methods (isomorphous replacement, anomalous diffraction and molecular replacement) to determine this structure, but all failed until we straightforwardly succeeded thanks to both *AlphaFold* and *RoseTTAFold* models. Here, we describe how these new models helped us to solve this structure and conclude that in our case the *AlphaFold* model largely outcompetes the other models. We also discuss the importance of search-model generation for successful molecular replacement.

Implications of *AlphaFold*2 for crystallographic phasing by molecular replacement

Airlie J. McCoy,[a]* Massimo D. Sammito[a] and Randy J. Read[a]

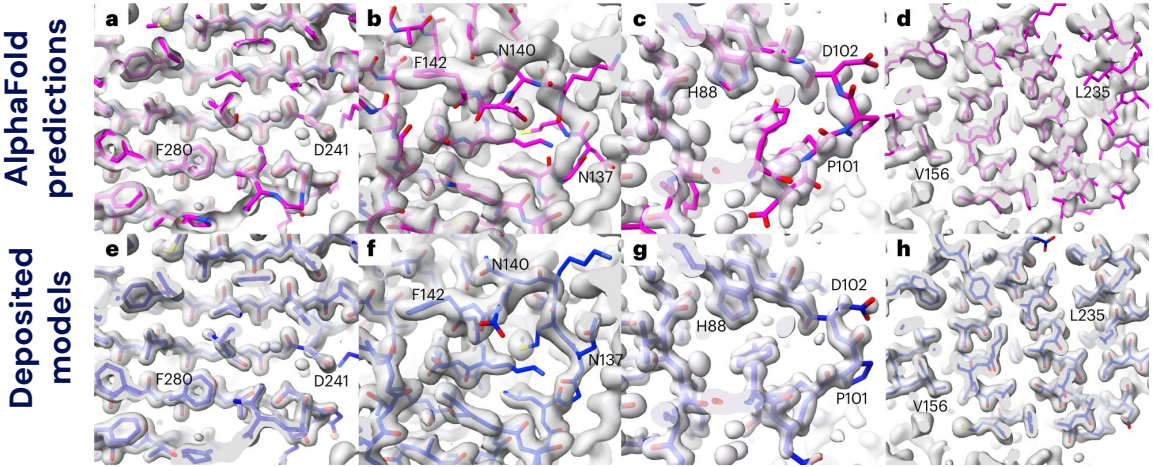
[a]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Hills Road, Cambridge CB2 0XY, United Kingdom
*Correspondence e-mail: ajm201@cam.ac.uk

Edited by D. J. Rigden, University of Liverpool, United Kingdom (Received 18 May 2021; accepted 13 November 2021)

The *AlphaFold*2 results in the 14th edition of Critical Assessment of Structure Prediction (CASP14) showed that accurate (low root-mean-square deviation) *in silico* models of protein structure domains are on the horizon, whether or not the protein is related to known structures through high-coverage sequence similarity. As highly accurate models become available, generated by harnessing the power of correlated mutations and deep learning, one of the aspects of structural biology to be impacted will be methods of phasing in crystallography. Here, the data from CASP14 are used to explore the prospects for changes in phasing methods, and in particular to explore the prospects for molecular-replacement phasing using *in silico* models.

AlphaFold predictions are valuable hypotheses

- Main chains:
  - Median RMSD values for the AlphaFold predictions is 1.0 Å, versus 0.6 Å for experimental structures.
  - pLDDT<70 median RMSD roughly 3.5 Å.
  - pLDDt > 90,median RMSD is 0.6 Å.
- Side chains:
  - 20% of the side chains are substantially different, 7% were clearly incompatible with the experimental data.



All the residues shown were predicted with very high confidence (pLDDT > 90) and the density maps range in resolution from 1.1 Å to 1.6 Å.

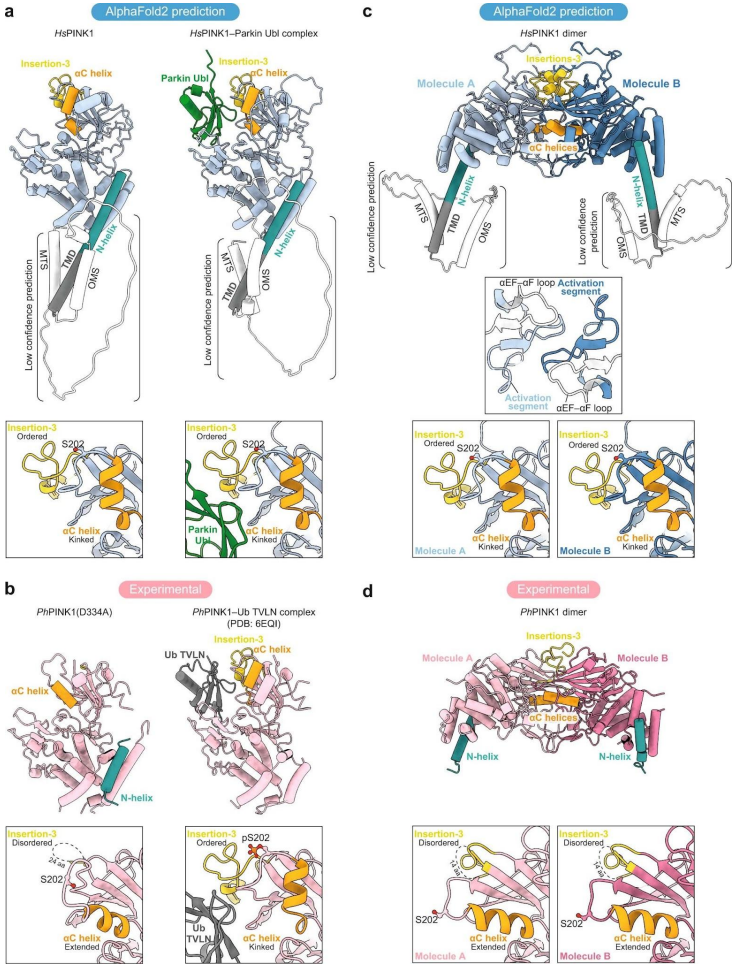
Google DeepMind

Terwilliger, T.C et al. AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. Nat Methods (2024)

PDBe Protein Data Bank in Europe EMBL-EBI

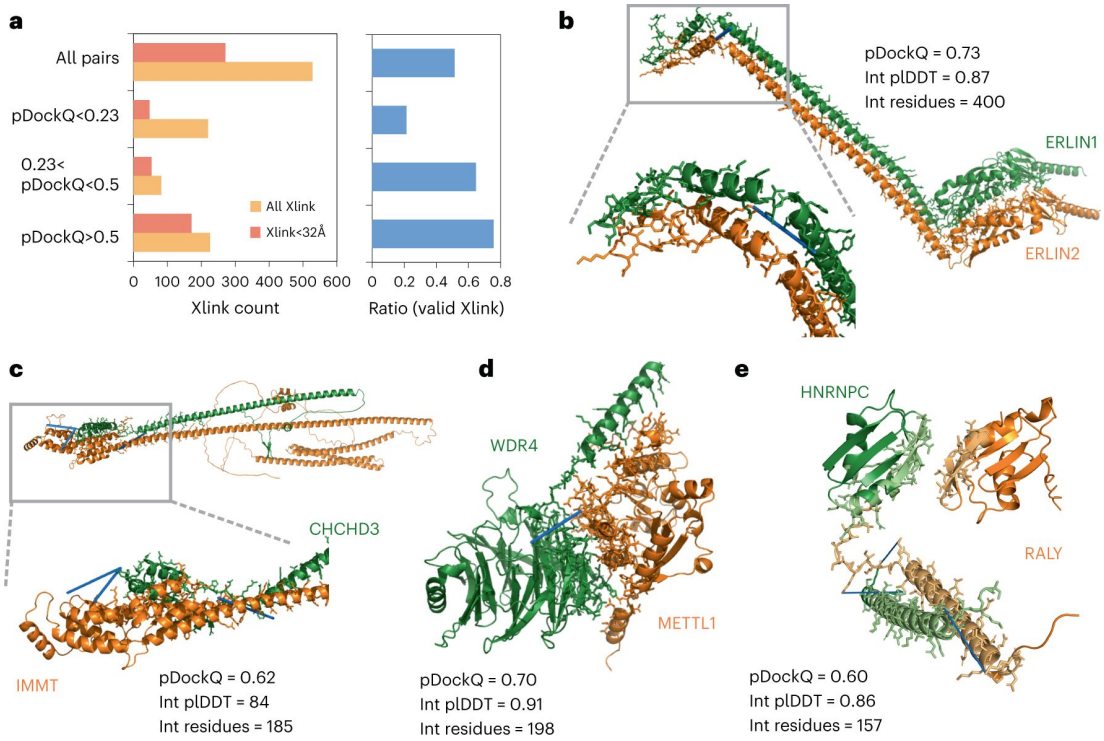
Propose mechanisms of protein action

- Mutations in the human protein PINK1 cause autosomal recessive early onset Parkinson's disease.
- Predicting the structure of HsPINK1 via AlphaFold 2 (a, left), resulted in a model remarkably similar to PhPINK1 from the PhPINK1–Ub TVLN complex.
- This provided evidence that human PINK1 uses the same mechanism as insect PINK1, which is better characterised both structurally and functionally.
- The human pathogenic mutations could then be analysed in the context of the protein's structure



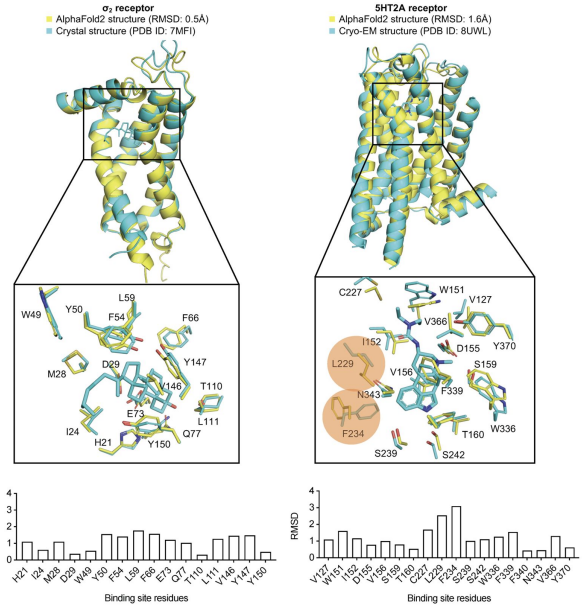
Large scale predictions of protein interactions

- Predicted the structures of over 65,000 pairs of unique human interactions, obtained 3,137 high-confidence pairs.
- Cross-linking validation:
  - Experimental evidence confirmed 171 out of 479 complexes predicted with high confidence.
  - 41 correspond to complex structures with no experimental structure or homology models,

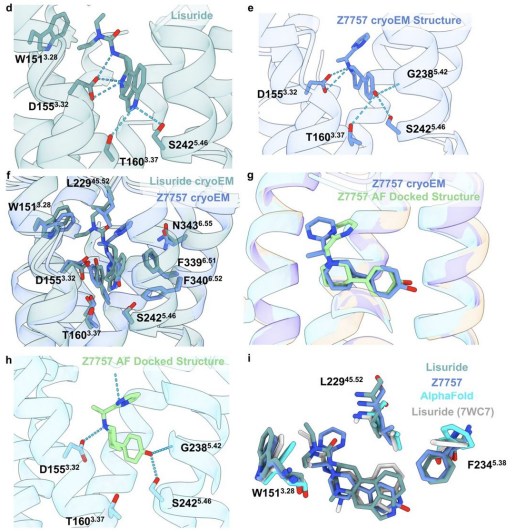


Burke, D. F. et al., Towards a structurally resolved human protein interaction network. Nat. Struct. Mol. Biol. (2023).

Template ligand discovery



Structural comparisons between the AFDB structure and the experimental structure.

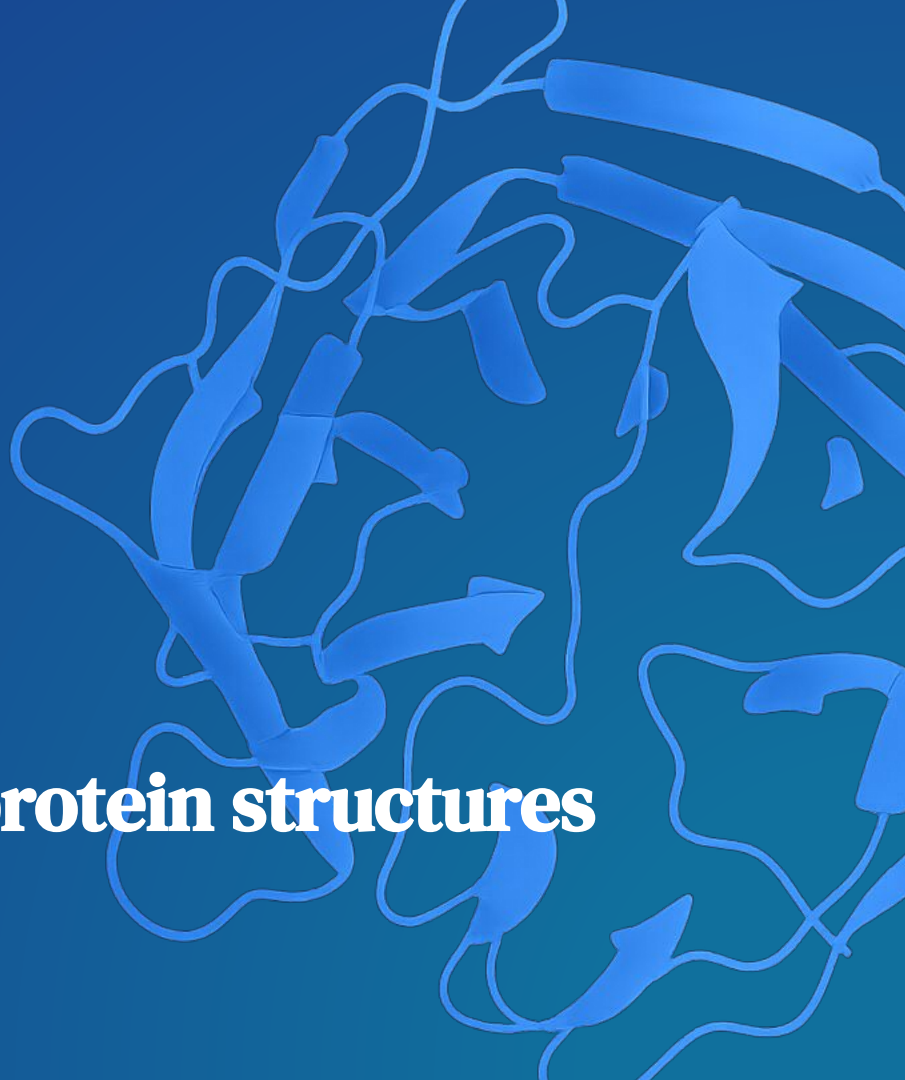


Structural characterisation of the 5-HT2A receptor in complex with Lisuride and Z7757.

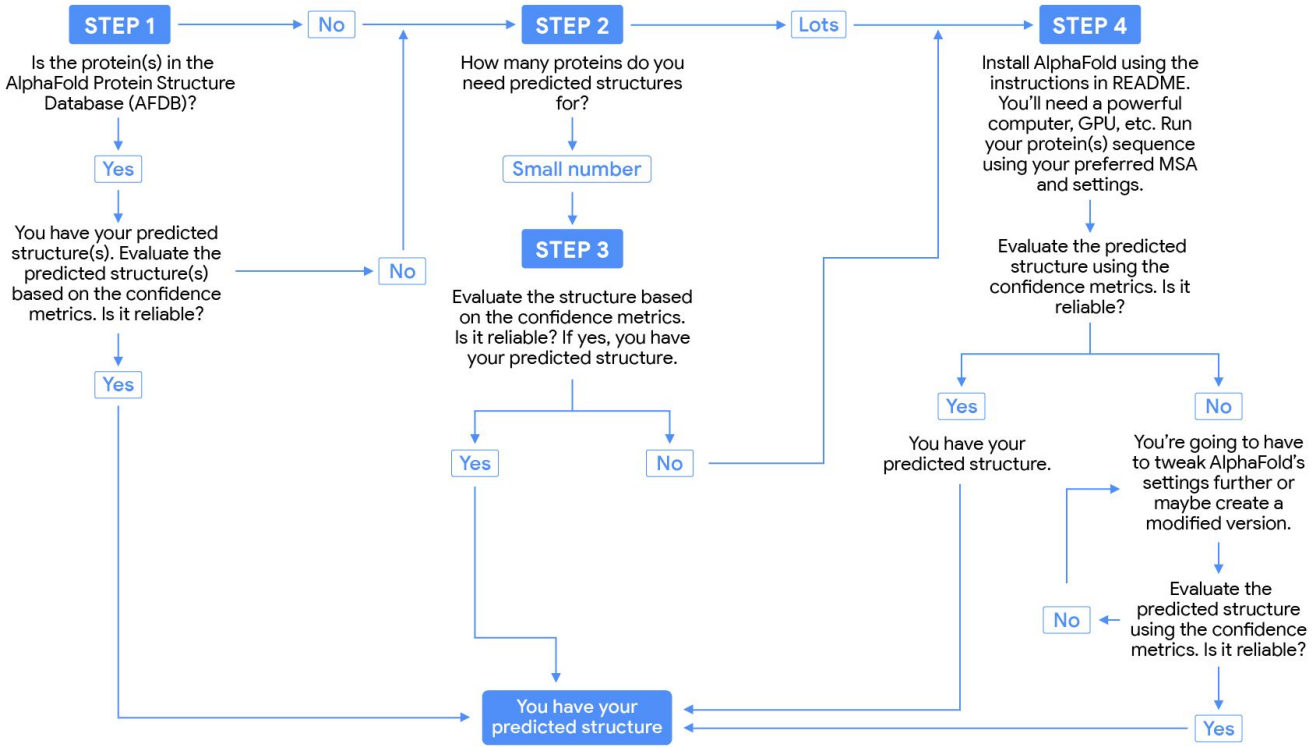
- Prospective large library docking campaigns against the AF2 models were no less effective than those against experimental structures.
- The AF2-derived docking hits are among the most potent and selective molecules to emerge from extensive structure-based campaigns against 5-HT2A.
- In the right circumstances, "*AF2 models can much expand the range of proteins targeted for structure-based discovery, by sampling conformations that are relevant for ligand discovery*".

Jiankun Lyu, et al, . AlphaFold2 structures template ligand discovery. bioRxiv 2023.

Accessing and predicting protein structures with AlphaFold 2

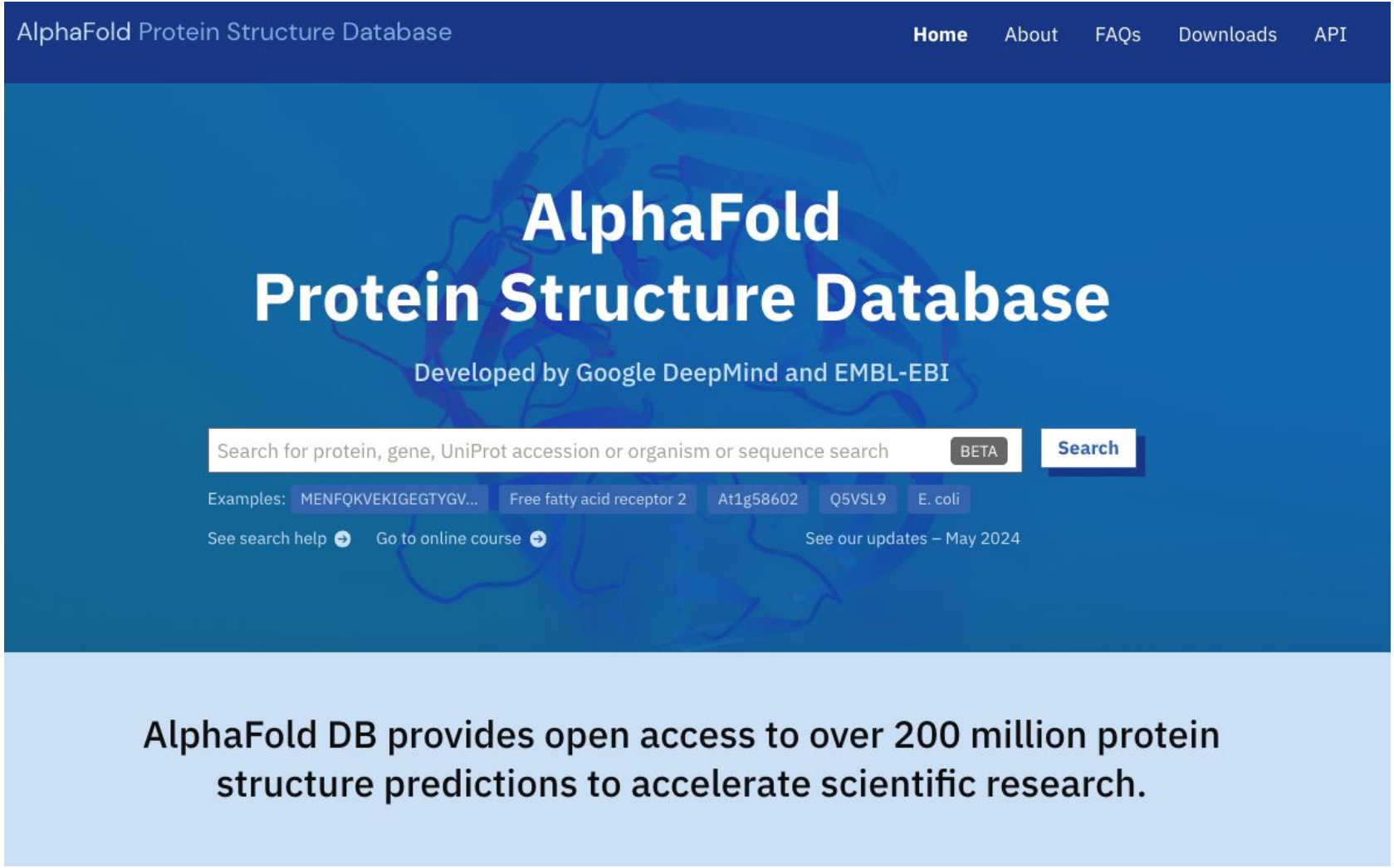


Choosing how to access AlphaFold2



Google DeepMind

PDBe Protein Data Bank in Europe EMBL-EBI


AlphaFold Protein Structure Database
Home
About
FAQs
Downloads
API
AlphaFold
Protein Structure Database
Developed by Google DeepMind and EMBL-EBI
Search for protein, gene, UniProt accession or organism or sequence search
BETA
Search
Examples:
MENFQKVEKIGEGTYGV...
Free fatty acid receptor 2
At1g58602
Q5VSL9
E. coli
See search help
Go to online course
See our updates – May 2024
AlphaFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research.

How can you search the database?

Seach Protein

Search

Showing all search results for

MENFQKVEKIGEGTYGVVYKARNKLTGEVVALKKIRLDTETEGVPSTAIREIS

1 - 20 of 2000 results from BLASTP

Filter by:

Status

Review

Reviewed (Swiss-Prot) (22)

Unreviewed (TrEMBL) (1978)

Reference proteome

Show predictions for sequences found only in UniProt reference proteomes (1794)

Average pLDDT score

Very low — Very high

Show results with the score between

0 and 100 Apply score range

Protein kinase domain-containing protein

AF-A0A4W3JAN5-F1-v4

Protein	Protein kinase domain-containing protein
Gene	cdk2
Source Organism	Callorhinchus milii search this organism
UniProt	A0A4W3JAN5 go to UniProt
Average pLDDT	82.19 (High)
Sequence length	332

Sequence match **HSP score:** 262 **E-value:** 6.12177e-26

Identity: 100% **Positives:** 53/53 (100%) **Gaps:** 0/53 (0%)

```
Your query      1  MENFQKVEKIGEGTYGVVYKARNKLTGEVVALKKIRLDTETEGVPSTAIREIS  53
                   MENFQKVEKIGEGTYGVVYKARNKLTGEVVALKKIRLDTETEGVPSTAIREIS
A0A4W3JAN5      1  MENFQKVEKIGEGTYGVVYKARNKLTGEVVALKKIRLDTETEGVPSTAIREIS  53
```

Sequence similarity-based search powered by the **Basic Local Alignment Search Tool (BLAST)**

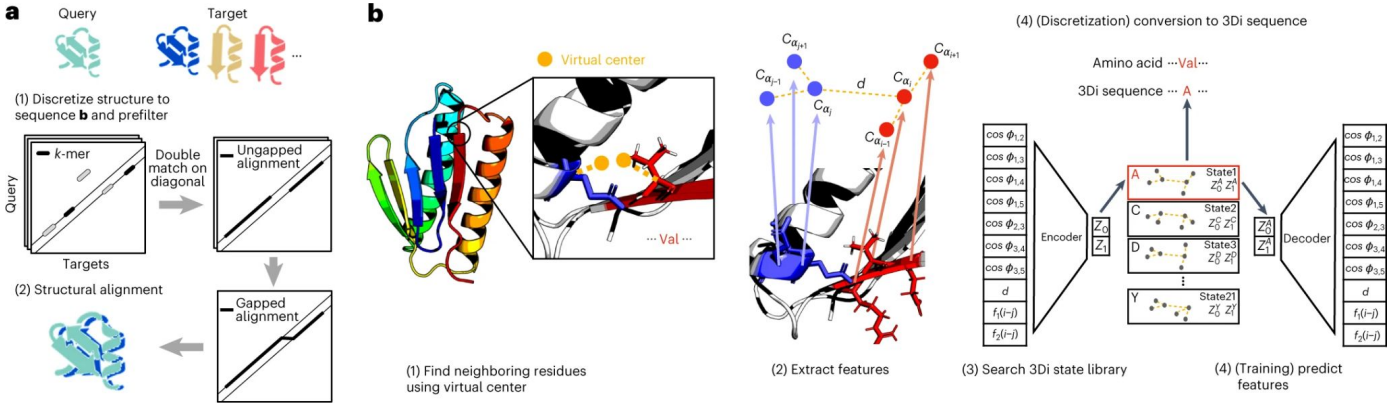


Google DeepMind

PDBe Protein Data Bank in Europe

Structure-based search in the AlphaFold Database using Foldseek





- Foldseek aligns the structure of a query protein against a database by describing tertiary amino acid interactions, using a unique '**3Di alphabet**'.

- This alphabet essentially translates the intricate local interactions within protein structures using the pre-trained 3Di substitution matrix. This process allows for rapid scanning of vast databases using MMseqs2, a highly sensitive and efficient sequence search tool, to pinpoint potential matches.

van Kempen, M., et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol (2023).

Prediction pages - Structure-based search

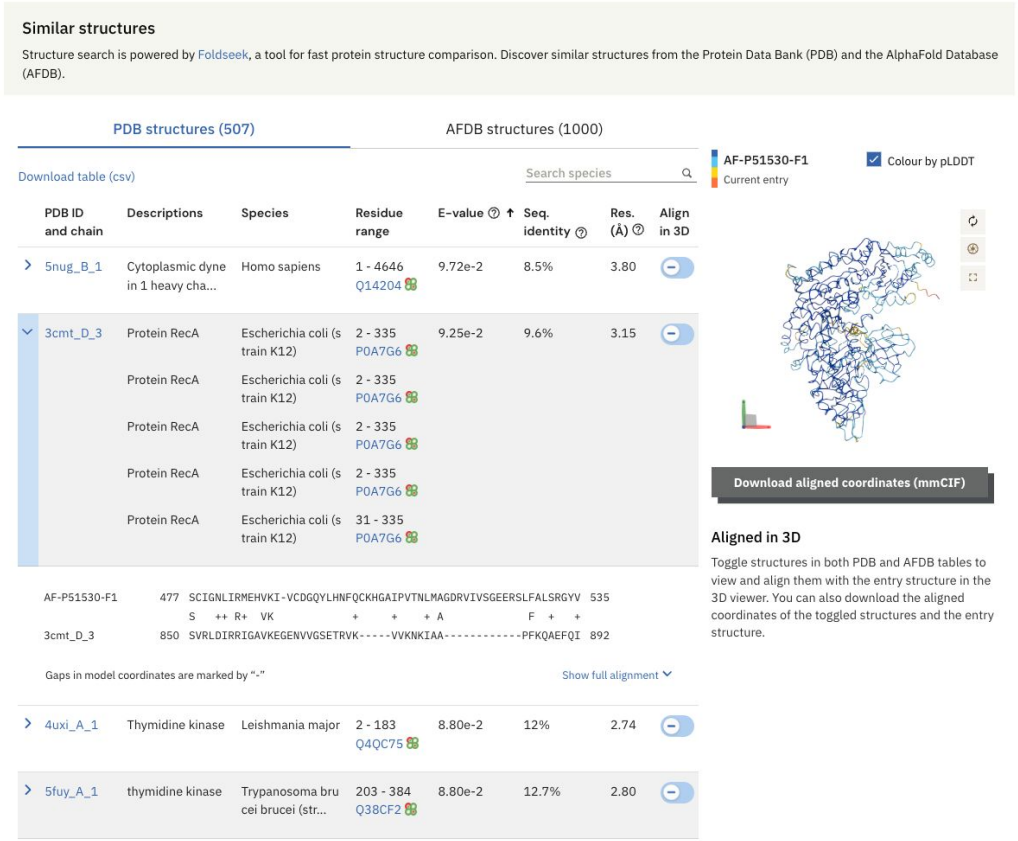
Similar functionality to structure similarity cluster table with added sorting capabilities with sequence identity and E-value

Search against:

AFDB50 (Clustered on 50% sequence identity)

PDBe (all experimental structures)(updated weekly)

Live search with Foldseek with an E-value threshold of 0.1



Google DeepMind

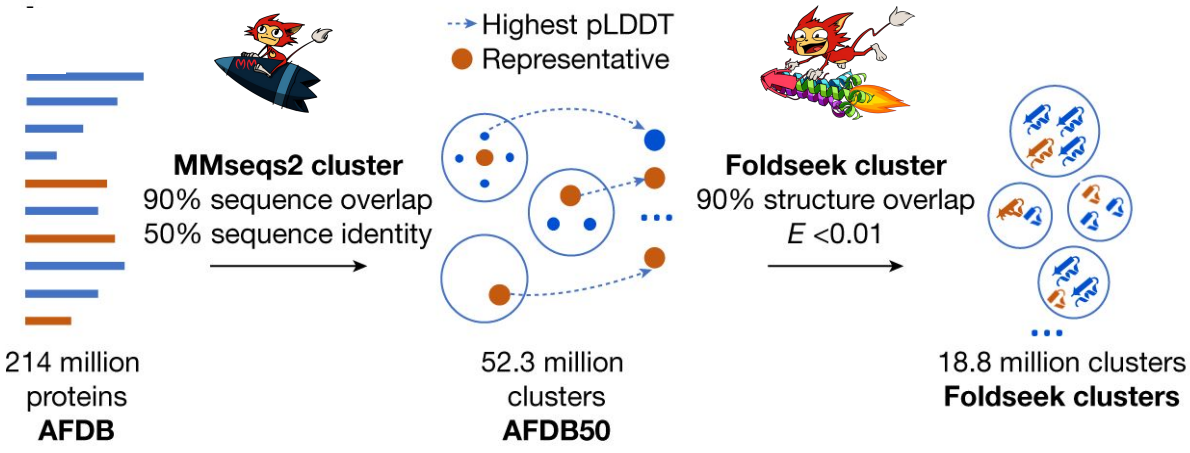
PDBe Protein Data Bank in Europe EMBL-EBI

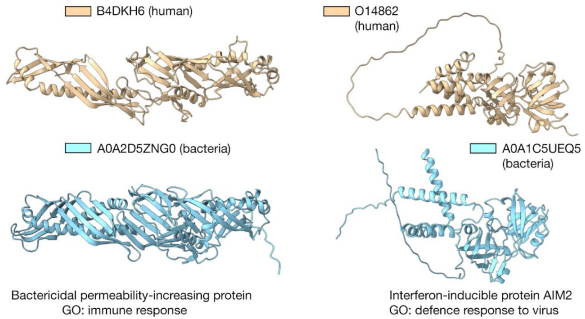
Clustering predicted structures at the scale of the known protein universe

Identified 2.30 M non-singleton structural clusters, of which 31% lack annotations representing probable previously undescribed structures.

>1.1M structures (49% of AFDB clusters) clusters were found, at least partially, similar to previously known structures in the PDB.

These clusters annotate only 4% of protein sequences, indicating that the vast majority of the protein structural space has been at least partially annotated.





The human innate immunity genes BPI (left) and AIM2 (right) encode structurally similar proteins in bacterial species, highlighting the potential for cross-kingdom sharing of immunity-related proteins.

Google DeepMind

Barrio-Hernandez, I., et al. Clustering predicted structures at the scale of the known protein universe. Nature (2023).

PDBe Protein Data Bank in Europe EMBL-EBI

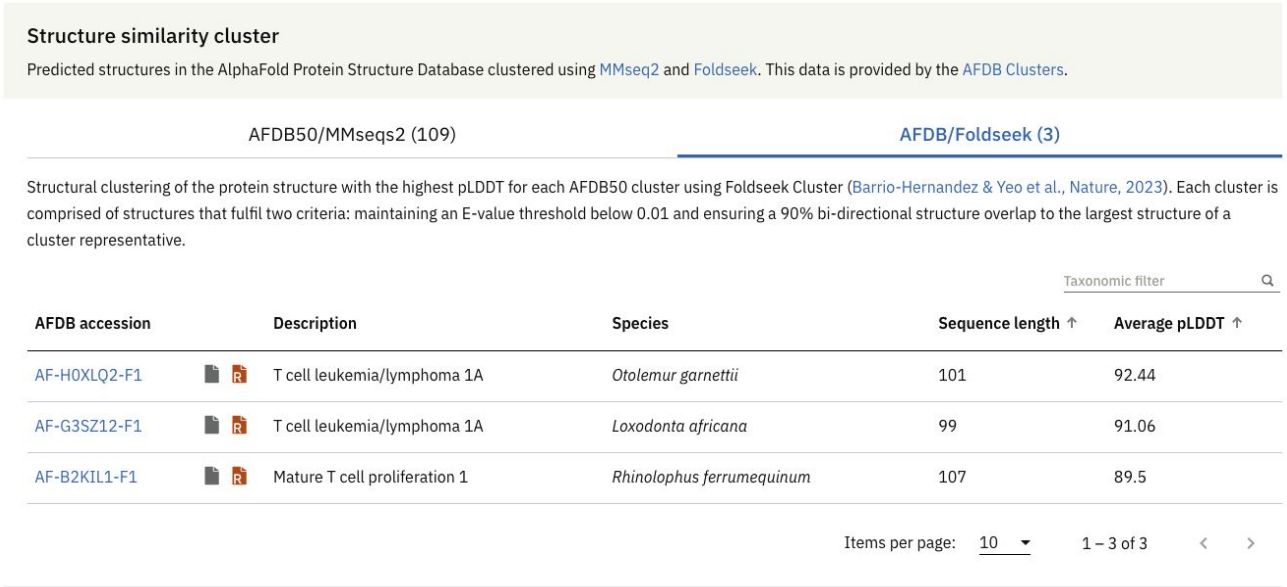
Shedding light on the structural organization of the whole AlphaFold database

Clicking between two tabs to navigate between the two clustering steps

Each entry includes the AFDB accession, UniProt description, Species, AF model sequence length and average pLDDT

Sorting by sequence length and average pLDDT

Taxonomic filter to filter by species



Structure similarity cluster

Predicted structures in the AlphaFold Protein Structure Database clustered using MMseq2 and Foldseek. This data is provided by the AFDB Clusters.

AFDB50/MMseqs2 (109) | AFDB/Foldseek (3)

Structural clustering of the protein structure with the highest pLDDT for each AFDB50 cluster using Foldseek Cluster (Barrio-Hernandez & Yeo et al., Nature, 2023). Each cluster is comprised of structures that fulfil two criteria: maintaining an E-value threshold below 0.01 and ensuring a 90% bi-directional structure overlap to the largest structure of a cluster representative.

Taxonomic filter

AFDB accession	Description	Species	Sequence length ↑	Average pLDDT ↑
AF-H0XLQ2-F1	T cell leukemia/lymphoma 1A	*Otolemur garnettii*	101	92.44
AF-G3SZ12-F1	T cell leukemia/lymphoma 1A	*Loxodonta africana*	99	91.06
AF-B2KIL1-F1	Mature T cell proliferation 1	*Rhinolophus ferrumequinum*	107	89.5

Items per page: 10 1 – 3 of 3

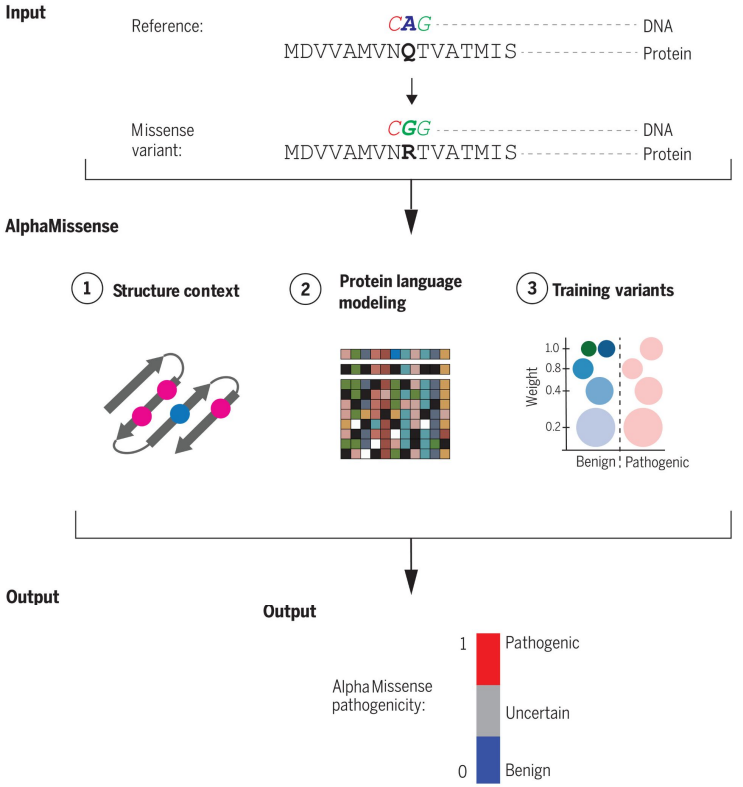
Protein p13 MTCP-1 (AF-P56278-F1)
Homo sapiens (Human)

Google DeepMind

PDBe Protein Data Bank in Europe EMBL-EBI

AlphaMissense: A catalogue of genetic mutations

- Fine-tuned AlphaFold on labels distinguishing variants seen in human and closely related primate populations.
- To predict whether a variant is pathogenic, AlphaMissense combines two types of data.
  - The degree to which the amino acid residue has been conserved over evolutionary history. If a residue is deeply conserved, it is likely to be important for protein function. Changes to such conserved residues are more likely to be pathogenic.
  - The structural context of the variant. Depending on where an amino acid is loca within a protein, it is more or less likely to be critical to function.



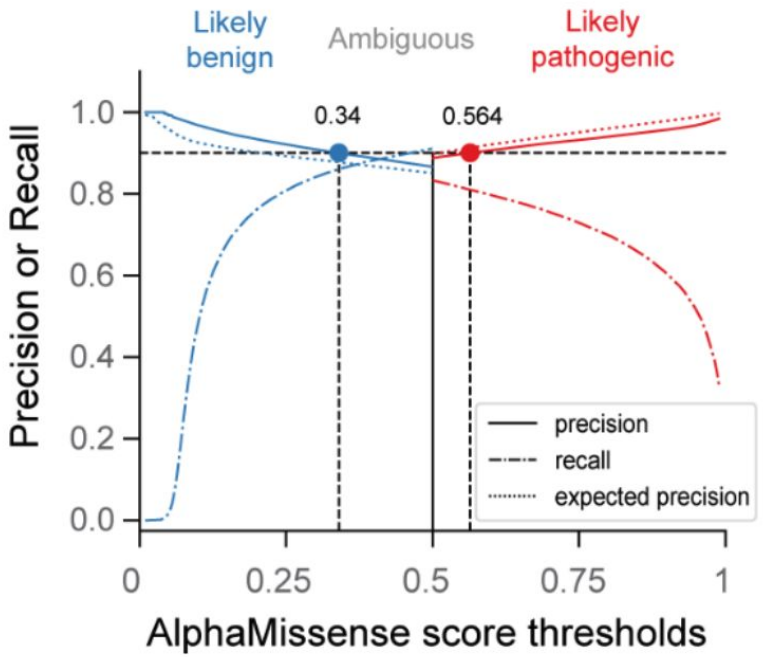
Google DeepMind

PDBe Protein Data Bank in Europe EMBL-EBI

Understanding pathogenicity scores from AlphaMissense

- The pathogenicity scores are generated by rescaling raw AlphaMissense data using a logistic regression model trained on data from ClinVar. The aim was to generate a single number for ease of interpretation.

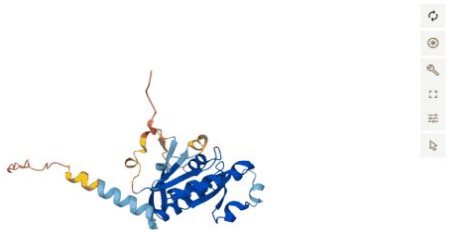
- To further facilitate interpretation, scores were divided into three categories:
  - 0 to 0.34: likely benign
  - 0.34 to 0.564: uncertain
  - 0.564 to 1: likely pathogenic

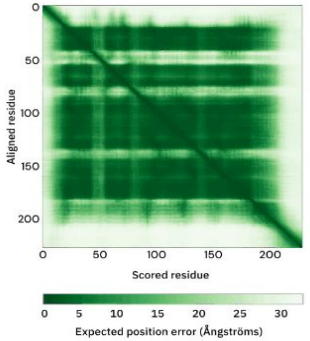


Precision and recall of AlphaMissense predictions on the ClinVar test set

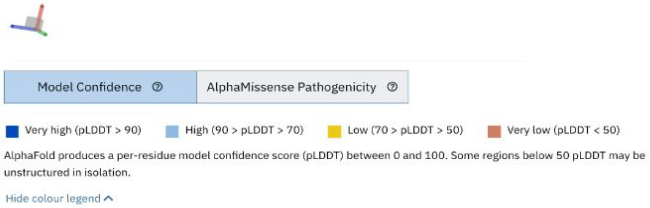
AlphaMissense in the AlphaFold Database

Rab-like protein 2A (AF-Q9UBK7-F1)





You can toggle between pLDDT and average AlphaMissense pathogenicity

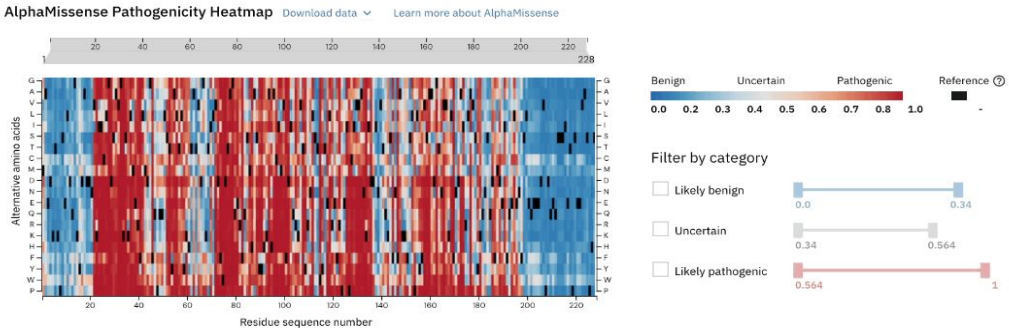


Predicted aligned error (PAE)

Click and drag a box on the PAE viewer to select regions of the structure and highlight them on the 3D viewer.

PAE data is useful for assessing inter-domain accuracy – go to Help section below for more information.

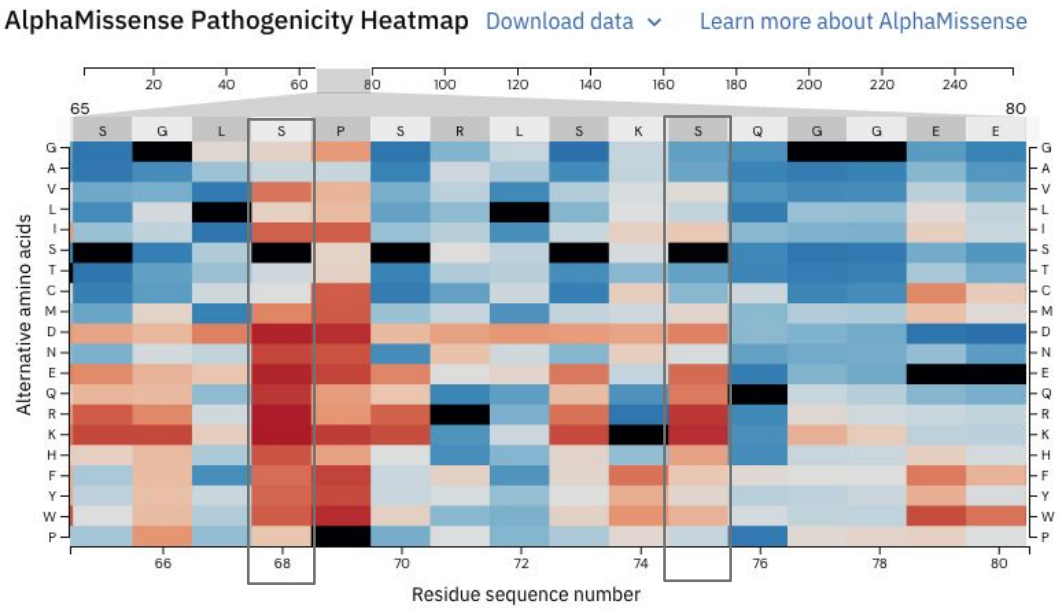
Interactive heatmap
Zoom in and select residue of interest

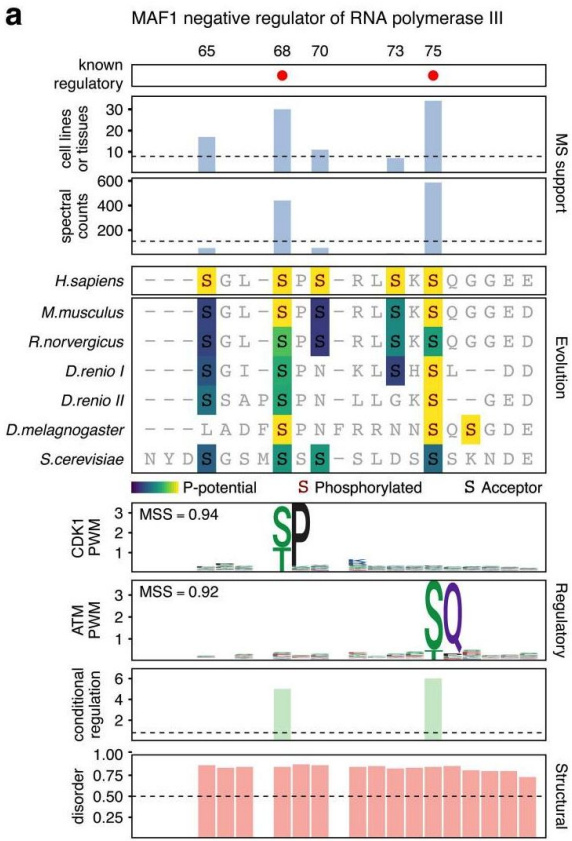


Filter heatmap view by category and adjust range

Google DeepMind

PDBe Protein Data Bank in Europe EMBL-EBI



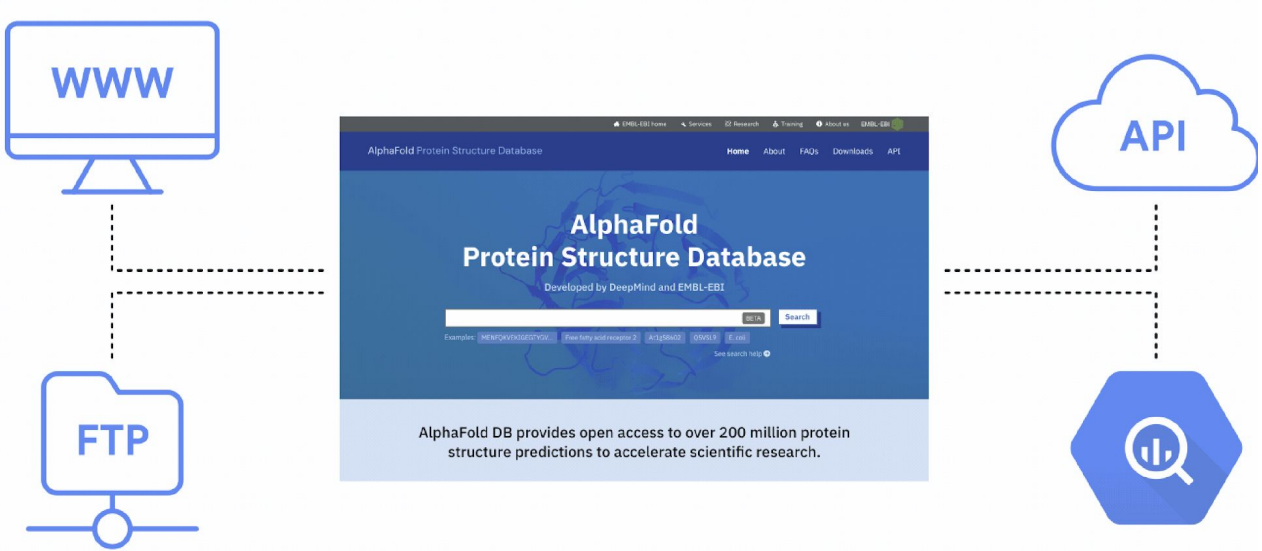


Repressor of RNA polymerase III transcription MAF1 homolog
(AF-Q9HO63-F1-v4)

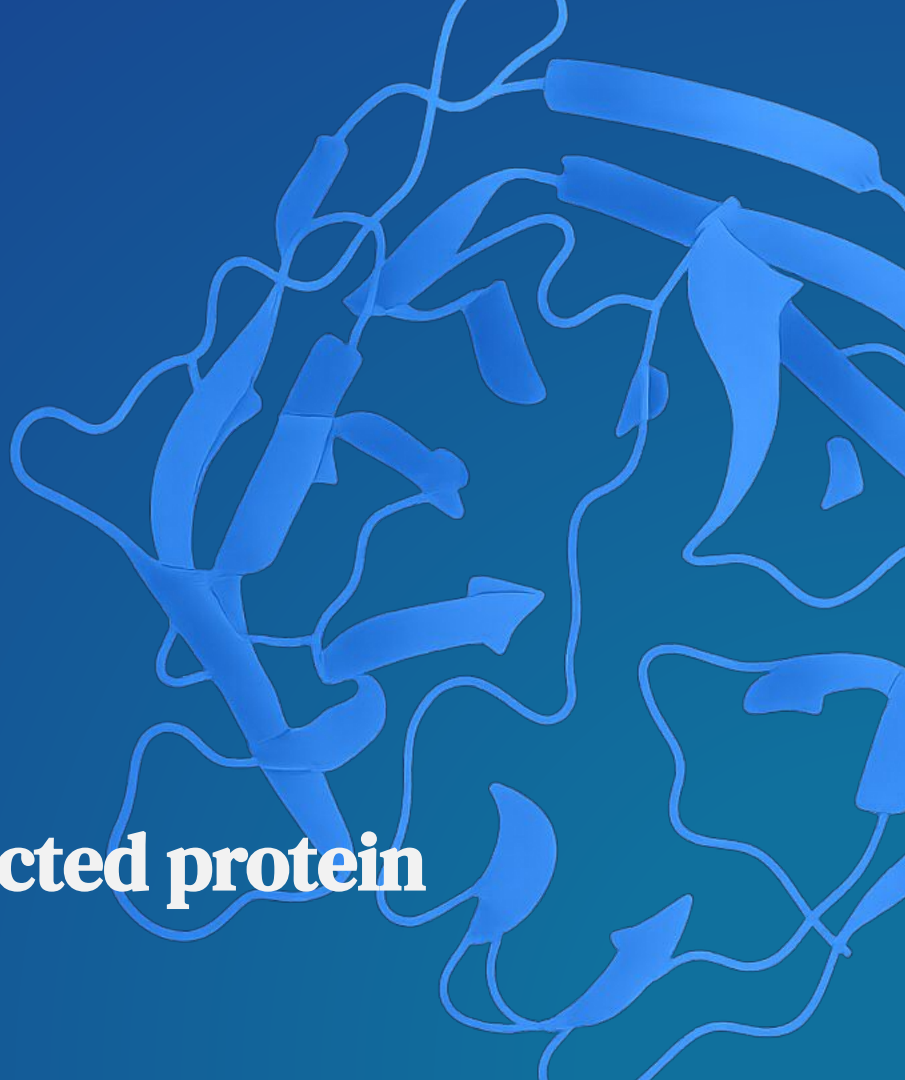
Limitations of AlphaMissense

- AlphaMissense **does not directly provide detailed structural change** information for altered sequences.
- AlphaMissense does not explicitly predict missense variants' impacts on **biophysical properties,** like, stability and binding affinity.
- AlphaMissense is primarily designed to predict the effects of amino acid substitutions in **individual proteins, does not cover insertions, deletions or epistatic interactions** of multiple missense variants.
- Polygenic traits often involve **gene-environment interactions.** Environmental exposures can modulate the phenotypic effects of mutations.
- Performance might be **limited** for proteins with limited evolutionary information .

How can you access the database?

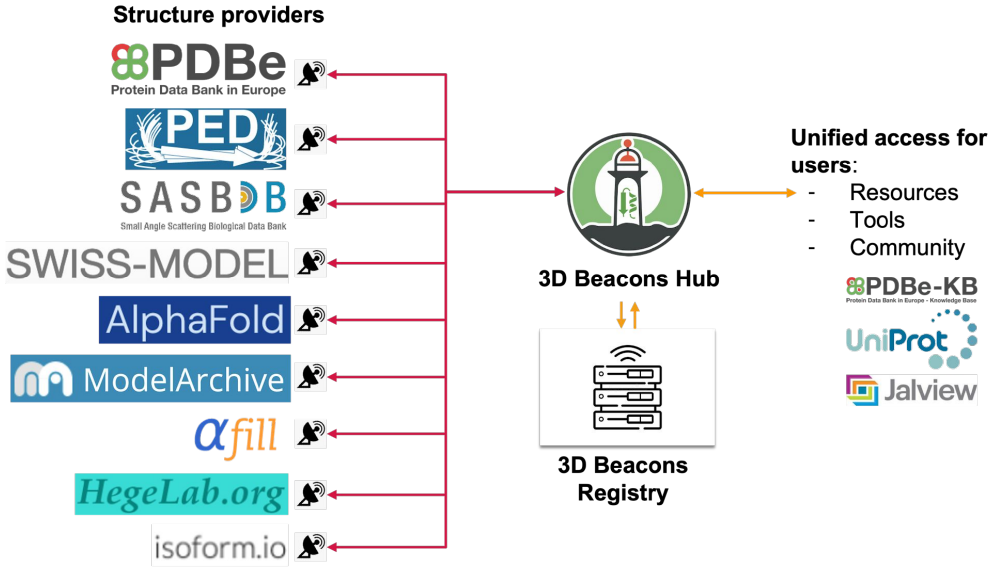


Other ways to access predicted protein structures



3D Beacons Network

- The 3D-Beacons Network is an open collaboration of macromolecular structure model providers.
- Simplified access to all models
    - Common API specification to access structure models and annotations
- Standardise mechanism to provide quality-assessment data for all structure models.
- All data on the 3D-Beacons Network is freely available for both academic and commercial use, under the Creative Commons Attribution 4.0 (CC-BY 4.0) licence.



M. Varadi, S. Nair, I. Sillitoe et al. 3D-Beacons: decreasing the gap between protein sequences and structures through a federated network of protein structure data resources. GigaScience (2022).

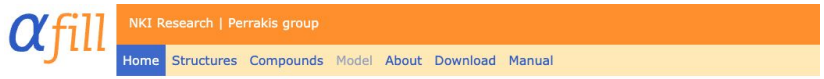
Other tools using AlphaFold

Databases:

- AlphaFill - “transplanting” ligands.
- TmAlphaFold Database - transmembrane proteins.
- O-GlcNAcome Database - mapping glycosylation.
- 3D-Beacons network.

Software:

- AlphaPulldown - computational pulldowns.
- StructureMap - map PTMs on AFDB.
- MoLPC - Monte Carlo tree for modelling large complexes.
- CombFold - building large complexes.



AlphaFill

AlphaFold models enriched with ligands and co-factors

Show an entry based on Uniprot ID

Uniprot ID | Search for Uniprot ID

Submit

Create your own AlphaFill model

Structure File | Choose file | No file chosen

Submit

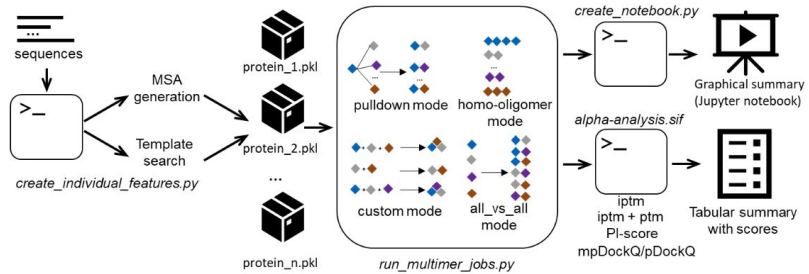
Hekkelman, M.L. et al. AlphaFill: enriching AlphaFold models with ligands and cofactors. Nat Methods (2023).

EMBL

Research Units > Structural biology > Kosinski group > **AlphaPulldown** Related: EMBL Hamburg

AlphaPulldown

AlphaPulldown, a Python package that streamlines protein-protein interaction screens and high-throughput modelling of higher-order oligomers using AlphaFold-Multimer. It provides a convenient command line interface, a variety of confidence scores, and a graphical analysis tool.

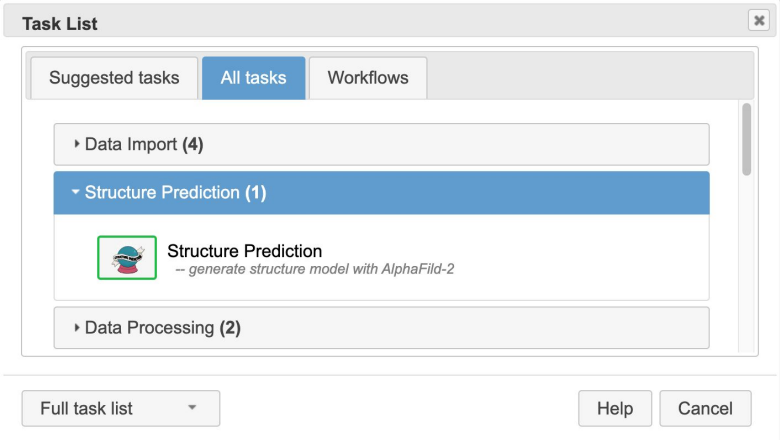


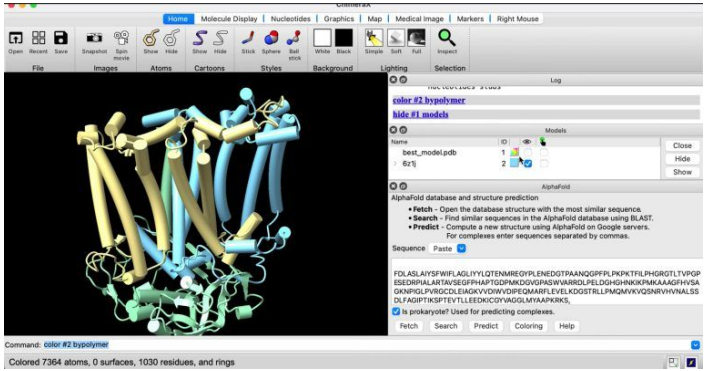
Yu, D. et al., AlphaPulldown—a python package for protein–protein interaction screens using AlphaFold-Multimer, Bioinformatics (2023)

Google DeepMind

PDBe Protein Data Bank in Europe EMBL-EBI

Software that integrated AlphaFold/ColabFold/AFDB



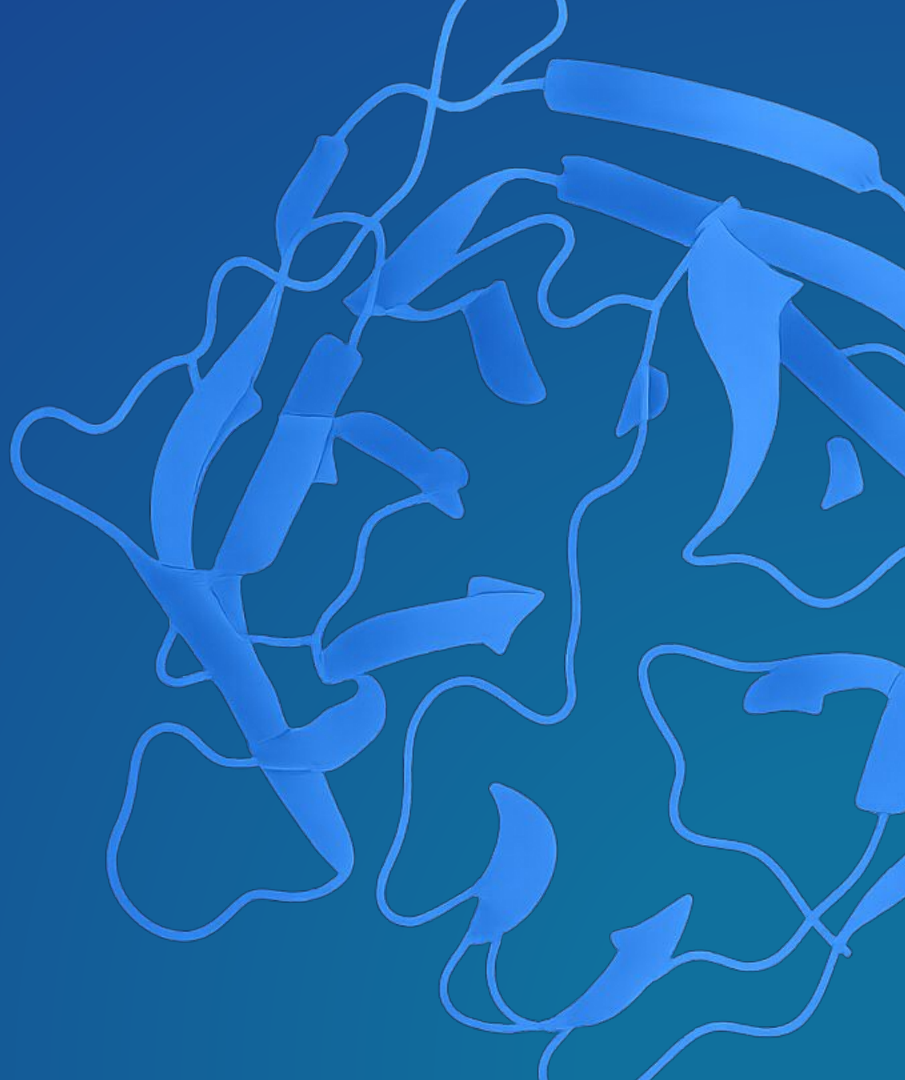


- CCP4 (MRBUMP, MRPARSE, Slice'n'Dice, CCP4Cloud, etc)
- PHENIX (phenix.process_predicted_model, etc)
- CCP-EM
- COOT
- PyMol plugin for AFDB import
- ChimeraX (interface to ColabFold)
- DALI server
- Foldseek
- Jalview
- Mol* viewer

Google DeepMind

PDBe Protein Data Bank in Europe EMBL-EBI

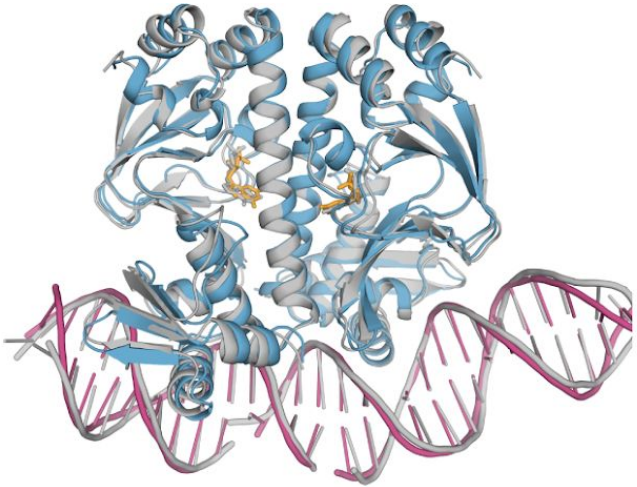
Introducing AlphaFold3 & AlphaFold Server



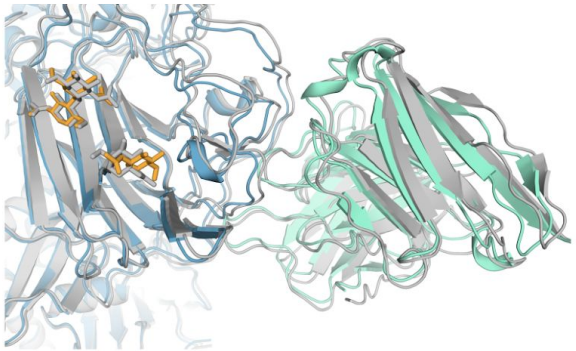
AlphaFold 3 (AF3)

- ✅ Proteins
- ✅ RNA
- ✅ DNA
- ✅ Small molecules
- ✅ Ions
- ✅ Modified residues
- ✅ Glycosylation
- ✅ ...
- ❌ Water
- ❌ Hydrogen

Example structures predicted using AF3



Bacterial CRP/FNR family transcriptional regulator protein bound to DNA and cGMP (PDB 7PZB; full-complex LDDT, 82.8; global distance test (GDT) 90.1).



Human coronavirus OC43 spike protein, 4,665 residues, heavily glycosylated and bound by neutralizing antibodies (PDB 7PNM; full-complex LDDT: 83.0; GDT: 83.1)

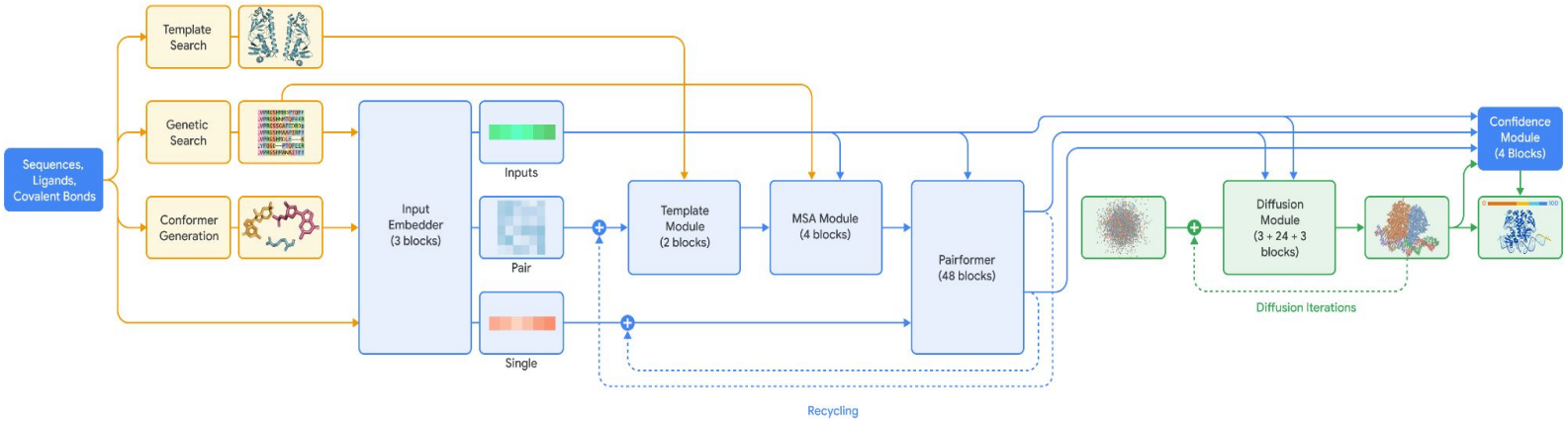
Google DeepMind

Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024).

PDBe Protein Data Bank in Europe EMBL-EBI

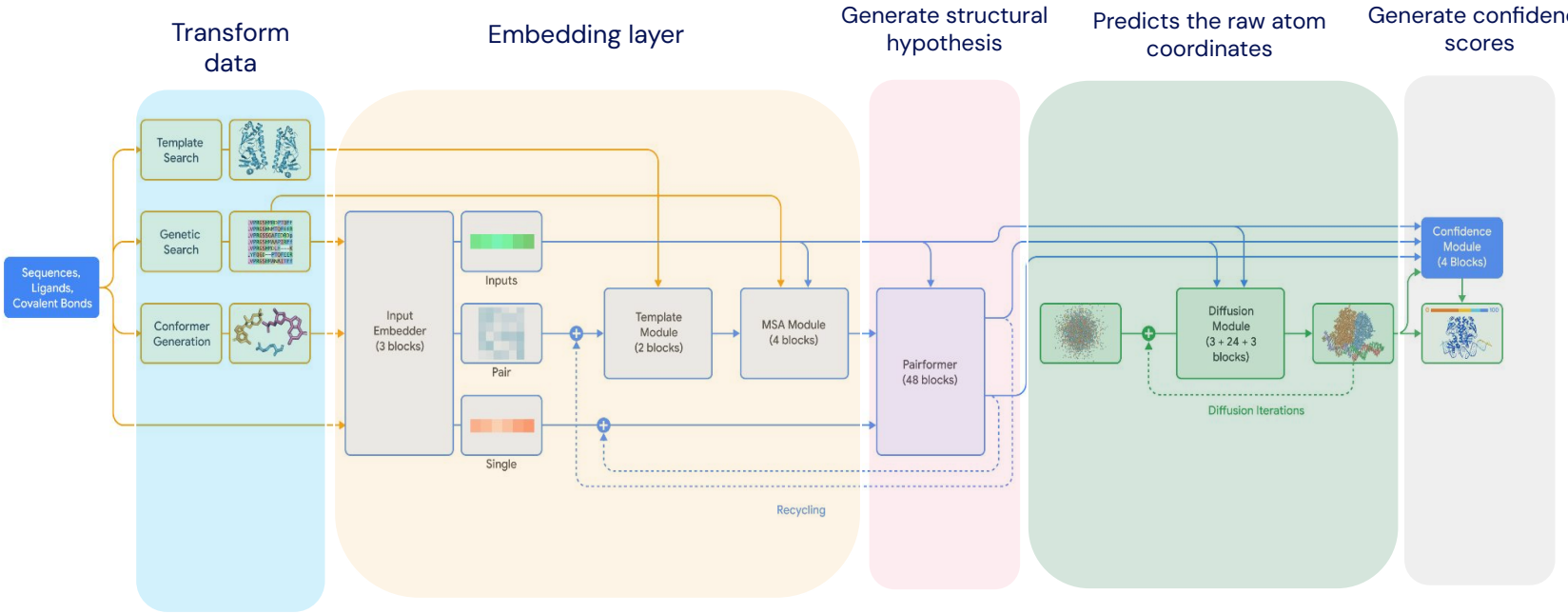
AlphaFold 3 architecture

AlphaFold 3 predicts raw atom coordinates using a diffusion module, making it a “generative” ML model. The diffusion module was trained to receive “noised” atomic coordinates, and to use them to predict the correct coordinates



Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024).

High-level overview

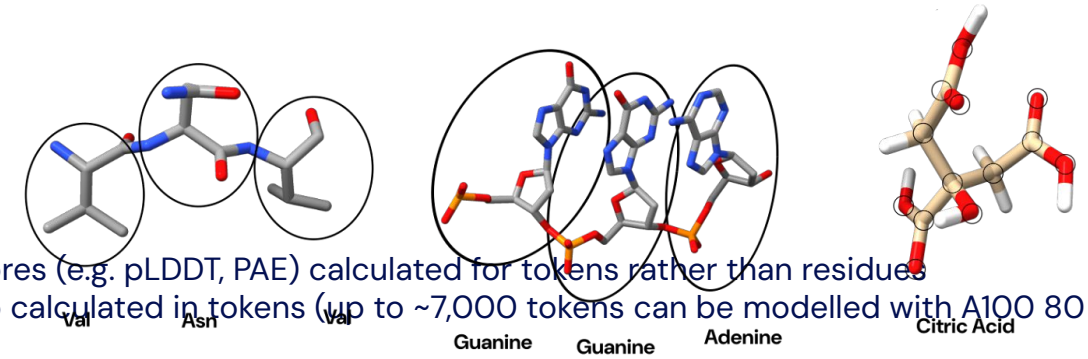


Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024).

AlphaFold 3 tokens

- Using transformer architecture, both AF2 and AF3 divide the complex of interest into “tokens”. In AF3, a token may correspond to:

One standard amino acid in the protein chain
One standard nucleotide in the nucleic acid chain
One atom of a ligand
One atom of an ion
One atom of a chemically-modified amino acid residue or nucleotide.

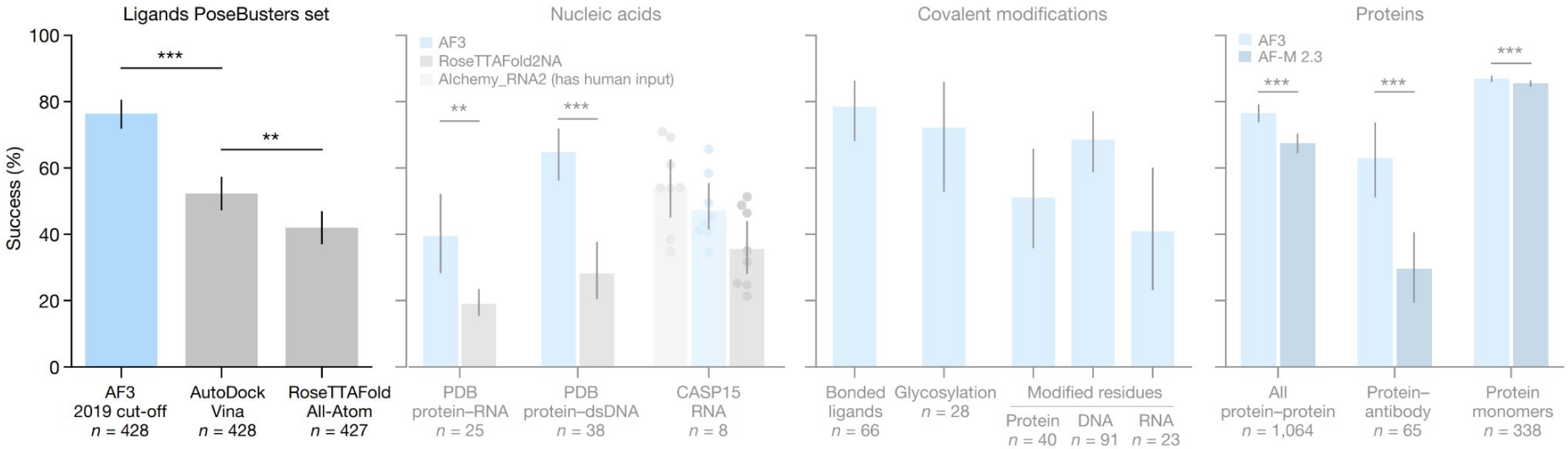


•All confidence scores (e.g. pLDDT, PAE) calculated for tokens rather than residues
•Structure size also calculated in tokens (up to ~7,000 tokens can be modelled with A100 80 GB)

Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024).

Accuracy across molecular types - Ligands

AF3 is light blue throughout



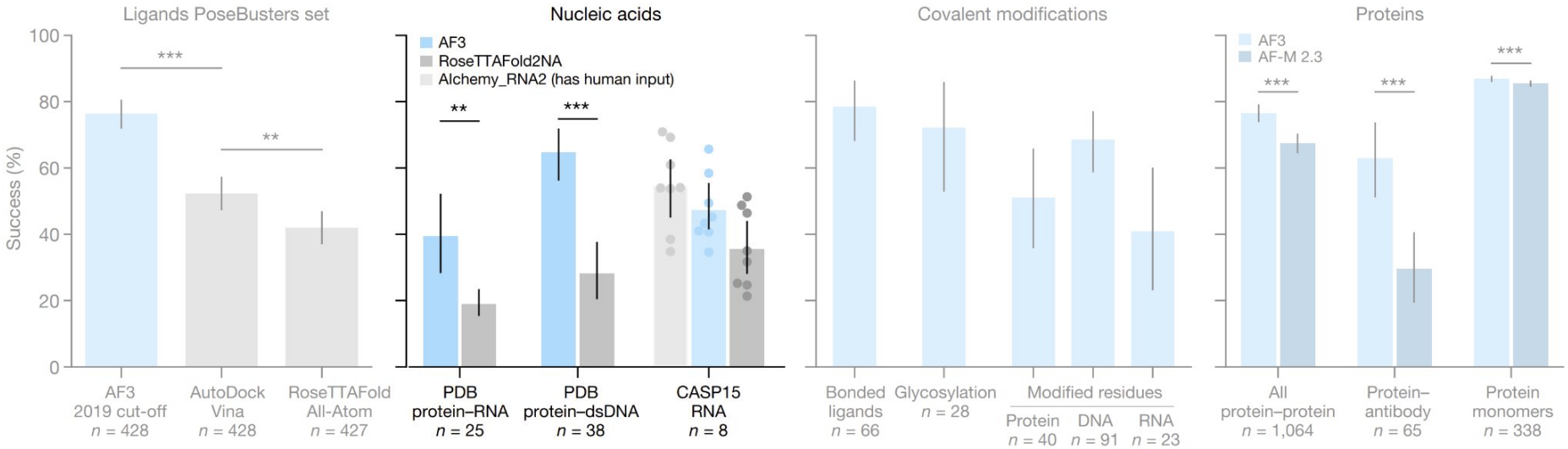
Classical docking tools and older ML models struggle on Posebusters benchmark for ligands

AF3 excels, with implications for drug discovery

Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024).

Accuracy across molecular types - Nucleic Acids

AF3 is light blue throughout



For protein-DNA and protein-RNA interactions, AF3 does well
For RNA alone human intervention is still better

Google DeepMind

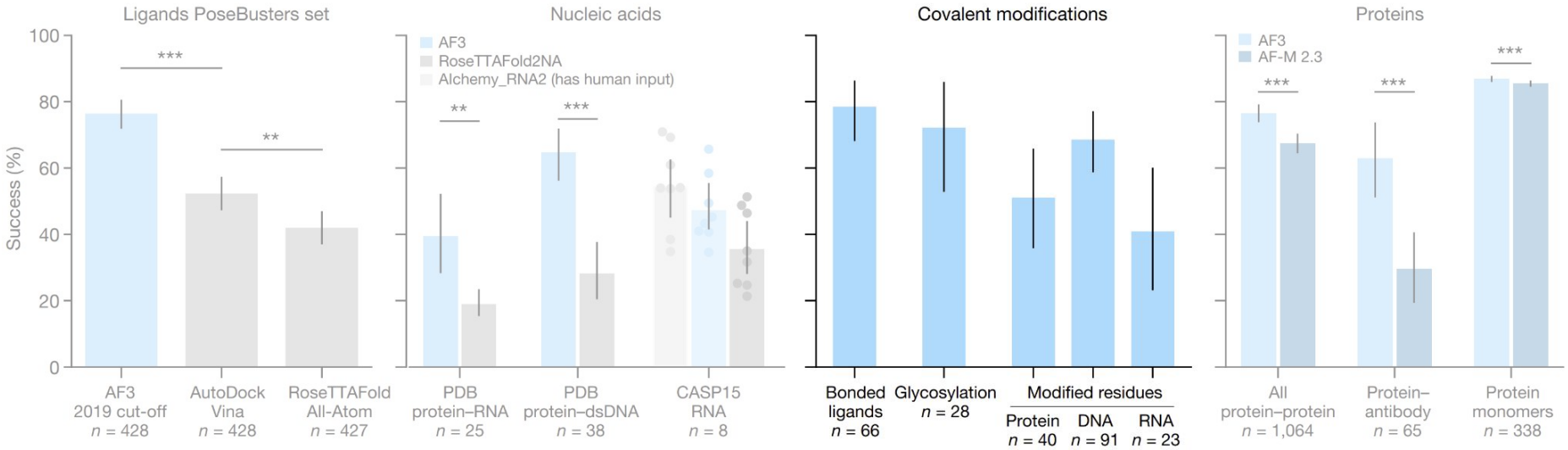
Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024).

PDBe Protein Data Bank in Europe EMBL-EBI

Accuracy across molecular types - Covalent modifications

AF3 is light blue throughout

AF3 is light blue throughout



No model comparison possible at time of writing, but accuracy is enough to be useful

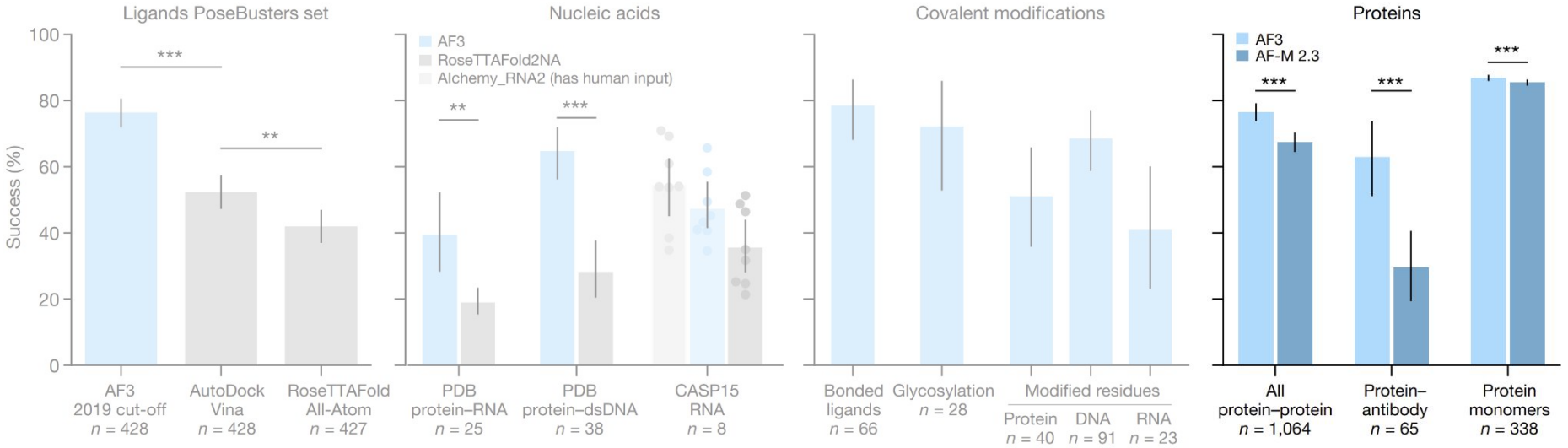
Google DeepMind

Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024).

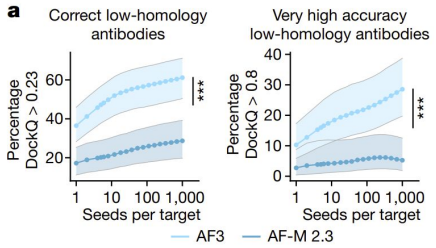
PDBe Protein Data Bank in Europe EMBL-EBI

Accuracy across molecular types - Proteins

AF3 is light blue throughout



AlphaFold 3 outperforms AlphaFold-Multimer 2.3 (the previous SotA) in all categories

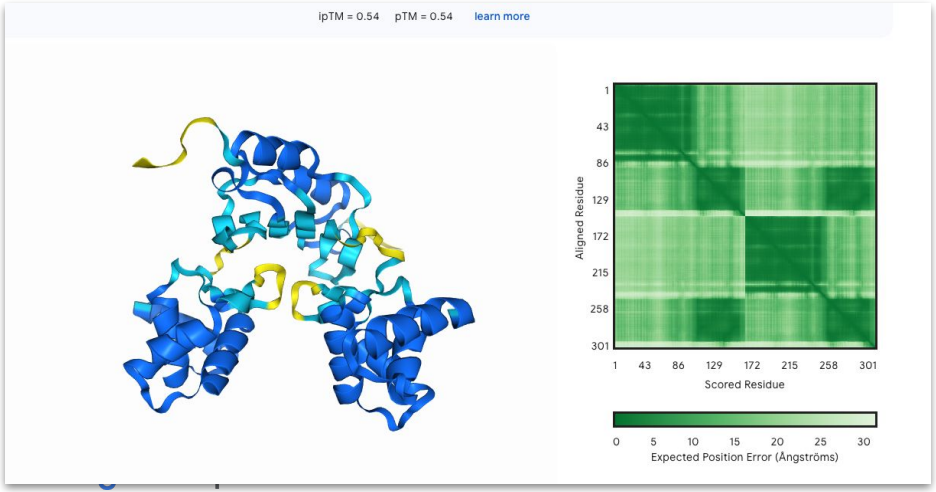


Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024).

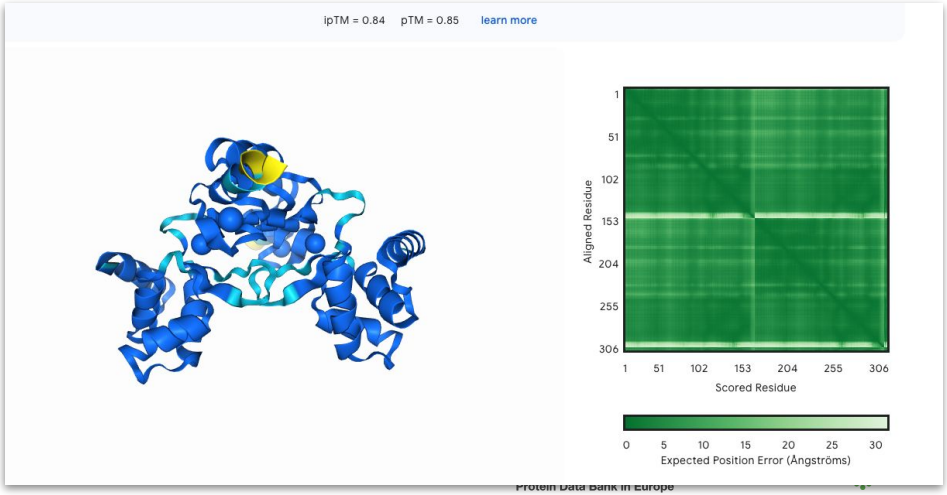
AlphaFold 3 looks for full structural context

- Confidence metrics of AlphaFold 3 (pLDDT, PAE, pTM, ipTM) may depend on having the full structural context:
  - For instance, if you model metal-binding protein **without metal ions**, it could be modelled with **lower confidence**, indicating that important parts of the structure are missing.
  - Adding metal ions will result in the confident prediction as all parts of the structure are in place.

no Zn, no Fe



plus Zn, plus Fe



How to run AlphaFold 3?

	AlphaFold Server	GitHub AlphaFold 3
Ease of use	User-friendly, designed for biologists	Designed for experienced power users
Functionality	Basic	Full
Speed	Fast	Depends on hardware used
Cost	Free	Free (but you need advanced hardware like latest GPU, SSD)
License ***(please check exact license terms)***	**Strictly for non-commercial use only;** also you can't use AlphaFold 3 output to train machine learning models or related technology for biomolecular structure prediction	

Specific limitations of AF3

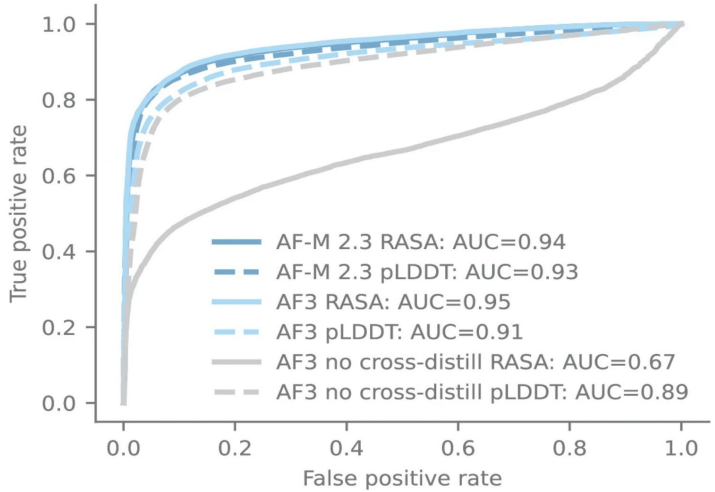
- There are general limitations such as:
- Not a model of protein dynamics
- Cannot model physics of folding or binding (e.g. it does not model Kd)
- Sometimes could be inaccurate - Usually this is communicated with poor confidence scores

Specific limitations of AF3

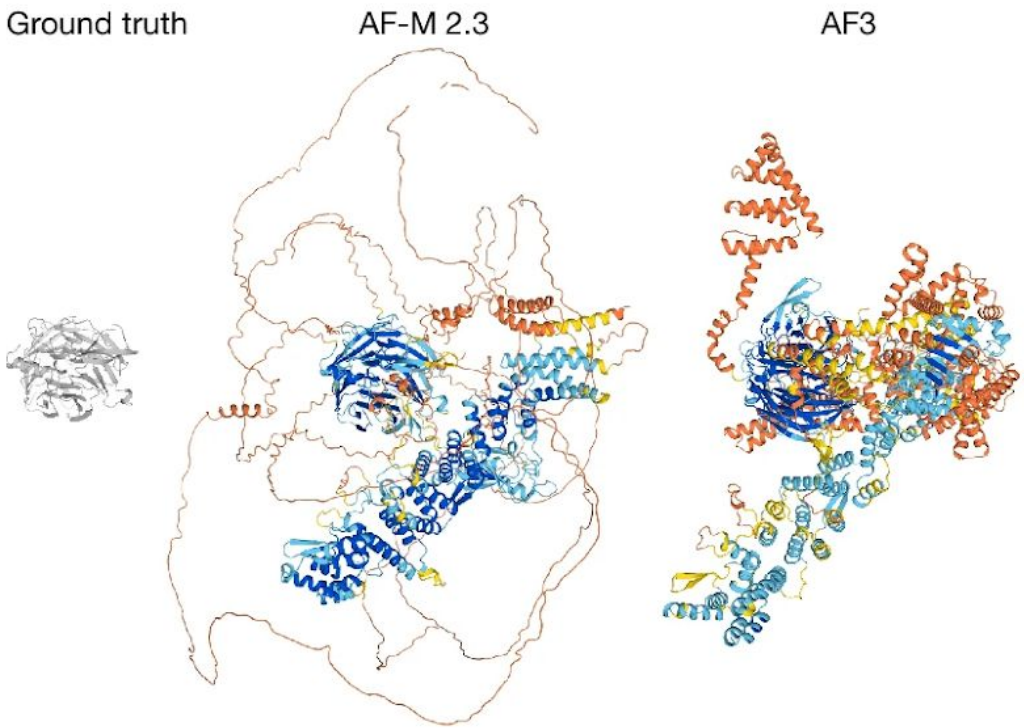
- Other tools now perform better for RNA – see recent CASP 16 results.
- Chirality errors - can be mitigated by adding a chiral error penalty to ranking score.
- Catastrophic clashing/overlapping regions can occur for structures with >2k tokens and large nucleic components.
- Covalent bonds for general ligands (i.e. not for glycans): leaving atoms not automatically removed and bond lengths exactly around the covalent bond can be wrong - can be mitigated by combining the base residue and bonded ligand into one large modified residue, or via relaxation.
- Not exactly a limitation: ligand predictions depend on bond order (i.e. atom A to B or atom B to A) due to the model not symmetrising them and potentially on exact atom name (which is arbitrary).

Specific limitations of AF3

- Hallucinations were not fully removed - the further from the training set, the more hallucinations. Hallucinations nearly always have very low pLDDT, but may still disrupt structure prediction.



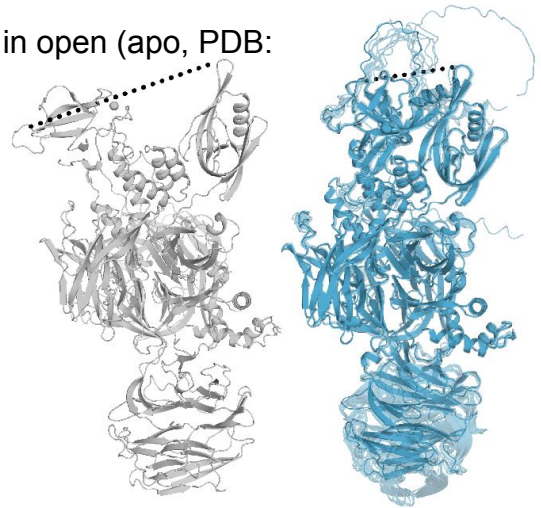
Specific limitations of AF3



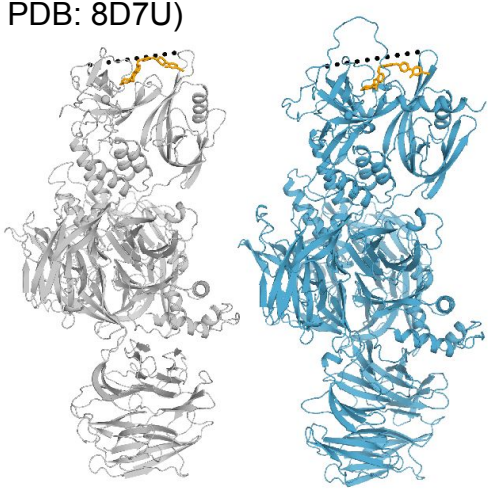
What AlphaFold 3 struggles with- Conformational diversity

E3 ubiquitin ligases natively adopt an open conformation in an apo (ligand-free) state and are only observed in a closed conformation when in a holo state (ligand-bound). Nevertheless, ***AlphaFold 3 exclusively predicts the closed conformation for both holo and apo states.***

Cereblon in open (apo, PDB: 8CVP)

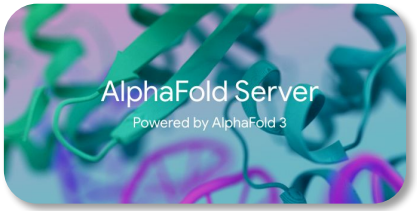


Closed (holo -bound, PDB: 8D7U)

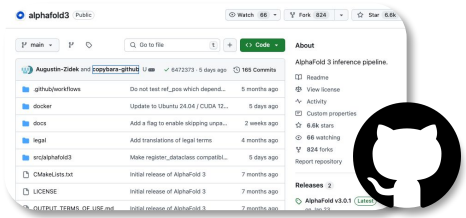


Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024).

Two ways of running AlphaFold 3



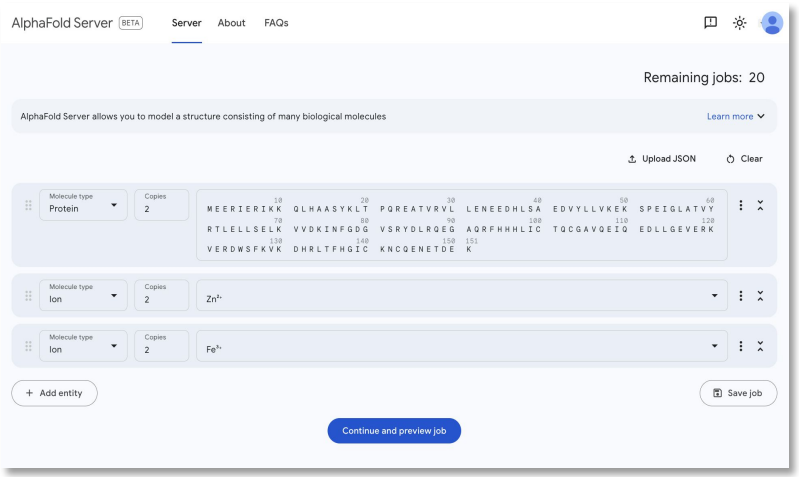
AlphaFold Server



GitHub AlphaFold 3

	AlphaFold Server	GitHub AlphaFold 3
Ease of use	User-friendly, designed for biologists	Designed for experienced power users
Functionality	Basic	Full
Speed	Fast	Depends on hardware used
Cost	Free	Free (but you need advanced hardware like latest GPU, SSD)
License *(please check exact license terms)*	**Strictly for non-commercial use only;** also you can't use AlphaFold 3 output to train machine learning models or related technology for biomolecular structure prediction	

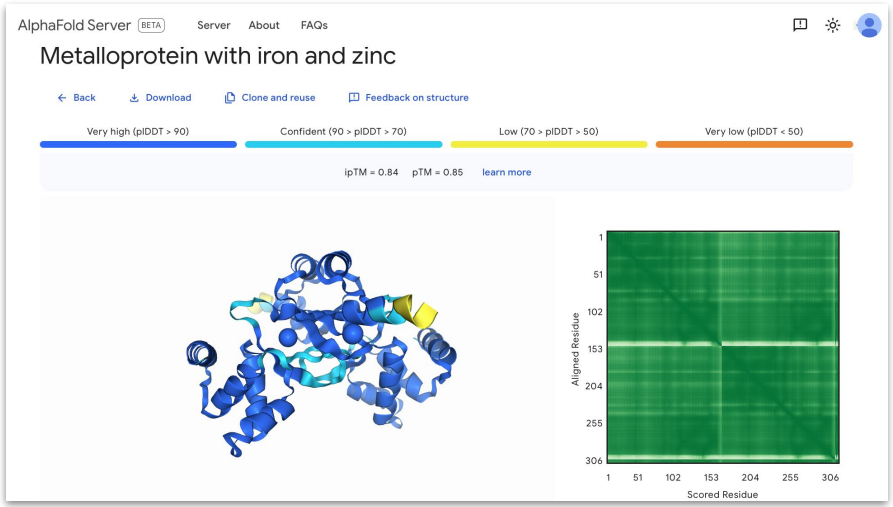
AlphaFold Server is made for biologists





AlphaFold Server provides a user-friendly and free platform for molecular and cell biologists, biochemists, and other life scientists to leverage AlphaFold 3 in their research

https://alphafoldserver.com/



Google DeepMind

PDBe Protein Data Bank in Europe EMBL-EBI

What AlphaFold Server can model?

https://alphafoldserver.com/



Any proteins*

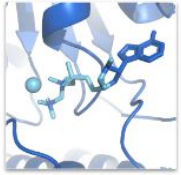
*except a few viral ones



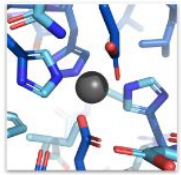
Any DNA



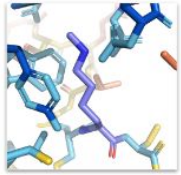
Any RNA



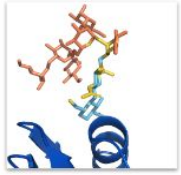
Common ligands ranging from ATP to chlorophyll



Common metal ions and chlorine ions



Common PTMs ranging from phosphorylation to citrullination



Protein glycosylation, including branched glycan chains

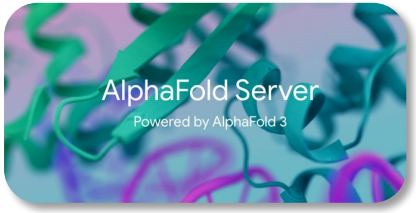


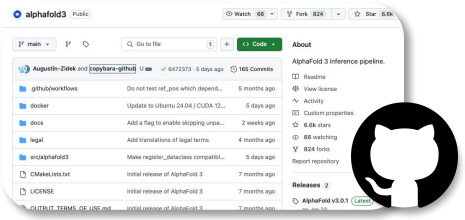
Common chemical modifications of nucleotides, such as methylation and formylation

Google DeepMind

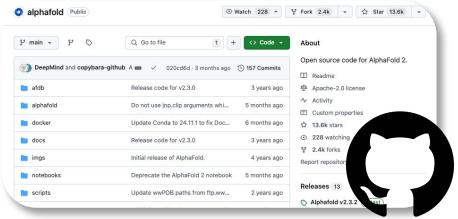
PDBe Protein Data Bank in Europe EMBL-EBI

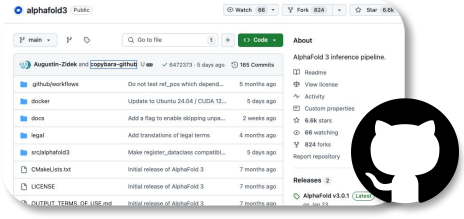
Features

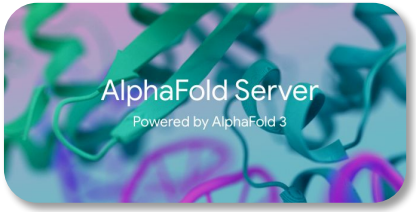




	AlphaFold Server	GitHub AlphaFold 3
Protein, RNA, and DNA chains	✔	✔
Custom multiple sequence alignment (MSA) for protein and RNA chains	Under development	✔
Custom template structures for protein chains	Under development	✔
Custom ligands (via CCD codes, SMILES and CIFs)	Selected ligands only	✔
Covalent modifications (PTMs, glycosylation, modified nucleotides)	Selected modifications only	✔
Custom covalent bonds between entities	❌	✔
Multiple random seeds	via multiple jobs	✔





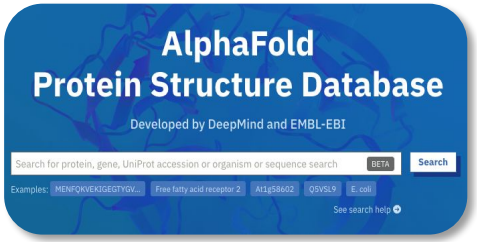


Feature	**AlphaFold 2**	**AlphaFold 3 GitHub**	**AlphaFold 3 Server**
Biomolecule types	Proteins and protein-protein complexes	Proteins, nucleic acids (DNA/RNA), ligands, ions, chemically modified residues and complexes between these molecule types.	Proteins and nucleic acids, some biologically common ligands, ions, chemical modifications and complexes.
Commercial use	Open-source (Apache 2.0 license); freely available for academic and commercial use.	Non-commercial use only (subject to AlphaFold Model Parameters terms of use).	Non-commercial use only (subject to AlphaFold Server Terms of Service).
Integration	Can be integrated into custom pipelines for large-scale screenings.	Can be integrated into custom pipelines for large-scale screenings	Web-based interface; no local integration.
Best use cases	- Large-scale protein-protein interaction screenings. - Commercial projects.	- Multi-molecular complexes. - Non-commercial research.	- Fast predictions for proteins, nucleic acids and common ligands, ions and modifications. - Non-commercial research.

Google DeepMind

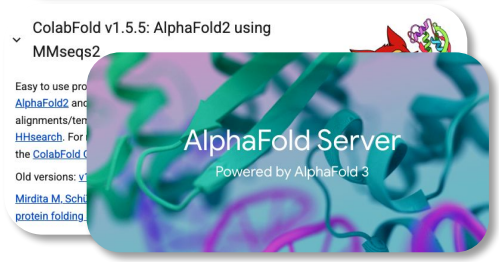
PDBe Protein Data Bank in Europe EMBL-EBI

Where to start? A Typical workflow



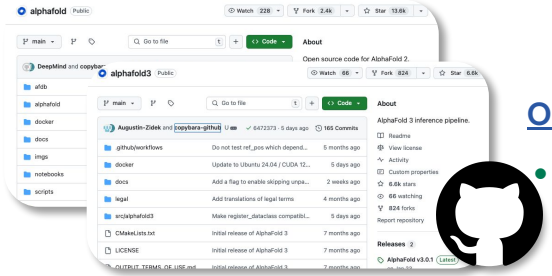
AlphaFold Database

- Pre-calculated models and confidence metrics.
- >214 million protein structures and 47 proteomes of key model organisms.



AlphaFold Server (AF3) or ColabFold (AF2)

- Starting point for predicting your protein or complex.



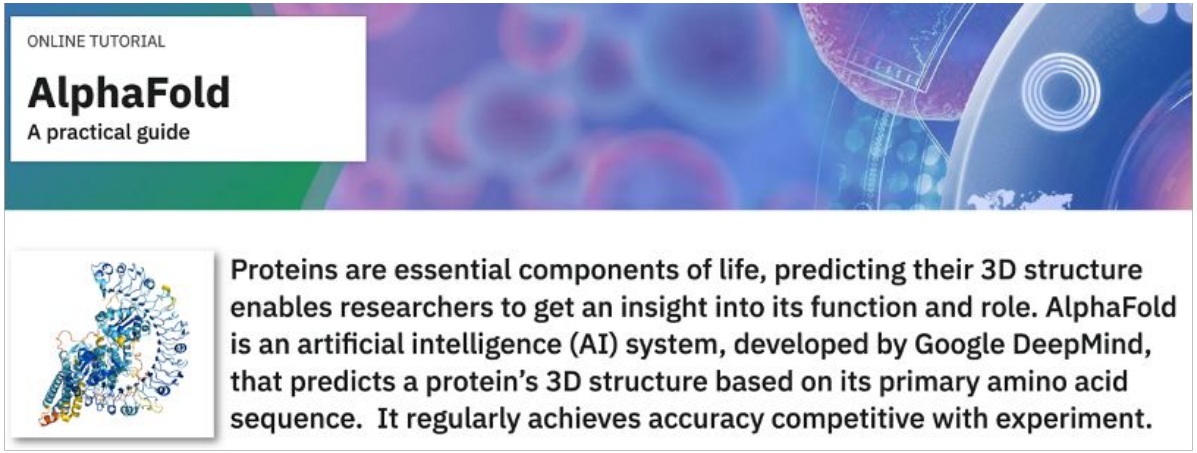
Open source code (AF2 & AF3)

- Total control over predictions.

Google DeepMind

PDBe Protein Data Bank in Europe EMBL-EBI

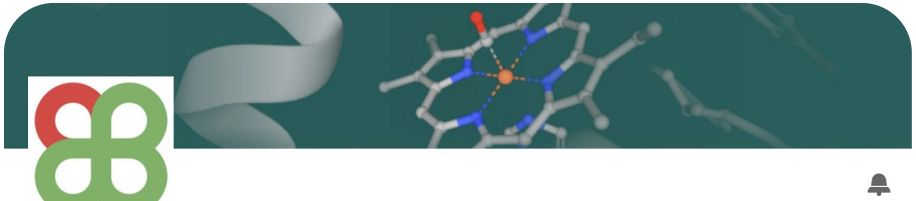
How can I learn more about AlphaFold?



Google DeepMind

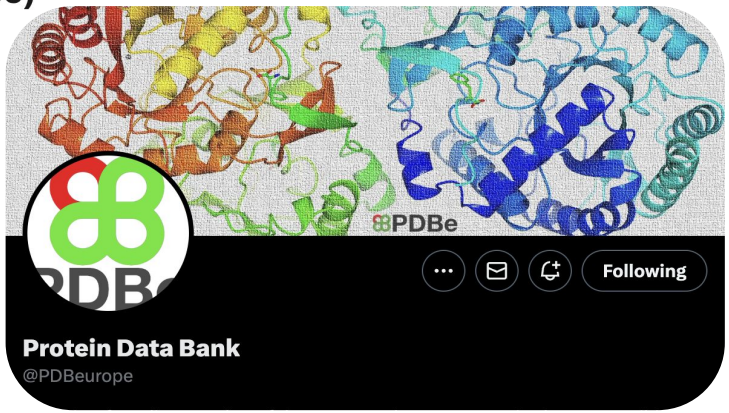
PDBe Protein Data Bank in Europe EMBL-EBI

How can I get updates?



Protein Data Bank in Europe (PDBe)

wwPDB | PDBe-KB | AlphaFold DB | 3D-Beacons



Google DeepMind

PDBe Protein Data Bank in Europe EMBL-EBI

Thank you to everyone who made AlphaFold possible!

We are very grateful to all the **authors** of the above mentioned research, **community** of AlphaFold users and **developers** building upon AlphaFold

AlphaFold 2

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger (Seoul NU), Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, Demis Hassabis

Database

Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon,
Sameer Velankar (EBI), Gerard J. Kleywegt (EBI), Alex Bateman (EBI), Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon Kohl, Anna Potapenko, Andrew J. Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Petersen, Andrew W. Senior, Koray Kavukcuoglu, Ewan Birney (EBI), Pushmeet Kohli, John Jumper, Demis Hassabis

AlphaFold 3

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew Ballard, Joshua Bambrick, Sebastian Bodenstein, David Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian Fuchs, Hannah Gladman, Rishub Jain, Yousuf Khan, Caroline Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, John Jumper

AlphaFold Server

Dhavanthi Hariharan, Meera Last, Hamish Tomlinson, Josh Abramson, Charles Beattie, Molly Beck, Richard Green, Tim Green, Kristian Holsheimer, Oleg Kovalevskiy, Dariusz Lasecki, Alex Morris, Antonia Paterson, George Thomas, Gregory Thornton, Augustin Žídek, Aliyah Bond, Alex Bridgland, Jana Čavojská, Alexander Cowen-Rivers, Andrew Cowie, Jaume Sanchez Elias, Jacob Kelly, Eugenia Kim, Evan Palmer, Kuba Perlin, Ollie Purkiss, Chongyang Shi, Akvilė Žemgulytė, Ben Coppin, John Jumper, Anna Koivuniemi, Agata Laydon, Sam Miller, Stig Petersen, and many more people at Google DeepMind

The wider team at Google DeepMind

DeepMind

Collaborators at EMBL-EBI

EMBL

The CASP community

CASP14

PDB & the biology community

YEARS OF Protein Data Bank