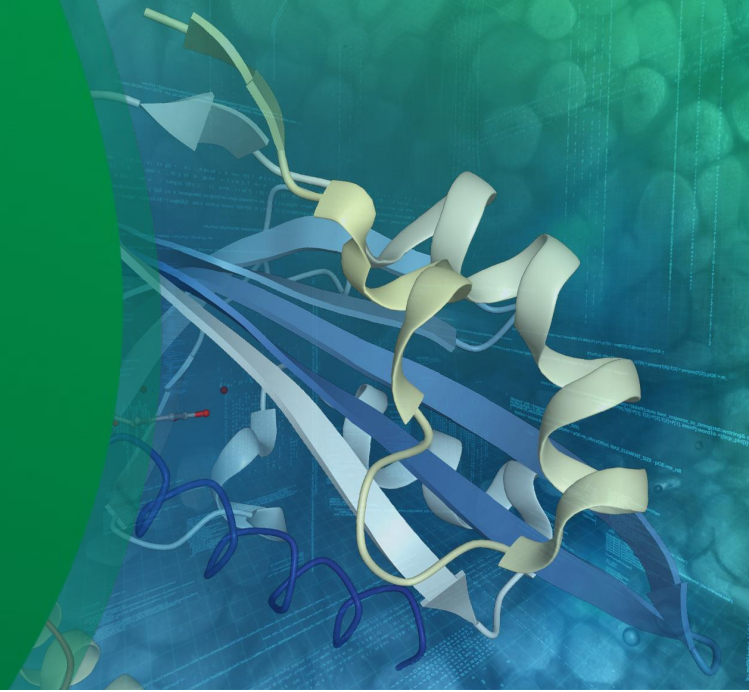
AlphaFold DB and 3D-Beacons

CCP-EM | Oxfordshire



Jennifer Fleming

PDBe coordinator

jfleming@ebi.ac.uk

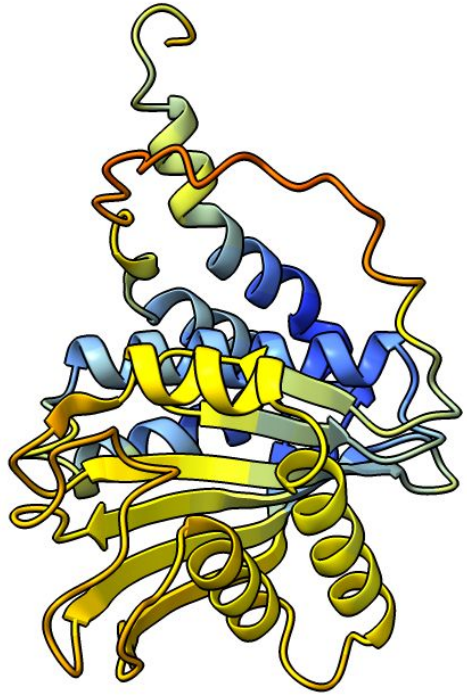
8th March, 2024

EMBL-EBI

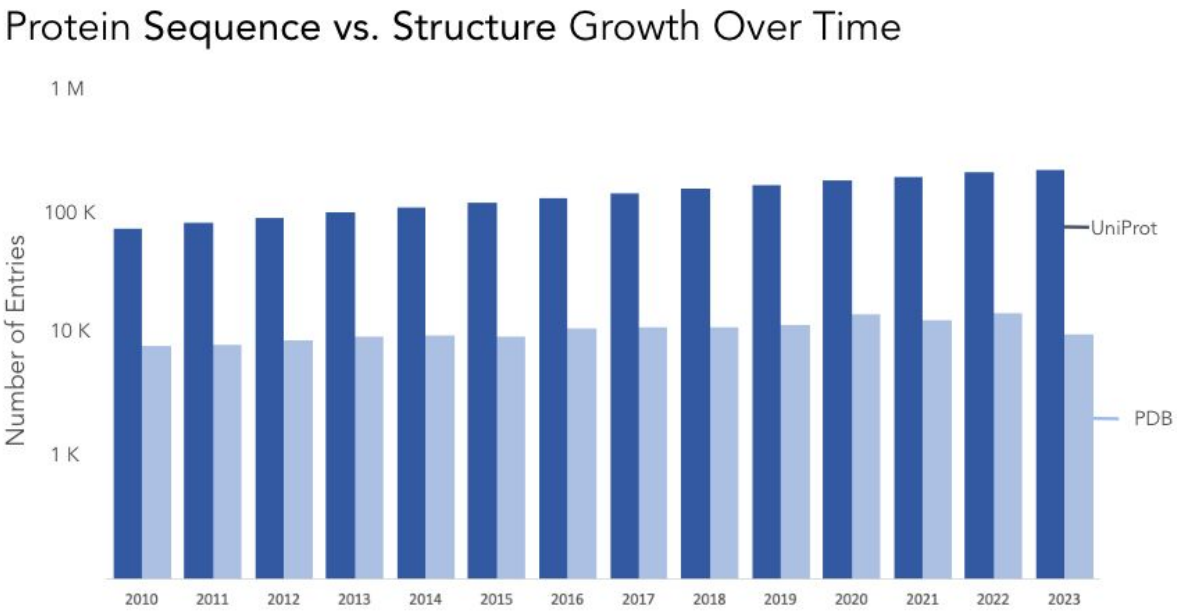
The structure prediction problem

- Solving structures experimentally takes months / years.
- But in principle, **sequence determines structure (Anfinsen's dogma),** so it could be predicted/computed.
- **Levinthal's paradox – too many possible conformations**, sequential search would surpass the time span of the Universe.

MHIEQRLEFLDTLPALPASSFTPRVAAFSSLIDPPVDPNKESGTVGPEMMNIYMPGVSVQNRKDVDNCKLLVQNAAT
QLYDPIKQLHEWYKYYTESLAKLGWVTQASQVKDITIRKTGLSMDAVAFEILQGLVGANAPQLLALAGKAVDGVKNN
EGLINIYNRNAKIGYEAKFDMSPVWQTREGSPMMILNCTSVDVRESTRGILWWKSTSQSTAVKSAASAVYLNVDTYD
QVRAAVLKKLGQAAEDFLDSIPGFQ



The gap between sequences and structures



Numbers are accurate as of 21st Sep 2023
Data Source: PDBe 2023

The UniProt database contains 252,170,925* protein sequences

The Protein Data Bank contains >294K* entries, corresponding to 62,790* unique proteins

This gap grew 10x within the last decade

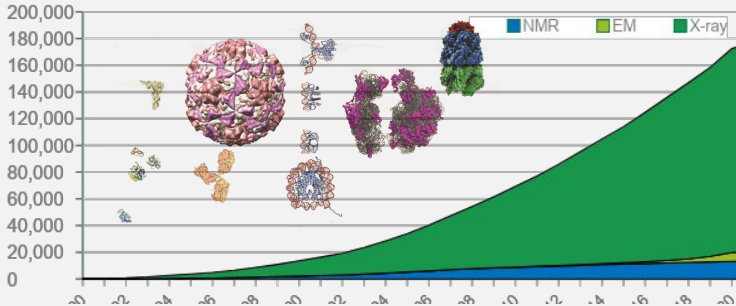
Predicted structures could help close the gap

* Numbers are accurate as of 4th Nov 2024

EMBL-EBI

Two elements that enabled the advances in structure prediction

- Open data
  - Protein structure data (PDB): Diversity of structures present in PDB.
  - Genome sequencing.
  - Annotated protein sequence/proteome data in UniProt ~240 million sequences.

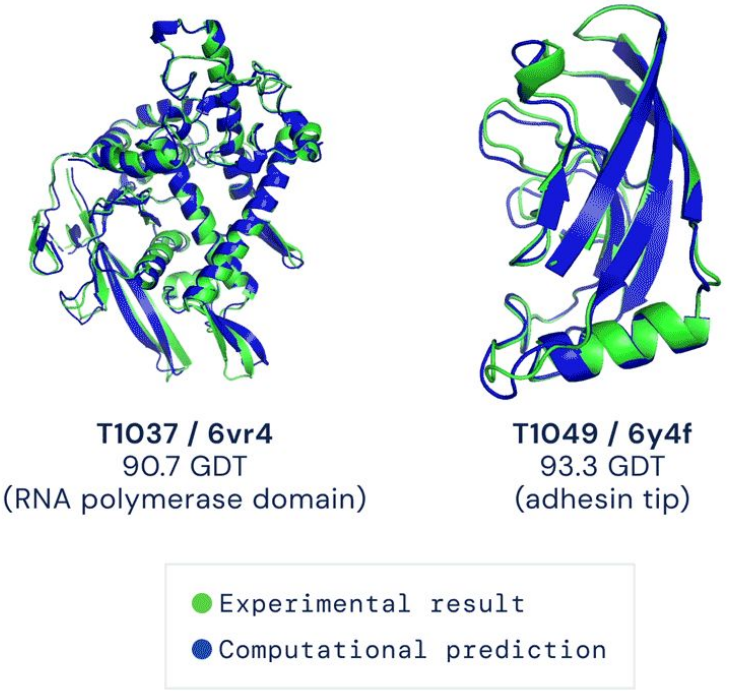


- Framework for assessing progress
  - John Moult established a meeting CASP to be held every 2 years, to assess progress: CASP1 was in 1994.
  - Many similar community assessment in life sciences – e.g. CAPRI, CAMEO, D3R, CAFA …..

Structure prediction breakthrough - AlphaFold

- In 2020, AlphaFold achieved an unprecedented accuracy score in CASP14
- **Disclaimer:**
  It is not a *real* solution to the folding problem,
  but **it can generate excellent predictions for protein fold based on sequence**

- AlphaFold was developed by Google DeepMind
  - It can create highly accurate theoretical models of protein structures using protein sequences
  - **AlphaFold was trained on the wwPDB archive (version downloaded 2018-04-30).**



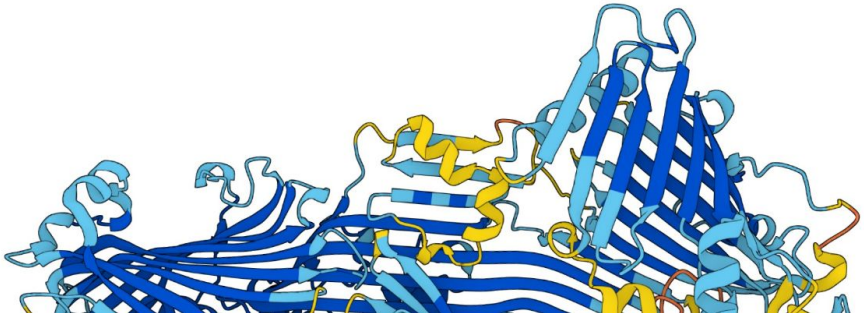
[1] Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021)
[2] Jumper, John et al. “Applying and improving AlphaFold at CASP14.” Proteins (2021)

AlphaFold Protein Structure Database

Developed by EMBL-EBI and Google DeepMind

PDBe

Protein Data Bank in Europe

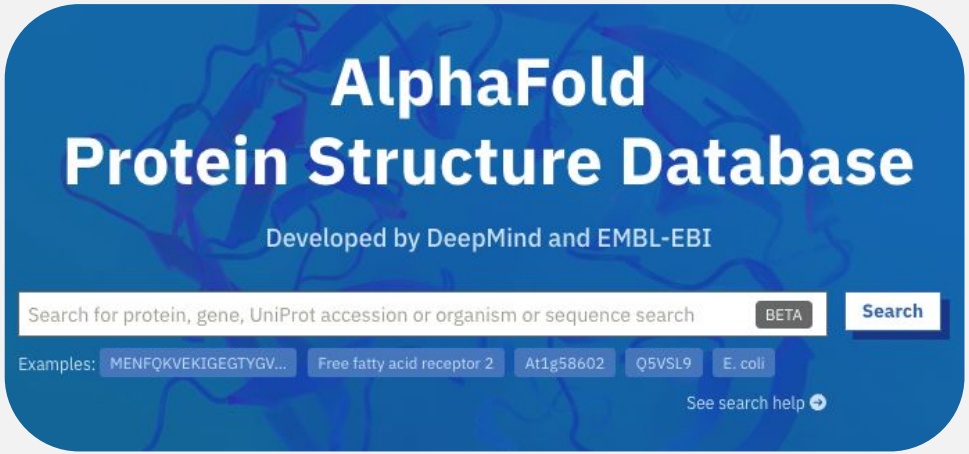


Google DeepMind

EMBL-EBI

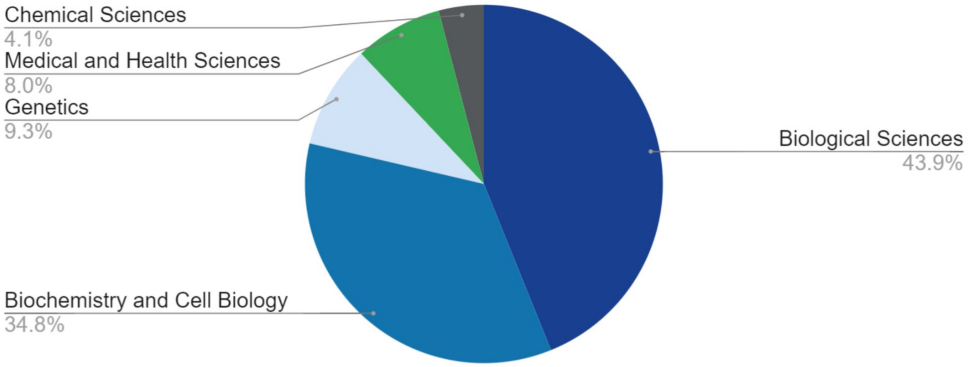
The AlphaFold Database

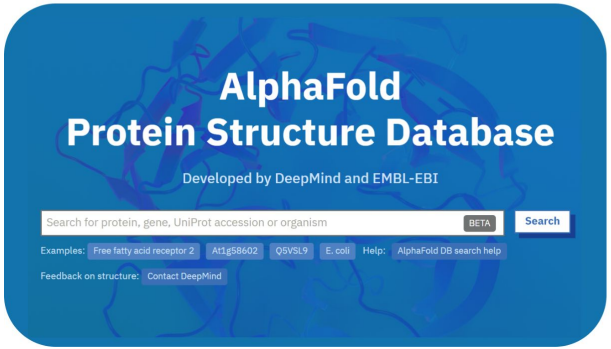
- The **AlphaFold database** at https://www.alphafold.ebi.ac.uk/
  - Holds pre-calculated models and confidence metrics
  - >214 million protein structures and 47 proteomes of key model organisms
  - Models predicted using **AlphaFold2**



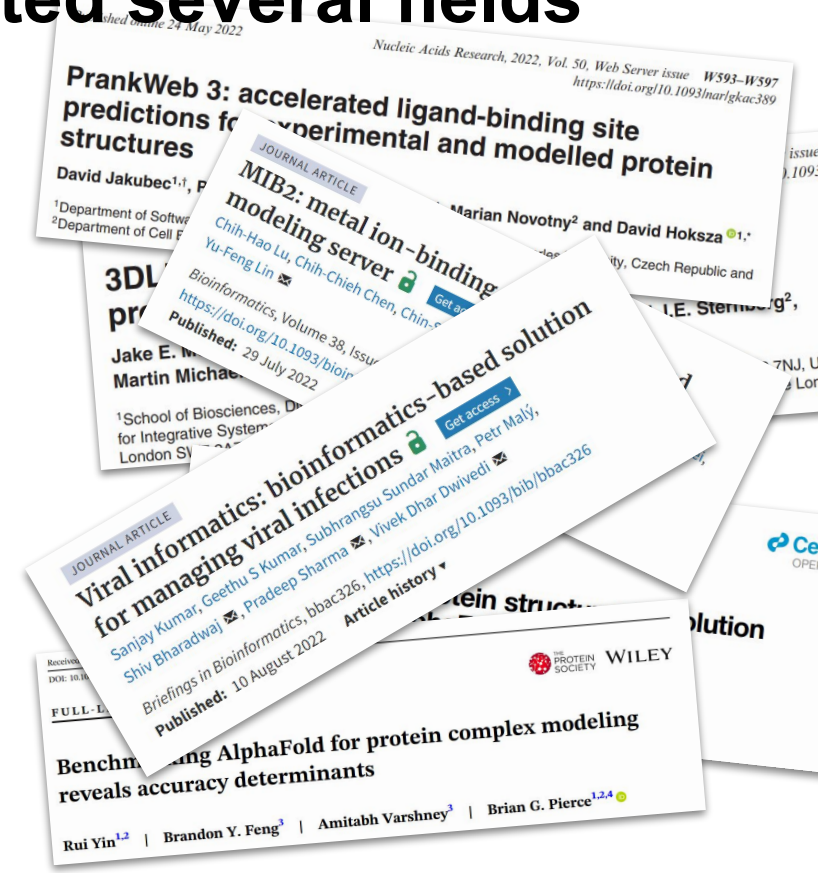
EMBL-EBI

Highly accurate models impacted several fields





- STRUCTURE DETERMINATION
- TEACHING AND TRAINING
- DATA PROVIDERS
- BIOINFORMATICS ANALYSES
- DRUG DISCOVERY

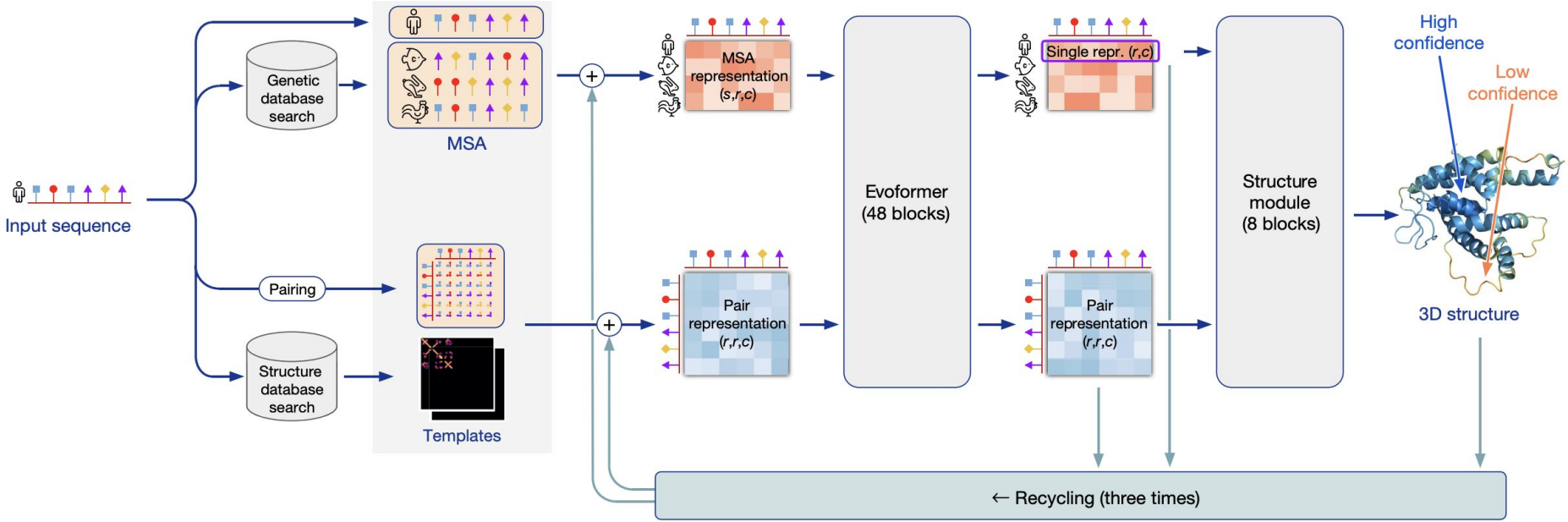


Google DeepMind

1,912 different publications

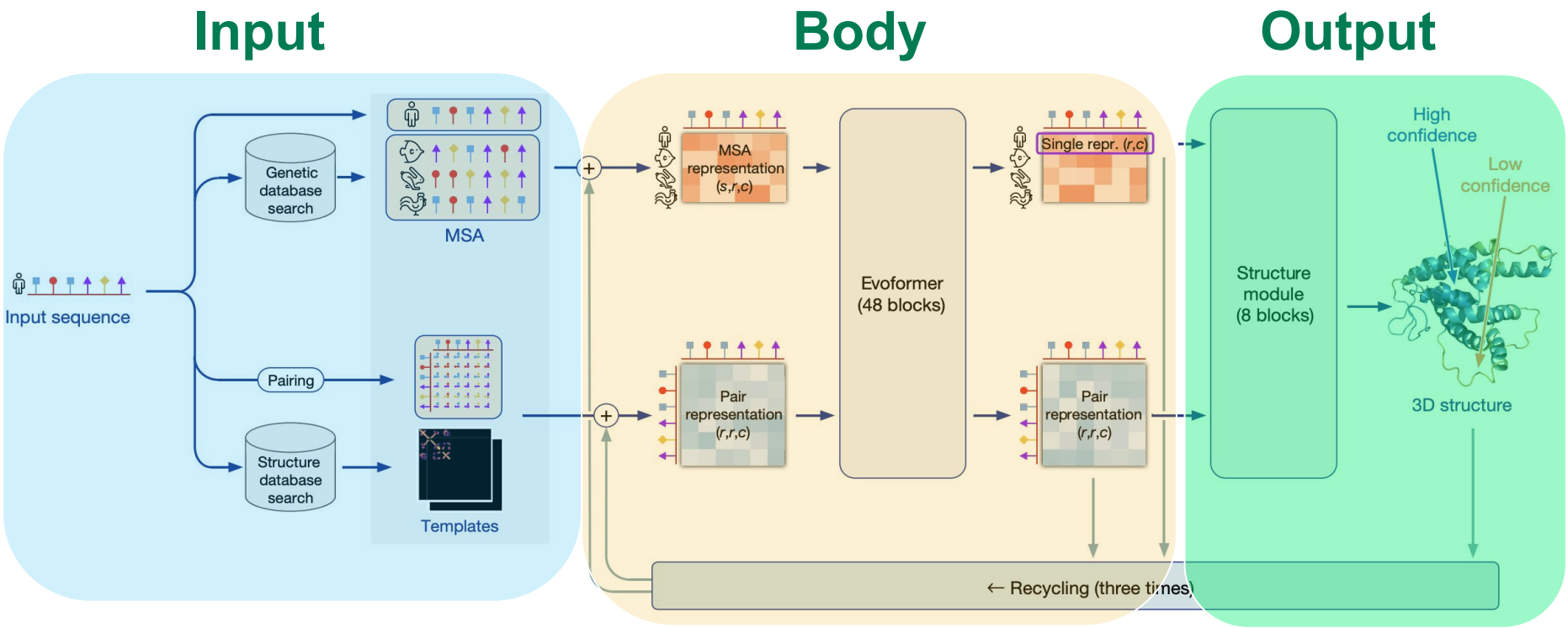
EMBL-EBI

What is AlphaFold? High-level overview



Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021). https://doi.org/10.1038/s41586-021-03819-2

How does AlphaFold work?

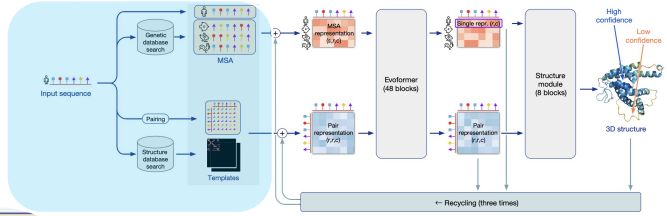


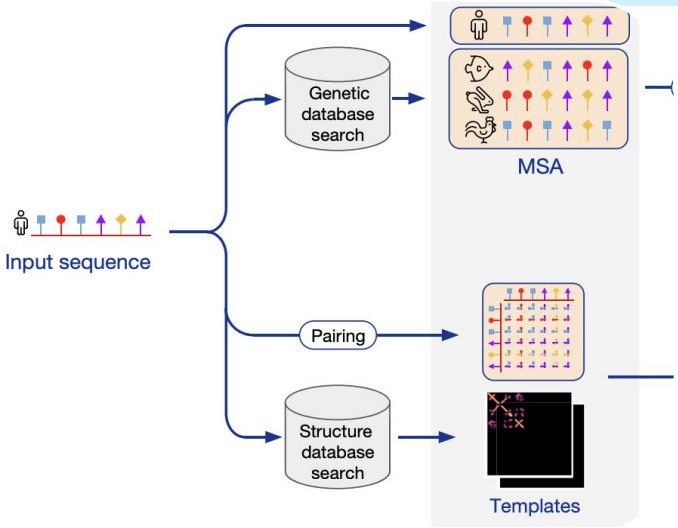
Google DeepMind

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

EMBL-EBI

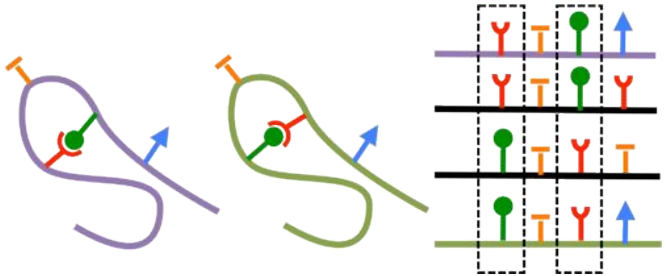
How does AlphaFold work?
Importance of the Multiple Sequence Alignment (MSA)





An MSA identifies similar, but not identical, sequences that have been identified in living organisms.

Pairs of residues that “change together” across evolution often denotes contact.



Google DeepMind

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

EMBL-EBI

Strengths of AlphaFold

- AlphaFold can predict protein structures **based only on amino acid sequences.**
  - It is a very good starting point for creating testable hypotheses.
  - AlphaFold models are routinely used as starting models for X-ray crystallography and cryo-EM.
  - ~850 entries associated with a paper that uses AlphaFold (>60% were cryo-EM structures) Jan 2023

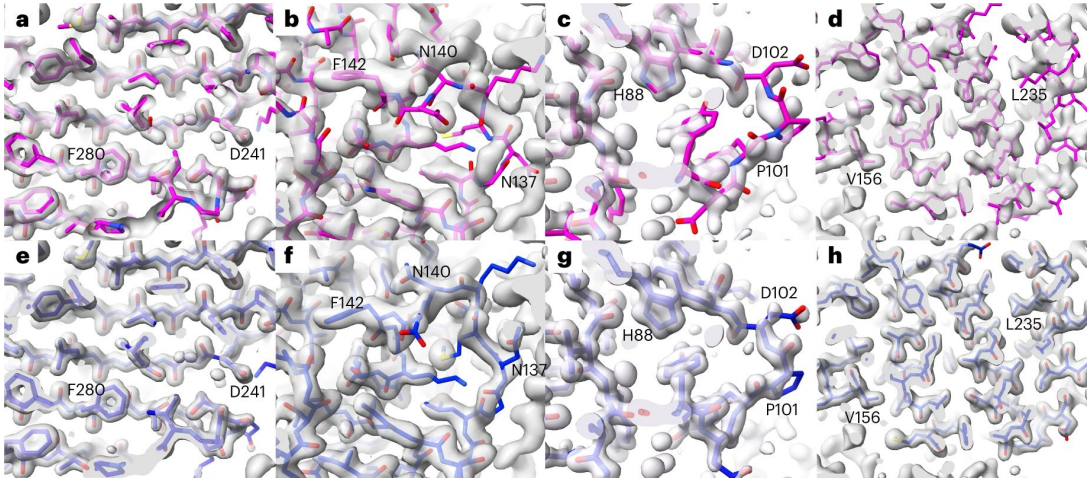
- If available, **AlphaFold can use templates** from the PDB.
  - But even if templates are available, AlphaFold may discard them and use only sequence data.

- AlphaFold produces **three independent outputs**:
  - Predicted atomic coordinates (PDB and mmCIF).
  - Confidence measure per amino acids (pLDDT).
  - Predicted Aligned Error (PAE).

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.



Google DeepMind

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

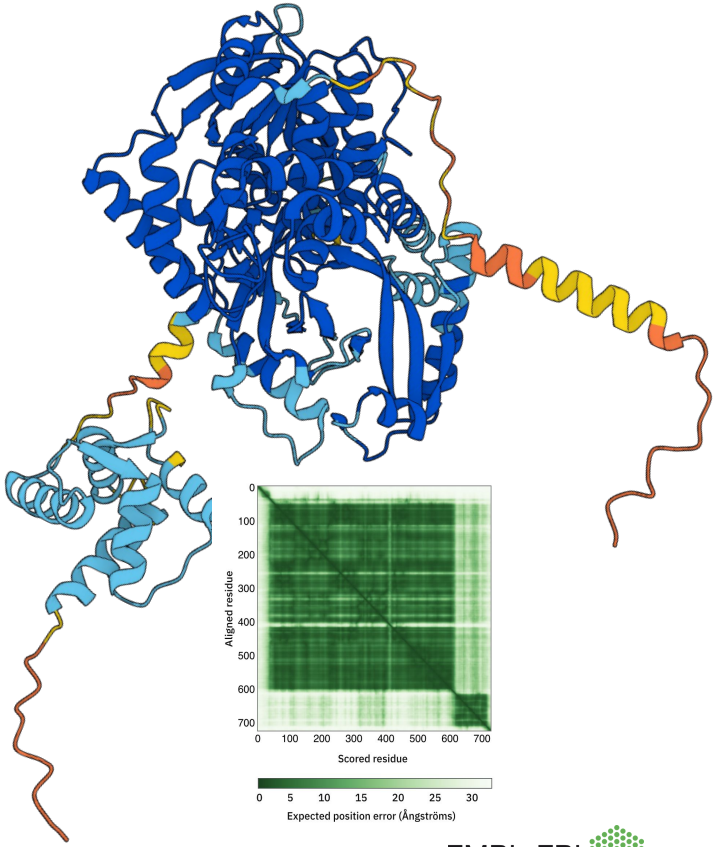
EMBL-EBI

What types of data does the AFDB store?

- For the 214,684,311 proteins:
  - Predicted atomic coordinates
  - Confidence metrics (pLDDT and PAE)
  - Metadata

What do we not (yet) have in the database?

  - There are no multiple conformations of protein structures.
  - There are no viral proteins.
  - There are no isoforms.
  - There are no assemblies.
  - There are no mutant structures.
  - There are no ligands
  - Only AlphaFold2 models



Glycerol-3-phosphate dehydrogenase, mitochondrial, P43304
https://alphafold.ebi.ac.uk/entry/P43304

Google DeepMind

EMBL-EBI

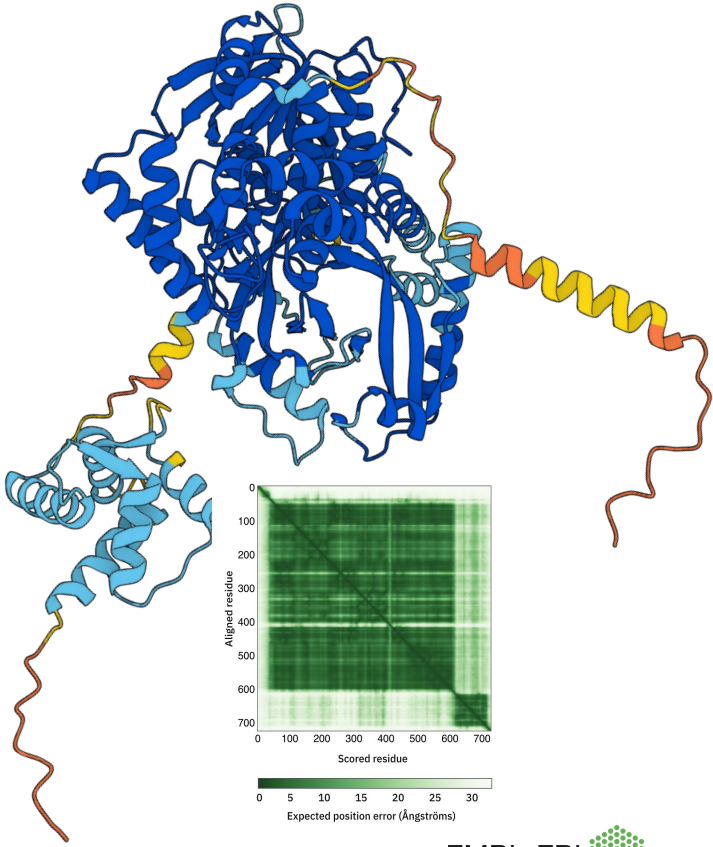
What types of data does the AFDB store?

- For the 214,684,311 proteins:
  - Predicted atomic coordinates
  - Confidence metrics (pLDDT and PAE)
  - Metadata

What do we not (yet) have in the database?

  - There are no multiple conformations of protein structures.
  - There are no viral proteins.
  - There are no isoforms.
  - There are no assemblies.
  - There are no mutant structures.
  - There are no ligands
  - Only AlphaFold2 models

See 3D-Beacons



Glycerol-3-phosphate dehydrogenase, mitochondrial, P43304
https://alphafold.ebi.ac.uk/entry/P43304

Google DeepMind

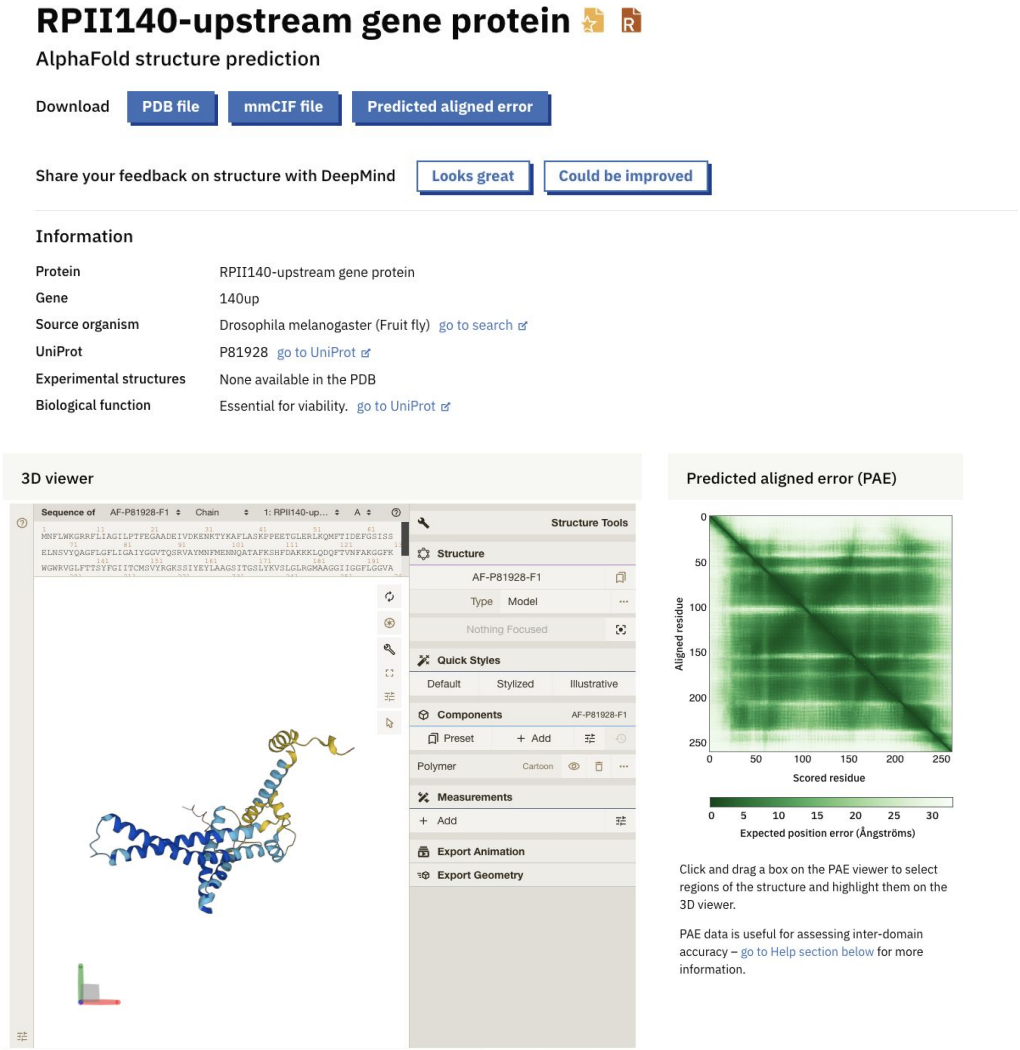
EMBL-EBI

Prediction pages

Protein predictions have dedicated pages.

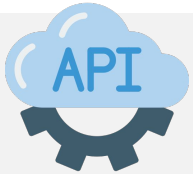
These pages provide download options and 2D/3D data visualisations for structures and confidence metrics.

We are expanding these pages to include protein similarity data and functional annotations.

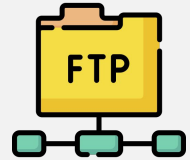


How can you access the database?

- **API** documentation: https://alphafold.ebi.ac.uk/api-docs, notebook
  - Provides access to metadata about all the archived AlphaFold predictions, as well as the URLs to model files (mmCIF, bCIF, and PDB) and models quality estimates.



- **FTP** site: https://ftp.ebi.ac.uk/pub/databases/alphafold/
  - Good option for users who need to download large amounts of data.



- Using **BigQuery,** notebook
  - Need to set up a BigQuery Sandbox. - Free tier allows up to querying 1TB
  - Good option for users who need to perform complex queries on the data.

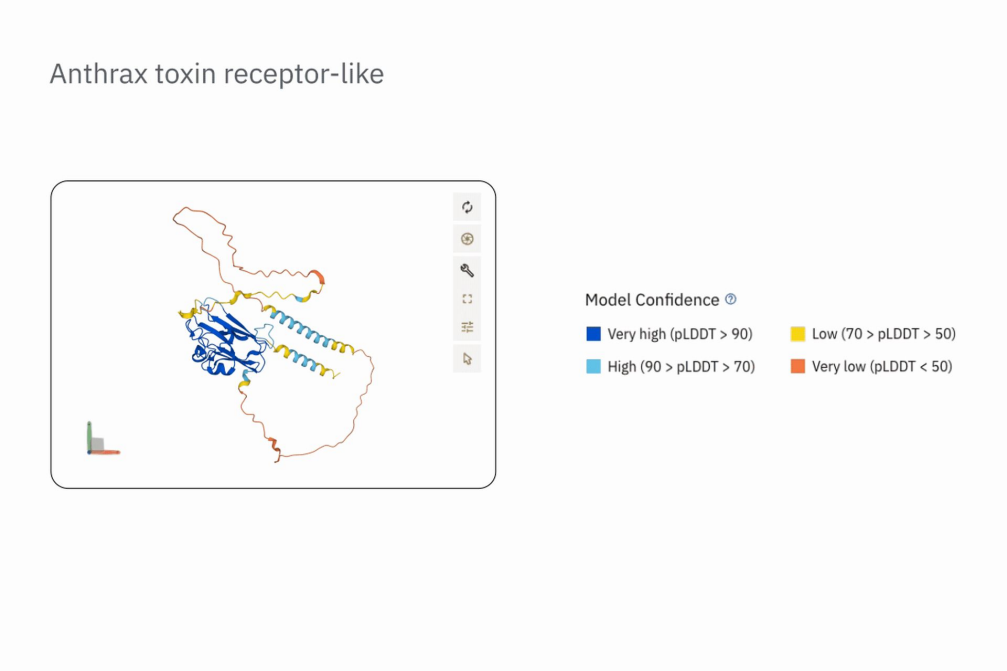
“Nothing in **biology** makes sense except in the light of **evolution** ”

Theodosius Dobzhansky, 1973

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

- Two major types of uncertainties:
  - of the position of an **amino acid** relative to its close neighbours (residue pLDDT score).
  - of the orientation of a **larger part** of the structure, e.g. a domain, relative to other domains (PAE plot).

pLDDT confidence metric



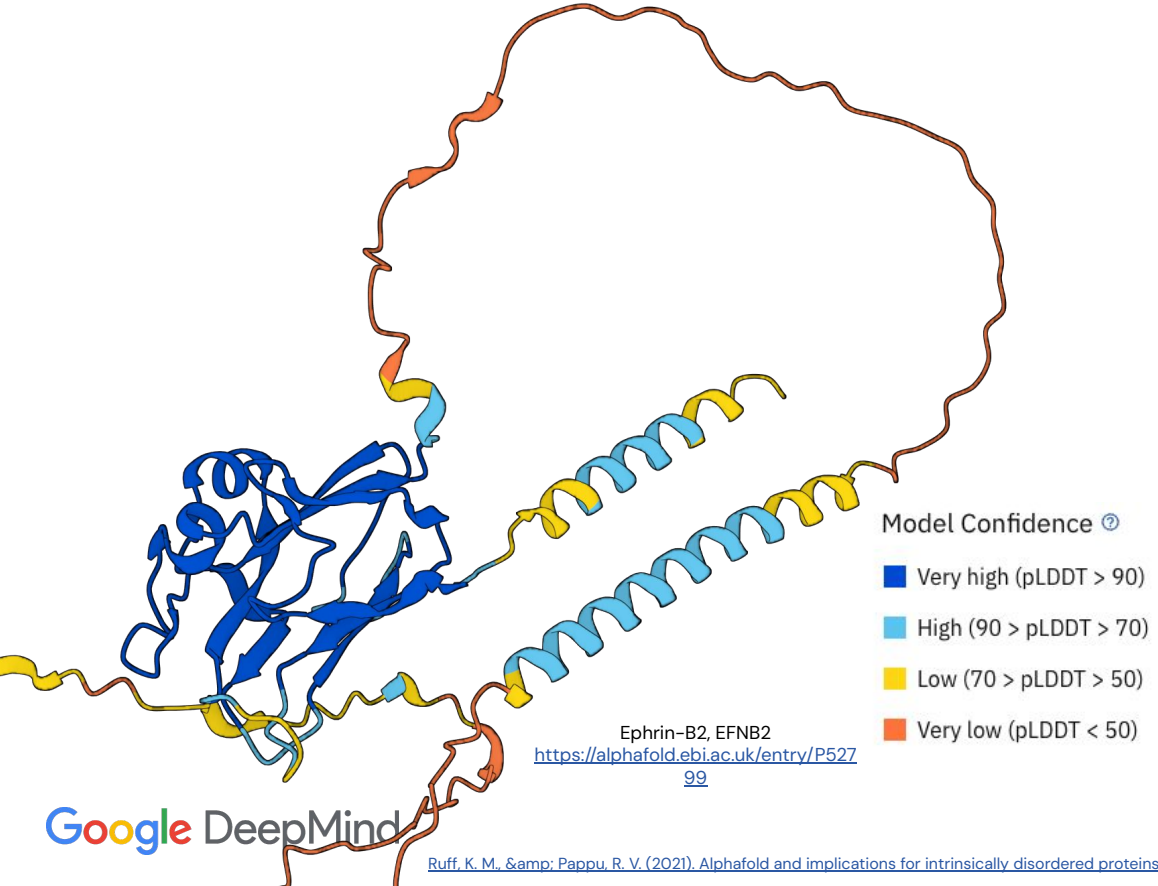
- Amino acid-level confidence measure called pLDDT.
- It is a **local accuracy metric. It** rewards **locally correct** structures and **getting individual domains right.**
- We **colour-code** residues based on their pLDDT scores.
- The range of pLDDT scores is between 0 and 100.

Google DeepMind

EMBL-EBI

Ruff, K. M., & Pappu, R. V. (2021). Alphafold and implications for intrinsically disordered proteins. Journal of Molecular Biology, 433(20), 167208. https://doi.org/10.1016/j.jmb.2021.167208

pLDDT confidence metric



Ephrin-B2, EFNB2
https://alphafold.ebi.ac.uk/entry/P52799

- pLDDT scores <50 are a reasonably strong predictor of disorder.
- It could encompass both regions that are intrinsically disordered and regions that are structured only in complex.

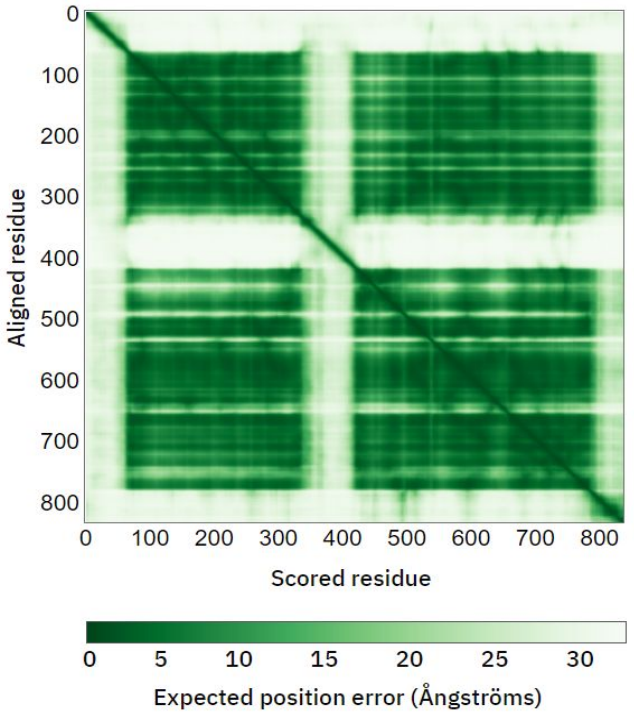
Google DeepMind

EMBL-EBI

Ruff, K. M., & Pappu, R. V. (2021). Alphafold and implications for intrinsically disordered proteins. Journal of Molecular Biology, 433(20), 167208. https://doi.org/10.1016/j.jmb.2021.167208

Predicted Aligned Error (PAE)

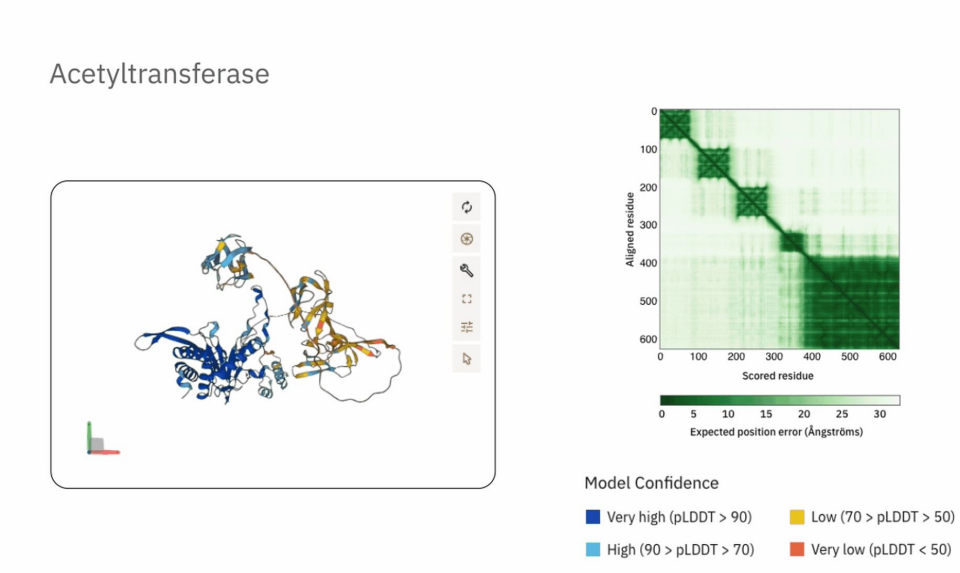
- PAE is pairwise, i.e. it has a value for every residue pair.
- The colour at (x, y) indicates AlphaFold's expected position error at residue x if the predicted and true structures were aligned on residue y.
- It measures the confidence in the relative position of two amino acids in Ångströms.



Google DeepMind

EMBL-EBI

Interpreting the relative positions of domains



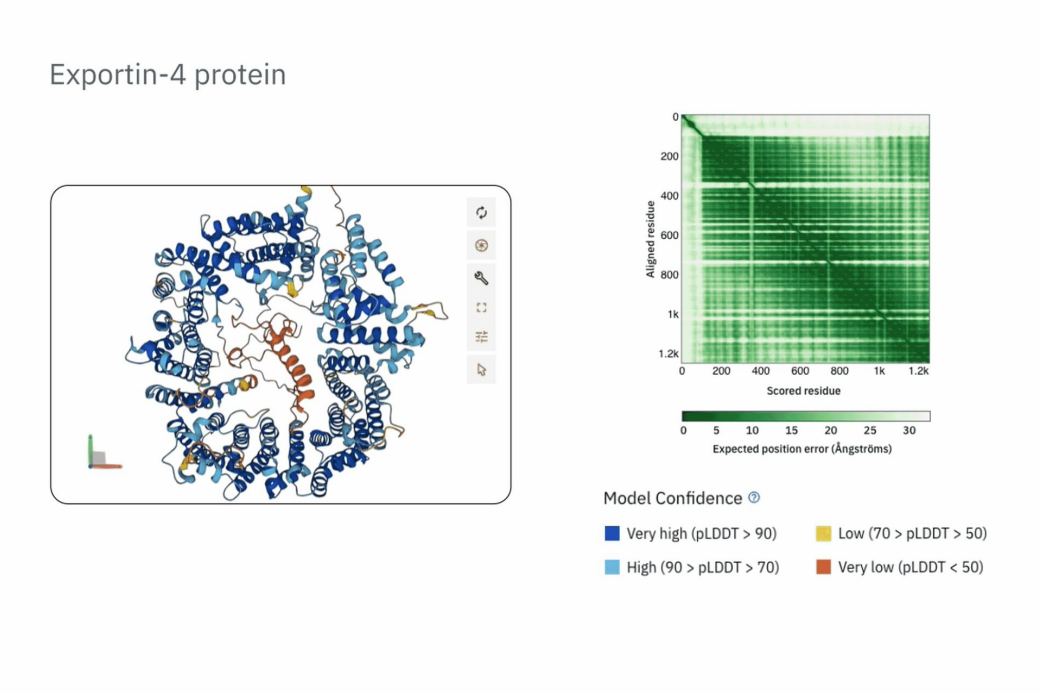
ESX-1 secretion system protein EccD1, P9WNQ7
https://alphafold.ebi.ac.uk/entry/P9WNQ7

- **PAE is mainly used to assess relative domain positions.**
- **Low PAE** for residue pairs from two different domains, indicates a **good** prediction of the relative positions and orientations for them.
- **High PAE** for residue pairs from two different domains, then indicates that the relative positions and/or orientations of these domains are **uncertain** and should not be interpreted.

Google DeepMind

EMBL-EBI

Important to use both outputs together



- Involved in protein export from the nucleus.
- Predicted to adopt a donut-like shape, with an N-terminal helix inside the hole of the donut.
- Should suggest a functional role of the helix?

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

Google DeepMind

EMBL-EBI

Limitations

- AlphaFold2 only **accepts** the **20 standard amino** acids in its input.
- AlphaFold2 (by default) predicts **five models per run**:
  - However, these models are generally very similar. In other words, AlphaFold2 usually **cannot predict conformational variability** in a protein.
- Cannot predict **assemblies** (AlphaFold-Multimer was trained to do this).
- Fails to predict the effects of **point mutations**.
- Not designed to predict the **binding of ligand molecules or post-translational modifications** (AlphaFill adds small molecules).
- **Not designed to fold nucleic acid** structures or model protein-DNA and protein-RNA complexes.
- Not aware of the **membrane plane** for the transmembrane proteins.
- Rarely predicts correct **antigen-antibody** interactions.
- Reduced performance for **orphan proteins.**

What can I do with these predicted structures?

- AlphaFold models are routinely used as **starting models** for X-ray crystallography and cryo-EM but can be used for further analysis
- **Guide mutational analysis**.
- **Generate and test hypotheses**
- **Propose mechanisms of protein action**, which are crucial for early-stage biomedical research.
- Map known **pathogenic mutations** on predicted structures.
- Uncover proteins' functions and **evolutionary relationships.**

Erickson, E., Gado, J.E., Avilán, L. et al. Sourcing thermotolerant poly(ethylene terephthalate) hydrolase scaffolds from natural diversity. Nat Commun 13, 7850 (2022).

What can I do with these predicted structures?

- AlphaFold models are routinely used as **starting models** for X-ray crystallography and cryo-EM but can be used for further analysis
- **Guide mutational analysis**.
- **Generate and test hypotheses**
- **Propose mechanisms of protein action**, which are crucial for early-stage biomedical research.
- Map known **pathogenic mutations** on predicted structures.
- Uncover proteins' functions and **evolutionary relationships.**

Generally requires additional tools -> new AFDB integrations

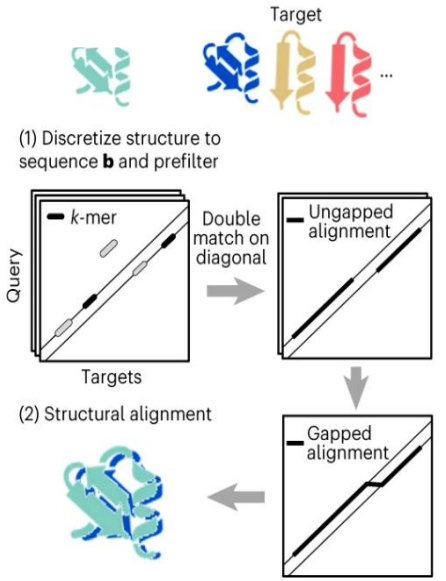
Erickson, E., Gado, J.E., Avilán, L. et al. Sourcing thermotolerant poly(ethylene terephthalate) hydrolase scaffolds from natural diversity. Nat Commun 13, 7850 (2022).

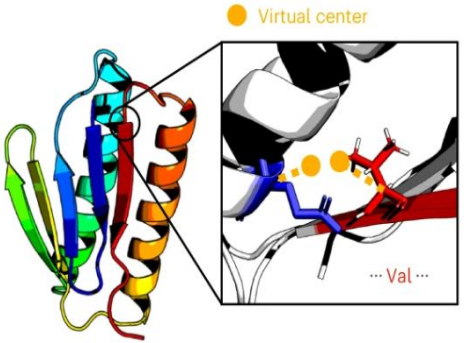
Advancing structure-Based Search: Foldseek

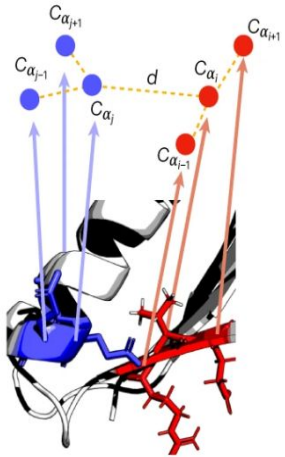

Jingi Yeo

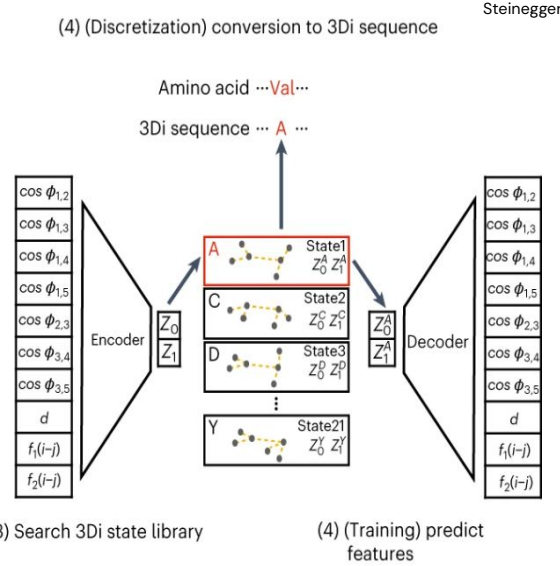

Milot Mirdita


Martin Steinegger









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

Prediction pages - Structure-based search

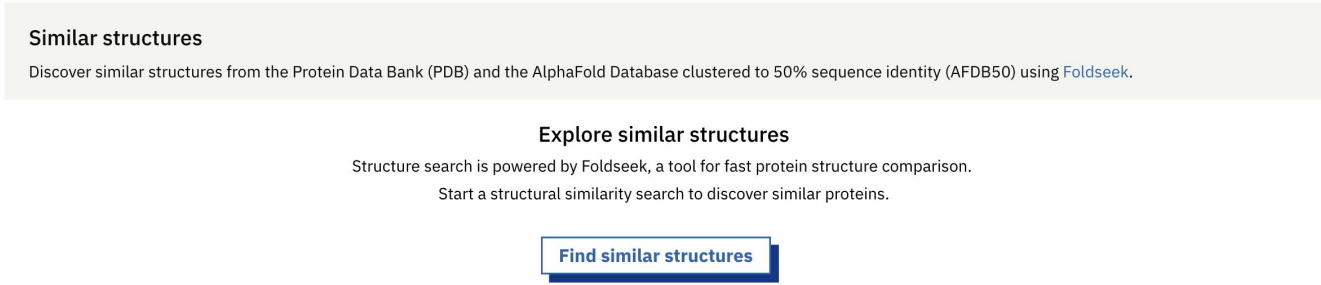
Search against:

AFDB50 (Clustered on 50% sequence identity) and

PDBe (all experimental structures)

Live search with Foldseek with an E-value threshold of 0.01

Smooth navigation between sequence and structural data



EMBL-EBI

Structure-based search with Foldseek

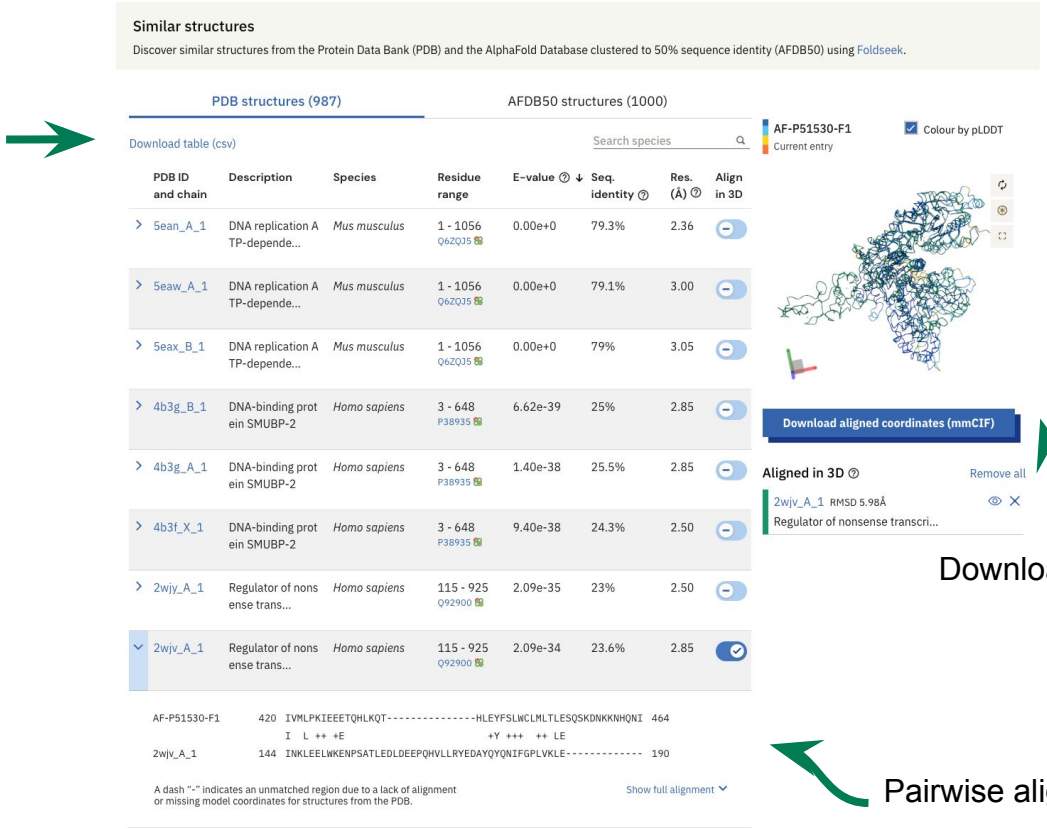
Search against:

AFDB50 (Clustered on 50% sequence identity) and

PDBe (all experimental structures)

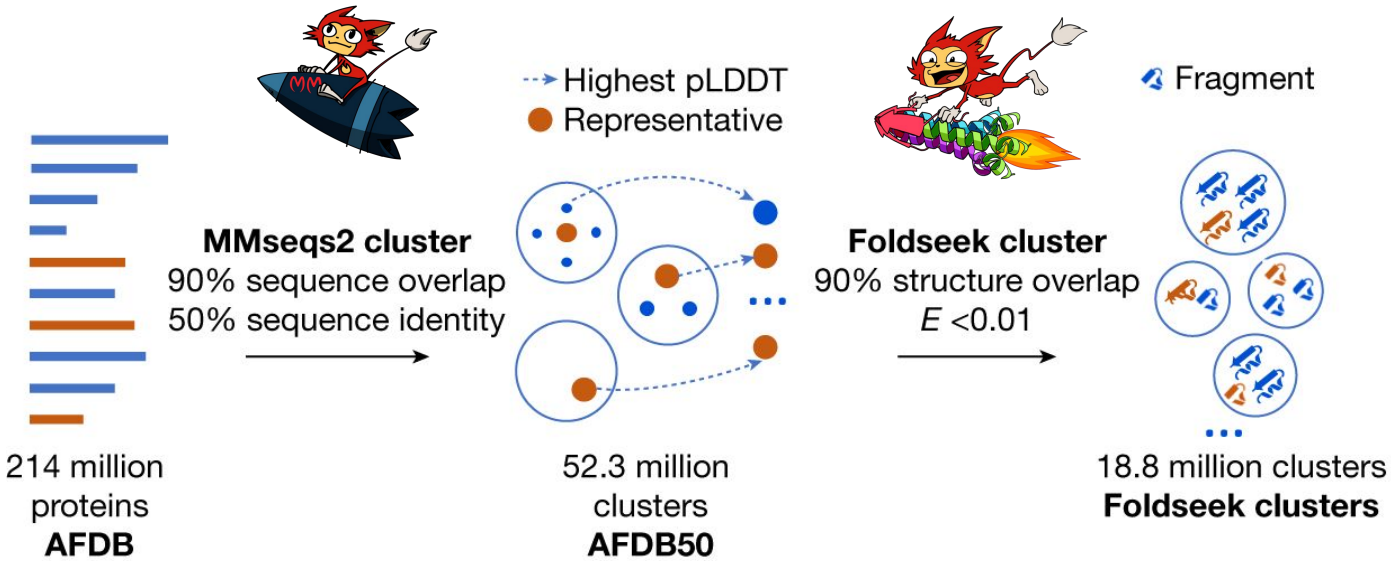
Live search with Foldseek with an E-value threshold of 0.01

Smooth navigation between sequence and structural data



EMBL-EBI

Scaling the protein universe



Google DeepMind

Barrio-Hernandez, I., Yeo, J., Jänes, J. et al. Clustering predicted structures at the scale of the known protein universe. Nature 622, 637–645 (2023).

EMBL-EBI

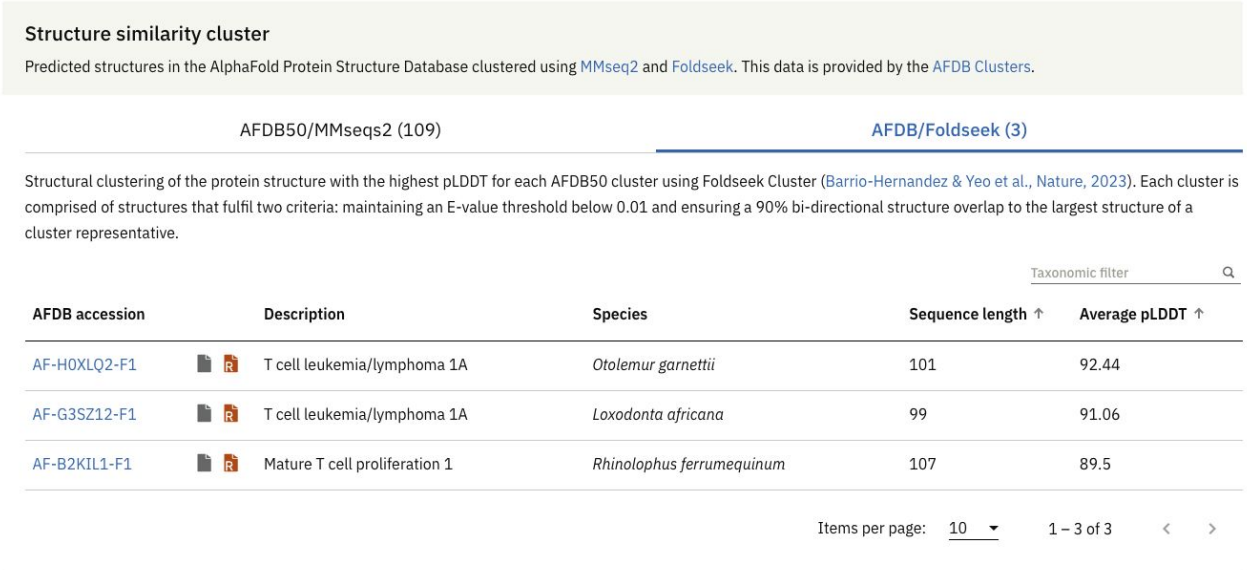
Shedding light on the structural organisation of the whole AlphaFold database - **the cluster table**

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

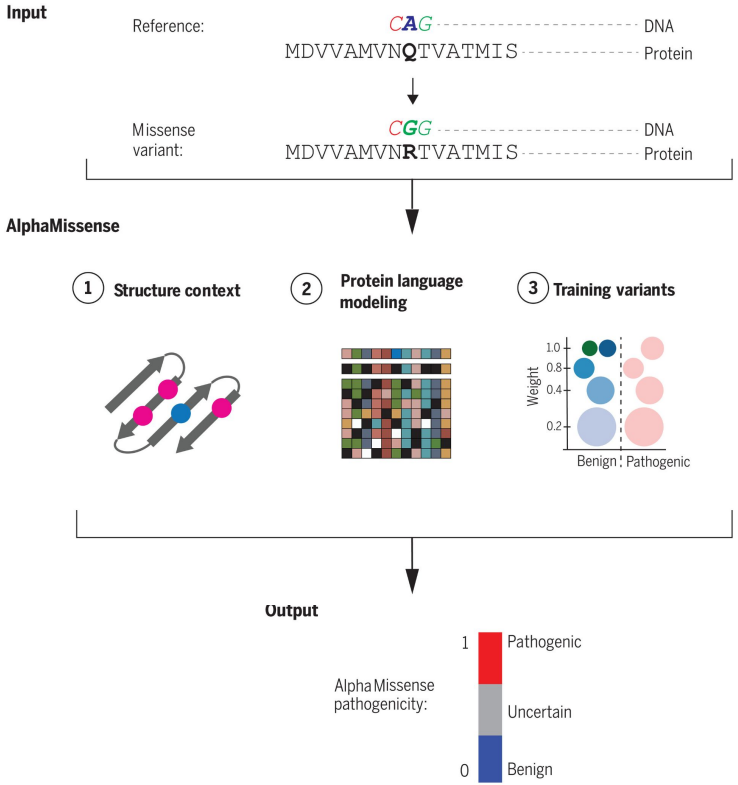
Taxonomic filter to filter by species

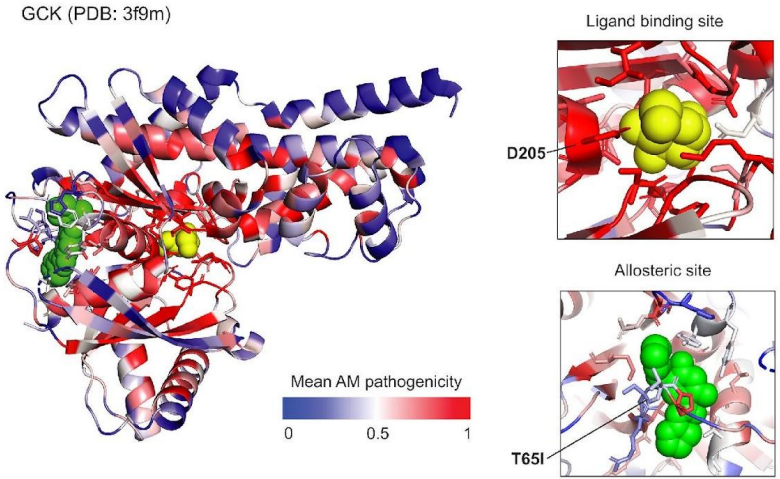


Protein p13 MTCP-1 (AF-P56278-F1)
Homo sapiens (Human)

EMBL-EBI

AlphaMissense, human proteome-wide missense prediction





Google DeepMind

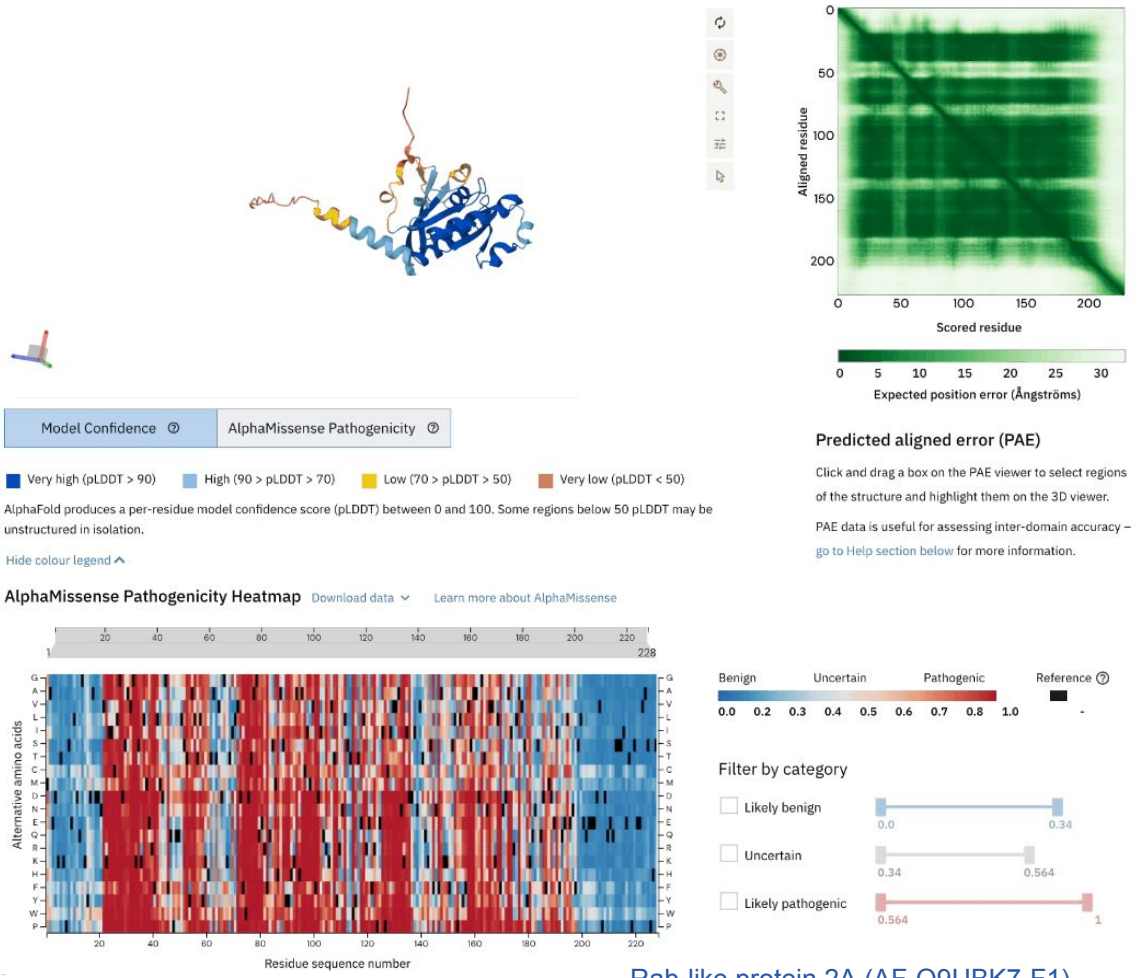
Cheng, J et al. ,Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science (2023)

EMBL-EBI

You can toggle between pLDDT and average AlphaMissense pathogenicity

Interactive heatmap
Zoom in and select residue of interest

Filter heatmap view by category and adjust range



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

Google DeepMir

EMBL-EBI

AlphaMissense Limitations

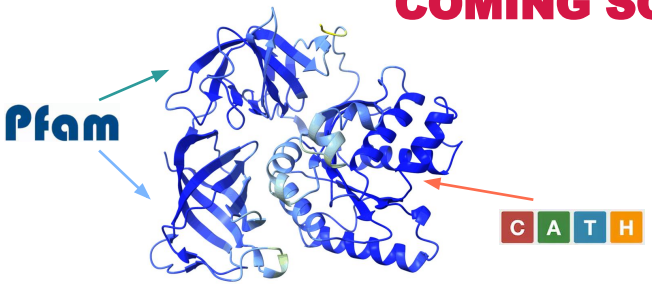
- AlphaMissense **does not directly provide detailed structural change** information for altered sequences.
- AlphaMissense does not explicitly predict missense variants' impacts on **biophysical properties,** like, stability and binding affinity.
- AlphaMissense is primarily designed to predict the 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 .

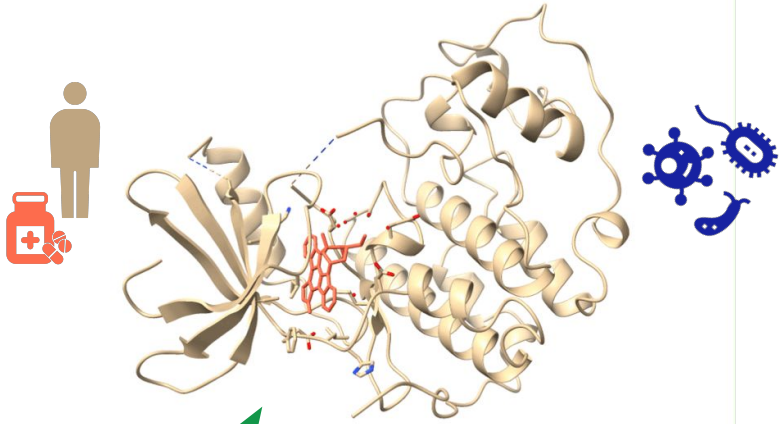
Enriching AlphaFold DB

Identify remote homologues

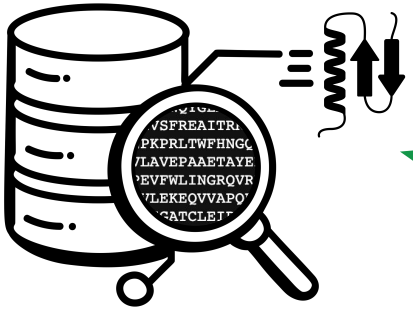
Sequence and structure annotations

COMING SOON

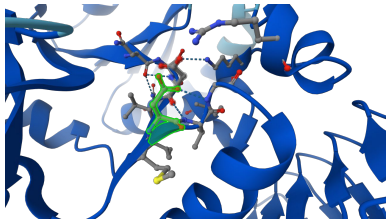




Sequence- and structure-based search



Residue level annotations

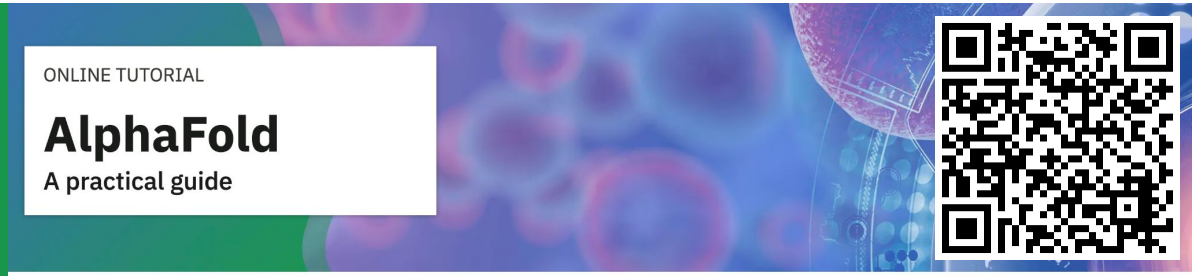


watch this space

EMBL-EBI

AlphaFold: A practical guide

- AlphaFold significantly impacts scientific research
- EMBL-EBI and Google DeepMind created an online training for AlphaFold adoption
- Training available on the EMBL-EBI platform
- Offers interactive learning for various expertise levels



Enter course

Mark as favourite

Proteins are essential components of life, predicting their 3D structure enables researchers to get an insight into its function and role. AlphaFold is an artificial intelligence (AI) system, developed by Google DeepMind, that predicts a protein's 3D structure based on its primary amino acid sequence. It regularly achieves accuracy competitive with experiment.

Google DeepMind

EMBL-EBI

AlphaFold DB FAQ and Help Desk

- Please first check the **FAQ page for common questions** here https://alphafold.ebi.ac.uk/faq.
- For **questions and feedback about the AlphaFold DB website**, please contact afdbhelp@ebi.ac.uk.
- For **sharing feedback on structure predictions or for questions about AlphaFold not directly related to the database**, please contact the AlphaFold team at alphafold@deepmind.com. We may not be able to respond to every query and there may be some delay before we can get back to you.
- For other **questions about AlphaFold not directly related to the database**, please contact the AlphaFold team at alphafold@deepmind.com. Please do not share anything confidential with Google DeepMind.
- For press enquiries, please contact press@deepmind.com or comms@ebi.ac.uk.

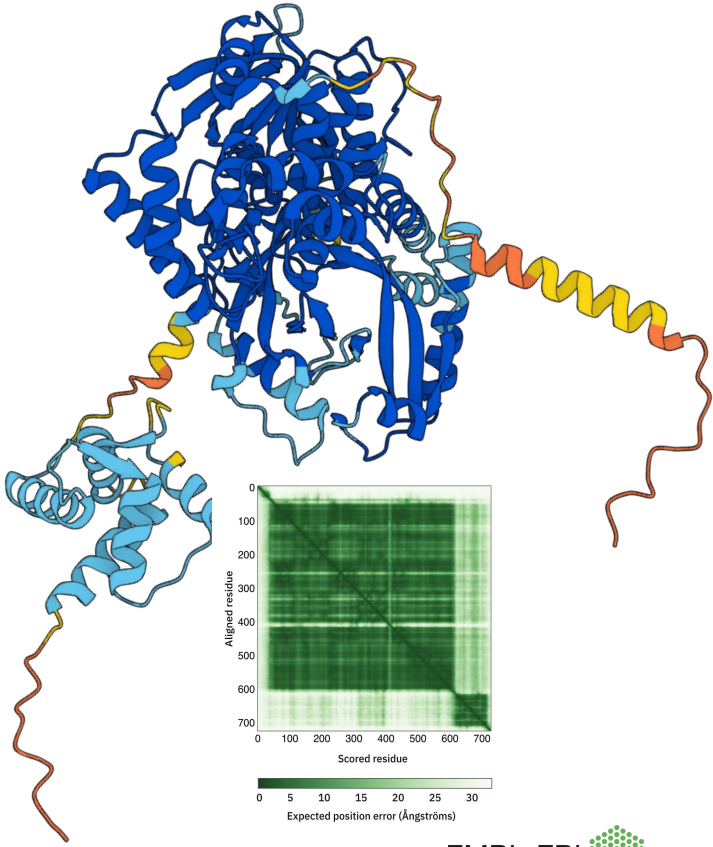
What types of data does the AFDB store?

- For the 214,684,311 proteins:
  - Predicted atomic coordinates
  - Confidence metrics (pLDDT and PAE)
  - Metadata

What do we not (yet) have in the database?

  - There are no multiple conformations of protein structures.
  - There are no viral proteins.
  - There are no isoforms.
  - There are no assemblies.
  - There are no mutant structures.
  - There are no ligands
  - Only AlphaFold2 models

See 3D-Beacons



Glycerol-3-phosphate dehydrogenase, mitochondrial, P43304
https://alphafold.ebi.ac.uk/entry/P43304

Google DeepMind

EMBL-EBI

3D-Beacons Network

3D-Beacons[1] is an open collaboration between providers of macromolecular structure models.

Unified programmatic access to experimentally determined and predicted structure models.

The main purpose of 3D-Beacons is to provide programmatic access to experimentally determined and theoretical protein structures.

Data providers include: AlphaFold DB, PDBe, Swiss-Model, ModelArchive, SASBDB, PED, AlphaFill and more

Provides API endpoints keyed on UniProt accessions

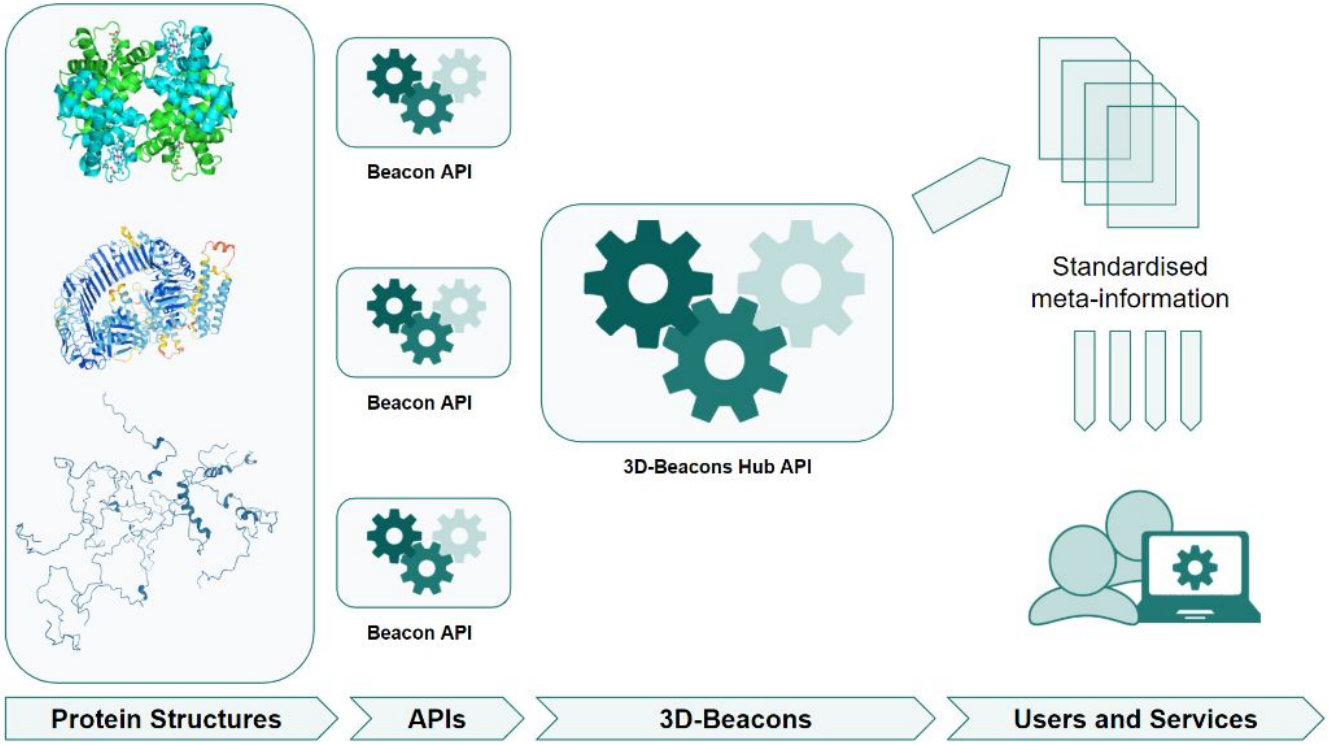
Supports sequence-based search

Sreenath Nair

[1] 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).

EMBL-EBI

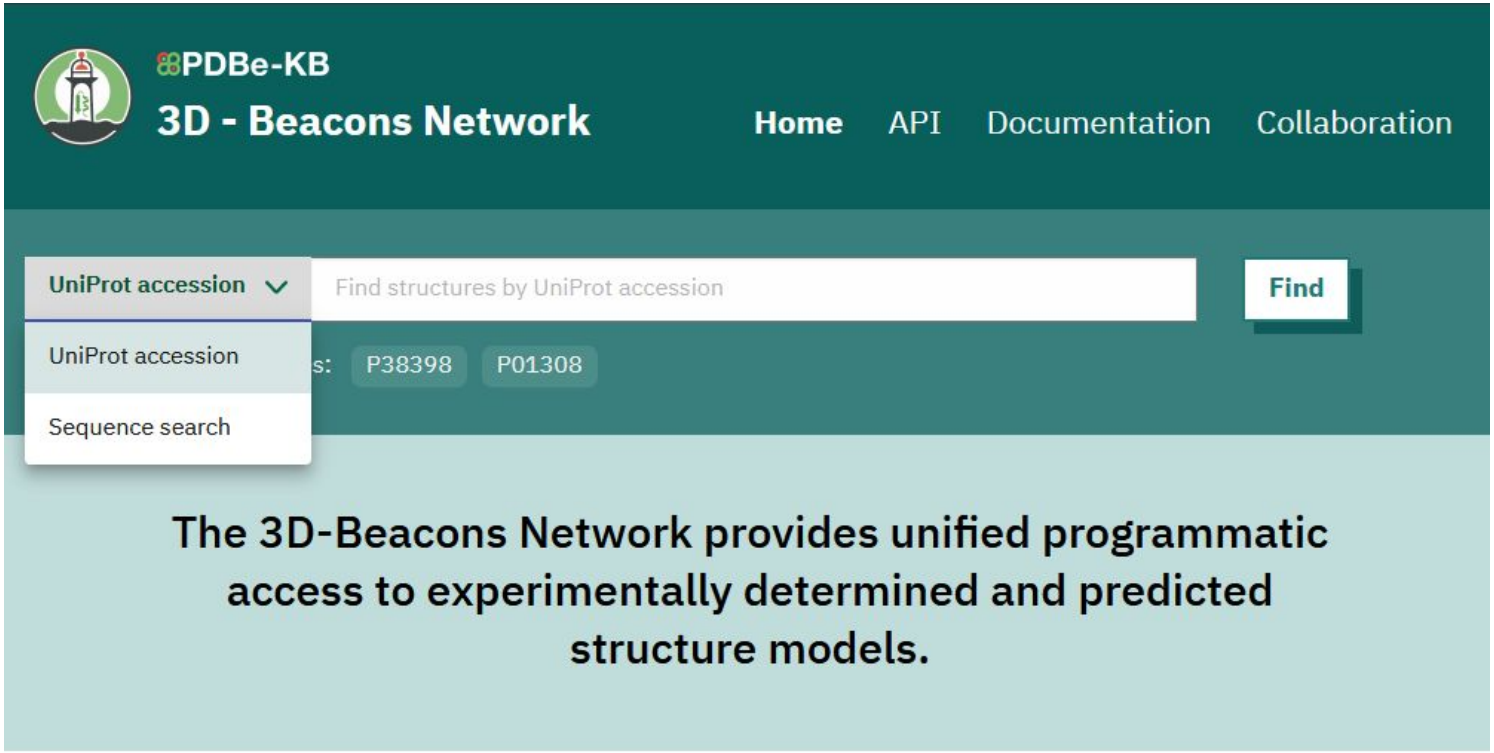
3D-Beacons Network



All of the data provided is freely available for both academic and commercial use under Creative Commons Attribution 4.0 (CC-BY 4.0)

The 3D Beacons API retrieves and combines data from member data providers, and returns the data in JSON format according to the 3D-Beacons API specification.

SEARCH:
Predicted & experimentally-determined



SWISS-MODEL

AlphaFold Protein Structure Database
Developed by DeepMind and EMBL-EBI

ModelArchive

isoform.io

HegeLab.org

LevyLab

SASBDB
Small Angle Scattering Biological Data Bank

PED Protein Ensemble Database

PDBe
Protein Data Bank in Europe

Ensembl

EMBL-EBI

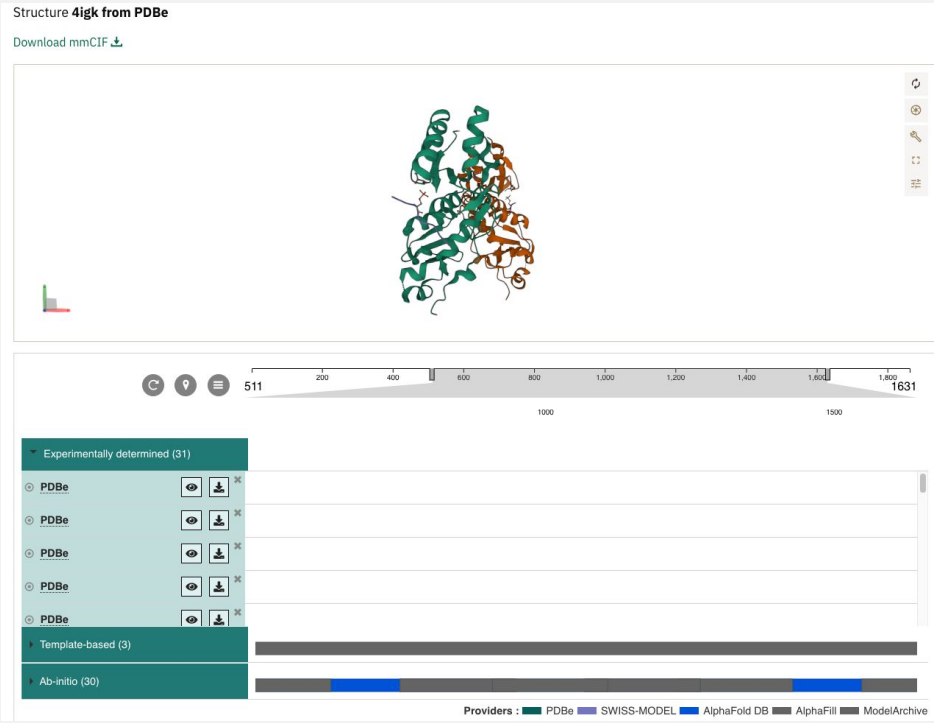
3D-Beacons: Finding computational models

P38398 (BRCA1_HUMAN) - 64 Structures available

Information

Protein	Breast cancer type 1 susceptibility protein Go to UniProt
Gene	BRCA1
Source organism	*Homo sapiens*
Biological function	E3 ubiquitin-protein ligase that specifically mediates the formation of 'Lys-6'-linked polyubiquitin chains and plays a central role in DNA repair by facilitating cellular responses to DNA ... Show more

31	0	3	30
Experimentally Determined Structures	Conformational Ensembles	Template-based models	Ab-initio Models



EMBL-EBI

SEARCH:
Predicted & experimentally-determined

Harnessing the 3D-Beacons Network:

A Comprehensive Guide to Accessing and Displaying Protein Structure Data

https://doi.org/10.1002/cpz1.1047

- Scripting assistance / example available *via* Google Research Colab
- Google Research Colab allows anybody to write and execute arbitrary python code through a browser.

Using 3D-beacons to:

- Basic (UniProt accession code) search
- Sequence-based search
- Download specific model

Acknowledgements

Resources

PDBe
Protein Data Bank in Europe

PDBe-KB
Protein Data Bank in Europe - Knowledge Base

AlphaFold
Protein Structure Database

PDBe
@PDBeurope
@PDBeurope
proteindatabank
proteindatabank
pdbeurope

The Protein Databank in Europe team

EMBL-EBI

Acknowledgements

Funding

EMBL

UKRI UK Research and Innovation

wellcome

Funded by the European Union

Google DeepMind

Collaborators

WORLDWIDE PDB PROTEIN DATA BANK

PDBe-KB Protein Data Bank in Europe - Knowledge Base

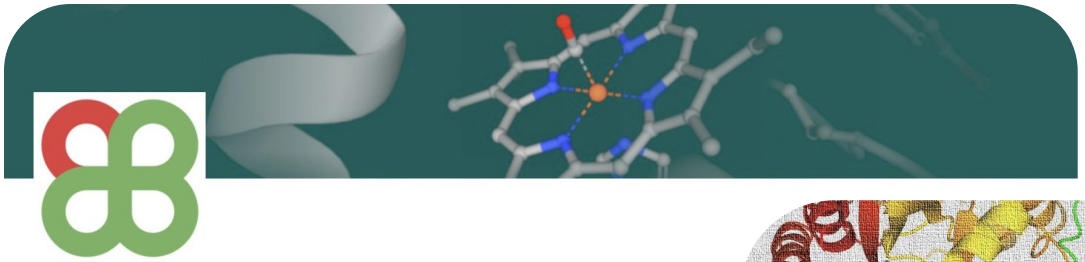
3D-Beacons

CEITEC Central European Institute of Technology BRNO | CZECH REPUBLIC

3D-BioInfo

Google DeepMind

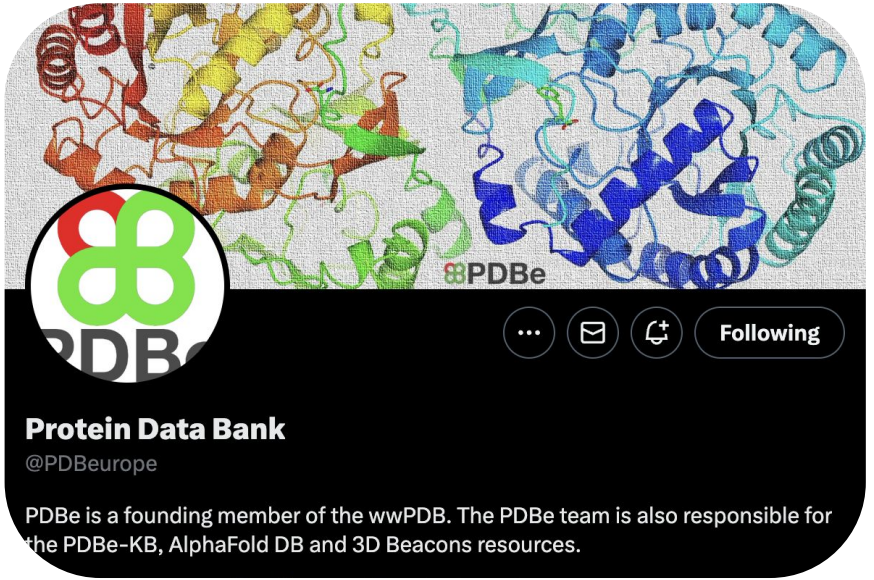
Follow the PDBe!



Protein Data Bank in Europe (PDBe)

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

Software Development · Hinxton, Cambridge · 545 followers · 11-50 employees



EMBL-EBI