AlphaFoldology
What is next in impact of AlphaFold?

Karel Berka

31st January 2025

AlphaFold(ology)

- Why?
  - Protein structure prediction
- What?
  - CASP
- How?
  - AFx - under the hood
- What next?
  - AlphaFoldology

Motto:

“Disruptive scientific breakthroughs **raise more questions than they answer**. They open new research avenues and can inspire entirely new fields of study. Just as the Human Genome moment marked the beginning of a revolution in genomics, so too **AlphaFold might usher in a new era in biology**.”

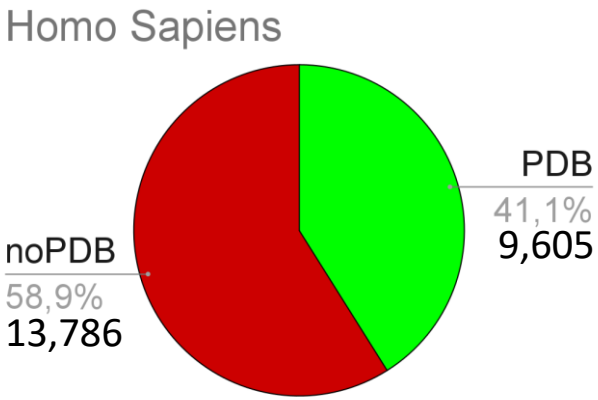
Janet Thornton, director emeritus EMBL-EBI - *22 July **2021**, Cambridge*

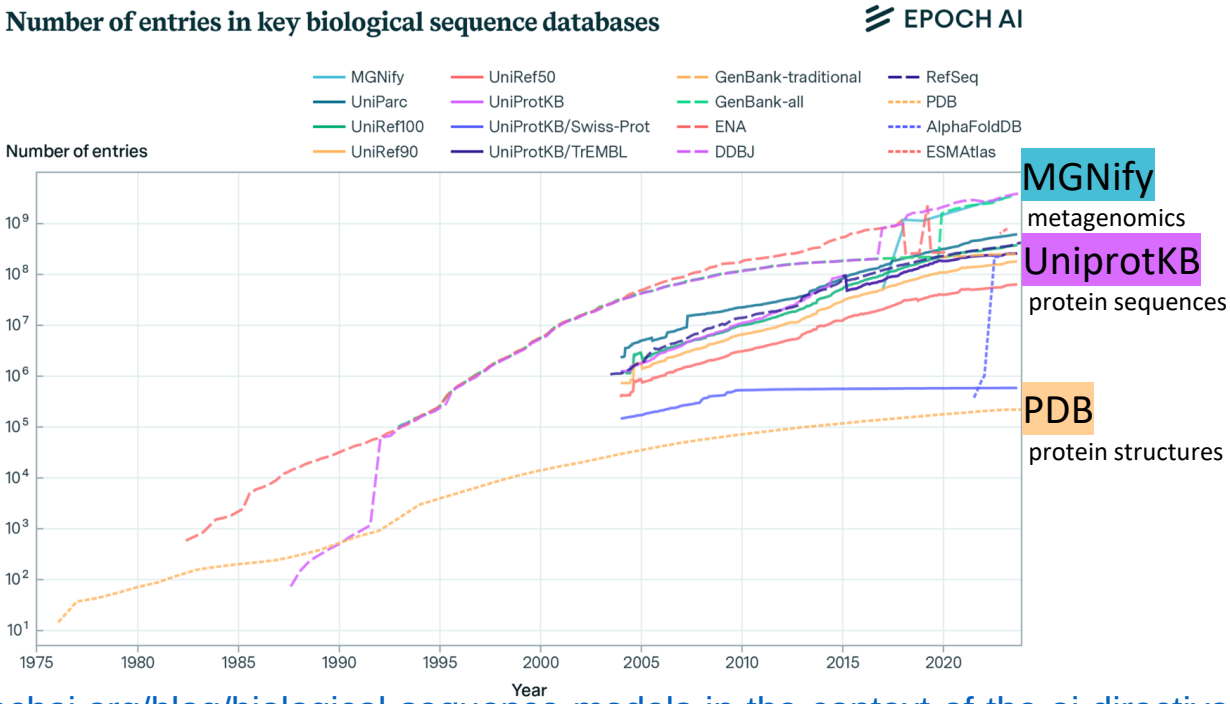
elixir CZECH REPUBLIC

Why?

Proteins are workers and their structure means function

Solving 3D structures is still difficult...





https://epochai.org/blog/biological-sequence-models-in-the-context-of-the-ai-directives

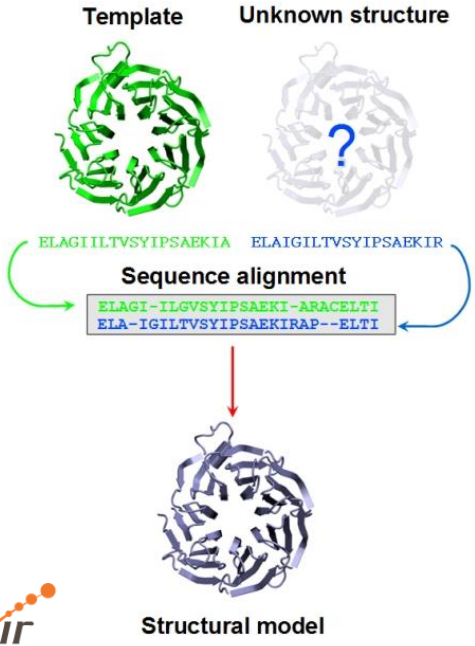
The gap between numbers of experimental structures and sequences is increasing over time

What?

Protein structure prediction problem

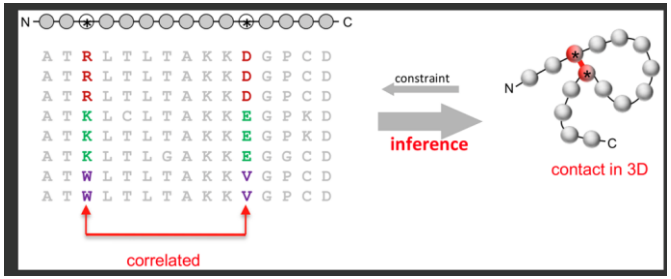
Principles of prediction from sequence

Template-based

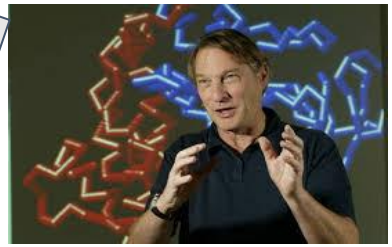


https://www.unil.ch/pmf/en/home/menuinst/technologies/homology-modeling.html

Covariance



NEED for VALIDATION

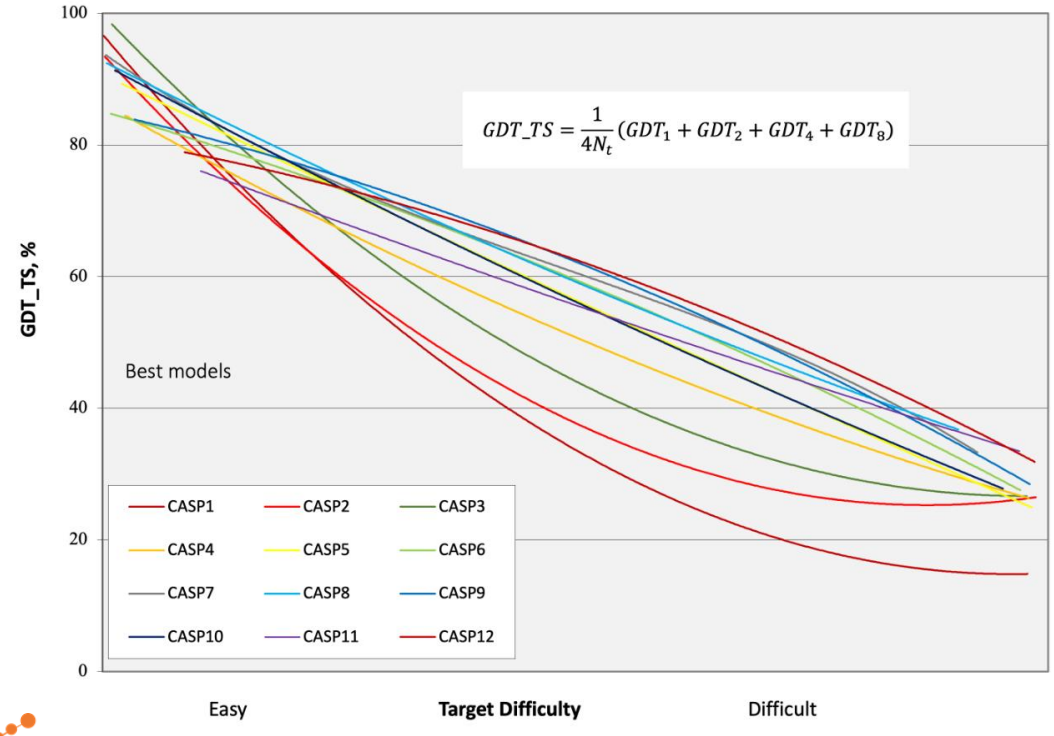


- **CASP - critical assessment of protein structure prediction**
- since 1994 biannually
- compare with experimentally solved structures in PDB

JOHN MOULT
CO-FOUNDER AND CHAIR OF **CASP**, UNIVERSITY OF MARYLAND

HELEN BERMAN **PHILIP BOURNE**
FORMER PDB DIRECTORS

How to compare structures?



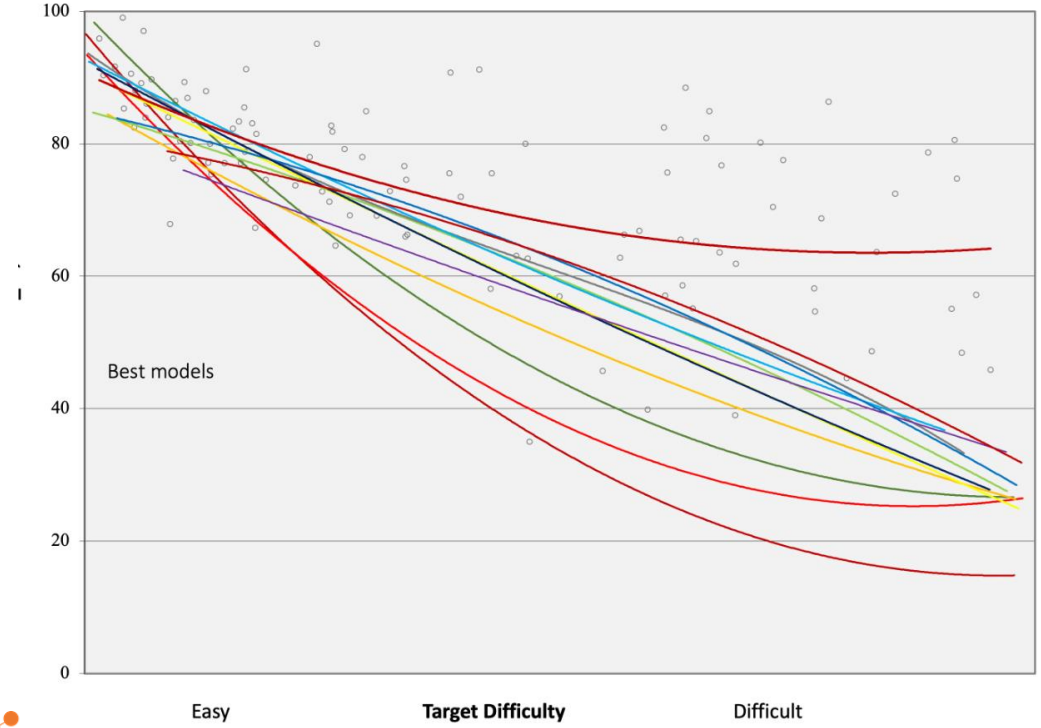
After precision growth CASP1(1994) -> CASP6(2006) protein prediction field was stuck

https://predictioncenter.org/casp14/doc/presentations/2020_11_30_CASP14_Introduction_Moult.pdf

GDT_TS = Global distance test - total score (max 100%)

The conventional GDT_TS total score in CASP is the average result of cutoffs at 1, 2, 4, and 8 Å falling within experimental position

CASP13(2018) - AlphaFold enters...



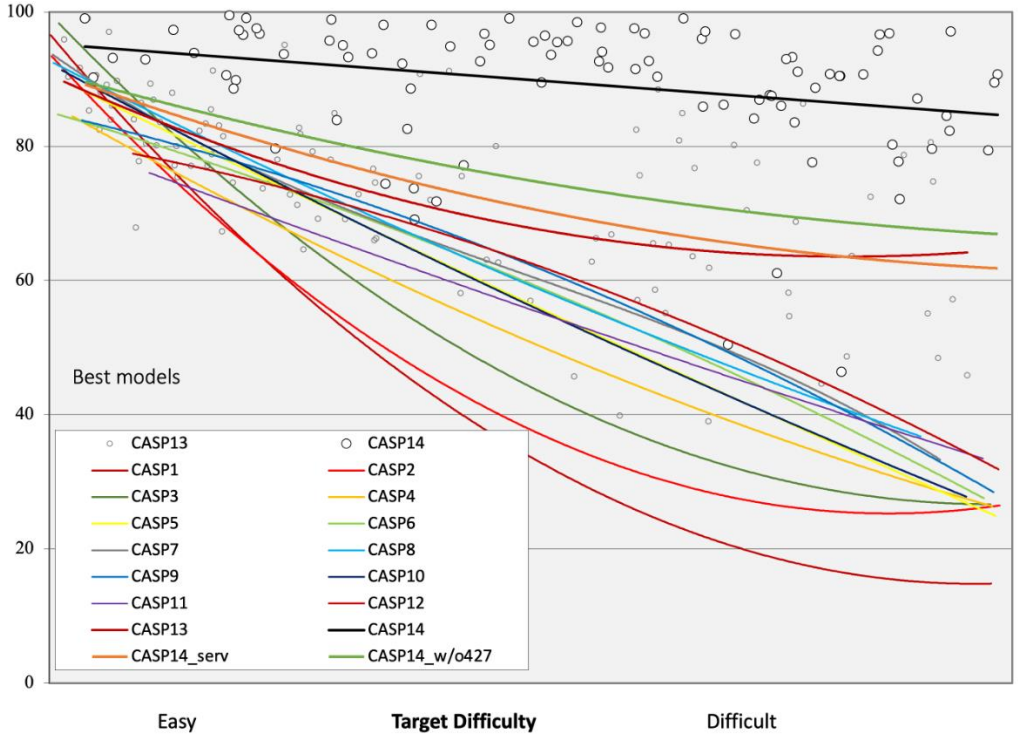
DEMIS HASSABIS
DEEPMIND,
FORMER BULLFROG PRODUCTIONS (THEME PARK)

DeepMind: 1st company to ever attend CASP
! not open model
(too expensive to use)

https://predictioncenter.org/casp14/doc/presentations/2020_11_30_CASP14_Introduction_Moult.pdf

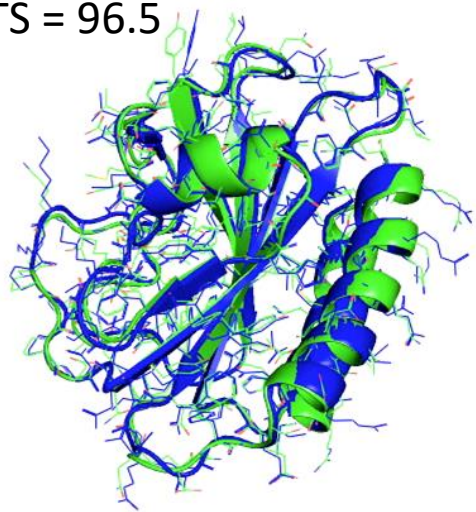
elixir CZECH REPUBLIC

CASP14(2020) - AlphaFold2 wins

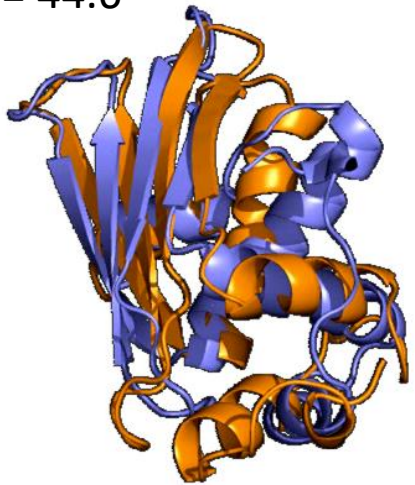


https://predictioncenter.org/casp14/doc/presentations/2020_11_30_CASP14_Introduction_Moult.pdf

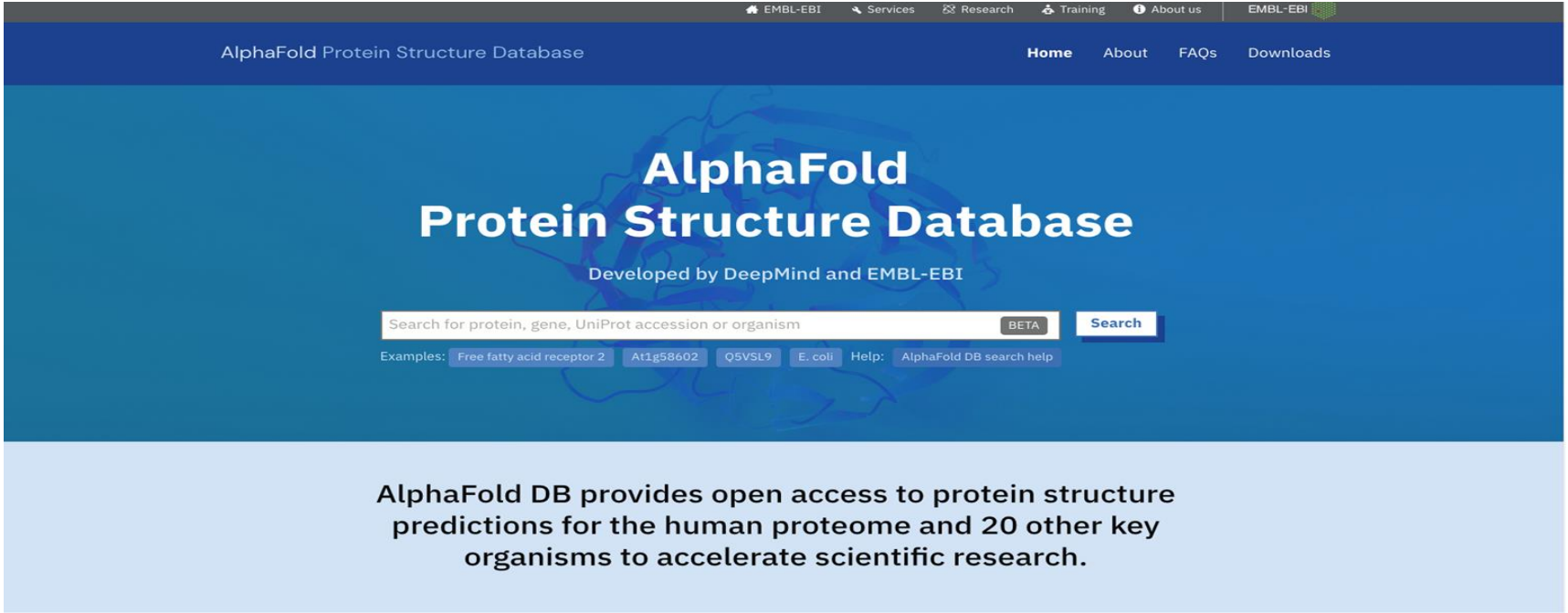
GDT_TS = 96.5



GDT_TS = 44.6



July 2021: AlphaFold2 open sourced with AFDB



"This will be one of the most important datasets since the mapping of the Human Genome."
Professor Ewan Birney
EMBL Deputy Director General and EMBL-EBI Director

https://www.alphafold.ebi.ac.uk/

CASP15(2022)

Best CASP15 broadly in line with best CASP14 but …

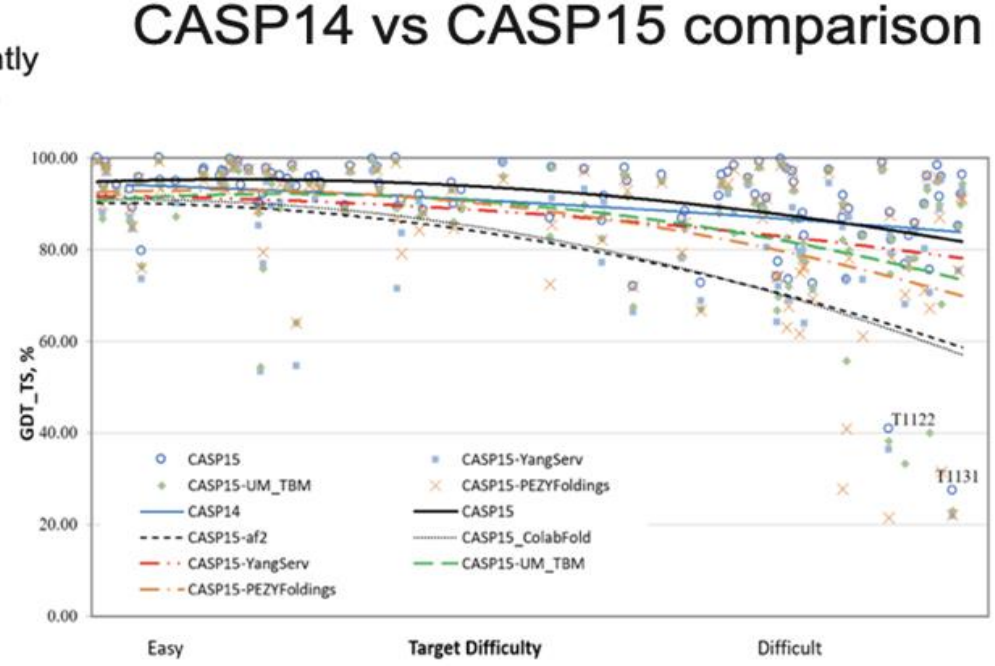
… best CASP14 (mainly AF2) consistently a little higher than best CASP15 groups

NBIS-af2-standard and ColabFold not performing at level of CASP14 DM AF2 submission

CASP invited DeepMind to informally model the set. Broadly this brings performance up to the best official CASP15 groups. vs AF2 'controls' they have

- retrained on current PDB
- increased sampling and crop size
- made some human interventions

So why persistent gap? Are CASP15 targets harder in ways not captured by this scale?



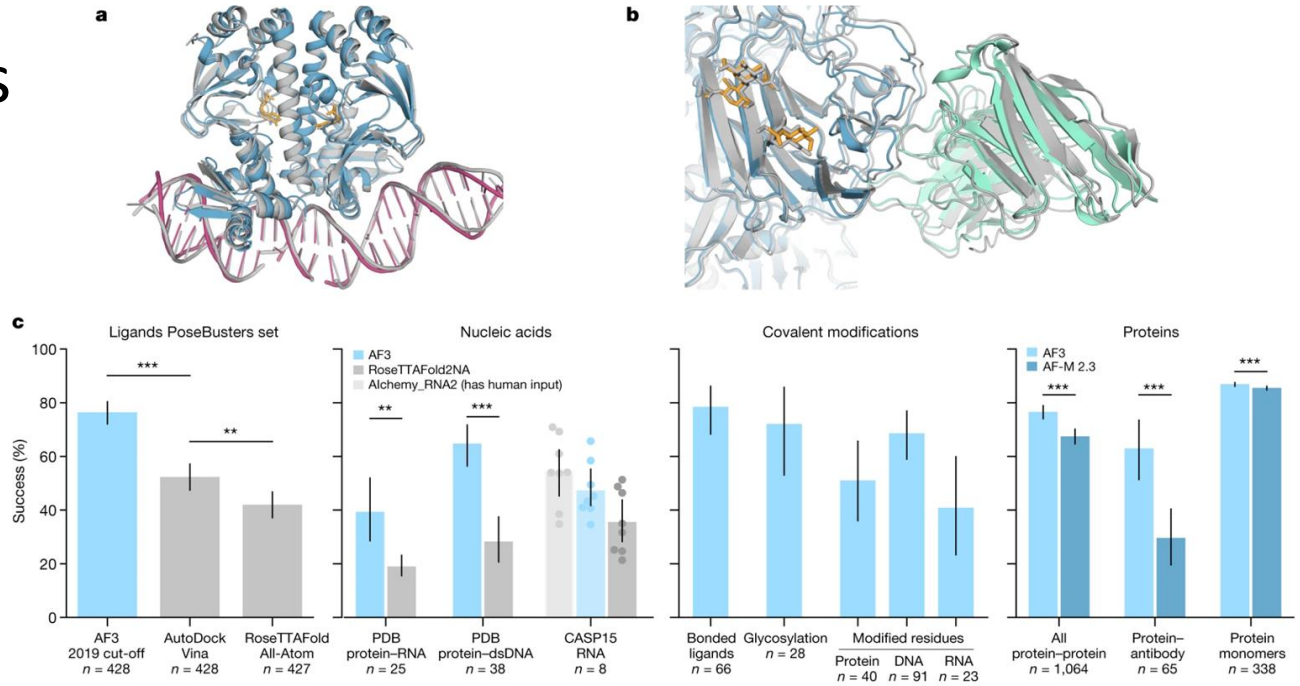
All successful tools based on AlphaFold

Dan Rigden ś slides from CASP 15 conference

May 2024:Return of the king - AlphaFold3

Model with ligands

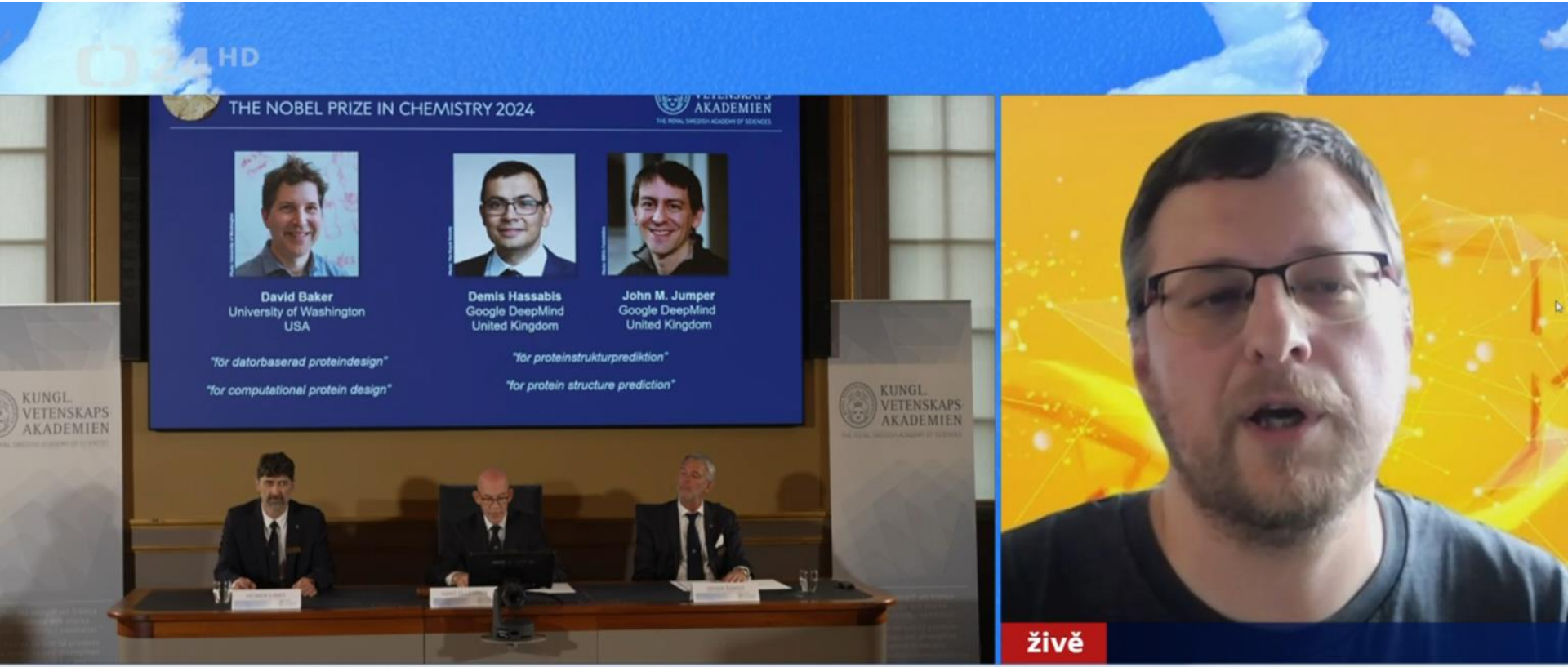
Public server



https://alphafoldserver.com/

Abramson, J., Adler, J., Dunger, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630, 493–500 (2024). https://doi.org/10.1038/s41586-024-07487-w

October 2024: Nobel prize in Chemistry



December 2024: CASP16

- Monomers

0. Single protein "folding problem" is still "solved": not a single fold is predicted incorrectly.

1. The current methods are known to be not sensitive to mutations and truncations. Future CASPs may focus on such cases.

2. Peripheral regions, irregular structures, regions involved in interactions have errors in predictions.

3. Viral and eukaryotic monomers are harder to predict well.

4. Proteins with shallow alignments are predicted worse.

5. Quality of experimental structures could be an issue, and they need to be checked by Mprob.

6. More extensive AF3 use gives better performance.

7. Progress compared to AF3 is measurable, but seemingly incremental (sorry!).

Qian Cong

https://predictioncenter.org/casp16/doc/presentations/Day-2/Day2-02-Cong-monomers_redacted.pdf

- Complexes

0. Exciting progress in antigen-antibody interactions. We may want more antibody targets in the future to more robustly evaluate the progress.

1. Protein complex modeling is not "solved": each group gets a subset correctly.

- CON'S:
  - Alternate conformers were ***not*** found
  - Nucleic acids remain problematic; DNA is easier than RNA
  - Scoring needs more development
    - MassiveFold "best" outperforms everybody, especially on "conventional" targets
  - Scorers recognize the *good* models, but not the *best* ones

Marc Lensink

https://predictioncenter.org/casp16/doc/presentations/Day-2/Day2-03-Lensink_CASP16_redacted.pdf

How?

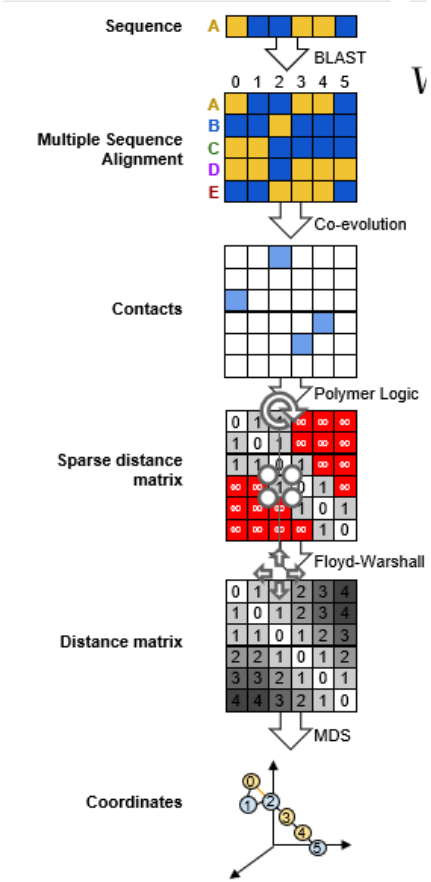
AlphaFold - under the hood*

* briefly - there are excellent step-by-step resources:
alphafold-decoded.com , alphafolding.ipynb, YouTube videos…

Alphafold 1 vz Alphafold 2

Alphafold 1

i) **co-evolutionary analysis** to map residue co-variation in protein sequence to physical contact in protein structure,
(ii) **deep neural networks** to identify patterns in protein sequence and co-evolutionary couplings and convert into distance-wise **contact maps**.
(iii) make **structure** from these contacts



Alphafold 2

i) **co-evolution** stands (MSA)

ii) deep neural networks exchanged for **transformer** - Evoformer

- **contact** maps exchanged for triangles

iii) structure module with MD optimization

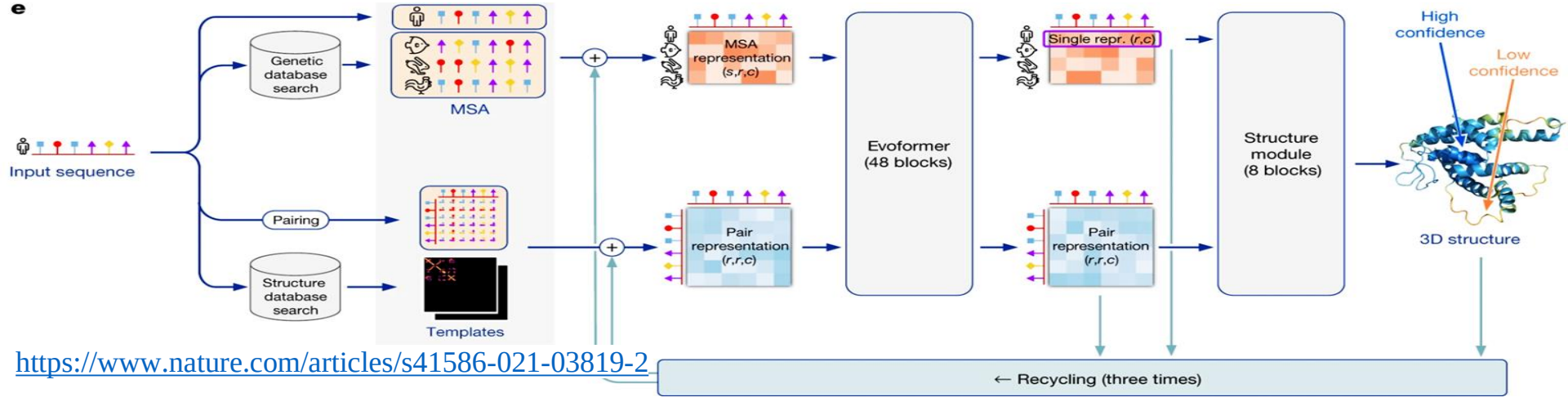
AlphaFold 2

Input: sequence

extended by **MSA** (at least 30 sequences) + **structural templates** (UniRef30 PDB-trained self distillation)

Evoformer and **Structure** model (w Amber MD simulation)

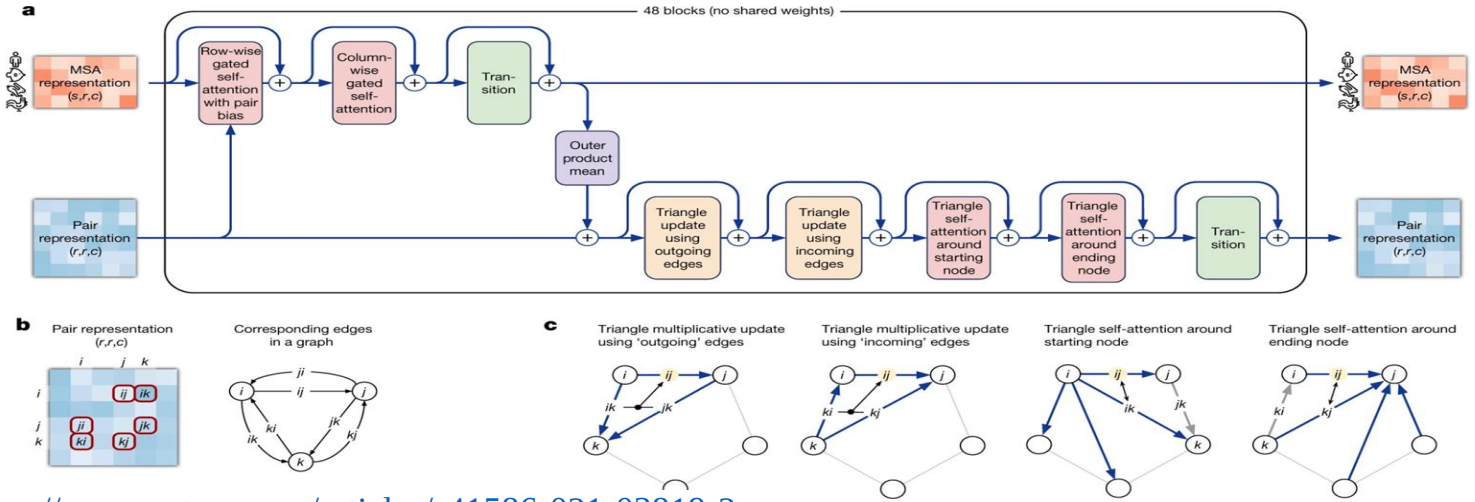
plDDT - predicted local confidence prediction



https://www.nature.com/articles/s41586-021-03819-2

EvoFormer

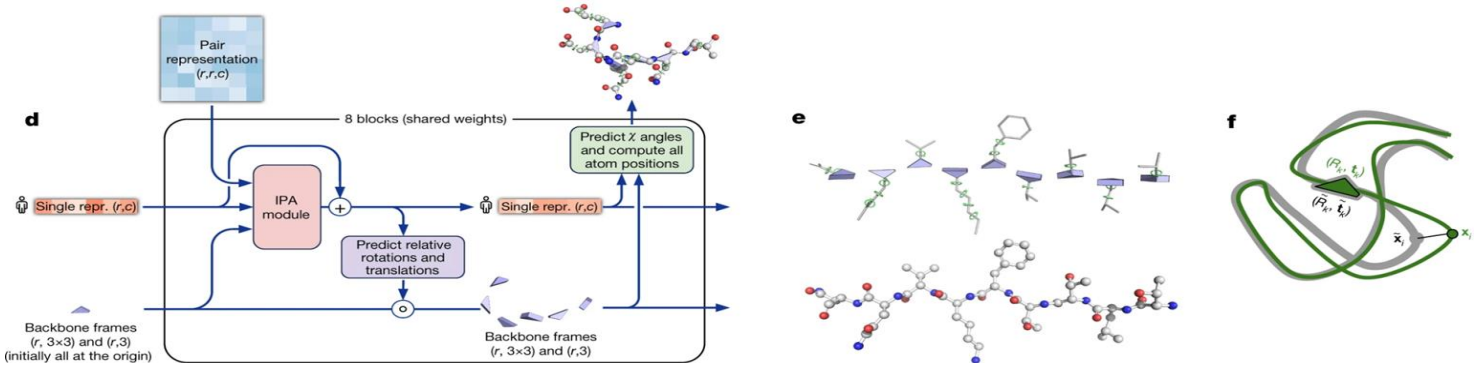
- mixing MSA and pairs via updates
- graph inference problem in 3D space
  - edges = residues in proximity
  - updates per each block (48 blocks) separately (AF1 updated all network at once)
- using triangles (instead of just pairs from contact map)



https://www.nature.com/articles/s41586-021-03819-2

Structure model

- prioritize backbone positions+orientations
  - residue gas - free floating rigid body rotations and translation
  - updates
    - IPA (invariant point attention) - neural activations only in rigid 3D
    - equivariant update using updated activations
- later fix backbone geometry
  - avoid loop closure problem)
- sidechain final refinement:
  - OpenMM with Amber 99sb forcefield

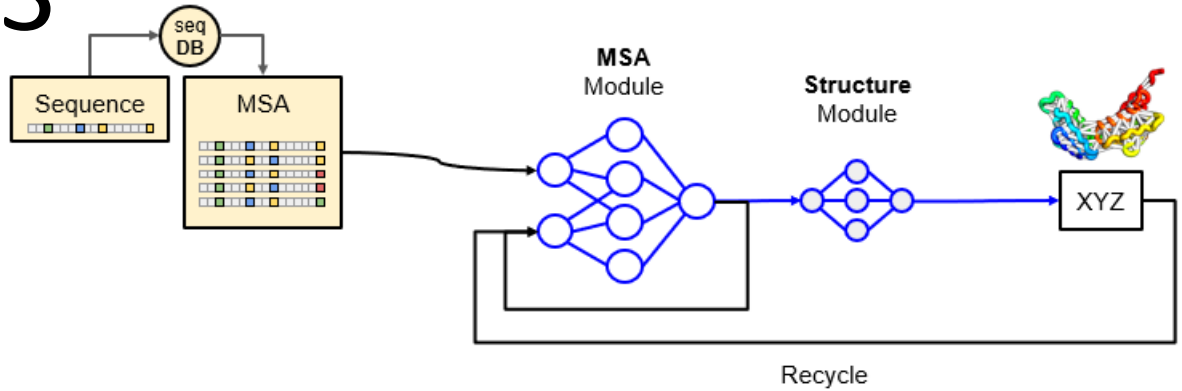


https://www.nature.com/articles/s41586-021-03819-2

AF2 vz AF3

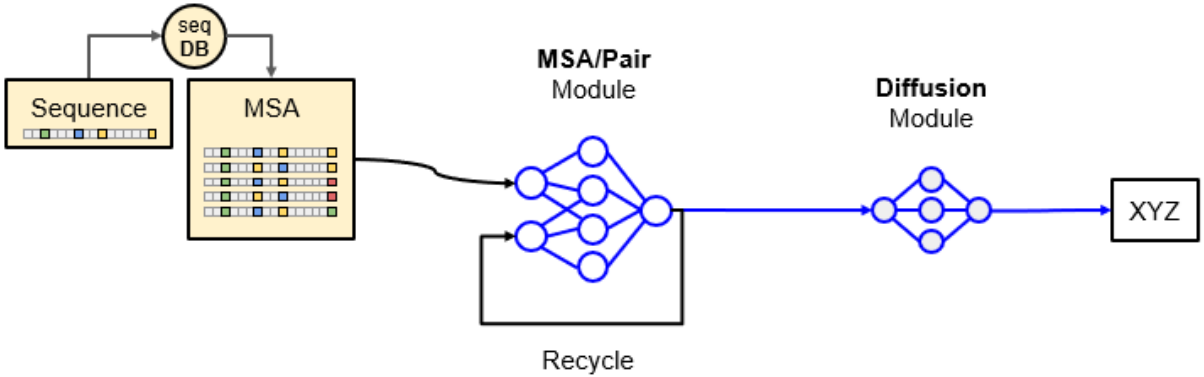
John Jumper

AF2



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

AF3



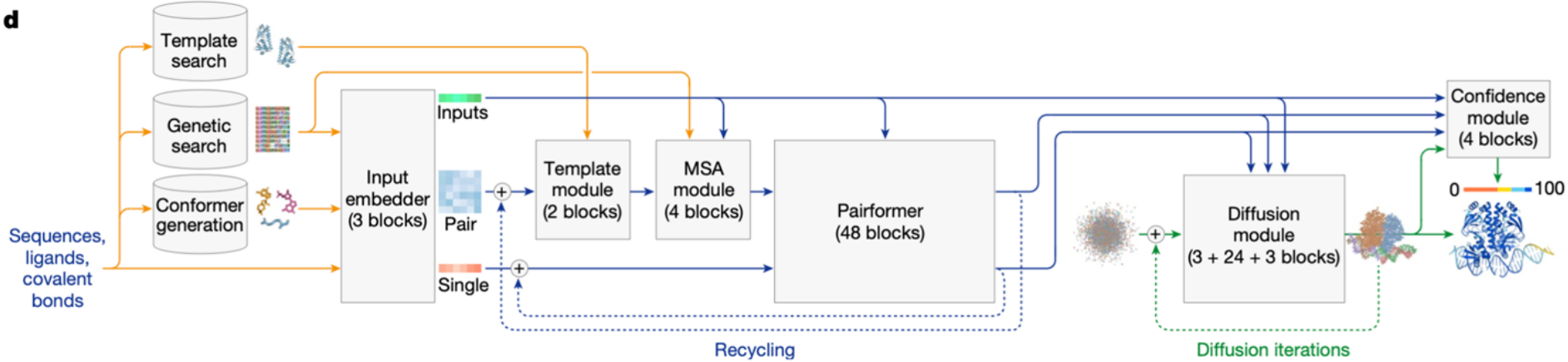
Abramson, J., Adler, J., Dunger, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630, 493–500 (2024). https://doi.org/10.1038/s41586-024-07487-w

AF3

diffusion module - enables to build **ligands**

- denoising towards structure

shortening MSA module -> speeding up calculations



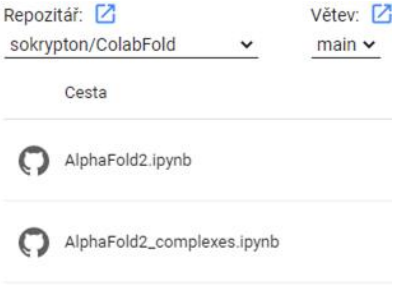
What next?

Usage - AlphaFoldology

Where to run AF?

https://alphafoldserver.com

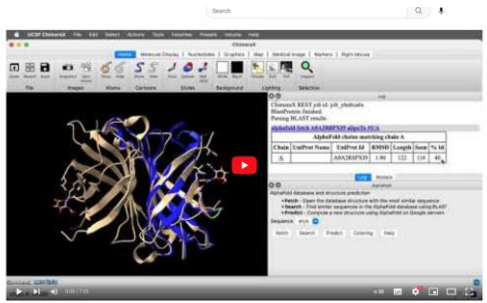
ColabFold in GoogleColab



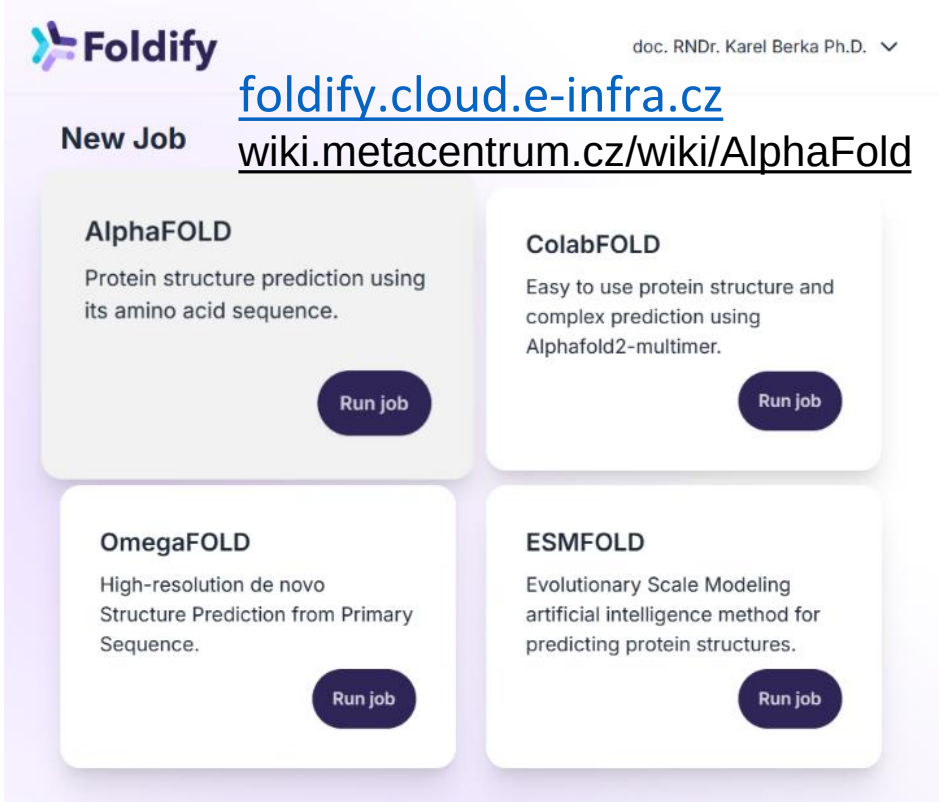
Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* 19, 679–682 (2022).
https://colab.research.google.com/github/sokrypton/ColabFold/

ChimeraX

- Fetch
- Search AFDB
- Predict



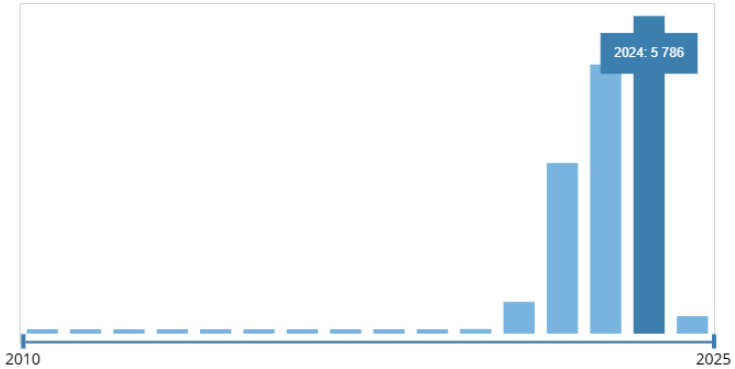
ELIXIR CZ/MetaCentrum/Galaxy



foldify.cloud.e-infra.cz
wiki.metacentrum.cz/wiki/AlphaFold

Alphafoldology

Alphafold led to enormous innovations in protein design



2018 2019 2020 2021 2022 2023 2024

CASP13
Alphafold 1

CASP14
AlphaFold 2
RoseTTAFold

AF2 code
AFDB
…

AFMultimer
AFConformer
ColabFold
AlphaFill
ProteinMPNN
…

AlphaMissense
ESMFold+Atlas
RoseTTAFoldNA
…
(all major DBs)

RoseTTAFoldAA
AlphaFold 3
Chai-1, Boltz-1 ,
Protenix,
NeuroPlexer, …
ChannelsDB 2.0

AF on monomers

and proteomes

SNW domain-containing protein 1

AlphaFold structure prediction

Download **PDB file** **mmCIF file** **Predicted aligned error**

Information

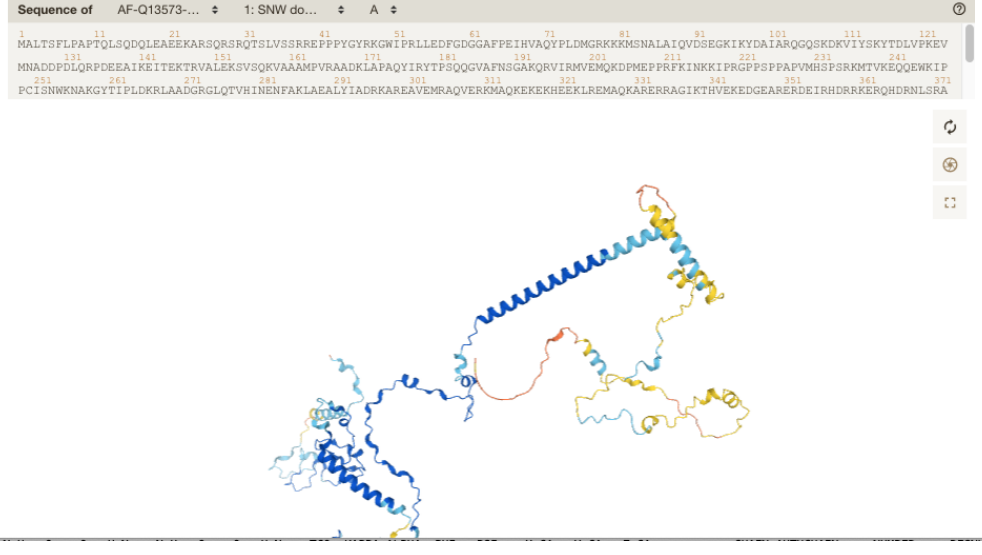
Protein	SNW domain-containing protein 1
Gene	SNW1
Source organism	Homo sapiens go to search
UniProt	Q13573 go to UniProt
Experimental structures	17 structures in PDB for Q13573 go to PDBe-KB
Biological function	(Microbial infection) Proposed to be involved in transcriptional activation by EBV EBNA2 of CBF-1/RBPJ-repressed promoters. go to UniProt

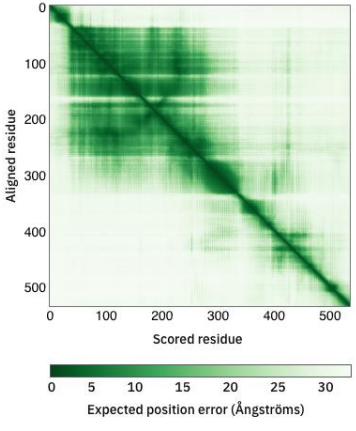
3D viewer

Model Confidence:

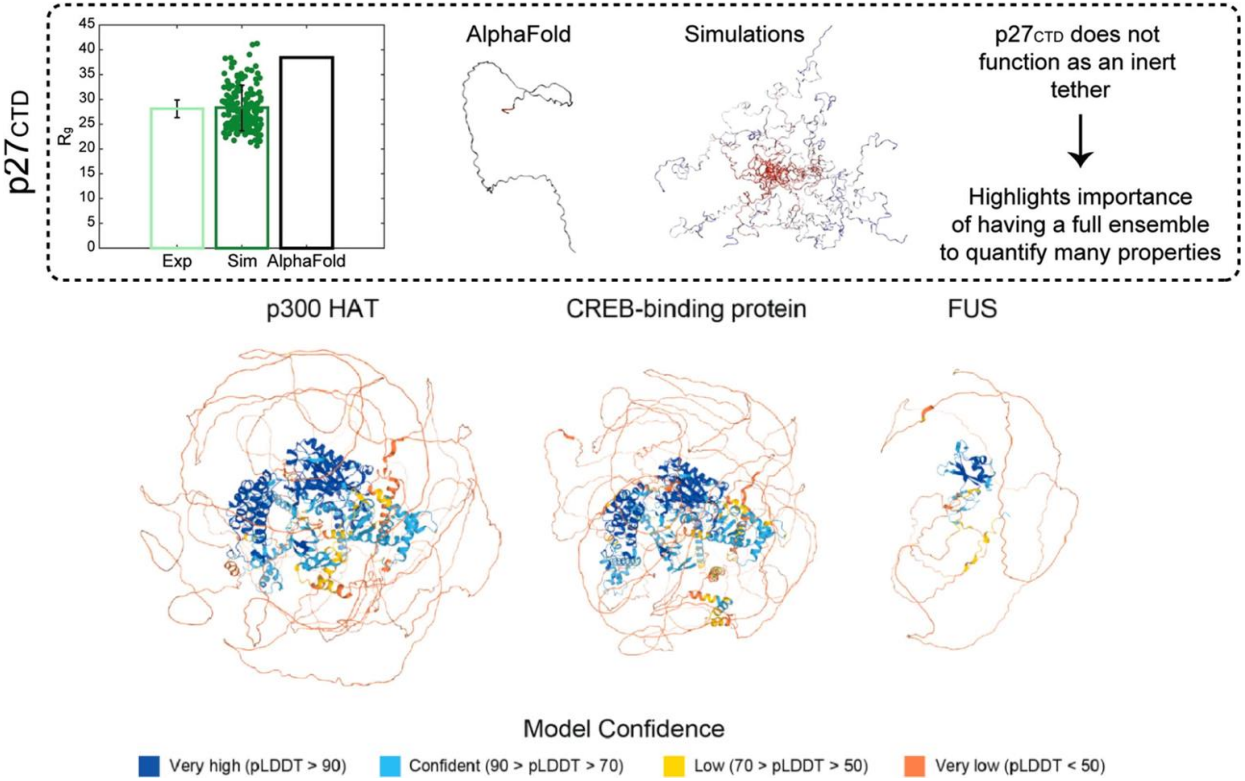
- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

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





AlphaFold and Intrinsically Disordered Proteins



Beware AF3!

Overstructurizes IDP

Ruff KM, Pappu RV ,AlphaFold and Implications for Intrinsically Disordered Proteins, J Mol Biol. 2021, https://doi.org/10.1016/j.jmb.2021.167208

channelsdb2.biodata.ceitec.cz

ChannelsDB 2.0: protein tunnels and pores in AlphaFold era

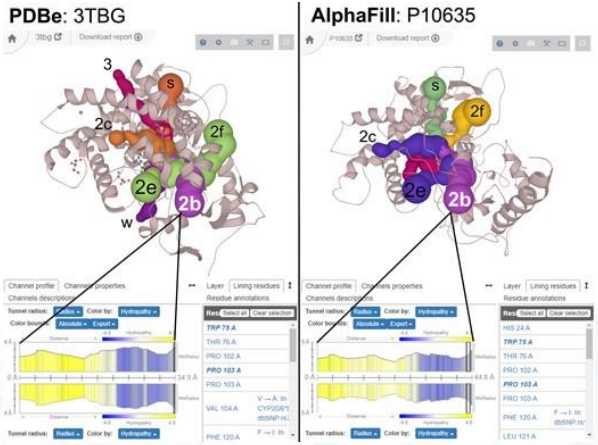
Major update of ChannelsDB database

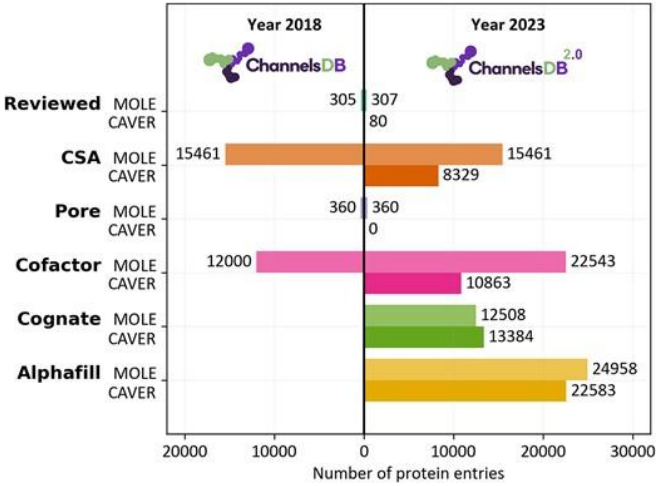
Addition of Caver on top of MOLE

Addition of AlphaFold DB -> similar tunnels

Cognate ligands

Next: Pores





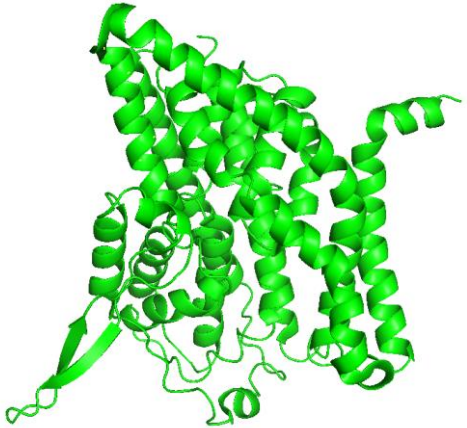
Špačková A, Vávra O, Raček T, Bazgier V, Sehnal D, Damborský J, Svobodová R, Bednář D, **Berka K** ChannelsDB 2.0: a comprehensive database of protein tunnels and pores in AlphaFold era, ***Nucleic Acids Research***, 52(D1), D413–D418, 2024, https://doi.org/10.1093/nar/gkad1012

AF on conformers

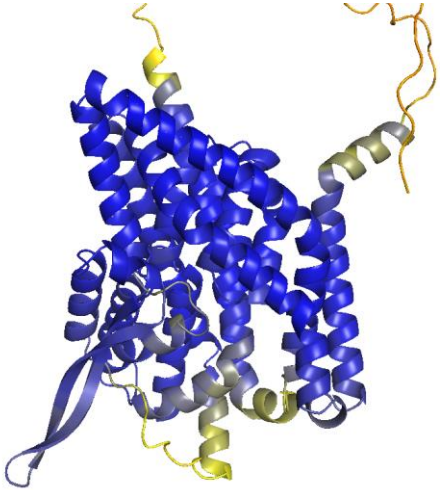
Alphafold can predict **dynamics**

pLDDT shows flexibility

SLC1A5



6mp6
Outward-Facing



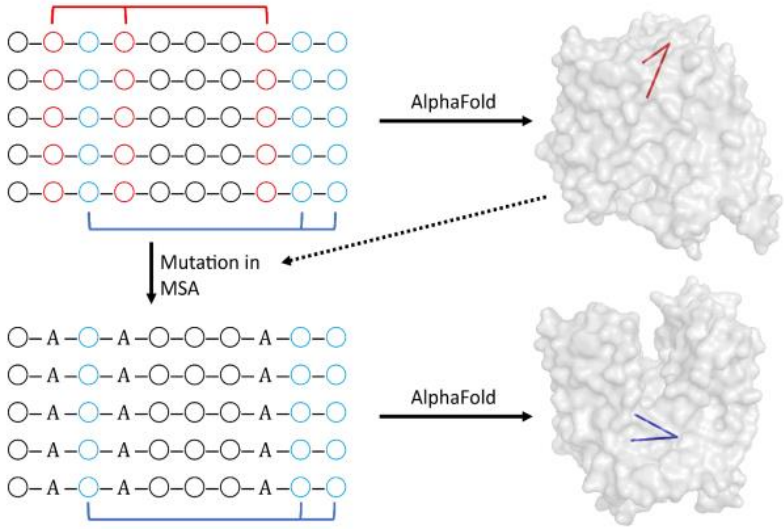
AlphaFold
similar to OF

6rvx
Inward-Facing

lower pIDDT values show **flexible regions**

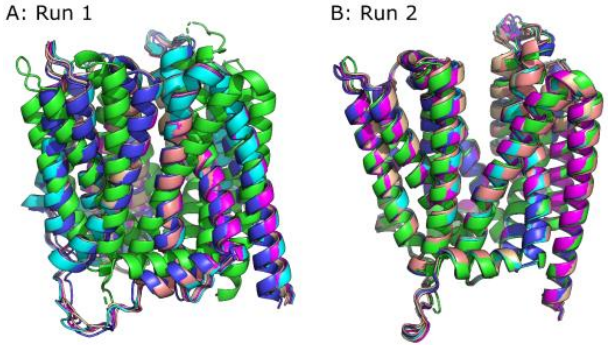
Alphafold can do **conformational changes**

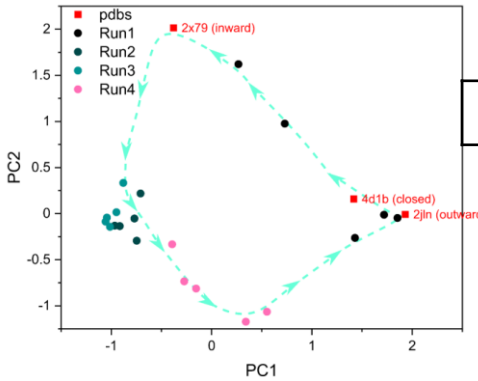
- manipulation with MSA allows selection of multiple conformers via mutation of contact points in MSA



LmrP transporter

default after mutation on interface





Mhp1

Modeling Alternate Conformations with Alphafold2 via Modification of the Multiple Sequence Alignment

Richard A. Stein, Hassane S. Mchaourab

doi: https://doi.org/10.1101/2021.11.29.470469

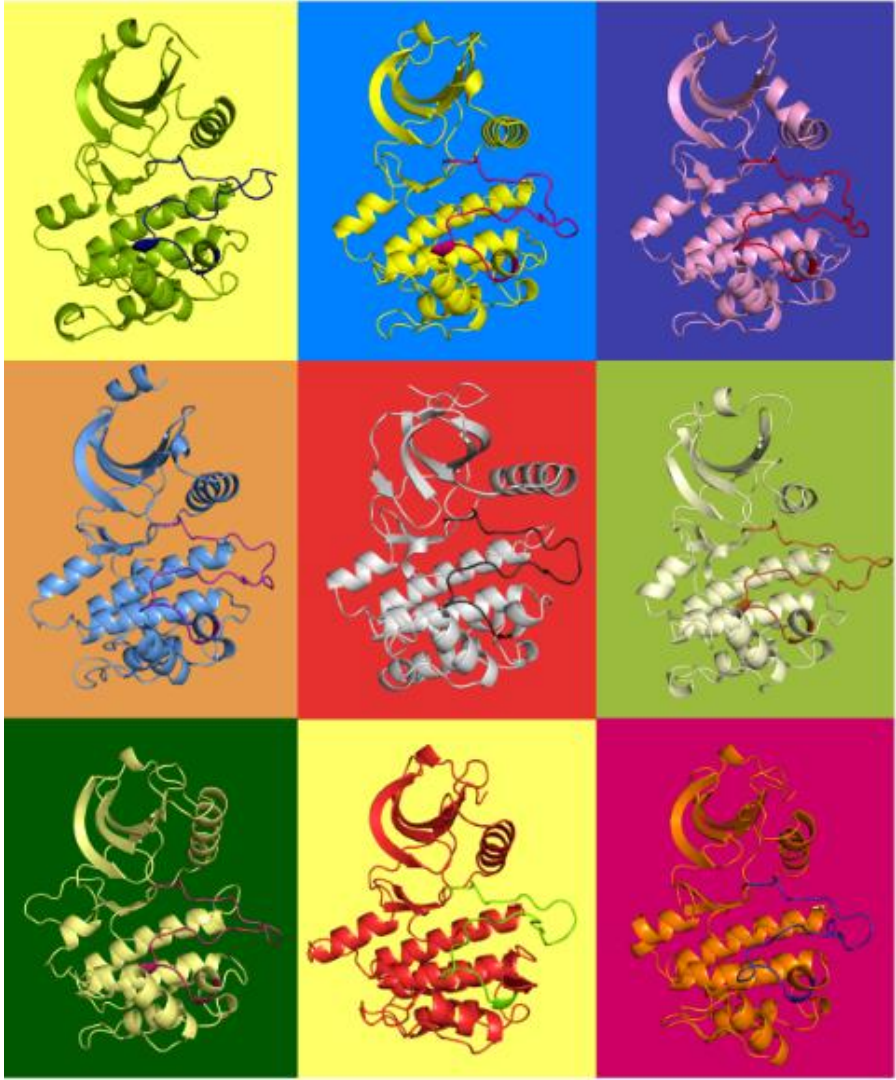
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

Kincore: AlphaFold2 models of the active form of human typical protein kinase domains

- Humans - 437 active kinases
- PDB - 268 kinases (155 actives)
- AFDB - 209 of the 437 (48%) catalytic human protein kinases have a fully active model in the EBI data set

pipeline to produce actives:

- MSA for templates in active forms (including non human kinases)
- multiple depths MSA (1-90 seqs) -> different models -> check active conformation -> combine models

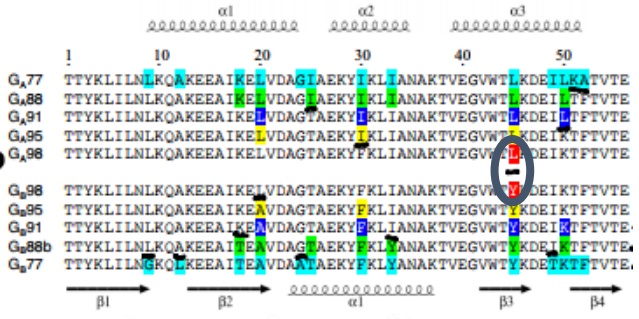


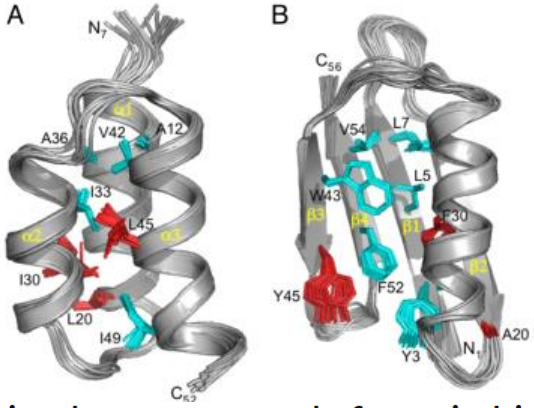
http://dunbrack.fccc.edu/kincore/activemodels

Faezov B, Dunbrack RL: AlphaFold2 models of the active form of all 437 catalytically competent human protein kinase domains. *bioRxiv* 2023.07.21.550125; https://doi.org/10.1101/2023.07.21.550125

AF on mutations

Alphafold can do **point-mutations effects**
Fold-switching proteins

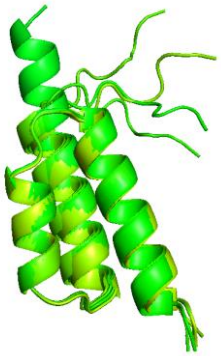




A minimal sequence code for switching protein structure and function

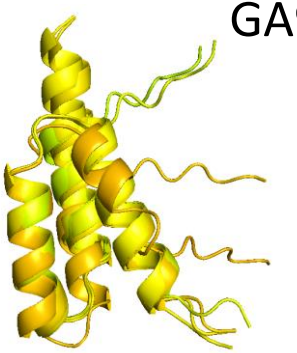
Patrick A. Alexander, Yanan He, Yihong Chen, +1, and Philip N. Bryan Authors Info & Affiliations

December 15, 2009 | 106 (50) 21149-21154 | https://doi.org/10.1073/pnas.0906408106

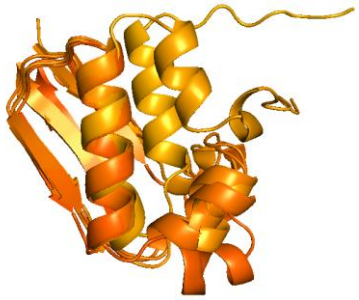


GA77

L45Y

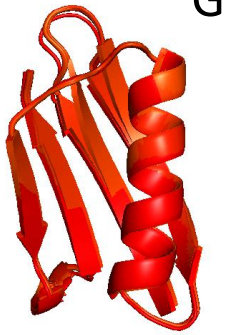


GA98



GB98

GB98 models shows mix between **3α** to **α+β**
own calculations



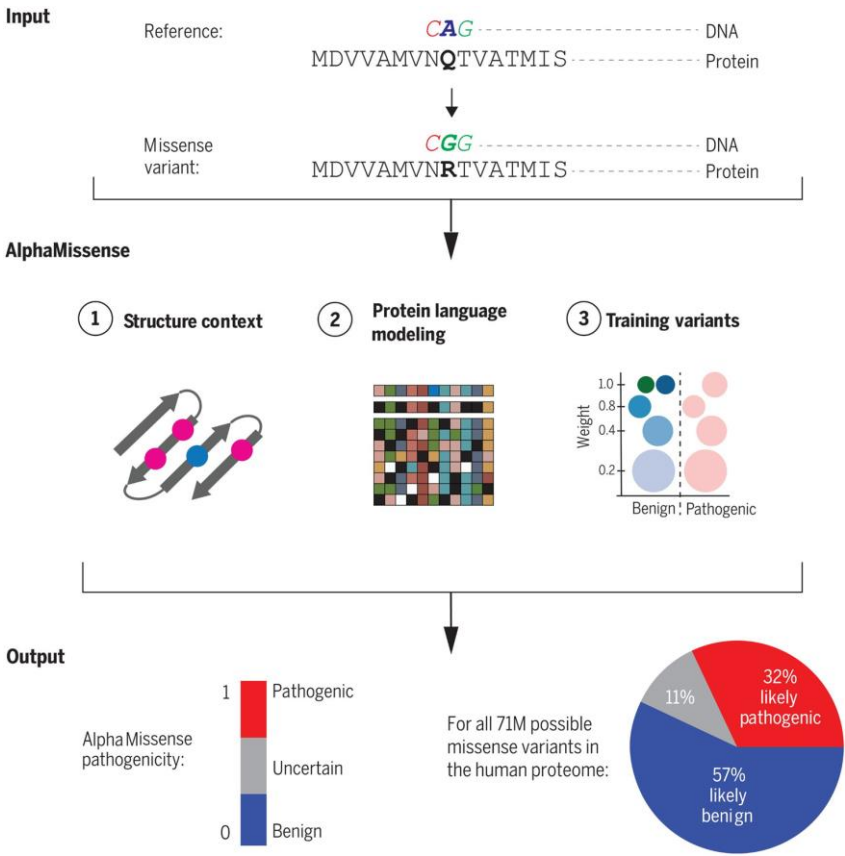
GB77

AlphaMissense

AlphaFold is not enough sensitive to mutations, but structural context for mutations is important

AlphaMissense adds protein language modelling on variants

- downloadable from Zenodo
- available in AFDB



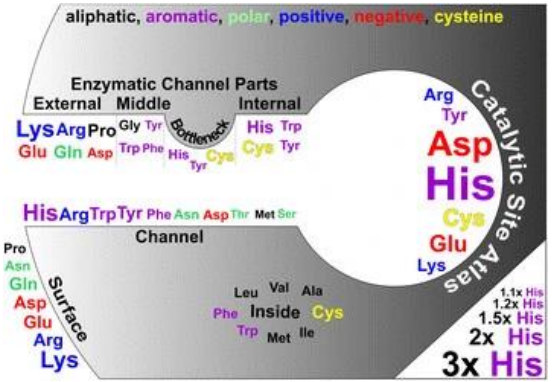
Jun Cheng *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense.*Science* **381**,eadg7492(2023).DOI:10.1126/science.adg7492

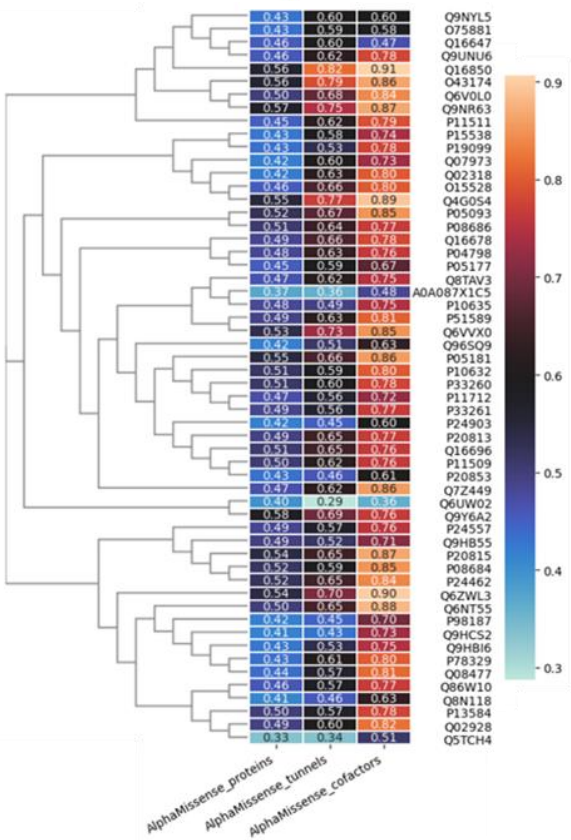
AlphaMissense in channels and TM parts show high pathogenicity

- cytochrome P450 family
- GLUT transporters

pathogenicity trend:

- **most: cofactor** environment
- **channels**
- **transmembrane** parts
- least: proteins in general





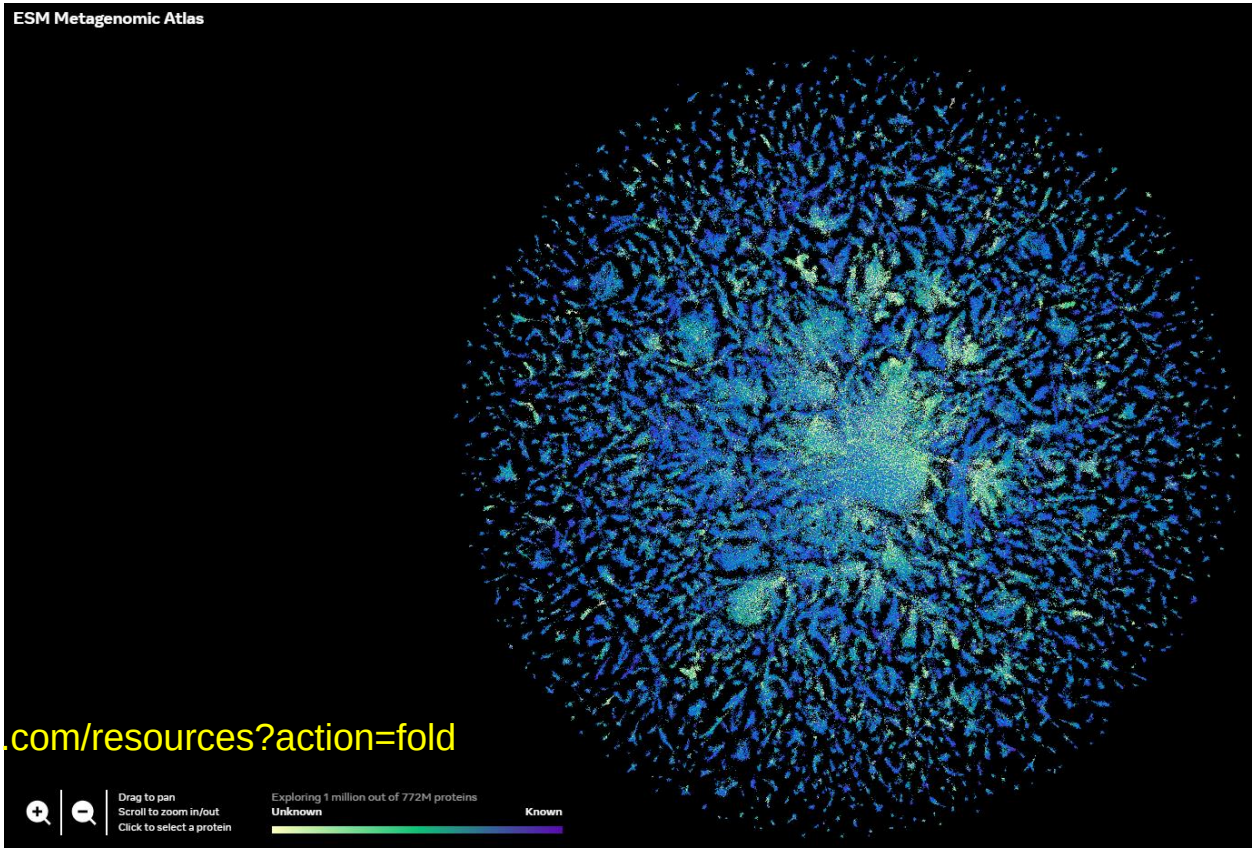
Pravda, L., **Berka, K.**, Svobodová Vařeková, R. *et al.* Anatomy of enzyme channels. *BMC Bioinformatics* 15, 379 (2014). https://doi.org/10.1186/s12859-014-0379-x

Špačková A, Kadášová N, Hutařová Vařeková I, Svobodová R, **Berka K**: Pathogenicity Patterns in Cytochrome P450 Family. *in preparation*

Kadášová N, Špačková A, Martinát D, **Berka K**: Understanding GLUT Proteins Pathogenicity: Integration of AlphaMissense, SIFT, and PolyPhen-2 Predictions, *in preparation*

AF on engineered proteins

Alphafold 2 requires MSA for start -> language models (pLM) - e.g. **ESMfold**



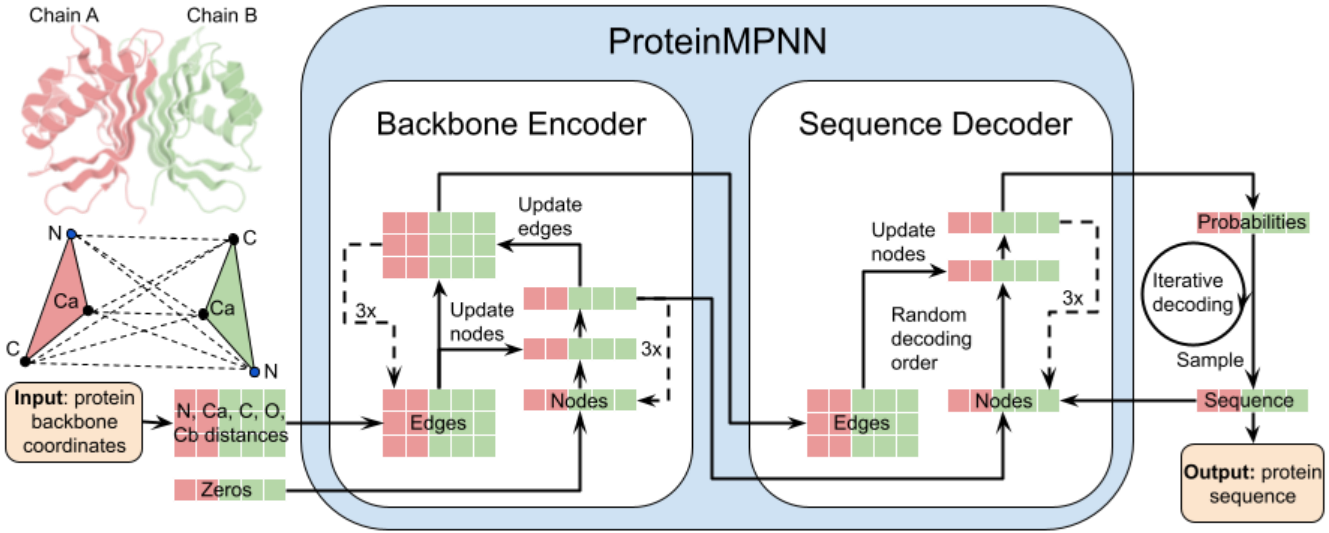
https://esmatlas.com/resources?action=fold

Zeming Lin *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science 379*, 1123-1130 (2023). DOI:10.1126/science.ade2574

Reversed prediction - ProteinMPNN

find sequence to a given structural feature

-> applicability to almost any protein sequence design problem

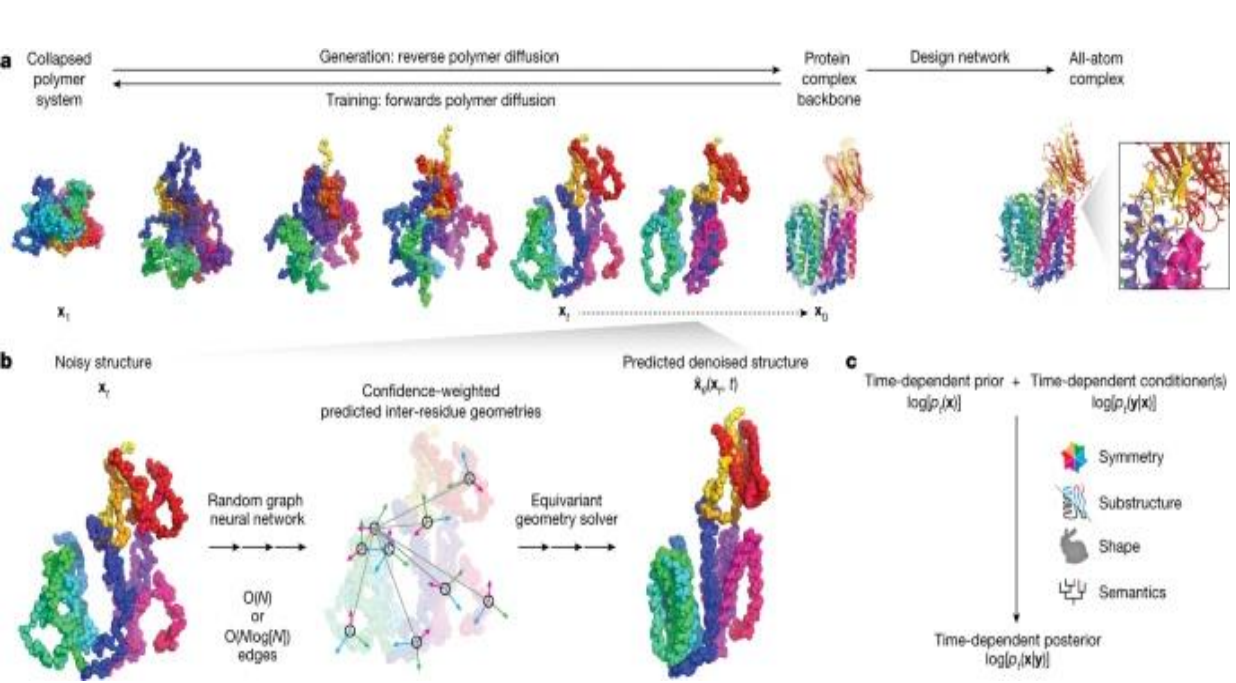


DAVID BAKER
INSTITUTE FOR PROTEIN DESIGN
UNIVERSITY OF WASHINGTON

J. Dauparas …, Baker D, *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49-56(2022).
DOI:10.1126/science.add2187

https://github.com/dauparas/ProteinMPNN

Chroma

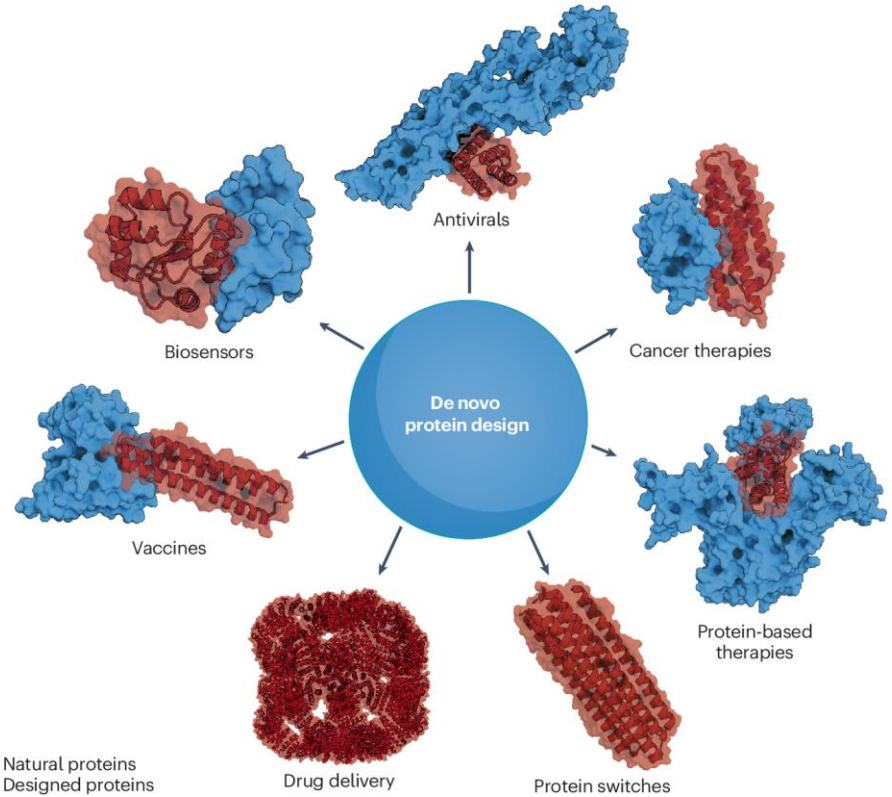


Ingraham, J.B et al *et al.* Illuminating protein space with a programmable generative model. *Nature* 623, 1070–1078 (2023). https://doi.org/10.1038/s41586-023-06728-8 https://github.com/generatebio/chroma

AF on multimers

PPI

What can be protein design used for?



Listov, D., Goverde, C.A., Correia, B.E. *et al.* Opportunities and challenges in design and optimization of protein function. *Nat Rev Mol Cell Biol* **25**, 639–653 (2024). https://doi.org/10.1038/s41580-024-00718-y

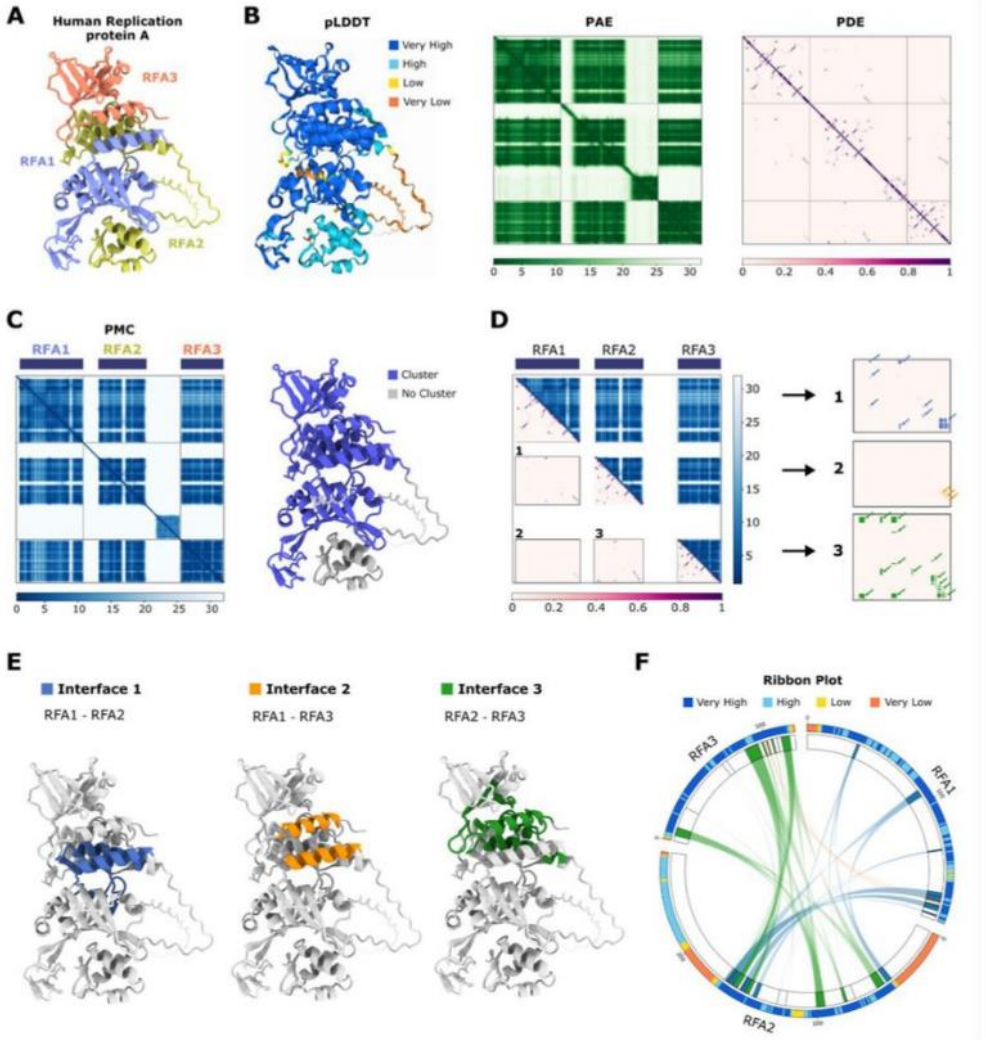
AlphaBridge: analysis of predicted macromolecular complexes

interactive platform where users can upload AlphaFold3 prediction files,

visualize the predicted 3D structures, and analyze contact interfaces through an integrated web viewer.

https://alpha-bridge.eu/

github.com/PDB-REDO/AlphaBridge



Álvarez-Salmoral D, et al. AlphaBridge: tools for the analysis of predicted macromolecular complexes. *bioRxiv* 2024.10.23.619601; doi: 10.1101/2024.10.23.619601

BindCraft

Protein–protein interactions (PPIs)

open-source and automated pipeline for *de novo* protein binder design with experimental success rates of 10-100%.

github.com/martinpacesa/BindCraft

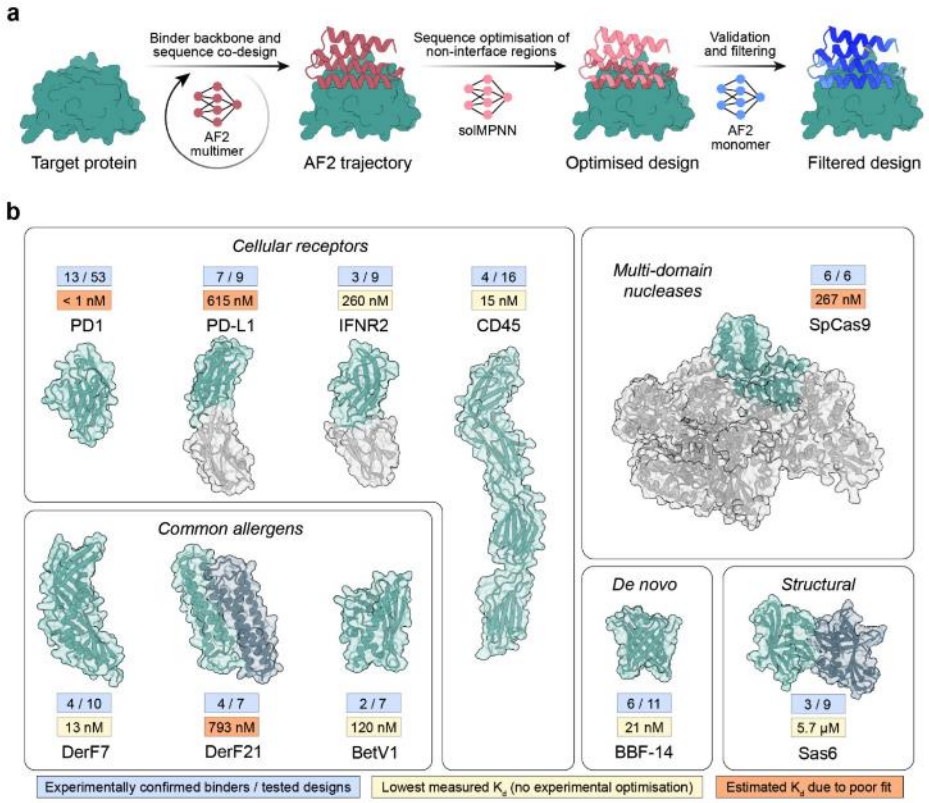
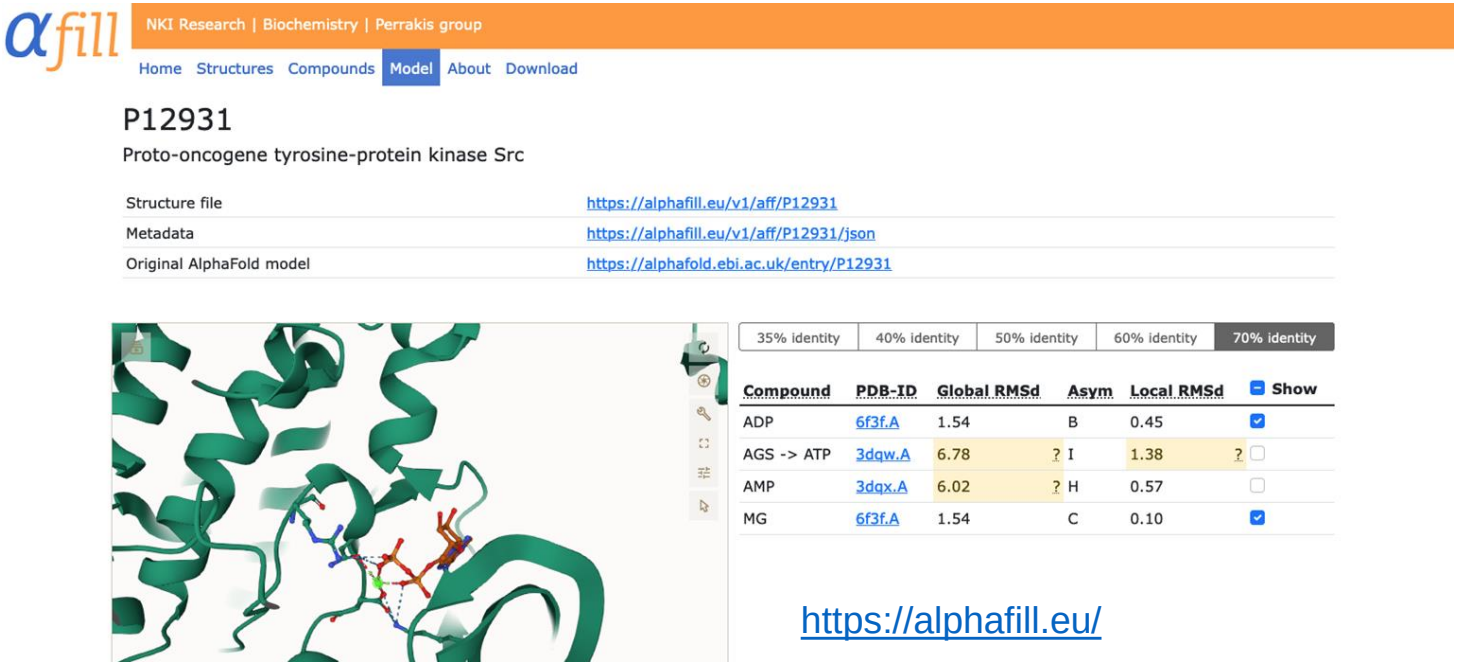


Figure 1 | *De novo* binder design using BindCraft. a, Schematic representation of the BindCraft binder design pipeline. Given a target protein structure, a binder backbone and sequence is generated using AF2 multimer, then the surface and core are optimized using $MPNN_{sol}$ while keeping the interface intact, and finally designs are filtered using the AF2 monomer model. **b,** Overview of protein targets for binder design. Parts of the model colored in green were considered during design, gray areas were excluded. Values in the blue box indicate the number of successful designs, where binding was observed on SPR measurement versus the total number of designs tested. Values in the yellow box indicate the measured K_d of the highest affinity binder without experimental sequence optimization, while values in orange boxes indicate estimated K_d values due to poor fit and will be re-measured.

Martin Pacesa, et al. BindCraft: one-shot design of functional protein binders. *bioRxiv* 2024.09.30.615802; doi:10.1101/2024.09.30.615802

AF on ligands

Alphafold can be filled with **ligands and cofactors**

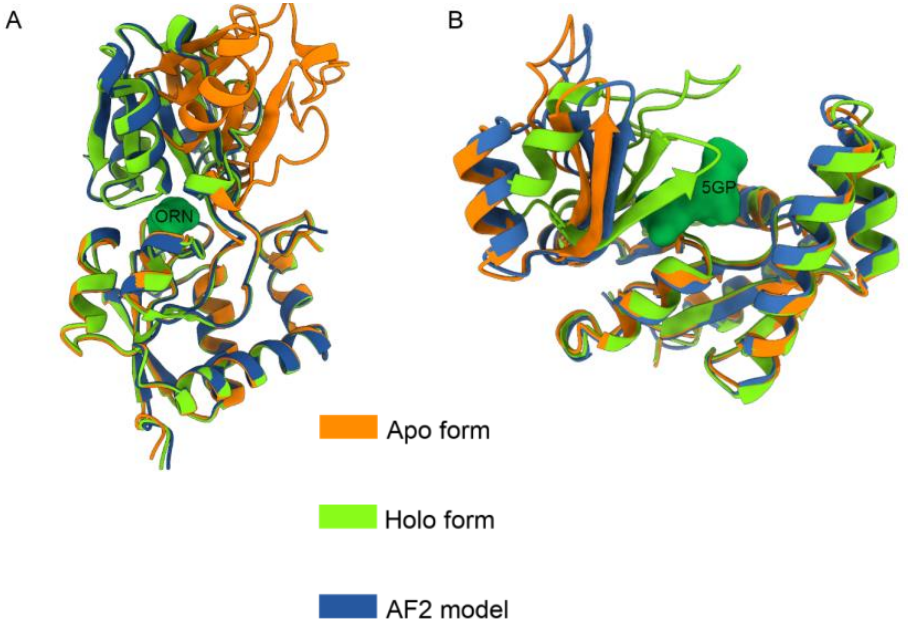


https://alphafill.eu/

Hekkelman, M.L., de Vries, I., Joosten, R.P. *et al.* AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat Methods* 20, 205–213 (2023). https://doi.org/10.1038/s41592-022-01685-y

AlphaFold models good enough for **drug design**?

- AlphaFold2 predicts **holo** protein in 70% => it can be used for drug designing
- pLDDT values in a single 3D model could be used to infer local conformational changes linked to ligand binding transitions.
- locally AlphaFold2 can be there - but it needs validation (as always)



bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

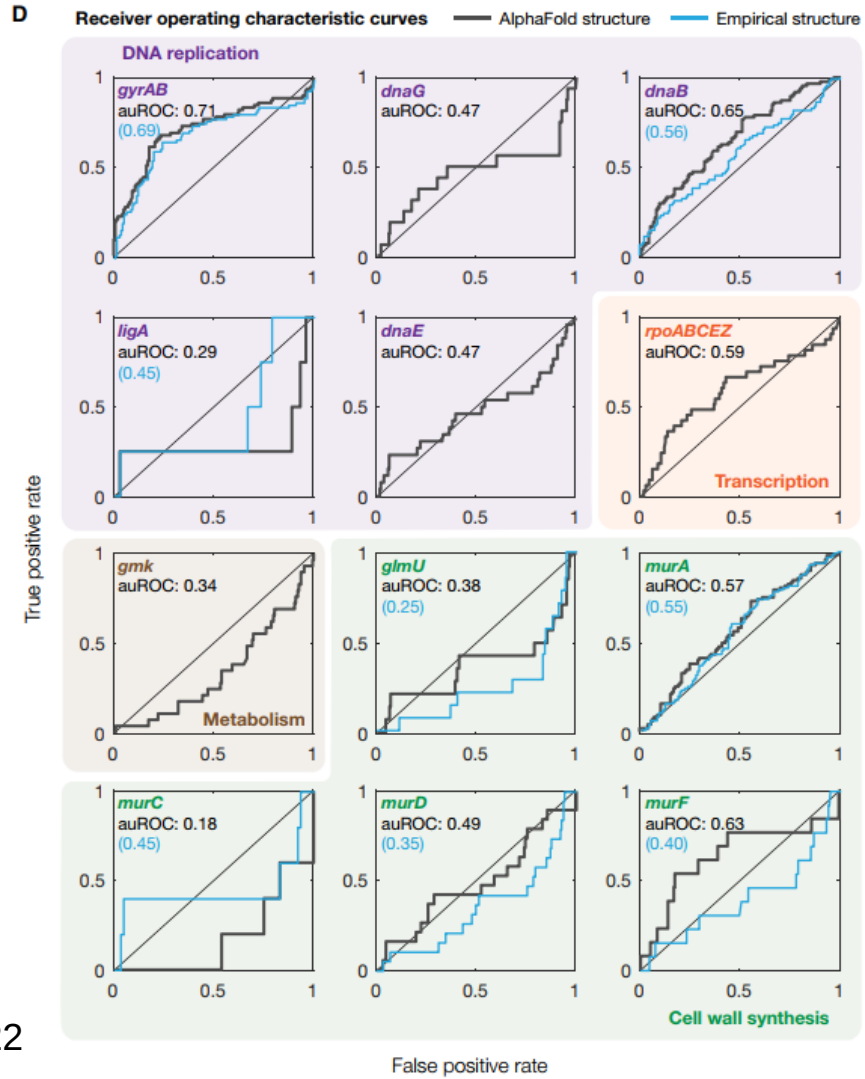
Impact of protein conformational diversity on AlphaFold predictions

Tadeo Saldaño, Nahuel Escobedo, Julia Marchetti, Diego Javier Zea, Juan Mac Donagh, Ana Julia Velez Rueda, Eduardo Gonik, Agustina García Melani, Julieta Novomisky Nechcoff, Martín N. Salas, Tomás Peters, Nicolás Demitroff, Sebastian Fernandez Alberti, Nicolas Palopoli, Maria Silvina Fornasari, Gustavo Parisi

doi: https://doi.org/10.1101/2021.10.27.466189

AlphaFold docking antibiotics example

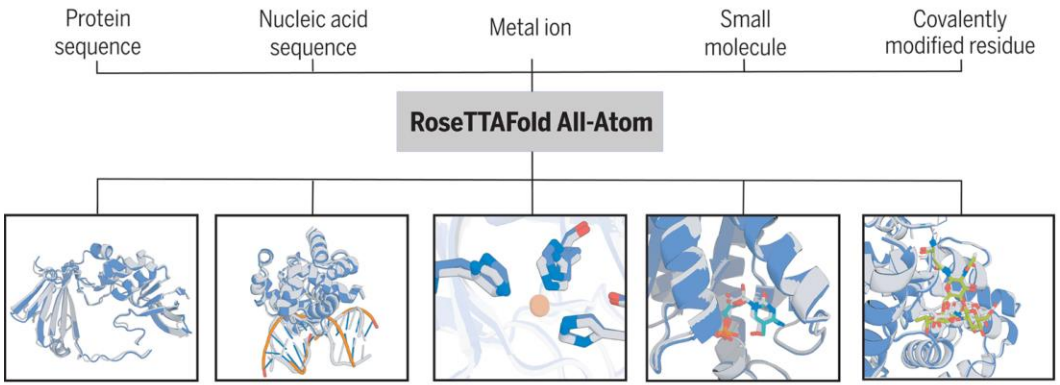
- benchmarking docking by metabolic activity of 12 essential proteins
- auROC = 0.48 (**Vina** on AF2)
- rescoring -> auROC 0.63
- auROC = 0.46 (**Vina** on experimental structures)
- **both bad** (auROC random is 0.5)

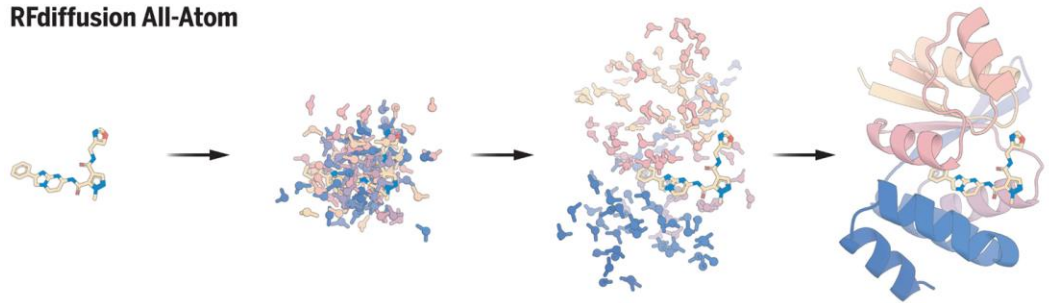


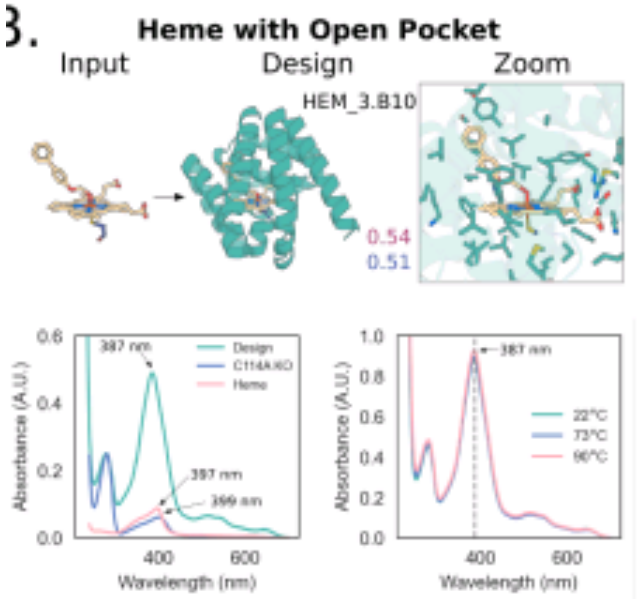
Felix Wong et al. Molecular Systems Biology 18: e11081 | 2022

RoseTTAFold All-Atom

building protein around ligand

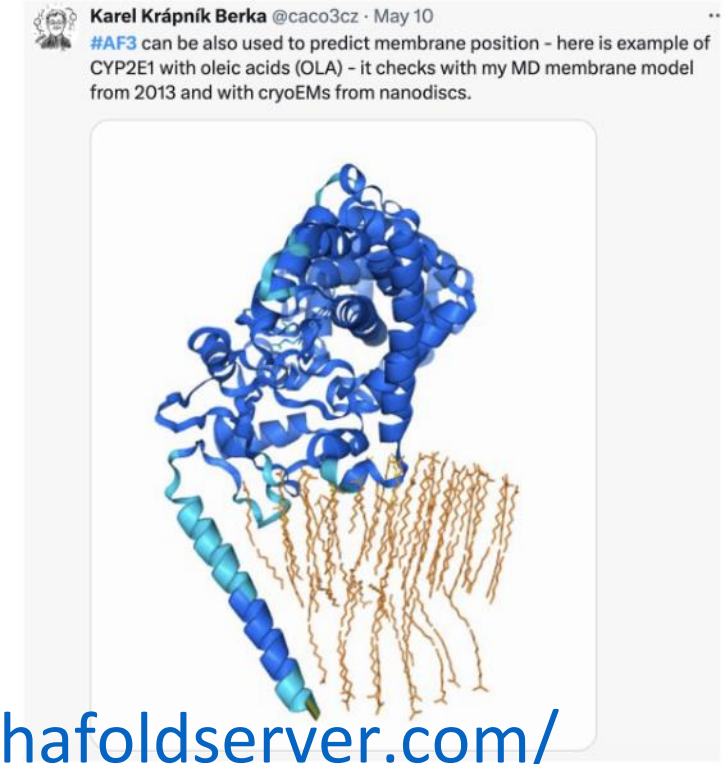


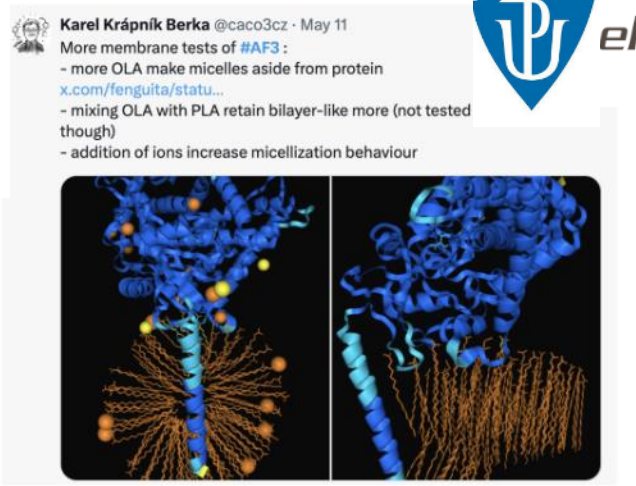


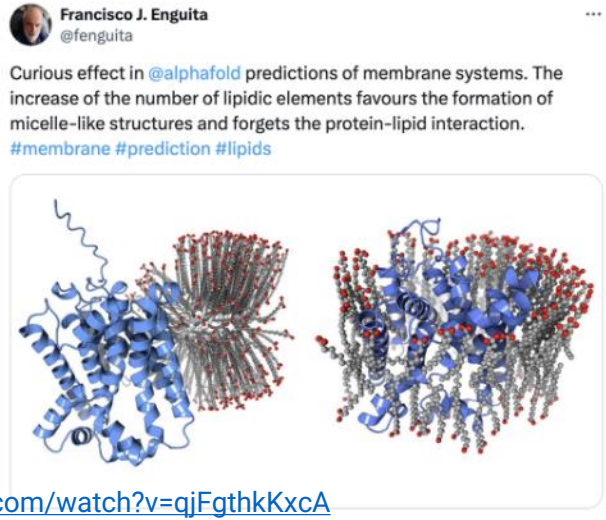


Rohith Krishna *et al.* Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**,eadl2528(2024).
DOI:10.1126/science.adl2528

Berka's hack – AF3 w lipids







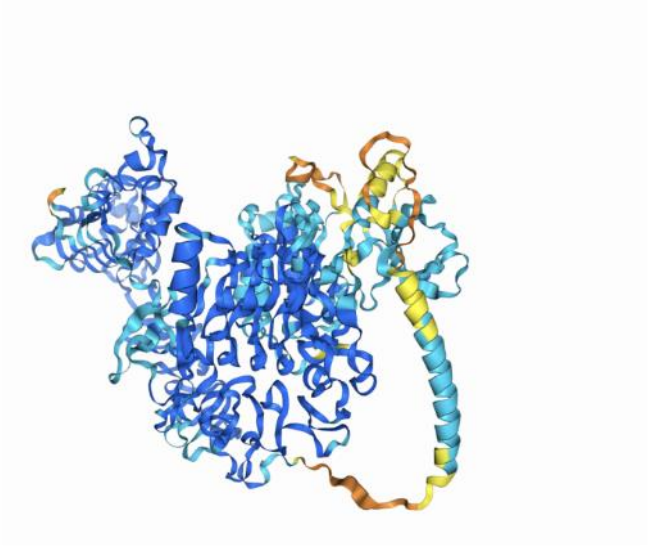
https://alphafoldserver.com/

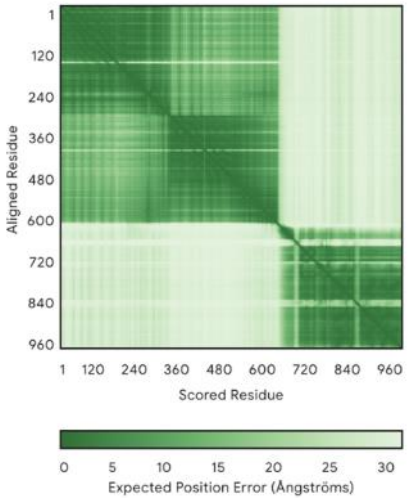
Sergey Ovchinnikov - Boston Protein Design and Modeling Club - https://www.youtube.com/watch?v=qjFgthkKxcA

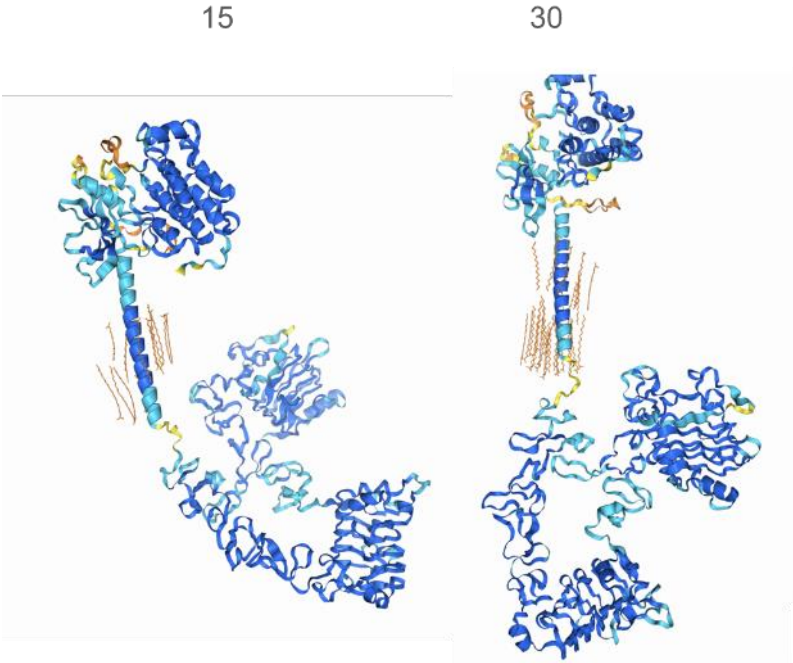
AF3 + lipids

In AF2 intracellular and extracellular domain touched (PAE was ok)

-> lipids in AF3 separates them

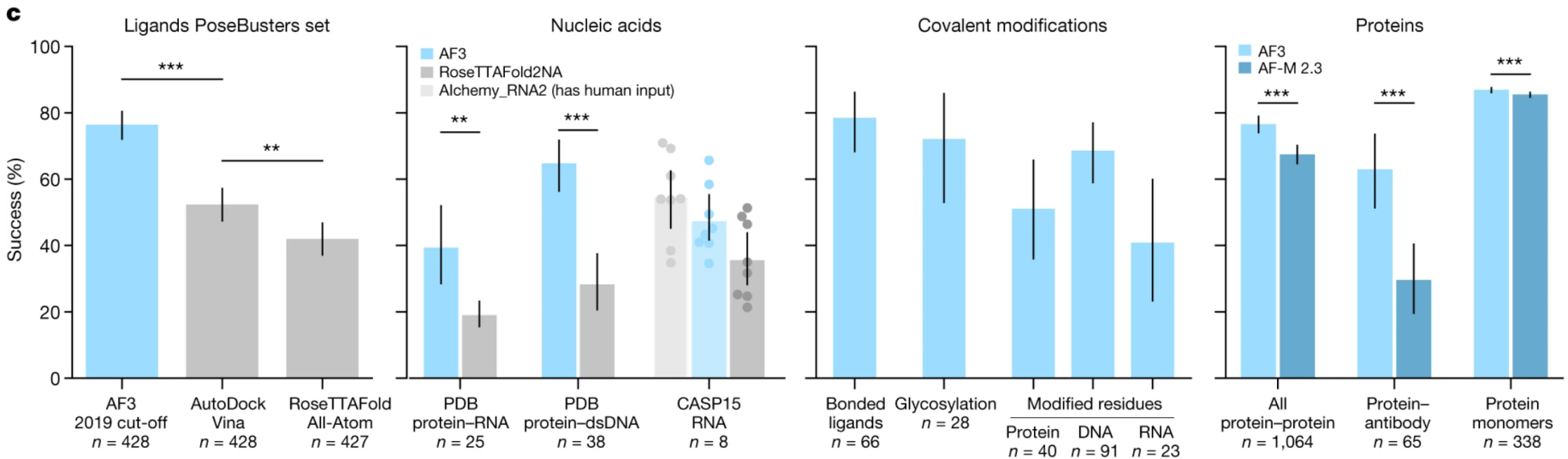






Sergey Ovchinnikov - Boston Protein Design and Modeling Club - https://www.youtube.com/watch?v=qjFgthkKxcA

AF3 performance



https://blog.booleanbiotech.com/alphafold3-boltz-chai1

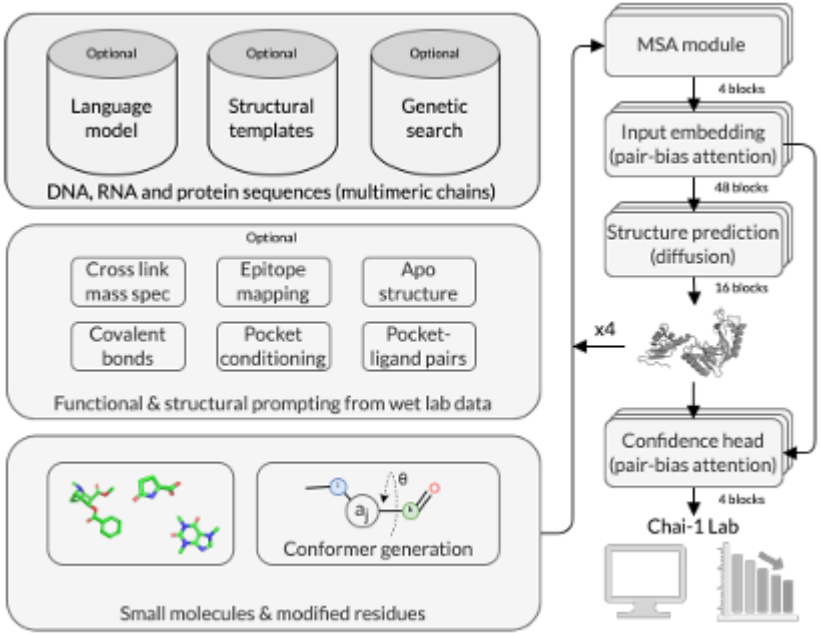
AF3 variant

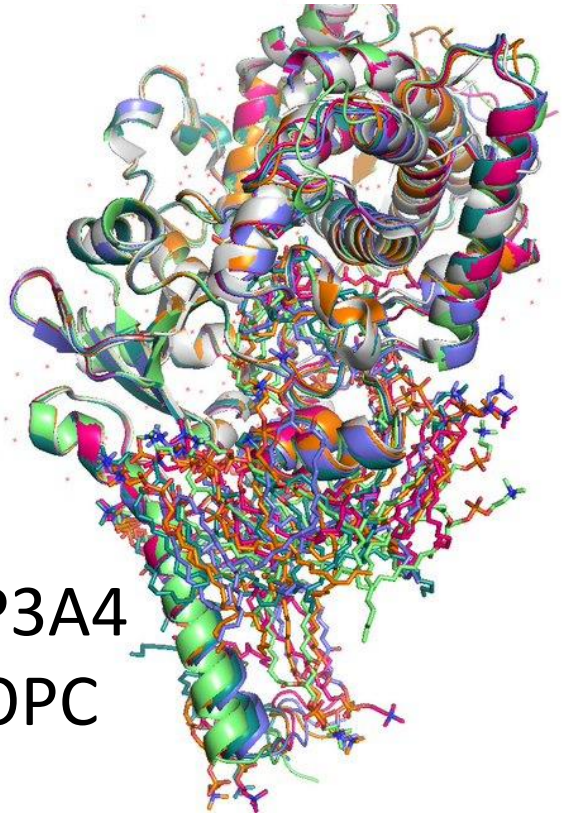
Date	Software	Code available?	Parameters available?	Lines of Python code
2024-05	Alphafold 3	✗ (CC-BY-NC-SA 4.0)	✗ (you must request access)	32k
2024-08	HelixFold3	✗ (CC-BY-NC-SA 4.0)	✗ (CC-BY-NC-SA 4.0)	17k
2024-10	Chai-1	✗ (Apache 2.0, inference only)	✓ (Apache 2.0)	10k
2024-11	Protenix	✗ (CC-BY-NC-SA 4.0)	✗ (CC-BY-NC-SA 4.0)	36k
2024-11	Boltz	✓ (MIT)	✓ (MIT)	17k

https://blog.booleanbiotech.com/alphafold3-boltz-chai1

- Chai-1
  - https://github.com/chaidiscovery/chai-lab
- Boltz-1
  - https://github.com/jwohlwend/boltz
- Protenix
  - https://github.com/bytedance/Protenix
- HelixFold3
  - https://github.com/PaddlePaddle/
- AlphaFold 3
  - https://github.com/google-deepmind/alphafold3

Chai-1 - open version of AF3



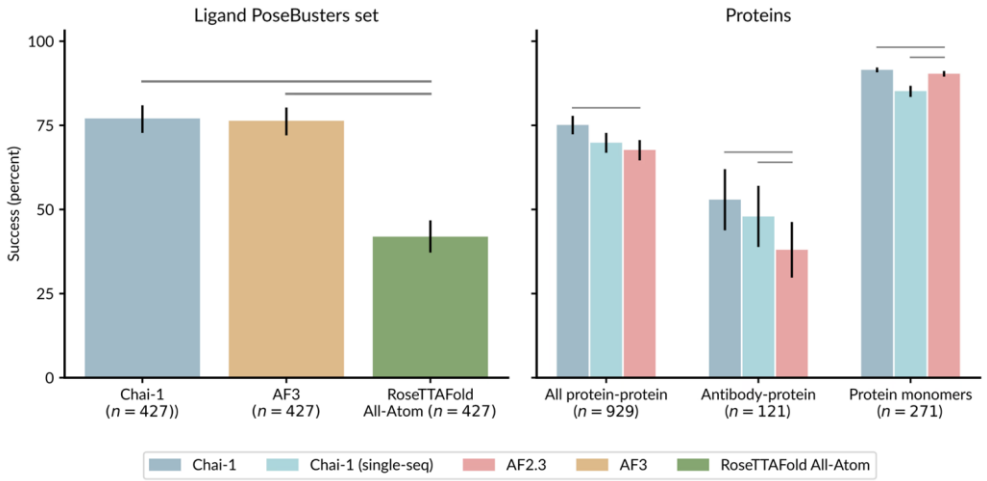


CYP3A4
+POPC

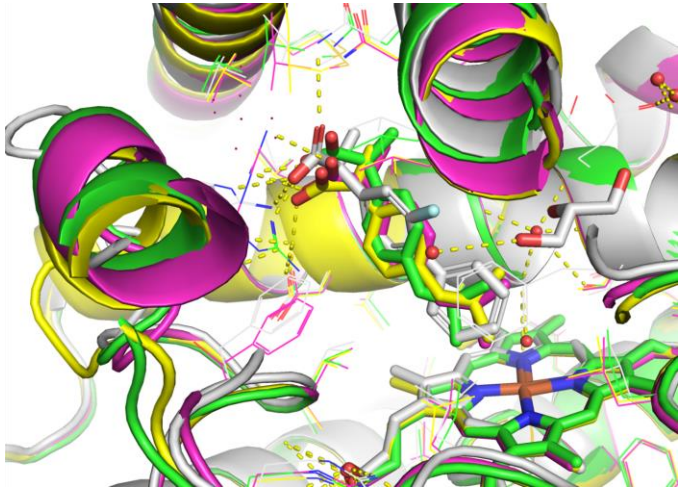
lab.chaidiscovery.com, github.com/chaidiscovery/chai-lab

Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, Kevin Wu: Chai-1: Decoding the molecular interactions of life, **bioRxiv** 2024.10.10.615955; https://doi.org/10.1101/2024.10.10.615955

Chai-1 - “docking”

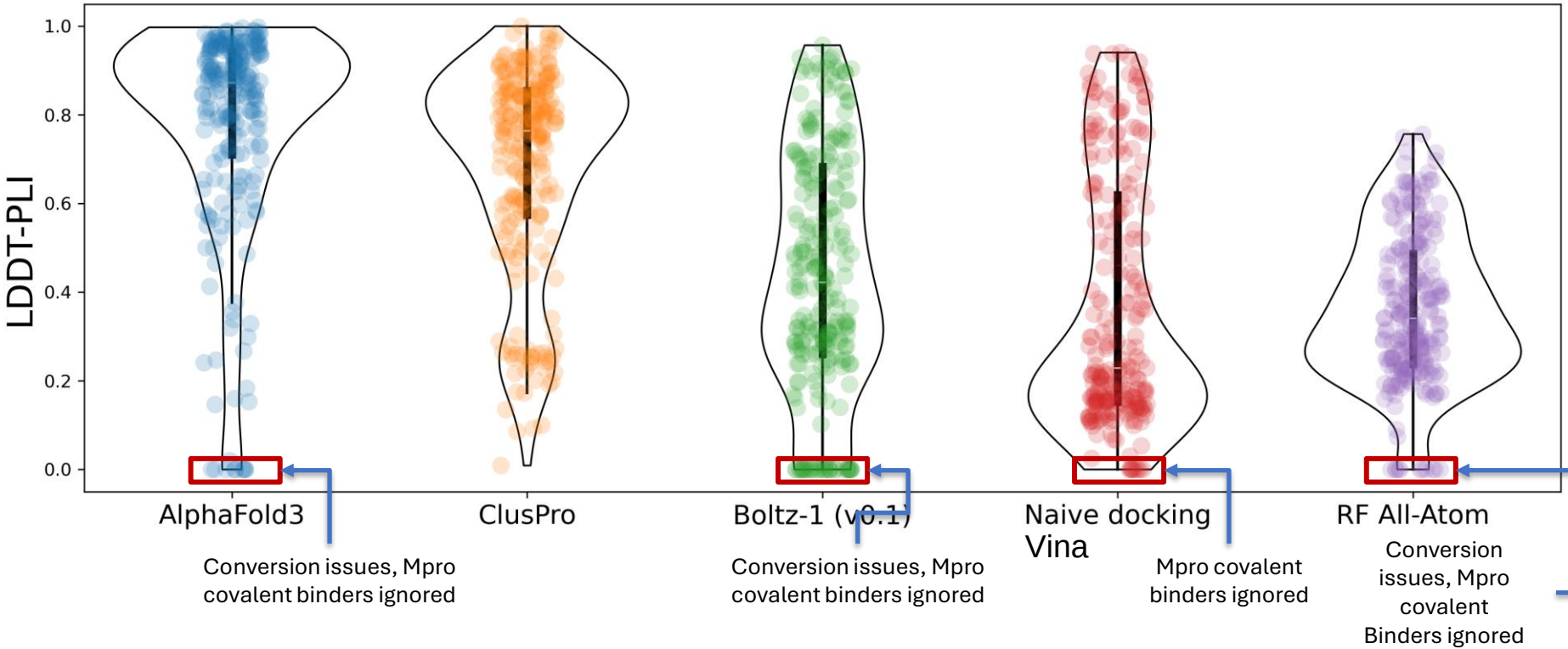


ibuprofen to CYP2C9 - copy flurbiprofen pose



Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, Kevin Wu: Chai-1: Decoding the molecular interactions of life, **bioRxiv** 2024.10.10.615955; https://doi.org/10.1101/2024.10.10.615955

Overall LDDT-PLI performances per docking method

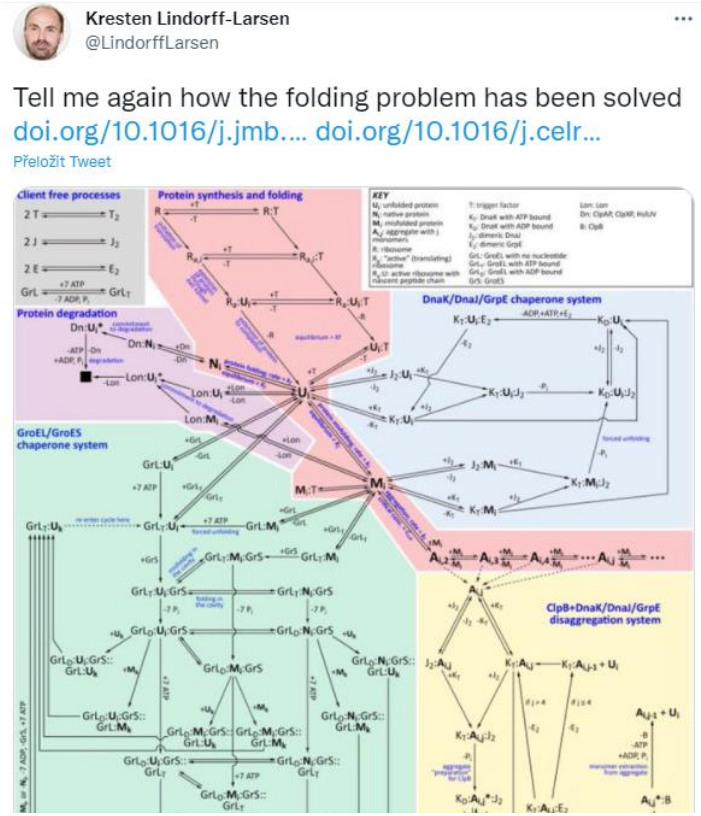


IT DOES NOT MEAN THAT IT IS GENERALIZABLE!

https://predictioncenter.org/casp16/doc/presentations/Day-3/Day3-09-Ligand_Assessment_PuntaCanaCASP-Gilson-1f.pptx

Summary

- Alphafold2 made a huge leap in **prediction accuracy** enabling to use protein structural models due to **quality predictor**
- Role of **open science and publicly available data** in **PDB** can not be overstated
- **CASP competition** was a driver of the change
- Alphafold is **publicly available** and can be run from many places including ELIXIR CZ
- Alphafold has **inspired many "Alphafoldology" tools and uses** already and this space **flourish with innovation**
- **Openness helps to quick innovation cycle**



Thank you
for your attention

Questions, please?

BioLists

github.com/biolists

📖 Table of contents

- Predictors
- Tools and Extensions
- Databases and Datasets
- Webservers
- Discontinued

abeebyekeen.com/biomodes-biomolecular-structure-prediction/

BioMoDes: A Repository of Tools for Biomolecular Modeling and Design

1. Biomolecular Structure Prediction Tools
2. Biomolecular Design Tools
3. Biomolecular Property Prediction and Analysis Tools
4. Protein Search, Alignment, & DB Management Tools
5. Small-molecule Design and Prediction Tools

Are structural biologists and bioinformaticians on the job market?

- Alphafold does not tell much about **folding process**
- Alphafold can not do **point mutations** - design of functions
- Alphafold is not usable for **drug design**
- Alphafold can not do **conformational changes** or **dynamics**
- Alphafold can not do **multiprotein complexes –** interactions
- Alphafold can not do effects of **post-translational protein modifications**
- Alphafold can not do **ligand effects**
- Alphafold is not good with **orphan sequences**
- **or is it?**

elixir CZECH REPUBLIC

Extra slides

Alphafold Decoded

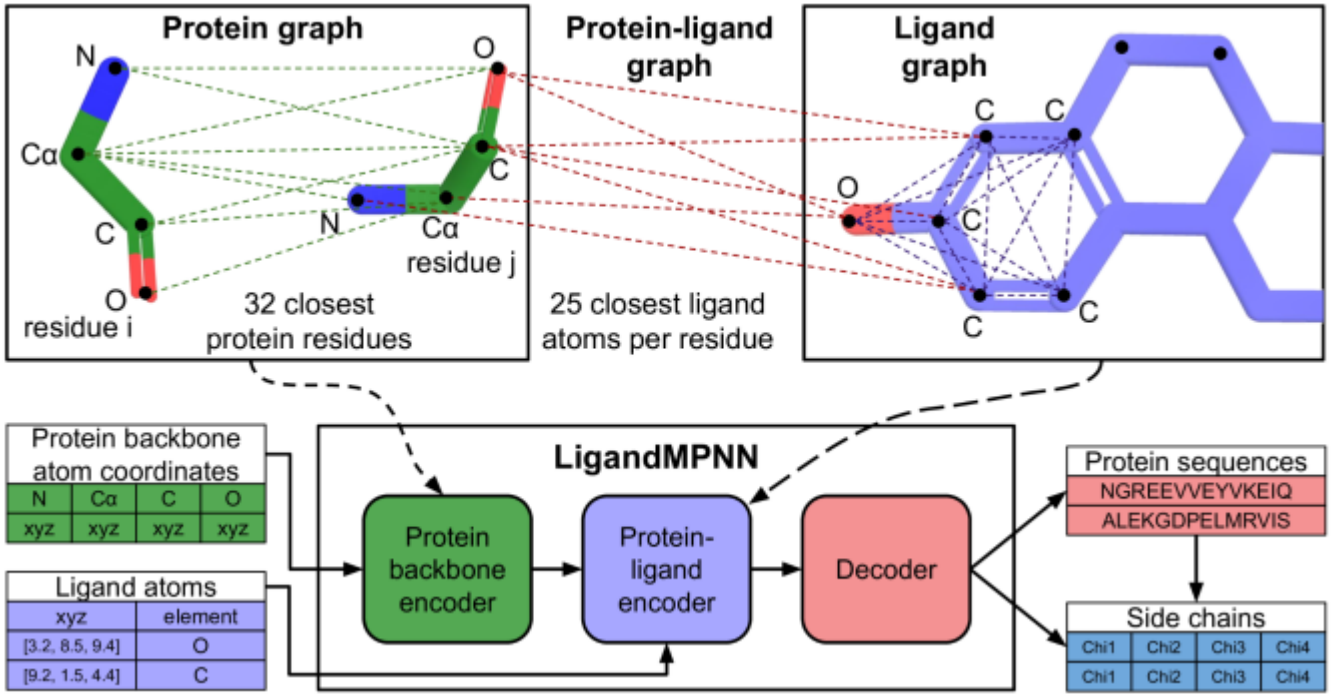
https://www.alphafold-decoded.com/

https://www.youtube.com/watch?v=7dS3nyEcOyE&list=PLJ0WcPQS7xJVJr6ceIPFSkAGAgrkmw1c9

Understand and implement cutting-edge AI for protein structure prediction. No prior knowledge assumed.

LigandMPNN

deep learning-based protein sequence design method that explicitly models all non-protein components of biomolecular systems

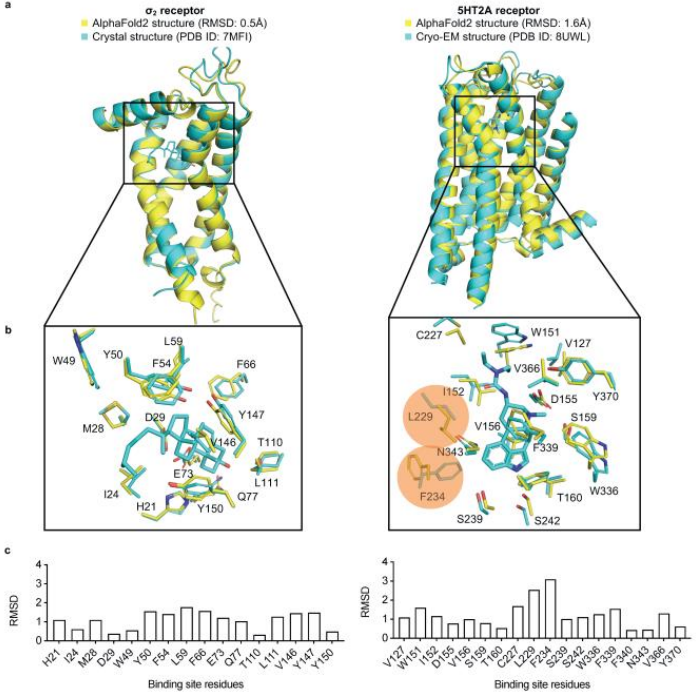


bioRxiv preprint https://doi.org/10.1101/2023.12.22.573103 https://github.com/dauparas/LigandMPNN

AlphaFold2 structures template ligand discovery

prospective screen

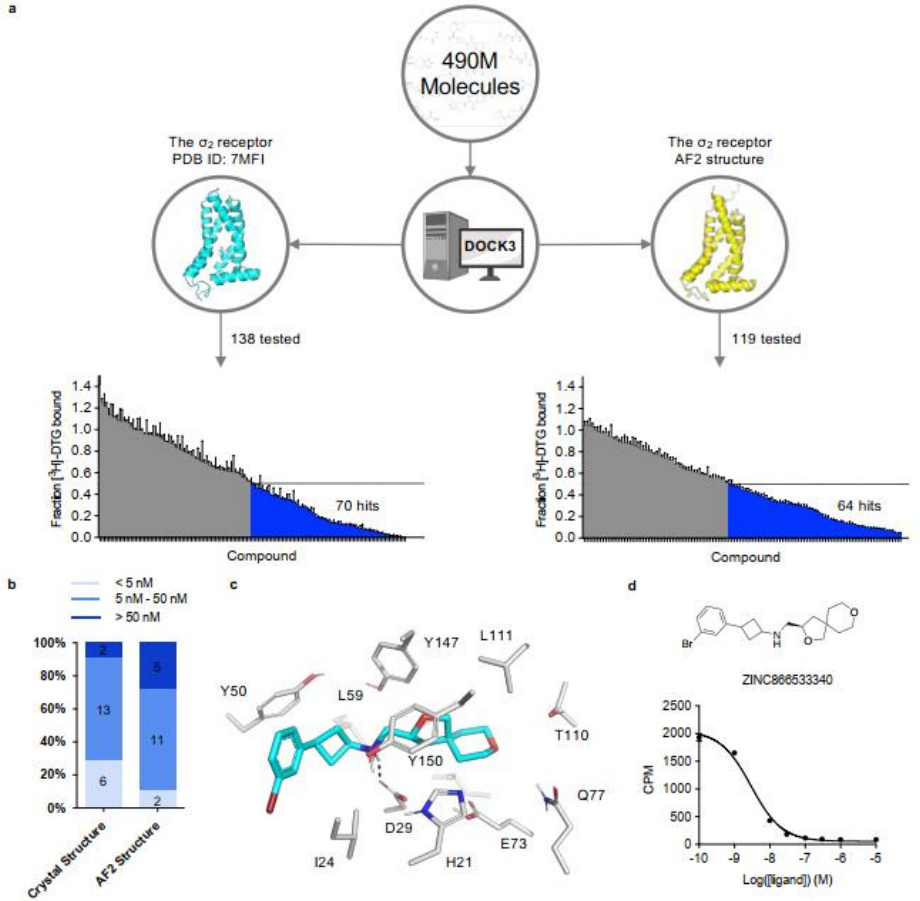
different binding site conformation



average Tc of 0.32, not far from random for this fingerprint. Consistent with the diversit

the most potent ligand from the AF2 campaign, ZINC866533340 (Ki 1.6 nM), represente

a chemotype previously unseen for the σ_2 receptor (**Fig 2c** and **2d**).



bioRxiv preprint doi: https://doi.org/10.1101/2023.12.20.572662

Alphafold can describe **folding process** to some level
Was Anfinsen right?

bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

New Results

Follow this preprint

State-of-the-Art Estimation of Protein Model Accuracy using AlphaFold

James P. Roney, Sergey Ovchinnikov

doi: https://doi.org/10.1101/2022.03.11.484043

This article is a preprint and has not been certified by peer review [what does this mean?].

Previous

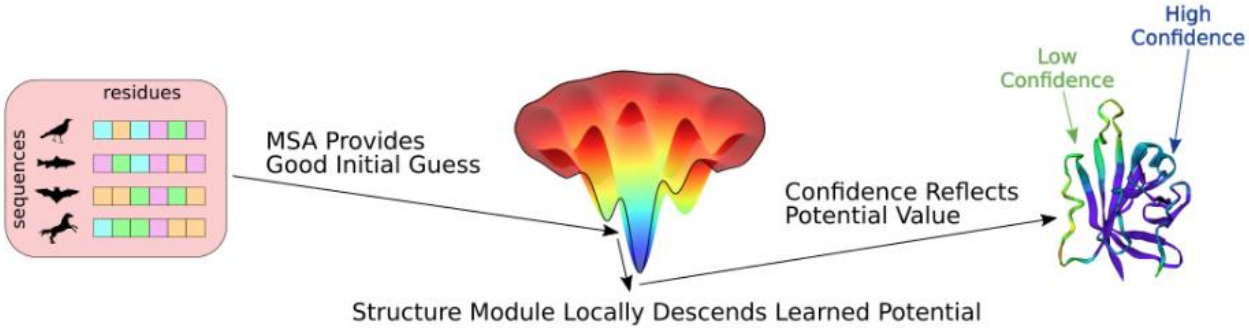
Posted March 24, 2022.

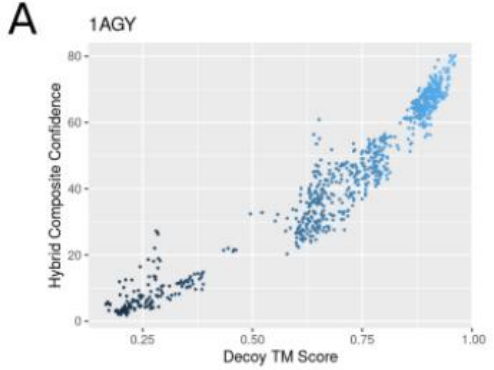
Download PDF

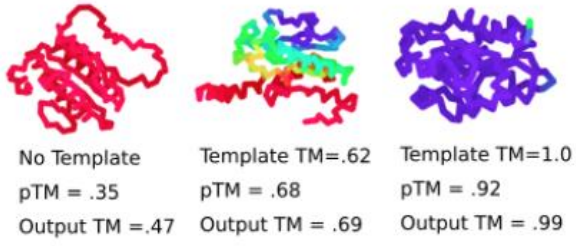
Print/Save Options

Data/Code

Revision Summary







Published September 19, 2023 | Version v2 Journal article Open

Predictions for AlphaMissense

Jun Cheng[1]; Guido Novati[1]; Joshua Pan[1]; Clare Bycroft[1]; Akvilė Žemgulytė[1]; Taylor Applebaum[1]; Alexander Pritzel[1]; Lai Hong Wong[1]; Michal Zielinski[1]; Tobias Sargeant[1]; Rosalia G. Schneider[1]; Andrew W. Senior[1]; John Jumper[1]; Demis Hassabis[1]; Pushmeet Kohli[1]; Žiga Avsec[1]

Show affiliations

Variant Effect Predictor Plugins

VEP can use plugin modules written in Perl to **add functionality** to the software.

Plugins are a powerful way to **extend**, **filter** and **manipulate** the VEP output.
They can be installed using VEP's installer script, run the following command to get a list of available plugins:

```
perl INSTALL.pl -a p -g list
```

Alternatively, VEP plugins and their dependencies are available in the Docker image. Read how to use Ensembl VEP in Docker and Singularity.

Some plugins are also available to use via the VEP web and REST interfaces.

Existing plugins

We have written several plugins that implement experimental functionalities that we do not (yet) include in the variation API, and these are stored in a public github repository:

https://github.com/Ensembl/VEP_plugins

Here is the list of the VEP plugins available:

Select categories: All categories

Plugin	Description	Category	External libraries	Developer
AlphaMissense	This plugin for the Ensembl Variant Effect Predictor (VEP) annotates missense variants with the pre-computed AlphaMissense pathogenicity scores. AlphaMissense is a deep learning model developed by	Pathogenicity predictions	-	Ensembl

From: Exome sequencing and analysis of 454,787 UK Biobank participants

Variant category	No. of variants (% with MAC=1)	Median number of variants per participant (IQR)
Coding regions[a]	12,326,144 (46.86)	19,895 (247)
Predicted function		
In-frame indels	75,096 (40.33)	115 (11)
Synonymous	3,457,173 (43.12)	10,273 (141)
Missense	7,878,586 (47.28)	9,292 (143)
Likely benign	1,532,129 (44.11)	6,561 (104)
Possibly deleterious	4,556,629 (47.23)	2,610 (70)
Likely deleterious	1,789,828 (50.1)	121 (16)
pLOF (any transcript)	915,289 (57.88)	214 (16)

Upload JSON Clear

Molecule type: Protein | Copies: 1

```
MGSNKSKPKD ASQRRRSLEP AENVHGAGGG AFPASQTPSK PASADGHRGP SAAFAPAAAE
PKLFGGFNSS DTVTSPQRAG PLAGGVTTFV ALYDYESRTE TDLSFKKGER LQIVNNTEGD
WWLAHSLSTG QTGYIPSNYV APSDSIQAEE WYFGKITRRE SERLLLNAEN PRGTFLVRES
ETTKGAYCLS VSDFDNAKGL NVKHYKIRKL DSGGFYITSR TQFNSLQQLV AYYSKHADGL
CHRLTTVCPT SKPQTQGLAK DAWEIPRESL RLEVKLGQGC FGEVWMGTWN GTTRVAIKTL
KPGTMSPEAF LQEAQVMKKL RHEKLVQLYA VVSEEPIYIV TEYMSKGSLL DFLKGETGKY
LRLPQLVDMA AQIASGMAYV ERMNYVHRDL RAANILVGEN LVCKVADFGL ARLIEDNEYT
ARQGAKFPIK WTAPEAALYG RFTIKSDVWS FGILLTELTT KGRVPYPGMV NREVLDQVER
GYRMPCPPEC PESLHDLMCQ CWRKEPEERP TFEYLQAFLE DYFTSTEPQY QPGENL
```

PTMs 419Y: O-Phospho-L-tyrosine

Molecule type: Ligand | Copies: 1

- ADP – Adenosine disphosphate
- ATP – Adenosine triphosphate
- AMP – Adenosine phosphate
- GTP – Guanosine-5'-triphosphate
- GDP – Guanosine-5'-diphosphate
- FAD – Flavin-adenine dinucleotide

+ Add entity

Save job

History ✓ Completed ✓ Saved

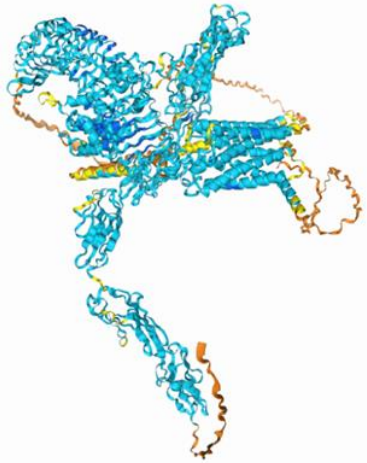
Name	Modified
2024-05-13_15:30	2024-05-13 15:35
2024-05-09_14:36	2024-05-09 14:42
2024-05-08_21:48	2024-05-08 22:03
2024-05-08_21:52	2024-05-08 22:00

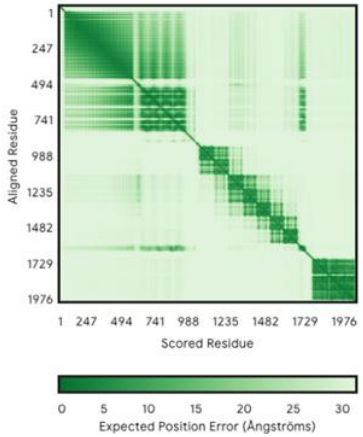
2024-05-08_21:48

← Back ⬇ Download Clone and reuse Feedback on structure

Very high (plDDT > 90) | Confident (90 > plDDT > 70) | Low (70 > plDDT > 50) | Very low (plDDT < 50)

ipTM = 0.39 pTM = 0.42 learn more

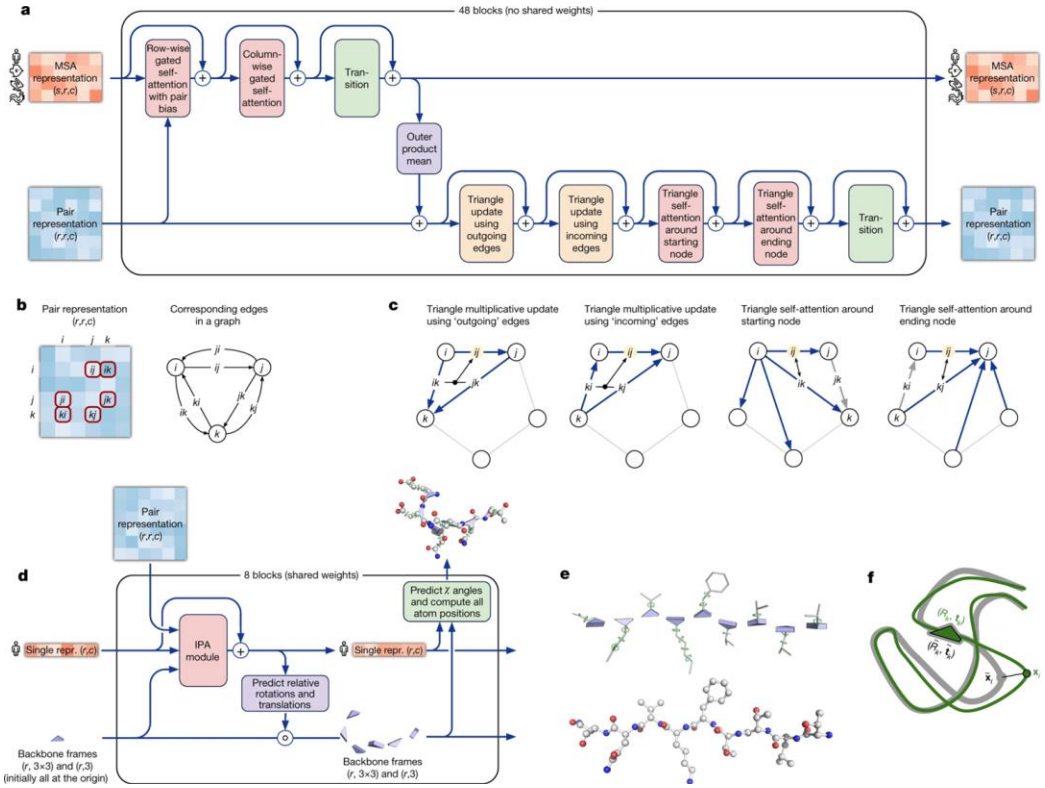




Information

Type	Copies	Sequence
Protein	1	MDTSRLGVLL SLPVLLQLAT GGSSPRSGVL LRGCPTHCHC EPDGRMLLRV DCSDLGLSEL PSNLSVFTSY LDLSMNNISQ LLPNPLPSLR FLEELRLAGN ALTYIPKGAF TGLYSLKVLM LQNNQLRHVP TEALQNLRSL QSLRLDANHI SYVPPSCFSG LHSLRHLWLD DNALTEIPVQ AFRSLSALQA MTLALNKIHH IPDYAFGNLS SLVVLHLHNN RIHSLGKKCF DGLHSLETLD

Architectural details AF2

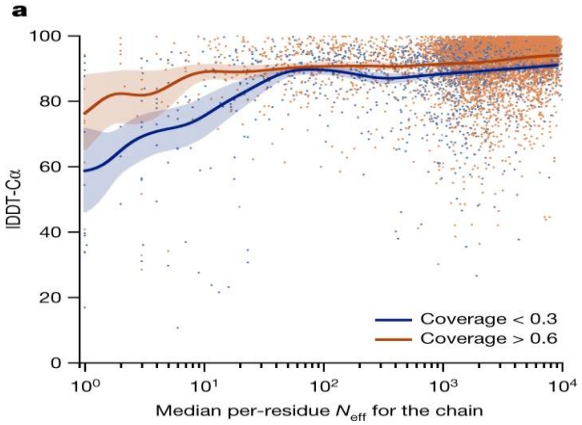


MSA - multiple sequence alignment

using standard tools - jackhmmer, HHBlits

- sequence DBs:
  - *UniRef90*
  - *UniClust30* = for sequence self-distilation
- metagenomicsDBs - to fully cover classes underepresented in UniRef90
  - *Big Fantastic database (BFD)* = 66M protein families from 2.2G protein sequences
  - clustered *MGnify*

needed at least 30 sequences per MSA otherwise quality deteriorated>



https://www.nature.com/articles/s41586-021-03819-2

Training

PDB database + PDB70 clusters

training db:

40% identity clusters, crop to 258 residues, batches by 128 per Tensor processing unit (TPU)

enhance accuracy by **noisy student self-distillation**

predict 350000 structures from UniRef30 using trained network

filter to high confidence subset

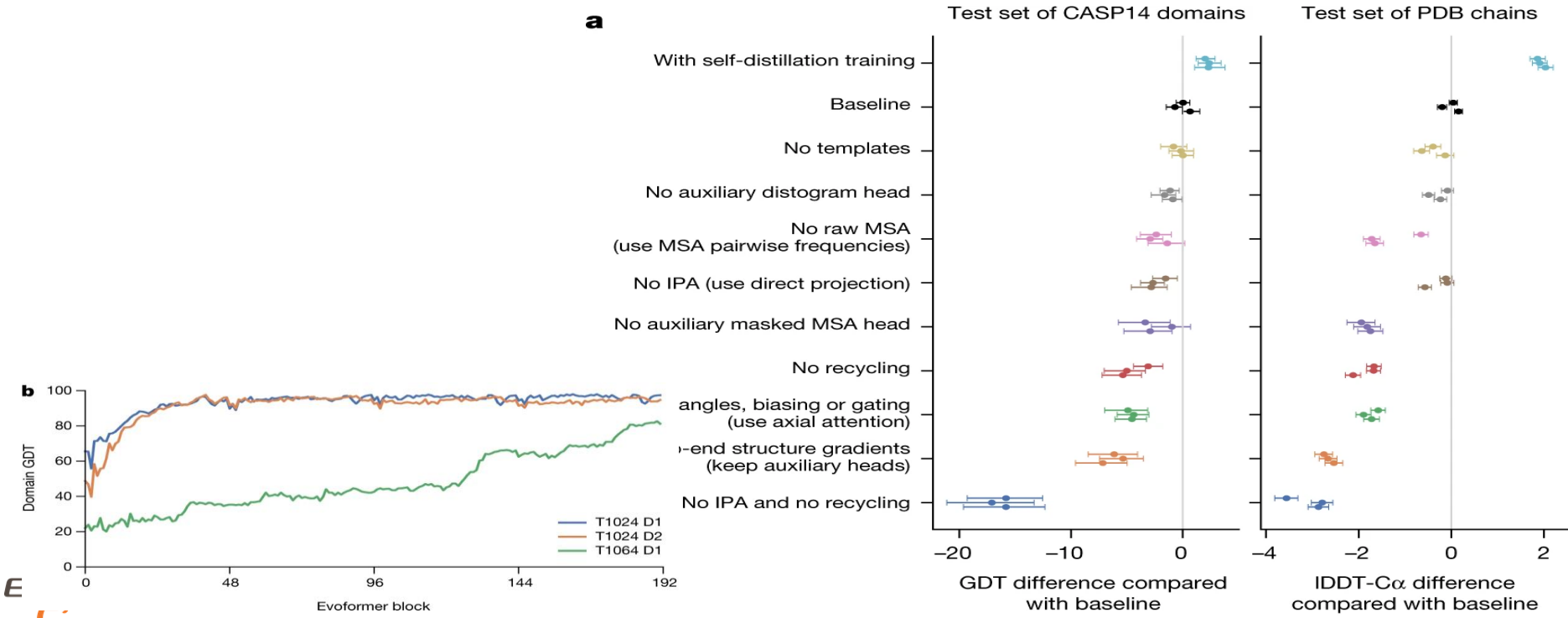
then train again from scratch with mixture of PDB and UniRef30

=> effective use of unlabelled sequence data

randomly mask or mutate individual residues from MSA using BERT (bidirectional encoder representations from Transformers => to predict masked elements within MSA

https://www.nature.com/articles/s41586-021-03819-2

Interpreting the neural network



depth of neural network - it is usually quick, but for challenging targets it can be quite deep

https://www.nature.com/articles/s41586-021-03819-2

What is "diffusion"?

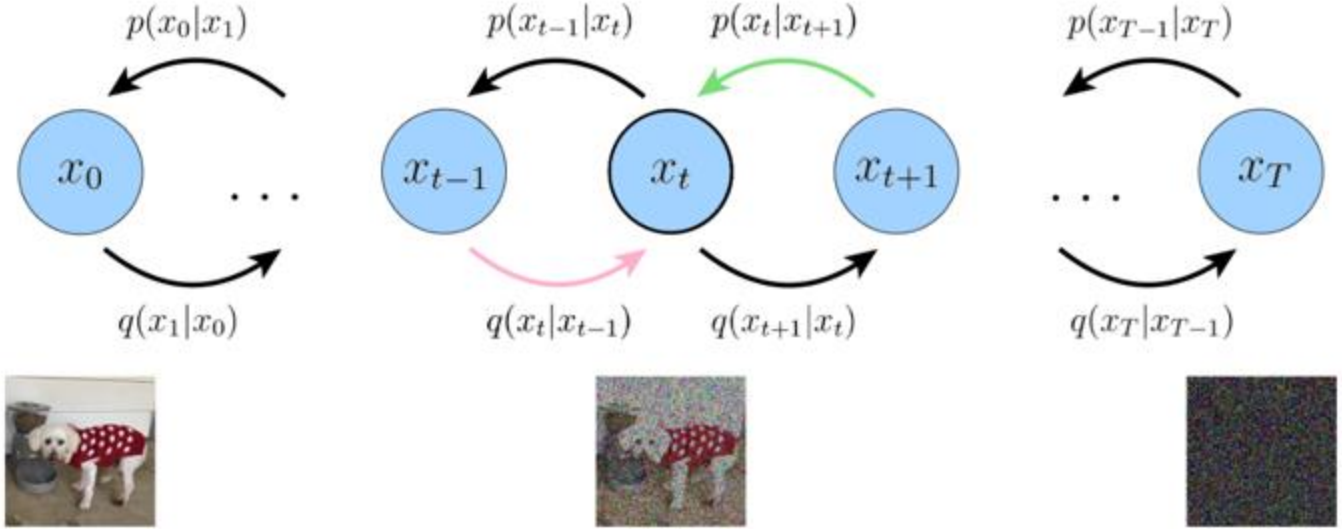
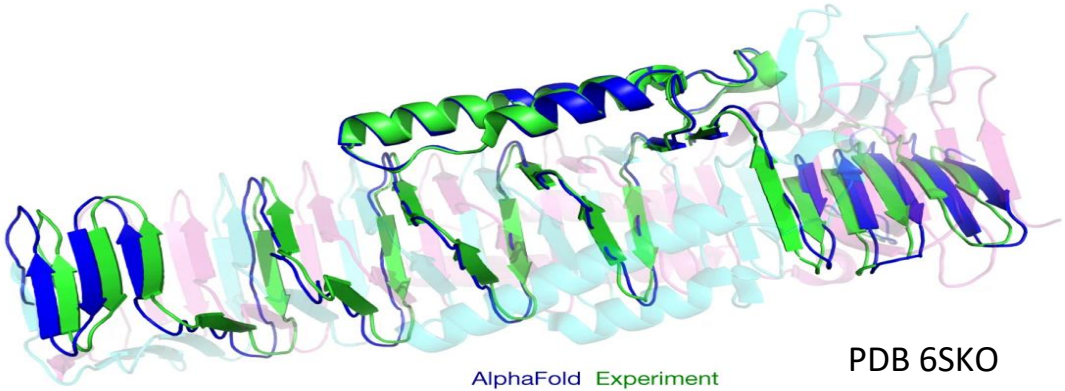


Image from Luo, 2022

Effect of cross-chain contacts.

prediction is worse for **heterotropic** contacts (large complexes where 3D structure is dictated by other chains in complex)

homotropics yields high-accuracy even when chains are intertwinned



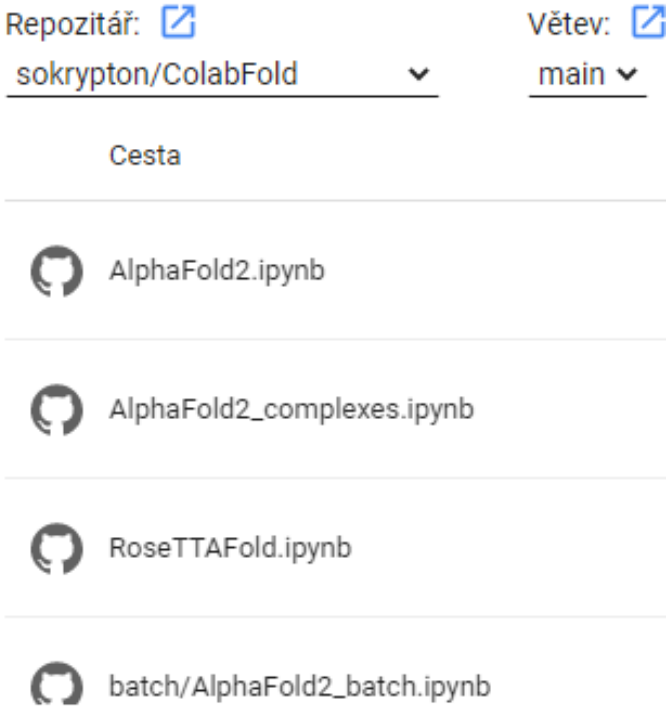
PDB 6SKO

AlphaFold in Google Colab

Github enabled JupyterNotebooks running in Google Colab environment

limitation in size (time)

Starting so from Chimera



Mirdita M, Ovchinnikov S, Steinegger M. ColabFold - Making protein folding accessible to all. *bioRxiv*, 2021 https://doi.org/10.1101/2021.08.15.456425

https://colab.research.google.com/github/sokrypton/ColabFold/

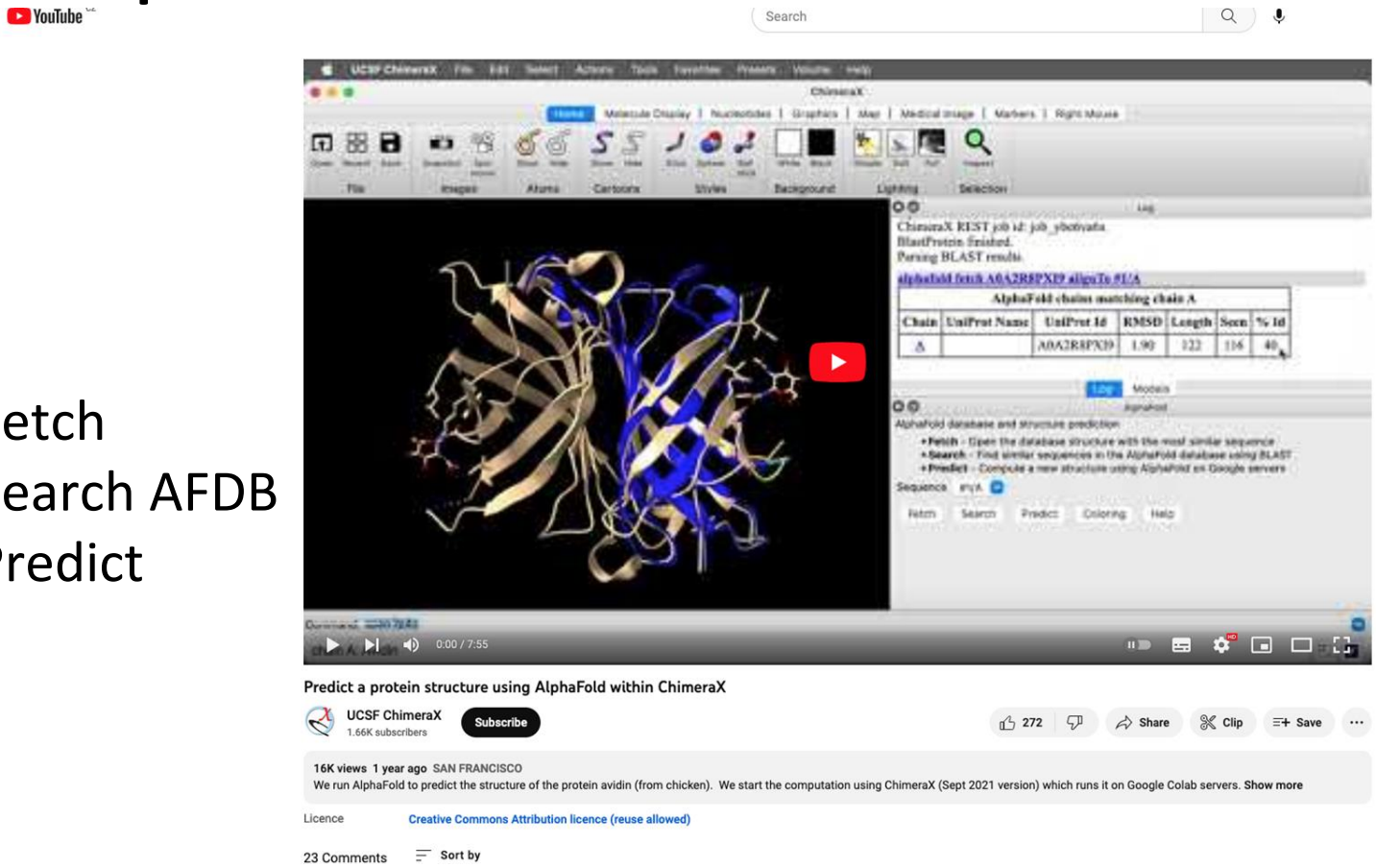
Alphafold 2 on ELIXIR CZ

- Alphafold “needs” TPU to run -> not many people have it on their PC
- Alphafold has been installed on Elixir CZ hardware
- Alphafold (Multimer) in the newest version 2.2.0 is accessible through Metacentrum
- speed is dependent on size of predicted protein (complex)

https://wiki.metacentrum.cz/wiki/AlphaFold

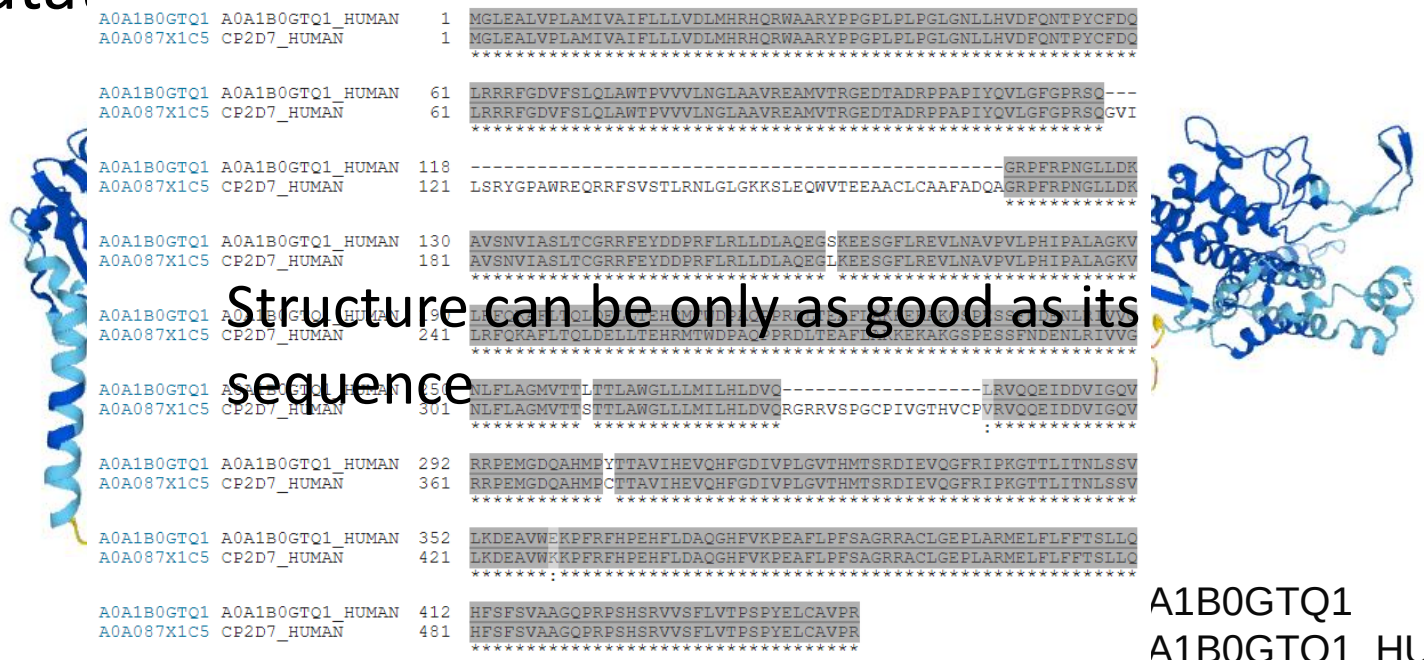
AlphaFold within ChimeraX

Fetch
Search AFDB
Predict



But one still needs to be careful…

putative human cytochrome P450 2C7

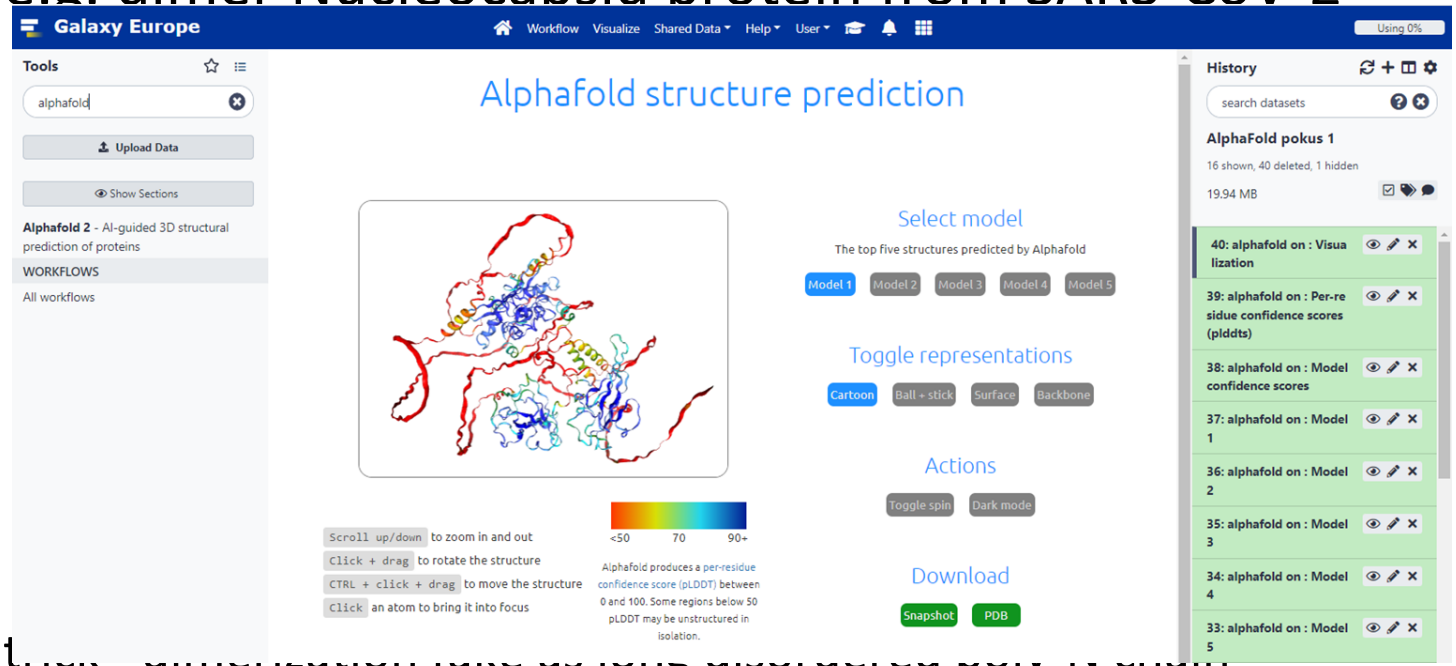


Structure can be only as good as its sequence

A1B0GTQ1
A1B0GTQ1_HUMAN

AlphaFold in UseGalaxy.eu

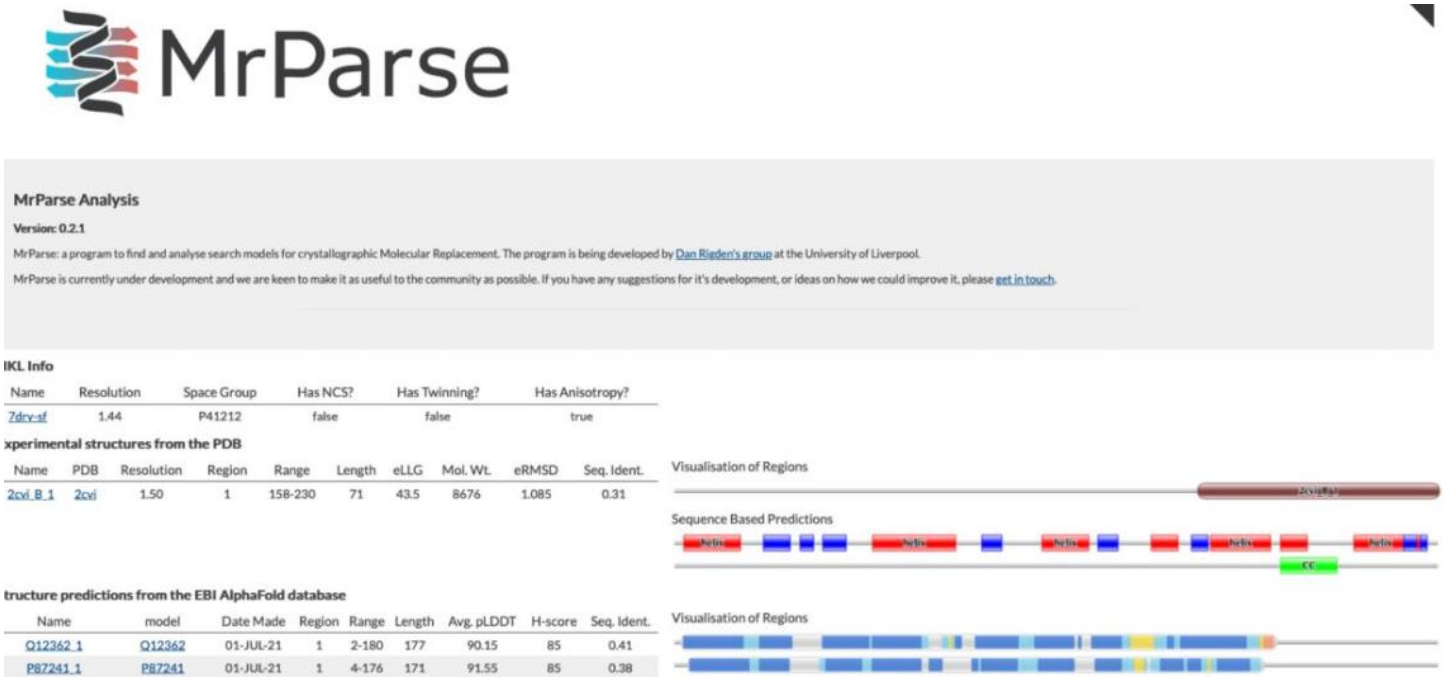
e.g. dimer Nucleocapsid protein from SARS-CoV-2



trick - dimerization fake as long disordered poly-N chain

https://usegalaxy.eu/tool_runner?tool_id=toolshed.g2.bx.psu.edu%2Frepos%2Fgalaxy-australia%2Falphafold2%2Falphafold%2F2.1.2%2Bgalaxy0

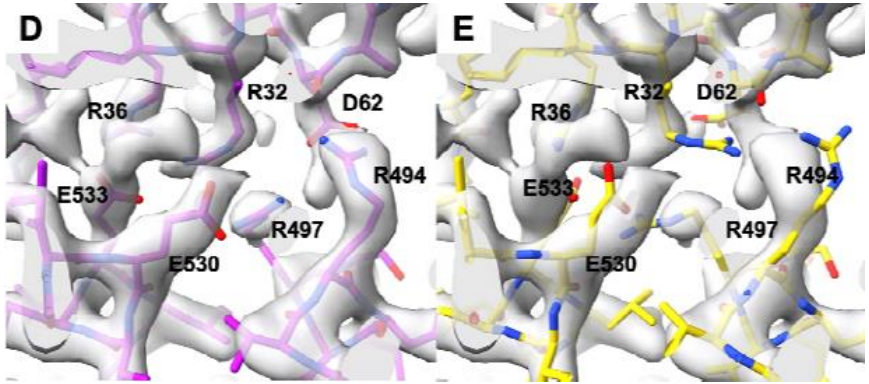
MrParse: Finding homologues in the PDB and the EBI AlphaFold database for Molecular Replacement and more



Adam J. Simpkin, Jens M. H. Thomas, Ronan M. Keegan, Daniel J. Rigden

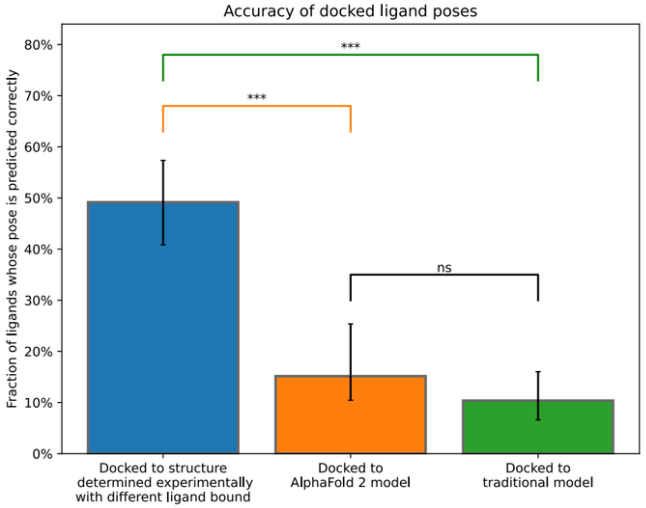
doi: https://doi.org/10.1101/2021.09.02.458604

How accurate are the models?



AlphaFold predictions are valuable hypotheses, and accelerate but do not replace experimental structure determination

Thomas C. Terwilliger, Dorothee Liebschner, Tristan I. Croll, Christopher J. Williams, Airlie J. McCoy, Billy K. Poon, Pavel V. Afonine, Robert D. Oeffner, Jane S. Richardson, Randy J. Read, Paul D. Adams



How accurately can one predict drug binding modes using AlphaFold models?

Masha Karelina, Joseph J. Noh, Ron O. Dror

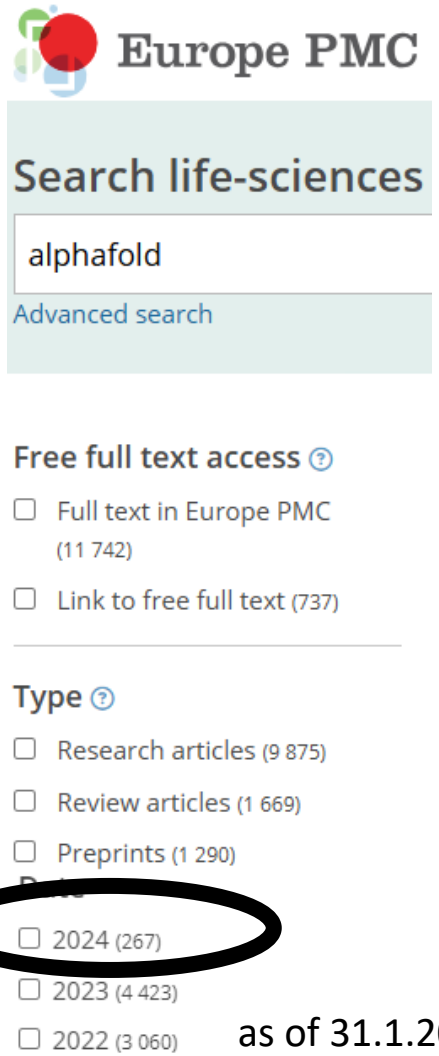
doi: https://doi.org/10.1101/2023.05.18.541346

This article is a preprint and has not been certified by peer review [what does this mean?].

Alphafold is just a start...

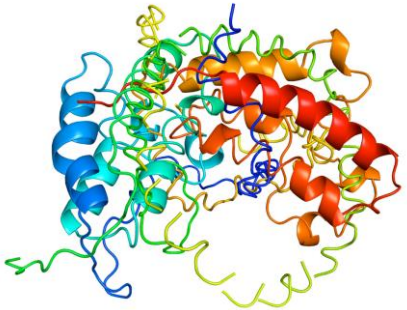
- use Alphafold ideas for development of their own 3D structure predictions
  - RoseTTAfold
  - ESMfold
  - OpenFold
  - Chroma
- prediction of designed proteins

...



as of 31.1.2024

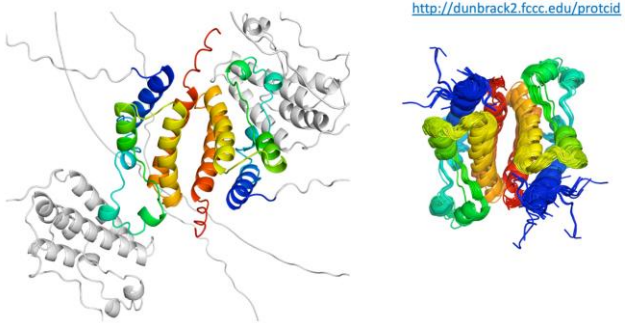
Alphafold can do **multiprotein complexes** – interactions



Alphafold-Multimer v2 reproduces dimer of Bromodomains BD2 of BET proteins observed in crystal structures

AF2mult_v2 homodimer of BRD2_HUMAN Bromodomain B2 in rainbow; BD1 in gray

ProtCID cluster of dimers of BD2 domains of human BRD2, BRD3, BRD4, mouse BRDT
http://dunbrack2.fccc.edu/protcid



https://twitter.com/RolandDunbrack/status/1502818748868317188

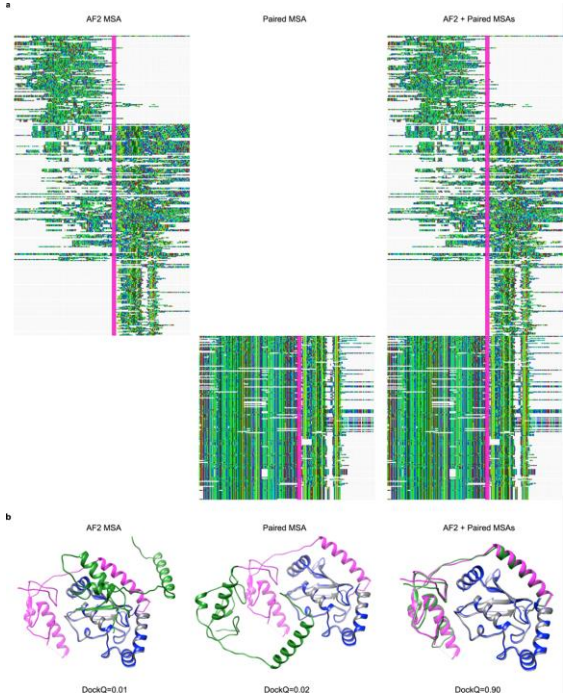
DeepMind

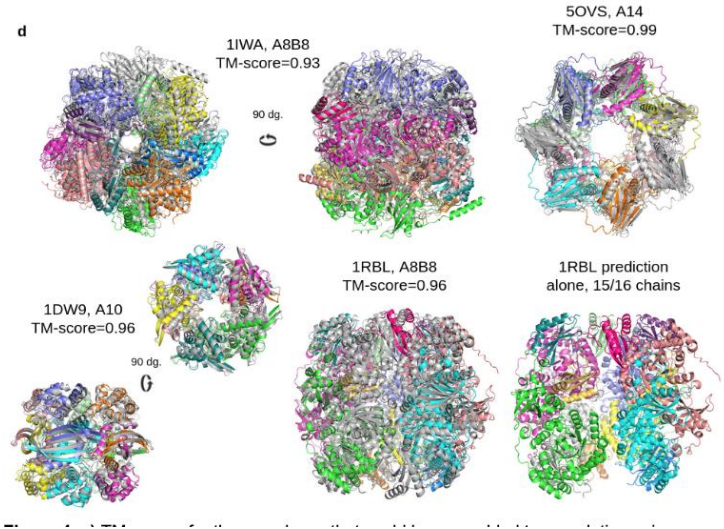
2022-03-10

Protein complex prediction with AlphaFold-Multimer

Richard Evans[1*], Michael O'Neill[1*], Alexander Pritzel[1*], Natasha Antropova[1*], Andrew Senior[1], Tim Green[1], Augustin Žídek[1], Russ Bates[1], Sam Blackwell[1], Jason Yim[1], Olaf Ronneberger[1], Sebastian Bodenstein[1], Michal

Alphafold can do **multiprotein complexes** – interactions





Article | Open Access | Published: 10 March 2022

Improved prediction of protein-protein interactions using AlphaFold2

Patrick Bryant, Gabriele Pozzati & Arne Elofsson

Nature Communications **13**, Article number: 1265 (2022) | Cite this article
6092 Accesses | **27** Altmetric | Metrics

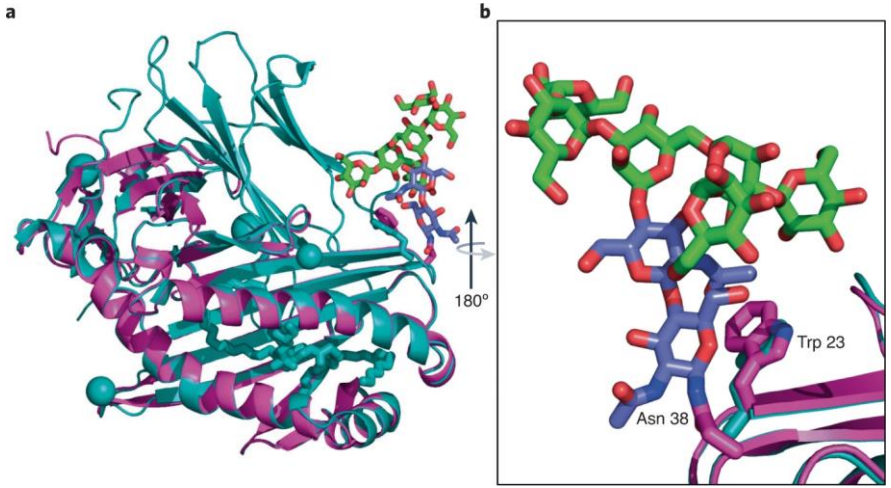
New Results

Follow this preprint

Predicting the structure of large protein complexes using AlphaFold and sequential assembly

Patrick Bryant, Gabriele Pozzati, Wensi Zhu, Aditi Shenoy, Petras Kundrotas, Arne Elofsson
doi: https://doi.org/10.1101/2022.03.12.484089

- Alphafold can not do effects of **post-translational protein modifications** (by itself)



Correspondence | Published: 29 October 2021

The case for post-predictional modifications in the AlphaFold Protein Structure Database

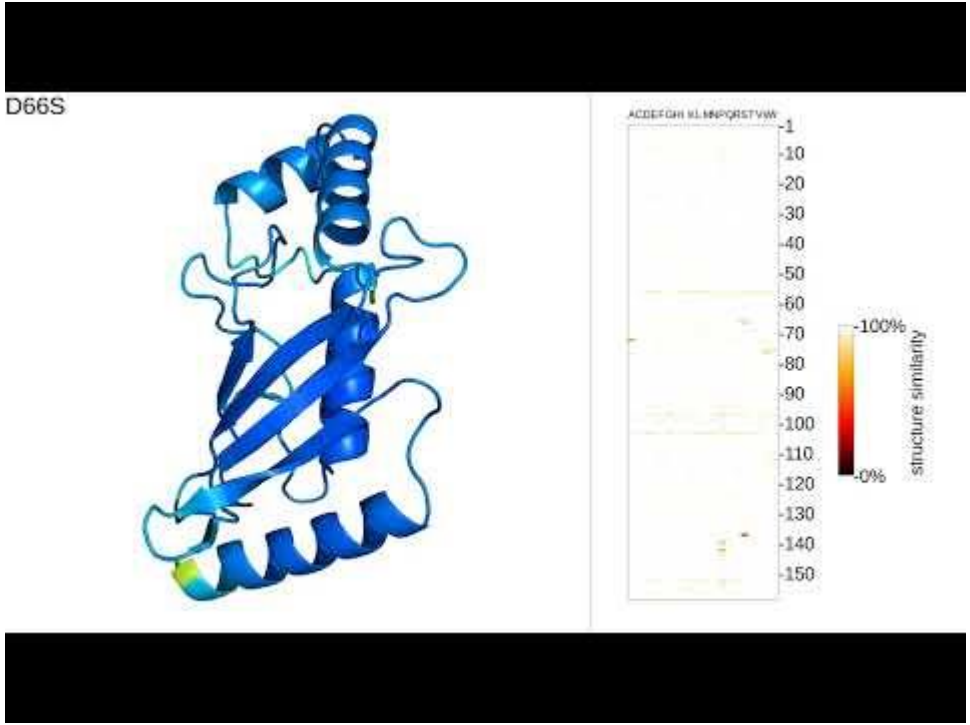
Haroldas Bagdonas, Carl A. Fogarty, Elisa Fadda & Jon Agirre

Nature Structural & Molecular Biology **28**, 869–870 (2021) | Cite this article

10k Accesses | **2** Citations | **151** Altmetric | Metrics

MutAmore

- generate all SNPs of protein
- using ESMfold/OpenFold



Rendering protein mutation movies with MutAmore

Konstantin Weissenow, Burkhard Rost

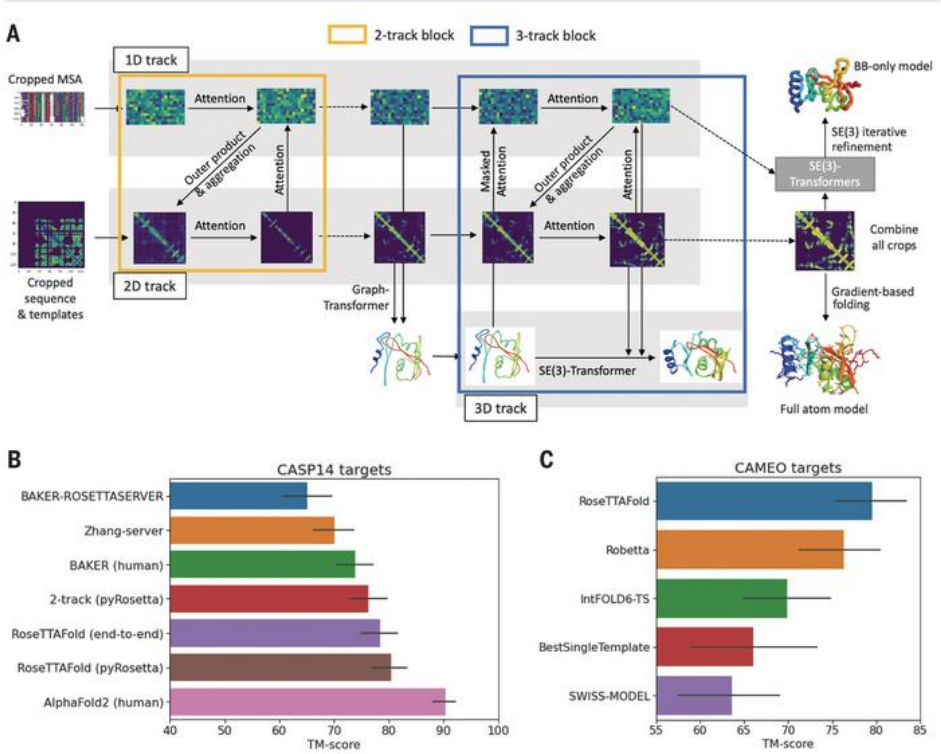
doi: https://doi.org/10.1101/2023.09.15.557870

https://www.youtube.com/watch?v=1XgiFXg-Xrs&list=PL0QUUE_zWBuJ6Y5NWtDoY93FUweUUGVuf

https://www.biorxiv.org/content/10.1101/2023.09.15.557870v1

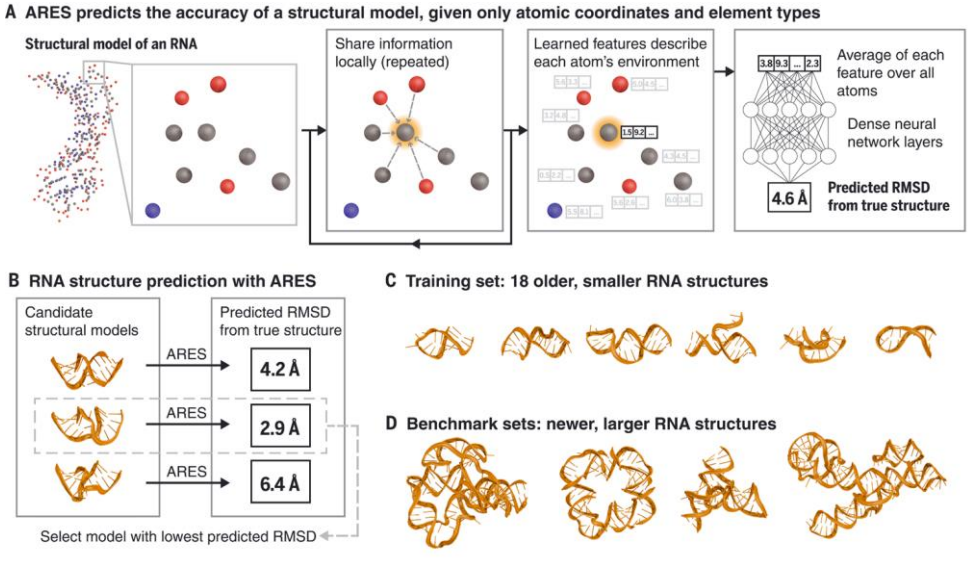
https://github.com/kWeissenow/MutAmore

Accurate prediction of protein structures and interactions using a three-track neural network



https://www.science.org/doi/full/10.1126/science.abj8754

Geometric deep learning of RNA structure

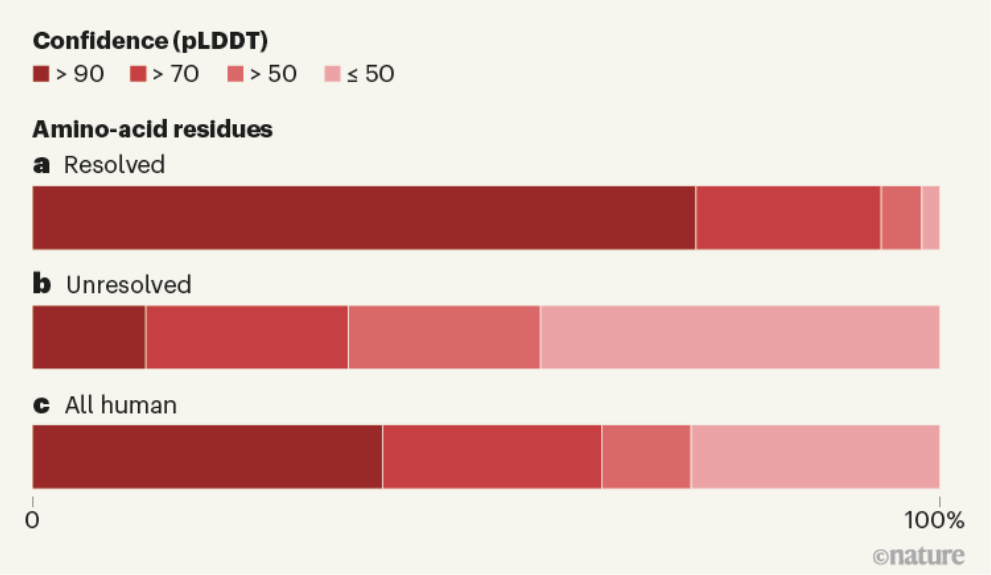


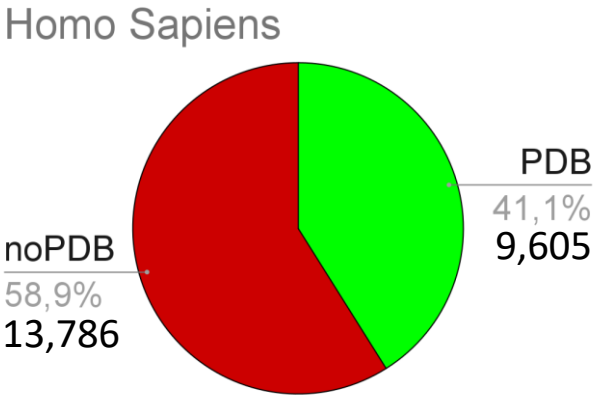
https://www.science.org/doi/10.1126/science.abe5650

Not used files

AF on proteomes

How good are the predictions of human proteins?





pLDDT

- quality metrics
- measure of disorder

pLDDT - per-residue estimate of its confidence on a scale from 0 - 100 model's predicted score on the lDDT-Cα metric (local superposition-free score for comparing protein structures and models using distance difference tests).

Tunyasuvunakool, K., Adler, J., Wu, Z. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* 596, 590–596 (2021). https://doi.org/10.1038/s41586-021-03828-1

TED

The Encyclopedia of Domains

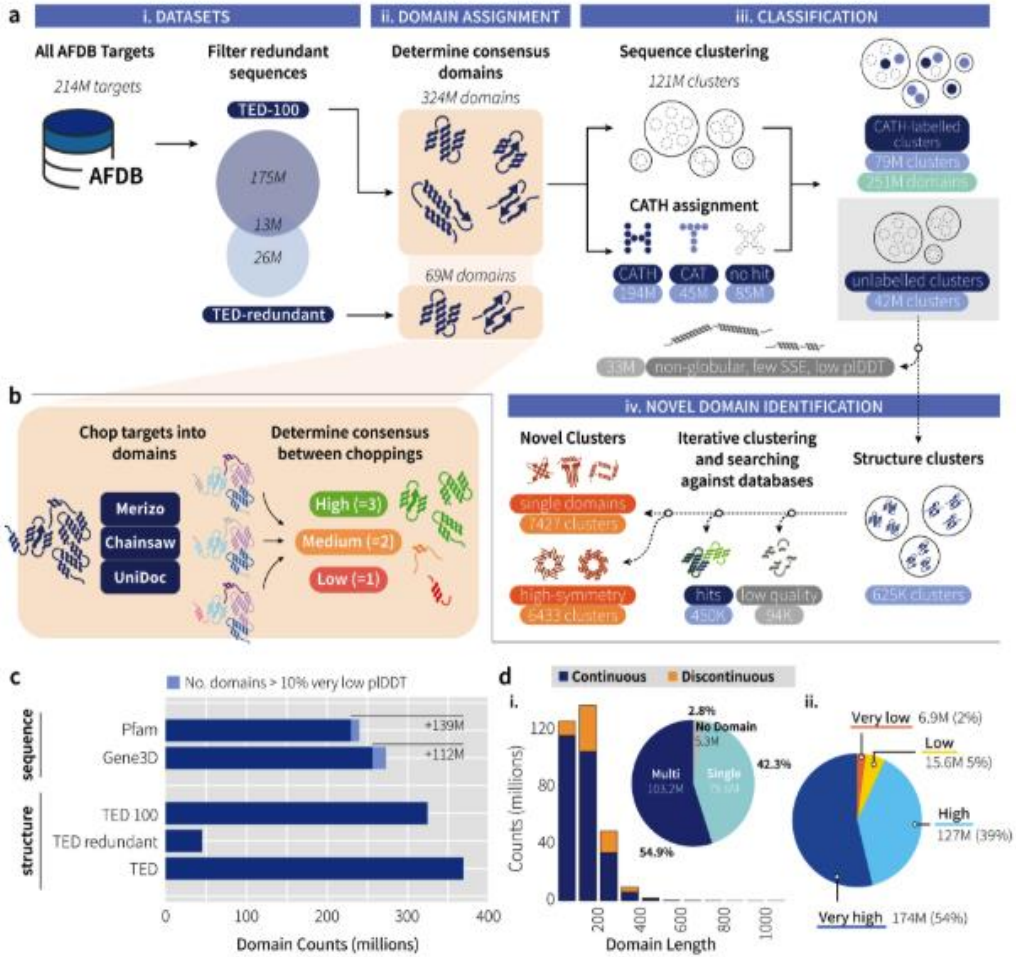
365 million domains

77% of the nonredundant domains are similar to known superfamilies

>10,000 new structural interactions between superfamilies and thousands of new folds across the fold space continuum

ted.cathdb.info/

zenodo.org/records/13908086



Andy M. Lau *et al.* Exploring structural diversity across the protein universe with The Encyclopedia of Domains. *Science* **386**,eadq4946(2024). DOI:10.1126/science.adq4946