AlphaFold2
ML Revolution in Structural Biology

Marian Novotný, Karel Berka

9th September 2021

elixir CZECH REPUBLIC

PŘÍRODOVĚDECKÁ FAKULTA
Univerzita Karlova

KATEDRA FYZIKÁLNÍ CHEMIE
Univerzita Palackého v Olomouci

Outline

- Protein structure prediction
- CASP14
- AlphaFold2 - under the hood
- Uses of AF2
- AF2 DB
- AF2 in MetaCentrum
- Other software (RosettaFold, ML)
- AF2 publically available servers - power of Jupyter notebooks
- Limitations and Future challenges

elixir CZECH REPUBLIC

Průlom v biologii. Umělá inteligence „vyřešila“ šmodrchání proteinů na 92 %

NEWS · 30 NOVEMBER 2020

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

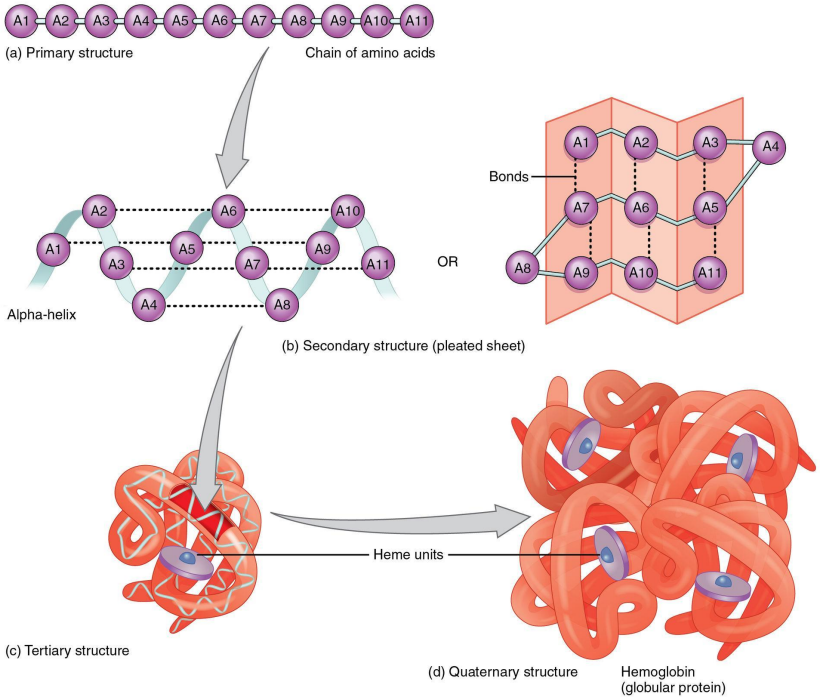
Umělá inteligence AlphaFold dosáhla vědeckého průlomu. Dovede stanovit tvar molekul proteinů

‘The game has changed.’ AI triumphs at solving protein structures

elixir CZECH REPUBLIC

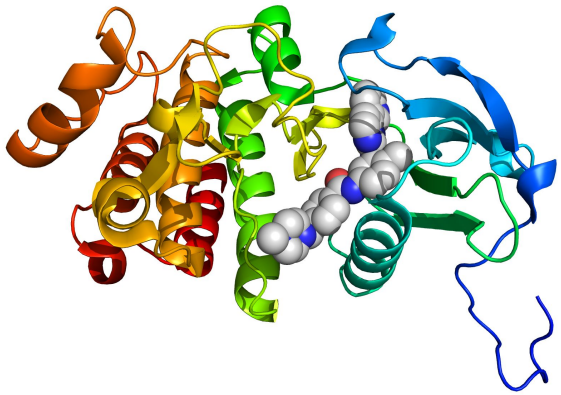
So what is a protein?

- elements of cells that actually do things
- responsible for almost everything
- composed of amino acids
- produced from RNA by ribosomes
- folding leads to a 3D structure
- human has around 20 000 different proteins

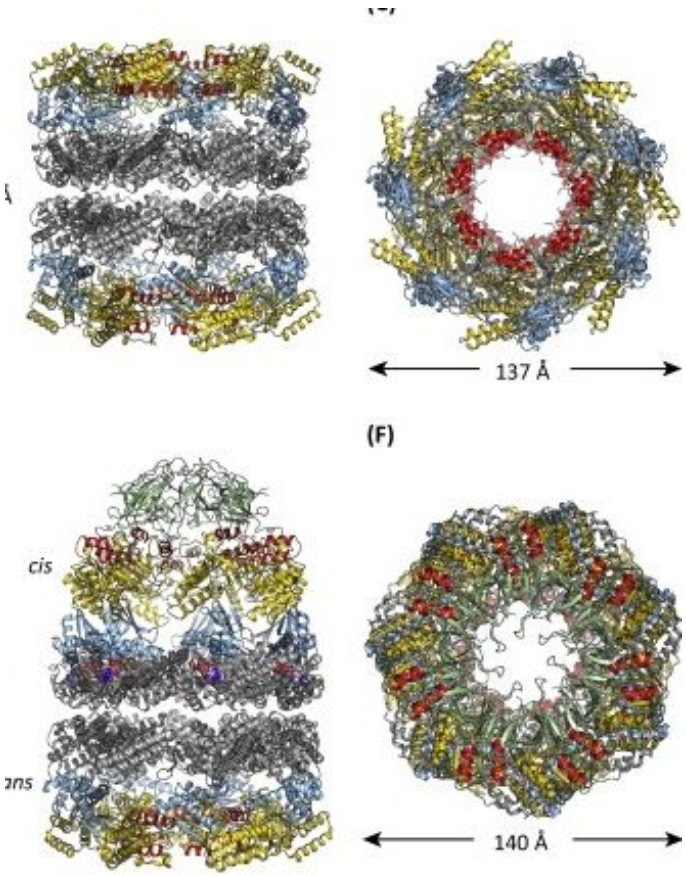


wikipedia

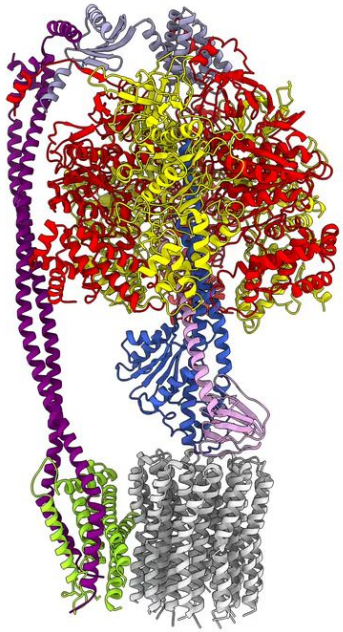
Knowing structure helps to understand the function



wikipedia



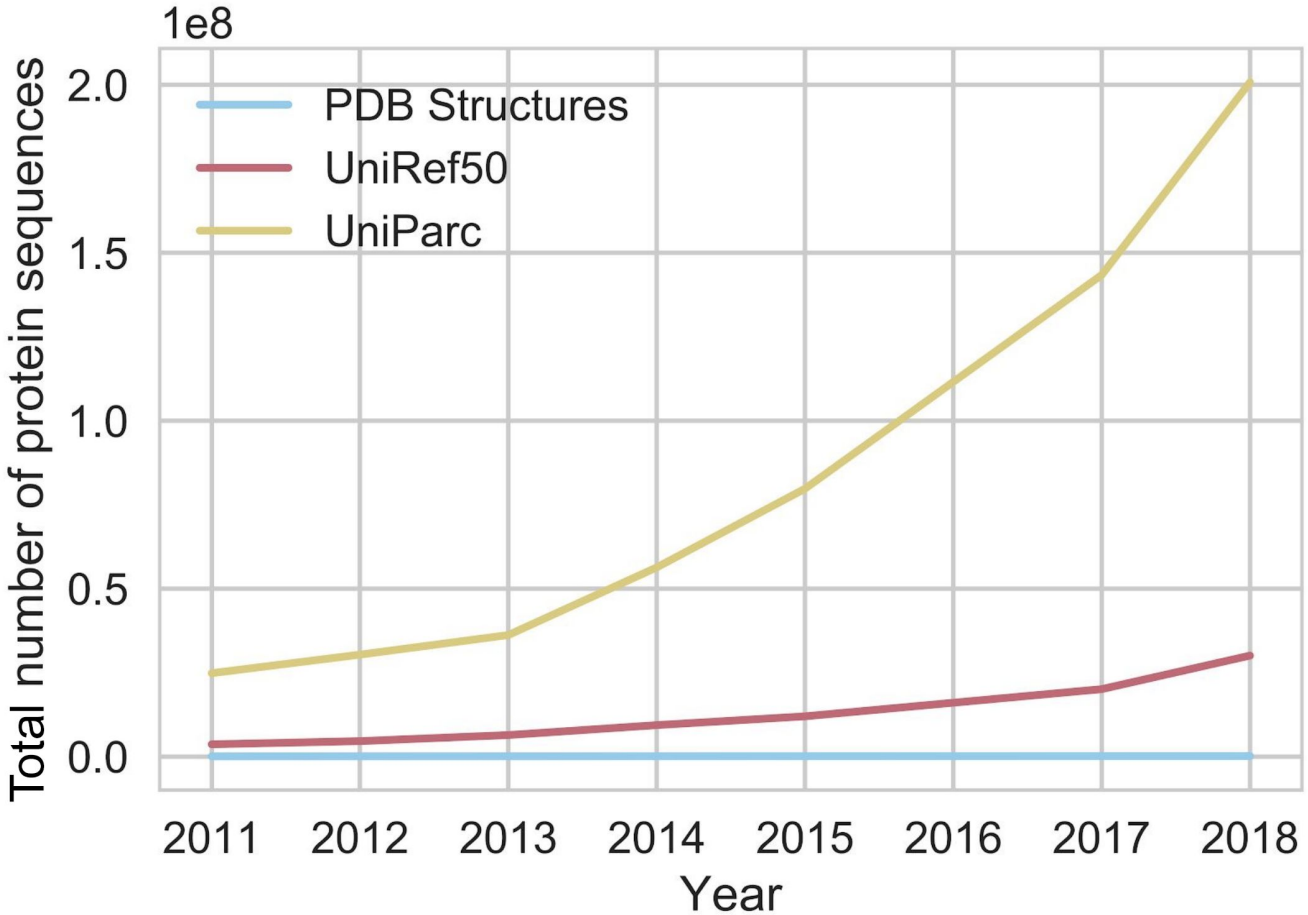
Hayer-Hartl et al., 2015



Guo et al., 2019

elixir CZECH REPUBLIC

Solving 3D structures is expensive...



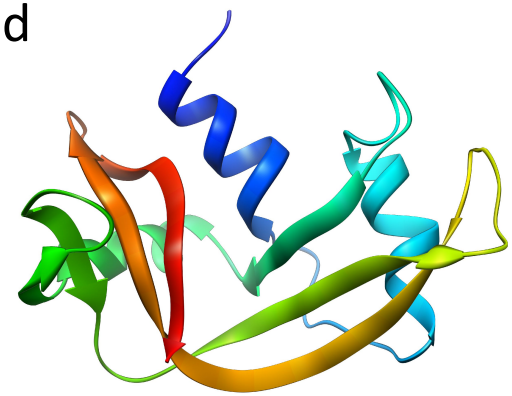
https://bair.berkeley.edu/blog/2019/11/04/proteins/

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

Can we use sequence to predict 3D structure?

- C.B. Anfinsen received Nobel prize in Chemistry (1972) for describing the relationship between sequence and structure

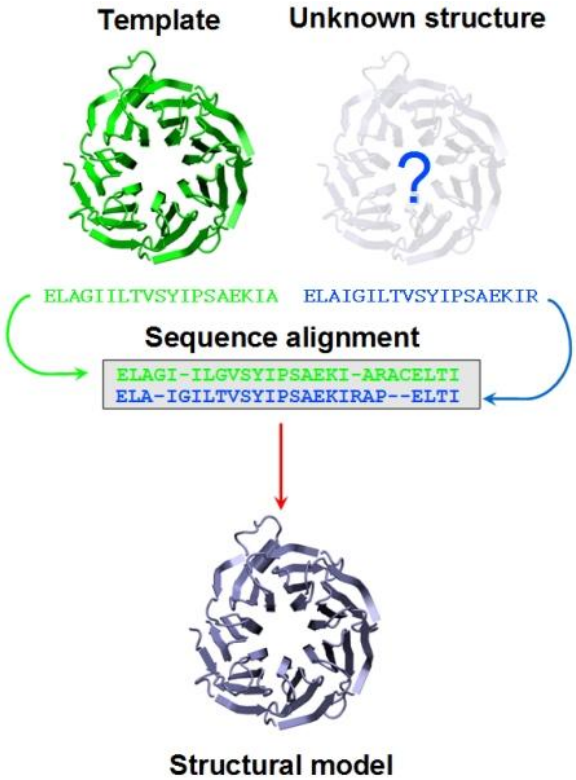
"The native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment."

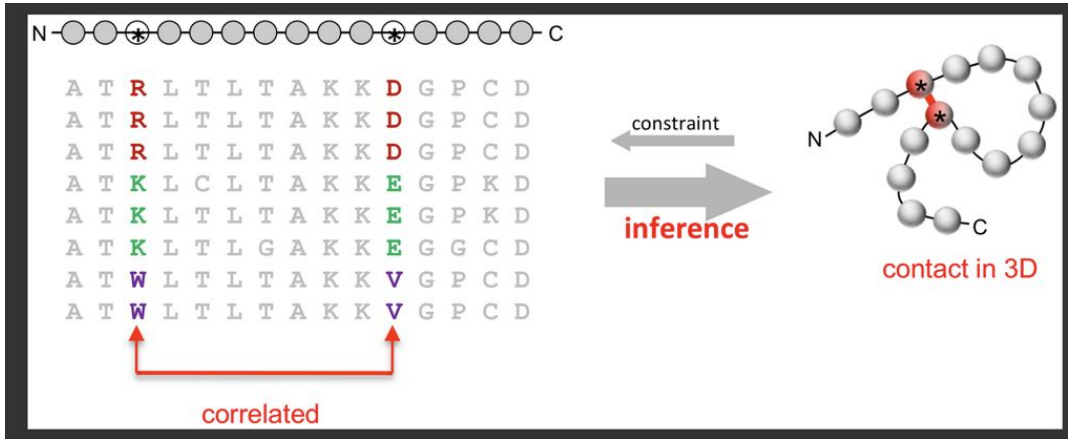


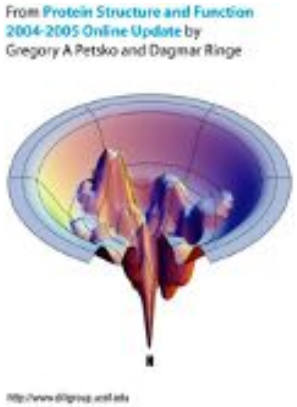
- it shall be possible to give
  to predict structure from sequence

z

Principles of prediction from sequence







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

elixir CZECH REPUBLIC

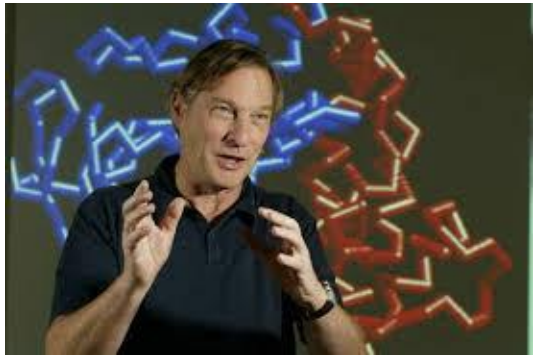
Structure prediction = simulation of protein folding

Levinthal's paradox - protein of 100 aa has 10^{70} available conformations -> it would take 10^{52} years at the speed of 10^{-11} s to sample one conformation to assume it native shape

elixir CZECH REPUBLIC

How to move the prediction field forward?

- transparent competition
- provide an “environment” for communication and exchange of experience
- develop metrics for careful examination of predicted structures

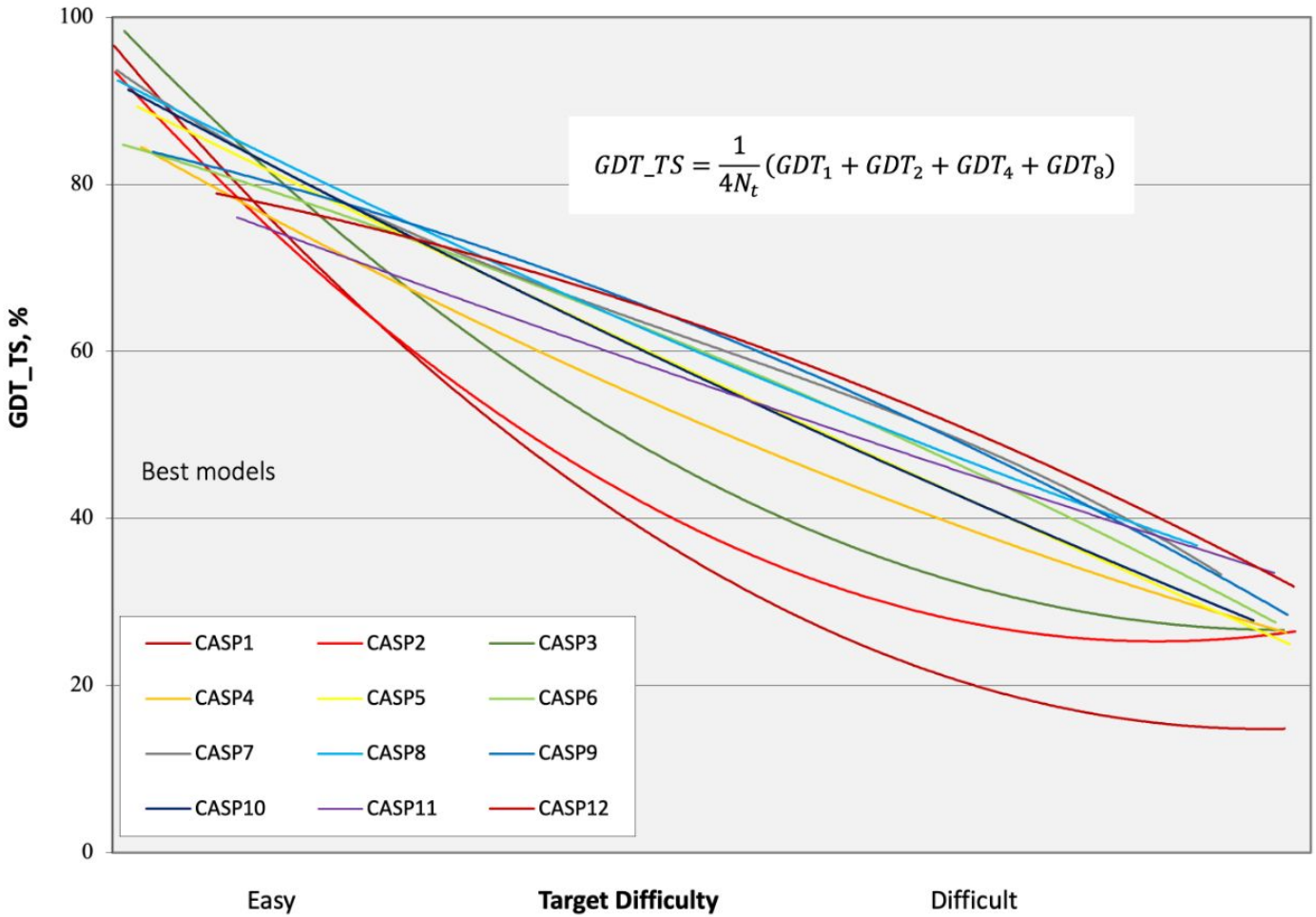


- **CASP** – critical assessment of protein structure prediction
- once in two years since1994
- compare with experimentally solved structures

elixir CZECH REPUBLIC

CASP

How to compare structures?

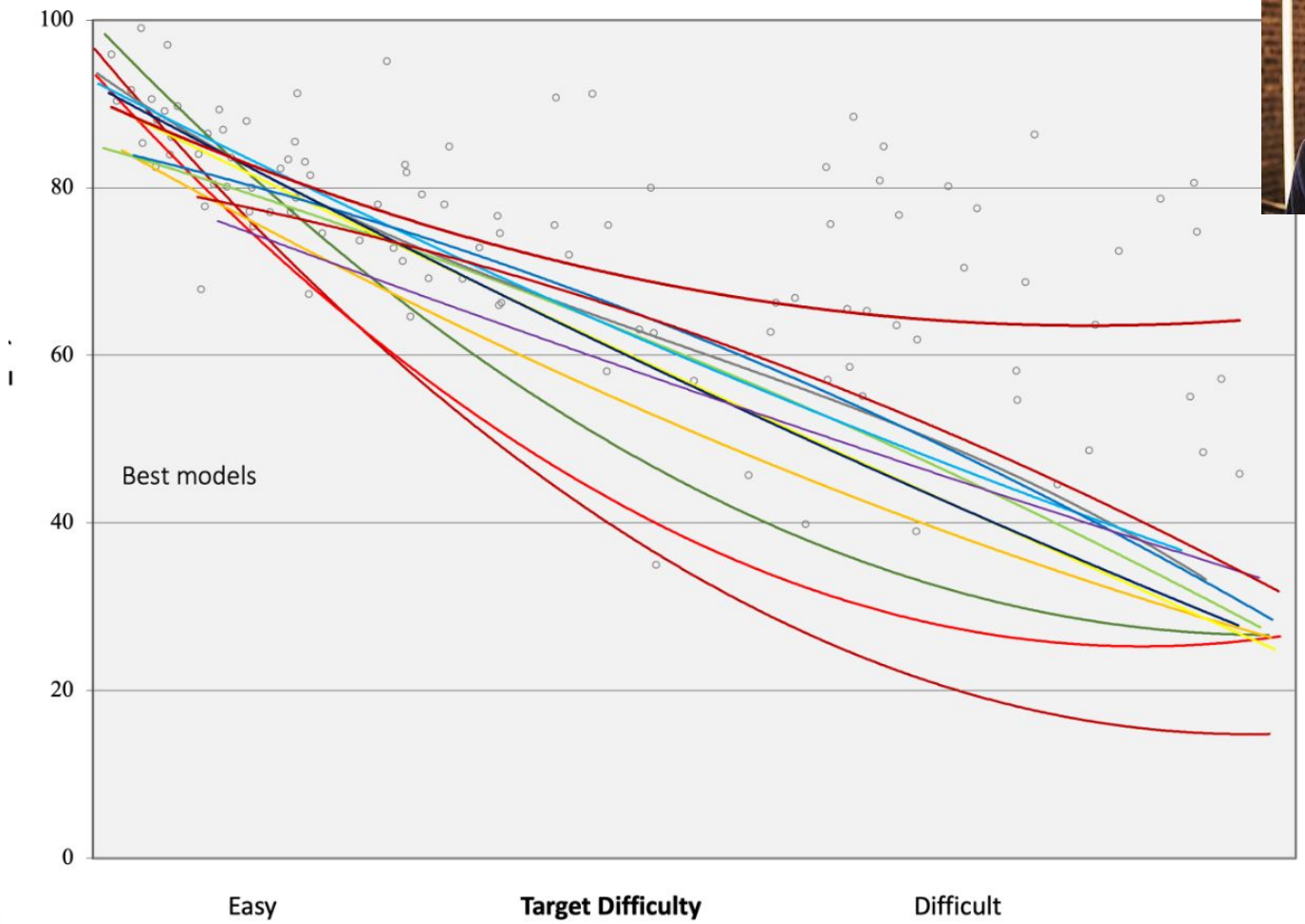


https://predictioncenter.org/casp14/doc/presentations/2020_11_30_CASP14_Introduction_Moult.pdf

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

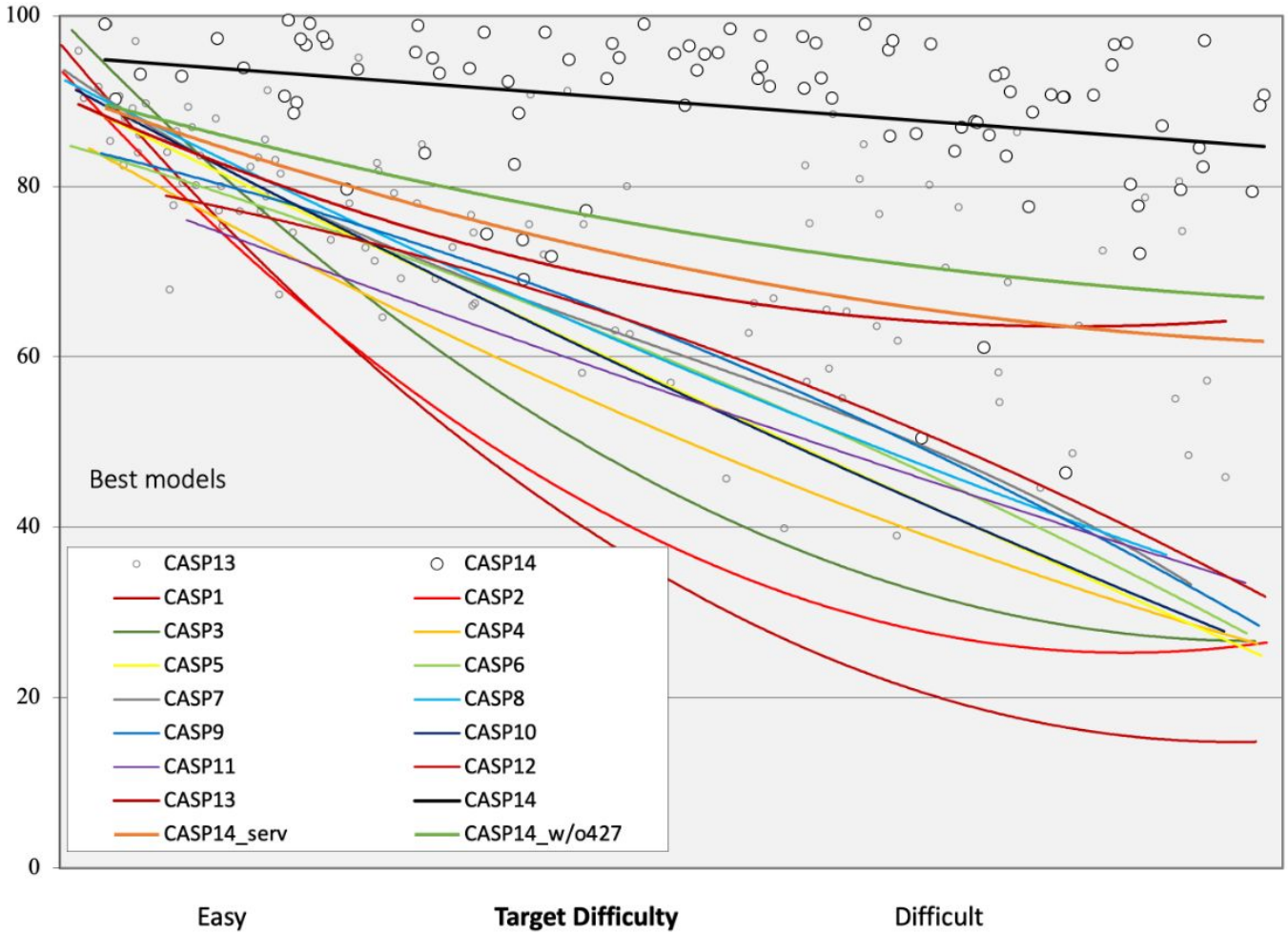
2018: AlphaFold enters...



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

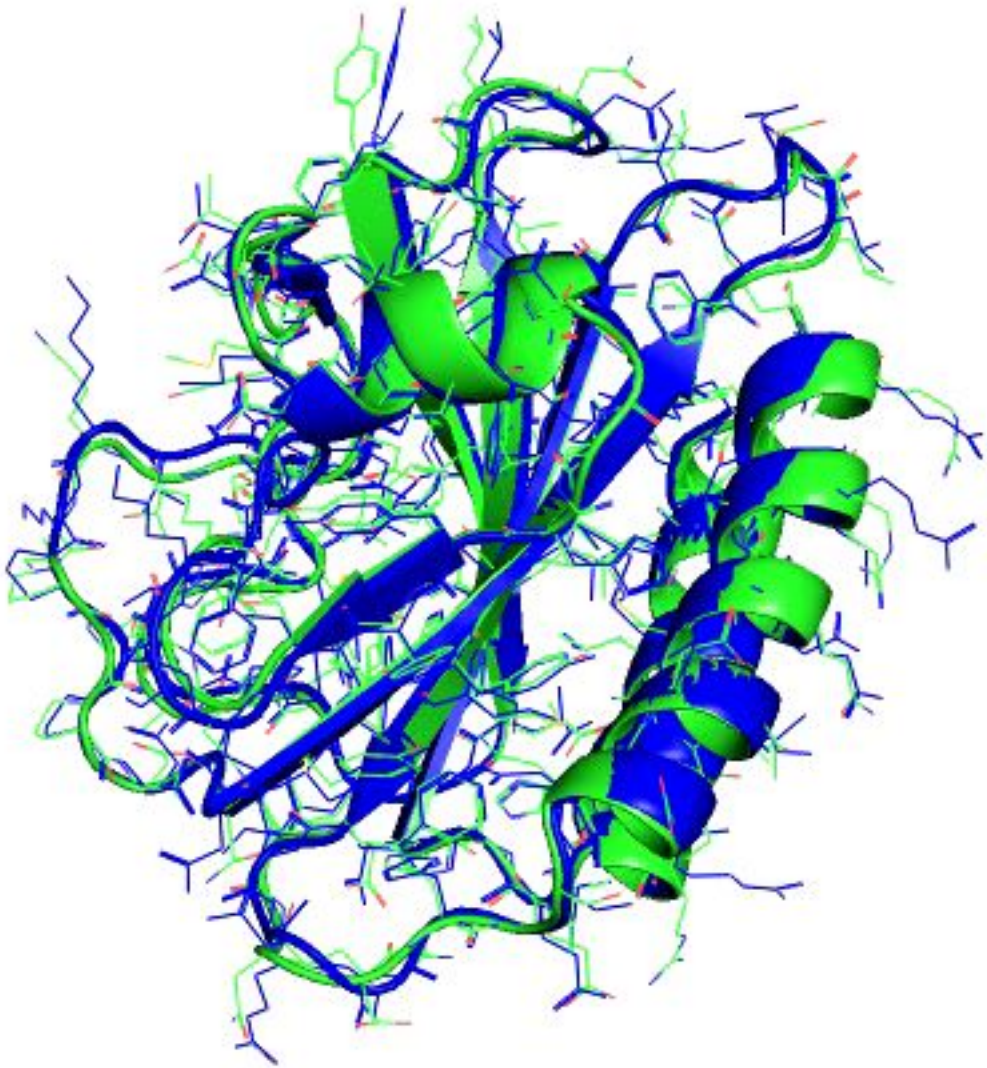
elixir CZECH REPUBLIC

2020: Alphafold2 wins



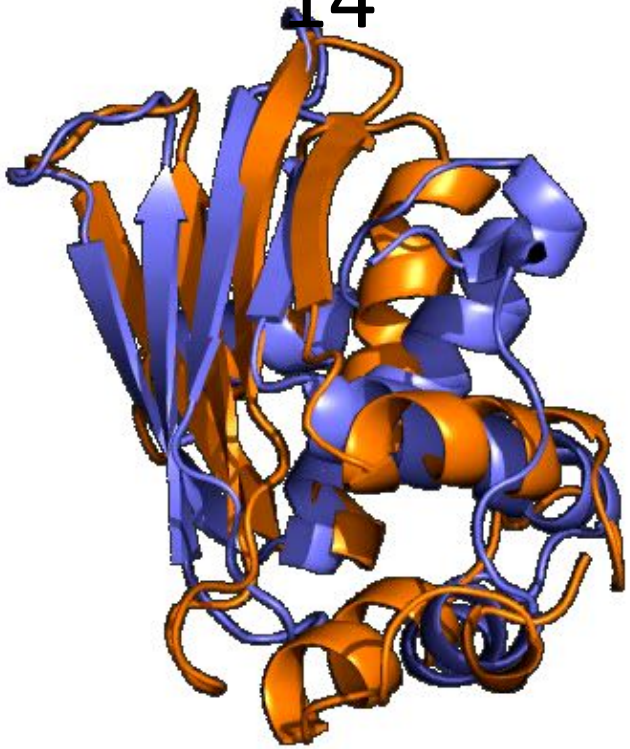
https://predictioncenter.org/casp14/doc/presentations/2020_11_30_CASP14_Introduction_Moult.pdf

How does good prediction look like?



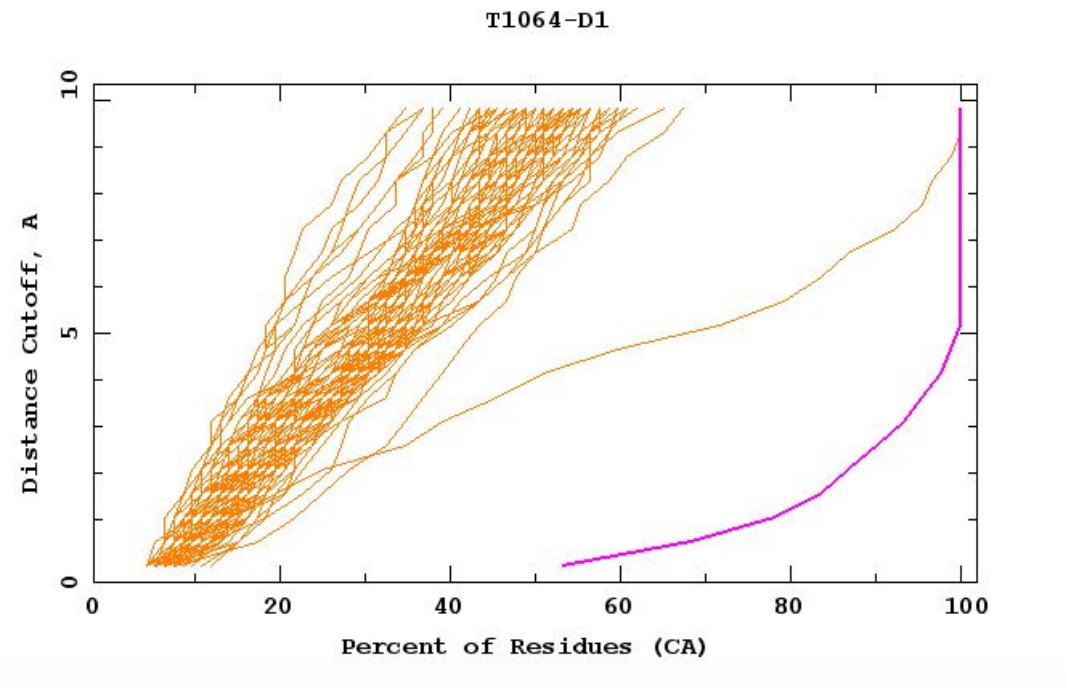
GDT_TS = 96.5

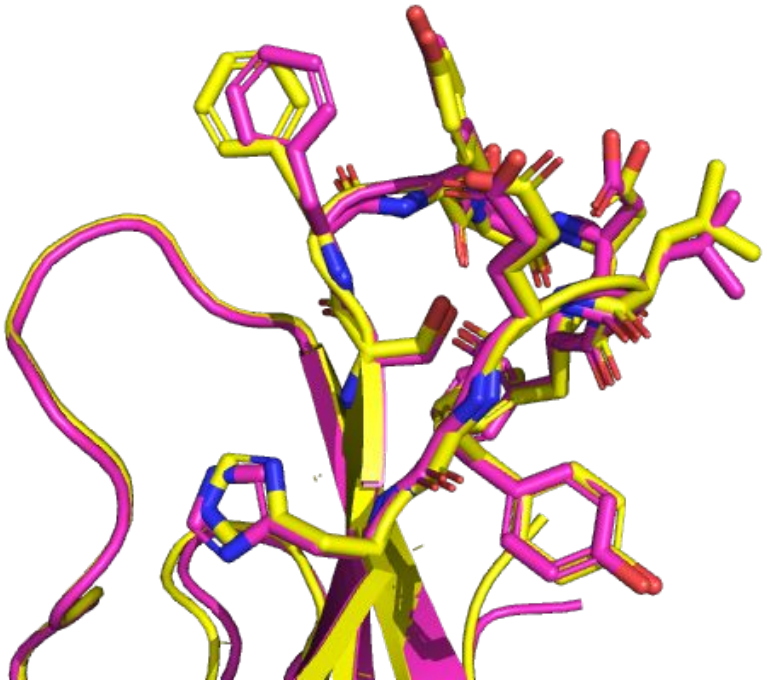
The worst prediction of Alphafold 2 in CASP 14



GDT_TS = 44.6

Side chain predictions– orf8 covid19





GDT_TS= 87

so how it works?

AlphaFold2
- under the hood

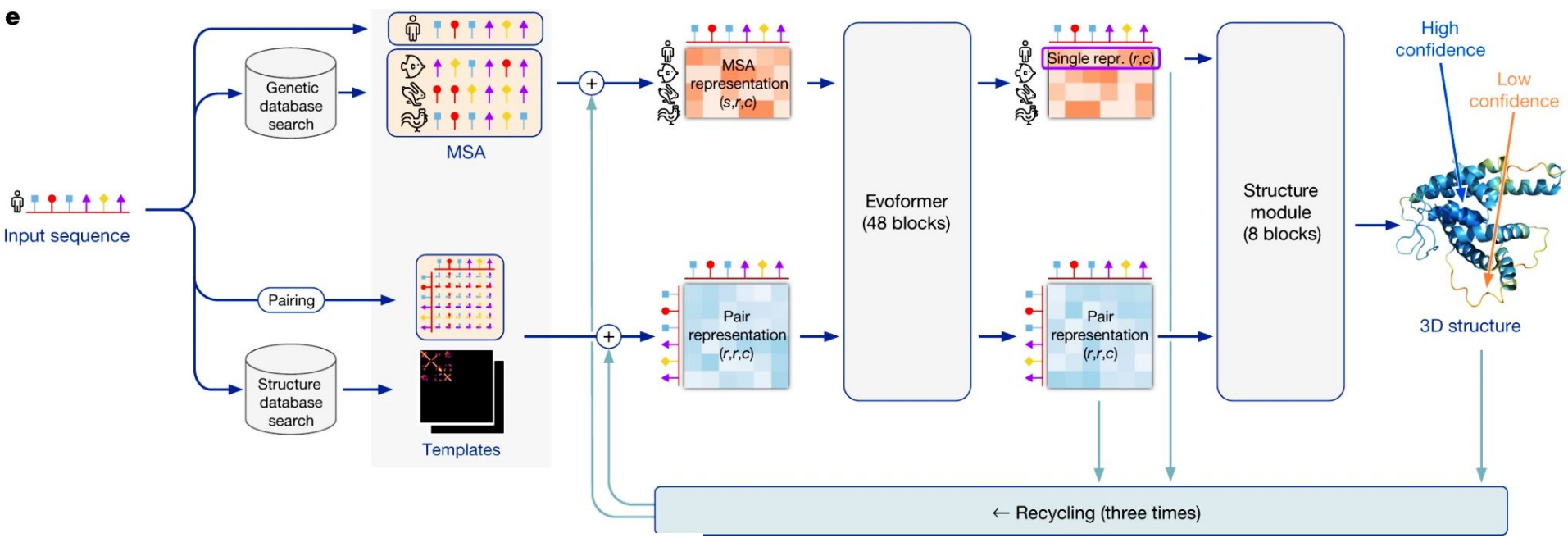
AlphaFold2

Input: sequence

extended by MSA + structural templates

Evoformer and Structure model (w MD simulation)

plDDT - predicted local confidence prediction



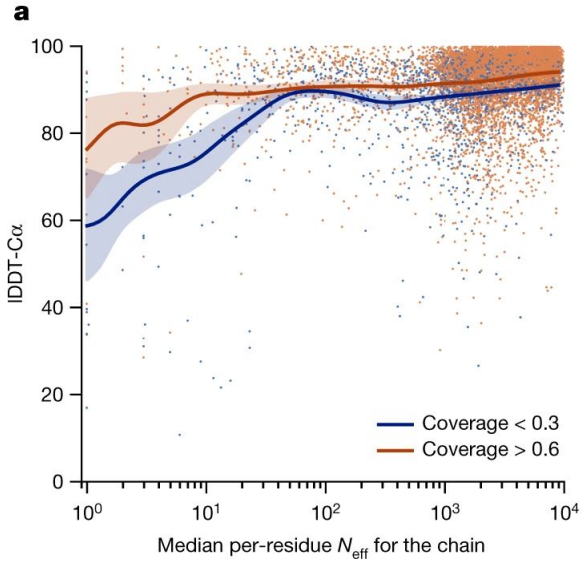
https://www.nature.com/articles/s41586-021-03819-2

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

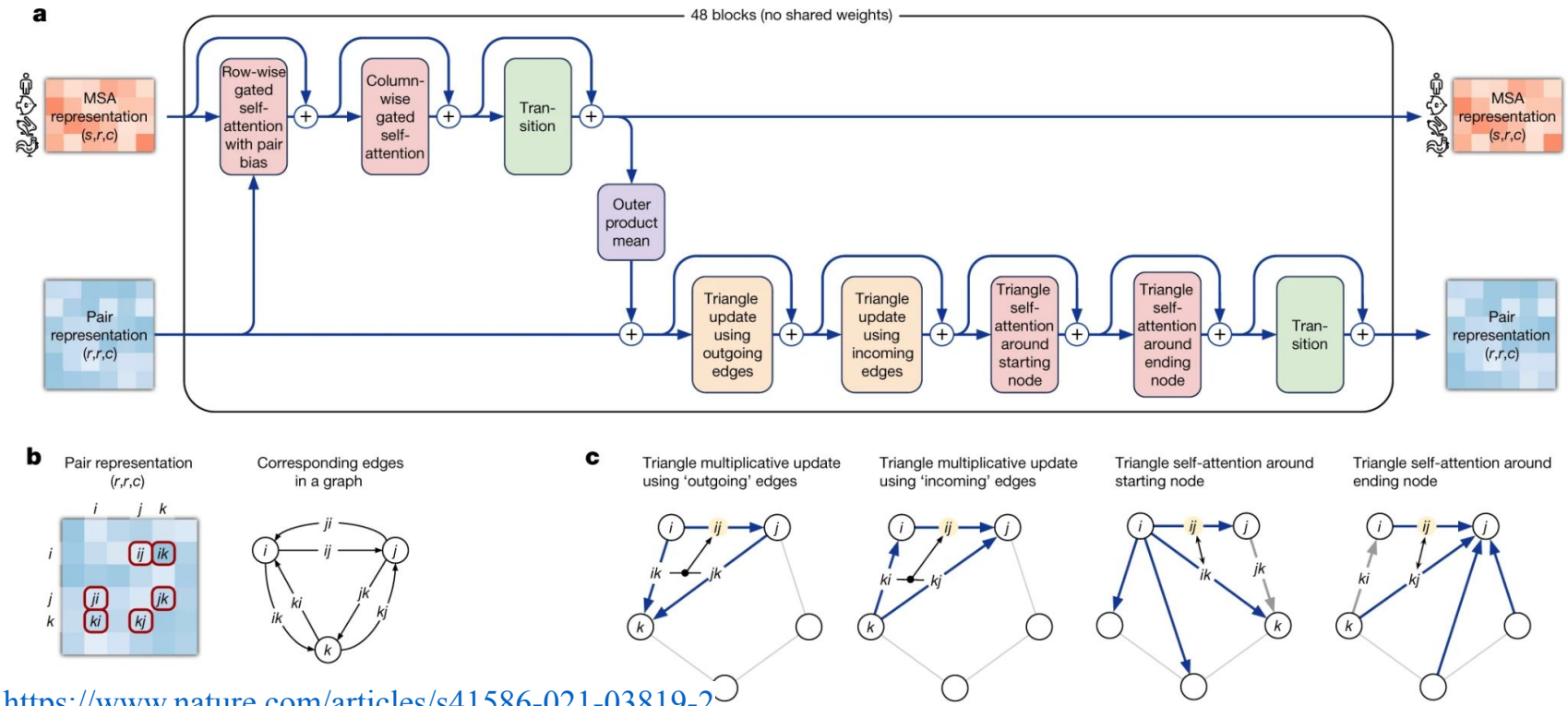
=> effective use of unlabelled sequence data

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

EvoFormer

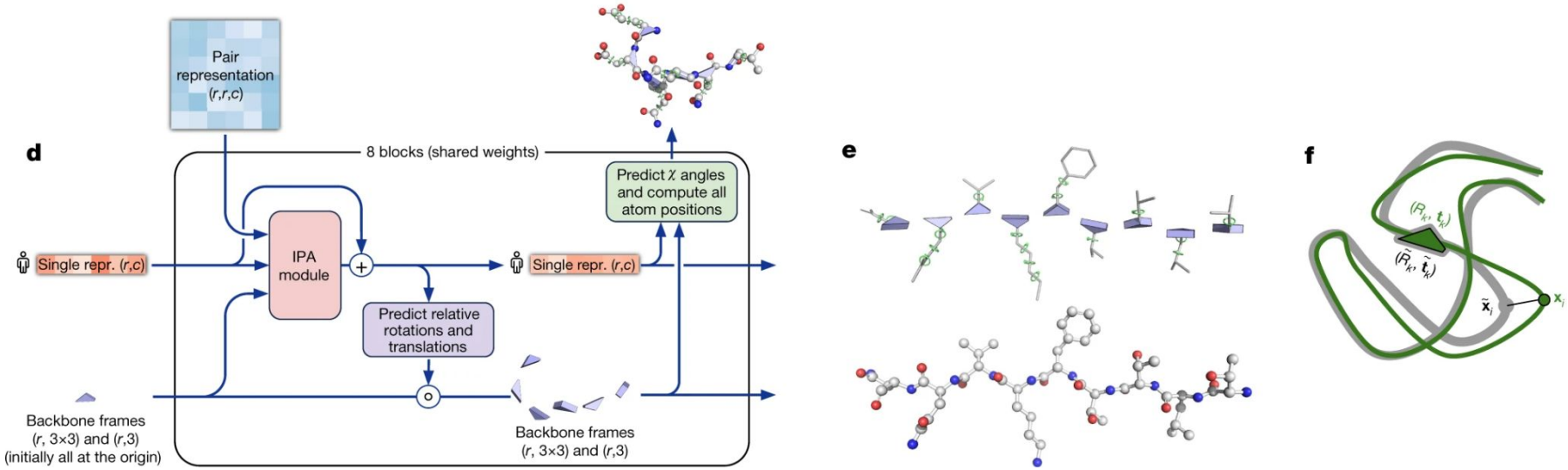
- mixing MSA and pairs via updates
- graph inference problem in 3D space
  - edges = residues in proximity
  - updates per each block (48 blocks) separately (AF1 updated all network at once)
- using triangles (instead of just pairs)



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

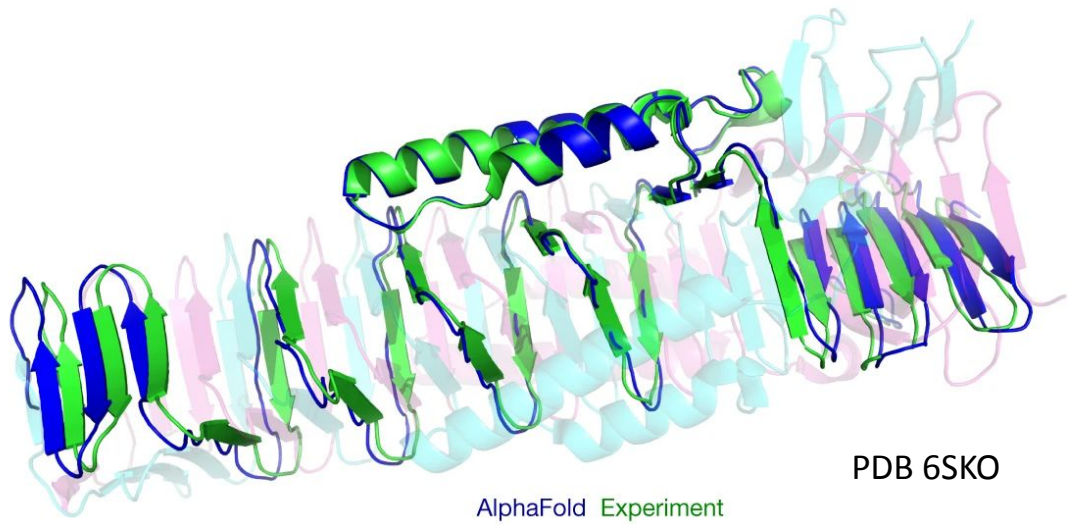


https://www.nature.com/articles/s41586-021-03819-2

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

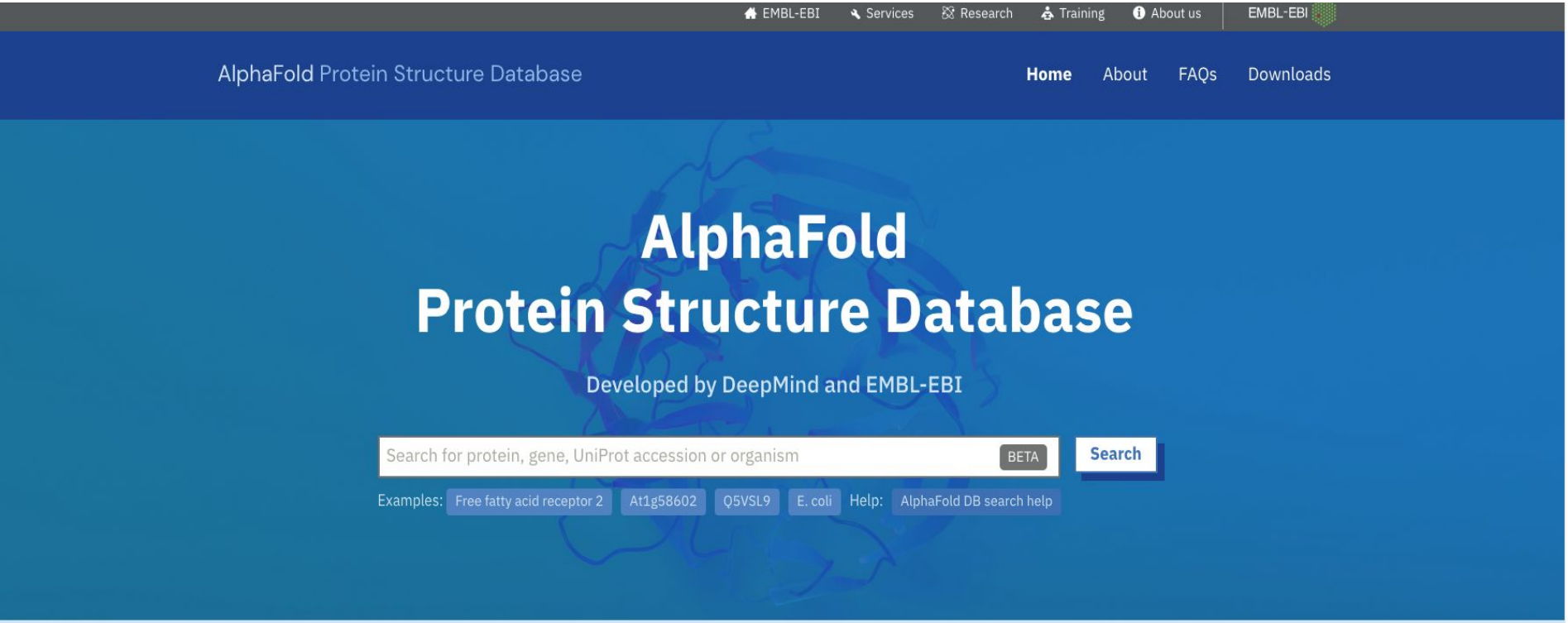
Timings

one GPU minute per model with 384 residues

=> allows proteome-scale studies

1500 residues trimer (SARS-CoV2 S protein) - about a day on ELIXIR CZ Metacentrum pipeline

AlphaFoldDB



AlphaFold DB provides open access to protein structure predictions for the human proteome and 20 other key organisms to accelerate scientific research.

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

elixir CZECH REPUBLIC

Complete structures of 20 model organisms

Species	Common Name	Reference Proteome	Predicted Structures	Download
Arabidopsis thaliana	*Arabidopsis*	UP000006548	27,434	Download (3642 MB)
Caenorhabditis elegans	Nematode worm	UP000001940	19,694	Download (2601 MB)
Candida albicans	*C. albicans*	UP000000559	5,974	Download (965 MB)
Danio rerio	Zebrafish	UP000000437	24,664	Download (4141 MB)
Dictyostelium discoideum	*Dictyostelium*	UP000002195	12,622	Download (2150 MB)
Drosophila melanogaster	Fruit fly	UP000000803	13,458	Download (2174 MB)
Escherichia coli	*E. coli*	UP000000625	4,363	Download (448 MB)
Glycine max	Soybean	UP000008827	55,799	Download (7142 MB)
Homo sapiens	Human	UP000005640	23,391	Download (4784 MB)
Leishmania infantum	*L. infantum*	UP000008153	7,924	Download (1481 MB)
Methanocaldococcus jannaschii	*M. jannaschii*	UP000000805	1,773	Download (171 MB)
Mus musculus	Mouse	UP000000589	21,615	Download (3547 MB)
Mycobacterium tuberculosis	*M. tuberculosis*	UP000001584	3,988	Download (421 MB)
Oryza sativa	Asian rice	UP000059680	43,649	Download (4416 MB)
Plasmodium falciparum	*P. falciparum*	UP000001450	5,187	Download (1132 MB)
Rattus norvegicus	Rat	UP000002494	21,272	Download (3404 MB)
Saccharomyces cerevisiae	Budding yeast	UP000002311	6,040	Download (960 MB)
Schizosaccharomyces pombe	Fission yeast	UP000002485	5,128	Download (776 MB)
Staphylococcus aureus	*S. aureus*	UP000008816	2,888	Download (268 MB)
Trypanosoma cruzi	*T. cruzi*	UP000002296	19,036	Download (2905 MB)

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

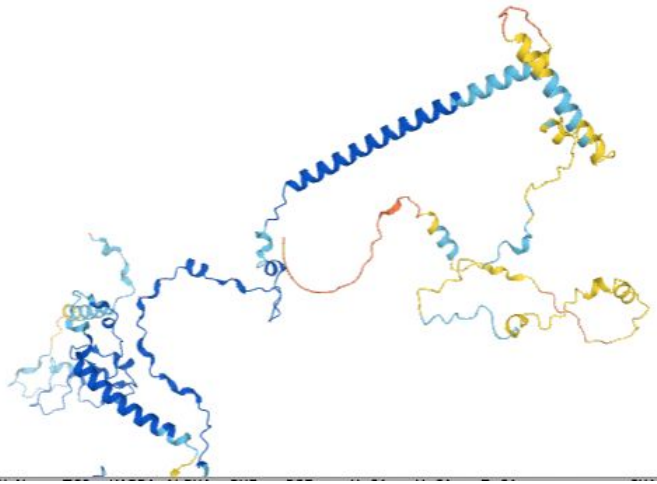
Model Confidence:

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

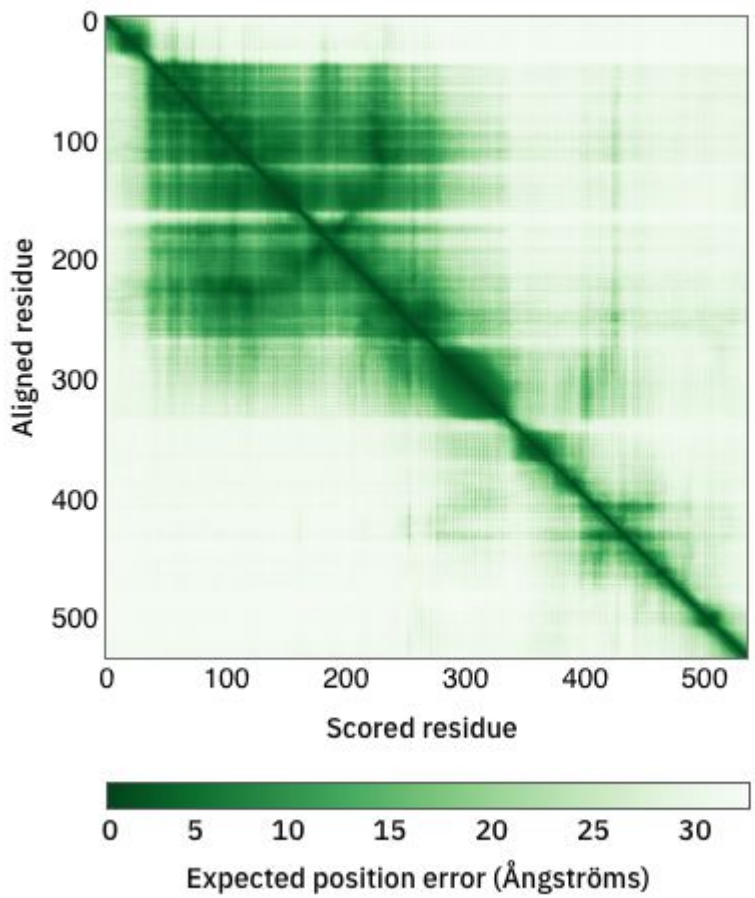
AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

Sequence of AF-Q13573-... 1: SNW do... A

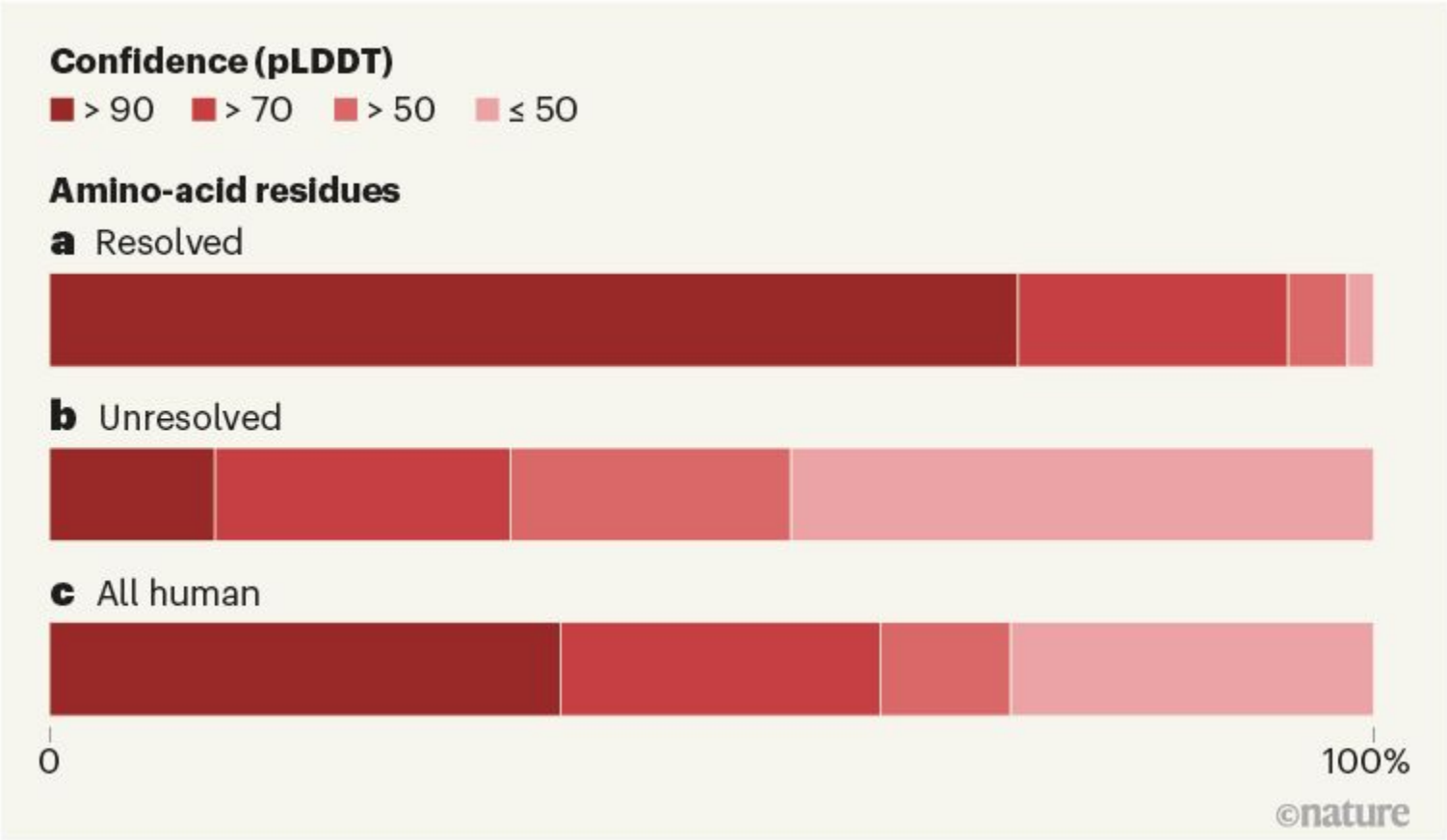
```
1         11        21        31        41        51        61        71        81        91       101       111       121
MALTSFLPAPTQLSQDQLEAEEKARSQRSRQTSLVSSRREPPPYGYRKGWIPRLLEDFGDGGAFPEIHVAQYPLDMGRKKKMSNALAIQVDSEGKIKYDAIARQGQSKDKVIYSKYTDLVPKEV
      131       141       151       161       171       181       191       201       211       221       231       241
MNADDPDLQRPDEEAIKEITEKTRVALEKSVSQKVAAAMPVRAADKLAPAQYIRYTPSQQGVAFNSGAKQRVIRMVEMQKDPMEPPRFKINKKIPRGPPSPPAPVMHSPSRKMTVKEQQEWKIP
  251       261       271       281       291       301       311       321       331       341       351       361       371
PCISNWKNAKGYTIPLDKRLAADGRGLQTVHINENFAKLAEALYIADRKAREAVEMRAQVERKMAQKEKEKHEEKLREMAQKARERRAGIKTHVEKEDGEARERDEIRHDRRKERQHDRNLSRA
```



Alphafold tells you where is it right!



How good are the predictions of human proteins?



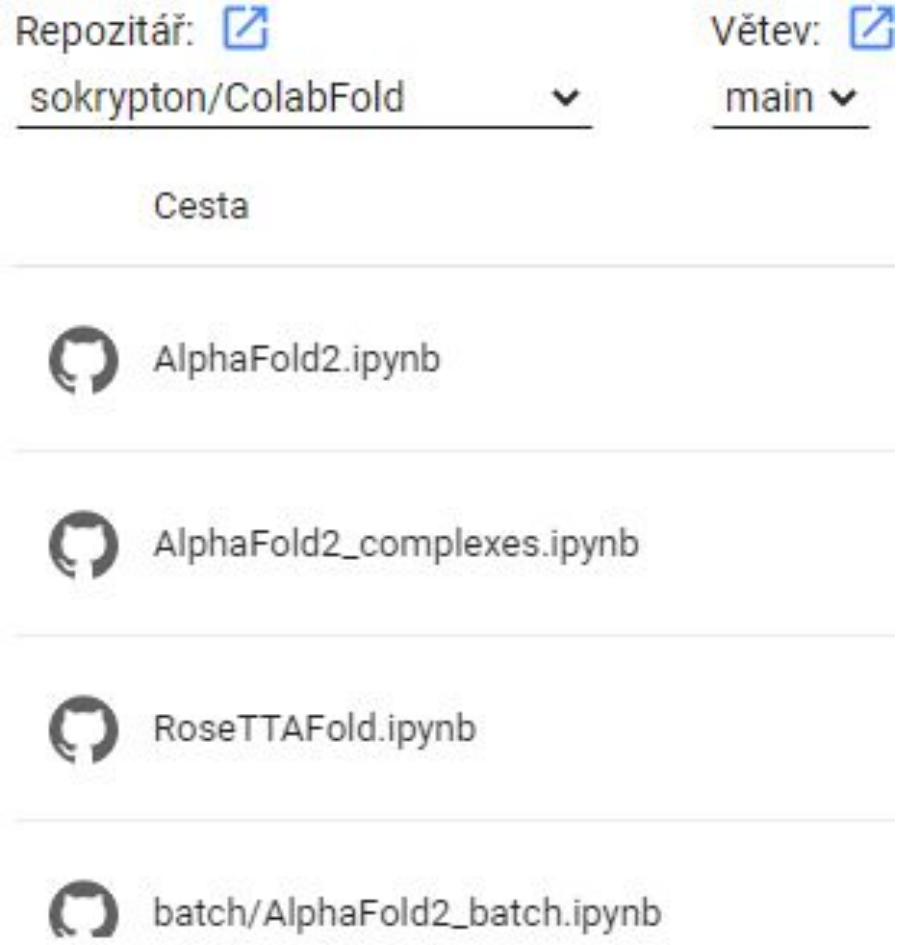
pLDDT - per-residue estimate of its confidence on a scale from 0 - 100
model's predicted score on the lDDT-Cα metric (local superposition-free score for comparing protein structures and models using distance difference tests).

Usages

AlphaFold in Google Colab

Github enabled JupyterNotebooks running in Google Colab environment

limitation in size



Mirdita M, Ovchinnikov S, Steinegger M. ColabFold - Making protein folding accessible to all. *bioRxiv*, 2021

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

Alphafold on ELIXIR CZ

- Alphafold needs GPU to run -> not many people have it on their PC
- Alphafold has been installed on Elixir CZ hardware
- Elixir is accessible through Metacentrum
- speed is dependent on size of predicted protein but can be in order of tens of minutes

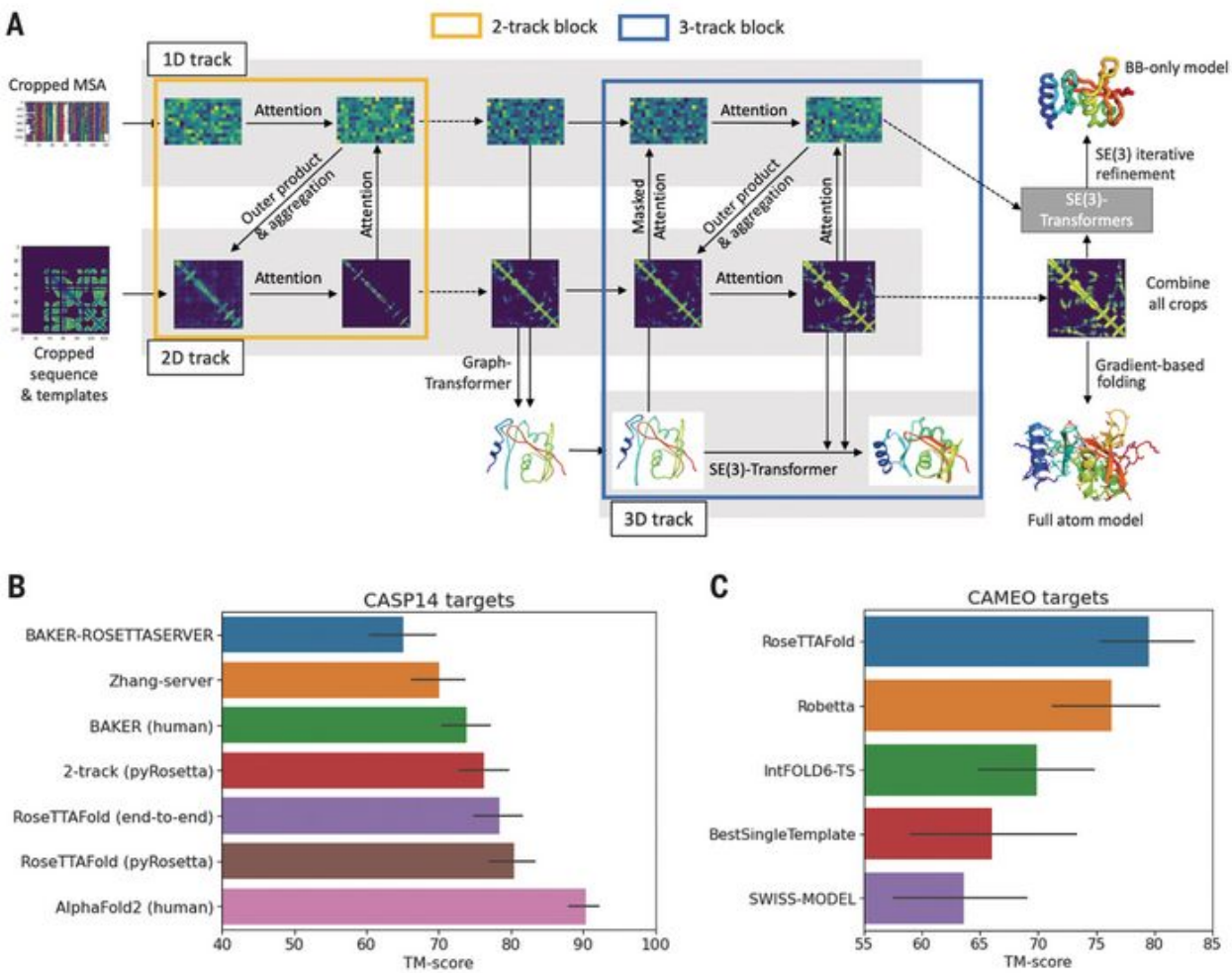
Alphafold is just a start...

- use Alphafold ideas for development of their own 3D structure predictions - RoseTTAfold
  - prediction of designed proteins
- prediction of RNA structures
- prediction of orphan proteins
  - molecular replacement
  - interpretation of cryoEM
  - pLDDT can act as IDP predictor

...

as of 9.9.2021

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



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

USING ALPHAFOLD FOR RAPID AND ACCURATE FIXED BACKBONE PROTEIN DESIGN

Lewis Moffat
Department of Computer Science
University College London
Gower St, London WC1E 6BT
lewis.moffat@cs.ucl.ac.uk

Joe G. Greener
Department of Computer Science
University College London
Gower St, London WC1E 6BT
j.greener@ucl.ac.uk

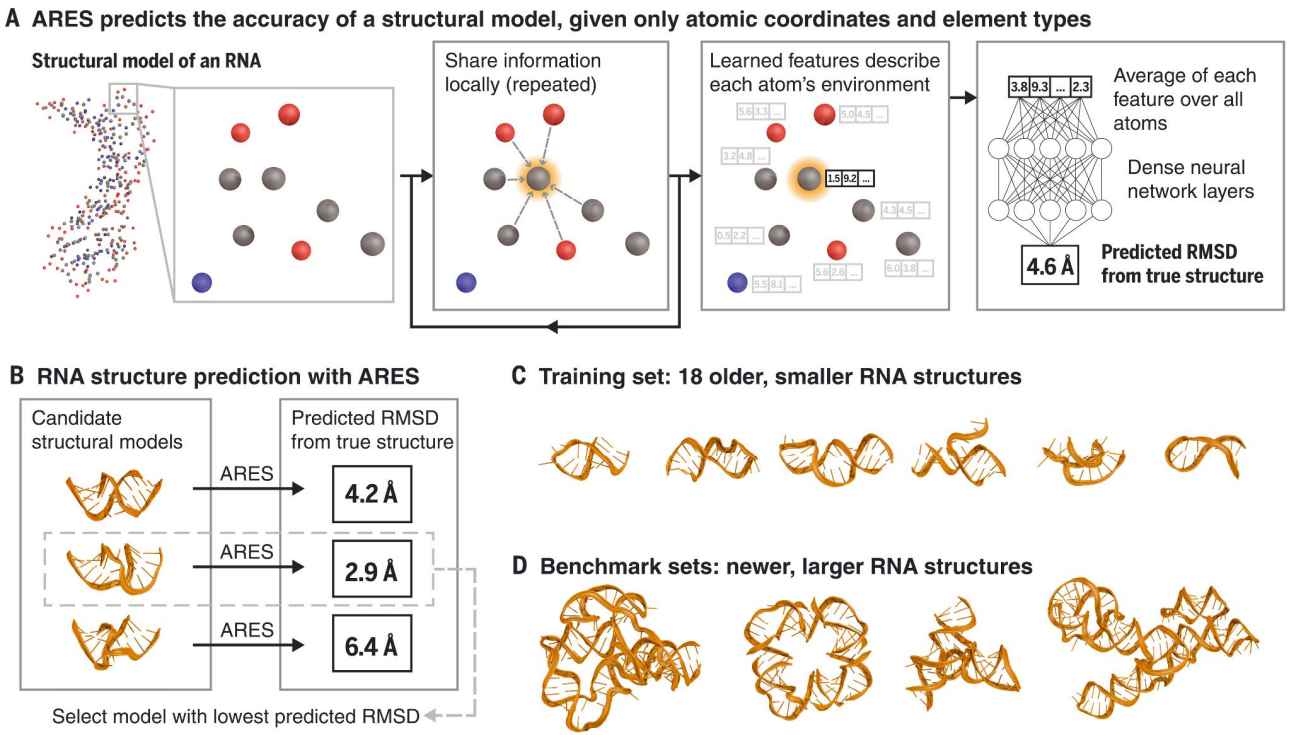
David T. Jones*
Department of Computer Science
University College London
Gower St, London WC1E 6BT
d.t.jones@ucl.ac.uk

ABSTRACT

The prediction of protein structure and the design of novel protein sequences and structures have long been intertwined. The recently released AlphaFold has heralded a new generation of accurate protein structure prediction, but the extent to which this affects protein design stands yet unexplored. Here we develop a rapid and effective approach for fixed backbone computational protein design, leveraging the predictive power of AlphaFold. For several designs we demonstrate that not only are the AlphaFold predicted structures in agreement with the desired backbones, but they are also supported by the structure predictions of other supervised methods as well as *ab initio* folding. These results suggest that AlphaFold, and methods like it, are able to facilitate the development of a new range of novel and accurate protein design methodologies.

**To whom correspondence should be addressed*

Geometric deep learning of RNA structure



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

Single-sequence protein structure prediction using language models from deep learning

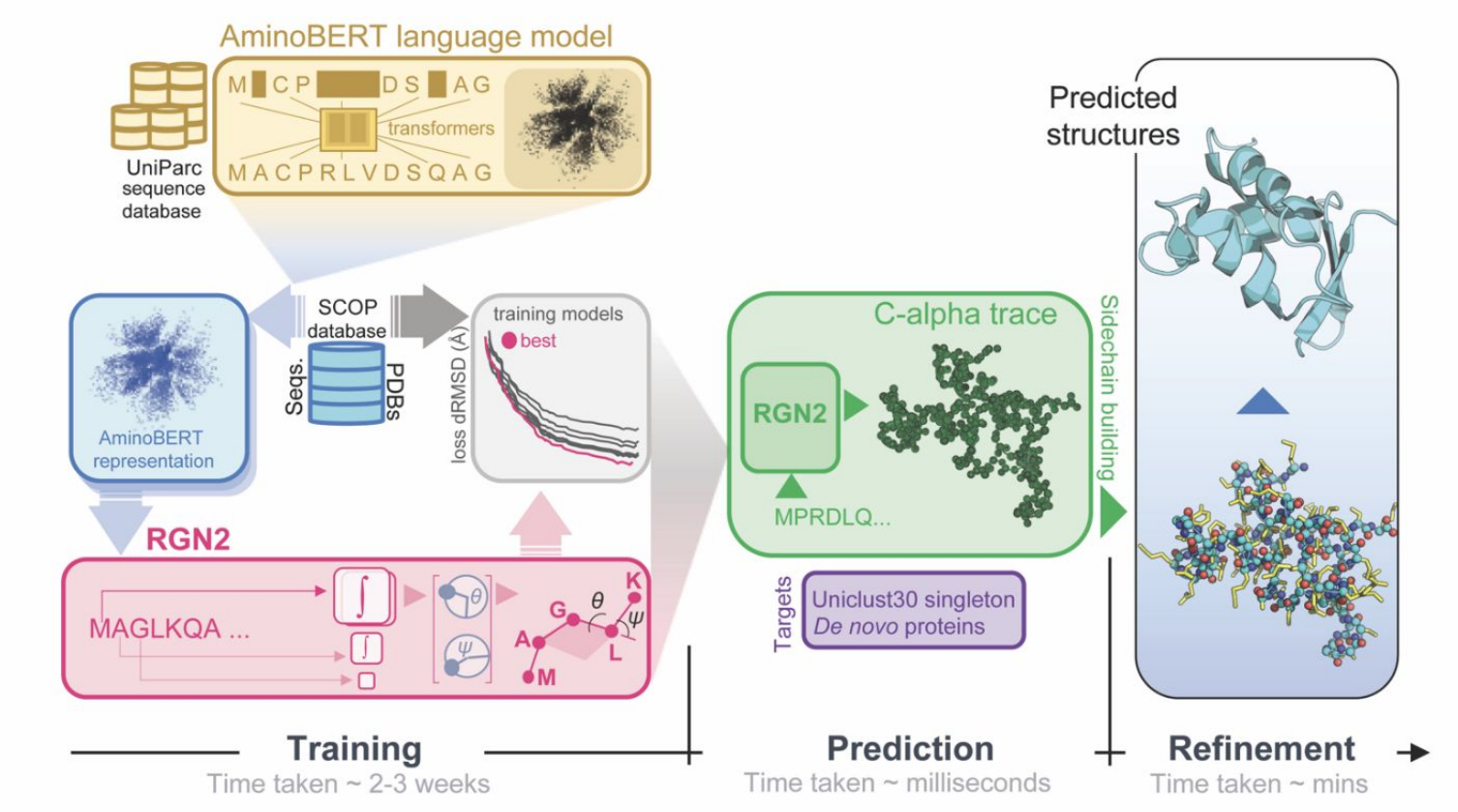
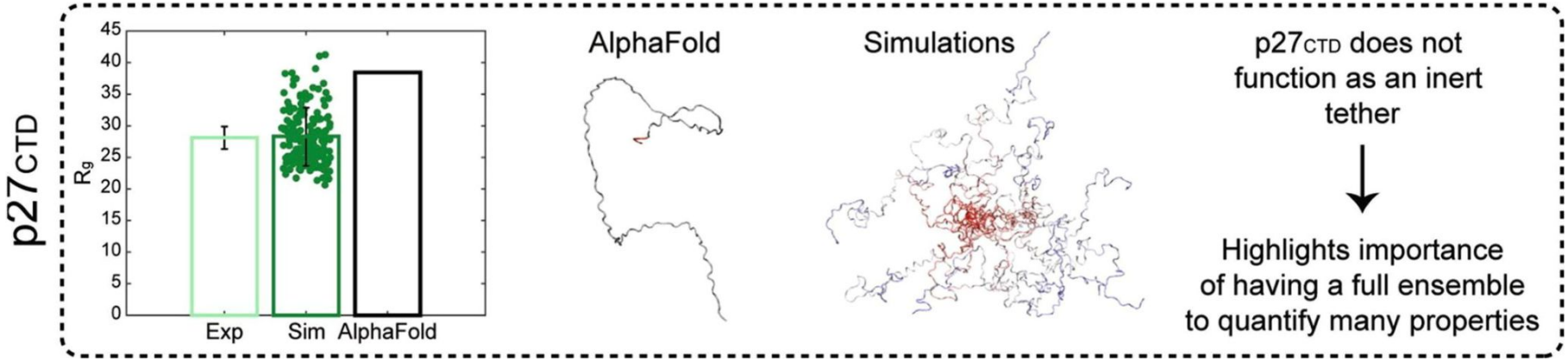
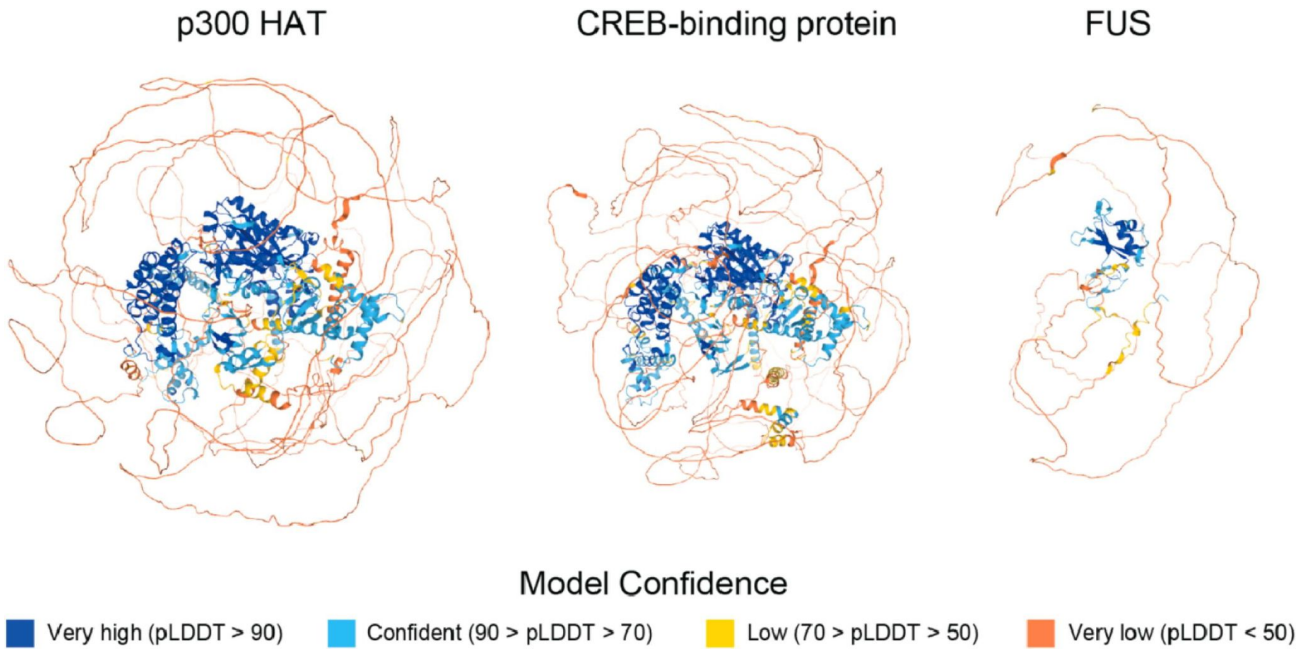


Figure 1. Organization and application of RGN2. RGN2 combines a Transformer-based protein language model (AminoBERT) with a recurrent geometric network that utilizes Frenet-Serret frames to generate the backbone structure of a protein. Placement of side chain atoms and refinement of hydrogen-bonded networks are subsequently performed using the Rosetta energy function.

https://www.youtube.com/watch?v=eobc7cMMpeY&feature=youtu.be

AlphaFold and Implications for Intrinsically Disordered Proteins





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

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

MrParse

MrParse Analysis

Version: 0.2.1

MrParse: a program to find and analyse search models for crystallographic Molecular Replacement. The program is being developed by Dan Rigden's group at the University of Liverpool.

MrParse is currently under development and we are keen to make it as useful to the community as possible. If you have any suggestions for it's development, or ideas on how we could improve it, please get in touch.

IKL Info

Name	Resolution	Space Group	Has NCS?	Has Twinning?	Has Anisotropy?
7dry-sf	1.44	P41212	false	false	true

xperimental structures from the PDB

Name	PDB	Resolution	Region	Range	Length	eLLG	Mol. Wt.	eRMSD	Seq. Ident.
2cvi_B_1	2cvi	1.50	1	158-230	71	43.5	8676	1.085	0.31

Visualisation of Regions

Sequence Based Predictions

tructure predictions from the EBI AlphaFold database

Name	model	Date Made	Region	Range	Length	Avg. pLDDT	H-score	Seq. Ident.
Q12362_1	Q12362	01-JUL-21	1	2-180	177	90.15	85	0.41
P87241_1	P87241	01-JUL-21	1	4-176	171	91.55	85	0.38

Visualisation of Regions

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

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

Limitations

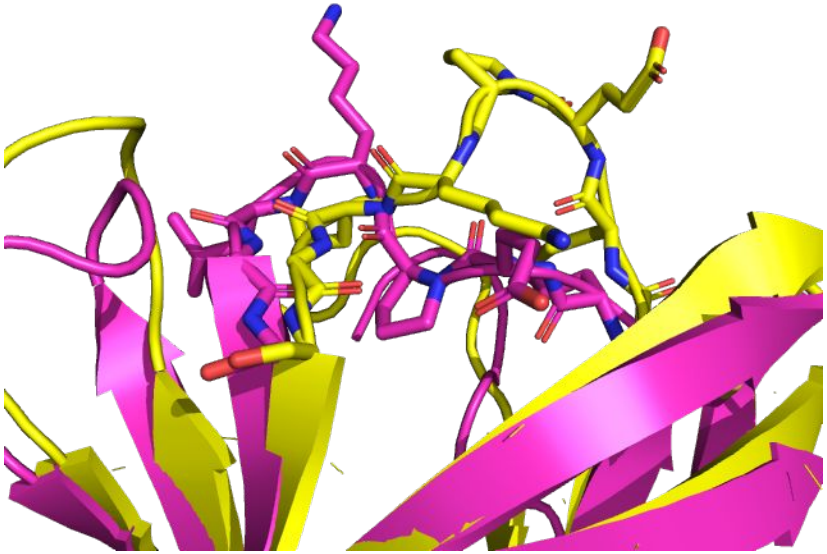
Are structural biologists and bioinformaticians on the job market?

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

Are the models good enough for drug design?

- we do not know yet
- average RMSD for Alphafold2 models is 1.3 Å
- average RMSD of X-Ray structures is 0.3 - 0.5 Å
- best Alphafold prediction has RMSD 0.6 Å

- locally AlphaFold2 might be there



T1064

elixir CZECH REPUBLIC

Summary

- Alphafold2 made a huge leap in prediction accuracy
- Role of open science and publicly available data 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 tools already
- Alphafold limits are yet to be fully described

Thank you for your attention.

Any questions?

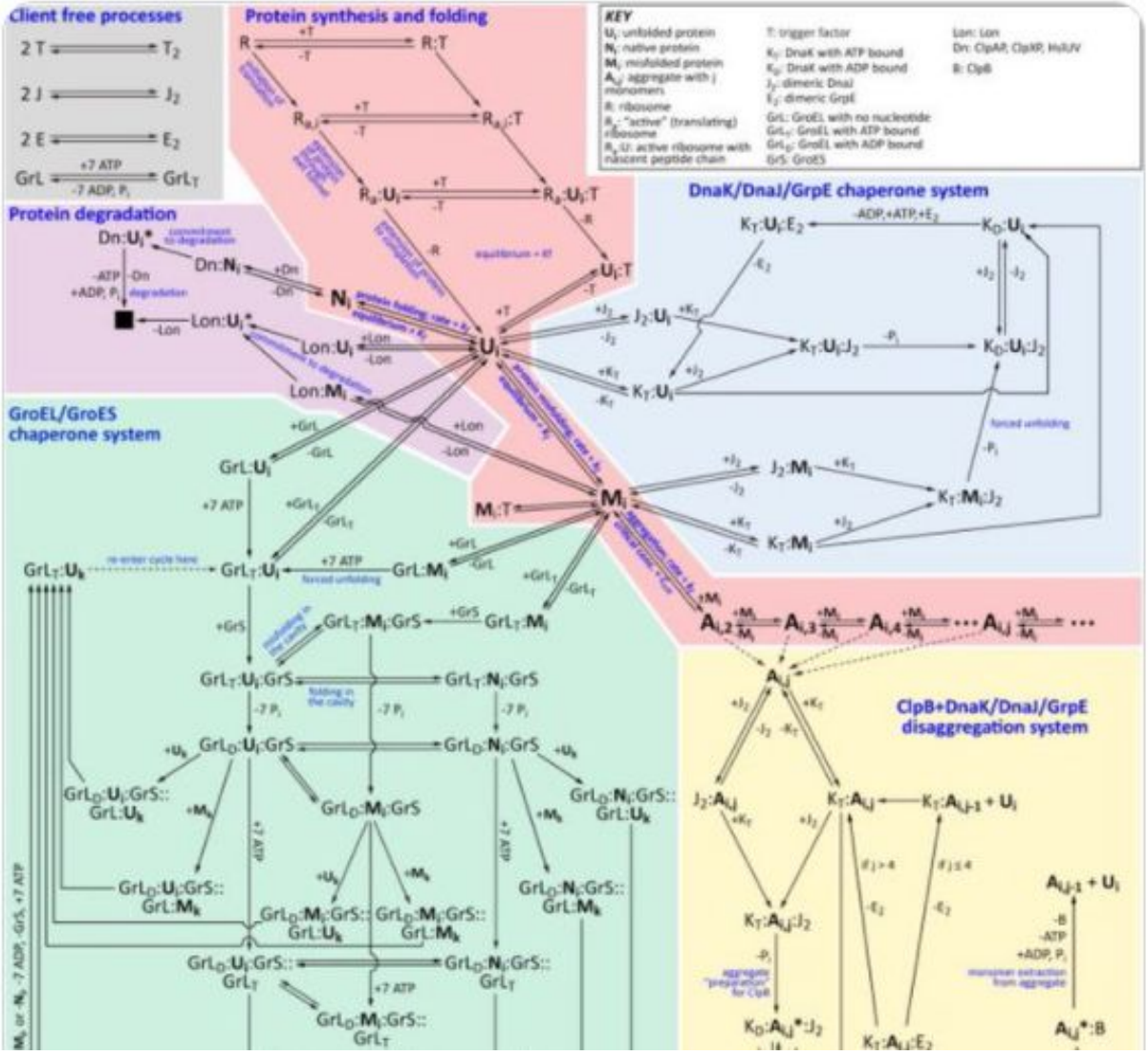
MINISTRY OF EDUCATION, YOUTH AND SPORTS

elixir CZECH REPUBLIC

Kresten Lindorff-Larsen
@LindorffLarsen

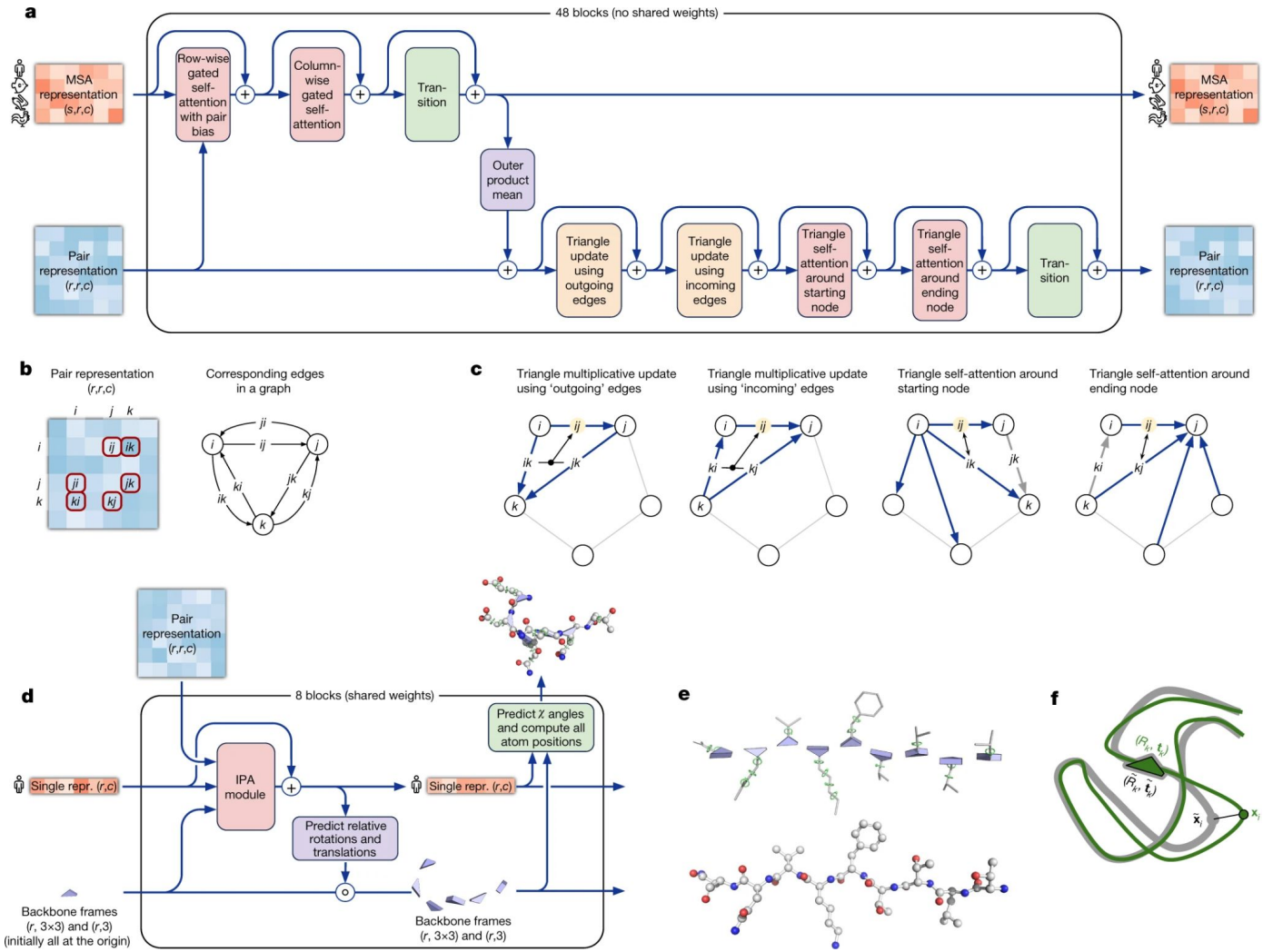
Tell me again how the folding problem has been solved
doi.org/10.1016/j.jmb.... doi.org/10.1016/j.celr...

Přeložit Tweet

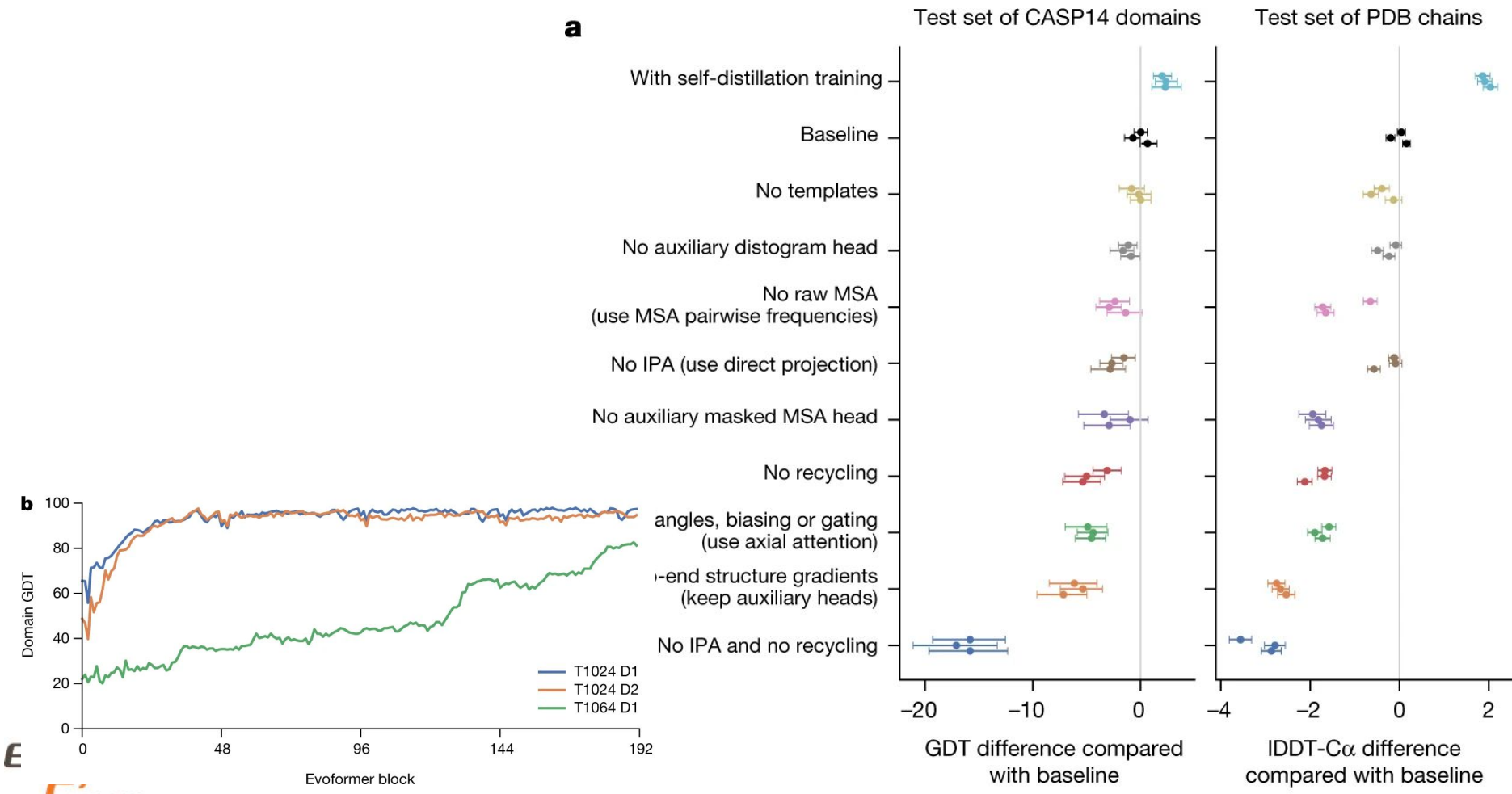


Extra slides

Architectural details.



Interpreting the neural network



CZECH REPUBLIC

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