KFC/STBI
Structural Bioinformatics

Structure Alignment and Prediction

Karel Berka

Outline

- Structural alignment
- Structure prediction
  - Homology modelling
    - SwissMODEL, Modeller,
  - threading
    - I-TASSER
  - de novo modelling
    - Robbeta, Quark
  - molecular mechanics
    - protein folding
    - Folding@Home, FoldIt
  - Evolutionary coupling
    - EVcoupling
  - Machine learning
    - AlphaFold
    - RosettaFoldAA

Structural Alignment

Are Those Structures Similar?

- By eye
- By algorithm:
  - Structural alignment

Structural alignment of thioredoxins from humans (red) and Drosophila (yellow) PDBID: 3TRX and 1XWC.

Structural Alignment

- To find **the best** pairing between two structures
- *The best*
  -> “smallest RMSD”
- Problem:
  - Quite oftenly it is possible to find only subset of dissimilar atoms – how to discern them?

Structural Alignment

- Other problems:
  - Nr of aminoacids in both chains
  - fit and RMSD calculation
  - Identity between „aligned residues“
  - Nr of „gaps“
  - Size of proteins
  - Conserved sequence sites

- There are no universal criteria

Structural alignment

- Warning:
- It is different to RMSD calculation -
- there is not easy correspondance between atoms. => Z Analysis of all possible correspondences
- RMSD is just tool to analyse similarity

Why to use structural alignment?

Structure is usually more conserved than sequence (there is smaller number of structural fold families in contrast with sequence clusters)

1. Homologous proteins (same ancestor)
   - „gold standard“ for sequence alignment
2. Nonhomologous proteins
   - similar substructures (domains)
3. Classification to clusters
   - structural similarities (CATH)
   - sequence similarities (Pfam)

CATH vz Pfam

http://www.cathdb.info/

Latest Release Statistics

Info

	CATH v4.1	CATH-B
PDB Release	01-01-2015	about 7 hours ago
Domains	308999	417837
Superfamilies	2737	6344
Annotated PDBs	108378	123625

	Gene3D v14
Cellular Genomes	19,471
Protein Sequences	43,387,462
CATH Domain Predictions	53,479,436

Sillitoe I, et al. **CATH: comprehensive structural and functional annotations for genome sequences.** *Nucleic Acids Res. 2015* doi: 10.1093/nar/gku947

http://pfam.xfam.org/

- **Pfam 30.0**
  July 1, 2016
- 16,306 families
  - 22 new and 11 killed families
- 17.7M sequences
  - 11.9M sequences in Pfam 29

R.D. Finn et al. *The Pfam protein families database: towards a more sustainable future*: , ***Nucleic Acids Research*** (2016) 44:D279-D285

Types of Structural Alignment

point methods

- **CE (**Combinatorial Extension)

  rigid structures, start – largest group of sequentially equivalent atoms
- **DALI** (Holm, Sander)

  matrix of distances to search for similar patterns to fit correspondences between atoms (without sequence)

secondary structure methods

- **VAST**

  alignment of secondary structures
- **FATCAT**

  protein is not rigid - hinges can bend

CE (Combinatorial Extension)

- pairs of protein segments by 8 AA
- comparison by local geometry
- pairs are further enlarged
- results:
  - RMSD
  - z-statistics

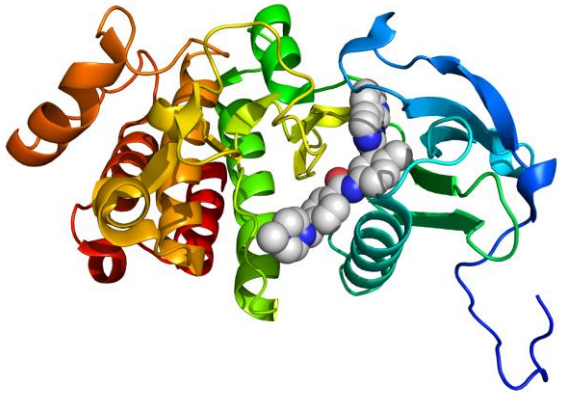
(standard) z-score is $$z = \frac{x - \mu}{\sigma}$$

where:

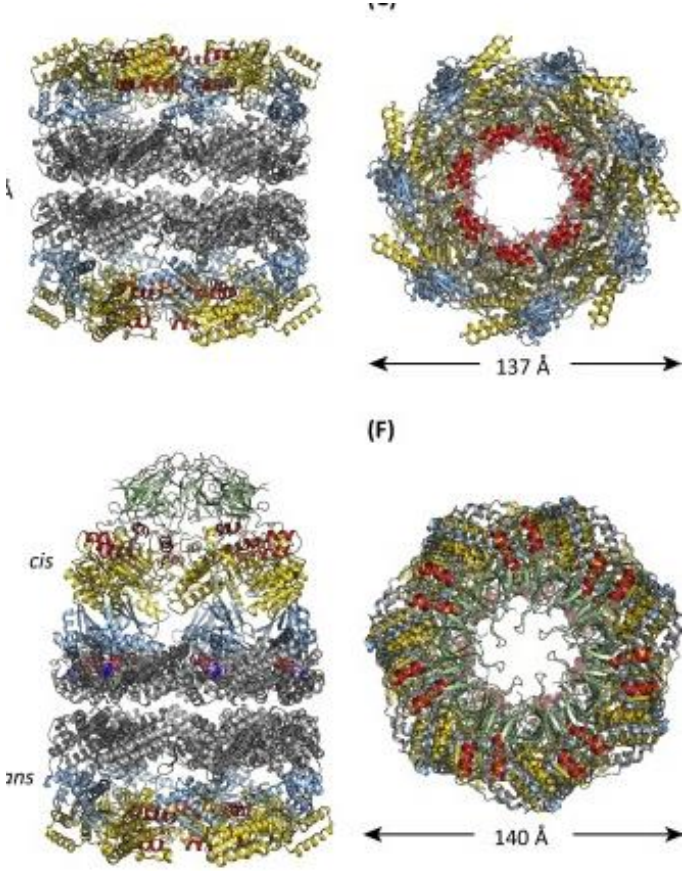
x is a raw score to be standardized;
μ is the mean of the population;
σ is the standard deviation of the populat

Structure Prediction

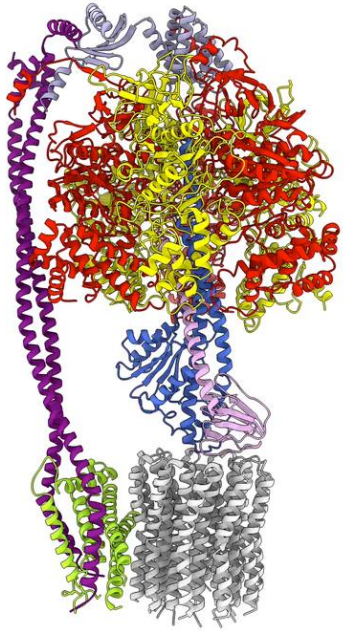
Knowing structure helps to understand the function



wikipedia



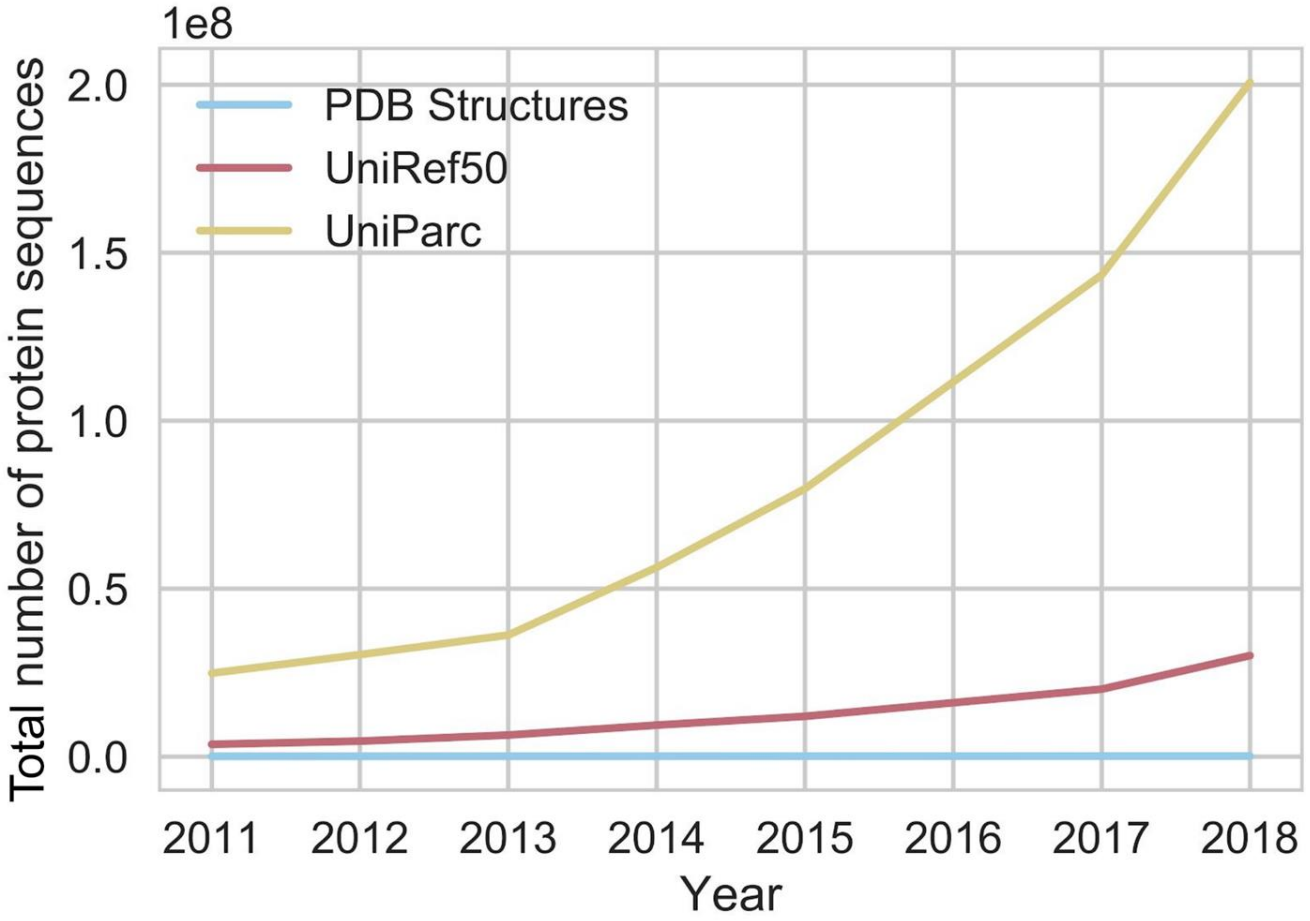
Hayer-Hartl et al., 2015



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

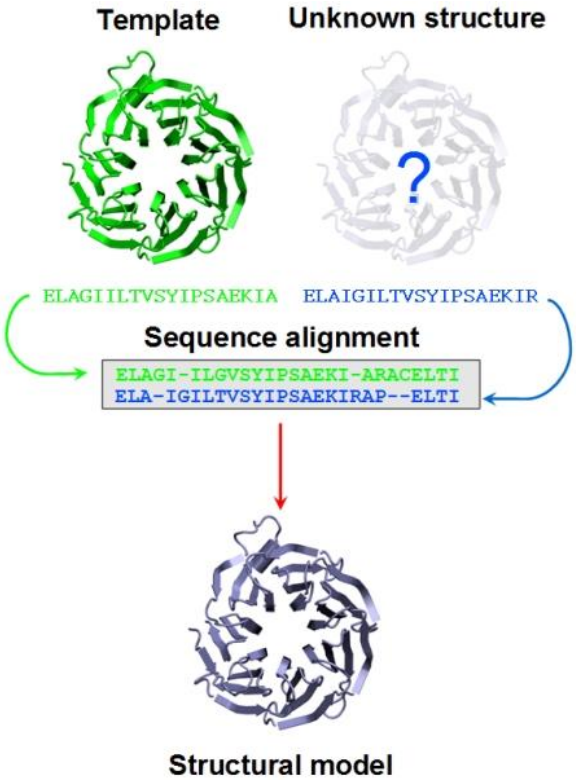
- C.B. Anfinsen received Nobel prize in Chemistry (1972) for describing the relationship between sequence and structure

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

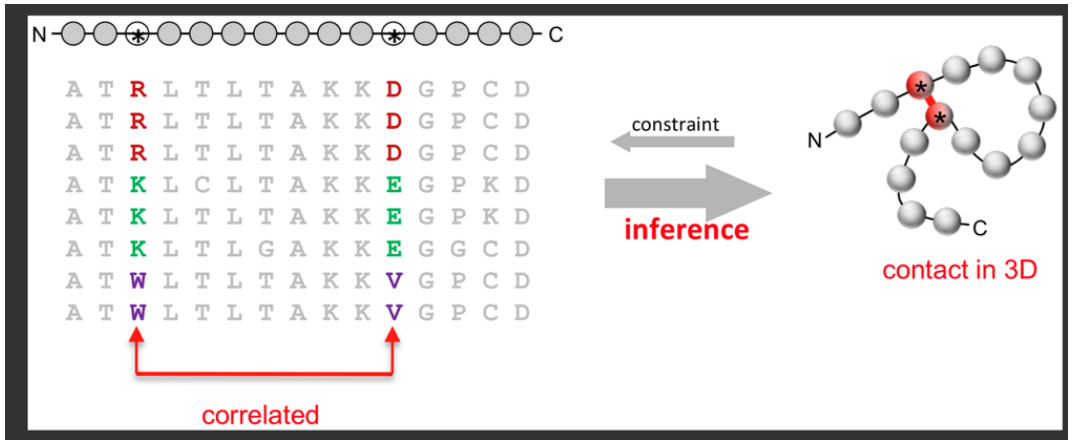
- it shall be possible to give to **predict structure from sequence**

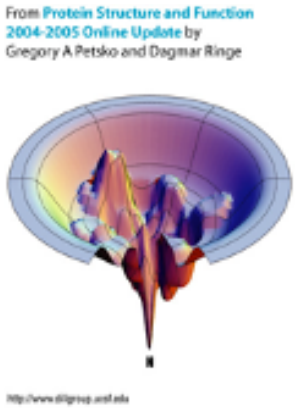
ribonuclease

Principles of prediction from sequence



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





Structure prediction = simulation of protein folding? Nope

Levinthal's paradox

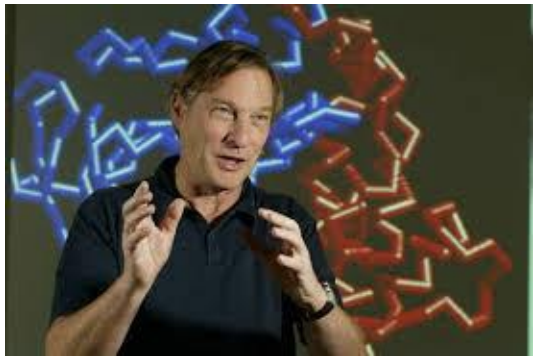
- protein of 100 aa has 10^{70} available conformations

-> 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



John Moult - father of CASP

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

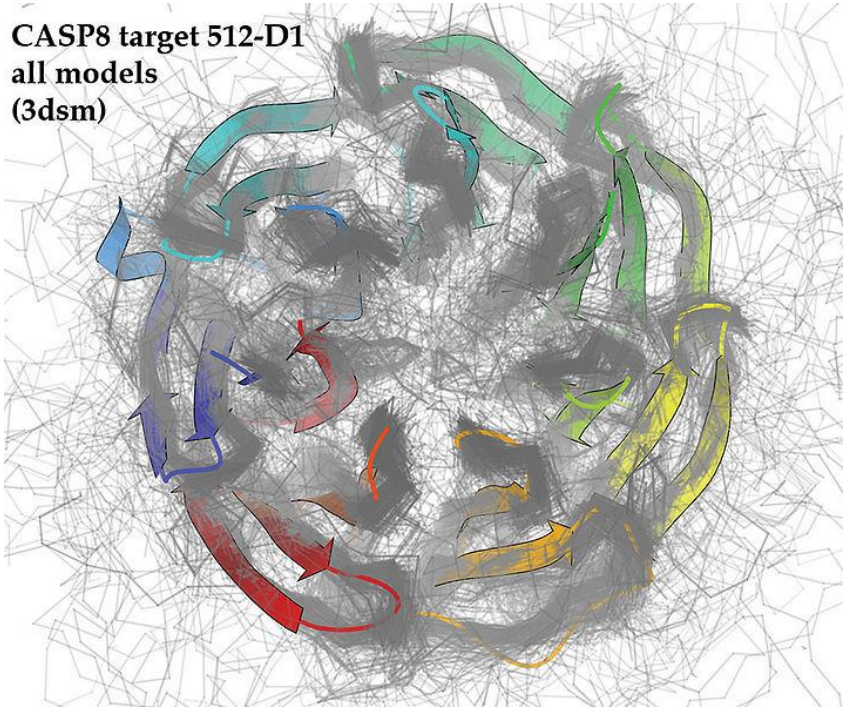
elixir CZECH REPUBLIC

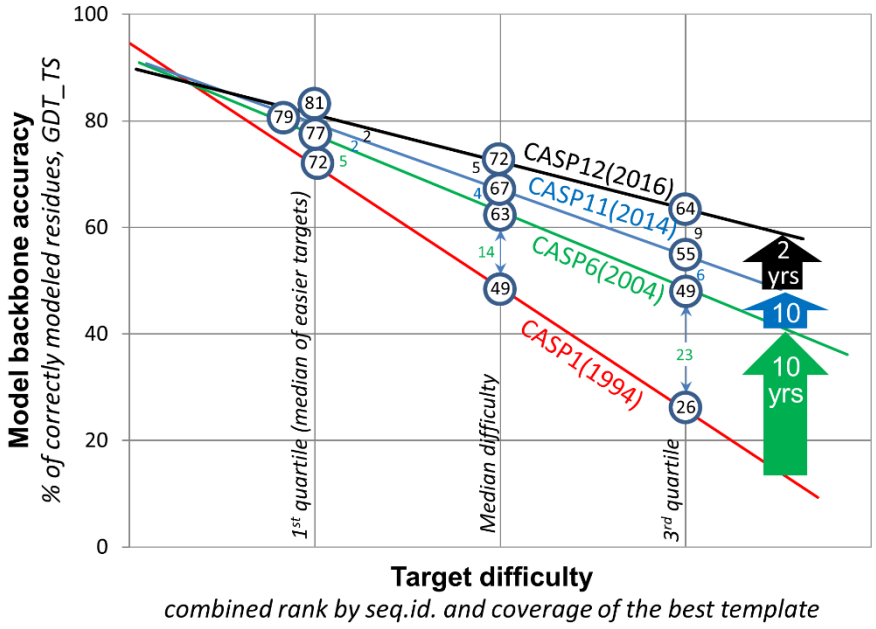
CASP

Critical Assessment of protein Structure Prediction

CASP

- **Critical Assessment of Techniques for Protein Structure Prediction**
- Comparison with prepublished x-ray data
- no prior information for predictors (double-blind)



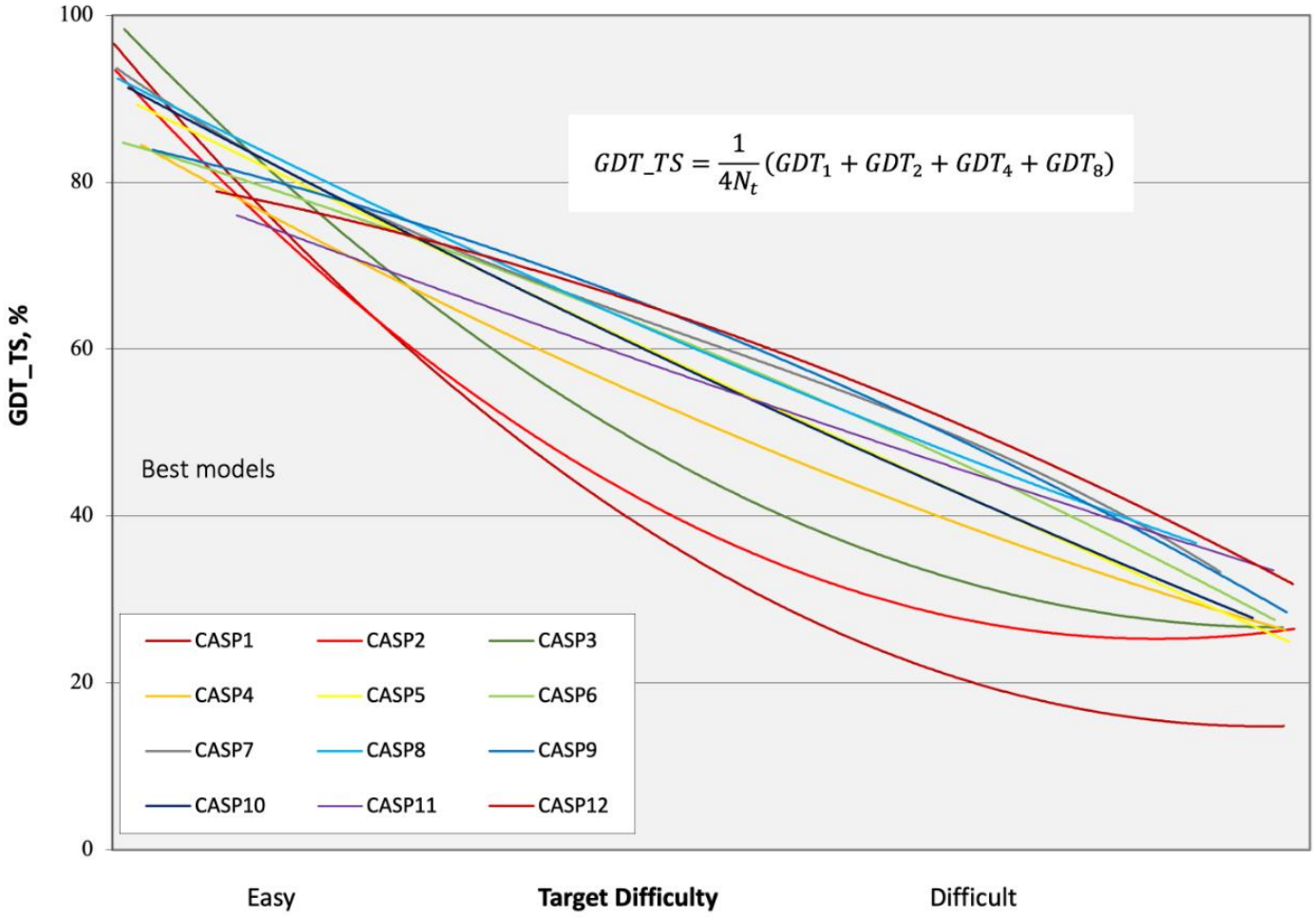


http://predictioncenter.org/casp9/index.cgi
Proteins: Structure, Function, and Bioinformatics Volume 77, Issue S9, Pages 1-228 (2009)

CASP

- tertiary structure prediction (all CASPs)
- secondary structure prediction (dropped after CASP5)
- prediction of structure complexes (CASP2 only; a separate experiment - CAPRI - carries on this subject)
- residue-residue contact prediction (starting CASP4)
- disordered regions prediction (starting CASP5)
- domain boundary prediction (CASP6-CASP8)
- function prediction (starting CASP6)
- model quality assessment (starting CASP7)
- model refinement (starting CASP7)
- high-accuracy template-based prediction (starting CASP7)

How to compare structures?

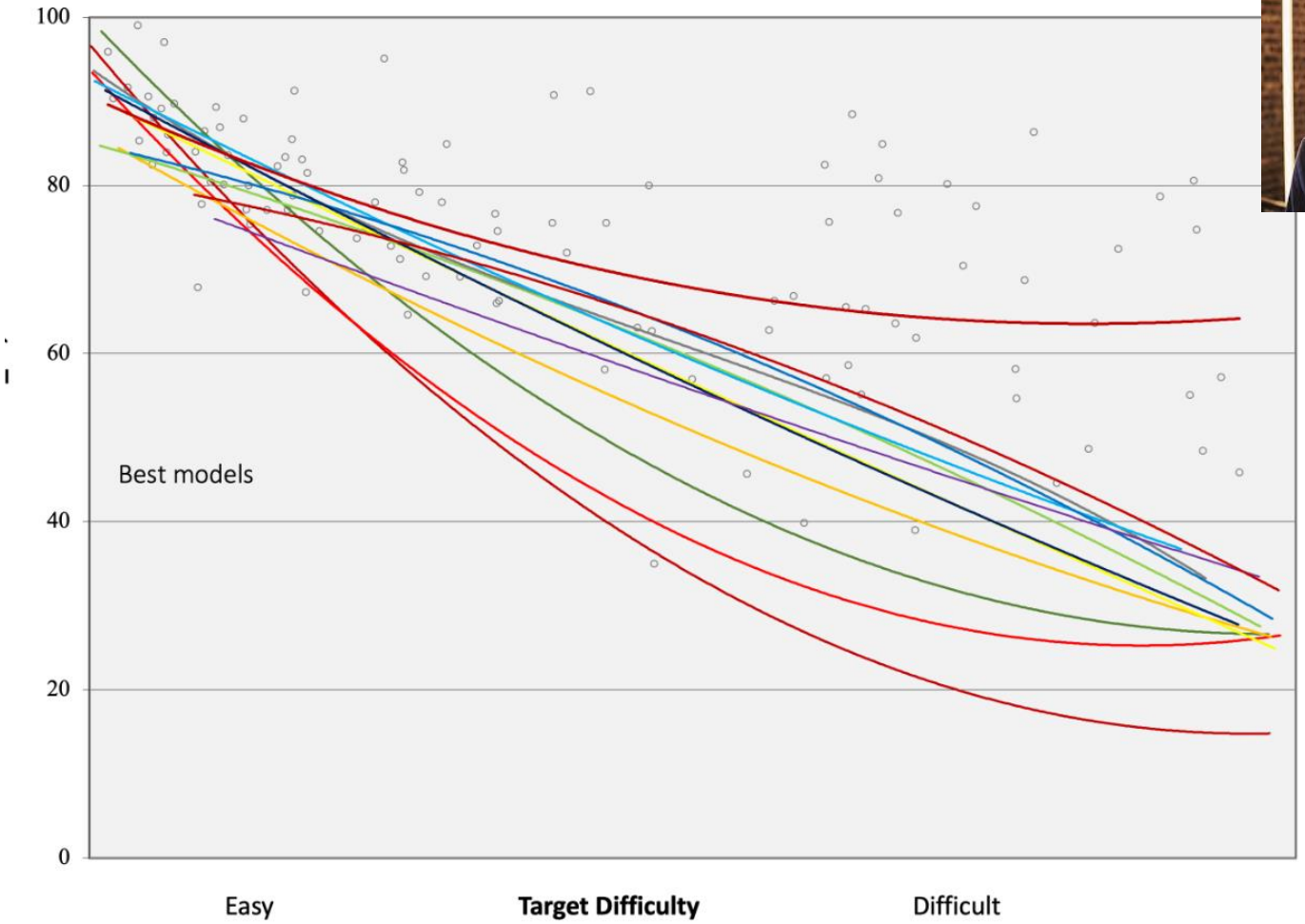


https://predictioncenter.org/casp14/doc/presentations/2020_11_30_CASP14_Introduction_Moult.pdf

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

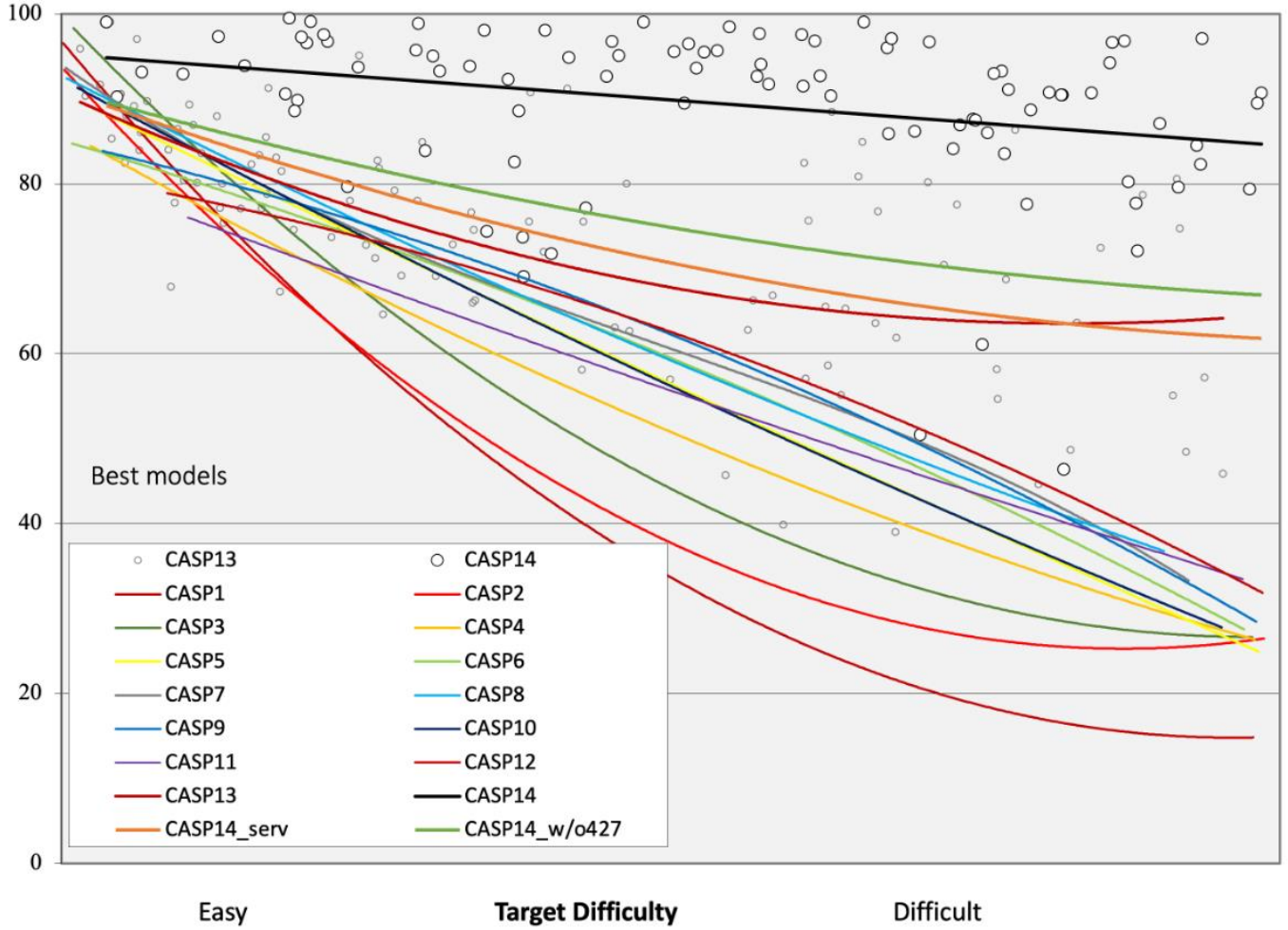
2018: AlphaFold enters...



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

elixir CZECH REPUBLIC

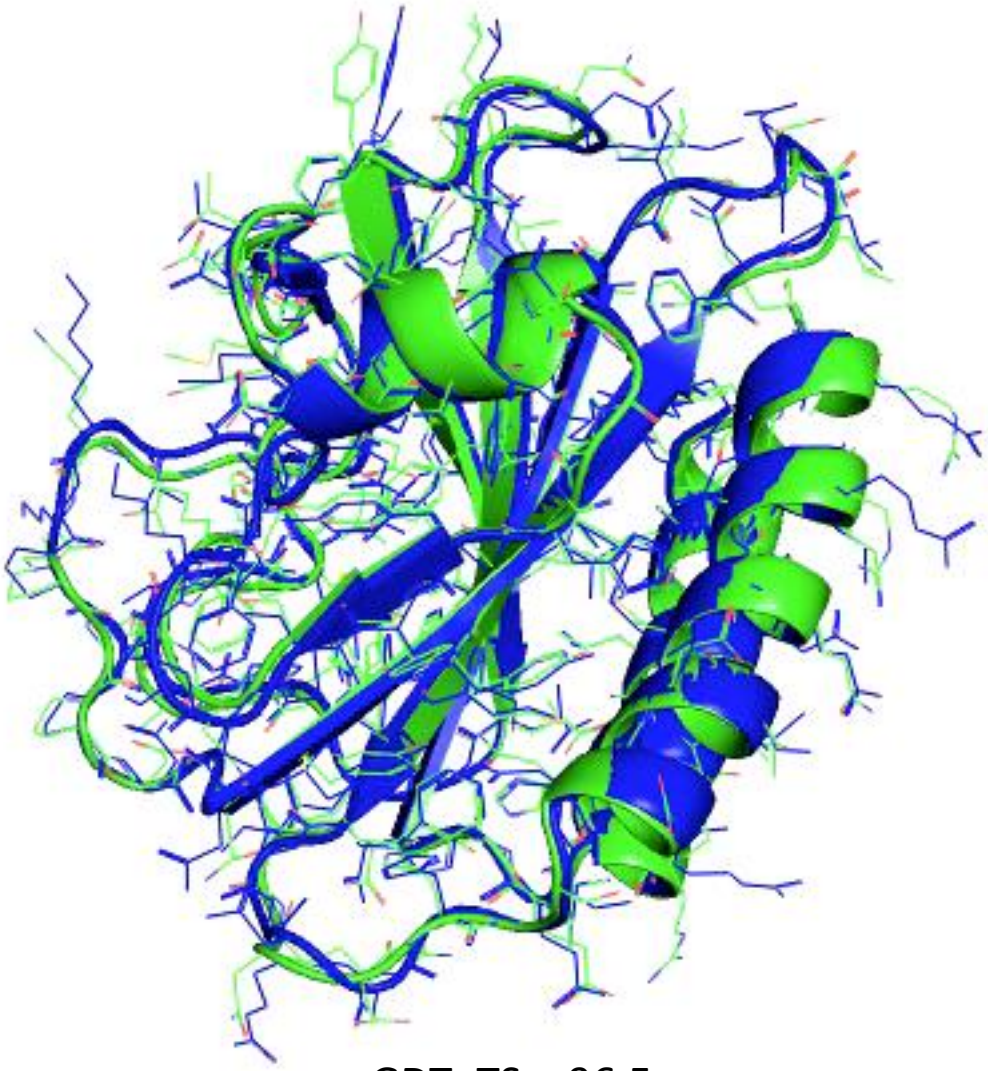
2020: AlphaFold2 wins



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

elixir CZECH REPUBLIC

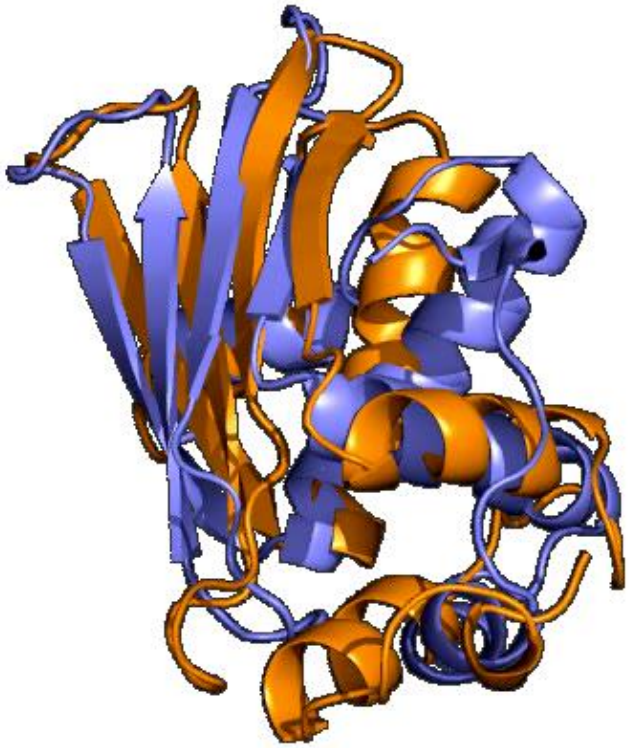
How does good prediction look like?



GDT_TS = 96.5

elixir CZECH REPUBLIC

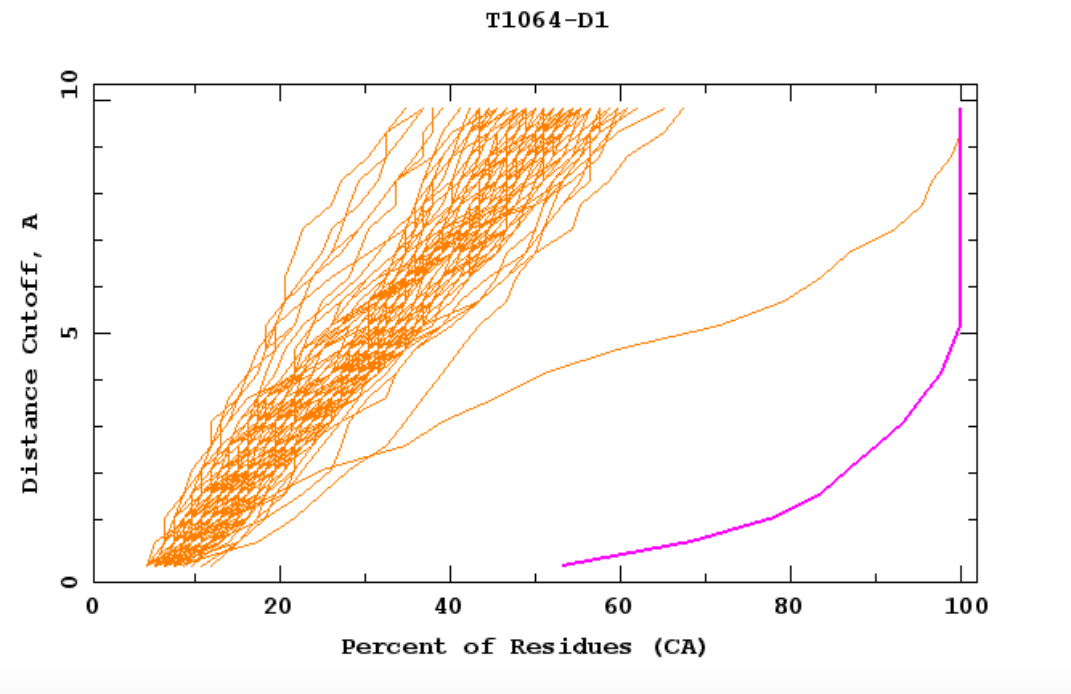
The worst prediction of AlphaFold 2 in CASP 14

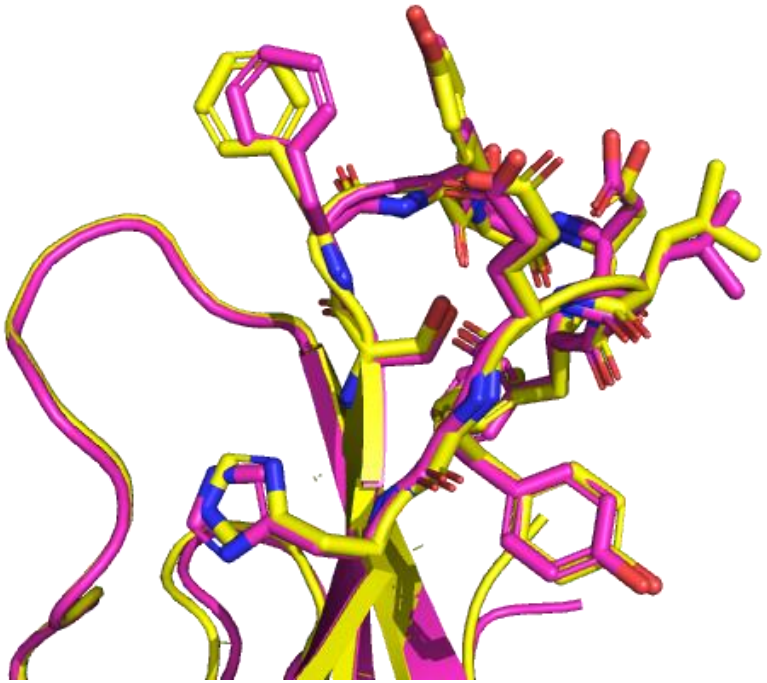


GDT_TS = 44.6

elixir CZECH REPUBLIC

Side chain predictions– orf8 covid19





GDT_TS= 87

so how it works?

Prediction of Protein Tertiary Structure

- Structure Prediction
  - from known structures
    - Homology modelling
      - SwissMODEL, I-TASSER
    - threading
      - Modeller
  - from physical models
    - de novo modelling (ab initio)
      - Quark, Robbeta,
      - protein folding
      - Folding@Home, FoldIt

- Protein evolution
  - Evolutionary coupling
- Machine learning
  - AlphaFold
  - Xfolds…

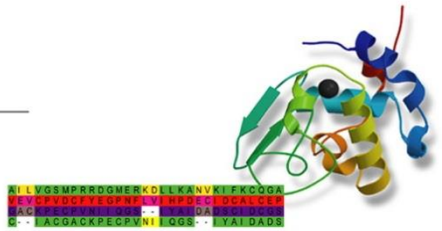
Homology modelling

- also known as comparative or knowledge-based modelling
- based on template structure

- Swiss-MODEL
  - http://www.expasy.org/spdbv/
- Modeller
  - http://salilab.org/modeller/

Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints

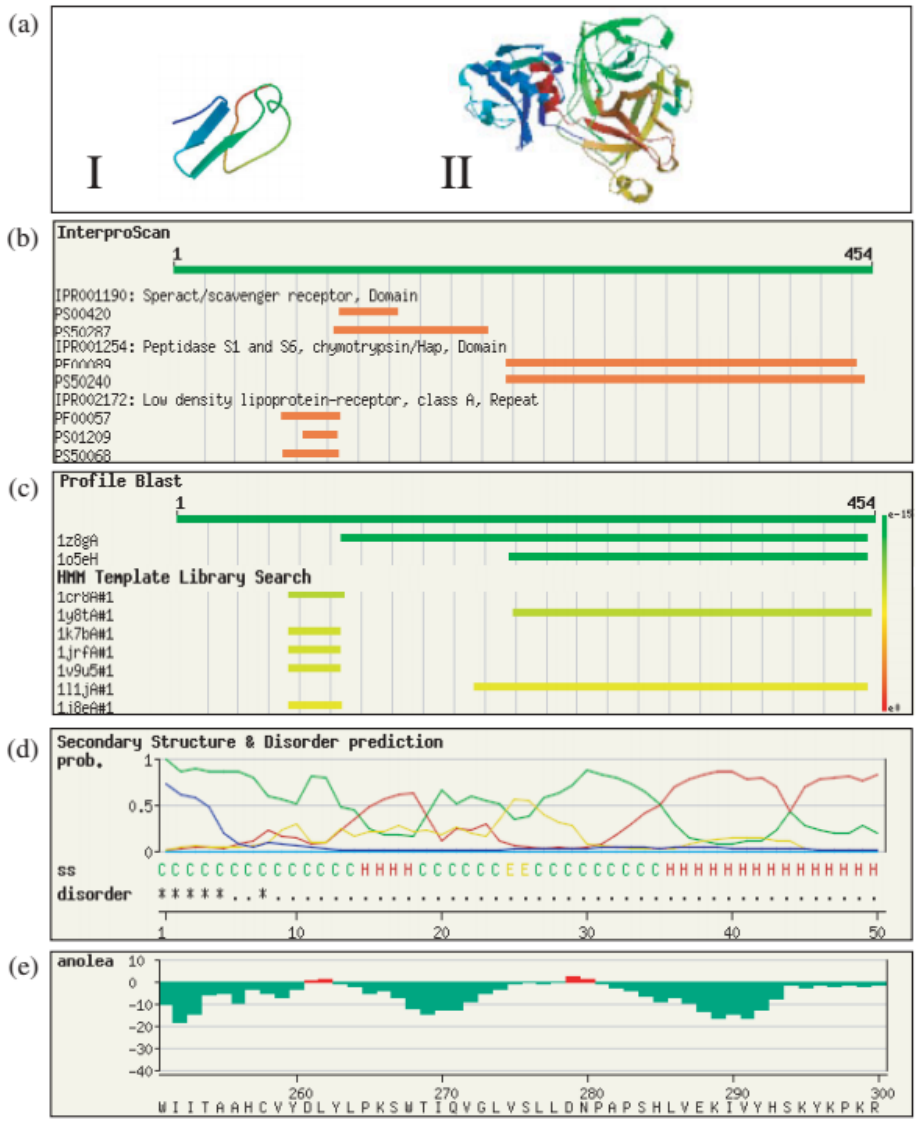


Typical protocol

- template selection
  - sequence alignment
- target-template alignment
  - pair comparisons (usually iterative refinement of alignment)
- model construction
  - main chain construction
  - loops
  - side chain construction -> rotamers
  - energy minimization
- model assessment
  - stereochemical control (PROCHECK, Ramachandran plot)
  - statistics – scoring function, z-score, probability of failure...

SwissModel

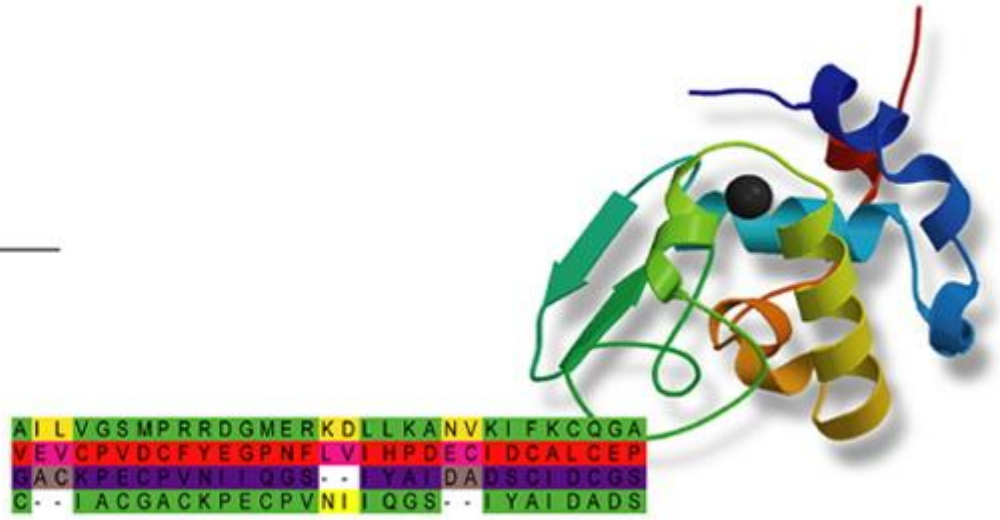
- Model representation
  - visual
  - InterproScan to detect domains
  - Sequence-based searches of the template library
  - Secondary structure and disorder prediction of the target protein.
  - Anolea mean force potential plot allows for quality assessment



Arnold et al. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics (2016) 22, 195–201. doi:10.1093/bioinformatics/bti770

Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints



- homology modelling with constraints (NMR, EM, apod)
  - i-Sites (short pieces with known structure)

http://salilab.org/modeller/

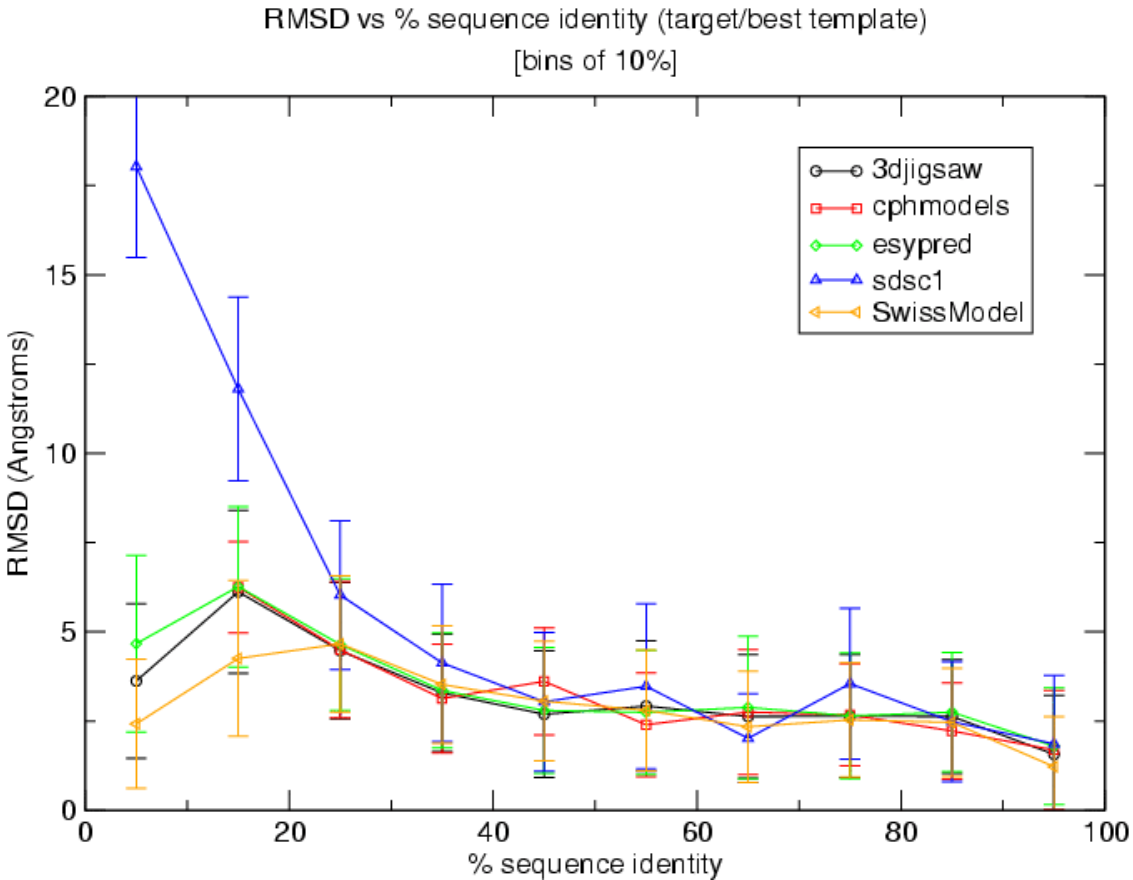
Comparative protein modelling by satisfaction of spatial restraints. Šali A, Blundell TL. *J Mol Biol.* 1993 Dec 5;234(3):779-815.

When to use Homology modelling?

- if there is enough similarity between target and template sequences

- CASP
  - at least

  30% IDENTITY



Threading

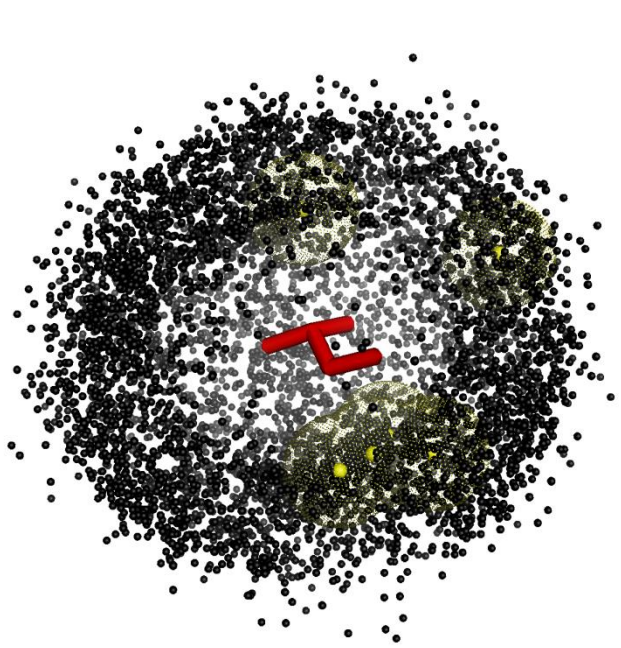
- or fold recognition
- tries to model onto common folds (not just one target) and tries to find out which one is the best

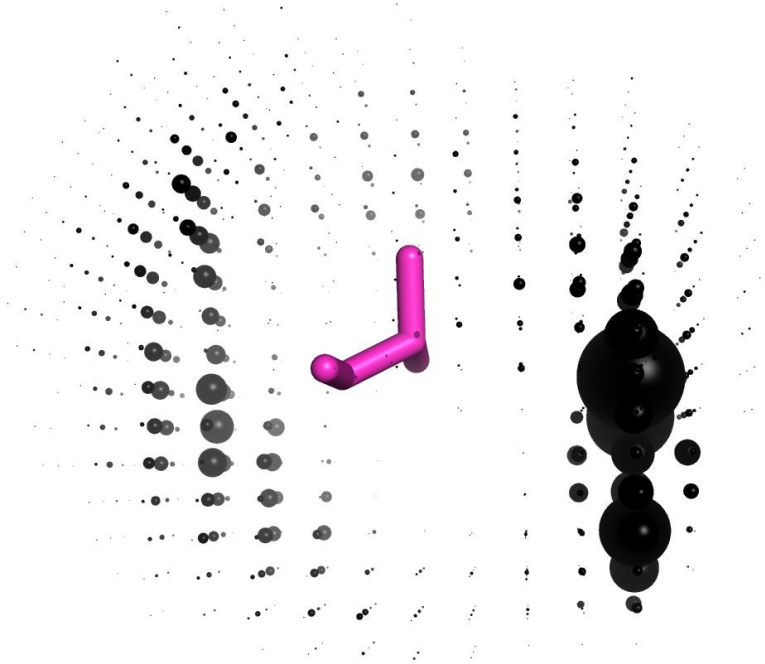
- I-Tasser

http://zhanglab.ccmb.med.umich.edu/I-TASSER

threading function

- energy-like function to find out amino acid preference for specific positions

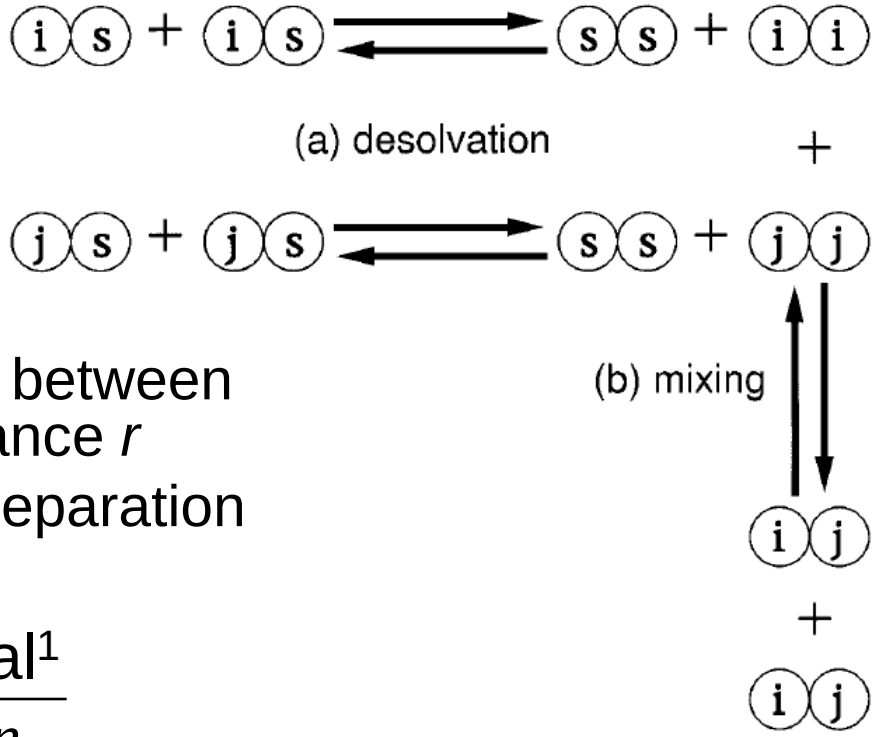




threading function

- Boltzmann formula

$$w_{ij}(r) = -kT \ln\left(\frac{\rho_{ij}(r)}{\rho^*}\right)$$



  - w_{ij} – free energy
  - $\rho_{ij}(r)$ – density of contacts between aminoacids *i* and *j* at distance *r*
  - ρ^* - reference density at separation

- Phenomenological potential[1]

$$e_{ij} = A \cdot \ln \frac{\overline{n_{ij}} \cdot \overline{n_{oo}}}{\overline{n_{io}} \cdot \overline{n_{jo}}}$$

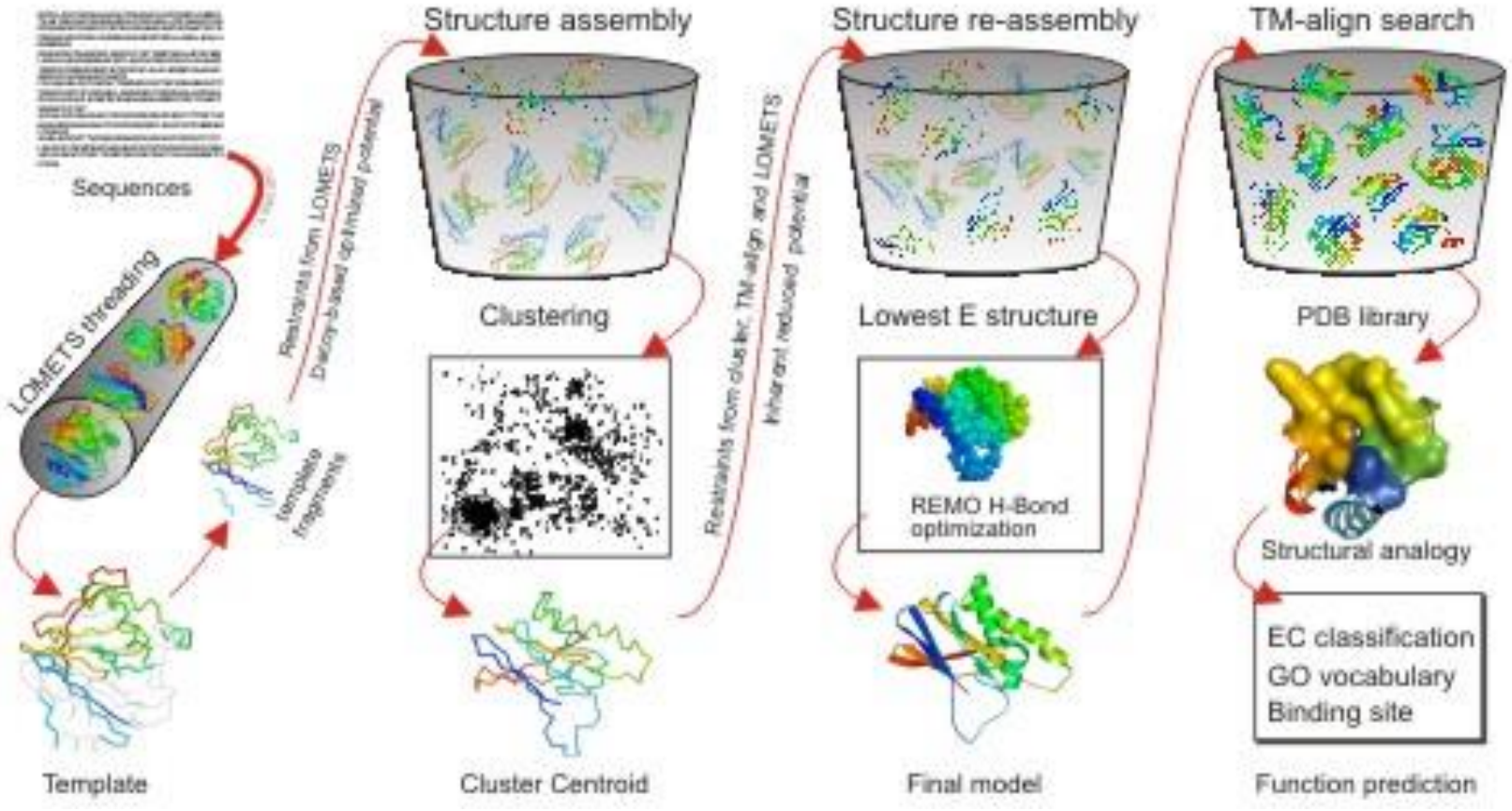
[1] Miyazawa, S. & Jernigan, R.L. *PROTEINS,***1999**, (36)357–369

Protein threading

- **DB of structural templates**
  - from PDB, FSSP, SCOP, or CATH, after **removing** protein structures with high sequence similarities.
- **Scoring function preparation**
  - measure the fitness between target sequences and templates based on the knowledge of the known relationships between the structures and the sequences.
  - A good scoring function should contain **mutation potential**, **environment fitness potential**, **pairwise potential**, **secondary structure compatibilities**, and **gap penalties**.
  - The quality of the energy function is closely related to the prediction accuracy, especially the alignment accuracy.
- **Threading alignment**
  - Align the target sequence with each of the structure templates by optimizing the designed scoring function.
  - solving the optimal alignment problem derived from a scoring function considering pairwise contacts.
- **Threading prediction**
  - statistically most probable alignment => threading prediction
  - construct a structure model for the target by placing the **backbone atoms** of the target sequence at their aligned backbone positions of the selected structural template.

I-TASSER

- Best automated server for prediction of 3D structure



- http://zhanglab.ccmb.med.umich.edu/I-TASSER/

I-TASSER ONLINE
Protein Structure & Function Predictions

I-TASSER

1. LOMETS
   – **metaserver** pro 8 methods of prediction for tertiary structure of fragments
2. use of fragments from identified templates
   - replica-exchange Monte Carlo simulations
   - threading of not homologous regions (loops)
     with ab initio modeling
3. SPICKER – clustering of best results
4. again LOMETS on individual clusters
5. TM-align - sequence-order independent protein structure alignment

I-TASSER ONLINE
Protein Structure & Function Predictions

energy terms

contact_cut.comm: Residue contact cutoff parameters
contact_profile.comm: Side-chain contacts environment profile
contact3.comm: Orientation-dependent side-chain contact potential
CA13.comm: Short-range C-alpha correlation of (i,i+2)
CA14.comm: Short-range C-alpha correlation of (i,i+3)
CA15.comm: Short-range C-alpha correlation of (i,i+4)
CA14s.comm: Short-range C-alpha correlation of (i,i+3) for strands
CA14h.comm: Short-range C-alpha correlation of (i,i+3) for helices
CA15s.comm: Short-range C-alpha correlation of (i,i+4) for strands
CA15h.comm: Short-range C-alpha correlation of (i,i+4) for helices
CB.comm: C-beta positions
sidechain.comm: Sidechain center positions

When to use threading?

- If there is not enough sequence identity to one template
- by individual domains
- fold recognition by consensus from several programs – meta methods
- use as much as experimental evidence as possible – to discern which fold is true

Ab initio modeling

- ab initio = without template
- masive search for right conformation
- (pseudo-)physical energy function for free energy

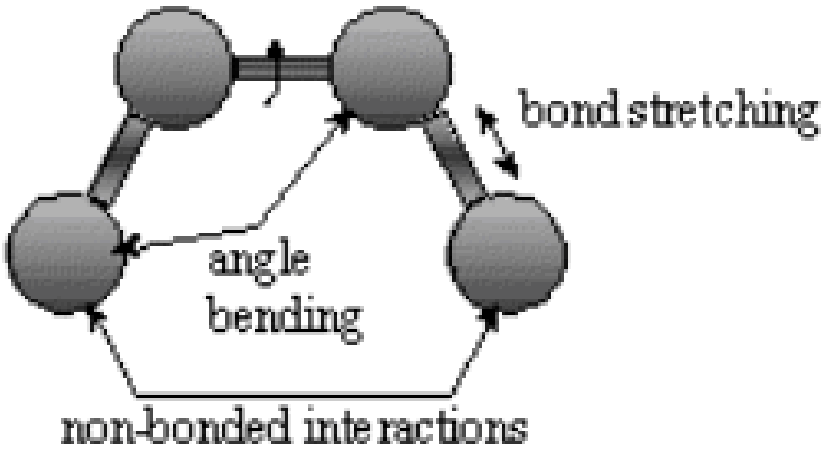
- http://robetta.bakerlab.org/
- *ab initio* and comparative models of protein domains

- The least precise, but the only one which can be use when no template is known

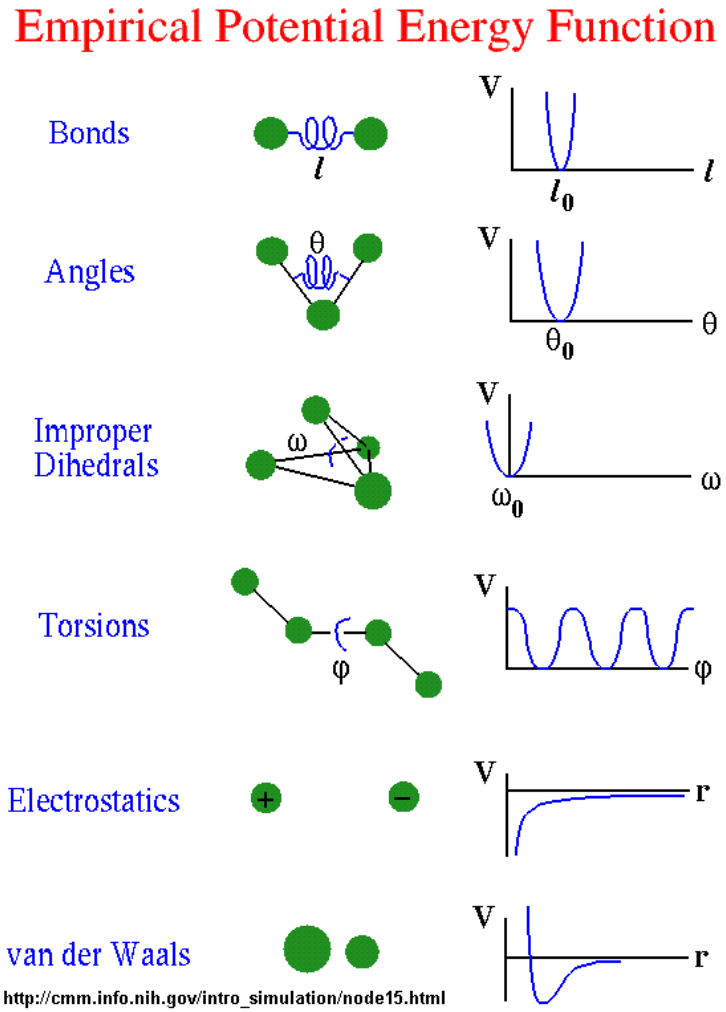
Molecular mechanics

total energy is function of atom positions

$$E = f(\mathbf{R}) = E_b + E_a + E_t + E_c + E_{vdw}$$



Force-field



$$E_b = \frac{k_r}{2}\left(r - r_0\right)^2$$

$$E_a = \frac{k_\theta}{2}\left(\theta - \theta_0\right)^2$$

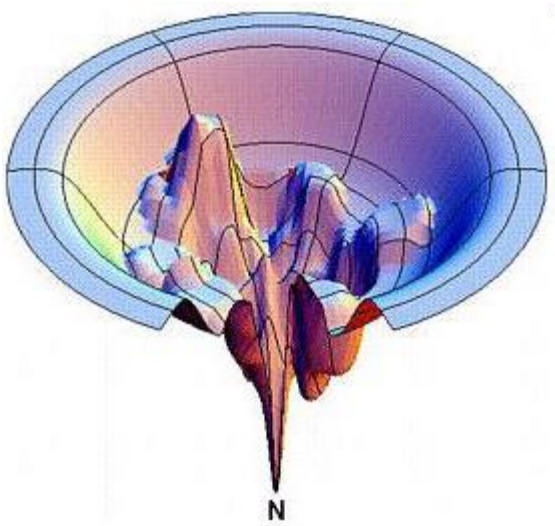
$$E_t = \frac{k_\theta}{2}\left(1 + \cos\left(n\phi - \phi_0\right)\right)$$

$$E_c = \frac{1}{4\pi\varepsilon_0\varepsilon_r}\frac{q_i q_j}{r_{ij}}$$

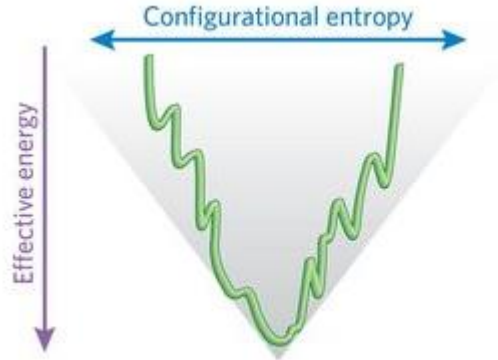
$$E_{vdw} = -2\varepsilon_{ij}\left(\frac{r_{ij}^*}{r_{ij}}\right)^6 + \varepsilon_{ij}\left(\frac{r_{ij}^*}{r_{ij}}\right)^{12}$$

Conformations

- Potential energy surface (PES)
  - barriers, minima
  - global vz. local



- folding -> movement over PES
  - folding funnel
  - methods:
    - energy optimization
    - molekular dynamics
    - simulated heating
    - metadynamics
    - Monte Carlo



Use of MM/MD

- **protein folding**
  - search for global minima
  - too resource intensive
  - usable only for small proteins

  \+ can be used to study process of folding and its kinetics

  \+ model raffination

Distributed computing projects

Folding@Home

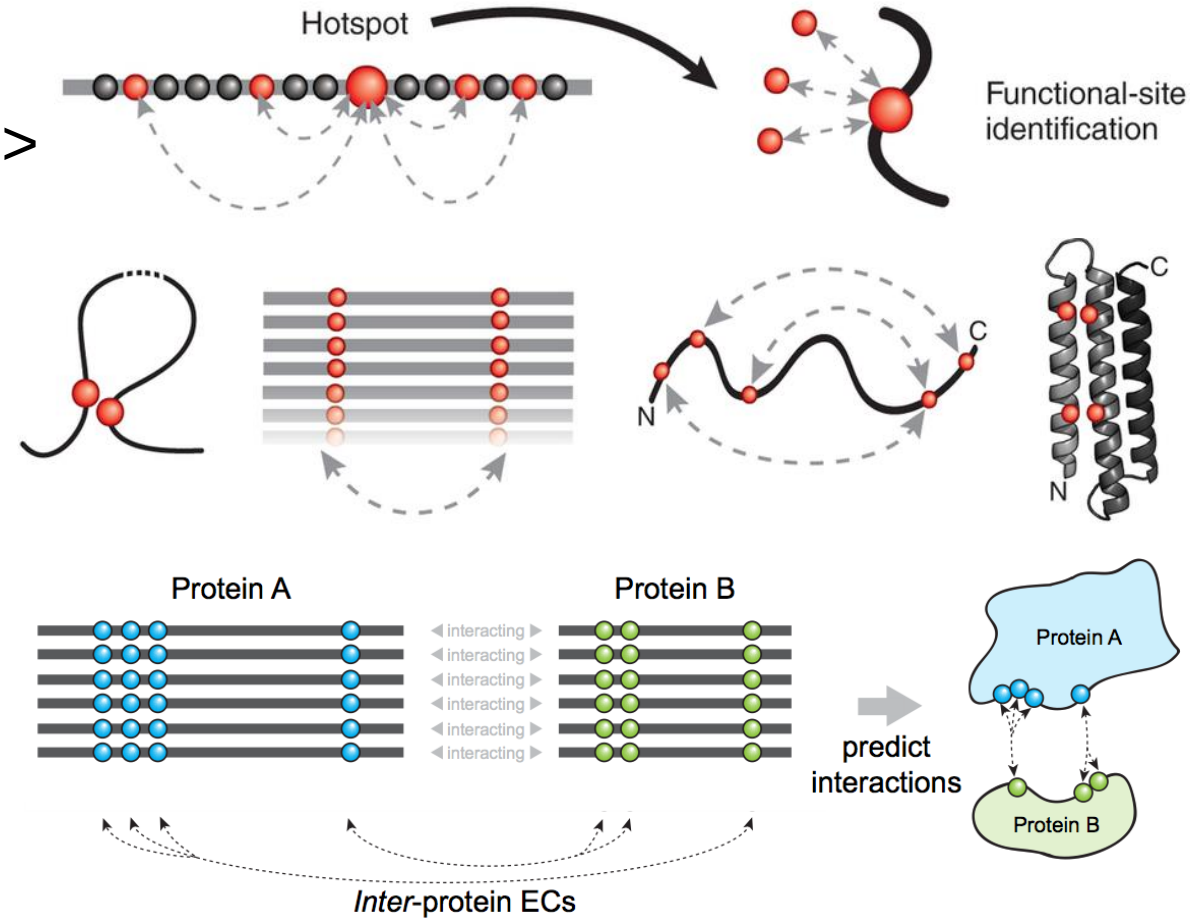
- http://foldingathome.stanford.edu/
- simulates protein folding, computational drug design, and other types of molecular dynamics
- determine the mechanisms of protein folding
- Pande Lab at Stanford U
- 100 petaFLOPS on May 11, 2016

Rossetta@Home

- http://boinc.bakerlab.org/
- protein structure prediction
- on the Berkeley Open Infrastructure for Network Computing (BOINC)
- Baker lab at UWashington
- predict protein–protein docking and design new proteins with the help of
- ~60 000 active computers over 210 teraFLOPS on average as of July 29, 2016

Evolutionary Couplings

- EVcouplings
  - multiple alignment -> functional site
- EVfold
  - Folds the protein when unknown structure
- EVcomplex
  - Finds and joins partner sequences from the two MSAs



http://evfold.org/

ALPHAFOLD2

under the hood

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

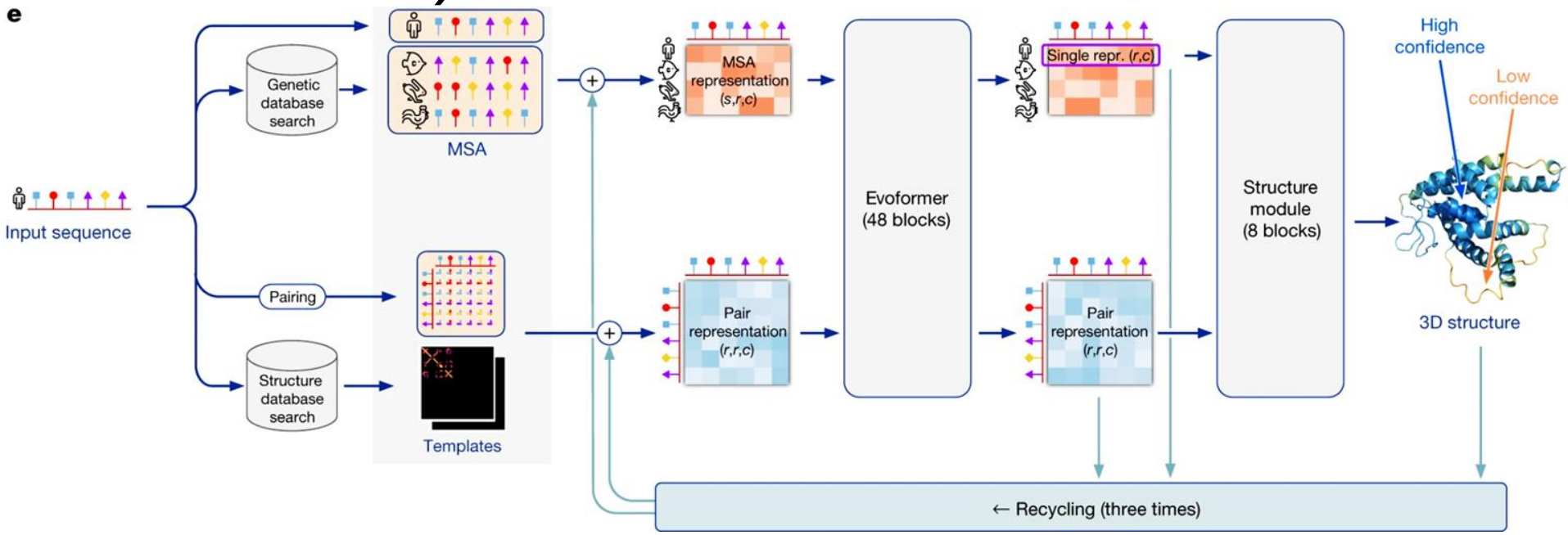
elixir CZECH REPUBLIC

AlphaFold2

Input: sequence

extended by MSA + structural templates

Evoformer and Structure model (w MD simulation)

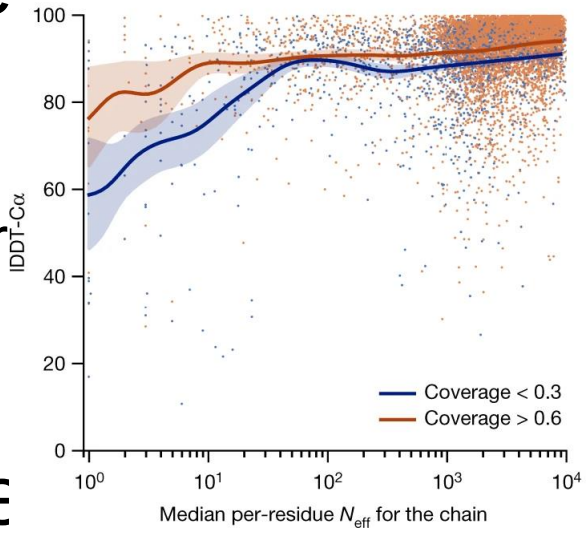


https://www.nature.com/articles/s41586-021-03819-2

MSA - multiple sequence alignment

using standard tools - jackhammer, HHBlits

- sequence DBs:
  - *UniRef90*
    - reference sequences from UniProt
  - *UniClust30*
    - for sequence self-distillation
- metagenomicsDBs - fully cove
  UniRef90
  - *Big Fantastic database (BFD)*
    - 66M protein families from 2.2G pr
  - clustered *MGnify*
    - more than 260k genomes



needed at least 30 sequences pe

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

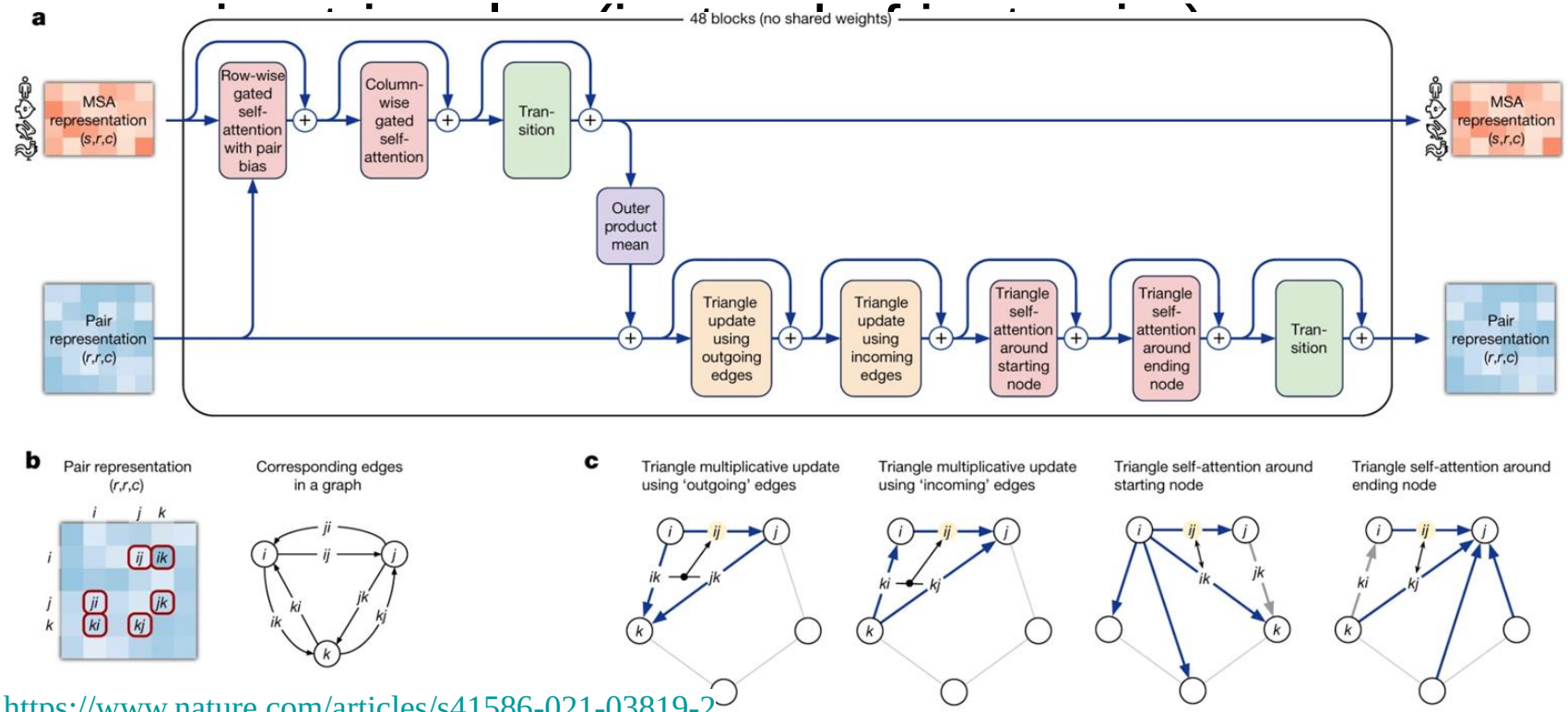
then train again from scratch with mixture of PDB and UniRef30

=> effective use of unlabelled sequence data

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

EvoFormer

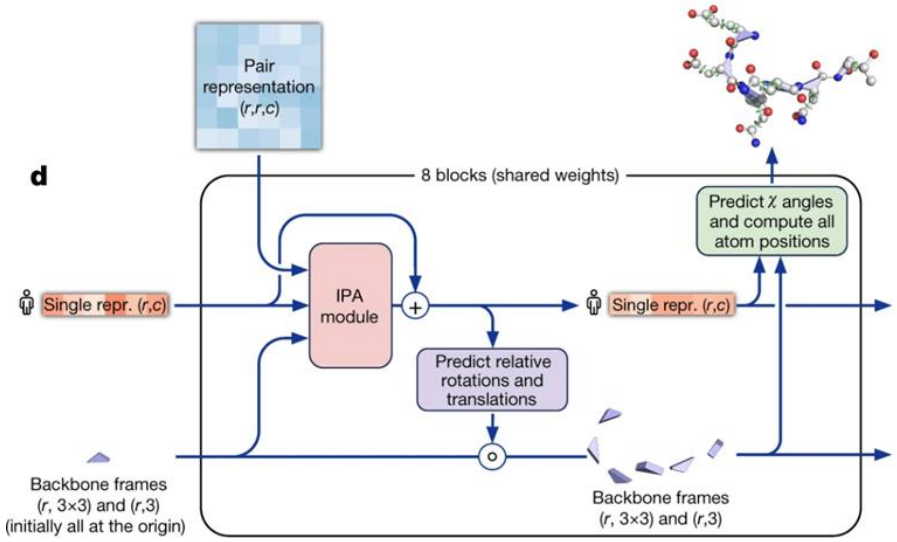
- mixing MSA and pairs via updates
- graph inference problem in 3D space
  - edges = residues in proximity
  - updates per each block (48 blocks) separately (AF1 updated all network at once)

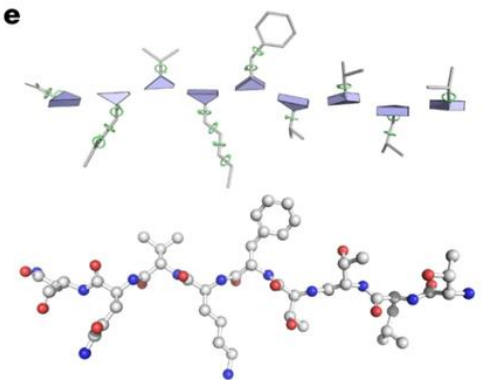


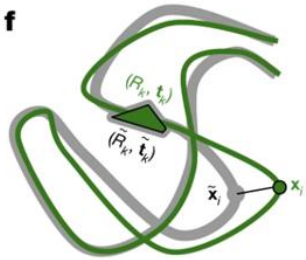
https://www.nature.com/articles/s41586-021-03819-2

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





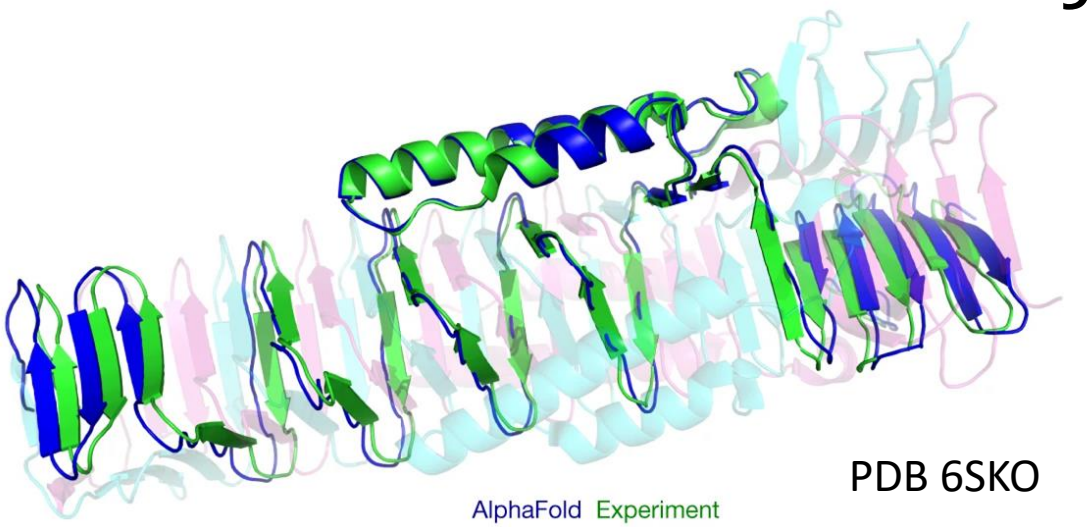


https://www.nature.com/articles/s41586-021-03819-2

Effect of cross-chain contacts

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

homotro[illegible]en when ch[illegible]



PDB 6SKO

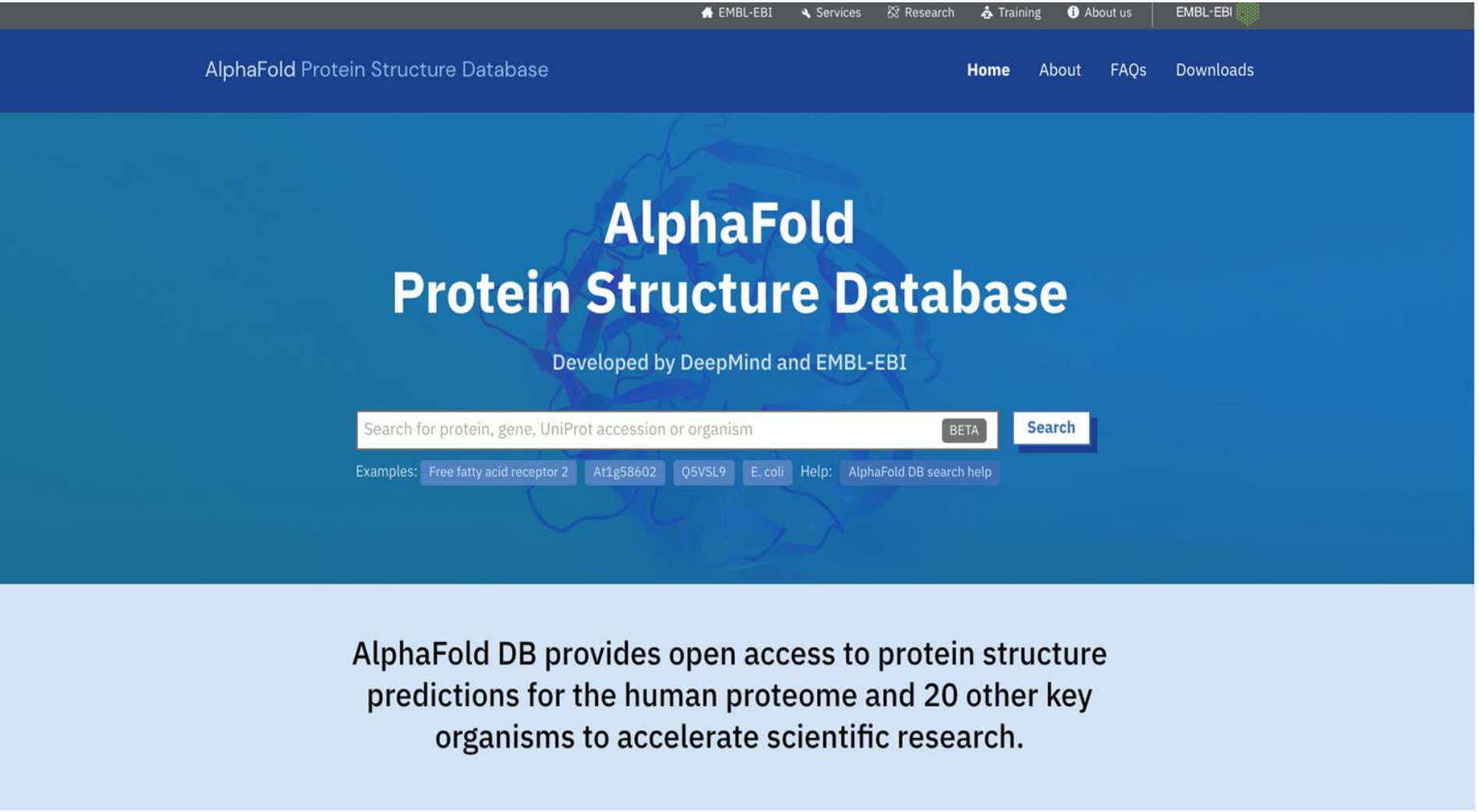
Timings

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



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

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

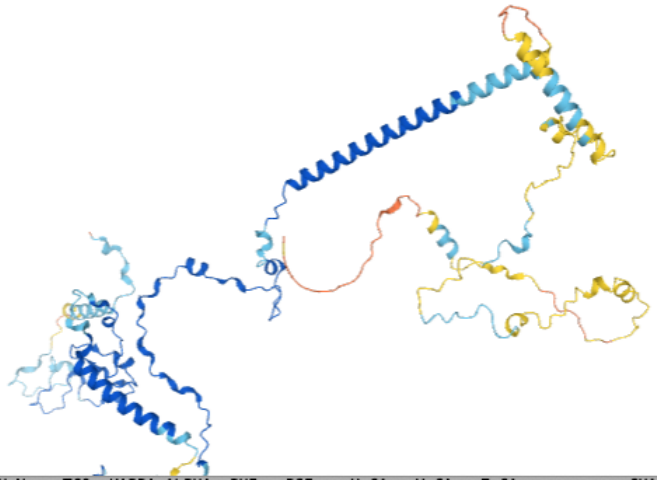
Model Confidence:

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

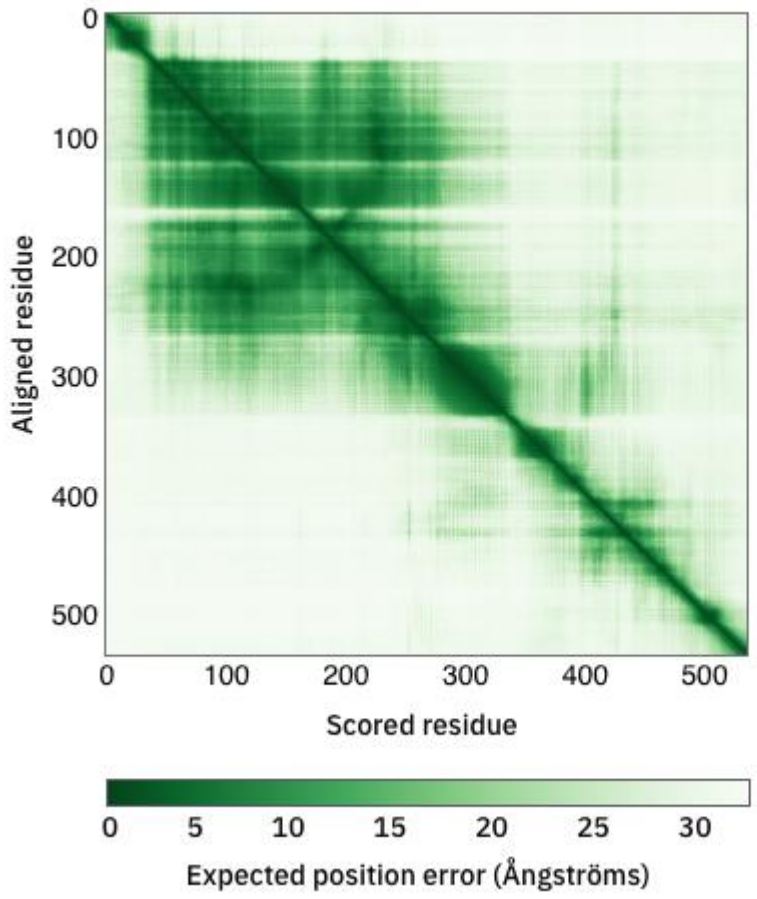
AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

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

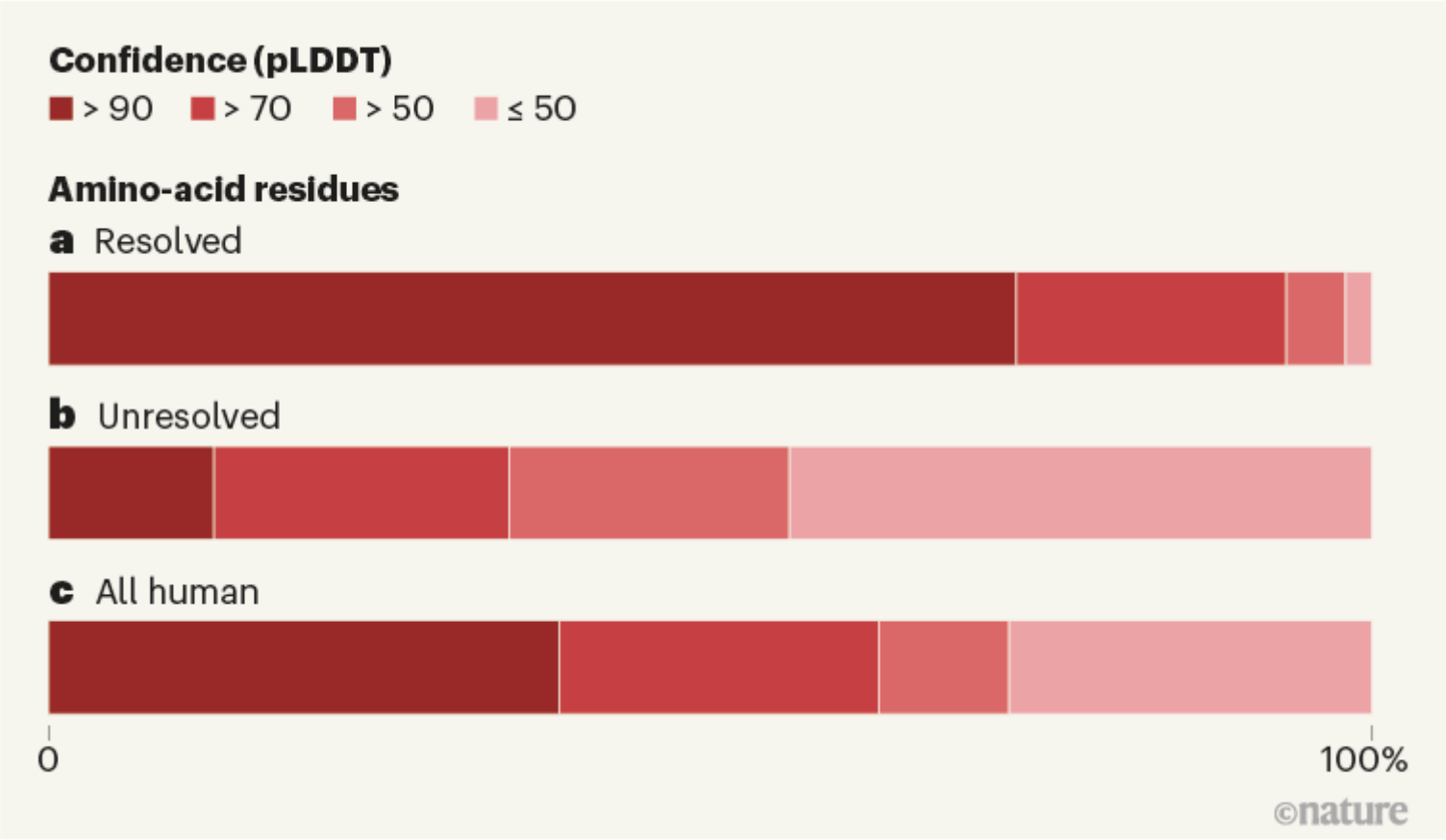
```
MALTSFLPAPTQLSQDQLEAEEKARSQRSRQTSLVSSRREPPPYGYRKGWIPRLLEDFGDGGAFPEIHVAQYPLDMGRKKKMSNALAIQVDSEGKIKYDAIARQGQSKDKVIYSKYTDLVPKEV
MNADDPDLQRPDEEAIKEITEKTRVALEKSVSQKVAAAMPVRAADKLAPAQYIRYTPSQQGVAFNSGAKQRVIRMVEMQKDPMEPPRFKINKKIPRGPPSPPAPVMHSPSRKMTVKEQQEWKIP
PCISNWKNAKGYTIPLDKRLAADGRGLQTVHINENFAKLAEALYIADRKAREAVEMRAQVERKMAQKEKEKHEEKLREMAQKARERRAGIKTHVEKEDGEARERDEIRHDRRKERQHDRNLSRA
```



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

Repozitář: sokrypton/ColabFold

Větev: main

Cesta

AlphaFold2.ipynb

AlphaFold2_complexes.ipynb

RoseTTAFold.ipynb

batch/AlphaFold2_batch.ipynb

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 good GPU/TPU to run
  -> not many people have it on their PC

- Alphafold has been installed on ELIXIR CZ hardware
  - `/storage/brno11-elixir/projects/alphafold`

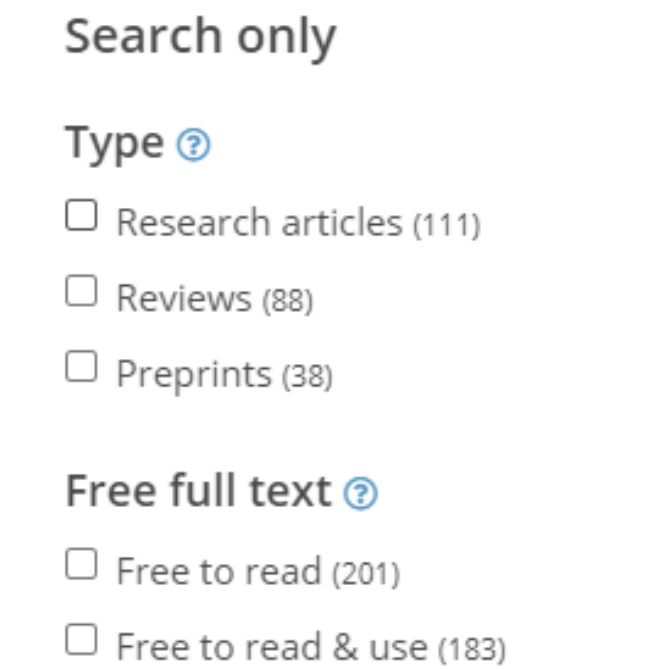
- Elixir is accessible through Metacentrum
  - https://wiki.metacentrum.cz/wiki/AlphaFold

- speed is dependent on size of predicted protein but can be in order of tens of minutes

Alphafold is just a start...

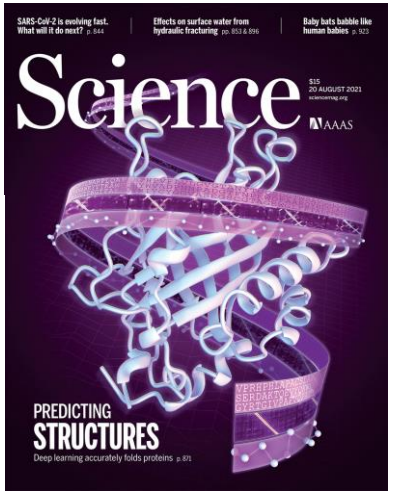
- use Alphafold ideas for development of their own 3D structure predictions
  - RoseTTAfold
  - prediction of designed proteins
- prediction of RNA structures
- prediction of orphan proteins
  - molecular replacement
  - interpretation of cryoEM
  - pLDDT can act as IDP predictor

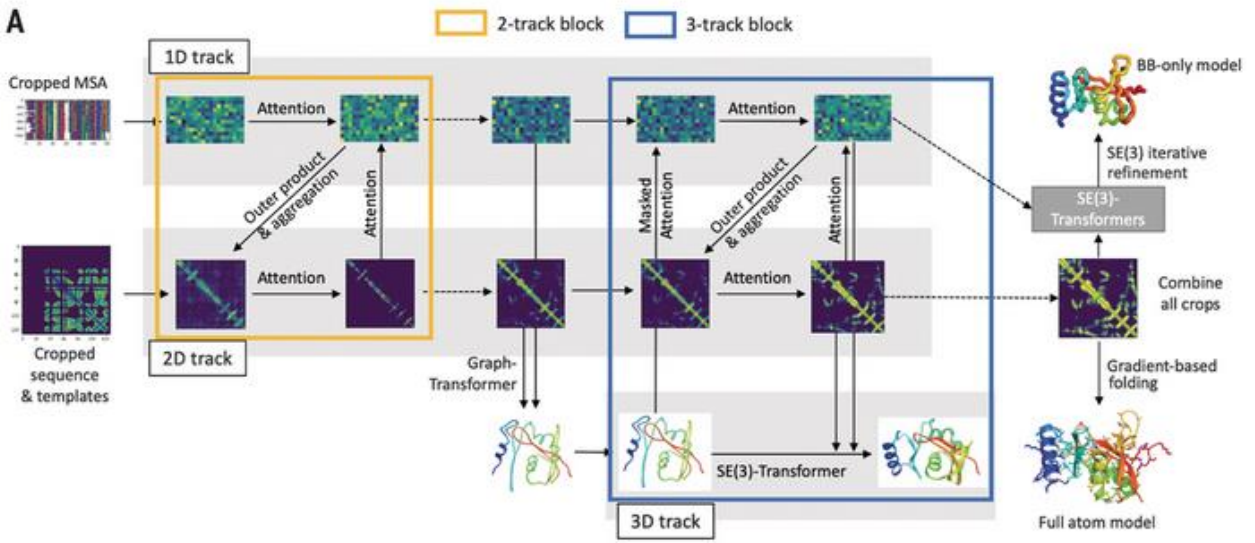
...

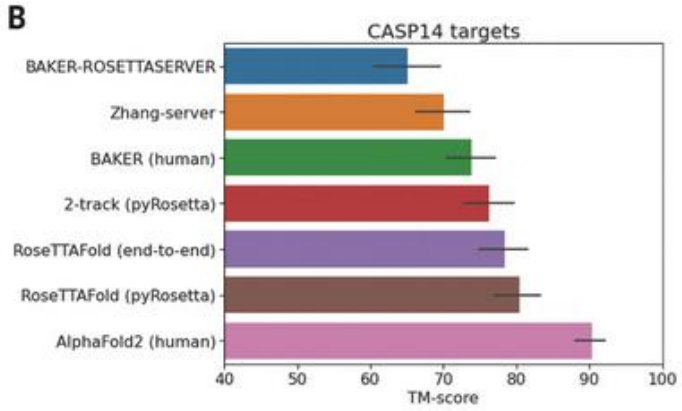


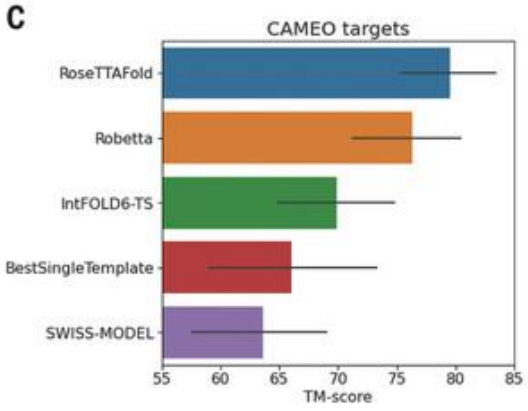
as of

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

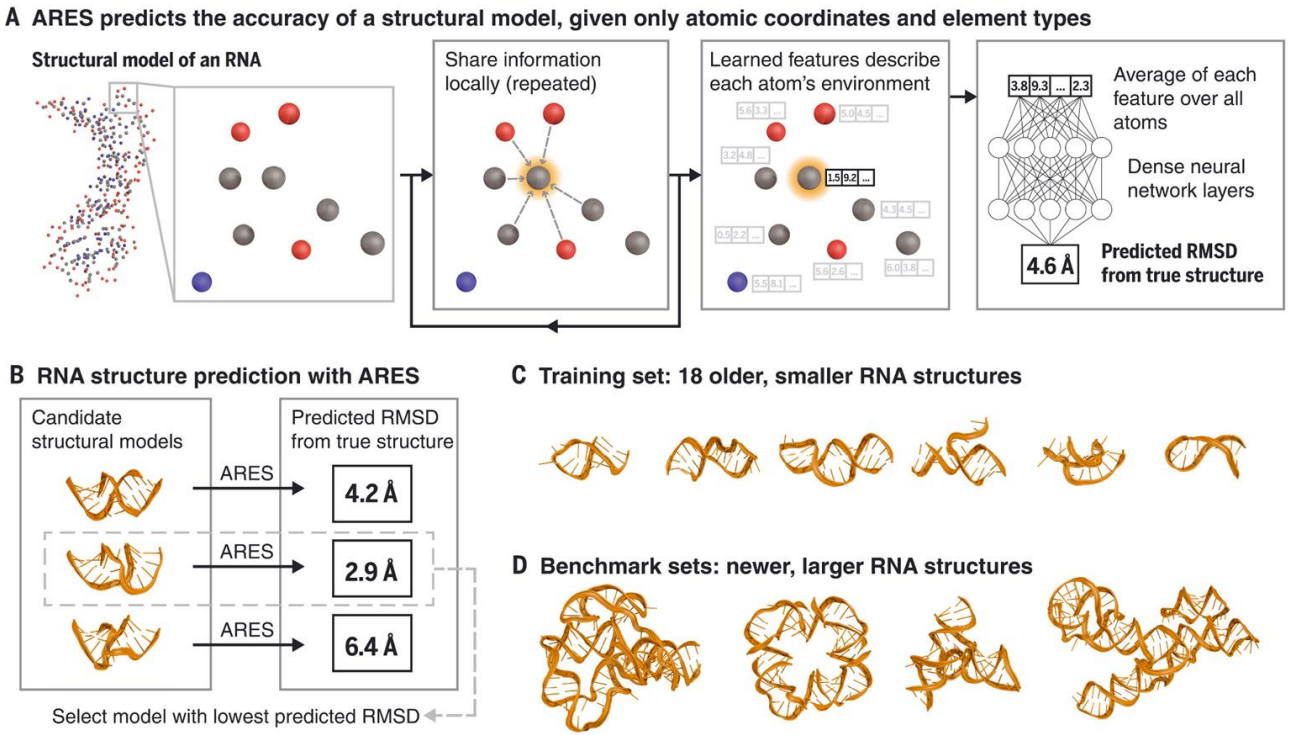
David T. Jones*
Department of Computer Science
University College London
Gower St, London WC1E 6BT
d.t.jones@ucl.ac.uk

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

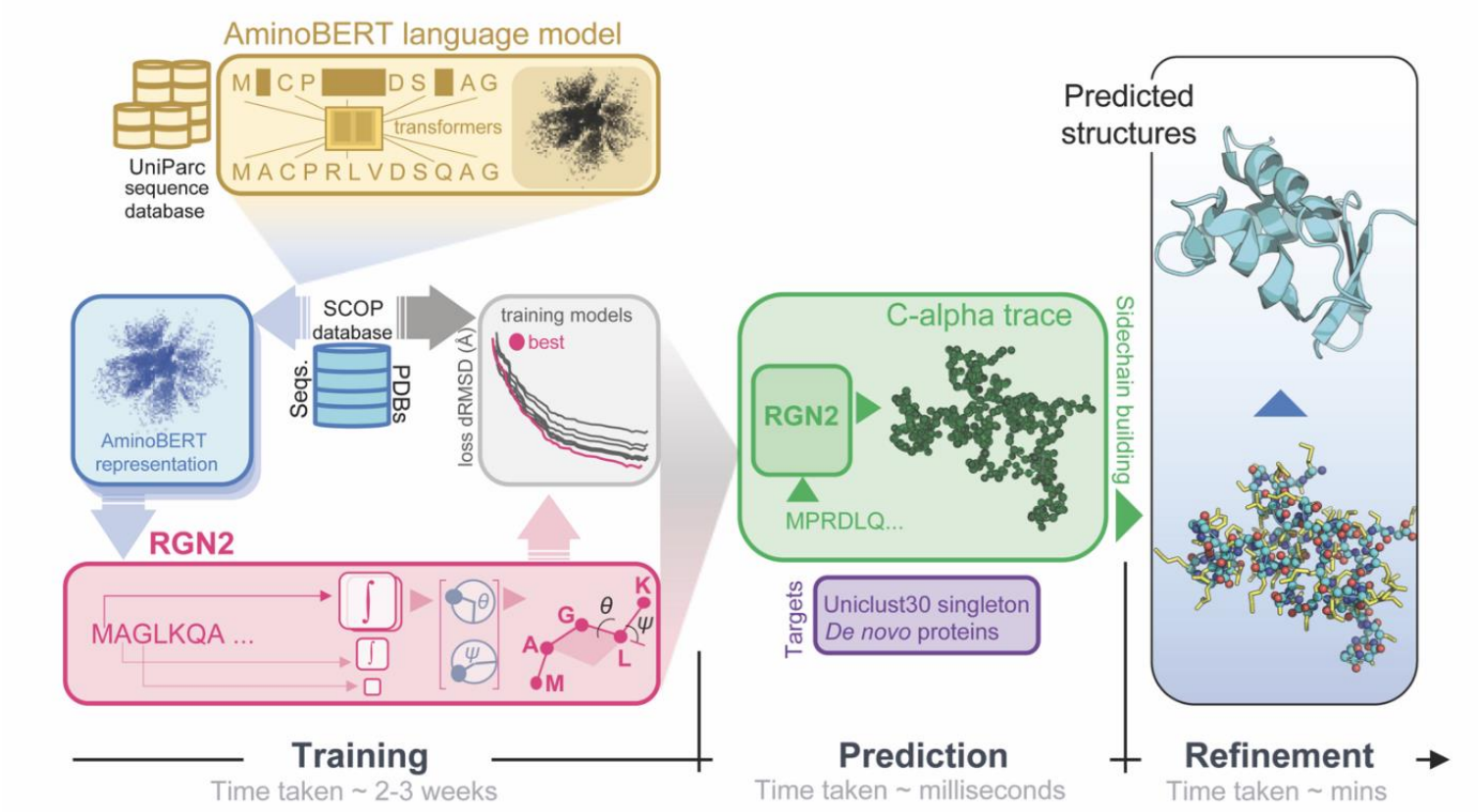
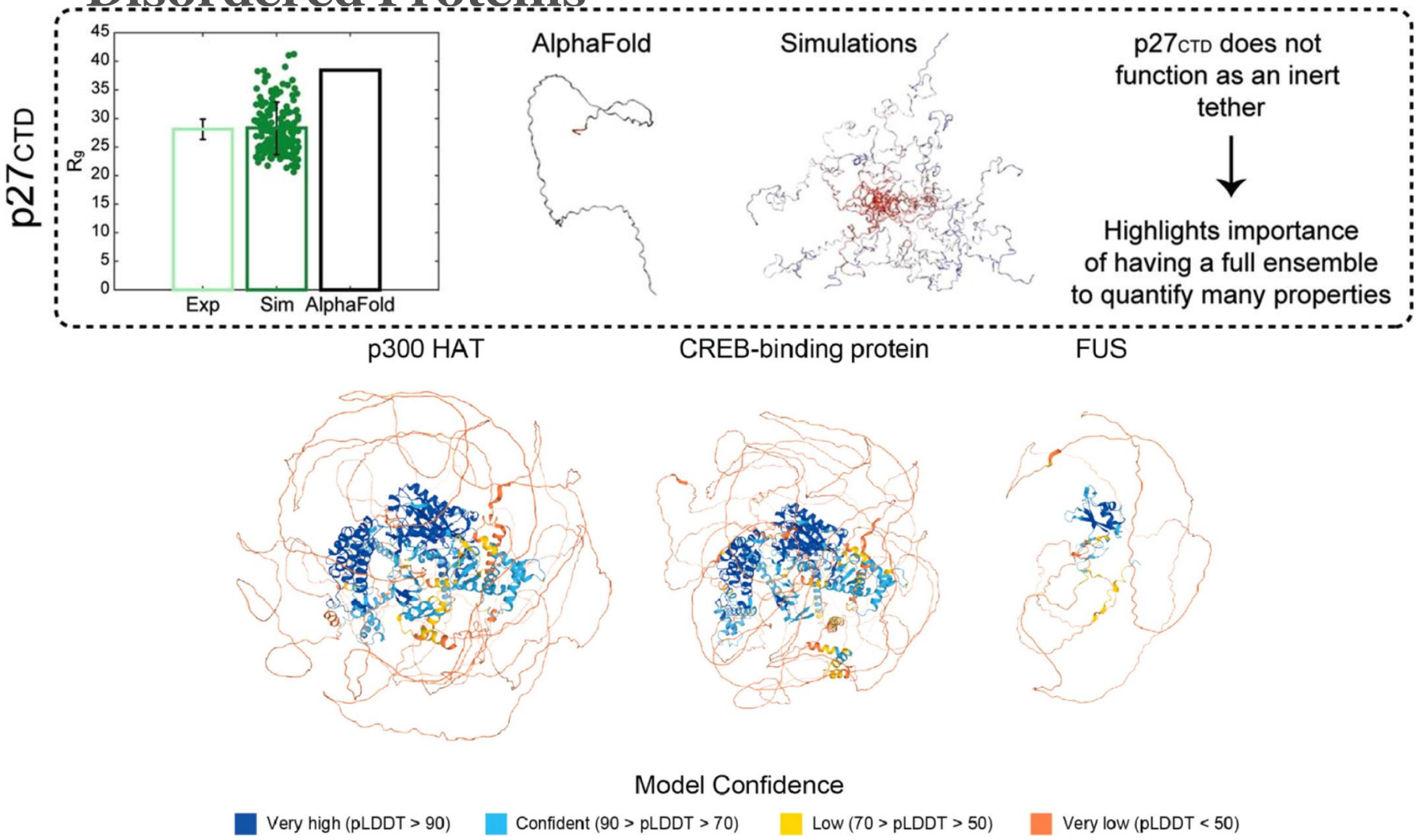


Figure 1. Organization and application of RGN2. RGN2 combines a Transformer-based protein language model (AminoBERT) with a recurrent geometric network that utilizes Frenet-Serret frames to generate the backbone structure of a protein. Placement of side chain atoms and refinement of hydrogen-bonded networks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

AlphaFold in MobiDB

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

ALPHAFOLD LIMITATIONS

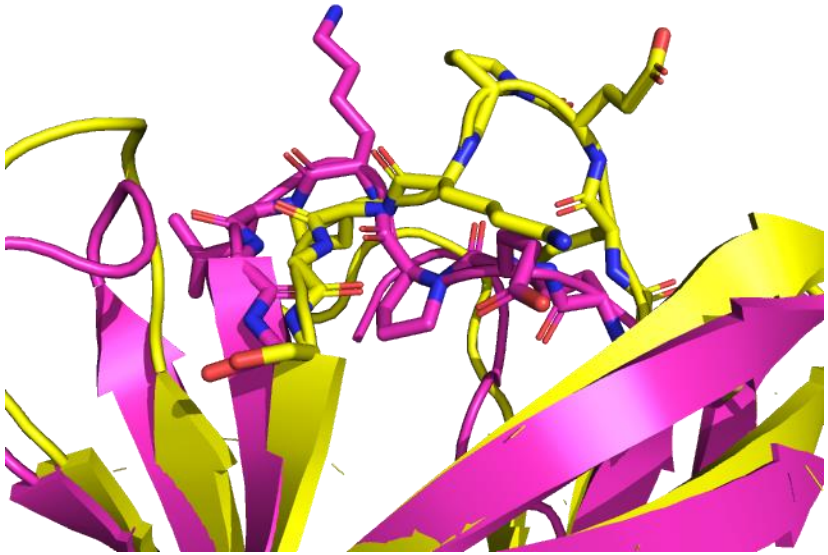
Are structural biologists and bioinformaticians on the job market?

- 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
- Alphafold2 is **publicly available** and can be run from many places including ELIXIR CZ
- Alphafold has inspired many tools already – ESMfold, OpenFold, ColabFold, …
- Alphafold limits are yet to be fully described

Quality Control

CAMEO, CASP

CAPRI

CAFA

Quality control of protein models - CASP, CAMEO

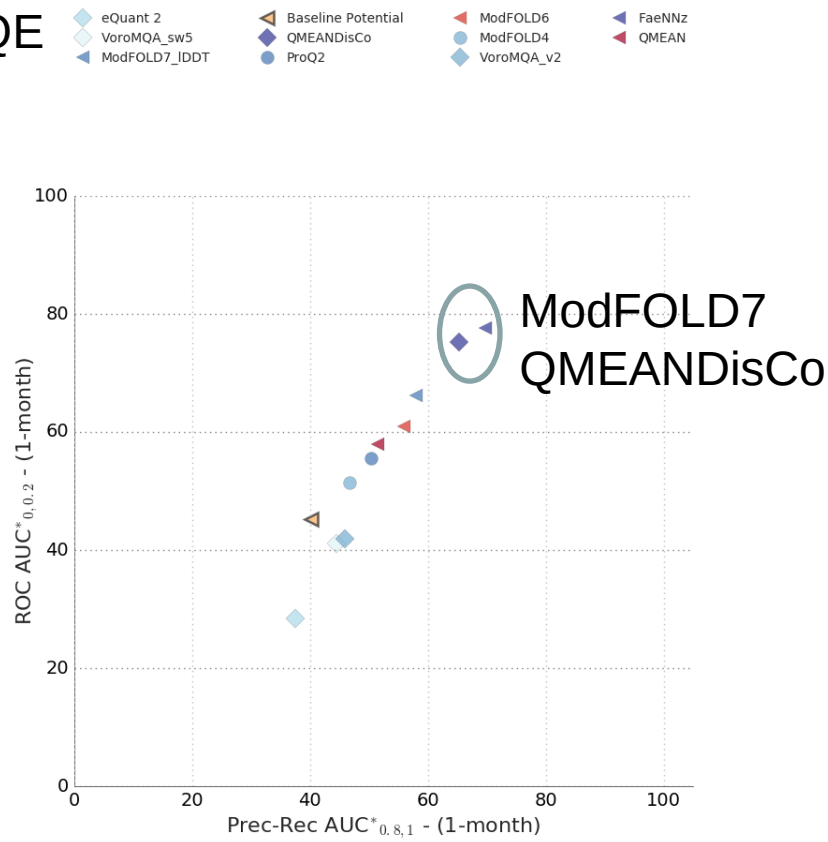
- **CASP** - Critical Assessment of techniques for protein Structure Prediction
  - runs every two years -
  http://www.predictioncenter.org/casp13
  - large QA for whole protein modelling field - establishes the criteria and ranks prediction teams, programs and servers
  - last round not yet published - special Proteins 2019 issue
- **CAMEO** - Continuous Automated Model EvaluatiOn
  - runs every week - https://www.cameo3d.org/
  - tests **3D structure** prediction, **Quality of Model** Estimation, **Contact Prediction**
  - 3D - based on IDDT (Local Distance Difference Test - model vz exp. 0-100(best)
  - QE - predicted IDDT>60

CAMEO - best 3D and QE

- 3D

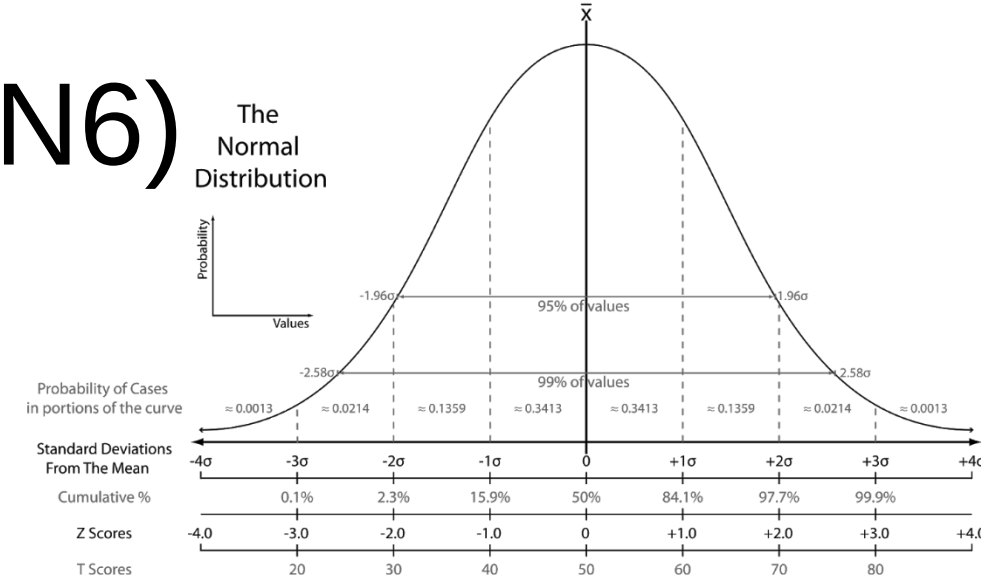
Server	time	IDDT	IDDT-BS
Robetta	23 h	**69.1 ± 13.6**	65.77
IntFOLD5-TS	35 h	67.7 ± 15.5	**71.61**
RaptorX	10 h	66.8 ± 15.3	67.87
HHpredB	38 min	63.9 ± 17.2	67.14
SwissModel	**7 min**	63.1 ± 21.1	69.24
...			
Phyre2	2 h	52.9 ± 21.6	65.12

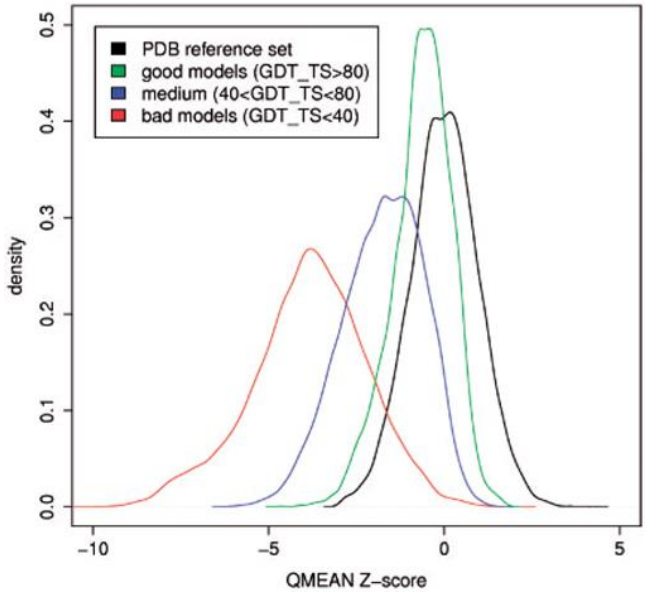
QE



QMEAN4 (QMEAN6)

- composite scoring of aspects of QA of model
  - all-atoms -
  - Cβ -
  - solvation -
  - torsions -
  - (ss_aggreement) -
  - (acc_agreement) -
- Benchmarked to PDB reference set (cross validated)
  z-score - good models ~ -1, bad models < -4

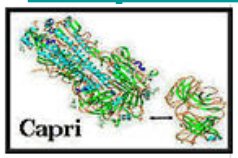




Bioinformatics, 27(3) 2011, 343–350, https://doi.org/10.1093/bioinformatics/btq662

CAPRI

- Critical Assessment of Prediction of Interactions
- http://www.ebi.ac.uk/msd-srv/capri/



Community wide experiment on the comparative evaluation of protein-protein docking for structure prediction

Hosted By EMBL/EBI-PDBe Group

Round 35

Round 35 ID mapping from group number to Accessor Number for Target 107
Round 35 UPLOADER ID mapping for Target 107
Round 35 SCORER ID mapping for Target 107

Full results:

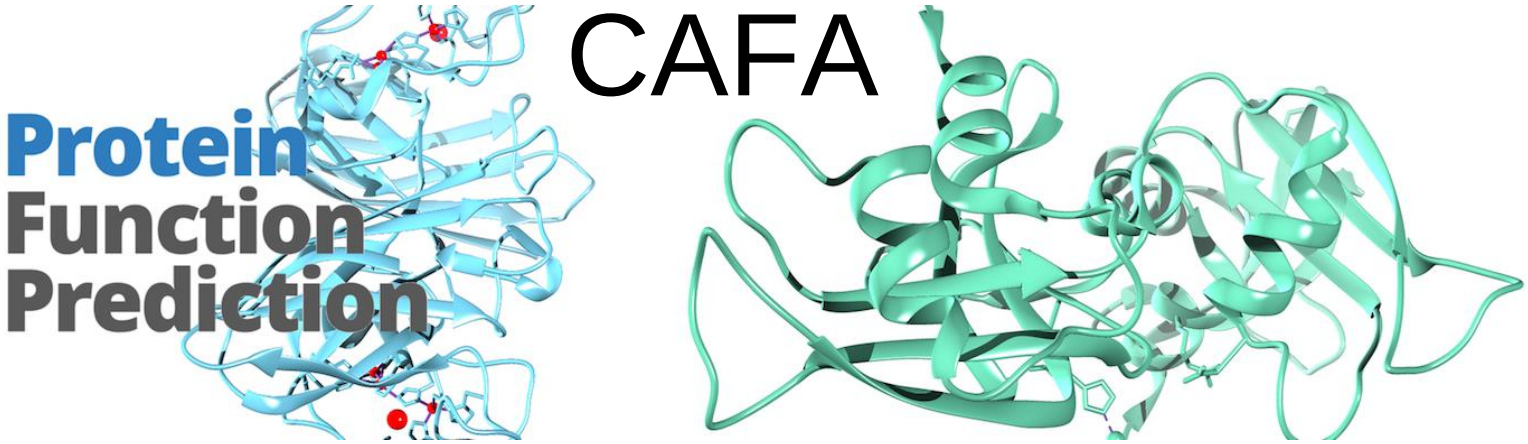
Target 107

Results T107 - click here to see the results

Clash threshold	53.61
average	16.81
std dev	18.40

	Predictor	Uploader	Scorer
Nr groups	25	14	14
High Accuracy (***)	0 (0)	0 (0)	0 (0 <- 0)
Medium (**)	0 (0)	0 (0)	0 (0 <- 0)
Acceptable (*)	0 (0)	0 (0)	0 (0 <- 0)
Incorrect	239 (25)	1075 (13)	137 (14 <- 13)
Clashes	11 (4)	135 (5)	2 (2 <- 2)
Low ID	0 (0)	0 (0)	0 (0 <- 0)
Total	250 (25)	1210 (13)	139 (14 <- 13)

Numbers in parentheses refer to the number of different source groups producing these models (ie z scorer groups recognize y out of x uploader groups)



CAFA

- Critical Assessment of Function Annotation
  - large-scale assessment of computational methods dedicated to predicting protein function.
  - http://biofunctionprediction.org/

P. Radivojac et al A large-scale evaluation of computational protein function prediction. Nat Methods, 10(3):221–227, 2013.

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
- Alphafold2 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

Acknowledgement

Marian Novotný

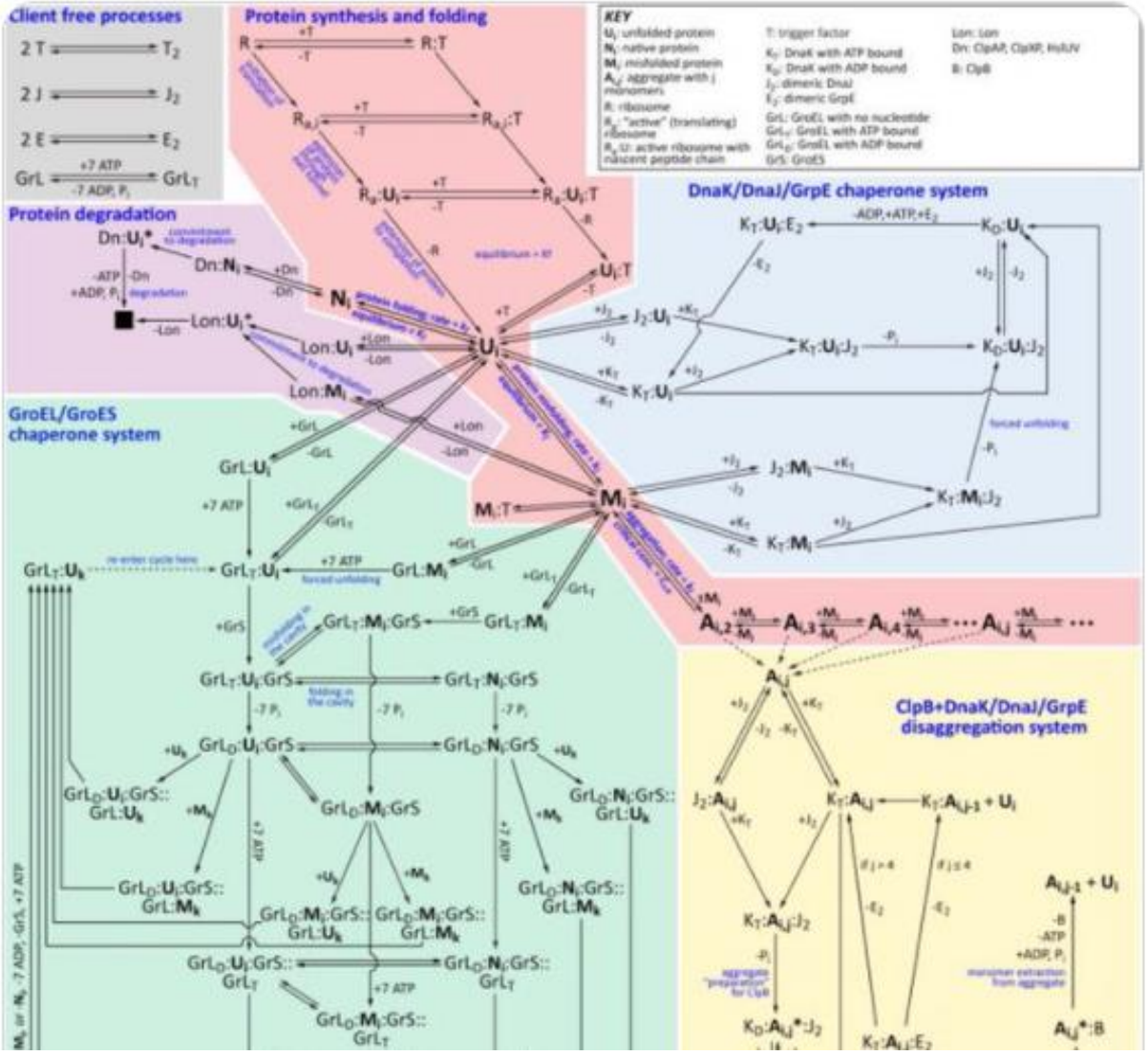
PŘÍRODOVĚDECKÁ
FAKULTA
Univerzita Karlova

And now something completely different...

Kresten Lindorff-Larsen
@LindorffLarsen

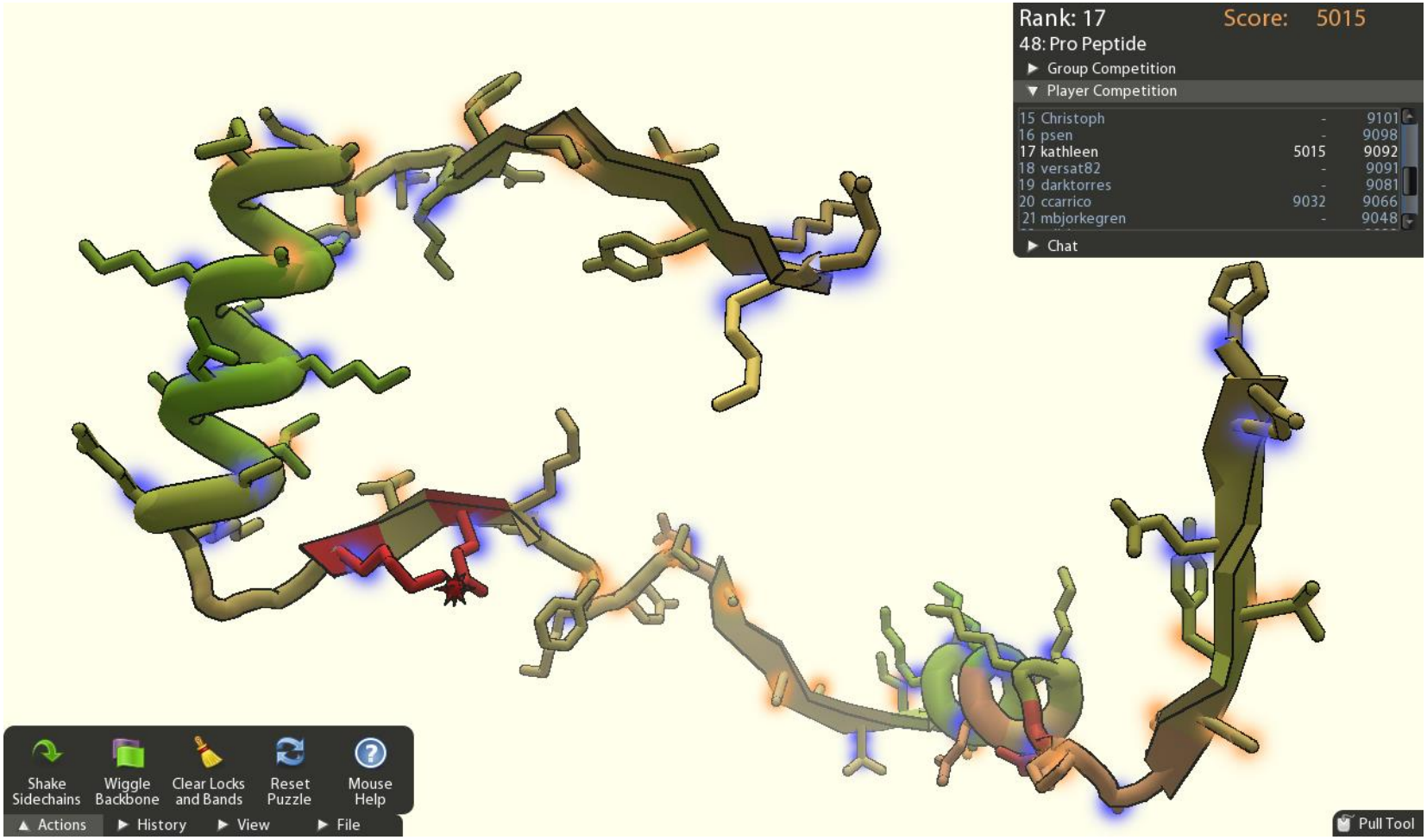
Tell me again how the folding problem has been solved doi.org/10.1016/j.jmb.... doi.org/10.1016/j.celr...

Přeložit Tweet



FoldIt

- protein folding as a game

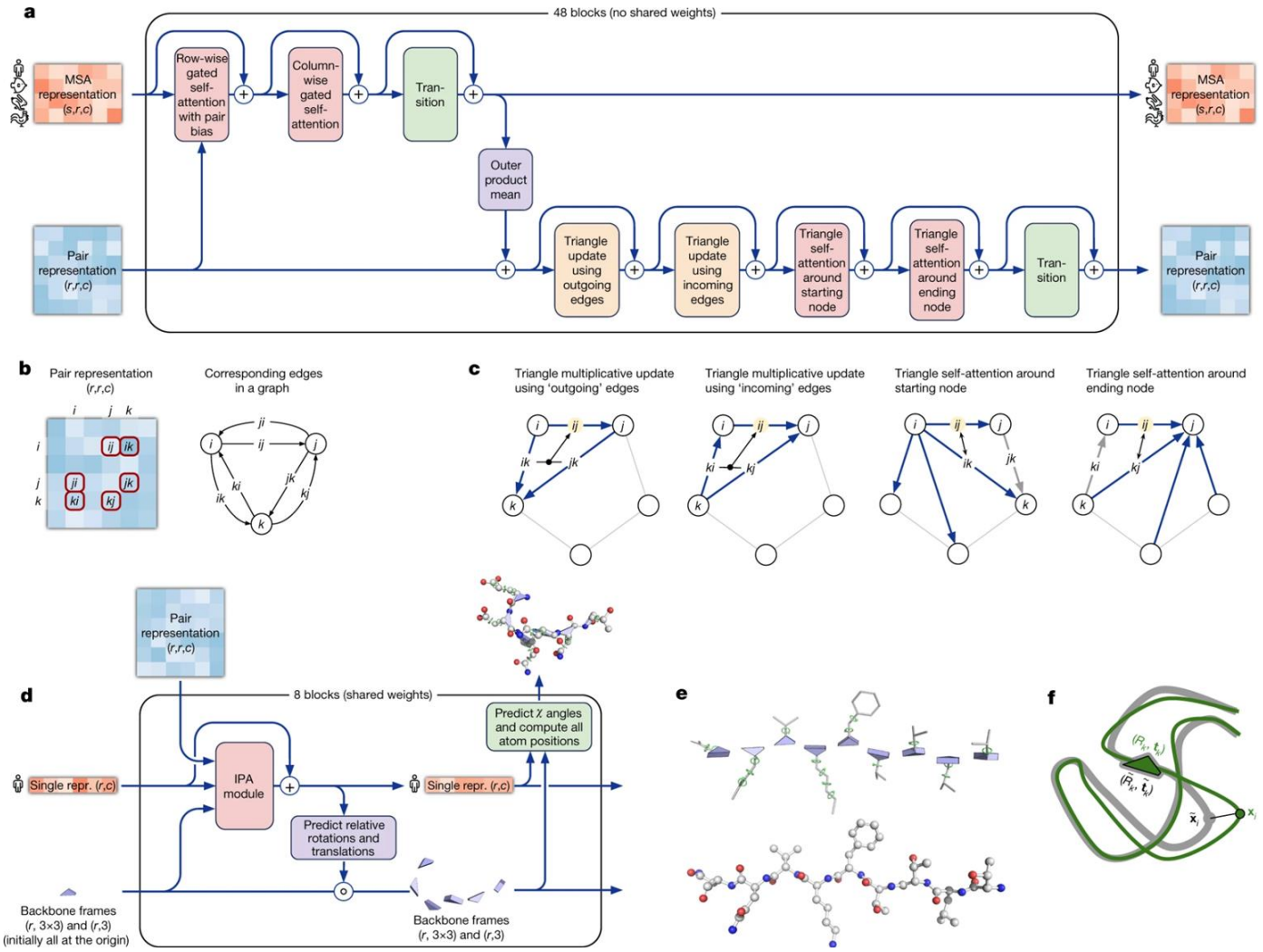


- http://fold.it/portal/

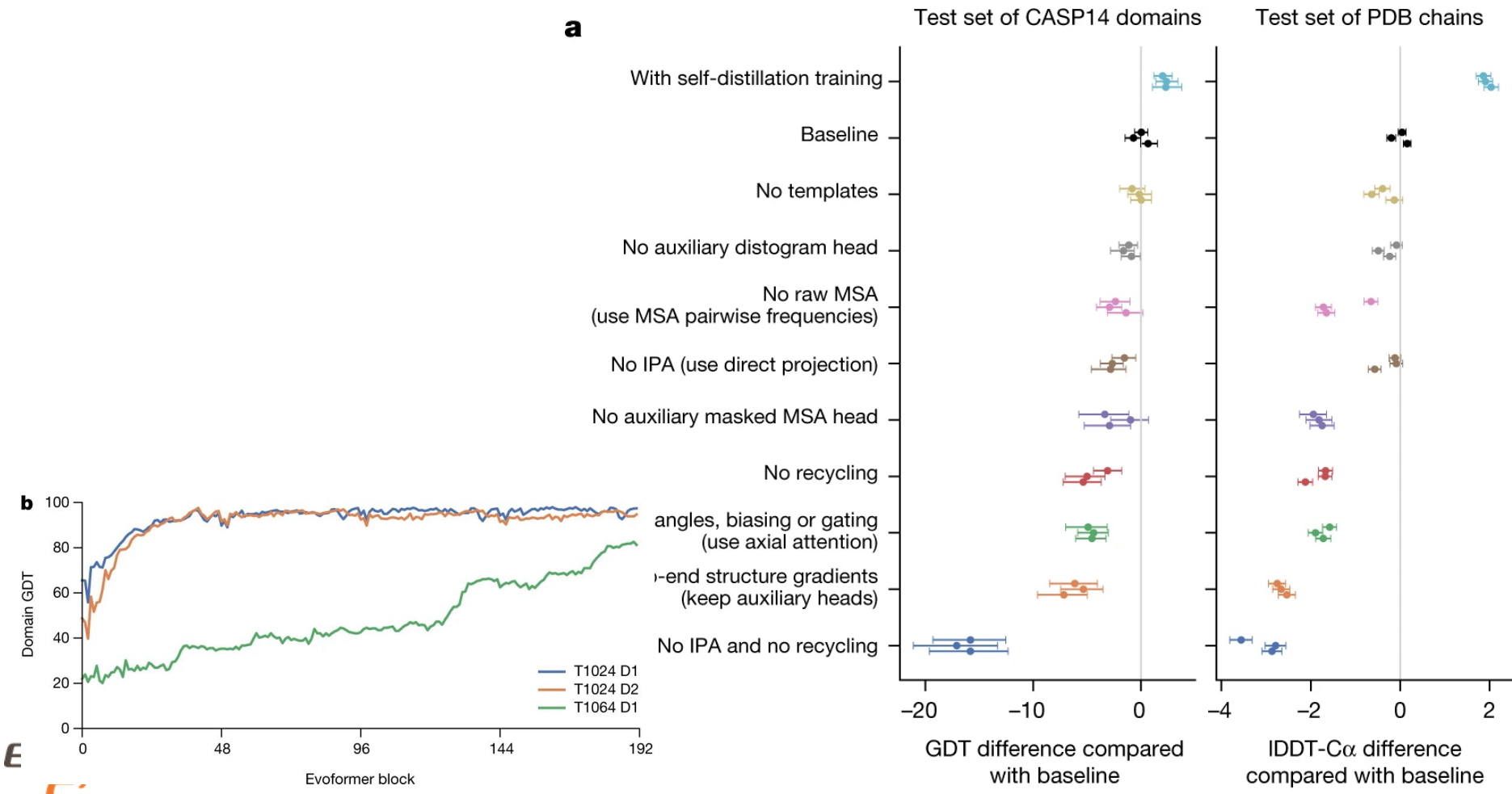
EXTRA SLIDES

AlphaFold

Architectural details.



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