AlphaFold: predicts or recognizes the protein structure?

XXVIII Symposium on Bioinformatics and Computer-Aided Drug Discovery

24 May 2022

Dmitry Ivankov
Assistant Professor
Skoltech

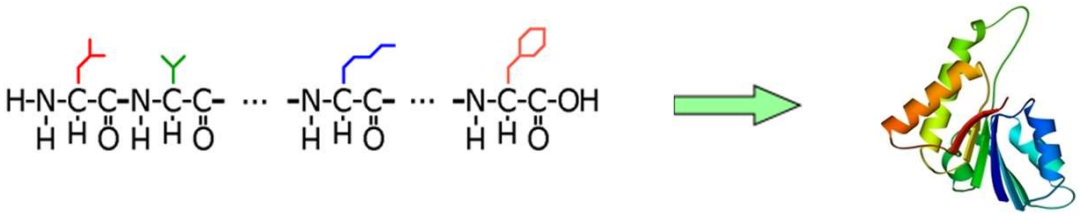
Marina Pak
PhD, 2nd year
Skoltech

Alexei Finkelstein
Professor
Institute of Protein Research

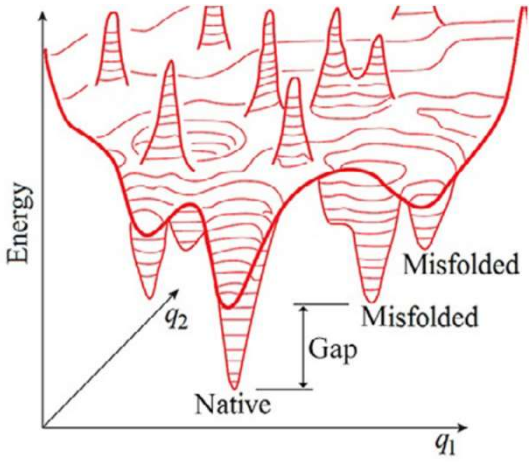
https://ru.wikipedia.org/wiki/Финкельштейн,_Алексей_Витальевич

Anfinsen's experiments

- Protein sequence defines 3D protein structure



- 3D protein structure is the free energy minimum



Therefore:

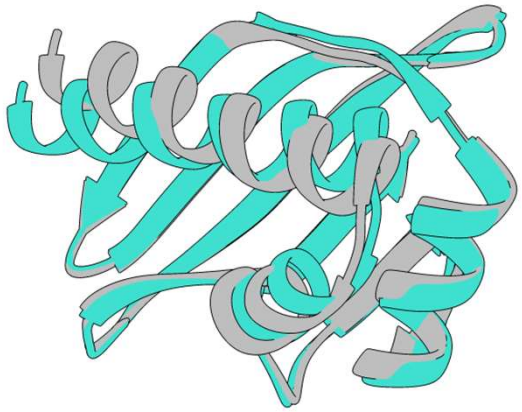
- Having sequence we have all the information to predict 3D structure
- We look for the most stable structure

Anfinsen C. et al. (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. PNAS, 47, 1309–1314.
Anfinsen C. (1973) Principles that govern the folding of protein chains. Science, 181, 223–230.
Finkelstein A., Ptitsyn O. (2016) Protein physics.

Measures of protein 3D structure comparison

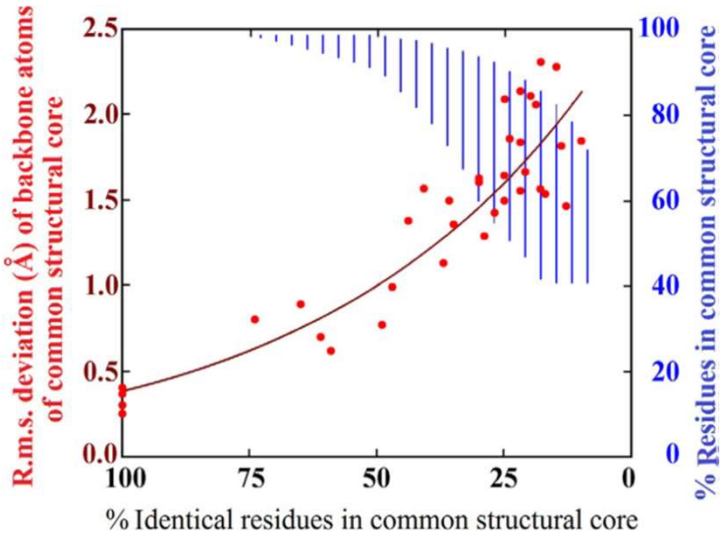
- RMSD – Root Mean Square Deviation
- TM score
- GDT_TS – Global Distance Test Total Score
- LDDT – Local Distance Difference Test

RMSD = 1.55
TM-score = 0.88
GDT-TS-score = 0.86
LDDT = 0.85



Key ideas in protein structure prediction

- Homology modeling
- Threading
- Ab initio folding
- Correlated mutations
- Molecular dynamics



Query: structure unknown

```
Query  133  RFIINDWVKTHTKGMISNLLGKGAVDQLTRLVLVNALYFNGQWKTPFPDSSTHRRLFHKS  192
            R +IN W   HT GMIS  L  G + +LTRLV +NAL+F+G WKTPF   +T  +LFH
Sbjct  125  RQVINSWTSDHTDGMISEFLPSGVLSELTRLVFLNALHFHGVWKTPFDPRNTREQLFHTV  184
```

Subject has known structure

- Copy-paste subject structure
- Rename residues
- Rebuild side-chains
- Make relaxation

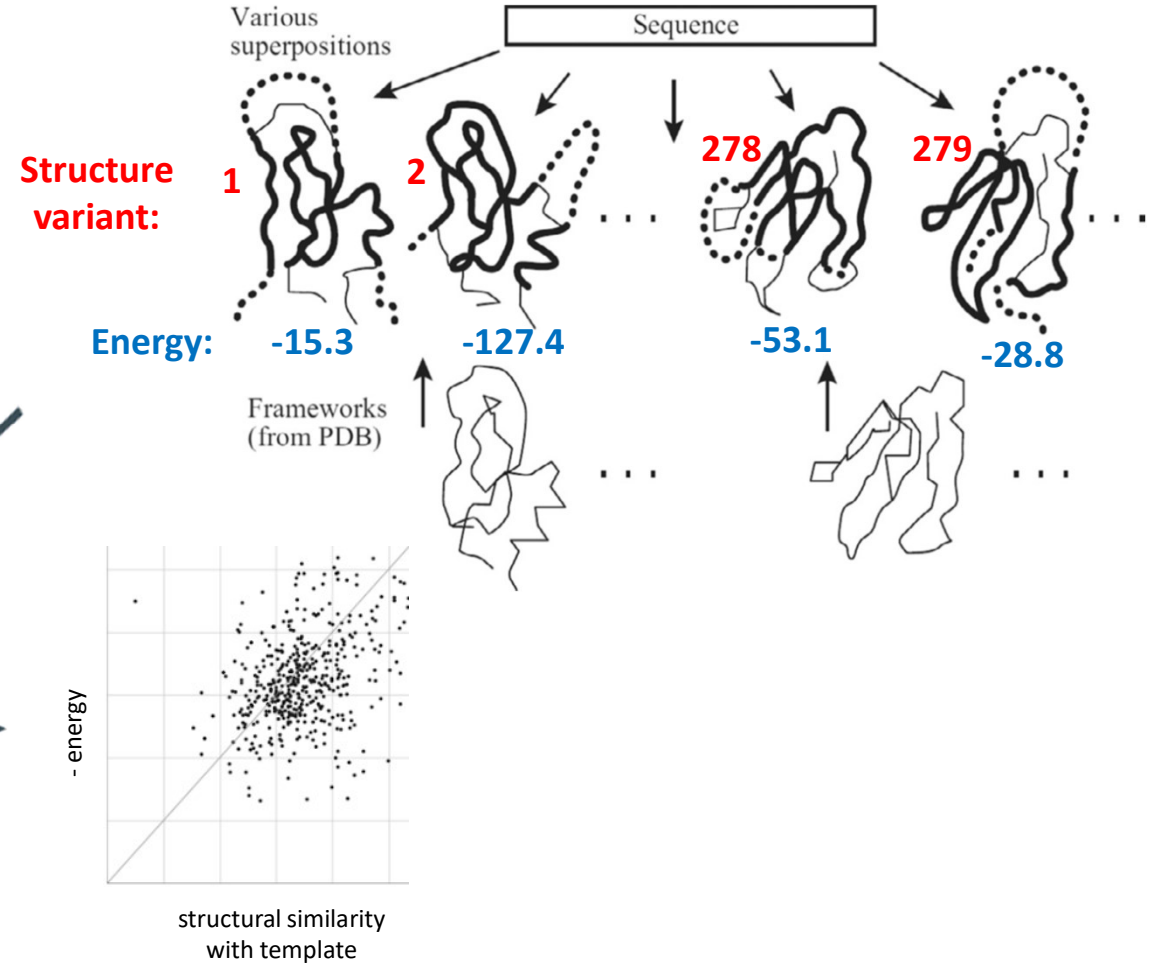
Predicted query structure

Lesk A., Chothia C. (1986) The relation between the divergence of sequence and structure in proteins. EMBO J., 5, 823-826.
Finkelstein A., Ptitsyn O. (2016) Protein physics.

Key ideas in protein structure prediction

- Homology modeling
- Threading
- Ab initio folding
- Correlated mutations
- Molecular dynamics

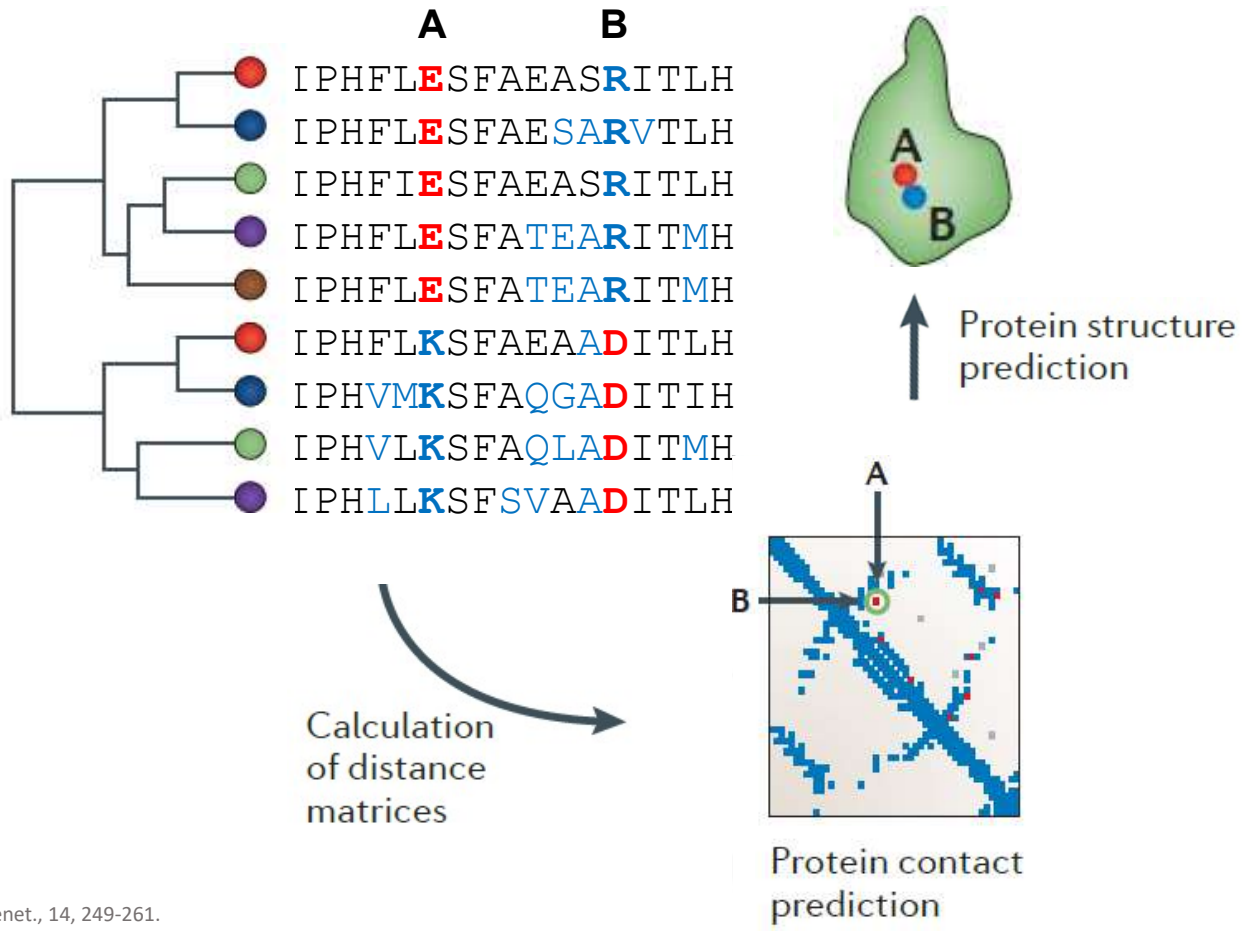
Structure variant	Energy
1	-15.3
2	-127.4
…	…
278	-53.1
279	-28.8
…	…



Finkelstein A., Ptitsyn O. (2016) Protein physics.

Key ideas in protein structure prediction

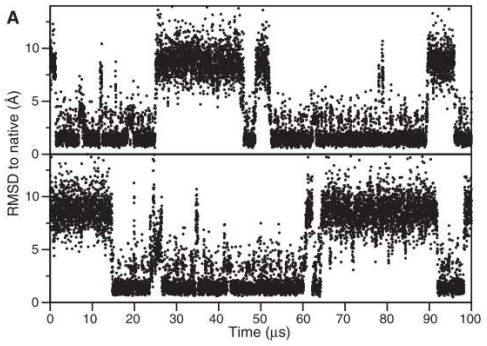
- Homology modeling
- Threading
- Ab initio folding
- **Correlated mutations**
- Molecular dynamics

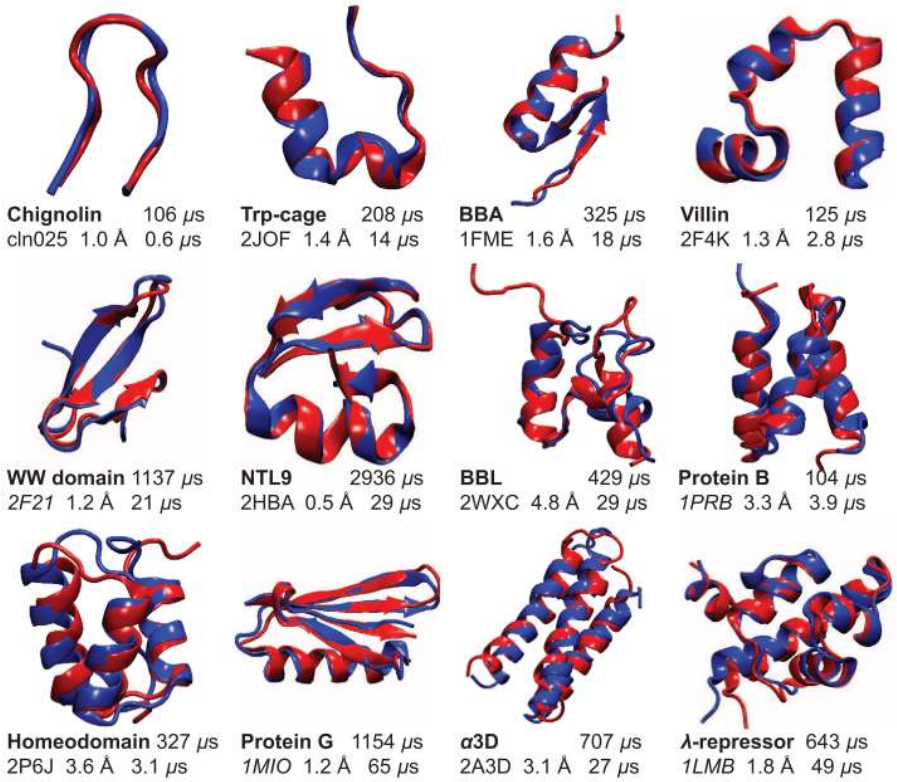


de Juan et al. (2013) Emerging methods in protein co-evolution. Nat. Rev. Genet., 14, 249-261.

Key ideas in protein structure prediction

- Homology modeling
- Threading
- Ab initio folding
- Correlated mutations
- Molecular dynamics



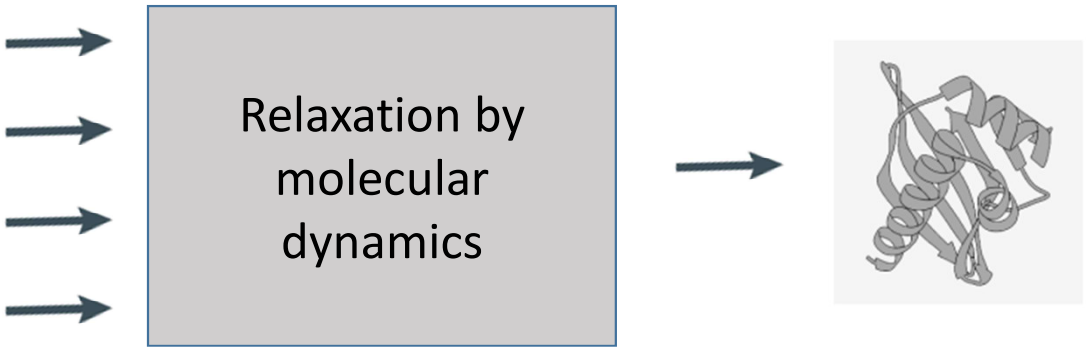


Shaw D. et al. (2010) Atomic-level characterization of the structural dynamics of proteins. Science, 330, 341-346.

Linforff-Larsen K. et al. (2011) How fast-folding proteins fold. Science, 334, 517-520.

Key ideas in protein structure prediction

- Homology modeling
- Threading
- Ab initio folding
- Correlated mutations
- Molecular dynamics



Relaxation box contains potential energy function, e.g. Amber99sb force field

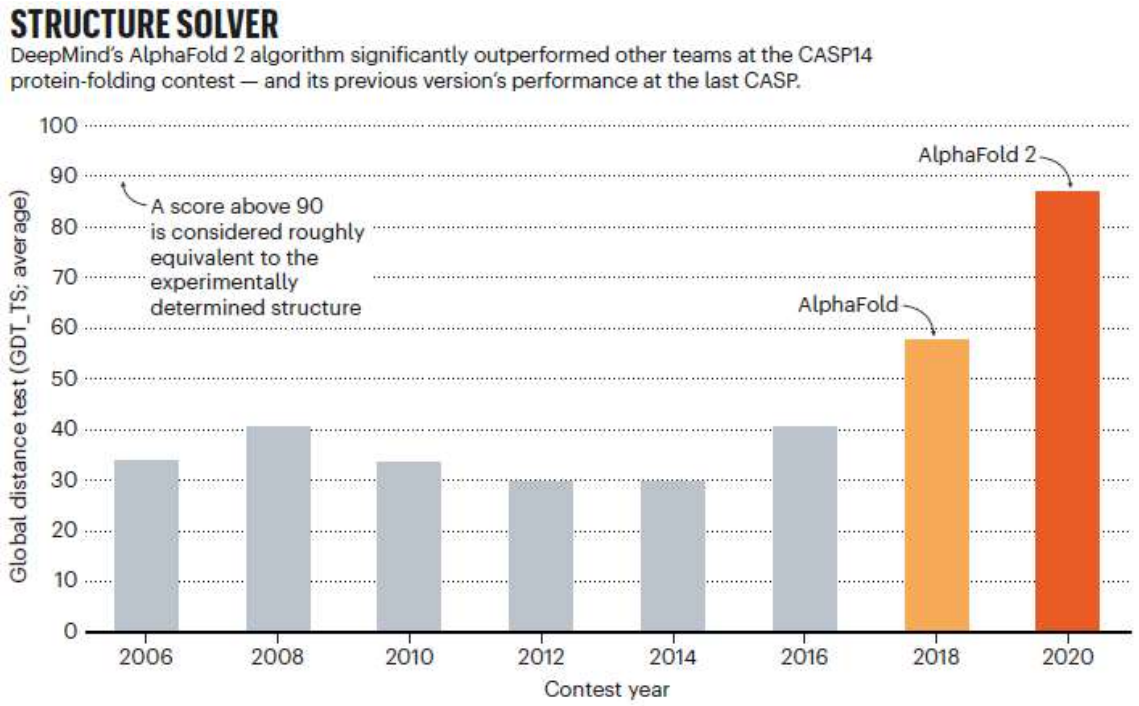
Structure without clashes

Note:

- AlphaFold uses relaxation as well.

CASP experiment

- CASP: Critical Assessment of Protein Structure Prediction
- Since 1994 bi-annual blind competition on protein structure prediction

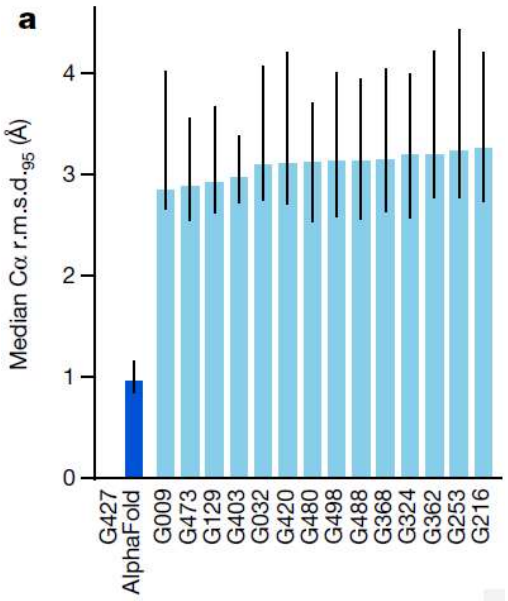


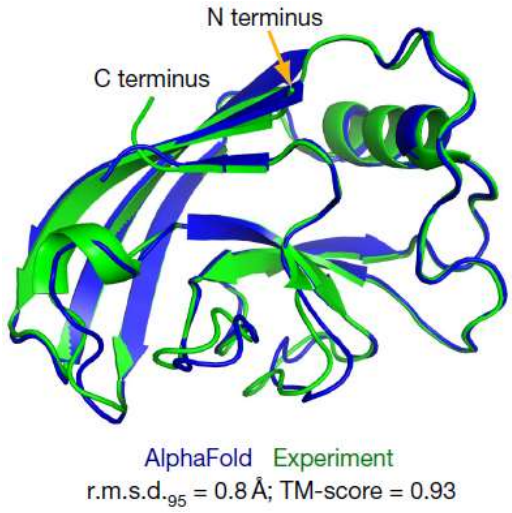
https://predictioncenter.org

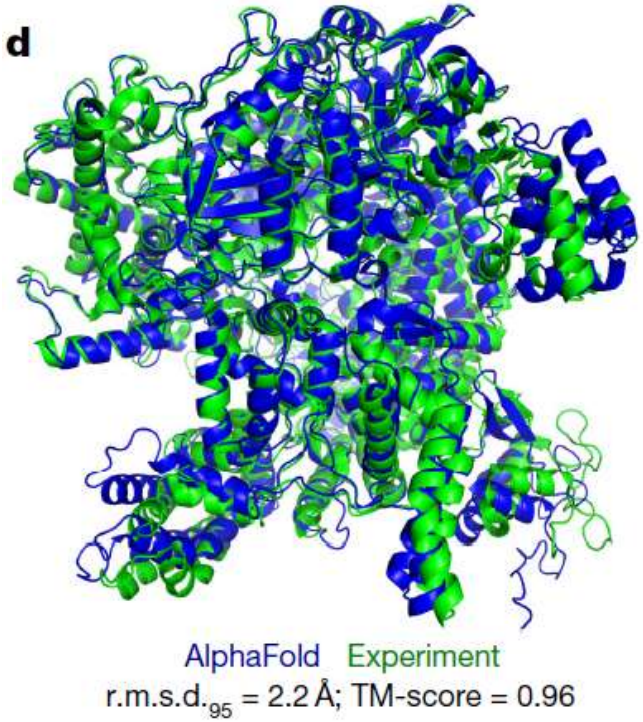
Callaway E. (2020) 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures Nature, 588, 203–204.

AlphaFold performance in CASP14

- Deep learning algorithm
- Trained on PDB structures published before April 30, 2018
- Uses multiple sequence alignments (MSA) and PDB







Jumper J. et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589.

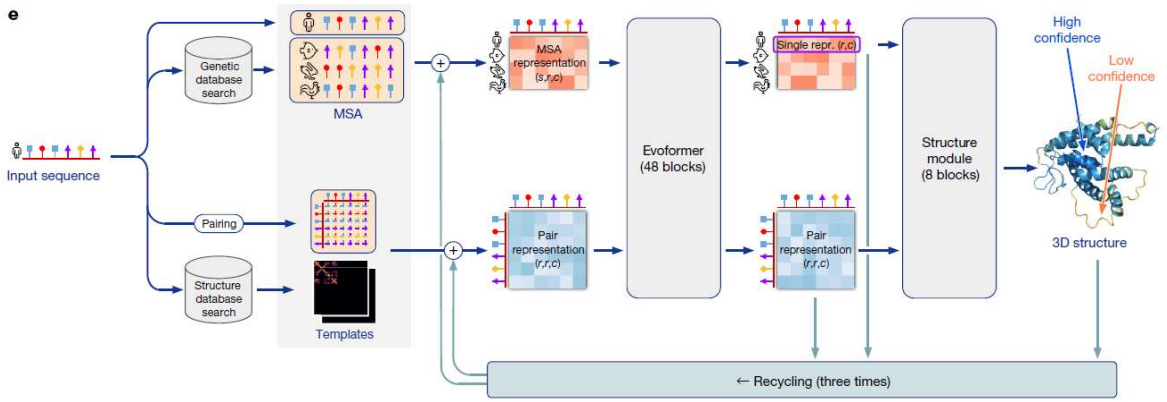
Questions:

- What is the main reason for this success?
- What does AlphaFold actually do:
  - does it predict 3D protein structure from the physics of protein chain,

  or
  - does it recognize the 3D structure by the similarity of the amino acid sequence in question to sequences with already known 3D structures?

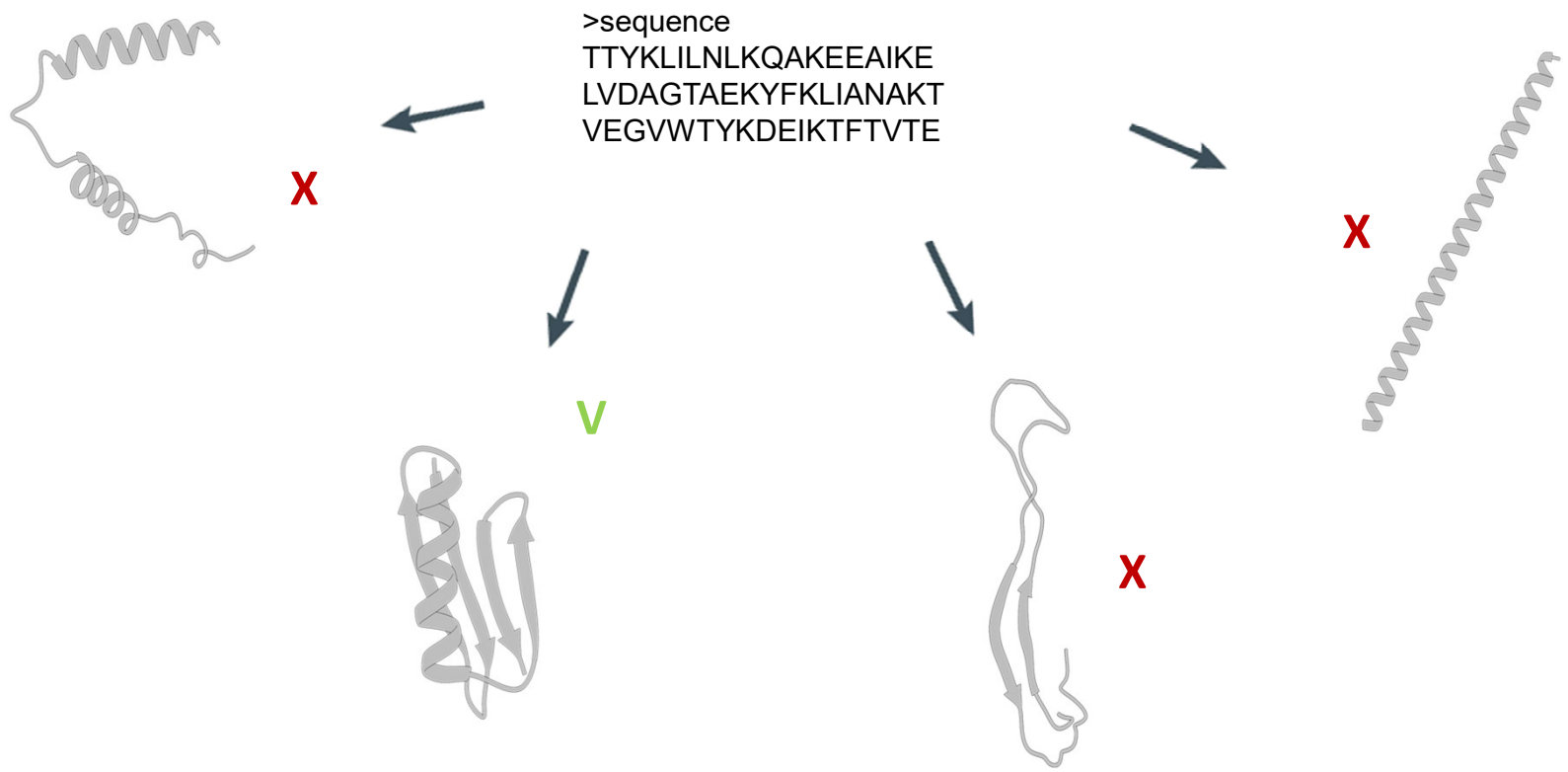
Notes:

- No way to ask AlphaFold directly
- We do not ask about the physics that is in the relaxation module: [almost] every tool uses it



Jumper J. et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589.

How do we understand physics of proteins?

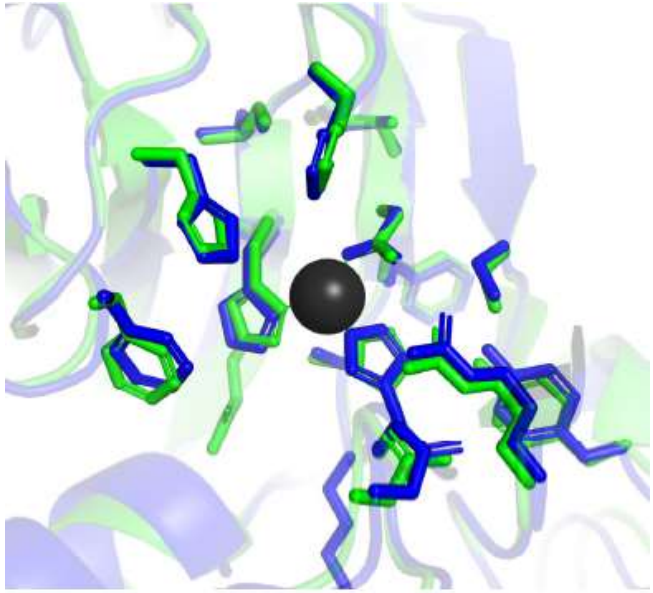


Note:

- We ask about “extra” physics that allows AlphaFold to outperform other tools

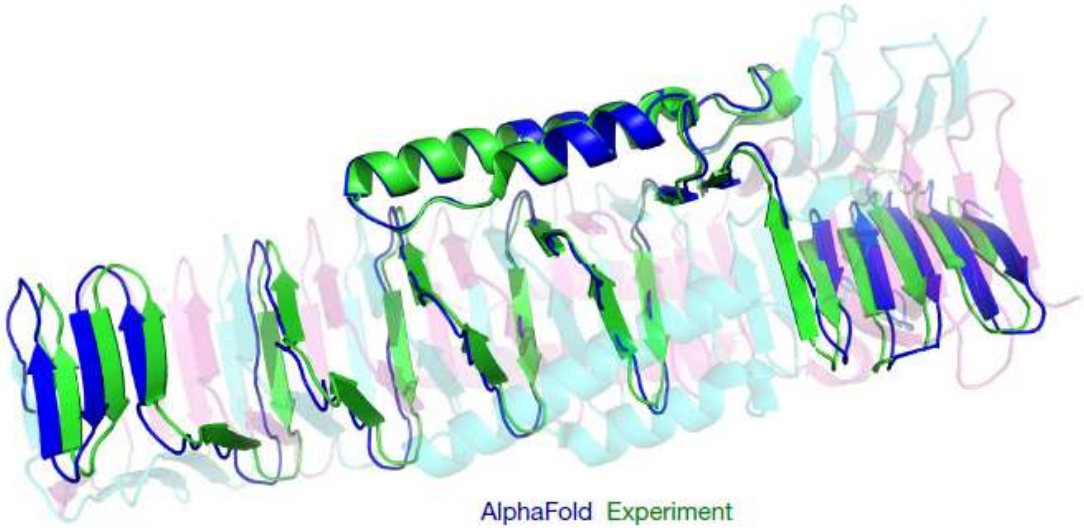
Reasons to be sceptic: physics looses to statistics

Perfect prediction in the absence of metal ion



AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn

One chain from a multimer



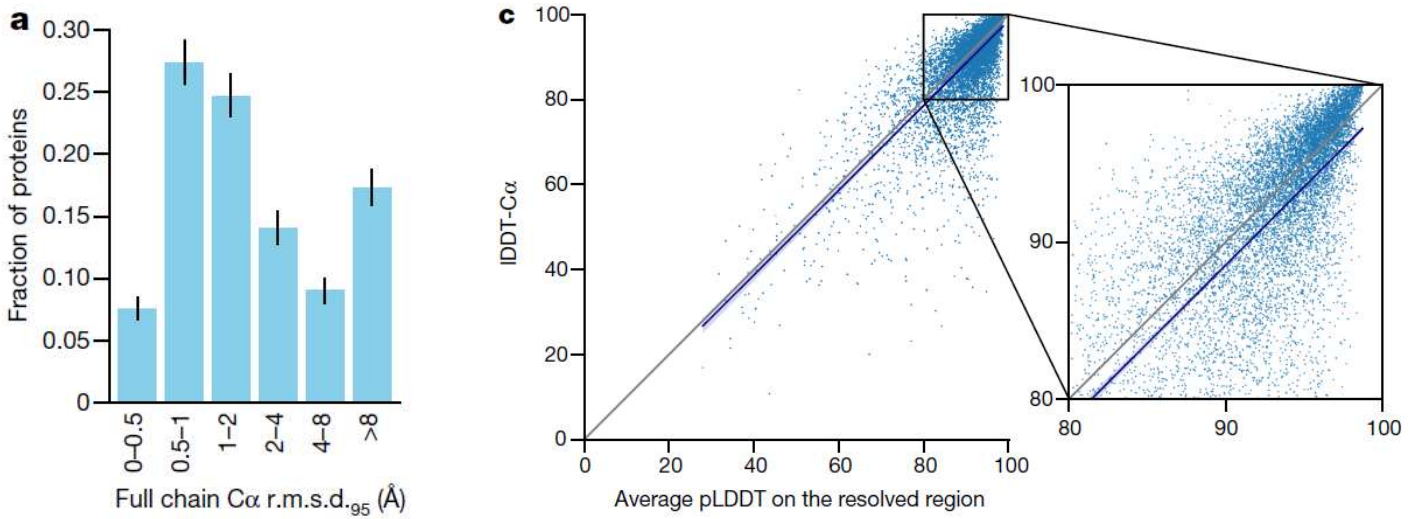
AlphaFold Experiment

"An intertwined homotrimer (PDB 6SK0) is correctly predicted without input stoichiometry and only a weak template (blue is predicted and green is experimental)."

Jumper J. et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589.

AlphaFold performance for unseen proteins

- Proteins from PDB:
  - after April 30, 2018
  - id < 40% covering more than 1% of the sequence



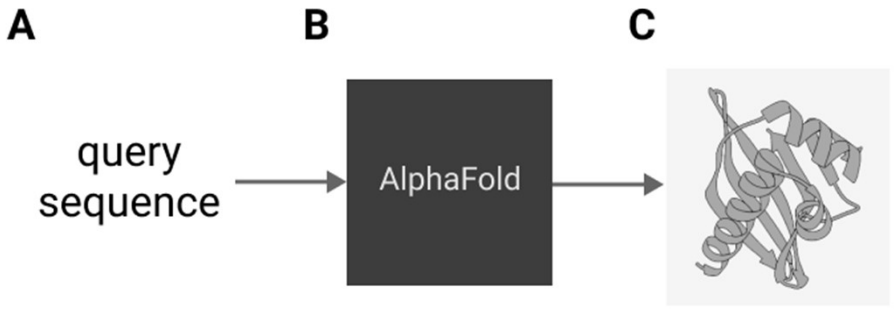
Still:

- Multiple sequence alignments were used
- There could be some structures with similar 3D structures but dissimilar sequences

Jumper J. et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589.

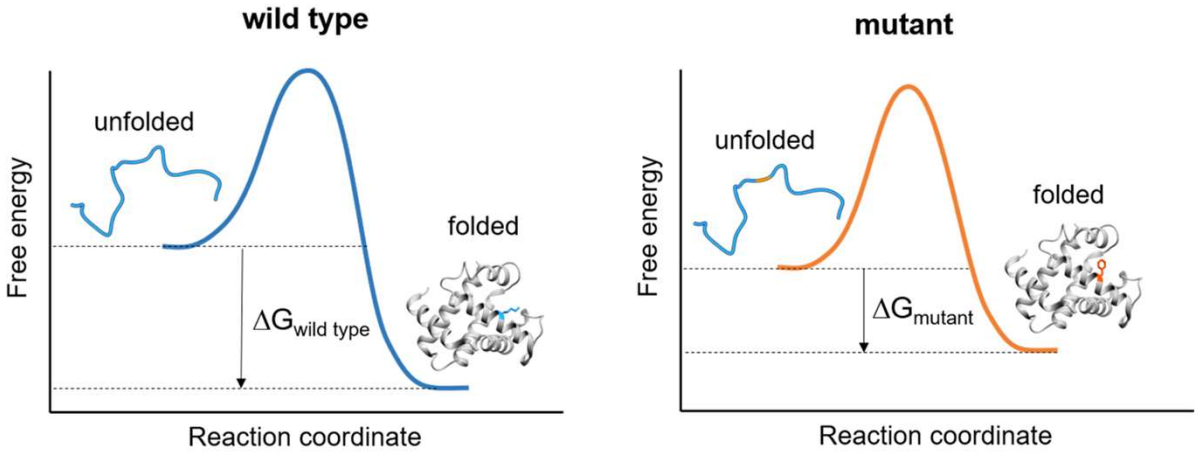
What do we have from AlphaFold?

- Output structure
- pLDDT: per-residue predicted LDDT
- Average pLDDT
- pTM: Predicted TM-score (highly correlates with average pLDDT)



Do AlphaFold metrics correlate with ΔΔG?

- Disclaimer of AlphaFold:
  - "[AlphaFold] has not been validated for predicting the effect of mutations"
- However, native structure is native because it is the most stable.
- David Jones with colleagues:
  - *"Amino acids in the sequence that lead to low confidence predictions are less likely to lead to a stable structures."*

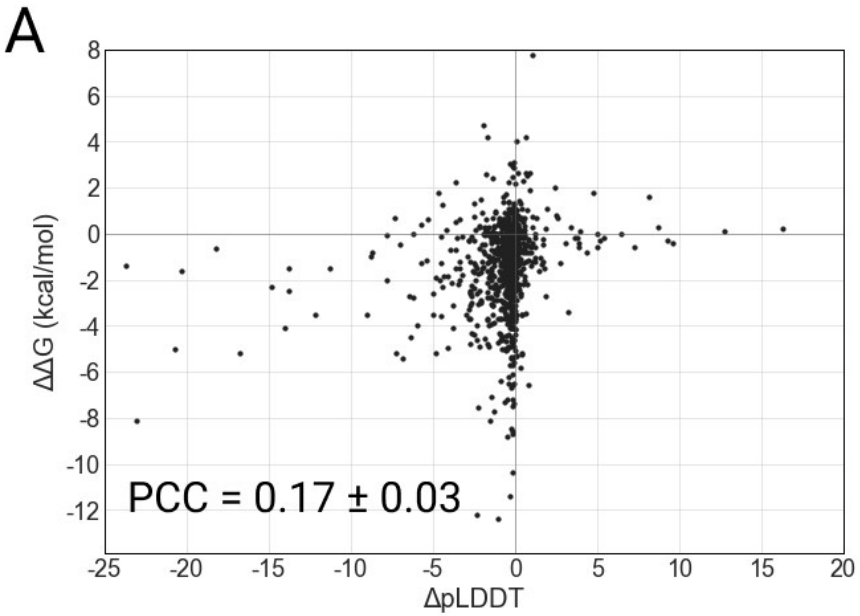


https://alphafold.ebi.ac.uk/faq

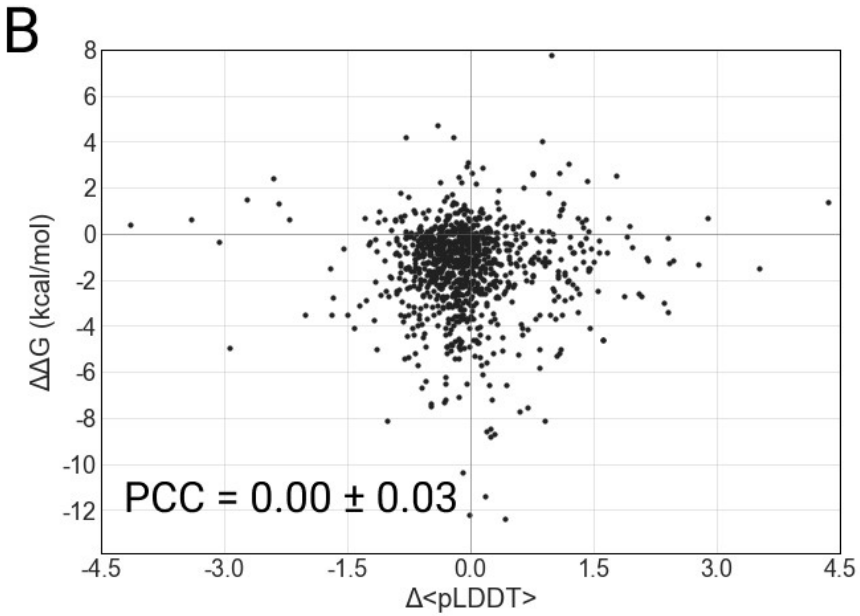
Moffat L. et al. (2021) Using AlphaFold for Rapid and Accurate Fixed Backbone Protein Design. bioRxiv, https://doi.org/10.1101/2021.08.24.457549

AlphaFold pLDDT vs. ΔΔG

pLDDT of the mutated residues



Average pLDDT of the whole structre



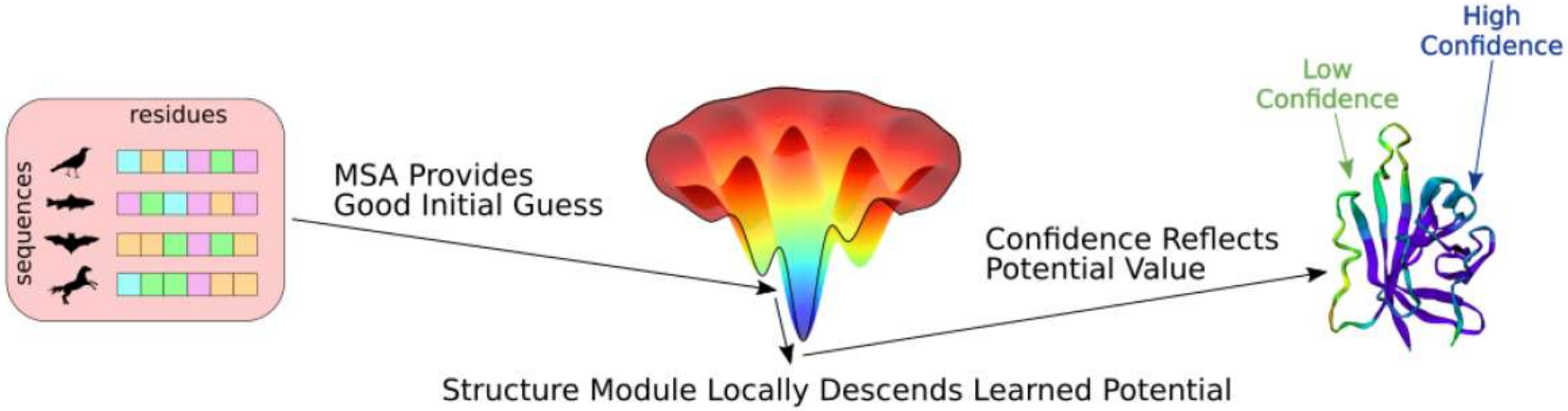
AlphaFold cannot predict the energy changes due to single mutations

Pak M. et al. (2021) Using AlphaFold to predict the impact of single mutations on protein stability and function. bioRxiv, https://www.biorxiv.org/10.1101/2021.09.19.460937

What about far distances?

Roney and Ovchinnikov:

- Hypothesis: "*AlphaFold has learned an accurate potential function … but … the MSA is necessary to locate an approximate global minimum*"

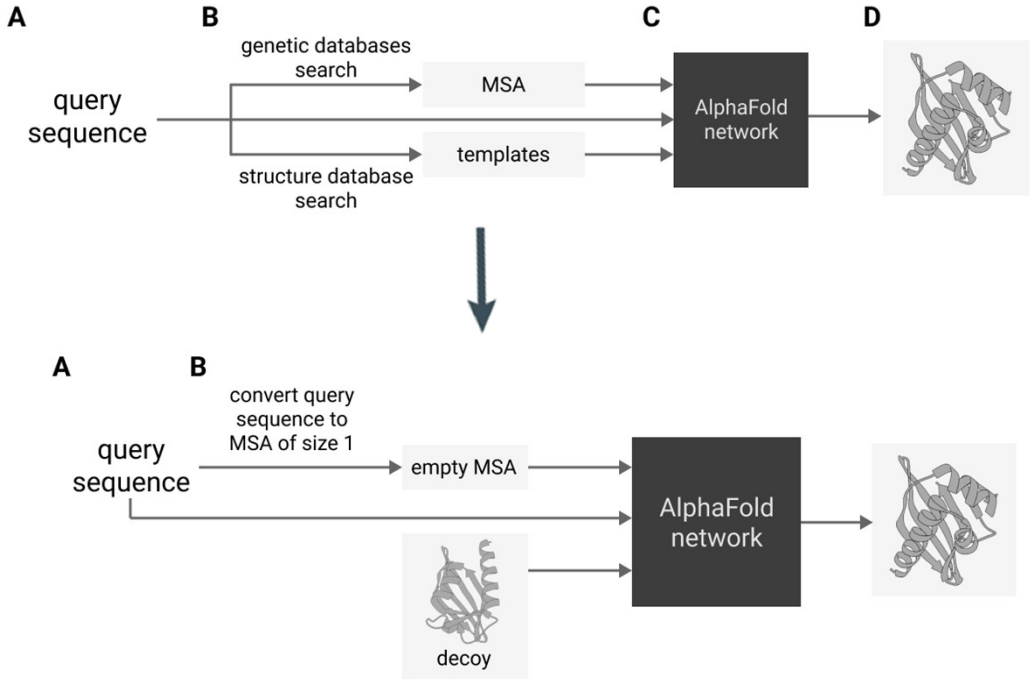


Solution: "*to score the plausibility of the target amino acid sequence adopting the geometry given by the decoy structure.*"

Roney J., Ovchinnikov S. (2022) State-of-the-art estimation of protein model accuracy using AlphaFold. bioRxiv, doi: 10.1101/2022.03.11.484043.

What about far distances?

- AlphaFold: if to look under the hood:



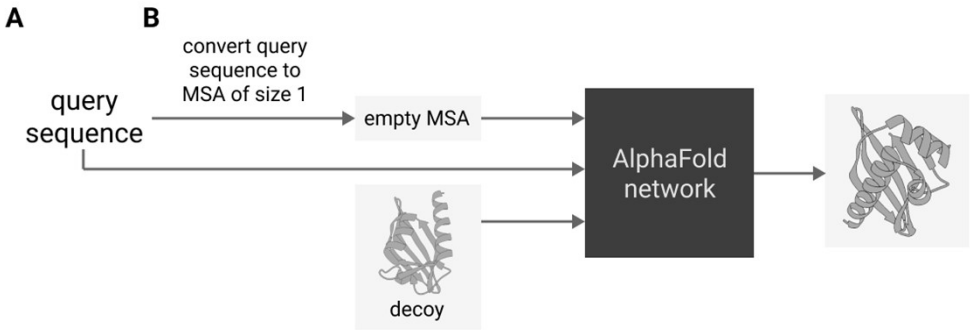
- Scenarios:
  - default
  - no MSA
  - no templates
  - no MSA, no templates
- Decoy structure:
  - query structure
  - alpha-helical structure
- Decoy sequence:
  - query sequence
  - poly-alanine sequence

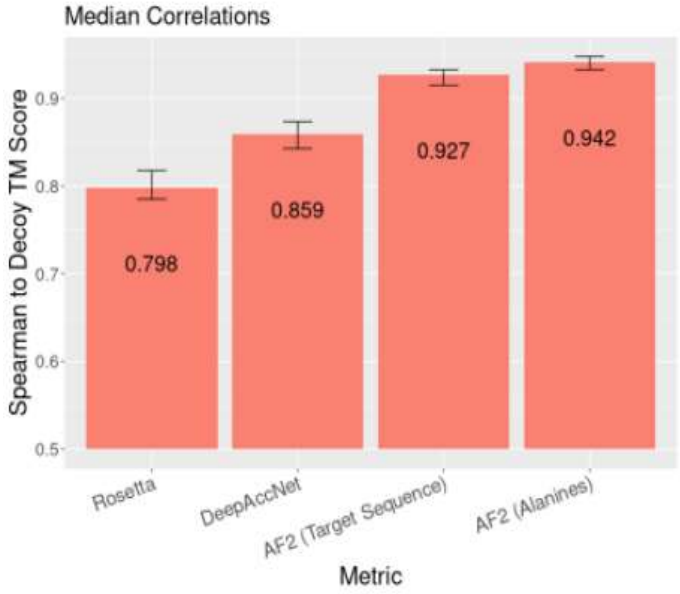
Jumper J. et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589.
Roney J., Ovchinnikov S. (2022) State-of-the-art estimation of protein model accuracy using AlphaFold. bioRxiv, doi: 10.1101/2022.03.11.484043.

What about far distances?

Set of proteins:

- “Novel fold” proteins
- Decoys from Rosetta decoy set





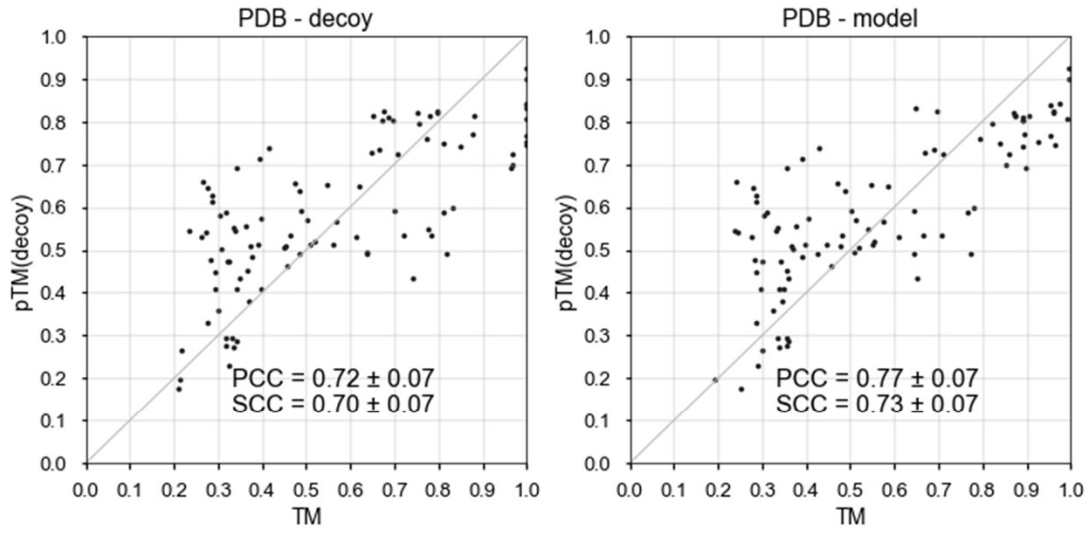
BUT: AlphaFold has seen both proteins’ and decoy structures: they are all in PDB

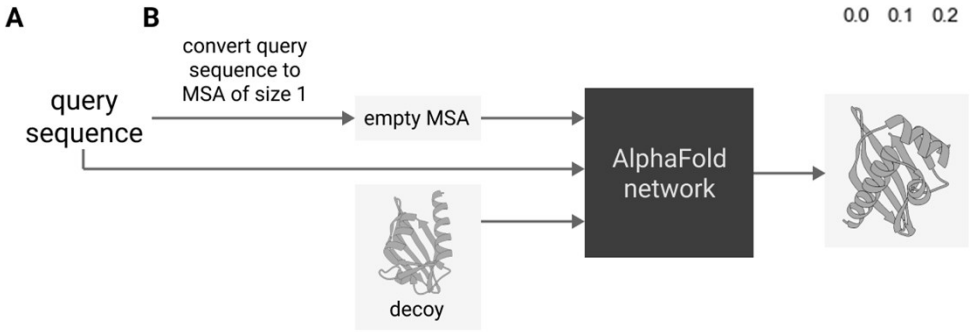
Roney J., Ovchinnikov S. (2022) State-of-the-art estimation of protein model accuracy using AlphaFold. bioRxiv, doi: 10.1101/2022.03.11.484043.

We used different set of proteins

Set of proteins:

- Structures from PDB:
  - Released after April 30, 2018
  - “Novel fold” proteins
  - TM-score to PDB < 0.5





Note: distinguishing closer structures is more important

Is it possible to predict well with no physics?

What is the expected similarity of a random sequence S to the most similar to it chain S' from the set Σ_N of N other random sequences?

or

Is the set Σ_N large enough to include a sequence S', which is so similar to S that their 3D structures are very similar?

Is it possible to predict well with no physics?

Probability that the random sequence S_n of the length n matches in m positions another random sequence of the same length n is

$$P_{m,n} = \frac{n!}{m!\,(n-m)!} p^m (1-p)^{n-m}$$

Stirling's approximation:

$$\frac{n!}{m!(n-m)!} p^m (1-p)^{n-m} \approx \left(\frac{pe}{m/n}\right)^m e^{-pn}$$

And:

$$\left(\frac{M/n}{pe}\right)^{\frac{M/n}{pe}} = N^{\frac{1}{npe}} e^{-1/e}$$

Is it possible to predict well with no physics?

Domains of $\boldsymbol{n} \approx 100$:

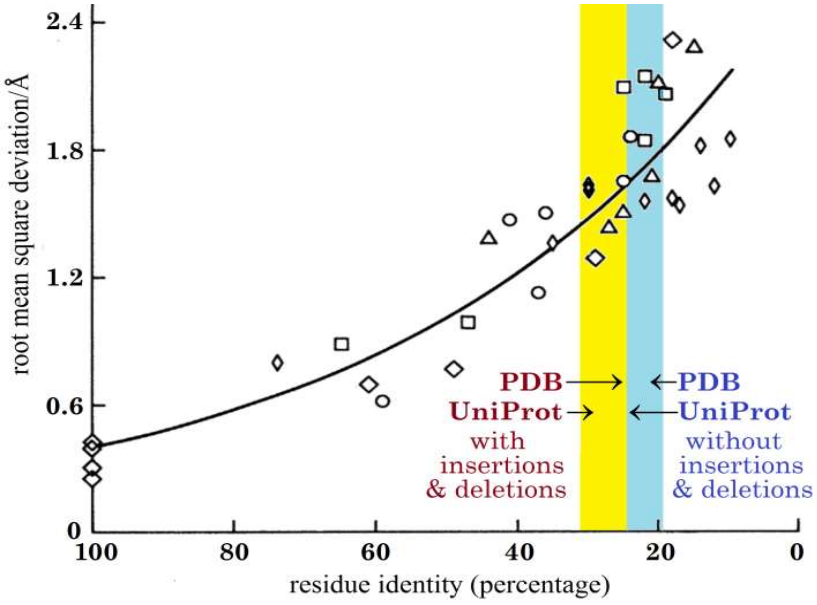
PDB: $\boldsymbol{N} \approx 1.6 \cdot 10^5$ => $\boldsymbol{M/n} \approx 0.19$

UniProt: $\boldsymbol{N} \approx 2 \cdot 10^8$ => $\boldsymbol{M/n} \approx 0.24$

19% and 24%

With insertions/deletions this shifts to

25% and 32%



Conclusions

- From structurally close proteins AlphaFold cannot choose more stable structure at all
- AlphaFold's ranking of structurally different proteins seems to be comparable with other methods
- The conceptual reason of tremendous AlphaFold success is that databases cover (almost) all protein superfamilies existing in nature
- Overall, we stick to the null hypothesis:
  AlphaFold does not know energy potential function better than other programs

AlphaFold: predicts or recognizes the protein structure?

XXVIII Symposium on Bioinformatics and Computer-Aided Drug Discovery

24 May 2022

Dmitry Ivankov
Assistant Professor
Skoltech

Marina Pak
PhD, 2nd year
Skoltech

Alexei Finkelstein
Professor
Institute of Protein Research

https://ru.wikipedia.org/wiki/Финкельштейн,_Алексей_Витальевич