Introduction to AlphaFold

Samuel Murail
CMPLI, RPBS, Université Paris Cité

Plan

- CASP/CAPRI
- Alphafold 2
- An ecosystem around Alphafold
- Going Beyond Alphafold
  - Sampling
  - Scoring
  - Pruning
- Alphafold 3

Protein Folding

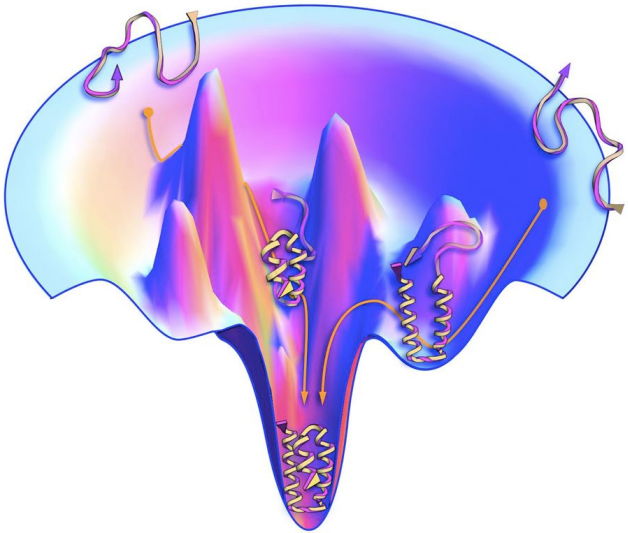
Anfinsen’s dogma :

- native structure is determined only by the protein's aa sequence
- native structure is a unique, stable and kinetically accessible minimum of the free energy

if one could model this energy function with sufficient accuracy, then one could predict protein structures

Issues:

- Accurate model the energy function that governs protein folding in computationally tractable human timsescale
- Searching for the optimum is a difficult global optimization task



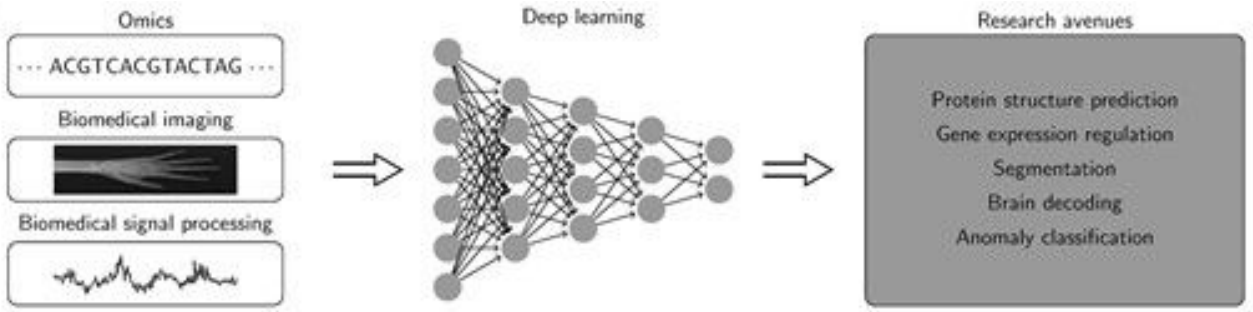
Deep Learning

Problem:

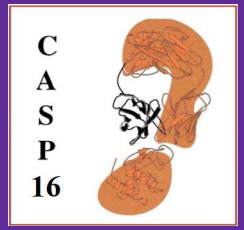
- High-throughput analysis
- Massive set of annotated data (genomic, images ...)
- Too large and too complex to be understood by the human brain.

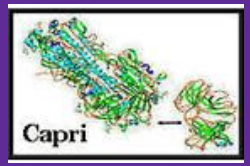
Applications:

- Processing of high-dimensional biological data
- Classification
- Predictive tool
- Image analysis, genomics, drug discovery, ...
-



CASP/CAPRI



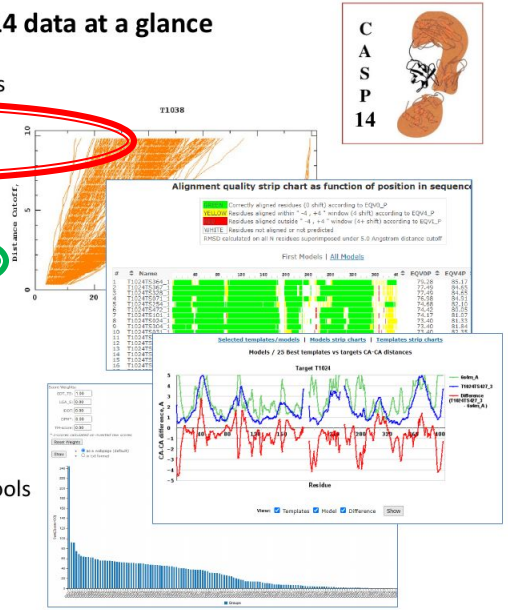


CASP is a Big Deal

Reference for structure prediction
Around for over 25 years

CASP14 data at a glance

- >200 prediction methods
- ~100 research centers
- ~350 predictors
- >80 targets
- >67,000 models
- >5,000,000 scores
- ~430 GB of data
- >30 different software tools
- >20 visualization tools



CAPRI was modeled after CASP; specialized in the prediction of protein complexes

Slide Courtesy: Marc F. Lensink

What is CAPRI/CASP?

CAPRI		CASP	
Since 2001		Since 1994	
Critical Assessment of PRedicted Interactions		**Critical Assessment of Structure Predictions**	
Joint prediction rounds since 2014:			
25 Targets	Round 30	CASP11	2014
10 Targets	Round 37	CASP12	2016
21 Targets	Round 46	CASP13	2018
12 Targets	Round 50	CASP14	2020
37 Targets	Round 54	CASP15	2022
30+ Targets	Round 57	CASP16	2024
Prediction rounds on a "rolling" basis		Prediction season	
Fits with publication schedule		Intense 2 to 3 months	
3 to 4 weeks per prediction round			
Fixed assessor team, established metrics		Varying assessors, varying metrics	
Difference in targets			
Mostly hetero-dimers or -trimers Peptides, sugars, water positions		Mostly obligate, many homo-oligomers Very large assemblies	
Incites method development		**Large-scale testing of methodologies**	

Slide Courtesy: Marc F. Lensink

Metrics

- RMSD

$$\mathrm{RMSD}(\mathbf{v},\mathbf{w}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|v_i - w_i\|^2}$$
$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{ix}-w_{ix})^2 + (v_{iy}-w_{iy})^2 + (v_{iz}-w_{iz})^2)}$$

- TMScore

$$\text{TM-score} = \max\left[\frac{1}{L_{\text{target}}}\sum_{i}^{L_{\text{common}}}\frac{1}{1+\left(\frac{d_i}{d_0(L_{\text{target}})}\right)^2}\right]$$

  - weights smaller distance errors more strongly than larger distance errors
  - L_{target} is the length of the target sequence , and L_{common} is the number of residues that appear in both structures. d_i is the distance between the i th pair of residues in the template and target structures, and d_0 a distance scale that normalizes distances ~28.

- GDT (global distance test), GDT_TS for “total score” ranging from 0 to 100
  - % of 20 consecutive distance cutoffs of Calpha atoms (0.5 Å, 1.0 Å, 1.5 Å, ... 10.0 Å)
  - The conventional GDT_TS total score in CASP is the average result of cutoffs at 1, 2, 4, and 8 Å.

Metrics (2)

- LDDT: Local Distance Difference Test (lDDT) measures how well the environment in a reference structure is reproduced in a protein model (bad 0 to 100 perfect)
  - "computed over all pairs of atoms in the reference structure at a distance closer than a predefined threshold R_o (called inclusion radius), and not belonging to the same residue" ,
  - A pair is considered conserved if it is within a threshold distance (0.5 Å, 1 Å, 2 Å and 4 Å)

- DockQ (Basu and Wallner 2016 Plos One)

$$DockQ(F_{nat}, LRMS, iRMS, d_1, d_2) = (F_{nat} + RMS_{scaled}(LRMS, d_1) + RMS_{scaled}(iRMS, d_2))/3$$

With:

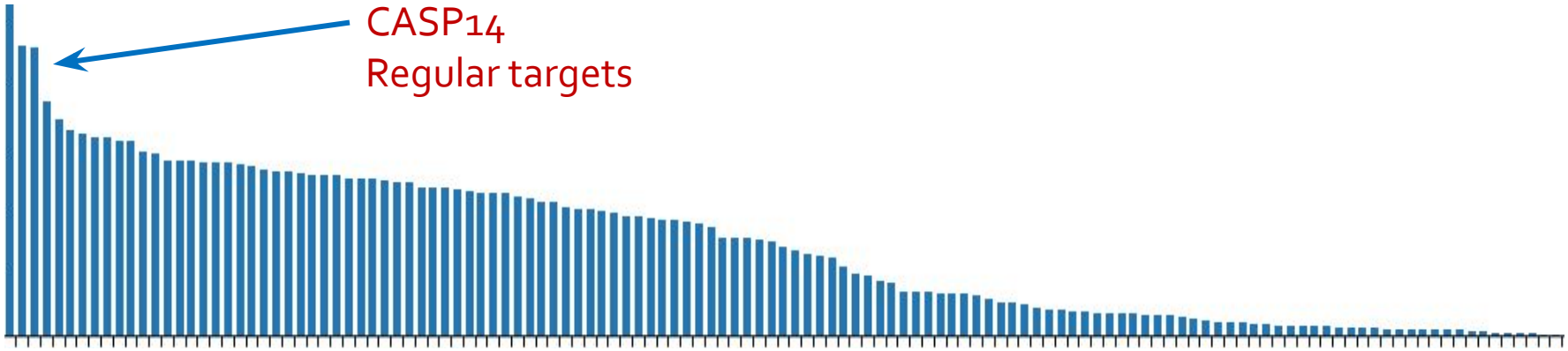
$$RMS_{scaled}(RMS, d_i) = \frac{1}{1 + \left(\frac{RMS}{d_i}\right)^2}$$

d_1 = 8.5Å and d_2 = 1.5Å

CASP13 (2018)

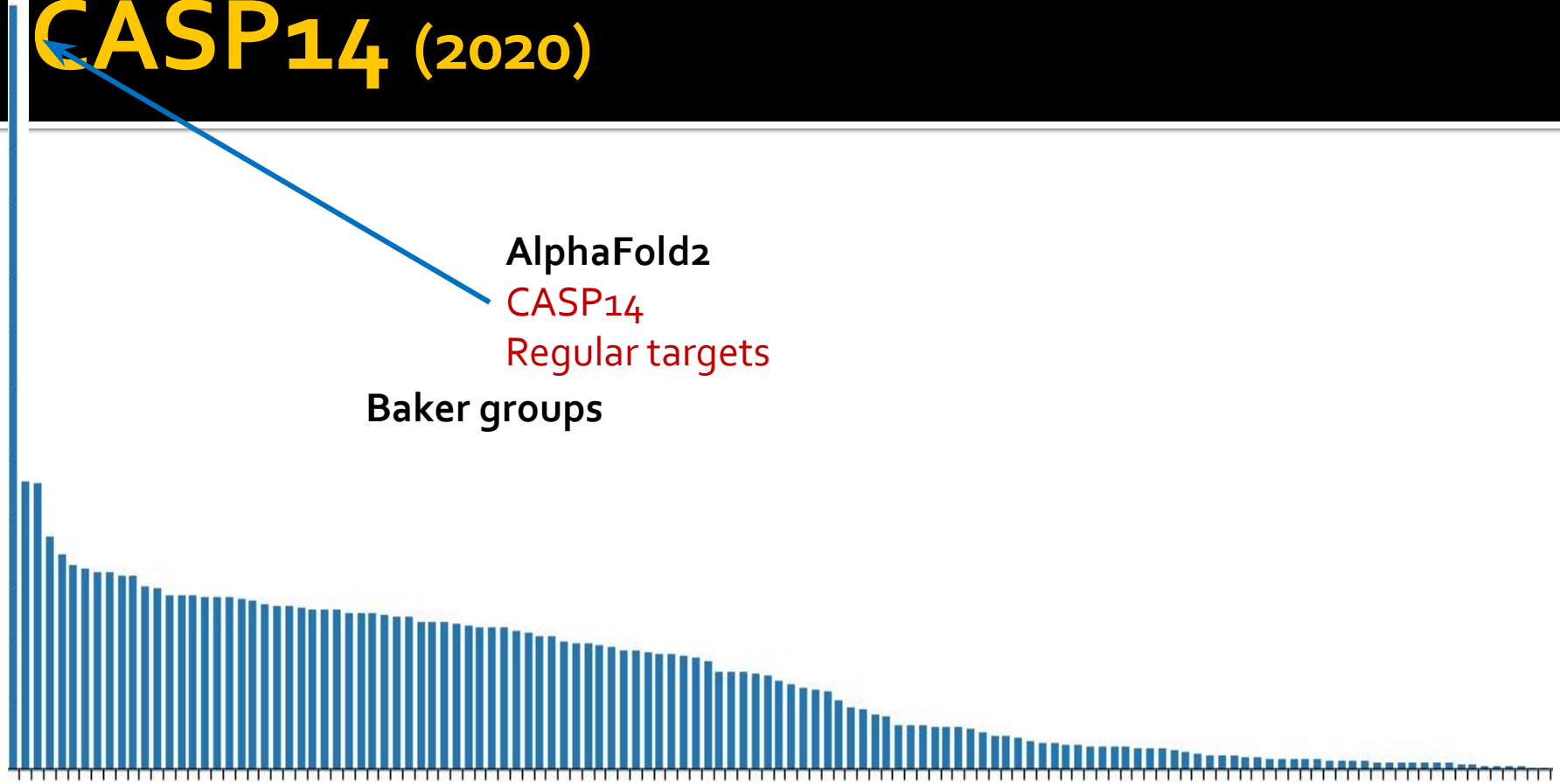
Slide Courtesy: Marc F. Lensink

CASP14 (2020)



Slide Courtesy: Marc F. Lensink

CASP14 (2020)



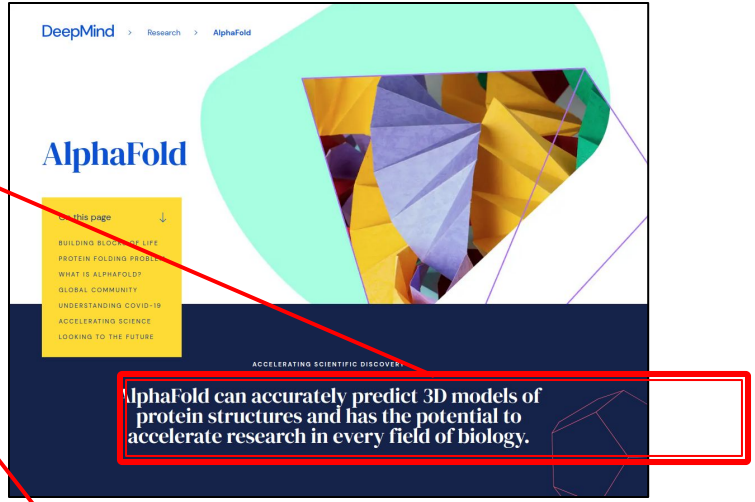
Slide Courtesy: Marc F. Lensink

Alphafold 2

A revolution ?

"AlphaFold can accurately predict 3D models of protein structures and has the potential to accelerate research in every field of biology."

"AlphaFold: a solution to a 50-year old grand challenge in biology."



www.nature.com › news

'It will change everything': DeepMind's AI makes gigantic leap ...

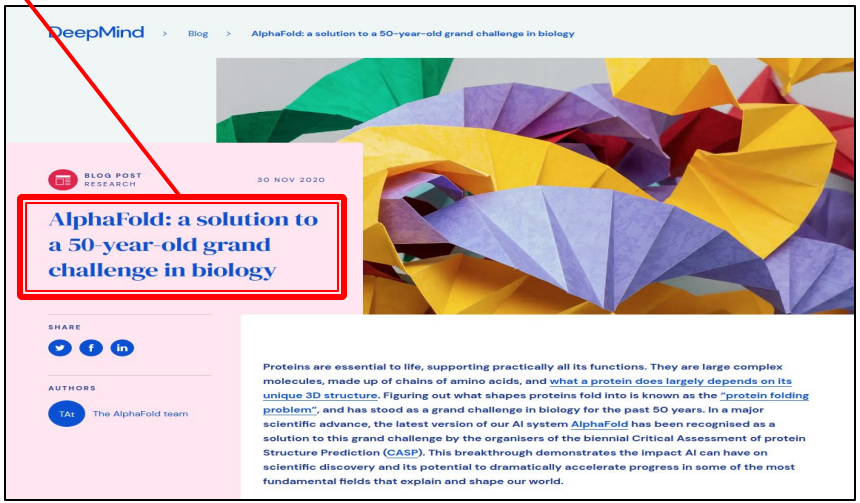
Nov 30, 2020 — DeepMind's **AlphaFold 2** algorithm outperformed other teams at the CASP14 protein. Source: DeepMind. DeepMind's 2018 performance at ...

deepmind.com › blog › article › alphafold-a-solution-t...

AlphaFold: a solution to a 50-year-old grand challenge in ...

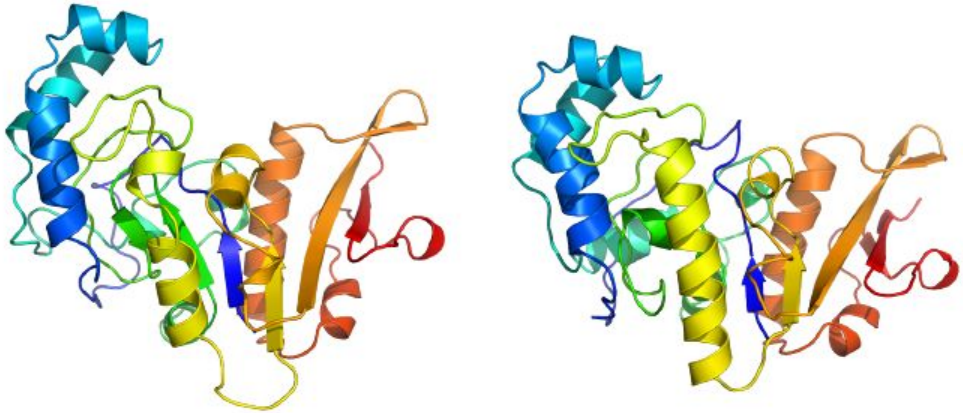
Nov 30, 2020 — **AlphaFold**: The making of a scientific breakthrough · Improvements in the median accuracy of predictions in the free modelling category for the ...

AlphaFold: Using AI for ... · AlphaFold · Computational predictions of ...



Slide Courtesy: Marc F. Lensink

So how well did they really do?



'So, either this group is close to solving the folding problem or they cheated somehow.'

Nick Grishin

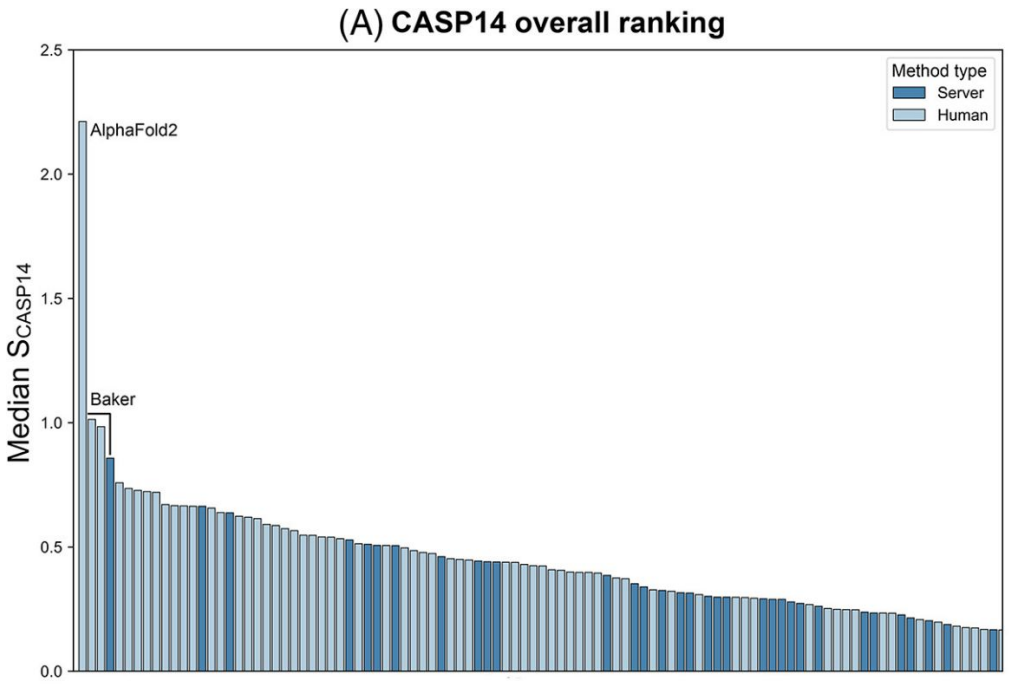
Slide Courtesy: Marc F. Lensink

CASP 14

Novembre 2020, DeepMind indisputably won the CASP14 competition.

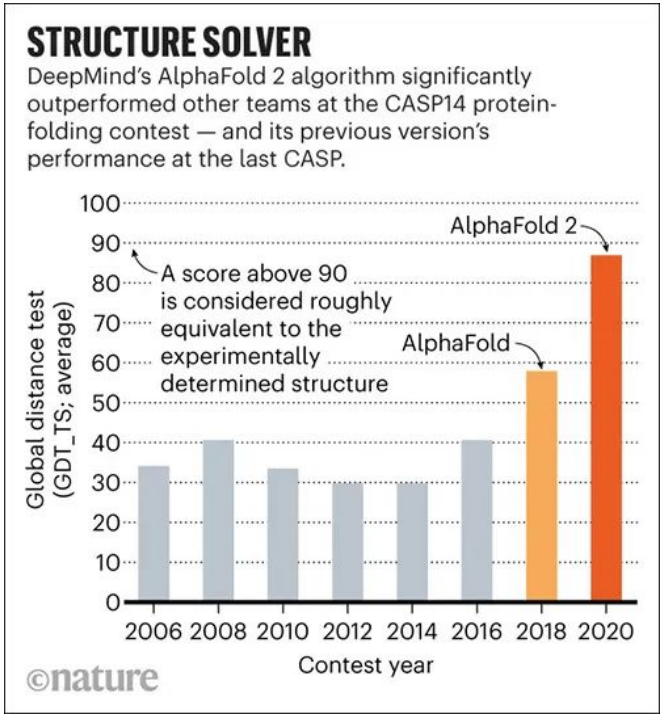
We add to wait until summer of 2021 to access to:

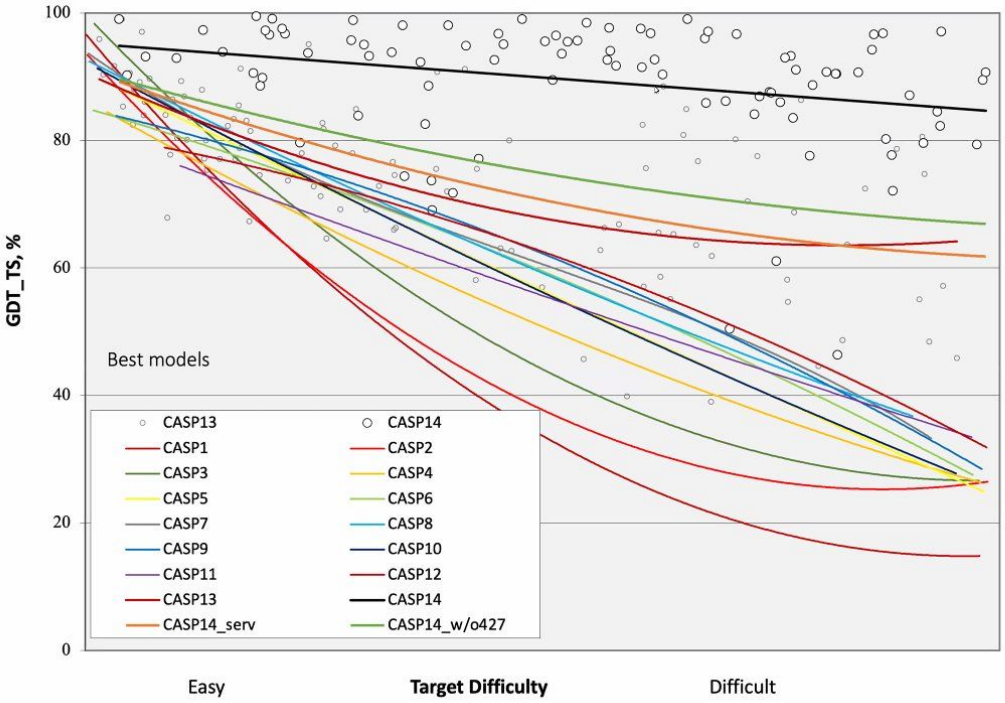
- The article (Jumper *et al*. 2021 Nature) and its 60+ pages of SI
- The [github code page]



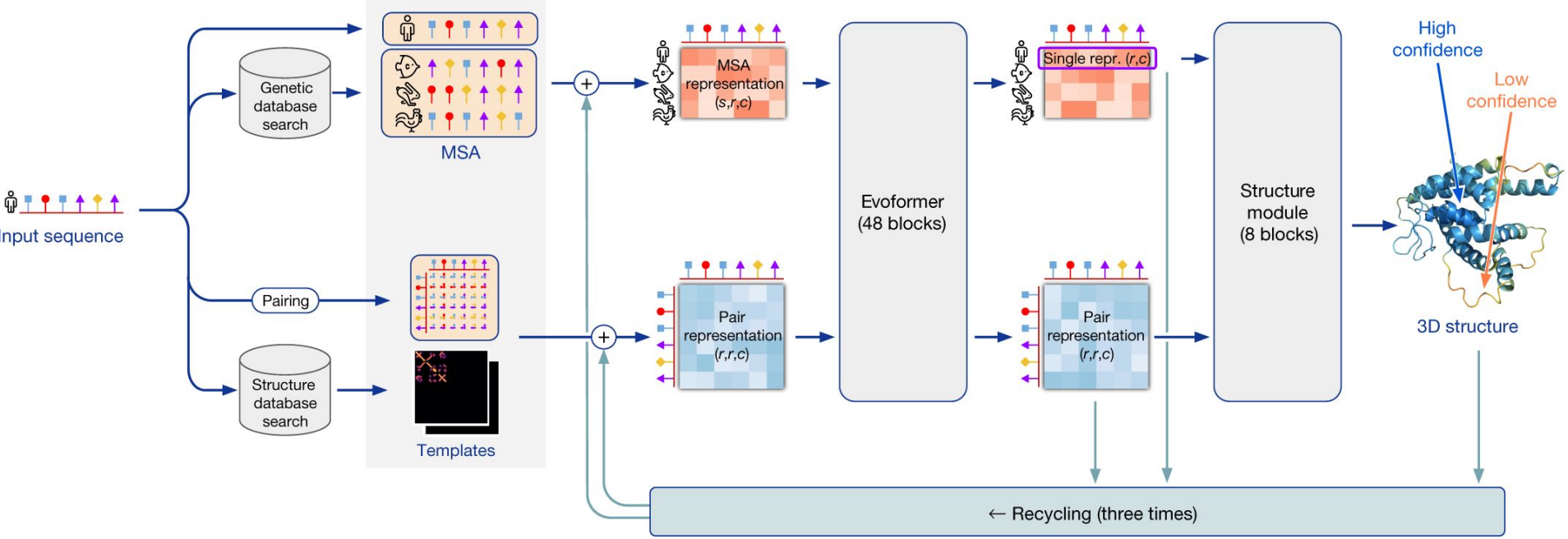
Pereira, *et al*. Proteins 2021

CASP 14


STRUCTURE SOLVER
DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.
Global distance test (GDT_TS; average)
100
90
80
70
60
50
40
30
20
10
0
A score above 90 is considered roughly equivalent to the experimentally determined structure
AlphaFold
AlphaFold 2
2006 2008 2010 2012 2014 2016 2018 2020
Contest year
©nature

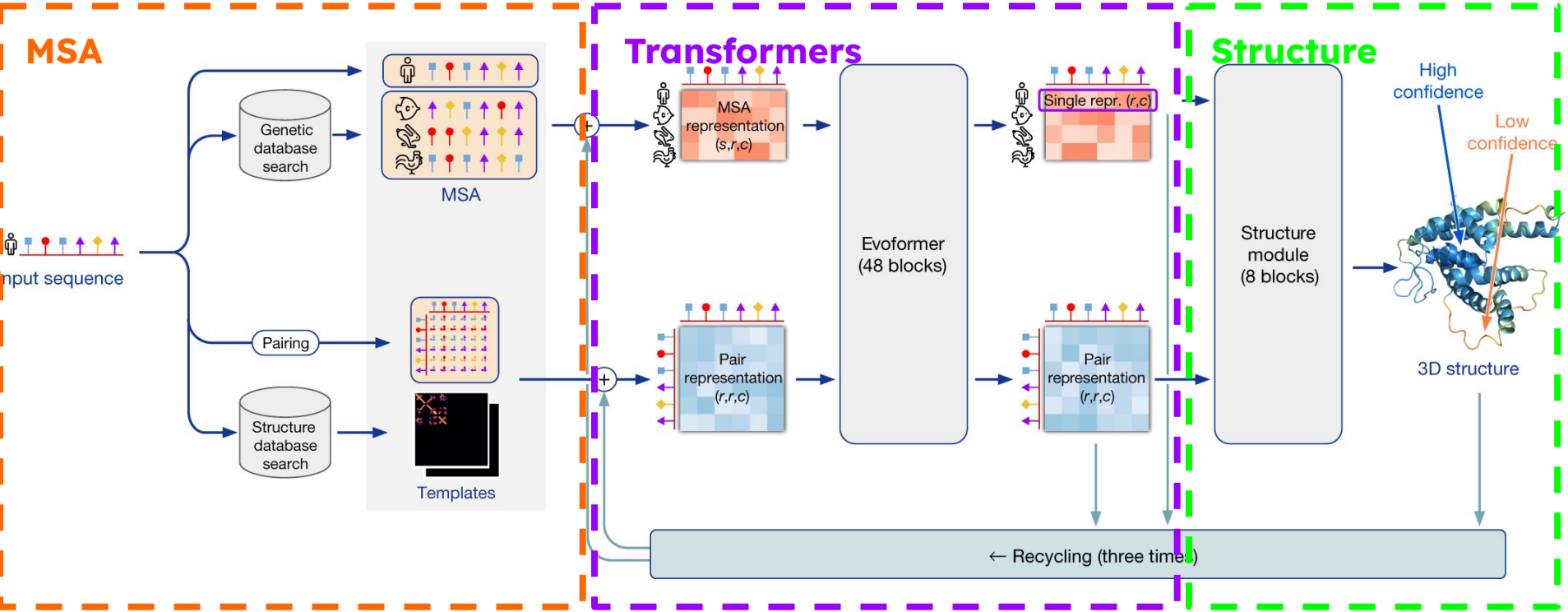

GDT_TS, %
100
80
60
40
20
0
Best models
CASP13
CASP14
CASP1
CASP2
CASP3
CASP4
CASP5
CASP6
CASP7
CASP8
CASP9
CASP10
CASP11
CASP12
CASP13
CASP14
CASP14_serv
CASP14_w/o427
Easy
Target Difficulty
Difficult

AlphaFold 2



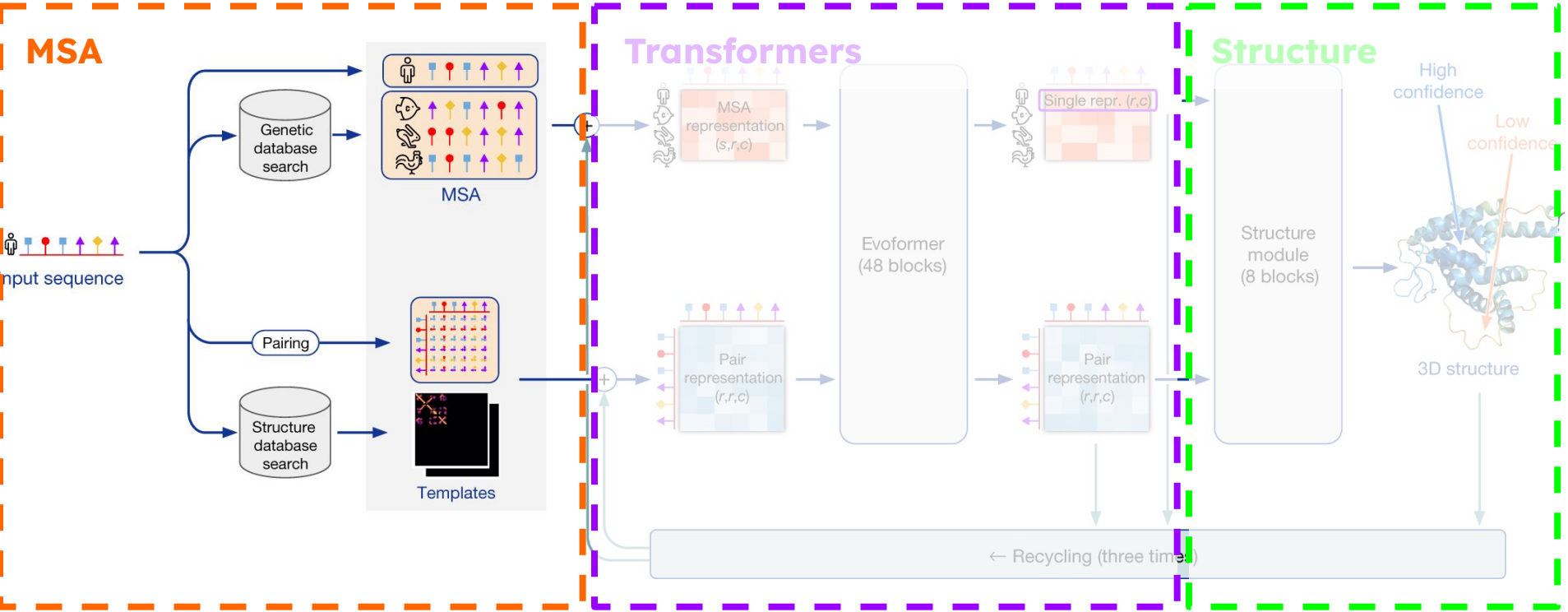
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. and Bridgland, A., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), pp.583-589.

AlphaFold 2 Architecture



Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. and Bridgland, A., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), pp.583-589.

AlphaFold 2 Architecture

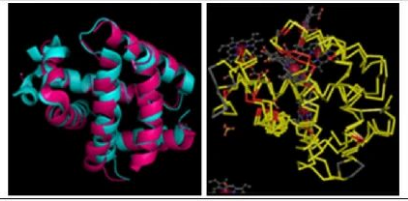
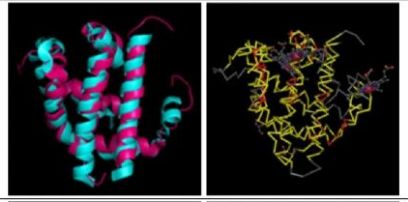
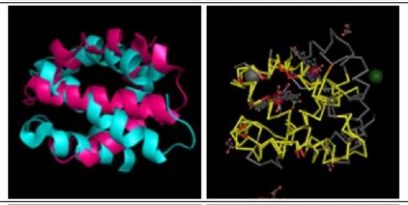
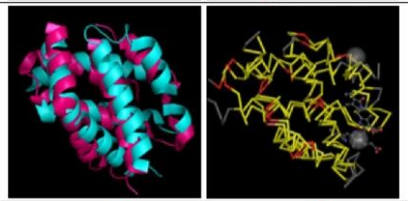
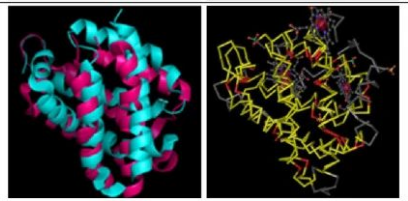


Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. and Bridgland, A., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), pp.583-589.

MSA Basic Principles

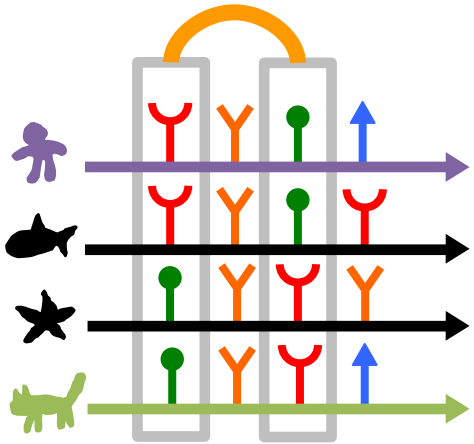
“protein structures are three to ten times more conserved than the amino acid sequence”

Sousounis, et al. Hum. Genomics 6, 10 (2012)

PDB # - Function	RMSD - % Identity	Images from PyMOL and Cn3D
Example 1 - 2GTL_Chain A: *Lumbricus terrestris* (annelide) hemoglobin part of a 3.6million Dalton protein. Transports oxygen. - 1H97_Chain A: *Paramphistomum epiclitum* (trematode) monomeric hemoglobin. High affinity to oxygen.	RMSD: 2.3 Identity: 12.1%	
Example 2 - 2GNW_Chain B: Found in plants. Its role is not yet determined. *Oryza sativa.* - 2W31_Chain A: detects oxygen and transmits signal. *Geobacter sulfurreducens.*	RMSD: 3.2 Identity: 13.4%	
Example 3 - 2GLN_Chain A: nitric oxide scavenging. *Mycobacterium tuberculosis.* - 2ZS1_Chain A: extracellular giant Hb. Cooperative oxygen binding via inorganic cations. *Oligobrachia mashikoi.*	RMSD: 2.4 Identity: 6.7%	
Example 4 - 1KN1_Chain A: allophycocyanin, absorbs light, part of phycobilisomes and phycobilisome structural family. *Pyropia yezoensis.* - 2BNL_Chain C: Non heme, regulates s factor after environmental stress. *Bacillus subtilis*	RMSD: 2.9 Identity:11.4%	
Example 5 - 2VEB_Chain A: Found in archae, role is not yet determined. *Methanosarcina acetivorans.* - 1OJ6_Chain A: A neuroglobin found in human brain. Binds to oxygen. *Homo sapiens.*	RMSD: 2.9 Identity: 12.7%	

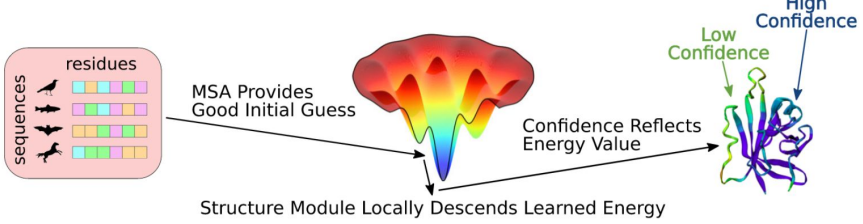
MSA Basic Principles

Contacts in proteins are evolutionarily conserved and encoded in a **MSA** (Multiple Sequence Alignment) due to **coevolution**

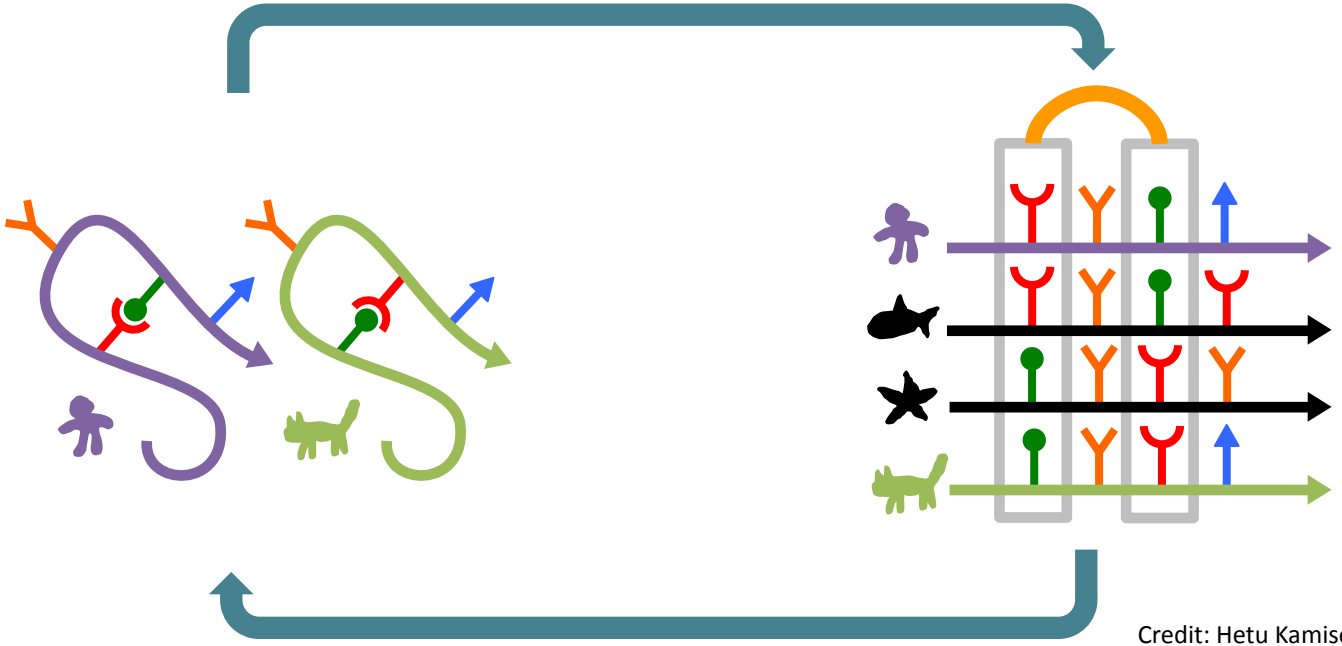


Credit: Hetu Kamisetty, Sergei Ovchinnikov

MSA Basic Principles



Contacts in proteins are evolutionarily conserved and encoded in a **MSA** (Multiple Sequence Alignment) due to **coevolution**



Credit: Hetu Kamisetty, Sergei Ovchinnikov

Multiple Sequence Alignment

The principle is not new (90s) ! Similar approach for all CASP14 participant.

Query the AA sequence in several DB:

- BFD,
- MGnify,
- PDB70,
- PDB (structures in the mmCIF format),
- PDB seqres – only for AlphaFold-Multimer,
- Uniclust30,
- UniProt – only for AlphaFold-Multimer,
- UniRef90.

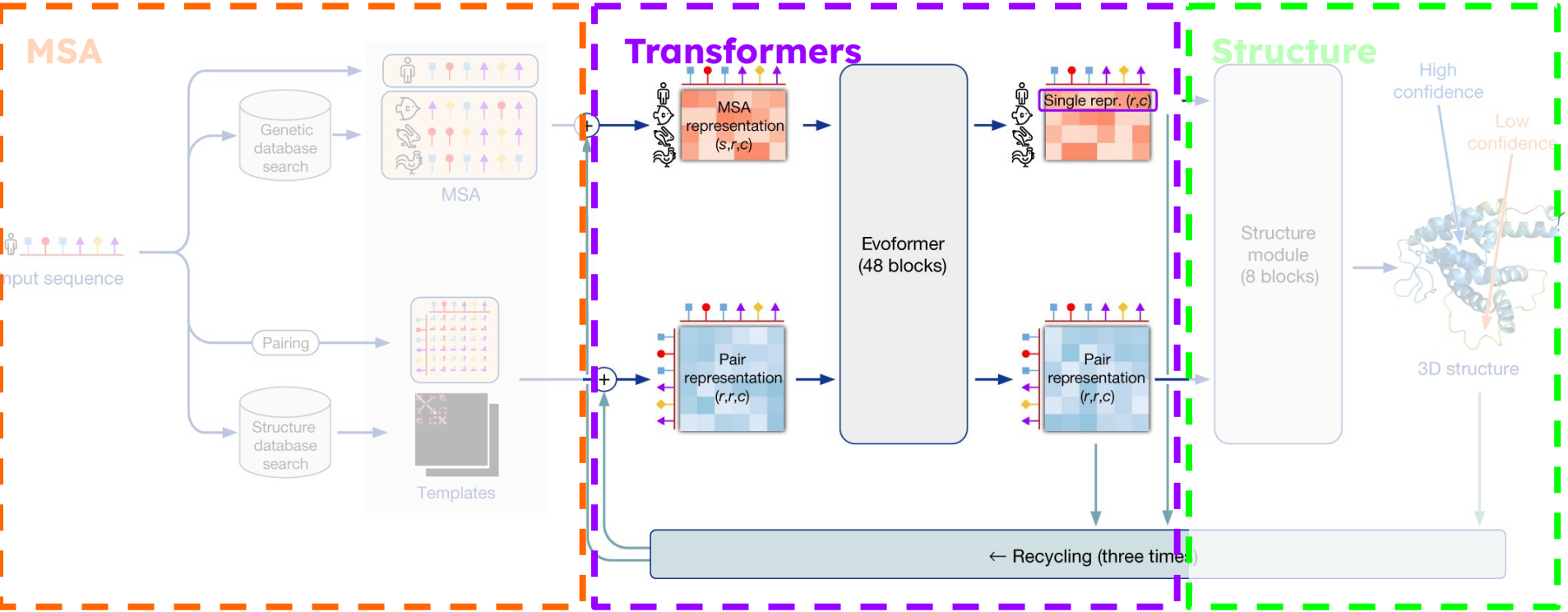
! Big size ~ 2.2 TB

MSA : $N_{seq} \times N_{res}$ array (N_{seq}, number of sequences; N_{res}, number of residues)

Templates

- Similar structure to the input sequence are scan in the PDB
- Gives initial representation of the structure or “pair representation”

AlphaFold 2 Architecture



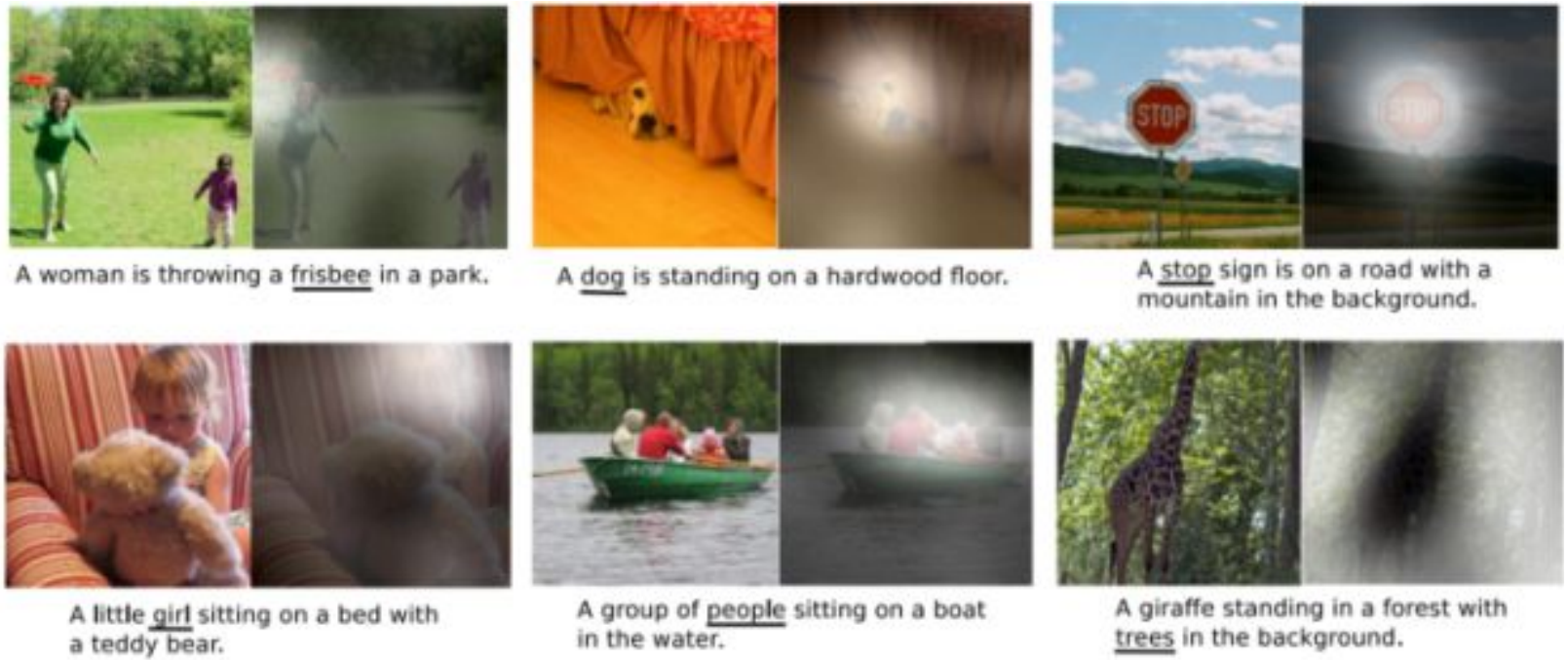
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. and Bridgland, A., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), pp.583-589.

Evoformer, an evolutionary transformer?

- Transformer Wikipedia definition:
  - A **transformer** is a deep learning model that adopts the mechanism of self-attention, **differentially weighting** the significance of each part of the **input data**.
- Two transformers run parallely and communicating together:
  - A MSA transformer
  - A Structure transformer (pair representation)

Transformers

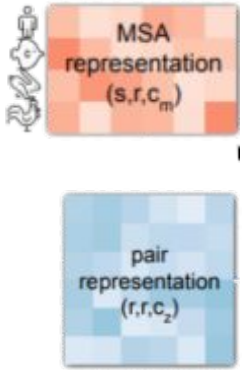
- Intro
(Atte
2017
- “The
whic
impo
netw
- But:



Kelvin Xu et al. (2016)

Evoformer, an evolutionary transformer?

- Transformer Wikipedia definition:
  - A **transformer** is a deep learning model that adopts the mechanism of self-attention, **differentially weighting** the significance of each part of the **input data**.
- Two transformers run parallely and communicating together:
  - A MSA transformer
  - A Structure transformer (pair representation), encode information about the **relation between the residues**



- Pair representation is both the **product** and an **intermediate layer** (new).

- At every cycle (48), hypothesis based on MSA and PR are tested to improve MSA and PR.

- Both representation MSA and PR exchange until network reach solid inference.

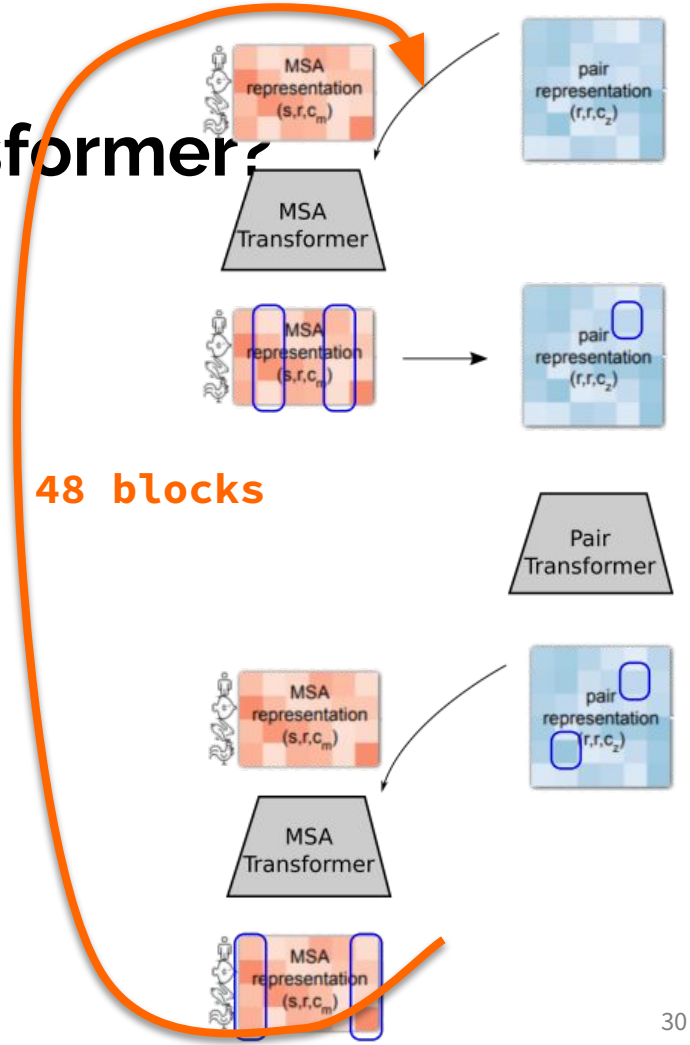
Evoformer, an evolutionary transformer?

- Transformer Wikipedia definition:
  - A **transformer** is a deep learning model that adopts the mechanism of self-attention, **differentially weighting** the significance of each part of the **input data**.
- Two transformers run parallely and communicating together:
  - A MSA transformer
  - A Structure transformer (pair representation)

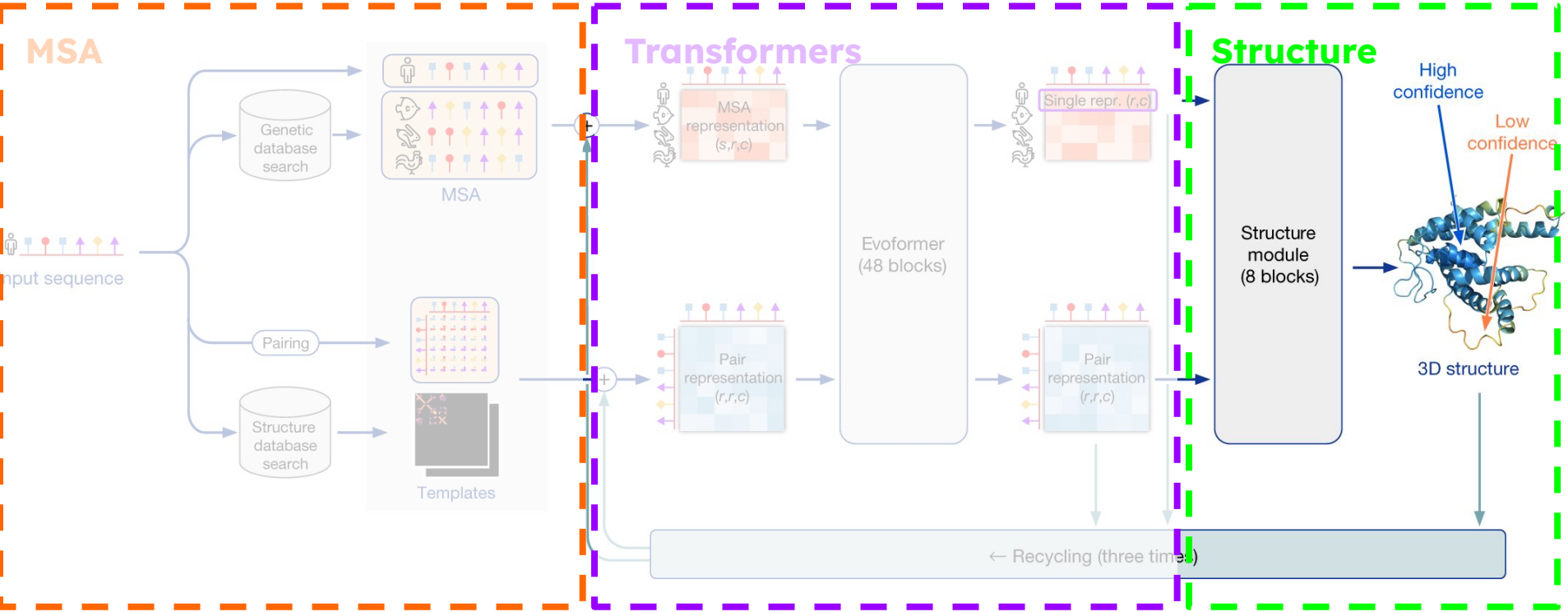
- Pair representation is both the **product** and an **intermediate layer** (new).

- At every cycle (48), hypothesis based on MSA and PR are tested to improve MSA and PR.

- Both representation MSA and PR exchange until network reach solid inference.



AlphaFold 2 Architecture



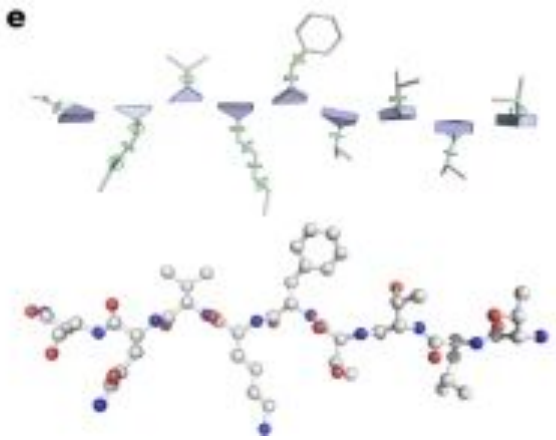
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. and Bridgland, A., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), pp.583-589.

The structure Module

After 48 iteration AF has build:

- MSA ~ sequence variation
- “pairs representation”, ~ residues interaction map

Need to translate to a 3D protein structure



Structure Module considers the protein as a “residue gas”

Each aa is modelled as a triangle, representing the three atoms of the backbone (C_α, N, C)

The structure Module

- The 3D backbone structure is represented as $\mathbf{N}_{\mathbf{res}}$ **independent rotations and translations**
- Initialized in a trivial state:
  - rotations set to identity
  - positions set to origin
- Breaking the chain structure to allow simultaneous **local refinement** of all parts of the structure
- AlphaFold 2's attention mechanism is much simpler than the equivariant transformer that underlies RoseTTAFold
- Eventually model side chains.

Recycle

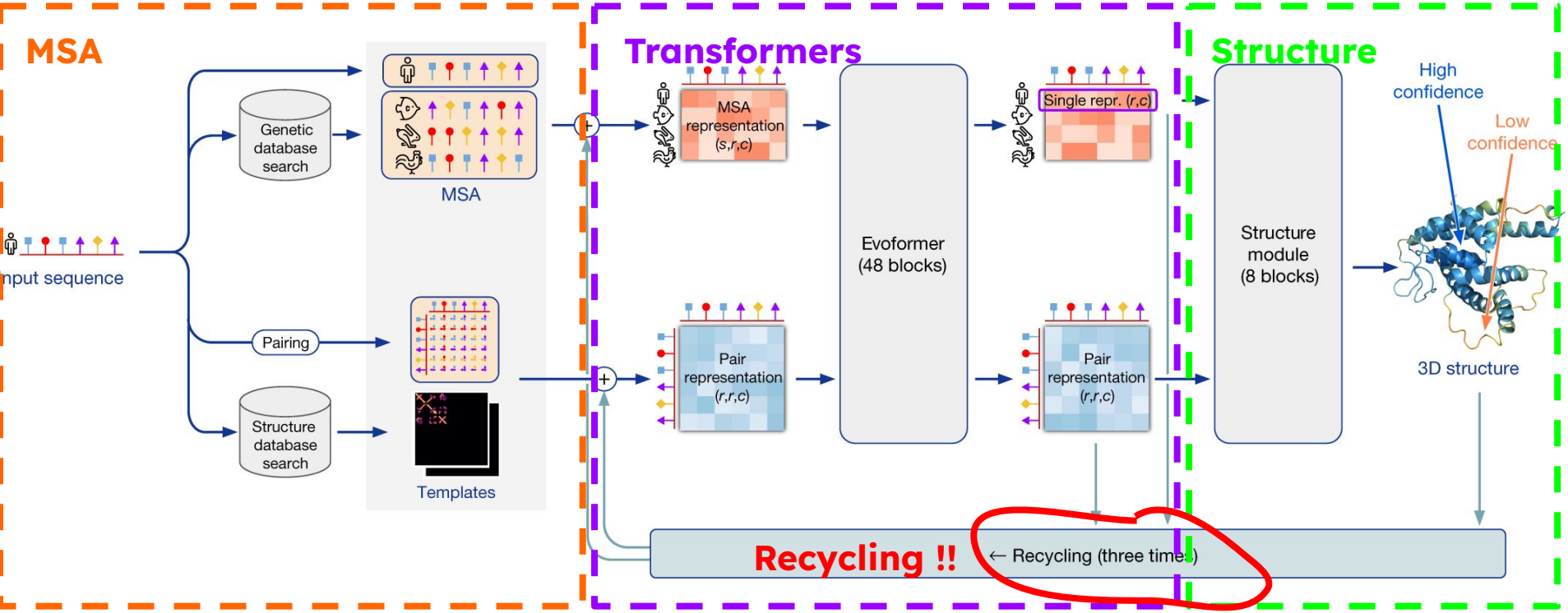
$$\mathbf{M} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Image taken from BrainVoyager.

The structure Module

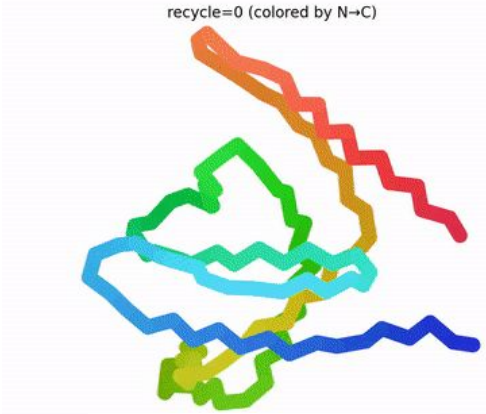
*"Conversely, the peptide bond geometry is **completely unconstrained** and the network is observed to frequently violate the chain constraint during the application of the structure module as breaking this constraint enables the local refinement of all parts of the chain without solving complex loop closure problems. Satisfaction of the **peptide bond geometry** is encouraged during fine-tuning **by a violation loss term**. Exact enforcement of peptide bond geometry is only achieved in the **post-prediction relaxation of the structure by gradient descent** in the Amber force field. Empirically, this final **relaxation does not improve** the accuracy of the model as measured by the global distance test (GDT) or lDDT-Cα but does remove distracting stereochemical violations without the loss of accuracy."*

AlphaFold 2 Architecture

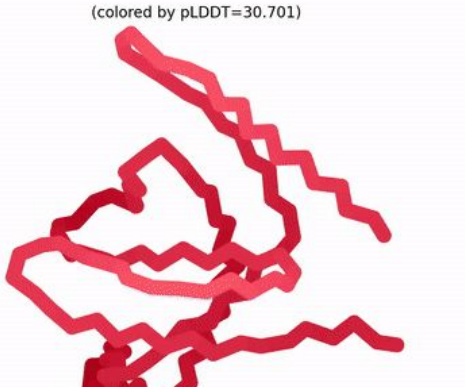


Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. and Bridgland, A., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), pp.583-589.

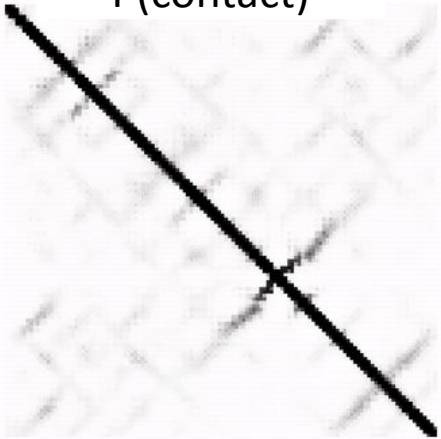
Recycles



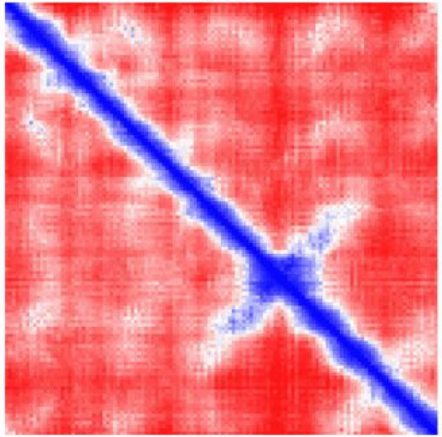
Model Confidence
predicted LDDT



P(contact)



predicted Aligned Error

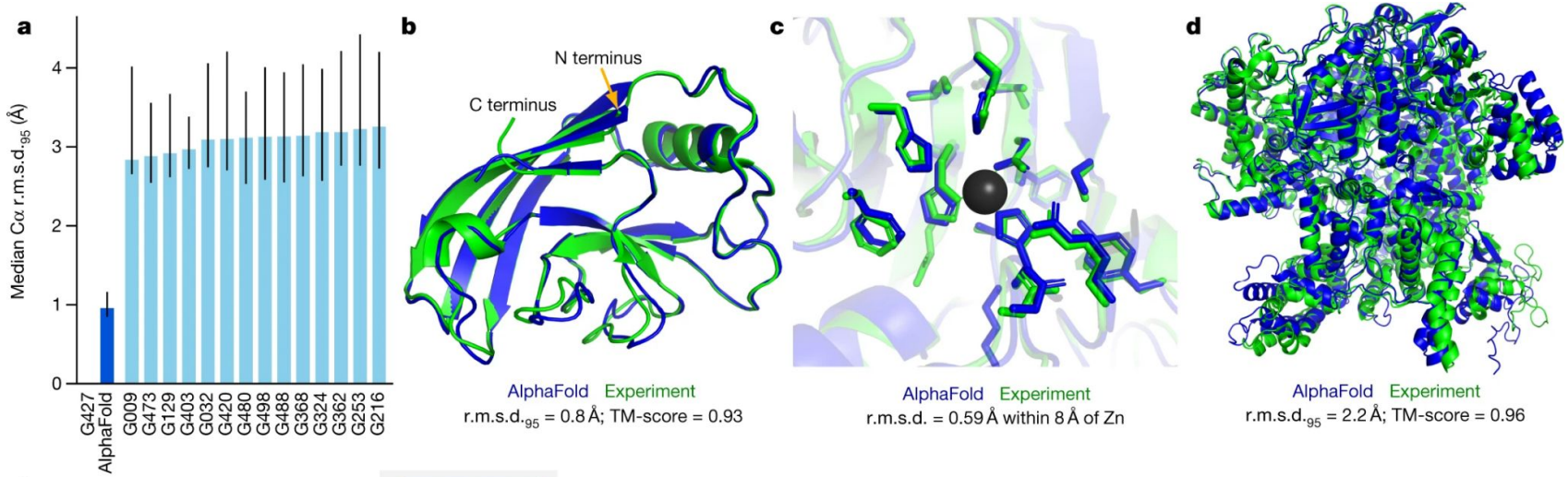


Slide courtesy of Sergey Ovchinnikov

AlphaFold 2 Results

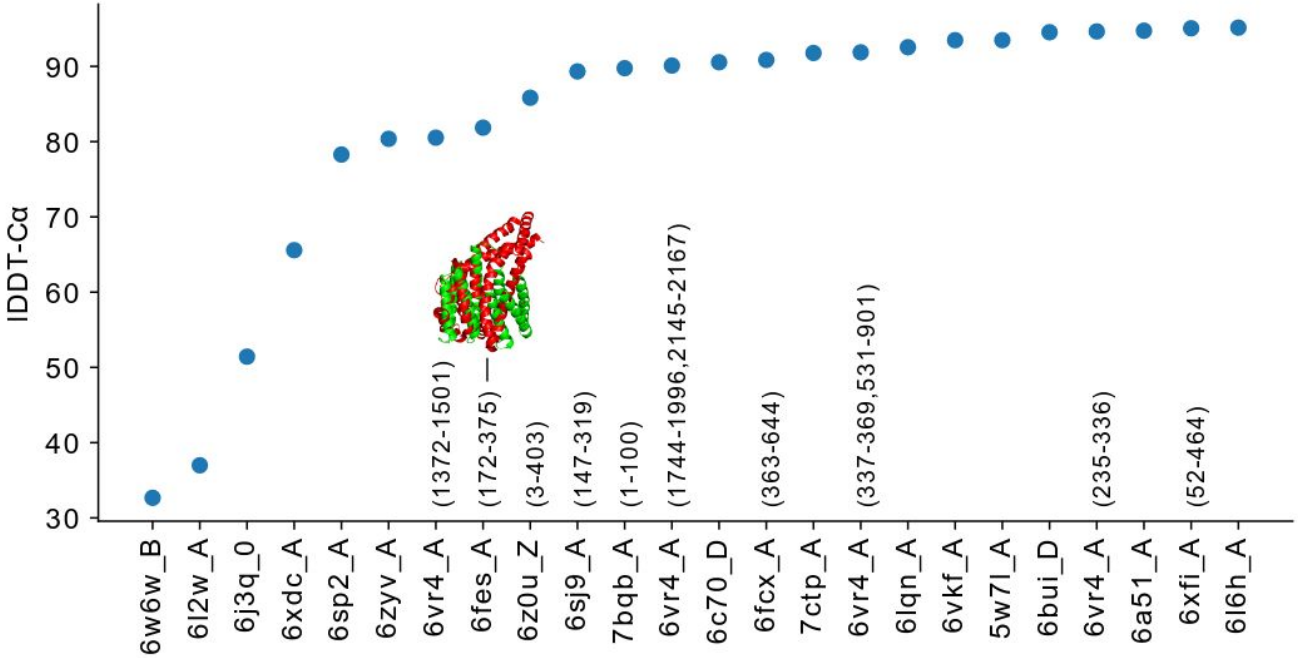
CASP14:

- Median backbone accuracy of 0.96 Å Cα $RMSD_{95}$ *vs.* next best performing accuracy of 2.8 Å Cα $RMSD_{95}$(Baker Lab)
- High accurate side chains: All-atom accuracy of AlphaFold 1.5 Å $RMSD_{95}$ (vs. 3.5 Å)



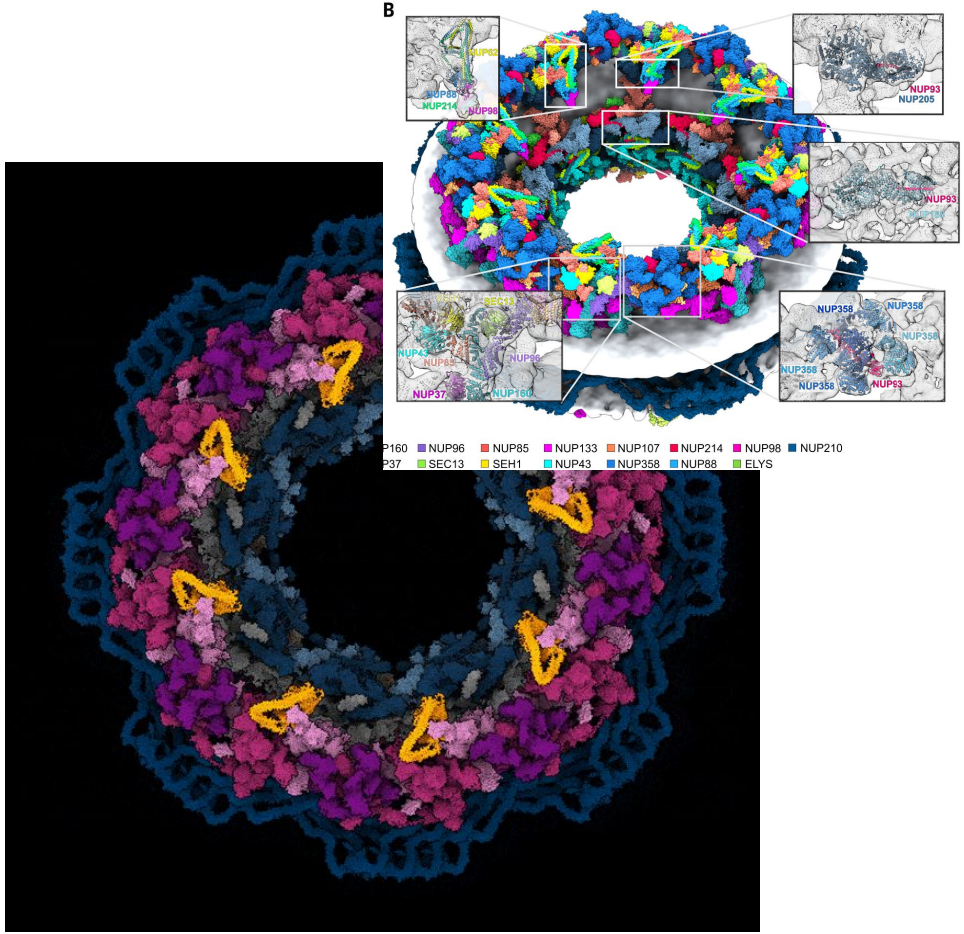
Just memorizing PDB ?

Results with not learned structures :



Nuclear Pore Structure

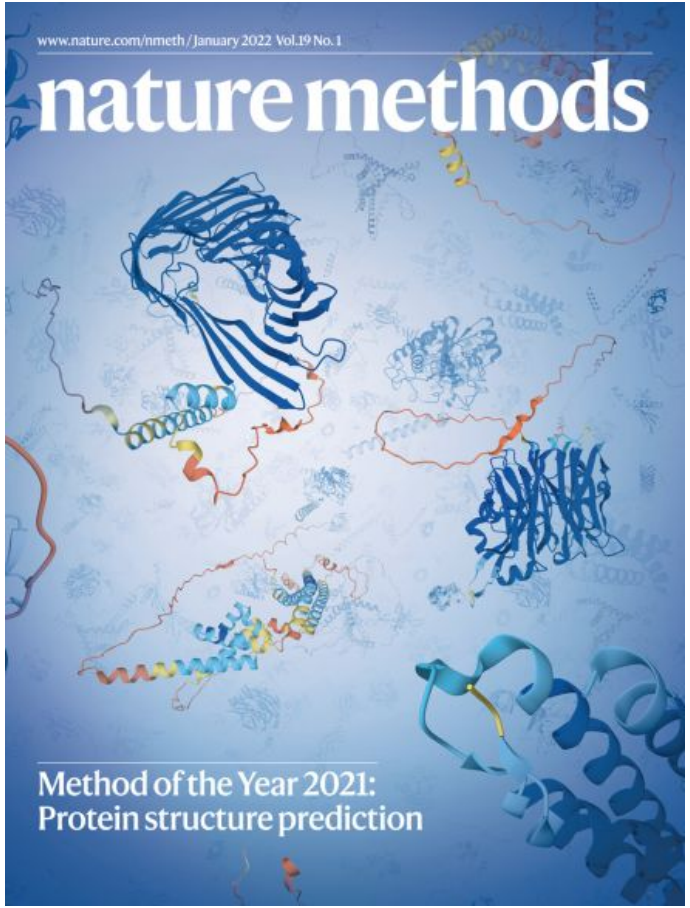
- +500 proteins
- 120 nm
- Alphafold2 allow passing from 30 to 90 % of structural covering of EM map.



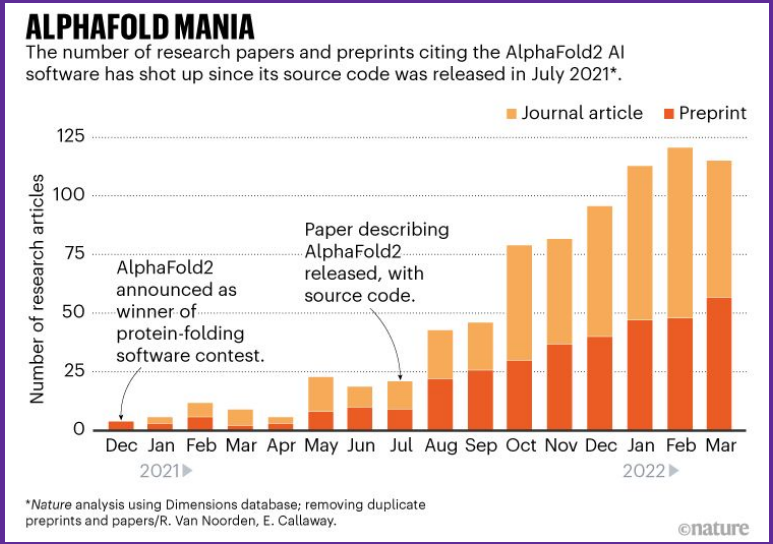
Mosalaganti, S. *et al.* Science 2022

Alphafold 2.0 limits

- Mostly **One single structural states** of a protein
- Large-flexible proteins
  - Individual domains OK, But **arrangement is random**
- TM-proteins
  - AF2 does not know what a membrane is
- DNA-binding protein complexes
- **No ability to train AlphaFold**
- Not integrating experimental data
- No protein dynamics
- No post-translational modifications

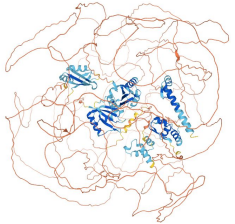


An ecosystem around Alphafold

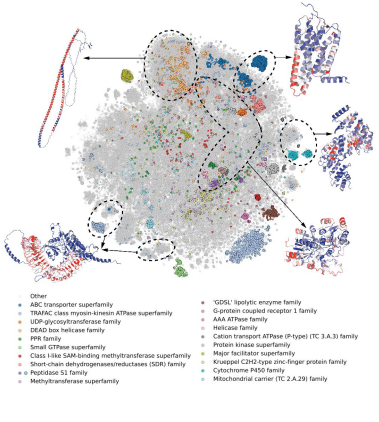


Uses of AlphaFold

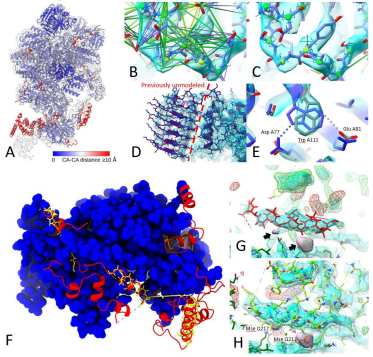
Predicting disorder



Understanding structural space

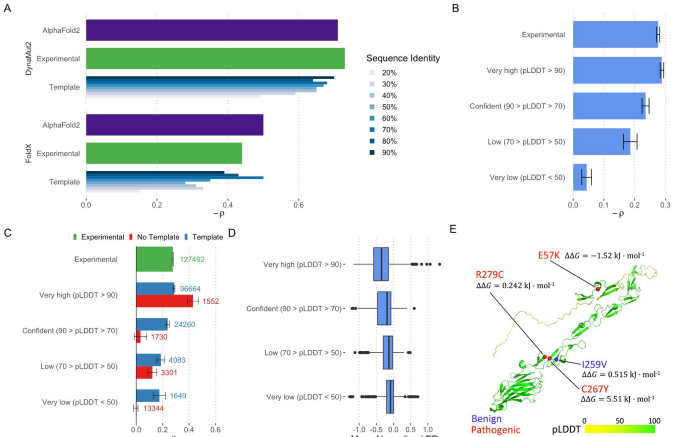


Solving structures

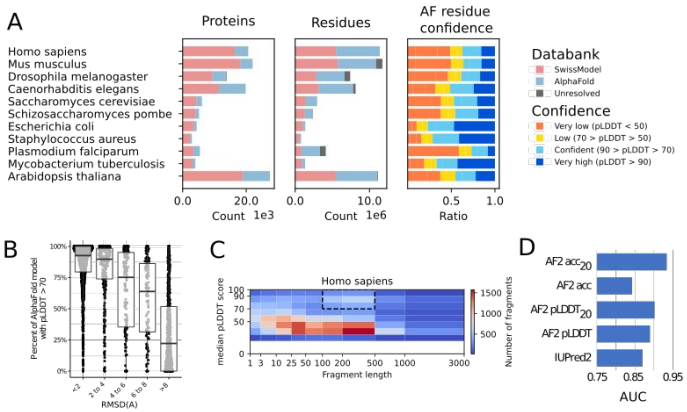


RoseTTTAFold and DMPfold2 add value

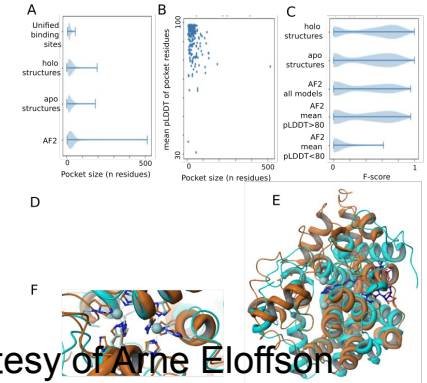
Predicting variants



Increased Structural coverage



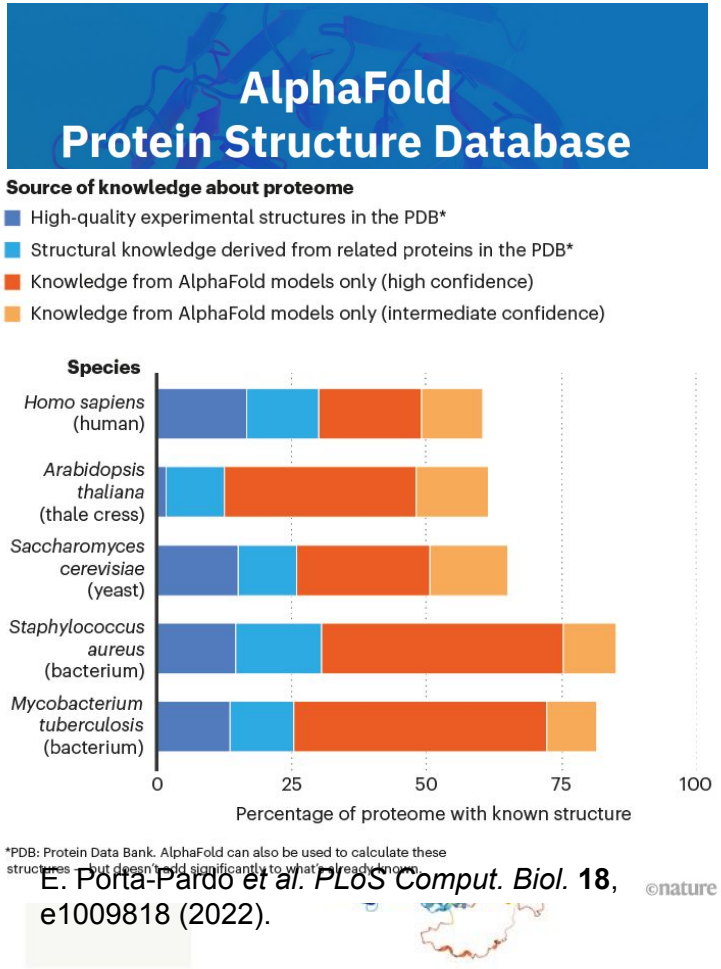
Drug design



Slide courtesy of Arne Elofsson

EMBL-EBI's Alphafold DB

- + 21 model organism proteomes
- UniProt
- +200 million protein structure predictions
- Only monomers !



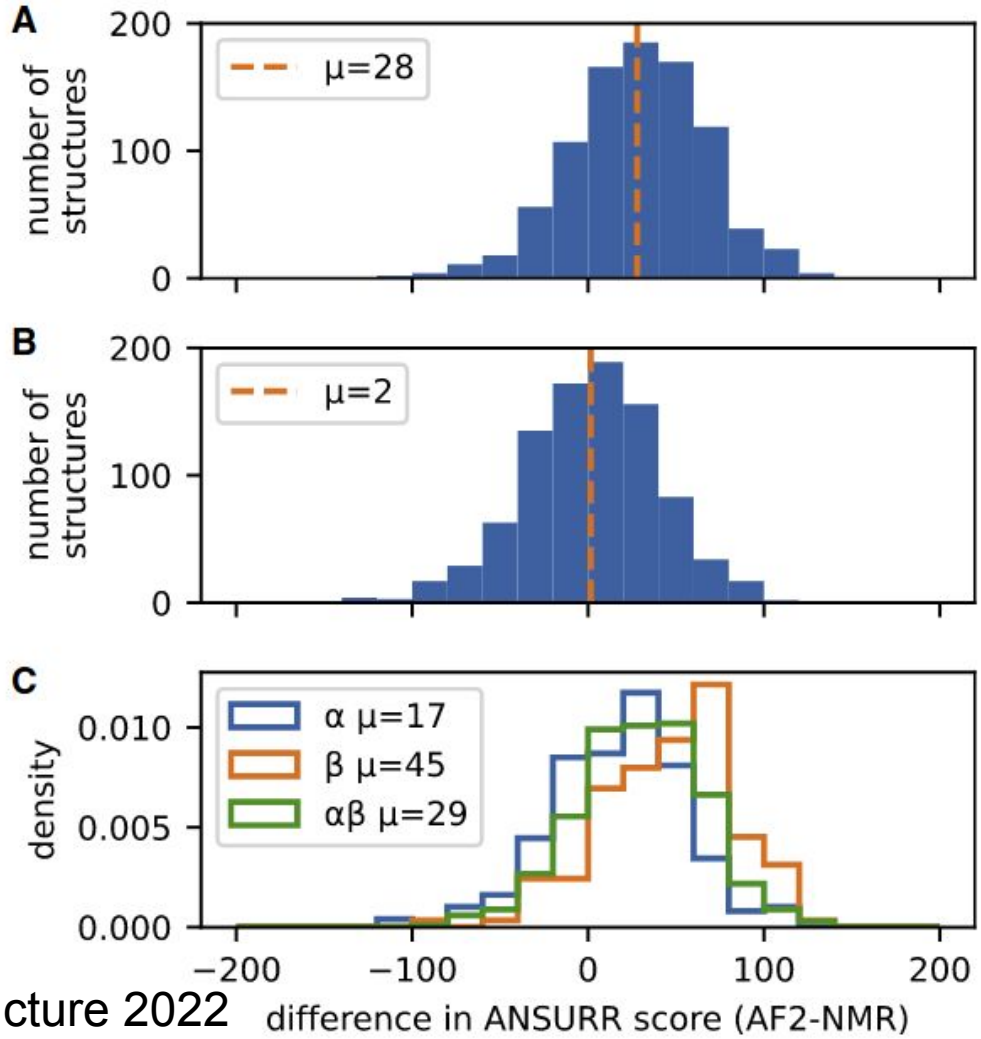
E. Porta-Pardo *et al. PLoS Comput. Biol.* **18**, e1009818 (2022).

Varadi et al. NAR 2022

Alphafold *vs.* NMR

Highlights

- 904 human proteins with both Alpha-Fold and NMR structures
- Alpha-Fold predictions are usually more accurate than NMR structures
- NMR can be better than Alpha-Fold where there are local dynamics
- NMR is useful to validate Alpha-Fold predictions and refine where necessary

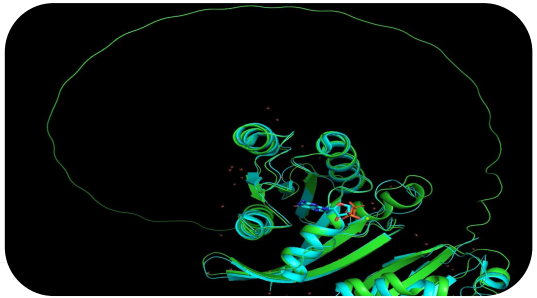


Fowler et al. Structure 2022

Can predict protein-protein/peptide interactions

Yoshitaka Moriwaki @Ag_smith · Jul 19

AlphaFold2 can also predict heterocomplexes. All you have to do is input the two sequences you want to predict and connect them with a long linker.



G-linker!

Hiroki Onoda
@onoda_hiroki

Unknown linker may be useful for multimer prediction on the local Alphafold2!!

UNK-linker!

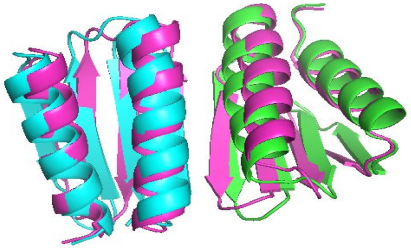
Slide courtesy of Sergey Ovchinnikov

Minkyung Baek
@minkbaek

Don't actually need a G-linker!

Adding a big enough number for "residue_index" feature is enough to model hetero-complex using AlphaFold (green&cyan: crystal structure / magenta: predicted model w/ residue_index modification).
#AlphaFold #alphafold2

```
# add big enough number to residue index to indicate chain breaks
idx_res = feature_dict['residue_index']
L_prev = 0
# Ls: number of residues in each chain
for L_i in Ls[:-1]:
    idx_res[L_prev+L_i:] += 200
    L_prev += L_i
feature_dict['residue_index'] = idx_res
```

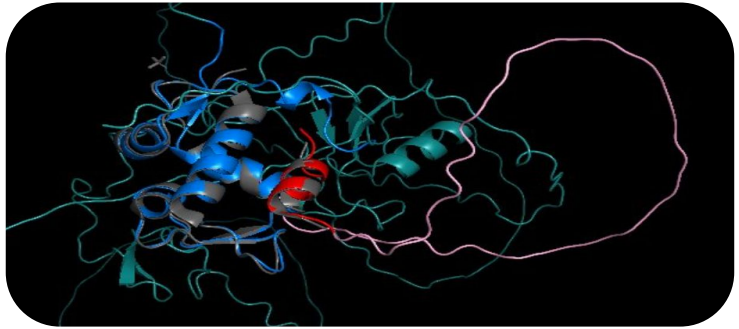


大上雅史 | Ohue M 2.5G
@tonets

あ、AlphaFold2でペプチドドッキングでき

Translated from Japanese by Google

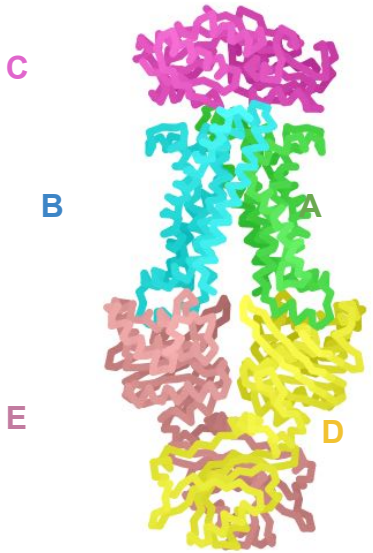
Oh, I was able to dock the peptide with AlphaFold2

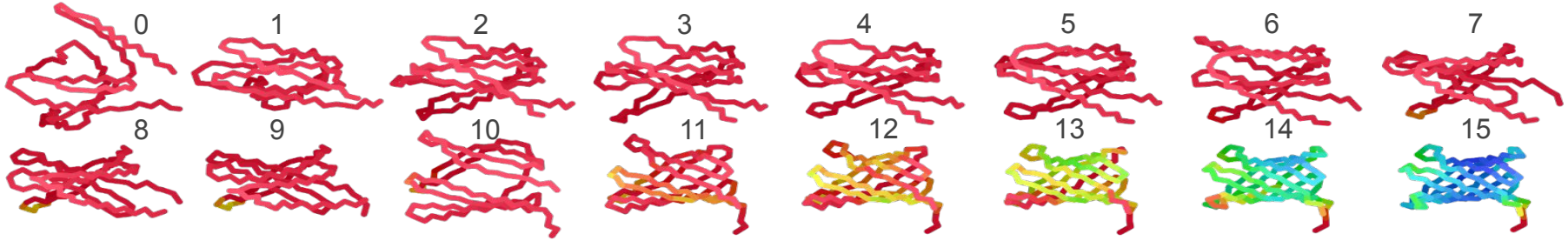


Protein-peptide interaction

ColabFold - Advanced options

- **github.com/sokrypton/ColabFold**
- Modify MSA input
  - Custom or MMseqs2 (much faster)
  - Trim
- **Complexes**
  - **Homo-oligomers**
  - **Hetero-oligomers**
- Fine control
  - Number of recycles
- Sample (Output more than 5 models)
  - Generate ensembles by iterating through random seeds, enabling dropout.





Slide courtesy of Sergey Ovchinnikov

ColabFold

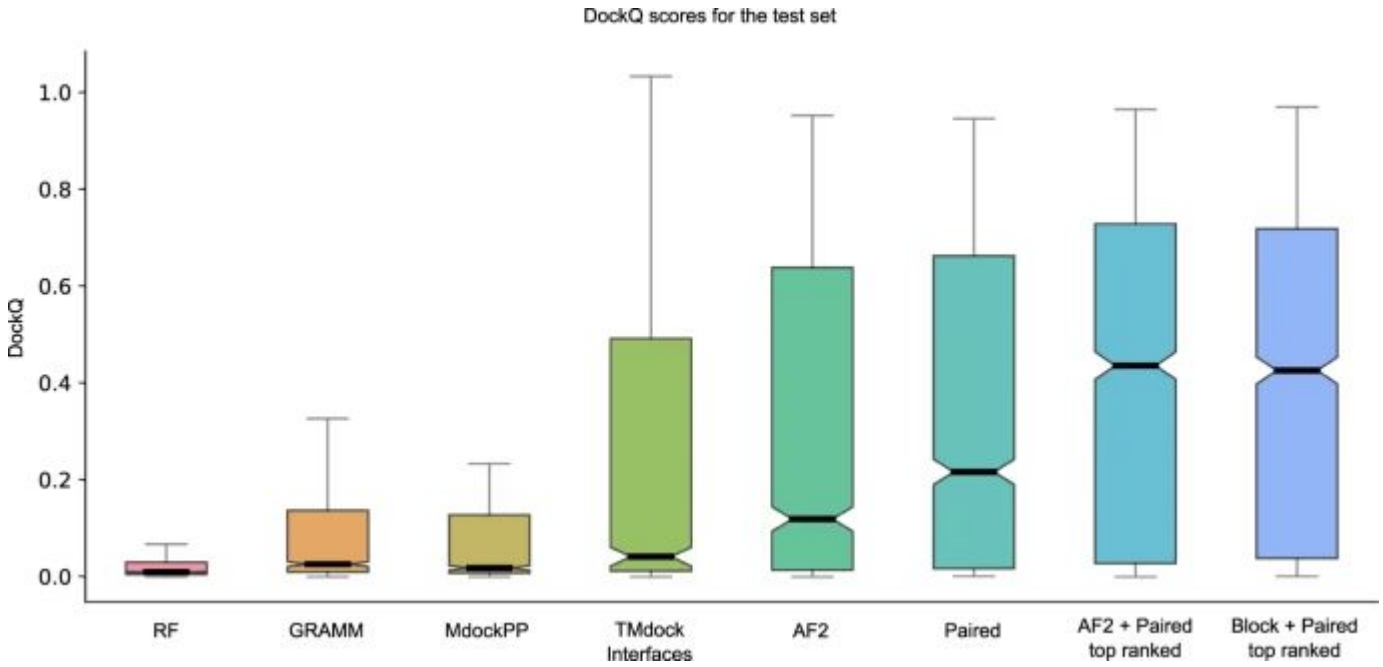
Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S and Steinegger M. ColabFold: Making protein folding accessible to all. Nature Methods (2022) doi: 10.1038/s41592-022-01488-1

https://github.com/sokrypton/ColabFold

Making Protein folding accessible to all via Google Colab!

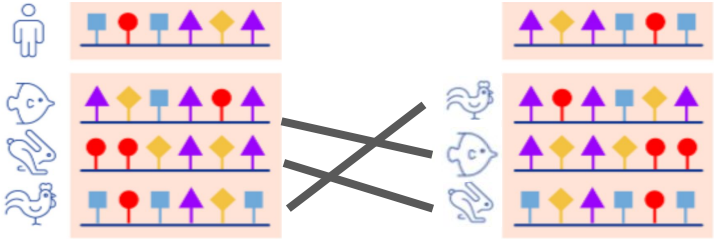
Notebooks	monomers	complexes	mmseqs2	jackhmmer	templates
AlphaFold2_mmseqs2	Yes	Yes	Yes	No	Yes
AlphaFold2_batch	Yes	Yes	Yes	No	Yes
RoseTTAFold	Yes	No	Yes	No	No
AlphaFold2 (from Deepmind)	Yes	Yes	No	Yes	No
ESMFold	Yes	Maybe	No	No	No
BETA (in development) notebooks					
AlphaFold2_advanced	Yes	Yes	Yes	Yes	No
OmegaFold	Yes	Maybe	No	No	No

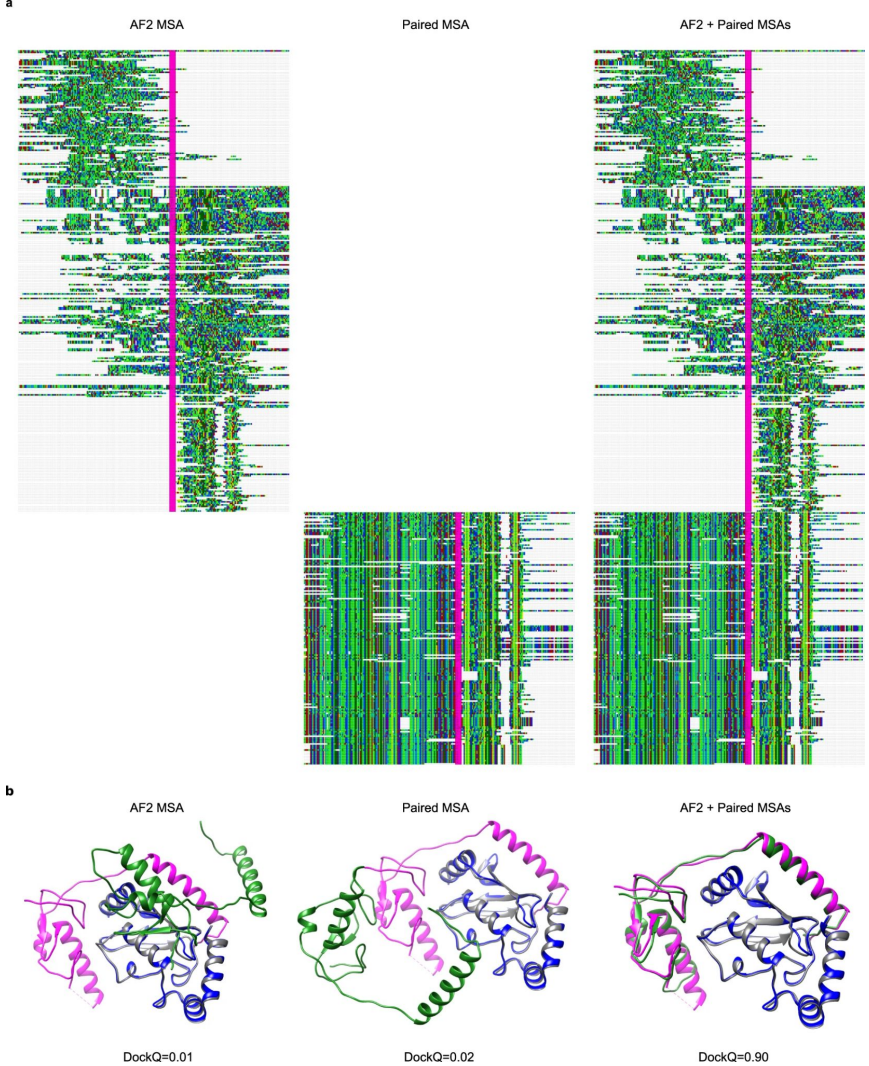
Protein-Protein Docking



Bryant et al. Nature Comm. 2022

Protein-Protein Docking





Bryant et al. Nature Comm. 2022

Alphafold-Multimer

- New weights
- Force sampling
- High accuracy
- ipTM

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

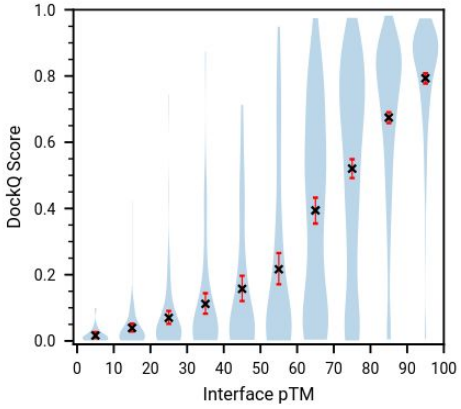
Follow this preprint

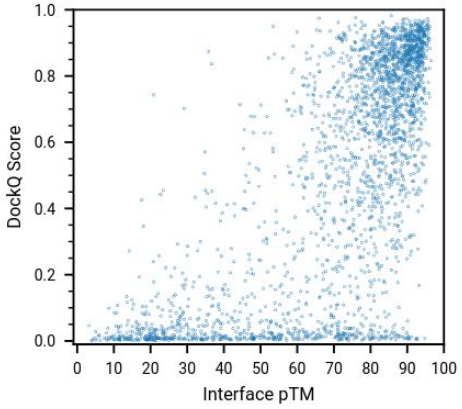
Protein complex prediction with AlphaFold-Multimer

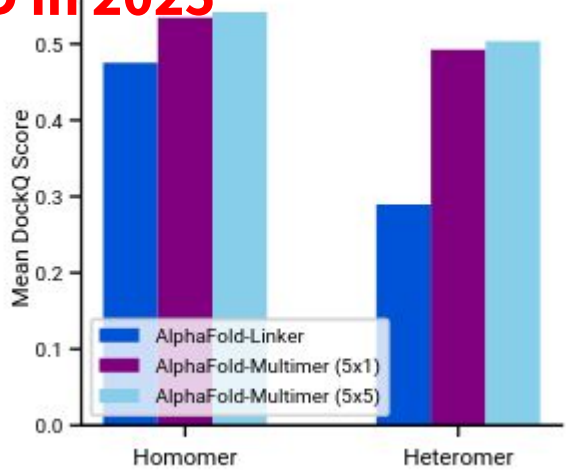
Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, Demis Hassabis

doi: https://doi.org/10.1101/2021.10.04.463034

STILL NOT REVIEWED in 2025



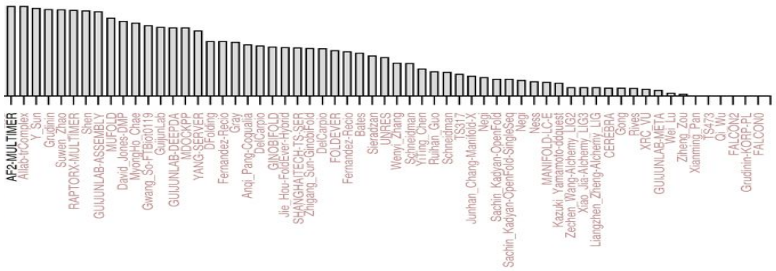




Going beyond Alphafold performances

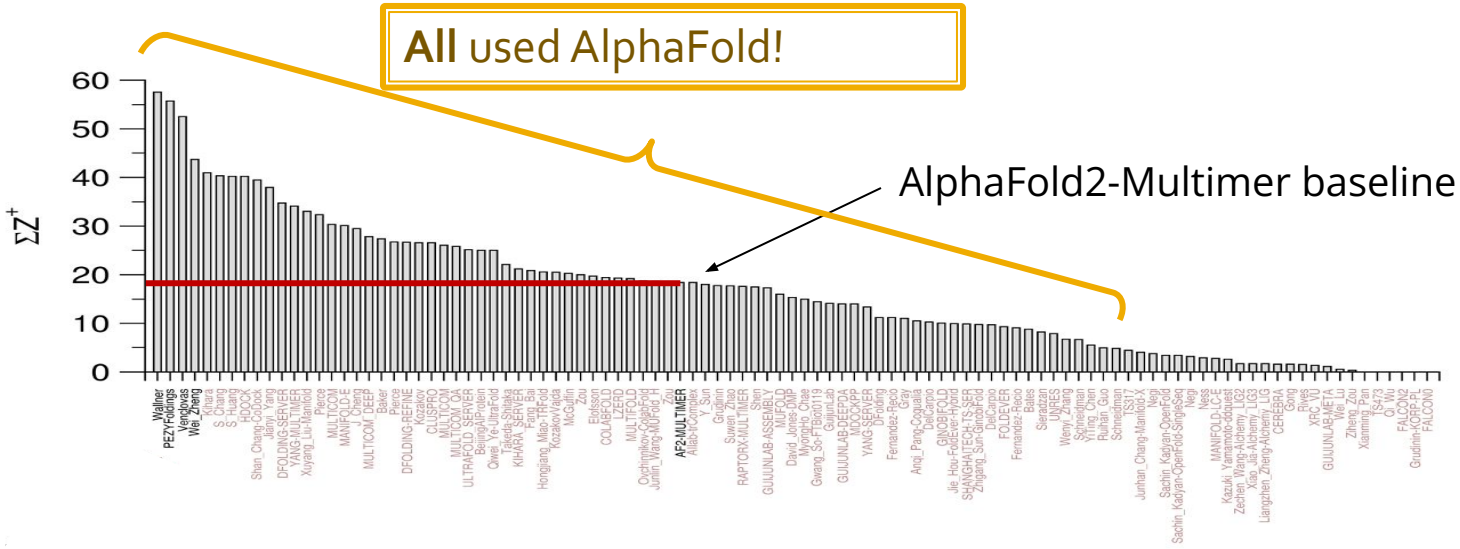
CASP15/CAPRI Assembly prediction

AlphaFold2-Multimer baseline



Slide Courtesy: Marc F. Lensink

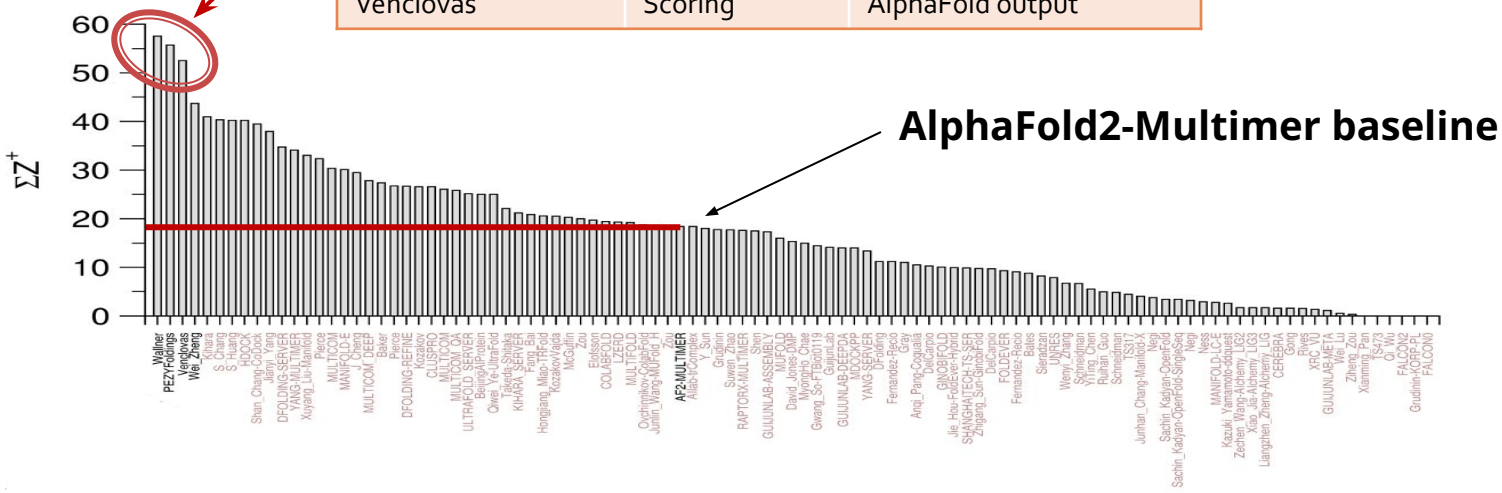
CASP15/CAPRI Assembly prediction



Slide Courtesy: Marc F. Lensink

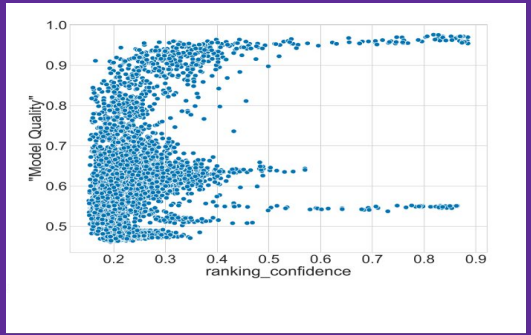
CASP15/CAPRI Assembly prediction

Wallner	Sampling	AlphaFold
PEZYFoldings	Pruning	AlphaFold input
Venclovas	Scoring	AlphaFold output



Slide Courtesy: Marc F. Lensink

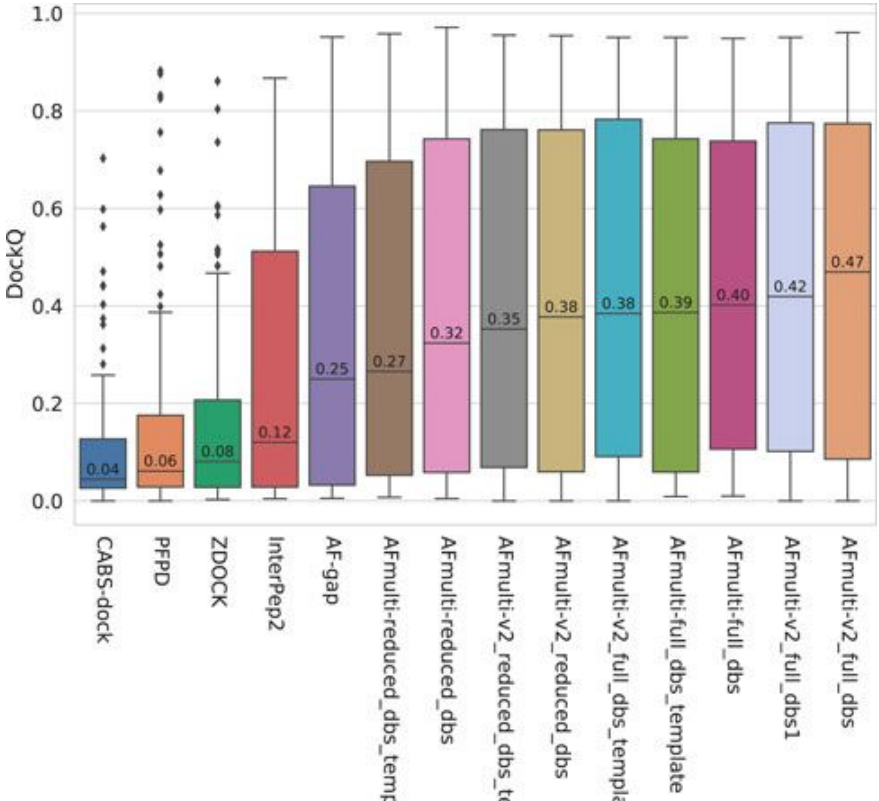
Sampling


1.0
0.9
0.8
0.7
0.6
0.5
"Model Quality"
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
ranking_confidence

Protein-Peptide Docking

- acceptable or better quality (DockQ ≥0.23) for 66 of the 112 complexes
- 25 of which were high quality (DockQ ≥0.8).
- predict whether a peptide and a protein will interact.

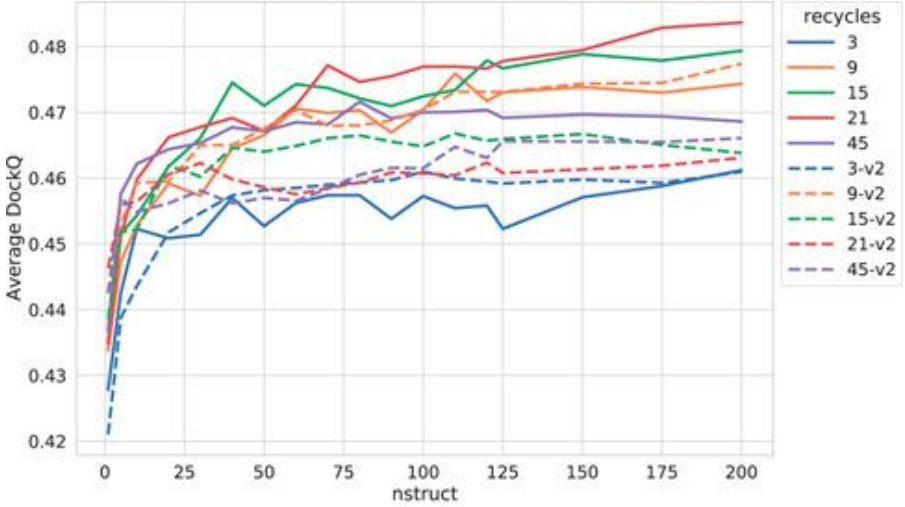
DockQ	Model quality
≥ 0.23	Acceptable
≥ 0.50	Medium
≥ 0.80	High

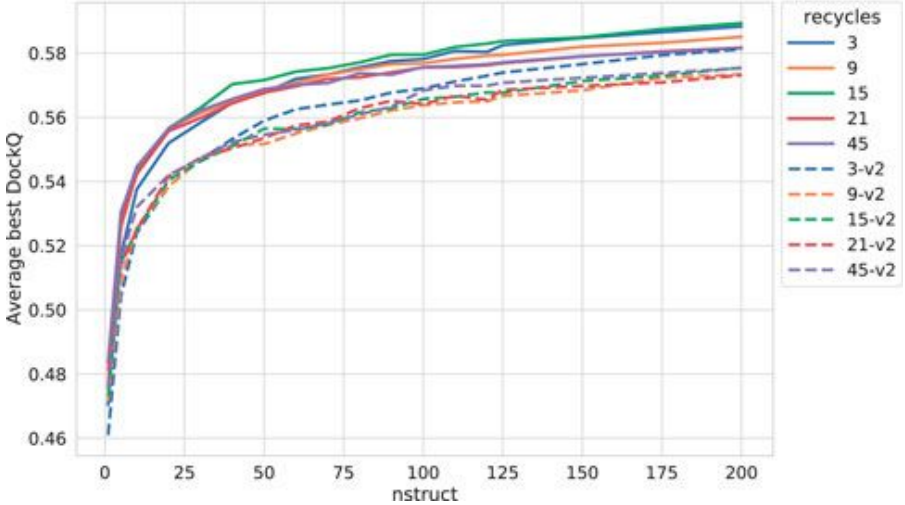


Åkhe and Wallner, Front. Bioinform. 2022

Protein-Peptide Docking

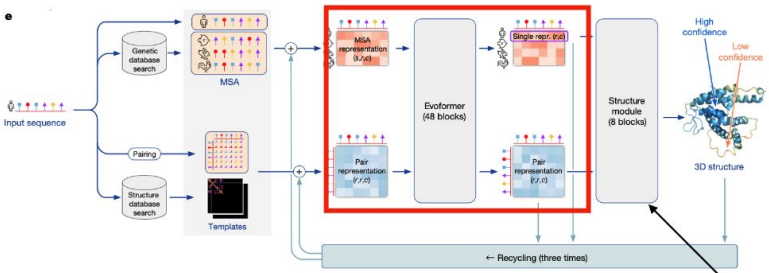
- Forced Sampling:
  - increases the number of acceptable models from 66 to 75
  - improves the median DockQ from 0.47 to 0.55 (17%)
  - best possible DockQ improves from 0.58 to 0.72 (24%)



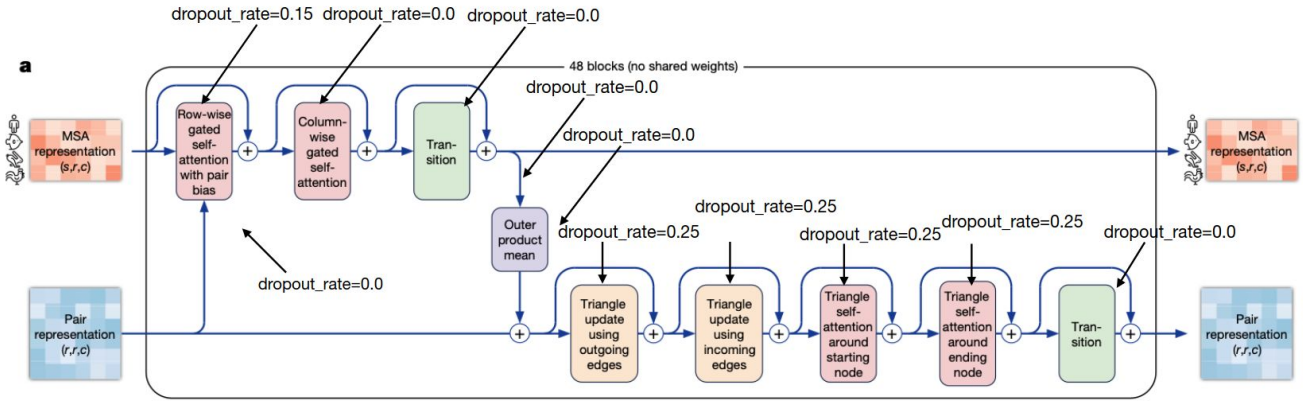


Åkhe and Wallner, Front. Bioinform. 2022

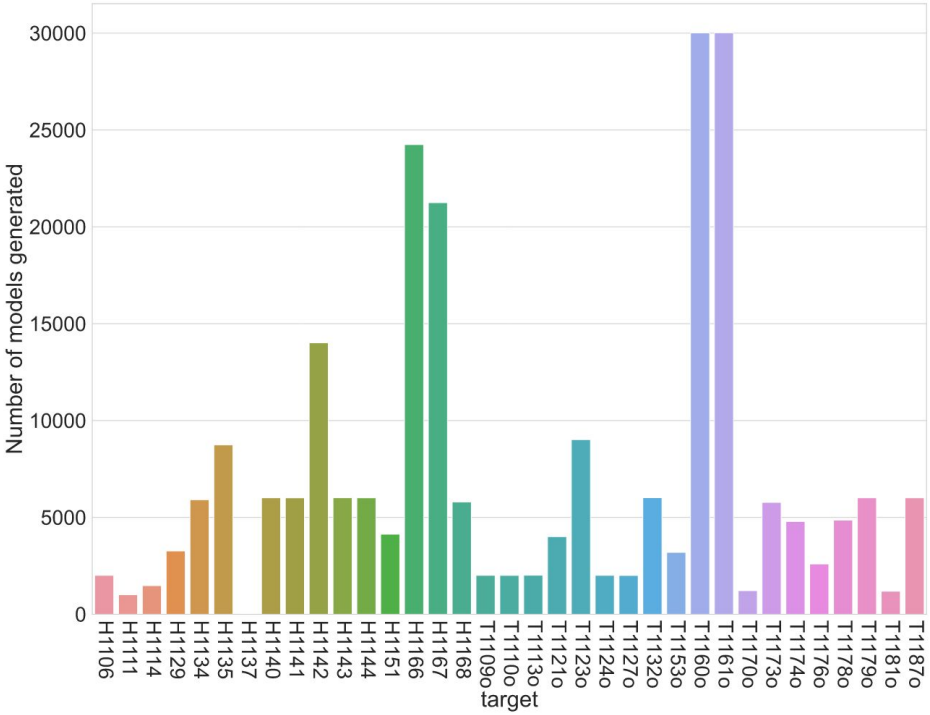
Dropout

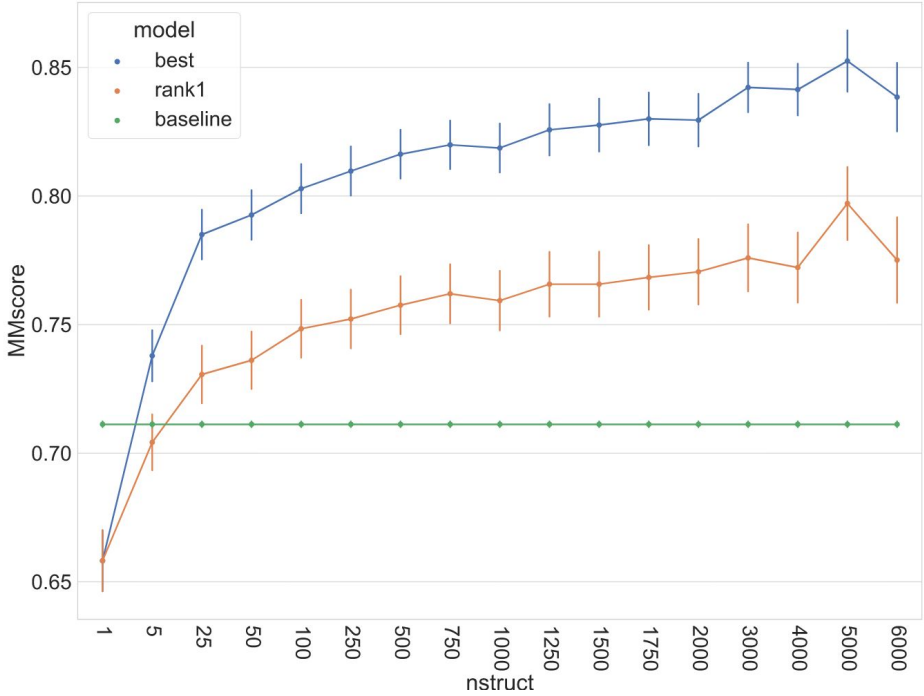


dropout_rate=0.1



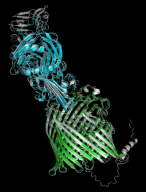
Amount of needed sampling





H1144 Antibody 3/6,000 models with `ranking_confidence>0.8`

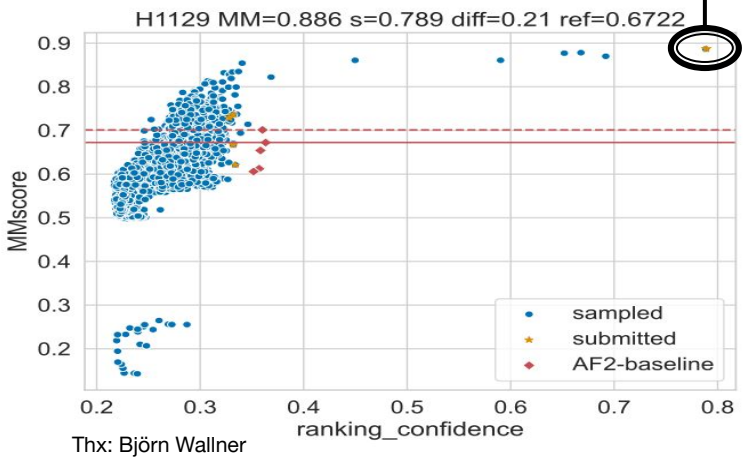
T200 – membrane protein



CASP15
H1129

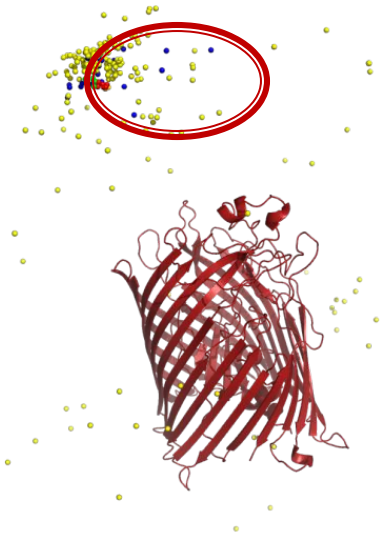
1. Increase sampling / variation
2. Rely on confidence ranking

71 Groups	
Top-1	Top-5
4	5
0	0
6	15



Thx: Björn Wallner

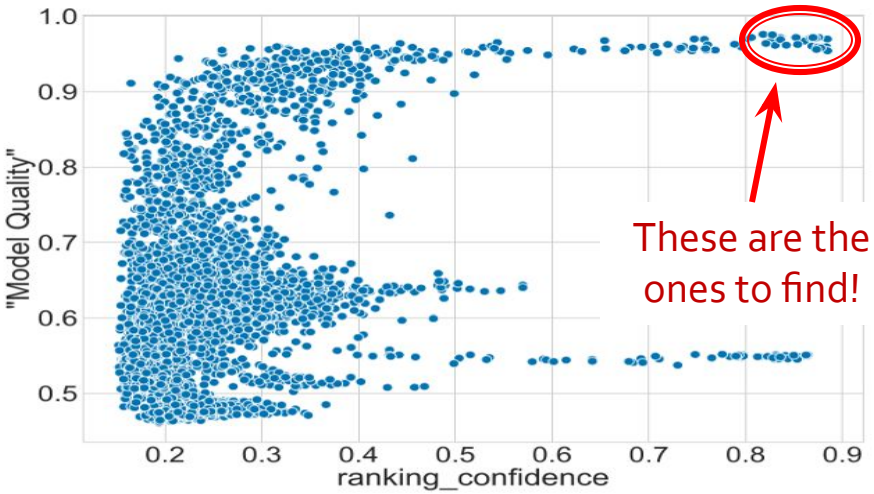
Distribution of Wallner models



Distribution of all submitted models

Slide Courtesy: Marc F. Lensink

Going beyond basic AlphaFold



Further progress is possible by "massive" exploration of the space of all possible structures

Every point in the graph is a single AlphaFold prediction

The main docking steps:

1. Producing a **huge** number of models
2. Scoring these models to find the near-native ones

Slide Courtesy: Marc F. Lensink

Scoring

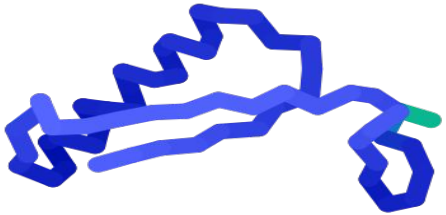
$$\mathrm{pDockQ} = \frac{L}{1+\exp[-k*(X-X_0)]} + b$$

$$X = <\mathrm{pLDDT}>_{\mathrm{int}} * \log(N_{\mathrm{int}}),$$

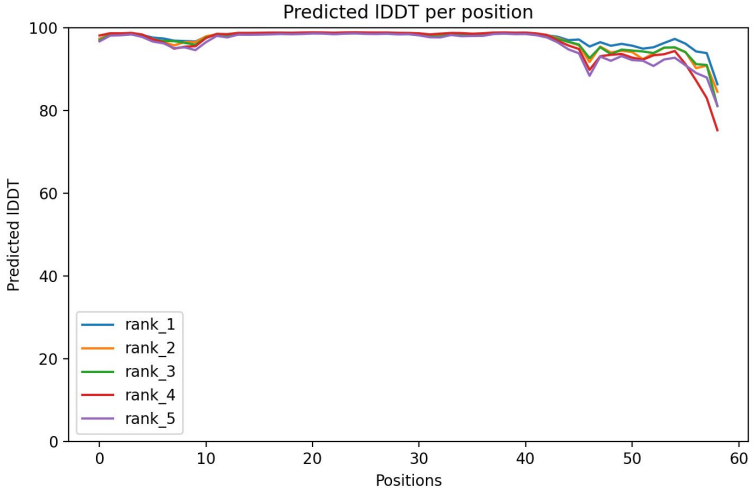
$$X_i = <\frac{1}{1+\left(\frac{\mathrm{PA\ E_{int}}}{d_0}\right)^2}> * <\mathrm{pLDDT}>_{\mathrm{int}}.$$

Confidence metrics

- **pLDDT** - "local" confidence per position
  - range 0 to 100 (higher better)
  - **Very low** (<50), **Low** (60), **OK** (70), **Confident** (80), **Very high** (>90)
  - Useful for deciding which local features (loops etc) are poorly modeled



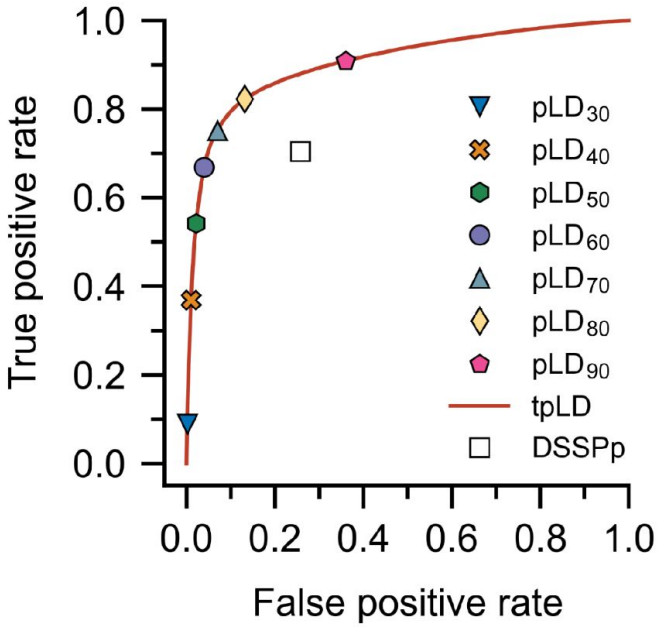
pLDDT 96.1



Slide courtesy of Sergey Ovchinnikov

pLDDT a measure of Intrinsic disorder ?

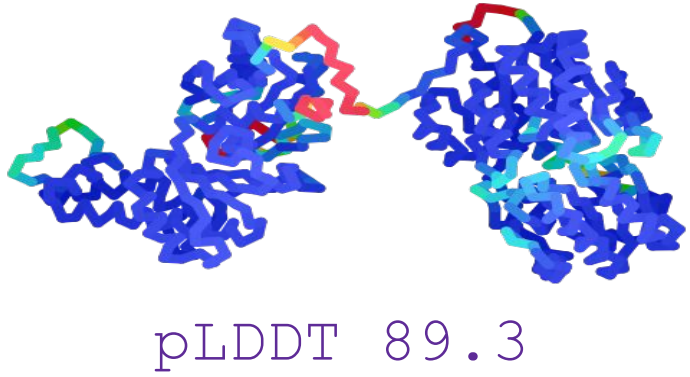
- tpLDDT = 1 − pLDDT/100
  - (1 is disordered and 0 is ordered)
- accurate metric for determining global and local disorder content
- Best metric so far
- NMR and MD simulation better tools

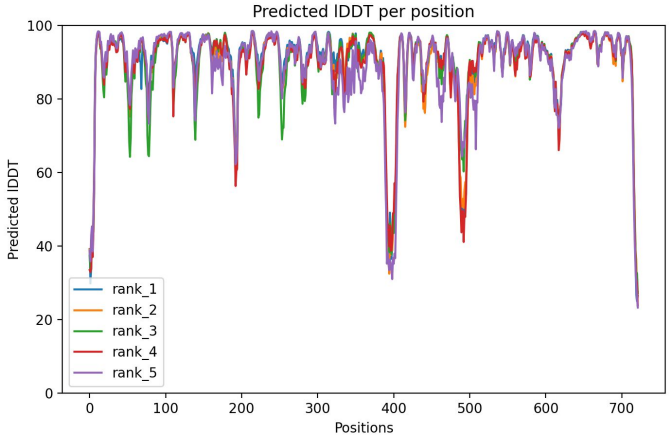


Wilson et al. 2022 Molecular Science

Confidence metrics

- **pLDDT** - "local" confidence per position
  - range 0 to 100 (higher better)
  - **Very low** (<50), **Low** (60), **OK** (70), **Confident** (80), **Very high** (>90)
  - Useful for deciding which local features (loops etc) are poorly modeled



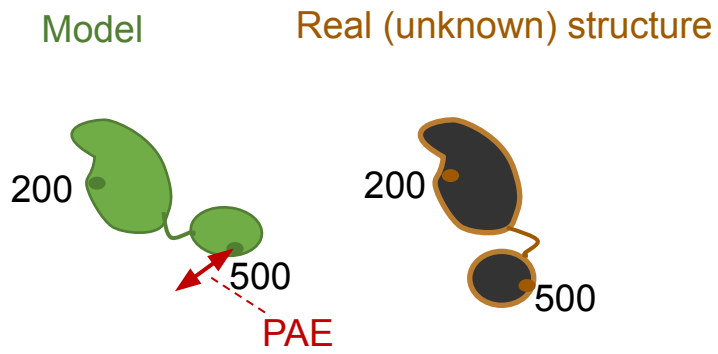


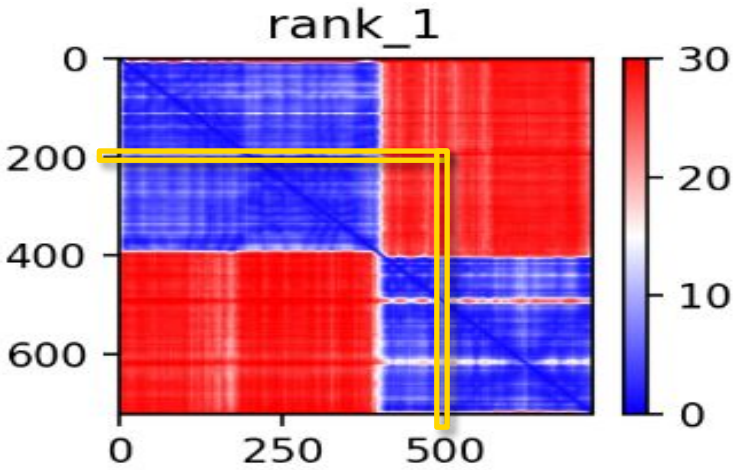
But wait... do we trust the domain-domain interaction?

Slide courtesy of Sergey Ovchinnikov

Confidence metrics

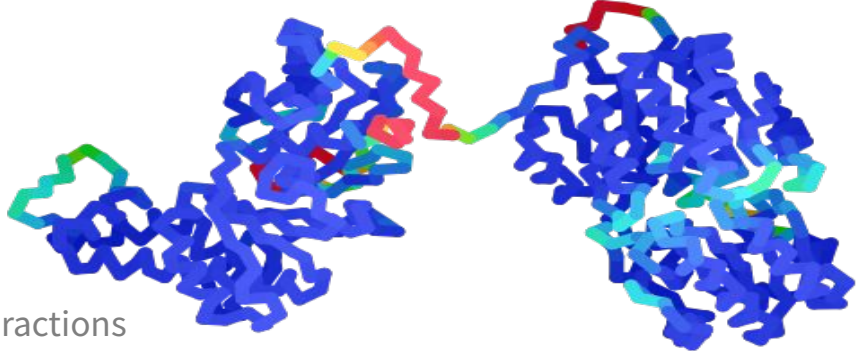
- **pAE** - confidence for every pair of positions
  - range 0 to 30 (lower better, in angstroms)
  - Useful for domain-domain or protein-protein interactions





AphaFold predicts that PAE without knowing the Real Structure!

Confidence metrics

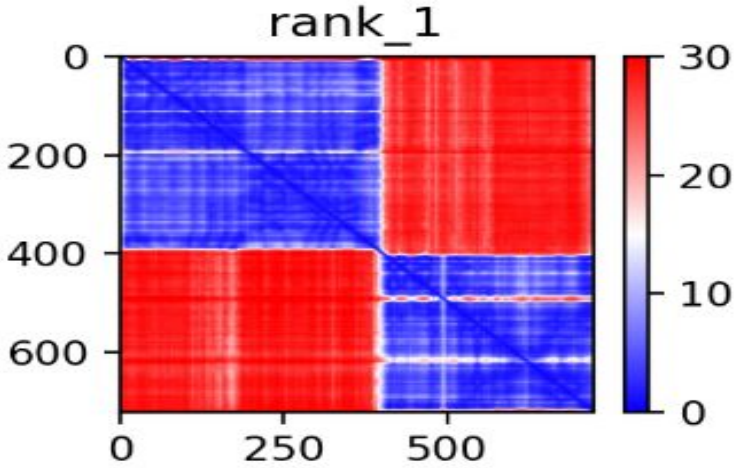


- **pAE** - confidence for every pair of positions
  - range 0 to 30 (lower better, in angstroms)
  - Useful for domain-domain or protein-protein interactions
- pTM - predicted TMscore (integrates pAE values)
  - range 0 to 1 (higher better)
  - good as a single value to tell you how good the overall structure is.
  - recommend value for confident structure > 0.7

pLDDT 89.3
pTMscore **0.577**

Multimer: ipTMscore !
(calculated from PAE)

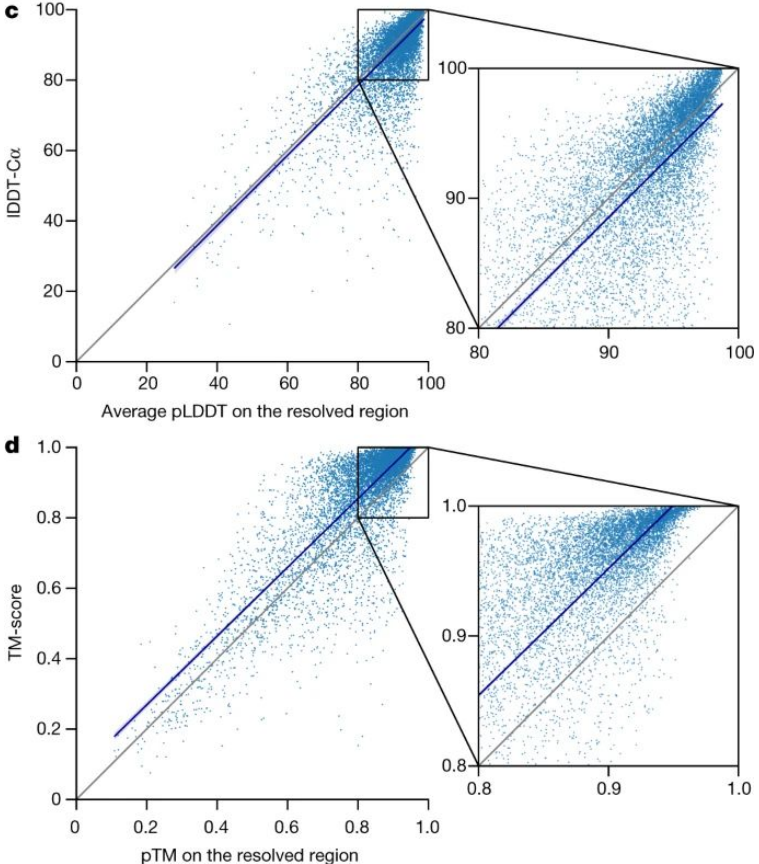
Ranking confidence = 0.8 ipTM + 0.2 pTM



Slide courtesy of Sergey Ovchinnikov

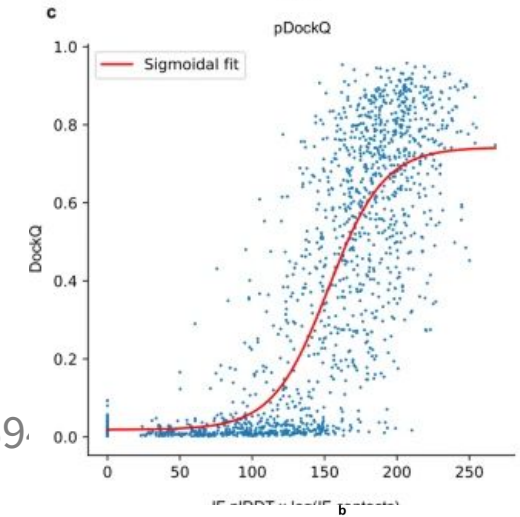
How can you trust the prediction

pLDDT and pTM are extremely accurate !



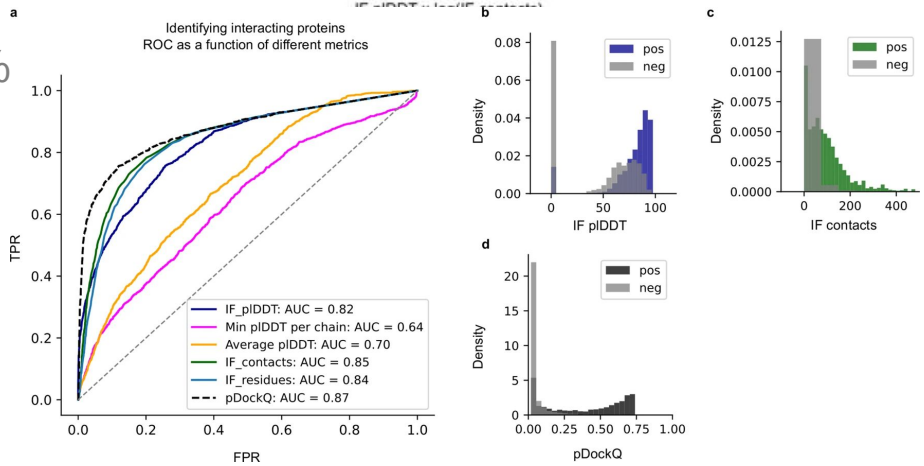
pDockQ

- models with acceptable quality (DockQ ≥0.23) for 63% of the dimers
- pDockQ allow to discriminate of interacting (n=1481) and non-interacting (n=569 proteins
- identify 51% of all interacting pairs at 1% FPR



$$\text{pDockQ} = \frac{L}{1 + e^{-k(x - x_0)}} + \text{b}$$

x = average interface plDDT · log(number of interface contacts)
with L = 0.724, x_0 = 152.611, k = 0.052 and b = 0.018



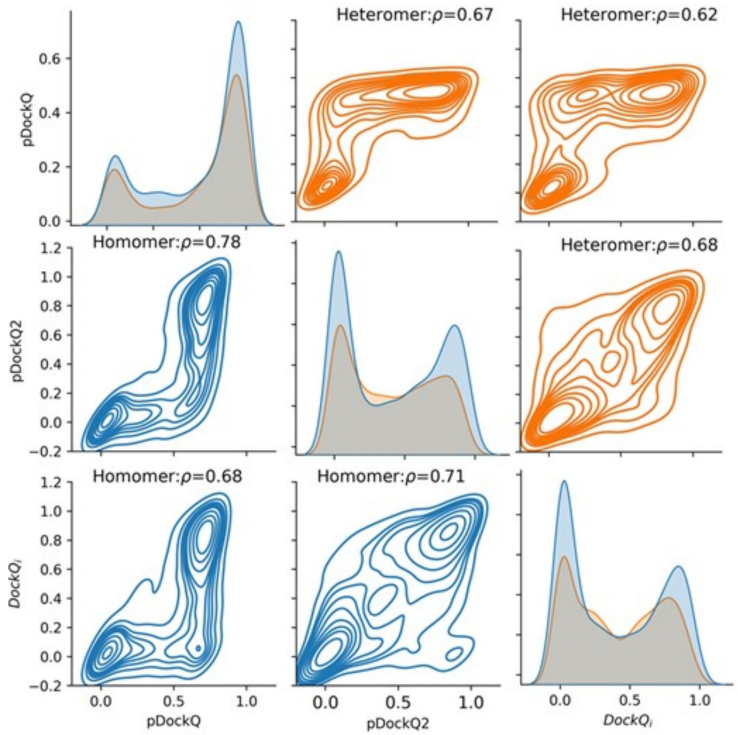
Bryant et al. Nature Comm. 2022

pDockq2

$$\mathrm{pDockQ} = \frac{L}{1+\exp[-k*(X-X_0)]} + b$$

$$X = <\mathrm{pLDDT}>_{\mathrm{int}} * \log(N_{\mathrm{int}}),$$

$$X_i = <\frac{1}{1+\left(\frac{\mathrm{PA\ E_{int}}}{d_0}\right)^2}> * <\mathrm{pLDDT}>_{\mathrm{int}}.$$

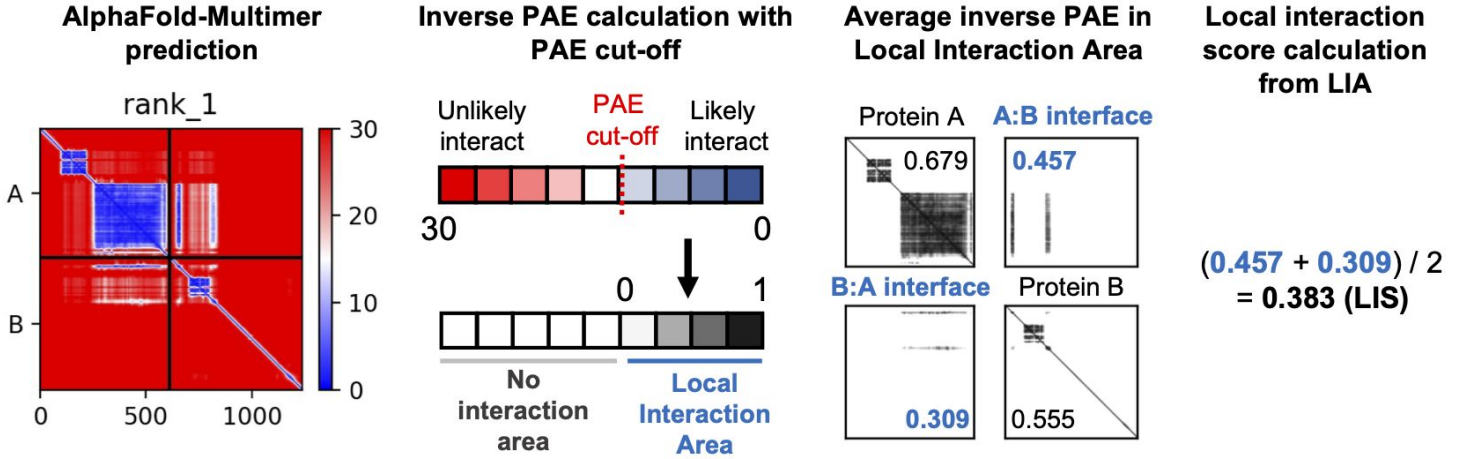


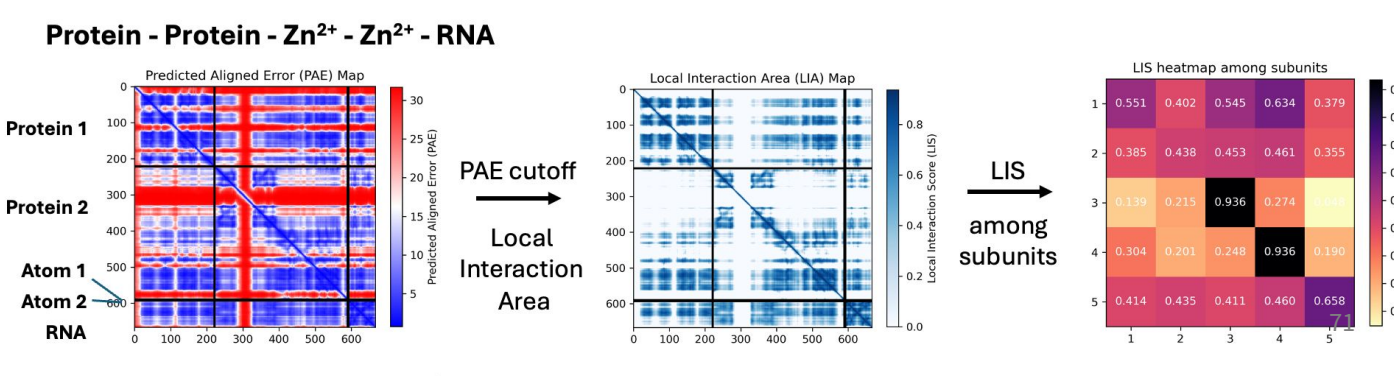
Zhu et al. Bioinformatics 2023

Local Interaction Score (LIS)

Kim et al. bioRxiv (2024)

One of the best score for multimer





AF_analysis

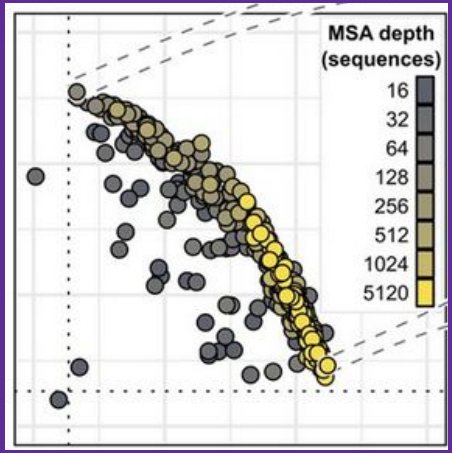
https://github.com/samuelmurail/af_analysis

- Analysis of Alphafold 2/3 and Colabfold models
- Python package
- GitHub (open source)
- Extract all models in a pandas Dataframe
- Allow pdockq, pdockq2, LIS calculation

```
import af2_analysis
my_data = af2_analysis.Data('MY_AF2_RESULTS_DIR')

# Extracted data are available in the
# df attribute of the Data object.

my_data.df

# Compute pdockQ and pdockQ2:
my_data.compute_pdockq()
my_data.compute_pdockq2()

# plot msa
my_data.plot_msa()

# plot plddt:
my_data.plot_plddt([0,1])

# plot PAE:
best_idx = my_data.df['ranking_confidence'].idxmax()
my_data.plot_pae(best_idx)

# show 3D structure (nglview required):
my_data.show_3d(best_idx)
```

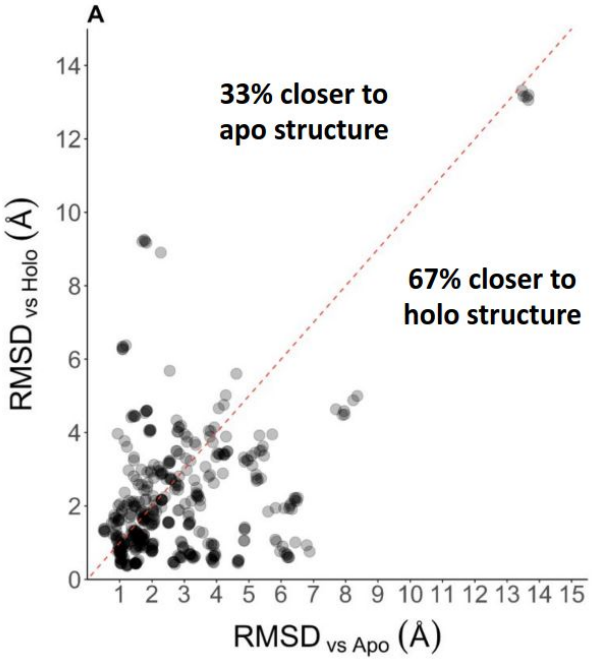
Regei and Murail JOSS 2025

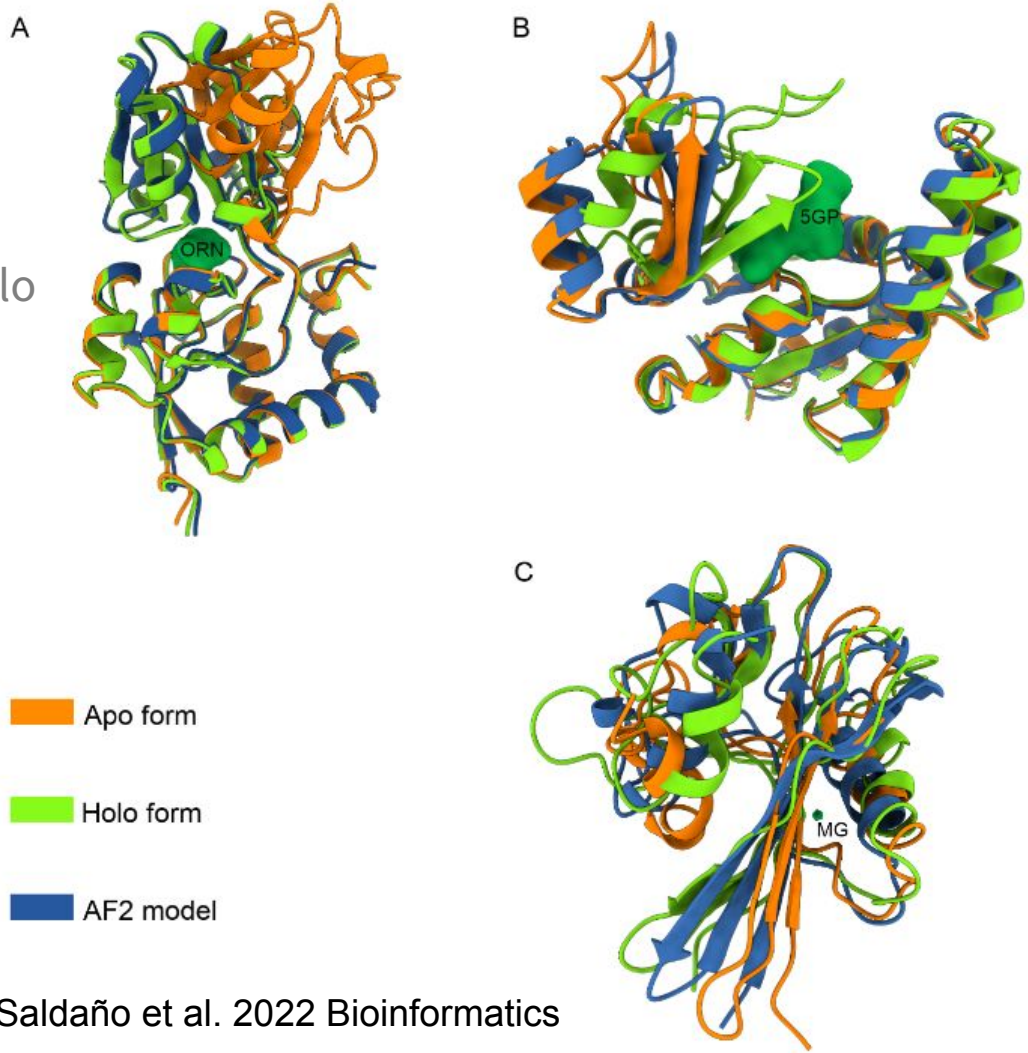
MSA Pruning



Apo vs. Holo

mostly indistinguishable from the holo form (67%)

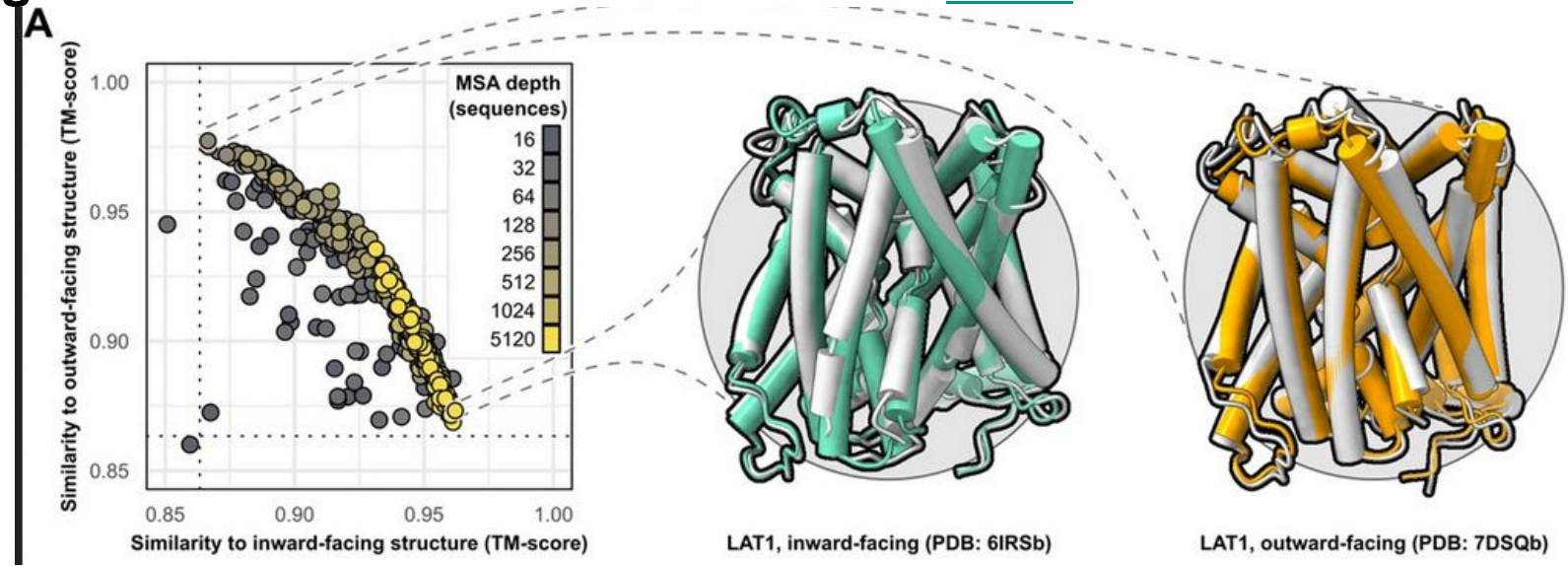




Saldaño et al. 2022 Bioinformatics

Sampling alternative conformations

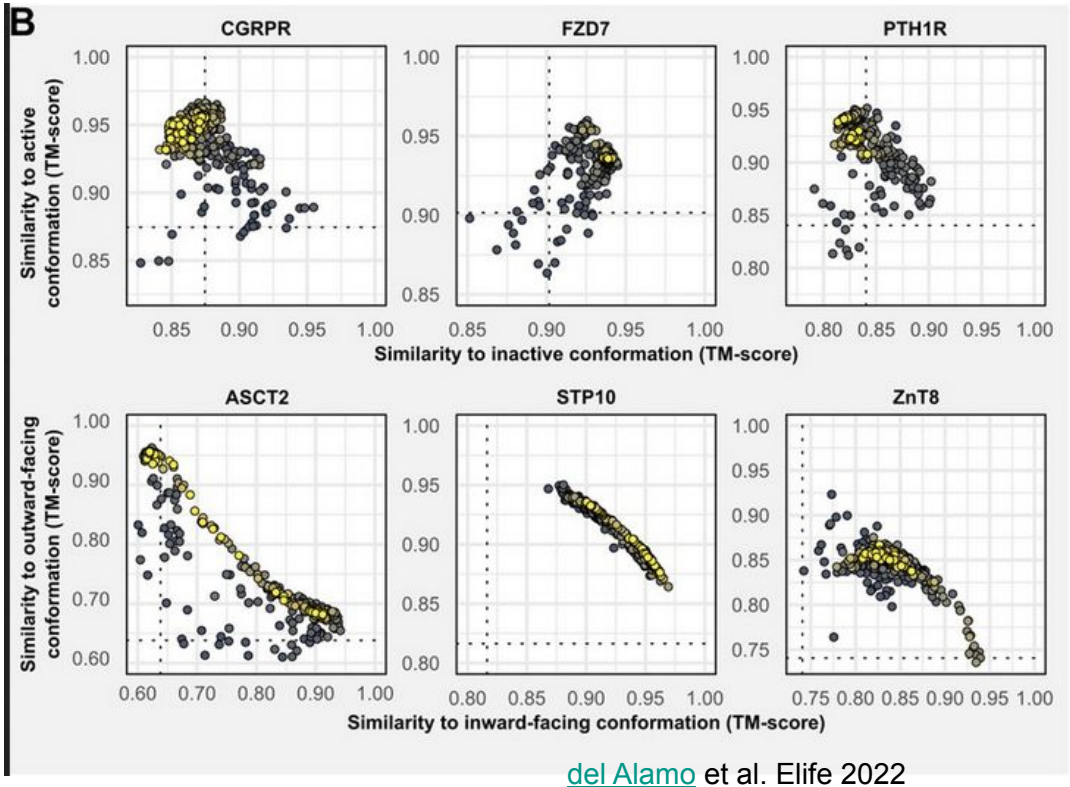
del Alamo et al. Elife 2022

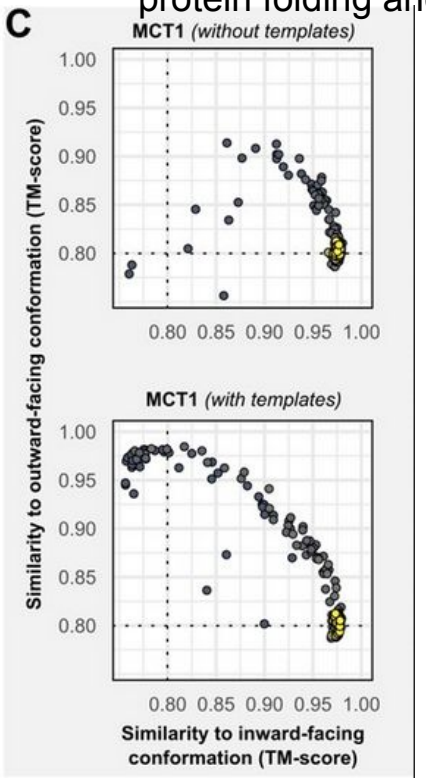


- GPCR
- reducing the depth of the input MSA rise the the conformational sampling
- argue against an optimal one-size-fits-all approach
- limited success when applied to transporters

Sampling alternative conformations (2)

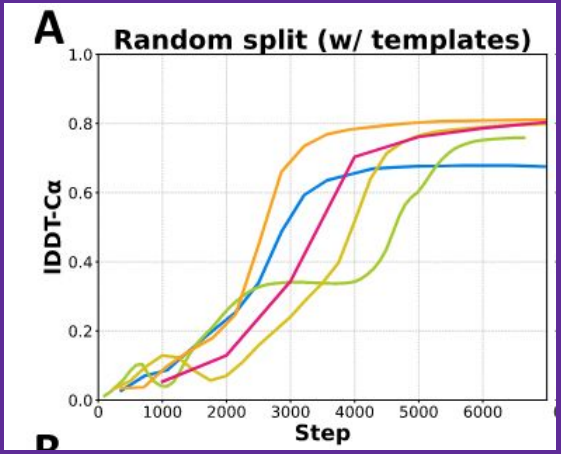
"Several preprints have provided evidence that AF2, despite its accuracy, likely does not learn the energy landscapes underpinning protein folding and function"





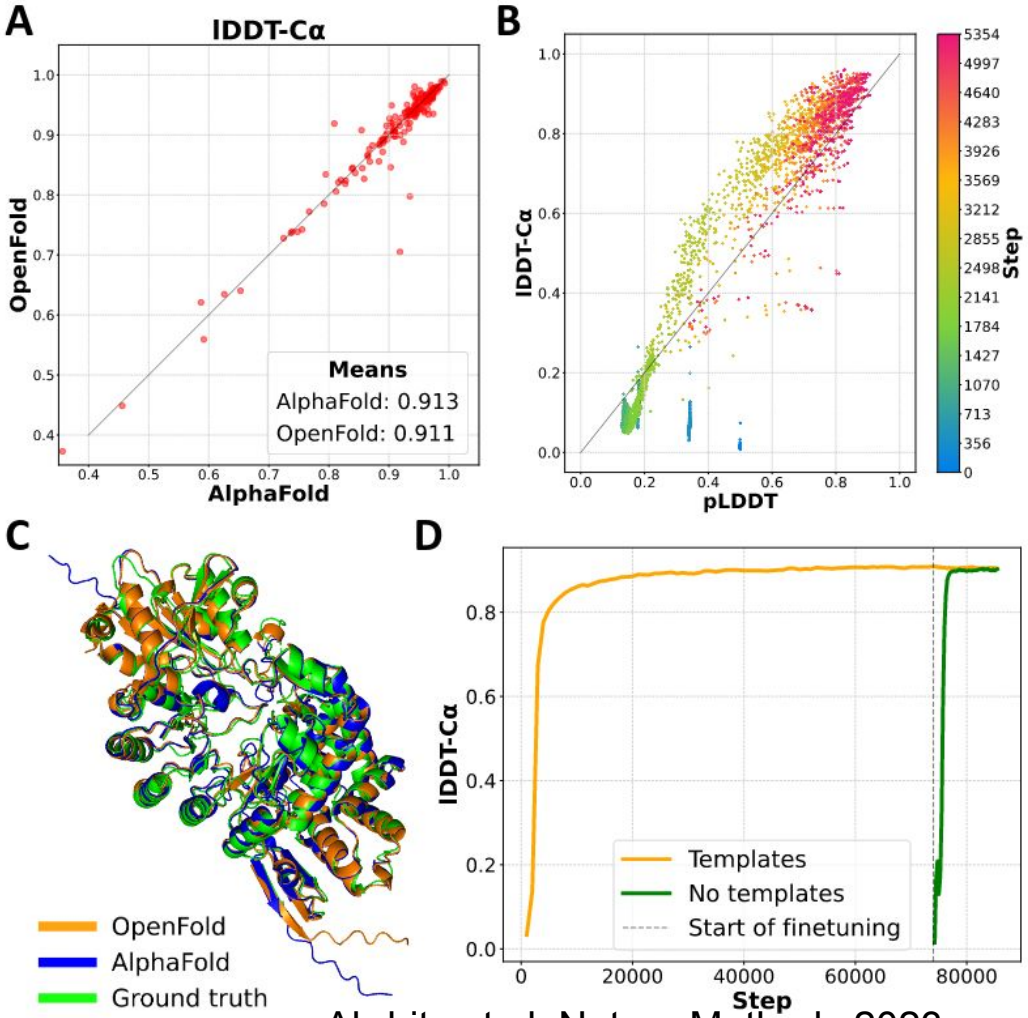
del Alamo et al. Elife 2022

Truly open source alternatives



OpenFold

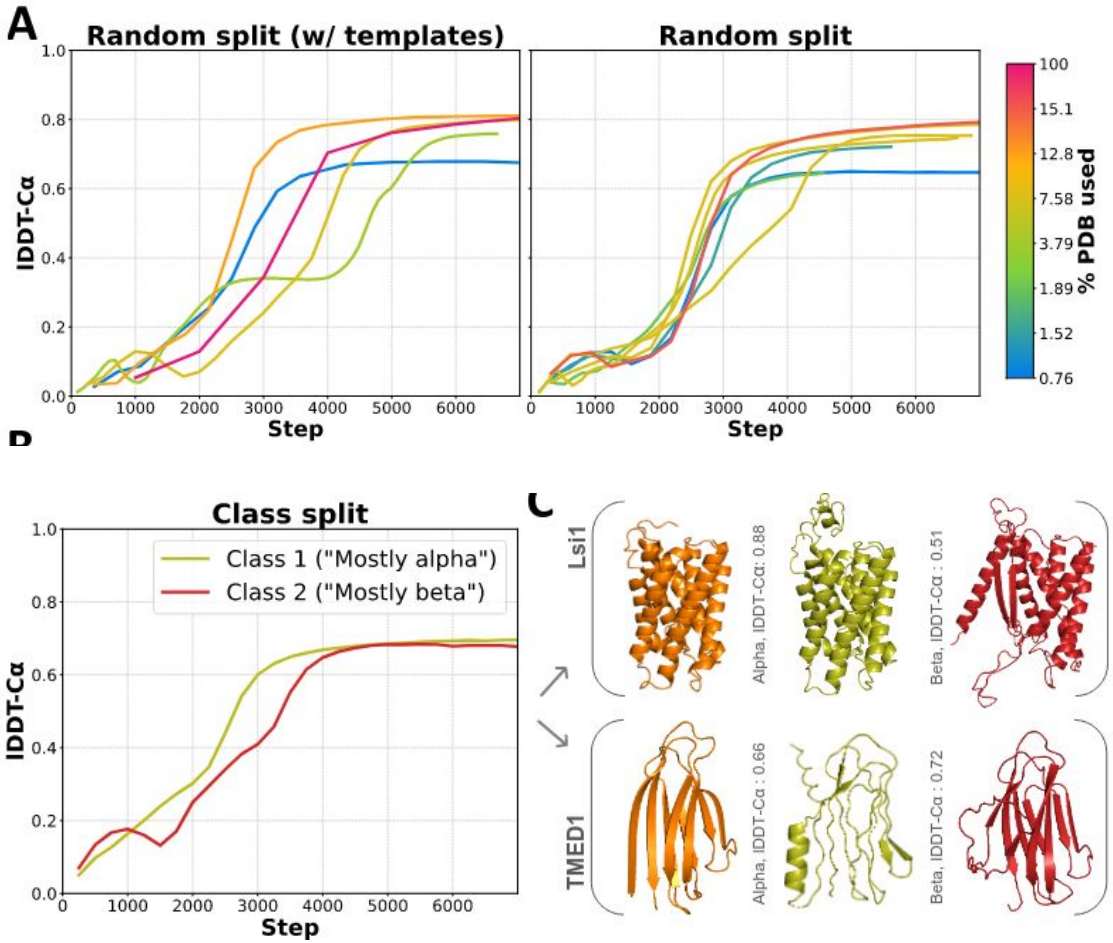
- OpenFold matches AlphaFold2 in accuracy
- predicts a sequence of (physically implausible) structures of increasing dimensionality
- https://figshare.com/articles/media/Folding_animations/21561939?file=38222889
- three times faster than AlphaFold2 for chains of length < 1100



Ahdritz et al. Nature Methods 2023

OpenFold (2)

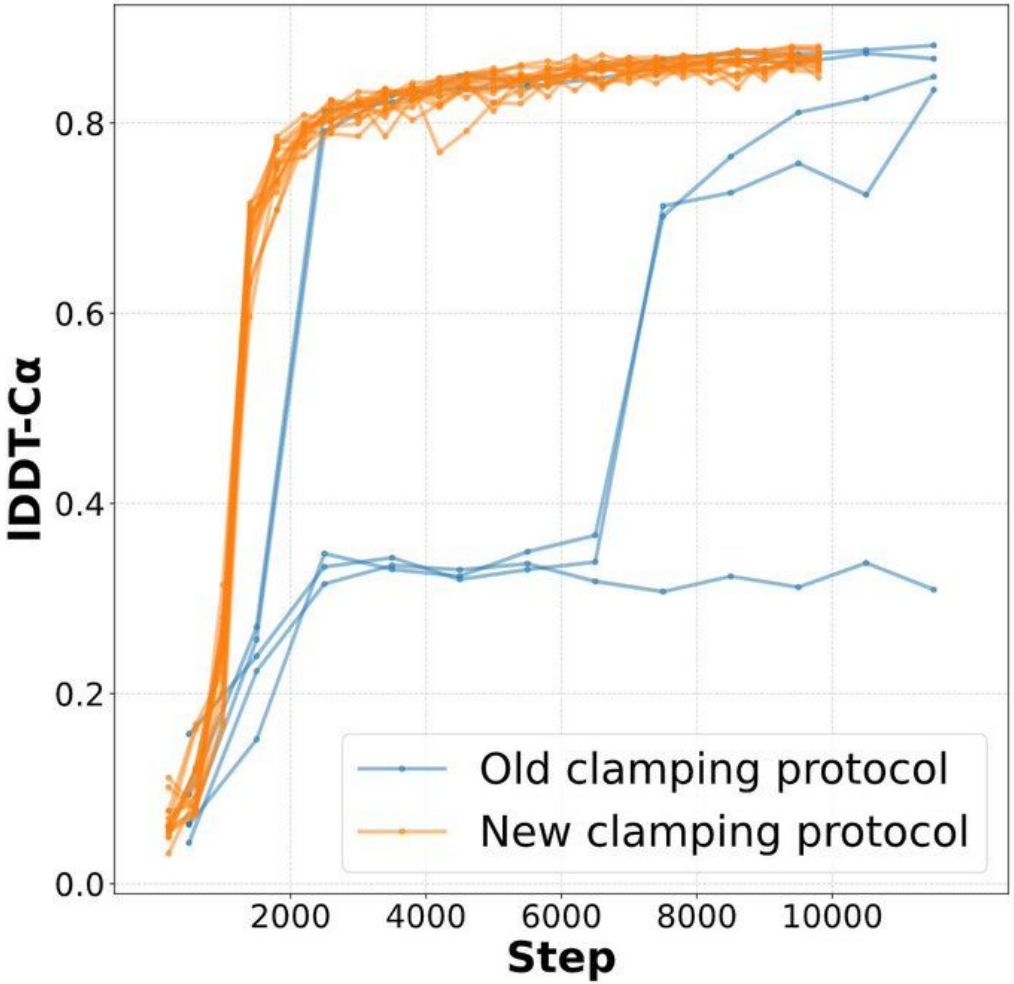
- 10, 000 protein chains—about 7.6% of all training data
- 1,000 protein chains, (0.76%) ~ Alphafold 1
- only 1,664 RNA structures.



Ahdritz et al. Nature Methods 2023

OpenFold (3)

- more efficient and trains more stably than AlphaFold2

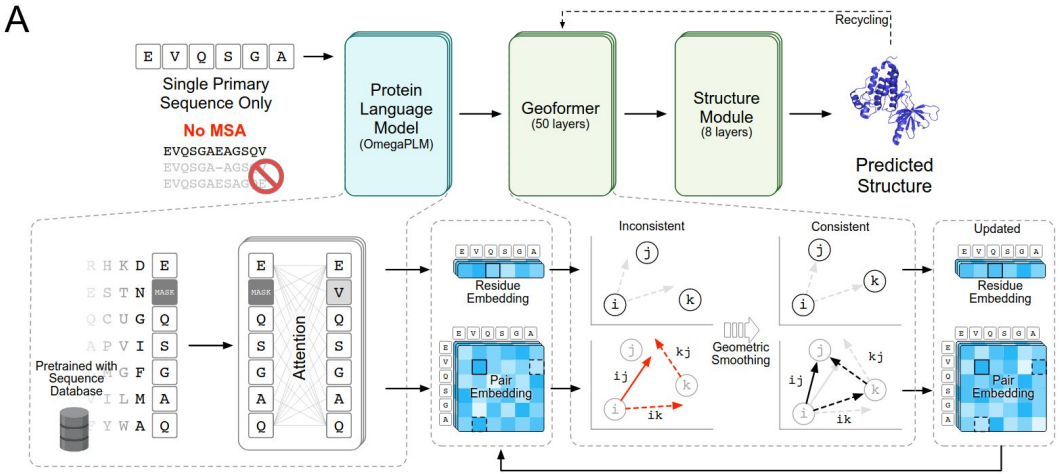


Ahdritz et al. Nature Methods 2023

OmegaFold

- No MSA preprocessing
- Deep language models

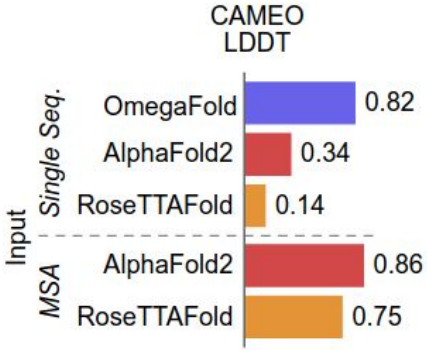
- capture structural and functional information encoded in the aa sequences
- evolutionary information may well be encoded in primary sequences



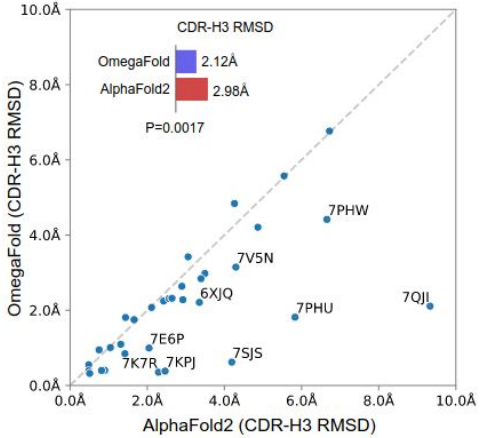
Wu et al. BioRxiv 2022

OmegaFold (2)

OmegaFold outperforms RoseTTAFold and AlphaFold2 on single-sequence inputs

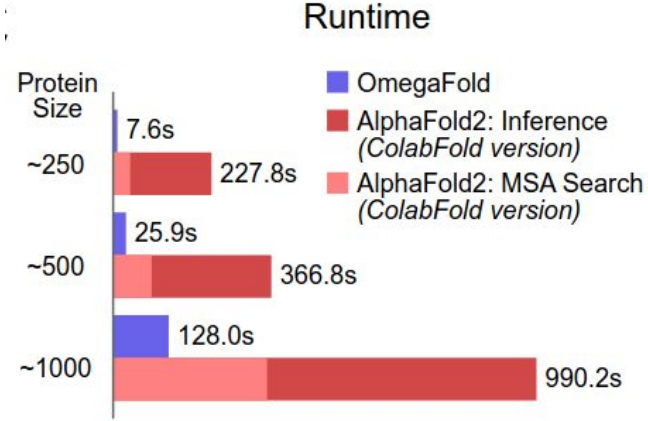


antibody loops

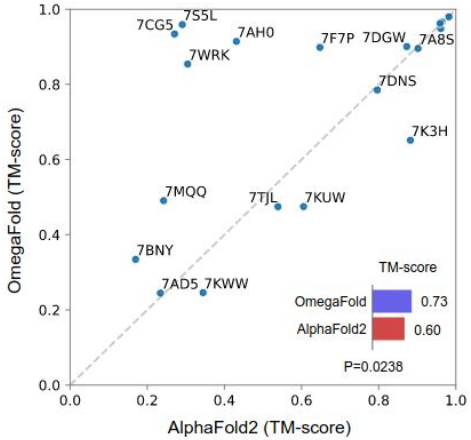


Very efficient on :

- antibody loops
- Orphan protein



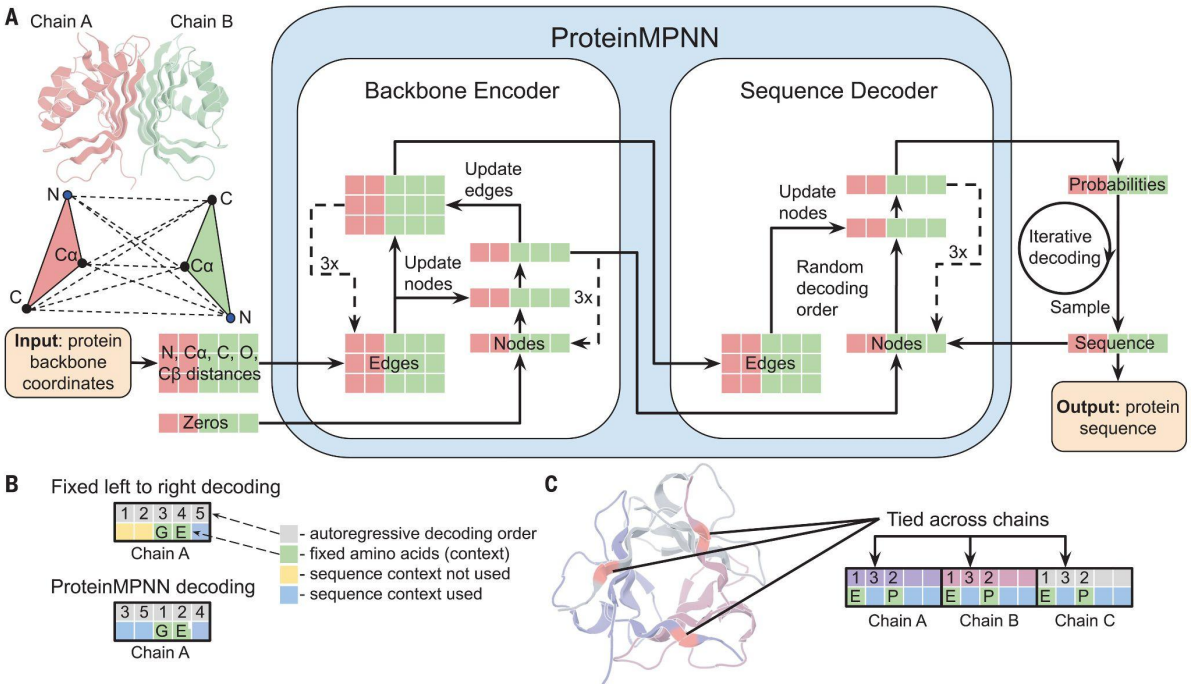
Orphan proteins

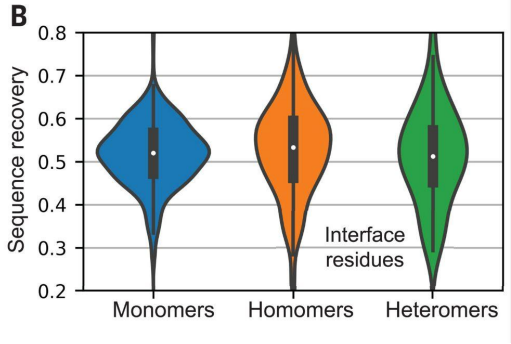


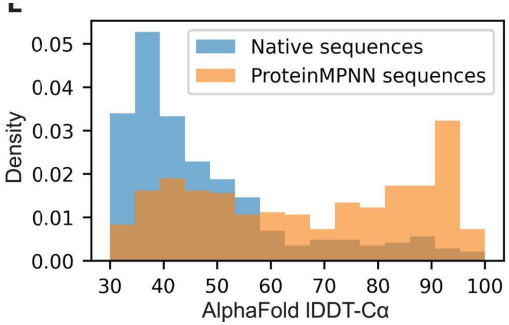
Wu et al. BioRxiv 2022

STILL NOT REVIEWED in 2024

ProteinMPNN

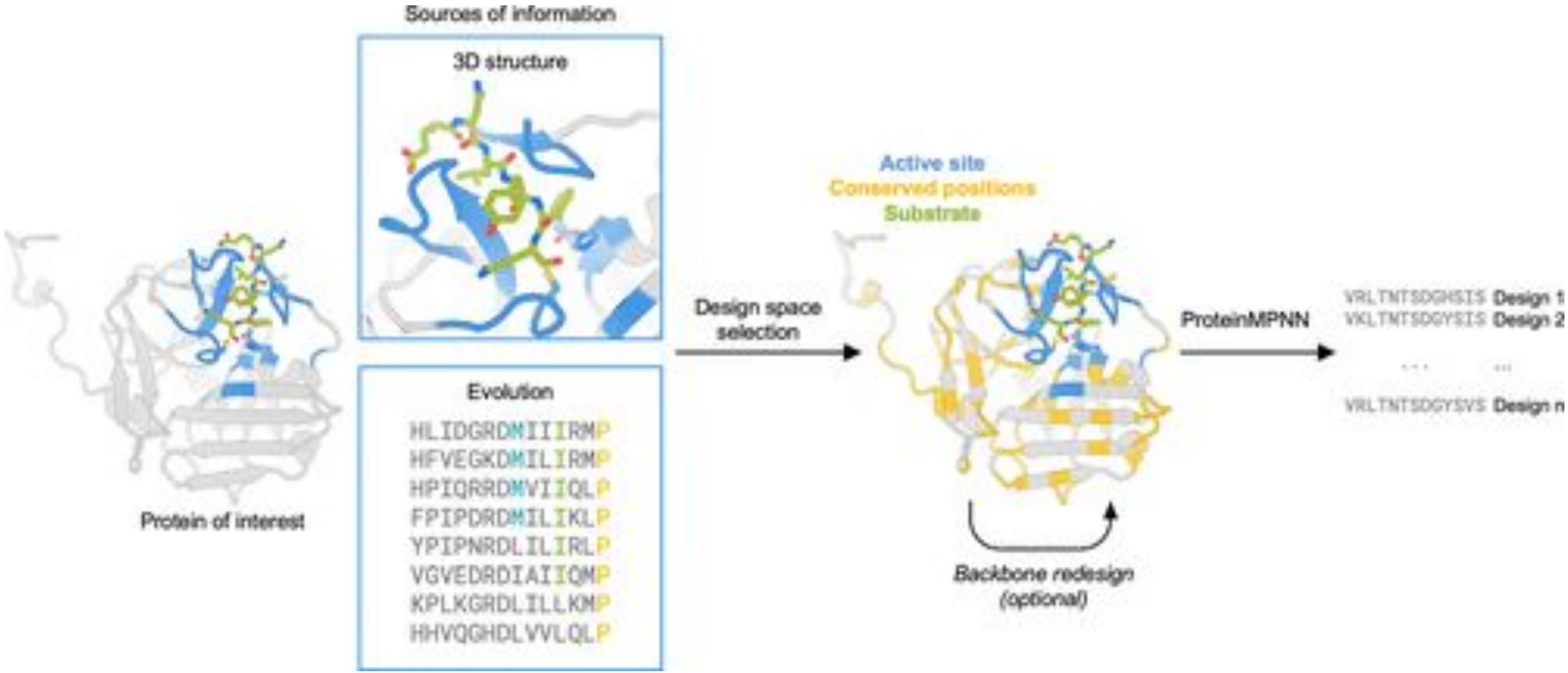






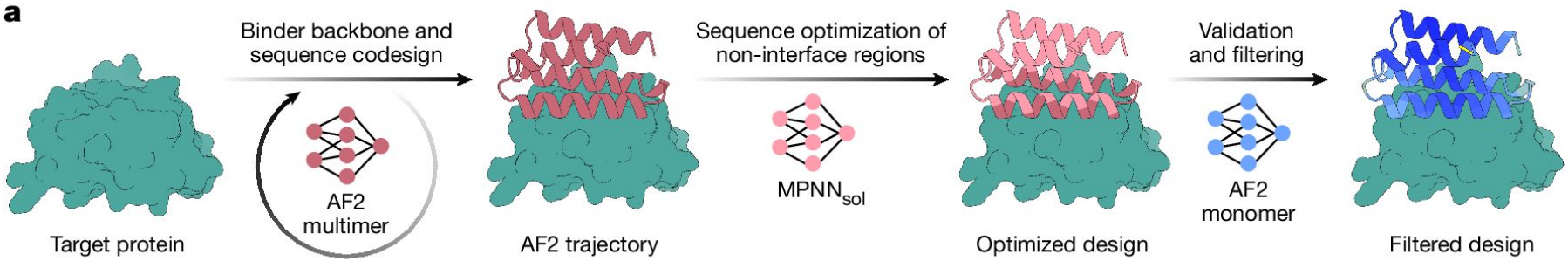
Dauparas et al. Science 2022

Protein MPNN (2)

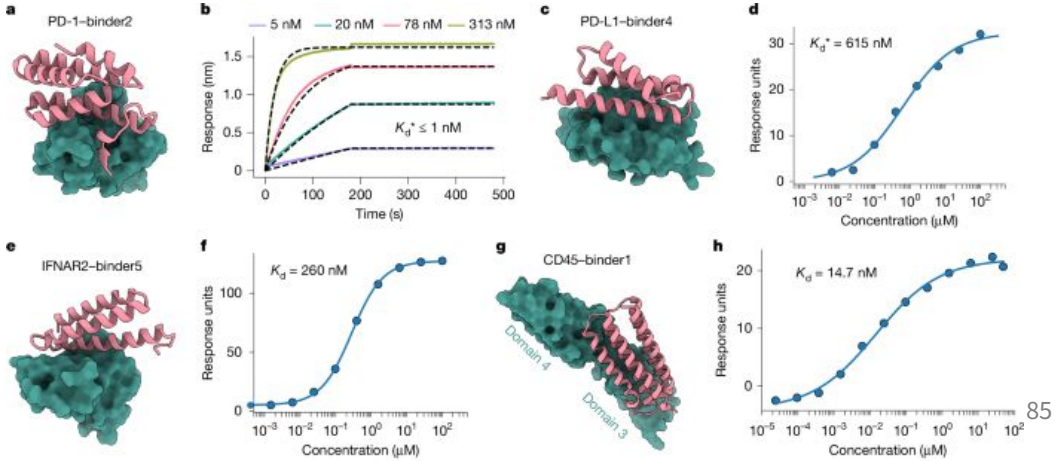


Sumida et al. JACS 2024

BindCraft

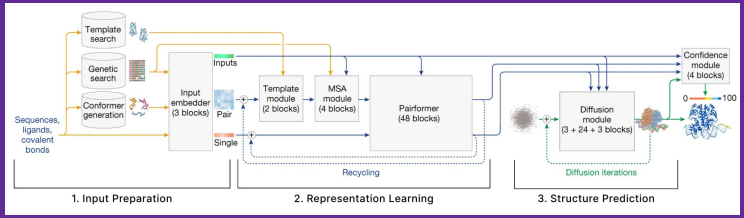


- Denovo Protein Binder designs
- 10–100% of success



Pacesa et al. Nature 2025

Alphafold 3

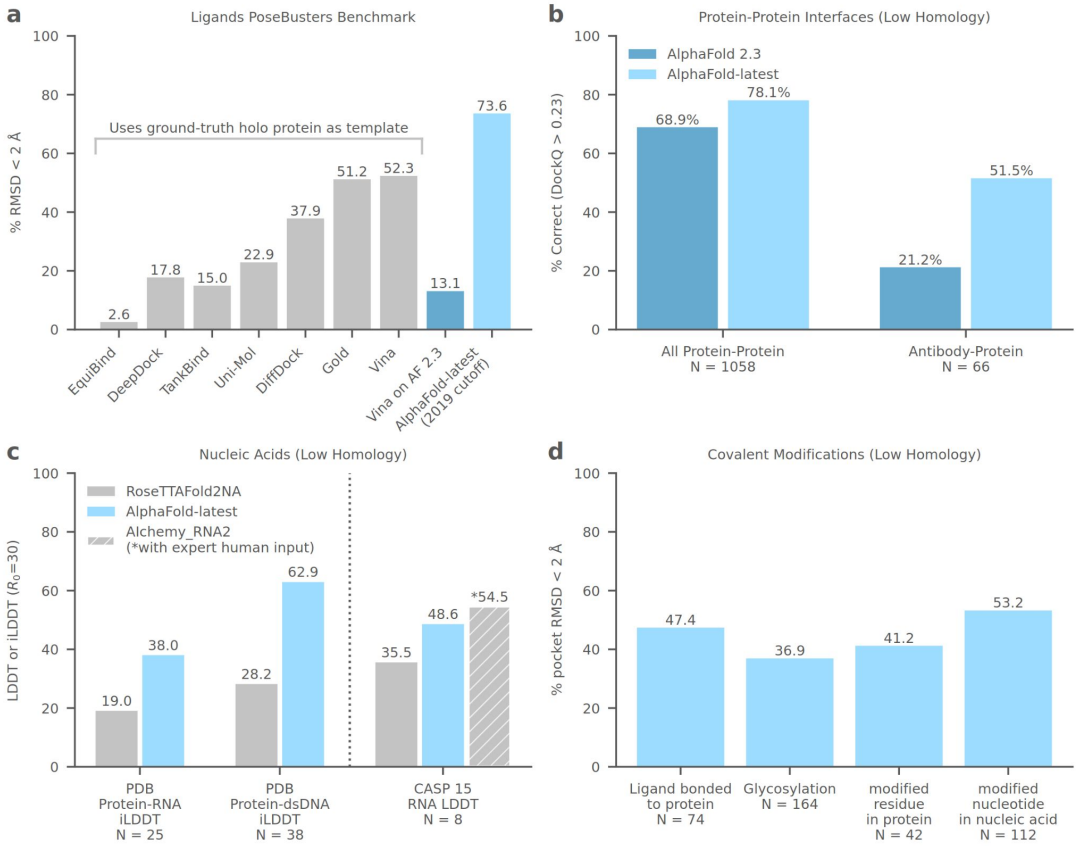

Template search
Genetic search
Conformer generation
Sequences, ligands, covalent bonds
Input embedder (3 blocks)
Inputs
Pair
Single
Template module (2 blocks)
MSA module (4 blocks)
Pairformer (48 blocks)
Recycling
Diffusion module (3 + 24 + 3 blocks)
Diffusion iterations
Confidence module (4 blocks)
0
100
1. Input Preparation
2. Representation Learning
3. Structure Prediction

AlphaFold 3.0

Support:

- nucleic acids
- small molecules
- Ions
- modified residues

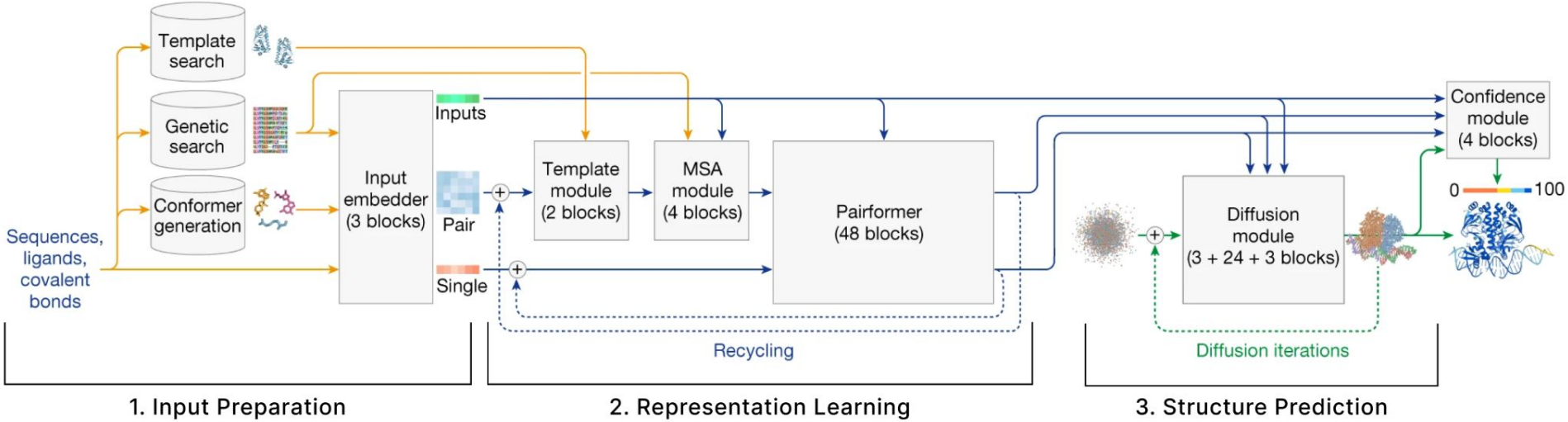
Protein-DNA and protein-RNA complexes, small molecule docking.



Abramson *et al.* Nature (2024)

DeepMind

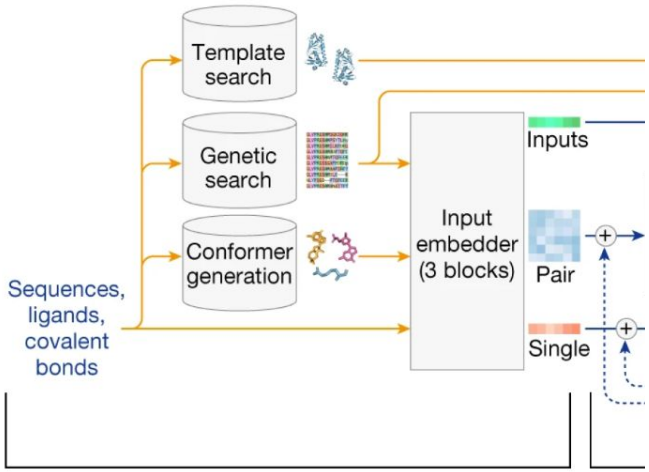
AlphaFold 3.0 architecture



Input Preparation

- Create MSA Similar to AF2, but MSA also for RNA, DNA
- Find out templates
- MSA webserver creation faster that AF2 !
- Ligand conformer with Rdkit

Create Atom-Level Representations and pair distance representation

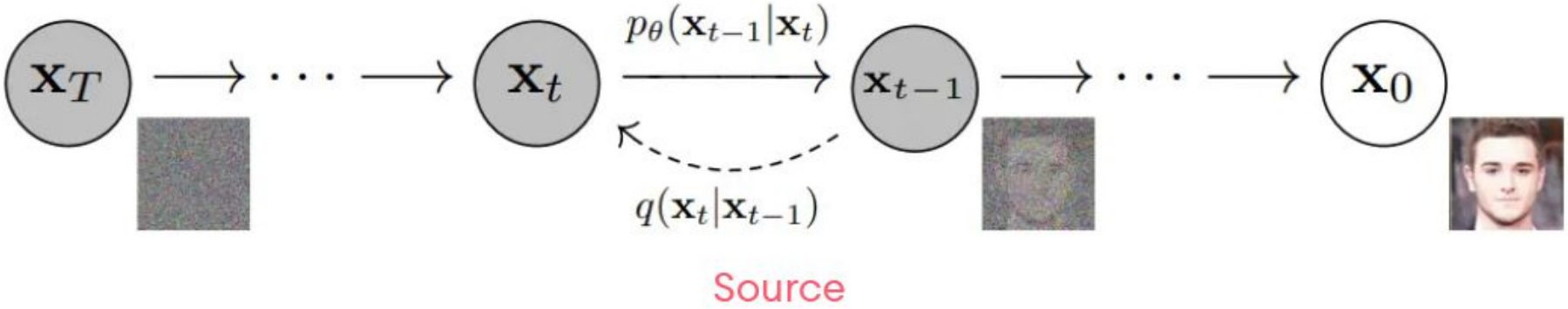


1. Input Preparation

Notes

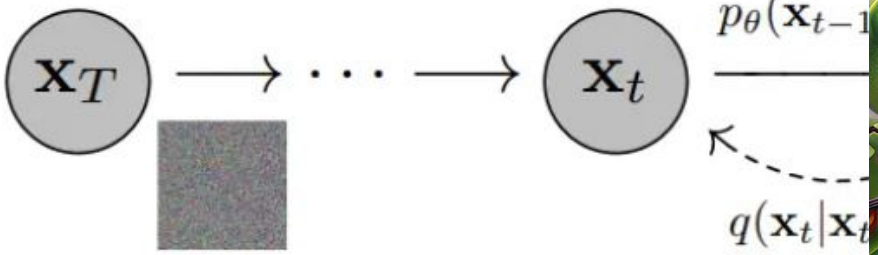
- The MSA module is smaller than in AlphaFold2.
- A Pairformer module replaces the Evoformer module of AlphaFold2. This module only processes the single and pair representations but not the MSA representations.
- The structure model in AlphaFold2 is replaced by a Diffusion model.
  - diffusion gives a distribution of structures instead of a single structure with uncertainty
  - no physics-based minimisation is needed as performed with AMBER
  - cross-distillation was used with training data from AlphaFold-Multimer v2.3

Diffusion Model



Deep Unsupervised Learning using Nonequilibrium Thermodynamics, arXiv:1503.03585, 2015

Diffusion Model



So

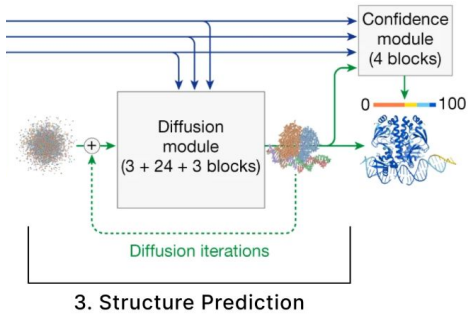
Deep Unsupervised Learning using Nonequilibr

Structure Module

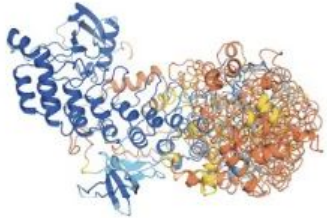
structure prediction is based on
atom-level diffusion

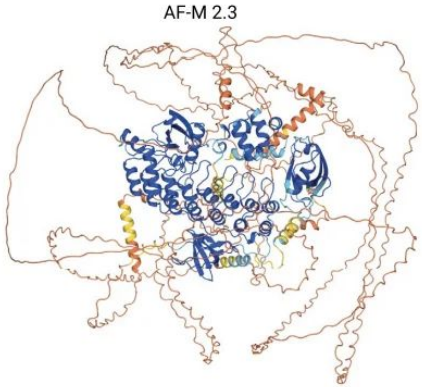
“AF3 uses a mix of synthetic training data generated by itself (via self-distillation) but also by AF2, via cross-distillation.”

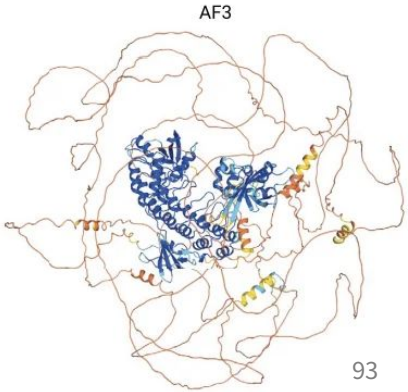
“If your previous model is doing one specific thing better than your new model, you can try cross-distillation to get the best of both worlds!”



AF3 no cross-distillation



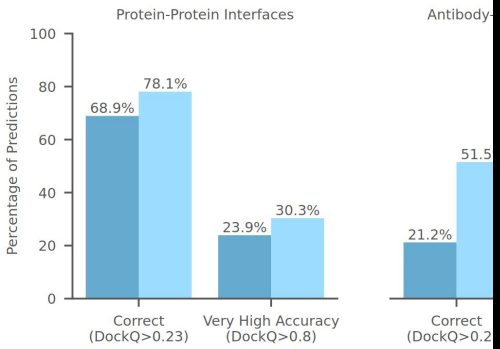


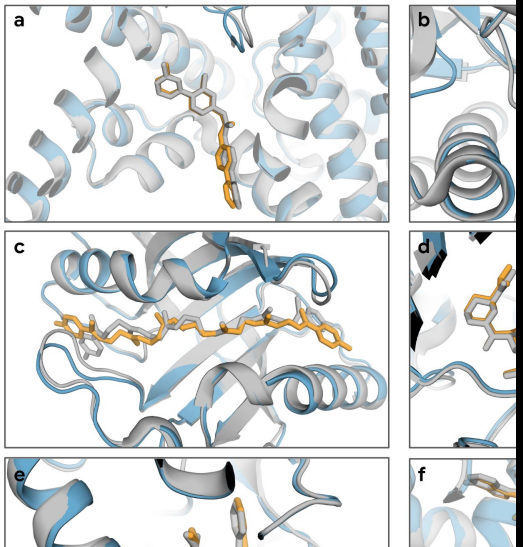


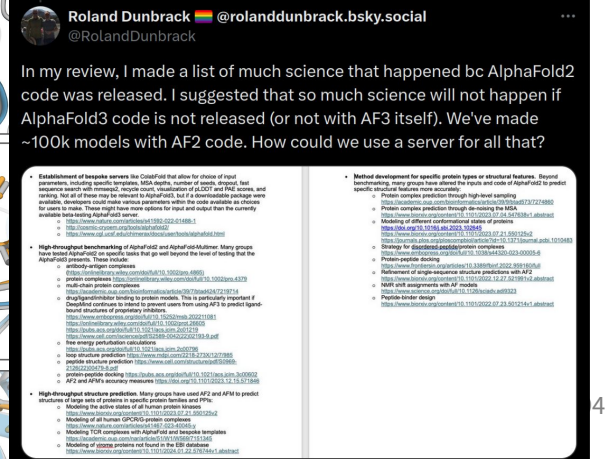
AlphaFold 3.0

- source code 1 year later
  - Deep mind said in 6 month …
  - No training
- Only webserver (20 model per days
- Licence
  - free for non commercial use

- https://alphafoldserver.com/





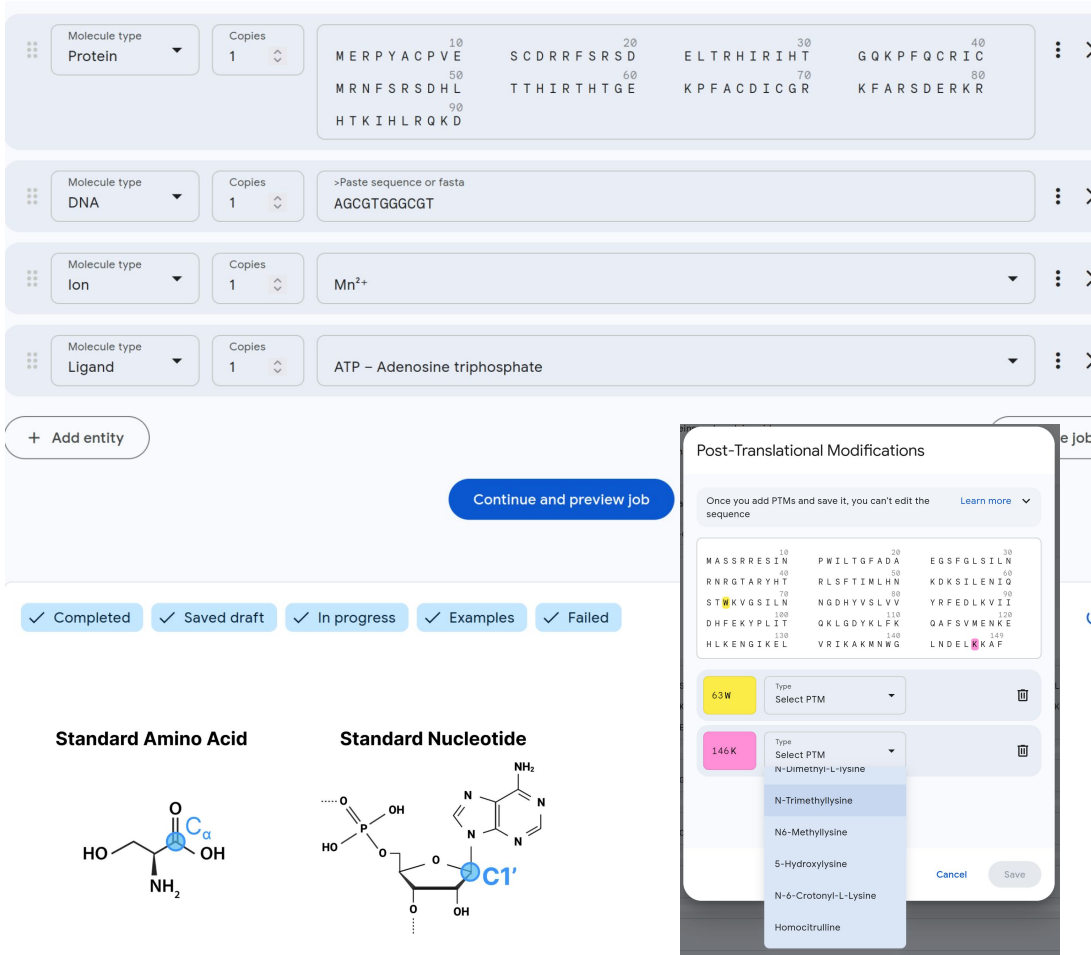


Input

- Sequence:
  - Protein **+ PTM**
  - DNA + Mod.
  - RNA + Mod.
- Ligands, limited to 20 choices (ATP, ADP, Heme, …)
- Ions, 10 choices (Ca^{2+}, K^{+}, Cl^{-},…)

Tokenisation:

- Protein C_{α}, Nucleid Acid C1’
- Ligand: 1 token per heavy atom



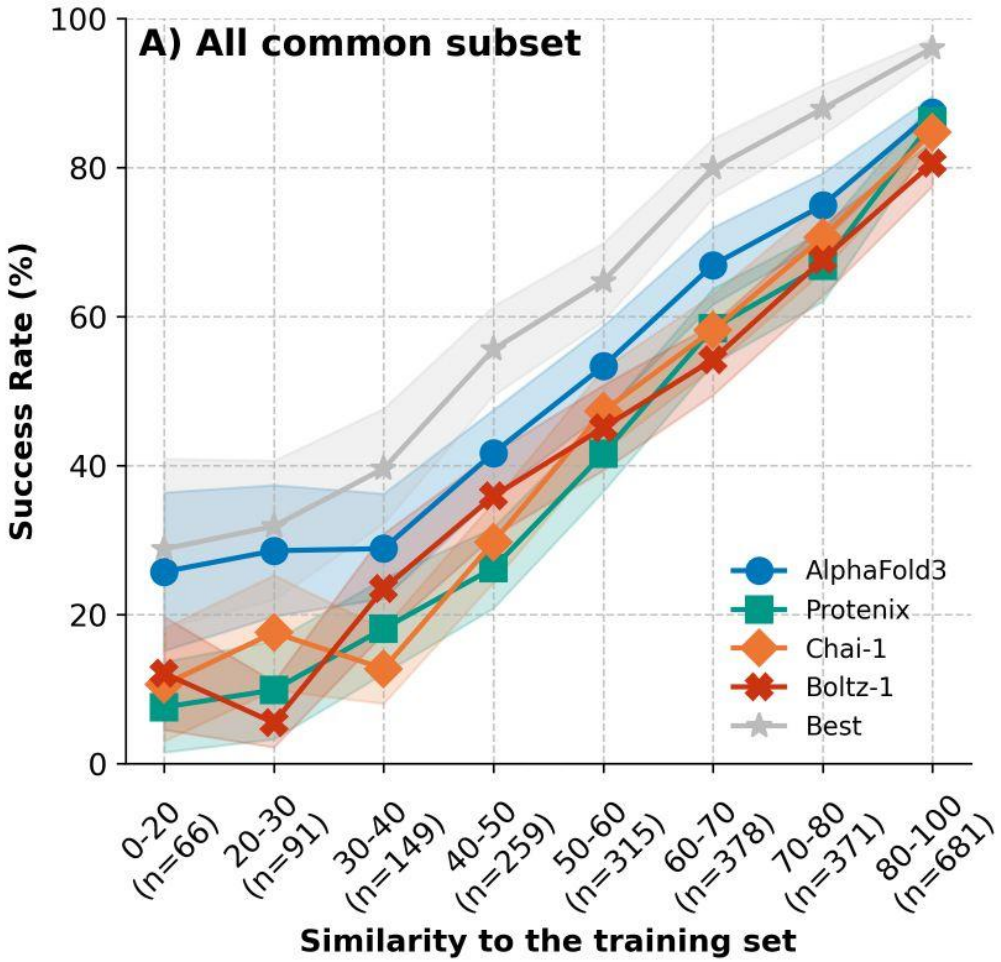
Alphafold 3 Replicates

Several alternatives in developpement

Free and truly open source

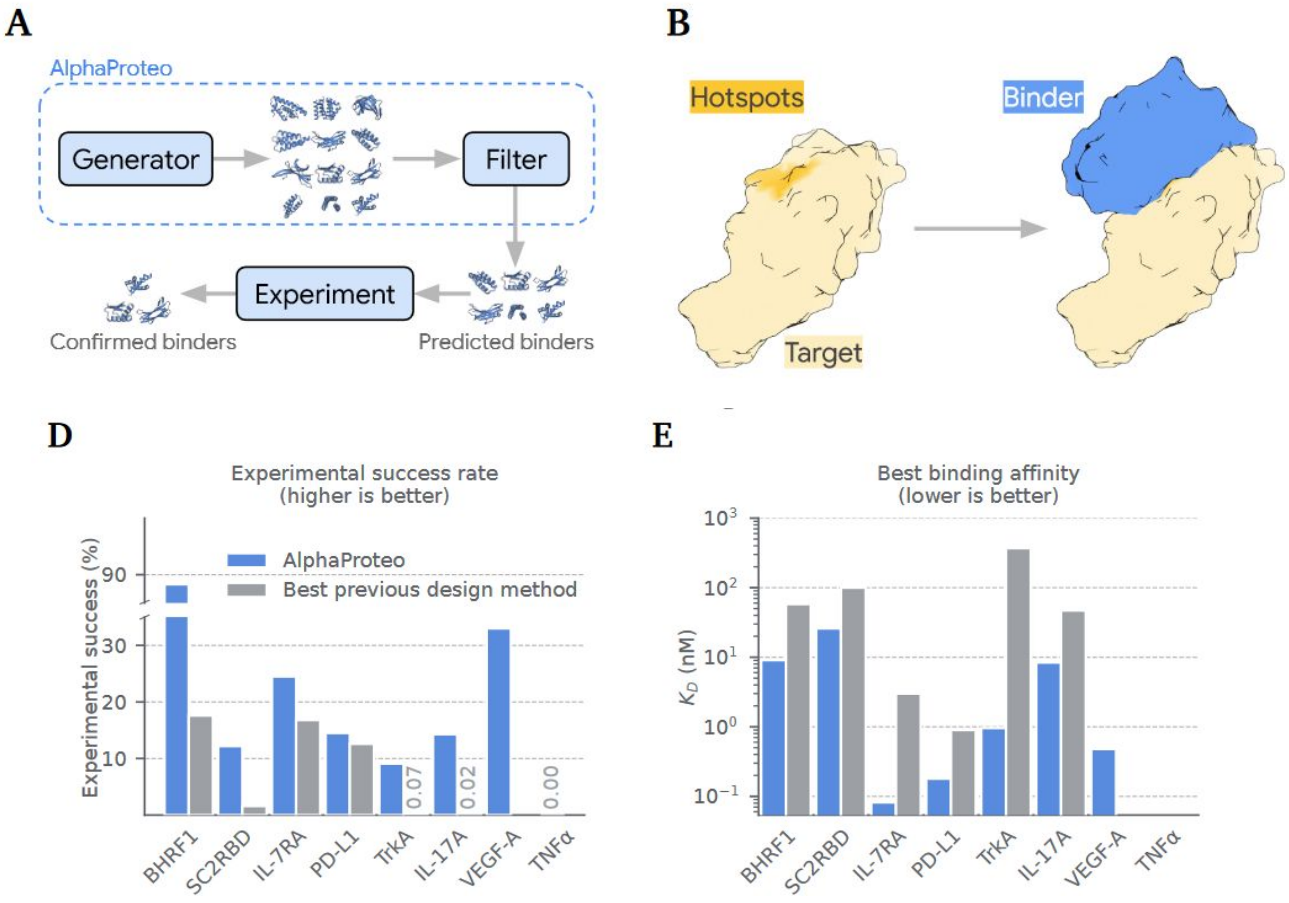
Škrinjar *et al.*

https://www.biorxiv.org/content/10.1101/2025.02.03.636309v1



AlphaProteo

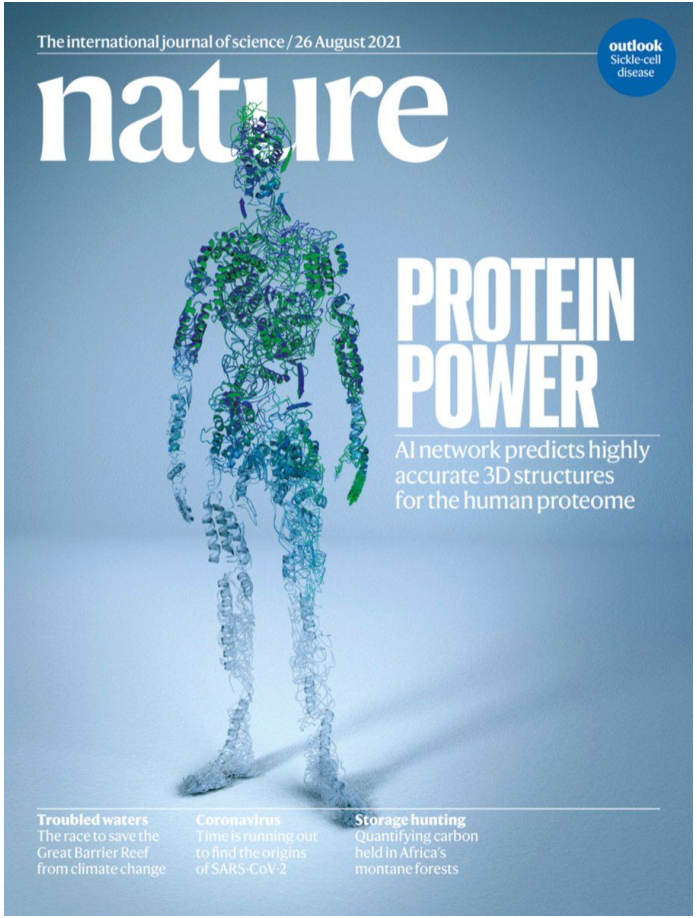
- Zambaldi et al.
- No source code
- No reviewed paper
- Only a report



DeepMind

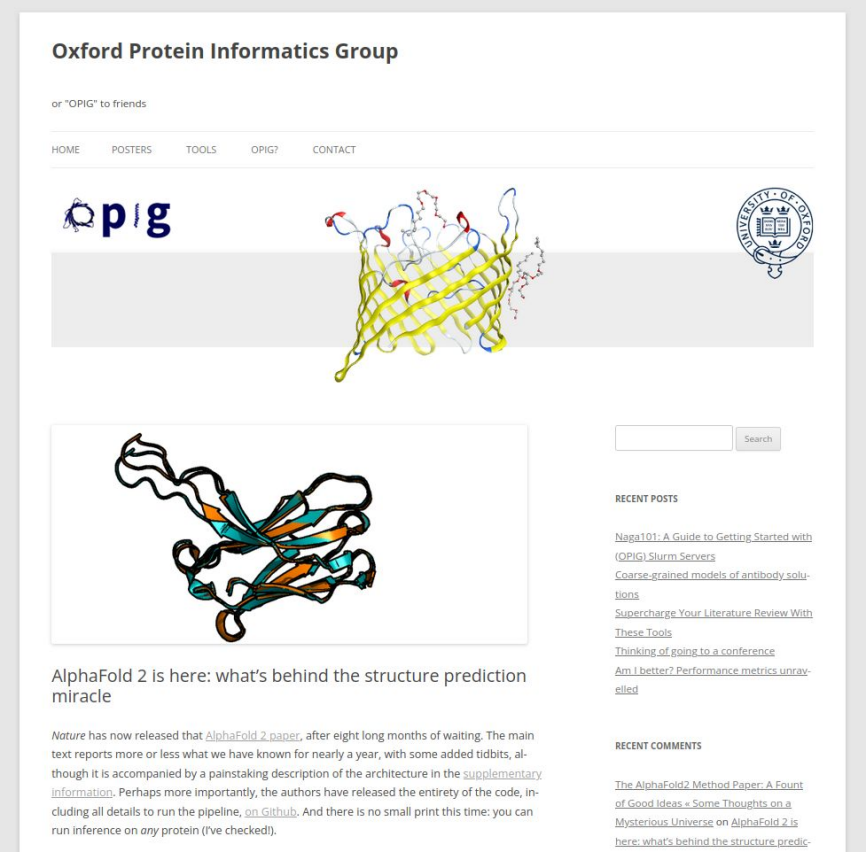
Take Home Message

- AF is a revolution !
- Its accuracy is without precedent
- Confidence metrics are extremely precise
- Accessible for free through:
  - AF EMBL EBI's DB
  - Google Colab
  - Source code with GPU
- Key to improved Alphafold accuracy:
  - Higher Sampling combined with Dropout
  - MSA sampling
  - Scoring



Bibliography

- Excellent Blog article from Carlos Outeiral Rubiera (Oxford)
- https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/
- Blog from Elana P. Simon (Stanford)
- https://elanapearl.github.io/blog/2024/the-illustrated-alphafold/
- Additional slides
- https://twitter.com/jankosinski/status/1565803556547993606



Going further AlphaFold 3.0

- https://medium.com/@falk_hoffmann/alphafold3-and-its-improvements-in-comparison-to-alphafold2-96815ffbb044
- https://elanapearl.github.io/blog/2024/the-illustrated-alphafold/
-

Acknowlegment

- Julien Rey for the deployment of the AlphaFold 2 server at RPBS.

- Nicolas Chevrollier, Gabriel Tourillon, Gautier Moroy and Pierre Tuffery

- Alaa Reguei for contributing to **af2_analysis**

- The RPBS platform for computational resources.

- IdEx Université Paris Cité n°ANR-18-IDEX-0001 projet GPU-APBS 2023

- Julien Dumont (IJM) as Beta-tester