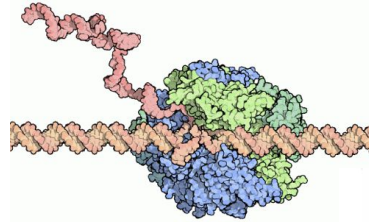
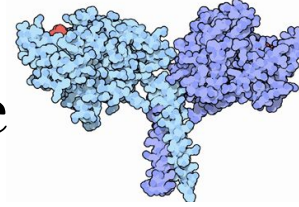


AI-based modeling of biomolecular interactions

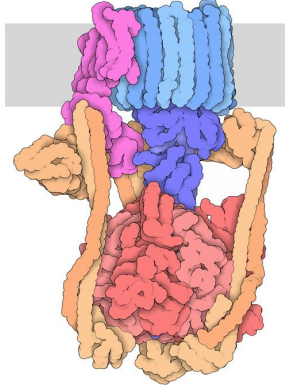


RNA polymerase

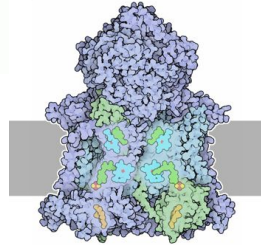
Molecules of life



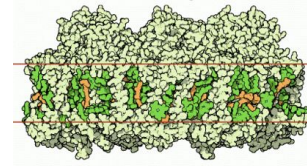
Kinesin



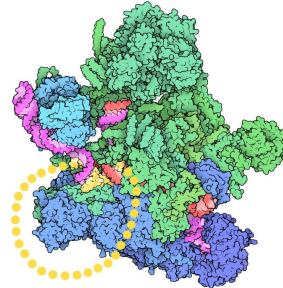
ATP synthase



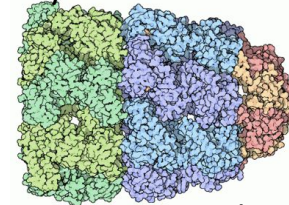
Cytochrome bc1



Photosystem I

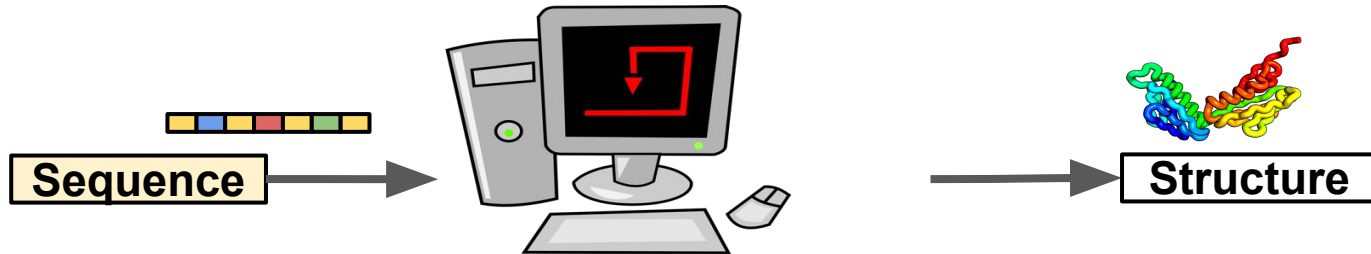
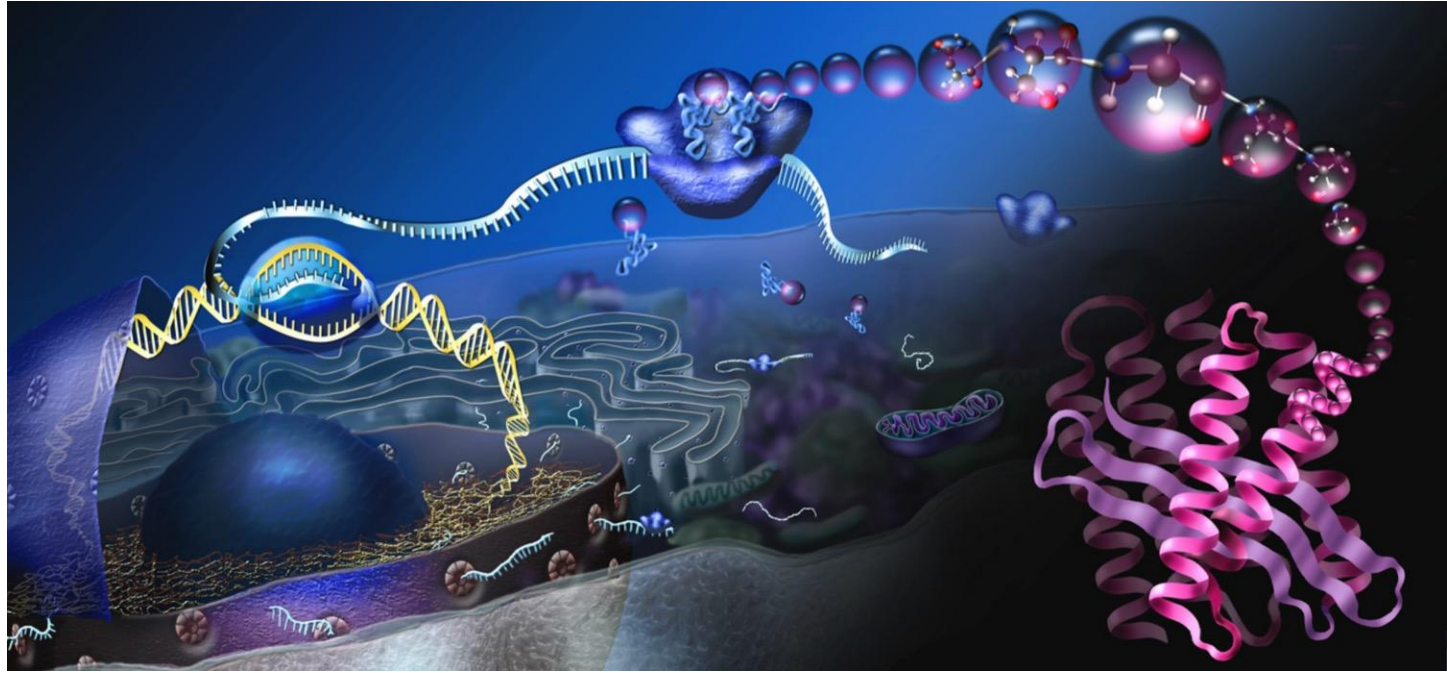


Spliceosomes



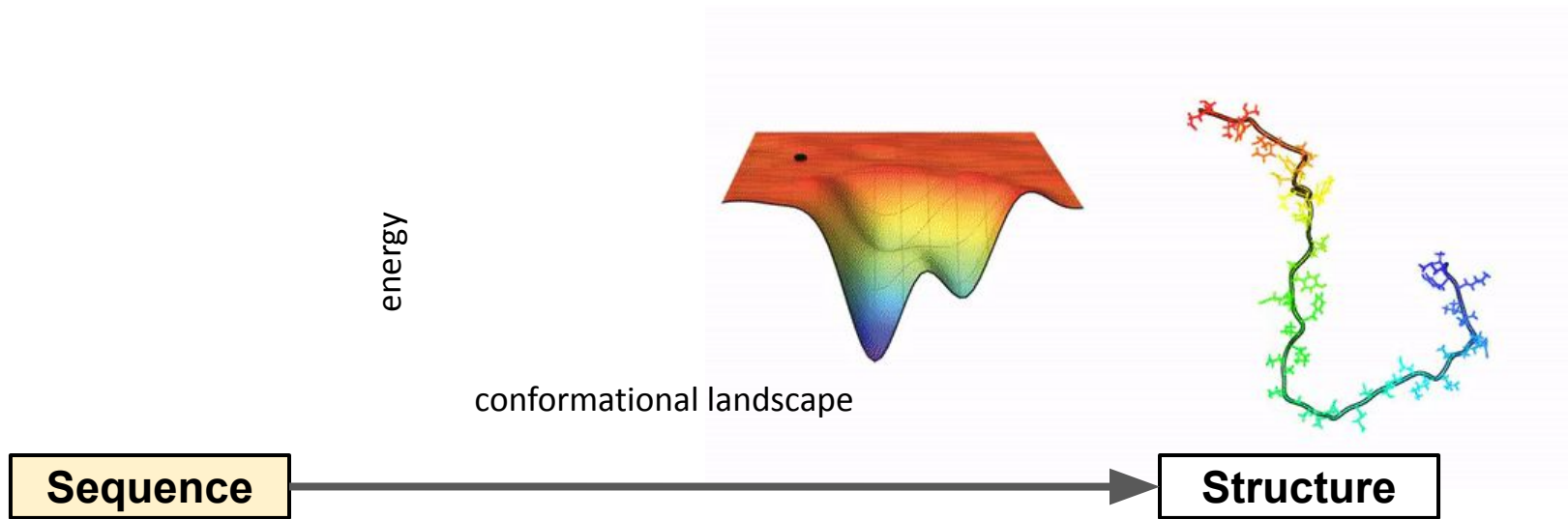
Chaperones

Protein Structure by AlphaFold



Protein folding problem:

For a given **sequence**, find **structure** with lowest free energy



Anfinsen's dogma

1. Uniqueness (single global minimum)
2. Stability (funnel)
3. Kinetic accessibility (path is smooth, no “knots” etc.)

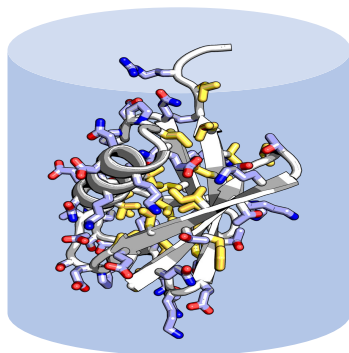
[Video credit: C. Fennell]

Dill, K.A. and MacCallum, J.L., 2012. The protein-folding problem, 50 years on. *science*, 338(6110), pp.1042-1046.

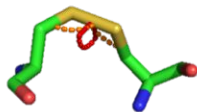
Typical Energy Function

Physics-Based and Knowledge-Based Potentials to Build the Energy Function

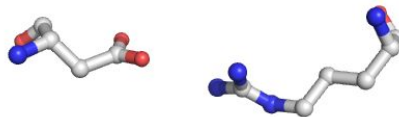
Solvation



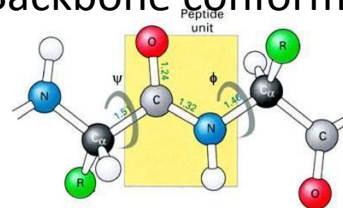
Disulfide bond



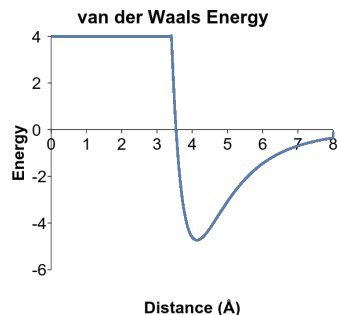
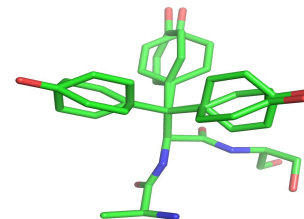
Ionic bond



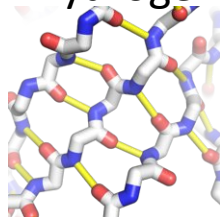
Backbone conformations



Sidechain conformations

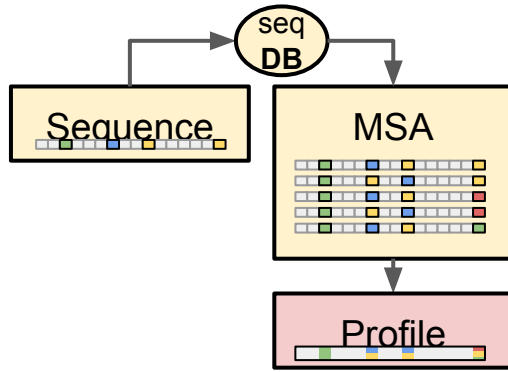


Hydrogen bond



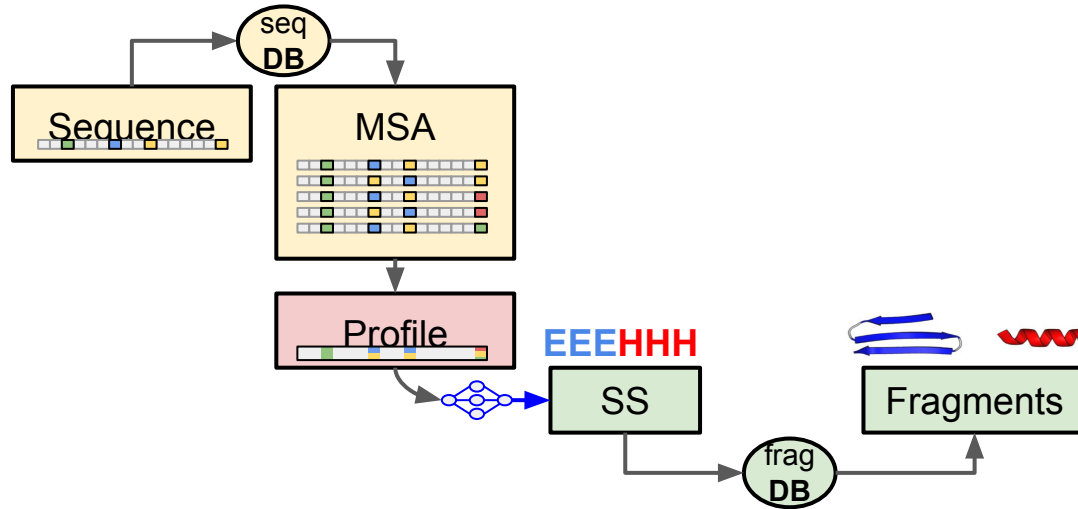
Typical structure prediction pipeline

MSA = multiple sequence alignment
DB = database
Profile = conservation

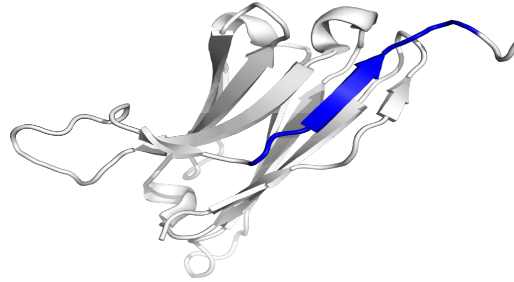


Typical structure prediction pipeline

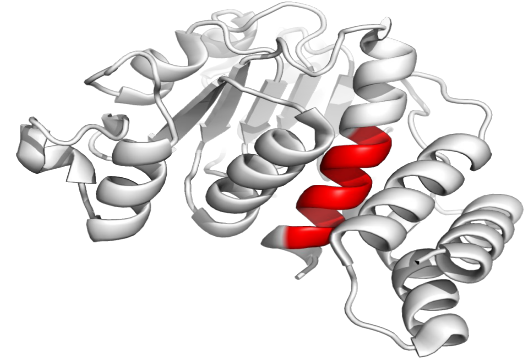
MSA = multiple sequence alignment
DB = database
Profile = conservation
SS = secondary structure prediction



Same sequence, β sheet or α helix, depending on context



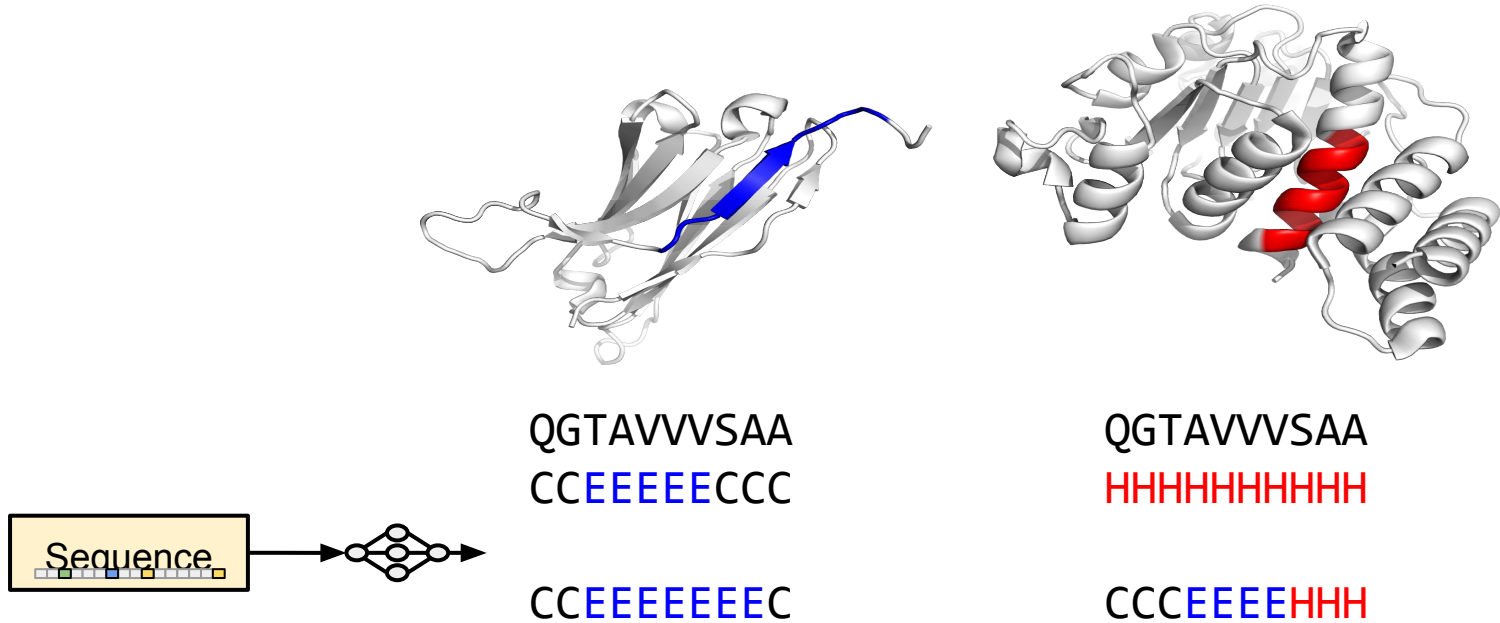
QGTAVVVSAA
CCEEEEECCC



QGTAVVVSAA
HHHHHHHHHH

Can we add evolutionary information?

Secondary structure prediction fails



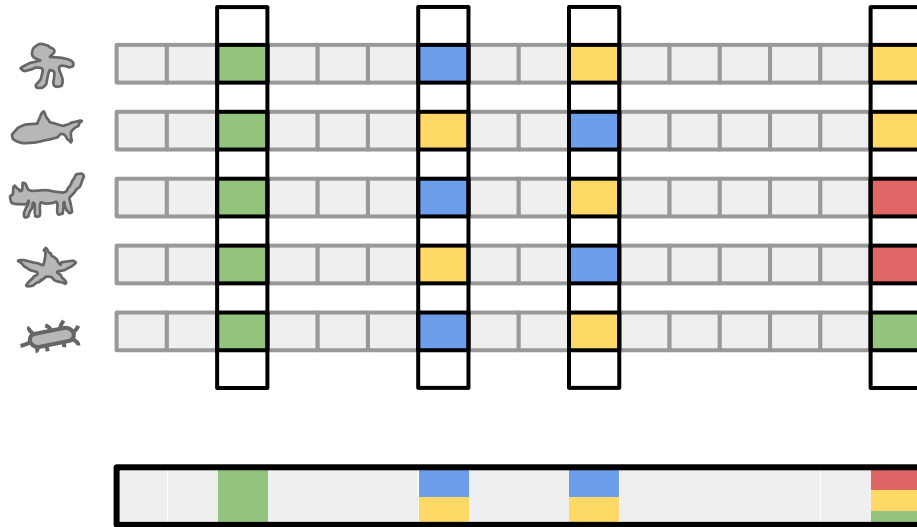
Use evolutionary information!



Search against a database of sequences

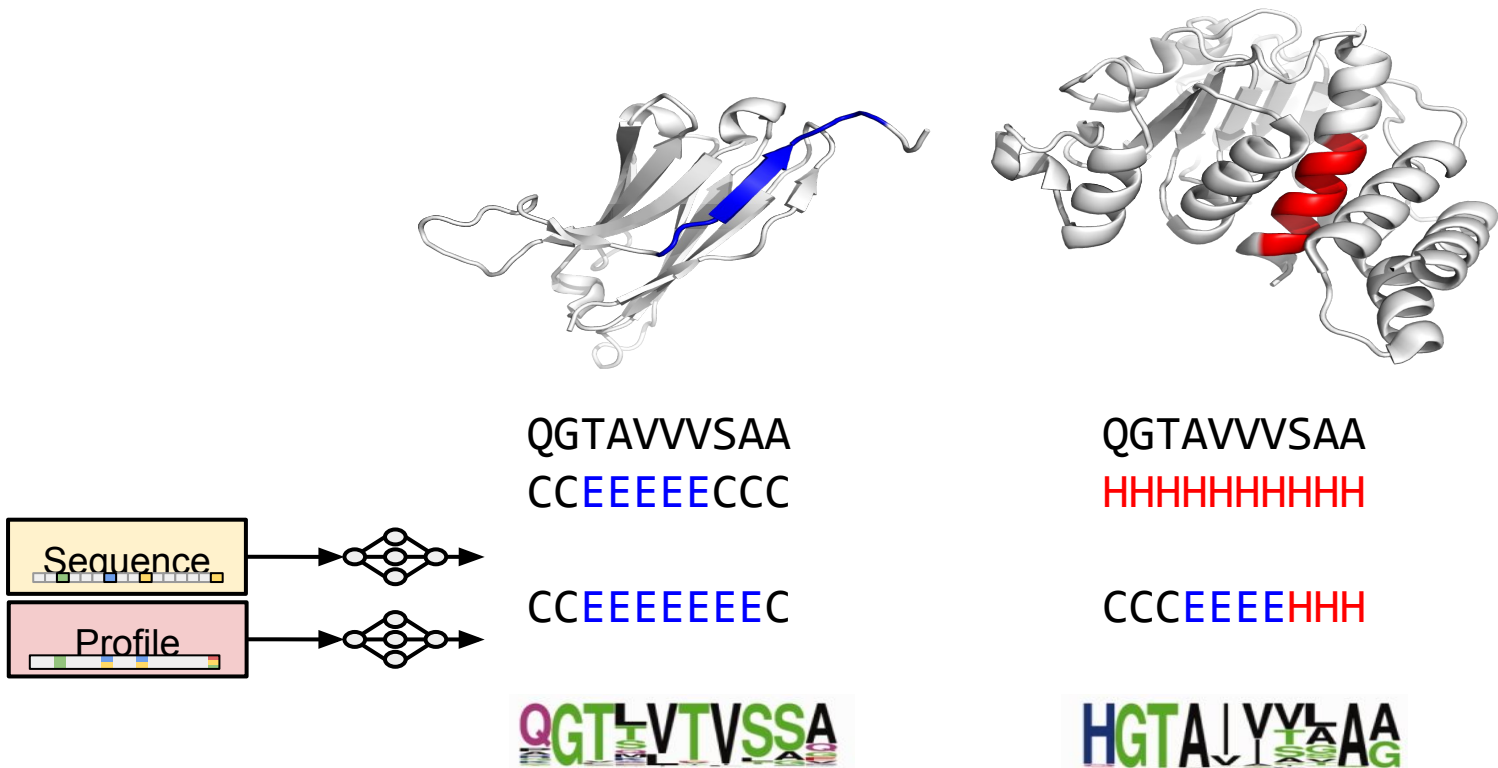


Analyze the MSA for conservation



Profile = Position-specific-scoring matrix

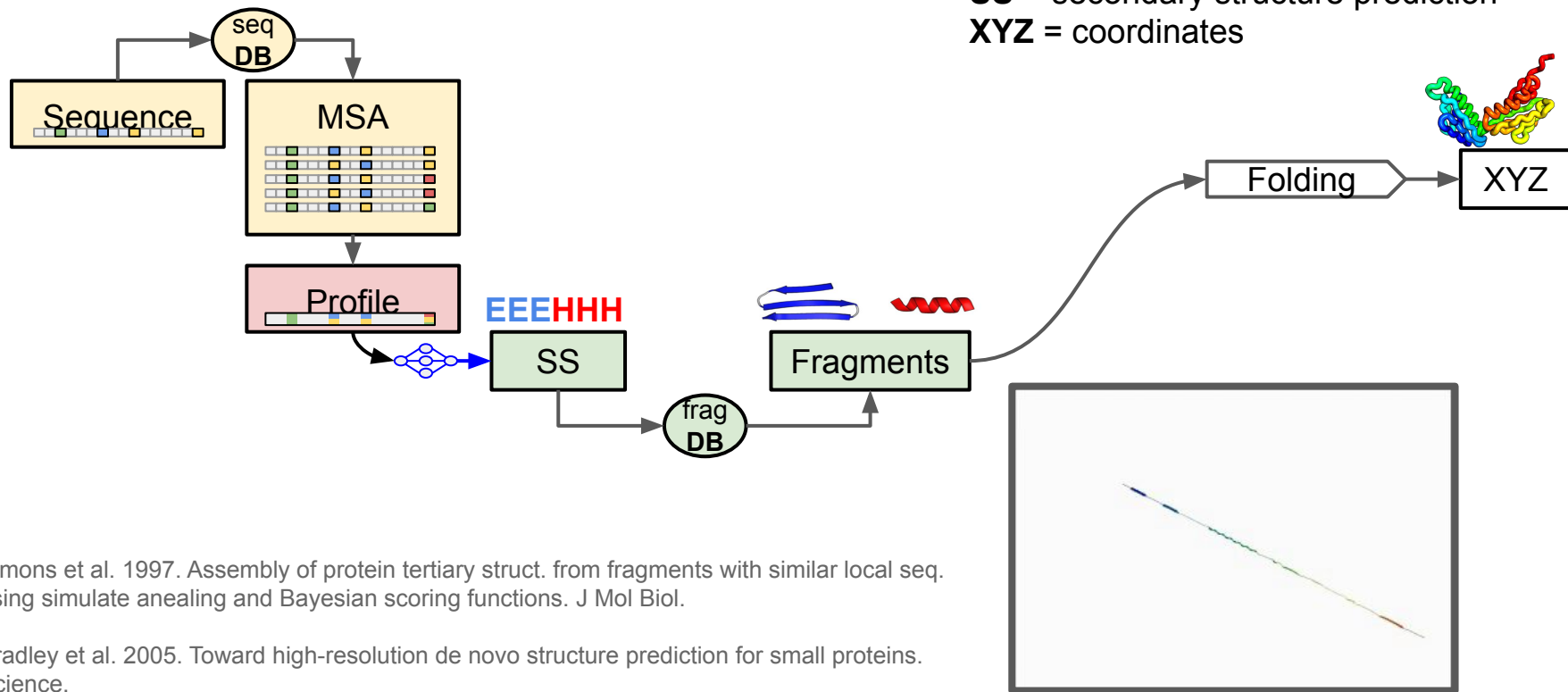
profile info improves secondary structure prediction



example from:

Li et al. 2015. ChSeq: A database of chameleon sequences. Protein Science

Typical structure prediction pipeline



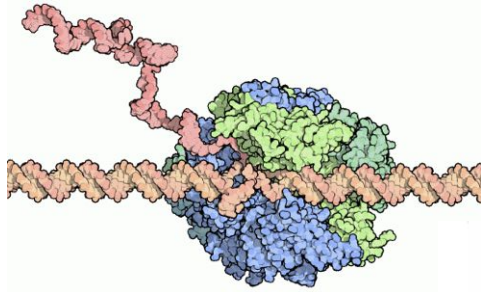
MSA = multiple sequence alignment
DB = database
Profile = conservation
SS = secondary structure prediction
XYZ = coordinates

Simons et al. 1997. Assembly of protein tertiary struct. from fragments with similar local seq. using simulate annealing and Bayesian scoring functions. J Mol Biol.

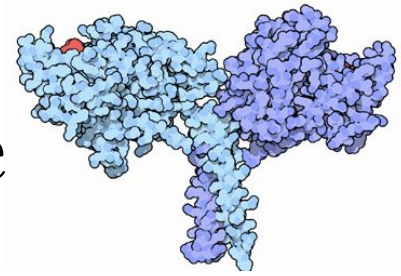
Bradley et al. 2005. Toward high-resolution de novo structure prediction for small proteins. Science.

Raman et al. 2009. Structure prediction for CASP8 with all-atom refinement using Rosetta.

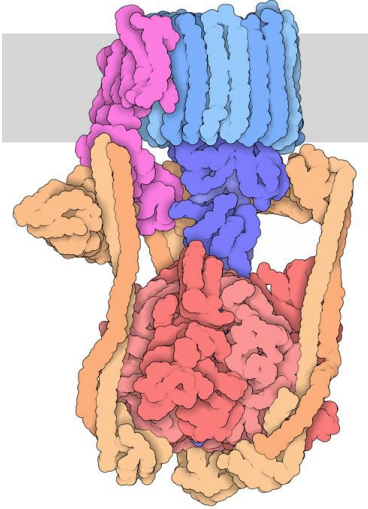
Molecules of life



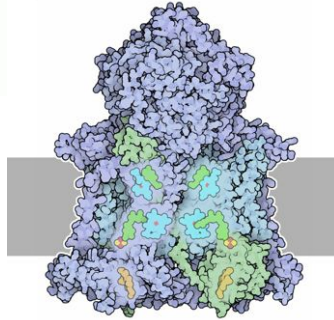
RNA polymerase



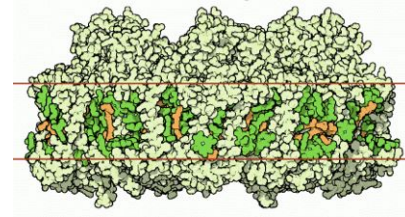
Kinesin



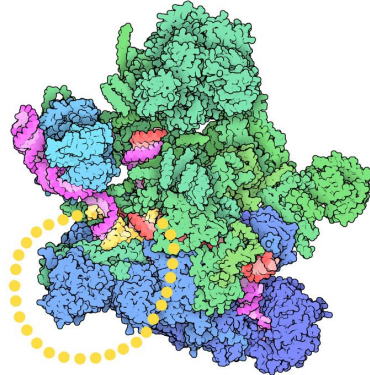
ATP synthase



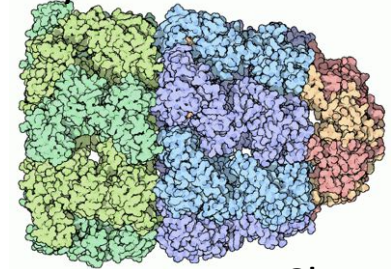
Cytochrome bc1



Photosystem I



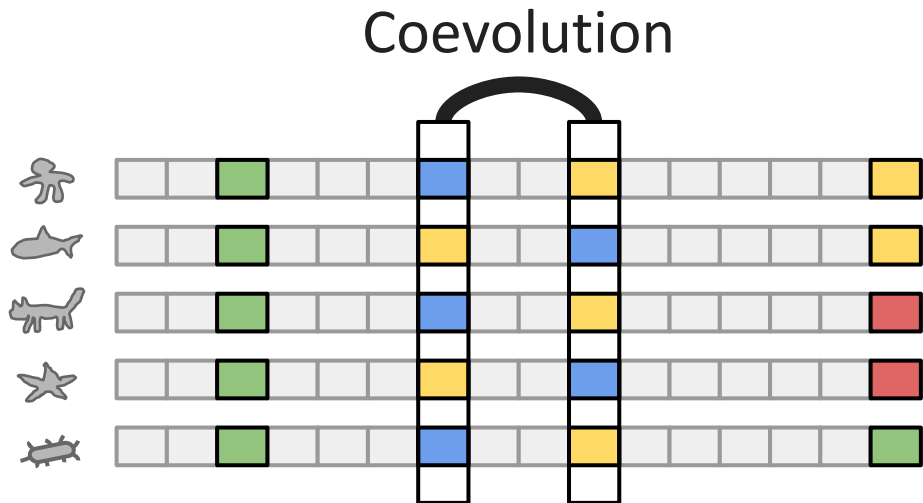
Spliceosomes



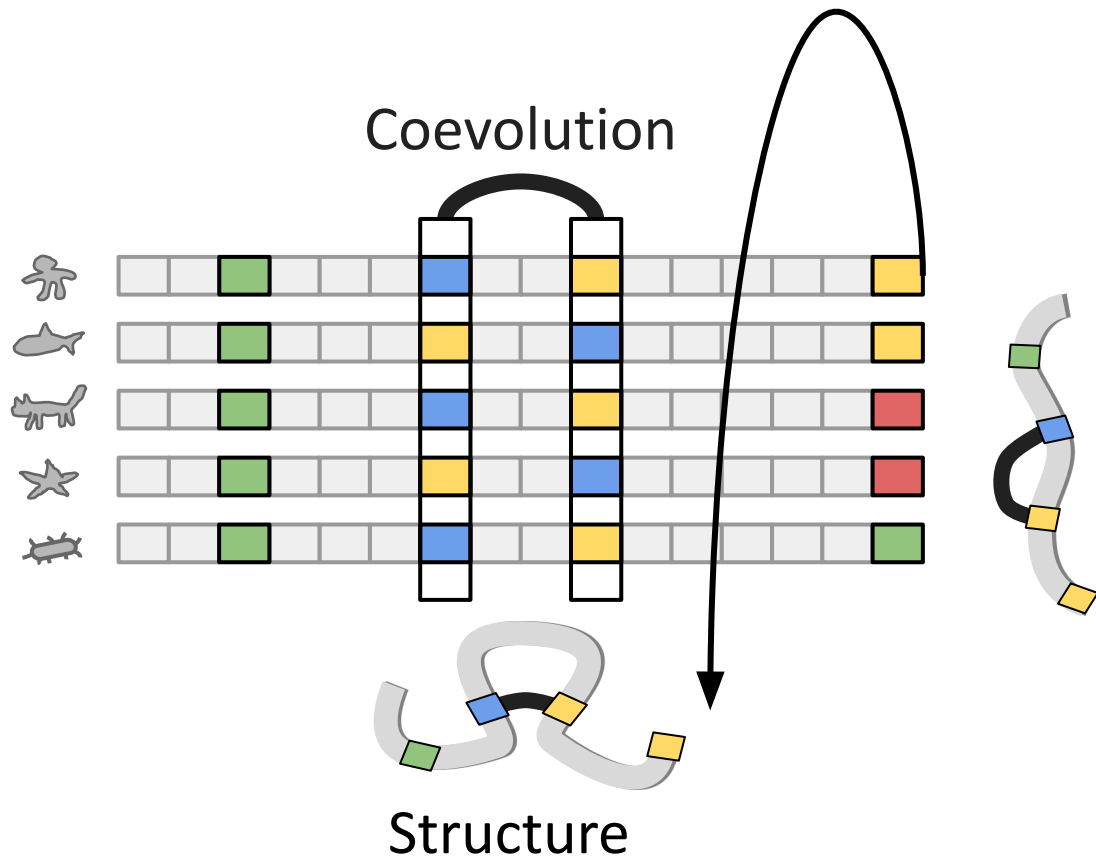
Chaperones

16

Analyze the MSA for coevolution

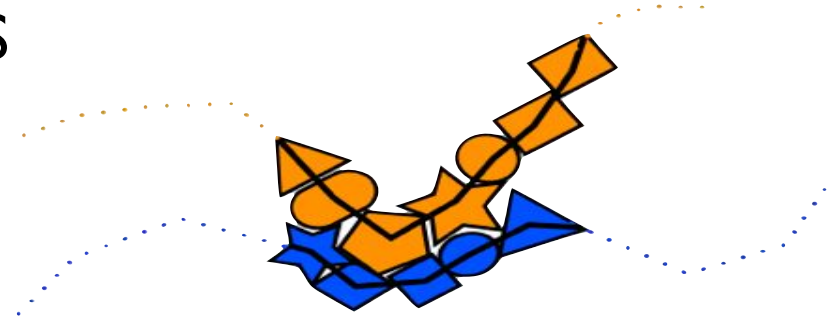


Use the as restraints in folding simulations!

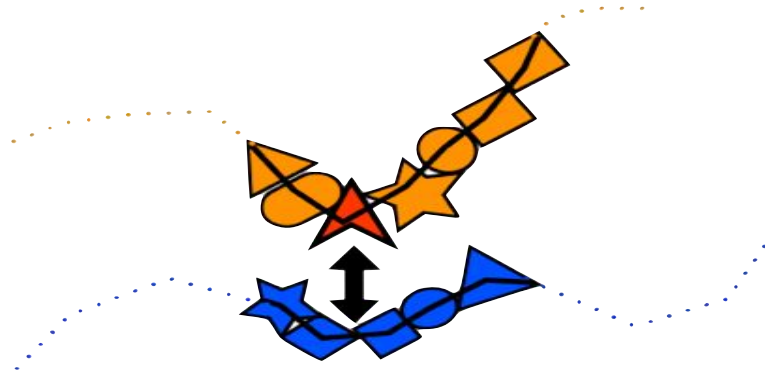


SPATIAL PROXIMITY INDUCES SEQUENCE COUPLING

Native interactions

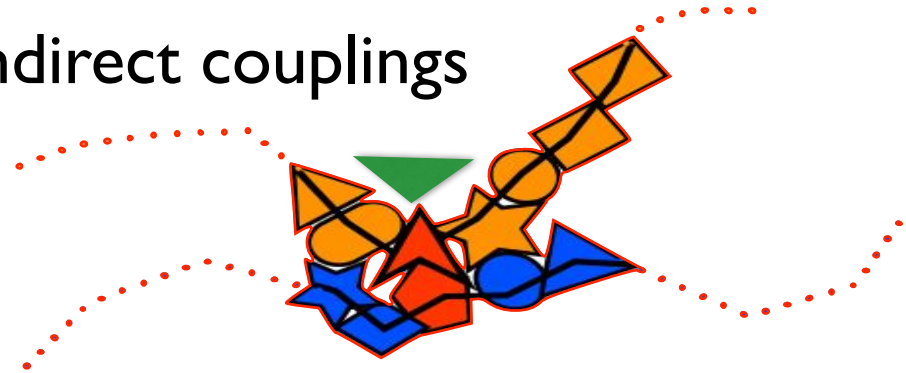


Unfavourable mutation



Compensating mutation

Indirect couplings



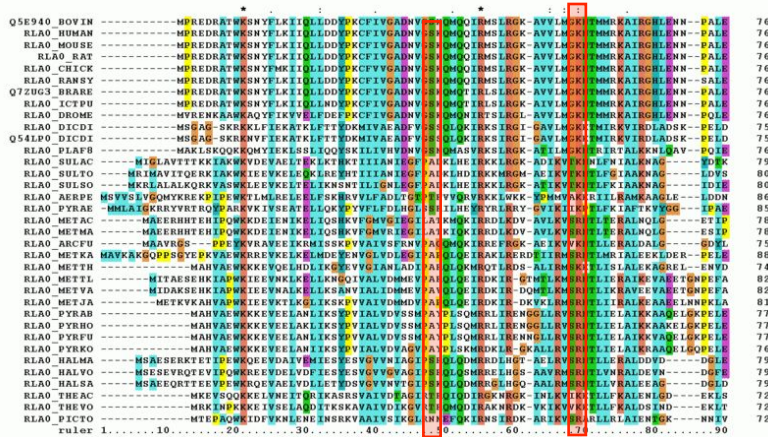
BASIS FOR CONTACT PREDICTION

Q5E940_BOVIN	-----MPREDRATWKSNYFLKIIQLDDYKCFIVGADNVGSKOMQIRMSLRGK-AVVLGKKTMMRKAIRGHLENN--PALE	76
RLA0_HUMAN	-----MPREDRATWKSNYFLKIIQLDDYKCFIVGADNVGSKOMQIRMSLRGK-AVVLGKKTMMRKAIRGHLENN--PALE	76
RLA0_MOUSE	-----MPREDRATWKSNYFLKIIQLDDYKCFIVGADNVGSKOMQIRMSLRGK-AVVLGKKTMMRKAIRGHLENN--PALE	76
RLA0_RAT	-----MPREDRATWKSNYFLKIIQLDDYKCFIVGADNVGSKOMQIRMSLRGK-AVVLGKKTMMRKAIRGHLENN--PALE	76
RLA0_CHICK	-----MPREDRATWKSNYFLKIIQLDDYKCFIVGADNVGSKOMQIRMSLRGK-AVVLGKKTMMRKAIRGHLENN--PALE	76
RLA0_RANSY	-----MPREDRATWKSNYFLKIIQLDDYKCFIVGADNVGSKOMQIRMSLRGK-AVVLGKKTMMRKAIRGHLENN--PALE	76
Q7ZUG3_BRARE	-----MPREDRATWKSNYFLKIIQLDDYKCFIVGADNVGSKOMQIRMSLRGK-AVVLGKKTMMRKAIRGHLENN--PALE	76
RLA0 ICTPU	-----MPREDRATWKSNYFLKIIQLDDYKCFIVGADNVGSKOMQIRMSLRGK-AVVLGKKTMMRKAIRGHLENN--PALE	76
RLA0_DROME	-----MVRENKAQYFIRKVVLFDEFKCFIVGADNVGSKOMQIRMSLRGK-AVVLGKKTMMRKAIRGHLENN--PALE	76
RLA0_DICDI	-----MSGAG-SKRKKLFIKATKLFITTDKMIKAEADVFGS-SOLQIRKKSIRGI-GAVLMGKKTMIKRVIRDLDADSK--PELD	75
Q54LP0_DICDI	-----MSGAG-SKRKNVFIKATKLFITTDKMIKAEADVFGS-SOLQIRKKSIRGI-GAVLMGKKTMIKRVIRDLDADSK--PELD	75
RLA0_PLAF8	-----MAKLSKQKKQMYIEKLSLIIQYKSLIVHVDNVGSKOMQIRMSLRGK-AVVLGKKTMMRKAIRGHLENN--PALE	76
RLA0_SULAC	-----MIGLAVTTTKKIAKWKVDEVAELTEKLTKHTIIIANIEGFPADKLIHIRKKLRGK-ADIKVTKINLFNIALKNAG----DYK	79
RLA0_SULTO	-----MRIMAVITQERKIAKWKIEEVKELEKLRKHTIIIANIEGFPADKLIHIRKKLRGK-ADIKVTKINLFNIALKNAG----DYK	80
RLA0_SULSO	-----MKRLALALKQKQKVASWKELEVKELEKLRKHTIIIANIEGFPADKLIHIRKKLRGK-ADIKVTKINLFNIALKNAG----DYK	80
RLA0_AERPE	MSVVSILGQMYKREKPIPEWKTLMLELELFSKHRVVLFDLTCPTFFVYQVRKKLWKK-YPMVVAKRITLRAKKAAGLE--LDDN	86
RLA0_PYRAE	MMLAIGKRRYVRTQYQARKYKIVSEATELLQKQYVYVFLDLHLGSLRIIHHYRRLRY-GVIKIKIPFLFKIAFTKVYGG--IPAE	85
RLA0_METAC	MAEERHHTETIPQWKDEIENIKELIQSHKVFQGMVIEGILATKQIRRDLDKV-AVLKVSRTTLERALNQLG----ETIP	78
RLA0_METMA	MAEERHHTETIPQWKDEIENIKELIQSHKVFQGMVIEGILATKQIRRDLDKV-AVLKVSRTTLERALNQLG----ESIP	78
RLA0_ARCFU	MAAVRGS--PDEYKVRAVEEIKRMISSEKVVVAIVSFERNYPAGOMQIRREFRGK-AEIKVTKITLLEALDAG----GDYL	75
RLA0_METKA	MAVKAKGQPPSGYEPKVAEWRKREVVELKELMDEYENVGLVDLEGIPAPOLQIRAKLRERDITIRMRRTILMRITALEEKLDER--PELE	88
RLA0_METTH	MAHVAEWKKKEVQELHDLIKGEYVVGIAWADIPAROLQIRMQTLRDS-ALIRMKRTILISLAKELAGREL--ENVV	74
RLA0_METTL	MITAESEHKIAPWKIEEVNKLKELKNGQIVALVDMMEVPAROLQIRDKIR-GTMTLKMSRTTLERAIKEVAETGNPEFA	82
RLA0_METVA	MIDAKSEHKIAPWKIEEVNKLKELKNSANVIALIDMMEVPAROLQIRDKIR-DQMTLKMSRTTLERAIKEVAETGNPEFA	82
RLA0_METJA	METKVAHVAPWKIEEVNKLKELKNSANVIALIDMMEVPAROLQIRDKIR-DKVKLRMSRTTLERAIKEVAETGNPEFA	81
RLA0_PYRAB	MAHVAEWKKKEVEELANLKSYPVIALVDVSSMPAYPISQMRRLIRENGLLRVSRITLIELAIKKAAGELGKPELE	77
RLA0_PYRHO	MAHVAEWKKKEVEELANLKSYPVIALVDVSSMPAYPISQMRRLIRENGLLRVSRITLIELAIKKAAGELGKPELE	77
RLA0_PYRFU	MAHVAEWKKKEVEELANLKSYPVIALVDVSSMPAYPISQMRRLIRENGLLRVSRITLIELAIKKAAGELGKPELE	77
RLA0_PYRKO	MAHVAEWKKKEVEELANLKSYPVIALVDVAGVPAYPISQMRRLIRENGLLRVSRITLIELAIKKAAGELGKPELE	76
RLA0_HALMA	MSAESERKTETIPWKQEEVDIVMIESYESYGVVNIACIPEROLQIRMRDLHGT-AELRVSRITTLERALDDVD----DGLF	79
RLA0_HALVO	MSSEYVRQETETIPWKQEEVDIVMIESYESYGVVNIACIPEROLQIRMRDLHGT-AELRVSRITTLERALDDVD----DGLF	79
RLA0_HALSA	MSAEQRITETETIPWKQEEVDIVMIESYESYGVVNIACIPEROLQIRMRDLHGT-AELRVSRITTLERALDDVD----DGLF	79
RLA0_THEAC	MKEYSQKKELVNEITQIRKASRSVAIVDAGIRIRQIRIRGKNRGK-INLVYIKKTLFLKALENLGD----EKLK	72
RLA0_THEVO	MRKNPKKKEIVSELAGDITKSKAVAIVDIKGVRRQIRIRKASRSVAIVDAGIRIRQIRIRGKNRGK-INLVYIKKTLFLKALENLGD----EKLK	72
RLA0_PICTO	MTEPAQWKIDFYKNLENEINSRKYAAIVSIKGLRNNETQIRIRNSIRDK-ARIKVSARLLRLAIENTGK----NNIV	72
ruler	1.....10.....20.....30.....40.....50.....60.....70.....80.....90	

Residues in contact tend to co-evolve

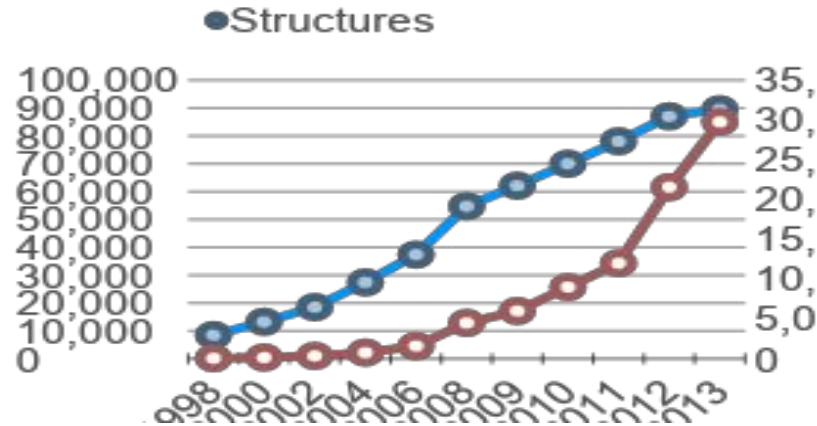
Reasons for Advances in Ab-Initio Protein Folding

CONTACT PREDICTION



Residues in contact
tend to co-evolve

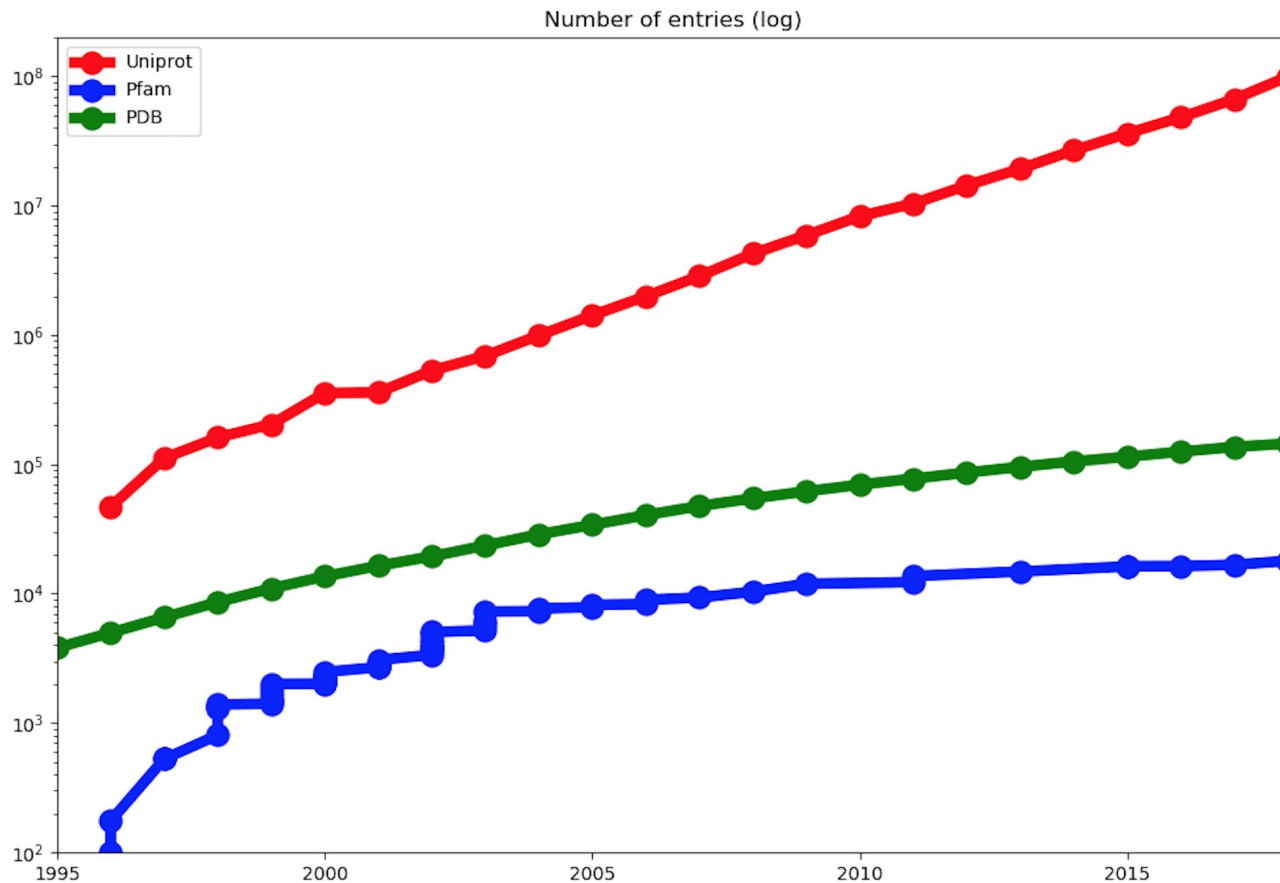
Sequence Growth



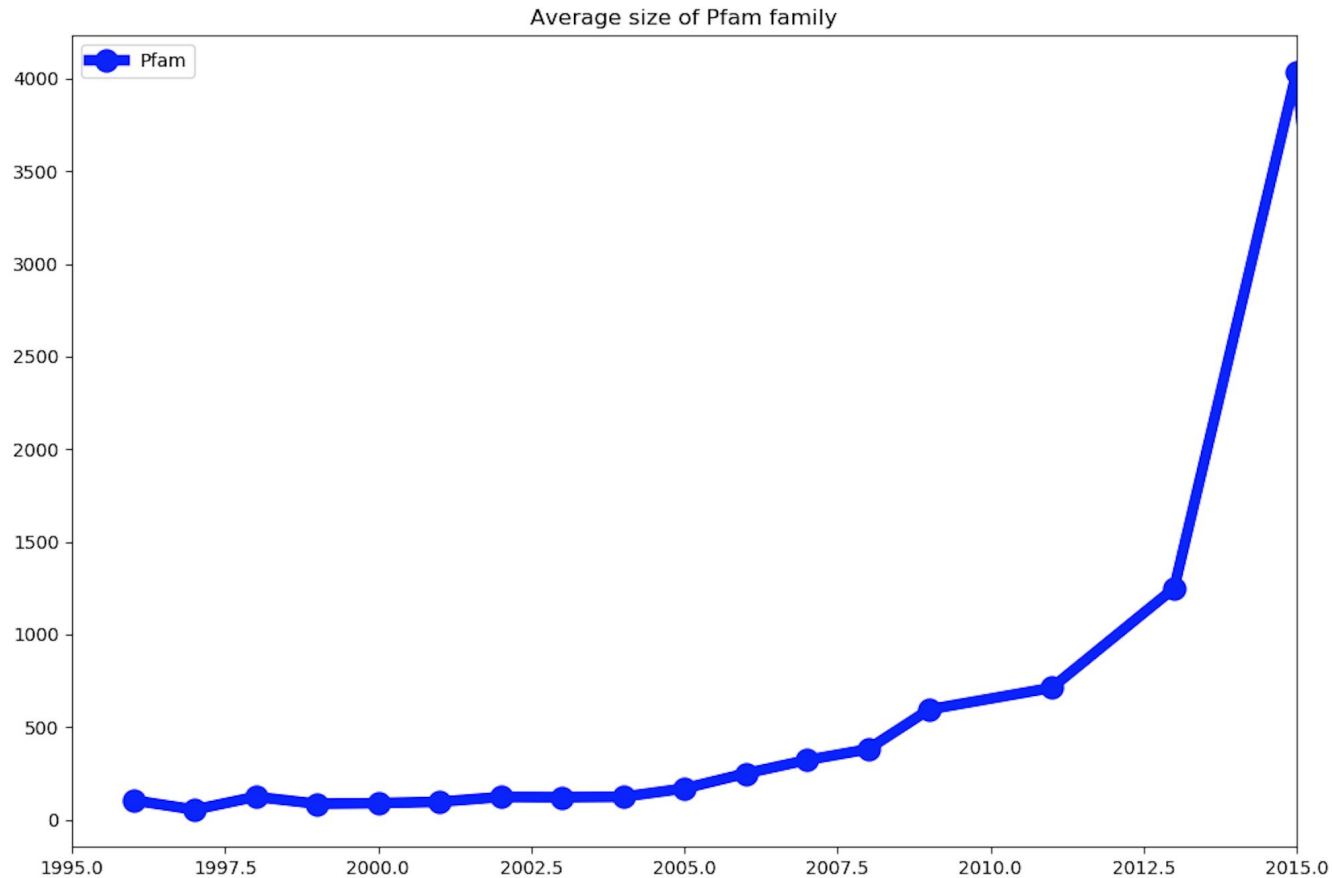
Decoupling direct interactions
from indirect ones

Giraud, Phys Rev E Stat 1999

Much faster increase in sequences.

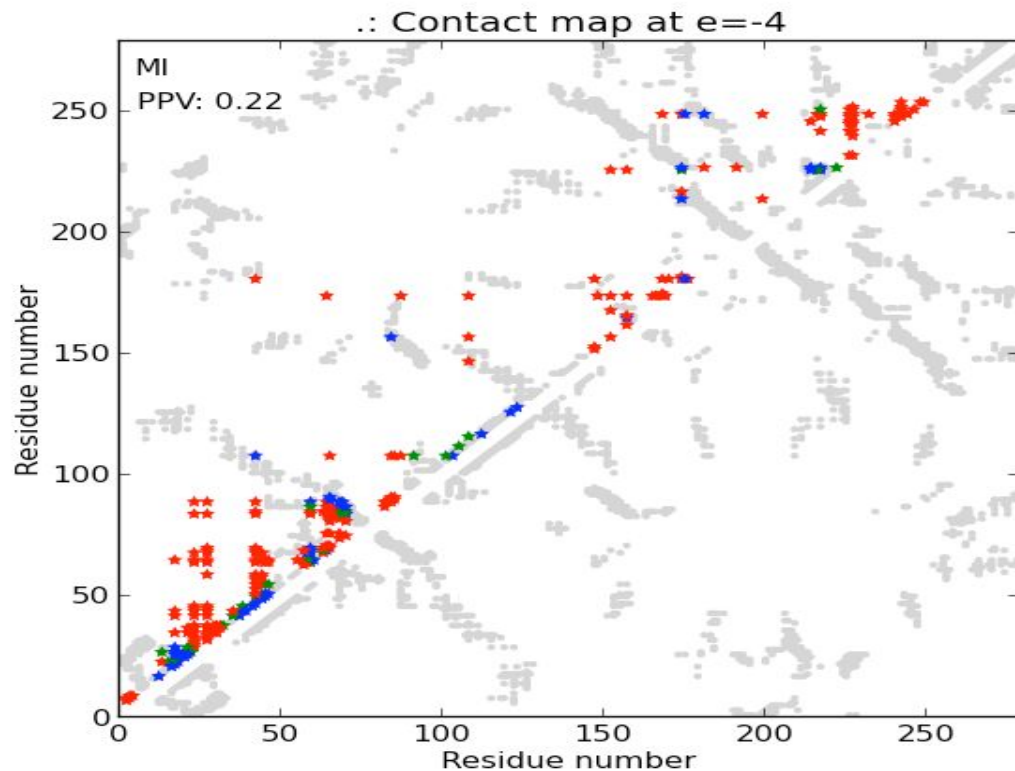


Sequence families are getting bigger and bigger



Direct information contact
prediction methods.

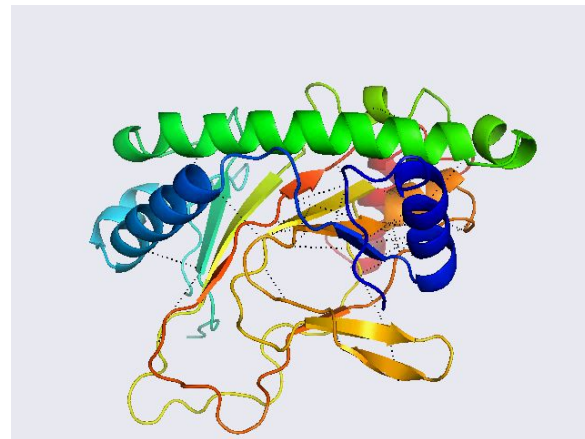
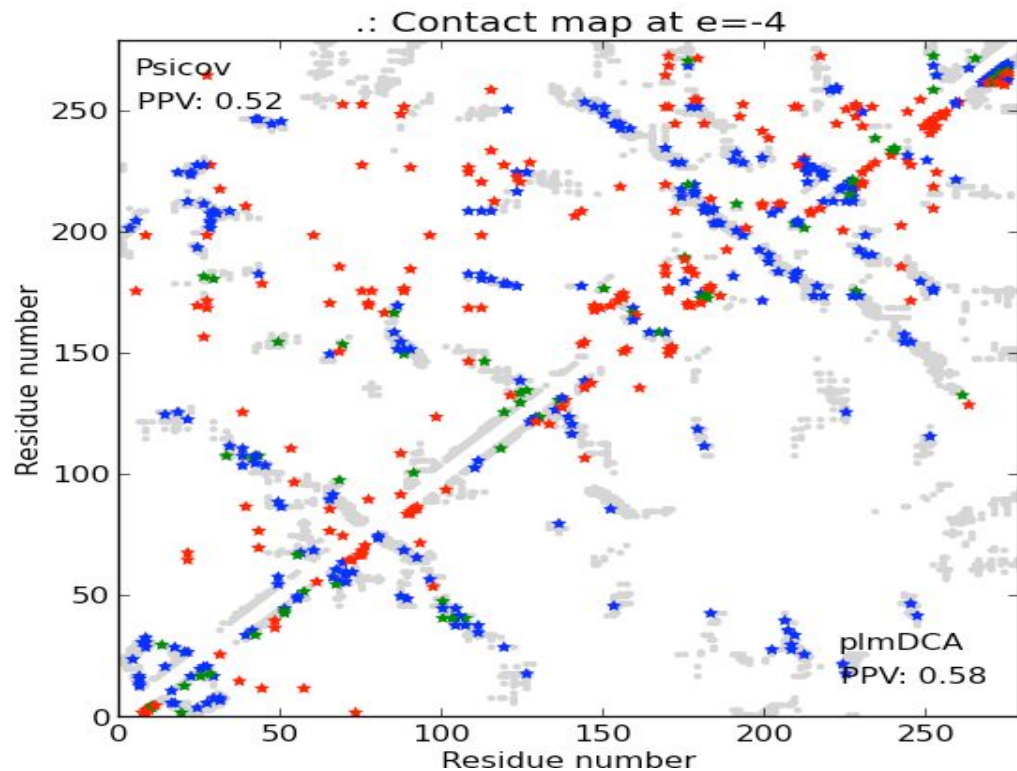
MUTUAL INFORMATION



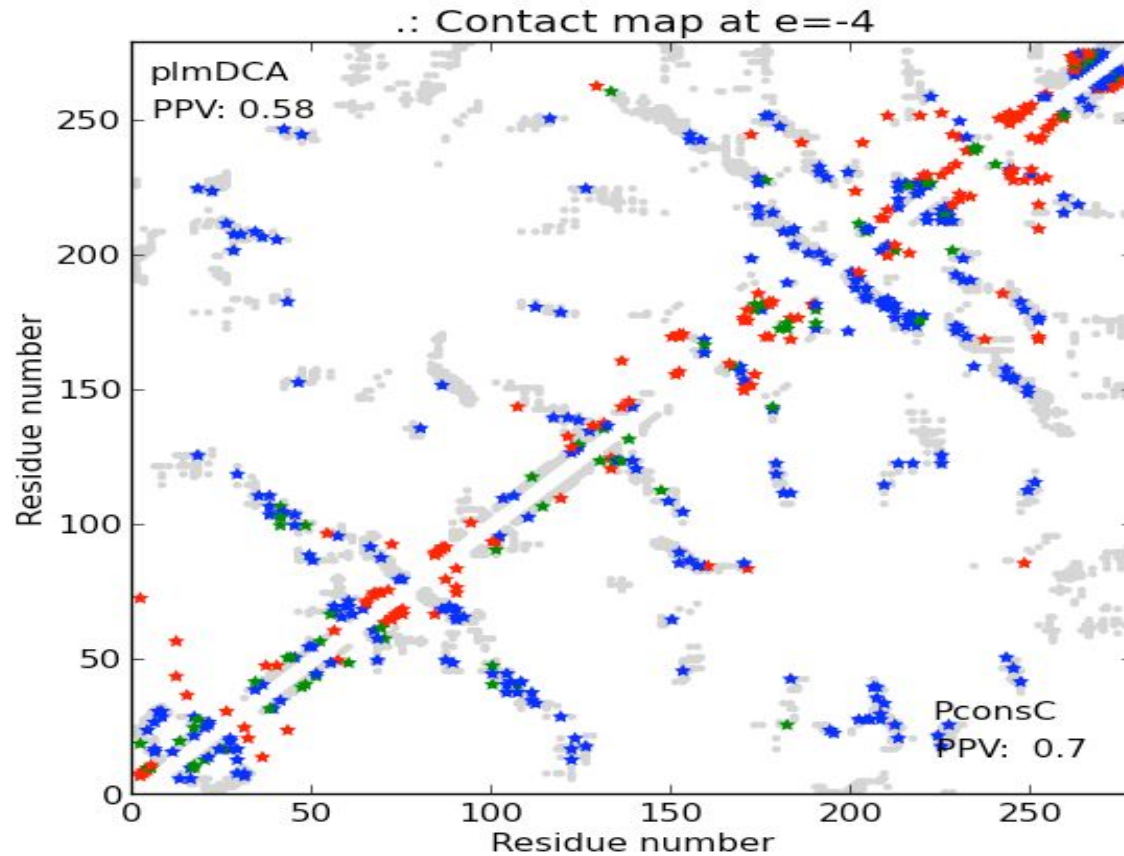
Fodor, A.A. and Aldrich, R.W.

Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. (2004).

PLMDCA

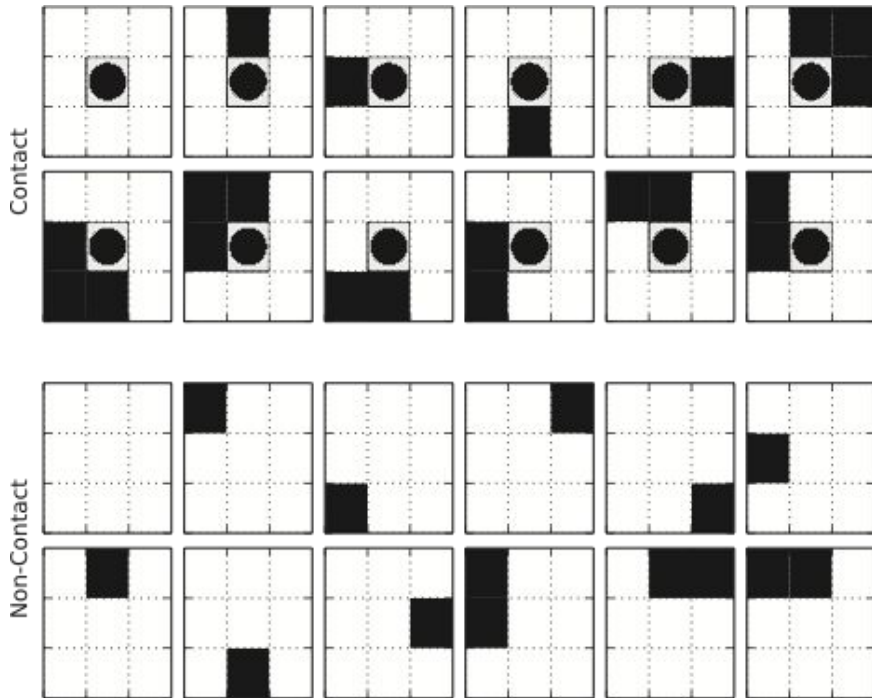


Contacts in contact maps are not randomly distributed



Most frequent 3x3 contact patterns

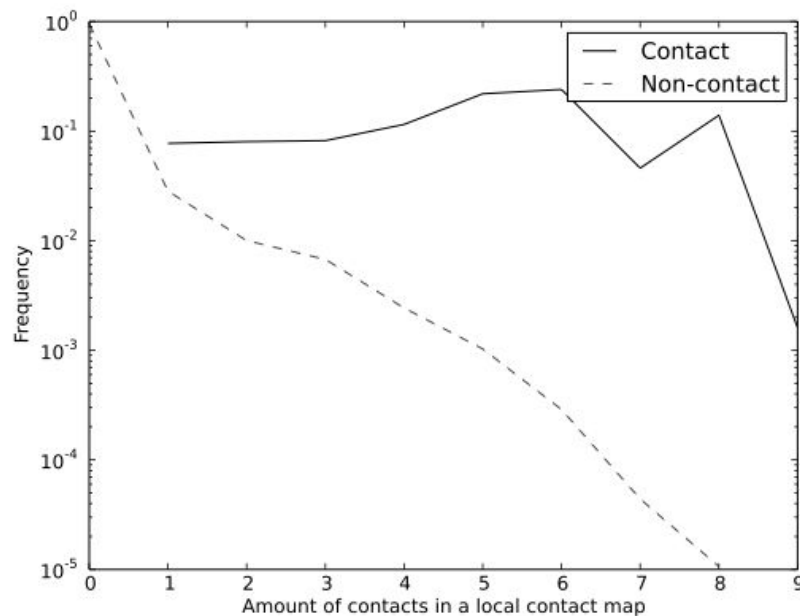
- Some patterns in a map are more frequent than others
- Some patterns are forbidden.
- Contact predictors should use this, by:
 - Specific rules
 - Machine Learning



Contacts and non-contacts

3x3 map

- Features of contact maps:
 - Contacts close to other contacts
 - Mainly empty
 - Visible diagonal patterns



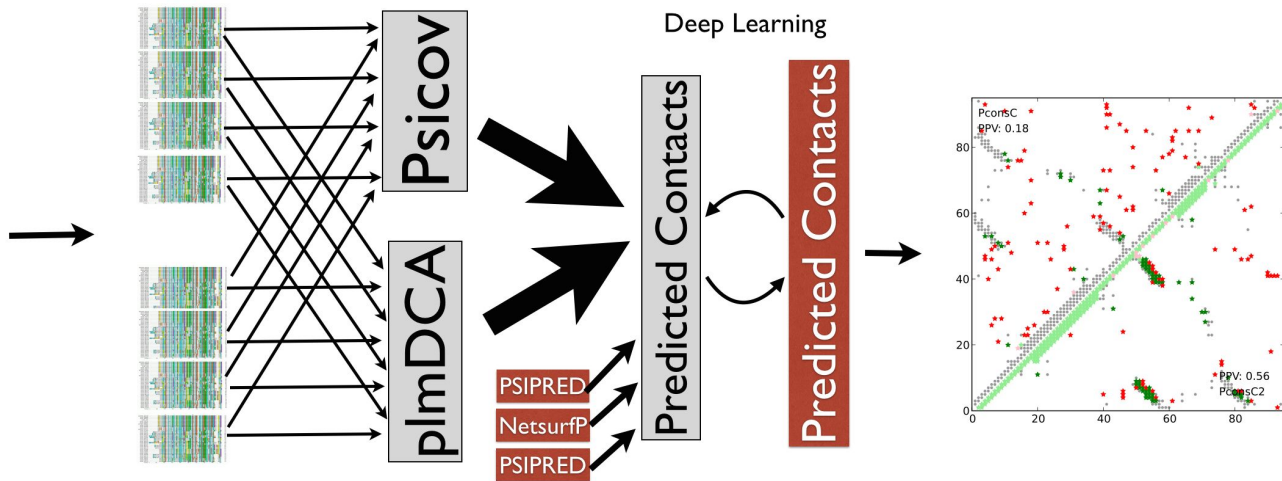
PconsC2 pipeline

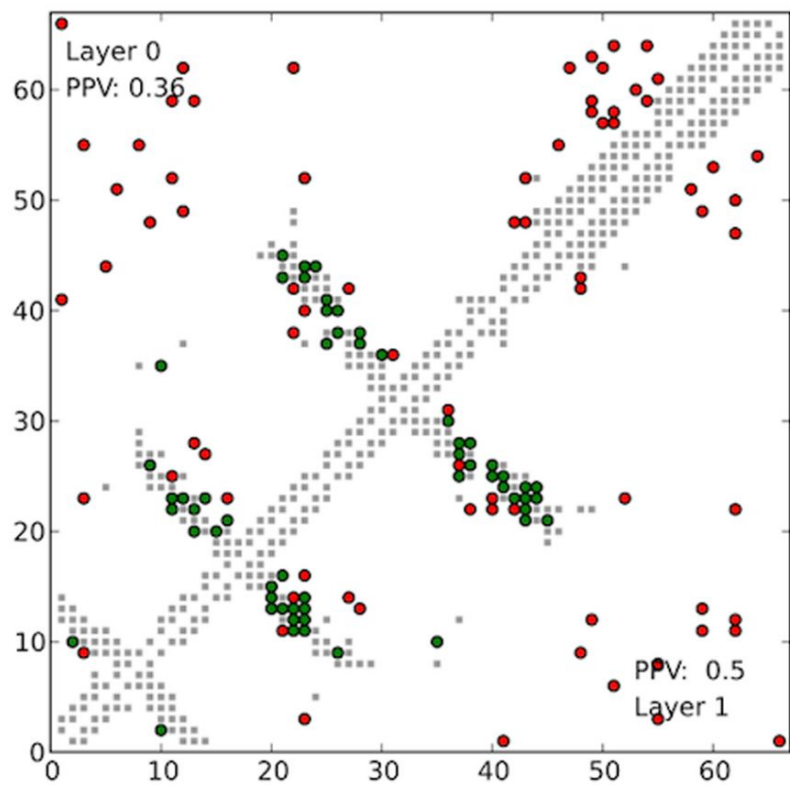
8 Multiple Sequence Alignments

16 Primary Contact Prediction

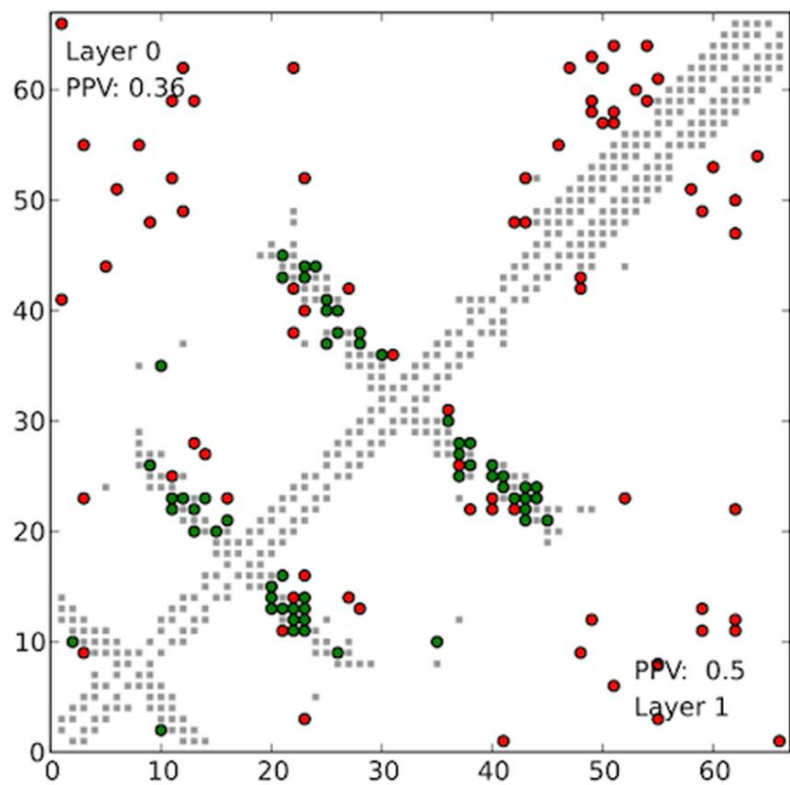
Final Contact Prediction

Sequence

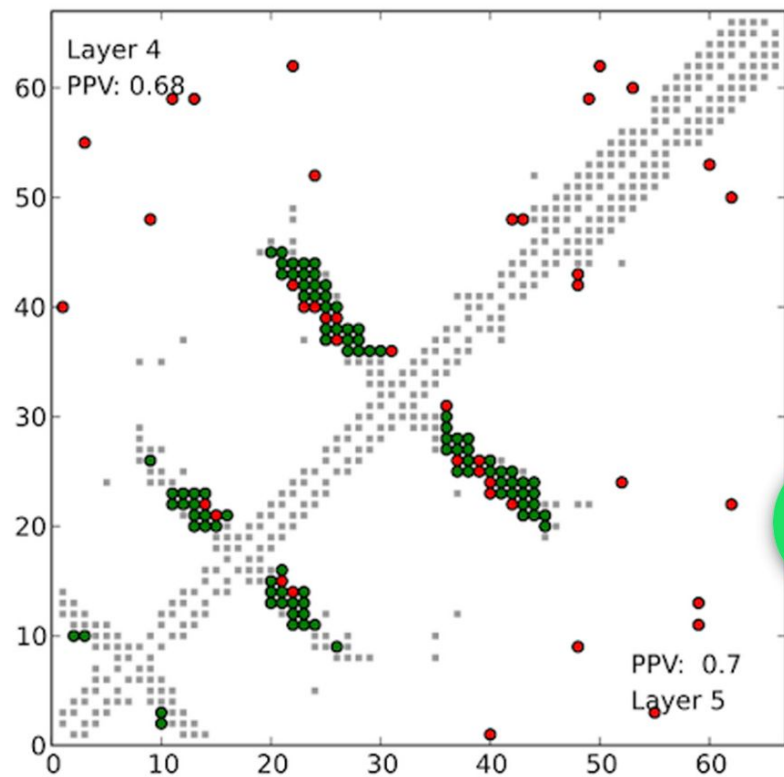




(b) 1pcf:A Layer 0-1



(b) 1pcf:A Layer 0-1



(d) 1pcf:A Layer 4-5

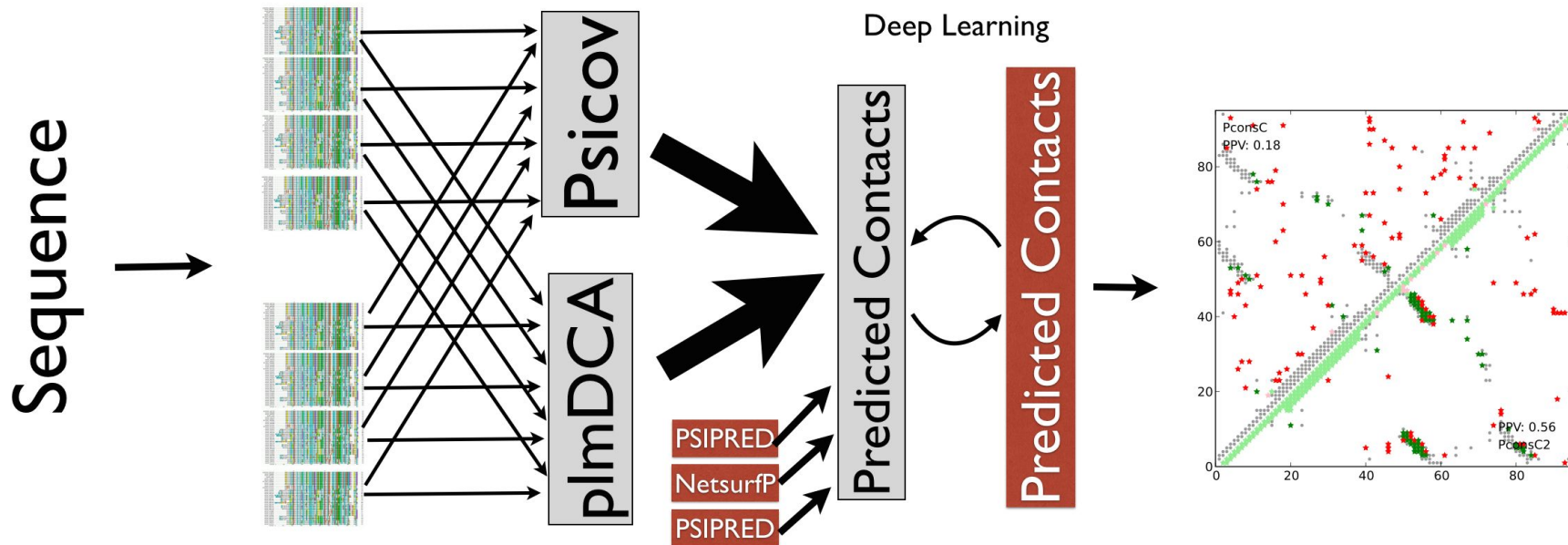


PconsC2 (2014)

8 Multiple Sequence Alignments

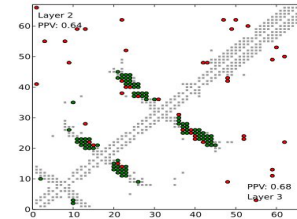
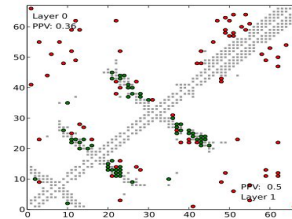
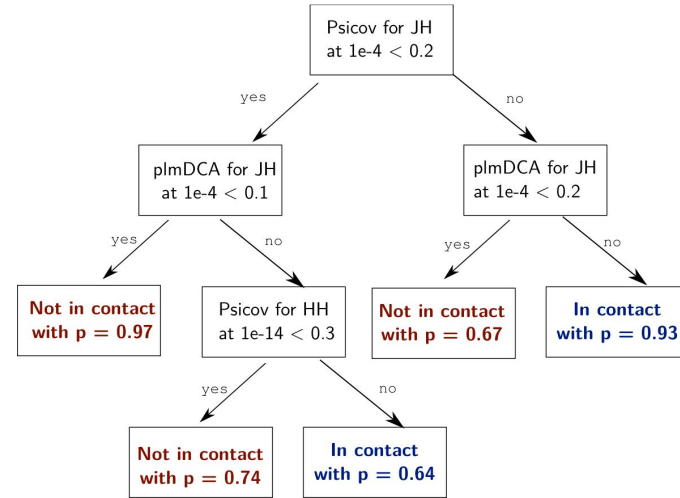
16 Primary Contact Prediction

Final Contact Prediction



Deep Learning in PconsC2

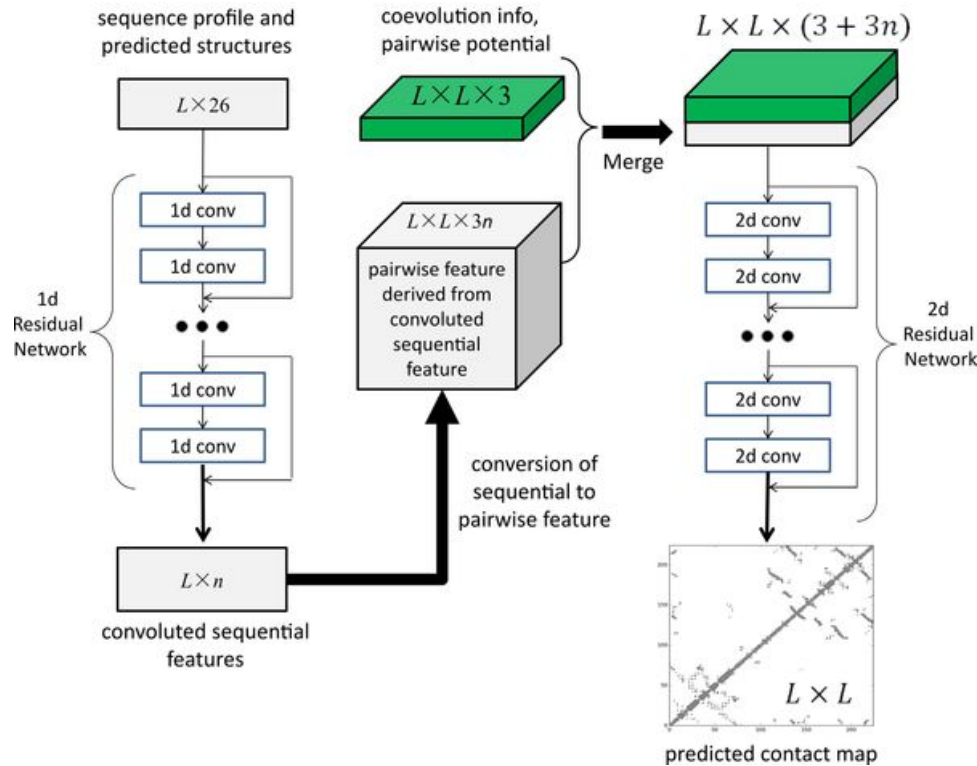
- Random forests with 100 trees and a minimum of 500 samples
- The same learner for all pairs
- Iterative refinement of map
- Inputs
 - Contact predictions
 - Additional features
 - **Receptive field 11x11**



Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model

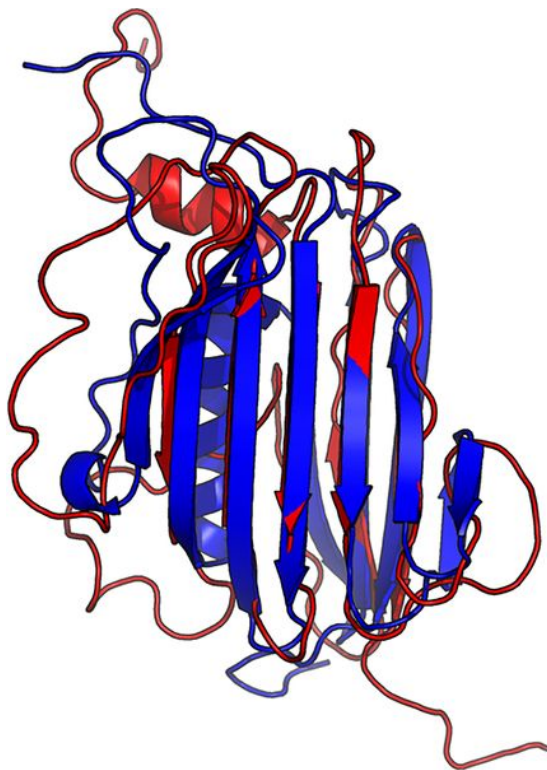
Sheng Wang , Siqi Sun , Zhen Li, Renyu Zhang, Jinbo Xu

Fig 1. Illustration of our deep learning model for contact prediction where L is the sequence length of one protein under prediction.



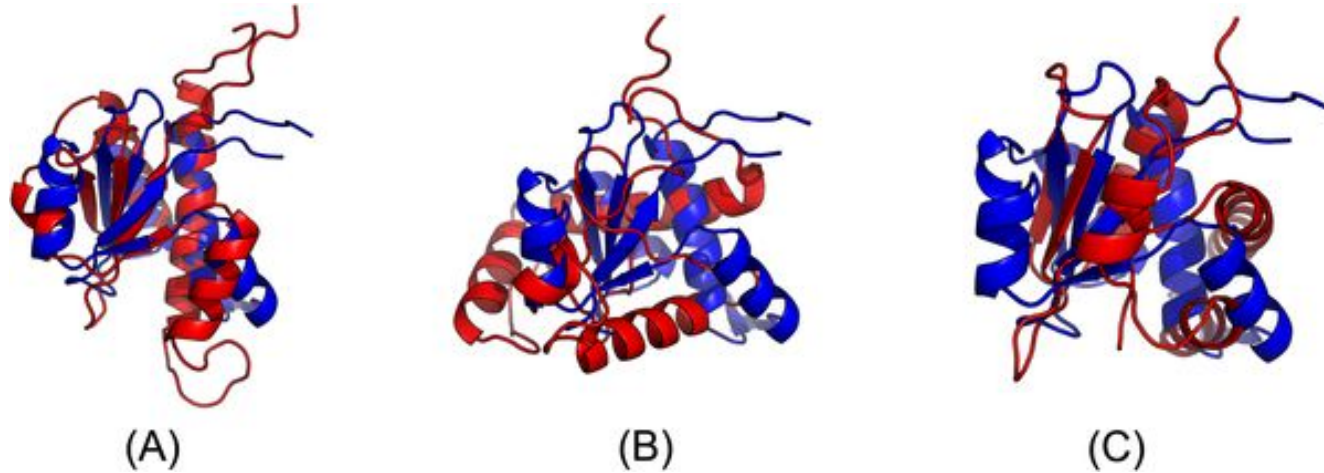
Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLOS Computational Biology 13(1): e1005324. <https://doi.org/10.1371/journal.pcbi.1005324>
<http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005324>

Fig 7. Superimposition between our predicted model (red) and its native structure (blue) for the CAMEO test protein (PDB ID 2nc8 and chain A).



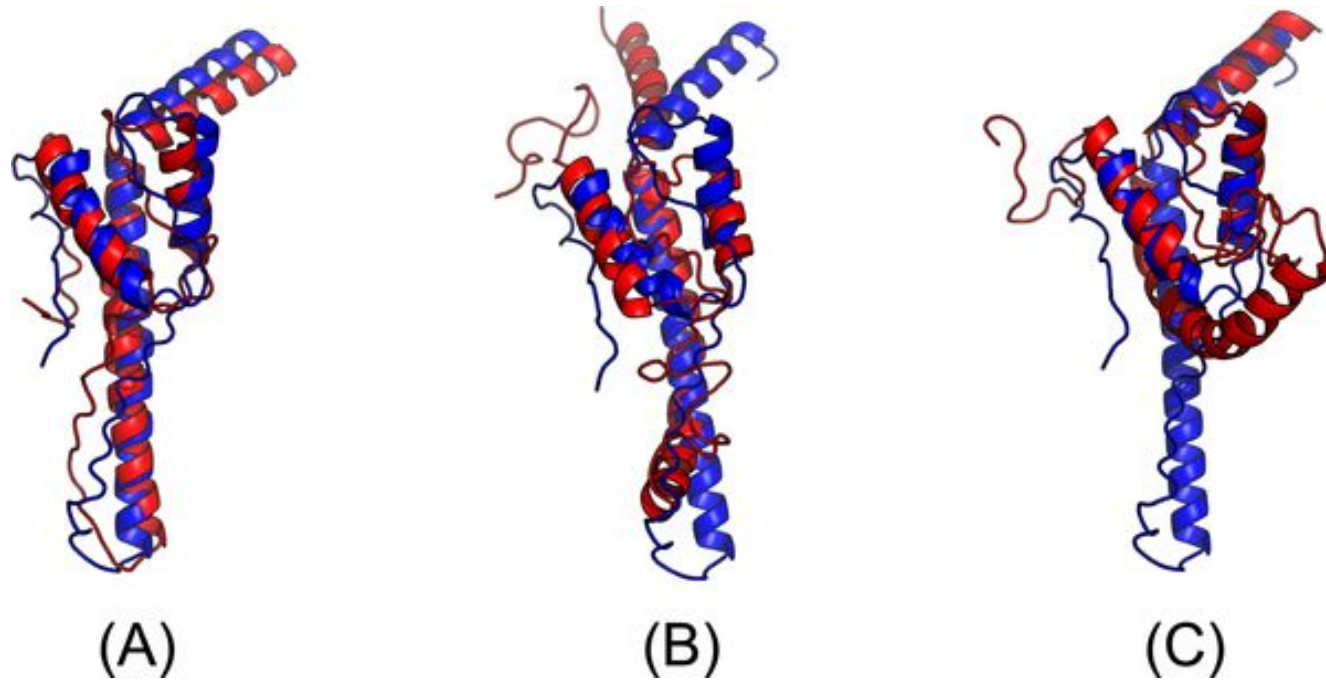
Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLOS Computational Biology 13(1): e1005324. <https://doi.org/10.1371/journal.pcbi.1005324>
<http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005324>

Fig 10. Superimposition between the predicted models (red) and the native structure (blue) for the CAMEO test protein (PDB ID 5dcj and chain A).



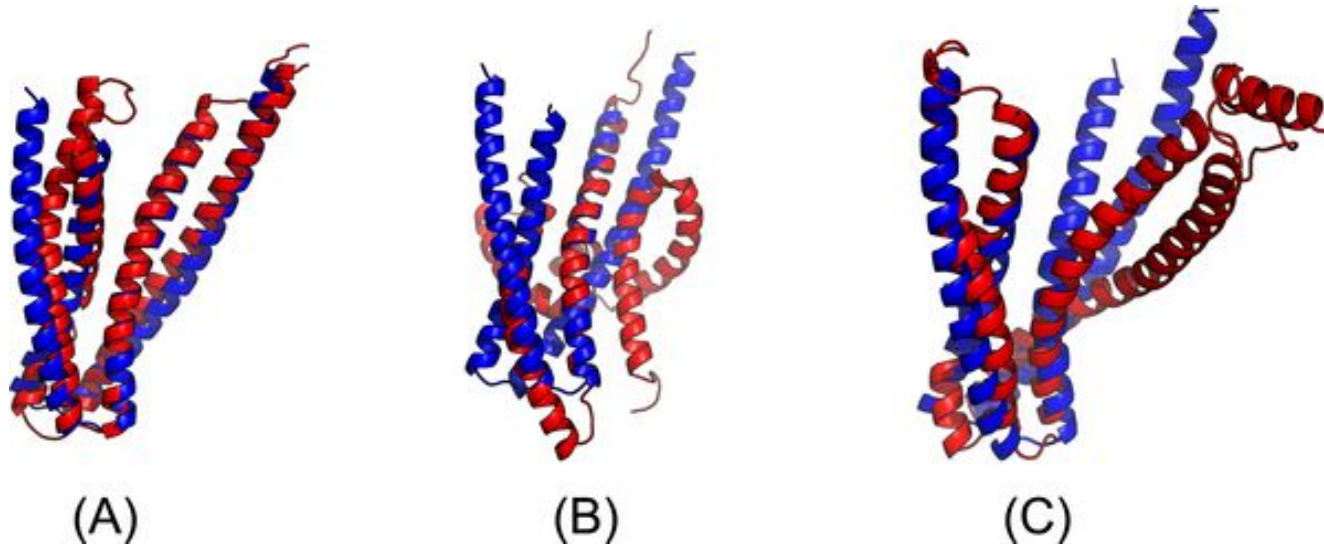
Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLOS Computational Biology 13(1): e1005324. <https://doi.org/10.1371/journal.pcbi.1005324>
<http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005324>

Fig 13. Superimposition between the predicted models (red) and the native structure (blue) for the CAMEO test protein (PDB ID 5dje and chain B).



Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLOS Computational Biology 13(1): e1005324. <https://doi.org/10.1371/journal.pcbi.1005324>
<http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005324>

Fig 16. Superimposition between the predicted models (red) and the native structure (blue) for the CAMEO target 5f5pH.

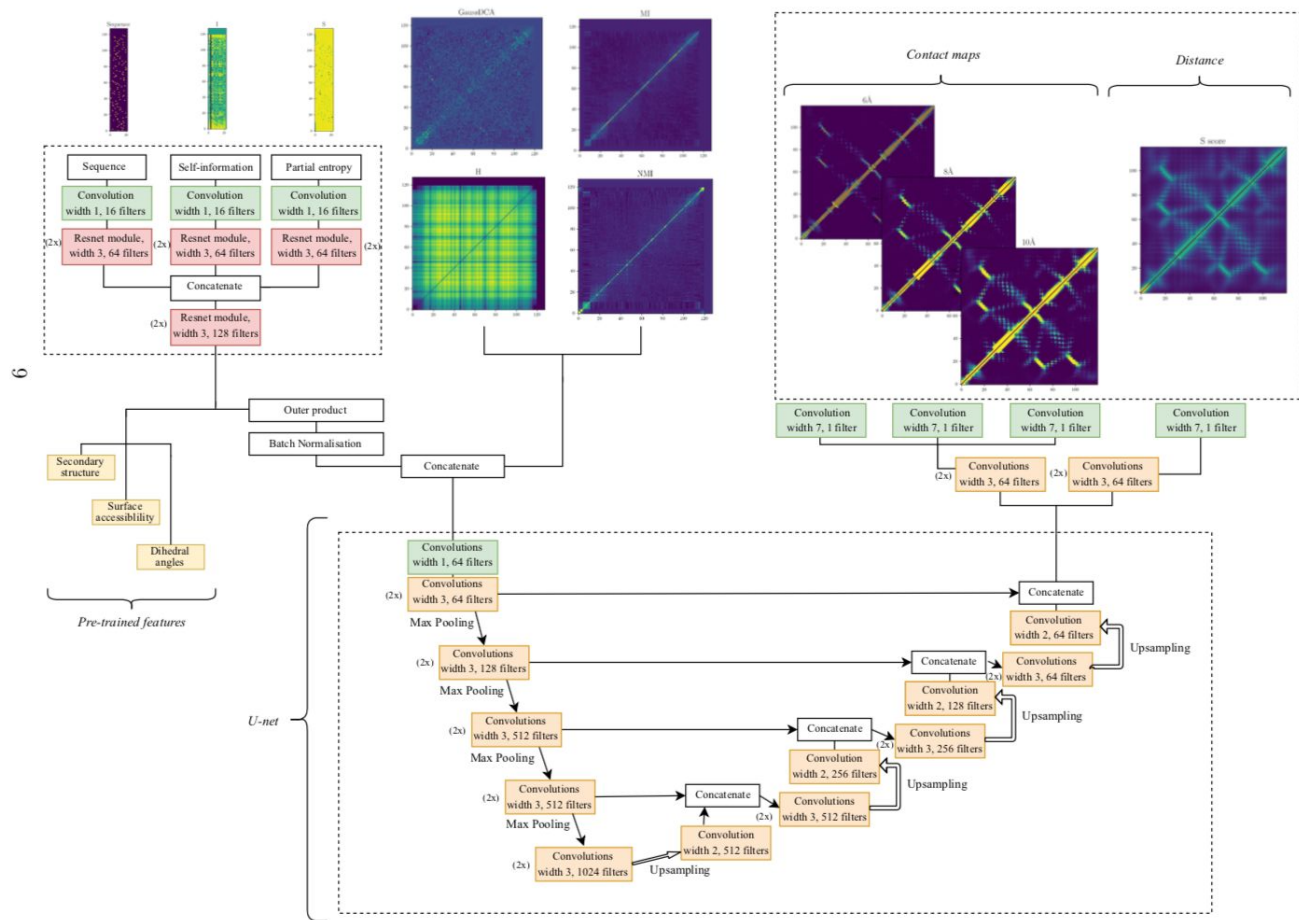


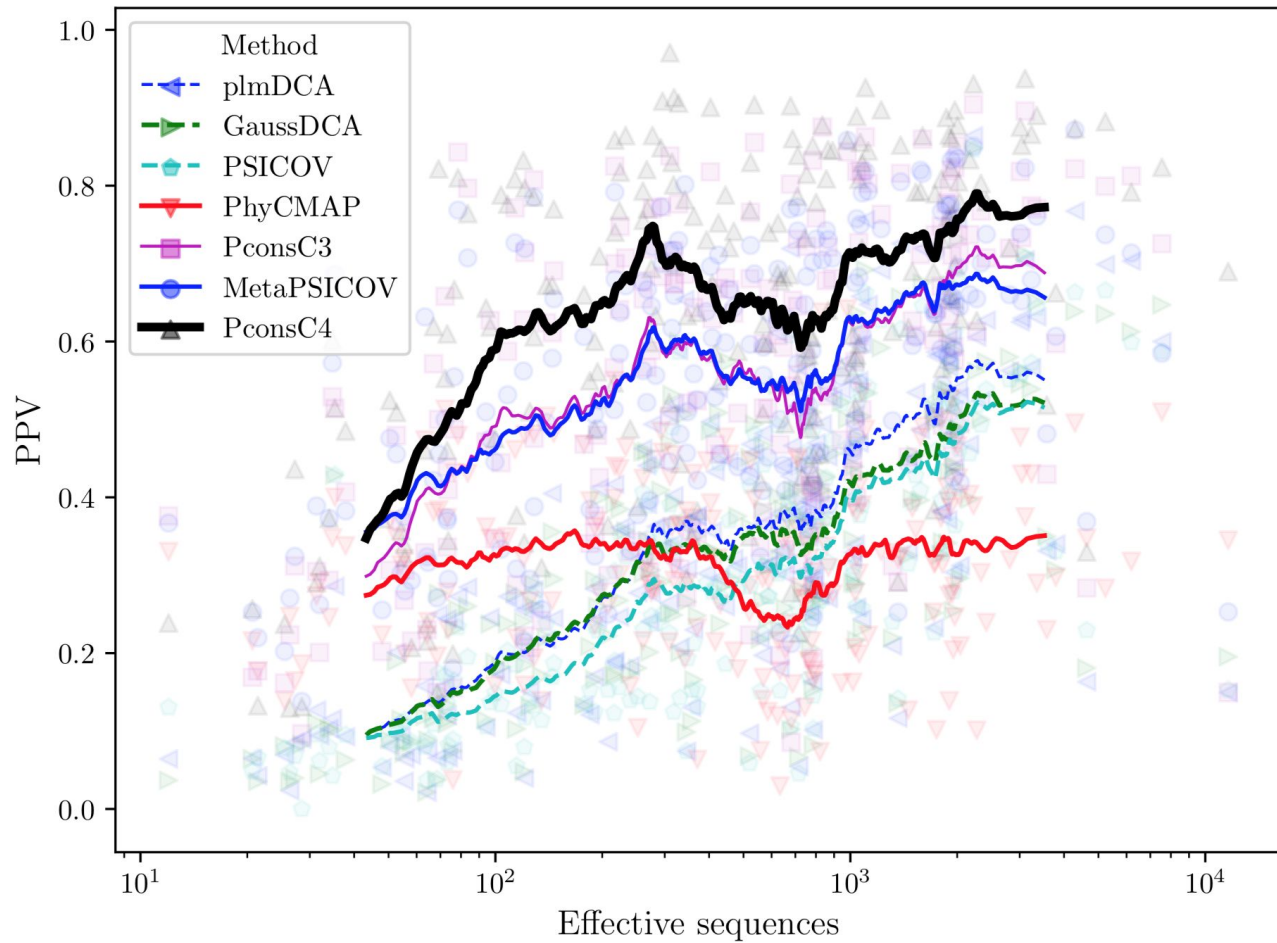
Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLOS Computational Biology 13(1): e1005324. <https://doi.org/10.1371/journal.pcbi.1005324>
<http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005324>

PconsC4

David Menendez Hurtado

PconsC4 pipeline





PconsC4 - fast, easy and accurate

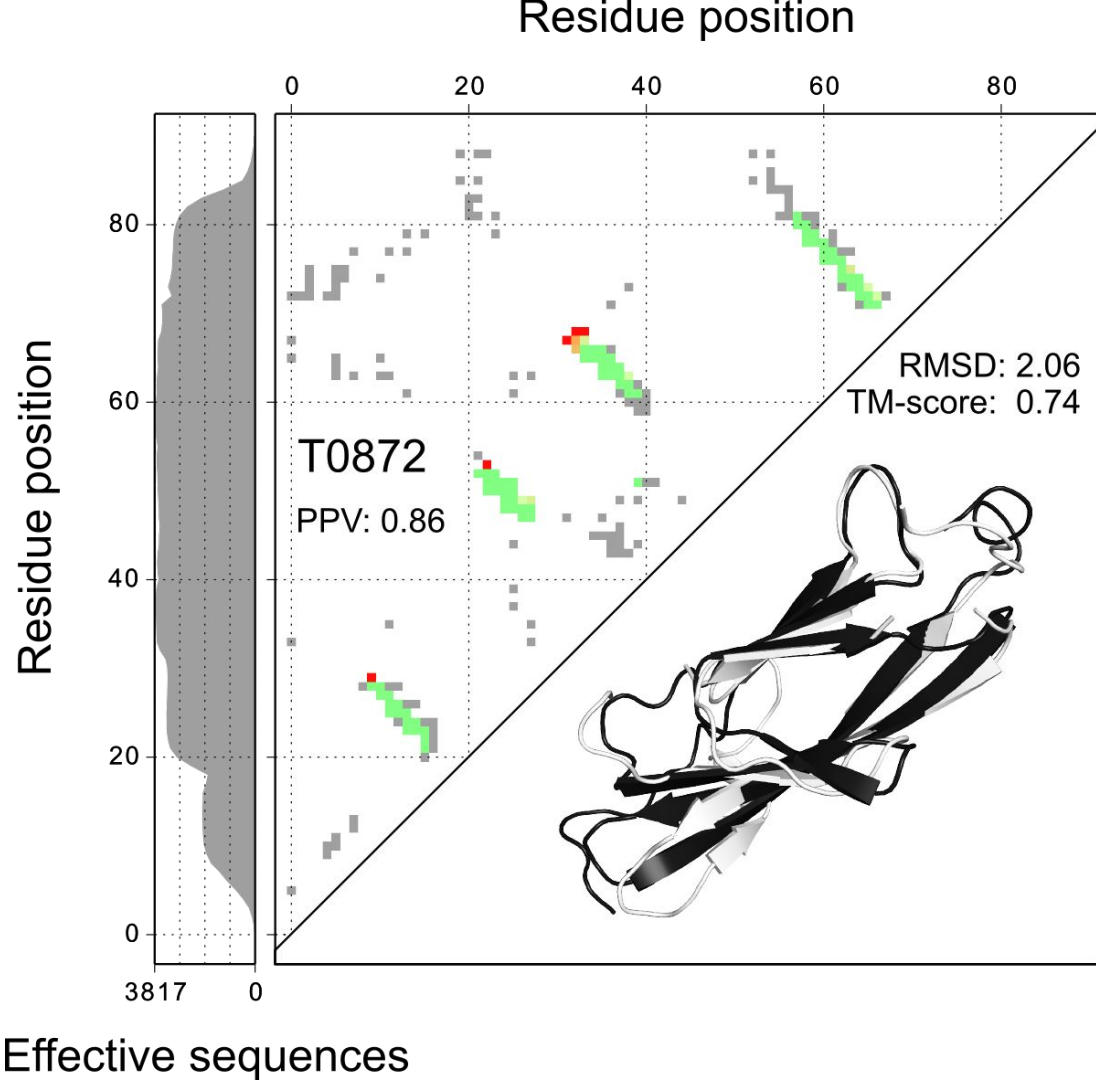
pip3 install -U tensorflow

pip3 install pconsc4

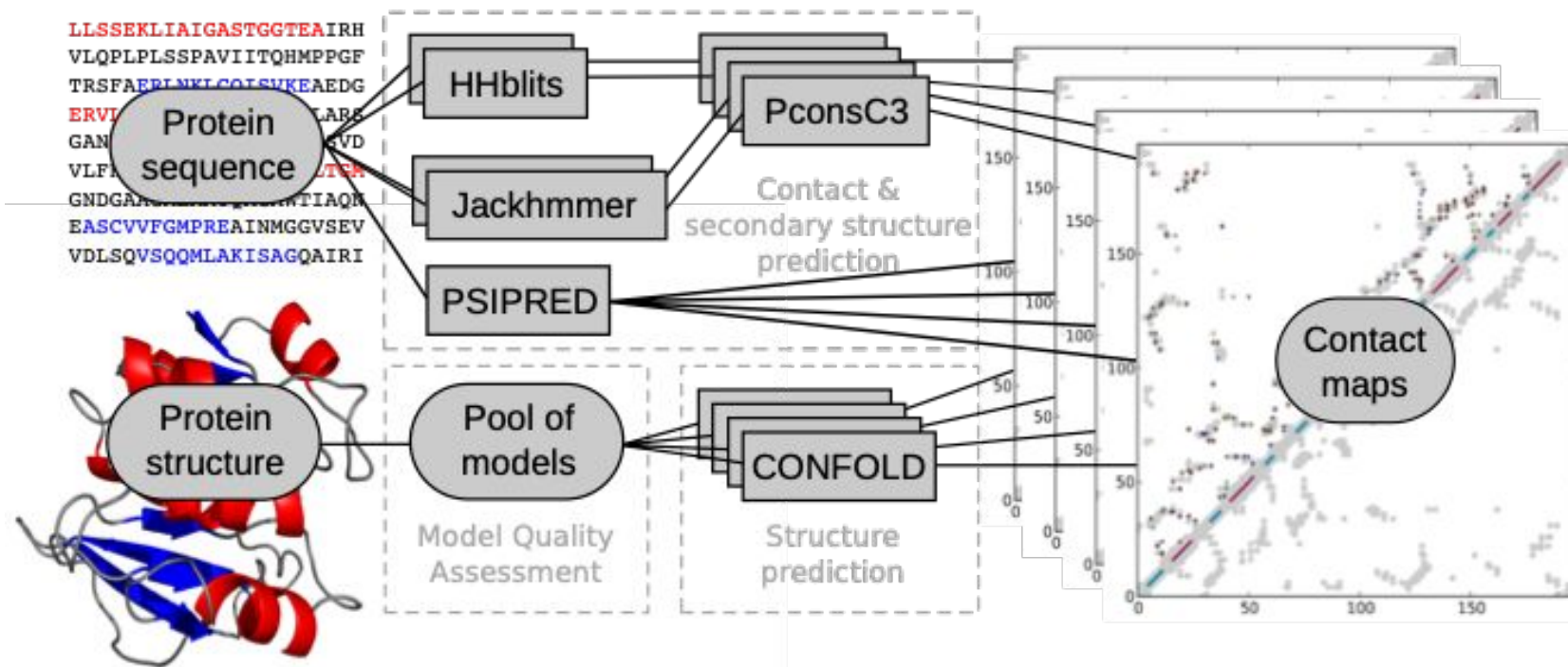
	All	Small	Short	Med	Long	Time [s]	Deps.
PlmDCA	0.36	0.12	0.14	0.15	0.26	84	-
GaussDCA	0.34	0.12	0.14	0.15	0.25	27	-
PSICOV	0.34	0.11	0.13	0.14	0.22	114	-
PhyCMAP	0.32	0.29	0.24	0.19	0.18	718	1,3
PconsC3	0.57	0.36	0.25	0.25	0.37	2931	1,2,3,4,5,6
Metapsicov	0.59	0.39	0.27	0.26	0.36	1158	1,3,7,8,9
PconsC4	0.65	0.42	0.29	0.28	0.45	12 (56*)	-

1: PSIPRED, 2: NetSurfP, 3: PSI-BLAST, 4: PhyCMAP, 5: PlmDCA,
6: GaussDCA, 7: PSICOV, 8: Freecontact, 9: CCMpred.

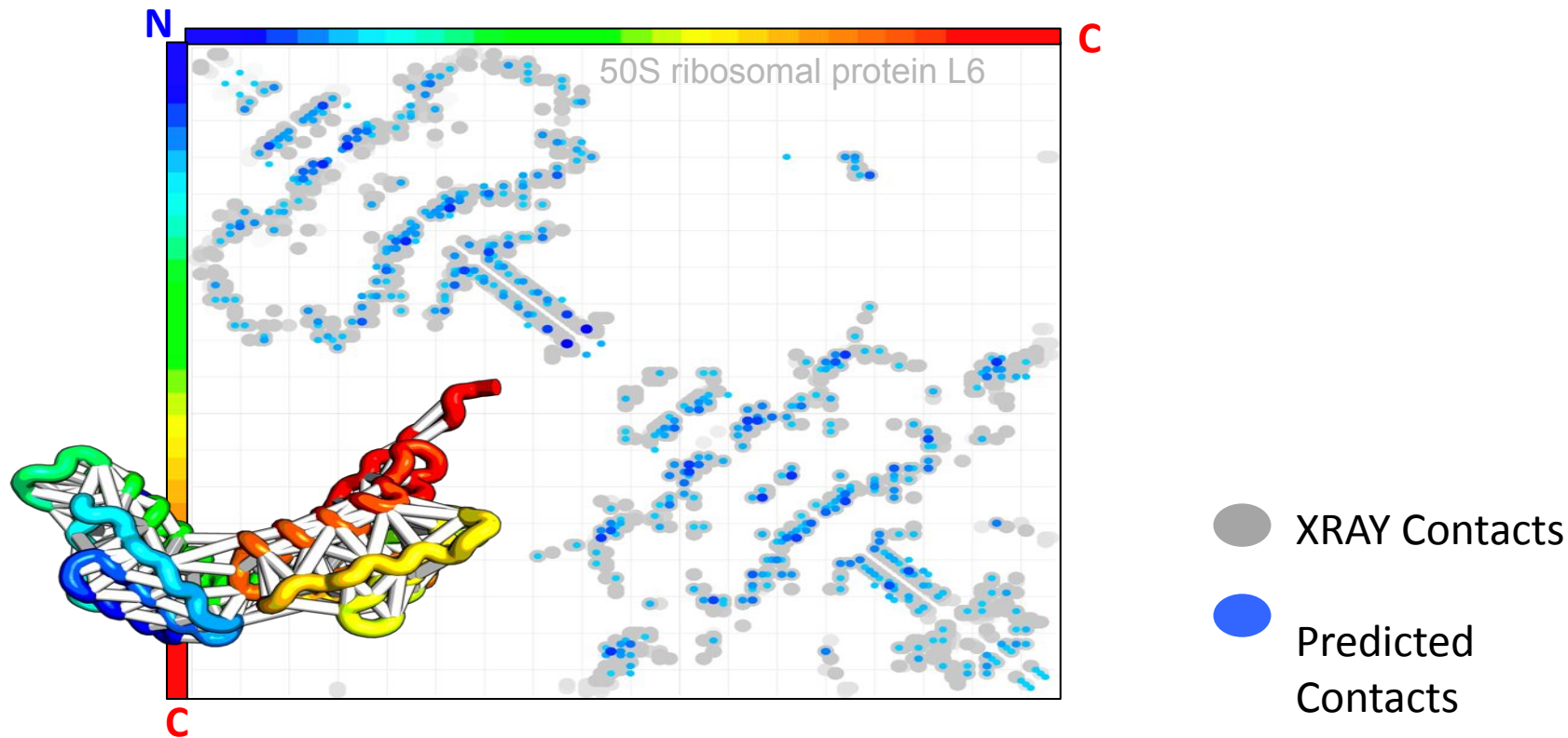
* Timing without pre-loading the network.



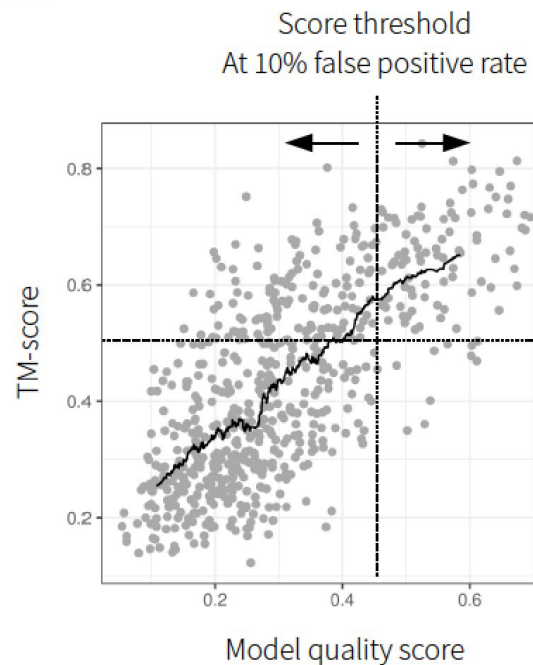
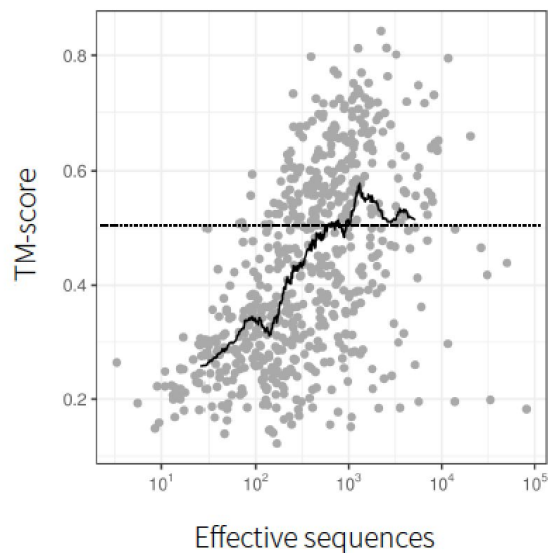
PconsFold flowchart



Overlay of predicted contacts on real contacts



Model quality assessment

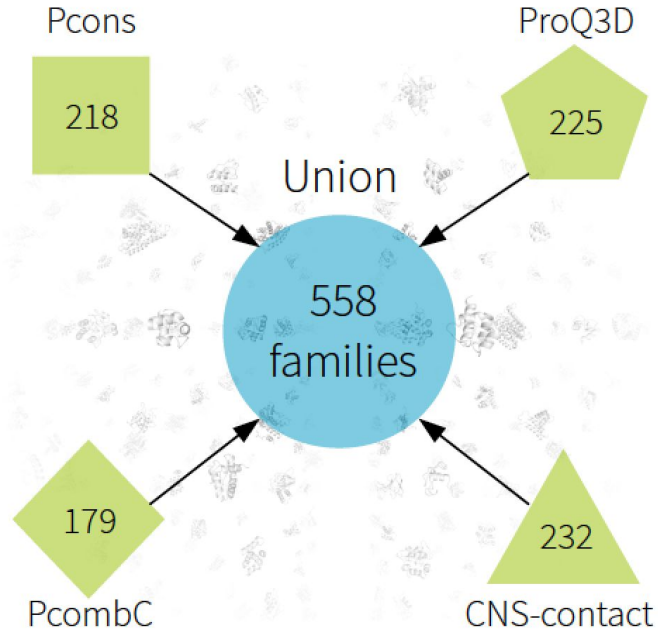


Model quality assessment methods:
Pcons, ProQ3D, PcombC, CNS-contact

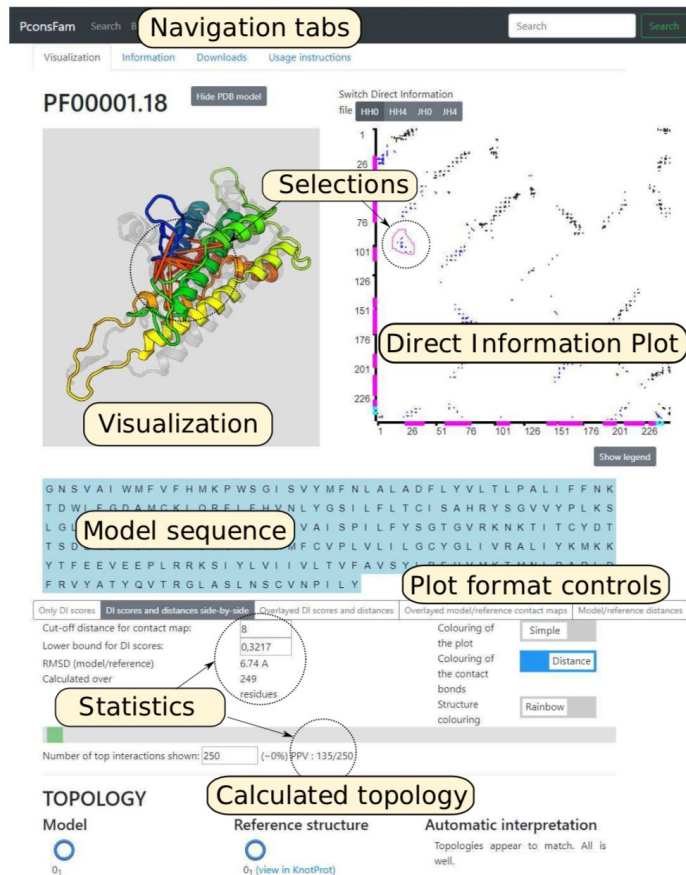
Model quality assessment

...on Pfam families without known structure

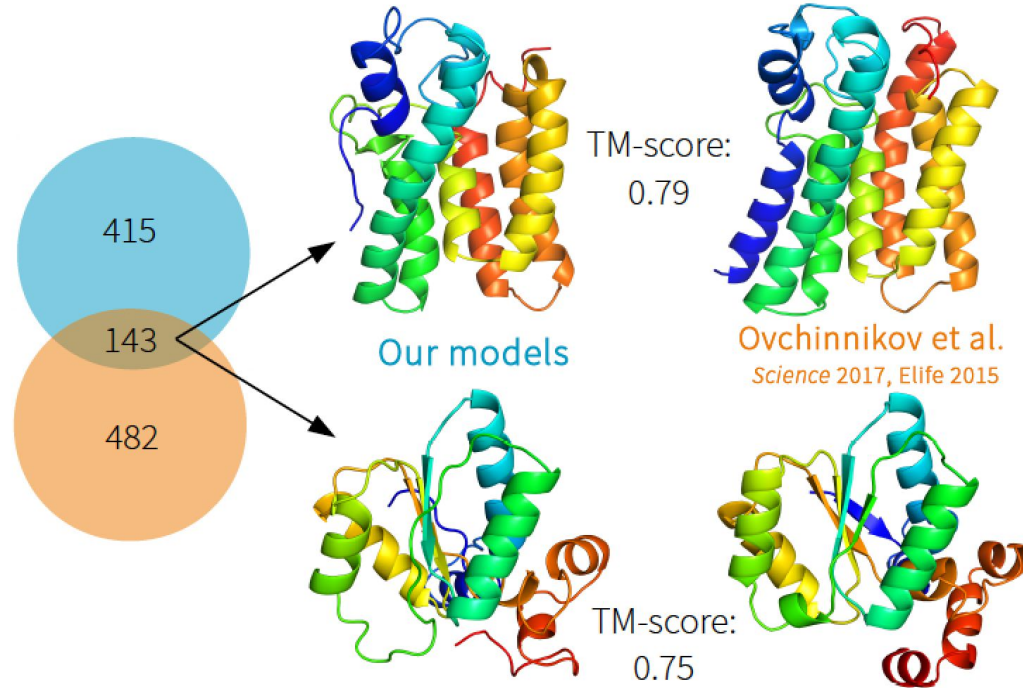
Using a score threshold
at 10% false positive rate



PconsFam database



Complementary to similar studies



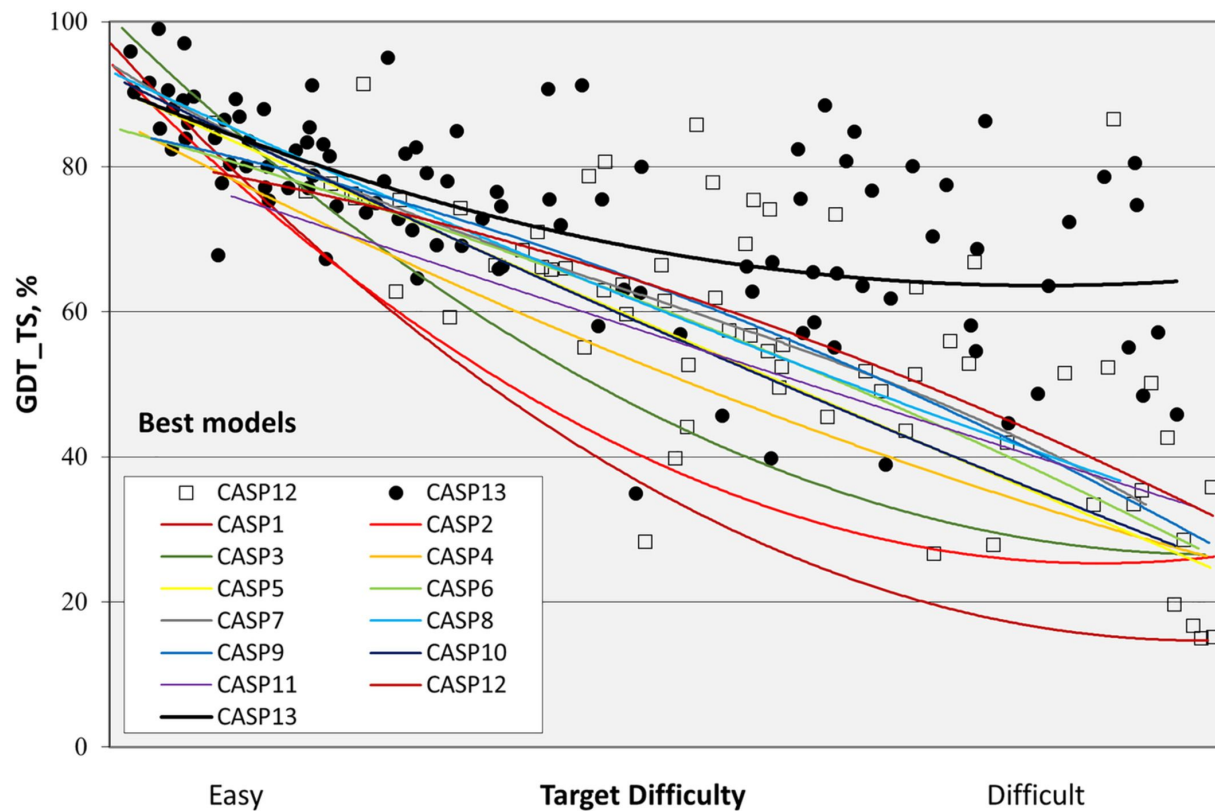
AlphaFold 1

Deepmind in CASP13

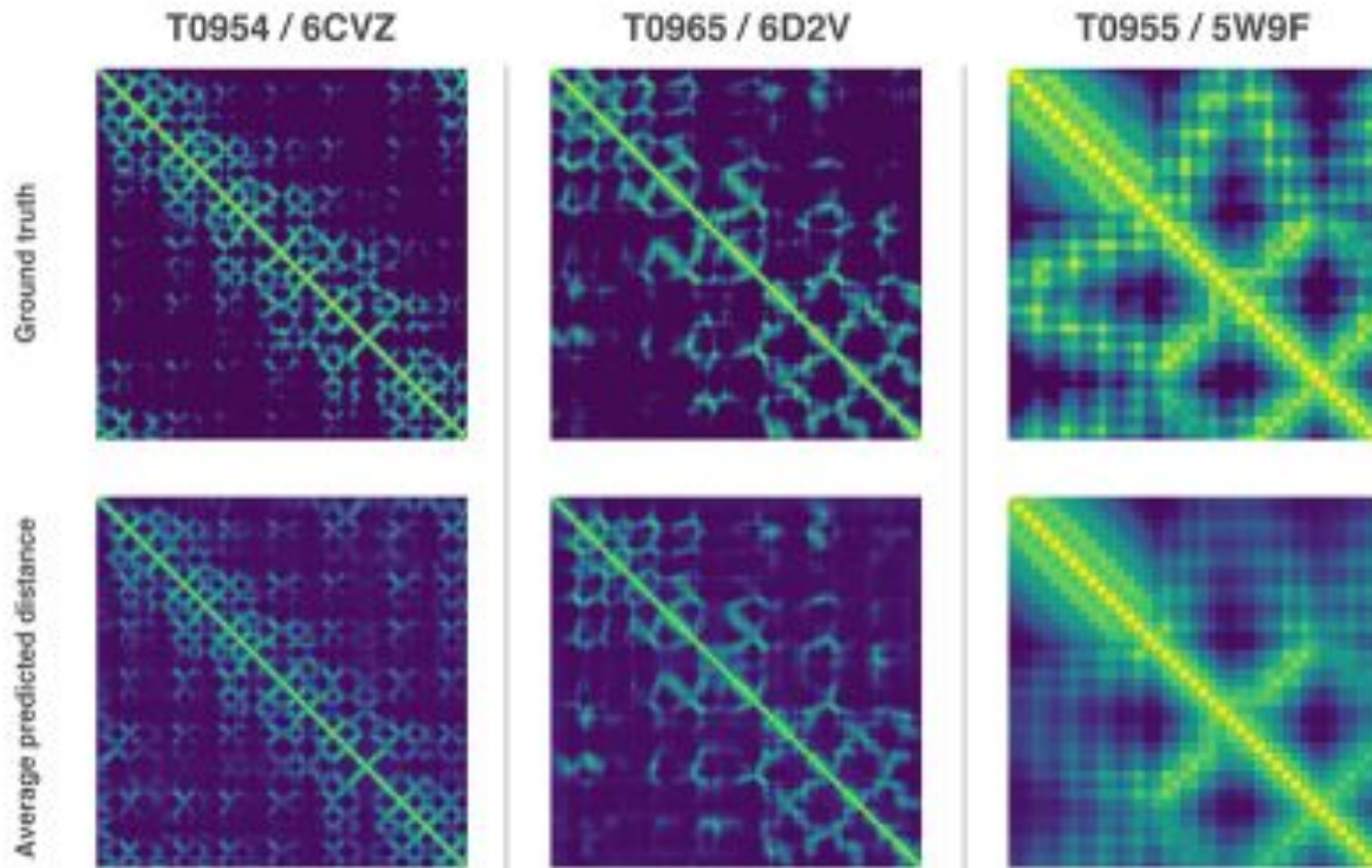
DeepMind algorithm

1. Predict distance distributions not only contacts.
2. Similar to PconsC4 inputs (plmDCA
3. Deeper and faster converting network.
4. Predicting fragments from distances.
5. Steepest descent minimization protocol

The leap in performance at CASP13

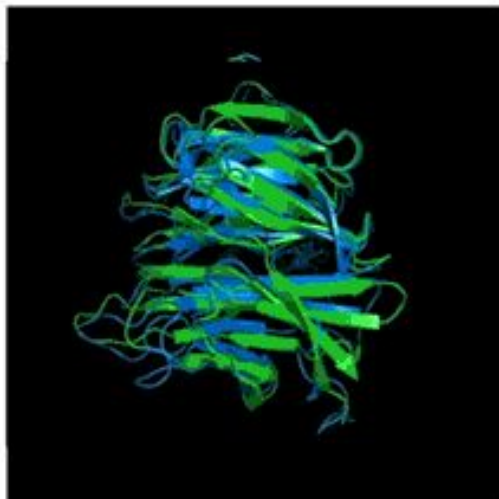


Contact predictions

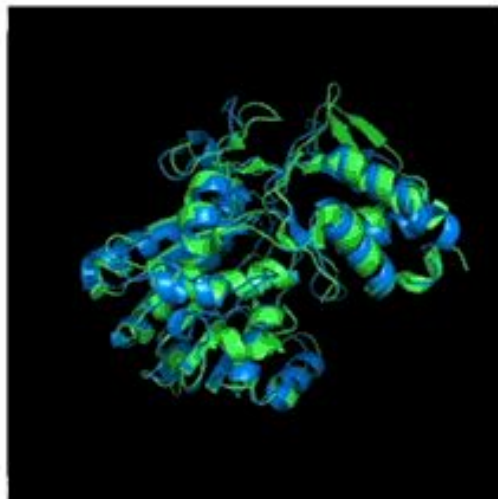


Structures:
Ground truth (green)
Predicted (blue)

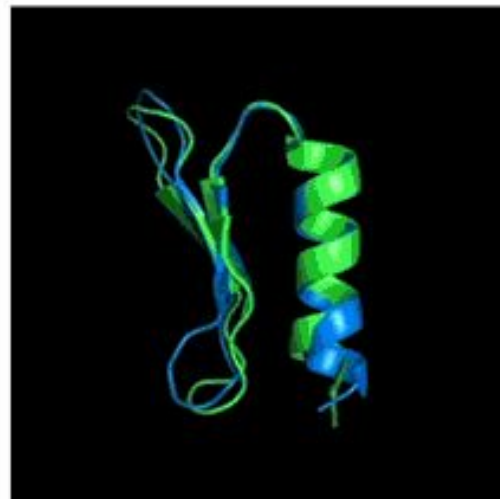
T0954 / 6CVZ



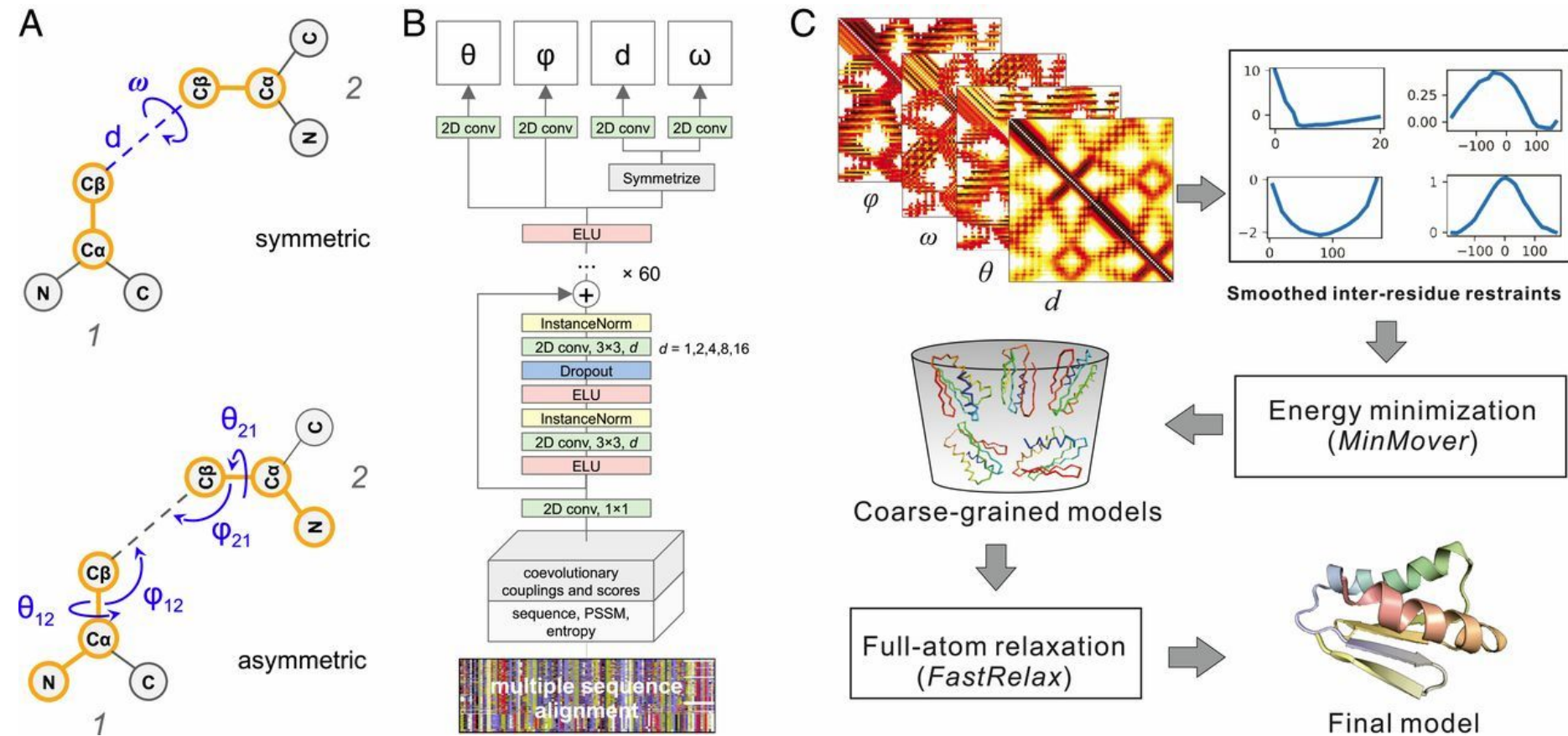
T0965 / 6D2V



T0955 / 5W9F



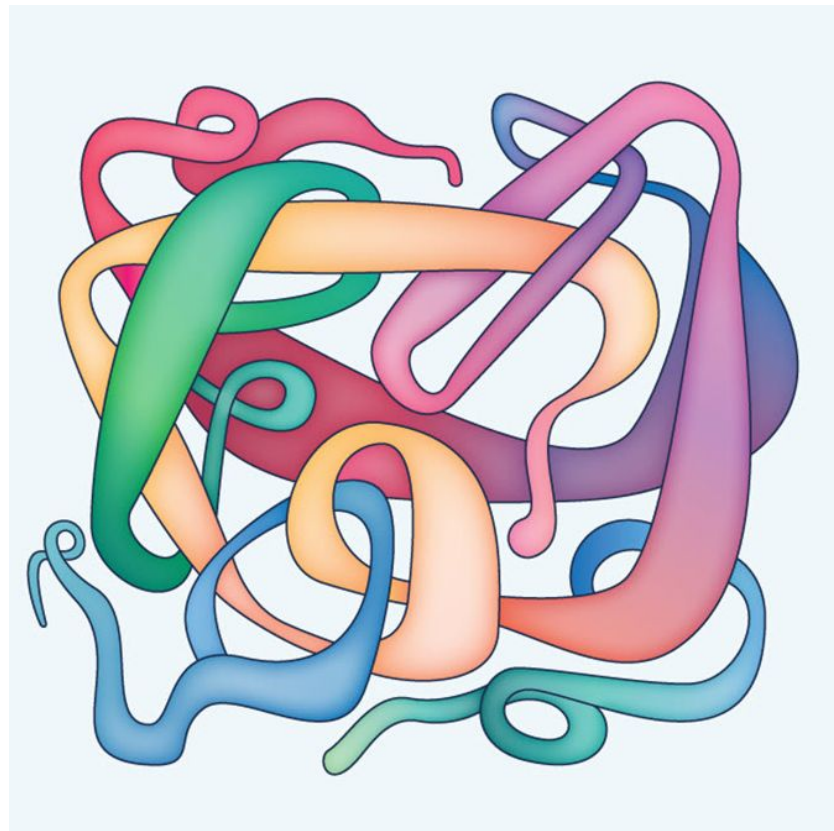
trRosetta



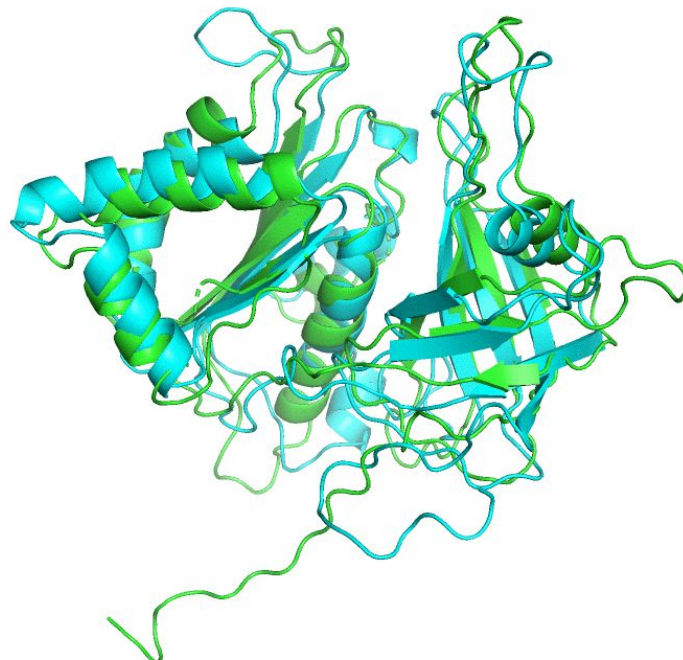
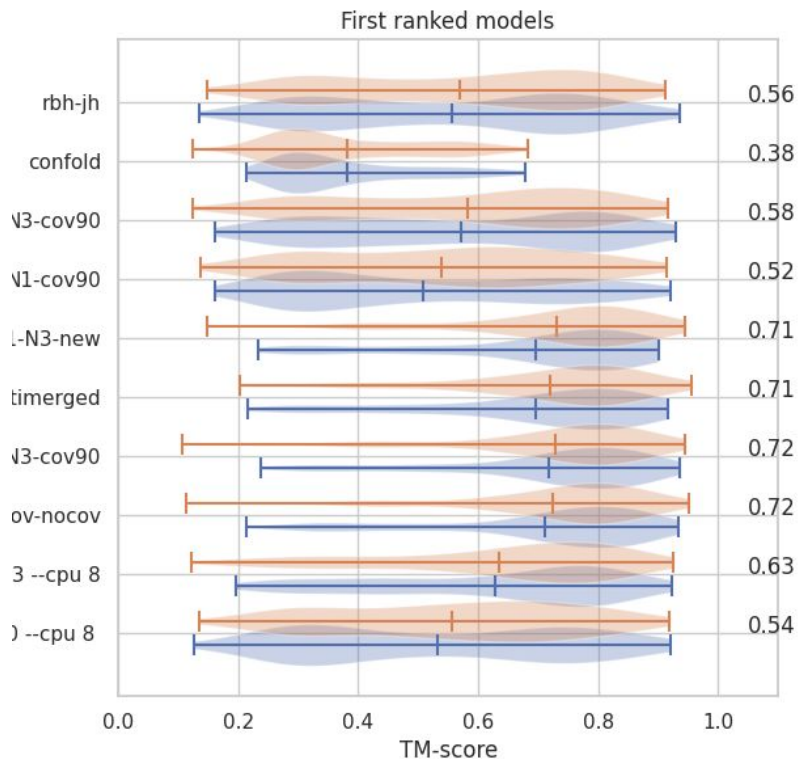
Insights from CASP13



- Deep learning is better
- Most single domain proteins can be predicted quite accurately
- Predictions of protein-protein interactions will become important



Folding - an (almost) solved problem in CASP13.

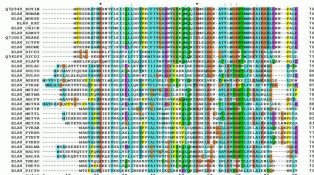


[illegible]

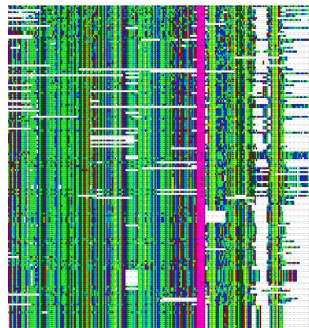
```
stpred/1ewy u1 A-1ewy u2 A.npz NumLongContacts: 118 LongPPV: 0.0
```



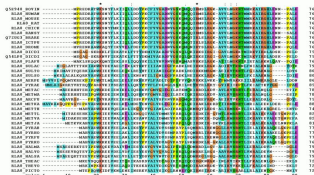
MSA 1



Jackhmmer



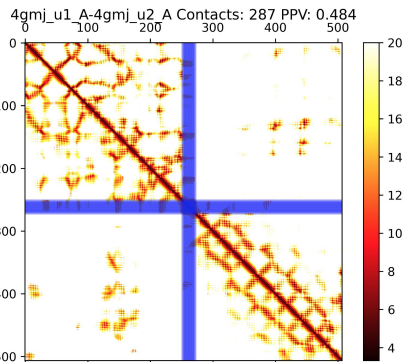
MSA 2



Jackhmmer

Merged MSA

Top hit in each
proteome



Distance
prediction

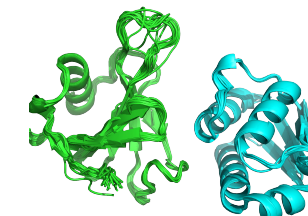
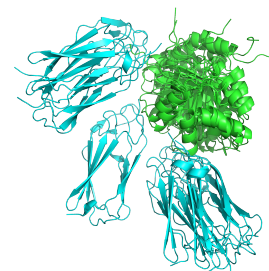
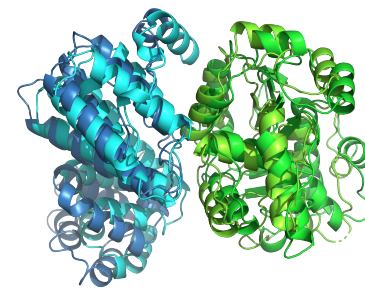
trRosetta

Fold and Dock

pyRosetta

Scoring

PconsDock

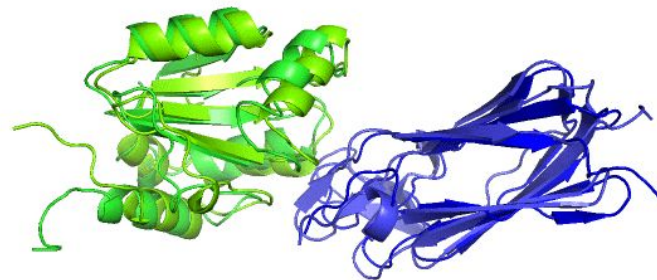
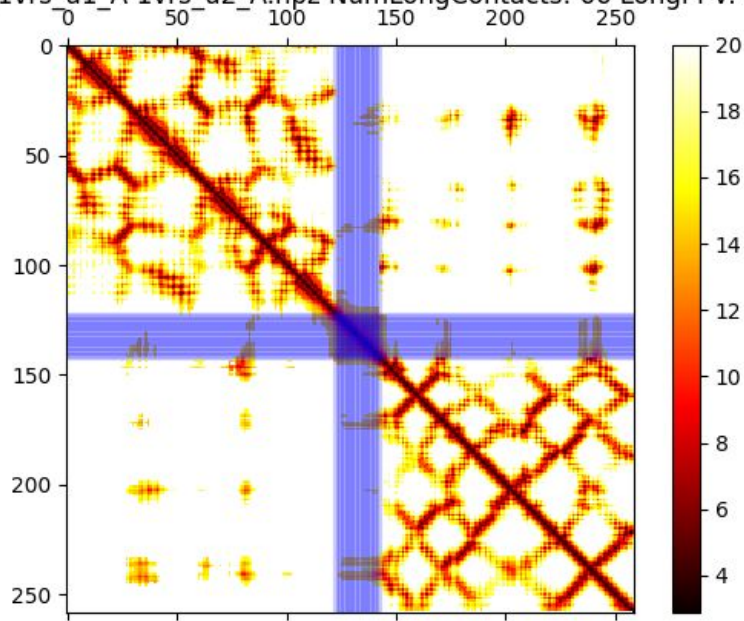


Limits and potential of combined folding and docking using PconsDock

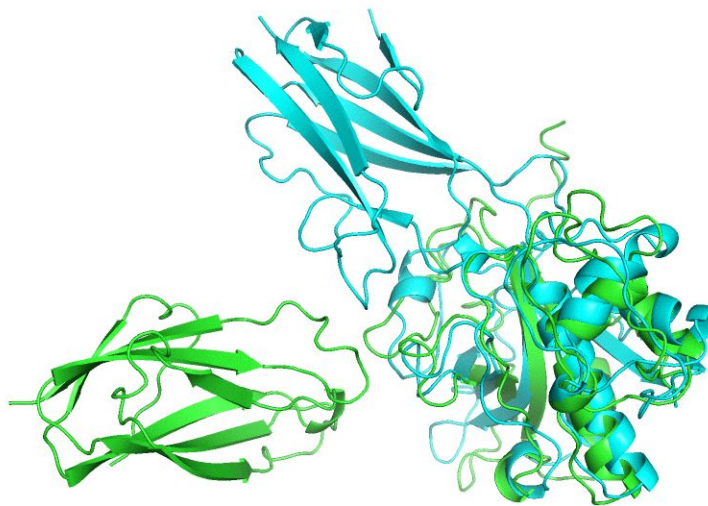
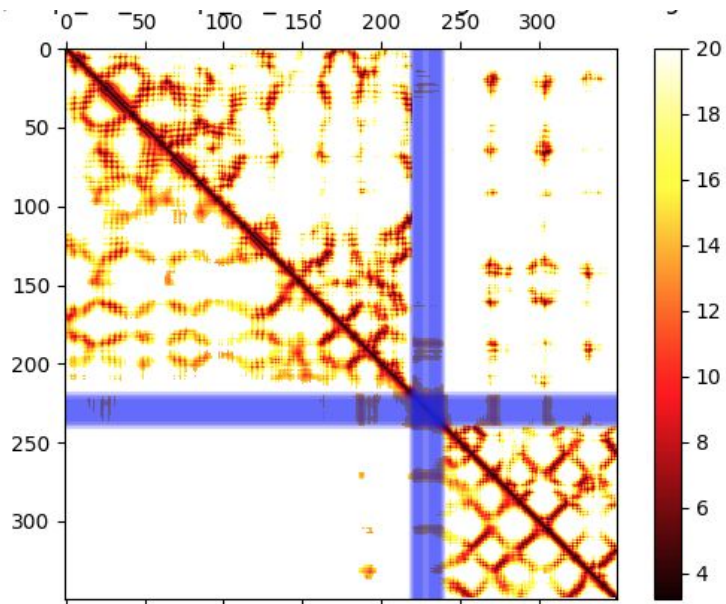
Good Example 1vrs (RMSD 2.65Å)



istpred/1vrs_u1_A-1vrs_u2_A.npz NumLongContacts: 66 LongPPV: 0.818



Failures 2nqd

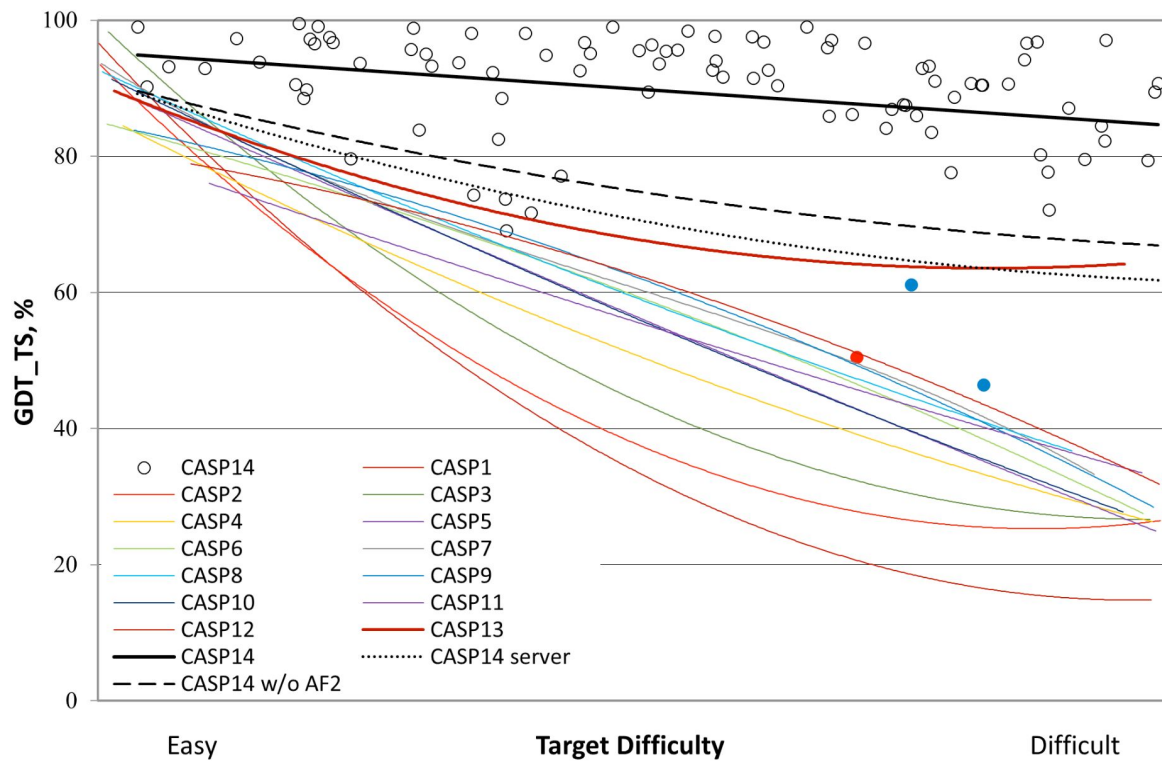


Breakout 1

To discuss

- How much did growth of databases contribute to AlphaFold1
- What is really co-evolution
- What does a contact-predictor deep neural network learn.

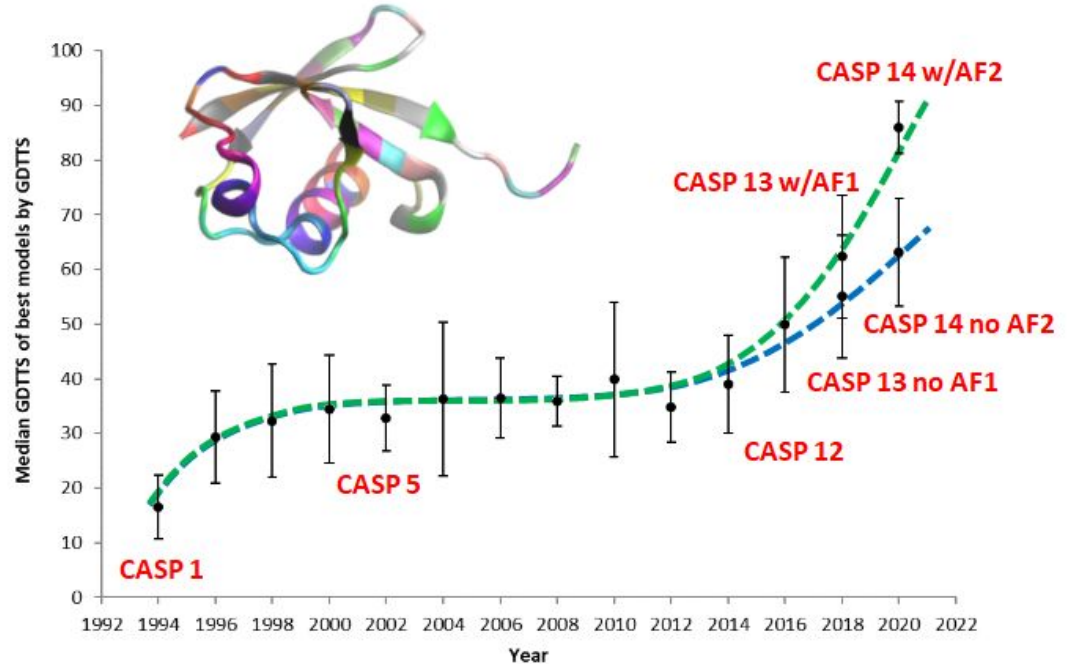
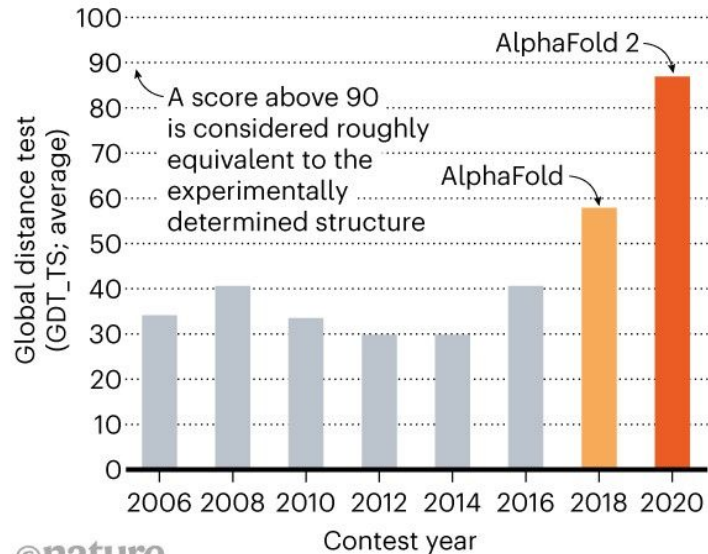
CASP14



AlphaFold2 - CASP14 (2020)

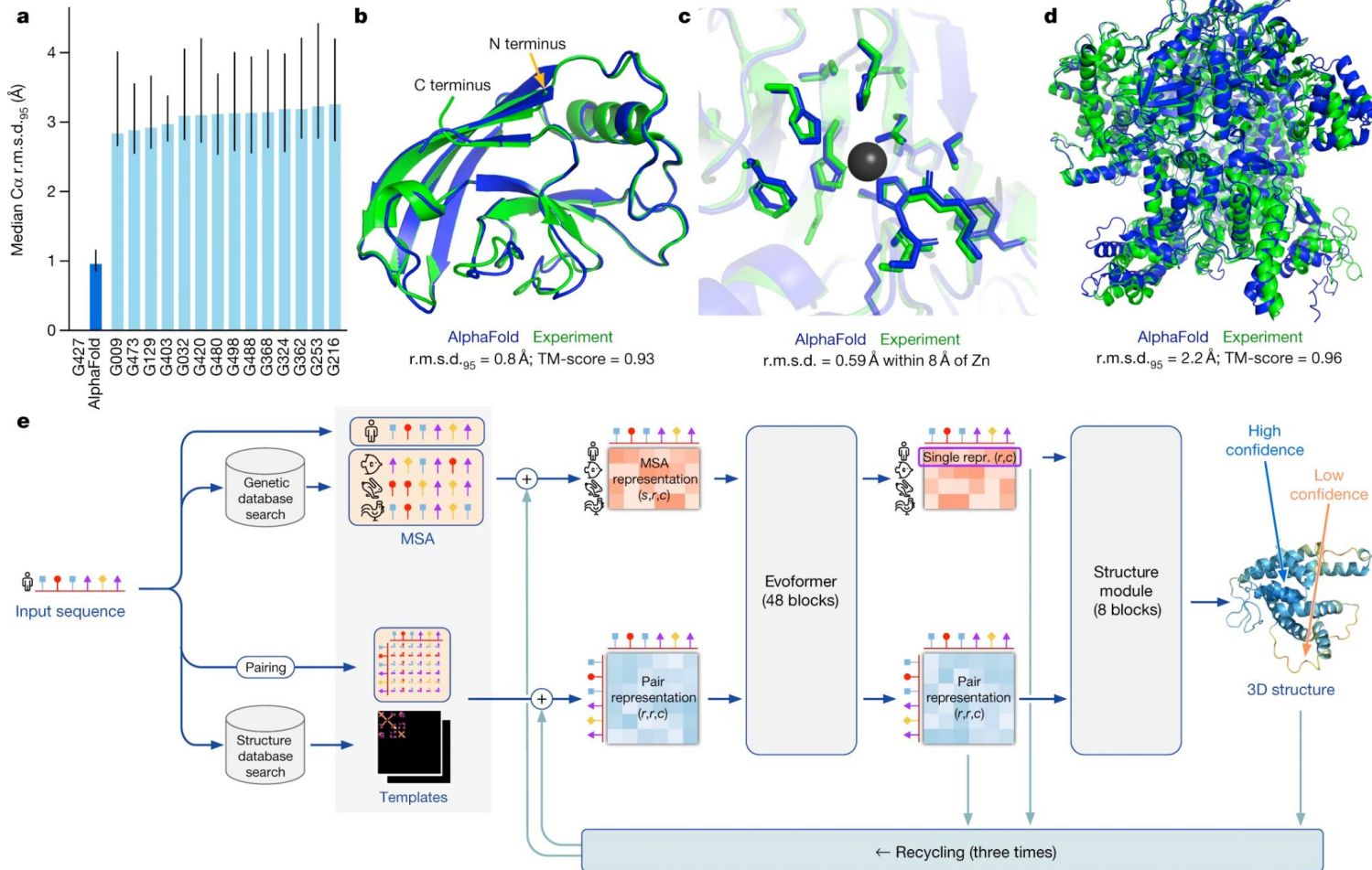
STRUCTURE SOLVER

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.



What is AlphaFold2?

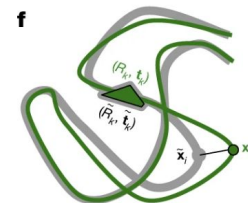
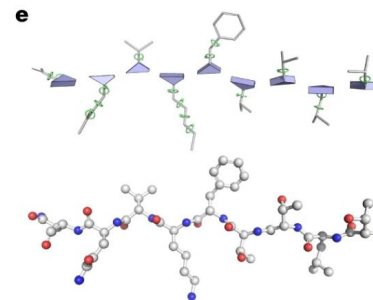
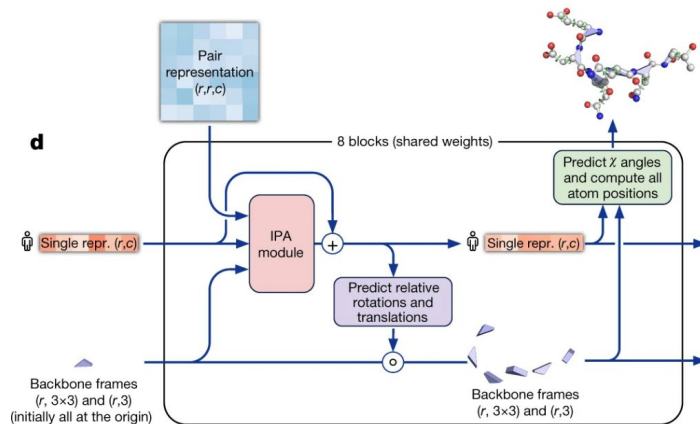
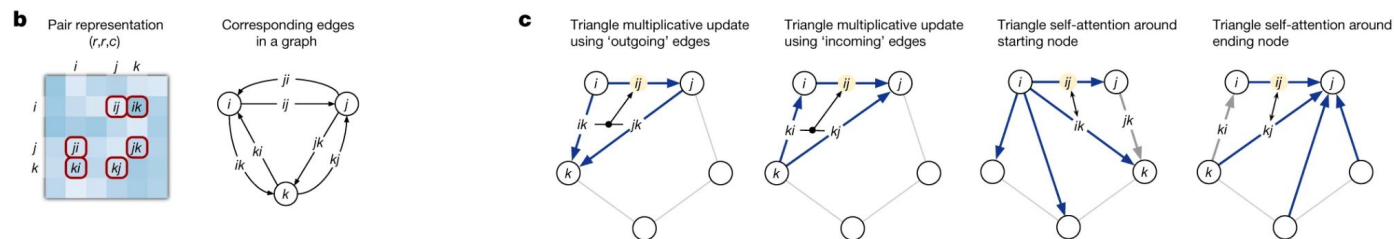
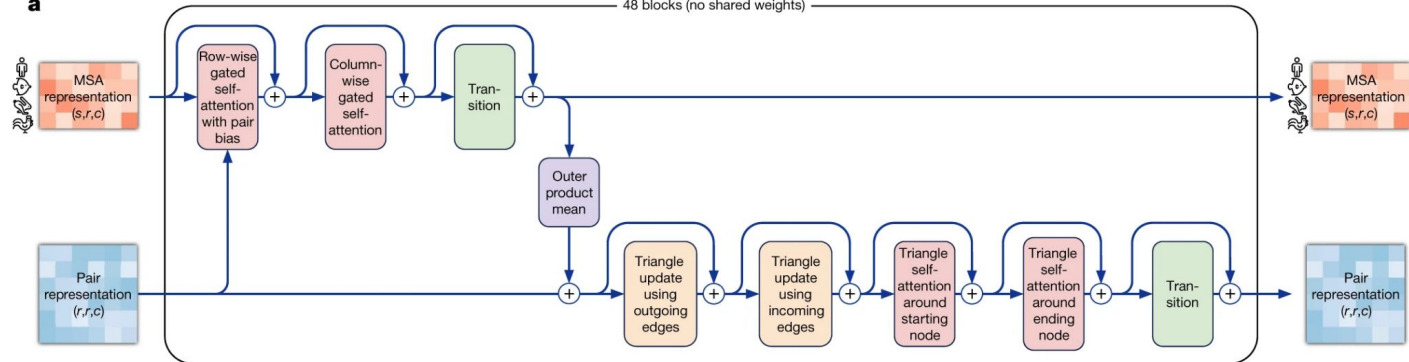
“This leaves one to wonder exactly how much computational power went into this project. In their original announcement, DeepMind claimed that they used “128 TPUv3 cores or roughly equivalent to ~100-200 GPUs”. Although this amount of compute seems beyond the wildest dreams of most academic researchers, my friends at Google tell me it is not uncommon for individual Googlers to access similar resources on a regular basis. How many times more computational power did this team have, in comparison with all the other CASP14 contenders combined?



Reviewed in. Protein sequence-to-structure learning: Is this the end(-to-end revolution)?

<https://doi.org/10.1002/prot.26235>

Evoformer

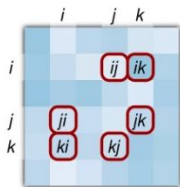


<https://www.blogig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/>

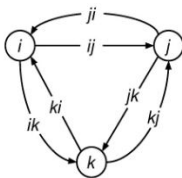
Evoformer - triangular updates



b Pair representation
(r, r, c)

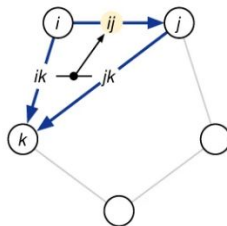


Corresponding edges
in a graph

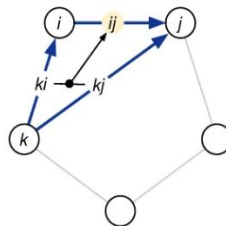


c

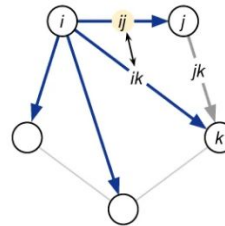
Triangle multiplicative update
using 'outgoing' edges



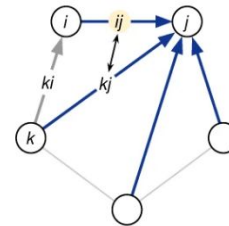
Triangle multiplicative update
using 'incoming' edges



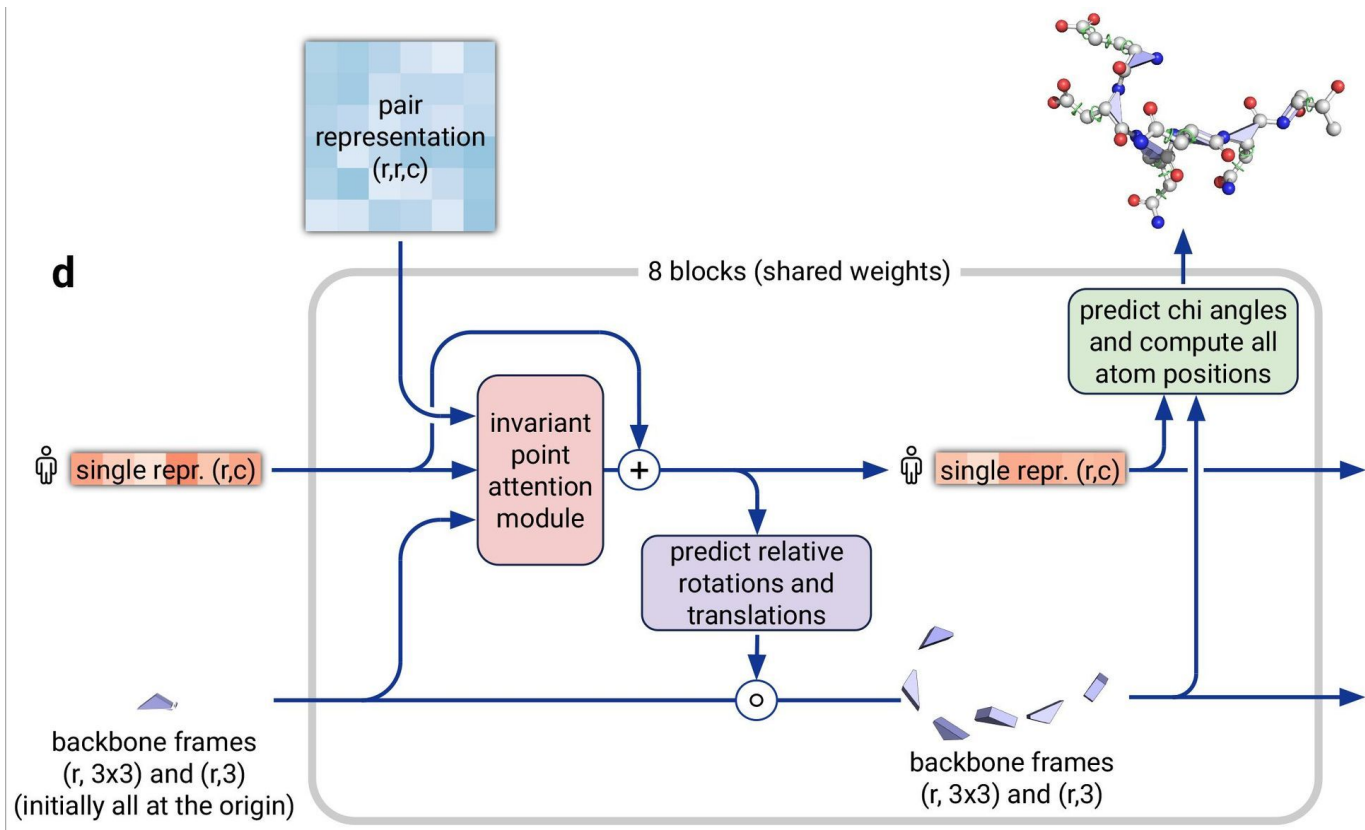
Triangle self-attention around
starting node



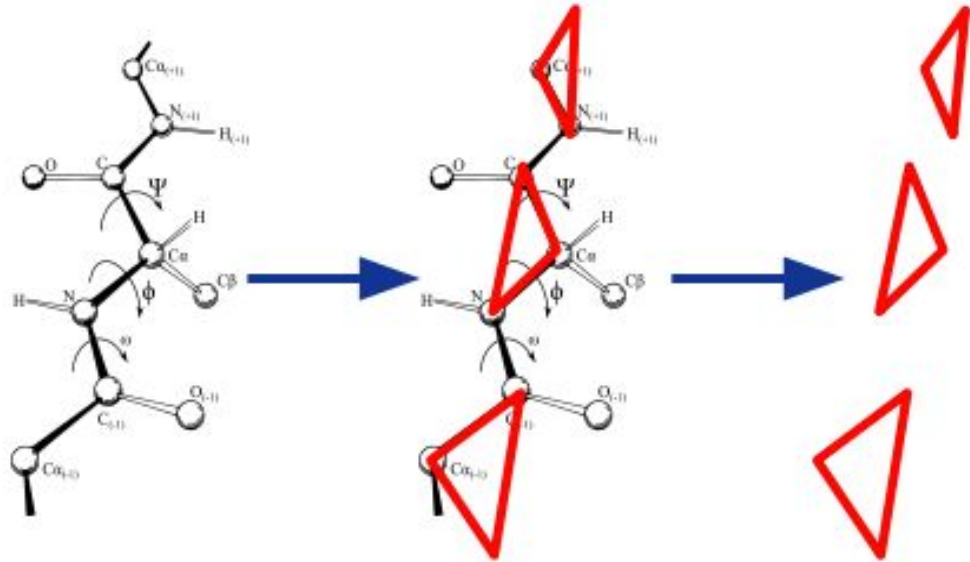
Triangle self-attention around
ending node



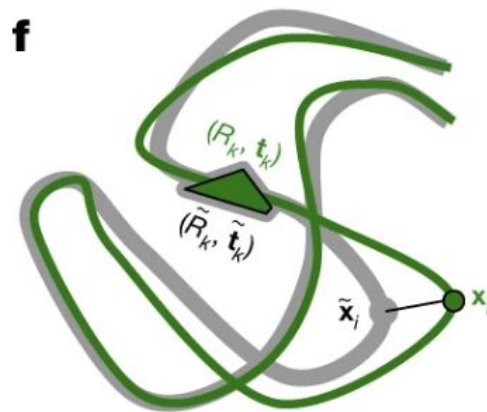
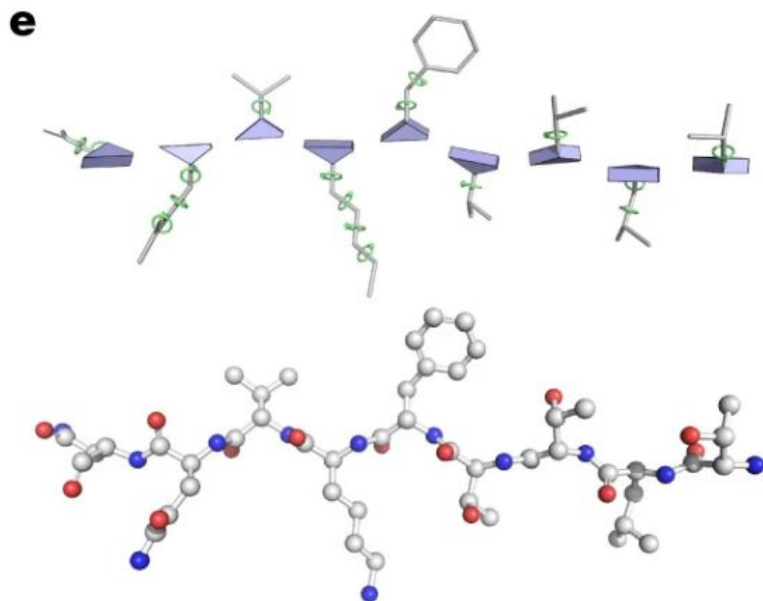
Structural module



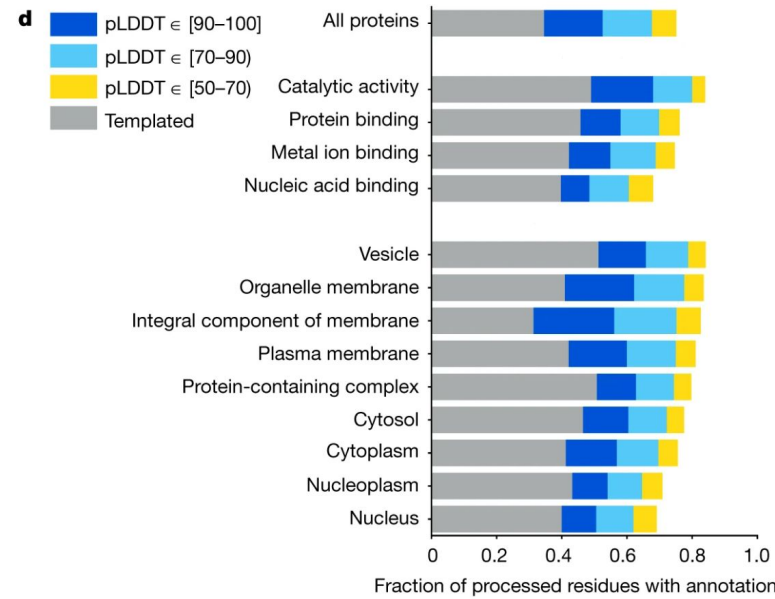
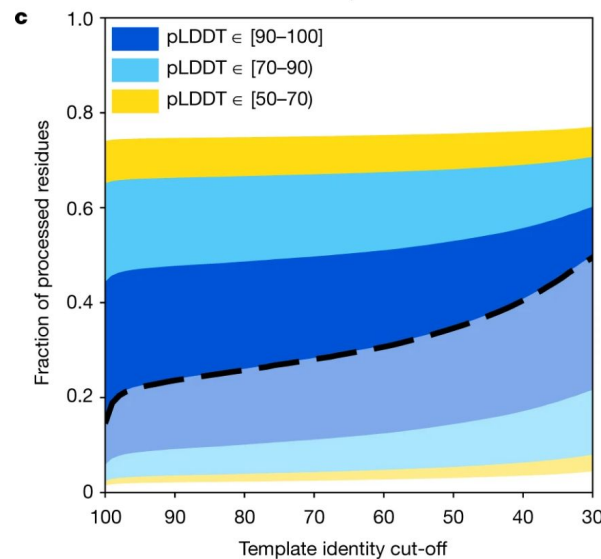
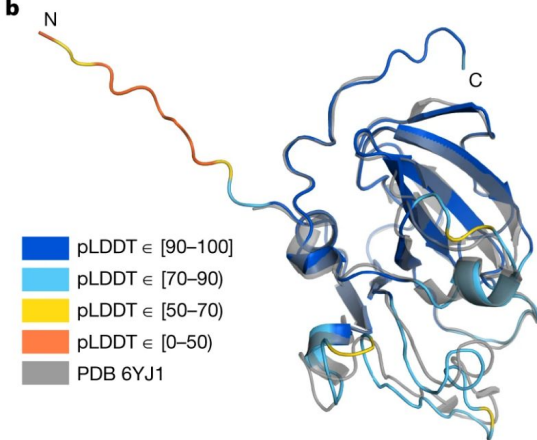
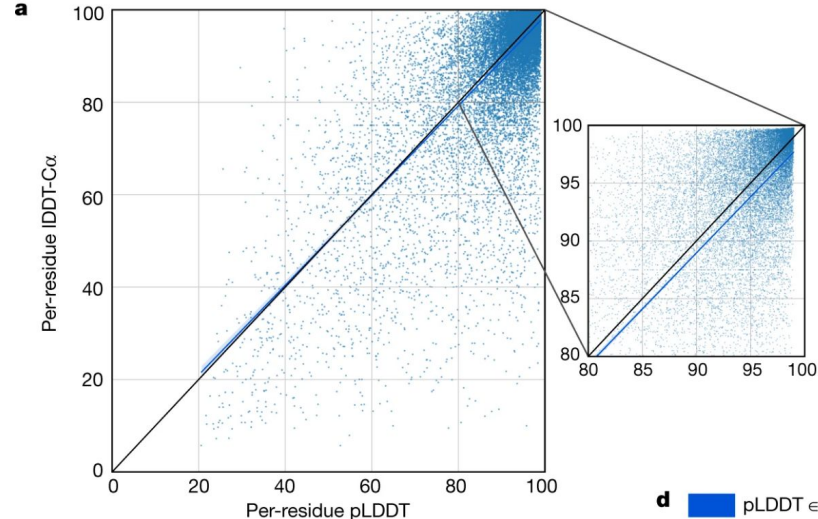
Structure Module



Structure Module



How good is AlphaFold2 for individual proteins.



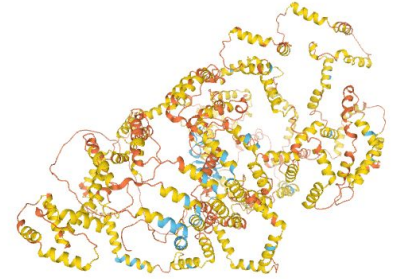
Known limits to AlphaFold

- Predicting if a particular chain folds or not.
 - But models can be used as input to other programs
- Predicting structural states of a protein
 - Sometimes you get two states, sometimes not
 - Most likely easier to get more states if you have shallower alignments
- Predicting structure of large - flexible (repeated) proteins
 - Individual domains are often OK, but arrangement is random
- Predicting structure of single spanning TM-proteins
 - AF2 does not know what a membrane is
- Predicting structure of DNA-binding protein complexes
 - Sometimes OK
 - Sometimes it puts another protein in the place of the DNA (or other large ligand)



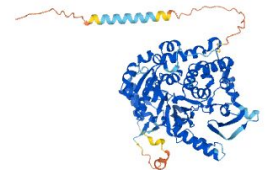
Sequence of AF-F1R833-F1 1: Nebulin-re... A

```
MNVQKARGGFVYPAEKVLIGQNHKACFCEVCKMVLTAANNVSHQKRPYCQVHNPKNITFTSVYETPUNINAKKQSEAVSENKYKEEGERFSTHYDMRSKEIEQARKASQLAQGVQTEFVQDQSFYSGSVTSQEIVRVSTQKTIISOVEYKRGHETRMLQFTSVADSPVELHAKTGVSLASDVKYTEEYQSGKSGSFAMITPGYEMAKKANTLASNLEYKKGHEERVSKYTTVDT
```



Sequence of AF-QSA687-F1 1: Alpha-1, 2... A

```
NGSIFKDGRRILVRPKSLIICLISITFTQLIRYQQLADEVQPTINEHSSQSLKMTKLNSTRSSSPISPPKSLNKLQSFEMHETNIFEINKFDDLPKLIKYPKEQLTKTKITRWLLSKANTHKDITROYHESVLRLNPKSLPKSVYTPHYIYITIGGNFYSWMAYIQLQLRLKLSNLPEVLILPSIEDYKAEHFCQHVLPQYNAKCLLVPKLGFWAKMWFSSYQKFA
```



Development during the first year



- RoseTTAFold
 - <https://doi.org/10.1126/science.abj8754>
- Protein-protein Interactions
 - <https://doi.org/10.1038/s41467-022-29480-5>
 - <https://doi.org/10.1101/2021.10.04.463034>
- Protein-peptide interactions
 - <https://doi.org/10.1038/s41467-021-27838-9>
- Building Large complexes
 - <https://doi.org/10.1038/s41467-022-33729-4>
- RNA
 - <https://doi.org/10.1101/2022.05.15.491755>
- Proteins, RNA and DNA
 - <https://doi.org/10.1101/2022.09.09.507333>
- Protein Design
 - <https://doi.org/10.1126/science.add2187>
- Ligand binding
 - <https://arxiv.org/abs/2209.15171>
- Reimplementations
 - <https://doi.org/10.1101/2022.08.04.502811>
- Language models
 - <https://doi.org/10.1101/2022.07.21.500999>
 - <https://doi.org/10.1101/2022.07.20.500902>
- Dynamics
 - <https://doi.org/10.1101/2021.11.16.468810>
- Disorder
 - <https://doi.org/10.1016/j.jmb.2021.167208>

Running AlphaFold 2

← → ↺ 🏠

🔒 🔑 🔍 https://colab.research.google.com/github/sokrypton/ColabFold/blob/main

🌐 ⭐

📧 ⬇️ 🐦 🟩 🟪 🟢 🎓 ⚙️ 📄 ☰

+ Code + Text

Connect 🔻 🔗 ⚙️ 🔻

☰ 🔍 <> 📁

▾ ColabFold: AlphaFold2 w/ MMseqs2

• **21Aug2021: The MSA/Templates issues should now be resolved! Please report any errors you see.**

Easy to use AlphaFold2 ([Jumper et al. 2021](#)) protein structure prediction using multiple sequence alignments generated through an MMseqs2 API. For details, refer to our manuscript: [Mirdita M, Ovchinnikov S, Steinegger M. ColabFold - Making protein folding accessible to all. bioRxiv, 2021](#)

- This notebook provides basic functionality, for more advanced options (such as modeling heterocomplexes, increasing recycles, sampling, etc.) see our [advanced notebook](#).
- This notebook replaces the homology detection of AlphaFold2 with MMseqs2. For a comparison against the [Deepmind Colab](#) and the full [AlphaFold2](#) system read our [preprint](#).

For more details, see [bottom](#) of the notebook and checkout the [ColabFold GitHub](#).

▶ Input protein sequence, then hit Runtime -> Run all

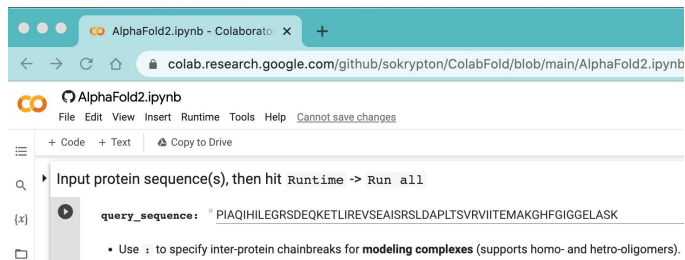
query_sequence: "PIAQIHILEGRSDEQKETLIREVSEAISSRLDAPLTSVRVIITEMAKGHFGIGGELASK"

jobname: "test"

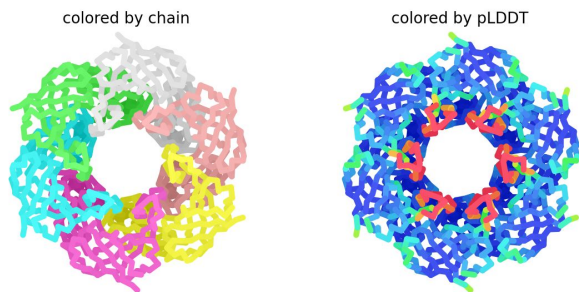
ColabFold

Making Protein folding accessible to all via Google Colab

<https://github.com/sokrypton/ColabFold>



<https://github.com/sokrypton/ColabFold>



Milot
Mirdita



Konstantin
Schütze



Yoshitaka
Moriwaki



Lim Heo

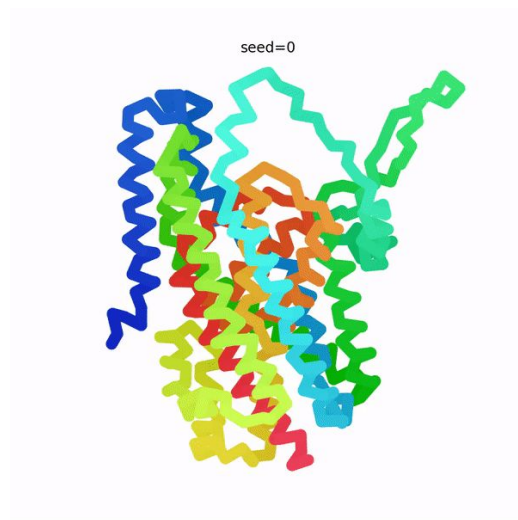


Martin
Steinegger

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M.
2022. ColabFold. *Nature Methods*

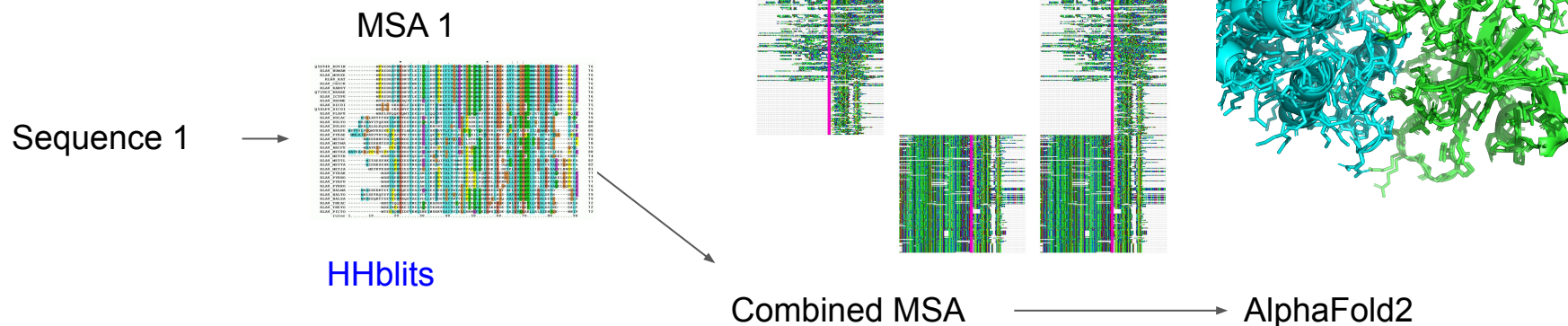
Things ColabFold can do:

- Fast MSA generation using MMseqs2 (minutes instead of hours)
- Model protein-protein complexes (using both AF and AF-multimer)
- Increase number of recycles
- Sample from the uncertainty (enable dropout, iterate through seeds, subsample MSA)
- Custom MSA/Template inputs



Notebooks	monomers	complexes	mmseqs2	jackhmmer	templates
AlphaFold2_mmseqs2	Yes	Yes	Yes	No	Yes
AlphaFold2_batch	Yes	Yes	Yes	No	Yes
AlphaFold2 (from Deepmind)	Yes	Yes	No	Yes	No
relax_amber (relax input structure)					
ESMFold	Yes	Maybe	No	No	No
BETA (in development) notebooks					
RoseTTAFold2	Yes	Yes	Yes	No	WIP
Boltz	Yes	Yes	Yes	No	No
BioEmu	Yes	No	Yes	No	No
OmegaFold	Yes	Maybe	No	No	No
AlphaFold2_advanced_v2 (new experimental notebook)	Yes	Yes	Yes	No	Yes

AlphaFold2: FoldDock

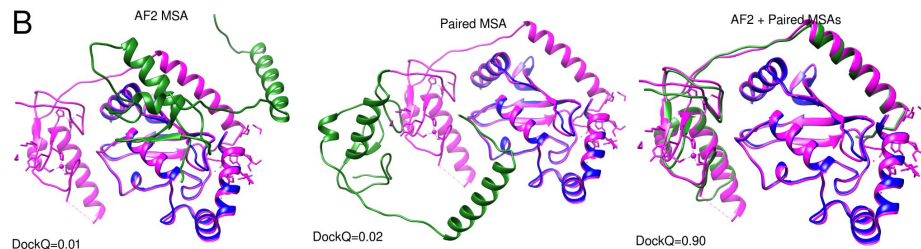
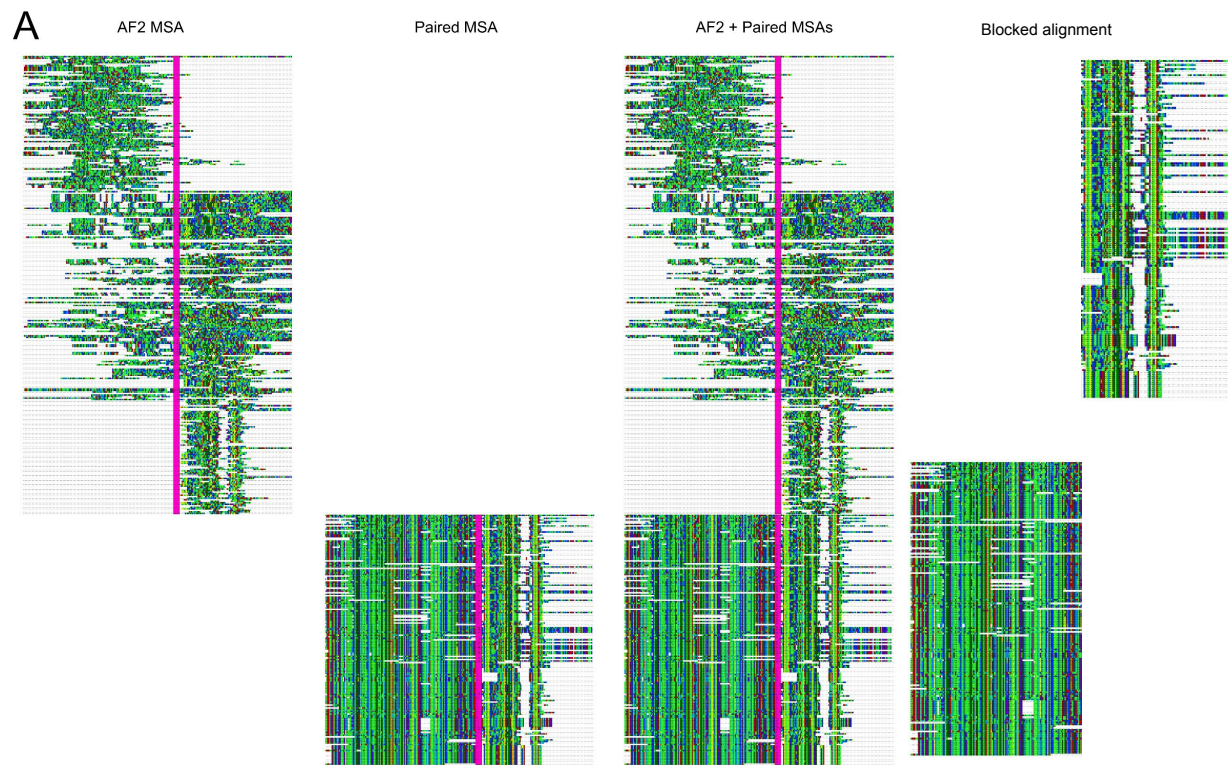


Improved prediction of protein-protein interactions using AlphaFold2 and extended multiple-sequence alignments

P. Bryant, G. Pozzati, A. Elofsson

doi: <https://doi.org/10.1101/2021.09.15.460468>

Importance of the Multiple sequence alignments in prediction of Protein-protein interactions

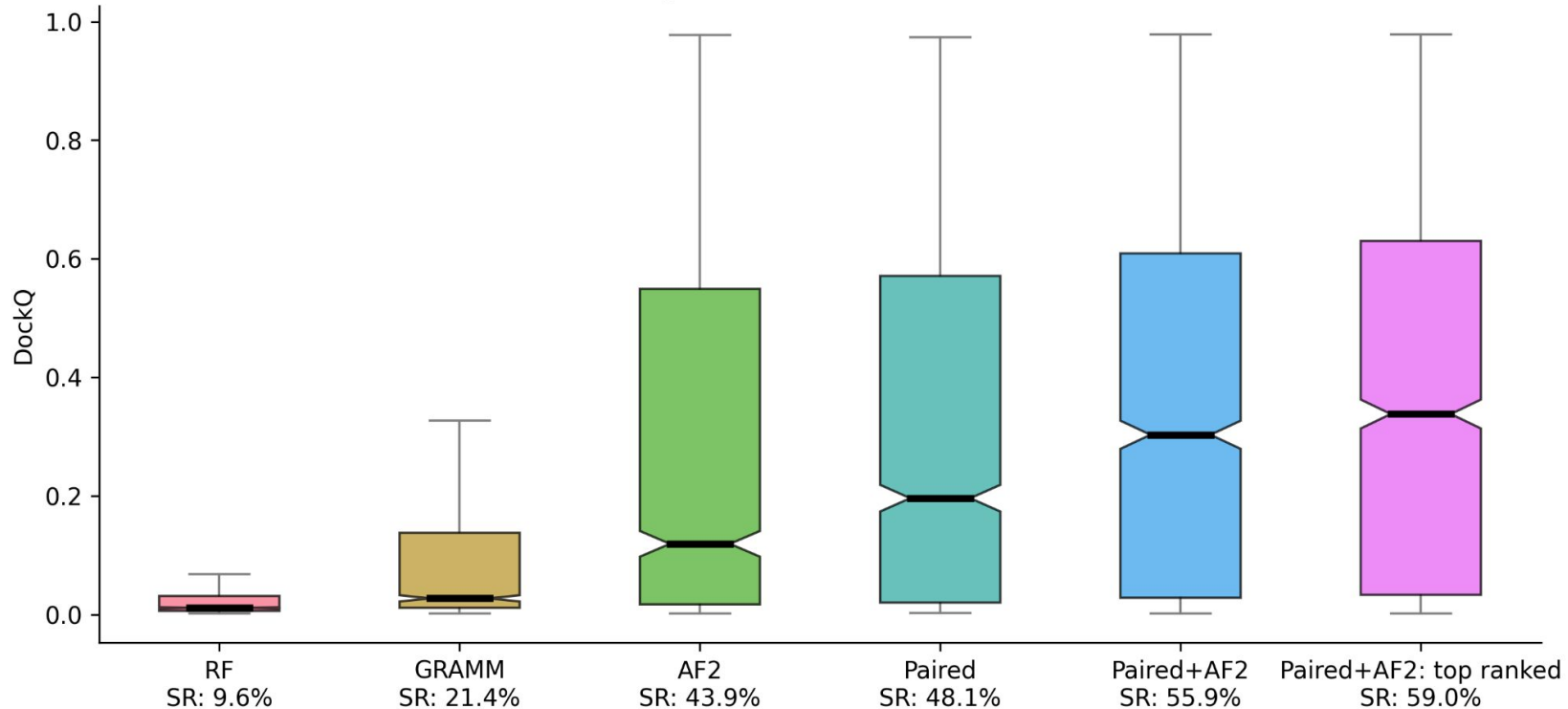


60% of protein-protein interactions are correct



A

DockQ scores for test set



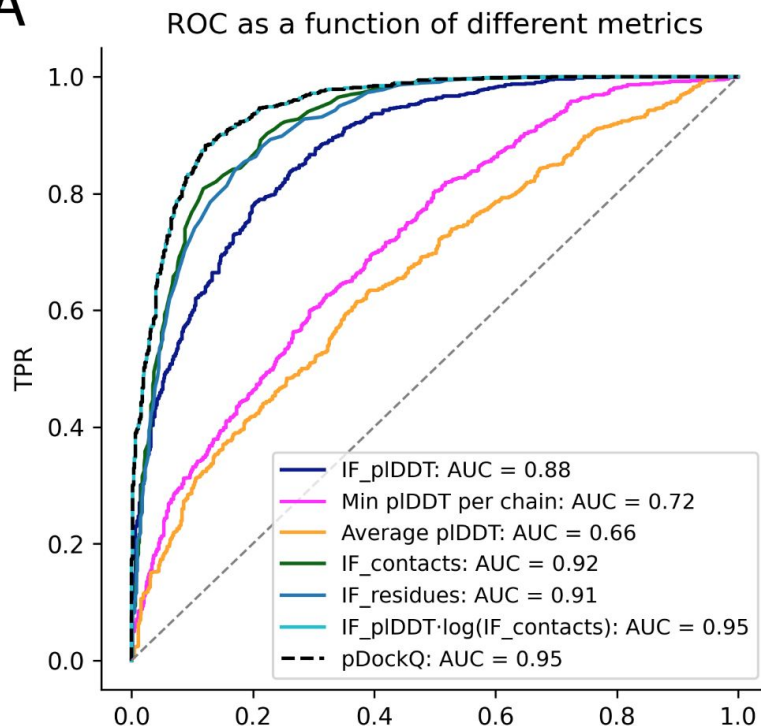
pDockQ is a good measure for AF2.0



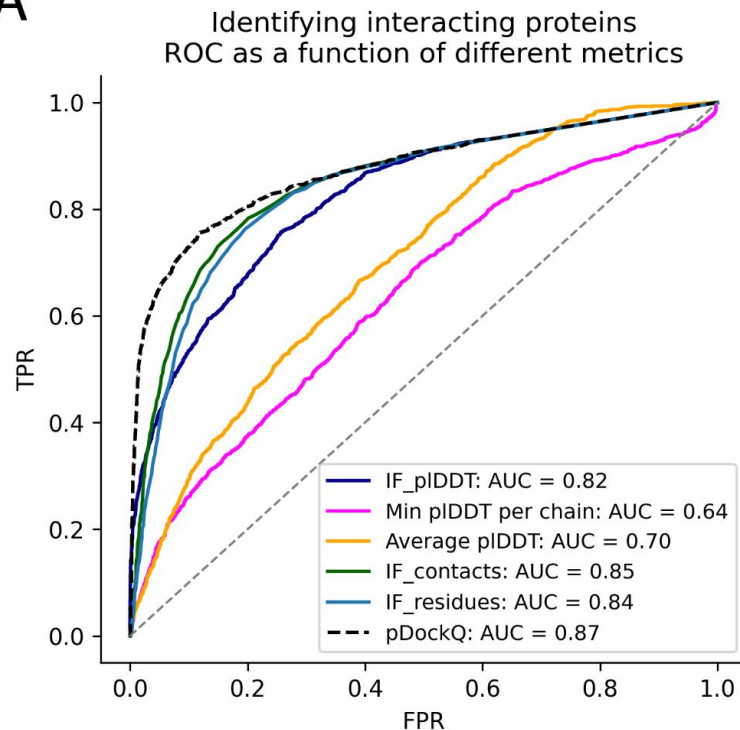
Good/bad models (dockQ>0.23)

Binding/non-binding proteins

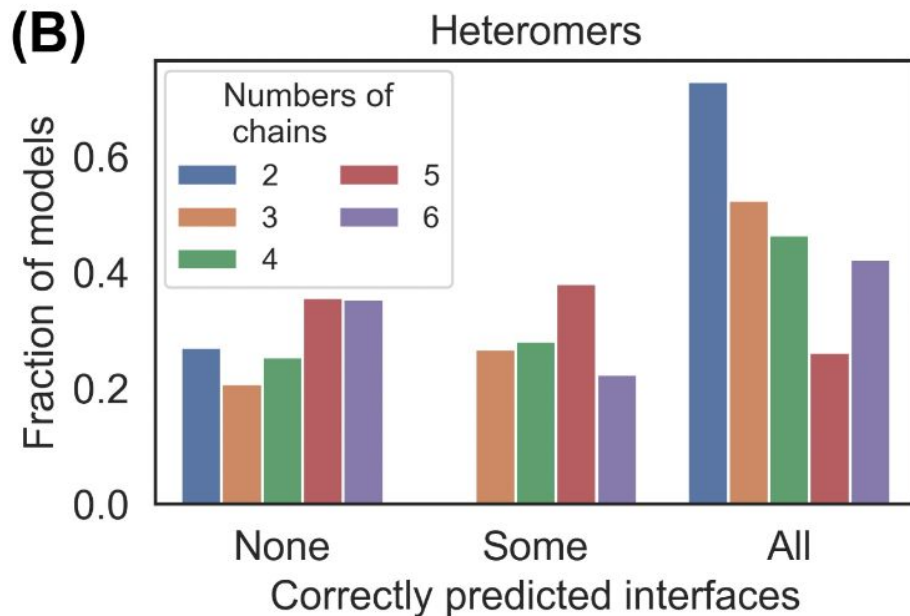
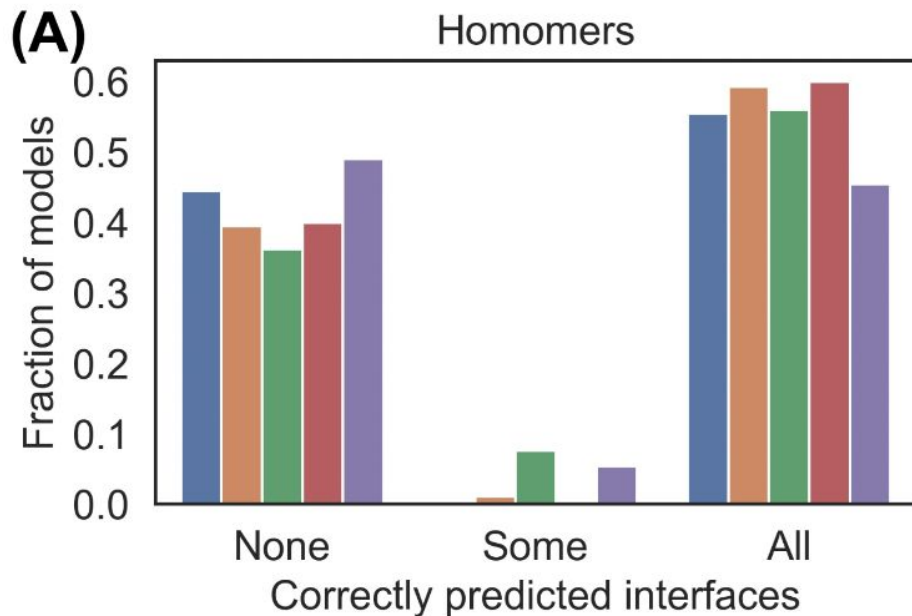
A



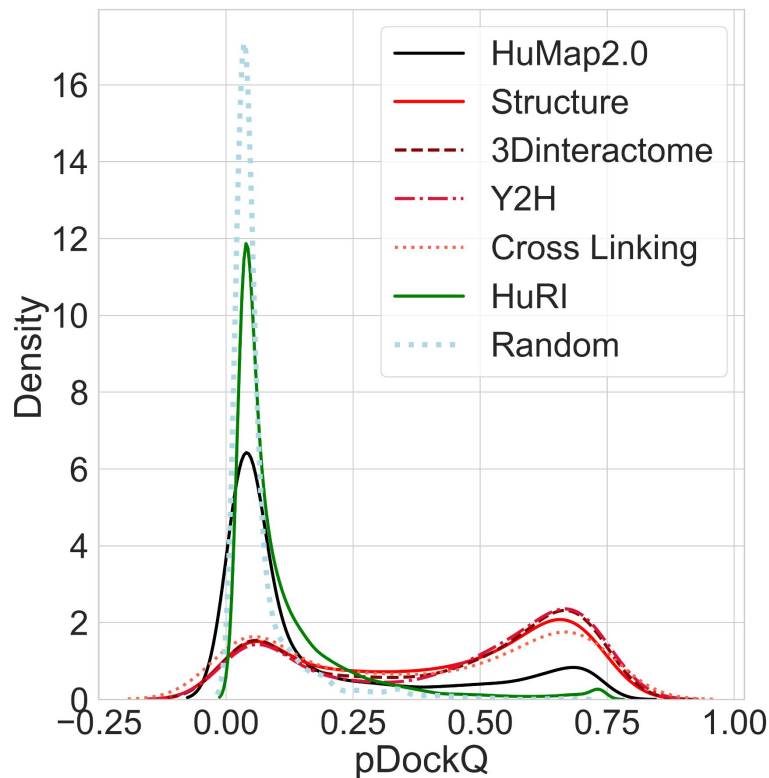
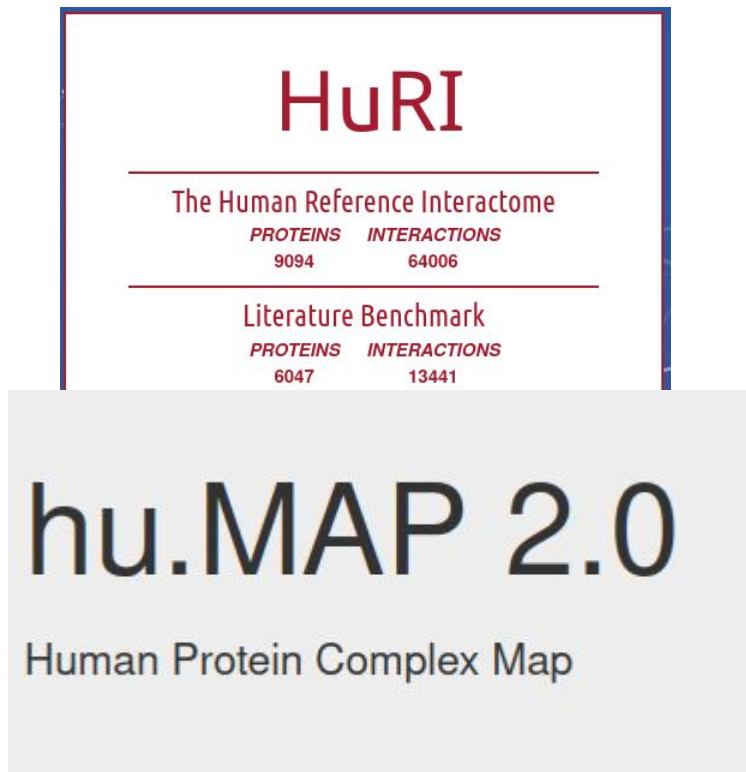
A



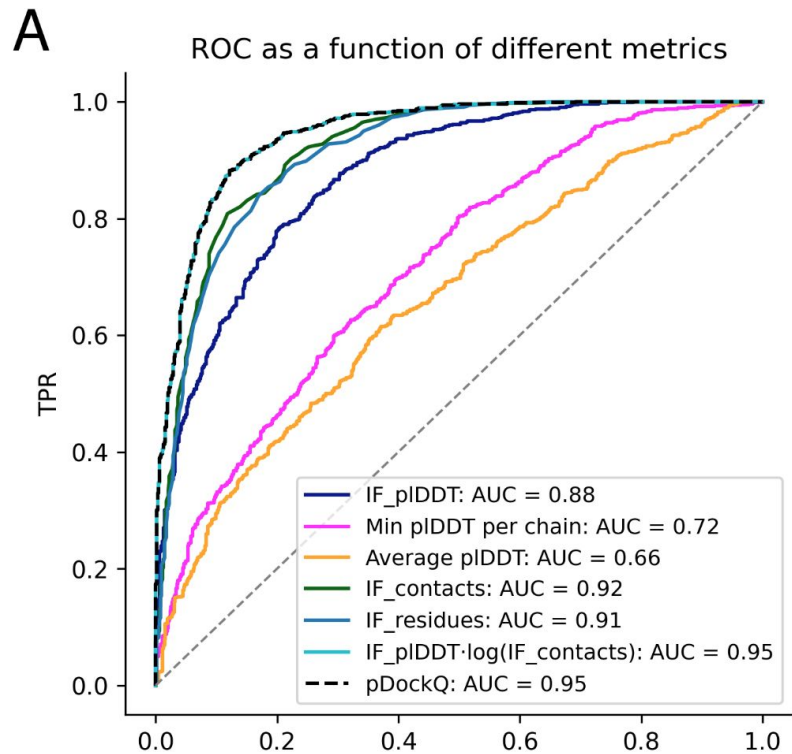
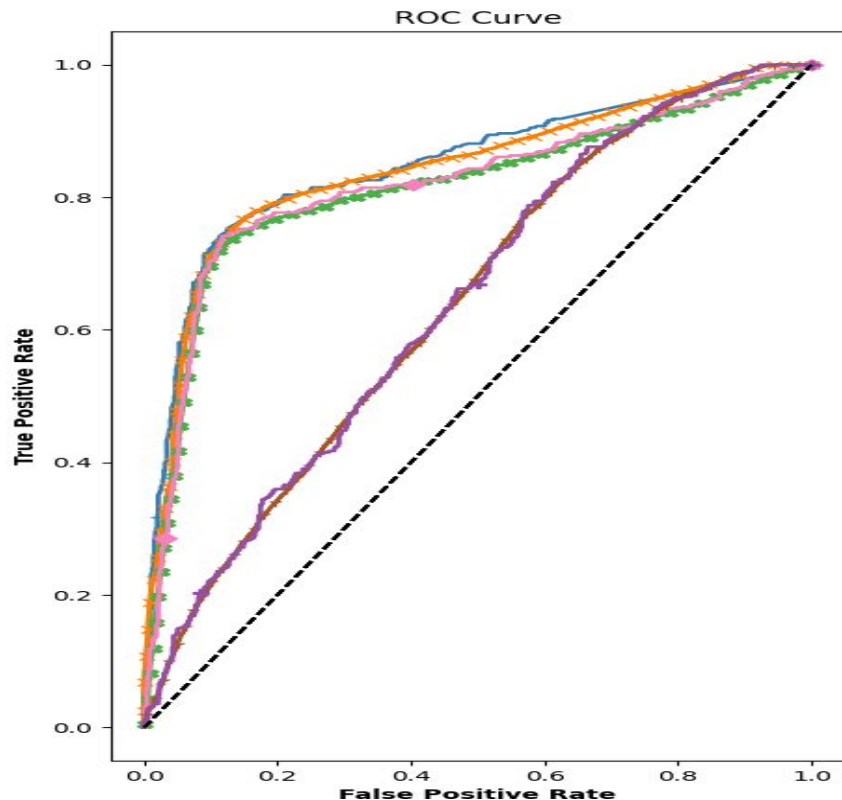
All or none for homomers



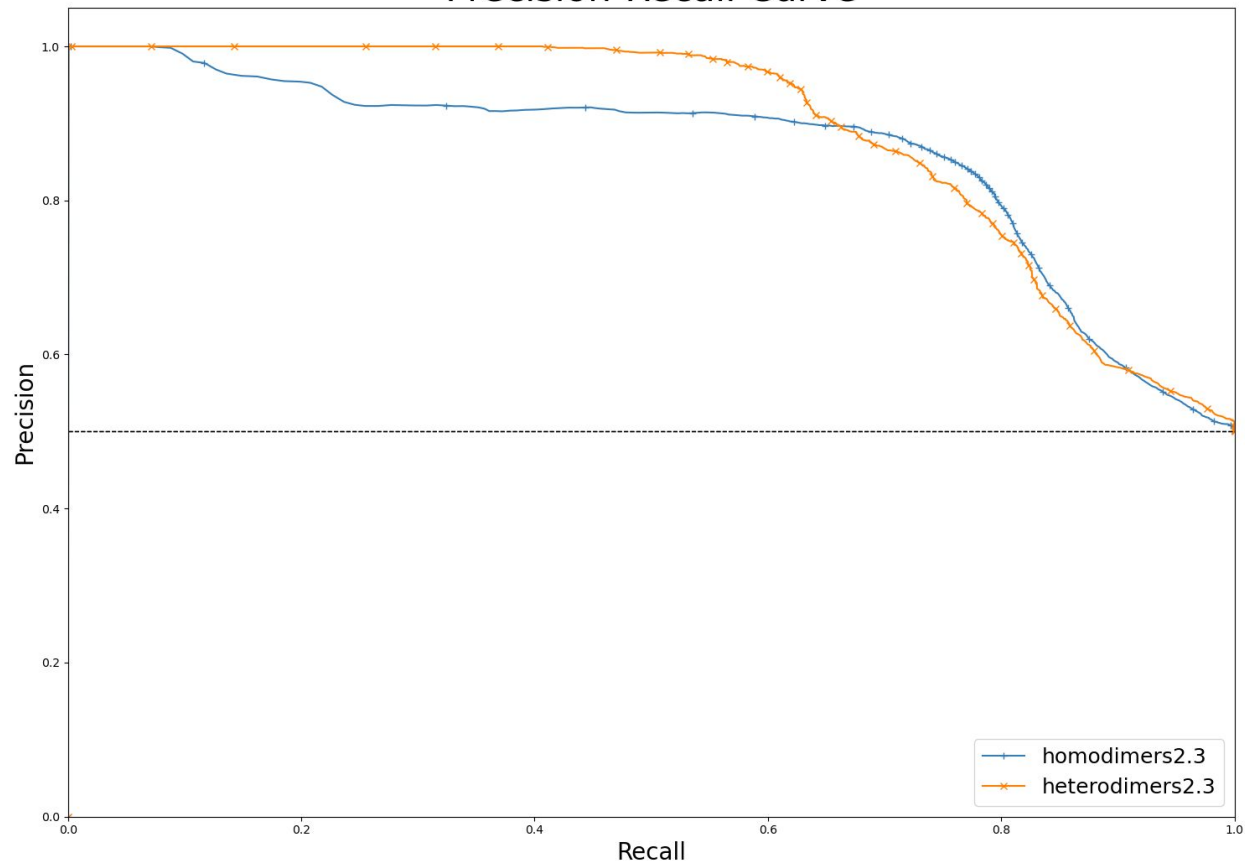
HuMap (fractionation) vs HuRI (Yeast 2 Hybrid)



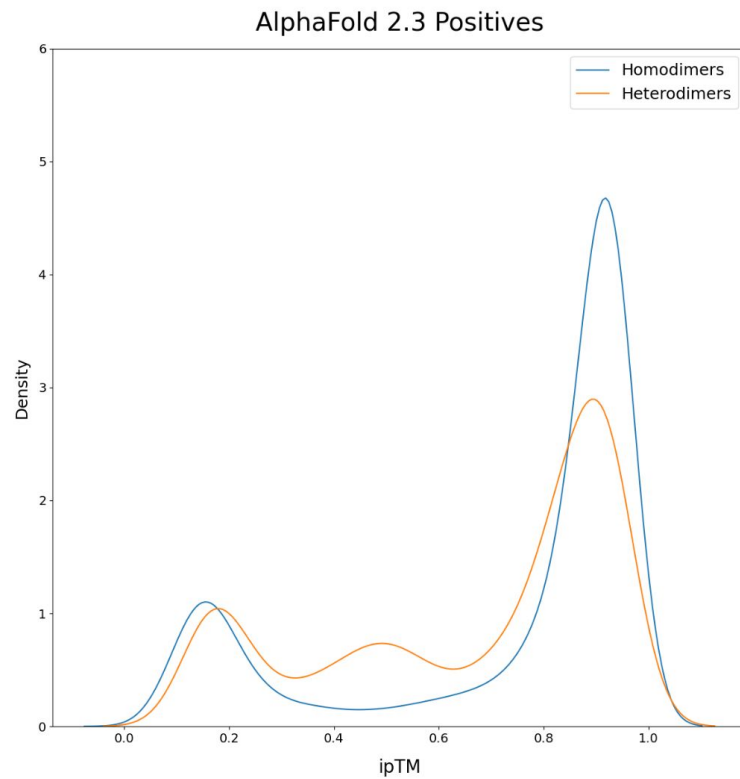
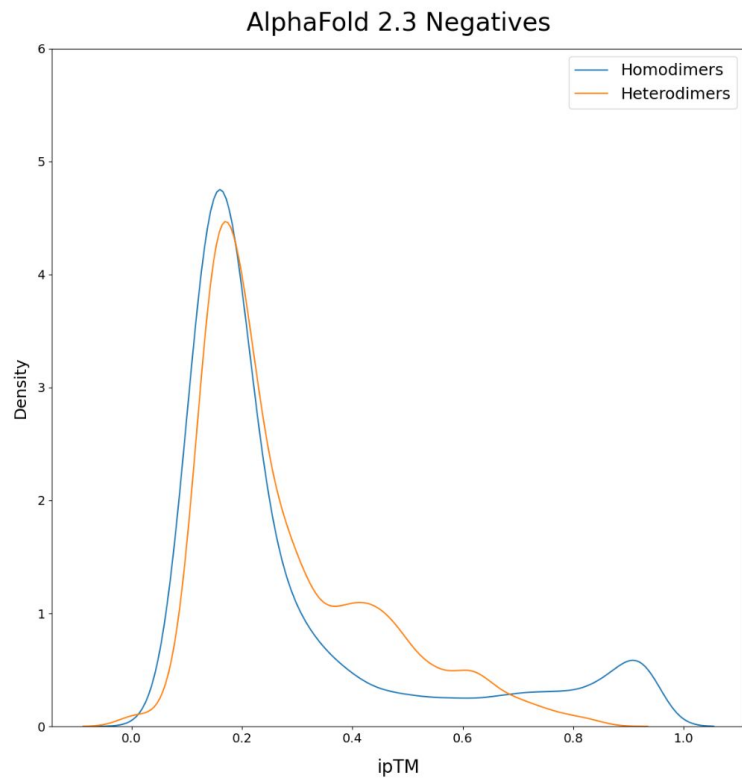
Homomers are harder to identify



Precision-Recall Curve



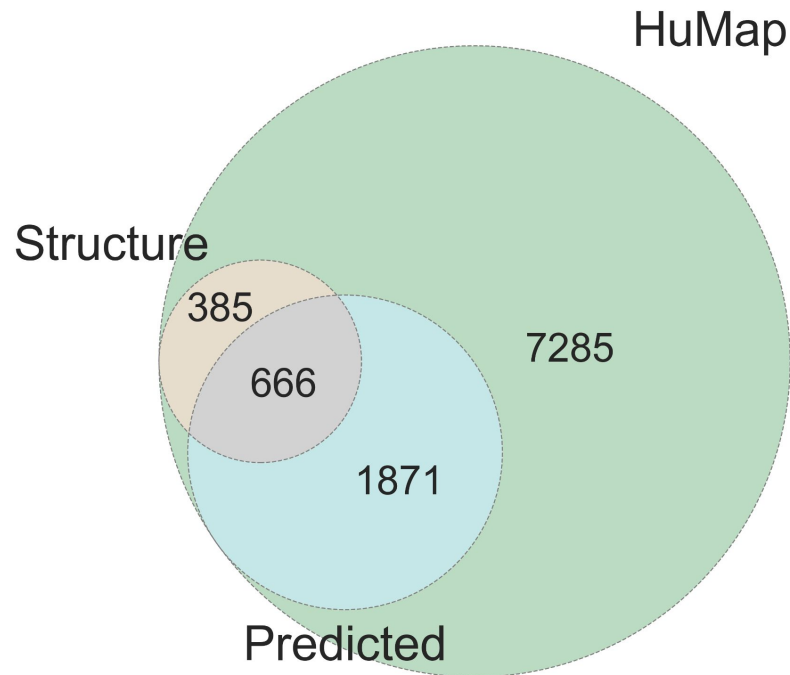
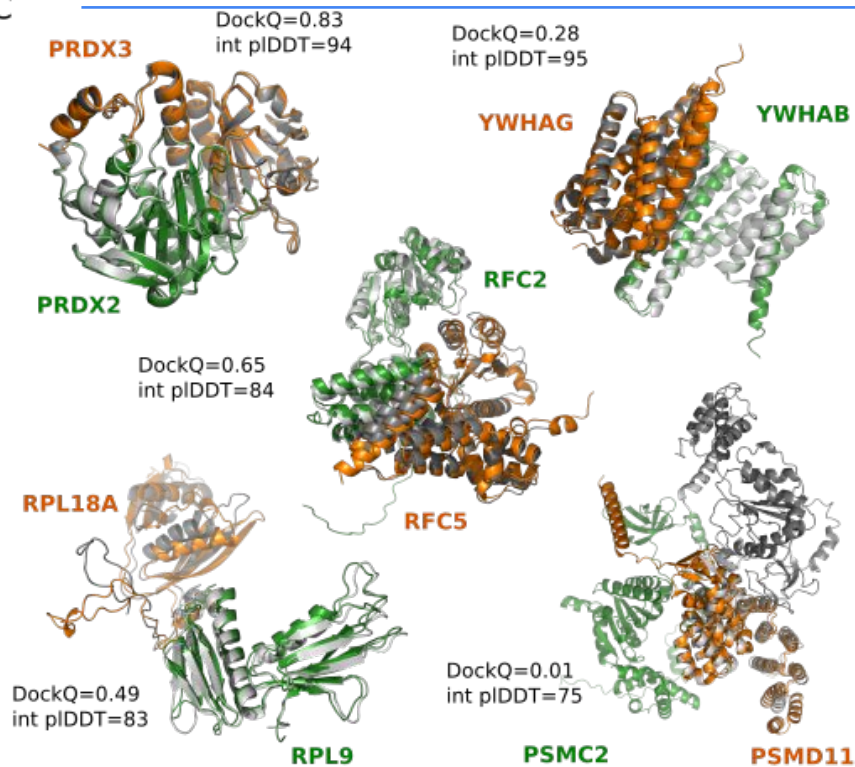
Homodimer ipTM



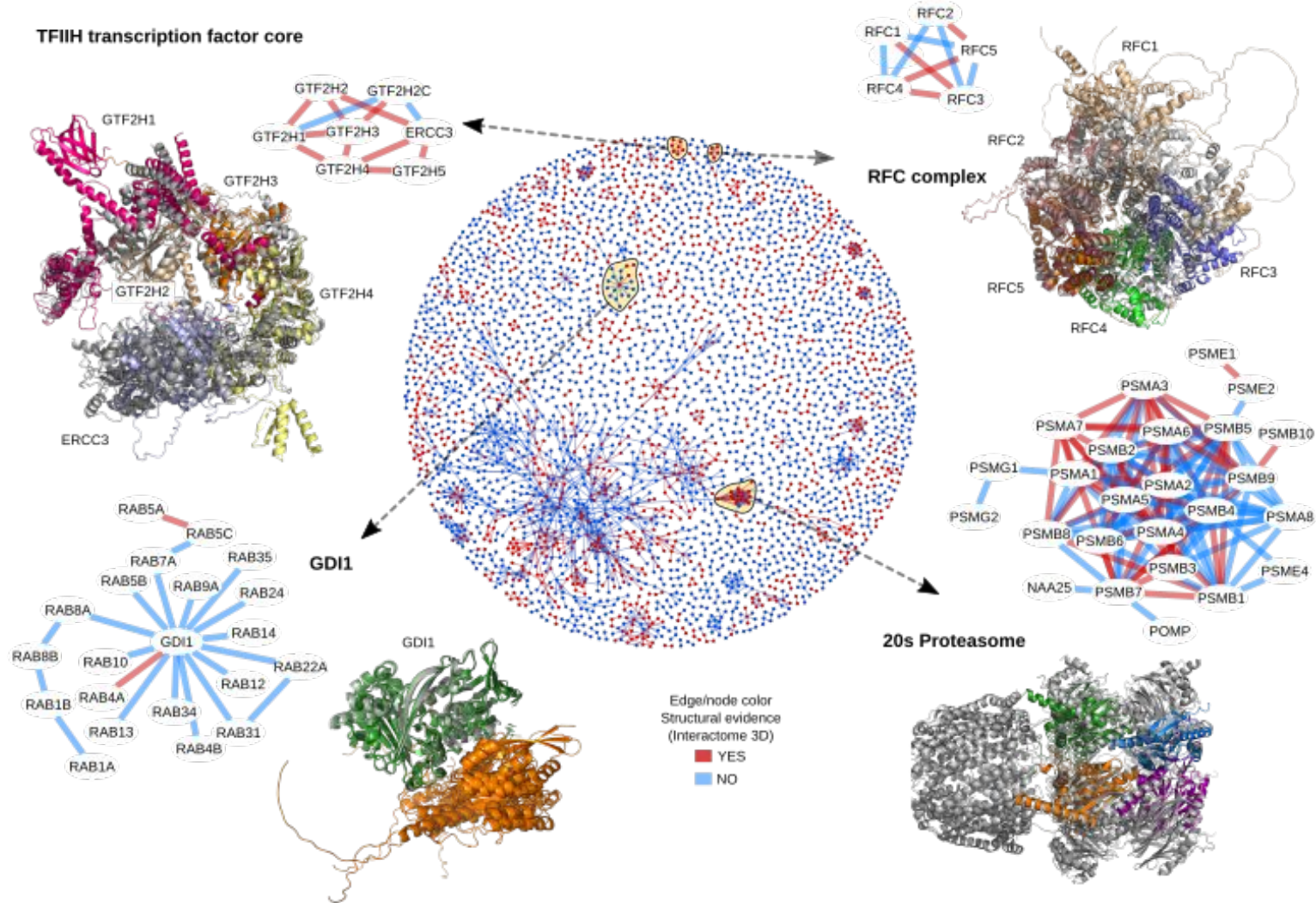
The Human interactome



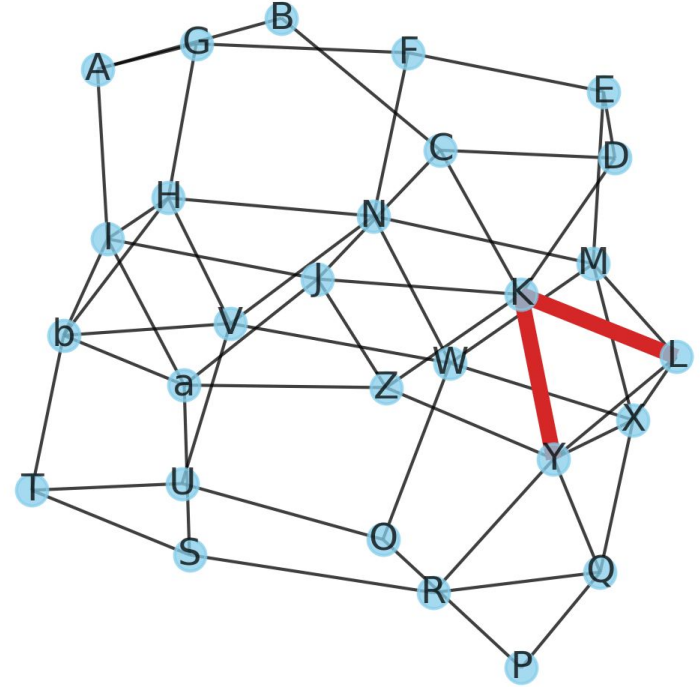
C



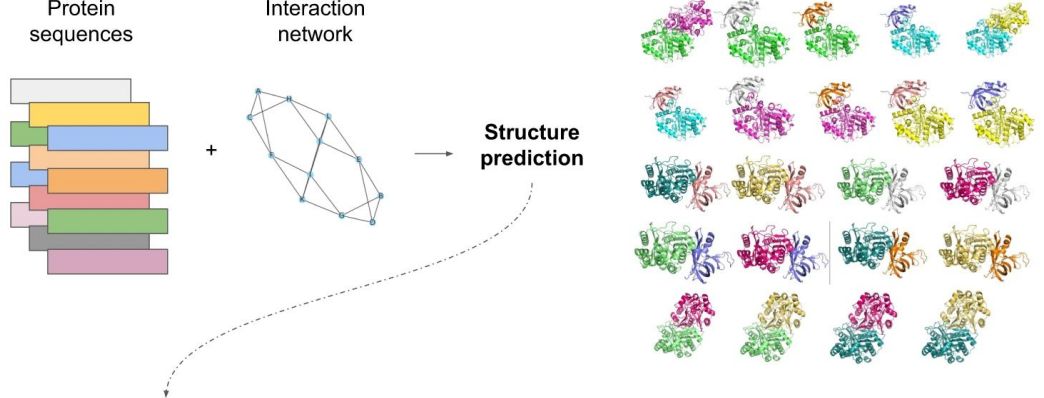
Interaction network



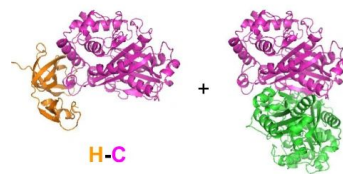
20S Proteasome



Predicting the structure of large complexes



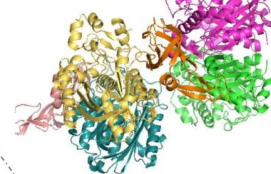
Assembly



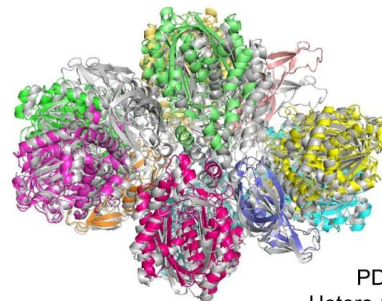
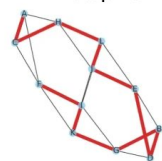
Step 1



Step 4

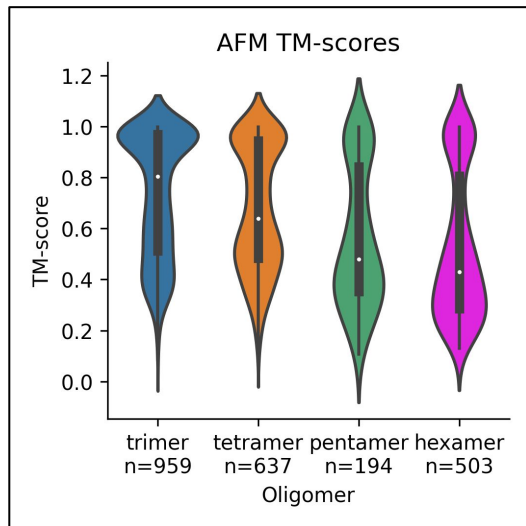


Step 10

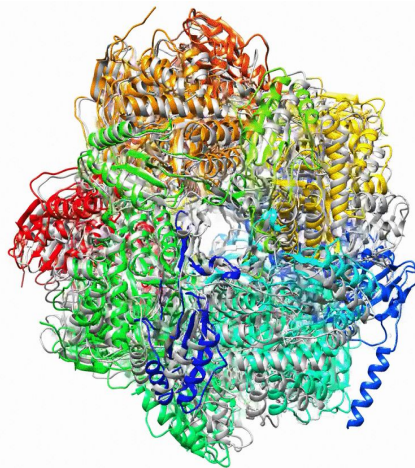


90 dg.
0

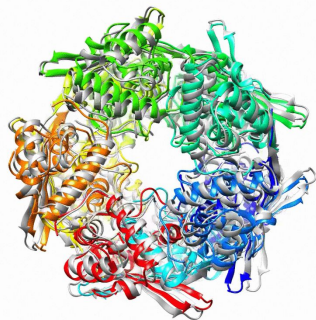
PDB ID: 6ESQ
Hetero 12-mer - A4B4C4
Dihedral - D2
TM-score=0.93



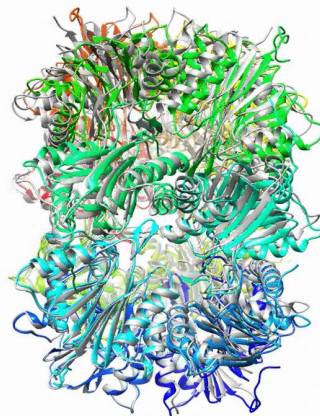
1IWA, A8B8
TM-score=0.93



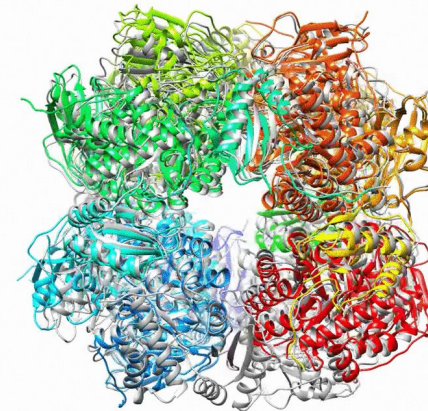
1DW9, A10
TM-score=0.96



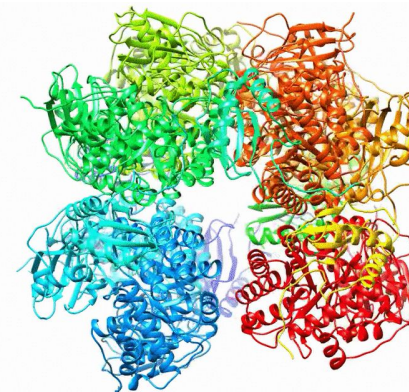
5OVS, A14
TM-score=0.99



1RBL, A8B8
TM-score=0.96

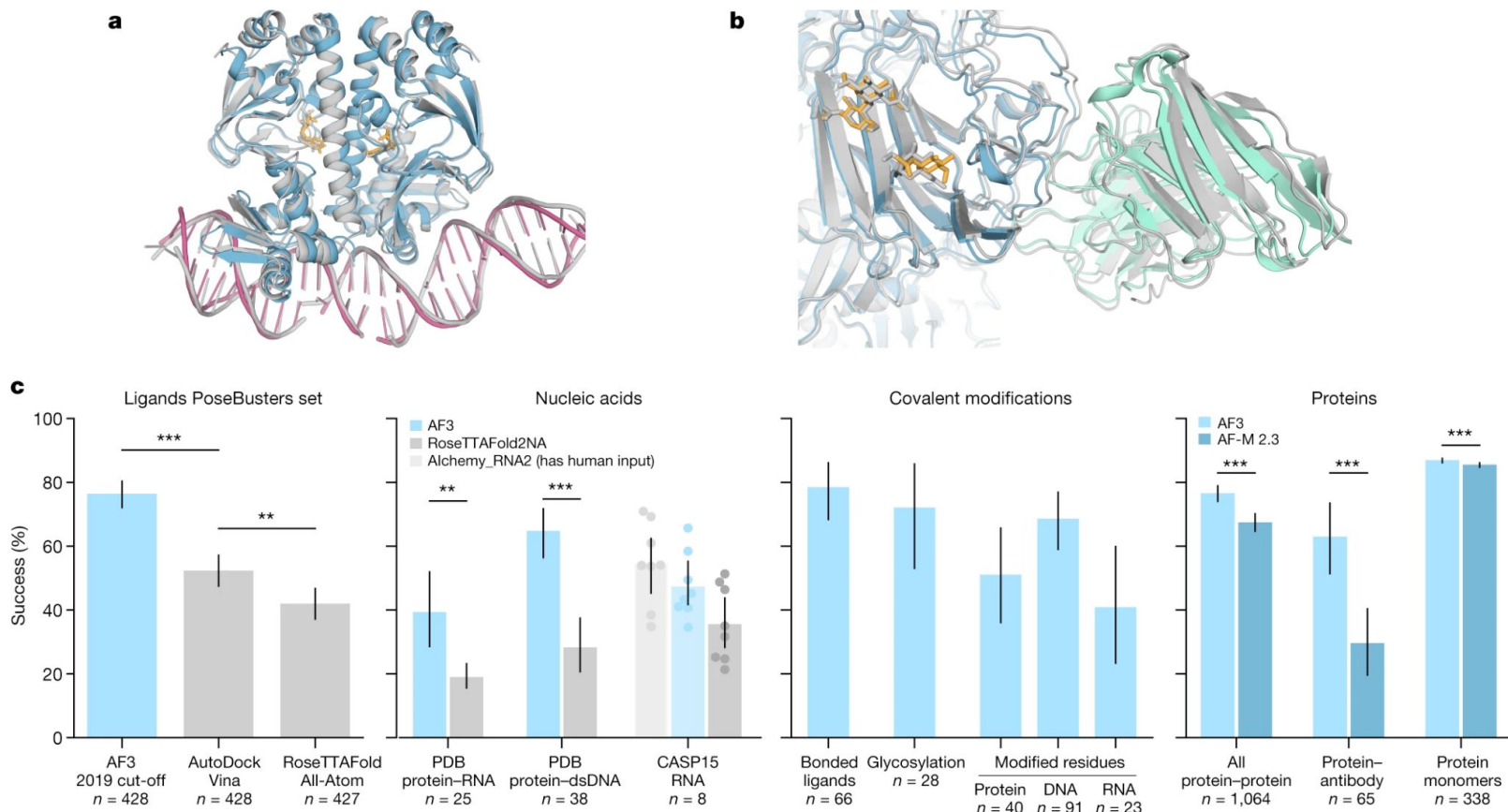


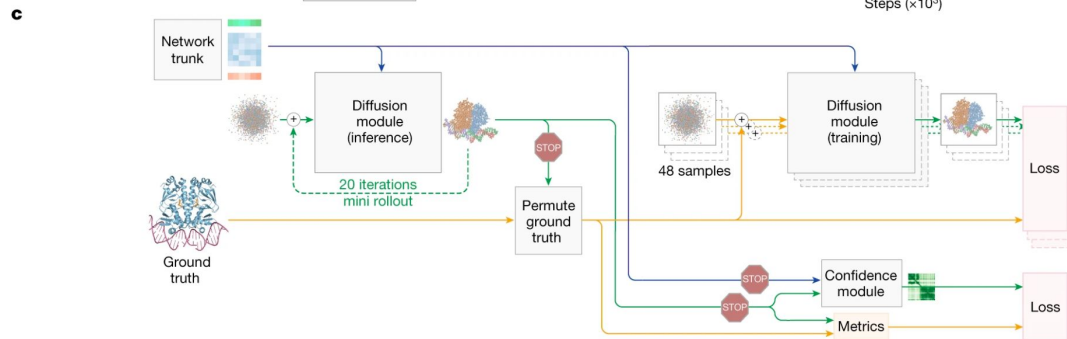
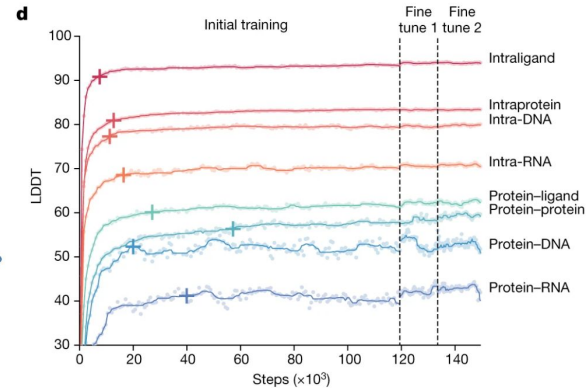
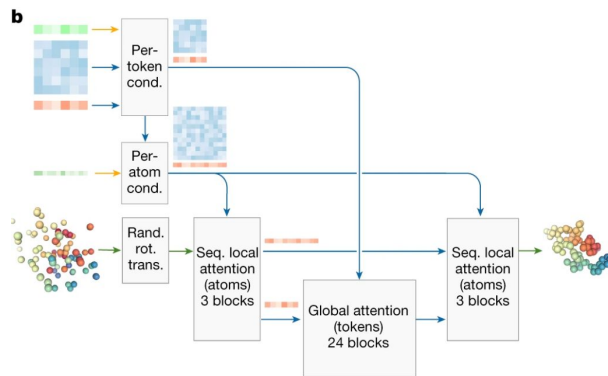
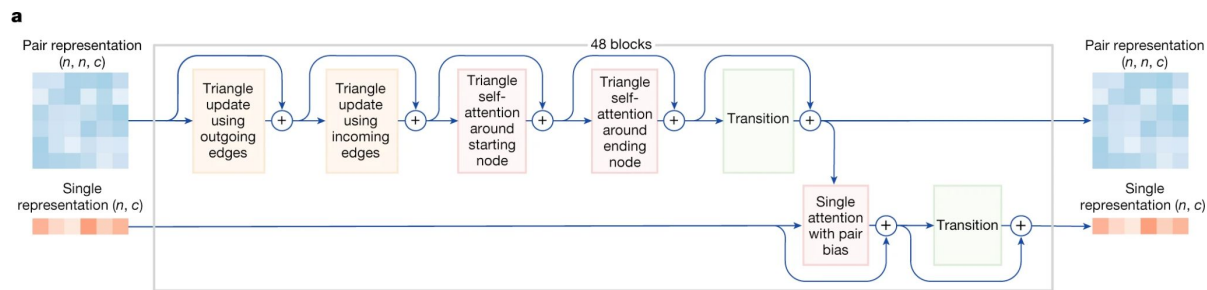
1RBL prediction
alone, 15/16 chains



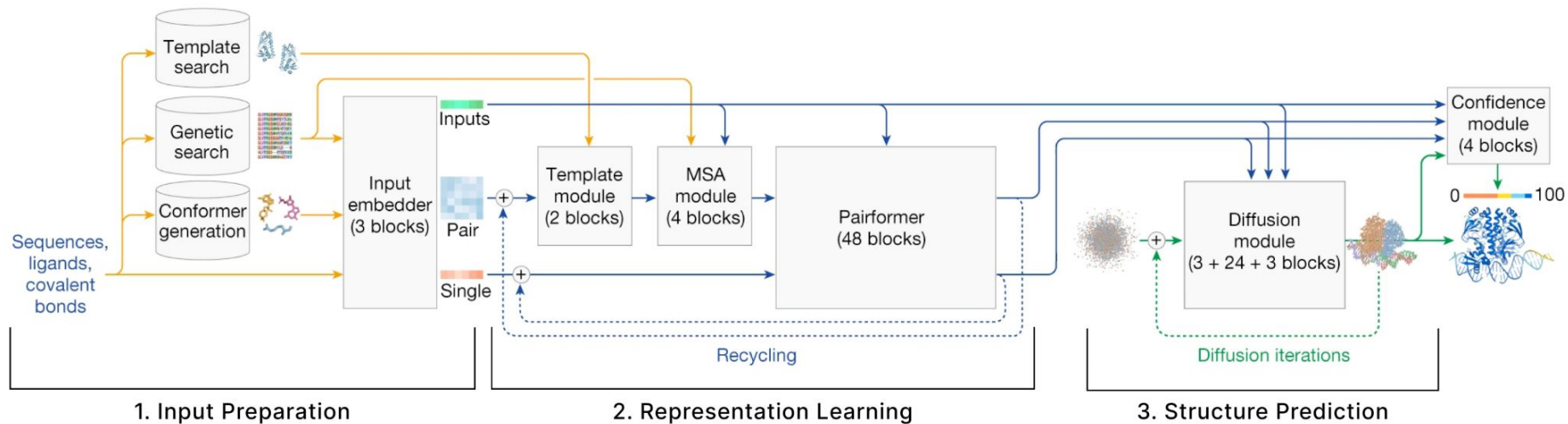
23/175 complexes are good
(TM-score>0.9 and complete)

AlphaFold3



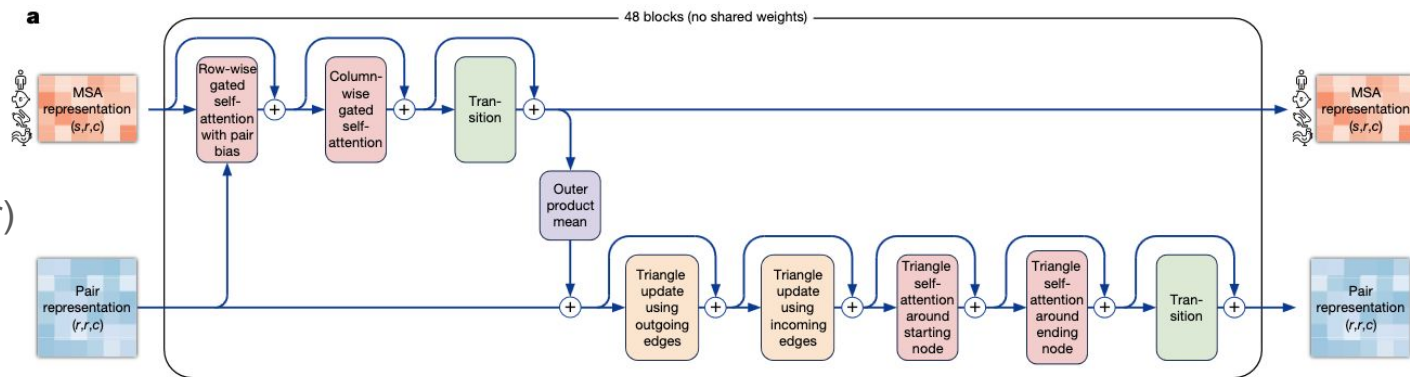


AF3

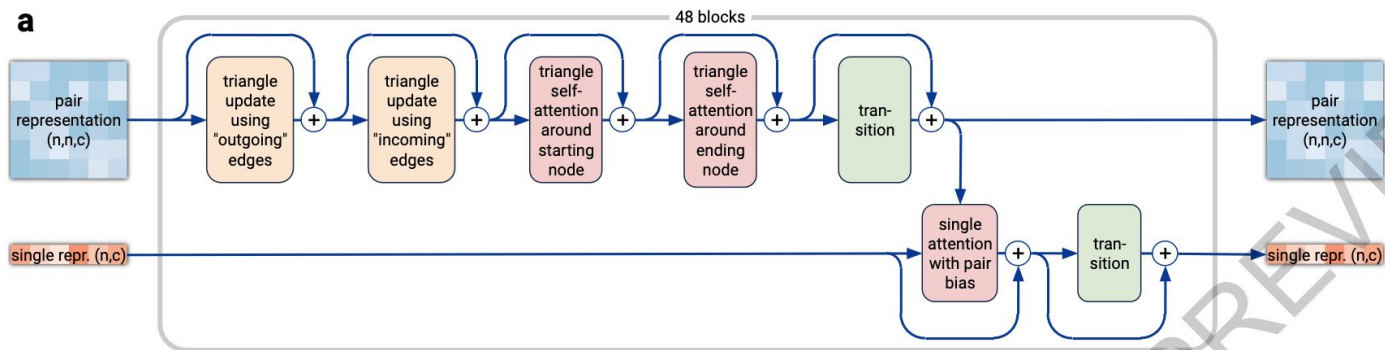


Pairformer

AF2
(Evoformer)



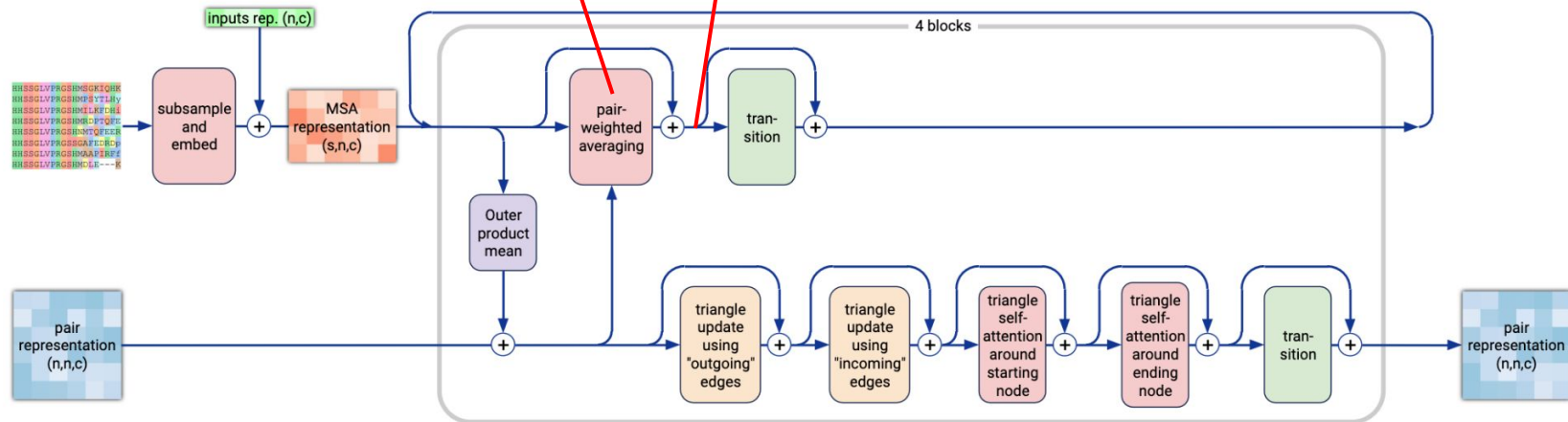
AF3



MSA module

(AF2) Row-wise attention with bias

Column global attention



Q: In Pairformer, we removed the outer product mean that the Evoformer had, meaning there is no information flow from the single to the pair representation. While in Supp Fig. 2 it is clear that the outer product mean is still here and still flowing information from the single to the pair representation? (and the algorithm is also there, numbered 9).

Q: In AF3, they removed the Column Attention, any comment on that?

MSA module

AF2

Algorithm 7 MSA row-wise gated self-attention with pair bias

def MSARowAttentionWithPairBias($\{\mathbf{m}_{si}\}, \{\mathbf{z}_{ij}\}, c = 32, N_{\text{head}} = 8$) :

Input projections

- 1: $\mathbf{m}_{si} \leftarrow \text{LayerNorm}(\mathbf{m}_{si})$
- 2: $\mathbf{q}_{si}^h, \mathbf{k}_{si}^h, \mathbf{v}_{si}^h = \text{LinearNoBias}(\mathbf{m}_{si})$ $\mathbf{q}_{si}^h, \mathbf{k}_{si}^h$
- 3: $b_{ij}^h = \text{LinearNoBias}(\text{LayerNorm}(\mathbf{z}_{ij}))$
- 4: $\mathbf{g}_{si}^h = \text{sigmoid}(\text{Linear}(\mathbf{m}_{si}))$

Attention

- 5: $a_{ij}^h = \text{softmax}_j \left(\frac{1}{\sqrt{c}} \mathbf{q}_{si}^{h\top} \mathbf{k}_{sj}^h + b_{ij}^h \right)$
- 6: $\mathbf{o}_{si}^h = \mathbf{g}_{si}^h \odot \sum_j a_{sij}^h \mathbf{v}_{sj}^h$

Output projection

- 7: $\tilde{\mathbf{m}}_{si} = \text{Linear}(\text{concat}_h(\mathbf{o}_{si}^h))$
 - 8: **return** $\{\tilde{\mathbf{m}}_{si}\}$
-

AF3

Algorithm 10 MSA pair weighted averaging with gating

def MSAPairWeightedAveraging($\{\mathbf{m}_{si}\}, \{\mathbf{z}_{ij}\}, c = 32, N_{\text{head}} = 8$) :

Input projections

- 1: $\mathbf{m}_{si} \leftarrow \text{LayerNorm}(\mathbf{m}_{si})$
- 2: $\mathbf{v}_{si}^h = \text{LinearNoBias}(\mathbf{m}_{si})$
- 3: $b_{ij}^h = \text{LinearNoBias}(\text{LayerNorm}(\mathbf{z}_{ij}))$
- 4: $\mathbf{g}_{si}^h = \text{sigmoid}(\text{LinearNoBias}(\mathbf{m}_{si}))$

Weighted average with gating

- 5: $w_{ij}^h = \text{softmax}_j(b_{ij}^h)$
- 6: $\mathbf{o}_{si}^h = \mathbf{g}_{si}^h \odot \sum_j w_{ij}^h \mathbf{v}_{sj}^h$

Output projection

- 7: $\tilde{\mathbf{m}}_{si} = \text{LinearNoBias}(\text{concat}_h(\mathbf{o}_{si}^h))$
 - 8: **return** $\{\tilde{\mathbf{m}}_{si}\}$
-

Q: Why did they change some ReLU to SwiGLU (Transition blocks)?

MSA pairing

- In AlphaFold-Multimer, **unpaired MSA** sequences across chains were presented to the network in a **block-diagonal fashion**.
- In AlphaFold 3 these “unpaired” sequences are presented in a **dense fashion** - so that the network sees many more MSA sequences per-chain.



Q: - Can you please clarify the processing of unpaired sequences for complex prediction?

Q: "Unlike in AlphaFold-Multimer, the rest of the MSA is constructed in a dense fashion", can you explain this?

Q: Can you explain “Weighted PDB dataset in SI 2.5.1” ? What is “chain-pair interfaces”?

Atom encoder

Algorithm 5 Atom attention encoder

def AtomAttentionEncoder($\{\mathbf{f}^*\}, \{\mathbf{r}_l\}, \{\mathbf{s}_i^{\text{trunk}}\}, \{\mathbf{z}_{ij}\}, c_{\text{atom}}, c_{\text{atompair}}, c_{\text{token}} \)$:

Create the atom single conditioning: Embed per-atom meta data

1: $\mathbf{c}_l = \text{LinearNoBias}(\text{concat}(\tilde{\mathbf{f}}_l^{\text{ref_pos}}, \mathbf{f}_l^{\text{ref_charge}}, \mathbf{f}_l^{\text{ref_mask}}, \mathbf{f}_l^{\text{ref_element}}, \mathbf{f}_l^{\text{ref_atom_name_chars}}))$

Embed offsets between atom reference positions

2: $\tilde{\mathbf{d}}_{lm} = \tilde{\mathbf{f}}_l^{\text{ref_pos}} - \tilde{\mathbf{f}}_m^{\text{ref_pos}}$

3: $v_{lm} = (\mathbf{f}_l^{\text{ref_space_uid}} - \mathbf{f}_m^{\text{ref_space_uid}})$

4: $\mathbf{p}_{lm} = \text{LinearNoBias}(\tilde{\mathbf{d}}_{lm}) \cdot v_{lm}$

Embed pairwise inverse squared distances, and the valid mask.

5: $\mathbf{p}_{lm} += \text{LinearNoBias}\left(\frac{1}{1 + \|\tilde{\mathbf{d}}_{lm}\|^2}\right) \cdot v_{lm}$

6: $\mathbf{p}_{lm} += \text{LinearNoBias}(v_{lm}) \cdot v_{lm}$

Initialise the atom single representation as the single conditioning.

7: $\mathbf{q}_l = \mathbf{c}_l$

$\mathbf{c}_l \in \mathbb{R}^{c_{\text{atom}}}$
 $l \in \{1, \dots, N_{\text{atoms}}\}$

$\tilde{\mathbf{d}}_{lm} \in \mathbb{R}^3$

$v_{lm} \in \mathbb{R}$

$\mathbf{p}_{lm} \in \mathbb{R}^{c_{\text{atompair}}}$

$\mathbf{q}_l \in \mathbb{R}^{c_{\text{atom}}}$

If provided, add trunk embeddings and noisy positions.

8: **if** $\{\mathbf{r}_l\} \neq \emptyset$ **then**

Broadcast the single and pair embedding from the trunk.

9: $\mathbf{c}_l += \text{LinearNoBias}(\text{LayerNorm}(\mathbf{s}_{\text{tok_idx}(l)}^{\text{trunk}}))$

10: $\mathbf{p}_{lm} += \text{LinearNoBias}(\text{LayerNorm}(\mathbf{z}_{\text{tok_idx}(l) \text{ tok_idx}(m)}))$

Add the noisy positions.

11: $\mathbf{q}_l += \text{LinearNoBias}(\mathbf{r}_l)$

12: **end if**

Add the combined single conditioning to the pair representation.

13: $\mathbf{p}_{lm} += \text{LinearNoBias}(\text{relu}(\mathbf{c}_l)) + \text{LinearNoBias}(\text{relu}(\mathbf{c}_m))$

Run a small MLP on the pair activations.

14: $\mathbf{p}_{lm} += \text{LinearNoBias}(\text{relu}(\text{LinearNoBias}(\text{relu}(\text{LinearNoBias}(\text{relu}(\mathbf{p}_{lm}))))))$

Cross attention transformer.

15: $\{\mathbf{q}_l\} = \text{AtomTransformer}(\{\mathbf{q}_l\}, \{\mathbf{c}_l\}, \{\mathbf{p}_{lm}\}, N_{\text{block}} = 3, N_{\text{head}} = 4)$

Aggregate per-atom representation to per-token representation

16: $\mathbf{a}_i = \text{mean}_{\substack{l \in \{1, \dots, N_{\text{atoms}}\} \\ \text{tok_idx}(l) = i}}(\text{relu}(\text{LinearNoBias}(\mathbf{q}_l)))$

$\mathbf{a}_i \in \mathbb{R}^{c_{\text{token}}}$

17: $\mathbf{q}_l^{\text{skip}}, \mathbf{c}_l^{\text{skip}}, \mathbf{p}_{lm}^{\text{skip}} = \mathbf{q}_l, \mathbf{c}_l, \mathbf{p}_{lm}$

18: **return** $\{\mathbf{a}_i\}, \{\mathbf{q}_l^{\text{skip}}\}, \{\mathbf{c}_l^{\text{skip}}\}, \{\mathbf{p}_{lm}^{\text{skip}}\}$

Atom decoder

Algorithm 6 Atom attention decoder

def AtomAttentionDecoder($\{\mathbf{a}_i\}, \{\mathbf{q}_l^{\text{skip}}\}, \{\mathbf{c}_l^{\text{skip}}\}, \{\mathbf{p}_{lm}^{\text{skip}}\}$) :

Broadcast per-token activations to per-atom activations and add the skip connection

1: $\mathbf{q}_l = \text{LinearNoBias}(\mathbf{a}_{\text{tok_idx}(l)}) + \mathbf{q}_l^{\text{skip}}$

Cross attention transformer.

2: $\{\mathbf{q}_l\} = \text{AtomTransformer}(\{\mathbf{q}_l\}, \{\mathbf{c}_l^{\text{skip}}\}, \{\mathbf{p}_{lm}^{\text{skip}}\}, N_{\text{block}} = 3, N_{\text{head}} = 4)$

Map to positions update.

3: $\mathbf{r}_l^{\text{update}} = \text{LinearNoBias}(\text{LayerNorm}(\mathbf{q}_l))$

4: **return** $\{\mathbf{r}_l^{\text{update}}\}$

Q: - Are the AtomAttention encoder and decoder acting as downscaling and upscaling networks, from atoms to tokens and back?

Diffusion

Algorithm 18 Sample Diffusion

def SampleDiffusion($\{\mathbf{f}^*\}$, $\{\mathbf{s}_i^{\text{inputs}}\}$, $\{\mathbf{s}_i^{\text{trunk}}\}$, $\{\mathbf{z}_{ij}^{\text{trunk}}\}$, Noise Schedule $[c_0, c_1, \dots, c_T]$,
 $\gamma_0 = 0.8$, $\gamma_{\min} = 1.0$, noise scale $\lambda = 1.003$, step scale $\eta = 1.5$):

- 1: $\vec{\mathbf{x}}_l \sim c_0 \cdot \mathcal{N}(\vec{\mathbf{0}}, \mathbf{I}_3)$ $\vec{\mathbf{x}}_l \in \mathbb{R}^3$
 - 2: **for all** $c_\tau \in [c_1, \dots, c_T]$ **do**
 - 3: $\{\vec{\mathbf{x}}_l\} \leftarrow \text{CentreRandomAugmentation}(\{\vec{\mathbf{x}}_l\})$
 - 4: $\gamma = \gamma_0$ if $c_\tau > \gamma_{\min}$ else 0
 - 5: $\hat{t} = c_{\tau-1}(\gamma + 1)$
 - 6: $\vec{\xi}_l = \lambda \sqrt{\hat{t}^2 - c_{\tau-1}^2} \cdot \mathcal{N}(\vec{\mathbf{0}}, \mathbf{I}_3)$ $\vec{\xi}_l \in \mathbb{R}^3$
 - 7: $\vec{\mathbf{x}}_l^{\text{noisy}} = \vec{\mathbf{x}}_l + \vec{\xi}_l$
 - 8: $\{\vec{\mathbf{x}}_l^{\text{denoised}}\} = \text{DiffusionModule}(\{\vec{\mathbf{x}}_l^{\text{noisy}}\}, \hat{t}, \{\mathbf{f}^*\}, \{\mathbf{s}_i^{\text{inputs}}\}, \{\mathbf{s}_i^{\text{trunk}}\}, \{\mathbf{z}_{ij}^{\text{trunk}}\})$
 - 9: $\vec{\delta}_l = (\vec{\mathbf{x}}_l - \vec{\mathbf{x}}_l^{\text{denoised}}) / \hat{t}$
 - 10: $dt = c_\tau - \hat{t}$
 - 11: $\vec{\mathbf{x}}_l \leftarrow \vec{\mathbf{x}}_l^{\text{noisy}} + \eta \cdot dt \cdot \vec{\delta}_l$
 - 12: **end for**
 - 13: **return** $\{\vec{\mathbf{x}}_l\}$
-

Diffusion

Algorithm 20 Diffusion Module

def DiffusionModule($\{\bar{\mathbf{x}}_l^{\text{noisy}}\}, \hat{t}, \{\mathbf{f}^*\}, \{\mathbf{s}_i^{\text{inputs}}\}, \{\mathbf{s}_i^{\text{trunk}}\}, \{\mathbf{z}_{ij}^{\text{trunk}}\},$
 $\sigma_{\text{data}} = 16, c_{\text{atom}} = 128, c_{\text{atompair}} = 16, c_{\text{token}} = 768$) :

Conditioning

1: $\{\mathbf{s}_i\}, \{\mathbf{z}_{ij}\} = \text{DiffusionConditioning}(\hat{t}, \{\mathbf{f}^*\}, \{\mathbf{s}_i^{\text{inputs}}\}, \{\mathbf{s}_i^{\text{trunk}}\}, \{\mathbf{z}_{ij}^{\text{trunk}}\}, \sigma_{\text{data}})$

Scale positions to dimensionless vectors with approximately unit variance.

2: $\mathbf{r}_l^{\text{noisy}} = \bar{\mathbf{x}}_l^{\text{noisy}} / \sqrt{\hat{t}^2 + \sigma_{\text{data}}^2}$ $\mathbf{r}_l^{\text{noisy}} \in \mathbb{R}^3$

Sequence-local Atom Attention and aggregation to coarse-grained tokens

3: $\{\mathbf{a}_i\}, \{\mathbf{q}_l^{\text{skip}}\}, \{\mathbf{c}_l^{\text{skip}}\}, \{\mathbf{p}_{lm}^{\text{skip}}\} = \text{AtomAttentionEncoder}(\{\mathbf{f}^*\}, \{\mathbf{r}_l^{\text{noisy}}\}, \{\mathbf{s}_i^{\text{trunk}}\}, \{\mathbf{z}_{ij}\}, c_{\text{atom}}, c_{\text{atompair}}, c_{\text{token}})$
 $\mathbf{a}_i \in \mathbb{R}^{c_{\text{token}}}$

Full self-attention on token level.

4: $\mathbf{a}_i += \text{LinearNoBias}(\text{LayerNorm}(\mathbf{s}_i))$

5: $\{\mathbf{a}_i\} \leftarrow \text{DiffusionTransformer}(\{\mathbf{a}_i\}, \{\mathbf{s}_i\}, \{\mathbf{z}_{ij}\}, \beta_{ij} = 0, N_{\text{block}} = 24, N_{\text{head}} = 16)$

6: $\mathbf{a}_i \leftarrow \text{LayerNorm}(\mathbf{a}_i)$

Broadcast token activations to atoms and run Sequence-local Atom Attention

7: $\{\mathbf{r}_l^{\text{update}}\} = \text{AtomAttentionDecoder}(\{\mathbf{a}_i\}, \{\mathbf{q}_l^{\text{skip}}\}, \{\mathbf{c}_l^{\text{skip}}\}, \{\mathbf{p}_{lm}^{\text{skip}}\})$

Rescale updates to positions and combine with input positions

8: $\bar{\mathbf{x}}_l^{\text{out}} = \sigma_{\text{data}}^2 / (\sigma_{\text{data}}^2 + \hat{t}^2) \cdot \bar{\mathbf{x}}_l^{\text{noisy}} + \sigma_{\text{data}} \cdot \hat{t} / \sqrt{\sigma_{\text{data}}^2 + \hat{t}^2} \cdot \mathbf{r}_l^{\text{update}}$

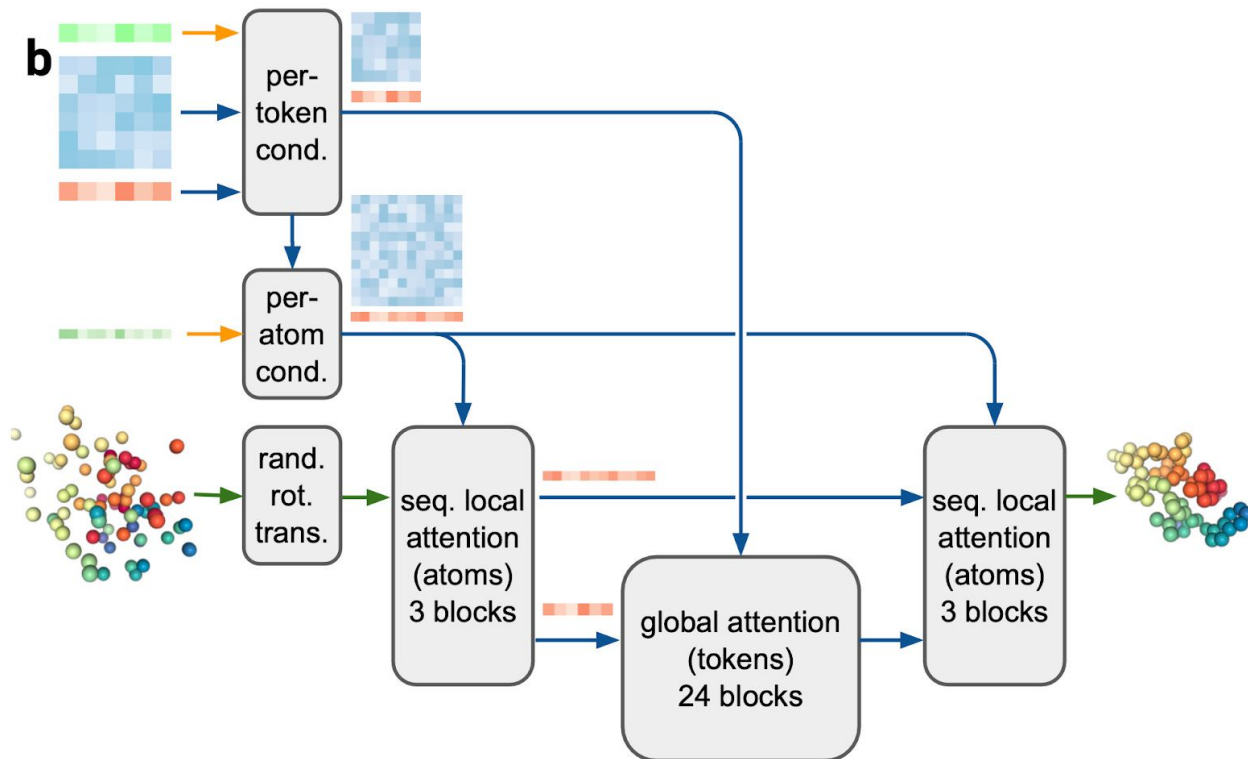
9: **return** $\{\bar{\mathbf{x}}_l^{\text{out}}\}$

AF3 training data

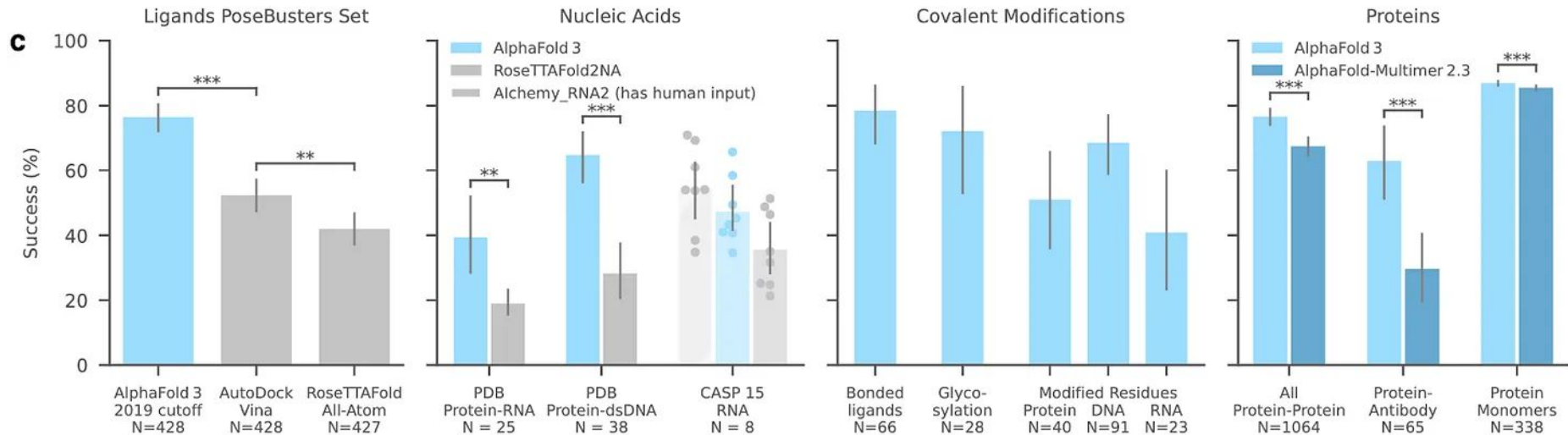
Table 3 | Sampling of training examples: Examples are drawn from a mixture of datasets with associated weight fractions (weight column). Within a dataset, examples are drawn either according to our weighted procedure or uniformly (Sampling strategy column).

Name	Description	Sampl. strategy	Weight
Weighted PDB	Ground truth PDB structures	weighted	0.5
Disordered protein PDB distillation	Proteins with unresolved residues	weighted	0.02
Protein monomer distillation	Protein monomer predictions from MGnify	uniform	0.495
Short protein monomer distillation	Protein short monomer predictions from MGnify	uniform	0.005
RNA distillation	RNA monomer predictions from Rfam	uniform	0.05
Transcription factor negatives	MGnify protein + random DNA	uniform	0.01 ¹
Transcription factor positives	DNA+protein predictions from JASPAR	uniform	0.02 ¹

AF3 diffusion module



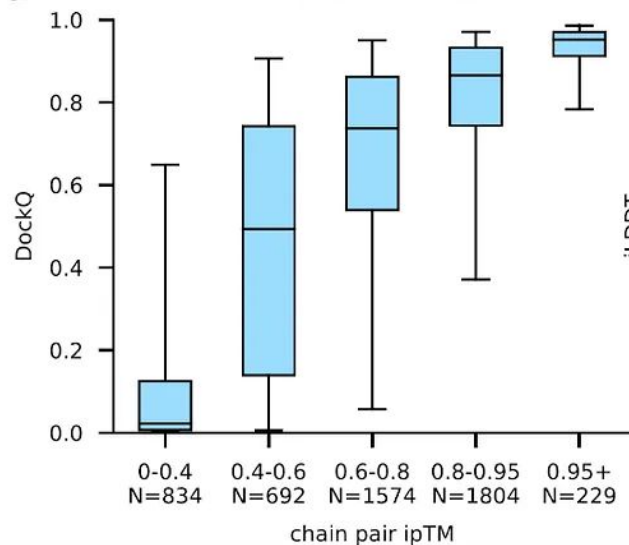
AF3 performance



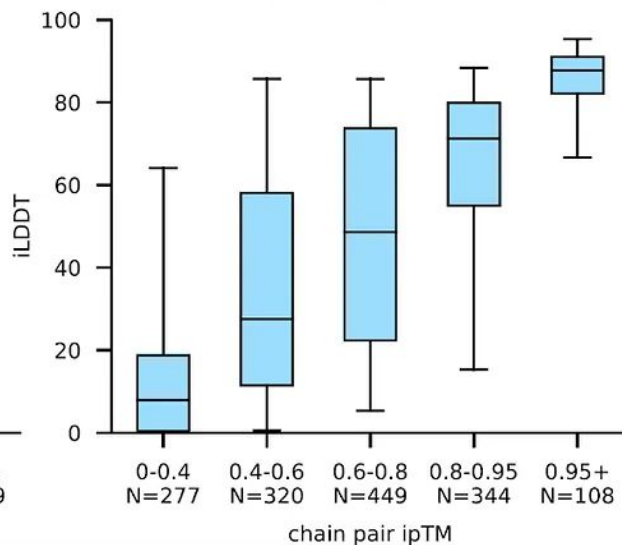
AF3 performance

a

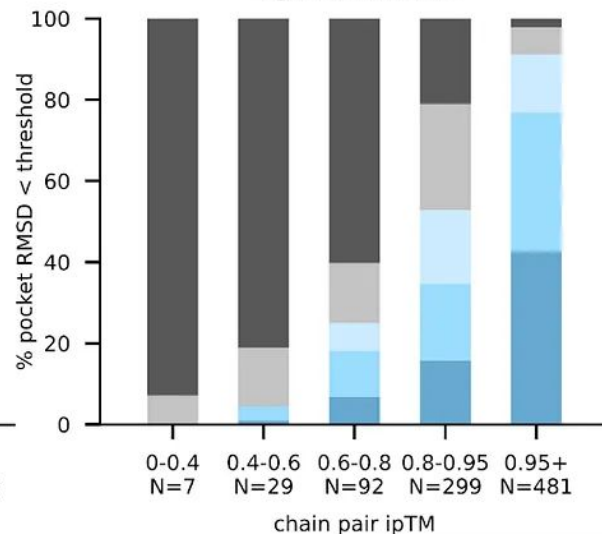
Protein-Protein



Nucleic Acid-Protein

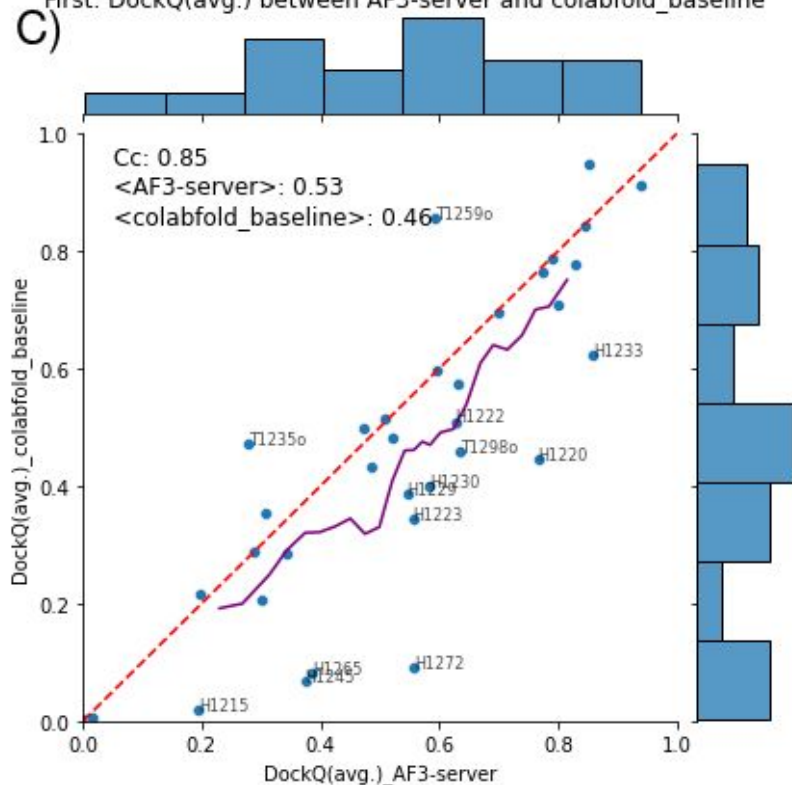


Ligand-Protein

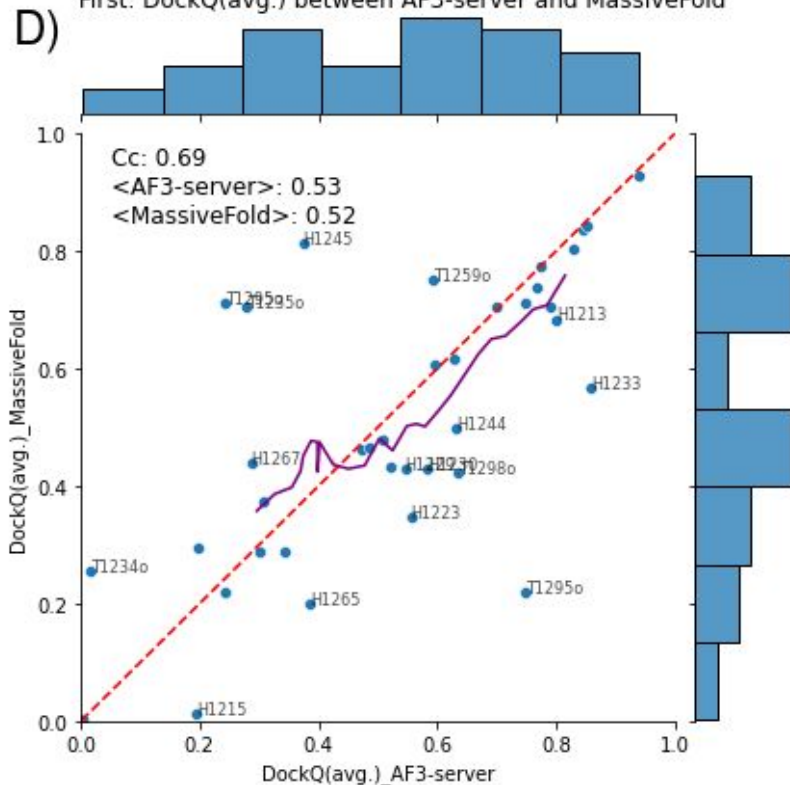


CASP16: AF3 vs AF2 - Top Ranked Predictions - DockQ

First: DockQ(avg.) between AF3-server and colabfold_baseline



First: DockQ(avg.) between AF3-server and MassiveFold

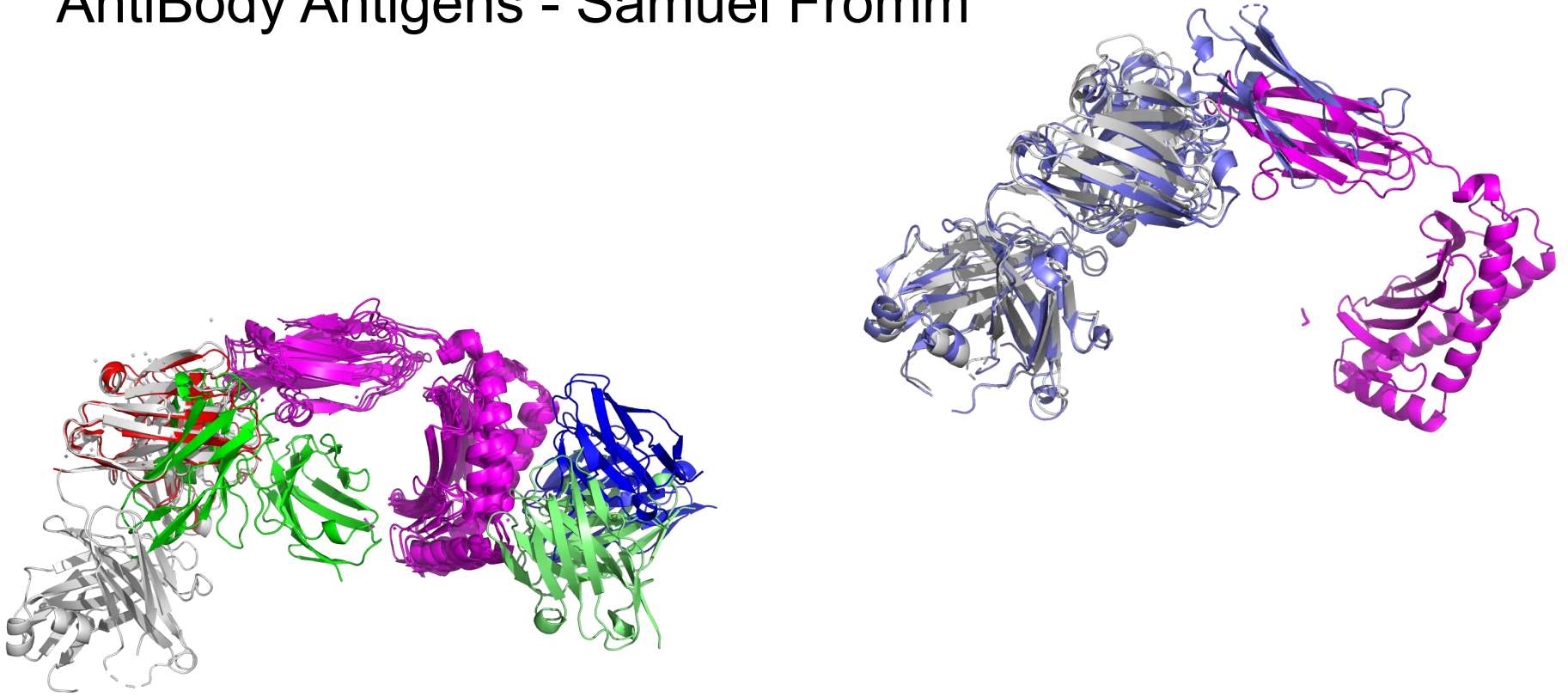


Loss

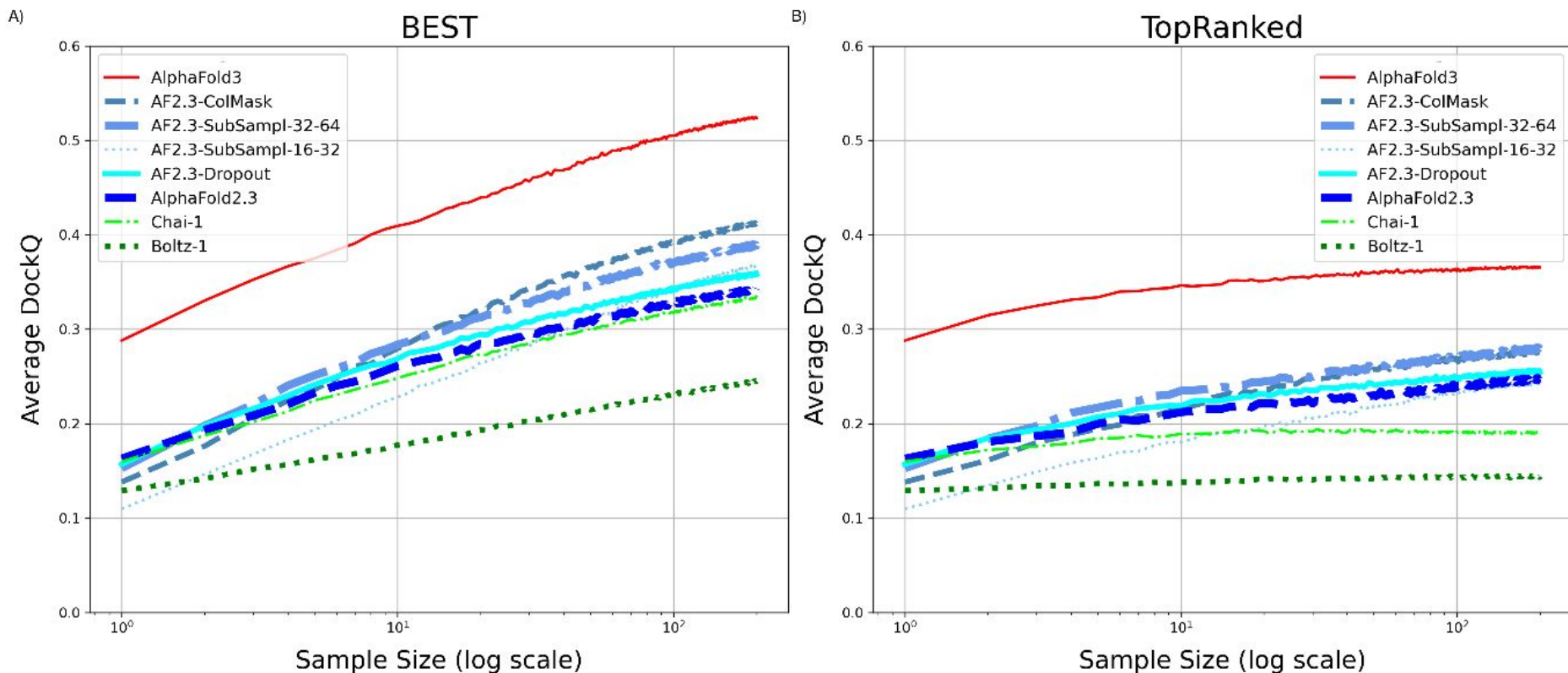
The details of the losses from each module are given in the relevant sections, these are then combined into the final loss: $L_{\text{loss}} = \alpha_{\text{confidence}} \cdot (L_{\text{plddt}} + L_{\text{pde}} + L_{\text{resolved}} + \alpha_{\text{pae}} \cdot L_{\text{pae}}) + \alpha_{\text{diffusion}} \cdot L_{\text{diffusion}} + \alpha_{\text{distogram}} \cdot L_{\text{distogram}}$ (15)

Where $\alpha_{\text{confidence}} = 10^{-4}$, $\alpha_{\text{diffusion}} = 4$, $\alpha_{\text{distogram}} = 3 \cdot 10^{-2}$ and $\alpha_{\text{pae}} = 0$ for all except for the final training stage, where it is set to 1.

AntiBody Antigens - Samuel Fromm

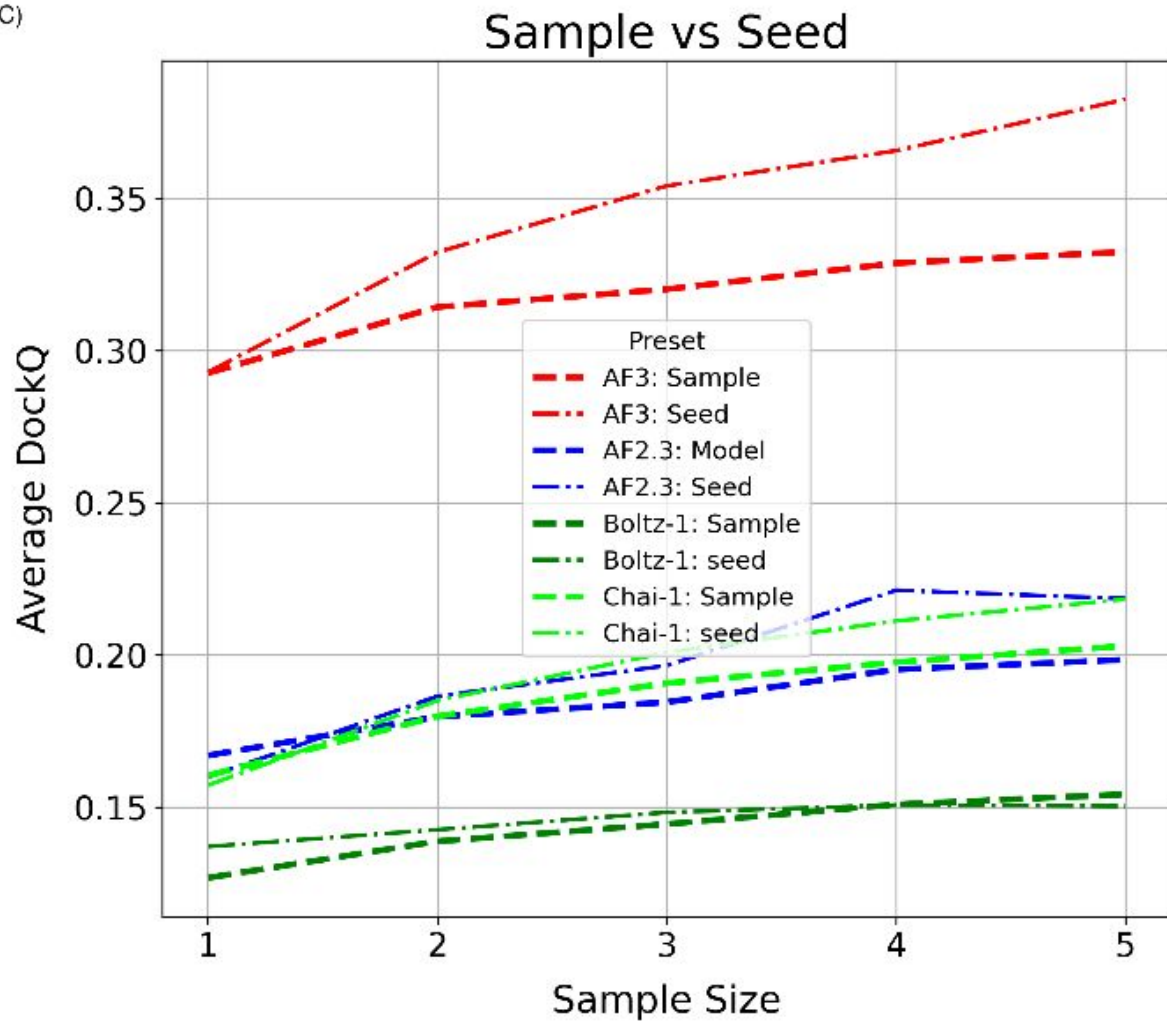


Performance

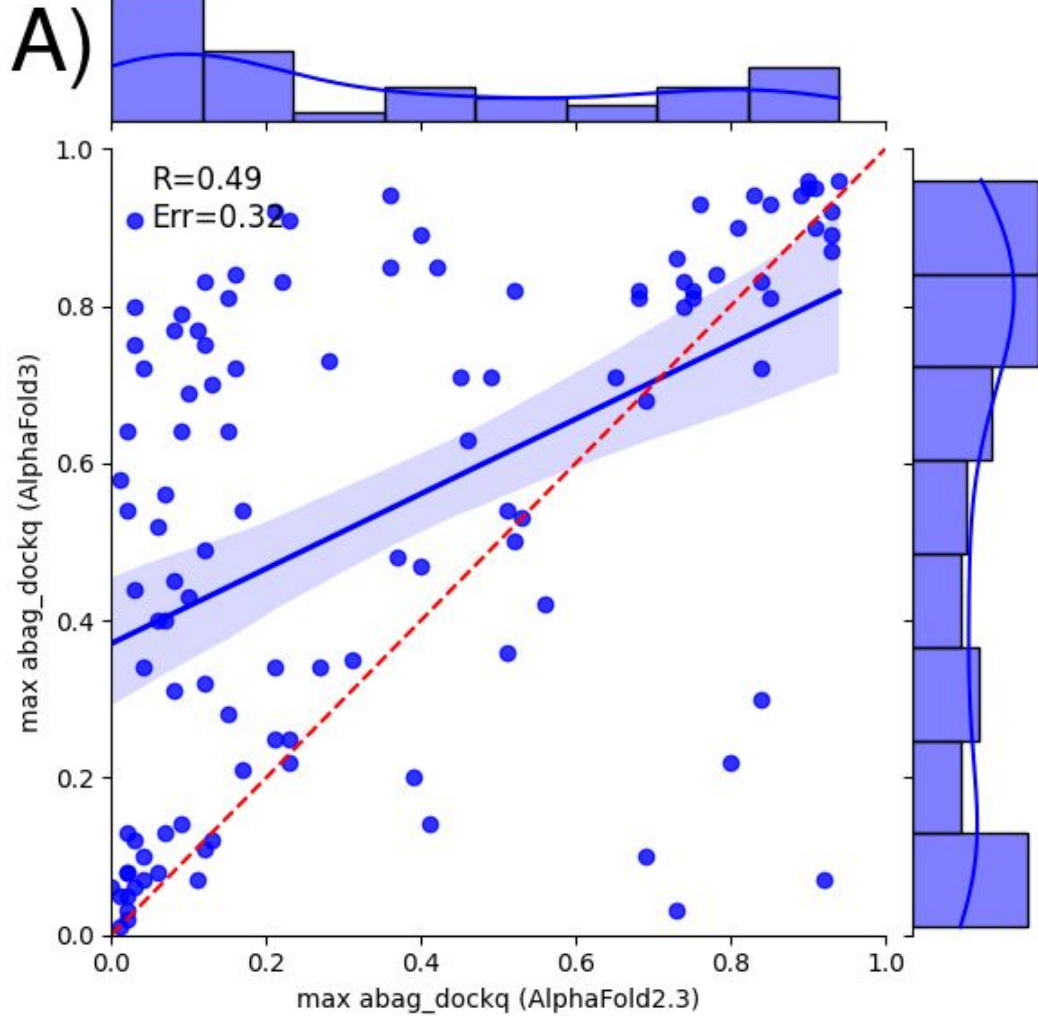


Performance

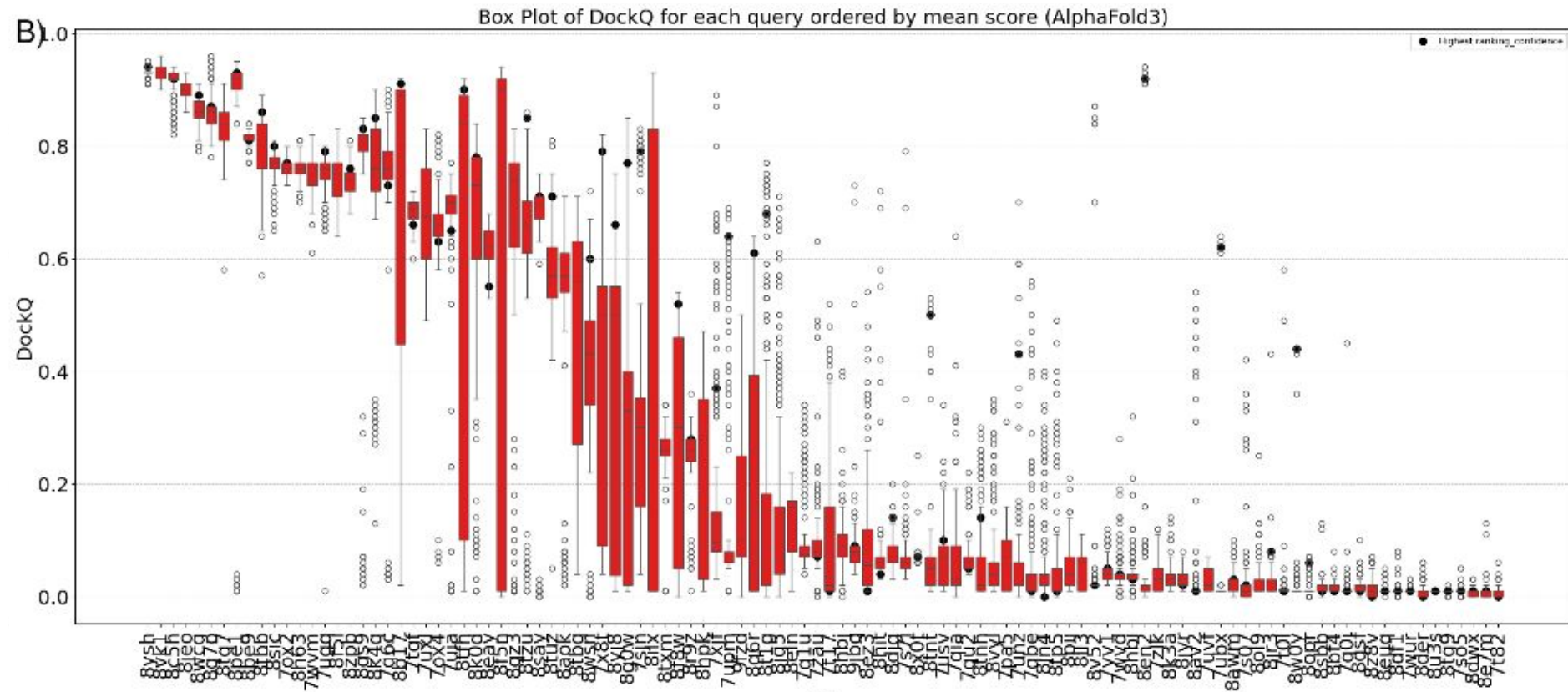
c)



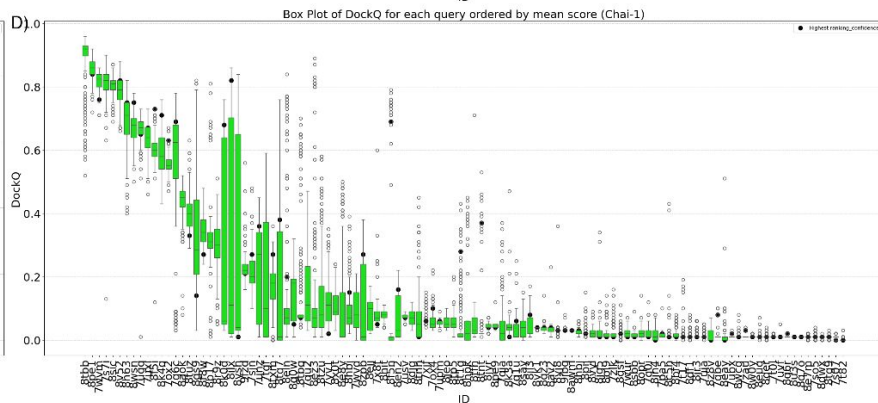
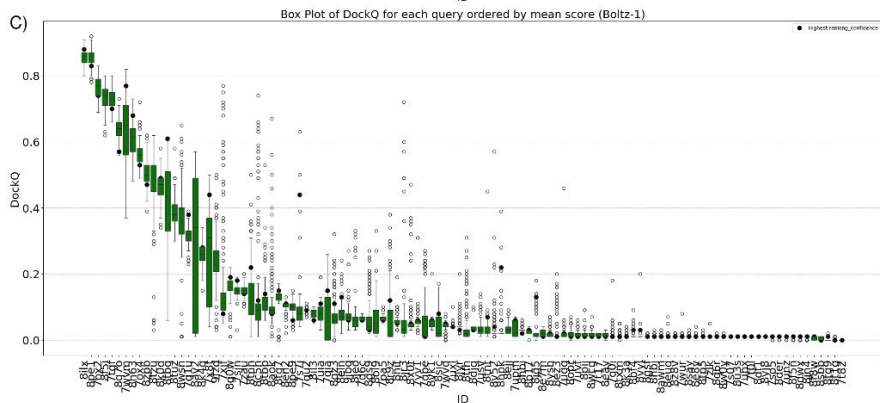
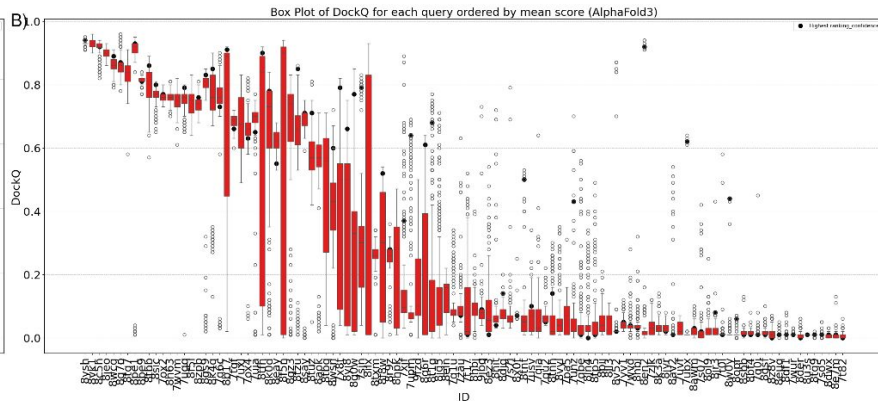
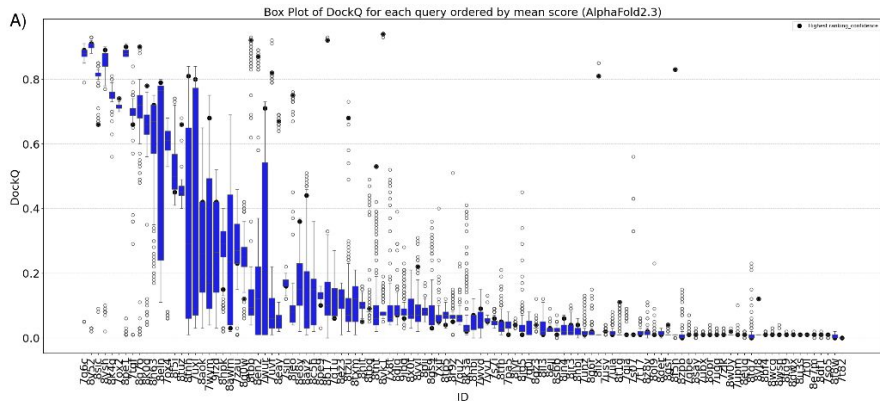
AF3 vs AF2.3



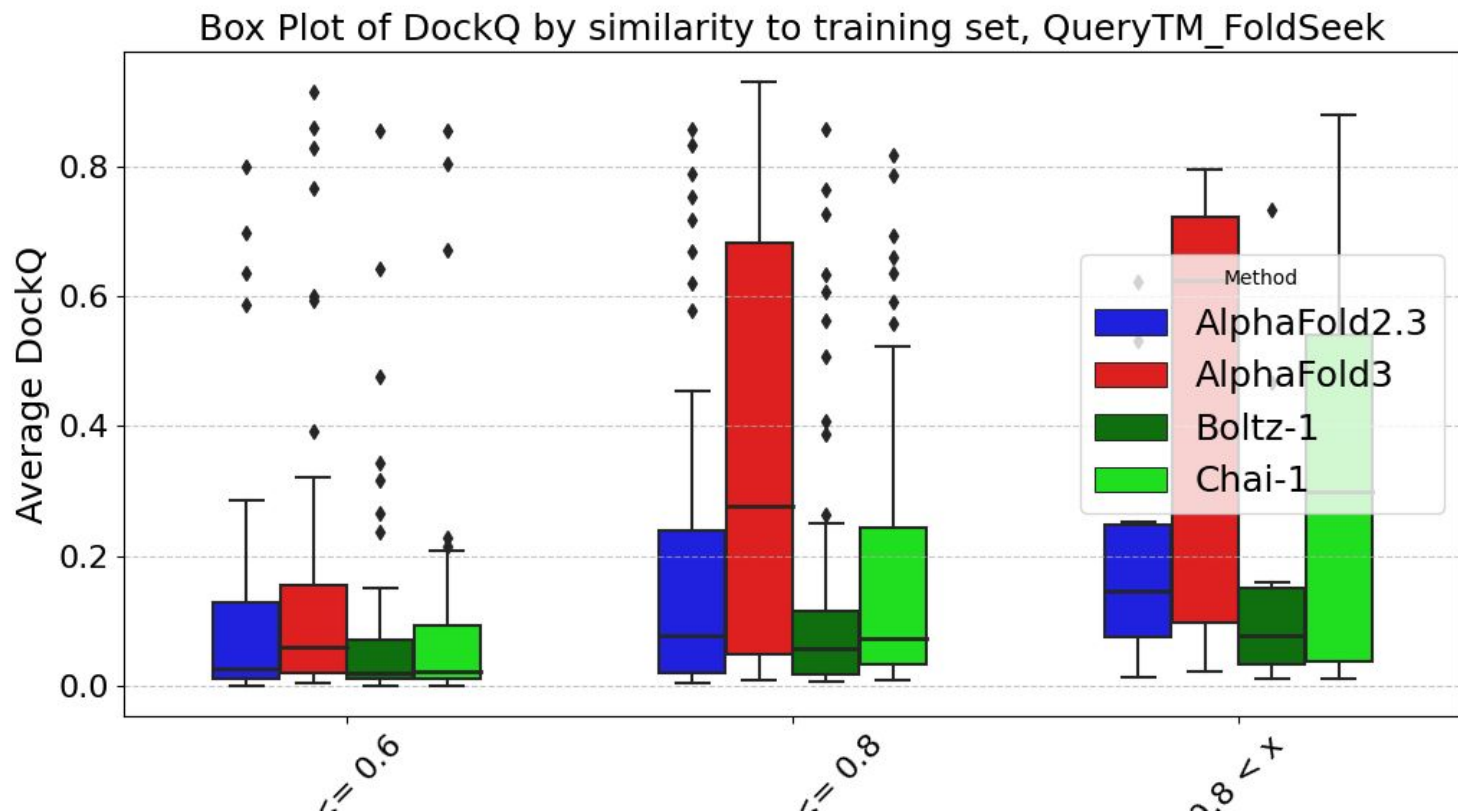
Performance per target



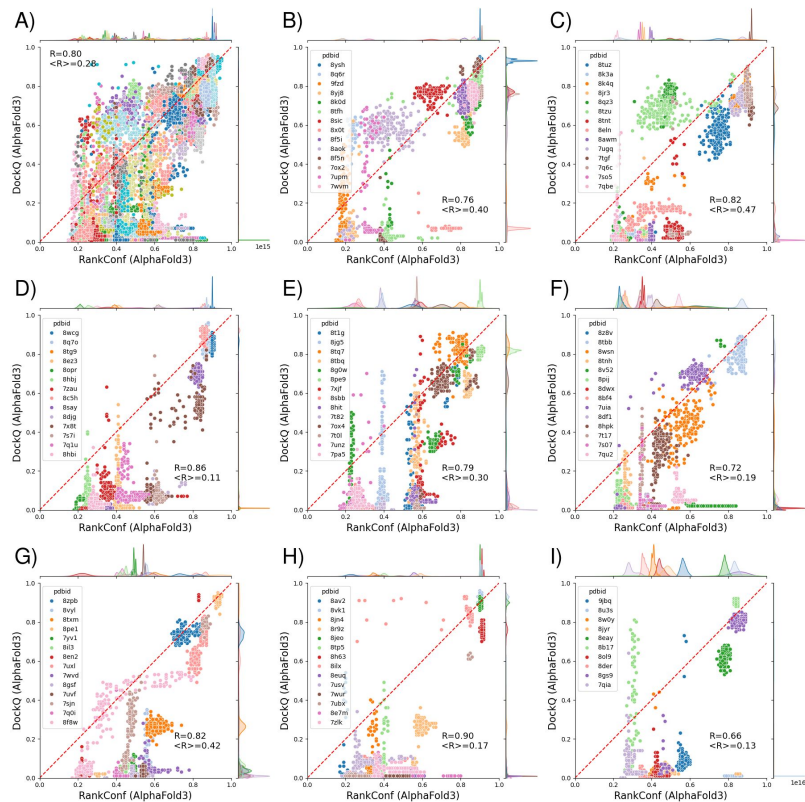
Type of targets



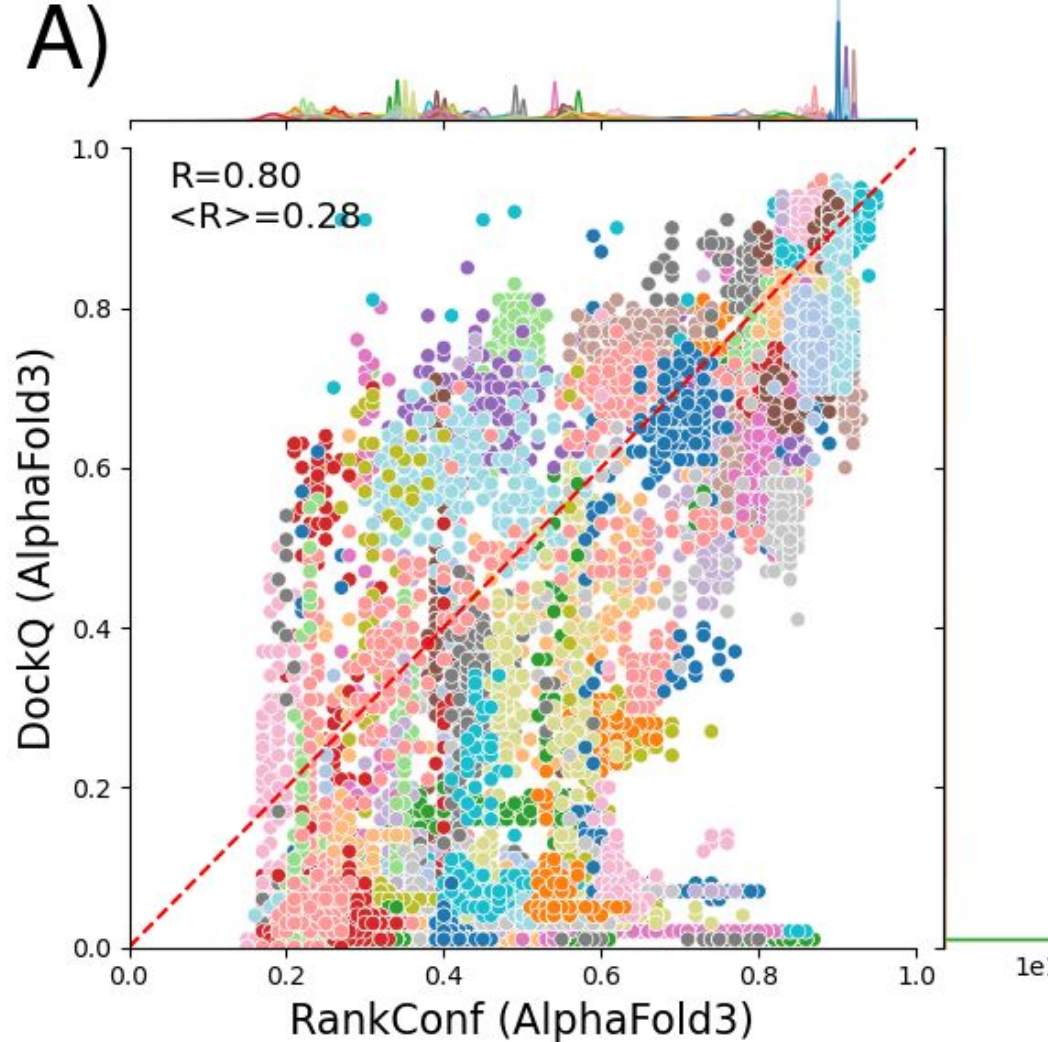
Similarity to PDB



Per target

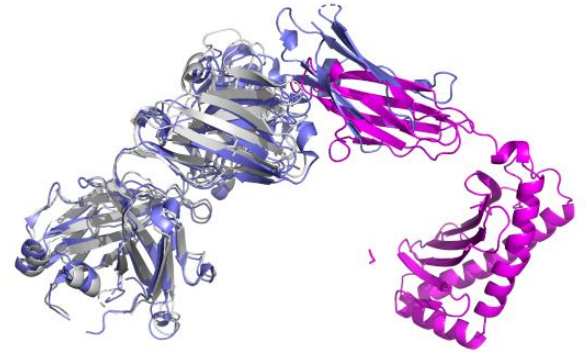


A)



AlphaFold3 - summary 2

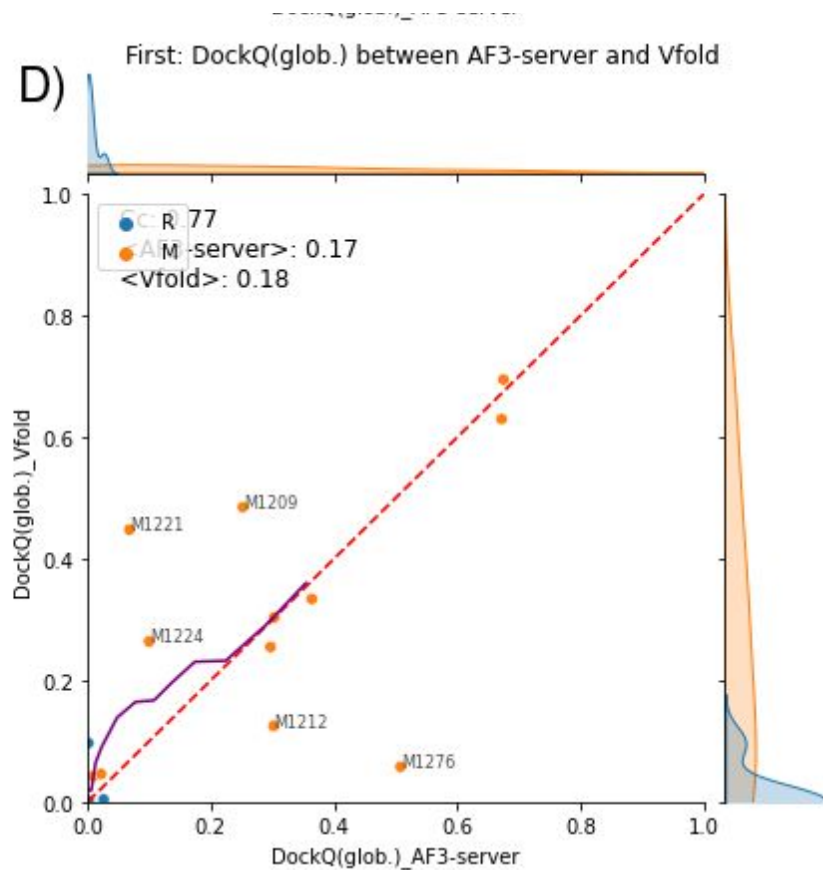
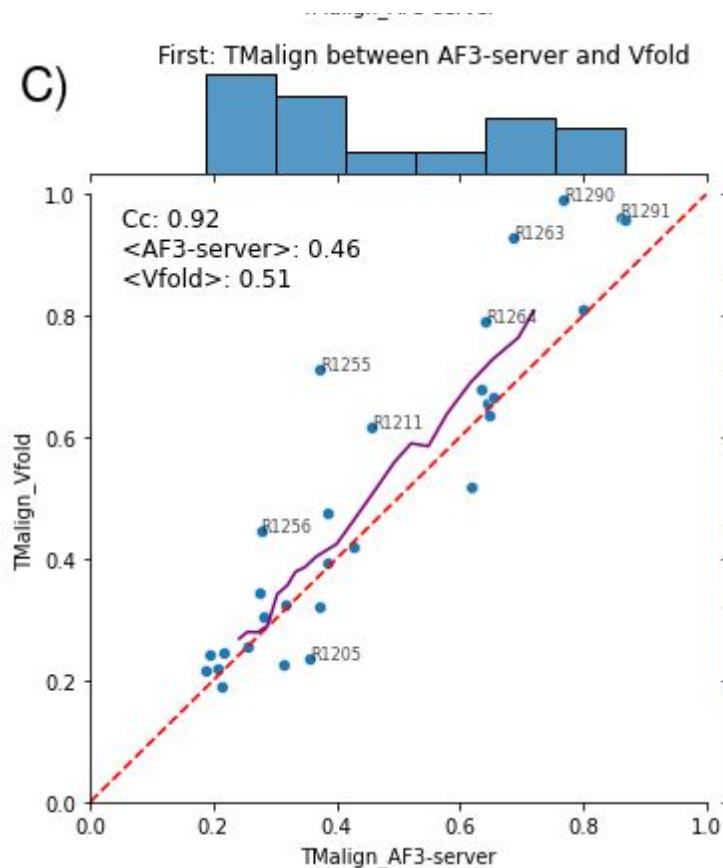
- When there is no co-evolutionary signal it does better than earlier methods
- However, not when there is no structural similarity to the training set.
- Memorization (not necessary a bad thing)
- Unclear exactly what it memorises.



RNA - Marko Ludaic

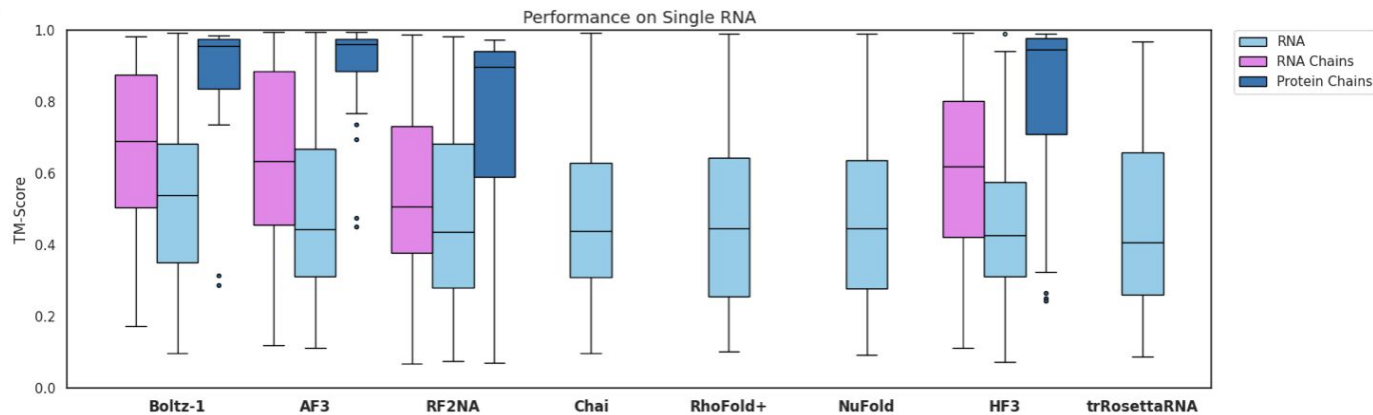
METHOD	DESCRIPTION	RNA
AlphaFold 3	A deep learning method consisting of a pairformer for residue interaction modeling and a diffusion module for denoising raw atomic coordinates, removing global rotational/translation constraints to enhance prediction flexibility and accuracy.	Single, complex
Boltz-1	AlphaFold 3-inspired architecture with modified order of operations in the MSAModule and transformer layer.	Single, complex
Chai-1	AlphaFold 3-inspired architecture incorporating residue-level embeddings from a large language model for proteins, also introducing experimental constraint features like pocket, contact, and docking constraints. Trained without RNA MSAs.	Single, complex
HelixFold3	AlphaFold 3-inspired architecture using PaddlePaddle library allowing for faster model training.	Single, complex
NuFold	A deep learning method built upon AlphaFold architecture design to predict RNA structures by incorporating nucleic acid sequences, secondary structure inputs, and RNA-specific inter-base angles, distances, and atom positions.	Single
RhoFold+	A deep learning model for RNA structure prediction that utilizes an RNA language model, alongside several deep learning modules, including an IPA module for modeling 3D positions.	Single
RoseTTAFoldNA	A three-track architected model that simultaneously updates sequence (1D), residue-pair (2D), and structural (3D) representations to predict macromolecular structures..	Single, complex
trRosettaRNA	A deep learning model consisting of MSA and pair representations fed into a transformer network (RNAformer) predicting 1D and 2D geometries, refined in the final 3D structure folding step.	Single

CASP16: RNA predictions

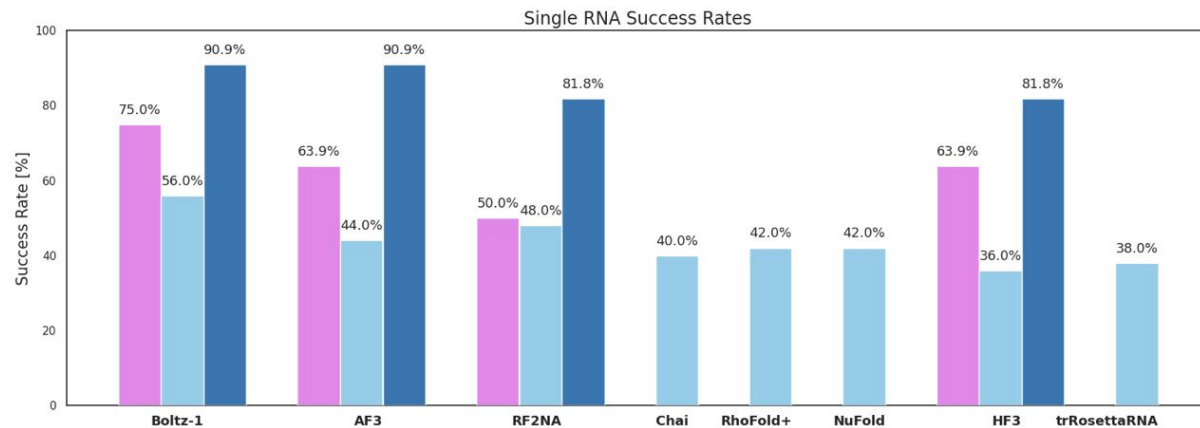


Performance - single RNA

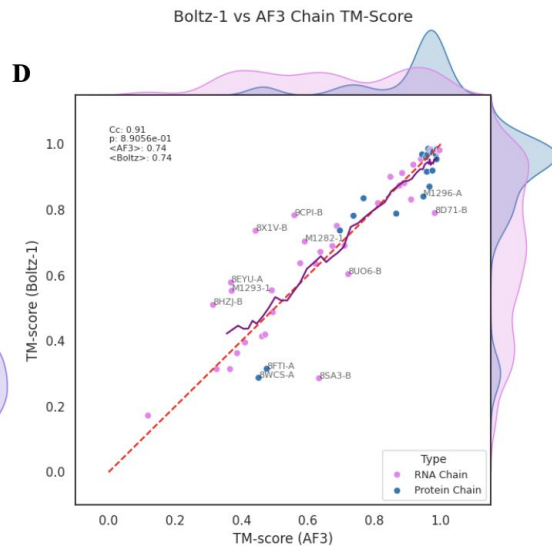
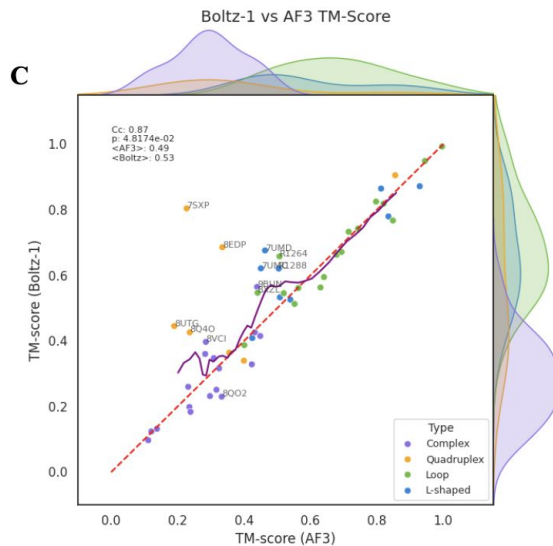
A



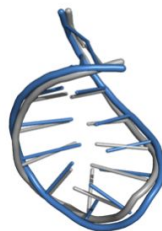
B



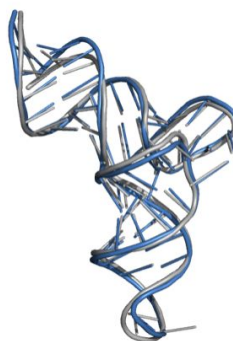
Performance - single RNA



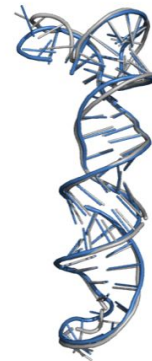
AlphaFold 3 | **7UCR**
TM-Score **0.996**
RMSD **0.31**



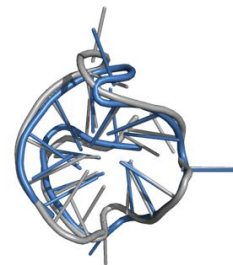
NuFold | **8CLR**
TM-Score **0.950**
RMSD **1.19**



AlphaFold 3 | **8VT5**
TM-Score **0.929**
RMSD **1.43**

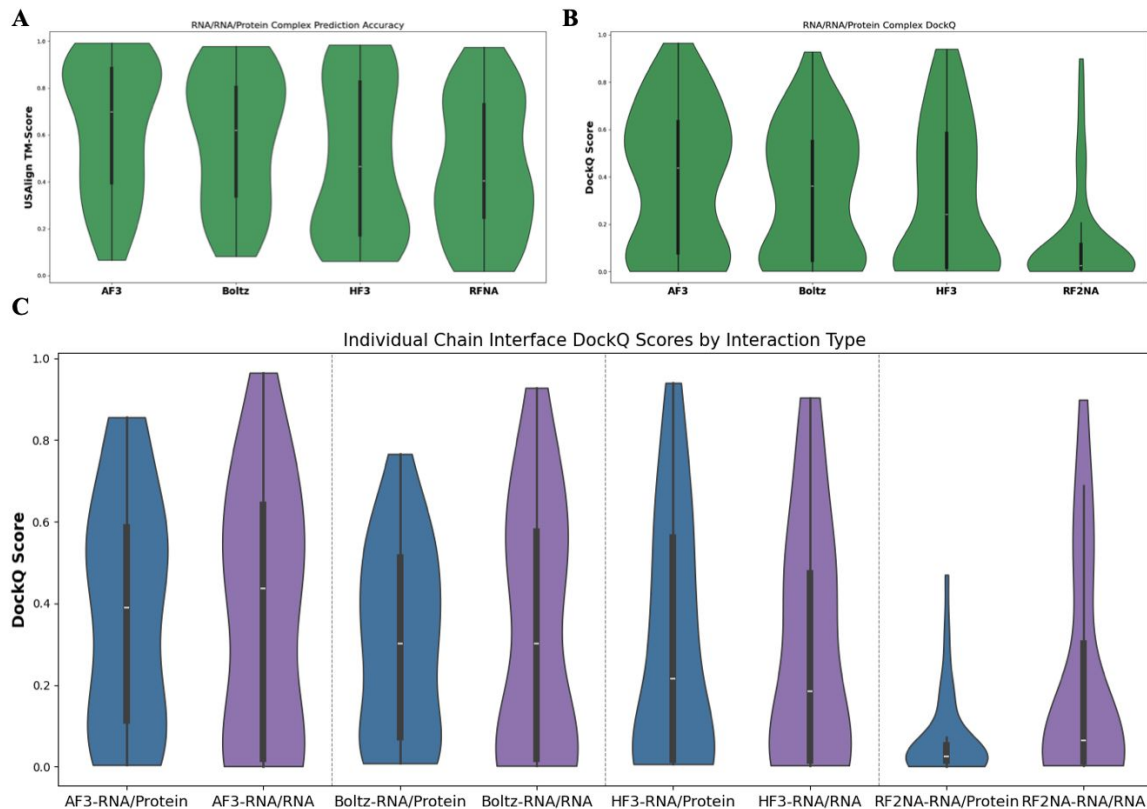
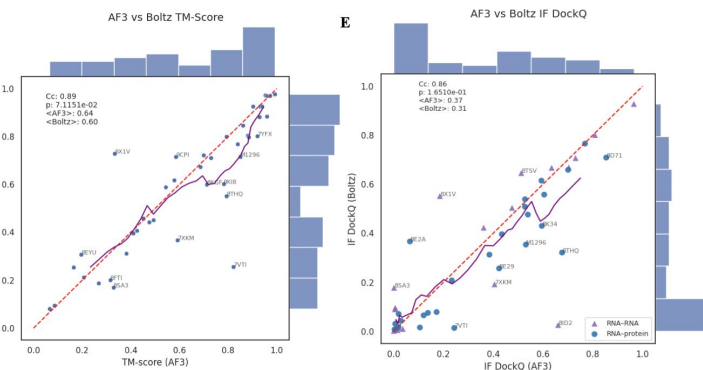


Boltz-1 | **8WA4**
TM-Score **0.905**
RMSD **1.78**



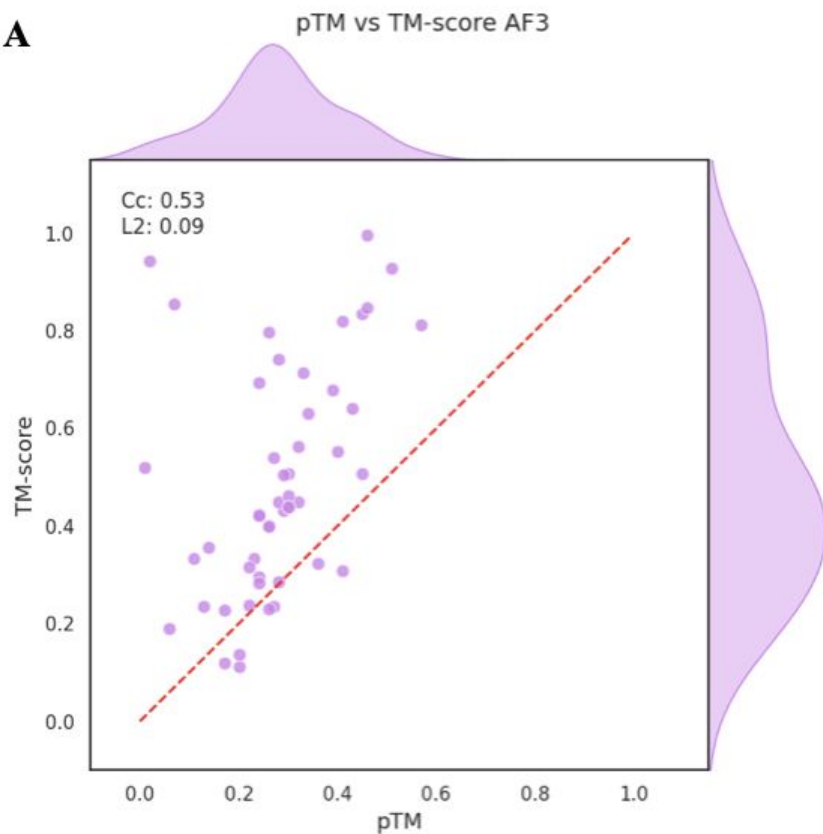
RF2NA | **R1264**
TM-Score **0.905**
RMSD **0.47**

Performance complexes

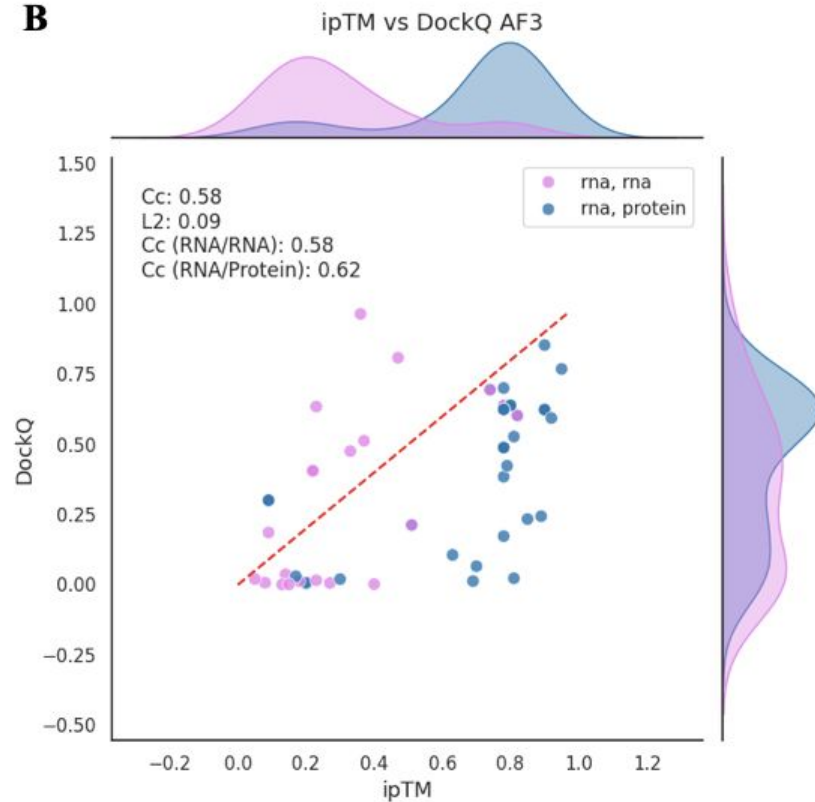


pTM vs TM

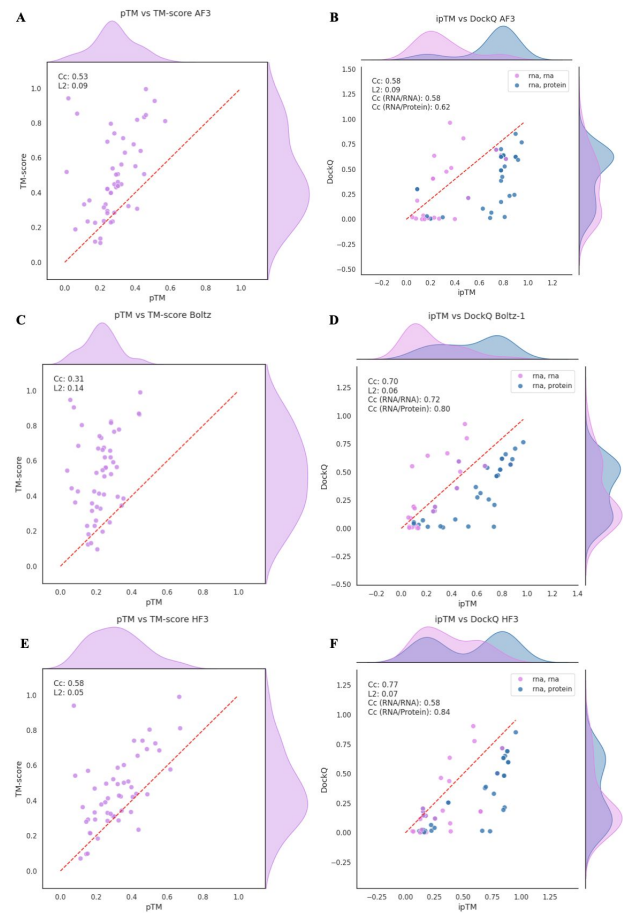
A



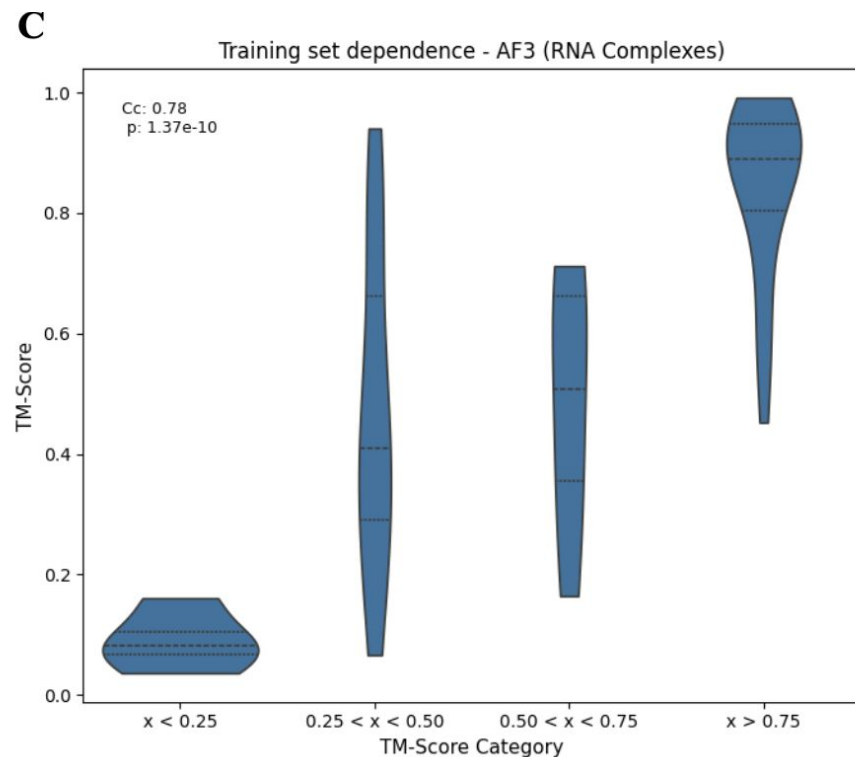
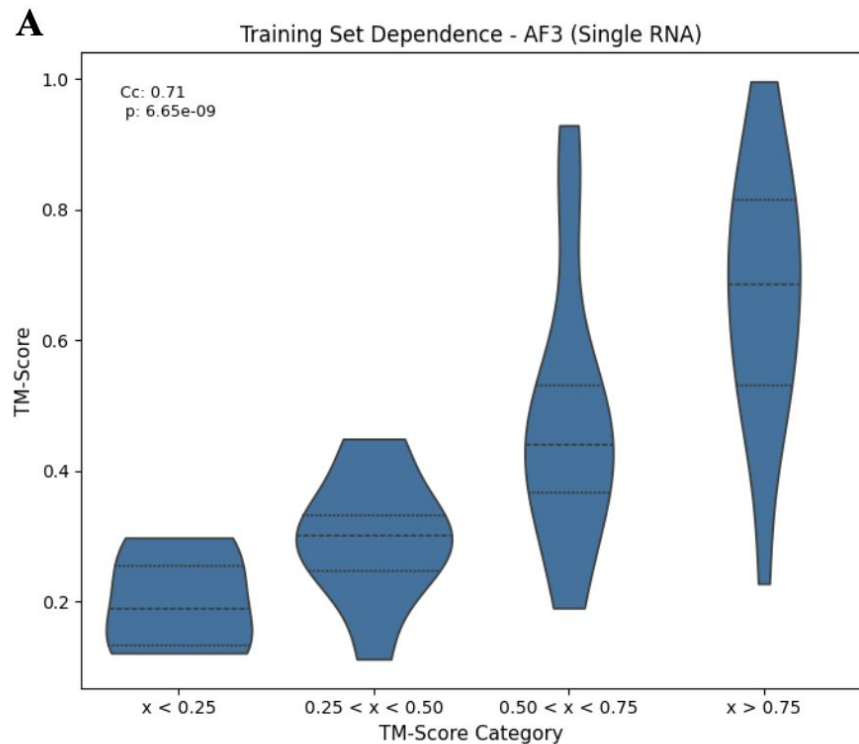
B



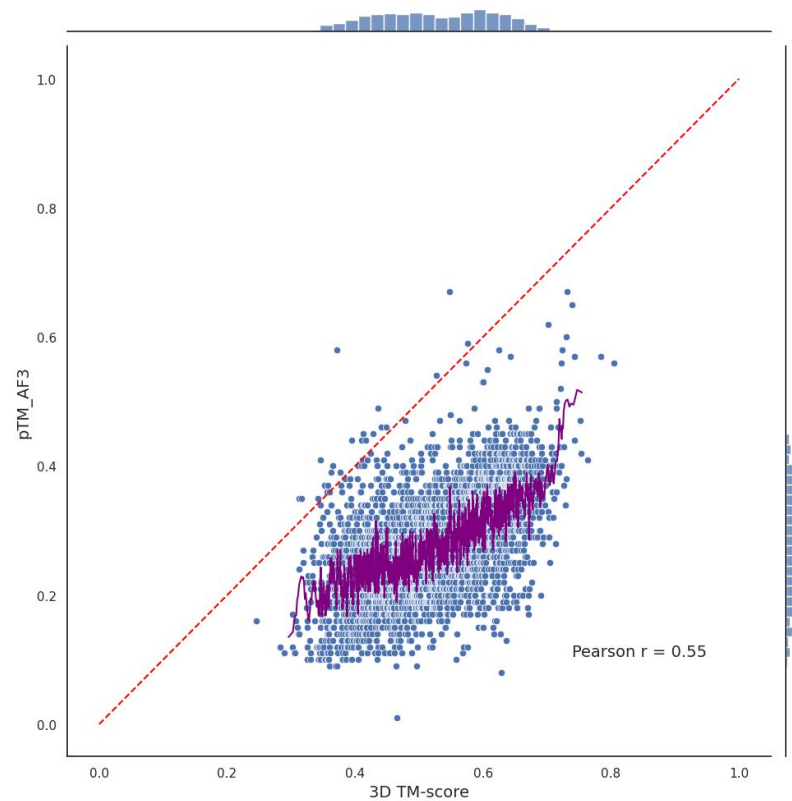
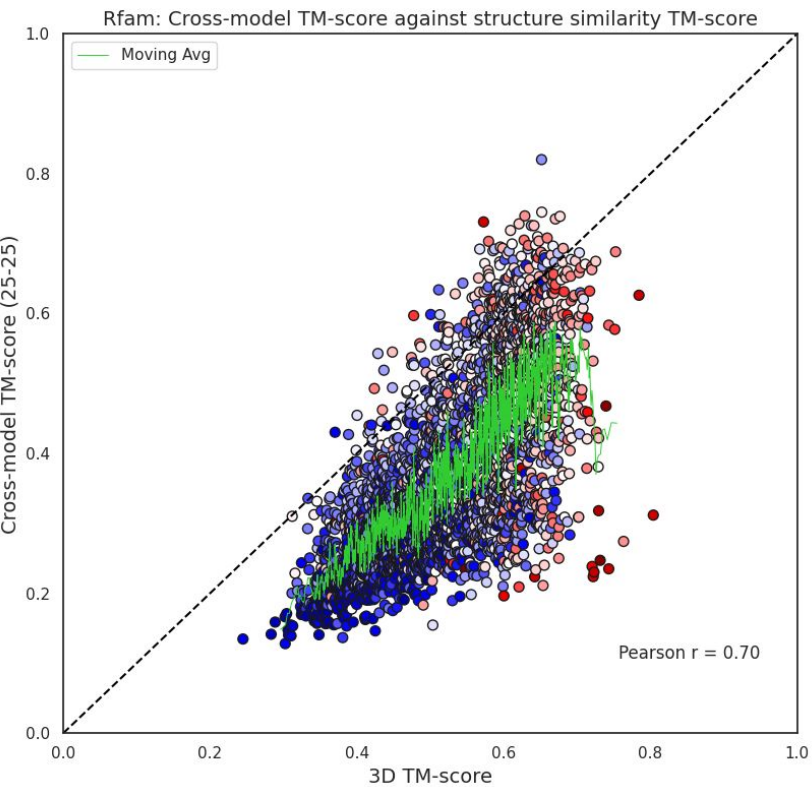
ipTM vs TM



AF3 vs PDB similarity



A hope...



AlphaFold3 - RNA

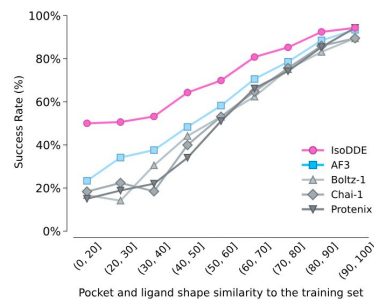
- Some good predictions
- Only for RNAs similar to what is seen in training.
- Underestimates the quality (and bad estimations)
-

IsoDDE

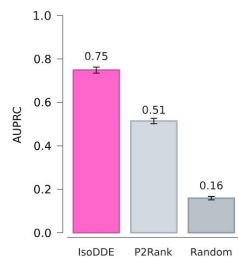
IsoDDE Capabilities



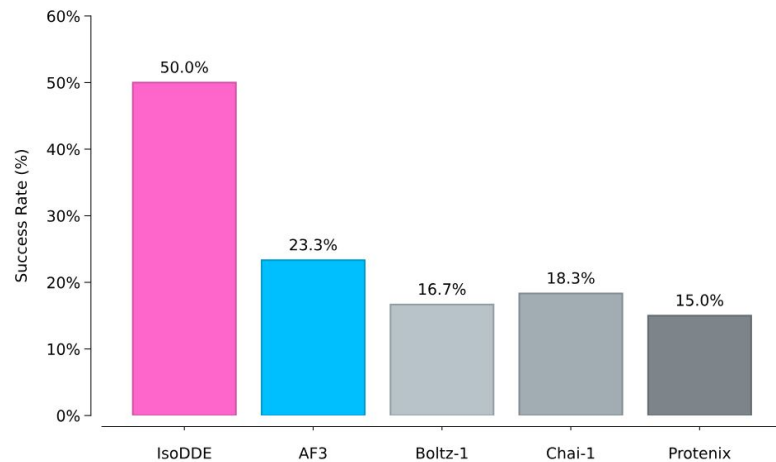
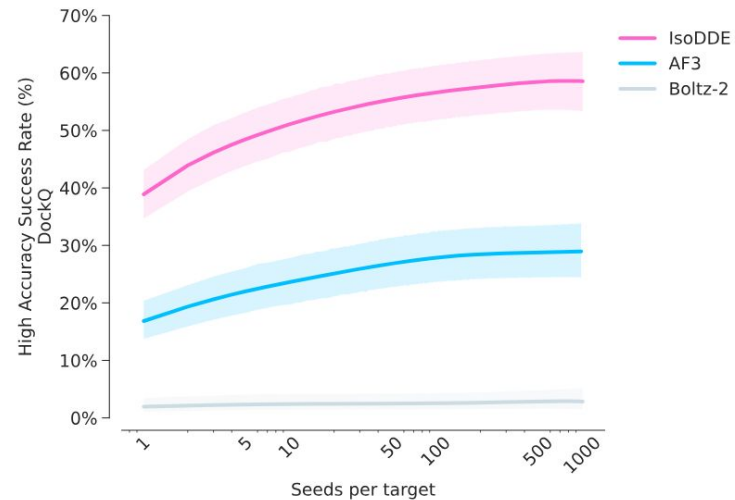
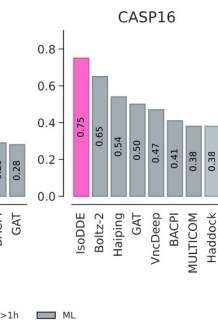
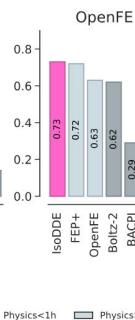
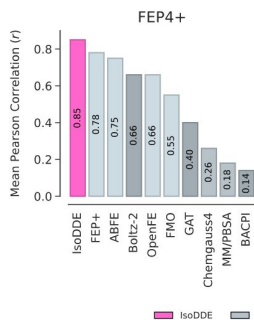
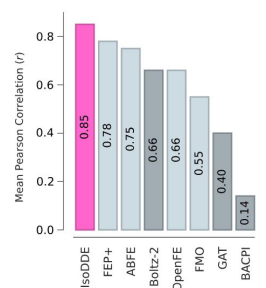
Structure Prediction



Pocket Identification



Binding Affinity



Acknowledgement



Elofsson group, Stockholm University

- Sarah Narrowe Danielsson
 - Matteo Tadiello
 - Vsevolod Viliuga
 - Marko Ludaic
 - Samuel Fromm
- PhD student positions
 - Postdoc positions
 - Contact: arne@bioinfo.se

Petras Kundrotas, Kansas / Stockholm



Future

- **Key Takeaways:**

- AlphaFold revolutionized protein structure prediction.
- Deep learning surpassed traditional bioinformatics approaches.
- AlphaFold3 pushes boundaries with multi-molecule modeling.

- **Final Thought:**

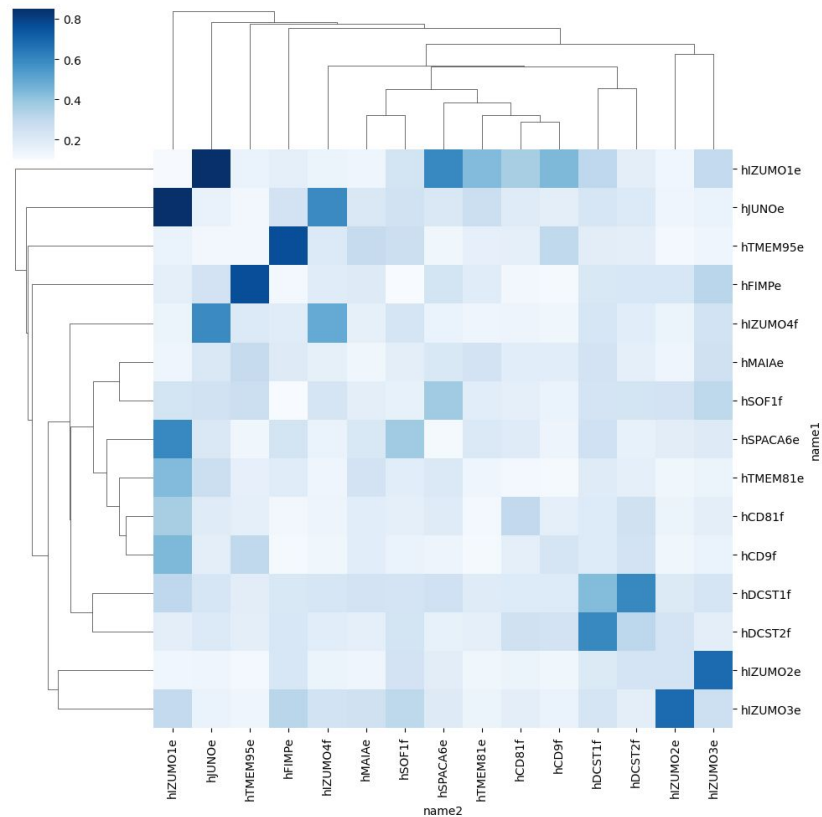
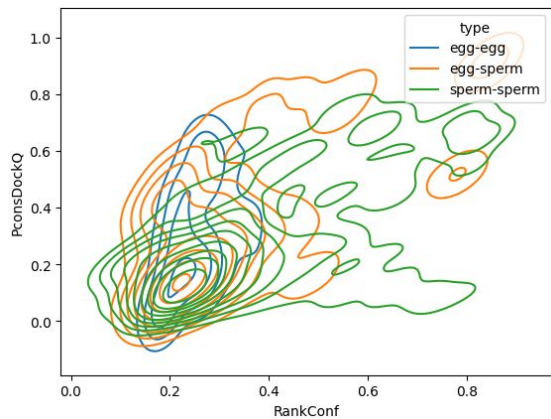
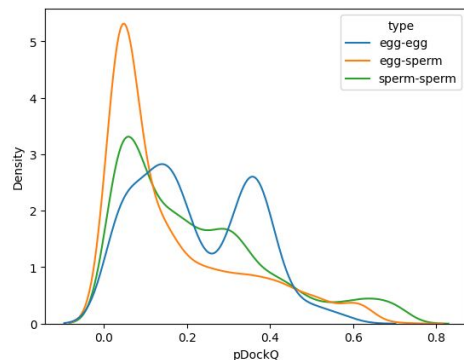
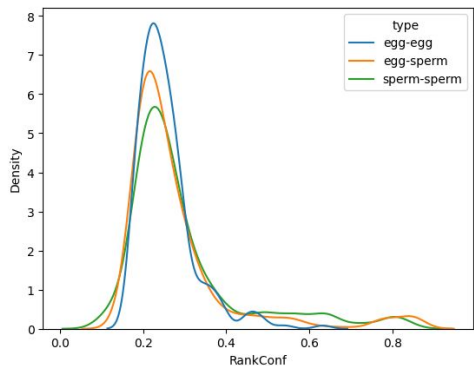
- AI-driven structural biology will accelerate drug discovery, synthetic biology, and personalized medicine.

Deep learning insights into the architecture of the mammalian egg-sperm fusion synapse

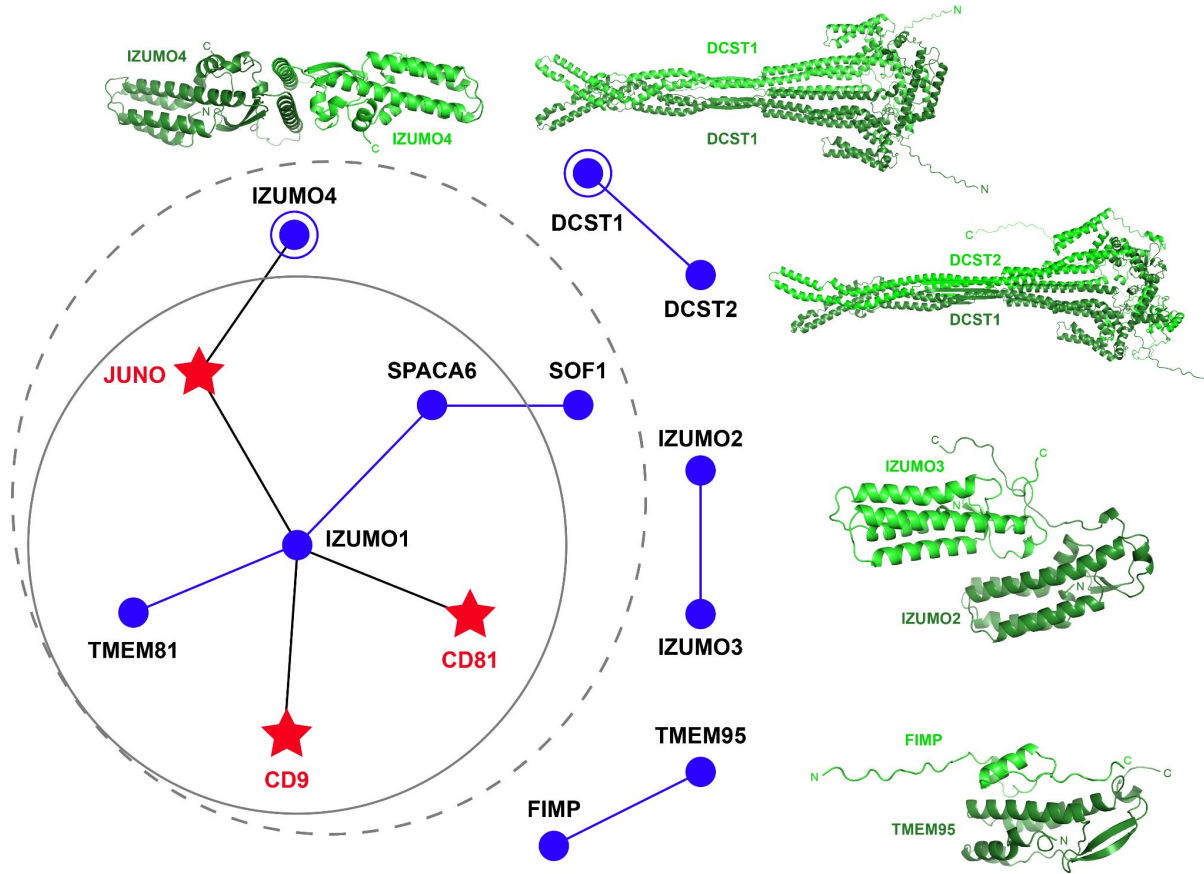
Arne Elofsson Ling Han Enrica Bianchi Gavin J Wright Luca Jovine

<https://elifesciences.org/articles/93131>

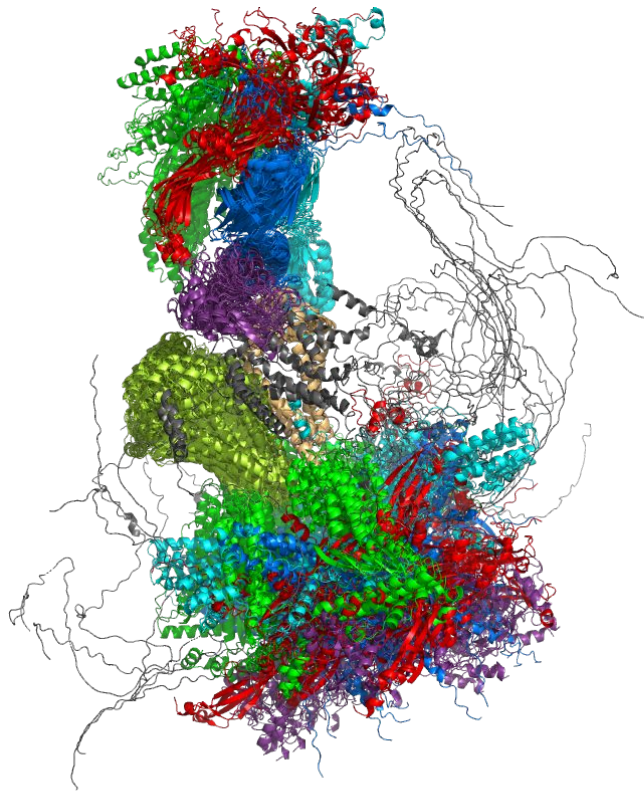
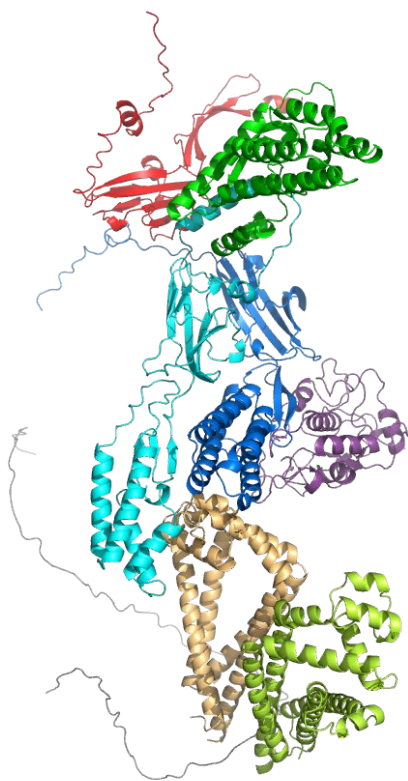
All against all docking of 14 proteins

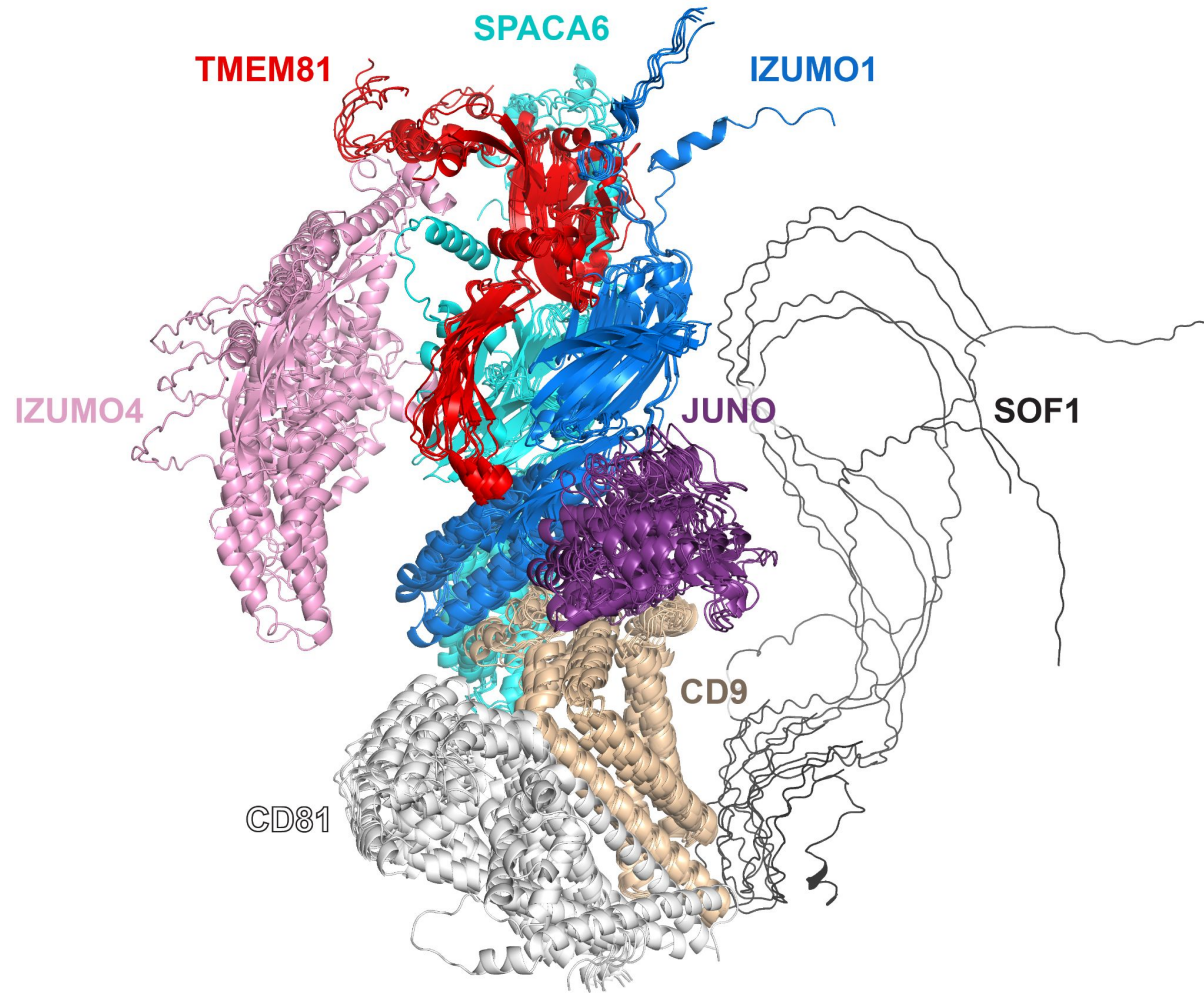


Identification of interacting partners

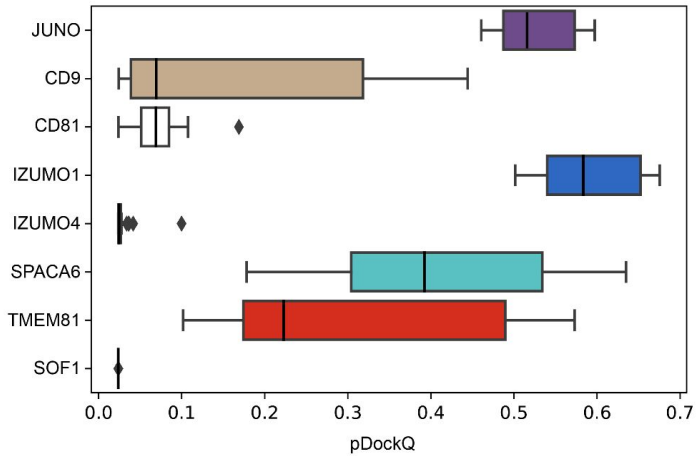


8-chain complex - not consistent

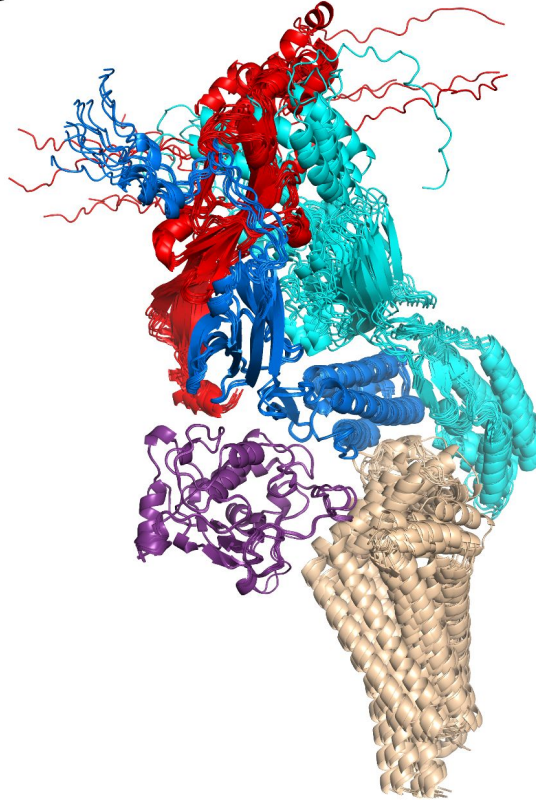




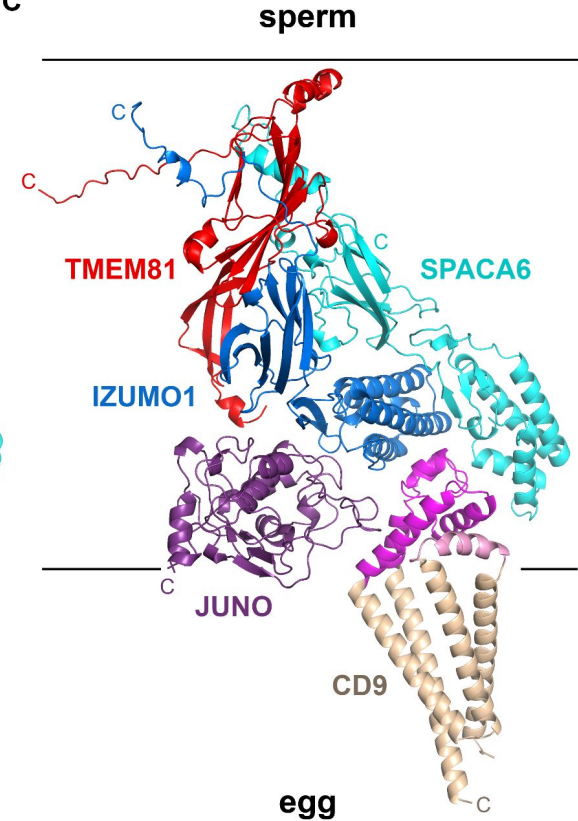
Five chain complex



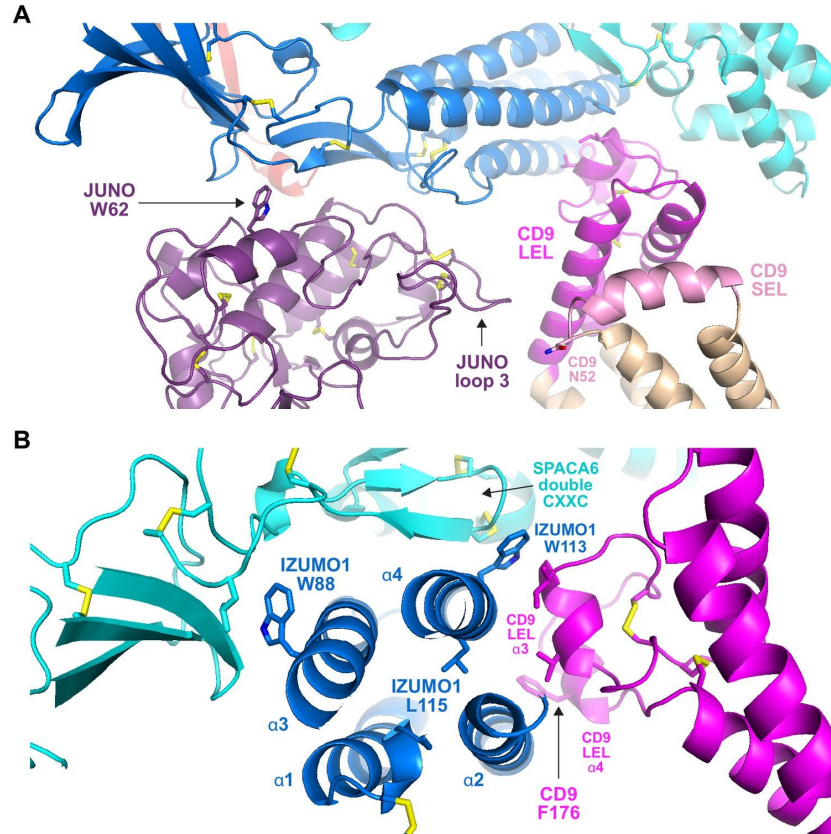
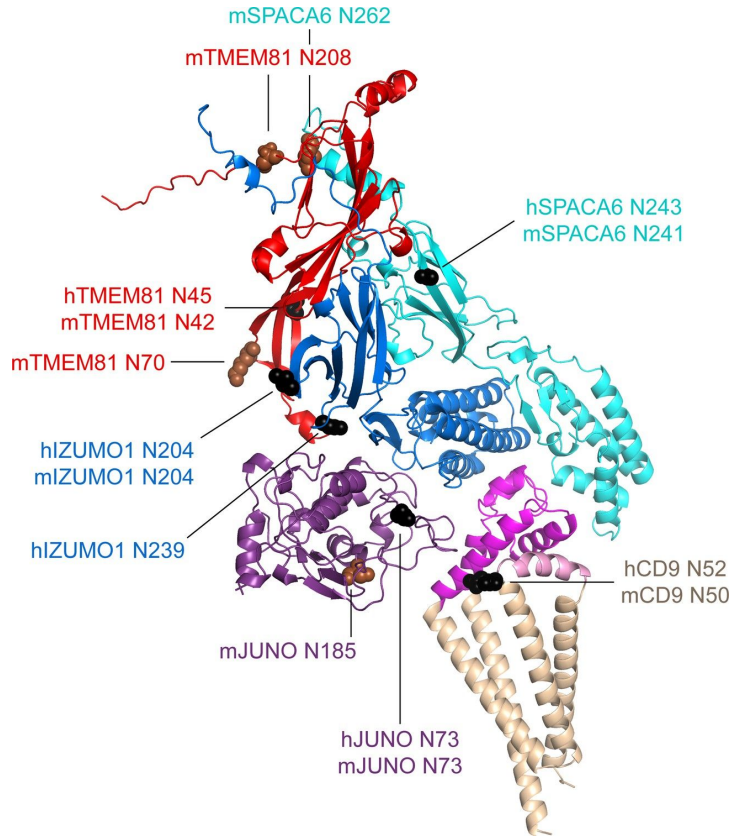
B



C



Glycosylation and mutation sites



Questions

- Given the success of AlphaFold in structure prediction, how do models like OmegaFold and ESMFold compare in terms of speed vs. accuracy trade-offs?
 - Faster due to no MSA - accuracy slightly worse for single proteins - much worse for complexes.
- Does AlphaFold perform equally well across different species, and if not, what challenges arise when predicting protein structures from different organisms, especially for organisms with limited experimental data?
 - Basically yes. There is a small drop in performance for proteins with shallow MSAs. Also proteins in different kingdoms varies (eukaryotic are longer and have more disorder) which affects the overall performance
- Do Evoformer's attention mechanisms correspond to real biological interactions, or are they black-box statistical correlations?
 - They do correlate with contacts (or at least a few of the channels in the last layer do)

- Has (or will) AlphaFold become the main resource for the first steps in resolving protein structure?
 - Yes it has already become the default method
- What are its downfalls and what can other technologies/methods provide?
 - Some of the main problems are antibody-antigen interactions and host-pathogen interactions.
- in the triangle attention block of the AlphaFold model, does it consider all possible positions of A,B and the distance to C at any given atom of the protein or only certain atoms, maybe from the backbone or at highly preserved residues?
 - The entire EvoFormer block only works on residues level (frames), so only backbone.
- When AlphaFold's predictions differ from experimentally determined structures, what criteria can be used to decide if this difference reflects limitations in the prediction model or issues with the experimental data?
 - Checking homologs and quality from AF2 predictions,
- How does AlphaFold use deep learning and attention mechanisms in order to predict protein structures, and what are its implications for fields like drug discovery and bioengineering?
 - Attention is in the EvoFormer - the whole models is quite Deep

- What does the structure of the Evoformer tell us about creating Neural Networks for other biological problems?
 - Not much, I think. It is quite specific for 3D-macromolecules. For RNA you can use the same architecture.
- Is known that AlphaFold predicts static protein structures, but is also known that many proteins function dynamically. What advancements would be needed to predict conformational changes related to this dynamic interactions and how can the response to different environments be studied?
 - Work is happening in the field (AlphaFlow etc). Data is limited (MD-simulations)?
- How do you think AlphaFold will develop in the future?
 - Better for RNA, Ligands, Antibodies

- How does AlphaFold3's Pairformer architecture enhance the modeling of pairwise residue interactions compared to AlphaFold2's Evoformer?
 - It is supposed to focus more on structure than co-evolution,
- If AlphaFold can predict structures with high accuracy, will we still need experimental methods like X-ray diffraction in the future?
 - Biology is not single static structures and large scale complexes are still not predicted accurately, in particular in all active states
- How does AlphaFold use multiple sequence alignment to improve protein structure prediction, and why is evolutionary information important in this process?
 - The basis is co-evolution (but not only)
- How is AlphaFold integrating with experimental techniques like cryo-EM and X-ray crystallography? Could AlphaFold eventually replace these methods, or will it always be a complementary tool?
 - Biology is not single static structures and large scale complexes are still not predicted accurately, in particular in all active states

- What exactly are covariation features?
 - DCA etc
- Can AlphaFold predict structural changes related to pH, ionic strength and solvent that are relevant for pharmacological formulations of monoclonal antibodies or peptide based drugs?
 - No it is trained on all structures - ignoring this. Some other work (BioEMU etc) is trying to take this into account - but they are very different (Atom based). Water is important
- How does alphafold make up for the occurrence of certain ligands that have been shown to exist in the ground truth?
 - AF2 - learns this implicitly. In AF3 it needs to be included specifically
- What are some of the challenges associated with developing the models for predicting protein ligand interactions or more multi chain complexes?
 - See recent paper <https://www.biorxiv.org/content/10.1101/2025.02.03.636309v2>

- Can tools like alpha fold be used to design proteins that do not aim to fix a condition but act as enhancements that improve normal functions?
 - Not sure I understand..
- How can Alphafold3 be improved in the future?
 - More data ? Dynamics ? Better performance?
- What is the main advantage of Alphafold3 over Alphafold2? Is there any logistic changing?
 - You can now handle all type of molecules, not only proteins
- What computational trade-offs exist between model accuracy, interpretability, and scalability when predicting large protein complexes?
 - Limits of size in AF2/3 about 5000 tokens.

- When the protein interaction prediction results come out with AlphaFold, what are the threshold of ipTM and pTM values that we take to consider that there is actual interact between two proteins?
 - ipTM approximately 0.5 - but other scores might be better
- In ligand binding and drug discovery, can we use the results from AlphaFold to reduce some animal testing? If it could help minimize animal testing, how can we enhance it to perform better in drug design?
 - Most likely not - more likely to replace in vitro experimental work - animals are still very complex
- How to interpret the relationship between pTM/ipTM scores and protein interactions ?
 - Bimodal distribution cutoff avout 0.5

- In the YouTube video linked, it was mentioned that the weights remain the same in the recycling steps. How can this improve the model's performance?
 - The new input comes mainly from providing a “template” - the distance distribution. Nudges the predictions in one (better) direction
- How does AlphaFold tackle the challenges related to Multiple Sequence Alignments (MSAs), specifically the issues of poor performance for "orphan proteins" and the difficulty in distinguishing between direct and indirect coevolutionary signals? By transformer in evoformer and pairformer? then how do the alphafold predict multi-state protein
 - Indirect signals it has learned. It was trained on some “bad” alignment to not to be too sensitive

- What is the main issue that future alphafold iterations would want to address?
In other words, what would be the number one limitation of Alphafold that needs to be solved?
 - Better performance for ligands RNA
- In what scenarios would traditional structural biology methods be preferred due to some limitations of AlphaFold3?
 - When you want to not only know the structure but study effects of environment etc.
- How do diffusion models contribute to accuracy of AlphaFold3, and what advantages do they offer over previous methods?
 - Most likely not that much (but they are faster)
- What's the importance of knowing if the coevolution is direct or indirect?
 - Necessary for contact predictions

Excercises etc

Video with more details about AF2

<https://www.youtube.com/watch?v=92j-7b-sRPw>

Reading a more biological perspective:

<https://www.nature.com/articles/s41592-021-01365-3>

Using the server

The objective is to get familiar with the [AlphaFold2](#) and the [AlphaFold2 database](#) and [AlphaFold3](#)

You will have to generate a model of protein structure from the primary amino acid sequence using either [ColabFold](#) or AlphaFold3. This should not take more than about 15 minutes. Most likely AlphaFold3 is faster and easier to use, but less flexible. The performance for these targets will be similar.

Exercise 1

Run AlphaFold2/3 on a monomeric sequence

>1SHW_1|Chain A|Ephrin-A5|Mus musculus (10090)

VADRYAVYWNSSNPRFQRGDYHIDVCINDYLDVFCPHYEDSVPEDKTERYVLYMVNFDGYSACDHTSKGFKRWECNRPHSPNGPLKF
SEKFQLFTPFSLGFEFRPGREYFYISSAIPDNGRRSCLKLKVFVRPTNSCM

Generate the protein structure using ColabFold/AlphaFold3 . This could take up to 10-15 minutes so you can try answering questions involving the AlphaFold Database in the mean time. If you are using AlphaFold3 you can go ahead and submit the other sequences while you waiting. You are allowed to run 20 jobs per day.

How many sheets and helices have this protein?

Exercise 2

Run AlphaFold2 on dimeric sequences

>1SHW_1|Chain A|Ephrin-A5|Mus musculus (10090)

VADRYAVYWSSNPRFQRGDYHIDVCINDYLDVFCPHYEDSVPEDKTERYVLYMVNFDGYSACDHTSKGFKRWECNRPHSPNGPLKFSEKFQLFT
PFSLGFEFRPGREYFYISSAIPDNGRRSCLKLKVFRPTNSCM

>1SHW_2|Chain B|Ephrin type-B receptor 2|Mus musculus (10090)

VEETLMDSTTATAELGWMVHPPSGWEEVSGYDENMNTIRTYQVCNVFESSQNNWLRTKFIRRRGAHRIHVEMKFSVRDCSSIPSVPGSCKETFNL
YYYEADFDLATKTFPNWMENPWVKVDITAADESFSQVDLGGVMKINTEVRSFGPVSRRNGFYLAQDYGGCMSLIAVRVFYRKCPR

- Generate the protein structure using ColabFold/AlphaFold3 and look at the interface between the two proteins. How would you describe it.
- Report the pTM [ptm] scores for the dimeric AlphaFold model generated.
- What purpose does MMseqs2 and HHsearch serve in the ColabFold notebook?
-

Multimers

Run AlphaFold2 on dimeric sequences

>1SHW_1|Chain A|Ephrin-A5|Mus musculus (10090)

VADRYAVYWNSSNPRFQRGDYHIDVCINDYLDVFCPHYEDSVPEDKTERYVLYMVNFDGYSACDHTSKGFKRWEENRPHSPNGPLKF
SEKFQLFTPFSLGFEFRPGREYFYISSAIPDNGRRSCLKLKVFRPTNSCM

>1SHW_2|Chain B|Ephrin type-B receptor 2|Mus musculus (10090)

VEETLMDSTTATAELGWMVHPPSGWEEVSGYDENMNTIRTYQVCNVFESSQNNWLRTKFIRRRGAHRIHVEMKFESVRDCSSIPSVPGS
CKETFNLYYYEADFDLATKTFPNWMENPWVKVDTIAADESFSQVDLGGVMKINTEVRSFGPVSRNGFYLAQDYGGCMSLIAVRVY
RKCPR

Generate the protein structure using ColabFold or AlphaFold3. Make sure you have a white background and upload a high resolution (dpi=300) png image. This could take up to 10-15 minutes so you can try answering questions involving the AlphaFold Database in the mean time.

Exercise 3 (AlphaFoldDB)

- Looking at the predicted model for the sequence from UniProt entry P94485, in AlphaFold3v which residue does SER 11 form a hydrogen bond with?
- What are the max and min length of the protein sequences predicted in AlphaFold database?
- Which terminal is disordered in AF2 model for Uniprot Q00768?

More excersizes

EMBL-EBI

<https://www.ebi.ac.uk/training/online/courses/alphafold/accessing-and-predicting-protein-structures-with-alphafold/predicting-protein-structures-with-colabfold-and-alphafold-colab/>