Single-sequence protein structure prediction by integrating protein language models

Xiaoyang Jing, Fandi Wu, Xiao Luo, and Jinbo Xu

Presented by:
Sumit Tarafder

Landscape of structure prediction

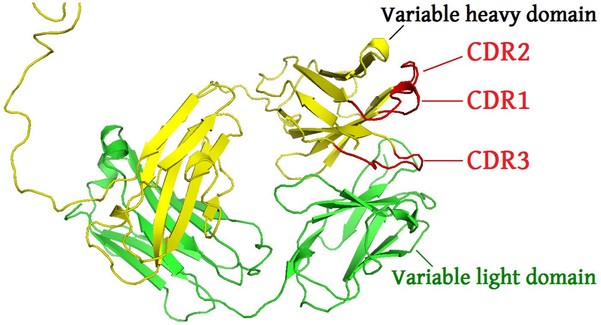
- Advances in computational structure prediction
  - Deep learning
  - Co-evolutionary information (MSA)

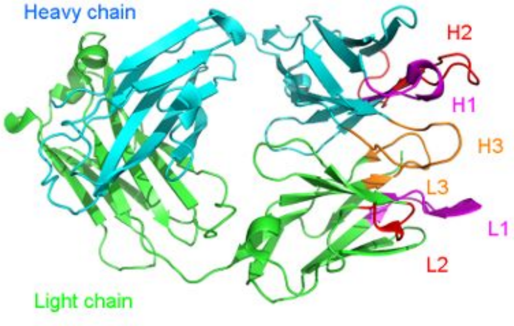
- Caveats
  - Protein folds in the absence of sequence homologs
  - Time and complexity of sequence search
  - MSA, non natural

- Lack of efficiency
  - Flexible region prediction such as loop or CDR regions
    - Weak presence of evolutionary information in these regions
  - Single point mutation effects

Complementarity-determining region (CDR)

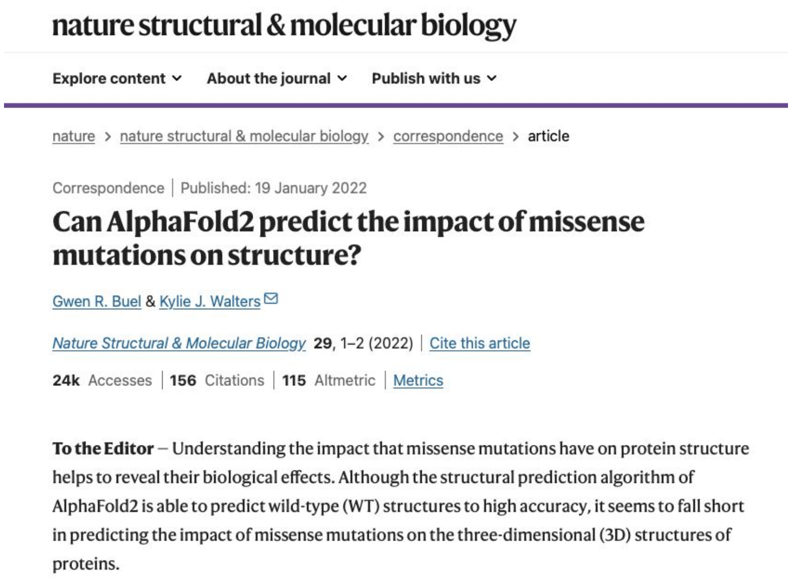
- Antibody (**Ab**) or immunoglobulin (**Ig**)
    - Responsible to bind to antigens
    - 4 chains (2 heavy, 2 light)
    - Constant structure in the framework region (**Fr**)
    - Large structure variability in the CDR regions
- CDR
    - Highly variable regions in antibody
    - Shape complements that of an antigen.
    - Classified using ANARCI tool
- CDR3
    - Highly variable among the three regions





Wild type vs missense mutation

- Potential limitation of AF2
  - Structure-disruptive folding
  - Trained one WT or homologus sequences
- Missense mutations
  - Frequently associate with human diseases and single amino-acid mutations



nature structural & molecular biology

Explore content
About the journal
Publish with us
nature > nature structural & molecular biology > correspondence > article

Correspondence | Published: 19 January 2022

Can AlphaFold2 predict the impact of missense mutations on structure?

Gwen R. Buel & Kylie J. Walters

Nature Structural & Molecular Biology **29**, 1–2 (2022) | Cite this article

24k Accesses | **156** Citations | **115** Altmetric | Metrics

To the Editor — Understanding the impact that missense mutations have on protein structure helps to reveal their biological effects. Although the structural prediction algorithm of AlphaFold2 is able to predict wild-type (WT) structures to high accuracy, it seems to fall short in predicting the impact of missense mutations on the three-dimensional (3D) structures of proteins.

RaptorX-Single & RaptorX-Single-Ab

- **RaptorX-Single**
  - MSA free method
    - Leverages multiple language model information
    - ESM-1b, ESM-1v, ProtTrans
  - Outperforms AlphaFold2 in
    - Orphan protein structure prediction
    - Single mutation effect prediction
    - Comparatively scalable

- **RaptorX-Single-Ab**
  - Focused on antibody structure prediction
    - Outperforms all other methods
  - Incorporates fine-tuning

SOTA methods for single-sequence prediction

- ESMFold
- OmegaFold
- trRosettaX-Single
- HelixFold-Single
- RGN
- AlphaFold2 (Single)

Baseline methods in this work

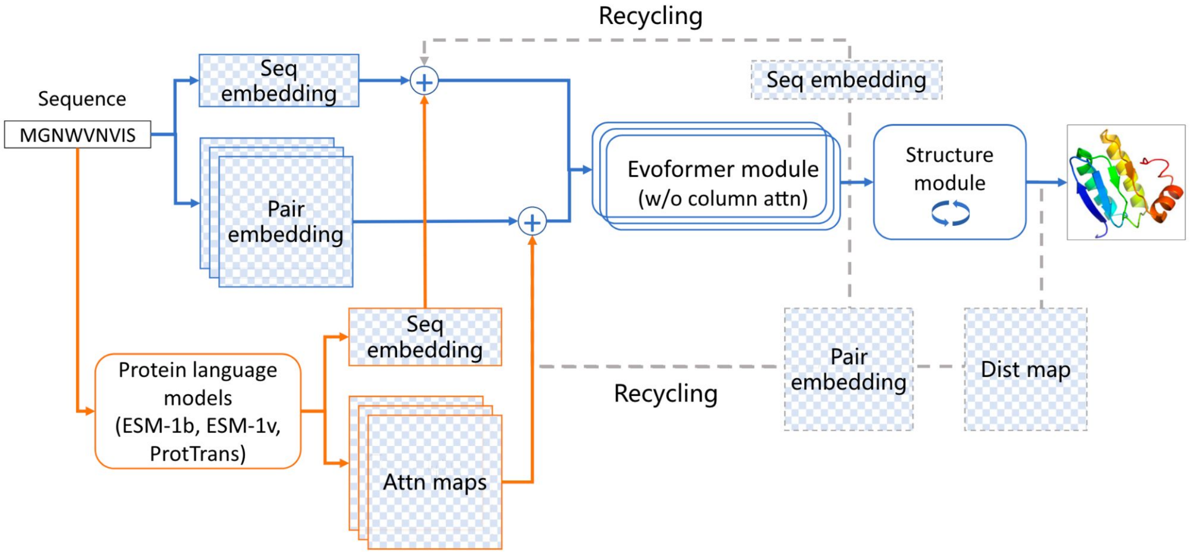
- ESMFold
- OmegaFold
- HelixFold-Single
- AlphaFold2 (MSA)
    - Without templates
- AlphaFold2 (Single)
    - No MSA, template

Antibody specific methods

- DeepAb
- IgFold
- EquiFold
  - Solely depends on sequence for prediction

Architecture

- Modified Evoformer
  - 24 layers
- Structure module
  - Linear layer to integrate attention values
- Initial pair embedding
  - Relative positional encoding in the pairwise embedding



pLMs

- ESM-1b (~650 M parameters)
  - UniRef50 - 27.1 million protein sequences
- ESM-1v
  - Uniref90 with 98 million protein sequences
- ProtTrans (3 billion parameters)
  - Newer UniRef50 of 45 million sequences

Training dataset

- The training data consist of ~340 k proteins.
    - 80,852 proteins released before January 2020 in PDB
        - 40% sequence identity clusters (BC100By40)
    - The remaining 264 k proteins - predicted by AlphaFold2 (denoted as distillation data)
        - Extracted from Uniclust30_2018_08
        - < 30% sequence similarity
- Each epoch
    - One protein is randomly sampled
        - From each cluster in BC100By40
        - From distillation data by the ratio of 1:3 between BC100By40 and the distillation data.

Benchmark datasets

- Three antibody datasets
  - SAbDab-Ab (202 Ab)
  - IgFold-Ab (67 Ab)
  - Nanobody (60 Ab)
- One orphan protein dataset
  - **11** proteins released between 01 January 2020, and 12 April 2022
  - No homologs in BFD, MGnify, Uniref90 and Uniclust30
- Rocklin dataset: Single mutation effects dataset
  - 14 native and de novo designed proteins and their stability measures of 10,674 single mutations.
  - The stability was evaluated using thermal and chemical denaturation.

Training

- Training losses
  - Pairwise loss (trRosetta)
    - Distogram loss
    - Distance loss
    - Orientation
  - Structure loss
    - Frame Aligned Point Error loss with a clamp of 20 Å
    - pLDDT loss.
- Recycling
  - Randomly sampled from 0 to 3
- 150 epochs
- RaptorX-Single (1b) - ESM-1b
- RaptorX-Single (1v) - ESM-1v
- RaptorX-Single (pt) - ProtTrans
- RaptorX-Single (All 3)

Fine-tuning for antibody prediction

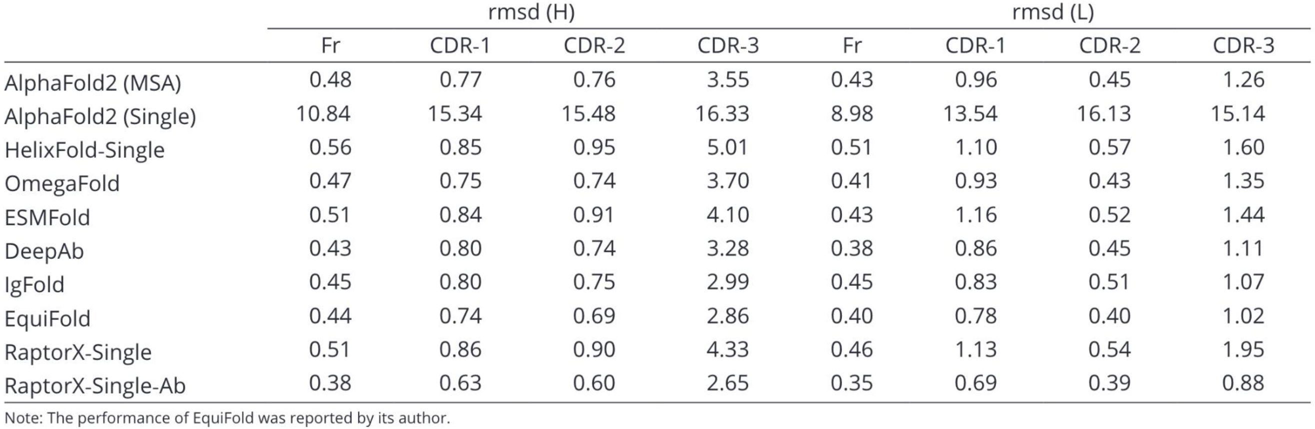
- An antibody training set for fine-tuning.
    - Experimental structures from SAbDab (20) released before 2021/03/31
    - 5,033 heavy and light chains.
    - Validation set - 178 antibody structures
- All four models 50 epochs
    - RaptorX-Single-Ab (1b)
    - RaptorX-Single-Ab (1v)
    - RaptorX-Single-Ab (pt)
    - RaptorX-Single-Ab.

Evaluation metrics

- For antibodies
  - Backbone rmsd (Using PyRosetta)
    - Framework (Fr)
    - CDR (CDR-1, CDR-2, and CDR-3); Heavy and light chains separately
- For orphan targets
  - TM-score
  - Global distance test–total score (GDT_TS)
  - Global distance test–high accuracy (GHT_HA)
- Single mutation effect prediction
  - Pearson correlation coefficient
    - Between the predicted structure changes and the stability data
    - Structure change = ΔTMscore

Average rmsd of on the IgFold-Ab dataset

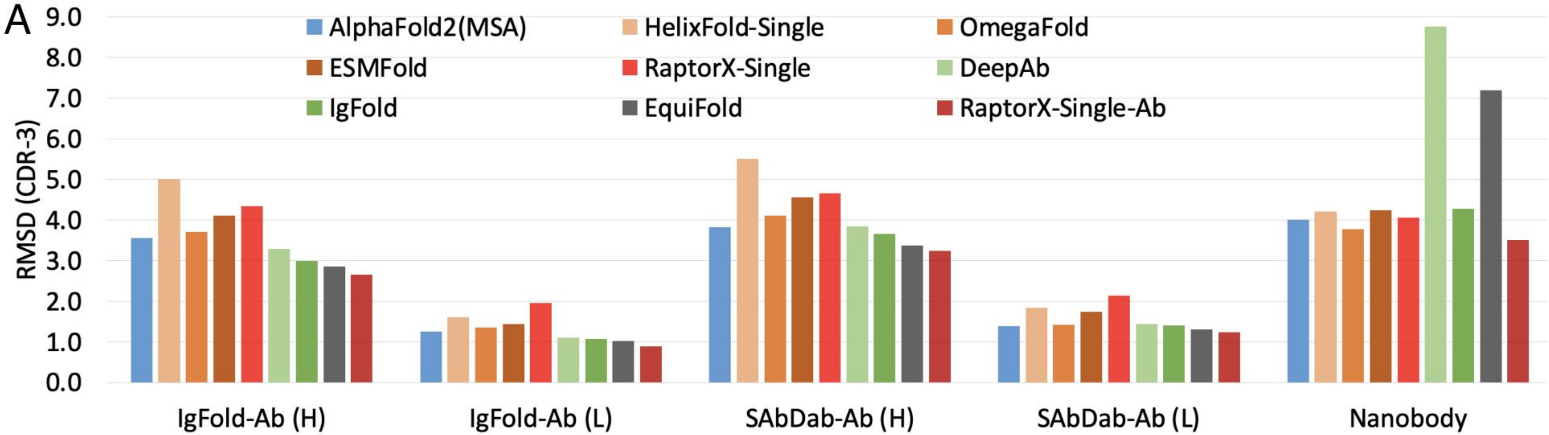
- AF2 (MSA) not as good as Ab-specific methods
- Difference in fine-tuning vs trivial methods



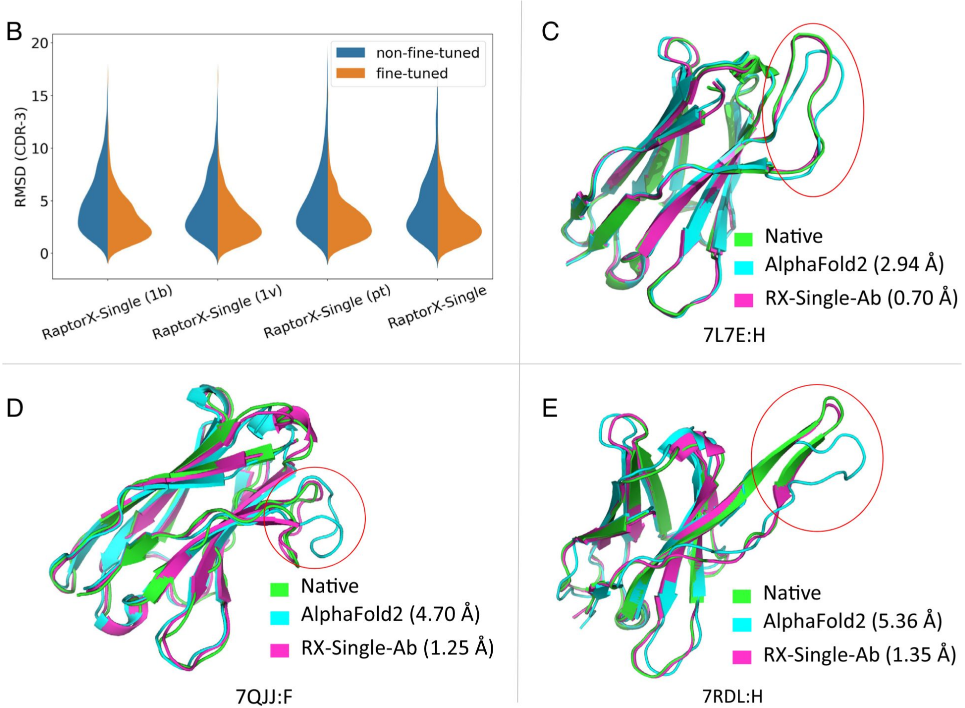
	rmsd (H)				rmsd (L)			
	Fr	CDR-1	CDR-2	CDR-3	Fr	CDR-1	CDR-2	CDR-3
AlphaFold2 (MSA)	0.48	0.77	0.76	3.55	0.43	0.96	0.45	1.26
AlphaFold2 (Single)	10.84	15.34	15.48	16.33	8.98	13.54	16.13	15.14
HelixFold-Single	0.56	0.85	0.95	5.01	0.51	1.10	0.57	1.60
OmegaFold	0.47	0.75	0.74	3.70	0.41	0.93	0.43	1.35
ESMFold	0.51	0.84	0.91	4.10	0.43	1.16	0.52	1.44
DeepAb	0.43	0.80	0.74	3.28	0.38	0.86	0.45	1.11
IgFold	0.45	0.80	0.75	2.99	0.45	0.83	0.51	1.07
EquiFold	0.44	0.74	0.69	2.86	0.40	0.78	0.40	1.02
RaptorX-Single	0.51	0.86	0.90	4.33	0.46	1.13	0.54	1.95
RaptorX-Single-Ab	0.38	0.63	0.60	2.65	0.35	0.69	0.39	0.88

Note: The performance of EquiFold was reported by its author.

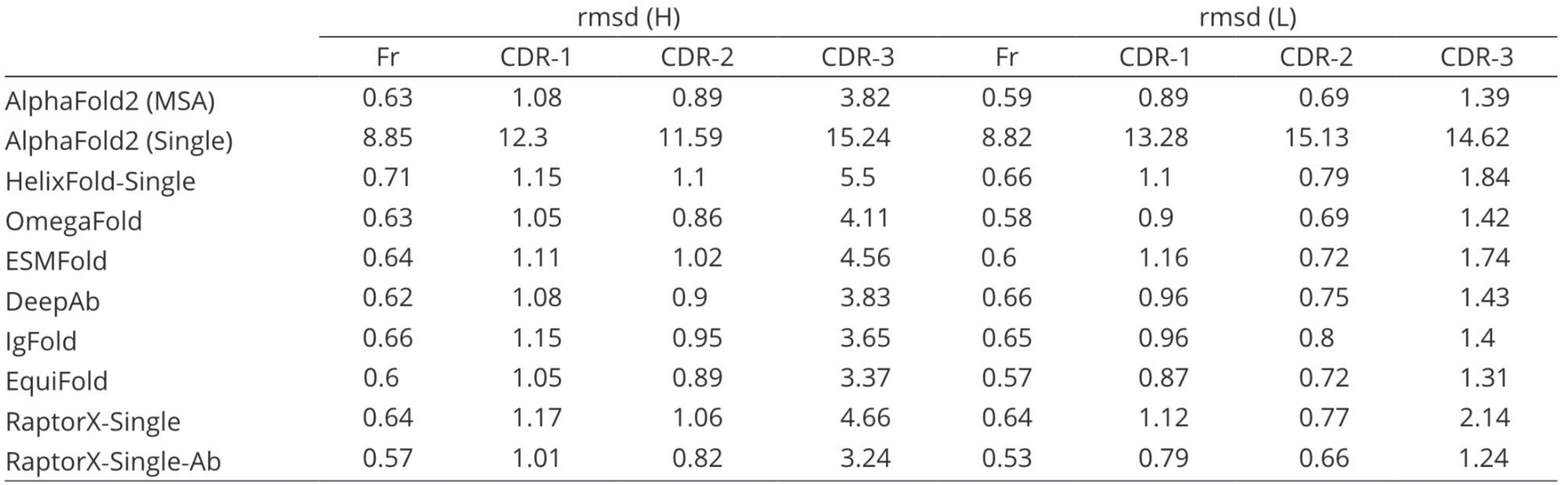
Average rmsd of predicted CDR-3 regions



Performance comparison on antibody structure prediction



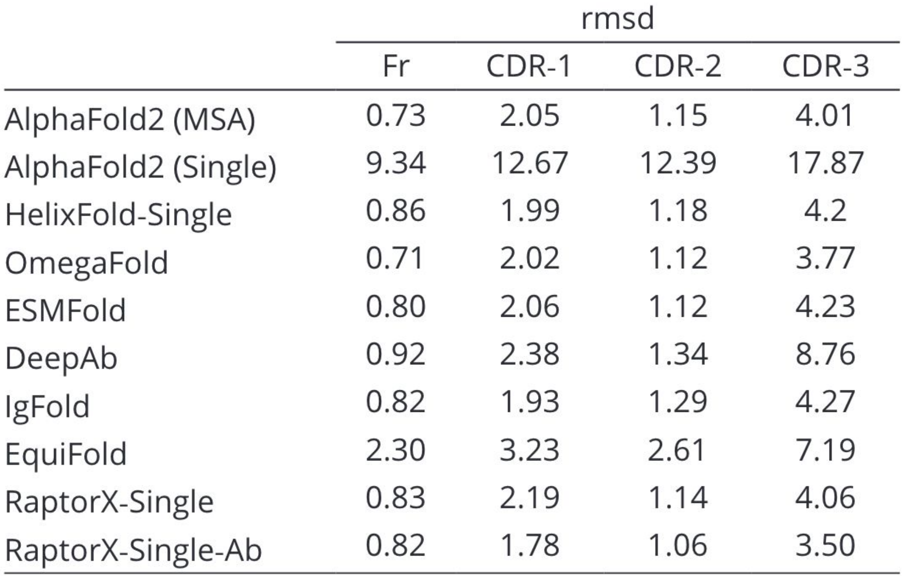
The average rmsd on the SAbDab-Ab dataset



	rmsd (H)				rmsd (L)			
	Fr	CDR-1	CDR-2	CDR-3	Fr	CDR-1	CDR-2	CDR-3
AlphaFold2 (MSA)	0.63	1.08	0.89	3.82	0.59	0.89	0.69	1.39
AlphaFold2 (Single)	8.85	12.3	11.59	15.24	8.82	13.28	15.13	14.62
HelixFold-Single	0.71	1.15	1.1	5.5	0.66	1.1	0.79	1.84
OmegaFold	0.63	1.05	0.86	4.11	0.58	0.9	0.69	1.42
ESMFold	0.64	1.11	1.02	4.56	0.6	1.16	0.72	1.74
DeepAb	0.62	1.08	0.9	3.83	0.66	0.96	0.75	1.43
IgFold	0.66	1.15	0.95	3.65	0.65	0.96	0.8	1.4
EquiFold	0.6	1.05	0.89	3.37	0.57	0.87	0.72	1.31
RaptorX-Single	0.64	1.17	1.06	4.66	0.64	1.12	0.77	2.14
RaptorX-Single-Ab	0.57	1.01	0.82	3.24	0.53	0.79	0.66	1.24

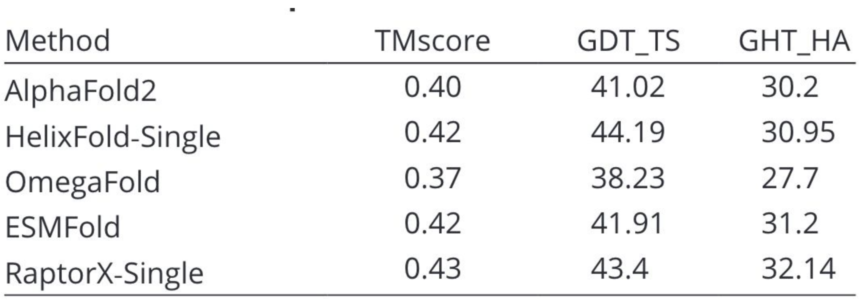
The average rmsd on the Nanobody dataset

- Nanobody is an increasingly popular modality for therapeutic development.
- Lacks a second Ig chain
- Increased CDR3 loop length,
  - Challenging
- EquiFold fails
  - Significance of pLMs



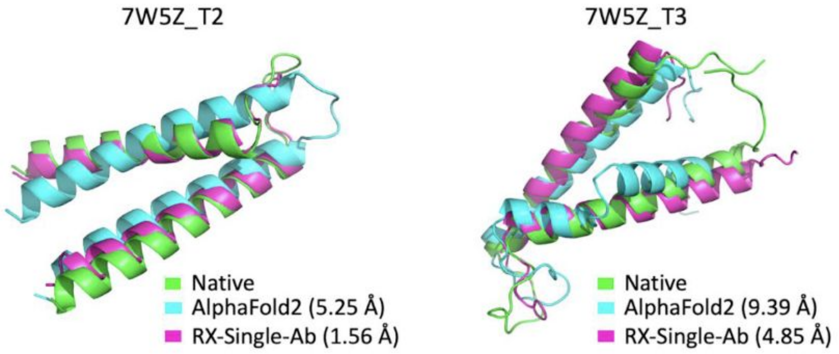
	rmsd			
	Fr	CDR-1	CDR-2	CDR-3
AlphaFold2 (MSA)	0.73	2.05	1.15	4.01
AlphaFold2 (Single)	9.34	12.67	12.39	17.87
HelixFold-Single	0.86	1.99	1.18	4.2
OmegaFold	0.71	2.02	1.12	3.77
ESMFold	0.80	2.06	1.12	4.23
DeepAb	0.92	2.38	1.34	8.76
IgFold	0.82	1.93	1.29	4.27
EquiFold	2.30	3.23	2.61	7.19
RaptorX-Single	0.83	2.19	1.14	4.06
RaptorX-Single-Ab	0.82	1.78	1.06	3.50

Average model quality on Orphan dataset



Method	TMscore	GDT_TS	GHT_HA
AlphaFold2	0.40	41.02	30.2
HelixFold-Single	0.42	44.19	30.95
OmegaFold	0.37	38.23	27.7
ESMFold	0.42	41.91	31.2
RaptorX-Single	0.43	43.4	32.14

- Why RaptorX-Single-Ab in figure?
- Superior in loop and alpha-helix region
- Neither MSA nor language model can predict the fold
  - **Implicitly MSA dependent**



Mutational effect prediction

- RaptorX-single outperforms on 9 out of 14 targets
- AF2 (single) outperforms AF2 (MSA)
  - Advantage of single-seq method in this type of studies

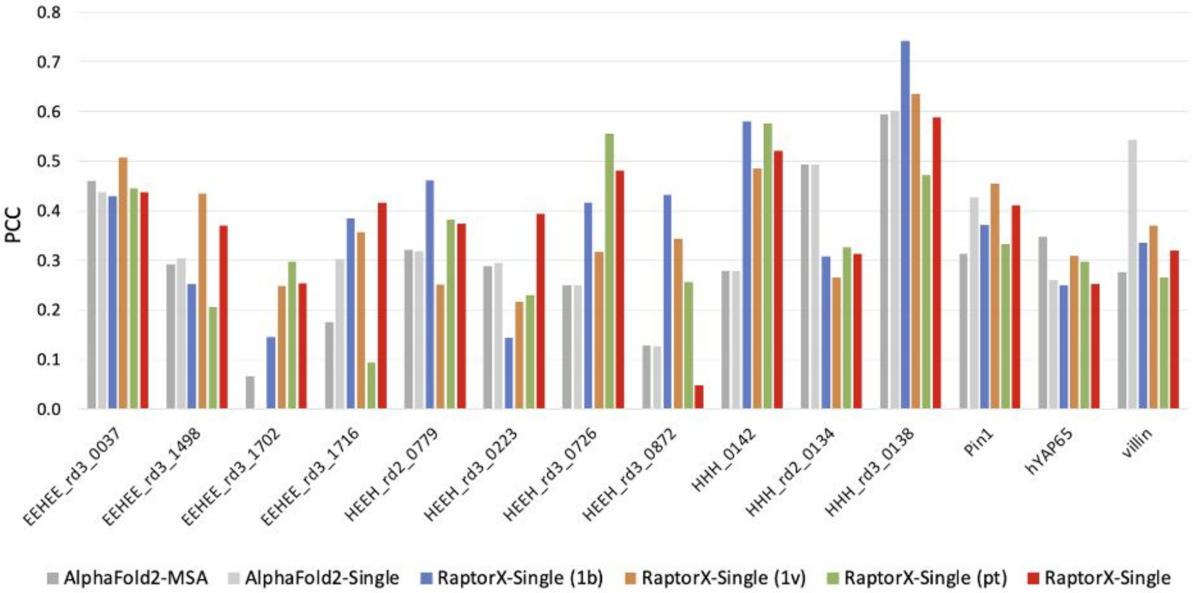
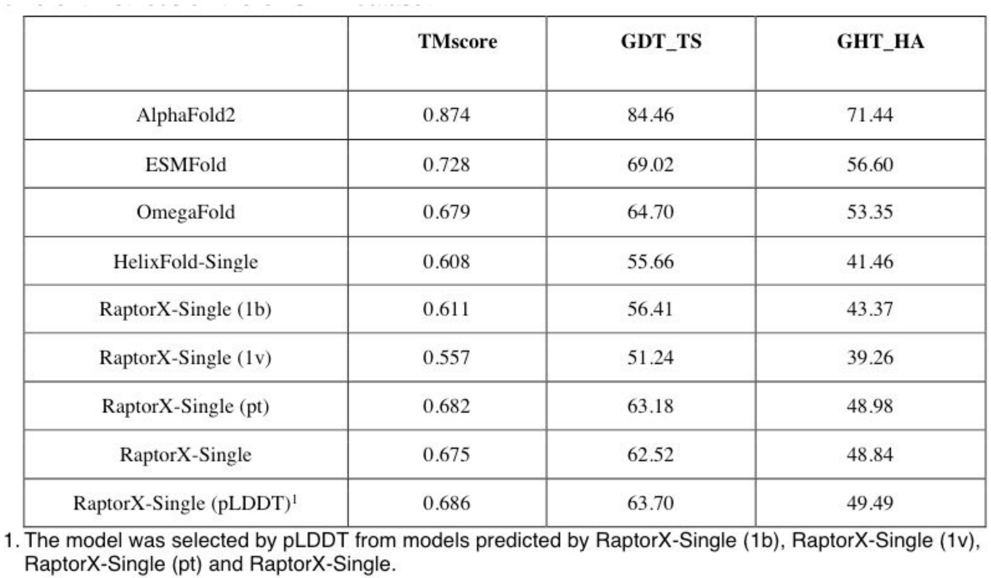


Fig : The PCC between predicted structure change and stability change of all targets.

Performance on CASP14 dataset (60 targets)

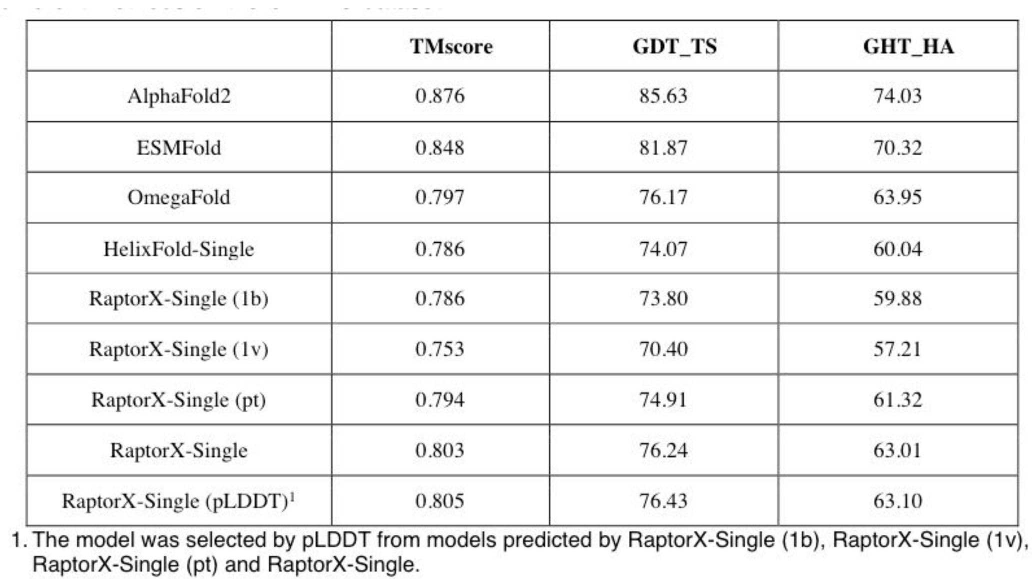
- AlphaFold2 is the best
  - Importance of MSA
- ESMFold outperforms other single-seq methods
  - Importance of pLMs
- RaptorX-single (pt) is better than other two pLMs.



	TMscore	GDT_TS	GHT_HA
AlphaFold2	0.874	84.46	71.44
ESMFold	0.728	69.02	56.60
OmegaFold	0.679	64.70	53.35
HelixFold-Single	0.608	55.66	41.46
RaptorX-Single (1b)	0.611	56.41	43.37
RaptorX-Single (1v)	0.557	51.24	39.26
RaptorX-Single (pt)	0.682	63.18	48.98
RaptorX-Single	0.675	62.52	48.84
RaptorX-Single (pLDDT)[1]	0.686	63.70	49.49

[1] The model was selected by pLDDT from models predicted by RaptorX-Single (1b), RaptorX-Single (1v), RaptorX-Single (pt) and RaptorX-Single.

Performance on CAMEO dataset (194 targets)

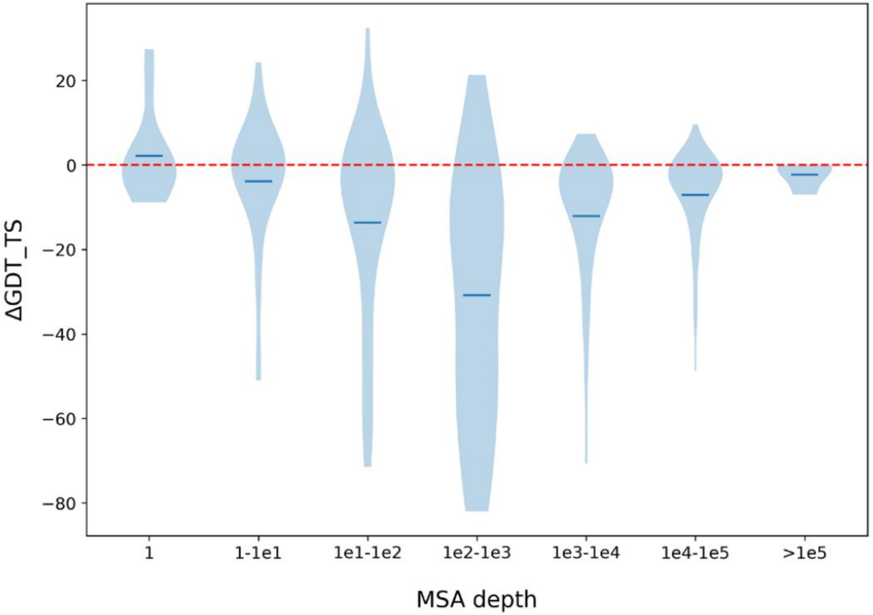


	TMscore	GDT_TS	GHT_HA
AlphaFold2	0.876	85.63	74.03
ESMFold	0.848	81.87	70.32
OmegaFold	0.797	76.17	63.95
HelixFold-Single	0.786	74.07	60.04
RaptorX-Single (1b)	0.786	73.80	59.88
RaptorX-Single (1v)	0.753	70.40	57.21
RaptorX-Single (pt)	0.794	74.91	61.32
RaptorX-Single	0.803	76.24	63.01
RaptorX-Single (pLDDT)[1]	0.805	76.43	63.10

1. The model was selected by pLDDT from models predicted by RaptorX-Single (1b), RaptorX-Single (1v), RaptorX-Single (pt) and RaptorX-Single.

Effect of MSA depth on prediction quality

- Are single-seq methods implicitly making use of homologs?
- Comparison of RaptorX-Single with AF2 (MSA)
  - CASP14 and CAMEO [homolog rich]
  - 99 targets more; no homolog in Uniclust30
- ΔGDT-TS = RaptorX-Single - AF2



Observations:

- Significantly underperforms for depth = 100-1000
- Comparable for low and high depths
- pLMs implicitly learn coevolution information of large-sized protein
  - Avg. length of >1e4 = 411

Limitations or Future works

- Only outperforms Alphafold2 after fine-tuning
- No comparison with other stability prediction methods
- Did not include RGN despite mentioning in the paper
- Choice of pLMs
- Interconverting states in solution
    - Range of states with likelihood

Future works

- VH-VL complex for antibody structure prediction
- No method can predict the fold of orphan proteins
    - Implicit use of homologs through pLMs
    - Prediction directly from sequence