OVERVIEW OF TOOLS FOR PREDICTING THE 3D STRUCTURE OF A PROTEIN BASED ON THE SEQUENCE
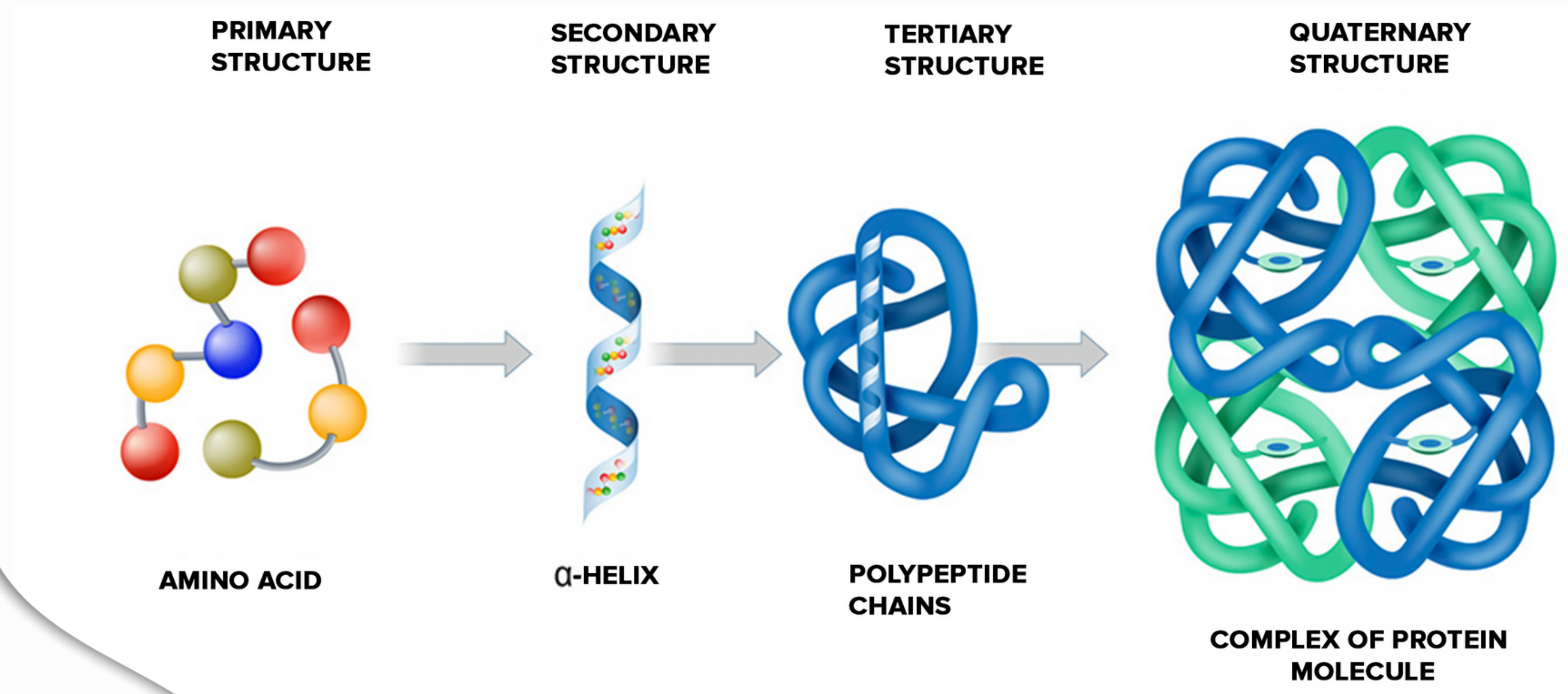
Presented by Marta Korpacz
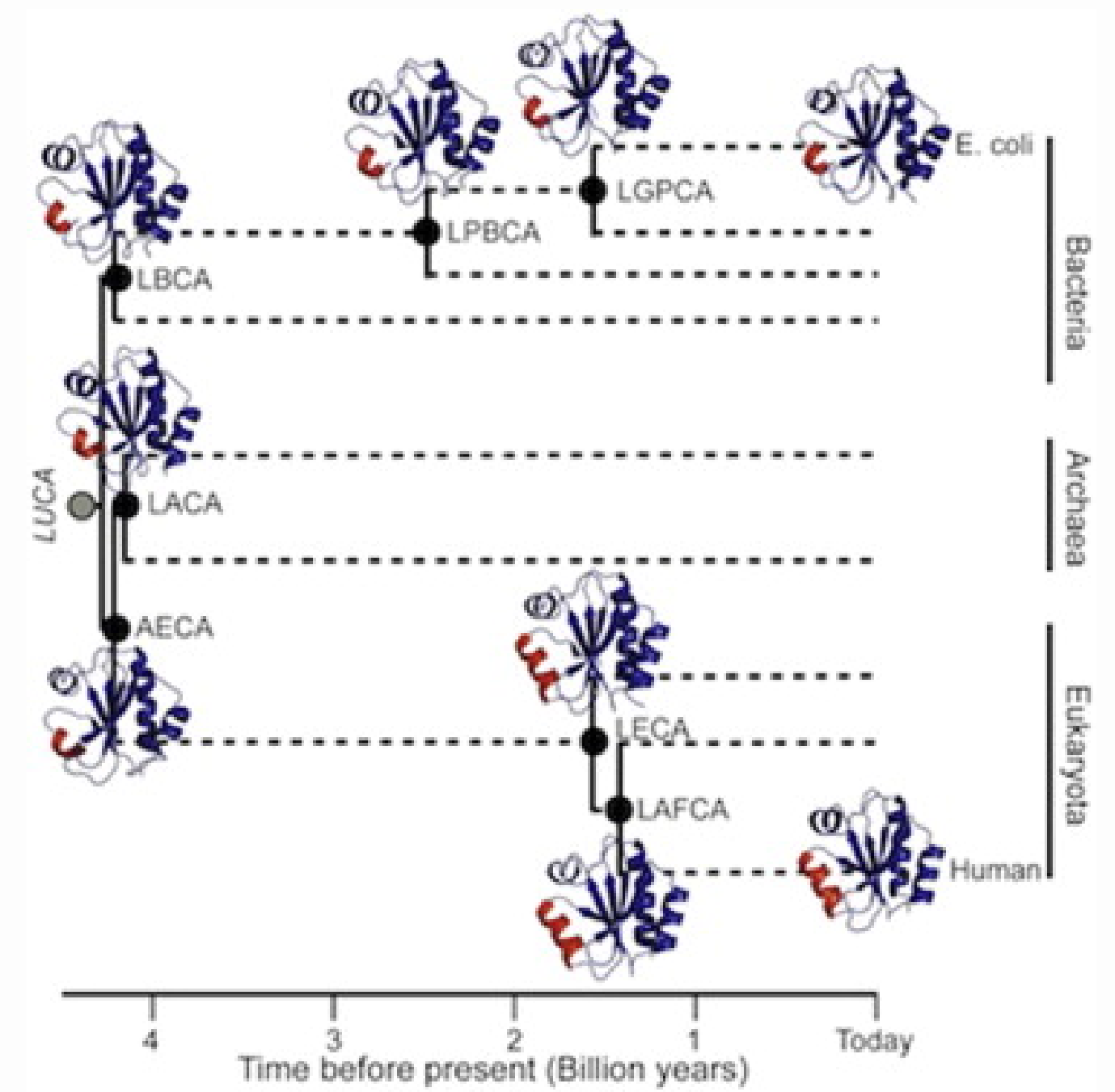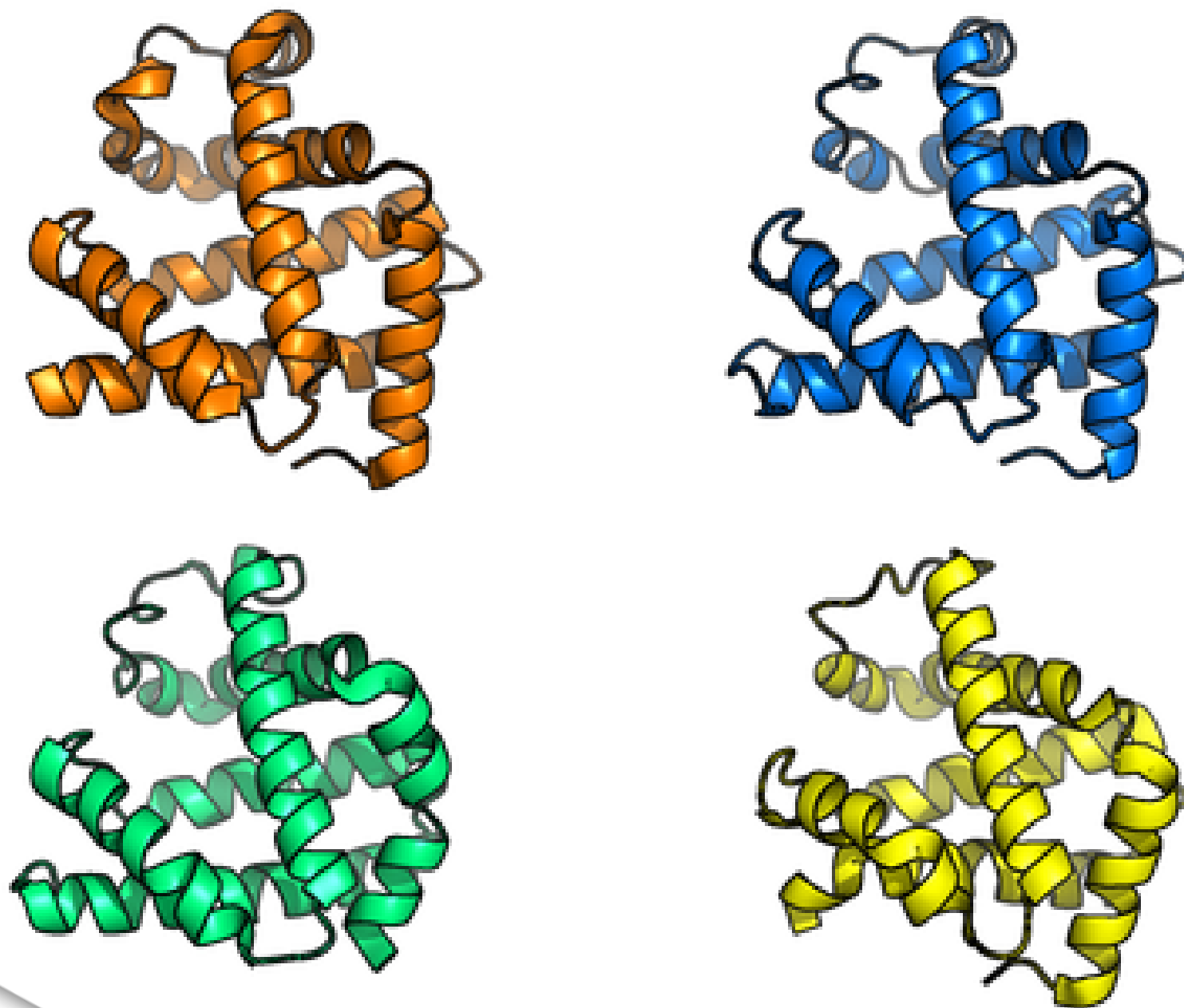
The native structure is determined only by the protein's amino acid sequence.

PRIMARY STRUCTURE

SECONDARY STRUCTURE

TERTIARY STRUCTURE

QUATERNARY STRUCTURE

AMINO ACID

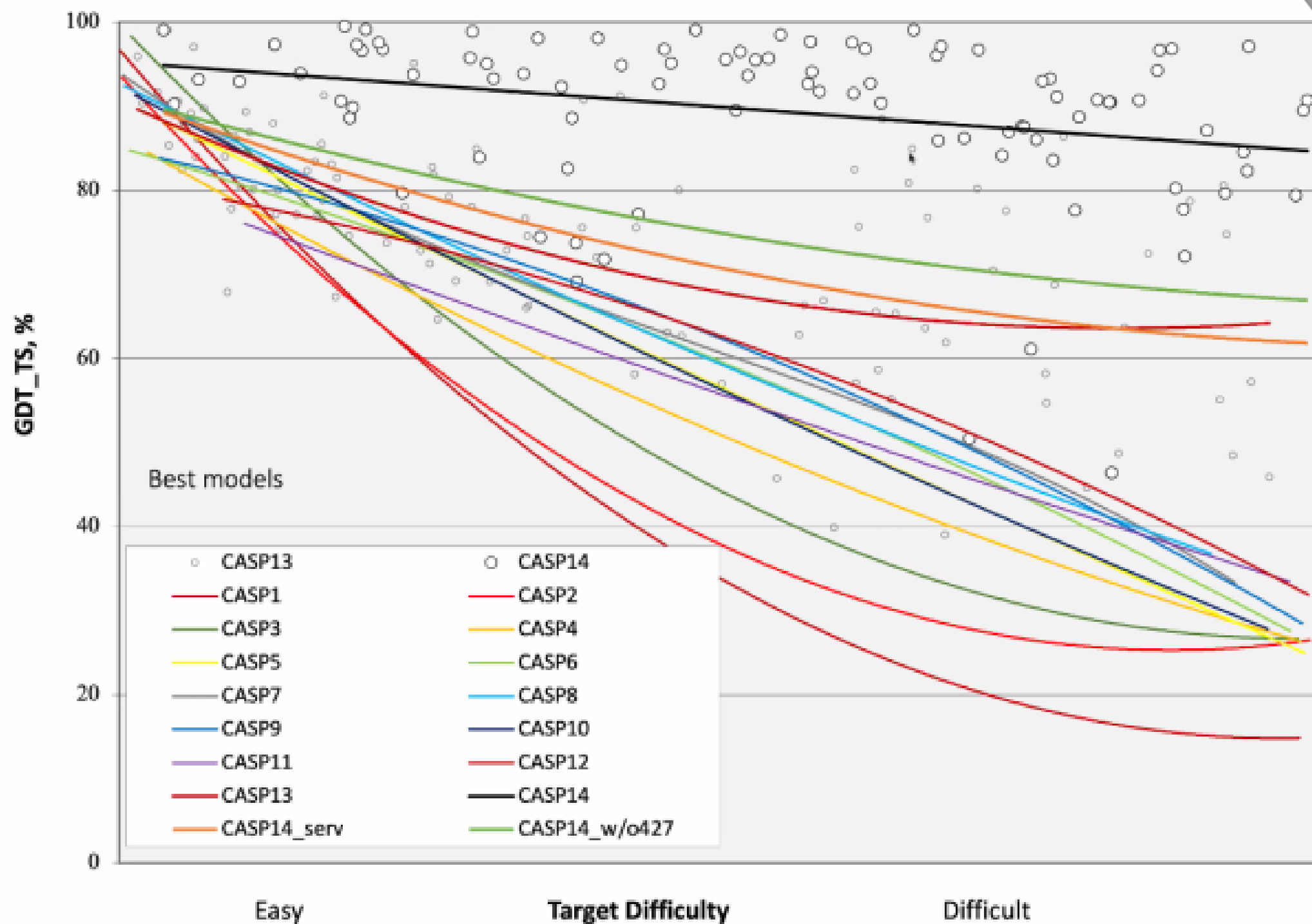α-HELIX

POLYPEPTIDE CHAINS

COMPLEX OF PROTEIN MOLECULE

CONSERVATION

Protein structure and function are more highly conserved than sequence!

CASP8 target 512-D1
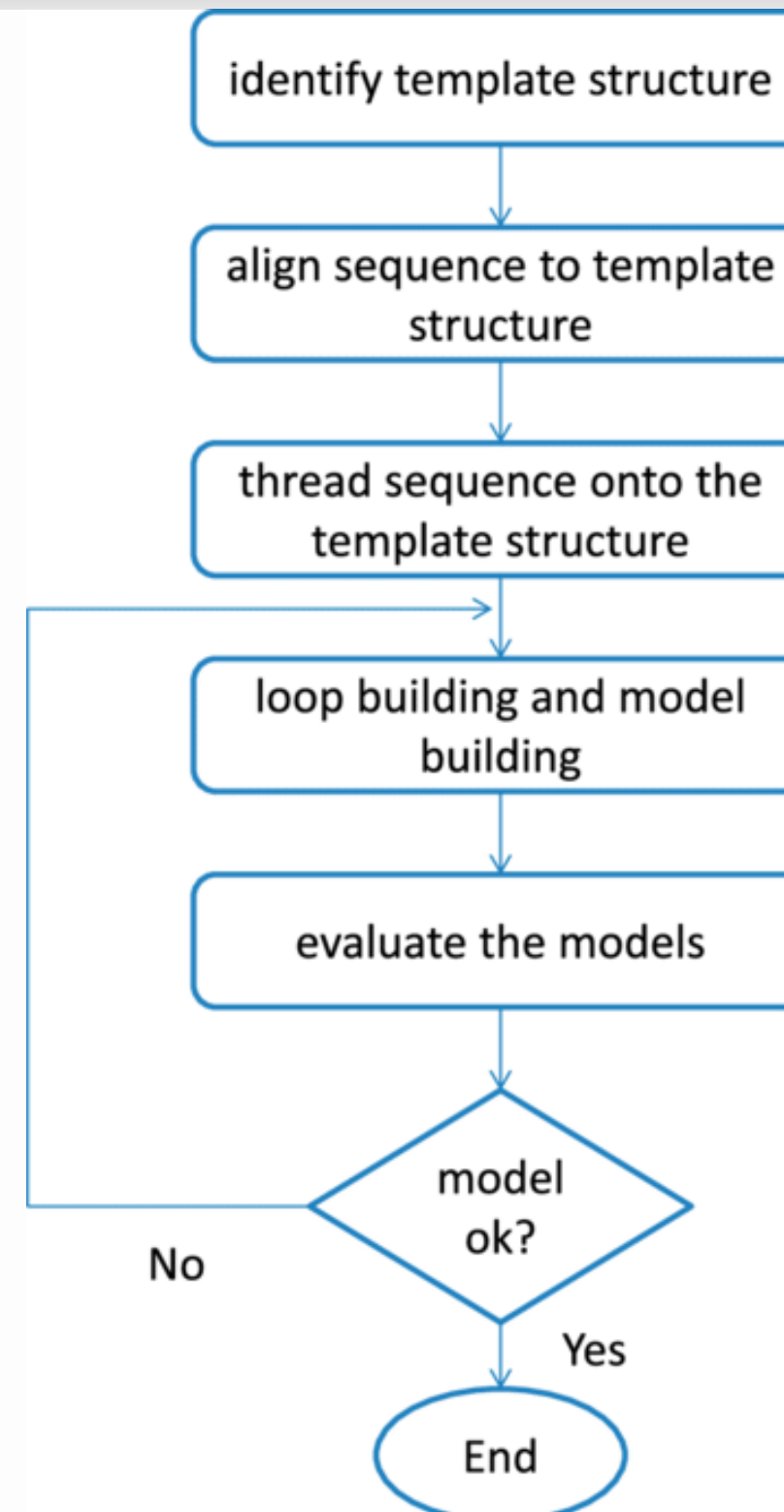all models
(3dsm)

Best models

- CASP13
- CASP14
- CASP1
- CASP2
- CASP3
- CASP4
- CASP5
- CASP6
- CASP7
- CASP8
- CASP9
- CASP10
- CASP11
- CASP12
- CASP13
- CASP14
- CASP14_serv
- CASP14_w/o427
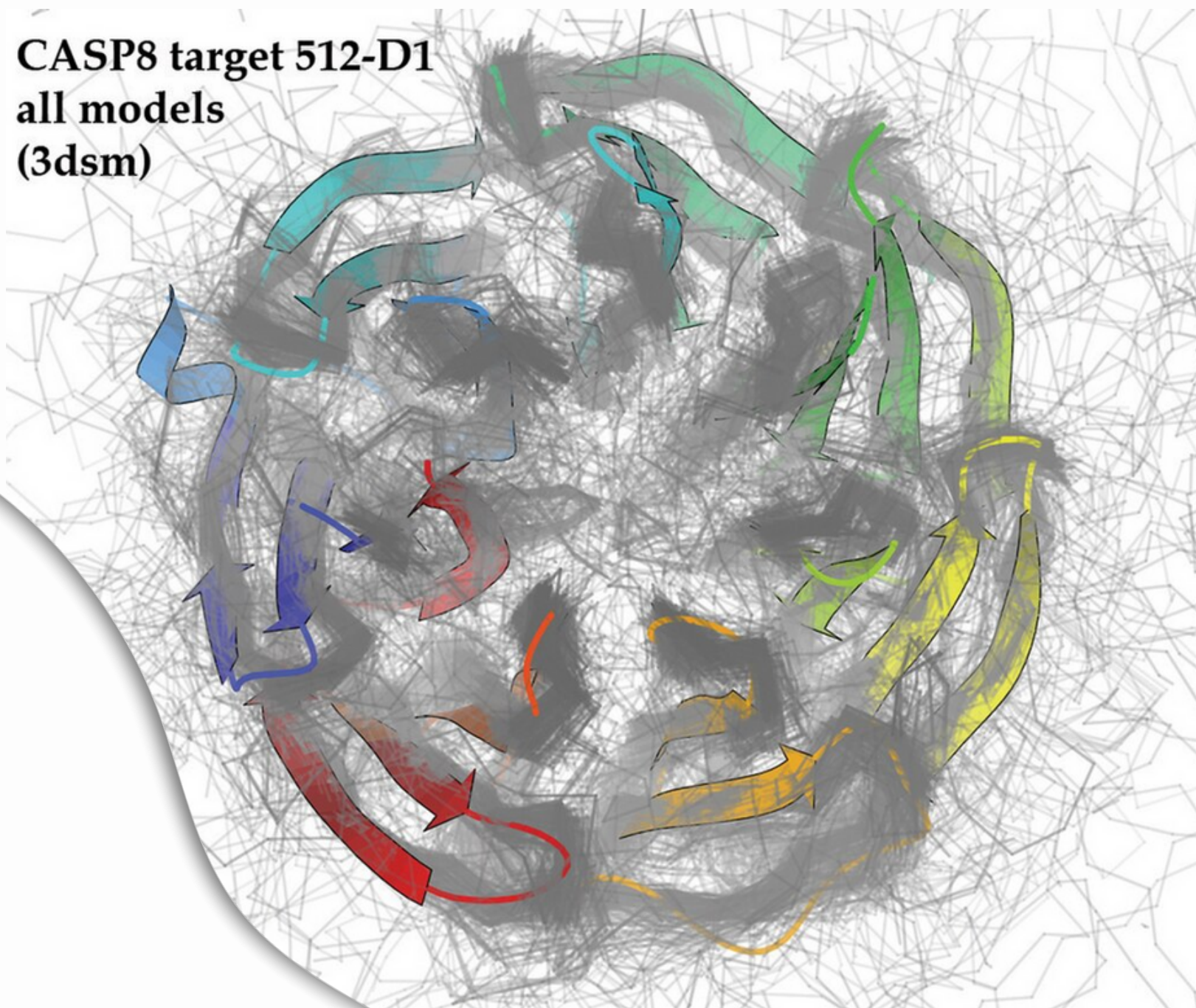
HOMOLOGY MODELLING

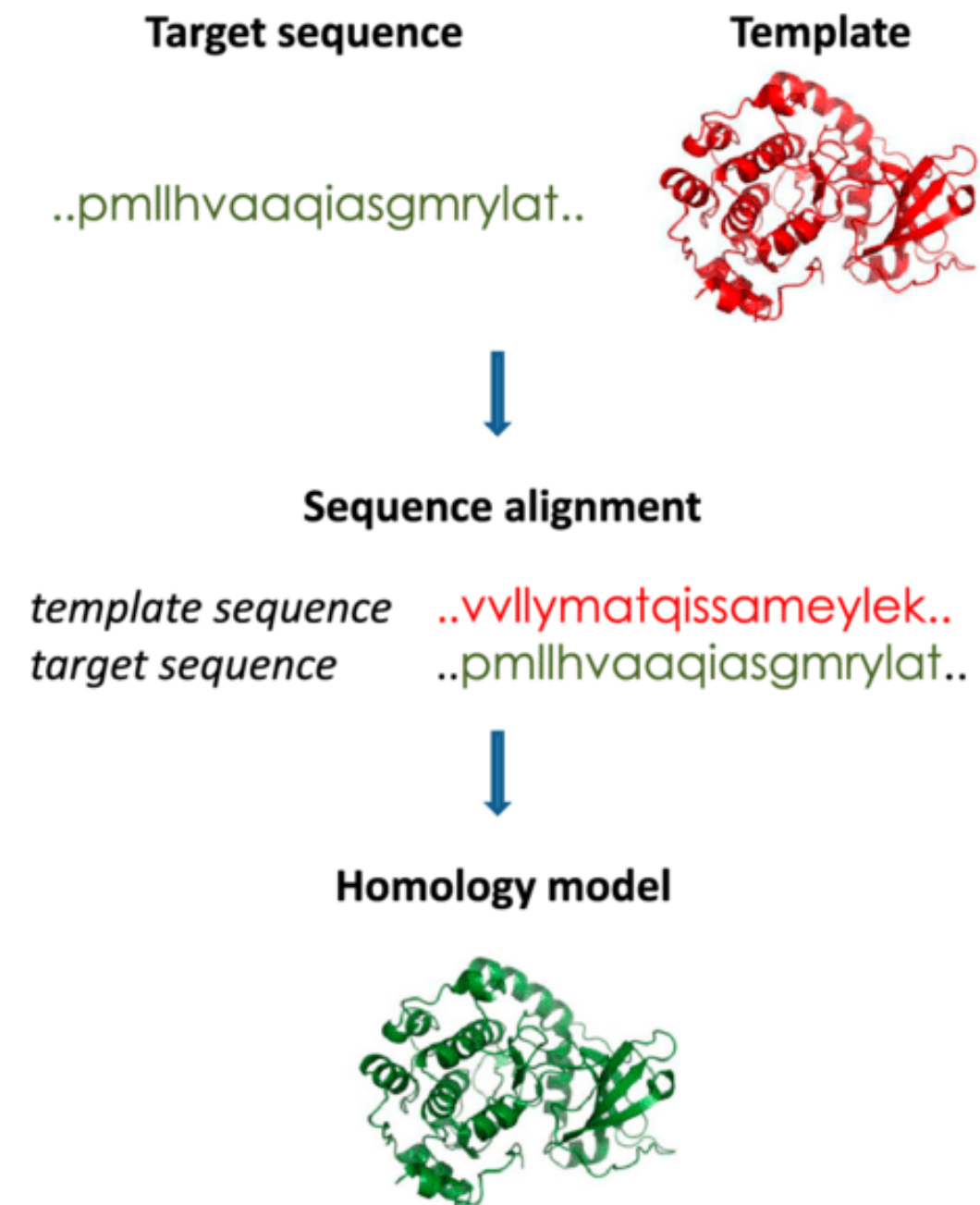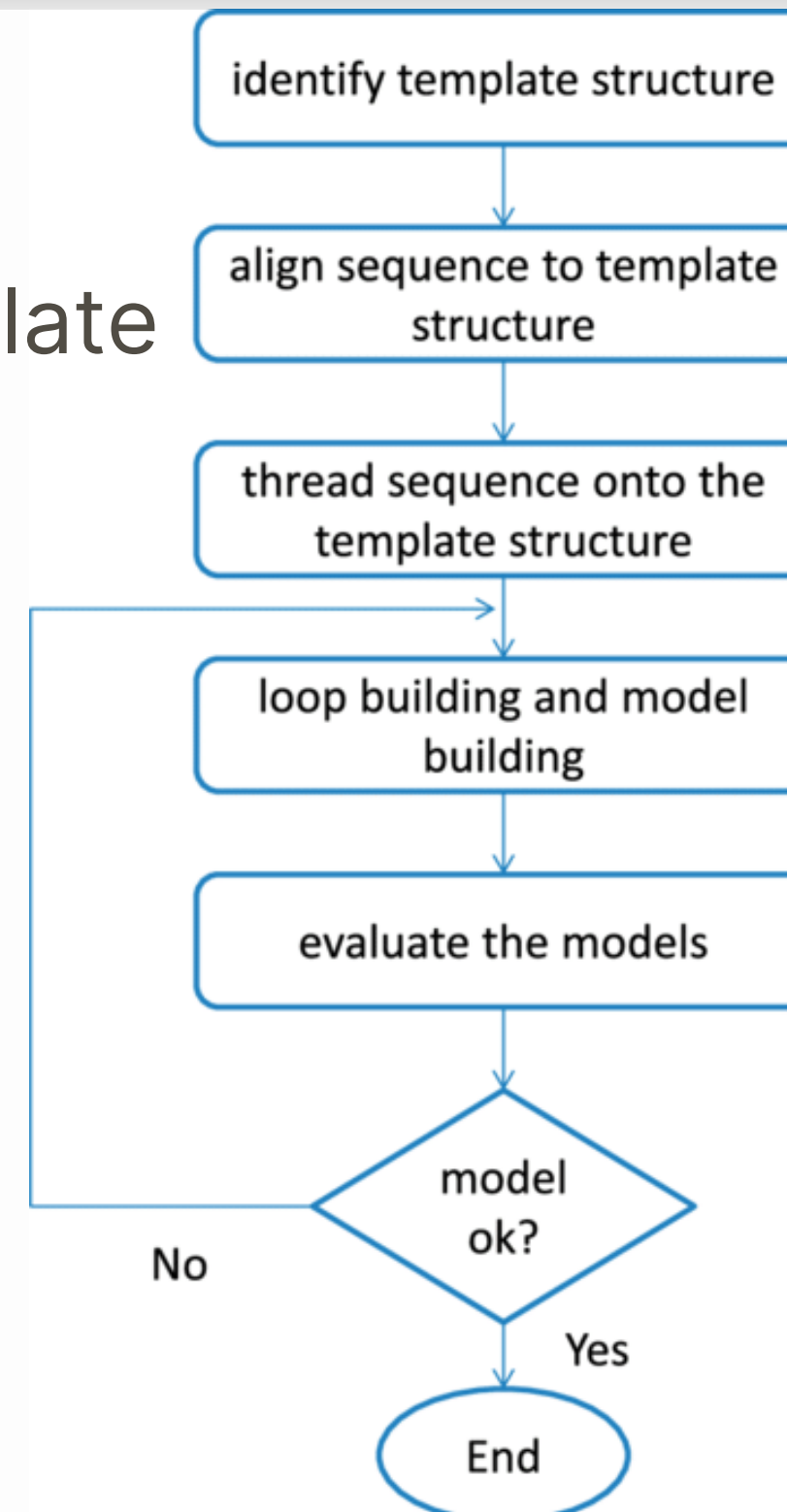an intuitive solution, but ….
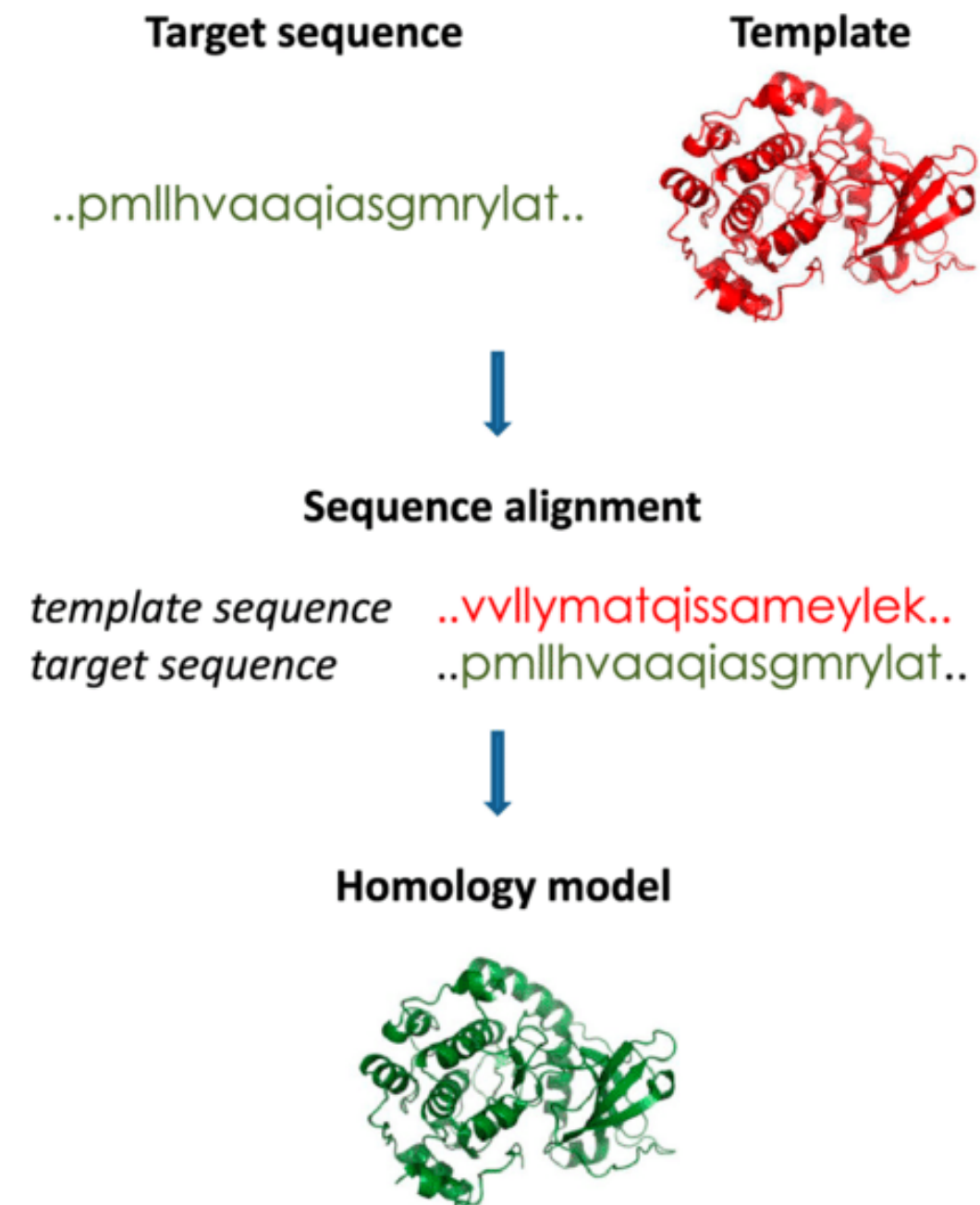
H O M O L O G Y   M O D E L L I N G
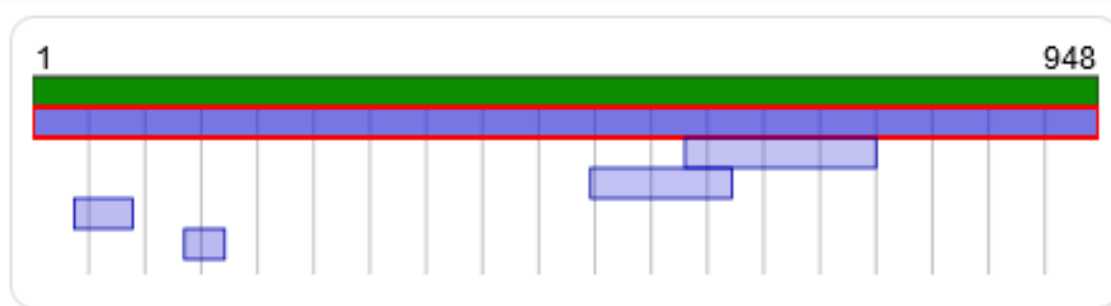
an intuitive solution, but ....

it's limited to known folds and template

availability!

Many protein families has no

representant in PDB!

SWISS MODEL

homology modeller using templates from

AlphaFoldDB and/or PDB

AlphaFold

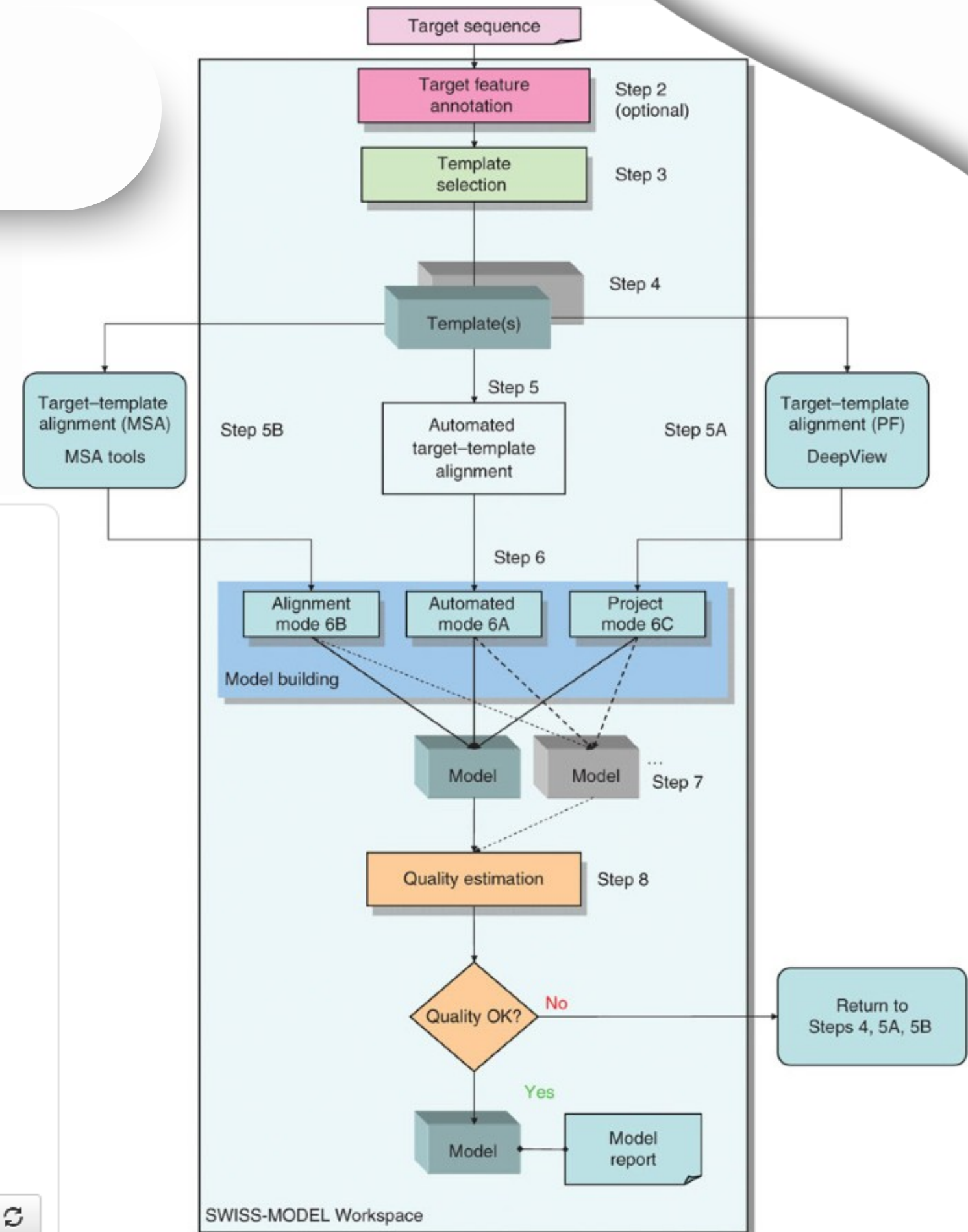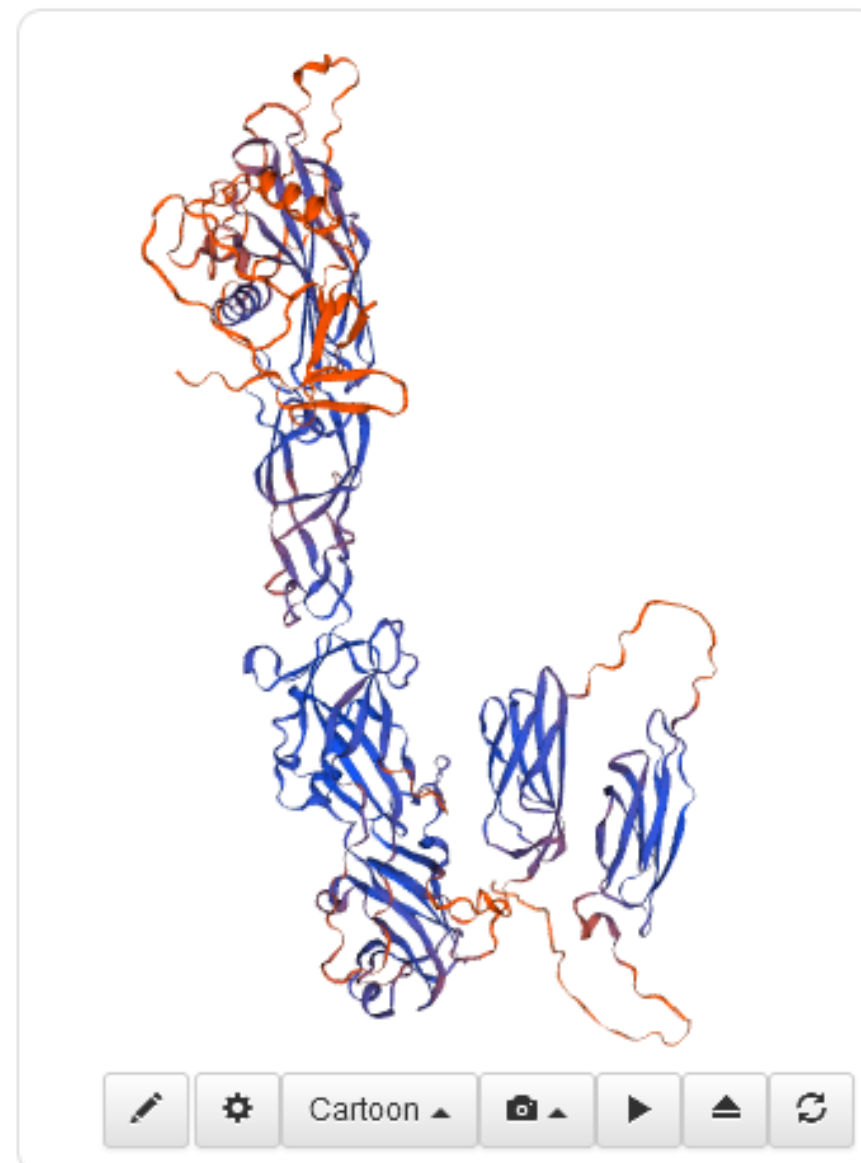A. user input   B. database search   C. prediction model   D. relaxation + output
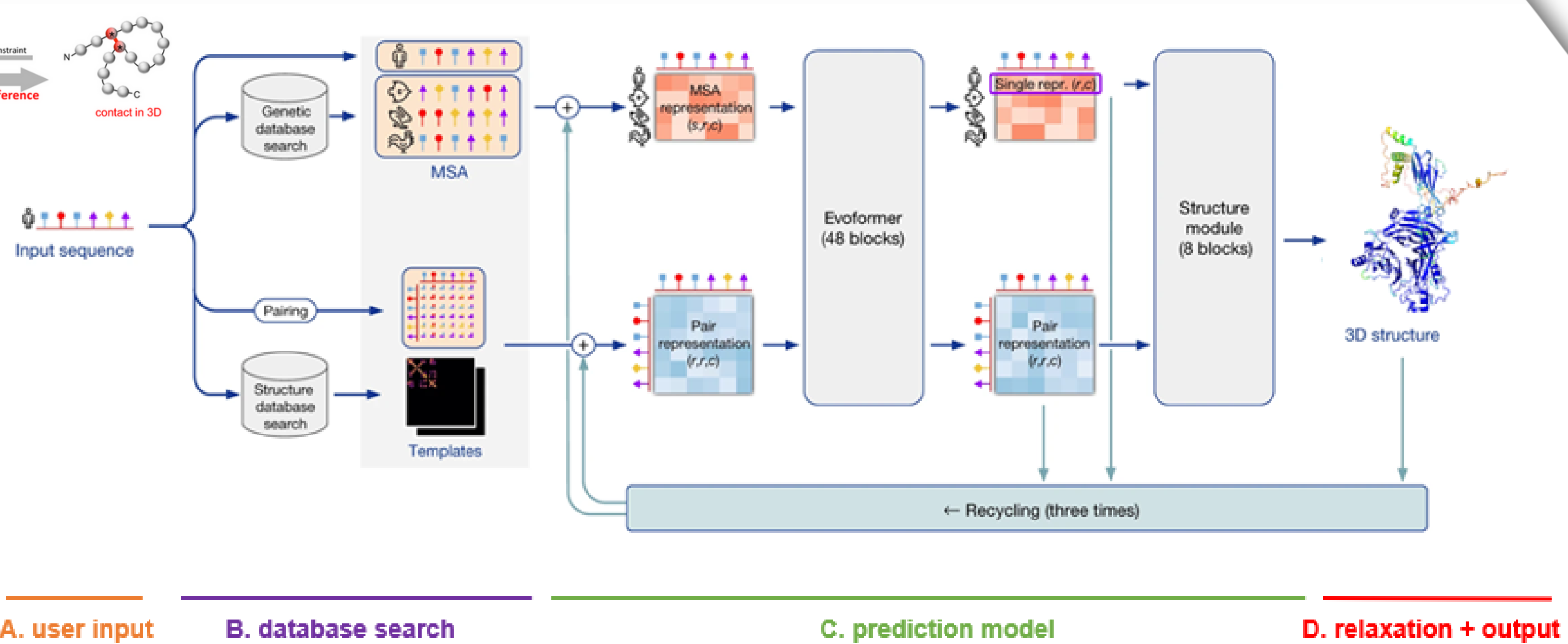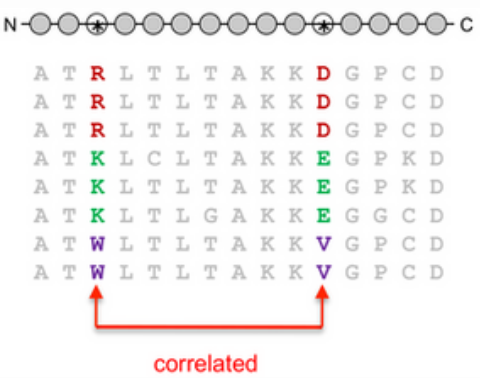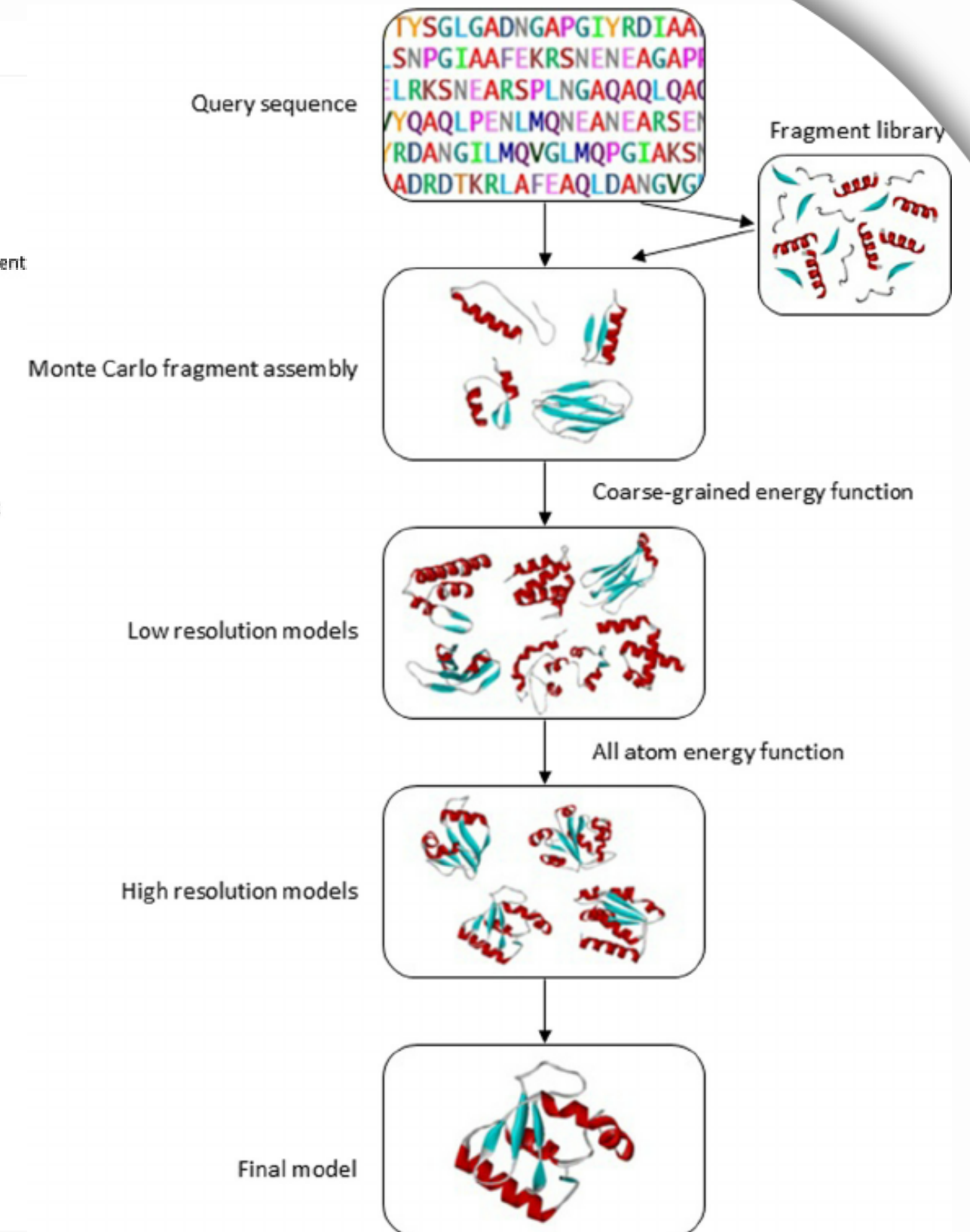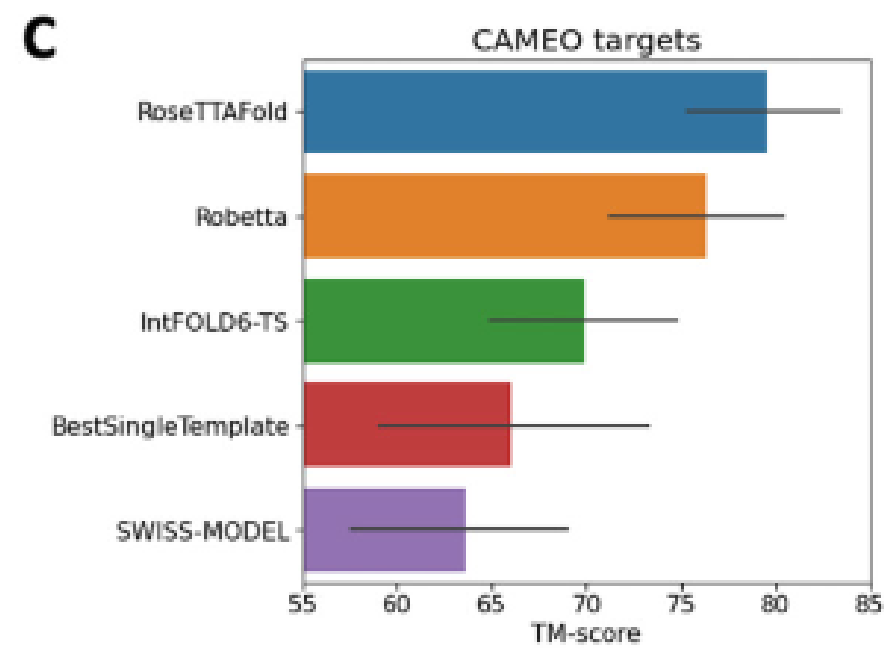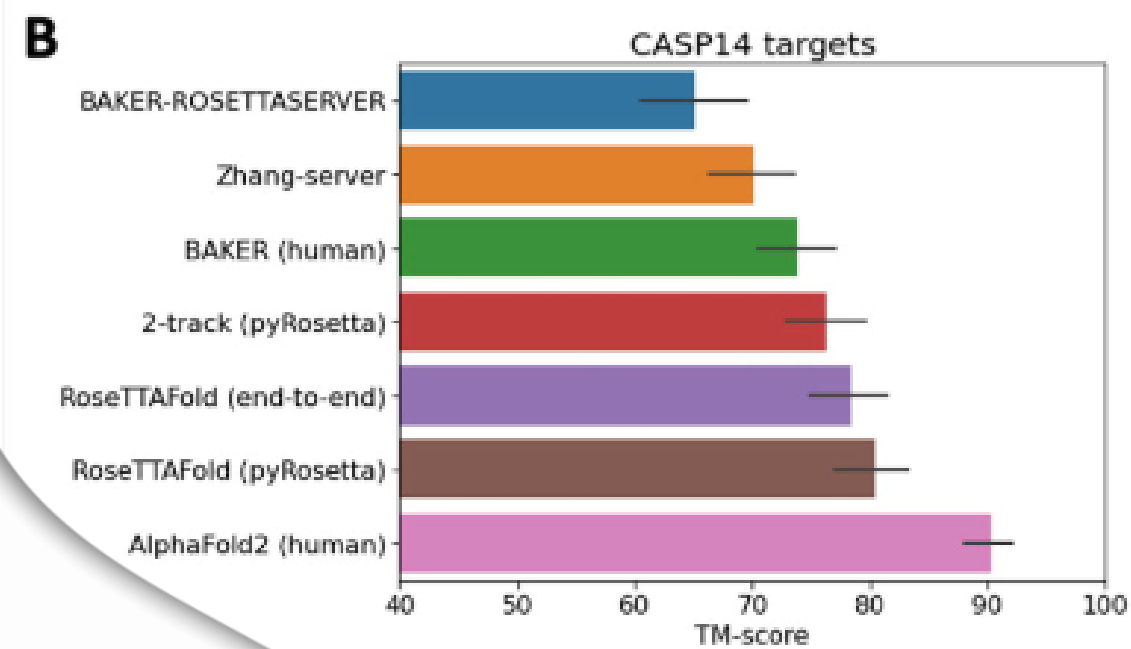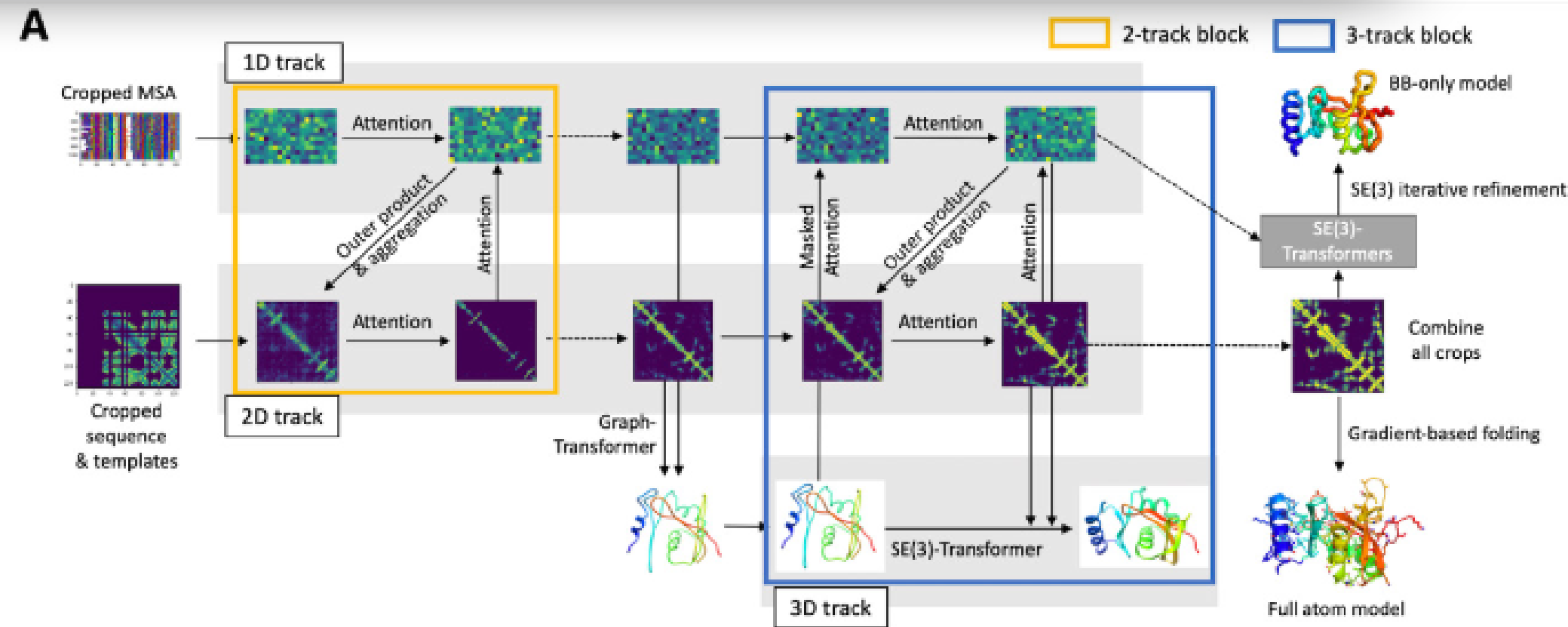
RoseTTAFold

ESMfold
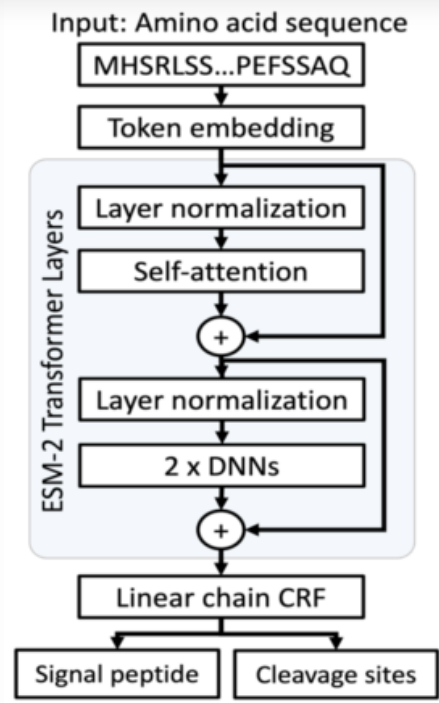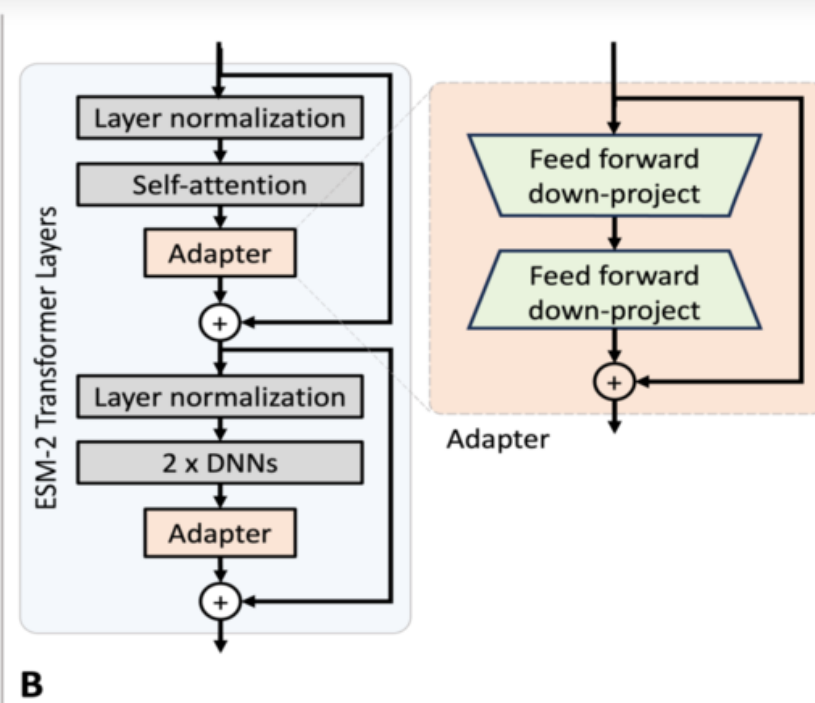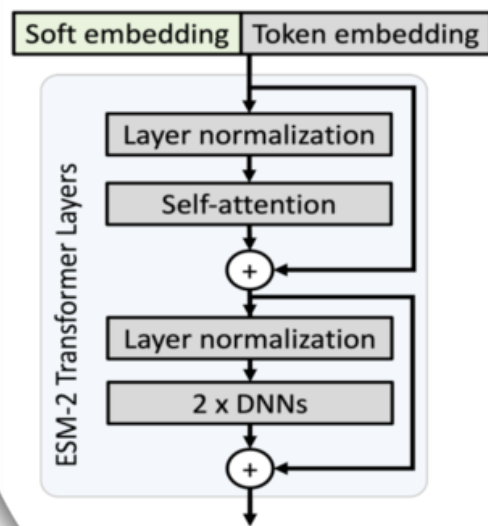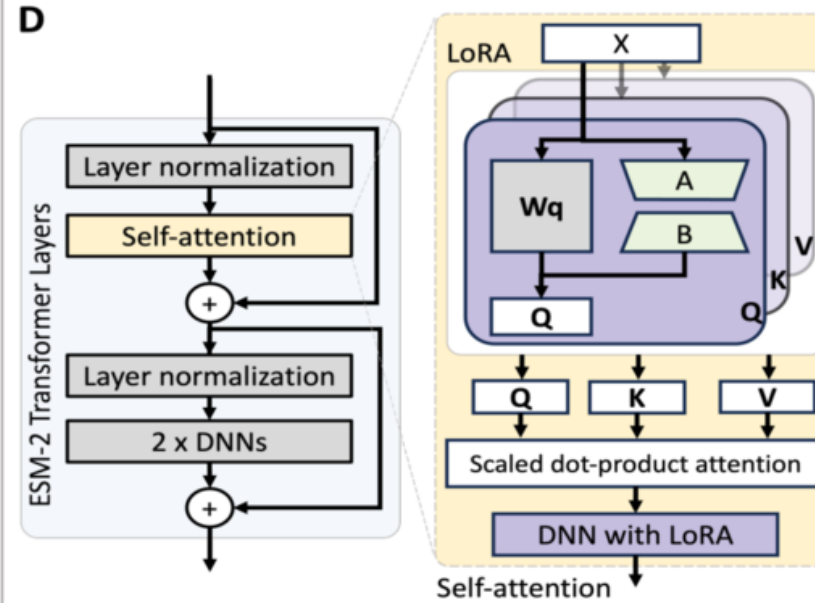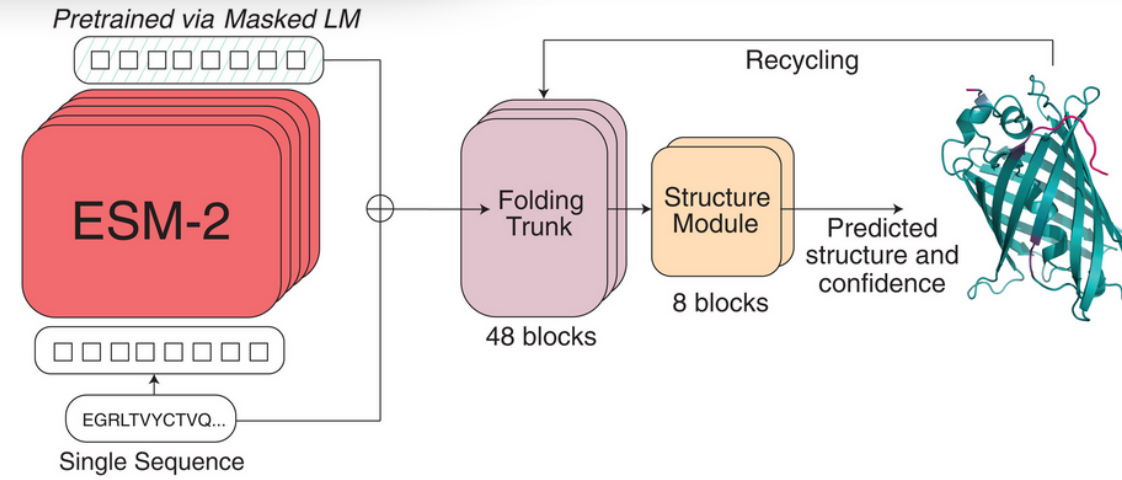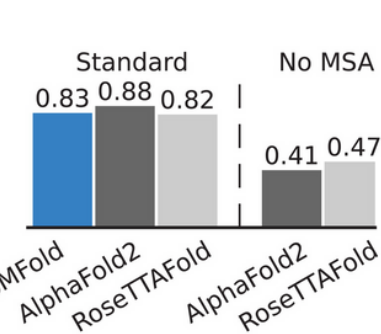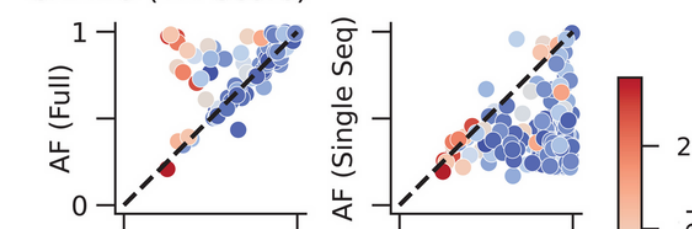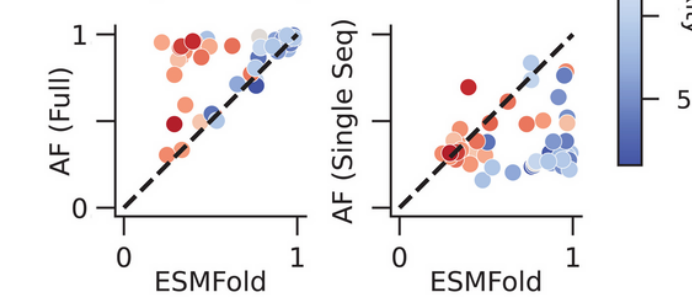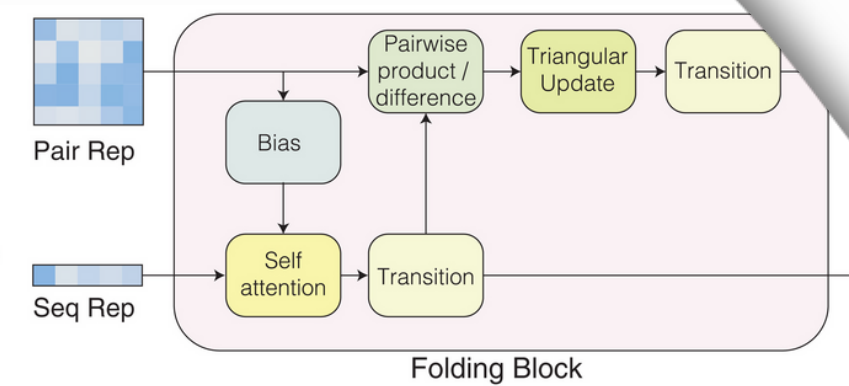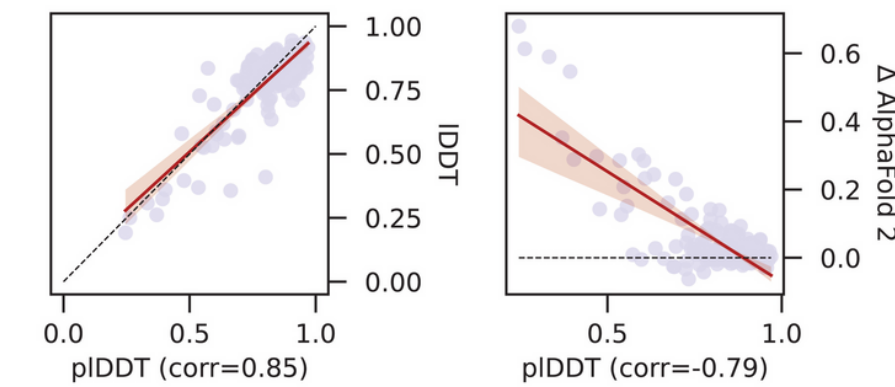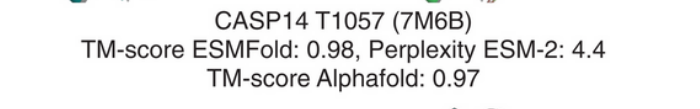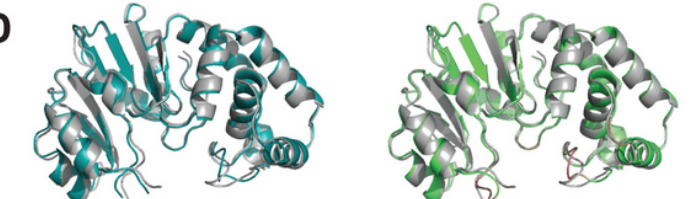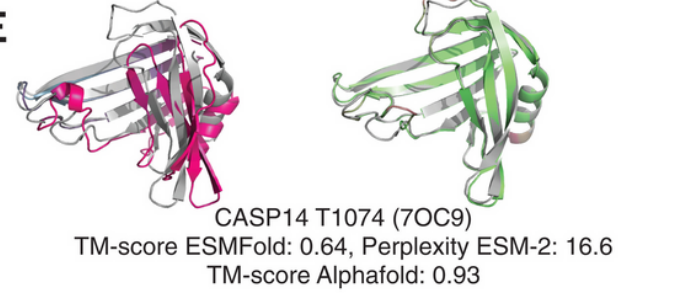
OTHER TOOLS

ColabFold

Offers the ability to run slightly simplified ESM, AlphaFold and RoseTTaFold models in Google Colab.

OpenFold

An AlphaFold based model using Torch and some adjustments in terms of computing complexity.

I-Tasser

hierarchical approach to protein structure prediction and structure-based function annotation from Zhang Lab. Possibility to define own template or to exclude some templates from PDB.

OpenComplex

OpenComplex support almost all features from

Alphafold 2 and OpenFold, and introduces some new

features:

- Reimplemented Alphafold-Multimer models.

- RNA and protein-RNA complex models with high

  precision.

- Kernel fusion and optimization on >=Ampere GPUs,

  brings 16%

Comparison

AVAILABILITY

Models	GitHub URL	Web servers to compute user sequences
ColabFold	https://github.com/sokrypton/ColabFold	Can be easily used on Colab
AlphaFold	https://github.com/google-deepmind/alphafold	-
ESM	https://github.com/facebookresearch/esm	https://esmatlas.com/resources?action=fold
RosettaFold	https://github.com/RosettaCommons/RoseTTAFold	https://robetta.bakerlab.org/ (requires account)
OpenComplex	https://github.com/baaihealth/OpenComplex	-
SwissModel	-	https://swissmodel.expasy.org/
I-Tasser	https://zhanggroup.org/I-TASSER/download/ (required account)	https://zhanggroup.org/I-TASSER/

Models	Database URL	Data
AlphaFold	https://alphafold.ebi.ac.uk/	UniprotKB
ESM	https://esmatlas.com/	MGnify
SwissModel	https://swissmodel.expasy.org/repository	UniprotKB
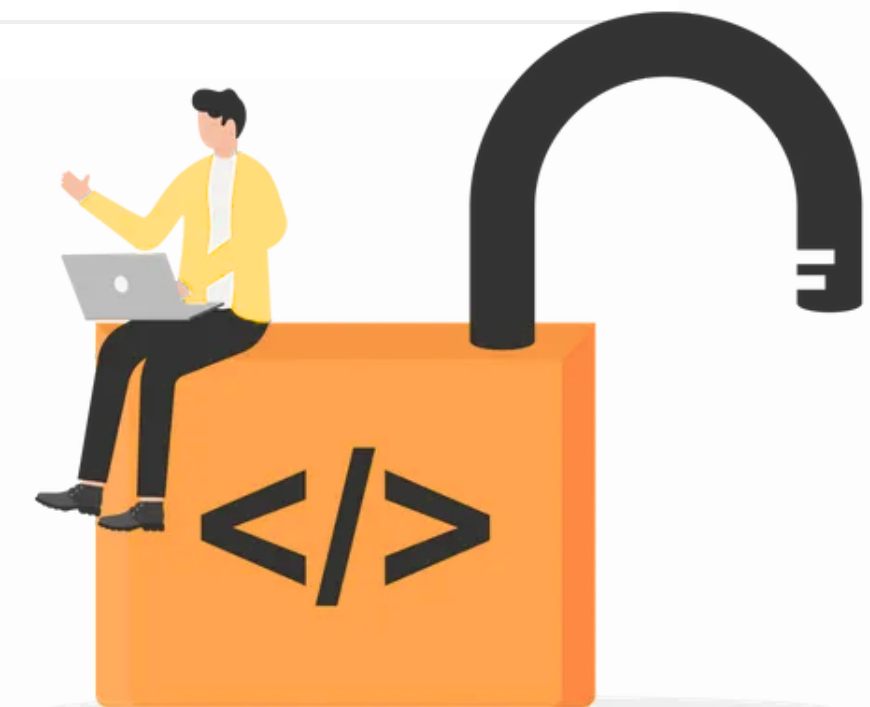
BEST QUALITY?

| Server Name			Targets				Average IDDT	
	Common Subset - Start Comparison	Avg. response time (hh:mm:ss) ^	#Submitted ^	#Modeled ^	#Submitted Oligo ^	#Modeled Oligo ^	All	Modeled ^
OpenComplex	○ ☐	22:42:03	187	184	81	0	82.3	83.7
SWISS-MODEL	○ ☐	00:06:50	187	184	81	38	77.1	78.4
IntFOLD7	○ ☐	29:53:27	187	169	81	0	72.0	79.6
Naive AlphaFoldDB 90	○ ☐	00:08:33	187	157	81	0	69.3	82.5
RoseTTAFold	○ ☐	04:46:54	187	175	81	0	66.6	71.1
IntFOLD6-TS	○ ☐	27:51:06	187	185	81	0	62.7	63.4
IntFOLD5-TS	○ ☐	34:25:53	187	185	81	0	62.1	62.8
BestSingleStructuralTemplate	○ ☐	04:48:44	187	163	81	0	58.9	67.6
Phyre2	○ ☐	00:32:17	187	182	81	0	50.5	51.9
Naive AlphaFoldDB 100	○ ☐	00:08:41	187	103	81	0	46.5	84.4
NaiveBLAST	○ ☐	00:00:31	187	152	81	0	42.8	52.6

CAMEO (Continuous Automated Model EvaluatiOn): models are evaluated against reference structures released by the PDB on a weekly basis.
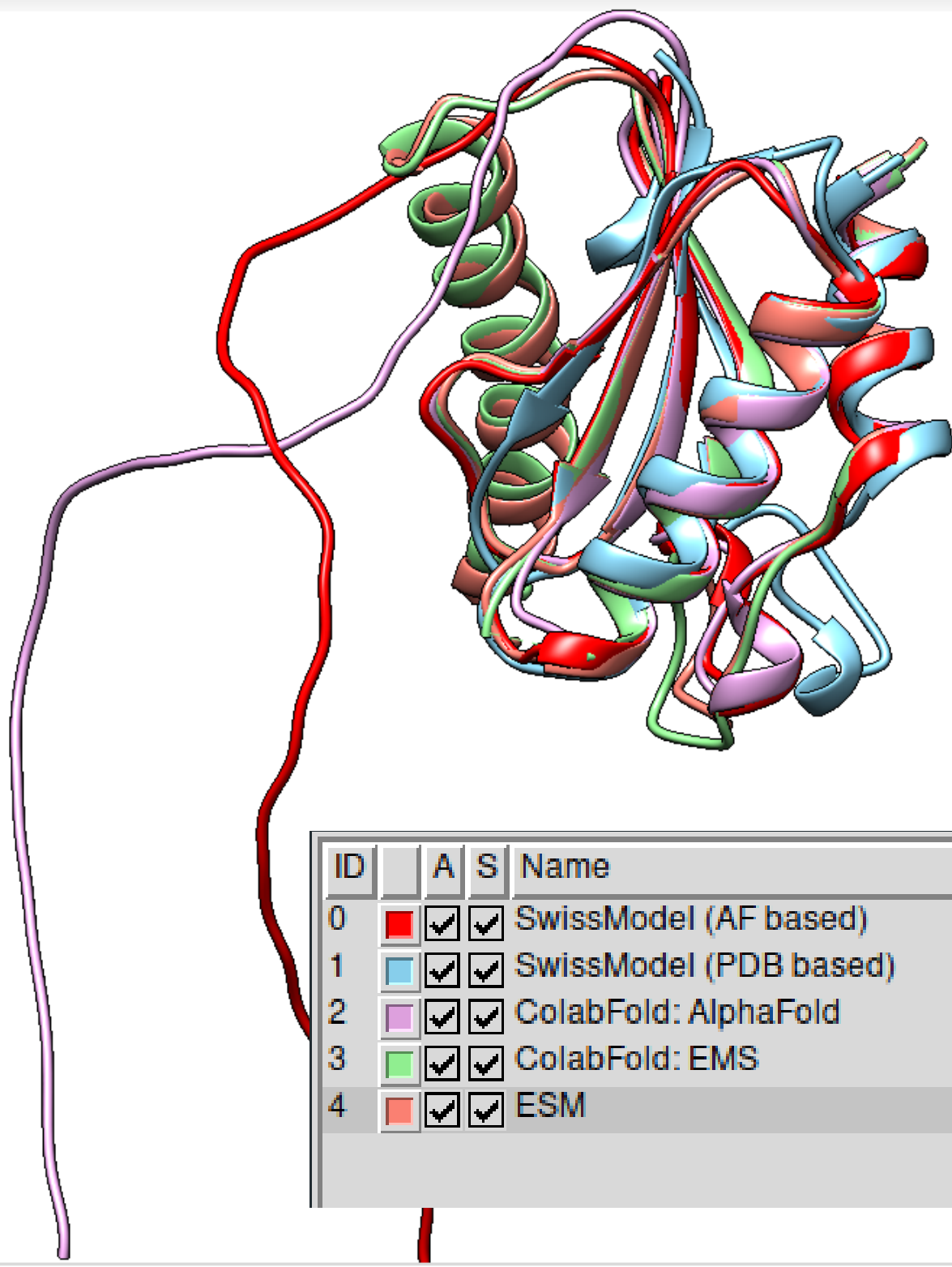

CAMEO-3D 3-months Performance (2024-01-05 - 2024-03-30)

TIME

Inference time of ESMFold / AlphaFold
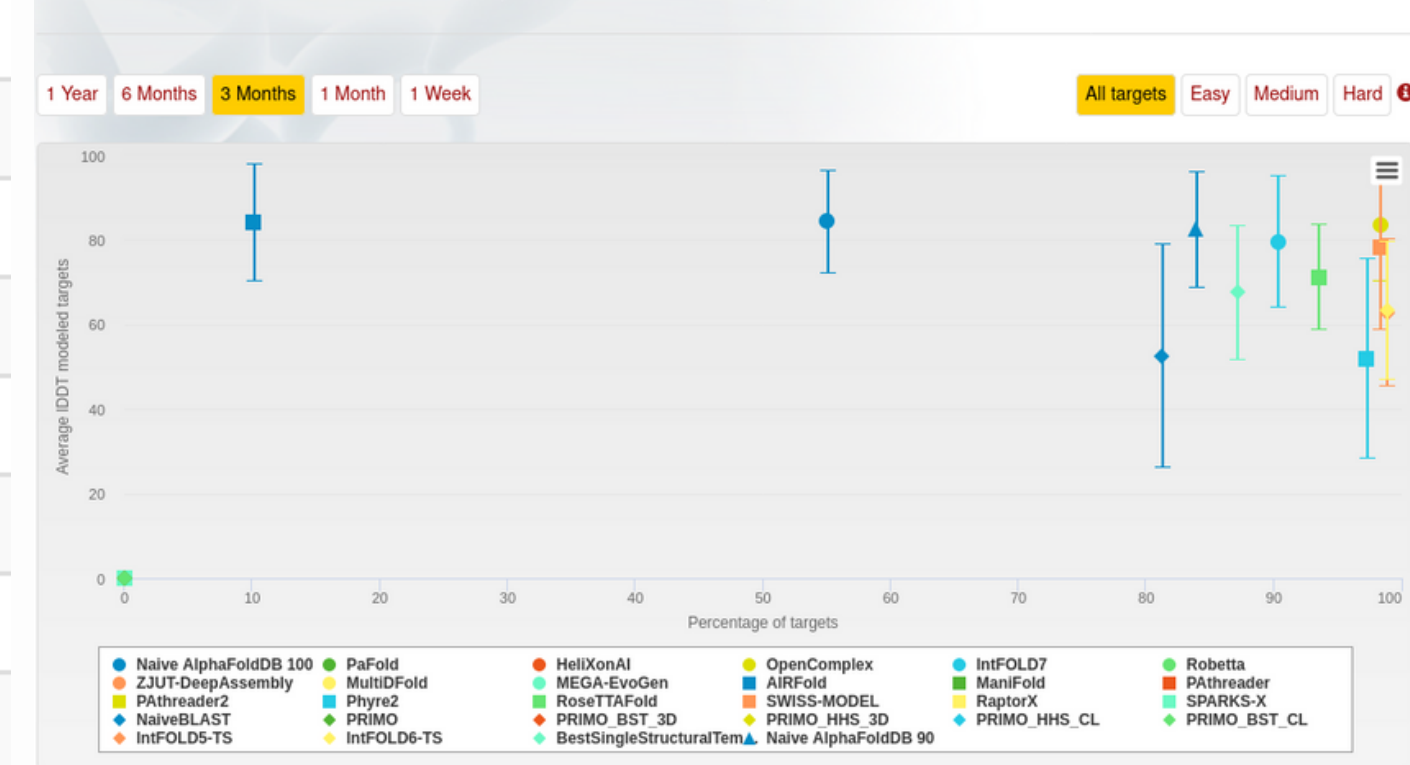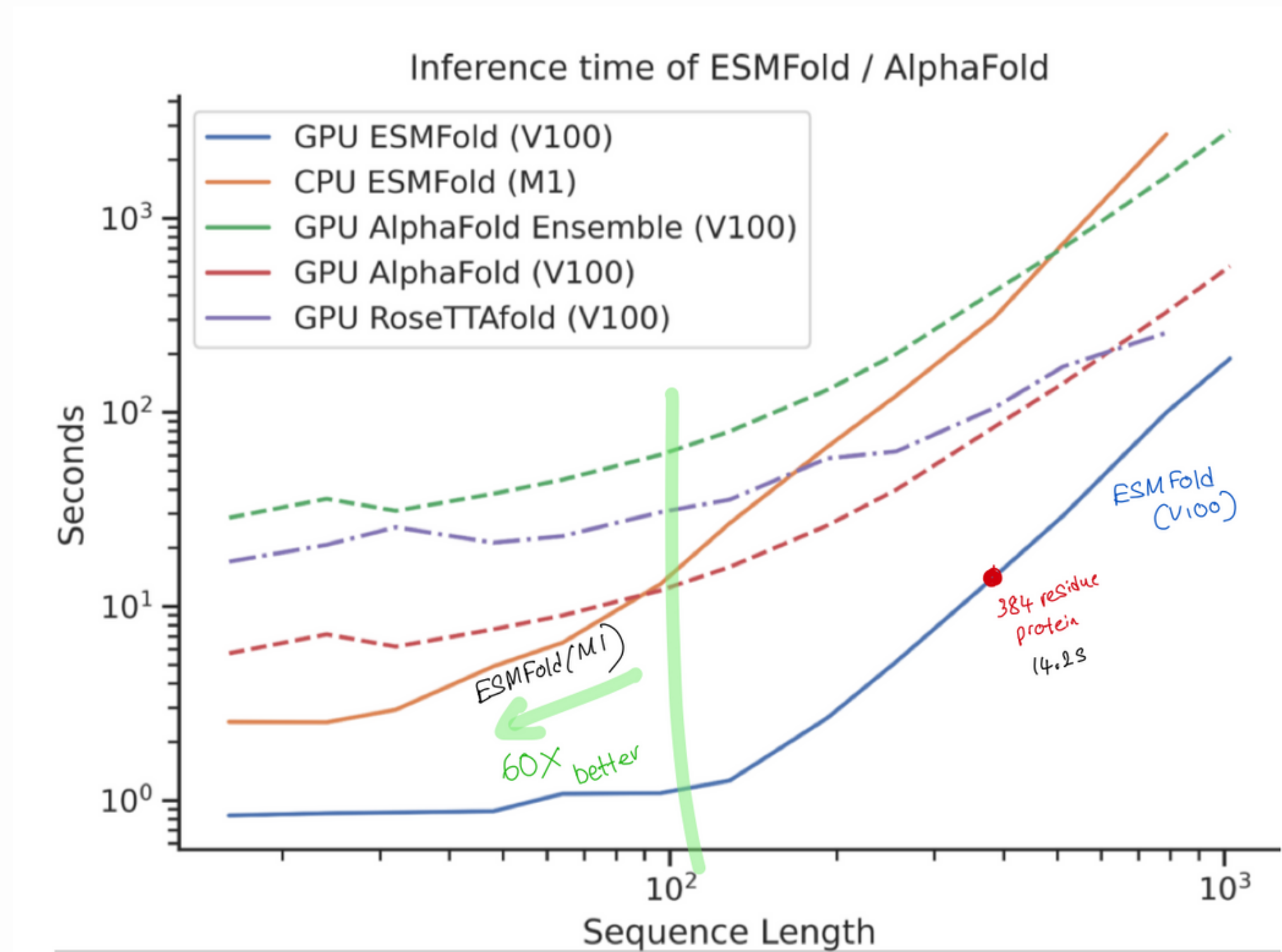
Q8IZQ5- MY TEST

Structure[i]

No structure information available for Q8IZQ5

3D structure databases

ModBase   Search... ⬏

SWISS-MODEL- | Submit a new modelling project... ⬏
Workspace

Protein[i]	Selenoprotein H
Gene[i]	SELENOH
Status[i]	UniProtKB reviewed (Swiss-Prot)
Organism[i]	Homo sapiens (Human)

Amino acids	122 (go to sequence)
Protein existence[i]	Evidence at protein level
Annotation score[i]	2/5

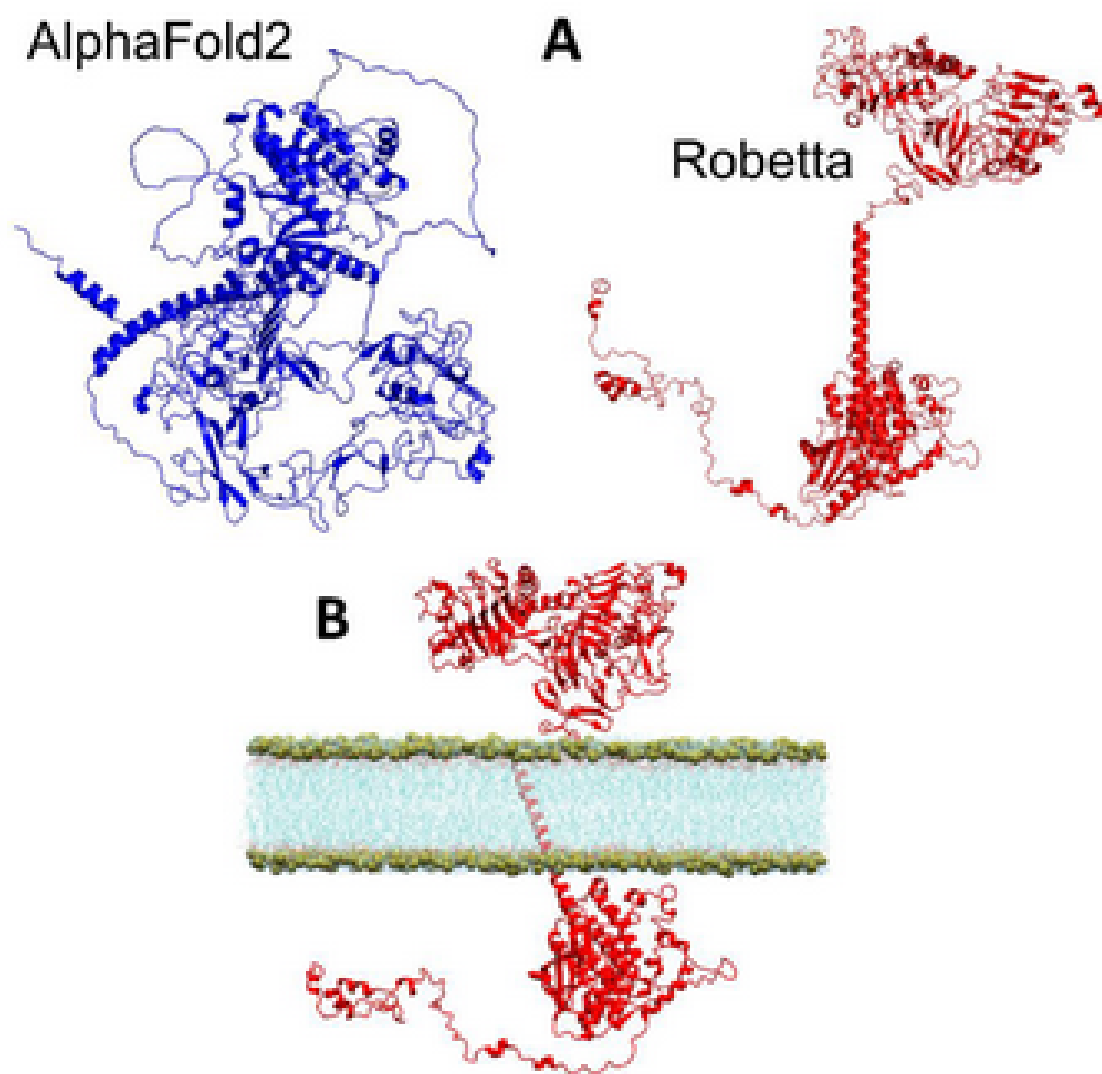| ID	A	S	Name
0	✓	✓	SwissModel (AF based)
1	✓	✓	SwissModel (PDB based)
2	✓	✓	ColabFold: AlphaFold
3	✓	✓	ColabFold: EMS
4	✓	✓	ESM

SwissProt: 3 minutes

ColabFold: about 5 minutes to download packages +

CF (AF): 3 minutes, CF (ESM): 1,5 minutes
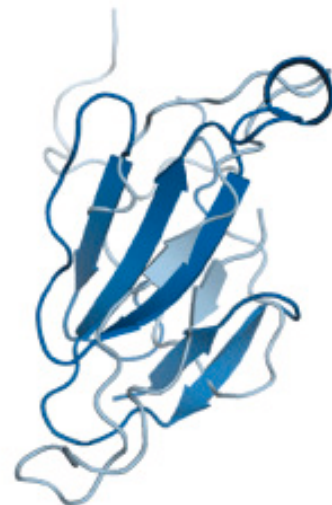
ESM through webserver: seconds

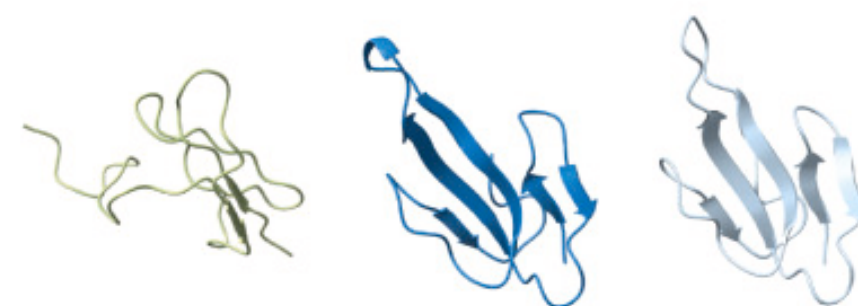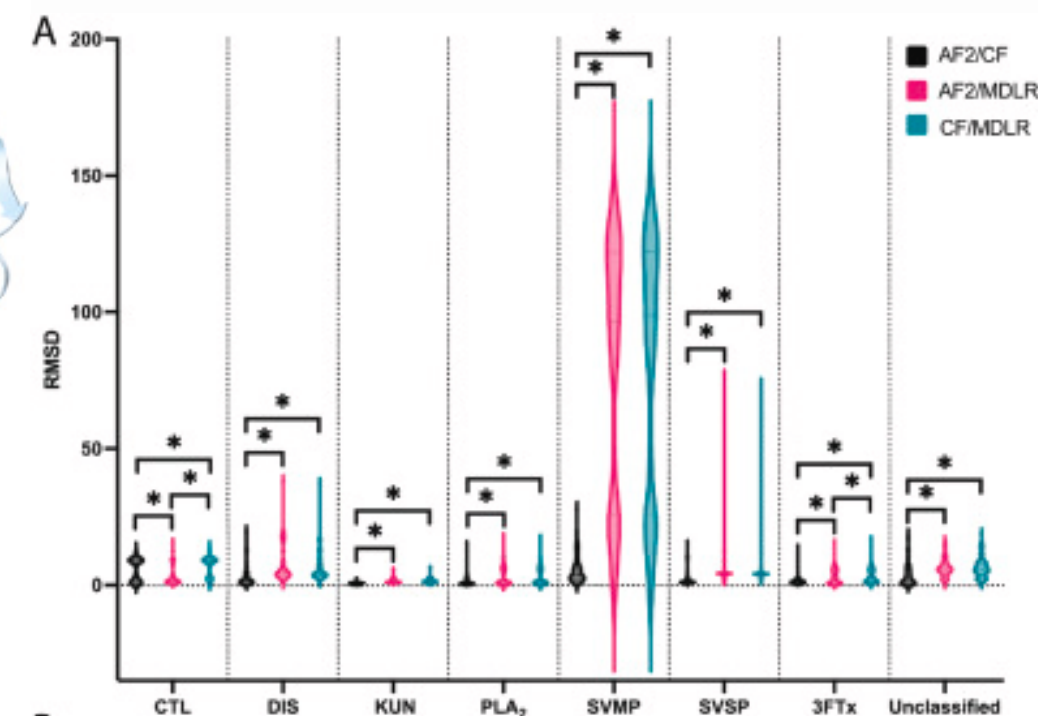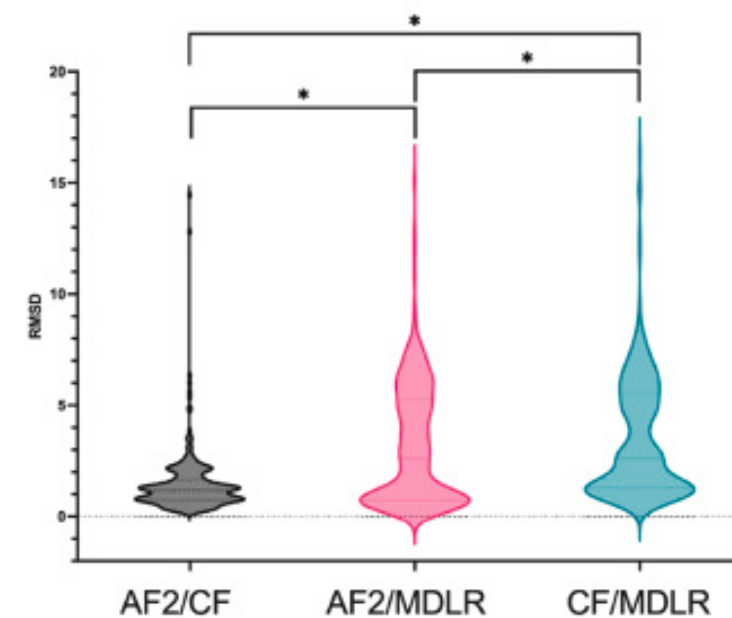| Knot type	Category	Uniprot	Organism	pLDDT knotcore
3₁	AF4	A0A6H5H865-F1	Nesidiocoris tenuis	81.5
5₂	ESM1	A0A6H5H865-F1	Nesidiocoris tenuis	79.2

HOW TO CHOOSE THE TOOL?

- check the availability of structures you need in databases, analyze their quality

- analyze the number of structures you want to predict, the number of amino acids in them (ESM is limited to 400, in AF Db is 1280) and the computer resources (some models need modern GPUs, Google Collab has limitations etc.)

- consider the time you have (some tools need minutes, some webservers need hours...)

- consider the nature of the proteins you want to analyse

REFERENCES

https://www.science.org/doi/full/10.1126/science.ade2574?casa_token=D9NpbpC-

js4AAAAA%3ABI_654xMkbVyo632UOnRpqjR2ovEW_0yh7f2NJDA4V4qT3PrPwq_7Eix1435vz7o3SNkCQqTKkmZDRA

https://ramith.fyi/esm-2-evolutionary-scale-prediction-of-atomic-level-protein-structure-with-a-language-model/

https://www.cameo.com/

https://www.uniprot.org/

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

https://www.sciencedirect.com/science/article/pii/S0041010123003707

https://predictioncenter.org/

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