Carnegie Mellon University

Protein Folding Prediction

Capstone Partner : Merck & Co., Inc

Joon Jung, Britney Wang, Fangzhou Yuan

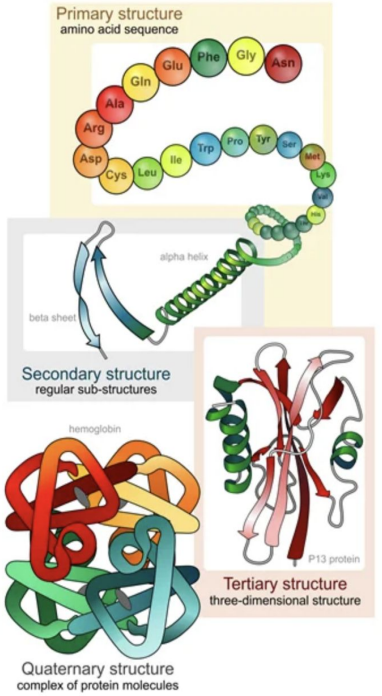
Agenda

I. Introduction
II. Model & Setup
III. Evaluation Metrics
IV. Results, Visualization
V. Future Directions

Introduction

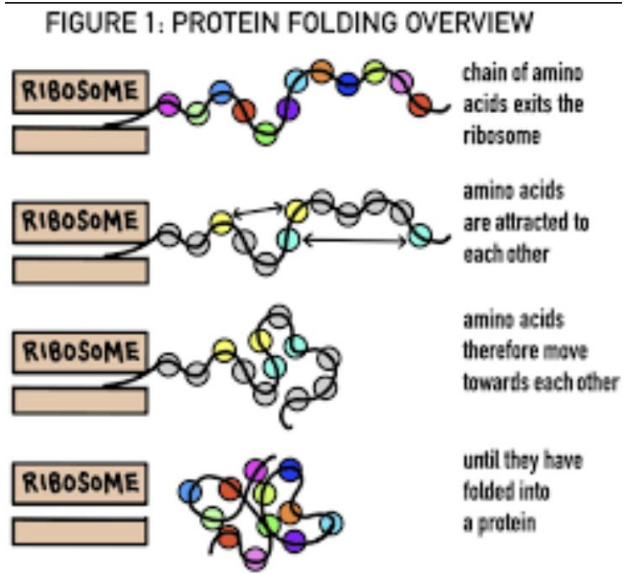
What is Protein Folding?

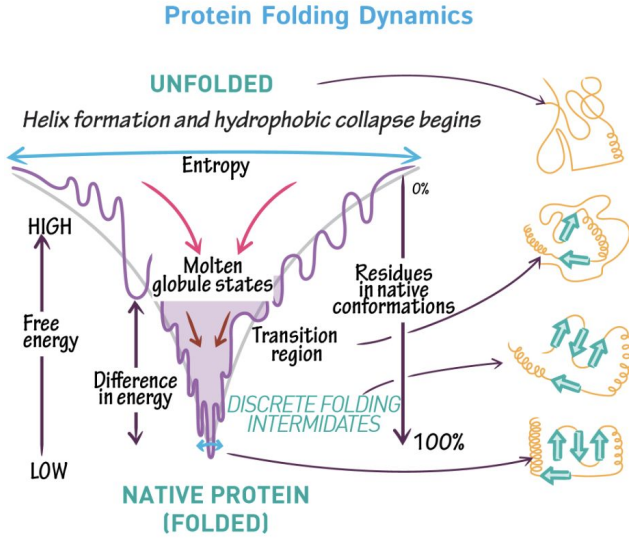
- **Process where linear sequence of amino acids folds into structure**
- **Crucial for its functionality:** Correct folding of protein is crucial for its functionality
  - Enzymes binding to substrates, signaling molecules to receptors
  - Neurodegenerative diseases : Alzheimer's, Parkinson's disease
- **Important to understand and be able to predict correctly**
  - Biological function, disease mechanism, drug design



https://www.news-medical.net/life-sciences/Protein-Folding.aspx

Protein Folding Prediction





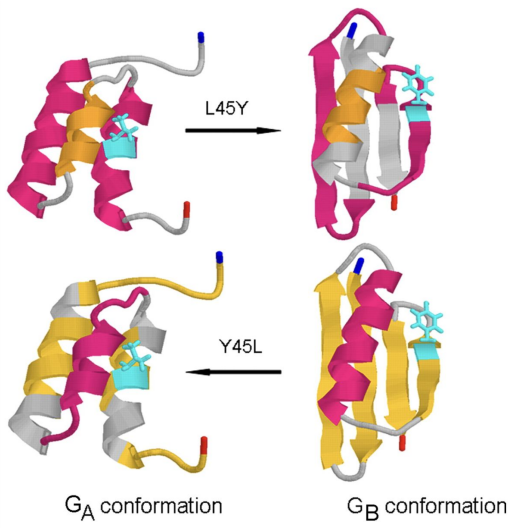
https://ditki.com/course/biochemistry/glossary/biochemical-pathway/protein-folding-dynamics

Protein Structure Prediction Tasks

- **Structure Prediction:** prediction of how the backbone and side chains fold in 3D space.
  - Primary, Secondary, **Tertiary**, Quaternary
- **Prediction of Protein Disorder Regions:** some regions of proteins (eg: RNA) are intrinsically disordered or highly flexible, which can be critical for their function.
- **Antibody-antigen Structure Prediction**
- **Effects of Mutations Prediction:** how changes in the amino acid sequence (due to mutations) can affect the overall structure and function of the protein.

Single, Double Mutation

- Changes in Amino acid sequence of protein
  - Single : One replaced by other
  - Double : Two different sites, simultaneously or independently
- Each mutation can have drastic impact on Folding stability and protein functionality
- Always a bad thing? No!
  - Intentionally to enhance antigenicity of protein
  - Helps us get better understanding of protein folding mechanism



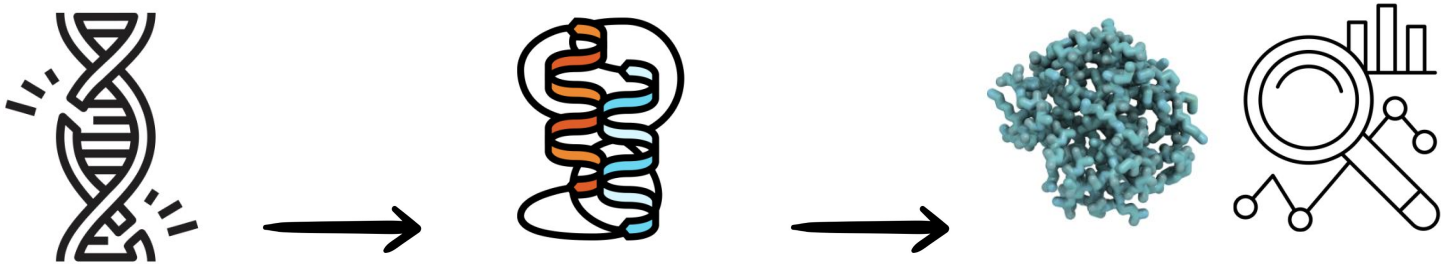
https://www.pnas.org/doi/10.1073/pnas.0912370107

Why is it so hard?

- **Complexity of Protein folding process:** Highly complex process that involves numerous interatomic interactions
    - Hydrogen bonds, hydrophobic interactions, van der Waals force, etc
- **Limited Experimental data:** Collecting data through experiments are time-consuming and expensive.
- **Conformational Flexibility:** Proteins adopt into multiple conformations under different environmental conditions or with other molecules
- **Computational Complexity:** complexity of prediction increases exponentially as the sequence gets long and size of protein increases (degrees of freedom, local minima)

Models & Setup

Process diagram



Step 1: Merck discusses with scientists and use code to generate where to perform mutations

Step 2: Protein Folding Structure prediction tools

Step 3: Visualization and Quantitative metrics

Model Comparison (Research)

- **AlphaFold2 (DeepMind):** network-based model, relies on Multiple Sequence Alignments (MSAs)
  - Winner of CASP13 (2018), CASP14 (2020)
  - High accuracy, got a GDT score above 90
- **RoseTTAFold (UW):** inspired by AF2
  - Accuracy scores are comparable to AF2's in CASP14
- **ESMFold (Meta AI)**: large-scale language based model
  - Requires only a single input sequence
  - Faster prediction speed, but lower accuracy
- **OmegaFold (Helixon)**: DL-based method that uses only single primary sequence and no MSAs
  - Incorporating Physics based approach
  - better suited for proteins that have low sequence coverage

Folding Models Setup

OmegaFold (OmegaFold v1.1.0)

- A single .fasta file (containing all sequences) -> A list of .pdb files (containing 3D structure information)

ESMFold (ESM-2 v1.0.3)

- Compute embeddings in bulk from FASTA, faster than OmegaFold

AlphaFold & RosettaFold

- Using ColabFold versions

Evaluation Metrics

Model Evaluation Metrics

Structural Accuracy

- **GDT_TS** and **GDT_HA**: overall shape and fold of a protein (Used in CASP)
- **TM-score:** overall topological similarity rather than specific atomic positions
- **plDDT:** confidence in the accuracy of local structural features

Backbone Conformation / Structural Similarity

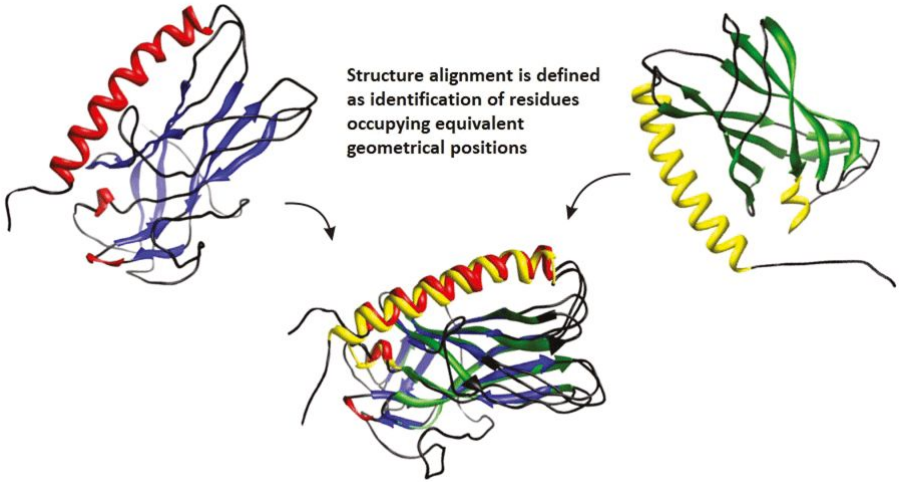
- **RMSD:** average distance between the atoms (usually the backbone atoms)

Mutation Effects

- **Effective Strain(ES):** measure local deformation
- **Log Odds Ratio:** how often the mutant (altered) amino acid appears at a specific position in a set of related proteins

Evaluation on Structural Similarity - RMSD

Predictions -> Structure Alignment (Superposition) -> Quantify the Overlap



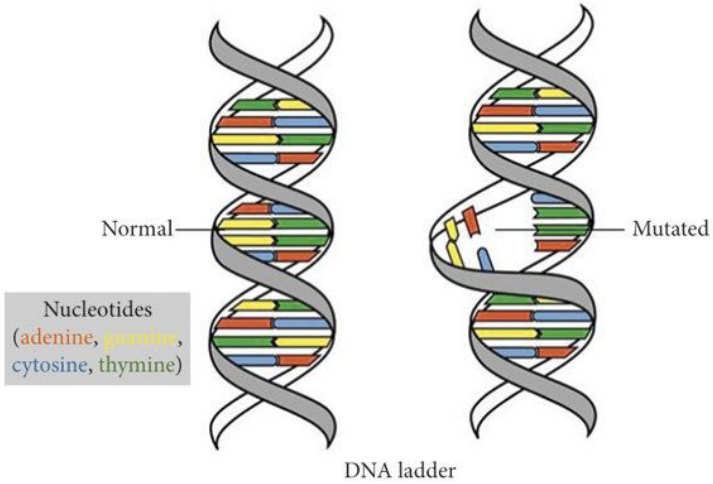
RMSD provides a quantitative value that reflects how closely two structures resemble each other post-superposition.

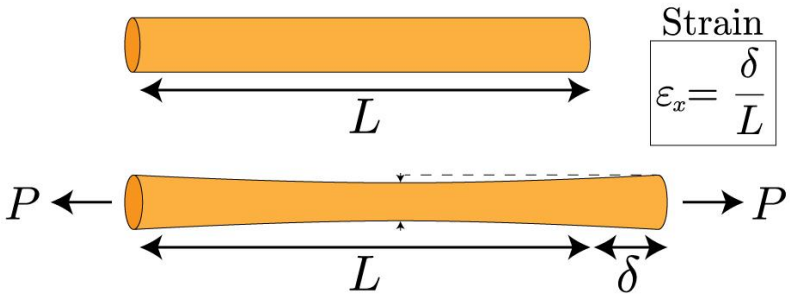
A lower RMSD value indicates a closer match and higher structural similarity

Evaluation on Mutations Effects - Local Deformation

Mutation -> Deformation -> Strain

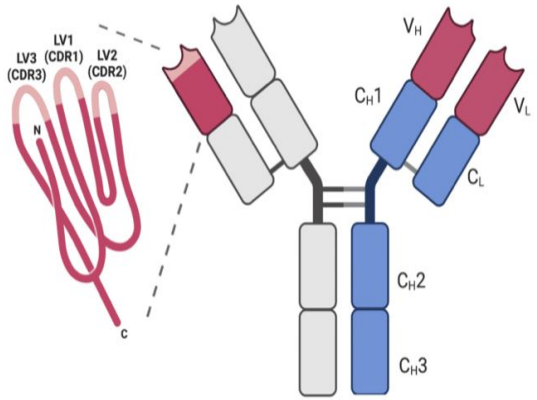
Metrics: Effective Strain - A measure of structural change in a protein or antibody due to mutations. It quantifies how much the structure of the mutant differs from the original structure.

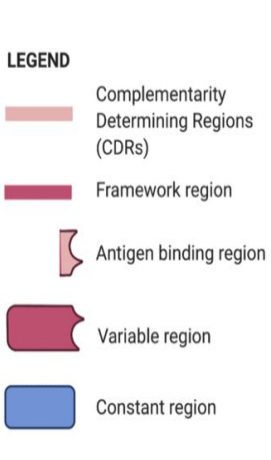


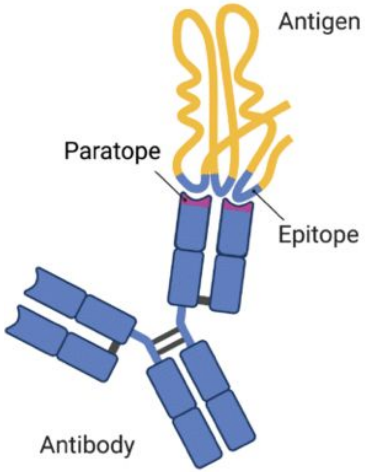


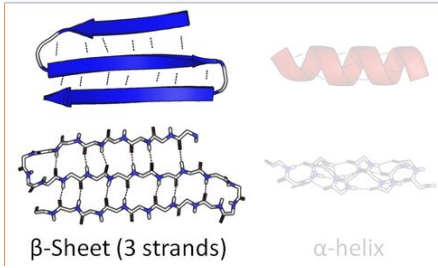
Results, Visualization

CDRs, String, α-Helix, β-Strands





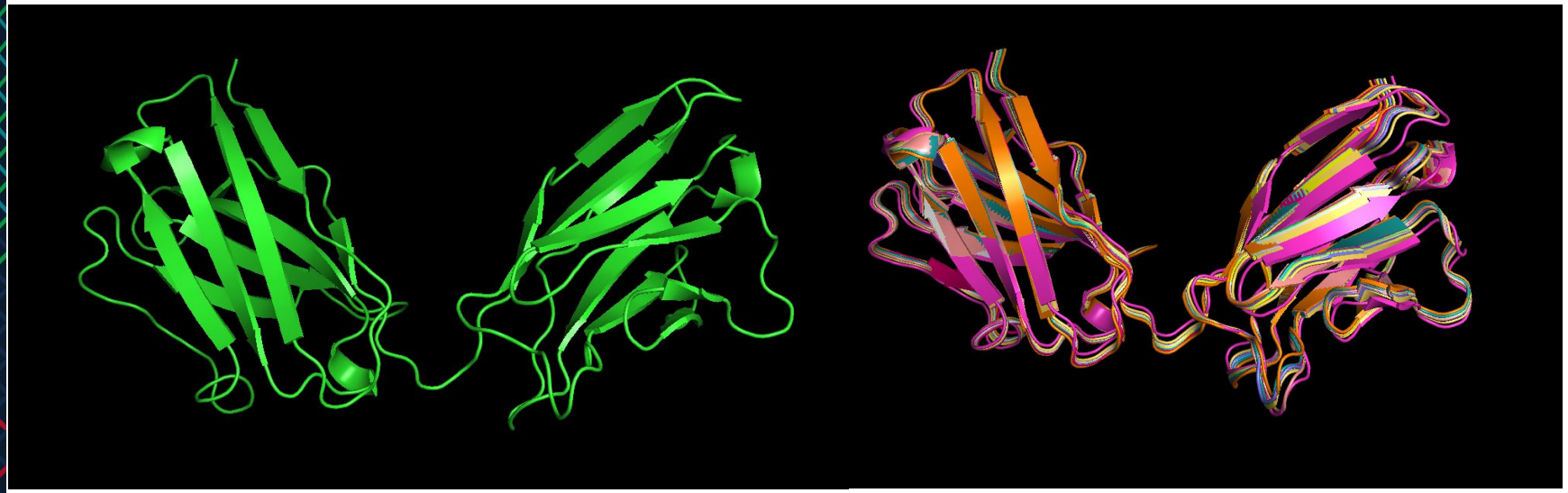




The CDRs include: {'GFTFSNYGMS': (26, 35), 'TISYGGSYTYYPDNIKG': (50, 66), 'GYGYDTMDY': (99, 107), 'KASQSVSFAGTGLMH': (142, 156), 'RASNLEA': (172, 178), 'QQSREYPWT': (211, 219)}.

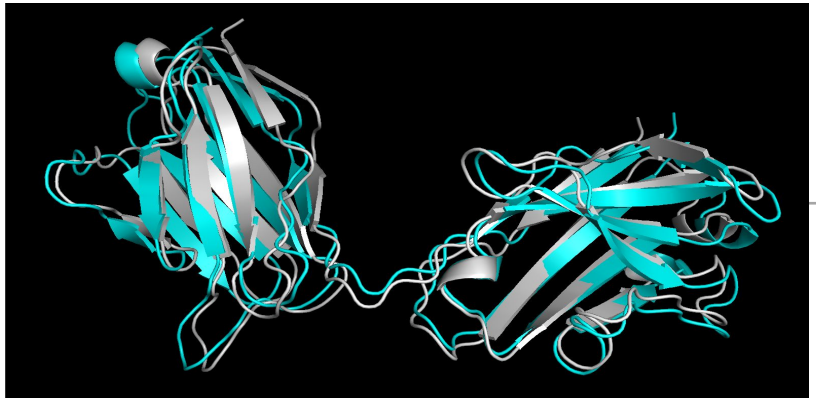
The CDRs include: {'SSFIH': (31, 35), 'RIDPAFGATEYNPAFQG': (50, 66), 'YHYAASHFDA': (99, 108), 'KSSQSVTNDLT': (143, 153), 'YASQRYT': (169, 175), 'QQDYASPFT': (208, 216)}.

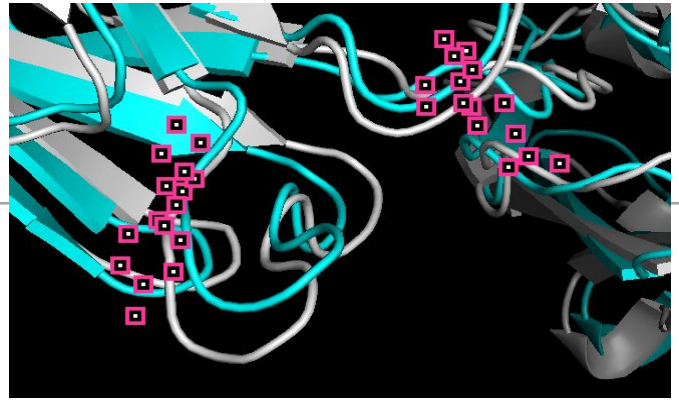
Pymol - Visualization 1

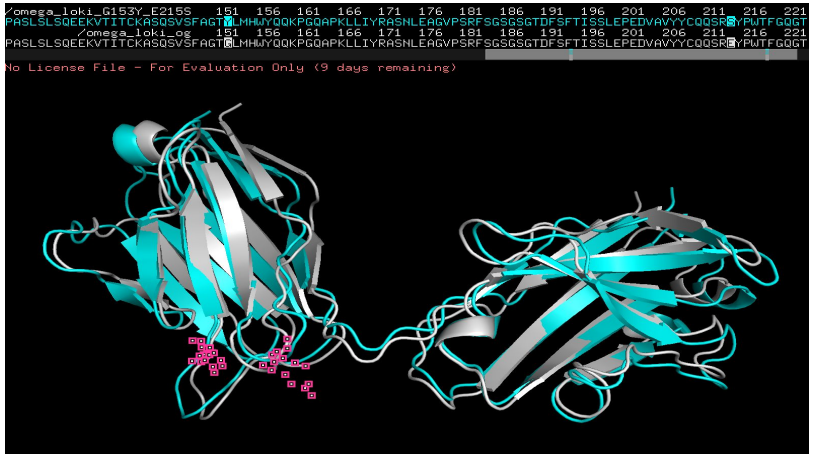


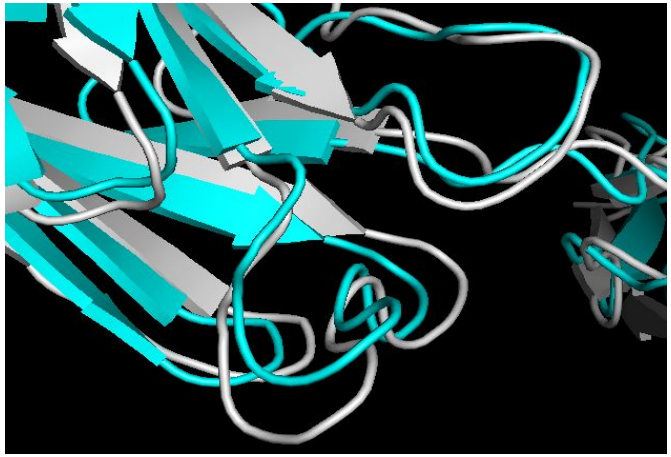
2 Sequences, 11-12 Single Mutation, 5-6 Double Mutation, 4 Tools, 144 Models

Pymol - Visualization 2

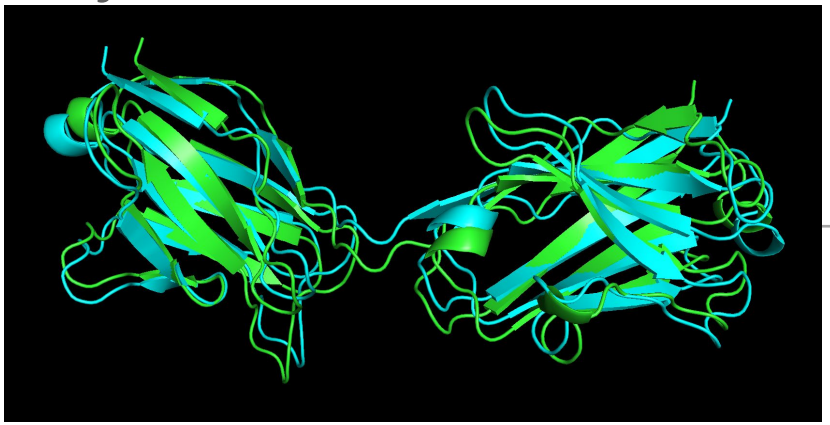






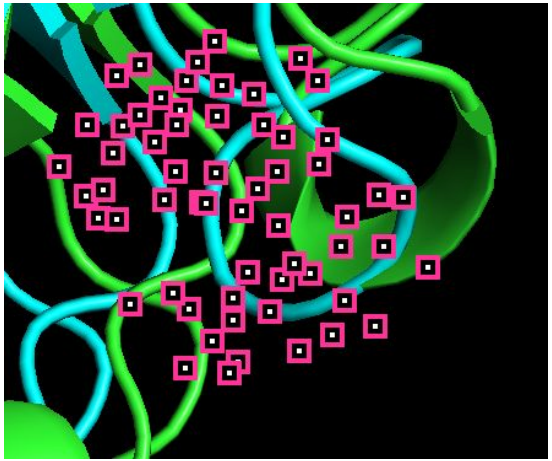


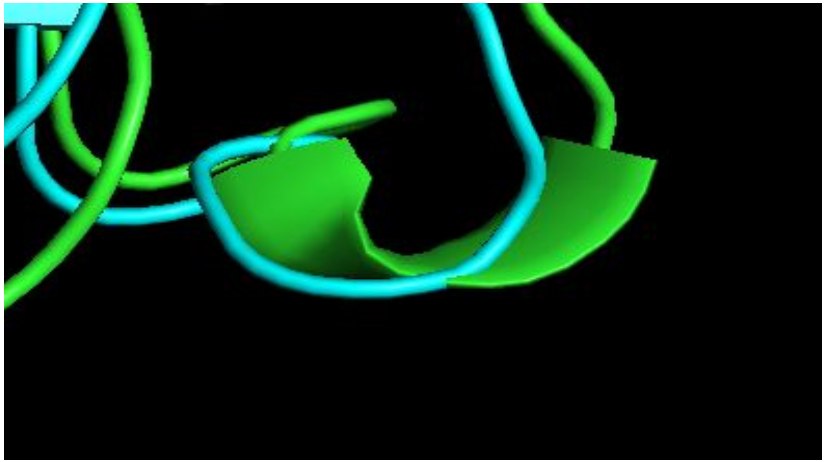
Pymol - Visualization 3



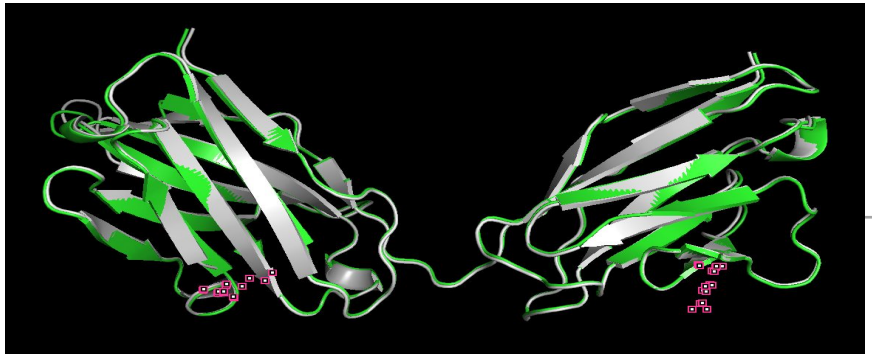
- rf2_loki_og
- rf2_loki_N63S_I64V

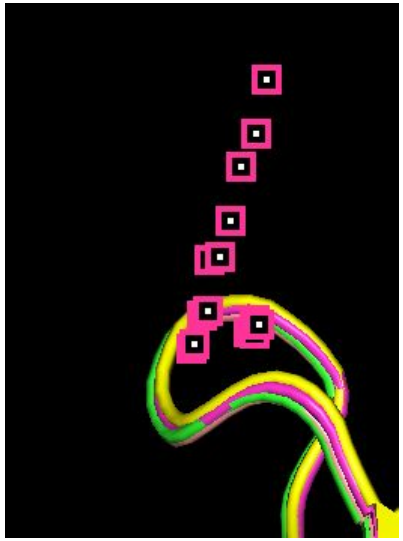
```
 56    61    66    7
GSYTYYPDNIKGRFTIS
 56    61    66    7
GSYTYYPDSVKGRFTIS
```

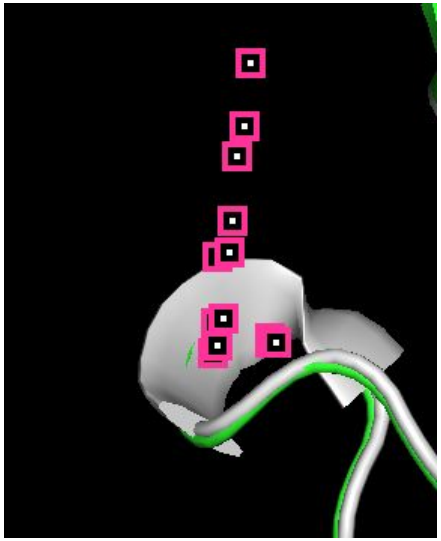




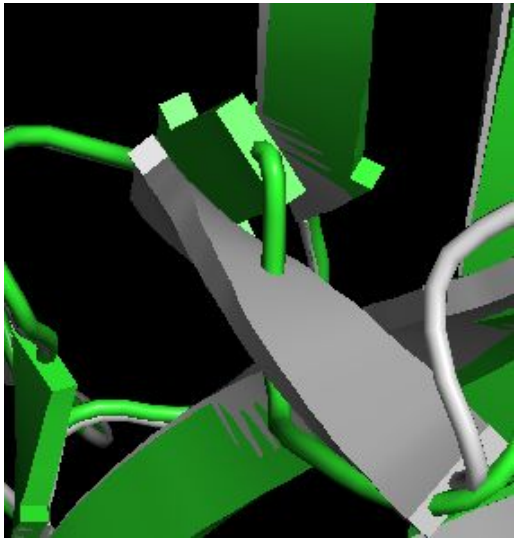
Pymol - Visualization 4



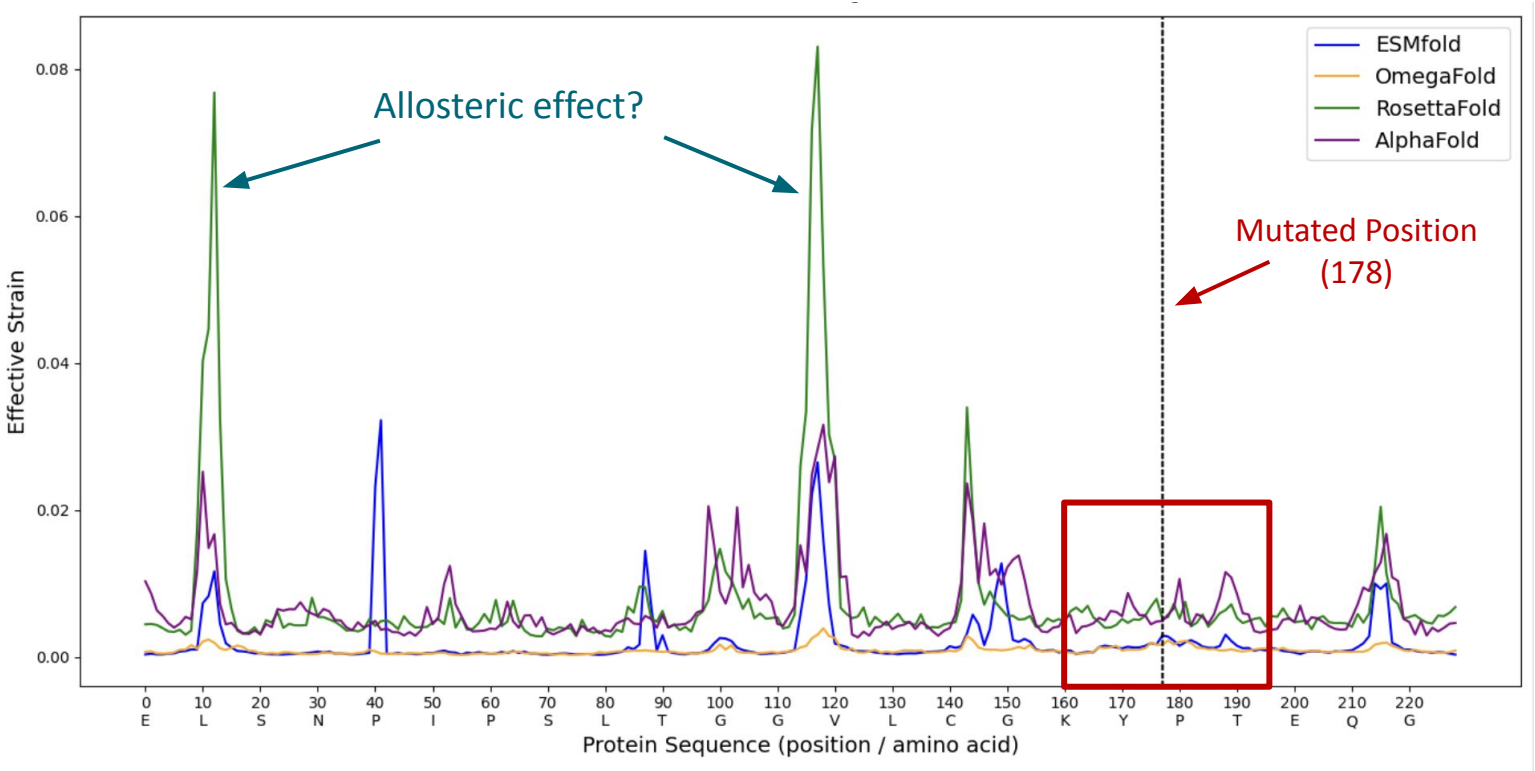




- Alpha_Nola_original_21b65_unrela
- Alpha_Nola_double_A108Y_Y174A_7
- Alpha_Nola_double_D151Y_E59N_5f
- Alpha_Nola_double_F55G_Q172T_78
- Alpha_Nola_double_F55S_S144A_38
- Alpha_Nola_double_Q172N_A63K_

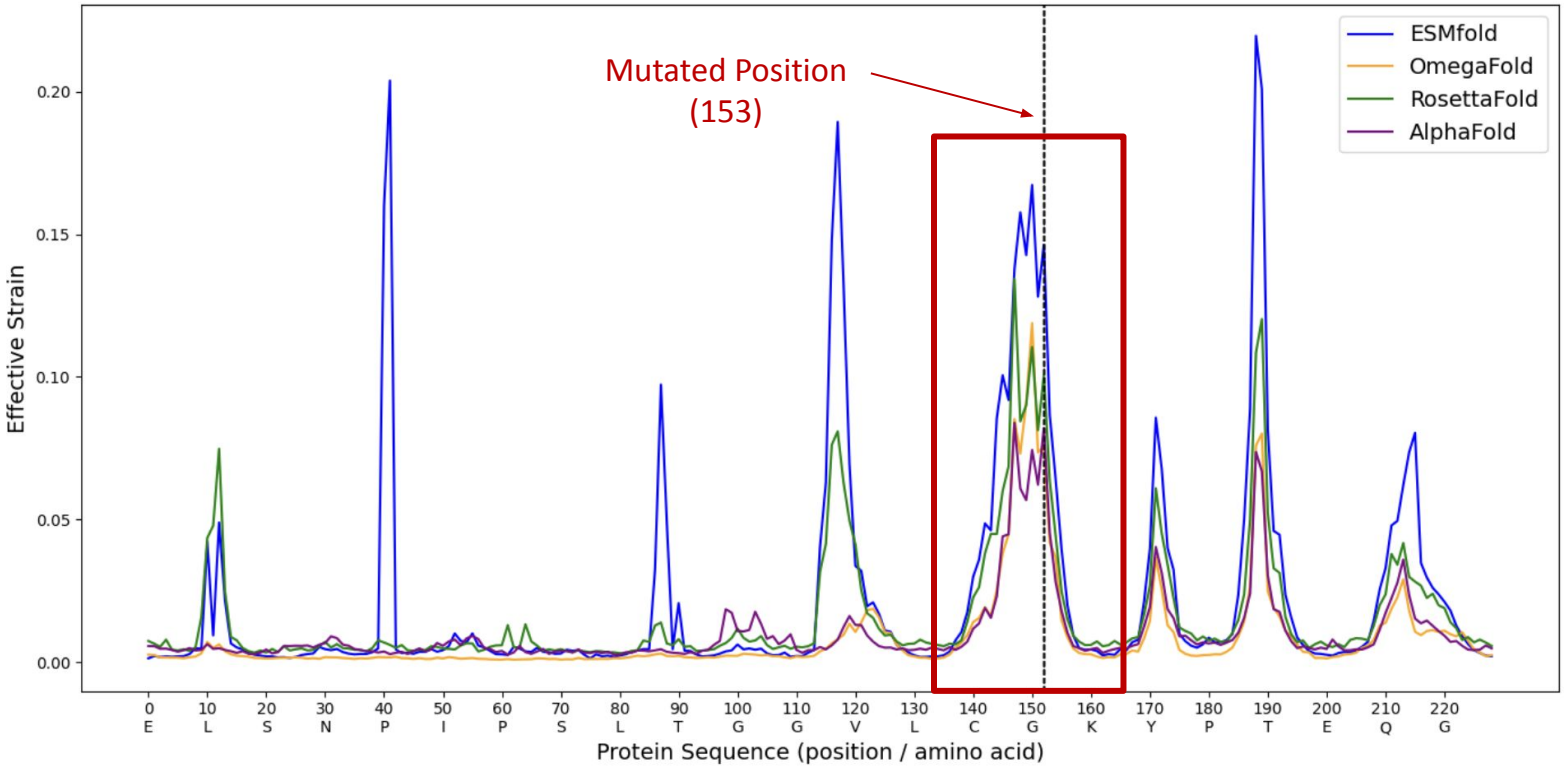


Effective Strain Plot

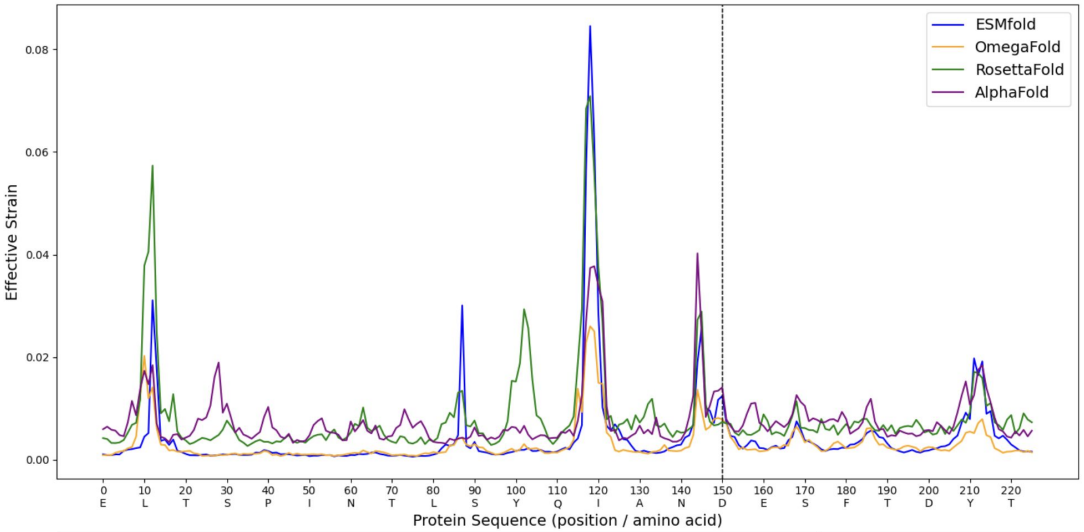


Sequences: Loki_Original vs. Loki_Single_A178S

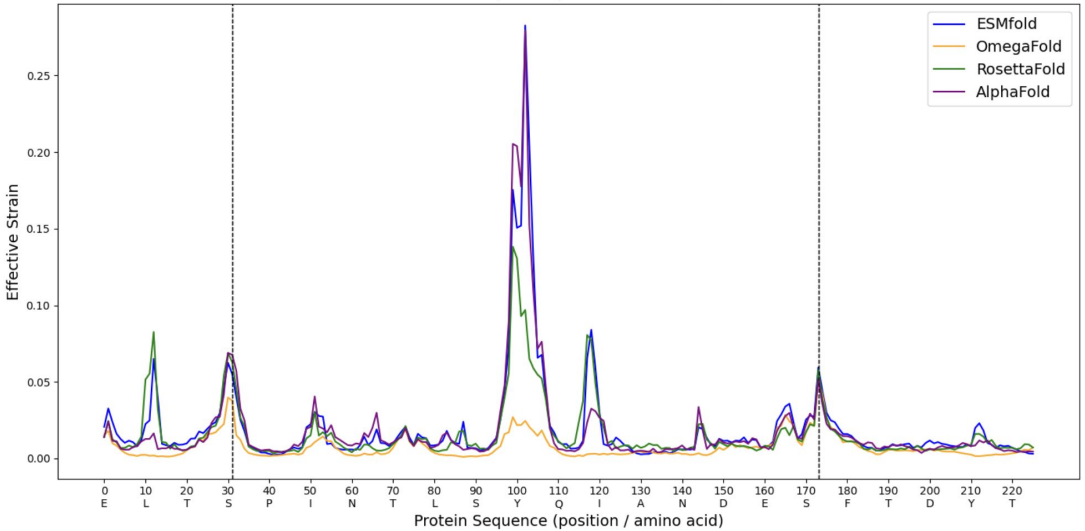
Effective Strain Plot



Sequences: Loki_Original vs. Loki_Single_G153Y



Nola_Original
vs.
Nola_Single_D151Y



Nola_Original
vs.
Nola_Double_Y174Q_S32Y

Effective Strain table - (eg: ESMFold)

Nola Antibody - Sequence	Effective Strain	Mutation Type
IL31R_Nola_Double_3_A108Y_Y174A.pdb	0.007261	Double
IL31R_Nola_Double_2_D151Y_E59N.pdb	0.005433	Double
IL31R_Nola_Double_5_F55G_Q172T.pdb	0.004125	Double
IL31R_Nola_Double_6_Q172N_A63K.pdb	0.004439	Double
IL31R_Nola_Double_1_F55S_S144A.pdb	0.003921	Double
IL31R_Nola_Double_4_Y174Q_S32Y.pdb	0.019521	Double
IL31R_Nola_Single_1_S144A.pdb	0.002632	Single
IL31R_Nola_Single_11_Q172N.pdb	0.002546	Single
IL31R_Nola_Single_4_E59N.pdb	0.002587	Single
IL31R_Nola_Single_2_F55S.pdb	0.001994	Single
IL31R_Nola_Single_12_A63K.pdb	0.002778	Single
IL31R_Nola_Single_10_Q172T.pdb	0.002906	Single
IL31R_Nola_Single_5_A108Y.pdb	0.006273	Single
IL31R_Nola_Single_7_Y174Q.pdb	0.005846	Single
IL31R_Nola_Single_8_S32Y.pdb	0.015124	Single
IL31R_Nola_Single_3_D151Y.pdb	0.004325	Single
IL31R_Nola_Single_9_F55G.pdb	0.003840	Single
IL31R_Nola_Single_6_Y174A.pdb	0.002176	Single

Loki Antibody - Sequence	Effective Strain	Mutation Type
IL31_Loki_Double_2_A178S_P61A.pdb	0.008177	Double
IL31_Loki_Double_4_N31S_F149S.pdb	0.012979	Double
IL31_Loki_Double_1_N63S_I64V.pdb	0.007787	Double
IL31_Loki_Double_3_L154Y_M155V.pdb	0.008842	Double
IL31_Loki_Double_5_G153Y_E215S.pdb	0.023643	Double
IL31_Loki_Single_7_N31S.pdb	0.001825	Single
IL31_Loki_Single_10_E215S.pdb	0.005597	Single
IL31_Loki_Single_2_I64V.pdb	0.006505	Single
IL31_Loki_Single_9_G153Y.pdb	0.023064	Single
IL31_Loki_Single_3_A178S.pdb	0.002029	Single
IL31_Loki_Single_4_P61A.pdb	0.008576	Single
IL31_Loki_Single_11_R172Y.pdb	0.003362	Single
IL31_Loki_Single_6_M155V.pdb	0.004824	Single
IL31_Loki_Single_8_F149S.pdb	0.010909	Single
IL31_Loki_Single_1_N63S.pdb	0.003108	Single
IL31_Loki_Single_5_L154Y.pdb	0.008540	Single

Effective Strain table - Summary

Model Name	Antibody	Mutation Type	Avg ES Score
ESMFold	Nola	Single	0.0044190
		Double	0.0074500
	Loki	Single	0.0071217
		Double	0.0122855
OmegaFold	Nola	Single	0.0034844
		Double	0.0052293
	Loki	Single	0.0044219
		Double	0.0080781
RosettaFold	Nola	Single	0.0082048
		Double	0.0101770
	Loki	Single	0.0087595
		Double	0.0143115
AlphaFold	Nola	Single	0.011178
		Double	0.010762
	Loki	Single	0.008258
		Double	0.010278

Takeaways:

- Model Variations: RosettaFold and Alphafold generally predict higher strain values
- Mutation Type: all models show that double mutations result in higher ES scores than single mutations
- Antibody Type: Loki shows higher ES scores on average compared to Nola

RMSD - across different tool - Nola

	esmfold_vs_ omegafold	esmfold_vs_ rosettafold	esmfold_vs_ alphafold	omegafold_vs_ rosettafold	omegafold_vs_ alphafold	rosettafold_vs_ alphafold
nola_double_1	6.327	2.515	3.953	4.341	3.182	1.651
…	…	…	…	…	…	…
nola_single_1	6.242	2.547	5.603	4.250	1.510	3.328
…	…	…	…	…	…	…
nola_original	6.214	1.745	3.863	5.000	3.137	2.282
nola_average	6.382	2.596	4.368	4.395	2.720	2.052

ESMFold ≈ RosettaFold; RosettaFold ≈ AlphaFold

RMSD - across different tool - Loki

	esmfold_vs_ omegafold	esmfold_vs_ rosettafold	esmfold_vs_ alphafold	omegafold_vs_ rosettafold	omegafold_vs_ alphafold	rosettafold_vs_ alphafold
loki_double_1	11.899	11.004	12.456	5.858	1.641	7.255
…	…	…	…	…	…	…
loki_single_1	10.990	10.680	11.892	2.656	2.762	4.981
…	…	…	…	…	…	…
loki_original	11.509	10.624	12.234	3.902	2.178	5.835
loki_average	11.570	10.568	12.302	3.839	2.239	5.708

ESMFold gives very different results from other models
OmegaFold produces similar results as AlphaFold.

Future Directions

Step 3:
Visualization and
Quantitative
metrics

Step 4 - 1:
More Mutated Sequences/
Different Mutation Locations

Step 4 - 2:
Comparison/Validation with
the experimental data

Reference

Scholarly Articles

McBride JM, et al. AlphaFold2 Can Predict Single-Mutation Effects, 2023.

McDonald, Eli Fritz, et al. Benchmarking Alphafold2 on Peptide Structure Prediction, 2022.

Zeming Lin, et al. Evolutionary-scale prediction of atomic level protein structure with a language model, 2022.

Milot Mirdita, et al. ColabFold: making protein folding accessible to all, 2022.

GitHub Repositories

https://github.com/HeliXonProtein/OmegaFold

https://github.com/facebookresearch/esm

https://github.com/sokrypton/ColabFold?tab=readme-ov-file

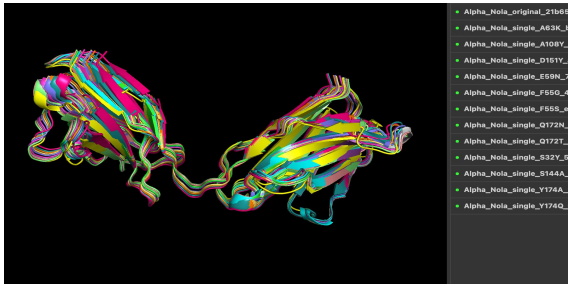
Thank you

Questions?

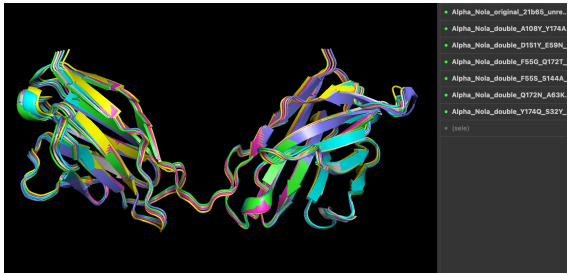
We want to thank CMU, PSC, and Merck.co team for the resources and time to allow us do this capstone project.

Reference

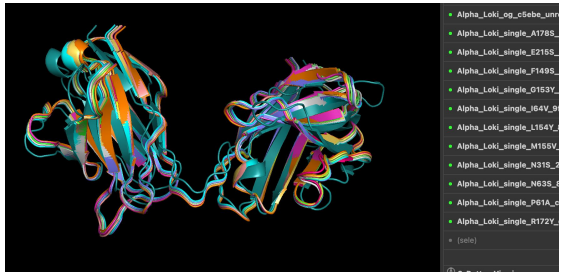
Pymol Visualizations - Alpha Fold2



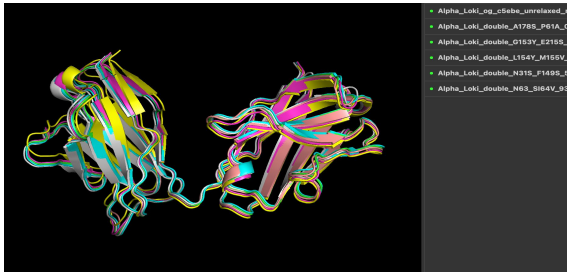
Nola - single mutation



Nola - double mutation



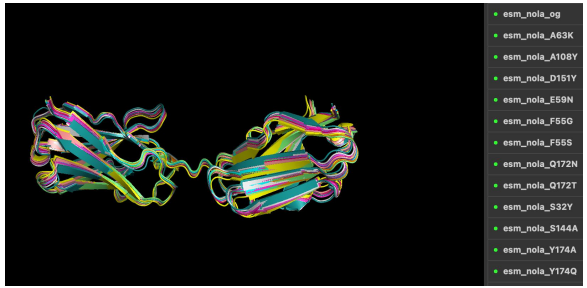
Loki - single mutation



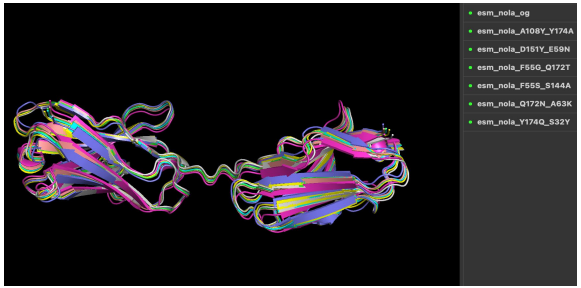
Loki - double mutation

Reference

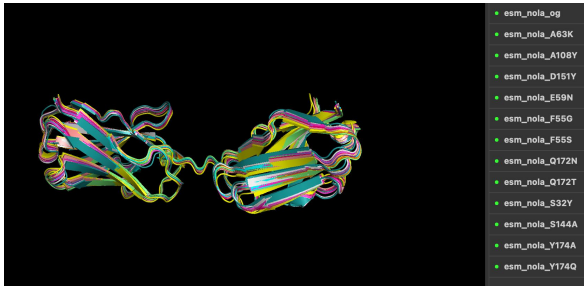
Pymol Visualizations - ESM Fold



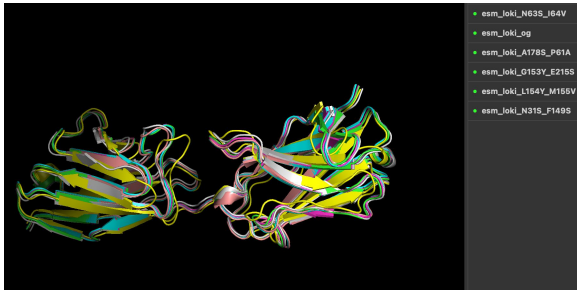
Nola - single mutation



Nola - double mutation



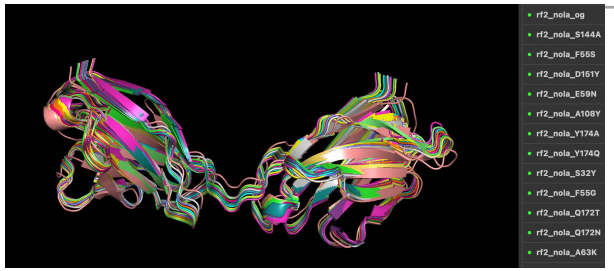
Loki - single mutation



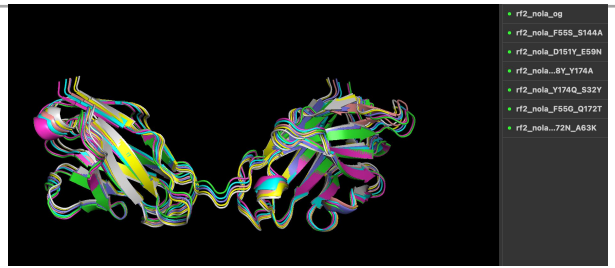
Loki - double mutation

Appendix

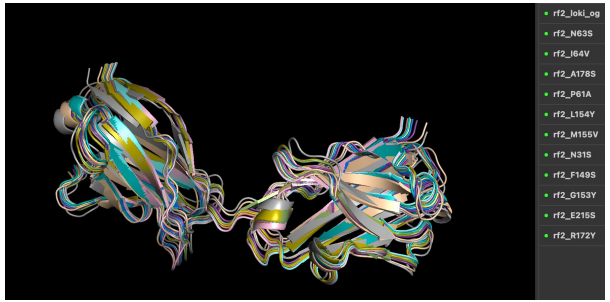
Pymol Visualizations - RoseTTA Fold



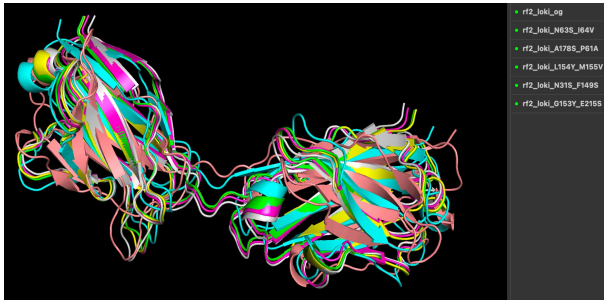
Nola - single mutation



Nola - double mutation



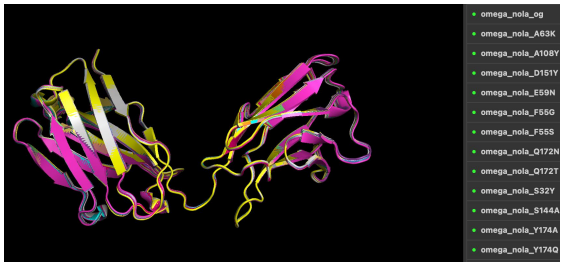
Loki - single mutation



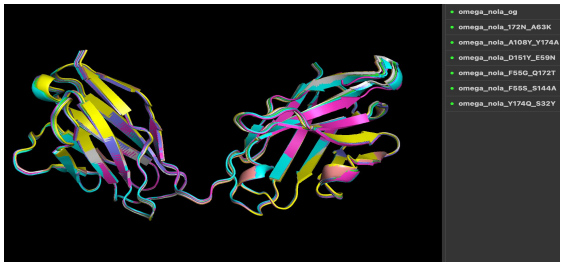
Loki - double mutation

Reference

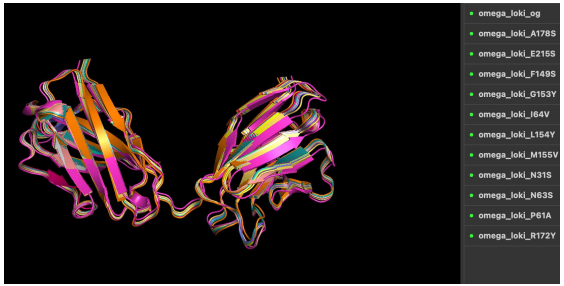
Pymol Visualizations - Omega Fold



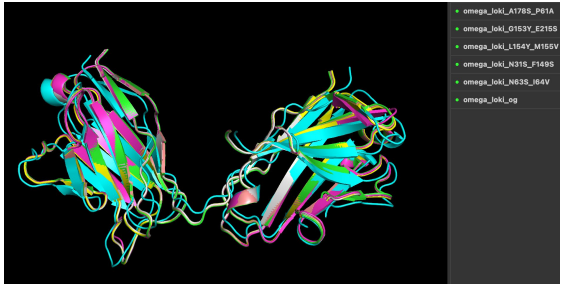
Nola - single mutation



Nola - double mutation



Loki - single mutation



Loki - double mutation