Dipartimento di Fisica e Astronomia Galileo Galilei
UNIVERSITÀ DEGLI STUDI DI PADOVA

UNIVERSITÀ DEGLI STUDI DI PADOVA

Transformer from Scratch

Group 2410 July 2024

Mandana Goudarzi	2122279
Maryam Feizi	2091504
Sandra Elsa Sanjai	2113951
Mihriban Yavas	2106188

- **Introduction**
  - -Main Problem
  - -Protein Structure
  - -Our Data
- Technical Approaches
- Results
- Conclusion

Introduction

UNIVERSITÀ DEGLI STUDI DI PADOVA

- **THE PROBLEM** → Find angles
- **MAIN SOURCE OF APPROACH** → Paper ”Enhancing protein backbone angle prediction by using simpler models of deep neural networks”

- **HOW WE CHANGE IT** → Using transformer instead
- **OTHER TRANSFORMERS** → They are more complicate

Protein structure Explanation

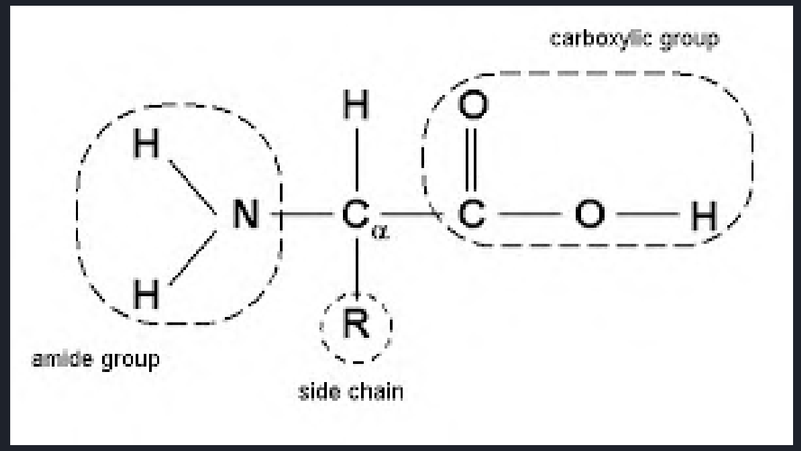


Figure 1. Amino acid

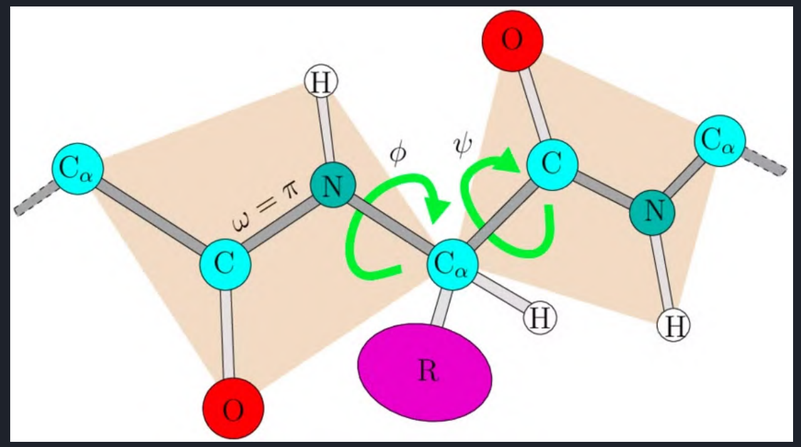


Figure 2.

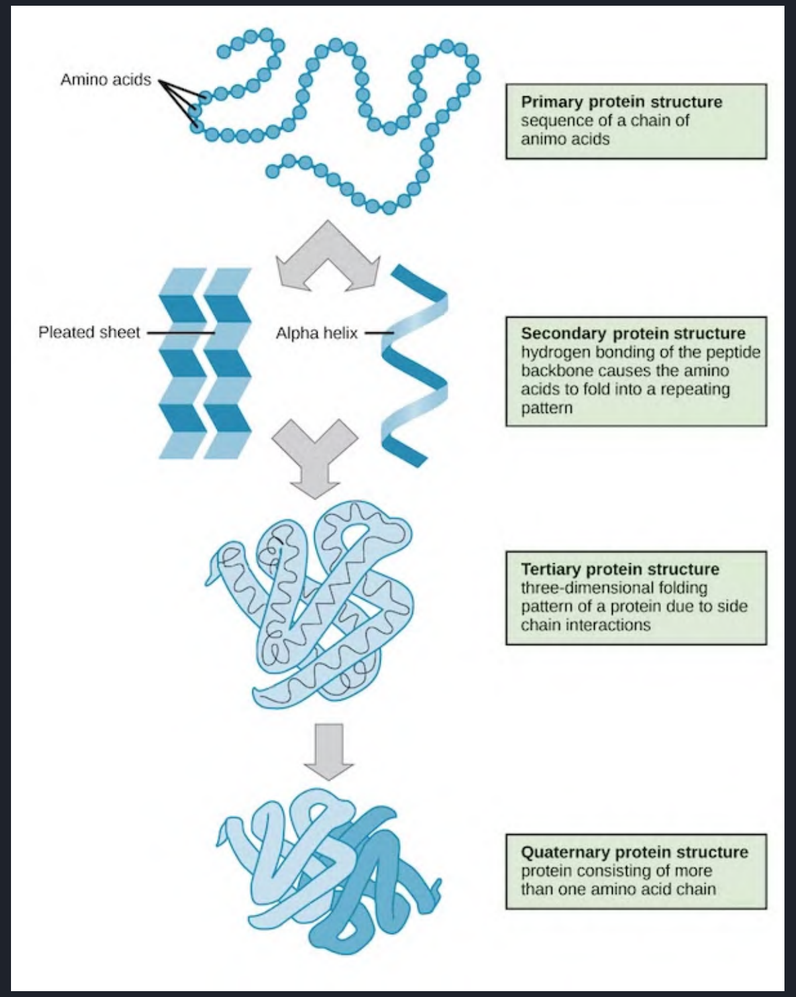


Figure 3. Protein folding

Our Data : Pisces

Università degli Studi di Padova

• Sequence identities for PDB sequences are determined by creating a hidden Markov model for every unique PDB sequence with the program HHblits (Soding et al.) and searching the resulting collection of HMMs with each individual HMM with the program HHsearch.

• For each calculated list, the server provides an output list of accession IDs (e.g., 1ABCA) with sequence length, structure determination method, resolution, and R-factor (if available) and a file of the sequences in FASTA format.

Specification

Resolution: 0.0- 2.0 • R-factor: 0.25 • Sequence length : 40-200 • Sequence percentage identity: ≤ 30.0 • NMR entries: Included • Chains with chain breaks: Excluded • Chains with disorder: Excluded.

Filtering

We filter sequences to have a **maximum length of 129**. This led to a total dataset of 1712 proteins.

Tokenizer

For encoding the protein into vectors, we used the ProtBert tokenizer.
matrix of size = length of the sequence + 2 by **1024** (which is cropped afterwards).

- Introduction
- **Technical Approaches**
  - -Prot-bert
  - -Transformer Architecture
  - -Self Attention
  - -Custom Loss
  - -Masking
- Results
- Conclusion

Bert Model

Is a model specifically designed for protein sequence analysis

Is an "encoder only" model

Prot-Bert is a model for predicting proteins structures

Layers Within Architecture:

- Multi-Head Self-Attention
- Normalization
- Feed-Forward Neural Network
- Residual Connections
- Dropout

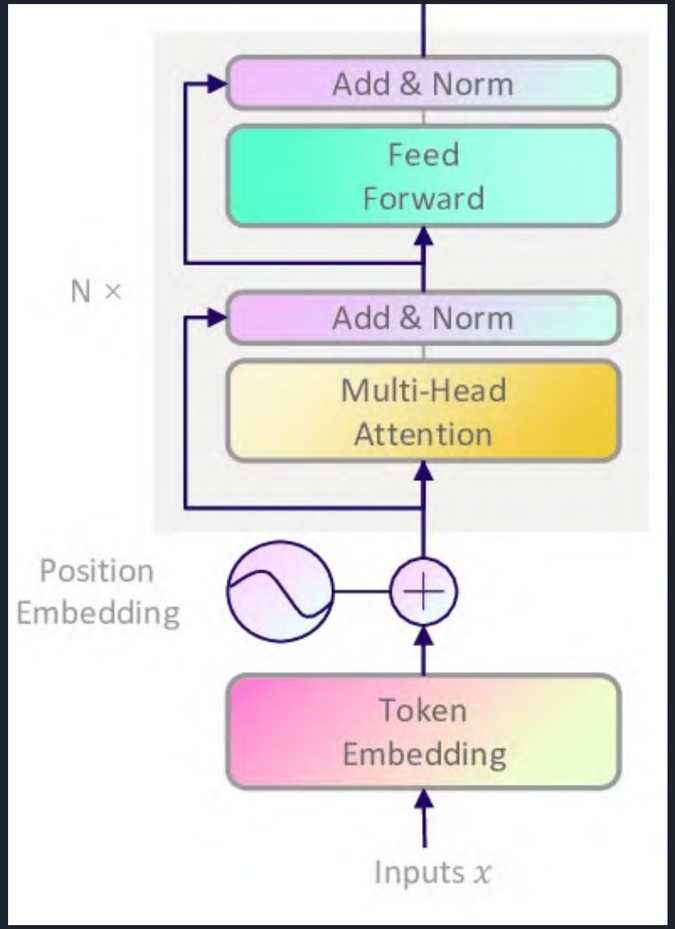
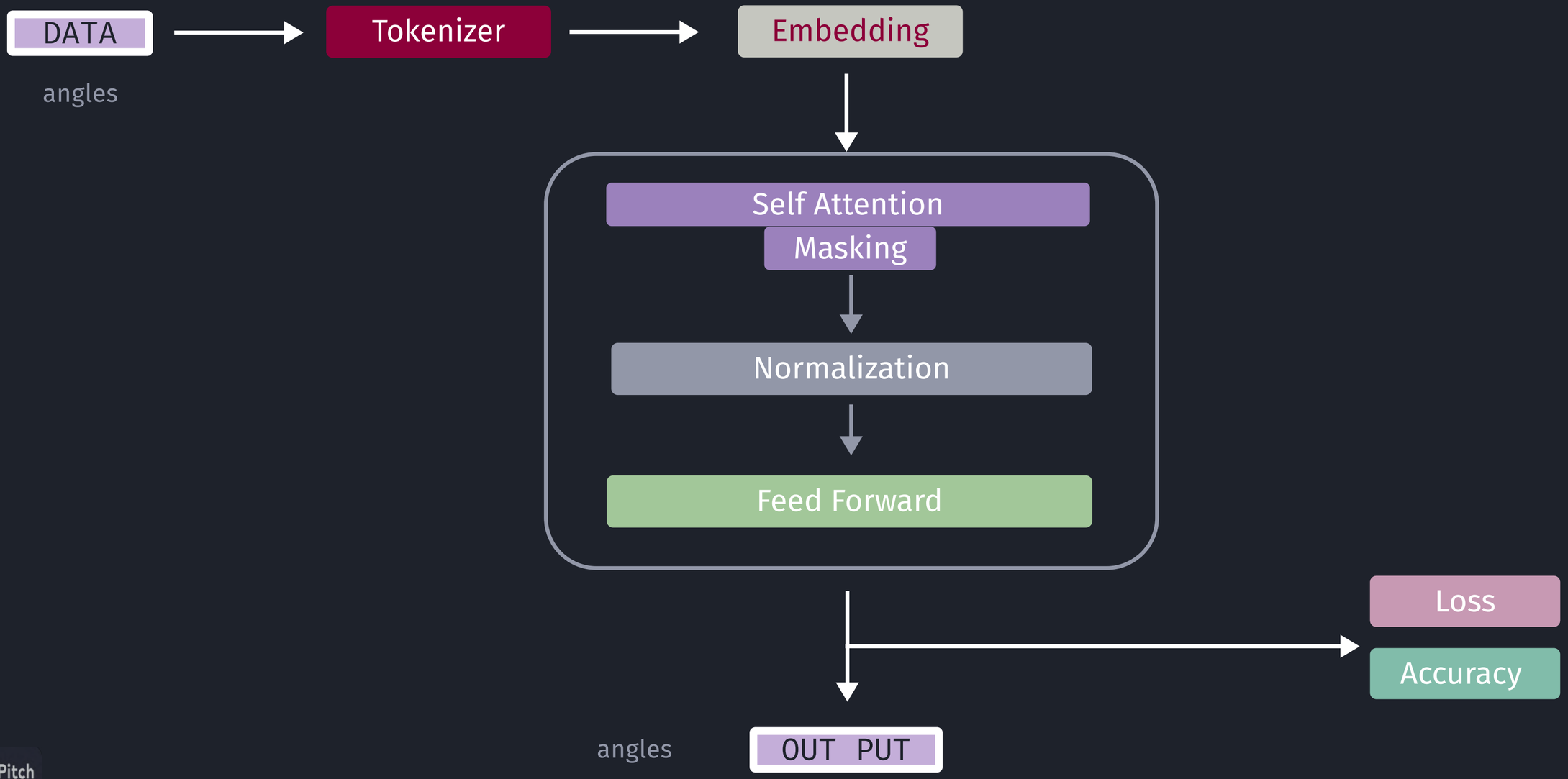


Figure 4. Prot-Bert architecture

Transformer Architecture



Self-Attention Mechanism

UNIVERSITÀ DEGLI STUDI DI PADOVA

Since we are doing a translation problem, (sequence to angles) we do not need masking. But for training we padded the sequences and angles to have a consistent length, so we needed to find a way to not let these padded values effect the translation.

PROCESS:

- Calculate Q, K, and V matrices
- Compute attention scores
- Apply the attention mask
- Compute Attention output

$$\text{attention\_output} = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}} + \text{mask}^T\right) \cdot V$$

Eq 1. Attention with causal mask

Loss Function

TORUS DISTANCE AS LOSS FUNCTION

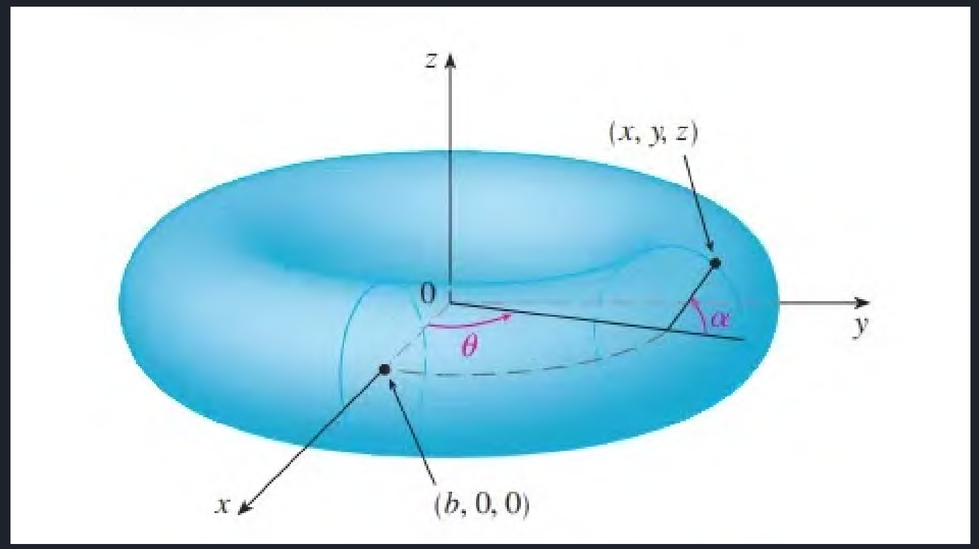


Figure 5. Thorus

$$loss = \sum_{i=1}^{N} min(\psi_{pred} - \psi_{true}, 360 - (\psi_{pred} - \psi_{true}))$$

Eq 2. Thorus distance function used as angle loss

Masking

- Masking is crucial to handle padded sequences properly.
- Ensures the model does not attend to learn irrelevant positions, improving training efficiency and performance.

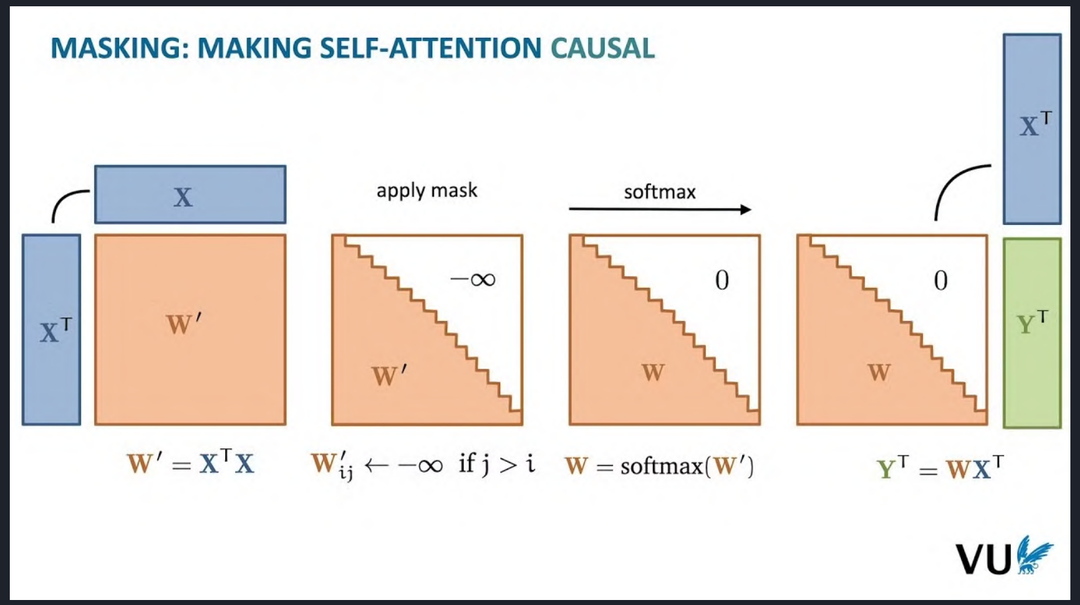


Figure 6. Causal Self-Attention

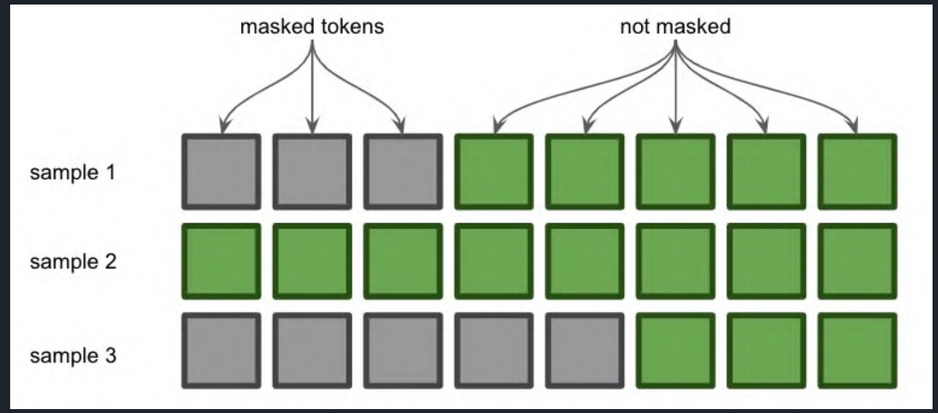


Figure 7. Masking

- Introduction
- Technical Approaches
- **Results**
    - -Ramachandran
    - - Predictions
- Conclusion

Ramachandran Plots : Pisces

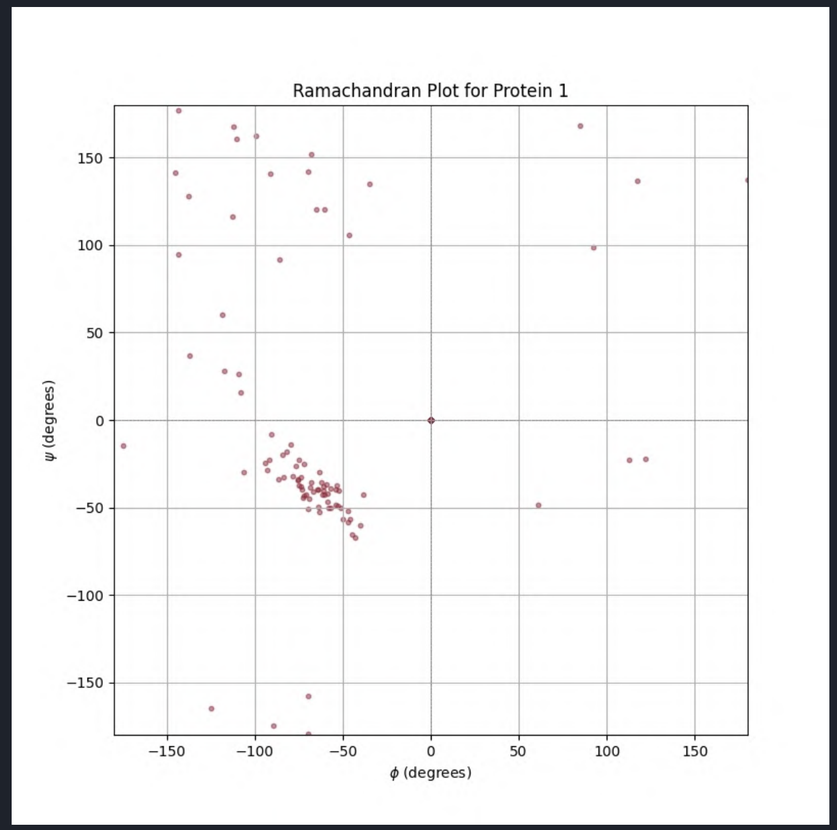


Figure 8. Original Pisces angles

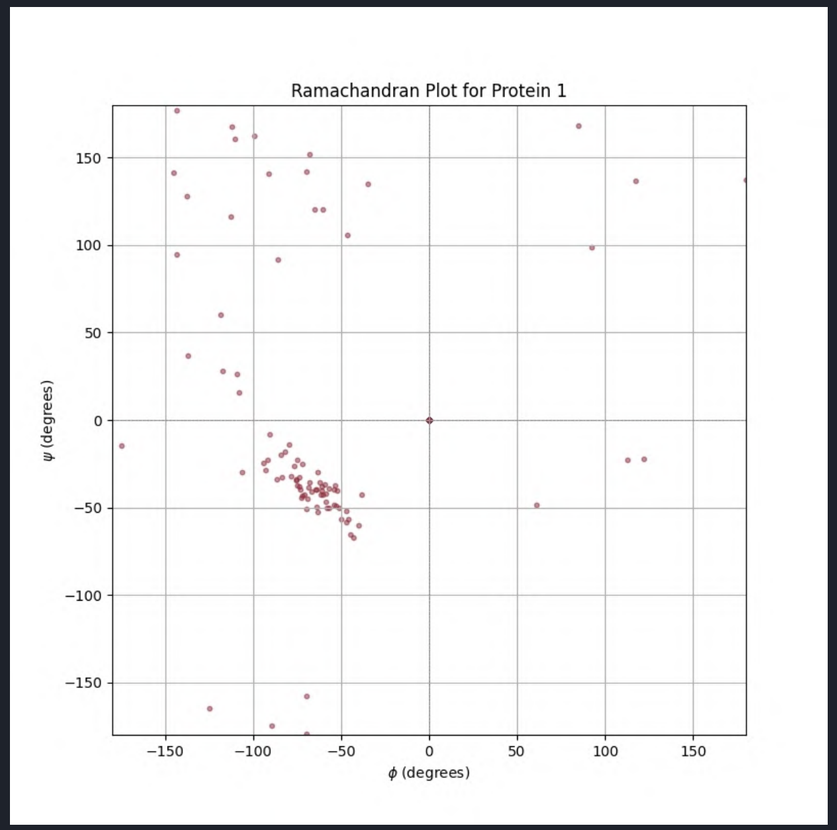


Figure 9. Predicted Pisces angles

Training with Pisces dataset

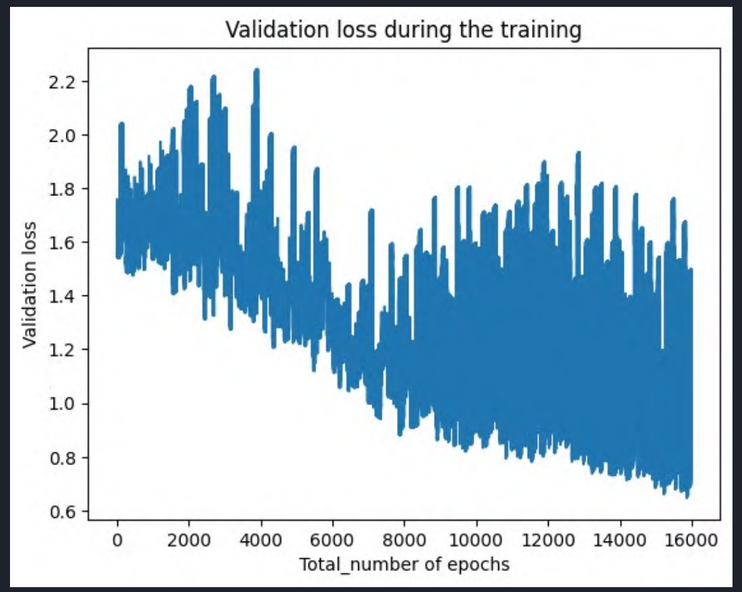


Figure 10. 800 data points with 200 epochs

- Angle-based loss: **50.4574**
- Mean absolute error for phi: **17.8034**
- Mean absolute error for psi: **16.9545**

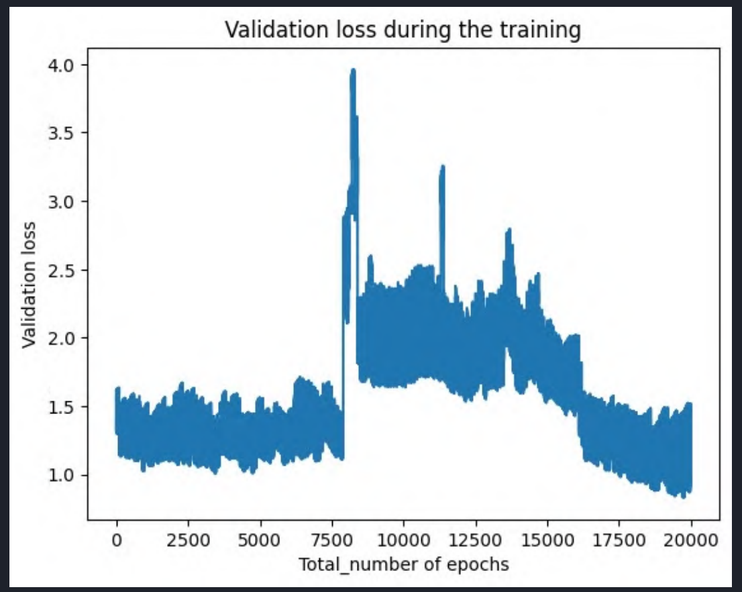


Figure 11. 1000 data points with 200 epochs

- Angle-based loss: **79.9202**
- Mean absolute error for phi: **36.1058**
- Mean absolute error for psi: **25.2288**

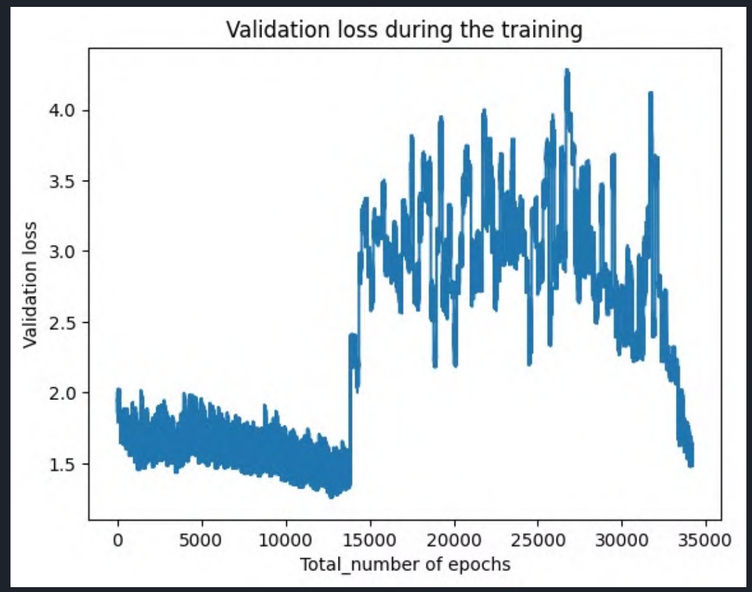


Figure 12. Full data points with 200 epochs

- Angle-based loss: **80.7814**
- Mean absolute error for phi: **39.5436**
- Mean absolute error for psi: **31.6467**

Evolution of training

Φ ANGLE DISTRIBUTION BEFORE AND AFTER TRAINING FOR A SINGLE PROTEIN

Low training, epochs = 10

Another example of this bounded angles can be seen in the angle plots. In the early epochs, the predicted angles are tightly bound and do not cover the full range of actual angles. The model likely focuses on minimising the overall loss by predicting the more probable angles.

High training, epochs = 100

As the training progresses, the model becomes better at generalizing

Secondary vs Tertiary

The model at first is better at predicting secondary structures than tertiary. Tertiary structure required more data. We can theorise that this is due to multiple points available for secondary structure even in a single protein, but the critical folding points are lesser in comparison.

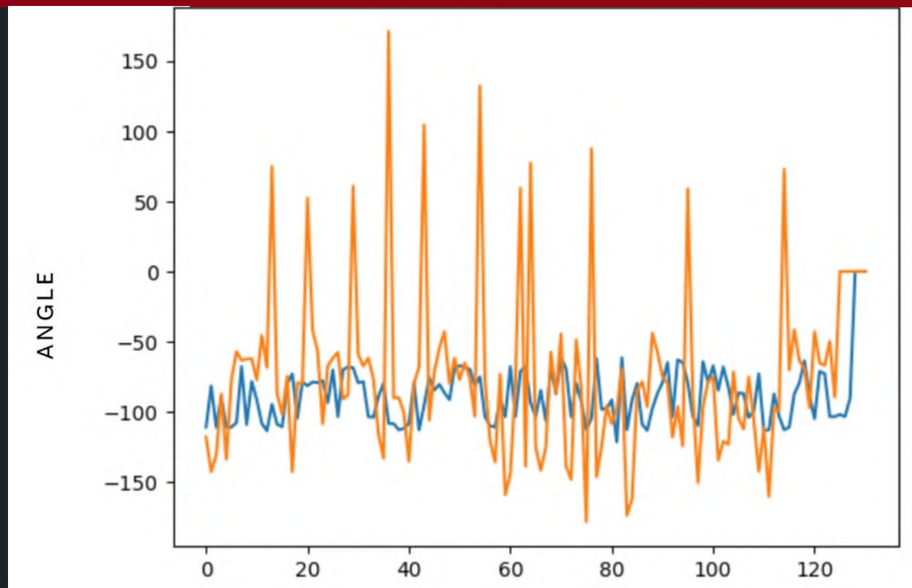
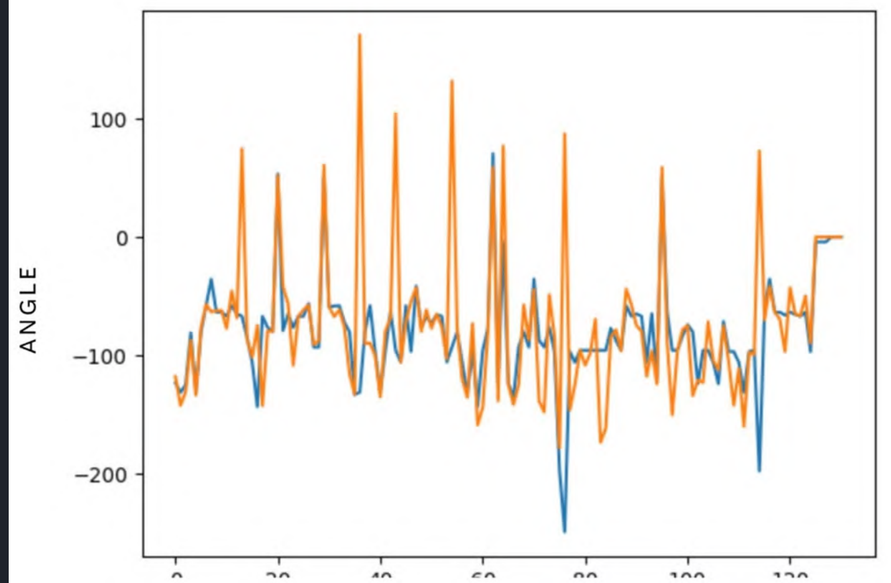


Figure 13.



Ramachandran Plots : AlphaFold

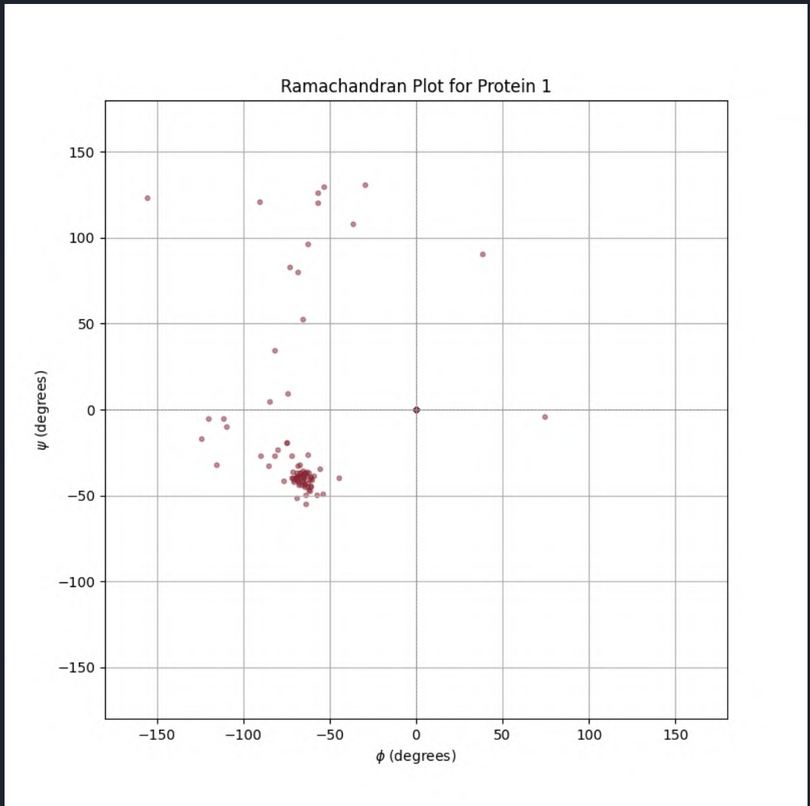


Figure 14. Original AlphaFold angles

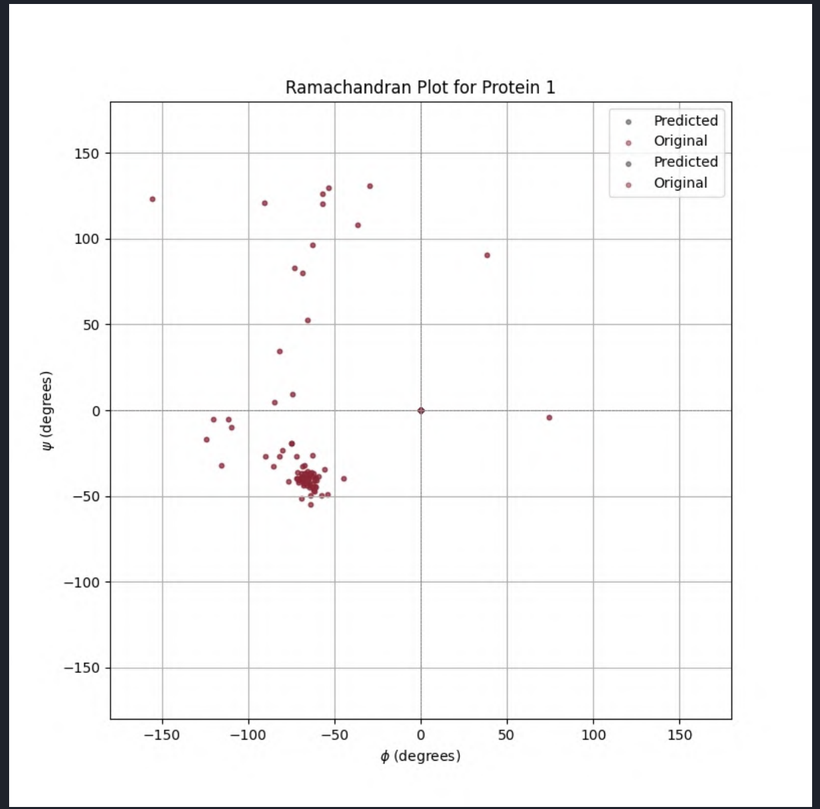


Figure 15. Predicted Alphafold angles

Training with AlphaFold dataset

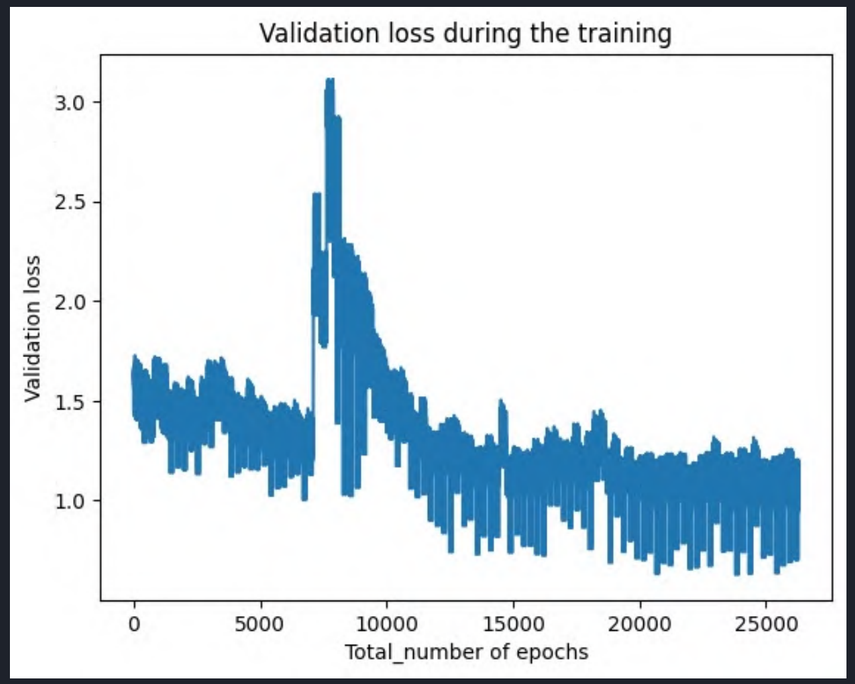


Figure 16. Full data points with 100 epochs

- Angle-based loss: **66.1113**
- Mean absolute error for phi: **31.4715**
- Mean absolute error for psi: **18.0226**

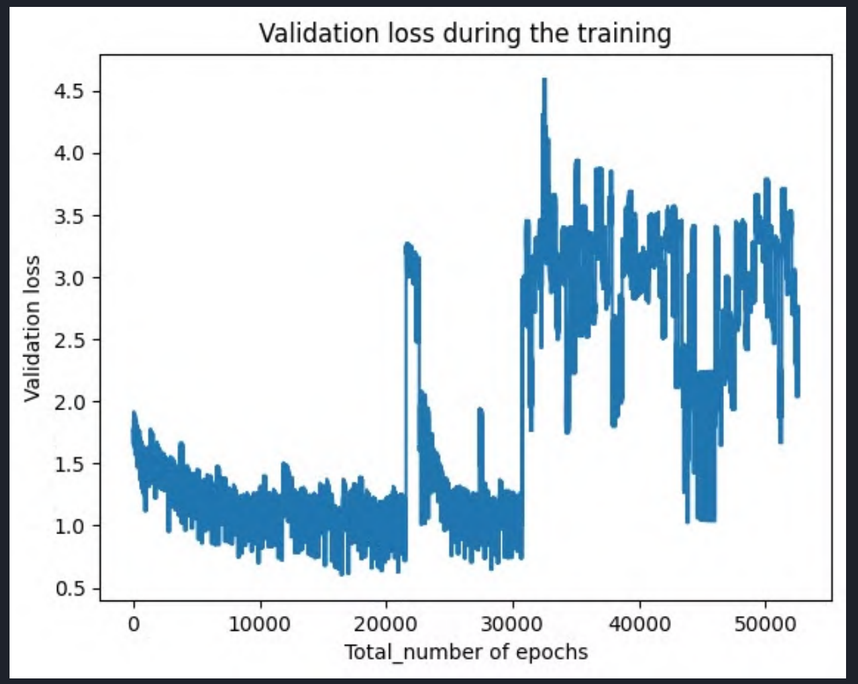


Figure 17. Full data points with 200 epochs

- Angle-based loss: **86.7574**
- Mean absolute error for phi: **41.6457**
- Mean absolute error for psi: **37.9755**

First Pisces then AlphaFold

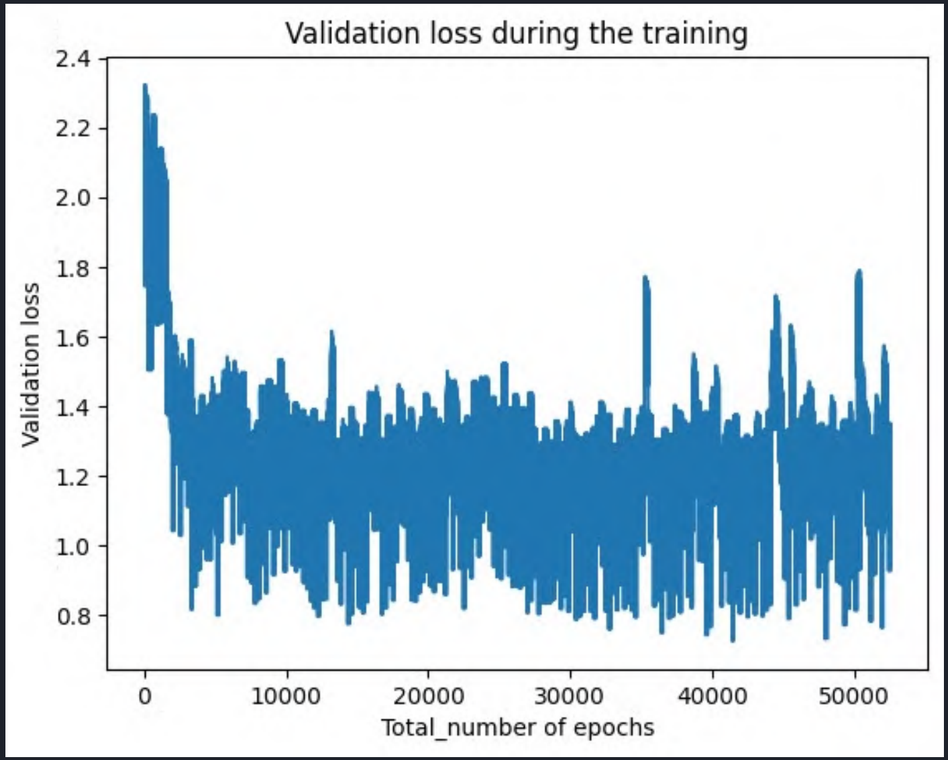


Figure 18. Model trained with Pisces and then with AlphaFold

- Angle-based loss: **61.0769**
- Mean absolute error for phi: **26.9673**
- Mean absolute error for psi: **21.2135**

First AlphaFold then Pisces

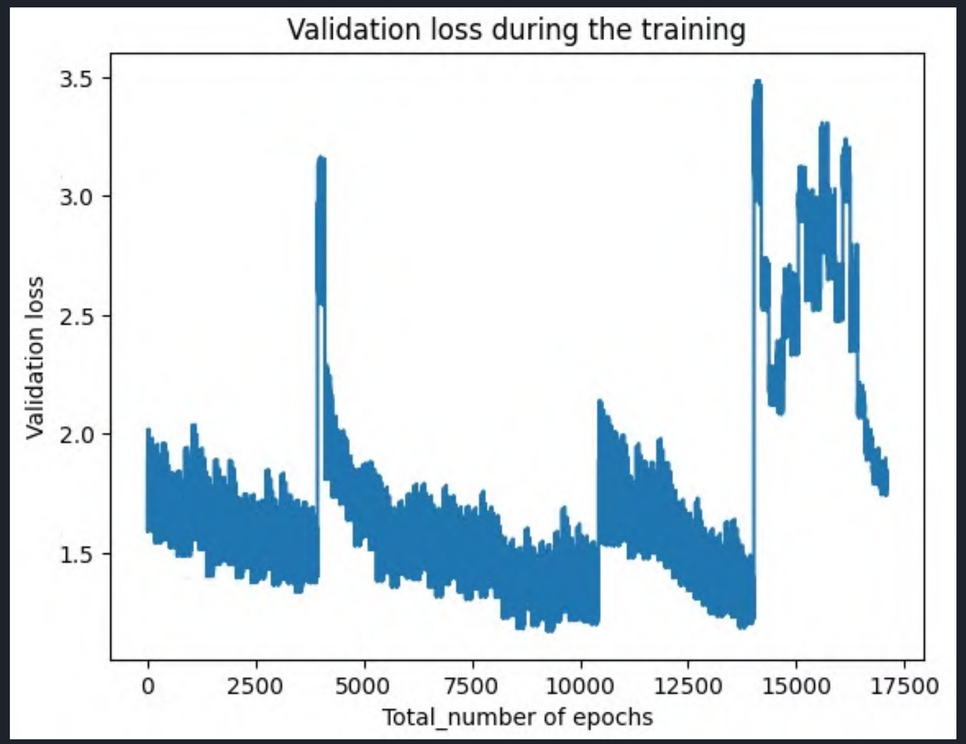


Figure 19. Model trained with AlphaFold and then trained with Pisces

- Angle-based loss: **94.3603**
- Mean absolute error for phi: **44.0252**
- Mean absolute error for psi: **41.4135**

Distribution of Psi

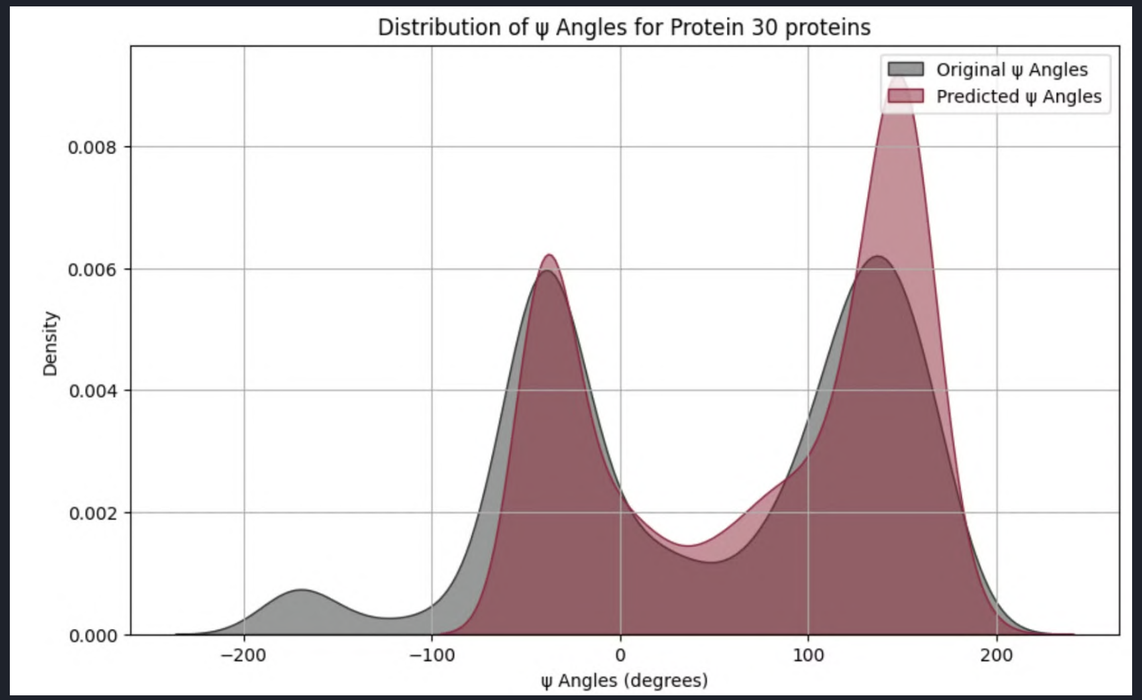


Figure 20.a Distribution of Psi for Pisces

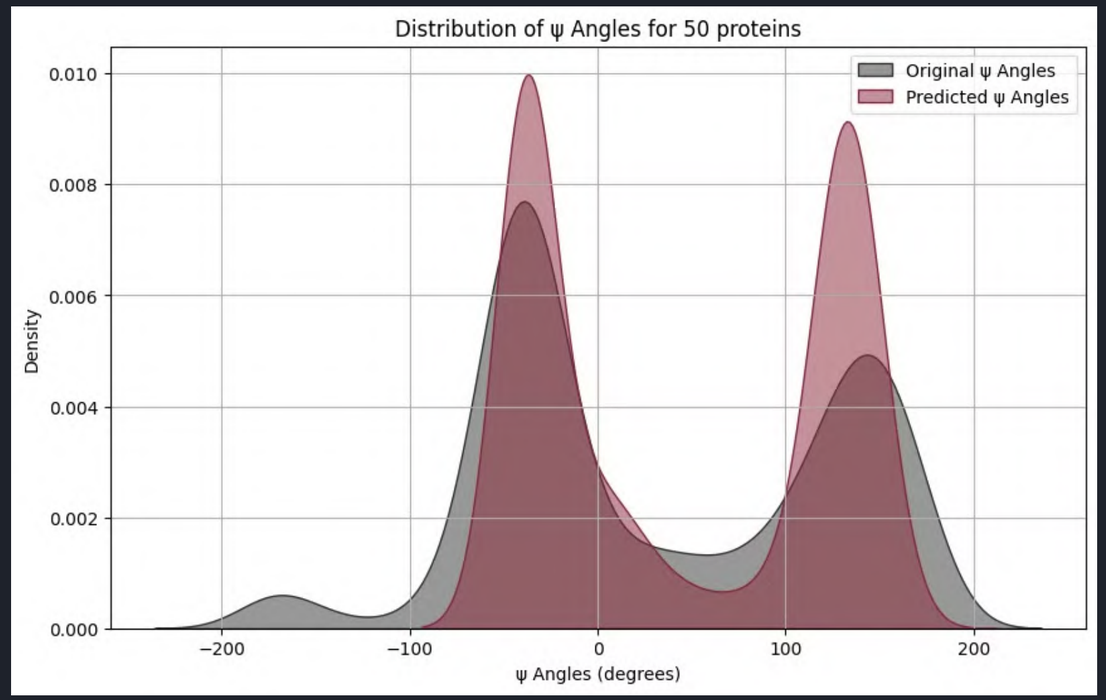


Figure 20.b Distribution of Psi for AlphaFold

Visualisation in PyMol

UNIVERSITÀ DEGLI STUDI DI PADOVA

PROTEIN 1UG7- TRAIN

Figure 21.a

Original

Expand on it here. Why is it important? Why does it matter?

Figure 21.b

50 proteins

You already know that it's important. But what about your listeners?

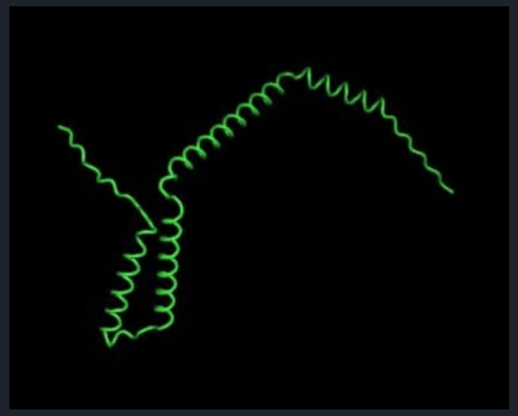


Figure 21.c

500 proteins

notice the improvement in the helix

Figure 21.d

1711 proteins

Better Tertiary structure when increasing data not epoch

Visualisation in PyMol

UNIVERSITÀ DEGLI STUDI DI PADOVA

PROTEIN 2FM4 - TRAIN

Figure 22.a

original

You already know that it's important. But what about your listeners?

Figure 22.b

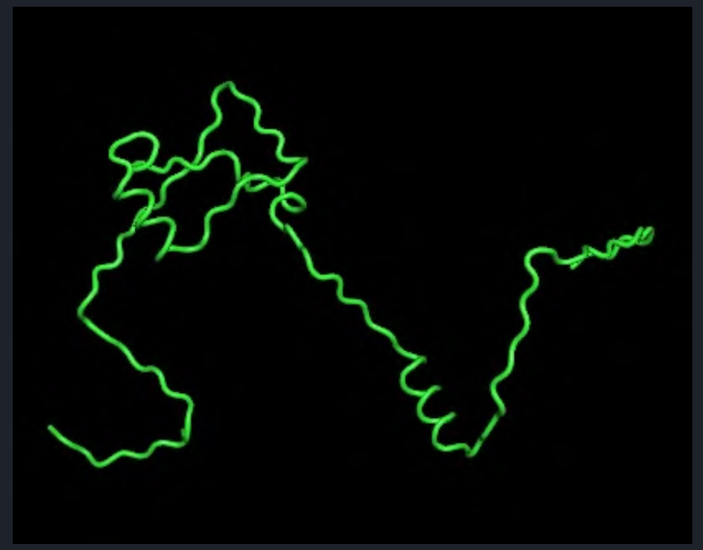


Figure 22.c

1711 proteins

Convince the audience, both with facts and with GIFs.

Visualisation in PyMol

UNIVERSITÀ DEGLI STUDI DI PADOVA

PROTEIN 2MYJ - TEST

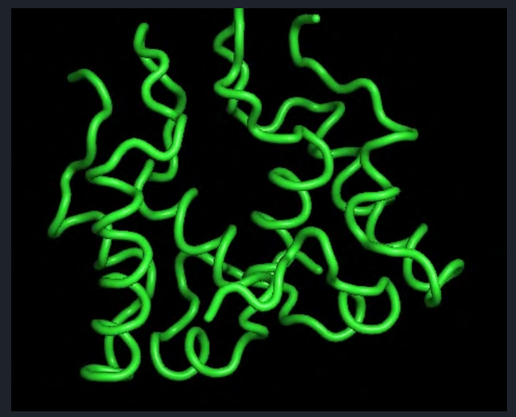


Figure 23.a

Original

a symmetric protein

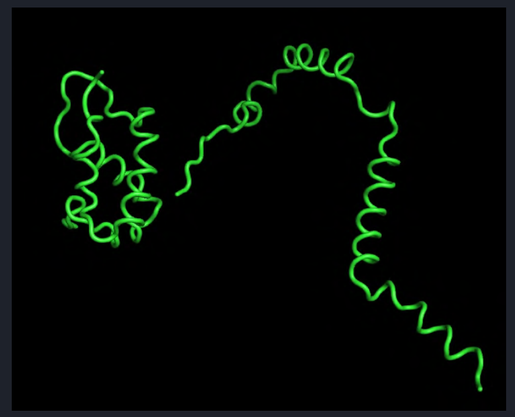


Figure 23.b

Predicted

notice the improvement in the helix

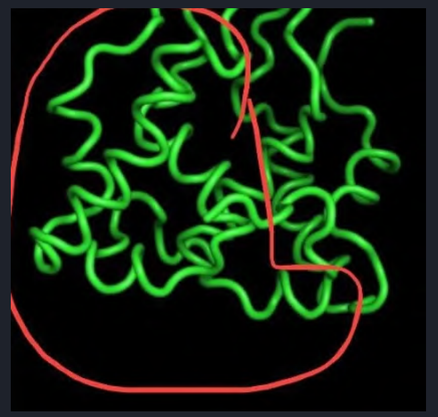


Figure 23.c

Original different angle

You already know that it's important. But what about your listeners?

Figure 23.d

Predicted Another angle

notice the improvement in the helix

Comparing

UNIVERSITÀ DEGLI STUDI DI PADOVA

PROTEIN 2MYJ - TEST USING PISCES

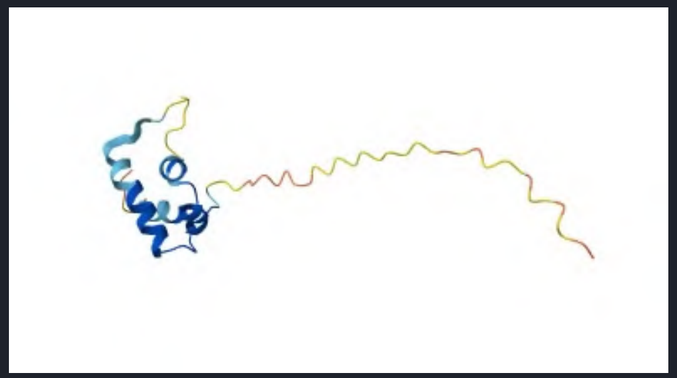


Figure 24.a

AlphaFold model

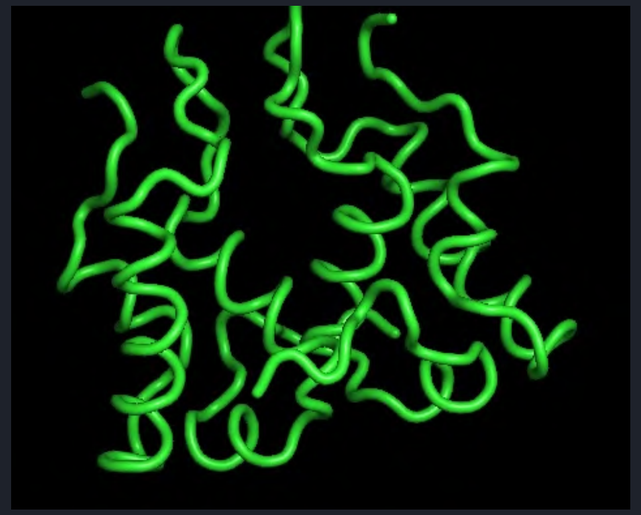


Figure 24.b

Original Model

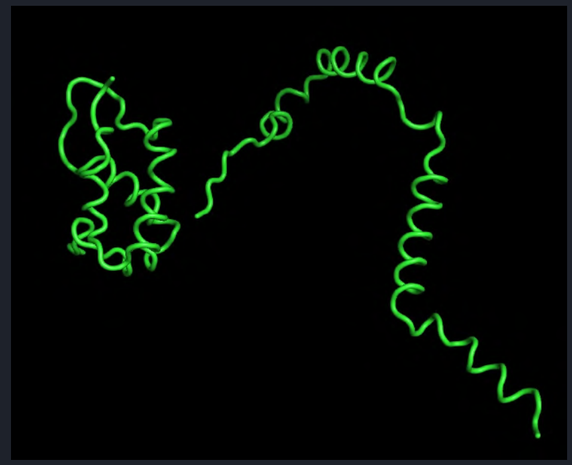


Figure 24.c

Our Result

Comparing

UNIVERSITÀ DEGLI STUDI DI PADOVA

PROTEIN P78946 - TEST

Figure 25.a

AlphaFold model

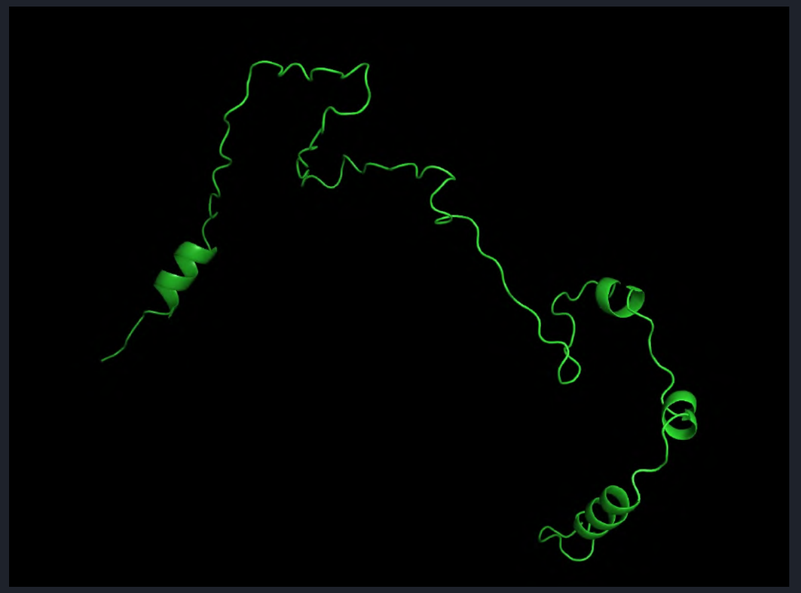


Figure 25.b

Our Result

Comparing

UNIVERSITÀ DEGLI STUDI DI PADOVA

PROTEIN Q57787 - TEST

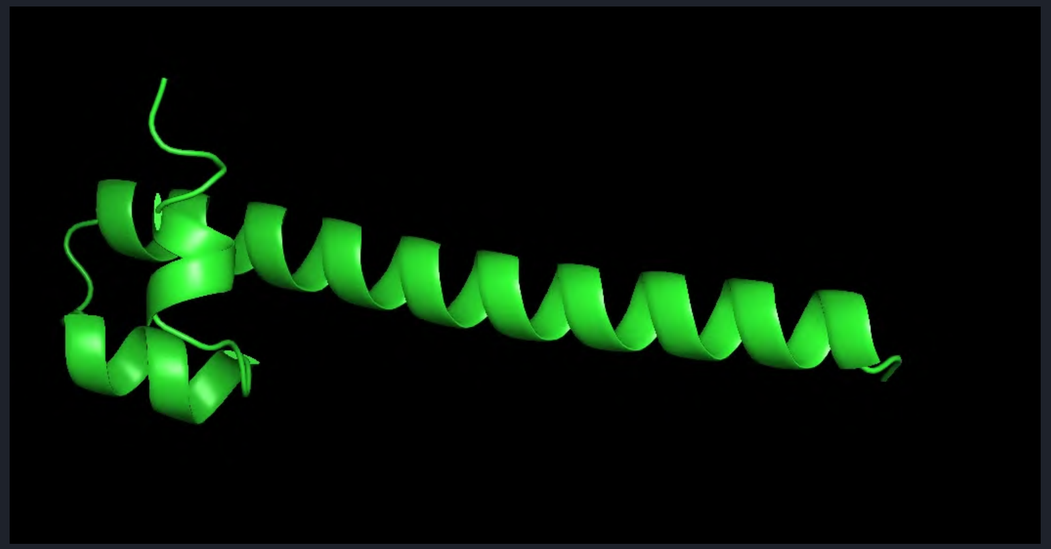


Figure 26.a

AlphaFold model

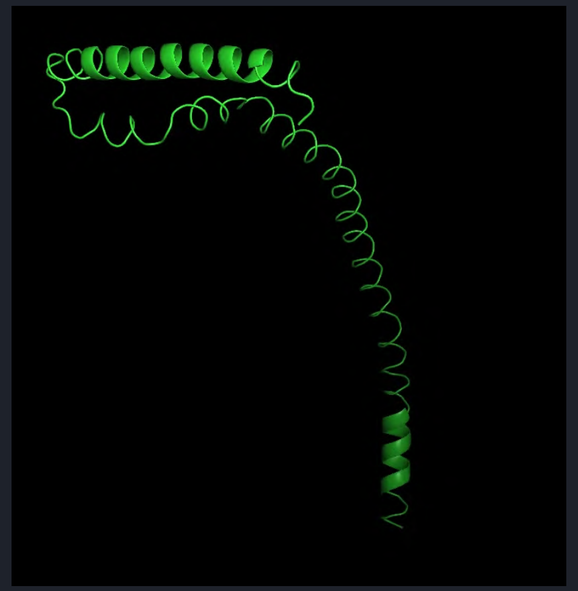


Figure 26.b

Our Result

Comparing

UNIVERSITÀ DEGLI STUDI DI PADOVA

PROTEIN V9HVX0 - TEST

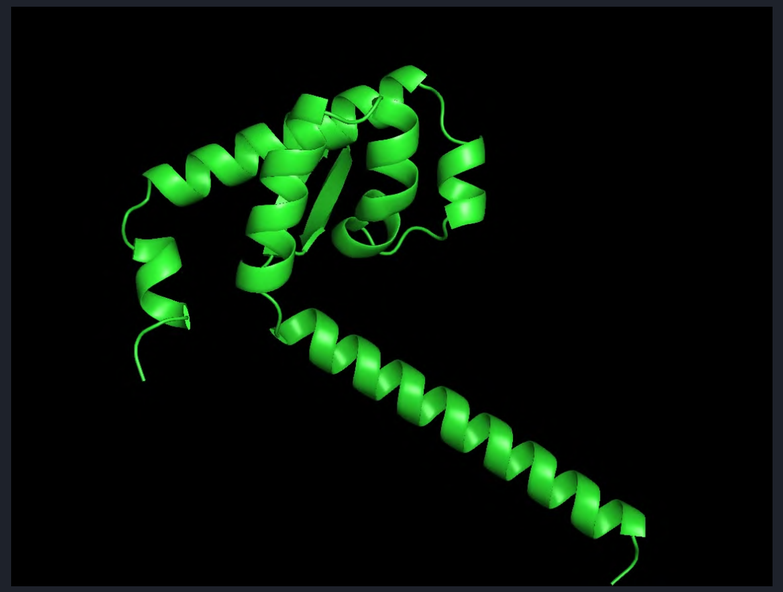


Figure 27.a

AlphaFold model

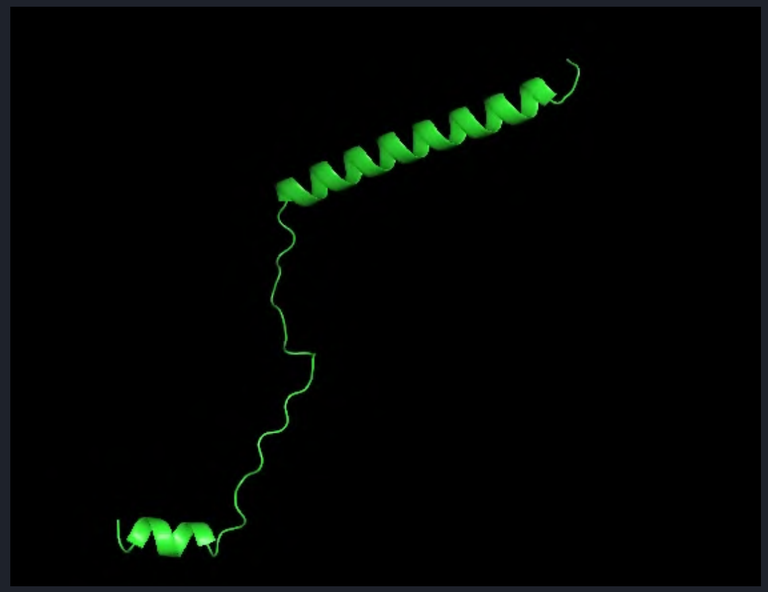


Figure 27.c

Our Result

Comparing

UNIVERSITÀ DEGLI STUDI DI PADOVA

PROTEIN 1UG7 - TEST

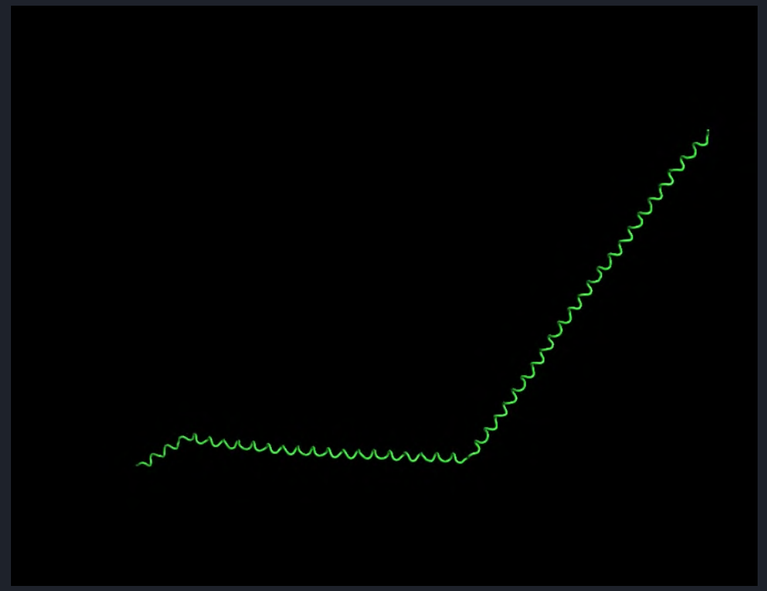


Figure 28.a

Our Result with Alphafold- Pisces

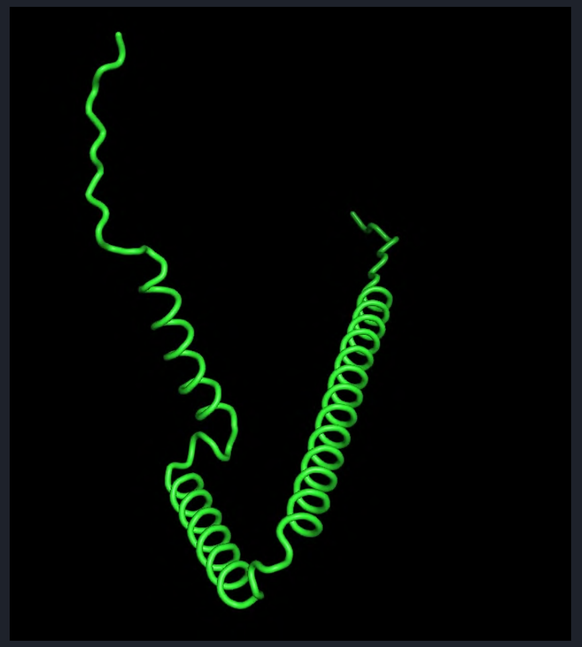


Figure 28.b

Our Result with Pisces-Alphafold

Comparing

UNIVERSITÀ DEGLI STUDI DI PADOVA

PROTEIN P78946 - TEST

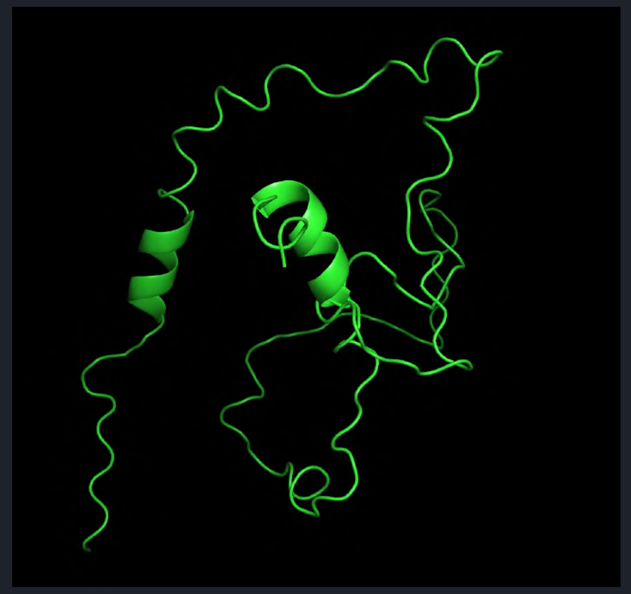


Figure 29.a

Our Result with Alphafold- Pisces

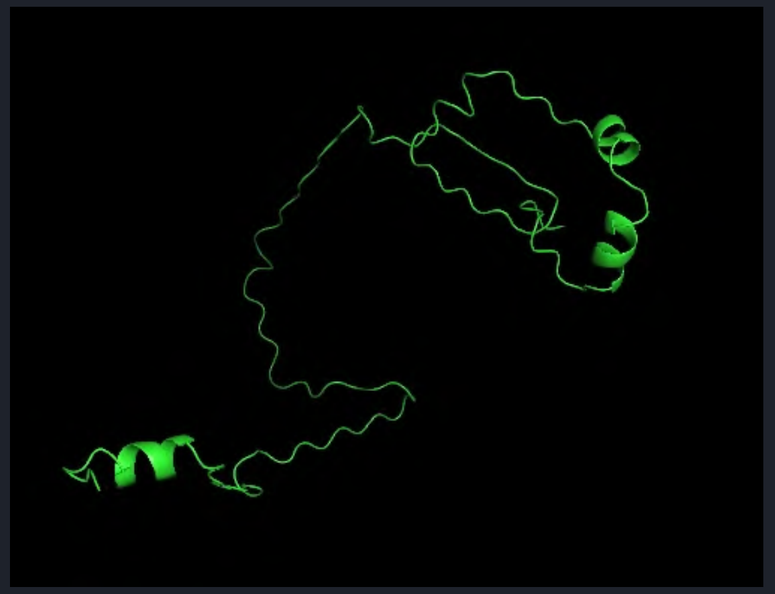


Figure 29.b

Our Result with Pisces-Alphafold

- Introduction
- Technical Approaches
- Results
- **Conclusion**

Conclusion

UNIVERSITÀ DEGLI STUDI DI PADOVA

- Good at secondary structure prediction
- Need more data for tertiary structure maybe even another loss
- A good embedding space is one of the keys

WHAT CAN BE IMPROVED:

- Datasets with only similar lengths (i.e. all close to 129)
- Multiple transformer layers
- More prior information (weights)

Thank you!

Pitch

Want to make a presentation like this one?

Start with a fully customizable template, create a beautiful deck in minutes, then easily share it with anyone.

Create a presentation (It's free)