# Protein Structure Prediction and Analysis Application

Menna Allah Whdan Esraa Mahmoud
Under Supervision of Prof. Waleed Eid

# Table of Contents

# Introduction

- Protein structure prediction is a field of research that aims to predict the three-dimensional structure of proteins using computational methods.
- The structure of a protein is essential for understanding its function, interactions, and potential implications in diseases.
- AlphaFold is machine learning algorithm that predicts the protein structure.
- It helps to understand the diseases and develop cure.

# Evolutionary Scale Modeling (ESM)

- Evolutionary Scale Modeling (ESM) employs artificial intelligence (AI) to analyze patterns in protein sequences.
- The primary goal is  to learn statistical patterns and relationships among amino acids within these sequences.
- It has demonstrated success in capturing information related to the folded structure and function of proteins.

# Environmental Sustainability Model Fold (ESM Fold)

- Focuses on predicting the three-dimensional structures that proteins adopt.

- Essential for studying protein structure-function relationships and drug discovery.

- Applications of the Resulting ESM Fold Model:

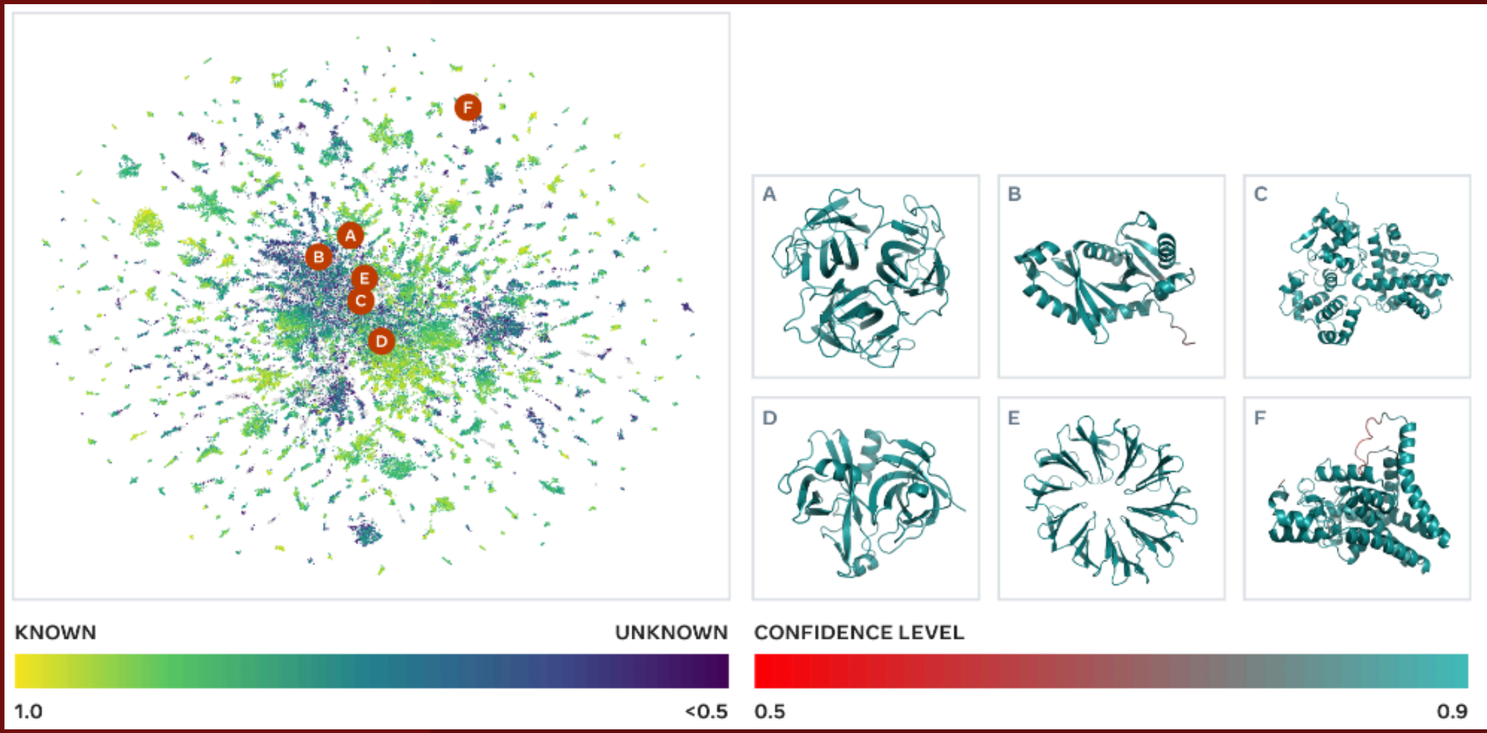
    1. Predicting protein evolution

    2. Understanding protein structure

    3. Contributing to diverse biological and medical applications

# Unlocking a Hidden Natural World

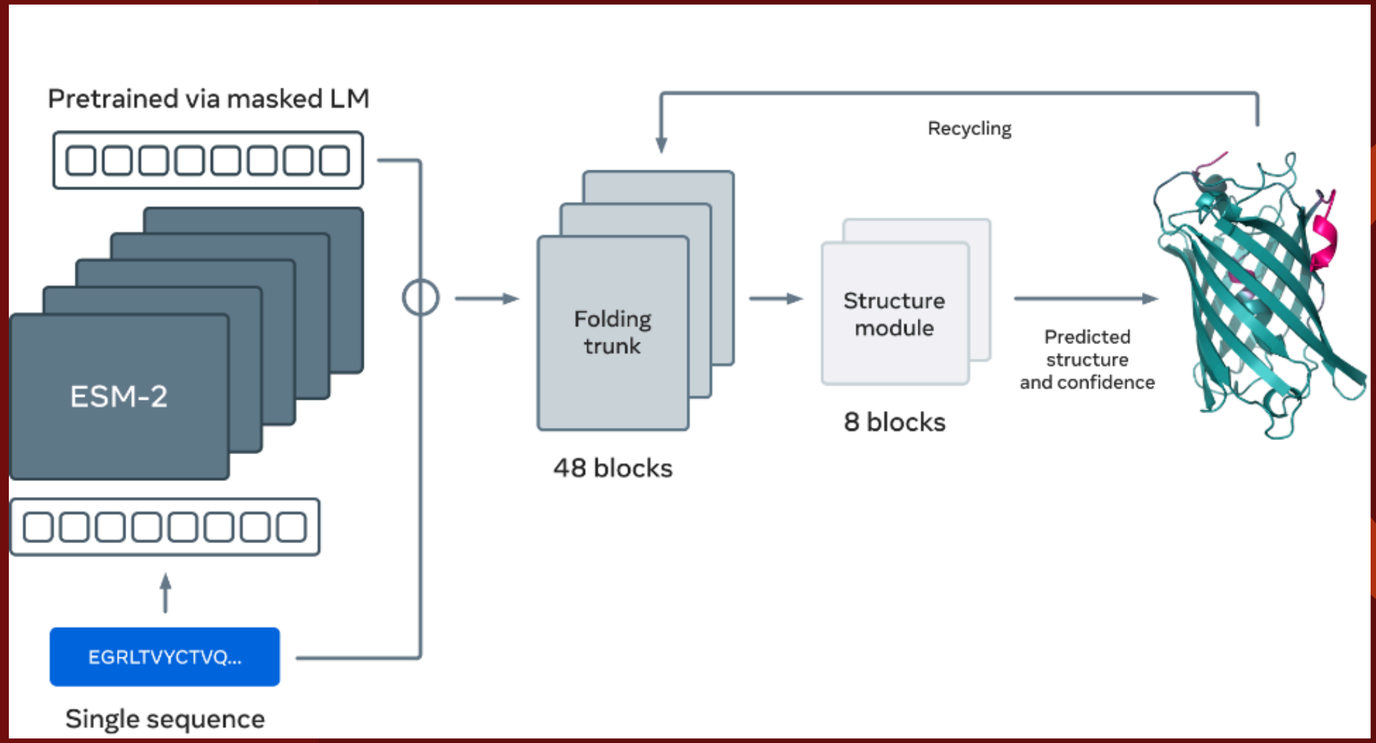
- Made it possible to catalog billions of **metagenomic** protein sequences.
- Determining the three-dimensional structures .
- Help in **discover new proteins** that can be useful in medicine and other applications.




https://ai.meta.com/blog/protein-folding-esmfold-metagenomics/

# Protein Folding with Language Modeling

- **Uses artificial intelligence (AI) to analyze and understand patterns in protein sequences.**
- **Which are sequences of amino acids, the building blocks of proteins.**
- **These sequences, with 20 possible amino acids at each position.**




https://ai.meta.com/blog/protein-folding-esmfold-metagenomics/

# Application Interface Overview

ESMFold

Protein Structure Prediction and Analysis Application

Input sequence

Predict

Enter protein sequence data!

ESMFold

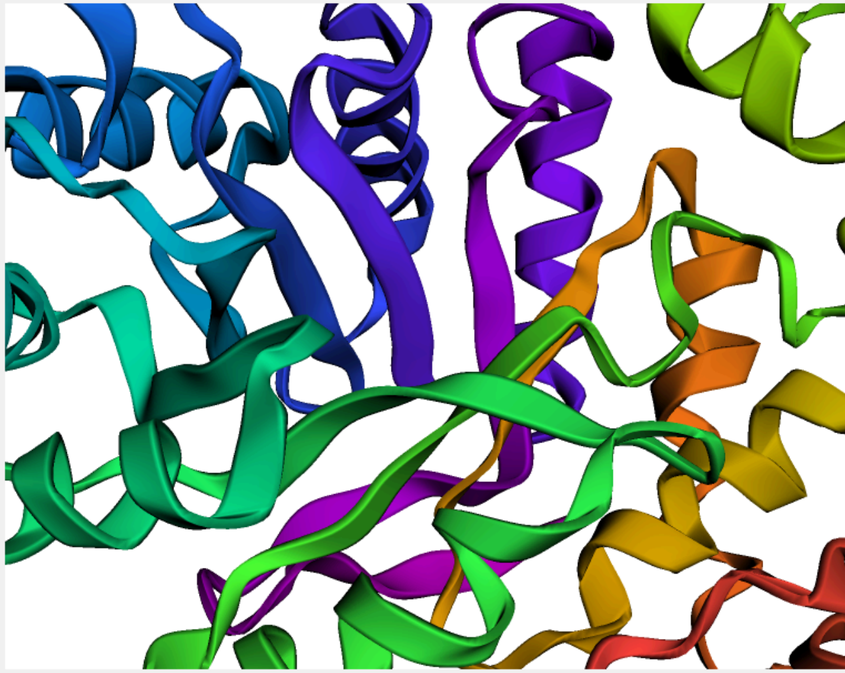
Protein Structure Prediction and Analysis Application

Input sequence

QFYRNLGKSGLRVSCLGLGTWVTFGGQI
TDEMAEHLMTLAYDNGINLFDTAEVYAA
GKAEVVLGNIIKKKGWRRSSLVITTKIFW
GGKAETERGLSRKHIIEGLKASLERLQLE
YVDVVFANRPDPNTPMEETVRAMTHVI
NQGMAMYWGTSRWSSMEIMEAYSVAR
QFNLIPPICEQAEYHMFQREKVEVQLPE
LFHKIGVGAMTWSPLACGIVSGKYDSGI
PPYSRASLKGYQWLKDKILSEEGRRQQA
KLKELQAIAERLGCTLPQLAIAWCLRNE
GVSSVLLGASNAEQLMENIGAIQVLPKL
SSSIVHEIDSILGNKPYS

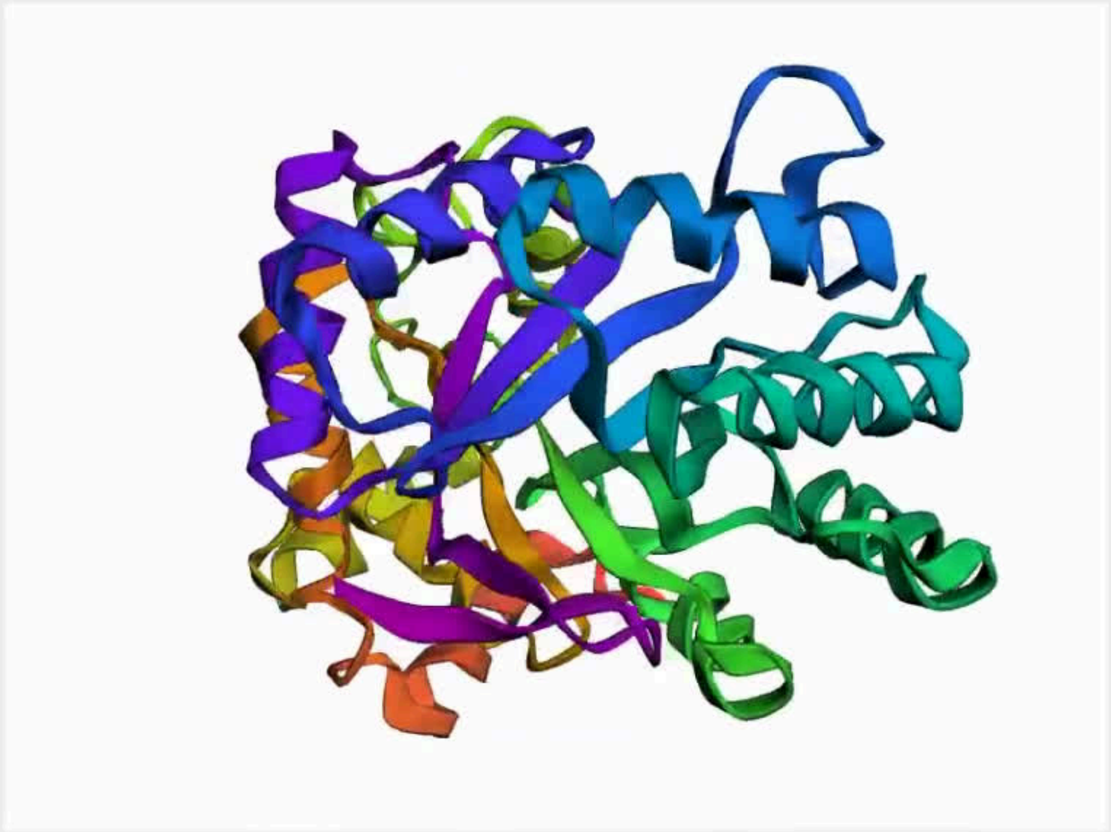
Predict

Visualization of predicted protein structure



# Amyloid-Beta Protein in Alzheimer's Disease Visualization

Visualization of predicted protein structure



# Amyloid-Beta Protein in Alzheimer's Disease Prediction

## plDDT

plDDT is a per-residue estimate of the confidence in prediction on a scale from 0-100.

plDDT: 0.8896

## GRAVY Score

GRAVY is a measure of the overall hydrophobicity or hydrophilicity of the protein sequence.

GRAVY Score: -0.1549

# Amyloid-Beta Protein in Alzheimer's Disease Prediction, Continue...

- **Identification of binding sites is essential for drug discovery.**
- **Binding sites are important in biochemistry and molecular biology, as they play a crucial role in regulating the interactions between molecules and their functions.**



## Predicted Binding Sites

Predicted binding sites based on ligands or co-factors in the structure.

Chain: A, Residue: (' ', 1, ' ')

Chain: A, Residue: (' ', 2, ' ')

Chain: A, Residue: (' ', 3, ' ')

Chain: A, Residue: (' ', 4, ' ')

Chain: A, Residue: (' ', 5, ' ')

Chain: A, Residue: (' ', 6, ' ')

Chain: A, Residue: (' ', 325, ' ')

# Amyloid-Beta Protein in Alzheimer's Disease Analysis

- The isoelectric point refers to the pH at which a molecule or a substance is electrically neutral.
- The closer the pH is to the isoelectric point, the less charged the protein is.

## Protein Properties

Basic properties of the protein sequence.

Amino Acid Composition: {'A': 0.07692307692307693, 'C': 0.015384615384615385, 'D': 0.024615384615384615, 'E': 0.08, 'F': 0.024615384615384615, 'G': 0.08615384615384615, 'H': 0.018461538461538463, 'I': 0.07076923076923076, 'K': 0.05846153846153846, 'L': 0.10153846153846154, 'M': 0.033846153846153845, 'N': 0.036923076923076927, 'P': 0.036923076923076927, 'Q': 0.046153846153846156, 'R': 0.052307692307692305, 'S': 0.07384615384615385, 'T': 0.043076923076923075, 'V': 0.06153846153846154, 'W': 0.024615384615384615, 'Y': 0.033846153846153845}

Molecular Weight: 36285.4811

Isoelectric Point: 8.218316841125489

Secondary Structure Fraction: (0.3507692307692308, 0.25846153846153846, 0.36)

# Amyloid-Beta Protein in Alzheimer's Disease Analysis, Continue...

- **Substrate specificity of a protein refers to its ability to selectively recognize and interact with specific amino acids or sequences of amino acids.**
- **Purity refers to the degree to which the isolated protein sample is free from contaminants.**
- **Yield represents the efficiency of the purification process in retaining the target protein.**

## Substrate Specificity

Predicted substrate specificity based on amino acid composition in binding sites

Substrate Specificity: G, T, A, O, Y, M, L, H, E, R, S, U, I, C, P, N, V

## Protein Purification Results

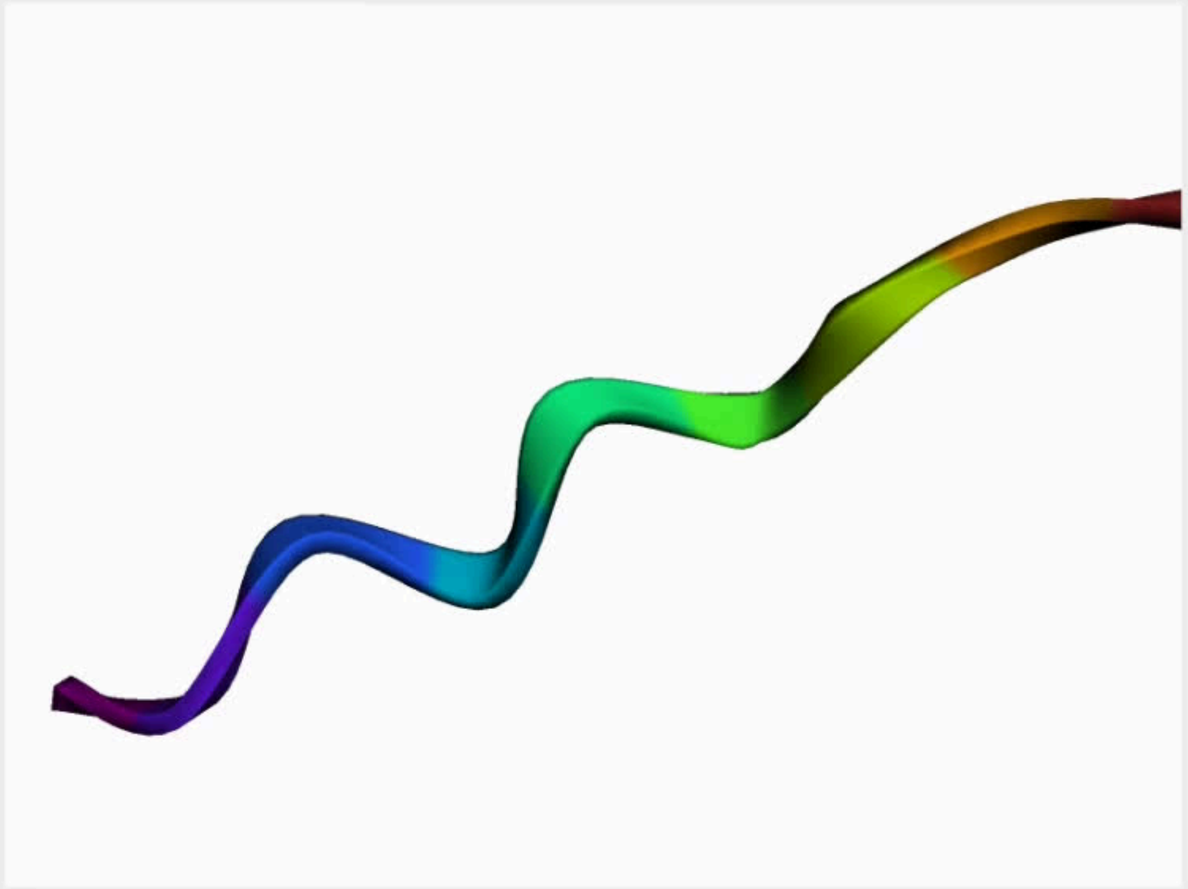
Purity Level: 95.0%

Yield Percentage: 80.0%

Download PDB

# Tau Protein in Alzheimer's Disease Visualization

Visualization of predicted protein structure



# Tau Protein in Alzheimer's Disease Prediction

## plDDT

plDDT is a per-residue estimate of the confidence in prediction on a scale from 0-100.

plDDT: 0.7224

## GRAVY Score

GRAVY is a measure of the overall hydrophobicity or hydrophilicity of the protein sequence.

GRAVY Score: -0.1889

# Tau Protein in Alzheimer's Disease Analysis

## Protein Properties

Basic properties of the protein sequence.

Amino Acid Composition: {'A': 0.0, 'C': 0.0, 'D': 0.0, 'E': 0.0, 'F': 0.0, 'G': 0.0, 'H': 0.0, 'I': 0.2222222222222222, 'K': 0.3333333333333333, 'L': 0.1111111111111111, 'M': 0.0, 'N': 0.1111111111111111, 'P': 0.0, 'Q': 0.1111111111111111, 'R': 0.0, 'S': 0.0, 'T': 0.0, 'V': 0.1111111111111111, 'W': 0.0, 'Y': 0.0}

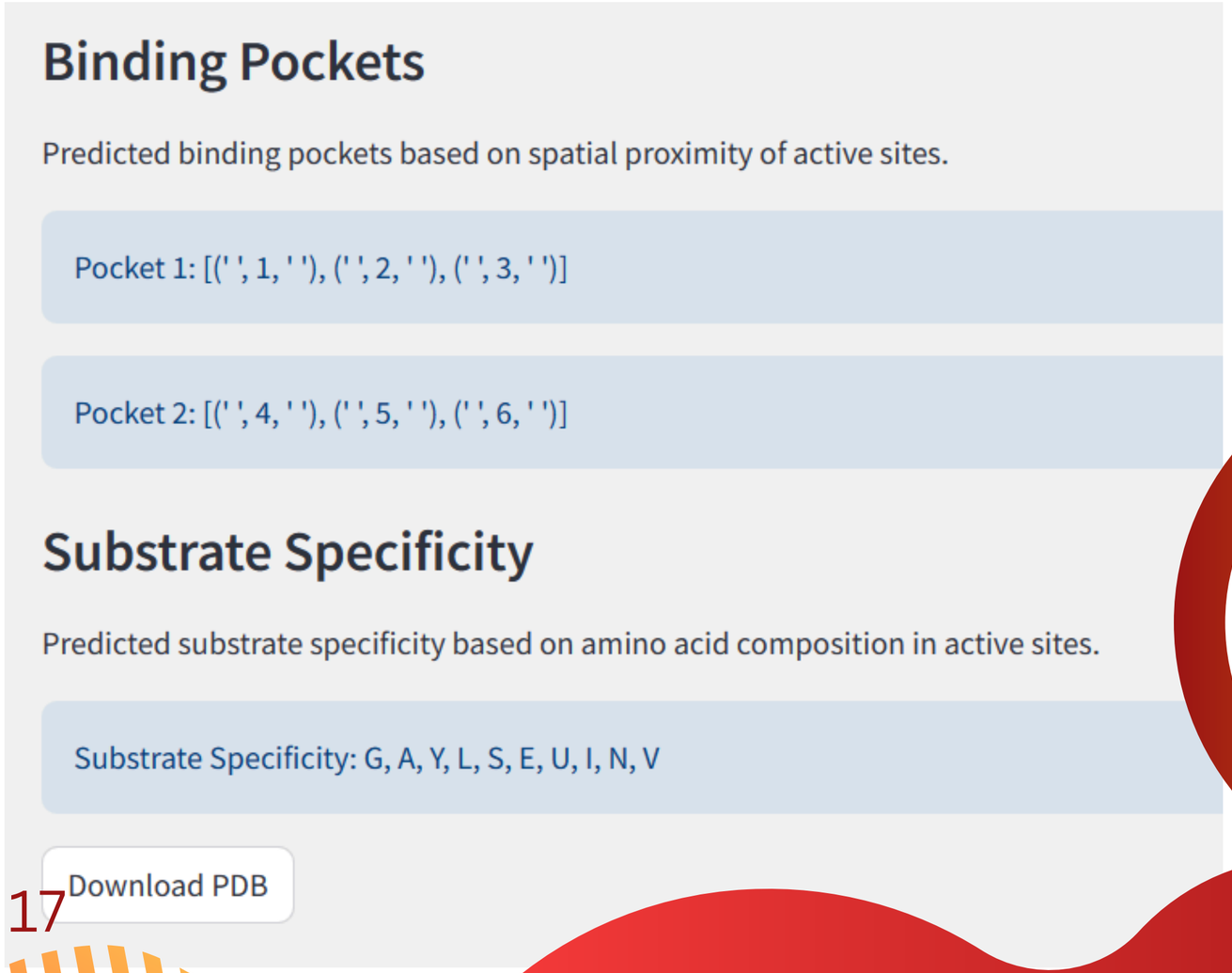
Molecular Weight: 1083.3678

Isoelectric Point: 10.302063941955566

Secondary Structure Fraction: (0.4444444444444444, 0.1111111111111111, 0.4444444444444444)

# Tau  Protein in Alzheimer's Disease Analysis, Continue...

- A binding pocket is a specific, three-dimensional crevice or cavity on the surface of a biomolecule where a ligand can bind.
- Ligands can be small molecules, ions, or other proteins, fit into the binding pocket with a degree of specificity.



# PDB File Overview

## Aβ Protein

```
ATOM      1  N   GLN A   1     -15.170  -9.997   2.354  1.00  0.80
ATOM      2  CA  GLN A   1     -13.877  -9.605   2.904  1.00  0.83
ATOM      3  C   GLN A   1     -13.282 -10.720   3.760  1.00  0.83
ATOM      4  CB  GLN A   1     -14.009  -8.324   3.729  1.00  0.76
ATOM      5  O   GLN A   1     -13.997 -11.372   4.524  1.00  0.78
ATOM      6  CG  GLN A   1     -12.696  -7.841   4.330  1.00  0.67
ATOM      7  CD  GLN A   1     -11.697  -7.399   3.277  1.00  0.63
ATOM      8  NE2 GLN A   1     -10.525  -6.958   3.721  1.00  0.55
ATOM      9  OE1 GLN A   1     -11.975  -7.454   2.075  1.00  0.67
ATOM     10  N   PHE A   2     -12.012 -10.978   3.656  1.00  0.86
ATOM     11  CA  PHE A   2     -11.324 -11.921   4.529  1.00  0.89
ATOM     12  C   PHE A   2     -10.085 -11.283   5.146  1.00  0.89
ATOM     13  CB  PHE A   2     -10.934 -13.185   3.757  1.00  0.84
ATOM     14  O   PHE A   2      -9.682 -10.188   4.750  1.00  0.87
ATOM     15  CG  PHE A   2      -9.992 -12.930   2.612  1.00  0.80
ATOM     16  CD1 PHE A   2     -10.475 -12.524   1.374  1.00  0.75
ATOM     17  CD2 PHE A   2      -8.623 -13.098   2.773  1.00  0.78
ATOM     18  CE1 PHE A   2      -9.606 -12.287   0.312  1.00  0.73
ATOM     19  CE2 PHE A   2      -7.748 -12.864   1.716  1.00  0.72
ATOM     20  CZ  PHE A   2      -8.241 -12.459   0.486  1.00  0.73
ATOM     21  N   TYR A   3      -9.603 -11.919   6.118  1.00  0.93
ATOM     22  CA  TYR A   3      -8.492 -11.390   6.901  1.00  0.93
ATOM     23  C   TYR A   3      -7.262 -12.281   6.774  1.00  0.92
ATOM     24  CB  TYR A   3      -8.889 -11.256   8.375  1.00  0.92
ATOM     25  O   TYR A   3      -7.382 -13.496   6.603  1.00  0.91
ATOM     26  CG  TYR A   3      -9.946 -10.208   8.626  1.00  0.89
```

## Tau Protein

```
N ATOM      1  N   LYS A   1      -4.552  -8.093 -11.786  1.00  0.78           N
C ATOM      2  CA  LYS A   1      -3.513  -7.317 -11.116  1.00  0.78           C
C ATOM      3  C   LYS A   1      -3.968  -6.871  -9.729  1.00  0.79           C
C ATOM      4  CB  LYS A   1      -3.125  -6.099 -11.957  1.00  0.73           C
O ATOM      5  O   LYS A   1      -4.911  -6.087  -9.602  1.00  0.73           O
C ATOM      6  CG  LYS A   1      -2.198  -6.420 -13.120  1.00  0.67           C
C ATOM      7  CD  LYS A   1      -1.823  -5.166 -13.897  1.00  0.63           C
N ATOM      8  CE  LYS A   1      -0.984  -5.497 -15.124  1.00  0.60           C
O ATOM      9  NZ  LYS A   1      -0.721  -4.288 -15.960  1.00  0.51           N
N ATOM     10  N   VAL A   2      -3.935  -7.500  -8.682  1.00  0.82           N
C ATOM     11  CA  VAL A   2      -4.365  -7.363  -7.294  1.00  0.81           C
C ATOM     12  C   VAL A   2      -3.464  -6.367  -6.569  1.00  0.81           C
C ATOM     13  CB  VAL A   2      -4.355  -8.723  -6.561  1.00  0.78           C
O ATOM     14  O   VAL A   2      -2.242  -6.392  -6.732  1.00  0.75           O
C ATOM     15  CG1 VAL A   2      -4.762  -8.551  -5.098  1.00  0.66           C
C ATOM     16  CG2 VAL A   2      -5.282  -9.714  -7.263  1.00  0.65           C
N ATOM     17  N   GLN A   3      -4.023  -5.304  -6.230  1.00  0.81           N
C ATOM     18  CA  GLN A   3      -3.415  -4.191  -5.509  1.00  0.80           C
C ATOM     19  C   GLN A   3      -3.448  -4.428  -4.002  1.00  0.79           C
C ATOM     20  CB  GLN A   3      -4.123  -2.879  -5.850  1.00  0.77           C
C ATOM     21  O   GLN A   3      -4.470  -4.848  -3.456  1.00  0.75           O
N ATOM     22  CG  GLN A   3      -3.576  -2.190  -7.093  1.00  0.72           C
C ATOM     23  CD  GLN A   3      -4.294  -0.891  -7.407  1.00  0.70           C
C ATOM     24  NE2 GLN A   3      -3.780  -0.150  -8.383  1.00  0.63           N
C ATOM     25  OE1 GLN A   3      -5.303  -0.556  -6.779  1.00  0.66           O
O ATOM     26  N   ILE A   4      -2.410  -4.766  -3.430  1.00  0.81           N
C ATOM     27  CA  ILE A   4      -2.174  -4.931  -2.000  1.00  0.81           C
```

# Conclusion

- Evolutionary Scale Modeling has shown great potential in predicting protein folding and analyzing proteins in the context of Alzheimer's disease.

- There is still much future work to be done in fully understanding the underlying mechanisms and developing new applications.

# Future Work

- Pre-drug design analysis.
- Drug design for amlyloid -beta and tau proteins in Alzheimer's Disease.

# Refrences


- Carpenter, A. (2022, January 1). Visualizing and Analyzing Proteins in Python - Towards Data Science. Medium. Retrieved December 6, 2023, from https://towardsdatascience.com/visualizing-and-analyzing-proteins-in-python-bd99521ccd.

- Bank, R. P. D. (n.d.). RCSB PDB - 6NK4: KVQIINKKL, Crystal Structure of a Tau Protein Fragment. Retrieved December 6, 2023, from https://www.rcsb.org/structure/6NK4.

- Prospr. PyPI. (n.d.). https://pypi.org/project/prospr/.

- ESM Metagenomic Atlas: The first view of the “dark matter” of the protein universe. AI at Meta. (n.d.). https://ai.meta.com/blog/protein-folding-esmfold-metagenomics/.

- UniProt. (n.d.). https://www.uniprot.org/uniprotkb/Q54A46/entry.

- Dataprofessor. (n.d.). Dataprofessor/Esmfold. GitHub. https://github.com/dataprofessor/esmfold/.

# Any questions?

# Thank You