Predicting the Structure of Synaptophysin and Microtubule-Associated Protein Tau

DD2402 Advanced Individual Course in Computational Biology

Joakim Abdinur Iusuf

ABSTRACT

This project uses ColabFold, a tool that uses AlphaFold for its predictions, to predict the 3D structures of synaptophysin and tau protein across humans, bovines and mice. Using ChimeraX for analysis, these predictions have been evaluated through root-mean-square deviation (RMSD) and predicted local distance difference test (pLDDT) scores. The findings indicate high-confidence predictions for synaptophysin across humans, bovines, and mice. Findings also show that the structure of synaptophysin across humans, bovines, and mice is largely similar. Conversely, tau protein, known for being an inherently disordered protein, shows predominantly low confidence in structure predictions and its structure varies across species. Future work could delve deeper into the lower-confidence regions, such as the yellow-colored alpha helix in synaptophysin predictions, and perform detailed analyses on specific tau protein segments instead of looking at the entire protein.

Keywords: ColabFold, AlphaFold, Synaptophysin, Tau

1 INTRODUCTION

The purpose of this project is to deepen my understanding of AlphaFold [21], which is one of the biggest breakthroughs in machine learning. To learn more about AlphaFold, I will predict the 3D structure of the synaptic protein synaptophysin [40] and the neurodegenerative disease-associated protein microtubule-associated protein tau [11] using ColabFold [25] [21] [26] [27] [28], a tool that uses AlphaFold for its predictions.

I first introduce AlphaFold, a deep learning-based system that is designed to predict protein structures with high accuracy [21]. This is followed by a brief overview of protein structure, as well as an overview of two specific proteins: synaptophysin and the microtubule-associated protein tau. In the *Methods and Materials* section, I present the use of ColabFold and ChimeraX [35], along with the metrics and procedure used in this project. Finally, I present and discuss the results.

1.1 AlphaFold, Proteins, and Protein Structure Prediction

In the paper *Highly accurate protein structure prediction with AlphaFold* [21], researchers at Google DeepMind [6] present AlphaFold. AlphaFold is a deep learning-based system that is designed to predict the structure of proteins with a high level of accuracy. Proteins, which are essential to life and support practically all biological functions, are large and complex molecules composed of amino acid chains. Determining the 3D shape of a protein is important because a protein's unique 3D structure determines its function, and many of the worlds greatest challenges, such as developing treatments for diseases, are tied to proteins and the role they play. Determining the structure of proteins, therefore, allows for a greater understanding of what they do and how they work, which is crucial for solving many problems [32].

Traditionally, experimental methods like X-ray crystallography [12] and cryo-electron microscopy [24] have been used to study these structures. As the AlphaFold team explains [32], the issue with these methods is that they often require years and costly equipment to determine the structure of one single protein. The development of AlphaFold is significant because it provides a solution to a longstanding problem in biology, known as the protein folding problem, which involves predicting the 3D shape of a protein based solely on the 1D sequence of amino acids that make it up. The protein folding problem was first suggested by Nobel laureate Christian Anfinsen in 1972 [14], where he postulated that a protein's amino acid sequence fully dictates its structure. This problem has, for more than five decades, driven research into predicting protein structures computationally as a faster and cheaper alternative to experimental methods [32].

AlphaFold solves this problem, and can accurately predict protein structures with atomic accuracy and is competitive with experimental methods and also greatly outperforms other methods [21]. AlphaFold was validated through its performance in the 14th Critical Assessment of protein Structure Prediction (CASP14) [3] competition and was entered under the name ”AlphaFold2” (an improved model from the CASP13 AlphaFold system [21]). CASP is a competition that is carried out biennially and serves as the gold-standard assessment for the accuracy of structure prediction [29] [22].

AlphaFold produces predictions in minutes to hours depending on the length of a protein sequence [21]. In 2021, DeepMind applied AlphaFold to the entire human proteome [33]. Before AlphaFold, there were about 190 000 protein structures stored in the Protein Data Bank [10] [20]. Google DeepMind in partnership with EMBL's European Bioinformatics Institute [7] created a new database called the AlphaFold Protein Structure Database [4]. It hosts over 200 million structures, including structures for plants, animals, bacteria, and other organisms [20]. More than half a million researchers have used the AlphaFold database to look at over 2 million structures, and it has opened up opportunities for researchers to tackle a wide range of issues, such as issues related to sustainability, food insecurity, and neglected diseases [20]. Before giving an overview of synaptophysin and tau, a brief overview of the structure of a protein is given as this is essential for interpreting and understanding the results of this project effectively.

1.2 Protein Structure

As explained in [2], the structure of a protein is fundamental to its function and is determined at several levels of complexity. At its core, the primary structure of a protein is its amino acid sequence. This sequence is crucial because it guides the folding and intramolecular bonding of the protein, shaping its unique 3D shape. As the linear chain of amino acids folds, hydrogen bonds form between amino groups and carboxyl groups in neighboring regions. This interaction leads to the creation of alpha helices and beta sheets, the stable patterns that constitute a protein's secondary structure. Most proteins feature a combination of these helices and sheets, alongside other less common folding patterns. In the context of this project, the primary structure of the proteins whose structure we predict (the amino acid sequences) is important as well as their secondary structure, since the visual representation of the protein structures presented in section 3 are made up of alpha helices and beta sheets.

1.3 Synaptophysin - an overview

Synaptophysin, which is also known as the major synaptic vesicle protein p38, is encoded by the SYP gene in humans. It is situated on the short arm of the X chromosome and results in a protein of 313 amino acids [40] [36]. Synaptophysin is present in a wide array of neuroendocrine cells and neurons within the brain and spinal cord. It plays a role in synaptic transmission and acts as a marker for neuroendocrine tumors. Research has shown that deactivating its gene in animals affects behaviors like exploratory activity, leads to challenges in object novelty detection, and reduces spatial learning capabilities. Synaptophysin has also been linked to X-linked intellectual disability [40].

As mentioned, synaptohysin is a marker for neuroendocrine tumors, but it is also a marker in neuronal tumors, including some of the most basic forms like medulloblastoma. However, since synaptophysin can be detected in various tumors, it is important to use multiple markets in tumor identification and not rely solely on synaptophysin [30].

The precise function of synaptophysin is unknown [40], but it potentially plays a role in organizing membrane components or directing vesicles towards the plasma membrane, contributing to the regulation of both short-term and long-term synaptic plasticity [36].

1.4 Microtubule-Associated Protein Tau - an overview

Tau plays a crucial role in stabilizing microtubules in axons. Tau is predominantly found in the neurons of the central nervous system (CNS), especially within the cerebral cortex, and to a lesser extent in astrocytes and oligodendrocytes of the CNS. Tau was identified in 1975 for it's critical function in microtubule assembly, and is known as an intrinsically disordered protein. Tau has been linked to neurological pathologies and dementias, including Alzheimer's and Parkinson's diseases, where it forms hyperphosphorylated, insoluble aggregates known as neurofibrillary tangles [11].

In relation to Alzheimer's disease, the tau hypothesis suggests that when tau protein gets abnormally phosphorylated, it changes from its normal state into something that can cause harm in the brain. This process involves the tau protein clustering into neurofibrillary tangles, which are closely linked to the progression of Alzheimer's. These tangles accumulate inside neurons, disrupting their function and ultimately leading to cell death. The presence of hyperphosphorylated tau, identified in these tangles, is a hallmark of Alzheimer's, indicating a breakdown in the cell's internal transport system. Recent research has pointed towards the potential of a blood test for a specific form of tau, known as p-tau-217, to diagnose Alzheimer's decades before the symptoms of dementia appear, which could lead to early detection and management of the disease [11]. Similarly, a researcher associated with KTH Royal Institute of Technology has developed an medical technology that holds promise for the early detection and management of Alzheimer's disease. It uses geometric and cloud-based AI modelling, in conjunction with factors such as cognitive ability, age and gender to predict the brain's condition five years into the future with a high degree of accuracy [13].

Tau is also associated with traumatic brain injury. Recurring minor traumatic brain injuries, commonly seen in contact sports like American football and from military explosions, can result in chronic traumatic encephalopathy (CTE). This condition is marked by abnormal tau protein accumulations in the brain. Furthermore, following a severe traumatic brain injury, elevated tau protein levels in the brain's extracellular fluid are associated with worse health outcomes [11]. The genetic background of tau, particularly the microtubule associated protein tau haplotype, could also play a role in CTE. There is research that explores how genetic variations might influence the development of CTE [23].

2 METHODS AND MATERIALS

2.1 ColabFold

In this project I have used ColabFold [25] [21] [26] [27] [28] as an alternative to the full AlphaFold [18] system since the full system requires up to 3 TB of disk space and extensive computational resources for running predictions. As explained in [25], to use the full AlphaFold system, researchers need to build diverse multiple sequence alignments (MSAs) [9] which involves searching massive public reference and environmental databases with sensitive detection tools like HMMer [19] and HHblits [31] that use profile hidden Markov models. Searching for a single protein, critical for identifying homologous sequences, can take several hours and requires more than 2 TB of storage space. Moreover, executing deep neural networks for structure prediction with AlphaFold demands high-performance GPUs with extensive RAM [25].

ColabFold is a free and fast alternative to the full AlphaFold system for predicting protein structures, and is accessible via Google Colaboratory [5]. It speeds up the prediction of single proteins by replacing AlphaFold's homology search with MMseqs2 (Many-against-Many sequence searching) [26]. MMseqs2 is a software suite designed for fast and efficient sequence searching and alignment. It enables the rapid comparison of protein sequences by utilizing many-against-many sequence searching techniques, and makes the process of identifying homologous sequences and building MSAs 40-60 times faster [25].

ColabFold has been shown to match the full AlphaFold system on CASP14 targets, and is composed of three key components: an MMseqs2-based server for doing homology searches and building MSAs; a Python library that interfaces with the MMseqs2 server, prepares input for structure prediction by AlphaFold's deep learning models, and visualizes outcomes; and Jupyter notebooks that offer templates for basic, advanced, and batch processing tasks through the Python library [25].

2.2 ChimeraX

I have used ChimeraX [35] to compare protein structures. ChimeraX is a next-generation molecular visualization software developed by the Resource for Biocomputing, Visualization, and Informatics (RBVI) at the University of California, San Francisco. It enables researchers to visualize and analyze molecular structures, sequences, and associated data. One of the many features of ChimeraX is that it is able to compare protein structures using RMSD [34] [17] as a metric, which is explained in section 2.3.

To compare the structures I have used the MatchMaker tool [1] in ChimeraX. The command used is *mm #2 to #1 showAlignment true*. MatchMaker initiates the process of superimposing proteins by creating a sequence alignment that takes residue types into consideration and also aims to align helices with helices and strands with strands. It then aligns the sequence-matched residues in three dimensions by pairing their C-alpha atoms. The option 'showAlignment true' enables the display of this sequence alignment. By default, the alignment excludes pairs that are significantly distant from each other and ensures that the most closely matching segments align more precisely. The final matching statistics are detailed in the Log, and the pairs involved in the ultimate alignment are indicated by light orange boxes on the sequence alignment. Clicking on these boxes highlights the segments of the proteins that were used in the final alignment.

2.3 Metric 1 - Root-Mean-Square Deviation

Structural biologists typically measure the similarity between two protein structures by superimposing them optimally and then computing the Root-Mean-Square Deviation (RMSD) [17] of the C-alpha atoms from corresponding residues [8]. For this reason, I have used RMSD as a metric for comparing protein structures. RMSD serves as a quantitative measure to assess the similarity between two or more protein structures by calculating the average distance between the atoms of superimposed proteins. This method is widely recognized in the field, notably utilized by the Critical Assessment of Structure Prediction (CASP) competition to evaluate the accuracy of protein structure predictions. The lower the RMSD value, the closer a modeled structure is to the target structure, indicating a higher level of accuracy in the prediction. Using RMSD allows for an objective comparison of the protein structures predicted with ColabFold against known structures, providing a clear, numerical indication of prediction quality and similarity to the target [17].

2.4 Metric 2 - Predicted Local Distance Difference Test

AlphaFold uses a per-residue confidence score, known as predicted local distance difference test (pLDDT), which ranges from 0 to 100, to assess the reliability of the predicted protein structures [36] [8]. The pLDDT scores help with distinguishing the confidence levels of various regions within the protein structure:

- 🟦 **Very High Confidence**: Regions with pLDDT scores above 90.
- 🟦 **Confident**: Scores between 70 and 90.
- 🟨 **Low Confidence**: Scores between 50 and 70.
- 🟧 **Very Low Confidence**: Scores below 50.

Regions with low confidence scores may be disordered, and these regions often have a ribbon-like appearance [8]. This metric was used to evaluate the quality of the predicted structures, and to see if some regions yielded higher confidence scores than others.

2.5 Procedure

I started by navigating to the UniProt website and searching for ”Synaptophysin.” I then went to the entry for SYPH_HUMAN, and clicked the ”go to sequence” link located in the upper right section of the page. Following this, I downloaded the amino acid sequence and copied it. This sequence was then inputted into ColabFold in the ”query_sequence” field, after which I executed the prediction by selecting ”runtime” and then ”run all.” The process to predict the structure of human synaptophysin took approximately 21 minutes, and a zip file was automatically downloaded, which I subsequently extracted. The same procedure was applied to both SYPH_BOVINE and SYPH_MOUSE, with each prediction requiring around 14 minutes. The same procedure was used for TAU_HUMAN, TAU_BOVINE, and TAU_MOUSE, but with the search term changed to ”Tau.” The predictions varied in duration: 41 minutes for human tau, 19 minutes for bovine tau, and 50 minutes for mouse tau.

Each prediction generated a zip file containing five distinct structural models, ranked from 1 to 5. To conduct a comparative analysis of two structures, I accessed them in ChimeraX by selecting the .pdb files associated with the model of rank 1. With both structures loaded in ChimeraX, I executed the command *mm #2 to #1 showAlignment true* to align and compare them. For additional details on this process, please refer to section 2.2.

I compared the predicted structure of synaptophysin obtained with ColabFold with a 3D structure predicted with AlphaFold that can be found on UniProt [36]. The main reason for this is that there is no experimental structure for synaptophysin, so I compared it with the structure predicted by AlphaFold. I compared the structure of human synaptophysin with its counterparts in bovines [15] and mice [16] to explore the differences between the species. The structures for bovine and mouse synaptophysin were also predicted using ColabFold.

I compared the tau protein structure predicted using ColabFold with an experimental structure available on UniProt [37], which has been determined using the complete amino acid sequence of tau. While there are other experimental structures available [37], many have not used the full amino acid sequence for determining tau’s structure. Given that my prediction with ColabFold used the entire amino acid sequence, this specific experimental structure from UniProt seemed like a good structure to compare against. I also compared it with the structure of human tau found on Uniprot [37] that was predicted using AlphaFold. Moreover, I compared the structure of human tau with its counterparts in bovines [39] and mice [38], which again were also predicted using ColabFold.

After completing the comparisons, I assessed the pLDDT scores for each ColabFold-predicted structure. I loaded each structure individually into ChimeraX, and then ran the command *color bfactor #1 palette alphafold*.

3 RESULTS

3.1 Synaptophysin

The RMSD between the aligned structures in Figure 1 is 0.261 Å for 220 pruned atom pairs. The alignment utilized amino acids 14 to 233 in the sequence for the calculation. Across all 313 pairs, the calculation gave a RMSD of 14.721 Å.

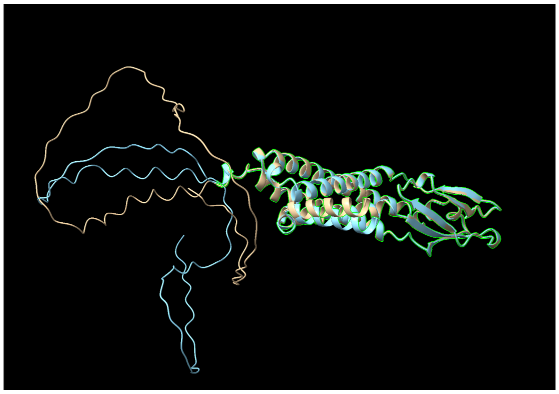

Figure 1. Structural comparison of human synaptophysin predicted by ColabFold versus the AlphaFold model retrieved from UniProt. The green highlights show the segments of the proteins that were aligned using the MatchMaker tool in ChimeraX, highlighting the regions used in the final structural alignment.

The RMSD value between the human and bovine synaptophysin structures in Figure 2 is 0.310 Å for 226 pruned atom pairs, specifically including residues 9-10, 13-14, 16-88, 90-234, 238, 254-255, and 309. The RMSD calculation across all 313 pairs resulted in a value of 8.177 Å.

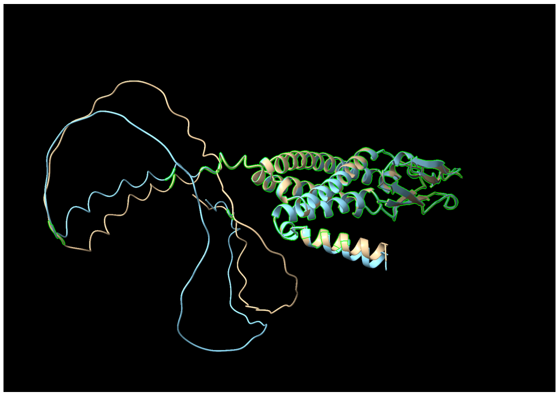

Figure 2. Structural comparison of human synaptophysin predicted by ColabFold with its bovine counterpart also predicted using ColabFold. The green highlights show the segments of the proteins that were aligned using the MatchMaker tool in ChimeraX, highlighting the regions used in the final structural alignment.

The RMSD value between the human and mouse synaptophysin structures in Figure 3 is 0.292 Å for 214 pruned atom pairs, specifically including residues 21-22, 24-87, 90-127, 129-236, and 238-239. The RMSD calculation across all 313 pairs resulted in a value of 8.401 Å.

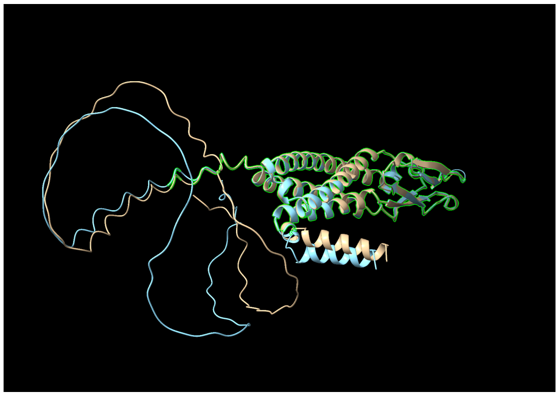

Figure 3. Structural comparison of human synaptophysin predicted by ColabFold with its mouse counterpart also predicted using ColabFold. The green highlights indicate the segments of the proteins that were aligned using the MatchMaker tool in ChimeraX, highlighting the regions used in the final structural alignment.

In the predicted structure in Figure 4, most of the alpha-helices and beta-strands are colored in blue, suggesting high confidence in these regions. Notably, one alpha-helix is highlighted in yellow, indicating lower confidence in its conformation. One part of the protein shows orange coloring, reflecting very low confidence scores. This part of the protein has a ribbon-like appearance and could potentially be disordered, as mentioned in section 2.4.

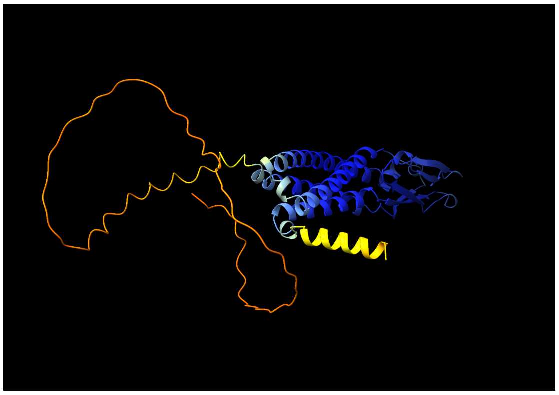

Figure 4. Predicted structure of human synaptophysin using ColabFold. The color coding indicates the per-residue confidence score (pLDDT) with dark blue representing very high confidence regions (pLDDT > 90), light blue indicating confident regions (pLDDT 70-90), yellow for regions of low confidence (pLDDT 50-70), and orange denoting very low confidence regions (pLDDT < 50).

Just like the predicted structure of human synaptophysin, most of the alpha-helices and beta-strands in Figure 5 are colored in blue, suggesting high confidence in these regions. For the predicted structure of bovine Synaptophysin, one alpha-helix is also highlighted in yellow, indicating lower confidence in its conformation. Moreover, one part of the protein also shows orange coloring, reflecting very low confidence scores.

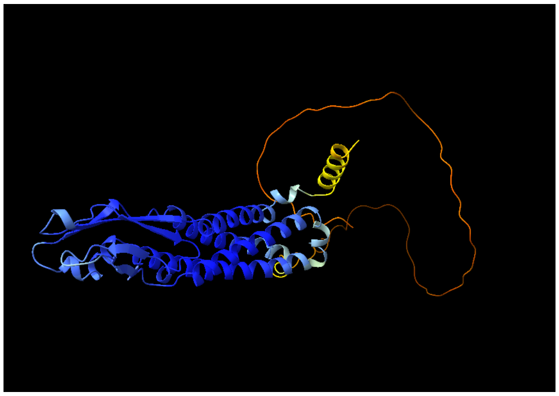

Figure 5. Predicted structure of bovine synaptophysin using ColabFold. The color coding indicates the per-residue confidence score (pLDDT) with dark blue representing very high confidence regions (pLDDT $>$ 90), light blue indicating confident regions (pLDDT 70-90), yellow for regions of low confidence (pLDDT 50-70), and orange denoting very low confidence regions (pLDDT $<$ 50).

In the predicted structure of mouse synaptophysin, seen in Figure 6, most of the alpha-helices and beta-strands are also colored in blue, suggesting high confidence in these regions. We also observe one alpha-helix that is highlighted in yellow, indicating lower confidence in its conformation, and one part of the protein that shows orange coloring, reflecting very low confidence scores in this region. The pLDDT scores for human, bovine, and mouse Synaptophysin seem to be similar. One notable difference is that we see some more yellow highlighting in mouse Synaptophysin in regions close to the alpha-helix with a low confidence score.

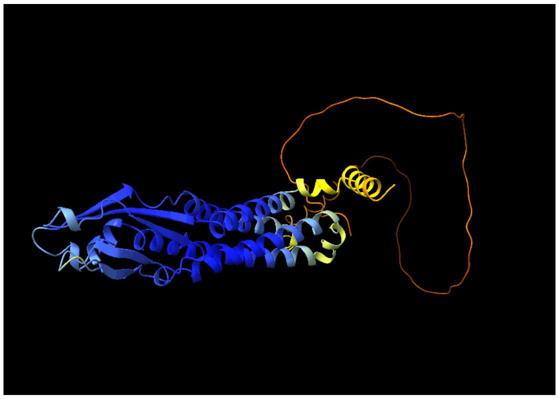

Figure 6. Predicted structure of mouse synaptophysin using ColabFold. The color coding indicates the per-residue confidence score (pLDDT) with dark blue representing very high confidence regions (pLDDT $>$ 90), light blue indicating confident regions (pLDDT 70-90), yellow for regions of low confidence (pLDDT 50-70), and orange denoting very low confidence regions (pLDDT $<$ 50).

3.2 Microtubule-Associated Protein Tau

The RMSD between the human tau predicted using ColabFold and the experimental model retreived from UniProt, found in Figure 7, is 0.998 Å for 5 pruned atom pairs. The alignment utilized amino acids 635 to 639 in the sequence for the calculation. Across all 77 pairs, the calculation gave a RMSD of 46.370 Å.

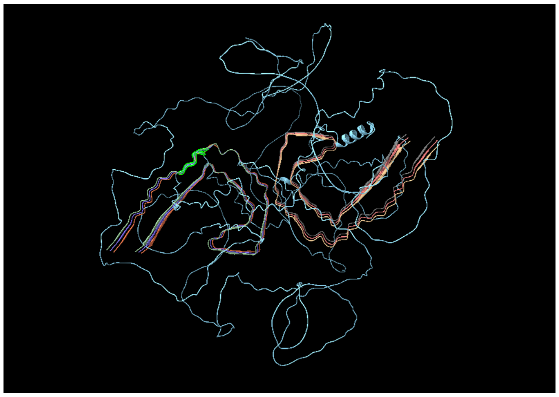

Figure 7. Structural comparison of human tau predicted by ColabFold with the experimental model retrieved from UniProt. The green highlights indicate the segments of the proteins that were aligned using the MatchMaker tool in ChimeraX, highlighting the regions used in the final structural alignment.

The RMSD between the aligned structures in Figure 8 is 1.236 Å for 12 pruned atom pairs. The alignment utilized amino acids 501 to 506 and 578 to 583 in the sequence for the calculation. Across all 758 pairs, the calculation gave a RMSD of 58.745 Å.

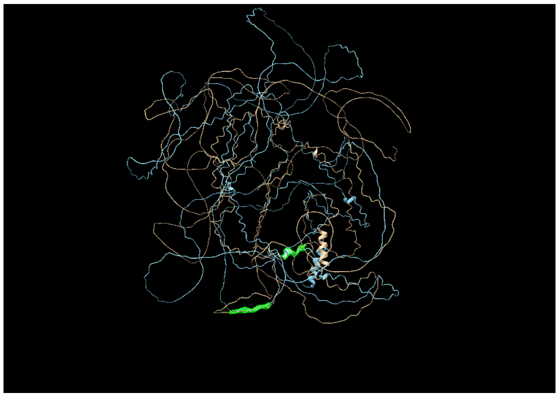

Figure 8. Structural comparison of human tau predicted by ColabFold versus the AlphaFold model retrieved from UniProt. The green highlights indicate the segments of the proteins that were aligned using the MatchMaker tool in ChimeraX, highlighting the regions used in the final structural alignment.

The RMSD value between the human and bovine tau structures in Figure 9 is 1.072 Å for 33 pruned atom pairs. This aligment includes residues 588-589, 591-608, and 610-622. The RMSD calculation across all 430 pairs resulted in a value of 34.521 Å.

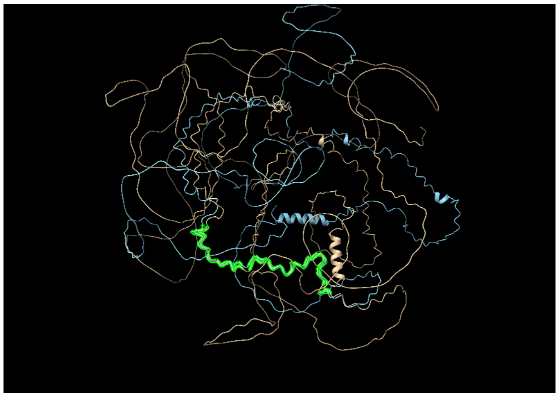

Figure 9. Structural comparison of human tau predicted by ColabFold with its bovine counterpart. The green highlights indicate the segments of the proteins that were aligned using the MatchMaker tool in ChimeraX, highlighting the regions used in the final structural alignment.

The RMSD value between the human and mouse tau structures in Figure 10 is 1.107 Å for 50 pruned atom pairs. The used residues were 586-587, 590-591, 593-602, and 643-678. The RMSD calculation across all 724 pairs resulted in a value of 51.933 Å.

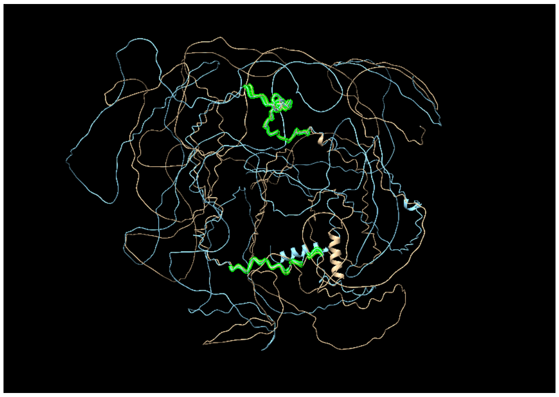

Figure 10. Structural comparison of human tau predicted by ColabFold with its mouse counterpart. The green highlights indicate the segments of the proteins that were aligned using the MatchMaker tool in ChimeraX, highlighting the regions used in the final structural alignment.

The color coding in Figure 11 suggests a low confidence level in the structure of human tau, as it is mostly orange and yellow. The central part of the alpha-helix is light blue, indicating a slightly higher confidence score in this region compared to the rest of the protein. Moreover, there are no beta-strands present, only one alpha-helix and other secondary structures that are neither helices nor strands.

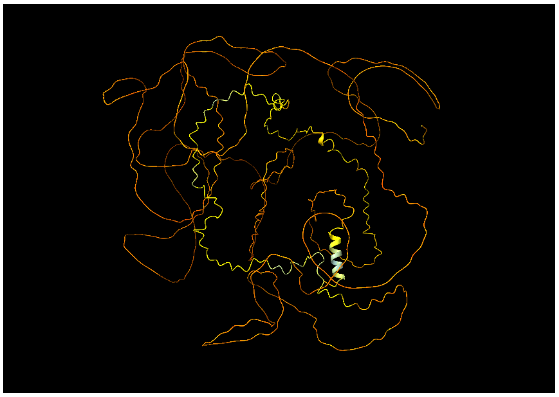

Figure 11. Predicted structure of human tau using ColabFold. The color coding indicates the per-residue confidence score (pLDDT) with dark blue representing very high confidence regions (pLDDT > 90), light blue indicating confident regions (pLDDT 70-90), yellow for regions of low confidence (pLDDT 50-70), and orange denoting very low confidence regions (pLDDT < 50).

The confidence level in the structure of bovine tau, seen in Figure 12, is also low, as it is predominantly orange and yellow. Just like human Tau, the central segment of the alpha-helix is light blue. Unlike human Tau, we see some light blue segments in the coils and loops that make up the rest of the protein structure. However, the overall confidence level seems low.

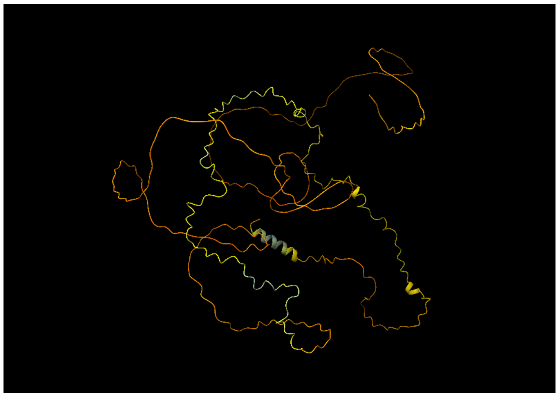

Figure 12. Predicted structure of bovine tau using ColabFold. The color coding indicates the per-residue confidence score (pLDDT) with dark blue representing very high confidence regions (pLDDT > 90), light blue indicating confident regions (pLDDT 70-90), yellow for regions of low confidence (pLDDT 50-70), and orange denoting very low confidence regions (pLDDT < 50).

Just like bovine tau and human tau, Figure 13 suggests that the confidence level in the structure of mouse Tau is low. As with bovine Tau, the central segment of the alpha-helix is light blue and we see some light blue segments in the coils and loops that make up the rest of the protein structure. However, the structure is predominantly orange and yellow.

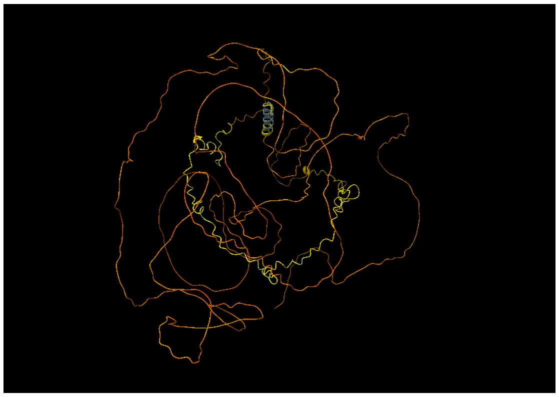

Figure 13. Predicted structure of mouse tau using ColabFold. The color coding indicates the per-residue confidence score (pLDDT) with dark blue representing very high confidence regions (pLDDT $>$ 90), light blue indicating confident regions (pLDDT 70-90), yellow for regions of low confidence (pLDDT 50-70), and orange denoting very low confidence regions (pLDDT $<$ 50).

4 DISCUSSION

In the exploration of the protein structures of synaptophysin and tau using ColabFold, we see a clear pattern in their predicted confidence scores and structural comparisons. For synaptophysin, the pLDDT scores in all three species (humans, bovines, and mice) mostly fall in the very confident and confident ranges. This suggests that ColabFold has predicted a large portion of synaptophysin with a high degree of accuracy. These parts could be suitable for any application that benefits from high accuracy [8]. Interestingly, we also observe the presence of a ribbon-like part within synaptophysin for all three species which is yellow and orange, meaning it is characterized by lower confident scores. A pLDDT score that is under 50 is a reasonably strong predictor of disorder, meaning this region of synaptophysin is either unstructured under physiological conditions or only becomes structured within a complex [8].

In contrast to synaptophysin, the prediction for tau exhibit predominantly low confidence scores, as many areas are colored yellow and orange. This aligns with the understanding that tau is an intrinsically disordered protein, mentioned in section 1.4. The results suggest that tau is disordered in all three species, meaning tau is either unstructured under physiological conditions or only becomes structured within a complex.

The comparisons between structures further emphasize these points. The synaptophysin comparisons showed low RMSD values and used a large number of residues in their comparisons. The similarity in RMSD values between human, bovine and mouse synaptophysin seems to suggest that synaptophysin is a protein who's structure in large part has been preserved through evolution. Conversely, the comparisons between tau-structures yielded poor RMSD values, which could be due to it being disordered. Notably, a smaller number of residues were used when comparing the ColabFold-predicted human tau structure to the experimental and AlphaFold structures available on UniProt (5 and 12 residues, respectively) in contrast to the comparison with bovine and mouse tau obtained with ColabFold, where 33 and 50 residues were used, respectively. Future research could explore if this was due to the fact that all three structures were predicted with ColabFold, and whether similar patterns emerge when ColabFold is used to predict the structure of other proteins.

The main limitation of this project stems from its scope and my background. As I have limited experience in biology and the project is limited in scope, my analysis might not capture all the nuances of these protein's structures and functions. Future work could investigate why certain regions, such as the yellow-colored alpha helix in the synaptophysin predictions, exhibit lower confidence. Future work could also conduct a more detailed analysis focusing on specific regions of tau, and analyze specific segments rather than the entire protein.

REFERENCES

[1] Structure analysis and comparison. Accessed: 2024-03-18.

[2] Protein structure. https://www.nature.com/scitable/topicpage/protein-structure-14122136, 2014. Accessed: 2024-04-07.

[3] Critical assessment of techniques for protein structure prediction. https://www.predictioncenter.org/casp14/doc/CASP14_Abstracts.pdf, 2020. Accessed: 2024-04-07.

[4] Alphafold protein structure database. https://alphafold.ebi.ac.uk/, 2024. Accessed: 2024-04-07.

[5] Colabfold v1.5.5: Alphafold2 using mmseqs2. https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb, 2024. Accessed: 2024-04-08.

[6] Deepmind. https://deepmind.google/, 2024. Accessed: 2024-04-05.

[7] Embl's european bioinformatics institute. https://www.ebi.ac.uk/, 2024. Accessed: 2024-04-07.

[8] FAQ - AlphaFold Database. https://alphafold.ebi.ac.uk/faq, 2024. Accessed: 2024-04-08.

[9] Multiple sequence alignment. https://en.wikipedia.org/wiki/Multiple_sequence_alignment, 2024. Accessed: 2024-04-08.

[10] RCSB Protein Data Bank. https://www.rcsb.org/, 2024. Accessed: 2024-04-07.

[11] Tau protein. https://en.wikipedia.org/wiki/Tau_protein, 2024. Accessed 2024-04-05.

[12] X-ray crystallography, 2024. Accessed: 2024-04-05.

[13] Katarina Ahlfort. Early detection of alzheimer's thanks to groundbreaking medical technology. https://www.kth.se/en/om/nyheter/centrala-nyheter/alzheimers-upptacks-tidigt-genom-banbrytande-medicinteknik-1.1276443, 2023. Accessed: 2024-04-08.

[14] Christian B. Anfinsen. Studies on the principles that govern the folding of protein chains. https://www.nobelprize.org/uploads/2018/06/anfinsen-lecture.pdf, 1972. Nobel Lecture, Accessed: 2024-04-05.

[15] UniProt Consortium. P20488 syph bovin. https://www.uniprot.org/uniprotkb/P20488/entry, 2024. Accessed: 2024-03-18.

[16] UniProt Consortium. Q62277 syph mouse. https://www.uniprot.org/uniprotkb/Q62277/entry, 2024. Accessed: 2024-03-18.

[17] Wikipedia contributors. Root-mean-square deviation of atomic positions. https://en.wikipedia.org/wiki/Root-mean-square_deviation_of_atomic_positions, 2023. Accessed: 2024-02-12.

[18] DeepMind. Alphafold, 2021. Accessed: 2024-02-11.

[19] EMBL-EBI. Hmmer: Biosequence analysis using profile hidden markov models. https://www.ebi.ac.uk/Tools/hmmer/, 2024. Accessed: 2024-04-08.

[20] Demis Hassabis. Alphafold reveals the structure of the protein universe. https://deepmind.google/discover/blog/alphafold-reveals-the-structure-of-the-protein-universe/, 2021. Accessed: 2024-04-07.

[21] John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David A. Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021.

[22] Schwede T. Topf M. Fidelis K. Moult J. Kryshtafovych, A. Critical assessment of methods of protein structure prediction (casp)-round xiii. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6927249/, 2019. Accessed: 2024-04-07.

[23] Jesse Mex and Bobak Abdolmohammadi. Chapter 7 - genetics of chronic traumatic encephalopathy. pages 83–90, 2018. Accessed: 2024-04-08.

[24] Borgnia M. J. Bartesaghi A. Tran E. E. Earl L. A. Schauder D. M. Lengyel J. Pierson J. Patwardhan A. Subramaniam S Milne, J. L. Cryo-electron microscopy–a primer for the non-microscopist. *The FEBS journal*, 2013.

[25] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: Making Protein folding accessible to all. *Nature Methods*, 2022.

[26] Milot Mirdita, Martin Steinegger, and Johannes S"oding. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics*, 35(16):2856–2858, 2019.

[27] Milot Mirdita, Lars von den Driesch, Clovis Galiez, Maria J. Martin, Johannes S"oding, and Martin Steinegger. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.*, 45(D1):D170–D176, 2017.

[28] Alex L Mitchell, Alexandre Almeida, Martin Beracochea, Miguel Boland, Josephine Burgin, Guy Cochrane, Michael R Crusoe, Varsha Kale, Simon C Potter, Lorna J Richardson, Ekaterina Sakharova, Maxim Scheremetjew, Anton Korobeynikov, Alex Shlemov, Olga Kunyavskaya, Alla Lapidus, and Robert D Finn. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*, 2019.

[29] Judson R Fidelis K. Moult J, Pedersen JT. A large-scale experiment to assess pro-

tein structure prediction methods. https://pubmed.ncbi.nlm.nih.gov/8710822/, 1995. Accessed: 2024-04-07.

[30] Arie Perry and Daniel J. Brat. Practical surgical neuropathology. pages 1–14, 2010. Accessed: 2024-04-08.

[31] Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature Methods*, 9:173–175, 2012.

[32] The AlphaFold team. Alphafold: A solution to a 50-year-old grand challenge in biology. https://deepmind.google/discover/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biolo 2024. Accessed: 2024-04-05.

[33] Adler J. Wu Z. et al. Tunyasuvunakool, K. Highly accurate protein structure prediction for the human proteome. https://www.nature.com/articles/s41586-021-03828-1#citeas, 2021. Accessed: 2024-04-07.

[34] Visualization UCSF Resource for Biocomputing and Informatics. Command: rmsd. https://www.cgl.ucsf.edu/chimerax/docs/user/commands/rmsd.html, 2024. Accessed: 2024-04-08.

[35] UCSF Resource for Biocomputing, Visualization, and Informatics. Ucsf chimerax. https://www.cgl.ucsf.edu/chimerax/, 2024. Accessed: 2024-02-12.

[36] UniProt Consortium. P08247 syph human. https://www.uniprot.org/uniprotkb/P08247/entry#structure, 2024. Accessed: 2024-02-12.

[37] UniProt Consortium. P10636 tau human. https://www.uniprot.org/uniprotkb/P10636/entry#function, 2024. Accessed: 2024-02-12.

[38] UniProt Consortium. P10637 tau mouse. https://www.uniprot.org/uniprotkb/P10637/entry, 2024. Accessed: 2024-02-12.

[39] UniProt Consortium. P29172 tau bovin. https://www.uniprot.org/uniprotkb/P29172/entry, 2024. Accessed: 2024-02-12.

[40] Wikipedia. Synaptophysin. https://en.wikipedia.org/wiki/Synaptophysin, 2023. Accessed 2024-04-05.