HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 3 (2024)

Assessing AlphaFold AI's Protease Enzyme Structure Prediction Accuracy

Iris Zhang[1] and Janet Wolfe[#]

[1]Conestoga High School, USA
[#]USA

ABSTRACT

This study analyzed AlphaFold AI's ability in accurately predicting protease enzyme structures. AlphaFold uses machine learning, taking amino acid sequences and using physical and scientific knowledge of protein structures to generate a protein structure prediction. Past studies have comfirmed AlphaFold's general abilities but have identified limitations in certain factors, like post translational modifications, ligands, and other environmental factors. However, there have not been studies assessing AlphaFold in predicting protease enzyme structures specifically. Quantitative data was collected using ex-post facto and correlational methods, which compared the RMSD score between AlphaFold and Protein Data Bank structures of the same protease enzyme. Furthermore, correlational trends were searched for between protein complexity and length with the RMSD score. 77% of the 30 protease enzymes assessed were found to be accurate, with more complex structures lowering in accuracy. Protein length was not a factor in AlphaFold's prediction accuracy. By utilizing these findings, researchers in the pharmaceutical industry can consider the weak points of AlphaFold, conduct further studies identifying more factors that contribute to AlphaFold's accuracy, and work on improving the program based on the results.

Introduction

As artificial intelligence continues to advance in many facets of life, the need for its use in the field of pharmaceutical development grows. According to Giorgio Buttazzo, from the department of Excellence on Robotics and AI at Sant'Anna School of Advanced Studies, artificial intelligence is progressing at a rapid pace, surpassing human capability in various tasks, including pursuits within the medical industry (Buttazzo, 2023). This topic has been further expanded with the development of the machine learning software, AlphaFold, which was released in July of 2021 (Jumper, 2021). AlphaFold is a deep learning program created by DeepMind that serves to predict protein structures (Jumper, 2021). Since the software's release, AlphaFold has used primary amino acid sequences and aligned sequences of homologues as inputs to predict protein structures, and it has made over 350,000 models available for public access (J.M Thornton, J Jumper). Furthermore, according to Adarsh Sahu, from the Department of Pharmaceutical Sciences at Harisingh Gour University, Artificial intelligence has been able to improve the process of pharmaceutical development, the design method and reduce research time, development time, and costs. Traditional methods for drug development employ experiments such as X-ray crystallography, are lengthy, costly, and have a high failure rate (Dhakal, 2021). Therefore, with the rising development of machine learning software such as AlphaFold, methods of drug development have the potential to be assisted with Artificial intelligence. Artificial intelligence can possibly reduce the costs, development time, and errors in developing pharmaceuticals, benefiting patients with various medical conditions by helping them get access to treatments faster (Kolluri, 2022). However, the accuracy of AlphaFold may be limited depending on the type of protein, according to Janet Thornton of European Bioinformatics Institute. Furthermore, there is a lack of consistent research results which determine the accuracy of AlphaFold in predicting protease

enzyme structures. This leads to the research question: to what extent can the machine learning software Alpha Fold confidently predict protease enzyme structures in comparison to experimentally determined structures from the Protein Data Bank? The Protein Data Bank is a database that contains the 3D structural data of proteins and nucleic acids (Berman, 2000). Researchers are able to input experimentally determined protein structures into the PDB, making it a valuable resource for the field of pharmaceutical development. Additionally, PDB provides information about the coordinates of atoms in a molecule, allowing them to be visualized and analyzed, serving as a crucial reference for validating computational predictions, like those generated by AlphaFold (Berman, 2000).

Literature Review

At its current rate of progression, Artificial intelligence has the potential to revolutionize drug development, assisting with the lengthy and costly development time of traditional development methods. Traditionally, the production of Pharmaceuticals is a highly lengthy and expensive process (Kolluri, 2022). It can take decades and billions of dollars to produce a new drug/pharmaceutical (Dhakal, 2021). However, according to Ashwin Dhakal from the Bioinformatics & Machine Learning Lab at the University of Missouri, using AI to predict the most optimal protein structures could possibly drastically reduce costs and development time. Similarly, as explained by Adarsh Sahu, Artificial intelligence has improved the design method and reduced research time, as target proteins can also be identified using artificial intelligence, enhancing the success rate of drug development. Additionally, AI can data mine based on pharmaceutical data and machine learning and has been used in de novo drug design (Sahu, 2021). AI can play an impactful role in streamlining and optimizing the process of protein prediction and has the potential for real-world impact. As mentioned by Cade Metz, a technology reporter at the New York Times, Alphafold AI's protein prediction is on the rise, and has generated over 350,000 protein predictions by 2021. Researchers have used some of these predictions to develop possible malaria vaccines and treatments for other conditions (Metz, 2022). As Artificial Intelligence continues to grow and evolve, its integration into the biotech field and pharmaceutical development has the potential to help resolve longstanding challenges within the industry.

Furthermore, AlphaFold and artificial intelligence have already grown and shown success in certain areas of protein predictions. Amir Shanehsazzadeh, an AI researcher and colleagues at Absci, a generative AI drug creation company, carried out a study using generative artificial intelligence deep learning models to design antibodies against three targets, which demonstrated the frequent binding rates for specific regions of the antibodies. The experiment carried out by Shanehsazzadeh and colleagues confirms the effectiveness of these AI-designed antibodies, as they exhibited diversity, low sequence similarity to already known antibodies, and favorable developability profiles (Shanehsazzadeh, 2024). These results demonstrate that there is potential for the development of new therapeutic antibody targets using generative artificial intelligence and high-throughput experimentation. Additionally, according to John Jumper, a senior research scientist at DeepMind Technologies, AlphaFold has shown not only success, but improvement in the accuracy of the program's structure prediction by "incorporating novel neural network architectures and training procedures based on the evolutionary, physical and geometric constraints of protein structures" (Jumper, 2021). As AlphaFold continues to show success and improvement in protein structure prediction as well as the program's developers continuing to improve its algorithms, its value within the biotech industry increases.

However, there are various limitations to Alphafold's prediction accuracy, and there is insufficient research available on the program's ability to predict protease enzymes. This is a significant gap, according to Carios Lopez-Otin, a Professor in the Biochemistry at the University of Oviedo, protease enzymes are significant to the biomedical field and "regulate the fate, localization, and activity of many proteins, modulate protein-protein interactions, create new bioactive molecules, contribute to the processing of cellular information, and generate, transduce, and amplify molecular signals"(Lopez-Otin, 2008). Furthermore, there have been various

inaccuracies found in AlphaFold's algorithms that refute other studies in which AlphaFold was found to be relatively accurate. According to Haroldas Bagdonas, who specializes in structural bioinformatics at the University of York, the AlphaFold algorithm does not consider post-translational modifications and cofactors that are essential to the protein's structure and functions (Bagdonas, 2021). To further highlight this gap, Devlina Chakravarty, the Managing Director at Artemis Global Life Sciences Limited, tested alpha fold predictions on 98-fold switching proteins which had at least two distinct and stable secondary and tertiary structures. With the results of the experiment, Chakravarty discovered that approximately 94% of alpha fold's predictions successfully predicted "one experimentally determined conformation but not the other" (Chakravarty, 2022). More shortcomings of the program are emphasized by Thomas Terwilliger, a researcher at the Los Alamos National Laboratory, who asserts that "The accuracies of these predictions vary, however, and they do not take into account ligands, covalent modifications or other environmental factors." These findings demonstrate a gap in AlphaFold's ability to accurately predict certain types of protein structures, as many partner regions are unavailable in the computational folding process. In addition, there is little information known about AlphaFold's accuracy in predicting protease enzymes, which are important in many biological processes.

Overall, there is a lack of solid conclusion on the ability of alpha fold to accurately predict protease enzyme structures despite their significance in the pharmaceutical development industry. With the rise and growth of AI, alpha fold could prove to be significantly beneficial to the medical field, saving time and costs. In the context of the current body of knowledge related to AlphaFold's accuracy in protein prediction, research on the accuracy of Alpha Fold in predicting protease enzyme structures in relation to protein complexity could help workers in the pharmaceutical development industry draw conclusions on AlphaFold's ability in predicting protease enzymes and can possibly incorporate the use of Alpha fold into the process of drug development. For the method of inquiry, a mix of ex post facto and correlational methods will be used. The software ChimeraX will be utilized to compare the alpha fold predicted structures of 30 proteins in comparison to experimentally determined structures from Protein Data Bank, taking the root mean square deviation, or the difference between each pair. ChimeraX is a molecular visualization and analysis software, allowing users to analyze structures such as proteins in 3D (Goddard, 2017). In addition, the software offers various functions such as structural comparison, alignment, and the calculation of structural metrics like Root Mean Square Deviation (RMSD), which will be calculated in this study to measure the similarity between the predicted and experimental structures (Goddard, 2017). The structures of the enzymes from both AlphaFold and PDB will be loaded onto ChimeraX. The predicted structure will then be aligned with the experimentally determined structure for accurate comparison. The RMSD data will then be calculated and collected into a data table for analysis. In addition, the protein structures can be visualized and inspected for structural differences (Goddard, 2017). To create an efficient process, a script/loop will be created to automate the process for each set of predicted and experimental structure. It is hypothesized that the accuracy of alpha fold's protease enzyme predictions will decrease as protein complexity increases.

Methods

This study explored the ability of the AI program AlphaFold in predicting protease enzyme structures. The goal was to provide valuable insights on the capabilities of AlphaFold, as the software has the potential to provide researchers with more structural knowledge of protein structures, assisting them in developing pharmaceuticals, as they must effectively bind to sites on protein structures. This is important as current methods for determining protein structures require traditional methods such as X-ray crystallography, which are lengthy and costly, slowing down drug hypothesis testing. Using AI programs like AlphaFold to understand protein structures could expedite hypothesis testing, improving pharmaceutical accessibility. For the method of inquiry, a two-part mix of ex post facto and correlational methods were used to carry out the research study, collecting qualitative data. The chosen ex post facto and correlational research methods align with the topic and goals of the research study,

as data and variables that already exist with no manipulation were analyzed, which were AlphaFold's predicted protein structures and the Protein Data Bank's experimentally determined structures. Analyzing pre-existing variables provide valuable insights into the current capability of the AlphaFold software, whether it is proving that AlphaFold is capable in predicting protease enzyme structures, or finding places where the software can be improved. Additionally, correlations were searched for between protein length in amino acids and complexity and prediction accuracy to find valuable insights on what factors influence AlphaFold and its ability to predict protease enzymes, potentially benefiting the field of pharmaceutical development. Furthermore, protease enzymes are valuable subjects in assessing AlphaFold's value to the pharmaceutical industry, as they are significant to the biomedical field and "regulate the fate, localization, and activity of many proteins, modulate protein-protein interactions, create new bioactive molecules, contribute to the processing of cellular information, and generate, transduce, and amplify molecular signals"(Lopez-Otin, 2008).

The Protein Data Bank and AlphaFold were the primary sources from where the data was collected in the study. The Protein Data Bank, or PDB, is a database widely used by researchers that contains the 3D structures of biological molecules. The protein structures in the database are collected by various scientists and researchers using traditional methods like X-ray crystal structure determination, NMR, and cryoelectronic microscopy. The experimental structures used in this study were collected from the PDB and downloaded in a mmCIF format. The AlphaFold database was then used to collect the AI predicted structures. Alphafold is an artificial intelligence program developed by DeepMind, which uses deep learning to predict the structures of proteins from their amino acid sequences. Once each protein with an available PDB and AlphaFold structure were identified, the AlphaFold structure was measured for accuracy in comparison to the experimental PDB structure. The software, ChimeraX, a molecular visualization and analysis program, created by UC San Diego, was the main program used in the study to analyze the protease enzyme structures.

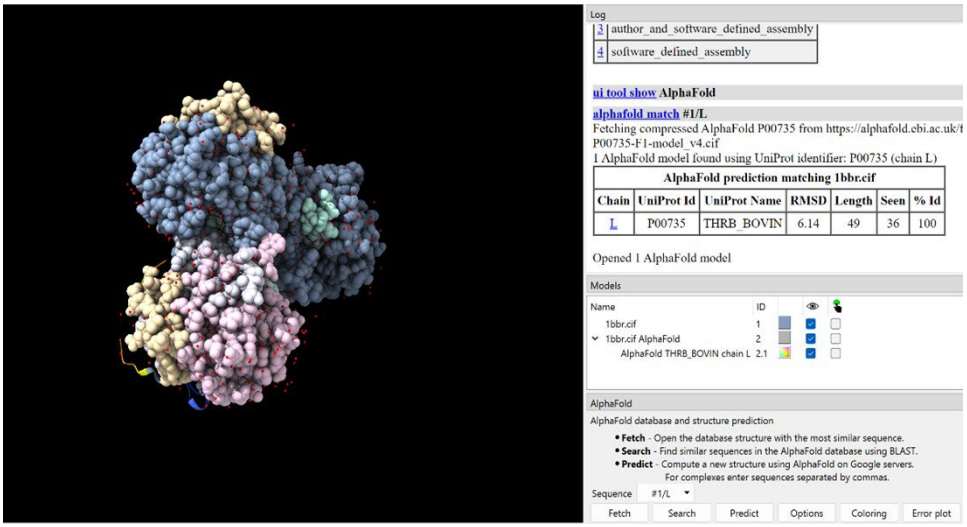

Figure 1. A screenshot displaying the ChimeraX software comparing a PDB and AlphaFold structure.

Shown in figure 1 is an example of the data collected for one protein in ChimeraX. ChimeraX allows users to analyze structures such as proteins in 3D (Goddard, 2017). The protein structures can also be visualized and inspected for structural differences with ChimeraX (Goddard, 2017). In addition, the software offers various functions such as structural comparison, alignment, and the calculation of structural metrics like Root Mean Square Deviation (RMSD), which was calculated in this study to measure the similarity between the predicted and experimental structures (Goddard, 2017). ChimeraX was used to compare the Alphafold predicted structures of 15-20 protease enzymes in comparison to experimentally determined structures from Protein Data Bank,

taking the root mean square deviation, or the RMSD score. The RMSD is the measure of the average distance between the atoms of proteins in Angstroms and is widely used by researchers to measure the similarity of multiple protein structures. According to Castro-Alvarez and colleagues, who carried out a similar protein study using RMSD scores, a RMSD score of less than or equal to two Angstroms is acceptable in most cases. For this study, a similar criterion was used and RMSD scores greater than two Angstroms were considered relatively inaccurate. The structures of the protease enzymes from the Protein Data Bank were then loaded onto the ChimeraX software. The experimental structure was then aligned with the correlated or most similar structure from AlphaFold for accurate comparison. The RMSD score was then calculated and collected into a data table for analysis. More protein properties provided by ChimeraX were also recorded, including the number of unique protein chains and chain length in amino acids. In addition, the pLDDT score of each protein from AlphaFold was also recorded and analyzed. The pLDDT is AlphaFold's measure scale in confidence for its protein predictions, and the score ranges from 0-100, 0 being the lowest and 100 the highest. According to the ChimeraX ALphaFold tool, a pLDDT score of 90-100 suggests high accuracy, while scores from 0-50 indicate a disordered protein that should not be interpreted. Furthermore, to create an efficient process, a script/loop was created to automate the process for each set of predicted and experimental structure. The protease enzymes were randomly selected off the Protein Data Bank and were picked from five major groups of protease enzymes, which were serine, cysteine, threonine, aspartic, and glutamic enzymes, to maintain diversity and to prevent sampling bias. In total, 30 diverse protease enzymes from varying groups were analyzed through ChimeraX. The A chain of each protein was analyzed.

Some limitations occurred if a corresponding AlphaFold structure to the PDB A chain structure was not yet available. When this was the case, a different protease enzyme was identified and analyzed in place of the randomly selected protein. However, varying resolutions of structures within the PDB and AlphaFold may have also caused inaccuracies in the RMSD, as higher resolutions provide more detail of the atomic structure, leading to a higher RMSD score. Lower resolutions could also block out details within a protein structure, also raising RMSD scores. Additionally, AlphaFold predictions may not include the whole protein structure for larger and more complex proteins. In these cases, RMSD scores will only be available for the predicted region, potentially adding bias to the data. Furthermore, AlphaFold does not account for post-translational modifications that many proteins go through, changing their structure. This may cause the AlphaFold structure to vary more from the PDB structure, raising RMSD scores.

Results

Overall, RMSD scores and additional protein data from 35 protease enzymes were collected from the ChimeraX software. The sample consisted of randomized proteins, varying in protease groups, length, and number of complex protein chains. The RMSD score ranged from 0.24 to 12.08 angstroms while protein lengths ranged from 13 to 1045 amino acids. The number of complex protein chains in each structure ranged from 1 to 10 chains. To answer the research question, "to what extent can the machine learning software AlphaFold confidently predict protease enzyme structures in comparison to experimentally determined structures from the protein data bank?", the proteins were classified into 3 groups based on their RMSD score, which were good, acceptable, and inaccurate.

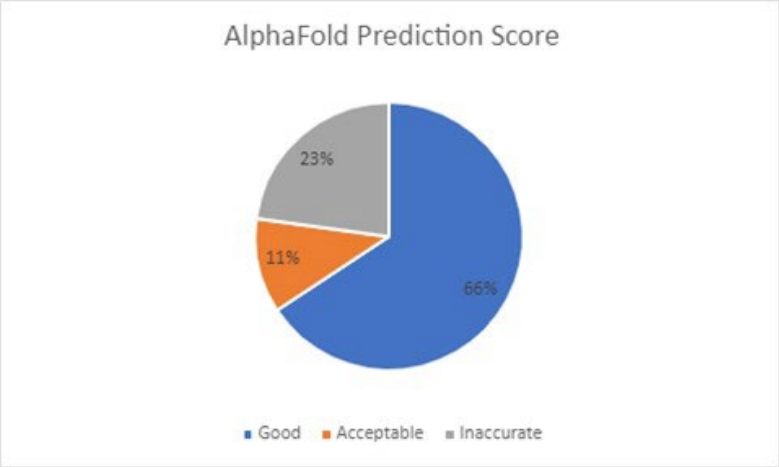

Figure 2. A pie chart displaying the accuracy scores of the sample protease enzymes

Shown in Figure 2 is a pie chart displaying the accuracy scores of 35 Protease Enzymes. After RMSD scores were collected, it is shown in the figure that the majority of the structural predictions were either good or acceptable. It was found that the RMSD score of 65.71% of the protease enzymes collected were less than or equal to 2 angstroms, classifying them as good and generally accurate protein predictions. 11.43% of the protease enzymes analyzed were classified as acceptable, with RMSD scores ranging from 2 to 3 angstroms. 22.86% were found to be generally inaccurate, with RMSD scores greater than 3 angstroms. To further analyze possible causes of the prediction inaccuracies of 22.86% of the structures, the RMSD, or accuracy score of the protease enzymes were compared to the number of complex protein chains in each structure and protein length. Possible correlations between these structural factors and AlphaFold prediction accuracy were searched for.

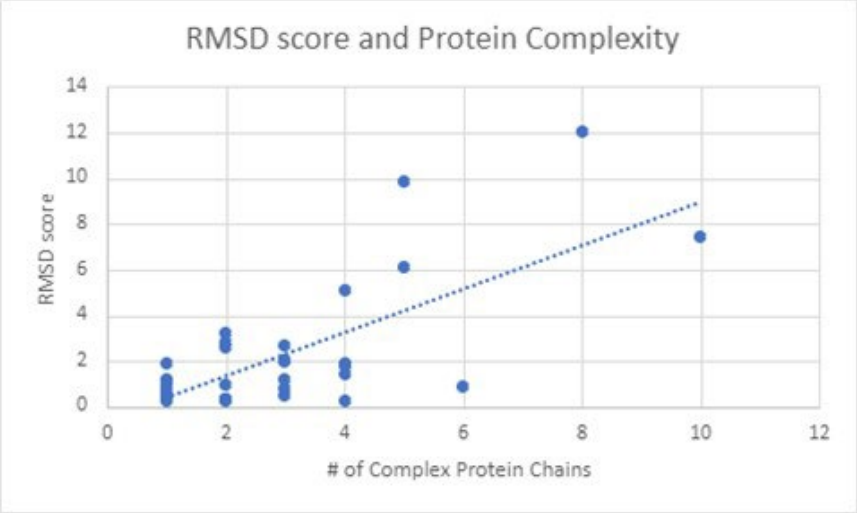

Figure 3. A scatterplot displaying the correlation between complex protein chains and RMSD score.

Shown in figure 3 is a visual scatterplot displaying a possible correlation between AlphaFold prediction inaccuracies and the number of complex protein chains within the structure. The average protease enzyme examined had 3 complex protein chains, and the enzymes ranged from 1 to 10 complex protein chains. According

to the trendline, there is a positive slope, signaling a positive correlation between protein complexity and prediction inaccuracy, with a positive slope of 0.96. Additionally, a correlational coefficient for the scatter plot was calculated at 0.723, indicating a strong positive correlation. However, the correlation is not entirely strong, with multiple outliers. Protein 3ZRY was an outlier, with a good AlphaFold prediction score of 0.85 despite having 6 complex protein chains. Protein 5T4O's RMSD score of 12.08 was significantly higher than the trend line, scoring 5.99 angstroms higher than the predicted RMSD score value. Additionally, there was a lack of data in the Protein Data Bank of protease enzymes with 6-10 complex protein chains. This absence of data may have changed the trend line and correlational strength. Despite these limitations, the positive correlation between RMSD score and protein complexity is apparent. However, there may be other factors that could contribute to a higher RMSD score, correlating with the outliers.

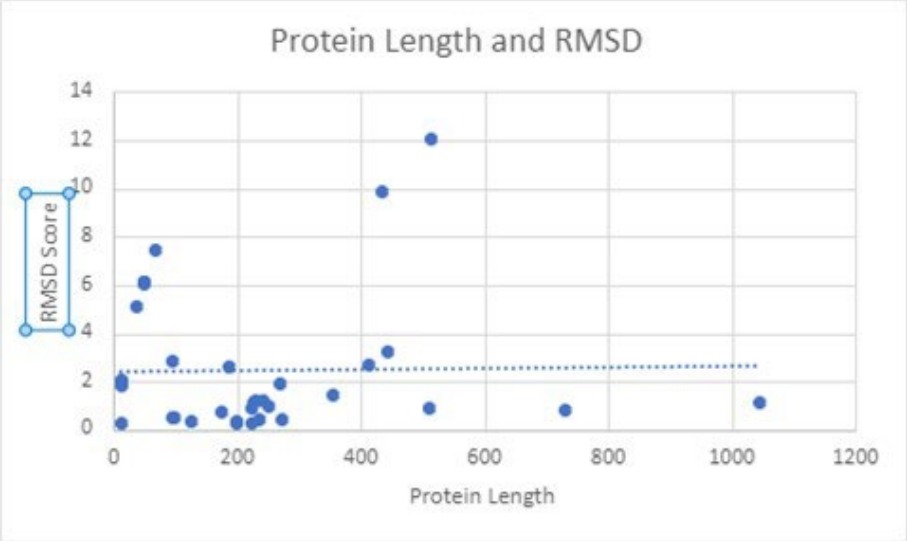

Figure 4. A scatterplot displaying the correlation between protein length and RMSD score.

Shown in Figure 4 is a visual scatterplot displaying the relationship between protein length and AlphaFold prediction accuracy. The average protease enzyme length was 231.77 and the sample ranged from 0.24 to 12.08 amino acids. According to the scatterplot, there are clearly no noticeable correlations between protease enzyme length and prediction accuracy due to the flat trend line. A correlational coefficient was calculated for the scatterplot as 0.017, indicating no evidence of correlation between protein length and prediction accuracy. There are both proteins with good RMSD scores and high lengths and proteins with inaccurate predictions and shorter lengths. Out of these 35 protease enzyme structures, it was determined that protein length was not a significant factor in determining AlphaFold's prediction accuracy.

Analysis

Overall, it was found that the majority of the protease enzymes structures were predicted accurately, with 77.14% of the sample being classified as accurate or good and 22.86% of the sample classified as inaccurate. It was demonstrated that AlphaFold was able to accurately predict the majority of the protease enzyme structures, especially those with less complex protein chains. There was a strong correlation between the number of complex protein chains and RMSD score, but there was no significant correlation between protein length and RMSD score. A possible cause of the strong positive correlation between the number of complex protein chains and RMSD score is that Proteases with multiple complex chains are more structurally complex due to their three

dimensional and interactive nature. Predicting these complex structures requires AlphaFold to understand molecule-molecule interactions and accurately predict spatial arrangements of multiple complex chains, posing challenges to the algorithm and leading to a higher RMSD score. Additionally, Proteases with more complex protein chains may be more structurally flexible because of the multiple interacting components. This increased flexibility could result in larger conformational differences between experimental and predicted structures, leading to higher RMSD scores. Additionally, proteases with more complex chains often have more complex and important functions in cellular processes, making them more susceptible to different conformational changes that AlphaFold's algorithms may not be able to predict, causing a larger RMSD score. Furthermore, no correlations were found between protease length and RMSD score. A possible reason for this absence of association is that while longer protein chains are larger and therefore may be more complex, AlphaFold's prediction accuracy may not deteriorate with protein chain length and longer chains can still have fairly simple structures.

Limitations

Some possible limitations to the data and trends examined may have arisen during the study. The quality or resolution of protein data bank could have varied from the AlphaFold predicted structure, contributing to possibly higher RMSD scores if the low resolution left out structural details. This may have been the case for larger and more complex structures. Additionally, AlphaFold predictions may not include the whole protein structure for larger and more complex proteins. In these cases, RMSD scores will only be available for the predicted region, potentially adding bias to the data. Furthermore, AlphaFold does not account for post-translational modifications that many proteins go through, changing their structure. This may cause the AlphaFold structure to vary more from the PDB structure, raising RMSD scores. Additionally, there may have been varying experimental methods used to determine the Protein Data Bank structures. This could introduce variability that may not correlate with the AlphaFold prediction, leading to increased RMSD scores. Furthermore, there were few protease enzymes with 6-10 complex protein chains available in the protein data bank, causing Figure 1 to lack data points in the 6-10 complex protein chains range. This absence of data may have left out data points that could have influenced the trend line and correlation observed.

Conclusion

The research question, "to what extent can the machine learning software AlphaFold confidently predict protease enzyme structures in comparison to experimentally determined structures from the protein data bank?", was answered in this study. From the structural comparisons of predicted and experimental protease enzyme structures in my method, tt was found that the majority of AlphaFold's protease enzymes prediction were accurate, at 77.14% accuracy. While other factors may influence the accuracy of AlphaFold's protein predictions, it was found through my correlation studies between RMSD score and protein complexity and length that it is apparent that the number of complex protein chains in a structure have an impact on the RMSD score and accuracy of the predictions. However, protein length did not play a role in shaping RMSD scores. Therefore, my original hypothesis that RMSD scores would increase as the complexity and length increased was partially correct, as RMSD was affected by the number of complex protein chains but was not affected by protein length. These results are significant as AI is becoming increasingly prevalent and useful in many aspects of life, including the medical field. Current methods for drug development like X-ray crystallography are lengthy and costly with a high failure rate. If AlphaFold can be properly and accurately used to predict protein structures for drug development, it can greatly reduce developments time and costs. The results of the study show that AlphaFold is generally reliable in predicting simpler protease enzyme structures with less complex chains, and protein length does not affect the prediction accuracy. These results are useful to the biotech and pharmaceutical fields, as they

provide valuable insights on the capabilities of AlphaFold in predicting protease enzymes. With these factors in mind, they can help researchers and scientists determine whether or not AlphaFold would be valuable and reliable for their specific protein projects using protease enzymes. With this new information on AlphaFold's protease enzyme prediction capabilities, researchers can further research more factors that may contribute to AlphaFold's inaccuracies and decide whether to develop AlphaFold for more complex protease predictions, benefiting the fields of biotech and pharmaceutical development.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Bagdonas, H., Fogarty, C.A., Fadda, E. *et al.* The case for post-predictional modifications in the AlphaFold Protein Structure Database. *Nat Struct Mol Biol* 28, 869–870 (2021). https://doi.org/10.1038/s41594-021-00680-9

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic acids research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Buttazzo G. (2023). Rise of artificial general intelligence: risks and opportunities. *Frontiers in artificial intelligence*, *6*, 1226990. https://doi.org/10.3389/frai.2023.1226990

Chakravarty D, Porter LL (2022). AlphaFold2 fails to predict protein fold switching. *Protein Science*. 31(6):e4353. https://doi.org/10.1002/pro.4353

Castro-Alvarez, A., Costa, A. M., & Vilarrasa, J. (2017). The Performance of Several Docking Programs at Reproducing Protein-Macrolide-Like Crystal Structures. *Molecules (Basel, Switzerland)*, *22*(1), 136. https://doi.org/10.3390/molecules22010136

Dhakal, A., McKay, C., Tanner, J. J., & Cheng, J. (2021). Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions. *Briefings in Bioinformatics*, *23*(1). https://doi.org/10.1093/bib/bbab476\

Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H., & Ferrin, T. E. (2018). UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein science : a publication of the Protein Society*, *27*(1), 14–25. https://doi.org/10.1002/pro.3235

Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

Kolluri, S., Lin, J., Liu, R., Zhang, Y., & Zhang, W. (2022). Machine Learning and Artificial Intelligence in Pharmaceutical Research and Development: a Review. *The AAPS Journal*, *24*(1). https://doi.org/10.1208/s12248-021-00644-3

López-Otín, C., & Bond, J. S. (2008). Proteases: multifunctional enzymes in life and disease. *The Journal of biological chemistry*, *283*(45), 30433–30437. https://doi.org/10.1074/jbc.R800035200

Metz, C. (2022, July 28). A.I. Predicts the Shape of Nearly Every Protein Known to Science. *The New York Times*. https://www.nytimes.com/2022/07/28/science/ai-deepmind-proteins.html

Sahu, A., Mishra, J., & Kushwaha, N. (2021). Artificial Intelligence (AI) in Drugs and Pharmaceuticals. *Combinatorial Chemistry & High Throughput Screening*, *25*. https://doi.org/10.2174/1386207325666211207153943

Shanehsazzadeh, A. (2024). Unlocking de novo antibody design with generative artificial intelligence [Review of *Unlocking de novo antibody design with generative artificial intelligence*]. Cold *Spring Harbor Laboratory*.

Terwilliger, Thomas C et al. "AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination." *Nature methods* vol. 21,1 (2024): 110-116. doi:10.1038/s41592-023-02087-4

Thornton, J.M., Laskowski, R.A. & Borkakoti, N. AlphaFold heralds a data-driven revolution in biology and medicine. *Nat Med* 27, 1666–1669 (2021). https://doi.org/10.1038/s41591-021-01533-0