Deciphering ERAP1 in Cancer: Machine Learning for Protein Detection and Therapeutic Implications

Chloe Cheng[1,*]

[1]*Queen Margaret College, Wellington, New Zwaland*
[*]*Corresponding author*

Abstract: *ERAP1 plays a pivotal role in processing antigenic peptides for presentation on major histocompatibility complex (MHC) class I molecules (Li et al. [1]). Disruptions in ERAP1 expression or function have been associated with a range of diseases, including cancer. Recent investigations have unveiled ERAP1's potential involvement in regulating tumor cell growth and immune evasion in cancers like lung cancer, melanoma, and breast cancer (Stratikos et al. [2]). Inhibiting ERAP1 activity has emerged as a promising therapeutic approach for combating these types of cancers (Bufalieri et al. [3]).The dataset used for this study comprises diverse proteins, representing various protein families. Among these, ERAP1 is a member of the aminopeptidase M1 family. This research is centered around the utilization of ERAPNet to identify Endoplasmic Reticulum Aminopeptidase 1 (ERAP1) within cancer cells. Both Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models were used and trained using TensorFlow, a deep learning framework. During the training process, optimizations were undertaken to fine-tune the models' parameters and weights, enhancing their proficiency in detecting ERAP1 within the samples. The models underwent post-training evaluation using a testing dataset. Their accuracy was assessed using metrics including precision, recall, and the F1 score. Additionally, confusion matrices were generated to provide insight into the models' performance concerning the detection of ERAP1 protein. This evaluation process ensures reliability in the models' predictive capabilities.*

Keywords: *ERAP1, antigenic peptides, major histocompatibility complex (MHC) class I, cancer, tumor cell growth*

1. Introduction

Protein structure prediction has been a subject of intense research for several decades, and significant progress has been made in this field. Researchers have developed various techniques to predict protein structures, such as homology modeling, ab initio modeling, and hybrid methods (Seffernick and Lindert [4]). Homology modeling is the most widely used method and is based on the assumption that proteins with similar sequences have similar structures. Ab initio modeling, on the other hand, predicts protein structures from scratch, without any prior knowledge of the protein's sequence or structure (Lee, Wu, & Zhang. [5]). Hybrid methods combine both homology modeling and ab initio modeling techniques to improve the accuracy of predictions (Seffernick and Lindert [4]). This progress in the field of protein structure prediction has enabled researchers to understand the structure and function of proteins at a molecular level, leading to the development of new drugs and therapies. Protein model assessment is an important aspect of protein structure prediction, as it helps to evaluate the quality of the predicted models. Various metrics have been developed to assess the accuracy of protein models, such as root-mean-square deviation (RMSD), global distance test (GDT), and MolProbity score (Kufareva and Abagyan [6]). Traditional biological methods such as X-ray crystallography, NMR spectroscopy, and electron microscopy are also used to determine the structures of proteins (PDB, RCSB. [7]). However, these methods are expensive, time-consuming, and require large amounts of protein samples. In recent years, machine learning and AI methods have been applied to protein structure prediction and model assessment, leading to faster and more efficient techniques. For instance, deep learning-based methods such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have shown promising results in protein structure prediction and model assessment.

Despite the progress made in the field of protein structure prediction and model assessment, several challenges remain. One of the biggest challenges is the accurate prediction of protein-protein interactions, which play a crucial role in many biological processes (Y. Li et al. [8]). Another challenge is the prediction of protein structures with multiple domains or regions, which often require a combination of different modeling techniques (Xia et al. [9]). Additionally, the accuracy of predicted models decreases with the increase in protein size, making it difficult to predict the structures of large proteins. To address these challenges, researchers are exploring new machine learning and AI methods, such as reinforcement learning and generative models, which have the potential to improve the accuracy and efficiency of protein structure prediction and model assessment (Alquraishi. [10]).

While substantial strides have been made in the field of protein structure prediction and model assessment, there remains opportunity for refinement. A critical hurdle lies in the validation of projected protein structures, a pivotal step in guaranteeing their precision and dependability (Bagaria et al. [11]). Conventional validation techniques, including X-ray crystallography and NMR spectroscopy, are resource-intensive, time-consuming, and demand specialized expertise. Thus, there is an imperative to cultivate validation methods that are both more streamlined and accessible. To address this challenge, researchers have turned to machine learning and deep learning methods to develop faster and more accurate validation techniques. For example, one approach is to use machine learning algorithms to predict the quality of protein models based on their physical and chemical properties. Several studies have reported promising results in predicting protein quality using machine learning algorithms such as random forests, support vector machines, and artificial neural networks.

Another promising approach is to use deep learning methods, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to predict the quality of protein models. These methods can analyze large datasets of protein structures and extract complex features that can be used to predict the quality of new protein models. For example, a recent study used a deep neural network to predict the quality of protein models based on their Ramachandran plot scores, which are a measure of the stereochemical quality of protein structures.

In addition to machine learning and deep learning methods, there is also a need to develop new validation techniques that can assess the quality of protein models in real-time. One promising approach is to use cloud-based computing platforms that can rapidly validate protein models using parallel computing techniques.

2. Introduction of the Program Background

ERAPNet is a machine learning-based approach that uses spectral imaging and deep learning algorithms to accurately identify ERAP1 in human cells and tissues. The development of ERAPNet was motivated by the need for a more precise and efficient method to detect ERAP1, a human protein involved in the processing and presentation of antigens to the immune system.

ERAP1 plays a crucial role in shaping the immune response by trimming and modifying peptides to fit into the binding groove of MHC molecules (Li et al. [1]). Dysregulation of ERAP1 has been implicated in various diseases, including autoimmune disorders and cancer. However, traditional methods of detecting ERAP1 are time-consuming and often require complex sample preparation. ERAPNet was developed to address these limitations by leveraging the power of machine learning and spectral imaging to identify ERAP1 in human cells and tissues with high accuracy and efficiency. This technology has the potential to significantly improve our understanding of ERAP1's role in disease pathology and aid in the development of targeted therapies for various diseases.

2.1. Reasons for Doing This

There are several reasons for employing ERAPNet for detecting ERAP1:

1) Accuracy: ERAPNet is a highly accurate method for detecting ERAP1. By using this method and training the algorithm on a dataset of known ERAP1 samples, accurate results are ensured.

2) Efficiency: ERAPNet uses spectral imaging and deep learning algorithms to identify ERAP1 quickly and efficiently, which is particularly useful when analyzing large datasets.

3) Complexity: ERAP1 is a complex protein with many functions and interactions. Traditional detection methods can be time-consuming and require complex sample preparation. ERAPNet simplifies the process by providing a straightforward, reliable way to detect ERAP1.

4) Evaluation: Using a confusion matrix allows for evaluating the performance of the ERAPNet algorithm. By comparing the predicted and actual values of the classification model, the accuracy of the

algorithm can be determined, and any areas needing improvement can be identified.

2.2. Question that Arises

How can the information obtained from using ERAPNet for the detection of ERAP1 be applied to further our understanding of ERAP1's role in disease and inform the development of targeted therapies?

2.3. Purpose of This Research

The purpose of this experiment is to use ERAPNet, a machine learning-based approach that uses spectral imaging and deep learning algorithms, to accurately detect the presence of ERAP1 in samples of human cells and tissues. The algorithm's performance can be evaluated using a confusion matrix, which can help researchers assess the accuracy and efficiency of the approach. Ultimately, this experiment aims to improve our understanding of ERAP1 and its role in disease, which can inform the development of targeted therapies.

3. Hypothesis/Expectations

ERAPNet can accurately detect and quantify the presence of ERAP1 in human cells and tissues, and the generated confusion matrix can provide insights into the performance and limitations of the algorithm. This information can contribute to a better understanding of ERAP1's role in disease and potentially lead to the development of more effective targeted therapies.

4. Results and Graphs

Confusion Matrix 1

The confusion matrix analysis in Figure 1 reveals the model's proficiency in predicting various protein categories. In the HYDROLASE category, the model accurately predicts over 70% of instances in the test set, demonstrating a strong performance. Similarly, the model shows over 70% accuracy in the Immune System category. Moving along the diagonal, the model achieves above 70% accuracy for OXIDOREDUCTASE instances, while LYASE instances are accurately classified at over 60%.

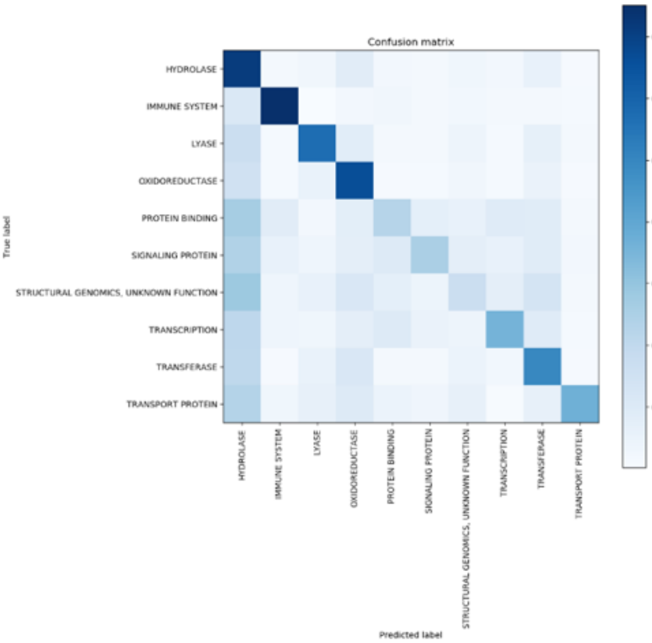

Figure 1: Confusion Matrix of the Classification Model

Training and Validation 1

Figure 2 illustrates the performance of a machine learning model during training. The training accuracy is notably higher than the validation accuracy (approximately 0.5 to 0.6), indicating potential overfitting to the training data. Initially, the training loss decreases before plateauing, suggesting effective learning from the training data. However, the validation loss increases after about 5 epochs, indicating deteriorating performance on the validation data, further suggesting overfitting.

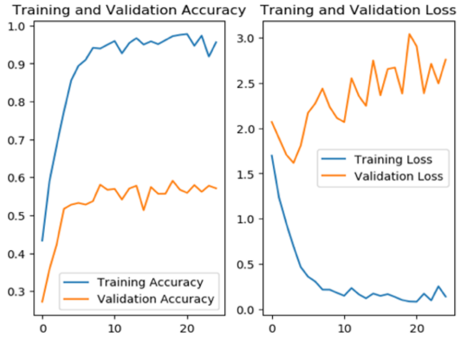

Figure 2: Training and Validation Metrics with accuracy (left) and loss (right) over 25 epochs

Confusion Matrix 2

For the HYDROLASE category in Figure 3, the model in figure 3 achieves over 60% accuracy. In the Immune System category, it reaches over 70% accuracy. The model demonstrates impressive performance in predicting OXIDOREDUCTASE and TRANSFERASE instances, with accuracy over 70%. LYASE instances are predicted with over 50% accuracy, highlighting the model's capability across diverse categories.

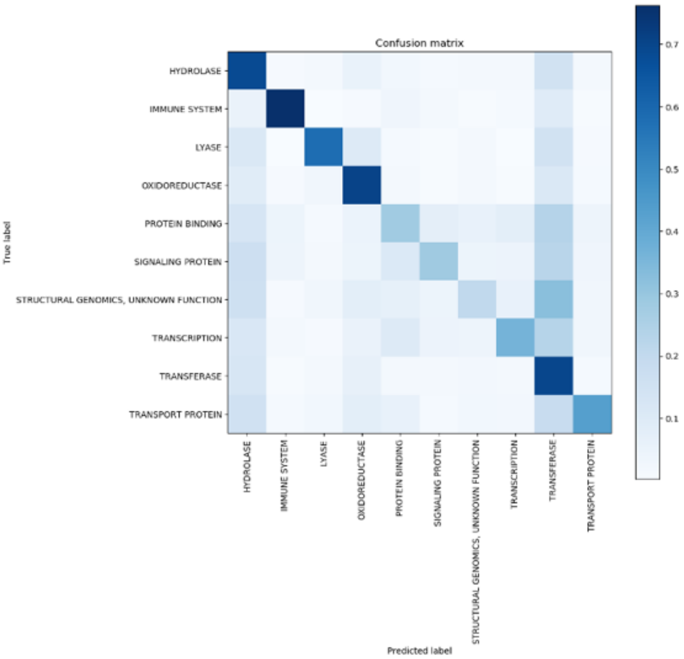

Figure 3: Confusion Matrix of Model Predictions

Training and Validation 2

The training accuracy graph in Figure 4 starts low and increases to around 0.9 after about 10 epochs, indicating effective learning. However, the validation accuracy plateaus at around 0.5 to 0.6, suggesting weaker generalization. The training loss decreases rapidly to below 0.25, while the validation loss initially decreases but starts to increase after 10 epochs, indicating overfitting. The subsequent decrease in validation loss at 20 epochs suggests a slight recovery.

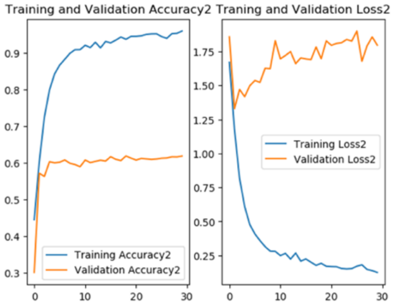

Figure 4: Training and Validation Metrics

Confusion Matrix 3

The model in Figure 5 achieves over 70% accuracy in predicting HYDROLASE instances. It also shows over 70% accuracy in the Immune System category. The accuracy for OXIDOREDUCTASE and TRANSFERASE categories exceeds 70%, and LYASE instances are predicted with over 50% accuracy, demonstrating robustness in diverse categories.

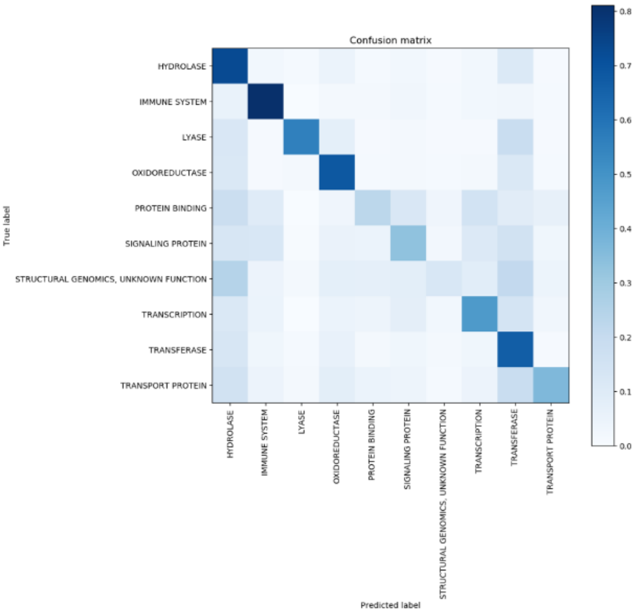

Figure 5: Training and Validation Metrics

Training and Validation 3

Training accuracy shown in Figure 6 increases to about 0.9 after 10 epochs, indicating effective learning. The validation accuracy initially increases to about 0.6 but then plateaus, suggesting limited improvement on unseen data. Training loss decreases steadily to below 0.25, while the validation loss decreases initially but increases after 5 epochs, indicating overfitting.

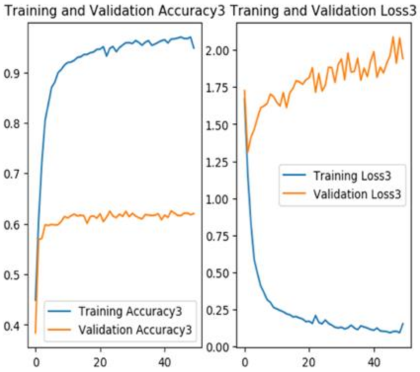

Figure 6: Training and Validation Performance Over 50 Epochs

Confusion Matrix 4

The model in Figure 7 predicts over 60% of HYDROLASE instances accurately. It achieves over 70% accuracy in the Immune System category and over 60% in the TRANSFERASE category, showing reliable predictive capability.

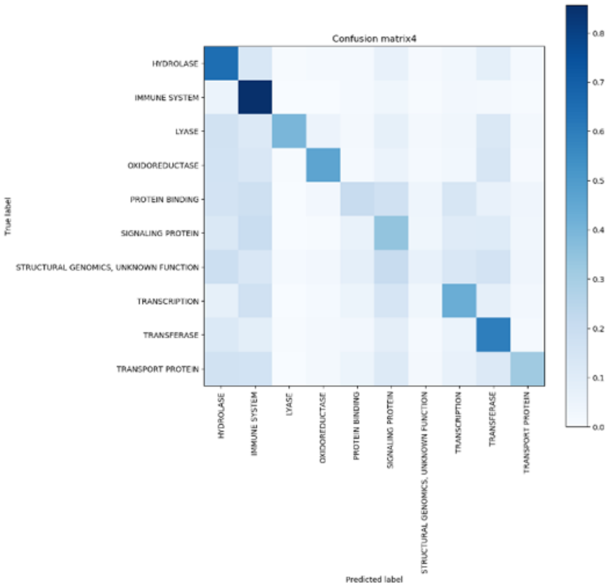

Figure 7: Confusion Matrix of the Classification Model

Training and Validation 4

Training accuracy in Figure 8 increases to around 0.95 in the first 10 epochs and then remains flat, indicating effective learning. The validation accuracy increases to about 0.6 around 5 epochs and remains stable, suggesting reasonable generalization. Training loss decreases significantly to about 0.25, while the validation loss shows an erratic pattern, indicating inconsistent performance on validation data.

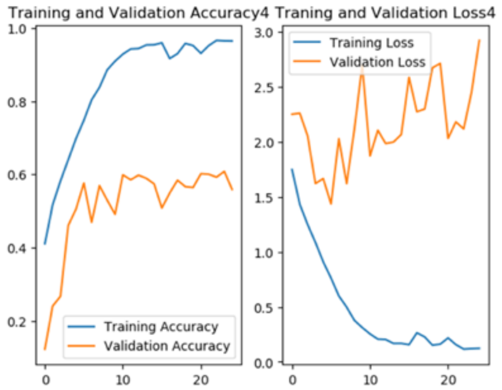

Figure 8: Training and Validation Accuracy and Loss over 25 Epochs

Confusion Matrix 5

The model shown in Figure 9 achieves over 70% accuracy in the HYDROLASE category and over 80% in the Immune System category. It also reaches over 70% accuracy in OXIDOREDUCTASE and TRANSFERASE categories. LYASE instances are predicted with more than 50% accuracy, indicating strong performance across categories.

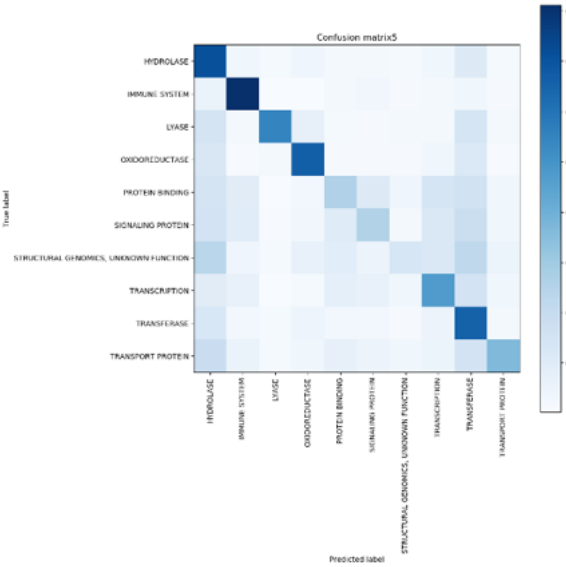

Figure 9: Confusion Matrix of the Classification Model

Training and Validation 5

Training accuracy in Figure 10 increases to about 0.95 and then plateaus, indicating effective learning. Validation accuracy shows fluctuations around 0.6, indicating inconsistent performance. Training loss decreases to about 0.2, while validation loss fluctuates between 2.5 and 3.0, showing inconsistent generalization.

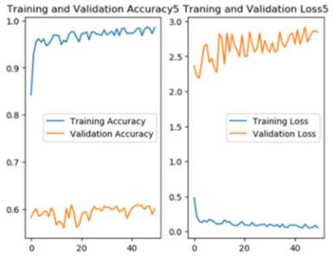

Figure 10: Training and Validation Metrics Over 50 Epochs

5. Discussion

The primary objective of this research was to utilize the ERAPNet model, incorporating both Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architectures, to detect ERAP1 protein within diverse protein families. This study aimed to address the need for an accurate method to identify ERAP1, which plays a crucial role in antigen processing and presentation, particularly in the context of diseases like cancer.

Both CNN and LSTM models demonstrated commendable performance. The CNN model excelled in detecting ERAP1 in spectral images, achieving over 70% accuracy in specific protein categories. The LSTM model, on the other hand, effectively analyzed amino acid sequences, achieving similar accuracy levels in different categories. These outcomes align with the initial expectations, showcasing the models' suitability for their respective data types.

However, the divergence between training and validation metrics suggests potential overfitting, especially in the CNN model. This could be related to the model's architecture or training parameter settings. Fine-tuning these aspects could mitigate overfitting and enhance generalization.

These findings have practical value, as the AI-driven approach can streamline the preliminary screening process, reducing the workload associated with wet experiments. The model's ability to provide valuable reference indicators for protein structure design also supports targeted and efficient research.

Despite the promising results, limitations include the dataset's size and diversity, which may affect the models' generalization to unseen data. Future research should focus on collecting larger and more diverse datasets and fine-tuning model parameters. Incorporating attention mechanisms and deep learning techniques could further improve accuracy.

Overall, the results highlight the complementary strengths of the CNN and LSTM models. The CNN model excels in handling spectral images, while the LSTM model is effective in processing amino acid sequences. Model 3 (LSTM) appears to be the more reliable choice for detecting ERAP1 protein in diverse protein families.

6. Conclusion

This study utilized the ERAPNet model, using both CNN and LSTM architectures, to identify ERAP1 in diverse protein families. The LSTM model demonstrated consistent accuracy in predicting OXIDOREDUCTASE and TRANSFERASE instances, while the CNN model excelled in predicting HYDROLASE and Immune System instances, despite potential overfitting. The findings have significant implications for understanding ERAP1's role in diseases like cancer and can aid in developing targeted therapies.

Future work should explore integrating advanced modules and techniques, such as transfer learning and domain-specific feature extractors, to enhance model performance. Semi-supervised learning approaches could also improve classification accuracy by leveraging unlabeled protein sequences. These improvements will refine predictive models, enhancing their ability to discern nuanced structural patterns within protein families and advancing our understanding of their functional implications.

References

[1] Li, Lenong, et al. "ERAP1 Enzyme-Mediated Trimming and Structural Analyses of MHC I–Bound Precursor Peptides Yield Novel Insights into Antigen Processing and Presentation." The Journal of Biological Chemistry, vol. 294, no. 49, Dec. 2019, pp. 18534–44. PubMed Central, https://doi.org/10.1074/jbc.RA119.010102.

[2] Stratikos, Efstratios, et al. "A Role for Naturally Occurring Alleles of Endoplasmic Reticulum Aminopeptidases in Tumor Immunity and Cancer Pre-Disposition." Frontiers in Oncology, vol. 4, Dec. 2014. Frontiers, https://doi.org/10.3389/fonc.2014.00363.

[3] Bufalieri, Francesca, et al. "ERAP1 Promotes Hedgehog-Dependent Tumorigenesis by Controlling USP47-Mediated Degradation of BTrCP." Nature Communications, vol. 10, no. 1, 24 July 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6656771/, https://doi.org/10.1038/s41467-019-11093-0.

[4] Seffernick, Justin T., and Steffen Lindert. "Hybrid Methods for Combined Experimental and Computational Determination of Protein Structure." The Journal of Chemical Physics, vol. 153, no. 24, 28 Dec. 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7773420/, https://doi.org/10.1063/5.0026025.

[5] Li Z , Miao Q , Yan F ,et al.Machine Learning in Quantitative Protein–peptide Affinity Prediction: Implications for Therapeutic Peptide Design[J].Current Drug Metabolism, 2019.DOI: 10.2174/ 1389200219666181012151944.

[6] Kufareva, Irina, and Ruben Abagyan. "Methods of Protein Structure Comparison." Methods in Molecular Biology (Clifton, N.J.), vol. 857, 2012, p. 231. www.ncbi.nlm.nih.gov, https://doi.org/ 10. 1007/978-1-61779-588-6_10.

[7] PDB, RCSB. "PDB101: Learn: Guide to Understanding PDB Data: Methods for Determining Structure." RCSB: PDB-101, 2016, pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/methods-for-determining-structure.

[8] Li, Yang, et al. "Robust and Accurate Prediction of Protein–Protein Interactions by Exploiting Evolutionary Information." Scientific Reports, vol. 11, no. 1, 19 Aug. 2021, p. 16910, www.nature.com/articles/s41598-021-96265-z, https://doi.org/10.1038/s41598-021-96265-z.

[9] Xia, Yuhao, et al. "Multi-Domain and Complex Protein Structure Prediction Using Inter-Domain Interactions from Deep Learning." Communications Biology, vol. 6, no. 1, 1 Dec. 2023, https://doi.org/10.1038/s42003-023-05610-7.

[10] Su T , Hasan S M S , Nahab F B ,et al.Abstract TMP98: Identifying Stroke Patients At Risk For Atrial Fibrillation Using Electronic Health Record Data And Machine Learning[J].Stroke, 2023, 54(Suppl_1):ATMP98-ATMP98.DOI:10.1161/str.54.suppl_1.TMP98.

[11] Bagaria, Anurag, et al. "Protein Structure Validation by Generalized Linear Model Root-Mean-Square Deviation Prediction." Protein Science: A Publication of the Protein Society, vol. 21, no. 2, Feb. 2012, pp. 229–38. PubMed, https://doi.org/10.1002/pro.2007.