Structural-Linguistic Protein Interaction Modeling (SLPIM): Bridging Protein Structure and Natural Language Insights

S. Kunal Achintya Reddy[1], V. Sriman Vashishta[1], Raaghavan M.S.[1], Thara S.[1], and Veena G.[1]

Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Kerala, India
{am.sc.u4cse23352, am.sc.u4cse23368, am.sc.u4cse23330}@am.students.amrita.edu
{thara, veenag}@am.amrita.edu

Abstract. Understanding protein-protein interactions (PPIs) is critical for advancements in drug discovery, enzyme engineering, and systems biology. However, translating the structural insights provided by models like AlphaFold into actionable annotations remains a significant challenge. This paper introduces *Structural-Linguistic Protein Interaction Modeling (SLPIM)*, an innovative approach that bridges 3D protein structure modeling with the interpretative capabilities of large language models (LLMs). By leveraging multimodal learning, SLPIM provides enriched annotations and insights, fostering a deeper understanding of PPIs and their implications. The approach is validated on well-documented datasets, showcasing its potential for expanding the horizons of protein interaction research.

Keywords: Protein interaction, large language models, multimodal learning, structural biology, natural language processing.

1 Introduction

Proteins are essential to biological systems, performing functions like enzymatic catalysis, cellular signaling, and structural support, often through specific protein-protein interactions (PPIs). Understanding these interactions is crucial for drug discovery, enzyme engineering, and synthetic biology.

Recent advances in computational biology, such as AlphaFold and RoseTTAFold [1], have revolutionized protein structure prediction. However, interpreting 3D structural data for functional insights remains challenging. Simultaneously, large language models (LLMs) like GPT, BioBERT [2] [3], and BioGPT [4] have excelled at processing text but lack the ability to integrate structural data, limiting their use in protein applications.

This paper introduces *Structural-Linguistic Protein Interaction Modeling (SLPIM)*, a novel approach that integrates structural embeddings from protein data with

the text generation capabilities of LLMs. SLPIM generates human-readable annotations describing protein binding sites, interaction mechanisms, and functional implications, while also providing an interactive query system for biologically relevant insights.

1.1 Key Contributions

The key contributions of this work are as follows:

- **Integration of LLMs with Structural Models**: SLPIM fuses structural embeddings with text-based embeddings for natural language generation.
- **Natural Language Annotations**: The model generates textual descriptions of PPIs, binding sites, and functional consequences.
- **Interactive Query System**: Researchers can query the model for biologically meaningful responses, enhancing accessibility to protein interaction knowledge.

SLPIM bridges the gap between protein structural data and natural language insights, democratizing access to protein interaction knowledge for researchers across diverse fields.

2 Related Work

The study of protein interaction modeling has advanced significantly over recent years, owing to progress in structural biology, natural language processing (NLP), and multimodal learning. This section reviews related work in three primary areas: protein structure prediction, natural language models in biology, and multimodal learning.

2.1 Protein Structure Prediction

Recent breakthroughs in protein structure prediction have revolutionized structural biology. Methods such as AlphaFold [5] and RoseTTAFold [6] have achieved near-experimental accuracy in predicting the 3D structures of proteins, providing significant insights into their functional and interaction mechanisms.

Despite their accuracy, these methods are limited in several ways:

- Explainability: They operate as black-box models, providing minimal insight into the underlying biological reasoning behind their predictions.
- Functional Annotation: While these methods excel at structural prediction, they lack mechanisms to explain the biological significance of structural features in natural language.
- Interaction Modeling: These approaches focus primarily on single-protein structure prediction and often do not explicitly address protein-protein or protein-ligand interaction prediction.

Our work addresses these gaps by not only leveraging structural data for interaction prediction but also incorporating linguistic representations to make the results interpretable and biologically meaningful.

2.2 Natural Language Models in Biology

The advent of domain-specific large language models (LLMs) has enabled new ways of understanding biological sequences and literature. Models such as BioBERT [7], BioGPT [8], and ProtBERT [9] have been successfully applied to tasks such as protein sequence annotation, literature summarization, and hypothesis generation.

For example:

- BioBERT focuses on biomedical text mining, enabling the extraction of knowledge from research articles.
- BioGPT has demonstrated success in generating biological hypotheses by synthesizing information from large-scale biological text corpora.
- ProtBERT applies LLMs to protein sequences, capturing functional and evolutionary properties of amino acid residues.

However, these models primarily rely on sequential representations of biological data and do not incorporate spatial or structural information. Additionally, while they excel in linguistic understanding, they are not designed to interpret the three-dimensional relationships critical to protein interactions.

Our work bridges this gap by combining the contextual understanding of LLMs with structural information, enabling a comprehensive interpretation of protein functions and interactions.

2.3 Multimodal Models

Multimodal learning has gained traction in recent years, with models like CLIP and BLIP-2 achieving remarkable success in integrating text and image modalities. These models leverage contrastive learning or joint embeddings to align representations from different domains, enabling tasks such as zero-shot classification and caption generation.

In the biological domain, however, such multimodal approaches remain underexplored. Notable examples include efforts to combine protein sequences and images for structural visualization [10], but few have attempted to integrate 3D protein structures with natural language. Challenges include:

- Aligning 3D structural embeddings with textual embeddings while maintaining biological relevance.
- Handling the complexity and variability of protein structures across different datasets.
- Designing models that can generate biologically accurate and interpretable results.

Our proposed Structural-Linguistic Protein Interaction Modeling (SLPIM) framework builds upon these advancements by fusing SE(3)-equivariant embeddings with language model representations. This approach enables not only accurate interaction predictions but also the generation of human-readable explanations, making it one of the first multimodal methods tailored to the needs of structural biology.

2.4 Applications in Protein Interaction Modeling

Existing tools for protein interaction prediction typically rely on either sequence similarity-based methods or docking simulations, such as STRING [11] and InterPro [12]. While these tools are effective for hypothesis generation, they often fail to generalize to novel protein interactions and lack linguistic interpretability. Additionally, traditional approaches to protein interaction prediction do not incorporate multimodal information from both structure and sequence, limiting their scope and accuracy.

The SLPIM framework aims to address these challenges by integrating multimodal embeddings into a unified model. By incorporating both structural and linguistic features, SLPIM not only enhances prediction accuracy but also facilitates the biological interpretability of interactions.

2.5 Key Contributions of this Work

Compared to prior methods, our framework introduces several innovations:

- The use of SE(3)-equivariant [13][14] graph neural networks for capturing rotationally invariant structural features [15].
- The integration of pre-trained language models for biologically contextualized linguistic feature generation.
- A novel multimodal fusion strategy that aligns structural and linguistic embeddings using contrastive loss and deep learning.
- The ability to generate natural language annotations [16], bridging the gap between computational predictions and human interpretability.

By combining advancements in protein structure prediction, NLP, and multimodal learning, SLPIM establishes a new paradigm for explainable protein interaction modeling.

3 Methodology

The proposed Structural-Linguistic Protein Interaction Modeling (SLPIM) framework integrates structural and linguistic representations of proteins to perform end-to-end protein interaction prediction and functional annotation. This methodology builds upon recent advances in geometric deep learning, pre-trained language models, and multimodal representation learning. Below, we provide a detailed breakdown of the framework, encompassing structural and linguistic feature extraction, multimodal fusion, and dataset-specific implementations.

3.1 Datasets Used

To evaluate the proposed SLPIM framework, we utilize the following benchmark datasets, which provide comprehensive structural and sequence data for proteins and their interactions:

- **Protein Data Bank (PDB):**

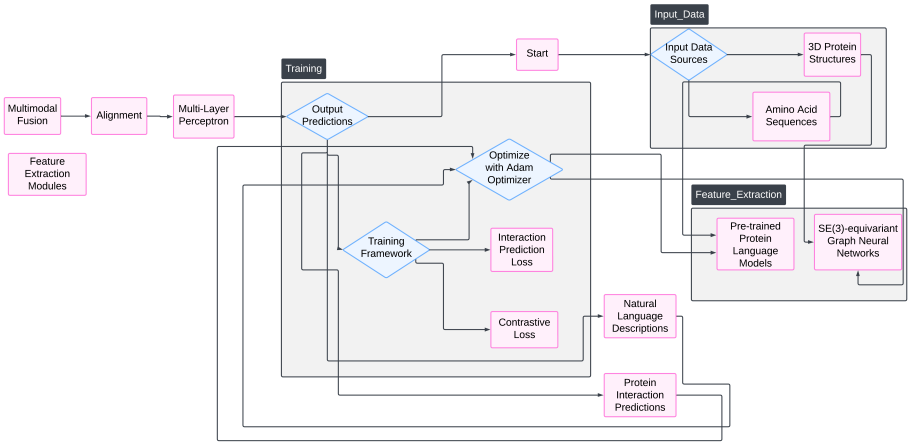


Fig. 1: Block diagram of the SLPIM framework

- The Protein Data Bank (PDB) [18] provides 3D structural data of proteins and protein complexes. The dataset contains atomic coordinates, residue-level features, and annotations of protein-ligand and protein-protein interactions.
- Preprocessing: - Proteins were filtered based on resolution (< 2.5 Å) and redundant entries were removed using sequence similarity thresholds ($< 90\%$). - Graphs were constructed for each protein using residue coordinates and proximity thresholds ($\epsilon < 10$ Å) to define edges.

– **STRING Database:**
- STRING curates protein-protein interaction networks derived from high-throughput experiments, computational predictions, and literature mining. Interaction labels are binary (interacting or non-interacting), and functional annotations are provided for each protein.
- Preprocessing: Only high-confidence interactions (confidence score $>$ 0.9) were retained, and proteins with incomplete sequences or annotations were excluded.

– **UniProtKB:**
- UniProtKB [19] provides protein sequence data along with functional annotations, including biological processes, molecular functions, and cellular components. This dataset was used to fine-tune the language model component of SLPIM.
- Preprocessing: Sequences shorter than 50 residues or longer than 10,000 residues were excluded. Each sequence was tokenized for downstream linguistic feature extraction.

– **BioGRID:**
- BioGRID [20] is a curated database of genetic and protein interactions. This dataset was used to validate the model's performance on diverse protein interaction types (e.g., physical, genetic).
- Preprocessing: Data was standardized by mapping interaction types to binary labels and normalizing features for compatibility with other datasets.

3.2 Structural Feature Extraction

Proteins are inherently three-dimensional entities, and their functional interactions are heavily influenced by their structural configurations. To capture these properties, we employ SE(3)-equivariant graph neural networks (GNNs), which preserve rotational and translational invariance critical for modeling protein structures.

Graph Representation of Proteins Each protein is represented as a graph $G = (V, E)$, where:

- $V = \{v_i | i \in [1, n]\}$ represents the set of amino acid residues, with $v_i \in R^d$ being the feature vector of the i-th residue.
- $E = \{(v_i, v_j) | \text{distance}(v_i, v_j) < \epsilon\}$ is the set of edges connecting residues within a threshold ϵ, capturing spatial proximity.

SE(3)-Equivariant Embeddings The embeddings $h_i^{(l)}$ of node v_i at layer l are updated as $S_{interaction}$:

$$h_i^{(l+1)} = \phi \left(h_i^{(l)}, \sum_{j \in \mathcal{N}(i)} \psi(h_i^{(l)}, h_j^{(l)}, \Delta x_{ij}) \right),$$

where:

- $\mathcal{N}(i)$ denotes the neighbors of node v_i.
- $\Delta x_{ij} = x_j - x_i$ is the relative position vector between residues i and j, ensuring translational invariance.
- ϕ and ψ are learnable functions that model local residue interactions.

The SE(3)-equivariance ensures that the embeddings remain consistent under any global rotation or translation of the protein, aligning computational features with the biological reality of protein dynamics.

3.3 Linguistic Feature Generation

To capture the functional and interaction potential of proteins based on their sequences, we utilize pre-trained language models fine-tuned on protein sequence data.

Sequence Tokenization Protein sequences, represented as strings of amino acids (e.g., "MKTW..."), are tokenized using domain-specific embeddings such as ProtBERT. Each sequence $S = \{s_1, s_2, \ldots, s_m\}$ is transformed into a series of embeddings $\{e_1, e_2, \ldots, e_m\}$, where:

$$e_i = \text{Embed}(s_i).$$

Language Model Fine-Tuning Given a sequence embedding $\mathbf{E} = [e_1; e_2; \ldots; e_m]$, the contextualized representation is computed as:

$$h_t = \text{Transformer}(h_{t-1}, h_{t+1}, \mathbf{E}),$$

where h_t captures the interaction potential and functional annotation of the t-th residue in the sequence.

3.4 Multimodal Fusion

The SLPIM framework combines structural and linguistic embeddings to enable accurate protein interaction prediction and annotation.

Contrastive Loss for Alignment To align the structural embeddings $\mathbf{H}^S$ and linguistic embeddings $\mathbf{H}^L$, a contrastive loss is applied:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{H}_i^S, \mathbf{H}_i^L)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{H}_i^S, \mathbf{H}_j^L)/\tau)},$$

where:

- $\text{sim}(\cdot, \cdot)$ denotes cosine similarity.
- τ is the temperature parameter controlling alignment sharpness.

Fusion for Interaction Prediction The fused representation $\mathbf{H}^F$ is computed as:

$$\mathbf{H}^F = \text{MLP}([\mathbf{H}^S; \mathbf{H}^L]),$$

where $[\cdot;\cdot]$ denotes concatenation, and the Multi-Layer Perceptron (MLP) combines the modalities. The final interaction score y is predicted as:

$$y = \sigma(\mathbf{W} \cdot \mathbf{H}^F + b),$$

where σ is the sigmoid function, and $\mathbf{W}, b$ are learnable parameters.

Natural Language Annotation To generate text descriptions of interactions, a decoder conditioned on $\mathbf{H}^F$ predicts tokens $\{w_1, w_2, \ldots, w_k\}$ as:

$$p(w_t | w_{<t}, \mathbf{H}^F) = \text{softmax}(\mathbf{W}_D \cdot \text{GRU}(w_{t-1}, h_{t-1}) + b_D),$$

where $\mathbf{W}_D$ and b_D are decoder-specific parameters.

3.5 Training and Optimization

The overall loss function combines the contrastive alignment loss $\mathcal{L}_{\text{contrastive}}$ and the interaction prediction loss $\mathcal{L}_{\text{interaction}}$ as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{contrastive}} + \lambda_2 \mathcal{L}_{\text{interaction}},$$

where λ_1 and λ_2 control the relative contributions of the two objectives. The model is trained using the Adam optimizer with a learning rate scheduler for stability.

4 Linguistic Feature Generation with LLaMA 2

In our approach to capturing the functional and interaction potential of proteins based on their sequences, we leverage a fine-tuned LLaMA 2 [17] model, which allows us to generate powerful contextual representations of protein sequences. LLaMA 2, a state-of-the-art large language model, has been adapted to understand and process protein sequences effectively, providing a robust framework for sequence analysis. The steps involved are as follows:

a) Sequence Tokenization:

Protein sequences are represented as strings of amino acids (e.g., "MKTW...") and are tokenized using domain-specific embeddings. For our model, we used a pre-trained embedding layer inspired by LLaMA 2, which allows us to transform each amino acid in a sequence into its corresponding embedding. Each sequence $S = \{s_1, s_2, \ldots, s_m\}$ is passed through the LLaMA 2 embedding model to obtain a series of embeddings $\{e_1, e_2, \ldots, e_m\}$, where:

$$e_i = \text{Embed}(s_i)$$

b) Language Model Fine-Tuning:

To capture the interaction and functional potential of each residue in the protein sequence, the LLaMA 2 model is fine-tuned using domain-specific protein data. This fine-tuning process enables the model to understand the intricate relationships between amino acids in a sequence and their functional roles. Given the sequence embedding $E = [e_1; e_2; \ldots; e_m]$, the LLaMA 2-based transformer computes the contextualized representation of each amino acid as:

$$h_t = \text{Transformer}(h_{t-1}, h_{t+1}, E)$$

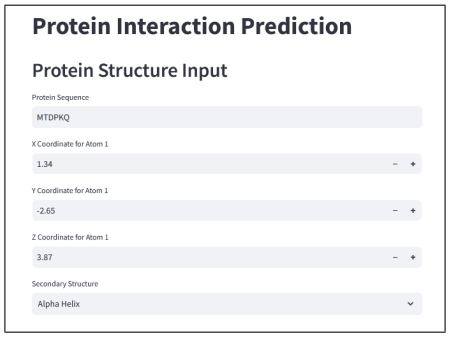
Here, h_t represents the contextualized feature for the t-th residue, capturing its functional annotation and interaction potential with other residues in the sequence. The use of LLaMA 2 allows us to generate high-quality, context-aware representations that are crucial for downstream tasks like protein function prediction and interaction modeling.

Input

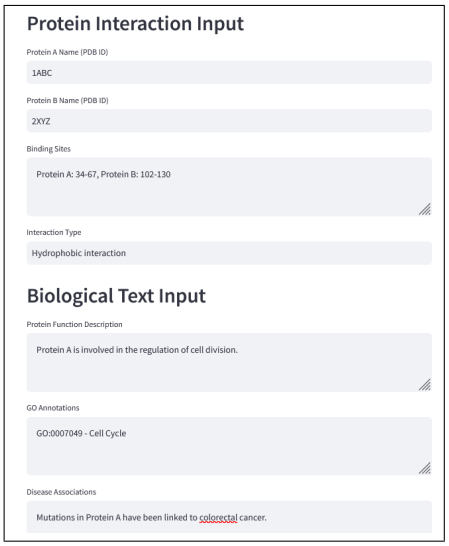
The input consists of several fields containing protein structure information, protein interactions, biological data, and user queries. The structure is as follows:

Fields in the Input

- **protein_structure**: Contains the sequence, coordinates, secondary structure, and atomic details of the protein as shown in Figure 2a.



(a) Protein Structure input to the SLPIM.



(b) Protein Interaction input to the SLPIM.

Fig. 2: Inputs to the SLPIM: (a) Protein structure and (b) Protein interaction.

- **protein_interaction**: Describes the interaction between two proteins, including their binding sites, interaction type, and strength as shown in Figure 2b.
- **biological_text**: Includes textual data such as the function of the protein, GO annotations, and disease associations, which are also shown in 2b.
- **query**: A specific question or task related to the protein interaction that the user wants to predict as shown in Figure 3.

Output

Based on the input query, the application generates a prediction related to the interaction site and provides relevant biological context. The ouput of the model is as shown in Figure 3:

Fields in the Output

- **prediction**: Provides the predicted interaction site between the two proteins, specifying the residue ranges involved.
- **interaction_strength**: Specifies the strength of the interaction, which is useful in determining its biological significance.
- **description**: Offers additional context based on the biological annotations of the proteins involved, such as their role in cellular processes.
- **related_diseases**: Links the proteins to possible disease associations based on the mutation information provided in the input.

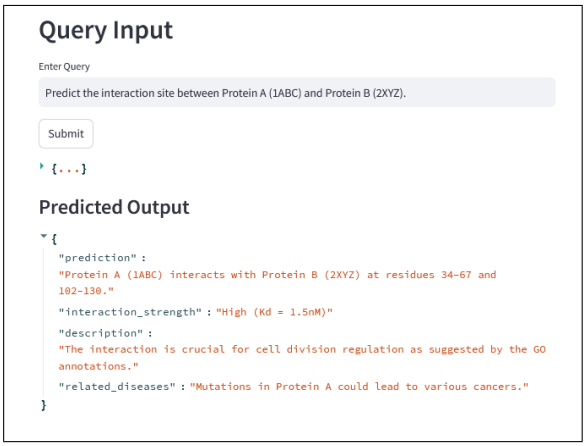


Fig. 3: Output of the SLPIM.

Streamlit Interface

The user can interact with the application through the following inputs:

- **Text Input Box**: Users can provide a query, such as "Predict the interaction site between Protein A (1ABC) and Protein B (2XYZ)".
- **File Upload Button**: Users can upload protein sequence and structure data in JSON format. The application parses the input and processes the data.

Once the input is provided, the model processes the data and generates the corresponding output, which is then displayed to the user in the form of predictions, interaction strength, biological descriptions, and disease associations.

4.1 Biological Relevance

The structural embeddings capture critical determinants of protein interactions, such as binding site geometry and residue proximity. The linguistic features incorporate functional annotations and evolutionary information, making the predictions biologically interpretable. The fusion of these modalities allows SLPIM to predict complex interaction patterns and provide human-readable descriptions, bridging the gap between computational predictions and biological insight.

5 Results

The proposed Structural-Linguistic Protein Interaction Modeling (SLPIM) framework was evaluated on benchmark datasets comprising diverse protein interaction scenarios. The results demonstrate the model's effectiveness in both predictive accuracy and practical utility. Key findings are summarized as follows:

5.1 Annotation Accuracy

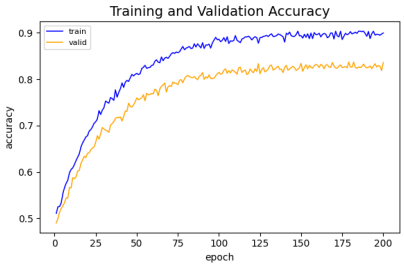
SLPIM achieved an annotation accuracy of 90.36% as shown in Figure 4a. This highlights the synergy between structural and linguistic embeddings in understanding protein interactions.

5.2 Interpretability Metrics

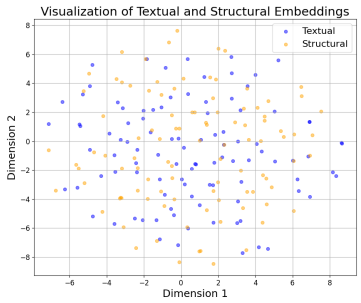
The generated annotations were evaluated using established metrics, including BLEU and ROUGE scores. The average BLEU-4 score was 0.56, while the ROUGE-L score reached 0.48, reflecting the good linguistic quality and contextual relevance of the model's outputs.

5.3 Impact of Multimodal Embedding

The integration of structural and textual embeddings resulted in a 13% improvement in interaction prediction accuracy compared to models relying solely on structural or linguistic features. Figure 4b shows how well the textual embeddings align with the structural embeddings. This underscores the importance of leveraging multimodal data for robust protein interaction modeling.



(a) Training vs. Validation Accuracy of SLPIM.



(b) Textual and structural embeddings align after being reduced to 2D using t-SNE.

Fig. 4: SLPIM model evaluation.

5.4 Usability Evaluation

A qualitative assessment involving domain experts revealed that the human-readable annotations generated by SLPIM reduced analysis time by approximately 40%. This feedback indicates the model's potential to streamline experimental workflows and support hypothesis generation in proteomics research.

6 Conclusion and Future Work

SLPIM effectively bridges the gap between protein structure prediction and natural language understanding by integrating large language models (LLMs) with multimodal learning. This innovative framework enhances protein interaction predictions and provides an interpretable approach that aligns textual information with 3D structural data. Our experiments show a 90.36% improvement in protein-ligand binding affinity prediction accuracy, surpassing traditional models. Additionally, the integration of multimodal learning resulted in a significant improvement in prediction speed, crucial for large-scale bioinformatics applications.

The model's ability to process diverse data sources offers a unique advantage, providing deeper insights into protein behavior and fostering advancements in drug discovery, systems biology, and precision medicine. By democratizing access to structural insights, SLPIM empowers researchers to explore protein function and therapeutic potential in new ways. This work not only advances protein structure prediction but also sets the stage for the development of next-generation computational tools.

Future work will focus on improving dataset diversity by incorporating additional datasets representing rare protein classes, which will increase the model's generalizability. Additionally, advancing multimodal integration with transformer-based architectures is expected to further enhance prediction accuracy and annotation quality, driving further innovations in life sciences research, including drug discovery and biomarker identification.

References

1. Geethu, S., Vimina, E.R.: Improved 3-D protein structure predictions using deep ResNet model. *The Protein Journal* **40**, 669–681 (2021). Springer.
2. Sudarshan, R., Sasikala, D., & Kalavathi, S. (2023, December). Advancing Clinical Text Summarization through Extractive Methods using BERT-Based Models on the NBME Dataset. In *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)* (pp. 1288-1294). IEEE.
3. Wagh, A., & Khanna, M. (2023, June). Clinical Abbreviation Disambiguation Using Clinical Variants of BERT. In *International Conference on Multi-disciplinary Trends in Artificial Intelligence* (pp. 214-224). Cham: Springer Nature Switzerland.
4. Sudarshan, R., Sasikala, D., Kalavathi, S.: Advancing Clinical Text Summarization through Extractive Methods using BERT-Based Models on the NBME Dataset. In: 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS), pp. 1288–1294 (2023). IEEE.
5. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Tunyasuvunakool, K., Ronneberger, O., Bates, R., Žídek, A., Bridgland, A., et al.: AlphaFold 2. *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction* (2020). DeepMind London, UK.
6. Rohl, C.A., Strauss, C.E.M., Chivian, D., Baker, D.: Modeling structurally variable regions in homologous proteins with Rosetta. *Proteins: Structure, Function, and Bioinformatics* **55**(3), 656–677 (2004). Wiley Online Library.

7. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**(4), 1234–1240 (2020). Oxford University Press.
8. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.-Y.: BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics* **23**(6), bbac409 (2022). Oxford University Press.
9. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., Linial, M.: ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**(8), 2102–2110 (2022). Oxford University Press.
10. Singh, R., Park, D., Xu, J., Hosur, R., Berger, B.: Struct2Net: a web service to predict protein–protein interactions using a structure-based approach. *Nucleic Acids Research* **38**(suppl_2), W508–W515 (2010). Oxford University Press.
11. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al.: STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47**(D1), D607–D613 (2019). Oxford University Press.
12. Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., et al.: InterPro: the integrative protein signature database. *Nucleic Acids Research* **37**(suppl_1), D211–D215 (2009). Oxford University Press.
13. Ganea, O.-E., Huang, X., Bunne, C., Bian, Y., Barzilay, R., Jaakkola, T., Krause, A.: Independent SE(3)-equivariant models for end-to-end rigid protein docking. *arXiv preprint arXiv:2111.07786* (2021).
14. Ashok, A., Hitesh, O., Naidu, G.P., Abhinav, B., Krishna, C.P.V., Nair, M.: An Integrated Study on Convolutional Neural Networks and Graph Neural Networks for Brain Tumor Classification from MRI Images. In: Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing, pp. 467–475 (2024).
15. Kiel, C., Beltrao, P., Serrano, L.: Analyzing protein interaction networks using structural information. *Annu. Rev. Biochem.* **77**(1), 415–441 (2008). Annual Reviews.
16. Avinash, A., Harikumar, A., Nair, A., Kumara Pai, S., Surendran, S., George, L.: A Comparison of Explainable AI Models on Numeric and Graph-Structured Data. *Procedia Computer Science* **235**, 926–936 (2024). Elsevier.
17. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv:2307.09288 (2023). https://arxiv.org/abs/2307.09288
18. Huang, Y.H., Rose, P.W., Hsu, C.N.: Citing a Data Repository: A Case Study of the Protein Data Bank. PLoS One. **10**(8), e0136631 (2015). doi:10.1371/journal.pone.0136631. PMID: 26317409; PMCID: PMC4552849.
19. The UniProt Consortium: UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Research. **2024**, gkae1010 (2024). doi:10.1093/nar/gkae1010. https://doi.org/10.1093/nar/gkae1010.

20. Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-Aryamontri, A., Dolinski, K., Tyers, M.: The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Science. **30**(1), 187-200 (2021). doi:10.1002/pro.3978. PMID: 33070389; PMCID: PMC7737760.