MS in Bioengineering
BIO467 Scientific Literature Analysis in Bioengineering

Ecole polytechnique fédérale de Lausanne

## Individual Literature Search Report (no more than 3 pages)

Name of the student:

[redacted]

Topic:

Predicting protein-protein interaction with machine learning

**Briefly summarize the state of the art of the field you chose:**

Protein-protein interactions (PPIs) are essential for cellular processes, and their accurate prediction is crucial for understanding biological mechanisms, identifying disease biomarkers, and advancing targeted therapeutics. Experimental techniques, while instrumental, are often time-consuming, costly, and prone to errors. To address this, computational approaches have emerged as scalable and cost-effective alternatives.

Early methods relied on classical machine learning (ML) techniques such as support vector machines (SVMs) and random forests (RF), which required extensive manual feature engineering. The emergence of deep learning (DL) has transformed PPI prediction by enabling models to extract complex features automatically from raw protein data, such as sequences and structures.

State-of-the-art sequence-based DL models extract features from raw protein sequences or embeddings obtained from natural language processing. DELPHI does this utilizing convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [1]. Moreover, models leveraging attention mechanisms have allowed models to focus on critical regions of protein sequences, as seen in SDNN-PPI [2].

Meanwhile, structure-based models leverage 3D protein structural data, as seen in Struct2Graph , which applies graph attention networks to structures from the Protein Data Bank [3]. Recent advancements, have further improved performance by incorporating structural data from AlphaFold (AF) [4], which expanded the available dataset to over 200 million protein structures. For instance, SGPPI employs a Siamese network architecture with graph convolutional networks (GCNs) to identify binding interfaces to predict PPIs [5]. Bridging the gap between the two, hybrid models combine sequence and structure features to achieve superior accuracy. For example, GraphPPIS integrates evolutionary information and spatial structural features using GCN [6].

**List some major publications, groups working on the topic, and demonstrations:**

Recent advancements in protein-protein interaction (PPI) prediction have introduced innovative machine learning frameworks tailored to both structure-based and hybrid approaches. "*SGPPI: Structure-based deep learning framework for PPI prediction*" [5] employs graph convolutional networks (GCNs) to focus on binding interfaces, leveraging AlphaFold2 (AF2) monomer structures. This enables accurate predictions for novel proteins with the added benefit of cross-species transferability. Similarly, "*MEG-PPIS: a fast protein–protein interaction site prediction method based on multi-scale graph information and equivariant graph neural network*" [7] incorporates E(n)-equivariant graph neural networks (EGNNs) to explicitly handle 3D spatial symmetries of protein structures. By integrating multi-scale graph information, MEG-PPIS captures both global and local structural features, addressing limitations seen in earlier models. Additionally, "*De novo design of protein interactions with learned surface fingerprints*" [8] demonstrates the ability to computationally engineer protein interactions using surface fingerprint features. Meanwhile, "*Accurate structure prediction of biomolecular interactions with AlphaFold 3*" [9] represents a major breakthrough, extending beyond structure prediction to accurately model the joint structure of complexes for not only PPIs but also other biomolecular interactions, such as with DNA, RNA, and small molecules. For an overview of models, challenges, and trends in the field, "Machine learning on protein–protein interaction prediction: models, challenges, and trends" provides a comprehensive perspective.

Several leading research groups and institutions have made significant contributions to PPI prediction using machine learning. Google DeepMind has been at the forefront, particularly with its groundbreaking AlphaFold models. Chinese institutions, such as China Agricultural University and Tsinghua University, have published a substantial number of relevant papers in the field. The Laboratory of Protein Design and Immunoengineering, led by Bruno Correia at EPFL, focuses on machine learning approaches to study protein surfaces and interactions. At the University of Washington, the Institute for Protein Design (David Baker) extends its work beyond protein design, with numerous contributions to PPI research. Similarly, the Laboratory of Computational Structural and Systems Biology at Seoul National University has advanced computational methods for studying protein interactions.

Several studies showcase the practical applications of machine learning in predicting and analyzing PPIs. For example, in "Discovering Protein Interactions and Repurposing Drugs in SARS-CoV-2 Using Machine Learning," [10] researchers used a graph-based machine learning approach to analyze PPIs, genetic data, and virus-host interactions, identifying potential drugs for repurposing against SARS-CoV-2. Another publication, *"Learning to Design Protein-Protein Interactions with Enhanced Generalization,"* [11] introduces PPIformer, a SE(3)-equivariant model pre-trained on PPIRef, a large-scale 3D protein structure dataset. The study uses mutations on protein interactions to enhance the binding affinity of SARS-CoV-2 antibodies. It also demonstrates improving the activity of staphylokinase, a bacterial protein effective in breaking down blood clots, highlighting its potential for therapeutic applications.

**Indicate timeline and statistics on the publication volume (can include figures):**

Over the past two decades, predicting PPIs with ML has evolved significantly. Early approaches relied on classical ML models like SVM and RF, which were effective but limited in scalability. Starting around 2017, the rise of large-scale datasets and high-throughput technologies led to a shift toward deep learning (DL) models such as Deep Neural Networks (DNNs) and CNNs, exemplified by models like DeepPPI [12] and PIPR [13]. Sequence-based methods dominated early research due to the widespread availability of protein sequence data. Structure-based approaches for PPI prediction developed more gradually, as they were constrained by experimentally derived 3D structures. Struct2Graph is an example of such methods, leveraging graph-based techniques [3]. AlphaFold development paved the way to implement more structure-based models, as it dramatically increased the availability of structural data. Consequently, structure-based methods began to gain even more traction from 2021. Around this time, hybrid DL models also started to emerge as a promising approach. A timeline of proposed machine learning models until 2022 is seen in figure 1 [14]. Recent advancements, from 2022 to 2024, include graph-based neural networks like DL-PPI for sequence-based data [15], SGPPI [5], and MEG-PPIS [7] for structure-based models. The release of AlphaFold3 extends beyond structure prediction to accurately model protein-protein interactions [9], and other biomolecular complexes, making it a significant advancement in the field of interaction prediction.

In terms of the publication volume, figure 2 shows a clear upward trend in the number of machine learning-based studies for PPI prediction. Between 2002 and 2010, research output remained relatively low. However, in 2012 there is a noticeable rise, likely driven by advancement in computational power and high-throughput technologies, which increased protein sequence data available. A significant increase in trend is observed from 2015-2017 until now. This aligns with the adoption of DL techniques. The rapid increase from 2021 onward, corresponds to the increase in structural data availability due to AF-3, as well as the post COVID-19 pandemic, which highlighted importance of understanding PPIs.

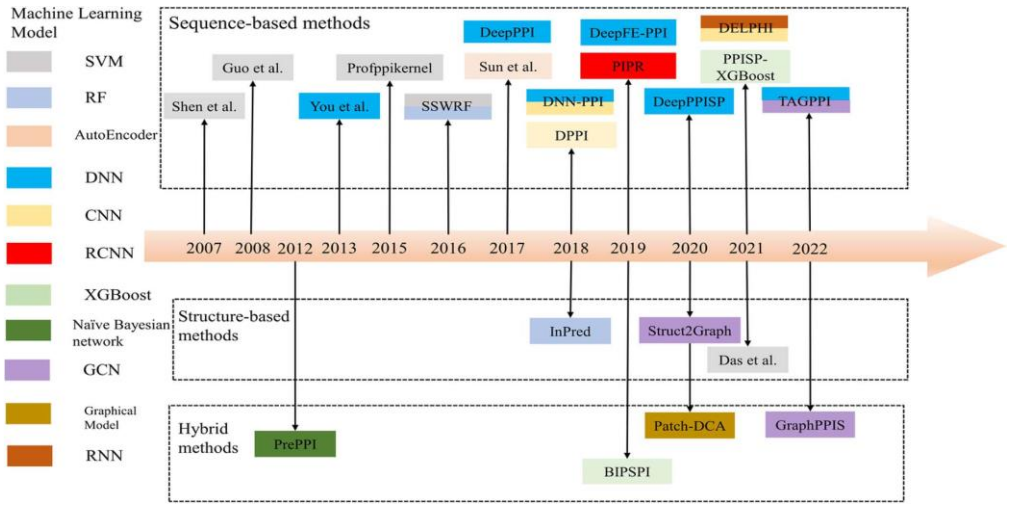


*Figure 1: Timeline of proposed machine learning models until 2022. [14]*

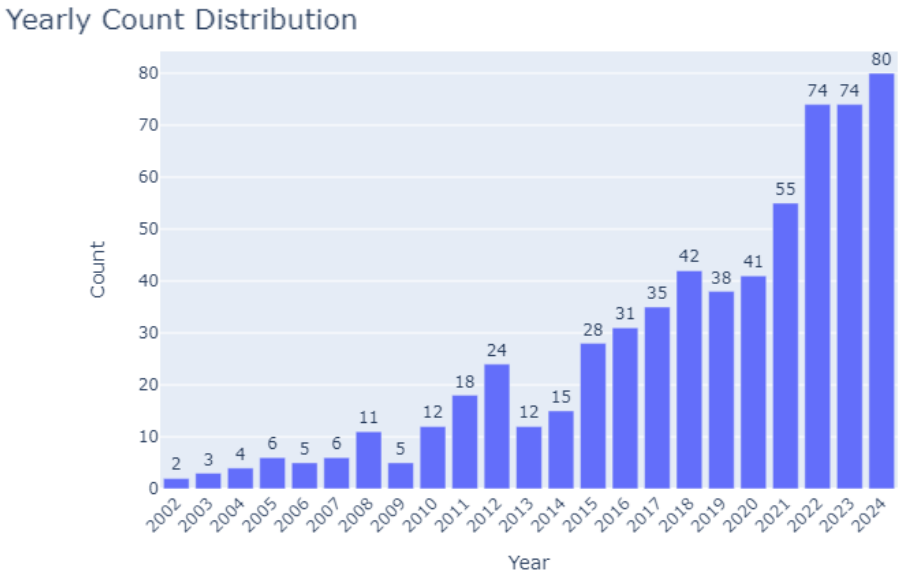


*Figure 2: Bar chart of approximate publication volume per year, extracted from PubMed.*

**Based on your scientific literature search, what are the future directions that you foresee and/or the most important open questions?**

Based on the literature research, there are several feature directions and open questions that are emerging. Firstly, hybrid approaches that combine sequence and structural data will have a pivotal role, especially as AF continues expanding the availability of high-quality structural data. Second, graph-based models need to be improved to better represent more complex protein interaction networks. This is particularly important for addressing the computational limitations encountered when processing large-scale graphs, which may come across in applications such as signaling pathway interaction maps. Some studies have explored approaches that might help, such as graph attention networks.

Another promising direction involves the use of generative models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), to simulate high-quality interaction data, such as synthetic interaction networks.

Moreover, domain-specific application of PPI predictions, such as abnormal PPIs associated with certain diseases like cancer, or neurodegenerative diseases, should remain a critical focus for developing targeted therapies and improving drug discovery.

Finally, there are still open questions regarding the dynamic nature of PPIs, which are influenced by environmental factors and cell cycle phases. Current models often assume static interactions, highlighting the need for approaches that incorporate dynamic modelling.

**BIBLIOGRAPHY**

[1] Y. Li, G. B. Golding, and L. Ilie, 'DELPHI: accurate deep ensemble model for protein interaction sites prediction', *Bioinformatics*, vol. 37, no. 7, pp. 896–904, May 2021, doi: 10.1093/bioinformatics/btaa750.

[2] X. Li, P. Han, G. Wang, W. Chen, S. Wang, and T. Song, 'SDNN-PPI: self-attention with deep neural network effect on protein-protein interaction prediction', *BMC Genomics*, vol. 23, no. 1, p. 474, Jun. 2022, doi: 10.1186/s12864-022-08687-2.

[3] M. Baranwal *et al.*, 'Struct2Graph: a graph attention network for structure based predictions of protein–protein interactions', *BMC Bioinformatics*, vol. 23, no. 1, p. 370, Sep. 2022, doi: 10.1186/s12859-022-04910-9.

[4] J. Jumper *et al.*, 'Highly accurate protein structure prediction with AlphaFold', *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.

[5] Y. Huang, S. Wuchty, Y. Zhou, and Z. Zhang, 'SGPPI: structure-aware prediction of protein–protein interactions in rigorous conditions with graph convolutional network', *Briefings in Bioinformatics*, vol. 24, no. 2, p. bbad020, Mar. 2023, doi: 10.1093/bib/bbad020.

[6] Q. Yuan, J. Chen, H. Zhao, Y. Zhou, and Y. Yang, 'Structure-aware protein-protein interaction site prediction using deep graph convolutional network', *Bioinformatics*, vol. 38, no. 1, pp. 125–132, Dec. 2021, doi: 10.1093/bioinformatics/btab643.

[7] H. Ding *et al.*, 'MEG-PPIS: a fast protein–protein interaction site prediction method based on multi-scale graph information and equivariant graph neural network', *Bioinformatics*, vol. 40, no. 5, p. btae269, May 2024, doi: 10.1093/bioinformatics/btae269.

[8] P. Gainza *et al.*, 'De novo design of protein interactions with learned surface fingerprints', *Nature*, vol. 617, no. 7959, pp. 176–184, May 2023, doi: 10.1038/s41586-023-05993-x.

[9] J. Abramson *et al.*, 'Accurate structure prediction of biomolecular interactions with AlphaFold 3', *Nature*, vol. 630, no. 8016, pp. 493–500, Jun. 2024, doi: 10.1038/s41586-024-07487-w.

[10] X. Li, A. Ovanessians, and H. Wang, 'Discovering Protein Interactions and Repurposing Drugs in SARS-CoV-2 (COVID-19) via Learning on Robust Multipartite Graphs', in *2023 IEEE International Conference on Data Mining (ICDM)*, Dec. 2023, pp. 289–298. doi: 10.1109/ICDM58522.2023.00038.

[11] A. Bushuiev *et al.*, 'Learning to design protein-protein interactions with enhanced generalization', Mar. 16, 2024, *arXiv*: arXiv:2310.18515. doi: 10.48550/arXiv.2310.18515.

[12] X. Du, S. Sun, C. Hu, Y. Yao, Y. Yan, and Y. Zhang, 'DeepPPI: Boosting Prediction of Protein–Protein Interactions with Deep Neural Networks', *J. Chem. Inf. Model.*, vol. 57, no. 6, pp. 1499–1510, Jun. 2017, doi: 10.1021/acs.jcim.7b00028.

[13] M. Chen *et al.*, 'Multifaceted protein–protein interaction prediction based on Siamese residual RCNN', *Bioinformatics*, vol. 35, no. 14, pp. i305–i314, Jul. 2019, doi: 10.1093/bioinformatics/btz328.

[14] 'Machine learning on protein–protein interaction prediction: models, challenges and trends | Briefings in Bioinformatics | Oxford Academic'. Accessed: Dec. 18, 2024. [Online]. Available: https://academic.oup.com/bib/article/24/2/bbad076/7069757

[15] J. Wu, B. Liu, J. Zhang, Z. Wang, and J. Li, 'DL-PPI: a method on prediction of sequenced protein–protein interaction based on deep learning', *BMC Bioinformatics*, vol. 24, no. 1, p. 473, Dec. 2023, doi: 10.1186/s12859-023-05594-5.