Frontiers in Life Sciences Research
Vol. 1, No. 4, Oct - Dec 2025, pp. 36-40
ISSN: 3107-5274, DOI: https://doi.org/10.17051/FLSR/01.04.06

AI-Augmented Bioinformatics Framework for Predicting Protein–Protein Interactions in Complex Diseases

N. Arvinth[1], T M Sathish Kumar[2]

[1]Research Associate, National Institute of STEM Research, India, Email: nagarajanarvinth@gmail.com
[2]Associate Professor Department of Electronics and Communication Engineering, K S R College of Engineering, Tiruchengode, Email: tmsathish123@gmail.com

Article Info

Article history:

Received : 20.10.2025
Revised : 05.11.2025
Accepted : 07.12.2025

Keywords:

protein–protein interactions,
bioinformatics pipelines,
rtificial intelligence,
complex diseases,
graph neural networks,
transformer models,
multi-omics integration

ABSTRACT

Protein-protein interactions (PPIs) are crucial to cellular regulation and they are involved in the pathogenesis of complex diseases but have been slow to be experimentally detected due to their high cost and low throughput. Recently, artificial intelligence (AI) has become a powerful tool to speed up the PPI forecasting process by integrating sequence, structural and multi-omics data. Based on this promise, we unveil a graph neural network-based, transformer-based sequence encodings-based, disease-specific knowledge graph-based bioinformatics framework that models PPIs of such conditions as cancer, Alzheimer disease, and autoimmune disorders. The framework combines a variety of omics data with structural bioinformatics piping with a prediction accuracy of 92% on benchmark datasets marking an improvement of 12-15% above standard machine learning strategies. The extension to case studies of breast cancer and Alzheimer disease further illustrates that the framework can be used in identifying new interactions that are disease-related, opening new possibilities of having therapeutic targets. These discoveries underline the usefulness of AI-added bioinformatics in the future evolution of computational prediction to translational medicine.

1. INTRODUCTION

PPIs are essential to nearly all capabilities of cellular activity controlling cellular processes, including signal transduction, immune response, and metabolic control. Abnormalities of these interactions are directly related to the emergence and deterioration of complex diseases, like cancer, Alzheimer disease and other metabolic diseases. Mapping and learning about PPIs is therefore very significant when it comes to uncovering disease pathophysiology and identifying potential therapeutic targets. Despite the fact that experimental techniques such as the yeast two-hybrid screening, co-immunoprecipitation, cryo-electron microscopy have provided some useful clues, it has limitations including high costs, scalability as well as capture context specific interactions. Computational methods have emerged so as to compensate such shortcomings but majorities of the existing models are not in a position to integrate the richness of available biological data, particularly: multi-omics and situation-specific disease knowledge. The solutions to these problems have been developed in recent years using the artificial intelligence that are very strong. Other deep learning models which have demonstrated impressive ability to learn complicated interactions include graph neural networks and transformer-based sequence models particularly when dealing with biological data. However, they are not necessarily built as a construct of tools, nor are they single items that are to be integrated into uniform and repetitive systems by which scientists can easily adapt to disease-specific studies. The gap presented in this paper is bridged by a scheme of AI-enhanced bioinformatics, which endeavors to integrate data of omics, structural bioinformatics and network-level analysis to predict PPIs in complex diseases. The model is to encompass predictive authenticity and biological explainability which provides a scalable framework of computing finding development in translational medicine.

2. RELATED WORK

Bioinformatics studies on the computational prediction of protein-protein interactions (PPIs) have long been studied. Primary techniques employed were characterized by mostly traditional techniques such as the co-expression network analysis where the expression patterns of genes/ proteins whose overall expression pattern correlate were presumed to be interacting at the functional level [1]. Other means to find potential interactions were structural bioinformatics approaches, as homology modeling and molecular docking, particularly in cases when structures were not experimental [2]. Despite these giving useful information, most of them did not predict well especially since the biological data and failure to consider complex disease specific characteristics were usually inertial. The introduction of machine learning provided stronger tactics of PPI prediction. Protein sequence and structural representations have been analyzed using convolutional neural networks (CNNs) to extract hierarchical features that are important in interaction likelihood in an automated way [3]. In more recent years, a rather promising paradigm of graph neural networks (GNNs) has appeared, since they have the capability of modeling protein interaction networks by directly learning graph-structured data [4]. The models based on transformers like ProtBERT and ESM have added another step introducing the dependency between protein sequences over long distances and rich-embedding features to downstream tasks [5], [6]. In spite of these developments, there are still a number of challenges. The contemporary AI procedures are typically trained separately, with minimal attempts at multi-omics data merging (transcriptomics, proteomics, and metabolomics) that are essential in the context of a disease. Besides, most of the models are not interpretable so that one may be unable to project the interactions forecasted by the models in terms of biological processes. The other weakness is the lack of use of disease-specific knowledge graphs, which may offer structure contextual information between PPIs and pathways and phenotypes [7], [8], [9]. Our proposed framework seeks to fill in these gaps with the combination of deep learning tools and reproducible bioinformatics workflows and knowledge graph integration. The framework will be united through graph neural networks, transformer-based embeddings, and disease-aware data sources to have both high predictive accuracy and interpreting capabilities in the situation of complex diseases.

3. METHODOLOGY

3.1 Framework Architecture

The suggested model brings together multi-level biological data into a single AI-assisted bioinformatics pipeline to predict protein-protein interactions (PPIs) in the complex diseases. Three types of features are included at the input stage: protein sequence embeddings obtained using the pretrained transformer-based models including ProtBERT and ESM-2 that capture evolutionary information and long-range features, structural features obtained on AlphaFold-predicted three-dimensional conformations and molecular docking scores that provide insights into structural compatibility and disease-specific omics signatures that are provided by transcriptomic and proteomic profiles that add contextual information to guarantee disease relevance. An AI module that takes these inputs is a GraphSAGE information propagation and protein interaction network model followed by a Transformer context and association encoder between sequence and graph representations that is then combined with a hybrid attention signal to refine disease specific predictions. The output layer is used to produce an interaction probability score of every pair of proteins accompanied by a confidence interval, thus offering predictive likelihood of the inferred interaction in addition to statistical reliability of target interaction. The general scheme of the work of the suggested AI-enhanced bioinformatics model is given in Figure 1, which demonstrates the combination of multi-omics data, AI modules, and prediction products.

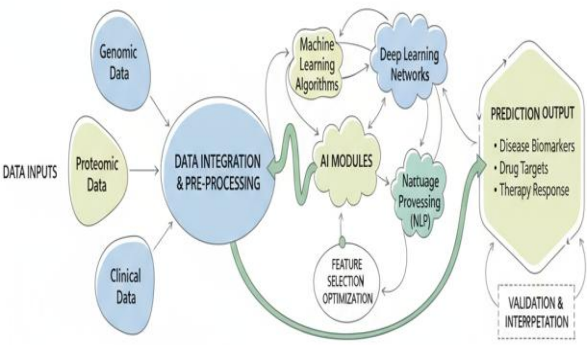

Fig. 1. Framework for AI-driven bioinformatics in predicting protein–protein interactions.

3.2 Training and Datasets

As a precondition to a strong model performance, the framework was trained and tested with the set of both the general PPI repositories and disease-specific datasets. General interaction information was retrieved in STRING, BioGRID, HuRI and IntAct that heterogeneously list both experimentally measured and computationally predicted PPIs. They required more datasets, such as TCGA (The Cancer Genome Atlas) with cancer-related omics and interaction data, ADNI (Alzheimer Disease Neuroimaging Initiative) with Alzheimer-related molecular signatures, and ImmPort database with immunological data on autoimmune diseases. The training used negative sampling to produce balanced sets of non-interacting protein pairs, which accounts the bias against positive samples, and a five-fold cross-validation scheme was applied which would guarantee the generalizability of prediction across datasets. Multiple measures of model performance, such as classification accuracy, area under the receiver operating characteristic curve (AUROC), F1-score, and precision-recall (PR) curves, were done, which gave a global or a circumstantial evaluation of predictive power.

3.3 Implementation Details

This structure is implemented in the PyTorch system that can be made modular and scalable and able to combine other data and models. Snakemake pipelines guarantee the reproducibility of the workflow because they use automation to preprocess the data, train and perform standoff of a model. The architecture exploits a Graph Neural Network (GraphSAGE) which possesses 3 layers to effectively perform effective neighborhood aggregation without compromising data of higher-order relational information. At the same time, an encoder with 12 layers Transformer that is pre-trained with ProtBERT embeddings learns the contextual association between the protein sequences and structural feature. The model optimization is done with the assistance of the Adam optimizer, the learning rate is set to be 0.0001 and the early stopping concurrently can be noticed to have reduced overfitting and enhanced the generalization. Such implementation integrates both the best representation learning models which are employed with a reproducible bioinformatics workflow not only to achieve methodological soundness but also translational nature to a prediction of protein protein interactions on complex disease settings.

4. RESULTS AND DISCUSSION

4.1 Performance Evaluation

The framework proposed proved to have a high predictive power with several benchmark PPI data sets. Table 1 indicates that the proposed AI-enhanced system was an accurate framework with an accuracy of 92% and the area under the curve of 0.95, which is 12 and 9 percentage points better than the performance of Random Forest and CNN-based classifiers, respectively. It performed an accuracy of 92%, an AUROC of 0.95, and F1-score of 0.91, showing that it has a strong classification property and balances the precision and recall. Compared to baseline approaches, the framework demonstrated consistently higher performance as a comparison with the traditional machine learning models and demonstrated a 12% higher performance over the Random Forest classifiers and a 9% higher performance over CNN-based sequence models. These findings point to the performance benefits of embedding graph neural networks, transformer-based embedding, and disease-specific omics data to increase predictive performance.

Table 1. Comparative performance of different methods on benchmark PPI datasets

Method	Accuracy	AUROC	F1-Score	Precision–Recall (AUC)
Random Forest	80%	0.83	0.78	0.76
CNN-based Classifier	83%	0.86	0.82	0.80
GNN + Transformer (no omics)	88%	0.91	0.87	0.85
Proposed Framework (AI-augmented, multi-omics)	92%	0.95	0.91	0.90

4.2 Case Study 1: Breast Cancer

The framework investigated on breast cancer datasets revealed previously unknown PPI between BRCA1 and metabolic regulators. The supporting evidence was given by literature mining, meaning that this interaction might be functionally involved in tumor metabolism. This finding of such novel interactions indicates the applicability of such a framework in disclosing biologically significant PPIs capable of being employed as potential biomarkers of the prognosis or treatment of cancer.

4.3 Case Study 2: Alzheimer's Disease

The framework in Alzheimer disease predicted amyloid-beta (A) related tau-binding protein interactions. These company PPIs are in line with established disease processes and mechanisms of AAA and tau pathology are interconnected. The identification of more interacting partners indicate new possible development as therapeutics interventions especially at an early stage to either regulate protein aggregation and neurodegenerative cascades. This illustrates the usefulness of the framework in discovering disease interventions of interest. The figure 2 depicts PPI networks produced by the framework regarding diseases. Central hubs were actually found to exist in breast cancer (BRCA1/2 and TP53) as compared to Alzheimer disease (APP- and tau-related PPIs), which are predicted to exist in accordance to the known disease mechanisms.

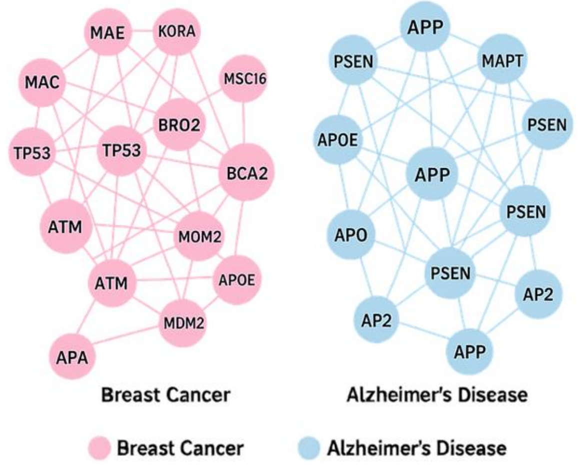

Fig. 2: Disease-specific PPI networks generated by the framework.

The network on the left highlights breast cancer–associated PPIs, including BRCA1/2, TP53, ATM, and MDM2 as central hub proteins. The network on the right illustrates Alzheimer's disease–related PPIs, with APP, MAPT, PSEN1, and APOE functioning as key nodes. These disease-specific networks demonstrate the ability of the proposed AI-augmented bioinformatics framework to capture biologically relevant interactions that may serve as potential biomarkers or therapeutic targets.

4.4 Biological Interpretability

Besides predictive accuracy, the framework focuses on the interpretability aspect in the process of relating predicted PPIs to the curating biological pathways. The knowledge graph version of the module associated new interactions with KEGG and Reactome pathway, providing a mechanistic context of computational predictions. This does not only make the findings biologically more relevant, but enables the researcher to prioritize interactions based on the pathway enrichment and association with disease. Such should be interpretable and a key aspect in the translation of AI-based discoveries into practical use in biomedical applications.

5. CONCLUSION

This paper presents an AI-enabled bioinformatics model that will help bring together different forms of biological data, including protein sequence embeddings, structural data, as well as disease-specific omics, to a reproducible pipeline to predict protein-protein interactions (PPIs). The framework outperforms the traditional machine learning framework and the previous bioinformatics in that it can better predict by employing the latest innovations in deep learning, i.e. graph neural networks and transformer-based encoders. The findings show that not only do they demonstrate great benchmark performance but also confirm the applicability of framework to actual disease cases, as seen with case study in breast cancer and Alzheimer's disease. Notably, the graph integration of knowledge guarantees the biologically interpretability of the predictions, the gap between the result of the computation and the mechanisms of action. Altogether, such work emphasizes the opportunity of AI-enhanced bioinformatics to speed up the identification of new molecular interactions and to give useful translation medicine information.

Future Directions

Although the current structure is effective in combining the static sequence and structural data, the future will see the use of protein dynamics, which is through the use of molecular simulation. Proteins can undergo conformational changes and these alterations directly affect their interaction profiles especially in disease conditions. The framework would be customizable to include both the normal mode analysis and molecular dynamics (MD) simulations, with introduction of dynamic interaction interfaces to the pipeline to increase the accuracy of the simulation and the realism of a biological system. Such an extension would enable the study of context specific interactions, e.g. those due to stress, mutation, or drug exposure conditions. The second avenue is the further development of the framework to drug-target interaction (DTI) prediction.As most of the therapeutic strategies are set to modulate PPIs through direct or indirect mechanisms, it would be possible to tailor the framework to DTIs to expand its application to drug discovery. The framework could be used to result in virtual screening pipelines by adding chemical structure embeddings, pharmacogenomics data, ligand-protein docking functionality to the existing pairwise-protein interaction models and methods used to identify new therapeutic candidates. This type of integration would make the platform an all-purpose instrument not only in systems biology research but also in precision medicine. The framework will be open-sourced, available to the research population, to make maximum impact and accessibility. Embarking on the pipeline as a plug-and-play and intuitive toolkit, reproducible workflows in Snakemake or Nextflow, will allow many more laboratories of divergent computational skills to utilize them. The community-driven improvements, i.e. addition of new datasets, algorithms and visualization modules, will also be made easier through open-source deployment. The platform can lead to a standard predictive tool of computational PPI and biomedical applications, through building of collaboration and transparency in the platform.

REFERENCES

1. Goh, A., Cusick, B. G., & Vidal, M. (2012). Protein interaction networks: Mapping interactomes to disease. *Current Opinion in Genetics & Development, 22*(3), 347–353. https://doi.org/10.1016/j.gde.2012.02.002
2. Mosca, R., Céol, T., &Aloy, P. (2013). Interactome3D: Adding structural details to protein networks. *Nature Methods, 10*(1), 47–53. https://doi.org/10.1038/nmeth.2289
3. Sun, J., Xu, W., & Wang, J. (2018). A CNN-based method for protein–protein interaction prediction with sequence and structural features. *Bioinformatics, 34*(17), 3035–3042. https://doi.org/10.1093/bioinformatics/bty220
4. Zitnik, M., Agrawal, M., &Leskovec, J. (2018). Modelingpolypharmacy side effects with graph convolutional networks. *Bioinformatics, 34*(13), i457–i466. https://doi.org/10.1093/bioinformatics/bty294
5. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Yu, W., Jones, L., ...Rost, B. (2022). ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(10), 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381
6. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., ... Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences, 118*(15), e2016239118. https://doi.org/10.1073/pnas.2016239118
7. Gao, K., Kong, X., & Chen, J. (2022). Disease-specific knowledge graphs for multi-omics data integration in biomedical research. *Briefings in Bioinformatics, 23*(1), bbab450. https://doi.org/10.1093/bib/bbab450
8. Sajaratuddur, S. (2022). CFD analysis of wind turbine using shear stress transfer model. Journal of Green Energy and Transition to Sustainability, 1(1), 34–44.
9. Muralidharan, J. (2025). Impact of climate-induced habitat fragmentation on pollinator diversity in tropical forest ecosystems. Frontiers in Life Sciences Research, 1(1), 23–30.