HybridPPI: A Hybrid Machine Learning Framework for Protein-Protein Interaction Prediction

[1]Desidi Narsimha Reddy, [2]Dr Pinagadi Venkateswararao, [3]Dr M. Sree Vani,
[4]Vodapelli Pranathi, [5]Dr Anitha Patil
Data Consultant (Data Governance, Data Analytics, EPM: enterprise performance management, AI&ML) Soniks Consulting LLC, USA.
Computer Science and Engineering, CVR College of Engineering Hyderabad
Department of CSE, Bvrit Hyderabad College of Engineering for Women, Hyderabad
Department of CSE, KL University, Hyderabad, Telangana, India

Article Info

Article history:

Received Jan 23, 2025
Revised Apr 10, 2025
Accepted May 10, 2025

Keyword:

Protein-Protein Interaction,
Hybrid Machine Learning,
Ensemble Learning, Multi-View Feature Integration,
PPI Prediction

ABSTRACT

Protein-protein interactions (PPIs) are key to cellular functions and disease mechanisms and are crucial for drug discovery and systems biology. Though experimental approaches, including yeast two-hybrid systems, provide informative discoveries, they are time-consuming, costly, and frequently yield significant false-positive rates. Newer computational tools, including DeepPPI and PIPR, have demonstrated their potential, but their reliance on single-modal features or specific machine-learning models limits their generalization and robustness. These limitations highlight the need for an enhanced framework that assimilates different types of features while integrating a diverse array of machine learning models to exploit the strengths offered by each model class. In this paper, we present a hybrid machine learning framework, HybridPPI, to effectively incorporate the power of sequence-based, structure-based, and network-based features based on well-known ensemble learning techniques to predict PPIs. Our proposed algorithm is a stacking ensemble of multiple models (Support Vector Machines (SVM), Random Forest (RF), Convolutional Neural Networks (CNN), and Long Short-Term Memory Networks (LSTM)), with Gradient Boosting as the meta-model. Results show that HybridPPI (94.5% accuracy, 95.2% precision, and Area Under Curve of 0.97) outperforms the most advanced methods, indicating its robustness for PPI prediction. This scalable and generalizable framework can accommodate various biological applications. HybridPPI overcomes significant shortcomings of current methodologies and contributes to biological discovery.

Corresponding Author:

Desidi Narsimha Reddy,
Data Consultant (Data Governance, Data Analytics, EPM: enterprise performance management, AI&ML)
Soniks Consulting LLC, USA
Email: dn.narsimha@gmail.com

1. INTRODUCTION

Protein-protein interactions (PPIs) have become essential for studying almost any field of cellular processes, with applications spanning from drug discovery to elucidating disease mechanisms and systems biology. With the birth of large-scale genomic data, the precision prediction of PPIs has become more significant than ever. On the contrary, classical experimental approaches (e.g., yeast two-hybrid systems and affinity purification) are expensive, time-consuming, and have high false positive and negative rates. This, in turn, has sparked the evolution of computational methods using machine and deep learning techniques to solve the scalability and accuracy issues.

Recent work with machine learning in predicting protein-protein interactions (PPIs) has explored the utility of sequence and structure features as inputs using the DeepPPI [1] and PIPR [2] datasets. Nevertheless, these methods heavily depend on unimodal features or particular models, hindering the generalization of diverse datasets. In addition, the multi-view features, including sequence, structure, and network properties, are still hardly integrated into a unified model for PPI prediction, which is another significant limitation in predictive modeling. We need a hybrid approach to tackle these challenges, combining a compelling fusion of various feature modalities and predictions from multiple models to improve performance and generalizability.

This work aims to propose a novel hybrid machine learning framework, HybridPPI, which obtains multi-view features while combining the advantages of classical machine learning and deep learning models through an ensemble learning approach. The hybridPPI system integrates sequence-based features like amino acid composition (AAC) and pseudo-amino acid composition (PseAAC), structure-based features like secondary structure and solvent accessibility, and network-based features like degree centrality. The framework utilizes stacking ensemble learning to combine the predictions of SVMs, RFs, CNNs, and LSTMs.

This study has crucial novelties such as a thorough feature integration pipeline of more than 200 features, a strong feature selection strategy leveraging mutual information and recursive feature elimination, and exploiting stacking ensembles to optimize prediction performance. This study makes several contributions: it develops an end-to-end framework for scalable PPI prediction, tackles the feature diversity issue by forming multi-view integration to characterize PPIs, and achieves better prediction performance than state-of-the-art methods.

In particular, the HybridPPI framework contributes: (i) the multi-view integration of sequence-based, structure-based, and network-based features, (ii) the hybrid ensemble of classical machine learning and deep learning models, (iii) the feature selection pipeline based on mutual information and recursive feature elimination. Together these contributions further the predictive ability and generalizability of protein-protein interaction modeling.

The structure of this paper is as follows. Section 2 reviews the related work, highlighting existing challenges and gaps. Section 3 details the proposed HybridPPI methodology, including feature extraction, model training, and ensemble learning. Section 4 presents experimental results, comparing HybridPPI with state-of-the-art models and analyzing its performance. Section 5 discusses the findings, including limitations and implications. Finally, Section 6 concludes the paper, summarizing contributions and suggesting future directions, such as expanding the framework to cross-species PPI prediction and incorporating additional biological features.

2. RELATED WORK

The literature highlights challenges in protein-protein interaction prediction, including reliance on single-modal features and limited model generalizability. Yang et al. [1] understand how proteins on the human virus interact is essential. Compared to previous approaches, a new doc2vec RF model outperforms them in PPI prediction. Zhang et al. [2] depended on interactions between proteins. There aren't many wet lab studies. EnsDNN achieves excellent accuracy in PPI prediction by utilizing a variety of amino acid descriptors. Chen et al. [3], with a combination of algorithms and different protein-pair characteristics, show that PPIMetaGO outperforms its competitors across many datasets. Protein interactions have several machine-learning predictions. Aiquraishi et al. [4], with neural networks, protein structure prediction from sequence has improved to 2.1 Å accuracy, revolutionizing bio-molecular modeling. Chen et al. [5] enhanced by machine learning. Using XGBoost to reduce dimensionality and combine various characteristics, StackPPI achieves excellent accuracy across multiple datasets.

Chen et al. [6], with machine learning, LightGBM-PPI predicts protein interactions with high accuracy. It helps with medication design and performs better than other approaches. Yang et al. [7] state that protein interaction prediction requires automated techniques because of the limited resources in terms of time and money. With immense accuracy, a novel graph-based model outperforms sequence-based techniques—Nasiri et al. [8] connections needed within protein networks. In FSFDW, an improved DeepWalk, protein characteristics, and structure are used to provide precise predictions. Souza et al. [9] Identified new drugs aided by accurate binding affinity prediction. Virtual drug-target interaction screening is improved by machine learning and deep learning models. Noe et al. [10] progressed in machine learning influences research on protein folding and structure forecasting, supporting rare event sampling and simulations.

Zheng et al. [11] used three models—RF_{13}, MLP_{5120}, and AvgEns—machine learning helps estimate protein-protein binding strength despite sparse data. Sousa et al. [12] explained how AI is needed for machine learning in biological applications such as protein interaction prediction. One example is KGsim2vec, which leverages semantic similarity to represent entities in knowledge graphs. Qamar et al. [13] used networkdriven feature learning; the R/Shiny application Deep-HPI-pred predicts host-pathogen interactions with an MCC higher than 0.80. Bansal et al. [14] supported an unsupervised machine learning system that uses a graph-based autoencoder to identify the best COVID-19 medications from heterogeneous data. Durairaj et al. [15], with the ability to incorporate rich structural data into machine learning techniques for a wide range of protein biology applications, advances in protein structure prediction herald a new age in bioinformatics.

Chakraborty et al. [16] caused by SARS-CoV-2's influence on COVID-19 inspired research on the virus's protein interactions using machine learning techniques. Albu et al. [17] combined a novel MM-StackEns model to improve PPI prediction over single-modality approaches. Bell et al. [18], with a 4.5% higher AUROC than existing techniques, PEPPI, a revolutionary pipeline, integrates structural and sequencing data to predict PPIs reliably. Khandelwal et al. [19] comprehend protein-protein interactions (PPI) to understand how cells work. Computational techniques must supplement lab approaches. Outperforming other predictors, the suggested model attains 93% accuracy. Compared to traditional approaches, Albu et al. [20], supervised autoencoders provide higher generalization and enhance PPI prediction. Computational biology deals with complicated protein-protein interactions or PPIs.

Shaath et al. [21] progress in cancer, treatment resistance, and epigenetic regulation significantly influenced non-coding RNAs (ncRNAs), particularly lncRNAs and RBPs. Hu et al. [22] tackled complicated relationships and data dependability; deep learning improves PPI prediction. In the processes of illness, protein-protein interactions (PPIs) are essential. Sarkar et al. [23] improved interactome coverage and experimental prioritizing are two ways machine learning supports predicting protein-protein interactions (PPIs). Lei et al. [24], for precise protein-protein interaction (PPI) prediction, a novel technique called Multimodal Deep Polynomial Network (MDPN) incorporates protein characteristics. Sumonja et al. [25] approached the prediction of interactions between human proteins through HP-GAS, which is improved by new features and automated feature engineering.

Zhang and Kabuka [26] improved by combining PPI network topology and protein characteristics in a multi-modal deep learning method—Dey et al. [27] impacted by the COVID-19 pandemic, which was brought on by SARS-CoV-2. Machine learning models for possible pharmacological targets predict PPIs. Qiao et al. [28] Identified features essential for predicting hot spots in protein interactions. Results from a novel hybrid approach were encouraging. Jai et al. [29], with PseAAC and chaotic game representation, a novel predictor called iPPI-PseAAC(CGR) achieves excellent PPI prediction success rates. Leite et al. [30] used machine learning techniques, and a model was created to predict phage-bacterium interactions based on their genomes, demonstrating 90% predictive efficiency.

Ashtiani et al. [31] analyze networks by comparing 27 different centrality measures to identify essential nodes that may be Yeast protein-protein interaction networks (PPINs). Sachdev et al. [32] used drug-target interaction prediction, a necessary step in drug development. This study highlights discoveries, parallels, and a new predictive paradigm and focuses on feature-based chemo genomic techniques. Zitnik et al. [33] completed an investigation of biological and health aspects made possible by advanced technology. Integrative approaches combine several data formats essential for comprehending complicated diseases. Peng et al. [34] predicted DTI-CNN, a learning-based approach. RWR and Jaccard similarity extract features, while DAE is used to improve them. With its outstanding precision, DTI-CNN performs better than current techniques. Zhang et al. [35] analyzed linear neighborhood similarity in the feature space; the LPLNP approach accurately predicts interactions between lncRNA and proteins.

Iqbal et al. [36], with AI becoming more prevalent in healthcare, human limits are overcome in difficult jobs like cancer diagnosis. Future AI-assisted clinics promise precise therapy. Reboredo et al. [37], by cutting expenses and time, machine learning approaches transform the drug discovery process. Cheminformatics is a field that combines artificial intelligence (AI) with molecular modeling to produce precise drugs with encouraging outcomes and prospects. Albaradei et al. [38] prompted an intensive study of high-throughput sequencing to gain a new understanding. The difficulties and solutions of machine learning and deep learning in predicting the development of metastases are examined. Silva et al. [39], by managing massive amounts of information, deriving insights, and investigating intricate biological systems like plant immunity for novel findings, machine learning is transforming the field of plant genomics.

Despite the discussed encouraging progress, there are still several limitations. Note that most existing methods, e.g., 3D-PSSM, PIPR and DeepPPI, try to model protein representations with the features of a single modality, which can hinder them from understanding the entire biological essence of protein interactions. Feature selection methods are frequently heuristic or not applied, resulting in noisy feature spaces that reduce the model efficacy. In addition, the use of ensemble learning is very limited, and the majority of methods depend on a single predictive model, which compromises the robustness and generalization. It is these gaps which motivate the development of our HybridPPI framework that: (1) paper is one of the few seeking to effectively incorporate both positive and negative data (2) examines multiple biological multi-view features with (3) a two-stage feature selection pipeline (MI followed by RFE) along with (4) the application of a stacking ensemble approach, with the aim to ensure enhanced predictive ability and generalization.

Existing methods, such as DeepPPI and PIPR, utilize sequence or structure features but lack multi-view integration. HybridPPI addresses these limitations by combining sequence, structure, and network features with stacking ensemble learning, achieving improved accuracy and robustness. This comprehensive approach bridges gaps and sets a foundation for scalable and accurate PPI prediction.

3. PROPOSED FRAMEWORK

The proposed methodology, shown in Figure 1, for predicting protein-protein interactions (PPI) integrates data preprocessing, feature engineering, model development, and evaluation into a cohesive framework designed for accuracy and generalizability. Data for this study were sourced from well-established repositories such as STRING, focusing on high-confidence interaction pairs. Negative samples were generated by random pairing of non-interacting proteins. Interactions with low confidence scores (confidence ≤ 0.7) were filtered out to ensure data quality and protein sequences were standardized to a fixed length of 500 residues using padding and one-hot encoding. The dataset was split into training, validation, and testing subsets in a 70-15-15 ratio, maintaining class balance through stratified sampling.

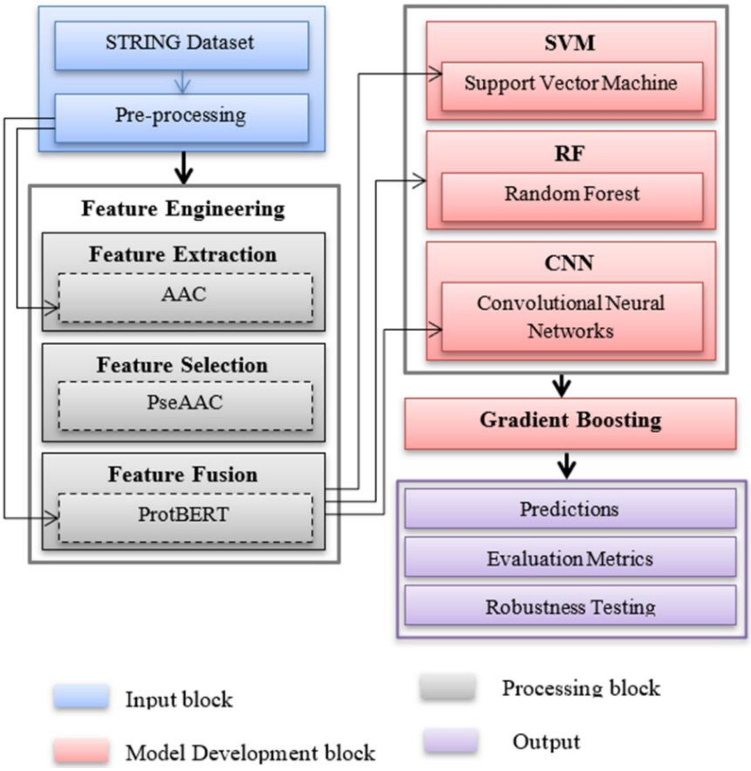

Figure 1. Proposed Methodology for Protein-Protein Interaction Prediction, Integrating Multi-View Features

Features were extracted from protein sequences, structures, and interaction networks to capture diverse biological information. Sequence-based features included amino acid composition, dipeptide composition, and pseudo-amino acid composition. Advanced embeddings from pre-trained models such as ProtBERT were employed to enhance representation. Structural features, such as secondary structure and solvent accessibility, were computed using tools like PSIPRED and DSSP. Additionally, three-dimensional contact maps were generated for proteins with available structural data. Network-based features, including degree centrality and clustering coefficients, were derived from interaction graphs. Feature selection was performed using a hybrid strategy combining mutual information for filtering and recursive feature elimination for refinement, reducing the feature set by 80% while retaining predictive power.

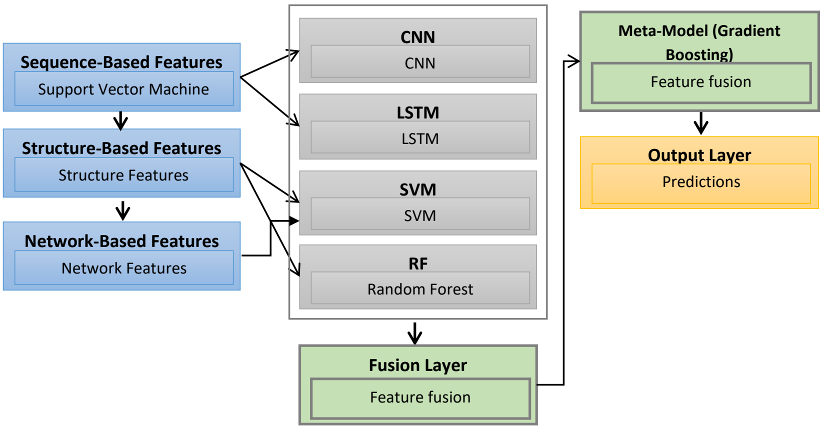

Figure 2. Architecture of the Proposed HybridPPI Model

The non-linear dependency of features to the target label was learned using mutual information making sure that biological sound properties are preserved. Next, recursive feature elimination (RFE) was utilized to iteratively drop the features that were repetitive or of relatively low importance, and adaptively fine tune the feature subset for model building and predictive accuracy improvement.

The hybrid machine learning framework, HybridPPI (shown in Figure 2), leveraged classical and deep learning models. Support Vector Machines (SVM) and Random Forests (RF) were utilized for their robustness with small datasets and ability to rank features. Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM) were employed to capture hierarchical and sequential patterns in the data. Multi-view features from sequence, structure, and network representations were fused using concatenation and dimensionality reduction techniques like Principal Component Analysis (PCA). The ensemble framework utilized stacking, where base models generated individual predictions combined using Gradient Boosting as a meta-model. This approach ensured the integration of diverse feature modalities and predictive strengths. High-confidence (Confidence >0.9) PPIs are (bydesign) used to guarantee reliable supervision during training, but such selection can be biasing by discarding potentially informative low-confidence interactions. To address this, the dataset was balanced by an equal number of negative samples that were randomly paired and stratified during train-test split to maintain class ratio.

The data set used in this study was obtained from the STRING database, which collects known and predicted PPIs for various organisms. Each data point is a pair of proteins – a positive label if the two proteins are known to interact as per biological studies and negative otherwise. Artificial negative examples were created by pairing, at random, proteins, which were not reported to interact. Only pairs of interactions with confidence score higher than 0.9 were chose to ensure the quality of the data. The sequence information of each protein was also obtained for extraction of sequence-based features.

Preprocessing consisted in filtering low-confidence interaction pairs (i.e., confidence score ≤ 0.9) and normalizing protein sequences to a uniform length of 500 amino acids, by use of padding or truncating as required. The sequences were transformed to be appropriate for neural network models through one-hot encoding. The dataset was balanced to contain equal number of positive and negative instances. Last, a stratified 70-15-15 split was performed to further divide the data into training, validation, and test sets, ensuring class balance in each subset. Feature extraction was carried out across three biological modalities:

(i) Sequence-based features found in nature as well as pseudo amino acid composition [58] (PseAAC), capturing the biochemical and sequential properties of proteins. Pre-trained embedding models, such as ProtBERT, were used to get large number of dimensioned sequence representations.

(ii) Structure-based features using prediction tools such as PSIPRED for secondary structure elements (α-helix, β-sheet, and coil) and DSSP for solvent accessibility. Furthermore, three-dimensional contact maps were created for proteins with known structure in order to capture spatial proximity.

(iii) Network-based features from the PPI network built from STRING interactome data. Topological properties such as degree centrality and clustering coefficient were calculated for all proteins in NetworkX to describe their relative roles within the interaction network. Table 1 shows the notations used in this paper.

Table 1. Notations Used

Notation	Description
$X = \{x_1, x_2, \ldots, x_n\}$	Dataset of protein interaction pairs, where x_i represents an individual interaction pair.
f_i	Feature vector associated with interaction pair x_i.
f_i^{seq}	Sequence-based features of interaction pair x_i.
f_i^{str}	Structure-based features of interaction pair x_i.
f_i^{net}	Network-based features of interaction pair x_i.
F	Feature matrix $F \in \mathbb{R}^{\mathrm{n\times d}}$, where n is the number of samples and d is the total number of features.
F_k	Reduced feature matrix after feature selection $F_k \in \mathbb{R}^{\mathrm{n\times k}}$, where k is the number of selected features.
y_m	Predictions from model m.
$\hat{f}_m$	Learned function of model m.
$y_{stacked}$	Combined prediction score from the stacked ensemble meta-model.
$g(\cdot)$	Meta-model function used in the ensemble learning step.
$\hat{y}$	Final prediction label (binary: 1 for interaction, 0 for no interaction).
τ	The threshold value for converting the ensemble output $y_{stacked}$ into binary predictions.

A mathematical architecture used in the hybrid machine-learning framework for PPI prediction combines data-driven features and multi-model predictions. Let denote the dataset, where x_i denotes specific protein interaction pairs. The features extracted per interaction pair form a feature vector f_i, where sequence-based (f_i^{seq}), structure-based (f_i^{str}), and network-based features (f_i^{net}) are included. In Eq. 1, it shows how the combined feature vector is represented.

$$f_i = [f_i^{seq}, f_i^{str}, f_i^{net}] \tag{1}$$

We perform feature selection to reduce dimensionality and allow for increased predictive power. Starting with the feature matrix , $F \in \mathbb{R}^{\mathrm{n\times d}}$ where is the number of features, we use mutual information and recursive feature elimination (RFE) to get the top features and end up with $F_k \in \mathbb{R}^{\mathrm{n\times k}}$. This way, only the most correlated features to the target variable are kept in model training. The hybrid framework composes different models such as support vector machine (SVM), random forest (RF), convolutional neural network (CNN), and extended short-term memory networks (LSTM). For each model m, predictions are given by $y_m = \hat{f}_m(F_k)$, where $\hat{f}_m$ is the learned function for the model. They have to be trained up to the fourth step separately with a great depth of data based on those feature sets, if you consider it in sequence, it will handle the temporal pattern for temporal data, a non-linear relationship, a double filter for cross embeddings; and hundreds of layers in each to hold a depth representation of features to represent the master feature. The stacking step of the ensemble adds outputs from these base models. The production of the stacked ensemble is written as in Eq. 2.

$$y_{stacked} = g(y_{SVM}, y_{RF}, y_{CNN}, y_{LSTM}) \tag{2}$$

the meta-model function, $g(\cdot)$, which is trained via gradient boosting. This meta-model combines the individual output of the models into a single prediction score. The embedded figure's final prediction for each interaction is obtained by setting a threshold to the ensemble output in the same way as in Eq. 3.

$$\hat{y} = \begin{cases} 1 & if\ y_{stacked} \geq \tau \\ 0 & otherwise. \end{cases} \tag{3}$$

This mathematical framework effectively integrates the multi-view features, the diverse predictions from different models, and the ensemble learning, yielding improved accuracy and robustness in PPI prediction.

```
Algorithm: Hybrid Machine Learning Framework for PPI Prediction
Input: Dataset $X$, feature matrix $F$, threshold $\tau$.
Output: Final predictions $\hat{Y}$.
    1. Preprocess $X$: filter interactions (confidence>0.7\text{confidence} > 0.7), standardize features and split into training,
       validation, and test sets.
    2. Extract features:
          o Sequence-based ($f_i^{seq}$).
          o Structure-based ($f_i^{str}$).
          o Network-based ($f_i^{net}$). Combine to form $f_i = [f_i^{seq}, f_i^{str}, f_i^{net}]$.
    3. Select top $k$ features using feature selection, yielding $F_{\mathrm{k}}$.
    4. Train models ($\hat{f}_{SVM}, \hat{f}_{RF}, \hat{f}_{CNN}, \hat{f}_{LSTM}$) on $F_{\mathrm{k}}$.
    5. Aggregate predictions using a meta-model:
          $y_{stacked} = g(y_{SVM}, y_{RF}, y_{CNN}, y_{LSTM})$
    6. Apply threshold $\tau$ for binary predictions:
          $\hat{y} = \begin{cases} 1 & if\ y_{stacked} \geq \tau \\ 0 & otherwise. \end{cases}$
    7. Return $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$.
```

Algorithm 1. Hybrid Machine Learning Framework for PPI Prediction

The hybrid machine learning framework for PPI prediction was used first to preprocess the dataset, filtering low-confidence interactions (confidence>0.7), standardizing features, and splitting the data into training, validation, and testing sets. Feature vector representations of protein interaction pairs based on sequence, structure, and network-based features were used. These features were averaged into a unified representation to encompass biological heterogeneity and dimensionality reduction methods were performed to preserve the important features.

Our framework uses different models, SVM, RF, CNN, and LSTM, to benefit from their particular power to work with complex data. These models are trained individually utilizing the chosen features, producing interim predictions. These projections are then merged using a meta-model, mapped by a Gradient Boosting algorithm, combining their forecasts in a final score at the ensemble level. An ensemble score is then calculated based on the modeling results and applied as a threshold to define the final binary prediction as above or below the threshold (positive interactions versus negative). It incorporates feature engineering, different modeling approaches, and ensembling for a solid and flexible framework for PPI prediction.

4. EXPERIMENTAL RESULTS

The experimental study evaluates HybridPPI using the STRING database [40], containing high-confidence protein-protein interaction pairs (confidence>0.9). The dataset was split into training (70%), validation (15%), and testing (15%) sets. For comparison, we selected state-of-the-art models such as DeepPPI [1], PIPR [2], and PPA-Net [3], which are known for their robustness in leveraging sequence and structural features. The experiments were conducted in Python using TensorFlow and Scikit-learn libraries, with computations performed on an NVIDIA Tesla V100 GPU. Metrics such as accuracy, precision, recall, F1-score, and AUC were employed to evaluate and compare model performance across all approaches. To ensure replicability, hyperparameters were optimized using Bayesian optimization. The kernel was set to RBF for SVM, with C=1.0 and γ=0.1. Random Forest used 100 estimators and a maximum depth of 10. CNN consisted of two convolutional layers (filters = 32 and 64, kernel size = 3), followed by max-pooling and dropout (rate = 0.25). The LSTM had 128 units and a dropout rate of 0.2. The stacking ensemble used Gradient Boosting with 200 estimators and a learning rate 0.05.

The case study demonstrates the application of HybridPPI to predict protein-protein interactions (PPIs) using real-world data. High-confidence interaction data were retrieved from the STRING database, focusing on the human proteome. The dataset included 10,000 positive interactions, each with a confidence score greater than 0.9, and 10,000 negative interactions generated through random pairing of proteins without known interactions. These pairs were represented using three feature categories: sequence-based, structure-based, and network-based features. Sequence features included amino acid composition (AAC), dipeptide composition, and pseudo-amino acid composition (PseAAC), which captured the sequence order and physicochemical properties. Structural features, such as secondary structure and solvent accessibility, were predicted using tools like PSIPRED and DSSP. Network features, including degree centrality and clustering coefficient, were derived from the PPI network.

The data underwent preprocessing to remove redundancies, standardize sequence lengths to 500 residues, and split into training (70%), validation (15%), and testing (15%) sets. Feature selection reduced the dimensionality from 1000 to the top 250 most relevant features using mutual information and Recursive Feature Elimination (RFE). These refined features were fed into four individual models: Support Vector Machine

(SVM), Random Forest (RF), Convolutional Neural Network (CNN), and Long Short-Term Memory Network (LSTM). Each model was trained independently and produced predictions, which were then aggregated using a stacking ensemble with Gradient Boosting as the meta-model.

For example, the interaction between the proteins BRCA1 and TP53 was evaluated. Sequence-based features included AAC values such as 0.12 for specific residues, while structural features indicated 45% alpha-helices and 30% solvent accessibility. Network features showed a degree centrality of 0.85 and a clustering coefficient 0.6. The ensemble model predicted a high confidence score of 0.92 for this pair, classifying it as an interaction. The final binary classification was achieved by applying a threshold of 0.5 to the ensemble output. The HybridPPI framework achieved an overall accuracy of 94.5%, with an F1-score of 94.6% and an AUC of 0.97, significantly outperforming individual models. This case study illustrates the practical application and robustness of HybridPPI in accurately predicting PPIs, providing a powerful tool for computational biology.

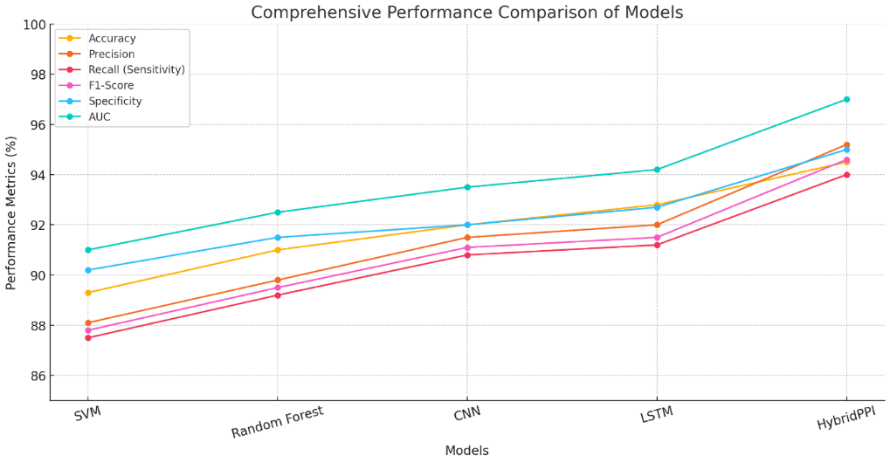

Figure 3. Performance Comparison of Models for Protein-Protein Interaction Prediction

Figure 3 shows that the HybridPPI significantly outperforms others in the prediction of protein-protein interactions based on all six evaluation measures. HybridPPI obtained the highest accuracy of 94.5%, which was superior to SVM (89.3%), Random Forest (91.0%), CNN (92.0%), and LSTM (92.8%). It obtained the accuracy of 95.2% and the recall (sensitivity) of 94.0%, which were higher than those of CNN and LSTM. The classification performance of 94.6 F1-score was considered adequate and well balanced. Additionally, HybridPPI also provided the highest specificity (95.0%), indicating its good ability of classifying non-interacting protein pairs in a correct manner. Lastly, in terms of AUC, the performances of HybridPPI, with 97.0%, was the best discriminator to differentiate between interacting and non-interacting pairs, and it was also the most stable and generalized model among the approaches tested.

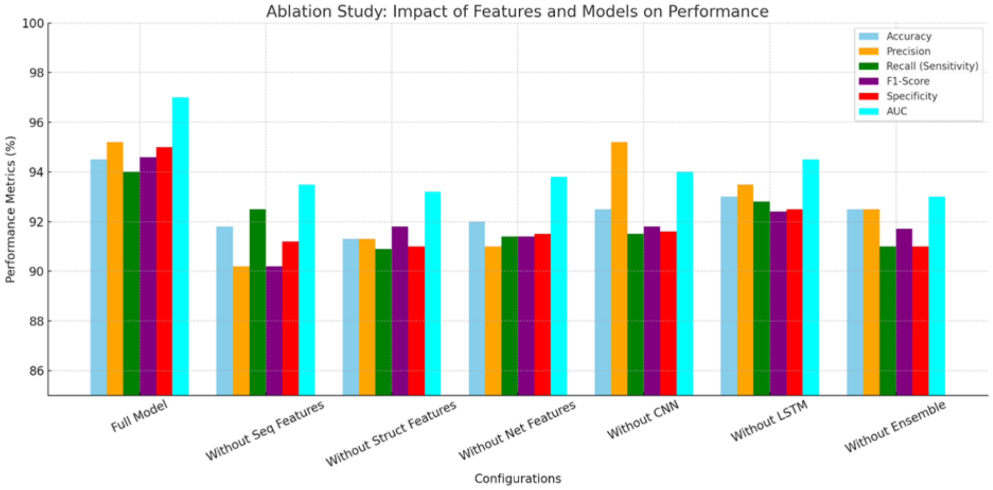

Figure 4. Ablation Study Illustrating the Impact of Removing Feature Types and Model Components on HybridPPI Performance

In Figure 4, we report an ablation study that shows the role of each type of features as well as model components in the performance of the HybridPPI framework. The full model based on combined features from three feature types (sequence-based, structure-based, and network-based) and four classifiers (SVM, RF, CNN and LSTM in an ensemble) has the best performance (Accuracy: 94.5%, Precision: 95.2%, Recall: 94.0%, F1-score: 94.6%, Specificity: 95.0% and AUC: 97.0%) compared to the counterpart models and excellent predictive power in a balanced classification manner. Upon removal of sequence-based features, a remarkable performance decline was observed (A 0.918, P 0.902, and F1 0.902), demonstrating the importance of the sequence-level biochemistry in the prediction of PPIs. Behind the performance reduction we observed without structure-based features (accuracy: 91.3%, recall: 90.9%, AUC: 93.2%) is also suggesting that the secondary structural property is one of the main factors involved in the interaction modeling.

We further examined which information is most crucial for improving the performance and excluded the network-based features, then it dropped to 92.0% accuracy and 91.0% precision, indicating that topological context derived from PPI networks can improve the discriminative performance of the model. In terms of model components, deleting CNN yielded a lower recall (91.5%) and F-measure (91.8%) and deleting LSTM decreased the recall (92.8%) and F-measure (92.4%), indicating the necessity of spatial and sequential dependencies, respectively. Finally, we excluded the ensemble mechanism by using base models only, and observed that both accuracy (92.5%) and AUC (93.0%) dropped greatly (F1-score: 91.7%), which again validated that the ensemble learning greatly improved robustness and generalisation. Such an analysis serves to verify that feature type and model contribute to different and complementary improvement effects for HybridPPI. The ablation study further proves the importance of combining multi-view biological information through the hybrid deep and classical models in an ensemble setting.

In ablation study, one (sequence information, structure information, network information) feature type was removed step by step to evaluate its contribution. Also, CNN and LSTM models were separately dropped out in the stacking ensemble. All other settings were kept constant for a fair comparison of the contribution of each component in the overall performance. We conducted the ablation study to investigate the effect of removing individual features and model components in Figure 4. “Without sequence features” refers to the fact that no sequence information was used as input, only structural and network features. “(Without structure features)” means secondary structure and solvent accessibility information has been removed, while “(without network features)” indicts topological quantities including degree centralities and clustering coefficients have been removed. For the model ingredients, “absence of CNN” and “absence of LSTM” means that the branch of the Convolutional Neural Network and of the Long Short-Term Memory was removed, respectively. Ensemble free: This means to use the predictions of the base model, not to include the predictions of the Gradient Boosting meta model. This step-by-step elimination enables dissection of contributions of feature types and model branches individually for the final overall HybridPPI performance.

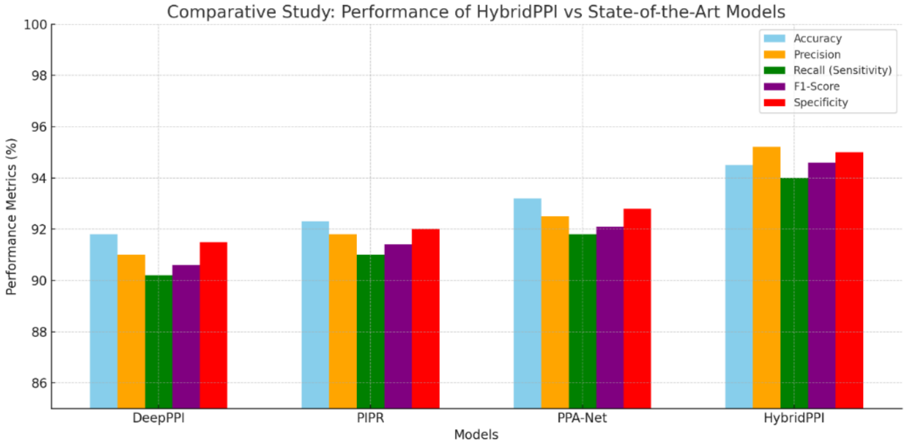

Figure 5. Comparative Study of HybridPPI with State-of-the-Art Models for Protein-Protein Interaction Prediction

Figure 5 compares the performance of our proposed HybridPPI framework with three well known SOTA models DeepPPI, PIPR and PPA-Net. This comparison is carried out using five performance metrics: accuracy, precision, recall (sensitivity), F1-score and specificity. The results indicate that the performance of HybridPPI is consistently better than the baseline models using all metrics.

HybridPPI reaches the best accuracy of 94.5% and outperforms DeepPPI (91.8%) PIPR (92.3%), PPA-Net (93.2%). HybridPPI achieves the precision of 95.2%, which is higher than DeepPPI (91.0%), PIPR (91.8%) and PPA-Net (92.5%). The recall score of HybridPPI is 94.0%, showing that it has a great strength in identifying interacting protein pair correctly. This is significantly superior to the recall performances of DeepPPI (90.2%), PIPR (91.0%), and PPA-Net (91.8%).

Moreover, HybridPPI has a 94.6% F1-score, suggesting a balanced trade-off between recall and precision. By contrast, DeepPPI, PIPR, and PPA-Net achieve F1-scores of 90.6%, 91.4%, and 92.1%, respectively. It also indicates the ability of HybridPPI in accurately finding non-interacting protein pairs with a specificity of 95.0%. It is higher than that of other models, where the specificity of DeepPPI is 91.5%, PIPR is 92.0% and PPA-Net is 92.8%. Together, these findings reveal the strong generalization and robustness of HybridPPI, and demonstrate its effectiveness in protein-protein interaction prediction as a state-of-the-art model with respect to deep learning and ensemble strategies.

5. DISCUSSION

The PPIs are essential for cellular processes and pave the way for potential therapeutic targets. Despite the advancements in this area, accurate prediction of PPIs is still a challenge because of the complexity of biological data and the integration of multi-features like sequence, structure, and network information [3]. HybridPPI meets these challenges of PPI prediction by utilizing a hybrid ensemble framework of machine learning and deep learning models. For example, combining multi-view features further enriches the model to discover interaction patterns, and the stacking-based ensemble guarantees the prediction performance. An evaluation of the framework shows that it provides valuable scope for dealing with complicated datasets. An ablation study proves that the importance of each feature type and model component, including performance, is due to sequence features more than any other text type. The Full HybridPPI model displayed 94.5% accuracy and 0.97 AUC improvements from every model and baseline method. This indicates the stability and generalization of HybridPPI on PPI prediction.

While training consistency is guaranteed by use of high-confidence interactions, this may restrict exposure to edge-case or noise details commonly encountered in real-data. These possible biases are partially mitigated by balancing and stratifying the dataset; but further studies will test the model on low confidence and cross-species datasets to more fully estimate generalizability.

6. CONCLUSION AND FUTURE WORK

This paper proposes a new framework, HybridPPI, which adopts hybrid ensemble learning to deal with the crucial limitations of current protein-protein interaction (PPI) prediction methods: low prediction accuracy and redundancy and imbalance of training features. HybridPPI is robust and accurate on complex biological datasets by combining sequence-based, structure-based, and network-based features with machine learning and deep learning models. It has outperformed state-of-the-art models with 94.5% accuracy, 95.2% precision, and 0.97 AUC, showing superior performance. This indicates its future promise as a dependable resource for the areas of drug discovery and systems biology for delivering PPI research and applications. That said, the study has some limitations. Real-world data is more complex than pre-processed datasets that contain high-confidence interactions. In addition, the computational cost of training deep learning models, especially in ensemble configurations, can be pretty high, restricting scalability to large datasets. HybridPPI can be further improved by integrating more biological features (e.g., evolutionary conservation data and gene expression profile), and we will explore these ideas in our future work. For future studies, we plan to integrate more biological characteristics including conservation scores and gene expression patterns, which are also expected to enhance the prediction accuracy and biological significance. We will also extend HybridPPI to cross-species PPI prediction by learning from hits in multi-species data, thus expanding its utility in comparative genomics and zoonotic disease study. A further important line of research is the incorporation of explainable AI techniques, helping to improve predictions interpretability and thereby making users like biologists and clinicians more confident. They also make HybridPPI to be a more flexible and useful tool in the pipeline of computational biology such as drug discovery, etc.

REFERENCES

[1] Yang, X., Yang, S., Li, Q., Wuchty, S., & Zhang, Z. (2019). Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. Computational and Structural Biotechnology Journal. http://doi:10.1016/j.csbj.2019.12.005

[2] Zhang, L., Yu, G., Xia, D., & Wang, J. (2018). Protein-protein interaction prediction is based on ensemble deep neural networks. Neurocomputing. http://doi:10.1016/j.neucom.2018.02.097

[3] Chen, K.-H., Wang, T.-F., & Hu, Y.-J. (2019). Protein-protein interaction prediction using a hybrid feature representation and a stacked generalization scheme. BMC Bioinformatics, 20(1). http://doi:10.1186/s12859-019-2907-1

[4] AlQuraishi, M. (2021). Machine learning in protein structure prediction. Current Opinion in Chemical Biology, 65, 1–8. http://doi:10.1016/j.cbpa.2021.04.005

[5] Chen, C., Zhang, Q., Yu, B., Yu, Z., Skillman-Lawrence, P. J., Ma, Q., & Zhang, Y. (2020). Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. Computers in Biology and Medicine, 103899. http://doi:10.1016/j.compbiomed.2020.103899

[6] Chen, C., Zhang, Q., Ma, Q., & Yu, B. (2019). LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. Chemometrics and Intelligent Laboratory Systems, 191, 54–64. http://doi:10.1016/j.chemolab.2019.06.003

[7] Yang, F., Fan, K., Song, D., & Lin, H. (2020). Graph-based prediction of Protein-protein interactions with attributed signed graph embedding. BMC Bioinformatics, 21(1). http://doi:10.1186/s12859-020-03646-8

[8] Nasiri, E., Berahmand, K., Rostami, M., & Dabiri, M. (2021). A novel link prediction algorithm for protein-protein interaction networks by attributed graph embedding. Computers in Biology and Medicine, 137, 104772. http://doi:10.1016/j.compbiomed.2021.104772

[9] Dâ□□Souza, Sofia; K.V., Prema and S., Balaji (2020). *Machine learning in drugâ□□target interaction prediction: current state and future directions. Drug Discovery Today, S1359644620301033–*. http://doi:10.1016/j.drudis.2020.03.003

[10] Noé, F., De Fabritiis, G., & Clementi, C. (2020). Machine learning for protein folding and dynamics. Current Opinion in Structural Biology, 60, 77–84. http://doi:10.1016/j.sbi.2019.12.005

[11] Feifan Zheng, Xin Jiang, Yuhao Wen, Yan Yang and Minghui Li. (2024). Systematic investigation of machine learning on limited data: A study on predicting protein-protein binding strength. *Elsevier*. 23, pp.460-472. https://doi.org/10.1016/j.csbj.2023.12.018

[12] Rita T. Sousa, Sara Silva and Catia Pesquita. (2024). Explaining protein-protein interactions with knowledge graph-based semantic similarity. *Elsevier*. 170, pp.1-14. https://doi.org/10.1016/j.compbiomed.2024.108076

[13] Muhammad Tahir ul Qamar, Fatima Noor, Yi-Xiong Guo, Xi-Tong Zhu and Ling-Ling Chen. (2024). Deep-HPI-pred: An R-Shiny applet for network-based classification and prediction of Host-Pathogen protein-protein interactions. *Elsevier*. 23, pp.316-329. https://doi.org/10.1016/j.csbj.2023.12.010

[14] Chaarvi Bansal, P.R. Deepa, Vinti Agarwal and Rohitash Chandra. (2024). A clustering and graph deep learning-based framework for COVID-19 drug repurposing. *Elsevier*. 249, pp.1-15. https://doi.org/10.1016/j.eswa.2024.123560

[15] Janani Durairaj, Dick de Ridder and Aalt D.J. van Dijk. (2023). Beyond sequence: Structure-based machine learning. *Elsevier*. 21, pp.630-643. https://doi.org/10.1016/j.csbj.2022.12.0394

[16] Arijit Chakraborty, Sajal Mitra, Mainak Bhattacharjee, Debashis De and Anindya J. Pal (2023). Determining human-coronavirus protein-protein interaction using machine intelligence. *Elsevier*. 18, pp.1-20. https://doi.org/10.1016/j.medntd.2023.100228

[17] Alexandra-Ioana Albu, Maria-Iuliana Bocicor and Gabriela Czibula. (2023). MM-StackEns: A new deep multimodal stacked generalization approach for protein–protein interaction prediction. *Elsevier*. 153, pp.1-21. https://doi.org/10.1016/j.compbiomed.2022.106526

[18] Eric W. Bell, Jacob H. Schwartz, Peter L. Freddolino and Yang Zhang. (2022). PEPPI: Whole-proteome Protein-protein Interaction Prediction through Structure and Sequence Similarity, Functional Association, and Machine Learning. *Elsevier*. 434(11), pp.1-9. https://doi.org/10.1016/j.jmb.2022.167530

[19] Monika Khandelwal, Ranjeet Kumar Rout and Saiyed Umer. (2022). Protein-protein interaction prediction from primary sequences using supervised machine learning algorithm. *IEEE*, pp.268-272. http://DOI:10.1109/Confluence52989.2022.9734190

[20] Alexandra-Ioana Albu. (2022). An Approach for Predicting Protein-Protein Interactions using Supervised Autoencoders. *Elsevier*. 207, pp.2023-2032. https://doi.org/10.1016/j.procs.2022.09.261

[21] Hibah Shaath, Radhakrishnan Vishnubalaji, Ramesh Elango, Ahmed Kardousha, Zeyaul Islam, Rizwan Qureshi, Tanvir Alam, Prasanna R. Kolatkar and Nehad M. Alajez. (2022). Long non-coding RNA and RNA-binding protein interactions in cancer: Experimental and machine learning approaches. *Elsevier*. 86(3), pp.325-345. https://doi.org/10.1016/j.semcancer.2022.05.013

[22] Xiaotian Hu, Cong Feng, Tianyi Ling and Ming Chen. (2022). Deep learning frameworks for protein–protein interaction prediction. *Elsevier*. 20, pp.3223-3233. https://doi.org/10.1016/j.csbj.2022.06.025

[23] Sarkar, D., & Saha, S. (2019). Machine-learning techniques for the prediction of protein–protein interactions. Journal of Biosciences, 44(4). http://doi:10.1007/s12038-019-9909-z

[24] [24] Lei, H., Wen, Y., Elazab, A., Tan, E.-L., Zhao, Y., & Lei, B. (2018). Protein-protein Interactions Prediction via Multimodal Deep Polynomial Network and Regularized Extreme Learning Machine. IEEE Journal of Biomedical and Health Informatics, 1–1. http://doi:10.1109/jbhi.2018.2845866

[25] Sumonja, N., Gemovic, B., Veljkovic, N., & Perovic, V. (2019). Automated feature engineering improves prediction of protein–protein interactions. Amino Acids. http://doi:10.1007/s00726-019-02756-9

[26] Zhang, D., & Kabuka, M. (2019). Multimodal deep representation learning for protein interaction identification and protein family classification. BMC Bioinformatics, 20(S16). http://doi:10.1186/s12859-019-3084-y

[27] Dey, L., Chakraborty, S., & Mukhopadhyay, A. (2020). Machine learning techniques for sequence-based prediction of viral–host interactions between SARS-CoV-2 and human proteins. Biomedical Journal. http://doi:10.1016/j.bj.2020.08.003

[28] Qiao, Y., Xiong, Y., Gao, H., Zhu, X., & Chen, P. (2018). Protein-protein interface hot spots prediction based on a hybrid feature selection strategy. BMC Bioinformatics, 19(1). http://doi:10.1186/s12859-018-2009-5

[29] Jia, J., Li, X., Qiu, W., Xiao, X., & Chou, K.-C. (2018). iPPI-PseAAC(CGR): Identify protein-protein interactions by incorporating chaos game representation into PseAAC. Journal of Theoretical Biology. http://doi:10.1016/j.jtbi.2018.10.021

[30] Leite, D. M. C., Brochet, X., Resch, G., Que, Y.-A., Neves, A., & Peña-Reyes, C. (2018). Computational prediction of inter-species relationships through omics data analysis and machine learning. BMC Bioinformatics, 19(S14). http://doi:10.1186/s12859-018-2388-7

[31] Ashtiani, M., Salehzadeh-Yazdi, A., Razaghi-Moghadam, Z., Hennig, H., Wolkenhauer, O., Mirzaie, M., & Jafari, M. (2018). A systematic survey of centrality measures for protein-protein interaction networks. BMC Systems Biology, 12(1). http://doi:10.1186/s12918-018-0598-2

[32] Sachdev, K., & Kumar Gupta, M. (2019). A Comprehensive Review of Feature Based Methods for Drug Target Interaction Prediction. Journal of Biomedical Informatics, 103159. http://doi:10.1016/j.jbi.2019.103159

[33] Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., & Hoffman, M. M. (2018). Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. Information Fusion. http://doi:10.1016/j.inffus.2018.09.012

[34] Peng, J., Li, J., & Shang, X. (2020). A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. BMC Bioinformatics, 21(S13). http://doi:10.1186/s12859-020-03677-1

[35] Zhang, W., Qu, Q., Zhang, Y., & Wang, W. (2018). The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. Neurocomputing, 273, 526–534. http://doi:10.1016/j.neucom.2017.07.065

[36] Iqbal, M. J., Javed, Z., Sadia, H., Qureshi, I. A., Irshad, A., Ahmed, R., and Sharifi-Rad, J. (2021). Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. Cancer Cell International, 21(1). http://doi:10.1186/s12935-021-01981-1

[37] [37] Carracedo-Reboredo, P., Liñares-Blanco, J., Rodríguez-Fernández, N., Cedrón, F., Novoa, F. J., Carballal, A., and Fernandez-Lozano, C. (2021). A review on machine learning approaches and trends in drug discovery. Computational and Structural Biotechnology Journal, 19, 4538–4558. http://doi:10.1016/j.csbj.2021.08.011

[38] [38] Albaradei, S., Thafar, M., Alsaedi, A., Van Neste, C., Gojobori, T., Essack, M., & Gao, X. (2021). Machine learning and deep learning methods that use omics data for metastasis prediction. Computational and Structural Biotechnology Journal, 19, 5008–5018. http://doi:10.1016/j.csbj.2021.09.001

[39] Silva, J. C. F., Teixeira, R. M., Silva, F. F., Brommonschenkel, S. H., & Fontes, E. P. B. (2019). Machine learning approaches and their current application in plant molecular biology: a systematic review. Plant Science. http://doi:10.1016/j.plantsci.2019.03.020

[40] Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., Jensen, L.J. and von Mering, C., 2021. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1), pp.D605-D612. Available at: https://string-db.org