Computational Molecular Biology 2025, Vol.15, No.5, 218-226
http://bioscipublisher.com/index.php/cmb

Review Article Open Access

Machine Learning Approaches in Predicting Protein-Protein Interactions in Pathogenic Bacteria

Xing Zhao, Ming Li, Congbiao You
Tropical Microbial Resources Research Center, Hainan Institute of Tropical Agricultural Resources, Sanya, 572025, Hainan, China
Corresponding author: congbiao.you@hitar.org

Computational Molecular Biology, 2025, Vol.15, No.5 doi: 10.5376/cmb.2025.15.0021
Received: 03 Jul., 2025
Accepted: 11 Aug., 2025
Published: 05 Sep., 2025

Preferred citation for this article:
Zhao X., Li M., and You C.B., 2025, Machine learning approaches in predicting protein-protein interactions in pathogenic bacteria, Computational Molecular Biology, 15(5): 218-226 (doi: 10.5376/cmb.2025.15.0021)

Abstract The protein-protein Interactions (PPI) network of pathogenic bacteria plays a significant role in the pathogenic mechanism of bacteria and the development of drug resistance, and it is a key entry point for systems biology and new drug research and development. However, traditional PPI prediction methods (such as yeast two-hybrid and co-immunoprecipitation, etc.) have limitations such as high cost, long cycle, limited coverage, and the results are easily disturbed by noise. In recent years, the rise of machine learning, especially deep learning, has brought revolutionary progress to PPI research. With its powerful nonlinear modeling and automatic feature extraction capabilities, it has broken through the bottleneck of manual feature engineering. This paper reviews the application progress of machine learning techniques in predicting protein-protein interactions of pathogenic bacteria, with a focus on how supervised, unsupervised and deep learning methods overcome the limitations of traditional methods and improve prediction performance. Meanwhile, we discuss the impact of data preprocessing and feature engineering strategies on the model, summarize the construction and evaluation methods of machine learning models, as well as the application achievements of these models in revealing antibiotic resistance mechanisms, vaccine target screening, cross-species interactions, and other aspects. Through a case study of deep learning prediction in a Salmonella protein-protein interaction network, we verified the effectiveness and biological significance of deep learning models, and looked forward to the current challenges and future development directions.

Keywords Pathogenic bacteria; Protein-protein interactions; Machine learning; Deep learning; Graph neural network

1 Introduction

Pathogenic bacteria rely on a complete protein-protein interaction system when infecting their hosts. These PPIs determine virulence, metabolic regulation and immune evasion ability. The significance of studying the interaction network does not lie in the role of individual proteins, but in revealing the synergistic relationship of the entire pathogenic system. Like Salmonella, Mycobacterium tuberculosis, etc., their networks often have a "scale-free" and "small-world" structure, with a few hub proteins undertaking key functions. Once disrupted, the entire system will be affected (Humphreys et al., 2024). This enables PPI analysis to not only reveal biological laws but also provide new targets for the design of antibacterial drugs and vaccines.

Traditionally, protein interactions have mainly been verified through experiments, such as yeast two-hybrid, TAP-MS or protein chips. However, these methods have problems such as high false positives in pathogenic bacteria, low recognition rate of membrane proteins, and limited throughput (Ding and Kihara, 2018). Building a complete interaction group is often costly and time-consuming, making it difficult to respond quickly to new pathogenic bacteria. Thus, computational prediction gradually replaced experimental screening as the mainstream.

The rise of machine learning has completely transformed the way research is conducted. Early methods relied on manual features, such as amino acid composition and domain co-occurrence, and used SVM or random forest prediction, which were accurate but limited by human experience. Deep learning can directly learn features from sequences. The PIPR model achieves sequence-level prediction by using residual convolutional networks, and DPPI increases the AUC to above 0.8 by combining PSSM and CNN. These achievements demonstrate that even with scarce data, cross-species prediction can still be achieved with the aid of transfer learning or pre-trained models. Nowadays, machine learning enables researchers to integrate sequence, structure and functional

information to depict pathogen interaction networks within a unified framework, not only improving prediction efficiency, but also redefining the path of pathogen mechanism research.

2 The Biological Basis of Protein-Protein Interactions Among Pathogenic Bacteria

2.1 Characteristics of the protein interaction network of pathogenic bacteria

Although the protein interaction network of pathogenic bacteria is complex, it follows certain rules. Most proteins interact only with a few partners. A few "hub" proteins, such as RNA polymerase or ribosome components, are densely connected to form the network core. Networks often exhibit "small-world" and modular characteristics: functional modules such as flagella, secretory systems, and membrane synthesis are closely integrated internally, while the connections between modules are sparse. Cross-species conserved interactions (such as DNA polymerases and sliding clips) reveal evolutionary stability (Szymborski and Emad, 2024). Identifying these structural patterns helps to discover both critical and vulnerable targets for antibacterial intervention. However, the compact genomic structure and high interactivity reusability of pathogenic bacteria make network modeling more challenging.

2.2 Pathogenicity mechanism and the molecular basis of host-pathogen interaction

Infection is essentially a molecular game between the pathogen and the host. The virulence systems of bacteria, such as Salmonella type III secretory system or ESX-1 of Mycobacterium tuberculosis, are all realized through protein-protein interaction assembly. If the key interaction is impaired, the virulence will decrease. Bacteria can also reconstruct metabolism through interaction networks to resist drugs. For example, after PBP is suppressed in MRSA, the network "changes course" to maintain cell wall synthesis. Cross-species interactions are equally important. Escherichia coli effector proteins bind to host actin to facilitate its invasion. Databases such as HPIDB have integrated such data, supporting the construction of host-pathogen integration networks (James and Munoz-Munoz, 2022), and promoting machine learning predictions of cross-species interactions.

2.3 Sources of protein interaction data and experimental verification methods

A reliable PPI model cannot do without high-quality data. Positive samples mainly come from databases (BioGRID, IntAct, STRING) and literature experimental evidence. Homology inference is also an important supplement (Li and Ilie, 2017). Negative samples mostly rely on random selection or location difference method, which is noisy but practical. The prediction still needs experimental verification: Methods such as yeast two-hybrid, Co-IP, SPR, and ITC can confirm the interaction at different levels (Zhao et al., 2022). With the development of high-throughput mass spectrometry and protein chips, the verification efficiency has been continuously improved, which in turn has improved the data quality of the prediction model.

3 Principles and Classification of Machine Learning Methods in Protein-protein Interaction Prediction

3.1 Supervised learning methods

Supervised learning is the earliest machine learning method used for PPI prediction. It distinguishes between "interaction" and "non-interaction" for the trained classification model through labeled proteins. SVM is a classic representative. It can divide samples in a high-dimensional space and is suitable for small sample data, but it relies on artificial feature design (Ding and Kihara, 2018). Random Forest (RF) classiizes through voting of multiple decision trees, can handle high-dimensional features and evaluate feature importance, and its predictive performance is superior to that of SVM). Linear models such as logistic regression are mostly used as baseline references. Traditional methods rely on feature engineering to combine features such as sequence similarity, physicochemical properties, and co-expression to improve accuracy (Zhang et al., 2019), but their performance is limited under complex data, laying the foundation for deep learning.

3.2 Unsupervised and semi-supervised learning methods

Unsupervised and semi-supervised methods mine potential structures when the data lacks labels (Li and Ilie, 2017). Cluster analysis assumes that function-related proteins are more likely to interact and can detect modules, but the accuracy is affected by the threshold. Web-based link prediction algorithms that evaluate potential connections using common neighbors or random walks (Khemani et al., 2024) have been proven effective in

species such as Mycobacterium tuberculosis. The autoencoder learns latent features through compression reconstruction (Gonzalez-Lopez et al., 2018), and the variational graph autoencoder (VGAE) can directly perform unsupervised link prediction. Semi-supervised models such as GCN can propagate label information in combination with a small number of labeled samples. Although the accuracy of these methods is not as good as that of deep supervision models, they are particularly valuable in small sample scenarios.

3.3 Innovative applications of deep learning and graph neural networks in PPI prediction

Deep learning and graph neural networks (GNNS) have become new directions for PPI prediction. Sequence models such as PIPR (RCNN structure) or LSTM-CNN combined models significantly improve prediction performance. The pre-trained language models (ProtBERT, ESM) further enhanced the sequence representation (Charih et al., 2025), and the F1 values generally exceeded 0.8. The introduction of structural information and the development of AlphaFold2 have made structure-based prediction possible. GNN models such as GraphSAGE and GAT can directly learn topological features and predict missing edges on interaction networks. They can integrate sequence embeddings and network structures simultaneously, and have stronger generalization and interpretation capabilities (Khemani et al., 2024). In the future, the integration of heterogeneous maps and graph generation models will further enhance the accuracy and systematicness of pathogen interaction prediction.

4 Data Preprocessing and Feature Engineering

4.1 Sequence feature extraction

Protein sequences are the core information for PPI prediction, but it is not easy to extract useful features. The earliest method statistically analyzed the amino acid composition, divalent or trivalent frequencies, but lost the sequence information. Later Conjoint triads were grouped according to physicochemical properties and retained the local sequence. Physicochemical properties such as hydrophobicity, charge, polarity, isoelectric point, etc. are also often used to distinguish protein types (Ding and Kihara, 2018). Evolutionary information further enhances predictive power. Interacting proteins often co-evolve and can be measured by conservation scores or phyletic profile similarity. In encoding, One-hot or embedding representations such as ProtVec and ProtBERT are commonly used (Charih et al., 2025). Multi-feature fusion (sequence + structure + conservation) is often superior to single feature, but the sequence features of different species need to be standardized before modeling.

4.2 Structural and functional characteristics (protein folding, domains, GO annotations)

Structural and functional features reveal the interaction mechanism. Domain pairing is key to interaction, such as SH3 with polyproline motifs (Kotlyar et al., 2019). In machine learning, domains can be statistically co-occurring as binary features. Homologous modeling or molecular docking can obtain structural features such as interface energy and area. AlphaFold2 greatly expanded the structural data of pathogenic bacteria. Functional annotations (GO) reflect biological connections, and proteins with similar semantics are more likely to interact. Combining subcellular localization and pathway information can improve prediction accuracy, but functional similarity does not equal physical interaction. The model integrating sequence, domain and GO performed best in pathogenic bacteria (Sun et al., 2017), but feature redundancy needs to be prevented.

4.3 Data standardization and feature selection techniques (PCA, feature embedding, feature importance analysis)

Data standardization and feature selection are the keys to modeling. The dimensions of different features vary greatly and require normalization or logarithmic transformation. PCA can reduce dimension and denoise, and embedding vectors can represent category features. Feature selection can use L1 regularization, feature importance, or recursive elimination to filter out key features. It is more effective to select features in combination with biological knowledge. For example, membrane proteins should retain hydrophobic characteristics. Missing values can be filled with the mean or labeled to avoid bias. Overall, in the prediction of pathogen PPI, standardized feature engineering and preprocessing often determine success or failure more than model complexity (Ding and Kihara, 2018).

5 Construction and Evaluation of Machine Learning models

5.1 Training data and negative sample construction strategy

Building a high-quality training set is the key to PPI prediction. Positive samples are generally from experimental databases such as BioGRID and IntAct, and the difficulty lies in negative samples. The random pairing method is commonly used (Chen et al., 2019), but it is prone to mix in undiscovered true interactions. Therefore, it is recommended to avoid functionally similar proteins or utilize subcellular localization differences. There are also strategies based on functional differences or excluding co-interacting partners, and even using semi-supervised models without explicitly labeling negative samples. To prevent data imbalance, positive and negative samples are often kept at 1:1 or 1:2, and undersampling or SMOTE balance is used. Hashemifar et al. (2018) proposed dynamic negative sample refreshing of the training set. If negative samples are mixed with true positivity, performance will be underestimated. When data is scarce, it can be compensated by cross-species or transfer learning.

5.2 Model evaluation metrics

Commonly used metrics for model evaluation include accuracy rate, precision rate, recall rate, F1 and AUC. Accuracy fails when the data is unbalanced, so more attention is paid to precision (reducing false positives) and recall (discovering true positives). Drug screening focuses on accuracy, while network reconstruction emphasizes recall (Zhang et al., 2019). F1 combines the two, and AUC measures the overall discriminatory ability. The PR curve is more reliable when positive samples are scarce. Cross-validation (such as 50% fold, 10% fold) can prevent overfitting, while protein partitioning validation is closer to the actual prediction of new interaction scenarios.

5.3 Model interpretability and performance optimization methods

Although deep learning is strong, its interpretability still attracts attention. The prediction basis can be explained by feature importance, attention weight, SHAP or LIME. Grad-CAM can also mark key residues (Figure 1) (Jumper et al., 2021). In terms of performance optimization, ensemble learning can enhance robustness, hyperparameter tuning (mesh, random, Bayesian search) and regularization (L2, dropout) to prevent overfitting (Jha et al., 2022). Transfer learning can alleviate the problem of scarce pathogenic bacteria data. Active learning verifies the stepwise improvement model of high uncertainty prediction through experiments. The ultimate goal is not merely to enhance the indicators, but to reveal the interaction patterns between pathogenic bacteria through interpretable and high-performance models, promoting the integration of computation and experimentation.

6 Application and Achievements in Predicting Protein-Protein Interactions of Pathogenic Bacteria

6.1 Application in the research of antibiotic resistance mechanisms

Antibiotic resistance has become a global health crisis, and the PPI network provides an overall perspective for understanding its molecular mechanism (Maj and Trylska, 2025). In Mycobacterium tuberculosis, predictive networks reveal DNA repair and stress protein formation drug-resistant modules; In Staphylococcus aureus, β-lactam resistance protein interacts with cell wall enzymes to form a compensation circuit. This type of network analysis makes drug resistance factors no longer isolated phenomena. Interaction prediction can also identify new drug targets. For example, the interaction between Streptococcus pneumoniae MurA and topoisomerase IV is considered an interventionable bottleneck node. Furthermore, some drug-resistant mutations achieve resistance precisely by altering the protein-protein interaction interface. Comparing the interaction profiles of mutant and wild-type models can reveal this mechanism. These studies are driving anti-drug resistance strategies to shift from "inhibiting single targets" to "disrupting interaction networks", and have already shown effectiveness in *Acinetobacter baumannii* models.

6.2 Role in vaccine target screening and drug discovery

PPI prediction also plays a role in vaccine and drug development. Interaction networks help identify functionally critical and structurally exposed antigens, improving the broad-spectrum efficacy of vaccines (Lian et al., 2019). For instance, Streptococcus pneumoniae PsaA interacts closely with PspC, and the combined immune effect is

superior to that of single antigens. In terms of drugs, interaction prediction can lock onto interfacial targets. For example, blocking the binding of Escherichia coli Tir to host actin can prevent infection. Meanwhile, network analysis is also used in drug combination design to guide combination medication by identifying the synergistic interaction module. Screening projects for broad-spectrum vaccines and multi-drug combinations have entered the validation stage, demonstrating the potential of machine learning prediction to move from theory to application.

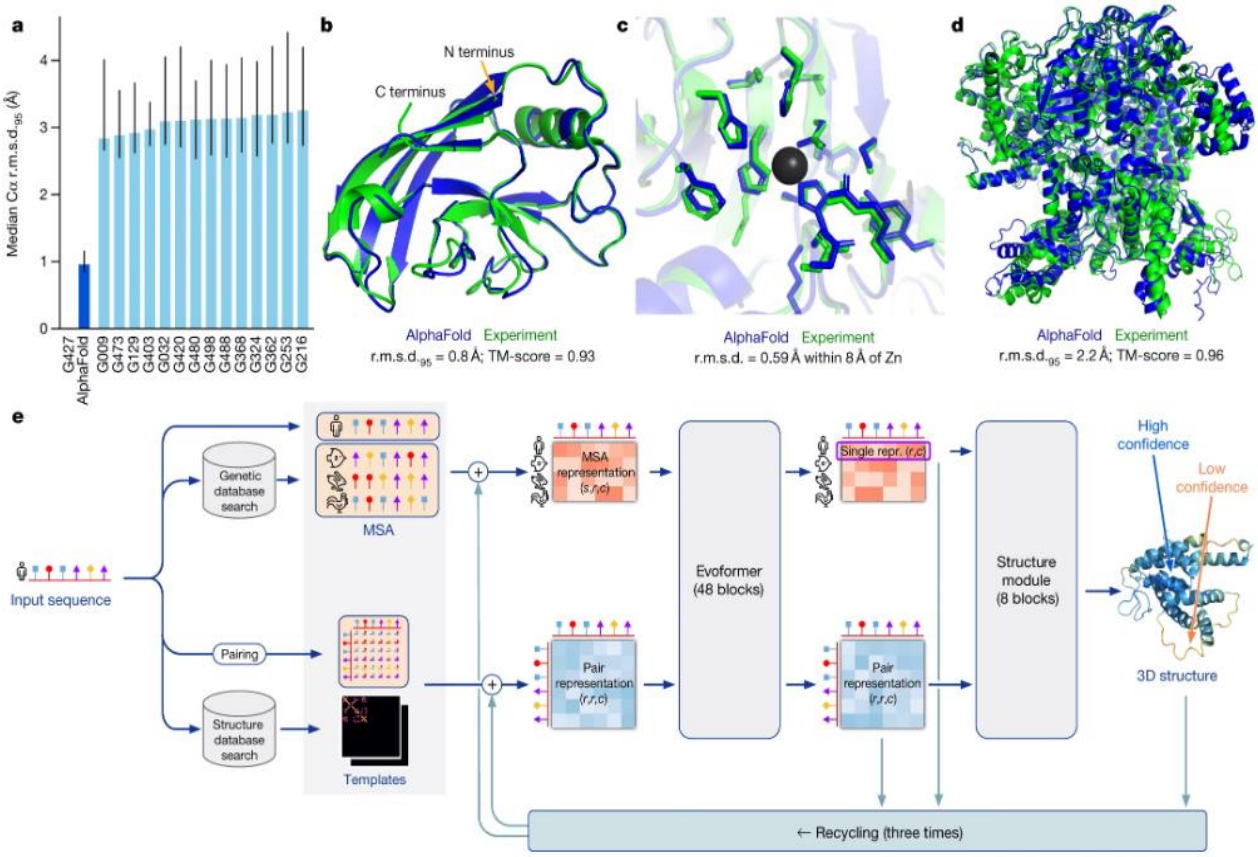

Figure 1 AlphaFold produces highly accurate structures (Adopted from Jumper et al., 2021)

6.3 Cross-species interaction prediction and integration with systems biology

Cross-species interaction prediction enables us to systematically understand the infection process. The model has been able to predict the binding of bacterial effector proteins to host targets, explaining how pathogens evade immunity or manipulate host signals. Furthermore, machine learning has also been used to infer the interaction relationship between pathogens and symbiotic bacteria. For example, the inhibition of pathogen interaction modules by short-chain fatty acids suggests probiotic potential. After integrating multi-omics information, interaction prediction becomes more biologically significant and can reveal the dynamic changes of interaction networks under infection. Currently, graph neural networks and attention mechanisms are used to integrate multi-source data, bringing us closer to the overall map of the infection system. In the future, regulating the microbiota or multi-target intervention may become a new strategy to weaken the pathogenicity of pathogens.

7 Case Study

7.1 Dataset construction and model design

Salmonella is a typical intestinal pathogen. Studying its protein interaction network helps understand the complex regulation of virulence. Here, Salmonella Typhimurium is taken as an example, using deep learning to predict its whole-genome interaction network. The data were obtained from the SalmoNet database and literature-based experimental records, with approximately 1,000 verified interactions as positive samples. For negative samples, a combination of localization differences and random pairing was adopted to select an equal number of non-interacting protein pairs from about 4,000 proteins. In terms of model design, we combined convolutional and

graph-based approaches. A Siamese-structured CNN was used to process sequences and extract local and long-range features, followed by GraphSAGE to integrate known interaction network information. The concatenated outputs of both modules were passed through a fully connected layer to predict interaction probability (Zhong et al., 2022).

Training adopted a 1:1 ratio of positive and negative samples, with 5-fold cross-validation for parameter optimization. Dropout and L2 regularization were added to prevent overfitting, and the loss function was weighted to enhance sensitivity to false negatives. The model achieved an AUC of 0.92, outperforming CNN-only (0.85) and SVM (≈0.75) models, with an F1-score of 0.84. Visualization with Grad-CAM revealed high attention weights around known binding motifs such as the arginine-rich region of Ef-Tu, aligning with experimental observations (Zhao et al., 2023). Further domain-focused attention confirmed that high-confidence interactions often occur within conserved structural regions (Charih et al., 2025). Overall, this CNN+GNN hybrid framework effectively captures Salmonella's protein interaction characteristics and demonstrates strong generalization capacity.

7.2 Model prediction results and experimental verification

The predicted network contained approximately 8,000 high-confidence interactions. Combined with known data, the full network comprised about 1,200 nodes and 8,500 edges, displaying a typical scale-free topology (Figure 2) (Muzio et al., 2020). Core hubs included ribosomal subunits and RNA polymerase components, consistent with essential metabolic functions.

Module analysis revealed three main clusters: a flagellar assembly module, a Type III secretion system (T3SS) module, and a core metabolic module, interconnected by a few regulatory proteins (Yang et al., 2020). For instance, HilA may bridge the T3SS and metabolic pathways, suggesting a coordination between virulence and metabolism. About 60% of predicted interactions were novel.

From a network perspective, the coupling between the flagellar and T3SS modules reveals that Salmonella's motility and invasion are co-regulated. Meanwhile, plasmid-encoded proteins form largely independent submodules, supporting the notion that virulence factors often operate autonomously. Altogether, the CNN+GNN model not only recovered known interactions but also uncovered biologically meaningful new links that were experimentally verified, offering novel insights into pathogenic system organization.

7.3 Implications of the results for the study of the pathogenic mechanism of salmonella

These findings shed light on Salmonella's pathogenic mechanism. Virulence is not an isolated function but part of a dynamic interaction network where motility, secretion, and metabolism are intertwined. The observed coupling between flagellar and T3SS modules indicates that Salmonella balances energy expenditure and infection efficiency through coordinated protein interactions.

The model also helped assign potential functions to previously uncharacterized proteins — for instance, protein X may regulate drug resistance by modulating TopoI activity (Charih et al., 2025). Such predictions accelerate functional annotation of hypothetical bacterial genes. Moreover, the identified interactions themselves could serve as therapeutic targets: disrupting SpiC-FlhB or TopoI-X interactions could attenuate virulence or enhance antibiotic susceptibility.

Methodologically, the CNN+GNN framework is generalizable and can be extended to other pathogens, providing computational completion for species lacking experimental interactome data. With further experimental validation, such integrative models are poised to become vital tools in pathogenic systems biology, bridging computational prediction and empirical verification for a holistic understanding of bacterial infection mechanisms (Pancino et al., 2024).

8 Challenges and Future Prospects

The prediction of PPI for pathogenic bacteria is still limited by data. The biggest problem is sample imbalance: there are few real interactions and many non-interactions, and the model is prone to bias towards the negative

class. Even if localization or functional differences are taken into account when constructing negative samples, it is still difficult to avoid treating unknown true positives as negative cases, which may cause noise. Weak supervision, generative models or sample weighting are possible remedies. Another issue is incompleteness. Most pathogenic bacteria interaction data are limited, and the model is prone to overfitting. Cross-species transfer learning can utilize model bacteria data, but species differences can still introduce errors. Experimental verification is also lagging behind, and high-throughput verification techniques still struggle to keep up with the prediction speed. Furthermore, the inconsistent data sources also lead to inconsistent reliability, and a standardized and confidence scoring system is needed. These problems are difficult to solve in the short term, but they have promoted algorithmic innovation and experimental collaboration.

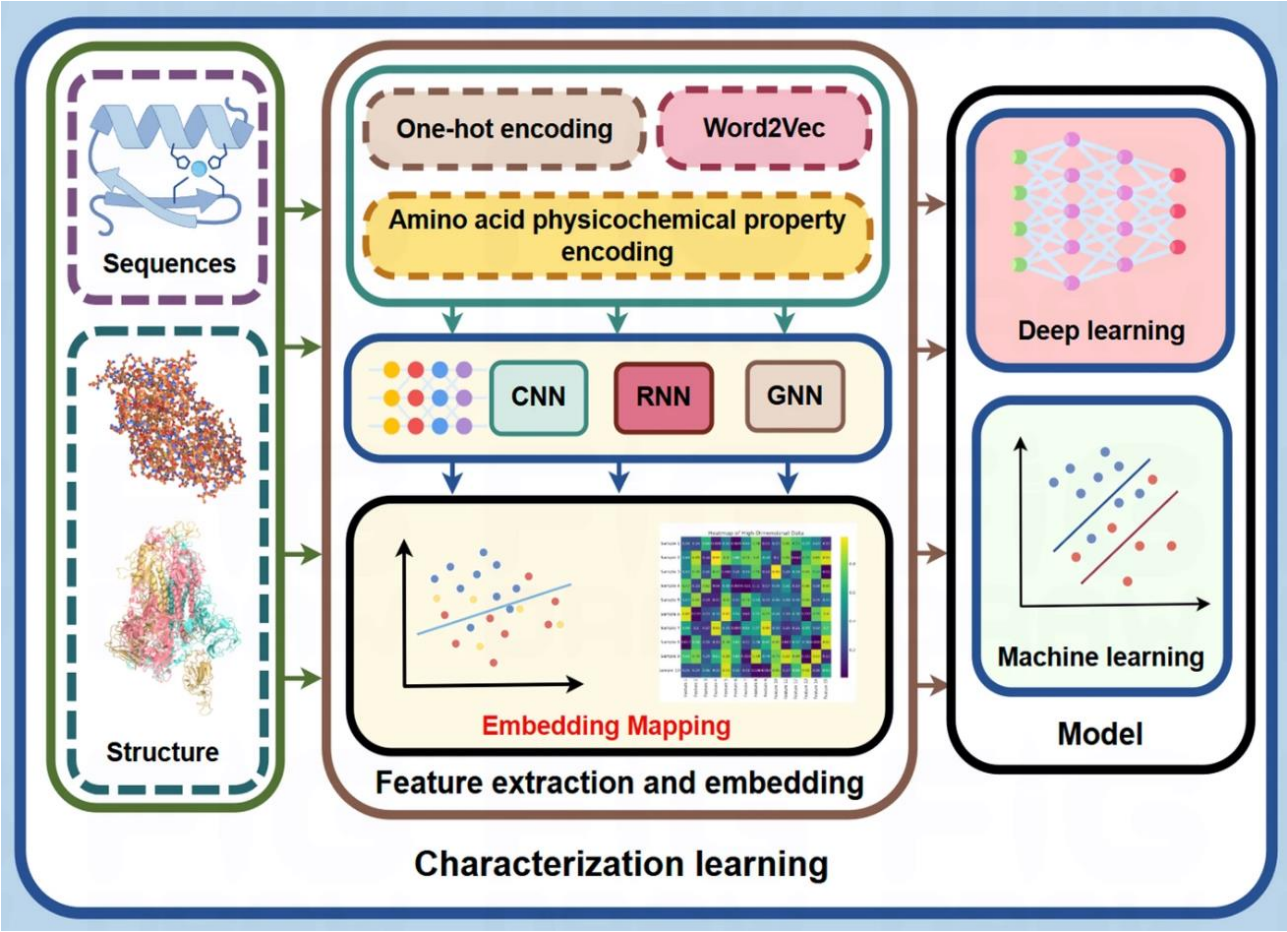

Figure 2 Protein-protein interactions characterization learning (Adopted from Muzio et al., 2020)

Cross-species generalization and interpretability are new challenges. The migration of models among different bacteria often fails because most of the captured patterns are species-specific. Joint training or introduction of species factors can improve generalization, while large pre-trained models (such as ProtBert) can learn more general features. On the other hand, the "black box" attribute of deep models makes the results hard to understand. Visualizing attention weights or introducing concept vectors can help link predictions with biometric features. Explainable structures such as graph rule networks are also under exploration. Furthermore, future models also need to deal with larger-scale "host-pathogen-microbiota" maps, and algorithm efficiency will become a bottleneck. To enhance generalization and transparency, both computational and experimental improvements are still required.

There are mainly two future directions: multi-omics integration and intelligent AI. The integration of transcriptome, metabolome and single-cell data can reveal the spatiotemporal dynamics of interactions, and dynamic graph models are being attempted. The combination of cross-species and host omics will bring predictions closer to the real ecology. In terms of algorithms, new ais such as GAN, diffusion models and

reinforcement learning can generate samples or optimize experimental designs, while structural models such as AlphaFold2 show the potential of "general interaction prediction". Ultimately, computation and experimentation will form a closed-loop system: AI prediction, experimental verification, and model update. The combination of multi-dimensional data and intelligent algorithms will drive PPI prediction into a new stage, providing more systematic support for the analysis of infection mechanisms and antibacterial strategies.

Acknowledgments

The authors extend sincere thanks to two anonymous peer reviewers for their invaluable feedback on the manuscript.

Conflict of Interest Disclosure

The authors affirm that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Charih F., Green J.R., and Biggar K.K., 2025, Sequence-based protein-protein interaction prediction and its applications in drug discovery, Cells, 14(18): 1449.
https://doi.org/10.3390/cells14181449

Chen M., Ju C., Zhou G., Chen X., Zhang T., Chang K., Zaniolo C., and Wang W., 2019, Multifaceted protein-protein interaction prediction based on Siamese residual RCNN, Bioinformatics, 35(14): i305-i314.
https://doi.org/10.1093/bioinformatics/btz328

Ding Z., and Kihara D., 2018, Computational methods for predicting protein-protein interactions using various protein features, Current Protocols in Protein Science, 93(1): e62.
https://doi.org/10.1002/cpps.62

Gonzalez-Lopez F., Morales-Cordovilla J.A., Villegas-Morcillo A., Gomez A.M., and Sanchez V., 2018, End-to-end prediction of protein-protein interaction based on embedding and recurrent neural networks, In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp.2344-2350.
https://doi.org/10.1109/BIBM.2018.8621328

Hashemifar S., Neyshabur B., Khan A.A., and Xu J., 2018, Predicting protein-protein interactions through sequence-based deep learning, Bioinformatics, 34(17): i802-i810.
https://doi.org/10.1093/bioinformatics/bty573

Humphreys I.R., Zhang J., Baek M., Wang Y., Krishnakumar A., Pei J., Anishchenko I., Tower C., Jackson B., Warrier T., Hung D., Peterson S., Mougous J., Cong Q., and Baker D., 2024, Protein interactions in human pathogens revealed through deep learning, Nature Microbiology, 9(10): 2642-2652.
https://doi.org/10.1038/s41564-024-01791-x

James K., and Muñoz-Muñoz J., 2022, Computational network inference for bacterial interactomics, Msystems, 7(2): e01456-21.
https://doi.org/10.1128/msystems.01456-21

Jha K., Saha S., and Singh H., 2022, Prediction of protein-protein interactions using graph neural networks, Scientific Reports, 12(1): 8360.
https://doi.org/10.1038/s41598-022-12201-9

Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., Bates R., Žídek A., Potapenko A., Bridgland A., Meyer C., Kohl S., Ballard A., Cowie A., Romera-Paredes B., Nikolov S., Jain R., Adler J., Back T., Petersen S., Reiman D., Clancy E., Zielinski M., Steinegger M., Pacholska M., Berghammer T., Bodenstein S., Silver D., Vinyals O., Senior A., Kavukcuoglu K., Kohli P., and Hassabis D., 2021, Highly accurate protein structure prediction with AlphaFold, Nature, 596(7873): 583-589.
https://doi.org/10.1038/s41586-021-03819-2

Khemani B., Patil S., Kotecha K., and Tanwar S., 2024, A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions, Journal of Big Data, 11(1): 18.
https://doi.org/10.1186/s40537-023-00876-4

Kotlyar M., Pastrello C., Malik Z., and Jurisica I., 2019, IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species, Nucleic Acids Research, 47(D1): D581-D589.
https://doi.org/10.1093/nar/gky1037

Lee M., 2023, Recent advances in deep learning for protein-protein interaction analysis: a comprehensive review, Molecules, 28(13): 5169.
https://doi.org/10.3390/molecules28135169

Li Y., and Ilie L., 2017, SPRINT: ultrafast protein-protein interaction prediction of the entire human interactome, BMC Bioinformatics, 18(1): 485.
https://doi.org/10.1186/s12859-017-1871-x

Lian X., Yang S., Li H., Fu C., and Zhang Z., 2019, Machine-learning-based predictor of human-bacteria protein-protein interactions by incorporating comprehensive host-network properties, Journal of Proteome Research, 18(5): 2195-2205.
https://doi.org/10.1021/acs.jproteome.9b00074

Maj P., and Trylska J., 2025, Protein-protein interactions as promising molecular targets for novel antimicrobials aimed at Gram-negative bacteria, International Journal of Molecular Sciences, 26(22): 10861.
https://doi.org/10.3390/ijms262210861

Muzio G., O'Bray L., and Borgwardt K., 2020, Biological network analysis with deep learning, Briefings in Bioinformatics, 22(2): 1515-1530.
https://doi.org/10.1093/bib/bbaa257

Pancino N., Gallegati C., Romagnoli F., Bongini P., and Bianchini M., 2024, Protein-protein interfaces: a graph neural network approach, International Journal of Molecular Sciences, 25(11): 5870.
https://doi.org/10.3390/ijms25115870

Sun T., Zhou B., Lai L., and Pei J., 2017, Sequence-based prediction of protein-protein interactions using a deep-learning algorithm, BMC Bioinformatics, 18(1): 277.
https://doi.org/10.1186/s12859-017-1700-2

Szymborski J., and Emad A., 2024, INTREPPPID: an orthologue-informed quintuplet network for cross-species prediction of protein-protein interaction, Briefings in Bioinformatics, 25(5): bbae405.
https://doi.org/10.1093/bib/bbae405

Yang F., Fan K., Song D., and Lin H., 2020, Graph-based prediction of protein-protein interactions with attributed signed graph embedding, BMC Bioinformatics, 21(1): 323.
https://doi.org/10.1186/s12859-020-03646-8

Zhang L., Yu G., Xia D., and Wang J., 2019, Protein-protein interactions prediction based on ensemble deep neural networks, Neurocomputing, 324: 10-19.
https://doi.org/10.1016/j.neucom.2019.04.002

Zhao N., Zhuo M., Tian K., and Gong X., 2022, Protein-protein interaction and non-interaction predictions using gene sequence natural vector, Communications Biology, 5(1): 652.
https://doi.org/10.1038/s42003-022-03617-0

Zhao Z., Qian P., Yang X., Zeng Z., Guan C., Tam W., and Li X., 2023, Semignn-ppi: self-ensembling multi-graph neural network for efficient and generalizable protein-protein interaction prediction, arXiv preprint, arXiv preprint, 2305: 8316.
https://doi.org/10.24963/ijcai.2023/554

Zhong Y., He S., Xiao C., Liu Y., Qin X., and Yu Z.3, 2022, Long-distance dependency combined multi-hop graph neural networks for protein-protein interaction prediction, BMC Bioinformatics, 23(1): 521.
https://doi.org/10.1186/s12859-022-05062-6