OXFORD

Briefings in Bioinformatics, 2025, **26(5)**, bbaf461

https://doi.org/10.1093/bib/bbaf461

Problem Solving Protocol

Graph neural network integrated with pretrained protein language model for predicting human–virus protein–protein interactions

Linyang Jiang[1], Xiaodi Yang[2], Xiaokun Guo[1], Dianke Li[1], Jiajun Li[1], Stefan Wuchty[3,4,5], Wenyu Shi[1,*], Ziding Zhang[1,*]

[1]State Key Laboratory of Animal Biotech Breeding, College of Biological Sciences, China Agricultural University, No. 2 Yuanmingyuan West Road, Haidian District, Beijing 100193, China

[2]Department of Hematology, Peking University First Hospital, No. 8 Xishiku Street, Xicheng District, Beijing 100034, China

[3]Department of Computer Science, University of Miami, 310A Ungar Building, 1365 Memorial Drive, Coral Gables, FL 33146-4245, USA

[4]Department of Biology, University of Miami, 215 Cox Science Center, 1301 Memorial Drive, Coral Gables, FL 33124-0421, USA

[5]Sylvester Comprehensive Cancer Center, University of Miami, 1475 NW 12th Ave, Miami, FL 33136, USA

*Corresponding authors. Wenyu Shi, State Key Laboratory of Animal Biotech Breeding, College of Biological Sciences, China Agricultural University, No. 2 Yuanmingyuan West Road, Haidian District, Beijing 100193, China. E-mail: shiwy@cau.edu.cn; Ziding Zhang, State Key Laboratory of Animal Biotech Breeding, College of Biological Sciences, China Agricultural University, No. 2 Yuanmingyuan West Road, Haidian District, Beijing 100193, China. E-mail: zidingzhang@cau.edu.cn.

Abstract

The systematic identification of human-virus protein–protein interactions (PPIs) is a critical step toward elucidating the underlying mechanisms of viral infection, directly informing the development of targeted interventions against existing and emerging viral threats. In this work, we presented DeepGNHV, an end-to-end framework that integrated a pretrained protein language model with structural features derived from AlphaFold2 and leveraged graph attention networks to predict human-virus PPIs. In comparison to other state-of-the-art approaches, DeepGNHV exhibited superior predictive performance, especially when applied to viral proteins absent from the training process, indicating its strong generalization capability for detecting newly emerging virus-related PPIs. We further demonstrated DeepGNHV's robustness across diverse perturbations and its practical application under high-confidence thresholds. Additionally, we conducted extensive predictions of human-HPV PPIs, which were supported by multiple lines of evidence and identified several host factors that specifically interact with high-risk HPV. To further explore the biological significance of DeepGNHV, we provided a case study to pinpoint specific residues that play critical roles in facilitating the corresponding PPIs. The source code of DeepGNHV and related data is publicly available on GitHub (https://github.com/bioboy0415/DeepGNHV).

Keywords: human-virus PPIs; graph neural network; pretrained protein language model; AlphaFold2; human papillomavirus

Introduction

Humans are exposed to numerous viruses over their lifetimes, most of which are effectively managed by the host immune system [1]. Nevertheless, millions of deaths are still caused by viral infections each year due to the vast diversity of viruses [2]. Protein–protein interactions (PPIs) between humans and viruses are crucial for understanding viral infections and host immune response [3]. In-depth investigations of them contribute to a better understanding of the complex human-virus relationships, offering clues for developing novel strategies for the prevention and treatment of viral infections [4]. In recent years, widely used techniques such as yeast two-hybrid, affinity purification mass spectrometry (AP-MS), and protein microarrays have resulted in a growing number of human-virus PPIs, considerably advancing efforts to develop vaccines and antiviral therapies [5–7]. However, the distribution of these interactions remains markedly imbalanced, with a disproportionate number discovered for well-studied viruses such as SARS-CoV-2, influenza, and HIV. In contrast, many other viruses remain underexplored, pointing to numerous uncovered potential human-virus PPIs. Given the time-intensive and costly nature of the experimental PPI identification, the ability to preselect more likely interacting protein pairs through computational methods can enhance the overall efficiency of the human-virus interactomic studies.

A series of traditional methods have been developed for PPI prediction, including interolog mapping [8, 9], domain–domain–based interaction inference [10] and protein docking-based interaction inference [11–13]. However, these approaches frequently demonstrate inadequate predictive accuracy and limited applications. Many machine learning (ML) and deep learning (DL) methods are widely applied to predict PPIs within species and achieve higher predictive performance than traditional approaches. In recent years, several ML-based methods have been developed to tackle the challenge of predicting interspecies interactions between human and viral proteins. For instance, our team developed the HVPPI [14] by integrating Doc2vec with random forest (RF) models and TransPPI [15] by leveraging position-specific scoring matrices with a Siamese convolutional neural

Received: April 23, 2025. **Revised:** July 20, 2025. **Accepted:** August 12, 2025

network (CNN) and transfer learning strategies. Other notable contributions include DeepViral [16], which combines information from protein sequences, functional phenotypes, and taxonomic families alongside CNN for human-virus PPI prediction. Tsukiyama *et al.* further advanced the field by combining cross-attention mechanisms with 1D-CNN to develop the Cross-attention PHV model (CAPHV) [17]. More recently, Zhang *et al.* (2024) proposed HBFormer [18], which integrates ProtT5 embeddings with a hybrid attention mechanism for improved human–virus PPI prediction. Nevertheless, as a consequence of the limited availability of structural data for viral proteins, interspecies predictions of human-virus PPIs largely depend on sequence-based approaches. However, the spatial arrangement of amino acid residues within protein structures is also crucial for these interactions. Additionally, the information in the previously used encoding schemes is relatively limited, restricting the models' ability to capture the intricate details of these interactions.

Over the past few years, pretrained protein language models, such as ESM2 [19], ESM-MSA-1b [20], ProtTrans [21], ProteinBERT [22], SaProt [23], and TAPE [24], were shown to capture complex relationships between amino acids through unsupervised training on large-scale protein sequence or structure datasets and enhance the performance of downstream prediction tasks significantly, such as PPI prediction [18], variant effect prediction [25], protein engineering [26]. Furthermore, AlphaFold2 [27] achieved a considerable breakthrough in predicting protein tertiary structures, with a released comprehensive database encompassing over 200 million predicted protein monomer structures for the human proteome and 47 other key organisms [28].

Historically, the predominant models developed for predicting human-virus PPIs have mainly encompassed RF, CNN, LSTMs, and attention-based networks, all exhibiting commendable efficacy. Nevertheless, these approaches are more suitable for dealing with protein sequence data. In recent years, graph neural networks (GNNs) have demonstrated exceptional adaptability to leverage protein structural data. There are a growing number of studies applying GNNs to structural bioinformatics tasks, including PPI prediction [29], binding site identification [30], and protein function annotation [31]. In this study, we aim to develop a generalizable model for human-virus PPI prediction, focusing on interactions involving previously uncharacterized viruses. We also systematically explore PPIs associated with human papillomavirus (HPV) across multiple subtypes to identify shared and unique interaction patterns as well as key factors that contribute to a deeper understanding of the underlying mechanisms. Furthermore, we pinpoint some key residue sites that play a crucial role in the interaction predictions independently of any residue-level annotations.

Materials and methods

Dataset construction

We compiled experimentally validated human-virus PPIs (i.e. positive samples) from six public databases, including BIOGRID [32], IntAct [33], HPIDB [34, 35], MINT [36], PHISTO [37], and VirHostNet [38]. To ensure data quality, we removed duplicates, nonphysical PPIs, proteins shorter than 30 amino acids, and PPIs identified only once through large-scale MS experiments. As a result, we obtained 31 142 nonredundant human-viral PPIs between 6729 human proteins and 1154 viral proteins.

For dataset splitting, we employed a virus protein-based five-fold cross-validation strategy. The viral proteins in the positive samples were randomly divided into five parts. In each iteration, four parts were used to construct the training set (70%), validation set (15%), and C1Virus test set (15%), while the remaining part served as the C2Virus test set. The C2Virus test set was specifically constructed to simulate scenarios where the models predict PPIs involving novel viral proteins. This process was repeated five times to mitigate the bias introduced by the selection of viral proteins, ensuring that each viral protein appeared in the C2Virus test set only once across all iterations (Supplementary Fig. 1 and Supplementary Table 1). Regarding negative sample selection, we first randomly paired human and viral proteins from the positive samples. After excluding the known PPIs, we applied "Dissimilarity-based negative sampling" method [39] (sequence identity < 60%) to minimize potential false negative samples while randomly selecting negative samples at a quantity 10 times that of the positive samples for each dataset.

We obtained the predicted structural models of human proteins from the AlphaFold Protein Structure Database [28]. Considering that the structural models for most viral proteins are not available in the database, we predicted the structures of these viral proteins using the locally deployed AlphaFold2 [27]. The predicted local distance difference test distributions of human/viral proteins and the corresponding AlphaFold2 parameter settings in predicting viral protein structures are provided in Supplementary Fig. 2.

Protein representation strategies

To capture the intricate dependencies and inherent patterns among amino acid residues, we employed the ProtT5-XL-Uniref50 model (ProtT5) [21] with 3 billion parameters, pretrained on the Big Fasta Database dataset and fine-tuned on the UniRef50 dataset for protein embedding. Specifically, we downloaded the source code and parameters of ProtT5 from https://github.com/agemagician/ProtTrans and executed it locally. For a human or viral protein sequence of length N, we input it into the frozen ProtT5 model and obtained an $N\times1024$ matrix from the final neural network layer after removing the end tokens, where each row of the matrix corresponds to the embedding of a specific amino acid.

We then used the graph structure $G = (V, E)$ to represent the 3D structure of a human or virus protein, where V denoted amino acid residues in the protein, and E indicated the contact relationships between these residues. The contact relationships were determined by assessing whether the Euclidean distance between the C_α atoms of any two residues in the protein monomer model was less than a specified threshold of 8 Å. This criterion was used to construct the adjacency matrix, while the embedding matrix from ProtT5 was used to encode the feature matrix of V for residue-level annotations. We also explored the impact of different structural representations on the model's generalization performance, including focusing solely on surface residues or transforming the undirected graph into a directed one by modifying the adjacency matrix. The Undirected graph uses all amino acid residues in human or viral proteins to construct the adjacency matrix and embedding vectors, with bidirectional edges between neighboring residues. The Surface graph only includes surface residues, excluding those with a relative solvent accessibility lower than 0.3. The Semidirected graph extends the Undirected graph by modifying the adjacency matrix in the final layer, removing edges to buried residues while retaining self-connections for all nodes.

Model architecture

In this study, we constructed a two-layer Graph Attention Network (GATConv) [40] to extract key structural information of PPI prediction while progressively reducing the dimensionality of

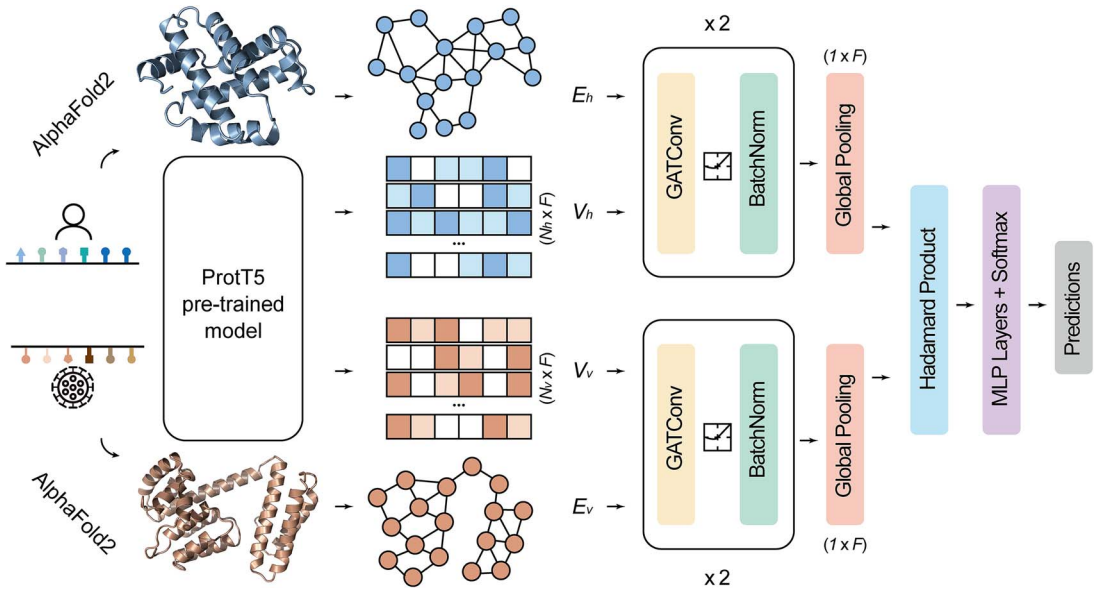

Figure 1. Workflow of human-virus PPI prediction based on structural models and pretrained protein language model. Firstly, we obtain the structural models of human and viral proteins predicted by AlphaFold2. Each protein is represented as a graph, where a embedding matrix generated by ProtT5 encodes information about the amino acid residues, and an adjacency matrix represents the contact relationship between all residues. Then, we employ a two-layer weight-sharing Siamese GATConv to progressively update the protein embeddings with the help of the adjacency matrix. After applying global max pooling operations, we compute the Hadamard product of both protein representations and submit them to the Multilayer Perceptron (MLP) layers for PPI predictions.

embedding matrices (from 1024 to 720, and then to 360) to ease the computational load. Given that human-virus PPI prediction is a pairwise heterogeneous input problem, we utilized a weight-sharing Siamese network updating strategy for the GATConv. The updated embedding matrices of human and viral proteins were normalized through max pooling operations followed by Hadamard product to capture local information and submitted to the fully connected layers for classification (Fig. 1).

Specifically, for a human or viral protein of length N, the amino acid embedding matrices $H = \left\{\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_N\right\}$, $\vec{h}_i \in R^F$ are updated using the GATConv [40] provided by PyTorch Geometric [41]. The updated embedding matrices are represented as $H' = \left\{\vec{h}'_1, \vec{h}'_2, \ldots, \vec{h}'_N\right\}$, $\vec{h}'_i \in R^{F'}$, where F and F' denote the feature dimensions before and after transformation, respectively. A fully connected layer $W \in R^{F' \times F}$ projects the embedding matrices to the specified dimensionality for attention score calculation. The attention score represents the strength of the association between amino acids i and j, and is calculated as $\boldsymbol{a}^T\left[W\vec{h}_i \| W\vec{h}_j\right]$, where $\boldsymbol{a} \in R^{2F'}$ is a fully connected layer. The attention scores are then normalized using a softmax function, which is applied exclusively to neighboring amino acids. To introduce non-linearity and improve generalization, a LeakyReLU activation is applied to the attention scores. This process is encapsulated in the following formula:

$$\alpha_{ij} = \frac{\exp\left(LeakyReLU\left(\boldsymbol{a}^T\left[W\vec{h}_i \| W\vec{h}_j\right]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(LeakyReLU\left(\boldsymbol{a}^T\left[W\vec{h}_i \| W\vec{h}_j\right]\right)\right)} \quad (1)$$

The obtained attention scores α_{ij} are utilized for the linear integration of the embedding matrices of all amino acids adjacent to amino acid i. This integration is repeated k times, and the resulting embedding matrices are concatenated to capture information from multiple perspectives, followed by a nonlinearity σ:

$$\vec{h}_{i'} = \Big\|_{k=1}^{K}\sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}^{(k)}W^{(k)}\vec{h}_j\right) \quad (2)$$

Batch normalization is applied after each GATConv layer to improve the model's generalization performance. After updating the human and viral protein embedding matrices using the weight-sharing two-layer GATConv, global max pooling is performed on the resulting features, followed by the Hadamard product $\odot$, and then passing them into a prediction module consisting of two fully connected layers f to obtain final interaction predictions:

$$\hat{y} = Softmax\left(f\left(\vec{h}_{g(human)''} \odot \vec{h}_{g(virus)''}\right)\right) \quad (3)$$

In the training step, we used a batch size of 64 and a dropout rate of 0.3 to mitigate overfitting. The GATConv employed 12 attention heads. Optimization was performed using the AdamW optimizer with an initial learning rate of 3e-4. Additionally, we adopted a ReduceLROnPlateau scheduler with a reduction factor of 0.5, a patience of two epochs, and a minimum learning rate of 5e-5. All experiments were conducted on a single NVIDIA RTX 3090 GPU (24GB memory).

Performance evaluation

To assess the performance of different predictive models on the benchmark datasets, we plotted the Precision-Recall (PR) curves to visualize the precision and recall values at various thresholds. Moreover, we calculated the area under the PR curve (AUPRC) for each model, which is particularly suitable for assessing imbalanced classification tasks. We also drew the Receiver Operating Characteristic (ROC) curve to visualize the trade-off between false positive rate (FPR) and the true positive rate. Likewise, the area under the ROC curve (AUROC) for each model was also calculated.

The computations were conducted using the scikit-learn [42] package in Python. Additionally, we employed several other common evaluation metrics to supplement the model performance assessment, as described below:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = TPR = \frac{TP}{TP + FN}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

where TP, FP, TN and FN indicate the numbers of correct and incorrect predictions for positive and negative samples, respectively.

Results and discussion

The overall performance of DeepGNHV

To assess the performance of our proposed DeepGNHV model, we compared it with five state-of-the-art methods (HBFormer, TransPPI, HVPPI, CAPHV, and DeepViral) for predicting human-virus PPIs, focusing on the performance of both the C1Virus dataset and the more challenging C2Virus dataset. To ensure fair comparison, we obtained and locally implemented all competing methods using their publicly available source codes from GitHub. The results showed that DeepGNHV achieved the highest AUPRC on the C1Virus dataset, slightly outperforming TransPPI and considerably surpassing the other models (Fig. 2A and B). While all models exhibited reduced prediction performance on the C2Virus dataset compared to the C1Virus dataset, this reflected the challenge of accurately predicting interactions for novel viral proteins. DeepGNHV achieved the highest AUPRC of 0.573, followed by HBFormer, TransPPI, HVPPI, CAPHV, and DeepViral (Fig. 2C and D). The above result indicated DeepGNHV's superior predictive performance, particularly on the more challenging dataset, thereby contributing significantly to the real application.

Recently, the latest version of AlphaFold (i.e. AlphaFold3) [43] achieved a massive success in predicting the 3D complex structure of two interacting proteins, where the parameter of interface pTM score (ipTM) was reported to judge if two proteins can interact or not from the predicted 3D complex structure. Generally, a higher ipTM points to a higher interaction likelihood. Therefore, it is also interesting to compare the proposed DeepGNHV and other existing human-virus PPI predictors against AlphaFold3 in distinguishing protein interactions from non-interactions. Considering the current AlphaFold3 web server only allows the daily processing of 30 predictions, we benchmarked its performance on a subset of the C2Virus dataset, comprising 100 positive and 200 negative samples randomly selected. The results indicate that AlphaFold3 performs inferior to DeepGNHV and other existing human-virus PPI predictors (Supplementary Table 2). Such a result implies that the simple metric inferred from AlphaFold3 cannot detect human-virus PPIs accurately, although AlphaFold3 was explicitly designed for 3D structure prediction of interacting proteins.

Analysis of different graph representations indicated that the Undirected graph (incorporating all amino acid residues) slightly outperformed both the Surface graph and the Semidirected graph (Supplementary Table 3). It is widely acknowledged that surface residues play a crucial role in protein physical interactions; however, buried residues within a protein structure contribute to the integrity of the protein's structure and can provide useful information in PPI prediction. Considering that the information of buried residue is neglected in the Surface graph and the Semidirected graph, the Undirected graph yielded more favorable performance.

We examined the performance of various methods across different virus categories in the C2Virus dataset (Fig. 3A). Overall, DeepGNHV demonstrated superior predictive performance in most virus categories compared to other methods, except for Hepatitis. On the other hand, the performance on different viruses is different. Notably, all methods achieved exceptionally high performance on human-HIV PPI prediction, which is likely attributed to the substantially larger training set size ($n = 9347$ verified interactions) compared to other virus categories. However, this pattern does not extend to influenza A virus , which showed poorer performance despite having the second-highest number of PPIs, indicating that prediction difficulty depends on factors beyond mere interaction counts. Thus, we further examined the relationship between intra-virus protein sequences and their interaction profiles in HIV and other viruses (Supplementary Fig. 3). We found that HIV is characterized by a substantial number of highly similar HIV protein pairs sharing the same human host interacting partners, which is less frequently observed in other viruses, particularly in Hepatitis and SARS-CoV. Due to this potential dataset bias, predicting the PPIs between Hepatitis/SARS-CoV and their human host yielded a relatively weak performance.

The practicality and robustness assessment

While ROC and PRC curves demonstrated the comprehensive predictive performance of the models across various thresholds, we specifically evaluated model utility in high-confidence prediction scenarios relevant to real-world applications. We calculated the number of true positives predicted by each model on the C2Virus dataset at a FPR control of 0.01 and visualized the results using UpSetR [44] (Fig. 3B). The results showed that DeepGNHV not only predicted the highest number of positive PPIs in high-confidence scenarios but also yielded the most correctly predicted PPIs shared with other methods.

To assess the robustness of the DeepGNHV model, we introduced various levels of perturbations into the training dataset to simulate potential false positives in public databases, as well as possible false negatives in the constructed negative samples. Specifically, we randomly swapped the labels between positive and negative samples in the training set to increase noise. For instance, a 20% perturbation involved swapping the labels of 20% of the positive samples with an equal number of negative samples. The models were then retrained and evaluated on the joint test dataset derived from the C1Virus and C2Virus datasets. While the predictive performance of all models gradually declined as the perturbation proportion increased. DeepGNHV exhibited a superior noise tolerance compared to other models (Fig. 3C). This systematic evaluation confirms DeepGNHV's enhanced robustness to data quality issues commonly encountered in PPI studies.

Human-HPV protein–protein interaction prediction and analysis

To exemplify the real application of DeepGNHV, we selected HPV for large-scale interaction predictions to enhance our understanding of its infection mechanisms. HPV is a non-enveloped double-stranded DNA virus that primarily enters basal layer cells through wounds in the skin and mucous membranes where it initiates replication. By hijacking host cellular machinery, HPV establishes a low-copy persistent infection in basal cells and

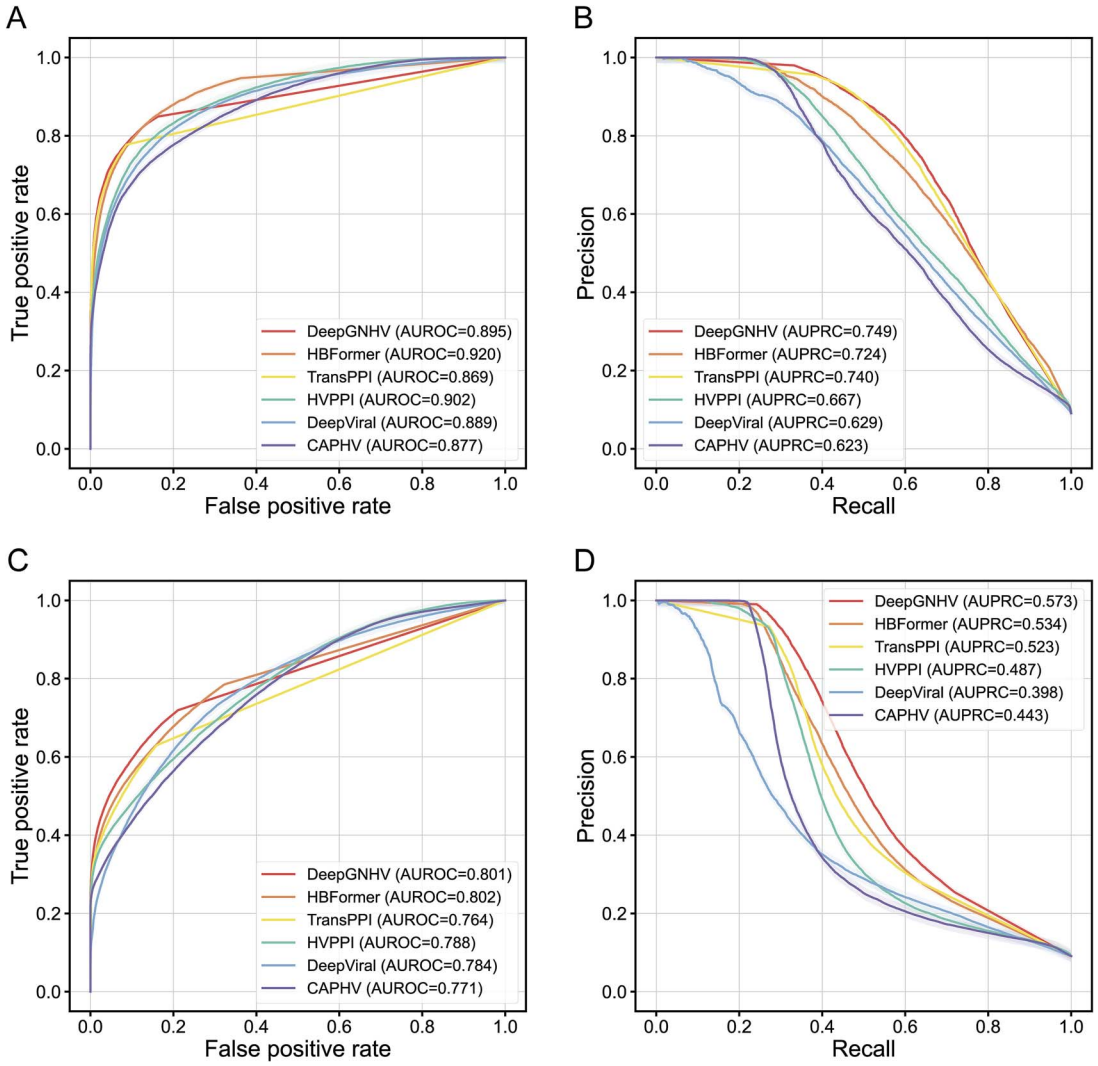

Figure 2. Performance of DeepGNHV compared to other existing methods in predicting human-virus PPIs. Panels A and B plot ROC and PRC curves on the C1Virus dataset, while panels C and D plot ROC and PRC curves on the C2Virus dataset.

produces viral particles within terminally differentiated epithelial cells, which are eventually shed to release the virus. Based on their pathogenicity and phylogeny, HPV can be mainly classified into high-risk HPV (HR-HPV) and low-risk HPV (LR-HPV). HR-HPV is closely associated with cancer, particularly cervical cancer, while LR-HPV primarily causes benign warts [45].

We conducted large-scale human-HPV PPI predictions across the human proteome, with six common HR-HPV and six LR-HPV subtypes. Due to the considerable differences in the number of recorded PPIs across HPV subtypes, we selected the top 500 high-confidence predictions per subtype for further analysis, covering 2521 human proteins and 87 HPV proteins. Initially, we analyzed the degree distribution of human proteins related to the predicted PPIs (Fig. 4A). The results indicated that most human proteins were targeted by an individual viral protein, while a small number of human proteins were targeted by multiple viral proteins. Indeed, it follows a typical power law distribution, which is in line with previous host-pathogen PPI network topologies, thereby indirectly supporting the reliability of our predictions [46]. Subsequently, we integrated the predicted and experimentally validated PPIs from various HPV subtypes to perform a comprehensive Gene Ontology (GO) enrichment analysis on the targeted human proteins, followed by the calculation of the Jaccard similarity matrix and the generation of a heatmap along with clustering analysis (Fig. 4B). The result suggests that HR-HPV subtypes tend to cluster more tightly than LR-HPV subtypes, which may reflect underlying biological differences. HR-HPV subtypes, associated with more severe diseases, tend to interact with more specific and functionally coherent host factors. In contrast, LR-HPV subtypes, generally linked to milder pathogenicity, appear to rely on more diverse and inconsistent interaction partners, leading to a more scattered clustering pattern. Focusing on HR-HPV, we identified 23 human proteins that specifically interacted with at least five HR-HPV subtypes, 22 of which have been experimentally validated to interact with HR-HPV, with the exception of LRRC7 (Supplementary Table 4). Interestingly, LRRC7 plays a key role in postsynaptic density complex localization and NMDA receptor-dependent long-term depression [47], which could be indirectly exploited by HR-HPV to modulate the host cell's metabolic and proliferative states, thereby facilitating its replication and persistence. Subsequently, we conducted GO enrichment analysis on the 23 human genes that specifically interact with HR-HPV (Fig. 4C). This analysis revealed several pathways that have been previously reported to be closely associated with the viral entry, persistence, and oncogenic transformation of HR-HPV infection, including the establishment or maintenance of epithelial cell

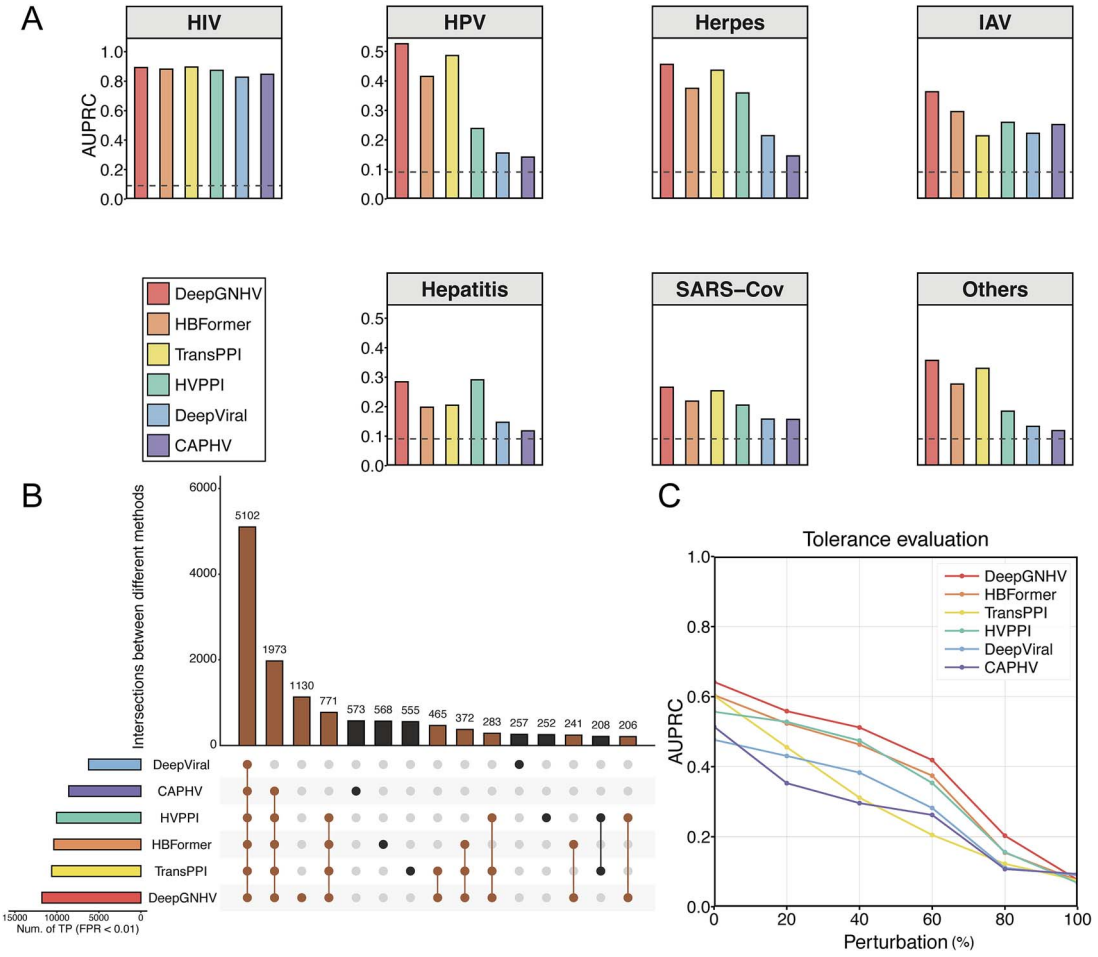

Figure 3. The practicality and robustness analysis of DeepGNHV for human-virus PPI predictions. (A) Performance across C2Virus datasets associated with different virus-related proteins. Dashed lines indicate the AUPR baselines of random classifiers for each virus. (B) Under high-confidence conditions (FPR < 0.01), the number of true-positive PPIs predicted by different models on the C2Virus test dataset is shown. (C) Robustness evaluation presents the AUPRC across different prediction methods under various degrees of training dataset perturbations, where positive and negative samples are randomly replaced at varying ratios. Then, the models were evaluated on the joint test datasets (i.e. C1Virus + C2Virus).

apical/basal polarity [48], receptor localization to synapse [49], and PDZ domain binding [50]. These findings provide functional support for the validity of our predictions and illustrate the utility of DeepGNHV for identifying biologically meaningful interactions.

As E6 is one of the most functionally divergent proteins between HR-HPV and LR-HPV, we performed multiple sequence alignment (MSA) on the E6 proteins of 12 selected HPV subtypes (Supplementary Fig. 4). Based on this alignment, we defined representative residues as those at positions where a specific amino acid appears with over 50% in either HR-HPV or LR-HPV. Positions with the same representative residues in both groups are labeled as Conserved Regions; those with differing representative residues or present in only one group were labeled as Distinct Regions; other positions are labeled as Noise Regions. We then selected 104 PPIs targeted by the E6 protein of HR-HPV with the 23 human proteins previously identified and introduced single-amino acid perturbations (using zero embedding) to assess the absolute changes in interaction scores (Fig. 4D). To assess the mutation effects, positions with a score change exceeding 0.02 were defined as notable alterations (Supplementary Table 5). The results showed that Conserved Regions exhibited a significantly higher proportion of notable alterations compared to Noise Regions (Fisher's exact test, $P = 0.0004$). Similarly, Distinct Regions also showed an elevated proportion relative to Noise Regions (Fisher's exact test, $P = 0.0326$).

Analysis of the key residues for DeepGNHV interaction prediction

Protein interaction prediction is typically regarded as a binary classification problem. Existing methods for predicting human-virus PPIs primarily focus on the predictive accuracy of the models while neglecting the investigation of key amino acid residues responsible for PPI prediction. In this study, we applied the GNNExplainer [51] method to the DeepGNHV model to identify key residues for predicting human-virus PPIs and introduced a Smooth Module to mitigate excessive sparsity (Fig. 5A). Specifically, we customized the loss function and retrained the amino acid weight parameters while keeping all other parameters within DeepGNHV fixed, ensuring accurate identification of the key residues. We took an interaction between the human T-cell surface glycoprotein CD4 and the HIV-1 Envelope glycoprotein gp120 as a case study to demonstrate the identified key residues. Structural visualization of this interaction was based on the protein complex structure (PDB ID: 3J70) [52]. Based on the GNNExplainer method, we identified 22 and 43 key residues for human and viral proteins within the solved structure (residues

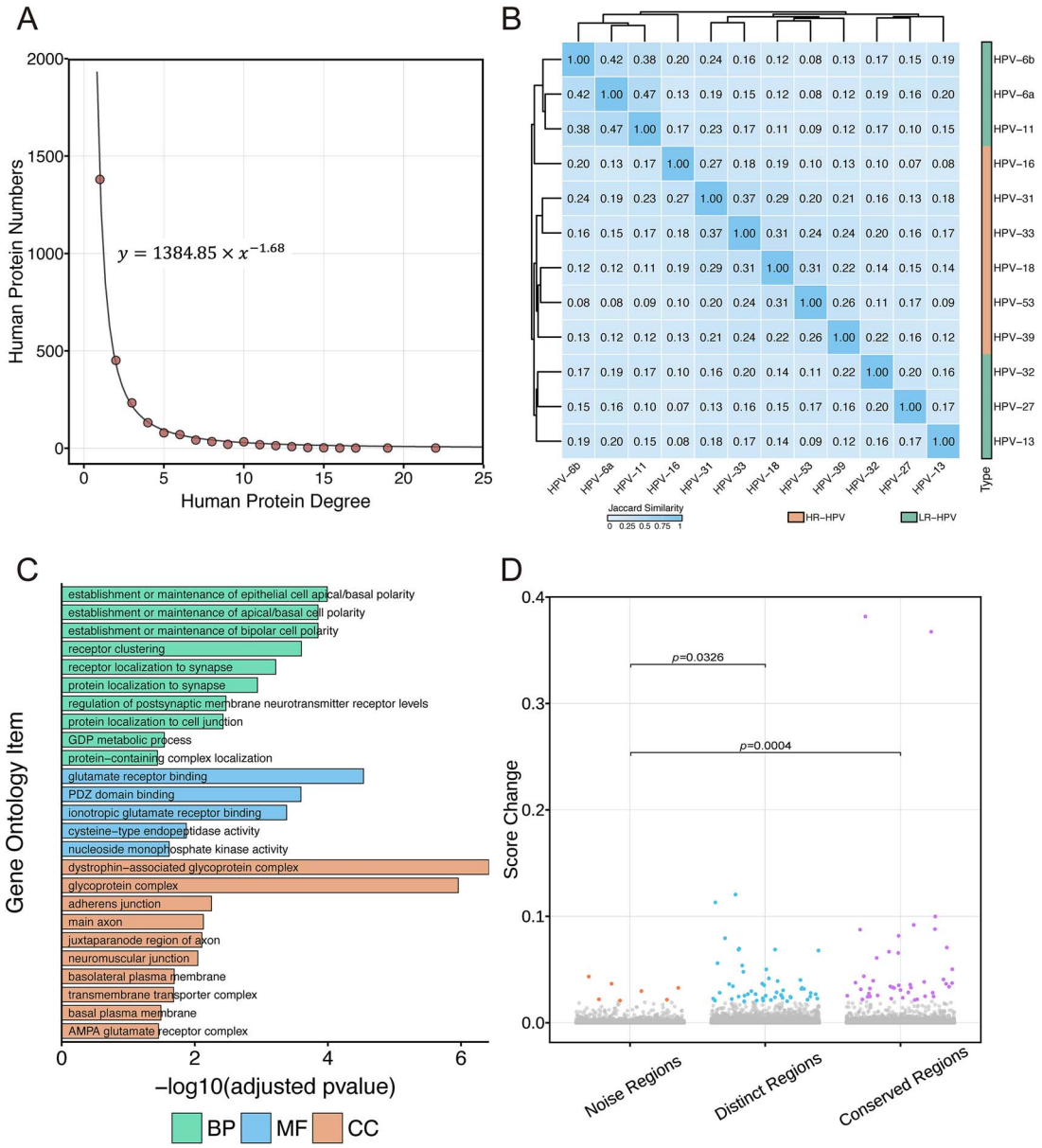

Figure 4. Large-scale human-HPV PPI predictions by DeepGNHV. (A) The distribution of human proteins predicted by DeepGNHV to interact with various HPV subtypes follows a power-law distribution. (B) The heatmap of the Jaccard similarity matrix from GO enrichment analysis, combining prediction results with known PPIs, shows a clear separation between HPV subtypes of different risk levels. (C) GO enrichment results for human proteins identified as specifically interacting with HR-HPV. (D) Analysis of the impact of single-amino acid perturbations on the HR-HPV E6 protein interactions in Conserved Regions, Distinct Regions, and Noise Regions.

with an important score >0.5 are defined as key residues), respectively (Fig. 5B). We further visualized the distribution of the important scores and their corresponding positions in protein sequences (Fig. 5C). Analysis of the 3J70 crystal structure revealed that human CD4 protein (181 resolved residues) and viral gp120 protein (470 resolved residues) include 18 and 22 experimentally verified interface residues, respectively. Briefly, the residues from two interacting proteins were assigned as interface residues if their Cα distance was less than or equal to 8.0Å. In particular, 10 of 22 human key residues identified by DeepGNHV belong to the interface residues. Although only one of the 43 key residues identified in the HIV gp120 protein directly overlaps with the 22 interface residues, nine additional key residues were found near the interaction interface (see region ① in Fig. 5B and C, the corresponding minimum Cα distance to human CD4 protein residues was <16Å), suggesting their potential role in mediating protein interactions.

Evaluation of informative model attributes

This section systematically explores and compares several key hyperparameter choices that play a crucial role in the model architecture. These include the distance threshold for constructing the adjacency matrix, amino acid embedding from different pretrained protein language models, the selection of GNN types, the application of a Siamese GAT strategy, pooling methods, and degree-preserving perturbation analysis of the adjacency matrix. First, we compared the performance of different distance thresholds for adjacency matrix construction. The results indicated that the model's predictive performance was slightly better when using an 8Å distance threshold compared

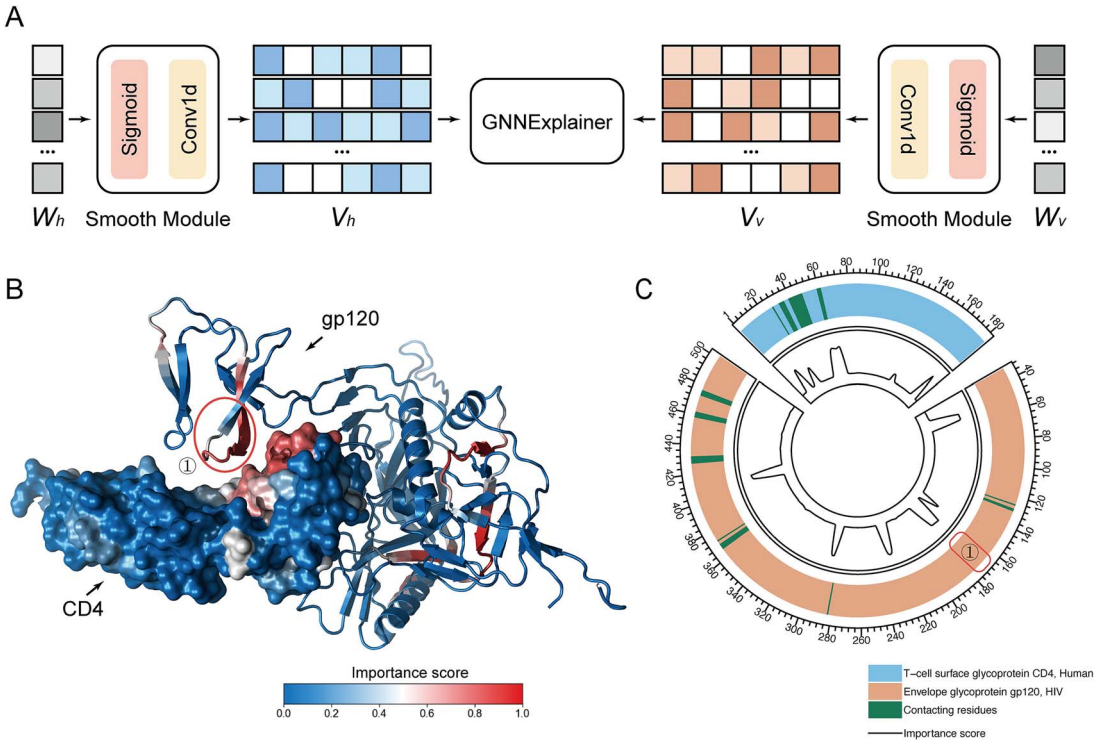

Figure 5. Analyzing the key residues for DeepGNHV interaction prediction. (A) The strategies for identifying key residues for DeepGNHV based on GNNExplainer and Smooth Module. (B) Surface (CD4) and cartoon (gp120) representations of 3J70 with importance scores in Pymol [53]. (C) A Circos plot illustrating the importance scores of individual residues in 3 J70 and their proximity to the binding interface.

to other options (Supplementary Table 6). We also compared the results of various pretrained models (SaProt [23], ESM2 [19], and TAPE [24]) as alternatives to ProtT5 [21] for encoding protein amino acid residues. Our findings revealed that their predictive effectiveness was slightly lower than that of the DeepGNHV model using ProtT5 [21] (Supplementary Table 7). In addition, we compared the predictive performance of five GNN architectures (GATConv, GCNConv, TransformerConv, GraphSAGE, and GIN) (Supplementary Table 8), and the results indicated that GATConv performed relatively better in the current prediction task. Furthermore, we compared Siamese and non-Siamese designs within the GAT framework and the results showed that Siamese architecture achieved slightly higher AUPRC (Supplementary Table 9). We also compared max and mean pooling and found that max pooling performed significantly better, likely because it captures localized sequence or structural features important for human–virus interactions (Supplementary Table 10). Finally, by conducting degree-preserving perturbation analysis on the adjacency matrix (keeping the degree constant while adjusting the amino acid position indices) to assess its impact on predictions, we found that the model's AUPRC value decreased by approximately 16 percentiles after perturbation (Supplementary Table 11). This underscores the importance of correctly constructing the adjacency matrix for interaction prediction.

Conclusion

In this work, we proposed DeepGNHV, a GNN framework that integrates a pretrained protein language model with structural information for predicting human virus PPIs. Our approach demonstrates superior performance compared to existing methods, particularly in predicting interactions involving novel viral proteins—a critical capability for emerging pathogen research. To explore the real application of the proposed DeepGNHV model, we carried out extensive PPI prediction and analysis for the human-HPV system, offering new insights into the human-HPV biology. Furthermore, the biological significance of DeepGNHV was exemplified by the identification of key residues in a human-HIV PPI prediction. Taken together, we hope the proposed end-to-end DeepGNHV model can effectively predict human-virus PPIs and thus provide new hints to decipher the host-virus relationships.

Nonetheless, the proposed DeepGNHV still suffers from several limitations. First, its performance still leaves room for further improvement. In particular, the model demonstrates limited generalization ability on certain viruses, probably due to the inherent data imbalance across different virus types. Second, our current framework relies on static structures predicted by AlphaFold2, which are inherently limited in capturing the dynamic nature of proteins. Moreover, post-translational modifications (PTMs) can profoundly influence PPIs in ways that static structural models cannot capture. However, public interaction databases such as BioGRID and IntAct rarely annotate the presence of specific PTM types in human virus PPIs, limiting the integration of PTM information into predictive frameworks. Finally, many viral proteins function as multimers, resulting in interaction stoichiometries that extend beyond simple one-to-one relationships. This highlights an important avenue for future methodological development: predictive models should simultaneously infer whether interactions occur or not and their multimeric states. Through the joint efforts of experimental scientists and computational biologists, we believe the above issues can be solved step by step, and the practical applicability of AI-driven human-virus PPI prediction methods will be continuously enhanced.

Key Points

- DeepGNHV is a deep learning framework based on structural information and pre-trained protein language models for human-virus PPI prediction.
- DeepGNHV can effectively predict human-virus PPIs, especially for novel viral proteins, enhancing the understanding of viral infection mechanisms.
- DeepGNHV provides new insights into key factors involved in the infection processes of HPV viruses with different risk levels.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: The authors declare that they have no conflicts of interest.

Funding

This work was supported by the National Natural Science Foundation of China (32270703). We appreciate the support of High-performance Computing Platform of China Agricultural University for providing computational resources.

Data availability

The source code of DeepGNHV and related data are publicly available on GitHub (https://github.com/).

References

1. Rouse BT, Sehrawat S. Immunity and immunopathology to viruses: What decides the outcome? *Nat Rev Immunol* 2010;**10**:514–26. https://doi.org/10.1038/nri2802.
2. Casanova J-L, Abel L. Mechanisms of viral inflammation and disease in humans. *Science* 2021;**374**:1080–6. https://doi.org/10.1126/science.abj7965.
3. Durmuş Tekir SD, Ülgen KÖ. Systems biology of pathogen-host interaction: Networks of protein-protein interaction within pathogens and pathogen-human interactions in the post-genomic era. *Biotechnol J* 2013;**8**:85–96. https://doi.org/10.1002/biot.201200110.
4. Baltoumas FA, Zafeiropoulou S, Karatzas E. *et al.* Biomolecule and bioentity interaction databases in systems biology: A comprehensive review. *Biomolecules* 2021;**11**:1245. https://doi.org/10.3390/biom11081245.
5. Nicod C, Banaei-Esfahani A, Collins BC. Elucidation of host–pathogen protein–protein interactions to uncover mechanisms of host cell rewiring. *Curr Opin Microbiol* 2017;**39**:7–15. https://doi.org/10.1016/j.mib.2017.07.005.
6. Shah PS, Link N, Jang GM. *et al.* Comparative Flavivirus-host protein interaction mapping reveals mechanisms of dengue and Zika virus pathogenesis. *Cell* 2018;**175**:1931–1945.e18. https://doi.org/10.1016/j.cell.2018.11.028.
7. Gordon DE, Jang GM, Bouhaddou M. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 2020;**583**:459–68. https://doi.org/10.1038/s41586-020-2286-9.
8. Matthews LR, Vaglio P, Reboul J. *et al.* Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "Interologs". *Genome Res* 2001;**11**:2120–6. https://doi.org/10.1101/gr.205301.
9. Yu H, Luscombe NM, Lu HX. *et al.* Annotation transfer between genomes: Protein–protein Interologs and protein–DNA Regulogs. *Genome Res* 2004;**14**:1107–18. https://doi.org/10.1101/gr.1774904.
10. Dyer MD, Murali TM, Sobral BW. Computational prediction of host-pathogen protein–protein interactions. *Bioinformatics* 2007;**23**:i159–66. https://doi.org/10.1093/bioinformatics/btm208.
11. Pierce BG, Wiehe K, Hwang H. *et al.* ZDOCK server: Interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics* 2014;**30**:1771–3. https://doi.org/10.1093/bioinformatics/btu097.
12. Yu J, Vavrusa M, Andreani J. *et al.* InterEvDock: A docking server to predict the structure of protein–protein interactions using evolutionary information. *Nucleic Acids Res* 2016;**44**:W542–9. https://doi.org/10.1093/nar/gkw340.
13. Van Zundert GCP, Rodrigues JPGLM, Trellet M. *et al.* The HADDOCK2.2 web server: User-friendly integrative Modeling of biomolecular complexes. *J Mol Biol* 2016;**428**:720–5. https://doi.org/10.1016/j.jmb.2015.09.014.
14. Yang X, Yang S, Li Q. *et al.* Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. *Comput Struct Biotechnol J* 2020;**18**:153–61. https://doi.org/10.1016/j.csbj.2019.12.005.
15. Yang X, Yang S, Lian X. *et al.* Transfer learning via multi-scale convolutional neural layers for human–virus protein–protein interaction prediction. *Bioinformatics* 2021;**37**:4771–8. https://doi.org/10.1093/bioinformatics/btab533.
16. Liu-Wei W, Kafkas Ş, Chen J. *et al.* DeepViral: Prediction of novel virus–host interactions from protein sequences and infectious disease phenotypes. *Bioinformatics* 2021;**37**:2722–9. https://doi.org/10.1093/bioinformatics/btab147.
17. Tsukiyama S, Kurata H. Cross-attention PHV: Prediction of human and virus protein-protein interactions using cross-attention–based neural networks. *Comput Struct Biotechnol J* 2022;**20**:5564–73. https://doi.org/10.1016/j.csbj.2022.10.012.
18. Zhang L, Wang S, Wang Y. *et al.* HBFormer: A single-stream framework based on hybrid attention mechanism for identification of human-virus protein–protein interactions. *Bioinformatics* 2024;**40**:btae724.
19. Lin Z, Akin H, Rao R. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. https://doi.org/10.1126/science.ade2574.
20. Rives A, Meier J, Sercu T. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA* 2021;**118**:e2016239118.
21. Elnaggar A, Heinzinger M, Dallago C. *et al.* ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2022;**44**:7112–27. https://doi.org/10.1109/TPAMI.2021.3095381.
22. Brandes N, Ofer D, Peleg Y. *et al.* ProteinBERT: A universal deep-learning model of protein sequence and function. *Bioinformatics* 2022;**38**:2102–10. https://doi.org/10.1093/bioinformatics/btac020.
23. Su J, Han C, Zhou Y. *et al.* SaProt: Protein language modeling with structure-aware vocabulary. *bioRxiv:101101/20231001560349* 2023.
24. Rao R, Bhattacharya N, Thomas N. *et al.* Evaluating protein transfer learning with TAPE. *Adv Neural Inf Proces Syst* 2019;**32**:9689–701.

25. Li G, Yao S, Fan L. ProSTAGE: Predicting effects of mutations on protein stability by using protein embeddings and graph convolutional networks. *J Chem Inf Model* 2024;**64**:340–7. https://doi.org/10.1021/acs.jcim.3c01697.
26. Zhang Q, Chen W, Qin M. *et al.* Integrating protein language models and automatic biofoundry for enhanced protein evolution. *Nat Commun* 2025;**16**:1553.
27. Jumper J, Evans R, Pritzel A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2.
28. Varadi M, Bertoni D, Magana P. *et al.* AlphaFold protein structure database in 2024: Providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* 2024;**52**:D368–75. https://doi.org/10.1093/nar/gkad1011.
29. Huang Y, Wuchty S, Zhou Y. *et al.* SGPPI: Structure-aware prediction of protein–protein interactions in rigorous conditions with graph convolutional network. *Brief Bioinform* 2023;**24**: bbad020.
30. Xie Z, Xu J. Deep graph learning of inter-protein contacts. *Bioinformatics* 2022;**38**:947–53. https://doi.org/10.1093/bioinformatics/btab761.
31. Gligorijević V, Renfrew PD, Kosciolek T. *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat Commun* 2021;**12**:3168.
32. Stark C. BioGRID: A general repository for interaction datasets. *Nucleic Acids Res* 2006;**34**:D535–9. https://doi.org/10.1093/nar/gkj109.
33. Orchard S, Ammari M, Aranda B. *et al.* The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 2014;**42**:D358–63.
34. Ammari MG, Gresham CR, McCarthy FM. *et al.* HPIDB 2.0: A curated database for host–pathogen interactions. *Database* 2016;**2016**:baw103. https://doi.org/10.1093/database/baw103.
35. Kumar R, Nanduri B. HPIDB - a unified resource for host-pathogen interactions. *BMC Bioinformatics* 2010;**11**:S16.
36. Licata L, Briganti L, Peluso D. *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 2012;**40**:D857–61. https://doi.org/10.1093/nar/gkr930.
37. Durmuş Tekir S, Çakır T, Ardıç E. *et al.* PHISTO: Pathogen–host interaction search tool. *Bioinformatics* 2013;**29**:1357–8. https://doi.org/10.1093/bioinformatics/btt137.
38. Guirimand T, Delmotte S, Navratil V. VirHostNet 2.0: Surfing on the web of virus/host molecular interactions data. *Nucleic Acids Res* 2015;**43**:D583–7. https://doi.org/10.1093/nar/gku1121.
39. Eid F-E, ElHefnawi M, Heath LS. DeNovo: Virus-host sequence-based protein–protein interaction prediction. *Bioinformatics* 2016;**32**:1144–50. https://doi.org/10.1093/bioinformatics/btv737.
40. Veličković P, Cucurull G, Casanova A. *et al.* Graph attention networks. *arXiv:171010903* 2018.
41. Fey M, Lenssen JE. Fast graph representation learning with PyTorch geometric. *arXiv:190302428* 2019.
42. Pedregosa F, Varoquaux G, Gramfort A. *et al.* Scikit-learn: Machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
43. Abramson J, Adler J, Dunger J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;**630**: 493–500. https://doi.org/10.1038/s41586-024-07487-w.
44. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 2017;**33**:2938–40. https://doi.org/10.1093/bioinformatics/btx364.
45. McBride AA. Human papillomaviruses: Diversity, infection and host interactions. *Nat Rev Microbiol* 2022;**20**:95–108. https://doi.org/10.1038/s41579-021-00617-5.
46. Wuchty S, Siwo G, Ferdig MT. Viral Organization of Human Proteins. *PLoS One* 2010;**5**:e11796. https://doi.org/10.1371/journal.pone.0011796.
47. Izawa I, Nishizawa M, Ohtakara K. *et al.* Densin-180 interacts with δ-catenin/neural plakophilin-related armadillo repeat protein at synapses. *J Biol Chem* 2002;**277**:5345–50. https://doi.org/10.1074/jbc.M110052200.
48. Ellenbroek SIJ, Iden S, Collard JG. Cell polarity proteins and cancer. *Semin Cancer Biol* 2012;**22**:208–15. https://doi.org/10.1016/j.semcancer.2012.02.012.
49. Letian T, Tianyu Z. Cellular receptor binding and entry of human papillomavirus. *Virol J* 2010;**7**:2.
50. Chi CN, Bach A, Engström Å. *et al.* Biophysical characterization of the complex between human papillomavirus E6 protein and synapse-associated protein 97. *J Biol Chem* 2011;**286**:3597–606. https://doi.org/10.1074/jbc.M110.190264.
51. Ying R, Bourgeois D, You J. *et al.* GNNExplainer: Generating explanations for graph neural networks. *arXiv:190303894* 2019.
52. Rasheed M, Bettadapura R, Bajaj C. Computational refinement and validation protocol for proteins with large variable regions applied to model HIV env spike in CD4 and 17b bound state. *Structure* 2015;**23**:1138–49. https://doi.org/10.1016/j.str.2015.03.026.
53. Schrödinger L, Delano W. *PyMOL*. 2020; Retrieved from http://www.pymol.org/pymol.

Briefings in Bioinformatics, 2025, **26(5)**, bbaf461
https://doi.org/10.1093/bib/bbaf461
Problem Solving Protocol