

AraVirusPPI: a protein language model-powered method for predicting *Arabidopsis thaliana*–virus protein–protein interactions

Shaoke Zhang ¹, Jiajun Li ¹, Mengdi Yuan ¹, Linyang Jiang ¹, Zhigang Li ², Chunyang Wang ³, Wenyu Shi ^{1*}, and Ziding Zhang ^{1*}

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

²Sanya Institute of Breeding and Multiplication/School of Tropical Agriculture and Forestry, Hainan University, No. 58 Renmin Avenue, Meilan District, Haikou 570228, China

³College of Life Sciences, Hebei Agricultural University, No. 2596 Lekai South Street, Lianchi District, Baoding 071000, China

*Corresponding author: 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. Email: shiw@cau.edu.cn; or 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. Email: zidingzhang@cau.edu.cn

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (<https://academic.oup.com/plphys/pages/General-Instructions>) is: Shaoke Zhang.

Abstract

Plant–virus protein–protein interactions (PPIs) play crucial roles in viral infection and host immune responses. However, large-scale experimental characterization of plant–virus PPIs is costly and labor-intensive, suggesting the need to develop computational methods. Recent advances in biological sequence language models are offering unprecedented opportunities to develop machine learning-based plant virus PPI predictors. In this study, we collected publicly available PPI data between *Arabidopsis thaliana* (*A. thaliana*) and viruses, establishing a robust dataset for method development. Based on this dataset, we developed AraVirusPPI, a machine learning framework for predicting *A. thaliana* virus PPIs. AraVirusPPI employs the protein language model ESM Cambrian (ESMC) to encode sequence features and combines these representations with Extreme Gradient Boosting (XGBoost) to build the prediction model. Furthermore, we systematically evaluated multiple protein and genomic language models combined with various classification algorithms, thereby justifying the optimal model architecture adopted in AraVirusPPI. We applied AraVirusPPI to predict PPIs in the *A. thaliana*–Cabbage leaf curl virus (*CabLCV*) system. Predicted viral target proteins were compared with experimental datasets from *A. thaliana*–Turnip mosaic virus (*TuMV*) and *A. thaliana*–*Pseudomonas syringae* (*Psy*) systems. Network topology and functional enrichment analyses showed that the predicted *CabLCV* targets in *A. thaliana* were more similar to *TuMV* than *Psy* targets. These results suggest potential shared and distinct host-targeting patterns among pathogens and provide computational evidence supporting the biological significance of the proposed AraVirusPPI predictor.

Introduction

Plant diseases caused by pathogens have long been recognized as major threats to agricultural productivity and global food security (Savary et al. 2019; Stukenbrock and Gurr 2023). Among these, plant viral diseases are particularly notable for their high transmission efficiency, wide epidemic ranges, and substantial challenges in control, leading to severe losses in crop yield and quality (Savary et al. 2019; He and Creasey Krainer 2020). Under natural conditions, plant viruses are typically transmitted by sap-feeding insects and other vectors, and their spread is further facilitated by crop trade and agricultural activities, resulting in expanding host ranges and geographic distributions (Ray and Casteel 2022; Wu et al. 2024; Liang et al. 2025). The viral life cycle involves genome replication, protein expression, virion

assembly, as well as local and systemic movement within the host. Each of these processes requires precise coordination within the host cellular environment (Rodriguez-Peña et al. 2021; Jiang and Zhou 2023). Owing to their small genomes and limited coding capacity, plant viruses often rely on their multifunctional proteins to manipulate key host biological processes (BPs) and complete their life cycles (Wu et al. 2022, 2024; Atabekova et al. 2023; Xavier and Whitfield 2023). In response, plants have evolved multilayered antiviral immune systems that restrict viral infection, replication, and spread (Dangl et al. 2013; Ceulemans et al. 2021; Jones et al. 2024; Yu et al. 2024). This dynamic interplay between viral manipulation and host defense highlights the central role of plant–virus protein–protein interactions (PPIs) in shaping infection outcomes (Xavier and Whitfield 2023; Wu et al. 2024).

Received: May 8, 2026. Accepted: October 5, 2026

© The Author(s) 2026. Published by Oxford University Press on behalf of American Society of Plant Biologists. All rights reserved. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

The development of various experimental strategies has greatly advanced the study of plant–virus PPIs. Techniques such as fluorescence-based approaches, coimmunoprecipitation, yeast two-hybrid assays (Y2H), etc., have enabled the identification of interactions between viral and plant proteins (Thiébeauld et al. 2009; Ibrahim et al. 2020; Tilsner and Kriechbaumer 2022; Rosas-Diaz et al. 2023). High-throughput approaches, including affinity purification coupled with mass spectrometry and proximity labeling (PL), have further facilitated the systematic characterization of plant–virus interaction networks (Martínez et al. 2016; Huang et al. 2023). With growing experimental efforts, an increasing number of plant–virus PPIs have been identified, providing valuable resources for understanding how viruses manipulate host processes. Among these, potyviruses are probably the most extensively studied plant virus groups, with *Turnip mosaic virus* (TuMV) yielding the most abundant plant–virus PPI data (Martínez et al. 2023). As a representative DNA virus, *Cabbage leaf curl virus* (CabLCV) is also an important system for studying plant–virus interactions (Xavier and Whitfield 2023; Wu et al. 2024). Nevertheless, experimental screening remains labor-intensive and time-consuming, limiting its large-scale applications. Therefore, integrating experimental data with computational prediction approaches has become increasingly essential. Experimentally validated interactions offer reliable datasets for model development, while computational methods enable rapid, proteome-wide prediction of potential interactions, thereby facilitating systematic analysis of plant–virus PPI networks.

In recent years, machine learning-powered PPI prediction methods based on sequence and structural features have been widely developed across various biological systems (Hu and Ohue 2024; Jiang et al. 2025). Machine learning models targeting plant–pathogen PPIs in bacteria, fungi, and oomycetes have been used to complement experimental screening (Yang et al. 2019; Lei et al. 2024). In contrast, PPI prediction methods specifically designed for plant–virus systems remain relatively underdeveloped, which precludes proteome-wide large-scale plant–virus PPI predictions. A major reason is that experimental plant–virus PPI data are scarce and scattered across multiple databases and literature sources, which limits the development of predictive models. Despite these challenges, machine learning approaches offer unique advantages in plant–virus PPI prediction. By systematically integrating protein sequence features, evolutionary information, and known interaction data, computational models can identify potential interaction candidates at the proteome scale, thereby greatly reducing the search space for experimental validation. Therefore, there is an urgent need to develop new machine learning predictors specifically tailored to plant–virus PPIs.

Inspired by advances in deep learning-powered natural language processing, biological sequence language models, such as protein language models (PLMs) and genomic language models (GLMs), have been rapidly developed in the past several years. These models are typically built on deep neural networks and pretrained in a self-supervised manner on large-scale sequence datasets. They can capture contextual dependencies, evolutionary constraints, and latent structural and functional signals from biological sequences, thereby generating informative representations. Pretrained PLMs, such as the ESM series (Rives et al. 2021; Lin et al. 2023; Candido et al. 2026) and the ProtTrans series (Elnaggar et al. 2022), have demonstrated advantages in diverse bioinformatics tasks, including PPI predictions (Jiang et al. 2025; Liu et al. 2025). In parallel, pretrained models for genome-scale sequence modeling have also undergone rapid development, including DNABERT (Ji et al. 2021; Zhou et al. 2023), GPN (Benegas et al. 2023), GROVER (Sanabria et al. 2024), megaDNA

(Shao and Yan 2024), and PlantCAD (Zhai et al. 2026, 2025). For plant–virus PPI prediction, these models offer distinct advantages. Plant viral genomes are compact and encode multifunctional proteins, yet experimental annotations remain limited. Pretrained models can transfer generalizable patterns learned from large-scale datasets and capture subtle sequence features underlying plant–virus interactions. Therefore, using PLMs/GLMs as sequence representations holds great promise for advancing the development of machine learning-based plant–virus PPI predictors.

Here, we focus on PPI prediction between *A. thaliana* and its associated plant viruses. We integrate experimentally validated plant–virus PPIs to construct a standardized dataset and develop an optimal machine learning-based computational framework that employs pretrained PLMs/GLMs as feature representations (Fig. 1a). The resulting computational framework, termed AraVirusPPI, has achieved reasonable performance in rigorous benchmarking experiments. Thus, it can accelerate the identification of plant–virus interaction candidates, offering new insights into plant–virus biology.

Results

Performance comparison of computational frameworks with PLMs

In this work, we aim to develop an *A. thaliana* virus PPI prediction method by screening the optimal combination among 4 pretrained PLM encodings (ESM-1b, ESM-2, ESMC, and ProtT5) and 4 machine learning classifiers, including Extreme Gradient Boosting (XGBoost), Random Forest (RF), support vector machine (SVM), and convolutional neural network (CNN) (Fig. 1; Table S2). To this end, each method was implemented as an ensemble model, in which 5 predictive models were trained, and their predictions on the independent test set were averaged to obtain the final prediction. The area under the precision-recall curve (AUPRC) was used as the primary performance metric, with other metrics serving as supplementary evaluation measures. More details on dataset preparation, feature engineering, model training, and performance assessment are provided in the Materials and Methods section.

In general, substantial performance differences were observed among the combinations of PLMs and classifiers (Fig. 1). Among all combinations, XGBoost- and CNN-based models consistently outperformed RF and SVM. Under the XGBoost framework, ESMC achieved the best performance, with an AUPRC of 0.662 on the independent test set, followed by ProtT5 (0.651), whereas ESM-1b and ESM-2 showed relatively lower performance (0.613 and 0.612) (Fig. 1b). In contrast, RF and SVM models yielded consistently lower AUPRC values when using the same encoding methods (Fig. 1c, 1d). Although CNN models achieved comparable AUPRC values in some cases (0.594 to 0.608), they did not surpass the best XGBoost-based combinations (Fig. 1e). Based on the above comparisons, the combination of XGBoost and ESMC was identified as the best-performing framework. Therefore, we selected this computational framework (ESMC_XGBoost) as the final prediction model and designated it as AraVirusPPI.

Performance comparison of computational frameworks with GLMs

Since protein sequence information is also encoded in the corresponding DNA sequence, DNA-derived sequence features of protein pairs have been used to predict PPIs (Zhou et al. 2012). Here, we also

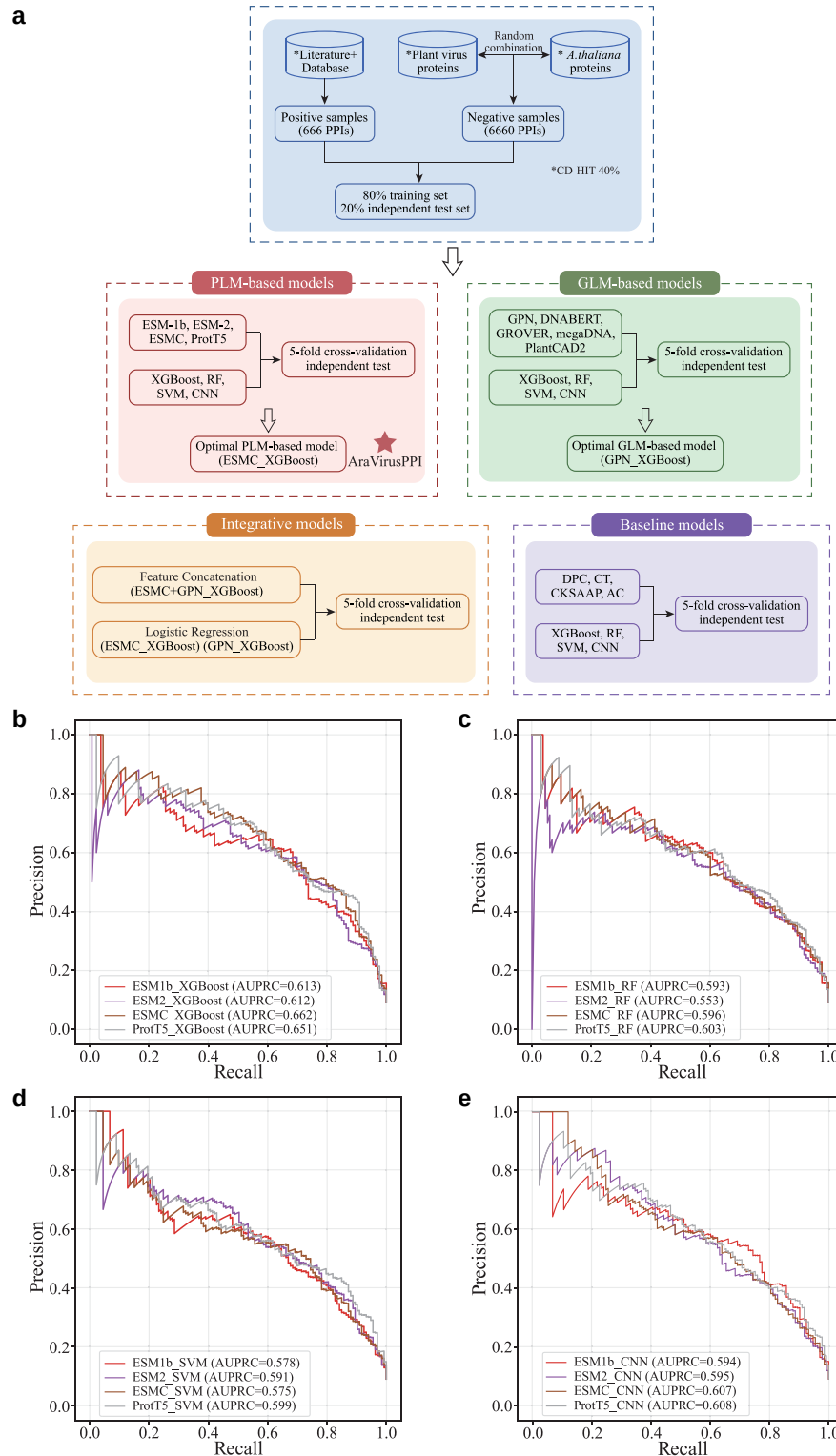


Figure 1 Workflow for model construction and performance comparison of PLM-based representations across 4 classifiers. a, Workflow of data collection, model construction, and performance comparison in this study. b–e, PR curves of models trained with different PLM embeddings (ESM-1b, ESM-2, ESMC, and ProtT5) using XGBoost (b), RF (c), SVM (d), and CNN (e), respectively.

investigated the performance of computational frameworks with GLMs (GPN, DNABERT, DNABERT-2, GROVER, megaDNA, and PlantCAD2) in predicting *A. thaliana*-virus PPIs (Fig. 2; Table S2). Among all combinations, XGBoost-based models again achieved the

best overall performance (Fig. 2). Within this framework, GPN achieved the highest AUPRC (0.597), comparable to PlantCAD2 (0.595) and exceeding the other 4 encodings by more than 5% (Fig. 2a; Table S2). Consistent trends were observed for RF, SVM, and

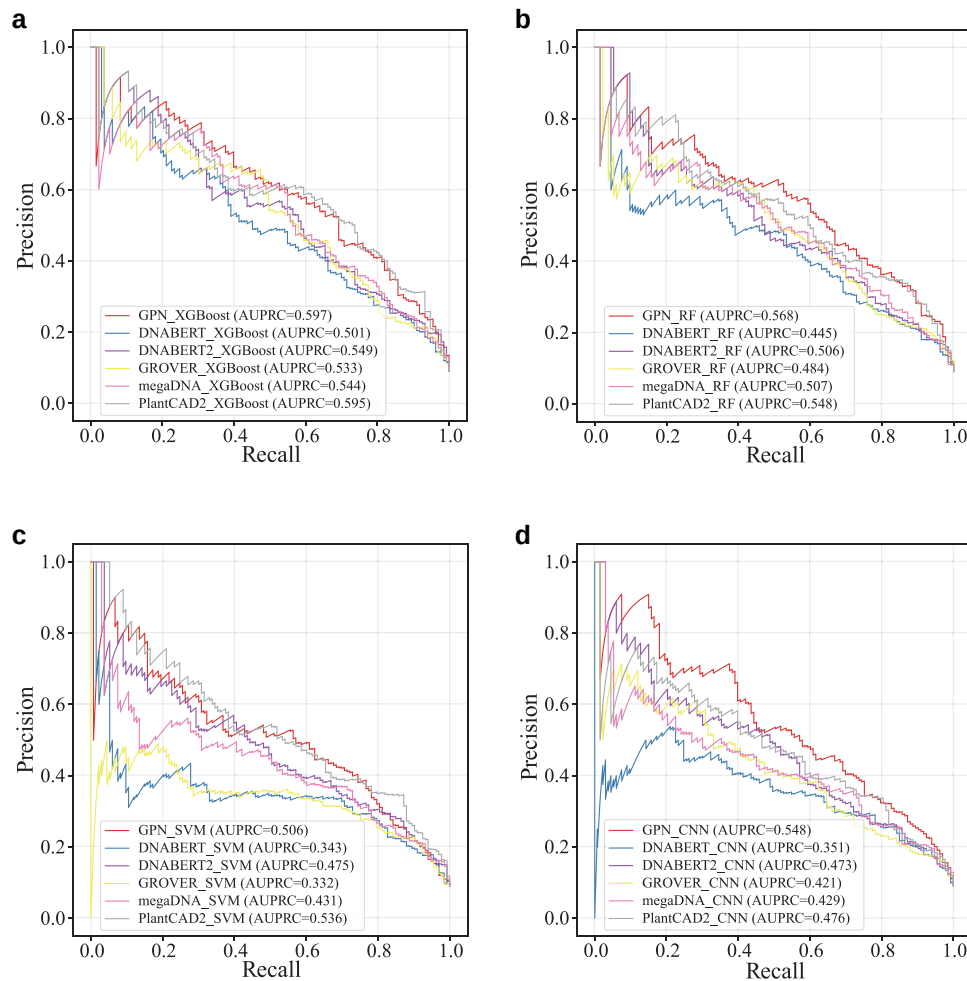


Figure 2 Performance comparison of GLM-based representations across 4 classifiers. a–d, PR curves of models trained with different GLM embeddings (GPN, DNABERT, DNABERT2, GROVER, megaDNA and PlantCAD2) using XGBoost (a), RF (b), SVM (c), and CNN (d), respectively.

CNN classifiers, where GPN and PlantCAD2 also showed superior performance (Fig. 2b–d). The favorable performance of plant-specific genomic models is likely attributable to their pretraining on the *A. thaliana* genome, which may enhance their ability to capture species-specific genomic features relevant to *A. thaliana*–virus interactions. Despite this competitive performance, GLM-based models generally underperformed PLM-based models under the same evaluation framework. In addition, GLM-based models exhibited greater performance variability across classifiers, with AUPRC values ranging from 0.332 to 0.597, and some combinations (eg, DNABERT_CNN) showed limited ability to identify positive interaction pairs (Table S2). Overall, these results suggest that while GLMs capture informative sequence features, protein sequence representations remain more effective for *A. thaliana*–virus protein interaction prediction.

Performance comparison of integrative models of PLMs and GLMs

Given the potential complementarity between PLMs and GLMs, we further investigated whether integrating these 2 representations via 2 different strategies could improve prediction performance. In the first strategy, the optimal protein features derived from ESMC and the optimal genomic features derived from GPN were concatenated to form a

3,328-dimensional representation, which was subsequently used for classification with XGBoost. This model (ESMC+GPN_XGBoost) achieved an AUPRC of 0.658 on the test set (Fig. 3a, 3b; Table S2). Compared with the optimal PLM-based model (ie, AraVirusPPI), AUPRC showed no substantial increase. In the second strategy, a model-level ensemble approach was adopted. Prediction scores generated by the protein-based model (ESMC_XGBoost) and the genomic model (GPN_XGBoost) were used as input features for a logistic regression (LR) classifier. This ensemble model achieved an AUPRC of 0.647 on the test set (Fig. 3a, 3b; Table S2). Overall, although integrative models improved certain performance metrics (eg, recall), they did not substantially outperform AraVirusPPI. In addition, these approaches introduced increased computational cost and model complexity.

Performance comparison of traditional protein sequence encoding methods

To further benchmark model performance, we compared AraVirusPPI with multiple baseline models based on traditional protein sequence encoding schemes [ie, dipeptide composition (DPC), conjoint triad (CT), composition of k-spaced amino acid pairs (CKSAAP), and auto-correlation (AC)] combined with XGBoost, RF, SVM, and CNN classifiers. Among these, the CKSAAP_XGBoost model

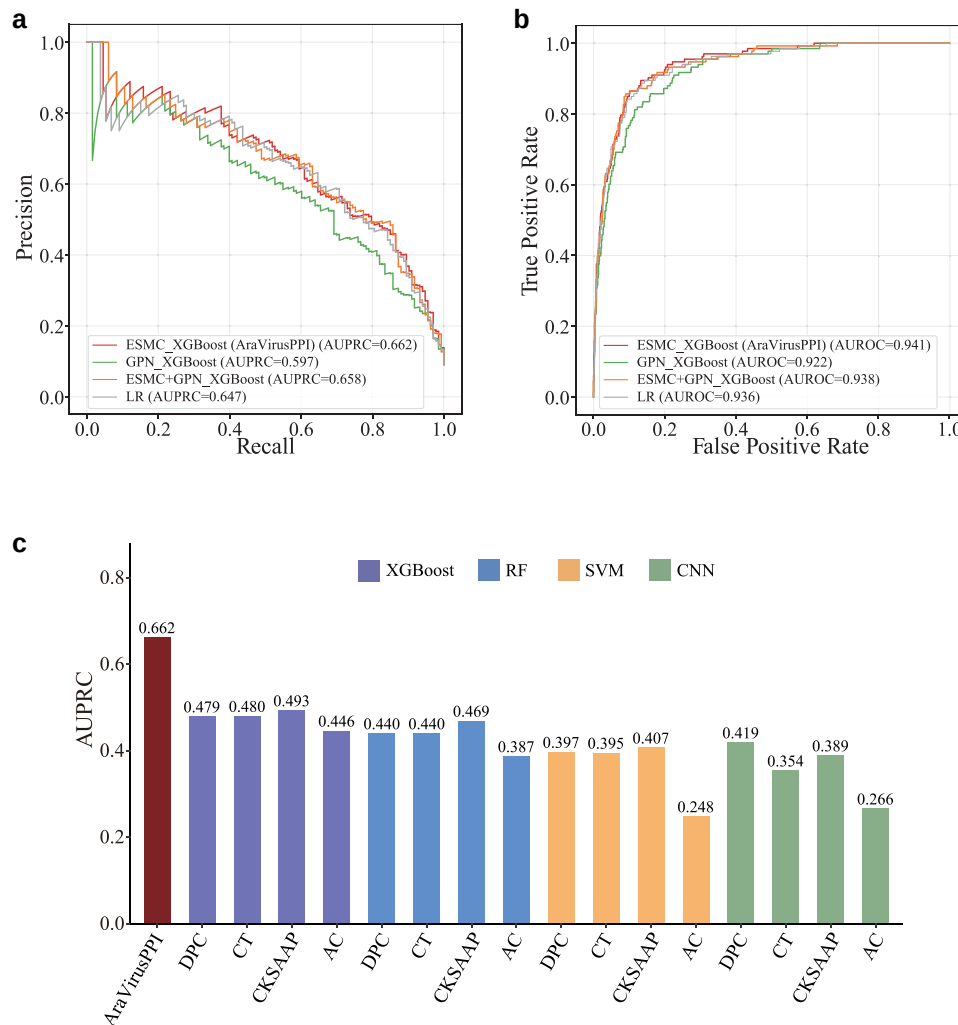


Figure 3 Performance of integrative models and traditional sequence encodings. a, PR curves of two integrative models, including ESMC+GPN_XGBoost (ie, the XGBoost model trained with the concatenated ESMC and GPN features) and the LR model trained using ESMC_XGBoost and GPN_XGBoost predictions. For comparison, PR curves of ESMC_XGBoost and GPN_XGBoost were also plotted. b, ROC curves for the corresponding models shown in (a). c, AUPRC comparison of AraVirusPPI against 4 machine learning methods using traditional protein sequence encoding schemes (ie, DPC, CT, CKSAAP, and AC).

achieved the best performance, with an AUPRC of 0.493 and an area under the receiver operating characteristic curve (AUROC) of 0.901 (Table S2). However, this performance remained substantially lower than that of AraVirusPPI, which achieved an AUPRC of 0.662 (Fig. 3c). Across different classifiers, traditional encoding methods generally performed better under the XGBoost framework, but their overall performance remained limited (Table S2). These results demonstrate that pretrained PLM-based representations provide a clear advantage over traditional sequence encoding schemes for *A. thaliana*–virus PPI prediction. Taken together, these benchmarking experiments support the selection of the AraVirusPPI framework as the optimal computational framework for the current dataset.

Performance comparison of residue-level representations and fusion strategies

We additionally evaluated whether residue-level representations and interaction-aware feature fusion strategies could provide performance improvements over the AraVirusPPI framework. A multi-

head cross-attention model was developed based on residue-level ESMC representations, and 3 protein pair feature fusion strategies, including concatenation, Hadamard product, and element-wise difference, were compared. Compared with pooled representations, residue-level embeddings retain local sequence information and may better capture important interaction-related residues (Elnaggar et al. 2022; Candido et al. 2026). Within the residue-level models, element-wise difference and Hadamard product achieved better performance than concatenation but did not outperform the original AraVirusPPI model (Figure S1a). This may be related to the limited size of the current training dataset, where the advantages of more complex model architectures may not be fully exploited. We further replaced the concatenation strategy in the ESMC_XGBoost framework with Hadamard product and element-wise difference, but neither strategy outperformed AraVirusPPI (Figure S1b). These results suggest that, although more complex residue-level representations and interaction-aware fusion strategies were explored, the simpler ESMC_XGBoost framework remained suitable for the current dataset.

Table 1 Overview of the compiled *A. thaliana*–virus PPI data.

Category	Virus name	Abbreviation	Genome	No. of positive PPIs	No. of negative PPIs
RNA	Turnip mosaic virus (strain Japanese)	TuMV	ssRNA(+)	319	3,190
	Tobacco etch virus	TEV	ssRNA(+)	214	2,140
	Cucumber mosaic virus (strain Q/FNY)	CMV	ssRNA(+)	14	140
	Potato virus X	PVX	ssRNA(+)	6	60
	Tobacco mosaic virus (strain vulgare/U1)	TMV	ssRNA(+)	4	40
	Beet western yellows virus	BWYV	ssRNA(+)	2	20
	Alfalfa mosaic virus (strain 425/isolate Leiden)	AMV	ssRNA(+)	1	10
	Alternanthera mosaic virus (Potexvirus alternantherae)	AltMV	ssRNA(+)	1	10
	Barley yellow dwarf virus (isolate PAV)	BYDV	ssRNA(+)	1	10
	Groundnut rosette virus (strain MC1)	GRV	ssRNA(+)	1	10
	Potato mop-top virus (isolate Potato/Sweden/Sw)	PMTV	ssRNA(+)	1	10
	Rice stripe virus (isolate T)	RSV	ssRNA(–)	1	10
	Turnip crinkle virus	TCV	ssRNA(+)	1	10
	Cabbage leaf curl virus (isolate Jamaica)	CabLCV	ssDNA(+/–)	79	790
	Cauliflower mosaic virus (strain Strasbourg/S-Japan)	CaMV	dsDNA-RT	8	80
	Tomato golden mosaic virus (strain Yellow vein)	TGMV	ssDNA(+/–)	6	60
	Beet curly top virus (strain California/Logan)	BCTV	ssDNA(+/–)	3	30
DNA	Mungbean yellow mosaic India virus	MYMIV	ssDNA(+/–)	2	20
	Tomato yellow leaf curl Sardinia virus	TYLCSaV	ssDNA(+/–)	1	10
	Tomato yellow leaf curl virus	TYLCV	ssDNA(+/–)	1	10
Total				666	6,660

Evaluation of model generalization across unseen virus species

To evaluate generalization across unseen viruses, we conducted 2 more rigorous computational experiments to further test the performance of the AraVirusPPI framework. Due to the limited number of virus species and PPIs, virus species with fewer than 5 experimentally validated PPIs were used as a test set (Table 1), while the remaining viruses were used for training. The model achieved an AUPRC of 0.378 in this evaluation (Figure S2a). We further evaluated AraVirusPPI on an additional independent dataset collected from published literature and not included in the original training or test sets (Table S4), achieving an AUPRC of 0.444 (Figure S2b). Overall, performance in both computational experiments was lower than the original test performance (AUPRC = 0.662), which is likely due to the challenges of cross-virus prediction, including evolutionary differences among viruses and insufficient interaction data used in training. Nevertheless, both evaluations remained substantially higher than random expectation (AUPRC = 0.091), suggesting that AraVirusPPI retains predictive ability on the evaluated unseen-virus datasets.

Comparative analysis of *A. thaliana*–virus and *A. thaliana*–bacterium PPI networks

To further explore the biological significance of the predicted interactions, we performed a comparative analysis of 3 *A. thaliana*–pathogen PPI datasets: 2 viruses (ie, CabLCV and TuMV), and 1 bacterium (ie, *Pseudomonas syringae*, Psy), which is one of the best-studied plant bacterial pathogens and serves as a model for understanding plant–pathogen interactions (Chen et al. 2022). CabLCV interactions were derived from high-confidence predictions generated by the AraVirusPPI model in this study (500 interaction pairs), whereas TuMV interactions were obtained from experimentally validated *A. thaliana*–virus datasets. Experimentally validated *A. thaliana*–Psy interactions were also included for cross-pathogen comparison. More details about these datasets are available in the Materials and Methods section. We first compared the host target protein sets of the 3 pathogens. Notably, no *A. thaliana* protein was targeted by all 3 pathogens, indicating

substantial differences in host target protein selection among distinct pathogen types. Comparatively, the 2 viruses (CabLCV and TuMV) shared considerably more host target proteins than either virus shared with Psy. This difference was further supported by Jaccard similarity, which measures the relative overlap between 2 target protein sets. The higher Jaccard similarity between the 2 viral target sets suggests potential similarities in their inferred host targeting patterns (Fig. 4a).

We next examined the network properties of pathogen target proteins by constructing an *A. thaliana* host PPI network comprising 55,517 interactions among 11,411 proteins. The pathogen target proteins were then mapped onto this network to examine their potential organizational patterns and convergent targeting behaviors. Using random sampling simulations, we systematically evaluated convergent targeting among pathogen effectors (Fig. 4b). For both viruses and the bacterium, the observed numbers of unique host-targeted proteins were significantly lower than random expectations, indicating that multiple proteins from the same pathogen tend to target overlapping host proteins, exhibiting pronounced intraspecies convergence (Fig. 4c). This phenomenon has been previously reported for bacterial effectors (Weßling et al. 2014), and our results demonstrate that plant viruses display similar convergent targeting patterns. The number of shared host target proteins between different pathogens was consistently higher than expected by chance, suggesting that distinct pathogens tend to target common key nodes in the host interaction network. Notably, the predicted CabLCV and TuMV target proteins overlapped significantly more with each other than either did with Psy, further suggesting distinct targeting patterns between viral and bacterial pathogens (Fig. 4d).

Based on the host PPI network, we further analyzed the topological positions and distance distributions of pathogen target proteins. The average shortest-path distances among target proteins were significantly shorter than those between targeted and nontargeted proteins for all 3 pathogens (Fig. 5a, d), indicating that pathogen-targeted proteins are not randomly distributed but tend to form locally clustered modules. This clustering suggests coordinated perturbation of specific functional modules involved in host immunity and regulation. Moreover, targeted proteins showed higher betweenness centrality than nontargeted proteins for CabLCV, TuMV, and Psy, whereas degree differences varied among pathogens (Fig. 5b, c; Table S6).

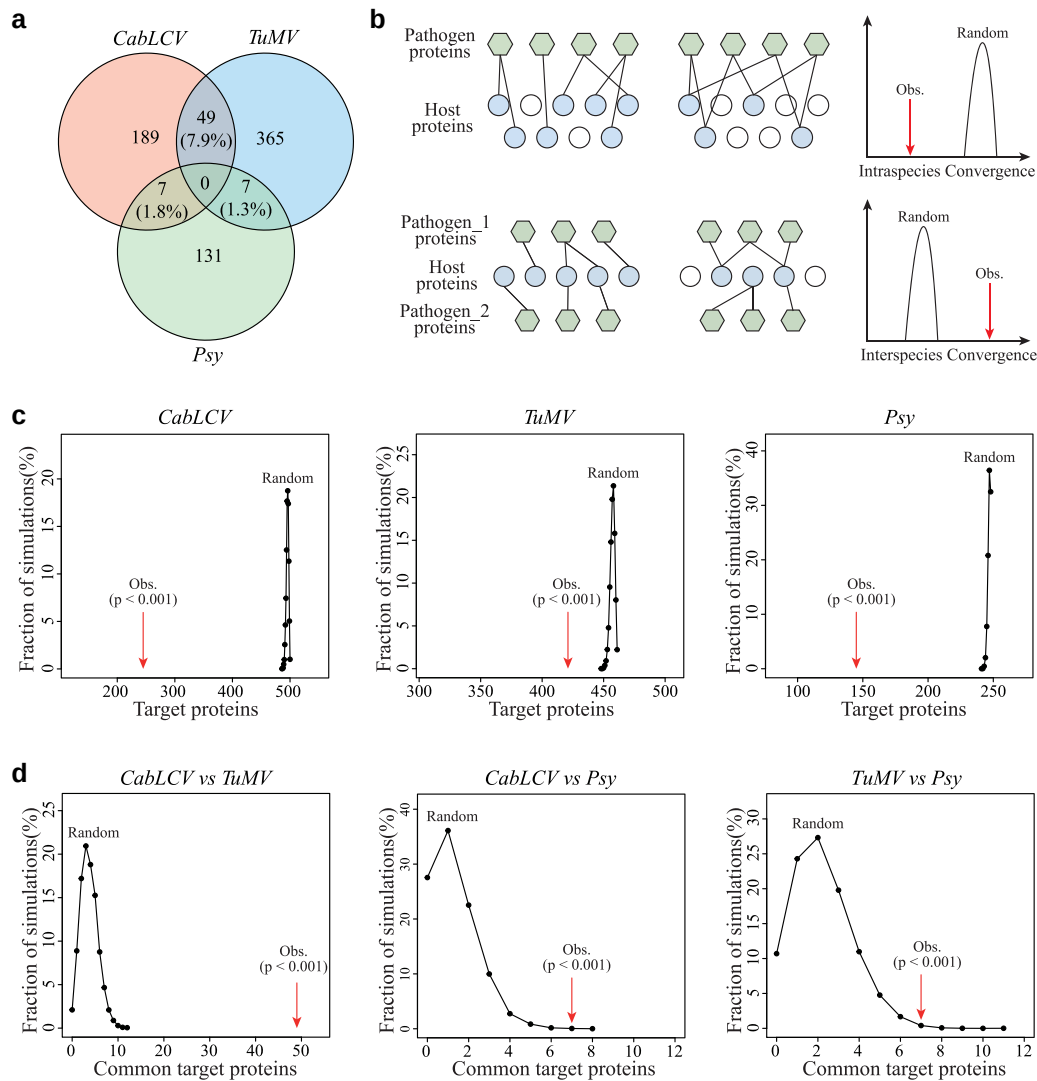


Figure 4 Network-based analysis of *A. thaliana* proteins targeted by different pathogens. **a**, Venn diagram showing the overlap of host proteins targeted by *CabLCV*, *TuMV*, and *Psy*. Percentages represent pairwise Jaccard similarity between target protein sets. **b**, Network rewiring analysis between pathogen proteins (hexagons) and host proteins (circles). Lines indicate interactions between pathogen proteins and host proteins. This analysis was performed to assess the likelihood of random and intraspecies convergence (top) and interspecies convergence (bottom) in host–pathogen interactions. **c**, Distribution of simulated numbers (lines) and observed numbers (arrows) of shared *A. thaliana* proteins targeted by *CabLCV*, *TuMV*, and *Psy* proteins. **d**, Distribution of simulated numbers (lines) and observed numbers (arrows) of shared *A. thaliana* proteins targeted by *CabLCV*, *TuMV*, and *Psy* in pairwise comparisons.

Repeated random sampling analyses showed that betweenness centrality differences remained significant in >99% of iterations for all 3 pathogens, while degree differences were significant in 52.1%, 51.8%, and 100% of iterations for *CabLCV*, *TuMV*, and *Psy*, respectively (Table S6). However, clear differences were observed among pathogens. *Psy*-targeted proteins exhibited significantly higher degree and betweenness centrality than viral targets (Fig. 5e, f), suggesting that bacterial pathogens tend to directly interfere with core regulatory hubs, whereas computationally inferred viral targets may be more likely located in secondary functional modules. Overall, these results provide computational evidence for distinct targeting patterns between viral and bacterial pathogens within the *A. thaliana* protein interaction network. These support the potential applicability of AraVirusPPI for large-scale plant–virus PPI analysis, although the predicted *CabLCV* interactions require further experimental validation to confirm their biological relevance.

Functional enrichment of host proteins targeted by viruses and bacteria

We further investigated the functional roles of host proteins targeted by *CabLCV*, *TuMV*, and *Psy* using Gene Ontology (GO) enrichment analysis (Fig. 6). A set of significantly overlapping functional terms was identified across the 3 pathogens. At the BP level, all 3 target protein sets were enriched in terms related to “processes involved in symbiotic interactions” and “protein folding”. At the molecular function (MF) level, common enrichment was detected for “heat shock protein binding” and “copper ion binding.” These results indicate that both viral and bacterial pathogens preferentially target host proteins involved in proteostasis maintenance and stress responses. Notably, the consistent enrichment of protein folding and molecular chaperone-related terms suggests that host chaperone systems may be associated with pathogen infection

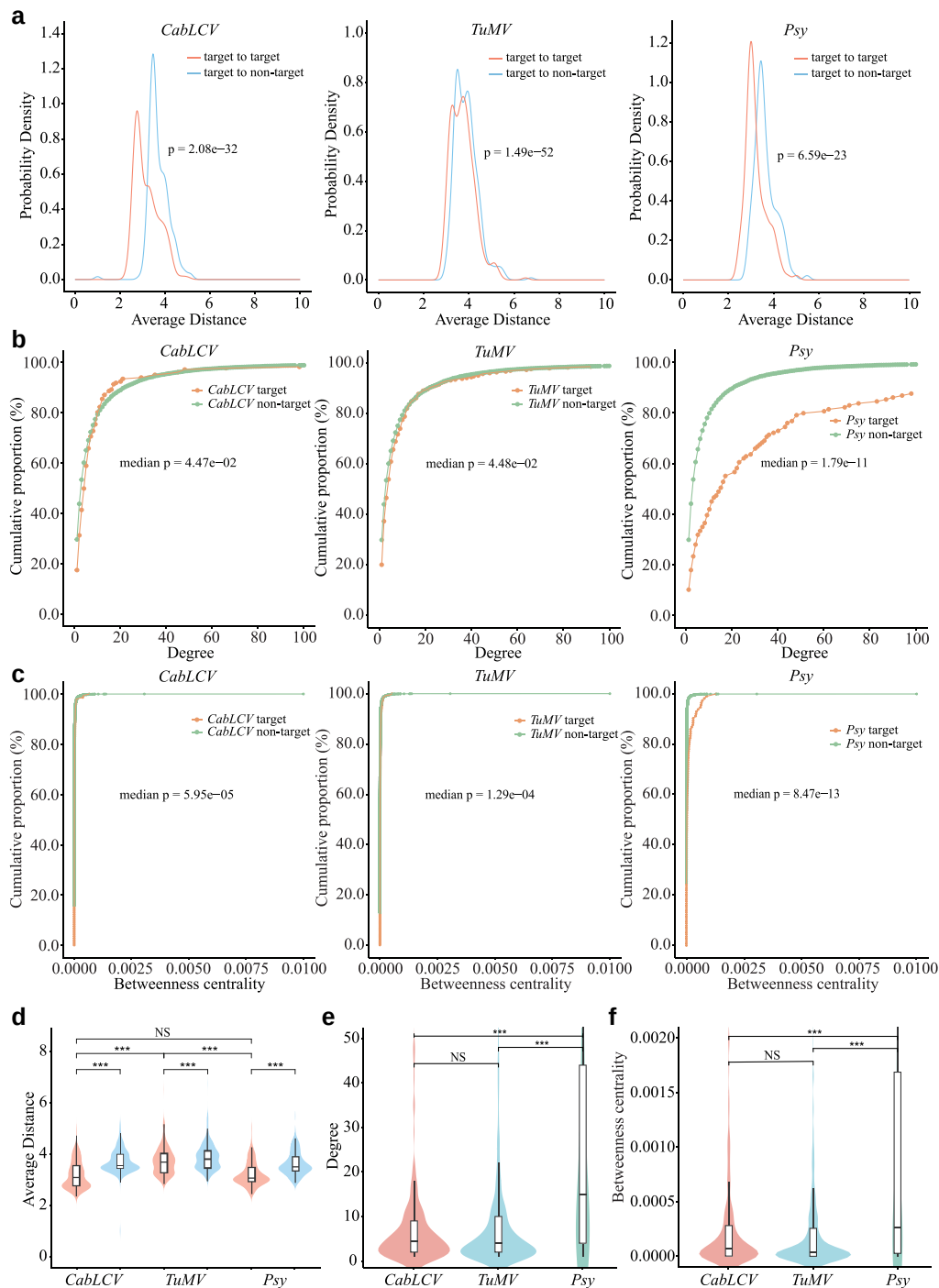


Figure 5 Topological characteristics of *A. thaliana* proteins targeted by plant pathogens in PPI networks. a, Average shortest path distances among target *A. thaliana* proteins and between target and nontarget *A. thaliana* proteins for *CabLCV*, *TuMV*, and *Psy*. b, Cumulative proportion distribution of degree ratios for target and nontarget *A. thaliana* proteins. c, Cumulative proportion distribution of betweenness centrality ratios for target and nontarget *A. thaliana* proteins. d, Boxplots corresponding to panel (a), showing the average shortest path distances among target *A. thaliana* proteins and between target and nontarget *A. thaliana* proteins. e, Boxplots showing the degree of *A. thaliana* proteins targeted by *CabLCV*, *TuMV*, and *Psy*. f, Boxplots showing the betweenness centrality of *A. thaliana* proteins targeted by *CabLCV*, *TuMV*, and *Psy*. Statistical significance was evaluated using the Wilcoxon rank-sum (two-sided) test unless otherwise specified. Intraspecies distance comparisons in panels (a) and (d) were assessed using a paired Wilcoxon signed-rank (two-sided) test. Exact P-values are shown in (a–c), while (d–f) use significance levels (*** $P < 0.001$; NS, not significant).

and cellular stress responses (Zhang and Wang 2012; Herath et al. 2020). Moreover, the shared enrichment for copper ion binding across the 3 target protein sets suggests that metal ion-associated

processes play important roles during both viral and bacterial infections. This observation is consistent with previous reports highlighting the involvement of metal ions in animal immunity

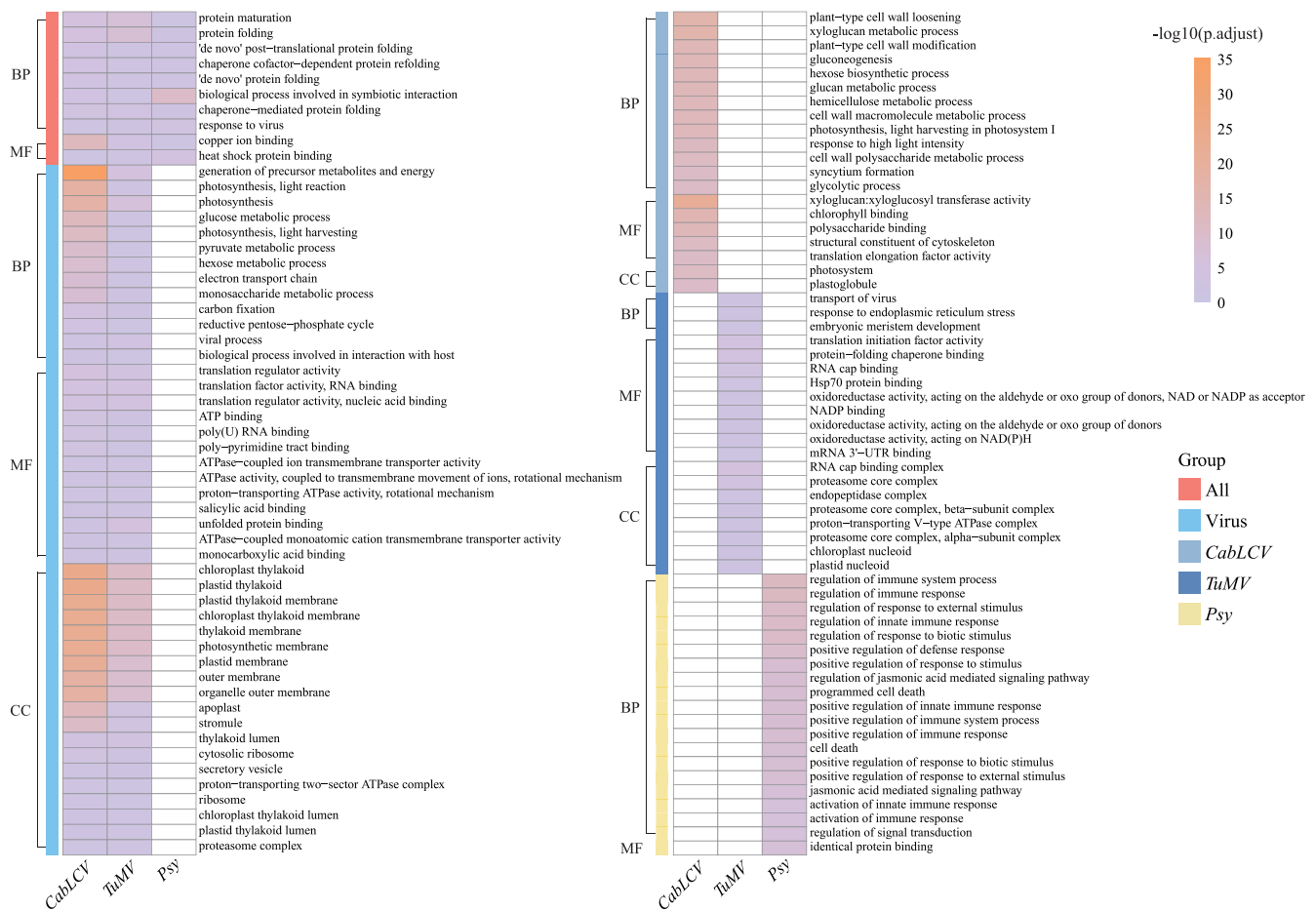


Figure 6 Heatmap of GO functional enrichment analysis for *A. thaliana* proteins targeted by *CabLCV*, *TuMV*, and *Psy*. The left panel shows GO terms for *A. thaliana* proteins targeted by all 3 pathogens and by *CabLCV* and *TuMV* only. The right panel displays the top 20 significantly enriched GO terms associated with *A. thaliana* proteins uniquely targeted by each pathogen (*CabLCV*, *TuMV*, and *Psy*). Color intensity represents $-\log_{10}(p.adjust)$, with higher values indicating stronger enrichment.

(Lu et al. 2024), suggesting that copper-regulatory pathways may represent a common functional module in host responses to diverse pathogens.

In addition to these shared features, viral and bacterial target proteins exhibited distinct functional patterns. At the BP level, viral target proteins were significantly enriched in terms related to “viral processes”, “photosynthesis”, “electron transport chain”, “photosystem II repair”, and “pigment metabolism”, indicating a close association between viral infection and host chloroplast function and photosynthetic activity (Li et al. 2021). By contrast, bacterial target proteins were more strongly enriched in processes related to immune responses and stress responses, including “regulation of immune response”, “regulation of response to external stimulus”, and “positive regulation of defense response”. These enrichment differences suggest distinct host response pathways targeted by viruses and bacteria. Consistent with this, although both viral and bacterial infections can induce reactive oxygen species (ROS) production, the enrichment patterns may indicate differences in their sources and regulatory mechanisms. Bacterial target proteins are primarily associated with immune regulatory processes, suggesting that ROS production may be linked to pattern-triggered immunity (PTI)-related signaling (Dangl et al. 2013; Jones et al. 2024), whereas viral targets are predominantly enriched in chloroplast-related and photosynthetic electron

transport processes, indicating that chloroplast-derived ROS may play a more prominent role during viral infection.

Notably, the DNA virus *CabLCV* and the RNA virus *TuMV* also exhibited distinct functional enrichment patterns, reflecting fundamental differences in their replication and gene expression strategies. The predicted *CabLCV*-targeted proteins were mainly enriched in cell wall organization and metabolic processes, including carbohydrate metabolism and photosynthesis, suggesting a potential role in remodeling the host cellular environment and altering resource allocation. This interpretation is consistent with previous studies showing that begomovirus proteins interact with host factors involved in intracellular trafficking, nuclear transport, transcriptional regulation, and antiviral signaling (Carvalho and Lazarowitz 2004; Gouveia-Mageste et al. 2021). In contrast, *TuMV*-targeted proteins were primarily enriched in functions associated with RNA metabolism, translation-related processes, protein homeostasis, and related cellular processes. These functions are involved in multiple stages of *TuMV* infection, including viral replication, intracellular trafficking, and host cellular regulation (Li et al. 2016; Martínez et al. 2023). Collectively, these differences suggest that DNA and RNA viruses may adopt distinct strategies to manipulate host cellular functions. Based on the predicted host targets, the DNA virus examined here appears to target broader cellular processes associated with metabolism and cellular remodeling, whereas the experimentally

validated *TuMV* targets are more prominently associated with host factors involved in RNA metabolism and translation-related processes.

Discussion

Plant viruses possess small genomes, and they typically employ a limited number of multifunctional viral proteins to precisely manipulate key host molecular networks, thereby facilitating infection and proliferation (Wu et al. 2022; Atabekova et al. 2023). Compared with other host–pathogen interaction systems, studies on plant–virus protein interactions lag far behind. Most existing studies have focused on individual viruses or single host factors, while a comprehensive, system-level understanding of plant–virus PPI networks remains lacking. In this context, we systematically compiled publicly available *A. thaliana*–virus protein interaction data reported before 2024. This effort partially alleviated the limitations caused by data fragmentation. We further employed experimentally validated interactions as a reliable reference set and introduced a computational framework to develop and evaluate plant host–virus PPI prediction models. The established framework not only enables the identification of potential host–virus interactions that have not yet been experimentally validated, but also provides a feasible complementary approach for expanding the coverage of plant–virus interactomes under limited experimental conditions.

Regarding dataset preparation, selecting negative samples is a critical issue in developing machine learning-based PPI predictors. Since experimentally validated noninteracting protein pairs are difficult to define, random sampling from large protein pair spaces remains a commonly used strategy in PPI prediction (Lian et al. 2021). However, this strategy does not explicitly incorporate subcellular localization information. On the one hand, negative pairs without annotated subcellular colocalization are expected to have a lower probability of physical interaction. On the other hand, such negative pairs may represent “easy negatives”, which artificially boosts the model’s performance. To examine this issue, we analyzed the localization annotations of randomly generated negative pairs. Among pairs with available localization information, 48.9% shared at least one subcellular localization annotation (ie, colocalization), whereas 51.1% did not share any localization annotation (ie, noncolocalization) (Figure S3). We acknowledged that the subcellular localization constraint of negative samples was not taken into account in the current work. Although how to incorporate subcellular localization information into negative sample selection remains an open question, it provides a potential direction for future method improvement.

In terms of methodology, PLM-derived representations demonstrated clear advantages for predicting *A. thaliana*–virus PPIs. Compared with GLMs and traditional protein sequence encoding schemes, these representations exhibited superior performance. This finding suggests that representations learned directly from protein sequences are more effective in capturing biologically relevant features associated with protein interactions. During large-scale pretraining, PLMs learn contextual relationships among amino acids and implicitly encode structural constraints and functional characteristics (Rives et al. 2021; Elnaggar et al. 2022; Lin et al. 2023; Candido et al. 2026). By comparison, traditional protein sequence encodings mainly capture manually defined sequence patterns (Shen et al. 2007; Chen et al. 2011; Kao et al. 2020), but may have limited ability to represent complex and long-range dependencies within protein sequences. GLM-derived representations capture genomic-level sequence information, including

nucleotide composition and evolutionary patterns (Ji et al. 2021; Benegas et al. 2023; Zhou et al. 2023; Sanabria et al. 2024; Shao and Yan 2024; Zhai et al. 2026, 2025). Comparatively, plant-specific genomic models showed relatively better performance in plant–virus interaction prediction. However, GLMs may also capture genomic features that are less relevant to protein interactions. This limitation may be particularly relevant for plant viruses with compact and complex genomes, where overlapping open reading frames, ribosomal frameshifting, and poly-protein processing complicate the relationship between genomic sequences and mature viral protein functions (Wu et al. 2022, 2024; Atabekova et al. 2023; Xavier and Whitfield 2023). Consequently, these learned representations are inferior to PLM-derived representations in predicting plant–virus PPIs. Notably, the performance differences among different sequence representations were consistent across classification algorithms, and XGBoost performed best for the current plant–virus PPI prediction task.

Integrative analyses based on the host PPI network revealed that different pathogen types adopt distinct interference strategies when targeting host proteins. Both viruses and bacteria tend to converge on functionally related host proteins, indicating that pathogens do not disrupt host systems in a broad, random manner but instead preferentially target specific regulatory modules (Weßling et al. 2014; Li et al. 2017; Osborne et al. 2023). This modular targeting strategy may enable pathogens to efficiently manipulate key host processes despite having a limited number of proteins, particularly in viruses. Further comparisons revealed clear differences between viruses and bacteria in the hierarchical levels at which they interfere with host networks. Bacteria tend to target central regulatory nodes within host interaction networks, consistent with their relatively larger and more functionally diverse repertoires of effector proteins. In contrast, viruses, which encode fewer proteins, are more likely to influence host cellular processes by coordinately targeting relatively peripheral yet functionally connected modules. Consistent with these shared network-level patterns, functional enrichment analysis also revealed common host-targeting features across pathogen groups, particularly related to proteostasis and stress responses (Zhang and Wang 2012; Herath et al. 2020). In contrast, virus- and bacterium-targeted proteins exhibited distinct functional preferences: viruses were primarily associated with chloroplast and energy-related processes (Li et al. 2021), whereas bacteria were more closely linked to immune regulation (Jones et al. 2024). In addition, virus-specific functional signatures indicate that different viral pathogens may adopt distinct strategies to modulate host cellular processes. Collectively, the concordance between network topology and functional profiles provides computational evidence supporting the potential biological significance of the predicted interactions and highlights the value of integrating network and functional analyses to understand the global landscape of plant–pathogen PPI networks.

In summary, the AraVirusPPI predictor developed in this study represents the first machine learning-based plant–virus PPI prediction method for the community. The highlight of AraVirusPPI should be ascribed to the introduction of a pretrained PLM (ie, ESMC), which captures rich semantic information to reflect the interaction potential between plant–virus protein pairs. To explore the real-world application of AraVirusPPI, we conducted proteome-wide PPI predictions between *A. thaliana* and *CabLCV*, providing a global understanding of the *A. thaliana*–*CabLCV* PPI network. Nevertheless, the current AraVirusPPI still suffers from several limitations. First, the training and evaluation of the prediction model rely on currently available

plant-virus interaction data, which remain relatively limited. This limitation may restrict the model's ability to learn rare or species-specific interaction patterns. Second, the interactions predicted by AraVirusPPI are purely computational inferences and require further experimental validation to establish their biological validity. For instance, the *A. thaliana*-CabLCV PPI network analysis was based on predicted PPIs, and the resulting network patterns should be interpreted with caution. Moreover, extending this framework to additional plant species and diverse pathogen types will facilitate broader comparative analyses of pathogen interference strategies within host systems. Taken together, we hope that AraVirusPPI can serve as an important tool to accelerate the identification of plant-virus PPIs, further guiding hypothesis-driven experimental efforts to decipher the plant-virus relationship.

Materials and methods

Data collection and preparation

To construct a high-quality PPI dataset between *A. thaliana* and plant viruses, 25 experimentally validated PPIs were first collected from multiple public databases, including BioGRID (<https://thebiogrid.org>), HPIDB (Ammari et al. 2016), IntAct (<https://www.ebi.ac.uk/intact>), and VirusMentha (Calderone et al. 2015). Moreover, 940 experimentally validated PPIs were collected from the literature before 2024. CD-HIT v4.8.1 (Li and Godzik 2006) at a 40% identity threshold was used to reduce sequence redundancy in the collected PPIs. As a result, 666 nonredundant PPIs were obtained, which cover 13 RNA viruses and 7 DNA viruses (Table S1).

The complete proteome of *A. thaliana* based on the Araport11 annotation release was downloaded from the TAIR database (<http://arabidopsis.org>), and viral proteomes were obtained from UniProt (<https://www.uniprot.org/>). Detailed pathogen strain/isolate information is provided in Table 1. These datasets were further clustered using CD-HIT with the same threshold, reducing *A. thaliana* proteins from 27,562 to 14,930 and viral proteins from 134 to 80. For negative sample construction, a random pairing strategy was adopted to generate 6,660 plant-virus protein pairs with the exclusion of positive samples, resulting in a 1:10 ratio of positives to negatives. The distribution of positive and negative samples across different virus species is summarized in Table 1. For model training and evaluation, the compiled dataset was randomly split into a training and a testing set at an 8:2 ratio. The training set contained 5,860 pairs (533 positives and 5,327 negatives), and the testing set contained 1,466 pairs (133 positives and 1,333 negatives). An additional test dataset was also compiled using 19 experimentally validated PPIs from 12 virus species not included in the original training or test sets (Table S4). Negative samples were generated using the same random pairing strategy described above.

Protein embedding extraction

Protein language models

In this study, 4 transformer-based PLMs were employed to extract feature representations from protein sequences, including ESM-1b (esm1b_t33_650M_UR50S) (Rives et al. 2021), ESM-2 (esm2_t33_650M_UR50D) (Lin et al. 2023), ESMC (esm2-600m-2024-12) (Candido et al. 2026), and ProtT5 (prot_t5_xl_half_uniref50-enc; from the ProtTrans framework) (Elnaggar et al. 2022). ESM-1b contains approximately 650 million parameters and consists of 33 transformer

layers, enabling the generation of context-aware embeddings for amino acid residues. ESM-2 also contains approximately 650 million parameters and incorporates improvements over ESM-1. ESMC represents a new generation of the ESM models, with optimized architecture and training strategies for efficient and accurate sequence representation. Likewise, ProtT5 is pretrained on large-scale protein sequence datasets to capture contextual information. For all the above models, a unified feature extraction pipeline was applied. Residue-level embeddings were first obtained for each input sequence, followed by mean pooling along the sequence dimension to generate fixed-length protein-level vectors. As a result, each protein sequence was represented by a vector of fixed dimensionality (ESM-1b/ESM-2: 1280; ESMC: 1152; ProtT5: 1024). These vectors were subsequently used as input for PPI prediction. For each protein pair, the corresponding feature vector was constructed by horizontally concatenating the embeddings of the 2 proteins.

Traditional protein sequence encoding schemes

In this study, 4 classical protein sequence encoding schemes were also employed to characterize sequence features, including DPC, CT, CKSAAP, and AC. DPC (Bhasin and Raghava 2004; Kao et al. 2020) denotes the frequency distribution of adjacent amino acid pairs in a protein sequence, which is a 400-dimensional feature vector. CT (Shen et al. 2007) calculates the frequencies of consecutive triad category combinations. Since the 20 standard amino acids were classified into 7 residue types, CT is a 343-dimensional feature vector. CKSAAP encoding (Wang et al. 2009; Chen et al. 2011) captures the composition features of amino acid pairs with different spacing distances. The maximum spacing parameter was set to $k=2$, and features from different spacing distances were concatenated to form a 1200-dimensional vector, followed by normalization to eliminate the influence of sequence length. AC encoding (Guo et al. 2008) characterizes protein sequences based on the autocorrelation of physicochemical properties. Each amino acid is represented by a vector of 7 predefined physicochemical attributes, converting a sequence of length L into an $L \times 7$ numerical matrix. The *lag* parameter is set from 1 to 30, where each *lag* defines a positional interval between residues (i and $i + \text{lag}$). For each *lag* and each property, the autocorrelation is calculated by subtracting the mean value and averaging the products of values at positions i and $i + \text{lag}$ across the sequence. This yields 7 values per *lag*, and concatenating across all *lags* results in a 210-dimensional feature vector.

Gene embedding extraction

In this study, 6 pretrained GLMs, including GPN, DNABERT, DNABERT-2, GROVER, megaDNA, and PlantCAD2, were employed to extract feature representations from gene sequences. GPN (gpn-brassicales) is a CNN-based GLM pretrained on *A. thaliana* and other Brassicales genomes (Benegas et al. 2023). To obtain fixed-length sequence representations and eliminate the influence of gene length variation, mean pooling was applied along the sequence dimension to all base embeddings, yielding a 512-dimensional vector per gene. Both DNABERT (6-new-12w-0) and DNABERT-2 (DNABERT-2-117M) are based on transformer architectures (Ji et al. 2021; Zhou et al. 2023). Token-level embeddings were averaged across the sequence dimension, resulting in 768-dimensional gene representations. GROVER is also a pretrained transformer-based model (Sanabria et al. 2024). Hidden states from the final encoder layer were extracted as sequence representations. Mean pooling was then performed along the

sequence length dimension to generate a 609-dimensional embedding vector for each gene. megaDNA (megaDNA_phage_145M) is a transformer-based GLM designed for long-sequence modeling (Shao and Yan 2024). The model-generated block-level embeddings were aggregated using mean pooling, yielding a 512-dimensional representation for each gene. PlantCAD2 (PlantCAD2-Medium-148-d1024) is a plant-specific GLM pretrained on angiosperm genomes using a long-context Mamba2 architecture (Zhai et al. 2026). Mean pooling was applied to generate 1024-dimensional gene representations. For the integrated protein-genomic representation, ESMC embeddings of the plant and virus proteins were concatenated with the corresponding GPN embeddings of their genes, generating a 3328-dimensional representation ($1,152 \times 2 + 512 \times 2$) for each interaction pair.

Model architecture and parameter optimization

This study employed multiple machine learning and deep learning classifiers, including XGBoost, RF, SVM, and CNN. The XGBoost model was implemented using the Python xgboost package v2.0.3 (<https://xgboost.readthedocs.io/>), while the RF and SVM models were implemented using scikit-learn v1.4.2 (<https://scikit-learn.org/>). For these 3 traditional machine learning models, class weight balancing was applied during training. Key parameters were systematically optimized via grid search (GridSearchCV), focusing on the number of trees, maximum tree depth, learning rate, and feature subsampling ratio for tree-based models, as well as the regularization coefficient and kernel-related parameters for SVM. Parameter tuning was evaluated using 5-fold cross-validation, and the hyperparameter search ranges and optimized parameters are provided in Table S3. The 2 hyperparameter optimization strategies “roc_auc” and “average_precision” were evaluated for XGBoost, and “roc_auc” was selected for the final optimization (Figure S4). Using the optimized parameters, the training set was stratified into 5 folds. Five models were then trained, each on 4 folds with a different fold held out. These 5 models were combined into an ensemble and evaluated on the independent test set, with their prediction scores averaged to obtain the final prediction. The CNN model was constructed using the TensorFlow v2.17.0 framework (<https://www.tensorflow.org/>), adopting a one-dimensional convolutional architecture for feature learning. The CNN model consisted of 2 Conv1D layers with 32 and 64 filters (kernel size = 3), respectively, each followed by a ReLU activation function and a max-pooling layer. The extracted features were flattened and passed through a fully connected layer with 128 neurons, followed by dropout (0.3) and a sigmoid output layer. The model was trained using the Adam optimizer (learning rate = 0.001) with binary cross-entropy loss for a maximum of 100 epochs with a batch size of 32. Early stopping was applied with a patience of 10 epochs, and the best-performing model weights were restored for evaluation. All models were trained and evaluated on the same dataset splits and assessment criteria to ensure comparability across different methods.

Performance evaluation

Model performance was systematically evaluated using multiple commonly adopted classification metrics (ie, accuracy, precision, recall, and F1-score), which are defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

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

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

$$\text{F1 - score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Here, TP, FP, TN, and FN represent the numbers of true positives, false positives, true negatives, and false negatives, respectively, with a classification threshold of 0.5. To further evaluate overall discriminative ability across different decision thresholds, receiver operating characteristic (ROC) and precision recall (PR) curves were plotted, and the corresponding areas under the curves (AUROC and AUPRC) were computed. Given the pronounced imbalance between positive and negative samples in this study, the PR curve and its associated AUPRC were used as key indicators to better reflect the model's predictive performance.

Construction of *A. thaliana* PPI and *A. thaliana*-pathogen PPI networks

To construct the *A. thaliana* PPI network, interaction data published before 2025 were collected from BioGRID, IntAct, and TAIR. Note that only experimentally validated physical interactions among *A. thaliana* proteins were retained, resulting in a high-confidence host PPI network (55,517 PPIs among 11,411 proteins). In parallel, *A. thaliana*-pathogen PPI networks were further constructed for 2 viruses, *CabLCV* and *TuMV*, and 1 bacterium, *Psy*. *A. thaliana*-*CabLCV* interactions were predicted using the developed AraVirusPPI model. Predicted interactions with scores above 0.5 were further ranked, and the top 500 high-confidence interactions covering 245 *A. thaliana* proteins and 5 *CabLCV* proteins were further retained (Table S5). Four hundred and twenty-one *A. thaliana*-*TuMV* interactions between 461 *A. thaliana* proteins and 11 *TuMV* proteins were derived from the experimentally validated plant-virus PPI dataset compiled in this study, whereas *A. thaliana*-*Psy* interactions (248 PPIs between 145 *A. thaliana* proteins and 65 *Psy* proteins) were obtained from the experimentally validated data compiled in AraPathogen 2.0 (Lei et al. 2024).

Network analysis

All network analyses focused on host proteins targeted by pathogens within the *A. thaliana* PPI network. Convergence analysis was conducted to evaluate the overlap of host target proteins within and between pathogens. The number of shared target proteins was calculated for intra- and inter-species comparisons. An empirical null distribution was generated by randomly sampling proteins with replacement from the *A. thaliana* protein set, matching the number of observed targets in each simulation. This procedure was repeated 10,000 times. Statistical significance was assessed by comparing observed values to the null distribution. For intraspecies comparisons, the *P*-value was calculated as the proportion of simulations in which the number of shared targets was less than or equal to the observed value. For interspecies comparisons, the *P*-value was defined as the proportion of simulations in which the value was greater than or equal to the observed. In both cases, $P = n/10,000$, where *n* is the number of simulations satisfying the corresponding criterion. $P < 0.001$ was reported when $n < 10$ (Osborne et al. 2023).

Shortest path distances were calculated using the `shortest_paths` function in `igraph` v2.1.4 (<https://igraph.org>), and the average distance for each protein was obtained after excluding unreachable pairs. In intraspecies distance analysis, to control for differences in sample size, nontarget proteins were randomly sampled without replacement 100 times to match the number of target proteins, and the average was used. Network topological features, including degree and betweenness centrality, were calculated using the corresponding functions in `igraph`. For comparison with target proteins, nontarget proteins were randomly sampled 1,000 times, with each sampling matched to the number of target proteins. The cumulative distribution curves of nontarget proteins were generated by averaging the curves from the 1,000 random samplings. Statistical significance was assessed using the Wilcoxon test (two-sided), and visualization was performed using `ggplot2` v3.5.1 (Wickham 2016).

Functional enrichment analysis

We conducted GO enrichment analyses of the host proteins targeted by the 3 pathogens. All enrichment analyses were performed using the R package `clusterProfiler` v4.10.1 (Wu et al. 2021), with `enrichGO` used for GO enrichment analysis. GO annotations were derived from `org.At.tair.db` (<https://bioconductor.org/>) using TAIR gene identifiers. Significance thresholds were set to `pvalueCutoff`=0.05 and `qvalueCutoff`=0.05. Multiple testing correction was performed using the Benjamini–Hochberg (BH) method, and adjusted *P*-values (`p.adjust`) were used to control the false discovery rate. GO terms with *P.adjust* < 0.05 were considered significantly enriched. To facilitate functional comparisons among pathogen target proteins, significantly enriched GO terms from different pathogens were visualized using the `heatmap` package v1.0.12 (<https://CRAN.R-project.org/package=heatmap>).

Author contributions

S.Z., W.S., and Z.Z. designed the research. S.Z., J.L., M.Y., and L.J. developed the algorithms. S.Z., J.L., and M.Y. analyzed the data. S.Z. and Z.Z. wrote the first draft of the paper. S.Z., J.L., M.Y., L.J., Z.L., C.W., W.S., and Z.Z. reviewed the paper.

Supplementary material

Supplementary material is available at *Plant Physiology* online.

Funding

This work was supported by the National Natural Science Foundation of China (31271414, 31471249, and 31970645) and the Chinese Universities Scientific Fund (2026TC049). We appreciate the support of the High-performance Computing Platform of China Agricultural University for providing computational resources.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Data availability

The authors confirm that the data supporting the findings of this study are available within the article and its [supplementary materials](#). The source code and datasets used in this study are available on GitHub (<https://github.com/skzhang-zsk/AraVirusPPI>).

References

- Ammari MG, Gresham CR, Mccarthy FM, Nanduri B. 2016. HPIDB 2.0: a curated database for host-pathogen interactions. *Database*. 2016:baw103. <https://doi.org/10.1093/database/baw103>.
- Atabekova AK, Solovieva AD, Chergintsev DA, Solovyev AG, Morozov SY. 2023. Role of plant virus movement proteins in suppression of host RNAi defense. *Int J Mol Sci*. 24:9049. <https://doi.org/10.3390/ijms24109049>.
- Benegas G, Batra SS, Song YS. 2023. DNA language models are powerful predictors of genome-wide variant effects. *Proc Natl Acad Sci U S A*. 120:e2311219120. <https://doi.org/10.1073/pnas.2311219120>.
- Bhasin M, Raghava GP. 2004. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J Biol Chem*. 279:23262–23266. <https://doi.org/10.1074/jbc.M401932200>.
- Calderone A, Licata L, Cesareni G. 2015. VirusMentha: a new resource for virus-host protein interactions. *Nucleic Acids Res*. 43:D588–D592. <https://doi.org/10.1093/nar/gku830>.
- Candido S et al. 2026 Jun 4. Language modeling materializes a world model of protein biology [preprint]. *bioRxiv* 2026.06.03.729735. <https://doi.org/10.64898/2026.06.03.729735>.
- Carvalho MF, Lazarowitz SG. 2004. Interaction of the movement protein NSP and the Arabidopsis acetyltransferase AtNSI is necessary for cabbage leaf curl geminivirus infection and pathogenicity. *J Virol*. 78:11161–11171. <https://doi.org/10.1128/JVI.78.20.11161-11171.2004>.
- Ceulemans E, Ibrahim HMM, De Coninck B, Goossens A. 2021. Pathogen effectors: exploiting the promiscuity of plant signaling hubs. *Trends Plant Sci*. 26:780–795. <https://doi.org/10.1016/j.tplants.2021.01.005>.
- Chen H, Chen J, Zhao Y, Liu F, Fu ZQ. 2022. *Pseudomonas syringae* pathogens. *Trends Microbiol*. 30:912–913. <https://doi.org/10.1016/j.tim.2022.03.002>.
- Chen Z et al. 2011. Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PLoS One*. 6:e22930. <https://doi.org/10.1371/journal.pone.0022930>.
- Dangl JL, Horvath DM, Staskawicz BJ. 2013. Pivoting the plant immune system from dissection to deployment. *Science*. 341:746–751. <https://doi.org/10.1126/science.1236011>.
- Elnaggar A et al. 2022. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell*. 44:7112–7127. <https://doi.org/10.1109/TPAMI.2021.3095381>.
- Gouveia-Mageste BC et al. 2021. A plant-specific syntaxin-6 protein contributes to the intracytoplasmic route for the begomovirus CabLCV. *Plant Physiol*. 187:158–173. <https://doi.org/10.1093/plphys/kiab252>.
- Guo Y, Yu L, Wen Z, Li M. 2008. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res*. 36:3025–3030. <https://doi.org/10.1093/nar/gkn159>.
- He S, Creasey Krainer KM. 2020. Pandemics of people and plants: which is the greater threat to food security? *Mol Plant*. 13:933–934. <https://doi.org/10.1016/j.molp.2020.06.007>.
- Herath V, Gayral M, Miller RK, Verchot J. 2020. BIP and the unfolded protein response are important for potyvirus and potexvirus infection. *Plant Signal Behav*. 15:1807723. <https://doi.org/10.1080/15592324.2020.1807723>.
- Hu W, Ohue M. 2024. SpatialPPI: three-dimensional space protein-protein interaction prediction with AlphaFold multimer. *Comput Struct Biotechnol J*. 23:1214–1225. <https://doi.org/10.1016/j.csbj.2024.03.009>.
- Huang YW et al. 2023. A viral movement protein co-opts endoplasmic reticulum luminal-binding protein and calreticulin to promote intracellular movement. *Plant Physiol*. 191:904–924. <https://doi.org/10.1093/plphys/kiac547>.

- Ibrahim A *et al.* 2020. Plant SNAREs SYP22 and SYP23 interact with tobacco mosaic virus 126 kDa protein and SYP2s are required for normal local virus accumulation and spread. *Virology*. 547:57–71. <https://doi.org/10.1016/j.virol.2020.04.002>.
- Ji Y, Zhou Z, Liu H, Davuluri RV. 2021. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*. 37:2112–2120. <https://doi.org/10.1093/bioinformatics/btab083>.
- Jiang L *et al.* 2025. Graph neural network integrated with pretrained protein language model for predicting human-virus protein-protein interactions. *Brief Bioinform*. 26:bbaf461. <https://doi.org/10.1093/bib/bbaf461>.
- Jiang T, Zhou T. 2023. Unraveling the mechanisms of virus-induced symptom development in plants. *Plants*. 12:2830. <https://doi.org/10.3390/plants12152830>.
- Jones JDG, Staskawicz BJ, Dangl JL. 2024. The plant immune system: from discovery to deployment. *Cell*. 187:2095–2116. <https://doi.org/10.1016/j.cell.2024.03.045>.
- Kao HJ, Nguyen VN, Huang KY, Chang WC, Lee TY. 2020. SuccSite: incorporating amino acid composition and informative k-spaced amino acid pairs to identify protein succinylation sites. *Genomics Proteomics Bioinformatics*. 18:208–219. <https://doi.org/10.1016/j.gpb.2018.10.010>.
- Lei C *et al.* 2024. AraPathogen2.0: an improved prediction of plant-pathogen protein-protein interactions empowered by the natural language processing technique. *J Proteome Res*. 23:494–499. <https://doi.org/10.1021/acs.jproteome.3c00364>.
- Li H, Zhou Y, Zhang Z. 2017. Network analysis reveals a common host-pathogen interaction pattern in Arabidopsis immune responses. *Front Plant Sci*. 8:893. <https://doi.org/10.3389/fpls.2017.00893>.
- Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 22:1658–1659. <https://doi.org/10.1093/bioinformatics/btl158>.
- Li Y, Xiong R, Bernards M, Wang A. 2016. Recruitment of Arabidopsis RNA helicase ATRH9 to the viral replication complex by viral replicase to promote turnip mosaic virus replication. *Sci Rep*. 6:30297. <https://doi.org/10.1038/srep30297>.
- Li Z *et al.* 2021. NSvc4 encoded by rice stripe virus targets host chloroplasts to suppress chloroplast-mediated defense. *Viruses*. 14:36. <https://doi.org/10.3390/v14010036>.
- Lian X, Yang X, Yang S, Zhang Z. 2021. Current status and future perspectives of computational studies on human-virus protein-protein interactions. *Brief Bioinform*. 22:bbab029. <https://doi.org/10.1093/bib/bbab029>.
- Liang P *et al.* 2025. A plant virus manipulates both its host plant and the insect that facilitates its transmission. *Sci Adv*. 11:eadr4563. <https://doi.org/10.1126/sciadv.adr4563>.
- Lin Z *et al.* 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. 379:1123–1130. <https://doi.org/10.1126/science.ade2574>.
- Liu D *et al.* 2025. PLM-interact: extending protein language models to predict protein-protein interactions. *Nat Commun*. 16:9012. <https://doi.org/10.1038/s41467-025-64512-w>.
- Lu J *et al.* 2024. Copper regulates the host innate immune response against bacterial infection via activation of ALPK1 kinase. *Proc Natl Acad Sci U S A*. 121:e2311630121. <https://doi.org/10.1073/pnas.2311630121>.
- Martínez F *et al.* 2016. Interaction network of tobacco etch potyvirus NIa protein with the host proteome during infection. *BMC Genomics*. 17:87. <https://doi.org/10.1186/s12864-016-2394-y>.
- Martínez F *et al.* 2023. A binary interaction map between turnip mosaic virus and *Arabidopsis thaliana* proteomes. *Commun Biol*. 6:28. <https://doi.org/10.1038/s42003-023-04427-8>.
- Osborne R *et al.* 2023. Symbiont-host interactome mapping reveals effector-targeted modulation of hormone networks and activation of growth promotion. *Nat Commun*. 14:4065. <https://doi.org/10.1038/s41467-023-39885-5>.
- Ray S, Casteel CL. 2022. Effector-mediated plant-virus-vector interactions. *Plant Cell*. 34:1514–1531. <https://doi.org/10.1093/plcell/koac058>.
- Rives A *et al.* 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A*. 118:e2016239118. <https://doi.org/10.1073/pnas.2016239118>.
- Rodríguez-Peña R, Mounadi KE, García-Ruiz H. 2021. Changes in subcellular localization of host proteins induced by plant viruses. *Viruses*. 13:677. <https://doi.org/10.3390/v13040677>.
- Rosas-Díaz T *et al.* 2023. The transcriptional regulator JAZ8 interacts with the C2 protein from geminiviruses and limits the geminiviral infection in Arabidopsis. *J Integr Plant Biol*. 65:1826–1840. <https://doi.org/10.1111/jipb.13482>.
- Sanabria M, Hirsch J, Joubert PM, Poetsch AR. 2024. DNA language model GROVER learns sequence context in the human genome. *Nat Mach Intell*. 6:911–923. <https://doi.org/10.1038/s42256-024-00872-0>.
- Savary S *et al.* 2019. The global burden of pathogens and pests on major food crops. *Nat Ecol Evol*. 3:430–439. <https://doi.org/10.1038/s41559-018-0793-y>.
- Shao B, Yan J. 2024. A long-context language model for deciphering and generating bacteriophage genomes. *Nat Commun*. 15:9392. <https://doi.org/10.1038/s41467-024-53759-4>.
- Shen J *et al.* 2007. Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A*. 104:4337–4341. <https://doi.org/10.1073/pnas.0607879104>.
- Stukenbrock E, Gurr S. 2023. Address the growing urgency of fungal disease in crops. *Nature*. 617:31–34. <https://doi.org/10.1038/d41586-023-01465-4>.
- Thiébaud O *et al.* 2009. A new plant protein interacts with eIF3 and 60S to enhance virus-activated translation re-initiation. *EMBO J*. 28:3171–3184. <https://doi.org/10.1038/emboj.2009.256>.
- Tilsner J, Kriechbaumer V. 2022. Reticulons 3 and 6 interact with viral movement proteins. *Mol Plant Pathol*. 23:1807–1814. <https://doi.org/10.1111/mpp.13261>.
- Wang XB, Wu LY, Wang YC, Deng NY. 2009. Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs. *Protein Eng Des Sel*. 22:707–712. <https://doi.org/10.1093/protein/gzp055>.
- Weßling R *et al.* 2014. Convergent targeting of a common host protein-network by pathogen effectors from three kingdoms of life. *Cell Host Microbe*. 16:364–375. <https://doi.org/10.1016/j.chom.2014.08.004>.
- Wickham H. 2016. *Ggplot2: elegant graphics for data analysis*. Springer-Verlag.
- Wu G *et al.* 2022. Acidic dileucine motifs in the cylindrical inclusion protein of turnip mosaic virus are crucial for endosomal targeting and viral replication. *Mol Plant Pathol*. 23:1381–1389. <https://doi.org/10.1111/mpp.13231>.
- Wu J *et al.* 2024. Plant virology in the 21st century in China: recent advances and future directions. *J Integr Plant Biol*. 66:579–622. <https://doi.org/10.1111/jipb.13580>.
- Wu T *et al.* 2021. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation*. 2:100141. <https://doi.org/10.1016/j.xinn.2021.100141>.
- Xavier CAD, Whitfield AE. 2023. Plant virology. *Curr Biol*. 33:R478–R484. <https://doi.org/10.1016/j.cub.2023.03.038>.
- Yang S, Li H, He H, Zhou Y, Zhang Z. 2019. Critical assessment and performance improvement of plant-pathogen protein-protein interaction prediction methods. *Brief Bioinform*. 20:274–287. <https://doi.org/10.1093/bib/bbx123>.
- Yu XQ *et al.* 2024. PTI-ETI synergistic signal mechanisms in plant immunity. *Plant Biotechnol J*. 22:2113–2128. <https://doi.org/10.1111/pbi.14332>.
- Zhai J *et al.* 2025. Cross-species modeling of plant genomes at single-nucleotide resolution using a pretrained DNA language model. *Proc Natl Acad Sci U S A*. 122:e2421738122. <https://doi.org/10.1073/pnas.2421738122>.

- Zhai J *et al.* 2026. PlantCAD2: a DNA foundation model for interpreting genomes across flowering plants. *Cell Genom.* 6:101329. <https://doi.org/10.1016/j.xgen.2026.101329>.
- Zhang L, Wang A. 2012. Virus-induced ER stress and the unfolded protein response. *Front Plant Sci.* 3:293. <https://doi.org/10.3389/fpls.2012.00293>.
- Zhou Y, Zhou YS, He F, Song J, Zhang Z. 2012. Can simple codon pair usage predict protein-protein interaction? *Mol Biosyst.* 8:1396–1404. <https://doi.org/10.1039/C2MB05427B>.
- Zhou Z *et al.* 2023. DNABERT-2: efficient foundation model and benchmark for multi-species genome [preprint]. arXiv, arXiv:2306.15006. <https://doi.org/10.48550/arXiv.2306.15006>