the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2025) **124,** e70593 doi: 10.1111/tpj.70593

TECHNICAL ADVANCE

PPEPFinder: A deep learning framework integrating sequence embeddings and structural graph representations for predicting fungal and oomycete effector proteins

Mengdi Yuan, Shaoke Zhang, Jiajun Li, Jingyan Zheng, Chenping Lei, Zhuojun Wang, Miao Zhao, Wenyu Shi* and Ziding Zhang*

State Key Laboratory of Animal Biotech Breeding, College of Biological Sciences, China Agricultural University, Beijing 100193, China

Received 10 July 2025; revised 28 September 2025; accepted 4 November 2025.
*For correspondence (e-mail shiwy@cau.edu.cn and zidingzhang@cau.edu.cn).

SUMMARY

Plant pathogens, such as fungi and oomycetes, primarily rely on the secretion of effector proteins to successfully colonize host plants and suppress immune responses, ultimately leading to disease symptoms. Therefore, accurately identifying these effector proteins is crucial for understanding the pathogenic mechanisms and developing strategies for disease resistance. However, the current effector protein identification mainly depends on experimental screening methods, which are labor-intensive and costly. In recent years, the rapid development of deep learning approaches and the emergence of protein language models (PLMs) have presented unprecedented opportunities for developing new bioinformatics methods to identify effector proteins. In this context, we propose Plant Pathogen Effector Protein Finder (PPEPFinder), an integrated deep learning framework designed to predict effector proteins of fungi and oomycetes. PPEPFinder consists of three individual predictive models: (i) a sequence-based transformer model that utilizes rich semantic embeddings generated by a pre-trained PLM, Evolutionary Scale Modeling (ESM), as input; (ii) two structure-based Graph Attention Network models that represent protein structures as residue contact graphs using the ESM-generated embeddings or the structure pre-trained model SaProt embeddings as node features. To maximize the predictive performance, PPEPFinder employs a logistic regression model to aggregate the three predictors' outputs into a final prediction score. Compared to existing state-of-the-art effector prediction tools, PPEPFinder demonstrates superior performance by jointly leveraging sequence and structural information. We have made PPEPFinder and its associated datasets freely accessible to the community through an online prediction server (http://zzdlab.com/PPEPFinder/) and GitHub (https://github.com/mdiyuan/PPEPFinder).

Keywords: effector protein, prediction, plant pathogens, fungi, oomycetes, deep learning, protein language models.

INTRODUCTION

Plant pathogens, including fungi and oomycetes, threaten global food security and ecosystem stability (Stukenbrock & Gurr, 2023). Although fungi and oomycetes are phylogenetically unrelated, they have evolved similar filamentous growth forms and infection structures through convergent evolution. They can invade plant cells, extract nutrients, and disrupt normal plant growth and development. In response to the invasion of plant pathogens, plants have evolved a sophisticated immune system over time. Likewise, pathogens have developed diverse strategies to circumvent plant immune defenses and modulate host cellular signaling pathways, facilitating their invasion, proliferation, and colonization. One such approach involves the spatiotemporally regulated secretion of effector proteins into the host. These effectors act in the apoplast or within host cells to promote pathogen entry, suppress plant immune responses, and reprogram host metabolism (Lo Presti et al., 2015; Toruño et al., 2016).

It has been common knowledge that some conserved motifs/domains exist in oomycete effector proteins. For example, the conserved RXLR motif, which frequently

appears at the N-terminus of cytoplasmic effectors, is typically located immediately downstream of the signal peptide cleavage site. This motif may play a role in the proteolytic processing of the effectors before secretion and mediate their subsequent translocation into host plant cells (Ellis & Dodds, 2011; Govers & Bouwmeester, 2008; Wawra et al., 2017). Although motif screening or Hidden Markov Model (HMM)-based domain assignment approaches have become a routine strategy in predicting oomycete cytoplasmic effectors, these methods fail to identify potential effectors that lack typical motifs or domains (Wood et al., 2020). In contrast, predicting fungal effectors is more challenging due to their higher sequence diversity and lack of clear conserved motifs. Early studies typically relied on some general features of effectors, such as small molecular size and high cysteine content (Sperschneider et al., 2016, 2018; Stergiopoulos & de Wit, 2009). However, these feature-based screening methods have low predictive efficiency and are prone to yield biased results.

With the advancement of computational technologies, machine learning methods have become increasingly popular for predicting effector proteins. To date, various approaches, including Support Vector Machines (SVM), Random Forests (RF), HMM, and deep learning, have been successfully applied to predict bacterial effectors, yielding satisfactory results (Dong et al., 2015; Eichinger et al., 2016; Hui et al., 2020; Zhang et al., 2022, 2023).

In contrast, the machine learning-based prediction of fungal and oomycete effector proteins is still in its infancy, although some predictive methods have been elegantly developed (Kristianingsih & MacLean, 2021; Li et al., 2024; Mishra et al., 2019; Nur et al., 2023; Sperschneider et al., 2016, 2018; Sperschneider & Dodds, 2022; Zhao et al., 2024). EffectorP is probably the first machine learning-based predictor for fungal effectors, employing a Naive Bayes model combined with amino acid physicochemical properties, and achieving a prediction accuracy exceeding 70% (Sperschneider et al., 2016). Its upgraded version, EffectorP 2.0, incorporates ensemble learning techniques and a larger training dataset to increase predictive performance (Sperschneider et al., 2018). The latest release, EffectorP3.0, further introduces subcellular localization prediction for secreted proteins, distinguishing between apoplastic and cytoplasmic effectors (Sperschneider & Dodds, 2022). As a convolutional neural network (CNN)-based classifier, Deepredeff is developed for predicting bacterial, fungal, and oomycete effectors (Kristianingsih & MacLean, 2021). EffectorO-ML is an oomycete effector protein prediction tool trained on experimentally validated datasets, leveraging biochemical properties of N-terminal protein sequences (Nur et al., 2023). More recently, our team developed POOE, a machine learning framework that integrates a pre-trained protein language model (PLM) called ProtTrans with an SVM classifier, achieving a prediction accuracy of 87.4% for oomycete effectors (Zhao et al., 2024). Very recently, Fungtion, a predictor of fungal effector proteins, applied the ESM1b language model to generate protein-embedding vectors, which were subsequently classified by SVM, achieving a prediction accuracy of 89.7% (Li et al., 2024).

Despite the progress achieved by these methods in identifying plant pathogen effector proteins, they are still suffering from several limitations. First, the datasets employed for model training in existing methods remain relatively small, potentially resulting in biased models with limited generalizability. Second, most existing methods still rely on conventional sequence encoding schemes and classical machine learning algorithms, with relatively limited incorporation of deep learning models. Furthermore, protein structural information has yet to be systematically integrated into effector prediction frameworks. Consequently, considerable room remains for method improvement in developing effector protein prediction models.

Although numerous feature engineering-based methods have been proposed for effector protein prediction, routine encoding schemes may fail to capture critical information related to effector proteins, thereby limiting model accuracy. Protein sequences are inherently similar to natural language, wherein amino acids combine in diverse patterns to form functional structures, much like letters form meaningful words and sentences. As a result, deep learning-powered natural language processing techniques have been increasingly adopted in the past several years to deal with protein data (Ofer et al., 2021). By applying unsupervised learning to large-scale protein sequence datasets, a series of PLMs have been developed (Elnaggar et al., 2022; Heinzinger et al., 2019; Lin et al., 2023; Madani et al., 2023; Rives et al., 2021). These models embed individual residues and entire protein sequences into fixed-dimensional feature vectors, effectively capturing the intrinsic semantic information within sequences and uncovering potential relationships among sequence, structure, function, and evolutionary connections. A representative example is ESM1b, a transformer-based PLM trained on a dataset comprising 250 million protein sequences and 86 billion amino acid residues (Rives et al., 2021). More recently, large-scale PLMs pre-trained on protein structural data have also been introduced. For instance, SaProt incorporates a 'structure-aware vocabulary' by integrating residue tokens with structure tokens derived from 3D structural encoding using Foldseek (Su et al., 2024; van Kempen et al., 2024). The semantic representations generated by these pre-trained models, when coupled with deep learning architectures, have demonstrated remarkable performance across a variety of tasks, including protein function prediction, subcellular localization, and structure modeling (Kang et al., 2023; Olsen et al., 2022; Singh et al., 2022; Thumuluri et al., 2022; Zhang et al., 2022).

Nevertheless, applying these advanced techniques to identify plant pathogen effector proteins remains in its early stages.

In the past several years, we have also witnessed that AI-driven protein structure prediction algorithms, such as AlphaFold3 (Abramson et al., 2024), have significantly improved the accuracy of three-dimensional (3D) structure prediction, providing unprecedented opportunities for protein function prediction and analysis. Compared with sequence information, 3D structural information of a protein is more directly associated with its biological function. A widely adopted strategy for leveraging structural information involves representing the protein backbone as a graph, where each amino acid residue is treated as a node and spatial distance relationships between residues are encoded as edges. Based on this graph-based representation, graph neural networks (GNNs) have been extensively applied to tasks such as protein function prediction and protein–protein interaction identification (Huang et al., 2023; Ran et al., 2024; Réau et al., 2023; Wong et al., 2024). Recent studies have demonstrated that GNNs consistently outperform conventional models, including multilayer perceptron (MLP) and CNN, across various structure-related bioinformatics tasks (Chen et al., 2023). Among these GNN architectures, the graph attention network (GAT) (Veličković et al., 2018) has gained popularity for its incorporation of attention mechanisms into the message-passing framework, utilizing multi-head attention to stabilize training and capture diverse structural and contextual dependencies within the graph. Such methodologies facilitate the integration of intricate spatial relationships and residue-level interactions into predictive models, thereby advancing our understanding of structure–function relationships in proteins. However, prediction frameworks based on accurate 3D structural data remain unexplored in the context of plant pathogen effector protein identification.

In this study, we propose an integrated prediction framework, Plant Pathogen Effector Protein Finder (PPEPFinder), designed to accurately identify effector proteins from fungi and oomycetes by combining protein sequence and 3D structural information. The framework comprises three individual models (Figure 1a): (i) a sequence-based transformer model, which utilizes the ESM embeddings as input; (ii) two structure-based GAT models, which construct residue contact networks based on the predicted protein 3D structures, using the ESM and SaProt embeddings as node features, respectively. To effectively integrate the predictive capabilities of these complementary models, PPEPFinder employs a logistic regression (LR) classifier to fuse the outputs of the three models, thereby generating a comprehensive prediction score. Assessment experiments demonstrate that this ensemble strategy consistently outperforms any individual model, yielding superior performance in predicting fungal and oomycete effectors. Further benchmarking experiments showed that PPEPFinder significantly outperforms existing prediction methods, demonstrating outstanding predictive accuracy and robust generalization performance.

RESULTS AND DISCUSSION

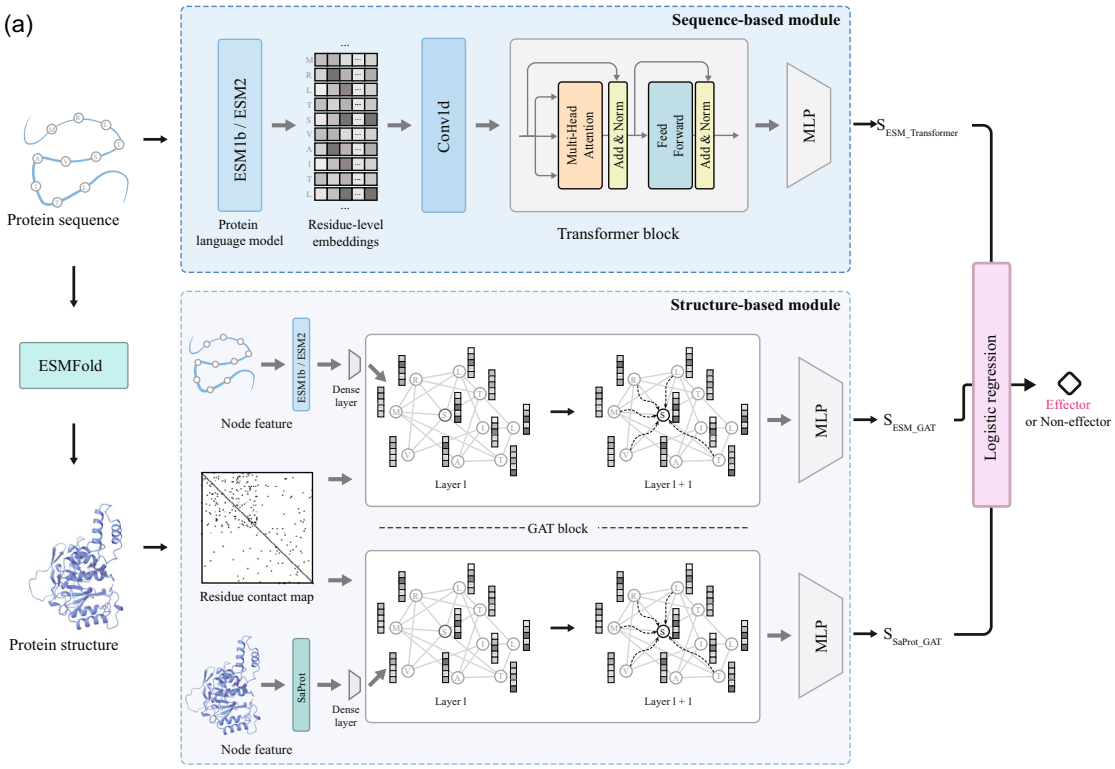
Overall architecture of PPEPFinder

The proposed PPEPFinder framework is composed of three complementary modules designed to jointly leverage protein sequence and structural information for effector prediction (Figure 1a). First, a transformer-based sequence prediction module is employed, which adopts the architecture of DeepSecE (Zhang et al., 2023). In this module, protein sequence embeddings generated by the ESM model are fed into a single-layer transformer encoder for further contextual feature learning. The remaining two structure-based GAT modules are designed to extract topological and geometric features from protein 3D structures. Both modules construct residue contact networks based on 3D structures predicted by ESMFold (Lin et al., 2023), using spatial contact distances as edge features. However, the two modules differ in the types of node features they utilize. One uses 1280-dimensional embeddings generated by the pre-trained PLM ESM (Lin et al., 2023; Rives et al., 2021), while the other employs 1280-dimensional embeddings derived from the structure-aware pre-trained model SaProt (Su et al., 2024). This design allows the two models to capture complementary structural representations independently. Finally, the prediction scores from these three individual models are integrated using an LR classifier, which effectively fuses the diverse representations to generate a unified prediction score. This ensemble strategy enhances the robustness and generalization of effector protein prediction.

To train and test the proposed PPEPFinder framework, we first collected experimentally validated effector proteins from fungi and oomycetes as positive samples. Then, we sampled negative examples to construct two datasets (Fungi_dataset and Oomycetes_dataset) with a ratio of 1:3 positive to negative samples, reflecting the actual class distribution in the real world. Each dataset was divided such that 80% was used for training with five-fold cross-validation, while the remaining 20% served as an independent test set to provide a more rigorous assessment of model performance. More methodological details about PPEPFinder and the preparation of the datasets are described in the Materials and Methods section.

Overall performance of PPEPFinder

To comprehensively assess model performance, we employed several widely used evaluation metrics, including accuracy, precision, recall, and F1-score. Additionally, precision-recall (PR) curves and receiver operating



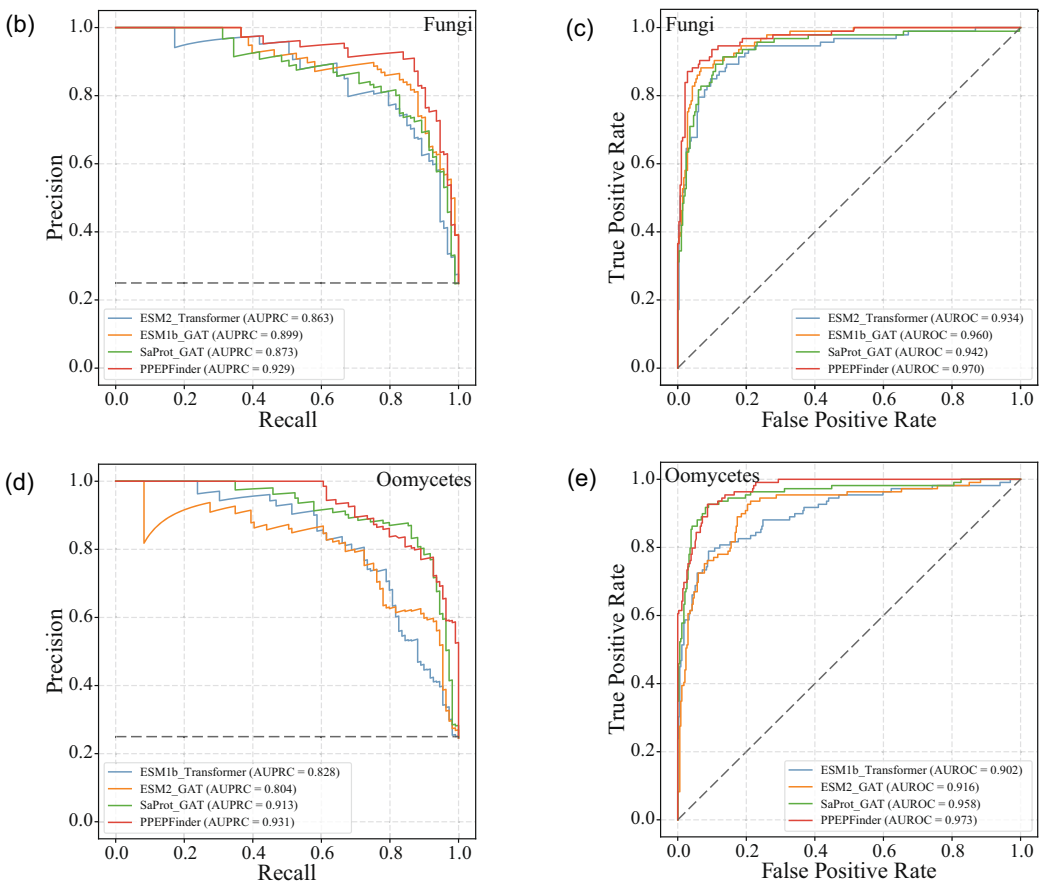


Figure 1. Overview of the Plant Pathogen Effector Protein Finder (PPEPFinder) computational framework and its performance.
(a) The PPEPFinder framework consists of three individual predictive models: a transformer model leveraging ESM, a GAT model based on ESM, and a GAT model utilizing SaProt embeddings. The final prediction is obtained by integrating the outputs of the three models using a logistic regression classifier.
(b, c) Precision-recall (PR) and receiver operating characteristic (ROC) curves for fungal effector prediction.
(d, e) PR and ROC curves for oomycete effector prediction. Dash lines in (b–e) stand for random predictions.

Table 1 Overall performance of our models on the independent test set for fungal and oomycete effectors prediction

	Model	AUROC	AUPRC	Accuracy[a]	Precision[a]	Recall[a]	F1-score[a]
Fungi	ESM2_Transformer	0.934	0.863	0.893	0.805	0.753	0.778
	ESM1b_GAT	0.960	0.899	0.906	0.892	0.710	0.790
	SaProt_GAT	0.942	0.873	0.898	0.867	0.699	0.774
	ESM2_Transformer+ESM1b_GAT	0.956	0.903	0.912	0.905	0.720	0.802
	PPEPFinder	0.970	0.929	0.912	0.917	0.710	0.800
Oomycetes	ESM1b_Transformer	0.902	0.828	0.883	0.807	0.688	0.743
	ESM2_GAT	0.916	0.804	0.885	0.796	0.716	0.754
	SaProt_GAT	0.958	0.913	0.910	0.888	0.725	0.789
	ESM1b_Transformer+SaProt_GAT	0.965	0.915	0.907	0.886	0.716	0.792
	PPEPFinder	0.973	0.931	0.905	0.904	0.688	0.781

[a]The corresponding values were calculated at a prediction threshold of 0.5.

characteristic (ROC) curves were plotted, and the areas under these curves (AUPRC and AUROC) were calculated to provide robust, threshold-independent performance indicators. Notably, the PR curve is particularly informative for evaluating classifier performance on imbalanced datasets, as is often the case in effector protein prediction.

We mainly relied on the independent test dataset to evaluate the predictive performance of PPEPFinder. The performance of three individual predictors and the final integrative model (i.e., PPEPFinder) was summarized in Table 1. The corresponding PR and ROC curves were also plotted for each model (Figure 1b–e). Notably, among the three individual modules, the GAT models incorporating 3D structural information generally outperformed the sequence-based transformer model for both fungal and oomycete effector protein prediction tasks. This result underscores the importance of leveraging 3D structural information in effector protein prediction. Interestingly, the integrated LR model achieved highly accurate predictive performance and outperformed all individual predictors, demonstrating the effectiveness of our model integration strategy. In particular, the AUPRC score reached 0.929 for fungi and 0.931 for oomycetes in the final integrative models, highlighting the advantage of combining complementary sequence and structural information through multiple predictors. Additionally, the results of five-fold cross-validation are listed in Table S1, which are in line with the independent test results.

To investigate the model's interpretability in distinguishing fungal/oomycete effector proteins from non-effector proteins, we employed *t*-distributed stochastic neighbor embedding (*t*-SNE) to perform dimensionality reduction and visualize the high-dimensional embedding vectors generated from different models. Briefly, the sequence embeddings generated after processing by the transformer layer in the sequence-based module, and two structure-based representations after processing the GAT layer in the structure-based modules, were used for *t*-SNE analysis. As shown in Figure 2a–c, all three embedding vectors of fungal effectors/non-effectors reveal clear clustering patterns in the two-dimensional space, intuitively indicating their ability to distinguish effectors from non-effectors. Generally, the embeddings generated by GATs reveal more discriminative features than sequence-based embeddings. Comparatively, the GAT-based ESM embedding seems to yield the best clustering result, which agrees with the performance of the three individual predictors. Likewise, the clustering patterns are also observed in the oomycete effector dataset (Figure 2d–f). Here, the GAT-based SaProt embeddings appear to be the most informative, which explains the performance rank of the three individual modules in predicting oomycete effectors.

Ablation studies

For the sequence-based effector prediction task model, we followed the DeepSecE framework, which utilizes feature vectors generated by the pretrained model as input to a transformer module for feature extraction and learning. Here, the pretrained models employed were ESM1b, ESM2, and ProtTrans (specifically ProtT5). In addition, fine-tuning is a frequently used strategy to adapt the pretrained models to task-specific training. Given that ESM combined with transformer consistently outperformed ProtT5, we further implemented a lightweight fine-tuning scheme on

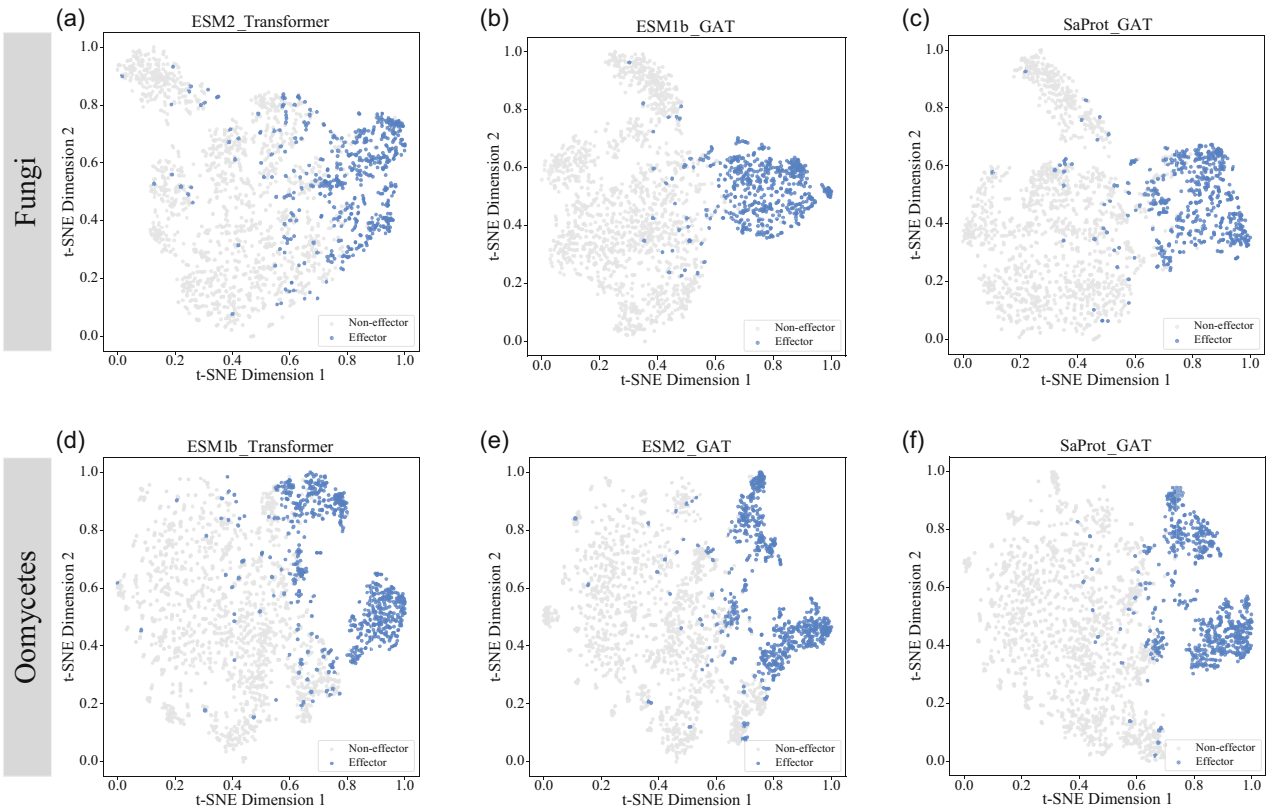


Figure 2. *t*-Distributed stochastic neighbor embedding (*t*-SNE) visualizations of protein embeddings for effector and non-effector learned from training data. Panels (a–c) plot the *t*-SNE projections of fungal effector and non-effector embeddings. (a) shows ESM2 embeddings processed through a transformer layer, while (b, c) display ESM1b and SaProt embeddings, respectively, processed through GAT layers. Panels (d–f) plot the *t*-SNE projections of oomycete effector and non-effector protein embeddings. (d) shows ESM1b embeddings processed through a transformer layer, while (e, f) display ESM2 and SaProt embeddings, respectively, processed through GAT layers.

the ESM models and systematically compared their performance with the corresponding transformer-based counterparts. Briefly, all layers of the ESM models were initially frozen to preserve the generalizable representations learned from large-scale protein sequence data. Then, only the final layer responsible for representation extraction was unfrozen, allowing it to be updated during task-specific training. Furthermore, we attached a trainable MLP classifier to the extracted sequence representations to perform additional feature transformation and generate the final predictions.

Regarding the fungal effector prediction, the ESM2_Transformer combination achieved the highest performance on the independent test set, with an AUPRC of 0.863 (AUROC = 0.934), outperforming ESM1b_Transformer (AUPRC = 0.859 and AUROC = 0.929) and ProtT5_Transformer (AUPRC = 0.833 and AUROC = 0.923) (Figure 3a,b). Other metrics also indicated improvements of ESM2_Transformer over ESM1b_Transformer and ProtT5_Transformer (Table S2). Notably, fine-tuned ESM1b and ESM2 models also yielded excellent performance but were lower than the corresponding transformer models, indicating that pretraining alone is informative in predicting fungal effectors (Table S2). For the oomycete effector protein prediction task, sequence-based models exhibited a different trend. The ESM1b_Transformer model outperformed other combinations, achieving the highest AUPRC of 0.828, which was superior to ESM2_Transformer (AUPRC = 0.823) and ProtT5_Transformer (AUPRC = 0.797) (Figure 4a,b). Other performance metrics confirmed this advantage (Table S2). In contrast to the prediction task of fungal effectors, pre-trained transformer models outperformed the corresponding fine-tuned models considerably (Table S2).

For the structure-based GAT models, we employed various pre-trained models to generate embedding vectors as node features for the residue contact networks. The selected pre-trained models included ESM1b, ESM2, ProtTrans (specifically ProtT5), and SaProt. For each pre-trained model, we determined the optimal distance threshold for constructing the residue contact networks under fivefold cross-validation. Detailed results are presented in Tables S3 and S4. Based on the optimal distance thresholds, we constructed residue contact networks and evaluated model performance on the independent test

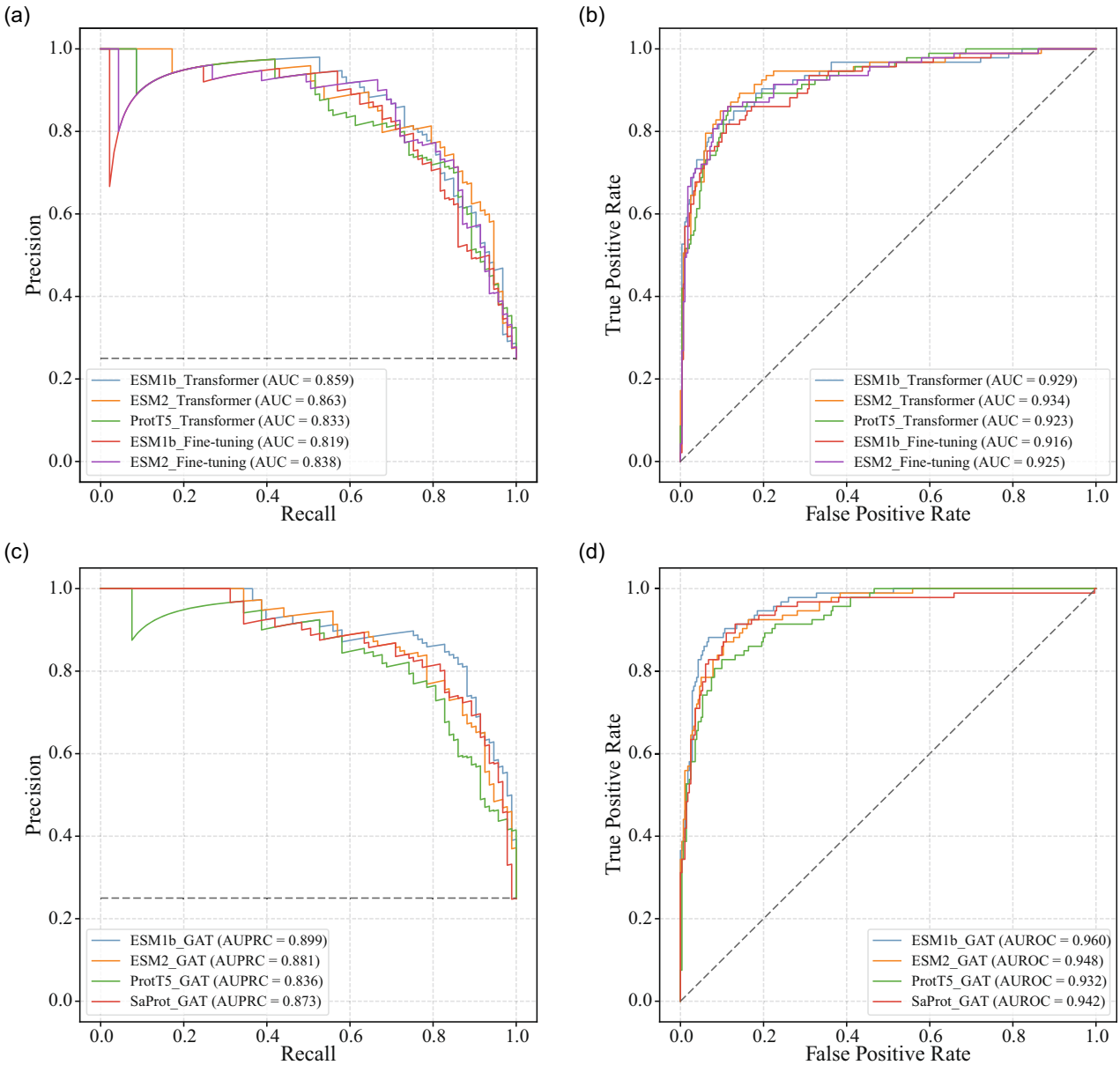


Figure 3. Predictive performance of different models for fungal effector prediction in ablation experiments.
(a, b) Precision-recall (PR) and receiver operating characteristic (ROC) curves for fungal effector prediction using sequence-based features extracted by the transformer-based model.
(c, d) PR and ROC curves for fungal effector prediction using different embeddings processed by GAT-based models. Dash lines stand for random predictions.

set. Regarding the prediction of fungal effectors, the ESM1b_GAT combination achieved the best performance (AUPRC = 0.899 and AUROC = 0.960), followed by ESM2 (AUPRC = 0.881 and AUROC = 0.948), SaProt (AUPRC = 0.873 and AUROC = 0.942), and ProtT5 (AUPRC = 0.836 and AUROC = 0.932) (Figure 3c,d; Table S5). Regarding the oomycete effector prediction, the SaProt_GAT model achieved the best performance with an AUPRC of 0.913, outperforming ESM2 (AUPRC = 0.804), ESM1b (AUPRC = 0.781), and ProtT5 (AUPRC = 0.699) (Figure 4c,d; Table S5).

Considering that the concatenation of different features into a larger feature vector is also a popular strategy in feature engineering, we also evaluated a single GAT model with concatenated ESM and SaProt embeddings as node features. However, this new GAT model only yielded intermediate performance. It outperformed SaProt_GAT but underperformed ESM1b_GAT in predicting fungal effectors (Table S5), while it exceeded ESM2_GAT but was inferior to SaProt_GAT for oomycete effector prediction (Table S5). Therefore, our choice of dual GAT architecture rather than a single model integrating two embeddings was justified.

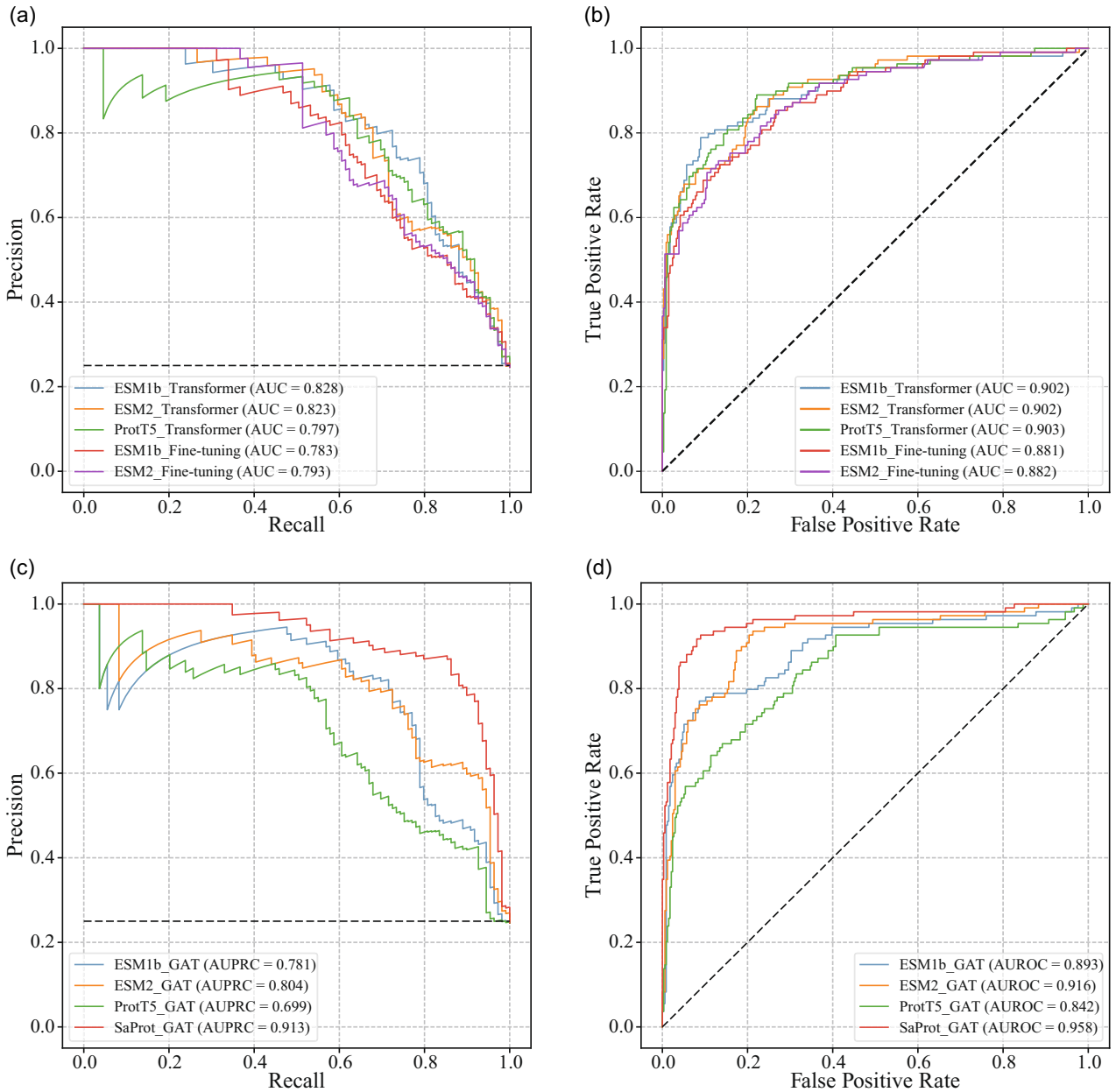


Figure 4. Predictive performance of different models for oomycete effector prediction in ablation experiments.
(a, b) Precision-recall (PR) and receiver operating characteristic (ROC) curves for oomycete effector prediction using sequence-based features extracted by the transformer-based model.
(c, d) PR and ROC curves for oomycete effector prediction using different embeddings processed by GAT-based models under different structural encoding strategies. Dash lines stand for random predictions.

To maximize the prediction accuracy, we integrated three individual predictors via LR, and achieved a reasonable improvement in overall predictive performance compared to the corresponding best individual model alone (Figure 1b–e; Table 1). Specifically, an approximately 3% AUPRC increase was observed for fungi, while an approximately 2% AUPRC increase was observed for oomycetes. Likewise, the final PPEPFinder yielded more powerful performance than the best combination of two single models (ESM2_Transformer+ESM1b_GAT for fungi and ESM1b_Transformer+SaProt_GAT for oomycetes), further demonstrating that these three individual predictors are complementary to some extent.

Additionally, we investigated the effect of protein sequence length on model performance. Due to computational resource constraints, this analysis was limited to the

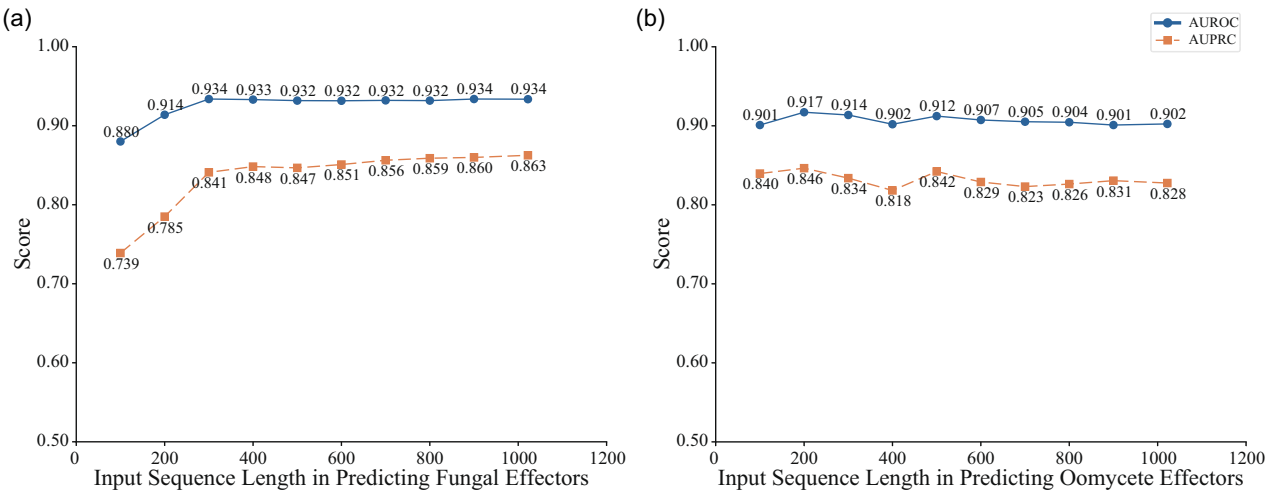


Figure 5. Impact of input sequence length on transformer-based model performance.
The fungal (a) and oomycete (b) effector prediction performance based on different input sequence lengths was evaluated through AUPRC and AUROC values.

transformer-based sequence model. In the fungal effector protein prediction task, model performance steadily improved with increasing protein sequence length (Figure 5a). This trend likely reflects the benefit of incorporating longer sequence information, which provides richer feature information and enables the model to achieve better predictive performance (i.e., higher AUPRC). In the oomycete effector protein prediction task, a maximal performance judged by AUPRC was observed when the input sequence length reached 200 residues (Figure 5b). This finding is consistent with the biological characteristics of many oomycete effector proteins, which typically possess an N-terminal signal peptide for extracellular secretion (Whisson et al., 2007).

Comparison of PPEPFinder against baseline models, existing predictors, and BLAST searching

We first compared PPEPFinder against traditional machine learning approaches based on protein sequences, with all models evaluated on the independent test set. Specifically, we employed common sequence encoding schemes, including dipeptide composition (DPC) (Zhou et al., 2012), conjoint triad (CT) (Shen et al., 2007), and composition of K-spaced amino acid pairs (CKSAAP) (Chen et al., 2018), combined with two conventional classifiers (SVM and RF), to construct multiple baseline models. As shown in Table S6, the sequence-based transformer model in PPEPFinder substantially outperformed these traditional machine learning methods across all evaluation metrics, highlighting the advantage of leveraging pre-trained PLMs with transformer architectures for effector prediction.

To further assess the generalization ability of our model, we manually curated a set of experimentally validated fungal and oomycete effector proteins from recent literature published within the past 2 years. Then, we compiled two additional independent test sets (i.e., Fungi_additional_dataset and Oomycetes_additional_dataset) to compare the performance of PPEPFinder against several existing prediction methods. More details about the preparations of these two additional test sets are available in the Materials and Methods section.

In this study, we benchmarked four state-of-the-art fungal effector protein prediction tools, including EffectorP3.0/EffectorP3.0-fungi (https://effectorp.csiro.au/), Fungtion (https://step3.erc.monash.edu/Fungtion/), and Deepredeff (https://github.com/TeamMacLean/ruth-effectors-prediction). These tools were compared with our PPEPFinder model using Fungi_additional_dataset, which comprises 54 fungal effector proteins and 162 non-effector proteins. We uploaded the test dataset to the web servers of EffectorP3.0, EffectorP3.0-fungi, and Fungtion to obtain the prediction results. Regarding Deepredeff, we downloaded the best model for fungal effector prediction (i.e., the CNN-LSTM model) and ran it locally. As shown in Figure 6a, PPEPFinder outperformed existing fungal effector protein prediction methods, revealing the highest AUPRC and AUROC scores. Probably due to the introduction of the ESM1b embedding as an input feature, it should be mentioned that Fungtion also performs excellently, second only to PPEPFinder. EffectorP3.0 is a popular machine learning-based tool for predicting fungal and oomycete effectors, and it achieved a reasonably good performance. It is also interesting to note that EffectorP3.0-fungi outperforms EffectorP3.0 (AUPRC = 0.754 versus 0.721), as it is specially tailored for fungal effectors. Comparatively, the performance of Deepredeff remained limited, with an AUPRC of 0.380 and an AUROC of 0.597.

We further compared the performance of the PPEPFinder model on Oomycetes_additonal_dataset (72 positive

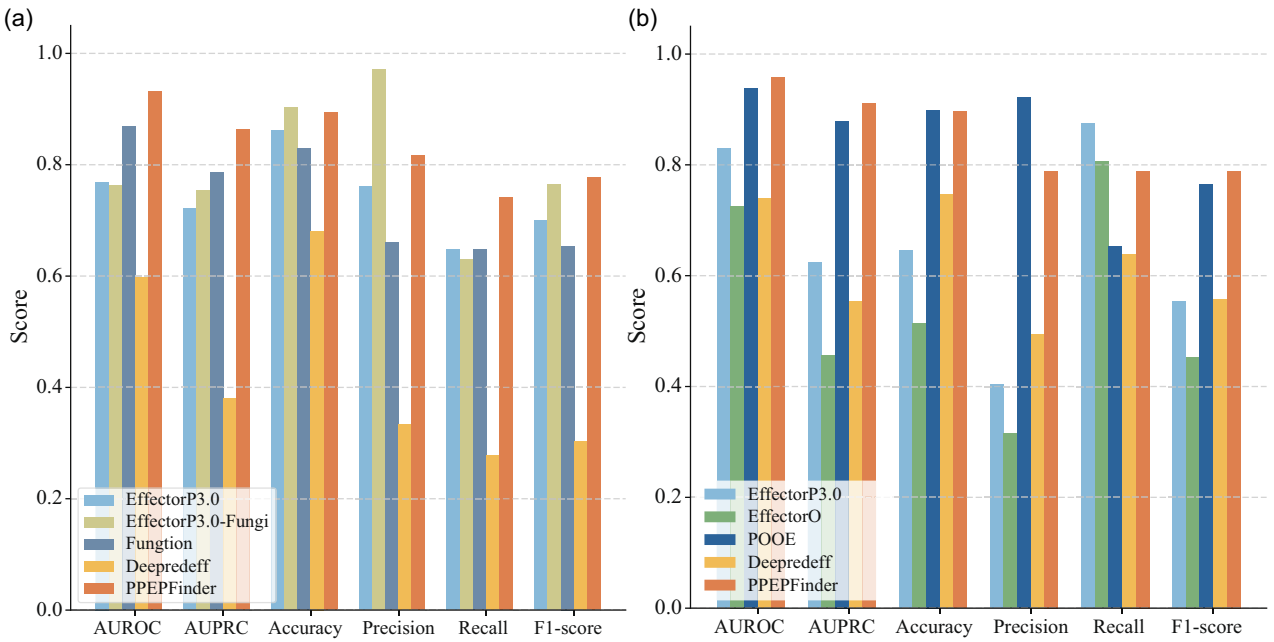


Figure 6. Performance comparison of Plant Pathogen Effector Protein Finder against existing methods.
(a) Fungal effector predictions; (b) oomycete effector predictions.

samples and 216 negative samples) with four existing oomycete effector protein prediction methods, including EffectorP3.0, EffectorO (https://bremia.ucdavis.edu/effectorO.php), POOE (http://zzdlab.com/pooe/), and Deepredeff. Similarly, the test data were uploaded to the web servers of EffectorP3.0, EffectorO, and POOE. For the Deepredeff predictor, we chose the CNN-LSTM model again, which was reported to be more competent in predicting oomycete effectors. Again, PPEPFinder demonstrated the best performance (Figure 6b), followed by POOE, EffectorP3.0, Deepredeff, and EffectorO.

Considering that fungal and oomycete effectors may share sequence similarity or contain specific sequence motifs, classical sequence-based searches have been widely used to identify potential novel effectors in practical applications. Therefore, it is natural to compare PPEPFinder with this conventional effector detection strategy. We compared PPEPFinder with BLAST (Altschul et al., 1997), one of the most commonly used sequence similarity search tools. To this end, all the effector sequences from the training data (fungi or oomycetes) were used to construct a sequence searching database, while sequences from the corresponding additional independent test set served as query proteins. For each query, the top hit was retained, and the corresponding *E*-value was defined as the BLAST prediction score. By default, a query was predicted as an effector if the *E*-value was less than 1.0×10^{-3}; otherwise, it was predicted as a non-effector. The results showed that BLAST achieved moderate performance but was still considerably inferior to PPEPFinder (Table S7).

Together, the above comparison results demonstrate that PPEPFinder consistently outperforms routine baseline methods, existing state-of-the-art predictors, and BLAST searching, highlighting the effectiveness of integrating pre-trained PLMs with transformer architectures and GAT for accurate effector prediction.

Online prediction platform

To facilitate the research community, we have developed an online web server implementing the PPEPFinder model, which is freely accessible at http://zzdlab.com/PPEPFinder/. The web server is deployed on CentOS 7.4 with Apache 2.4.6. To run the web server, users first select the pathogen type they wish to predict (i.e., fungus or oomycete). To explore the full functionality of PPEPFinder, users are required to provide the sequence and the 3D structure of the query protein in PDB format. Due to our limited computational resources, we are not able to integrate the ESMFold structure prediction into our web server. To facilitate the users, we provide an online link to ESMFold for structure prediction; however, please note that it can only predict proteins shorter than 400 residues. For users who wish to predict structures on their local machines, they should download the prediction code from ESMFold's GitHub repository (https://github.com/facebookresearch/esm). Alternatively, predictions can be made solely using the sequence-based transformer model, which is particularly useful for processing multiple query proteins. Query proteins with a prediction score greater than the default prediction threshold value (i.e., 0.5) are classified as

effectors. According to the benchmark experiment on the additional independent test datasets, the corresponding precision values of PPEPFinder for predicting fungal and oomycete effectors are 0.816 (recall = 0.741) and 0.781 (recall = 0.792), respectively. Comparatively, the solely sequence-based transformer model achieved a precision value of 0.771 (recall = 0.685) and 0.713 (recall = 0.792) for predicting fungal and oomycete effectors, respectively.

PPEPFinder offers three modes for users to retrieve their submitted tasks: by job name, job ID, or submission date. To enable this, users must assign a job name when submitting their sequences. Once the task is completed, the result page will display the raw prediction scores and classification result (i.e., whether the protein is predicted to be a fungal or oomycete effector). Additionally, the source code and datasets used in this study are freely available for download from our web server or the GitHub repository (https://github.com/mdiyuan/PPEPFinder).

It is important to note that PPEPFinder does not explicitly predict or report specific effector-related characteristics such as RXLR motifs, WY domains, Cys-rich regions, or the presence of a signal peptide. Although these features were represented in the training dataset and may implicitly contribute to the model's learning process, they are not individually assessed during prediction. Users who wish to include these characteristics in their analyses may pre-screen their protein sequences using existing tools prior to submission, or examine the predicted effector sequences afterward with specialized bioinformatics programs.

CONCLUSION

In this study, we present PPEPFinder, a novel computational framework for predicting fungal and oomycete effector proteins. It should be emphasized that PPEPFinder seamlessly combines the strengths of a transformer-based model utilizing protein sequence embeddings with two GAT models that leverage protein structural information, achieving a highly accurate predictive performance. Rigorous benchmarking experiments demonstrate that PPEPFinder considerably outperforms existing state-of-the-art effector prediction methods. To facilitate the research community, we have also developed a publicly accessible web server. We believe that PPEPFinder will serve as a valuable tool to accelerate the discovery of novel effector proteins and advance our understanding of plant–pathogen biology.

MATERIALS AND METHODS

Dataset collection and preparation

Fungal effector dataset

Fungal effector proteins (i.e., positive samples) were derived from the experimentally validated effector sequences that were originally used as training sets in previous fungal effector prediction tools. Specifically, we compiled positive samples from EffectorP 2.0, EffHunter, FunEffector-Pred, Predector, and WideEffHunter (Carreón-Anguiano et al., 2020, 2022; Jones et al., 2021; Sperschneider et al., 2018; Wang et al., 2020). Additionally, 1101 experimentally validated fungal effector sequences published before December 31, 2022, were manually curated from the literature. To eliminate redundancy, all sequences were clustered using CD-HIT (Li & Godzik, 2006), resulting in a final positive dataset containing 469 non-redundant effector sequences (see Table S8 for species distribution). Previous studies have reported that canonical effector characteristics include protein length ≤400 amino acids and cysteine-rich content (≥4 cysteines) (Carreón-Anguiano et al., 2022). In our dataset, the average sequence length of fungal effectors is 290, and approximately 70% of effectors contained at least four cysteine residues, which is in line with the known sequence patterns of fungal effectors.

Negative samples for fungal effectors were derived from three primary sources: (1) secreted fungal proteins from the UniProt database (UniProt Consortium, 2025), (2) secreted proteins from saprophytic fungi that are non-pathogenic to plants, and (3) fungal proteins annotated in the PHI-base database as having 'unaffected pathogenicity' (Urban et al., 2020). All candidate negative sequences were also clustered using CD-HIT (Li & Godzik, 2006) with a 40% identity threshold. Sequences were retained only when predicted by SignalP 5.0 (Almagro Armenteros et al., 2019) to contain a signal peptide and lacked transmembrane domains according to TMHMM 2.0 (Krogh et al., 2001). To further minimize potential bias, proteins sharing more than 95% identity with any positive sample were excluded. Additionally, potential false-negatives were filtered out using EffectorP 3.0 (Sperschneider & Dodds, 2022), resulting in a high-confidence negative set containing 5647 non-effector fungal proteins.

In total, the dataset (i.e., Fungi_dataset) consisted of 469 positive and 5647 negative samples. To address the imbalance between positive and negative samples, 1407 negative sequences were randomly selected to ensure a 1:3 ratio of positive to negative samples. This dataset was then randomly divided into a training set (80%) and an independent test set (20%). Specifically, the training set consisted of 375 positive and 1125 negative samples, while the test set comprised 94 positive and 282 negative samples.

Oomycete effector dataset

The positive/negative samples of oomycete effector proteins used in this study were obtained from our previous work on developing an oomycete effector predictor called POOE (Zhao et al., 2024). Briefly, positive sequences were collected from literature published before January 1, 2022, and redundancy was removed using a 40% identity threshold. Furthermore, species with fewer than 10 effectors were removed, and the retained positive samples contained 549 effector sequences from eight species (Table S8). Likewise, the known sequence patterns of oomycete effectors were observed in these positive samples. For instance, we searched the first 110 amino acids of each effector sequence and identified 266 sequences containing the canonical RXLR motif. Furthermore, HMM-based domain searches revealed the presence of WY domains in 65 oomycete effectors.

Negative samples were constructed from the full proteomes of the same species. Sequences annotated as 'effector' or containing canonical effector motifs were removed. Similar to the negative sample construction in Fungi_dataset, redundant sequences (>40% identity) and sequences highly similar (>95%) to positive samples were excluded. Secreted proteins and transmembrane proteins were further filtered out. Negative samples were then randomly selected at a 1:3 ratio relative to positive samples,

resulting in a dataset of 1670 non-effectors. This dataset (i.e., Oomycetes_dataset) was subsequently divided into a training set (80%) and an independent test set (20%), resulting in a training set containing 439 effectors and 1336 non-effectors proteins, and a test set consisting of 110 effectors and 334 non-effectors.

Ideally, the negative samples should cover diverse sequence and structural space to reduce potential biases in the proposed models. Here, we analyzed the sequence and structural similarities of negative samples within each species for fungi and oomycetes. Pairwise sequence identities were calculated using MAFFT (Nakamura et al., 2018), and structural similarities were assessed with TM-align (Zhang & Skolnick, 2005) on protein structures predicted by ESMFold. We calculated the sequence identity and TM-score of within-species protein pairs in fungi and oomycetes. The results showed that 73.5% of fungal pairs had sequence identity <0.3 and TM-score <0.5, while 79.1% of oomycete pairs fell into this low-similarity region. These results indicate the negative samples cover a broad spectrum of protein sequences, structures, and functions, ensuring the robustness and generalizability of the established predictive models.

Additional independent test sets

To further assess the generalization capability of our model and benchmark its performance against existing prediction tools, we manually assembled additional independent test sets comprising newly reported, experimentally validated effector proteins from fungi and oomycetes.

For the fungal dataset, 54 newly reported effector proteins were manually collected from publications published between January 1, 2023 and December 31, 2024. The oomycete dataset included 72 newly reported effectors extracted from recent literature published during the same period. Negative samples for both fungal and oomycete additional test sets were selected using the same criteria as those used for the training dataset. Negative samples were randomly selected at a 1:3 positive-to-negative ratio to ensure consistency. As a result, two additional independent test sets were constructed, including Fungi_addtional_dataset (54 positive and 162 negative samples) and Oomycetes_additional_dataset (72 positive and 216 negative samples).

Protein structure prediction

In this work, ESMFold (Lin et al., 2023) was employed to predict the 3D structures of effector proteins from the fungal and oomycete datasets. Compared to AlphaFold, ESMFold offers a substantial advantage by eliminating the requirement for multiple sequence alignment searches, thereby significantly reducing computational cost and inference time. We employed the official GPU-accelerated Linux version of ESMFold. We calculated the pLDDT values for all predicted fungal and oomycete protein structures. As shown in Figure S1, the predicted structures exhibited high pLDDT values (e.g., approximately half of the fungal/oomycete protein structures with a pLDDT >70), indicating the reasonable accuracy of structure prediction. To maximize the training datasets, we did not apply any filtering cutoffs based on pLDDT.

Protein embedding extraction using PLM

Embedding extraction with ESM1b

ESM1b (https://github.com/facebookresearch/esm) (Rives et al., 2021) is a transformer-based PLM developed by Facebook AI Research, which was pre-trained in an unsupervised manner on the UniRef50 database. In this study, we utilized ESM1b to generate 1280-dimensional sequence embeddings for effector proteins. These embeddings served as input features in two distinct predictive models within the PPEPFinder framework. For the structure-based fungal effector prediction model, the ESM1b embeddings were used as node features in residue contact networks. In parallel, for the sequence-based oomycete effector prediction model, the same embeddings were fed into a transformer architecture to learn discriminative sequence patterns.

Embedding extraction with ESM2

ESM2 (Lin et al., 2023) is an upgraded version of the ESM series, built upon an improved transformer architecture with a significantly larger model scale. In particular, when integrated with structural prediction frameworks such as ESMFold, ESM2 enables end-to-end prediction of protein 3D structures directly from sequences. This model provides a powerful and generalizable representation tool for protein research and bioinformatics applications. In this study, we employed the esm2_t33_650M_UR50D model to generate 1280-dimensional protein sequence embeddings, which were subsequently used as input features for downstream predictive tasks. Within the PPEPFinder framework, these embeddings served two distinct roles: they were used as sequence-level input features in the transformer-based model for oomycete effector prediction, and simultaneously, they were incorporated as node features in the residue contact networks of the structure-based GAT model for fungal effector prediction.

Embedding extraction with ProtTrans

ProtTrans (Elnaggar et al., 2022) consists of a series of pretrained PLMs, which were trained on 393 billion residues from UniRef and BFD (Big Fantastic Database). Among these models, we employed the ProtT5-XL-UniRef50 autoencoder model (ProtT5), which was downloaded from https://huggingface.co/Rostlab/ProtT5 and configured for local deployment. ProtT5 generated 1024-dimensional embedding vectors for each protein. In the sequence-based model, these embeddings were fed into the transformer. In the structure-based model, the embeddings were used as node features in the residue contact network, which were then input to the GAT.

Embedding extraction with SaProt

SaProt (Su et al., 2024) is a large-scale, structure-aware PLM specifically designed to incorporate structural information into protein sequence representations. SaProt produces 1280-dimensional embeddings that capture rich structural and contextual features. In this study, the input protein 3D structures in PDB format were first converted into structure-aware sequences by invoking the Foldseek binary for structural tokenization. The resulting sequences were then processed using the SaProt_650M_AF2 model to generate protein embeddings, which were subsequently employed as node features within the residue contact network for downstream tasks.

Transformer-based predictor

In this study, inspired by the DeepSecE model (Zhang et al., 2023), we integrated a transformer-based framework built upon PLMs into PPEPFinder to enhance the identification of effector proteins (Figure 1a). Specifically, the model employs ESM as a pre-trained language model module to extract initial embedding representations from input protein sequences. The input sequences are truncated to no more than 1022 amino acid residues and encoded as

token sequences. Both ESM1b and ESM2 generate 1280-dimensional context-aware embedding vectors for each residue through their 33-layer stacked transformer architecture.

To improve training efficiency and reduce computational cost, a one-dimensional convolutional layer is introduced after the pre-trained model to compress the residue-level embedding dimension from 1280 to 256. To further enhance the model's ability to capture secretion-related features potentially associated with effector proteins, PPEPFinder incorporates an additional transformer module consisting of a single transformer layer with four attention heads. Finally, a fully connected layer is used to classify each protein sequence as an effector or a non-effector.

Two GAT-based predictors

As shown in Figure 1a, two structure-based GAT models first rely on spatial information from predicted protein structures to construct residue contact graphs. In the graphs, a node represents residue-level features, while an edge represents spatial relationships between residues. Then the two GAT models utilize different node feature representations: one employs ESM embeddings derived from protein sequences, while the other uses SaProt embeddings generated from protein structures. By leveraging node features and edge relationships through a self-attention mechanism, the GAT models dynamically adjust the influence of neighboring nodes, thereby learning weighted and context-aware representations for each node. In this architecture, the representation of each node is computed based not only on its own features but also on those of its neighboring nodes. Attention coefficients determine the relative contribution of each neighboring node to the current node's representation. A fully connected layer is used at the final stage to predict whether the input protein is an effector.

In the constructed contact graph, spatial relationships are defined based on the distances between Cα atoms. If the distance between two Cα atoms falls below a predefined threshold, they are considered to be in contact. This contact information is used to construct the adjacency matrix, which is then utilized by the GAT model to learn both local structural and graph-level features. A multi-head attention mechanism is applied to enhance the representational capacity of node embeddings. The learned node representations are aggregated using global average pooling to obtain graph-level feature vectors, which are subsequently passed to an output layer for final effector protein prediction.

Fivefold cross-validation was used to train the GAT models for fungal and oomycete effector prediction. The batch size was set to 32, and the maximum number of epochs was set to 500. The network was configured with 1 GAT layer, 10 hidden units per layer, 6 attention heads, and a learning rate of 1×10^{-4}. Early stopping with a patience of 50 epochs was applied to prevent overfitting, and the best-performing model weights were restored for evaluation. The model was implemented using TensorFlow 2.13.1 with the Keras API (CUDA version: 12.4, NVIDIA GeForce RTX 3090).

LR model

To further improve predictive performance, an LR model is employed to perform weighted integration of prediction scores generated by multiple independent models, thereby producing a final composite prediction score. Specifically, the prediction scores generated from three individual models (i.e., $S_{ESM\_Transformer}$, S_{ESM_GAT}, and S_{SaProt_GAT}) are used as input features to train the LR model, which is implemented using the scikit-learn library in Python (https://scikit learn.org) (Pedregosa et al., 2011). To prevent overfitting and ensure robust generalization performance, hyperparameter optimization for the regularization strength parameter was performed using the GridSearchCV function combined with fivefold cross-validation.

Baseline models

The baseline models were constructed by combining traditional sequence-based encoding schemes with machine learning algorithms. Three sequence encoding methods (i.e., DPC, CT, and CKSAAP) were taken into account. DPC (Zhou et al., 2012) represents the composition of two consecutive amino acids (i.e., dipeptides) across the entire protein sequence and transforms a protein into a 400-dimensional vector. CT (Shen et al., 2007) stands for the frequencies of three consecutive amino acid groups. Regarding the CT encoding, the 20 amino acids are categorized into seven groups (AGV, C, DE, FILP, HNQW, KR, and MSTY) based on their side-chain physicochemical properties, and the frequencies of three consecutive amino acid groups are computed to yield a 343-dimensional ($7 \times 7 \times 7$) vector. CKSAAP (Chen et al., 2018) encodes protein sequences by calculating the frequencies of amino acid pairs separated by k residues, with k set to 3 in this study. For classification, SVM and RF models were also implemented using the scikit-learn package. Hyperparameters were optimized using GridSearchCV, which performs an exhaustive grid search with five-fold cross-validation on the training data. Specifically, the parameters C, gamma, and the kernel function were tuned for SVM, while the parameters n_estimators, min_samples_split, min_samples_leaf, max_features, and max_depth were optimized in the model training of RF.

Performance evaluation

In both fivefold cross-validation and independent testing, the classification performance of the model was evaluated using several standard metrics, including accuracy, precision, recall, and F1-score. These evaluation metrics are defined as follows:

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

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

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

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

where TP, FP, TN, and FN represent the numbers of true-positives, false-positives, true-negatives, and false-negatives, respectively. To provide a more comprehensive assessment of model performance, we plotted the ROC curve and the PR curve, and computed their respective area under the curve values, namely AUROC and AUPRC. In general, the closer the AUROC and AUPRC values are to 1, the better the model's performance. Notably, for imbalanced datasets, the PR curve and its associated AUPRC value typically provide a more accurate reflection of a model's predictive ability.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest associated with this work.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (31271414, 31471249, and 31970645).

DATA AVAILABILITY STATEMENT

All data used in this study were manually collected from published literature. The datasets used for model training and evaluation are available for download from the PPEP-Finder (http://zzdlab.com/PPEPFinder/) and our GitHub repository (https://github.com/mdiyuan/PPEPFinder).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Distribution of pLDDT values of protein structures predicted by ESMFold.

Table S1. Overall performance of our models via fivefold cross-validation for fungal and oomycete effector prediction.

Table S2. Performance of transformer- and fine-tuning-based predictors on the independent test set.

Table S3. Evaluation of distance thresholds for residue contact networks with various pre-trained embeddings for fungal effector prediction.

Table S4. Evaluation of distance thresholds for residue contact networks with various pre-trained embeddings for oomycete effector prediction.

Table S5. Fungal and oomycete effector prediction performance of different pre-trained models on the independent test set using the optimal residue contact network.

Table S6. Performance of traditional sequence-based machine learning methods on the independent test set.

Table S7. Performance comparison of PPEPFinder and existing fungal effector predictors on Fungi_additional_dataset and Oomycetes_additional_dataset.

Table S8. Species distribution of fungal and oomycete effector proteins.

REFERENCES

Abramson, J., **Adler, J.**, **Dunger, J.**, **Evans, R.**, **Green, T.**, **Pritzel, A.** ***et al.*** (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, **630**(8016), 493–500. Available from: https://doi.org/10.1038/s41586-024-07487-w

Almagro Armenteros, J.J., **Tsirigos, K.D.**, **Sønderby, C.K.**, **Petersen, T.N.**, **Winther, O.**, **Brunak, S.** ***et al.*** (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology*, **37**(4), 420–423. Available from: https://doi.org/10.1038/s41587-019-0036-z

Altschul, S.F., **Madden, T.L.**, **Schäffer, A.A.**, **Zhang, J.**, **Zhang, Z.**, **Miller, W.** ***et al.*** (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402. Available from: https://doi.org/10.1093/nar/25.17.3389

Carreón-Anguiano, K.G., **Islas-Flores, I.**, **Vega-Arreguín, J.**, **Sáenz-Carbonell, L.** & **Canto-Canché, B.** (2020) EffHunter: a tool for prediction of effector protein candidates in fungal proteomic databases. *Biomolecules*, **10**(5), 712. Available from: https://doi.org/10.3390/biom10050712

Carreón-Anguiano, K.G., **Todd, J.N.A.**, **Chi-Manzanero, B.H.**, **Couoh-Dzul, O.J.**, **Islas-Flores, I.** & **Canto-Canché, B.** (2022) WideEffHunter: an algorithm to predict canonical and non-canonical effectors in fungi and oomycetes. *International Journal of Molecular Sciences*, **23**(21), 13567. Available from: https://doi.org/10.3390/ijms232113567

Chen, Z., **Ji, M.** & **Qi, Y.** (2023) Advances in artificial intelligence algorithms for protein structure design. *Synthetic Biology*, **4**(3), 464–487. Available from: https://doi.org/10.12211/2096-8280.2023-008

Chen, Z., **Zhao, P.**, **Li, F.**, **Leier, A.**, **Marquez-Lago, T.T.**, **Wang, Y.** ***et al.*** (2018) iFeature: A Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics (Oxford, England)*, **34**(14), 2499–2502. Available from: https://doi.org/10.1093/bioinformatics/bty140

Dong, X., **Lu, X.** & **Zhang, Z.** (2015) BEAN 2.0: an integrated web resource for the identification and functional analysis of type III secreted effectors. *Database*, **2015**, bav064. Available from: https://doi.org/10.1093/database/bav064

Eichinger, V., **Nussbaumer, T.**, **Platzer, A.**, **Jehl, M.A.**, **Arnold, R.** & **Rattei, T.** (2016) EffectiveDB--updates and novel features for a better annotation of bacterial secreted proteins and type III, IV, VI secretion systems. *Nucleic Acids Research*, **44**(D1), D669–D674. Available from: https://doi.org/10.1093/nar/gkv1269

Ellis, J.G. & **Dodds, P.N.** (2011) Showdown at the RXLR motif: serious differences of opinion in how effector proteins from filamentous eukaryotic pathogens enter plant cells. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(35), 14381–14382. Available from: https://doi.org/10.1073/pnas.1111668108

Elnaggar, A., **Heinzinger, M.**, **Dallago, C.**, **Rehawi, G.**, **Wang, Y.**, **Jones, L.** ***et al.*** (2022) ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**(10), 7112–7127. Available from: https://doi.org/10.1109/tpami.2021.3095381

Govers, F. & **Bouwmeester, K.** (2008) Effector trafficking: RXLR-dEER as extra gear for delivery into plant cells. *Plant Cell*, **20**(7), 1728–1730. Available from: https://doi.org/10.1105/tpc.108.062075

Heinzinger, M., **Elnaggar, A.**, **Wang, Y.**, **Dallago, C.**, **Nechaev, D.**, **Matthes, F.** ***et al.*** (2019) Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, **20**(1), 723. Available from: https://doi.org/10.1186/s12859-019-3220-8

Huang, Y., **Wuchty, S.**, **Zhou, Y.** & **Zhang, Z.** (2023) SGPPI: structure-aware prediction of protein-protein interactions in rigorous conditions with graph convolutional network. *Briefings in Bioinformatics*, **24**(2), bbad020. Available from: https://doi.org/10.1093/bib/bbad020

Hui, X., **Chen, Z.**, **Lin, M.**, **Zhang, J.**, **Hu, Y.**, **Zeng, Y.** ***et al.*** (2020) T3SEpp: an integrated prediction pipeline for bacterial type III secreted effectors. *MSystems*, **5**(4), e00288-20. Available from: https://doi.org/10.1128/mSystems.00288-20

Jones, D.A.B., **Rozano, L.**, **Debler, J.W.**, **Mancera, R.L.**, **Moolhuijzen, P.M.** & **Hane, J.K.** (2021) An automated and combinative method for the predictive ranking of candidate effector proteins of fungal plant pathogens. *Scientific Reports*, **11**(1), 19731. Available from: https://doi.org/10.1038/s41598-021-99363-0

Kang, Y., **Elofsson, A.**, **Jiang, Y.**, **Huang, W.**, **Yu, M.** & **Li, Z.** (2023) AFTGAN: prediction of multi-type PPI based on attention free transformer and graph attention network. *Bioinformatics*, **39**(2), btad052. Available from: https://doi.org/10.1093/bioinformatics/btad052

Kristianingsih, R. & **MacLean, D.** (2021) Accurate plant pathogen effector protein classification ab initio with Deepredeff: an ensemble of convolutional neural networks. *BMC Bioinformatics*, **22**(1), 372. Available from: https://doi.org/10.1186/s12859-021-04293-3

Krogh, A., **Larsson, B.**, **von Heijne, G.** & **Sonnhammer, E.L.** (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, **305**(3), 567–580. Available from: https://doi.org/10.1006/jmbi.2000.4315

Li, J., **Ren, J.**, **Dai, W.**, **Stubenrauch, C.**, **Finn, R.D.** & **Wang, J.** (2024) Fungtion: a server for predicting and visualizing fungal effector proteins. *Journal of Molecular Biology*, **436**(17), 168613. Available from: https://doi.org/10.1016/j.jmb.2024.168613

Li, W. & **Godzik, A.** (2006) CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**(13), 1658–1659. Available from: https://doi.org/10.1093/bioinformatics/btl158

Lin, Z., **Akin, H.**, **Rao, R.**, **Hie, B.**, **Zhu, Z.**, **Lu, W.** ***et al.*** (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, **379**(6637), 1123–1130. Available from: https://doi.org/10.1126/science.ade2574

Lo Presti, L., **Lanver, D.**, **Schweizer, G.**, **Tanaka, S.**, **Liang, L.**, **Tollot, M.** ***et al.*** (2015) Fungal effectors and plant susceptibility. *Annual Review of Plant Biology*, **66**, 513–545. Available from: https://doi.org/10.1146/annurev-arplant-043014-114623

Madani, A., **Krause, B.**, **Greene, E.R.**, **Subramanian, S.**, **Mohr, B.P.**, **Holton, J.M.** ***et al.*** (2023) Large language models generate functional protein

sequences across diverse families. *Nature Biotechnology*, **41**(8), 1099–1106. Available from: https://doi.org/10.1038/s41587-022-01618-2

Mishra, B., **Kumar, N.** & **Mukhtar, M.S.** (2019) Systems biology and machine learning in plant-pathogen interactions. *Molecular Plant-Microbe Interactions*, **32**(1), 45–55. Available from: https://doi.org/10.1094/MPMI-08-18-0221-FI

Nakamura, T., **Yamada, K.D.**, **Tomii, K.** & **Katoh, K.** (2018) Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics (Oxford, England)*, **34**(14), 2490–2492. Available from: https://doi.org/10.1093/bioinformatics/bty121

Nur, M., **Wood, K.** & **Michelmore, R.** (2023) EffectorO: motif-independent prediction of effectors in oomycete genomes using machine learning and lineage specificity. *Molecular Plant-Microbe Interactions*, **36**(7), 397–410. Available from: https://doi.org/10.1094/mpmi-11-22-0236-ta

Ofer, D., **Brandes, N.** & **Linial, M.** (2021) The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, **19**, 1750–1758. Available from: https://doi.org/10.1016/j.csbj.2021.03.022

Olsen, T.H., **Moal, I.H.** & **Deane, C.M.** (2022) AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, **2**(1), vbac046. Available from: https://doi.org/10.1093/bioadv/vbac046

Pedregosa, F., **Varoquaux, G.**, **Gramfort, A.**, **Michel, V.**, **Thirion, B.**, **Grisel, O.** ***et al.*** (2011) Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, **12**, 2825–2830. Available from: https://doi.org/10.5555/1953048.2078195

Ran, Z., **Wang, C.**, **Sun, H.**, **Pan, S.** & **Li, F.** (2024) Characterizing secretion system effector proteins with structure-aware graph neural networks and pre-trained language models. *IEEE Journal of Biomedical and Health Informatics*, **28**(9), 5649–5657. Available from: https://doi.org/10.1109/jbhi.2024.3413146

Réau, M., **Renaud, N.**, **Xue, L.C.** & **Bonvin, A.** (2023) DeepRank-GNN: a graph neural network framework to learn patterns in protein-protein interfaces. *Bioinformatics*, **39**(1), btac759. Available from: https://doi.org/10.1093/bioinformatics/btac759

Rives, A., **Meier, J.**, **Sercu, T.**, **Goyal, S.**, **Lin, Z.**, **Liu, J.** ***et al.*** (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **118**(15), e2016239118. Available from: https://doi.org/10.1073/pnas.2016239118

Shen, J., **Zhang, J.**, **Luo, X.**, **Zhu, W.**, **Yu, K.**, **Chen, K.** ***et al.*** (2007) Predicting protein-protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(11), 4337–4341. Available from: https://doi.org/10.1073/pnas.0607879104

Singh, J., **Paliwal, K.**, **Litfin, T.**, **Singh, J.** & **Zhou, Y.** (2022) Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment. *Scientific Reports*, **12**(1), 7607. Available from: https://doi.org/10.1038/s41598-022-11684-w

Sperschneider, J. & **Dodds, P.N.** (2022) EffectorP 3.0: prediction of apoplastic and cytoplasmic effectors in fungi and oomycetes. *Molecular Plant-Microbe Interactions*, **35**(2), 146–156. Available from: https://doi.org/10.1094/mpmi-08-21-0201-r

Sperschneider, J., **Dodds, P.N.**, **Gardiner, D.M.**, **Singh, K.B.** & **Taylor, J.M.** (2018) Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. *Molecular Plant Pathology*, **19**(9), 2094–2110. Available from: https://doi.org/10.1111/mpp.12682

Sperschneider, J., **Gardiner, D.M.**, **Dodds, P.N.**, **Tini, F.**, **Covarelli, L.**, **Singh, K.B.** ***et al.*** (2016) EffectorP: predicting fungal effector proteins from secretomes using machine learning. *New Phytologist*, **210**(2), 743–761. Available from: https://doi.org/10.1111/nph.13794

Stergiopoulos, I. & **de Wit, P.J.** (2009) Fungal effector proteins. *Annual Review of Phytopathology*, **47**, 233–263. Available from: https://doi.org/10.1146/annurev.phyto.112408.132637

Stukenbrock, E. & **Gurr, S.** (2023) Address the growing urgency of fungal disease in crops. *Nature*, **617**(7959), 31–34. Available from: https://doi.org/10.1038/d41586-023-01465-4

Su, J., **Han, C.**, **Zhou, Y.**, **Shan, J.**, **Zhou, X.** & **Yuan, F.** (2024) SaProt: protein language modeling with structure-aware vocabulary. *bioRxiv*, 1663. https://doi.org/10.1101/2023.10.01.560349

Thumuluri, V., **Almagro Armenteros, J.J.**, **Johansen, A.R.**, **Nielsen, H.** & **Winther, O.** (2022) DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Research*, **50**(W1), W228–w234. Available from: https://doi.org/10.1093/nar/gkac278

Toruño, T.Y., **Stergiopoulos, I.** & **Coaker, G.** (2016) Plant-pathogen effectors: cellular probes interfering with plant defenses in spatial and temporal manners. *Annual Review of Phytopathology*, **54**, 419–441. Available from: https://doi.org/10.1146/annurev-phyto-080615-100204

UniProt Consortium. (2025) UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, **53**(D1), D609–D617. Available from: https://doi.org/10.1093/nar/gkae1010

Urban, M., **Cuzick, A.**, **Seager, J.**, **Wood, V.**, **Rutherford, K.**, **Venkatesh, S.Y.** ***et al.*** (2020) PHI-base: the pathogen-host interactions database. *Nucleic Acids Research*, **48**(D1), D613–d620. Available from: https://doi.org/10.1093/nar/gkz904

van Kempen, M., **Kim, S.S.**, **Tumescheit, C.**, **Mirdita, M.**, **Lee, J.**, **Gilchrist, C.L.M.** ***et al.*** (2024) Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, **42**(2), 243–246. Available from: https://doi.org/10.1038/s41587-023-01773-0

Veličković, P., **Cucurull, G.**, **Casanova, A.**, **Romero, A.**, **Liò, P.** & **Bengio, Y.** (2018) Graph attention networks. *arXiv*. https://doi.org/10.48550/arXiv.1710.10903

Wang, C., **Wang, P.**, **Han, S.**, **Wang, L.**, **Zhao, Y.** & **Juan, L.** (2020) FunEffector-Pred: identification of fungi effector by activate learning and genetic algorithm sampling of imbalanced data. *IEEE Access*, **8**, 57674–57683. Available from: https://doi.org/10.1109/ACCESS.2020.2982410

Wawra, S., **Trusch, F.**, **Matena, A.**, **Apostolakis, K.**, **Linne, U.**, **Zhukov, I.** ***et al.*** (2017) The RxLR motif of the host targeting effector AVR3a of *Phytophthora infestans* is cleaved before secretion. *Plant Cell*, **29**(6), 1184–1195. Available from: https://doi.org/10.1105/tpc.16.00552

Whisson, S.C., **Boevink, P.C.**, **Moleleki, L.**, **Avrova, A.O.**, **Morales, J.G.**, **Gilroy, E.M.** ***et al.*** (2007) A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature*, **450**(7166), 115–118. Available from: https://doi.org/10.1038/nature06203

Wong, F., **Zheng, E.J.**, **Valeri, J.A.**, **Donghia, N.M.**, **Anahtar, M.N.**, **Omori, S.** ***et al.*** (2024) Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, **626**(7997), 177–185. Available from: https://doi.org/10.1038/s41586-023-06887-8

Wood, K.J., **Nur, M.**, **Gil, J.**, **Fletcher, K.**, **Lakeman, K.**, **Gann, D.** ***et al.*** (2020) Effector prediction and characterization in the oomycete pathogen *Bremia lactucae* reveal host-recognized WY domain proteins that lack the canonical RXLR motif. *PLoS Pathogens*, **16**(10), e1009012. Available from: https://doi.org/10.1371/journal.ppat.1009012

Zhang, Y., **Guan, J.**, **Li, C.**, **Wang, Z.**, **Deng, Z.**, **Gasser, R.B.** ***et al.*** (2023) DeepSecE: a deep-learning-based framework for multiclass prediction of secreted proteins in gram-negative bacteria. *Research (Washington, D.C.)*, **6**, 0258. Available from: https://doi.org/10.34133/research.0258

Zhang, Y. & **Skolnick, J.** (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, **33**(7), 2302–2309. Available from: https://doi.org/10.1093/nar/gki524

Zhang, Y., **Zhang, Y.**, **Xiong, Y.**, **Wang, H.**, **Deng, Z.**, **Song, J.** ***et al.*** (2022) T4SEfinder: a bioinformatics tool for genome-scale prediction of bacterial type IV secreted effectors using pre-trained protein language model. *Briefings in Bioinformatics*, **23**(1), bbab420. Available from: https://doi.org/10.1093/bib/bbab420

Zhao, M., **Lei, C.**, **Zhou, K.**, **Huang, Y.**, **Fu, C.**, **Yang, S.** ***et al.*** (2024) POOE: predicting oomycete effectors based on a pre-trained large protein language model. *mSystems*, **9**(1), e0100423. Available from: https://doi.org/10.1128/msystems.01004-23

Zhou, Y., **Zhou, Y.S.**, **He, F.**, **Song, J.** & **Zhang, Z.** (2012) Can simple codon pair usage predict protein-protein interaction? *Molecular BioSystems*, **8**(5), 1396–1404. Available from: https://doi.org/10.1039/c2mb05427b