Journal of proteome research

pubs.acs.org/jpr

Technical Note

AraPathogen2.0: An Improved Prediction of Plant−Pathogen Protein−Protein Interactions Empowered by the Natural Language Processing Technique

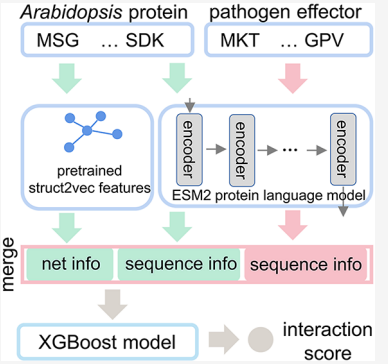
Chenping Lei, Kewei Zhou, Jingyan Zheng, Miao Zhao, Yan Huang, Huaqin He, Shiping Yang,* and Ziding Zhang*

Cite This: *J. Proteome Res.* 2024, 23, 494−499

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Plant−pathogen protein−protein interactions (PPIs) play crucial roles in the arm race between plants and pathogens. Therefore, the identification of these interspecies PPIs is very important for the mechanistic understanding of pathogen infection and plant immunity. Computational prediction methods can complement experimental efforts, but their predictive performance still needs to be improved. Motivated by the rapid development of natural language processing and its successful applications in the field of protein bioinformatics, here we present an improved XGBoost-based plant−pathogen PPI predictor (i.e., AraPathogen2.0), in which sequence encodings from the pretrained protein language model ESM2 and *Arabidopsis* PPI network-related node representations from the graph embedding technique struc2vec are used as input. Stringent benchmark experiments showed that AraPathogen2.0 could achieve a better performance than its precedent version, especially for processing the test data set with novel proteins unseen in the training data.



KEYWORDS: *plant−pathogen interaction, protein−protein interaction, machine learning, natural language processing*

1. INTRODUCTION

Plants face the threat of pathogens throughout their lives, and they have evolved a two-layered immune defense system to fight against the potential pathogens.[1] The first layer is pattern-triggered immunity (PTI), in which pattern recognition receptors of plant surface cells can recognize pathogen-associated molecular patterns and activate the immune response. Pathogens can secrete effector proteins into plant cells to interact with host proteins and thus to subvert the PTI response. In turn, plants evolved resistance proteins (R proteins) to directly or indirectly recognize effectors and trigger the second layer of the plant immune system called effector-triggered immunity (ETI).[2] Generally, the interactions between plant proteins and pathogen effectors are heavily involved in the ETI process, and the identification of these protein−protein interactions (PPIs) is important to decipher the molecular mechanism of plant−pathogen relationships.[3]

Large-scale experimental identification of *Arabidopsis*−pathogen PPIs has generated a considerable amount of PPI data, including the PPIN-1,[4] PPIN-2,[5] and EffectorK[6] data sets. To supplement experimental efforts, generic computational methods such as interolog mapping and domain−domain interaction-based inference have been widely used to predict plant−pathogen PPIs.[7] On the other hand, the accumulation of known *Arabidopsis*−pathogen PPI data provides a valuable resource for developing machine learning (ML)-based predictors. In 2019, our team developed AraPathogen1.0 to predict the interactions between *Arabidopsis* proteins and pathogen effectors, which utilizes Random Forest (RF) to integrate multiple sequence encoding schemes and host network features.[8] Until now, more than 3,000 queries from ∼40 countries have been processed. Despite the good achievement of AraPathogen1.0, it often performs poor on predicting protein pairs containing novel plant or pathogen proteins unseen in the training data. Therefore, performance improvement on the prediction of *Arabidopsis*−pathogen PPIs is still urgently needed.

In the past years, we have witnessed the rapid development of natural language processing (NLP). Typically, the word/document embedding algorithm can convert a word/document into a semantically rich vector representation through the training of a corpus, which has been easily adapted to characterize protein sequences.[9] Very recently, protein language models (PLMs) with unsupervised training were further investigated to extract features from a large volume of protein sequences. Interestingly, such pretrained large PLMs yield protein features containing rich structural and functional

Received: June 18, 2023
Revised: October 6, 2023
Accepted: November 14, 2023
Published: December 9, 2023

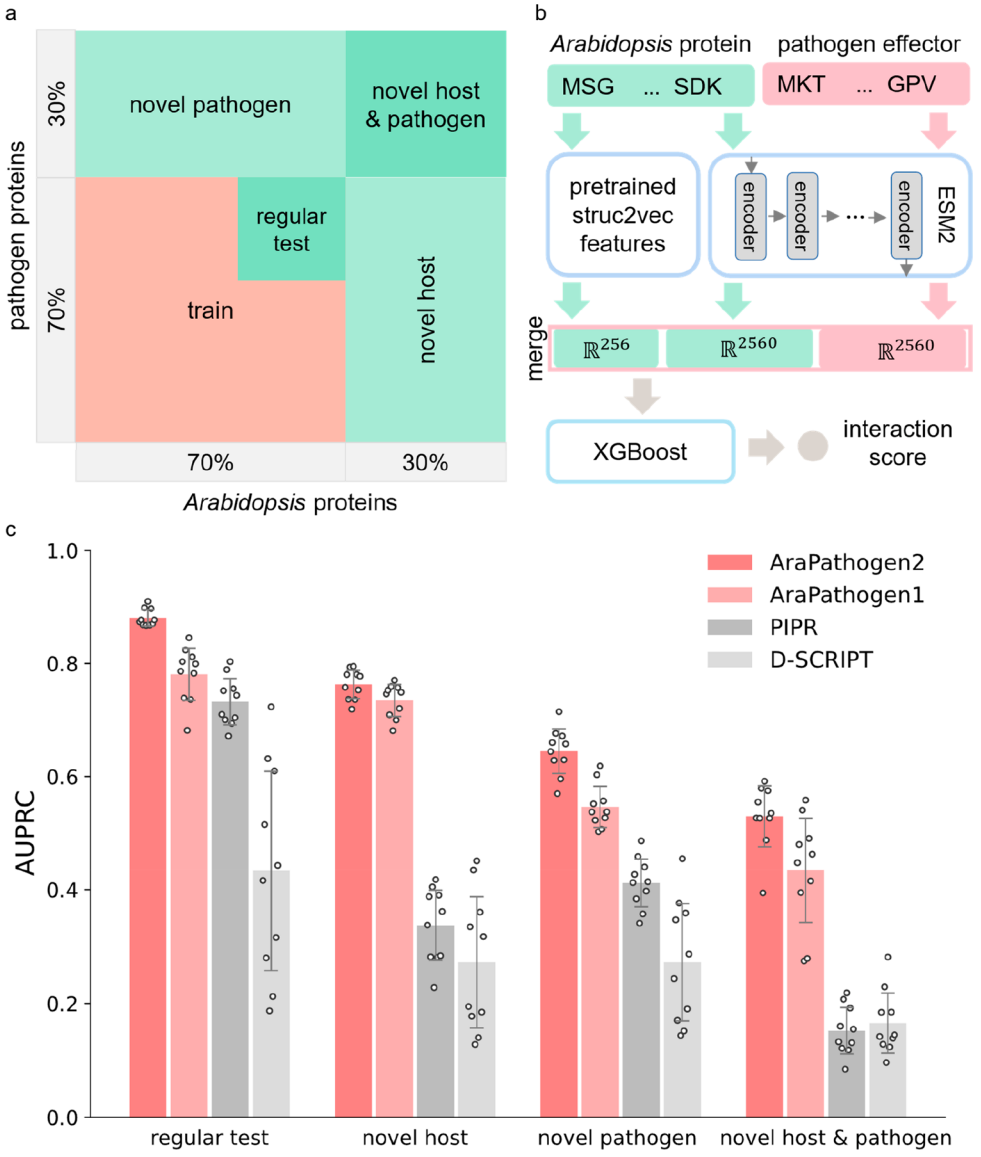


Figure 1. Data set partition (A), computational framework (B), and performance assessment (C) of the proposed AraPathogen2.0.

properties of proteins, and they have been proven to be very powerful in many protein prediction tasks,[10,11] such as predictions of secondary structure and mutational effects.[12] One representative PLM is ESM2, which was trained on ∼65 million protein sequences by using Transformers.[10] Also inspired by NLP, the graph embedding techniques, such as node2vec[13] and struc2vec,[14] were proposed to convert nodes in a biological network into vector representations. By doing so, the comprehensive topology properties of nodes in the network are effectively captured. In this context, it is very natural to evolve AraPathogen1.0 into an upgraded version (i.e., AraPathogen2.0) by taking advantage of the NLP technique. In general, AraPathogen2.0 shares the similar prediction strategy with its predecessor, but it is an extreme gradient boosting (XGBoost)-based predictor using ESM2 to extract sequence features and struc2vec to characterize host network properties.

2. MATERIALS AND METHODS

2.1. Data Set Preparation

We collected 1,387 PPIs between 564 *Arabidopsis* proteins and 286 pathogen effectors from PPIN-1, PPIN-2, and EffectorK, which were considered as positive samples. Compared with AraPathogen1.0, 928 PPIs were newly collected. To construct negative samples (i.e., non-PPIs), we randomly paired the noninteracting proteins between nonredundant effectors and *Arabidopsis* proteins and further selected 10 times the number of positive samples as the original negative samples (Supplemental Table S1). Thus, we obtained an original data set containing 1,387 PPIs and 13,870 non-PPIs, which cover 8,505 *Arabidopsis* proteins and 872 pathogen effectors. To allow for an unbiased model training and assessment, we used a modified training-test data set partition method proposed by Park and Marcotte.[15] Briefly, we randomly selected 70% *Arabidopsis* proteins and 70% pathogen effectors from our original data set, and all the potential protein pairs in this subset were further randomly divided into two sets (90% protein pairs were used to construct one training data set

https://doi.org/10.1021/acs.jproteome.3c00364
J. Proteome Res. 2024, 23, 494−499

named "training" and 10% protein pairs were used to construct one test set named "regular test") (Figure 1a). If protein pairs in these two data sets were not assigned as positive or negative samples, they were skipped. The remaining proteins from the 30% *Arabidopsis* proteins and 30% pathogen effectors were used to construct another three test sets named "novel host", "novel pathogen", and "novel host and pathogen". As illustrated in Figure 1a, the "regular test" means sampling test data without considering protein existence in the training data, while "novel host" and "novel pathogen" should not contain host and pathogen proteins in the training data, respectively. The "novel host and pathogen" means that both host and pathogen proteins are unseen in the training data. By doing so, the ratio of the training set and four individual test sets is roughly controlled at 44:5:21:21:9 (Figure 1a), and all the data set partitions were repeated ten times (Supplemental Table S2).

2.2. Model Architecture and Performance Evaluation

The AraPathogen2.0 model contains three modules (Figure 1b), namely the pretrained ESM2 to infer sequence embeddings, the pretrained struc2vec to extract host network features, and the XGBoost classifier. XGBoost is an accurate and efficient ensemble learning algorithm based on gradient boosting decision trees, which has been proved to be suitable for handling high-dimensional data in bioinformatics tasks. ESM2 consists of 36 encoders, each of which contains a multihead and a multilayer perceptron, while EsmMean is the feature vector of the final layer averaged over the protein length.[10] We used EsmMean to encode the sequences of *Arabidopsis* proteins and pathogen effectors, which was implemented using a script provided by the ESM authors (https://github.com/facebookresearch/esm/). The EsmMean embedding for a protein has a dimensionality of 2,560, which allows us to convert a protein pair into a 5,120 dimensional vector. More details about the EsmMean encoding are available in Table 1. We collected 42,236 *Arabidopsis* PPIs from TAIR (https://www.arabidopsis.org/) and UniProt (https://www.uniprot.org/), and compiled them into an *Arabidopsis* PPI network (i.e., AraNet). The struc2vec embedding is derived from a pretrained struc2vec model on the *Arabidopsis* PPI network, which was implemented using the GraphEmbedding software (https://github.com/shenweichen/GraphEmbedding). In brief, struc2vec first generates a multilayer graph based on the PPI network, then produces node paths by walking between or within layers of the graph, and trains word2vec on these node paths as the corpus to generate node representation vectors.

To benchmark the proposed AraPathogen2.0, we applied three other sequence-based encodings [i.e., ProtTrans,[11] dipeptide composition (DPC)[16] and composition of k-spaced amino acid pairs (CKSAAP)[16,17]] and another one network-based encoding (i.e., node2vec) as the baseline encoding schemes. Brief descriptions of these baseline encoding schemes are shown in Table 1. We also employed three other popular ML methods [i.e., RF, support vector machine (SVM), and multilayer perceptron (MLP)] as the baseline methods. The implementation platforms of different ML methods and the corresponding model parameter and optimization are shown in Supplemental Tables S3 and S4.

Table 1. Brief Descriptions of Different Encoding Schemes

Encodings	Dimension	Description
EsmMean	2,560	The EsmMean encoding is the vector output (L × 2,560) of the final layer (i.e., the 36th layer) of ESM2 (esm2_t36_3B_UR50D) averaged over the protein length (L). To infer the EsmMean encoding, we employed the feature extract script available at (https://github.com/facebookresearch/esm/blob/main/scripts/extract.py).
ProtTrans	1,024	ProtTrans used the UniRef and BFD (Big Fantastic Database) data set as the corpus and employed autoregressive and autoencoder models to generate protein representations. In our work, we downloaded an autoencoder model called ProtT5 from https://github.com/agemagician/ProtTrans and installed it for local use. For each protein, the ProtT5 model was used to generate a final representation of L × 1,024, where L is the length of the protein. In this work, the ProtTrans encoding is a vector of the protT5 output (L × 1,024) averaged by length.
struc2vec	256	The graph node embedding method struc2vec can extract both the topological and neighbor information of nodes in the network. In this work, the struc2vec embedding is inferred from a pretrained model on the *Arabidopsis* PPI network (i.e., AraNet) using GraphEmbedding implementation (https://github.com/shenweichen/GraphEmbedding).
node2vec	256	The node embedding method node2vec can extract the node neighbor information in a network. It generates node paths through random walks in the network, and then the node embedding is extracted from the training of word2vec with node paths serving as a corpus. In this work, the node2vec technique was used to represent the proteins in the *Arabidopsis* PPI network, which was implemented through GraphEmbedding (https://github.com/shenweichen/GraphEmbedding).
DPC	400	DPC is a conventional sequence encoding scheme, which represents the frequency of dipeptides in a protein sequence.
CKSAAP	1,600	CKSAAP represents the frequency of k-spaced dipeptide combinations in a protein sequence. In this work, k = 0, 1, 2, and 3 were taken into account.

https://doi.org/10.1021/acs.jproteome.3c00364
J. Proteome Res. 2024, 23, 494−499

Table 2. Performance Comparison of Different Machine Learning Methods Based on the EsmMean and struc2vec Encodings

Test set	ML	Precision	Recall	Specificity	F1	AUPRC
regular test	XGBoost	0.921 ± 0.041	0.680 ± 0.041	0.994 ± 0.004	0.781 ± 0.024	0.881 ± 0.015
	RF	0.871 ± 0.052	0.501 ± 0.067	0.992 ± 0.004	0.632 ± 0.053	0.769 ± 0.041
	SVM	0.713 ± 0.075	0.548 ± 0.061	0.978 ± 0.006	0.619 ± 0.064	0.685 ± 0.076
	MLP	0.793 ± 0.094	0.705 ± 0.109	0.979 ± 0.014	0.734 ± 0.060	0.810 ± 0.038
novel host	XGBoost	0.881 ± 0.023	0.403 ± 0.040	0.995 ± 0.001	0.552 ± 0.036	0.763 ± 0.025
	RF	0.906 ± 0.036	0.199 ± 0.050	0.998 ± 0.001	0.323 ± 0.067	0.703 ± 0.034
	SVM	0.765 ± 0.088	0.105 ± 0.036	0.997 ± 0.002	0.182 ± 0.057	0.544 ± 0.065
	MLP	0.714 ± 0.079	0.466 ± 0.107	0.980 ± 0.011	0.551 ± 0.068	0.647 ± 0.048
novel pathogen	XGBoost	0.798 ± 0.045	0.362 ± 0.038	0.991 ± 0.003	0.497 ± 0.038	0.645 ± 0.039
	RF	0.790 ± 0.081	0.301 ± 0.046	0.992 ± 0.004	0.433 ± 0.053	0.593 ± 0.040
	SVM	0.665 ± 0.044	0.449 ± 0.028	0.977 ± 0.004	0.536 ± 0.029	0.580 ± 0.035
	MLP	0.605 ± 0.045	0.425 ± 0.089	0.971 ± 0.011	0.491 ± 0.054	0.526 ± 0.045
novel host and pathogen	XGBoost	0.760 ± 0.072	0.078 ± 0.036	0.997 ± 0.001	0.139 ± 0.058	0.530 ± 0.054
	RF	0.864 ± 0.174	0.033 ± 0.018	0.999 ± 0.001	0.063 ± 0.034	0.461 ± 0.063
	SVM	0.618 ± 0.288	0.062 ± 0.033	0.997 ± 0.002	0.112 ± 0.059	0.442 ± 0.057
	MLP	0.573 ± 0.108	0.259 ± 0.082	0.980 ± 0.009	0.350 ± 0.083	0.425 ± 0.077

3. RESULTS AND DISCUSSION

We compiled four test sets to comprehensively assess the performance of AraPathogen2.0. Considering that the positive and negative samples are highly imbalanced, we plotted the Precision−Recall curve and mainly used the area under the PR curve (AUPRC) to characterize the predictive performance. In addition, we also introduced four common performance metrics (i.e., Precision, Recall, Specificity, and F1-score) for method evaluation. As shown in Figure 1c and Table 2, AraPathogen2.0 achieves excellent performance and the corresponding AUPRC values in the four test sets are 0.881, 0.763, 0.645 and 0.530, respectively. When benchmarked against three popular ML algorithms, XGBoost revealed the best performance in all the four test sets (Table 2). In the evaluation of different sequence encodings, the EsmMean and ProtTrans features derived from pretrained PLMs yielded much higher AUPRC values than the other two conventional encoding schemes (i.e., DPC and CKSAAP) on the four test sets (Supplemental Table S5). Comparatively, the performance of EsmMean is slightly better than that of ProtTrans. Although the individual features in EsmMean did not contain clear biologically meaningful information, it is still interesting to investigate the contributions of different EsmMean features. Indeed, ∼2,800 out of the 5,120 EsmMean features were ranked as important features (i.e., the important score inferred from the XGBoost model >0.0), suggesting different EsmMean features contribute differently to plant−pathogen PPI prediction. Indeed, when only top ranking features were used to train the XGBoost model, the resulting performance in four different tests is very close to that using all the features (Supplemental Figure S1). Regarding the network-based encodings, struc2vec performed better than node2vec, especially on the "novel host" and "novel host and pathogen" test sets (Supplemental Table S6), which is probably because struc2vec incorporates more topological information on nodes to ensure a better node representation. Altogether, the above computational experiments confirmed that the XGBoost model with the feature combination of EsmMean and Struc2vec achieved the best performance among all the computational frameworks we tested.

We further compare AraPathogen2.0 with its predecessor AraPathogen1.0. As shown in Figure 1c and Supplemental Table S7, AraPathogen2.0 revealed considerable performance improvement in all the four test sets. To comprehensively evaluate the performance of the AraPathogen2.0 model, we also compared it with two state-of-the-art generic PPI predictors, PIPR[18] and D-SCRIPT.[19] PIPR integrates the word2vec embedding and one-hot encoding with a 5-layer recurrent neural network and achieved excellent performance in predicting PPIs. D-SCRIPT utilizes a pretrained PLM to obtain structurally informative embeddings as input and conducts the PPI prediction through the integration of bidirectional LSTM and CNN, exhibiting good generalization performance in cross-species PPI prediction tasks. The results showed that both AraPathogen2.0 and AraPathogen1.0 considerably outperformed PIPR and D-SCRIPT on the four test sets (Supplemental Table S7). The more favorable performance of AraPathogen2.0 over PIPR and D-SCRIPT should be ascribed to the following two reasons. First, AraPathogen2.0 has introduced host PPI network information complementary to sequence information compared to PIPR and D-SCRIPT. Second, the sequence embedding technique used in AraPathogen2.0 (i.e., ESM2) is more advanced than those used in PIPR and D-SCRIPT, since ESM2 adopts a much larger corpus and a more powerful architecture (Transformers) for model training. Taken together, the above computational experiments indicated that AraPathogen2.0 outperforms existing methods in predicting plant−pathogen PPIs.

4. CONCLUSION

In this work, we developed AraPathogen2.0 for plant−pathogen PPI prediction. AraPathogen2.0 is an XGBoost-based predictor trained on a comprehensive *Arabidopsis*−pathogen PPI data set. More importantly, it adopted powerful ESM2 and struc2vec to construct the sequence and network representations. Rigorous benchmark experiments clearly and consistently show that we have considerably improved the prediction performance of plant−pathogen PPIs, especially for those PPIs with proteins unseen in training data. We anticipate that AraPathogen2.0 can serve as a useful tool to identify potential interactions between *Arabidopsis* proteins and pathogen effectors, further guiding hypothesis-driven experimental efforts to decipher plant−pathogen relationships.

https://doi.org/10.1021/acs.jproteome.3c00364
J. Proteome Res. 2024, 23, 494−499

■ ASSOCIATED CONTENT

Data Availability Statement

Currently, AraPathogen2.0 is freely accessible at http://zzdlab.com/arapathogen2/index.php. The source codes and datasets are also available at https://github.com/miderxi/arapathogen2.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.3c00364.

Table S1: Information of positive and negative samples. Table S2: Sizes of ten repeated data sets. Table S3: Optimization and setting of model parameters in different machine learning algorithms. Table S4: Optimization and setting of model parameters in training two network embedding methods. Table S5: Performance comparison of different sequence encoding schemes using XGBoost. Table S6: Performance comparison of two network embeddings using XGBoost. Table S7: Performance comparison of AraPathogen2.0 with three existing methods. Figure S1: The performance of the XGBoost model when using only top N important EsmMean features (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Shiping Yang − *State Key Laboratory of Plant Environmental Resilience, College of Biological Sciences, China Agricultural University, Beijing 100193, China*; Email: shi_ping_yang@163.com

Ziding Zhang − *State Key Laboratory of Animal Biotech Breeding, College of Biological Sciences, China Agricultural University, Beijing 100193, China*; orcid.org/0000-0002-9296-571X; Email: zidingzhang@cau.edu.cn

Authors

Chenping Lei − *State Key Laboratory of Animal Biotech Breeding, College of Biological Sciences, China Agricultural University, Beijing 100193, China*; orcid.org/0009-0001-0938-4042

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

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

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

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

Huaqin He − *College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.3c00364

Author Contributions

C.L. developed the prediction model and drafted the paper. K.Z., J.Z., M.Z., and Y.H. participated in the data analysis. H.H. gave key advice in the model construction and evaluation. Z.Z. and S.Y. supervised the study and significantly revised the paper. All authors have read and agreed to the submitted version of the paper.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful for the stimulating discussion with Dr. Stefan Wuchty regarding the prediction and network analysis of host−pathogen interaction. This work was supported by the National Natural Science Foundation of China (31471249 and 31970645).

■ ABBREVIATIONS

PTI, pattern-triggered immunity; ETI, effector-triggered immunity; PPIs, protein−protein interactions; ML, machine learning; NLP, natural language processing; RF, random forest; XGBoost, extreme gradient boosting; SVM, support vector machine; MLP, multilayer perceptron; LSTM, long short-term memory; CNN, convolutional neural network; PLMs, protein language models; AUPRC, the area under the precision-recall curve; CKSAAP, composition of k-spaced amino acid pairs; DPC, dipeptide composition

■ REFERENCES

(1) Jones, J. D. G.; Dangl, J. L. The Plant Immune System. *Nature* **2006**, *444* (7117), 323−329.

(2) Zhang, S.; Li, C.; Si, J.; Han, Z.; Chen, D. Action Mechanisms of Effectors in Plant-Pathogen Interaction. *Int. J. Mol. Sci.* **2022**, *23* (12), 6758.

(3) Dong, A.; Wang, Z.; Huang, J.; Song, B.; Hao, G. Bioinformatic Tools Support Decision-Making in Plant Disease Management. *Trends Plant Sci.* **2021**, *26* (9), 953−967.

(4) Mukhtar, M. S.; Carvunis, A.; Dreze, M.; Epple, P.; Steinbrenner, J.; Moore, J.; Tasan, M.; Galli, M.; Hao, T.; Nishimura, M. T.; Pevzner, S. J.; Donovan, S. E.; Ghamsari, L.; Santhanam, B.; Romero, V.; Poulin, M. M.; Gebreab, F.; Gutierrez, B. J.; Tam, S.; Monachello, D.; Boxem, M.; Harbort, C. J.; McDonald, N.; Gai, L.; Chen, H.; He, Y.; European Union Effectoromics Consortium; Vandenhaute, J.; Roth, F. P.; Hill, D. E.; Ecker, J. R.; Vidal, M.; Beynon, J.; Braun, P.; Dangl, J. L. Independently Evolved Virulence Effectors Converge onto Hubs in a Plant Immune System Network. *Science* **2011**, *333* (6042), 596−601.

(5) Weßling, R.; Epple, P.; Altmann, S.; He, Y.; Yang, L.; Henz, S. R.; McDonald, N.; Wiley, K.; Bader, K. C.; Gläßer, C.; Mukhtar, M. S.; Haigis, S.; Ghamsari, L.; Stephens, A. E.; Ecker, J. R.; Vidal, M.; Jones, J. D. G.; Mayer, K. F. X.; Ver Loren van Themaat, E.; Weigel, D.; Schulze Lefert, P.; Dangl, J. L.; Panstruga, R.; Braun, P. Convergent Targeting of a Common Host Protein-Network by Pathogen Effectors from Three Kingdoms of Life. *Cell Host Microbe* **2014**, *16* (3), 364−375.

(6) GonzálezFuente, M.; Carrère, S.; Monachello, D.; Marsella, B. G.; Cazalé, A.; Zischek, C.; Mitra, R. M.; Rezé, N.; Cottret, L.; Mukhtar, M. S.; Lurin, C.; Noël, L. D.; Peeters, N. EffectorK, a Comprehensive Resource to Mine for Ralstonia, Xanthomonas, and Other Published Effector Interactors in the Arabidopsis Proteome. *Mol. Plant Pathol.* **2020**, *21* (10), 1257−1270.

(7) Loaiza, C. D.; Kaundal, R. PredHPI: An Integrated Web Server Platform for the Detection and Visualization of Host-Pathogen Interactions Using Sequence-Based Methods. *Bioinformatics* **2021**, *37* (5), 622−624.

(8) Yang, S.; Li, H.; He, H.; Zhou, Y.; Zhang, Z. Critical Assessment and Performance Improvement of Plant-Pathogen Protein-Protein Interaction Prediction Methods. *Brief. Bioinform.* **2019**, *20* (1), 274−287.

(9) Le, Q.; Mikolov, T. Distributed Representation of Sentences and Documents. In *International Conference on Machine Learning, Vol 32 (cycle 2)*; Xing, E. P., Jebara, T., Eds.; Jmlr-Journal Machine Learning Research: San Diego, 2014; Vol. *32*, pp 1188−1196.
(10) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; Dos Santos Costa, A.; Fazel Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379* (6637), 1123−1130.
(11) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; Rost, B. ProtTrans: Toward Understanding the Language of Life through Self-Supervised Learning. *Ieee Trans. Pattern Anal. Mach. Intell.* **2022**, *44* (10), 7112−7127.
(12) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118* (15), e2016239118.
(13) Grover, A.; Leskovec, J.Node2vec: Scalable Feature Learning for Networks. In *Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*; Assoc Computing Machinery: New York, 2016; pp 855−864.
(14) Ribeiro, L. F. R.; Saverese, P. H. P.; Figueiredo, D. R. Struc2vec: Learning Node Representations from Structural Identity. In *Kdd'17: Proceedings of the 23rd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*; Assoc. Computing Machinery: New York, 2017; pp 385−394.
(15) Park, Y.; Marcotte, E. M. A Flaw in the Typical Evaluation Scheme for Pair-Input Computational Predictions. *Nat. Methods* **2012**, *9* (12), 1134−1136.
(16) Chen, Z.; Zhao, P.; Li, F.; Leier, A.; Marquez-Lago, T. T.; Wang, Y.; Webb, G. I.; Smith, A. I.; Daly, R. J.; Chou, K.-C.; Song, J. iFeature: A Python Package and Web Server for Features Extraction and Selection from Protein and Peptide Sequences. *Bioinformatics* **2018**, *34* (14), 2499−2502.
(17) Hong, X.; Lv, J.; Li, Z.; Xiong, Y.; Zhang, J.; Chen, H.-F. Sequence-Based Machine Learning Method for Predicting the Effects of Phosphorylation on Protein-Protein Interactions. *Int. J. Biol. Macromol.* **2023**, *243*, 125233.
(18) Chen, M.; Ju, C. J.-T.; Zhou, G.; Chen, X.; Zhang, T.; Chang, K.-W.; Zaniolo, C.; Wang, W. Multifaceted Protein-Protein Interaction Prediction Based on Siamese Residual RCNN. *Bioinformatics* **2019**, *35* (14), i305−i314.
(19) Sledzieski, S.; Singh, R.; Cowen, L.; Berger, B. D-SCRIPT Translates Genome to Phenome with Sequence-Based, Structure-Aware, Genome-Scale Predictions of Protein-Protein Interactions. *Cell Syst.* **2021**, *12* (10), 969−982.

https://doi.org/10.1021/acs.jproteome.3c00364
J. Proteome Res. 2024, 23, 494−499