the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2023) doi: 10.1111/tpj.16188

TECHNICAL ADVANCE

Deep learning-assisted prediction of protein–protein interactions in *Arabidopsis thaliana*

Jingyan Zheng[1,†], Xiaodi Yang[2,†], Yan Huang[1], Shiping Yang[3], Stefan Wuchty[4,5,6,7] and Ziding Zhang[1,*]

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

[2]*Department of Hematology, Peking University First Hospital, Beijing 100034, China,*

[3]*State Key Laboratory of Plant Physiology and Biochemistry, College of Biological Sciences, China Agricultural University, Beijing 100193, China,*

[4]*Department of Computer Science, University of Miami, Miami, FL, 33146, USA,*

[5]*Department of Biology, University of Miami, Miami, FL, 33146, USA,*

[6]*Sylvester Comprehensive Cancer Center, University of Miami, Miami, FL, 33136, USA, and*

[7]*Institute of Data Science and Computing, University of Miami, Miami, FL, 33146, USA*

Received 17 October 2022; revised 20 February 2023; accepted 9 March 2023.

*For correspondence (e-mail zidingzhang@cau.edu.cn).

†These authors contributed equally to this work.

SUMMARY

Currently, the experimentally identified interactome of Arabidopsis (*Arabidopsis thaliana*) is still far from complete, suggesting that computational prediction methods can complement experimental techniques. Motivated by the prosperity and success of deep learning algorithms and natural language processing techniques, we introduce an integrative deep learning framework, DeepAraPPI, allowing us to predict protein–protein interactions (PPIs) of Arabidopsis utilizing sequence, domain and Gene Ontology (GO) information. Our current DeepAraPPI comprises: (i) a word2vec encoding-based Siamese recurrent convolutional neural network (RCNN) model; (ii) a Domain2vec encoding-based multiple-layer perceptron (MLP) model; and (iii) a GO2vec encoding-based MLP model. Finally, DeepAraPPI combines the prediction results of the three individual predictors through a logistic regression model. Compiling high-quality positive and negative training and test samples by applying strict filtering strategies, DeepAraPPI shows superior performance compared with existing state-of-the-art Arabidopsis PPI prediction methods. DeepAraPPI also provides better cross-species predictive ability in rice (*Oryza sativa*) than traditional machine learning methods, although the overall performance in cross-species prediction remains to be improved. DeepAraPPI is freely accessible at http://zzdlab.com/deeparappi/. In the meantime, we have also made the source code and data sets of DeepAraPPI available at https://github.com/zjy1125/DeepAraPPI.

Keywords: *Arabidopsis thaliana*, protein–protein interaction, deep learning, prediction, GO annotation, domain.

INTRODUCTION

Proteins are indispensable macromolecules in living organisms that act in concert with each other through physical interactions. In particular, many basic cellular processes, such as cellular metabolism, transport and regulation, depend on protein–protein interactions (PPIs) (Berggard et al., 2007; Keskin et al., 2016), indicating that the proteome-wide identification of PPIs is of great significance for our systematic understanding of cellular functions. Furthermore, PPIs also play a crucial role in the identification of therapeutic targets and the design of novel drugs (Petta et al., 2016; Shin et al., 2017; Skrabanek et al., 2008).

To reliably measure PPIs, many experimental techniques have been developed and are classified as low- and high-throughput detection methods (Gul & Hadian, 2014; Lian et al., 2021; Peng et al., 2017; Petschnigg et al., 2011). Low-throughput methods include X-ray crystallography,

surface plasmon resonance (SPR) and pull-down assay, whereas high-throughput methods use yeast two-hybrid (Y2H), affinity purification coupled with mass spectrometry (AP-MS) and protein microarrays. In general, these experimental methods are time-consuming, laborious and costly, and have their own advantages and limitations. For instance, high-throughput methods can identify PPIs on a large scale but suffer from high false-positive rates.

As Arabidopsis is an important model plant, knowledge of comprehensive PPI networks is necessary to understand molecular mechanisms such as organ formation, signal transduction and stress response (Lin et al., 2009). Although roughly 300 000 PPIs are estimated to exist in Arabidopsis (Arabidopsis Interactome Mapping Consortium, 2011), the coverage of experimentally verified interactions is rather limited (Ding & Kihara, 2019). Currently, a variety of computational methods have been applied to predict PPIs in Arabidopsis, including interolog mapping (Geisler-Lee et al., 2007), structure-based methods (e.g. molecular docking; Dong et al., 2019), integrative methods of multiple features (Cui et al., 2008; De Bodt et al., 2009; Xu et al., 2010) and machine learning-based approaches, such as random forest (RF) (Zhang et al., 2016) and support vector machines (SVM) (Ding & Kihara, 2019). As different methods have their own strengths and weaknesses, interolog mapping is limited in its prediction abilities to known interactions in other species, whereas structure-based methods are restricted by the need to obtain highly accurate structural information.

As for deep learning methods to predict PPIs, Sun et al. introduced stacked autoencoder (SAE) to develop a sequence-based human PPI predictor (Sun et al., 2017). Du et al. proposed DeepPPI, which uses deep neural networks to efficiently learn the representation of protein pairs to predict PPIs (Du et al., 2017). Hashemifar et al. introduced DPPI, a Siamese-like convolutional neural network (CNN), to extract features from sequences, capturing complex and nonlinear relationships in PPIs (Hashemifar et al., 2018). Chen et al. proposed PIPR, an end-to-end framework to predict PPIs based only on sequence information, utilizing robust local features and contextualized information within protein sequences (Chen et al., 2019). Furthermore, natural language processing (NLP) methods have also been applied, as protein sequences are represented as a string of amino acids. For example, Wu et al. divided each peptide sequence into *k*-mers using a sliding window, embedded proteins through vectors of such *k*-mers using word2vec and applied a deep learning model to predict therapeutic peptides (Wu et al., 2019). Yang, Yang, Li, et al. (2020) converted protein sequences into fixed dimensional feature vectors through doc2vec, an unsupervised sequence embedding technique, to predict human-virus PPIs with an RF model (Yang, Yang, Li, et al., 2020). Zeng et al. (2019) developed DeepEP to identify essential proteins based on a convolutional neural network framework that uses the node2vec technique to automatically learn the topological and semantic features of each protein in the PPI network. Pan et al. introduced a deep learning-based method, termed ToxDL, for predicting protein toxicity. Its framework consists of two modules, including a Skip-gram model (i.e. Domain2vec) to find protein domain embeddings and a CNN to process input sequences (Pan et al., 2021). Zhong et al. proposed GO2vec to learn about the feature embedding of Gene Onotolgy (GO) terms and applied the node2vec model on an established GO graph. As a result, each node in the GO graph was represented by a fixed-dimensional feature vector (Zhong et al., 2019).

Recently, deep learning algorithms have also been applied to predict plant-specific PPIs. In particular, Pan et al. proposed DWPPI, integrating sequence features and PPI network embedding as the input of deep neural networks to predict the PPIs of Arabidopsis, *Oryza sativa* (rice) and *Zea mays* (maize) (Pan et al., 2022). Although experimentally determined plant PPI data have been significantly accumulating in recent years, providing a good foundation for developing deep learning predictive models, source codes or webservers of existing deep learning-based plant PPI prediction methods are rarely available to the community.

To further develop deep learning-based plant PPI predictors, we constructed an integrative deep learning framework, DeepAraPPI, to predict Arabidopsis PPIs by using sequence, domain and GO information (Figure 1a). In particular, DeepAraPPI is based on three predictors, where: (i) sequence information between interacting proteins is captured through a word2vec encoding-based Siamese recurrent convolutional neural network (RCNN) model (Figure 1b); (ii) DeepAraPPI uses Domain2vec to provide domain embeddings of protein pairs as the input to a multiple-layer perceptron (MLP) to predict protein interactions (Figure 1c); and (iii) to capture functional information, DeepAraPPI embeds protein pairs by using GO2vec encodings and predicts PPIs through an MLP. To integrate the power of these predictors, DeepAraPPI employs a logistic regression (LR) model to obtain a comprehensive prediction score, allowing us to show that the integrative LR model performed better than any single predictor. In comparison with existing state-of-the-art Arabidopsis PPI prediction methods, DeepAraPPI provides superior performance.

RESULTS AND DISCUSSION

Overall performance of DeepAraPPI

To train and assess the performance of our proposed DeepAraPPI model, we collected high-quality experimental Arabidopsis PPIs as positive samples. Furthermore, we sampled negative training data by randomly selecting

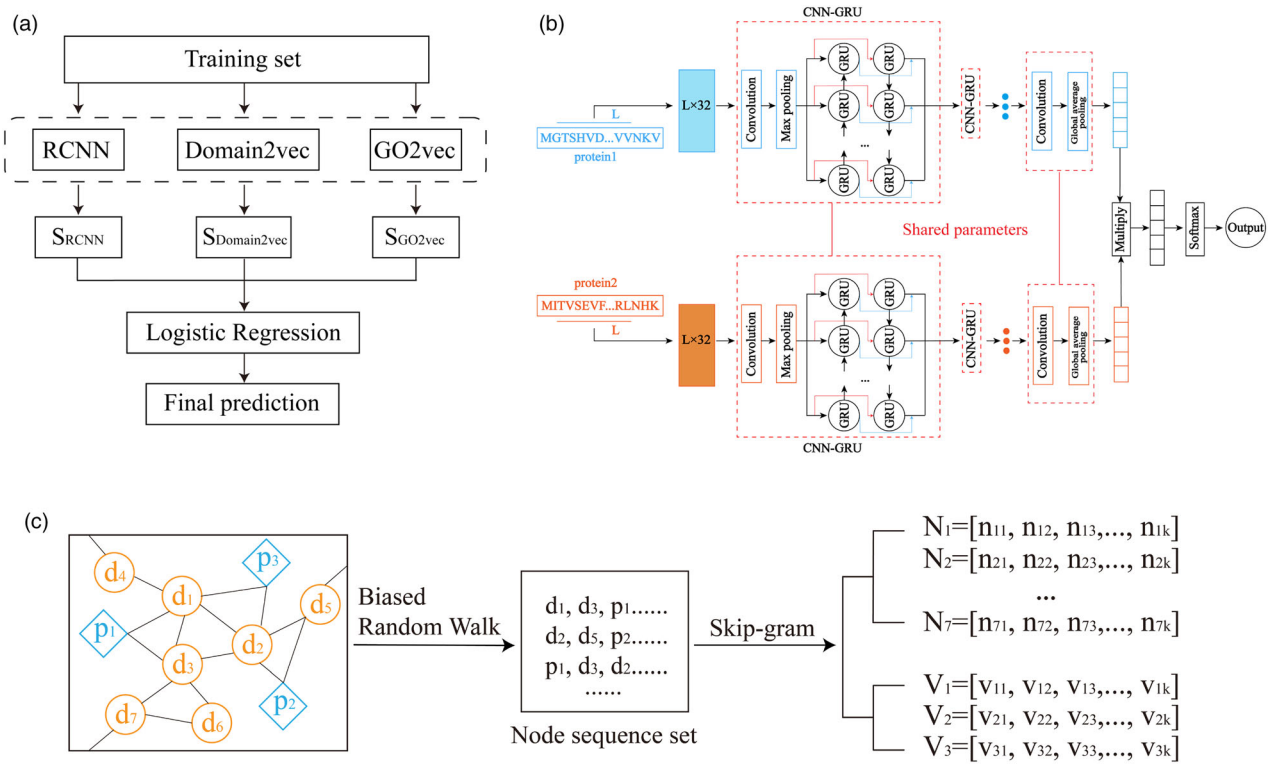


Figure 1. Overview of the computational framework of DeepAraPPI. (a) DeepAraPPI integrates sequence, domain and Gene Ontotlogy (GO) information to predict protein–protein interactions (PPIs) in Arabidopsis. Three baseline models (i.e. RCNN, Domain2vec and GO2vec) are separately trained to obtain the corresponding predictive scores (i.e. S_{RCNN}, $S_{Domain2vec}$ and S_{GO2vec}) that are combined into a vector to train a logistic regression model as our final predictive model. (b) The overall architecture of our Siamese RCNN predictor. (c) Graphic illustration of Domain2vec on the network graph formed by the DDIs and domain annotations of proteins. Briefly, node2vec first generates sequences of node paths by biased random walks, which are further inputted to the SkipGram model of word2vec to obtain the feature vector of each node. d_i labels a domain and N_i denotes its k-dimensional vector, where n_{ij} is the j-th element of n_i. Furthermore, p_m labels a protein and V_m denotes its k-dimensional vector, where v_{mn} is the n-th element of v_m.

Arabidopsis protein pairs that do not co-occur in the same subcellular compartment. Moreover, the ratio of positive to negative samples was set as 1:10. Considering that the performance of PPI prediction is strongly linked to the benchmarking data sets, we designed three tasks through dataset partition (i.e. Task1, Task2 and Task3, corresponding to low, medium and high difficulty level, respectively) to rigorously test the prediction performance of our model. To quantify performance, we utilize commonly used measurements such as TPR (true-positive rate, also called recall), FPR (false-positive rate) and precision. To achieve a more comprehensive performance assessment, we plotted the precision–recall (PR) curve and quantified the performance through the corresponding area under the PR curve (AUPRC), which is commonly used to evaluate classification performance when positive and negative samples are imbalanced. Further details about the preparation of the data set and the methodology of DeepAraPPI are available in the Experimental procedures.

Assessing the model performance of DeepAraPPI using benchmarking data sets at three different difficulty levels, Table 1 summarizes the AUPRC of RCNN, Domain2vec, GO2vec and the integrated LR model on independent test sets. In particular, we observed that with increasing prediction difficulty, the performance of all single models decreased, indicating that prediction accuracy suffers when facing unknown proteins. Comparing all three individual prediction models, GO2vec uniformly performs best, suggesting that our GO2vec model captures functional similarity between interacting proteins effectively. Although RCNN shows sensitivity to increasing prediction difficulty, Domain2vec demonstrates robust performance in all prediction tasks, suggesting that Domain2vec effectively integrates domain information to infer PPIs. Notably, the integrated LR model performed better than any individual model, indicating that the integrative strategy is generally effective. The prediction ability of RCNN in dealing with unknown proteins suffers, as evinced by low AUPRC values in Task2 and Task3, suggesting that sequence information is more sensitive to PPI predictions than GO or domain information.

Model interpretability of DeepAraPPI

We conducted computational experiments to investigate the model interpretability of DeepAraPPI. Out of the three baseline models in DeepAraPPI, RCNN captures raw

Table 1 Overall performance of our models measured by AUPRC for Task1, Task2 and Task3

	Task1	Task2	Task3
RCNN	0.925	0.746	0.481
Domain2vec	0.868	0.780	0.681
Go2vec	0.939	0.871	0.803
Logistic Regression	0.965	0.897	0.825

sequence data through multiple hidden layers to improve classification performance. Using the high-dimensional feature visualization technique, *t*-distributed stochastic neighbor embedding (*t*-SNE), we visualize raw features and sequence pair vectors obtained by multiplying the pair of protein-embedding vectors after convolution. Initially raw data are disorderly distributed (Figure 2a), but positive and negative samples are clearly separated after RCNN processing (Figure 2b), indicating that the RCNN model has effectively learned the features to distinguish whether protein pairs interact.

We further explained the RCNN model based on gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2020). Grad-CAM utilizes the gradients flowing into the convolutional layer to calculate different weights for each neuron without changing the architecture or retraining, pointing to the region of the sequence that ultimately helps the model to make decisions. Here, we used an interacting protein pair (i.e. P56761–P83755) with experimentally verified three-dimensional (3D) structure to further exemplify the interpretability of the model. We used the Grad-CAM technology to calculate the class-discriminative localization maps of two proteins. As a result, a class-discriminative localization value was obtained for each residue in these two proteins. The larger the value, the greater the influence of the underlying residue on the model decision. As annotated by PlaPPISite (Yang, Yang, Qi, et al., 2020), we found that 80 out of the 353 residues in protein P56761 are involved in binding to protein P83755. In turn, 28 out of the top 80 residues that we located in the class-discriminative localization map of P56761 significantly overlapped with the protein-interacting residues (Fisher's exact test, $P = 0.0017$). Whereas protein P83755 has 81 interacting residues over a total of 353 residues, we found 21 overlapping residues in its class-discriminative localization map (Fisher's exact test, $P = 0.090$). Presenting the observed overlap between the residues with top class-discriminative localization values and interacting residues in these two proteins through a surface representation of the 3D structure of the interacting proteins (Figure 2c,d), our example clearly suggests that the RCNN model can effectively capture sequence regions that affect protein interactions to implement its classification task.

As the GO2vec model has the best overall performance among the three predictors, we further explain the GO2vec model. Using the R package GOSemSim (Csardi & Nepusz, 2006), we calculated the GO similarity between two interacting proteins in the independent test set of Task1. Moreover, we obtained fixed dimensional vectors for each protein using the GO2vec model, and further determined the similarity of two interacting proteins using cosine similarity. Notably, we observed a significant correlation between the protein similarity yielded by the GO2vec model and the GO similarity (Figure 2e; $R^2 = 0.6023$, $P = 4.3 \times 10^{-21}$), indicating that the GO2vec endowed each protein with rich semantic information.

Comparison with other methods

As RF classifiers often outperform other conventional machine learning methods, we compared the performance of our deep learning method with RF-based approaches (Chen et al., 2019; Wu et al., 2009; Yang, Yang, Li, et al., 2020). In particular, we used three baseline encoding schemes, auto covariance (AC), conjoint triad (CT) and dipeptide composition (DPC), to represent interacting pairs of proteins. PR curves of the three test tasks clearly indicate that DeepAraPPI outperforms any RF-based methods (Figure 3a–c).

We further compared the performance of our DeepAraPPI platform with three existing Arabidopsis PPI prediction methods, including AraPPINet (https://netbio.sjtu.edu.cn/arappinet) (Zhang et al., 2016), AtPIN (https://atpin.bioinfoguy.net/cgi-bin/atpin.pl) (Brandao et al., 2009) and AtPID (http://119.3.41.228/atpid/webfile) (Li et al., 2011). AraPPINet utilized 3D structure and function information to generate four structural features and seven non-structural features to predict Arabidopsis PPIs based on an RF model (Zhang et al., 2016). AtPIN is a user-friendly resource that aggregates Arabidopsis PPIs, ontology and subcellular localization information and provides Arabidopsis PPI predictions through an interolog mapping method (Brandao et al., 2009). AtPID uses a variety of computational methods to predict PPIs, including interolog mapping, gene expression data, genomic context, gene fusion, phylogenetic profiles and GO annotation. Finally, these prediction data sources are integrated via a Bayesian network method (Li et al., 2011).

As these three existing predictors were published before 2018, we repartitioned the 11 858 high-quality PPIs into two subsets, where we used 8997 PPIs published before 2018 as the positive samples for the training set and 2861 PPIs published after 2018 as the positive samples for the independent test set. We sampled negative data according to the same strategy used in Task1, Task2 and Task3. We retrained our DeepAraPPI method using the newly compiled training data set and obtained its performance on the independent test set.

We submitted all the protein pairs in the independent test set to the websites of these three existing methods to

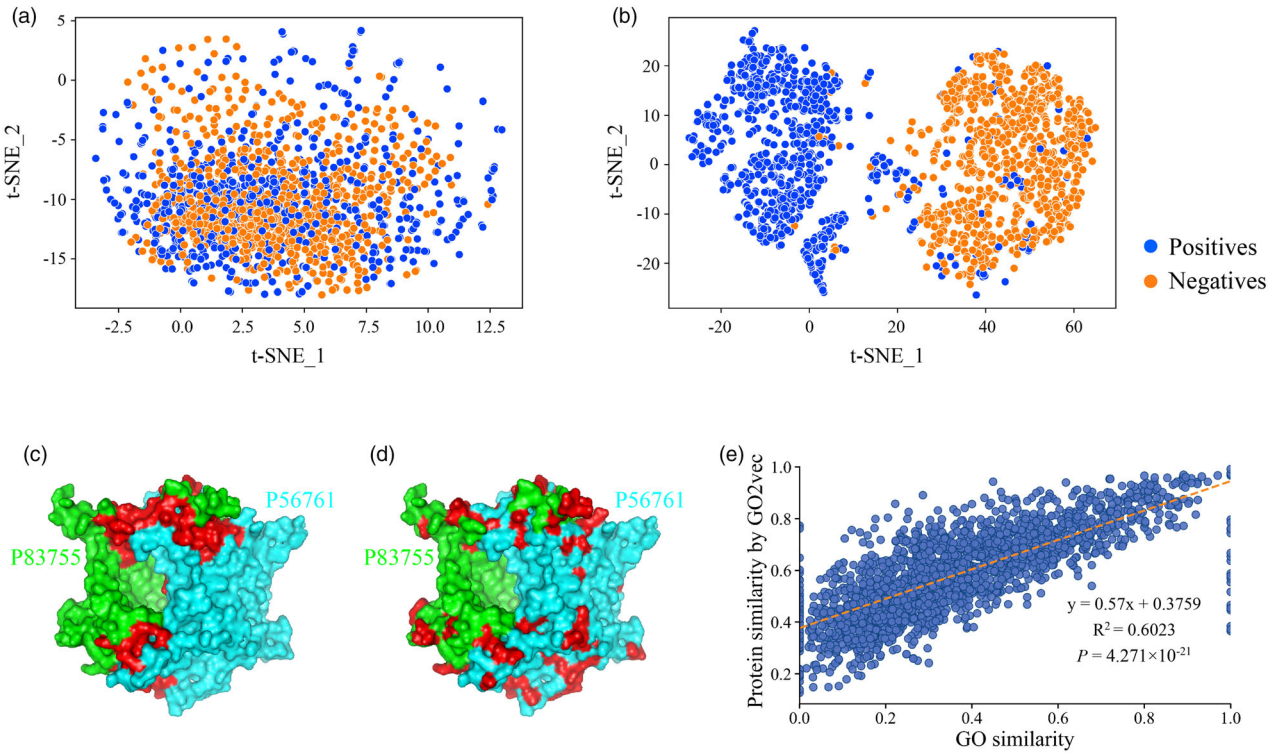


Figure 2. Model INTERPRETABILITY of DeepAraPPI. (a, b) Visualization of features learned from the recurrent convolutional neural network (RCNN) model. We randomly select 1000 positive and 1000 negative samples from the independent test set of Task1, visualizing such samples using the tsnecuda PYTHON library (https://pypi.org/project/tsnecuda). (a) Raw input data. (b) Visualization of sequence pair vectors produced by the multiplication of a pair of protein embedding vectors after convolution. (c, d) Examples showing that RCNN can effectively capture sequence regions relevant to protein interacting residues to implement the classification task. The images present surface representations of the interaction between proteins P56761 (blue) and P83755 (green) (PDB code: 5MDX). The red-colored residues in (c) are interacting residues between P56761 and P83755 (80 residues for P56761 and 81 residues for P83755), whereas the red-colored residues in (d) are the top residues in the class-discriminative localization maps (80 residues for P56761 and 81 residues for P83755). (e) Significant correlation between Gene Ontology (GO) similarity and protein similarity inferred by GO2vec.

obtain the corresponding prediction results. The FPR and TPR values obtained by AraPPINet were 0.063% and 9.1%, respectively. When the FPR value of our proposed model was set at 0.063%, the TPR value of our proposed model was 24.1%, which is considerably higher compared with AraPPINet (Figure 3d). Using a similar strategy, we set the same FPR value as that of AtPIN or AtPID, and our predictive model achieved a much higher TPR value (Figure 3d), clearly indicating that our model outperforms these three state-of-the-art methods.

It is also interesting to benchmark DeepAraPPI against state-of-the-art non-plant-specific PPI predictors. To this end, we compared DeepAraPPI with a recently developed human PPI predictor called D-SCRIPT (Sledzieski et al., 2021). In brief, D-SCRIPT utilizes a pre-trained protein language model to obtain structurally informative protein embeddings as input and implements the PPI prediction through a deep learning architecture. To allow for a fair comparison, we downloaded the source code of D-SCRIPT, and retrained and evaluated the model using the data sets compiled for comparing the three existing plant PPI predictors. As shown in Figure S1, DeepAraPPI performs better than D-SCRIPT (AUPRC = 0.828 vs 0.708).

Cross-species prediction

We collected experimentally verified rice PPIs from four public databases (DIP, MINT, BioGRID and IntAct) and the literature (Wierbowski et al., 2020). After removing self-interactions, non-physical interactions and redundant interactions, 611 rice PPIs between 555 proteins were retained as positive samples. Protein pairs other than known rice PPIs were randomly selected as negative samples, keeping the ratio of positives to negatives at 1:10. We used our deep learning models (i.e. RCNN, Domain2vec, GO2vec and the integrated LR model) that we trained on Arabidopsis data to predict rice PPIs. Specifically, we employed the DPC encoding-based RF model inferred from Arabidopsis as a baseline predictor to assess the cross-species prediction performance.

In Figure 4, we found that the AUPRC values of RCNN, Domain2vec and GO2vec are 0.248, 0.279 and 0.265, respectively. The AUPRC of the integrated LR model is

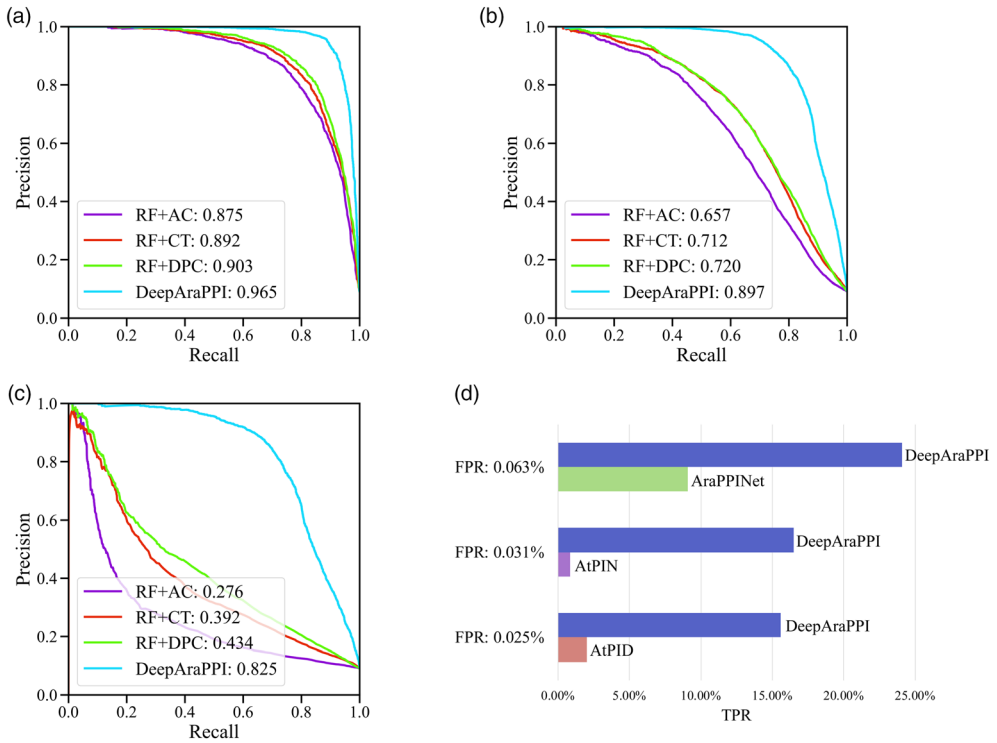


Figure 3. Comparison of DeepAraPPI with other methods. (a–c) Precision–recall curves indicate that DeepAraPPI outperformed random forest (RF) models with three encoding schemes. Panels (a), (b) and (c) display the performance of Task1, Task2 and Task3, respectively. (d) Performance comparison of DeepAraPPI with three existing Arabidopsis PPI prediction methods. By tuning the false-positive rate (FPR) resulting from AraPPINet, AtPIN and AtPID, DeepAraPPI showed a much higher true-positive rate (TPR).

0.305, which is 0.026 higher than the best individual model (i.e. Domain2vec), indicating that the LR integrated model performs better than the three individual models. In comparison, the RF model performs worst, with an AUPRC of 0.171. Although the performance rankings of our models to predict rice PPIs show the same trend compared with Arabidopsis, the current cross-species performance is dramatically inferior to the counterpart in Arabidopsis (cf. Figure 4 and Table 1). We speculate that the decreased performance in cross-species prediction can be partly attributed to the insufficient PPI data in the rice test set. On the other hand, the models trained on Arabidopsis are still highly biased to deal with Arabidopsis proteins and their generalizability to other plant species is limited.

We further randomly divided the data set of rice into a training data set capturing 80% of the data, with the remainder serving as an independent test set, and combined the rice-specific training set with the data set of Arabidopsis as a new hybrid training set. As the GO2vec model has favorable performance among the three predictors, we retrained the GO2vec model on this hybrid training set and predicted the independent test set of rice with an AUPRC of 0.561. Comparatively, the original GO2vec model of Arabidopsis only yielded an AUPRC of 0.285 in

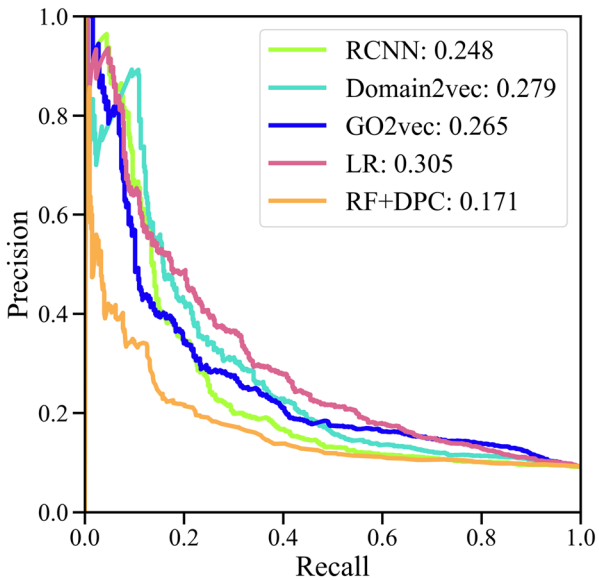


Figure 4. Performance comparison of various classifiers in predicting rice protein–protein interactions (PPIs). Areas under the precision–recall curves (AUPRC) indicate that the logistic regression (LR) integrated model provided the best prediction performance compared with the recurrent convolutional neural network (RCNN), Domain2vec, GO2vec and random forest (RF) models.

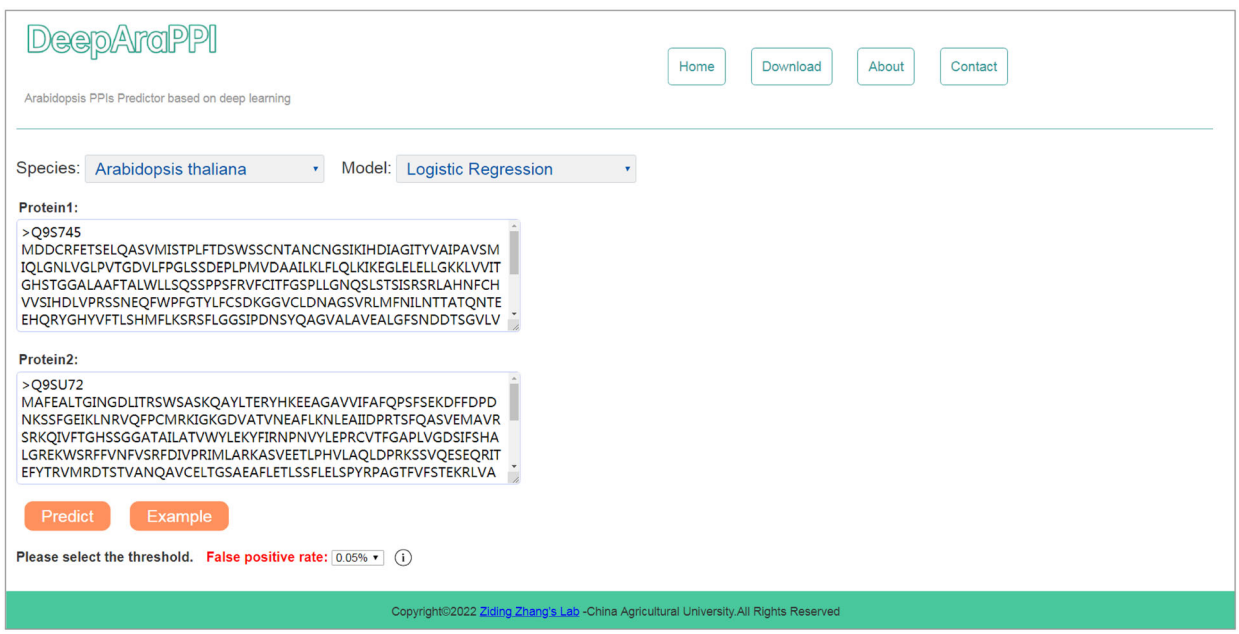
predicting the independent test set of rice, indicating that the prediction performance improves when the training set presents features similar to the independent test set. In real applications, training using a hybrid data set is a potential alternative for predicting PPIs in rice. We wish to emphasize that the generalizability of the model for cross-species application remains an open issue in PPI prediction.

Online prediction platform

We provide an online PPI prediction platform that currently supports the prediction of Arabidopsis and rice (http://zzdlab.com/deeparappi). The webserver is implemented with CentOS 7.4 and Apache 2.4.6. Users can choose RCNN, Domain2vec, GO2vec and the integrated LR model to predict PPIs. We also provide alternative choices of different FPR thresholds. It is important to note here that Domain2vec, GO2vec and the integrated LR model have a pre-trained corpus. If the protein inputted by the user is not in our corpus, it cannot be predicted using the above models. In this case, the user can choose RCNN as RCNN is only based on sequence information. We have also made the source code and all the data sets of DeepAraPPI downloadable from our online platform and Github (https://github.com/zjy1125/DeepAraPPI).

Presenting two cases to guide users to properly access the online platform, we first predict the probability of interaction between the Arabidopsis proteins Q9S745 and Q9SU72 through our webserver (Figure 5a) that has been experimentally observed (Feys et al., 2001; Pruitt et al., 2021). Both proteins are lipase-like proteins that play important roles in plant disease resistance and are essential for plant defense to enhance the accumulation of the molecule salicylic acid (Pruitt et al., 2021). As for the prediction, we first select the species 'Arabidopsis thaliana' and the model 'Logistic Regression', subsequently input the corresponding protein sequences in Fasta format and select an FPR threshold of 0.05%. As shown in Figure 5b, both prediction scores generated by the RCNN and Domain2vec models are relatively low, whereas the score yielded by GO2vec is close to 1. The prediction score of the final LR model is 1, indicating that the two proteins interact. The successful prediction is rooted in the ability of the GO2vec model to capture similar functions of the two proteins. Predicting another experimentally known interaction between proteins F4K5K0 and Q9SDY5 as our second

(a)



(b)

Protein1 ID	Protein2 ID	Model	RCNN score	Domain2vec score	GO2vec score	Prediction score	Interaction
Q9S745	Q9SU72	Logistic Regression	0.0413	0.3468	0.9991	1.0000	Yes

(c)

Protein1 ID	Protein2 ID	Model	RCNN score	Domain2vec score	GO2vec score	Prediction score	Interaction
F4K5K0	Q9SDY5	Logistic Regression	0.3218	0.9130	0.2304	1.0000	Yes

Figure 5. Usage of our online prediction platform. (a) The input page of our web server. (b) Prediction result of the protein pair Q9S745–Q9SU72. (c) Prediction result of the protein pair F4K5K0–Q9SDY5.

case (Figure 5c), we observed that the prediction scores of RCNN and GO2vec are relatively low, whereas the Domain2vec model contributes a high score (i.e. 0.913). In particular, the known interacting domain pair ('UQ_con'–'zf-rbx1') was found between the two interacting proteins, which suggests that the Domain2vec model can accurately predict interactions mediated by domain–domain interactions (DDIs). Taken together, the DeepAraPPI webserver is easy to use, and can jointly utilize sequence, domain and GO information to maximize predictive performance.

CONCLUSION

To provide a reliable tool that allows PPI predictions in plants, we propose a deep learning model for predicting PPIs in Arabidopsis. Our DeepAraPPI framework effectively integrates sequence, domain and GO information to represent protein features. The benchmarking experiments demonstrate that DeepAraPPI performs well on Arabidopsis PPI data sets at different difficulty levels. To further verify its effectiveness, we compare the performance of DeepAraPPI with multiple competitive baseline methods, indicating that DeepAraPPI considerably outperforms existing state-of-the-art Arabidopsis PPI prediction methods. At the same time, we verify the feasibility of the current strategy for cross-species prediction performance. Although DeepAraPPI performs better than traditional machine learning methods in predicting rice PPIs, its performance is still far behind that of Arabidopsis, meaning that there is still room for improvement in real applications. In the future, with the increasing availability of experimental PPI data and other data associated with protein interactions, it will be possible to achieve more accurate cross-species prediction. First, with the emergence of Alphafold2 (Jumper et al., 2021), protein structural information is easily accessible, which can be used as an important input feature to develop predictive models with better performance in cross-species prediction. Second, with the advent of large pre-trained protein language models (Rives et al., 2021), protein sequences can be converted to semantically rich feature representations, which have been successfully used for diverse protein bioinformatics prediction tasks (Rives et al., 2021). The application of protein language models in predicting PPIs is promising (Sledzieski et al., 2021; Yang, Yang, Li, et al., 2020), and further efforts are needed to explore the potential in cross-species PPI prediction. Last but not least, the high-throughput experimental determination of plant PPIs is proceeding at an accelerating rate, which will directly boost the development of plant PPI predictors. For instance, the recently published maize PPIs (Han et al., 2023) will become an important data resource to train new plant PPI predictors with improved performance in cross-species settings.

EXPERIMENTAL PROCEDURES

Data set construction and partition

We collected experimentally verified PPIs of Arabidopsis from BioGRID (https://thebiogrid.org) (Chatr-Aryamontri et al., 2017), DIP (https://dip.doe-mbi.ucla.edu/dip/Main.cgi) (Salwinski et al., 2004), IntAct (https://www.ebi.ac.uk/intact) (Orchard et al., 2014), MINT (https://mint.bio.uniroma2.it) (Licata et al., 2012) and TAIR (https://www.arabidopsis.org) (Lamesch et al., 2012). To unify protein IDs from these different databases, protein IDs of different types were converted to UniProt IDs. We further discarded self-interactions, redundant interactions, non-physical interactions and PPIs containing proteins with fewer than 40 amino acids or with non-standard amino acids. Finally, we obtained 49 398 experimentally verified PPIs between 10 330 Arabidopsis proteins.

To obtain high-quality PPIs as positive samples, we adopted the Human Integrated Protein–Protein Interaction rEference (HIPPIE) scoring scheme to assess the confidence of the collected Arabidopsis PPIs (Schaefer et al., 2012). For each PPI, a quality score ranging from 0 to 1 was assigned by accounting for: (i) the experimental methods for the PPI determination; (ii) the number of articles in the literature reporting the PPI; and (iii) the species included in the PPI. Preliminarily, we determined a threshold of 0.72 to classify 49 398 PPIs into a high-quality subset (11 858 PPIs with a score of $\geq$0.72) and a low-quality subset (37 540 PPIs with a score of $<$0.72). More details about the justification of the threshold value (0.72) are available in Figure S2 and Table S1.

We downloaded the reference proteome sequences of Arabidopsis from UniProt (https://www.uniprot.org) (The UniProt Consortium, 2021) and removed protein sequences with fewer than 40 amino acids or non-standard amino acids, totaling 28 361 sequences. We sampled negative training data by randomly selecting protein pairs from this pool. In particular, we constrained the sample proteins to neither share the same subcellular localization nor belong to the pool of known interactions. Furthermore, we set the ratio of positive to negative samples at 1:10.

Considering that the PPI prediction is based on paired input, its performance is significantly affected by different data set partitions (Park & Marcotte, 2012). Therefore, we designed three tasks at different difficulty levels to assess the prediction performance of our model. As for Task1, we considered 11 858 high-quality PPIs as positive samples, whereas we randomly sampled 118 580 protein pairs as negative samples following our negative sampling strategy. Finally, training data comprise a random sample of 80% of the PPIs, with the remaining 20% of PPIs serving as an independent test set. As for Task2 and Task3, we followed the data set partition method proposed by (Park & Marcotte, 2012). First, the high-quality PPI data set was segmented into three subsets (C1, C2 and C3), where the corresponding number of PPIs are 2844, 6005 and 3009, respectively. Subset C1 is used as the positive training set in Task2 and Task3. PPIs in C2 were used as positive samples of the independent test set in Task2. Note that each PPI in C2 shared only one protein with C1. The PPIs in C3 were used as positive samples of the independent test set in Task3, where each PPI in C3 does not share any protein with C1. Negative samples in C1, C2 and C3 still follow the sampling strategy of Task1.

RCNN-based predictor

We employ a Siamese RCNN model to predict Arabidopsis PPIs based on sequence information. In particular, the RCNN predictor consists of a pre-trained amino acid embedding module, a mixed neural network module of CNN and gated recurrent units (GRUs),

and a prediction module (Figure 1b). To represent protein sequences as feature vectors, RCNN pre-trains the embedding of 20 standard amino acids based on word2vec. The CNN model is used to extract sequence features, whereas the GRU model captures long-term dependency information of the sequence. Based on the representations generated in the above steps, the prediction module outputs a prediction score through a soft-max function, assessing the interaction probability of a pair of proteins.

Pre-trained amino acid embedding module

In more detail, word2vec transforms amino acids of the protein sequences into numerical embedding vectors. Specifically, we utilized protein sequences from the Uniref50 database and adopted the continuous bag-of-words (CBOW) model architecture to train the word2vec model, allowing us to represent each word by surrounding context words. The word2vec model training was implemented through the PYTHON library Gensim (https://pypi.org/project/gensim). The prediction results of fivefold cross-validation in Task1 were used to evaluate the performance of word2vec with different hyper-parameters, allowing us to find an optimal window size = 3 and represent each amino acid through a 32-dimensional embedding vector. Furthermore, we truncated long sequences into a fixed length L (i.e. 2000) and zero-padding short sequences (Min et al., 2017), representing each protein as an $L \times 32$ array.

Siamese CNN–GRU module

As stacking multiple CNN and GRU layers allows us to better capture patterns of interacting proteins, we construct a Siamese CNN–GRU architecture with two identical CNN–GRU subnetworks sharing the same parameters (Chen et al., 2019; Hashemifar et al., 2018). High-dimensional local features are first captured by a CNN, which are further modeled by a GRU unit to reflect sequential and contextualized information.

In particular, we use 1D convolution layers with a kernel size of 3 to extract the feature of an $n \times s$ dimensional array X at any position, where n represents length and s is the dimension of the features (i.e. channels), set to 50. After each convolution operation, the maximum pooling layer is used to reduce the dimension and ensure the invariance of features. The GRU unit is a variant of long short-term memory (LSTM) modules (Cho et al., 2014). However, unlike LSTMs, GRU does not introduce additional memory units. The GRU introduces an update gate to manage how much information the current state needs to retain from the historical state and how much new information it needs to receive from the candidate state. To better capture context information, we use the bidirectional GRU layer to process data from the 50 channels of the previous CNN.

Prediction module

The prediction module is composed of one multiplication layer and a subsequent MLP with three fully connected layers. First, element-wise multiplication is performed on the embedding vectors of a pair of proteins to eliminate the bias caused by the order of protein input in the pairwise output. Then, a sequence pair vector is obtained as the input of the fully connected layers in the MLP to calculate the probability that two proteins interact using the softmax function.

Domain2vec-based predictor

Considering that the physical interaction between two proteins is often mediated by DDI, we also developed a Domain2vec-based predictor using domain information as the input. To this end, we downloaded all known DDI information from the 3did database (http://3did.irbbarcelona.org) (Mosca et al., 2014) and annotated the domain information of proteins through HMMER (http://hmmer.org) (Potter et al., 2018). We integrated the DDIs and protein domain annotation to generate a network consisting of 37 342 nodes and 80 123 edges as the input to node2vec (Grover & Leskovec, 2016) to effectively convert each protein into a feature vector. Inspired by the Skip-Gram model, node2vec takes nodes as words and extracts node sequences from the underlying network as sentences and transforms the network into a document with an ordered sequence of nodes. In our case, we integrate DDI and domain annotations of proteins into a network (Figure 1c), where nodes are proteins or domains. Edges represent DDIs or protein–domain annotations (i.e. if protein A contains domain m, the association between protein A and domain m is represented by an edge). After training the node2vec model, each protein is represented by an embedding vector concatenated into a sequence pair vector, which serves as the input of an MLP containing three fully connected layers to predict the interaction probability of a protein pair.

GO2vec-based predictor

The GO annotation system establishes a standard functional vocabulary for genes and their products, which can be regarded as a directed acyclic graph structure. Each node (i.e. GO term) in the annotation system is a functional description of a gene or protein, connected by three strict relationships between nodes (i.e. 'is_a', 'part_of' and 'regulates') (Zhong et al., 2019). We integrated the relationships between GO terms and GO annotations of proteins into a hybrid network (i.e. GO graph) of 92 359 nodes and 294 460 edges, where each node is embedded in a vector after training with node2vec. After concatenating the embedding vectors of two proteins into a protein pair vector, we used an MLP with three fully connected layers to calculate the probability of whether two proteins interact.

Logistic regression model

To maximize prediction performance, we combined these individual scores (i.e. S_{RCNN}, $S_{\mathrm{Domain2vec}}$ and S_{GO2vec}) into a vector and trained an LR model reflecting the overall interaction probability of each query protein pair. The LR algorithm implemented in this study was based on the scikit-learn PYTHON library (https://scikit-learn.org) (Pedregosa et al., 2011), using L2 penalty and a linear solver. The optimal hyperparameters of the model were determined by using the GridSearchCV function with fivefold cross-validation.

RF model

As a benchmark of traditional machine learning-based prediction methods, we also constructed RF predictive models. The basic units of RF are decision trees, achieving predictions through the votes from all decision trees. We set the number of optimal trees in the forest (n_estimators) to 79 and retained the other default parameters. In this work, the RF algorithm was also implemented using the scikit-learn PYTHON library (https://scikit-learn.org) (Pedregosa et al., 2011). Here, the RF predictive models mainly accounted for three sequence-based encoding schemes, including AC, CT and DPC.

Auto covariance (AC)

Auto covariance (AC) encoding represents the properties of amino acids using seven indices, including hydrophobicity, hydrophilicity, polarity, polarizability, side-chain volumes of amino

acids, solvent-accessible surface area and net charge index of residue side chains (Guo et al., 2008). Furthermore, AC encoding accounts for neighboring effects between amino acids at a certain distance. In particular, we determine the AC score of a protein P of length L as:

$$S_{AC}(lag, L) = \frac{1}{L-lag}\sum_{i=1}^{L-lag}\left(R_{i,j}-\frac{1}{L}\sum_{k=1}^{L}R_{k,j}\right) \times\left[R_{(i+lag),j}-\frac{1}{L}\sum_{k=1}^{L}R_{k,j}\right],$$

where *lag* represents the sequence distance between residues. $R_{i,j}$ and $R_{k,j}$ denotes the j^{th} physicochemical property value of the i^{th} residue and k^{th} residue, respectively. Here, we set *lag* ranging from 1 to 30, allowing us to transform a protein pair into a $30 \times 7 \times 2 = 420$ dimensional vector.

Conjoint triad (CT)

After binning 20 amino acids into seven groups based on the physicochemical properties of residue side chains, CT considers the proportions of three consecutive amino acid groups in a protein sequence (Shen et al., 2007), defined as:

$$S_{CT}(G_iG_jG_k) = \frac{N_{G_iG_jG_k}}{L-2}, i, j, k \in (1, 2, \ldots, 7),$$

where G_i, G_j and G_k represent the groups of residue i, j and k, $N_{G_iG_jG_k}$ denotes the number of the CT ($G_iG_jG_k$) in the sequence and L is the length of the sequence. As a consequence, each protein pair is represented by a $7 \times 7 \times 7 \times 2 = 686$ dimensional vector.

Dipeptide composition (DPC)

The DPC represents the proportion of two consecutive amino acids in a protein (Zhou et al., 2012), defined as:

$$S_{DPC}(A_iA_j) = \frac{N_{A_iA_j}}{L-1}, i, j \in (1, 2, \ldots, 20),$$

where A_i and A_j represent two of the 20 amino acids, $N_{A_iA_j}$ denotes the number of dipeptide (A_iA_j) in the sequence and L is the length of the sequence. As a consequence, a protein pair is represented by a $20 \times 20 \times 2 = 800$ dimensional vector.

ACKNOWLEDGEMENTS

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

CONFLICT OF INTEREST

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

DATA AVAILABILITY STATEMENT

All relevant data can be found at http://zzdlab.com/deeparappi/download.html. The code is available at https://github.com/zjy1125/DeepAraPPI.

SUPPORTING INFORMATION

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

Figure S1. Precision–recall curves of DeepAraPPI and D-SCRIPT.

Figure S2. The PPI distribution for different quality scores.

Table S1. Overlap of the Y2H data set with two scored PPI subsets.

REFERENCES

Arabidopsis Interactome Mapping Consortium. (2011) Evidence for network evolution in an arabidopsis interactome map. *Science*, **333**, 601–607.

Berggard, T., Linse, S. & James, P. (2007) Methods for the detection and analysis of protein-protein interactions. *Proteomics*, **7**, 2833–2842.

Brandao, M.M., Dantas, L.L. & Silva-Filho, M.C. (2009) AtPIN: Arabidopsis thaliana protein interaction network. *BMC Bioinformatics*, **10**, 454.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N.K. *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, **45**, D369–d379.

Chen, M., Ju, C.J., Zhou, G., Chen, X., Zhang, T., Chang, K.W. *et al.* (2019) Multifaceted protein-protein interaction prediction based on siamese residual RCNN. *Bioinformatics*, **35**, i305–i314.

Cho, K., Merrienboer, B.V., Gulcehre, C., Ba Hdanau, D., Bougares, F., Schwenk, H. *et al.* (2014) *Learning phrase representations using rnn encoder–decoder for statistical machine translation*. Doha, Qatar: Association for Computational Linguistics, pp. 1724–1734. Available from: https://doi.org/10.3115/v1/D14-1179

Csardi, G. & Nepusz, T. (2006) The igraph software package for complex network research. *Interjournal Complex Systems*, **1695**, 1–9.

Cui, J., Li, P., Li, G., Xu, F., Zhao, C., Li, Y. *et al.* (2008) ATPID: Arabidopsis thaliana protein interactome database–an integrative platform for plant systems biology. *Nucleic Acids Research*, **36**, D999–D1008.

De Bodt, S., Proost, S., Vandepoele, K., Rouze, P. & Van de Peer, Y. (2009) Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression. *BMC Genomics*, **10**, 288.

Ding, Z. & Kihara, D. (2019) Computational identification of protein-protein interactions in model plant proteomes. *Scientific Reports*, **9**, 8740.

Dong, S., Lau, V., Song, R., Ierullo, M., Esteban, E., Wu, Y. *et al.* (2019) Proteome-wide, structure-based prediction of protein-protein interactions/new molecular interactions viewer. *Plant Physiology*, **179**, 1893–1907.

Du, X., Sun, S., Hu, C., Yao, Y., Yan, Y. & Zhang, Y. (2017) DeepPPI: boosting prediction of protein-protein interactions with deep neural networks. *Journal of Chemical Information and Modeling*, **57**, 1499–1510.

Feys, B.J., Moisan, L.J., Newman, M.A. & Parker, J.E. (2001) Direct interaction between the Arabidopsis disease resistance signaling proteins, eds1 and pad4. *The EMBO Journal*, **20**, 5400–5411.

Geisler-Lee, J., O'Toole, N., Ammar, R., Provart, N.J., Millar, A.H. & Geisler, M. (2007) A predicted interactome for arabidopsis. *Plant Physiology*, **145**, 317–329.

Grover, A. & Leskovec, J. (2016) Node2vec: scalable feature learning for networks. *KDD*, **2016**, 855–864.

Gul, S. & Hadian, K. (2014) Protein-protein interaction modulator drug discovery: past efforts and future opportunities using a rich source of low- and high-throughput screening assays. *Expert Opinion on Drug Discovery*, **9**, 1393–1404.

Guo, Y., Yu, L., Wen, Z. & Li, M. (2008) Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Research*, **36**, 3025–3030.

Han, L., Zhong, W., Qian, J., Jin, M., Tian, P., Zhu, W. *et al.* (2023) A multi-omics integrative network map of maize. *Nature Genetics*, **55**, 144–153.

Hashemifar, S., Neyshabur, B., Khan, A.A. & Xu, J. (2018) Predicting protein-protein interactions through sequence-based deep learning. *Bioinformatics*, **34**, i802–i810.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

Keskin, O., Tuncbag, N. & Gursoy, A. (2016) Predicting protein-protein interactions from the molecular to the proteome level. *Chemical Reviews*, **116**, 4884–4909.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R. *et al.* (2012) The Arabidopsis information resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, **40**, D1202–D1210.

Li, P., Zang, W., Li, Y., Xu, F., Wang, J. & Shi, T. (2011) AtPID: the overall hierarchical functional protein interaction network interface and analytic platform for Arabidopsis. *Nucleic Acids Research*, **39**, D1130–D1133.

Lian, X., Yang, X., Yang, S. & Zhang, Z. (2021) Current status and future perspectives of computational studies on human-virus protein-protein interactions. *Briefings in Bioinformatics*, **22**, bbab029.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E. *et al.* (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, **40**, D857–D861.

Lin, M., Hu, B., Chen, L., Sun, P., Fan, Y., Wu, P. *et al.* (2009) Computational identification of potential molecular interactions in Arabidopsis. *Plant Physiology*, **151**, 34–46.

Min, X., Zeng, W., Chen, N., Chen, T. & Jiang, R. (2017) Chromatin accessibility prediction via convolutional long short-term memory networks with k-mer embedding. *Bioinformatics*, **33**, i92–i101.

Mosca, R., Céol, A., Stein, A., Olivella, R. & Aloy, P. (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*, **42**, D374–D379.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F. *et al.* (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, **42**, D358–D363.

Pan, J., You, Z.H., Li, L.P., Huang, W.Z., Guo, J.X., Yu, C.Q. *et al.* (2022) DWPPI: a deep learning approach for predicting protein-protein interactions in plants based on multi-source information with a large-scale biological network. *Frontiers in Bioengineering and Biotechnology*, **10**, 807522.

Pan, X., Zuallaert, J., Wang, X., Shen, H.B., Campos, E.P., Marushchak, D.O. *et al.* (2021) ToxDL: deep learning using primary structure and domain embeddings for assessing protein toxicity. *Bioinformatics*, **36**, 5159–5168.

Park, Y. & Marcotte, E.M. (2012) Flaws in evaluation schemes for pair-input computational predictions. *Nature Methods*, **9**, 1134–1136.

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.

Peng, X., Wang, J., Peng, W., Wu, F.X. & Pan, Y. (2017) Protein-protein interactions: detection, reliability assessment and applications. *Briefings in Bioinformatics*, **18**, 798–819.

Petschnigg, J., Snider, J. & Stagljar, I. (2011) Interactive proteomics research technologies: recent applications and advances. *Current Opinion in Biotechnology*, **22**, 50–58.

Petta, I., Lievens, S., Libert, C., Tavernier, J. & De Bosscher, K. (2016) Modulation of protein-protein interactions for the development of novel therapeutics. *Molecular Therapy*, **24**, 707–718.

Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R. & Finn, R.D. (2018) HMMER web server: 2018 update. *Nucleic Acids Research*, **46**, W200–w204.

Pruitt, R.N., Locci, F., Wanke, F., Zhang, L., Saile, S.C., Joe, A. *et al.* (2021) The EDS1-PAD4-ADR1 node mediates Arabidopsis pattern-triggered immunity. *Nature*, **598**, 495–499.

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**, e2016239118.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. & Eisenberg, D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Research*, **32**, D449–D451.

Schaefer, M.H., Fontaine, J.F., Vinayagam, A., Porras, P., Wanker, E.E. & Andrade-Navarro, M.A. (2012) HIPPIE: integrating protein interaction networks with experiment based quality scores. *PLoS One*, **7**, e31826.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2020) Grad-cam: visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, **128**, 336–359.

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**, 4337–4341.

Shin, W.H., Christoffer, C.W. & Kihara, D. (2017) In silico structure-based approaches to discover protein-protein interaction-targeting drugs. *Methods*, **131**, 22–32.

Skrabanek, L., Saini, H.K., Bader, G.D. & Enright, A.J. (2008) Computational prediction of protein-protein interactions. *Molecular Biotechnology*, **38**, 1–17.

Sledzieski, S., Singh, R., Cowen, L. & Berger, B. (2021) D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Systems*, **12**, 969–982.

Sun, T., Zhou, B., Lai, L. & Pei, J. (2017) Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics*, **18**, 277.

The UniProt Consortium. (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, **49**, D480–D489.

Wierbowski, S.D., Vo, T.V., Falter-Braun, P., Jobe, T.O., Kruse, L.H., Wei, X. *et al.* (2020) A massively parallel barcoded sequencing pipeline enables generation of the first ORFeome and interactome map for rice. *Proceedings of the National Academy of Sciences of the United States of America*, **117**, 11836–11842.

Wu, C., Gao, R., Zhang, Y. & De Marinis, Y. (2019) PTPD: predicting therapeutic peptides by deep learning and word2vec. *BMC Bioinformatics*, **20**, 456.

Wu, J., Liu, H., Duan, X., Ding, Y., Wu, H., Bai, Y. *et al.* (2009) Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics*, **25**, 30–35.

Xu, F., Li, G., Zhao, C., Li, Y., Li, P., Cui, J. *et al.* (2010) Global protein interactome exploration through mining genome-scale data in Arabidopsis thaliana. *BMC Genomics*, **11**(Suppl 2), S2.

Yang, X., Yang, S., Li, Q., Wuchty, S. & Zhang, Z. (2020) Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. *Computational and Structural Biotechnology Journal*, **18**, 153–161.

Yang, X., Yang, S., Qi, H., Wang, T., Li, H. & Zhang, Z. (2020) Plappisite: a comprehensive resource for plant protein-protein interaction sites. *BMC Plant Biology*, **20**, 61.

Zeng, M., Li, M., Wu, F.X., Li, Y. & Pan, Y. (2019) DeepEP: a deep learning framework for identifying essential proteins. *BMC Bioinformatics*, **20**, 506.

Zhang, F., Liu, S., Li, L., Zuo, K., Zhao, L. & Zhang, L. (2016) Genome-wide inference of protein-protein interaction networks identifies crosstalk in abscisic acid signaling. *Plant Physiology*, **171**, 1511–1522.

Zhong, X., Kaalia, R. & Rajapakse, J.C. (2019) GO2vec: transforming GO terms and proteins to vector representations via graph embeddings. *BMC Genomics*, **20**, 918.

Zhou, Y., Zhou, Y.S., He, F., Song, J. & Zhang, Z. (2012) Can simple codon pair usage predict protein-protein interaction? *Molecular BioSystems*, **8**, 1396–1404.