Plant Methods

RESEARCH Open Access

Pre-trained protein language model sheds new light on the prediction of Arabidopsis protein–protein interactions

Kewei Zhou[1], Chenping Lei[1], Jingyan Zheng[1], Yan Huang[1] and Ziding Zhang[1*]

Abstract

Background Protein–protein interactions (PPIs) are heavily involved in many biological processes. Consequently, the identification of PPIs in the model plant Arabidopsis is of great significance to deeply understand plant growth and development, and then to promote the basic research of crop improvement. Although many experimental Arabidopsis PPIs have been determined currently, the known interactomic data of Arabidopsis is far from complete. In this context, developing effective machine learning models from existing PPI data to predict unknown Arabidopsis PPIs conveniently and rapidly is still urgently needed.

Results We used a large-scale pre-trained protein language model (pLM) called ESM-1b to convert protein sequences into high-dimensional vectors and then used them as the input of multilayer perceptron (MLP). To avoid the performance overestimation frequently occurring in PPI prediction, we employed stringent datasets to train and evaluate the predictive model. The results showed that the combination of ESM-1b and MLP (i.e., ESMAraPPI) achieved more accurate performance than the predictive models inferred from other pLMs or baseline sequence encoding schemes. In particular, the proposed ESMAraPPI yielded an AUPR value of 0.810 when tested on an independent test set where both proteins in each protein pair are unseen in the training dataset, suggesting its strong generalization and extrapolating ability. Moreover, the proposed ESMAraPPI model performed better than several state-of-the-art generic or plant-specific PPI predictors.

Conclusion Protein sequence embeddings from the pre-trained model ESM-1b contain rich protein semantic information. By combining with the MLP algorithm, ESM-1b revealed excellent performance in predicting Arabidopsis PPIs. We anticipate that the proposed predictive model (ESMAraPPI) can serve as a very competitive tool to accelerate the identification of Arabidopsis interactome.

Keywords Arabidopsis, Protein–protein interactions, Machine learning, Pre-trained language model, Natural language processing

Background

Protein–protein interactions (PPIs) are heavily involved in cellular biological processes, including signal transduction, transcriptional activation, and regulations of expression and metabolism [1]. Thus, it is critical to identify whether two proteins interact or not to help understand protein functions. Traditional experiments [e.g., isothermal titration calorimetry [2], pull-down assay [3], and surface plasmon resonance [4]] are low-throughput

*Correspondence:
Ziding Zhang
zidingzhang@cau.edu.cn
[1] State Key Laboratory of Animal Biotech Breeding, College of Biological Sciences, China Agricultural University, Beijing 100193, China

and time-consuming. With the development of high-throughput techniques, such as in vitro yeast two-hybrid screening [5] and affinity purification coupled with mass spectrometry [6], the identification of PPI data has been significantly accelerated, and the bioinformatics applications of PPI data have also been widely explored [7]. On the one hand, the experimental PPI data are often compiled as PPI interaction networks [8–10], and functionally unknown proteins in the networks can be annotated through network clustering and analysis [11–13]. On the other hand, the experimental PPI data can also be used to train PPI prediction models [14–16]. In this regard, machine learning is an increasingly popular computational method to learn data features deposited in known PPIs and build predictive models to predict unknown interactions.

Since protein interactions are mainly determined by their primary sequences, many efforts have focused on developing sequence-based PPI predictors. To build a machine learning model for PPI prediction, the key step is conducting feature engineering, which converts protein sequences into fixed-dimensional vectors. The frequently used sequence encoding schemes include amino acid composition (AAC), dipeptide composition (DPC), conjoint triad (CT), and composition of k-spaced amino acid pairs (CKSAAP). These descriptive representations are often combined with traditional machine learning methods, such as support vector machine (SVM) [17] and random forest (RF) [18], to develop effective PPI predictors. Regarding deep learning methods, primitive information without feature engineering can be used to extract more abstract representations. For instance, one-hot encoding and position-specific scoring matrix (PSSM) representation have been integrated with the framework of convolutional neural network (CNN) to achieve higher performance [19, 20].

By analogy to natural language, a protein sequence can be deemed a sentence, in which residue segments are regarded as words. Based on this hypothesis, natural language processing (NLP) methods have been used in protein representation. For instance, the typical word/sentence embedding techniques in NLP (e.g., word2vec and doc2vec) have been applied to protein sequence representations [21–23]. Although the word2vec/doc2vec models are either too shallow or trained with the corpus containing a limited number of existing proteins, they have revealed very promising results in many protein bioinformatics tasks. As one of the self-supervised language models, Transformer, released in 2017 by Google [24], solved the problem of memory capacity and processing speed. A typical Transformer is comprised of Attention modules focusing on vital information from global to local, and it often showed significant performance improvement when trained on large datasets [25]. Considering the advantages of Transformer in NLP, Rives et al. used this technique to generate a protein language model (pLM) for the purpose of protein sequence embeddings [26]. ESM is a deep Transformer language model trained on UniRef50, which can learn multiscale representations, including biochemical properties, remote homology, and alignment within a protein family. Researchers have applied similar representation in different prediction tasks [27–31].

Large-scale identification of PPIs in the model plant Arabidopsis is of significance to decipher plant gene regulatory relationships, deeply understand plants' growth and development, and promote the basic research of crop improvement and breeding. Although many known Arabidopsis PPIs have been stored in public databases, the Arabidopsis interactome remains incomplete. Thus, developing effective machine learning methods trained on existing PPI data to promptly predict unknown PPIs will accelerate the determination of Arabidopsis interactome data, reduce the experimental cost and provide new hints for plant functional genomics. To our best knowledge, however, the pLM-based embeddings have not been employed for predicting Arabidopsis PPIs. Here, we further explored the application of ESM representation in predicting Arabidopsis PPIs. Through a series of computational experiments, we observed that the combination of ESM-1b representation (one representative ESM model) with multilayer perceptron (MLP), termed ESMAraPPI, yielded more powerful performance than the predictive models inferred from other pLMs or conventional sequence features. In the meantime, the proposed model also revealed better performance than several generic or plant-specific PPI predictors.

Results and discussion

The computational framework and benchmarking datasets of ESMAraPPI

The flowchart of the proposed prediction method is illustrated in Fig. 1. To train and assess the performance of different PPI prediction models, we collected high-quality experimental Arabidopsis PPIs as positive samples. Additionally, we compiled negative training data by randomly selecting Arabidopsis protein pairs, and the ratio of positive to negative samples was set as 1:10. To train and evaluate model performance, we followed Park and Marcotte's advice [32] to construct one training dataset (i.e., C1) and two independent test sets (i.e., C2 and C3) (Fig. 1). Specifically, only one protein in each pair from C2 is allowed to be appeared in C1, whereas both proteins in each pair from C3 are unseen in C1. The representation of a protein was extracted from ESM-1b directly, which resulted in a feature vector of 1280 dimensionality. Since

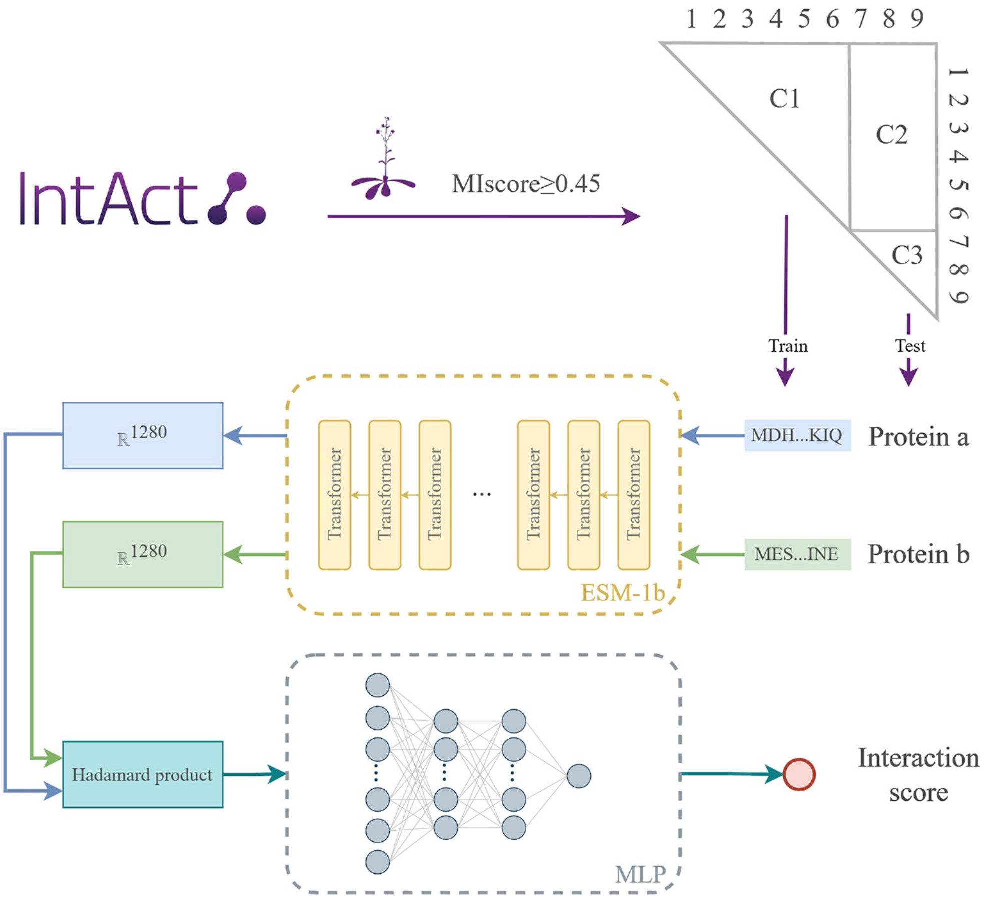

Fig. 1 The schematic diagram of ESMAraPPI. Arabidopsis PPIs from the IntAct database with MIscore ≥ 0.45 were collected as positive samples. We also compiled 10 times negative samples to construct an original dataset. Then, we divided the original dataset into three datasets (i.e., C1, C2 and C3). C1 was the training dataset, while C2 and C3 were two independent test datasets. The representations of protein pair were extracted from ESM-1b, and Hadamard product was applied before inputting to a 4-layer MLP. The final output was an interaction score between 0 and 1 (a prediction score ≥ 0.5 corresponded to a positive interaction)

the PPI prediction is a pair-input problem, Hadamard product was applied before inputting representations of protein pairs to a 4-layer MLP for model training. More details about the dataset preparation, pLM feature vector construction, and machine learning algorithm implementation are available in the Methods section.

ESM-1b coupled with MLP performed best in predicting Arabidopsis PPIs

Nine different pLMs from ESM embed each protein sequence to a vector of 1280 dimensionality. We combined these pLMs with three machine learning algorithms (MLP, RF and SVM) to seek the best combination. The 4-layer MLP computational framework, which contains 1024, 512, 128, and 16 nodes, was optimally selected. Concerning RF and SVM, the corresponding parameters were optimized through grid search. Considering that positives and negative samples are highly imbalanced in this work, we mainly quantified the performance by plotting the precision–recall (PR) curve and calculating the corresponding area under the PR curve (AUPR). As shown in Fig. 2, MLP-based models yielded the highest AUPR values, followed by RF- and SVM-based models. In particular, the combination of MLP and ESM-1b (i.e., esm1b_t33_650M_UR50S) achieved the best performance (AUPR = 0.834 on C2 and 0.810 on C3). To supplement the AUPR-based assessment, we also plotted the receiver operating characteristic (ROC) curve and calculated the corresponding area under the ROC curve (AUROC) for each combination. Again, the MLP and ESM-1b combination resulted in the largest AUROC value (Additional file 1: Fig. S1).

We further compared ESM-1b with three pLMs, ProtTrans, UniRep, and TAPE. Note that these three pLMs were trained using different training strategies from ESM-1b. Trained on data from UniRef and BFD covering up to 393 billion amino acids, an auto-encoder model (ProtT5-XL-U50) from ProtTrans for the first time outperformed

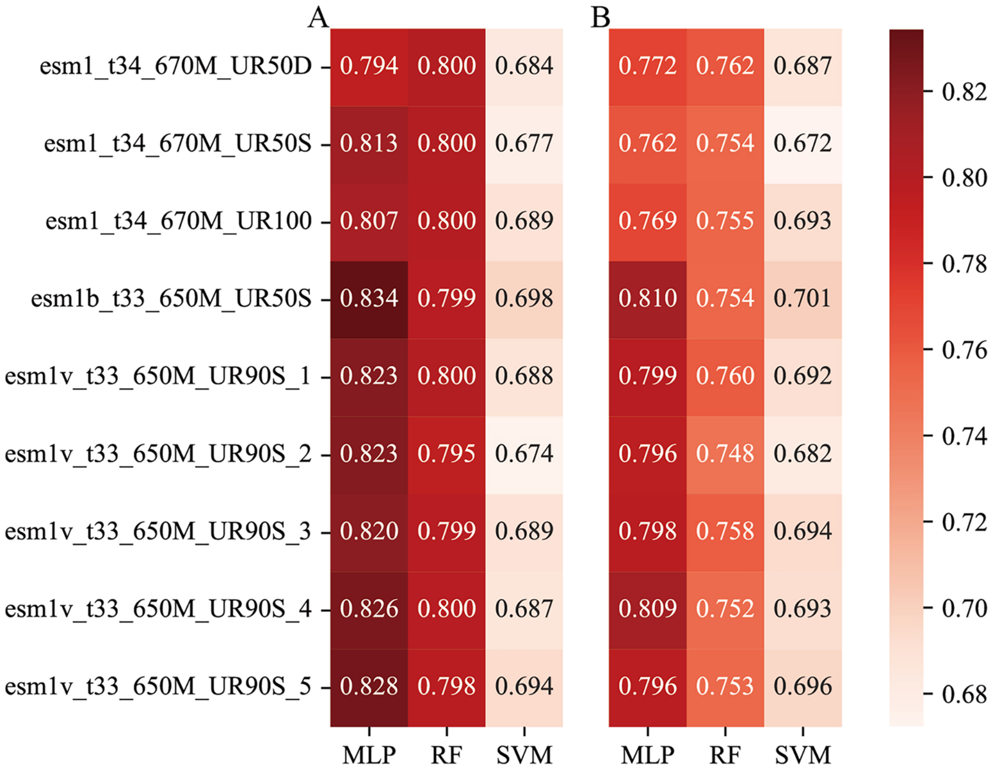

Fig. 2 AUPR values of combinations between nine pLMs from ESM and three machine learning algorithms. Of the different ESM models, ESM-1v was fine-tuned for predicting variant effects and contained five models with different random seeds. ESM-1b differs from ESM-1 mainly in higher learning rate, dropout after word embedding, learned positional embeddings, final layer norm before the output, and tied input/output word embedding. **A** The results from the independent dataset C2 where only one protein in each pair appeared in the training dataset (i.e., C1), while **B** corresponds to the results from the independent dataset C3 where no protein in each pair appeared in the training dataset (i.e., C1)

existing methods without the need of multiple sequence alignments (MSAs) or evolutionary information in secondary structure prediction [33]. UniRep was based on multiplicative LSTM (mLSTM) and was trained on UniRef50 [34]. It was found that the amino-acid embeddings learned by UniRep contained physiochemically meaningful clusters. TAPE was a small Transformer trained on UniRef50 [35], which embedded each protein sequence to a vector of 768 and achieved comparable performance with UniRep on protein fluorescence and stability prediction. Of the three machine learning algorithms under investigation, the MLP algorithm again achieved the best performance in combination with these three pLMs judged by AUPRC or AUROC values. Interestingly, ESM-1b also outperformed three other pLMs in the computational framework of MLP (Fig. 3). We further compared ESM-1b with two baseline sequence encoding schemes (i.e., AAC and DPC). AAC stands for the compositions of each amino acid in the whole protein sequence, which transforms a protein into a vector of 20 dimensionality. DPC represents the compositions of two continuous amino acids in the whole protein sequence, which was used to convert a protein into a vector of 400 dimensionality. As shown in Additional file 1: Table S1, the combination of AAC and SVM seems to be optimal (AUPR = 0.519 on C2 and 0.481 on C3; AUROC = 0.852 on C2 and 0.824 on C3), while the combination of DPC and RF achieves the best performance (AUPR = 0.646 on C2 and 0.564 on C3; AUROC = 0.884 on C2 and 0.845 on C3). Comparatively, the optimal performance of these two traditional encoding schemes is much inferior to that of ESM-1b.

Comparison to existing generic PPI prediction methods

We compared our method with four generic PPI prediction methods, including three sequence-based methods [i.e., D-SCRIPT [16], RAPPPID [36], and PIPR [37]] and one structure-based method [i.e., TAGPPI [38]]. D-SCRIPT first applied a pre-trained model to generate structurally informative feature representations of proteins, and then estimates an interaction probability of protein pairs based on these features. RAPPPID is a deep learning-based PPI predictor implemented through a twin averaged weight-dropped LSTM network employing multiple regularization methods in the training step to learn generalized weights. When

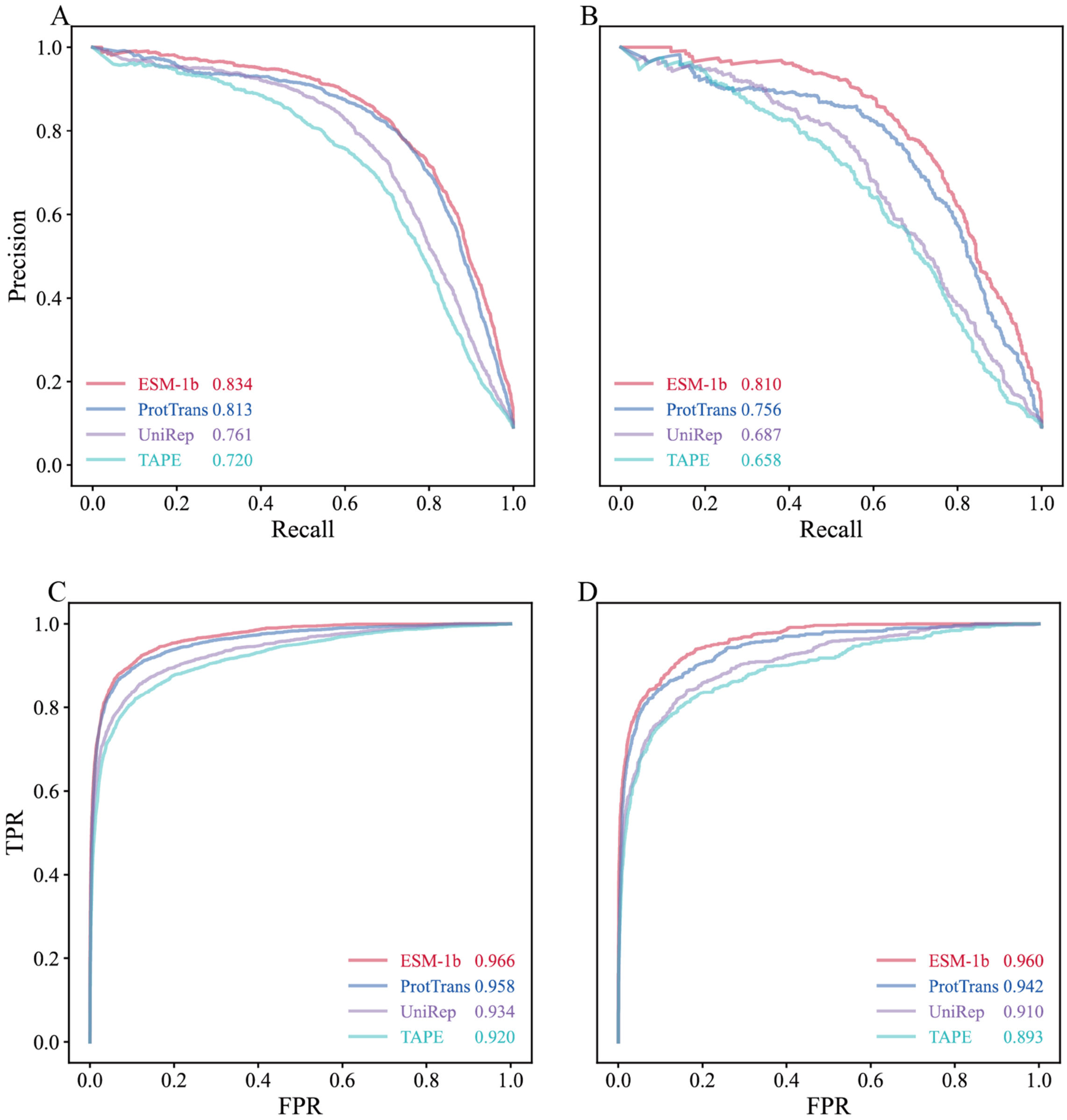

Fig. 3 PR and ROC curves of the predictive models from four pLMs in combination with MLP. **A** plots the PR curves on the independent test set C2, while **B** plots the PR curves on the independent test set C3. Parameters in the legends of **A** and **B** denote the corresponding AUPR values. **C** plots the ROC curves on the independent test set C2, while **D** plots the ROC curves on the independent test set C3. Parameters in the legends of **C** and **D** denote the corresponding AUROC values

tested on stringent PPI datasets containing proteins unseen in the training dataset, it reveals excellent performance. PIPR is a sequence-based PPI predictor combining pre-trained amino acid embeddings with a Siamese recurrent convolutional neural network (RCNN) architecture. TAGPPI is an end-to-end computational framework for PPI prediction, in which multi-dimensional features by employing 1D convolution operation on protein sequences and graph learning method on contact maps constructed from AlphaFold2 are considered.

We downloaded the source codes of PIPR, D-SCRIPT, RAPPPID, and TAGPPI and retrained the corresponding predictive models using the C1 dataset. Moreover, we tested their performance on our two independent datasets (C2 and C3). As shown in Fig. 4, the proposed ESMAraPPI considerably outperformed the four existing PPI predictors in terms of

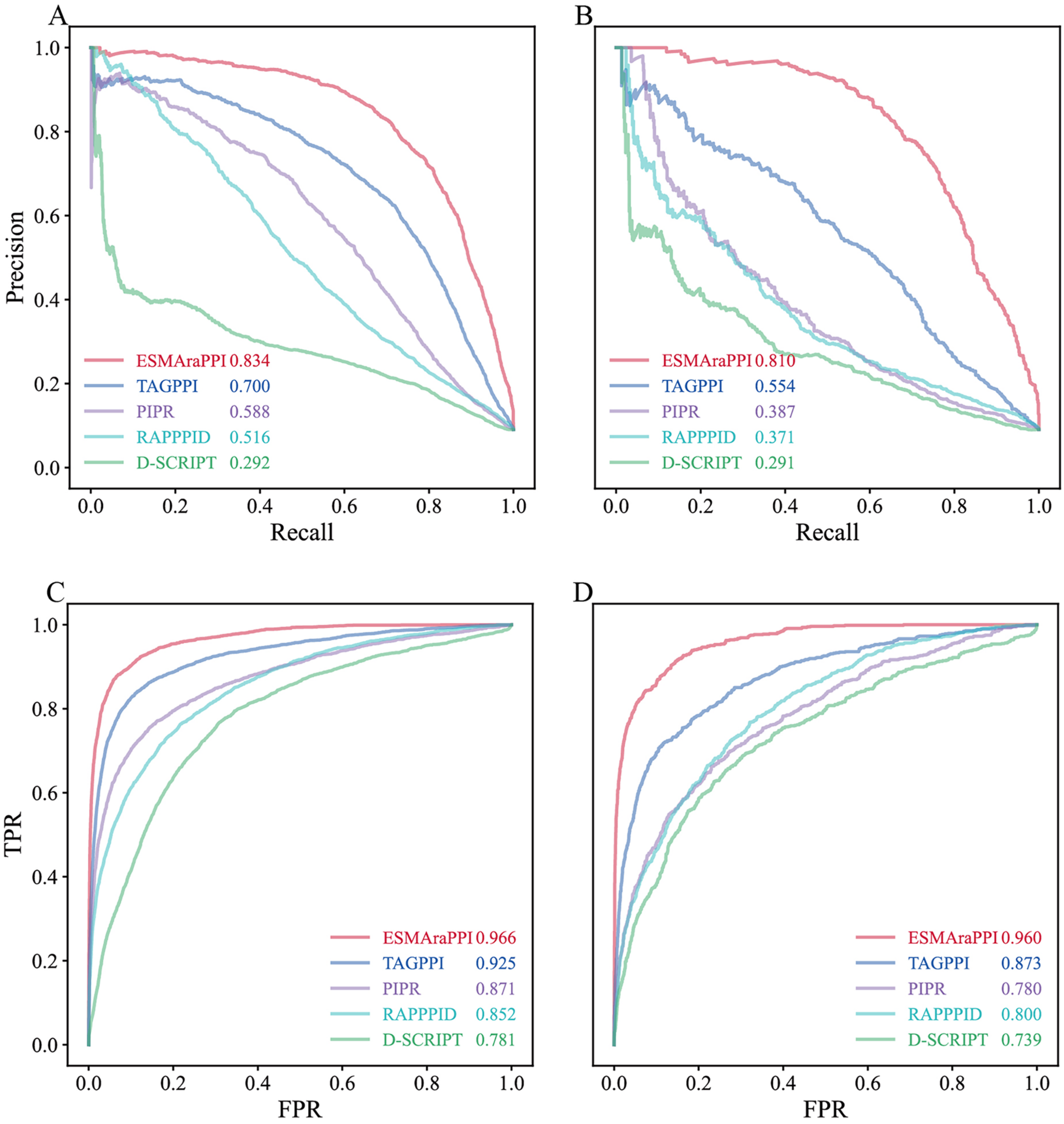

Fig. 4 PR and ROC curves of ESMAraPPI and four existing generic PPI predictors in predicting Arabidopsis PPIs. **A** plots the PR curves on the independent test set C2, while **B** plots the PR curves on the independent test set C3. Parameters in the legends of **A** and **B** denote the corresponding AUPR values. **C** plots the ROC curves on the independent test set C2, while **D** plots the ROC curves on the independent test set C3. Parameters in the legends of **C** and **D** denote the corresponding AUROC values

AUPRC or AUROC values. In particular, the proposed ESMAraPPI reveals robust performance on the C3 test set. For the three existing sequence-based methods, the performance of PIPR is ranked as the best, followed by RAPPPID and D-SCRIPT. When tested on the C3 test set, the performance ranking remains the same, but the performance of PIPR and RAPPPID dropped rapidly. Considering the predicted protein structural information used in PPI prediction, TAGPPI considerably surpassed these three pure sequence-based models on both C2 and C3, although it also dropped sharply on C3.

In addition to the better performance of ESM-1b embedding with simple MLP, its computational efficiency is also high. In either the model training or prediction steps, ESMAraPPI showed more rapid computational speed (Table 1).

Table 1 Computational time required in different methods[a]

	ESM-1b + MLP	TAGPPI	RAPPPID	PIPR	D-SCRIPT
Training epoch	40	10	20	20	10
Total training time	56 s	9.29 h	1.12 h	700 s	7.22 h
Total predicting time[b]	0.1 s	583 s	18 s	5 s	82 s

[a] All the training and prediction procedures were processed on a high-performance computer with 20 cores CPU, 256G RAM, and Tesla V100 GPU

[b] Total predicting time means the computational time required for processing the C3 test dataset

Comparison to existing Arabidopsis PPI prediction methods

We compared the proposed ESMAraPPI with two existing Arabidopsis PPI prediction methods [i.e., AraPPINet [39] and DeepAraPPI [40]]. AraPPINet was inferred from three-dimensional structures and functional evidence, encompassing 316,747 high-confidence interactions among 12,574 proteins. It exhibited high predictive power for discovering protein interactions at a 50% true positive rate. To allow for a fair comparison between our model and AraPPINet, we submitted our test datasets (C2 and C3) directly to their web server, and the default threshold (0.5) reported by AraPPINet was also used to distinguish interactions and non-interactions. Since AraPPINet did not release the predictions at different threshold values, we were not able to compare ESMAraPPI and AraPPINet through AUPRC or AUROC values. Thus, the routine measurements, such as Accuracy, Matthews correlation coefficient (MCC), Recall [i.e., true positive rate (TPR)], Specificity [i.e., 1-false positive rate (FPR)] and Precision, were employed for performance comparison (Table 2). Judged by the MCC value, which is a more comprehensive measurement than the other parameters, ESMAraPPI outperforms AraPPINet in both C2 and C3 datasets (Table 2).

We finally compared ESMAraPPI with DeepAraPPI, which was recently developed in our team. As an integrative Arabidopsis PPI prediction method, DeepAraPPI comprises three individual predictors, (i) a word2vec encoding-based Siamese RCNN model, (ii) a Domain2vec encoding-based MLP model, and (iii) a GO2vec encoding-based MLP model [40]. The final DeepAraPPI model combined the prediction results of the three individual predictors through a Logistic regression model. We also tested DeepAraPPI on our datasets. As shown in Table 3, ESMAraPPI outperformed the individual predictors of DeepAraPPI (i.e., RCNN, Domain2vec, and Go2vec) in both test sets (C2 and C3). With respect to C3, ESMAraPPI surpassed the final DeepAraPPI model, which means our new method was more competitive and will be more reliable in practical applications.

Case study

To explore the real application of ESMAraPPI, we provided a case study related to the interaction prediction of two proteins (BIN2 and SOS2) involved in the salt overly sensitive (SOS) pathway. In 2020, Li et al. showed that BIN2 functions as a negative regulator of primary root growth under salt stress by phosphorylating and inhibiting SOS2 [41]. It should be emphasized that the interaction between BIN2 and SOS2 was consistently determined by the yeast two-hybrid assay, the split-LUC assay and the BiFC assay in Li et al.'s work, which has not been included in any public database. Using the ESMAraPPI model, BIN2 and SOS2 were predicted to interact (prediction score = 0.592), indicating the proposed method has practical application in predicting Arabidopsis PPIs.

Table 3 AUPR and AUROC values of DeepAraPPI and ESMAraPPI on the C2 and C3 test sets[a]

Method	AUPR		AUROC	
	C2	C3	C2	C3
DeepAraPPI_RCNN	0.541	0.331	0.852	0.778
DeepAraPPI_Domain2vec	0.706	0.639	0.884	0.845
DeepAraPPI_Go2vec	0.771	0.709	0.942	0.917
DeepAraPPI	**0.871**	0.785	**0.978**	0.944
ESMAraPPI	0.824	**0.810**	0.966	**0.960**

[a] Figure in bold font indicates the corresponding model achieved the maximal AUPR and AUROC value

Table 2 Comparison of ESMAraPPI and AraPPINet on the C2 and C3 test sets

Methods	C2					C3				
	Accuracy	Specificity	MCC	Recall	Precision	Accuracy	Specificity	MCC	Recall	Precision
ESMAraPPI	0.957	0.994	0.708	0.589	0.901	0.954	0.994	0.688	0.557	0.902
AraPPINet	0.939	0.999	0.551	0.337	0.966	0.937	0.999	0.534	0.318	0.966

Conclusion

In this work, we found that sequence representations directly generated by large-scale pre-trained pLMs without any further feature engineering can be successfully used to develop machine learning-based Arabidopsis PPI predictors. We have shown that the proposed ESMAraPPI (i.e., ESM-1b + MLP) model yielded a highly accurate performance in predicting Arabidopsis PPIs. On the one hand, it achieved dramatic performance improvement in comparison to the models inferred from baseline sequence encoding schemes. On the other hand, it also revealed better performance than several state-of-the-art generic or plant-specific PPI predictors. The success of ESMAraPPI should be ascribed to the fact that the large-scale pre-trained pLMs can capture rich semantic information regarding protein sequence-structure-evolution relationships. To facilitate the research community, we have made all our codes and datasets freely available at https://github.com/keiwo/ESMAraPPI. We believe that the application of pLMs in protein sequence representation is providing a very promising way to deal with feature engineering in PPI prediction.

Methods

Data collection and preprocessing

Experimental Arabidopsis PPIs were first downloaded from IntAct (https://www.ebi.ac.uk/intact/home), and only PPIs with the type of direct interaction or physical association were further retained. Moreover, PPIs with MIscore < 0.45 were removed. Finally, we obtained 7729 PPIs, which are regarded as positive samples in this work. To construct negative samples, we first removed proteins in positive samples from the complete Arabidopsis protein list, and the remaining Arabidopsis proteins sharing ≥ 40% sequence identity with proteins in positive samples were further filtered out. Then, we removed redundant proteins by applying a sequence identity cutoff of 40%, and 8382 proteins were retained. After that, we obtained a protein list by mixing these 8382 proteins and proteins in positive samples, which were used to construct negative samples through random pairing. By controlling the ratio of positive and negative samples as 1:10, 77,290 random protein pairs that were not experimentally identified as PPIs were selected as negative samples. Finally, an original dataset containing 7729 positive samples (i.e., PPIs) and 77,290 negative samples (i.e., non-PPIs) was compiled in this work. To conduct model training and evaluation, we followed Park and Marcotte's advice to divide the original dataset into three datasets (i.e., C1, C2, and C3). C1 was the train dataset, while C2 and C3 were two independent test datasets. More details about the sizes of the three datasets are listed in Table 4.

Table 4 Statistics of the C1, C2 and C3 datasets

Dataset	#positive samples	#proteins involved in positive samples	#negative samples	#proteins involved in negative samples
C1	3519	1415	35,190	7068
C2	3404	1781	34,040	10,586
C3	806	551	8060	3534

Protein representation

The ESM models are available at https://github.com/facebookresearch/esm/tree/v1.0.2. There are 13 models in the ESM version we used. Of them, esm_msa1_t12_100M_UR50S and esm_msa1b_t12_100M_UR50S require extra MSAs as input, which were not further considered. In the remaining 11 models, nine models encode each protein sequence into a vector of 1280 dimensionality, which were chosen for further investigation. We downloaded these nine pre-trained ESM models and followed the ESM authors' instructions to run them locally. After extracting the final layer's hidden parameters, the matrix was averaged on the first dimension to generate 1280 features for each sequence. The ProtTrans model is available at https://github.com/agemagician/ProtTrans. We downloaded prot_t5_xl_uniref50 (ProtT5), which converted each protein sequence to a vector of 1024 dimensionality. The TAPE and UniRep models are available at https://github.com/songlab-cal/tape. Similarly, we downloaded TAPE and UniRep that embed each protein sequence to a vector of 1900 and 768 dimensionality, respectively.

Machine learning algorithms

Multilayer perceptron (MLP)

Through the PyTorch machine learning framework, we implemented a 4-layer MLP, which contains 1024, 512, 128, and 16 nodes. To avoid the order bias from protein pairs, the Hadamard product of two protein features, rather than their concatenation, was used as model input. The sigmoid function was applied to the final output to yield a prediction score between 0 and 1 (a prediction score ≥ 0.5 corresponded to a positive interaction). Then, the binary cross entropy (BCE) loss function was implemented.

Support vector machine (SVM)

We implemented SVM based on the sklearn package in Python, and the parameters were optimized by grid search. The kernel function was set as 'rbf', the regularization parameter was set to 1, and the kernel coefficient was set as 'scale'. The other parameters were set as default. The model input was Hadamard product of two protein features. A prediction score ≥ 0.5 was thought to be interaction.

Random forest (RF)

We implemented RF based on the sklearn package in Python. The parameters were optimized by grid search. The n_estimators was set as '100', and max_depth was set as 'None'. The model input was Hadamard product of two protein features. A prediction score ≥ 0.5 was thought to be interaction.

Performance evaluations

Accuracy, Specificity, Precision, Recall, and MCC were used to evaluate the prediction performance. These parameters are defined as follows:

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

$$Specificity = 1 - FPR = \frac{TN}{TN + FP}$$

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

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

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

where TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively. To provide a comprehensive performance assessment for each predictive model, the PR curve was plotted, and the AUPRC value was also calculated to quantify the performance. In the meantime, the ROC curve, which plots the TPR value against the FPR value at different thresholds, and the corresponding AUROC value were also employed for performance assessment.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13007-023-01119-6.

Additional file 1.

Acknowledgements
We are grateful for the stimulating discussion with Dr. Stefan Wuchty at Miami University regarding the prediction and network analysis of plant interactomes.

Author contributions
KZ developed the prediction model and drafted the manuscript. CL participated in the model construction. JZ and YH provided suggestions on data collection, model construction, and performance assessment. ZZ supervised the study and significantly revised the manuscript. All authors read and approved the final manuscript.

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

Availability of data and materials
The codes and datasets supporting the conclusions of this article are available in the GitHub repository, https://github.com/keiwo/ESMAraPPI.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 29 August 2023 Accepted: 28 November 2023
Published online: 07 December 2023

References

1. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014;42(D1):D358–63.
2. Velazquez-Campoy A, Freire E. ITC in the post-genomic era… ? Priceless. Biophys Chem. 2005;115(23):115–24.
3. Louche A, Salcedo SP, Bigot S. Protein-protein interactions: pull-down assays. Methods Mol Biol. 2017;1615:247–55.
4. Jung SO, Ro HS, Kho BH, Shin YB, Kim MG, Chung BH. Surface plasmon resonance imaging-based protein arrays for high-throughput screening of protein-protein interaction inhibitors. Proteomics. 2005;5(17):4427–31.
5. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A. 2001;98(8):4569–74.
6. Kocher T, Superti-Furga G. Mass spectrometry-based functional proteomics: from molecular machines to protein networks. Nat Methods. 2007;4(10):807–15.
7. Nakajima N, Akutsu T, Nakato R. Databases for Protein-Protein Interactions. Methods Mol Biol. 2021;2361:229–48.
8. Li TB, Wernersson R, Hansen RB, Horn H, Mercer J, Slodkowicz G, et al. A scored human protein-protein interaction network to catalyze genomic interpretation. Nat Methods. 2017;14(1):61–4.
9. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. Nucleic Acids Res. 2017;45(1):D408–14.
10. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021;49(D1):D605–12.
11. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Soding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 2017;45(D1):D170–6.
12. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast genome-wide functional annotation through orthology sssignment by eggNOG-mapper. Mol Biol Evol. 2017;34(8):2115–22.

13. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015;43(D1):D261–9.
14. Gainza P, Sverrisson F, Monti F, Rodola E, Boscaini D, Bronstein MM, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods. 2020;17(2):184–92.
15. Li YW, Golding GB, Ilie L. DELPHI: accurate deep ensemble model for protein interaction sites prediction. Bioinformatics. 2021;37(7):896–904.
16. 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):969–82.
17. Pugalenthi G, Kandaswamy KK, Suganthan PN, Archunan G, Sowdhamini R. Identification of functionally diverse lipocalin proteins from sequence information using support vector machine. Amino Acids. 2010;39(3):777–83.
18. Manavalan B, Shin TH, Kim MO, Lee G. AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest. Front Pharmacol. 2018;9(1):276.
19. Wang L, Wang HF, Liu SR, Yan X, Song KJ. Predicting protein-protein interactions from matrix-based protein sequence using convolution neural network and feature-selective rotation forest. Sci Rep. 2019;9(1):9848.
20. Le NQK, Huynh TT, Yapp EKY, Yeh HY. Identification of clathrin proteins by incorporating hyperparameter optimization in deep learning and PSSM profiles. Comput Meth Prog Bio. 2019;177:81–8.
21. Koca MB, Nourani E, Abbasoglu F, Karadeniz I, Sevilgen FE. Graph convolutional network based virus-human protein-protein interaction prediction for novel viruses. Comput Biol Chem. 2022;101(1):10775.
22. Yang XD, Yang SP, Li QMG, Wuchty S, Zhang ZD. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. Comput Struct Biotechnol J. 2020;18:153–61.
23. Cox S, Dong XL, Rai RH, Christopherson L, Zheng WF, Tropsha A, et al. A semantic similarity based methodology for predicting protein-protein interactions: Evaluation with P53-interacting kinases. J Biomed Inform. 2020;111: 103579.
24. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al., editors. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 December; Long Beach, California, USA.
25. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. P Ieee. 1998;86(11):2278–324.
26. Rives A, Meier J, Sercu T, Goyal S, Lin ZM, Liu JS, et al. 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.
27. Nie LP, Quan LJ, Wu TF, He RJ, Lyu Q. TransPPMP: predicting pathogenicity of frameshift and non-sense mutations by a Transformer based on protein features. Bioinformatics. 2022;38(10):2705–11.
28. Wang L, Zhong H, Xue Z, Wang Y. Res-Dom: predicting protein domain boundary from sequence using deep residual network and Bi-LSTM. Bioinform Adv. 2022;2(1):060.
29. Lin PC, Yan YM, Huang SY. DeepHomo2.0: improved protein-protein contact prediction of homodimers by transformer-enhanced deep learning. Brief Bioinform. 2022;24(1):499.
30. Shashkova TI, Umerenkov D, Salnikov M, Strashnov PV, Konstantinova AV, Lebed I, et al. SEMA: antigen B-cell conformational epitope prediction using deep transfer learning. Front Immunol. 2022;13:960985.
31. Kang Y, Elofsson A, Jiang Y, Huang W, Yu M, Li Z. AFTGAN: prediction of multi-type PPI based on attention free transformer and graph attention network. Bioinformatics. 2023;39(2):052.
32. Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. Nat Methods. 2012;9(12):1134–6.
33. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell. 2022;44(10):7112–27.
34. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods. 2019;16(12):1315–22.
35. Rao RS, Bhattacharya N, Thomas N, Duan Y, Chen X, Canny J, et al. Evaluating protein transfer learning with TAPE. Adv Neural Inf Process Syst. 2019;32(1):9689–701.
36. Szymborski J, Emad A. RAPPPID: towards generalizable protein interaction prediction with AWD-LSTM twin networks. Bioinformatics. 2022;38(16):3958–67.
37. Chen MH, Ju CJT, Zhou GY, Chen XL, Zhang TR, Chang KW, et al. Multifaceted protein-protein interaction prediction based on Siamese residual RCNN. Bioinformatics. 2019;35(14):I305–14.
38. Song BS, Luo XY, Luo XL, Liu YS, Niu ZM, Zeng XX. Learning spatial structures of proteins improves protein-protein interaction prediction. Brief Bioinform. 2022;23(2):558.
39. Zhang FY, Liu SW, Li L, Zuo KJ, Zhao LX, Zhang LD. Genome-wide inference of protein-protein interaction networks identifies crosstalk in abscisic acid signaling. Plant Physiol. 2016;171(2):1511–22.
40. Zheng JY, Yang XD, Huang Y, Yang SP, Wuchty S, Zhang ZD. Deep learning-assisted prediction of protein-protein interactions in Arabidopsis thaliana. Plant J. 2023;114(4):984–94.
41. Li JF, Zhou HP, Zhang Y, Li Z, Yang YQ, Guo Y. The GSK3-like Kinase BIN2 Is a Molecular Switch between the Salt Stress Response and Growth Recovery in Arabidopsis thaliana. Dev Cell. 2020;55(3):367–80.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.