Received: 10 July 2021 | Revised: 30 August 2021 | Accepted: 21 September 2021 | Accepted article online: 27 September 2021

DOI: 10.1002/pmic.202100161

Proteomics
Proteomics and Systems Biology

RESEARCH ARTICLE

prPred-DRLF: Plant R protein predictor using deep representation learning features

Yansu Wang[1,2] | **Lei Xu**[1] | **Quan Zou**[2,3] | **Chen Lin**[4]

[1] School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen, China

[2] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China

[3] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, Zhejiang, China

[4] School of Informatics, Xiamen University, Xiamen, China

Correspondence
Chen Lin, School of Informatics, Xiamen University, Xiamen 361005, China.
Email: chenlin@xmu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 91935302, 61922020, 61972328; Sichuan Provincial Science Fund for Distinguished Young Scholars, Grant/Award Number: 2021JDJQ0025; Special Science Foundation of Quzhou, Grant/Award Number: 2020D003; China Postdoctoral Science Foundation, Grant/Award Number: 2021M690029; Post-doctoral Foundation Project of Shenzhen Polytechnic, Grant/Award Number: 6021330009K; Joint Innovation Research Program of Fujian Province China, Grant/Award Number: 2020R0130

Abstract

Plant resistance (R) proteins play a significant role in the detection of pathogen invasion. Accurately predicting plant R proteins is a key task in phytopathology. Most plant R protein predictors are dependent on traditional feature extraction methods. Recently, deep representation learning methods have been successfully applied in solving protein classification problems. Motivated by this, we propose a new computational approach, called prPred-DRLF, which uses deep representation learning feature models to encode the amino acids as numerical vectors. The results show that the fused features of bidirectional long short-term memory (BiLSTM) embedding and unified representation (UniRep) embedding have a better performance than other features for plant R protein identification using a light gradient boosting machine (LGBM) classifier. The model was evaluated using an independent test achieving an accuracy of 0.956, F1-score of 0.933, and area under the receiver operating characteristic (ROC) curve (AUC) of 0.997. Meanwhile, compared with the state-of-the-art prPred and HMMER method, prPred-DRLF shows an overall improvement in accuracy, F1-score, AUC, and recall. prPred-DRLF is a higher-performance plant R protein prediction tool based on two kinds of deep representation learning technologies and offers a user-friendly interface for inspecting possible plant R proteins. We hope that prPred-DRLF will become a useful tool for biological research. A user-friendly webserver for prPred-DRLF is freely accessible at http://lab.malab.cn/soft/prPred-DRLF. The Python script can be downloaded from https://github.com/Wangys-prog/prPred-DRLF.

KEYWORDS
bidirectional long short-term memory, deep representation learning, light gradient boosting, plant R proteins, unified representation

1 | INTRODUCTION

During evolution, plants have developed a sophisticated defense immune system to recognize pathogens, and resistance (R) proteins play an important role in the plant defense process. Plant R proteins are divided into two categories. One is PRRs, cell surface pattern-recognition receptors, which contain various ligand-binding ectodomains. PRRs include two classes: receptor-like kinases (RLKs) and receptor-like proteins (RLPs) [1]. The other class is the NBS-LRR, which comprises a central nucleotide-binding site (NBS), a variable amino-terminal domain, and leucine-rich repeats (LRRs). The NBS is part of a nucleotide binding (NB)-ARC domain. It functions as a molecular switch and participates in ATP hydrolytic processes and pathogen recognition. The N-terminal domain often possesses either

Abbreviations: AUC, area under the ROC curve; BERT, bidirectional encoder representations from transformers; BiLSTM, bidirectional long short-term memory; ET, extra tree; LGBM, light gradient boosting machine; LRRs, leucine-rich repeats; LSTM, long-short-term memory; mLSTM, multiplicative long-/short-term-memory; MRMD, max-relevance-max-distance; NBS, nucleotide-binding site; R proteins, resistance proteins; RF, random forest; RLKs, receptor-like kinases; RLPs, receptor-like proteins; RNNs, recurrent neural networks; SVM, support vector machine; ROC, receiver operating characteristic; TAPE, tasks assessing protein embedding; UMAP, uniform manifold approximation and projection; UniRep, unified representation

a toll/interleukin-1 receptor-like (TIR) domain or a coiled coil (CC) domain. The C-terminal LRR domain is highly involved in pathogen recognition specificity and protein–protein interactions [2, 3].

Six predictors have been developed for plant R protein detection, including NLR-parser [4], RGAugury [5], Restrepo-Montoya's pipeline [6], NBSPred [7], DRPPP [8], and prPred [9]. Protein sequence conversion is the most important step for building predictors. To accurately infer the structural properties of a protein based on its amino acid sequence, many feature extraction methods have been proposed, including amino acid composition, pseudo acid composition, composition/transition/distribution, autocorrelation, and profile-based descriptors. Among the existing predictors for plant R protein prediction, NBSPred, DRPPP, and prPred are all based on traditional feature extraction methods to generate various numerical representation schemes to represent input sequences. Although these methods play an important role in understanding protein function, there remain many unknown protein properties.

Researchers have developed deep representation learning methods for protein feature engineering construction. Cui et al. [10] introduced the main approaches that are used to encode or embed amino acids into numerical vectors. They grouped the methods into several categories, including non-contextual embedding models [11, 12], long short-term memory (LSTM)-based models and transformer-based models [13, 14]. Recurrent neural networks (RNNs) are a type of deep learning technology that have been successfully applied in different areas; in recent years, RNNs have been widely applied in natural language processing (NLP), human action recognition, and biological sequence analysis [15–17]. As a particular subclass of RNNs, LSTM was proposed by Hochreiter and Schmidhuber [18] and it is a popular method for mapping amino acid sequences to vectors. It has been widely applied in bioinformatic prediction and has achieved state-of-the-art results in protein remote homology detection [19]. Bidirectional long short-term memory (BiLSTM) means that the input sequence should be passed through the pretrained language model in both the forward and reverse directions, and BiLSTM has also been used in protein disorder [20] and protein contact map prediction [21]. Multiplicative long-/short-term memory (mLSTM) RNNs are another successful representation model and have obtained state-of-the-art performance on many tasks [22]. UniRep is a deep representation learning method that adopts a 1900-hidden unit mLSTM to pretrain amino acid characteristics on UniRef50 and has been found to increase the efficiency in protein engineering problems [23–25]. Tasks assessing protein embedding (TAPE) are semi-supervised learning on protein sequences based on the BERT model that was proposed by Rao et al. [26]. Their results demonstrated that the representative models performed well for protein learning.

In this paper, we evaluated three deep representation learning methods to represent multiple characteristics of sequences for predicting plant R proteins. The experimental results demonstrated that the fused feature vector of BiLSTM and UniRep has excellent performance compared to other feature combinations. Three feature selection techniques, including random forest (RF), LGBM, and max-relevance-max-distance (MRMD), were applied to select an optimal feature subset from the fused BiLSTM + UniRep feature, and then the feature subset was used as input to four classifiers, including support vector machine (SVM), LGBM, RF, and extra tree (ET). After cross-validation and independent testing, LGBM shows better superiority and competitiveness relative to the other classifiers. The predictor is named prPred-DRLF, and its flowchart is illustrated in Figure 1.

Statement of significance

Plants R proteins play an important role in plant defense immune system, hence, developing an accurate computational tool for identifying plant R proteins is crucial for biological researches. However, most of the existing predictors for plant R protein prediction are all based on traditional feature extraction methods. It is necessary to develop higher-performance plant R prediction methods by extracting deep-seated sequence features. Here, we present prPred-DRLF, a computational tool based on two kinds of deep representation learning technologies that has already demonstrated success in classification of plant R proteins.

2 | METHODS

2.1 | The basic procedure of prPred-DRLF

The dataset employed in this study is based on Wang et al. [9] and can be downloaded from https://github.com/Wangys-prog/prPred-DRLF/tree/master/dataset. After removing the redundant sequences using CD-HIT [27] with a 30% identity cutoff, the dataset contained 456 protein sequences and included 152 positive samples and 304 negative samples. The number of training samples was 364, and the number of independent test samples was 92. We use 1 to represent the positive samples and 0 to identify the negative samples. The protein sequences are embedded into numerical vectors by three protein representation learning models, including BiLSTM, UniRep, and TAPE-BERT using eFeature (http://lab.malab.cn/soft/eFeature/index.html). The obtained multidimensional feature vectors include 3605D for BiLSTM, 1900D for UniRep, and 768D for TAPE-BERT. The performance of each representation learning model and their combinations are evaluated in four machine learning models. The embedding procedure of BiLSTM, TAPE-BERT, and UniRep is illustrated in Figure 2.

2.2 | The architecture of BiLSTM

LSTM is made up of a forget gate, input gate, and output gate, and its details are shown in Figure 2A. The forget gate determines which information should be retained. The sigmoid function is the activation function of the three gates, and it receives two inputs: the information of $x_{(t)}$, where the input is at the current time step, and the $h_{(t\text{-}1)}$, where it is generated from the previous step. The signal state of the LSTM

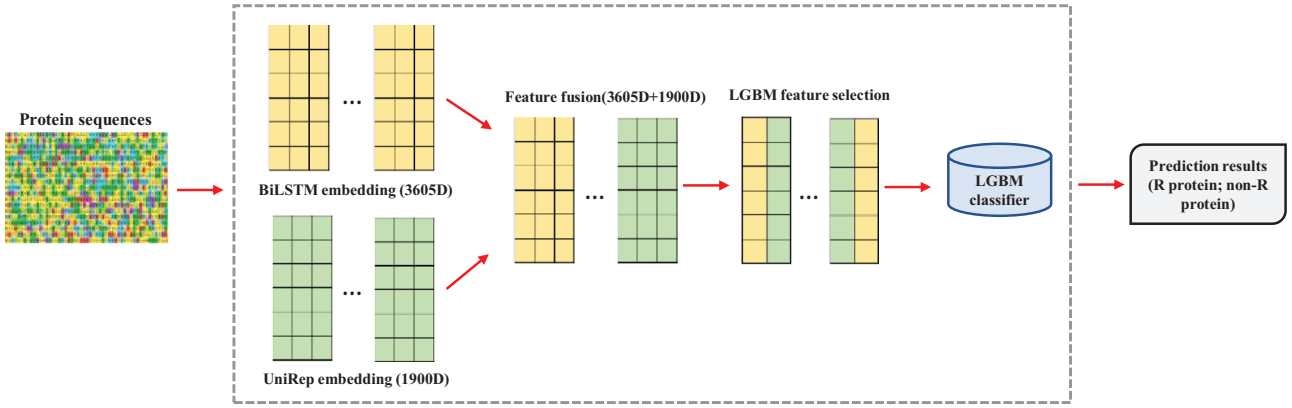

FIGURE 1 Overview of the prPred-DRLF procedure. The BiLSTM and UniRep embedding models encode protein sequences into 3605 dimension (D) and 1900 dimension (D) feature vectors, respectively. The fused BiLSTM + UniRep feature dimension is 5505. BiLSTM, bidirectional long short-term memory; UniRep, unified representation

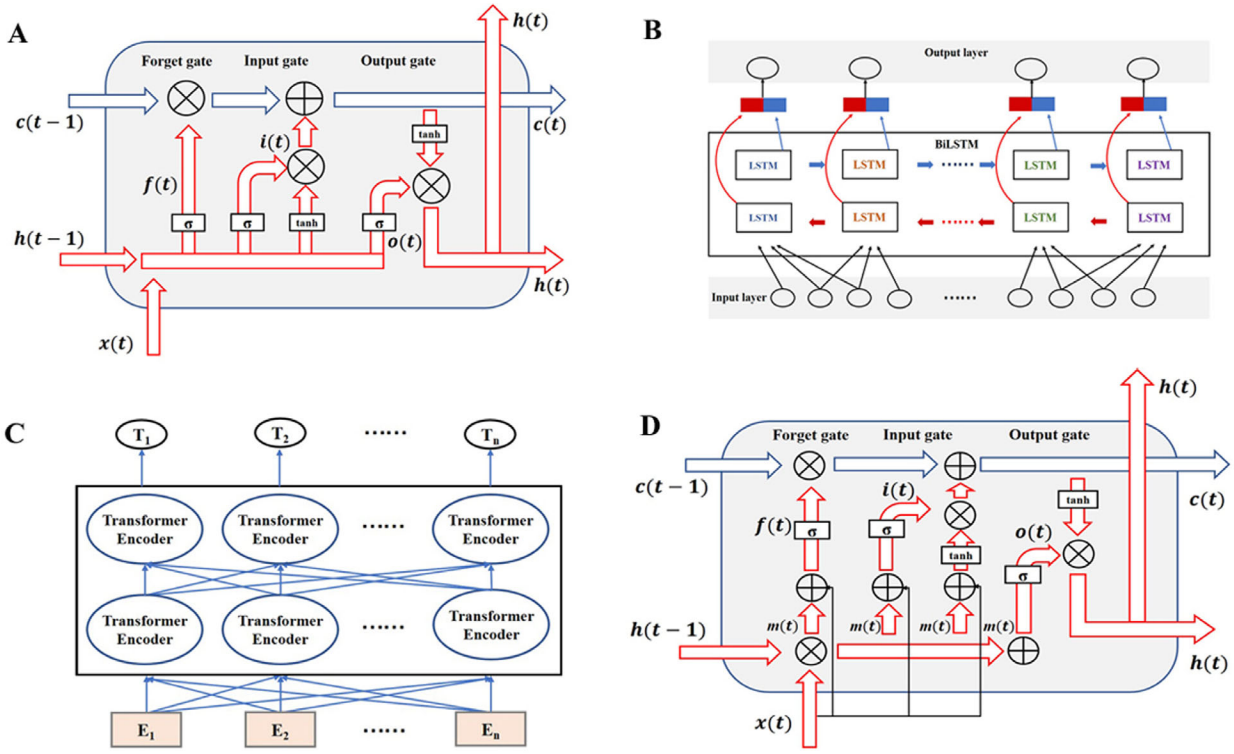

FIGURE 2 Illustration of the LSTM (A), BiLSTM model (B), BERT model (C), and mLSTM model (D). σ denotes the sigmoid function, tan$h$ is the hyperbolic tangent function, $\otimes$ represents elementwise multiplication, x indicates the matrix of the input subsequence and t means the LSTM cell architecture at the tth time step. h is the output hidden state. The internal cell state is maintained and updated by the coordination of the input gate and forget gate. BiLSTM extends a second layer based on the unidirectional LSTM, where bidirectional connections flow through the sequence before passing on to the next layer. BiLSTM, bidirectional long short-term memory; LSTM, long short-term memory; mLSTM, multiplicative long-/short-term memory

memory cell can be updated from $c_{(t-1)}$ to $c_{(t)}$ through the output of the input gate and forget gate [28, 29]. To acquire comprehensive bidirectional sequence information, BiLSTM proposed by Bepler and Berger [30] is adopted in this study. BiLSTM is a two-layer bidirectional model that comprises two reversed unidirectional LSTMs [19, 31], and its structure is shown in Figure 2B.

2.3 | The architecture of TAPE-BERT

Recently, Rao et al. [26] introduced TAPE to systematically evaluate semi-supervised protein learning. The TAPE implements five biologically relevant supervised tasks to assess the relative merits of five sequence representation models, which include an LSTM [32],

a transformer, a dilated residual network (ResNet), BiLSTM, and UniRep. They found that transformer-based adoption performed well for sequence modeling. The BERT network model is the bidirectional encoder representation from transformers and it contains a multilayer transformer encoder structure [14] and has been successfully applied to identify DNA enhancers [33] and bitter peptides [34]. Details of the BERT architecture are shown in Figure 2C.

2.4 | The architecture of UniRep

UniRep is another type of deep sequence representation learning method that was proposed by Alley et al. [23]. It uses a unidirectional mLSTM [35] with 1900 hidden units to represent protein sequences, and the learned representations based on the UniRep model are semantically rich. mLSTM is a hybrid architecture that combines LSTM and a multiplicative recurrent neural network (mRNN) and it adds the $m_{(t)}$ that is the mRNN's intermediate state to the gating units of LSTM (Figure 2D). The LSTM in mLSTM is responsible for controlling information flow using multiplicative gates, and mRNN is designed to allow flexible input-dependent transitions.

2.5 | Feature selection

The feature vector generated from deep representation learning methods is the high-dimensional feature space, and it is necessary to extract the optimal feature subset using feature selection approaches. In this study, we compared three feature selection methods: RF, LGBM, and MRMD3.0 [36, 37] (http://lab.malab.cn/soft/MRMD3.0/index.html). The first algorithm is an ensemble learning method consisting of multiple decision trees, the second algorithm performs gradient boosting on decision trees, and the third algorithm consists of various feature selection methods, such as MIC and mRMR, and ranks features based on PageRank, LeaderRank, Hits, or TrustRank algorithm. Fivefold cross-validation on the dataset is used to examine the effectiveness of each model.

2.6 | Evaluation metrics

To evaluate the performance of the models, four evaluation measures are used, including precision, recall, accuracy (ACC), and F1-score. Their equations are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$F1 - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where TP, FP, TN, and FN represent the numbers of true positives (predicted R proteins are R proteins), false positives (predicted R proteins are non-R proteins), true negatives (predicted non-R proteins are non-R proteins), and false negatives (predicted non-R proteins are R proteins).

Additionally, we used the area under the ROC curve (AUC) to diagnose the performance of the different models. The ROC curve refers to the receiver operating characteristic (ROC) curve that plots the true positive rates versus the false positive rates for all thresholds.

3 | RESULTS

3.1 | Initial performances of the different classifiers

The results presented in Table S1 illustrate how well the different classifiers were at distinguishing plant R and non-R proteins with each combination of features without feature selection. As Table S1 shows, the fivefold cross-validation accuracy, F1-score and AUC in the training set ranged from 0.926 to 0.959, 0.881 to 0.933, and 0.942 to 0.984, respectively. To better understand the performance of these classifiers, we set three thresholds according to the value of accuracy, F1-score and AUC: 0.950 for accuracy, 0.920 for F1-score, and 0.980 for AUC. The results show that 16 groups' accuracy values were greater than or equal to 0.950, 14 groups' F1-score values were greater than or equal to 0.920, and three groups' AUC values were greater than or equal to 0.980. It is clear that the accuracy values of the classifiers that used the fused features of BiLSTM and UniRep were all greater than 0.950 (Table S1). Moreover, among the four classifiers, the LGBM and ET classifiers offered their advantages over different feature combinations for plant R protein prediction based on the accuracy and F1-score values, among which the LGBM classifier performed better than the other three classifiers based on the AUC value.

3.2 | Comparison with feature selection technologies based on RF, LGBM, and MRMD3.0

To determine the best feature vector subset, features were ranked by feature importance values generated by the RF, LGBM, and MRMD3.0 methods. Then, we used five measures to evaluate the performance of the two feature subsets selected using RF, LGBM, and MRMD3.0 (Tables S2–S4). We found that the LGBM method could improve the performance of the model more effectively than RF and MRMD3.0 and it achieved the highest accuracy (Table 1). For the LGBM classifier, the average accuracy was improved 0.63% for BiLSTM, 0.32% for TAPE-BERT, 0.63% for UniRep, 1.16% for BiLSTM + TAPE-BERT, 1.46% for BiLSTM + UniRep, 1.16% for TAPE-BERT + UniRep, and 1.79% for TAPE-BERT + BiLSTM + UniRep (Table 1 and Figure 3). The F1-score was improved 0.86% for BiLSTM, 0.22% for TAPE-BERT, 0.86% for UniRep, 1.84% for BiLSTM + TAPE-BERT, 2.47% for BiLSTM + UniRep, 1.85% for TAPE-BERT + UniRep, and 2.82% for TAPE-BERT + BiLSTM + UniRep after LGBM feature selection (Tables S1–S2). The value

TABLE 1 Fivefold cross-validation comparison among different feature combinations and different feature selection classifiers on the training dataset

Classifier	Feature	ACC	ACC (LGBM)	ACC (RF)	ACC (MRMD3.0)
SVM	BiLSTM	0.945 ± 0.02^{ab}	0.939 ± 0.02^{a}	0.942 ± 0.02^{ab}	0.948 ± 0.03^{ab}
SVM	TAPE-BERT	0.948 ± 0.03^{ab}	0.951 ± 0.02^{ab}	0.95 ± 0.02^{ab}	0.948 ± 0.02^{ab}
SVM	UniRep	0.948 ± 0.03^{ab}	0.953 ± 0.03^{ab}	0.953 ± 0.02^{ab}	0.945 ± 0.03^{ab}
SVM	BiLSTM + TAPE-BERT	0.948 ± 0.03^{ab}	0.95 ± 0.03^{ab}	0.95 ± 0.03^{ab}	0.948 ± 0.03^{ab}
SVM	BiLSTM + UniRep	0.951 ± 0.03^{ab}	0.948 ± 0.03^{ab}	0.953 ± 0.02^{ab}	0.956 ± 0.04^{ab}
SVM	TAPE-BERT + UniRep	0.951 ± 0.02^{ab}	0.961 ± 0.02^{ab}	0.945 ± 0.02^{ab}	0.956 ± 0.02^{ab}
SVM	TAPE-BERT + BiLSTM + UniRep	0.951 ± 0.02^{ab}	0.956 ± 0.03^{ab}	0.964 ± 0.02^{b}	0.953 ± 0.03^{ab}
LGBM	BiLSTM	0.953 ± 0.02^{ab}	0.959 ± 0.02^{ab}	0.948 ± 0.03^{ab}	0.956 ± 0.03^{ab}
LGBM	TAPE-BERT	0.945 ± 0.05^{ab}	0.948 ± 0.03^{ab}	0.939 ± 0.03^{a}	0.937 ± 0.02^{a}
LGBM	UniRep	0.956 ± 0.02^{ab}	0.962 ± 0.02^{ab}	0.948 ± 0.02^{ab}	0.951 ± 0.02^{ab}
LGBM	BiLSTM + TAPE-BERT	0.950 ± 0.02^{ab}	0.961 ± 0.03^{ab}	0.948 ± 0.03^{ab}	0.950 ± 0.03^{ab}
LGBM	BiLSTM + UniRep	0.956 ± 0.02^{ab}	**0.97** ± 0.01^{b}	0.948 ± 0.02^{ab}	0.961 ± 0.02^{b}
LGBM	TAPE-BERT + UniRep	0.951 ± 0.01^{ab}	0.962 ± 0.02^{ab}	0.95 ± 0.02^{ab}	0.959 ± 0.02^{ab}
LGBM	TAPE-BERT + BiLSTM + UniRep	0.950 ± 0.03^{ab}	0.967 ± 0.02^{ab}	0.953 ± 0.02^{ab}	0.964 ± 0.02^{b}
RF	BiLSTM	0.942 ± 0.02^{ab}	0.948 ± 0.02^{ab}	0.942 ± 0.03^{ab}	0.956 ± 0.02^{ab}
RF	TAPE-BERT	0.926 ± 0.03^{a}	0.942 ± 0.03^{ab}	0.939 ± 0.04^{ab}	0.942 ± 0.03^{ab}
RF	UniRep	0.948 ± 0.02^{ab}	0.945 ± 0.02^{ab}	0.948 ± 0.02^{ab}	0.951 ± 0.02^{ab}
RF	BiLSTM + TAPE-BERT	0.948 ± 0.03^{ab}	0.956 ± 0.02^{ab}	0.948 ± 0.02^{ab}	0.956 ± 0.03^{ab}
RF	BiLSTM + UniRep	0.951 ± 0.02^{ab}	0.951 ± 0.02^{ab}	0.953 ± 0.02^{ab}	0.956 ± 0.02^{ab}
RF	TAPE-BERT + UniRep	0.948 ± 0.02^{ab}	0.951 ± 0.02^{ab}	0.953 ± 0.02^{ab}	0.959 ± 0.02^{ab}
RF	TAPE-BERT + BiLSTM + UniRep	0.948 ± 0.02^{ab}	0.956 ± 0.02^{ab}	0.956 ± 0.03^{ab}	0.956 ± 0.02^{b}
ET	BiLSTM	0.953 ± 0.02^{ab}	0.959 ± 0.02^{ab}	0.951 ± 0.01^{ab}	0.959 ± 0.02^{ab}
ET	TAPE-BERT	0.945 ± 0.02^{ab}	0.956 ± 0.03^{ab}	0.948 ± 0.02^{ab}	0.950 ± 0.02^{ab}
ET	UniRep	0.956 ± 0.02^{ab}	0.959 ± 0.02^{ab}	0.959 ± 0.02^{ab}	0.959 ± 0.02^{ab}
ET	BiLSTM + TAPE-BERT	0.956 ± 0.02^{ab}	0.956 ± 0.02^{ab}	0.956 ± 0.02^{ab}	0.956 ± 0.02^{ab}
ET	BiLSTM + UniRep	0.956 ± 0.03^{ab}	0.962 ± 0.02^{ab}	0.962 ± 0.02^{ab}	0.962 ± 0.02^{b}
ET	TAPE-BERT + UniRep	0.956 ± 0.03^{ab}	0.959 ± 0.02^{ab}	0.959 ± 0.02^{ab}	0.962 ± 0.02^{b}
ET	TAPE-BERT + BiLSTM + UniRep	0.959 ± 0.02^{ab}	0.959 ± 0.02^{ab}	0.959 ± 0.02^{ab}	0.964 ± 0.02^{b}

The accuracy values in the third column indicate the initial performances of different classifiers. The accuracy values in the fourth column indicate the performances of different classifiers after LGBM feature selection. The accuracy values in the fifth column indicate the performances of different classifiers after RF feature selection. The accuracy values in the sixth column indicate the performances of different classifiers after MRMD3.0 feature selection. The bold fonts represent the highest value of accuracy. Superscript a and b represent the significance level at 0.05. The same letters in the same column represent in apparent differences. The different letters represent significant differences. ACC, accuracy; BiLSTM, bidirectional long short-term memory; ET, extra tree; LGBM, light gradient boosting machine; MRMD, max-relevance-max-distance; RF, random forest; SVM, support vector machine; TAPE, tasks assessing protein embedding; UniRep, unified representation.

of the average accuracy had less or no improvements after RF and MRMD3.0 feature selection (Figure 3). Moreover, these phenomena were also observed for other classifiers (Table 1). Although no significant increase was achieved in accuracy after feature selection, LGBM feature selection technology performed better than the other methods and reduced the computing resource consumption (Figure 3). Hence, the LGBM feature selection method was utilized to build the prediction model.

The model using the LGBM classifier and LGBM feature selection based on the BiLSTM + UniRep feature vector achieved the highest accuracy and F1-score values, which were 0.970 and 0.953, respectively (Tables 1 and S2). The model estimated by fivefold cross-validation had an ROC = 0.988. For further comparison of the performance with and without the LGBM feature selection method, we use the uniform manifold approximation and projection (UMAP) method [38] to visualize the distribution profiles of positive and negative samples based on the BiLSTM + UniRep feature type. We found that the two-dimensional UMAP space separated the non-R and R proteins into distinct clusters, and feature vector dimensionality was reduced by the LGBM method (Figure 4). According to the cross-validation results, the model using the BiLSTM + UniRep feature based on the LGBM feature selection method and the LGBM classifier shows excellent performance compared to the other models. Then, we further evaluated the performance of the model in the independent test dataset.

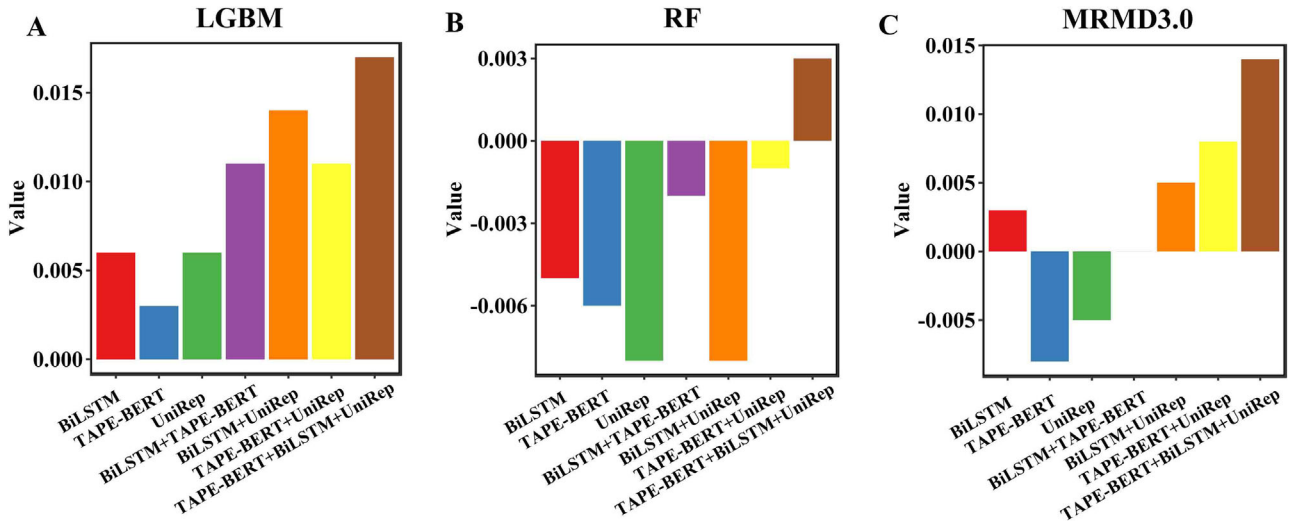

FIGURE 3 Comparison of LGBM (A), RF (B), and MRMD3.0 (C) feature selection technology based on the LGBM classifier. The y-axis represents the improvement or decrease in accuracy after feature selection. The LGBM feature selection technology performed better than the other methods. LGBM, light gradient boosting machine; MRMD, max-relevance-max-distance; RF, random forest

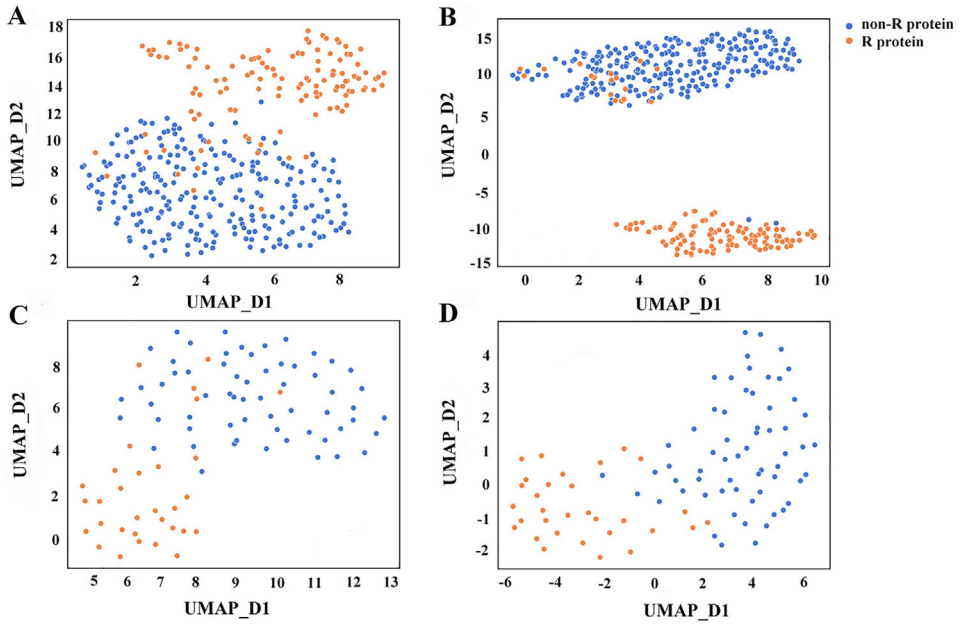

FIGURE 4 Feature visualization by UMAP for dimension reduction. (A) BiLSTM + UniRep without feature selection on the training dataset; (B) BiLSTM + UniRep after LGBM feature selection on the training dataset; (C) BiLSTM + UniRep without feature selection on the independent test dataset; and (D) BiLSTM + UniRep after LGBM feature selection on the independent test dataset. BiLSTM, bidirectional long short-term memory; LGBM, light gradient boosting machine; UMAP, uniform manifold approximation and projection; UniRep, unified representation

3.3 | Comparison with the predictors based on different deep representation feature types in the independent test dataset

We established different predictors based on different feature combinations for the identification of plant R proteins to explore whether the best predictor selected based on the cross-validation results still has the best effect on plant R protein classification in independent tests. Table 2 lists the average accuracy, precision, recall, F1-score, and AUC scores reported by the predictors in the independent test after feature selection using LGBM. From the results, we found that the scores achieved by the fusion feature of BiLSTM + UniRep were higher than those of the other features. For example, BiLSTM + UniRep significantly outperformed TAPE-BERT in terms of all scores, improving by 8.51%, 10.39%, 16.62%, 14.20%, and 5.28% in accuracy, precision, recall, F1-score, and AUC, respectively, and it exceeded the individual

TABLE 2 Performance of models with different feature combinations based on the LGBM classifier on the independent test dataset

Models	Accuracy	Precision	Recall	F1-score	AUC
BiLSTM	0.923	0.933	0.838	0.882	0.965
TAPE-BERT	0.881	0.876	0.776	0.817	0.947
UniRep	0.923	0.967	0.805	0.875	0.989
BiLSTM + TAPE-BERT	0.923	0.933	0.838	0.882	0.992
BiLSTM + UniRep	**0.956**	**0.967**	**0.905**	**0.933**	**0.997**
TAPE-BERT + UniRep	0.945	0.967	0.871	0.915	0.992
TAPE-BERT + BiLSTM + UniRep	0.923	0.943	0.838	0.884	0.989

AUC, area under the receiver operating characteristic (ROC) curve; BiLSTM, bidirectional long short-term memory; LGBM, light gradient boosting machine; TAPE, tasks assessing protein embedding; UniRep, unified representation. The bold font represents the best performance of models.

BiLSTM and UniRep feature type, accuracy, precision, recall, F1-score, and AUC, which were increased by 3.58% (3.58%), 3.64% (0.00%), 8.00% (12.42%), 5.78% (6.63%), and 3.32% (0.81%), respectively. Thus, from the above results, the predictor for plant R proteins was chosen, where both the classification and feature selection algorithms were LGBM, and the deep representation feature type was BiLSTM + UniRep.

3.4 | Comparison with the prPred classifier and HMMER method

We then used an independent test dataset to evaluate the ability of prPred-DRLF and prPred that we previously established. The results in Figure 5A show that prPred-DRLF performed better than prPred that extracted feature representations using traditional feature representation methods. prPred-DRLF increased the ACC by 2.20%, F1-score by 4.29%, and AUC by 4.91%. Similar to a previous study by Lv et al. [24], the deep representation learning methods demonstrated excellent performance in protein classification and detection.

Next, we benchmarked the performance of prPred-DRLF against the alignment method HMMER. The alignment method identifies R proteins based on domains and motifs, such as LRR, CC, Toll/interleukin-1 receptor (TIR), NBS (NB-ARC), transmembrane (TM), serine/threonine and tyrosine kinase (STTK), and lysin motif (LysM) [5]. Because most protein sequences from the negative training set are annotated incompletely using the HMMER method, we only evaluated whether the sequences from the positive training set belong to potential R proteins based on the domain structures and obtained the recall values. From the comparative results, it can be concluded that the method we proposed can perform better than the HMMER method (Figure 5B).

3.5 | Web server implementation

To help users analyze their protein sequences in a user-friendly manner, we designed a web server for prPred-DRLF. With our web server, we provide an easy interface that allows the users to input

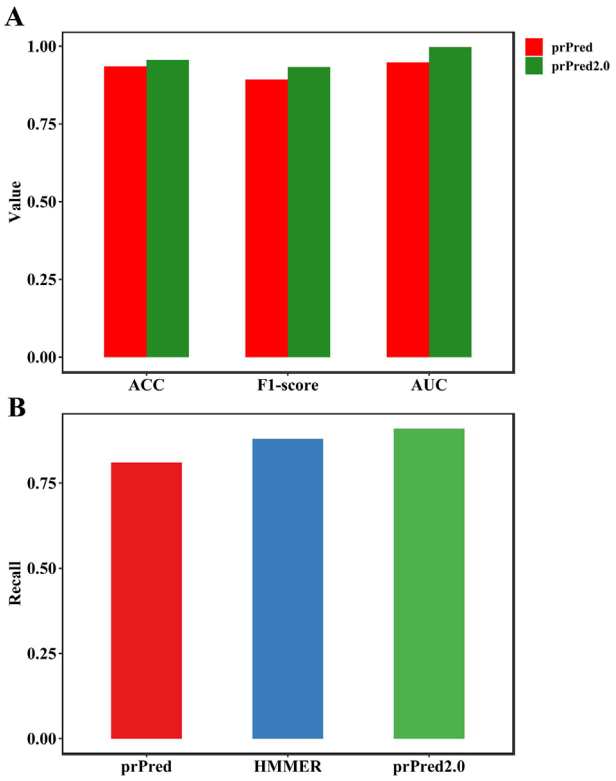

FIGURE 5 Performance of prPred-DRLF, prPred, and HMMER

their sequences in FASTA format to run the program. The web server can load models automatically and detect the probability of whether a protein is a plant R protein. The homepage of the server is shown in Figure 6.

4 | CONCLUSION

In this study, we proposed a new predictor, prPred-DRLF, for plant R protein detection based on BiLSTM and the UniRep feature learning method. The experimental results showed that prPred-DRLF yielded better prediction quality than the existing method prPred, which extracted protein features by traditional machine learning methods.

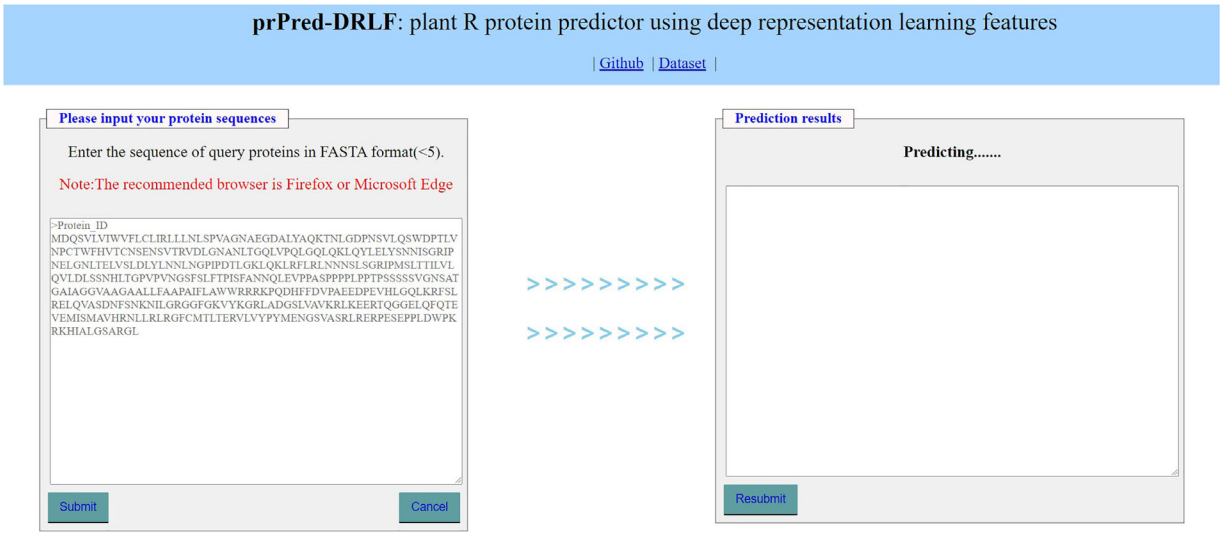

FIGURE 6 Homepage of the prPred-DRLF web server

Breeding disease-resistant varieties has proven to be the most effective and economical means to control plant disease and increase crop yield and quality [38]. Diverse strategies for breeding durable resistance varieties have been adopted, such as pyramiding [40], mixtures [41], and multilines [42]. Identification of resistance genes or proteins is a critical step for molecular resistance breeding, and this strategy has been successfully applied for breeding new wheat varieties containing resistance genes to resist wheat stem rust [43]. We hope that prPred-DRLF will become a useful tool to help plant breeders develop disease-resistant varieties.

Although deep representation learning feature methods capture sequence information more effectively and achieve better performance than traditional feature extraction methods, they usually require an elevated level of computing resources. In future studies, we will attempt to incorporate parallel computing into our program to achieve high computational efficiency.

ACKNOWLEDGEMENTS

The authors thank the editor and anonymous reviewers for their useful comments.

The work was supported by the National Natural Science Foundation of China (No.91935302, 61922020), the Sichuan Provincial Science Fund for Distinguished Young Scholars (2021JDJQ0025), and the Special Science Foundation of Quzhou (2020D003). China Postdoctoral Science Foundation (No. 2021M690029) and Post-doctoral Foundation Project of Shenzhen Polytechnic (No. 6021330009K).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Chen Linand Quan Zou designed this research. Yansu Wang performed the experiment and drafted the manuscript. Lei Xu helped complete the data processing and discussion. All authors revised and approved the final manuscript.

DATA AVAILABILITY STATEMENT

All codes and data files are available on GitHub https://github.com/Wangys-prog/prPred-DRLF.

ORCID

Yansu Wang https://orcid.org/0000-0003-1274-4958
Quan Zou https://orcid.org/0000-0001-6406-1142
Chen Lin https://orcid.org/0000-0002-2275-997X

REFERENCES

1. Altenbach, D., & Robatzek, S. (2007). Pattern recognition receptors: From the cell surface to intracellular dynamics. *Molecular Plant-Microbe Interactions*, *20*(9), 1031–1039.
2. Takken, F. L., Albrecht, M., & Tameling, W. I. (2006). Resistance proteins: Molecular switches of plant defence. *Current Opinion in Plant Biology*, *9*(4), 383–390.
3. Dubey, N., & Singh, K. (2018). Role of NBS-LRR proteins in plant defense. In A. Singh& I. K. Singh (Eds.), *Molecular aspects of plant-pathogen interaction* (pp. 115–138). Springer Singapore.
4. Steuernagel, B., Jupe, F., Witek, K., Jones, J. D., & Wulff, B. B. (2015). NLR-parser: Rapid annotation of plant NLR complements. *Bioinformatics*, *31*(10), 1665–1667.
5. Li, P., Quan, X., Jia, G., Xiao, J., Cloutier, S., & You, F. M. (2016). RGAugury: A pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics (Electronic Resource)*, *17*(1), 1–10.
6. Restrepo-Montoya, D., Brueggeman, R., McClean, P. E., & Osorno, J. M. (2020). Computational identification of receptor-like kinases "RLK" and receptor-like proteins "RLP" in legumes. *BMC Genomics (Electronic Resource)*, *21*(1), 1–17. https://doi.org/10.1186/s12864-020-06844-z
7. Kushwaha, S. K., Chauhan, P., Hedlund, K., & Ahrén, D. (2016). NBSPred: A support vector machine-based high-throughput pipeline for plant resistance protein NBSLRR prediction. *Bioinformatics*, *32*(8), 1223–1225.
8. Pal, T., Jaiswal, V., & Chauhan, R. S. (2016). DRPPP: A machine learning based tool for prediction of disease resistance proteins in plants. *Computers in Biology and Medicine*, *78*, 42–48.
9. Wang, Y., Wang, P., Guo, Y., Huang, S., Chen, Y., Xu, L., prPred: A Predictor to Identify Plant Resistance Proteins by Incorporating k-Spaced Amino Acid (Group) Pairs *Frontiers in Bioengineering and Biotechnology* 2021, *8*, https://doi.org/10.3389/fbioe.2020.645520

10. Cui, F., Zhang, Z., & Zou, Q. (2021). Sequence representation approaches for sequence-based protein prediction tasks that use deep learning. *Briefings in Functional Genomics*, *20*(1), 61–73.
11. Le, N. Q. K., & Huynh, T.-T. (2019). Identifying SNAREs by incorporating deep learning architecture and amino acid embedding representation. *Frontiers in Physiology*,*10*, 1501.
12. Do, D. T., & Le, N. Q. K. (2020). Using extreme gradient boosting to identify origin of replication in Saccharomyces cerevisiae via hybrid features. *Genomics*, *112*(3), 2445–2451.
13. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018). Deep contextualized word representations. Paper presented at the Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana.
14. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
15. Liu, A.-A., Su, Y.-T., Jia, P.-P., Gao, Z., Hao, T., & Yang, Z.-X. (2014). Multiple/single-view human action recognition via part-induced multitask structural learning. *IEEE Transactions on Cybernetics*, *45*(6), 1194–1208.
16. Morchid, M. (2018). Parsimonious memory unit for recurrent neural networks with application to natural language processing. *Neurocomputing*,*314*, 48–64.
17. Hawkins, J., & Bodén, M. (2005). The applicability of recurrent neural networks for biological sequence analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *2*(3), 243–253.
18. Hochreiter, S., & Schmidhuber, J. (1997). LSTM can solve hard long time lag problems. In *Advances in neural information processing systems*, MA: MIT Press, Cambridge, (pp. 473–479).
19. Liu, B., & Li, S. (2018). ProtDet-CCH: Protein remote homology detection by combining long short-term memory and ranking methods. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*,*16*(4), 1203–1210.
20. Hanson, J., Yang, Y., Paliwal, K., & Zhou, Y. (2017). Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*, *33*(5), 685–692.
21. Hanson, J., Paliwal, K., Litfin, T., Yang, Y., & Zhou, Y. (2018). Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics*, *34*(23), 4039–4045.
22. Radford, A., Jozefowicz, R., & Sutskever, I. (2017). Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*.
23. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, *16*(12), 1315–1322.
24. Lv, Z., Cui, F., Zou, Q., Zhang, L., & Xu, L. (2021). Anticancer peptides prediction with deep representation learning features. *Briefings in Bioinformatics*, *22*(5), bbab008. https://doi.org/10.1093/bib/bbab1008
25. Lv, Z., Wang, P., Zou, Q., & Jiang, Q. (2020). Identification of sub-Golgi protein localization by use of deep representation learning features. *Bioinformatics*, *36*(24), 5600–5609. https://doi.org/10.1093/bioinformatics/btaa1074
26. Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., & Song, Y. S. (2019). Evaluating protein transfer learning with tape. *Advances in Neural Information Processing Systems*,*32*, 9689.
27. Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152.
28. Gers, F. A., Schmidhuber, J. & Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, *12*(10), 2451–2471. https://doi.org/10.1162/089976600300015015
29. Graves, A. (2013). Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850.
30. Bepler, T., & Berger, B. (2019). Learning protein sequence embeddings using information from structure. *arXiv preprint arXiv:1902.08661*.
31. Li, S., Chen, J., & Liu, B. (2017). Protein remote homology detection based on bidirectional long short-term memory. *BMC Bioinformatics*, *18*(1), 1–8.
32. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*,*9*(8), 1735–1780.
33. Le, N. Q. K., Ho, Q.-T., Nguyen, T.-T.-D., & Ou, Y.-Y. (2021). A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information. *Briefings in Bioinformatics*,*22*(5). bbab005. https://doi.org/10.1093/bib/bbab005
34. Charoenkwan, P., Nantasenamat, C., Hasan, M. M., Manavalan, B., & Shoombuatong, W. (2021). BERT4Bitter: A bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. *Bioinformatics*, *37*(17), 2556–2562. https://doi.org/10.1093/bioinformatics/btab133
35. Krause, B., Lu, L., Murray, I., & Renals, S. (2016). Multiplicative LSTM for sequence modelling. *arXiv preprint arXiv:1609.07959*.
36. He, S., Guo, F., & Zou, Q. (2020). MRMD2. 0: A python tool for machine learning with feature ranking and reduction. *Current Bioinformatics*, *15*(10), 1213–1221.
37. Zou, Q., Zeng, J., Cao, L., & Ji, R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing*, *173*, 346–354.
38. Dorrity, M. W., Saunders, L. M., Queitsch, C., Fields, S., & Trapnell, C. (2020). Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nature Communications*, *11*(1), 1–6.
39. Ashkani, S., Rafii, M. Y., Shabanimofrad, M., Miah, G., Sahebi, M., Azizi, P., Tanweer, F. A., Akhtar, M. S., & Nasehi, A.(2015). Molecular breeding strategy and challenges towards improvement of blast disease resistance in rice crop. *Frontiers in Plant Science*,*6*, 886.
40. Ramalingam, J., Raveendra, C., Savitha, P., Vidya, V., Chaithra, T. L., Velprabakaran, S., Saraswathi, R., Ramanathan, A., Pillai, M. P. A., Arumugachamy, S., & Vanniarajan, C.(2020). Gene pyramiding for achieving enhanced resistance to bacterial blight, blast, and sheath blight diseases in rice. *Frontiers in Plant Science*, *11*,1662.
41. Zhu, Y., Chen, H., Fan, J., Wang, Y., Li, Y., Chen, J., Fan, J., Yang, S., Hu, L., Leung, H., Mew, T. W., Teng, P. S., Wang, Z., & Mundt, C. C.(2000). Genetic diversity and disease control in rice. *Nature*, *406*(6797), 718–722.
42. Li, W., Deng, Y., Ning, Y., He, Z., & Wang, G.-L. (2020). Exploiting broad-spectrum disease resistance in crops: From molecular dissection to breeding. *Annual Review of Plant Biology*, *71*, 575–603.
43. Bansal, U., Bariana, H., Wong, D., Randhawa, M., Wicker, T., Hayden, M., & Keller, B., (2014). Molecular mapping of an adult plant stem rust resistance gene Sr56 in winter wheat cultivar Arina. *Theoretical and Applied Genetics*,*127*(6), 1441–1448.

SUPPORTING INFORMATION

Additional supporting information may be found online https://doi.org/10.1002/pmic.202100161 in the Supporting Information section at the end of the article.

How to cite this article: Wang, Y., Xu, L., Zou, Q., &Lin, C. (2022). prPred-DRLF: Plant R protein predictor using deep representation learning features. *Proteomics*, *22*, e2100161. https://doi.org/10.1002/pmic.202100161