Proteomics

WILEY-VCH

Proteomics
Proteomics and Systems Biology

RESEARCH ARTICLE

Prediction of Plant Resistance Proteins Using Alignment-Based and Alignment-Free Approaches

Pushpendra Singh Gahlot | Shubham Choudhury | Nisha Bajiya | Nishant Kumar | Gajendra P. S. Raghava

Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India

Correspondence: Gajendra P. S. Raghava (raghava@iiitd.ac.in)

Received: 30 July 2024 | **Revised:** 21 October 2024 | **Accepted:** 11 November 2024

Funding: This study was funded by Department of Biotechnology (DBT) grant BT/PR40158/BTIS/137/24/2021.

Keywords: alignment-based approach | ensemble method | machine learning | PDR

ABSTRACT

Plant disease resistance (PDR) proteins are critical in identifying plant pathogens. Predicting PDR protein is essential for understanding plant–pathogen interactions and developing strategies for crop protection. This study proposes a hybrid model for predicting and designing PDR proteins against plant-invading pathogens. Initially, we tried alignment-based approaches, such as Basic Local Alignment Search Tool (BLAST) for similarity search and MERCI for motif search. These alignment-based approaches exhibit very poor coverage or sensitivity. To overcome these limitations, we developed alignment-free or machine learning (ML)-based methods using compositional features of proteins. Our ML-based model, developed using compositional features of proteins, achieved a maximum performance area under the receiver operating characteristic curve (AUROC) of 0.91. The performance of our model improved significantly from AUROC of 0.91–0.95 when we used evolutionary information instead of protein sequence. Finally, we developed a hybrid or ensemble model that combined our best ML model with BLAST and obtained the highest AUROC of 0.98 on the validation dataset. We trained and tested our models on a training dataset and evaluated them on a validation dataset. None of the proteins in our validation dataset are more than 40% similar to proteins in the training dataset. One of the objectives of this study is to facilitate the scientific community working in plant biology. Thus, we developed an online platform for predicting and designing plant resistance proteins, "PlantDRPpred" (https://webs.iiitd.edu.in/raghava/plantdrppred).

1 | Introduction

A wide range of pathogens, including fungi, bacteria, nematodes, viruses, protozoa, and insects, can destroy or affect the growth of plants. Over the years, plants have evolved complex and dynamic immune systems essential for survival, growth, and development. Broadly, the plant immune system can be categorized into two main types: cell surface or pattern-triggered immunity (PTI) and intracellular or effector-trigger immunity (ETI). Pattern recognition receptors (PRRs) are specialized proteins that play a pivotal role in PTI by recognizing and responding to pathogen-associated molecular patterns (PAMPs) and damage-associated molecular patterns (DAMPs) [1].Disease resistance proteins, products of resistance (R) genes, enable plants to recognize specific pathogen effectors in ETI. These effectors are molecules produced by pathogens to promote their growth within host tissues. The interaction between R genes and pathogen avirulence (Avr) gene products determines

Abbreviations: AAC, amino acid composition; AUROC, area under the receiver operating characteristic curve; BC, bagging classifier; BLAST, Basic Local Alignment Search Tool; DPC, dipeptide composition; ET, extra trees; MCC, Matthew correlation coefficient; PDR, plant disease resistance; PSSM, position-specific scoring matrix; RF, random forest; SVC, support vector classifier; XGB, eXtreme gradient boosting.

Summary

- Development of a machine-learning model for resistance protein prediction.
- Used alignment-based and alignment-free ensemble methods.
- Web server development and standalone package.
- Prediction and design of plant disease resistance (PDR) proteins.

whether a plant is resistant or susceptible to a pathogen attack [2].

R genes in plants exhibit diverse structural features, reflecting the complex nature of plant–pathogen interactions. At the molecular level, R genes typically encode plant disease resistance (PDR) proteins with conserved domains such as nucleotide-binding sites (NBSs) and leucine-rich repeats (LRRs). The NBS domain, which can be either N-terminal or centrally located within the protein, is involved in ATP or GTP binding and hydrolysis, essential for signal transduction in plant defense responses. The LRR domain, often located at the C-terminus, is responsible for protein–protein interactions and pathogen recognition, as it forms a versatile scaffold for binding to pathogen-derived molecules [3]. Additionally, many PDR proteins contain other domains, such as coiled-coil (CC) motifs or Toll/interleukin-1 receptor (TIR), which mediate downstream signaling events upon pathogen recognition [4, 5]. The modular structure of PDR proteins allows for diverse recognition specificities and signaling pathways, contributing to the robustness of plant immunity.

Current prediction methods such as NBSPred and NLR-parser utilize sequence similarity or domain-based approaches for predicting PDR proteins [6, 7]. These methods may fail if new proteins are not similar to known annotated proteins. Several machine learning (ML)-based methods have been developed to overcome these challenges, like DRPPP, prPred, and stackRPred [8, 9, 10]. Most of these methods have been developed on old and outdated data. Thus, there is a need to develop highly accurate and reliable models for predicting PDR proteins. In this study, we have systematically attempted to create the dataset of PDR proteins from the database PRGdb and non-PDR proteins from Swiss-Prot. To create a nonredundant dataset, we created clusters using CD-Hit at 40%; these clusters were partitioned into training and validation datasets. In contrast to existing methods, we developed a hybrid or ensemble method that combines alignment-based and alignment-free approaches (Figure 1).

2 | Materials and Methods

2.1 | Dataset Formation

In this study, we acquired 199 PDR proteins from the PRGdb 4.0 database called positive proteins [11]. Similarly, we obtained non-PDR proteins from Swiss-Prot, called negative proteins [12]. Since the PDR protein sequences of Bryophyta and Angiosperm are available in positive sequences, we used sequences from the parent clade Embryophyta to generate the negative dataset, as it includes PDR proteins from both Bryophyta and Angiosperm. We specifically selected sequences that had been previously reviewed. Subsequently, we filtered this dataset to include sequences with lengths more than 63 and less than 1826. We randomly sampled 199 protein sequences from this filtered pool to constitute our negative dataset and ensure they are not involved in defense-related functions based on available information. To generate nonredundant subsets while preserving the number of sequences, we adopted an approach used in previous studies [13–16]. We used CD-HIT [17] software at a cutoff of 40% to create clusters for PDR proteins and non-PDR, where no two proteins have a sequence similarity of more than 40%; 118 clusters for PDR proteins and 186 clusters for non-PDR were obtained. The steps involved in creating a nonredundant dataset are illustrated in Figure 2.

2.2 | Feature Generation

We utilized the tool Pfeature [18] to extract composition-based features and the POSSUM tool [19] to extract the evolutionary feature (position-specific scoring matrix [PSSM]). Composition-based features include amino acid composition (AAC) and dipeptide composition (DPC). AAC is represented as a 20-length vector, where each element signifies the fraction of a particular residue type within the sequence. In contrast, DPC is a 400-length vector that captures the occurrence of amino acid pairs within the protein sequence. The evolutionary features of proteins are recognized to provide vital insights beyond the primary sequence features [20]. These features are typically derived through the calculation of PSSM profiles using the Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) [21]. The PSSM represents a matrix with dimensions of 20 × sequence length for protein or peptide sequences. However, to develop ML models, which require fixed-length vectors, we employed the composition of the PSSM profile as PSSM-400, a fixed-length vector containing 400 elements as evolutionary features [15].

2.3 | Cross-Validation

The performance of the modules developed in this study was assessed using a five-fold cross-validation technique. This approach divided the dataset into positive and negative subsets to form training and test sets. Specifically, four positive subsets and their corresponding four negative subsets were combined to create the training set. The remaining positive subset and negative subset were used to form the test set. In this instance, the ML models were assessed using k-fold cross-validation (five-fold cv) on the training data. This involves randomly dividing the full dataset into k subsets, with one subset used for testing the classifier and the remaining k–1 subsets used for training. This method is iterated k times until every subset has been utilized to test the classifier precisely once. The classifier's overall performance is measured by calculating the average of the classification accuracies obtained from each step of the cross-validation process. Using different train/test splits can result in significant differences in accuracy. Therefore, this method provides a more generalizable estimate of classifier performance compared to using only a single split. Additionally, this approach helps reduce overfitting and underfitting of the model by ensuring

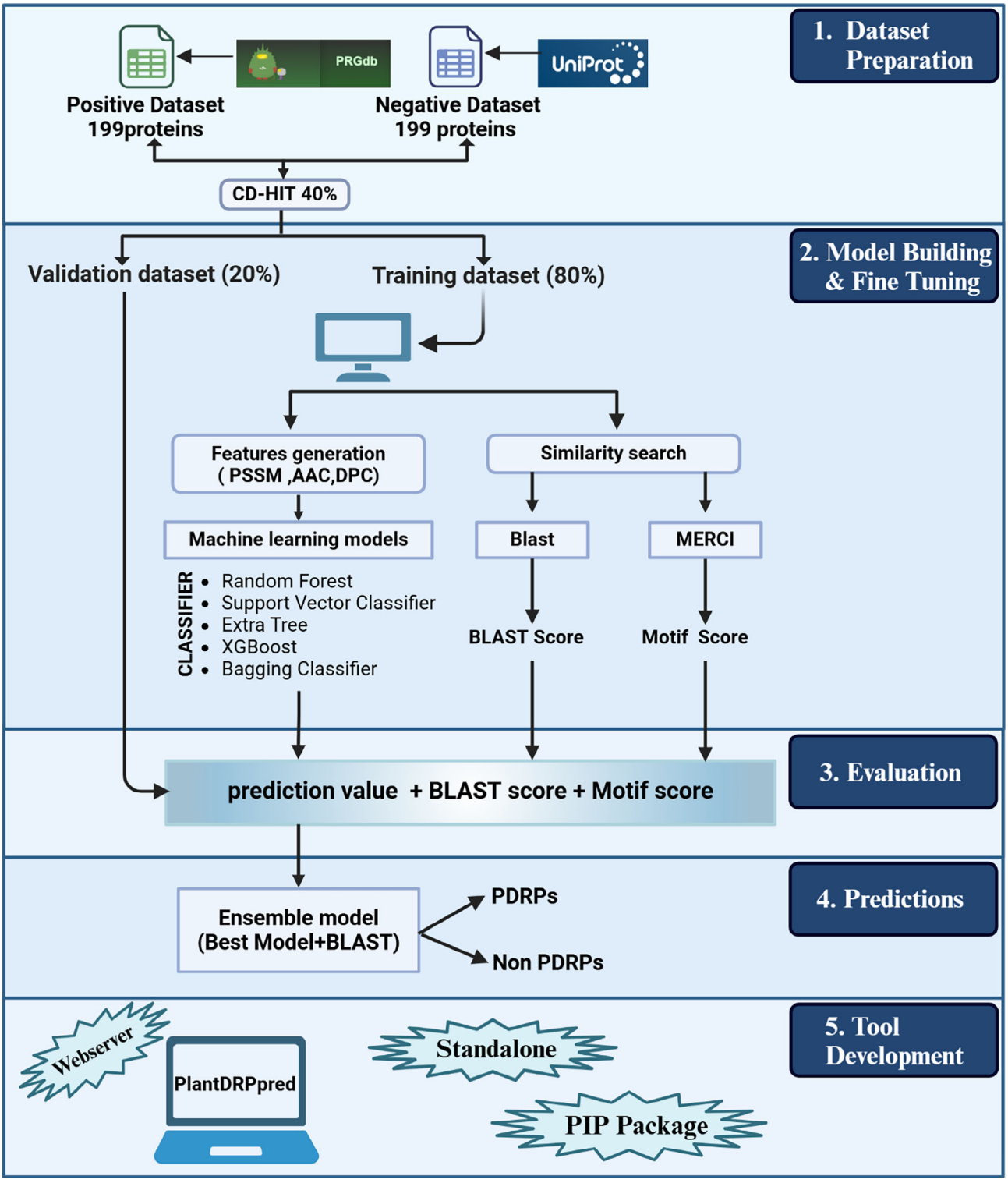


FIGURE 1 | Workflow shows study architecture from data collection to web server development.

that the classifier is trained and validated on different subsets of data, promoting a more balanced learning process.

2.4 | Alignment Based Method

We employed BLAST [22] for similarity searches. First, we constructed a database using the training dataset. Then, we performed similarity searches against this database using sequences from the validation dataset. This approach allows us to compare sequences in the validation dataset against those in the training dataset, aiding in tasks such as classification or identification of similar sequences. Identifying functional motifs within protein sequences is crucial for functional annotation and distinguishing between positive and negative datasets. This study employed the Motif Emerging with Classes Identification (MERCI) program

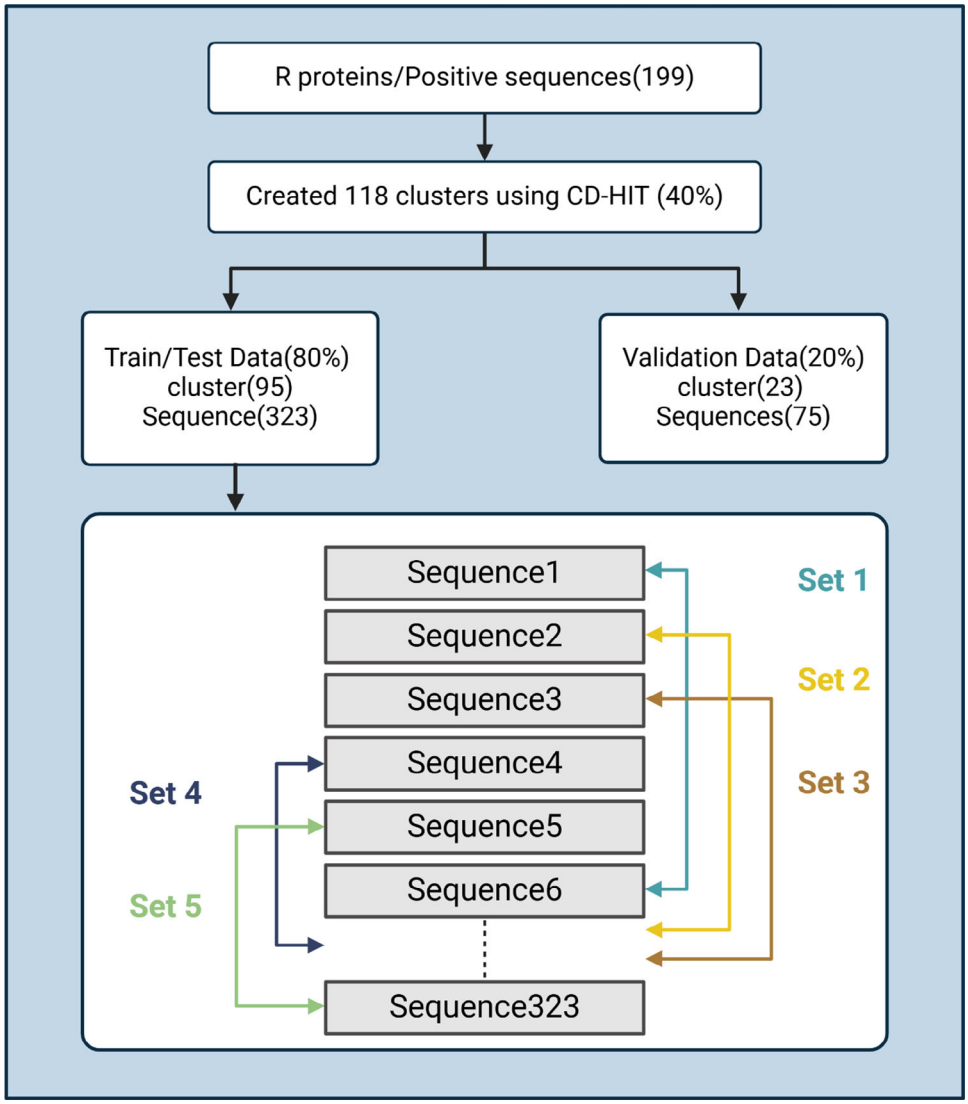


FIGURE 2 | The flowchart illustrates the generation of nonredundant training, validation, and testing of datasets of plant disease resistance (PDR) proteins. Initially, training and validation datasets are created by separating all clusters of PDR proteins. Subsequently, the training dataset's sequences are divided into five subsets.

to identify motifs within PDR and non-PDR protein sequences [23]. We explored the extraction of motifs using various k values. Afterward, we use parameter $k = 20$ to extract motifs (Table S2) that are exclusively present in PDR proteins and not non-PDR proteins.

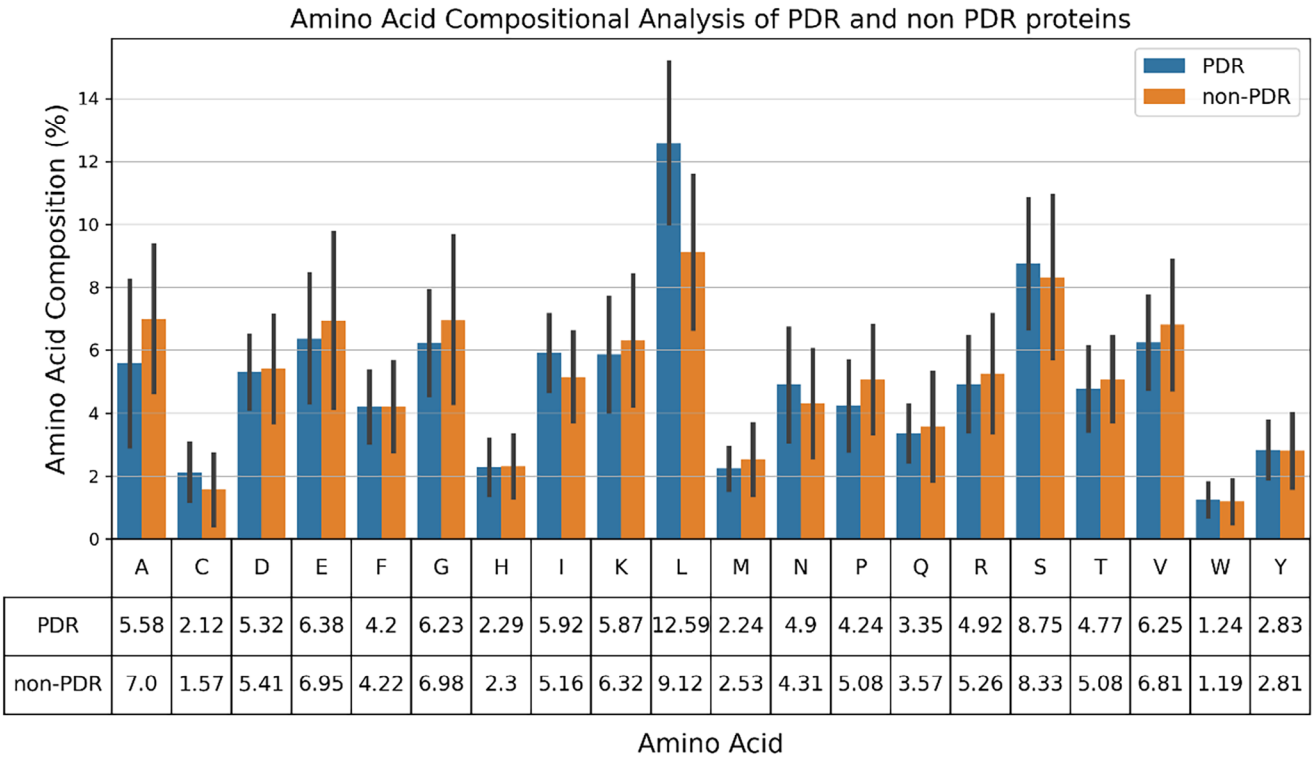
2.5 | Alignment-Free Methods

2.5.1 | Machine Learning Techniques

We implemented ML classifiers using the Python sci-kit learn package to build our prediction models. To enhance model performance, we fine-tuned hyperparameters on the training dataset with the help of sklearn's GridSearch package. After determining the optimal hyperparameters, the best model was evaluated on the validation set. We used a five-fold cross-validation to execute the entire process and averaged the results across all five folds. Using optimized feature vectors, we developed prediction models with various classifiers, including support vector classifier (SVC), extra trees (ET), random forest (RF), Gradient Boosting (GB), and bagging classifier (BC).

2.6 | Ensemble or Hybrid Approach for Classification

This study used a hybrid or an ensemble approach to augment the model's predictive capabilities. This ensemble methodology adopts a weighted scoring mechanism, incorporating three distinct methods: (i) ML approach, (ii) similarity-based technique employing BLAST, and (iii) motif-based technique using MERCI. Scoring system in the BLAST approach, a weight of "+0.5" was assigned for positive predictions, "−0.5" for negative predictions, and "0" for no hits using BLAST. A scoring system was devised for MERCI classification to assign values based on various conditions



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
PDR	5.58	2.12	5.32	6.38	4.2	6.23	2.29	5.92	5.87	12.59	2.24	4.9	4.24	3.35	4.92	8.75	4.77	6.25	1.24	2.83
non-PDR	7.0	1.57	5.41	6.95	4.22	6.98	2.3	5.16	6.32	9.12	2.53	4.31	5.08	3.57	5.26	8.33	5.08	6.81	1.19	2.81

FIGURE 3 | The percent amino acid compositional analysis for plant disease resistance (PDR) or non-PDR proteins.

where, when a motif is found in a sequence, a score of "+0.1" is assigned. Additionally, for each additional motif found in the same sequence, an extra "0.1" is added. If a motif is found in a negative sequence, a score of "−0.1" is added. If no motif is found in a sequence, a score of 0 is assigned.

2.7 | Performance Metrics Calculation

The ML models employed in this investigation were assessed using a variety of performance metrics, encompassing parameters both dependent and independent of the threshold. The evaluation metrics include specificity, sensitivity, Matthew correlation coefficient (MCC), the area under the receiver operating characteristic curve (AUROC), and accuracy. Notably, AUROC is threshold-independent, while the remaining parameters, such as specificity, sensitivity, and MCC, are threshold-dependent and were optimized to identify the threshold yielding maximum values. An ML model's accuracy is a measure that shows how often it delivers accurate predictions. The computation involves dividing the total number of guesses by the number of correct forecasts. Sensitivity measures the level of accuracy in identifying the true positives. Specificity quantifies the ratio of accurately detected actual negatives from the total number of negatives. Matthews correlation coefficient (MCC) is a statistical measure that provides a balanced assessment by considering all four components of a confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). AUROC measures the area under the ROC curve, providing a single value that summarizes the model's performance across all possible classification thresholds. These metrics have been widely employed in previous studies to gauge the performance of ML models [16, 24, 26–28]. The following Equations (1–4) were used to calculate these:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{2}$$

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

$$\text{Mcc} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{4}$$

where TP is the true positive, FP is the false positive, TN is the true negative, and FN is the false negative.

3 | Results

3.1 | Composition Analysis

We calculated AAC for PDR and non-PDR proteins, as shown in Figure 3. We observed that amino acids leucine (L), serine (S), and isoleucine (I) are more abundant in PDR proteins. Similarly, amino acids alanine (A), proline (P), and glycine (G) are widespread in non-PDR proteins. We calculated AAC using the formula shown in Equation 5.

$$\text{AAC}_i = \frac{R_i}{L} \tag{5}$$

where is AAC_i is the amino acid composition of residue type i, R_i is a number of amino acids of type i, and L is the length of the sequence.

TABLE 1 | The hits obtained by employing BLAST on the validation datasets of PDR and non-PDR proteins at different *e* values.

	Correct hits (PDRs)		Incorrect hits (non-PDRs)	
BLAST (*e* value)	**PDRs**	**Non-PDRs**	**Non-PDRs**	**PDRs**
1.00E-06	31	0	9	0
1.00E-05	31	0	9	0
0.0001	31	0	9	0
0.001	31	0	10	0
0.01	31	0	11	1
0.1	31	0	12	3
1	34	0	17	8
10	36	0	21	18
100	36	0	21	18

Abbreviations: BLAST, Basic Local Alignment Search Tool; PDR, plant disease resistance; PDRP, plant disease resistance protein.

TABLE 2 | The performance of different machine learning techniques-based models using the AAC feature of protein sequences.

	Training dataset					Validation dataset				
Models	**Sen**	**Spec**	**Acc**	**AUROC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUROC**	**MCC**
RF	0.81	0.86	0.83	0.91	0.68	0.86	0.77	0.81	0.88	0.63
SVC	0.81	0.87	0.84	0.92	0.68	0.75	0.87	0.81	0.91	0.63
ET	0.86	0.85	0.85	0.92	0.71	0.83	0.77	0.80	0.90	0.60
XGB	0.81	0.81	0.81	0.90	0.63	0.86	0.72	0.79	0.88	0.58
BC	0.79	0.85	0.82	0.91	0.64	0.86	0.79	0.83	0.88	0.66

Abbreviations: AAC, amino acid composition; Acc, accuracy; AUROC, area under the receiver operating characteristic curve; BC, bagging classifier; ET, extra trees; MCC, Matthew correlation coefficient; RF, random forest; Sen, sensitivity; Spec, specificity; SVC, support vector classifier; XGB, eXtreme gradient boosting.

3.2 | BLAST Search

Initially, we utilized BLAST (version 2.15.0+) for similarity searches by constructing a database of training datasets. Subsequently, we employed this dataset to search for similarities within the validation sequences using various *E* values. The performance of the BLAST module deteriorates when greater *e* values are used, as BLAST permits random matches at higher *e* values (see Table 1). Consequently, as the *e* values increase, the number of hits also increases, but this negatively impacts performance because it allows for false hits.

3.3 | Compositional-Based Model

We utilized the AAC feature to construct ML models employing RF, ET SVC, eXtreme gradient boosting (XGB), and BC. The results are summarized in Table 2. The highest AUROC on the training set was achieved with SVC, reaching 0.92; similarly, SVC attained the highest AUROC of 0.91 on the validation dataset.

Similarly, models were developed utilizing DPC along with various ML techniques. Details of these models can be found in Table S1. The best performance is the RF model with AUROC, which is 0.90 on the training set, and the validation dataset is 0.90.

3.4 | PSSM Feature-Based Model

Previous studies have demonstrated the enhanced information provided by sequence profiles compared to individual sequences alone [16, 29, 30]. Therefore, in this study, we initially generated PSSM-400 composition profiles corresponding to each protein utilizing POSSUM software and used them as feature vectors to develop classification models. Similar to the AAC and DPC-based methods, we employed various classifiers, such as ET, RF, SVC, XGB, BC, etc. As depicted in Table 3, models based on evolutionary information exhibited a maximum AUROC of 0.96 on the training dataset. The PSSM profile was not generated for the sequence in the validation dataset. We used the model based on the second-best feature, AAC, to extract prediction values for that sequence. The maximum AUROC achieved was 0.95 on the validation set.

3.5 | Ensemble Approach

The previous results indicate that both the similarity-based approach and ML-based models have their own advantages and disadvantages. Therefore, we endeavored to create a technique that integrates the strengths of both approaches. We explored hybrid methodologies, combining ML with BLAST scores and ML with BLAST and motif scores. we observed increased performance by combining the ML with the BLAST score using an *E*

TABLE 3 | The performance of different machine learning techniques-based models using the PSSM feature of protein sequences.

	Training dataset					Validation dataset				
Models	**Sen**	**Spec**	**Acc**	**AUROC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUROC**	**MCC**
RF	0.86	0.95	0.90	0.96	0.80	0.86	0.90	0.88	0.95	0.76
SVC	0.85	0.90	0.88	0.92	0.76	0.86	0.87	0.86	0.90	0.73
ET	0.86	0.96	0.91	0.95	0.82	0.86	0.95	0.91	0.95	0.81
XGB	0.88	0.90	0.89	0.94	0.78	0.86	0.87	0.86	0.95	0.73
BC	0.87	0.91	0.88	0.94	0.77	0.86	0.90	0.88	0.93	0.76

Abbreviations: Acc, accuracy; AUROC, area under the receiver operating characteristic curve; BC, bagging classifier; ET, extra trees; MCC, Matthew correlation coefficient; PSSM, position-specific scoring matrix; RF, random forest; Sen, sensitivity; Spec, specificity; SVC, support vector classifier; XGB, eXtreme gradient boosting.

TABLE 4 | The performance of the hybrid method developed by combining machine learning and BLAST-based approach, evaluated on a validation dataset.

Methods		**Performance**				
Models	**Features**	**Sen**	**Spec**	**Acc**	**AUROC**	**MCC**
SVC	AAC	0.86	0.90	0.88	0.96	0.76
BC	DPC	0.89	0.82	0.85	0.95	0.71
RF	PSSM	0.89	0.90	0.89	0.98	0.78

Abbreviations: AAC, amino acid composition; Acc, accuracy; AUROC, area under the receiver operating characteristic curve; BC, bagging classifier; BLAST, Basic Local Alignment Search Tool; DPC, dipeptide composition; MCC, Matthew correlation coefficient; PSSM, position-specific scoring matrix; RF, random forest; Sen, sensitivity; Spec, specificity; SVC, support vector classifier.

TABLE 5 | The performance of the hybrid method developed by combining machine learning and motif-based approach on a validation dataset.

Methods		**Scores**				
Models	**Features**	**Sen**	**Spec**	**Acc**	**AUROC**	**MCC**
SVC	AAC	0.78	0.87	0.83	0.90	0.65
RF	DPC	0.83	0.82	0.83	0.92	0.65
RF	PSSM	0.86	0.87	0.87	0.95	0.73

Abbreviations: AAC, amino acid composition; Acc, accuracy; AUROC, area under the receiver operating characteristic curve; DPC, dipeptide composition; MCC, Matthew correlation coefficient; PSSM, position-specific scoring matrix; RF, random forest; Sen, sensitivity; Spec, specificity; SVC, support vector classifier.

value of 10^{-3}. Specifically, when utilizing RF with features derived from PSSM, the AUROC improved to 0.98 on the validation dataset. Employing the ML model alongside exclusive positive motifs also resulted in an AUROC of 0.95 on the validation dataset. Combining ML with BLAST and motifs yielded an AUROC of 0.98 on the validation dataset. The performance of different ensemble approaches has been illustrated in Tables 4–6.

4 | Web Server Development for PDR Classification

We have constructed a web server called "PlantDRPpred" (https://webs.iiitd.edu.in/raghava/plantdrppred) for categorizing proteins into plant disease resistance proteins (PDRPs) or non-PDRPs groups, utilizing the top-performing model identified in our study. The server incorporates various modules: "Predict," "Design," "Protein scan," and "BLAST scan." Users can categorize provided sequences as PDR or non-PDR using the "Predict" module. The "Design" module allows users to construct all possible PDR analogs that can be created from the provided sequence. Through the utilization of the "Protein Scan" module, users have the ability to scan or identify specific sections within an amino acid sequence that correspond to either PDR or non-PDR. The "Blast scan" module allows users to search their query sequence against a database containing established PDR information. The classification of the query sequence as either PDR or non-PDR is determined by whether or not there is a match or hit in the database. If the sequence corresponds to or aligns with a recognized protein in the database, it is classified as PDR; if no alignment is detected, it is classified as non-PDR. Users have the option to submit protein sequences for analysis in two different formats. The user can input a file in FASTA format or directly paste multiple sequences. The server will then generate predictions and detailed reports using the integrated modules. The web server can handle up to 50 queries concurrently. Users

TABLE 6 | The performance of a hybrid method that combines machine learning, motif, and BLAST-based approaches on a validation dataset.

Methods		Scores				
Models	**Features**	**Sen**	**Spec**	**Acc**	**AUROC**	**MCC**
SVC	AAC	0.86	0.90	0.88	0.96	0.76
BC	DPC	0.89	0.82	0.85	0.95	0.71
RF	PSSM	0.89	0.95	0.92	0.98	0.84

Abbreviations: AAC, amino acid composition; Acc, accuracy; AUROC, area under the receiver operating characteristic curve; BC, bagging classifier; BLAST, Basic Local Alignment Search Tool; DPC, dipeptide composition; MCC, Matthew correlation coefficient; PSSM, position-specific scoring matrix; RF, random forest; Sen, sensitivity; Spec, specificity; SVC, support vector classifier.

TABLE 7 | Benchmarking of existing tools and tools proposed in this study on dataset not used in training of models.

Models	Sen	Spec	Acc	AUROC	MCC
prPred-DRLF	0.85	0.83	0.84	0.93	0.68
prPred	0.24	0.98	0.64	0.86	0.34
PlantDRPpred[a]	0.89	0.95	0.92	0.98	0.84

Abbreviations: Acc, accuracy; AUROC, area under the receiver operating characteristic curve; MCC, Matthew correlation coefficient; Sen, sensitivity; Spec, specificity.

[a]PlantDRPpred results demonstrate that our best model runs on the validation dataset.

can process a large number of protein sequences on their own computational resources by utilizing either the pip package or the standalone package.

5 | Comparison With Other Tools

Our study evaluated the performance of various available tools for predicting R proteins. Some tools, NBSpred [6], DRPPP [25], ResCap [28], and StackRPred [10], are not available for evaluation. Available tools are trained on outdated datasets encompassing a limited range of plant disease-resistance protein classes, which restricts their ability to accurately predict all PDR protein classes. Our webserver, PlantDRPpred, overcomes these limitations and accurately predicts plant disease-resistance proteins. Table 7 represents the comparison of our model with the existing methods. The protein sequences in our validation dataset may already be present in the training dataset of existing tools. We have generated a dataset comprising proteins not utilized in the training or testing of existing methods and used that for benchmarking.

6 | Discussion

Plants encounter various types of biotic stress, including attacks from pathogens such as fungi, bacteria, insects, and viruses. These biotic stressors can severely affect plant health, reduce crop yields, and compromise food security. In response to these challenges, plants have evolved sophisticated defense mechanisms in which PDR plays a crucial role, recognizing specific pathogen-derived molecules and activating defense responses. This recognition triggers a cascade of signaling events to activate various defense mechanisms. Recent advancements in genomics and biotechnology have facilitated the identification and characterization of numerous PDR proteins across multiple plant species, but they are a time-consuming process. In the past few years, some approaches, such as DRPPP, prPred, stackRPred, and NBSpred DRPPP, trained their model with a smaller number of protein sequences and typically relied on either ML-based approaches or similarity searches alone and did not eliminate the redundancy of positive data. NBSpred, DRPPP, ResCap, and StackRPred tools are no longer available for public use, which restricts their accessibility and utility for researchers. Only prPred-DRLF provides a web server, but it is no longer available. This study aims to establish an in-silico approach to predict PDRPs using the ensemble model, so we retrieved 199 sequences from PRGdb4.0 as positive data. Negative data was extracted from UniProt, focusing on the Embryophyta clade since the positive data encompassed Angiosperm and Bryophyta. From the 5674 negative sequences retrieved from UniProt, we selected 199 sequences to generate a balanced dataset, ensuring that those involved in the defense system were excluded.

This study explored various methods to predict PDR proteins. We developed ML-based models to distinguish PDRs from non-PDRs by utilizing multiple features. These features included composition-based properties such as AAC and DPC, as well as evolutionary information-based properties derived from PSSM. We employed a variety of classifiers, including SVC, RF, XGB, ET, and BC, to achieve this. Additionally, we used alignment-based approaches, such as BLAST and the Motif-search approach, to annotate protein sequences. However, these approaches demonstrated low sensitivity or coverage. Consequently, we adopted an ensemble approach that integrated the ML model with BLAST with motif search, exploring all possible combinations of methods to improve the accuracy and reliability of our predictions. The highest performance was observed with a hybrid model combining the PSSM-based ML model and BLAST score. Our hybrid approach combines ML techniques with BLAST, effectively capturing both sequence

similarity and evolutionary characteristics. Although BLAST provides strong similarity but limited coverage, ML techniques have their own limitations. By leveraging the strengths of both techniques, our hybrid strategy significantly improves performance. For instance, our model's AUROC increased from 0.95 to 0.98, demonstrating the enhanced predictive power of our approach, and we optimized the threshold by minimizing the difference between the sensitivity and specificity of our best model (RF + BLAST). This approach led us to select a threshold of 0.38, which provides a balanced performance by equitably managing both true positive and true negative rates. This hybrid approach was implemented in the freely accessible web server.

The model is trained on a dataset of PDR proteins, which may not represent the full diversity of PDR proteins across different plant species. This limitation in dataset size may affect the model's ability to generalize its predictions to other PDR proteins. PlantDRPpred may not fully capture the complex interactions and regulatory mechanisms of PDR proteins. Experimental validation is essential to confirm the biological relevance of the model's predictions. Despite its ability to determine if an input protein sequence is a PDR, PlantDRPpred does not offer details on binding locations or affinity scores. We have constructed a comprehensive model to predict PDR by incorporating sequences from different plants. However, the wide range of functional and structural variations among these proteins implies that PDRs may possess distinct characteristics and perform specific roles. Hence, the utilization of function and structure-specific techniques has the potential to improve the accuracy of PDR predictions.

We believe that accurately predicting PDR proteins can play a pivotal role in crop protection strategies. By identifying key resistance genes, these predictions can guide breeding programs or genetic engineering efforts to enhance crop resilience against various pathogens. This approach has already demonstrated its effectiveness in developing new wheat varieties with resistance genes to fight against wheat stem rust [31]. This has the potential to contribute to sustainable agricultural practices by reducing the reliance on chemical pesticides and enhancing the natural defense mechanisms of crops. Understanding the specific roles and interactions of these proteins in plant immune responses can provide new insights into plant–pathogen interactions, guiding the development of innovative approaches to plant disease management.

7 | Conclusion

Our methodology for accurately predicting PDR proteins improves our capacity to recognize them better than the available methods. We used the ML model and BLAST approach for the final ensemble model. We have created a web server called PlantDRPpred and a standalone tool to assist researchers in identifying PDR proteins. If query sequences resembling PDR proteins are present, a prediction score based on similarity will be provided. We believe our work will contribute to the annotation of PDRs and provide valuable support for research in plant pathology.

Author Contributions

P.S.G. and N.K. collected and processed the dataset. P.S.G. developed computer programs and implemented the algorithms and prediction models. P.S.G. and S.C. created the front-end and back-end of the web server. P.S.G., N.B., and G.P.S.R. wrote the manuscript. G.P.S.R. conceived and coordinated the project and provided overall supervision.

Acknowledgments

The authors are thankful to the University Grants Commission (UGC), Council of Scientific and Industrial Research (CSIR), and Department of Science & Technology (DST) for their generous fellowships and financial support and to Indraprastha Institute of Information Technology Delhi for infrastructure. The authors would like to acknowledge the Department of Biotechnology (DBT) for the infrastructure grant awarded to the institute. Furthermore, they would like to acknowledge BioRender.com for creating the figures utilized in this work.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The dataset used in this study is accessible from the "PlantDRPpred" web server at https://webs.iiitd.edu.in/raghava/plantdrppred/downloads.php. The source code can be obtained from https://github.com/raghavagps/PlantDRPpred.

BioRxiv doi

https://doi.org/10.1101/2024.07.22.604583

References

1. J. D. G. Jones and J. L. Dangl, "The Plant Immune System," *Nature* 444 (2006): 323–329.

2. K. E. Hammond-Kosack and K. Kanyuka, "Resistance Genes (*R* Genes) in Plants," *Encyclopedia of Life Sciences. Portico.* (2007), https://doi.org/10.1002/9780470015902.a0020119.

3. C. Ameline-Torregrosa, B.-B. Wang, M. S. O'bleness, et al., "Identification and Characterization of Nucleotide-Binding Site-Leucine-Rich Repeat Genes in the Model Plant *Medicago truncatula*," *Plant Physiology* 146 (2008): 5–21.

4. B. C. Meyers, A. Kozik, A. Griego, H. Kuang, and R. W. Michelmore, "Genome-Wide Analysis of NBS-LRR–Encoding Genes in Arabidopsis[W]," *Plant Cell* 15 (2003): 809–834.

5. B. C. Meyers, M. Morgante, and R. W. Michelmore, "TIR-X and TIR-NBS Proteins: Two New Families Related to Disease Resistance TIR-NBS-LRR Proteins Encoded in Arabidopsis and Other Plant Genomes," *Plant Journal* 32 (2002): 77–92.

6. S. K. Kushwaha, P. Chauhan, K. Hedlund, and D. Ahrén, "NBSPred: A Support Vector Machine-Based High-throughput Pipeline for Plant Resistance Protein NBSLRR Prediction," *Bioinformatics* 32 (2016): 1223–1225.

7. B. Steuernagel, F. Jupe, K. Witek, J. D. G. Jones, and B. B. H. Wulff, "NLR-Parser: Rapid Annotation of Plant NLR Complements," *Bioinformatics* 31 (2015): 1665–1667.

8. T. Pal, V. Jaiswal, and R. S. Chauhan, "DRPPP: A Machine Learning Based Tool for Prediction of Disease Resistance Proteins in Plants," *Computers in Biology and Medicine* 78 (2016): 42–48.

9. Y. Wang, P. Wang, Y. Guo, S. Huang, Y. Chen, and L. Xu, "PrPred: A Predictor to Identify Plant Resistance Proteins by Incorporating k-Spaced

Amino Acid (Group) Pairs," *Frontiers in Bioengineering and Biotechnology* 8 (2020): 645520.

10. Y. Chen, Z. Li, and Z. Li, "Prediction of Plant Resistance Proteins Based on Pairwise Energy Content and Stacking Framework," *Frontiers in Plant Science* 13 (2022): 912599.

11. J. Calle García, A. Guadagno, A. Paytuvi-Gallart, et al., "PRGdb 4.0: An Updated Database Dedicated to Genes Involved in Plant Disease Resistance Process," *Nucleic Acids Research* 50 (2022): D1483–D1490.

12. A. Bateman, M.-J. Martin, S. Orchard, et al., "UniProt: The Universal Protein Knowledgebase in 2021," *Nucleic Acids Research* 49 (2021): D480–D489.

13. A. Garg and G. P. S. Raghava, "A Machine Learning Based Method for the Prediction of Secretory Proteins Using Amino Acid Composition, Their Order and Similarity-Search," *In Silico Biology* 8 (2008): 129–140.

14. J. D. Bendtsen, L. J. Jensen, N. Blom, G. Von Heijne, and S. Brunak, "Feature-Based Prediction of Non-Classical and Leaderless Protein Secretion," *Protein Engineering, Design & Selection* 17 (2004): 349–356.

15. N. Sharma, S. Patiyal, A. Dhall, A. Pande, C. Arora, and G. P. S. Raghava, "AlgPred 2.0: An Improved Method for Predicting Allergenic Proteins and Mapping of IgE Epitopes," *Briefings in Bioinformatics* 22 (2021): bbaa294.

16. D. Kaur, C. Arora, and G. P. S. Raghava, "A Hybrid Model for Predicting Pattern Recognition Receptors Using Evolutionary Information," *Frontiers in Immunology* 11 (2020): 71.

17. L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: Accelerated for Clustering the Next-Generation Sequencing Data," *Bioinformatics* 28 (2012): 3150–3152.

18. A. Pande, S. Patiyal, A. Lathwal, et al., "Pfeature: A Tool for Computing Wide Range of Protein Features and Building Prediction Models," *Journal of Computational Biology* 30 (2023): 204–222.

19. J. Wang, B. Yang, J. Revote, et al., "POSSUM: A Bioinformatics Toolkit for Generating Numerical Sequence Feature Descriptors Based on PSSM Profiles," *Bioinformatics* 33 (2017): 2756–2758.

20. M. Kumar, M. M. Gromiha, and G. P. Raghava, "Identification of DNA-binding Proteins Using Support Vector Machines and Evolutionary Profiles," *BMC Bioinformatics* 8 (2007): 463.

21. S. Altschul, "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs," *Nucleic Acids Research* 25 (1997): 3389–3402.

22. C. Camacho, G. Coulouris, V. Avagyan, et al., "BLAST+: Architecture and Applications," *BMC Bioinformatics* 10 (2009): 421.

23. C. Vens, M.-N. Rosso, and E. G. J. Danchin, "Identifying Discriminative Classification-Based Motifs in Biological Sequences," *Bioinformatics* 27 (2011): 1231–1238.

24. A. Dhall, S. Patiyal, N. Sharma, S. S. Usmani, and G. P. S. Raghava, "Computer-Aided Prediction and Design of IL-6 Inducing Peptides: IL-6 Plays a Crucial Role in COVID-19," *Briefings in Bioinformatics* 22 (2021): 936–945.

25. S. K. Kushwaha, I. Åhman, and T. Bengtsson, "ResCap: Plant Resistance Gene Prediction and Probe Generation Pipeline for Resistance Gene Sequence Capture," *Bioinformatics Advances* 1 (2021): vbab033.

26. A. Arora, S. Patiyal, N. Sharma, N. L. Devi, D. Kaur, and G. P. S. Raghava, "A Random Forest Model for Predicting Exosomal Proteins Using Evolutionary Information and Motifs," *Proteomics* 24 (2024): e2300231.

27. Q.-H. Kha, T.-O. Tran, T.-T.-D. Nguyen, V.-N. Nguyen, K. Than, and N. Q. K. Le, "An Interpretable Deep Learning Model for Classifying Adaptor Protein Complexes From Sequence Information," *Methods (San Diego, Calif.)* 207 (2022): 90–96.

28. Q.-H. Kha, V.-H. Le, T. N. K. Hung, N. T. K. Nguyen, and N. Q. K. Le, "Development and Validation of an Explainable Machine Learning-Based Prediction Model for Drug-Food Interactions From Chemical Structures," *Sensors (Basel)* 23 (2023): 3962.

29. S. Patiyal, A. Dhall, K. Bajaj, H. Sahu, and G. P. S. Raghava, "Prediction of RNA-Interacting Residues in a Protein Using CNN and Evolutionary Profile," *Briefings in Bioinformatics* 24 (2023): bbac538.

30. S. Zhou, Y. Zhou, T. Liu, J. Zheng, and C. Jia, "PredLLPS_PSSM: A Novel Predictor for Liquid-Liquid Protein Separation Identification Based on Evolutionary Information and a Deep Neural Network," *Briefings in Bioinformatics* 24 (2023): bbad299.

31. U. Bansal, H. Bariana, D. Wong, et al., "Molecular Mapping of an Adult Plant Stem Rust Resistance Gene Sr56 in Winter Wheat Cultivar Arina," *Theoretical and Applied Genetics* 127 (2014): 1441–1448.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.