Computers in Biology and Medicine 170 (2024) 108081

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

DRBpred: A sequence-based machine learning method to effectively predict DNA- and RNA-binding residues

Md Wasi Ul Kabir, Duaa Mohammad Alawad, Pujan Pokhrel, Md Tamjidul Hoque *

Department of Computer Science, University of New Orleans, New Orleans, LA, USA

ARTICLE INFO

Keywords:
DNA-Binding proteins
RNA-Binding proteins
Machine learning

ABSTRACT

DNA-binding and RNA-binding proteins are essential to an organism's normal life cycle. These proteins have diverse functions in various biological processes. DNA-binding proteins are crucial for DNA replication, transcription, repair, packaging, and gene expression. Likewise, RNA-binding proteins are essential for the post-transcriptional control of RNAs and RNA metabolism. Identifying DNA- and RNA-binding residue is essential for biological research and understanding the pathogenesis of many diseases. However, most DNA-binding and RNA-binding proteins still need to be discovered. This research explored various properties of the protein sequences, such as amino acid composition type, Position-Specific Scoring Matrix (PSSM) values of amino acids, Hidden Markov model (HMM) profiles, physiochemical properties, structural properties, torsion angles, and disorder regions. We utilized a sliding window technique to extract more information from a target residue's neighbors. We proposed an optimized Light Gradient Boosting Machine (LightGBM) method, named DRBpred, to predict DNA-binding and RNA-binding residues from the protein sequence. DRBpred shows an improvement of 112.00 %, 33.33 %, and 6.49 % for the DNA-binding test set compared to the state-of-the-art method. It shows an improvement of 112.50 %, 16.67 %, and 7.46 % for the RNA-binding test set regarding Sensitivity, Mathews Correlation Coefficient (MCC), and AUC metric.

1. Introduction

Protein-DNA and protein-RNA interactions are important in various biological processes. This includes DNA replication and repair, gene regulation, transcription, post-transcriptional control of RNAs and RNA metabolism, and other DNA-related and RNA-related biological activities [1–4]. Understanding how and why proteins interact with DNA and RNA requires the identification of DNA-binding and RNA-binding proteins. Many experimental techniques, such as nuclear magnetic resonance, X-ray crystallography, and chromatin immunoprecipitation on microarrays, can identify DNA-binding and RNA-binding proteins [5]. However, the experimental techniques to determine DNA-binding and RNA-binding proteins are time-consuming and labor-intensive [6]. Given the limitations of wet experiments for determining DNA-binding and RNA-binding proteins, computational methods for identifying putative DNA-binding and RNA-binding proteins have become increasingly important in recent years. Recent breakthroughs in genomic and proteomic techniques have recently generated numerous DNA-binding and RNA-binding protein sequences [7]. For example, in 2014, there were more than ten times as many DNA-binding proteins in the UniProt database as in 2000 [8]. These massive amounts of data lay the groundwork for research into computational approaches for identifying DNA-binding and RNA-binding proteins.

In the existing literature, many recent methods [9,10] rely not just on the protein sequence but also on the protein's experimentally derived or predicted 3D structure. However, the number of experimentally derived structures (DNA and RNA complexes) is limited. Addressing this gap, our work introduces DRBpred, a novel approach for predicting DNA-binding and RNA-binding residues using only protein sequences. We studied several properties of protein sequences, including amino acid composition, evolutionary profiles (PSSM and HMM values of amino acids), physicochemical properties, structural properties, torsion angles, and disorder values. We ranked the features to determine which features contribute most to our trained model. We employed a recursive feature elimination (RFE) technique combined with SHAP (Shapley Additive exPlanations) values to select a subset of important features. The sliding window technique was used to obtain as much information as possible about the target and context residues. The features were concatenated to

* Corresponding author.
E-mail addresses: mkabir3@uno.edu (M.W.U. Kabir), dmalawad@uno.edu (D.M. Alawad), ppokhre1@uno.edu (P. Pokhrel), thoque@uno.edu (M.T. Hoque).

https://doi.org/10.1016/j.compbiomed.2024.108081
Received 16 August 2023; Received in revised form 12 January 2024; Accepted 27 January 2024
Available online 29 January 2024

achieve superior predictive performance. In addition, an optimized LightGBM (Light Gradient Boosting Machine) classifier-based predictor was trained as the machine-learning method for the classification task. We found that the proposed method outperformed the existing state-of-the-art methods. Finally, we used Local Interpretable Model-Agnostic Explanations (LIME) to explain the trained model and analyze the effects of features.

2. Related work

Several methods have been proposed in the literature to identify DNA-binding and RNA-binding sites in proteins. The three types of features used in these prediction methods are sequence, structure, and evolutionary. Using evolutionary features was hard to compute due to the lack of computing power. The structural and sequence-based features were mostly used for prediction. Ahmad et al. utilized only sequence features to predict protein-DNA-binding [9]. Cai and Lin employed the SVM algorithm to predict DNA-binding proteins, utilizing a protein's amino acid composition, hydrophobicity, and solvent-accessible surface area correlations as input features [11]. In more recent work, Zou et al. introduced a sequence-based protocol that integrates informative features from different scales to train an SVM model for the prediction of DNA-binding proteins [12]. The random forest (RF) algorithm, which is a useful machine learning classifier, was also used to predict DNA-binding proteins. Lou et al. applied the RF algorithm for the prediction of DNA-binding proteins, with predicted relative solvent accessibility, predicted secondary structure, and position-specific scoring matrix serving as the primary sequence features [13]. Zhang et al. proposed DNA-Prot, a predictor for DNA-binding proteins. It employs an SVM classifier and a comprehensive set of features categorized into six groups: primary sequence-based, evolutionary profile-based, predicted relative solvent accessibility-based, predicted secondary structure-based, physicochemical property-based, and biological function-based features [14]. In addition, Yan et al. presented the DRNApred tool [15] that can distinguish between DNA-binding and RNA-binding residues and proteins. It employs a collection of features extracted from a diverse set of sources of sequence-derived information extracted from a dataset with both DNA-binding and RNA-binding proteins. This information contains amino acid types, amino acid physicochemical properties, evolutionary profiles, potential intrinsic disorder, secondary structure, and solvent accessibility. DRNApred lowers cross predictions and predicts potentially higher-quality false positives near-native binding residues. Moreover, Seungwoo et al. introduced DP-Bind, a method for predicting DNA-binding sites in a DNA-binding protein based on the protein's amino acid sequence. DP-Bind implements three machine learning methods: support vector machine (DP-Bind(SVN)), kernel logistic regression (DP-Bind (klr)), and penalized logistic regression(DP-Bind (plr)) [16]. In DP-Bind, predictions can be made using either the input sequence alone or an autonomously created profile of the input sequence's evolutionary conservation in the form of a PSI-BLAST position-specific scoring matrix (PSSM). Wang et al. proposed the BindN + method, which employs two SVM models to predict RNA-binding and DNA-binding sites; each model performs better on its respective type of proteins [17].

Several studies have indicated the significance of evolutionary features in the detection of DNA-binding proteins [18–20]. Methods lacking these evolutionary features tend to exhibit lower accuracy, often resulting in classifier bias due to imbalanced sample numbers. Thus, the inclusion of evolutionary information to predict DNA-binding residues can improve accuracy. Computing power has increased dramatically in the last decade, which makes it much easier to compute evolutionary features, which are often time-consuming. Position-Specific Scoring Matrix (PSSM) is used to represent evolutionary features. They are usually calculated in one of two ways: (a) Concatenation methods that encode the residues by concatenating PSSM scores in a sliding window (b) Combination methods, which encode residues by combining PSSM scores with other physiochemical properties such as hydrophobicity, molecular mass, torsion angles, and other frequency profiles in a sliding window. Zhou et al. [21] introduce a residue-encoding technique called Position Specific Score Matrix Relation Transformation (PSSM-RT), which encodes residues by considering their evolutionary relationships. Deng et al. proposed the PDRLGB method that predicts binding residues in protein-DNA complexes using a light gradient-boosting machine (LightGBM) [9]. The author used an incremental feature selection with the random forest algorithm to find the best subset of features and trained a light gradient boosting machine. However, their method is dependent not only on the protein sequence but also on the experimentally derived 3D structure of the protein. They extracted structural features from the three-dimensional protein structure using the DSSP [22]. Zhang et al. [23] proposed the StackPDB method for predicting DNA-binding Proteins. The StackPDB method extracts pseudo amino acid composition (PseAAC), pseudo-position-specific scoring matrix (PsePSSM), position-specific scoring matrix-transition probability composition (PSSM-TPC), evolutionary distance transformation (EDT), and residue probing transformation (RPT) features from protein sequences. The authors selected a subset of the features using extreme gradient boosting-recursive feature elimination (XGB-RFE) and employed a stacked ensemble classifier consisting of XGBoost, LightGBM, and SVM for DNA-binding protein prediction. Ali et al. introduce DP-BINDER, a computational method for identifying DBPs based on physicochemical and evolutionary information. It involves extracting key features from protein sequences using normalized Moreau-Broto autocorrelation (NMBAC), position-specific scoring matrix-transition probability composition (PSSM-TPC), and pseudo position-specific scoring matrix (PsePSSM). These features are refined using support vector machine recursive feature elimination and correlation bias reduction (SVM-RFE + CBR) and analyzed using random forest (RF) and support vector machine (SVM). DP-BINDER demonstrated an accuracy of 92.46 % with the jackknife method.

Many research works also apply deep learning methods to predict the DNA-binding and RNA-binding Residues. Hendrix et al. [24] constructed and evaluated a deep-learning model to estimate the likelihood that a voxel on the protein surface is a DNA-binding site. Based on three distinct evaluation datasets, the results indicate that the model beats a number of earlier methods on two widely used datasets. In Ref. [25]the authors presented EL LSTM, an approach for DNA-binding residue prediction that consists of two main components: Long Short-Term Memory (LSTM) and an ensemble learning-based classifier. LSTM uses a bi-gram model to learn pairwise relationships between residues before learning feature vectors for all residues. Then, an ensemble learning-based classifier is developed to address the data imbalance problem in binding residue predictions. To achieve balanced samples, they used a variant of the bagging strategy in ensemble learning. Despite the existence of numerous methods, the classification score remains low, indicating room for improvement. Additionally, some of these methods rely on the three-dimensional structure of proteins. Certain methods only offer protein-level predictions rather than residue-level predictions. This motivates us to explore this problem further and develop a machine-learning method capable of accurately predicting DNA-binding and RNA-binding residues.

3. Proposed method

This section formally discusses the data collection methods, feature extraction, machine learning methods, feature selection, and performance evaluation metrics for predicting DNA-binding and RNA-binding residues.

3.1. Dataset

Throughout this study, we used the processed dataset that was used in Ref. [15]. In Fig. 1, we summarized the steps the authors in Ref. [15]

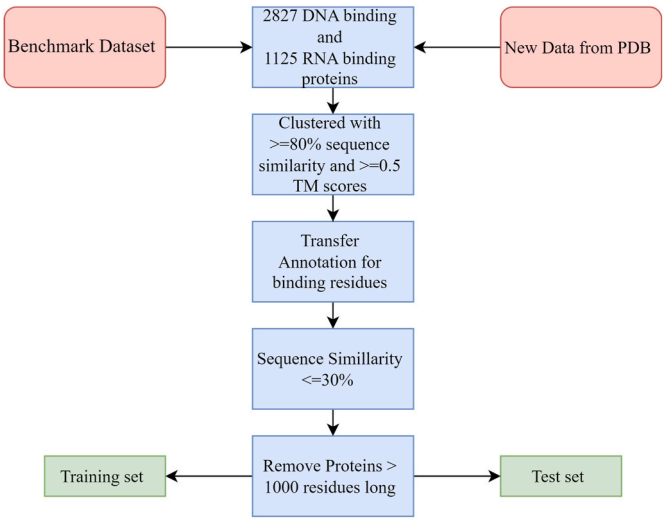

Fig. 1. Illustrates the steps of creating the Training and Test datasets. The dataset was created with both benchmark datasets and data collected from the Protein Data Bank (PDB).

had used to prepare the dataset. The original dataset was collected from 564 protein–DNA, 72 protein–RNA, and 16 protein–DNA–RNA high-resolution (better than 2.5) complexes PDB. Then, an extra 892 DNA-binding and 145 RNA-binding chains with the previous dataset were added, yielding 2827 DNA-binding and 1125 RNA-binding chains.

Next, the dataset is clustered to select proteins that share ≥80 % sequence similarity and ≥0.5 TM scores. Annotations were moved between proteins in the same cluster [15]. All chains' DNA-binding and RNA-binding residues are transferred in the same cluster into a representative chain with the largest number of binding residues. To reduce the sequence similarity between training and test datasets, the test proteins are filtered by removing every sequence that shares >30 % sequence similarity with any training sequence based on pairwise sequence similarity [15]. Finally, the long proteins from the training and test datasets were removed because of the existing predictors of DNA- and RNA-binding residues that could not complete predictions for proteins that are over 1000 residues long [15]. A version of the test dataset was also created without transferring annotations of binding residues. Table 1 summarizes the number of proteins, RNA-binding, and DNA-binding residue annotations.

3.2. Feature extraction

We extracted a variety of features to represent proteins. We encompassed important properties such as sequence information, predicted structural details, and evolutionary information. These features offer relevant insights into the characteristics of the residues. Previous studies in the literature have indicated that information concerning the correct folding of a protein is embedded within its amino acid sequence and the disorder contents [26]. Furthermore, details pertaining to the binding affinity of proteins are encoded within the evolutionary information, along with other structural and physicochemical properties [27–30]. Consequently, these features were integrated with the evolutionary attributes to enhance prediction accuracy.

We collected a total of 119 features using various feature encoding techniques, as depicted in Fig. 2. Utilizing different feature-encoding techniques, the subsequent section briefly describes the collected features.

Table 1
The number of DNA- and RNA-binding residues in the Training and Test Dataset.

Dataset	No. of proteins	No. of Non-binding residues	No. of DNA-binding residues	No. of RNA-binding residues
Training	488	95161	7823 (7.6 %)	4699 (4.6 %)
Test	82	17925	968 (5.1 %	808 .2 %)

3.2.1. Physiochemical properties

The physiochemical characteristics of a protein are the inherent properties of its constituent amino acids. Previous research studies [31, 32] have shown the influence amino acid physiochemical properties have on the activity of transcription factors and how they regulate their interactions with other proteins. In this study, we have extracted seven concise numerical patterns from the work of [31] to represent key aspects of amino acid properties. The include polarity, secondary structure propensity, molecular volume, codon diversity, and electrostatic charge. These patterns serve as features to capture the distinctive attributes of each amino acid.

3.2.2. Residue properties

We represented each of the 20 standard amino acid (AA) types with a unique feature to effectively capture the amino acid composition of individual residues within a protein sequence. Previous research studies [33–35] have highlighted the significance of this feature in addressing bioinformatics problems. We encoded terminal residues, specifically those five residues located from the N and C termini. The values ranged from −1.0 to −0.2 and + 0.2 to +1.0, creating a distinctive feature for each residue [33].

3.2.3. Evolutionary properties

As demonstrated in previous research, the evolutionary profile is a crucial factor in post-translational modifications (PTM). This includes DNA-binding and RNA-binding activity [27–30]. In our study, we acquired the evolutionary profile of the protein sequence through a normalized position-specific scoring matrix (PSSM) obtained from BLAST (PSI-BLAST) [36]. This PSSM is represented with a 20-dimensional matrix, capturing evolutionary patterns in multiple alignments and storing scores for each position in the alignment. High scores indicate highly conserved positions, while scores near zero or negative values indicate weakly conserved positions. We extended the PSSM scores to calculate monogram (MG) and bi-gram (BG) features. MG and BG features can be used to describe a segment of a protein sequence that exhibits conservation in terms of transition probabilities from one amino acid to another [37]. We extracted 1-dimensional MG features and 20-dimensional BG features from the DisPredict2 program and incorporated them into our analysis. We calculated the close neighbor correlation coefficient based on the PSSM scores. We obtained 30 Hidden Markov Model (HMM) profile-based evolutionary features for the protein sequence. To identify distantly related sequences, profile Hidden Markov Models (HMMs) transform a multiple sequence alignment into a specialized scoring system tailored for searching databases [38]. Numerous studies have emphasized the importance of evolutionary features in characterizing protein properties [27–30,39,40]. Methods that do not consider evolutionary features typically exhibit lower accuracy. The classifier can be biased due to imbalanced sample numbers. Therefore, the inclusion of evolutionary information in predicting DNA-binding and RNA-binding residues can enhance accuracy.

3.2.4. Structural properties

Local structural characteristics, such as the predicted secondary structure (SS) and accessible surface area (ASA) of amino acids, have been widely employed in addressing various biological challenges, including DNA- and RNA-binding residue prediction. In our study, we utilized the Dispredict2 [35] and SPOT-Disorder2 [41] programs to acquire predicted ASA values and SS probabilities for helix (H), coil (C), and beta-sheet (E) at the individual residue level. Additionally, we obtained a separate set of SS probabilities for E, C, and E at the residue level from the Dispredict2 and SPOT-Disorder2 programs.

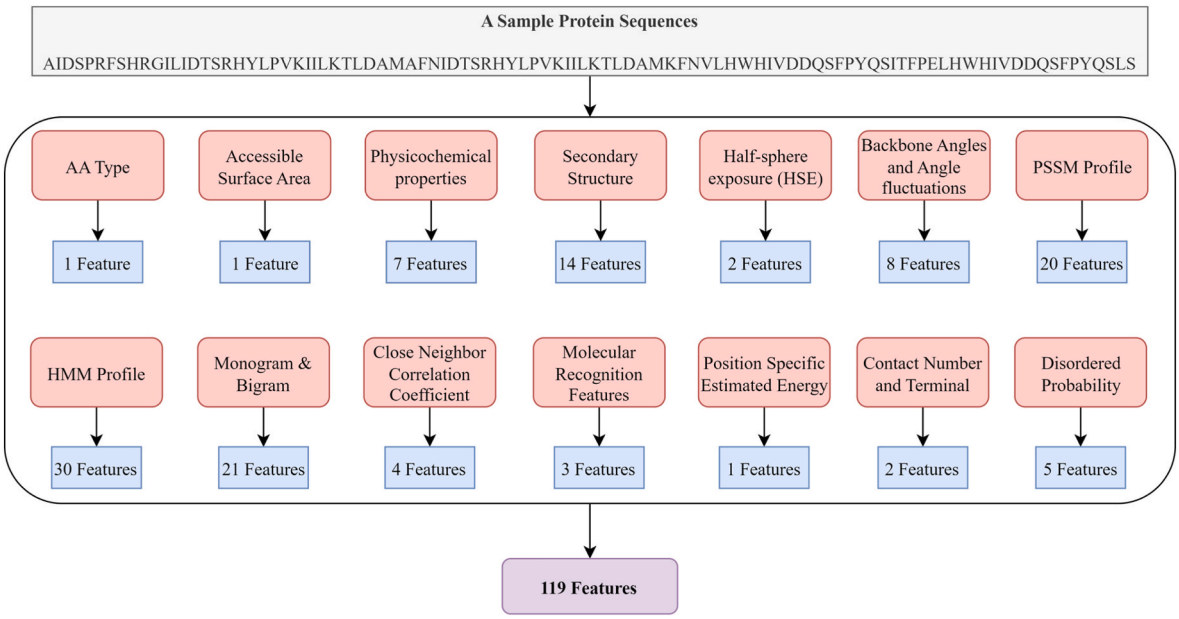

Fig. 2. Illustration of encoding the protein residues into a feature vector of 119 features utilizing various feature encoding techniques. The feature vector includes amino acid composition type, evolutionary features, physicochemical, structural properties, torsion angles, and disorder probabilities.

3.2.5. *Flexibility properties*

Protein molecules exhibit varying levels of flexibility within their 3D structures, often expressed as fluctuations in the Cartesian coordinates of the protein backbone and defined by two torsion angles Φ and Ψ. The fluctuations in backbone torsion angles have proven valuable in developing several computational methods [40,41]. We acquired two features related to backbone angle fluctuations, specifically dphi (ΔΦ) and dpsi (ΔΨ), using the Dispredict2 and SPOT-Disorder2 programs [35,41]. Previous research has established that intrinsically disordered regions (IDRs) contain post-translational modification (PTM) site-sorting signals and play a crucial role in regulating protein structures and functions, i.e., DNA- and RNA-binding proteins [42–44]. In our study, we represented each amino acid in a protein with a disorder probability obtained from a disorder predictor. We also included Molecular Recognition Features (MoRFs). These are short, interaction-prone segments of protein disorder that transition from disorder to order upon specific binding, representing a specific class of intrinsically disordered regions with molecular recognition and binding functions.

3.2.6. *Energy profile*

A method for estimating the position-specific estimated energy (PSEE) of amino acid residues solely based on sequence information was developed by Iqbal et al. [35]. The authors incorporate the contact energy and predict relative solvent accessibility (RSA) to determine the PSEE. Their work showcased how PSEE can effectively distinguish between structured and unstructured regions within a protein, including intrinsically disordered regions. Additionally, PSEE can be employed to identify functional binding regions within a protein. We incorporated the PSEE score per amino acid as a feature in our study.

3.3. *Machine learning algorithms*

In this study, we have explored the following seven Machine Learning Methods.

- K-nearest Neighbors Classifier (KNN): KNN learns from the K number of training samples in the feature space that are the closest distance to the target point. The classification decision is based on the neighbors' majority votes. K was set to 5 as a default value, and all neighbors were equally weighted [45].
- Random Forest Classifier (RF): Random forest [46] is a supervised learning algorithm that employs ensemble learning techniques for classification tasks. It is a meta-estimator that aggregates many decision trees (bagging). The random forest creates trees in parallel, and these trees have no interaction. At the training time, the algorithm creates a large number of decision trees and outputs the average prediction of the individual trees.
- Logistic Regression (LG): Logistic regression [47] is a statistical method employed in binary classification tasks to model the probability of a particular outcome based on the relationships with independent variables. It calculates the estimated probability of a categorical dependent variable's relationship with one or more independent variables.
- Extra Tree Classifier (ET): Extra Tree (ET), or extremely randomized tree, is an ensemble machine learning method [48]. The Extra Tree Classifier method improves predictive accuracy and controls over-fitting by averaging by fitting several randomized decision trees from the original learning sample.
- Support Vector Machine (SVM): The Support Vector Machine classifier determines how much error in the model is acceptable and selects a line or hyperplane that best fits the data [49]. We optimized the epsilon and cost parameter C using a Bayesian optimization algorithm.
- Light Gradient Boosting Machine (LGBM): Light GBM is a learning algorithm that uses a tree-based approach [50]. The algorithm grows the tree vertically and selects a leaf based on the loss. The gradient boosting framework is used in this project. LGBM is a quick algorithm with a small memory footprint that can handle large datasets.
- Categorical Gradient Boosting Classifier (CAT): CatBoost handles categorical features and outperforms existing publicly available gradient boosting implementations in terms of quality [51]. On ensembles of similar size, the library has a GPU implementation of the learning algorithm and a CPU implementation of the scoring algorithm, making it significantly faster than other gradient-boosting libraries.

3.4. *Feature selection*

The feature selection process can be considered a method of selecting a subset of variables from a large feature set and assessing their

accuracy. It is used for various reasons, including simplifying models to make them easier for researchers to interpret, reducing training times, avoiding the dimensionality curse, and improving data compatibility with a learning model class.

To identify which features are important, we have used SHAP (Shapley Additive exPlanations) importance scores [52]. SHAP is a state-of-the-art method used in machine learning to interpret the output of complex machine learning models. These scores are based on game theory, specifically the concept of Shapley values, which were developed to allocate the payout of a cooperative game fairly to its players based on their individual contributions [52]. SHAP importance scores provide a detailed and fair explanation of how each feature in a dataset influences the prediction of a machine learning model, enhancing transparency and interpretability in complex models. Figs. 3 and 4 show the SHAP importance scores for the DNA and RNA datasets. We found that the most important feature in both datasets is the Accessible Surface Area. A larger Accessible Surface Area likely provides more binding space with DNA and RNA. AA index, Monogram and Bigram, and HMM profile are some of the other features that contribute to the prediction of the proposed method.

When applying a feature selection technique, the fundamental assumption is that the dataset includes features that might be redundant or irrelevantand can be safely eliminated without substantial loss of information. Features that are not relevant or only partially relevant have the potential to affect the performance of a model; hence, feature selection becomes a crucial step in the model creation process. We have used a Recursive Feature Elimination technique (RFE) that allows you to reduce the number of features in the dataset while maintaining the model's predictive power. It removes the features with the lowest importance based on the SHAP importance scores. Recursive Feature Elimination offers several benefits, including the utilization of tree-based or linear models to detect the complex relations between features and the target. RFE can be implemented with SHAP importance scores, one of the most reliable ways to estimate the importance of features. Unlike many other techniques, it works with missing values and categorical variables. It also provides a list of features that should not be eliminated, e.g., in the case of prior knowledge. After using the Recursive Feature Elimination method, a total of 96 features out of 119 features are selected for both the DNA and RNA datasets. Figs. 5 and 6 show the selected features for each (DNA and RNA) dataset.

3.5. Performance evaluation metrics

In our study, the dataset is highly imbalanced, so we need to choose the evaluation metrics carefully. We have selected AUC, Recall, and MCC to evaluate our method. Area Under the "Receiver Characteristic Operator Curve" (AUC) is a widely used metric to find the performance of the machine learning method. AUC is not threshold-dependent, making it a robust metric to evaluate the model performance. We chose the Recall and MCC metrics because they are also used for imbalanced datasets, and existing methods use them for comparison. The following section shows the formula to calculate the recall and MCC.

$$Recall \Big/ Sensitivity = \frac{TP}{TP + FN}$$

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

Where TP is the number of correctly predicted binding residues (true positives)

TN is the number of correctly predicted non-binding residues (true negatives)

FP is the number of incorrectly predicted non-binding residues (false positives)

FN is the number of incorrectly predicted binding residues (false negatives)

4. Results

In this section, we first discuss the performance of Machine learning methods and then optimizing window size and hyperparameters. Finally, we compare the performance of DRBpred with the existing state-of-the-art methods.

4.1. Performance of machine learning methods on the training dataset

As discussed before, we have selected seven machine learning methods to find the best method suitable for this problem. Figs. 7 and 8 show the 10-fold cross-validation results for the DNA and RNA datasets, respectively. The Light Gradient Boosting Machine performs better for each dataset than the other methods in terms of AUC, Recall, and MCC, so we selected this method for the rest of the experiments.

Additionally, to assess the robustness and consistency of our model, we conducted a 5-fold cross-validation on the training dataset in terms of AUCROC. This approach is crucial to understand the model's performance variability between the training and testing phases. In cross-validation, the dataset is divided into five equal parts. In each fold, a different part of the dataset is held out for testing while the remaining four parts are used for training. This process is repeated five times, each

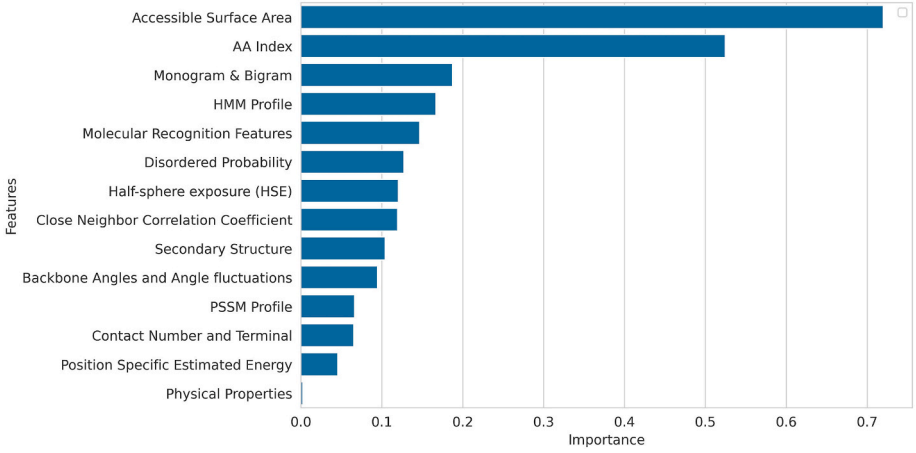

Fig. 3. Importance scores from SHAP (Shapley Additive exPlanations) for DNA-binding proteins. The Accessible Surface Area feature holds the highest feature importance score, followed by AA index. The evolutionary-based features, Monogram and Bigram, calculated from PSSM scores, have the third highest importance scores.

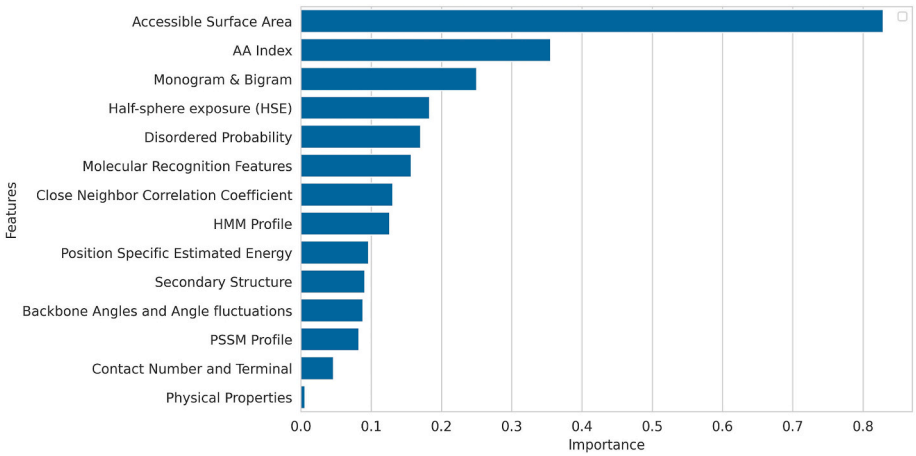

Fig. 4. SHAP (Shapley Additive exPlanations) Importance scores for RNA-binding proteins. Similar to DNA-binding proteins, the Accessible Surface Area feature possesses the highest feature importance score, succeeded by the AA index. Following these, the evolutionary-based features Monogram and Bigram, derived from PSSM scores, rank as the third most significant in terms of importance scores.

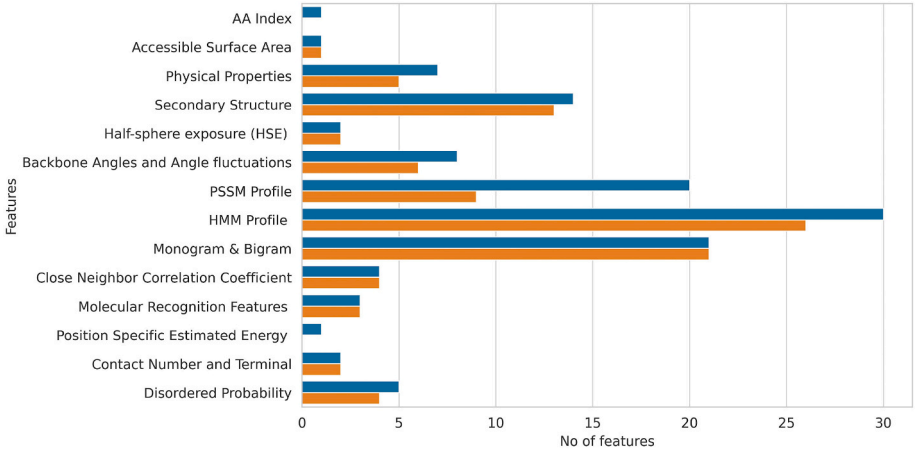

Fig. 5. Illustration of the number of selected features for the DNA dataset. The orange bar represents the selected features, and the blue bar represents the total number of features. The lower number of features are selected from PSSM and HMM profiles as they both represent the evolutionary features.

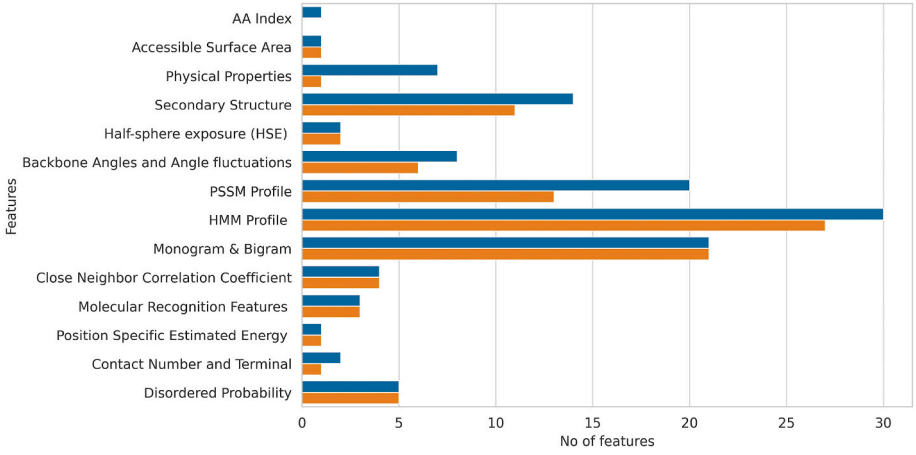

Fig. 6. Illustration of the number of selected features for the RNA dataset. The orange bar represents the selected features, and the blue bar represents the total number of features. The lower number of features are selected from PSSM and physical properties.

time with a different part being used as the test set, ensuring a comprehensive evaluation. Figs. 9 and 10 display the Receiver Operating Characteristic-Area Under Curve (ROC-AUC) for the DNA and RNA training and test sets. The ROC-AUC metric is a reliable indicator of the model's ability to distinguish between classes, with a value closer to 1 indicating higher accuracy. For our model, the training ROC-AUC score

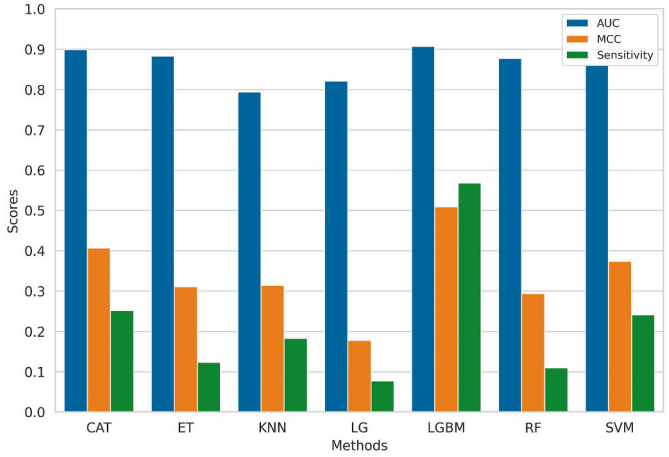

Fig. 7. 10-fold cross-validation results on the DNA Training dataset on different Machine learning methods. The Light Gradient Boosted Machine outperforms all the other methods in terms of AUC, MCC, and Sensitivity metrics. (CAT: Categorical Gradient Boosting Classifier, ET: Extra Tree Classifier, KNN: K-nearest Neighbors Classifier, LG: Logistic Regression, LGBM: Light Gradient Boosted Machine, RF: Random Forest Classifier, SVM: Support Vector Machine).

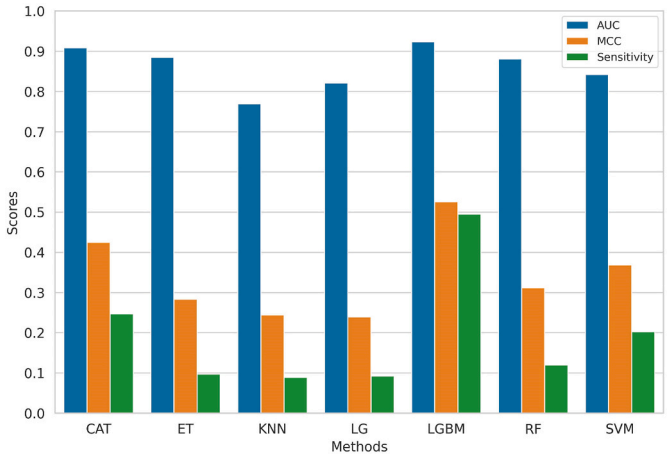

Fig. 8. 10-fold cross-validation results on the RNA Training dataset on different Machine learning methods. The Light Gradient Boosted Machine outperforms all the other methods in terms of AUC, MCC, and Sensitivity metrics. (CAT: Categorical Gradient Boosting Classifier, ET: Extra Tree Classifier, KNN: K-nearest Neighbors Classifier, LG: Logistic Regression, LGBM: Light Gradient Boosted Machine, RF: Random Forest Classifier, SVM: Support Vector Machine).

is notably high, approximately 0.91, which is a strong indication of the model's effectiveness in the training phase. Furthermore, this high score is consistent across all five folds, as indicated by the low variation in performance. This consistency is important as it implies that the model is not overly fitted to a specific part of the training data and can generalize well across the entire dataset. On the other hand, while the performance on the test set shows a decrease compared to the training set, it still yields good results. This decrease is a common observation, as models tend to perform slightly worse on unseen data. However, the fact that the model still shows good results on the test sets suggests that while there is a drop in performance, the model maintains a significant degree of its predictive power when applied to new, unseen data, which is a critical aspect of model reliability and usefulness in practical applications.

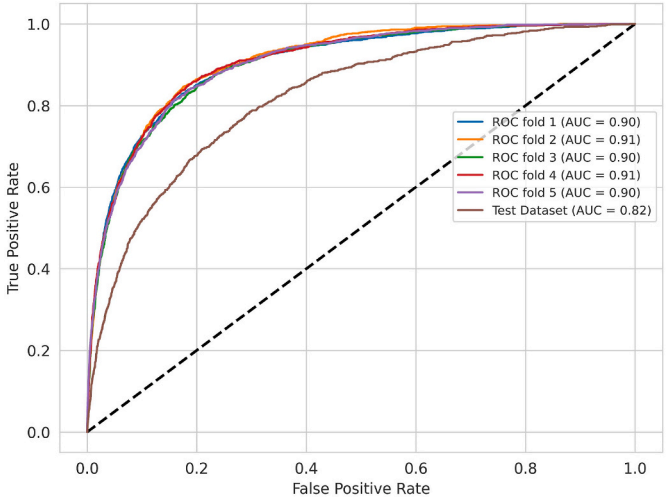

Fig. 9. The ROC-AUC curve for the DNA training and test dataset. The training ROC-AUC score is approximately 0.90 for five folds and shows low variation in performance. For unseen test set the ROC-AUC score is 0.82.

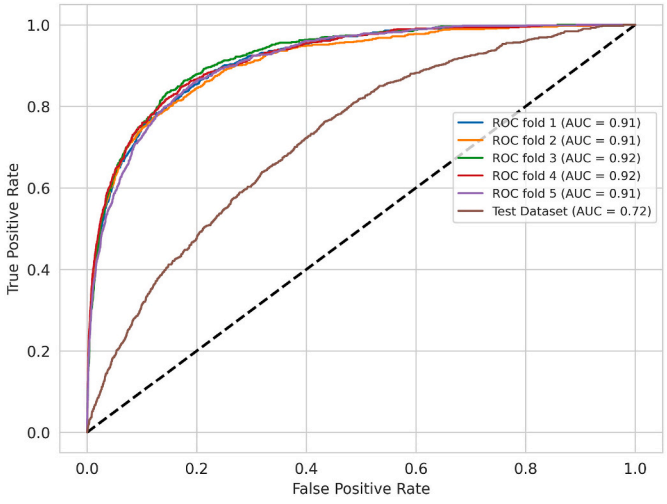

Fig. 10. The performance on the RNA training and test datasets is depicted by the ROC-AUC curve. The training phase achieves a consistent ROC-AUC score of around 0.91 across all five folds, indicating stable performance. On the unseen test set, the ROC-AUC score reaches 0.72.

4.2. Optimizing hyperparameters

Machine learning method performance highly depends on the selected hyperparameter. To improve the results of our method, we optimized the hyperparameter of the Light Gradient Boosting Machine. The parameters (n_estimators, learning_rate, num_leaves, max_depth, min_child_samples, max_bin, subsample, subsample_freq, and colsample_bytree) of LightGBM are optimized using a hyperparameter optimization framework (Optuna) [53]. The framework used a Tree-structured Parzen Estimator algorithm to optimize the hyperparameters. Table 2 shows the selected hyperparameters for both DNA and RNA datasets.

4.3. Selection of best window size

The residues/amino acids are interconnected within proteins. This means each residue's characteristics are influenced by its adjacent residues.That motivates us to represent residues not only with their own features but also the neighboring residue's features. We collected 96

Table 2
Selected best parameters LightGBM for DNA and RNA datasets. The parameters are selected with a Tree-structured Parzen Estimator algorithm.

Parameter Name	DNA	RNA
n_estimators	1000	1000
learning_rate	0.151	0.159
num_leaves	7	10
max_depth	3	4
min_child_samples	95	89
max_bin	102	118
subsample	0.72	0.83
subsample_freq	1	1
colsample_bytree	0.93	0.95

features to represent each residue/amino acid. Fig. 11 shows that the residues glycine(G) can be represented by concatenating the features from two of its neighbor's residues, lysine(K) and leucine(L). For window size 3, the length of the glycine(G) residue feature vector is 96 × 3 = 288 features. As the feature dimensions increase with the window size increase, we investigate the optimal window size for each model.

We investigated window sizes from 1 to 19 to find the optimal size for both DNA and RNA models. For the DNA dataset, Figs. 12 and 13 show that the optimal window size is 11 for the DNA and RNA models.

4.4. Performance on the test dataset

The model's performance is assessed by conducting an evaluation on the test dataset. The performance of DRBpred is presented in Table 3, where various metrics such as Sensitivity, Specificity, Balanced Accuracy (BACC), Matthews Correlation Coefficient (MCC), Accuracy (ACC), False Positive Rate (FPR), False Negative Rate (FNR), Precision, F1-score, and Receiver Operating Characteristic Area Under the Curve (ROCAUC) are reported. The results indicate that DRBPred exhibits strong performance, particularly excelling in terms of BACC, ACC, and ROCAUC.

4.5. Comparison with existing methods

We performed a comparative evaluation of our method against recent state-of-the-art methods, namely DRNApred, Pprint, RNABindR, and BindN+. The results of these state-of-the-art methods were gathered from the DRNApred paper. The performance of the RNA model is detailed in Table 4 and Fig. 14. Our proposed method has an improvement of 112.50 %, 16.67 %, and 7.46 % in terms of Sensitivity, MCC, and AUC compared with the best method DRNApred.

Similarly, we tested our method for DNA-binding prediction. Table 5 and Fig. 15 show the performance of the DNA model. Our proposed method has improved by 112.00 %, 33.33 %, and 6.49 % in Sensitivity, MCC, and AUC compared with the best method, DRNApred.

We plotted the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) analysis in Figs. 16 and 17. The ROC-AUC curve is a graphical representation of the model's ability to discriminate between positive and negative samples, where a larger area under the curve indicates better performance. These curves were constructed using data obtained from the findings presented in the paper [15], as some existing methods were not publicly available. Figs. 16 and 17 provide clear evidence that the DRBpred method surpasses the performance of currently established state-of-the-art techniques.

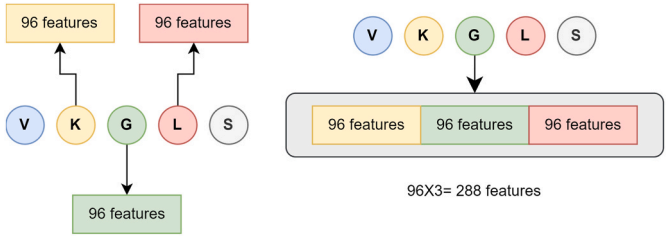

Fig. 11. Illustration of sliding window technique to incorporate neighbor residues information. After feature selection, each residue is represented by 96 features. For sliding window size 3, the residues glycine(G) can be represented by concatenating the features from two of its neighbor's residues, lysine(K) and leucine(L), and the feature vector length is 96x3 = 288 features.

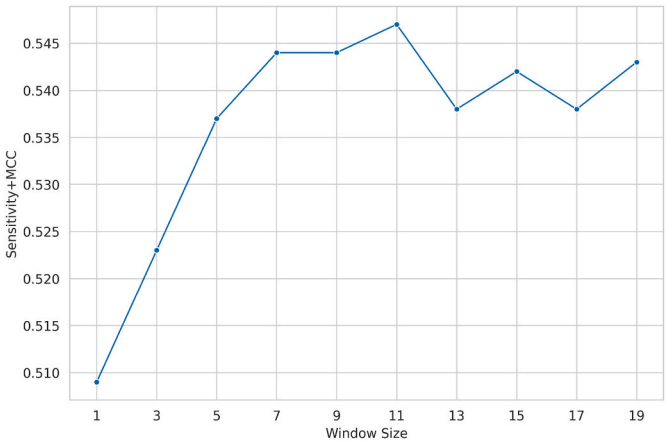

Fig. 12. Selection of sliding window size for DNA dataset. To maximize the objective function (Sensitivy + MCC), the model performance for window sizes 1–19 has been evaluated. The model performs best for window size 11.

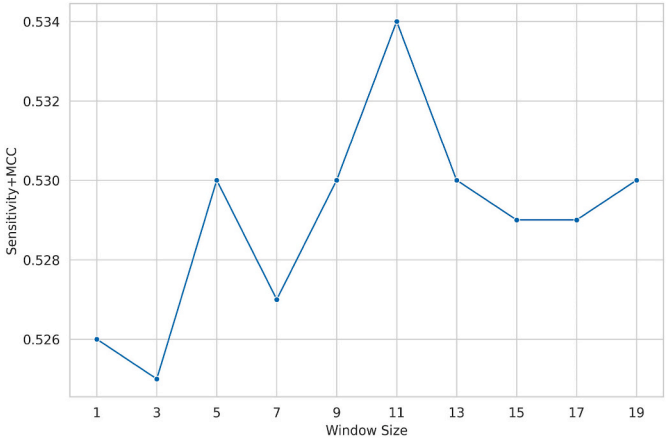

Fig. 13. Selection of sliding window size for RNA dataset. To maximize the objective function (Sensitivy + MCC), the model performance for window sizes 1–19 has been evaluated. The model performs best for window size 11.

5. Case study

We conducted LIME analysis [54] on the independent test samples. LIME, an acronym for Local Interpretable Model-Agnostic Explanations, is used to approximate local, interpretable models that can explain individual predictions for black-box machine learning models [54]. For machine learning models, it is crucial for models to be explainable to gain the trust of users. LIME allows users to understand what happens within these black-box machine-learning models and aids in the identification of possible concerns, including issues related to information leakage, model bias, robustness, and causality [52,54]. LIME introduces perturbations to the original data points, inputs them into the black-box model, and observes the resulting outputs [54]. The method then assigns weights to these new data points based on their proximity to the original point. Subsequently, it creates a surrogate model on the dataset, incorporating these weighted variations [54]. This surrogate model is then

Table 3
Classification scores for DNA and RNA model.

Datasets	Sensitivity	Specificity	BACC	MCC	ACC	FPR	FNR	Precision	F1-score	ROCAUC
DNA	52.58	89.53	71.06	0.28	87.64	0.11	0.47	21.33	0.30	82.00
RNA	34.03	88.82	61.43	0.14	86.47	0.11	0.66	11.98	0.18	72.36

Table 4
Performance comparison of DRBpred with existing methods in the RNA Test dataset. DRBpred method shows promising results compared to the existing methods.

Methods	Sensitivity	MCC	AUC
DRNApred	0.16	0.12	0.67
Pprint	0.15	0.11	0.66
RNABindR	0.14	0.10	**0.73**
BindN+	0.12	0.08	0.67
DRBpred	**0.34**	**0.14**	0.72
(Imp%)	112.50 %	16.67 %	7.46 %

The best score values are **bold-faced.** (Imp%) shows improvement compared to the best method (DRNApred).

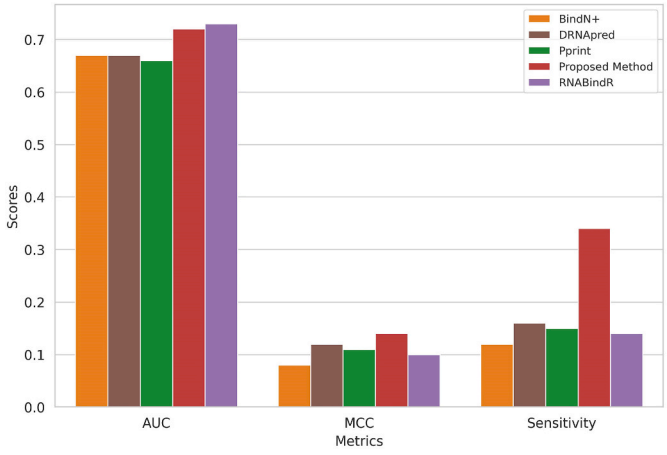

Fig. 14. Performance comparison of DRBpred with existing methods in the RNA Test dataset. The red bar shows the performance of the DRBpred method in terms of AUC, MCC, and sensitivity metrics.

Table 5
Performance comparison of DRBpred with existing methods in the DNA Test dataset. DRBpred performs better compared to the existing methods.

Methods	Sensitivity	MCC	AUC
DRNApred	0.25	0.21	0.77
BindN+	0.22	0.18	0.79
DP-Bind(svm)	0.24	0.20	0.75
DP-Bind(klr)	0.24	0.20	0.76
DP-Bind(plr)	0.22	0.18	0.74
DBS-PSSM	0.21	0.17	0.77
DRBpred	**0.53**	**0.28**	**0.82**
(Imp%)	112.00 %	33.33 %	6.49 %

The best score values are **bold-faced.** (Imp%) shows improvement compared to the best method (DRNApred).

used to explain each original data point individually.

We randomly selected two amino acids for DNA-binding and RNA-binding predictive models from the test dataset for LIME analysis. Fig. 18 provides insights into the top five features influencing the prediction of Valine (V) as a DNA-binding residue for protein 3POV0. The predicted probability for the DNA-binding class is 0.91, whereas the non DNA-binding class has a probability of 0.09. The model correctly predicts the label for this test sample. Notably, the feature importance

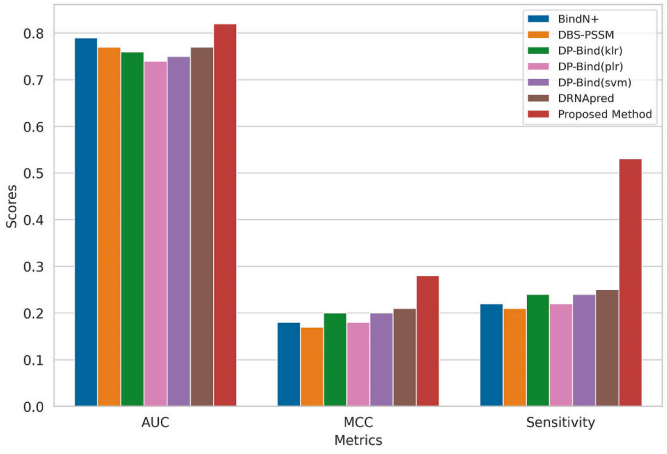

Fig. 15. Performance comparison of DRBpred with existing methods in the DNA Test dataset. The red bar shows the performance of the DRBpred method in terms of AUC, MCC, and sensitivity metrics.

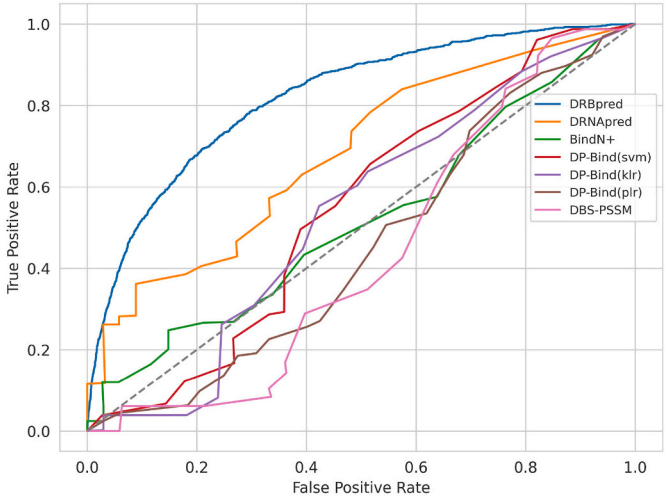

Fig. 16. The ROC-AUC curve for the DNA test dataset. DRBpred achieves an AUC of 0.72, is represented by the blue color. Among the evaluated methods, the second-best performance is demonstrated by DRNApred, with an AUC score of 0.77. DRBpred surpasses the performance of the currently established state-of-the-art methods, indicating its superior accuracy in classifying DNA-binding proteins.

scores for this particular sample reveal that the HMM profile (L) has a 5 % importance score, followed by Bigram (R) with 4 %, Accessible Surface Area with 3 %, and HMM profile (E) also has a 3 % importance score toward the DNA-binding class. On the other hand, the Secondary Structure (P(8-T)) has a 3 % importance score toward the non DNA-binding class.

Fig. 18(b) and (d) visualize the range of local interpretability predictions for the Valine (V) sample. They indicate that the HMM profile (L) for this specific instance exceeds 5658, the Accessible Surface Area is greater than 0.5, Bigram (R) falls within the range of 0.52–1.54, and the HMM profile (L) falls within the range of 4805–6758, contributing to the

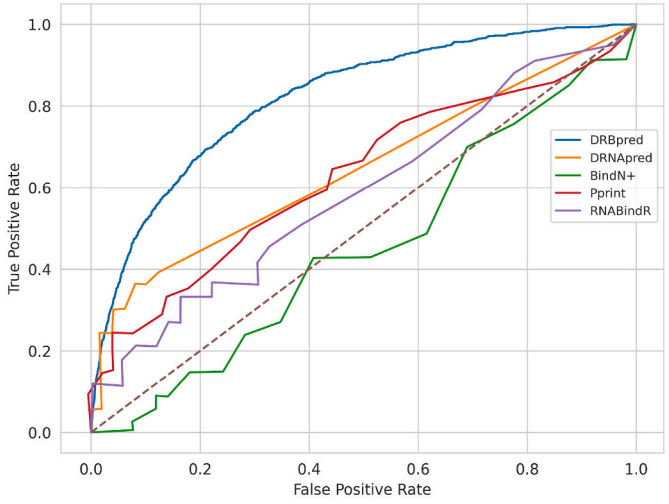

Fig. 17. The ROC-AUC curve for the RNA test dataset. DRBpred achieves an AUC of 0.82, which is shown by the blue color. DRBpred outperforms the performance of the state-of-the-art methods, demonstrating effectiveness in the classification of RNA-binding proteins.

DNA-binding class. It is extremely difficult to relate these features to the prediction of DNA-binding model. However, one feature importance score aligns with our hypothesis. Residues tend to exhibit DNA-binding tendencies when they possess a higher accessible surface area. In this instance, the Accessible Surface Area is greater than 0.5, contributing to this sample being identified as a DNA-binding residue. Furthermore, during our analysis of feature importance scores, we observed that the Accessible Surface Area is the most crucial feature in our trained model for both DNA and RNA datasets.

We further investigated the contribution of features for RNA-binding protein prediction. Fig. 19 provides insights into the top five features influencing the Serine (S) prediction at position 261 as an RNA-binding residue for protein 3ZH22. The predicted label for this instance is non RNA-binding with a probability of 0.99. The true label is non RNA-binding. Notably, the feature importance scores for this particular sample reveal that HMM profile (E) has a less than 1 % importance score, followed by Bigram (V), Bigram (F), and Bigram (P), all with less than 1 %. In contrast, the Accessible Surface Area is less than 0.11 for this particular sample and contributed to the non RNA-binding prediction. These feature importance scores align with our hypothesis that residues exhibit non RNA-binding tendencies when they possess a lower accessible surface area.

6. Conclusions

In this study, we developed a new method, DRBpred, to predict DNA-binding and RNA-binding residues from protein sequences. This method involves gathering relevant features and employing a recursive feature elimination (RFE) technique along with SHAP values to select a subset of features. Additionally, a sliding window technique was utilized to extract additional information from neighboring residues, and an optimized LightGBM classifier was trained to predict the binding residues. DRBpred demonstrated significant improvements across various evaluation metrics compared to the state-of-the-art method. Specifically, for the DNA-binding test dataset, DRBpred exhibited enhancements of 112.00 % in sensitivity, 33.33 % in Matthews's correlation coefficient (MCC), and 6.49 % in the area under the curve (AUC). Similarly, improvements of 112.50 % in sensitivity, 16.67 % in MCC, and 7.46 % in AUC were observed for the RNA-binding test dataset. These results clearly indicate that the optimized LightGBM method surpasses the performance of the existing state-of-the-art approach. The limitation of the proposed approach is that feature extraction is computationally expensive. DRBpred depends on other existing methods for feature extraction, some of which are time intensive. To mitigate this, we plan to employ parallel processing strategies involving multiple CPUs in future developments. Additionally, we aim to incorporate three-dimensional predicted structural information in the future. Moreover, large language models (LLMs) could be employed to extract important features,

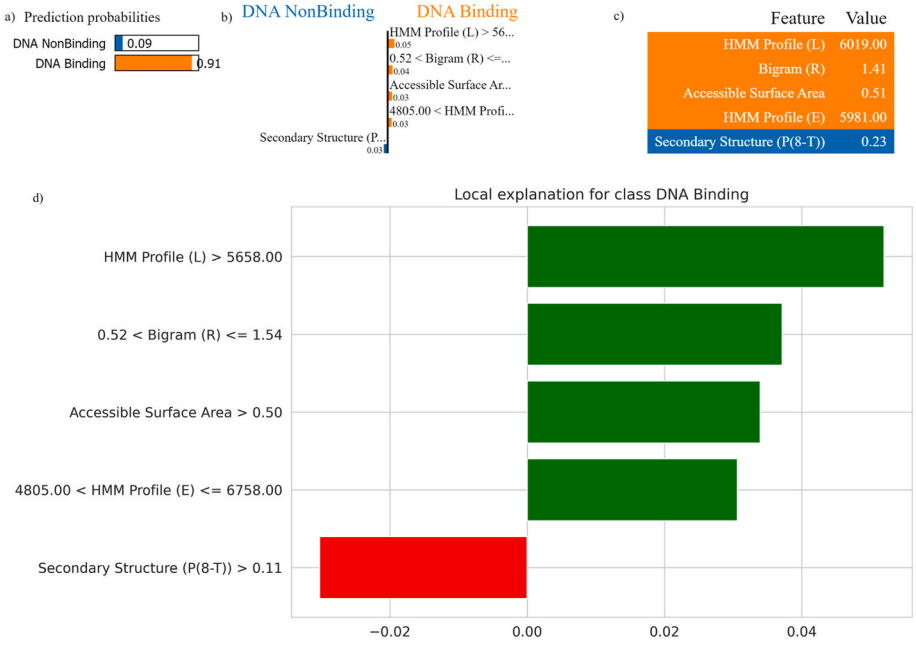

Fig. 18. The figure illustrates the features influencing the prediction of the amino acid valine (V) as a DNA-binding residue in protein ID 3POV0. (a) Displays the prediction probabilities of the model for DNA-binding (orange) and non-DNA-binding (blue) classes. (b) Highlights the top five significant features. (c) the top five features and their corresponding values. (d) same as figure (b) and shows each feature's contribution to the prediction of the selected amino acid, with their relative importance denoted by floating-point numbers on the x-axis. Features contributing to DNA-binding are shown in green, and those contributing to non DNA-binding in red.

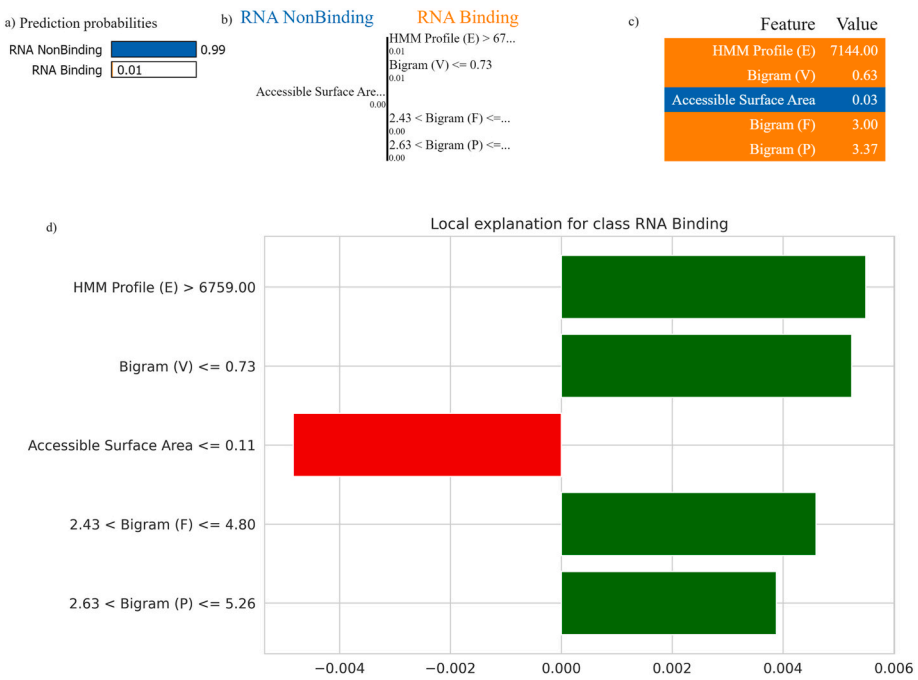

Fig. 19. The key features influencing the identification of Serine (S) in protein 3ZH22 as an RNA-binding residue. (a) Showcases the model's predictive probabilities: RNA-binding in orange and non-RNA-binding in blue. (b) Displays the five most crucial features. (c) Shows the top five features with their values. (d) Illustrates each feature's role in predicting the specific amino acid, with their significance quantified by floating-point values on the x-axis. Features contributing to RNA-binding are shown in green and those contributing to non-RNA binding in red.

potentially improving prediction accuracy. We believe that the DRBpred method will assist researchers in predicting DNA- and RNA-binding residues, enabling a better understanding of the roles played by DNA- and RNA-binding proteins in the life cycle of organisms.

Data Availability

The DRBpred webserver is available at https://bmll.cs.uno.edu. The data, including the code related to the development of DRBpred can be found here https://github.com/wasicse/DRBpred.

CRediT authorship contribution statement

Md Wasi Ul Kabir: Data curation, Formal analysis, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Duaa Mohammad Alawad:** Data curation, Software, Writing – original draft, Writing – review & editing. **Pujan Pokhrel:** Conceptualization, Data curation, Methodology, Writing – original draft, Writing – review & editing. **Md Tamjidul Hoque:** Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

There is no conflict of interest with any of the authors or the suggested reviewers (if any).

Acknowledgments

The authors would like to thank Christopher David Moore for thorough review of the manuscript. The research reported in the paper was partially supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P2O GM103424-21.

References

[1] J. Zhou, et al., EL_PSSM-RT: DNA-Binding Residue Prediction by Integrating Ensemble Learning with PSSM Relation Transformation, vol. 18, 2017.
[2] X. Dai, S. Zhang, K. Zaleta-Rivera, RNA: interactions drive functionalities, Mol. Biol. Rep. 47 (2) (2020) 1413–1434.
[3] D.D. Licatalosi, Roles of RNA-binding proteins and post-transcriptional regulation in Driving male germ cell development in the mouse, Adv. Exp. Med. Biol. 907 (2016) 123–151.
[4] F. Cozzolino, et al., Protein-DNA/RNA interactions: an overview of investigation methods in the -omics era, J. Proteome Res. 20 (6) (2021) 3018–3030.
[5] C.M. Kobras, A.K. Fenton, S.K. Sheppard, Next-generation microbiology: from comparative genomics to gene function, Genome Biol. 22 (1) (2021) 123.
[6] K. Li, et al., Prediction of hot spots in protein-DNA binding interfaces based on supervised isometric feature mapping and extreme gradient boosting, BMC Bioinf. 21 (Suppl 13) (2020) 381.
[7] M. Mesri, Advances in proteomic technologies and its contribution to the field of cancer, Advances in medicine 2014 (2014), 238045-238045.
[8] B. Faezov, R.L. Dunbrack Jr., PDBrenum: a webserver and program providing Protein Data Bank files renumbered according to their UniProt sequences, PLoS One 16 (7) (2021) e0253411.
[9] L. Deng, et al., PDRLGB: precise DNA-binding residue prediction using a light gradient boosting machine, BMC Bioinf. 19 (Suppl 19) (2018) 522.
[10] Q. Yuan, et al., AlphaFold2-aware protein–DNA binding site prediction using graph transformer, Briefings Bioinf. 23 (2) (2022) bbab564.
[11] Y.D. Cai, S.L. Lin, Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence, Biochim. Biophys. Acta 1648 (1–2) (2003) 127–133.
[12] C. Zou, J. Gong, H. Li, An improved sequence based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis, BMC Bioinf. 14 (1) (2013) 90.
[13] W. Lou, et al., Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes, PLoS One 9 (1) (2014) e86703.
[14] Y. Zhang, et al., newDNA-Prot: prediction of DNA-binding proteins by employing support vector machine and a comprehensive sequence representation, Comput. Biol. Chem. 52 (2014) 51–59.
[15] J. Yan, L. Kurgan, DRNApred, fast sequence-based method that accurately predicts and discriminates DNA-and RNA-binding residues, Nucleic acids research 45 (10) (2017) e84-e84.
[16] S. Hwang, Z. Gou, I.B. Kuznetsov, Dp-Bind, A web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins, Bioinformatics 23 (5) (2007) 634–636.
[17] L. Wang, et al., BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features, BMC Syst. Biol. 4 (2010) 1–9.

[18] L. Wang, S.J. Brown, BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences, Nucleic Acids Res. 34 (2006) W243–W248 (Web Server issue).
[19] L. Wang, et al., BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features, BMC Syst. Biol. 4 (Suppl 1) (2010) S3. Suppl 1.
[20] S. Ahmad, A. Sarai, PSSM-based prediction of DNA binding sites in proteins, BMC Bioinf. 6 (1) (2005) 33.
[21] J. Zhou, et al., EL_PSSM-RT: DNA-binding residue prediction by integrating ensemble learning with PSSM Relation Transformation, BMC Bioinf. 18 (1) (2017) 379.
[22] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (12) (1983) 2577–2637.
[23] Q. Zhang, et al., StackPDB: predicting DNA-binding proteins based on XGB-RFE feature optimization and stacked ensemble classifier, Appl. Soft Comput. 99 (2021) 106921.
[24] S.G. Hendrix, et al., DeepDISE: DNA binding site prediction using a deep learning method, Int. J. Mol. Sci. 22 (11) (2021) 5510.
[25] J. Zhou, et al., EL_LSTM: prediction of DNA-binding residue from protein sequence by combining long short-term memory and ensemble learning, IEEE/ACM Trans Comput Biol Bioinform 17 (1) (2020) 124–135.
[26] D.T. Jones, J.J. Ward, Prediction of disordered regions in proteins from position specific score matrices, Proteins 53 (Suppl 6) (2003) 573–578, 6.
[27] X. Ma, et al., A SVM-Based Approach for Predicting DNA-Binding Residues in Proteins from Amino Acid Sequences, IEEE Xplore, 2009.
[28] B. Liu, et al., Using amino acid physicochemical distance transformation for fast protein remote homology detection, PLoS One 7 (9) (2012) e46633.
[29] H.L. Huang, et al., Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties, BMC Bioinf. (2011) S47.
[30] B. Liu, et al., Using amino acid physicochemical distance transformation for fast protein remote homology detection, PLoS One 7 (2012).
[31] L. Zhu, et al., Improving the accuracy of predicting disulfide connectivity by feature selection, J. Comput. Chem. 31 (7) (2010) 1478–1485.
[32] S. Niu, et al., Prediction of tyrosine sulfation with mRMR feature selection and analysis, J. Proteome Res. 9 (12) (2010) 6490–6497.
[33] S. Iqbal, A. Mishra, M.T. Hoque, Improved prediction of accessible surface area results in efficient energy function application, J. Theor. Biol. 380 (2015) 380–391.
[34] S. Iqbal, M.T. Hoque, PBRpredict-Suite: a suite of models to predict peptide-recognition domain residues from protein sequence, Bioinformatics 34 (19) (2018) 3289–3299.
[35] S. Iqbal, M.T. Hoque, Estimation of position specific energy as a feature of protein residues from sequence alone for structural classification, PLoS One 11 (9) (2016) e0161452.
[36] S.F. Altschul, et al., Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[37] A. Sharma, et al., Evaluation of sequence features from intrinsically disordered regions for the estimation of protein function, PLoS One 9 (2) (2014) e89890.
[38] S.R. Eddy, Profile hidden Markov models, Bioinformatics 14 (9) (1998) 755–763.
[39] A. Mishra, M.W.U. Kabir, M.T. Hoque, diSBPred: a machine learning based approach for disulfide bond prediction, Comput. Biol. Chem. 91 (2021) 107436.
[40] M.W. Kabir, et al., TAFPred: torsion angle fluctuations prediction from protein sequences, Biology 12 (2023), https://doi.org/10.3390/biology12071020.
[41] J. Hanson, et al., SPOT-Disorder2: improved protein intrinsic disorder prediction by ensembled deep learning, Dev. Reprod. Biol. 17 (6) (2019) 645–656.
[42] P.E. Wright, H.J. Dyson, Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, J. Mol. Biol. 293 (2) (1999) 321–331.
[43] J. Liu, H. Tan, B. Rost, Loopy proteins appear conserved in evolution, J. Mol. Biol. 322 (1) (2002) 53–64.
[44] P. Tompa, Intrinsically unstructured proteins, Trends Biochem. Sci. 27 (10) (2002) 527–533.
[45] S. Gattani, A. Mishra, M.T. Hoque, StackCBPred: a stacking based prediction of protein-carbohydrate binding sites from sequence, Carbohydr. Res. 486 (2019) 107857.
[46] E. Vigneau, et al., Random forests: a machine learning methodology to highlight the volatile organic compounds involved in olfactory perception, Food Quality 68 (2018) 135–145.
[47] P. Ranganathan, C.S. Pramesh, R. Aggarwal, Common pitfalls in statistical analysis: logistic regression, Perspect Clin Res 8 (3) (2017) 148–151.
[48] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (1) (2006) 3–42.
[49] D.M. Alawad, A. Mishra, M.T. Hoque, AIBH: accurate identification of brain hemorrhage using genetic algorithm based feature selection and stacking, Machine Learning Knowledge Extraction 2 (2) (2020) 56–77.
[50] G. Ke, et al., Lightgbm: a highly efficient gradient boosting decision tree, Adv. Neural Inf. Process. Syst. 30 (2017).
[51] A.V. Dorogush, V. Ershov, A. Gulin, CatBoost: Gradient Boosting with Categorical Features Support, 2018 arXiv preprint.
[52] Lundberg, S. and S.-I. Lee, A unified approach to interpreting model predictions, in Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4768–4777.
[53] T. Akiba, et al., Optuna: A Next-Generation Hyperparameter Optimization Framework, 2019 arXiv [cs.LG].
[54] M.T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you?, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Association for Computing Machinery, San Francisco, California, USA, 2016, pp. 1135–1144.