ACS OMEGA

Open Access

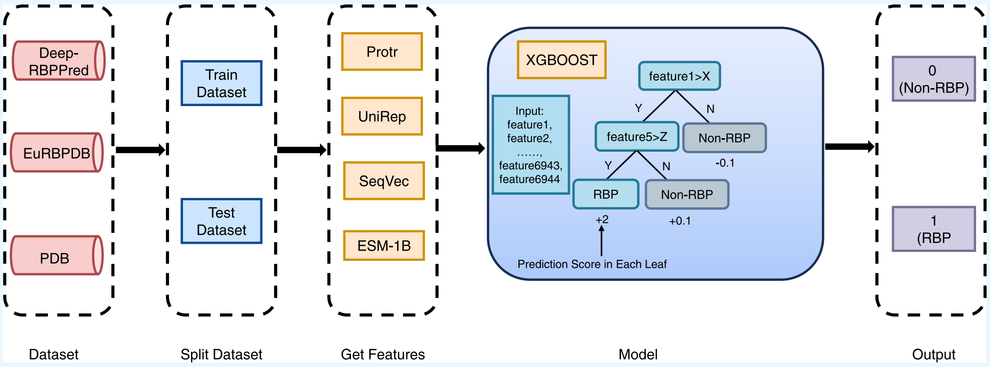
Seq-RBPPred: Predicting RNA-Binding Proteins from Sequence

Yuyao Yan, Wenran Li, Sijia Wang,* and Tao Huang*

Cite This: *ACS Omega* 2024, 9, 12734–12742

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: RNA-binding proteins (RBPs) can interact with RNAs to regulate RNA translation, modification, splicing, and other important biological processes. The accurate identification of RBPs is of paramount importance for gaining insights into the intricate mechanisms underlying organismal life activities. Traditional experimental methods to predict RBPs require a lot of time and money, so it is important to develop computational methods to predict RBPs. However, the existing approaches for RBP prediction still require further improvement due to unidentified RBPs in many species. In this study, we present Seq-RBPPred (predicting RBPs from sequence), a novel method that utilizes a comprehensive feature representation encompassing both biophysical properties and hidden-state features derived from protein sequences. In the results, comprehensive performance evaluations of Seq-RBPPred its superiority compare with state-of-the-art methods, yielding impressive performance including 0.922 for overall accuracy, 0.926 for sensitivity, 0.903 for specificity, and Matthew's correlation coefficient (MCC) of 0.757 as ascertained from the evaluation of the testing set. The data and code of Seq-RBPPred are available at https://github.com/yaoyao-11/Seq-RBPPred.

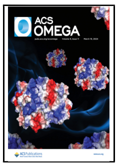
1. INTRODUCTION

RNA-binding proteins (RBPs) are a class of proteins that exhibit the capacity to interact with mRNA as well as noncoding RNA molecules. These RBPs hold significant prominence in the regulation of diverse metabolic processes within the biological system, exerting their influence through intricate RNA-related mechanisms, including translation,[1] modification,[2] splicing,[3] and transport.[4] The identification of RBPs remains a significant challenge as numerous species harbor a substantial number of RBPs that are yet to be characterized. Achieving accurate identification of RBPs assumes utmost importance in comprehending the functional intricacies underlying organismal processes.

With the advancement of biotechnology, the availability of genomic data has grown exponentially, leading to an increasing trend of utilizing machine learning techniques to explore the intricacies of the human genome. In recent years, there has been a proliferation of applications employing machine learning in genomics. These applications encompass a wide range of areas, including the prediction of binding sites for DNA and RNA binding proteins,[5−9] the identification of *cis*-regulatory elements such as promoters[10,11] and enhancers,[12,13] the prediction of DNA methylation patterns,[14−16] and histone modifications,[17,18] the determination of cellular localization,[19−21] the analysis of alternative splicing events,[22,23] and the assessment of the impact of genetic variants on gene expression,[24] among others.

Although machine learning has been widely applied in genomics, its utilization in predicting RBPs is still limited. Currently, two primary methodologies are employed for RBP prediction: structure-based approaches and sequence-based approaches. The structure-based approach leverages three-dimensional structural information on proteins to predict RBPs. For instance, NAbind[25] utilizes electrostatic information derived from protein structures in combination with the support vector machine (SVM)[26] as a machine-learning algorithm for RBP prediction. Another example is SPOT-

Received: October 26, 2023
Revised: December 18, 2023
Accepted: December 28, 2023
Published: March 4, 2024



ACS Publications

https://doi.org/10.1021/acsomega.3c08381

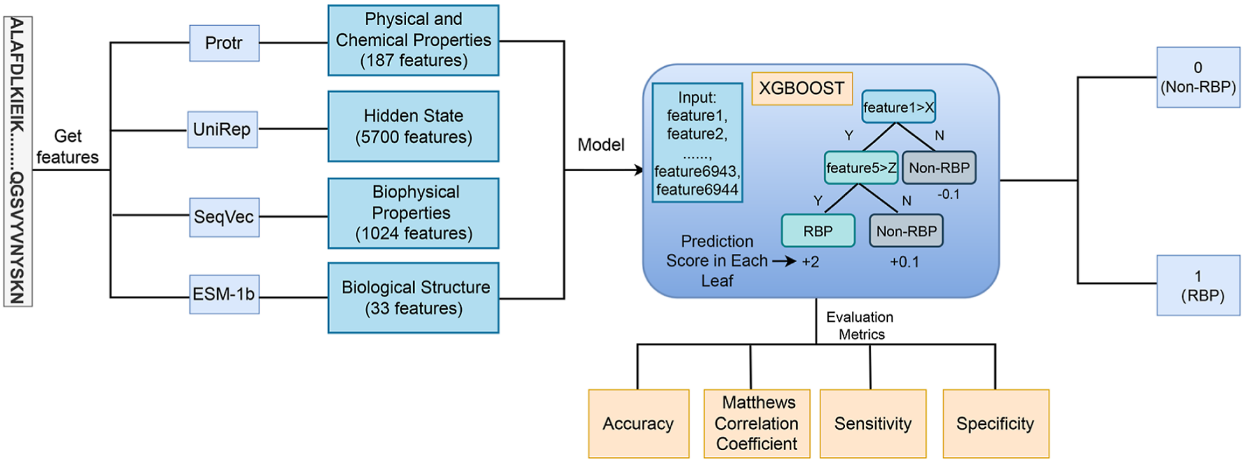

ALAFDLKIEIK......QGSVYVNYSKN
Get features
Protr
UniRep
SeqVec
ESM-1b
Physical and Chemical Properties (187 features)
Hidden State (5700 features)
Biophysical Properties (1024 features)
Biological Structure (33 features)
Model
XGBOOST
Input: feature1, feature2, ..., feature6943, feature6944
feature1>X
Y
N
feature5>Z
Non-RBP
-0.1
Y
N
RBP
Non-RBP
Prediction Score in Each Leaf
+2
+0.1
0 (Non-RBP)
1 (RBP)
Evaluation Metrics
Accuracy
Matthews Correlation Coefficient
Sensitivity
Specificity

Figure 1. Seq-RBPPred framework employs four distinct tools: Protr, UniRep, SeqVec, and ESM-1b, to extract a comprehensive set of features from protein sequences. Specifically, it extracts 187, 5700, 1024, and 33 types of features, culminating in a total of 6944 features. These features are then fed into the XGBoost algorithm, which is used to analyze and predict the proteins that act as RBPs. Within the processing mechanism of XGBoost, each feature of the input samples is evaluated, and scores are allocated based on their discriminative capacity. The final determination of whether a protein is an RBP is made by selecting the outcome with the highest score. More precisely, the output of Seq-RBPPred is a binary classification, where "1" signifies the presence of an RBP in the input protein and "0" denotes its absence, thereby categorizing it as non-RBP. Furthermore, we adopt ACC, MCC, SN, and SP as the evaluative metrics for our model.

stru,[27] which employs a template-based approach to predict RBPs by considering the folding recognition and binding affinity of protein structures. The structure-based approach provides valuable insights into the three-dimensional characteristics of proteins, offering direct information about protein—RNA interactions. However, it is reliant on the availability and accuracy (ACC) of protein structures. Currently, there is a relative scarcity of protein structure data, necessitating further advancements in certain methods employed for predicting protein structures.

The sequence-based approach primarily relies on amino acid sequence information for RBP prediction. This approach is more convenient and allows for prediction on a large scale of protein sequences. Currently, the most accurate method for predicting RBPs is Deep-RBPPred,[28] an improved version based on RBPPred.[29] RBPPred utilizes SVM for RBP prediction based on protein sequences. However, it takes a long time to run and requires significant computational resources. Deep-RBPPred uses a convolutional neural network (CNN) for RBP prediction, demonstrating enhanced prediction capabilities compared to RBPPred. Nevertheless, in certain data sets, the prediction rate of Deep-RBPPred is still not optimal.

In this paper, we developed Seq-RBPPred with the primary objective of enhancing the precision of predicting RBPs. To accomplish this, our approach involves the utilization of Protr,[30] UniRep,[31] SeqVec,[32] and ESM-1b[33] to extract a comprehensive set of 6944 features from protein sequences. Subsequently, the eXtreme gradient boosting (XGBoost)[34] was employed for training and predicting RBPs (Figure 1).

In this study, we utilized two distinct training sets to enhance the predictive capacity of our model. The first training set, obtained from Deep-RBPPred, consisted of 2780 RBPs and 7093 non-RBPs. To further enrich our training set, we extracted additional protein sequences from the same species present in EuRBPDB[35] and PDB.[36] By integrating these sequences with training set 1, we created training set 2, encompassing 4801 RBPs and 6243 non-RBPs. We partitioned training set 1 into ten subsets, employing eight subsets for training and two subsets for validation purposes. The Seq-RBPPred model was applied to these subsets, and its performance was compared against the Deep-RBPPred[28] model using the same testing set. Our findings revealed that Seq-RBPPred exhibited superior performance, thereby validating its efficacy over Deep-RBPPred. Furthermore, we employed Seq-RBPPred to train and predict using the training set 2. To facilitate comparative analysis, we applied the methodology employed in Deep-RBPPred and the aforementioned machine learning approach to predict the same testing set. Remarkably, the performance of Seq-RBPPred consistently surpassed that of Deep-RBPPred, achieving an ACC of 0.922, a Matthew's correlation coefficient (MCC) of 0.757, a sensitivity (SN) of 0.923, and a specificity (SP) of 0.903. To assess the discriminative ability of the model across different samples, we employed the receiver operating characteristic curve and calculated the corresponding area under the curve (AUROC). The AUROC provides a comprehensive evaluation of the model's performance, depicting the relationship between the true positive rate and the false positive rate. Our results demonstrated that Seq-RBPPred achieved an AUROC value of 0.971, further affirming its high predictive ACC.

2. METHODS

2.1. Data Preprocessing.

We collected data on RBPs from the EuRBPDB,[35] and non-RBPs from the PDB,[36] focusing on the same species present in both databases (Figure S1a). EuRBPDB encompasses 162 species, while PDB encompasses 2072 species. Given the dissimilarity between the species in EuRBPDB and the PDB, we carefully curated a set of 30 species, encompassing both RBPs and non-RBPs, to maintain a balanced representation of positive and negative samples for further analysis.

EuRBPDB is a comprehensive repository of eukaryotic RBPs, serving as a comprehensive database encompassing a total of 311,571 RBPs derived from various eukaryotic organisms, including humans, mice, Drosophila, worms, and other representatives from 162 eukaryotic species. It furnishes a precise and exhaustive compilation of RBPs for each eukaryotic organism under consideration. Essential files, namely, RBPlist and totalFa, are acquired by downloading data from EuRBPDB. The RBPlist comprises a collection of individual RBP sequences extracted from totalFa. It is noteworthy that multiple corresponding sequences are often encountered, necessitating their comprehensive documentation while giving precedence to the longest sequence.

To identify non-RBPs, PISCES[37] in the PDB database is utilized for screening purposes. Our processing methodology aligns with that employed by RBPPred[29] and Deep-RBPPred.[28] Non-RBPs conforming to the subsequent processing criteria are included: a sequence identity threshold of 0.25, exclusion of sequences shorter than 50 or longer than 10,000 residues, and an X-ray resolution superior to 3.0 Å. Following the aforementioned procedures, a total of 11,606 chains were obtained.

Based on an analysis of 11,606 data strands, we have selected non-RBPs that belong to the same species as EuRBPDB. Protein chains obtained from PDB are excluded from our selection if their titles contain any of the following terms: “ribosomal”, “RNA”, “nucleoprotein”, “unknown function”, “uncharacterized”, or “hypothetical”. Consequently, we have identified a final set of 2777 protein sequences, which are classified as non-RBPs.

To eliminate redundancy between RBPs from EuRBPDB and non-RBPs from PDB, we merged the two data sets and employed the psi-cd-hit program within the CD-HIT[38] package. This program allowed us to remove redundant sequences with a sequence identity of 25% or higher. As a result, our nonredundant data set consists of 6618 RBPs and 1565 non-RBPs. To ensure consistency in the prediction results and minimize the influence of length variations, we adjusted the length of RBPs to match that of non-RBPs, restricting both to a range of 50 to 10,000 amino acids. Additionally, any proteins labeled as “fragment” were excluded from the RBPs data set.

2.2. Training Set and Testing Set. For a fair comparison with the previous methods, we trained on two data sets and tested on the same testing set.

Training set 1, obtained from Deep-RBPPred,[28] comprised 2780 RBPs and 7093 non-RBPs. To enhance the diversity of our training set, we partitioned the processed data from Section 2.1 into three distinct subsets. Two-thirds of these data was allocated for training purposes and combined with training set 1. To eliminate redundancy, CD-HIT[38] was employed, ensuring a sequence identity threshold of ≤30% (Figure S1b). As Deep-RBPPred did not explicitly provide a testing set, we employed the remaining third of the data as the testing set (Figure S1a). To maintain the integrity of training set 1, we excluded the same data present in the testing set from training set 1, resulting in our final testing set.

We employed two training sets and one testing set for feature extraction, and subsequently, the extracted data were combined. The process of feature extraction is described in Section 2.3. Ultimately, training set 1 comprises 2412 RBPs and 6961 non-RBPs, while training set 2 consists of 4801 RBPs and 6243 non-RBPs. As for the testing set, it encompasses 1626 RBPs and 329 non-RBPs.

2.3. Protein Features and Encoding. The features of protein sequences hold paramount significance in acquiring insights into the fundamental attributes of proteins. In this regard, we employ four distinct methodologies to extract a total of 6944 features from each protein sequence. The availability of these discerning features facilitates the seamless progression toward subsequent stages of model training.

2.3.1. Protr: Obtaining a Series of Physical and Chemical Properties. Protr[30] is an R-package specifically designed for extracting a variety of protein descriptors, including amino acid composition (AAC), pseudo amino acid composition (PAAC), and composition, transition, distribution (CTD) descriptors. AAC describes the proportions of various amino acids in a protein, reflecting the fundamental characteristics of its structure and function. PAAC considers the impact of the sequence position on protein properties, adding spatial structure considerations to traditional AAC. Composition details the proportions of various amino acids in a protein, reflecting its basic compositional characteristics. Transition analysis analyzes the frequency of transitions from one amino acid characteristic to another in the protein sequence, revealing dynamic changes in the sequence. Distribution focuses on the distribution patterns of specific amino acid properties (such as hydrophobicity or hydrophilicity) throughout the entire sequence. The comprehensive analysis of these features allows us to understand the structure and function of proteins from different perspectives. By analyzing these distribution patterns in the sequence, we obtain a 147-dimensional vector that deeply reveals the biological functions of proteins. Amino acids are divided into three different categories based on their unique chemical properties (Table S1), further enriching our understanding of the relationship between protein structure and function. Finally, using this comprehensive information in the protein backbone sequence, we constructed a feature vector with 187 dimensions for each protein sequence (Figure S2), providing rich information for further biological research.

2.3.2. UniRep: Obtaining the Hidden State of a Protein. UniRep[31] uses a multiplicative long short-term memory network to learn the statistical representations of protein sequences from UniRef50,[39] focusing on a profound understanding of biological characteristics. Through this learning mechanism, UniRep can precisely encode input sequence length vectors. The protein sequence features learned from the samples encompass three dimensions: the average hidden layer, the final hidden layer, and the final cell, each with a feature dimension of 1900. The aggregation of these features reflects the complex biological nature of proteins including their structural characteristics, evolutionary information, biophysical properties, and statistical representations. Here, the structural characteristics refer to the three-dimensional arrangement and conformation of amino acids in proteins, which are crucial for their functional efficacy. In terms of statistical representations, UniRep transforms the fundamental features of proteins into semantically rich statistical data through deep learning, thereby capturing the essence of their structures and functions.

Integrating these features, each protein sequence is endowed with a feature vector of 5700 dimensions, emphasizing an in-depth analysis of the biological functions of proteins. In practical applications, the use of UniRep for feature extraction has further validated its utility and ACC in the field of bioinformatics, as evidenced by the changes in the number of protein samples in the training and testing sets. Upon utilizing UniRep for feature extraction, the number of RBPs in training set 1 is reduced from 2780 to 2,682, while the non-RBPs decrease from 7093 to 6986. In the EuRBPDB[35] set, the number of RBPs experiences a change from 4243 to 3,953, and the non-RBPs in the PDB[36] are reduced from 1043 to 1039.

Additionally, the number of RBPs in the testing set is altered from 2122 to 1,988, and the non-RBPs show a decrease from 522 to 521 (Figure S3).

2.3.3. SeqVec: Effectively Capturing the Biophysical Properties. SeqVec[32] utilizes ELMo (embeddings from language models),[40] a natural language processing language model, for modeling protein sequences and uses continuous vectors to represent these sequences. It adeptly captures the essence of proteins by leveraging large-scale, unlabeled biophysical properties, including the hydrophobicity which determines amino acids' interactions with water, charge characteristics that influence the attraction or repulsion among amino acids, and molecular size that impacts the positioning and folding within the protein structure. This approach facilitates the generation of a 1024-dimensional feature vector for each protein sequence.

2.3.4. ESM-1b: Predicting Structure, Function, and Other Protein Properties Directly from a Single Sequence. ESM-1b[33] uses a self-supervised language modeling approach, is effectively applied to various natural language processing tasks, and is suitable for unlabeled amino acid sequences within protein data. This model profoundly learns from a vast protein sequence database, capturing biological structural features at the amino acid level and extending to the entire protein structure. Additionally, ESM-1b reflects the evolutionary relationships between proteins, revealing sequence homology. It also internalizes and represents secondary and tertiary structural information on proteins, crucial aspects of spatial protein structure. Notably, during the pretraining process, the model relies solely on the sequences themselves without any external learning signals, indicating that these biological characteristics emerge naturally in an unsupervised environment. In the case of ESM-1b, the author selected protein sequences with a length below 1,023, thus our data set exclusively includes protein sequences meeting this criterion. Consequently, training set 1 consists of 2418 RBPs and 7067 non-RBPs, while the EuRBPDB[35] training set contains 3505 RBPs and the PDB[36] training set comprises 1038 non-RBPs. As for the testing set, it encompasses 1769 RBPs and 519 non-RBPs. Each protein sequence is associated with a 33-dimensional feature vector.

Tables 1−3 respectively present the number of RBPs and non-RBPs retained after employing different methods for feature extraction in training set 1, training set 2, and the testing set. The sequences processed through diverse methods are combined, and their intersection is determined. Consequently, the final training set 1 consists of 2412 RBPs and 6961 non-RBPs. In training set 2, the EuRBPDB[35] contains 3379 RBPs and the PDB[36] includes 1034 non-RBPs.

Table 1. Number of Positive and Negative Samples After the Training Set 1 Uses Different Methods to Obtain Features

method	RBPs (after feature extraction)	non-RBPs (after feature extraction)
Protr	2767	7003
UniRep	2682	6986
SeqVec	2780	7093
ESM-1b	2418	7067
obtaining intersection	2412	6961

The sequences processed through diverse methods are combined, and their intersection is determined. Consequently, the final training set 1 consists of 2412 RBPs and 6961 non-RBPs. In training set 2, the EuRBPDB[35] contains 3379 RBPs and the PDB[36] includes 1034 non-RBPs. These data sets are merged with the training data set 1, followed by the removal of redundant entries, resulting in training data set 2, comprising 4801 RBPs and 6243 non-RBPs. Regarding the testing set, EuRBPDB contributes 1708 RBPs and PDB contains 517 non-RBPs. We exclude the data present in the Deep-RBPPred sequences from both EuRBPDB and PDB, yielding a final testing set comprising 1626 RBPs and 329 non-RBPs. Each protein sequence within the final testing and training sets is associated with a 6944-dimensional feature vector.

Table 2. Number of Positive and Negative Samples After the Training Set 1 Uses Different Methods to Obtain Features

method	RBPs (after feature extraction)	non-RBPs (after feature extraction)
Protr	4102 (from EuRBPDB)	1039 (from PDB)
UniRep	3953 (from EuRBPDB)	1039 (from PDB)
SeqVec	4243 (from EuRBPDB)	1043 (from PDB)
ESM-1b	3505 (from EuRBPDB)	1038 (from PDB)
obtaining intersection	3379 (from EuRBPDB)	1034 (from PDB)
redundancy after merging with training set 1	4801	6243

Table 3. Number of Positive and Negative Samples After the Training Set 1 Uses Different Methods to Obtain Features

method	RBPs (after feature extraction)	non-RBPs (after feature extraction)
Protr	2053 (from EuRBPDB)	520 (from PDB)
UniRep	1988 (from EuRBPDB)	521 (from PDB)
SeqVec	2122 (from EuRBPDB)	522 (from PDB)
ESM-1b	1769 (from EuRBPDB)	519 (from PDB)
obtaining intersection	1708 (from EuRBPDB)	517 (from PDB)
delete the same sequences as the training set 1	1626	329

2.4. Machine Learning Analysis.

Seq-RBPPred, integrates XGBoost[34] as the central component of its algorithm, harnessing the robust capabilities of XGBoost to execute its designated tasks with efficiency and ACC. Additionally, to underscore the exceptional performance of XGBoost in binary classification models, we include random forest,[41] deep forest,[42] and DmLab[43] as classifiers for comparative analysis.

2.4.1. XGBoost. XGBoost[34] represents a boosting algorithm implementation that effectively incorporates the gradient-boosting decision tree algorithm. It demonstrates remarkable proficiency in addressing both classification and regression problems. Due to its exemplary performance, simplicity, and speed, XGBoost has gained substantial popularity in numerous data competitions and has found extensive utilization across various industries.

For training purposes, we employ the xgb.XGBClassifier module within the scikit-learn[44] framework. This enables us to generate binary classification outputs on both the validation

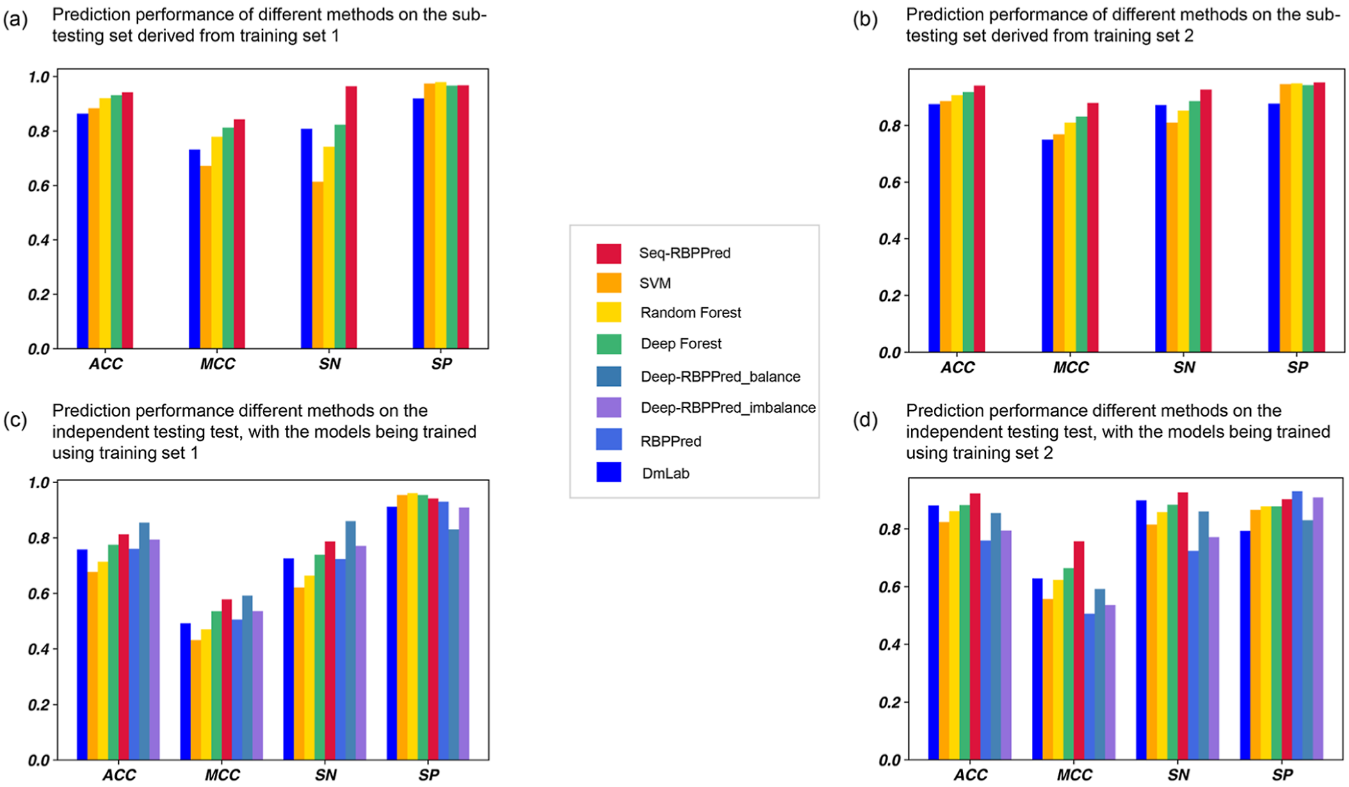


Figure 2. Performance of RBPPred, Deep-RBPPred, SVM, random forest, deep forest, DmLab, and Seq-RBPPred on the subtesting set and independent testing set is evaluated. (a,b) Seq-RBPPred achieves the best performance in the subtesting set, suggesting that XGBoost can be more suitable for predicting RBPs than the other methods mentioned in the article. (c) Performance of the model trained on training set 1 on the testing set. (d) Performance of the model trained on training set 2 on the testing set, where Seq-RBPPred outperformed others in terms of ACC, MCC, and SN, while RBPPred scored the highest in SP. Overall, Seq-RBPPred proves to be more suitable for predicting RBPs than the other methods.

and testing sets alongside individual scores for each protein sequence within the testing set.

2.4.2. SVM Classifier. SVM,[26] known as a linear classifier, primarily tackles binary classification problems through supervised learning, wherein it classifies data based on provided training examples.

In this study, we employ the SVM algorithm in the scikit-learn[44] library, utilizing the radial basis function (RBF) kernel. This facilitates the production of binary classification outputs on the validation and testing sets as well as individual scores for each protein sequence within the testing set.

2.4.3. Random Forest. The random forest[41] classifier constitutes an ensemble classifier that comprises numerous decision trees. By training on a subset of randomly selected training samples and variables, a notable performance.

Within our investigation, the RandomForestClassifier module within scikit-learn[44] is employed for data training. The resulting model is subsequently utilized to generate binary classification outputs on the subtesting set and the testing set, enabling the determination of whether a given protein sequence corresponds to an RBP. Additionally, scores are obtained for each protein sequence within the testing set.

2.4.4. Deep Forest. Zhou[42] proposed the deep forest method, which consistently yields favorable outcomes despite its modest number of layers and basic forest trees. In this method, the random forest serves as the fundamental unit where a single random forest consistently surpasses the performance of an individual neuron.

We use CascadeForestClassifier in the deep forest[42] package for training on the designated training set. Subsequently, binary classification outputs are derived for the validation and testing sets, accompanied by scores for each protein sequence within the testing set.

2.4.5. DmLab: Training and Getting a Set of Rules. DmLab[43] represents software capable of ranking features based on their importance and identifying interdependencies between them. Its core principle revolves around employing Monte Carlo feature selection techniques. In DmLab, the random seed is set to 2022, and the program processes the data to obtain a series of feature rules. By applying these obtained rules to the test set, the prediction ACC for said test set can be determined.

2.5. Performance Evaluation.

In this study, the effectiveness of the model was further assessed through the utilization of cross-validation and independent testing set validation. DmLab[43] uses the 5-fold cross-validation. The training group is randomly divided into ten parts in Seq-RBPPred, SVM,[26] random forest,[41] and deep forest.[42] Specifically, eight subsets were designated as subtraining sets while the remaining two subsets served as subtesting sets.

Each of the aforementioned methods was evaluated on an independent testing set. To ascertain the model's performance, various metrics including ACC, MCC, SN, and SP were employed. These metrics were defined as follows

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}} \tag{1}$$

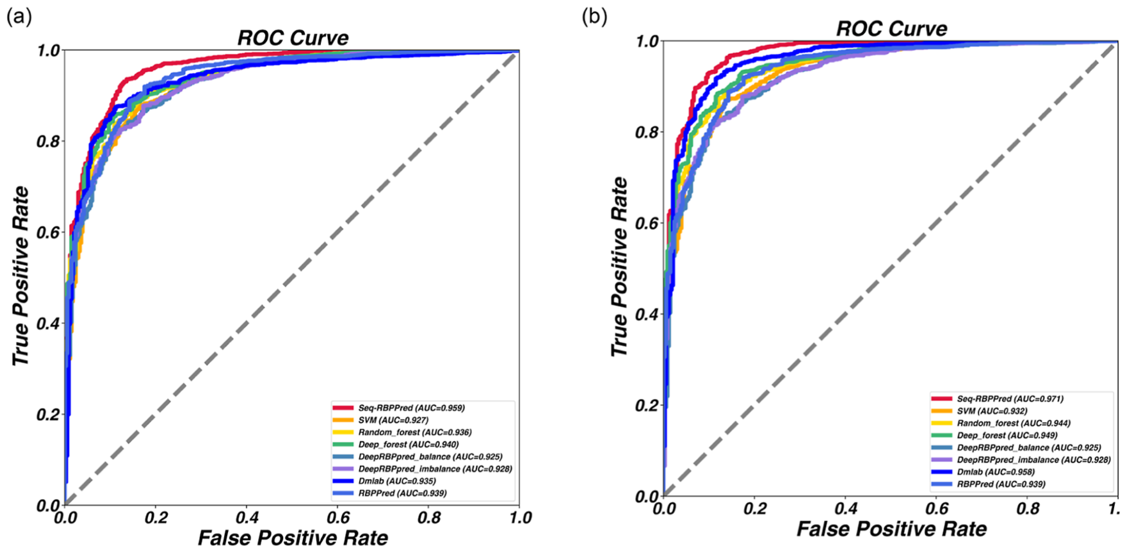


Figure 3. Performance of RBPPred, Deep-RBPPred, SVM, random forest, deep forest, DmLab, and Seq-RBPPred on the testing set. (a) AUROC of the models trained on training set 1, including Deep-RBPPred, SVM, random forest, deep forest, DmLab, and Seq-RBPPred, when tested on an independent test set. Among them, Seq-RBPPred achieved the highest AUC with a value of 0.959. (b) AUROC of the models trained on the training set 2, including SVM, random forest, deep forest, DmLab, and Seq-RBPPred, when evaluated on an independent test set. Once again, Seq-RBPPred exhibited the highest AUC with a value of 0.971.

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN})\cdot(\mathrm{TP} + \mathrm{FP})\cdot(\mathrm{TN} + \mathrm{FP})\cdot(\mathrm{TN} + \mathrm{FN})}} \tag{2}$$

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{4}$$

In the above equation, TP is true positive, FN is false negative, TN is true negative, and FP is false positive.

At the same time, to make the evaluation index more objective, we employed the receiver operating characteristic curve and calculated the AUROC. The larger the AUROC, the better the performance.

3. RESULTS

We present Seq-RBPPred, a sequence-based approach for the prediction of RBPs. To predict RBPs, we extracted relevant features from each protein sequence. These features are then encoded and utilized to train machine learning models, which are subsequently compared to previous prediction methods. Additionally, we employ SVM,[26] random forest,[41] deep forest,[42] and DmLab[43] as alternative classifiers for training and prediction to demonstrate the superior performance of XGBoost[34] in binary classification tasks.

3.1. Performance on the Training Set. The training set is divided randomly into ten parts, with eight sections serving as subtraining sets and the remaining two sections as subtesting sets.

During training test 1, Seq-RBPPred achieves an ACC of 0.942, an MCC of 0.843, an SN of 0.864, and an SP of 0.968. DmLab yields an ACC of 0.864, an MCC of 0.732, an SN of 0.808, and an SP of 0.919. SVM obtains an ACC of 0.884, an MCC of 0.672, an SN of 0.614, and an SP of 0.974. Random forest produces an ACC of 0.920, an MCC of 0.779, an SN of 0.742, and an SP of 0.979. Deep forest demonstrates an ACC of 0.931, an MCC of 0.813, an SN of 0.823, and an SP of 0.967.

In training test 2, Seq-RBPPred achieves an ACC of 0.940, an MCC of 0.879, an SN of 0.926, and an SP of 0.951. DmLab yields an ACC of 0.875, an MCC of 0.750, an SN of 0.872, and an SP of 0.877. SVM obtains an ACC of 0.885, an MCC of 0.769, an SN of 0.810, and an SP of 0.945. Random forest produces an ACC of 0.906, an MCC of 0.810, an SN of 0.852, and an SP of 0.948. Deep forest demonstrates an ACC of 0.917, an MCC of 0.831, an SN of 0.885, and an SP of 0.942.

Figure 2a,b demonstrates the predictive performance of various methods on subtesting sets within training sets 1 and 2. Notably, Seq-RBPPred, employing XGBoost, exhibited superior performance, highlighting the advantages of XGBoost in RBP prediction over other methods. The strengths of XGBoost are attributable to its gradient boosting framework, which effectively addresses various types of data biases, a crucial aspect for complex bioinformatics tasks such as RBP prediction. Additionally, the regularization component incorporated in its training process significantly reduces the risk of overfitting, which is essential for dealing with the high complexity and variability inherent in RNA sequence data. Furthermore, XGBoost's efficiency and scalability in handling large data sets underscore its superiority in RBP prediction.

3.2. Performance of Seq-RBPPred on the Independent Testing Data Set. When applying the aforementioned method to an independent testing set consisting of 1626 RBPs and 329 non-RBPs, Seq-RBPPred consistently achieves the highest performance. Deep-RBPPred,[28] incorporates two methods, namely "balance" and "imbalance", for testing the protein sequences within our independent testing set.

Within training test 1, the independent testing set results for Seq-RBPPred reveal an ACC of 0.813, MCC of 0.578, SN of 0.787, and SP of 0.942. Deep-RBPPred generates a score for each sequence within the independent testing set with a threshold of 0.5. Protein sequences with scores above 0.5 are considered RBPs, while those below 0.5 are deemed non-RBPs.

In the balance model, the ACC is 0.855, the MCC is 0.592, the SN is 0.860, and the SP is 0.830. In the imbalance model, the ACC is 0.794, the MCC is 0.536, the SN is 0.771, and the SP is 0.909. It is observed that the balance model within Deep-RBPPred outperforms Seq-RBPPred in terms of ACC, MCC, and SN. However, Seq-RBPPred demonstrates superior ACC, MCC, SN, and SP when compared with the unbalanced model within Deep-RBPPred. Our evaluation utilizes the entire data of Deep-RBPPred and not its balanced data set, we focus on comparing it with the imbalanced model of Deep-RBPPred.

Satisfactory results were achieved when training our method on the data from training set 1, thereby validating the effectiveness of Seq-RBPPred in the RBP prediction. Consequently, we proceeded to employ the model trained using training set 2 on an independent testing set to evaluate the performance of Seq-RBPPred. In training test 2, Seq-RBPPred within the independent testing set yields an ACC of 0.922, an MCC of 0.757, an SN of 0.926, and an SP of 0.903.

Similarly, we evaluated the results of SVM, random forest, deep forest, and DmLab as model classifiers while ensuring consistent input data (Tables S2 and S3). The findings indicate that Seq-RBPPred performs better within the binary classification model.

Figure 2c,d demonstrates the performance of the model trained using training sets 1 and 2 on the independent testing set. In training test 2, Seq-RBPPred demonstrates superior performance by achieving the highest values for ACC, MCC, and SN. Additionally, RBPPred exhibited the highest score in SP. This could be attributed to Seq-RBPPred's design emphasis on enhancing the ACC of positive samples (true positives), thereby achieving superior performance in terms of ACC, MCC, and SN. However, this focus may lead to a comparatively poorer classification of negative samples (true negatives), subsequently reducing the SP score. The prioritization of ACC for positive instances may result in a trade-off with the classification effectiveness for negative instances, impacting the overall SP score in the evaluation metrics. Overall, Seq-RBPPred is more suitable for predicting RBPs than other methods.

To reduce the impact of score thresholds on the comparative analysis of results, this study employs ROC curves for a more comprehensive evaluation of the predictive capabilities of various methodologies on the independent testing data set. Seq-RBPPred demonstrates optimal performance on both training set 1 and training set 2, achieving area under the curve (AUC) values of 0.959 and 0.971, respectively. Figure 3 distinctly illustrates the ROC curves for diverse methodologies on the independent testing set, using training sets 1 and 2 as the training data sets. These results suggest that enhancing the feature representation of amino acid sequences significantly aids Seq-RBPPred in more effectively learning RBP, thereby improving the ACC and efficiency of predictions.

4. DISCUSSION

As the volume of genomic data continues to expand and advancements in machine learning techniques progress, computational approaches are increasingly being utilized in the field of genomics. However, the application of these methods in predicting RBPs remains somewhat limited. RBPs play a pivotal role in the biological processes of organisms. Presently, there are two primary methods for RBP prediction: one based on protein structural information and the other on protein sequence information. Among these, Deep-RBPPred, which uses protein sequence information, has shown a higher ACC. However, it suffers from an extended computational time and requires further enhancement in ACC.

To overcome these limitations, we propose a novel XGBoost-based method, named Seq-RBPPred, for RBP prediction. This method integrates 6944 protein sequence features and employs machine learning techniques as described in prior research to improve predictive performance (code available at https://github.com/yaoyao-11/Seq-RBPPred). Our analysis indicates that Seq-RBPPred surpasses Deep-RBPPred in training and prediction results for both RBPs and non-RBPs. Additionally, Seq-RBPPred demonstrated favorable performance in an independent testing set. To ensure a fair evaluation, we compared the AUROC curves of DmLab, SVM, random forest, deep forest, and Seq-RBPPred on an independent testing set. Notably, Seq-RBPPred shows the largest AUC, indicating its superior performance. These results affirm the advantages of this method over Deep-RBPPred in terms of ACC and computational efficiency. Consequently, we postulate numerous potential RBPs await discovery using this approach.

While Seq-RBPPred has achieved some progress, there is still room for improvement. Currently, we have employed only conventional machine learning methods and have not utilized deep learning as the classifier in our model. Therefore, future research could explore the use of deep learning techniques for predicting RBPs.

■ ASSOCIATED CONTENT

Data Availability Statement

The data and documentation underlying this article are available at: https://github.com/yaoyao-11/Seq-RBPPred.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c08381.

> Figure S1: Build training and testing sets. Figure S2: Flowchart of feature extraction using Protr. Figure S3: The process diagram for feature extraction using UniRep. Table S1: Classification of amino acid properties. Table S2: Prediction performance of RBPPred, Deep-RBPPRed, DmLab, SVM, random forest, deep forest, and Seq-RBPPred on the independent testing test, with the models being trained using training set 1. Table S3: Prediction performance of RBPPred, Deep-RBPPRed, DmLab, SVM, random forest, deep forest, and Seq-RBPPred on the independent testing test, with the models being trained using training set 2 (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Sijia Wang − *CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai 200021, China*; Email: wangsijia@sinh.ac.cn

Tao Huang − *CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai 200021, China*; Email: huangtao@sinh.ac.cn

Authors

Yuyao Yan − *CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai 200021, China;* orcid.org/0009-0000-1226-0879

Wenran Li − *CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai 200021, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c08381

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by Strategic Priority Research Program of Chinese Academy of Sciences [XDB38050200, XDB38040202, XDA26040304], CAS Young Team Program for Stable Support of Basic Research [YSBR-077], National Natural Science Foundation of China (NSFC) [32200472], China Postdoctoral Science Foundation [2021M693274, BX2021336], National Key R&D Program of China [2022YFF1203202], and Self-supporting Program of Guangzhou Laboratory [SRPG22-007].

■ REFERENCES

(1) Gebauer, F.; Hentze, M. W. Molecular mechanisms of translational control. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 827−835.

(2) Roundtree, I. A.; Evans, M. E.; Pan, T.; He, C. Dynamic RNA modifications in gene expression regulation. *Cell* **2017**, *169*, 1187−1200.

(3) Huelga, S. C.; Vu, A. Q.; Arnold, J. D.; Liang, T. Y.; Donohue, J. P.; Shiue, L.; Hoon, S.; Brenner, S.; Ares, M.; Yeo, G. W. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep.* **2012**, *1*, 167−178.

(4) Li, P.; Banjade, S.; Cheng, H. C.; Kim, S.; Chen, B.; Guo, L.; Llaguno, M.; Hollingsworth, J. V.; King, D. S.; Banani, S. F.; et al. Phase transitions in the assembly of multivalent signalling proteins. *Nature* **2012**, *483*, 336−340.

(5) Alipanahi, B.; Delong, A.; Weirauch, M. T.; Frey, B. J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **2015**, *33*, 831−838.

(6) Trabelsi, A.; Chaabane, M.; Ben-Hur, A. Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. *Bioinformatics* **2019**, *35*, i269−i277.

(7) Hassanzadeh, H. R.; Wang, M. D. DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins. In *2016 IEEE International conference on bioinformatics and biomedicine (BIBM)*, 2016, pp 178−183.

(8) Zhang, S.; Zhou, J.; Hu, H.; Gong, H.; Chen, L.; Cheng, C.; Zeng, J. A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Res.* **2016**, *44*, No. e32.

(9) Wang, F.; Chainani, P.; White, T.; Yang, J.; Liu, Y.; Soibam, B. Deep learning identifies genome-wide DNA binding sites of long noncoding RNAs. *RNA Biol.* **2018**, *15*, 1468−1476.

(10) Umarov, R. K.; Solovyev, V. V. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PloS One* **2017**, *12*, No. e0171410.

(11) Oubounyt, M.; Louadi, Z.; Tayara, H.; Chong, K. T. DeePromoter: robust promoter predictor using deep learning. *Front. Genet.* **2019**, *10*, 286.

(12) Kleftogiannis, D.; Kalnis, P.; Bajic, V. B. DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res.* **2015**, *43*, No. e6.

(13) Min, X.; Zeng, W.; Chen, S.; Chen, N.; Chen, T.; Jiang, R. Predicting enhancers with deep convolutional neural networks. *BMC Bioinf.* **2017**, *18*, 478.

(14) Wang, Y.; Liu, T.; Xu, D.; Shi, H.; Zhang, C.; Mo, Y.-Y.; Wang, Z. Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks. *Sci. Rep.* **2016**, *6*, 19598.

(15) Khanal, J.; Tayara, H.; Zou, Q.; Chong, K. T. Identifying dna n4-methylcytosine sites in the rosaceae genome with a deep learning model relying on distributed feature representation. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 1612−1619.

(16) Angermueller, C.; Lee, H. J.; Reik, W.; Stegle, O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **2017**, *18*, 67.

(17) Yin, Q.; Wu, M.; Liu, Q.; Lv, H.; Jiang, R. DeepHistone: a deep learning approach to predicting histone modifications. *BMC Genomics* **2019**, *20*, 193.

(18) Baisya, D. R.; Lonardi, S. Prediction of histone post-translational modifications using deep learning. *Bioinformatics* **2021**, *36*, 5610−5617.

(19) Gudenas, B. L.; Wang, L. Prediction of LncRNA subcellular localization with deep learning from sequence features. *Sci. Rep.* **2018**, *8*, 16385.

(20) Yan, Z.; Lécuyer, E.; Blanchette, M. Prediction of mRNA subcellular localization using deep recurrent neural networks. *Bioinformatics* **2019**, *35*, i333−i342.

(21) Almagro Armenteros, J. J.; Sønderby, C. K.; Sønderby, S. K.; Nielsen, H.; Winther, O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* **2017**, *33*, 3387−3395.

(22) Leung, M. K.; Xiong, H. Y.; Lee, L. J.; Frey, B. J. Deep learning of the tissue-regulated splicing code. *Bioinformatics* **2014**, *30*, i121−i129.

(23) Jha, A.; Gazzara, M. R.; Barash, Y. Integrative deep models for alternative splicing. *Bioinformatics* **2017**, *33*, i274−i282.

(24) Zhou, J.; Theesfeld, C. L.; Yao, K.; Chen, K. M.; Wong, A. K.; Troyanskaya, O. G. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **2018**, *50*, 1171−1179.

(25) Shazman, S.; Mandel-Gutfreund, Y. Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput. Biol.* **2008**, *4*, No. e1000146.

(26) Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Their Appl.* **1998**, *13*, 18−28.

(27) Zhao, H.; Yang, Y.; Zhou, Y. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. *RNA Biol.* **2011**, *8*, 988−996.

(28) Zheng, J.; Zhang, X.; Zhao, X.; Tong, X.; Hong, X.; Xie, J.; Liu, S. Deep-RBPPred: Predicting RNA binding proteins in the proteome scale based on deep learning. *Sci. Rep.* **2018**, *8*, 15264.

(29) Zhang, X.; Liu, S. RBPPred: predicting RNA-binding proteins from sequence using SVM. *Bioinformatics* **2017**, *33*, 854−862.

(30) Xiao, N.; Cao, D.-S.; Zhu, M.-F.; Xu, Q.-S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* **2015**, *31*, 1857−1859.

(31) Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **2019**, *16*, 1315−1322.

(32) Heinzinger, M.; Elnaggar, A.; Wang, Y.; Dallago, C.; Nechaev, D.; Matthes, F.; Rost, B. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinf.* **2019**, *20*, 723.

(33) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; et al. Biological structure and

function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, No. e2016239118.

(34) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp 785−794.

(35) Liao, J.-Y.; Yang, B.; Zhang, Y.-C.; Wang, X.-J.; Ye, Y.; Peng, J.-W.; Yang, Z.-Z.; He, J.-H.; Zhang, Y.; Hu, K.; et al. EuRBPDB: a comprehensive resource for annotation, functional and oncological investigation of eukaryotic RNA binding proteins (RBPs). *Nucleic Acids Res.* **2020**, *48*, D307−D313.

(36) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(37) Wang, G.; Dunbrack, R. L. PISCES: a protein sequence culling server in Bioinformatics. *Bioinformatics* **2003**, *19*, 1589−1591.

(38) Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658−1659.

(39) Suzek, B. E.; Wang, Y.; Huang, H.; McGarvey, P. B.; Wu, C. H.; UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **2015**, *31*, 926−932.

(40) Sarzynska-Wawer, J.; Wawer, A.; Pawlak, A.; Szymanowska, J.; Stefaniak, I.; Jarkiewicz, M.; Okruszek, L. Detecting formal thought disorder by deep contextualized word representations. *Psychiatr. Res.* **2021**, *304*, 114135.

(41) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.

(42) Zhou, Z.-H.; Feng, J. Deep forest. *Natl. Sci. Rev.* **2019**, *6*, 74−86.

(43) Dramiński, M.; Rada-Iglesias, A.; Enroth, S.; Wadelius, C.; Koronacki, J.; Komorowski, J. Monte Carlo feature selection for supervised classification. *Bioinformatics* **2008**, *24*, 110−117.

(44) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.