ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

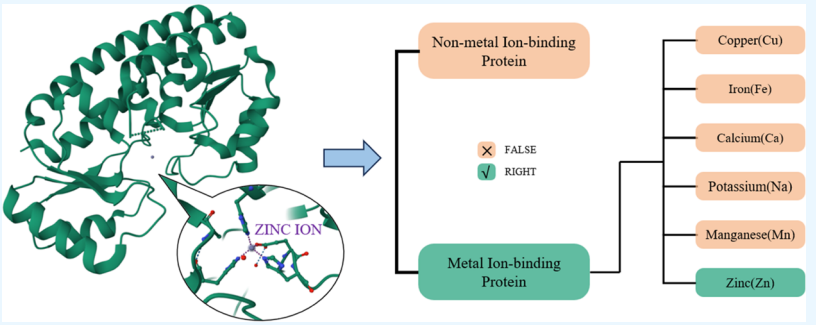
MIBPred: Ensemble Learning-Based Metal Ion-Binding Protein Classifier

Hong-Qi Zhang,[⊥] Shang-Hua Liu,[⊥] Rui Li, Jun-Wen Yu, Dong-Xin Ye, Shi-Shi Yuan, Hao Lin,* Cheng-Bing Huang,* and Hua Tang*

Cite This: *ACS Omega* 2024, 9, 8439−8447

Read Online

ACCESS | Metrics & More | Article Recommendations



ABSTRACT: In biological organisms, metal ion-binding proteins participate in numerous metabolic activities and are closely associated with various diseases. To accurately predict whether a protein binds to metal ions and the type of metal ion-binding protein, this study proposed a classifier named MIBPred. The classifier incorporated advanced Word2Vec technology from the field of natural language processing to extract semantic features of the protein sequence language and combined them with position-specific score matrix (PSSM) features. Furthermore, an ensemble learning model was employed for the metal ion-binding protein classification task. In the model, we independently trained XGBoost, LightGBM, and CatBoost algorithms and integrated the output results through an SVM voting mechanism. This innovative combination has led to a significant breakthrough in the predictive performance of our model. As a result, we achieved accuracies of 95.13% and 85.19%, respectively, in predicting metal ion-binding proteins and their types. Our research not only confirms the effectiveness of Word2Vec technology in extracting semantic information from protein sequences but also highlights the outstanding performance of the MIBPred classifier in the problem of metal ion-binding protein types. This study provides a reliable tool and method for the in-depth exploration of the structure and function of metal ion-binding proteins.

1. INTRODUCTION

Proteins serve as the executors of cellular functions and play a crucial role in cell growth.[1] According to their interaction with metal ions, proteins can be categorized into two major types: metal ion-binding proteins (MIBP) and nonmetal ion-binding proteins (NMIBP). Nearly 40% of proteins in the Protein Data Bank (PDB) have been found to bind to metal ions.[2,3] Metal ions are essential for maintaining the stability of protein structures as well as for cellular biological functions such as enzyme catalysis and gene expression regulation. For instance, copper ion-binding proteins such as LOX (lysine oxidase), MAP2K1, and SPARC are associated with promoting the proliferation and invasion of tumor cells in lung cancer and breast cancer.[4−7] Therefore, identifying metal ion-binding proteins and their types is conducive to understanding the mechanisms of related diseases and designing targeted drugs.[8]

Unfortunately, methods for identifying metal ion-binding proteins, such as nuclear magnetic resonance spectroscopy, gel electrophoresis, metal affinity column chromatography, electrophoretic mobility assays, absorbance spectroscopy, and mass spectrometry, often involve complex procedures and specialized equipment.[9,10] These methods are not suitable for unknown targets and are both time-consuming and labor-intensive. With the advancement of high-throughput technology, protein sequencing has become increasingly accessible. It is easy to retrieve various metal ion-binding protein sequences from diverse databases. Consequently, the development of an accurate and efficient classifier utilizing metal ion-binding protein sequences has become imperative.

Received: December 1, 2023
Revised: January 16, 2024
Accepted: January 22, 2024
Published: February 8, 2024

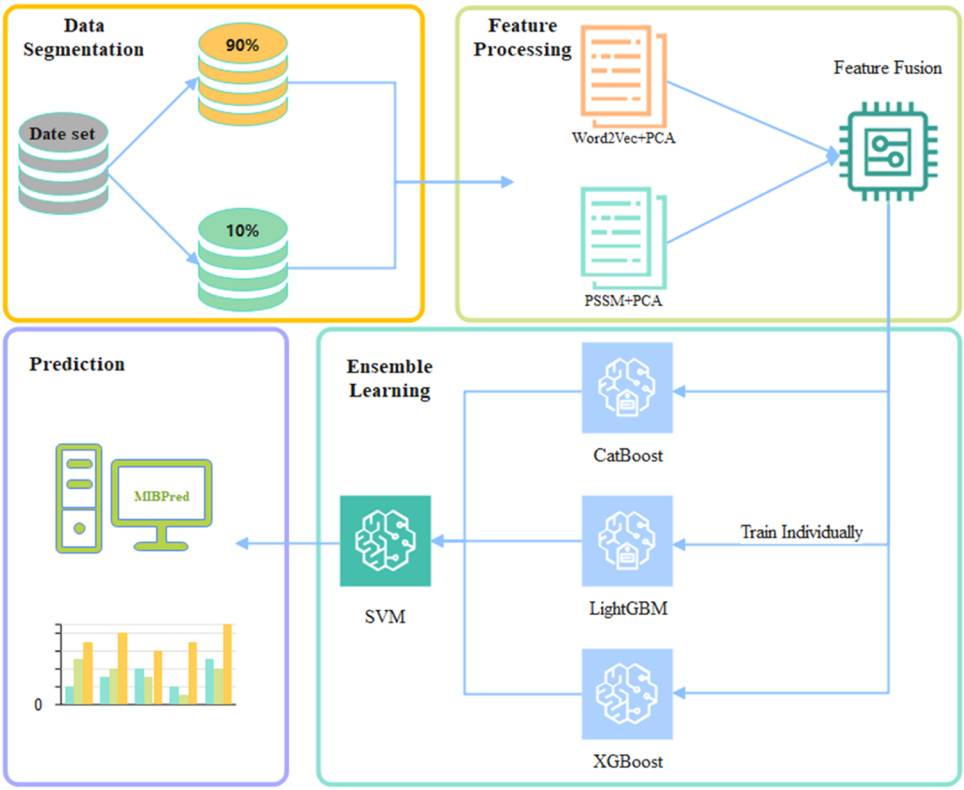


Figure 1. Model Workflow Architecture Diagram.

In the past, many machine learning methods used for predicting metal ion-binding proteins have shown outstanding performance. For example, experts have developed MetalPredator and RPCIBP classifiers for predicting proteins related to iron and sulfur clusters and copper ion-binding proteins, respectively.[11−13] In addition, some studies have also achieved favorable results in predicting metal ion-binding sites, including iron ion-binding sites, zinc ion-binding sites, and copper ion-binding sites.[14−18] However, there is currently no efficient and stable multimetal ion-binding protein classifier.

Ensemble learning is an excellent strategy for building predictive models. In ensemble learning, it is essential to first construct multiple basic learners, which can be parallel or sequential. Subsequently, these basic learners are combined and utilized in another machine learning method for prediction.[19−22] Common combination methods include majority voting for classification problems and weighted averaging for regression problems. However, we adopted an innovative approach that utilized support vector machines (SVMs)[23,24] to automatically learn the voting rules from the data to handle our prediction problem.[25−28]

Word2Vec technology is a crucial technique in the field of natural language processing.[29−32] It transforms vocabulary into computable low-dimensional vector representations, solving the issues of high dimensionality and inaccurate semantic similarity in traditional methods. In protein research, understanding the semantic information and functions of the protein sequence language is vital for revealing biological processes and disease mechanisms. Word2Vec technology maps protein sequences into high-dimensional vector spaces, preserving semantic relationships between sequence languages and offering a new avenue for protein research.[33]

In this study, we developed a classifier called MIBPred, based on Word2Vec and ensemble learning. This classifier successfully discriminates metal ion-binding proteins from nonmetal ion-binding proteins and further identifies the specific type of metal ion-binding proteins. Word2Vec effectively extracts information from protein sequences, significantly enhancing the predictive performance of the model.[34,35] The specific process is illustrated in Figure 1. We collected sequences of six types of metal ion-binding proteins and nonmetal ion-binding proteins, divided into training and independent test sets in a ratio of 9:1. Features were extracted using Word2Vec and reduced to 512 dimensions through principal component analysis (PCA).[36] Simultaneously, PSSM matrices were generated for the protein sequences and also reduced to 512 dimensions through PCA. These two types of features were then fused. The fused features were individually input into the XGBoost (extreme gradient boosting), LightGBM (light gradient boosting machine), and CatBoost (category boosting) machine learning algorithms.[37−41] The output values of these three algorithms were integrated into the SVM for ensemble learning. Ultimately, our classifier demonstrated excellent internal performance, and examination on the independent test set confirmed the strong generalization and robustness of the model. The following section will introduce the construction procession of the model in details.

2. MATERIALS AND METHODS

2.1. Datasets. In constructing a classifier with strong predictive performance, the use of diverse and representative datasets is crucial. We collected data from two widely used protein databases, PDB and Uniprot,[42,43] encompassing seven distinct protein categories, including six types of metal ion-binding proteins and nonmetal ion-binding proteins. Specifi-

https://doi.org/10.1021/acsomega.3c09587
ACS Omega 2024, 9, 8439−8447

cally, metal ion-binding proteins include calcium ion-binding proteins (4554), copper ion-binding proteins (570), iron ion-binding proteins (2000), potassium ion-binding proteins (986), manganese ion-binding proteins (1422), and zinc ion-binding proteins (6124), while nonmetal ion-binding proteins totaled 5477. For the purpose of model training, each protein category was divided randomly into a training set comprising 90% of the samples and a separate independent test set comprising 10% of the samples. Within this 90% training set, a random 20% of the data were selected to form a validation set, which was utilized for parameter fine tuning during the training of XGBoost, LightGBM, and CatBoost models. Subsequently, the aforementioned independent test set was further subdivided into two parts at 7:3 ratio. The initial portion was employed to train the SVM voting model, while the latter portion was set aside for the final independent data testing.

2.2. Design of the MIBPred Classifier. We employed the Word2Vec model to extract semantic features from the protein sequence language, representing each protein sequence as a low-dimensional vector of 512 dimensions (after PCA dimensionality reduction).[44] Additionally, we utilized PSSM features to reflect the evolutionary information on each amino acid, which were also reduced to 512 dimensions through PCA.[11] We adopted PCA dimensionality reduction technology to accelerate the model training speed. By connecting these two types of features, we obtained a 1024-dimensional feature vector. This fusion strategy allows our model to comprehensively capture protein feature information.

The fused features are independently trained by three machine learning algorithms (XGBoost, LightGBM, and CatBoost), which perform well in classification problems. These algorithms are capable of handling complex nonlinear relationships, significantly enhancing the predictive performance of the model. Each individual model generates a prediction, and we input the predictions from these three models into an SVM for voting. For the identification of metal ion-binding protein types, our SVM adopts a one-vs-one multiclassification method. Through the voting mechanism of the SVM, we obtained the final prediction by integrating the perspectives of each model. This voting fusion method effectively enhances the accuracy and stability of predictions, providing reliable results for the identification of metal ion-binding proteins and their types.

2.3. MIBPred Classifier Parameter Settings. In our study, we adopted the model of continuous bag of words (CBOW) as the foundational architecture for Word2Vec. In the parameter settings, we configured the word vector dimensions to be 50, the maximum distance between the current word and the predicted word to be 5, the learning rate to be 0.0001, and the number of iterations to be 5. Additionally, we meticulously tuned the parameters of the fundamental classifiers, namely, CatBoost, LightGBM, and XGBoost, to achieve optimal performance in our predictive model. For the CatBoost model, we set the number of iterations to 2500, the learning rate to 0.06, and the maximum depth to 7. We chose Logloss as the loss function and adjusted the 'border_count' parameter to 254 for optimizing feature split points. In the LightGBM model, we utilized gradient boosting decision trees as the boosting method, setting the number of leaf nodes in each tree to 31. The learning rate was set to 0.06, and the model was trained for 4000 iterations. To prevent overfitting, we specified a minimum of 20 samples for each leaf node and randomly selected 90% of the features during training. In the configuration of XGBoost, we specified 3000 iterations, a learning rate of 0.03, a maximum tree depth of 6, a minimum leaf node weight of 1, and Logloss as the loss function. In the ensemble approach, we employed support vector machines (SVMs) as the voter, setting the regularization parameter to 1 and utilizing the radial basis function as the kernel. We also set the polynomial kernel function degree to 3 and the kernel function coefficient to 'scale'. When predicting the types of metal ion-binding proteins, we made specific modifications to certain parameters of the CatBoost model. We increased the number of iterations to 5000, modified the maximum depth to 8, and adjusted the learning rate to 0.09. For the LightGBM model, we fine-tuned the number of iterations to 2500 and the learning rate to 0.01. In the XGBoost model, we set the number of iterations to 4500 and the learning rate to 0.01. These parameter choices were carefully balanced to achieve the optimal predictive performance across all models.

2.4. Performance Evaluation. The precision (P), recall (R), accuracy (ACC), and $F1$ score were computed to measure the performance of models across the prediction process.[45−54] According to the definition of these evaluation quantities, they can be expressed as follows

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{FP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{3}$$

$$F1 - \mathrm{score} = \frac{2PR}{P + R} \tag{4}$$

where TP, TN, FP, and FN represent the samples' true positive, true negative, false positive, and false negative, respectively.

Above are the computation precision (P), recall (R), and $F1$ score for binary classing. For multiclass classification, we used a weighted average calculation. Its specific formula is as follows

$$\mathrm{index}_{\mathrm{weighted}} = \sum_{1}^{6} (w_i \times \mathrm{index}_i) / \sum_{1}^{6} w_i \tag{5}$$

where $\mathrm{index}_{\mathrm{weighted}}$ represents the weighted average index, w_i represents the weight of the number of samples in each category, and index_i represents the metric of each category (including P, R, $F1$ score).

3. RESULTS AND DISCUSSION

3.1. Discriminating MIBP from NMIBP. In biological organisms, there are numerous metal ion-binding proteins that are crucial for maintaining vital life activities and many metabolic processes within the organism. However, there is still relatively insufficient research on the classification of various metal ion-binding proteins in this field. To address this challenge, we designed a high-performance classifier based on ensemble learning. This classifier can effectively determine whether a protein binds to metal ions based on its amino acid sequence. In our experiments, our classifier significantly outperformed various outstanding machine learning algorithms, including the SVM, decision trees, random forests, XGBoost, LightGBM, and CatBoost, in identifying metal ion-

https://doi.org/10.1021/acsomega.3c09587
ACS Omega 2024, 9, 8439−8447

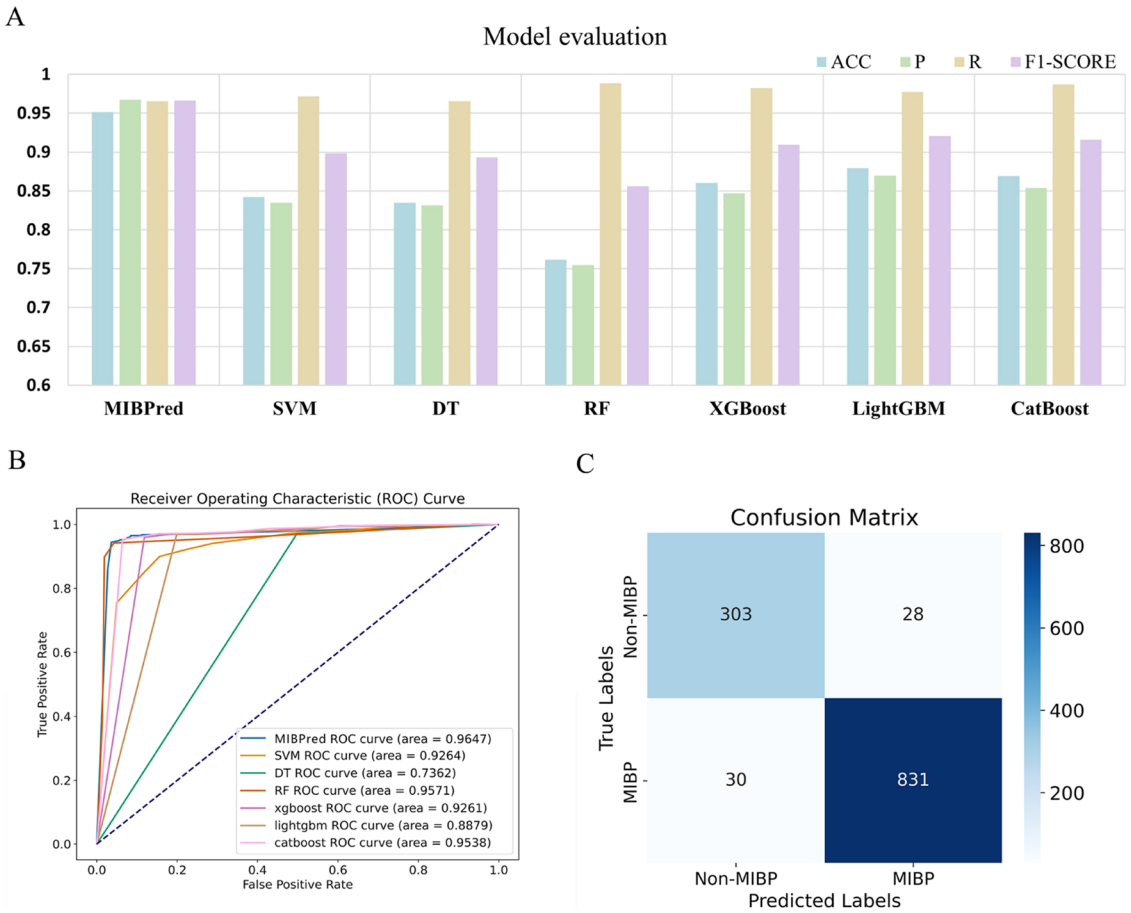


Figure 2. Results of MIBPred for the predictions of metal ion-binding proteins. (A) Compared with other classic machine learning models used to determine metal ion-binding proteins, MIBPred outperforms other models significantly, and four metrics exhibit greater stability in its performance. (B) ROC curves were used to measure the prediction performance of MIBPred and other classical machine learning models. (C) Confusion matrix results of MIBPred show that our model exhibits excellent performance in both positive and negative sample recognition tasks, indicating its excellent effectiveness.

binding proteins. Specifically, our classifier achieved an ACC of 95.13% in identifying metal ion-binding proteins, with an *F*1 score of 0.9663, a recall of 0.9652, and a precision of 0.9674 (Figure 2A and Table 1). The results demonstrate the excellent performance of our classifier in terms of ACC, marking a significant breakthrough. Additionally, the area under the receiver operating characteristic curve (ROC) of the proposed classifier was 0.9647, significantly higher than the area under the ROC (AUC) values of SVM, decision trees (DT), random forests (RF), XGBoost, LightGBM, and CatBoost (Figure 2B). The confusion matrix analysis reveals a well-balanced performance of our model in both positive and negative samples (Figure 2C). This result indicates that our model exhibits higher ACC and reliability in distinguishing positive and negative samples, making it suitable for specific prediction tasks related to this particular problem.

Table 1. Comparison of Different Methods on the Discrimination between MIBP and NMIBP[a]

classifier	ACC	*P*	*R*	*F*1 score
MIBPred	**0.9513**	**0.9674**	0.9652	**0.9663**
LightGBM	0.8791	0.8699	0.9775	0.9206
CatBoost	0.8691	0.8540	0.9869	0.9157
XGBoost	0.8603	0.8472	0.9822	0.9097
SVM	0.8422	0.8349	0.9719	0.8982
DT	0.8348	0.8313	0.9653	0.8933
RF	0.7616	0.7546	**0.9888**	0.8560

[a]Note: The bold font indicates the classifiers that work best.

3.2. Predicting the Types of MIBP. We further investigated the application of the MIBPred model in predicting the types of metal ion-binding proteins, including six common and important metal ions: Ca, Cu, Fe, K, Mn, and Zn. Through testing, we found that our classifier outperformed various classical machine learning models, such as SVM, decision trees, and random forests. Specifically, our model achieved satisfactory results in predicting the types of metal ion-binding proteins, with an ACC of 85.19%, an *F*1 score of 0.8520, a recall of 0.8519, and a precision of 0.8571 (Figure 3A and Table 2). Confusion matrix analysis revealed that our model has excellent predictive performance for each metal type, indicating that it performs well in distinguishing true positive rates and false positive rates in various tasks related to metal ion-binding proteins (Figure 3B). Our model provides a powerful tool for the prediction task of metal ion-binding protein types. This is of significant importance for understanding protein functions in living organisms, drug design, disease research, and other fields. By accurately predicting the types of metal ion-binding proteins, we can delve deeper into the biochemical processes within living organisms.

https://doi.org/10.1021/acsomega.3c09587
ACS Omega 2024, 9, 8439−8447

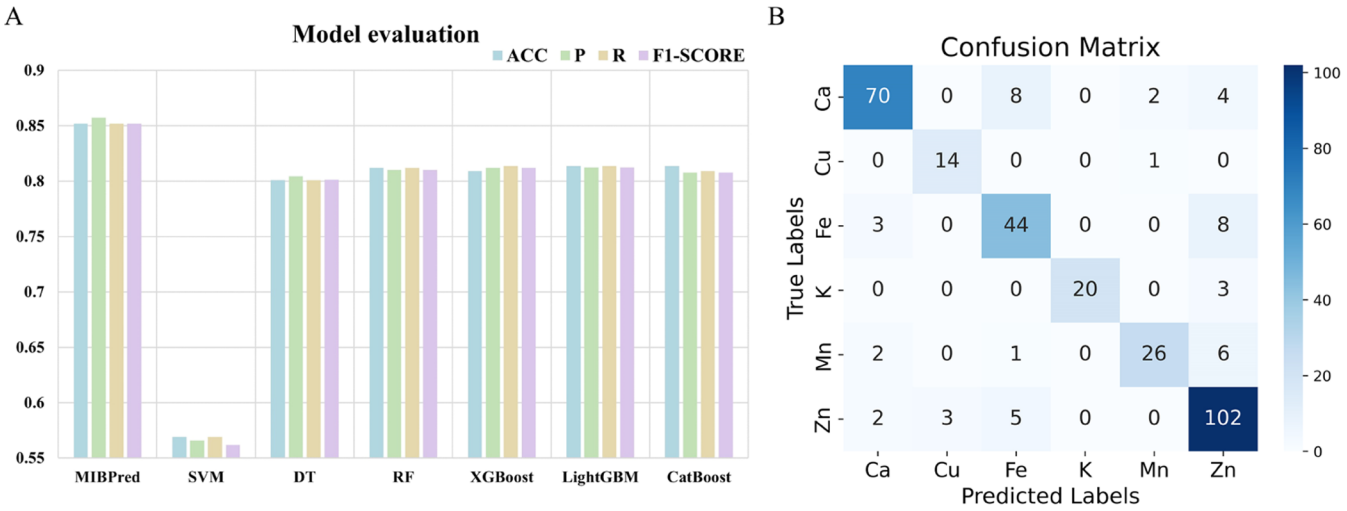


Figure 3. Results of MIBPred for predicting metal ion-binding protein types. (A) Compared with other classic machine learning models, MIBPred is significantly superior, and four metrics exhibit highly balanced performance. (B) Confusion matrix reveals that MIBPred performs well in various metal ion-binding protein recognition tasks.

Table 2. Comparison of Different Methods on the Predictions of Metal Ion-Binding Protein Types[a]

classifier	ACC	*P*	*R*	*F*1 score
MIBPred	**0.8519**	**0.8571**	**0.8519**	**0.8520**
CatBoost	0.8137	0.8076	0.8091	0.8075
LightGBM	0.8137	0.8121	0.8137	0.8122
RF	0.8119	0.8101	0.8119	0.8102
XGBoost	0.8091	0.8120	0.8137	0.8119
DT	0.8007	0.8042	0.8007	0.8011
SVM	0.5690	0.5659	0.5690	0.5617

[a]Note: The bold font indicates the classifiers that work best.

3.3. Evaluation on the Data with Different Ratios. We further examined the performance of our proposed model by gradually decreasing training set proportions from 90% down to 10% and compared it with the machine learning algorithms mentioned earlier, including SVM, decision trees, random forests, and others. The results demonstrate that our model exhibited significantly higher ACC (Figure 4A,C) and AUC (Figure 4B) values compared to other algorithms at data ratios from 90 to 10%. This highlights the robustness and stability of our model across varying training data sizes. Usually, as the proportion of the training set decreases, the performance of the model tends to deteriorate. However, our research indicates that even with relatively small datasets, our model can still extract valuable information, maintaining superior ACC and predictive capability, significantly outperforming other classical machine learning models. This outcome holds significant practical relevance; especially, in situations where data are limited or data collection scenarios are challenging, our model can provide reliable predictive results.

3.4. Effectiveness of Word2Vec for Protein Sequence Feature Extraction. In the data extraction phase, we employed Word2Vec for semantic feature extraction of protein sequences and fused it with PSSM features using a concatenation strategy. Experimental results demonstrate that the feature extraction method, which combines Word2Vec with PSSM features, is superior to using PSSM features alone (Figure 5A,B and Table 3). Word2Vec technology can convert protein sequences into low-dimensional vectors while preserving the semantic relationships of the sequence language. By integrating this structural information with PSSM features, our model can more accurately capture valuable protein features, thus improving the prediction performance. This data fusion strategy not only enriches the feature space but also enhances the model's deep understanding of protein sequence information, resulting in more accurate and reliable final predictions.

3.5. Discussion. In this study, we presented MIBPred, a novel computational model for predicting metal ion-binding proteins and their types. By combining structural features extracted using Word2Vec with PSSM features, our model exhibits exceptional performance. The robustness and stability of the model are particularly pronounced across varying data proportions, highlighting its reliability in situations of limited data availability, which is a common challenge in the field of bioinformatics research.[8,55−63]

The integration of Word2Vec structural features significantly enhances the predictive capabilities of our model. By transformation of protein sequences into low-dimensional vectors, Word2Vec preserves intricate semantic relationships, potentially capturing crucial information about interactions between amino acids. The combination of these features with PSSM provides a comprehensive perspective on protein sequence analysis, enabling precise predictions with our model. This innovative feature fusion method can be applied to various bioinformatics tasks and holds the potential to advance our understanding of complex biological phenomena in the field.

In the context of individual model training, we selected machine learning algorithms including XGBoost, LightGBM, and CatBoost. Subsequently, we employed a voting mechanism based on SVM training, which has been demonstrated to be a successful strategy. Ensemble learning techniques are highly regarded for their ability to leverage the advantages of multiple models and compensate for their individual weaknesses. In our study, this approach not only enhanced ACC but also bolstered the model's stability, rendering it highly suitable for practical application scenarios where reliable predictions are of paramount importance.

https://doi.org/10.1021/acsomega.3c09587
ACS Omega 2024, 9, 8439−8447

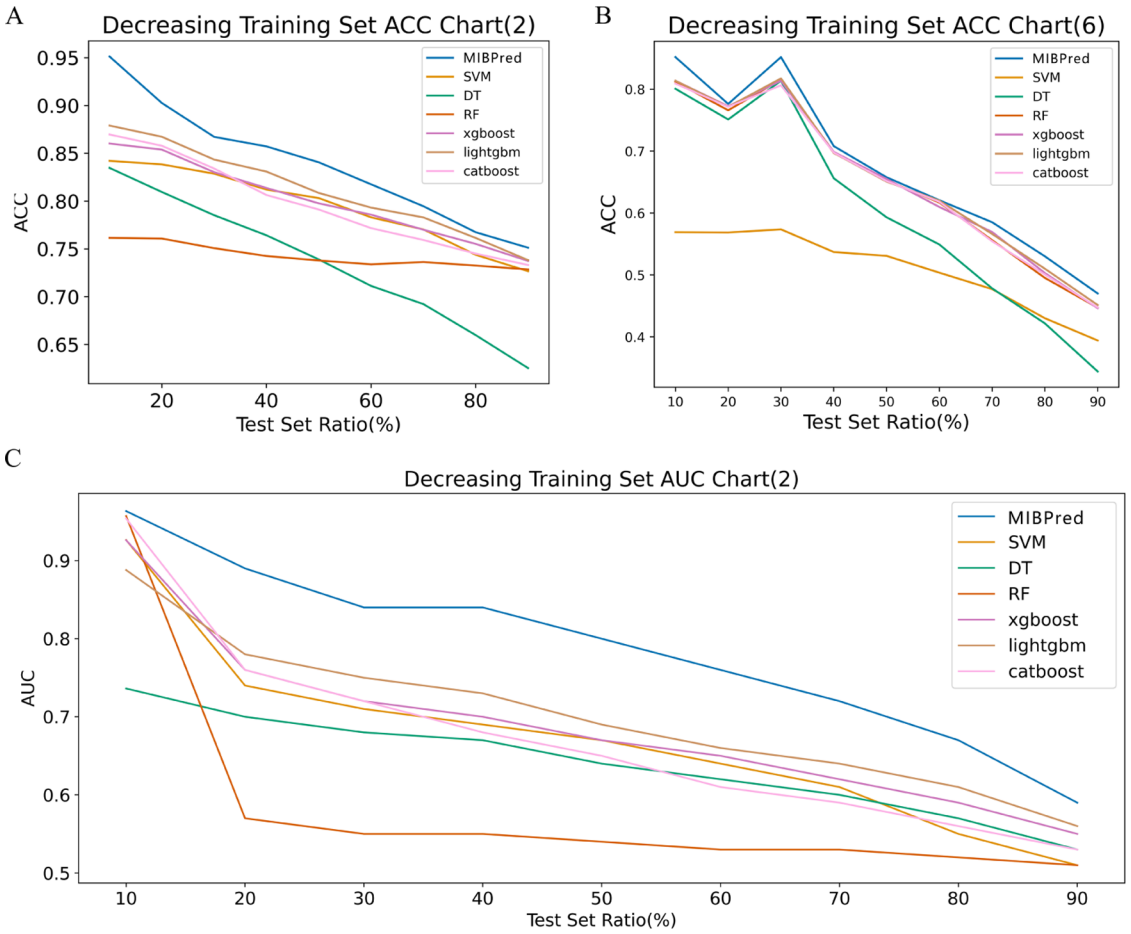


Figure 4. Performance evaluation chart of the MIBPred model at different data ratios. (A) In the process of gradually decreasing the proportion of training sets, we conducted a detailed comparison between the MIBPred model and other classical machine learning algorithms in predicting metal ion-binding proteins. The results showed that the ACC of the MIBPred model was significantly higher than that of other models. (B) Next, in the process of gradually reducing the proportion of training sets, we conducted an in-depth comparison between the MIBPred model and other classical machine learning algorithms in predicting the types of metal ion-binding proteins. The results indicated that the ACC of the MIBPred model was significantly higher than that of other models. (C) Simultaneously, we evaluated the AUC of this model in predicting metal ion-binding proteins, and the results similarly demonstrated the superior performance of the MIBPred model compared to other models. These observations underscore the outstanding performance of the MIBPred model in the task of predicting metal protein binding, even with a reduced training set proportion.

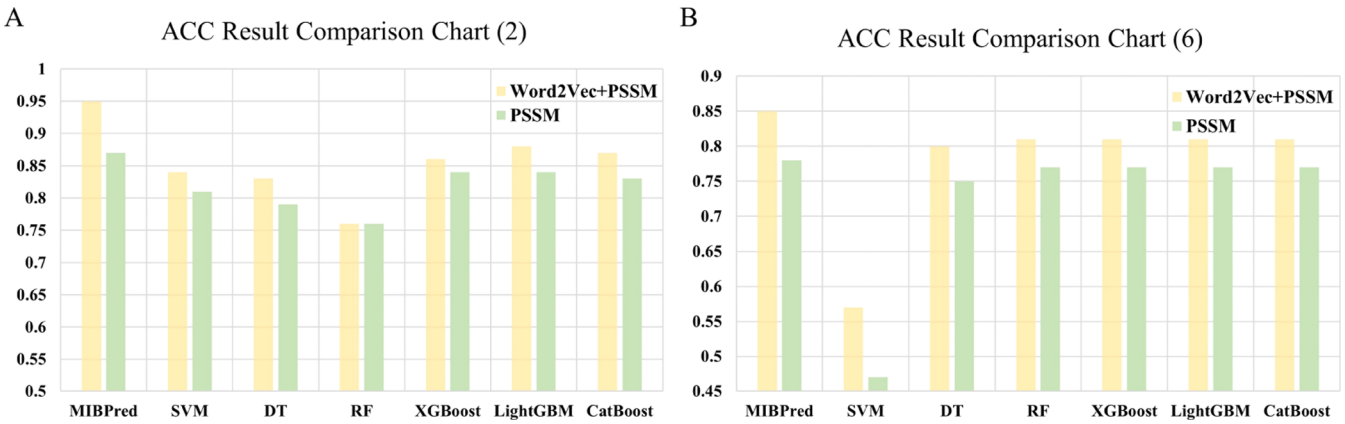


Figure 5. Effectiveness evaluation chart of Word2Vec in protein sequence feature extraction. (A) By comparing the ACC results of the MIBPred model with and without Word2Vec technology in the task of identifying metal ion-binding proteins, we observed that the classification performance of various models was significantly enhanced when using Word2Vec technology. (B) In the task of classifying metal ion-binding protein types, further validation of the ACC results of MIBPred with and without Word2Vec technology demonstrated that the classification performance was significantly improved when Word2Ver technology was also used in various models.

https://doi.org/10.1021/acsomega.3c09587
ACS Omega 2024, 9, 8439−8447

Table 3. Performance Comparison of the Fusion Features (Word2Vec+PSSM) and PSSM Alone[a]

	feature	MIBPred	SVM	DT	RF	XGBoost	LightGBM	CatBoost
MIBP-NMIBP	Word2Vec + PSSM	**0.95**	**0.84**	**0.83**	**0.76**	**0.86**	**0.88**	**0.87**
	PSSM	0.87	0.81	0.79	0.76	0.84	0.84	0.83
MIBP types	Word2Vec + PSSM	**0.85**	**0.57**	**0.80**	**0.81**	**0.81**	**0.81**	**0.81**
	PSSM	0.78	0.47	0.75	0.77	0.77	0.77	0.77

[a]Note: The bold font indicates the classifiers that work best. "MIBP-NMIBP" represents metal ion-binding protein and nonmetal ion-binding protein results. "MIBP types" represents the identification result of the metal ion-binding protein type.

Our model has the ability to predict the metal ion-binding proteins and their types, providing valuable insights into their sequence characteristics. This knowledge holds profound implications for targeted therapy development and novel drug design concerning diseases related to metal ion-binding proteins. Despite the significant progress achieved in our research, avenues for further investigation are present. Exploring additional feature engineering techniques, integrating diverse biological data sources such as structures,[64−70] and exploring alternative machine learning architectures may enhance the predictive capabilities of the model.

4. CONCLUSIONS

This study established a powerful tool called MIBPred to accurately predict metal ion-binding proteins and their types. The Word2Vec structural features were proposed to combine with the PSSM features to formulate protein samples. Subsequently, an innovative learning strategy based on ensemble learning techniques was employed that could provide a reliable model for understanding complex biological phenomena. This model is of great significance for disease treatment and drug design. The source code has been uploaded to GitHub and can be accessed at https://github.com/ZhangHongqi215/MIBPred.

AUTHOR INFORMATION

Corresponding Authors

Hao Lin − *School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China*; orcid.org/0000-0001-6265-2862; Email: hlin@uestc.edu.cn

Cheng-Bing Huang − *School of Computer Science and Technology, Aba Teachers University, Aba 623002, China*; Email: 20049607@abtu.edu.cn

Hua Tang − *School of Basic Medical Sciences, Southwest Medical University, Luzhou 646000, China; Central Nervous System Drug Key Laboratory of Sichuan Province, Luzhou 646000, China*; Email: huatang@swmu.edu.cn

Authors

Hong-Qi Zhang − *School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China*

Shang-Hua Liu − *School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China*

Rui Li − *School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China*

Jun-Wen Yu − *School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China*

Dong-Xin Ye − *School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China*

Shi-Shi Yuan − *School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c09587

Author Contributions

H.-Q.Z.: model design, model training, project management, visualization, and drafting of the original article. S.-H.L.: background research, data organization, model validation, and drafting of the original article. R.L.: figure preparation. J.-W.Y.: model validation. D.-X.Y.: validation and software. S.-S.Y.: validation and software. H.L.: funding acquisition, supervision, and writing—review and editing. C.-B.H.: writing—review and editing. H.T.: funding acquisition, supervision, and writing—review and editing.

Author Contributions

[⊥]H.-Q.Z. and S.-H.L. contributed equally to this work.

Funding

This work was supported by the National Natural Science Foundation of China (62250028, 62172343) and the Sichuan Science and Technology Program (Grant No. 2022YFS0614).

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Liu, D.; Li, G.; Zuo, Y. Function determinants of TET proteins: the arrangements of sequence motifs with specific codes. *Briefings Bioinf.* **2019**, *20* (5), 1826−1835, DOI: 10.1093/bib/bby053.

(2) Yuan, Q.; Chen, S.; Wang, Y.; Zhao, H.; Yang, Y. Alignment-free metal ion-binding site prediction from protein sequence through pretrained language model and multi-task learning. *Briefings Bioinf.* **2022**, *23* (6), No. bbac444, DOI: 10.1093/bib/bbac444.

(3) Jiang, Y.; Wang, R.; Feng, J.; Jin, J.; Liang, S.; Li, Z.; Yu, Y.; Ma, A.; Su, R.; Zou, Q. Explainable deep hypergraph learning modeling the peptide secondary structure prediction. *Adv. Sci.* **2023**, *10* (11), No. 2206151, DOI: 10.1002/advs.202206151.

(4) Ryan, A.; Nevitt, S. J.; Tuohy, O.; Cook, P. Biomarkers for diagnosis of Wilson's disease. *Cochrane Database Syst. Rev.* **2019**, *2019*, No. CD012267, DOI: 10.1002/14651858.CD012267.pub2.

(5) Doguer, C.; Ha, J. H.; Collins, J. F. Intersection of Iron and Copper Metabolism in the Mammalian Intestine and Liver. *Compr. Physiol.* **2018**, *8* (4), 1433−1461.

(6) Yang, K.; Gao, L.; Hao, H.; Yu, L. Identification of a novel gene signature for the prognosis of sepsis. *Comput. Biol. Med.* **2023**, *159*, No. 106958, DOI: 10.1016/j.compbiomed.2023.106958.

(7) Cao, C.; Wang, J.; Kwok, D.; Cui, F.; Zhang, Z.; Zhao, D.; Li, M. J.; Zou, Q. webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. *Nucleic Acids Res.* **2022**, *50* (D1), D1123−D1130.

(8) Zeng, X.; Xiang, H.; Yu, L.; Wang, J.; Li, K.; Nussinov, R.; Cheng, F. J. N. M. I. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nat. Mach. Intell.* **2022**, *4* (11), 1004−1016.
(9) Lu, C. H.; Lin, Y. F.; Lin, J. J.; Yu, C. S. Prediction of metal ion-binding sites in proteins using the fragment transformation method. *PLoS One* **2012**, *7* (6), No. e39252.
(10) Cao, X. Y.; Hu, X. Z.; Zhang, X. J.; Gao, S. J.; Ding, C. J.; Feng, Y. G.; Bao, W. H. Identification of metal ion binding sites based on amino acid sequences. *PLoS One* **2017**, *12* (8), No. e0183756. ARTN.
(11) Liu, S.; Liang, Y.; Li, J.; Yang, S.; Liu, M.; Liu, C.; Yang, D.; Zuo, Y. Integrating reduced amino acid composition into PSSM for improving copper ion-binding protein prediction. *Int. J. Biol. Macromol.* **2023**, *244*, No. 124993.
(12) Valasatava, Y.; Rosato, A.; Banci, L.; Andreini, C. MetalPredator: a web server to predict iron-sulfur cluster binding proteomes. *Bioinformatics* **2016**, *32* (18), 2850−2852.
(13) Wang, R.; Jiang, Y.; Jin, J.; Yin, C.; Yu, H.; Wang, F.; Feng, J.; Su, R.; Nakai, K.; Zou, Q. DeepBIO: an automated and interpretable deep-learning platform for high-throughput biological sequence prediction, functional annotation and visualization analysis. *Nucleic Acids Res.* **2023**, *51* (7), 3017−3029, DOI: 10.1093/nar/gkad055.
(14) Levy, R.; Edelman, M.; Sobolev, V. Prediction of 3D metal binding sites from translated gene sequences based on remote-homology templates. *Proteins* **2009**, *76* (2), 365−374.
(15) Liu, R.; Hu, J. HemeBIND: a novel method for heme binding residue prediction by combining structural and sequence information. *BMC Bioinf.* **2011**, *12*, 207.
(16) You, X.; Hu, X.; Feng, Z.; Wang, Z.; Hao, S.; Yang, C. Recognizing protein-metal ion ligands binding residues by random forest algorithm with adding orthogonal properties. *Comput. Biol. Chem.* **2022**, *98*, No. 107693.
(17) Cheng, X.; Xiao, X.; Wu, Z. C.; Wang, P.; Lin, W. Z. Swfoldrate: predicting protein folding rates from amino acid sequence with sliding window method. *Proteins* **2013**, *81* (1), 140−148.
(18) Li, J.; He, X.; Gao, S.; Liang, Y.; Qi, Z.; Xi, Q.; Zuo, Y.; Xing, Y. The Metal-binding Protein Atlas (MbPA): An Integrated Database for Curating Metalloproteins in All Aspects. *J. Mol. Biol.* **2023**, *435*, No. 168117, DOI: 10.1016/j.jmb.2023.168117.
(19) Guo, G.; Li, S.; Liu, Y.; Cao, Z.; Deng, Y. Prediction of Cavity Length Using an Interpretable Ensemble Learning Approach. *Int. J. Environ. Res. Public Health* **2023**, *20* (1), No. 702, DOI: 10.3390/ijerph20010702.
(20) Ao, C.; Jiao, S.; Wang, Y.; Yu, L.; Zou, Q. Biological Sequence Classification: A Review on Data and General Methods. *Research* **2022**, *2022*, No. 0011, DOI: 10.34133/research.0011.
(21) Su, Q. S.; Wang, F. S.; Chen, D.; Chen, G.; Li, C.; Wei, L. Y. Deep convolutional neural networks with ensemble learning and transfer learning for automated detection of gastrointestinal diseases. *Comput. Biol. Med.* **2022**, *150*, No. 106054, DOI: 10.1016/j.compbiomed.2022.106054.
(22) Sinha, D.; Dasmandal, T.; Yeasin, M.; Mishra, D. C.; Rai, A.; Archak, S. EpiSemble: A Novel Ensemble-based Machine-learning Framework for Prediction of DNA N6-methyladenine Sites Using Hybrid Features Selection Approach for Crops. *Curr. Bioinf.* **2023**, *18* (7), 587−597.
(23) Zhang, H. Y.; Zou, Q.; Ju, Y.; Song, C. G.; Chen, D. Distance-based Support Vector Machine to Predict DNA N6-methyladenine Modification. *Curr. Bioinf.* **2022**, *17* (5), 473−482.
(24) Wang, Y.; Zhai, Y.; Ding, Y.; Zou, Q. SBSM-Pro: Support Bio-sequence Machine for Proteins, arXiv:2308.10275, *arXiv preprint* 2023.
(25) Manavalan, B.; Shin, T. H.; Lee, G. PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. *Front. Microbiol.* **2018**, *9*, No. 476, DOI: 10.3389/fmicb.2018.00476.
(26) Zhu, W.; Yuan, S. S.; Li, J.; Huang, C. B.; Lin, H.; Liao, B. A First Computational Frame for Recognizing Heparin-Binding Protein. *Diagnostics* **2023**, *13* (14), No. 2465, DOI: 10.3390/diagnostics13142465.
(27) Wang, M. J.; Liang, Y. Q.; Hu, Z. Y.; Chen, S. Y.; Shi, B. B.; Heidari, A. A.; Zhang, Q.; Chen, H. L.; Chen, X. W. Lupus nephritis diagnosis using enhanced moth flame algorithm with support vector machines. *Comput. Biol. Med.* **2022**, *145*, No. 105435, DOI: 10.1016/j.compbiomed.2022.105435.
(28) Zhou, H. H.; Wang, H.; Ding, Y. J.; Tang, J. J. Multivariate Information Fusion for Identifying Antifungal Peptides with Hilbert-Schmidt Independence Criterion. *Curr. Bioinf.* **2022**, *17* (1), 89−100.
(29) Desai, A.; Zumbo, A.; Giordano, M.; Morandini, P.; Laino, M. E.; Azzolini, E.; Fabbri, A.; Marcheselli, S.; Giotta Lucifero, A.; Luzzi, S.; et al. Word2vec Word Embedding-Based Artificial Intelligence Model in the Triage of Patients with Suspected Diagnosis of Major Ischemic Stroke: A Feasibility Study. *Int. J. Environ. Res. Public Health* **2022**, *19* (22), No. 15295, DOI: 10.3390/ijerph192215295.
(30) Ao, C.; Ye, X.; Sakurai, T.; Zou, Q.; Yu, L. m5U-SVM: identification of RNA 5-methyluridine modification sites based on multi-view features of physicochemical features and distributed representation. *BMC Biol.* **2023**, *21* (1), No. 93, DOI: 10.1186/s12915-023-01596-0.
(31) Li, H.; Liu, B. BioSeq-Diabolo: Biological sequence similarity analysis using Diabolo. *PLoS Comput. Biol.* **2023**, *19* (6), No. e1011214.
(32) Li, H.; Pang, Y.; Liu, B. BioSeq-BLM: a platform for analyzing DNA, RNA, and protein sequences based on biological language models. *Nucleic Acids Res.* **2021**, *49* (22), No. e129.
(33) Zeng, X.; Wang, F.; Luo, Y.; Kang, S.-G.; Tang, J.; Lightstone, F. C.; Fang, E. F.; Cornell, W.; Nussinov, R.; Cheng, F. Deep generative molecular design reshapes drug discovery. *Cell Rep. Med.* **2022**, *4*, No. 100794, DOI: 10.1016/j.xcrm.2022.100794.
(34) Tsukiyama, S.; Hasan, M. M.; Fujii, S.; Kurata, H. LSTM-PHV: prediction of human-virus protein-protein interactions by LSTM with word2vec. *Briefings Bioinf.* **2021**, *22* (6), No. bbab228, DOI: 10.1093/bib/bbab228.
(35) Sharma, R.; Shrivastava, S.; Kumar Singh, S.; Kumar, A.; Saxena, S.; Kumar Singh, R. Deep-ABPpred: identifying antibacterial peptides in protein sequences using bidirectional LSTM with word2vec. *Briefings Bioinf.* **2021**, *22* (5), No. bbab065, DOI: 10.1093/bib/bbab065.
(36) Yang, H. J.; Gao, Y. L.; Kong, X. Z.; Liu, J. X. Tensor Decomposition Based on Global Features and Sparse Structure for Analyzing Cancer Multiomics Data. *Curr. Bioinf.* **2022**, *17* (10), 946−957.
(37) Lavrova, A. I.; Postnikov, E. B. An Improved Diagnostic of the Drug Resistance Status by Applying a Decision Tree to Probabilities Assigned by the CatBoost Multiclassifier of Matrix Metalloproteinases Biomarkers. *Diagnostics* **2022**, *12* (11), 2847. ARTN.
(38) Rufo, D. D.; Debelee, T. G.; Ibenthal, A.; Negera, W. G. Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM). *Diagnostics* **2021**, *11* (9), 1714. ARTN.
(39) Yang, H.; Luo, Y. M.; Ma, C. Y.; Zhang, T. Y.; Zhou, T.; Ren, X. L.; He, X. L.; Deng, K. J.; Yan, D.; Tang, H.; et al. A gender specific risk assessment of coronary heart disease based on physical examination data. *NPJ Digital Med.* **2023**, *6* (1), 136.
(40) Liu, M. C.; Guo, C. H.; Guo, S. J. An explainable knowledge distillation method with XGBoost for ICU mortality prediction. *Comput. Biol. Med.* **2023**, *152*, No. 106466, DOI: 10.1016/j.compbiomed.2022.106466.
(41) Liu, B.; Gao, X.; Zhang, H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* **2019**, *47* (20), No. e127.
(42) UniProt, C. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D523−D531.
(43) 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* (1), 235−242.

https://doi.org/10.1021/acsomega.3c09587
ACS Omega 2024, 9, 8439−8447

(44) Abdelhafez, O. H.; Othman, E. M.; Fahim, J. R.; Desoukey, S. Y.; Pimentel-Elardo, S. M.; Nodwell, J. R.; Schirmeister, T.; Tawfike, A.; Abdelmohsen, U. R. Metabolomics analysis and biological investigation of three Malvaceae plants. *Phytochem. Anal.* **2020**, *31* (2), 204−214.
(45) Zuo, Y.; Liang, P.; Wang, H.; Liang, Y.; Zhou, J.; Li, H. FeatureScML: An Open-source Python Package for the Feature Importance Visualization of Single-Cell Omics with Machine Learning. *Curr. Bioinf.* **2022**, *17* (7), 578−585.
(46) Dao, F. Y.; Lv, H.; Fullwood, M. J.; Lin, H. Accurate Identification of DNA Replication Origin by Fusing Epigenomics and Chromatin Interaction Information. *Research* **2022**, *2022*, No. 9780293.
(47) Charoenkwan, P.; Chiangjong, W.; Nantasenamat, C.; Hasan, M. M.; Manavalan, B.; Shoombuatong, W. StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides. *Briefings Bioinf.* **2021**, *22* (6), No. bbab172, DOI: 10.1093/bib/bbab172.
(48) Basith, S.; Lee, G.; Manavalan, B. STALLION: a stacking-based ensemble learning framework for prokaryotic lysine acetylation site prediction. *Briefings Bioinf.* **2022**, *23* (1), No. bbab376, DOI: 10.1093/bib/bbab376.
(49) Yu, L.; Yang, K.; He, X.; Li, M.; Gao, L.; Zha, Y. Repositioning linifanib as a potent anti-necroptosis agent for sepsis. *Cell Death Discovery* **2023**, *9* (1), 57.
(50) Hasan, M. M.; Tsukiyama, S.; Cho, J. Y.; Kurata, H.; Alam, M. A.; Liu, X.; Manavalan, B.; Deng, H. W. Deepm5C: A deep-learning-based hybrid framework for identifying human RNA N5-methylcytosine sites using a stacking strategy. *Mol. Ther.* **2022**, *30*, 2856.
(51) Jeon, Y. J.; Hasan, M. M.; Park, H. W.; Lee, K. W.; Manavalan, B. TACOS: a novel approach for accurate prediction of cell-specific long noncoding RNAs subcellular localization. *Briefings Bioinf.* **2022**, *23* (4), No. bbac243, DOI: 10.1093/bib/bbac243.
(52) Ren, S.; Yu, L.; Gao, L. Multidrug representation learning based on pretraining model and molecular graph for drug interaction and combination prediction. *Bioinformatics* **2022**, *38* (18), 4387−4394.
(53) Wang, Y.; Zhai, Y.; Ding, Y.; Zou, Q. SBSM-Pro: Support Bio-sequence Machine for Proteins, arXiv preprint arXiv:2308.10275. 2023 DOI: 10.48550/arXiv.2308.10275.
(54) Yan, K.; Lv, H.; Guo, Y.; Peng, W.; Liu, B. sAMPpred-GAT: Prediction of Antimicrobial Peptide by Graph Attention Network and Predicted Peptide Structure. *Bioinformatics* **2023**, *39* (1), No. btac715, DOI: 10.1093/bioinformatics/btac715.
(55) Kurata, H.; Tsukiyama, S.; Manavalan, B. iACVP: markedly enhanced identification of anti-coronavirus peptides using a dataset-specific word2vec model. *Briefings Bioinf.* **2022**, *23* (4), No. bbac265, DOI: 10.1093/bib/bbac265.
(56) Manavalan, B.; Patra, M. C. MLCPP 2.0: An Updated Cell-penetrating Peptides and Their Uptake Efficiency Predictor. *J. Mol. Biol.* **2022**, *434* (11), No. 167604.
(57) Shoombuatong, W.; Basith, S.; Pitti, T.; Lee, G.; Manavalan, B. THRONE: A New Approach for Accurate Prediction of Human RNA N7-Methylguanosine Sites. *J. Mol. Biol.* **2022**, *434* (11), No. 167549.
(58) Thi Phan, L.; Woo Park, H.; Pitti, T.; Madhavan, T.; Jeon, Y. J.; Manavalan, B. MLACP 2.0: An updated machine learning tool for anticancer peptide prediction. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 4473−4480.
(59) Bupi, N.; Sangaraju, V. K.; Phan, L. T.; Lal, A.; Vo, T. T. B.; Ho, P. T.; Qureshi, M. A.; Tabassum, M.; Lee, S.; Manavalan, B. An Effective Integrated Machine Learning Framework for Identifying Severity of Tomato Yellow Leaf Curl Virus and Their Experimental Validation. *Research* **2023**, *6*, No. 0016, DOI: 10.34133/research.0016.
(60) Pan, X.; Lin, X.; Cao, D.; Zeng, X.; Yu, P. S.; He, L.; Nussinov, R.; Cheng, F. J. Deep learning for drug repurposing: Methods, databases, and applications. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2022**, *12*, No. e1597, DOI: 10.1002/wcms.1597.
(61) Xu, J.; Xu, J.; Meng, Y.; Lu, C.; Cai, L.; Zeng, X.; Nussinov, R.; Cheng, F. Graph embedding and Gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data. *Cells Rep. Methods* **2023**, *3*, No. 100382, DOI: 10.1016/j.crmeth.2022.100382.
(62) Tang, Y.; Pang, Y.; Liu, B. IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning. *Bioinformatics* **2021**, *36* (21), 5177−5186.
(63) Chen, L.; Yu, L.; Gao, L. Potent antibiotic design via guided search from antibacterial activity evaluations. *Bioinformatics* **2023**, *39* (2), No. btad059, DOI: 10.1093/bioinformatics/btad059.
(64) Song, B.; Luo, X.; Luo, X.; Liu, Y.; Niu, Z.; Zeng, X. J. Learning spatial structures of proteins improves protein−protein interaction prediction. *Briefings Bioinf.* **2022**, *23* (2), No. bbab558, DOI: 10.1093/bib/bbab558.
(65) Zeng, X.; Tu, X.; Liu, Y.; Fu, X.; Su, Y. J. Toward better drug discovery with knowledge graph. *Curr. Opin. Struct. Biol.* **2022**, *72*, 114−126.
(66) Ni, P.; Moe, J.; Su, Z. J. Accurate prediction of functional states of cis-regulatory modules reveals common epigenetic rules in humans and mice. *BMC Biol.* **2022**, *20* (1), No. 221, DOI: 10.1186/s12915-022-01426-9.
(67) Chen, S.; Zhou, L.; Song, Y.; Xu, Q.; Wang, P.; Wang, K.; Ge, Y.; Janies, D. J. J. o. M. I. R. A novel machine learning framework for comparison of viral COVID-19−Related Sina Weibo and Twitter Posts: Workflow Development and Content Analysis. *J. Med. Int. Res.* **2021**, *23* (1), No. e24889.
(68) Li, Y.; Ni, P.; Zhang, S.; Li, G.; Su, Z. J. B. ProSampler: an ultrafast and accurate motif finder in large ChIP-seq datasets for combinatory motif discovery. *Bioinformatics* **2019**, *35* (22), 4632−4639.
(69) John, C.; Sahoo, J.; Madhavan, M.; Mathew, O. K. Convolutional Neural Networks: A Promising Deep Learning Architecture for Biological Sequence Analysis. *Curr. Bioinf.* **2023**, *18* (7), 537−558.
(70) Wu, S.; Yang, S.; Wang, M.; Song, N.; Feng, J.; Wu, H.; Yang, A.; Liu, C.; Li, Y.; Guo, F.; et al. Quorum sensing-based interactions among drugs, microbes, and diseases. *Sci. China: Life Sci.* **2023**, *66* (1), 137−151.

https://doi.org/10.1021/acsomega.3c09587
ACS Omega 2024, 9, 8439−8447