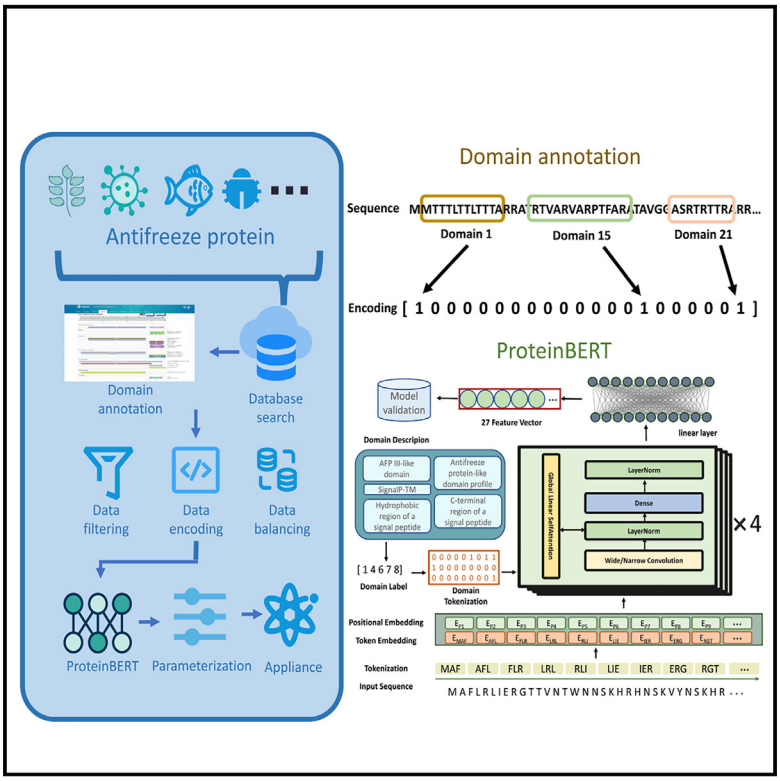
iScience

Article

BERT-DomainAFP: Antifreeze protein recognition and classification model based on BERT and structural domain annotation

Graphical abstract



Highlights

- BERT-DomainAFP achieves 98.48% accuracy, outperforming existing AFP methods
- Introduces domain annotations, enabling classification of most AFP types
- Pre-trained ProteinBERT reduces resources and accelerates model development
- Oversampling and undersampling improve performance on imbalanced datasets

Authors

Shengzhen Chen, Ping Zheng, Lele Zheng, ..., Longshan Lin, Xinhua Chen, Ruoyu Liu

Correspondence

chenxinhua@tio.org.cn (X.C.), liuruoyu13@mails.ucas.ac.cn (R.L.)

In brief

Molecular modelling; Protein; Bioinformatics

Chen et al., 2025, iScience 28, 112077
April 18, 2025
https://doi.org/10.1016/j.isci.2025.112077

CellPress

iScience

CellPress
OPEN ACCESS

Article

BERT-DomainAFP: Antifreeze protein recognition and classification model based on BERT and structural domain annotation

Shengzhen Chen,[1] Ping Zheng,[1] Lele Zheng,[1] Qinglong Yao,[1] Ziyu Meng,[1] Longshan Lin,[2] Xinhua Chen,[1,*] and Ruoyu Liu[1,3,*]

[1]State Key Laboratory of Mariculture Breeding, Key Laboratory of Marine Biotechnology of Fujian Province, Institute of Oceanology, College of Marine Sciences, Haixia Institute of Science and Technology, Fujian Agriculture and Forestry University, Fuzhou 350002, China
[2]Laboratory of Marine Biodiversity Research, Third Institute of Oceanography, Ministry of Natural Resources, Xiamen 361005, China
[3]Lead contact
*Correspondence: chenxinhua@tio.org.cn (X.C.), liuruoyu13@mails.ucas.ac.cn (R.L.)
https://doi.org/10.1016/j.isci.2025.112077

SUMMARY

Antifreeze proteins (AFPs) are crucial for organisms to adapt to low temperatures, with applications in medicine, food storage, aquaculture, and agriculture. Accurate AFP identification is challenging due to structural and sequence diversity. To improve prediction and classification, we propose BERT-DomainAFP, a deep learning model trained on the AntiFreezeDomains dataset created with a novel annotation strategy. The model uses pre-trained ProteinBERT and incorporates oversampling and undersampling techniques to handle unbalanced data, ensuring high predictive ability. BERT-DomainAFP achieves 98.48% accuracy, the highest among existing models, and can classify different AFP types based on structural domain features. This model outperforms current tools, offering a promising solution for AFP recognition and classification in research and applications.

INTRODUCTION

Antifreeze proteins (AFPs) are critical for enhancing the cold adaptation capabilities of organisms, which are mainly found in fish, plants, insects, and sea ice organisms living in extreme cold regions.[1] These proteins prevent cold-induced damage by lowering the freezing point, altering ice crystal morphology, and inhibiting ice crystal growth, thereby increasing an organism's ability to survive in low temperatures.[2–4] In recent years, AFPs have found widespread applications in medicine, frozen food storage and processing, aquaculture, and agriculture, showcasing significant potential for application and economic value.[1,5] Research on AFPs initially focused on fish and insects before expanding to microorganisms, animals, and plants. To date, five types of structurally distinct AFPs have been identified in fish: antifreeze glycoproteins (AFGPs), type I AFP (AFP I), type II AFP (AFP II), type III AFP (AFP III), and type IV AFP (AFP IV).[6–12] AFGPs are primarily composed of repetitive tripeptide units of alanine-alanine-threonine.[13] AFP I lacks glycosylation and tripeptide repeats, consisting of repeated polypeptide units with 11 amino acid residues, and features the smallest molecular weight and the simplest structure.[14] AFP II has the largest molecular weight and is homologous to animal C-type lectins.[13] AFP III is structurally unique due to the absence of major nonpolar amino acids and its globular protein form. Additionally, AFP IV is characterized not only by its richness in glutamic acid but also by its high content of glutamine.[5] AFPs in other species are not yet uniformly typed and are mainly classified by specific amino acids in the AFPs. Based on previous studies, the present study classified AFP according to five types from the structural properties of AFP: AFP type I, AFP type II, AFP type III, insect AFP, and AFGP.[15]

Accurately and reliably predicting AFPs is a prerequisite for their investigation and utilization. However, the significant heterogeneity in structure and sequence among different AFPs makes precise identification challenging.[16] Meanwhile, experimental identification of AFPs has long been challenging, primarily due to the limitations of classical methods, such as basic local alignment search tool (BLAST) and hidden Markov models, which struggle to accurately predict highly specific sequences and are computationally expensive. With the recent exponential growth in annotated protein sequences, the task of experimental identification has become even more formidable, as the vast volume of data makes comprehensive analysis using traditional methods increasingly impractical. To address these challenges, researchers have introduced machine learning-based prediction methods for AFP research. Since 2011, fifteen machine learning-based predictors have been developed for AFP identification.[13] For example, Yu et al. (2011) developed the iAFP predictor, which uses tripeptide composition (TPC), amino acid composition (AAC), and dipeptide composition (DPC) for feature extraction, employs genetic algorithms (GA) for key feature selection, and trains the model using a support vector machine (SVM).[17] Kandaswamy et al. (2011) developed the AFP-Pred predictor,

Table 1. Comparison of prediction accuracy between BERT-DomainAFP model and other models on the independent dataset

Methods	Classifier	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC
iAFP[17]	SVM	95.60	97.23	13.2	0.80
AFP-Pred[18]	RF	77.34	77.04	91.16	0.66
AFP-PSSM[19]	SVM	93.01	93.20	75.80	0.34
AFP-PseAAC[20]	SVM	84.75	84.74	88.08	0.80
afpCOOL[46]	SVM	96.00	98.00	72.00	NA
AFP-CKSAAP[29]	DNN	88.00	87.80	94.00	0.33
AFP-CMBPred[22]	SVM	91.65	91.98	75.14	0.85
Wang et al. (2021)[47]	DT	91.33	NA	NA	0.82
AFP-LSE[30]	Autoencoder	88.17	91.52	84.83	0.46
TargetFreeze[48]	SVM	90.95	91.78	90.11	0.81
VotePLMs-AFP[27]	Soft voting	94.20	94.10	96.70	0.95
iAFP-gap-SMOTE[49]	KNN, PNN, SVM	95.02	96.58	93.49	0.9
AFP-XGB[25]	Xgboost	99.5	99.48	99.74	0.97
Feng et al.[26]	Stacking classifier	98.3	96.6	100	0.96
BERT-DomainAFP	BERT	98.48	99.80	99.70	0.95

utilizing secondary structural features and physicochemical information as the feature set, and random forest (RF) as the classifier for model training.[18] Additionally, the AFP-PSSM (2012) method uses evolutionary information and SVM for AFP identification.[19] Mondel et al. (2014) developed the AFP-PseAAC method by extracting features using pseudo amino acid composition (PseAAC) and training the model with SVM.[20] Akbar et al. (2019) developed the ''iAFP-gap-SMOTE'' predictor, employing split AAC, G-gap di-peptide composition, and reduce amino acid alphabet composition for feature extraction, and synthetic minority over-sampling technique (SMOTE) to handle class imbalance.[21] The AFP-CMBPred (2021) model introduces a novel multi-block position-specific scoring matrix (MB-PSSM) approach, coupled with its consensus sequence-based (CS-MB-PSSM) feature descriptors, and uses SVM and RF for model training.[22] Khan et al. (2022) proposed the AFP-LXGB predictor, utilizing DPC, grouped amino acid composition (GAAC), position specific scoring matrix-segmentation-autocorrelation transformation (Sg-PSSM-ACT), and pseudo position specific scoring matrix tri-slicing (PseTS-PSSM) for feature extraction, and light extreme gradient boosting (LXGB) for model training.[23] Khan et al. (2023) also introduced the AFP-SPTS method, exploring features with dipeptide deviation from the expected mean (DDE), reduced amino acid alphabet (RAAA), grouped dipeptide composition (GDPC), and PseTS-PSSM, and training models with a combination of machine learning algorithms.[24] Dhiar et al. (2023) proposed a method combining n-gram feature vectors with machine learning models, with Xgboost as the predictor for identifying potential AFPs from sequences.[25] Feng et al. (2024) proposed a stacking-based classifier for AFP identification, with feature extraction from reduction properties, scalable PseAAC, and physicochemical properties, and training with LightGBM, XGBoost, and random forest algorithm.[26] Qi et al. (2024) developed the VotePLM AFP predictor, based on transformer embedding features and ensemble learning, extracting features from pre-trained protein language models (PLM).[27] Kumar et al. (2024) proposed machine learning models for AFP prediction using composition-based protein features and evolutionary information, and developed the ''AFPropred'' web server for user accessibility.[28]

In addition to the aforementioned models, there are other machine learning-based methods that have shown great promise in predicting antifungal peptides, antioxidant proteins, and antiviral peptides. For instance, the iAFPs-Mv-BiTCN model combines word embeddings and transformation-based features, employs SHAP methodology to reduce training costs, and is trained with BiTCN, achieving high accuracy and area under the ROC curve (AUC) values. The StackedEnC-AOP model integrates wavelet transform, evolutionary difference formulas, and complex rational properties, utilizes mRMR feature selection, and employs a stacked ensemble meta-model to achieve high accuracy and AUC values. The DeepAVP-TPPred model leverages image feature extraction and binary tree growth algorithms to optimize features and then constructs classification models through deep neural networks (DNNs), demonstrating excellent performance and generalization capabilities. The innovation of these models lies in their combination of various feature extraction techniques with advanced machine learning algorithms to enhance the accuracy and efficiency of predictions. Compared to machine learning models based on homology, rules, or probabilistic features, another approach to constructing AFP prediction models involves utilizing the rapidly advancing field of deep learning. Currently, deep learning models have also seen some applications in AFP prediction. The AFP-CKSAAP method leverages k-spaced amino acid pair (CKSAAP) features and DNNs to enhance prediction performance by learning the nonlinear mappings between protein sequence descriptors and class labels.[29] Additionally, the AFP-LSE method employs a latent space encoded deep autoencoder, combining feature compression and dimensionality reduction with a multi-layer perceptron (MLP) classifier to achieve efficient and accurate AFP prediction.[30] Although deep learning algorithms are currently being applied to AFP prediction, they still lag behind feature selection-based machine learning models in terms of prediction accuracy.

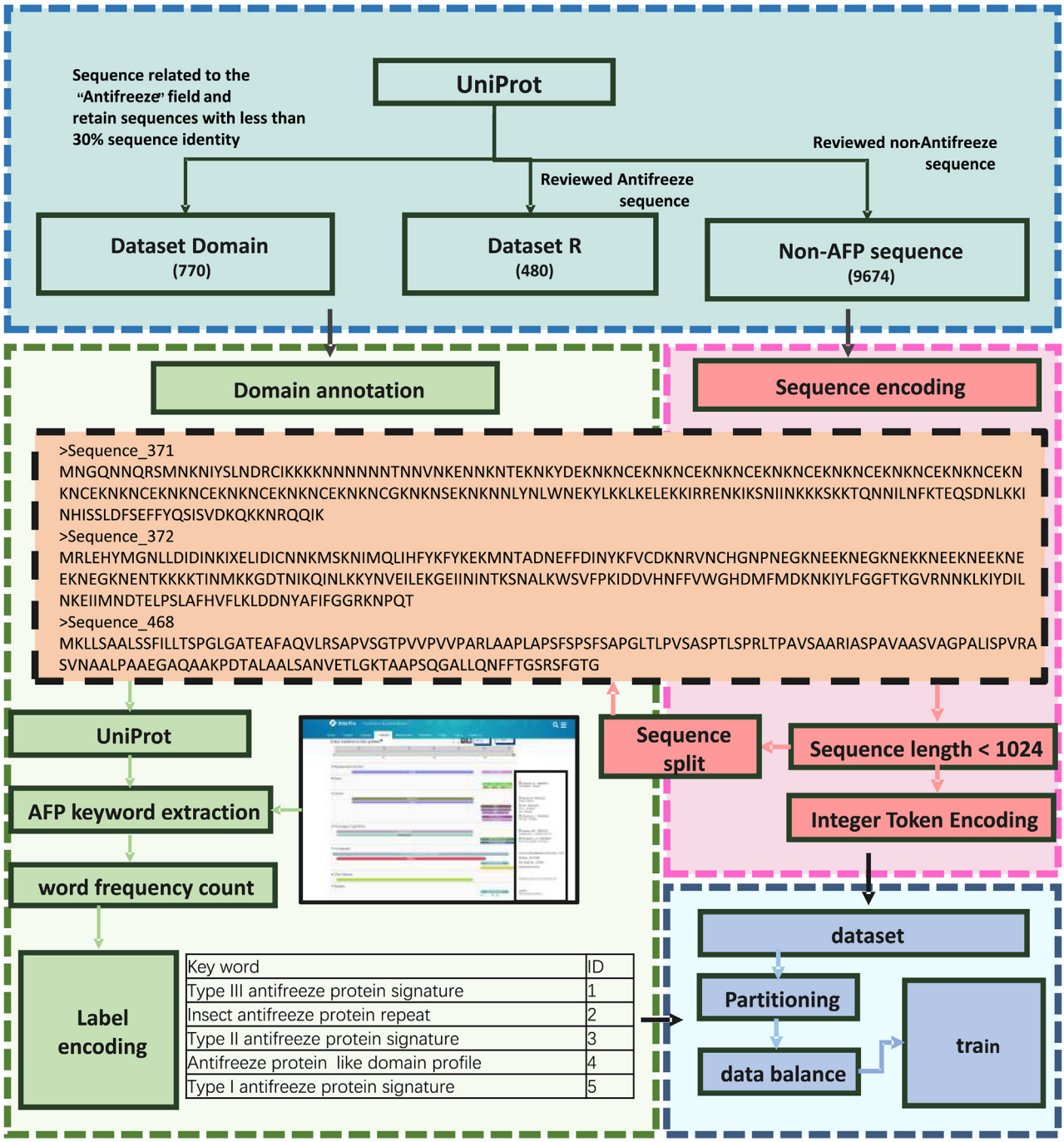


Figure 1. Workflow of the datasets collection and domain coding for AFPs

The process of constructing the positive dataset was first based on the existing DatasetR dataset and combined with sequences of AFPs retrieved from the InterPro database. All sequences were structurally domain annotated by InterPro, yielding a total of 37,970 annotation records. Five major primary AFP structural domain entries and 22 secondary entries closely related to AFP sequences were identified by manual screening. For AFP data longer than 1,024 amino acid residues, they were split into multiple shorter sequences and re-annotated with structural domains.

In the field of AFPs prediction, although existing prediction models have made significant progress in evaluation indexes, there is still room for further optimization in the classification precision and prediction accuracy of AFPs. In previous studies, the annotation of AFPs has often adopted a simplified dichotomous approach,[13] the proteins are classified into two categories, namely "antifreeze proteins" and "non-antifreeze proteins." This classification strategy may not adequately capture the complexity of

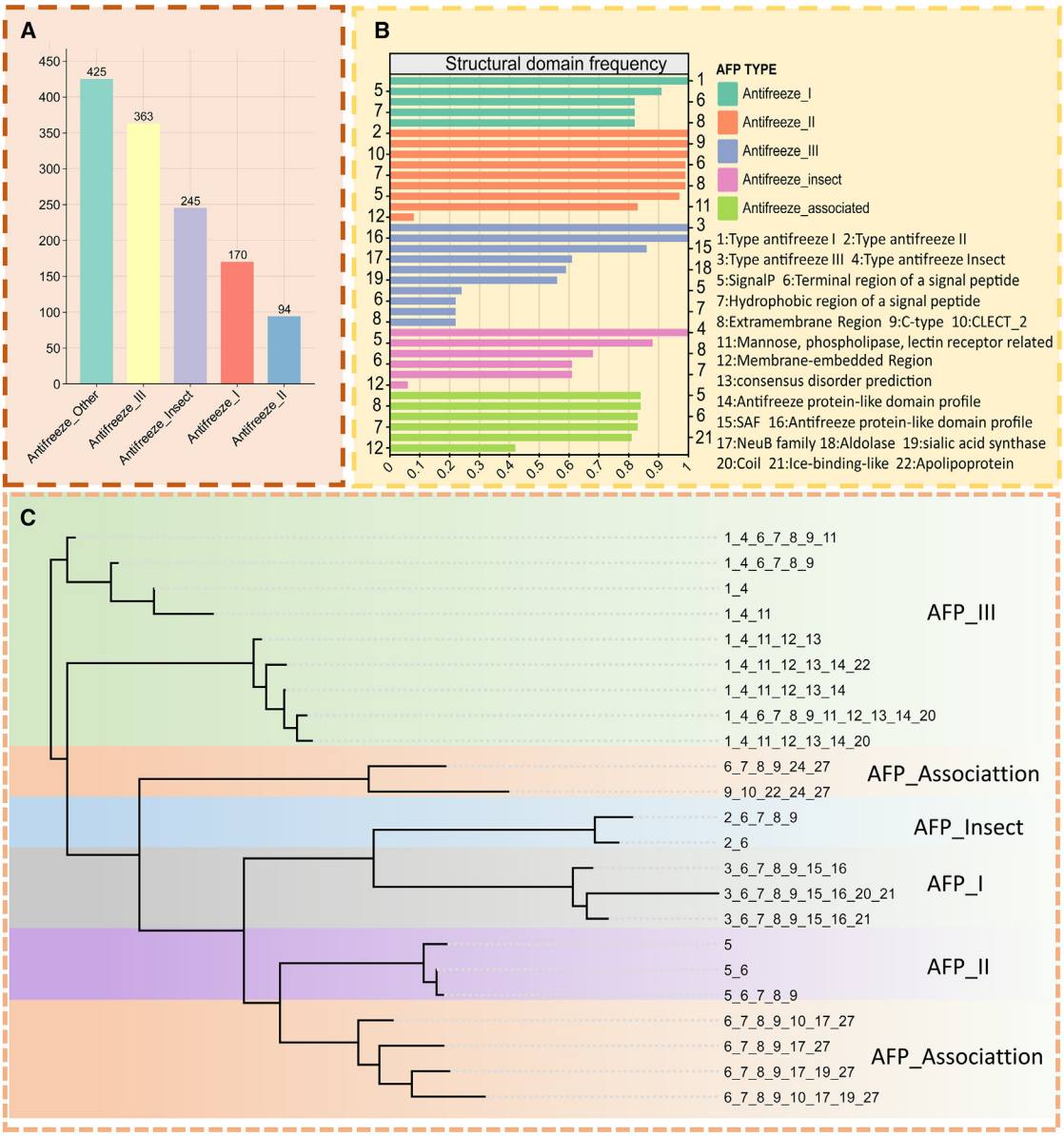


Figure 2. Statistical information on the AFP dataset

(A) Number of AFP sequences in the AFP dataset.

(B) Probability of domain annotation of various AFPs.

(C) Evolutionary tree of multiple types of AFPs containing multiple structural domains. The tree was constructed in MEGA 11 and IQ-TREE 2 software using the maximum likelihood (ML) option and 1,000 bootstrap replicates. Branches with different colored backgrounds represent different AFP types, and each leaf corresponds to its structural domain entry annotated in the structural annotation of the dataset (Table S1).

structurally non-conserved and sparse AFP types when constructing the training dataset, resulting in insufficient learning of these types of AFPs during model training. Furthermore, this simplified classification approach may ignore the diversity and complexity of AFPs, thereby limiting the model's ability to recognize and understand these few types of AFPs. Further, this annotation approach also fails to achieve an effective balance of data on various types of AFPs, which in turn affects the performance of the model on predictive metrics (Table 1), causing it to encounter a developmental bottleneck. There is still a need to seek more effective feature descriptors, feature selection methods, or learning models to develop more accurate predictors. In 2017, the transformer was proposed,[31] introducing self-attention mechanisms and position encoding to achieve global parallel processing of sequences, significantly improving computational efficiency and performance. Since its introduction, the transformer model has been widely applied in the field of biological sequence analysis. In miRNA precursor prediction tasks, transformer-based

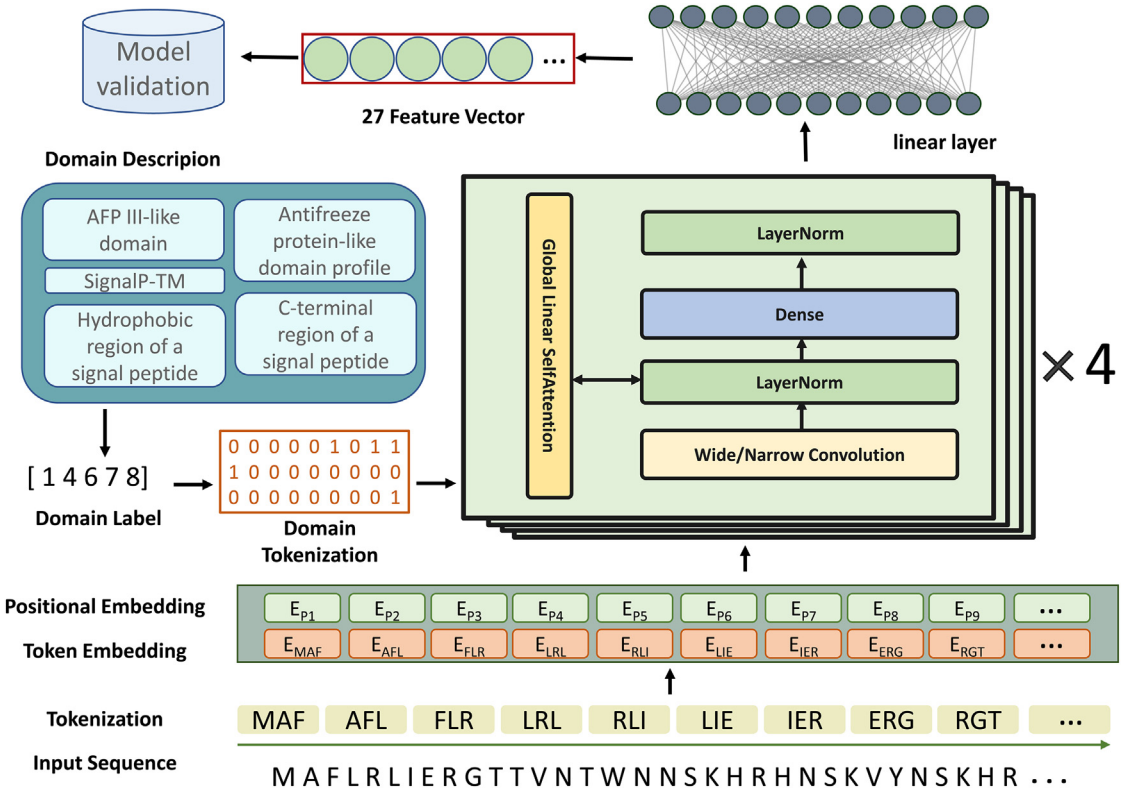


Figure 3. Architecture of the pre-training model used in this study

The ProteinBERT[39] model consists of 6 attention heads and 4 layers. It uses Gelu activation functions, with convolution kernel sizes of 9 and an expansion rate of 5. A linear layer is used for the output, and a manually set classification threshold is applied.

models have achieved an accuracy exceeding 98%, approximately 10% higher than previous models based on deep convolutional neural network (CNN), recurrent neural network (RNN), and long short memory network (LSTM), etc.[32–35] Furthermore, incorporating other technologies and features into deep learning models has significantly improved the accuracy of biological sequence prediction. The DeepCIP model optimizes the prediction of internal ribosome entry sites (IRES) sites within circRNAs through multimodal deep learning techniques.[36] Meanwhile, BERMP enhances its ability to predict m6A sites by combining the bidirectional gated recurrent unit (BGRU) with a RF classifier.[37] Moreover, most existing AFP sequence recognition models focus on binary classification tasks,[38] but there is limited research on classifying different AFP categories. Traditional models are often based on *ab initio* training, which can lead to suboptimal performance and accuracy when training samples are limited. Multiple label classification models for protein sequences, primarily using gene ontology (GO) annotation information, have been studied extensively.[39,40] These models train on large, de-duplicated protein sequence datasets, learning thousands of functional annotations and have already applied to tasks such as *HKT* gene identification in *Spartina alterniflora*.[41] Brandes et al. (2022) developed ProteinBERT based on the transformer framework, which has significant advantages in capturing long-distance dependencies and global information within the sequence and achieved leading accuracy in label training of GO annotations.[39,42]

To improve the prediction performance of AFPs and enable classification among different AFP types, we introduce a novel approach that leverages the transfer learning capabilities of the ProteinBERT model. Firstly, this model integrates the sequence set AntiFreezeDomains, which is characterized by AFP domains, and uses these domain annotations as training labels. Secondly, for handling of the unbalanced datasets, our BERT-DomainAFP model innovates employing oversampling and undersampling techniques. This ensures the model can fully learn the features of various data types during training and maintain high predictive ability even on more unbalanced datasets. Thirdly, to improve the predictive performance and training efficiency of the AFP model, the BERT DomainAFP model uses transfer learning based on the protein BERT model. And to evaluate the prediction performance, we compare it with existing models using binary classification evaluation metrics; results showed that the BERT-DomainAFP model has superior performance over existing models. Most importantly, the BERT-DomainAFP model introduces an innovative application of structural domain annotations as labels. Based on the unique structural domain features, the BERT-DomainAFP model can not only improve the accuracy of predicting the presence of AFP, but also can classify them into specific AFP types. The BERT-DomainAFP model provides a more nuanced understanding of AFP properties and functions, a capability not found in traditional models.

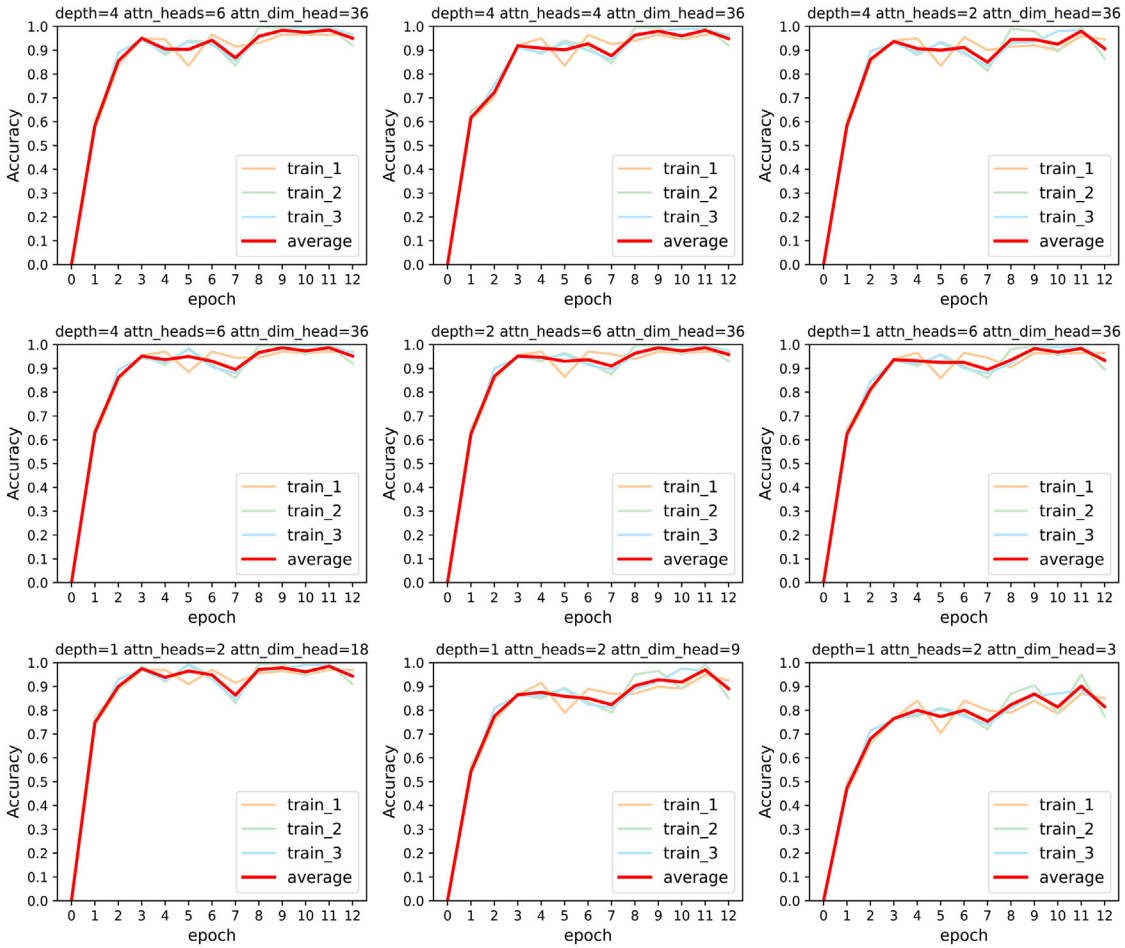


Figure 4. Predictive accuracy of the model for 12 rounds with different parameter configurations of depth, attn_heads, and attn_dim_head

The first row of line charts shows the predictive performance of the model trained with different configurations of attn_heads. The second row displays the predictive performance of the model trained with varying configurations of attn_depth. The third row illustrates the predictive performance of the model trained with different configurations of attn_dim_head.

The main contributions of our work are as follows.

(1) We firstly introduce a novel approach leveraging ProteinBERT's transfer learning for AFP prediction and classification.
(2) Employ oversampling and undersampling techniques to handle unbalanced datasets, ensuring effective learning and high predictive ability.
(3) Use transfer learning based on ProteinBERT to enhance prediction performance and training efficiency.
(4) Most importantly, we introduce an innovative application of structural domain annotations as labels for improved accuracy and specific AFP type classification.

RESULTS

Annotation results of AFP dataset

Through the data collection and preprocessing pipeline (Figure 1), we ultimately obtained the AntiFreezeDomain dataset, which contained 1,297 AFP entries. Additionally, a negative dataset comprising 9,637 non-AFP sequences from UniProtKB was constructed for training, with batches of 100 sequences processed for domain prediction and labeling. Notably, most of the negative data lacked structural domain annotations related to AFP, with negative sequences that did not receive any annotations accounting for ~95% of the total number of negative sequences, and negative sequences that were annotated but did not contain any annotations related to AFPs accounting for ~5% of the total number of negative sequences. We counted the annotated AFP sequences, and among the five AFP types (AFP I, AFP II, AFP III, AFP insect, and AFP other) (Figure 2A), the "AFP Other" category comprised the largest proportion of AFP sequences in the dataset, with 425 sequences. The predominant keyword associated with these proteins is "ice-bound-like." This category mainly includes unclassified antifreeze sequences from biological sources outside of fish and insects, which have not yet been assigned into specific AFP types. Based on this, we further counted the high-frequency annotation terms for each AFP type in Interpro annotations (Figure 2B). Finally, we randomly selected one protein sequence from each annotated structural domain

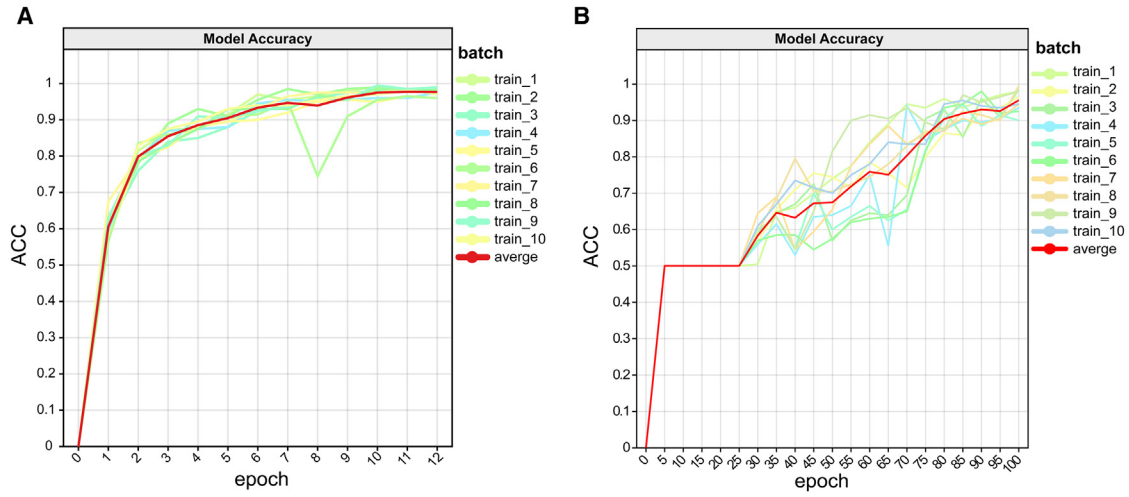


Figure 5. Accuracy evaluation and comparison of model predictions

(A) Model accuracy for the first 12 rounds with pre-trained models established through repeated random subsampling. The dataset was formed using 10 batches of random subsampling.

(B) Model prediction accuracy for the first 100 rounds without pre-training, also based on 10 batches of random subsampling.

type and constructed an evolutionary tree based on the structural domain codes (Figure 2C). The corresponding codes for the structural domains are listed in Table S1.

Pretrain model

We employed ProteinBERT,[39] a protein-specific deep learning model based on the BERT framework, for training. Different from the traditional deep learning network, the core idea of BERT model architecture was to use its attention mechanism to capture the dependencies between input sequences.[31,43] The deep learning model architecture in this experiment was trained on a pre-training model with 6 attention heads and 4 layers (Figure 3). During the training process, the original label was randomly modified with a 5% probability, annotations were randomly added with a 1% probability, and annotation information was randomly added with a 25% probability. To enhance the model's ability to process incomplete data and prevent overfitting, we used Gelu (Gaussian error linear unit) for all hidden layers and linear layers in the model, and the convolution kernel size for wide convolution layer (expansion rate = 5) and narrow convolution layer pairs was 9. Finally, the linear layer was directly used to output the prediction results of the model, and the classification threshold was manually set. During the training process, most of the remaining parameters used the default settings of the pre-training model.[39] Due to the difference between the training task and the training dataset, the model reduced the complexity and the number of layers.

We tested three key parameters in the model architecture—depth, attn_heads, and attn_dim_head—and for each configuration, we selected 3 randomly assigned datasets for training, ultimately obtaining the model's training results (Figure 4). The experiment showed that with attn_dim_head set to 36, most models were able to quickly fit the data, even when depth was reduced to 1, and still achieved high accuracy. However, when we gradually reduced attn_heads and attn_dim_head, the model's accuracy began to drop significantly, with the highest accuracy mean decreasing from 0.97 to 0.90. Therefore, we concluded that as long as the product of attn_heads and attn_dim_head is greater than 18, the model can achieve optimal performance on the current dataset. Furthermore, increasing depth, attn_heads, or attn_dim_head within a certain range (i.e., depth $\leq$ 4, attn_heads $\leq$ 6, attn_dim_head $\leq$ 36) does not cause significant fluctuations in the model's prediction accuracy.

Evaluation of predictive performance of the BERT-DomainAFP model for binary classification tasks

In existing research, researchers often choose pre-trained models on large amounts of data as a method to reduce model training epochs and accelerate model fitting speed.[39,44,45] The average accuracy reached 97% during the 10th round of training on the pre-trained model (Figure 5A). However, in models with the same structure but without pre-training, it took nearly 95 rounds of training to achieve the same accuracy level as the pre-trained model did by the 10th round (Figure 5B). Each model was validated using the test set with the highest accuracy saved during the training process. The final average accuracy obtained on the test set was 98.48%. Since our accuracy was based on the test results from multiple training and testing sets, we confidently concluded that the high accuracy of the model was not due to randomness in the dataset.

Accurately and reliably predicting AFPs is essential for their development and utilization. Many methods for predicting AFPs have already been developed.[13] However, due to the diversity of AFP sequences and structures, different prediction methods show varying performance. We compared the performance of the developed BERT-DomainAFP model with existing AFP prediction methods. Leveraging pre-trained models, our BERT-DomainAFP model achieved leading predictive performance compared to a range of existing models across various evaluation indicators. As shown in Table 1, the BERT-DomainAFP model

Table 2. Performance comparison of the BERT-DomainAFP model with contemporary deep learning-based methods tested on the training dataset (600) and test dataset (9,484)

Dataset	Methods	ACC (%)	Sensitivity (%)	Specificity (%)	MCC
training dataset (600)	AFP-CKAAP	88.01	94.43	87.90	0.33
	AFP-LSE	87.35	82.59	92.86	0.48
	VotePLMs-AFP	94.20	91.00	95.00	0.86
	BERT-DomainAFP	93.00	91.00	95.00	0.86
test dataset (9,484)	AFP-CKAAP	88.00	94.00	87.00	0.32
	AFP-LSE	93.70	86.7	93.90	0.52
	BERT-DomainAFP	94.64	94.56	94.65	0.47

attained the highest levels in four key indicators: accuracy, specificity, sensitivity, and Matthews correlation coefficient (MCC). Specifically, the BERT-DomainAFP model boasted an accuracy rate of 98.48%, significantly surpassing other models and indicating its high precision in distinguishing between AFPs and non-AFPs. Its specificity and sensitivity also reach 99.80% and 99.70%, respectively, demonstrating the model's high accuracy and coverage in identifying AFPs. Moreover, the MCC value of 0.95 further confirmed the classification performance advantage of the BERT-DomainAFP model.

In comparison, while other models performed well in certain aspects, they generally could not match the performance of BERT-DomainAFP. For instance, the SVM model used by iAFP[17] achieved an accuracy rate of 95.60%, but its specificity and sensitivity were both lower than those of BERT-DomainAFP. The RF model used by AFP-Pred[18] showed a notable sensitivity of 91.16%, yet its accuracy and specificity were comparatively low. Although afpCOOL[46] reached a specificity of 98.00%, its sensitivity was only 72.00%, which might imply that it could miss a significant number of positive samples in practical applications. The DNN model of AFP-CKSAAP[29] and the SVM model of AFP-CMBPred[22] had MCC values of only 0.33 and 0.85, respectively, significantly lower compared to the 0.95 of BERT-DomainAFP.

In previous research, two deep learning-based methods had been developed for predicting AFPs. Both studies utilized the same training dataset (consisting of 600 samples) and test dataset (9,484 samples) for model training and performance evaluation. To comprehensively assess the predictive capability of the BERT-DomainAFP model, we compared it with contemporary deep learning methods such as AFP-CKSAAP and AFP-LSE (Table 2). On the training dataset (600), the BERT-DomainAFP model achieved 93.00% accuracy (ACC), 91.00% sensitivity, 95.00% specificity, and an MCC value of 0.86. These results indicated that the BERT-DomainAFP model outperformed the AFP-CKAAP and AFP-LSE models on the training set, especially in terms of specificity and MCC values, demonstrating higher classification accuracy and better balance. On a larger test dataset (9,484), despite the MCC value of the BERT-DomainAFP model being 0.47, slightly lower than the AFP-LSE model (0.52), it was still significantly higher than the AFP-CKSAAP model (0.32). Moreover, the BERT-DomainAFP model achieved the highest scores among the three deep learning-based methods in the other three indicators, with an accuracy of 94.64%, sensitivity of 94.56%, and specificity of 94.65%. These results indicated that the BERT-DomainAFP model still performed well on the larger test dataset.

Assessment of the BERT-DomainAFP for AFP classification capability

Different types of AFPs have varying function mechanisms,[50,51] while existing AFP prediction methods primarily determined whether a protein was an AFP.[38] To provide more specific information for future investigations, our proposed BERT-DomainAFP model not only predicted AFPs but also attempted to classify different types of AFPs. The classification performance of BERT-DomainAFP model for AFPs was shown in Table 3, focusing on differentiating them from non-AFPs across two training sample ratios: 1:1 and 1:5. Notably, the model maintained high accuracy (ACC) and sensitivity (Recall) across all AFP types, particularly excelling with type AFP II and type AFP III. Specificity remained robust, indicating the model's proficiency in correctly identifying non-AFPs. The MCC and F1-Score were also high, suggesting a strong balance between precision and recall. These results underscored the model's reliability in accurately classifying AFPs, even when the training data were skewed toward non-AFPs.

We also evaluated the prediction accuracy of each AFP type and plotted the receiver operating characteristic curve (ROC) for each AFP type prediction, which was a comprehensive indicator of the sensitivity and specificity of the model to continuous variables for each AFP type (Figure 6A). The higher the value of the AUC, the better the classification ability of the prediction model. The average accuracy of the model in identifying each type of AFP was shown in Figure 6B. In the evaluation of the BERT-Domain AFP model, the accuracy of each prediction category demonstrated the model's ability to classify NON-AFPs and AFPs with different training sample ratios. At the ratio of NON-AFP to AFP of 1:1, the accuracy of non-AFP reached 98.45%, while the accuracy of different AFP types varied, with 97.5% for AFP I, 99.5% for AFP II, 99.8% for AFP III, 90.3% for AFP insect, and 95.8% for AFP associate. In the evaluation of the BERT-DomainAFP model in the face of an unbalanced dataset, we observed that the model's classification accuracy for non-AFP samples increased from 98.45% to 99.025% when the ratio of non-AFP to AFP training samples was adjusted from 1:1 to 1:5, showing the model's improved accuracy when dealing with non-AFP samples. For different types of AFPs, the change in prediction accuracy varies: the accuracy of AFP II and AFP III remains high at both sample scales, at 100% respectively, indicating that these types of AFPs have better stability and prediction accuracy in the model. However, the accuracy of AFP I decreased from 97.5% to 95%, as did the accuracy of AFP insect from 90.3% to 87.5%, and the accuracy of AFP associate from 95.8% to 92.4%.

DISCUSSION

Up to now, multiple machine learning algorithms have been developed for predicting AFPs, including two deep learning

Table 3. Model classification ability evaluation of the BERT-DomainAFP model (data are represented as mean ± SEM)

Training samples ratios	AFP Type	ACC	Sensitivity (recall)	Specifity	MCC	Precision	F1-Score
NON-AFP:AFP 1:1	NON-AFP	0.985 ± 0.006	0.998 ± 0.001	0.971 ± 0.021	0.969 ± 0.027	0.972 ± 0.008	0.985 ± 0.009
	AFP I		0.975 ± 0.02	0.986 ± 0.011	0.919 ± 0.054	0.882 ± 0.071	0.878 ± 0.088
	AFP II		0.995 ± 0.001	0.983 ± 0.009	0.922 ± 0.037	0.869 ± 0.105	0.874 ± 0.065
	AFP III		0.998 ± 0.001	0.983 ± 0.004	0.922 ± 0.061	0.867 ± 0.098	0.874 ± 0.093
	AFP insect		0.903 ± 0.077	0.994 ± 0.002	0.913 ± 0.064	0.94 ± 0.013	0.895 ± 0.029
	AFP associate		0.958 ± 0.009	0.984 ± 0.011	0.92 ± 0.036	0.876 ± 0.098	0.876 ± 0.023
NON-AFP:AFP 1:5	NON-AFP	0.990 ± 0.001	0.997 ± 0.002	0.957 ± 0.012	0.965 ± 0.019	0.991 ± 0.004	0.976 ± 0.01
	AFP I		0.95 ± 0.023	0.992 ± 0.003	0.865 ± 0.092	0.797 ± 0.042	0.97 ± 0.01
	AFP II		1.000 ± 0.000	0.99 ± 0.009	0.875 ± 0.09	0.774 ± 0.116	0.995 ± 0.002
	AFP III		1.000 ± 0.000	0.99 ± 0.007	0.875 ± 0.112	0.774 ± 0.068	0.995 ± 0.001
	AFP insect		0.875 ± 0.062	0.994 ± 0.005	0.852 ± 0.069	0.84 ± 0.136	0.931 ± 0.053
	AFP associate		0.924 ± 0.051	0.991 ± 0.006	0.867 ± 0.082	0.793 ± 0.028	0.975 ± 0.014

methods.[38] Although deep learning algorithms can avoid manual feature selection and demonstrate certain advantages in predicting AFP sequences, current practice shows that the prediction accuracy of these deep learning algorithms is not as good as some machine learning algorithms based on artificial feature selection. Research has found that deep learning models have difficulty balancing the learning of minority class samples in AFP sequences when processing existing datasets, leading to a decline in their predictive performance.[30,38] To solve this problem, we constructed a dataset named AntiFreezeDomains, adopting a collection method similar to that of current mainstream datasets (Figure 1). Notably, while AntiFreezeDomains nearly fully encompasses the existing AFP sequence datasets, DatasetS and DatasetR,[38] a novel strategy was employed for data filtering and label annotation. In previous models for predicting AFPs, researchers often treated the prediction task as a binary classification task, ignoring the sequence and quantity differences between different types of AFPs, which often made it difficult for the model to learn features from a few categories. We drew inspiration from the GO annotation task, where each protein sequence corresponds to an average of 2–3 pieces of annotation information.[39,52] Similarly, a protein sequence typically corresponds to multiple structural domains. Therefore, we conducted as comprehensive as possible structural domain annotation for each AFP through the InterPro website. This detailed annotation approach makes AntiFreezeDomains the most comprehensively annotated dataset available. The protein 3D and structural domain annotations (Figure 7) for AFPs reflect the diversity of AFP species and structural domains, allowing us to differentiate between a few categories and balance the data in the next step.

To address the challenge of data imbalance in predicting AFP domains, we first employed techniques such as multiple undersampling and over-sampling strategies. These methods ensure that the model can learn the features of all categories more comprehensively during the training process, thereby improving the recognition ability of minority classes and enhancing the overall performance of the model. Based on the existing structural domain prediction models,[53–55] We selected ProteinBERT as the pre-training model, and the BERT model framework is highly suitable for it as a pre-training model.[39] Its introduction of self-attention mechanism makes it a leader in current long text machine learning tasks.[42] ProteinBERT is different from other transfer learning biological models. ProteinBERT is currently one of the few multi-label classification models trained on a large number of protein sequences. Through multiple parameter tuning, we have reduced the complexity of the pre-trained ProteinBERT model and modified its output layer to fit our dataset. And by comparing with untrained models with the same structure, we found that pre-trained models can save ~90% of training time, which also indicates that the features in GO annotation tasks are highly similar to those in AFP prediction tasks. Finally, in order to conduct a comprehensive model evaluation, we trained the model on multiple datasets and ultimately achieved BERT DomainAFP.

For model evaluation, we first choose to evaluate our model using binary classification evaluation metrics, which test the model's ability to predict whether the protein sequence is an AFP sequence. Our model achieved the best prediction performance among current models, as shown in Table 1. Through leveraging a large volume of negative datasets and a balanced AFP dataset for training, we maximized the utilization of collected data and enhanced the model's generalization ability. When evaluating our multi-class task, we observed that the prediction accuracy of our trained model for each AFP category in the test set varied significantly. The prediction accuracy for insects and other types of AFPs was relatively low, at 90.3% and 95.8%, respectively, while AFP I, AFP II, and AFP III showed higher prediction accuracy, at 97.5%, 99.5%, and 99.8% (Table 3), respectively. We believe that this result may be related to the conservation level of AFP types. According to existing research, AFP I, AFP II, and AFP III are mainly distributed in fish and exhibit high conservation. In contrast, AFPs found in insects, bacteria, and other species, despite their highly similar functions, display significant diversity in terms of structure and sequence characteristics.[13,15,38] Furthermore, we conducted our research based on pre-trained models. Through comparison with models that have not undergone pre-training (Figure 5), we discovered that training with pre-trained models can significantly improve the efficiency of model training. In addition, in terms of performance, the model can annotate 16–32 sequences per

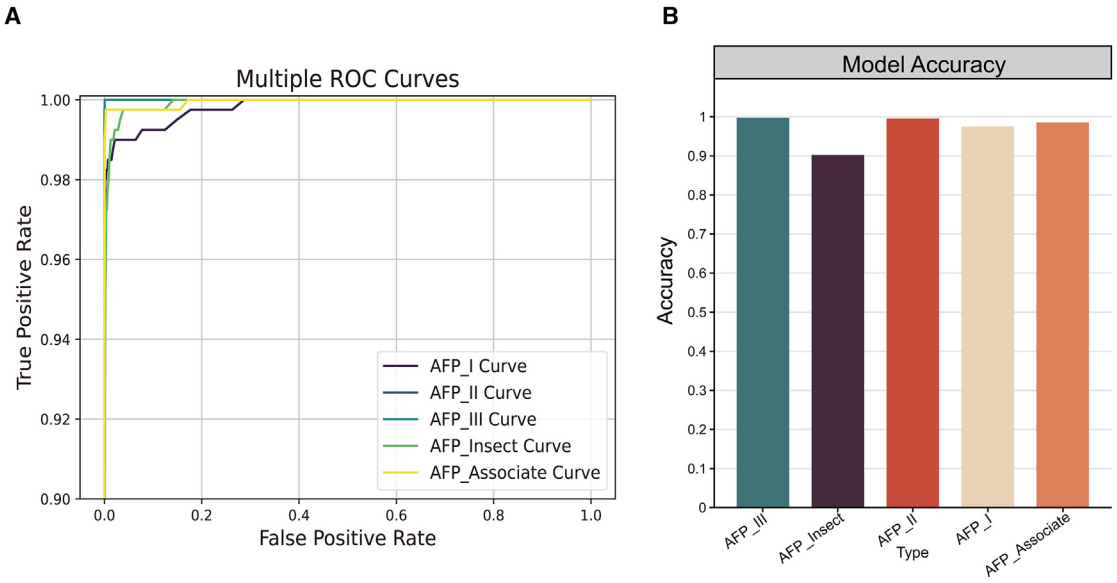


Figure 6. Evaluation of the BERT-DomainAFP model's prediction performance across different AFP types

(A) Receiver operating characteristic curve (ROC) for each AFP type prediction, which illustrates the sensitivity and specificity of the BERT-DomainAFP model to continuous variables of different AFP types. The area under the ROC curve (AUC) value reflects the classification ability of the model, and the higher the AUC value, the better the prediction effect.

(B) The average prediction accuracy for each AFP type shows the performance of the model in recognizing different types of AFP. The graphs show the accuracy for each type of AFP, focusing on the model's ability to classify non-AFPs and different AFP types with different proportions of training samples.

second on a single central processing unit (CPU), which is in line with the prediction performance of current mainstream deep learning models.[56–58]

Our research not only reveals a close correlation between structural domain annotation and prediction, as well as GO annotation prediction tasks, but also demonstrates that pre-training based on extensive data training can enhance the model's ability to handle specific tasks. We believe that this discovery provides valuable insights for future research work and is expected to drive more significant progress in related fields.

Conclusion

The accurate identification and classification of AFPs are of great importance due to their significant roles in enhancing cold tolerance in organisms and their broad applications in industries such as medicine, food storage, aquaculture, and agriculture. AFPs possess unique structural features that enable them to prevent ice formation, making them valuable for improving the preservation of biological materials and increasing the frost resistance of crops. However, the prediction of AFPs remains a challenging task due to the high structural and sequence diversity among different AFP types, coupled with the limitations of traditional prediction methods. Classical techniques such as sequence similarity searches (e.g., BLAST) and hidden Markov models struggle to accurately detect AFPs, particularly in large-scale datasets or when dealing with imbalanced classes. These methods are also computationally intensive and often fail to capture the complex, non-linear relationships inherent in protein sequences.

In response to these challenges, the BERT-DomainAFP model, as presented in this study, has proven to be a highly effective tool for the recognition and classification of AFPs. It achieves an unprecedented accuracy of 98.48% in identifying AFPs, outperforming all existing models. The model's ability to classify AFPs into different types based on their structural domain features is a significant advancement, offering deeper insights into the properties and functions of AFPs.

The integration of structural domain annotations into the training process has been key to the model's success, allowing it to capture the complexity and diversity of AFPs more effectively than simplified binary classification approaches. The use of transfer learning with the ProteinBERT model has also been instrumental, enabling the BERT-DomainAFP model to adapt quickly to the specific task of AFP prediction with high accuracy, even in the face of unbalanced datasets.

In summary, the BERT-DomainAFP model represents a major step forward in the field of AFP research, providing a powerful new tool for both basic research and practical applications. Its high accuracy, classification capabilities, and robustness against dataset imbalance make it an asset for advancing our understanding of AFPs and their roles in cold adaptation and beyond.

Future direction

Incorporation of SHAP

Integrating SHAP (Shapley additive explanations) into the analysis of the BERT-DomainAFP model will provide deeper insights into its decision-making process. SHAP will help identify the structural domains or amino acid sequences that significantly contribute to the model's predictions, thereby offering valuable

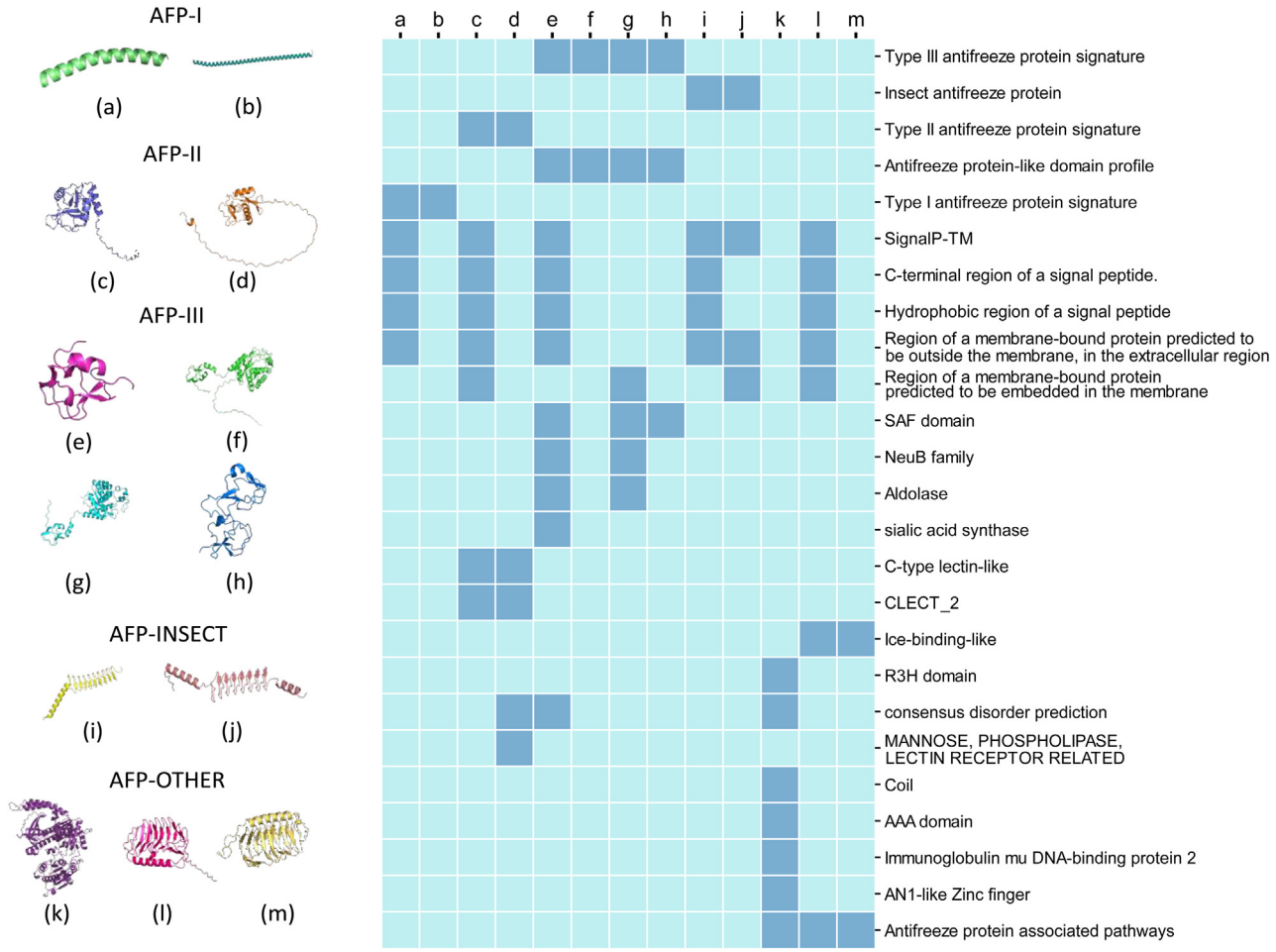


Figure 7. Overview of the predicted 3D structures and domain

Predicted three-dimensional structures of frequently occurring AFP sequences (a-m), predicted through Swiss-MODEL for online protein structure prediction, and their corresponding domain compositions (right), with each column in the heatmap representing one structural domain composition of the AFP. The fields in the heatmap are selected from the top 25 most frequent InterPro annotation strings.

information for model refinement and enhancing our understanding of AFP characteristics.

Optimization of AFP classification especially AFP type IV

Future efforts will focus on optimizing the classification of AFPs, with an emphasis on improving the discrimination of AFP type IV. As additional data on AFP structural domains becomes available, particularly for AFP type IV, the model can be trained on a broader range of sequences, leading to better classification accuracy and a more detailed understanding of the distinct properties and functions of AFP type IV.

Data balancing

Although the current approach of oversampling and undersampling has proven effective, we plan to explore advanced techniques, such as SMOTE and other sequence-based data augmentation methods, to further balance the training data. This will provide more diverse and comprehensive datasets, potentially improving the model's performance. Additionally, the data balancing methods used in the BERT-DomainAFP model, which combine transfer learning and structural domain annotation, may offer valuable insights for other machine learning tasks involving heterogeneous data.

Validation of model prediction results

In this study, the amino acid sequence data of *Cyclopterus lumpus* and *Pholis gunnellus* were predicted and the predictions are provided as a supplemental table (Table S5). It is hoped that the accuracy of AFP predictions for these sequences will be further verified in future studies.

Limitations of the study

The BERT-DomainAFP model exhibits superior performance in AFP recognition compared to existing methods. It represents the first AFP prediction model with the ability to classify AFPs into five types: AFP I, AFP II, AFP III, AFP insect, and AFP association, instead of merely distinguishing whether a protein is an AFP or non-AFP as previous models did. However, the current study is constrained by the limited quantity of reviewed data currently in the database. Additionally, InterPro has not yet incorporated and annotated the relevant sequences of

AFP IV and AFGPs. Since the validation of the model requires at least 20 experimental validation sequences as the validation dataset and a training set with 4–5 times the number of validation sets for training, the BERT-DomainAFP in this study is unable to classify AFP IV and AFGPs. In the future, the addition of more validated data for these two types of AFPs and the integration of additional published data will enable the BERT-DomainAFP to update and achieve more accurate prediction and classification.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Ruoyu Liu (liuruoyu13@mails.ucas.ac.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The BERT-DomainAFP model code used in this study is publicly available at: https://github.com/ChenShengZhen/BERT-DomainAFP/tree/main/dataset%26code/code.
- The training data and base data are stored in the following Figshare repository: https://figshare.com/articles/dataset/_b_AntiFreezeDomains_b_zip/28344692

ACKNOWLEDGMENTS

This work was supported by grants from the National Natural Science Foundation of China (32202995 and 42276247), National Key Research and Development Program of China (2022YFD2401001), and Natural Science Foundation of Fujian Province (2022J01135).

AUTHOR CONTRIBUTIONS

X.C., L.L., and R.L. conceived this project and coordinated research activities. S.C. and R.L. conducted the data collection, model construction, and drafted the manuscript. P.Z. and R.L. designed experiments and revised the manuscript. Z.M. and L.Z. conducted investigation and data collection. Q.Y. processed and analyzed the data.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ProteinBERT pre-trained model
- METHOD DETAILS
  - Datasets and domain coding
  - Evaluation method
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.112077.

Received: October 23, 2024
Revised: January 3, 2025
Accepted: February 17, 2025
Published: March 6, 2025

REFERENCES

1. Baskaran, A., Kaari, M., Venugopal, G., Manikkam, R., Joseph, J., and Bhaskar, P.V. (2021). Anti freeze proteins (Afp): Properties, sources and applications-A review. Int. J. Biol. Macromol. *189*, 292–305. https://doi.org/10.1016/j.ijbiomac.2021.08.105.
2. Du, N., Liu, X.Y., and Hew, C.L. (2003). Ice nucleation inhibition - Mechanism of antifreeze by antifreeze protein. J. Biol. Chem. *278*, 36000–36004. https://doi.org/10.1074/jbc.M305222200.
3. Ewart, K.V., Lin, Q., and Hew, C.L. (1999). Structure, function and evolution of antifreeze proteins. Cell. Mol. Life Sci. *55*, 271–283. https://doi.org/10.1007/s000180050289.
4. Tran-Guzman, A., Moradian, R., Walker, C., Cui, H., Corpuz, M., Gonzalez, I., Nguyen, C., Meza, P., Wen, X., and Culty, M. (2022). Toxicity profiles and protective effects of antifreeze proteins from insect in mammalian models. Toxicol. Lett. *368*, 9–23. https://doi.org/10.1016/j.toxlet.2022.07.009.
5. Tirado-Kulieva, V.A., Miranda-Zamora, W.R., Hernández-Martínez, E., Pantoja-Tirado, L.R., Bazán-Tantaleán, D.L., and Camacho-Orbegoso, E.W. (2022). Effect of antifreeze proteins on the freeze-thaw cycle of foods: fundamentals, mechanisms of action, current challenges and recommendations for future work. Heliyon *8*, e10973. https://doi.org/10.1016/j.heliyon.2022.e10973.
6. Bang, J.K., Lee, J.H., Murugan, R.N., Lee, S.G., Do, H., Koh, H.Y., Shim, H.E., Kim, H.C., and Kim, H.J. (2013). Antifreeze peptides and glycopeptides, and their derivatives: potential uses in biotechnology. Mar. Drugs *11*, 2013–2041. https://doi.org/10.3390/md11062013.
7. Hassas-Roudsari, M., and Goff, H.D. (2012). Ice structuring proteins from plants: Mechanism of action and food application. Food Res. Int. *46*, 425–436. https://doi.org/10.1016/j.foodres.2011.12.018.
8. Mahatabuddin, S., Hanada, Y., Nishimiya, Y., Miura, A., Kondo, H., Davies, P.L., and Tsuda, S. (2017). Concentration-dependent oligomerization of an alpha-helical antifreeze polypeptide makes it hyperactive. Sci. Rep. *7*, 42501. https://doi.org/10.1038/srep42501.
9. Nishimiya, Y., Kondo, H., Takamichi, M., Sugimoto, H., Suzuki, M., Miura, A., and Tsuda, S. (2008). Crystal structure and mutational analysis of Ca^{2+}-independent type II antifreeze protein from longsnout poacher, Brachyopsis rostratus. J. Mol. Biol. *382*, 734–746. https://doi.org/10.1016/j.jmb.2008.07.042.
10. Olijve, L.L.C., Meister, K., DeVries, A.L., Duman, J.G., Guo, S., Bakker, H.J., and Voets, I.K. (2016). Blocking rapid ice crystal growth through nonbasal plane adsorption of antifreeze proteins. Proc. Natl. Acad. Sci. USA *113*, 3740–3745. https://doi.org/10.1073/pnas.1524109113.
11. Sönnichsen, F.D., DeLuca, C.I., Davies, P.L., and Sykes, B.D. (1996). Refined solution structure of type III antifreeze protein: hydrophobic groups may be involved in the energetics of the protein–ice interaction. Structure *4*, 1325–1337. https://doi.org/10.1016/S0969-2126(96)00140-2.
12. Urbańczyk, M., Góra, J., Latajka, R., and Sewald, N. (2017). Antifreeze glycopeptides: from structure and activity studies to current approaches in chemical synthesis. Amino Acids *49*, 209–222. https://doi.org/10.1007/s00726-016-2368-z.
13. Khan, A., Uddin, J., Ali, F., Banjar, A., and Daud, A. (2023). Comparative analysis of the existing methods for prediction of antifreeze proteins. Chemometr. Intell. Lab. Syst. *232*, 104729. https://doi.org/10.1016/j.chemolab.2022.104729.
14. Gilbert, J.A., Davies, P.L., and Laybourn-Parry, J. (2005). A hyperactive, Ca^{2+}-dependent antifreeze protein in an Antarctic bacterium. FEMS Microbiol. Lett. *245*, 67–72. https://doi.org/10.1016/j.femsle.2005.02.022.

15. Eskandari, A., Leow, T.C., Rahman, M.B.A., and Oslan, S.N. (2020). Antifreeze Proteins and Their Practical Utilization in Industry, Medicine, and Agriculture. Biomolecules *10*, 1649. https://doi.org/10.3390/biom10121649.

16. Jia, Z., and Davies, P.L. (2002). Antifreeze proteins: an unusual receptor–ligand interaction. Trends Biochem. Sci. *27*, 101–106. https://doi.org/10.1016/s0968-0004(01)02028-x.

17. Xiao, X., Hui, M., and Liu, Z. (2016). iAFP-Ense: an ensemble classifier for identifying antifreeze protein by incorporating grey model and PSSM into PseAAC. J. Membr. Biol. *249*, 845–854. https://doi.org/10.1007/s00232-016-9935-9.

18. Kandaswamy, K.K., Chou, K.-C., Martinetz, T., Möller, S., Suganthan, P.N., Sridharan, S., and Pugalenthi, G. (2011). AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. J. Theor. Biol. *270*, 56–62. https://doi.org/10.1016/j.jtbi.2010.10.037.

19. Zhao, X., Ma, Z., and Yin, M. (2012). Using support vector machine and evolutionary profiles to predict antifreeze protein sequences. Int. J. Mol. Sci. *13*, 2196–2207. https://doi.org/10.3390/ijms13022196.

20. Mondal, S., and Pai, P.P. (2014). Pseudo amino acid composition improves antifreeze protein prediction. J. Theor. Biol. *356*, 30–35. https://doi.org/10.1016/j.jtbi.2014.04.006.

21. Shahid, A., Maqsood, H., Muhammad, K., and Muhammad, I. (2019). iAFP-gap-SMOTE: An Efficient Feature Extraction Scheme Gapped Dipeptide Composition is Coupled with an Oversampling Technique for Identification of Antifreeze Proteins. Lett. Org. Chem. *16*, 294–302. https://doi.org/10.2174/1570178615666180816101653.

22. Ali, F., Akbar, S., Ghulam, A., Maher, Z.A., Unar, A., and Talpur, D.B. (2021). AFP-CMBPred: Computational identification of antifreeze proteins by extending consensus sequences into multi-blocks evolutionary information. Comput. Biol. Med. *139*, 105006. https://doi.org/10.1016/j.compbiomed.2021.105006.

23. Khan, A., Uddin, J., Ali, F., Ahmad, A., Alghushairy, O., Banjar, A., and Daud, A. (2022). Prediction of antifreeze proteins using machine learning. Sci. Rep. *12*, 20672. https://doi.org/10.1038/s41598-022-24501-1.

24. Khan, A., Uddin, J., Ali, F., Kumar, H., Alghamdi, W., and Ahmad, A. (2023). AFP-SPTS: An Accurate Prediction of Antifreeze Proteins Using Sequential and Pseudo-Tri-Slicing Evolutionary Features with an Extremely Randomized Tree. J. Chem. Inf. Model. *63*, 826–834. https://doi.org/10.1021/acs.jcim.2c01417.

25. Dhibar, S., and Jana, B. (2023). Accurate Prediction of Antifreeze Protein from Sequences through Natural Language Text Processing and Interpretable Machine Learning Approaches. J. Phys. Chem. Lett. *14*, 10727–10735. https://doi.org/10.1021/acs.jpclett.3c02817.

26. Feng, C., Wei, H., Li, X., Feng, B., Xu, C., Zhu, X., and Liu, R. (2024). A stacking-based algorithm for antifreeze protein identification using combined physicochemical, pseudo amino acid composition, and reduction property features. Comput. Biol. Med. *176*, 108534. https://doi.org/10.1016/j.compbiomed.2024.108534.

27. Qi, D., and Liu, T. (2024). VotePLMs-AFP: Identification of antifreeze proteins using transformer-embedding features and ensemble learning. Biochim. Biophys. Acta. Gen. Subj. *1868*, 130721. https://doi.org/10.1016/j.bbagen.2024.130721.

28. Kumar, N., Choudhury, S., Bajiya, N., Patiyal, S., and Raghava, G.P.S. (2025). Prediction of Anti-Freezing Proteins From Their Evolutionary Profile. Proteomics *25*, e202400157. https://doi.org/10.1002/pmic.202400157.

29. Usman, M., and Lee, J.A. (2019). AFP-CKSAAP: Prediction of Antifreeze Proteins Using Composition of k-Spaced Amino Acid Pairs with Deep Neural Network. In 2019 IEEE 19th international conference on bioinformatics and bioengineering (BIBE) (IEEE).

30. Usman, M., Khan, S., and Lee, J.-A. (2020). Afp-lse: Antifreeze proteins prediction using latent space encoding of composition of k-spaced amino acid pairs. Sci. Rep. *10*, 7197. https://doi.org/10.1038/s41598-020-63259-2.

31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. *30*.

32. Gupta, S., and Shankar, R. (2023). miWords: transformer-based composite deep learning for highly accurate discovery of pre-miRNA regions across plant genomes. Brief. Bioinform. *24*, bbad088. https://doi.org/10.1092/bib/bbad088.

33. Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. Neural Comput. *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

34. Xuan, P., Guo, M., Liu, X., Huang, Y., Li, W., and Huang, Y. (2011). PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs. Bioinformatics *27*, 1368–1376. https://doi.org/10.1093/bioinformatics/btr153.

35. Zheng, X., Fu, X., Wang, K., and Wang, M. (2020). Deep neural networks for human microRNA precursor detection. BMC Bioinf. *21*, 17. https://doi.org/10.1186/s12859-020-3339-7.

36. Zhou, Y., Wu, J., Yao, S., Xu, Y., Zhao, W., Tong, Y., and Zhou, Z. (2023). DeepCIP: A multimodal deep learning method for the prediction of internal ribosome entry sites of circRNAs. Comput. Biol. Med. *164*, 107288. https://doi.org/10.1016/j.compbiomed.2023.107288.

37. Huang, Y., He, N., Chen, Y., Chen, Z., and Li, L. (2018). BERMP: a cross-species classifier for predicting m6A sites by integrating a deep learning algorithm and a random forest approach. Int. J. Biol. Sci. *14*, 1669–1677. https://doi.org/10.7150/ijbs.27819.

38. Miyata, R., Moriwaki, Y., Terada, T., and Shimizu, K. (2021). Prediction and analysis of antifreeze proteins. Heliyon *7*, e07953. https://doi.org/10.1016/j.heliyon.2021.e07953.

39. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics *38*, 2102–2110. https://doi.org/10.1093/bioinformatics/btac020.

40. Kulmanov, M., and Hoehndorf, R. (2021). DeepGOPlus: improved protein function prediction from sequence. Bioinformatics *37*, 1187. https://doi.org/10.1093/bioinformatics/btaa763.

41. Yang, M., Chen, S., Huang, Z., Gao, S., Yu, T., Du, T., Zhang, H., Li, X., Liu, C.M., Chen, S., and Li, H. (2023). Deep learning-enabled discovery and characterization of HKT genes in Spartina alterniflora. Plant J. *116*, 690–705. https://doi.org/10.1111/tpj.16397.

42. Devlin, J., Chang, M.W., Lee, K., Toutanova, K., and Assoc Computat, L. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In 2019 conference of the north american chapter of the association for computational linguistics: human language technologies (NAACL HLT 2019), VOL. 1.

43. Ji, S., Hölttä, M., and Marttinen, P. (2021). Does the magic of BERT apply to medical code assignment? A quantitative study. Comput. Biol. Med. *139*, 104998. https://doi.org/10.1016/j.compbiomed.2021.104998.

44. Park, M., Seo, S.-w., Park, E., and Kim, J. (2022). EpiBERTope: a sequence-based pre-trained BERT model improves linear and structural epitope prediction by learning long-distance protein interactions effectively. Preprint at bioRxiv. https://doi.org/10.1101/2022.02.27.481241.

45. Zhou, K., Lei, C., Zheng, J., Huang, Y., and Zhang, Z. (2023). Pre-trained protein language model sheds new light on the prediction of Arabidopsis protein–protein interactions. Plant Methods *19*, 141. https://doi.org/10.1186/s13007-023-01119-6.

46. Eslami, M., Zade, R.S.H., Takalloo, Z., Mahdevar, G., Emamjomeh, A., Sajedi, R.H., and Zahiri, J. (2018). afpCOOL: A tool for antifreeze protein prediction. Heliyon *4*, e00705. https://doi.org/10.1016/j.heliyon.2018.e00705.

47. Wang, S., Deng, L., Xia, X., Cao, Z., and Fei, Y. (2021). Predicting antifreeze proteins with weighted generalized dipeptide composition and multi-regression feature selection ensemble. BMC Bioinf. *22*, 340. https://doi.org/10.1186/s12859-021-04251-z.

48. He, X., Han, K., Hu, J., Yan, H., Yang, J.-Y., Shen, H.-B., and Yu, D.-J. (2015). TargetFreeze: identifying antifreeze proteins via a combination of weights using sequence evolutionary information and pseudo amino acid composition. J. Membr. Biol. *248*, 1005–1014. https://doi.org/10.1007/s00232-015-9811-z.

49. Akbar, S., Hayat, M., Kabir, M., and Iqbal, M. (2019). iAFP-gap-SMOTE: an efficient feature extraction scheme gapped dipeptide composition is coupled with an oversampling technique for identification of antifreeze proteins. Lett. Org. Chem. *16*, 294–302.

50. Davies, P.L. (2014). Ice-binding proteins: a remarkable diversity of structures for stopping and starting ice growth. Trends Biochem. Sci. *39*, 548–555. https://doi.org/10.1016/j.tibs.2014.09.005.

51. Xiang, H., Yang, X., Ke, L., and Hu, Y. (2020). The properties, biotechnologies, and applications of antifreeze proteins. Int. J. Biol. Macromol. *153*, 661–675. https://doi.org/10.1016/j.ijbiomac.2020.03.040.

52. The Gene Ontology Consortium (2019). The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res. *47*, D330–D338. https://doi.org/10.1093/nar/gky1055.

53. Bryant, P., and Elofsson, A. (2020). Decomposing Structural Response Due to Sequence Changes in Protein Domains with Machine Learning. J. Mol. Biol. *432*, 4435–4446. https://doi.org/10.1016/j.jmb.2020.05.021.

54. Kallberg, M., and Lu, H. (2009). A Machine Learning Protocol for Distinguish Intra-domain Peripheral Membrane Targeting Properties using Sequence and Structure. Biophys. J. *96*, 206a. https://doi.org/10.1016/j.bpj.2008.12.1837.

55. Lee, D., Park, I.-B., and Kim, K. (2024). An incremental learning approach to dynamic parallel machine scheduling with sequence-dependent setups and machine eligibility restrictions. Appl. Soft Comput. *164*, 112002. https://doi.org/10.1016/j.asoc.2024.112002.

56. Kulmanov, M., Khan, M.A., Hoehndorf, R., and Wren, J. (2018). DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics *34*, 660–668. https://doi.org/10.1093/bioinformatics/btx624.

57. Wang, L., and Zhou, Y. (2024). MRM-BERT: a novel deep neural network predictor of multiple RNA modifications by fusing BERT representation and sequence features. RNA Biol. *21*, 1–10. https://doi.org/10.1080/15476286.2024.2315384.

58. Zhang, F., Song, H., Zeng, M., Wu, F.-X., Li, Y., Pan, Y., and Li, M. (2021). A deep learning framework for gene ontology annotations with sequence-and network-based information. IEEE/ACM Trans. Comput. Biol. Bioinform. *18*, 2208–2217. https://doi.org/10.1109/TCBB.2020.2968882.

59. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

60. Huang, X., Schmelter, F., Irshad, M.T., Piet, A., Nisar, M.A., Sina, C., and Grzegorzek, M. (2023). Optimizing sleep staging on multimodal time series: Leveraging borderline synthetic minority oversampling technique and supervised convolutional contrastive learning. Comput. Biol. Med. *166*, 107501. https://doi.org/10.1016/j.compbiomed.2023.107501.

61. Shi, H., Wu, C., Bai, T., Chen, J., Li, Y., and Wu, H. (2023). Identify essential genes based on clustering based synthetic minority oversampling technique. Comput. Biol. Med. *153*, 106523. https://doi.org/10.1016/j.compbiomed.2022.106523.

62. Zhang, L., Bai, T., and Wu, H. (2023). sgRNA-2wPSM: Identify sgRNAs on-target activity by combining two-window-based position specific mismatch and synthetic minority oversampling technique. Comput. Biol. Med. *155*, 106489. https://doi.org/10.1016/j.compbiomed.2022.106489.

63. Hu, J., He, X., Yu, D.-J., Yang, X.-B., Yang, J.-Y., and Shen, H.-B. (2014). A new supervised over-sampling algorithm with application to protein-nucleotide binding residue prediction. PLoS One *9*, e107676. https://doi.org/10.1371/journal.pone.0107676.

64. Wang, X., Yu, B., Ma, A., Chen, C., Liu, B., and Ma, Q. (2019). Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. Bioinformatics *35*, 2395–2402. https://doi.org/10.1093/bioinformatics/bty995.

65. Zhang, S., and Duan, X. (2018). Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC. J. Theor. Biol. *437*, 239–250. https://doi.org/10.1016/j.jtbi.2017.10.030.

66. Kabir, M.W.U., Alawad, D.M., Pokhrel, P., and Hoque, M.T. (2024). DRBpred: A sequence-based machine learning method to effectively predict DNA-and RNA-binding residues. Comput. Biol. Med. *170*, 108081. https://doi.org/10.1016/j.compbiomed.2024.108081.

67. Kazemi, A., Rasouli-Saravani, A., Gharib, M., Albuquerque, T., Eslami, S., and Schüffler, P.J. (2024). A systematic review of machine learning-based tumor-infiltrating lymphocytes analysis in colorectal cancer: Overview of techniques, performance metrics, and clinical outcomes. Comput. Biol. Med. *173*, 108306. https://doi.org/10.1016/j.compbiomed.2024.108306.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
AntiFreezeDomain	This paper	https://figshare.com/articles/dataset/_b_AntiFreezeDomains_b_zip/28344692
DatasetR	Miyata et al.[38]	https://github.com/ryomiya/Prediction-and-analysis-of-antifreeze-proteins
Structural domain annotation data	This paper	https://github.com/ChenShengZhen/BERT-DomainAFP/tree/main/dataset%26code/base_data/Domain_annotation_Interpro
Software and algorithms		
BERT-DomainAFP	This paper	https://github.com/ChenShengZhen/BERT-DomainAFP/tree/main/dataset%26code/code
ProteinBERT	Brandes et al.[39]	https://github.com/lucidrains/protein-bert-pytorch/
Python version 3.8.12	Python Software Foundation	https://www.python.org

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

In this study, the primary experimental model is the ProteinBERT pre-trained model. Our study mainly focuses on gene and protein sequences and does not have animals, human participants, plants, microbe strains, cell lines, or primary cell as participants. Model details are in the "ProteinBERT pre-trained model" subsection, and gene/protein sequence information is in the "datasets and domain coding" subsection. The experimental protocol was approved by the Experimental Animal Ethics and Welfare Committee of Fujian Agriculture and Forestry University (PZCASFAFU22019 and PZCASFAFU24069).

ProteinBERT pre-trained model

In this study, we performed transfer learning based on the ProteinBERT deep learning model architecture. We modified the pre-trained model by removing its original output layer and replacing it with a new output layer suitable for the specific classification task. We employed the Gelu (Gaussian error linear unit) activation function in all hidden and linear layers of the model. The kernel size of both the dilated convolutional layer (dilation rate = 5) and the narrow convolutional layer was set to 9. The model's predictions were directly output via a linear layer, and a classification threshold of 0 was manually set.

METHOD DETAILS

Datasets and domain coding

After integration and screening, we combined 487 AFP sequences from datasetR[13] and 771 sequences from the InterPro database. These sequences were obtained using search terms such as "antifreeze" and related terms .To ensure that the model learns more features of the unconserved sequences from the dataset, we use CD-HIT[59] to remove redundancy in the dataset with a 30% homogeneity cutoff. As a result, we constructed a comprehensive dataset encompassing a total of 1,094 AFP sequences. For label processing of this dataset, we input batches of 100 AFP sequences into InterPro for domain annotation. The domain keyword frequencies of all AFP sequences were then adopted as coding labels. We extracted all AFP-related domains with a frequency greater than 10, categorizing AFP sequences into type I-III AFPs, insect AFPs, and other AFP-related domains into five main tags. Additionally, 22 domain entries that frequently appeared in antifreeze protein sequences were added as secondary tags. We selected 13 high-frequency antifreeze protein sequences from various categories to illustrate their three-dimensional structural diversity and used a heatmap to depict the domain composition of each sequence (Figure 7). Finally, these were encoded through an entry dictionary (Table S1). The encoded tag was a one-dimensional vector with a length of 27 for subsequent model prediction.

The antifreeze protein sequence is encoded as an integer tagged sequence. Twenty unique tags were used to represent the 20 canonical amino acids, as well as three additional tags (START, END, and PAD) used to identify the sequence start position, and the end position with sequence blank position fillers. For amino acid sequences longer than 1024, we split them into multiple sequences of less than 1024 based on the range of the domain.

The workflow for dataset collection and domain coding for AFPs is illustrated in Figure 1 were constructed a negative dataset using non-AFP sequences for training. The negative dataset included 9,637 reviewed non-AFP sequences searched in UniProtKB.[18] Each

sequence in the negative dataset was also divided into batches of 100 sequences for domain prediction and encoded as labels. Notably, for the 27 domains we focus on, most negative data failed to receive any domain annotation, and a small number of negative data that received domain annotation did not contain direct AFP domains. The ratio of the negative dataset with domain annotation to the negative dataset without any domain annotation was about 1:19. Finally, we divided the training set and the test set for testing. The ratio of the training set to the validation set to the test set was 7:1:2.

For protein sequence data, we encoded it as a one-dimensional vector with a length threshold of 1024. Twenty unique tags were used to represent the 20 canonical amino acids, as well as three additional tags (START, END, and PAD) used to identify the sequence start position, and the end position with sequence blank position fillers. For amino acid sequences with a length of more than 1024, we split them into multiple sequences with a length of less than 1024 according to the domain scope and assigned corresponding domain labels to each subsequence. In previous machine learning tasks, over-sampling and under-sampling have been proven to be effective methods for promoting data balance.[60–62] Consequently, facing the significant imbalance between positive and negative datasets in the number of AFP sequences, we adopted the strategy of over-sampling and under-sampling during the transfer learning phase of our model training, this method is often used in machine learning tasks with imbalanced data to further improve the performance of machine learning models.[63–65] The basic steps involved over-sampling the minority classes of AFP samples while under-sampling the majority classes of AFP samples and non-AFP samples to ensure the balance of different types of AFPs (Figure 1). However, to ensure a fair comparison with other models during binary validation, we do not apply data balancing techniques. Finally, in each training round, we used balanced AFP data along with newly randomly selected non-AFP data for training, to ensure that the model was continuously exposed to balanced AFP samples during the training process and to allow the model to extract more features from all the data.

Evaluation method

In machine learning tasks, selecting an appropriate evaluation metric system that fits the task can greatly assist in enhancing the predictive performance of a model.[66,67] Because of the imbalance between the data of different kinds of AFPs and the data of AFPs and non-AFPs, it was difficult to adopt k-fold cross-validation for all types of AFP data during the training process. In this experiment, repeated random subsampling was used to verify the model. Initially, the original data set was randomly divided into a training set, validation set and test set. Subsequently, models were established for multiple datasets at the same time, and each model was evaluated to exclude the chance of the model. We randomly sampled the dataset domain ten times, and each dataset contained an equal amount of training set, test set, and validation set. Each dataset was named train_1-10. To verify the predictive accuracy of the model in identifying BERT-DomainAFP in more imbalanced data, we added three training batches under imbalanced conditions, named train_imbalance_1-3. For the validation of the model, the accuracy, sensitivity, specificity, MCC, precision, recall, F1-Score were mainly used to evaluate the model:

$$Accuracy(ACC) = \frac{TP+TN}{TP+FP+TN+FN}$$

$$Recall = Sensitivity(Sn) = \frac{TP}{TP+FN}$$

$$Specifity(Sp) = \frac{TN}{TN+FP}$$

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

$$Precision = \frac{TP}{TP+FP}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision+Recall}$$

TN, TP, FN, and FP represented true negative, true positive, false positive, and false negative respectively. The validation of the model consists of two key components. Firstly, we considered the model as a binary classification model to compute the prediction accuracy for AFP sequences and non-AFP sequences. In this configuration, the test set maintained a 1:1 ratio of antifreeze to non-AFPs. To comparison with other models, we also included training and evaluation on the Dataset (600), which consisted of 300 validated AFP sequences and 300 non-AFPs instances. Additionally, the test dataset (9474) was used for validation to examine the generalization ability of the model, containing 181 AFPs and 9,293 non-AFPs instances. Secondly, we approached the model as a multi-label classification model, encompassing categories such as non-AFP, AFP type I, AFP type II, AFP type III, insect AFP,

and AFP association sequence. The ratio of non-AFPs to the number of each AFP type in the aforementioned datasets was 5:1, and in the case of imbalanced datasets, the ratio could be 20:1.

QUANTIFICATION AND STATISTICAL ANALYSIS

All computations were performed in the Python programming language. The graphic abstract were generated by Adobe Illustrator. Other plots appearing in this study were generated by the Python package and Adobe Illustrator.