WILEY-VCH

Proteomics
Proteomics and Systems Biology

RESEARCH ARTICLE

Prediction of Anti-Freezing Proteins From Their Evolutionary Profile

Nishant Kumar[1] | Shubham Choudhury[1] | Nisha Bajiya[1] | Sumeet Patiyal[1,2] | Gajendra P. S. Raghava[1]

[1]Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India | [2]Cancer Data Science Laboratory, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA

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

Received: 14 May 2024 | **Revised:** 29 August 2024 | **Accepted:** 29 August 2024

Funding: The current work has been supported by the Department of Biotechnology (DBT) grant BT/PR40158/BTIS/137/24/2021.

Keywords: alignment-based | alignment-free | antifreeze proteins | ensemble method | machine learning | PSSM profile

ABSTRACT

Prediction of antifreeze proteins (AFPs) holds significant importance due to their diverse applications in healthcare. An inherent limitation of current AFP prediction methods is their reliance on unreviewed proteins for evaluation. This study evaluates, proposed and existing methods on an independent dataset containing 80 AFPs and 73 non-AFPs obtained from Uniport, which have been already reviewed by experts. Initially, we constructed machine learning models for AFP prediction using selected composition-based protein features and achieved a peak AUROC of 0.90 with an MCC of 0.69 on the independent dataset. Subsequently, we observed a notable enhancement in model performance, with the AUROC increasing from 0.90 to 0.93 upon incorporating evolutionary information instead of relying solely on the primary sequence of proteins. Furthermore, we explored hybrid models integrating our machine learning approaches with BLAST-based similarity and motif-based methods. However, the performance of these hybrid models either matched or was inferior to that of our best machine-learning model. Our best model based on evolutionary information outperforms all existing methods on independent/validation dataset. To facilitate users, a user-friendly web server with a standalone package named "AFPropred" was developed (https://webs.iiitd.edu.in/raghava/afpropred).

1 | Introduction

In the 1950s, Scholander et al. found fish species that could survive in freezing temperatures, challenging traditional ideas about living in cold climates [1–3]. In 1969, DeVries and his team linked these adaptations to antifreeze proteins (AFPs) [4]. Since then, AFPs have been discovered in various species, such as insects, fungi, bacteria, and mammals. These proteins help these organisms to survive in extremely cold temperatures either by avoiding freezing or tolerating it [5–10]. AFPs sustain organisms from freezing stress by thermal hysteresis and preventing ice recrystallization [5, 11–16]. Due to their unique freeze-resistance property, AFPs have a wide range of applications, including food preservation [17–19], medicine [20, 21], human cryosurgery, and the production of yoghurt [22–26].

Numerous experimental techniques have been developed to identify and characterize AFPs. This includes ammonium precipitation and ice affinity chromatography for purification of AFPs [27–29], Ice-etching [30, 31], fluorescence-based ice plane affinity

Abbreviations: AAC, amino acid composition; Ala, alanine; AUROC, area under the receiver operating characteristic; BLAST, basic local alignment search tool; ET, extra tree classifier; FS, feature selection; GNB, Gaussian Naive Bayes; Ile, isoleucine; KNN, k-nearest neighbor; LR, logistic regression; MCC, Matthew's correlation coefficient; ML, machine learning; MLP, multi-layer perceptron classifier; PSI-BLAST, position-specific iterated BLAST; PSSM, position-specific scoring matrix; RF, random forest; Thr, threonine; Val, valine; XGB, extreme gradient boosting.

Summary

- Prediction of antifreeze proteins with high precision.
- Evaluation of prediction models on an independent dataset.
- Machine learning based models using sequence composition.
- Evolutionary information based prediction models.
- A webserver for predicting, scanning, and designing AFPs.

(FIPA) [32], and site-directed mutations [33–35] to determine ice-affinity planes to assess the mechanism of AFPs. These experimental techniques are time-consuming and costly [36]. Simple similarity-based techniques like BLAST or PSI-BLAST fail to detect distantly related AFPs [23, 25, 37].

Previously, several in silico methods have been used to discriminate AFPs and non-AFPs [38]; a brief description of existing methods is shown in Table 1. In 2011, Kandaswamy and Chou [23] developed AFP-Pred, a pioneered ML-based approach deploying random forest (RF) to classify AFPs. In the same year, Yu and Lu developed an SVM method, iAFP [39], utilizing a genetic algorithm (GA) for feature selection. Subsequently, various AFP prediction tools are AFP_PSSM [40], AFP-PseAAC [41], AFP-ensemble [42], TargetFreeze [43], iAFP-Ense [44], CryoProtect [45], afpCOOL [25], and RAFP-Pred [46]. Boosting methods also have been employed by PoGB-pred [47], Miyata et al. [26], and AFP-LXGB [24]. Recently, sparse representation-based classifiers such as AFP-LSE [16] and AFP-SRC [38] have also been developed for AFP Prediction.

The major limitation of the existing methods is their dataset, as these methods have been evaluated on unreviewed data. This study has made a systematic effort to create a validation dataset of AFPs obtained from Swiss-Prot (reviewed) for evaluation. We implement a robust strategy for AFP prediction by leveraging diverse approaches, including machine learning, sequence similarity, and pattern detection. In addition, ensemble methods have been developed where two or more than two approaches have been integrated to predict AFPs.

TABLE 1 | List of available methods of AFP prediction.

Name	Year	Source/Database	Dataset description
AFP-Pred[a] [23]	2011	Pfam database	481 AFPs and 9493 non-AFPs
iAFP[a] [39]	2011	PDB, UniProKB	Training: 44 AFPs and 3762 non-AFPs Testing: 369 AFPs
AFP_PSSM[a] [40]	2012	AFP-Pred	481 AFPs and 9493 non-AFPs
AFP-PseAAC [41]	2014	AFP-Pred	481 AFPs and 9493 non-AFPs
TargetFreeze[a] [43]	2015	AFP-Pred, Pfam database	Training: 300 AFPs and 300 non-AFPs Testing: 181 AFPs and 8293 non-AFPs
AFP-ensemble[a] [42]	2015	AFP-Pred	Training: 371 AFPs and 7266 non-AFPs Testing: 93 AFPs and 1817 non-AFPs
CryoProtect [45]	2016	AFP-Pred	478 AFPs and 9139 non-AFPs
RAFP-Pred[a] [46]	2018	AFP-Pred, iAFP, UniProKB	Data1: 481 AFPs and 9493 non-AFPs (AFP-Pred) Data2: 44 AFPs and 3762 non-AFPs (iAFP) Data3: 369 AFPs (Uniprot)
afpCOOL[a] [25]	2018	AFP-Pred, UniProtKB	Data1: 481 AFPs and 9493 non-AFPs Data2: 517 AFP and 517 non-AFPs
AFP-CKSAAP [48]	2019	AFP-Pred, Pfam	481 AFPs and 9493 non-AFPs
W-GDipc-LRMR-Ri [36]	2019	AFP-Pred, MemType-2L	Data1: 480 AFPs and 374 no-AFPs Data2: 7582 MemType-2L dataset
AFP-LSE [16]	2020	AFP-Pred	481 AFPs and 9493 non-AFPs
Sun et al. [49]	2020	AFP-Pred	481 AFPs and 9493 non-AFPs
AFP-CMBPred [50]	2021	AFP-Pred, Pfam database	Training: 300 AFPs and 300 non-AFPs Testing: 181 AFPs and 8293 non-AFPs
AFP-LXGB[a] [24]	2022	AFP-Pred, Pfam database	481 AFPs and 9193 non-AFPs
AFP-SPTS[a] [51]	2022	AFP-Pred, Pfam database	481 AFPs and 9193 non-AFPs
AFP-SRC [38]	2022	AFP-Pred	481 AFPs and 9493 non-AFPs

[a] Webserver is not working properly or not available.

2 | Material and Methods

2.1 | Dataset Compilation and Pre-Processing

The reliability of a method depends on the quality of the dataset used for training and evaluation. Ideally, both training and testing datasets should be experimentally validated [52, 53]. In this study, we have used two datasets, the main and the validation, which were used for model training and external validation, respectively.

2.1.1 | Validation Dataset

To create an independent/validation dataset, we extracted AFPs from reviewed sequences of the Swiss-Prot [54]. We retrieved 297 AFPs from UniProtKB using relevant keyword in reviewed entries, complete list of keywords is shown in Table S1. After eliminating redundant sequences and fragments, we got 80 non-redundant sequences out of the 297 sequences. These 80 AFPs having ranging from 16 to 2439 amino acids used to validate our models. In order to create set of non-AFP, we searched UniProtKB for the term "NOT_antifreeze_protein" and obtained 429,837. We got 81,339 non-redundant non-AFPs, after removing redundant sequence using CD-hit at cut-off 40%. We randomly picked an equal number of negative sequences and then removed sequences which are not in the range of more than 16 and less than 2439 amino acids. Finally, we have 73 non-AFPs and 80 AFPs in validation dataset for evaluating our models.

2.1.2 | Main Dataset

For the training of our model, we constructed the AFPs (positive) dataset by initially obtaining 49,352 (unreviewed) AFP sequences from the UniProt database [54]. Additionally, we applied the standard practices followed in bioinformatics, by employing the CD-HIT program [55] to reduce the sequence identity to 40%. After eliminating identical sequences and sequences that contain non-natural amino acids, we were left with 8134 positive sequences with lengths varying from 37 to 12,385. Consequently, the 9493 non-AFPs (negative) were collected from the existing tool, AFP-Pred. We were left with 9439 non-AFPs after removing the non-natural amino acids. Finally, we have 17,573 training sequences for model training and optimization.

2.2 | Composition Analysis

In this study, we have conducted an assessment of amino acid composition (AAC) for both positive and negative datasets, with a specific focus on AFPs and non-AFPs. AAC is the percentage frequency of the 20 different amino acids within a peptide or protein sequence. We have utilized the "Pfeature" [56] module to compute and calculate the AAC composition. To calculate AAC, we applied the following equation:

$$CR_i = \frac{NR_i}{TR} \tag{1}$$

CR_i represents the composition of residue i; NR_i is the total number of residues of type i; TR stands for the total number of residue in the sequence.

TABLE 2 | List of descriptors calculated using the Pfeature.

Type of feature	Vector size
AAC (amino acid composition)	20
DPC (dipeptide composition)	400
ATC (atomic composition)	5
BTC (bond composition)	4
PCP (physico-chemical properties based composition)	30
RRI (residue repeat information)	20
DDOR (distance distribution of residues)	20
SER (Shannon entropy for all residues)	20
SEP (Shannon entropy at protein level)	1
CTC (conjoint triad calculation)	343
PAAC (pseudo amino acid composition)	21
APAAC (amphiphilic pseudo amino acid composition)	23
QSO (quasi-sequence order)	42
SOCN (sequence order coupling number)	2
CeTD (composition enhanced transition and distribution)	189
SPC (Shannon entropy at property level)	25
Total vector size	1165

This analysis helps in gaining important insights into the significance of AAC for a better understanding of the quantity and distribution of particular amino acids across various datasets.

2.3 | Feature Generation Techniques

Discovering the discriminative features using appropriate techniques is important for designing an effective computational model for classifying the AFPs. In this regard, we have utilized the Pande et al. standalone software "Pfeature" [56] to extract the salient patterns from primary sequences of AFPs [24].

2.3.1 | Composition Based Features

In this study, we have utilized the composition-based feature module of "Pfeature" to compute the features/descriptors listed in Table 2. This module calculates various features based on the dataset's composition or proportion of specific elements.

2.3.2 | Evolutionary Based Features

The protein's evolutionary features are known to provide additional crucial information about proteins [57, 58]. The evolutionary information of proteins was retrieved from a position-specific scoring matrix (PSSM) profile generated using Position-Specific Iterated BLAST (PSI-BLAST) [59]. In this study, we have used the "pssm_composition" module from the POSSUM package [60] to generate the PSSM-400, a 20 × 20 dimension matrix for a protein sequence composition profile that measures the occurrence of 20 amino acids in the sequence. Using the Swiss-Prot, we generated a PSSM matrix for each sequence whose hit was found in the database. We have created an empty PSSM-400 with zero values for those sequences whose PSSM is not generated due to lack of homologous sequence. The complete dataset pool of sequences is in the Table S2.

2.4 | Feature Selection Techniques

In the majority of datasets, only a small number of features are relevant that contribute to identifying the endpoint. However, a vast number of irrelevant and redundant features remain a significant issue. To overcome this issue, we have applied the feature selection techniques to identify the relevant subset of features from the original set. In addition, numerous techniques for feature selection have been developed so far and effectively implemented in various fields [61]. In this study, we initially ranked the features based on mutual information by implementing an ensemble minimum redundancy–maximum relevance (mRMR) technique [49, 62]. The top-ranked features were most relevant to discriminating AFPs and non-AFPs and complementary to each other.

2.5 | Selection of Appropriate ML Models

To select an appropriate classifier for the prediction of AFPs, we have used different machine learning algorithms, as its prediction performance will depend not only on the feature representation method but also on the classifier used [24, 43]. This study used a different set of classifiers from the Scikit-learn [63] library, including k-nearest neighbor (KNN) [64], RF [65], logistic regression (LR) [66], Gaussian Naïve Bayes (GNB) [67], extremely randomized tree (ET) [68], multi-layer Perceptron classifier (MLP) [69], and extreme gradient boosting (XGB) [70].

2.6 | Cross-Validation Technique

We have adopted the k-fold cross-validation (CV) as a model selection criterion in which the whole dataset is divided into k-folds. In our study, $k = 5$, then k-1 folds are used for model construction, and the hold-out fold is allocated to model validation [71]. We have divided the complete dataset into 80:20 ratios, where 80% constitutes the training dataset, and 20% constitutes the validation dataset. The model was trained and evaluated based on five-fold CV and the independent validation dataset [49]. Finally, the mean of the performances of five iterations is used as the overall performance [58, 72].

2.7 | Evaluation Parameters

We have utilized libraries like Scikit-learn (sklearn), pandas, and numpy to build machine-learning models based on classification. From sklearn, we have imported the following classifiers: RF, ET, GB, MLP, and LR. XGB and other libraries were installed using the pip packages [73]. To evaluate our model's performance, we used the threshold-dependent metrics of sensitivity, specificity, accuracy, Mathews correlation coefficient (MCC), Youden's Index, balanced accuracy, F1-score, and threshold-independent metrics of area under the receiver operating characteristics curve (AUROC) [74]. These threshold-dependent parameters can be calculated as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{MCC} = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

$$\text{Youden's Index} = \text{Sensitivity} + \text{Specificity} - 1 \tag{6}$$

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{7}$$

$$F1\ \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

where, *TP* stands for true positive; *TN* stands for true negative; *FP* stands for false positive; *FN* stands for false negative

Once the model was evaluated, we chose our top-performing model for further analysis, in which we integrated the evolutionary features with composition-based features and the ML score with the BLAST score and named the hybrid methods.

2.8 | Motif Based Features

It is crucial to identify conserved motifs in biological sequences to discover common shared functions. This study uses the MERCI tool to identify degenerate motifs based on a given classification of amino acids according to their physico-chemical properties [75]. MERCI is employed to select distinct motifs from the positive dataset by analyzing input sequences of negative and positive sets. A number of MERCI parameters were applied to extract exclusive or inclusive motifs of both sets. The default value of the maximal frequency of the negative sequences (fn) is zero, which is used to create exclusive motifs—that is, motifs that are not shared by the positive and negative sets. We extended this number to fn = 8 to achieve inclusive motifs. We devised motifs from exclusive and inclusive motifs by determining values such as (a) No gap, (b) Gap = 1, (c) Gap = 2, and (d) Class = Koolman–Rohm [75]. Following that, the unique proteins possessing the motifs were

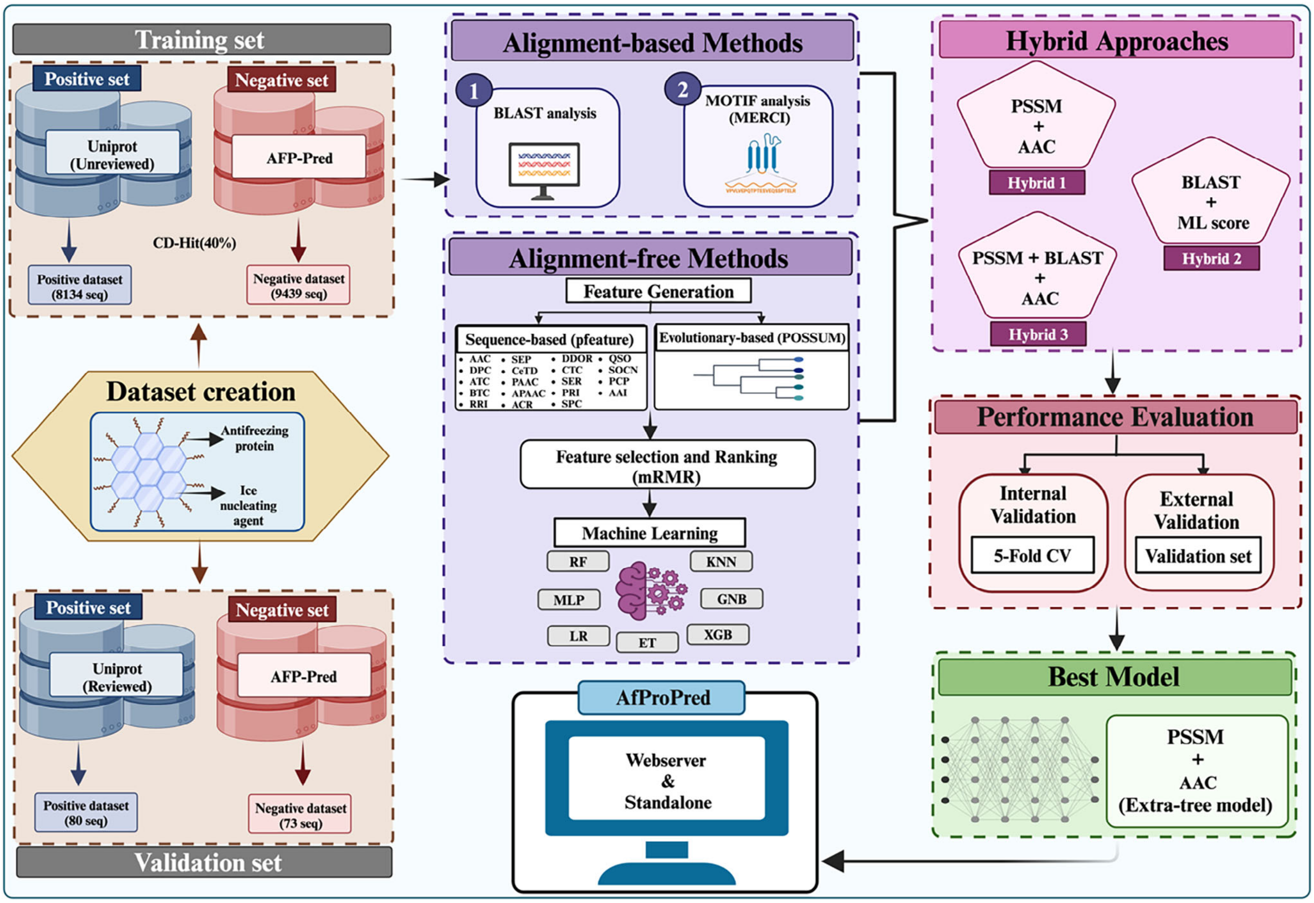


FIGURE 1 | The workflow of the study includes data collection, alignment-based and alignment-free methods, model development and evaluation, and development of web services.

identified to estimate the total coverage of motifs in the protein sequence [58].

2.9 | Webserver Architecture

In this study, we have developed a webserver named "AFProPred" (available at https://webs.iiitd.edu.in/raghava/afpropred/) to predict AFPs and non-AFPs. The webserver front and back end are developed using HTML5, CSS, Java, and PHP scripts. It is user-friendly, easily accessible, and compatible with almost all devices, including the desktop, tablet, and mobile phone. In the webserver, we have incorporated three major modules: (i) Predict, (ii) Design, and (iii) Scan.

3 | Results

In this study, we have divided the result section broadly into the following sub-sections: (i) Compositional analysis of AFPs, (ii) Alignment-based method, (iii) Alignment-free method, (iv) Hybrid approaches, (v) Benchmarking, and (vi) Webserver. A complete workflow of the study is shown in Figure 1, and a description of these subsections can be found below.

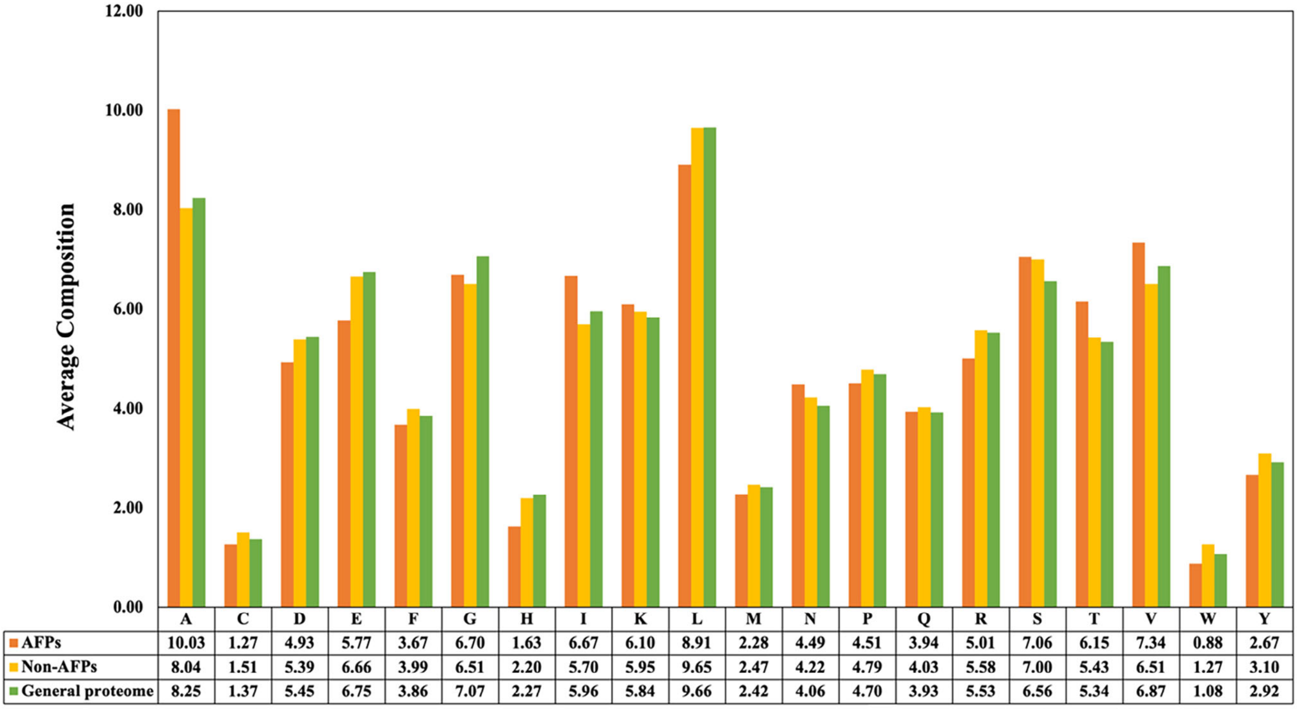
3.1 | Compositional Analysis

We have calculated the composition of individual amino acids to assess the prevalence of amino acid residues in each dataset. By analyzing the average composition of amino acids in protein sequence, the researcher can identify the potential AFPs. Here, we computed the average residue composition for APFs, non-AFPs, and general proteome in Figure 2, which exhibits that Alanine (A), Isoleucine (I), Valine (V), Threonine (T) are most abundant in AFPs (Positive dataset) while in non-AFPs Leucine (L), Glutamic acid (E) are highly conserved. Figure 3 represents the difference of average composition scores of AFPs compared to general proteome and no-nAFPs.

3.2 | Alignment-Based Method

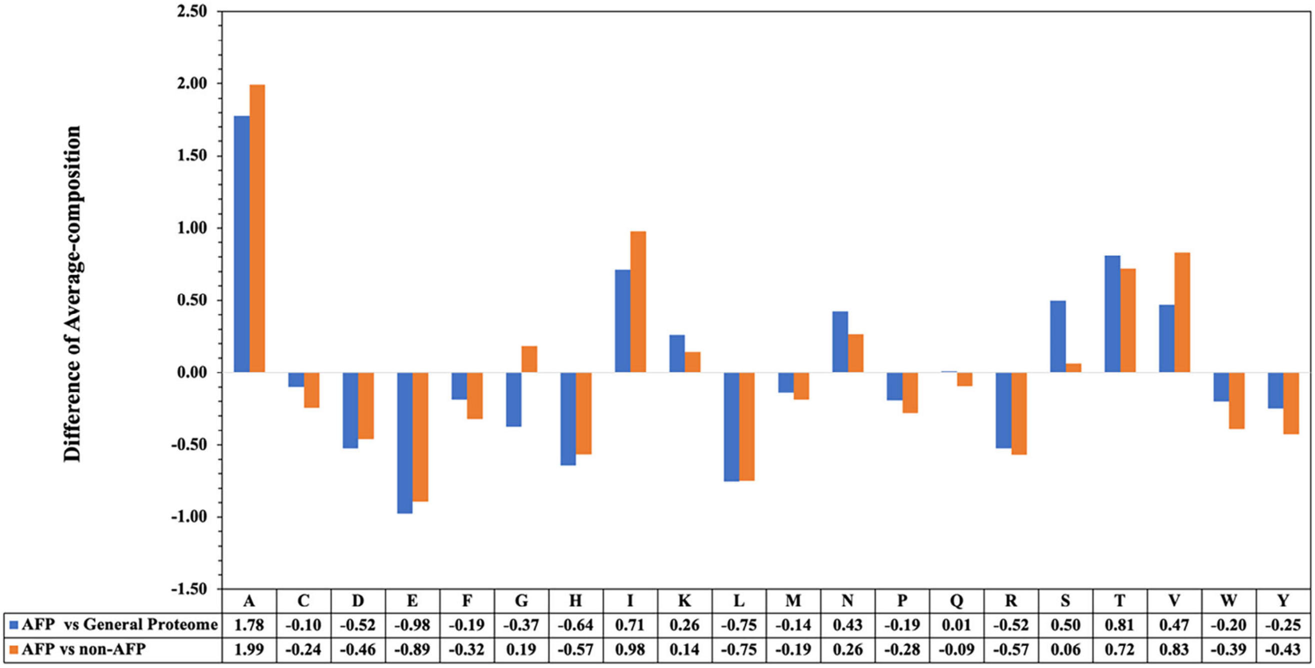
3.2.1 | Motifs Analysis

We aim to identify exclusive and inclusive motifs/patterns present in AFPs, which can be used to search AFPs. We employed the open-source/freeware available MERCI [75] software to discover motifs. As MERCI provides several options, we explore the various options or parameters: no gap (default), a gap of 1, a gap of 2, and Koolman–Rohm (Table S3). It is crucial to remember that MERCI utilizes an fn (maximum



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
🟧 AFPs	10.03	1.27	4.93	5.77	3.67	6.70	1.63	6.67	6.10	8.91	2.28	4.49	4.51	3.94	5.01	7.06	6.15	7.34	0.88	2.67
🟨 Non-AFPs	8.04	1.51	5.39	6.66	3.99	6.51	2.20	5.70	5.95	9.65	2.47	4.22	4.79	4.03	5.58	7.00	5.43	6.51	1.27	3.10
🟩 General proteome	8.25	1.37	5.45	6.75	3.86	7.07	2.27	5.96	5.84	9.66	2.42	4.06	4.70	3.93	5.53	6.56	5.34	6.87	1.08	2.92

FIGURE 2 | Comparison analysis of the average composition of amino acid residues in AFPs, non-AFPs, and general proteome.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
🟦 AFP vs General Proteome	1.78	-0.10	-0.52	-0.98	-0.19	-0.37	-0.64	0.71	0.26	-0.75	-0.14	0.43	-0.19	0.01	-0.52	0.50	0.81	0.47	-0.20	-0.25
🟧 AFP vs non-AFP	1.99	-0.24	-0.46	-0.89	-0.32	0.19	-0.57	0.98	0.14	-0.75	-0.19	0.26	-0.28	-0.09	-0.57	0.06	0.72	0.83	-0.39	-0.43

FIGURE 3 | The differences in average-composition scores of AFPs compared to general proteome and non-AFPs.

frequency in negative sequences) with a value of zero by default, producing exclusive motifs. We changed the fn value to 8 to widen our analysis and get an inclusive variety of motifs for both negative and positive datasets. Unfortunately, only limited number of sequences have motifs. Thus, a motif-based approach is not adequate to predict the function of all proteins.

3.2.2 | Similarity Search Approach

A commonly used software package, BLAST (BLAST+ 2.7.1), has been used for performing similarity searches where a query sequence is searched against a database [76]. A training dataset was used to build a target database for performing a blastp search, where query sequences (sequences in the test set) were searched

TABLE 3 | AUROC of machine learning based models developed using different types of protein composition on validation dataset (80 AFPs and 73 non-AFPs).

Features	XGB	RF	MLP	LR	KNN	GNB	ET
AAC	0.89	0.87	0.86	0.89	0.89	0.79	0.87
DPC	0.84	0.83	0.90	0.91	0.83	0.70	0.82
ATC	0.77	0.76	0.79	0.86	0.75	0.73	0.78
BTC	0.67	0.70	0.56	0.78	0.65	0.58	0.68
PCP	0.89	0.87	0.87	0.89	0.77	0.75	0.86
RRI	0.74	0.71	0.71	0.69	0.63	0.55	0.70
DDR	0.83	0.84	0.81	0.62	0.74	0.66	0.86
SER	0.88	0.87	0.87	0.87	0.89	0.77	0.86
SEP	0.67	0.65	0.68	0.68	0.62	0.69	0.64
CTC	0.75	0.73	0.79	0.77	0.70	0.65	0.74
PAAC	0.88	0.88	0.87	0.89	0.89	0.78	0.88
APAAC	0.87	0.89	0.87	0.90	0.89	0.80	0.89
QSO	0.87	0.86	0.87	0.91	0.89	0.77	0.83
SOCN	0.66	0.67	0.66	0.65	0.63	0.72	0.68
CeTD	0.80	0.72	0.77	0.82	0.75	0.66	0.74
SPCl	0.85	0.87	0.83	0.85	0.74	0.71	0.85

at different e-values. Each query sequence has been classified as AFP or non-AFP based on the top hit. For example, the query sequence is designated as an AFP if the query sequence has the top hit against an AFP. One of the challenges with this approach is assigning a class to a protein if it has no hits with AFPs or non-AFPs. Thus, a similarity-based approach is not adequate to predict the function of all proteins.

3.3 | Alignment-Free Method

As shown above, alignment-based techniques were unable to provide full coverage. To address this challenge, we applied various machine-learning techniques to construct models for predicting AFPs. These techniques include RF, MLP, LR, XGB, KNN, Extra Tree (ET), and Gaussian Naive Bayes (GNB). These models have been developed using a wide range of compositional features, evolutionary based features or PSSM profile, combined features, and selected features.

3.3.1 | Compositional Features

We build machine learning-based prediction models using different type of composition-based features that includes AAC as well as dipeptide composition (DPC). These features have been generated using standalone software "Pfeature," different type of features including number of descriptors is shown in Table 2. The performance of machine learning based models developed using different type of features have been shown in Table S4. The performance in terms of AUROC for these models on different types of protein compositions on our validation dataset that contain 80 AFPs and 73 non-AFPs is shown in Table 3.

3.3.2 | PSSM Profiles

It has been shown in several studies in the past that evolutionary information-based models perform better than single-sequence based models [58, 77–79]. Thus, this study also explores the potential of evolutionary information in predicting AFPs. We used the POSSUM package [60] to generate evolutionary information for each protein sequence in the form of a PSSM profile. While we managed to successfully generated PSSM profile for 146 out of 153 proteins in validation dataset, the method fails to generate PSSM profile for seven AFPs (Table S2). The performance of our machine learning based models developed using PSSM profile on validation dataset (146 proteins) is shown in Table S5. As shown in Table S5, models based on ET, XGB and RF achieved a maximum AUROsC of 0.92 on validation dataset containing 73 AFPs and 73 non-AFPs.

3.3.3 | Combined Features

Subsequently, we aggregated all compositional features, resulting in a vector of size 1165 for each sequence across the datasets. We then developed the prediction models employing the distinct classifier on the combined feature dataset, meticulously fine-tuning the parameters to optimize the AUROC. Our analysis showed that the ET and XGB perform almost equally in terms of AUROC, which is 0.86 and 0.87, respectively. Additionally, they achieved MCC scores of 0.54 and 0.63, correspondingly. Table 4 provides a comprehensive overview of performance metrics, encompassing both threshold-dependent and threshold-independent measures, for all the classifiers that were trained and validated across training and validation datasets.

3.3.4 | Selected Features

In a large number of feature sets, some might be unrelated, making it time-consuming to analyze and train the model. It is essential to select the more relevant features that result in a more simplified model by focusing only on the significant ones. In this study, we have generated 1165 features using Pfeature. We used mRMR [80, 81] feature selection technique to select most relevant features by removing redundant features. We got best set of 150 features after applying mRMR. These features were used to develop machine learning based models for discriminating AFPs and non-AFPs (Table 5). As shown in Table 5, ET and RF perform better other methods.

3.4 | Hybrid Approaches

As demonstrated in the preceding sections, we utilized alignment-based methods (Motif & BLAST) and alignment-free methods (machine learning techniques). Each approach has its advantages and drawbacks. Alignment-based methods offer high specificity but poor sensitivity. Their performance relies on the similarity or presence of motifs. On the other hand, machine learning-based models or alignment-free methods are more generalized, with performance unaffected by similarity. We developed ensemble or hybrid methods to capitalize on the

TABLE 4 | Performance of different models on the combined features on main (for training) and validation dataset (for validation).

Classifier	Sens (%)	Spec (%)	Acc (%)	AUROC	MCC	F1-score	BAC (%)	YI
				Training (8134 AFPs and 9439 non-AFPs)				
ET	81.72	81.39	81.54	0.90	0.63	0.80	81.55	0.63
GNB	35.28	82.74	60.78	0.67	0.21	0.45	59.01	0.18
KNN	58.78	73.87	66.89	0.72	0.33	0.62	66.33	0.33
LR	70.85	71.99	71.46	0.78	0.43	0.70	71.42	0.43
MLP	69.98	70.01	69.99	0.76	0.40	0.68	69.99	0.40
RF	81.72	82.1	81.92	0.90	0.64	0.81	81.91	0.64
XGB	82.70	82.53	82.61	0.91	0.65	0.82	82.62	0.65
				Validation (80 AFPs and 73 non-AFPs)				
ET	71.25	82.19	76.47	0.86	0.54	0.76	76.72	0.53
GNB	67.50	61.64	64.71	0.67	0.29	0.67	64.57	0.29
KNN	38.75	69.86	53.6	0.55	0.09	0.47	54.31	0.09
LR	61.25	79.45	69.94	0.82	0.41	0.68	70.35	0.41
MLP	56.25	90.41	72.55	0.78	0.49	0.68	73.33	0.47
RF	62.50	82.19	71.90	0.84	0.45	0.70	72.35	0.45
XGB	82.50	80.82	81.70	0.87	0.63	0.83	81.66	0.63

Abbreviations: Acc, accuracy; AUROC, area under receiver operating curve; BAC, balanced accuracy; ET, extra-trees classification; GNB, Gaussian Naive Bayes; KNNs, k-nearest neighbors; LR, logistic regression; MCC, Matthews correlation coefficient; MLP, multi-layer perceptron classifier; RF, random forest; Sens, sensitivity; Spec, specificity; XGB, XGBoost; YI, Youden's Index.

TABLE 5 | The performance of machine learning based models developed using 150 features select by mRMR technique.

Features	Sens (%)	Spec (%)	Acc (%)	AUROC	MCC	F1-score	BAC (%)	YI
				Training (8134 AFPs and 9439 non-AFPs)				
ET	79.41	79.09	79.24	0.88	0.58	0.78	79.25	0.58
GNB	56.06	78.15	67.93	0.74	0.35	0.62	67.11	0.34
KNN	67.09	72.51	70.00	0.76	0.40	0.67	69.80	0.40
LR	73.48	74.48	74.02	0.81	0.48	0.72	73.98	0.48
MLP	83.53	83.36	83.44	0.90	0.67	0.82	83.44	0.67
RF	82.59	83.49	83.08	0.91	0.66	0.82	83.04	0.66
XGB	80.98	80.62	80.79	0.89	0.62	0.80	80.80	0.62
				Validation (80 AFPs and 73 non-AFPs)				
ET	83.75	84.93	84.31	0.90	0.69	0.85	84.34	0.69
GNB	58.75	91.78	74.51	0.83	0.53	0.71	75.27	0.51
KNN	61.25	82.19	71.24	0.77	0.44	0.69	71.72	0.43
LR	78.75	82.19	80.39	0.87	0.61	0.81	80.47	0.61
MLP	53.75	86.30	69.28	0.80	0.42	0.65	70.03	0.40
RF	76.25	91.78	83.66	0.92	0.69	0.83	84.02	0.68
XGB	70.00	82.19	75.82	0.88	0.52	0.75	76.10	0.52

Abbreviations: Acc, accuracy; AUROC, area under receiver operating curve; BAC, balanced accuracy; ET, extra-trees classification; GNB, Gaussian Naive Bayes; KNNs, k-nearest neighbors; LR, logistic regression; MCC, Matthews correlation coefficient; MLP, multi-layer perceptron classifier; RF, random forest; Sens, sensitivity; Spec, specificity; XGB, XGBoost; YI, Youden's Index.

TABLE 6 | Performance on the PSSM composition profile combined with AAC composition on main (for training) and validation dataset (for validation).

Classifier	Sens (%)	Spec (%)	Acc (%)	AUROC	MCC	F1-score	BAC (%)	YI
				Training (8134 AFPs and 9439 non-AFPs)				
ET	82.91	83.66	83.32	0.92	0.67	0.82	83.29	0.67
GNB	51.48	78.79	66.15	0.73	0.32	0.59	65.14	0.30
KNN	82.35	82.07	82.20	0.90	0.64	0.81	82.21	0.64
LR	80.16	80.06	80.11	0.87	0.60	0.79	80.11	0.60
MLP	85.52	85.68	85.60	0.93	0.71	0.85	85.60	0.71
RF	83.38	83.19	83.28	0.91	0.67	0.82	83.29	0.67
XGB	84.13	84.37	84.26	0.92	0.68	0.83	84.25	0.69
				Validation (80 AFPs and 73 non-AFPs)				
ET	85.00	91.78	88.24	0.93	0.77	0.89	88.39	0.77
GNB	46.25	91.78	67.97	0.85	0.42	0.60	69.02	0.38
KNN	68.75	87.67	77.78	0.89	0.57	0.76	78.21	0.56
LR	47.50	79.45	62.75	0.79	0.28	0.57	63.48	0.27
MLP	70.00	73.97	71.90	0.80	0.44	0.72	71.99	0.44
RF	76.25	87.67	81.70	0.91	0.64	0.81	81.96	0.64
XGB	73.75	78.08	75.82	0.86	0.52	0.76	75.92	0.52

Abbreviations: Acc, accuracy; AUROC, area under receiver operating curve; BAC, balanced accuracy; ET, extra-trees classification; GNB, Gaussian Naive Bayes; KNNs, k-nearest neighbors; LR, logistic regression; MCC, Matthews correlation coefficient; MLP, multi-layer perceptron classifier; RF, random forest; Sens, sensitivity; Spec, specificity; XGB, XGBoost; YI, Youden's Index.

strengths of both alignment-free and alignment-based models. First, we developed ML models using a combination of PSSM-based features and AAC. The hybrid PSSM-based ET models perform better than others, achieving an AUROC of 0.93 with 0.77 MCC, as shown in Table 6. In addition, we developed machine learning techniques using a combination of 150 selected and the PSSM profile. It was observed that RF and ET attain almost similar performance in terms of AUROC of 0.92 (Table S6). We have combined BLAST-based approach with our ET-based models developed using AAC+PSSM model and achieved a maximum AUROC of 0.90 (Table S7). In this study, we have tried many combinations to develop hybrid models. We achieved a maximum performance AUROC of 0.93 on the validation dataset using the PSSM+AAC feature based on ET-model. Ultimately, we used this model for our study to develop a standalone package and web server.

3.5 | Benchmarking

To assess the significance of a newly developed method, it is important to compare its performance with available methods on the same dataset. Thus, we evaluate existing methods on our validation dataset. Unfortunately, not all existing methods are available. Therefore, we can predict the performance of those methods that are fully functional and available to the public. Our findings show that the existing methods CryoProtect [45], AFP-CKSAAP [48], AFP-LSE [16], and AFP-SRC [38] do not perform better on the validation dataset as compared to AFProPred, as shown in Table 7.

3.6 | Webserver Implementation

To provide a more user-friendly webserver for the scientific community, we developed AFProPred, which can be accessed at https://webs.iiitd.edu.in/raghava/afpropred/. This platform is designed to predict AFPs using our top-performing model. AFProPred offers several modules: Prediction, Design, and Protein Scan. The Prediction module effectively distinguishes between AFPs and non-AFPs, allowing users to submit protein sequences in FASTA format for prediction. The Design module enables users to generate all possible analogs of the input sequence and predict AFPs among analogs. The protein Scan module assists in identifying AFP regions within a given protein sequence. This platform is developed using a responsive template to browse on various devices, including smartphones, laptops, and desktops. Additionally, we have created a Python-based standalone package named "AFProPred" to facilitate users in predicting AFP proteins at the genome scale. This package is available via the "download module" on the web server at https://webs.iiitd.edu.in/raghava/afpropred/standalone.html.

4 | Discussion

Numerous methods have been developed in the past to improve the accuracy of prediction of AFPs (Table 1). Most of the existing studies used the AFP-Pred dataset, which contains 481 AFPs and 9493 non-AFPs. AFPs were mainly extracted from the Pfam database in a few studies (like iAFP) that extracted AFPs from PDB. One of the significant limitations of existing studies is their

TABLE 7 | Comparison of available methods on the validation dataset (80 AFPs and 73 non-AFPs).

Name	Year	Sens (%)	Spec (%)	Acc (%)	YI	MCC	AUROC	F1-score	BAC (%)
Cryoprotect	2017	0.79	0.43	60.13	0.22	0.23	0.61	NA	0.61
AFP-CKSAAP	2019	0.78	0.88	82.00	0.65	0.65	0.89	NA	0.83
AFP-LSE	2020	0.73	0.75	74.00	0.48	0.48	NA	0.74	0.74
AFP-SRC	2022	0.56	0.58	57.00	0.14	0.14	NA	0.58	0.57
AFProPred	—	0.85	0.92	88.24	0.77	0.77	0.93	0.89	0.90

Abbreviations: Acc, accuracy; AUROC, area under receiver operating curve; BAC, balanced accuracy; MCC, Matthews correlation coefficient; Sens, sensitivity; Spec, specificity; YI, Youden's Index.

dataset; none of the existing studies used experimentally validated AFPs or well-annotated AFPs available in Swiss-prot, which contain reviewed proteins. Thus, developing a method based on annotated or experimentally supported data is critical. In this study, we systematically attempted to create a dataset of well-annotated AFPs. We queried Swiss-prot for extracting annotated AFPs and got only 80 AFPs, which is insufficient for training, testing, and validating our prediction models. Thus, we used these 80 AFPs only to validate our model and existing models.

Similarly, we obtained 73 non-AFPs from Swiss-prot using the keyword "NOT_antifreeze_protein." Finally, we build an independent dataset that contains well-annotated 80 AFPs and 73 non-AFPs. This independent dataset is not used for training or optimization of hyperparameters of machine learning techniques.

Our primary analysis indicated common evolutionary relationships among AFPs where certain residues are conserved in AFPs (e.g., Ala, Ile, Val, and Thr). The amino acid Thr increases the activity of AFPs by adding hydrogen bonds to their surface area [45]. First, we utilized the similarity-based standard technique BLAST to identify AFPs. In this study, BLAST fails to discriminate between AFPs and non-AFPs due to poor similarity among AFPs. To overcome this challenge, we used generalized machine learning techniques to predict AFPs. These techniques are also called alignment-free techniques as they are not based on alignment. Since, most of the machine learning techniques need fixed length vectors and proteins have variable lengths, we have computed composition-based features [16]. Here, we tried various machine-learning techniques and achieved a maximum AUC of 0.90 on an independent dataset. It is a well-known fact that evolutionary information is important for predicting the function of proteins. Thus, in this study, we also developed models using evolutionary information obtained from the PSSM profile generated using PSI-BLAST. The performance of our model also improved from AUC 0.90 to 0.93 when PSSM was used instead of protein composition. Finally, we compared the performance of our method with the existing methods. In order to provide a fair evaluation, we compare the performance of our model as well as existing methods on an independent dataset of 80 AFPs and 73-non AFPs. As shown in the Section 3, our method performs better than existing methods. One of the major weakness of this study is the evaluation dataset which is too small for a robust evaluation. Despite our best effort, we failed to create a larger validation dataset for robust evaluation. We hope that in the future, researchers will generate large datasets of AFPs and non-AFPs for training, testing and validation of their models.

5 | Conclusion

To uncover the relationship between proteins and ice crystals and, more generally, the adaptation of organisms to their environments, depends on the ability to understand the evolution of AFPs [49]. Our findings revealed that the conservation of several essential amino acids showed opposite tendencies in AFPs and non-AFPs. This suggests that there has been a significant selection pressure related to these amino acids, leading to the differentiation between AFPs and non-AFPs in regards to their ice-binding capacities. Therefore, our hybrid approach AAC combined with PSSM profiles, had high performed and outperformed the state-of-the-art tools; hence, they are effective and beneficial for identifying new AFPs.

Author Contributions

G.P.S.R. collected the dataset. N.K. and S.P. processed the dataset. N.K., S.P., and G.P.S.R. implemented the algorithms and developed the prediction models. N.K., N.B., and G.P.S.R. analyzed the results. S.C. created the front-end, back-end of the webserver, and standalone of the method. N.K., N.B., S.P., S.C., and G.P.S.R. penned the manuscript. G.P.S.R. conceived and coordinated the project. All authors have read and approved the final manuscript.

Acknowledgments

Authors are thankful to the University Grants Commission (UGC), Department of Bio-Technology (DBT), Department of Science and Technology (DST-INSPIRE), and Council of Scientific & Industrial Research (CSIR) for fellowships and financial support, and the Department of Computational Biology, IIITD New Delhi for infrastructure and facilities. We would like to acknowledge that Figures were created using BioRender.com.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Data Availability Statement

All the datasets used in this study are available at the "AFProPred" web server, https://webs.iiitd.edu.in/raghava/afpropred/dataset.html.

BioRxiv Link

https://doi.org/10.1101/2024.04.28.591577

References

1. P. F. Scholander, L. van Dam, J. W. Kanwisher, H. T. Hammel, and M. S. Gordon, "Supercooling and Osmoregulation in Arctic Fish," *Journal of Cellular and Comparative Physiology* 49 (1957): 5–24, https://doi.org/10.1002/jcp.1030490103.

2. A. Sakai and W. Larcher, *Frost Survival of Plants* (Berlin, Heidelberg: Springer Berlin Heidelberg, 1987), https://doi.org/10.1007/978-3-642-71745-1.

3. M. Yoshida, J. Abe, M. Moriyama, S. Shimokawa, and Y. Nakamura, "Seasonal Changes in the Physical State of Crown Water Associated With Freezing Tolerance in Winter Wheat," *Physiologia Plantarum* 99 (1997): 363–370, https://doi.org/10.1111/j.1399-3054.1997.tb00548.x.

4. A. L. DeVries and D. E. Wohlschlag, "Freezing Resistance in Some Antarctic fishes," *Science* 163 (1969): 1073–1075, https://doi.org/10.1126/science.163.3871.1073.

5. J. M. Logsdon Jr and W. F. Doolittle, "Origin of Antifreeze Protein Genes: A Cool Tale in Molecular Evolution," *Proceedings of the National Academy of Sciences* 94 (1997): 3485–3487, https://doi.org/10.1073/pnas.94.8.3485.

6. M. Griffith, P. Ala, D. S. Yang, W. C. Hon, and B. A. Moffatt, "Antifreeze Protein Produced Endogenously in Winter Rye Leaves," *Plant Physiology* 100 (1992): 593–596, https://doi.org/10.1104/pp.100.2.593.

7. J. G. Duman and T. M. Olsen, "Thermal Hysteresis Protein Activity in Bacteria, Fungi, and Phylogenetically Diverse Plants," *Cryobiology* 30 (1993): 322–328, https://doi.org/10.1006/cryo.1993.1031.

8. J. A. Husby and K. E. Zachariassen, "Antifreeze Agents in the Body Fluid of Winter Active Insects and Spiders," *Experientia* 36 (1980): 963–964, https://doi.org/10.1007/bf01953821.

9. T. Sformo, F. Kohl, J. McIntyre, P. Kerr, J. G. Duman, and B. M. Barnes, "Simultaneous Freeze Tolerance and Avoidance in Individual Fungus Gnats, Exechia Nugatoria," *Journal of Comparative Physiology B* 179 (2009): 897–902, https://doi.org/10.1007/s00360-009-0369-x.

10. J. Levitt, *Responses of Plants to Environmental Stress, Volume 1 Chilling, Freezing, and High Temperature Stresses* (Cambridge: Academic Press, 1980), https://www.scirp.org/reference/referencespapers?referenceid=2682591.

11. K. V. Ewart, Q. Lin, and C. L. Hew, "Structure, Function and Evolution of Antifreeze Proteins," *Cellular and Molecular Life Sciences* 55 (1999): 271–283, https://doi.org/10.1007/s000180050289.

12. C. H. Cheng, "Evolution of the Diverse Antifreeze Proteins," *Current Opinion in Genetics & Development* 8 (1998): 715–720, https://doi.org/10.1016/s0959-437x(98)80042-7.

13. P. L. Davies and B. D. Sykes, "Antifreeze Proteins," *Current Opinion in Structural Biology* 7 (1997): 828–834, https://doi.org/10.1016/s0959-440x(97)80154-6.

14. M. E. Urrutia, J. G. Duman, and C. A. Knight, "Plant Thermal Hysteresis Proteins," *Biochimica Et Biophysica Acta* 1121 (1992): 199–206, https://doi.org/10.1016/0167-4838(92)90355-h.

15. X. M. Yu, M. Griffith, and S. B. Wiseman, "Ethylene Induces Antifreeze Activity in Winter Rye Leaves," *Plant Physiology* 126 (2001): 1232–1240, https://doi.org/10.1104/pp.126.3.1232.

16. M. Usman, S. Khan, and J.-A. Lee, "AFP-LSE: Antifreeze Proteins Prediction using Latent Space Encoding of Composition of k-Spaced Amino Acid Pairs," *Scientific Reports* 10 (2020): 7197, https://doi.org/10.1038/s41598-020-63259-2.

17. X. Zhan, D.-W. Sun, Z. Zhu, and Q.-J. Wang, "Improving the Quality and Safety of Frozen Muscle Foods by Emerging Freezing Technologies: A Review," *Critical Reviews in Food Science and Nutrition* 58 (2018): 2925–2938, https://doi.org/10.1080/10408398.2017.1345854.

18. J. G. Provesi, P. A. Valentim Neto, A. C. M. Arisi, and E. R. Amante, "Extraction of Antifreeze Proteins from Cold Acclimated Leaves of Drimys Angustifolia and their Application to Star Fruit (Averrhoa Carambola) Freezing," *Food Chemistry* 289 (2019): 65–73, https://doi.org/10.1016/j.foodchem.2019.03.055.

19. D. H. Song, M. Kim, E.-S. Jin, et al., "Cryoprotective Effect of an Antifreeze Protein Purified From Tenebrio Molitor Larvae on Vegetables," *Food Hydrocolloids* 94 (2019): 585–591, https://doi.org/10.1016/j.foodhyd.2019.04.007.

20. S. G. Lee, H. Y. Koh, J. H. Lee, S.-H. Kang, and H. J. Kim, "Cryopreservative Effects of the Recombinant Ice-Binding Protein from the Arctic Yeast Leucosporidium sp. on Red Blood Cells," *Applied Biochemistry and Biotechnology* 167 (2012): 824–834, https://doi.org/10.1007/s12010-012-9739-z.

21. M. S. Khan, S. M. Ibrahim, A. A. Adamu, et al., "Pre-Grafting Histological Studies of Skin Grafts Cryopreserved in α Helix Antarctic Yeast Oriented Antifreeze Peptide (Afp1m)," *Cryobiology* 92 (2020): 26–33, https://doi.org/10.1016/j.cryobiol.2019.09.012.

22. Y.-H. Zhang, Z. Li, L. Lu, et al., "Analysis of the Sequence Characteristics of Antifreeze Protein," *Life (Basel)* 11 (2021): 520, https://doi.org/10.3390/life11060520.

23. K. K. Kandaswamy, K.-C. Chou, T. Martinetz, et al., "AFP-Pred: A Random Forest Approach for Predicting Antifreeze Proteins from Sequence-Derived Properties," *Journal of Theoretical Biology* 270 (2011): 56–62, https://doi.org/10.1016/j.jtbi.2010.10.037.

24. A. Khan, J. Uddin, F. Ali, et al., "Prediction of Antifreeze Proteins Using Machine Learning," *Scientific Reports* 12 (2022): 20672, https://doi.org/10.1038/s41598-022-24501-1.

25. M. Eslami, R. Shirali Hossein Zade, Z. Takalloo, et al., "afpCOOL: A Tool for Antifreeze Protein Prediction," *Heliyon* 4 (2018): e00705, https://doi.org/10.1016/j.heliyon.2018.e00705.

26. R. Miyata, Y. Moriwaki, T. Terada, and K. Shimizu, "Prediction and Analysis of Antifreeze Proteins," *Heliyon* 7 (2021): e07953, https://doi.org/10.1016/j.heliyon.2021.e07953.

27. M. J. Kuiper, C. Lankin, S. Y. Gauthier, V. K. Walker, and P. L. Davies, "Purification of Antifreeze Proteins by Adsorption to Ice," *Biochemical and Biophysical Research Communications* 300 (2003): 645–648, https://doi.org/10.1016/s0006-291x(02)02900-5.

28. X. Ding, H. Zhang, H. Chen, L. Wang, H. Qian, and X. Qi, "Extraction, Purification and Identification of Antifreeze Proteins From Cold Acclimated Malting Barley (*Hordeum vulgare* L.)," *Food Chemistry* 175 (2015): 74–81, https://doi.org/10.1016/j.foodchem.2014.11.027.

29. B. Sharma, D. Sahoo, and R. Deswal, "Single-Step Purification and Characterization of Antifreeze Proteins From Leaf and Berry of a Freeze-Tolerant Shrub Seabuckthorn (*Hippophae rhamnoides*)," *Journal of Separation Science* 41 (2018): 3938–3945, https://doi.org/10.1002/jssc.201800553.

30. J. Cheng, Y. Hanada, A. Miura, S. Tsuda, and H. Kondo, "Hydrophobic Ice-Binding Sites Confer Hyperactivity of an Antifreeze Protein From a Snow Mold Fungus," *Biochemical Journal* 473 (2016): 4011–4026, https://doi.org/10.1042/BCJ20160543.

31. C. A. Knight, C. C. Cheng, and A. L. DeVries, "Adsorption of Alpha-Helical Antifreeze Peptides on Specific Ice Crystal Surface Planes," *Biophysical Journal* 59 (1991): 409–418, https://doi.org/10.1016/S0006-3495(91)82234-2.

32. K. Basu, C. P. Garnham, Y. Nishimiya, S. Tsuda, I. Braslavsky, and P. Davies, "Determining the Ice-Binding Planes of Antifreeze Proteins by Fluorescence-Based Ice Plane Affinity," *Journal of Visualized Experiments: JoVE* 83 (2014): e51185, https://doi.org/10.3791/51185.

33. H. Chao, F. D. Sönnichsen, C. I. DeLuca, B. D. Sykes, and P. L. Davies, "Structure-Function Relationship in the Globular Type III Antifreeze Protein: Identification of a Cluster of Surface Residues Required for Binding to Ice," *Protein Science* 3 (1994): 1760–1769, https://doi.org/10.1002/pro.5560031016.

34. M. C. Loewen, W. Gronwald, F. D. Sönnichsen, B. D. Sykes, and P. L. Davies, “The Ice-Binding Site of Sea Raven Antifreeze Protein is Distinct From the Carbohydrate-Binding Site of the Homologous C-Type Lectin,” *Biochemistry* 37 (1998): 17745–17753, https://doi.org/10.1021/bi9820513.

35. J. Baardsnes and P. L. Davies, “Contribution of Hydrophobic Residues to Ice Binding by Fish Type III Antifreeze Protein,” *Biochimica Et Biophysica Acta* 1601 (2002): 49–54, https://doi.org/10.1016/s1570-9639(02)00431-4.

36. S. Wang, L. Deng, X. Xia, Z. Cao, and Y. Fei, “Predicting Antifreeze Proteins With Weighted Generalized Dipeptide Composition and Multi-Regression Feature Selection Ensemble,” *BMC Bioinformatics [Electronic Resource]* 22 (2021): 340, https://doi.org/10.1186/s12859-021-04251-z.

37. A. Nath and K. Subbiah, “The Role of Pertinently Diversified and Balanced Training as Well as Testing Data Sets in Achieving the True Performance of Classifiers in Predicting the Antifreeze Proteins,” *Neurocomputing* 272 (2018): 294–305, https://doi.org/10.1016/j.neucom.2017.07.004.

38. M. Usman, S. Khan, S. Park, and A. Wahab, “AFP-SRC: Identification of Antifreeze Proteins Using Sparse Representation Classifier,” *Neural Computing & Applications* 34 (2022): 2275–2285, https://doi.org/10.1007/s00521-021-06558-7.

39. C.-S. Yu and C.-H. Lu, “Identification of Antifreeze Proteins and Their Functional Residues by Support Vector Machine and Genetic Algorithms Based on n-Peptide Compositions,” *PLoS ONE* 6 (2011): e20445, https://doi.org/10.1371/journal.pone.0020445.

40. X. Zhao, Z. Ma, and M. Yin, “Using Support Vector Machine and Evolutionary Profiles to Predict Antifreeze Protein Sequences,” *International Journal of Molecular Sciences* 13 (2012): 2196–2207, https://doi.org/10.3390/ijms13022196.

41. S. Mondal and P. P. Pai, “Chou's Pseudo Amino Acid Composition Improves Sequence-Based Antifreeze Protein Prediction,” *Journal of Theoretical Biology* 356 (2014): 30–35, https://doi.org/10.1016/j.jtbi.2014.04.006.

42. R. Yang, C. Zhang, R. Gao, and L. Zhang, “An Effective Antifreeze Protein Predictor With Ensemble Classifiers and Comprehensive Sequence Descriptors,” *International Journal of Molecular Sciences* 16 (2015): 21191–21214, https://doi.org/10.3390/ijms160921191.

43. X. He, K. Han, J. Hu, et al., “TargetFreeze: Identifying Antifreeze Proteins Via a Combination of Weights Using Sequence Evolutionary Information and Pseudo Amino Acid Composition,” *Journal of Membrane Biology* 248 (2015): 1005–1014, https://doi.org/10.1007/s00232-015-9811-z.

44. X. Xiao, M. Hui, and Z. Liu, “IAFP-Ense: An Ensemble Classifier for Identifying Antifreeze Protein by Incorporating Grey Model and PSSM Into PseAAC,” *Journal of Membrane Biology* 249 (2016): 845–854, https://doi.org/10.1007/s00232-016-9935-9.

45. R. Pratiwi, A. A. Malik, N. Schaduangrat, et al., “CryoProtect: A Web Server for Classifying Antifreeze Proteins From Nonantifreeze Proteins,” *Journal of Chemistry* 2017 (2017): 1–15, https://doi.org/10.1155/2017/9861752.

46. S. Khan, I. Naseem, R. Togneri, and M. Bennamoun, “RAFP-Pred: Robust Prediction of Antifreeze Proteins Using Localized Analysis of n-Peptide Compositions,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 15 (2018): 244–250, https://doi.org/10.1109/TCBB.2016.2617337.

47. A. Alim, A. Rafay, and I. Naseem, “PoGB-Pred: Prediction of Antifreeze Proteins Sequences Using Amino Acid Composition With Feature Selection Followed by a Sequential-Based Ensemble Approach,” *Current Bioinformatics* 16 (2021): 446–456, https://doi.org/10.2174/1574893615999200707141926.

48. M. Usman and J. A. Lee, “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, 2019), https://doi.org/10.1109/bibe.2019.00016.

49. S. Sun, H. Ding, D. Wang, and S. Han, “Identifying Antifreeze Proteins Based on Key Evolutionary Information,” *Frontiers in Bioengineering and Biotechnology* 8 (2020): 244, https://doi.org/10.3389/fbioe.2020.00244.

50. F. Ali, S. Akbar, A. Ghulam, Z. A. Maher, A. Unar, and D. B. Talpur, “AFP-CMBPred: Computational Identification of Antifreeze Proteins by Extending Consensus Sequences Into Multi-Blocks Evolutionary Information,” *Computers in Biology and Medicine* 139 (2021): 105006, https://doi.org/10.1016/j.compbiomed.2021.105006.

51. A. Khan, J. Uddin, F. Ali, H. Kumar, W. Alghamdi, and A. Ahmad, “AFP-SPTS: An Accurate Prediction of Antifreeze Proteins using Sequential and pseudo-Tri-Slicing Evolutionary Features With an Extremely Randomized Tree,” *Journal of Chemical Information and Modeling* 63 (2023): 826–834, https://doi.org/10.1021/acs.jcim.2c01417.

52. F. Ali, H. Kumar, W. Alghamdi, F. A. Kateb, and F. K. Alarfaj, “Recent Advances in Machine Learning-Based Models for Prediction of Antiviral Peptides,” *Archives of Computational Methods in Engineering* 30 (2023): 4033–4044, https://doi.org/10.1007/s11831-023-09933-w.

53. S. Seo, M. Oh, Y. Park, and S. Kim, “DeepFam: Deep Learning Based Alignment-Free Method for Protein Family Modeling and Prediction,” *Bioinformatics* 34 (2018): i254–i262, https://doi.org/10.1093/bioinformatics/bty275.

54. UniProt Consortium. (2021). UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Research* 49, D480–D489, https://doi.org/10.1093/nar/gkaa1100.

55. L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, “CD-HIT: Accelerated for Clustering the Next-Generation Sequencing Data,” *Bioinformatics* 28 (2012): 3150–3152, https://doi.org/10.1093/bioinformatics/bts565.

56. A. Pande, S. Patiyal, A. Lathwal, et al., “Pfeature: A Tool for Computing Wide Range of Protein Features and Building Prediction Models,” *Journal of Computational Biology* 30 (2023): 204–222, https://doi.org/10.1089/cmb.2022.0241.

57. M. Kumar, M. M. Gromiha, and G. P. S. Raghava, “Identification of DNA-Binding Proteins Using Support Vector Machines and Evolutionary Profiles,” *BMC Bioinformatics [Electronic Resource]* 8 (2007): 463, https://doi.org/10.1186/1471-2105-8-463.

58. A. Arora, S. Patiyal, N. Sharma, N. L. Devi, D. Kaur, and G. P. S. Raghava, “A Random Forest Model for Predicting Exosomal Proteins Using Evolutionary Information and Motifs,” *Proteomics* 24 (2024), 2300231, https://doi.org/10.1002/pmic.202300231.

59. S. F. Altschul, T. L. Madden, A. A. Schäffer, et al., “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs,” *Nucleic Acids Research* 25 (1997): 3389–3402, https://doi.org/10.1093/nar/25.17.3389.

60. J. Wang, B. Yang, J. Revote, et al., “POSSUM: A Bioinformatics Toolkit for Generating Numerical Sequence Feature Descriptors Based on PSSM Profiles,” *Bioinformatics* 33 (2017): 2756–2758, https://doi.org/10.1093/bioinformatics/btx302.

61. A. Benkessirat and N. Benblidia, “Fundamentals of Feature Selection: An Overview and Comparison,” in 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA) (IEEE, 2019), https://doi.org/10.1109/aiccsa47632.2019.9035281.

62. C. Ding and H. Peng, “Minimum Redundancy Feature Selection from Microarray Gene Expression Data,” in Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003 (IEEE Computer Society, 2004), https://doi.org/10.1109/csb.2003.1227396.

63. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, Scikit-Learn: Machine Learning in Python (2012), https://doi.org/10.48550/ARXIV.1201.0490.

64. L. Peterson, “K-Nearest Neighbor,” *Scholarpedia J* 4 (2009): 1883, https://doi.org/10.4249/scholarpedia.1883.

65. L. Breiman, “Random Forests,” *Machine Learning* 45 (2001): 5–32, https://doi.org/10.1023/a:1010933404324.

66. J. S. Cramer, “The Origins of Logistic Regression,” *Social Science Research Network Electronic Journal* (2003), https://doi.org/10.2139/ssrn.360300.

67. M. V. Anand, B. KiranBala, S. R. Srividhya, M. Y. Kavitha, and M. H. Rahman, “Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer,” *Mobile Information Systems* 2022 (2022): 1–7, https://doi.org/10.1155/2022/2436946.

68. P. Geurts, D. Ernst, and L. Wehenkel, “Extremely Randomized Trees,” *Machine Learning* 63 (2006): 3–42, https://doi.org/10.1007/s10994-006-6226-1.

69. M.-C. Popescu, V. Balas, L. Perescu-Popescu, and N. Mastorakis, “Multilayer Perceptron and Neural Networks,” *WSEAS Transactions on Circuits and Systems* 8 (2009).

70. T. Chen and C. Guestrin, “XGBoost,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA: ACM, 2016), https://doi.org/10.1145/2939672.2939785.

71. Y. Jung, “Multiple Predicting *K*-Fold Cross-Validation for Model Selection,” *Journal of Nonparametric Statistics* 30 (2018): 197–215, https://doi.org/10.1080/10485252.2017.1404598.

72. I. Baturynska and K. Martinsen, “Prediction of Geometry Deviations in Additive Manufactured Parts: Comparison of Linear Regression With Machine Learning Algorithms,” *Journal of Intelligent Manufacturing* 32 (2021): 179–200, https://doi.org/10.1007/s10845-020-01567-0.

73. A. Rawal, C. Kidchob, J. Ou, O. N. Yogurtcu, H. Yang, and Z. E. Sauna, “A Machine Learning Approach for Identifying Variables Associated With Risk of Developing Neutralizing Antidrug Antibodies to Factor VIII,” *Heliyon* 9 (2023): e16331, https://doi.org/10.1016/j.heliyon.2023.e16331.

74. X. Liang, F. Li, J. Chen, et al., “Large-Scale Comparative Review and Assessment of Computational Methods for Anti-Cancer Peptide Identification,” *Briefings in Bioinformatics* 22 (2021): bbaa312, https://doi.org/10.1093/bib/bbaa312.

75. C. Vens, M.-N. Rosso, and E. G. J. Danchin, “Identifying Discriminative Classification-Based Motifs in Biological Sequences,” *Bioinformatics* 27 (2011): 1231–1238, https://doi.org/10.1093/bioinformatics/btr110.

76. G. M. Boratyn, A. A. Schäffer, R. Agarwala, S. F. Altschul, D. J. Lipman, and T. L. Madden, “Domain Enhanced Lookup Time Accelerated BLAST,” *Biology Direct* 7 (2012): 12, https://doi.org/10.1186/1745-6150-7-12.

77. D. Kaur, C. Arora, and G. P. S. Raghava, “A Hybrid Model for Predicting Pattern Recognition Receptors Using Evolutionary Information,” *Frontiers in Immunology* 11 (2020): 71, https://doi.org/10.3389/fimmu.2020.00071.

78. R. Kumar, B. Panwar, J. S. Chauhan, and G. P. Raghava, “Analysis and Prediction of Cancerlectins Using Evolutionary and Domain Information,” *BMC Research Notes* 4 (2011): 237, https://doi.org/10.1186/1756-0500-4-237.

79. R. Verma, G. C. Varshney, and G. P. S. Raghava, “Prediction of Mitochondrial Proteins of Malaria Parasite Using Split Amino Acid Composition and PSSM Profile,” *Amino Acids* 39 (2010): 101–110, https://doi.org/10.1007/s00726-009-0381-1.

80. C. Ding and H. Peng, “Minimum Redundancy Feature Selection From Microarray Gene Expression Data,” *Journal of Bioinformatics and Computational Biology* 3 (2005): 185–205, https://doi.org/10.1142/s0219720005001004.

81. L. Qian, Y. Wen, and G. Han, “Identification of Cancerlectins Using Support Vector Machines With Fusion of G-Gap Dipeptide,” *Frontiers in Genetics* 11 (2020): 275, https://doi.org/10.3389/fgene.2020.00275.

Supporting Information

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