AFPFusionLM: A Hybrid Sequence–Structure Protein Language Model for Antifreeze Protein Function Prediction

Yijun Li [*1] **Chi Wang** [*1] **Xiaonan Wang** [2 3]

[*]Equal contribution [1]*Tanwei College, Tsinghua University* [2]*Department of Chemical Engineering, Tsinghua University* [3]*Beijing Key Laboratory of Artificial Intelligence for Advanced Chemical Engineering Materials*
. Correspondence to: Xiaonan Wang wangxiaonan@tsinghua.edu.cn.

1. Introduction

Antifreeze proteins (AFPs) regulate ice-crystal nucleation and growth, reducing freeze-induced mechanical damage to biological tissues and enabling applications in organ/cell cryopreservation, cryotherapy, and food preservation. Despite this promise, AFP discovery and validation remain largely dependent on wet-lab screening, an expensive, time-consuming process that cannot realistically traverse the immense protein sequence space.

Computational identification is further complicated by its pronounced sequence diversity and weak homolog, limiting the robustness and generalizability of homology-based searches and manually engineered sequence features. Recent self-supervised protein language models (PLMs) have advanced function prediction by learning transferable evolutionary and semantic representations from large-scale unlabeled sequences.

Building on these advances, we propose a hybrid sequence–structure framework that integrates evolutionary semantics from a sequence PLM with spatial constraints captured by a tokenized structure PLM. A lightweight fusion module combines both modalities and is trained with a joint objective of supervised classification and contrastive learning to improve representation separability. Our approach achieves state-of-the-art performance on a held-out test set and independent gold-standard benchmarks, and provides a reusable structure-aware module to support subsequent function-driven, closed-loop protein design.

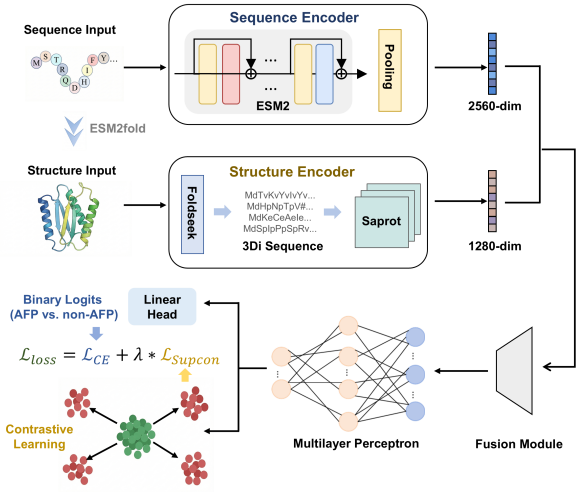

Fig. 1: The illustration of our hybrid sequence–structure PLM with a fusion module.

2. Methods

2.1 Related Work

Protein language models and structural representations. Self-supervised PLMs have become a core foundation for protein function prediction [1]. ESM2 is a representative sequence-based PLM that provides strongly transferable embeddings for a wide range of downstream tasks [1]. More recently, structure-aware modeling has introduced conformational constraints by discretizing 3D information into structural tokens; SaProt follows this direction and highlights the complementary value of structural signals for function-related prediction [2].

AFP classification and prediction. AFP identification has evolved from classical machine learning with manually engineered sequence descriptors to deep learning models leveraging PLM representations. Representative baselines include AFP-LSE [3], AFP-SRC [4], AFP-XGB [5], and BERT-DomainAFP [6], which incorporates a domain-adapted language model for AFP prediction.

2.2 Problem Setup and Data

We formulate antifreeze protein function prediction as a supervised binary classification task (AFP vs. non-AFP). Positive samples are curated from UniProt-annotated AFP entries, while negatives are drawn from non-AFP UniProt proteins with approximate matching on sequence length, organism, and subcellular localization to mitigate confounding effects. To reduce homology redundancy and potential information leakage, we apply CD-HIT with a sequence identity threshold of $\leq 40\%$, yielding 7,334/8,481 AFP/non-AFP sequences for training (15,815 total) and 800/958 for testing (1,758 total). In addition to the UniProt split, we evaluate on a SwissProt-reviewed gold-standard independent set (80/73, 153 total) to assess cross-database generalization.

2.3 Model architecture

Our approach treats *sequence semantics* and *structural constraints* as equally important sources of information, and performs AFP binary classification by fusing two complementary upstream representations. Concretely, the input to the pipeline is only the amino-acid sequence. The sequence branch extracts pretrained embeddings from ESM2 to capture co-evolutionary patterns and functional semantics. The structure branch first predicts a 3D structure from the input sequence using ESMFold, converts

Table 1: Performance comparison on the UniProt test set and the SwissProt-reviewed independent set.

	Test set (UniProt)					Independent set (SwissProt-reviewed)				
Method	Acc (%)	Sen (%)	Spe (%)	MCC	AUC (%)	Acc (%)	Sen (%)	Spe (%)	MCC	AUC (%)
AFP-LSE	79.41	74.38	87.21	0.64	83.25	74.46	73.08	75.71	0.49	77.83
AFP-SRC	81.77	83.61	79.94	0.71	85.39	57.28	57.04	58.27	0.36	60.12
AFP-XGB	92.60	90.00	94.78	0.86	97.46	86.93	85.00	89.04	0.74	95.05
BERT-DomainAFP	84.01	81.50	86.10	0.68	91.17	81.70	80.00	83.56	0.64	91.88
SaProt only	94.48	93.92	94.97	0.889	98.09	90.73	88.61	**93.06**	0.816	96.41
ESM2 only	94.78	**94.68**	94.86	0.895	98.65	90.73	89.87	91.67	0.815	95.90
Ours	**95.07**	94.05	**95.98**	**0.901**	**98.89**	**92.05**	**91.14**	**93.06**	**0.841**	**97.59**

the predicted structure into a 3Di representation via Foldseek, and then feeds the resulting 3Di tokens into SaProt to obtain structure-conditioned embeddings. We subsequently fuse the two embeddings and pass the fused representation through a multilayer perceptron to produce an intermediate feature vector, followed by a linear classification head that outputs binary logits.

We train the model with a joint objective that combines supervised classification and supervised contrastive learning, aiming to improve both predictive performance and the separability of the representation space. Let $\mathbf{f}_i$ denote the intermediate feature of sample i, and $\mathbf{z}_i = \mathbf{f}_i/\|\mathbf{f}_i\|_2$ its ℓ_2-normalized embedding. Let $P(i) = \{p \neq i \mid y_p = y_i\}$ be the set of positives sharing the same label as i. The supervised contrastive loss is defined as

$$\mathcal{L}_{\text{supcon}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\mathbf{z}_i^\top \mathbf{z}_p/\tau)}{\sum_{a\neq i}\exp(\mathbf{z}_i^\top \mathbf{z}_a/\tau)}, \tag{1}$$

where τ is a temperature parameter. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{supcon}}, \tag{2}$$

where $\mathcal{L}_{\text{CE}}$ denotes the binary cross-entropy classification loss and λ is a weighting coefficient. This joint objective improves robustness to hard negatives and distribution shifts, and yields more coherent class-wise clustering in the learned representation space.

2.4 Results

Compared with the open-source baselines compared here, our method delivers the strongest discrimination as measured by AUC on both benchmarks (98.89% on the UniProt and 97.59% on SwissProt-reviewed). Among these baselines, our model also attains the best accuracy, specificity, and MCC on the held-out UniProt test set, and the best accuracy, sensitivity, MCC, and AUC on the SwissProt-reviewed independent set, with shared-best specificity.

Ablations show that each modality is beneficial but insufficient alone. Structure-only (SaProt) reaches 98.09%/96.41% AUC (UniProt/SwissProt), and Sequence-only (ESM2) improves to 98.65%/95.90%, yet both remain below the fused model. Correspondingly, the fused model delivers the strongest overall trade-off across the two benchmarks, while ESM2-only attains the highest sensitivity on the held-out UniProt test set. These results indicate complementary sequence semantics and structural constraints, with fusion yielding more discriminative representations. In addition, we visualize the learned representations and decision boundaries (see Fig. A2).

2.5 Outlook Toward Closed-Loop Design

We further applied the predictor to high-throughput screening over protein databases: for high-scoring candidates, we quantified repeated structural domains and excised them into short peptides. We hypothesize that, among the predicted AFP candidates, more frequently occurring repeated segments are more likely to correspond to antifreeze-active sites. For the most abundant segment, we conducted molecular dynamics simulations and observed antifreeze-relevant signatures (see Fig. A1 and Fig. A3), providing preliminary computational support for this hypothesis.

Looking forward, we plan to operationalize the proposed sequence–structure fusion predictor as a scoring module in a closed-loop design pipeline: a generative model proposes candidate sequences, our predictor rapidly scores and ranks them, top candidates are validated via expression and activity assays, and experimental feedback is used to iteratively update both the generation strategy and the discriminator. This closed-loop paradigm may generalize beyond AFPs to broader protein/peptide function optimization tasks.

Acknowledgments

We thank Jialin Liu, Jinhao Yao, Yiheng Zhou and other collaborators for their valuable contributions to this research.

References

[1] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan

dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[2] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *The Twelfth International Conference on Learning Representations*, 2024.

[3] Muhammad Usman, Shujaat Khan, and Jeong-A Lee. AFP-LSE: Antifreeze Proteins Prediction Using Latent Space Encoding of Composition of k-Spaced Amino Acid Pairs. *Scientific Reports*, 10(1):7197, April 2020.

[4] Muhammad Usman, Shujaat Khan, Seongyong Park, and Abdul Wahab. AFP-SRC: identification of antifreeze proteins using sparse representation classifier. *Neural Computing and Applications*, 34(3):2275–2285, February 2022.

[5] Saikat Dhibar and Biman Jana. Accurate Prediction of Antifreeze Protein from Sequences through Natural Language Text Processing and Interpretable Machine Learning Approaches. *The Journal of Physical Chemistry Letters*, 14(48):10727–10735, December 2023.

[6] Shengzhen Chen, Ping Zheng, Lele Zheng, Qinglong Yao, Ziyu Meng, Longshan Lin, Xinhua Chen, and Ruoyu Liu. BERT-DomainAFP: Antifreeze protein recognition and classification model based on BERT and structural domain annotation. *iScience*, 28(4), April 2025.

Appendix A. Supplementary Materials

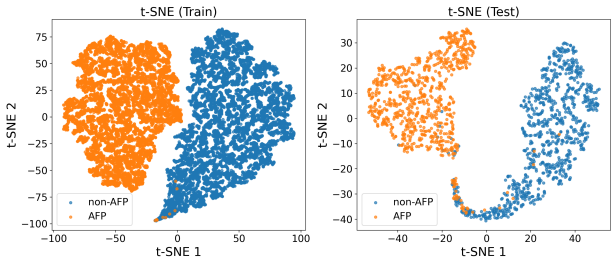

Fig. A1: **t-SNE visualization of learned representations on the training and test sets.**

t-SNE projections of the intermediate features learned by the model, shown separately for the training set and the held-out test set. The embeddings exhibit clearer class-wise clustering and separation between AFP and non-AFP samples in both splits, indicating discriminative representations with consistent generalization behavior.

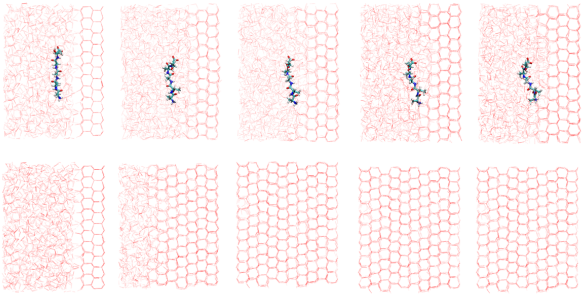

Fig. A2: **Molecular dynamics evidence of ice inhibition by an antifreeze peptide (0–100 ns).**

Comparison of crystallization behavior in a pure-water system versus a system with a selected antifreeze peptide over a 0–100 ns MD trajectory. Pure water exhibits clear ice nucleation and growth, whereas peptide addition delays and suppresses crystallization, maintaining a more disordered (liquid-like) structure for a longer duration, consistent with antifreeze-relevant ice-inhibition activity.

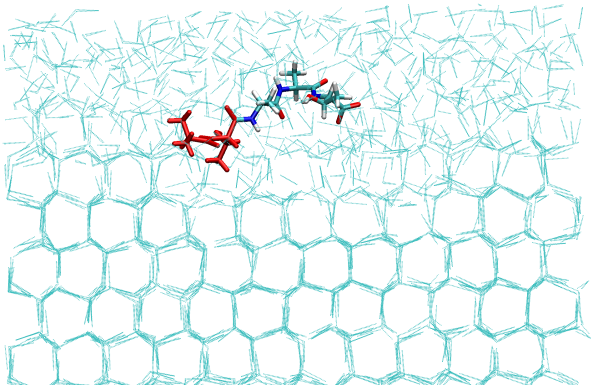

Fig. A3: **Molecular mechanism of ice inhibition by the antifreeze peptide in MD.**