JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

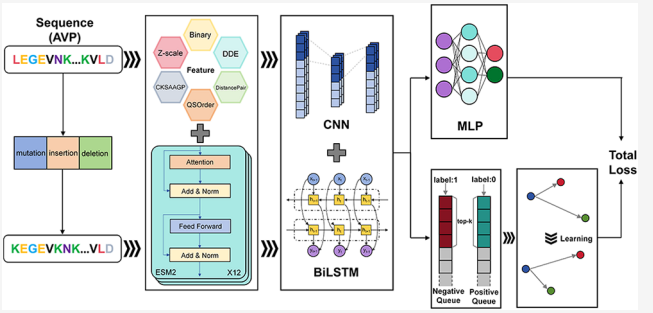
AVP-HNCL: Innovative Contrastive Learning with a Queue-Based Negative Sampling Strategy for Dual-Phase Antiviral Peptide Prediction

Yuanhao Li,[○] Aoyun Geng,[○] Zheyu Zhou, Feifei Cui, Junlin Xu, Yajie Meng, Leyi Wei, Quan Zou, Qingchen Zhang, and Zilong Zhang*

Cite This: J. Chem. Inf. Model. 2025, 65, 5868−5886 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Viral infections have long been a core focus in the field of public health. Antiviral peptides (AVPs), due to their unique mechanisms of action and significant inhibitory effects against a wide range of viruses, exhibit tremendous potential in protecting organisms from various viral diseases. However, existing studies on antiviral peptide recognition often rely on feature selection. As data volume continues to grow and task complexity increases, traditional methods are increasingly showing limitations in feature extraction capabilities and model generalization performance. To tackle these challenges, we propose an innovative two-stage predictive framework that integrates the ESM2 model, data augmentation, feature fusion, and contrastive learning techniques. This framework enables simultaneous identification of AVPs and their subclasses. By introducing a novel top-k queue-based contrastive learning strategy, the framework significantly improves the model's accuracy in distinguishing challenging positive and negative samples and its generalization performance. This approach provides robust theoretical support and technical tools for advancing research on antiviral peptides. Model evaluation results show that on Set 1-nonAVP, the framework achieves an accuracy of 0.9362 and a Matthews correlation coefficient (MCC) score of 0.8730. On the Set 2-nonAMP, the model achieves perfect accuracy (1.0000) and an MCC score of 1.0000. In addition, during the second stage, the model accurately predicts the antiviral activity of antiviral peptides against six major virus families and eight specific viruses. To further enhance accessibility for users, we have developed a user-friendly web interface, available at http://www.bioai-lab.com/AVP-HNCL.



INTRODUCTION

The limitations of current antiviral therapies and the insufficient advancement in viral pathogen research pose significant challenges to the prevention of diseases.[1,2] Antiviral peptides (AVPs), a class of small peptide molecules, have demonstrated effectiveness in inhibiting viral attachment and replication.[3−5] Compared to traditional antiviral drugs, AVPs offer notable advantages, including high specificity, low side effects, cost-effectiveness, ease of synthesis and modification, and strong sensitivity to viral resistance.[6,7] These attributes make AVPs promising candidates for the development of novel antiviral therapies. The discovery of AVPs is critically important from a biological perspective, as it enhances our understanding of host−pathogen interactions and innate immune mechanisms. AVPs are naturally occurring components of the innate immune response in various organisms, acting as first-line defense molecules that can rapidly neutralize viral pathogens. Studying AVPs can provide insights into evolutionary conserved antiviral mechanisms, aiding in deciphering the fundamental biological processes involved in viral inhibition and immunity. The discovery of AVPs is of great significance for both biological research and clinical application: to date, several AVPs have been successfully utilized in clinical settings. For example, the FDA-approved HIV inhibitor Enfuvirtide, along with Boceprevir and Telaprevir for hepatitis C treatment, has shown significant clinical efficacy.[8,9] However, the process of screening natural AVPs remains time-consuming and costly, underscoring the urgent need for more efficient, low-toxicity, and highly selective AVP screening methods.[10] Given that AVPs can target specific viruses, predicting their antiviral activity requires not only determining whether a peptide has antiviral potential but also accurately forecasting its efficacy against particular viral families or species. This dual prediction capability is

Received: February 13, 2025
Revised: May 23, 2025
Accepted: May 23, 2025
Published: June 6, 2025

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886

crucial for advancing AVP-based therapies, offering both scientific and practical benefits.

Extensive research has shown that machine learning (ML) methods are both efficient and cost-effective for predicting antiviral peptides (AVPs), attracting significant attention in recent years.[11−14] As the field has developed, various machine-learning-based AVP prediction models have been developed. For instance, Thakur et al.[15] developed AVPpred, a predictive tool that integrates amino acid profile and physicochemical characteristics using a support vector machine (SVM).[16] Chang and Yang[17] developed a random forest (RF)[18,19] model based on physicochemical properties, achieving excellent classification performance. Lissabet et al.[20] presented AntiVPP1.0 utilizing a random forest algorithm alongside features such as static charge, molecular weight, and hydrophilicity index for prediction. Schaduangrat et al.[21] created Meta-iAVP, an integrated classification method that combines the predicted outputs of six different machine learning methods to form an ensemble classifier. Although traditional machine learning methods have improved the prediction performance of antiviral peptides (AVPs) to some extent, they heavily rely on feature selection, which increases the dependence on domain expertise and limits their potential for generalization and automation. Recently, deep learning techniques have been increasingly applied to AVP prediction. For example, Li et al.[22] proposed DeepAVP, a dual-channel model that leverages deep learning to improve prediction performance.[23] Afterward, Akbar et al.[24] integrated discrete wavelet transform with k-segmentation techniques and employed Shapley additive explanations (SHAP) to identify the optimal features, which were subsequently validated using five machine learning classifiers. An ensemble model was then created by incorporating the predicted labels into a genetic algorithm. Most of these methods rely on data sets established by Thakur et al. in 2012. However, AVP databases have been updated since then, and most antimicrobial peptide databases now include AVPs. Current research predominantly focuses on the binary classification of AVPs. The AVP-IFT method proposed by Guan et al.[25] has achieved significant results in classifying specific antiviral activities, but it cannot more effectively distinguish between positive and negative samples in imbalanced data sets, limiting its practicality. This underscores the need for further optimization of existing models to handle complex cases more effectively.

Currently, contrastive learning has been widely adopted in sequence prediction tasks and has demonstrated promising prospects. The core idea of contrastive learning is to pull similar sample pairs closer while pushing dissimilar pairs farther apart in the feature space, thereby enhancing the feature extraction capability of neural networks. With its continuous development, an increasing number of contrastive-learning-based frameworks have been proposed for sequence modeling. For instance, Zhang et al. developed SiameseCPP,[26] which constructs a contrastive learning framework by pairing sequences with the same label as positive pairs and those with different labels as negative pairs, achieving outstanding performance. Yang et al.[27] designed a SimCLR-based framework, where two augmented views of the same sequence are treated as a positive pair, and other samples serve as negative pairs, for gene prediction tasks. Lee and Shin[28] introduced con_ACP, in which the same sequence is processed by two independent tokenizers to generate different index representations that form positive pairs. Guan et al. proposed a fragment-level contrastive learning strategy where embedding vectors of different fragments from the same sequence are treated as positive pairs, while those from different sequences serve as negative pairs, effectively improving the prediction performance for antiviral peptides. Although contrastive learning has been widely applied in most sequence prediction tasks, the potential impact of negative samples on the model performance has often been overlooked.

The core idea of transfer learning is to train a model on data related to the target task and transfer the learned knowledge to the target task. This approach helps reduce training time, lessens the reliance on large-scale labeled data, and enhances the model's generalization ability on new tasks. It is particularly suitable for scenarios in which the target task has limited samples or suffers from severely imbalanced class distributions. For instance, Hanson et al. proposed SPOT-MoRF,[29] which applies transfer learning at both the fully connected layer and the contextual modeling layer to generate ensemble predictions of residues. Xu et al. developed ACVPred,[30] pretraining the model on antiviral peptide (AVP) data and then fine-tuning it on a data-scarce ACVP data set, thereby achieving superior predictive performance. Tan et al.[31] leveraged the AlphaPept-Deep architecture,[32] pretraining it on a trypsin-derived data set to achieve accurate prediction of peptide half-lives. Given the extreme scarcity of AVPs targeting specific viruses, directly training models often encounters challenges such as insufficient samples and overfitting. Transfer learning, as an effective strategy for small-sample learning, pretrains models on large-scale AVP data sets and transfers the acquired knowledge to the identification of AVPs for target virus families or specific viruses, thereby significantly enhancing model performance in the second-stage tasks.

Considering real-life scenarios, there are typically more non-AVPs than AVPs, leading to inherently imbalanced data sets. If this issue is not properly addressed during model training, the model may become biased toward the majority class, ultimately compromising its predictive performance.[33] Therefore, designing models that can effectively handle imbalanced data sets is essential. By training models specifically on imbalanced data,[34] it is possible to enhance their ability to recognize minority classes, thereby improving their performance in practical applications.

Based on the above issues, we proposed an innovative tool for identifying antiviral peptides and their subclasses, named AVP-HNCL. The main contributions of this study are as follows:

(a) We improved the contrastive learning strategy by adopting a queue-based approach to select the top-k sequences as samples when forming negative pairs. This targeted selection of negative pairs for each anchor enhances the model's ability to learn sequence representations more effectively.

(b) In contrast, we alter the original sequence views through a multistep augmentation strategy, providing positive sample views for anchors in contrastive learning while incorporating large models and utilizing protein language models to understand and extract features from peptide sequences.

(c) We applied the concept of transfer learning in the second stage, initializing the model with the weights pretrained in the first stage. The model was then fine-tuned for antiviral peptides specific to different viruses,

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886

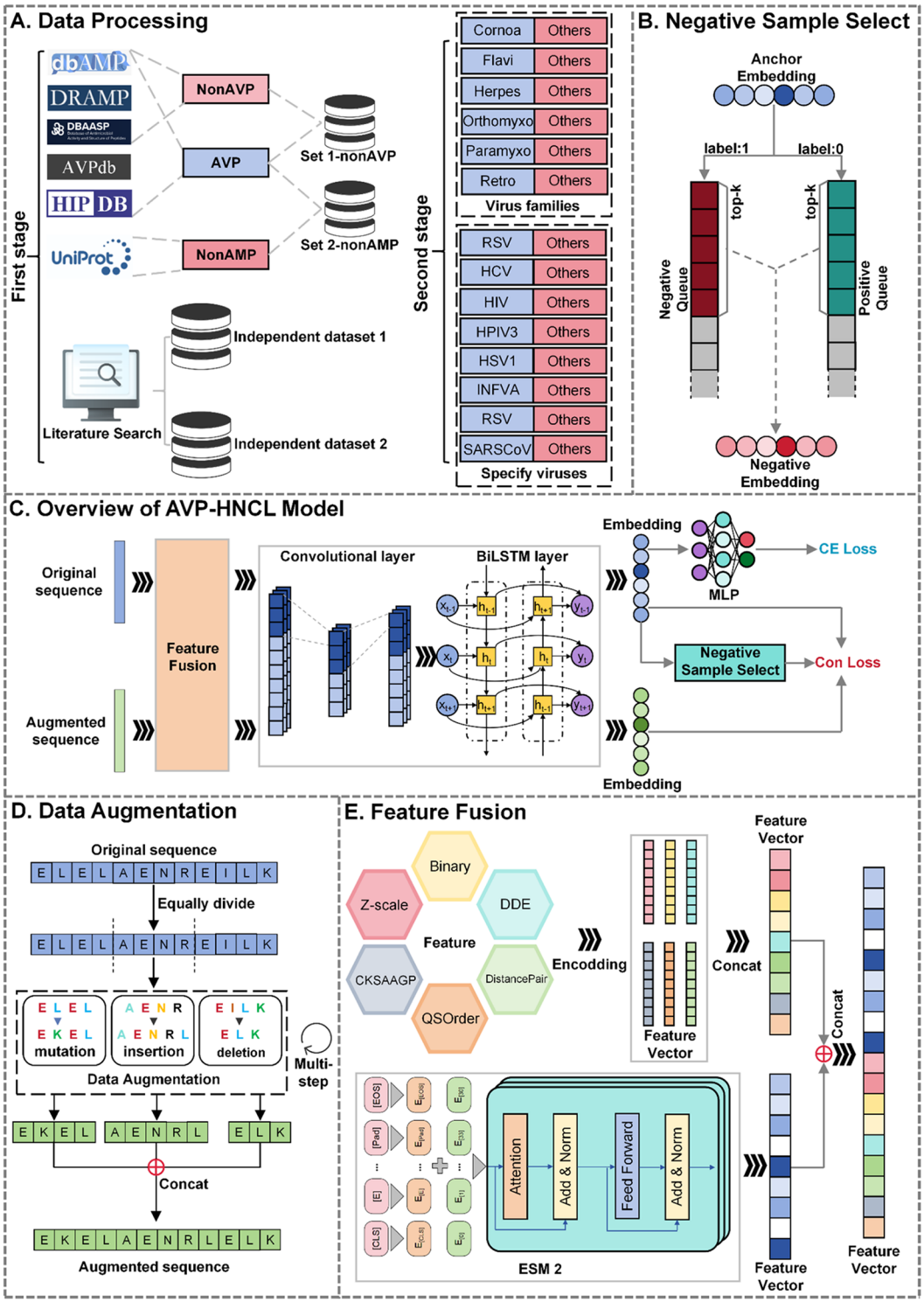


Figure 1. Framework of AVP-HNCL. (A) Data set processing: In the first stage, to assess the model's generalization ability, we selected four antiviral peptide data sets, two of which used antimicrobial peptides without antiviral activity as negative samples. In the second stage, to prevent label leakage during transfer learning, we redefined the training and test sets by categorizing antiviral peptides into different families and classes. (B) Negative sample select: We defined two separate queues to store data with different labels. Based on the label of the anchor, we selected the top-k sequences from the queue with the opposite label as negative samples, according to cosine similarity. (C) Overview of the AVP-HNCL model: In the model learning strategy, both augmented and original sequences were passed through feature fusion and then processed through convolutional layers and BiLSTM to capture local and global information, forming positive sample pairs. For the negative samples, a queue-based approach was used to improve sample selection, which optimizes the model's learnable parameters. The model was then trained using contrastive learning and cross-entropy loss. (D) Data augmentation module: The original sequences were enhanced using a multistep data augmentation approach, where mutation, insertion, and deletion strategies were applied to the data with a certain probability. The augmented sequences were then used to form positive sample pairs. (E) Feature fusion: Various feature encoding methods and protein language models were used to convert the sequences into vector embeddings.

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886

optimizing its ability to recognize specific virus families and individual viruses.

(d) Finally, we developed a dual-phase prediction system, as depicted in Figure 1. In the initial phase, the system determines whether a peptide qualifies as an AVP within complex sample sets. Subsequently, in the second phase, it classifies the AVPs into subclasses, encompassing six primary virus families (Coronaviridae, Flaviviridae, Herpesviridae, Orthomyxoviridae, Paramyxoviridae, and Retroviridae) and eight specific viruses (FIV, HCV, HIV, HPIV3, HSV1, IFVA, RSV, and SARS-CoV).

This study presents a more efficient method for the early identification of AVPs, significantly reducing labor costs. Compared with existing approaches, our model demonstrates substantial improvements in AVP prediction performance. Additionally, we achieved a high predictive accuracy for AVP subclasses, underscoring the robustness of our approach. By integrating multiple features, we enhanced the model's interpretability, providing deeper insights into AVP characteristics. This research also provides important guidance for the design and development of antiviral drugs in the future. With the continuous advancement of machine learning prediction tools, our model is well-positioned to play a pivotal role in developing innovative therapeutic strategies against various viruses.

METHODS

Data Set Processing. To predict antiviral peptides, this study utilized two AVP-IFT[15] data sets: non-AVPs and non-AMPs, both developed by Guan et al. These two data sets share the same 2662 positive samples, which were extracted from multiple databases, including AVPdb,[35] dbAMP,[36] DRAMP,[37] DBAASP,[38] and HIPdb.[39] However, the negative samples originate from different sources: the first data set's negative samples were extracted from dbAMP, DBAASP, and DRAMP, which do not include virus-specific entries, while the second data set's negative samples were obtained from the UniProt[40] database and filtered based on the following attribute keywords: "toxic", "membrane", "secretory", "defensive", "antibiotic", "anticancer", "antiviral", and "antifungal". While the positive samples are identical between the two data sets, the negative samples are distinct. To reduce redundancy, duplicate sequences were removed using the Cluster Database at High Identity with Tolerance (CD-HIT) tool, with a similarity threshold set at 40%.[41] This process resulted in a balanced data set, where the number of negative samples equaled the number of positive samples, yielding a total of 2662 sequences. Subsequently, the data set was split into a training set and a test set at a ratio of 4:1. Fivefold cross validation was performed on the training set, while the test set was used for final model performance evaluation. In our research, we call these two data sets "Set 1-nonAVP" and "Set 2-nonAMP". In addition, to evaluate the model's generalization capability, we employed two independent test sets: independent data set 1, which includes 604 experimentally validated AVPs and 452 non-AVPs, and independent data set 2, which contains 604 validated AVPs and 604 unvalidated non-AVPs. After removing sequences that contain invalid amino acids, the exact number of samples for the four data sets of first stage is detailed in Table 1.

Among AVPs, the majority of sequences are capable of targeting specific viruses and viral families. Based on the multifunctionality of AVPs, we classified the data into six viral families, including Coronaviridae, Retroviridae, Flaviviridae, Orthomyxoviridae, Paramyxoviridae, and Herpesviridae, and eight targeted viruses, including feline immunodeficiency virus (FIV), human immunodeficiency virus (HIV), hepatitis C virus (HCV), herpes simplex virus type 1 (HSV1), influenza A virus (INFVA), respiratory syncytial virus (RSV), and SARS-CoV. These viral families and targeted viruses were selected based on stringent criteria: each viral family contains at least 100 sequence records, and each targeted virus includes no fewer than 80 sequence records, ensuring the integrity of our training set and the feasibility of model development. Finally, to prevent data leakage during the use of transfer learning, we established six and eight multifunctionally classified data sets based on these viral families and targeted viruses, with detailed specifications provided in Tables 2 and 3. As a result, we obtained 14 distinct data sets covering six major virus families and eight specific viruses, as summarized in Tables 2 and 3. These data sets were also partitioned into training and test sets using the same 4:1 ratio.

Table 1. Summary of the First-Stage Data Set

data sets	training/test sets	positive samples	negative samples
Set 1-nonAVP	training	2129	2129
	test	553	553
Set 2-nonAMP	training	2129	2129
	test	553	553
independent data set 1	training	540	400
	test	60	44
independent data set 2	training	540	541
	test	60	58

Feature Fusion. In this study, we employed six sequence encoding methods alongside the ESM2 model to comprehensively extract features from peptide sequences.[42,43] The six encoding methods include Binary encoding, Z-scale encoding,[44] DistancePair encoding,[45] CKSAAGP encoding,[46] QSOrder encoding,[47,48] and DDE encoding.[49] Additionally, the ESM2 model was integrated to capture deep contextual and evolutionary information from the peptide sequences,[50] significantly enhancing the feature representation.

Binary Encoding. Each amino acid is uniquely encoded using a 20-dimensional binary vector, which effectively distinguishes it from the other 19 standard amino acids.

Z-scale Encoding. This approach measures the physicochemical properties of amino acids through z-score normalization and represents each amino acid as a five-dimensional vector.

DistancePair Encoding. It describes the pairwise distances between amino acids in a protein sequence, reflecting their relative positions and potential interactions. This encoding method is closely related to the structural properties of proteins.

CKSAAGP Encoding. The CKSAAGP encoding captures local patterns by partitioning overlapping amino acid groups and calculating their frequencies, aiding in the identification of functional regions or active sites in peptides.

QSOrder Encoding. The Quasi-Sequence Order (QSOrder) encoding integrates amino acid composition and positional information, effectively capturing both local and global sequence features.

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886

Table 2. Summary of Multifunctional Classified Data Sets—Viral family

viral family	Coronaviridae	Retroviridae	Herpesviridae	Paramyxoviridae	Orthomyxoviridae	Flaviviridae
positive samples	184	995	267	272	113	489
negative samples	2478	1667	2395	2390	2549	2173

Table 3. Summary of Multifunctional Classified Data Sets—Targeted Virus

targeted virus	FIV	HIV	HCV	HPIV3	HSV1	INFVA	RSV	SARS-CoV
positive samples	101	867	438	87	213	112	119	137
negative samples	2561	1795	2224	2575	2449	2550	2543	2525

DDE Encoding. The Dipeptide Deviation from Expected Mean (DDE) encoding reveals potential evolutionary, functional, or structural characteristics by comparing the actual frequency of dipeptides to their predicted frequency based on codon usage patterns.

ESM2 Encoding. ESM2 is a masked language model based on the Transformer architecture, trained by using a large number of protein sequences. During training, amino acids are masked and specific position mutations are evaluated, enabling the model to capture contextual information and long-range dependencies within protein sequences.

By combining the features extracted through six encoding methods and the embeddings generated by ESM2, each peptide sequence is transformed into a robust, multidimensional feature representation. This approach effectively captures contextual, physicochemical, and structural properties, providing a solid foundation for subsequent analyses. We used the iFeatureOmega package to obtain some sequence encodings, including DistancePair, CKSAAGP, and QSOrder.[51] Further details on peptide encodings can be found in the Supporting Information.

Data Augmentation. We propose an enhancement strategy for the raw amino acid sequences to assist in forming positive sample pairs in contrastive learning. Unlike the typical approach of handling features during encoding, we choose to enhance the original amino acid sequences directly, which better improves the model's generalization ability. Specifically, the sequences are evenly divided into three equal-length segments, and for each segment, one of three different enhancement operations (mutation, insertion, or deletion) is applied sequentially.

Mutation involves selecting an index within a sequence segment with a certain probability and replacing the amino acid at that position with another amino acid chosen with a probability of p. **Insertion** adds a new amino acid at a randomly selected position within the segment, increasing the diversity and complexity of the sequence. **Deletion** removes a few amino acids randomly from the segment, simulating data loss scenarios.

During the enhancement process, we set the number of augmentations as a hyperparameter, performing all three operations once for each enhancement. This multistep augmentation strategy not only preserves the main structural features of the original sequence but also significantly enhances the data set's diversity and the contrastive learning capabilities by introducing various types of mutations. The augmented segments are then recombined into a complete sequence for subsequent model training. By applying these random operations based on sequence segments, the data augmentation module effectively reduces the distributional gap between the augmented and original sequences, thereby improving the model's adaptability and robustness to unseen data. This method significantly improves the model's generalization performance while ensuring that the augmented data remain highly relevant to the original data.

Feature Engineering. This study employed a CNN[52] + BiLSTM[53] architecture for feature engineering to effectively learn both local and global information from sequence data.[54] In the convolutional layer, we used 1D convolution, which offers high computational efficiency when processing sequential information. Compared with more complex convolutional methods, Conv1D features a simpler structure with fewer parameters, reducing both computational load and memory consumption while improving model training efficiency. At the same time, BiLSTM, with its bidirectional propagation mechanism, can capture both forward and backward dependencies in sequences, thus better handling long-term dependencies and enhancing the model's contextual understanding. The combination of these two methods fully leverages their respective strengths, significantly improving the overall model performance.

We employed a one-dimensional convolutional neural network (1D-CNN) to process the concatenated encoded sequences and extract convolutional features using convolutional kernels. For an input sample X, the mathematical representation of the convolution layer is

$$y^k = \mathrm{ReLU}\left(\mathrm{Batchnorm}\left(\sum_{p=0}^{C_{in}-1}\sum_{q=0}^{K-1} \omega_{pq}^k \bullet x_{p+(q-\lfloor \frac{K}{2} \rfloor)} + b^k\right)\right) \quad (1)$$

where $\mathrm{ReLU}(x) = \max(0, x)$ and ω^k represents the weight tensor of the kth convolutional kernel, with dimensions $C_{in} \times K$, where C_{in} is the number of input channels and K is the kernel size.

The features obtained after the convolution operation are transposed to match the input format required by the BiLSTM.[36] LSTM, as a specialized type of recurrent neural network, excels at processing long sequential data and capturing dependencies across varying intervals and delays. BiLSTM further enhances this capability by simultaneously capturing forward and backward dependencies in the input sequence, thereby providing a more comprehensive representation of the sequence information. The core formulas for LSTM feature extraction are

$$f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_f) \quad (3)$$

$$\tilde{C}_t = \tanh(W_C \bullet [h_{t-1}, x_t] + b_C) \quad (4)$$

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886

$$C_t = f_t \bullet C_{t-1} + i_t \bullet \tilde{C}_t \tag{5}$$

$$o_t = f_t \bullet C_{t-1} + i_t \bullet \tilde{C}_t \tag{6}$$

$$h_t = o_t \bullet \tanh(C_t) \tag{7}$$

where f_t, i_t, and o_t represent the activation vectors for the forget gate, input gate, and output gate at each time step t, managing the sequential information flow. $\tilde{C}_t$ and C_t represent the potential and true cell states at time t, which store the network's memory, and h_t represents the hidden state vector that encapsulates the information processed up to time t. Weight matrices W_f, W_i, W_C, and W_o and bias vectors b_f, b_i, b_C, and b_o are essential elements that govern the operations of the gates and the updates of the cell states within the LSTM. These processes are controlled by the sigmoid function and the hyperbolic tangent function (tanh), which manage the activation and adjustment of cell states.

In the BiLSTM architecture, outputs are derived from both forward and backward passes, referred to as h_t^{forward} and h_t^{backward}, respectively. These hidden states are subsequently concatenated to form the final output.

After the BiLSTM layer, adaptive max pooling is applied to the features to extract the most significant features. The mathematical representation is as follows

$$\text{pooling}(H) = \max(H[t_1, 1], H[t_2, 1], ..., H[t_T, 1]) \tag{8}$$

where H represents the feature vector output by the BiLSTM layer and T is the total number of time steps.

Contrastive Learning with the Queue-Based Negative Sampling Strategy for Model Training. During the model training process, we performed feature engineering on both the positive sample sequences generated by the proposed data augmentation strategy and the original sequences after Sequence encodings.[55] Each individual sequence learns local context information and global features of the sequence during the feature engineering process. We also developed a queue-based negative sample pair selection strategy for contrastive learning. The goal of contrastive learning is to build a latent space where the distance between the anchor and positive samples is minimized while the distance between the anchor and negative samples is maximized. To achieve this, we used cosine similarity as the metric to measure the distance between samples. The formula is defined as follows

$$\text{sim}(x, y) = \frac{x \bullet y}{\|x\| \|y\|} \tag{9}$$

where x represents the embedding of the anchor sequence and y denotes the corresponding positive or negative sample.

In the process of negative sample selection, for each batch, we first divided the individual sample sequences into two queues based on their labels. Specifically, when the label of the anchor is 1, we retrieved the top-k samples from the queue with label 0, which are most similar to the anchor sample but have opposite labels. The main idea behind this strategy is to select negative samples that are similar to the anchor but have opposite labels, which helps to effectively increase the distance between positive and negative samples in the latent space. This approach not only improves the quality of the negative samples but also aids the model in learning more discriminative features. As a result, this negative sample selection method enhances the construction of the latent space, optimizes model training, and improves both the learning efficiency and generalization ability. The mathematical formulation is expressed as follows

$$N^* = \bigcup_{j=1}^{k} (\text{top} - \text{k}(\text{sim}(x, x_i^-), k) | x_i^- \in Q) \tag{10}$$

where $\cup_{j=1}^{k}$ represents the aggregation of the top-k most similar samples into a set and Q is a queue that stores the embedding of the negative sample sequence. The contrastive loss function is defined as follows

$$L_{\text{contrastive}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\max\left(0, \frac{N^*}{k} - \text{sim}(x, x^+) + \text{margin}\right)}{T} + \lambda \bullet |T| \tag{11}$$

where x^+ represents the feature representation of the augmented sample, margin is a key parameter in contrastive learning used to control the degree of separation between positive and negative samples in the feature space, and temperature (T) is another important hyperparameter in contrastive learning, used to scale the similarity distribution. $|T|$ is the absolute value of the temperature parameter, which penalizes the magnitude of the temperature parameter to prevent it from becoming too large or too small, ensuring the stability of model training. λ is the regularization coefficient, used to control the influence of the regularization term on the loss function.

Furthermore, to enhance classification performance, we incorporated class weights into the cross-entropy loss function.[56] By assigning different weights to each class, we can direct the model's attention more effectively to the minority class, thereby mitigating the issue of class imbalance. This approach not only improves the overall performance of the model but also ensures more reliable predictions across all categories. By fine-tuning the class weights, we further boost the model's ability to handle imbalanced data, resulting in better performance on such tasks. The weighted cross-entropy loss function used in our study is as follows

$$L_{\text{CE-w}} = -\frac{1}{N} \sum_{i=1}^{N} [\alpha_1 y_i \log(p_i) + \alpha_0 (1 - y_i) \log(1 - p_i)] \tag{12}$$

where N represents the total number of samples, y_i denotes the label of sample x_i, p_i is the predicted probability of sample x_i being positive, α_1 is the weight for the positive class, and α_0 is the weight for the negative class.

By combining these two, the model is able to enhance feature representation while effectively addressing class imbalance, ultimately improving overall classification performance and boosting the model's generalization ability. The final loss is defined as

$$L_{\text{total}} = L_{\text{CE-w}} + \alpha L_{\text{contrastive}} \tag{13}$$

where the α coefficient controls the proportion of contrastive loss in determining the final loss.

Transfer Learning. In this study, we designed a two-stage task. In the first stage, we determined whether a sequence is an antiviral peptide (AVP), and in the second stage, we classified the antiviral activity against specific types of viruses. Due to the limited data available for antiviral peptides targeting specific viruses, we adopted the concept of transfer learning to enhance the model's ability to learn under small sample conditions.

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886

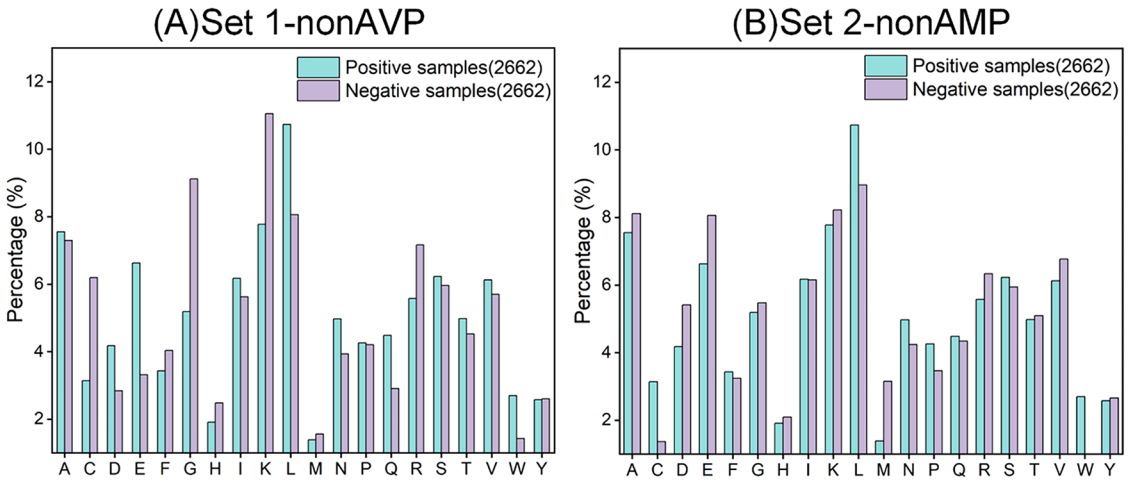


Figure 2. Amino acid composition of Set 1-nonAVP and Set 2-nonAMP in the first stage. (A) Percentage composition of amino acids in Set 1-nonAVP. (B) Percentage composition of amino acids in Set 2-nonAMP. (The *x*-axis represents the distribution of the 20 standard amino acids, and the *y*-axis indicates the percentage composition of each amino acid.)

During the pretraining phase, we trained the model using Set 2-nonAMP consisting of 2662 positive and negative samples. After training, we saved all of the parameters of the feature engineering module, which consists of CNN layers and BiLSTM layers.

In the fine-tuning phase, we constructed an independent binary classification task for each viral activity type and performed fine-tuning based on the parameters of the feature engineering module learned during pretraining. Additionally, we added a multilayer perceptron (MLP) module during fine-tuning to refine the classification of specific AVPs. To ensure the model's generalization ability, we strictly adhered to the principle that test samples in the fine-tuning phase should not appear in the pretraining phase training set. Specifically, we first labeled the positive and negative samples in the fine-tuning data set and split them into training and test sets in a 4:1 ratio. Then, all positive and negative samples from the fine-tuning data set were treated as positive samples for the pretraining phase, while non-AVP negative samples from Set 2 were split in a 4:1 ratio and added to the pretraining data set. This data processing strategy ensures that the samples predicted by the model during fine-tuning are sequences it has never encountered before, thereby effectively enhancing the model's generalization ability and accuracy in classifying specific antiviral peptides.

Performance Metrics. In this study, we employed various metrics to evaluate the performance of the proposed method, including accuracy (ACC), sensitivity (SN), specificity (SP), Matthews correlation coefficient (MCC),[57] the geometric mean of sensitivity and specificity (G-mean),[58] the area under the precision-recall curve (AUPRC), the area under the receiver operating characteristic curve (AUROC), and F1 score.[59] These metrics provide a comprehensive evaluation of the model's predictive performance and robustness from multiple perspectives

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{14}$$

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{15}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{16}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{17}$$

$$\mathrm{Gmean} = \sqrt{\mathrm{SN} \times \mathrm{SP}} \tag{18}$$

$$\mathrm{AUPRC} = \int_0^1 \mathrm{Precision} d(\mathrm{Recall}) \tag{19}$$

$$\mathrm{AUROC} = \int_0^1 \mathrm{TPR} d(\mathrm{FPR}) \tag{20}$$

$$\mathrm{F1} = 2 \bullet \frac{\mathrm{Precision} \bullet \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{21}$$

where TP is true positive, TN is true negative, FP is false positive, FN is false negative, and TPR (true positive rate) represents the proportion of actual positive samples correctly identified by the model. It indicates the model's sensitivity or recall ability to detect positive instances. FPR represents the proportion of actual negative samples that are incorrectly identified as positive by the model. It reflects the model's tendency to produce false alarms.

Model Interpretation. In this study, we employed two interpretable methods, Uniform Manifold Approximation and Projection (UMAP)[60] and Deep Learning Important Features (DeepLIFT),[61] to enhance model transparency and provide valuable biological insights. UMAP was used to visualize the distribution of features in the model's intermediate layers by reducing high-dimensional data to a two-dimensional space, facilitating the intuitive analysis of different AVP sample categories. DeepLIFT, implemented through the Captum library,[62] assessed the contribution of input features relative to reference values, determining the significance of amino acid positions. This approach effectively identified key amino acid residues that significantly impact AVP predictions and elucidated their decision-making processes. These findings not only improve the model's interpretability but also offer data-driven support for the functional design and optimization

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886

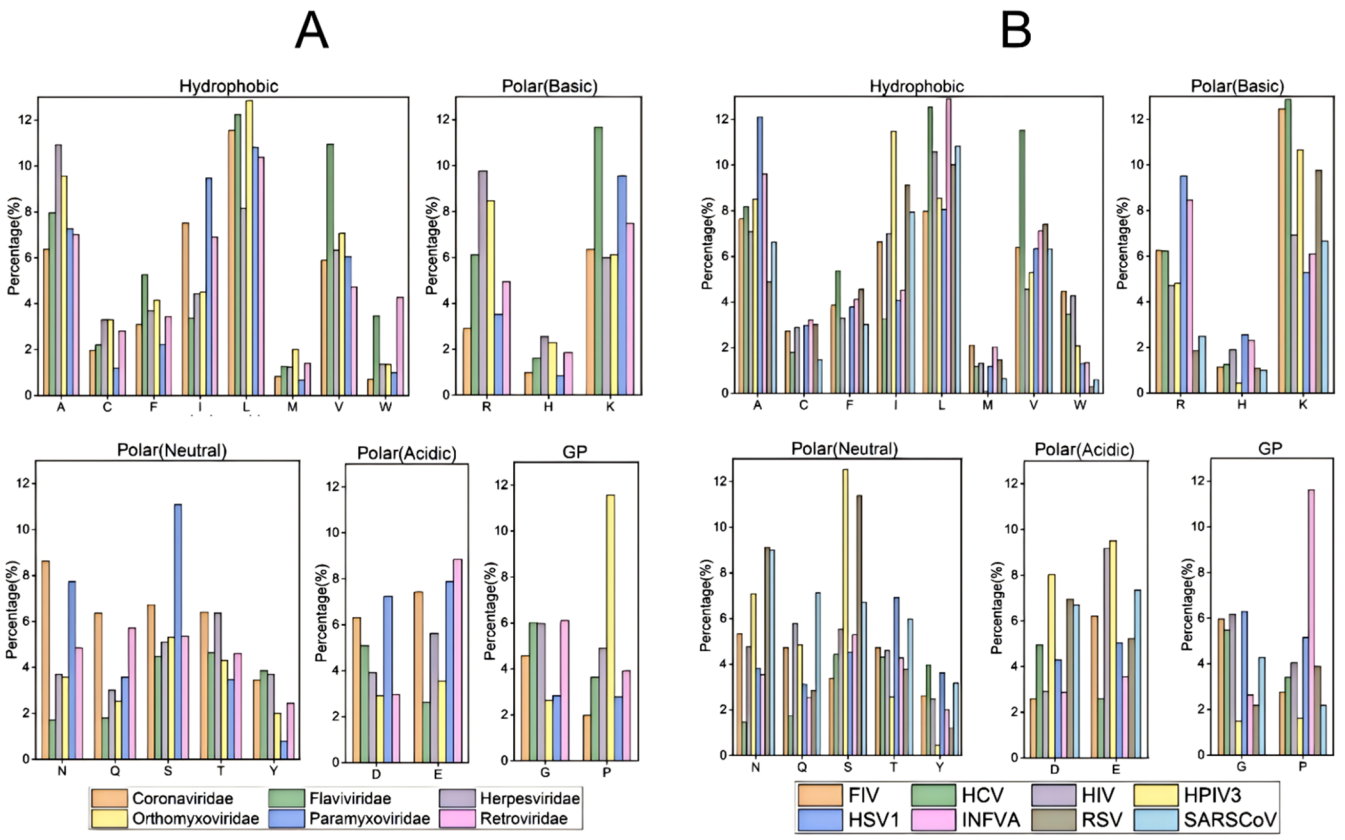


Figure 3. Amino acid composition of data sets in the second stage. (A) Amino acid composition of antiviral peptides (AVPs) targeting six major viral families: Coronaviridae, Retroviridae, Herpesviridae, Paramyxoviridae, Togaviridae, and Flaviviridae. (B) Amino acid composition of antiviral peptides (AVPs) targeting eight representative viruses: feline immunodeficiency virus (FIV), hepatitis C virus (HCV), human immunodeficiency virus (HIV), human parainfluenza virus type 3 (HPIV3), herpes simplex virus type 1 (HSV1), influenza A virus (INFVA), respiratory syncytial virus (RSV), and severe acute respiratory syndrome coronavirus (SARS-CoV).

of antiviral peptides. In conclusion, the combined application of UMAP and DeepLIFT enhances the model's transparency and reliability, providing a robust theoretical foundation for antiviral peptide design and advancing further research and development in the biomedical field.

RESULTS AND DISCUSSION

Overview of Amino Acid Distributions. To explore the differences between the samples, we analyzed the amino acid profiles of two data sets (Set 1-nonAVP and Set 2-nonAMP). As shown in Figure 2A, the positive samples in Set 1-nonAVP are rich in acidic amino acids, such as aspartic acid (D) and glutamic acid (E), and hydrophobic amino acids, including leucine (L) and tryptophan (W). In contrast, the negative samples predominantly contain basic amino acids, such as arginine (R), lysine (K), and glycine (G). As shown in Figure 2B, the positive samples in Set 2-nonAMP exhibit higher levels of hydrophobic amino acids, such as leucine (L) and tryptophan (W), and polar amino acids, like cysteine (C), while the negative samples show a higher proportion of acidic amino acids, particularly aspartic acid (D) and glutamic acid (E). Overall, the analysis suggests that AVP samples typically have a higher proportion of hydrophobic amino acids, especially leucine (L) and tryptophan (W), while non-AVP samples exhibit a higher proportion of basic amino acids, particularly arginine (R), lysine (K), and glycine (G). These differences may reflect the structural requirements of antiviral peptides (AVPs), which could enhance their interactions with target viruses and improve their antiviral activity.

To further investigate AVP subclasses, we classified amino acids based on their physicochemical properties and evaluated the compositional differences of AVPs across various viral families and target viruses. As shown in Figure 3A, AVPs targeting the Retroviridae family are enriched in glycine (G) and glutamic acid (E), which may be linked to the unique membrane structures, protein interactions, and key processes in the viral lifecycle. AVPs from the Orthomyxoviridae family contain higher levels of leucine (L), potentially associated with the closed structural nature of these viruses and their interaction with host cell membranes through hydrophobic amino acids. In the Flaviviridae family, AVPs show significantly higher levels of valine (V) and lysine (K), which may relate to the viral envelope structure, invasion mechanisms, and membrane protein affinity. AVPs from the Paramyxoviridae family are characterized by high selenite (S) content, likely playing a role in antigen epitopes and receptor binding regions. AVPs targeting the Herpesviridae family are rich in Arginine (R), a positively charged amino acid that may be involved in viral binding to host cell receptors. However, AVPs from the Coronaviridae family do not show any particular amino acid enrichment, suggesting a more diverse amino acid requirement. As shown in Figure 3B, AVPs targeting FIV are enriched in tryptophan (W), an aromatic amino acid important for maintaining protein structure and stability. In contrast, HCV shows significantly elevated levels of valine (V) and lysine (K),

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886

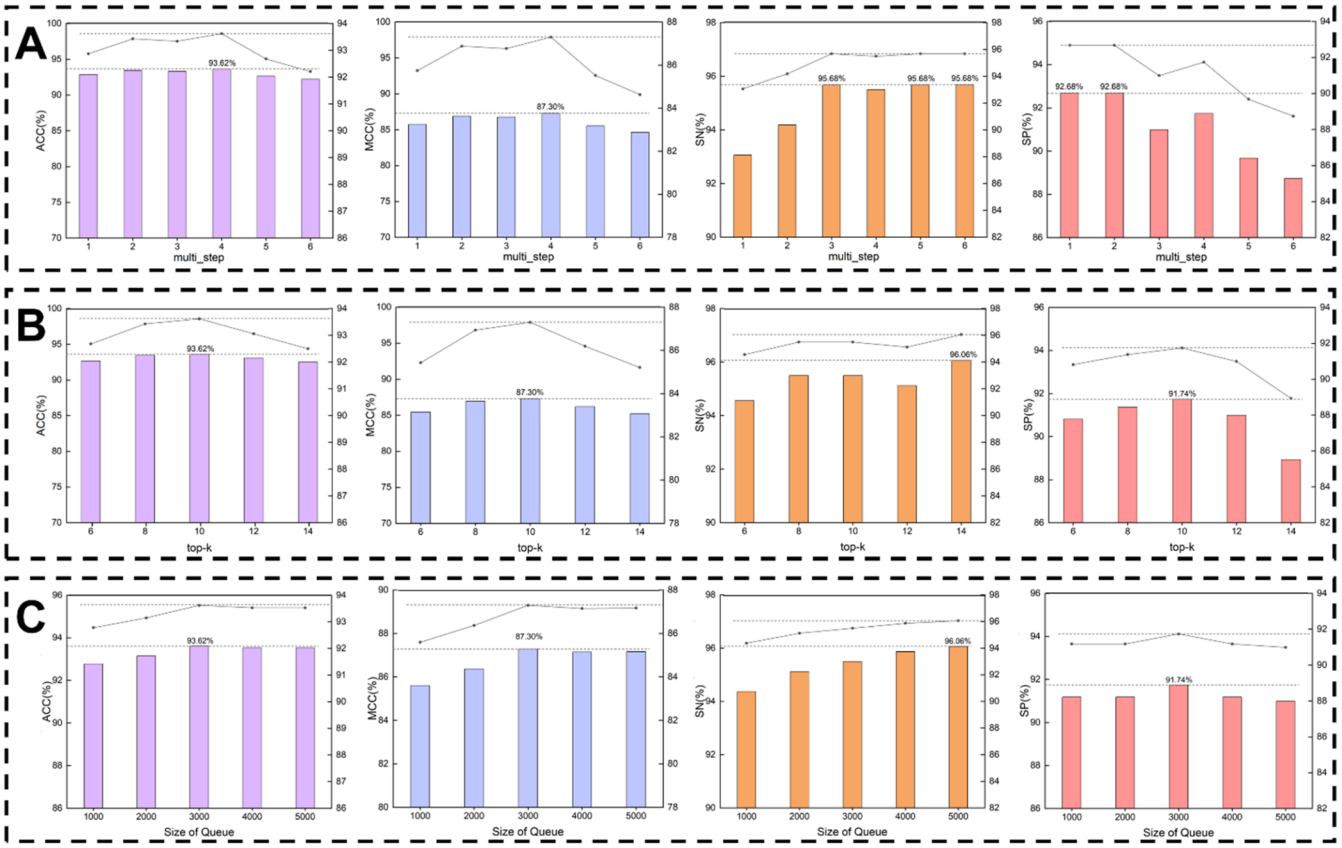


Figure 4. Results of the hyperparameter tuning experiments using Set 1-nonAVP. (A) Model performance with different numbers of multistep augmentations. (B) Values of various metrics under different k selection values. (C) Impact of different queue sizes on model performance. (From left to right, the metrics are ACC, MCC, SN, and SP, respectively. The bar plots illustrate the differences in performance metrics under various hyperparameter settings. The line plots depict the trends of these metrics as the hyperparameters change.).

which may relate to the hydrophobicity of its membrane and core proteins and their localization in the host cell membrane. INFVA is notable for its higher proline (P) content, which plays a role in protein conformation changes and turns, which is likely crucial for AVP function. HPIV3 and RSV exhibit higher levels of serine (S), correlating with receptor binding and invasion mechanisms. HPIV3 lacks cyclohemonine (C), methionine (M), and phenylalanine (F). HIV, HSV1, and SARS-CoV do not show significant amino acid enrichment. Additionally, our observations indicate that AVPs typically contain minimal amounts of cysteine (C), methionine (M), phenylalanine (F), tryptophan (W), and histidine (H). This low abundance may result from the chemical properties of these amino acids, affecting the AVP stability, solubility, or binding ability with target viruses.

These findings highlight the distinct amino acid composition of AVPs and provide insights into their antiviral mechanisms. AVPs targeting different viral families and specific viruses exhibit preferences for particular amino acids, which likely relate to their structural features, receptor binding mechanisms, and viral entry modes.

Tuning Key Parameters in the Proposed Model for Improved AVP Prediction. In our proposed model architecture, we investigated the impact of several key hyperparameters on model performance: (A) the *k* value for negative sample selection, (B) the queue size, and (C) the number of multistep data augmentation steps. First, the *k* value represents the number of negative samples selected during training. Increasing the number of negative samples enhances the complexity of the negative samples in the latent space, improving the model's ability to distinguish between positive and negative samples. However, a small *k* value may result in insufficient diversity of negative samples, while a large *k* value may introduce too many irrelevant samples, reducing the efficiency and stability of the training process. The queue size controls the available data pool for the negative samples. The queue's capacity determines the number of embeddings it can store, which, in turn, affects the diversity and comprehensiveness of the samples used for contrastive learning. A larger queue enhances the sample diversity, allowing the model to capture more intricate features. However, this comes at the cost of increased memory usage and computational demands. On the other hand, a smaller queue may limit the effectiveness of contrastive learning, as the model may struggle to adequately learn a wide range of features. Finally, the number of multistep data augmentation steps influences the diversity between anchor and positive samples. By applying multiple transformations to the input sequence, we generated augmented sequences, providing different views of the same sequence. This augmentation strategy improves the model's generalization ability, enabling it to handle data variations more robustly. However, too many augmentation steps can introduce excessive noise, disrupting the learning process.

In Figure 4, we demonstrate the hyperparameter tuning process using Set 1-nonAVP. It can be observed that as the *k* value increases, both the MCC and ACC curves fluctuate with the SN curve. This indicates that as the *k* value increases, the model becomes better at capturing the differences between

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886

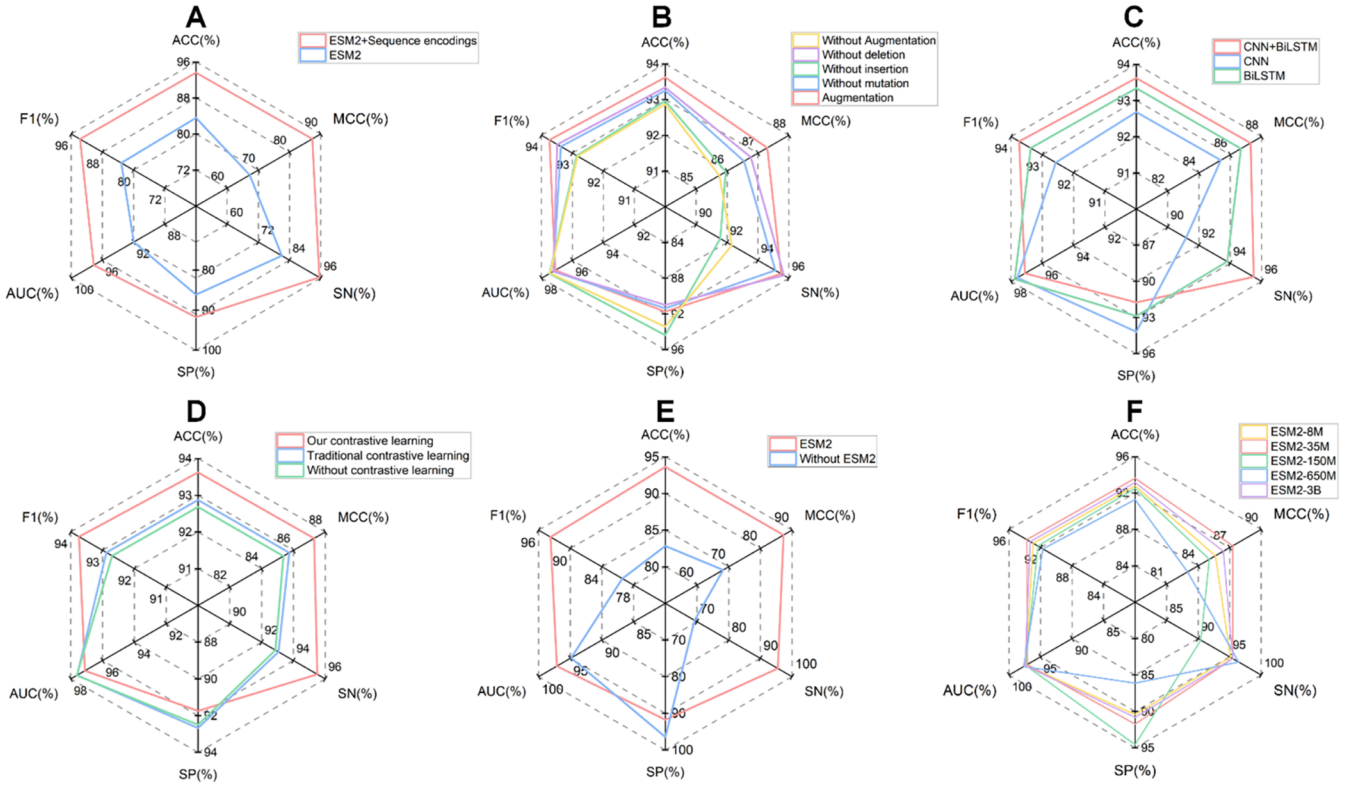


Figure 5. We conducted a series of ablation experiments using Set 1-nonAVP: (A) we validated the effectiveness of six sequence encoding methods, (B) we compared the model performance with and without different augmentation strategies in the multistep augmentation process, (C) we evaluated the performance differences among various model architectures, (D) we evaluated the performance gap when replacing our proposed contrastive learning negative sampling strategy with traditional strategies, compared to the model without contrastive learning, (E) we assessed the effectiveness of the pretrained protein model ESM2 for sequence encoding, and (F) we verified the optimality of the selected ESM2 model (ESM2_t12_35M_UR50D).

positive and negative samples. However, when the k value exceeds 10, the excessive number of negative samples leads to confusion in the model. Additionally, in the augmentation process, the number of multistep augmentations shows a clear impact on the selection of positive samples, as seen in the SN curve of Figure 4A. The more the augmentation steps applied, the better the model's robustness when facing unseen data. Regarding the queue size, since we fixed it at a certain length and each batch consists of a specific number of samples, data will be evicted from the queue in later batches. Thus, an appropriate queue size helps us to select suitable negative samples from the data pool. In Table S1, we list the hyperparameters used in the first stage for different data sets, all of which were selected based on model performance evaluation. Furthermore, in the second stage, we also provide the parameter selection for different viral families and antiviral activities in Tables S2 and S3.

Impact of Improved Contrastive Learning and the Effectiveness of Other Modules. To comprehensively assess the importance of our proposed negative sampling strategy and other modules, including feature encoding, augmentation strategies, and the use of pretrained protein large models, we conducted ablation experiments: (i) comparing the performance of the ESM model encoding versus ESM combined with six other feature encodings, (ii) validating the effectiveness of augmentation strategies, (iii) comparing model performance across different architectural designs, (iv) validating the effectiveness of contrastive learning, (v) validating the effectiveness of the ESM2 model, (vi) verifying the optimality of the selected ESM2 model, and (vii) comparing the effectiveness of second-stage pretraining.

Our experimental validation clearly shows that combining the ESM2 model with six other feature extraction methods outperforms using the ESM2 model alone, with an improvement of 0.4−1.5 units across five evaluation metrics, as shown in Figure 5A. Additionally, we conducted a separate classification task for each sequence embedding and compared their performance, with detailed results presented in Table S4. This improvement can be attributed to the integration of ESM2 with other feature extraction methods, which enhanced the model's ability to distinguish between positive and negative samples. Figure 5B demonstrates a significant improvement in model performance when three sequence augmentation strategies (mutation, insertion, and deletion) were incorporated. Although there was a slight decline in SP, the model showed significant improvements in four key metrics (ACC, MCC, SN, and F1), with particularly strong performance in MCC and SN. These results highlight the crucial role of sequence augmentation strategies in enhancing the model's generalization ability. To further validate the effectiveness of each augmentation strategy, we removed one of the strategies at a time. The results showed that each individual strategy remained effective, further confirming its contribution to model performance. By generating diverse training samples, these strategies help the model better handle variability and complexity in the data. Overall, the results emphasize the complementary and synergistic effects of the three sequence augmentation strategies. Combining these strategies signifi-

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886

Table 5. Summary of Performance on Independent Viral Family Data Sets with and without Transfer Learning

viral family	pretrain	ACC	SN	SP	MCC	G-mean
Coronaviridae	no	0.9475	**0.9189**	0.9496	0.7036	0.9341
	yes	**0.9700**	**0.9189**	**0.9738**	**0.8002**	**0.9460**
Flaviviridae	no	0.9362	**0.8571**	0.9540	0.7929	0.9043
	yes	**0.9456**	0.8469	**0.9678**	**0.8180**	**0.9054**
Herpesviridae	no	0.8968	0.8889	0.8977	0.6152	0.8933
	yes	**0.9437**	**0.9444**	**0.9436**	**0.7582**	**0.944**
Orthomyxoviridae	no	0.8874	0.8696	0.8882	0.438	0.8789
	yes	**0.9587**	**1.0000**	**0.9569**	**0.6993**	**0.9782**
Paramyxoviridae	no	0.9906	**0.9818**	0.9916	0.9509	**0.9867**
	yes	**0.9925**	0.9636	**0.9958**	**0.9595**	0.9796
Retroviridae	no	0.9024	**0.8844**	0.9132	0.793	**0.8987**
	yes	**0.9062**	0.8593	**0.9341**	**0.7985**	0.8959

cantly boosts the model's robustness and generalization ability, further verifying the key role of sequence augmentation in improving AVP classification model performance.

To validate the advantage of our chosen network architecture, we compared it with two network architectures: CNN and BiLSTM. As shown in Figure 5C, the CNN and BiLSTM model surpasses individual CNN or BiLSTM models in processing AVP sequences, effectively capturing the complex sequence features of antiviral peptides. We validated the effectiveness of the contrastive learning method we developed; we introduced traditional contrastive learning and tested it on Set 1-nonAVP. As shown in Figure 5D, the model using our contrastive learning method achieved an ACC of 0.9362 and an MCC of 0.8730, while the model using traditional contrastive learning had an ACC of 0.9287 and an MCC of 0.8574, and the model without contrastive learning achieved an ACC of 0.9268 and an MCC of 0.8537. The results suggest that traditional contrastive learning did not lead to a significant improvement in model performance. This may be attributed to the relatively simple architecture of the model used in this study. It is possible that more pronounced effects could emerge with increased model complexity or parameter count. Nevertheless, under the same conditions, our proposed model demonstrated a significant performance improvement compared to both the traditional contrastive learning approach and the baseline without contrastive learning. This improvement can be attributed to our method's focus on hard negative samples—specifically those highly similar to anchor samples. This approach helps the model better distinguish between positive and negative samples, thereby enhancing its learning capacity. Furthermore, the contrastive loss function emphasizes the relationships between samples, aiding in the learning of key features and further boosting classification performance.

To evaluate and compare the performance of the ESM2-35M model with a model without ESM2 on Set 1-nonAVP, specific metrics were analyzed. The analysis, with its detailed results presented in Figure 5E, indicates that the ESM2-35 M model outperforms the model without ESM2 on all key performance metrics, demonstrating the model's powerful feature extraction capabilities. By capturing deep-level information within sequences, ESM2 provides richer feature representations compared to traditional methods. This is particularly valuable for the challenging task of accurate classification on Set 1-nonAVP, where distinguishing between classes is difficult. The ESM2 model ensures the extraction of crucial features, effectively improving the accuracy and reliability of AVP prediction, further confirming its critical role in enhancing classification performance. To validate that the selected ESM2 model is the optimal choice, we conducted a comparative analysis of the performance impact of different scales of ESM2 models. As shown in Figure 5F, we selected ESM2_t12_35M_UR50D as the preferred model. Specifically, on Set 1-nonAVP, this model achieved an accuracy of 0.9362, a specificity of 0.9174, and an MCC score of 0.8730. Although it slightly underperformed in sensitivity compared to ESM2_t33_650M_UR50D and in specificity compared to ESM2_t30_150M_UR50D, the ESM2_t12_35M_UR50D model demonstrated a more balanced and superior overall performance. These results highlight that the selected ESM2_t12_35M_UR50D model is the optimal choice. Although in large-scale protein pretraining models, an increase in the number of parameters enhances the model's ability to learn deeper features, it also leads to overfitting. To balance model efficiency and performance, we selected a pretraining model with 35 M parameters.

Validating the Impact of Transfer Learning on the Enhancement of the Second Phase. In the specific classification of AVPs in the second phase, we employed transfer learning by transferring the parameters of the pretrained model from the first phase to the second phase. This approach helps address the issue of sparse data set samples for AVPs targeting specific viruses. Tables 5 and 6 present performance comparisons between models with and without transfer learning. The results show that the model achieved significantly improved performance when transfer learning was applied.

To further demonstrate the importance of this strategy in the second phase, we visualized the shift in sequence positions in the latent space using UMPA plots, comparing the results with and without transfer learning. In Figure 6, we present that the main text includes three key figures, while additional UMAP visualizations are provided in the Supporting Information (Figures S1 and S2). The results demonstrate that pretraining facilitates a clearer distinction between positive and negative samples in high-dimensional spaces, resulting in more compact feature clusters. This underscores pretraining's capability to identify essential features and fine-tune feature distributions, thereby enhancing classification accuracy and the model's ability to generalize in virus classification tasks. Moreover, fine-tuning enables the model to swiftly adjust to new data distributions, minimizing the need for extensive training data. This two-step methodology boosts classification performance, particularly in situations with limited data sets or intricate sample distributions, showcasing notable robustness

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886

Table 6. Summary of Performance on Independent Targeted Virus Data Sets with and without Transfer Learning

targeted virus	pretrain	ACC	SN	SP	MCC	G-mean
HCV	no	0.9493	**0.9432**	0.9506	0.8343	0.9469
	yes	**0.9719**	**0.9432**	**0.9775**	**0.9007**	**0.9602**
HIV	no	0.9156	**0.908**	0.9192	0.8128	**0.9136**
	yes	**0.9193**	0.8736	**0.9415**	**0.8163**	0.9069
HPIV3	no	0.8293	**0.9444**	0.8252	0.3471	0.8828
	yes	**0.9906**	**0.9444**	**0.9922**	**0.8697**	**0.968**
HSV1	no	0.9137	0.8605	0.9184	0.6033	0.8889
	yes	**0.9325**	**0.8837**	**0.9367**	**0.6656**	**0.9098**
INFVA	no	0.9381	0.913	0.9392	0.5836	0.926
	yes	**0.97**	**0.9565**	**0.9706**	**0.7412**	**0.9635**
RSV	no	0.9756	**0.9583**	0.9764	0.7826	**0.9673**
	yes	**0.9962**	0.9167	**1.0000**	**0.9556**	0.9574
SARS-CoV	no	0.9587	**0.9286**	0.9604	0.7063	**0.9443**
	yes	**0.9869**	0.8929	**0.9921**	**0.8704**	0.9412
FIV	no	0.8668	**0.9524**	0.8633	0.4236	0.9067
	yes	**0.9606**	0.9048	**0.9629**	**0.656**	**0.9334**

and versatility. The proposed framework effectively combines pretraining, feature encoding, and contrastive learning, leading to substantial improvements in classification outcomes. This novel approach addresses critical challenges in antiviral peptide classification and functional type prediction, allowing for the efficient extraction of key features even when sample sizes are constrained.

Model Performance and Comparison with Existing AVP Prediction Tools. To evaluate the generalization ability and fairness of AVP-HNCL, we conducted fivefold cross validation on the Set 1-nonAVP data set and compared it with the AVP-IFT method.[25] As shown in Table 7, AVP-HNCL outperformed AVP-IFT across key performance metrics on the Set 1-nonAVP data set, achieving higher accuracy (0.9213 vs 0.8940) and MCC score (0.8436 vs 0.7931). These results indicate that AVP-HNCL, by incorporating a queue-based negative sampling strategy, can more effectively distinguish between different classes with high similarity, thereby achieving superior classification performance. To more comprehensively assess our model's generalization capability, and considering the influence of the validation strategy proposed by Huseyin Tunc[63] in the drug-isolate-fold change (DIF) framework, we aimed to incorporate a similar evaluation approach. However, since the sequence data we used lacked explicit mutation information, we adopted a stratified cross-validation strategy instead. We first applied K-means clustering to perform unsupervised classification, partitioning the samples into n clusters for various values of n. Based on these clusters, we then carried out stratified sampling by assigning 80% of each cluster's samples to the training set and the remaining 20% to the validation set, thereby preserving the original distribution

Table 7. Comparison of the Performance of Existing Methods Based on Fivefold Cross Validation on the Set 1-nonAVP Data Set

method	ACC	SN	SP	MCC
AVP-IFT	0.8940	0.8670	**0.9210**	0.7931
our method	**0.9213**	**0.9357**	0.9070	**0.8436**

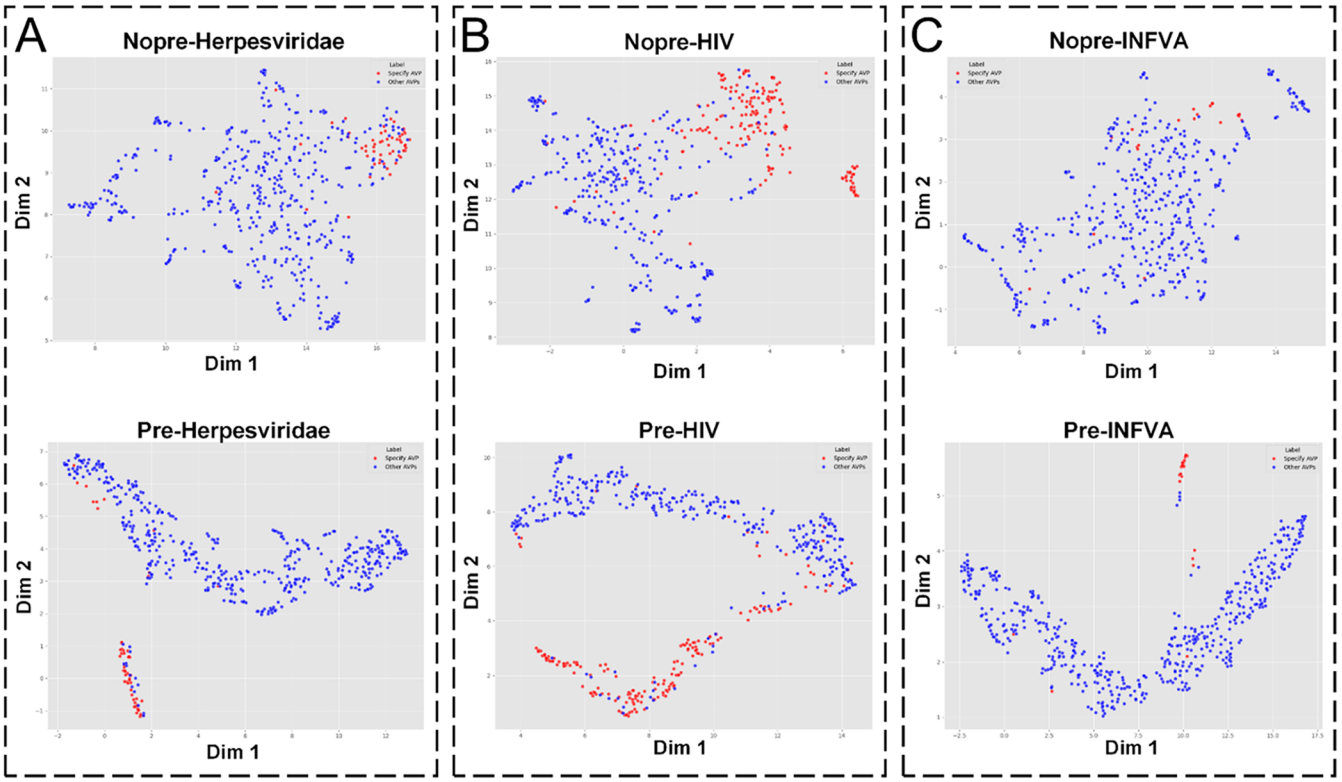


Figure 6. UMAP visualization of performance comparison between pretraining and nonpretraining in the second stage. (A) Herpesviridae. (B) HIV. (C) INFVA. (The top figure represents no pretraining, while the bottom figure represents pretraining; red dots represent specific antiviral peptides, while blue dots represent other antiviral peptides.)

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886

and ensuring a more robust evaluation. Table 8 presents the performance metrics for n = 2, 3, 4, and 5. As the number of clusters increases, the overall performance steadily improves, peaking at n = 5 with an accuracy (ACC) of 0.9322 and a Matthews correlation coefficient (MCC) of 0.8644. These results demonstrate that finer-grained stratification yields a more thorough assessment of the model's generalization ability.

Table 8. Comparison of the Performance of Existing Methods Based on Stratified Cross Validation on the Set 1-nonAVP Data Set

n	ACC	SN	SP	MCC
2	0.9156	0.9188	0.9123	0.8312
3	0.9145	0.9304	0.8983	0.8294
4	0.9287	0.9304	0.9269	0.8573
5	0.9322	0.9282	0.9362	0.8644

In order to demonstrate the effectiveness of our model, we compared the AVP-IFT method on Set 1-nonAVP and Set 2-nonAMP. On Set 1-nonAVP, as shown in Table 9, our method achieved higher scores across all metrics, including accuracy (0.9362 vs 0.9240), sensitivity (0.9550 vs 0.9343), specificity (0.9174 vs 0.9137), and MCC score (0.8730 vs 0.8482). These improvements demonstrate that our method is more effective in distinguishing AVPs from Set 1-nonAVP under complex conditions. On Set 2-nonAMP, our method demonstrated outstanding performance, achieving perfect scores across all metrics, including accuracy, sensitivity, specificity, and MCC, all reaching 1.0000, surpassing the AVP-IFT method. These results clearly highlight the significant advantages of our method in terms of predictive performance and generalization ability, further demonstrating its potential for practical applications.

Table 9. Comparison of the Performance of Existing Methods on Set 1-nonAVP and Set 2-nonAMP

Data set	method	ACC	SN	SP	MCC
set 1-nonAVP	AVP-IFT	0.9240	0.9343	0.9137	0.8482
	our method	**0.9362**	**0.9550**	**0.9174**	**0.8730**
set 2-nonAMP	AVP-IFT	0.9934	0.9944	0.9925	0.9869
	our method	**1.0000**	**1.0000**	**1.0000**	**1.0000**

Additionally, to validate the model's generalization ability, we compared it with other models on two independent test sets, including AVPpred,[15] Chang's method,[17] AntiVPP1.0,[20] Meta-iAVP,[21] DeepAVP,[22] PreTP-Stack,[64] and iAMPpred.[65] The comparative results are visualized in Figure 7, with the detailed values available in Tables S5 and S6. In the independent data set 1, our model outperformed all comparative methods, achieving the highest accuracy of 0.9615, which is 2−16 percentage points higher than other methods. Our model also demonstrated significant advantages in sensitivity and specificity, reaching values of 0.9667 and 0.9545, respectively. Regarding the MCC (Matthews correlation coefficient) score, our model achieved a score of 0.92, significantly surpassing other models. Similarly, in the independent data set 2, our model showed strong performance, with an accuracy of 0.9661, sensitivity of 0.9500, specificity of 0.9828, and an MCC score of 0.93, all of which were superior to those of other comparative methods. Although the

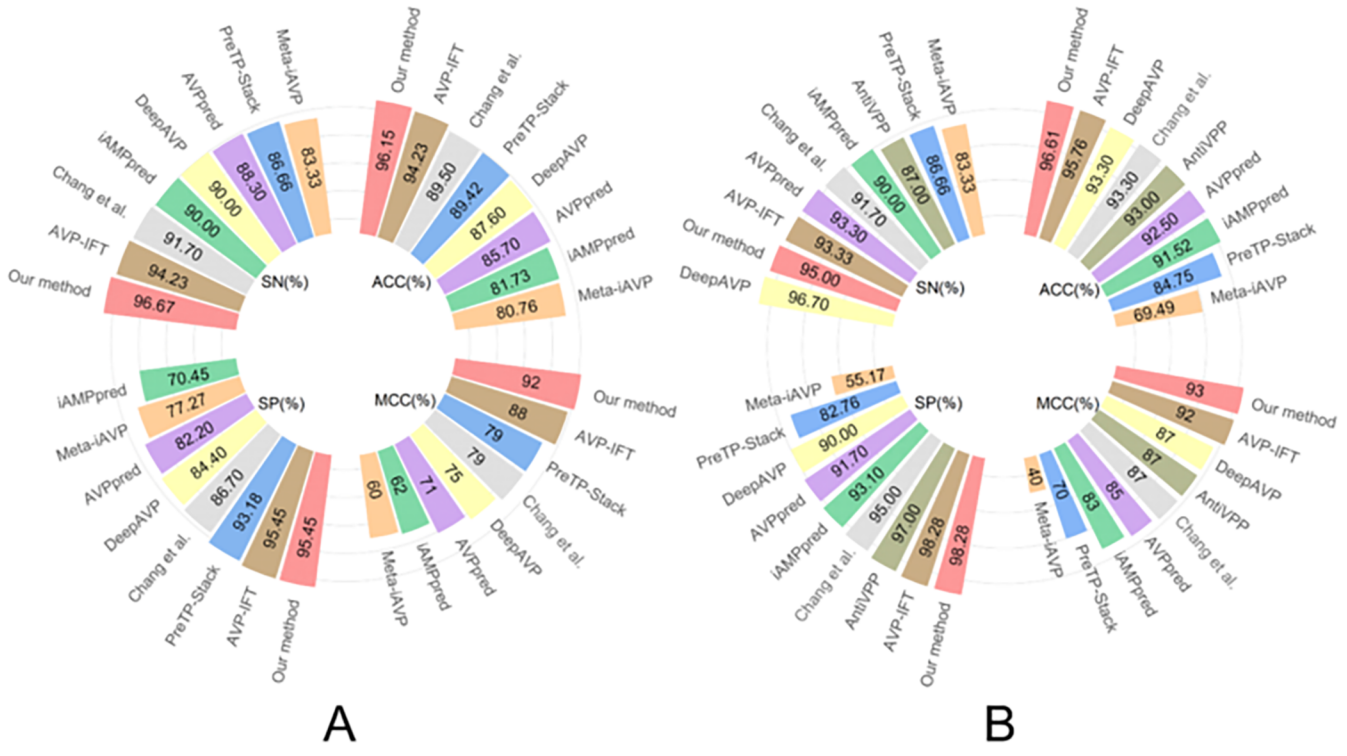


Figure 7. Comparison of our proposed model with SOTA models on two independent test sets (ACC, MCC, SN, SP) (the radar charts illustrate an overall increasing trend across performance metrics). (A) Comparison on the independent data set 1 (**AVP-HNCL** achieved the highest scores across all performance metrics). (B) Comparison on the independent data set 2 (except for SN, where **DeepAVP** achieved the highest score, **AVP-HNCL** outperformed in all other metrics).

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886

Table 10. Summary of Performance on Independent Viral Family Data Sets

viral family	method	ACC	SN	SP	MCC	G-mean
Coronaviridae	AVP-IFT	0.8979	**0.9348**	0.8952	0.5709	0.9148
	our method	**0.9700**	0.9189	**0.9738**	**0.8002**	**0.9460**
Flaviviridae	AVP-IFT	0.8378	0.9754	0.8070	0.6406	0.8872
	our method	**0.9456**	0.8469	**0.9678**	**0.8180**	**0.9054**
Herpesviridae	AVP-IFT	0.8483	0.8806	0.8447	0.5199	0.8625
	our method	**0.9437**	**0.9444**	**0.9436**	**0.7582**	**0.9440**
Orthomyxoviridae	AVP-IFT	0.8036	0.9655	0.7962	0.3654	0.8768
	our method	**0.9587**	**1.0000**	**0.9569**	**0.6993**	**0.9782**
Paramyxoviridae	AVP-IFT	0.9625	0.9118	0.9682	0.8152	0.9396
	our method	**0.9906**	**0.9636**	**0.9937**	**0.9498**	**0.9786**
Retroviridae	AVP-IFT	0.9009	0.9197	0.8897	0.7954	**0.9046**
	our method	**0.9062**	0.8593	**0.9341**	**0.7985**	0.8959

Table 11. Summary of Performance on Independent Targeted Virus Data Sets

targeted virus	method	ACC	SN	SP	MCC	G-mean
HCV	AVP-IFT	0.8442	0.8818	0.8369	0.5913	0.8591
	our method	**0.9719**	**0.9432**	**0.9775**	**0.9007**	**0.9602**
HIV	AVP-IFT	0.8818	0.8241	0.9089	0.7294	0.8654
	our method	**0.9193**	**0.8736**	**0.9415**	**0.8163**	**0.9069**
HPIV3	AVP-IFT	0.9641	**0.9999**	0.9629	0.6786	**0.9813**
	our method	**0.9906**	0.9444	**0.9922**	**0.8697**	0.9680
HSV1	AVP-IFT	0.8452	**0.9630**	0.8350	0.5161	0.8967
	our method	**0.9325**	0.8837	**0.9367**	**0.6656**	**0.9098**
INFVA	AVP-IFT	0.8504	**0.9643**	0.8454	0.4131	0.9029
	our method	**0.9700**	0.9565	**0.9706**	**0.7412**	**0.9635**
RSV	AVP-IFT	0.9414	**0.9999**	0.9386	0.6388	**0.9688**
	our method	**0.9962**	0.9167	**1.0000**	**0.9556**	0.9574
SARS-CoV	AVP-IFT	0.9414	**0.9999**	0.9386	0.6388	**0.9688**
	our method	**0.9869**	0.8929	**0.9921**	**0.8704**	0.9412
FIV	AVP-IFT	0.8559	**0.9200**	0.8534	0.3863	0.8861
	our method	**0.9606**	0.9048	**0.9629**	**0.656**	**0.9334**

sensitivity was slightly lower than that of DeepAVP, our model effectively balanced the sensitivity and specificity, thereby eliminating prediction bias.

In the second-phase comparison, considering the imbalance in the data set, the evaluation incorporated G-mean, AUPRC, and AUROC metrics. The results for other metrics can be found in Tables S7 and S8. Table 10 presents the model's performance on viral family classification using an imbalance learning strategy. The model achieved remarkable results across the six viral families with accuracies ranging from 0.90 to 0.99, G-mean scores between 0.89 and 0.97, AUPRC scores ranging from 0.89 to 0.97, and AUROC scores between 0.93 and 0.99. Notably, the Paramyxoviridae family achieved the best classification results, with an accuracy of 0.9906, a G-mean score of 0.9786, an AUPRC of 0.9765, and an AUROC of 0.9917.

For the classification of specific target viruses, Table 11 summarizes the model's performance. The model effectively classified target viruses, achieving accuracies between 0.91 and 0.99, G-mean scores between 0.90 and 0.96, AUPRC scores ranging from 0.80 to 0.97, and AUROC scores between 0.94 and 0.99. Particularly for respiratory syncytial virus (RSV), the model achieved an accuracy of 0.9962, a G-mean score of 0.9574, an AUPRC of 0.9482, and an AUROC of 0.9813. These results demonstrate the model's outstanding performance in accurately identifying AVPs targeting specific viral families or viruses.

Overall, our model performed excellently in distinguishing AVPs and non-AVPs, which can be attributed to several key factors: First, we incorporated a comprehensive set of peptide features, including composition, sequence order information, physicochemical properties, and evolutionary information, providing multidimensional insights into peptide characteristics for the model. Second, we employed advanced contrastive learning techniques that helped the model learn subtle but crucial differences by separating anchor samples from highly similar negative samples, thereby improving its ability to distinguish AVPs from non-AVPs. Additionally, by utilizing the ESM2 protein language model, the model effectively captured contextual and evolutionary information, enabling a more in-depth understanding of the peptide sequences.

Model Interpretation. To gain a deeper understanding of the behavior and effectiveness of the antiviral peptide classification model, we utilized UMAP as a key visualization technique. UMAP projects high-dimensional data into a two-dimensional space, allowing a clear examination of the distribution of antiviral peptide sequence features.[66] This method highlights the clustering and separation of AVPs and non-AVPs, providing valuable insights into the model's ability to accurately distinguish between these two categories. Figure 8A presents the interpretable results of the model.

We performed multilayer UMAP visualizations for each layer of the AVP-HNCL model to understand how it distinguishes

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886

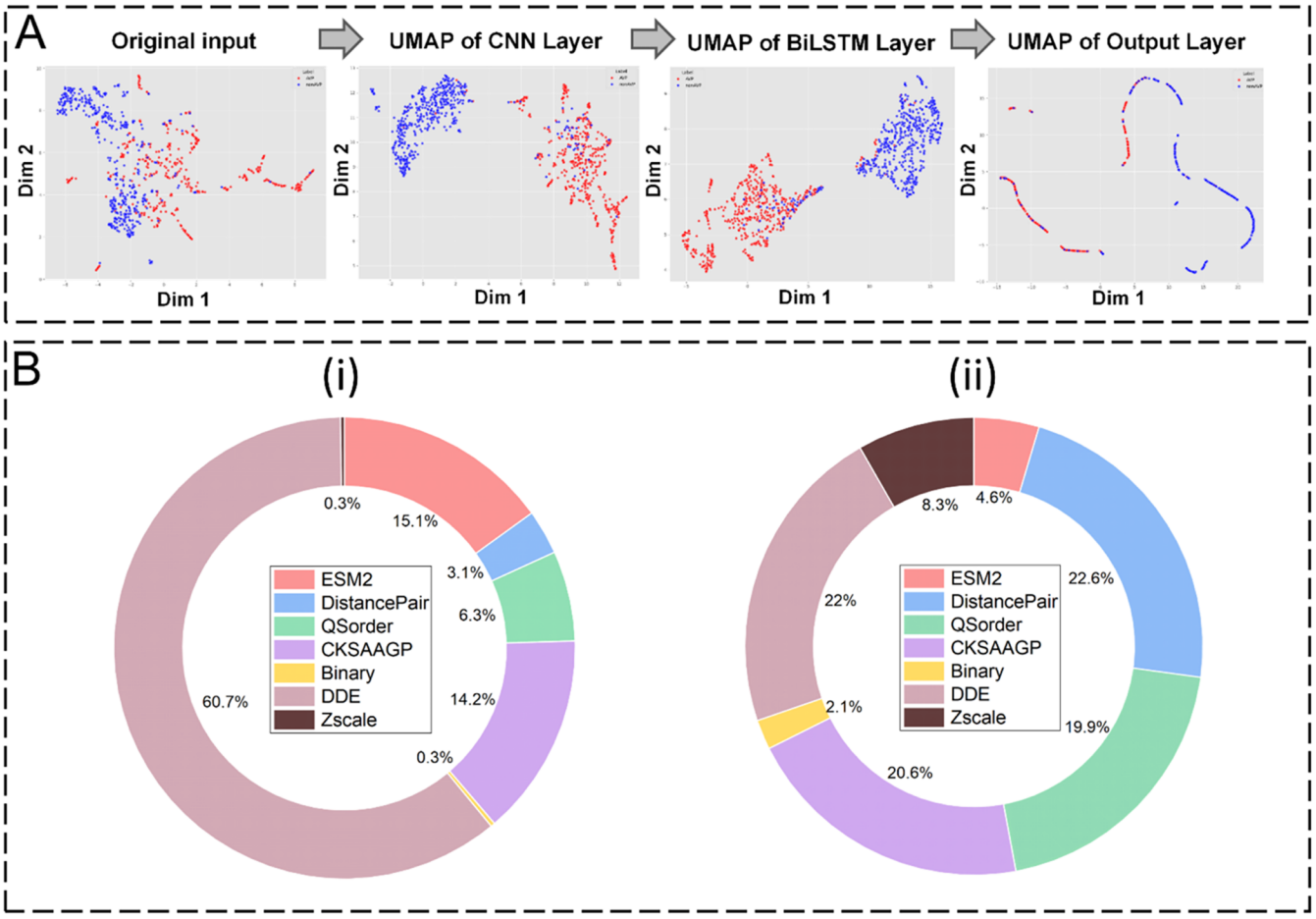


Figure 8. Presentation of the results of the interpretability analysis experiment. (A) Visualization of the discrimination capability in intermediate layers of the AVP-HNCL architecture (red dots represent antiviral peptides, and blue dots represent nonantiviral peptides). (B) Analysis of the overall contribution ratio of each feature to the prediction results and the normalized average contribution associated with their dimensionality.

the activity of antiviral peptides. Figure 8A illustrates the evolution of data points from the input layer to the output layer during the prediction process. Initially, the data points are clustered together, reflecting the complexity of the data. Through the CNN layer, the data points gradually separate as key local features are extracted. The BiLSTM layer further improves separation by capturing long-range correlations. Finally, the integration of the FC layer results in clearer classifications, enabling the output layer to effectively distinguish between AVP and non-AVP. This demonstrates the model's high accuracy in predicting antiviral peptides.

In this study, we applied the Deep Learning Important Features (DeepLIFT) method to interpret the features encoded by the sequences. Figure 8B presents the overall and average contributions of each descriptor during the initial identification phase. The results show that the DDE descriptor is particularly effective in capturing antiviral activity, as evidenced by its significant overall impact and high average importance. This effectiveness may stem from the marked differences in the frequency of specific dipeptides in antiviral peptides (AVPs) compared with their expected values. These frequency patterns likely represent key biochemical or structural features that distinguish AVPs from non-AVPs. Furthermore, the CKSAAGP descriptor also shows high overall and average contributions, suggesting that common subsequences within AVPs may represent crucial regions for antiviral activity. These subsequences likely correspond to conserved motifs or structural elements that play an essential role in interactions with viral targets.

Although the DistancePair and QSOrder descriptors have smaller feature dimensions, leading to lower contributions to the overall impact, they still exhibit high average importance. This suggests that while their overall influence is limited, these descriptors capture highly specific and critical information that enhances the model's ability to distinguish AVPs. DistancePair likely reflects spatial relationships between key residues, while QSOrder captures significant effects related to the sequence order.

Protein function is typically linked to the arrangement of amino acids, where the position of specific amino acids plays a crucial role. To quantitatively analyze the positional importance of amino acids within sequences, we employed DeepLIFT to interpret the results. This approach enables us to uncover the potential biological significance of amino acid positions in functional predictions.

Through statistical analysis, we found that the length distribution of antiviral peptides (AVPs) is primarily within the range of 0−60. To achieve a balanced distribution of lengths between positive and negative samples, we divided these sequences into six groups: 0−10, 10−20, 20−30, 30−40, 40−50, and 50−60. Unlike previous studies that primarily focused on interpreting both antiviral and nonantiviral

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886

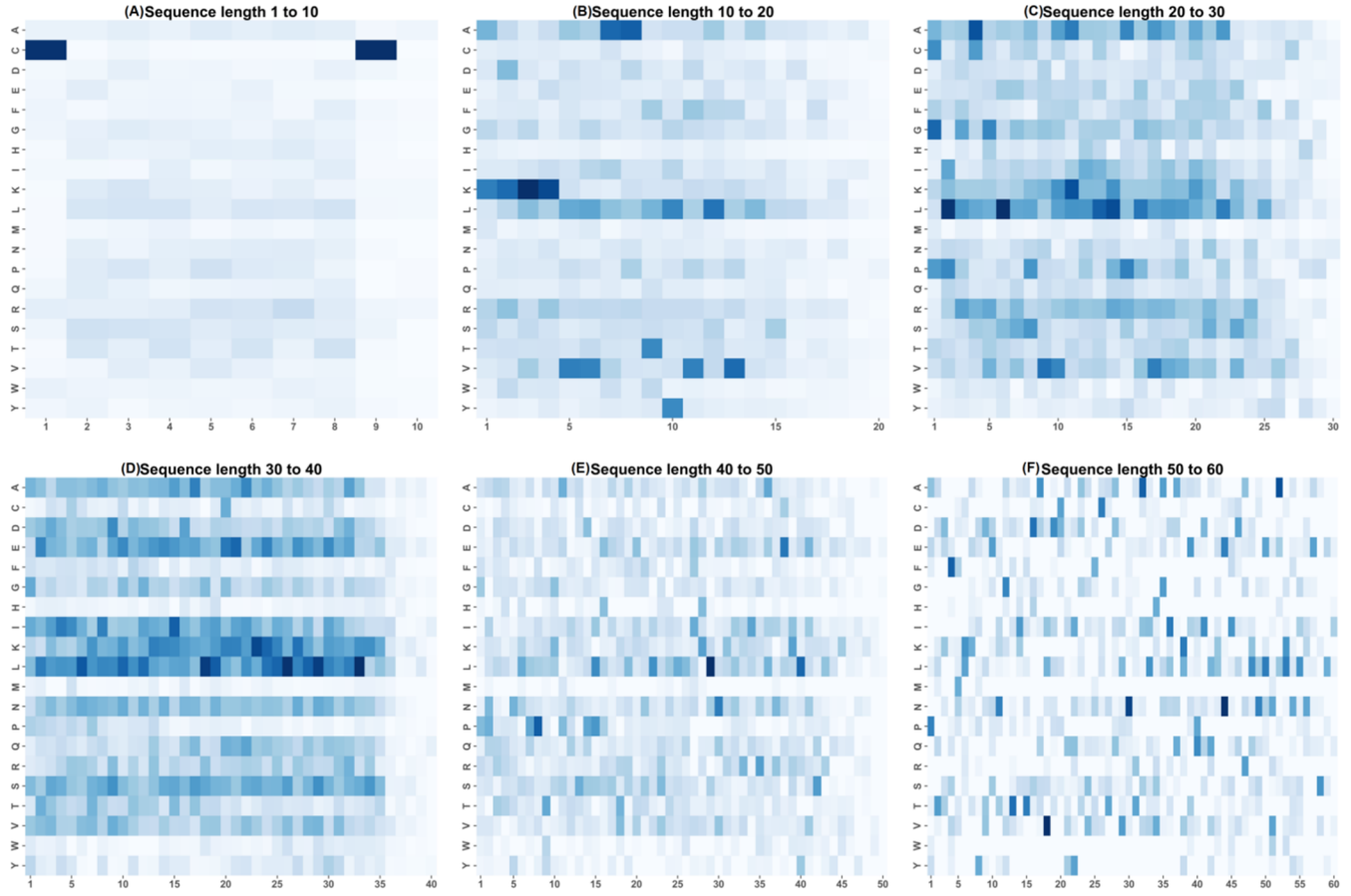


Figure 9. Importance of amino acid positions across various sequence length intervals. (A) Sequence length between 0 and 10. (B) Sequence length between 10 and 20. (C) Sequence length between 20 and 30. (D) Sequence length 30 and 40. (E) Sequence length between 40 and 50. (F) Sequence length between 50 and 60. (The darker the blue, the higher the importance; the lighter the blue, the lower the importance.)

sequences, our research specifically focused on the interpretability of antiviral peptides. This targeted approach allows for a more detailed and precise understanding of the key features contributing to the functionality of AVPs.

Figure 9 illustrates the impact of different amino acid positions on model predictions. In the heatmap, the steeper the blue gradient, the more significant the impact on AVP prediction. As shown in Figure 9A, for sequences with lengths between 0 and 10, amino acid C at positions 1 and 9 has a significant influence on AVP predictions. Amino acid C, a common structural stabilizer, may influence the peptide chain's three-dimensional structure or its interaction with viral targets through its unique chemical properties, such as the formation of disulfide bonds, thus becoming an important structural feature in AVP functionality. In Figure 9B–D, for sequences with lengths of 10–20, 20–30, and 30–40, we observed that amino acids K and L contribute significantly to the prediction results. K typically participates in electrostatic interactions, while L exhibits hydrophobicity, suggesting that the AVP functionality may be mediated by specific amino acid interactions. In Figure 9E, for sequences with lengths between 40 and 50, amino acid P at position 8 and amino acid L at position 29 show significant effects on AVP predictions. Proline (P) has a unique cyclic structure that may help stabilize AVPs or form specific structural domains, while leucine (L) may play a key role in interactions between the peptide chain and the membranes. Similarly, in Figure 9F, for sequences with lengths between 50 and 60, amino acid V (valine) at position 18 and amino acid N (asparagine) at position 39 have a significant influence on AVP prediction. The hydrophobicity of valine likely enhances the integration of AVPs with membranes, while the polarity of asparagine may aid in binding to specific targets.

The analysis of sequences in different length segments from 0 to 60 shows that for shorter peptides (such as the 0–10 length segment), the importance of amino acid C is more pronounced. This suggests that in shorter peptides, amino acids with higher reactivity or structural specificity (such as C) are more likely to influence function. For longer peptides (such as the 50–60 length segment), the influence of amino acids V and N indicates that, as peptide length increases, the interactions and spatial configuration of amino acids begin to have a more complex effect on AVP predictions.

Web Interface. To facilitate antiviral peptide prediction for researchers and users, we developed the proposed framework into a user-friendly online platform. Through this web server, users can simply input protein sequences directly or upload files in FASTA format to obtain prediction results without requiring any prior knowledge of machine learning or deep learning. The AVP-HNCL web server is freely accessible at: http://www.bioai-lab.com/AVP-HNCL.

■ CONCLUSIONS

Antiviral peptides exhibit a broad range of functional activities against various viruses, making the identification of these peptides and their specific subclasses crucial for the develop-

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868–5886

ment of targeted antiviral therapies. In this study, we propose a novel two-stage predictive framework for identifying antiviral peptides and their subclasses. This method offers several notable advantages: first, it can predict multiple functional types of antiviral peptides, thereby expanding its range of applications; second, we propose an innovative contrastive learning method that significantly improves the model's ability to distinguish challenging negative samples, enabling more precise discrimination between samples from different classes with high sequence similarity, and the multistep augmentation strategy effectively aggregates anchor samples with their augmented counterparts in the feature space, thereby improving the model's overall generalization capability; additionally, by employing a feature fusion strategy, the model captures more comprehensive features, improving prediction accuracy. Combined with UMAP and DeepLIFT techniques, the model's interpretability is enhanced, making the prediction results more intuitive and easier to understand. Meanwhile, our imbalance learning strategy aligns well with real-world application scenarios.

Although our model has achieved promising results, the current negative sample selection strategy has some limitations and significant room for improvement. For example, while the model can effectively distinguish negative samples highly similar to anchor sequences, during contrastive learning, some of these negative samples may already be clearly differentiated from the anchor samples, yet still get repeatedly used due to their high similarity. This phenomenon wastes significant computational resources. Moreover, we have yet to integrate structural information on the sample, which also limits the potential for further optimization of the model.

In the future, we plan to introduce negative sample selection methods based on attention mechanisms and uncertainty evaluation. By dynamically adjusting the importance weights of negative samples, we aim to prioritize those with higher learning value, thus improving training efficiency, avoiding resource waste, and further enhancing the model's learning capacity. Meanwhile, we also intend to incorporate structural information on the samples and conduct multimodal contrastive learning, fully integrating sequence and structural features to further improve the model's discriminative power and generalization performance.

■ ASSOCIATED CONTENT

Data Availability Statement

We have provided some data in the supporting files; please be sure to check them. Code and data sets of this study are available at https://github.com/kongjian408/AVP-HNCL.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c00306.

Detailed feature encoding methods from main text with mathematical formulas (including Binary, *Z*-scale, DistancePair, CKSAAGP, QSOrder, DDE encodings); Figures S1 and S2: UMAP visualizations of transfer learning impact on prediction in the second stage; Tables S1−S3: optimal hyperparameter settings; Table S4: optimality analysis of selected feature encodings; Tables S5 and S6: performance comparisons; Tables S7 and S8: prediction performance on antiviral peptide subtypes in the second stage (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Zilong Zhang − *School of Computer Science and Technology, Hainan University, Haikou 570228, China;* orcid.org/0000-0002-4934-1258; Email: zhangzilong@hainanu.edu.cn

Authors

Yuanhao Li − *School of Computer Science and Technology, Hainan University, Haikou 570228, China*

Aoyun Geng − *School of Computer Science and Technology, Hainan University, Haikou 570228, China*

Zheyu Zhou − *School of Computer Science and Technology, Hainan University, Haikou 570228, China*

Feifei Cui − *School of Computer Science and Technology, Hainan University, Haikou 570228, China;* orcid.org/0000-0001-7055-3813

Junlin Xu − *School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081 Hubei, China*

Yajie Meng − *School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200 Hubei, China*

Leyi Wei − *Centre for Artificial Intelligence driven Drug Discovery, Faculty of Applied Science, Macao Polytechnic University, Macao, SAR, China; School of Informatics, Xiamen University, Xiamen 361005, China*

Quan Zou − *Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China; Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324000, China;* orcid.org/0000-0001-6406-1142

Qingchen Zhang − *School of Computer Science and Technology, Hainan University, Haikou 570228, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c00306

Author Contributions

○Y.L. and A.G. contributed equally to this work.

Funding

The work is supported by the National Natural Science Foundation of China (No. 62450002).

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Hollmann, A.; Cardoso, N. P.; Espeche, J. C.; et al. Review of antiviral peptides for use against zoonotic and selected non-zoonotic viruses. *Peptides* **2021**, *142*, No. 170570.

(2) Niu, M.; Ju, Y.; Lin, C.; et al. Characterizing viral circRNAs and their application in identifying circRNAs in viruses. *Briefings Bioinf.* **2022**, *23* (1), No. bbab404.

(3) Ngai, P. H.; Ng, T. Phaseococcin, an antifungal protein with antiproliferative and anti-HIV-1 reverse transcriptase activities from small scarlet runner beans. *Biochem. Cell Biol.* **2005**, *83* (2), 212−220.

(4) Quintero-Gil, C.; Parra-Suescún, J.; Lopez-Herrera, A.; et al. In-silico design and molecular docking evaluation of peptides derivatives from bacteriocins and porcine beta defensin-2 as inhibitors of Hepatitis E virus capsid protein. *Virusdisease* **2017**, *28*, 281−288.

(5) Rao, B.; Han, B.; Wei, L.; et al. CFCN: An HLA-peptide Prediction Model based on Taylor Extension Theory and Multi-view Learning. *Curr. Bioinf.* **2024**, *19* (10), 977−990.

(6) Otvos, L., Jr Peptide-based drug design: here and now. In *Peptide-Based Drug Design*; Springer, 2008.
(7) Mehta, K.; Vyas, P.; Mujawar, S.; et al. Design and In-silico Screening of Short Antimicrobial Peptides (AMPs) as Anti-Tubercular Agents Targeting INHA. *Curr. Bioinf.* **2023**, *18* (9), 715−736.
(8) Agarwal, G.; Gabrani, R. Antiviral peptides: identification and validation. *Int. J. Pept. Res. Ther.* **2021**, *27* (1), 149−168.
(9) Matthews, T.; Salgo, M.; Greenberg, M.; et al. Enfuvirtide: the first therapy to inhibit the entry of HIV-1 into host CD4 lymphocytes. *Nat. Rev. Drug Discovery* **2004**, *3* (3), 215−225.
(10) Guan, J.; Yao, L.; Chung, C. R.; et al. Stackthpred: identifying tumor-homing peptides through gbdt-based feature selection with stacking ensemble architecture. *Int. J. Mol. Sci.* **2023**, *24* (12), No. 10348.
(11) Guan, J.; Yao, L.; Chung, C. R.; et al. Predicting anti-inflammatory peptides by ensemble machine learning and deep learning. *J. Chem. Inf. Model.* **2023**, *63* (24), 7886−7898.
(12) Yao, L.; Zhang, Y.; Li, W.; et al. DeepAFP: an effective computational framework for identifying antifungal peptides based on deep learning. *Protein Sci.* **2023**, *32* (10), No. e4758.
(13) Ao, C.; Jiao, S.; Wang, Y.; et al. Biological Sequence Classification: A Review on Data and General Methods. *Research* **2022**, *2022*, No. 0011.
(14) Yan, K.; Lv, H.; Shao, J.; et al. TPpred-SC: multi-functional therapeutic peptide prediction based on multi-label supervised contrastive learning. *Sci. China Inf. Sci.* **2024**, *67* (11), No. 212105.
(15) Thakur, N.; Qureshi, A.; Kumar, M. AVPpred: collection and prediction of highly effective antiviral peptides. *Nucleic Acids Res.* **2012**, *40* (W1), W199−W204.
(16) Kumar Meher, P.; Hati, S.; Sahu, T. K.; et al. SVM-Root: Identification of Root-Associated Proteins in Plants by Employing the Support Vector Machine with Sequence-Derived Features. *Curr. Bioinf.* **2024**, *19* (1), 91−102.
(17) Chang, K. Y.; Yang, J.-R. Analysis and prediction of highly effective antiviral peptides based on random forests. *PLoS One* **2013**, *8* (8), No. e70166.
(18) Ru, X.; Li, L.; Zou, Q. Incorporating Distance-Based Top-n-gram and Random Forest To Identify Electron Transport Proteins. *J. Proteome Res.* **2019**, *18* (7), 2931−2939.
(19) Shi, C.; He, J.; Pundlik, S.; et al. Low-cost real-time VLSI system for high-accuracy optical flow estimation using biological motion features and random forests. *Sci. China Inf. Sci.* **2023**, *66* (5), No. 159401.
(20) Beltrán Lissabet, J. F.; Belén, L. H.; Farias, J. G. AntiVPP 1.0: a portable tool for prediction of antiviral peptides. *Comput. Biol. Med.* **2019**, *107*, 127−130.
(21) Schaduangrat, N.; Nantasenamat, C.; Prachayasittikul, V.; et al. Meta-iAVP: a sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation. *Int. J. Mol. Sci.* **2019**, *20* (22), No. 5743.
(22) Li, J.; Pu, Y.; Tang, J.; et al. DeepAVP: A Dual-Channel Deep Neural Network for Identifying Variable-Length Antiviral Peptides. *IEEE J. Biomed. Health Inf.* **2020**, *24* (10), 3012−3019.
(23) Ullah, M.; Akbar, S.; Raza, A.; et al. DeepAVP-TPPred: identification of antiviral peptides using transformed image-based localized descriptors and binary tree growth algorithm. *Bioinformatics* **2024**, *40* (5), No. btae305.
(24) Akbar, S.; Ali, F.; Hayat, M.; et al. Prediction of antiviral peptides using transform evolutionary & SHAP analysis based descriptors by incorporation with ensemble learning strategy. *Chemom. Intell. Lab. Syst.* **2022**, *230*, No. 104682.
(25) Guan, J.; Yao, L.; Xie, P.; et al. A two-stage computational framework for identifying antiviral peptides and their functional types based on contrastive learning and multi-feature fusion strategy. *Briefings Bioinf.* **2024**, *25* (3), No. bbae208.
(26) Zhang, X.; Wei, L.; Ye, X.; et al. SiameseCPP: a sequence-based Siamese network to predict cell-penetrating peptides by contrastive learning. *Briefings Bioinf.* **2023**, *24* (1), No. bbac545.
(27) Yang, M.; Wang, Z.; Yan, Z.; et al. DNASimCLR: a contrastive learning-based deep learning approach for gene sequence data classification. *BMC Bioinf.* **2024**, *25* (1), No. 328.
(28) Lee, B.; Shin, D. Contrastive learning for enhancing feature extraction in anticancer peptides. *Briefings Bioinf.* **2024**, *25* (3), No. bbae220.
(29) Hanson, J.; Litfin, T.; Paliwal, K.; et al. Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning. *Bioinformatics* **2020**, *36* (4), 1107−1113.
(30) Xu, Y.; Liu, T.; Yang, Y.; et al. ACVPred: Enhanced prediction of anti-coronavirus peptides by transfer learning combined with data augmentation. *Future Gener. Comput. Syst.* **2024**, *160*, 305−315.
(31) Tan, X.; Liu, Q.; Fang, Y.; et al. Introducing enzymatic cleavage features and transfer learning realizes accurate peptide half-life prediction across species and organs. *Briefings Bioinf.* **2024**, *25* (4), No. bbae350.
(32) Zeng, W.-F.; Zhou, X. X.; Willems, S.; et al. AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. *Nat. Commun.* **2022**, *13* (1), No. 7238.
(33) He, H.; Garcia, E. A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21* (9), 1263−1284.
(34) Zhang, W.; Jiang, L.; Chen, Z.; et al. FNNWV: farthest-nearest neighbor-based weighted voting for class-imbalanced crowdsourcing. *Sci. China Inf. Sci.* **2024**, *67* (10), No. 202102, DOI: 10.1007/s11432-023-3854-7.
(35) Qureshi, A.; Thakur, N.; Tandon, H.; et al. AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic acids Res.* **2014**, *42* (D1), D1147−D1153.
(36) Jhong, J.-H.; Yao, L.; Pang, Y.; et al. dbAMP 2.0: updated resource for antimicrobial peptides with an enhanced scanning method for genomic and proteomic data. *Nucleic Acids Res.* **2022**, *50* (D1), D460−D470.
(37) Kang, X.; Dong, F.; Shi, C.; et al. DRAMP 2.0, an updated data repository of antimicrobial peptides. *Sci. Data* **2019**, *6* (1), No. 148.
(38) Pirtskhalava, M.; Amstrong, A. A.; Grigolava, M.; et al. DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res.* **2021**, *49* (D1), D288−D297.
(39) Qureshi, A.; Thakur, N.; Kumar, M. HIPdb: a database of experimentally validated HIV inhibiting peptides. *PLoS One* **2013**, *8* (1), No. e54908.
(40) Consortium, U. UniProt: a hub for protein information. *Nucleic Acids Res.* **2015**, *43* (D1), D204−D212.
(41) Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22* (13), 1658−1659.
(42) Lin, Z.; Akin, H.; Rao, R.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379* (6637), 1123−1130.
(43) Yan, C.; Zhang, Z.; Xu, J.; et al. CasPro-ESM2: Accurate identification of Cas proteins integrating pre-trained protein language model and multi-scale convolutional neural network. *Int. J. Biol. Macromol.* **2025**, *308*, No. 142309.
(44) Sandberg, M.; Eriksson, L.; Jonsson, J.; et al. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **1998**, *41* (14), 2481−2491.
(45) Liu, B.; Gao, X.; Zhang, H. BioSeq-Analysis2. 0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* **2019**, *47* (20), e127.
(46) Chen, K.; Jiang, Y.; Du, L.; et al. Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. *J. Comput. Chem.* **2009**, *30* (1), 163−172.
(47) Chou, K.-C.; Cai, Y.-D. Prediction of protein subcellular locations by GO−FunD−PseAA predictor. *Biochem. Biophys. Res. Commun.* **2004**, *320* (4), 1236−1239.
(48) Schneider, G.; Wrede, P. The rational design of amino acid sequences by artificial neural networks and simulated molecular

evolution: de novo design of an idealized leader peptidase cleavage site. *Biophys. J.* **1994**, *66* (2), 335−344.
(49) Saravanan, V.; Gautham, N. BCIgEPRED—A dual-layer approach for predicting linear IgE epitopes. *Mol. Biol.* **2018**, *52*, 285−293.
(50) Geng, A.; Luo, Z.; Li, A.; et al. ACP-CLB: An Anticancer Peptide Prediction Model Based on Multichannel Discriminative Processing and Integration of Large Pretrained Protein Language Models. *J. Chem. Inf. Model.* **2025**, *65* (5), 2336−2349.
(51) Chen, Z.; Liu, X.; Zhao, P.; et al. iFeatureOmega: an integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets. *Nucleic Acids Res.* **2022**, *50* (W1), W434−W447.
(52) Li, Z.; Liu, F.; Yang, W.; et al. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans. Neural Networks Learn. Syst.* **2022**, *33* (12), 6999−7019.
(53) Graves, A. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks* 2012; Vol. *385*, pp 37−45 .
(54) Xiao, C.; Zhou, Z.; She, J.; et al. PEL-PVP: Application of plant vacuolar protein discriminator based on PEFT ESM-2 and bilayer LSTM in an unbalanced dataset. *Int. J. Biol. Macromol.* **2024**, *277*, No. 134317.
(55) Jiao, S.; Ye, X.; Sakurai, T.; et al. Integrated convolution and self-attention for improving peptide toxicity prediction. *Bioinformatics* **2024**, *40* (5), No. btae297.
(56) Jadon, S. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*; IEEE, 2020.
(57) Chicco, D.; Tötsch, N.; Jurman, G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min.* **2021**, *14*, 1−22.
(58) Kubat, M.; Matwin, S. *Addressing the Curse of Imbalanced Training Sets: One-sided Selection*, ICML, 1997.
(59) Wang, Y.; Zhai, Y.; Ding, Y.; et al. SBSM-Pro: support bio-sequence machine for proteins. *Sci. China Inf. Sci.* **2024**, *67* (11), No. 212106.
(60) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction *arXiv*, arXiv preprint arXiv:1802.03426, 2018.
(61) Li, J.; Zhang, C.; Zhou, J. T.; et al. Deep-LIFT: Deep label-specific feature learning for image annotation. *IEEE Trans. Cybern.* **2022**, *52* (8), 7732−7741.
(62) Kokhlikyan, N.; Miglani, V.; Martin, M.et al. Captum: A unified and generic model interpretability library for pytorch. arXiv preprint arXiv:2009.07896, 2020.
(63) Tunc, H.; Yilmaz, S.; Darendeli Kiraz, B. N.; et al. Improving Predictive Efficacy for Drug Resistance in Novel HIV-1 Protease Inhibitors through Transfer Learning Mechanisms. *J. Chem. Inf. Model.* **2024**, *64* (20), 7844−7863.
(64) Yan, K.; Lv, H.; Wen, J.; et al. PreTP-Stack: prediction of therapeutic peptide based on the stacked ensemble learning. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2023**, *20* (2), 1337−1344.
(65) Meher, P. K.; Sahu, T. K.; Saini, V.; et al. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci. Rep.* **2017**, *7* (1), No. 42362.
(66) Yuan, J.; Wang, Z.; Pan, Z.; et al. DPNN-ac4C: a dual-path neural network with self-attention mechanism for identification of N4-acetylcytidine (ac4C) in mRNA. *Bioinformatics* **2024**, *40* (11), No. btae625.

https://doi.org/10.1021/acs.jcim.5c00306
J. Chem. Inf. Model. 2025, 65, 5868−5886