Artificial Intelligence Chemistry 4 (2026) 100114

Contents lists available at ScienceDirect

Artificial Intelligence Chemistry

journal homepage: www.journals.elsevier.com/artificial-intelligence-chemistry

Graph attention network integrating sequence semantics and structural constraint information for robust antiviral peptide prediction

Yuan Chongjun [a], Muhammad Alif Mohammad Latif [a,b], Mohd Basyaruddin Abdul Rahman [a], Bimo A. Tejo [a,*]

[a] Department of Chemistry, Faculty of Science, Universiti Putra Malaysia, Serdang 43400 UPM, Malaysia
[b] Centre for Foundation Studies in Science, Universiti Putra Malaysia, Serdang 43400 UPM, Malaysia

ARTICLE INFO

Keywords:
Graph neural network
Machine learning
Antiviral peptides classifier
ESMFold
Grad-CAM

ABSTRACT

Viral infections remain a major global health challenge, with current antiviral therapies often limited by drug resistance and high development costs. Antiviral peptides (AVPs) are promising alternatives to traditional antiviral drugs owing to their safety and broad-spectrum activity. Most existing AVP prediction classifiers rely solely on sequence-derived features while neglecting three-dimensional structural information, which limits their generalization ability under highly imbalanced virtual screening conditions. To overcome these limitations, we propose a graph-based deep learning framework that explicitly integrates residue-level three-dimensional structural information with sequence semantics for AVPs identification. Residue-level graphs are constructed from ESMFold-predicted structures and enriched with embeddings from the ESMC protein language model, enabling the model to capture both spatially proximal and sequentially distant interactions that are inaccessible to sequence-only approaches. These graphs are processed using a graph attention network with multiscale pooling to learn structure-aware representations. Evaluated on a large, imbalanced, and independent test set, our model demonstrates substantially improved robustness to class imbalance and structural variability, outperforming state-of-the-art sequence-based predictors. Notably, the proposed framework reduces false positives by 54% relative to Stack-AVP, improves the Matthews correlation coefficient by 29%, and achieves an accuracy of 84.1% and specificity of 84.8%. Furthermore, Grad-CAM-based interpretability analysis provides residue-level mechanistic insights, highlighting structurally and functionally relevant amino acids driving antiviral activity. By unifying sequence semantics with explicit structural constraints, this work advances AVPs prediction beyond sequence-only paradigms and provides a practical, interpretable tool for antiviral peptide discovery under realistic, imbalanced conditions.

1. Introduction

Viruses pose significant threats to global health because of their diverse transmission routes, high genetic variability, and ability to persist within host cells [1]. These characteristics contribute to a wide spectrum of infectious diseases, ranging from common colds to life-threatening conditions such as human immunodeficiency virus, hepatitis, and emerging outbreaks such as COVID-19 [2,3]. Current antiviral strategies—primarily small-molecule drugs and vaccines—face critical limitations, including the rapid emergence of drug-resistant strains, high development costs, variable clinical efficacy, and unintended side effects [4,5]. These constraints collectively hinder effective viral control and complicate public health protection efforts.

As alternatives, peptide-based therapeutics have emerged as promising substitutes for traditional protein drugs because of their lower production costs, favorable safety profiles, and high selectivity. Among these peptides, antiviral peptides (AVPs)—a specialized subclass of antimicrobial peptides (AMPs)—exhibit unique advantages over conventional antiviral agents. AVPs demonstrate effectiveness against drug-resistant viruses [6], while offering natural biodegradability and low toxicity. Crucially, they display broad-spectrum activity against diverse viruses [7,8]. These attributes have positioned AVPs as compelling candidates for next-generation antiviral therapeutics, garnering increasing attention for their potential in novel treatment strategies.

* Corresponding author.
E-mail addresses: gs66984@student.upm.edu.my (Y. Chongjun), aliflatif@upm.edu.my (M.A.M. Latif), basya@upm.edu.my (M.B.A. Rahman), bimo.tejo@upm.edu.my (B.A. Tejo).

https://doi.org/10.1016/j.aichem.2026.100114
Received 20 November 2025; Received in revised form 28 January 2026; Accepted 8 March 2026
Available online 13 March 2026

The post-genomic era has witnessed rapid growth in peptide sequence databases, creating an urgent need for efficient computational tools to identify novel AVPs. This demand has driven the development of numerous machine learning and deep learning models that provide faster and more cost-effective AVP predictions compared to traditional experimental approaches. The field of AVP prediction has undergone several key developments. AVPpred pioneered computational AVPs identification by employing support vector machines with physicochemical properties and amino acid composition [9]. AI4AVP employed PC6 encoding and convolutional neural networks, augmented by generative adversarial networks to expand positive training samples [10]. More recently, Stack-AVP introduced a stacked ensemble approach that leverages multi-perspective features incorporating both categorical and probabilistic information [11]. The state-of-the-art AVP-HNCL framework further advanced the field by combining ESM2 embeddings with novel top-k queue contrastive learning and integrated data augmentation techniques [12]. TargetAVP-DeepCaps employs contextual embeddings derived from large pre-trained language models combined with image-based descriptors and capsule networks to achieve high prediction accuracy [13]. DeepAVP-TPPred integrates transformed image-based localized descriptors with evolutionary features and a binary tree growth algorithm for feature optimization [14]. Deepstacked-AVPs adopts a multi-view feature fusion strategy by combining evolutionary profiles, word embeddings, and physicochemical descriptors with stacked ensemble classifiers [15]. In addition, pAVP_PSSMDWT-EnC utilizes transformed evolutionary descriptors, SHAP-based feature selection, and ensemble learning to enhance predictive performance [16]. These evolving methodologies demonstrate the field's progression from simple feature-based models to sophisticated architectures that integrate representation learning and data enhancement techniques. For comprehensive comparisons of AVP prediction models, we direct readers to recent review articles [17].

Although existing computational models have facilitated the identification of potential AVPs, several critical limitations remain unresolved. First, most current models are trained and optimized via small datasets of experimentally validated AVPs, and then evaluated on balanced datasets containing equal numbers of AVPs and non-antiviral peptides (non-AVPs). This approach fails to mimic real-world drug screening scenarios, where bioactive candidates are inherently rare compared with inactive compounds. Consequently, their predictive performance may not be robust in practical applications. Secondly, existing methods for antiviral peptide classifiers rely entirely on sequence-derived features (such as Stack-AVP and AVP-HNCL, etc.), while ignoring three-dimensional (3D) structural information. This oversight is particularly problematic given that antiviral activity often depends on peptide tertiary structure and target binding conformations—features that cannot be fully captured by sequence analysis alone.

Fortunately, recent breakthroughs in protein structure prediction and sequence modeling have made it feasible to extract such rich biological information computationally. The release of AlphaFold [18] in 2018 transformed protein science by enabling accurate structure prediction directly from sequences. Moreover, progress in natural language processing has led to powerful protein language models, which can extract meaningful, contextual features from sequences [19]. Together, these tools open up new possibilities for integrating structural and sequence-based information into unified machine learning frameworks.

In the context of these developments, graph-based modeling approaches have emerged as powerful tools for peptide representation. In these approaches, peptides are treated as graphs, where amino acids are nodes and edges represent structural or spatial connections. This representation allows models to go beyond simple sequence order and learn relationships between residues that are far apart but close in space. Graph neural network, especially graph attention network (GAT) [20], have shown strong performance in peptide classification. GAT can assign different importance levels to neighboring residues, enabling the model focus on structural parts that are most critical for function. This attention mechanism enables learning from both local and long-range interactions. Recent advances in structure-aware peptide prediction have demonstrated the power of integrating graph-based representations with deep learning. Pioneering this approach, sAMPpred-GAT was the first to combine predicted peptide structures, sequence features, and evolutionary information in a GAT framework, significantly improving AMP identification [21]. Building on this, esm-AxP-GDL eliminated reliance on multiple sequence alignment by using ESM-2 embeddings and ESMFold-predicted structures, achieving superior performance over 20 existing classifiers in AMP classification [22]. Further advancing the field, PGAT-ABPp leveraged AlphaFold2-predicted structures and ProtT5 embeddings within a GAT architecture, outperforming 14 state-of-the-art models in terms of accuracy, F1-score, and Matthews correlation coefficient, highlighting the critical role of spatial and semantic features in antibacterial peptide prediction [23].

In this study, we developed an end-to-end computational pipeline that uses both structural and sequence information to predict antiviral peptides. First, ESMFold predicts the 3D structure of input peptide sequences and constructs residue contact edges based on Cα atom Euclidean distances and thresholds. Simultaneously, the ESMC protein language model extracts 1152-dimensional residue-level semantic features from the same sequences. These complementary representations—structural contact graphs from ESMFold and sequence-derived embeddings from ESMC—are integrated into a unified graph. The 3-layer GAT processes this graph through successive attention layers to propagate features, followed by multiscale pooling (512D) to capture hierarchical patterns. The refined features are then classified through fully connected layers (256D→128D→2D) to predict antiviral activity. This dual-modality framework uniquely combines spatially constrained structural relationships (derived from ESMFold) with sequence semantics (captured by ESMC), enabling comprehensive peptide function prediction. By integrating structural and semantic representations, our model addresses two key limitations in current AVP prediction and offers a promising direction for peptide-based antiviral discovery.

2. Methods

2.1. Dataset collection and preparation

We obtained the training datasets for both AVPs and non-AVPs from two sources: (1) the Stack-AVP [11] training dataset, and (2) the Non-AVP Model and Non-AMP Model training sets in the AVP-HNCL [12]. Using identical training data sources ensured consistency and minimized potential biases in model development.

For the testing dataset, we collected AVP sequences from nine established databases: AVPdb [24], ACovPepDB [25], DRAVP [26], HIPdb [27], CAMP R4 [28], DBAASP v3 [29], dbAMP 3.0 [30], DRAMP 4.0 [31], and Peptipedia v2.0 [32]. The non-AVP test sequences were extracted from UniProt [33] via the following search query: “(length: [5−50]) AND (reviewed: true) NOT (keyword:"Antiviral defense") NOT (keyword:"Antiviral protein")”.

All datasets underwent uniform preprocessing where we retained only natural peptide sequences with 5–50 residues and removed redundant entries. Specifically, redundancy reduction for the training datasets was inherited from the original data sources: the Stack-AVP training dataset was constructed using CD-HIT [34] with a 90% sequence identity threshold, while the Non-AVP Model and Non-AMP Model training sets from AVP-HNCL were generated using a more stringent CD-HIT threshold of 40%. To ensure fair model evaluation and dataset independence, we implemented two key measures: (1) removing sequences with conflicting labels to ensure the uniqueness of sequence categories, (2) eliminating any duplicate sequences between the training set and the testing set, as well as within each set (including both AVP and non-AVP subsets), to ensure complete independence between the training and test sets with no data leakage. This rigorous approach

prevents data leakage and guarantees that performance comparisons reflect true predictive capabilities.

The final curated datasets contained 3091 AVP sequences and 3719 non-AVP sequences in the training set, along with 1403 AVP sequences and 7894 non-AVP sequences in the testing set (a summary of the datasets used in this study is presented in Table 1). This independent testing set construction enables objective benchmarking against existing prediction methods, as it contains completely novel sequences not encountered during model training.

2.2. Graph representation

2.2.1. Node feature extraction using ESMC embeddings

To numerically represent peptide sequences at the residue level, we employed the 600M-parameter ESMC model [35], a pretrained protein language model based on the Transformer architecture. ESMC is trained on large-scale protein sequence corpora using self-supervised learning objectives, enabling it to learn contextualized representations that encode evolutionary, biochemical, and structural signals directly from amino acid sequences. Specifically, ESMC captures residue-level contextual dependencies by modeling each amino acid in relation to its surrounding sequence context, allowing the resulting embeddings to reflect not only local sequence motifs but also long-range interactions that are relevant to protein structure and function.

In this work, ESMC was used to generate residue-level embeddings for each peptide sequence. For a peptide of length L, the model outputs an embedding matrix of size [L, D], where $D = 1152$ for ESMC. Each embedding vector serves as a node feature in the residue-level graph, providing a rich semantic description of individual residues that complements the explicit three-dimensional structural information used in graph construction. These embeddings form the foundation for downstream graph-based learning of antiviral activity.

2.2.2. Peptide 3D structure feature extraction

The three-dimensional structures of all peptide sequences were first predicted via ESMFold [36]. To enhance the understanding of amino acid interactions, the predicted structural information was transformed into a graph-based representation. Initially, we computed the coordinate differences for each pair of amino acids in terms of their 3D positions. Let $r_i = (x_i, y_i, z_i)$ and $r_j = (x_j, y_j, z_j)$ denote the Cartesian coordinates of the Cα atoms of residues i and j, respectively. The vector difference r_{ij} is given by:

$$r_{ij} = r_i - r_j = \left(x_i - x_j,\ y_i - y_j,\ z_i - z_j\right) \tag{1}$$

To establish residue contacts, we calculated the Euclidean distance d_{ij} between every pair of amino acids in the peptide:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + \left(y_i - y_j\right)^2 + (z_i - z_j)^2} \tag{2}$$

We defined residue-residue contacts based on spatial proximity between Cα atoms. Two residues were considered in contact if the Euclidean distance between their Cα atoms was below a threshold distance θ. This binary classification criterion allowed us to construct an adjacency matrix A, where each element A_{ij} represented the contact state between residues i and j:

$$A_{ij} = \left\{ \begin{array}{l} 1,\ d_{ij} < \theta \\ 0, otherwise \end{array} \right\} \tag{3}$$

Table 1
Overview of the training set and testing set.

Dataset	Number of AVP	Number of non_AVP	Total
Training set	3091 (45.4%)	3719 (54.6%)	6810
Testing set	1403 (15.1%)	7894 (84.9%)	9297

Each peptide was represented as an undirected graph $G = (V, E)$, where V denotes the set of residues (nodes) with corresponding feature vectors, and $E = \{(i, j) \mid d_{ij} < \theta\}$ encodes the residue-level spatial proximity relationships as edges.

Since the proposed framework relies on ESMFold-predicted peptide structures, we assessed the confidence of the predicted conformations using the predicted Local Distance Difference Test (pLDDT) scores. The distribution of average pLDDT scores across all predicted peptide structures was analyzed to evaluate structural reliability. Although short peptides are inherently flexible and may adopt multiple conformations in solution, the predicted structures exhibited a reasonable confidence range, with an average pLDDT score of approximately 69 as shown in Fig. S1. Importantly, rather than filtering structures based on a strict pLDDT threshold, we retained all predicted structures to preserve dataset completeness and diversity, while allowing the downstream graph-based model to learn robust representations under structural uncertainty.

2.3. Model architecture

We constructed GAT models with either two or three attention layers. Each layer performed attention-based message passing with a 128-dimensional hidden representation, followed by layer normalization and ReLU activation. A 30% dropout was applied after each non-final layer to mitigate overfitting. Before passing the features into the GAT layers, we projected them into 128 dimensions. This was done via a linear layer followed by batch normalization. After message passing, we used four types of global pooling to obtain graph-level features: max pooling, average pooling, additive pooling, and TopK pooling. For TopK, we retained 70% of the most important nodes. Multiple global pooling strategies, including max pooling, average pooling, additive pooling, and TopK pooling, were employed to capture complementary aspects of graph-level representations. Max pooling emphasizes the most informative residues, average pooling summarizes overall feature distributions, and additive pooling preserves global signal intensity across all nodes. TopK pooling further enables adaptive selection of the most structurally and functionally relevant residues by retaining the most important nodes based on learned importance scores. By combining these pooling operations, the model is able to integrate both local salient features and global structural context, leading to more robust peptide-level representations. These four pooled outputs were combined into one 512-dimensional vector. This vector was then passed into a three-layer classifier. The classifier used hidden sizes of 256 and 128, ending with a 2-class output. We used batch normalization, ReLU, and dropout (30% then 15%) between layers.

To evaluate model performance under different architectural and structural configurations, we trained and compared 2-layer and 3-layer GAT models across three distance thresholds (4.0 Å, 8.0 Å, and 12.0 Å), via stratified 5-fold cross-validation. The chosen thresholds are biologically significant: 4.0 Å captures short-range contacts, including covalent bonds and tight non-covalent interactions, 8.0 Å is representative of a commonly used cutoff for potential hydrogen bonding and medium-range interactions, and 12.0 Å encompasses longer-range van der Waals and electrostatic interactions. Since our training dataset was imbalanced, we employed a class-weighted CrossEntropyLoss during training. We trained all the models using the AdamW optimizer, with a learning rate of 0.0001 and a weight decay of 0.01. Each batch contained 64 graphs. We reduced the learning rate by a factor of 0.5 if the validation area under the receiver operating characteristic curve (AUC) did not improve for 10 consecutive epochs. Early stopping was triggered if no improvement in the AUC was observed after 20 consecutive epochs. All the code was developed in PyTorch and PyTorch Geometric.

To mitigate overfitting, several regularization strategies were employed during model training. Dropout was applied to the graph attention layers and fully connected layers to prevent co-adaptation of features. L2 weight decay was used to penalize large model parameters.

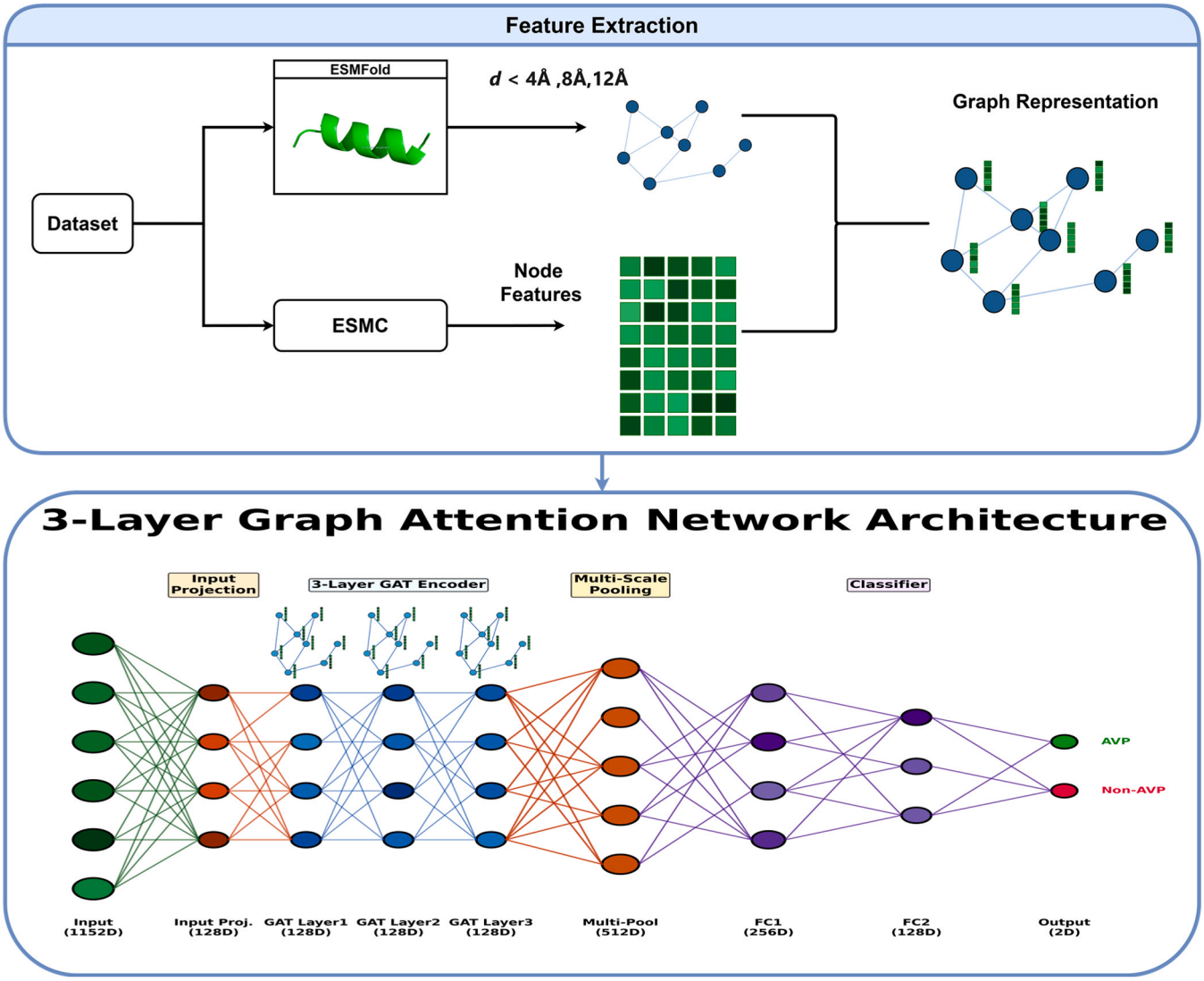

Fig. 1. Overview of the proposed graph-based AVP prediction framework. Peptide sequences are first encoded using ESMC embeddings and folded by ESMFold to obtain three-dimensional structures. Residue-level graphs are constructed based on spatial proximity, processed by a multi-layer graph attention network with multiscale pooling, and finally classified by a fully connected neural network.

In addition, early stopping based on validation loss was adopted to halt training when no further improvement was observed. Together, these strategies effectively improved the generalization performance of the proposed model.

An overview of the proposed model architecture is illustrated in Fig. 1, which summarizes the integration of sequence embeddings, structure-based graph construction, graph attention learning, and classification. The optimal hyperparameters used for training the proposed model are summarized in Table S1.

2.4. Evaluation metrics

To evaluate our proposed model and the state-of-the-art model, we employed multiple performance metrics. These include sensitivity (Sn), specificity (Sp), F1-score, Matthews correlation coefficient (MCC), and AUC. Each metric provides unique insights into model performance. Sn and Sp measure how well the models identify true positives (correctly predicted positive cases) and true negatives (correctly predicted negative cases), respectively. The F1-score balances precision (correct positive predictions) and sensitivity (actual positives detected) into a single metric. The MCC evaluates overall classification quality by considering

Table 2
Performance Metrics of Different-Layer GAT Models across Varying Distance Thresholds in the Training dataset.

Distance	GAT_Model	Accuracy	Sensitivity	Specificity	F1	MCC	AUC
4.0A	2Layer_GAT	0.812 ± 0.008	0.832 ± 0.029	0.796 ± 0.029	0.801 ± 0.009	0.626 ± 0.014	0.905 ± 0.005
4.0A	3Layer_GAT	0.820 ± 0.012	0.822 ± 0.024	0.819 ± 0.037	0.806 ± 0.008	0.640 ± 0.019	0.907 ± 0.007
8.0A	2Layer_GAT	0.815 ± 0.002	0.830 ± 0.021	0.804 ± 0.018	0.803 ± 0.004	0.631 ± 0.004	0.908 ± 0.001
8.0A	3Layer_GAT	0.816 ± 0.010	0.824 ± 0.025	0.809 ± 0.030	0.802 ± 0.009	0.632 ± 0.019	0.906 ± 0.004
12.0A	2Layer_GAT	0.808 ± 0.009	0.833 ± 0.017	0.787 ± 0.026	0.797 ± 0.007	0.618 ± 0.016	0.904 ± 0.005
12.0A	3Layer_GAT	0.812 ± 0.005	0.825 ± 0.014	0.802 ± 0.011	0.799 ± 0.006	0.624 ± 0.011	0.902 ± 0.003

all prediction outcomes (TP/FP/TN/FN), remaining reliable with imbalanced data. The AUC reflects the model's ability to rank positive samples higher than negative ones across all decision thresholds. A higher AUC indicates better separation between the two classes. These metrics are defined as follows:

$$\begin{aligned}
&\text{Accuracy (ACC)}: && \mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \\
&\text{Sensitivity (Sn)}: && \mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\
&\text{Specificity (Sp)}: && \mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\
&\mathrm{F1}-\text{score (F1)}: && F1 = \frac{\mathrm{TP}}{\mathrm{TP}+\frac{1}{2}(\mathrm{FP}+\mathrm{FN})} \\
&\text{Matthews Correlation Coefficient (MCC)}: && \mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \\
&\text{Area Under ROC Curve (AUC)}: && \text{Area under the receiver operating characteristic curve}
\end{aligned} \tag{4}$$

Here, TP (true positives) and TN (true negatives) represent the number of correctly predicted positive and negative cases, respectively. Conversely, FP (false positives) and FN (false negatives) indicate incorrectly predicted positive and negative cases, respectively. In addition to ROC-AUC, we employed precision–recall (PR) curves and the area under the PR curve (PR-AUC) to evaluate model performance under class-imbalanced conditions. Unlike ROC-AUC, PR-AUC focuses on the positive class and provides a more informative assessment when negative samples dominate the dataset, as is the case in realistic AVP screening scenarios.

2.5. Model performance visualization and interpretability analysis

To comprehensively assess the classification performance and feature representation ability of our deep learning model, we performed dimensionality reduction and visualization analyses on high-dimensional data. We used two popular dimensionality reduction techniques, t-distributed stochastic neighbor embedding (t-SNE) [37] and uniform manifold approximation and projection (UMAP) [38], to compare the features extracted by the protein language model (ESMC) with traditional manual features (AAIndex [39]) to intuitively demonstrate their advantages in distinguishing positive and negative samples. Additionally, we extracted and visualized features using the same methods from the three-layer GAT network and the final classification layer to illustrate the model's feature learning process. Furthermore, to understand the contribution of individual amino acids to the model's predictions, we applied gradient-weighted class activation mapping (Grad-CAM) [40]. These visual explanations help reveal how the model makes its classification decisions.

3. Results and discussion

3.1. Training set and testing set performance

Our systematic evaluation compared 2-layer and 3-layer GAT models across three distance thresholds (4.0 Å, 8.0 Å, 12.0 Å) using 5-fold cross-validation in the training dataset. As shown in Table 2, all models demonstrated strong discriminative ability, consistently achieving AUC values > 0.90 and accuracy $> 80\%$ across distance thresholds. This performance consistency reveals that both GAT architectures effectively learn graph representations for distinguishing AVPs from non-AVPs, regardless of specific model configurations. Among all settings, the 3-layer GAT model at the 4.0 Å distance threshold delivered the most robust performance, outperforming alternatives in four key metrics: accuracy (82.0% ± 1.2%), specificity (81.9% ± 3.7%), F1-score (0.806 ± 0.008), and MCC (0.640 ± 0.019). The improvement in specificity is especially important for real-world drug screening scenarios where non-AVPs vastly outnumber AVPs.

As shown in Table 3, the 4.0 Å 3-layer GAT model maintained its performance superiority in independent testing, demonstrating significant advantages in critical metrics: accuracy (0.841), specificity (0.848), F1-score (0.603), and MCC (0.537). Notably, the 3-layer GAT model at 4.0 Å exhibited a 3.5% absolute improvement in specificity (0.848) compared with the lowest-performing configuration (12.0 Å, 3-layer GAT, specificity = 0.813), which is a critical advantage for real-world screening where non-AVPs dominate sample libraries. While the sensitivity remains stable across models (±2% variation), the 3-layer/4.0 Å architecture's balanced performance profile (84.8% specificity with 80.2% sensitivity) establishes it as the most generalizable framework for practical applications. We subsequently benchmark this optimized model (we also call it AVP-3LGAT) against state-of-the-art predictors (Stacked-AVP and AVP-HNCL) to assess its competitive advantages.

The superior performance observed at the 4.0 Å distance cutoff can be attributed to the dominance of short-range residue–residue interactions in determining antiviral activity. Contacts within this range primarily capture local backbone constraints, side-chain packing, and specific non-covalent interactions, such as hydrogen bonds and salt bridges, which are critical for stabilizing functionally relevant peptide conformations and mediating target recognition. In contrast, larger distance thresholds (8.0 Å and 12.0 Å) introduce a substantial number of long-range and weakly informative contacts, which may dilute discriminative structural signals and increase graph connectivity noise. For short peptides, where functional activity is often governed by a limited set of key residues, emphasizing short-range structural contacts enables more precise representation learning and improves classification performance.

3.2. Benchmarking AVP-3LGAT against state-of-the-art AVP predictors

We benchmarked AVP-3LGAT against three state-of-the-art AVP prediction models: Stack-AVP and two variants of AVP-HNCL trained on non-AMP and non-AVP negative datasets, respectively. For comparative benchmarking, the Stack-AVP and AVP-HNCL models were evaluated using their original implementations and reported configurations. These comparative models were not retrained using the training set constructed in this study. Instead, their pre-trained versions were directly applied to the independent test set to assess generalization performance.

Table 3
Performance Metrics of Different-Layer GAT Models across Varying Distance Thresholds in the Test dataset.

Distance	GAT_Model	Accuracy	Sensitivity	Specificity	F1	MCC	AUC
4.0A	2-Layer GAT	0.819	0.815	0.82	0.576	0.508	0.911
4.0A	3-Layer GAT	0.841	0.802	0.848	0.603	0.537	0.913
8.0A	2-Layer GAT	0.826	0.822	0.827	0.588	0.522	0.912
8.0A	3-Layer GAT	0.835	0.806	0.841	0.597	0.53	0.913
12.0A	2-Layer GAT	0.829	0.823	0.83	0.593	0.528	0.914
12.0A	3-Layer GAT	0.814	0.819	0.813	0.57	0.502	0.909

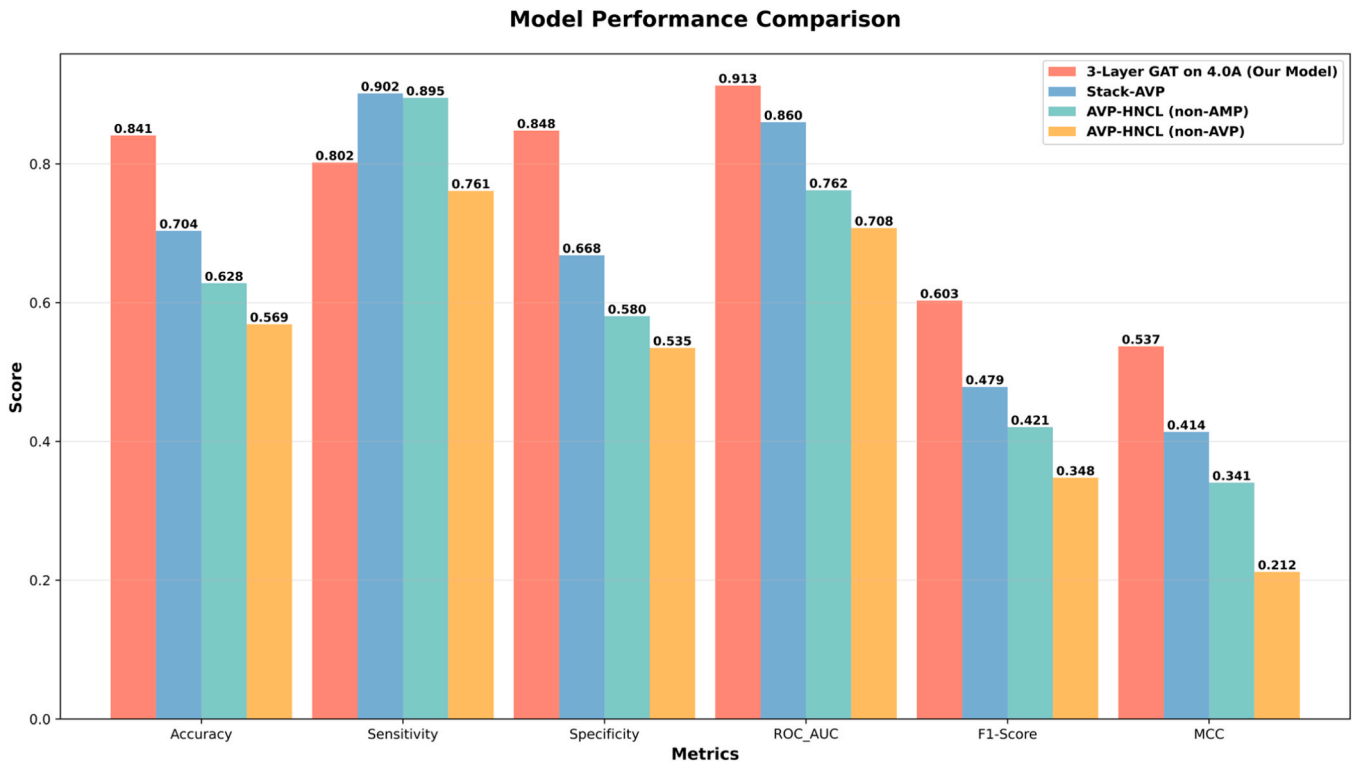

Fig. 2. Performance comparison of AVP-3LGAT versus Stack-AVP and AVP-HNCL in terms of Accuracy (ACC), Sensitivity (Sn), Specificity (Sp), AUC, F1-score, and Matthews correlation coefficient (MCC) on independent test dataset.

This evaluation strategy ensures a fair and realistic comparison under practical deployment scenarios, where pretrained predictors are typically applied to previously unseen peptide sequences. Moreover, the independent test set used in this study contains no overlapping sequences with the training data of our proposed model, thereby preventing data leakage and enabling an objective assessment of predictive performance.

As shown in Fig. 2, AVP-3LGAT outperformed all competing models across all evaluation metrics except sensitivity. Among the three external models, Stack-AVP shows the best overall performance, followed by AVP-HNCL trained on non-AMP data, whereas AVP-HNCL trained on non-AVP data performs the worst. Comparative results with both AVP-HNCL variants are shown in Figs. S2 and S3. A shared limitation of these models is their near-binary label output distribution: they tend to predict approximately half of the samples as AVPs and the other half as non-AVPs on the highly imbalanced test set (15% AVPs, 85% non-AVPs). This suggests that AVP-HNCL struggles to distinguish positive from negative samples under real-world class imbalance conditions. While Stack-AVP achieves the highest sensitivity (90.2%), this advantage comes with substantially reduced specificity (66.8%), which is a critical limitation for practical applications where non-AVPs dominate screening libraries. The main advantages of our model over Stack-AVP include the following: 1. 29.7% higher Matthews correlation coefficient (0.537 vs 0.414 for Stack-AVP); 2. 25.9% improvement in F1-score (0.603 vs 0.479); 3. 13.7% increase in accuracy. These results underscore AVP-3LGAT's strong generalization ability and practical relevance for AVP discovery pipelines.

To further assess whether the high specificity of AVP-3LGAT is merely an artifact of class imbalance, we performed a precision–recall (PR) curve analysis on the independent test set (Fig. 3). PR curves are particularly informative for imbalanced datasets, as they directly reflect the trade-off between precision and recall for the minority class (AVPs). As shown in Fig. 3, AVP-3LGAT consistently maintains substantially higher precision across a wide recall range compared with Stack-AVP and both variants of AVP-HNCL. Notably, the proposed model achieves a PR-AUC of 0.721, which is markedly higher than Stack-AVP (PR-AUC = 0.532), AVP-HNCL (non-AVP) (0.300), and AVP-HNCL (non-AMP) (0.284). These results indicate that AVP-3LGAT is able to effectively enrich true AVPs among predicted positives, even under severe class imbalance. Importantly, this demonstrates that the improved specificity of our model is not a consequence of bias toward the majority class, but rather reflects superior discriminative ability in identifying antiviral peptides under realistic screening conditions.

Our AVP-3LGAT model outperforms Stack-AVP across all key metrics except sensitivity. Confusion matrix analysis reveals that Stack-AVP misclassified 2618 non-AVPs, corresponding to 33% of all negative samples (Fig. 4A-B). In contrast, AVP-3LGAT reduced the number of false positives by 54%, with only 1203 misclassified non-AVPs (15%)

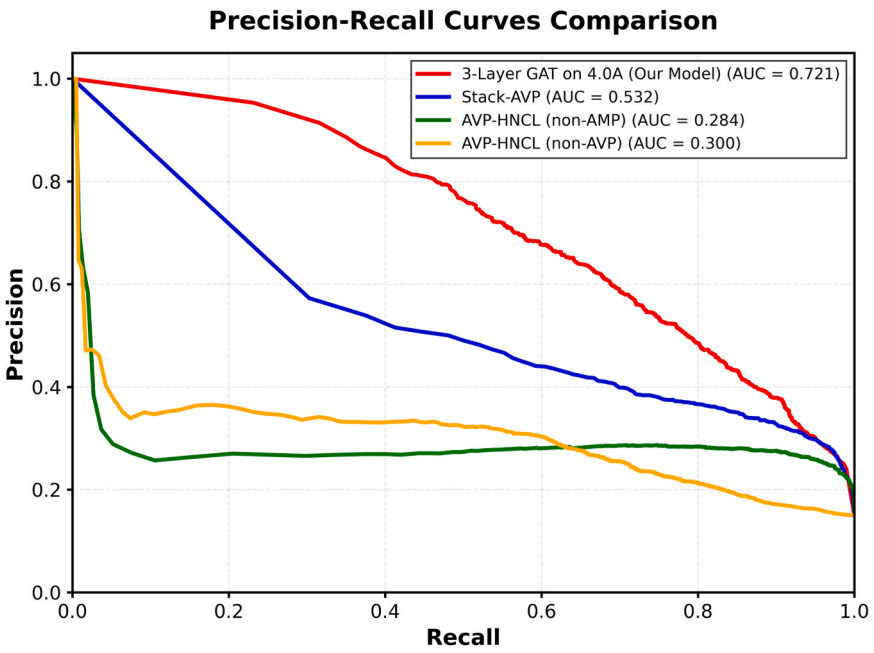

Fig. 3. Precision–recall curve comparison of AVP-3LGAT, Stack-AVP, and AVP-HNCL on the independent imbalanced test set.

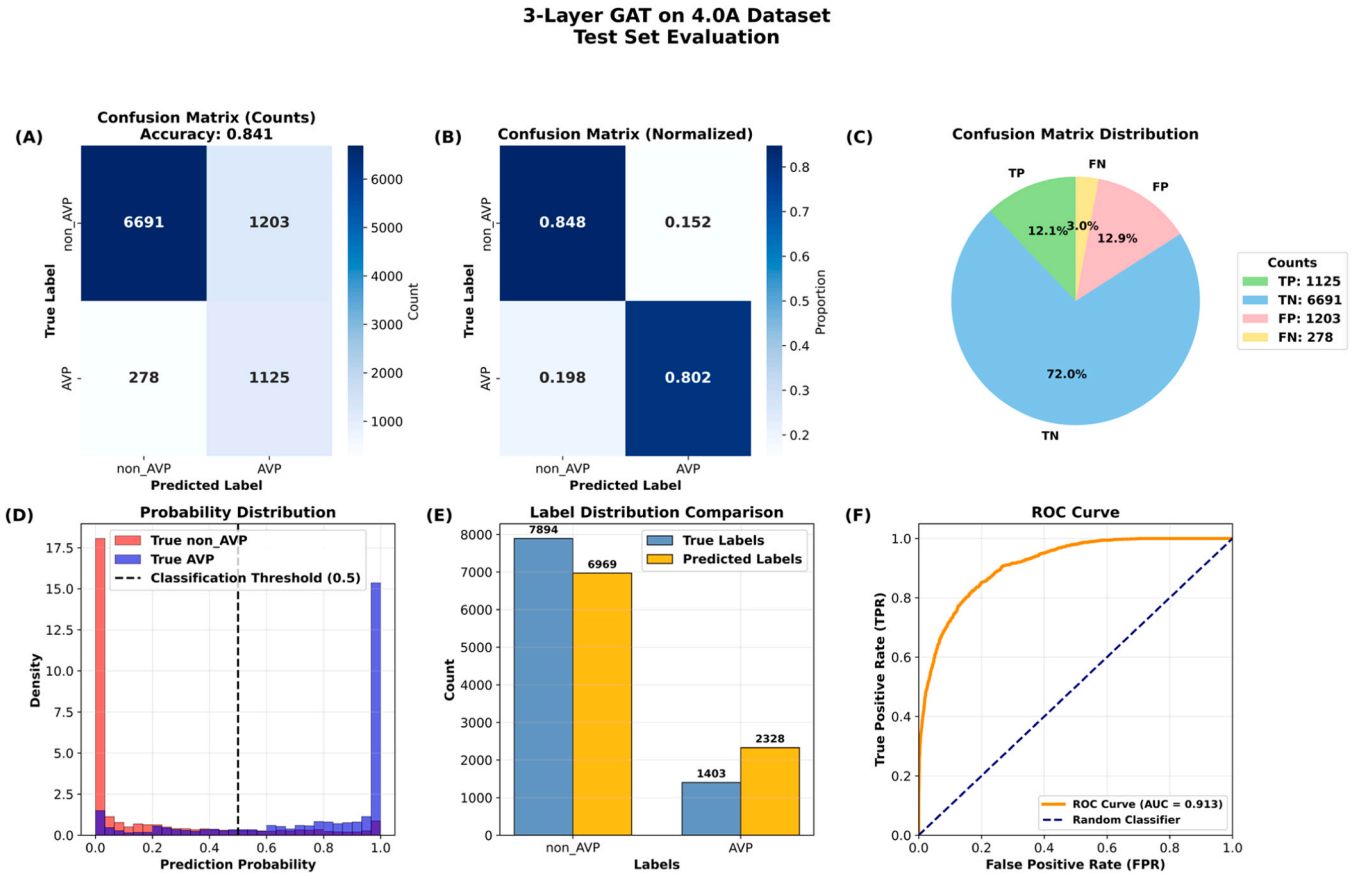

Fig. 4. Performance of the 3-layer GAT model with a 4.0 Å threshold on the independent test set. (A) Confusion matrix (raw counts). (B) Normalized confusion matrix. (C) Distribution of TP, TN, FP, and FN. (D) Prediction probability distribution. (E) True vs predicted label frequencies. (F) ROC curve (AUC = 0.913).

(Fig. 5A-B). Our model achieves an accuracy of 84.1% (TP:12.1%+TN:72%), whereas Stack-AVP only has an accuracy of only 70.4% (TP:13.6%+TN:56.7%) (Figs. 4C and 5C). The probability distribution plots and label distribution comparison confirm that our model better approximates the true label distribution, with tighter clustering of non-AVPs below the 0.5 threshold (Figs. 4D-E vs 5D-E). Most critically, AUC analysis demonstrates that our model achieves significantly greater discriminative ability (AUC 0.913 vs 0.860), reflecting its enhanced

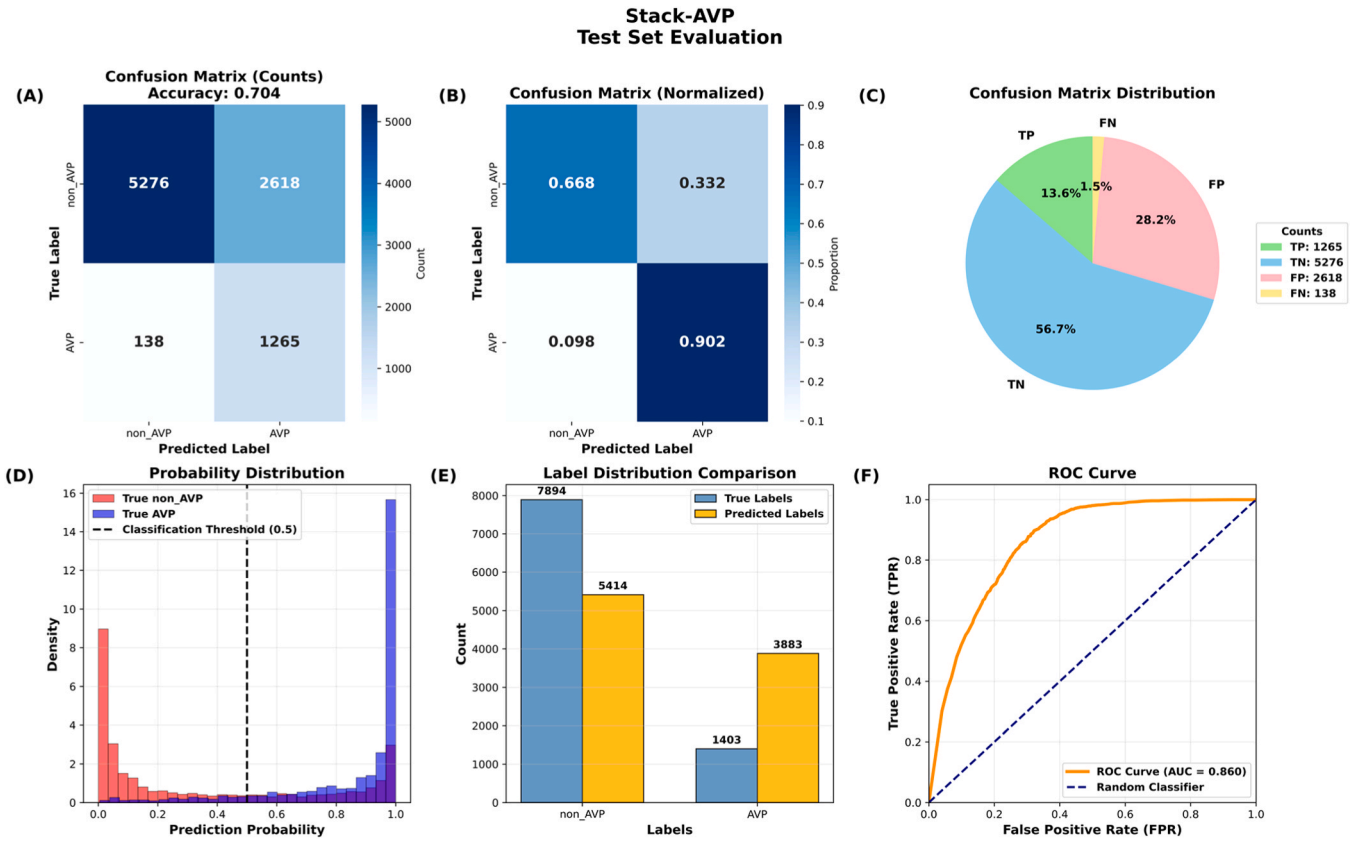

Fig. 5. Performance of the Stack-AVP model on the independent test set. (A) Confusion matrix (raw counts). (B) Normalized confusion matrix. (C) Distribution of TP, TN, FP, and FN. (D) Prediction probability distribution. (E) True vs predicted label frequencies. (F) ROC curve (AUC = 0.860).

ability to differentiate between AVPs and non-AVPs compared with Stack-AVP (Figs. 4F vs 5F). This robust separation of true positives from negatives underscores the model's reliability for accurate AVP identification.

Our findings indicate a clear trade-off in performance between the Stack-AVP model and our proposed model. While Stack-AVP showed higher sensitivity, meaning that it is better at identifying true antiviral peptides, this comes at the cost of significantly lower specificity. Specifically, Stack-AVP frequently flags peptides as antiviral when they are not, leading to many false positives. This high rate of false positives makes Stack-AVP impractical for real-world screening. Each false positive requires additional experimental validation, which is both time-consuming and expensive. Therefore, its high sensitivity does not translate into a useful tool for practical applications. In contrast, our model offers a more balanced performance. It maintains a good level of sensitivity while significantly improving specificity. This means that our model generates fewer false positives, making it a much more reliable tool for discovering antiviral peptides in real-world scenarios. Its consistent performance across both metrics makes it a better candidate for practical screening applications.

3.3. *Visualization of feature representations and model decision process*

The principle behind AVP-3LGAT involves using protein language models to extract key sequence patterns and employing GAT to capture complex multi-functional relationships. We can reveal the decision-making process by visualizing the distribution of peptide representations. To demonstrate the superiority of ESMC as node features, we compared peptide representation distributions extracted from both ESMC and AAIndex in the training and test sets. We specifically focused on 553 attributes from AAIndex version 9.2 that contained no NaN values. To compare embedding differences, we visualized the feature space using t-SNE (prioritizing local structure) and UMAP (preserving both local and global patterns). These techniques help project high-dimensional data into two-dimensional space, enabling intuitive visualization of data clustering and class separability.

The t-SNE plots (Fig. 6 A-D) and UMAP plots (Fig. 7 A-D) revealed different clustering patterns for each embedding in both the raining and test sets. T-SNE plots (Fig. 6 A-B) showed that the AAIndex exhibited substantial overlap between the AVPs and the non-AVPs categories in both the training and test sets, indicating limited precision in distinguishing functional differences. Similar phenomena were observed in the UMAP plots (Fig. 7 A-B). In contrast, ESMC demonstrated more distinct clustering in both t-SNE plots (Fig. 6 C-D) and UMAP plots (Fig. 7 C-D), showing clearer separation between AVP and non-AVP categories, which proved its strong discriminative ability. These observations indicate that features extracted by ESMC can more effectively cluster positive and negative samples, providing valuable insights for peptide classification model design.

To further assess how model processing refines these representations, we visualized intermediate features after the 3-layer GAT module and after the final fully connected layer. As shown in Fig. 6 E-F, positive and negative samples from the independent test set displayed more obvious clustering after processing through the 3-layer GAT and were clearly separated after the fully connected layer of the prediction module. In UMAP, although the initial ESMC feature distribution was relatively chaotic and disordered, the distinction between AVPs and non-AVPs became apparent as data passed through the model's 3-layer GAT module and classifier module (Fig. 7 E and F), demonstrating the effectiveness of our GAT model in improving classification accuracy. These visualizations demonstrate how the AVP-3LGAT progressively enhances feature discriminability, transforming raw embeddings into more class-informative representations. The consistent separation patterns across both the t-SNE and UMAP plots support the robustness and

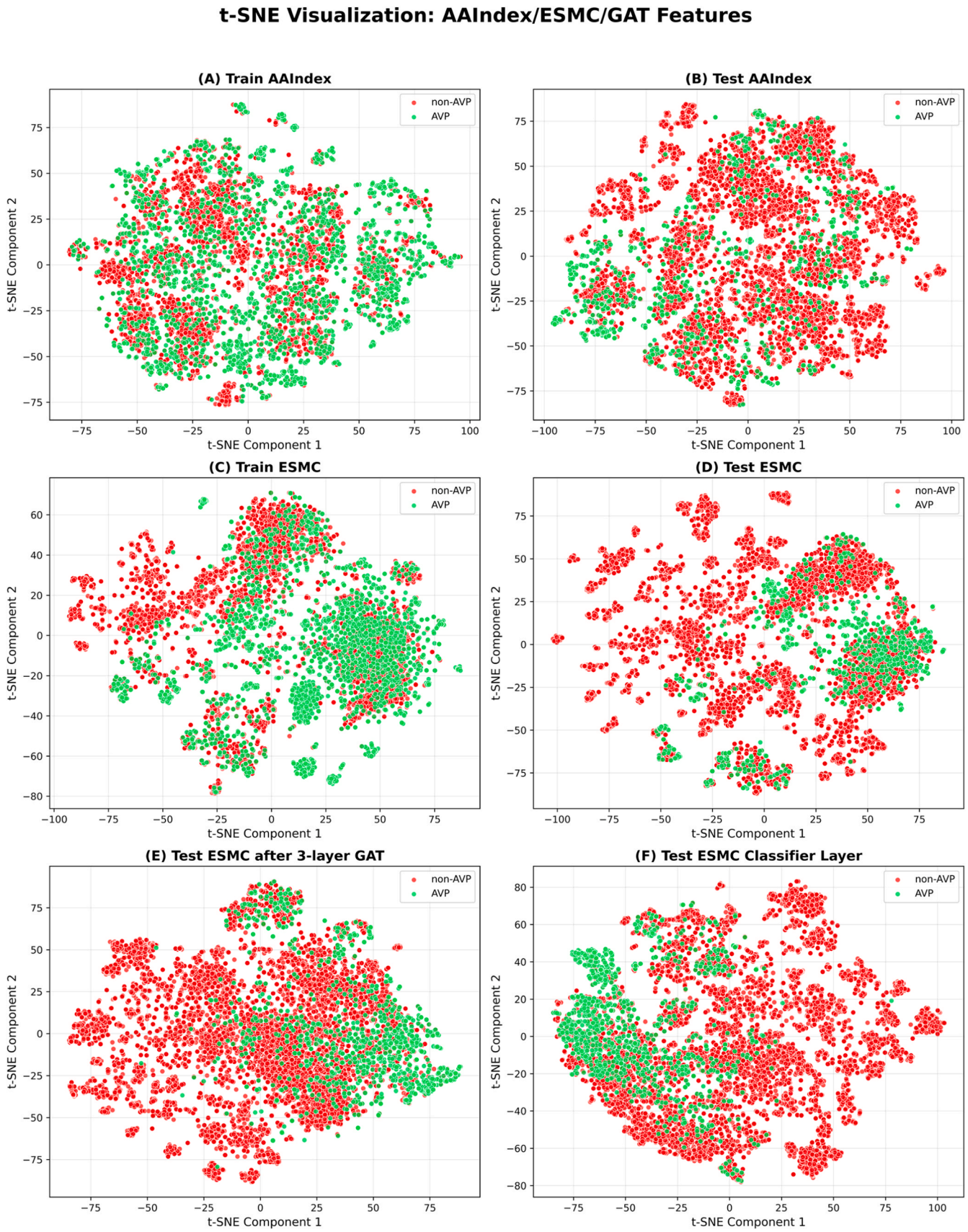

Fig. 6. T-SNE visualization of feature space separation for AVPs (green) versus non-AVPs (red) across different feature sets and model processing stages. (A) AAIndex features of the training set. (B) AAIndex features of the testing set. (C) ESMC features of the training set. (D) ESMC features of the testing set. (E) Testing set ESMC features after 3-layer GAT processing. (F) Testing set ESMC features in the classifier layer outputs.

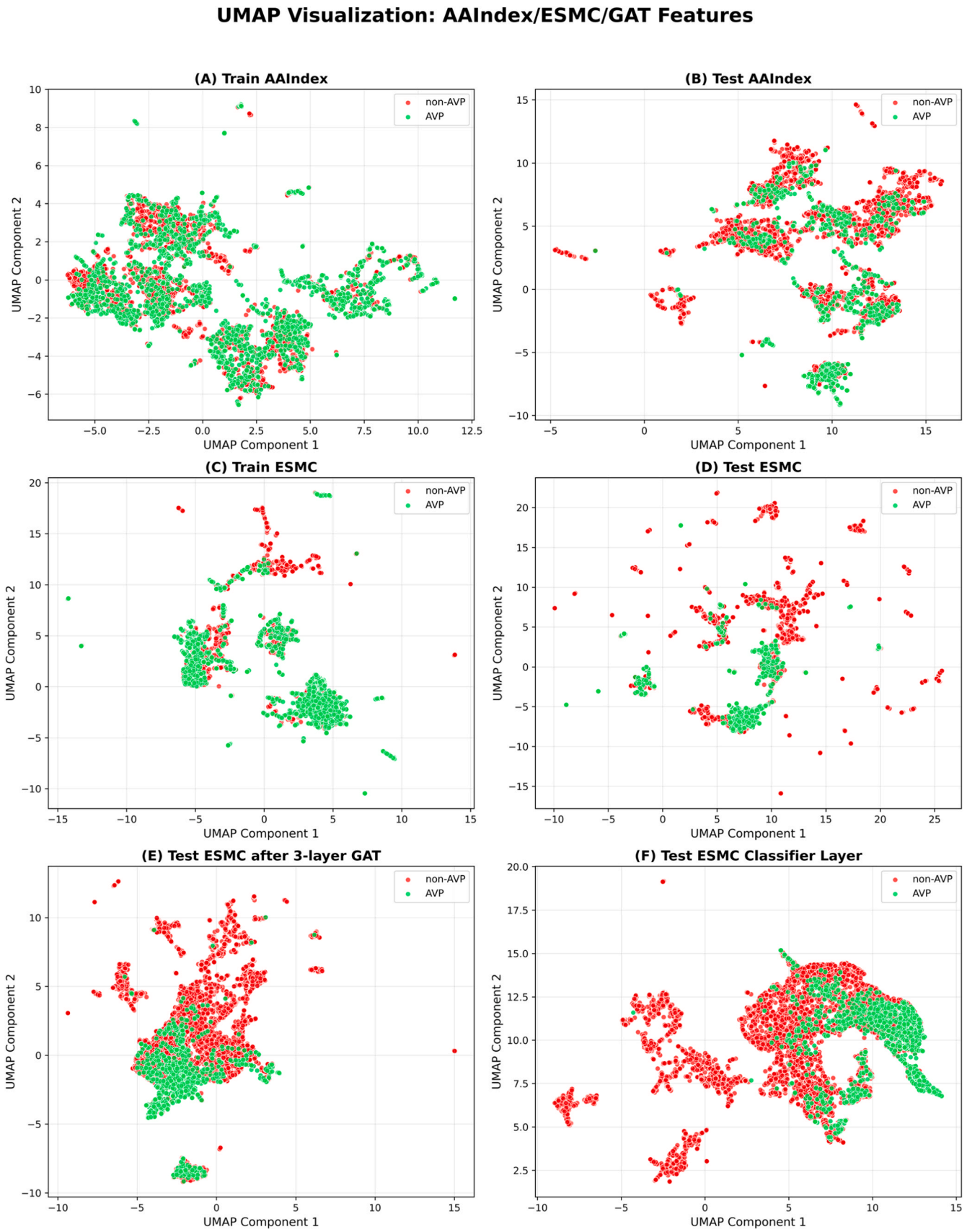

Fig. 7. UMAP visualization of feature space separation for AVPs (green) versus non-AVPs (red) across different feature sets and model processing stages. (A) AAIndex features of the training set. (B) AAIndex features of the testing set. (C) ESMC features of the training set. (D) ESMC features of the testing set. (E) Testing set ESMC features after 3-layer GAT processing. (F) Testing set ESMC features in classifier the layer outputs.

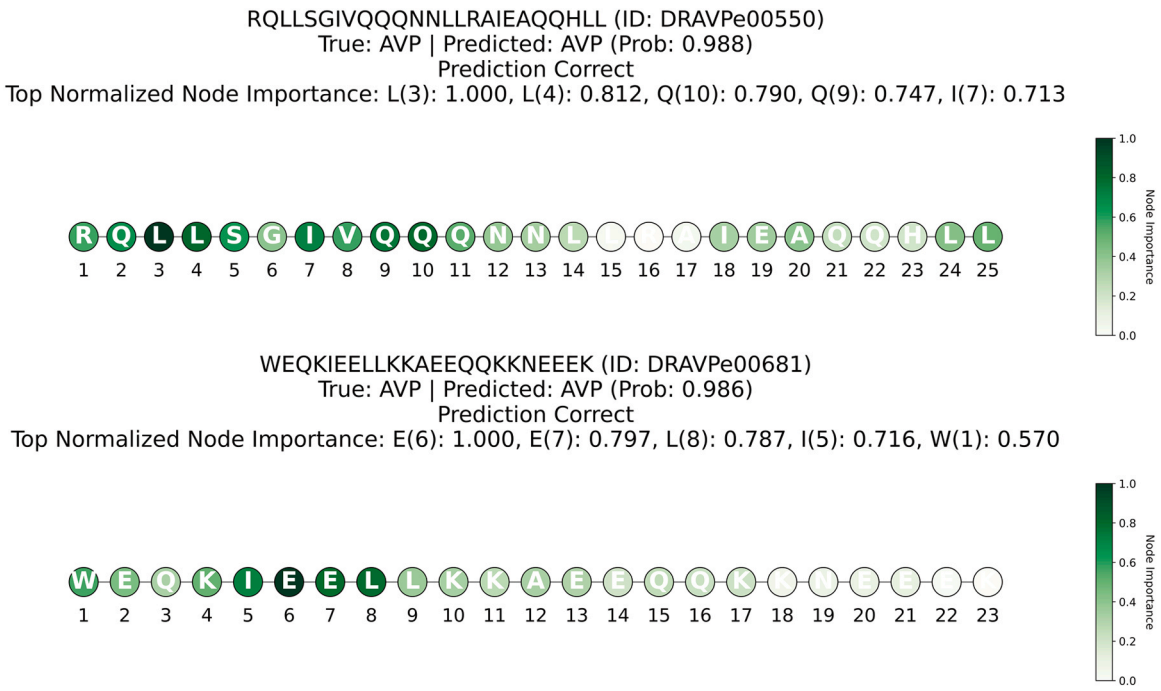

Fig. 8. The Grad-CAM importance plot of DRAVPe00550 and DRAVPe00681. The importance of the individual amino acid is represented from whitish (for a minor importance) to green (for a greater importance) for the development of AVP.

interpretability of the learned latent space.

3.4. Interpretability analysis

In this study, we used Grad-CAM to explore how individual amino acids influence the model's predictions. We applied Grad-CAM to the final layer of the GAT module. This method computes the gradient of the target class with respect to the layer's activations and uses these gradients to weigh the importance of each node. The result is a weighted importance map that highlights which residues are most influential in forming the final prediction.

The Grad-CAM implementation for our GAT models employed a systematic approach to quantify residue-level importance. By capturing both activations and gradients from the final GAT layer through PyTorch hooks, we computed normalized importance scores that reflect each amino acid's contribution to the prediction. The ensemble approach, averaging results across five independently trained models, provided robust importance estimates that mitigated individual model biases. To validate these importance scores, we conducted masking experiments. We replaced the features of both the top 20% and bottom 20% important residues with zeros. This helped us evaluate whether Grad-CAM correctly identified the most influential nodes for predictions.

We analyzed sample DRAVPe00550 as a representative case for Grad-CAM importance evaluation (Fig. 8). As shown in Table S2, the model initially assigned this peptide a high prediction score of 0.9883, indicating strong confidence in its AVP activity. When we masked the top 20% most important amino acid residues, the prediction score dropped sharply to 0.5582 - a substantial 43.01% decrease. This demonstrates these residues play a crucial role in the model's decision-making. In contrast, masking the least important 20% residues only caused a modest reduction to 0.7817 (20.66% decrease). The striking difference between these outcomes confirms that Grad-CAM effectively pinpoints the residues that most influence the prediction results.

To validate the biological relevance of our model's predictions, we compared the Grad-CAM importance map of DRAVPe00550 (C16-N25) with its established functional domains derived from structural and mutagenesis studies [41]. Strikingly, the residues deemed most important by our model largely coincide with the known functional sites of this class of Human Immunodeficiency Virus (HIV) fusion inhibitors:

(1) The N-terminal 'RQLL' Motif: The model assigned the highest importance scores to Leucine at position 3 (L3) and Leucine at position 4 (L4), which are part of the N-terminal 'RQLL' motif. This motif has been experimentally confirmed to be indispensable for fusion inhibitory activity, mutations in this motif completely abolish the peptide's activity, underscoring its critical role in the initial binding and stabilization of the inhibitor-target complex [41].
(2) Notably, our model identified residues 1, 3 and 8, 10 as critically important. These correspond to the 'e' and 'g' positions within the first and second heptad repeat units of the gp41 structure. These specific positions are located on the functional binding surface of the peptide and are directly involved in interactions with the target C-terminal heptad repeat (CHR) region of HIV gp41. This interaction is the fundamental mechanism by which this class of peptide inhibitors blocks the formation of the six-helix bundle, thereby preventing viral entry [41].

To further examine whether similar residue-level importance patterns were observed across different AVPs, we additionally analyzed peptide DRAVPe00681, which originates from a well-characterized class of T-20–derived lipopeptide HIV fusion inhibitors. As shown in Fig. 8, the Grad-CAM importance map of DRAVPe00681 reveals a pronounced enrichment of high-importance scores at its N-terminal region. In particular, residues corresponding to the N-terminal WEQK motif were consistently highlighted as critical by the model.

Notably, this finding is in strong agreement with previous structure–function and truncation studies, which demonstrated that the N-terminal WEQK motif is indispensable for maintaining potent and broad-spectrum antiviral activity, especially against T-20–resistant HIV-1 strains, HIV-2, and SIV. Experimental removal of this motif resulted in orders-of-magnitude loss of inhibitory activity, underscoring its functional significance [42]. The concordance between Grad-CAM–identified important residues and experimentally validated functional motifs

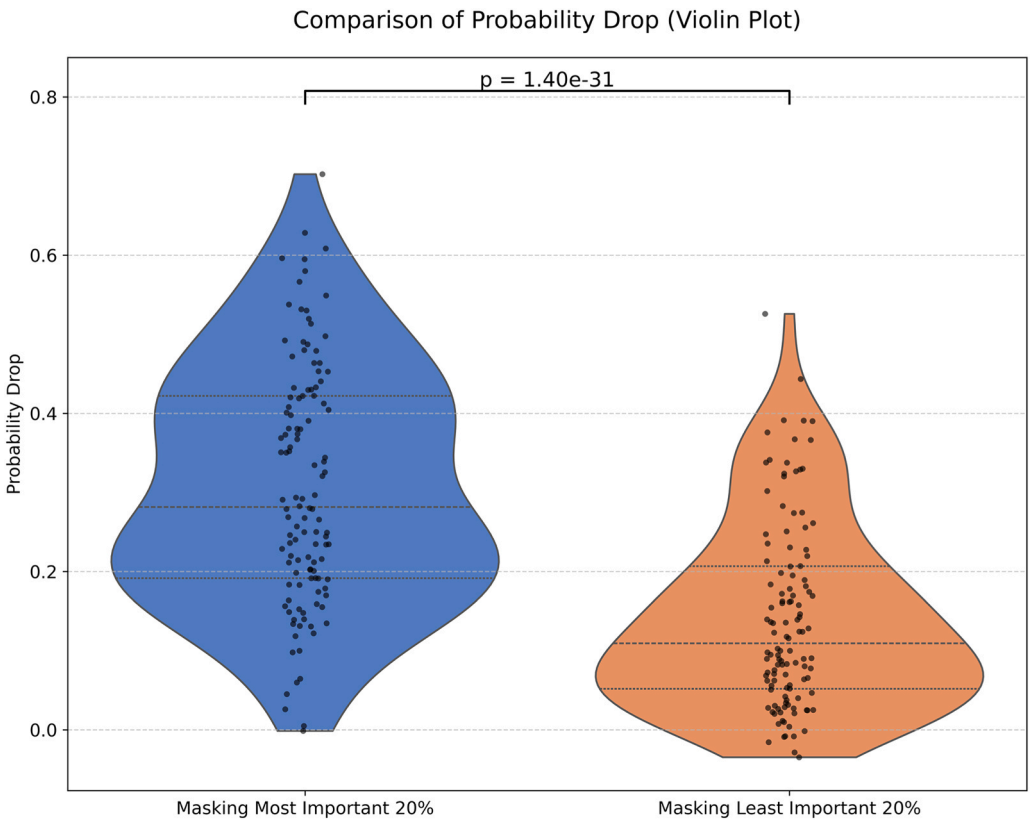

Fig. 9. Violin plot analysis of prediction probability changes upon masking top and bottom 20% important residues. Blue violin (left) represents the distribution of probability drops when masking the most important 20% of amino acids, while orange violin (right) shows the effect of masking the least important 20%, with highly significant separation between conditions ($p = 1.40 \times 10^{-31}$).

Table 4
Quartile comparison of prediction probability drops after masking top vs. bottom 20% important residues.

Statistic	Masking_Most_Important_20%	Masking_Least_Important_20%
25th_Percentile	0.1916	0.0519
50th_Percentile (Median)	0.2815	0.1094
75th_Percentile	0.4223	0.2067

in DRAVPe00681 further supports the generalizability and biological relevance of our residue-level interpretability analysis beyond a single representative peptide.

This compelling concordance between in silico prediction and experimental biological evidence demonstrates that our GAT-Grad-CAM framework does not merely make accurate predictions but effectively pinpoints the amino acid residues that govern biological function. It provides a powerful tool for de novo interpretation of AVP sequences and can guide the rational design of future inhibitors by focusing mutagenesis or modification efforts on these high-importance residues.

To test whether the Grad-CAM importance analysis was statistically significant, we included peptides in the statistical analysis if they met all of the following criteria: (1) the sequence length was greater than 15 amino acids; (2) the predicted probability was between 0.9 and 0.99; and (3) the model's prediction matched the true label. We then performed a paired *t*-test on these peptides. Our statistical validation confirms Grad-CAM reliably identifies important residues. The violin plot (Fig. 9) clearly shows masking the top 20% important residues (left distribution) causes much larger prediction drops than masking the bottom 20% unimportant residues (right distribution). This visual difference is striking - median drops were 0.2815 for important residues versus just 0.1094 for unimportant ones (Table 4). The pattern holds across all quartiles (25th percentile: 0.1916 vs 0.0519; 75th percentile: 0.4223 vs 0.2067). A paired *t*-test confirmed this difference is extremely significant ($p = 1.40e\text{-}31$). These results suggest that the Grad-CAM scores reliably identify residues that are critical for the model's prediction. The ability to highlight these key residues may help researchers better understand the model's decision process and guide future peptide design.

4. Conclusion

In this study, we introduced AVP-3LGAT, a structure-aware and sequence-informed model for predicting AVPs. Unlike most existing AVP prediction models, our approach combines semantic features from a protein language model (ESMC) with 3D structural relationships predicted by ESMFold. We showed that representing peptides as residue-level graphs allows the model to learn more informative patterns. Our 3-layer GAT architecture, paired with multiscale pooling, effectively captures both local and global residue interactions. On a highly imbalanced test set, the AVP-3LGAT achieved strong performance, especially in reducing false positives. This is critical for real-world screening where AVPs are rare. We also applied Grad-CAM to understand which residues contributed most to the model's decisions. These interpretability results help reveal how the model links peptide structure and function. Overall, the AVP-3LGAT offers a reliable and interpretable solution for AVP prediction. These findings can support early-stage antiviral research and guide the design of novel AVP therapeutics.

Author Contributions

YC conducted the experiments and drafted the manuscript. MAML and MBAR participated in experimental design, formal analysis, and manuscript drafting. BAT conceptualized and supervised the study. All

authors read and approved the final manuscript.

CRediT authorship contribution statement

Mohd Basyaruddin Abdul Rahman: Writing – original draft, Supervision, Formal analysis. **Muhammad Alif Mohammad Latif:** Writing – original draft, Supervision, Formal analysis. **Yuan Chongjun:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation. **Bimo A. Tejo:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Universiti Putra Malaysia under grant GP-IPS/2023/9747200.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.aichem.2026.100114.

Data availability

All the data and code used for this study are publicly available at https://github.com/YuanColab/AVP-3LGAT

References

[1] M.M. Islam, D. Koirala, Toward a next-generation diagnostic tool: A review on emerging isothermal nucleic acid amplification techniques for the detection of SARS-CoV-2 and other infectious viruses, Anal. Chim. Acta 1209 (2022) 339338, https://doi.org/10.1016/j.aca.2021.339338.
[2] T. Phan, Genetic diversity and evolution of SARS-CoV-2, Infect. Genet Evol. 81 (2020) 104260, https://doi.org/10.1016/j.meegid.2020.104260.
[3] S. Calvignac-Spencer, A. Dux, J.F. Gogarten, F.H. Leendertz, L.V. Patrono, A great ape perspective on the origins and evolution of human viruses, Adv. Virus Res 110 (2021) 1–26, https://doi.org/10.1016/bs.aivir.2021.06.001.
[4] E. De Clercq, G. Li, Approved Antiviral Drugs over the Past 50 Years, Clin. Microbiol Rev. 29 (2016) 695–747, https://doi.org/10.1128/CMR.00102-15.
[5] A. Hollmann, N.P. Cardoso, J.C. Espeche, P.C. Maffia, Review of antiviral peptides for use against zoonotic and selected non-zoonotic viruses, Peptides 142 (2021) 170570, https://doi.org/10.1016/j.peptides.2021.170570.
[6] A.S.K. Mahendran, Y.S. Lim, C.M. Fang, H.S. Loh, C.F. Le, The Potential of Antiviral Peptides as COVID-19 Therapeutics, Front Pharm. 11 (2020) 575444, https://doi.org/10.3389/fphar.2020.575444.
[7] L. Wang, N. Wang, W. Zhang, X. Cheng, Z. Yan, G. Shao, X. Wang, R. Wang, C. Fu, Therapeutic peptides: current applications and future directions, Signal Transduct. Target Ther. 7 (2022) 48, https://doi.org/10.1038/s41392-022-00904-4.
[8] L.C.P. Vilas Boas, M.L. Campos, R.L.A. Berlanda, N. de Carvalho Neves, O. L. Franco, Antiviral peptides as promising therapeutic drugs, Cell Mol. Life Sci. 76 (2019) 3525–3542, https://doi.org/10.1007/s00018-019-03138-w.
[9] N. Thakur, A. Qureshi, M. Kumar, AVPpred: collection and prediction of highly effective antiviral peptides, Nucleic Acids Res 40 (2012) W199–204, https://doi.org/10.1093/nar/gks450.
[10] T.T. Lin, Y.Y. Sun, C.T. Wang, W.C. Cheng, I.H. Lu, C.Y. Lin, S.H. Chen, AI4AVP: an antiviral peptides predictor in deep learning approach with generative adversarial network data augmentation, Bioinform Adv. 2 (2022) vbac080, https://doi.org/10.1093/bioadv/vbac080.
[11] P. Charoenkwan, P. Chumnanpuen, N. Schaduangrat, W. Shoombuatong, Stack-AVP: A Stacked Ensemble Predictor Based on Multi-view Information for Fast and Accurate Discovery of Antiviral Peptides, J. Mol. Biol. 437 (2025) 168853, https://doi.org/10.1016/j.jmb.2024.168853.
[12] Y. Li, A. Geng, Z. Zhou, F. Cui, J. Xu, Y. Meng, L. Wei, Q. Zou, Q. Zhang, Z. Zhang, AVP-HNCL: Innovative Contrastive Learning with a Queue-Based Negative Sampling Strategy for Dual-Phase Antiviral Peptide Prediction, J. Chem. Inf. Model 65 (2025) 5868–5886, https://doi.org/10.1021/acs.jcim.5c00306.
[13] S. Akbar, A. Raza, Q. Zou, W. Alghamdi, X. Kang, H. Ali, X. Luo, Accelerating Prediction of Antiviral Peptides Using Genetic Algorithm-Based Weighted Multiperspective Descriptors with Self-Normalized Deep Networks, J. Chem. Inf. Model 65 (2025) 9815–9830, https://doi.org/10.1021/acs.jcim.5c01777.
[14] M. Ullah, S. Akbar, A. Raza, Q. Zou, DeepAVP-TPPred: identification of antiviral peptides using transformed image-based localized descriptors and binary tree growth algorithm (pp.), Bioinformatics 40 (2024), https://doi.org/10.1093/bioinformatics/btae305.
[15] S. Akbar, A. Raza, Q. Zou, Deepstacked-AVPs: predicting antiviral peptides using tri-segment evolutionary profile and word embedding based multi-perspective features with deep stacking model, BMC Bioinforma. 25 (2024) 102, https://doi.org/10.1186/s12859-024-05726-5.
[16] S. Akbar, F. Ali, M. Hayat, A. Ahmad, S. Khan, S. Gul, Prediction of antiviral peptides using transform evolutionary & SHAP analysis based descriptors by incorporation with ensemble learning strategy, Chemom. Intell. Lab. Syst. 230 (2022) 104682, https://doi.org/10.1016/j.chemolab.2022.104682.
[17] M. Nawaz, Y. Huiyuan, F. Akhtar, M. Tianyue, H. Zheng, Deep learning in the discovery of antiviral peptides and peptidomimetics: databases and prediction tools (pp.), Mol. Divers (2025), https://doi.org/10.1007/s11030-025-11173-y.
[18] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S.A. A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature 596 (2021) 583–589, https://doi.org/10.1038/s41586-021-03819-2.
[19] D. Ofer, N. Brandes, M. Linial, The language of proteins: NLP, machine learning & protein sequences, Comput. Struct. Biotechnol. J. 19 (2021) 1750–1758, https://doi.org/10.1016/j.csbj.2021.03.022.
[20] V. Petar, C. Guillem, C. Arantxa, R. Adriana, L. Pietro, B. Yoshua, Graph attention networks, International conference on learning representations, (2018).
[21] K. Yan, H. Lv, Y. Guo, W. Peng, B. Liu, sAMPpred-GAT: prediction of antimicrobial peptide by graph attention network and predicted peptide structure (pp.), Bioinformatics 39 (2023), https://doi.org/10.1093/bioinformatics/btac715.
[22] G. Cordoves-Delgado, C.R. Garcia-Jacas, Predicting Antimicrobial Peptides Using ESMFold-Predicted Structures and ESM-2-Based Amino Acid Features with Graph Deep Learning, J. Chem. Inf. Model 64 (2024) 4310–4321, https://doi.org/10.1021/acs.jcim.3c02061.
[23] Y. Hao, X. Liu, H. Fu, X. Shao, W. Cai, PGAT-ABPp: harnessing protein language models and graph attention networks for antibacterial peptide identification with remarkable accuracy (pp.), Bioinformatics 40 (2024), https://doi.org/10.1093/bioinformatics/btae497.
[24] A. Qureshi, N. Thakur, H. Tandon, M. Kumar, AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses, Nucleic Acids Res 42 (2014) D1147–1153, https://doi.org/10.1093/nar/gkt1191.
[25] Q. Zhang, X. Chen, B. Li, C. Lu, S. Yang, J. Long, H. Chen, J. Huang, B. He, A database of anti-coronavirus peptides, Sci. Data 9 (2022) 294, https://doi.org/10.1038/s41597-022-01394-3.
[26] Y. Liu, Y. Zhu, X. Sun, T. Ma, X. Lao, H. Zheng, DRAVP: A Comprehensive Database of Antiviral Peptides and Proteins (pp.), Viruses 15 (2023), https://doi.org/10.3390/v15040820.
[27] A. Qureshi, N. Thakur, M. Kumar, HIPdb: a database of experimentally validated HIV inhibiting peptides, PLoS One 8 (2013) e54908, https://doi.org/10.1371/journal.pone.0054908.
[28] U. Gawde, S. Chakraborty, F.H. Waghu, R.S. Barai, A. Khanderkar, R. Indraguru, T. Shirsat, S. Idicula-Thomas, CAMPR4: a database of natural and synthetic antimicrobial peptides, Nucleic Acids Res 51 (2023) D377–D383, https://doi.org/10.1093/nar/gkac933.
[29] M. Pirtskhalava, A.A. Amstrong, M. Grigolava, M. Chubinidze, E. Alimbarashvili, B. Vishnepolsky, A. Gabrielian, A. Rosenthal, D.E. Hurt, M. Tartakovsky, DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics, Nucleic Acids Res 49 (2021) D288–D297, https://doi.org/10.1093/nar/gkaa991.
[30] L. Yao, J. Guan, P. Xie, C.R. Chung, Z. Zhao, D. Dong, Y. Guo, W. Zhang, J. Deng, Y. Pang, Y. Liu, Y. Peng, J.T. Horng, Y.C. Chiang, T.Y. Lee, dbAMP 3.0: updated resource of antimicrobial activity and structural annotation of peptides in the post-pandemic era, Nucleic Acids Res 53 (2025) D364–D376, https://doi.org/10.1093/nar/gkae1019.
[31] T. Ma, Y. Liu, B. Yu, X. Sun, H. Yao, C. Hao, J. Li, M. Nawaz, X. Jiang, X. Lao, H. Zheng, DRAMP 4.0: an open-access data repository dedicated to the clinical translation of antimicrobial peptides, Nucleic Acids Res 53 (2025) D403–D410, https://doi.org/10.1093/nar/gkae1046.
[32] G. Cabas-Mora, A. Daza, N. Soto-Garcia, V. Garrido, D. Alvarez, M. Navarrete, L. Sarmiento-Varon, J.H. Sepulveda Yanez, M.D. Davari, F. Cadet, A. Olivera-Nappa, R. Uribe-Paredes, D. Medina-Ortiz, Peptipedia v2.0: a peptide sequence database and user-friendly web platform. A major update (pp.), Database (Oxf.) 2024 (2024), https://doi.org/10.1093/database/baae113.
[33] S. Ahmad, L. Jose da Costa Gonzales, E.H. Bowler-Barnett, D.L. Rice, M. Kim, S. Wijerathne, A. Luciani, S. Kandasaamy, J. Luo, X. Watkins, E. Turner, M. J. Martin, C. UniProt, The UniProt website API: facilitating programmatic access to protein knowledge (pp.), Nucleic Acids Res (2025), https://doi.org/10.1093/nar/gkaf394.
[34] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT Suite: a web server for clustering and comparing biological sequences, Bioinformatics 26 (2010) 680–682, https://doi.org/10.1093/bioinformatics/btq003.

[35] E. Team, ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning, EvolutionaryScale Website, 2024.
[36] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. Dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model, Science 379 (2023) 1123–1130, https://doi.org/10.1126/science.ade2574.
[37] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[38] L. McInnes, J. Healy, J. Melville, UMAP: uniform manifold approximation and projection for dimension reduction. arXiv, arXiv preprint arXiv:1802.03426, 10 (2018).
[39] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, AAindex: amino acid index database, progress report 2008, Nucleic Acids Res 36 (2008) D202–205, https://doi.org/10.1093/nar/gkm998.
[40] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, Proc. IEEE Int. Conf. Comput. Vis. (2017) 618–626.
[41] Y. Wexler-Cohen, A. Ashkenazi, M. Viard, R. Blumenthal, Y. Shai, Virus-cell and cell-cell fusion mediated by the HIV-1 envelope glycoprotein is inhibited by short gp41 N-terminal membrane-anchored peptides lacking the critical pocket domain, Faseb J. 24 (2010) 4196–4202, https://doi.org/10.1096/fj.09-151704.
[42] H. Chong, Y. Zhu, D. Yu, Y. He, Structural and Functional Characterization of Membrane Fusion Inhibitors with Extremely Potent Activity against Human Immunodeficiency Virus Type 1 (HIV-1), HIV-2, and Simian Immunodeficiency Virus (pp.), J. Virol. 92 (2018), https://doi.org/10.1128/jvi.01088-18.